Open Access Bioinformatics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Using biomedical networks to prioritize gene–disease associations

This article was published in the following Dove Press journal:
Open Access Bioinformatics
24 August 2011

Joel P Arrais
José Luís Oliveira

Department of Electronics, Telecommunications and Informatics (DETI), Institute of Electronics and Telematics Engineering of Aveiro (IEETA), University of Aveiro, Portugal

Abstract: Understanding the genetic foundations of genetic diseases, such as cancer, Alzheimer disease, or Huntington's disease, is critical to the development of new diagnostics and treatments. Several computational methods have been used to speed up the discovery process, eg, by selecting the molecular targets for a given disease. However, despite the achievements obtained over recent years, better solutions are still required. This paper presents an innovative computational method that addresses the problem of using disperse biomedical knowledge to select the best candidate genes associated with a disease. The method uses a network representation of current biomedical knowledge that includes biomolecular concepts such as genes, diseases, pathways, and biological process. It also applies information extraction techniques to enrich the network with more dynamic and updated data. A biologically inspired algorithm is applied to this network in order to identify association levels between genes and diseases. The solution proposed here surpasses many limitations of previous methods such as the need for training data. The validation applied demonstrates that the proposed method has best overall results compared with state-of-the-art methods as it also performs especially well for the critical top-rank positions. We believe this method represents a major advance over previous work and that it will be a key tool for future gene–disease association studies.

Keywords: gene–disease, biomedical networks, prioritization, computational method

Introduction

The identification of the genes involved in human diseases is a first step to understand the molecular basis of a disease, its underlying mechanisms, diagnosis, and therapies. The current main issue is that the existing biomolecular methods, such as positional cloning or microarrays, lack precision as they return tens to hundreds of candidate genes involved in a particular disease. This problem has been tackled with the development of computational methods that help to identify the most relevant genes with the condition under study.[1,2] Benefiting from the enormous quantity of biomedical data available in public databases, these methods have been used to shorten the path from molecular evidence to therapy development.[3,4] The computational methods available are based on the biological principle that functionally related genes originate similar phenotypes. For instance, in the study of type 2 diabetes, the gene *KCNJ5* appears to be a good candidate because of its involvement in "potassium inwardly-rectifying channel", an important signaling pathway in diabetes, and because of its known interaction with the gene *ADRB2*, whose association with diabetes has already been documented.[1]

Several authors have already explored different approaches to this problem. The most common consist of using biomedical literature as the main source of information.

Correspondence: Joel Perdiz Arrais
University of Aveiro, 3810-193
Aveiro, Portugal
Tel +35 1234370500
Fax +35 1234370545
Email jpa@ua.pt

http://dx.doi.org/10.2147/OAB.S21325

In this case the association of genes is obtained by using gene-related concepts to construct the question over biomedical literature databases. This strategy was explored by Hristovski et al[5] who have used concepts from MeSH (Medical Subject Headings), by Yetisgen-Yildiz and Pratt[6] through the UMLS (Unified Medical Language System), by Perez-Iratxeta et al[7] combining data from Gene Ontology and MESH, and more recently by Frijters et al[8] enabling the discovery of hidden relationships between genes, diseases, and cellular processes.

As an alternative to the biomedical literature, several strategies use data from several biomedical sources. The most common are protein interaction, functional annotation of signaling pathways, expression data, sequence homologies, regulatory data, and disease. Examples include the strategies implemented by Adie et al[9] and by Masotti et al[10] which combine functional annotation data with expression patterns. Radivojac et al[11] combine data from functional annotation with genomic sequences and diseases. George et al[12] use data from signaling pathways, genomic sequences and protein interactions.

Despite the helpful results obtained with these methods, they present major limitations as they require training data (except for Tremblay et al[13] and Adie et al),[9] have constraints through using a high number of data sources, and restrictions in representing the inherent complexity of biomedical terms. The use of network-based methods consists of an interesting alternative to address the issue of studying gene–disease associations. Representing biomolecular concepts as nodes, such as genes or pathways, and their associations or interactions as edges, such as gene–disease or protein–protein interactions, is a simple yet powerful abstraction. One major advantage of using networks is the possibility to use the methods and tools available for graph or network theory. The typical procedure for using networks to establish gene–disease predictions is presented in Figure 1. First, a list of candidate genes is obtained. Second, those genes are mapped onto the previously obtained network that represents all biomedical knowledge. Depending on the method, additional information can be included. Finally, one of many methods can be applied to obtain the score of each gene–disease relationship. This last stage, scoring, is where most research work is currently focused. Although a comprehensive review can be found in Wu and Li,[14] most of the methods proposed over recent years fall into one of the following categories: proximity, similarity, or centrality. The proximity method consists of considering that the genes lying closer to genes whose association with the disease is already known have higher probabilities of also being involved in the disease. The Endeavour[15] and the Prioritizer[16] are two tools that implement the proximity approach. The similarity approach, instead of considering genes by their proximity to the relevant disease, considers their similarity to the disease. This approach has been explored by Lage et al[17] and by Miozzi et al.[18] The last approach, centrality, consists of selecting the genes based on how central they are in the network and therefore how informative they are for the gene–disease association. Gudivada et al[19] and Ozgur et al[20] were the first to present methods based on this approach.

This paper presents a solid contribution to the problem of using dispersed biomedical knowledge for selecting the best candidate genes associated with a disease. The proposed computational method uses network-based representation with an innovative biomedically inspired metric. It consists of a significant advance, surpassing many limitations of previous methods. The exhaustive validation scheme showed that the proposed method has best overall results when compared with state of the art methods as it also performs specially well for the critical top rank positions. We believe this method represents a major advance over previous work and that it will be a key tool for future gene–disease association studies.

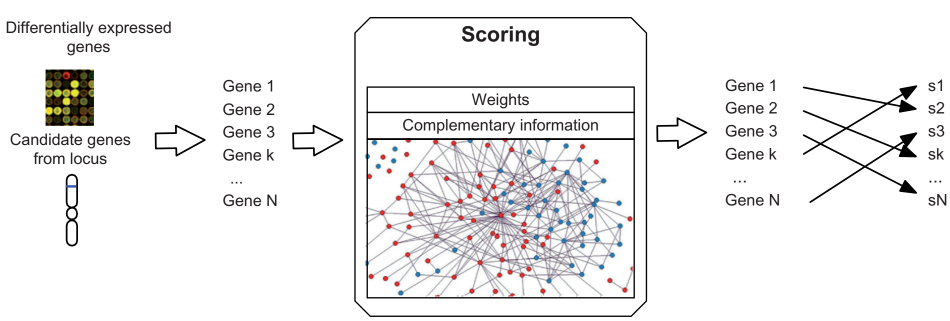

Figure 1 Schema of the network-based candidate gene–disease prioritization. Left to right: (1) from gene expression studies or positional studies, a list of genes is obtained for prioritization; (2) the genes are mapped against the network; (3) using the relations stored in the network the list of genes is ranked.

Methods

There are four main stages that underpin the development of the framework required to apply the proposed method. The first is integrated access of biomedical data including its validation, cleansing, and format adjustment. Secondly, the previously obtained data are used to assemble a network representation with current biomolecular knowledge. Additionally, the network is enriched with new associations obtained with information extraction tools. The last step is implementation of the algorithm that allows calculation of the gene–disease prioritization.

Data compilation and integration

The use of computational methods to help discover new gene–disease associations was only made possible with the access to a wealth of biomolecular data that are currently publicly available. There are already many frameworks specialized in the integration of biomedical data that differ from the implemented approach, in the number of resources included, or the number of organisms targeted. Examples include the BioWarehouse,[8] the BioCore,[21] and the GeNS.[22,23]

In the work presented here, access to biological data is supported by GeNS biomolecular data warehouse. It contains data for roughly 1000 species, representing over 7 million gene products with 70 million alternative gene/protein identifiers and 140 million associations with biological entities. For *Homo sapiens* it has 85,000 gene products that can be mapped to 704,000 synonyms, which also show 571,000 associations with biological entities, such as pathways, Gene Ontology terms, or homologs. Detailed information on the schema and the integrated databases is available in Arrais.[22] From the information available on GeNS, the following datasources were selected to construct the network:

– Genes: information on the synonyms of each gene. This is the result of merging the data available in three distinct databases, namely Entrez Gene, KEGG and UniProt;
– Diseases: list of diseases obtained from the OMIM database[24] and from the KEGG disease database;[25]
– Gene Ontology (GO): information on gene annotations to biological processes, molecular functions, and cellular component;[26]
– Pathways: obtained from the KEGG Pathway database;[27]
– Homologs: set of structural and functional components that can be used to classify genes. In this case, we have used the set of orthologs provided by the KEGG database;
– Literature: associations between genes and PubMed papers provided by the Entrez database.

Biomedical network modeling

The primary goal of the network is to represent the explicit and well-established relations among the biomedical terms from the previously presented databases. Figure 2 contains a representation of selected terms as the relations to be included in the graph.

Because the data stored in GeNS have already been cleaned and fused, the extracted data can be represented as the vector of terms $\vec{a}$.

$$\vec{a} = (a_1, a_2, \ldots, a_p), \tag{1}$$

with a_k representing the content of the *kth* term from the interval ($k \in [1, P] \subset \mathbb{N}$). Each term a_k is a tuple of four elements that can be represented as:

$$a_k : \{t_a, d_i, t_b, d_j\}, \tag{2}$$

where the element t_a from the type d_i has an association with the element t_b from the type d_j, where

$$(i, j \in [1, Q] \subset \mathbb{N}) \text{ and } (a, b \in [1, R] \subset \mathbb{N}).$$

The vector $\vec{a}$ can be modeled as a non-oriented weighted graph (BioMedical Graph) BMGi = (V_i, E_i) where:

– Each vertex $v_x \in V_i$ is obtained by identifying the unique entry t_a-d_i or t_b-d_j of all the association tuples contained in vector $\vec{a}$. The vertexes are labeled by their name and type;
– Each edge $e_x \in E_i$ connects vertexes (v_m, v_n) representing an association between the terms represented by the vertexes v_m and v_n contained in vector $\vec{a}$;
– The weight W_{v_m,v_n} of each edge e_x corresponds to its level of confidence. For the relations directly extracted from biomedical databases we assume have a confidence level of 1 and for the text pattern matching the weight correspond to the confidence on the association.

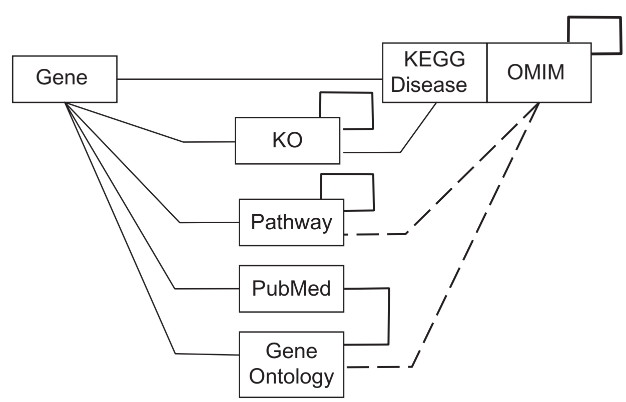

Figure 2 Biomolecular terms used and their relations. The dashed line represents terms obtained by pattern matching.
Abbreviations: KO, KEGG Orthology; OMIM, Online Mendelian Inheritance in Man.

Extend network terms relations through text pattern matching

The use of ontologies has as its main advantage the possibility to establish an association between two entities based on the premise that they share the same process or are somehow related. While specificity is good when the intention is to classify a biological entity, when the goal is to search for similar but not exactly equal entities, it may bring many challenges caused by narrow scope. For instance, for the *BRCA2* gene that is annotated in the KEGG Orthology as being a "Breast cancer 2 susceptibility protein", it is implicitly also part of a larger class named "Pathways in cancer" where many breast cancer relevant genes are also annotated.

To take full potential of the original data, the structure of each used terminology has been replicated into the network allowing this mapping to be included. This allows us to associate two terms not just by their direct relation but also by their possible indirect relations.

The previously described network contains updated information about current knowledge of gene–disease associations. These data are accurate, because they have been reported by manual/semi-manual validation. However it is static and also limited to the typical task in hand, since the ultimate goal is discovery of unknown gene–disease associations. One way to overcome this problem is by adding new relations to the network, which represent less accurate but still highly probable associations. This was done by comparing the textual description of two terms and including the matches in the network, for instance, by comparing the description of the Biological Process from the Gene Ontology with the synopses of the disease provided by OMIM. If a high level of correspondence is found, a new association can be added to the network.

For this purpose we have developed a customized interface that uses a non-deterministic approach to evaluate the level of association between two given terms. Generically, for a term t_a from the data source d_i we look for a correspondence in the description of each term t_b from database d_j. Using the TF-IDF[28] statistical measure we evaluate the number of occurrences of the term t_a in the description of t_b, the total number of times that t_a appears in all terms from d_j, and also the total number of terms from d_i and d_j. From the retrieved results, we have defined a cut-off of the top 10. For instance the Biological Process term "GO:0006629 – Lipid Metabolism" has a positive match with the OMIM disease term "275630 – Chanarin-Dorfman Syndrome". This match can be explained due to a mutation in the disease-related gene *PNPLA2* that is annotated as being involved in several lipid processes.

We have applied this approach to map KEGG metabolic pathways, Biological Process from the Gene Ontology, and genes to disease in OMIM. The results are presented in Table 1.

Algorithm description

The final goal of the algorithm is to produce a score that reflects the relevance of the association between two given biomedical terms. Usually this problem is a two-step procedure where the first is to evaluate the centrality of each vertex in the network and the second is to obtain the association level for each pair of terms.

This is done by exploring all possible paths between two terms in the graph. The final value obtained should correspond to the biological distance between the two given terms. By "biological distance" we mean quantification in the biomedical context. We believe that the importance of a node should not just be based on its connectivity (eg, short distance) but rather on its intrinsic biological relevance and how much information it can add to the path.

In this way, vertexes with relations to several distinct vertex types are promoted because they correspond to well-annotated entities. On the other hand, vertexes that point to several vertexes from the same type are de-promoted because they have lower specificity as they link to several different types of data. If we take, for instance, a pathway associated with 20 genes and one disease, the singleton association (disease) is more informative than each pathway–gene relationship.

One other measure considered is the biological context of each datatype. For instance, a gene associated with four pathways is more relevant than a pathway related to four genes. One way to address this is by giving distinct weights to each of the associations.

Finally, the length of the path is also accounted for in the algorithm. A higher number of intermediate vertexes between two concepts will decrease the association value.

Table 1 Total terms mapped from the Gene Ontology Biological Process and from KEGG metabolic pathways to OMIM

Association	Source terms	Target OMIM terms	% of mapped OMIM terms
GO:BP → OMIM	11,699	6,396	77%
Pathway → OMIM	114	186	1.4%

Notes: "Source terms" correspond to the total number of terms in source database with at least one match in the OMIM, "Target OMIM terms" correspond to the total number of OMIM terms with at least one match, and "% of mapped OMIM terms" corresponds to the percentage of total terms from OMIM mapped.

The final score is obtained based on the following premises:

– Importance of each vertex type;
– Importance of each pair source/destination edge;
– Distance between two vertexes – the algorithm should reflect that a long path has less probability of being relevant than a shorter one. We have also defined a maximum depth for the path search;
– Inverse frequency – if several vertexes of the same type point to the same node, the path relevance should decrease.

Based on the previous assumptions, the implemented algorithm is as follows:

For each path i between the vertexes v_m and v_n:

$$Score(k, e)_i = \sum_{e \in E_j} \frac{k_e \times Adj(e, k) \times W_t(e) \times decay}{Adj(e)}$$

where k_e corresponds to the number of distinct destination types that can be reached from the vertex e; $Adj(e)$ returns the total number of vertexes that can be reached from vertex e; $Adj(e, k)$ gives the number of vertex of type k that can be reached from e. Decay is a constant greater than 1 that reflects the penalization given to long paths. In this test we used decay = 1.2.

The final score is given by the sum of all scores from all path i between the vertexes v_m and v_n. A final step consists in normalizing the score values in order to enable direct comparisons over searches done over the same BMG_i. To accomplish this it is assumed that the score 1 represents two isolated vertexes that are directly associated.

Results

Systematic comparative analysis

To explore the feasibility of the proposed approach we have conducted the following comparative analysis. We started by randomly selecting 600 OMIM diseases with at least two associated genotype. For each disease we have created a list of 20 elements that include the gene known to be associated with the disease and 19 randomly selected genes. We obtain 600 lists with 20 genes each where only one gene per list is known to be associated with the disease. Next we delete all network edges that represent an explicit association between genes and diseases.

Combining the gene–disease pair, a total of 1200 prioritizations were performed. From these we calculate the sensitivity and specificity values. Sensitivity refers to the frequency of genes that are ranked above a particular threshold position and specificity to the percentage of genes ranked below this threshold. For instance a sensitivity/specificity value of 60/95 indicates that the correct disease gene is marked in the top 5% in 60% of all prioritizations.

Finally we are able to compare the power of each method to rank the list of genes and therefore reconstruct the disease. To assess the improvement of our method we also include in the comparison the method presented by Gudivada et al[19] and the shortest path that can be used as a baseline. The receiver operating characteristic (ROC) curve in Figure 3 compiles the results obtained.

The ROC curve shows that our method reached an AUC (area under curve) of 0.919 compared with 0.905 for the Gudivada method and 0.85 for the shortest path. While the improvements obtained (0.014 and 0.069) with the proposed method are not very expressive, it is particularly interesting to observe that fixing the specificity at 90% yields a sensitivity of 80% for the proposed method, 65% for the Gudivada method, and 35% for the shortest path. This means that 80% of the time, the proposed method ranks the disease in the top 10%. By fixing the specificity values at 95%, we observe sensitivity values of 43% for the proposed method, 34% for the Gudivada method, and 11% for the shortest path. These values show that the proposed method not only gives better overall results but has an improved performance in selecting the candidate genes for the restricted specificity values. This also means that the valid gene is, on average, ranked more often in the top positions. Indeed, for specificity values <65%, all three methods have similar sensitivity and for specificity values <85%, the proposed method and the Gudivada method have similar sensitivity. Finally, it is interesting to note that for specificity values ranging from 100% to 90%, corresponding to cases where the correct gene is marked in the two top positions, the proposed method presents a distinct advantage.

Contribution of individual components

Next we analyze the contribution of each component of the network to the final result. We iteratively remove each component and run the previous test for the 100 OMIM diseases. The three ROC curves in Figure 4 compile the results obtained for each case. First we set all weights to 1, meaning that all relations are considered to be of the same relevance (Figure 4A). Next we exclude from the network all relations obtained with the help of text matching (Figure 4B). Third we evaluate the influence of including the structure of the ontologies in the network (Figure 4C). The results obtained for the AUC were 0.90, 0.85, and 0.89.

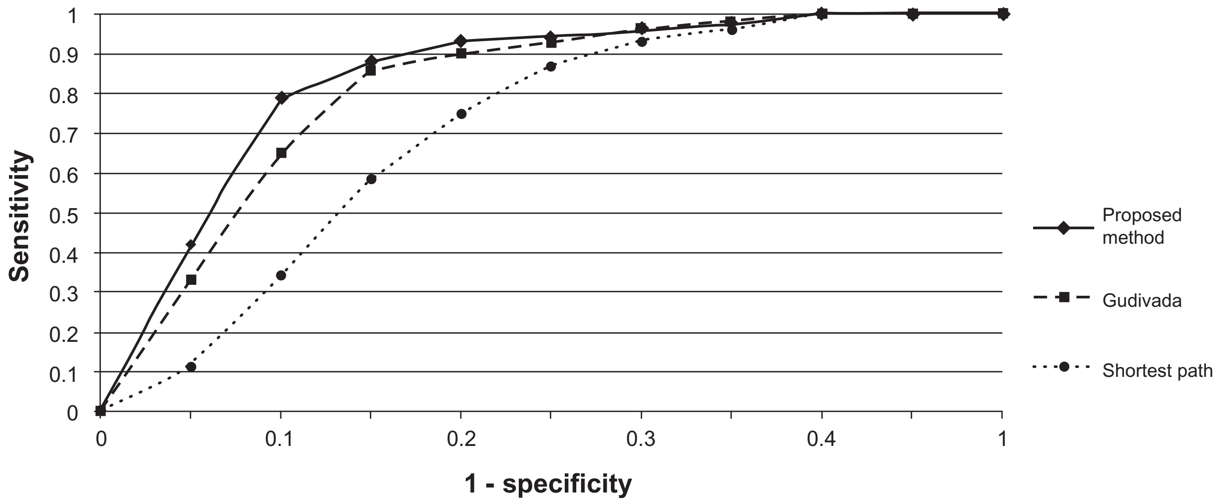

Figure 3 ROC curve for validating the proposed method against Gudivada et al[19] and shortest path.
Abbreviation: ROC, receiver operating characteristics.

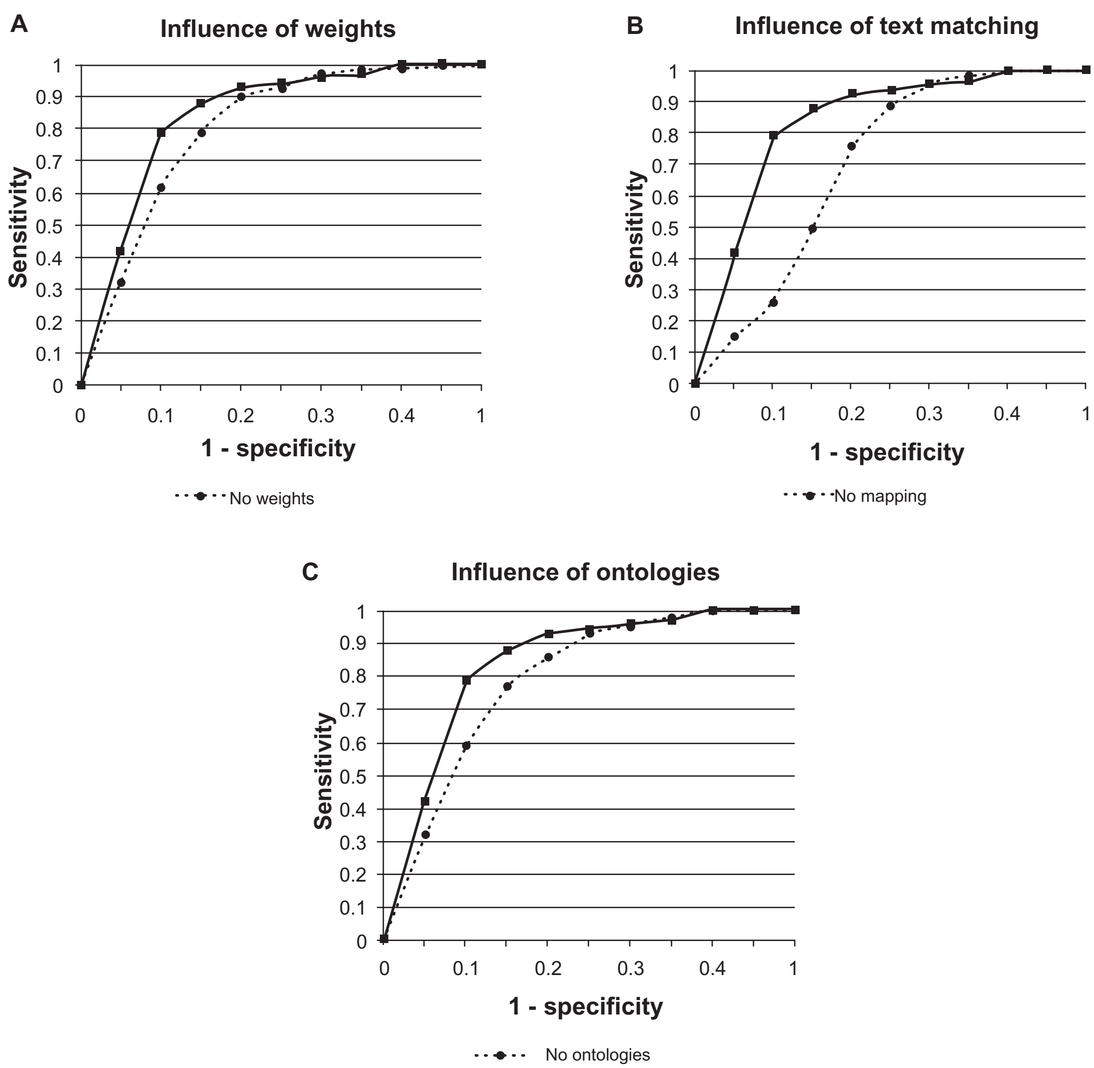

Figure 4 ROC curve for evaluating the effect of removing (**A**) weights, (**B**) information matching, and (**C**) ontologies. The black line represents the ROC curve from Figure 4 and the dotted line the ROC curve after removing each component.
Abbreviation: ROC, receiver operating characteristics.

Demonstration for microarray profiling

We have used our algorithm to prioritize the genes that were marked as differently expressed in the experiment "Transcription profiling of 47 human breast tumor cases" stored in ArrayExpress with the Accession number E-GEOD-3744. Table 2 contains the top ten ranked genes according to our algorithm. As expected, the gene at the top is *BRCA1* since several mutations lead to an increased risk of cancer. This gene was classified on top because it is directly associated with the disease. The *ADRA1A*, *HTR4*, and *OXTR* genes were ranked in top positions mainly due to their participation in the calcium signaling pathway whose association with cancer has already been documented.[29] The *FAS*, *FIGS*, and *STK4* genes are associated with pathways in cancer. Finally the *KCTD2*, *COL17A1*, and *MYB* genes are associated with two common biological processes in cancer.

Discussion

Compared with previous methods, the one proposed here has many advantages that to our knowledge make it unique. First it does not require any training data. Excluding G2D,[13] PROSPECTR,[30] and the method proposed by Gudivada et al[19] most of the available methods require training gene sets. The problem with this is that the data required for training are typically scarce and when available are frequently difficult to adapt to the specificities of the problem in hand.

One other limitation of previous approaches is the capacity to cope with an increase in the number of data sources and resources. This is evident in the method proposed by Adie et al[9] and by Masotti et al[10] which only use two data sources. Aerts et al points out the importance of adding more data sources but only include a maximum of four.[15] In contrast, the proposed methodology already uses six distinct data sources, allowing easy expansion to more. Additionally, network abstraction also facilitates the delimitation of complex relations among terms and also the use of weights to differentiate distinct levels of trust or accuracy between terms.

Table 2 Ranked genes from differentially expressed genes in human breast cancer

Rank	Gene symbol	Score
1	*BRCA1*	1.16
2	*OXTR*	2.24
3	*FAS*	2.36
4	*ADRA1A*	2.47
5	*STK4*	2.78
6	*HTR4*	3.18
7	*FIGS*	4.15
8	*KCTD2*	4.56
9	*COL17 A1*	6.22
10	*MYB*	6.43

In comparison with previous approaches that also use the network representation, a major advantage of the one proposed is that it reflects some of the intrinsic characteristics of the biomedical data. Examples of this include the association from one gene to multiple pathways and the reverse association. Unlike the method proposed by Gudivada et al we moved the focus of association measurement from the node to the edge. This tends to reflect more accurately the biological context.

The proposed method also benefits from the possibility to use the ontologies and also to incorporate the intrinsic structure of the ontology in the network. One last advantage of using a network-based approach is its flexibility to adapt to new contexts. In this paper we have explored the prioritization of genes for a given disease. The same framework can however be redirected to answer other research questions. Examples include selection of candidate genes for a given disease, analysis of the interference of pathways in a disease, or the effects of silencing/activating genes on a disease. Despite the major step forward of the work presented, we are aware of the overall limitations of using computational methods to identify gene–disease associations as well those of the approach followed. It is limited by the amount and quality of the data available online. Indeed, the lack of relations may ultimately lead to true negatives and inaccurate relations may lead to false positives. This means results should always be interpreted with caution, with experimental validation being required. We are also aware that the more accurate the available data become, the more information can be taken and previous experiments can even be re-evaluated.

We also faced challenges in directly comparing the results obtained with those previously published. In this context we selected two methods for direct comparison. The first, proposed by Gudivada et al[19] seemed to be closest to the current state of the art. The other was selected because it offers a good baseline. The obtained results reinforce the claimed improvements of the proposed method.

Conclusion

In this paper we have presented a new method to rank genes according to their level of association with a given disease. The proposed method works over a network that integrates biomedical data including several distinctive features such as the capacity to cope with an increase in the number of data sources and resources, and incorporate the structure of ontologies and associations based on text mapping between terms.

Also, unlike previous approaches, ours does not require any training dataset and offers an effective way to incrementally add new data without the need to reprocess the entire network. We also want to stress that the proposed method benefits from the intrinsic characteristics of the biomedical data.

The tests conducted clearly support our initial assumptions about the advantages of the proposed method. Indeed, this method represents a major advance over previous work and we believe it is of key importance for future gene–disease studies.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2012) under grant agreement nº 200754 – the GEN2PHEN project.

Disclosure

The authors declare no conflicts of interest in this work.

References

1. Tranchevent L-C, Capdevila FB, Nitsch D, et al. A guide to web tools to prioritize candidate genes. *Brief Bioinform*. 2010;12(1):22–32.
2. Zhu M, Zhao S. Candidate gene identification approach: progress and challenges. *Int J Biol Sci*. 2007;3(7):420–427.
3. Tiffin N, Adie E, Turner F, et al. Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res*. 2006;34(10):3067–3081.
4. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet*. 2002;3(5):391–397.
5. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Using literature-based discovery to identify disease candidate genes. *Int J Med Inform*. 2005;74(2–4):289–298.
6. Yetisgen-Yildiz M, Pratt W. Using statistical and knowledge-based approaches for literature- based discovery. *J Biomed Inform*. 2006; 39(6):600–611.
7. Perez-Iratxeta C, Wjst M, Bork P, Andrade MA. G2D: a tool for mining genes associated with disease. *BMC Genet*. 2005;6:45.
8. Frijters R, van Vugt M, Smeets R, van Schaik R, de Vlieg J, Alkema W. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput Biol*. 2010;6(9):pii. e1000943.
9. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*. 2006;22(6):773–774.
10. Masotti D, Nardini C, Rossi S, et al. TOM: enhancement and extension of a tool suite for in silico approaches to multigenic hereditary disorders. *Bioinformatics*. 2008;24(3):428–429.
11. Radivojac P, Peng K, Clark WT, et al. An integrated approach to inferring gene–disease associations in humans. *Proteins*. 2008;72(3):1030–1037.
12. George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, Wouters MA. Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res*. 2006;34(19):e130.
13. Tremblay K, Lemire M, Potvin C, et al. Genes to diseases (G2D) computational method to identify asthma candidate genes. *PLoS One*. 2008;3(8):e2907.
14. Wu X, Li S. Cancer gene prediction using a network approach. In: *Cancer Systems Biology*, Wang E, editor. Boca Raton, FL: Chapman and Hall/CRC Mathematical and Computational Biology; 2010.
15. Aerts S, Lambrechts D, Maity S, et al. Gene prioritization through genomic data fusion. *Nat Biotechnol*. 2006;24(5):537–544.
16. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*. 2006;78(6):1011–1025.
17. Lage K, Karlberg EO, Størling ZM, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*. 2007;25(3):309–316.
18. Miozzi L, Piro RM, Rosa F, et al. Functional annotation and identification of candidate disease genes by computational analysis of normal tissue gene expression data. *PLoS One*. 2008;3(6):e2439.
19. Gudivada RC, Qu XA, Chen J, Jegga AG, Neumann EK, Aronow BJ. Identifying disease-causal genes using Semantic Web-based representation of integrated genomic and phenomic knowledge. *J Biomed Inform*. 2008;41(5):717–729.
20. Ozgur A, Vu T, Erkan G, Radev DR. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*. 2008;24(13):i277–i285.
21. Kim HS, Kang SH, Park CH, et al. Genome-wide molecular characterization of mucinous colorectal adenocarcinoma using cDNA microarray analysis. *Oncol Rep*. 2010;25(3):717–727.
22. Arrais JP, Pereira JE, Fernandes F, Oliveira JL. GeNS: a biological data integration platform, in International Conference on Bioinformatics and Biomedicine. Venice, Italy; 2009.
23. Arrais J, Santos B, Fernandes J, Carreto L, Santos MAS, Oliveira JL. GeneBrowser: an approach for integration and functional classification of genomic data. *J Integr Bioinform*. 2007;4(3):82–91.
24. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 2005;33(Database issue): D514–D517.
25. Kanehisa M, Araki M, Goto S, et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*. 2008;36(Database issue): D480–D484.
26. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25–29.
27. Wixon J, Kell D. The Kyoto encyclopedia of genes and genomes – KEGG. *Yeast*. 2000;17(1):48–55.
28. Aizawa A. An information-theoretic perspective of tf-idf measures. *Inf Process Manag*. 2003;39(1):45–65.
29. Berridge MJ. Calcium signalling and cell proliferation. *Bioessays*. 1995;17(6):491–500.
30. Wang M, Zhang S, Huang QY. Computational biology strategy for identification of complex disease genes. *Yi Chuan*. 2009;31(6):581–586. Chinese.