Bioinformatics, 31(7), 2015, 1034–1043
doi: 10.1093/bioinformatics/btu766
Advance Access Publication Date: 18 November 2014
Original Paper

Systems biology

HyDRA: gene prioritization via hybrid distance-score rank aggregation

Minji Kim*, Farzad Farnoud and Olgica Milenkovic

Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on June 2, 2014; revised on October 26, 2014; accepted on November 13, 2014

Abstract

Summary: Gene prioritization refers to a family of computational techniques for inferring disease genes through a set of training genes and carefully chosen similarity criteria. Test genes are scored based on their average similarity to the training set, and the rankings of genes under various similarity criteria are aggregated via statistical methods. The contributions of our work are threefold: (i) first, based on the realization that there is no unique way to define an optimal aggregate for rankings, we investigate the predictive quality of a number of new aggregation methods and known fusion techniques from machine learning and social choice theory. Within this context, we quantify the influence of the number of training genes and similarity criteria on the diagnostic quality of the aggregate and perform in-depth cross-validation studies; (ii) second, we propose a new approach to genomic data aggregation, termed *HyDRA* (Hybrid Distance-score Rank Aggregation), which combines the advantages of score-based and combinatorial aggregation techniques. We also propose incorporating a new *top-versus-bottom* (TvB) weighting feature into the hybrid schemes. The TvB feature ensures that aggregates are more reliable at the top of the list, rather than at the bottom, since only top candidates are tested experimentally; (iii) third, we propose an iterative procedure for gene discovery that operates via successful augmentation of the set of training genes by genes discovered in previous rounds, checked for consistency.

Motivation: Fundamental results from social choice theory, political and computer sciences, and statistics have shown that there exists no consistent, fair and unique way to aggregate rankings. Instead, one has to decide on an aggregation approach using predefined set of desirable properties for the aggregate. The aggregation methods fall into two categories, score- and distance-based approaches, each of which has its own drawbacks and advantages. This work is motivated by the observation that merging these two techniques in a computationally efficient manner, and by incorporating additional constraints, one can ensure that the predictive quality of the resulting aggregation algorithm is very high.

Results: We tested HyDRA on a number of gene sets, including autism, breast cancer, colorectal cancer, endometriosis, ischaemic stroke, leukemia, lymphoma and osteoarthritis. Furthermore, we performed iterative gene discovery for glioblastoma, meningioma and breast cancer, using a sequentially augmented list of training genes related to the Turcot syndrome, Li-Fraumeni condition and other diseases. The methods outperform state-of-the-art software tools such as ToppGene and Endeavour. Despite this finding, we recommend as best practice to take the union of top-ranked items produced by different methods for the final aggregated list.

Availability and implementation: The HyDRA software may be downloaded from: http://web.engr.illinois.edu/∼mkim158/HyDRA.zip

Contact: mkim158@illinois.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Identification of genes that predispose an individual to a disease is a problem of great interest in medical sciences and systems biology (Adie *et al.*, 2006). The most accurate and powerful methods used for identification are experimental in nature, involving normal and disease samples (Cardon *et al.*, 2001). Experiments are time consuming and costly, complicated by the fact that typically, multiple genes have to be jointly mutated to trigger the onset of a disease. Given the large number of human genes ($\geq$25 000), testing even relatively small subsets of pairs of candidate genes is prohibitively expensive (Risch and Merikangas, 1996).

To mitigate this issue, a set of predictive analytical and computational methods have been proposed under the collective name *gene prioritization techniques*. Gene prioritization refers to the complex procedure of ranking genes according to their likelihoods of being linked to a certain disease. The likelihood function is computed based on multiple sources of evidence, such as sequence similarity, linkage analysis, gene annotation, functionality and expression activity, gene product attributes—all determined with respect to a set of training genes.

A wide range of tools has been developed for identifying genes involved in a disease (Köhler *et al.*, 2008; Kolde *et al.*, 2012; Pihur *et al.*, 2009), as surveyed (Tiffin *et al.*, 2006). Existing software includes techniques based on network information, such as GUILDify (Guney *et al.*, 2014) and GeneMANIA (Warde-Farley *et al.*, 2010), data mining and machine learning-based approaches as described in (Perez-Iratxeta *et al.*, 2002), POCUS (Turner *et al.*, 2003) SUSPECTS (Adie *et al.*, 2006) and (Yu *et al.*, 2008), and methods using statistical analysis, including Endeavour (Aerts *et al.*, 2006; De Bie *et al.*, 2007), ToppGene (Chen *et al.*, 2009) and NetworkPrioritizer (Kacprowski *et al.*, 2013). Here, we focus on statistical approaches coupled with new combinatorial algorithms for gene prioritization, and emphasize one aspect of the prioritization procedure: rank aggregation.

The problem of aggregating rankings of distinct objects or entities provided by a number of experts, voters or search engines has a rich history (Fishburn, 1970). One of the key findings is that various voting paradoxes arise when more than three candidates are to be ranked: it is frequently possible not to have a candidate that wins all pairwise competitions (the Condorcet paradox) and it is theoretically impossible to guarantee the existence of an aggregate solution that meets certain predefined set of criteria [such as those imposed by Arrow's impossibility theorem (Fishburn, 1970)]. These issues carry over to aggregation methods used for gene discovery, and as a result, the rank-ordered lists of genes heavily depend on the particular aggregation method used.

Two families of methods have found wide applications in rank aggregation: combinatorial methods (including score- and distance-based approaches) (Kemeny, 1959) and statistical methods. In the bioinformatics literature, the aggregation methods of choice are statistical in nature, relying on pre-specified hypotheses to evaluate the distribution of the gene rankings. One of the earliest prioritization softwares, Endeavour, uses the Q-statistics for multiple significance testing, and measures the minimum false discovery rate at which a test may be called significant. In particular, rankings based on different similarity criteria are combined via order statistics approaches. For this purpose, one uses the rank ratio (normalized ranking) of a gene g for m different criteria, $r_1(g),\ldots,r_m(g)$ and recursively computes the Q-value, defined as

$$Q_i(r_1(g),\ldots,r_m(g)) = m!\int_0^{r_1(g)}\int_{s_1}^{r_2(g)}\ldots\int_{s_{m-1}}^{r_m(g)} d_{s_m} d_{s_{m-1}}\ldots d_{s_1}.$$

Post-processed Q-values are used to create the resulting ranking of genes. The drawbacks of the method are that it is based on a null hypothesis that is difficult to verify in practice, and that it is computationally expensive, as it involves evaluating an m-fold integral. To enable efficient scaling of the method, Endeavour resorts to approximating the Q-integral. The influence of the approximation errors on the final ranking is hard to assess, as small changes in scores may result in significant changes of the aggregate orderings.

Likewise, ToppGene uses a well-known statistical approach, called the Fisher χ^2 method. It first determines the p-values of similarity score indexed by j, denoted by $p(j)$, for $j = 1,\ldots,m$. The p-values are computed through multiple preprocessing stages, involving estimation of the information contents (i.e. weights) of annotation terms, setting-up a similarity criteria based on Sugeno fuzzy measures (i.e. non-additive measures) (Popescu *et al.*, 2006), and performing meta-testing. The use of fuzzy measures ensures that all similarities are non-negative. Then, under the hypothesis of *independent tests*, ToppGene uses Fisher's inverse χ^2 result, stating that $-2\sum_{j=1}^{m}\log p(j) \rightarrow \chi^2(2m)$. Here, χ^2 $(2m)$ stands for the chi-square distribution with $2m$ degrees of freedom. The result is asymptotic in nature, and based on possibly impractical independence assumptions.

A number of methods, and additive scoring methods in particular, have the drawback that they tacitly or implicitly rely on the assumption that (i) only the total score matters, and the balance between the number of criteria that highly ranked the gene and those that ranked it very low is irrelevant. For example, outlier rankings may reduce the overall ranking of a gene to the point that it is not considered a disease gene candidate, while the outlier itself may be a problematic criterion. To illustrate this observation, consider a gene that was ranked 1st, 2nd, 1st, 20th by four criteria. At the same time, consider another gene that was ranked 6th by all four criteria. It may be unclear which of these two genes is more likely to be involved in the disease, given that additive score methods would rank the two genes equally (as one has $(1+2+1+20)/4=6$). Nevertheless, it appears reasonable to assume that the first candidate is a more reliable choice for a disease gene, as it had a very high ranking for three out of four criteria; and (ii) no distinction is made about the accuracy of ranking genes in any part of the list; i.e. the aggregate ranking has to be *uniformly accurate* at the top, middle and bottom of the list. Clearly, neither of the two aforementioned assumptions is justified in the gene prioritization process: there are many instances where genes similar only under a few criteria (such as sequence similarity or linkage distance) are involved in the same disease pathway. Furthermore, as the goal of prioritization is to produce a list of genes to be experimentally tested, only the highest ranked candidate genes are important and should have higher accuracy than other genes in the list. In addition, most known aggregation methods are highly sensitive to outliers and ranking errors.

We propose a new approach to gene prioritization by introducing a number of novel aggregation paradigms, which we collectively refer to as *HyDRA* (Hybrid Distance-score Rank Aggregation). The gist of HyDRA is to combine *combinatorial approaches* that have universal axiomatic underpinnings with *statistical evidence* pertaining to the accuracy of individual rankings. Our preferred distance measure for combinatorial aggregation is the Kendall distance (Kendall, 1938), which counts the number of pairwise disagreement between two rankings, and was axiomatically postulated by Kemeny (1959). The Kendall distance is closely related to the Kendall rank correlation coefficient (Dwork *et al.*, 2001; Kendall, 1948). As such, it has many properties useful for gene prioritization, such as monotonicity, reinforcement and Pareto efficiency (Thanassoulis, 2001). The Kendall distance can be generalized to take into account positional relevance of items, as was done in our companion article (Farnoud *et al.*, 2012, 2014). There, it was shown that by assigning weights to pairs of positions in rankings, it is possible to (i) eliminate negative outliers from the aggregation process, (ii) include quantitative data into the aggregate and (iii) ensure higher accuracy at the top of the ranking than at the bottom.

The contributions of this work are threefold. First, we introduce new weighted distance measures, where we compute the weights based on statistical evidence of a function of the difference between *p*-values of adjacently ranked items. Aggregation weights based on statistical evidence improve the accuracy of the combinatorial aggregation procedure and make them more robust to estimation errors. Second, we describe how to scale the weights obtained based on statistical evidence by a decreasing sequence of TvB (Top versus Bottom) multipliers that ensure even higher accuracy at the top of the aggregated list. As aggregation under the Kendall metric is NP-hard (Non-deterministic Polynomial-time hard) (Bartholdi *et al.*, 1989), and the same is true of the weighted Kendall metric, we propose a 2-approximation method that is stable under small perturbations. Aggregation is accomplished via weighted bipartite matching, such as the Hungarian algorithm and derivatives thereof (Kuhn, 1955). Third, we test HyDRA within two operational scenarios: cross-validation and disease gene discovery. In the former case, we assess the performance of different hybrid methods with respect to the choice of the weighting function and different number of test and training genes. In the latter case, we adapt aggregation methods to gene discovery via a new iterative re-ranking procedure.

2 Systems and methods

In our subsequent exposition, we use Greek lower case letters to denote complete linear orders (permutations), and unless explicitly mentioned otherwise, our findings also hold for partial (incomplete) permutations. Latin lower case letters are reserved for score vectors or scalar scores, and which of these entities we refer to will be clear from the context. The number of test genes equals n, while the number of similarity criteria equals m. Throughout the article, we also use $[k]$ to denote the set $\{1,\ldots,k\}$ and $\mathbb{S}_n$ to denote the set of all permutations on n elements—the symmetric group of order $n!$.

For a permutation $\sigma = (\sigma(1),\ldots,\sigma(n))$, the rank of element i in σ, $\mathrm{rank}_\sigma(i)$, equals $\sigma^{-1}(i)$, where σ^{-1} denotes the inverse permutation of σ. For a vector of scores $x = (x(i))_{i=1}^{\mathrm{n}} \in \mathbb{R}^{\mathrm{n}}$, σ_{x} represents a permutation describing the scores in decreasing order, i.e. $\sigma_{\mathrm{x}}(i) = \mathrm{argmax}_{k\in T_{\mathrm{i}}}\, x(k)$, where T_{i} is defined recursively as $T_{\mathrm{i}} = T_{i-1} \setminus \sigma_{\mathrm{x}}(i)$, with $T_0 = [n]$. For example, if $x = (2.5,\ 3.8,\ 1.1,\ 0.7)$, then $\sigma_x = (2,1,3,4)$. Note that if p is a vector of p-values, higher scores are associated with smaller p-values, so that argmax should be replaced by argmin.

The terms *gene* and *element* are used interchangeably, and each permutation is tacitly assumed to be produced by one similarity criteria. For a set of permutations $\Sigma = \{\sigma_1,\ldots,\sigma_m\}$, $\sigma_{\mathrm{i}} = (\sigma_{\mathrm{i}}(1),\ldots,\sigma_{\mathrm{i}}(n))$, an *aggregate permutation* σ^* is a permutation that optimally represents the rankings in Σ. Combinatorial aggregates may be obtained using *score-* and *distance-based* methods. Note that score and distance-based methods do not make use of quantitative information, such as, e.g., p-values (for the case of gene prioritization) or ratings (for the case of social choice theory and recommender systems). In what follows, we briefly describe score and distance-based methods and introduce their *hybrid* counterparts, which allow to integrate p-values and relevance constraints into combinatorial aggregation approaches.

2.1 Score-based methods

Score-based methods are the simplest and computationally least demanding techniques for rank aggregation. As inputs, they take a set of permutations or partial permutations, $\Sigma = \{\sigma_1,\ldots,\sigma_m\}$, $\sigma_{\mathrm{i}} = (\sigma_{\mathrm{i}}(1),\ldots,\sigma_{\mathrm{i}}(n))$. For each permutation $\sigma_i \in \Sigma$, the scoring rule awards $s(\sigma_{\mathrm{i}}(1), i)$ points to element $\sigma_{\mathrm{i}}(1)$, $s(\sigma_{\mathrm{i}}(2), i)$ points to element $\sigma_{\mathrm{i}}(2)$, and so on. For a fixed i, the scores are non-increasing functions of their first index. Each element $k \in [n]$ is assigned a cumulative score equal to $\sum_{j=1}^{m} s(k,j)$. The simplest scoring method is Borda's count, for which $s(k,j) = n - k + 1$ independent on j.

The Borda count and related scoring rules exclusively use positional information in order to provide an aggregate ranking. Ignoring actual p-values (ratings) may lead to aggregation problems, as illustrated by the next example.

Example 1: Assume that $n = 5$ elements were rated according to $x = (7.0, 7.01, 0.2, 0.45, 7.001)$. The ranking induced by this rating equals $\sigma_x = (2,5,1,4,3)$, indicating that element 2 received the highest rating, element 5 received the second highest rating and so on. According to the Borda rule, element 2 receives 5 points, element 5 receives 4 points, etc. Despite the fact that candidates 2 and 1 are almost tied with scores of 7.01 and 7.0, and that the difference in their scores may be attributed to computational imprecision, element 2 receives 5 points while element 1 receives only 3 points. As a result, very small differences in ratings may result in large differences in Borda scores.

One way to approach the problem is to quantize the score and work with rankings with ties, instead of full linear orders (i.e. permutations). Elements tied in their rank receive the same number of points in the generalized Borda scheme. A preferred alternative, which we introduce in this work, is the *Hybrid Borda method*.

Let $p(i,j)$ denote the p-value of gene i computed under similarity criteria $j, j = 1,\ldots,m$. The cumulative score of element i in the hybrid Borda setting is computed as

$$S_{\mathrm{i}} = \sum_{j=1}^{m}\left(\frac{\sum_{k\neq i}^{m} p(k,j)\mathbb{1}_{\{p(k,j)\geq p(i,j)\}}}{p(i,j)}\right).$$

The overall aggregate is obtained by ordering S in a descending order. It is straightforward to see that the previous score function extends Borda method in so far that it scores an element (gene) according to the total score of elements ranked lower than the element. Recall that in Borda's method, the element ranked i is awarded $n - i + 1$ points, as $n - i + 1$ elements are ranked below it, each receiving the same score 1. In our Hybrid Borda method, each element is awarded a score in *accordance with the p-values of elements ranked below it*.

Example 2: Let $n = 4$ and $m = 2$, where the two ratings equal to $p_1 = (0.2, 0.3, 0.01, 0.12)$ and $p_2 = (0.1, 0.4, 0.2, 0.35)$. The Hybrid Borda scores S_i for genes $i = 1, 2, 3, 4$ are computed as $S_1 = 0.3/0.2 + (0.4 + 0.2 + 0.35)/0.1 = 11$, $S_2 = 0$, $S_3 = (0.2 + 0.3 + 0.12)/0.01 + (0.4 + 0.35)/0.2 = 65.75$ and $S_4 = (0.2 + 0.3)/0.12 + 0.4/0.35 = 5.3$. By ordering the values S_i in a descending manner, we obtain the overall aggregate $\sigma_{HB} = (3, 1, 4, 2)$.

The hybrid Borda method can be extended further by adding a TvB feature, resulting in the *Weighted* hybrid Borda method. This is accomplished by including *increasing* (multiplier) weights into the score aggregates, thus stressing the top of the list more than the bottom. More precisely, the score of gene i is computed as:

$$S_i = \sum_{j=1}^{m} \left(\frac{\sum_{k \neq i} wm(k,j) p(k,j) \mathbb{1}_{\{p(k,j) \geq p(i,j)\}}}{wm(i,j) p(i,j)} \right).$$

where one simple choice for the weight multipliers that provides good empirical performance equals

$$wm(i,j) = \frac{1}{n - \text{rank}_{\sigma_j}(i) + 1}.$$

2.2 Distance-based methods

Another common approach to rank aggregation is distance-based rank aggregation. As before, assume that one is given a set of permutations $\Sigma = \{\sigma_1, \ldots, \sigma_m\}$. For a given distance function between two permutations σ and π, $d(\sigma, \pi)$, aggregation reduces to

$$\pi = \arg\min_{\sigma} \sum_{i=1}^{m} d(\sigma, \sigma_i)$$

The aggregate π is frequently referred to as the *median* of the permutations, and is illustrated in Figure 1.

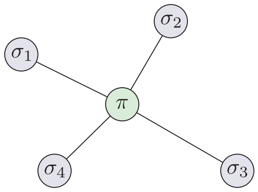

Fig. 1. Four rankings: $\sigma_1, \sigma_2, \sigma_3, \sigma_4$ and their aggregate (median) ranking π

One of the most important features of distance-based approaches is the choice of the distance function. Table 1 lists two of the most frequently used distances, the Kendall tau distance and the Spearman footrule. As may be seen from the table, the distance measures are combinatorial in nature, and do not account for scores or p-values. Furthermore, as already mentioned in the introduction, it is known that aggregation under the Kendall metric is computationally hard. Nevertheless, there exists a number of techniques which provide provable *approximation* guarantees for the aggregate, including the weighted Bipartite Graph Matching (WBGM) method (using the fact that the Spearman distance aggregate is a 2-approximation for the Kendall aggregate), linear programing (LP) relaxation and Page Rank/Markov chain (PR) methods (Dwork *et al.*, 2001; Farnoud *et al.*, 2012; Raisali *et al.*, 2013).

Table 1. Two frequently used distance measures for permutations, accounting for swaps or element-wise differences

Distance	Measurement	Example
Spearman's footrule	Sum of differences of ranks of elements.	$d_F(abc, cba) = 2 + 0 + 2 = 4$
Kendall	Minimum number of adjacent swaps of entries for transforming one ranking into another.	$d_K(abc, cba) = 3$

In the second example, the Kendall tau distance between the permutation $\sigma_1 = (a, b, c)$ and $\sigma_2 = (c, b, a)$ equals 3: one first swaps elements at positions 1 and 2 to get (b, a, c), then elements at positions 2 and 3 to get (b, c, a), and finally elements at positions 1 and 2 to get $\sigma_2 = (c, b, a)$. All swaps contribute the same weight (one) to the distance.

The Kendall distance also does not take into account the fact that the top of a list is more important than the remainder of the list. To overcome this problem, we introduced the notion of *weighted Kendall distances*, where each adjacent swap is assigned a cost, and where the cost is higher at the top of a list. This ensures that in an aggregate, strong showings of candidates are emphasized compared with their weaker showings, accounting for the fact that it is often sufficient to have strong similarity with respect to only a subset of criteria. Furthermore, such weights ensure that higher importance is paid to the top of the aggregate ranking.

The idea behind the weighted Kendall distance d_w is to compute this distance as the shortest path in a graph describing swap relationships between permutations. The key concepts are illustrated in Figures 2 and 3, where each edge is assigned a length proportional to its weight W. This weight depends on the swap being made at the top or at some other position in the ranking. Given that it is computationally demanding to aggregate under the weighted Kendall distance, we use a specialized approximation function $D_w(\sigma, \theta)$ for d_w, of the form

$$D_w(\sigma, \theta) = \sum_{i=1}^{n} w(\sigma^{-1}(i) : \theta^{-1}(i)), \tag{1}$$

where

$$w(k:l) = \begin{cases} \sum_{h=k}^{l-1} W(h, h+1), & \text{if } k < l, \\ \sum_{h=l}^{k-1} W(h, h+1), & \text{if } k > l, \\ 0, & \text{if } k = l, \end{cases} \tag{2}$$

denotes the sum of the weights of edges $W(\cdot)$ representing adjacent transpositions $(k\,k+1), (k+1\,k+2), \ldots, (l-1\,l)$, if $k < l$, the sum of the weights of edges $W(\cdot)$ representing adjacent transpositions $(l\,l+1), (l+1\,l+2), \ldots, (k-1\,k)$, if $l < k$, and 0, if $k = l$.

Example 3: Suppose that one is given four rankings, (1, 2, 3), (1, 2, 3), (3, 2, 1) and (2, 1, 3). There are two optimal aggregates according to the Kendall tau distance, namely (1, 2, 3) and (2, 1, 3). Both have cumulative distance four from the set of given permutations. If the transposition weights are non-uniform, say such that $W(12) > W(23)$, the solution becomes unique and equal to (1, 2, 3). If the last ranking is changed from (2, 1, 3) to (2, 3, 1), exactly three permutations are optimal from the perspective of Kendall tau aggregation: (1, 2, 3), (2, 1, 3) and (2, 3, 1). These three solutions give widely different predictions of what one should consider the top candidate. Nevertheless, by choosing once more $W(12) > W(23)$ the solution becomes unique and equal to (1, 2, 3).

It can be shown that for any non-negative weight function w, and for two permutations σ and θ, one has

$$1/2 D_w(\sigma, \theta) \leq d_w(\pi, \sigma) \leq D_w(\sigma, \theta)$$

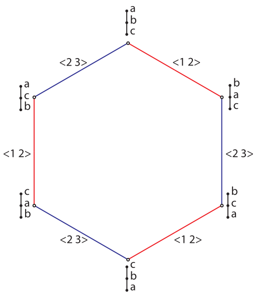

Fig. 2. The Kendall distance is the weight of the shortest path between two vertices labeled by two permutations, with each edge having length (weight) one. Edges are labeled by the adjacent swaps used to move between the vertex labels. For example, the two vertices labeled by *acb* and *cab* are connected via an edge bearing the label $< 12 >$, indicating that the two permutations differ in one swap involving the first and second element

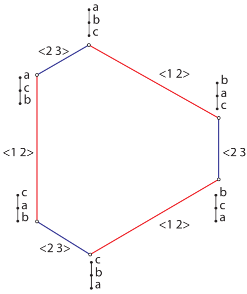

Fig. 3. The weighted Kendall distance is the weight of the shortest path between two permutations, with edges having possibly different lengths (weights). Edges are labeled by the adjacent swaps used to move along the vertices

In a companion article (Farnoud *et al.*, 2012), we presented extensions of the WBGM and PR aggregation methods for weighted Kendall distances. Here, we will pursue the WBGM framework, and propose a new method to compute the weights $W(\cdot)$ of edges (swaps) based on the *p*-values of the genes within each similarity criteria ranking. We refer to the resulting weighted model as the *Hybrid Kendall* method.

To start, arrange the *p*-values of all genes based on all similarity-criteria into an $n \times m$ matrix P. Next, rearrange the *p*-values of genes for each criteria in an increasing order, and denote the resultingrearranged matrix by $P^* = (p^*(i,j))$. We use the following $(n-1) \times m$ swap weight matrix $\mathcal{W}$, with entries

$$\mathcal{W}(i,j) = c\left(\frac{P^*(i+1,j) - P^*(i,j)}{P^*(i+1,j)}\right) \times d^{n-i},$$

indicating how much it costs to swap positions *i* and $i+1$ for criteria *j*. The parameters *c*, *d* are constants independent of *n* and *m*, used for normalization and for emphasizing the TvB constraint, respectively. For our simulations, we set $c=10$ and $d=1.05$, as these choices provided good empirical performance on synthetic data. The swap matrix assigns high weight to the top of the list.

To compute the aggregate based on the approximate distance $D_{\mathrm{w}}(\theta, \sigma)$, we only need to accumulate each of the contributions from the training permutations in Σ. This may be achieved by using a $n \times n$ total cost matrix C, with entry C(*i*, *j*) indicating how much it would 'cost' for gene *i* to be ranked at position *j*:

$$C(i,j) = \frac{1}{m}\sum_{k=1}^{m}\sum_{l=\min(j,\sigma_{\mathrm{p_k}}(i))}^{\max(j,\sigma_{\mathrm{p_k}}(i))-1} \mathcal{W}(l,k)$$

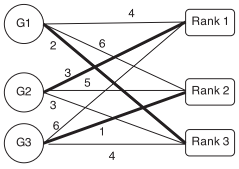

Fig. 4. A matching in a weighted bipartite graph.

The total cost matrix *C* is the input to the WBGM algorithm, where *C*(*i*, *j*) denotes the weight of an edge connecting gene *i* with position *j* (see Fig. 4 for an example of the bipartite graph, with the left-hand side nodes denoting genes and the right-hand side nodes denoting their possible positions; the minimum weight matching is represented by bold edges). To find the minimum cost solution, or the maximum weight matching, we used the classical *Hungarian algorithm* (Kuhn, 1955) implemented in (Melin, 2006).

Example 4: Let $n=4$ and $m=2$, where the two ratings equal to $p_1 = (0.2, 0.3, 0.01, 0.12)$ and $p_2 = (0.1, 0.4, 0.2, 0.35)$. Then

$$P^* = \begin{bmatrix} 0.01 & 0.1 \\ 0.12 & 0.2 \\ 0.2 & 0.35 \\ 0.3 & 0.4 \end{bmatrix}, \quad \mathcal{W} = \begin{bmatrix} 10.61 & 5.79 \\ 4.41 & 4.73 \\ 3.5 & 1.31 \end{bmatrix},$$

$$C = \begin{bmatrix} 7.51 & 5.1 & 5.23 & 7.67 \\ 15.18 & 6.98 & 2.4 & 0 \\ 2.9 & 5.3 & 9.88 & 12.28 \\ 10.57 & 2.37 & 2.2 & 4.61 \end{bmatrix}.$$

For example, since gene 3 was ranked 1st and 2nd by the two criteria, $C(3,3) = 1/2(10.61 + 4.41) + 1/2(4.73) = 9.88$. The minimum cost solution of the matching with cost matrix C, based on the Hungarian algorithm yields the aggregate $\sigma_{HK} = (3, 1, 4, 2)$.

2.3 The Lovász-Bregman divergence method

A previously reported distance measure represents another possible mean for performing HyDRA. The so called *Lovász-Bregman* method (Iyer and Bilmes, 2013) calls for a distance measure between real-valued vectors $x \in \mathbb{R}^n_{\geq 0}$ and permutations.

To define the Lovász-Bregman divergence that acts as a distance proxy between rankings and ratings, we start with a submodular set-function, i.e. a function *f* such that for a finite ground set *V*, $f: 2^V \to \mathbb{R}$, and for all $S, T \subset V$, it holds $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$. The Lovász extension of *f*, $f^L(x)$, equals

$$f^L(x) = \sum_{i=1}^{n} x(\sigma_x(i)) \left[f(S^{\sigma_x}_{\mathrm{j}}) - f(S^{\sigma_x}_{j-1})\right],$$

where $S^{\sigma_x}_{\mathrm{j}}$ denotes the set $\{\sigma_x(1),\ldots,\sigma_x(j)\}$. Note that under some mild conditions, the Lovász extension is convex. Let us next define the differential of *f* as

$$h^{\mathrm{f}}_{\sigma_x}(\sigma_x(j)) = f(S^{\sigma_x}_{\mathrm{j}}) - f(S^{\sigma_x}_{j-1})$$

Then the Lovász-Bregman divergence is defined via the dot product

$$d_{\mathrm{r}}(x||\sigma) = x \cdot (h^{\mathrm{f}}_{\sigma_x} - h^{\mathrm{f}}_{\sigma})$$

Despite its seemingly complex expression, the Lovász-Bregman divergence allows for *closed form* aggregation for a large class of submodular functions *f*. The optimal aggregate reduces to the ranking induced by the *sum of real-valued rating vectors*, ordered in a decreasing manner.

If, as before, $p(i, j)$ denotes the p-value of gene i under criteria j, we define the *normalized* Lovász-Bregman score for gene i as

$$\mathcal{L}(i) = \sum_{j=1}^{m} \frac{p(i,j)}{\frac{1}{n}\sum_{i=1}^{n} p(i,j)},$$

where the sum of p-values over criteria is normalized by the average of the p-values for each criterion. The aggregate equals $\sigma_{\mathcal{L}}$, where $\mathcal{L} = (\mathcal{L}(i))_{i=1}^{n}$.

Example 5: Let $n = 4$ and $m = 2$, where the two ratings equal to $p_1 = (0.2, 0.3, 0.01, 0.12)$ and $p_2 = (0.1, 0.4, 0.2, 0.35)$. Note that $1/n\sum_{i=1}^{\mathrm{n}} p(i,1) = 1/4(0.2 + 0.3 + 0.01 + 0.12) = 0.1575$, and $1/n \sum_{i=1}^{\mathrm{n}} p(i,2) = 1/4(0.1 + 0.4 + 0.2 + 0.35) = 0.2625$. The Lovász-Bregman scores $\mathcal{L}(i), i = 1, 2, 3, 4$, equal $\mathcal{L}(1) = 0.2/0.1575 + 0.1/0.2625 = 1.65$, $\mathcal{L}(2) = 0.3/0.1575 + 0.4/0.2625 = 3.43$, $\mathcal{L}(3) = 0.01/0.1575 + 0.2/0.2625 = 0.83$, $\mathcal{L}(4) = 0.12/0.1575 + 0.35/0.2625 = 2.1$. By ordering $\mathcal{L}(i)$ in an ascending manner, one arrives at $\sigma_{LB} = (3, 1, 4, 2)$.

3 Algorithms and implementation

We now turn our attention to testing different aggregation methods on lists of p-values generated by Endeavour and ToppGene. The aforementioned methods rely on a set of training genes known to be involved in a disease. The test genes are compared with all the training genes according to a set of similarity criteria, and the p-value of each comparison is computed in the process. For example, if the criterion is sequence similarity, the p-value reflects the z-value, describing the number of standard deviations above the mean for a given observation. Given the p-values, the question of interest becomes how to aggregate them into one ranking. Computing the p-values is a routine procedure, and the challenge of the prioritization process is to most meaningfully and efficiently perform the aggregation step.

There are two settings in which one can use the aggregation algorithms. The first setting is *cross-validation*, a verification step that compares the output of an aggregation algorithm with existing, validated knowledge. This mode of operation is aimed at discovering shortcomings and advantages of different methods. In the second setting, termed *gene discovery*, the aim is to identify sets of genes implicated in a disease which are not included in the database. Clearly, cross-validation studies are necessary first steps in gene discovery procedures, as they explain best aggregation strategies for different datasets and different similarity and training conditions.

For both methods, a list of genes involved in a certain disease (referred to as onset genes) was obtained from the publicly available databases Online Mendelian inheritance in Man (OMIM) (Hamosh *et al.*, 2005) and/or the Genetic Association Database (GAD) (Becker *et al.*, 2004). Both of these sources rely on the literature for genetic association for vast number of diseases, but OMIM typically provides a more conservative (i.e. shorter) list than the GAD. Onset genes were tested along with *random genes*, obtained by randomly permuting 19, 231 human genes in the GeneCards database (Safran *et al.*, 2002), and retaining the top portion of the list according to the chosen number of test genes.

3.1 Cross-validation

We performed a systematic, comparative performance analysis of the ToppGene and Endeavour aggregation algorithms and the newly proposed hybrid methods. Given a list of r onset genes, we first selected t onset genes to serve as target genes (henceforth referred to as *target onset genes*) for validation; we used the remaining $r{-}t$ onset genes as training genes. Of the n test genes, $n{-}t$ genes were selected randomly from GeneCards (Safran *et al.*, 2002). Our cross-validation procedure closely followed that of Endeavour and ToppGene: we fixed $t = 1$, and tested *all* r individual genes from the pool of onset genes, and then averaged the results. Averaging was performed as follows: we took target onset genes one-by-one and averaged their rankings over $\binom{r}{t}_{t=1} = r$ experiments. Note that in principle, one may also choose $t \geq 2$; in this case, the lowest ranking of the t genes (i.e. the highest positional value that a target onset gene assumed) should serve as a good measure of performance. One would then proceed to average the resulting rankings over $\binom{r}{t}$ experiments, producing a 'worst case scenario' for ranking of target onset genes. For fair comparison with Endeavour and ToppGene, we only used the first described method with $t = 1$ and the same set of p-values as inputs. As will be described in subsequent sections, we used $t \geq 2$ for gene discovery procedures.

3.2 Gene discovery

The ultimate goal of gene prioritization is to *discover* genes that are likely to be involved in a disease without having any prior experimental knowledge about their role. We describe next a new, iterative *gene discovery* method. The method uses aggregation techniques or combinations of aggregation techniques deemed to be most effective in the cross-validation study.

Given a certain disease with r onset genes, we first identify s *suspect genes*. Suspect genes are genes that are known to be involved in diseases *related* to that under study (as an example, a suspect gene for glioblastoma may be a gene known to be implicated in another form of brain cancer, say meningioma), but have not been tested in

Algorithm 1: Gene Discovery

```
Input: Set of onset genes, $O = \{o_1, o_2, \ldots, o_r\}$, set of suspect genes, $S = \{s_1, s_2, \ldots, s_s\}$, number of test genes, $n \in \mathbb{Z}^+$, a cut-off threshold, $\tau \in \mathbb{Z}^+$, and the number of allowed iterations, $I \in \mathbb{Z}$
Output: Set of potential disease genes, denoted by $A$
Initialization:
  • Set $i = 1$, $A = \emptyset$, $R = \{r_1, r_2, \ldots, r_{n-s}\}$ – a set of randomly chosen genes, training set $\mathrm{TR} = O$, test set $\mathrm{TS} = S \cup R$
For $i \leq I$ do
  1. Run a gene prioritization suite using the training set TR, test set TS, and $m$ similarity criteria
  2. Run $k$ aggregation methods on the p-values produced in Step 1, and denote the resulting rankings by $\sigma_1, \ldots, \sigma_k$
  3. Let $B = \{\sigma_1(1), \ldots, \sigma_1(\tau)\} \cup \cdots \cup \{\sigma_k(1), \ldots, \sigma_k(\tau)\}$
  4. $A \leftarrow A \cup B$; $\mathrm{TR} \leftarrow \mathrm{TR} \cup B$; $S \leftarrow S \backslash B$
  5. $\mathrm{TS} \leftarrow S \cup R'$, $R'$ = set of $n - |S|$ randomly chosen genes
  6. $i \leftarrow i + 1$
End
Return $A$
```

this possible role. Suspect genes are processed in an iterative manner, as illustrated in Algorithm 1. In the first iteration, r onset genes are used for training, and s suspect genes, along with $n - s$ randomly selected genes, are used as test genes. From the aggregate results provided by different hybrid algorithms, we selected q top-ranked genes and moved them to the set of training genes and simultaneously declared them as potential disease genes. The choice for the parameter q is governed by the number of training and test genes, as well as the empirical performance of the aggregation methods observed during multiple rounds of testing. The second iteration starts with $r + q$ training genes, $s - q$ suspect genes, and $n - s + q$ randomly selected genes; the procedure is repeated until a predetermined stopping criteria is met, such as the size of the set of potential disease genes exceeding a given threshold.

4 Results

We performed extensive cross-validation studies for eight diseases using both Endeavour- and ToppGene-generated p-values. Our results indicate that the similarity criteria that exhibits the strongest influence on the performance of the ToppGene and the Endeavour method is the PubMed and literature criteria, which award genes according to their citations in the disease related publications. In order to explore this issue further, we performed additional cross-validation studies for both ToppGene and Endeavour datasets to examine how exclusion of the literature criteria changes the performance of the two methods as well as our hybrid schemes. Our results reveal that HyDRA aggregation methods outperform Endeavour and ToppGene procedures for a majority of quality criteria, but they also highlight that each method offers unique advantages in prioritization for some specific diseases.

For gene discovery, we again used Endeavour and ToppGene p-values, and investigated three diseases—glioblastoma, meningioma and breast cancer—including all criteria available. We recommend as best practice a nested aggregation method, i.e. aggregating the aggregates of Endeavour, HyDRA and ToppGene, coupled with iterative training set augmentation.

4.1 Cross-validation

Cross-validation for HyDRA methods was performed on autism, breast cancer, colorectal cancer, endometriosis, ischaemic stroke, leukemia, lymphoma and osteoarthritis. Table 2 provides the summary of our results, pertaining to the average rank of one selected target gene. Table 2 illustrates that HyDRA methods offer optimal performance in 11 out of 16 tests when compared with ToppGene aggregates, and in 12 out of 16 cases when compared with Endeavour aggregates. In the former case, the Weighted Hybrid Kendall method outperformed all other techniques. A detailed review of our cross-validation results is given in the supplementary data Section S1. Note that in for all eight diseases, we performed two tests, in one of which we excluded those similarity criteria that contain strong prior information about disease genes, such as the 'Disease' and 'PubMed' category. Table 2 demonstrates the significant differences in average ranks of the target genes when literature information is excluded, suggesting that ToppGene and Endeavour both significantly benefit from this prior onset gene information when ranking the target genes. The Supplementary data Section S2 contains a detailed description of our results.

Another means for evaluating the performance of HyDRA algorithms compared with that of ToppGene and Endeavour is to examine the receiver operating characteristic (ROC) curves of the techniques. In this setting, we follow the same approach as used by both ToppGene and Endeavour. Sensitivity is defined as the frequency of tests in which prospect genes were ranked above a particular threshold position, and specificity as the percentage of prospect genes ranked below this threshold. As an example, a sensitivity/specificity pair of values 90/77 indicates that the presumably correct disease gene was ranked among the top-scoring $100 - 77 = 23\%$ of the genes in 90% of the prioritization tests. The ROCs plot the dependence between sensitivity and the reflected specificity, and the area under the curve (AUC) represents another useful performance measure. The higher the AUC and specificity, the better the performance of the method. Endeavour reported 90/74 sensitivity/specificity values for their chosen set of test and training genes, as well as an AUC score of 0.866. Similarly, ToppGene reported 90/77 sensitivity/specificity values and an AUC score of 0.916 for

Table 2. Cross-validation result of Endeavour, ToppGene and HyDRA methods for eight diseases

Disease	No. onset genes	ToppGene	Lovasz-Bregman	Hybrid Borda	Hybrid Kendall	Endeavour	Lovasz-Bregman	Hybrid Borda	Hybrid Kendall
Autism	40	7.275	11.2	9.75	**6.85**	17.96	19.3	17.78	**16.9**
Autism†	40	21.675	25.4	**19.775**	21.65	23.35	24.5	24.38	**21.78**
Breast cancer	10	4.6	7.1	12	**2.5**	14.4	15	**12.5**	15.7
Breast cancer†	10	**6.9**	17.8	8.1	7.1	16.6	**12.8**	15.5	17.8
Colorectal cancer	20	7.3	**5.2**	7.85	8.7	8.55	8.65	**7.8**	8.1
Colorectal cancer†	20	13.35	**9.5**	19.6	12.5	9.75	10.65	**9.55**	11.2
Endometriosis	43	**6.46**	8.63	10.63	7.74	5.3	6.37	**4.81**	5.65
Endometriosis†	43	**9.53**	9.76	15.84	9.7	**6.12**	7.63	6.86	6.6
Ischaemic stroke	44	**5.61**	7.25	9.25	6.05	6.18	7.3	7.07	**6.09**
Ischaemic stroke†	44	8.43	**7.5**	12.8	8.7	**7.95**	9.66	9.86	8.86
Leukemia	10	**5.5**	12	6.6	10.2	13.7	14.8	**7.1**	12.1
Leukemia†	10	20.8	22.8	24.3	**20.5**	19.5	19.9	**16.6**	21.3
Lymphoma	42	3.74	6.45	9.26	**2.93**	9.57	10.69	9	**8.81**
Lymphoma†	42	7.71	9.55	10.71	**6.76**	12.52	12.9	13.67	**11.67**
Osteoarthritis	41	6.44	6.51	13.54	**5.41**	**5.56**	6.32	7.46	6.29
Osteoarthritis†	41	8.73	8.32	14.1	**8.02**	**6.41**	7.41	6.51	7.22

Diseases without '†' refer to results using all 18 similarity categories both in Endeavour and ToppGene. Diseases indexed by '†' denote results which did not use the 'Human Phenotype, Mouse Phenotype, Pubmed, Drug, Disease' similarity criteria in ToppGene. Similarly, for Endeavour, the indexing by '†' corresponds to exclusion of similarity criteria 'Precalculated-Ouzounis, Precalculated-Prospectr, Text' on Endeavour data. The scores describing the best average rank are bolded and shaded.

Table 3. AUC and sensitivity/specificity values for ToppGene, Endeavour and HyDRA rankings, pertaining to diseases listed in table 2 using all criteria

	ToppGene	Lovasz-Bregman	Hybrid Borda	Hybrid Kendall		Endeavour	Lovasz-Bregman	Hybrid Borda	Hybrid Kendall
AUC	0.951	0.93	0.911	0.947	AUC	0.908	0.899	0.918	0.91
Sensitivity/Specificity	90/84	90/75	90/75	90/84	Sensitivity/specificity	90/69	90/63	90/79	90/72

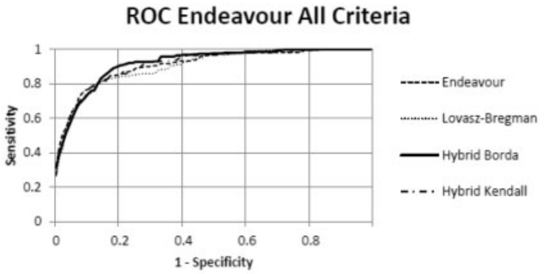

Fig. 5 Cross-validation results: ROC curves for disease listed in table 2 using all criteria and Endeavour data.

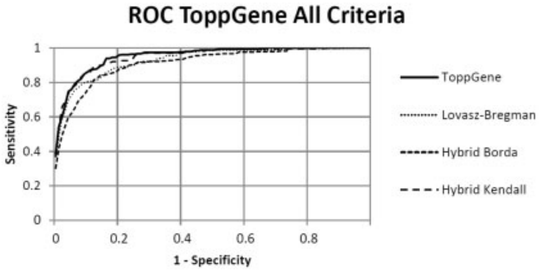

Fig. 6. Cross-validation results: ROC curves for disease listed in table 2 using all criteria and ToppGene data.

their tests of interest. Our specificity/sensitivity and AUC values are listed in Table 3, with best AUC and Sensitivity/Specificity values shaded in gray. Note that although the AUC values appear close in all cases, the HyDRA methods have very low overall computational complexity (Figs. 5 and 6).

4.2 Gene discovery

The genetic factors behind glioblastoma, the most common and aggressive primary brain tumor, are still unknown. We study this disease, as well as meningioma and breast cancer, in the gene discovery phase. Our choice is governed by the fact that few publications are available pointing towards the causes of this form of brain cancer, and by the fact that it is widely believed that the genetic base of this disease is related to the genetic base of the Von Hippel-Lindau (VHL), Li-Fraumeni (LF), and Turcot Syndromes (TS), Neurofibromatosis (N) and Tuberous Sclerosis (TS) (Kyritsis *et al.*, 2009). Furthermore, recent findings (Pandey, 2014) indicate that brain cancers and breast cancers share a common line of mutations in the family of Immunoglobulin GM genes, and that the Human cytomegalovirus puts patients at risk of both brain and breast cancer.

Consequently, we used genes documented to be involved in glioblastoma as training genes for three discovery tests. In the first test, for the suspect genes we selected a subset of 15 genes known to be implicated in the VHL, LF, TS, N and TS syndromes. We subsequently ran Algorithm1 with $l=3$, $s=15$, $n=100$, $\tau=3$. In the second test, we selected 18 genes known to be involved in breast cancer as suspect genes for glioblastoma, and run Algorithm1 with $l=3$, $s=18$, $n=100$, $\tau=3$. Finally, we performed the same analysis on suspect genes known to be involved in meningiomas, by setting the parameters of iterative HyDRA gene discovery to $l=3$, $s=19$, $n=100$, $\tau=3$. The results are shown in Table 4. Note that in our algorithmic investigation, we used $l=3$ (i.e. top-three) ranked genes, since this parameter choice offered a good trade-off between the size of the union of the top-ranked genes and the accuracy of the genes produced by the HyDRA discovery methods. The number of suspect genes was governed by the size of the available pool in OMIM/GAD and was targeted to be roughly 20% of the size of the test set. Such a percentage is deemed to be sufficiently high to allow for meaningful discovery, yet sufficiently low to prevent routine gene identification.

Table 4 reveals a number of results currently not known from the literature. The genes KRAS and CDH1, both implicated in breast cancer and meningioma, as well as CCND1 involved in meningioma (as well as in colorectal cancer) appear to be highly similar to genes implicated with glioblastoma. KRAS is a gene encoding for the K-Ras protein that is involved in regulating cell division, and hence an obvious candidate for being implicated in cancer. On the other hand, CDH1 is responsible for the production of the E-cadherin protein, whose function is to aid in cell adhesion and to regulate transmission of chemical signals within cells, and control cell maturation. E-cadherin also often acts as a tumor suppressor protein. GeneCards reveals that the CCND1 gene is implicated in altering cell cycle progression, and is mutated in a variety of tumors. Its role in glioma tumorogenesis appears to be well documented (Buschges *et al.*, 1999), but surprisingly, neither KRAS nor CDH1 nor CCND1 are listed in the OMIM/GAD database as potential glioblastoma genes.

Another interesting finding involves genes ranked among the top three candidates, but not identified as 'suspect' genes. For instance, according to GeneBank, GSTM2 regulates an individual's susceptibility to carcinogens and toxins and may suggest glioblastoma being in part caused by toxic and other environmental conditions; KAAG1 appears to be implicated with kidney tumors, while TP73 belongs to the p53 family of transcription factors and is known to be involved in neuroblastoma.

5 Discussion

We start by discussing the results in Table 2. The first observation is that the Lovász-Bregman method performs worse than any other aggregation method. This finding may be attributed to the fact that the p-values have a large span, and small values may be 'masked' by larger ones. Scaling all p-values may be a means to improve the performance of this technique, but how exactly to accomplish this task remains a question.

In almost all cases, except for Leukemia and Lymphoma, the average rankings produced by ToppGene and the Weighted Kendall distance appear to be almost identical. But *average* values may be misleading, as individual rankings of genes may vary substantially between the methods, as can be seen from the supplementary material. It is due to this reason that we recommend merging lists

Table 4. The union of top three ranked genes from ToppGene, Endeavour and HyDRA methods for the three suspect gene discovery sets, with the 'suspect' genes in bold

Test disease	Iteration 1	Iteration 2
Breast cancer	**AKT1, ATM, BRIP1, CDH1, CHEK2**, GSTM2, KAAG1, **RAD51**, TP73	**BARD1**, CASP7, ITGA4, **KRAS, PALB2, PHB**, SMAD7, UMOD
VHL, LF, TS, N, TS	**CCND1**, CD28, CD74, **CDK4, CHEK2, MLH1, MSH2, MSH6**, NBPF4, **PMS2**, PRNT, **TSC2**	ALCAM, **APC, MRC1**, NCL, **NF1, NF2**, SNCA, TAF7, TOPBP1, **TSC1, VHL**
Meningioma	**CCND1, HLA-DQB1**, KLF6, **KRAS, TGFB1, TGFBR2, XRCC5**	BAGE, **BAP1**, CAV1, CD4, **CDH1, NF2, PDGFB**, PSMC2, **RFC1**, SAMD9L, SERPING1, **SMARCB1**

In all cases, the training genes are genes implicated in glioblastoma. The 'Disease' category indicates from which family of diseases the test genes were drawn. The results of the third iteration may be found in the Supporting file Section S3.

generated by different methods as best aggregation practice. Another important observation is that HyDRA methods have significantly lower computational complexity than ToppGene and especially, Endeavour, and hence scale well for large datasets.

Another finding is the fact that the good performance of ToppGene and all other methods largely depends on including prior literature on the genes into the aggregation process. We observed situations where the rank of an element dropped by roughly 90 positions when this prior was not available. This implies that for gene discovery, it is risky to rely on any single method, and it is again good practice to merge top-ranked entries generated by different methods. Finally, it is not clear how to optimally choose the number of training genes for a given set of test genes, or vice versa. Choosing more training genes may appear to be beneficial at first glance, but it creates a more diverse pool of candidates for which some similarity criteria will inevitably fail to identify the right genes. In this case, we recommend using the Weighted Kendall to eliminate outliers, and in addition, we recommend the use of a fairly large TvB scaling parameter.

Acknowledgements

The work was supported in part by the National Science Foundation (NSF) under grants CCF 0809895, CCF 1218764, CSoI-CCF 0939370, and IOS 1339388.

Conflict of interest: none declared.

References

Adie,E.A. *et al.* (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, **22**, 773–774.

Aerts,S. *et al.* (2006) Gene prioritization through genomic data fusion. *Nat Biotechnol*, **24**, 537–544.

Bartholdi,J. *et al.* (1989) The computational difficulty of manipulating an election. *Soc. Choice Welfare*, **6**, 227–241.

Becker,K.G. *et al.* (2004) The Genetic Association Database. *Nat Genet*, **36**, 431–432.

Buschges,R. *et al.* (1999). Amplification and expression of cyclin D genes (CCND1 CCND2 and CCND3) in human malignant gliomas. *Brain Pathol.*, **9**, 435–442.

Cardon,L.R. *et al.* (2001) Association study designs for complex diseases. *Nat Rev Genet*, **2**, 91–99.

Chen,J. *et al.* (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res*, 37, W305–W311.

De Bie,T. *et al.* (2007) Kernel-based data fusion for gene prioritization. *Bioinformatics*, **23**, i125–i132.

Dwork,C. *et al.* (2001). Rank aggregation methods for the web. In: *Proceedings of the 10th international conference on World Wide Web (WWW10), ACM*. Hong Kong, China. pp. 613–622.

Farnoud,F. *et al.* (2012) Nonuniform vote aggregation algorithms. In: *Signal Processing and Communications (SPCOM), IEEE*. Bangalore, India. pp. 1–5.

Farnoud,F. *et al.* (2014), An axiomatic approach to constructing distances for rank comparison and aggregation. *IEEE Trans Inform Theory*, **60**, 6417–6439.

Fishburn,P. (1970) Arrow's Impossibility theorem: concise proof and infinite voters. *J Econ Theory*, **2**, 103–106.

Freudenberg,J. and Propping,P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, **18**, S110–S115.

Guney,E. *et al.* (2014) GUILDify: a web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms. *Bioinformatics*, **30**, 1789–1790.

Hamosh,A. *et al.* (2005) Online Mendelian inheritance in Man (OMIM), a knowledge base of human genes and genetic disorders. *Nucleic Acids Res*, **33**, D514–D517.

Iyer,R. and Bilmes,J.A. (2013) The Lovász-Bregman divergence and connections to rank aggregation, clustering, and web ranking. In: *Uncertainty in Artificial Intelligence (UAI)*, AUAI, Bellevue, Washington. pp. 1–10.

Kacprowski,T. *et al.* (2013) NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules. *Bioinformatics*, **29**, 1471–1473.

Kemeny,J.G. (1959) Mathematics without numbers. *Daedalus*, **88**, 577–591.

Kendall,M.G. (1938) A new measure of rank correlation. *Biometrika*, **30**, 81–93.

Kendall,M. (1948) *Rank Correlation Methods*. Charles Griffin and Company Limited, London.

Köhler,S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet*, **82**, 949.

Kolde,R. *et al.* (2012) Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, **28**, 573–580.

Kuhn,H.W. (1955) The Hungarian method for the assignment problem. *Nav Res Log*, **2**, 83–97.

Kyritsis,A.P. *et al.* (2009) Inherited predisposition to glioma. *Neuro Oncol*, **12**, 104–113.

Melin,A. (2006) The Hungarian algorithm. *MATLAB Central File Exchange*. http://www.mathworks.com/matlabcentral/fileexchange/11609-hungarian-algorithm (8 August 2006, retrieved).

Pandey,J.P. (2014) Immunoglobulin GM genes, cytomegalovirus immunoevasion, and the risk of glioma, neuroblastoma, and breast cancer. *Front Oncol*, **4.**, 238.

Perez-Iratxeta,C. *et al.* (2002) Association of genes to genetically inherited diseases using data mining. *Nat Genet*, **31**, 316–319.

Pihur,V. *et al.* (2009) RankAggreg, an R package for weighted rank aggregation. *BMC Bioinformatics*, **10**, 62.

Popescu,M. *et al.* (2006) Fuzzy measures on the Gene Ontology for gene product similarity. *IEEE/ACM Trans Comput Biol Bioinformatics*, **3**, 263–274.

Raisali,F. *et al.* (2013) Weighted rank aggregation via relaxed integer programming. In: *International Symposium on Information Theory (ISIT), IEEE*. Istanbul, Turkey. pp. 2765–2767.

Risch,N. and Merikangas,K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

Safran,M. *et al.* (2002) GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics*, **18**, 1542–1543.

Thanassoulis,E. (2001) *Introduction to the Theory and Application of Data Envelopment Analysis*. Kluwer Academic Publishers, Dordrecht.

Tiffin,N. *et al*. (2006) Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res*, **34**, 3067–3081.

Turner,F.S. *et al*. (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol*, **4**, R75–R75.

Warde-Farley,D. *et al*. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*, **38**, W214–W220.

Yu,S. *et al*. (2008) Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining. *Bioinformatics*, **24**, i119–i125.