Amino Acids (2010) 38:1237–1252
DOI 10.1007/s00726-009-0335-7

ORIGINAL ARTICLE

A mouse protein interactome through combined literature mining with multiple sources of interaction evidence

Xiao Li · Haoyang Cai · Jiabao Xu · Sancheng Ying · Yizheng Zhang

Received: 27 January 2009 / Accepted: 24 July 2009 / Published online: 8 August 2009

Abstract Protein–protein interactions (PPIs) play crucial roles in a number of biological processes. Recently, protein interaction networks (PINs) for several model organisms and humans have been generated, but few large-scale researches for mice have ever been made neither experimentally nor computationally. In the work, we undertook an effort to map a mouse PIN, in which protein interactions are hidden in enormous amount of biomedical literatures. Following a co-occurrence-based text-mining approach, a probabilistic model—naïve Bayesian was used to filter false-positive interactions by integrating heterogeneous kinds of evidence from genomic and proteomic datasets. A support vector machine algorithm was further used to choose protein pairs with physical interactions. By comparing with the currently available PPI datasets from several model organisms and humans, it showed that the derived mouse PINs have similar topological properties at the global level, but a high local divergence. The mouse protein interaction dataset is stored in the Mouse protein–protein interaction DataBase (MppDB) that is useful source of information for system-level understanding of gene function and biological processes in mammals. Access to the MppDB database is public available at http://bio.scu.edu.cn/mppi.

Keywords Interactome · Mouse · Protein interaction network · Protein–protein interaction

X. Li and H. Cai contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00726-009-0335-7) contains supplementary material, which is available to authorized users.

X. Li (✉) · H. Cai · J. Xu · Y. Zhang (✉)
Sichuan Key Laboratory of Molecular Biology and Biotechnology, Ministry of Education Key Laboratory for Bio-resource and Eco-environment, College of Life Sciences, Sichuan University, 610065 Chengdu, People's Republic of China
e-mail: lix@scu.edu.cn

Y. Zhang
e-mail: yizzhang@scu.edu.cn

S. Ying
Sichuan Animal Science Academy, 610066 Chengdu, People's Republic of China

Introduction

Protein–protein interactions (PPIs) are involved in biological processes at almost every level of cellular functions, such as DNA replication, transcription, signal transduction, cell cycle control, secretion and intermediary metabolism to name a few. Hitherto, large-scale genome-wide PPI studies have been carried out in several eukaryotic model organisms (yeast, fly and worm) and humans by yeast two-hybrid (Y2H) screens (Uetz et al. 2000; Ito et al. 2001; Giot et al. 2003; Li et al. 2004; Formstecher et al. 2005; Rual et al. 2005; Stelzl et al. 2005) and by co-affinity purification (co-AP) followed by mass spectrometry (MS) (Gavin et al. 2002; Ho et al. 2002; Ewing et al. 2007). These protein interaction networks (PINs) or "interactomes" have quickly become valuable resources to research protein function and understand the molecular mechanisms underlying diseases (Sharan et al. 2007).

Unfortunately, few large-scale researches on PPIs have ever been made in mouse, neither experimentally nor computationally. So far, only the small-scale experiments have been made for detecting mouse PPIs (Suzuki et al. 2001), and the data size still remains very small in the currently public-available databases. This might be due to

the expectation that a comprehensive mouse PIN could be mapped merely by being derived from interactions among orthologous proteins in humans, with whom they share 99% of homologous genes (Waterston et al. 2002; Xuan et al. 2003). For instance, most datasets of mouse PPIs in the current databases, like Bioverse (McDermott et al. 2005) and I2D/OPHID (Brown and Jurisica 2007), were predicted by mapping human PPIs to mouse. Indeed, the laboratory mice have been the most widely used model organism for studying human diseases and related phenotypes, as many similar physiological and pathological features between them (Cox and Brown 2003; Rosenthal and Brown 2007). It is also apparent that, however, there are clear genetic and morphological divergences between mice and humans that cast doubt on the expectation above. To date, little has been known about molecular mechanisms of these divergences. These divergences may be caused by their dissimilarities in many aspects, including transcription regulation, specific gene expression, alternative splicing as well as PPIs. For instance, the current comparisons between human and mouse gene co-expression networks revealed that there is a highly local divergence in gene expression patterns where only a very small fraction (<10%) of co-expressed gene pairs is conserved between the two species (Berg and Lassig 2006; Tsaparas et al. 2006). The previous studies demonstrated that proteins with physical interactions tend to be encoded by co-expressed genes (Ge et al. 2001; Wuchty et al. 2006), and the gene expression levels of interacting proteins show coordinated changes across species (Lemos et al. 2004). The gene expression divergence hints that there might also be a divergence in the PINs between human and mouse. In fact, the conservative protein interactions are extremely rare by analyzing the current public-available PPI datasets among humans and several model organisms (Gandhi et al. 2006; Beltrao and Serrano 2007). For this reason, the mouse PIN has to be mapped independently and renewably instead of simply deriving the map from that of human. As another point of view, once a comprehensive mouse PIN could be generated, it will be a useful resource to facilitate the studies for human protein interactions, and be helpful for further understanding on the molecular mechanisms of human disease at a system level, as well as biological network evolution among mammals.

Currently, most public-available mammalian PPI datasets were generated by four approaches: (1) based on large scans using Y2H assays (Rual et al. 2005; Stelzl et al. 2005), (2) extracted from the literatures (Alfarano et al. 2005; Ramani et al. 2005; Mishra et al. 2006), (3) derived from conserved interactions in other organisms (Lehner and Fraser 2004; Huang et al. 2007; Yellaboina et al. 2008) and (4) predicted by probabilistic or machine learning methods (Rhodes et al. 2005; Shen et al. 2007). It has been widely agreed that the protein interaction data mined from the literature has the highest confidence, especially for those from manually curated datasets. The Y2H-based mapping approaches can rapidly detect interactions between thousands of proteins, but might be compromised by large false-positive rates (over 50%) (von Mering et al. 2002). Similar with the Y2H approaches, the probabilistic or machine learning methods also produce a high rate of false positives. As these approaches have their obvious advantages as well as disadvantages, it seems to be an effective way in protein interaction mapping by combining different approaches.

In this work, we undertook an effort to map a mouse PIN, where protein interactions are embedded in enormous amount of biomedical literatures (Fig. 1). The co-occurrence approach appears to be the simplest and most comprehensive to find the association between biological entities, such as the protein relations, because two entities co-cited in a same text unit (e.g., a sentence or abstract) should have an underlying biological relationship (Stapley and Benoit 2000; Jenssen et al. 2001; Ramani et al. 2005; Li et al. 2006b). We then employed a probabilistic model (naïve Bayesian) for filtering protein pairs by integrating various types of genomic and proteomic data, including evolutionarily conserved protein interactions, underlying

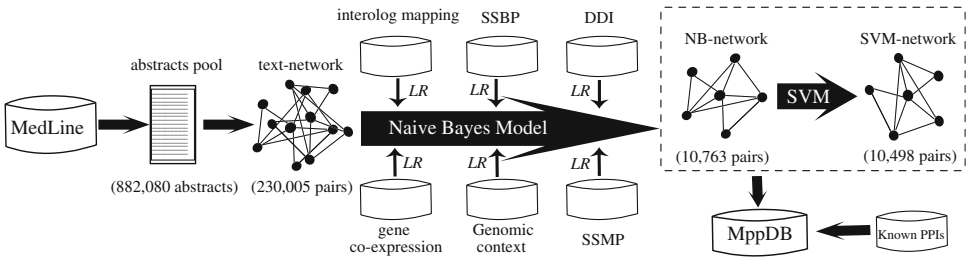

Fig. 1 A flowchart depicting the three-step approach for deriving mouse protein interaction network. Step 1 was to derive a protein pair dataset by text mining for MedLine titles and abstracts based on a tri-occurrence method. Step 2 was to filter protein pairs by integrating heterogeneous genomic and proteomic datasets based on a naïve Bayesian model. In step 3, a Support Vector Machine method was used to further choose protein pairs with direct physical interactions. The protein pairs from step 2, assembling the positive reference set in which protein pairs were collected from several protein–protein interaction databases, are stored into the database MppDB

domain–domain interactions, gene co-expression patterns, protein function and phenotype similarity. Up till now, the integrative approaches have been employed successfully in various species, including yeast (Lu et al. 2005), human (Jansen et al. 2003; Rhodes et al. 2005; Xia et al. 2006; Li et al. 2008), *Plasmodium falciparum* (Date and Stoeckert 2006; Wuchty and Ipsaro 2007) and *Arabidopsis thaliana* (Cui et al. 2008) to predict protein interactions mainly due to two remarkable advantages. First, it can integrate heterogeneous kinds of evidence and tolerate missing data among them. Second, it is a simple but highly efficient model to tackle with data in a large scale with short time consumption. In mouse, a similar probabilistic model also was applied to establish a functional network (MouseNET) which was demonstrated to be an invaluable resource for protein function prediction (Guan et al. 2008). Here we focus on the construction of mouse interactome—a map of PPIs. Therefore, a learning algorithm—support vector machine (SVM) was further used to validate protein pairs with direct physical associations. The derived mouse PPIs were stored in the Mouse protein–protein interaction DataBase (MppDB), which is public available, and the PINs involving interesting proteins are visualized online.

Materials and methods

Tri-occurrence-based text mining

The work started to derive a protein pair dataset by text mining for MedLine titles and abstracts. It is based on the assumption that two proteins should have an underlying biological relationship if they are co-cited in the same text unit. We performed the EFetch program, a perl script for batch retrieving NCBI data records (http://eutils.ncbi.nlm.nih.gov/entrez/query/static/efetch_help.html), to download 882,080 MedLine records (until February 2007) by using "mouse", "mice" and "Mus musculus" as keywords in all query fields. Protein names were recognized by the program NLProt (Mika and Rost 2004). NLProt is a tool to find protein and gene names in natural language text, and is capable of mapping protein names to their associated entries in the UniProtKB sequence database.

As co-occurrence may indicate many kinds of relationships, we used a tri-occurrence approach to extract protein pairs with potential physical interaction by adding a verb existence in the same sentence. We made a collection of 42 verbs, including their variation forms, to construct a verb list that can reflect protein–protein physical interactions (Supplementary Table 1). For over two protein names in a same sentence, any pair of proteins was considered as a potential interaction. Thus, 230,005 protein pairs among 16,226 proteins were retrieved to compose an initial network called text-network.

Bayesian probabilistic model

Reference datasets

A positive reference set (PRS) and a negative reference set (NRS) are needed when applying a naïve Bayesian classifier. We gathered mouse PPI data stored in five PPI databases: Database of Interacting Proteins (DIP; 103 interactions among 156 proteins) (Salwinski et al. 2004), Biomolecular Interaction Network Database (BIND; 2,283 interactions among 2,095 proteins) (Alfarano et al. 2005), Munich Information Center for Protein Sequences (MIPS; 230 interactions among 220 proteins) (Pagel et al. 2005), Molecular Interactions Database (MINT; 1188 interactions among 1,175 proteins) (Chatr-aryamontri et al. 2007), and IntAct (2,628 interactions among 2,059 proteins) (Kerrien et al. 2007), where proteins for each PPI datasets were unique identified by UniProtKB accessions. We mapped the UniProtKB accessions to Entrez GeneIDs for all proteins by BioMart (http://www.biomart.org/). After removing duplications and self-interactions, 5,162 distinct PPIs among 3,414 proteins unique identified by Entrez GeneIDs were used as the PRS. However, it is difficult to generate a NRS because there are no "gold standard" non-interactions. In general, there are two methods for generating NRS. A method is based on annotations of cellular localization, and the observation that pairs of proteins that have different localization patterns are unlikely to interact. The other is to select non-interacting pairs uniformly at random from the set of all protein pairs that are not known to interact. Here, we generated two NRS for training the naïve Bayesian classifier by using the two above-mentioned methods:

- As suggested previously (Jansen et al. 2003; Rhodes et al. 2005), a NRS can be defined as all the possible pairwise combinations, in which one protein is assigned to the plasma membrane and the other to the nucleus according to the Gene Ontology (GO) (Harris et al. 2006) cellular component annotation. By using the GO annotations for mouse genes from the NCBI (fttp://fttp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz) and without considering the GO evidence code IEA (Inferred from Electronic Annotation), we found 442 proteins localized in the plasma membrane and 1,441 proteins in the nucleus. By removing 42 proteins that are localized in both plasma membrane and nucleus, the NRS included 559,600 protein pairs among 1,799 proteins (400 proteins in the plasma membrane and 1,399 proteins in the nucleus). We named the NRS as Loc-NRS.

- We first generated a set of protein pairs by pairwise combining all of proteins that appeared in the PRS. Then we eliminated self pairs and interactions from the PRS. Finally, we selected protein pairs uniformly at random as a NRS where the number of protein pairs was 100 times more than that in the PRS (Ben-Hur and Noble 2006). The NRS was named Ran-NRS.

A likelihood ratio (LR) corresponding to a specific biological evidence (E_i) was used to measure the predictive power or confidence degree, and calculated as the ratio of the true positive rate (TPR) to the false-positive rate (FPR), where$\text{TPR} = |Ei \subset \text{PRS}|/|\text{PRS}|$ and $\text{FPR} = |Ei \subset \text{NRS}|/|\text{NRS}|$. In theory, $\text{LR}(E_i) > 1$ indicates that the biological evidence E_i is capable of identifying the true positives from a test.

Interology mapping

Interologies are conserved PPIs cross organisms (Matthews et al. 2001). Many PPIs were successfully predicted by mapping interologies, showing that interology provides a significant predictive power. We mapped three model organisms (yeast, worm and fly) and human PPI datasets to mouse protein pairs, using gene orthologs defined in the Inparanoid database that is based on an all versus all BLAST search following by clustering into orthologous groups (O'Brien et al. 2005). Four human interactome datasets were used to predict mouse PPIs: (a) two high-throughout Y2H datasets created by Stelzl et al. (2005) and Rual et al. (2005) that were named Y2H-1 and Y2H-2, respectively; (b) one dataset generated by MS method from Ewing et al. (2007); and (c) one literature-curated dataset from the Human Protein Reference Database (HPRD) that is a manually curated PPI database with high confidence (Mishra et al. 2006). The yeast (*Sacchromyces cerevisiae*), worm (*Caenorhabditis elegans*) and fly (*Drosophila melanogaster*) PPI datasets were queried from the DIP database on January 2007. The numbers of proteins and interactions for all PPI datasets are showed in Supplementary Table 2. For each data set, mouse protein pairs were divided into two bins: interolog or non-interolog, and then the LR for each bin was calculated.

Gene co-expression

As two interacting proteins often tend to be gene co-expressed (Ge et al. 2001; Wuchty et al. 2006), especially for those involved in the same protein complex and the same biochemical reaction, the gene expression profiles can be used to predict PPIs. We examined three high quality large-scale microarray datasets to calculate the co-expression data for each pair of genes by measuring the Pearson correlation coefficient (PCC). The three microarray datasets (GSE1986 (unpublished), GSE1133 (Su et al. 2004) and GSE3327 (Hovatta et al. 2005)) consisted of gene expression profiles cross a range of mouse tissues, strains and brain regions, and were downloaded from the gene expression omnibus (GEO) (Barrett et al. 2007). The values of PCC were divided uniformly into 20 bins, and then the LR for each bin of each dataset was calculated.

Domain–domain interaction

Proteins perform their biological functions often through domains as units, including interactions with other proteins. Therefore, novel PPIs may be predicted by identifying pairs of domains enriched among known interacting proteins. We downloaded two domain–domain interaction (DDI) databases, The databases of 3D Interaction Domains (3did) (Stein et al. 2005) and InterDom (Ng et al. 2003), both were derived mainly from structural information. For the dataset InterDom, we used the DDI scores from the original database to measure the co-occurrence of particular domain pairs (Xia et al. 2006; Li et al. 2008). And for the dataset 3did, we binned protein pairs into two groups depending on whether a protein pair contains the special domain pairs on the database 3did, "hit" for yes, and "no-hit" for no (Li et al. 2008).

Biological process similarity

Proteins involved in the same biological process (e.g., cell cycle and immune response) or with similar functional annotations are more likely to interact with each other than those in different biological processes. We introduced the smallest shared biological processes (SSBP) to quantify protein biological process similarity. The SSBP is based upon that a pair of proteins has a high likelihood to interact with each other if they share a small and specific functional annotation (Rhodes et al. 2005; Xia et al. 2006), and calculated by three steps: (1) search all the GO terms shared by the pair of proteins, (2) find the number of other proteins also sharing these GO terms, (3) get the GO term with the smallest protein count.

Mammalian phenotype annotation

It is well known that interacting proteins tend to result in similar phenotypes while losing their functions. Because functional loss of a protein may cause multiple phenotypes, proteins sharing more phenotype annotations are more likely to interact than those sharing less phenotype annotations. The mammalian phenotype (MP) ontology (Smith et al. 2005) has been constructed to provide standard terms

for annotating mammalian phenotypic data. We queried the MP annotations for all mouse proteins by the Mammalian Phenotype Browser (http://www.informatics.jax.org/searches/MP_form.shtml), and used the smallest shared mammalian phenotypes (SSMP) to quantify gene phenotypic similarity. The computational procedure for SSMP is similar with that for SSBP.

Genomic context

Genomic context has been used to predict PPIs in silico, including Rosetta Stone (RS), gene co-occurrence (GC) and gene neighborhoods (GN) (von Mering et al. 2002). Using the database Prolinks (Bowers et al. 2004), we searched three types of genomic context for each protein pair. Mouse protein pairs were grouped into two bins depending on whether a protein pair exists in each type of genomic context, "hit" for yes, and "no-hit" for no.

Integrating evidence by naïve Bayesian model

We integrated evidence from different sources by the naïve Bayesian probabilistic model described extensively in several studies previously (Jansen et al. 2003; Rhodes et al. 2005; Xia et al. 2006; Li et al. 2008). Briefly, we defined a pair of proteins that interact with each other as "positive" and those do not interact as "negative". Following the Bayesian theorem, the prior odds (O_{prior}) of finding a PPI can be calculated as:

$$O_{\text{prior}} = P_{\text{pos}}/P_{\text{neg}}$$

where P_{pos} is the probability that a pair of proteins interacts within all the possible protein pairs while P_{neg} stands for the possibility that a pair of proteins don't interacts. When considering the given n evidences (E), the posterior odds ($O_{\text{posterior}}$) of an interaction can be computed as:

$$O_{\text{posterior}} = \frac{P(\text{pos}|E_1,\ldots,E_n)}{P(\text{neg}|E_1,\ldots,E_n)}.$$

Let the likelihood ratio (LR) is defined as:

$$\text{LR}_{(E_1,\ldots,E_n)} = \frac{P(E_1,\ldots,E_n|\text{pos})}{P(E_1,\ldots,E_n|\text{neg})},$$

the posterior odds of an interaction can be calculated as the product of the prior odds and the likelihood ratio ($O\text{posterior} = O\text{prior} \times \text{LR}_{(E_1,\ldots,E_n)}$). If we assume that evidence is conditionally independent, the composite LR can be calculated simply as following:

$$\text{LR}_{(E_1,\ldots,E_n)} = \prod_{i=1}^{n} \text{LR}(E_i),$$

which is namely naïve Bayesian model (Jansen et al. 2003; Rhodes et al. 2005; Date and Stoeckert 2006; Xia et al. 2006). In the case, the evidence is the data sets used to infer PPI between the proteins. As the prior odds is a constant, the composite LR corresponding to a specific biological evidence can be used to measure the predictive power or confidence degree for predicting PPIs. A cutoff of likelihood ratio (LR_{cut}) is represented as an indicator whether a protein pair interacts (that is, yes if the composite LR is above the LR cutoff, no if not). By filtering in the naïve Bayesian model, the resulting protein pairs with the composite LR above the LR cutoff were identified as true positives and used to construct the second network named as NB network.

Receiver operating characteristic (ROC) curve

A ROC curve allows us to explore the relationship between the sensitivity and specificity of a binary classifier system for a variety of different cut points (Baldi et al. 2000). The ROC can also be represented equivalently by plotting the fraction of true positives (TPR true positive rate) versus the fraction of false positives (FPR false-positive rate). Sensitivity and specificity can measure the ability of a classifier to identify true positives and false positives in a test, and calculated as sensitivity = TP/positives, and specificity = 1 − (FP/negatives), where TP and FP are the number of true positives and false positives identified by a classifier, respectively, whereas positives and negatives are the total number of positives and negatives in a test. The area under the ROC curve (AUC) is an indicator of the efficacy of the assessment system. Thus, the performances of the different classifiers appear to be comparable by measuring the AUCs, that is to say, the larger the AUC, the better the performance. The SPSS software was performed to plot and smooth the ROC curves and calculate the AUCs (SPSS 1999). Besides the ROC curve, precision–recall curve is also widely used measure to evaluate the prediction performance for a classifier. Precision and recall are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \text{ and } \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

where FN is the number of false negatives. In the case, the precision and recall curve may be a more informative measure because the number of negatives is greatly larger than that of positives.

Determining physical interactions by support vector machine

We further performed a machine learning approach to pick up PPIs with directly physical association in NB network. Because some types of evidence integrated in the naïve Bayesian model above, such as gene co-expression,

function similarity and shared phenotype annotation, can only contribute to inferring protein pairs with functional relationship, the relationship type remains still uncertain for many protein pairs. We chose a SVM method developed by Shen et al. (2007) to achieve the purpose based on protein sequence information. Briefly, the amino acid features of a protein can be described as a conjoint triad according to its dipoles and volumes of the side chains, and the pair of proteins is then assigned into positive (interaction) or negative (non-interaction) based on the combinative features between them by using SVM. The detailed procedure about SVM algorithm can be found in Supplementary Materials or Vapnik (2005).

As a SVM method is sensitive to the quality of training sample, we refined these PPIs in the PRS. By mapping interologs between protein pairs in PRS and human PPIs in the HPRD database, 1,221 PPIs were chosen as the positive sample. Two sets of protein pairs with equal number of the positive sample were randomly extracted from Loc-NRS and Ran-NRS, respectively, and then were used to train the SVM classifier as the negative samples. Following the process described by Shen et al. the SVM classifier was implemented by the libsvm package 2.8 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/).

Results and discussion

Predictive power of individual dataset

Here we integrated six heterogeneous types of evidence from 17 data sets and used the likelihood ratios as the reliability of individual dataset to infer mouse PPIs by using the naïve Bayesian model. Each dataset was divided into several bins, and the LR for each bin was calculated over all datasets according to the gold standard reference set. For the datasets with discrete values, including SSBP, SSMP and InterDom, we binned manually the original scores and computed the LR for each bin. Figure 2 illustrates the correlations between the datasets and the corresponding LRs using Loc-NRS as the negative set, whereas those using Ran-NRS as the negative set are shown in Supplementary Fig. 1. We observed the clear correlations between 17 datasets and the LRs derived by the two negative sets, which indicates that the LRs can be taken as a relative measure for predicting mouse PPIs. For most of datasets, the LRs derived from Loc-NRS are higher than those from Ran-NRS. The highest LRs were contributed from the eight datasets by the interology mapping (Fig. 2a) and the DDI database InterDom (Fig. 2f). For gene co-expression, a significant correlation between the expression PCC and the LR was found when the PCC is above 0.6 (Fig. 2b).

Assessing conditional dependence between datasets

The naïve Bayesian classifier is based on the assumption that predictive evidence is strong independence. Here we assessed the statistical dependence between each pair of datasets by calculating the PCC (Lu et al. 2005). Because the values of datasets have various forms, we had to transform all of them to a same form for calculating PCC. For each dataset, we assigned a mouse protein pair to 1 if the evidence is present, otherwise to 0. An exception was that, for gene co-expression, protein pairs with the PCC of above 0.60 were assigned to 1, otherwise to 0. Table 1 lists the PCCs between each pair of datasets given the PRS and Loc-NRS (the PCCs for Ran-NRS is showed in Supplementary Table 3), and the highest absolute values are highlighted in bold. The result indicates that some datasets are statistical dependence or redundant. In general, the datasets belonging to a same evidence type appear dependence, including two DDI datasets, two microarray datasets and three genomic context datasets. Moreover, some datasets from different evidence types also display dependence, including SSBP, SSMP, 3did, InterDom and HPRD. This suggests that these evidence sources cannot be treated as conditionally independent predictors. However, the dependences among evidence sources within the NRS are weaker than those within the PRS, which is due to that the NRS, including Loc-NRS and Ran-NRS, has plenty of missing values for many sources of evidence.

To avoid bias, the datasets with statistical dependence had to be combined to a joint evidence source before implementing the naïve Bayesian classifier. Considering the data true of evidence sources, we used different strategies to combine the datasets with conditional dependence. For the two microarray datasets (GSE1986 and GSE1133), only the maximum PCC was assigned into a mouse protein pair. For genomic context, the joint evidence was considered as true if any one evidence type was present among the three evidence types of genomic context. However, it is difficult to combine the five datasets: SSBP, SSMP, 3did, InterDom and HPRD, because they belong to different evidence types and have different forms of data value. A decision tree algorithm J48 was used to combine them and to identify logical bins based on the five datasets (Rhodes et al. 2005). By the decision tree algorithm, we stratified them into 15 confidence bins as implemented in the Weka software package (Witten and Frank 2000) (Supplementary Fig. 2).

Note that the PCC only measures linear relationships between datasets. Therefore, we further assessed statistical dependence between datasets by comparing the true LRs with the expected LRs from the naïve Bayesian model (an expected LR is a product of LR of each dataset within a special bin). It is because if datasets are statistical

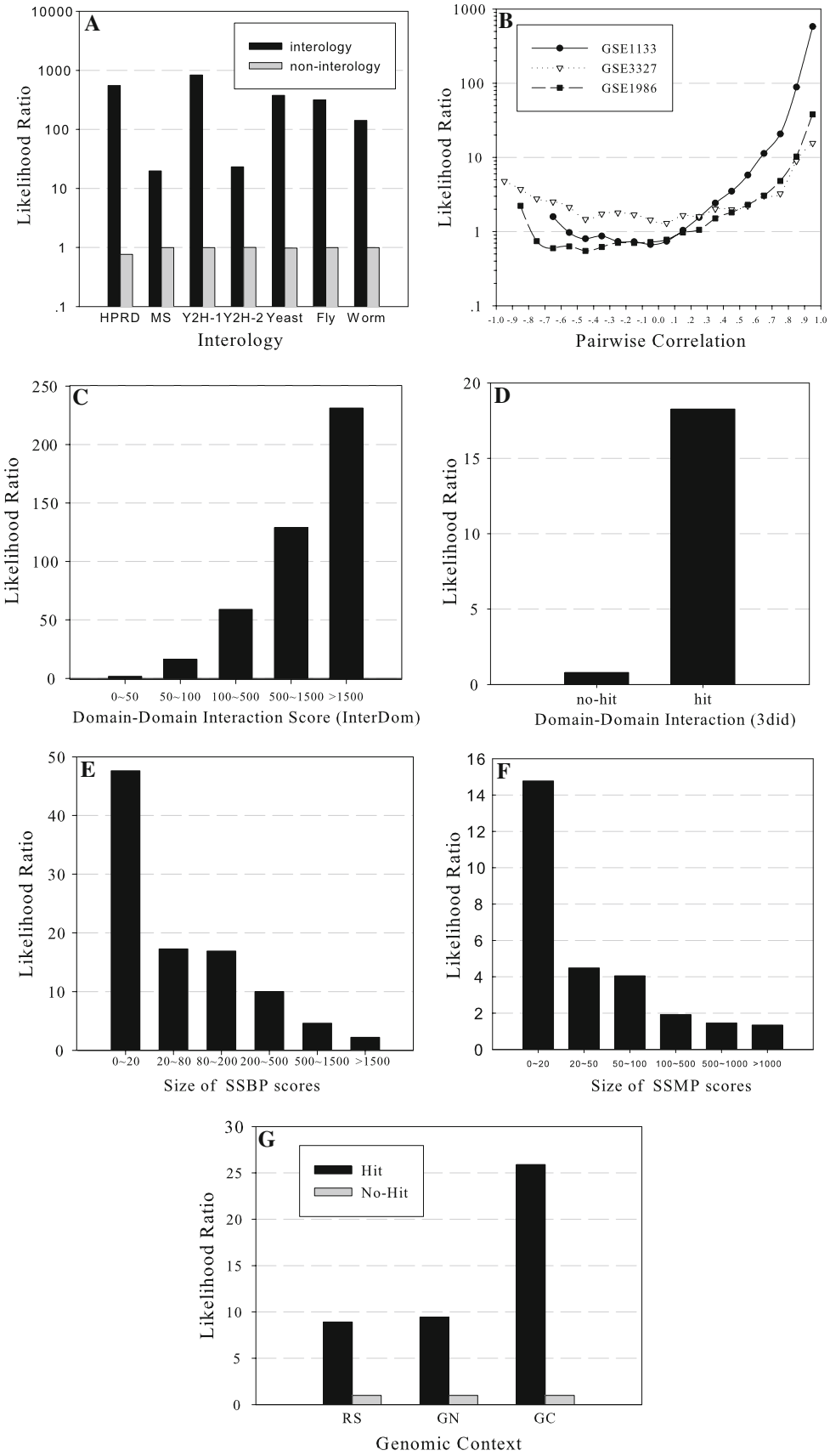

Fig. 2 Measurement of the contributions of diverse genomic and proteomic date sources for predicting mouse PPIs. The likelihood ratios were computed according to the positive reference set and the negative reference set Loc-NRS. **a** Large-scale PPI datasets from three model organisms and human. Human and model organism proteins were mapped to mouse orthologs using the Inparanoid database. **b** Three large-scale microarray datasets. For each protein pair, the PCC was calculated for each dataset by mapping to Entrez GeneIDs, and then all of protein pairs were binned by their PCC in each dataset. **c** DDI scores from the database InterDom. A pair of proteins was assigned a DDI score if they contain the special domain pairs from the DDI database InterDom. If different scores exist between a pair of proteins arising from different interacting domain pairs, the maximum of the scores was assigned to the pair. Protein pairs were grouped according their DDI scores. **d** DDI from the database 3did. **e** The SSBP was used to measure a pair of proteins' biological process similarity. The mouse protein annotation from the Gene Ontology Annotation (GOA) project was used to examine the SSBP for each protein pair. All of protein pairs were divided into several groups according to their SSBP. **f** The SSMP was used to measure a pair of proteins' phenotypic similarity. **i** Genome context. Three types of genomic context were used to in silico predict PPIs. *RS* Rosetta Stone, *GN* gene neighborhood, *GC* gene co-occurrence

Table 1 Pearson Correlation coefficients between evidence sources on the PRS and the negative set Loc-NRS

NRS*100	PRS*100																
	GSE 1133	GSE 3327	GSE 1986	SSMP	SSBP	3did	InterDom	HPRD	MS	Y2H-1	Y2H-2	Worm	Yeast	Fly	RS	GN	**GC**
GSE1133		6.8	**31.9**	5.1	13.1	6.7	2.6	4.9	1.3	3.1	0.2	0.2	7.3	0.7	0.7	2.5	6.9
GSE3327	3.2		9.4	0.2	5.2	3.7	1.9	5.7	1.1	0.7	0.8	0.4	1.0	0.5	1.8	0.8	0.9
GSE1986	12	3.2		8.8	9.5	7.5	6.4	3.2	1.7	0.8	2.5	0.3	9.1	1.3	1.9	1.3	3.7
SSMP	1.3	3.7	3.1		**19.9**	**15.9**	**20.0**	**19.2**	0.4	1.3	0.4	0.8	6.0	3.8	7.3	1.5	4.2
SSBP	0.3	0.9	0.9	8.4		**27.5**	**23.9**	**25.1**	3.0	2.1	6.2	0.9	12.5	2.2	10.8	4.4	9.2
3did	0.7	0.9	0.5	0.5	10.2		**36.6**	**14.5**	2.5	2.7	5.8	1.1	8.4	6.8	**18.6**	5.4	12.0
InterDom	0.6	1.2	1.2	2.5	5.2	**17.5**		**17.2**	1.9	2.8	3.6	3.9	10.7	4.6	13.1	1.8	7.8
HPRD	0.8	0.2	0.5	2.2	2.3	2.8	1.8		5.5	4.3	12.0	0.8	0.8	4.3	4.9	0.1	1.0
MS	0.1	0.5	0	0.3	0	0.1	0.3	0.7		0.1	5.0	6.0	4.3	8.7	2.3	0.2	0.4
Y2H-1	0	0.1	0.1	0.1	0.1	0.1	0.2	**22.5**	0		11.9	0.1	0.3	0.2	0.3	0.1	0.2
Y2H-2	1.4	0	0.5	0.1	0	0	0.3	11.2	0	0		9.8	3.2	0.5	0.8	0.2	0.5
Worm	0	0.2	0.1	0	0.1	0.1	0.4	0	0	0	0		0.8	12.5	2.0	0.2	4.3
Yeast	0.4	0.1	0	0.1	0.7	0.6	1	1.2	2.0	0	0	0		2.4	2.7	9.5	4.7
Fly	0	0.2	0.5	0.6	0.1	0	0.6	0	0	0	0	0	0		3.0	0.2	0.6
RS	0.8	0.5	0.1	0.7	6.5	15.6	7.2	1.2	0.1	0	0	0	0	0		**16.7**	**36.6**
GN	0.1	0.1	0	0.3	0.2	0.8	1.8	0	0	0	0	0	0	0	**15.2**		0.2
GC	0.4	0.5	0.1	0.8	6.7	10.4	2.9	0.6	0	0	0	0	0	0	**34.3**	0	

The highest absolute values in each category are highlighted in bold

PRS positive reference set, *NRS* negative reference set, *GSE1133, GSE3327, GSE1986* the Gene Expression Omnibus accessions, *SSMP* smallest shared mammalian phenotypes, *SSBP* smallest shared biological processes, *3did and InterDom* two domain–domain interaction datasets, *HPRD* the Human Protein Reference Database, *MS* a human PPI dataset generated by high-throughput MS technology, *Y2H-1 and Y2H-2* two human PPI datasets created by high-throughput Y2H method, *worm, yeast and fly* the worm (*Caenorhabditis elegans*), yeast (*Sacchromyces cerevisiae*), and fly (*Drosophila melanogaster*) PPI datasets from the Database of Interacting Proteins (DIP), *RS* Rosetta Stone, *GN* gene neighborhoods, *GC* gene co-occurrence

dependence, the expected LRs are higher than the true LRs (Rhodes et al. 2005). Table 2 lists the true LRs and expected LRs for the 15 bins created by J48 decision tree, showing the expected LRs are considerably higher than the true LRs for most of bins. The results further verified that the five datasets are strongly dependent and contain redundant information.

Performance evaluation of predictions

We used the ROC curve to evaluate the performances of predictions. To evaluate the overall performance of various predictions, the fivefold cross-validation was performed. First both the PRS and NRS were randomly divided into five subsets with equal number. Then we used four of the five subsets as the training set to compute the likelihood ratios of the individual evidence, and the remaining subset as the test set to identify the number of positives and negatives. A protein pair is predicted to be positive when its likelihood ratio is above a particular cutoff, to be negative otherwise. By repeating this process five times across the five test sets, we summed up the number of true positives and false positives under each of differential LR cutoff, and then calculated Sensitivity and Specificity for plotting the ROC curves.

In addition to the naïve Bayesian classifier that combined all datasets, we evaluated the performances of single dataset models where the confidence of each protein interaction is assigned by likelihood ratio of the confidence bins of individual dataset. Because the five datasets (SSBP, SSMP, 3did, InterDom and HPRD) were combined and binned by the J48 decision tree, we also tested the performance of prediction only by using the five datasets, and named the J48 model. The ROC curves of various prediction models are illustrated in Fig. 3. We observed that the predictive performance of the J48 model is higher than those of single evidence models, and the two naïve Bayesian models get the advantage over that of the J48 model, suggesting that the more datasets are combined, the better performance can be achieve. Moreover, we found that the prediction efficacy of the naïve Bayesian models on Loc-NRS is better slightly than that on Ran-NRS. Considering the number of negative sample is greatly larger than that of positive sample, a precision–recall curve was plotted to evaluate and compare overall predictive performance on two distinct negative sets (Supplementary

Table 2 The comparison between the real and the expected likelihood ratios for 15 bins grouped by the J48 decision tree

Bins	Real LR	Expected LR
G1	0.63	0.63
G2	7.60	10.87
G3	22.72	93.05
G4	758.85	9,160.59
G5	22.71	820.25
G6	6.99	39.05
G7	68.98	433.07
G8	25.29	651.99
G9	108.40	81,078.87
G10	424.59	41,547.80
G11	23.23	12,443.16
G12	22.99	472,153.85
G13	4,119.48	5632.09
G14	173.45	1,565.76
G15	553.83	553.83

Since the five datasets (SSBP, SSMP, 3did, InterDom and HPRD) are conditionally dependent, they are combined and stratified 15 bins (G1 to G15) by the J48 decision tree (Supplementary Fig. 2). For each bin, the true LR is observed in the PRS and Loc-NRS, whereas the expected LR is computed by the naïve Bayesian model (see text for details)

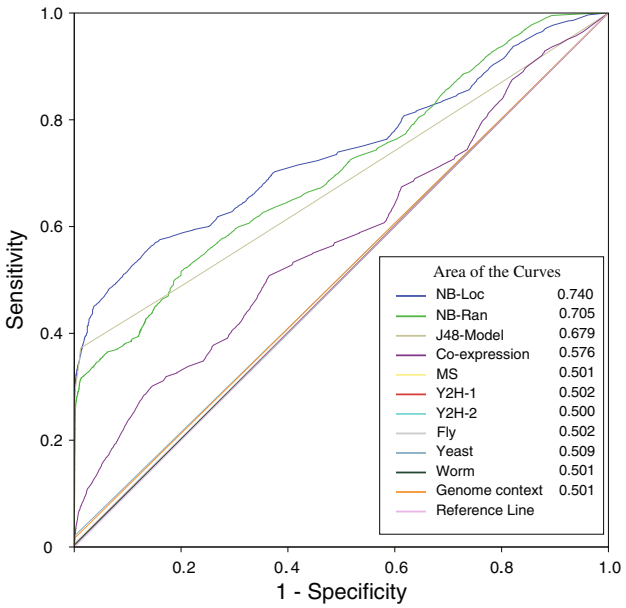

Fig. 3 ROC curves for evaluating the performances of various prediction models using fivefold cross-validations. The different prediction models are highlighted in different *colors*, and the area under each curve is present in the figure. NB-Loc and NB-Ran denote the naïve Bayesian models that integrate all of evidence sources and are based on the negative set Loc-NRS and Ran-NRS, respectively, whereas the J48 model combines the five evidence sources (HPRD, SSBP, SSMP, InterDom and 3did) that are conditional dependent and binned by the J48 decision tree. The others denote the single evidence models: *MS* a human PPI dataset generated by MS method, *Y2H-1 and Y2H-2* two human PPI dataset created by Y2H method, *yeast, fly and worm* the PPI datasets of three model organisms, *co-expression* a human large-scale gene expression dataset (GSE1133), and *genomic context* a joint evidence source from three evidence types of genomic context (Rosetta Stone, gene neighborhood and gene co-occurrence). Sensitivity and specificity are calculated using the fivefold cross-validations. All of prediction models are based on the negative reference set Loc-NRS except the NB-Ran model that uses Ran-NRS as the negative reference set

Fig. 3). As seem in the precision–recall curve, it is certain that the Loc-NRS has a slightly higher predictive performance than the Loc-NRS. We noticed that the result is not in accordance with that described from Ben-Hur and Noble (2006) where the authors considered that the negative set composed of proteins pairs that are not co-localized may lead to a biased assessment of classifier accuracy on the prediction of yeast protein interactions. This is mainly due to for different organisms, the accuracy and coverage rates of genomic datasets are different significantly. Moreover, the accuracy and quantity of protein interactions on the positive set also affects the assessment system, especially for the chose of negative samples. Indeed, a criticism of randomly chosen negatives is that they will contain true interactions. For mouse, the recent interaction data are still far from coverage and a very small quantity of interactions has been detected experimentally, which leads to the set of randomly chosen negatives is more likely to be contaminated with interacting proteins because we cannot remove effectively true interactions. Conversely, localization-derived negatives should be free of contamination because the subcellular localization annotations are relatively complete and accurate for mouse proteome (Sprenger et al. 2008). For this reason, we suggested that negative samples need to be chosen with case for PPI predictions of different organisms.

To get an appropriate composite LR cutoff, we plot the ratio of true positive to false positive (TP/FP) as the function of the cutoff of likelihood ratio. As showed in Fig. 4, given Loc-NRS as the negative set, the TP/FP ratios are apparently correlated with LR cutoffs, and they increase monotonically with the increase of cutoffs. But the correlation is not clear on the other negative set Ran-NRS especially when the cutoff of likelihood ratio is above 100, which is mainly due to the low likelihood ratios and the contamination of true interactions on the Ran-NRS. We have noted that most of genomic datasets generate low likelihood ratios based on the Ran-NRS in comparison with the Loc-NRS (Fig. 2; Supplementary Fig. 1). By the naïve Bayesian rule, the composite LR of a true interaction is significant lower on the Ran-NRS because it is calculated by multiplying individual LR from all datasets. Therefore, when setting a high cutoff value, only a small number of true interactions will be remained, and thus we will achieve a small number of true positives. Our investigation shows

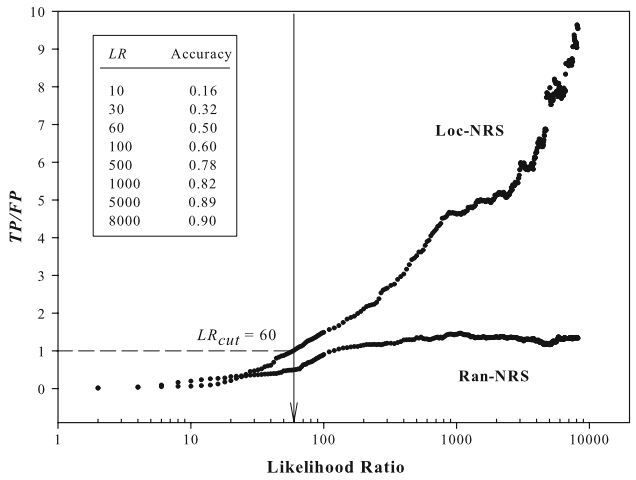

Fig. 4 TP/FP ratios (true positive versus false positive) at different LR cutoffs

that the numbers of true positives decrease significantly with the increase of cutoff value on the Ran-NRS. Furthermore, the Ran-NRS is inevitably contaminated with true interactions, and interacting protein pairs are still present because of the large data size of randomly chosen negatives. In general, they have a high likelihood ratio and thus will be classified as false positives even though a high cutoff value is set. Considering the limited number on the PRS, the contamination of true interactions has a significant influence on the TP/FP ratios even though the contamination rate is probably low. The unclear correlation on the Ran-NRS shows that the likelihood ratios can not be used to assess the confidence of the predicted protein interactions. Therefore, we used the Loc-NRS as negative set and selected the composite LR cutoff as 60 where the TP/FP is 1. That is to say, we can achieve the 50% accuracy of prediction at this resolution. Based on the naïve Bayesian model and the LR cutoff, 10,763 protein pairs among 5,299 proteins from text-network were remained to create the second network called NB network. Of course, users can filter out the higher confidence interactions by setting a higher threshold value of likelihood ratio.

As for the contribution of a special dataset to predicting PPIs, in fact, it is not sufficient if only the derived LR is considered, and the TPR has to be taken into account. For example, although the LR derived from the HPRD dataset is slightly weaker than that from Y2H-1 dataset, the TPR of the HPRD dataset is remarkably higher than that of Y2H-1 dataset, which is due to that the HPRD dataset also has a higher FPR than Y2H-1 dataset. This also means that the LR and TPR of predictive datasets are not always proportional. We inquired the TPR for all datasets on the PRS, text-network, NB network and SVM network (Supplementary Table 4), and found that the TPR of the seven datasets that derive the four types of evidence (gene co-expression, DDI, biological process similarity and phenotype similarity), are much higher than those of other datasets. This means that these datasets make greater contributions to predicting PPIs than other datasets. Within all types of evidence, many datasets can only provide the indirect evidence for inferring protein pairs associated with physical interactions, including SSBP, SSMP and gene co-expression. However, they have the high TPRs, indicating that NB network may still remain false-positive results where many protein pairs are only functional relative and not physical interactions. For this reason, the SVM approach is necessary for further filtering false positives in NB network.

Obviously, the LR and the TPR of a special dataset are affected by the quality and the comprehensiveness of both the reference sets and the dataset itself. Currently, many datasets are limited in accuracy and far from the complete coverage, such as the PPI data of model organisms, mouse protein function and phenotype annotations. To increase coverage and improve confidence in the prediction of mouse protein interactions, therefore, the naïve Bayesian model will need to be updated in the future as various datasets become more accurate and complete, as well as new predictive evidence sources are available.

The SVM predictions

We performed a SVM learning algorithm to identify protein pairs with physical interactions in NB network. The SVM prediction models were constructed based on the two negative samples, respectively. Moreover, we also conducted the fivefold cross-validation protocol to all tests for obtaining an appropriate negative sample, kernel function and a set of optimal parameters. The predictive performances of various SVM models are illustrated in Supplementary Fig. 4. We achieved the best prediction by using the negative sample extracted from Loc-NRS and linear function as the kernel function. The optimal values of C and γ for constructing the SVM model are 8 and 1.25, respectively. By the optimal SVM model, 97.54% of protein pairs in NB network were predicted to be positive, which may indicate that the quality of the NB network is high. We used the retaining 10,498 protein pairs among 5,245 proteins to construct the third network named SVM network.

The mouse PPI networks analysis

Overlap of PPIs between mouse PPI datasets

Three mouse PPI networks were generated by following the three-step approach, and named text-network, NB network and SVM network, respectively. SVM network is

a sub-network of NB network that is a sub-network of text-network. NB network and SVM network contain 10,763 PPIs among 5,299 proteins and 10,498 PPIs among 5245 proteins uniquely identified by the corresponding Entrez Gene IDs. We analyzed the shared proteins and interactions between the three predicted networks and the PRS that is composed of the known interactions. The numbers of common proteins are 2,390, 1,469, and 1,463 for text-network, NB network, and SVM network in comparison with the PRS, respectively. The overlap of interactions between the three derived networks and the PRS is 567 for text-network, 445 for NB network and SVM network, respectively. This shows that the proportion of true interactions on the SVM network is significantly higher than that on the text-network, and is slightly higher than that on the NB network (SVM network and NB network share most of proteins and interactions), suggesting that the SVM network has the highest confidence of these three network. The loss of proteins and interactions are caused by several factors. Many interactions in the PRS were detected by large-scale experiments so that the interacting proteins were commonly not written into abstracts of papers. Moreover, protein names can be recognized by NLProt with an only limited recall ($\sim$76%) (Mika and Rost 2004). As a result, the data size of our predicted PPIs is smaller than that in the analogous works for the genomic-scale interactome predictions, in which all of protein pairs were used to predict interactions. Compared with the previous works, however, our approach based on text-mining has a prominent advantage that each predicted PPI has at least a PubMed Identifier (PMID) by which uses can assess the confidence of the interaction by reading the corresponding paper. Moreover, two proteins existing in a same sentence have a much higher chance to interact with each other than those from a random protein pair; hence our approach can save more running time. However, the direct comparison with the PRS cannot provider the comprehensive estimation for our predicted networks because only a small fraction of proteins are shared between the predicted networks and the PRS. For this reason, we further compared the SVM network with random network. We paired proteins in the SVM network, and then a random network with equal number of protein pairs for the SVM network was constructed by extracting randomly protein pairs. The true interactions on the random network were counted by comparing the PRS. We repeated the process 100 times, and then calculated the mean of true interactions. Finally, we obtained only a mean of 3.4 true interactions on random networks, which is significantly smaller than that of the SVM network (455 true interactions). We also compared with the mouse PPI dataset from I2D/OPHID, a predicted dataset by mapping PPIs of several model organisms to mouse, and found that about 40.5% PPIs in the SVM

Table 3 The numbers of orthologous genes and conserved interactions between mouse and other model species PPI datasets

	HPRD		MS		Y2H-1		Y2H-2		Yeast		Fly		Worm	
	Gene	Interaction	Gene	Interaction	Gene	Interaction	Gene	Interaction	Gene	Interaction	Gene	Interaction	Gene	Interaction
PRS	1,158	1,221	24	15	7	3	45	23	169	104	41	32	31	17
SVM network	1,9911	3,651	55	80	24	16	61	43	292	278	65	67	66	99

PRS positive reference set, *HPRD* the Human Protein Reference Database, *MS* a human PPI dataset generated by high-throughput MS technology, *Y2H-1 and Y2H-2* two human PPI datasets created by high-throughput Y2H method, *worm, yeast and fly* the worm (*Caenorhabditis elegans*), yeast (*Sacchromyces cerevisiae*), and fly (*Drosophila melanogaster*) PPI datasets from the Database of Interacting Proteins (DIP)

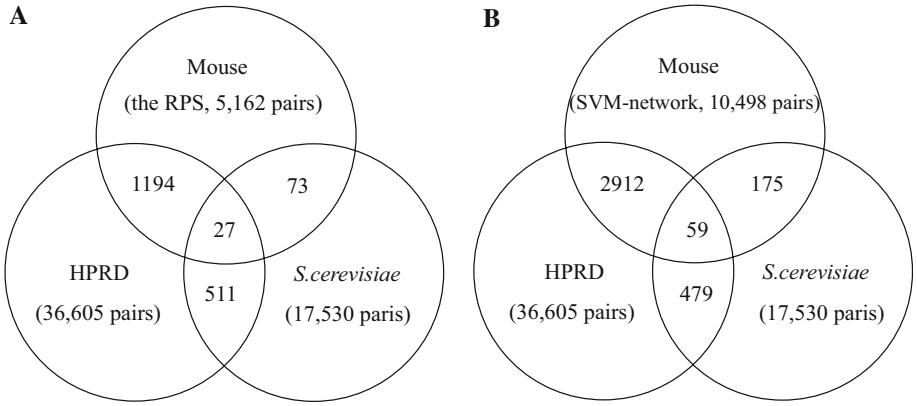

Fig. 5 Venn diagrams showing overlaps of PPI datasets from human and *S. cerevisiae* with the two mouse datasets. The overlap between human and *S. cerevisiae* datasets was compared independently by gene orthologs defined in the Inparanoid database. **a** The mouse PRS in which PPI data were obtained from several PPI databases. **b** The mouse PPI dataset of the SVM network

network are contained in the I2D/OPHID dataset. Overall, about 60% of predicted interactions are not present in other public-available large mouse interaction datasets, thus substantially increasing the coverage of the mouse interactome. Since text-network is obviously filled with a large number of false positives, and NB network and SVM network share most of interactions, only the SVM network was analyzed in the next section.

Interaction conservation between mouse and other model organisms

We surveyed the mouse PPIs that are conserved with those reported in the human, yeast, worm and fly datasets (about the data sources see Supplementary Table 2) in the final predicted network SVM network as well as the PRS. Table 3 lists the numbers of orthologous genes and conserved interactions between mouse and the four organisms PPI datasets. Surprisingly, the number of mouse protein interactions shared with other organisms is very small, even though in comparison with the literature-derived human dataset HPRD (only about 35%). We further examined the conserved PPIs cross-multiple organisms. As indicated by Venn diagrams showing the overlap between the mouse, human, and yeast interaction datasets (Fig. 5), only a very small fraction of interactions is conserved within the three species (27 PPIs for the PRS and 59 PPIs for SVM network). In the comparison, furthermore, we observed that no interaction is common to all of the five species. Our results are similar with the previous comparison that was performed in datasets from human, yeast, worm, and fly, and confirm the finding that the overlap among interaction datasets from different species is remarkably lacks (Gandhi et al. 2006). In addition, we at first carried out a comparison between datasets from human and mouse that both belong to mammalian species based on a limited accuracy and incomplete coverage, showing that human and mouse may only share a limited number of interactions, although they have short evolutionary distance (about 91 million years) (Hedges 2002) and significant sequence conservation at both protein-coding gene and whole genome levels.

Analysis of topological features

The topological parameters can be used to uncover potential bias in the PPI networks (Ramirez et al. 2007). A PIN can be modeled as a connectivity undirected graph, in which a node represents a protein and an edge is an interaction between two proteins. By using the NetworkAnalyzer plugin for Cytoscape (Shannon et al. 2003), we computed the degree distribution and clustering coefficient of the predicted network SVM network. By definition, the degree, k, of a given node is equal to the number of edges connected to it. Many biological, social, and technological networks are found to be scale-free that are characterized by a degree distribution, $P(k)$, that decays as a power law: $P(k) \sim 1/k^{\gamma}$ with decay exponent γ. In our analysis, the predicted network fits well a scale-free distribution (Fig. 6), and the value of decay exponent γ in SVM network is 2.03 which compares well with those of the yeast

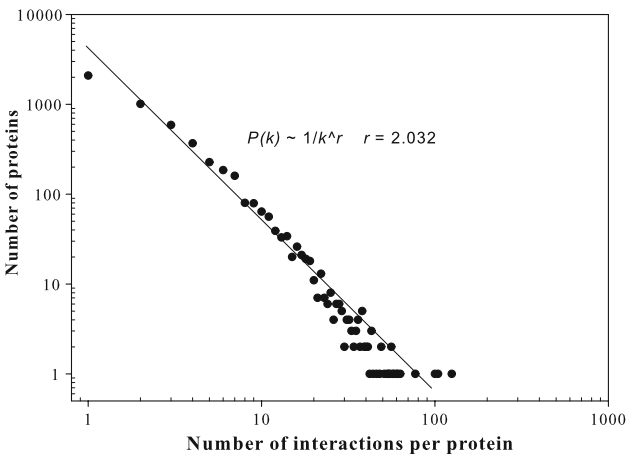

Fig. 6 Degree distribution of the SVM network

(1.8), worm (1.6), fly (2.0) (Li et al. 2006a) and humans (~2.0) (Barabasi and Oltvai 2004; Ramirez et al. 2007). The average clustering coefficient, a measure of interaction density, is 0.24 for SVM network which also is similar with those of other model organisms (Li et al. 2006a).

Application to functional inference of proteins

To explore the biological functions of mouse protein interactions, we grouped the PPIs of SVM network into ten categories according to the shared biological process for each interaction (Supplementary Table 5). The three top biological processes with the maximum number of interactions are DNA-dependent transcription, signal transduction and signal protein amino acid phosphorylation. We can further assign a more detailed function for each interaction by the shared more specific biological processes. Figure 7 shows a special sub-network corresponding to the Wnt receptor signaling pathway. The sub-network is composed of those interactions in which each shares the GO term "Wnt receptor signaling pathway" (GO: 0016055), and is clustered according to protein subcellular localization annotation using Cerebral (Barsky et al. 2007), a Java plugin for the Cytoscape. It is well known that the Wnt receptor signaling pathways are important for multiple biological processes during development and disease, and play a major role in bone cell differentiation, proliferation, apoptosis and tumorogenesis (for reviews see Clevers 2006; Gordon and Nusse 2006; van Amerongen and Berns 2006). Compared to the canonical Wnt signaling pathway in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Okuda et al. 2008), the sub-network shows that the most core interactions occurring in the pathway can be successfully detected by our approach. Some proteins and interactions that were validated to be involved in the canonical Wnt signaling pathway were added into the predicted network, though they were not annotated into the current KEGG Wnt signaling pathway. For instance, the protein Norrin (NDP) was found to activate the canonical Wnt signaling pathway through interacting with the Frizzled family receptor (FZD4) (Xu et al. 2004; Hendrickx and Leyns 2008); and the receptor tyrosine kinase ROR2 is a regulator of canonical Wnt pathway by interacting with WNT5A. However, a few unknown or false interactions are still remained, most of which are protein pairs both from the same protein family such as the Frizzled protein family (Winkel et al. 2008). It can be explained that the protein members from a same family always share a close gene expression profile, gene function similarity, and phenotype annotation so that the protein pairs both from a same family have a high composite LR.

Database and web service

We constructed a database named as Mouse protein–protein interaction DataBase (MppDB) for storing the predicted mouse PPIs, as well as assembling the public protein interaction data, and also provided a web-based query

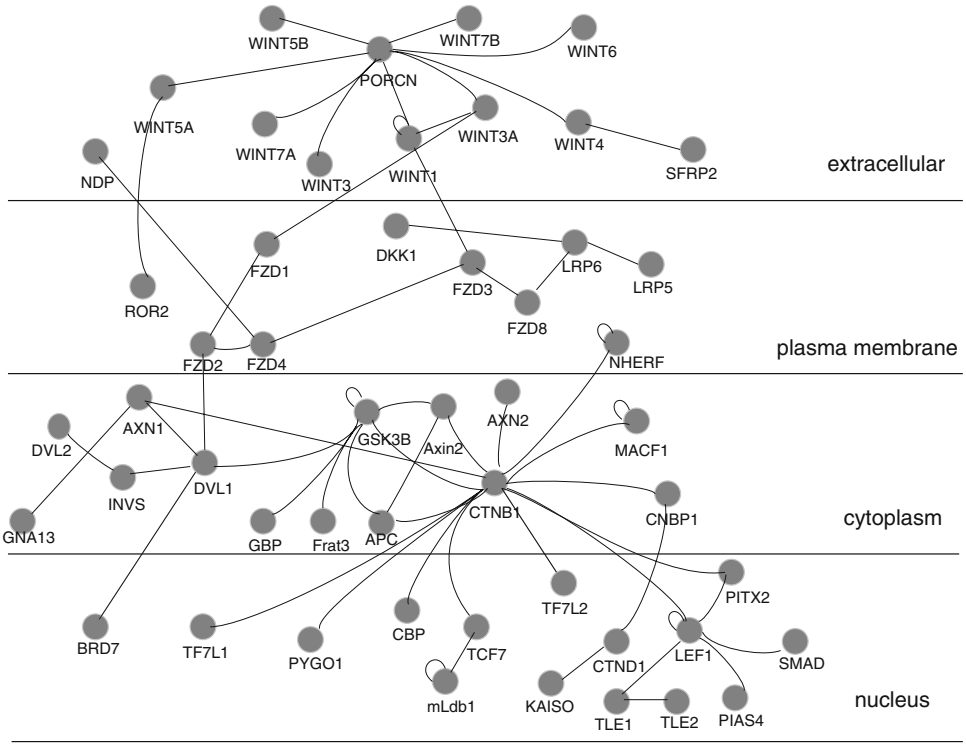

Fig. 7 A sub-network corresponding to the Wnt receptor signaling pathway. The subnet was extracted from SVM network, and composed of these proteins that all were annotated to GO biological process (GO: 0016055)

interface for users to search for potential PPIs among a list of query proteins or genes. The database is allowed to be queried by multiple IDs, including UniProtKB accession and Entrez GeneID. The database and web interface are publicly accessible at http://bio.scu.edu.cn/mppi. In addition, a sub-network centered on a query protein can be visualized online with a scalable vector graphics (SVG) file, so users can click other interesting nodes (proteins) to obtain more information about them through the sub-network (Supplementary Fig. 5). Three mouse PPIs datasets corresponding to the three networks (the PRS, NB network and SVM network) can also be downloaded as tab-delimited text files at our website.

Conclusions

In this study, we performed a three-step approach to derive a mouse PPI network. As each protein pair was extracted from the well-reviewed literatures and filtered by two prevalent approaches, the predicted protein interactions seem to be of good quality. The performances of the two prediction models (the naïve Bayesian and SVM model) were assessed based on the negative samples generated by the different methods. Moreover, the conditional independences between datasets were in detail assessed, showing many datasets are dependent or redundant though they are from different data sources. Note that the naïve Bayesian model can not only filter out the high confidence protein interactions mining from biomedical literatures, but also perform the genome-scale prediction in the future. We hope the public-available database (MppDB) can be a useful resource for the experimental study of the mouse interactome in future, and be helpful for further understanding on the molecular mechanisms of human disease at a system level.

Acknowledgments We are grateful to Dr. Han Hu for his assistance on the website construction. We also thank Prof. Yongsheng Liu, Dr. Bo Liu and Dr. Guan Song for reading the manuscript and their useful suggestions. This work was supported partially by Doctoral Fund of Ministry of Education of China (Grant No. 200806101013) and Important National Science & Technology Specific Projects of China (Grant No. 2009ZX10005-020).

References

Alfarano C, Andrade CE, Anthony K et al (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res 33:D418–D424. doi:10.1093/nar/gki051

Baldi P, Brunak S, Chauvin Y et al (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16:412–424. doi:10.1093/bioinformatics/16.5.412

Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101–113. doi:10.1038/nrg1272

Barrett T, Troup DB, Wilhite SE et al (2007) NCBI GEO: mining tens of millions of expression profiles—database and tools update. Nucleic Acids Res 35:D760–D765. doi:10.1093/nar/gkl887

Barsky A, Gardy JL, Hancock RE et al (2007) Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. Bioinformatics 23:1040–1042. doi:10.1093/bioinformatics/btm057

Beltrao P, Serrano L (2007) Specificity and evolvability in eukaryotic protein interaction networks. PLOS Comput Biol 3:e25. doi:10.1371/journal.pcbi.0030025

Ben-Hur A, Noble WS (2006) Choosing negative examples for the prediction of protein–protein interactions. BMC Bioinformatics 7(Suppl 1):S2. doi:10.1186/1471-2105-7-S1-S2

Berg J, Lassig M (2006) Cross-species analysis of biological networks by Bayesian alignment. Proc Natl Acad Sci USA 103:10967–10972. doi:10.1073/pnas.0602294103

Bowers PM, Pellegrini M, Thompson MJ et al (2004) Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol 5:R35. doi:10.1186/gb-2004-5-5-r35

Brown KR, Jurisica I (2007) Unequal evolutionary conservation of human protein interactions in interologous networks. Genome Biol 8:R95. doi:10.1186/gb-2007-8-5-r95

Chatr-aryamontri A, Ceol A, Palazzi LM et al (2007) MINT: the molecular INTeraction database. Nucleic Acids Res 35:D572–D574. doi:10.1093/nar/gkl950

Clevers H (2006) Wnt/beta-catenin signaling in development and disease. Cell 127:469–480. doi:10.1016/j.cell.2006.10.018

Cox RD, Brown SD (2003) Rodent models of genetic disease. Curr Opin Genet Dev 13:278–283. doi:10.1016/S0959-437X(03)00051-0

Cui J, Li P, Li G et al (2008) AtPID: *Arabidopsis thaliana* protein interactome database—an integrative platform for plant systems biology. Nucleic Acids Res 36:D999–D1008. doi:10.1093/nar/gkm844

Date SV, Stoeckert CJ Jr (2006) Computational modeling of the *Plasmodium falciparum* interactome reveals protein function on a genome-wide scale. Genome Res 16:542–549. doi:10.1101/gr.4573206

Ewing RM, Chu P, Elisma F et al (2007) Large-scale mapping of human protein–protein interactions by mass spectrometry. Mol Syst Biol 3:89. doi:10.1038/msb4100134

Formstecher E, Aresta S, Collura V et al (2005) Protein interaction mapping: a Drosophila case study. Genome Res 15:376–384. doi:10.1101/gr.2659105

Gandhi TK, Zhong J, Mathivanan S et al (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet 38:285–293. doi:10.1038/ng1747

Gavin AC, Bosche M, Krause R et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147. doi:10.1038/415141a

Ge H, Liu Z, Church GM et al (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat Genet 29:482–486. doi:10.1038/ng776

Giot L, Bader JS, Brouwer C et al (2003) A protein interaction map of *Drosophila melanogaster*. Science 302:1727–1736. doi:10.1126/science.1090289

Gordon MD, Nusse R (2006) Wnt signaling: multiple pathways, multiple receptors, and multiple transcription factors. J Biol Chem 281:22429–22433. doi:10.1074/jbc.R600015200

Guan Y, Myers CL, Lu R et al (2008) A genomewide functional network for the laboratory mouse. PLOS Comput Biol 4:e1000165. doi:10.1371/journal.pcbi.1000165

Harris MA, Clark JI, Ireland A et al (2006) The Gene Ontology (GO) project in 2006. Nucleic Acids Res 34:D322–D326. doi:10.1093/nar/gkj021

Hedges SB (2002) The origin and evolution of model organisms. Nat Rev Genet 3:838–849. doi:10.1038/nrg929

Hendrickx M, Leyns L (2008) Non-conventional frizzled ligands and Wnt receptors. Dev Growth Differ 50:229–243

Ho Y, Gruhler A, Heilbut A et al (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415:180–183. doi:10.1038/415180a

Hovatta I, Tennant RS, Helton R et al (2005) Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. Nature 438:662–666. doi:10.1038/nature04250

Huang TW, Lin CY, Kao CY (2007) Reconstruction of human protein interolog network using evolutionary conserved network. BMC Bioinformatics 8:152. doi:10.1186/1471-2105-8-152

Ito T, Chiba T, Ozawa R et al (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 98:4569–4574. doi:10.1073/pnas.061034498

Jansen R, Yu H, Greenbaum D et al (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. Science 302:449–453. doi:10.1126/science.1087361

Jenssen TK, Laegreid A, Komorowski J et al (2001) A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 28:21–28. doi:10.1038/88213

Kerrien S, Alam-Faruque Y, Aranda B et al (2007) IntAct–open source resource for molecular interaction data. Nucleic Acids Res 35:D561–D565. doi:10.1093/nar/gkl958

Lehner B, Fraser AG (2004) A first-draft human protein-interaction map. Genome Biol 5:R63. doi:10.1186/gb-2004-5-9-r63

Lemos B, Meiklejohn CD, Hartl DL (2004) Regulatory evolution across the protein interaction network. Nat Genet 36:1059–1060. doi:10.1038/ng1427

Li S, Armstrong CM, Bertin N et al (2004) A map of the interactome network of the metazoan *C. elegans*. Science 303:540–543. doi:10.1126/science.1091403

Li D, Li J, Ouyang S et al (2006a) Protein interaction networks of *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*: large-scale organization and robustness. Proteomics 6:456–461. doi:10.1002/pmic.200500228

Li S, Wu L, Zhang Z (2006b) Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. Bioinformatics 22:2143–2150. doi:10.1093/bioinformatics/btl363

Li D, Liu W, Liu Z et al (2008) PRINCESS, a protein interaction confidence evaluation system with multiple data sources. Mol Cell Proteomics 7:1043–1052. doi:10.1074/mcp.M700287-MCP200

Lu LJ, Xia Y, Paccanaro A et al (2005) Assessing the limits of genomic data integration for predicting protein networks. Genome Res 15:945–953. doi:10.1101/gr.3610305

Matthews LR, Vaglio P, Reboul J et al (2001) Identification of potential interaction networks using sequence-based searches for conserved protein–protein interactions or "interologs". Genome Res 11:2120–2126. doi:10.1101/gr.205301

McDermott J, Guerquin M, Frazier Z et al (2005) BIOVERSE: enhancements to the framework for structural, functional and contextual modeling of proteins and proteomes. Nucleic Acids Res 33:W324–325. doi:10.1093/nar/gki401

Mika S, Rost B (2004) NLProt: extracting protein names and sequences from papers. Nucleic Acids Res 32:W634–637. doi:10.1093/nar/gkh427

Mishra GR, Suresh M, Kumaran K et al (2006) Human protein reference database—2006 update. Nucleic Acids Res 34:D411–D414. doi:10.1093/nar/gkj141

Ng SK, Zhang Z, Tan SH et al (2003) InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. Nucleic Acids Res 31:251–254. doi:10.1093/nar/gkg079

O'Brien KP, Remm M, Sonnhammer EL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res 33:D476–D480. doi:10.1093/nar/gki107

Okuda S, Yamada T, Hamajima M et al (2008) KEGG Atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res 36:W423–426. doi:10.1093/nar/gkn629

Pagel P, Kovac S, Oesterheld M et al (2005) The MIPS mammalian protein–protein interaction database. Bioinformatics 21:832–834. doi:10.1093/bioinformatics/bti115

Ramani AK, Bunescu RC, Mooney RJ, 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. doi:10.1186/gb-2005-6-5-r40

Ramirez F, Schlicker A, Assenov Y et al (2007) Computational analysis of human protein interaction networks. Proteomics 7:2541–2552. doi:10.1002/pmic.200600924

Rhodes DR, Tomlins SA, Varambally S et al (2005) Probabilistic model of the human protein–protein interaction network. Nat Biotechnol 23:951–959. doi:10.1038/nbt1103

Rosenthal N, Brown S (2007) The mouse ascending: perspectives for human-disease models. Nat Cell Biol 9:993–999. doi:10.1038/ncb437

Rual JF, Venkatesan K, Hao T et al (2005) Towards a proteome-scale map of the human protein–protein interaction network. Nature 437:1173–1178. doi:10.1038/nature04209

Salwinski L, Miller CS, Smith AJ et al (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32:D449–D451. doi:10.1093/nar/gkh086

Shannon P, Markiel A, Ozier O et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504. doi:10.1101/gr.1239303

Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. Mol Syst Biol 3:88. doi:10.1038/msb4100129

Shen J, Zhang J, Luo X et al (2007) Predicting protein–protein interactions based only on sequences information. Proc Natl Acad Sci USA 104:4337–4341. doi:10.1073/pnas.0607879104

Smith CL, Goldsmith CA, Eppig JT (2005) The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol 6:R7. doi:10.1186/gb-2004-6-1-r7

Sprenger J, Lynn Fink J, Karunaratne S et al (2008) LOCATE: a mammalian protein subcellular localization database. Nucleic Acids Res 36:D230–D233. doi:10.1093/nar/gkm950

SPSS I (1999) SPSS Base 10.0 User's Guide. SPSS, Inc., Chicago

Stapley BJ, Benoit G (2000) Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. Pac Symp Biocomput 529–540

Stein A, Russell RB, Aloy P (2005) 3did: interacting protein domains of known three-dimensional structure. Nucleic Acids Res 33:D413–D417. doi:10.1093/nar/gki037

Stelzl U, Worm U, Lalowski M et al (2005) A human protein–protein interaction network: a resource for annotating the proteome. Cell 122:957–968. doi:10.1016/j.cell.2005.08.029

Su AI, Wiltshire T, Batalov S et al (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 101:6062–6067. doi:10.1073/pnas.0400782101

Suzuki H, Fukunishi Y, Kagawa I et al (2001) Protein–protein interaction panel using mouse full-length cDNAs. Genome Res 11:1758–1765. doi:10.1101/gr.180101

Tsaparas P, Marino-Ramirez L, Bodenreider O et al (2006) Global similarity and local divergence in human and mouse gene co-expression networks. BMC Evol Biol 6:70. doi:10.1186/1471-2148-6-70

Uetz P, Giot L, Cagney G et al (2000) A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. Nature 403:623–627. doi:10.1038/35001009

van Amerongen R, Berns A (2006) Knockout mouse models to study Wnt signal transduction. Trends Genet 22:678–689. doi:10.1016/j.tig.2006.10.001

Vapnik V (2005) The nature of statistical learning theory. Springer, New York

von Mering C, Krause R, Snel B et al (2002) Comparative assessment of large-scale data sets of protein–protein interactions. Nature 417:399–403. doi:10.1038/nature750

Waterston RH, Lindblad-Toh K, Birney E et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562. doi:10.1038/nature01262

Winkel A, Stricker S, Tylzanowski P et al (2008) Wnt-ligand-dependent interaction of TAK1 (TGF-beta-activated kinase-1) with the receptor tyrosine kinase Ror2 modulates canonical Wnt-signalling. Cell Signal 20:2134–2144. doi:10.1016/j.cellsig.2008.08.009

Witten IH, Frank E (2000) Data mining: practical machine learning techniques with Java implementations. Morgan Kaufmann, San Francisco

Wuchty S, Ipsaro JJ (2007) A draft of protein interactions in the malaria parasite *P. falciparum*. J Proteome Res 6:1461–1470. doi:10.1021/pr0605769

Wuchty S, Barabasi AL, Ferdig MT (2006) Stable evolutionary signal in a yeast protein interaction network. BMC Evol Biol 6:8. doi:10.1186/1471-2148-6-8

Xia K, Dong D, Han JD (2006) IntNetDB v1.0: an integrated protein–protein interaction network database generated by a probabilistic model. BMC Bioinformatics 7:508. doi:10.1186/1471-2105-7-508

Xu Q, Wang Y, Dabdoub A et al (2004) Vascular development in the retina and inner ear: control by Norrin and Frizzled-4, a high-affinity ligand-receptor pair. Cell 116:883–895. doi:10.1016/S0092-8674(04)00216-8

Xuan Z, Wang J, Zhang MQ (2003) Computational comparison of two mouse draft genomes and the human golden path. Genome Biol 4:R1. doi:10.1186/gb-2002-4-1-r1

Yellaboina S, Dudekula DB, Ko M (2008) Prediction of evolutionarily conserved interologs in Mus musculus. BMC Genomics 9:465. doi:10.1186/1471-2164-9-465