Supplement 2
Results Supplement – Computational Results

1. Performance comparison – AUC scores

Figure S2.1 compares the partial AUC scores [0] across all four classifiers for predicting the human receptor interactome. As can be seen, the random forest [4] method performs best on all partial AUC scores criteria. (For definitions of AUC scores, see Supplement S1.)

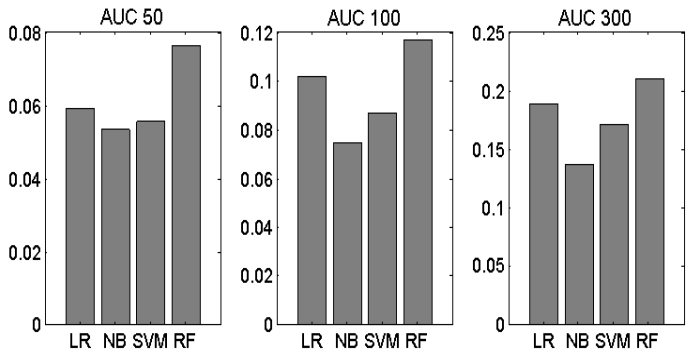

Figure S2.1 Performance comparison of human receptor protein interaction prediction task using partial AUC scores. Four classifiers are compared: Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression (LR) and Random Forest (RF).

There are several reasons that have contributed to the success of the RF [4] method when compared with the other three classifiers:

- The currently available direct and indirect protein interaction data is inherently noisy and contains many missing values. The randomization and ensemble strategies within RF make it more robust to noise when compared to other classification methods.
- Biological datasets are often correlated with each other and thus should not be treated as independent sources. Linear and non-linear regression models assume independence and may therefore perform worse than other classifiers in tasks where correlations among features are strong. In contrast, the RF classifier does not make any assumptions about the relationship between the data, which makes it more appropriate for the type of data available for the protein interaction prediction task.
- It is also important for the method to consider the feature correlation and missing value problems together. If a pair has values for one redundant feature but not the other, RF can still use this feature for the prediction process.

2. Performance comparison between the 'Receptor' protein interaction prediction task and the 'General' human protein interaction prediction task

In addition to comparing different methods for predicting the receptor interactome we also compared these methods with the more general task of predicting human protein interactions for all proteins (regardless of whether they are receptors). The experimental setup and the train-test size are the same as described in Supplement S1. The difference lies in the examples of the training data in the two cases. In the 'receptor protein interaction prediction task', the training pairs all relate to receptors. In contrast, for the 'general case', the training pairs are general protein interaction data from HPRD [2] and random negative pairs from all proteins excluding known interactions.

In Figure S2.2, black bars represent AUC scores for the 'receptor protein interaction prediction task' task while white bars describe the "general PPI" task training. As can be seen, the receptor specific training does improve the performance compared to using the more general human interaction data for training.

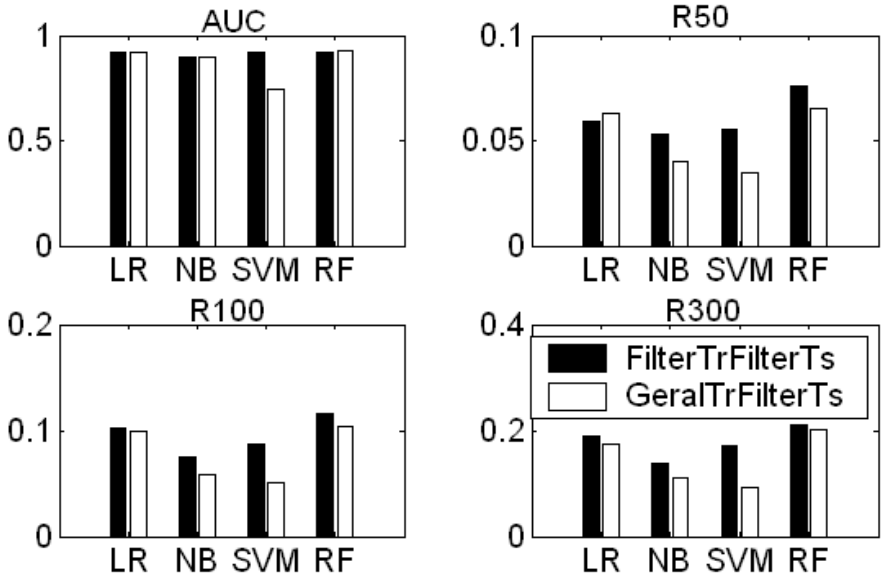

Figure S2.2 Performance comparison between 'receptor related only training' and the 'general human PPI training'. For 'receptor only' all training pairs contain at least one receptor protein whereas for the 'general case', the training pairs are any type of direct protein pairs from HPRD. The testing dataset is the same for both, i.e. every pair contains at least one receptor. In most cases the 'receptor related only' training (black bars) does improve the performance compared to using general human interaction data (white bars) for training.

3. Performance comparison between two settings of gold standard negatives

We used a random set of receptor-protein pairs excluding all known HPRD pairs as a negative training set. The drawback of the random negative set is that random proteins may be very easily distinguishable from interacting proteins simply because of their different functions. This may result in low performance of the classifier because it does not learn the fine distinctions between functionally related but not interacting proteins. We therefore also synthesized negatives from lists of random receptor-protein pairs (not in HPRD) having similar molecular functions (or similar cellular locations). The motivation is that proteins that are functionally related, but do not necessarily interact, might represent the negatives of the physical binding relationships better.

The following two figures give the AUC score comparison between the random setting and the random co-functional setting for negative sets. In Figure S2.3 we use the random receptor-protein pairs with the same cellular localizations (defined by Gene Ontology Slim version) for the co-functional setting. In Figure S2.4 biological functions (also reported in the main text) is used to synthesize the co-functional setting. Figure 2.4 shows that the two settings for gold standard negatives achieve comparable performance, with slightly less prediction power when using the co-functional gold standard negative (defined by the gene ontology biological function annotations) as compared to the random gold standard negatives. Similar conclusions could be drawn from Figure 2.3 as well when using the localization evidence to build the random co-functional setting.

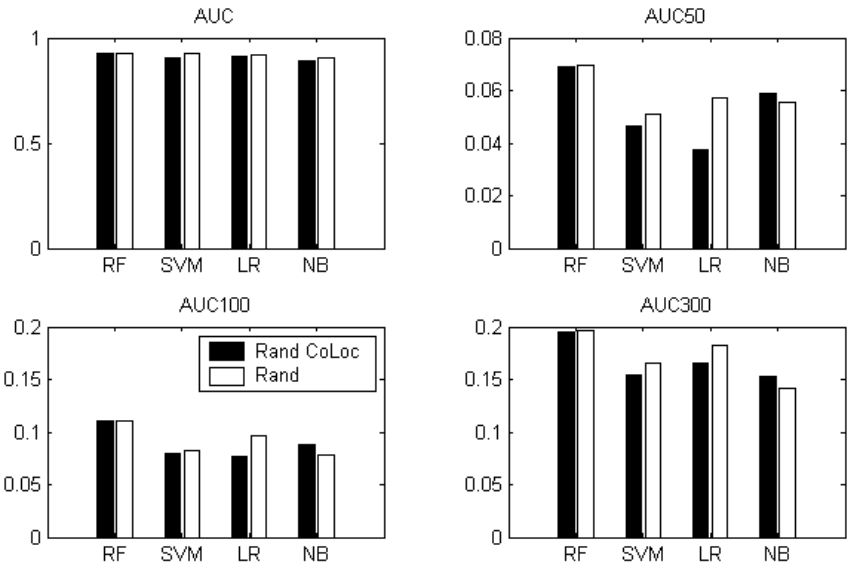

Figure S2.3 Performance comparison between two settings of gold standard negatives. For 'Rand CoLoc' (black bars) negative pairs use the random receptor-protein pairs with similar locations (defined by GO slim). For 'Rand' (white bars), negative pairs use the random receptor-protein pairs that are not in HPRD. The testing dataset is the same for both, i.e. every pair contains at least one receptor. In most cases the 'Rand CoLoc' training (black bars) does not improve the performance compared to using the random strategy (white bars) for training.

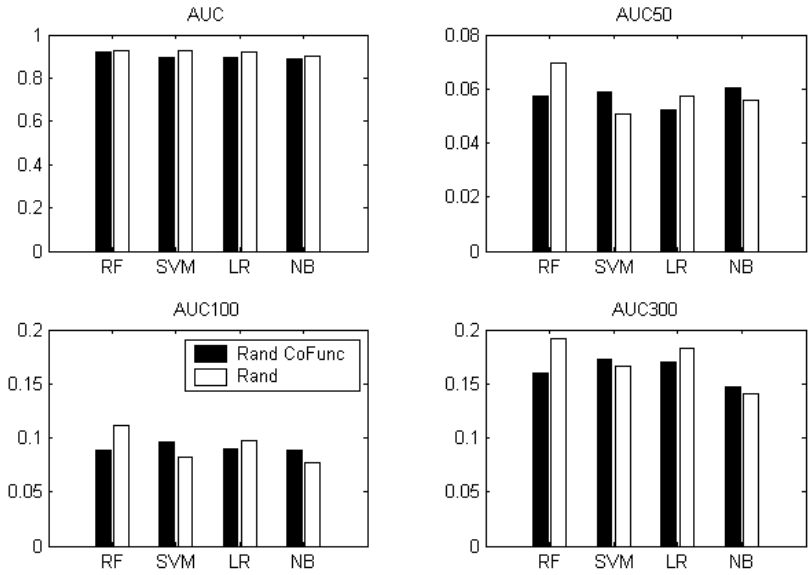

Figure S2.4 Performance comparison between two settings of gold standard negatives. For 'Rand CoFunc' (black bars) negative pairs use the random receptor-protein pairs with similar functions (defined by GO slim). For 'Rand' (white bars), negative pairs use the random receptor-protein pairs that are not in HPRD. The testing dataset is the same for both, i.e. every pair contains at least one receptor. In most cases the 'Rand CoFunc' training (black bars) does not improve the performance compared to using the random strategy (white bars) for training.

4. Performance comparison between two settings of gold standard positive with homology concerns

To address the concern that proteins with homology to receptor-related pairs might cause bias in the training as well as over-estimation of the performance of the method, we investigated the effect of homology using the ERBB receptors as a case study. ERBB receptors include EGFR / ERBB2 / ERBB3 / ERBB4. They are in sequence similar to each other.

We want to evaluate the prediction performance of EGFR interaction pairs. Thus we tried two kinds of gold standard training as following:

- Case I: In the training positive, we have all the known interaction pairs related to ERBB2 / ERBB3 / ERBB4. Also this positive set have a random sample (half) of the EGFR related known interaction.
- Case II: In the training positive, we have no known interaction pairs related to ERBB2 / ERBB3 / ERBB4. Also this positive set have the same random sample (half) of the EGFR related known interaction as above.

Then we trained two RF models with the above two different gold standard positive setting. Applying the two models on all possible pairs between EGFR and human proteins (excluding those pair used in the training), we could measure the resulting pair

list with precision-recall curves (also called "prediction accuracy vs. sensitivity" in main text). From the following Figure S2.6, we could see that the status of ERBB2-4 related pairs in the training or not does not affect the predictions of EGFR interaction pairs much, especially for the low recall regions (which is our primary targeted region).

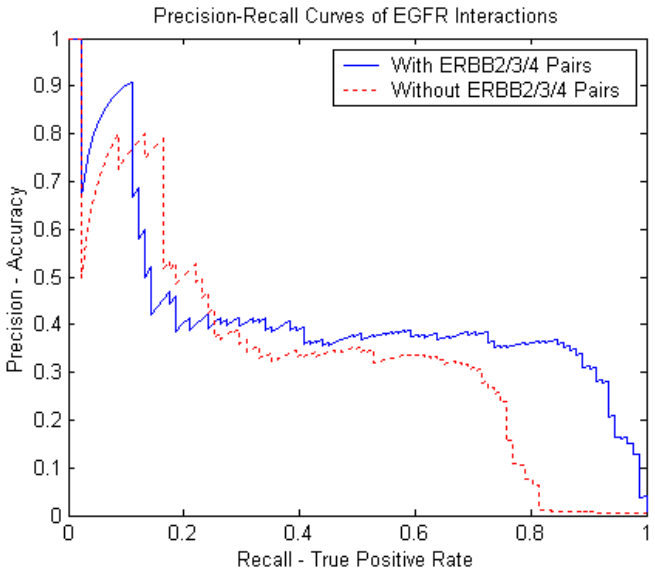

Figure S2.5 Performance comparison between two settings of gold standard positive. Precision vs. Recall curves from the two cases of gold standard positive related to ERBB receptors.

5. Feature importance

Biologically, it is of particular interest to identify the extent to which heterogeneous data sources carry information about protein interactions. This can help analyze what data source is most useful for determining interactions.

One way to determine such feature importance is to use the resulting RF trees [4, 5]. The RF classifier uses a splitting function called the *Gini* index to determine which attribute to split on during the tree learning phase. The *Gini* index measures the level of impurity / inequality of the samples assigned to a node based on a split at its parent. In our case, where there are two classes, let p represent the fraction of interacting pairs assigned to node m and $1\text{-}p$ the fraction of the non interacting pairs. Then, the Gini index at node m is defined as:

$$G_m = 2p(1-p).$$

The purer a node is, the smaller the Gini value. Every time a split of a node is made using a certain feature attribute, the Gini value for the two descendant nodes is less than the parent node. The sum of these Gini value decreases (from parent to sons) for each feature over all trees in the forest provides a simple and reliable estimate of the feature importance for this prediction task. The RF Gini feature importance selector is generally a popular metric used in a variety of feature selection tasks [5].

Table S2.1 lists RF Gini variable importance for the eight feature groups used in our Human receptor PPI prediction task. The sequence alignment is ranked as the highest feature. Among the top ten Gini ranked features, five of them are similarities of gene expressions. The other four features ranked in the top ten include domain-domain interaction features, the homology derived features from yeast, tissue positions and the biological process from GO. The sequence alignment and the homology based feature are extremely important, which (encouragingly) corresponds to current practice among experimentalists.

Table S2.1 RF Gini variable importance using the normalized Gini importance scores of the eight feature sources.

#	Feature Name	Gini score (normalized)	Top Ten Ranked
1	GO Function	0.0058	
2	GO Component	0.0086	
3	GO Process	0.0407	8
4	Co-Tissue	0.0635	5
5	Gene Expression	16 feature columns Max Gini: 0.0977 Min Gini: 0.0099	2: GeneExp2 3: GeneExp5 4: GeneExp10 6: GeneExp9 9: GeneExp3
6	BlastP E-value	0.1163	1
7	Homology PPI Yeast	5 feature columns Max Gini: 0.0544 Min Gini: 0.0041	7: homoYeast1
8	Domain Interaction	0.0262	10

6. Generating the receptor interactome and statistical significance analysis

To investigate global graph properties for the predicted receptor interactome, we made predictions for all human receptors. There are 904 receptor genes in the HPMR database [3] and for each one of them we identified their potential PPI partners from all possible 24380 human genes listed on NCBI [6].

For training the final classifier we used a positive set containing all known receptor interaction pairs (2522). The negative training set contains 250,000 random receptor pairs that do not have overlap with any of the HPRD pairs (the positive to negative pairs ratio 1:100 is found to be the best ratio for training data through performance evaluations). The parameters of our RF models are set as follows: The total number of trees was 200. The class cost factor was 5. The number of features to choose in each node was 7.

After deriving predicted interaction scores for all potential receptor related pairs, we applied a heuristic strategy to threshold the predicted RF scores. To estimate what RF cut-off we should use to generate a reliable membrane receptor interactome network graph, we investigated the distribution of predicted scores of testing pairs (from the previous train/testing runs) for known HPRD pairs and the remaining random receptor-protein pairs. From Figure4A (main text) we could see that 2.0 is a pretty stringent cut-off to split two classes. We also found that the stringent cut-off of 2.0 resulted in the recall rate range [15% to 20%] in train/test experiments which is reasonable. Thus we used this value to obtain the receptor interactome network. This cut-off is chosen based on the previous train-test experiments and it achieves 0.55 recall (this high recall rate is caused by that there is no out-of-training pairs) and 0.16 precision (only relying on current known positive) performance.

This thresholded graph contains ~9100 edges, which relates to 559 membrane receptors and 1750 non-receptors. Note: interactions between non-receptor genes are not evaluated and not considered in this graph. Several sub networks from this graph are visualized in Supplement S4.

To further investigate the statistical significance of the predicted pairs and related RF scores, we performed a t-test analysis further. The gold standard negative used for the training of the final RF model is randomly selected from random receptor-protein pairs (exclude HPRD). We repeated this random sampling process multiple times (six times here) and resulted in multiple versions of the negative sets. For each sampled random negative and all the HPRD receptor pairs, we trained a RF model with the above setting. For all potential receptor-protein pairs, we then predicted their RF scores using all the trained RF models. Finally for each receptor-protein pair, we have multiple RF scores related to random samples of gold standard negative sets. Essentially these RF scores should be similar to each other and roughly obey the normal distribution. Thus, t-test is used to measure the hypothesis that a pair's RF scores are normally distributed. The derived p-value is reported for each predicted pair. All receptor pairs in our predicted interactome with the RF cutoff 1.0 are shared in our Supplementary S6. Both their RF scores and the related p-values are also included in the shared EXCEL sheet.

7. Overlap and comparison to existing databases

We also make a comparison between our predicted interactome and other existing datasets, including four computational human PPI graph [O1-O4], one recent published experimental human PPI data [O5] (TAP-MSB07), one yeast experimental PPI data related to membrane proteins [O8] (YeastIyer05), one ERBB related study [O6] and one LUMIER system[O7]. In Table S2.2, the first four rows are for four computational PPI sets and the left for the experimental ones. We also extend our receptor pairs list by lowering the RF cutoff to 1.0. We could see the following table (the last column) that more overlapped pairs could be found for the RF1.0 derived pairs.

First, we compare all the data sets versus the HPRD receptor positive set in the columns [4-6]. We could see that our RFCut2.0 graph achieves the best performance. In our predicted 9144 receptor interactions, 1462 of them are known in HPRD, which means the accuracy is 16.0%. From our Precision-Recall curves reported in the main text (Figure 3), we have the average 20% testing coverage (the training coverage 58% = 1462/2522 HPRD receptor pairs) at the point of accuracy 16%. However for the receptor pairs of "RhodeBioTech05" [O1], it only achieves an accuracy of 3.2%, with coverage of 5.0%. Receptor interactions of "ScottBMCPPI07" [O2] achieve slightly better performance (accuracy 5.1%, coverage 11.5%), but are still worse than our predictions. Similar conclusions could be drawn from the table for STRING [O3] and RaminMSB08 [O4] datasets as well.

We also compared the computational predictions with available high-throughput experimental data ("TAP-MSB07" [7] and "YeastIyer05" [8]). All computational sets do not overlap with the "YeastIyer05" data through homology mapping. For the "TAP-MSB07" data, our interactome hit three pairs. In contrast, RhodeBioTech05 did not detect any of the pairs and the other data "ScottBMCPPI07" had one hit. The other two computational sets have no hits at all. If directly considering the performance of "TAP-MSB07" receptor pairs according to HPRD, the accuracy is just 2.2% with coverage 1.4% if we assume the related receptors are used to "attract" all possible partners in the affinity experiments. This performance is even worse than our computational predictions for receptors.

Overall our predicted interactome graph achieves better identifications of receptor interactions compared to related existing data.

Table S2.2 Datasets used for Overlapping Analysis.

O1.	Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM: **Probabilistic model of the human protein-protein interaction network**. *Nat Biotechnol* 2005, **8**:951-959.
O2.	Scott MS, Barton GJ: **Probabilistic prediction and ranking of human protein-protein interactions**. *BMC Bioinformatics* 2007, **8**:239.
O3.	von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P: **STRING 7--recent developments in the integration and prediction of protein interactions**. *Nucleic Acids Res* 2007, **35**(Database issue):D358-362.
O4.	Ramani AK, Li Z, Hart GT, Carlson MW, Boutz DR, Marcotte EM: **A map of human protein interactions derived from co-expression of human mRNAs and their orthologs**. *Mol Syst Biol* 2008, **4**:180.
O5.	Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, McBroom-Cerajewski L, Robinson MD, O'Connor L, Li M *et al*: **Large-scale mapping of human protein-protein interactions by mass spectrometry**. *Mol Syst Biol* 2007, **3**:89.
O6.	Jones RB, Gordus A, Krall JA, MacBeath G: **A quantitative protein interaction network for the ErbB receptors using protein microarrays**. *Nature* 2006, **439**(7073):168-174.
O7.	Barrios-Rodiles M, Brown KR, Ozdamar B, Bose R, Liu Z, Donovan RS, Shinjo F, Liu Y, Dembowy J, Taylor IW *et al*: **High-throughput mapping of a dynamic signaling network in mammalian cells**. *Science* 2005, **307**(5715):1621-1625.
O8.	Miller JP, Lo RS, Ben-Hur A, Desmarais C, Stagljar I, Noble WS, Fields S: **Large-scale identification of yeast integral membrane protein interactions**. *Proc Natl Acad Sci U S A* 2005, **102**(34):12123-12128.

Table S2.3 Overlap and performance comparison to existing data sets. Here, we compared our predicted interactome to other existing datasets, including four computational human PPI graph [O1-O4], one published experimental human PPI data [O5], one yeast experimental PPI data related to membrane proteins [O8], one experimental data related to ERBB [O6], and one experimental data from LUMIER system [O7].

Data Set	Receptors Related (902Total)	Receptor Interaction Pairs ([1]A: accuracy to HPRD) ([2]C: coverage to HPRD)	Overlapped HPRD Receptor Pairs	HPRD Receptor Pairs Covered In Test (2522 total)	Overlapped TAP-MSB07 Receptor Pairs (136 total)	Overlapped with Our (RF2.0) Interactome (9144 total)	Overlapped with Our (RF1.0) Interactome (42707 total)
RFCut2.0	551	**9144 (A: 16.0%; C: 58%)**	**1462**	2522	**3**	**9144**	**9144**
RhodeBioTech05[O1]	353	**3945 (A: 3.2%; C: 5.0%)**	**125**	2522	**0**	**257**	**739**
ScottBMCPPI07[O2]	380	**5625 (A: 5.1%; C: 11.5%)**	**289**	2522	**1**	**505**	**1099**
STRING08[O3]	581	**2422 (A: 6.4%; C: 6.2%)**	**156**	2522	**0**	**220**	**517**
RaminMSB08[O4]	38	**144 (A: 0%; C: 0%)**	**0**	2522	**0**	**0**	**1**
HPRD	475	2522	**2522**	2522	3	**1461**	**2253**
TAP-MSB07[O5]	27	**136 (A: 2.2%; C: 1.4%)**	**3**	209*	136	**3**	**3**
EGFR-nature06 [O6] (Four ERBB)	4	**181** (with four ERBB proteins) **(A: 22.1%; C: 26.5%)**	**40**	151*	0	**50**	**80**
Lumier05 [O7] (TGFB1)	4	**4** (with TGFB1) **(A: 25%; C: 2.8%)**	**1**	36*	0	**2**	**2**
YeastIyer06[O8]	12	**47** (homologous) **(A: 0%; C: 0%)**	**0**	29*	0	**0**	**2**

- * means the number is estimated

[1] Accuracy = Column_4 / Column_3 ;
[2] Coverage = Column_4 / Column_5;

8. Bibliography

1. Flach, P., The Many Faces of ROC Analysis in Machine Learning, ICML-04 Tutorial. 2004. Notes available from http://www.cs.bris.ac.uk/flach/ ICML04tutorial/
2. Mishra GR, Suresh M, Kumaran K, Kannabiran N, et al. and Pandey A. Human protein reference database--2006 update. Nucleic Acids Res 2006 Jan 1; 34(Database issue) D411-4.
3. Ben-Shlomo I, Yu Hsu S, Rauch R, Kowalski HW, Hsueh AJ., Signaling receptome: a genomic and evolutionary perspective of plasma membrane receptors involved in signal transduction. Sci STKE. 2003. 187: RE9
4. Breiman L. Random Forests. Machine Learning, 2001: 45, 5-32.
5. Guyon I. and Elisseeff A. Special issue on variable and feature selection, The Journal of Machine Learning Research, 2000
6. NCBI Taxonomy http://www.ncbi.nlm.nih.gov/Taxonomy (2005)
7. Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, McBroom-Cerajewski L, Robinson MD, O'Connor L, Li M, Taylor R, Dharsee M, Ho Y, Heilbut A, Moore L, Zhang S, Ornatsky O, Bukhman YV, Ethier M, Sheng Y, et al. (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry, Mol Syst Biol, 3, 89.
8. Miller JP, Lo RS, Ben-Hur A, Desmarais C, Stagljar I, Noble WS and Fields S (2005) Large-scale identification of yeast integral membrane protein interactions, Proc Natl Acad Sci U S A, 102, 12123-12128.
9. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A and Chinnaiyan AM (2005) Probabilistic model of the human protein-protein interaction network, Nat Biotechnol., 8, 951-959.
10. Scott MS and Barton GJ (2007) Probabilistic prediction and ranking of human protein-protein interactions, BMC Bioinformatics, 8, 239.