BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 11 2011, pages 1521–1528
doi:10.1093/bioinformatics/btr154

Systems biology

Advance Access publication March 30, 2011

A novel network-based method for measuring the functional relationship between gene sets

Qianghu Wang[1,†], Jie Sun[1,†], Meng Zhou[1,†], Haixiu Yang[1], Yan Li[1], Xiang Li[1], Sali Lv[1], Xia Li[1,*] and Yixue Li[1,2]

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081 and [2]Bioinformatics Center, Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: In the functional genomic era, a large number of gene sets have been identified via high-throughput genomic and proteomic technologies. These gene sets of interest are often related to the same or similar disorders or phenotypes, and are commonly presented as differentially expressed gene lists, co-expressed gene modules, protein complexes or signaling pathways. However, biologists are still faced by the challenge of comparing gene sets and interpreting the functional relationships between gene sets into an understanding of the underlying biological mechanisms.
Results: We introduce a novel network-based method, designated corrected cumulative rank score (CCRS), which analyzes the functional communication and physical interaction between genes, and presents an easy-to-use web-based toolkit called GsNetCom to quantify the functional relationship between two gene sets. To evaluate the performance of our method in assessing the functional similarity between two gene sets, we analyzed the functional coherence of complexes in functional catalog and identified protein complexes in the same functional catalog. The results suggested that CCRS can offer a significant advance in addressing the functional relationship between different gene sets compared with several other available tools or algorithms with similar functionality. We also conducted the case study based on our method, and succeeded in prioritizing candidate leukemia-associated protein complexes and expanding the prioritization and analysis of cancer-related complexes to other cancer types. In addition, GsNetCom provides a new insight into the communication between gene modules, such as exploring gene sets from the perspective of well-annotated protein complexes.
Availability and Implementation: GsNetCom is a freely available web accessible toolkit at http://bioinfo.hrbmu.edu.cn/GsNetCom.
Contact: lixia@hrbmu.edu.cn
Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 20, 2010; revised on March 3, 2011; accepted on March 19, 2011

1 INTRODUCTION

Numerous systematic biological studies have revealed that cellular function in a biological system normally involves the participation and interaction of multiple genes (Subramanian *et al*., 2005; Zhang *et al*., 2006). In the functional genomic era, a large number of gene sets have been identified through the application of high-throughput genomic, proteomic technologies, such as microarray, mass spectrometry and CHIP-on-chip assays and next-generation sequencing technologies (Huang da *et al*., 2009b; Morozova and Marra, 2008). Gene sets of interest may be related to disorders or disease phenotypes and the biological interpretation of gene set data may therefore be of great importance.

During the past several years, an increasing number of computational tools has been developed and played an important role in helping biologists explore these gene sets of interest. Of the tools contributing to the functional analysis of gene sets, most are enrichment tools and as reviewed by Huang da *et al*. (2009a). A significant portion of these bioinformatic enrichment tools is based on Gene Ontology (GO) (Ashburner *et al*., 2000), and only allow users to submit a single gene set and identify over-represented GO terms in 'interesting' gene set compared with the background through statistical analysis. Examples of such tools include GOstat (Beissbarth and Speed, 2004), GO::TermFinder (Boyle *et al*., 2004) and GOEAST (Zheng and Wang, 2008). Recently, some new tools and improved versions of previous tools, which integrate diverse and heterogeneous data content (e.g. KEGG pathways, gene expression data) have been released for the comprehensive functional analysis of single gene set, examples of which include Gazer (Kim *et al*., 2007), GeneTrail (Backes *et al*., 2007), DAVID (Huang da *et al*., 2009) and GSEA (Subramanian *et al*., 2005).

Within a biological system, no gene set functions in an isolated manner, even in fairly complete pathways. One gene set is interconnected with other gene sets through complex mechanisms, and these relationships may affect related disorders and phenotypes. Therefore, a common challenge faced by experimental biologists is to gain a better understanding of the functional relationships between different gene sets. A few computational tools have been developed to compare gene sets, most of which are based on GO, such as FatiGO (Al-Shahrour *et al*., 2007) and ProfCom (Antonov *et al*., 2008). Some GO-based semantic similarity methods can also be used to compare two gene sets by averaging the pairwise distances between the elements (Resnik, 1999; Sevilla *et al*., 2005; Wang *et al*., 2007). Currently, most of the GO terms have been assigned a

[*]To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first Three authors should be regarded as joint First Authors.

highly skewed distribution (Kim *et al*., 2007; Pesquita *et al*., 2009). In essence, GO is a language-based annotation, and it is difficult to finely tune GO annotation terms for genes to reflect the actual complexity of biological functions and relationships (Kim *et al*., 2007; Pesquita *et al*., 2009). Another approach is to use literature keywords to compare gene sets, as described by Martini (Soldatos *et al*., 2010). However, these types of computational tools have been implemented based on the same idea of mapping biological knowledge on sets of genes using GO or literatures and only found GO terms or keywords that are significantly over-represented in one set of genes versus a second reference set to compare two gene sets, and did not provide functional similarity score between two gene sets provided by users.

Proteins generally do not function in isolation, but rather function as part of a molecular machine. Biological and cellular functions are performed in a modular and hierarchical fashion (Barabasi and Oltvai, 2004; Pinkert *et al*., 2010; Qi *et al*., 2008). Previous studies have shown that protein–protein interaction (PPI) networks can reflect functional communication among proteins (Jiang and Keating, 2005; Lin *et al*., 2004; Lu *et al*., 2005) and the closer the two proteins are in a network, the more similar their functions are likely to be (Sharan *et al*., 2007). From many perspectives, this information is more suitable as it is expected that the functional relationship between gene sets can be well exhibited (Antonov *et al*., 2008; Lubovac, 2009). In this study, we present a novel network-based method, the corrected cumulative rank score (CCRS), for understanding how gene sets communicate at the higher protein interaction network level. Based on the protein complexes database CORUM (the Comprehensive Resource of Mammalian protein complexes, http://mips.helmholtz-muenchen.de/genre/proj/corum), we evaluated the performance of the CCRS method. We presented two case studies to demonstrate that the CCRS method can offer a significant advance in addressing the functional relationships between different gene sets. GsNetCom is freely accessible at http://bioinfo.hrbmu.edu.cn/GsNetCom.

2 MATERIALS AND METHODS

2.1 Data sources

PPI data were obtained from the HPRD (Keshava Prasad *et al*., 2009), BioGRID (Stark *et al*., 2006), IntAct (Kerrien *et al*., 2007), MINT database (Ceol *et al*., 2010), DIP (Salwinski *et al*., 2004) and by the co-citation of text mining (Ramani *et al*., 2005). We derived a non-redundant human PPI network comprising 69 331 interactions between 11 305 proteins. The topological characteristics of PPI network were summarized in Table 1. To assess the performance of our method, protein complex data was used. Experimentally verified protein complexes from human were downloaded from the CORUM database (Ruepp *et al*., 2010) at the Munich Information Center for Protein Sequences (MIPS) (Mewes *et al*., 2008). The CORUM database provides a resource of manually annotated protein complexes from mammalian organisms (Ruepp *et al*., 2010). The CORUM dataset is available in two alternative versions, the core dataset and the complete dataset, for searching and downloading (Ruepp *et al*., 2010). The core dataset is a reduced dataset which is essentially free of redundant entries, whereas the complete dataset consists of all annotated protein complexes. The function of protein complexes in the CORUM database is annotated using the MIPS Functional Catalog (FunCat) (Ruepp *et al*., 2004, 2010). There are 1343 human protein complexes in the core dataset and 1828 human protein complexes in the complete dataset. All nine cancer gene expression profiles were downloaded from NCBI Gene Expression Omnibus. To exclude potential platform-related bias, we restricted data acquisition to Affymetrix HG-U133A and HGU-133 Plus 2.0 arrays for cancer gene expression profiles. A detailed description of all datasets can be found in Supplementary Table S1.

Table 1. The topological characteristics of PPI network

Database	Proteins	Interactions	Degree		Shortest path	
	No. (%)	No. (%)	Max	Average	Diameter	Average
HPRD	8776 (77.6)	33 823 (48.8)	254	7.7	12	5.1
IntAct	6562 (58.0)	21 029 (30.3)	446	6.4	12	4.2
BioGRID	6412 (56.7)	19 851 (28.6)	193	6.2	14	4.6
Cocit	5419 (47.9)	21 740 (31.4)	136	8.0	15	5.1
MINT	4566 (40.4)	10 319 (14.9)	285	4.5	12	4.5
DIP	977 (8.6)	1197 (1.7)	27	2.5	14	5.9
Intergrated	11 305 (100)	69 331 (100)	640	12.3	12	3.9

2.2 Statistical analysis

To obtain the statistical significance of CCRS, we performed gene set sampling analysis and made the empirical distribution (with simulation) of the CCRS. A large number of simulated gene set pairs of the same size as given gene sets pair was randomly sampled from all human genes and the CCRS values were recomputed for each random gene set pair. ScoreP denoted the functional similarity score between a given gene set pair, N was the number of the simulated gene set pairs and M denoted the number of sampled gene set pairs having an equal or larger CCRS value than ScoreP. The estimate of the empirical P-value was obtained as $P = M/N$. The empirical P-value based on such randomizations represented the probability of obtaining a score greater than a given score by chance.

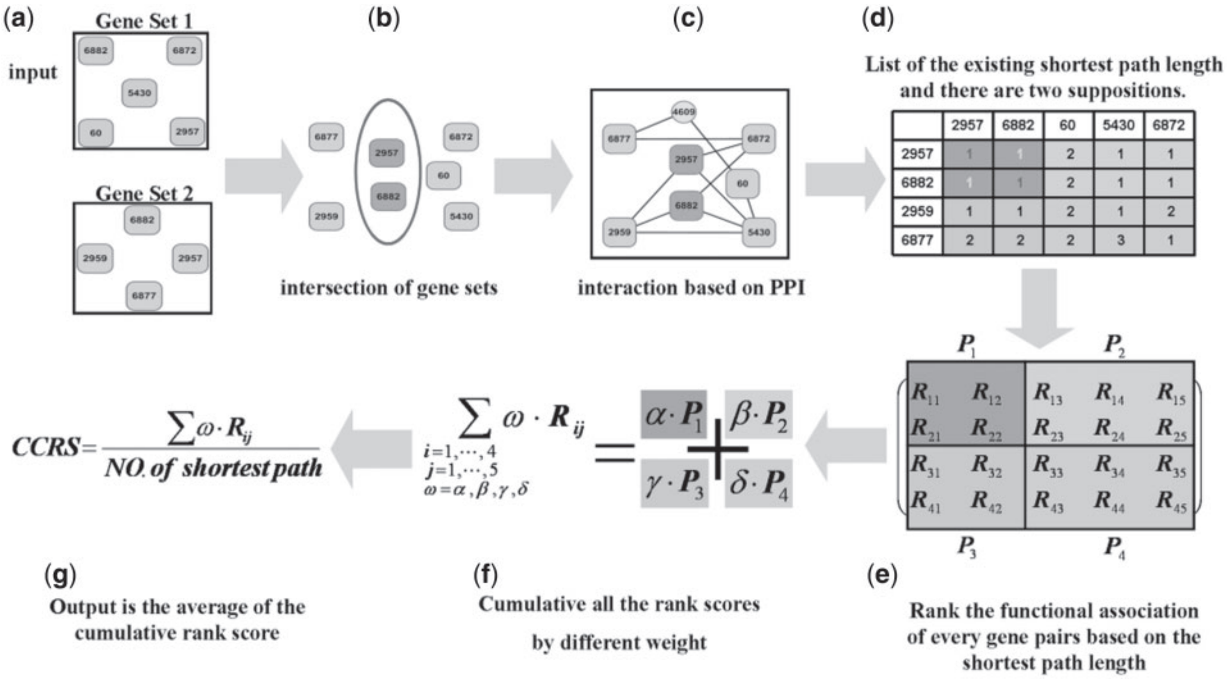
2.3 Input format of GsNetCom

In this study, we present an easy-to-use web-based toolkit called GsNetCom for assessing the functional similarity of two gene sets. This software enables a new insight into the communication between gene modules and allows for the exploration of gene sets from the perspective of well-annotated protein complexes based on the CCRS method and PPIs. GsNetCom requires text-formatted input of two lists of genes of interest. For the functional annotation of gene sets, one list of genes can be used as an input. GsNetCom supports many gene or gene product identifiers such as Gene Symbol (Sayers *et al*., 2009), Entrez Gene ID (Sayers *et al*., 2009), RefSeq Protein ID (Sayers *et al*., 2009), SwissProt/Uniprot and UniGene (Sayers *et al*., 2009).

3 RESULTS

3.1 Functional similarity measures (CCRS)

A schematic representation of the CCRS method is provided in Figure 1. This method measures the functional similarity between two gene sets considering the functional communication and physical interaction between these genes. We defined the 'function distance' between two genes as the shortest path length from one gene to another through the existing paths of the PPI network (the shorter the distance, the more similar the function between the two genes). So we are able to rank the functional similarity between two genes using the reciprocal of the 'function distance'. For comparing two gene sets, we cumulated the rank scores for every gene pair between the two sets. As we know, a gene set must be entirely functional consistent with itself. Thus when compared two identical gene sets, two assumptions were made. First, we assumed that there


(a)
Gene Set 1
input
6882
6872
5430
60
2957
Gene Set 2
6882
2959
2957
6877
(b)
6877
2957
6882
6872
60
2959
5430
intersection of gene sets
(c)
4609
6877
2957
6872
60
6882
2959
5430
interaction based on PPI
(d)
List of the existing shortest path length
and there are two suppositions.
(e)
Rank the functional association
of every gene pairs based on the
shortest path length
P_1
P_2
P_3
P_4
R_{11} R_{12} R_{13} R_{14} R_{15}
R_{21} R_{22} R_{23} R_{24} R_{25}
R_{31} R_{32} R_{33} R_{34} R_{35}
R_{41} R_{42} R_{43} R_{44} R_{45}
(f)
$\sum_{\substack{i=1,\cdots,4\\ j=1,\cdots,5\\ \omega=\alpha,\beta,\gamma,\delta}} \omega\cdot R_{ij} = \alpha\cdot P_1 + \beta\cdot P_2 + \gamma\cdot P_3 + \delta\cdot P_4$
Cumulative all the rank scores
by different weight
(g)
$CCRS=\frac{\sum\omega\cdot R_{ij}}{NO.\ of\ shortest\ path}$
Output is the average of the
cumulative rank score

	2957	6882	60	5430	6872
2957	1	1	2	1	1
6882	1	1	2	1	1
2959	1	1	2	1	2
6877	2	2	2	3	1

Fig. 1. Diagram of the CCRS method. (**a**) Two gene sets were used as the input datasets. (**b**) Extracting the intersection of two gene sets. (**c**) The interactions between two gene sets. (**d**) A list of all existing shortest path lengths between Gene Set 1 and Gene Set 2. (**e**) Based on the shortest path length, we calculated the functional association of every gene pairs and formed a block matrix. (**f**) All the rank scores by different weights on the four sub-blocks were cumulated. (**g**) Calculating CCRS as the functional relationship measurement between two gene sets.

is a direct interaction between a gene and itself. Second, if there is a path between two genes which belonged to the same gene set, we presumed that the 'function distance' is one regardless of the length of the path. Based on these assumptions, we were able to correct the value of the cumulative rank score between two gene sets when the gene sets overlapped.

In this section, we demonstrated how to compute the CCRS of two gene sets, G_1 and G_2. The intersection of G_1 and G_2 is denoted by G. There are n paths among the m nodes of the intersection G, based on the PPI network, for two genes p and q that are taken from G_1 and G_2, respectively.

$$\mathrm{CCRS}(G_1,G_2)=\frac{\sum\limits_{p\in G,q\in G}\alpha\cdot R_{pq}+\sum\limits_{p\in G,q\in G_1-G}\beta\cdot R_{pq}+\sum\limits_{p\in G,q\in G_2-G}\gamma\cdot R_{pq}+\sum\limits_{p\in G_1-G,q\in G_2-G}\delta\cdot R_{pq}}{N}$$

Where r_{pq} represents the 'function distance' (the length of the shortest path) between gene p and gene q, and R_{pq} is the rank score between them, $R_{pq}=1/r_{pq}$ or $R_{pq}=\exp(-r_{pq})$. N is the number of the existing 'function distance' values [including the 'function distance' of the overlapping nodes (m) from the node to itself]. When there is no path from gene p to gene q, we can define $r_{pq}=\infty$ and in this case, the rank score (R_{pq}) is zero. In our study, we set $\alpha=\beta=\gamma=\delta=1$and chose $R_{pq}=1/r_{pq}$. In the definition of the CCRS method, we learn that $\Sigma_{p,q}\in_G\alpha\bullet R_{pq}=\Sigma_{p,q}\in_G(1/r_{pq})=m+n$. Our algorithms not only used network topology information but also considered biological information. Here, we chose the reciprocal of harmonic mean as the score rather than the reciprocal of arithmetic mean to minimize the cost of edges between two proteins at greater distances (Krauthammer *et al.*, 2004; Ma *et al.*, 2007).

3.2 Analysis of the performance of the CCRS method

To evaluate the performance of our CCRS method in assessing the functional similarity between two gene sets, we conducted experiments on the functional similarity of human protein complexes. The protein complex data were derived from the CORUM database (Ruepp *et al.*, 2008). The CORUM database is a biological annotation resource for protein complexes based on the Functional Catalo annotation scheme, which contains 1343 human protein complexes in the core dataset and 1828 human protein complexes in the complete dataset (one protein complex will be regarded as one gene set). The CCRS functional similarity measure was analyzed in terms of the functional coherence of complexes in Functional Catalog and the identification of protein complexes which are in the same Functional Catalog.

First, we used the CCRS method to exploit the hierarchy of function for each complex and determine whether this method can offer a significant advance in evaluating the functional relationship between two gene sets. In our analysis, we used all 1343 core protein complexes in human available in the CORUM database and assembled them into 901 153 (C^2_{1343}) pairs of complexes. Specifically, we expect our functional similarity measure to exhibit relatively high CCRS values for the protein complex pairs that are annotated in the same functional categories, and low values for the pairs in different categories. Based on FunCat (MIPS Functional Catalogue), we organized the complex pairs into two

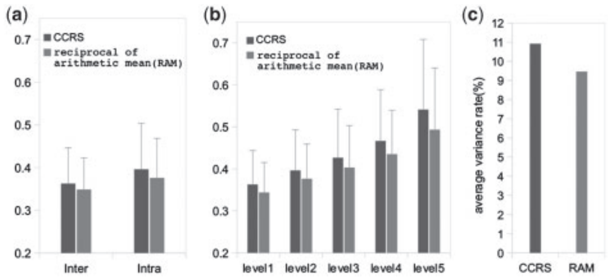


Fig. 2. Bar graph of the average CCRS scores derived from different categories and their subgroups compared to those derived from the RAM method. The error bars are taken from their SD.

groups: (i) intercategory pairs, when the two complexes in a complex pair were not annotated as belonging to a common catalog; (ii) intracategory pairs, when the two complexes in a complex pair were annotated as belonging to a common catalog. The number of complex pairs in these two groups was 431 494 and 469 659, respectively. The average of CCRS values for these two groups are shown in Figure 2A. Based on the CCRS method, the average CCRS value of the intercategory complex pairs was below that of the intra-category complex pairs. Furthermore, because every main functional category is organized as a hierarchical, tree-like structure, we sorted the 469 659 intra-category pairs according to the category level. For example, based on the FunCat Catalog, the annotations of complex 75 are FunCat 12.07 and 14.07.03 and the annotation of complex 81 is FunCat 14.07.05. The common functional category of complex 75 and complex 81 is FunCat 14.07. We classified this complex pair as category subgroup level2. According to the criteria described above, we obtained five category subgroups: level1, level2, level3, level4 and level5. Pairs classified into the most specific levels (level6 and above) were rare, so were grouped into category level5. Figure 2B shows the average CCRS values of the five subgroups. The more specific the Functional Catalog of the complex pairs, the higher the CCRS values.

Next, we performed above analysis by defining a score function as the reciprocal of arithmetic mean (RAM), and made a comparison with our CCRS. Based on RAM, the average similarity score of the intra-category pairs was higher than that of the intercategory pairs (Fig. 2A). The same was found to be true for the average similarity values of the five subgroups (Fig. 2B). To further assess the performance of the CCRS and RAM, we normalized the variance of the average scores of the subgroups (ASS). The average variance rate (AVR) was defined as follows:

$$\mathrm{AVR}=\sum_{i=1}^{4}\frac{\mathrm{ASS}(L_{i+1})-\mathrm{ASS}(L_i)}{4\mathrm{ASS}(L_i)}$$

where L_i is the level-i category. The AVR of the CCRS method was 10.92% and the AVR of the RAM was 9.47% (Fig. 2C), indicating that the CCRS method is more effective at distinguishing the functional hierarchy of complexes than RAM.

To further evaluate the performance of the CCRS method in measuring the functional relationship between two gene sets, we used a cross-validation method to examine how effectively the CCRS method can predict function-associated complexes. We randomly selected a fine Functional Catalog (FunCat 11.02.03.01.01) as the positive complex set. There are 13 human complexes in FunCat 11.02.03.01.01. Every complex in the positive set was selected as the target complex in turn. Then, we randomly extracted 87 protein complexes from the database to comprise the negative complex set. There were no common partners between the positive and negative complex sets. The complexes selected were all core complexes. Based on the results of analysis with the CCRS method, we determined the function-associated complexes (positive complexes) identified by the target complex. Receiver operating characteristic (ROC) curves were used to evaluate the sensitivity and specificity of identification of the function-associated complexes with the target complex. Sensitivity measures the proportion of actual positives complexes which are correctly identified, and specificity measures the proportion of negative complexes which are correctly identified.

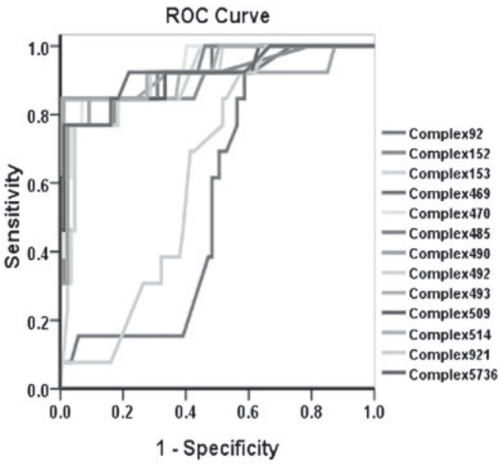


Fig. 3. The ROC curve showing the validation of identified function-associated complexes by the CCRS method.

Table 2. The AUC values for all the 13 cases

Target complex ID	AUC	Target complex ID	AUC
Complex92	0.565	Complex492	0.939
Complex152	0.886	Complex493	0.921
Complex153	0.898	Complex509	0.922
Complex469	0.939	Complex514	0.938
Complex470	0.935	Complex921	0.636
Complex485	0.939	Complex5736	0.922
Complex490	0.893		

Figure 3 shows the results of cross-validation via the ROC curves obtained by calculating the sensitivity [sensitivity = TP/(TP + FN)] and 1−specificity [specificity = FP/(TN + FP)] at different cutoffs, with ROC area under the curve (AUC) values between 0.886 and 0.940, except for complexes 92 and 921. The AUC values for all 13 cases are shown in Table 2. In the PPI network, we found that members of complexes 92 and 921 scarcely interacted with members of other complexes in FunCat 11.02.03.01.01 (Fig. 4). This finding may have been due to the sparsity of PPI data.

3.3 Comparisons with similar tools or algorithms

Finally, we also performed a comparison between the CCRS method and several other tools or algorithms of similar functionality, some

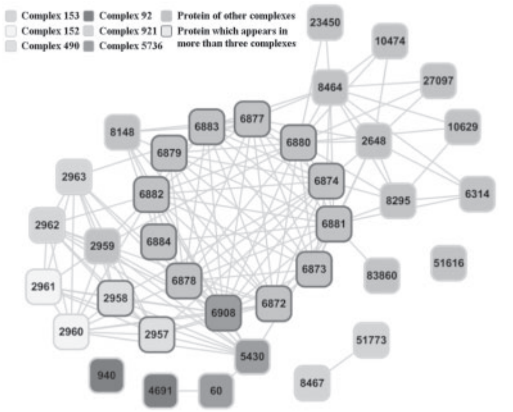


Fig. 4. The sub-network constructed by the core complexes subunits in FunCat 11.02.03.01.01.

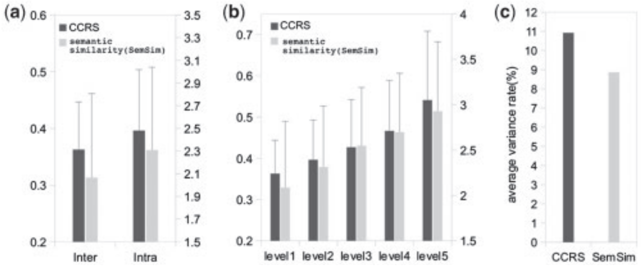


Fig. 5. Bar graph of the average CCRS scores derived from different groups and subgroups compared to those derived from the GO-based pairwise similarity measure method. The error bars are taken from their SD.

of which were based on GO, to examine GO annotations such as FatiGO and ProfCom. Another existed approach, such as Martini, is based on the use of literature keywords to compare gene sets. However, these existed tools found GO terms or literature keywords that are significantly over-represented in one set of genes versus a second reference set to reveal the difference between two gene sets rather than provide functional similarity measure between two gene sets provided by users. In comparison to these tools, the CCRS provides a quantitative analysis for exploring the functional relationship between two gene sets in the context of PPI networks and performs statistical tests on the analysis results. We also provide a comparison of the CCRS method with the GO-based pairwise similarity measure method which was used to compare two gene sets by averaging the pairwise distances between the elements using Resnik's algorithm which is the most commonly used measure for the functional prediction/validation of GO-based semantic similarity measures (Pesquita *et al*., 2009; Resnik, 1999; Sevilla *et al*., 2005; Wang *et al*., 2007). Using pairwise semantic similarity of GO terms, we computed the functional similarity score of 901 153 pairs of complexes. Then, we assembled the complex pairs into different groups according to the FunCat Catalog. The same classification was used. Based on pairwise semantic similarity measures, the average similarity score of the intra-category pairs was higher than that of the intercategory pairs (Fig. 5A). The same was found to be true for the average similarity values of the five subgroups (Fig. 5B). Again, we exploited the AVR to further assess the performance of the CCRS and GO-based pairwise semantic similarity measures.

The AVR of the CCRS method was 10.92% and the average AVR of the GO-based semantic similarity measures was 8.86% (Fig. 5C), indicating that the CCRS method is more effective at distinguishing the functional hierarchy of complexes than the GO-based semantic similarity measures. In summary, these results confirm that the CCRS method offers a significant advance in addressing the functional relationships between different gene sets.

3.4 Extended function: functional analysis of gene sets based on protein complex data

Bioinformatic enrichment tools based on diverse and heterogeneous data content (e.g. GO, KEGG pathways and gene expression data) have been widely used in the functional analysis of single gene set. Here, we show that functional annotation of gene sets based on protein complex data can be performed effectively using the GsNetCom software. When user inputs a query gene set, GsNetCom uses the CCRS method to analyze the functional associations between the query gene set and known complexes, and provide the user with a ranked complex list showing the potential function of the query gene set.

3.5 Case study

As an example, to illustrate the application of comparing gene sets using GsNetCom, we selected a known disease gene set (leukemia-related genes) as the input dataset and identified disease-related protein complexes using the CCRS method. The 141 leukemia-related genes were extracted from the Genetic Association Database (GAD) (Becker *et al*., 2004), an archive of human genetic association studies of complex diseases and disorders. We computed the functional similarity scores between the leukemia-related gene set and every protein complex using the CCRS method. The 10 protein complexes with the highest functional similarity scores with the leukemia gene set were listed and shown in Table 3. Among these 10 leukemia-related complexes, we found that complexes 2892 and 2895 had already been annotated as leukemia-related complexes in the CORUM database. Furthermore, complexes 2681 and 2679 had been reported as leukemia-related complexes in the literatures (Borellini and Glazer, 1993; Puil *et al*., 1994). The remaining six complexes are new candidate leukemia-related complexes.

Based on the above observations, we expanded the case of leukemia to other cancer types using the CCRS method for prioritizing cancer-related complexes. We applied the CCRS method to the differentially expressed gene sets for ovarian cancer, renal carcinoma, hepatocellular carcinoma, squamous cell lung carcinoma, prostate carcinoma, papillary thyroid cancer, breast carcinoma, urinary bladder cancer and colorectal adenoma, and found several common cancer-related complexes. The results seem to imply that there exist common mechanisms in the cancer biology.

We applied the significance analysis of microarray (SAM) method to identify differentially expressed genes between cancer samples and corresponding controls (Supplementary Table S2) (Tusher *et al*., 2001). All genes with a q-value <0.001 were considered as differentially expressed genes. Then we made the differentially expressed gene set as input gene set in turn and identified the cancer-related complexes using the CCRS method. We computed the functional similarity scores between the differentially expressed

Table 3. The top 10 complexes which are the most relevant with the disease gene set of leukemia

Complex name	CCRS	FunCat	*P*-value	References	Leukemia-related complexes
P53 homotetramer complex	0.5446	Transcriptional control DNA binding	< 0.001	(Johnson *et al.*, 2009; Zenz *et al.*, 2008)	*
BCR-ABL(p185 fusion protein)-GRB2 complex	0.5043	Tyrosine kinase G-protein-mediated signal transduction	< 0.001	(Faderl *et al.*, 1999; Puil *et al.*, 1994)	YES
P53-SP1 complex	0.5001	Transcriptional control DNA binding	< 0.001	(Borellini and Glazer, 1993)	*
Er-alpha-p53-hdm2 complex	0.4900	Enzymatic activity regulation / enzyme regulator	< 0.001	(Greiner *et al.*, 2003)	CANDIDATE
p300-MDM2-p53 protein complex	0.4898	Transcriptional control Enzymatic activity regulation / enzyme regulator	< 0.001	(Shima *et al.*, 2008)	CANDIDATE
BRCA1-cABL complex	0.4865	DNA repair DNA damage response	< 0.001	(Deutsch *et al.*, 2003)	CANDIDATE
SHC-GRB2 complex	0.4783	Enzyme-mediated signal transduction	< 0.001	(Faderl *et al.*, 1999; Puil *et al.*, 1994)	YES
BRCA1-SMAD3 complex	0.4745	DNA repair Cell cycle Transcriptional control TGF-beta-receptor signaling pathway	< 0.001	(Greiner *et al.*, 2003)	CANDIDATE
P53-BARD1-Ku70 complex	0.4686	Apoptosis (type I programmed cell death)	< 0.001	(Johnson *et al.*, 2009)	CANDIDATE
GRB2-SHP-2 complex, PDGF stimulated	0.4644	Transmembrane receptor protein tyrosine kinase signaling pathways	< 0.001	(Faderl *et al.*, 1999)	CANDIDATE

The protein complex has been annotated in CORUM database as leukemia-related complex and designated as 'YES'. The protein complex has been reported as leukemia-related complex in the literatures and designated as '*'.The protein complex which has been identified as candidate leukemia-related complex using CCRS method was designated as 'CANDIDATE'.

cancer gene set and every complex in CORUM and ranked the complexes in descending order of the scores. The top 20 protein complexes with every cancer gene set were listed and shown in Supplementary Table S3.

Out of these cancer-related complexes, six (16%) are shared among these nine cancer types: BCR-ABL (p185 fusion protein)-GRB2 complex (2892), EGFR-CBL-GRB2 complex (2542), SHC-GRB2 complex (2895), p53 homotetramer complex (2861), p53-SP1 complex (2679) and TRAF6 oligomer complex (2704). Based on FunCat, these complexes were involved in the cellular communication/signal transduction mechanism, transcription or regulation of metabolism and protein function. There are 19 (51%) cancer-related complexes which are at least involved in the biological process of five different cancer types.

Gene TP53 is a tumor suppressor gene, which is mutated in >50% of the human tumor. These mutations encode distinct isoforms of protein p53, which can regulate p53 transcriptional activity. This is consistent with our study that p53 homotetramer complex is the cancer-related complex with all the cancer types. Based on the pathways in cancer of KEGG, we found that BCR-ABL (p185 fusion protein)-GRB2 complex (2892), EGFR-CBL-GRB2 complex (2542) and SHC-GRB2 complex (2895) are the components of the ErbB signaling pathway, and these complexes indirectly affect the sustained angiogenesis, evading apoptosis and proliferation which are important features of cancer biological process. TRAF6 (TRAF6 oligomer complex) is a member of TRAF family, which has been implicated in the activation of these transcription factors by the tumor necrosis factor superfamily. The six cancer-related complexes mentioned above, are associated with all of nine cancer types. In our study, we also found two specific prostate cancer-related complexes: complex 5460 and complex 5464, which are involved in the prostate cancer pathway of KEGG. In conclusion, we found that most cancer-related complexes obtained by the CCRS method are shared by different cancer types. Our study revealed common network patterns in different cancer types.

3.6 Implementation

GsNetCom is a freely available web accessible toolkit which is implemented on a JavaEE framework and run on the Tomcat 6.0 container, so no software installation effort is required for the user. The request and response structure, based on the most commonly used web framework Struts2, can dispatch and handle a custom request friendly and quickly. All the logic data of GsNetCom is stored in MySQL 5 DBMS and the server-side is implemented in java 1.6 scripts. The GsNetCom system uses JGraphT 0.8.2 to implement its core analysis algorithm. This software is freely available to all users at http://bioinfo.hrbmu.edu.cn/GsNetCom.

The computational complexity of GsNetCom is $O(mn)$, where m, n is the size of query gene set 1 and gene set 2, respectively. We used Dijkstra's algorithm to find the shortest path between two different proteins in the integrated PPI network based on adjacent matrix. As we know, Dijkstra's algorithm runs in $O(V^2)$ for networks with V vertices. To make the efficiency of GsNetCom acceptable, we preprocessed integrated PPI network with Dijkstra's algorithm, and prestored all shortest paths between every two different proteins. So the computational complexity of GsNetCom is independent of the size of involved PPI network, and only in proportion to the size of query gene list.

4 DISCUSSION

Via the application of various technologies, biologists often identify gene sets of interest which may be involved with related disorders or phenotypes. However, comparing two gene sets and determining the functional relationships between them remains a challenging and daunting task. Several computational tools have been developed to compare gene sets. However, these tools generally found GO terms or keywords that are significantly over-represented in one set of genes versus a second reference set to compare two gene sets.

To remedy these problems, we developed a novel network-based method, designated CCRS, which takes into account the functional communication and physical interaction of genes, and presented an easy-to-use web-based toolkit called GsNetCom (http://bioinfo.hrbmu.edu.cn/GsNetCom/). The GsNetCom software quantifies the functional relationships between gene sets and performs statistical tests on the analysis results. Based on the protein complex database CORUM, we evaluated the performance of the CCRS method. In comparison to other available tools or algorithms, CCRS provides a significant advance in exploring the functional relationship between gene sets in the context of PPI network and performs statistical tests on the analysis results. Also, in the case study of cancer-associated protein complexes, CCRS successfully prioritized cancer-associated complexes. In addition, GsNetCom provides a new insight into the communication from gene modules, such as exploring gene sets from the perspective of well-annotated protein complexes. In its current version, the CCRS method computes the shortest path of any two genes between two gene sets as the 'function distance'. However, 'function distance' could more effectively be modeled using edge capacitance or optimization of edge constraints. In the near further we plan to improve GsNetCom by incorporating edge capacitance or optimization of edge constraints and some other biological information to comprehensively address the functional relationships between two gene sets. Furthermore, the annotation of gene sets could be improved by the supplementation of protein complex data.

Funding: The National Natural Science Foundation of China (Grant nos 30871394 and 61073136); the National High Tech Development Project of China; the 863 Program (Grant no. 2007AA02Z329); National Science Foundation of Heilongjiang Province (Grant no. QC2009C23); the Science Foundation of Educational Commission of Heilongjiang Province (Grant no. 11551233).

Conflict of Interest: none declared.

REFERENCES

Al-Shahrour,F. *et al*. (2007) FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Res.*, **35**, W91–W96.

Antonov,A.V. *et al*. (2008) ProfCom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data. *Nucleic Acids Res.*, **36**, W347–W351.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Backes,C. *et al*. (2007) GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Res.*, **35**, W186–W192.

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

Becker,K.G. *et al*. (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.

Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.

Borellini,F. and Glazer,R.I. (1993) Induction of Sp1-p53 DNA-binding heterocomplexes during granulocyte/macrophage colony-stimulating factor-dependent proliferation in human erythroleukemia cell line TF-1. *J. Biol. Chem.*, **268**, 7923–7928.

Boyle,E.I. *et al*. (2004) GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.

Ceol,A. *et al*. (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.

Deutsch,E. *et al*. (2003) Down-regulation of BRCA1 in BCR-ABL-expressing hematopoietic cells. *Blood*, **101**, 4583–4588.

Faderl,S. *et al*. (1999) The biology of chronic myeloid leukemia. *N. Engl. J. Med.*, **341**, 164–172.

Greiner,J. *et al*. (2003) Characterization of several leukemia-associated antigens inducing humoral immune responses in acute and chronic myeloid leukemia. *Int. J. Cancer*, **106**, 224–231.

Huang da,W. *et al*. (2009a) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Huang da,W. *et al*. (2009b) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Jiang,T. and Keating,A.E. (2005) AVID: an integrative framework for discovering functional relationships among proteins. *BMC Bioinformatics*, **6**, 136.

Johnson,G.G. *et al*. (2009) A novel type of p53 pathway dysfunction in chronic lymphocytic leukemia resulting from two interacting single nucleotide polymorphisms within the p21 gene. *Cancer Res.*, **69**, 5210–5217.

Kerrien,S. *et al*. (2007) IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.

Keshava Prasad,T.S. *et al*. (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

Kim,S.B. *et al*. (2007) GAzer: gene set analyzer. *Bioinformatics*, **23**, 1697–1699.

Krauthammer,M. *et al*. (2004) Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc. Natl Acad. Sci. USA*, **101**, 15148–15153.

Lin,N. *et al*. (2004) Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, **5**, 154.

Lu,L.J. *et al*. (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.*, **15**, 945–953.

Lubovac,Z. (2009) Investigating topological and functional features of multimodular proteins. *J. Biomed. Biotechnol.*, **2009**, 472415.

Ma,X. *et al*. (2007) CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics*, **23**, 215–221.

Mewes,H.W. *et al*. (2008) MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Res.*, **36**, D196–D201.

Morozova,O. and Marra,M.A. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics*, **92**, 255–264.

Pesquita,C. *et al*. (2009) Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.*, **5**, e1000443.

Pinkert,S. *et al*. (2010) Protein interaction networks—more than mere modules. *PLoS Comput. Biol.*, **6**, e1000659.

Puil,L. *et al*. (1994) Bcr-Abl oncoproteins bind directly to activators of the Ras signalling pathway. *EMBO J.*, **13**, 764–773.

Qi,Y. *et al*. (2008) Protein complex identification by supervised graph local clustering. *Bioinformatics*, **24**, i250–i258.

Ramani,A.K. *et al*. (2005) Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.*, **6**, R40.

Resnik (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artificial Intell. Res.*, **11**, 93–130.

Ruepp,A. *et al*. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.*, **32**, 5539–5545.

Ruepp,A. *et al*. (2008) CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.*, **36**, D646–D650.

Ruepp,A. *et al*. (2010) CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res.*, **38**, D497–D501.

Salwinski,L. *et al*. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.

Sayers,E.W. *et al*. (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.

Sevilla,J.L. *et al*. (2005) Correlation between gene expression and GO semantic similarity. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **2**, 330–338.

Sharan,R. *et al*. (2007) Network-based prediction of protein function. *Mol. Syst. Biol.*, **3**, 88.

Shima,Y. *et al*. (2008) PML activates transcription by protecting HIPK2 and p300 from SCFFbx3-mediated degradation. *Mol. Cell. Biol.*, **28**, 7126–7138.

Soldatos,T.G. *et al*. (2010) Martini: using literature keywords to compare gene sets. *Nucleic Acids Res*., **38**, 26–38.

Stark,C. *et al*. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Wang,J.Z. *et al*. (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**, 1274–1281.

Zenz,T. *et al*. (2008) Chronic lymphocytic leukemia and treatment resistance in cancer: the role of the p53 pathway. *Cell Cycle*, **7**, 3810–3814.

Zhang,P. *et al*. (2006) Gene functional similarity search tool (GFSST). *BMC Bioinformatics*, **7**, 135.

Zheng,Q. and Wang,X.J. (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res.*, **36**, W358–W363.