Ma et al. BMC Genomics 2015, 16(Suppl 9):S3
http://www.biomedcentral.com/1471-2164/16/S9/S3

BMC Genomics

RESEARCH Open Access

An unsupervised learning approach to find ovarian cancer genes through integration of biological data

Christopher Ma[1*†], Yixin Chen[1], Dawn Wilkins[1], Xiang Chen[2], Jinghui Zhang[2]

From IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2014)
Belfast, UK. 2-5 November 2014

Abstract

Cancer is a disease characterized largely by the accumulation of out-of-control somatic mutations during the lifetime of a patient. Distinguishing driver mutations from passenger mutations has posed a challenge in modern cancer research. With the advanced development of microarray experiments and clinical studies, a large numbers of candidate cancer genes have been extracted and distinguishing informative genes out of them is essential. As a matter of fact, we proposed to find the informative genes for cancer by using mutation data from ovarian cancers in our framework. In our model we utilized the patient gene mutation profile, gene expression data and gene gene interactions network to construct a graphical representation of genes and patients. Markov processes for mutation and patients are triggered separately. After this process, cancer genes are prioritized automatically by examining their scores at their stationary distributions in the eigenvector. Extensive experiments demonstrate that the integration of heterogeneous sources of information is essential in finding important cancer genes.

Introduction

Understanding cancer biology and the mechanism behind cancer progression has always been an important branch of cancer research. Thanks to the advancement of computational technology, a huge amount of biological data such as microarray gene expression is readily available. Researchers make use of gene expression profiles to predict clinical outcome of breast cancer and identify several cancer subtypes. This can help in elucidating the association between several molecular levels which enable us to identify the biological relationships and understand the molecular processes driving the cancer. This could potentially lead to improvement in cancer diagnosis and patient survival analysis prediction. Consequently, other types of high throughput biological data are produced which are of great interest and as a matter of fact, we aim to integrate different sources of data in a unified framework with an objective to locate important biomarkers responsible for cancer progression through a ranking method.

Our framework is encoded with various heterogeneous sources of data including 1) Protein protein interaction (PPI) Network: Cancer is not a disease of individual mutation but a group of genes interacting together in a molecular network. Hence incorporating PPI network and pathway interaction information in cancer studies is critical in discovering interactions among genes and deciphering the molecular pathway of cancer. 2) Gene Expression profiles: DNA microarray-based technology has provided researchers ample opportunity to perform comprehensive molecular and genetic profiling of cancer by simultaneously studying how thousands of genes were being expressed in hundreds of patients. We used the gene expression data and performed the Pearson correlation coefficients calculation to determine the correlation of the gene expressions of various genes in a bid to identify co-expressed genes which are responsible for cancer development. 3) Patient Somatic Mutation Profiles: A table which records the mutation profile of each patient is also included in our framework as background information.

* Correspondence: cma@go.olemiss.edu
† Contributed equally
[1]Department of Computer and Information Science, Weir Hall, University of Mississippi, University, MS 38677, USA
Full list of author information is available at the end of the article

BioMed Central

© 2015 Ma et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Separate Markov Chains on the genes and patients are defined and random walks are performed in order to obtain the results. Random walk based ranking work on cancer modules can be found in [2,4,5]. In [1], the authors utilize both the random walk and random walk with restart to rank genes with respect to their likelihood of being a member of each cancer module through the functional interaction network globally and interactions between genes in each cancer module locally. In [3], Erten, Bebek and Koyuturk investigate the topological similarity in PPI networks and suggest a random walk based algorithm to find genes with similar disease. They came up with a measure to calculate the topological profiles between the candidate genes and the driver genes. In [8], the authors utilize random walk and network community analysis for the identification of cancer-associated modules in gene expression data. In [9], Sharan, Ulitsky and Shamir survey a number of random walk related approaches, including direct methods and module-assisted methods. The algorithms propagate functional information and functional modules within the network, which are inferred for annotation purpose. Other techniques related to random walk are Markov field based propagation and Guassian Random Field propagation. The authors in [6,7] utilized those network propagation techniques in protein function prediction.

In [10], Zhang and Wei extend the general network propagation algorithm to consider graphs with nodes and edges to be positive and negative numbers for the sake of detecting differential gene expressions and DNA copy number variations (CNV). Gene up/down regulation or amplification/deletion CNV events are modelled to be positive and negative respectively. By exploiting the weighted connections between genes, gene labels are propagated sucessfully within the network. This method is capable of identifying hidden clusters to eliminate false positives and recover false negatives. However, it may also explore very weak similarities between genes as well. In our proposed random walk based framework, inspired by the algorithm of Google PageRank, randomness is introduced by permitting each gene to choose one patient to hop and each individual patient to choose one gene to hop randomly for minuscule amount of probability. In this method, both the gene nodes which are strongly connected and gene nodes with poor connectivity can also be exploited so as to discover biomakers globally through some noise introduced by random teleportation.

Our overall framework comprises five models which essentially differ from each other by the sequence and order the random walk is performed on our overall patient gene network. A multigraph is introduced in which the gene gene interaction network and the gene correlation network are merged thus multiple edges between a pair of gene nodes are allowed. Moreover, different sequences of traversal of this multigraph results in different transition matrices for the genes and patients respectively which culminates in our five different models. Our works sucessfully incorporates multiple heterogeneous data sources in a graph to find the cancer genes by computing the major eigenvector of each individual stationary matrix in each model and each gene is ranked in accordance with the corresponding value in the eigenvector rank. Comprehensive experiments demonstrate that the integration of heterogeneous sources of information is useful in discovering cancer genes and all six of the proposed models are able to rank those confirmed cancer genes as reported from other literature within top positions in the rank. The remainder of the paper is organized as follow: We will present the Methodology next. Afterwards, extensive experiments are performed and results are reported and tabulated. Lastly, we present conclusion and future work.

Methods

In this section, we illustrate how to represent the three sources of information.

Hetergeneous Sources of Data

1) Gene Gene Interaction: The gene gene interaction networks are encoded as an undirected graph $G(V, E)$ where V stands for the genes and edges $(i, j) \in E$ are weighted by a weight matrix W, whose element w_{ij} is the weight of the edge $(i, j) \in E$ which represents the strength of interaction between gene i and gene j using two sources of gene gene interaction networks described below. Two sources of protein protein interaction networks are utilized: Pathway-Commons and HumanNet v.1. PathwayCommons is a database of biological pathway information compiled from multiple sources related to PPI interactions and functional relationships between genes in signalling pathways. Only human genes and interactions in Pathway-Commons are utilized in our framework. HumanNet is a probabilistic functional gene network constructed using naive Bayesian method to weigh different types of data evidence collected in humans, yeast and worms in accordance with their functionality in Homo-sapiens. A single interaction score is calculated as a result. HumanNet v.1 consists of 18,714 validated protein encoding genes in total.

2) Gene Expression Profiles: Gene expression is measured through high throughput microarray experiments which show the expression level of a gene on each person. The gene expression data show the expression levels of genes in both the tumor and normal samples which are used for evaluating the similarity between genes, where genes with similar gene expression are often perceived to also carry similar functionality. We capitalize on the gene expression data and construct a gene correlation graph/network. To summarize, a gene correlation graph/network is a graph H(V, E), where V represents genes and an edge

$(i, j) \in E$ is weighted by calculating the Pearson correlation coefficient between their gene expression values.

3) Patient Mutation Profile: Patient-Mutation Profile is a two dimensional binary matrix with columns representing the genes and rows representing patients. Each entry is either 0 or 1, a 1 indicates that a mutation has occured in the tumor relative to the germline on that patient, a 0 otherwise.

Mutual Informative Model

In Mutual Reinforcement Model, (shown in Figure 1) each mutation (gene) is assigned a driver score μ_i and each patient is assigned a patient score π_i. Each patient is allowed to cast a vote on each mutation (gene) and vice versa. As a result, the driver score of a mutation (gene) is in proportion to the total votes the mutation (gene) received and the total votes received by the patient determine the patient score. Therefore, a high driver score means that the mutation is possessed by patients with high patient scores and a high patient score means that the patient possesses mutations with high driver scores. With the introduction of patient mutation profile, some notations can be laid out. Adjacency matrix B of a bipartite graph is defined as $B_{ij} = 1$ if and only if patient i has gene j mutated and 0 otherwise we use m to represent the total numbers of patients and n the total genes numbers. As the driver score of a mutation (gene) μ_i is directly prorportional to the number of patients possessing that mutation and the patient score π_i is directly proportional to the total number of mutations possesed by the patient, the mutation score and patient score are mutually defined relative to each other and the equations below are justified.

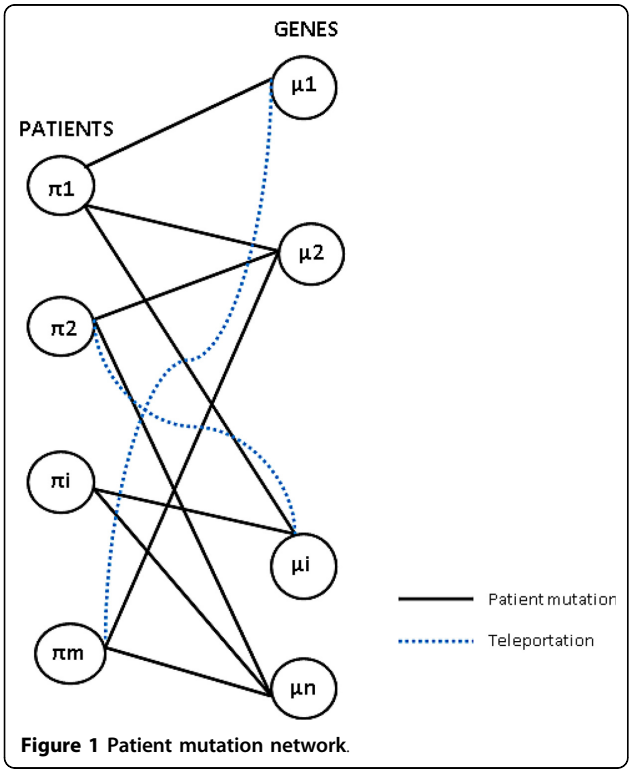

Figure 1 Patient mutation network.

$$\mu_i \propto \sum_{i \in l: B_{kj=1}} \pi_i$$

$$\pi_i \propto \sum_{j \in k: B_{ik=1}} \mu_j$$

To start with, the probability in which a mutation (gene) j traverses to patient i is defined by the following matrix:

$$B_c|i,j| = \frac{B_{ij}}{\sum_{k=1}^{m} B_{kj}}$$

Likewise, the probability in which a patient i traverses to mutation (gene) j is governed by the following matrix:

$$B_r|i,j| = \frac{B_{ij}}{\sum_{k=1}^{n} B_{ik}}$$

Notice that B_r is a row stochastic matrix whereas B_c is a column stochastic matrix. Randomness is introduced by allowing each patient to select arbitrarily a mutation (gene) to teleport for a small amount of time besides following the incident edges and hop to one of his neighbors in the gene partite set with the probability governed by matrix B_r for most of the time. The factor $1 - \alpha$ defines the probability in which the patient relinquishes the matrix B_r for traversal and use teleportation for traversal. Since each mutation (gene) has equal probability to be chosen by each patient. The following transition matrix for patients is justified:

$$C_r|i,j| = \alpha * B_r + (1-\alpha)\frac{1}{n} I_{m*n}$$

where I_{m*n} is an m by n matrix with all entries equal 1. Likewise for mutations (genes), each mutation (gene) is allowed to select arbitraily a different patient to teleport for a small percentage of time and the following transition matrix for the mutations (genes) is defined.

$$C_c[i,j] = \alpha * B_c + (1-\alpha)\frac{1}{m} I_{m*n}$$

Next, we incorporate the gene gene interaction network and gene correlation network in our model for the mutation of genes. As described above, a gene gene interaction network can be encoded as an undirected graph $G(V, E)$ where V represents the genes and edges $(i, j) \in E$ are weighted by a weight matrix W, whose element w_{ij} represents the interaction strength between gene i and gene j. Afterwards, normalization of matrix

W is performed in order to define the transition probability matrix within the genes and hence the transition probability matrix $Q = D^{-1}W$ is a result. Note that D ia a diagonal matrix with diagonal elements $d_{ii} = \Sigma_j\ w_{ij}$ and the elements q_{ij} of Q defines the probability of a random hop from gene i to gene j. The matrix Q satisfies the probabilistic constraint $\Sigma_j\ q_{ij} = 1$. As a result, the transition matrix of the genes following the PPI interaction network is defined:

$$C_g[i,j] = Q - D^{-1}W$$

Afterwards, gene expression data as shown in Figure 2 can come into play in the framework by constructing a gene correlation network. Each edge in the gene correlation network is encoded by matrix H whose elements $H(i, j)$ represents the Pearson's correlation coefficient between the gene expression of gene i and gene j. To derive, we let β_u to be the gene expression vector of gene u on the patients and so

$$H(u,v) = corr(\beta_u, \beta_v)$$

$$= \frac{\Sigma t \in V(B_u(t) - \frac{1}{V})(\beta_v(t) - \frac{1}{V})}{\sqrt{\Sigma t \in V(\beta_u(t) - \frac{1}{V})^2}\sqrt{\Sigma t \in V(\beta_v(t) - \frac{1}{V})^2}}$$

where corr(X, Y) represents the Pearson correlation of random variable × and random variable Y. The intuition behind it is that the gene expression of two genes may be correlated to each other if they are both partcipating in the same cancer pathway. Next, normalization of matrix H to define the transition probability is performed as above which results in the following transition matrix of the genes for the gene correlation network:

$$C_h = T = D^{-1}H$$

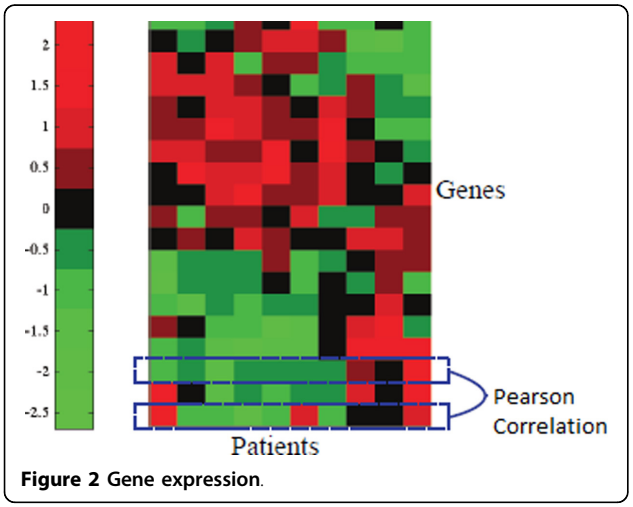

Figure 2 Gene expression.

Markov Chains

In this section, a series of models using the transition probability matrices are defined. Figure 3 shows the overall models used in our framework.

1. Random Walk Multiplicative Model Gene Correlation Start(RW-MMGCS):

The model starts with a random walk on gene correlation network and followed by PPI network and then to patient and back to gene correlation network and iterate. The stationary distribution of the random walk is defined as follow:

$$\begin{aligned} \mu &= C_r^T C_c C_g^T C_h^T \mu \\ \pi &= C_c C_g^T C_h^T C_r^T \pi\,. \end{aligned} \tag{1}$$

2. Random Walk Multiplicative Model Gene Interaction Start(RW-MMGIS):

The model starts with random walk on gene gene interaction network and followed by gene correlation network and then forward to patient and back to PPI network and reiterate. The stationary distribution of the random walk is defined as follow:

$$\begin{aligned} \mu &= C_r^T C_c C_h^T C_g^T \mu \\ \pi &= C_c C_h^T C_g^T C_r^T \pi\,. \end{aligned} \tag{2}$$

3. Random Walk Additive Model (RW-AM):

This random walk framework is slightly different from the above random walk settings in the sense that the transition matrix for the mutations is defined as a linear combination of the transition matrix of gene gene interaction network, patient mutation profile and the gene correlation network with each transition matrix responsible for a part of the overall transition matrix. It is

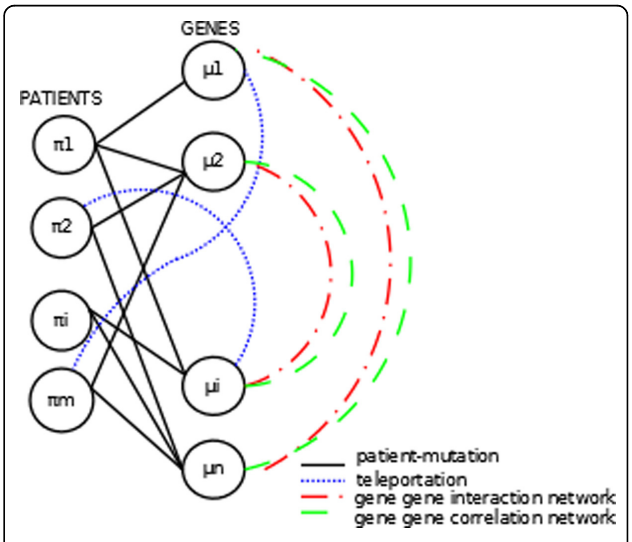

Figure 3 Overall patient mutation network.

defined as follow:

$$\begin{aligned} \mu &= (\alpha * C_r^T C_c + \beta * C_g^T + \gamma * C_h^T)\mu \\ 1 &= \alpha + \beta + \gamma \\ \pi &= C_c C_g^T C_h^T C_r^T \pi . \end{aligned} \tag{3}$$

4. Random Walk Multiplicative Model Penalised: (RW-MMP)

In this model, the gene gene interaction network and gene correlation network is combined into one network. As described above, the gene gene interaction network is encoded by a weight matrix W, whose element w_{ij} represents the interaction strength between gene i and gene j and the gene correlation network is weighted by matrix H whose elements H_{ij} is the Pearson's correlation coefficients between the gene expression of gene i and gene j. Each edge in the PPI network encoded by matrix W is penalized by the expontential value of its corresponding gene correlation value weighted by matrix H divided by the average. Then the random walk starts with overall gene network and then to gene patient and then back to overall gene network and repeats. The following equations are justified.

$$\begin{aligned} \sigma &= \text{mean of all entries of matrix H} \\ W_{ij} &= W_{ij} \exp(H_{ij}/\sigma) \\ C_g[i,j] &= Q = D^{-1} W \\ \mu &= C_r^T C_c C_g^T \mu \\ \pi &= C_c C_g^T C_r^T \pi . \end{aligned} \tag{4}$$

5. Random Walk Multiplicative Model Average: (RW-MMA)

In this model, we take the average output from the random walk on gene gene interaction network, gene correlation network and patient mutation profile in each iteration. The algorithm is tabulated in Algorithm 1.

6. Random Walk Multiplicative Patient Profile Start (RW-MMPFS):

The model starts with random walk on gene-patient network, followed by PPI network and then to gene correlation network and then back to gene correlation network and iterate. The stationary distribution of the random walk is defined as follow:

$$\begin{aligned} \mu &= C_h^T C_g^T C_r^T C_{c\mu} \\ \pi &= C_c C_g^T C_h^T C_r^T \pi . \end{aligned} \tag{5}$$

Algorithm 1 Random Walk Multiplicative Model Average algorithm (RW-MMA)

procedure Random Walk Multiplicative Model Average (*l, m, r, Cg, Cr, Cc, Ch*)

R_0 $\Re$ all entries are 1/n

for t = 1 to max(*l,m,r*) **do**

if t <= l **then** $R_{left} = C_r^T C_c * R_{t-1}$

end if

if t <= m **then** $R_{mid} = C_g^T * R_{t-1}$

end if

if t <= r **then** $R_{right} = C_h^T * R_{t-1}$

end if

$$R_t = \frac{(\sigma_{t<=l} * R_{left} + \sigma_{t<=m} * R_{mid} + \sigma_{t<=r} * R_{right})}{\sigma_{t<=l} + \sigma_{t<=m} + \sigma_{t<=r})}$$

σ_t <= x = 1 if t <= x and 0 otherwise

end for

return ^{Rt}

end procedure

In the aftermath of defining various models, it remains to demonstrate that all the Markov chains in all the five proposed models are valid and all the corresponding transition matrices converge to unique stationary matrices which result in a unique eigenvector as our ranking vector in each model.

Lemma: All the above transition matrices define valid Markov Chains that converge to a unique stationary eigenvectors.

For the sake of simplicity, the proof of the model Random Walk Multiplicative Model Gene Interaction Start (RWMMGIS) is outlined as below, the rest of the models can be proved similarly. Convergence: To prove convergence, we must prove the Markov chain defined by the transition matrix $C_r^T C_c C_h^T C_g^T$ is irreducible and aperiodic. Notice that each mutation is permitted to teleport to any patient and each patient is permitted to teleport to any mutation with a small probability. Coupled with the definitions of B_r and B_c, all entries in matrix C_r and C_c are strictly positive. Since C_h and C_g are also positive stochastic with nonnegative entries, the transition matrix defined by $C_r^T C_c C_h^T C_g^T$ are all strictly greater than 0 in all entries. This proves that every state in the state space S can be reached from every other state in the state space in a finite number of moves with positive probability which proves irreducibility. For aperiodicty, notice the fact that each P_{ii} >0 which implies that the minimum number of steps from each state i returning to itself is 1 which proves aperiodicity. Uniqueness: To prove uniqueness, notice that C_r is a row stochastic matrix and hence C_r^T is column stochastic, in addition, C_r^T, C_r^T, C_c, C_g^T are all positive column stochastic and hence the product of positive column stochastic matrices is also positive column stochastic. By Perron-Frobenius Theorem, 1 is an eigenvalue of multiplicity one of the matrix $C_r^T C_c C_h^T C_g^T$ which is the largest and all the other eigenvalues are in modulus smaller than 1. Furthermore the eigenvector corresponding to eigenvalue 1 has all entries positive. In particular, for the eigenvalue

1 there exists a unique eigenvector with the sum of its entries equal to 1. This gives us a unique eigenvector as our rank for the genes. Similar arguments can be applied for the proof of the existence of our patient rank.

Result

The data sets used for the experiment were taken from the study of Integrated Genomic Analyses of Ovarian Carcinoma led by the Cancer Genome Atlas. The associated results and discussions were published in Nature 2011 [29]. The analysis of 489 clinically annotated stage III-V HGS-OvCa samples and its corresponding normal DNA were reported in the article and posted on its associated website. The data incorporates the age at diagnosis, stage, tumour grade and surgical outcome of patients diagnosed with HGS-OvCa. We downloaded the TCGA-OV-mutations data and the unified expression profiles from the TCGA Data Portal website for our purpose. In the aftermath of data cleaning, we retain mutations containing insertion, deletion and alternation of base only. Finally a patient mutation profile table which comprises 316 patients and 8404 genes is obtained. Similar procedures were carried out on obtaining the gene expressions data from the website. Pearsons correlation coefficients are calculated on the gene expression data in pairwise fashion to obtain the gene correlation value between each pair of genes and the gene corelation graph is constructed. We utilized two different protein protein interaction networks for our experments. HumanNet is a probabilistic functional gene network which consists of 18,714 protein encoding genes and 476399 interactions between the genes of Homo sapiens. Pathway Commons is a collection of publicly available metabolic pathway database in conjunction with interactions from multiple organisms. It was filtered to retain human genes and interactions for the sake of our experiments. We obtained the required data through its web portal for download and query.

Ground Truth Data

We compiled a set of genes published in various literature on several cancer studies which are certified to be ovarian cancer genes to be our ground truth cancer genes in the evaluation of our proposed models. Afterwards, the experiments on our five proposed models are run. A gene scoring vector (gene rank μ) for each of the six models is obtained. We then evaluate our proposed models by the rankings of the ground truth genes in each of the six proposed models' gene scoring vector μ and demonstrate the effects of integrating more background information in ranking. Precision/Recall graph and the top 25 genes appeared in each of the gene scoring vector (gene rank μ) of the five proposed models are presented in subsequent sections. Table 1 below tabulates the collection of ovarian cancer genes (ground truth genes) and the associated references.

Table 1 Ground Truth Genes.

GENE	Literatures
BRCA1	[12,13,29]
BRCA2	[14]
BMPR1A	[17,20]
BRIP1	[25]
MLH1	[15]
FHIT	[14,35]
TFRC	[16]
FGFR2	[18,19]
GATA3	[21]
MYST4	[34]
PTEN	[22]
FAS	[23,24]
RB1	[25]
SEPT9	[26]
YWHAE	[33]
TP53	[29]
PIK3CA	[27]
BRAF	[27]
KRAS	[27]
AIB1	[28]
MSH2	[15]
BMP4	[31,32]
TRIP1	[30]
MYC	[30]
EP300	[30]

Experimental Results

We run the experiments on our six proposed models using the data set we obtained. In our experiments, we set α = 0.75. For the additive model (RW-AM), we set α = 0.3, β = 0.3 andγ = 0.4. Three benchmark models are utilized to evaluate our proposed models. The first one is frequency based in which each gene is awarded a rank in accordance with the occurence of mutation which means the higher the frequency of occurence of mutations on that gene, the higher rank will be awarded. The other two benchmark models are random walk based in which we perform random walk on gene correlation network (RW-GC) and patient mutation (RW-PM) network respectively and a gene scoring rank vector μ for each network is attained. We present the total number of appearances of ground truth genes in the top 1 percent of the gene rank μ of each model as follows in Table 2:

Table 2 Top 1 percent of the gene ranks.

Models	Numbers Of Appearance	Average Rank
RW-MMGIS	14	40
RW-MMGCS	17	38
RW-MMPFS	14	44
RW-AM	15	39
RW-MMP	16	40
RW-MMA	16	42
RW-GC	9	62
RW-PG	4	17
FREQUENCY BASE	4	18

The six proposed models outperform all the benchmark models. This can be demonstrated from the above table that the number of occurences of ground truth genes in the above six models outnumbers the three benchmark models. We found that incorporating heterogeneous sources of biological information enhances the performance of identifying ovarian cancer genes. In the nine models, RW-MMGCS yields the best performance, followed by RW-MMA and then RW-MMP and then RW-AM and then RW-MMGIS and then RW-MMPFS and then followed by three benchmark models at last: RW-GC, RW-PG and FREQUENCY BASE. Please note that a larger gap occurs between the results of two benchmark models with RW-GC outperforming RW-PG. This underscores that the gene expression data is more informative than patient mutation profile in locating ovarian cancer genes. All in all, integrating various heterogeneous sources of information helps in locating ovarian cancer genes.

Evaluation

In this subsection, Precision/Recall graph by adjusting the threshold on the rank of the ground truth genes is presented. Precision is defined as the fraction of the ground truth genes among all genes ranked above each threshold. Recall is defined as the fraction of ground truth genes which are ranked above each threshold among all known ground truth genes. 25 ground truth genes in each experiment are used and the results are tabulated in Figure 4.

All six proposed models outperform the three benchmark models. RWMMGCS yields the best performance, followed by RW-MMA and then RW-MMP and then RW-MMGIS and then RW-AM. RW-MMGCS, RW-MMA and RW-MMP show a very high precision rate at recall rates running from 0.1 to 0.2. This demonstrates that they are able to locate several true positive genes (ground truth genes) in topmost positions within the ranked list. Since we use only 25 ground truth genes in our experiments, we expect to achieve a better result if more candidate cancer genes are included. Almost all the models decrease their performances monotonically towards the higher recall rate except FREQUENCY BASE and RW-PG in which their precision increases a little towards a little higher recall rate and then plummets sharply. This can be explained by the fact that these two models discover a multitude of false positive at low recall rate while they obtain a little better precision towards higher recall rate when they are able to rank a few ground truth genes below the top ranked genes. Above all, we demonstrate that the integration of more hetergeneous background information in the ranking helps achieve a better recall/precision rate.

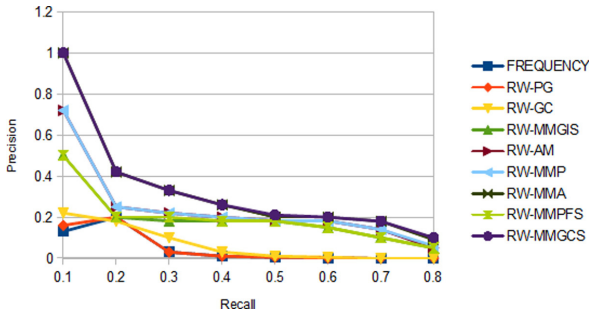

Figure 4 Recall/precision.

There is one parameter α in our proposed models (RWMMGCS, RW-MMA, RW-MMP, RW-MMGIS) which is the probability of teleportation of genes and patients. We performed an experiment on adjusting the value of α from 0 to 1 to inspect its relation to the average rank of the ground truth genes. The result is tabulated in Figure 5.

From above, the best α obtained is around 0.8 which achieves the lowest average ground truth genes ranking. Subsequently, in our additive model(RW-AM), we have three parameters α, β, γ that have to be determined. To evaluate these three parameters, we fix one of the

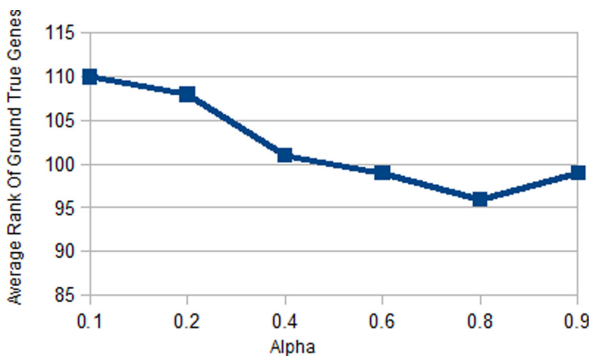

Figure 5 Average rank of ground truth genes by adjusting teleportation parameter alpha.

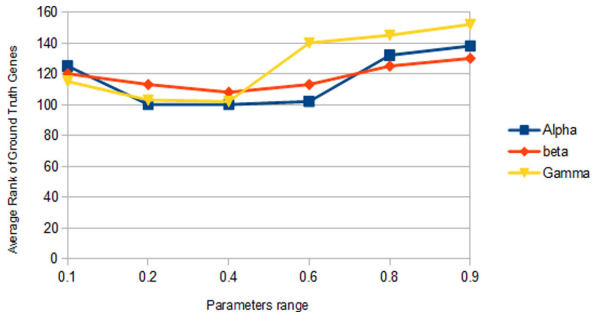

Figure 6 Average rank of ground truth genes achieved by fixing each parameter in RW-AM.

parameters each time and adjust the other two parameters and record the best average ground truth genes ranking and the result is tabulated in Figure 6.

Conclusion

In this paper, a Markov Chain Model for discovering important cancer genes through integration of heterogeneous sources of information are proposed: patient mutation profile, gene gene interaction network and gene correlation network in an unsupervised manner. Experimental results demonstrate that our proposed models outperform all benchmark models. Our future work will focus on developing graph Laplacian in learning cancer genes priority.

Grants

The publication of this work was partly funded by NSF under award numbers EPS-0903787, and EPS-1006883, and MCB-1027989.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Christopher Ma, Yixin Chen, Dawn Wilkins, Xiang Chen, and Jinghui Zhang planned the research. Christopher Ma prepared the computer code and ran the experiments. Christopher Ma, Yixin Chen, and Dawn Wilkins analyzed the results. Christopher Ma wrote the manuscript with assistance from Yixin Chen and Dawn Wilkins. All authors reviewed the manuscript before submission.

This article has been published as part of *BMC Genomics* Volume 16 Supplement 9, 2015: Selected articles from the IEE International Conference on Bioinformatics and Biomedicine (BIBM 2014): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/16/S9.

Authors' details

[1]Department of Computer and Information Science, Weir Hall, University of Mississippi, University, MS 38677, USA. [2]St. Jude Children's Research Hospital, Department of Computational Biology, Memphis, 262 Danny Thomas Place, Memphis, TN 38105-3678, Memphis, USA.

Published: 17 August 2015

References

1. Matteo , Re , Valentini Giorgio: **Random Walking on Functional Interaction Networks to Rank Genes Involved in Cancer.** *Artificial Intelligence Applications and Innovations* Springer Berlin Heidelberg; 2012, 66-75.
2. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, *et al*: **Gene prioritization through genomic data fusion.** *Nature Biotechnology* 2006, **24(5)**:537-544.
3. Erten S, Bebek G, Koyutürk M: **Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks.** *Journal of Computational Biology* 2011, **18(11)**:1561-1574.
4. McDermott J, Bumgarner R, Samudrala R: **Functional annotation from predicted protein interaction networks.** *Bioinformatics* 2005, **21(15)**:3217-3226.
5. Re M, Valentini G: **Cancer module genes ranking using kernelized score functions.** *BMC Bioinformatics* 2012, **13(Suppl 14)**:S3.
6. Mostafavi S, Debajyoti R, Warde-Farley D, Grouios C, Morris Q, *et al*: **GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function.** *Genome Biol* 2008, **9(Suppl 1)**:S4.
7. Deng M, Chen T, Sun F: **An integrated probabilistic model for functional prediction of proteins.** *Journal of Computational Biology* 2004, **11(2-3)**:463-475.
8. Petrochilos D, Shojaie A, Gennari J, Abernethy N: **Using random walks to identify cancer-associated modules in expression data.** *BioData Mining* 2013, **6(1)**:17.
9. Sharan R, Ulitsky I, Shamir R: **Network based prediction of protein function.** *Molecular Systems Biology* 2007, **3**:88.
10. Zhang , Wei , *et al*: **Signed network propagation for detecting differential gene expressions and DNA copy number variations.** *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine* ACM; 2012, 337-344.
11. Mackay HJ, Cameron D, Rahilly M, Mackean MJ, Paul J, Kaye SB, Brown R, *et al*: **Reduced MLH1 expression in breast tumors after primary chemotherapy predicts disease-free survival.** *Journal of Clinical Oncology* 2000, **18(1)**:87-93.
12. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, *et al*: **A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1.** *Science* 1994, **266(5182)**:66-71.
13. Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, *et al*: **Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families.** *The American Journal of Human Genetics* 1998, **62(3)**:676-689.
14. Dhillon VS, Shahid M, Husain SA: **CpG methylation of the FHIT, FANCF, cyclin-D2, BRCA2 and RUNX3 genes in Granulosa cell tumors (GCTs) of ovarian origin.** *Mol Cancer* 2004, **3**:33.
15. Samimi G, Fink D, Varki NM, Husain A, Hoskins WJ, Alberts DS, *et al*: **Analysis of MLH1 and MSH2 expression in ovarian cancer before and after platinum drug-based chemotherapy.** *Clinical Cancer Research* 2000, **6(4)**:1415-1421.
16. Majidzadeh-A K, Esmaeili R, Abdoli N: **TFRC and ACTB as the best reference genes to quantify Urokinase Plasminogen Activator in breast cancer.** *BMC Research Notes* 2011, **4**:215.
17. Shepherd TG, Nachtigal MW: **Identification of a putative autocrine bone morphogenetic protein-signaling pathway in human ovarian surface epithelium and ovarian cancer cells.** *Endocrinology* 2003, **144(8)**:3306-3314.
18. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, *et al*: **A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer.** *Nature Genetics* 2007, **39(7)**:870-874.
19. Meyer KB, Maia AT, O'Reilly M, Teschendorff AE, Chin SF, Caldas C, *et al*: **Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer.** *PLoS Biology* 2008, **6(5)**:e108.
20. Alarmo EL, Kuukasjarvi T, Karhu R, Kallioniemi A, *et al*: **A comprehensive expression survey of bone morphogenetic proteins in breast cancer highlights the importance of BMP4 and BMP7.** *Breast Cancer Research and Treatment* 2007, **103(2)**:239-246.
21. Mehra R, Varambally S, Ding L, Shen R, Sabel MS, Ghosh D, *et al*: **Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis.** *Cancer Research* 2005, **65(24)**:11259-11264.
22. Obata K, Morland SJ, Watson RH, Hitchcock A, Chenevix-Trench G, Thomas EJ, *et al*: **Frequent PTEN/MMAC mutations in endometrioid but**

not serous or mucinous epithelial ovarian tumors. *Cancer Research* 1998, **58(10)**:2095-2097.
23. Meinhold-Heerlein I, Stenner-Liewen F, Liewen H, Kitada S, Krajewska M, Krajewski S, *et al*: **Expression and potential role of Fas-associated phosphatase-1 in ovarian cancer.** *Am J Pathol* 2001, **158(4)**:1335-1344.
24. Baldwin RL, Tran H, Karlan BY: **Primary ovarian cancer cultures are resistant to Fas-mediated apoptosis.** *Gynecologic Oncology* 1999, **74(2)**:265-271.
25. Song H, Ramus SJ, Kjaer SK, Hogdall E, Dicioccio RA, Whittemore AS, *et al*: **Tagging single nucleotide polymorphisms in the BRIP1 gene and susceptibility to breast and ovarian cancer.** *PLoS One* 2007, **2(3)**:e268.
26. Scott M, McCluggage WG, Hillan KJ, Hall PA, Russell SE, *et al*: **Altered patterns of transcription of the septin gene, SEPT9, in ovarian tumorigenesis.** *International Journal of Cancer* 2006, **118(5)**:1325-1329.
27. Fadare Oluwole, Khabele D: **Molecular Profiling of Epithelial Ovarian Cancer.** *My Cancer Genome* [http://www.mycancergenome.org/content/disease/ovarian-cancer/].
28. Anzick SL, Kononen J, Walker RL, Azorsa DO, Tanner MM, Guan XY, *et al*: **AIB1, a steroid receptor coactivator amplified in breast and ovarian cancer.** *Science* 1997, **277(5328)**:965-968.
29. Cancer Genome Atlas Research Network, Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, *et al*: **Integrated genomic characterization of endometrial carcinoma.** *Nature* 2013, **497(7447)**:67-73.
30. Hwang TH, Atluri G, Kuang R, Kumar V, Starr T, Silverstein KA, *et al*: **Large-scale integrative network-based analysis identifies common pathways disrupted by copy number alterations across cancers.** *BMC Genomics* 2013, **14**:440.
31. McLean K, Buckanovich RJ: **BMPs morph into new roles in ovarian cancer.** *Cell Cycle* 2013, **12(3)**:389-389.
32. McLean K, Gong Y, Choi Y, Deng N, Yang K, Bai S, *et al*: **Human ovarian carcinoma-associated mesenchymal stem cells regulate cancer stem cells and tumorigenesis via altered BMP production.** *J Clin Invest* 2011, **121(8)**:3206-3219.
33. Gagne JP, Gagne P, Hunter JM, Bonicalzi ME, Lemay JF, Kelly I, *et al*: **Proteome profiling of human epithelial ovarian cancer cell line TOV-112D.** *Molecular and Cellular Biochemistry* 2005, **275(1-2)**:25-55.
34. Vignati S, Albertini V, Rinaldi A, Kwee I, Riva C, Oldrini R, *et al*: **Cellular, Molecular Consequences of Peroxisome Proliferator-Activated Receptoralpha Activation in Ovarian Cancer Cells.** *Neoplasia* 2006, **8(10)**:851-IN12.
35. Z¨ochbauer-Mu¨ller , Sabine , *et al*: **5′ CpG island methylation of the FHIT gene is correlated with loss of gene expression in lung and breast cancer.** *Cancer research* 2001, **61(9)**:3581-3585.

doi:10.1186/1471-2164-16-S9-S3
Cite this article as: Ma *et al.*: An unsupervised learning approach to find ovarian cancer genes through integration of biological data. *BMC Genomics* 2015 **16**(Suppl 9):S3.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central