Research Article

RecRWR: A Recursive Random Walk Method for Improved Identification of Diseases

Joel Perdiz Arrais[1] **and José Luís Oliveira**[2]

[1] *Department of Informatics Engineering (DEI), Centre for Informatics and Systems of the University of Coimbra (CISUC), University of Coimbra, 3030-290 Coimbra, Portugal*

[2] *Department of Electronics, Telecommunications and Informatics (DETI), Institute of Electronics and Telematics Engineering of Aveiro (IEETA), University of Aveiro, 3810-193 Aveiro, Portugal*

Correspondence should be addressed to Joel Perdiz Arrais; jpa@dei.uc.pt

High-throughput methods such as next-generation sequencing or DNA microarrays lack precision, as they return hundreds of genes for a single disease profile. Several computational methods applied to physical interaction of protein networks have been successfully used in identification of the best disease candidates for each expression profile. An open problem for these methods is the ability to combine and take advantage of the wealth of biomedical data publicly available. We propose an enhanced method to improve selection of the best disease targets for a multilayer biomedical network that integrates PPI data annotated with stable knowledge from OMIM diseases and GO biological processes. We present a comprehensive validation that demonstrates the advantage of the proposed approach, Recursive Random Walk with Restarts (RecRWR). The obtained results outline the superiority of the proposed approach, RecRWR, in identifying disease candidates, especially with high levels of biological noise and benefiting from all data available.

1. Introduction

A major research domain in molecular biology is the study of the causal association between genomic variations and clinical phenotypes [1–3]. Classical methods use a manual approach where one or a limited number of genomic targets are individually tested. However, due to the resources needed to systematically perform this procedure and due to the difficulty in controlling all experimental variables, improved strategies were required. The possibility to use computational methods to identify the best disease candidates to be further validated was a major breakthrough [4–8]. A common constraint of most methods is the need for training data, which is scarce and difficult to validate.

A recent research trend consists of exploiting the topological properties of protein-protein interaction (PPI) networks combined with other biological data to envisage the underlying mechanisms of genetic diseases. Barabási et al.

[9] and Joy et al. [10] discuss the role of proteins with high betweenness as mediators of relevant metabolic processes. Ma and Zeng [11] explore the use of the closeness centrality to quickly identify the top central metabolites in large scale networks. Approaches proposed by Erten et al. [12] and Arrais and Oliveira [13] explore the potentialities of the nodes with high degree for the prioritization of disease-associated genes.

While the previous methods focus on evaluation of the weights given to each node, a complementary strategy consists of evaluating the proximity of two given nodes in the network. Some common methods to conduct this task are the shortest path, log-likelihood, propagation matrix, and the RWR (Random Walk with Restarts). Previous studies confirm that the RWR outperforms other methods [14–18]. One common limitation of these studies is that they assume the graph is single concept, meaning that every node is equally treated. However, as we demonstrate in this study,

those methods are poor when the graph integrates nodes from distinct data types.

In this paper, we propose a novel method to improve selection of the best disease targets for multiconcept graphs. Towards this aim we build a multilayer biomedical graph that stores PPI data, annotated with stable knowledge from OMIM diseases and Biological Process from Gene Ontology. The inherent improvements of the proposed method are (a) the use of multilayer networks formed with PPI data and by the terms' associations; (b) the combination of this data to establish new associations among nodes; and (c) the use of degree-based methods to evaluate node weights.

Finally, we present comprehensive validation that demonstrates the superiority of the proposed approach, Recursive Random Walk with Restarts (RecRWR).

2. Methods

The method proposed herein uses a graph representation of biomedical knowledge centred on proteins enriched with biomedical terms. The first step consisted of selecting and curating the required data and using it to construct the graph. For performance issues this network is represented as a matrix of adjacencies. Based on this ground-based biomedical graph we apply a modified version of the Hubs and Authorities (HITS) [13] algorithm adapted to this particular subject, in order to obtain a normalized and more accurate association among relations. Although here we are interested in tuning to protein-disease association, it is important to stress that it can be extended to the study of general association of many-to-many biomedical terms. Finally, we formulate how the proposed method, RecRWR, can be applied to this subject.

2.1. Multiconcept Graph Modelling. A graph-based representation is used to store the relations among the biomedical terms. Since we are integrating three distinct data sources, three interconnected subgraphs are obtained.

(i) PPI data are retrieved from STRING database [19], where the average confidence level is considered. A filter is applied to select only human.

(ii) Disease data are extracted from OMIM morbid map [20] data where the genotype-phenotype associations are preserved. The morbid map is also used to extract the mapping relation for known protein diseases.

(iii) Biological Process from Gene Ontology (GO) [21] Directed Acyclic Graph structure is extracted and replicated. The GO-GO mapping is also retrieved and stored.

For each of the previous data sources a curated set of terms $\vec{a}$ is extracted:

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

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

$$a_k : \{t_a, t_b, w\}, \tag{2}$$

where the element t_a has an association with the element t_b, with a confidence score w where ($a, b \in [1, R] \subset \mathbb{N}$) and ($w \in [0, 1000] \subset \mathbb{N}$).

The set of vectors $\vec{a}$ are modelled as a nonoriented weighted graph $G = (V, E, W)$.

(i) Each vertex $v_x \in V$ is obtained by identifying the unique entry t_a-t_b or t_b-t_a of all the association tuples contained in vector $\vec{a}$. The vertices are labelled by their unique identifier.

(ii) Each edge $e_x \in E$ connects vertexes (v_m, v_n) representing an association between the terms represented by the vertices v_m and v_n contained in vector $\vec{a}$.

(iii) The weight w_{v_m, v_n} of each edge e_x corresponds to the score between two nodes.

The graph $G = (V, E, W)$ is then mapped to an adjacency matrix representation that consists of a $|V| \times |V| = n \times n$ matrix A:

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}, \quad n = |V|, \ a_{i,j} \in [0, 1000] \subset \mathbb{N}. \tag{3}$$

Because the graph is undirected the adjacency matrix is symmetric and therefore $a_{ij} = a_{ji}$.

The compiled graph resulted in 60.000 nodes with an average degree of 5. The memory space required to represent the graph is $\Theta(|E|)$, which is realistically equivalent to a memory space of 6.0 MB, excluding hash tables required for node mapping. The adjacency matrix requires a memory space of $\Theta(|V|^2)$, 7.2 GB.

2.2. RecRWR: Recursive Random Walk with Restarts. Next we formulate the RecRWR algorithm including a detailed pseudocode description of the algorithm (Algorithm 1). The three main components are

(i) Random Walk with Restarts;

(ii) recursive cross subgraph mapping;

(iii) node replacement.

2.3. Random Walk with Restarts. The final probability vector of Random Walker is defined as

$$\vec{p}^{t+1} = (1 - r) W \vec{p}^t + r \vec{p}^0, \tag{4}$$

where W is the column-normalized adjacency matrix A and $\vec{p}^t$ is a vector in which the ith element holds the probability of being at node i at time step t. The vector p^0 holds the probability of the initial states and is constructed such that equal probabilities are assigned to the list of seed nodes where the sum of the probabilities is equal to 1. This is obtained by a given list L of seed nodes, where $L \subset V$.

```
Algorithm Recursive Random Walk with Restarts
Input: Adjancecy matrix A of size N × N;
       Vector of size N with nseed proteins p0;
       Restart probability r
Output: Sorted list of candidate diseases
(1) Let m1, ..., mk be the binary vectors of length N
(2) Let W be a column normalized of the Amatrix of size N × N
(3)
(4) Let p := p0;
(5) Let i := 1, j := 1;
(6)
(7) While ((m1' *|p − pt − 1|) >= rho)
(8)         Process (i, j, p)
(9)
(10) Let x := top(p * diag(m1));
(11) Output x;
(12)
(13)
(14) Process (i, j, p)
(15)      If (i == j)
(16)           While (j <= k)
(17)                 p := Process (i, j, p);
(18)
(19)      If (i != j)
(20)           p := Process (j, j, p)
(21)
(22)      p := p' * Wmij;
(23)      Wmij := diag(mi) * W * diag(mj)
(24)      While ((mi' * (delta(p)) > ro)
(25)           p := (r * p0 + (1 − r)Wij * p);
(26)
(27)      p := p' * Wmij;
(28)      p := SelectTop(p, n);
(29)
(30)      return p;
```

ALGORITHM 1: Pseudocode for the RecRWR method.

2.4. Recursive Cross Subgraph Mapping. We extend the previous formulation to a symmetric matrix composed of $k^2/2$ of submatrixes, where k corresponds to the number of data sources. The submatrix that corresponds to the mapping between the subgraphs i and j is obtained by

$$W_{ij} = \text{diag}\left(\overrightarrow{m_i}\right) * W * \text{diag}\left(\overrightarrow{m_j}\right), \qquad (5)$$

where $\overrightarrow{m_i}$ and $\overrightarrow{m_j}$ are binary vectors with n elements that represent the mask of the source and target subgraphs where $i, j \in [1, k] \subset \mathbb{N}, |\overrightarrow{m_i}| = |\overrightarrow{m_j}| = n$.

The result of each iteration of the Random Walk with Restarts is given by

$$p_i^{t \to \infty} = \prod_{t=0}^{\infty} \left((1 - r) W_{ij} p^t + r p^0\right), \qquad (6)$$

where in fact the algorithm stabilizes when the following condition is met:

$$\left(\overrightarrow{m_1} \left| \vec{p}^t - \vec{p}^{t-1} \right|\right) < \varrho, \qquad (7)$$

where $\overrightarrow{m_1}$ is disease mask vector and $\vec{p}^t$ is weight vector at time t. The product will result in a scalar that corresponds to the sum of the differences between two iterations. The condition is true when this value is less than a given constant ϱ.

2.5. Node Replacement. The recursive iteration of the cross subgraph mapping returns a new term. A node replacement strategy is used to replace the genes to be used. The selection of the node index a_m to be replaced by the node index a_n is given by the minimum value of $a_m = \min_{0 \le m \le |V|} p_i^{t \to \infty}$ and where the candidate node is given by $a_n = \max_{0 \le i \le |V|} W_{ij} * p_j^{t \to \infty}$.

3. Results and Discussion

In this section we explore and evaluate the performance of the proposed method. We present a systematic evaluation using a synthetic datasets based on well-known disease profiles.

TABLE 1: Comparison of the AUC for the analysed methods.

	0%	20%	40%	60%
AUC-RWR	0.9866	0.9453	0.8894	0.8435
Δ (%)		−4.36%	−6.29%	−5.44%
AUC-RWR all data	0.9866	0.9417	0.8838	0.8115
Δ (%)		−4.77%	−6.55%	−8.91%
AUC-RecRWR	0.9856	0.9834	0.9534	0.9072
Δ (%)		−0.22%	−3.15%	−5.09%

We also present how the results of RecRWR can be used to explore the resemblances mechanisms on breast cancer.

3.1. Validation Procedure. For each selected phenotype entry on the OMIM database we created a dataset with the associated genotypes. We have selected 100 phenotype diseases with at least 10 associated genotypes each. Then, we iteratively replace genes from the original dataset by random ones, in 20% increments, and the dataset is progressively converted to a fully random dataset. We use each of these protein datasets as seed nodes on the graph. We end up with a test space of 600 gene sets (6 random step levels plus 100 diseases).

3.2. Information Paradox. Previous use of RWR on molecular biology typically concentrates on PPI networks. One would expect that including additional data would contribute to an improved overall result. Figure 1 presents a comparison of the relative frequency of the ranks for each of the analysed datasets, for two of the tested methods (RWR over only PPI data and RWR over the whole network) and for four levels of randomness. From analysis of this graph it is clear there is no improvement with including external annotations on the original PPI network. Indeed for original dataset, with random effect, there are no perceptible differences between the two methods. This statement is even sharper when we test progressive levels of randomness. For instance, when 20% of the genes on the dataset are random, 55% of the RWR over PPI ranks the disease in the top 3, while with the RWR over all data this frequency drops to 48%. For 60% randomness, 35% of the RWR over PPI ranks the disease in the top 5, while with the RWR over all data the frequency drops to 23%. These results were the primary motivation for the work presented in this paper, as they clearly show that the RWR method is not suitable for dealing with multiple biological data.

3.3. RecRWR Results on Synthetic Datasets. We evaluate the performance of the RecRWR method using the receiver operating characteristic (ROC) curves where each curve contains the results for each level of randomness. A higher AUC (area under curve) corresponds to a better overall performance. Figure 2 and Table 1 compile the obtained results.

With 0% randomness the AUC is approximately the same for the three methods, the proposed one having the lowest minimum value, which can be perceived visually. This means that in the absence of biological noise the protein annotation data does not contribute to improving the final
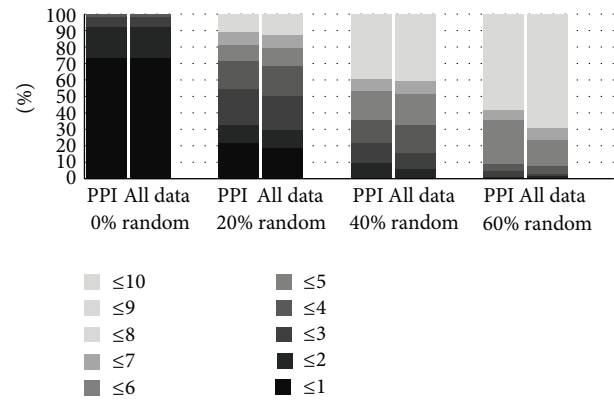

FIGURE 1: Comparison of the RWR method using PPI data and PPI enriched in biological terms.

result. However, if randomness is introduced the proposed method shows a strong improvement.

With 20% randomness the RecRWR AUC is 0.9834, which compared to 0.9453 on the RWR corresponds to a 4.0% of improvement. Comparing the behaviour of the RecRWR the 20% randomness reflects no real impact (−0.22%) on the obtained AUC.

For 40% and 60% the difference is even higher (7.1% for 40% and 7.6% for 60%) demonstrating the greater capability of the proposed method.

It is also relevant to note a 1.0 TPR (true positive rate, y-axis on the graphs from Figure 2), meaning that the disease is always correctly identified and is consistently obtained at the expense of a lower FPR (false positive rate, x-axis).

3.4. RecRWR Results on Breast Cancer. Breast cancer (MIM: 114480) is considered a complex disorder having 23 known genotypes that are shared with other cancer-related disorders. We have used RecRWR over the common expression profiles of breast cancer to explore the network of diseases that share common mechanism. The diseases most closely related to breast cancer are hepatocellular carcinoma, bladder cancer, and lung cancer.

From analysing the network of associations, we can see that the proteins most related with breast cancer are responsible for important cellular functions, such as DNA repairing, cell cycle arrest and its regulation, induced cell death (apoptosis), and tumor suppression. Also, we can see that the more closely GO terms associated with breast cancer are protein binding and apoptotic process. This means
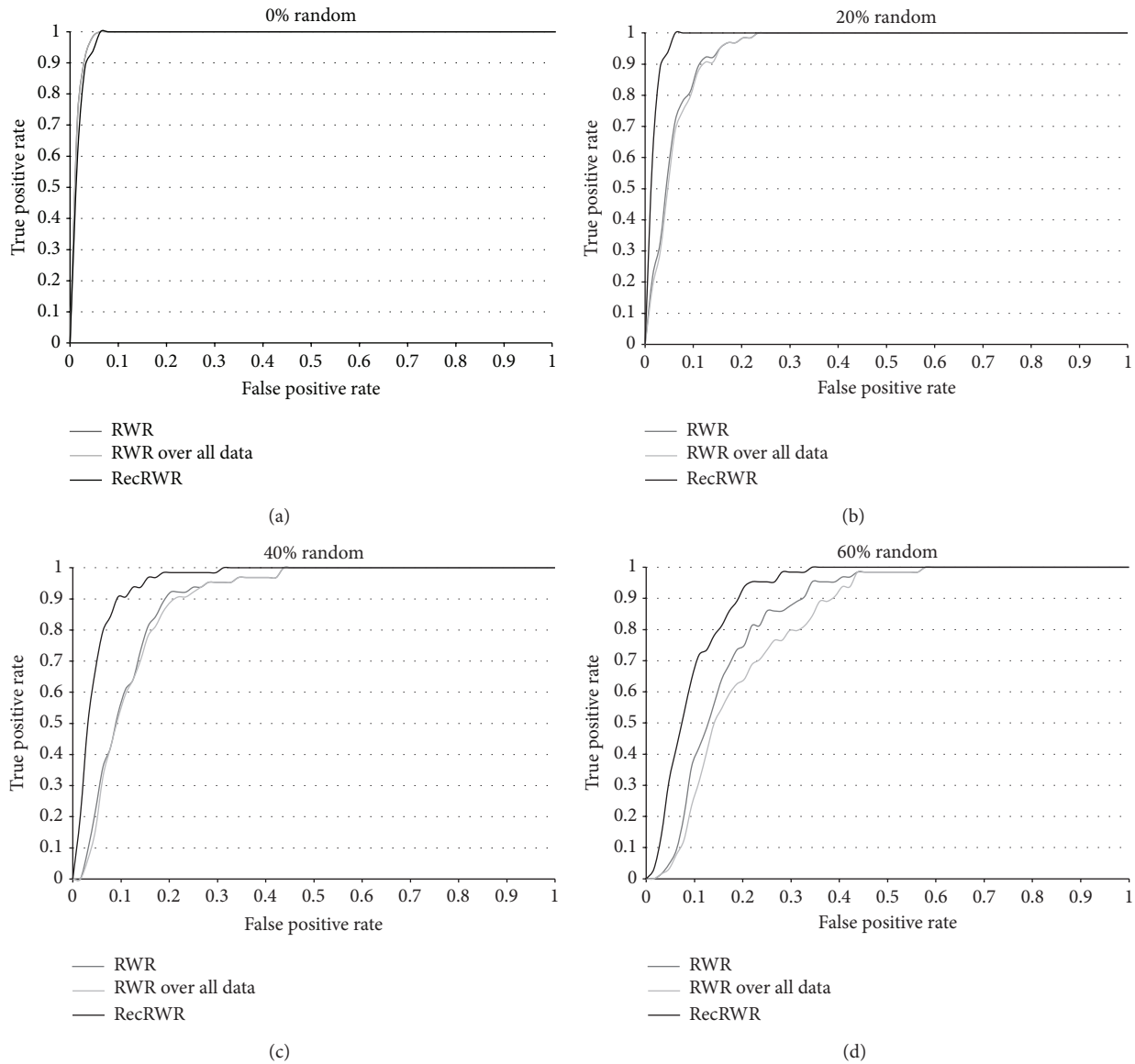
Figure 2: ROC curves with the comparison of the overall performance of RecRWR against existent methods.

that the probable causes of breast cancer are related to the impairment of all these functions. For instance, a genetic mutation causing loss of function on a tumor suppressor gene (such as the cellular tumor antigen p53, P04637) product would result in unrestrained cellular proliferation. Conversely, the transformation of a protooncogene (a gene that participates in a cell-growth pathway) into an oncogene (a protein that can induce cancer on animals) requires a gain-of-function mutation that will allow its permanent activation. An example of this is the epidermal growth factor receptor (EGFR, P00533), also present in Figure 3. EGFR is involved in the conversion of extracellular stimulus to cellular responses. Also, transcription errors are usually immediately corrected by DNA repairing proteins, like the DNA repair and recombination protein RAD54-like (Q92698), shown in the network. A mutation in this gene would result in

the defective proteins, and subsequently the correction of transcription and translation errors would cease. Finally, the protein caspase-8 also seems to be a possible cause of breast cancer. Since caspase-8 is involved in the apoptotic process, impairment of this protein would result in the absence of apoptosis, and defective cells would not be destroyed.

The shortest path between the two diseases is mediated by the cellular tumor antigen p53. There are however other connections between the two nodes. For instance, the proteins receptor tyrosine-protein kinase erbB-2 (P04626), GTPase KRas (P01116), and caspase-8 (Q14790) also connect the two cancers. The influence of caspase-8 mutations on the onset of cancer was explained above. ERBB2 is a protooncogene, with the potential of being converted into an oncogene and inducing cancer. The GTPase KRas is involved in a great variety of important biological processes, including regulation of both
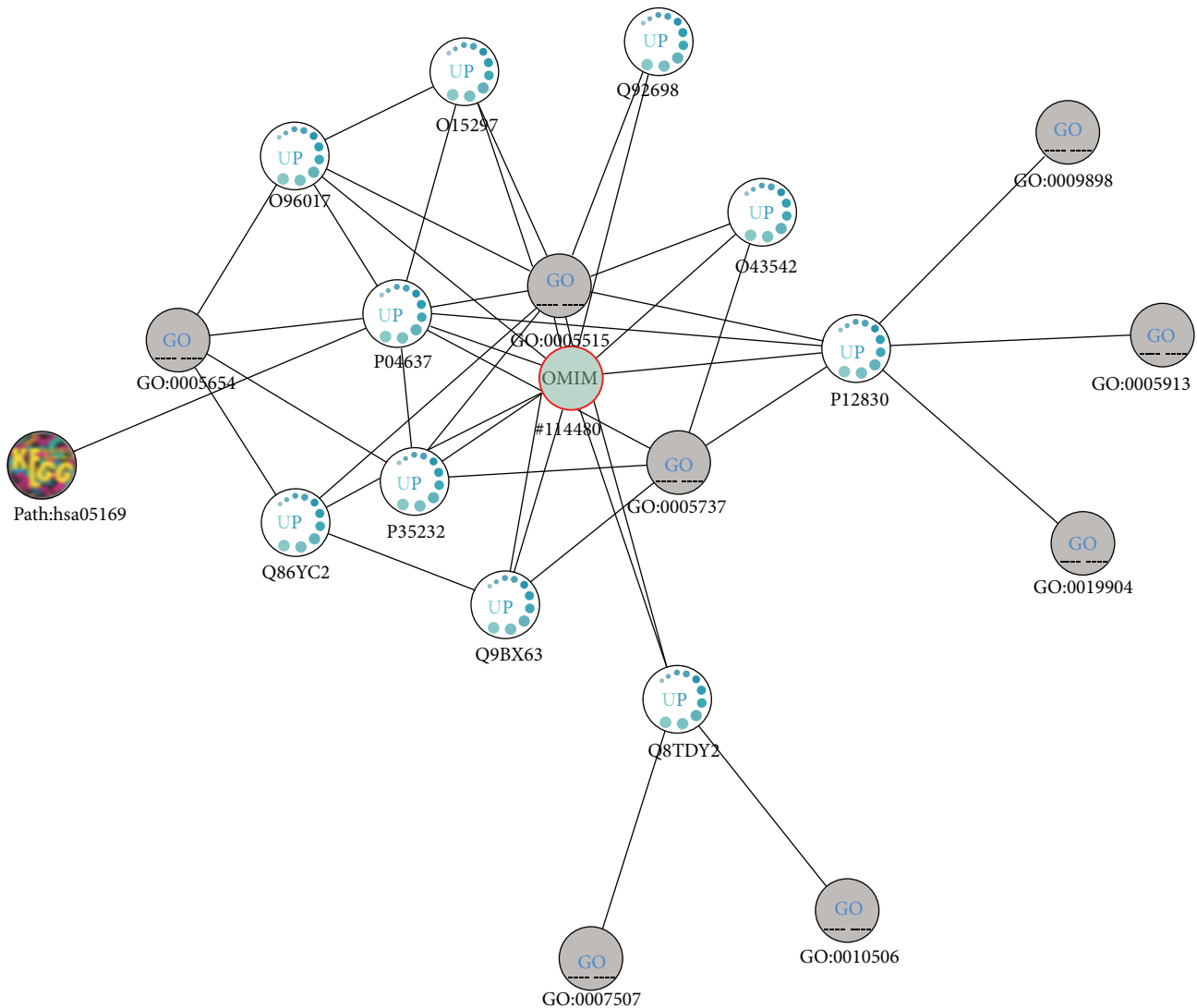
Figure 3: Network of biological concepts associated with breast cancer.

of cell proliferation and gene expression, signal transduction, and cell signalling. The majority of the proteins analysed here are part of the same KEGG pathways: pathways in cancer (hsa05200), neurotrophin signaling pathway (hsa04722), and focal adhesion (hsa04510). The first pathway consists of an integration of the various cancer pathways. The neurotrophin signalling pathway is responsible for the differentiation and survival of neural cells. However, this second pathway is heavily regulated by other intracellular signalling cascades, in which some of the proteins presented in Figure 3 participate. The focal adhesion pathway plays important roles in the proliferation, differentiation, and survival of cells and in gene expression. In case of compromise of any of the proteins involved on this pathway cellular communication becomes defective, which can also result in cancer.

4. Conclusion

In this paper, we have proposed a graph-based approach to address the problem of selecting the best disease targets for multiconcept graphs. Towards this aim we build a multilayer biomedical graph that stores PPI data, annotated with stable knowledge from OMIM diseases and Biological Process from Gene Ontology. The inherent improvements of the proposed method are the use of multilayer networks formed with the PPI data and by the terms' associations; combination of this data to establish new associations among nodes; and use of degree-based methods for evaluating node weights.

Finally, we have presented comprehensive validation that demonstrates the superiority of the proposed approach, Recursive Random Walk with Restarts (RecRWR). The obtained results outline the superiority of the proposed approach in identifying disease candidates, especially with high levels of biological noise and benefiting from all data available.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Joel P. Arrais and José Luís Oliveira contributed equally to the work presented here.

Acknowledgments

References

[1] C. Giallourakis, C. Henson, M. Reich, X. Xie, and V. K. Mootha, "Disease gene discovery through integrative genomics," *Annual Review of Genomics and Human Genetics*, vol. 6, pp. 381–406, 2005.

[2] H. G. Brunner and M. A. van Driel, "From syndrome families to functional genomics," *Nature Reviews Genetics*, vol. 5, no. 7, pp. 545–551, 2004.

[3] C. Auffray, Z. Chen, and L. Hood, "Systems medicine: the future of medical genomics and healthcare," *Genome Medicine*, vol. 1, no. 1, article gm2, 2009.

[4] J. Chen, H. Xu, B. J. Aronow, and A. G. Jegga, "Improved human disease candidate gene prioritization using mouse phenotype," *BMC Bioinformatics*, vol. 8, article 392, 2007.

[5] S. Aerts, D. Lambrechts, S. Maity et al., "Gene prioritization through genomic data fusion," *Nature Biotechnology*, vol. 24, no. 5, pp. 537–544, 2006.

[6] S. Rossi, D. Masotti, C. Nardini et al., "TOM: a web-based integrated approach for identification of candidate disease genes," *Nucleic Acids Research*, vol. 34, pp. W285–W292, 2006.

[7] Y. Moreau and L.-C. Tranchevent, "Computational tools for prioritizing candidate genes: boosting disease gene discovery," *Nature Reviews Genetics*, vol. 13, no. 8, pp. 523–536, 2012.

[8] J. P. Arrais, J. Fernandes, J. Pereira, and J. L. Oliveira, "GeneBrowser 2: an application to explore and identify common biological traits in a set of genes," *BMC Bioinformatics*, vol. 11, article 389, 2010.

[9] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[10] M. P. Joy, A. Brock, D. E. Ingber, and S. Huang, "High-betweenness proteins in the yeast protein interaction network," *Journal of Biomedicine & Biotechnology*, vol. 2005, no. 2, pp. 96–103, 2005.

[11] H.-W. Ma and A.-P. Zeng, "The connectivity structure, giant strong component and centrality of metabolic networks," *Bioinformatics*, vol. 19, no. 11, pp. 1423–1430, 2003.

[12] S. Erten, G. Bebek, R. M. Ewing, and M. Koyutürk, "DADA: degree-aware algorithms for network-based disease gene prioritization," *BioData Mining*, vol. 4, no. 1, article 19, 2011.

[13] J. P. Arrais and J. L. Oliveira, "Using biomedical networks to prioritize gene–disease associations," *Open Access Bioinformatics*, vol. 2011, no. 3, pp. 123–130, 2011.

[14] K. Macropol, T. Can, and A. K. Singh, "RRW: repeated random walks on genome-scale protein networks for local cluster discovery," *BMC Bioinformatics*, vol. 10, article 283, 2009.

[15] L. Yu, L. Gao, and K. Li, "A method based on local density and random walks for complexes detection in protein interaction networks," *Journal of Bioinformatics and Computational Biology*, vol. 8, supplement 1, pp. 47–62, 2010.

[16] D.-H. Le and Y.-K. Kwon, "GPEC: A Cytoscape plug-in for random walk-based gene prioritization and biomedical evidence collection," *Computational Biology and Chemistry*, vol. 37, pp. 17–23, 2012.

[17] M. Re, M. Mesiti, and G. Valentini, "A fast ranking algorithm for predicting gene functions in biomolecular networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 6, pp. 1812–1818, 2012.

[18] S. Kohler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *The American Journal of Human Genetics*, vol. 82, no. 4, pp. 949–958, 2008.

[19] D. Szklarczyk, A. Franceschini, M. Kuhn et al., "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research*, vol. 39, no. 1, pp. D561–D568, 2011.

[20] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Research*, vol. 33, pp. D514–D517, 2005.

[21] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.