Le *Algorithms for Molecular Biology* (2015) 10:14
DOI 10.1186/s13015-015-0044-6

ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

A novel method for identifying disease associated protein complexes based on functional similarity protein complex networks

Duc-Hau Le

Abstract

Background: Protein complexes formed by non-covalent interaction among proteins play important roles in cellular functions. Computational and purification methods have been used to identify many protein complexes and their cellular functions. However, their roles in terms of causing disease have not been well discovered yet. There exist only a few studies for the identification of disease-associated protein complexes. However, they mostly utilize complicated heterogeneous networks which are constructed based on an out-of-date database of phenotype similarity network collected from literature. In addition, they only apply for diseases for which tissue-specific data exist.

Methods: In this study, we propose a method to identify novel disease-protein complex associations. First, we introduce a framework to construct functional similarity protein complex networks where two protein complexes are functionally connected by either shared protein elements, shared annotating GO terms or based on protein interactions between elements in each protein complex. Second, we propose a simple but effective neighborhood-based algorithm, which yields a local similarity measure, to rank disease candidate protein complexes.

Results: Comparing the predictive performance of our proposed algorithm with that of two state-of-the-art network propagation algorithms including one we used in our previous study, we found that it performed statistically significantly better than that of these two algorithms for all the constructed functional similarity protein complex networks. In addition, it ran about 32 times faster than these two algorithms. Moreover, our proposed method always achieved high performance in terms of AUC values irrespective of the ways to construct the functional similarity protein complex networks and the used algorithms. The performance of our method was also higher than that reported in some existing methods which were based on complicated heterogeneous networks. Finally, we also tested our method with prostate cancer and selected the top 100 highly ranked candidate protein complexes. Interestingly, 69 of them were evidenced since at least one of their protein elements are known to be associated with prostate cancer.

Conclusions: Our proposed method, including the framework to construct functional similarity protein complex networks and the neighborhood-based algorithm on these networks, could be used for identification of novel disease-protein complex associations.

Keywords: Disease protein complex, Functional similarity protein complex network, Neighborhood-based algorithm, Prostate cancer

Correspondence: hauldhut@gmail.com
School of Computer Science and Engineering, Water Resources University, 175 Tay Son, Dong Da, Hanoi, Vietnam

Background

Protein complexes are formed by non-covalent interactions among proteins and have specific biological functions. These protein complexes and their cellular functions have been concurrently identified by a number of methods based on protein interaction networks [1-3] and affinity purification-mass spectrometry experiments [4]. However, their particular roles in terms of causing disease have not yet been well-determined. Indeed, all protein complexes in a most updated database of protein complexes CORUM [5] have been well functionally annotated and categorized; however, few of them have a comment on their association with diseases.

It is shown that interactions among proteins forming protein complexes do not only provide a better understanding of cellular functions, but also improve our understanding about human diseases [6-9]. A number of studies have shown the association between protein complexes and specific diseases. For instance, a protein complex of SCRIB, NOS1AP and VANGL1 is associated with breast cancer progression [10], TWIST/Mi2/NuRD protein complex has an essential role in cancer metastasis [11], aberrant protein complex consisting of prostaglandin-d-synthase (PDS) and transthyretin (TTR) is a biomarker of Alzheimer's disease [12]. In addition, past studies show that mutations in multiple proteins that form a protein complex may lead to the same disease phenotype. Therefore, protein complexes can be used to predict phenotypic effects of gene mutation and identify human disease genes [8]. Some early studies made use of protein complexes to predict novel disease genes [13,14]. However, they did not use actual protein complexes, but those simply assembled by neighboring proteins [13] or generated from densely connected subsets of ranked proteins [14]. In other words, the formation of such protein complexes was mainly based on topological properties rather than functional similarities of their protein elements. In addition, biological relationships between protein complexes were also omitted in those studies. Considering an observation that if two protein complexes have biological relationships (e.g., they share a number of common protein elements or their protein elements are highly physically connected), the mutations of genes in one protein complex can lead to same or similar phenotypes of the other protein complex [15], the functional interaction between protein complexes can play an important role in predicting phenotypic effects of gene mutation. Indeed, a recent study [16] used a heterogeneous network consisting of a global protein complex network layer and phenotype similarity network layer to predict novel disease phenotype-gene associations. In that study, the protein complex network layer was constructed using existing human protein complexes and a human protein interaction network. Then, a network propagation algorithm was applied on the heterogeneous network to prioritize candidate genes. Ultimately, they reported that their method outperformed other methods which were solely based on the human protein interaction network and phenotype similarity network for the prediction of novel disease phenotype-gene associations [17,18]. Taken together, these studies indicate that the protein complexes can be used to improve predictability of novel disease phenotype-gene associations. However, identification of novel direct disease-protein complex associations has not yet been well-focused. Indeed, only a few studies have directly focused on this problem recently [19-22]. For instance, study [19] used a complicated heterogeneous network including three layers (i.e., a phenotypic similarity network layer, a tissue-specific protein interaction network layer, and a protein complex membership layer), and then applied a network propagation algorithm on that network to discover disease-associated protein complexes. However, the phenotype similarity network was collected from a relatively old published study [23]; therefore it is not up-to-date. In addition, they were also limited in the prediction of disease of which a tissue-specific protein interaction network [19] or gene expression data [22] exist. Having the same limitation in using the out-of-date phenotype similarity network as in [19], study [20] prioritized protein complexes implicated in human diseases using a maximum information flow algorithm on a heterogeneous network which was constructed by combining a protein interaction network and the phenotypic similarity network. Recently, we also introduced a method for identification of disease associated protein complexes using a random walk with restart (Shortly called RWR) algorithm, which yields a global similarity measure, on a constructed functional similarity protein complex network. This method achieved relative high performance [21]. However, the functional similarity between protein complexes in the constructed protein complex network was only based on their shared protein elements.

In this study, we propose a novel method to identify novel disease-protein complex associations. First, we presented a framework to construct functional similarity protein complex networks where each node is a protein complex and two protein complexes are functionally connected if they either share protein elements, share annotating gene ontology (GO) [24] term or are connected by protein interactions. Then, we proposed a novel neighborhood-based algorithm (Shortly called NBH), which yields a local similarity measure, to prioritize disease candidate protein complexes. We compared the performance of our algorithm to two state-of-the-art network propagation algorithms, RWR [21] and PRINCE [14], on the three constructed functional similarity protein complex networks. The performance of each algorithm was

assessed based on a set of known disease-protein complex associations using a leave-one-out cross validation method. The comparative results showed that NBH statistically significantly outperformed that of the RWR and PRINCE algorithms in predicting novel disease-protein complex associations. In addition, NBH consumed less running time in ranking candidate protein complexes than that of the RWR and PRINCE algorithms. Moreover, relatively high performances achieved for all the constructed functional similarity protein interaction networks and the three algorithms indicated the stability and feasibility of our proposed method for the identification of novel disease associated protein complexes. Furthermore, a case study on prostate cancer was performed. As a result, 69 out of top 100 highly ranked protein complexes were shown to be associated with prostate cancer.

Methods

Databases of protein complexes and disease phenotype-gene associations

First, we obtained 1,704 human protein complexes from a database of mammalian organisms protein complex CORUM [5] (See in Additional file 1: Table S1). These protein complexes were manually annotated with their functions, localization, subunit composition and literature references. Then, we used known disease phenotype-gene associations from OMIM [25] to construct known disease phenotype-protein complex associations.

Construction of known disease phenotype-protein complex associations

To our knowledge, there is no standard database of disease-protein complex associations in public resources. Therefore, based on an observation from a number of studies that mutations in multiple proteins that form a protein complex may lead to the same disease [7,8,26-29], we defined a known disease phenotype-protein complex association as follows: a protein complex is associated with a disease phenotype if at least one of its protein elements is associated with the disease phenotype. Based on this definition of the association, the set of collected human protein complexes and the set of known disease phenotype-gene associations, we constructed 282 disease phenotypes and their known associated protein complexes (See in Additional file 1: Table S2).

Construction of functional similarity protein complex networks

A protein complex is formed by structurally and functionally related protein elements. Indeed, protein elements within a protein complex have higher GO-based semantic similarities than that on all proteins [30]. In addition, protein complexes often correspond to functionally and structurally cohesive substructures/densely connected regions in protein interaction network [31,32]. Therefore, to construct functional similarity protein complex networks where each node is a protein complex, we defined a functional similarity interaction between two protein complexes based on: i) shared protein elements; ii) shared annotating GO terms; and iii) protein interactions. Figure 1 shows an illustrative example of the construction of functional similarity interactions between two protein complexes.

Given two protein complexes $Ci = \{p_k | k = 1 \ldots m\}$ and $Cj = \{p_l | l = 1 \ldots n\}$, where p_k, p_l are protein elements of Ci and Cj, m and n are the number of protein elements belonging to Ci and Cj, respectively, we here defined three kinds of functional similarity interactions between Ci and Cj based on shared protein elements, shared annotating GO terms and protein interactions.

A functional similarity interaction between two protein complexes based on their shared protein elements

A protein complex is formed by functionally related protein elements. Therefore, it could be accepted that the more number of protein elements two protein complexes share, the more functionally related they are. Indeed, it was shown that this kind of interaction not only reflects physical interaction of complexes, but may also represent common regulation, localization, turnover or architecture [33]. Therefore, as we did in our previous study [21], we defined a functional similarity interaction between two protein complexes based on their shared protein elements as follows: Two protein complexes are functionally interacted with each other if they share at least one protein element. And, a weight of this interaction is number of the shared protein elements normalized by number of elements of a protein complex which has a smaller number of elements. It is formally defined as follows:

$$S_C = \frac{|Ci \cap Cj|}{min(|Ci|, |Cj|)}$$

Based on this definition and the set of collected human protein complexes, we constructed a functional similarity protein complex network including 1,579 nodes and 16,981 interactions (shortly called *SharedProteinComNet*).

A functional similarity interaction between two protein complexes based on their shared annotating GO terms

Functions of each protein elements are described by annotating GO terms. Therefore, we additionally constructed a functional similarity interaction between two protein complexes based on shared annotating GO terms, where a list of genes and annotating GO terms were collected from NCBI Entrez Gene FTP site (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz). More specifically, we defined the interaction based on a measure of gene functional similarity

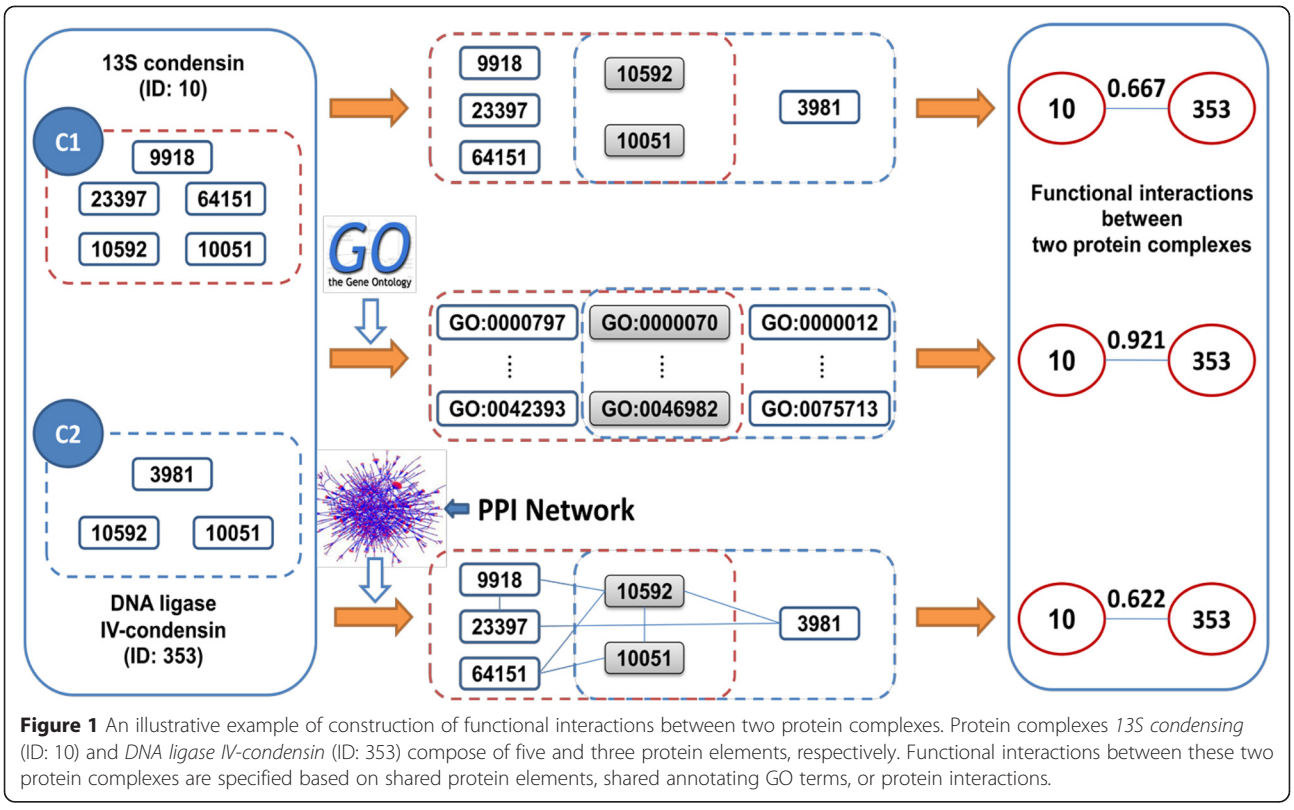

Figure 1 An illustrative example of construction of functional interactions between two protein complexes. Protein complexes *13S condensing* (ID: 10) and *DNA ligase IV-condensin* (ID: 353) compose of five and three protein elements, respectively. Functional interactions between these two protein complexes are specified based on shared protein elements, shared annotating GO terms, or protein interactions.

introduced in [34] as follows: Two protein complexes are functionally interacted with each other if they share at least one annotating GO term, where annotating GO terms include direct annotating GO terms and their ancestors in the GO directed acyclic graph. Then, weight of the interaction is defined as number of the shared annotating GO terms normalized by number of annotating GO terms of a protein complex whose protein elements were annotated with smaller number of annotating GO terms. Formally, a functional similarity interaction between *Ci* and *Cj* is defined based on their shared annotating GO terms as follows:

$$S_C = \frac{\left|GO_{Ci} \cap GO_{Cj}\right|}{min\left(\left|GO_{Ci}\right|,\ \left|GO_{Cj}\right|\right)}$$

where GO_{Ci}, GO_{Cj} are sets of GO terms annotating to protein elements in *Ci* and *Cj*, respectively.

As in [34], GO_{Ci} and GO_{Cj} are defined as follows:

$$GO_{Ci} = \bigcup_{k}^{m}(GO_k \cup anc(GO_k))$$

And

$$GO_{Cj} = \bigcup_{l}^{n}(GO_l \cup anc(GO_l))$$

where: GO_k and GO_l are set of direct annotating GO terms of protein p_k and p_l, respectively; $anc(GO_k)$ and $anc(GO_l)$ are ancestors of GO_k and GO_l excluding root terms (i.e., GO:0008150 for biological process, GO:0005575 for cellular component and GO:0003674 for molecular function) in GO directed acyclic graph of each sub-category, respectively.

For each GO sub-category, we constructed a functional similarity protein complex network. Then we integrated them using "per-edge average" method as follows:

$$\bar{S}_C = \frac{1}{M}\sum_{k=1}^{M}(S_C)_k$$

where M is number of networks containing interaction between protein complex C_i and C_j. $(S_C)_k$ is functional similarity interaction between C_i and C_j in network k.

Consequently, we constructed a functional similarity protein complex network including 1,683 nodes and 1,415,266 interactions based on annotating GO terms (shortly called *SharedGOTermComNet*).

A functional similarity interaction between two protein complexes based on their shared protein interactions

A past study proposed a method to measure a functional relationship between two gene sets based on protein interaction network [35]. By considering a protein complex a special case of a gene set, the study showed that protein complexes with high functional similarities tend to be involved in the same functional catalogue and

these functional similarities were successfully used in prioritizing candidate cancer-associated protein complexes [35]. In this study, we additionally used that method to define a functional similarity interaction between two protein complexes based on protein interaction network. To this end, we collected a human physical protein interaction network (PPI) consisting of 10,486 genes and 50,791 interactions from the NCBI Entrez Gene FTP site (ftp://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/interactions.gz). This network was constructed by integrating BIND [36], BioGRID [37] and HPRD [38]. Formally, a functional similarity interaction between *Ci* and *Cj* was defined based on protein interactions among protein elements belonging to the two protein complexes as follows:

$$S_C = \frac{\sum_{k=1}^{m}\sum_{l}^{n}\frac{1}{SP(p_k,\, p_l)}}{m \times n}$$

Where

$$SP(p_k,\, p_l) = \begin{cases} 1 & \text{if } p_k \equiv p_l \text{ or } p_k,\ p_l \in C_i \cap C_j \\ \text{Length of shortest path between } p_k \text{ and } p_l \end{cases}$$

Based on this definition, the set of collected human protein complexes and the human physical protein interaction network, we constructed a functional similarity protein complex network including 1,681 nodes and 1,412,040 interactions based on protein interactions (shortly called *SharedPPIComNet*).

Network-based ranking algorithms

Given a connected weighted graph G(V, E) with a set of nodes $V = \{v_1, v_2, \ldots, v_N\}$ and a set of links $E = \{(v_i, v_j) |\ v_i, v_j \in V\}$, a set of source nodes $S \subseteq V$ and a $N \times N$ adjacency matrix W of link weights. Here, we are going to introduce our proposed neighborhood-based algorithm. In addition, we also briefly describe the RWR algorithm, which was used in our previous study [21], and the PRINCE algo rithm [14]. These three algorithms will be used for measuring relative similarity of node v_i to S. By modelling a functional similarity protein complex network as a graph (i.e., nodes present protein complexes, links present functional interactions among protein complexes, W presents pair-wise similarities between protein complexes, and S presents known disease-associated protein complexes), a ranking of candidate protein complexes based on their relative similarity to S is to predict novel disease-associated protein complexes. The relative similarity also measures how relevant to a disease a candidate protein complex is.

The proposed neighborhood-based algorithm

The neighborhood-based algorithm (shortly called NBH) was based on direct neighbors of source nodes (S). Formally, the relative similarity of a node v_i to a set of source nodes (S) was defined as following:

$$p_i = \sum_{j \in S} w_{ij}$$

where w_{ij} is weight of link (v_i, v_j). This score is 0 for nodes not connected to any source nodes.

Random Walk with Restart (RWR)

RWR is a variant of the random walk [39] and it mimics a walker that moves from a current node to a randomly selected adjacent node or goes back to source nodes with a back-probability $\gamma \in (0, 1)$. RWR can be formally described as follows:

$$P^{t+1} = (1-\gamma) W' P^t + \gamma P^0$$

where P^t is a $N \times 1$ probability vector of $|V|$ nodes at a time step t of which the ith element represents the probability of the walker being at node $v_i \in V$, and P^0 is the $N \times 1$ initial probability vector where the value of an element corresponding to a non-source node or a source node is zero or $1/|S|$, respectively. The matrix W' is represented by a transition probability matrix and thus $(W')_{ij}$, the (i, j) element in W', denotes a probability with which a walker at v_i moves to v_j among $V \backslash \{v_i\}$. Formally, it is defined as follows:

$$(W')_{ij} = \frac{(W)_{ij}}{\sum_{k \in (V_{out})_i} (W)_{ik}}$$

where $(V_{out})_i$ is a set of outgoing nodes of v_i.

All nodes in the network are eventually ranked according to the steady-state probability vector P^{∞}, which is obtained by repeating the iterations until $||P^{t+1} - P^t|| < 10^{-6}$ in this study.

PRINCE

Study [14] introduced PRINCE for disease gene prediction. Similar to RWR, PRINCE is a kind of network propagation algorithm, it simulates a process where nodes for which prior information exists pump information to their neighbors though an iteration process. At each iteration, every node propagates the information received at the previous iteration to its neighbors. PRINCE can be formally described as follows:

$$P^{t+1} = \alpha W' P^t + (1-\alpha) P^0$$

where P^0 represents a prior knowledge function. Similar to RWR, in this study, it is the $N \times 1$ initial probability vector where the value of an element corresponding to a non-source node or a source node is zero or $1/|S|$, respectively. The first part of the equation represents the smooth propagation process which assigns similar values

to adjacent nodes, while the second part represents prior knowledge. Trade-off parameter $\alpha \in (0, 1)$ weighs the relative importance of these constraints with respect to one another. Different from RWR, The matrix W' is represented by a row-normalized matrix and defined formally as follow:

$$(W')_{ij} = \frac{(W)_{ij}}{\sqrt{\sum_{k\in(V_{in})_i}(W)_{ki} \times \sum_{l\in(V_{in})_j}(W)_{lj}}}$$

where $(V_{in})_i$ and $(V_{in})_j$ are a set of incoming nodes of v_i and v_j, respectively.

Similarly to the RWR algorithm, all nodes in the network are eventually ranked according to the steady-state probability vector P^{∞}, which is obtained by repeating the iterations until $||P^{t+1}\text{-}P^{t}|| < 10^{-6}$ in this study.

Performance evaluation

Ranking performance was assessed through the leave-one-out cross-validation (Shortly called LOOCV) process. Let us assume that a functional similarity protein complex network $G(V, E)$, a set of known disease-associated protein complexes ($D \subseteq V$) and a set of candidate protein complexes (C) are given. A protein complex $s \in D$ was held out for validation and the remaining known disease-associated protein complexes were specified to a set of source nodes (*i.e.*, $S = D \backslash \{s\}$). The network-based ranking algorithms were used to prioritize all the candidate protein com plexes. This process was repeated by setting every $s \in D$ to a held-out protein complex. For a reliable performance comparison, we drew the receiver operating characteristic (ROC) curves and computed the area under the curve (AUC) based on the rank of held-out protein complex s and candidate protein complexes in set $C \cup \{s\}$. More specifically, given a threshold τ, we counted *TP* (true positives), *FN* (false negatives), *FP* (false positives), and *TN* (true negatives), which were formally defined as following:

$$TP = \sum_{s\in D} I(rank(s) \le \tau) \qquad FN = \sum_{s\in D} I(rank(s) > \tau)$$

$$FP = \sum_{c\in C} I(rank(c) \le \tau) \qquad TN = \sum_{c\in C} I(rank(c) > \tau)$$

where $rank(s)$, $rank(c)$ and $I(\cdot)$ denote the rank of s, the rank of a candidate protein complex c out of the set $C \cup \{s\}$ and the indicator function, respectively. Then, we defined *sensitivity* and (1-*specificity*) as follows:

$$sensitivity = \frac{TP}{TP+FN} \qquad 1\text{-}specificity = \frac{FP}{FP+TN}$$

By varying τ from one to the number of protein complexes in the set $C \cup \{s\}$, the relationship between *sensitivity* and (1-*specificity*) was plotted. The ROC curve is the curve constructed based on those pairs of values, and the AUC is the area under the ROC curve. In this study, we considered candidate protein complexes set as all protein complexes that are not known to be associated with the disease (*i.e.*, $V \backslash D$) in G.

Results and discussion

Performance comparison

Due to using LOOCV method, we collected only 270 disease phenotypes which are known to be associated with at least two protein complexes to assess the performance of each algorithm. For RWR and PRINCE algorithm, we varied back-probability and trade-off parameters respectively in a range of [0.1, 0.9]. The AUC values were calculated for each disease phenotype on each individual functional similarity protein complex network. Then, the performance of each algorithm was averaged over those disease phenotypes for each individual functional similarity protein complex network. Figure 2 shows that the performance of NBH was statistically significantly better than that of RWR and PRINCE (i.e., All P-values < 0.05 using two sample t-Test. See more detail in Table 1).

Another advance of NBH compared to RWR and PRINCE is that NBH is free of parameters, whereas the performance of RWR and PRINCE is controlled by back-probability and trade-off parameters, respectively. Indeed, Figure 2 shows that the performance of RWR was improved when back-probability increases (i.e., slopes of regression lines for RWR in Figure 2(a), (b) and (c) are 0.00230, 0.22108 and 0.20141 with P-values are 0.00307, 0.00409, 0.005 using ANOVA test for regression, respectively). In contrast, the performance of PRINCE declined when trade-off increases (i.e., slopes of regression lines for PRINCE in Figure 2(a), (b) and (c) are −0.00147, −0.03838 and −0.08100 with P-values are 0.00015, 0.02758, 0.00714 using ANOVA test for regression, respectively). This opposite trends of RWR and PRINCE is because RWR and PRINCE algorithms are generally similar in that the first part of RWR and PRINCE represents the smooth propagation process which assigns similar values to adjacent nodes, while the second part represents prior knowledge. However, the parameters affect inversely for each algorithm (i.e., when back-probability increases the random walker of RWR tend to go back to the source nodes and therefore give higher score for nodes nearby source nodes. In contrast, when trade-off increases the random walker of PRINCE tends to go far from source nodes and therefore assign lower score for nodes nearby source nodes). Figure 2 also shows that RWR and PRINCE achieved best performance when back-probability and trade-off parameters are set to 0.9 and 0.1, respectively. This observation implied that by linking protein complexes by functional interactions, protein complexes associated with the same

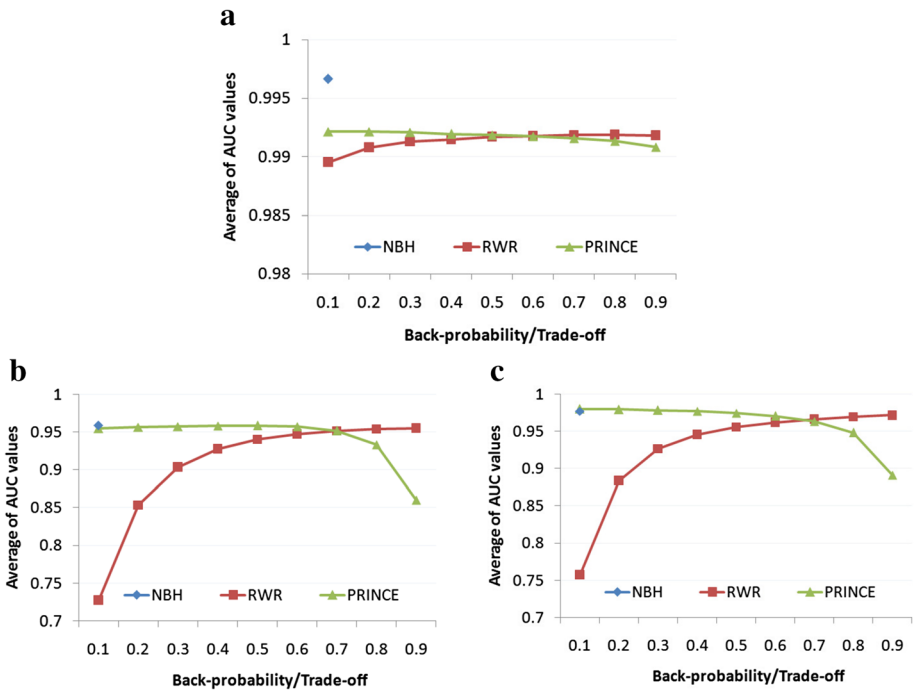

Figure 2 Performance comparison between network-based ranking algorithms on different functional similarity protein complex networks. **(a)** *SharedProteinComNet*. **(b)** *SharedGOTermComNet*. **(c)** *SharedPPIComNet*. For RWR and PRINCE, back-probability and trade-off parameters were varied in a range of [0.1, 0.9], respectively. Vertical axis represents average AUC values over 270 disease phenotypes.

or similar disease phenotypes tend to be connected closely. This is also the reason why NBH, which is only based on neighbors of known disease protein complexes, achieved better performance than that of RWR and PRINCE.

In addition, we observed that the performance of RWR shows less positive trend in *SharedProteinComNet* than that in *SharedGOTermComNet* and *SharedPPIComNet* (i.e., the slopes of regression lines are 0.00230, 0.22108 and 0.20141, respectively), and PRINCE was shown less as a negative trend in *SharedProteinComNet* than that in *SharedGOTermComNet* and *SharedPPIComNet* (i.e., the slopes of regression lines are −0.00147, −0.03838 and −0.08100, respectively). Considering the network density of *SharedProteinComNet* was much lower than that in *SharedGOTermComNet* and *SharedPPIComNet* (i.e., the network densities are 0.014, 0.999 and 1, respectively), this observation indicated that disease protein complexes are more connected in *SharedGOTermComNet* and *SharedPPIComNet*. However, networks with very high densities such as *SharedGOTermComNet* and *SharedPPIComNet*, may contain unreliable functional interactions that can reduce the performance of the methods. Indeed, Figure 3 shows that the performance of all algorithms

Table 1 Performance comparison between network-based ranking algorithms on different functional similarity protein complex networks

#	Protein complex network	NBH	RWR	PRINCE
1	*SharedProteinComNet*	0.99665 (±0.00988)	0.99134 (±0.01663)	0.99174 (±0.01633)
			(P-value = 4.37×10^{-14})	(P-value = 1.7×10^{-12})
2	*SharedGOTermComNet*	0.95869 (±0.07599)	0.90671 (±0.10804)	0.94305 (±0.08432)
			(P-value = 1.28×10^{-27})	(P-value = 8.2×10^{-4})
3	*SharedPPIComNet*	0.97704 (±0.05926)	0.92646 (±0.12794)	0.96257 (±0.08166)
			(P-value = 1.28×10^{-27})	(P-value = 1.51×10^{-4})

–P-values represent the statistical significance of performance comparison between NBH and RWR/PRINCE algorithms. They were calculated using two-sample t-Test for mean assuming unequal variances.
–Data in each cell represent mean (±standard deviation).
–Performance of NBH is statistically significantly better than that of RWR and PRINCE for all three functional similarity protein complex networks. Data row #1, #2 and #3 are detail performance comparison of the three algorithms on *SharedProteinComNet*, *SharedGOTermComNet* and *SharedPPIComNet* networks corresponding to Figure 2(a), (b) and (c), respectively.

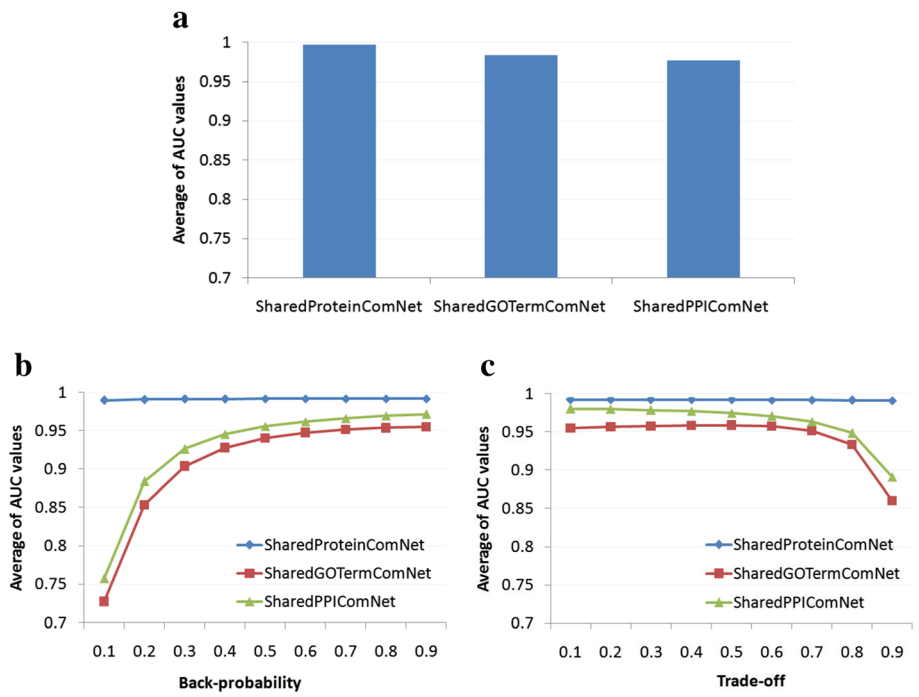

Figure 3 Performance comparison between functional similarity protein complex networks on different network-based ranking algorithms. **(a)** NBH. **(b)** RWR. **(c)** PRINCE. For RWR and PRINCE, back-probability and trade-off parameters were varied in a range of [0.1, 0.9], respectively. Vertical axis represents average AUC values over 270 disease phenotypes.

on *SharedProteinComNet* was better than that on *SharedGOTermComNet* and *SharedPPIComNet* (i.e., All P-values < 0.05. See more detail in Table 2).

To test the hypothesis of whether density affect the performance of the algorithms, we varied a threshold (t) of interaction weight to extract different functional similarity protein complex networks of which interactions having weight no less than t. Particularly, for *SharedGOTermComNet* and *SharedPPIComNet*, we additionally extracted five networks corresponding to t = 0.1, 0.3, 0.5, 0.7, and 0.9. Figure 4 shows the performance of each algorithm on *SharedGOTermComNet* and *SharedPPIComNet* with different thresholds. As observed, for NBH (Figure 4(a) and (b)), the best performance was achieved with "All" (i.e., *SharedGOTermComNet* and *SharedPPIComNet*) and decreased when the threshold increased. However, for RWR (Figure 4(c) and (d)) and PRINCE (Figure 4(e) and (f)), it performed best and most stable with t = 0.7 or t = 0.9. This result indicated these two algorithms perform better when unreliable functional interactions were eliminated.

In addition to comparison of prediction performance of the proposed algorithm with RWR and PRINCE, we here show its advance in running time. To this end, we ran each method for 282 disease phenotypes, where their known disease associated protein complexes were used as source nodes. For RWR and PRINCE, we also varied back-probability and trade-off parameters in a range of [0.1, 0.9] and took the average of running time over this range. The final running time for each disease phenotype was averaged over the set of 282 disease phenotypes. Table 3 shows the comparison on the running time of different algorithms. It was obvious that our proposed algorithm run faster than both RWR and PRINCE about 32 times averaging on all the three functional similarity protein complex networks.

To our knowledge, only a few studies have directly been proposed for identification of disease-associated protein complexes [19-22]. In which, as showed in the previous section, NBH algorithm outperformed RWR, which was used in our previous study [21], in both prediction performance and running time for all the three constructed functional similarity protein complex networks. Meanwhile, the study [22] was specifically proposed to identify cancer-associated protein complexes based on gene expression data and it did not show the over prediction performance. For the two remaining studies, they used network propagation [19] and a maximum information flow [20] algorithms on complicated heterogeneous networks. For instance, study [19] constructed a heterogeneous network including three layers including a disease phenotype similarity network layer, a tissue-specific protein interaction network layer and a protein complex membership layer, then they applied the RWR algorithm on this heterogeneous network. Besides using an out-of-date disease phenotype similarity

Table 2 Performance comparison between functional similarity protein complex networks on different network-based ranking algorithms

#	Algorithm	SharedProteinComNet	SharedGOTermComNet	SharedPPIComNet
1	*NBH*	0.99665 (±0.00988)	0.95869 (±0.07599)	0.97704 (±0.05926)
			(P-value = 6.74×10^{-15})	(P-value = 8.43×10^{-8})
2	*RWR*	0.99134 (±0.01663)	0.90671 (±0.10804)	0.92646 (±0.12794)
			(P-value = 1.06×10^{-252})	(P-value = 7.61×10^{-122})
3	*PRINCE*	0.99174 (±0.01633)	0.94305 (±0.08432)	0.96257 (±0.08166)
			(P-value = 7.14×10^{-151})	(P-value = 1.14×10^{-63})

–P-values represent the statistical significance of performance comparison between *SharedProteinComNet* and *SharedGOTermComNet*/*SharedPPIComNet* networks. They were calculated using two-sample t-Test for mean assuming unequal variances.
–Data in each cell represent mean (±standard deviation).
–Performance of *SharedProteinComNet* is statistically significantly better than that of *SharedGOTermComNet* and *SharedPPIComNet* for all three network-based ranking algorithms. Data row #1, #2 and #3 are detail performance comparison of the three functional similarity protein complex networks by NBH, RWR and PRINCE algorithms corresponding to Figure 3(a), (b) and (c), respectively.

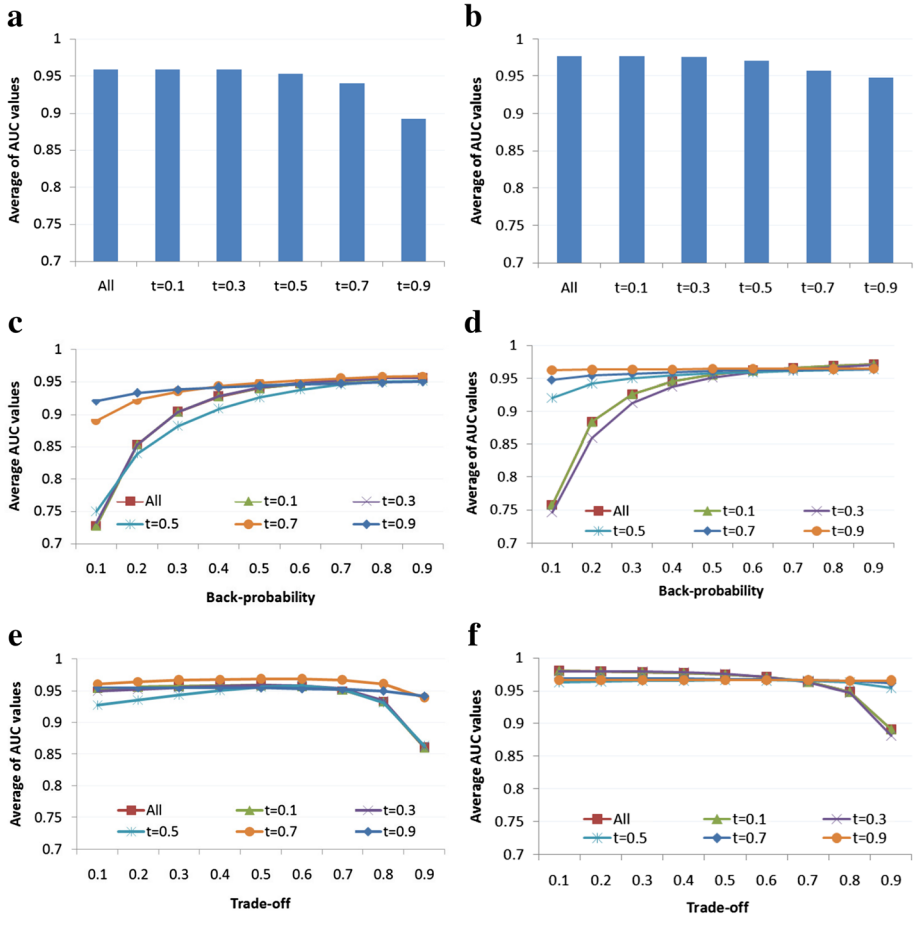

Figure 4 Investigation of the performance of each method on different functional similarity protein complex networks extracted with different thresholds (t). **(a)**, **(c)** and **(e)** are the performance of NBH, RWR and PRINCE for networks which are extracted from *SharedGOTermComNet*, respectively; **(b)**, **(d)** and **(f)** are the performance of NBH, RWR and PRINCE for networks which are extracted from *SharedPPIComNet*, respectively. For RWR and PRINCE, back-probability and trade-off parameters were varied in a range of [0.1, 0.9], respectively. Vertical axis represents average AUC values over 270 disease phenotypes. "All" is the original functional similarity protein complex networks (i.e., *SharedGOTermComNet* in **(a)**, **(c)** and **(e)**; *SharedPPIComNet* in **(b)**, **(d)** and **(f)**).

Table 3 Running time (millisecond) comparison between network-based ranking algorithms on different functional similarity protein complex networks

	NBH (A)	RWR (B)	PRINCE (C)	Speedup (B/A)	Speedup (C/A)
SharedProteinComNet	2	79	81	41.8	42.6
SharedGOTermComNet	110	3369	3428	30.6	31.2
SharedPPIComNet	115	2544	2497	22.2	21.8
Average				**31.5**	**31.9**

–Platform: Intel Core i3 3240 CPU 3.4GHz, 4GB RAM.
–Average running time for ranking all candidate protein complexes on functional similarity protein complex networks of a disease phenotype.

network collected from literature [23], this study can apply to only disease phenotypes of which a tissue-specific protein interaction network exist. Similarly, in the study [20], they also constructed a heterogeneous network including a disease phenotype similarity network layer and protein interaction network layer, in which the disease phenotype similarity network was also collected from literature [23]. It is obvious that, these two methods based on networks that are different from ours. Therefore, it is unfeasible to compare directly the performance of our proposed method with them. However, the best reported performances of these methods are inferior to ours (i.e., the performance in term of AUC value is about 0.88 and 0.92 in the study [19] and study [20], respectively; whereas the worst case for NBH is about 0.96 (See Table 1)).

Case study: prostate cancer

Prostate cancer (MIM ID: 176807) is a complex disease and there were 22 genes associated with it as published in OMIM [25]. Following the definition of a known disease phenotype-protein complex association (See Methods section), we found that 12 protein complexes were known to be associated with prostate cancer (See in Additional file 1: Table S2). These associated protein complexes were set as source nodes and all others as candidates in *SharedProteinComNet*. After applying NBH to rank all candidates, we selected 100 highly ranked candidate protein complexes. By searching associations between 219 genes coding proteins involved in those selected protein complexes with prostate cancer on GeneRIF [40] (ftp://ftp.ncbi.nih.gov/gene/GeneRIF/generifs_basic.gz), we found 28 of them reported to be associated with prostate cancer (See in Additional file 1: Table S3). These protein-coding genes are involved in 69 protein complexes in the top 100 selected protein complexes. For instance, overexpression of Skp2 in protein complex "Ubiquitin E3 ligase (SKP1A, SKP2, CUL1, RBX1)" (ID: 1051) is associated with recurrence following radical prostatectomy in prostate cancer [41]. In addition, lower levels of nuclear beta-catenin in protein complex "JUN-TCF4-CTNNB1 complex" (ID: 1816) is associated with prostate cancer progression [42]. Elevated BRCA1 and NBN truncating mutation in protein complex "BRCA1-RAD50-MRE11-NBS1 complex" (ID: 202) are associated with prostate cancer [43,44] . Beside the 69 evidenced protein complexes, others in the top 100 ranked protein complexes can be candidates for future validation (See in Additional file 1: Table S4).

To make the results for identification of novel prostate cancer-associated protein complexes more convincing, we additionally randomly selected 200 sets of 100 protein complexes among protein complexes in the same functional similarity protein complex network. Then, we repeated the same procedure as we did for the 100 highly ranked candidate protein complexes to find novel prostate cancer-associated protein complexes for each set. The result showed that, on average, about 28 protein complexes in each set were enriched with genes which are associated with Prostate cancer (See in Additional file 1: Table S5). Therefore, the set of 100 highly ranked candidate protein complexes has more prostate cancer associations than expected by chance (P-values = 2.9×10^{-200}, using one-sample t-Test).

Conclusions

Protein complexes play important roles in cellular functions; however, their roles in causing disease have not been paid enough attention because there have been a few studies directly focusing on identifying disease-protein complex associations. In this study, we have proposed a novel method for identification of novel disease associated protein complexes. Comparing to our previous study [21], which used RWR algorithm on a functional similarity protein complex network built based on shared protein elements, in this study, we presented a framework to construct functional similarity protein complex networks based on not only shared protein elements, but also shared annotating GO terms and protein interactions. In addition, we proposed a novel neighborhood-based algorithm to prioritize candidate disease protein complexes. Comparing the performance of our proposed algorithm with that of the two state-of-the-art algorithms (i.e., RWR and PRINCE, which yields global similarity measure), we found that our proposed algorithm outperformed these two algorithms for all the three constructed functional similarity protein complex networks. In addition, it also consumed less running time than these algorithms in

ranking candidate protein complexes for each disease phenotype. Moreover, our proposed algorithm is free of parameters; meanwhile the performance of RWR and PRINCE is controlled by back-probability and trade-off parameters, respectively. Comparing the performance of our proposed method with that of other methods also proposed for identification of disease associated protein complexes, we found that our proposed method is simpler since it is only based on homogeneous network of protein complexes (i.e., the three functional similarity protein complex networks only contain protein complexes as nodes), meanwhile other methods were based on a complicated heterogeneous networks in which a node can be either protein, protein complex or disease phenotype. In addition, the best reported performances of these methods are inferior to our proposed method. Finally, using the proposed method to identify novel prostate cancer associated protein complexes; we found that 69 out of top 100 highly ranked candidate protein complexes have evidences about the association with prostate cancer from literature.

Additional file

Additional file 1: Table S1. List of 1,704 collected human protein complexes. **Table S2.** List of 282 disease phenotypes and their known associated protein complexes. **Table S3.** List of 28 protein coding genes (involved in 100 highly ranked candidate protein complexes) having evidences about the association with prostate cancer from literature. **Table S4.** List of 100 highly ranked candidate protein complexes. 69 out of them have evidences about the association with prostate cancer from literature. **Table S5.** List of protein complexes having evidences about the association with prostate cancer from literature in 200 randomly selected sets of 100 protein complexes.

Competing interests

The author declares that there is no conflict of interests regarding the publication of this article.

Acknowledgements

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2014.21.

Received: 30 June 2014 Accepted: 1 April 2015
Published online: 28 April 2015

References

1. Li X, Wu M, Kwoh C-K, Ng S-K. Computational approaches for detecting protein complexes from protein interaction networks: a survey. BMC Genomics. 2010;11:S3.
2. Mukhopadhyay A, Ray S, De M. Detecting protein complexes in a PPI network: a gene ontology based multi-objective evolutionary approach. Mol BioSyst. 2012;8:3036–48.
3. Ray S, Bandyopadhyay S, Mukhopadhyay A, Maulik U. Incorporating fuzzy semantic similarity measure in detecting human protein complexes in PPI network: A multiobjective approach. In Fuzzy Systems (FUZZ), 2013 IEEE International Conference on; 7-10 July 2013. 2013: 1-8.
4. Choi H. Computational detection of protein complexes in AP-MS experiments. PROTEOMICS. 2012;12:1663–8.
5. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, et al. CORUM: the comprehensive resource of mammalian protein complexes - 2009. Nucleic Acids Res. 2010;38:D497–501.
6. Wang Q, Liu W, Ning S, Ye J, Huang T, Li Y, et al. Community of protein complexes impacts disease association. Eur J Hum Genet. 2012;20(11):1162–7.
7. Dudley AeM, Janse DM, Tanay A, Shamir R, Church GM. A global view of pleiotropy and phenotypically derived gene function in yeast. Molecular Systems Biology. 2005;1(1).
8. Fraser H, Plotkin J. Using protein complexes to predict phenotypic effects of gene mutation. Genome Biol. 2007;8(11):R252.
9. Texier Y, Kinkl N, Boldt K, Ueffing M. From quantitative protein complex analysis to disease mechanism. Vis Res. 2012;75:108–11.
10. Anastas JN, Biechele TL, Robitaille M, Muster J, Allison KH, Angers S, et al. A protein complex of SCRIB, NOS1AP and VANGL1 regulates cell polarity and migration, and is associated with breast cancer progression. Oncogene. 2012;31:3696–708.
11. Fu J, Qin L, He T, Qin J, Hong J, Wong J, et al. The TWIST/Mi2/NuRD protein complex and its essential role in cancer metastasis. Cell Res. 2011;21:275–89.
12. Lovell MA, Lynn BC, Xiong S, Quinn JF, Kaye J, Markesbery WR. An aberrant protein complex in CSF as a biomarker of Alzheimer disease. Neurology. 2008;70:2212–8.
13. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotech. 2007;25(3):309–16.
14. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol. 2010;6(1), e1000641.
15. Brunner HG, van Driel MA. From syndrome families to functional genomics. Nat Rev Genet. 2004;5:545–51.
16. Yang P, Li X, Wu M, Kwoh C-K, Ng S-K. Inferring gene-phenotype associations via global protein complex network propagation. PLoS One. 2011;6(7), e21502.
17. Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. Mol Syst Biol. 2008;4:1.
18. Li Y, Patra JC. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics. 2010;26(9):1219–24.
19. Jacquemin T, Jiang R. Walking on a tissue-specific disease-protein-complex heterogeneous network for the discovery of disease-related protein complexes. BioMed Res Int. 2013; 2013.
20. Chen Y, Jacquemin T, Zhang S, Jiang R. Prioritizing protein complexes implicated in human diseases by network optimization. BMC Syst Biol. 2014;8 Suppl 1:S2.
21. Le D-H, Uy NQ, Dung PQ, Binh HTT, Kwon Y-K. Towards the identification of disease associated protein complexes. Procedia Comput Sci. 2013;23:15–23.
22. Zhao J, Lee SH, Huss M, Holme P. The Network Organization of Cancer-associated Protein Complexes in Human Tissues. Sci Rep. 2013;3.
23. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM. A text-mining analysis of the human phenome. Eur J Hum Genet. 2006;14(5):535–42.
24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Gene Ontol Consort Nat Genet. 2000;25(1):25–9.
25. Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM®). Nucleic Acids Res. 2009;37 suppl 1:D793–6.
26. Parsons AB, Lopez A, Givoni IE, Williams DE, Gray CA, Porter J, et al. Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast. Cell. 2006;126(3):611–25.
27. Gavin A-C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, et al. Proteome survey reveals modularity of the yeast cell machinery. Nature. 2006;440(7084):631–6.
28. Gurtan AM, D'Andrea AD AD. Dedicated to the core: Understanding the Fanconi anemia complex. DNA Repair. 2006;5(9â€"10):1119–1125.
29. Lim LE, Campbell KP. The sarcoglycan complex in limb-girdle muscular dystrophy. Curr Opin Neurol. 1998;11(5):443–52.
30. Guzzi PH, Milano M, Veltri P, Cannataro M. Semantic similarities as discriminative features of protein complexes. Curr Bioinforma. 2013;8(3):347–56.
31. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402(6761 Suppl):C47–52.
32. Tong AHY, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, et al. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science. 2002;295(5553):321–4.
33. Gavin A-C, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. 2002;415(6868):141–7.
34. Mistry M, Pavlidis P. Gene ontology term overlap as a measure of gene functional similarity. BMC Bioinformatics. 2008;9(1):327.

35. Wang Q, Sun J, Zhou M, Yang H, Li Y, Li X, et al. A novel network-based method for measuring the functional relationship between gene sets. Bioinformatics. 2011;27(11):1521–8.
36. Bader GD, Betel D, Hogue CWV. BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. 2003;31(1):248–50.
37. Breitkreutz B-J, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, et al. The BioGRID interaction database: 2008 update. Nucleic Acids Res. 2008;36(suppl_1):D637–40.
38. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human protein reference database–2009 update. Nucleic Acids Res. 2009;37(suppl_1):D767–72.
39. Lovász L. Random walks on graphs: a survey. Combinatorics, Paul erdös is eighty. 1993;2(1):1–46.
40. Mitchell JA, Aronson AR, Mork JG, Folk LC, Humphrey SM, Ward JM. Gene Indexing: Characterization and Analysis of NLM's GeneRIFs. In: Proceedings of AMIA 2003 Symposium. American Medical Informatics Association; 2003.
41. Nguyen PL, Lin DI, Lei J, Fiorentino M, Mueller E, Weinstein MH, et al. The impact of Skp2 overexpression on recurrence-free survival following radical prostatectomy. Urol Oncol. 2011;29:302–8.
42. Horvath LG, Henshall SM, Lee CS, Kench JG, Golovsky D, Brenner PC, et al. Lower levels of nuclear β-catenin predict for a poorer prognosis in localized prostate cancer. Int J Cancer. 2005;113:415–22.
43. Zuhlke K, Johnson A, Okoth L, Stoffel E, Robbins C, Tembe W, et al. Identification of a novel NBN truncating mutation in a family with hereditary prostate cancer. Familial Cancer. 2012;11:595–600.
44. Schayek H, Haugk K, Sun S, True LD, Plymate SR, Werner H. Tumor suppressor BRCA1 is expressed in prostate cancer and controls insulin-like Growth Factor I Receptor (IGF-IR) gene transcription in an androgen receptor-dependent manner. Clin Cancer Res. 2009;15:1558–65.