Gene Prioritization via Weighted Kendall Rank Aggregation

Minji Kim, Fardad Raisali, Farzad Farnoud (Hassanzadeh) and Olgica Milenkovic
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, Urbana-Champaign, Illinois 61801
Email: mkim158@illinois.edu

***Abstract*—Gene prioritization is a class of methods for discovering genes implicated in the onset and progression of a disease. As candidate genes are ranked based on similarity to known disease genes according to different set of criteria, the overall aggregation of these ranked datasets is a vital step of the prioritization procedure. Aggregation of different lists of ordered genes is accomplished either via classical order statistics analysis or via combinatorial ordinal data fusion. We propose a novel approach to combinatorial gene prioritization via Linear Programming (LP) optimization and use the recently introduced weighted Kendall τ distance to assess similarities between rankings. The weighted Kendall τ distance allows for constructing aggregates that have higher accuracy at the top of the ranking, usually tested experimentally, and it can also accommodate ties in rankings and handle negative outliers. In addition, the Kendall distance does not use quantitative data which in many instances may be unreliable. We illustrate the performance of the prioritization method on a set of test genes pertaining to the Bardet-Biedl syndrome, schizophrenia, and HIV and show that the combinatorial method matches or outperforms state-of-the art algorithms such as ToppGene.**

I. Introduction

It is known that humans have roughly $25,000$ genes, some of which – when mutated – may lead to a host of diseases, conditions and abnormal phenotypes. Despite decades of intense research focus, the underlying gene aberrations that lead to even the most frequently encountered diseases are not completely known. Usually, the main impediment to identifying disease genes is the time-consuming and costly process of testing a working hypothesis, further exacerbated by alternative splicing and by the fact that typically, multiple genes have to be jointly mutated to trigger the onset of a disease. Even for experiments involving only up to three genes, one would have to test as many as 4×10^{12} combinations of genes in order to check if they are linked to a given disease. This is clearly an infeasible experimental endeavor which will remain difficult to accomplish for decades to come.

One approach to mitigate the problem is to preprocess available biological side-information about genes and then reduce the set of test genes accordingly. The problem of identifying a small subset of genes likely to be causally linked with a disease is known as the *gene prioritization problem*, and the class of algorithmic solutions for solving the problem are known as prioritization algorithms. Prioritization algorithms are typically based on using experimentally confirmed disease genes and identifying different qualitative evidence that associates the disease genes with target test genes. For this purpose, linkage analysis, sequence similarity, functional annotation, marker and pedigree analysis are all combined. The evidence obtained establishes the ranking of candidate genes based on the extent of their relationship – or similarity – to the training set of disease genes.

In the past few years, a number of sophisticated computational gene prioritization tools were proposed in [1], [4], [6], [11]. Most of these methods are statistical and *quantitative* in nature. Although offering significant improvements over random search methods, most such methods suffer from the fact that they tacitly or implicitly rely on the assumptions that a) a test gene has to be close to the training genes *under all similarity criteria*; in other words, the top-ranked genes have to be highly ranked in all individual lists reflecting different criteria for comparison; and b) no distinction is to be made about the accuracy of ranking genes in any part of the list; in other words, 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 pathways. Given that the goal of prioritization is to produce a list of genes to be experimentally tested in a wet lab, only highly relevant candidate genes are to be considered, and consequently, such genes have to be ranked with higher accuracy than other genes on the list. Furthermore, aggregation of rankings based on statistical methods is often highly sensitive to outliers and ranking errors.

To overcome the above issues of classical prioritization approaches, we employ a combinatorial median approach to ordinal data fusion using the weighted Kendall τ distance, first introduced by the authors in [9]. The aggregation approach is henceforth referred to as the *generalized Kemeny approach*. The ranking obtained using the weighted Kendall τ distance is more influenced by top positions in the rankings obtained from different criteria so it is robust to negative outliers – i.e., a small number of low rankings of some candidate gene. These properties are useful for gene prioritization, as weighted Kendall τ distance does not penalize genes for not being similar to training genes under *every* possible similarity criteria, and it allows for fusing weak orders in which several candidate genes may be ranked the same, which helps in resolving frequent scoring ambiguities. Although fundamental results from social choice theory and political sciences have shown that there exists no "optimal" rank aggregation method that is consistent, fair, and impossible-to-manipulate [15], the Kemeny method is one of the few aggregation solutions that provably offers

978-1-4673-3146-3/13/$31.00 ©2013IEEE

a large number of performance guarantees. The properties of the generalized Kemeny method were investigated in our companion papers [9], [13].

We apply the generalized Kemeny approach to lists of rankings generated by Endeavour and ToppGene [1], [4], using criteria such as sequence similarity, CisReg modules, expression profiles, transcription factor binding sites, annotation in different databases, pathways, etc. Our sets of test genes pertain to the Bardet-Biedl syndrome (a genetic condition affecting cellular cilia and causing obesity, retinal failure and sometime, mental retardation), schizophrenia, and HIV (Human Immunodeficiency Virus) infections. Despite the fact that generalized Kemeny aggregation is purely combinatorial in nature and hence discards all quantitative information in data, i.e., it does not make use of the *p*-values but only the underlying *rankings* of genes, it usually outperforms Endeavour [1] and matches/outperforms ToppGene [4]. In many instances, it produces ties in the rankings, indicative of potentially insufficient evidence to accurately discern the most similar genes (note that ToppGene and Endeavour always produce complete linear orders).

A. Background

Assume that one is given a set of n genes, ranked according to N different similarity criteria. For simplicity, one may assume that the genes are indexed by the positive integers $[n] = \{1, 2, \ldots, n\}$. Each ranking without ties may be viewed as a permutation over $[n]$, i.e., an element of the symmetric group $\mathbb{S}_n$. Similarly, a ranking with ties may be viewed as an ordered set partition, i.e., an ordered partition of the set $[n]$ into classes, where all genes in the same class are considered to have the same rank. As an example, for $n = 6$, $\sigma = (1, 5, 4, 3, 2, 6)$ is a ranking without ties, while $\sigma = (\{2, 3\}, \{1\}, \{4, 5, 6\}) = (2 - 3, 1, 4 - 5 - 6)$ is a ranking with ties. In the latter case, genes indexed by 2 and 3 share the first position, i.e. they are the top ranked genes. Usually, we represent ranking with ties through their *median scores*, defined as the average position of an element within a part. For the previous example, 2 and 3 have a median score of 1.5, given that they occupy the 1st and 2nd position, and $(1 + 2)/2 = 1.5$.

The inverse of a permutation σ is denoted by σ^{-1}. For each i, $\sigma^{-1}(i)$ denotes the rank of i in σ. In the example above, $\sigma^{-1}(4) = 3$. We use lower-case Greek letters for both permutations and rankings with ties, although it should be clear from the context to which entity we are referring to. The set of all rankings produced by different similarity criteria is denoted by $\Sigma = \{\sigma_1, \ldots, \sigma_N\}$.

In *distance-based* rank aggregation, the goal is to find a ranking, called the aggregate ranking, that is simultaneously as "close" as possible to all the votes in Σ. Closeness is measured via a chosen distance function over $\mathbb{S}_n$. We focus on aggregation using the Kendall τ distance, since this distance function has many desirable properties not matched by the Cayley distance, Spearman's footrule, and Spearman's rank correlation distances [7]. The Kendall τ distance between two permutations π and σ, denoted by $\mathrm{d}_K(\pi, \sigma)$, equals

$$\mathrm{d}_K(\pi, \sigma) = \left|\{(i, j) : \pi^{-1}(i) < \pi^{-1}(j), \sigma^{-1}(j) < \sigma^{-1}(i)\}\right|.$$

Alternatively, the Kendall distance between two permutations σ and π equals the smallest number of adjacent swaps of elements required to convert σ into π or vice versa.

The aggregate ranking π^* is formally computed as

$$\pi^* = \arg\min_{\pi \in \mathbb{S}_n} \sum_{\sigma \in \Sigma} \mathrm{d}_K(\pi, \sigma). \tag{1}$$

As already pointed out, a solution to the aforementioned problem is known as a Kemeny aggregate. It is known that computing the Kemeny aggregate is NP-hard [2]. To overcome the challenge of solving an NP-hard problem, a number of algorithms for approximate aggregation were proposed in the literature, including PageRank (PR), Weighted Bipartite Graph Matching (WBGM) [8], and Integer Programming (IP) relaxations/Linear Programming (LP) methods [5]. PR methods for rank aggregation mimic the algorithm used for ranking webpages by Google, and they reduce to computing equilibrium probabilities of Markov chains. WBGM algorithms are based on the fact that the Kendall distance may be approximated up to a multiplicative constant by the ℓ_1 norm of permutations.

We focus our attention on an alternative formulation of the Kemeny aggregation rule, described in what follows. Let $\sigma \in \mathbb{S}_n$, $i, j \in [n]$, and let

$$x_{ij}(\sigma) = \begin{cases} 1, & \text{if } \sigma^{-1}(i) < \sigma^{-1}(j), \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

denote the set of pairwise preference variables of σ. Note that there is a one-to-one correspondence between points x of the form above and permutations $\pi \in \mathbb{S}_n$, since $\pi^{-1}(i) < \pi^{-1}(j)$ if and only if $x_{ij} = 1$. Whenever clear from the context, we omit the symbol σ from the notation.

Straightforward computations show that the objective function of the Kemeny optimization procedure may be rewritten as [5]

$$\begin{aligned} \sum_{\sigma \in \Sigma} \mathrm{d}_K(\pi, \sigma) = \sum_{\sigma \in \Sigma} \mathrm{d}_K(x(\pi), \sigma) &= \sum_{\sigma \in \Sigma} \sum_{i,j} x_{ij} \sigma_{ji} \\ &= \sum_{i,j} c_{ij}\, x_{ij}, \end{aligned} \tag{3}$$

where $c_{ij} = \sum_{\sigma \in \Sigma} \sigma_{ji}$ and were we encoded a permutation π via its pairwise preference variables $x(\pi) = (x_{ij})$.

Furthermore, the constraints of Kemeny aggregation may be captured via the set of pairwise preference variables, $x = (x_{ij})$, as follows

$$\begin{aligned} x_{ij} + x_{ji} &= 1, && \text{for distinct } i, j \in [n], \\ x_{ij} + x_{jk} + x_{ki} &\leq 2, && \text{for distinct } i, j, k \in [n], \\ x_{ij} &\in \{0, 1\}, && \text{for distinct } i, j \in [n], \\ x_{ii} &= 0, && \text{for } i \in [n]. \end{aligned} \tag{4}$$

By relaxing the constraint in (4) to $x_{ij} \in [0, 1]$, we arrive at a Liner Programming (LP) approximation for a Kemeny aggregation solution.

In our companion paper [13], we showed how to extend the LP approximation framework for the weighted Kendall distance, first introduced in [9]. The main assumption behind the definition of this distance is that adjacent swaps $(i\ i + 1)$

are endowed with non-negative weights ρ_i. The ρ – weighted Kendall distance, $\mathrm{d}_\rho(\pi,\sigma)$, equals the smallest cost of any sequence of adjacent swaps needed to transform π into σ.

For a linearly decreasing weight function of the form

$$\rho_i = 1 + \frac{\epsilon}{n-2}(n-1-i),$$

with $\epsilon \geq 0$, it can be shown that the LP relaxation of the corresponding aggregation problem equals

$$\min_{w\in W} \frac{1}{n-2} \sum_{i,j,k} \alpha_{ijk} w_{ijk}, \tag{5}$$

where W represents the set of points $w = (w_{ijk})$, with $i,j,k \in [n]$, satisfying

$$\begin{aligned}
\textstyle\sum_{(r,s,t)\in\mathcal{T}_{i,j,k}} w_{rst} &= 1, && \text{for distinct } i,j,k \in [n],\\
w_{ijk} + w_{ikj} + w_{kij} &= x_{ij}, && \text{for distinct } i,j,k \in [n],\\
x_{ij},\, w_{ijk} &\in [0,1], && \text{for distinct } i,j,k \in [n],\\
w_{ijk} &= 0, && \text{for } i,j,k \text{ not distinct.}
\end{aligned}$$

Here, the variables x_{ij} have the same interpretation as in the classical Kemeny aggregation framework, $\mathcal{T}_{r,s,t} \equiv \mathbb{S}_3 = \{(r,s,t),(r,t,s),(s,r,t),(s,t,r),(t,r,s),(t,s,r)\}$, and

$$\alpha_{ijk} = \sum_{(r,s,t)\in\mathcal{T}_{i,j,k}} \mathrm{d}_\rho((r,s,t),(i,j,k))\, d_{rst},$$

where d_{rst} denotes the number of $\sigma \in \Sigma$ that rank r higher than s higher than t. Note that for the given linear choice of the weight ρ, it suffices to use $\mathcal{T}_{r,s,t}$ on triples of variables only. Furthermore, this definition easily extends to rankings with ties, by replacing $\mathcal{T}_{r,s,t}$ with

$$\begin{aligned}
\mathcal{T}^{(*)}_{r,s,t} = {} & \{(r,s,t),(r,t,s),(s,r,t),(s,t,r),(t,r,s),(t,s,r)\}\\
& \cup\ \{(r,s-t),(s,r-t),(t,r-s)\}\\
& \cup\ \{(r-s,t),(r-t,s),(s-t,r),(r-s-t)\},
\end{aligned}$$

and defining $\mathrm{d}_\rho(\pi_1,\pi_2)$, for $\pi_1,\pi_2 \in \mathcal{T}^{(*)}_{r,s,t}$, as the shortest path between π_1 and π_2 in the graph shown in Figure 1. Due to space limitations, the lengthy description of the constraint set for the aggregation problem, as well as detailed derivations and rigorous proofs behind the statements are relegated to the full version of the paper.

As a final remark, we observe that the LP program for weighted aggregation with ties of lists of n genes involves $O(n^3)$ constraints and $O(n^2)$ variables. Still, the constraints are sparse, which allows for efficient computational savings.

II. Prioritization Methods

One of the earliest gene prioritization software is known under the name Endeavour [1]. For different criteria, Endeavour ranks the candidate test genes based on their similarity to a set of known training genes. For each similarity criteria, Endeavour first calculates the average p-value with respect to the training genes, i.e., the probability of obtaining a test statistic as extreme as the one observed, under suitably chosen null hypotheses (the method which Endeavour uses to calculate the p-values is beyond the scope of this paper). It subsequently ranks the test genes from lowest to highest p-values. The

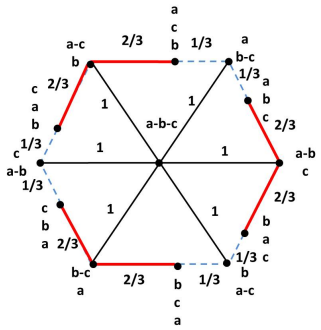

Fig. 1: A graph for the weighted Kendall distance between rankings with ties, involving three elements. The weights of the edges have to satisfy certain symmetry constraints, as described in [9,13]. The weights in our example are chosen to illustrate this symmetry property. To avoid confusion between the numerical values of the weight and the identity of candidates, we used the set $\{a,b,c\}$ to represent the candidates.

rankings are aggregated via the Q-statistic, calculated from all rank ratios r_i, $i = 1,\ldots,N$, using the joint cumulative distribution of an N – dimensional order statistic,

$$Q(r_1,r_2,...,r_N) = N! \int_0^{r_1} \int_{s_1}^{r_2} \cdots \int_{s_{N-1}}^{r_N} ds_N ds_{N-1} \ldots ds_1.$$

Here, the indices i refer to data sources, where N is the total number of data sources. Also, $r_0 = 0$.

ToppGene, a more recent software described in [4], also ranks candidate genes according to average p-values for different criteria, but the choice of criteria and the aggregation method differ from that proposed in Endeavour. The main difference is that ToppGene employs Human and Mouse phenotypes as one of the criteria, because direct comparison of human and mouse phenotypes provides vital information for identifying disease genes [3]. ToppGene aggregates rankings via Fisher's inverse chi-square method, which aggregates the p-values of different criteria, p_i, $i = 1,\ldots,N$, into $-2\sum_{i=1}^{N} \log p_i$. Assuming that the p-values $p_i, i = 1,\ldots,N$, come from independent tests and that the null hypotheses are all true, one has $-2\sum_{i=1}^{N} \log p_i \to \chi^2(2n)$, where $\chi^2(2n)$ denotes a χ^2 distribution with $2n$ degrees for freedom. Despite the fact that the p-values of gene prioritization criteria may not be independent, ToppGene currently appears to be the state-of-the-art prioritization method in terms of accuracy.

One of the most recently developed prioritization methods, NetworkPrioritizer [11], uses distances between genes in regulatory networks as additional criteria, and performs combinatorial aggregation based on Weighted Borda Fuse (WBF), Weighted AddScore Fuse (WASF), and MaxRank Fuse. However, these methods have the same drawbacks as the classical aggregation methods and differ substantially from the generalized Kemeny approach pursued in this paper.

III. Results

We tested the generalized Kemeny method on three diseases, and compared the overall rankings with those of ToppGene and Endeavour. For each disease, we obtained a list of phenotype genes on OMIM (Online Mendelian Inheritance in Man) [10], some of which are labeled as "training genes"

and some as "test genes". For example, OMIM lists 14 genes known to be involved in the Bardet-Biedl syndrome, 11 of which are listed as "training genes" in table I, and 3 genes, colored in red – TTC8, CEP290, MKS1– are part of the 12 "test genes". These phenotype test genes are expected to be ranked high in the overall aggregate, since there is strong evidence that they are similar to the training genes. The rest of the test genes are selected from GeneCards (www.genecards.org) [14] such that they are not known to be related to the disease. Although the sets of training and test genes are identical for Endeavour and ToppGene, the criteria used by Endeavour and ToppGene are different. For fairness of comparison, we took the *intersection* of Endeavour and ToppGene criteria. From the ToppGene suite, we used GO: Molecular Function, GO: Biological Process, GO: Cellular Component, Domain, Pathway, Pubmed, Interaction, Transcription Factor Binding Site, Gene Family. From the Endeavour suite, we used GeneOntology, Interpro, Kegg, Motif, Text.

We performed generalized Kemeny aggregation with ties via the LP method of Section II; the results are shown in tables I-III. The first two columns label the gene symbols with numbers, and those "Gene numbers" are used throughout columns 4-6. Note that column 3 simply indexes the ranking from 1 to 12, and the numbers are *not* gene numbers. Columns 4-6 contain rankings of genes according to ToppGene, generalized Kemeny, and Endeavour, respectively. In the case of the Bardet-Biedl syndrome, the generalized Kemeny method matches the performance of ToppGene, as it ranked the three phenotype genes at the top, and it outperforms Endeavour. A similar result is true for schizophrenia. The HIV results are interesting in so far that both ToppGene and Endeavour placed the three phenotype genes between the 2nd and 6th position, whereas the generalized Kemeny approach ranked all three phenotype genes at the top, tied along with 3 other non-phenotype genes.

TABLE I: Results for training genes CCDC28B, BBS5, ARL6, BBS7, BBS12, TMEM67, TRIM32, BBS1, BBS10, BBS4, BBS2, implicated with the **Bardet-Biedl syndrome**.

Gene #	HGNC Symbol	Rank #	ToppGene	Generalized Kemeny	Endeavour
1	TTC8	1	1	1	1
2	CEP290	2	2	3	2
3	MKS1	3	3	2	9
4	APP	4	4	5	3
5	ASPM	5	5	4	7
6	IL10	6	6	10 - 11	8
7	MYOD1	7	7		5
8	BDNF	8	8	7	11
9	SRY	9	9	9	12
10	CD4	10	10	12	10
11	SDHD	11	11	8	4
12	ZBTB7A	12	12	6	6

ACKNOWLEDGMENT

This work was supported in part by NSF grants CCF 0809895, CCF 1218764 and CSoI-CCF 0939370.

REFERENCES

[1] S. Aerts *et al.*, "Gene prioritization through genomic data fusion," *Nat. Biotechnol.*, vol. 24, pp. 537–544, 2006.

[2] J. Bartholdi, C. Tovey, and M. Trick, "The computational difficulty of manipulating an election," *Social Choice and Welfare*, vol. 6, no. 3, pp. 227–241, 1989.

[3] J. Chen*et al.*, "Improved human disease candidate gene prioritization using mouse phenotype," *BMC Bioinformatics*, vol. 8, pp.398, 2007.

TABLE II: Results for training genes MTHFR, CHI3L1, DISC1, SYN2, DRD3, DTNBP1, HTR2A, RTN4R, APOL4, implicated with **schizophrenia**.

Gene #	HGNC Symbol	Rank #	ToppGene	Generalized Kemeny	Endeavour
1	AKT1	1	1	1	1
2	HCN4	2	2	2	4
3	DAO	3	3	3	6
4	ADCY3	4	4	4	5
5	EPO	5	5	5 - 6	12
6	SOX3	6	6		7
7	LRAT	7	7	7	3
8	FGG	8	8	8	9
9	FGD3	9	9	9	2
10	NNT	10	10	10	8
11	ACLY	11	11	11	11
12	ICOS	12	12	12	10

TABLE III: Results for training genes CX3CR1, TLR3, HLA-C, CXCL12, IFNG, IL4R, CCL2, implicated with **HIV**.

Gene #	HGNC Symbol	Rank #	ToppGene	Generalized Kemeny	Endeavour
1	CXCR4	1	1	1 - 2 - 3 - 4 - 5 - 6	1
2	IL10	2	2		3
3	OSM	3	3		2
4	CRH	4	4		5
5	CD209	5	5		6
6	KIR3DL1	6	6		7
7	HFE	7	7	7	9
8	APC	8	8	10	8
9	RHO	9	9	8 - 9	11
10	SLC18A2	10	10		4
11	ABO	11	11	11	10
12	MCM6	12	12	12	12

[4] J. Chen *et al.*, "ToppGene Suite for gene list enrichment analysis and candidate gene prioritization," *Nucleic Acids Res.*, vol. 37, pp. W305–W311, 2009.

[5] V. Conitzer, A. Davenport, and J. Kalagnanam, "Improved bounds for computing Kemeny rankings," in *Proc. of the 21st National Conf. on Artificial Intelligence*, (Boston, Massachusetts), 2006.

[6] T. De Bie *et al.*, "Kernel-based data fusion for gene prioritization," *Bioinformatics*, vol. 23, pp. i125–i132, 2007.

[7] P. Diaconis, "Group representations in probability and statistics," *Lecture Notes-Monograph Series*, vol. 11, 1988.

[8] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation revisited." Manuscript, Available: http://www.eecs.harvard.edu/∼michaelm/CS222/rank2.pdf, 2001.

[9] F. Farnoud, O. Milenkovic, and B. Touri, "A Novel Distance-Based Approach to Constrained Rank Aggregation," CoRR abs/1212.1471, 2012.

[10] A. Hamosh *et al.*, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Research*, vol. 33, pp. D514-D517, 2005.

[11] T. Kacprowski *et al.*, "NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules," *Bioinformatics*, vol. 29, pp. 1471–1473, 2013.

[12] S. Köhler *et al.*, "Walking the interactome for prioritization of candidate disease genes," *Am. J. Hum. Genet.*, vol. 82, pp. 949, 2008.

[13] F. Raisali, F. Farnoud, and O. Milenkovic, "Weighted rank aggregation via relaxed integer programming," *Proceedings of the ISIT*, 2013.

[14] M. Rebhan *et al.*, "GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support," *Bioinformatics*, vol. 14, pp.656–664, 1998.

[15] D. Saari and R. Merlin, "A geometric examination of Kemeny's rule," *Social Choice and Welfare*, pp. 403-438, 2002.

[16] S. Yu *et al.*,"Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining," *Bioinformatics*, vol. 24, pp. i119–i125, 2008.