Deng *et al. BMC Systems Biology* 2013, **7**(Suppl 2):S8
http://www.biomedcentral.com/1752-0509/7/S2/S8

BMC
Systems Biology

RESEARCH Open Access

# ppiPre: predicting protein-protein interactions by combining heterogeneous features

Yue Deng[1,2], Lin Gao[1*], Bingbo Wang[1]

*From* The 6th International Conference on Computational Systems Biology (ISB2012)
Xi'an, China. 18-20 August 2012

## Abstract

**Background:** Protein-protein interactions (PPIs) are crucial in cellular processes. Since the current biological experimental techniques are time-consuming and expensive, and the results suffer from the problems of incompleteness and noise, developing computational methods and software tools to predict PPIs is necessary. Although several approaches have been proposed, the species supported are often limited and additional data like homologous interactions in other species, protein sequence and protein expression are often required. And predictive abilities of different features for different kinds of PPI data have not been studied.

**Results:** In this paper, we propose ppiPre, an open-source framework for PPI analysis and prediction using a combination of heterogeneous features including three GO-based semantic similarities, one KEGG-based co-pathway similarity and three topology-based similarities. It supports up to twenty species. Only the original PPI data and gold-standard PPI data are required from users. The experiments on binary and co-complex gold-standard yeast PPI data sets show that there exist big differences among the predictive abilities of different features on different kinds of PPI data sets. And the prediction performance on the two data sets shows that ppiPre is capable of handling PPI data in different kinds and sizes. ppiPre is implemented in the R language and is freely available on the CRAN (http://cran.r-project.org/web/packages/ppiPre/).

**Conclusions:** We applied our framework to both binary and co-complex gold-standard PPI data sets. The detailed analysis on three GO aspects suggests that different GO aspects should be used on different kinds of data sets, and that combining all the three aspects of GO often gets the best result. The analysis also shows that using only features based solely on the topology of the PPI network can get a very good result when predicting the co-complex PPI data. ppiPre provides useful functions for analysing PPI data and can be used to predict PPIs for multiple species.

## Background

Although different experimental methods [1,2] have already generated a large amount of PPI for many model species in recent years [3], these existing PPI data are incomplete and contain many false positive interactions. In order to refine these PPI data, computational approaches are urgently needed.

Some recent researches have shown that PPIs can be integrated with other kinds of biological data in using supervised learning to predict PPIs [4-7]. In supervised learning, a classifier is trained using truly interacting protein pairs (positive samples) and protein pairs which are not interacting with each other (negative samples). Then the trained classifier is able to recover false negative interactions and remove false positive interactions from the PPIs input by users.

Existing studies are mainly differing in the selection of features used in the prediction framework. In these studies, different biological evidences are extracted and used as features training the classifier, including Gene Ontology (GO) functional annotations [8,9], protein sequences [10] and co-expressed proteins [11]. For the organisms or

* Correspondence: lgao@mail.xidian.edu.cn
[1]School of Computer Science and Technology, Xidian University, Xi'an 710071, PR China
Full list of author information is available at the end of the article

BioMed Central

© 2013 Deng et al.; licensee BioMed Central Ltd. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

proteins which are lack of research, biological features may don't work well, so features based on network topology are also needed to integrate [12-14].

Although some frameworks and tools have also been proposed for predicting PPIs [15-20], they have two disadvantages in general. First, most of the frameworks only support a few well studied model organisms. Second, these frameworks often need users to provide additional biological data along with the PPIs. Moreover, different species often require different features, which make these existing frameworks not very convenient to use.

In this paper, we describe ppiPre, an open-source framework for the PPI prediction problem. The framework is implemented in the R language so it can work together with other R packages dealing with biological data and network [21], which is different from other tools accessed via web services. ppiPre integrates features extracted from multiple heterogeneous data sources, including GO [22], KEGG [23] and topology of the PPI network. Users don't need to provide additional biological data other than gold-standard PPI data. ppiPre provides functions for measuring the similarity between proteins and for predicting PPIs from the existing PPI data.

## Methods

Heterogeneous features are integrated in the prediction framework of ppiPre, including three GO-based semantic similarities, one KEGG-based similarity indicating the proteins are involved in the same pathways and three topology-based similarities using only the network structure of the PPI network.

We chose these three features to be integrated in our framework because they are highly available for the PPIs of different species and can be easily accessed in the R environment. Not like other methods and software tools, ppiPre did not integrate biological features that may not be available for the species or proteins which are not well studied, such as structural and domain information.

### GO-based semantic similarities

Proteins are annotated by GO with terms from three aspects: biological process (BP), molecular function (MF), and cellular component (CC). Directed acyclic graphs (DAGs) are used to describe these aspects. It is known that interacting protein pairs are likely to be involved in similar biological processes or in similar cellular component compared to those non-interacting proteins [2][24][25]. Thus if two proteins are semantically similar based on GO annotation, the probability that they actually interact is higher than two proteins that are less similar.

Several similarity measures have been developed for evaluating the semantic similarity between two GO terms [26-28]. The information content (IC) of GO terms and the structure of the GO DAG are often used in these measures.

The IC of a term $t$ can be defined as follows:

$$IC(t) = -\log(p(t)) \quad (1)$$

where p(t) is the probability of occurrence of the term $t$ in a certain GO aspect. Two IC-based semantic similarity measures proposed recently are integrated in ppiPre, which are Topological Clustering Semantic Similarity (TCSS) [29] and IntelliGO [30].

#### *TCSS*

In TCSS, the GO DAGs are divided into subgraphs. A PPI is scored higher if the two proteins are in the same subgraph. The algorithm is made up of two major steps.

In the first step, a threshold on the ICs of all terms is used to generate multiple subgraphs. The roots of the subgraphs are the terms which are below the previously defined threshold. If roots of two subgraphs have similar IC values, these two subgraphs are merged. Overlapping subgraphs may occur because some GO terms have more than one parent terms. In order to remove overlap between subgraphs, edge removal and term duplication are processed. Transitive reduction of GO DAG is used to remove overlapping edges by generating the smallest graph that has the same transitive closure as the original subgraph. After edge removal, if a term is included in two or more subgraphs, it will be duplicated into each subgraph. More details are described in [29].

After the first step, a meta-graph is constructed by connecting all subgraphs. Then the second step called normalized scoring is processed. For two GO terms, normalized semantic similarity is calculated based on the meta-graph rather than the whole GO DAG so that more balanced semantic similarity scores can be obtained.

Using the frequency of proteins that are annotated to GO term $t$ and its children, the information content of annotation (ICA) for a GO term $t$ is:

$$ICA(t) = -\ln\left(\frac{\left|P_t \bigcup_{c\in N(t)} P_c\right|}{\sum_{t\in O}\left|P_t \bigcup_{c\in N(t)} P_c\right|}\right) \quad (2)$$

where $P_t$ is the proteins that are annotated by $t$ in aspect $O$ and $N(t)$ is the child terms of $t$.

The information content of subgraph (ICS) for term $t_m^s$ in the $m^{th}$ subgraph $G_m^s$ is defined as follows:

$$ICS(t_m^s) = \frac{ICA(t_m^s)}{\max_{t_m^s\in G_m^s} ICA(t_m^s)} \quad (3)$$

The information content of meta-graph (ICM) for a term $t_n^m$ in meta-graph $G^m$ is defined as follows:

$$ICM\left(t_n^m\right) = \frac{ICA\left(t_n^m\right)}{\max\limits_{t_n^m \in G^m} ICA\left(t_n^m\right)} \tag{4}$$

Finally, the similarity between two proteins $i$ and $j$ is defined as:

$$Sim_{TCSS}(i,j) = \max_{s_m,t_n \in T_i,T_j} \begin{cases} ICM_{\max}\left(LCA\left(s_m,t_n\right)\right) \ \textit{if } s_m \in G_m^s \textit{ and } t_n \in G_n^s \\ ICS_{\max}\left(LCA\left(s_m,t_n\right)\right) \ \textit{if } s_m,t_n \in G_n^s \end{cases} \tag{5}$$

where $LCA(s_m,t_n)$ is the common ancestor of the terms $s_m$ and $t_n$ with the highest IC. $T_i$ and $T_j$ are two sets of GO terms which annotate the two proteins $i$ and $j$ respectively.

### *IntelliGO*

The IntelliGO similarity measure introduces a novel annotation vector space model. The coefficients of each GO term in the vector space consider complementary properties. The IC of a specific GO term and its evidence code (EC) [31] are used to assign this GO term to a protein. The coefficient $\alpha_t$ given to term $t$ is defined as follows:

$$\alpha_t = w\left(g, t\right) * IAF\left(t\right) \tag{6}$$

where $w(g, t)$ is the weight of the EC which indicates the annotation origin between protein $g$ and GO term $t$, and IAF (Inverse Annotation Frequency) represents the frequency of term $t$ occurred in all the proteins annotated in the aspect where $t$ belongs.

For two proteins $i$ and $j$, the IntelliGO uses their vectorial representation $\vec{i}$ and $\vec{j}$ to measure their similarity, which is defined as follows:

$$Sim_{IntelliGO}\left(i,j\right) = \frac{\vec{i} * \vec{j}}{\sqrt{\vec{i} * \vec{i}} * \sqrt{\vec{j} * \vec{j}}} \tag{7}$$

The detailed explanation of the definition can be found in [30].

### *Wang's method*

The similarity measure proposed by Wang [32] is also implemented in the ppiPre package, which is based on the graph structure of GO DAG.

In the GO DAG, each edge has a type which is "is-a" or "part-of". In Wang's measure, a weight is given to each edge according to its type. $DAG_t = (t,T_t,E_t)$ represents the subgraph made up of term $t$ and its ancestors, where $T_t$ is the set of the ancestor terms of $t$ and $E_t$ is the set of edges in $DAG_t$.

In $DAG_t$, $S_t(n)$ measures the semantic contribution of term $n$ to term $t$, which is defined as:

$$\begin{cases} S_t\left(t\right) = 1 \\ S_t\left(n\right) = \max\left\{w_e{}^*S_t\left(n'\right) \middle| n' \in \textit{childrenof}\left(n\right)\right\} \text{ if } t \neq n \end{cases} \tag{8}$$

The similarity between two GO term $m$ and term $n$ is defined as:

$$Sim_{Wang}\left(m, n\right) = \frac{\sum\limits_{t \in T_m \cap T_n} S_m\left(t\right) + S_n\left(t\right)}{SV\left(m\right) + SV\left(n\right)} \tag{9}$$

where $SV(m)$ is the sum of the semantic contribution of all the terms in $DAG_m$.

The semantic similarity between two proteins $i$ and $j$ is defined as the maximum value of all the similarity between any term that annotate $i$ and any term that annotate $j$.

## KEGG-based similarity

Proteins that work together in the same KEGG pathway are likely to interact[33][34]. The KEGG-based similarity between proteins $i$ and $j$ is calculated using the co-pathway membership information in KEGG. The similarity is defined as:

$$Sim_{KEGG}\left(i,j\right) = \frac{\left|P\left(i\right) \cap P\left(j\right)\right|}{\left|P\left(i\right) \cup P\left(j\right)\right|} \tag{10}$$

where $P(i)$ is the set of pathways which protein $i$ involved in the KEGG database.

## Topology-based similarities

In order to deal with the proteins that haven't got any annotations in GO or KEGG database, topology-based similarity measures are also integrated. In ppiPre, three different topological similarities are implemented.

The Jaccard similarity [35] between two proteins $i$ and $j$ is defined as:

$$Sim_{Jac}\left(i,j\right) = \frac{\left|N\left(i\right) \cap N\left(j\right)\right|}{\left|N\left(i\right) \cup N\left(j\right)\right|} \tag{11}$$

where $N(i)$ is set of all the direct neighbours of protein $i$ in PPI network.

Adamic-Adar(AA) similarity [36] punishes the proteins with high degree by assigning more weights to the nodes with low degree in PPI network. The AA similarity between two proteins $i$ and $j$ is defined as:

$$Sim_{AA}\left(i,j\right) = \sum_{n \in N(i) \cap N(j)} \frac{1}{\log k_n} \tag{12}$$

where $k_n$ is the degree of protein $n$.

Resource Allocation (RA) similarity [37] is similar to AA similarity and considers the common neighbours of two nodes as resource transmitters. The RA similarity between two proteins $i$ and $j$ is defined as:

$$Sim_{RA}\left(i,j\right) = \sum_{n \in N(x) \cap N(y)} \frac{1}{k_n} \tag{13}$$

### Prediction framework

The data of interacting protein pairs verified by experiments are very incomplete and the non-interacting protein pairs far outnumber interacting protein pairs. So the classical SVM [38] which is able to handle small and unbalanced data is chosen to integrate different features in ppiPre. We have tested different kernels in e1071 and the results showed no significant difference, so the default kernel and parameters are used in ppiPre.

The prediction framework of ppiPre is presented in Figure 1. Heterogeneous features are calculated for the gold-standard PPI data set which is given by users, and the SVM classifier is trained by the gold-standard positive and negative data set (solid arrows). After the classifier is trained, the features are calculated from the query PPIs input by users, and the trained classifier can predict false positive and false negative PPIs from the input data (hollow arrows).

## Results and discussion

Since all the features are calculated within the package, users don't need to provide additional biological data for different species. When users use ppiPre to predict the PPIs, they only need to provide both the gold-standard positive and negative training set and the test set. In this paper, we test the performance of ppiPre in yeast using two yeast gold-standard positive data sets which are a high quality binary data set provided by Yu's research [39] and the MIPS data set [40]. Self-interactions and duplicate interactions were removed previously. The detail of the two gold-standard data sets is shown in Table 1.

Non-interacting pairs were randomly selected from the proteins in gold-standard positive data sets as gold-standard negative data sets. The positive and negative data sets are set to the same size. In order to minimize the impact to the topological characteristics of the PPI network, the degree of each protein was maintained.

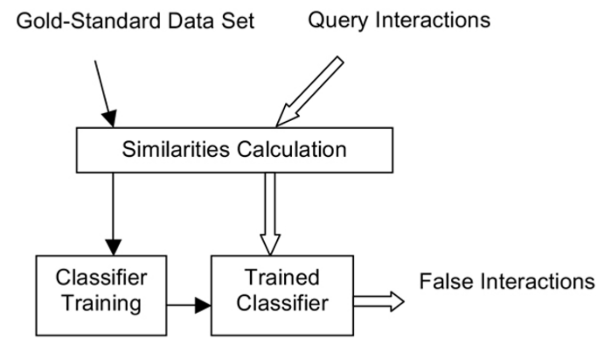

**Figure 1 Graphical overview of the prediction framework**. First, SVM is trained using the gold-standard PPI data sets (solid arrows). Then the trained classifier can be used to predict PPIs from the input PPI data (hollow arrows).

**Table 1 Gold-standard positive yeast protein interaction data sets**

| Data set | Number of Interactions | Number of Proteins | Interaction Type |
|---|---|---|---|
| Yu | 1263 | 1078 | binary |
| MIPS | 8250 | 871 | co-complex |

10-fold cross validation was used to evaluate the performance of the prediction framework.

### Predictive abilities of GO-based similarities

First, the predictive abilities of the three aspects of GO on different data sets were evaluated. We analysed the prediction performance using only one of the BP, MF and CC aspects. The receiver operating characteristic (ROC) curves are shown in Figure 2 and Figure 3. In order to assess these results quantitatively, the area under the ROC curve (AUC) of each ROC curve was calculated. The result is shown in Table 2.

For the binary data set, the BP aspect shows the best performance among all three aspects in ROC analysis of three GO-based semantic similarities (Figure 2, Table 2). This result is expected. The BP aspect is related to protein interaction and thus can be used to predict them.

For the co-complex data set, the CC aspect shows the best performance in ROC analysis of three GO-based semantic similarities (Figure 3, Table 2). Since the MIPS data set is composed of protein complexes, and a protein complex can only be formed if its proteins are localized within the same compartment of the cell, terms in the CC aspect correctly reflect the functional grouping of proteins in these complexes.

We then analysed the prediction performance using a combination of GO aspects. The ROC curves of a combination of two aspects are shown in Figure 4 and Figure 5. The ROC curves of combination three aspects are shown in Figure 6. The AUCs of the ROC curves are shown in Table 3. The results show that by combing more than one GO aspect, our method could get a better prediction performance than using a single aspect for both binary data set and co-complex data set. And the overall best performance was achieved by combing all the three GO aspects. So it is necessary to integrate all the three GO aspects in the prediction framework.

### Predictive abilities of KEGG-based and topological similarities

Then, the predictive abilities of KEGG-based similarity and three topological similarities were evaluated. For binary and co-complex data sets, the performance of KEGG-based similarity shows no big difference (Figure 7, Table 4). On the contrary, three topological similarities work perfectly for co-complex data set, but show only modest effects for binary data set. This is because the

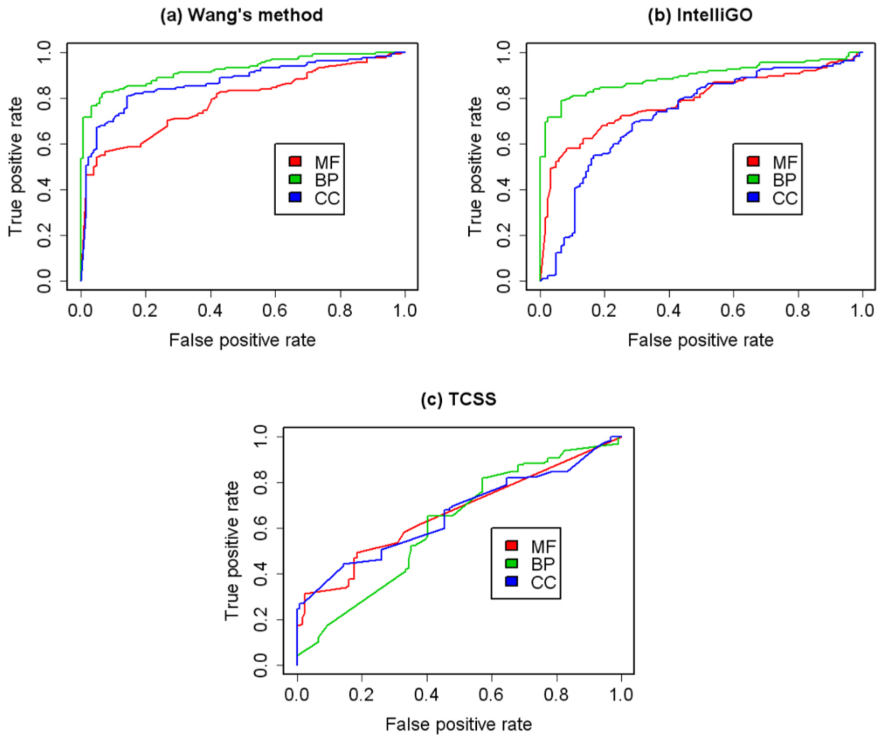

**Figure 2 ROC curves for binary data set using single GO aspect**. ROC evaluations of three GO aspects with three semantic similarity measures on the binary PPI data set are shown. The evaluation was performed using only one GO aspect at a time. BP shows the overall best predictive abilities in three aspects in GO.

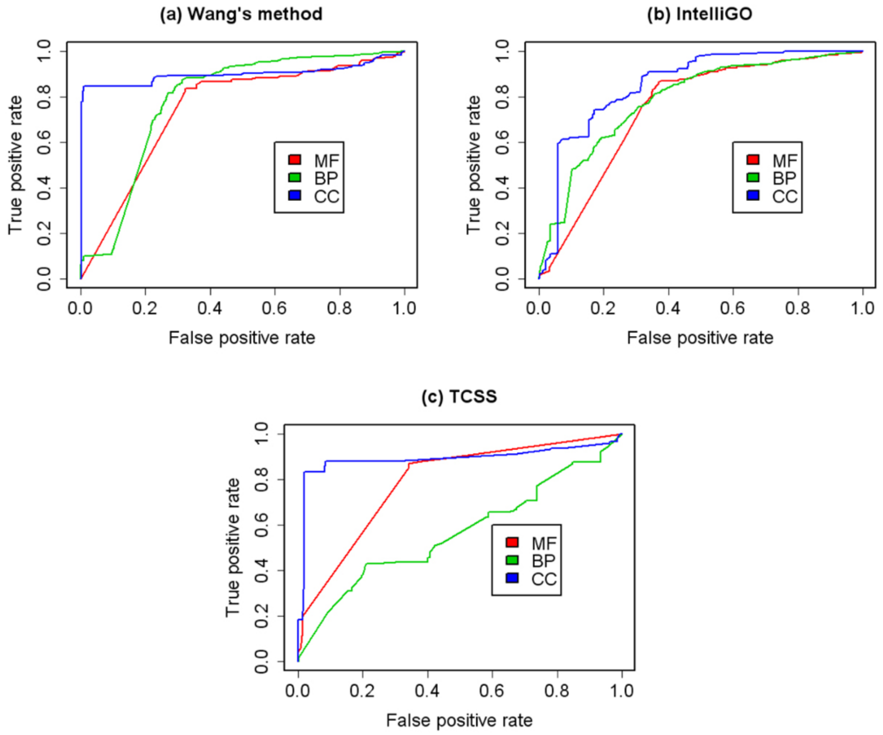

**Figure 3 ROC curves for co-complex data set using single GO aspect**. ROC evaluations of three GO aspects with three semantic similarity measures on the MIPS co-complex data set are shown. The evaluation was performed using only one GO aspect at a time. CC shows the overall best predictive abilities in three aspects of GO.

**Table 2 AUC for the yeast gold-standard PPI data sets using single GO aspect**

| | Binary data set | | | Co-complex data set | | |
|---|---|---|---|---|---|---|
| | **BP** | **MF** | **CC** | **BP** | **MF** | **CC** |
| Wang | **0.9246** | 0.7867 | 0.8696 | 0.7875 | 0.7482 | **0.8994** |
| IntelliGO | 0.8932 | 0.7842 | 0.7283 | 0.7882 | 0.7477 | 0.8551 |
| TCSS | 0.6178 | 0.6659 | 0.6628 | 0.5646 | 0.7891 | 0.8896 |

Tests were performed separately for biological process (BP), molecular function (MF) and cellular component (CC) ontologies in two data sets. We define similarity between two proteins as the maximum similarity found between any two GO terms that annotate them. The best ROC scores for each data set are in bold.

MIPS co-complex data set is composed of multi-protein complexes, and the interacting pairs are all in the same complex. The co-complex data set represents several unconnected subgraphs in the corresponding PPI network, meaning that two proteins from different complexes had no common neighbours in the PPI network. So the topological similarities of two proteins from two different complexes are zero while topological similarities of two proteins from the same complexes are not.

### Integration of biological and topological similarities

After analysing biological and topological features separately, we integrated these heterogeneous features together.

The ROC curves of two kinds of PPI data sets using GO-based, KEGG-based and topological similarities are shown in Figure 8. The AUC of binary and co-complex PPI data sets are 0.958 and 0.999.

The result shows that integrating biological and topological similarities can improve the prediction performance. So, it’s necessary to integrate heterogeneous features together when dealing with the PPI prediction problem. All the features are integrated in ppiPre.

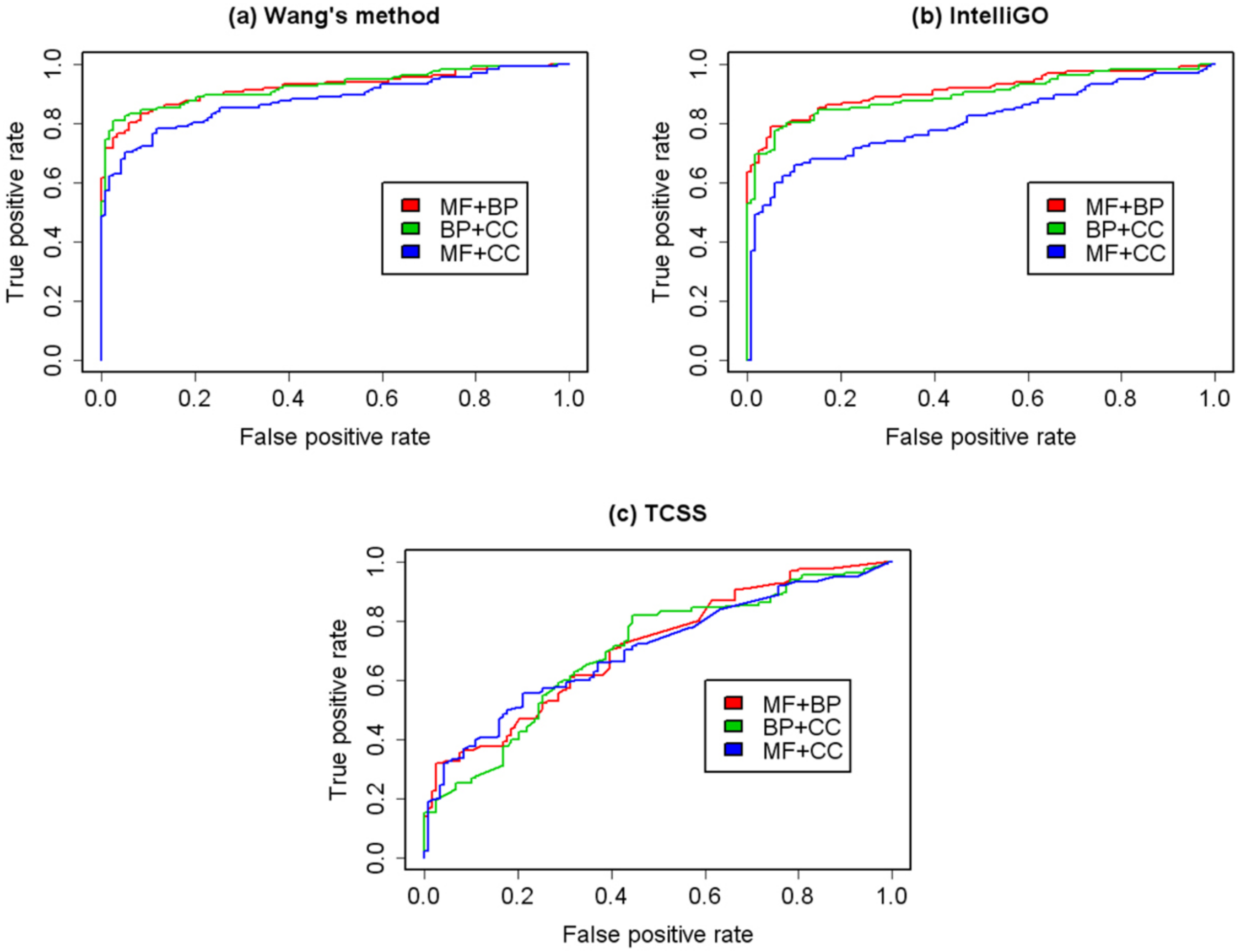

**Figure 4 ROC curves for binary data set using two GO aspects**. ROC evaluations of the combination of two GO aspects with three semantic similarity measures on the binary PPI data set are shown. The evaluation was performed using two of the three GO aspects at a time. In general, the prediction performance is better than that using one aspect.

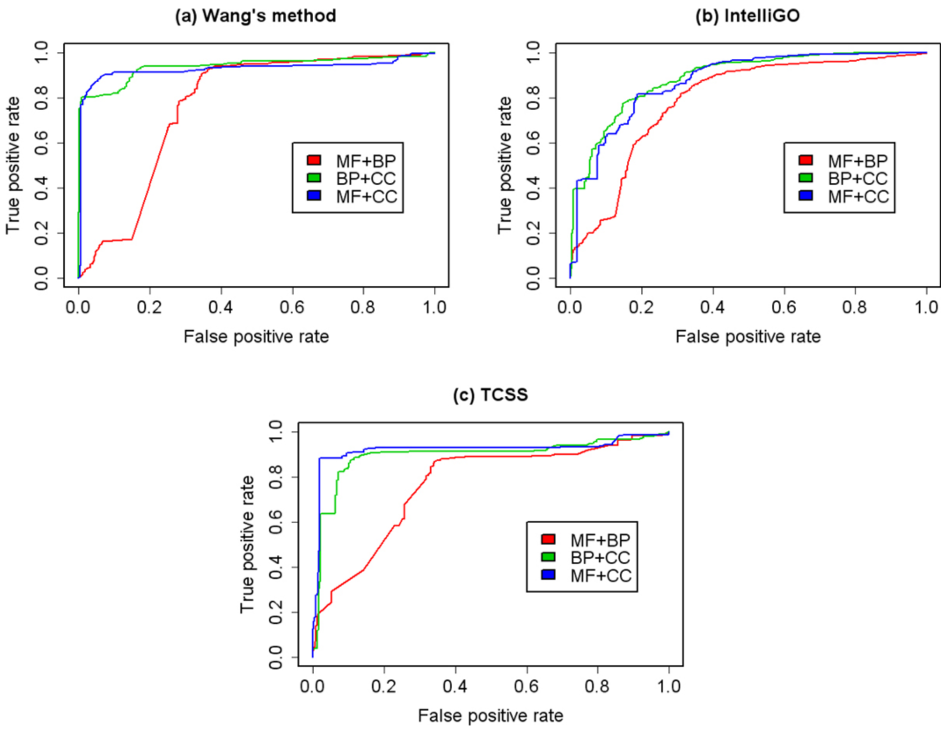

**Figure 5 ROC curves for co-complex data set using two GO aspects**. ROC evaluations of the combination of two GO aspects with three semantic similarity measures on the MIPS co-complex PPI data set are shown. The evaluation was performed using two of the three GO aspects at a time. In general, the prediction performance is better than that using one aspect.

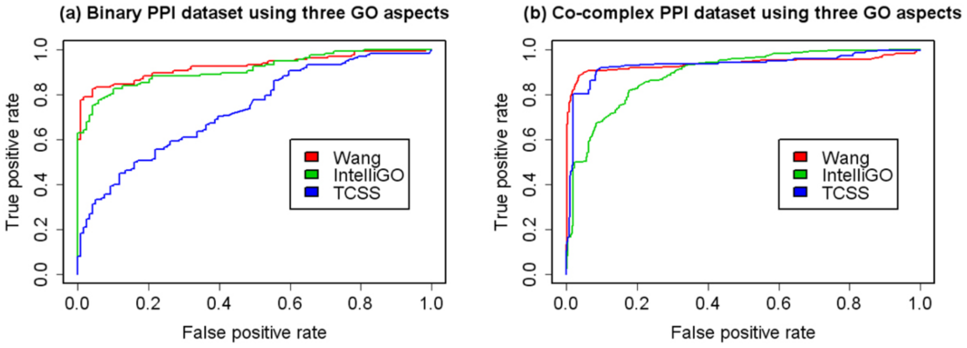

**Figure 6 ROC curves using three GO aspects**. ROC evaluations of the combination of all the three GO aspects with three semantic similarity measures on the binary and co-complex PPI data set are shown. In general, the prediction performance is better than that using one or two aspects.

**Table 3 AUC for the yeast gold-standard PPI data sets using a combination of GO aspects**

| | Binary data set | | | | Co-complex data set | | | |
|---|---|---|---|---|---|---|---|---|
| MF | √ | √ | | √ | √ | √ | | √ |
| BP | √ | | √ | √ | √ | | √ | √ |
| CC | | √ | √ | √ | | √ | √ | √ |
| Wang | 0.924 | 0.880. | 0.926 | **0.927** | 0.768 | 0.929 | **0.940** | 0.938 |
| IntelliGO | 0.912 | 0.804 | 0.899 | 0.914 | 0.792 | 0.877 | 0.890 | 0.895 |
| TCSS | 0.712 | 0.702 | 0.699 | 0.735 | 0.768 | 0.923 | 0.897 | 0.934 |

Tests were performed using a combination of three GO aspects. We define simi-larity between two proteins as the maximum similarity found between any two GO terms that annotate them. The best ROC scores for each data set are in bold.

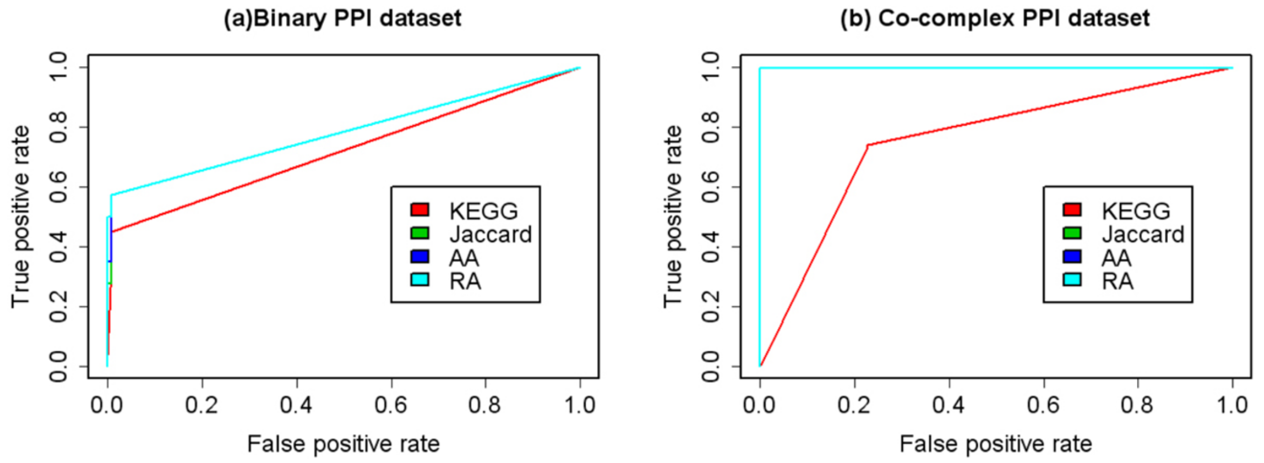

**Figure 7 ROC curves using KEGG-based and topological features**. ROC evaluations of the KEGG-based similarity (KEGG), Jaccard similarity (Jaccard), Adamic-Adar similarity (AA) and Resource Allocation similarity (RA) on the binary and co-complex PPI data sets are shown. The result shows that topological similarities work very well for the co-complex data set.

**Table 4 AUC for the yeast gold-standard PPI data sets using KEGG-based and different topological similarities**

| | KEGG | Jaccard | AA | RA |
|---|---|---|---|---|
| Binary data set | 0.7201 | 0.7819 | 0.7825 | **0.7838** |
| Co-complex data set | 0.7558 | **0.9988** | **0.9988** | **0.9988** |

Tests were performed separately for KEGG-based similarity (KEGG), Jaccard similarity (Jaccard), Adamic-Adar similarity (AA) and Resource Allocation similarity (RA) in two data sets. The best ROC scores for each data set are in bold.

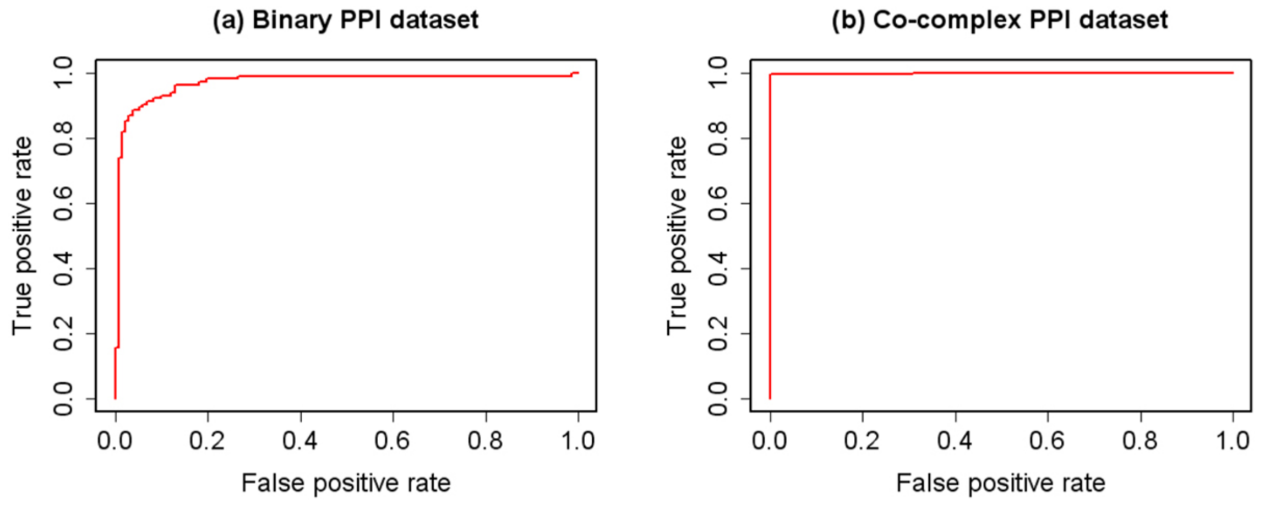

**Figure 8 ROC curves using a combination of GO-based, KEGG-based and topological features**. ROC evaluations of the integration of GO-based, KEGG-based and topological similarity measures on the binary and co-complex PPI data sets are shown. The result shows that integrating heterogeneous features can improve the prediction performance.

**Table 5 Functions provided in ppiPre**

| Name | Description |
|---|---|
| AASim | Computes the Adamic-Adar similarity |
| ComputeAllEvidences | Computes biological and topological similarities |
| FNPre | Predict false negative interactions using topological similarities |
| GOKEGGSims | Computes KEGG-based similarity and GO-based similarities |
| IntelliGOGeneSim | Computes IntelliGO semantic similarity |
| JaccardSim | Computes the Jaccard similarity |
| KEGGSim | Computes KEGG-based similarity |
| RASim | Computes the Resource Allocation similarity |
| SVMPredict | Trains the SVM classifier, and then predict false interactions |
| TCSSGeneSim | Computes TCSS semantic similarity |
| TopologicSims | Computes all the three topological similarities |
| WangGeneSim | Computes Wang's semantic similarity |

For proteins with unknown annotations in GO and KEGG, the GO-based and KEGG-based similarity measures cannot work. But the impact on these two data sets can be ignored since interactions without annotations are only 2 in the binary data set (0.19%) and 16 in MIPS data set (1.84%). However, when ppiPre is used on a large amount of proteins that are poorly annotated in GO, users should consider that the performance of ppiPre may be hampered under such situation.

### Implementation and usage

The current version of ppiPre supports 20 species. The detail of the species supported and IC data used in GO-based semantic similarities are described in [41]. The annotation data of GO and KEGG are got from the packages GO.db and KEGG.db.

ppiPre has been submitted to CRAN (Comprehensive R Archive Network) and can be installed and loaded easily in the R environment. ppiPre provides functions for calculating similarities and predicting PPIs. A summary of the functions available is shown in Table 5. Detailed descriptions and examples for all the functions are contained in the manual provided within ppiPre.

## Conclusions

An open-source framework ppiPre for PPI prediction is proposed in this paper. Several heterogeneous features are combined in ppiPre, including three GO-based similarities, one KEGG-based similarity and three topology-based similarities. To make the prediction, users don't need to provide additional biological data other than gold-standard PPI data.

ppiPre can be integrated into existing bioinformatics analysis pipelines in the R environment. Other features will be evaluated and integrated in future work, and the framework will be tested on PPI data of more species especially those poorly annotated in GO.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

LG conceived of the study, supervised the project and revised the manuscript. YD implemented the framework, performed the experiments, and drafted the manuscript. BW participated in the data analysis. All authors read and approved the final manuscript.

**Acknowledgements**
We thank the anonymous reviewers for many helpful suggestions and comments. We thank Dr. Yong Wang of AMSS, CAS for insightful discussions. We also thank Rongjie Shao and Gang Wang of Xidian University for technical support. A preliminary version of this paper was published in the proceedings of IEEE ISB2012 [41].

**Declarations**
The publication of this article was funded by the NSFC (91130006, 60933009, 61202174) and the Fundamental Research Funds for the Central Universities (K5051223003, K5051223005).
This article has been published as part of *BMC Systems Biology* Volume 7 Supplement 2, 2013: Selected articles from The 6th International Conference of Computational Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcsystbiol/supplements/7/S2.

**Authors' details**
[1]School of Computer Science and Technology, Xidian University, Xi'an 710071, PR China. [2]Institute of Software Engineering, Xidian University, Xi'an 710071, PR China.

Published: 14 October 2013

### References

1. Gavin A-C, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon A-M, Cruciat C-M, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier M-A, Copley RR, Edelmann A, Querfurth E, Rybin V, *et al*: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415**:141-147.
2. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.** *Nature* 2000, **403**:623-627.
3. De Las Rivas J, Fontanillo C: **Protein-Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks.** *PLoS Comput Biol* 2010, **6**:e1000807.

4. Ben-Hur A, Noble WS: **Kernel methods for predicting protein-protein interactions.** *Bioinformatics* 2005, **21**:i38-46.
5. Chen X-W, Liu M: **Prediction of protein-protein interactions using random decision forest framework.** *Bioinformatics* 2005, **21**:4394-4400.
6. Patil A, Nakamura H: **Filtering high-throughput protein-protein interaction data using a combination of genomic features.** *BMC Bioinformatics* 2005, **6**:100.
7. Lin X, Liu M, Chen X: **Assessing reliability of protein-protein interactions by integrative analysis of data in model organisms.** *BMC Bioinformatics* 2009, **10(Suppl 4)**:S5.
8. Mahdavi M, Lin Y-H: **False positive reduction in protein-protein interaction predictions using gene ontology annotations.** *BMC Bioinformatics* 2007, **8**:262.
9. Kuchaiev O, Rašajski M, Higham DJ, Pržulj N: **Geometric De-noising of Protein-Protein Interaction Networks.** *PLoS Comput Biol* 2009, **5**.
10. Wang C, Cheng J, Su S: **Prediction of Interacting Protein Pairs from Sequence Using a Bayesian Method.** *The Protein Journal* 2009, **28**:111-115.
11. Qi Y, Klein-Seetharaman J, Bar-Joseph Z: **A mixture of feature experts approach for protein-protein interaction prediction.** *BMC Bioinformatics* 2007, **8(Suppl 10)**:S6.
12. Lü L, Zhou T: **Link prediction in complex networks: A survey.** *Physica A: Statistical Mechanics and its Applications* 2011, **390**:1150-1170.
13. Guimerà R, Sales-Pardo M: **Missing and spurious interactions and the reconstruction of complex networks.** *Proceedings of the National Academy of Sciences* 2009, **106**:22073-22078.
14. Chua HN, Ning K, Sung W-K, Leong HW, Wong L: **Using indirect protein-protein interactions for protein complex prediction.** *J Bioinform Comput Biol* 2008, **6**:435-466.
15. Kim S, Shin S-Y, Lee I-H, Kim S-J, Sriram R, Zhang B-T: **PIE: an online prediction system for protein-protein interactions from text.** *Nucleic Acids Research* 2008, **36(Web Server)**:W411-W415.
16. Guo Y, Li M, Pu X, Li G, Guang X, Xiong W, Li J: **PRED_PPI: a server for predicting protein-protein interactions based on sequence data with probability assignment.** *BMC Research Notes* 2010, **3**:145.
17. Li D, Liu W, Liu Z, Wang J, Liu Q, Zhu Y, He F: **PRINCESS, a Protein Interaction Confidence Evaluation System with Multiple Data Sources.** *Mol Cell Proteomics* 2008, **7**:1043-1052.
18. Michaut M, Kerrien S, Montecchi-Palazzi L, Chauvat F, Cassier-Chauvat C, Aude J-C, Legrain P, Hermjakob H: **InteroPORC: Automated Inference of Highly Conserved Protein Interaction Networks.** *Bioinformatics* 2008, **24**:1625-1631.
19. Pitre S, Dehne F, Chan A, Cheetham J, Duong A, Emili A, Gebbia M, Greenblatt J, Jessulat M, Krogan N, Luo X, Golshani A: **PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs.** *BMC Bioinformatics* 2006, **7**:365.
20. McDowall MD, Scott MS, Barton GJ: **PIPs: human protein-protein interaction prediction database.** *Nucleic Acids Research* 2009, **37(Database)**:D651-D656.
21. Csárdi G, Nepusz T: **The igraph software package for complex network research.** *InterJournal Complex Systems* 2006, **1695**.
22. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene Ontology: tool for the unification of biology.** *Nat Genet* 2000, **25**:25-29.
23. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Research* 2000, **28**:27-30.
24. Lehner B, Fraser AG: **A first-draft human protein-interaction map.** *Genome Biology* 2004, **5**:R63.
25. Jansen R: **A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data.** *Science* 2003, **302**:449-453.
26. Resnik P: **Using Information Content to Evaluate Semantic Similarity in a Taxonomy.** *IJCAI* 1995, 448-453.
27. Jiang J, Conrath D: **Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy.** *International Conference Research on Computational Linguistics (ROCLING X)* 1997, 9008.
28. Lord PW, Stevens RD, Brass A, Goble CA: **Semantic similarity measures as tools for exploring the gene ontology.** *Pac Symp Biocomput* 2003, 601-612.
29. Jain S, Bader G: **An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology.** *BMC Bioinformatics* 2010, **11**:562.
30. Benabderrahmane S, Smail-Tabbone M, Poch O, Napoli A, Devignes M-D: **IntelliGO: a new vector-based semantic similarity measure including annotation origin.** *BMC Bioinformatics* 2010, **11**:588.
31. Rogers MF, Ben-Hur A: **The use of gene ontology evidence codes in preventing classifier assessment bias.** *Bioinformatics* 2009, **25**:1173-1177.
32. Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F: **A new method to measure the semantic similarity of GO terms.** *Bioinformatics* 2007, **23**:1274-1281.
33. Qi Y, Bar-Joseph Z, Klein-Seetharaman J: **Evaluation of different biological data and computational classification methods for use in protein interaction prediction.** *Proteins* 2006, **63**:490-500.
34. van Noort V, Snel B, Huynen MA: **Exploration of the omics evidence landscape: adding qualitative labels to predicted protein-protein interactions.** *Genome Biology* 2007, **8**:R197.
35. Jaccard P: **Étude comparative de la distribution florale dans une portion des Alpes et des Jura.** *Bull Soc Vaud Sci Nat* 1901, **37**:541.
36. Adamic LA, Adar E: **Friends and neighbors on the Web.** *Social Networks* 2003, **25**:211-230.
37. Zhou T, Lü L, Zhang Y-C: **Predicting missing links via local information.** *The European Physical Journal B - Condensed Matter and Complex Systems* 2009, **71**:623-630.
38. Vapnik VN: *The Nature of Statistical Learning Theory* Springer; 2000.
39. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual J-F, Dricot A, Vazquez A, Murray RR, Simon C, Tardivo L, Tam S, Svrzikapa N, Fan C, de Smet A-S, Motyl A, Hudson ME, Park J, Xin X, Cusick ME, Moore T, Boone C, Snyder M, Roth FP, *et al*: **High-Quality Binary Protein Interaction Map of the Yeast Interactome Network.** *Science* 2008, **322**:104-110.
40. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han J-DJ, Bertin N, Chung S, Vidal M, Gerstein M: **Annotation Transfer Between Genomes: Protein-Protein Interologs and Protein-DNA Regulogs.** *Genome Research* 2004, **14**:1107-1118.
41. Deng Y, Gao L: **ppiPre - an R package for predicting protein-protein interactions.** *2012 IEEE 6th International Conference on Systems Biology (ISB)* 2012, 333-337.

doi:10.1186/1752-0509-7-S2-S8
**Cite this article as:** Deng *et al.*: ppiPre: predicting protein-protein interactions by combining heterogeneous features. *BMC Systems Biology* 2013 **7**(Suppl 2):S8.

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

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

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central