A Link Prediction based Unsupervised Rank Aggregation Algorithm for Informative Gene Selection

Kang Li*, Nan Du† and Aidong Zhang‡
Department of Computer Science and Engineering
State University of New York at Buffalo
Emails: {kli22, nandu† and azhang‡}@buffalo.edu*

***Abstract*—Informative Gene Selection is the process of identifying relevant genes that are significantly and differentially expressed in biological procedures. The microarray experiments conducted for this purpose usually implement only less than a hundred of samples to rank the relevance of over thousands of genes. Many irrelevant genes thus may gain statistical importance due to the randomness caused by the *small sample problem*, while relevant genes may lose focus in the same way. Overcoming such a problem goes beyond what a single microarray dataset can offer and stresses the use of multiple experiment results, which is defined as rank aggregation. In this paper, we propose a novel link prediction based rank aggregation algorithm for the purpose of informative gene selection. Each rank is transferred into a fully connected and weighted network, in which the nodes represent genes and the weights of links stand for priorities between connected nodes (genes). The integration of multiple gene ranks is then formulated as an optimization problem of link prediction on multiple networks, with criterion function favoring the maximization of weighted consensus among each network. We solve the problem through iterative estimation of weights and maximization of consensus among them. In the experimental evaluation, we demonstrate our method on the *Prostate Cancer Dataset* and compare it with other baseline methods. The results show that our link prediction based rank aggregation method remarkably outperforms all the compared methods, which proves the effectiveness of our framework in finding informative genes from multiple microarray experimental results.**

***Keywords*-Informative Gene Selection; Rank Aggregation; Link Prediction;**

I. Introduction

Informative Gene Selection is crucial in analysis of relationships between inherent genes and external biological processes. It is claimed as one of the most challenge problems in post-array analysis, and attracts intensive research interests despite a decade of research. Microarray datasets used for this purpose usually contain more than several thousand genes. However, due to the high cost of DNA microarray assessment, usually only less than a hundred of samples are used to discriminate genes expressing differentially under normal and reference conditions, and to rank genes according to how informative they are. The machine learning techniques involved in the discriminating and ranking usually suffer from the *small sample problem*, in which small amount of samples easily overfit large number of features. As a result, irrelevant genes may obtain statistical relevance and relevant genes may lose focus due to the randomness caused by the overfitting.

Addressing the problem goes beyond what a single microarray dataset can offer and stresses the importance of integrating multiple studies. As the rapid development of bioinformatics and biomedicine research, many microarray experiments are conducted toward the same biological procedure, providing the opportunity of achieving higher performance through integration. However, directly combining samples from various sources is usually impossible because of the different gene formats and experimental conditions used in various studies. Existing researches thus focus on integrating gene ranks from multiple experiments, which is defined as rank aggregation in machine learning.

As a classical problem, rank aggregation aims at obtaining a better ordering through aggregating different ranks on the same set of candidates or alternatives. Extensive studies have been conducted in this context, where each rank comes from different experiments, conditions, sources or voters. For this problem, Borda [1] proposed a solution by assigning a score to the order of each candidate and sorting them by minimizing the total score. Out of a similar intuitive, Kemeny proposed to optimize ordering σ by minimizing the sum of the "bubble sort" distances $\sum_{i=1}^{k} K(\sigma_i, \tau_i)$, with ordering τ_1, ..., τ_k on alternatives $\{1, 2, ..., n\}$. However, solving the Kemeny optimization function is NP-hard [2].

In this paper, we propose a novel link prediction based rank aggregation method for informative gene selection from multiple ranks. We transfer each rank into a fully connected and weighted network, in which the nodes represent genes and the weights of links represent priorities between the connected nodes (genes). Rank aggregation is then formulated into an optimization problem of link prediction from multiple networks, in which we learn a new network which maximizes the weighted consensus to all the original networks. With estimation of the weights and consensus through mutual information theory, we propose an iterative solution for the deduced optimization problem.

The major contributions of this paper are:

• We propose a novel link prediction based rank aggregation model to solve the problem of informative genes selection

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

from multiple microarray ranks.

• It is the first work to investigate the rank aggregation problem from the link prediction point of view. In this process, we also have given out a novel model for link prediction from multidimensional weighted networks. Since our model is general, it can be further extended to other link prediction and rank aggregation problems.

II. METHOD

Suppose for a specified biological process/function, there are n candidate genes $G = \{g_1,\ g_2,\ ...,\ g_n\}$ to be ranked according to their informativeness. m independent experiments are performed on different subsets $S_1, S_2, ... , S_m$ of G and output the ranked lists of G. We will call these outputs as source ranks $r^1,\ r^2,\ ...,\ r^m$. Without loss of generality, we assume that $S_1 \bigcup S_2 \bigcup ... \bigcup S_m = G$, which states that a valid evidence-based rank aggregation method can output rank for each candidate only if there exists rank information of it in at least one of the source ranks. In each output rank, top-K genes are deemed to be informative.

A. Model Formulation

Our goal is to transform each of the m source ranks into a network, and then learn the aggregated rank by link prediction on the m formed networks $A^1,\ A^2,\ ...,\ A^m$. To avoid the difficulty introduced by different dimensionality of them, A^i for $i = 1,\ 2,\ ...,\ m$ contains the complete set of G as nodes. We set each network to be fully connected and use A^i_{jk} to denote the weight between node j (gene j) and node k (gene k) in network A^i. The weights between pairs of nodes will represent the relative ordering of the connected nodes. Ideally, if S_i is the complete set of G, weight between any pair of genes could be calculated as $A^i_{jk} = R^i(g_k) - R^i(g_j)$, in which $R^i(g_j)$ is the output order of gene j in r^i. However, since each experiment investigates only a subset of G, the main difficulty in transforming each rank to the proposed network is in calculation of the weights between connected nodes when one or two of them are uninvestigated. The most popular idea to handle this problem in existing works [3], [4] is to view the orders of uninvestigated genes as $l^i + 1$, in which l^i is the size of the investigated subset S_i. Such methods ignore the fact that informative genes may be uninvestigated, and thus rank uninvestigated genes right after all investigated ones in each ordering. The presumption will cause high bias against real facts. To avoid this, we compromise it with utilization of information provided from other ranks.

Since the task is to select top-K informative genes from the pool G, if a gene g_j does not appear in the ranked list r^i of the experiment i, there are three possible situations: 1) it is investigated and identified as uninformative; 2) it is uninvestigated and informative in ground truth; and 3) it is uninvestigated and uninformative in ground truth. For simplicity here, without any prior knowledge about source experiment settings, we view whether a gene g_j is informative and whether it is investigated as two independent distributions, and denote them as INF and INV, respectively. The probability of situation 1) is then formulated as $p^i_1 = p^i(INV = 1, INF = 0) = p^i(INV = 1)p(INF = 0) = \frac{l^i}{n} \cdot \frac{(n-K)}{n}$. In similar ways, we can get $p^i_2 = p^i(INV = 0, INF = 1) = \frac{(n-l^i)}{n} \cdot \frac{K}{n}$ for 2) and $p^i_3 = p^i(INV = 0, INF = 0) = \frac{(n-l^i)}{n} \cdot \frac{(n-K)}{n}$ for situation 3). Then if g_j does not appear in r^i, its rank $R^i(g_j)$ is estimated as:

$$R^i(g_j) = \frac{(p^i_1 + p^i_3)}{\sum_{q=1}^{3} p^i_q} \cdot R^i_{uinf} + \frac{p^i_2}{\sum_{q=1}^{3} p^i_q} \cdot R^i_{inf}(g_j), \quad (1)$$

where R^i_{uinf} and R^i_{inf} are estimated ranks for uninformative and informative cases; probabilities p^i_q for $q \in [1, 3]$ is normalized so as summed up to 1. In our case, we follow the existing methods to set $R_{uinf} = l^i + 1$, and calculate $R_{inf}(g_j)$ through averaging ranks of g_j in other experiments. The **Equation 1** is then formulated as:

$$R^i(g_j) = \alpha^i \cdot (l^i + 1) + (1 - \alpha^i) \cdot \frac{\sum_{k=1}^{m} (\delta(r^k, g_j) * R^k(g_j))}{\sum_{k=1}^{m} \delta(r^k, g_j)}, \quad (2)$$

in which, $\alpha^i = \frac{n(n-K)}{n^2 - l^i K}$; $\delta(r, g) = 1$ if g appear in r, else $\delta(x) = 0$.

Notice that the original purpose of rank aggregation is to achieve a better rank with information from multiple experimental ranks. After transforming them into networks, the general objective of link prediction based rank aggregation is to ensemble networks from different sources to one optimal network, and then calculate the rank based on the optimal network. The integration of networks is formally named as link prediction on multidimensional networks, in which we try to obtain an optimal pattern that is as close to each network as possible. Existing works on this topic usually do not consider the bias of each dimension and perform an unweighted combination of all the networks. In light of this, we come up with a novel weighted link prediction method. The intuition of our framework can be expressed as: Given a set of networks $\{A^1, A^2, A^3,\ ...,\ A^m | A^i \in R^{n \times n}\}$ and a set of non-negative weights $W = \{w_i | w_i \in R_+, i = 1,\ 2,\ ...,\ m\}$, the optimal embedding network $B \in R^{n \times n}$ is given by the objective function as:

$$J(W, B) = \max_{B, W\, s.t.\ \sum w_i = 1} \sum_{i=1}^{m} w_i \cdot \|d(A^i, B)\|, \quad (3)$$

in which $d(A, B)$ is a measure of the consensus between network A and B.

The reverse converting from optimal network B to output rank r^o can be achieved by sorting the array $\{\sum_{j=1}^{n} B_{ij} | i \in [1, n]\}$ descendingly. We denote this operation as $r^o = sort(B_i)_{i=1}^{n}$.

B. Link Prediction based Rank Aggregation

By the above definition, a weight function for each source rank and a distance function capturing the consensus between two networks are needed to solve the above objective. Formally, we use mutual information for the latter one. In information theory, it measures the information shared between two objects. We assume the more information two networks share, the higher consensus they have. Theoretically, the mutual information between two matrices representing networks X and Y can be defined as:

$$I(X,Y) = \int\int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy. \tag{4}$$

A highly related measurement is the Kullback-Leibler (KL) divergence, which calculates the distance between two distribution $p(x)$ and $q(x)$ as:

$$KL(p\|q) = \int p(x)\log\frac{p(x)}{q(x)}dx. \tag{5}$$

The mutual information between two networks X and Y can then be expressed as a special case of KL-divergence:

$$I(X,Y) = KL(p(X,Y)\|p(X)p(Y)). \tag{6}$$

Set $d(A^i,B) = I(A^i,B)$ for $i \in [1,m]$, the objective function is formulated as:

$$J(W,B) = \max_{B,W\, s.t.\ \sum w_i=1} \sum_{i=1}^{m} w_i \cdot KL(p(A^i,B)\|p(A^i)p(B)). \tag{7}$$

However, calculation of KL-divergence is very time-consuming and also requires the probability distributions $p(A^i,B)$, $p(A^i)$ and $p(B)$ of networks A^i and B, which are not applicable in our case. Luckily, we don't need to accurately calculate them. Instead, we only need to maximize the objective function. In information theory, there are many divergences that can achieve the same maximization purpose with KL-divergence. One of such divergence [5] is $I(X;Y) = \int\int (p(x,y) - p(x)p(y))^2 dxdy$. By this divergence, we could divide the probability distribution calculation on networks into integration of that on vectors, and the results will always be non-negative and symmetric.

Setting weight for each source network manually is the easiest way to handle the bias of contributions over them. However in reality, we seldom have enough prior knowledge for it. Invoked by the probability divergence calculation, we use prior probability to represent each weight, and adjust them during the optimization process. Thus we have: $w_i = p(A^i)$, for $i \in [1,m]$. Then the objective function can be finalized as:

$$J(W,B) = \max_{B,W} \sum_{i=1}^{m} p(A^i)\cdot\sum_{j=1}^{n}(p(A^i,B_j) - p(A^i)p(B_j))^2, \tag{8}$$

in which B_j is the j-th column of network matrix B.

In the above criterion function $J(W,B)$, $p(B_j)$ can be viewed as the importance of each column in the optimized matrix B, which can be generally assumed to be equal, thus $p(B_j) = \frac{1}{n}$. Besides this, we need to set an estimation of variables' probability density function. After combining the mutual information with the Parzen Window method which makes it nonparametric, we set the variables' probability density function to a Gaussian Kernel as: $p(x) = \frac{1}{n}\sum_{i=1}^{n} G(x - x_i, \sigma^2)$, where $G(x - x_i, \sigma^2) = \frac{1}{2\pi\sigma^{n/2}} exp\{\frac{-(x-x_i)^2}{2\sigma^2}\}$, and σ is the size of the Parzen window. Then $p(B_j|A^i) = \frac{1}{n}\sum_{k=1}^{n} G(B_j - A_k^i, \sigma^2)$. The objective function is then further formulated as:

$$\begin{aligned} J(W,B) &= \max_{B,W} \sum_{i=1}^{m} p(A^i)\cdot\sum_{j=1}^{n}(p(A^i,B_j) - p(A^i)p(B_j))^2 \\ &= \max_{B,W} \sum_{i=1}^{m} p(A^i)\cdot\sum_{j=1}^{n}(p(B_j|A^i)p(A^i) - p(A^i)p(B_j))^2 \\ &= \max_{B,W} \sum_{i=1}^{m} p(A^i)^3\cdot\sum_{j=1}^{n}(\sum_{k=1}^{n} G(B_j - A_k^i, \sigma^2) - 1)^2. \end{aligned} \tag{9}$$

C. Model Inference

To maximize the criterion function, we present an iterative EM algorithm, in which we iteratively estimate the prior probability of each source network then maximize the criterion function. The framework can be divided into:

E-Step: In each iteration, we estimate the prior probability of each source network by the ratio of mutual information as:

$$p(A^i) = \frac{I(A^i,B)}{\sum I(A^i,B)} = \frac{\sum_{j=1}^{n}\sum_{k=1}^{n} G(B_j - A_k^i, \sigma^2)}{\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{n} G(B_j - A_k^i, \sigma^2)}. \tag{10}$$

By this definition, the sum of possibilities of the source networks will be guaranteed to be 1 so that the condition $\sum_{i=1}^{m} w_i = 1$ is met. Furthermore, it satisfies the intuition that the higher the mutual information a network has, the higher prior probability it will gain. In this extreme case when $B = A^i$, it will reach the maximum 1.

M-Step: By the estimation of prior possibilities, we can iteratively calculate the optimal matrix B by the following derivation:

$$\begin{aligned} \frac{dJ(W,B)}{dB_j} = \sum_{i=1}^{m} p(A^i)^3\cdot(\sum_{k=1}^{n} G(B_j - A_k^i, \sigma^2) - 1)\cdot \\ (\sum_{k=1}^{n}\frac{(A_k^i - B_j)}{\sigma^2}\cdot G(B_j - A_k^i, \sigma^2)). \end{aligned} \tag{11}$$

Set $\frac{dJ(W,B)}{dB_j} = 0$, we get:

$$B_j(t+1) = \frac{\sum_{i=1}^{m} p(A^i)^3\cdot(M_{ij} - 1)\cdot N_{ij}}{\sum_{i=1}^{m} p(A^i)^3\cdot(M_{ij} - 1)\cdot M_{ij}}, \tag{12}$$

in which $B(t)$ represents the optimal matrix B at the $t-th$ iteration, $M_{ij} = \sum_{k=1}^{n} G(B_j(t) - A_k^i, \sigma^2)$ and $N_{ij} = \sum_{k=1}^{n} A_k^i \cdot G(B_j(t) - A_k^i, \sigma^2)$.

By the above EM process, we can iteratively optimize the matrix B to weightedly ensemble network matrices from different source ranks. To start the iteration, we initialize the network prior probability $p(A^i) = \frac{1}{m}$, for $i = 1,\ 2,\ ...,\ m$, and $B(1) = \sum_{i=1}^{m} p(A^i) \cdot A^i$. The maximal change of weights between two iteration $\epsilon(t) = max(abs(p(A^i, t) - p(A^i, t-1)))$ will then be used as the stop condition: if $\epsilon(t) \leq ERROR$ then stop, else continue, where $abs(X)$ converts X to its absolute values, $max(X)$ outputs the maximal value in X and $ERROR$ is an experience-based error bound.

III. Experiments

In this section, we carry out experiments to evaluate the effectiveness of our link prediction based rank aggregation method on the **Prostate Cancer Dataset**, which includes five microarray informative gene ranked lists. We compare our approach to five baseline methods, and show that our method significantly outperforms the comparative approaches on informative gene identification.

A. Dataset Description

The **Prostate Cancer Dataset** has already been used for rank aggregation in several researches including [3], [4] and [6]. It consists of output ranked lists from five microarray experiments, each of which investigates top-25 ranked informative genes that were suspected to be differently and significantly expressed in prostate tumors studies which are Luo [7], Welsh [8], Dhana [9], True [10] and Singh [11]. Since we don't have any information on the total number of informative genes existing in the candidate pool, we attempt to obtain an optimal top-25 ranked list based on the five given microarray experimental results. We present the details of this dataset in **Table I**.

As mentioned by Deconde et al in [12], 9 genes have already been identified as informative for prostate cancer, which include HPN, AMACR, FASN, GUCY1A3, ANK3, STRA13, CCT2, CANX and TRAP1. These identified informative genes will be used as ground truth to help evaluate the performance of each rank aggregation method.

B. Evaluation Metrics

Given the ground truth of informative genes, it is very straight forward that we could evaluate the performance of each rank aggregation method by the number of known informative genes appearing in the top-25 ranked list, and by average ranks of all informative genes.

C. Baseline Approaches

We compare our experimental results with five baselines, including: four types of Borda [1] (denoted as Borda1, Borda2, Borda3 and Borda4), three types of Markov Chain (MC) [12], [13] (denoted as MC1, MC2 and MC3), four types of Cross Entropy Monte Carlo (CEMC) [6] (denoted as CEMC1, CEMC2, CEMC3 and CEMC4), Artificial Fish Swarm Algorithm (AFSA) [3] and Graph-based Consensus Maximization (GCM) [4]. The results of these five comparative algorithms on **Prostate Cancer Dataset** are collected from [4] and [3].

D. Experimental Results and Discussions

Given the five experimentally ranked lists on the **Prostate Cancer Dataset**, we applied our model to integrate them with parameters $ERROR = 0.01$ and $\sigma^2 = 2000$.

Table I shows the experimental inputs and the output of our approach, in which $Rank$ denotes the order of each candidate, columns 2 to 6 represent the input experimental ranked lists and the $result$ is the top-25 informative genes obtained through our model. We present the results evaluated by the proposed metric in **Table II**, where we denote number of known informative genes appearing in the list for each method by $Appear\#$, and represent average ranks of the 9 informative genes by $Avg.Rank$. In calculation of the average ranks, if a known informative gene does not appear in the top-k ranked list, we follow the setting in [3], [4] to assign its order to be $k+1$ (26 in this case).

The iterative updating strategy will finally achieve its optimal result when the weights for the source lists are 0.1392, 0.2892, 0.2339, 0.1695 and 0.1682. An interesting finding is that the rank of these weights is very close to the ranks of their $Appear\#$ and $Avg.Rank$ as shown in **Table II**. Welsh has the smallest $Avg.Rank$, largest $Appear\#$ and highest weight over all; Dhana performs second to Welsh in the proposed evaluation metrics and obtains less weight than it; and Singh and Luo also have such coincidences. The only counter-case is True. It has similar $Appear\#$ and $Avg.Rank$ as Luo, but has much higher weight than it. Moreover, it even obtains slightly higher weight than Singh, which has one more known informative gene and much less $Avg.Rank$ than it. The singularity of True can be explained by the following two observations: first, every source experiment ranks AMACR in the top-2 except for Singh, which does not even include it into the top-25 ranked list. Such disagreements with other sources will drive down the consensus of it and hereby cause lower weight; and second, as shown in **Table I**, True shares 11, 5 and 8 candidates with Welsh, Dhana and Singh, respectively, while Luo only shares 3, 3, and 1 candidates with them. It has significantly less consensus than True, thus obtains much less weight.

As we can see from **Table II**, our approach not only contains largest $Appear\#$ (7 out of 9), but also achieves the lowest average rank when compared to others. Since the intuition of rank aggregation is to achieve an optimal rank through integrating information from multiple ranks, a

Table I: Detail Experimental Input and Output Ranks

Rank	Luo	Welsh	Dhana	True	Singh	Result
1	**HPN**	**HPN**	OGT	**AMACR**	**HPN**	**HPN**
2	**AMACR**	**AMACR**	**AMACR**	**HPN**	SLC25A6	**AMACR**
3	CYP1B1	OACT2	**FASN**	NME2	EEF2	**FASN**
4	ATF5	GDF15	**HPN**	CBX3	SAT	OACT2
5	BRCA1	**FASN**	UAP1	GDF15	NME2	GDF15
6	LGALS3	**ANK3**	**GUCY1A3**	MTHFD2	LDHA	KRT18
7	MYC	KRT18	OACT2	MRPL3	**CANX**	UAP1
8	PCDHGC3	UAP1	SLC19A1	SLC25A6	NACA	**ANK3**
9	WT1	GRP5	KRT18	NME1	**FASN**	NME1
10	TFF3	PPIB	EEF2	COX6C	SND1	**STRA13**
11	MARCKS	KRT7	**STRA13**	JTV1	KRT18	GRP58
12	OS-9	NME1	ALCAM	CCNG2	RPL15	OGT
13	CCND2	**STRA13**	GDF15	AP3S1	TNFSF10	PPIB
14	NME1	DAPK1	NME1	EEF2	SERP11	KRT7
15	DRRK1A	TMEM4	CALR	RAN	GRP58	EEF2
16	**TRAP1**	**CANX**	SND1	PPKACA	ALCAM	**GUCY1A3**
17	FM05	TRA1	STAT6	RAD23B	GDF15	**CANX**
18	ZHX2	PRSS8	TCEB3	PSAP	TMEM4	DAPK1
19	RPL36AL	EMTPD6	EIF4A1	**CCT2**	**CCT2**	TMEM4
20	ITPR3	PPP1CA	LMAN1	G3BP	SLC39A6	NME2
21	GCSH	ACADSB	MAOA	EPR3	RPL5	SLC25A6
22	DDB2	PTPLB	ATP6VOB	CKAP1	RPS13	CBX3
23	TFCP2	TMEM23	PPIB	LIG3	MTHFD2	ALCAM
24	TRAM1	MRPL3	FM05	SNX4	G3BP2	SAT
25	YTHDF3	SLC19A1	SLC7A5	NSMAF	UAP1	TRA1

Table II: Experimental Results

Methods	Appear #	Avg. Rank
Luo	3	19.4
Welsh	6	13.4
Dhana	5	14.1
True	3	19.8
Singh	4	18.4
Borda1	5	15.2
Borda2	4	17.4
Borda3	**7**	15.8
Borda4	5	15.2
MC1	5	16.1
MC2	6	15.5
MC3	6	14.7
CEMC1	6	14.1
CEMC2	6	14.4
CEMC3	5	14.8
CEMC4	6	14.6
AFSA	**7**	13.7
GCM	**7**	13.2
Result	**7**	**12.1**

solid method should output a ranked list that is better than any individual source list. But as we can conclude from the results in the table, Borda, MC, CEMC and AFSA can not achieve this goal in both of the two proposed evaluation metrics. Borda count method does include more informative genes than any of the initial list if sorting by geometric mean, its average rank is higher than Welsh's, and so does AFSA method. MC and CEMC can beat Luo, True and Singh in both evaluation metrics, and outperform Dhana on $Appear\#$ in some cases, but still perform significantly worse than Welsh. GCM and our approach are the only two that achieve the goal of better than single source list in both $Appear\#$ and $Avg.Rank$, and both of these methods are handling multiple sources from the graph theory point of view, which proves the benefits of doing rank aggregation under the framework of networks. When comparing our detailed rank to the best rank of each method, which are Borda3, MC3, CEMC1, AFSA and GCM, Borda3 successfully includes 7 informative genes, but it ranks ANK3, STRA13 and GUCY1A3 relatively lower (larger orders), causing higher average rank; the ranks of MC and CEMC are very close to each other, but both of them miss ANK3 and GUCY1A3 which lift up the $Avg.Rank$. Our method finds the same

number of informative genes as AFSA and GCM. Our method misses CCT2 and AFSA and GCM miss GUCY1A3 and ANK3, respectively. However, AFSA ranks ANK3 much lower than others, causing higher $Avg.Ranks$ in comparison. In GCM, the orders of CANX and GUCY1A3 are much larger than our rank, and in the overall comparison they lift up the $Avg.Ranks$. As a result, our approach remarkably outperforms all comparative methods in informative gene selection through rank aggregation.

Another interesting finding is that all of these aggregation approaches identify NME2, GDF15, KRT18, EEF2, OACT2, SLC25A6 and GRP58 as the top-25 informative genes, which suggests that they may have higher potential to be experimentally identified as informative in further biological studies. In the contrast, none of these listed methods successfully includes TRAP1 to the top-25. We may need to integrate more knowledge, such as gene ontology, into the aggregation framework to solve the problem in future studies.

IV. Conclusions

In this paper, we have introduced a link prediction based rank aggregation method for informative gene selection, which is an information theory based weighted combination of transformed networks that ranks candidate genes by simultaneously considering orders of them in multiple ranked lists. The main advantages of our method are the representation of different lists by network matrices, and the assignment of different weights to each of them, which are automatically calculated through the information theory, providing robustness towards noisy or low quality ranks. Also, a single straightforward EM procedure is used to iteratively optimize the proposed objective function and achieve the maximization of consensus between the optimal rank to each of the source ranked lists, without any prior knowledge of its importance.

The proposed framework has been tested on the well-known **Prostate Cancer Dataset** and compared with plenty of baseline methods. In general, the results verify the superiority of handling rank aggregation in the network theory point of view, and identifying informative genes through the proposed framework. Its performance is significantly better than any compared method with selecting the most informative genes in the top-25 ranked list and achieving much lower average ranks.

References

[1] J. C. Borda, "Mmoire sur les lections au scrutin," *Comptes rendus de lAcadmie des sciences traduit par Alfred de Grazia comme Mathematical Derivation of a election system Isis vol 44 pp 4251*, vol. 2, p. 85, 1781.

[2] J. Bartholdi III, C. Tovey, and M. Trick, "Voting schemes for which it can be difficult to tell who won the election," *Social Choice and Welfare*, vol. 6, no. 3, pp. 157–165, 1989. [Online]. Available: http://www.springerlink.com/index/10.1007/BF00303169

[3] N. Du, D. Suprita, B. Bindukumar, A. Stanley, B. Chiu, and A. Zhang, "An artificial fish swarm based supervised gene rank aggregation algorithm for informative genes studies." *Proceeding of Computational Intelligence and Bioinformatics*, vol. 753, no. 19, 2011. [Online]. Available: http://www.actapress.com/Abstract.aspx?paperId=452882

[4] L. Ge, N. Du, and A. Zhang, "Finding informative genes from multiple microarray experiments: A graph-based consensus maximization model," pp. 506–511, 2011.

[5] X. H. Dang and J. Bailey, "A hierarchical information theoretic technique for the discovery of non linear alternative clusterings," *Entropy*, vol. 12, no. 7, pp. 573–582, 2010. [Online]. Available: http://nms.sagepub.com/cgi/doi/10.1177/1461444809355648

[6] S. Lin and J. Ding, "Integration of ranked lists via cross entropy monte carlo with applications to mrna and microrna studies." *Biometrics*, vol. 65, no. 1, pp. 9–18, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18479487

[7] J. Luo, D. J. Duggan, Y. Chen, J. Sauvageot, C. M. Ewing, M. L. Bittner, J. M. Trent, and W. B. Isaacs, "Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling." *Cancer Research*, vol. 61, no. 12, pp. 4683–4688, 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11406537

[8] J. B. Welsh, L. M. Sapinoso, A. I. Su, S. G. Kern, J. Wang-rodriguez, C. A. Moskaluk, H. F. Frierson, and G. M. Hampton, "Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer advances in brief analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer," *Analysis*, vol. 1, no. 858, pp. 5974–5978, 2001. [Online]. Available: http://cancerres.aacrjournals.org/content/61/16/5974.full

[9] S. M. Dhanasekaran, T. R. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan, "Delineation of prognostic biomarkers in prostate cancer." *Nature*, vol. 412, no. 6849, pp. 822–826, 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11518967

[10] L. True, I. Coleman, S. Hawley, C.-Y. Huang, D. Gifford, R. Coleman, T. M. Beer, E. Gelmann, M. Datta, E. Mostaghel, and et al., "A molecular correlate to the gleason grading system for prostate adenocarcinoma," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 29, pp. 10 991–10 996, 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17421063

[11] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. DAmico, J. P. Richie, and et al., "Gene expression correlates of clinical prostate cancer behavior." *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12086878

[12] R. P. DeConde, S. Hawley, S. Falcon, N. Clegg, B. Knudsen, and R. Etzioni, "Combining results of microarray experiments: a rank aggregation approach." *Statistical Applications in Genetics and Molecular Biology*, vol. 5, no. 1, p. Article15, 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17049026

[13] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," *Proceedings of the tenth international conference on World Wide Web WWW 01*, vol. 16, no. 2, pp. 613–622, 2001. [Online]. Available: http://portal.acm.org/citation.cfm?doid=371920.372165