SCIENCE CHINA
Life Sciences

• RESEARCH PAPER •

June 2015 Vol.58 No.6: 579–589
doi: 10.1007/s11427-014-4766-3

PathPPI: an integrated dataset of human pathways and protein-protein interactions

TANG HaiLin[1†], ZHONG Fan[2†], LIU Wei[1], HE FuChu[2,3*] & XIE HongWei[1*]

[1]*College of Mechanical & Electronic Engineering and Automatization, National University of Defense Technology, Changsha 410073, China;*
[2]*Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China;*
[3]*State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing 102206, China*

Received April 5, 2014; accepted July 20, 2014; published online January 14, 2015

Integration of pathway and protein-protein interaction (PPI) data can provide more information that could lead to new biological insights. PPIs are usually represented by a simple binary model, whereas pathways are represented by more complicated models. We developed a series of rules for transforming protein interactions from pathway to binary model, and the protein interactions from seven pathway databases, including PID, BioCarta, Reactome, NetPath, INOH, SPIKE and KEGG, were transformed based on these rules. These pathway-derived binary protein interactions were integrated with PPIs from other five PPI databases including HPRD, IntAct, BioGRID, MINT and DIP, to develop integrated dataset (named PathPPI). More detailed interaction type and modification information on protein interactions can be preserved in PathPPI than other existing datasets. Comparison analysis results indicate that most of the interaction overlaps values (O_{AB}) among these pathway databases were less than 5%, and these databases must be used conjunctively. The PathPPI data was provided at http://proteomeview.hupo.org.cn/PathPPI/PathPPI.html.

databases integration, pathway, protein-protein interaction, proteomics

Citation: Tang HL, Zhong F, Liu W, He FC, Xie HW. PathPPI: an integrated dataset of human pathways and protein-protein interactions. Sci China Life Sci, 2015, 58: 579−589, doi: 10.1007/s11427-014-4766-3

It is increasingly clear that biological functions are mainly performed through diverse protein interactions. As such, studies on protein interactions have become important undertakings over the last few years [1–4]. Protein interactions can be mainly obtained from two types of databases: pathway and protein-protein interaction (PPI). Pathway databases contain more diverse interactions, such as signalling, transcription regulation and metabolism, which generally present a clear *in vivo* interaction type and higher confidence. By contrast, the PPIs from high throughput experiments always feature vague *in vivo* information and lower confidence.

Studies have shown that large, redundant and complementary information exists in pathway and PPI databases [5,6]. Integration of these pathways will result in more information that could lead to new biological insights [7]. For example, Ahn et al. [8], who explored the molecular mechanism and biomarker identification of prostate cancer by integrating diverse pathway and PPI databases, suggested that an integrated network could provide more detailed and interpretable roles of cancer-related genes in prostate cancer cells. Kirouac et al. [9] explored the inflammatory network structure by integrating the GeneGo, Cell MAP, PID and Reactome pathway databases and two other PPI databases. Several attempts have been made to develop databases that integrate pathway and PPI data, such as CPDB [10] and Pathway Commons [11] recently, while previous work fo-

†Contributed equally to this work
*Corresponding author (email: hefc@nic.bmi.ac.cn; xhwei_65@163.com)

cused only on PPI datasets [12,13].

Many challenges remain in terms of the integration of heterogeneous data from pathways and PPIs because such data are generally represented by different standards. PPIs are usually represented by a simple binary model, such as SIF (simple interaction format) and PSI MITAB [14] no matter how they were generally captured as *n*-ary (i.e., complex) or binary data, whereas pathways are represented by more complicated models, such as BioPAX [15], CellML [16] and SBML [17]. Although several models have been previously reported, the transformation method must be standardised to allow representation of integrated interactions to become more unified and convenient. In this study, we first build a binary model, and develop a series of rules to transform protein interaction from pathway to binary models. Some important information of interactions that were always neglected in earlier studies can be preserved in our binary model. Then, we integrate seven pathway and five PPI databases, and name this integrated dataset PathPPI, which can be obtained from http://proteomeview.hupo.org.cn/PathPPI/PathPPI.html.

1 Materials and methods

1.1 Pathway and PPI databases

We mainly focused on pathway databases with the widely used BioPAX standard. In this study, six BioPAX-modelled pathway databases, including PID (contains only the data curated by PID, version 2012.03.17), BioCarta (from PID, version 2010.08.11), Reactome (version 2012.03.14) [18], NetPath (downloaded 2012.04.14) [19], INOH (version 2011.01.31) [20] and SPIKE (version 2011.03.22) [21], were utilized. The non-metabolic pathway portion of KEGG (downloaded 2009.10.12) [22] was also included because of its large data size. The five integrated PPI databases included HPRD (version R9) [23], IntAct (downloaded 2013.10.22) [24], BioGRID (version 3.2.105) [25], MINT (version 2013.03.26) [26] and DIP (version 2013.07.07) [27]. Integrated databases contained STRING (version 9.1) [28], Pathway common (downloaded 2014.06.14) and CPDB (downloaded 2014.06.14).

1.2 Categorisation of PathPPI

The interaction type of PathPPIs must be depicted in a unified manner. Although PSI-MI provides molecular interaction ontology for PPIs, it cannot fully depict PathPPIs, such as its structured way of indicating the outcome of an interaction. We developed a new categorisation based on BioPAX Level 3 standards and provided eight types of PathPPI interactions: BiochemicalReactionRegulation (BRR), TransportRegulation (TR), TransportWithBiochemicalReactionRegulation (TBRR), ComplexAssemblyRegulation (CAR), ExpressionRegulation (ER), ComplexAssemblyInteraction (CAI), GeneticInteraction (GI) and MolecularInteraction (MI) (Table 1). The meaning of each types can be obtained through the BioPAX standard [15] and the following transferring rules. Five regulation interactions, including BRR, TR, TBRR, CAR and ER, contain an effect parameter that denotes regulation effects by Activation, Inhibition or Unspecified. BRR and TBRR also contain a modification parameter. A total of 22 common modifications with corresponding de-modifications were included in this work (Table S1 in Supporting Information). Several other rare modifications, such as cholesterol modification, will be added in further versions of our categorisation.

The six PathPPI types, BRR, TR, TBRR, CAR, ER and CAI, are BiolPPIs depicted from a biological perspective and generally with clear *in vivo* mechanisms. In our categorisation, GI and MI are treated as TechPPIs since they are produced by certain PPI detection technologies and without clear *in vivo* information.

1.3 Transformation of protein interactions from BioPAX into a binary model

BioPAX (Biological Pathway Exchange, http://www.biopax.org/) is a standard language to represent biological pathways at the molecular and cellular level and to facilitate the exchange of pathway data by defining an open file format specification for the exchange of biological pathway data. BioPAX covers all major concepts familiar to biologists studying pathways, including metabolic and signaling pathways, gene regulatory networks and genetic and molec-

Table 1 Categorisation of PathPPIs[a)]

	PathPPI	Effect*	Modification**	Directionality
BiolPPI	BiochemicalReactionRegulation (BRR)	●	●	Directed
	TransportRegulation (TR)	●		Directed
	TransportWithBiochemicalReactionRegulation (TBRR)	●	●	Directed
	ComplexAssemblyRegulation (CAR)	●		Directed
	ExpressionRegulation (ER)	●		Directed
	ComplexAssemblyInteraction (CAI)			Undirected
TechPPI	GeneticInteraction (GI)			Undirected
	MolecularInteraction (MI)			Undirected

a) *, With three status: Activation, Inhibition and Unspecified. **, With 22 pairs of modification currently (Table S1 in Supporting Information).

ular interactions. The BioPAX language uses a discrete representation of biological pathways. Dynamic and quantitative aspects of biological processes, including temporal aspects of feedback loops and calcium waves, are not supported. BioPAX Level 3 defines five types of molecular interactions: Control, Conversion, GeneticInteraction, MolecularInteraction and TemplateReaction (Figure 1A).

Conversion represents reactions in which one or more entities are physically transformed into other entities. The entities of Conversion can be protein, complex, DNA, RNA or small molecule, and we only focused on protein or complex. Conversion has five subclasses: BiochemicalReaction, Transport, TransportWithBiochemicalReaction, Degradation and ComplexAssembly. The input and output entities of BiochemicalReaction, Transport and TransportWithBiochemicalReaction are always the same proteins but with different modification or subcellular status. Degradation generally has no protein product. Thus, we discarded these four types of interaction and dealt only with ComplexAssembly. Each input entity was defined to have a CAI with each output entity in the binary model (Figure 1B). Given that ComplexAssembly representations are always reversible, the transformed CAI was non-directional. Thus, A has a CAI interaction with B, suggesting that A can produce B through non-covalent interactions with other molecules or through the decomposition of A and vice versa.

Control contains three subclasses: Catalysis, Modulation and TemplateReactionRegulation. Modulation describes an interaction in which a small molecule alters the ability of an enzyme to catalyze a specific reaction. Thus, Modulation was discarded because our focus was protein interactions. Catalysis is a type of interaction in which a physical entity (a catalyst) accelerates a Conversion interaction by lowering its activation energy. For Catalysis, we defined the controller have interactions with each output entity of the catalyzed Conversion interaction. The transformed PathPPI type depends on the Conversion interaction type. For example, if the Conversion interaction is BiochemicalReaction in BioPAX, the transformed type will be BiochemicalReaction-Regulation in PathPPI (Figure 1C). Effect and modification

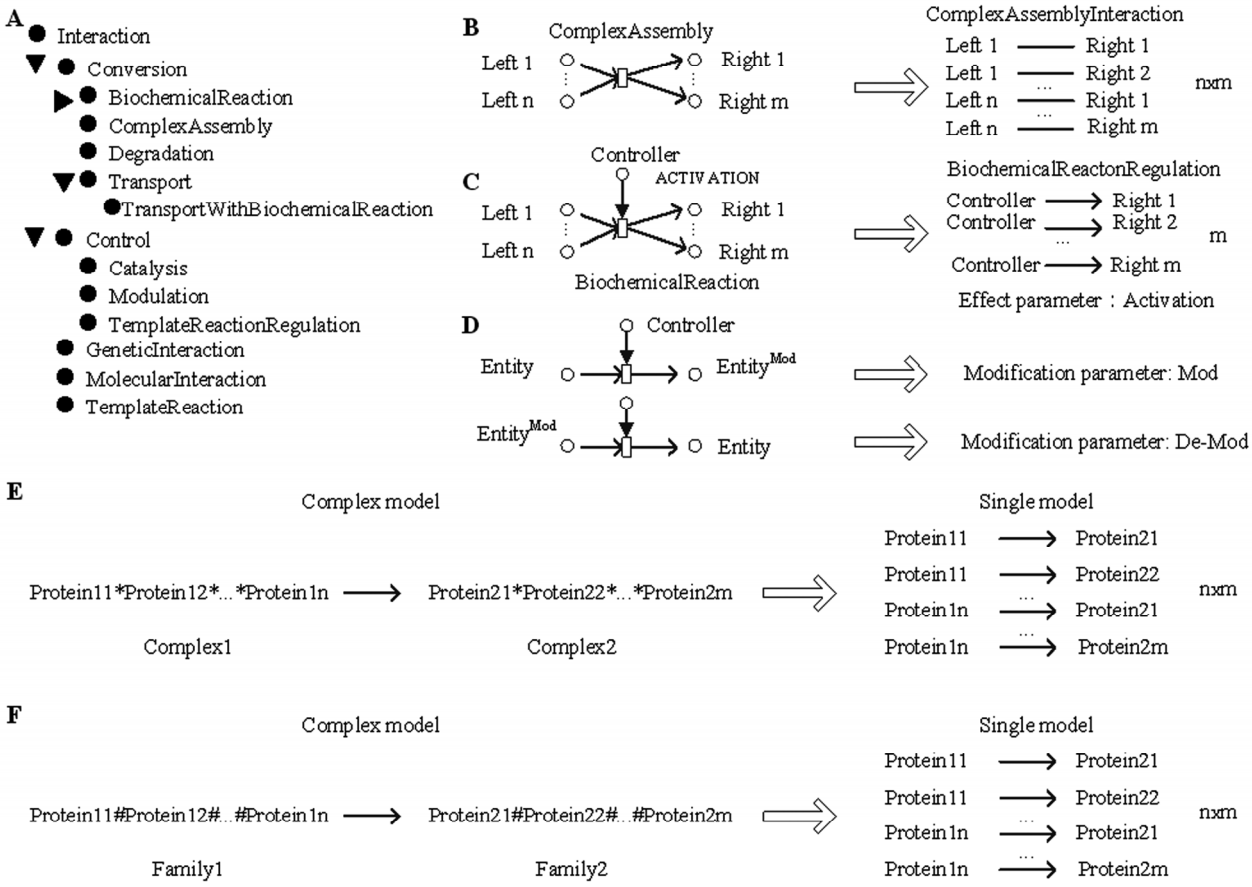

Figure 1 Illustration of the transformation from BioPAX to PathPPI model. A, BioPAX Level 3 contains five types of molecular interactions. Control and Conversion have subclasses. B, For ComplexAssembly, we specified that each input entity has a CAI with each output entity in the PathPPI model. C, For Control, we defined five types of interactions between the controller and products of controlled interaction for each controlled Conversion interaction. For example, if the controlled interaction is BiochemicalReaction, the controller has a BiochemicalReactionRegulation interaction with each product. Effect parameters can be obtained from controlType. D, Modification parameters can be obtained by comparing the modification state of the entity before and after reaction. E and F, Complex model allows PathPPI entities to be complexes or families, in contrast to a single model where each protein from one entity in the complex model interacts with each protein from another entity.

parameters can be obtained from the controlType and interactionType items of Catalysis, respectively. All six pathway databases provide controlType but only NetPath and INOH provide interactionType (Table S2 in Supporting Information). Other than the interactionType approach, we obtained modification information from the changes of protein modification states during BiochemicalReaction in PID and Reactome. For example, if a protein is initially without modification but with phosphorylation after BiochemicalReaction, then we assigned the modification parameter Phosphorylation. For simplicity, in current version we only extracted modification information for the BiochemicalReaction interactions, in which there is only one input entity and its corresponding modification state is output or vice versa (Figure 1D). The modification information of the BiochemicalReaction interactions that contain multiple entities were not extracted. Modifications in PID and Reactome were identified by MOD ID [29]. We developed mapping relationships between MOD IDs and PathPPI modification parameters (Table S2 in Supporting Information). BioCarta and SPIKE do not provide modification information. The rules for the Catalysis with other types of Conversion subclasses were similar to that with BiochemicalReaction. TemplateReactionRegulation depicts an interaction in which the controller regulates a TemplateReaction (gene expression) interaction in the BioPAX model. The controller (transcription factor) was defined to have an ER interaction with the products (target genes) of gene expression.

GeneticInteraction and MolecularInteraction are represented in binary form in BioPAX; thus, they were reserved in PathPPI categorisation. Although no GI was extracted from any of the six databases, we reserved GI in the PathPPI frame for further upgrades. TemplateReaction was ignored because its input and output are always the same. Considering that we focused only on the protein interactions, we discarded the interactions containing interactor of small molecule, non-coding DNA or RNA. If a complex contained small molecules, the proteins were preserved while small molecules were ignored.

1.4 Transforming protein interactions from KGML into a binary model

Given that KGML (KEGG Markup Language) adopts a binary-pair model, we can map KGML to PathPPI categories through lexical transformation. KGML categorisation is a two-level hierarchical structure containing four top-level categories: ECrel, PCrel, GErel and PPrel. ECrel depicts an enzyme-enzyme relation that two enzymes catalysing successive reaction steps. PCrel depicts protein-compound interactions. These two types of interaction were discarded. GErel represents the relation of transcription factor and target gene product. GErel has four interaction sub-types: repression, expression, indirect effect and missing interaction. The missing interaction depicts missing interactions in mutation, and it was ignored. The three remaining subtypes have corresponding types in PathPPI. PPrel is the most complicated category in KGML and its subtypes can be classified into three groups. The effect group contains two subtypes, activation and inhibition. The biological event group contains compound, phosphorylation, dephosphorylation, methylation, demethylation, glycosylation, ubiquitination, binding/association and dissociation. Three other subtypes, including state change, missing interaction and indirect effect, were classified as a third group. We developed a mapping relationship between PathPPI categories and all PPrel subtypes except for compound, dissociation, state change and missing interaction. We can obtain PathPPI effects and modification parameters based on the effect and biological event groups, respectively. For example, the interactions of FYN and MAP4K2 are annotated with activation and phosphorylation in KGML, and we assigned the effect and modification parameters of this interaction Activation and Phosphorylation in PathPPI model. More mapping relationships are shown in Table S3 in Supporting Information.

1.5 Complex and single models of PathPPIs

The entities (nodes) of interaction in current pathway databases can be complex or family (that is, the node represents multiples proteins), or their more sophisticated combinations, differing from traditional PPIs where the entities are only proteins. Thus, two types of model were proposed in PathPPI to represent binary interactions with complex or family entities (Figure 1E and F): the complex model, which allows the entities of PathPPIs to be complexes or families, and the single model, which assumes that all the proteins in a complex or family participate in interactions. For example, if a complex with three proteins interacts with another complex with two proteins in the complex model, the corresponding single model will represent 3×2=6 interactions. The complex model is more accurate in depicting real protein interactions, whereas the single model is more convenient for network analysis.

1.6 Unification of the protein identifier

Different databases sometimes use different naming systems for gene names. KEGG (human) has its object identifiers (hsa). HPRD uses official gene symbol. BioGRID and SPIKE use gene ID, while the remaining databases (PID, BioCarta, Reactome, NetPath, INOH, IntAct, MINT and DIP) all adopt UniProt AC. For the purpose of reducing redundancy in gene level, we transformed all other heterogeneous IDs to gene symbol utilizing their built-in id mapping systems.

1.7 Confidence scores of PathPPI

An existing method, Intscore (http://intscore.molgen.mpg.de/), was used to assign a confidence score for each pairs. IntScore is a web server for confidence scoring of biological interactions. It provides six methods for confidence scoring, as well as the possibility to integrate the method-specific scores. We submitted all PathPPI (single model) to Intscore (default parameter) and obtained their integrated scores for each pairs.

2 Results

Seven pathway databases were transformed to extract a large number of binary interactions. We combined these pathway-derived binary interactions with traditional PPIs from five other PPI databases into PathPPIs. After the transformation of the seven pathway databases, we obtained seven binary interactions data sets, the proteins of which are identified by gene symbols or gene names if there are no corresponding gene symbols. PID, BioCarta, Reactome and INOH only contained biological PPIs (BiolPPIs), whereas KEGG, NetPath and SPIKE contained technical PPIs (TechPPIs) (for depiction of BiolPPI and TechPPI see the section of "Categorisation of PathPPI"). In addition, we obtained 169,203 non-redundant TechPPIs from the five other PPI databases (HPRD, IntAct, BioGRID, MINT and DIP). Finally, PathPPI integrated 22,737 BiolPPIs and 174,770 TechPPIs in the complex model, and involved 16,768 human genes. The BiolPPI part contained 10,627 BRRs, 239 TRs, 132 TBRRs, 948 CARs, 3,938 ERs and 6,853 CAIs (Table 2).

2.1 Protein and interaction scales in BiolPPI datasets

First, we searched for the total number of unique proteins represented in the BiolPPI datasets. Results showed that SPIKE contains the highest number of proteins, reaching as high as 3,578. PID, Reactome and KEGG, which have similar scales, contained 2,666, 2,580 and 2,489 proteins, respectively. BioCarta, INOH and NetPath contained 1,436, 843 and 385 proteins, respectively. The total number of BiolPPI proteins was 7,012, covering 1/3 of all human protein-coding genes. For entities that can be considered family or complex in the complex model, we examined the proportion of complexes and families occupying whole entities in each dataset. Results showed that there are high complex proportions in INOH (56.44%), PID (51.73%), Reactome (45.77%) and BioCarta (40.54%), as well as high protein family proportions in INOH (41.33%), KEGG (32.65%) and Reactome (28.80%) (Table S4 in Supporting Information).

SPIKE contributed the most BRRs (4,504) and ERs (1,776) to PathPPI (complex model). PID contributed the highest number of other BiolPPI types, the second most BRR (2,160), and ERs (1,700) (Table 2). Reactome, NetPath and SPIKE did not have CAI and CAR interactions because Reactome and NetPath treat the ComplexAssembly as BiochemicalReaction and SPIKE does not provide ComplexAssembly in its BioPAX file, suggesting that there exists a large difference in the usage of BioPAX standards. NetPath and INOH did not have ER interactions. These differences show that different pathway databases build pathways with different details. The number of interactions expanded sharply with the single model, especially in Reactome and INOH, because of their high complex and family proportions (Table S5 in Supporting Information).

The proportions of the effect parameter Activation are higher than those of Inhibition in all datasets but these proportions were variant. The Activation:Inhibition ratio was approximately 20:1 in PID but less than 2:1 in SPIKE (Table S6 in Supporting Information). Interestingly, many BRRs existed with conflicting annotations in KEGG and SPIKE. For example, the AKT family (AKT1, AKT2 and AKT3) have both activation and inhibition interactions with CHUK, IKBKB and IKBKG. About 85% of the modification parameters were Phosphorylation, followed by Dephosphorylation (~7%), Ubiquitination (~6%) and Acetylation (~1%). No modification information was obtained from SPIKE, BioCarta and Reactome (Table S6 in Supporting Information).

Table 2 BiolPPI composition of the seven BiolPPI datasets (Complex model)[a)]

Dataset	BRR	TR	TBRR	CAR	ER	CAI
PID	2,160	200	131	709	1,700	5,195
BioCarta	1,149	46	1	256	242	1,143
Reactome	1,519	0	0	0	237	0
NetPath	582	0	0	0	0	0
INOH	198	2	0	1	0	691
KEGG	1,823	0	0	0	284	0
SPIKE	4,504	0	0	0	1,776	0
BiolPPI	10,627	239	132	948	3,938	6,853

a) It was noted that Reactome and NetPath treat the ComplexAssembly as BiochemicalReaction and SPIKE does not provide ComplexAssembly in its BioPAX file.

2.2 Overlaps among the BiolPPI datasets

We used the index O_{AB} to measure the overlap of sets A and B, which is equal to the arithmetic average of the ratios of the element numbers of $(A \cap B)$ divided by sets A and B, respectively [30]. The protein O_{AB} values of each two BiolPPI data sets ranged from 35.74% to 51.29% among the seven pathway datasets, whereas most of the interaction O_{AB} values were less than 5.00% in the complex model (Table 3) and also low in the single model (Table S7 in Supporting Information). These findings indicate that high protein overlaps exist in these datasets but the interaction overlaps are unexpectedly low, considering their common

origin of being curated from the literature. This suggests that the coverage of each pathway database is far from complete.

2.3 Overlaps between TechPPIs and BiolPPIs

TechPPIs were mainly from high throughput technologies with high false positive rates, while BiolPPIs were generally produced by molecular biological experiments and could be served as real *in vivo* interactions. Here, we calculated the overlaps of different types of BiolPPIs and TechPPIs, which can represent the real *in vivo* interaction-detecting capabilities of the high-throughput technologies in some way. We also defined sensitivity to measure the detecting capability of the detection technology, which is equal to the ratio of the element number of (BiolPPI ∩ TechPPI) divided by BiolPPI group.

BiolPPIs were classified based on their interaction types. TechPPIs annotated with 'two hybrid' (Y2H), 'anti bait co-immunoprecipitation' (ABCoIP) and 'affinity technology' (AffTech) were selected as three technology groups. In addition, interactions with '*in vivo*' annotation from HPRD that were mainly detected by small-scale experiments were treated as a control technology group.

Results showed that only a small part of the TechPPIs overlapped with BiolPPIs (Table 4), suggesting that the biological functions of most TechPPIs remain unclear, or the high-throughput experiments have surprisingly high false positive rate. In addition, all of the sensitivity values of three technology groups were very low.

2.4 Top degree hub proteins in BiolPPI

In the signalling network (BRR) of BiolPPI, ATM had the highest out-degree (number of directed downstream entities), reaching as high as 849, far more than any other entity. RHOA and RAC1 had the highest in-degree (number of directed upstream entities). In the transcription regulation network (ER) of BiolPPI, TP53 had the highest out-degree (174), whereas KLK3 had the highest in-degree (77). TP53 also has the 3rd highest of in-degree in signalling network, indicating it is a vital transcription factor of jointing the end of signalling cascade and the target genes (Table S8 in Supporting Information). The top-degree proteins in TR, TBRR, CAR and CAI networks were shown in Tables S9–S12.

We also examined the contributions of these degrees from the original seven pathway databases. Many of these

Table 3 Protein and interactionsoverlaps among the seven BiolPPI datasets

	PID 9,898	BioCarta 2,790	Reactome 1,739	NetPath 581	INOH 892	KEGG 2,249	SPIKE 6,164	Interaction scale[a)]
PID: 2666		335 (7.70%)	59 (1.99%)	77 (7.02%)	75(4.58%)	126 (3.44%)	390 (5.13%)	PID: 9,898
BioCarta: 1436	879(47.09%)		23 (1.07%)	52 (5.41%)	40(2.96%)	82 (3.29%)	165 (4.30%)	BioCarta: 2,790
Reactome: 2580	988(37.68%)	677 (36.69%)		9 (1.03%)	10(0.85%)	45 (2.29%)	52 (1.92%)	Reactome: 1,739
NetPath: 385	325 (48.30%)	217 (35.74%)	254 (37.91%)		26 (3.69%)	40 (4.33%)	349 (32.87%)	NetPath: 581
INOH: 843	619 (48.32%)	418 (39.35%)	459 (36.12%)	210 (39.73%)		47 (3.68%)	52 (3.34%)	INOH: 892
KEGG: 2489	1,194 (46.38%)	730 (40.08%)	993 (39.19%)	283 (42.44%)	646 (51.29%)		383 (11.62%)	KEGG: 2,249
SPIKE: 3578	1,490 (48.77%)	850(41.47%)	1,136 (37.89%)	336 (48.33%)	582(42.65%)	1,193 (40.64%)		SPIKE: 6,164
Protein scale	PID 2,666	BioCarta 1,436	Reactome 2,580	NetPath 385	INOH 843	KEGG 2,489	SPIKE 3,578	

a) Different types of interaction linking two identical entities were counted only once. O_{AB} values are in brackets. The upper right triangle part depicts the overlaps of interactions and the lower left triangle part depicts the overlaps of proteins.

Table 4 Overlaps of interactions among different types of BiolPPIs and TechPPIs (single model)[a)]

BiolPPI group	TechPPI group				
	Y2H 33901	ABCoIP 12766	AffTech 81860	*In vivo* 19118	AllPPI 65737
BRR: 71,535	1130 (1.58%)	783 (1.09%)	3,754 (5.52%)	3,662 (5.12%)	8,070 (11.28%)
TBRR: 358	23 (6.42%)	23 (6.42%)	55 (15.36%)	64 (17.88%)	110 (30.73%)
TR: 784	30 (3.83%)	26 (3.32 %)	116 (14.80%)	76 (9.69%)	195 (24.87%)
CAR: 4,375	130 (2.97%)	93 (2.13%)	462 (10.56%)	425 (9.71%)	838 (19.15%)
ER: 7,158	118 (1.65%)	61 (0.85%)	291 (4.07%)	210 (2.93%)	528 (7.38%)
CAI: 48,768	1,183 (2.43%)	667 (1.37%)	3,030 (6.21%)	2,674(5.48%)	5,820 (11.93%)
BiolPPI: 125,348	2,019 (1.61%)	1,214 (0.97%)	5,964 (4.76%)	5,339(4.26%)	12,458 (9.94%)

a) The sensitivity values, which equal to the ratios of the element number of (BiolPPI ∩ TechPPI) divided by BiolPPI group, are in brackets. AllPPI group contains the PPIs from five PPI databases. The BiolPPIs are depicted in single model.

high degree nodes (hubs) were cooperatively contributed from multiple sources other than a dominant one. Especially, out-degree of MAPK1 and in-degree of HRAS#KRAS# NRAS in BiolPPI signalling network were contributed no more than 50% from any original pathway database alone (Figure 2). The same situations occurred in in-degrees of CDKN1A, MYC, CDKN2A, BCL2 and IL8 in BiolPPI transcription regulation network (Figure 2; Table S13 in Supporting Information). The result suggested the necessary of dataset integration.

2.5 Comparison with existing databases

We compared our PathPPI categorisation and preserved information with those of Reactome, STRING, CPDB and Pathway Commons (Table 5). Reactome developed very simple rules of transforming the protein interaction from pathway to binary model and defined only four interaction types [18]. STRING and CPDB contained six and three interaction types respectively, but both of them did not provide the rules publicly. Pathway Commons developed the most detailed rules and defined nine interaction types to cover as many relationships as possible, such as the CO_CONTROL interaction to depict the relationship of two entities that control the same reaction and SEQUENTIAL_CATALYSIS for two entities catalyzing two neighbour reactions [11]. PathPPI covered fewer relationship types than Pathway Commons but preserves more information after model transformation, which is very valuable in network analysis. For example, the effect and modification information reserved in PathPPI were missing in all three methods. Another advantage of our PathPPI is that a large amount of traditional PPIs are integrated as TechPPIs.

These five datasets were further divided into subsets based on TechPPI, complex, signaling and transcription interaction (Table 5). PathPPI contained four interaction types. STRING and Pathway common contains three types, while Reactome and CPDB had two and one respectively (CPDB did not provide signalling, transcription and complex interactions for downloading). CPDB had the most TechPPT, up to 468,598 (Figure S1 in Supporting Information). Next were STRING and PathPPI. PathPPI had the most complex and signalling interactions. The interaction O_{AB} values of each two interaction data sets ranged from

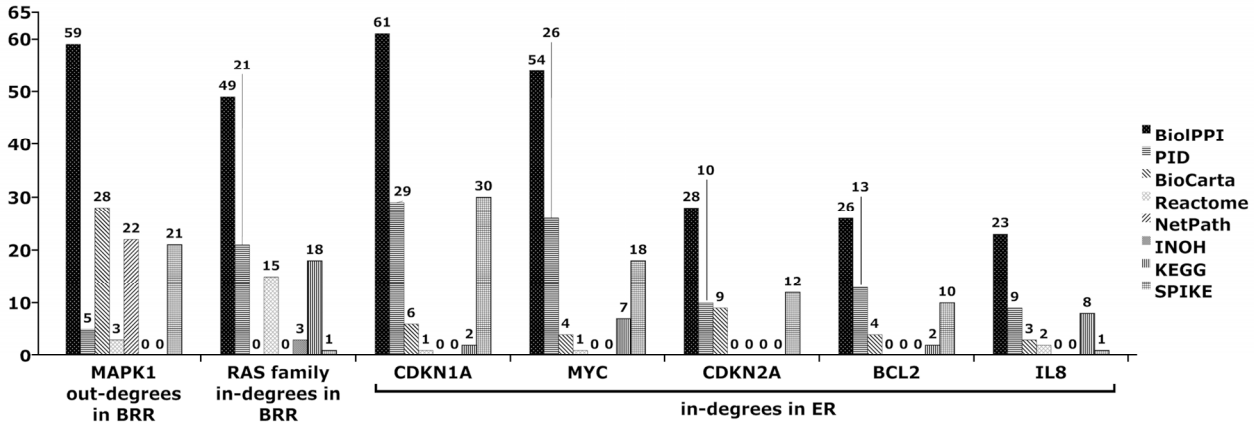

Figure 2 Degrees of the seven nodes that from top 10 signalling or transcription regulation degree list, and are with no more 50% from any original pathway database alone.

Table 5 Comparison of PathPPI with four existing interaction categorisations

Parameters		PathPPI	Reactome	STRING	Pathway commons	CPDB
Preserved information	Directionality	Yes	Yes	Yes	Yes	Yes
	Effect	Activation/ Inhibition/Unspecified	–	–	–	
	Modification	22 pairs of modification	–	–	–	
Interaction categorisation	TechPPI	GI, MI	–	binding	INTERACTS_WITH	Physical interaction
	Signaling	BRR, TBRR, CAR	neighbour-ing_reaction reaction	activation reaction ptmod	STATE_CHANGE CO_CONTROL	Biochemical reaction
	Expression	ER	–	expression	–	Gene regulation
	Complex	CAI	direct_complex indirect_complex	–	COMPONENT_OF IN_SAME_COMPONENT	Biochemical reaction
	Transport	TR, TBRR	–	–	–	
	Metabolic	–	reaction	catalysis	METABOLIC_CATALYSIS SEQUENTIAL_CATALYSIS REACTS_WITH	Biochemical reaction

0.29% to 94.75% among the 13 datasets (Table S14 in Supporting Information). The O_{AB} values between PathPPI_PPI and STRING_PPI, PC_PPI were more than 20.00%, but only 9.52% for CPDB_PPI. PathPPI_Cmp had higher O_{AB} values with PC_Cmp than Reactome_Cmp. The lower O_{AB} values between PathPPI_Sig and PC_Sig resulted mainly from their difference interaction transformation approaches from pathway to pairs model. Totally, the O_{AB} values indicated that there existed a significant difference between PathPPI and other four databases.

2.6 An example in disease genes identification

Screening of genes resulting in specific diseases has long been one of the major tasks in human genetics studies. Recently, numerous methods based on network, that can provide interpretability to a gene or protein, and gene-expression data have been proposed to screen potential disease genes [31,32]. Here, Liver cancer metastasis genes were screened based on PathPPI and two gene expression datasets (Ye's [33] and Zhang's data [34]) by Chuang's method [35]. Ye's data contained 7163 gene expression values for 36 primary metastasis and 26 non-metastasis samples by microarray technology. Zhang's data provided the 7794 protein quantitative values for two non-metastasis and six metastasis cells by mass spectrum technology. Two datasets were normalized by median normalization method. Then, using Chuang's method we screened 5, 1 the co-expression subnetworks ($P<10^{-5}$) from BRR and ER interaction network respectively in Ye's dataset (Table S15 in Supporting Information). and 5, 0 subnetworks ($P<10^{-5}$) in Zhang dataset (Table S16 in Supporting Information). Furthermore, we found that three subnetwork containing at least two known metastasis-association genes (Figure 3). These three subnetworks can be paid more attention in further research.

3 Discussion

Integration of pathway and PPI data can provide more information that could lead to new biological insights. Considering that pathways and PPIs are generally represented by different standards, model transformations are inevitable. One solution to this problem is to transform pathway into binary models. In this study, we developed a series of rules to realise such transformations. Some important information of interactions that is always neglected in earlier studies, such as modification information, is preserved in PathPPI.

A categorisation was established to depict pathway-derived binary interactions and traditional PPIs. Similar to the Reactome, STRING and Pathway Commons categorisations, PathPPI categorisation is also not fully systematic, but it is helpful to deal with current pathway data. A more systematic categorisation based on biological events is desired. In addition, a number of special relationships (e.g., CO_CONTROL in Pathway Commons) in BioPAX that are not extracted in the current PathPPI version will be considered in our future work.

Two fundamental problems exist in the transformation of interactions from the pathway model to the binary model. One is the lack of efficient transformability between the pathway and PPI models. BioPAX and PSI-MI are the mainstream standards of pathways and PPIs, respectively. Although BioPAX is designed to contain all PSI-MI functions, it does not consider transformability with PSI-MI. Multiple data models will inevitably exist for respective advantages and suitable fields. Thus, the efficient transformability between BioPAX and PSI-MI should be improved. Another problem is that some pathway databases do not provide enough information or show divergent use of standards. For example, Reactome and PID do not provide the conversionType items of BiochemicalReaction and Bio-

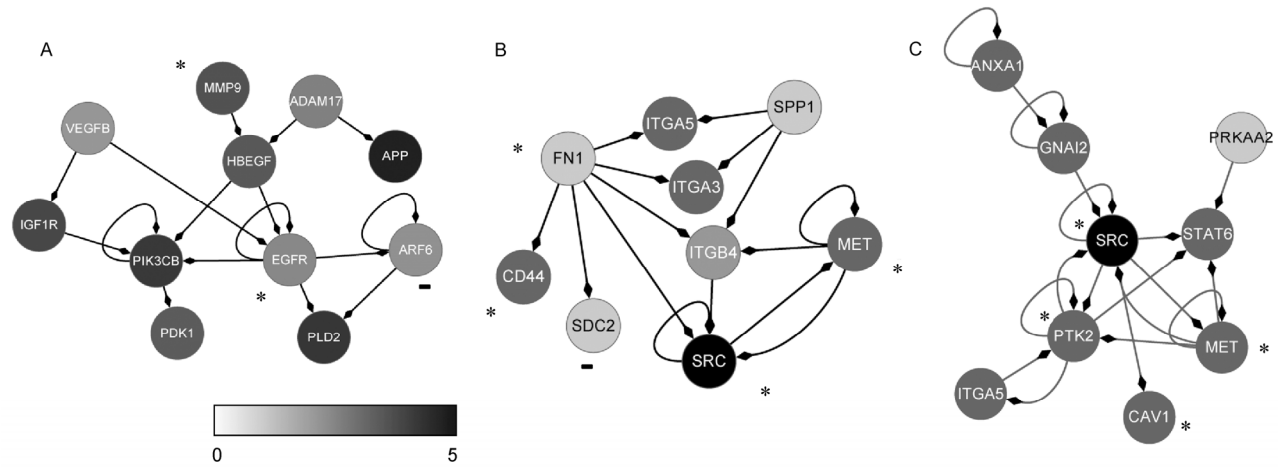

Figure 3 Three co-expression modules with known cancer metastasis genes. Known cancer metastasis and down-regulated genes were marked by "*" and "–" respectively. The color represented the –log of *P* values of each gene (rank-sum test). ADAM17, EGFR, HBEGF, IGF1R, MMP9, PLD2, and VEGFB in the second module were associated with cell migration (Gene Ontology).

Carta and SPIKE do not contain modification information in the BioPAX model. The divergent use of BioPAX may bring about errors or information loss. For example, the interaction, that a set of physical entities aggregated via non-covalent interactions, should be depicted by Complex-Assembly, but it is BiochemicalReaction in Reactome. The latter should be used to depict covalent reactions.

We propose a combination of the complex and single models to represent PathPPIs. The entity in complex models may be very complicated, for example, the entity GGT1*GGT1#GGT2*GGT2#GGT5*GGT5#GGT6*GGT6 #GGT7*GGT7 in Reactome, which represents a protein complex family (the marks * and # are used to separate the members of the complex and family, respectively). The complex model is more accurate in interaction description but inconvenient for analysis. The single model is convenient for analysis but an inaccurate and cumbersome approach for depicting interactions. We thus recommend the complex model equipped with corresponding data extracting methods.

4 Conclusion

This study aimed to provide a method to standardise the integration of pathway and PPI data. We obtained numerous BiolPPIs by transformation of seven pathway databases that are highly complementary with traditional PPIs. These BiolPPIs were integrated with traditional PPIs (TechPPIs) into the PathPPI dataset, which can provide great knowledge with hierarchical information content and confidence. Although PathPPI requires further improvement because of data source constraints, it is a starting point for the standardisation of pathway and PPI data integration. More complete standardisation requires collaboration from communities around the world.

This work was supported by the National High Technology Research and Development Program of China (2012AA020201), National Basic Research Program of China (2013CB910802, 2010CB912700), International Science & Technology Cooperation Program of China (2014DFB30020), National Natural Science Foundation of China (31000379, 31000587, 31000591), and Chinese State Key Project Specialized for Infectious Diseases (2012ZX10002012-006).

1 Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet, 2004, 5: 101–113
2 Han JDJ. Understanding biological functions through molecular networks. Cell Res, 2008, 18: 224–237
3 Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet, 2011, 12: 56–68
4 Tang H, Zhong F, Xie H. A quick guide to biomolecular network studies: construction, analysis, applications, and resources. Biochem Biophys Res Commun, 2012, 424: 7–11
5 Mathivanan S, Periaswamy B, Gandhi TKB, Kandasamy K, Suresh S, Mohmood R, Ramachandra YL, Pandey A. An evaluation of human protein-protein interaction data in the public domain. BMC Bioinformatics, 2006, 7: S19
6 Soh D, Dong D, Guo Y, Wong L. Consistency, comprehensiveness, and compatibility of pathway databases. BMC Bioinformatics, 2010, 11: 449
7 Lu LJ, Sboner A, Huang YJ, Lu HX, Gianoulis TA, Yip KY, Kim PM, Montelione GT, Gerstein MB. Comparing classical pathways and modern networks: towards the development of an edge ontology. Trends Biochem Sci, 2007, 32: 320–331
8 Ahn J, Yoon Y, Park C, Shin E, Park S. Integrative gene network construction for predicting a set of complementary prostate cancer genes. Bioinformatics, 2011, 27: 1846–1853
9 Kirouac DC, Saez-Rodriguez J, Swantek J, Burke JM, Lauffenburger DA, Sorger PK. Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. BMC Syst Biol, 2012, 6: 29
10 Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herwig R. ConsensusPathDB: toward a more complete picture of cell biology. Nucleic Acids Res, 2011, 39: D712–717
11 Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res, 2011, 39: D685–690
12 Chaurasia G, Iqbal Y, Hanig C, Herzel H, Wanker EE, Futschik ME. UniHI: an entry gate to the human protein interactome. Nucleic Acids Res, 2007, 35: D590–594
13 Jayapandian M, Chapman A, Tarcea VG, Yu C, Elkiss A, Ianni A, Liu B, Nandi A, Santos C, Andrews P, Athey B, States D, Jagadish HV. Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together. Nucleic Acids Res, 2007, 35: D566–571
14 Kerrien S, Orchard S, Montecchi-Palazzi L, Aranda B, Quinn AF, Vinod N, Bader GD, Xenarios I, Wojcik J, Sherman D, Tyers M, Salama JJ, Moore S, Ceol A, Chatr-Aryamontri A, Oesterheld M, Stumpflen V, Salwinski L, Nerothin J, Cerami E, Cusick ME, Vidal M, Gilson M, Armstrong J, Woollard P, Hogue C, Eisenberg D, Cesareni G, Apweiler R, Hermjakob H. Broadening the horizon—level 2.5 of the HUPO-PSI format for molecular interactions. BMC Biol, 2007, 5: 44
15 Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D'Eustachio P, Schaefer C, Luciano J, Schacherer F, Martinez-Flores I, Hu Z, Jimenez-Jacinto V, Joshi-Tope G, Kandasamy K, Lopez-Fuentes AC, Mi H, Pichler E, Rodchenkov I, Splendiani A, Tkachev S, Zucker J, Gopinath G, Rajasimha H, Ramakrishnan R, Shah I, Syed M, Anwar N, Babur O, Blinov M, Brauner E, Corwin D, Donaldson S, Gibbons F, Goldberg R, Hornbeck P, Luna A, Murray-Rust P, Neumann E, Reubenacker O, Samwald M, van Iersel M, Wimalaratne S, Allen K, Braun B, Whirl-Carrillo M, Cheung KH, Dahlquist K, Finney A, Gillespie M, Glass E, Gong L, Haw R, Honig M, Hubaut O, Kane D, Krupa S, Kutmon M, Leonard J, Marks D, Merberg D, Petri V, Pico A, Ravenscroft D, Ren L, Shah N, Sunshine M, Tang R, Whaley R, Letovksy S, Buetow KH, Rzhetsky A, Schachter V, Sobral BS, Dogrusoz U, McWeeney S, Aladjem M, Birney E, Collado-Vides J, Goto S, Hucka M, Le Novere N, Maltsev N, Pandey A, Thomas P, Wingender E, Karp PD, Sander C, Bader GD. The BioPAX community standard for pathway data sharing. Nat Biotech, 2010, 28: 935–942
16 Lloyd CM, Halstead MD, Nielsen PF. CellML: its future, present and past. Prog Biophys Mol Biol, 2004, 85: 433–450
17 Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novere N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics, 2003, 19: 524–531
18 Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy

M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H, D'Eustachio P, Stein L. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res, 2011, 39: D691–697
19 Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GS, Venugopal AK, Telikicherla D, Navarro JD, Mathivanan S, Pecquet C, Gollapudi SK, Tattikota SG, Mohan S, Padhukasahasram H, Subbannayya Y, Goel R, Jacob HK, Zhong J, Sekhar R, Nanjappa V, Balakrishnan L, Subbaiah R, Ramachandra YL, Rahiman BA, Prasad TS, Lin JX, Houtman JC, Desiderio S, Renauld JC, Constantinescu SN, Ohara O, Hirano T, Kubo M, Singh S, Khatri P, Draghici S, Bader GD, Sander C, Leonard WJ, Pandey A. NetPath: a public resource of curated signal transduction pathways. Genome Biol, 2010, 11: R3
20 Yamamoto S, Sakai N, Nakamura H, Fukagawa H, Fukuda K, Takagi T. INOH: ontology-based highly structured database of signal transduction pathways. Database, 2011, 2011: bar052
21 Paz A, Brownstein Z, Ber Y, Bialik S, David E, Sagir D, Ulitsky I, Elkon R, Kimchi A, Avraham K, Shiloh Y, Shamir R. SPIKE: a database of highly curated human signaling pathways. Nucleic Acids Res, 2011, 39: D793–799
22 Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res, 2012, 40: D109–114
23 Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. Human Protein Reference Database—2009 update. Nucleic Acids Res, 2009, 37: D767–772
24 Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S, Hermjakob H. The IntAct molecular interaction database in 2012. Nucleic Acids Res, 2012, 40: D841–846
25 Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res, 2006, 34: D535–539
26 Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, Castagnoli L, Cesareni G. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res, 2012, 40: D857–861
27 Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res, 2004, 32: D449–451
28 Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res, 2011, 39: D561–568
29 Montecchi-Palazzi L, Beavis R, Binz PA, Chalkley RJ, Cottrell J, Creasy D, Shofstahl J, Seymour SL, Garavelli JS. The PSI-MOD community standard for representation of protein modification data. Nat Biotechnol, 2008, 26: 864–866
30 Futschik ME, Chaurasia G, Herzel H. Comparison of human protein-protein interaction maps. Bioinformatics, 2007, 23: 605–611
31 Zhang Y, Cheng Y, Jia K, Zhang A. A generative model of identifying informative proteins from dynamic PPI networks. Sci China Life Sci, 2014, 57: 1080–1089
32 Li M, Li Q, Ganegoda GU, Wang J, Wu F, Pan Y. Prioritization of orphan disease-causing genes using topological feature and GO similarity between proteins in interaction networks. Sci China Life Sci, 2014, 57: 1064–1071
33 Ye QH, Qin LX, Forgues M, He P, Kim JW, Peng AC, Simon R, Li Y, Robles AI, Chen Y, Ma ZC, Wu ZQ, Ye SL, Liu YK, Tang ZY, Wang XW. Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. Nat Med, 2003, 9: 416–423
34 Zhang Y, Yan G, Zhai L, Xu S, Shen H, Yao J, Wu F, Xie L, Tang H, Yu H, Liu M, Yang P, Xu P, Zhang C, Li L, Chang C, Li N, Wu S, Zhu Y, Wang Q, Wen B, Lin L, Wang Y, Zheng G, Zhou L, Lu H, Liu S, He F, Zhong F. Proteome atlas of human chromosome 8 and its multiple 8p deficiencies in tumorigenesis of the stomach, colon, and liver. J Proteome Res, 2013, 12: 81–88
35 Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol, 2007, 3: 140

Supporting Information

Table S1 Twenty-two types of modifications in our current categorization of PathPPI

Table S2 Mapping relationships between interactionType values/MOD ID and PathPPI modification categories

Table S3 Mapping relationships between KEGG and PathPPI interaction categories

Table S4 Proportions of different types of entity in the 7 BiolPPI datasets (complex model)

Table S5 Distribution of different types of interaction in the 7 BiolPPI datasets (single model)

Table S6 Distribution of effects and modifications in the 7 BiolPPI datasets.

Table S7 Overlaps of interactions among the 7 BiolPPI datasets (single model)

Table S8 Top 10 proteins with the highest BiolPPI out- or in-degree in signalling and transcription network

Table S9 Top 10 proteins with highest degree in different BiolPPI networks (TransportRegulation)

Table S10 Top 10 proteins with highest degree in different BiolPPI networks (TransportWithBiochemicalReactionRegulation)

Table S11 Top 10 proteins with highest degree in different BiolPPI networks (ComplexAssemblyRegulation)

Table S12 Top 10 proteins with highest degree in different BiolPPI networks (ComplexAssemblyInteraction)

Table S13 Comparison of BiolPPI and the seven original databases for the top degree nodes in signalling and transcription regulation networks

Table S14 Interaction overlaps of 13 subsets from PathPPI, Pathway common, Reactome, STRING and CPDB

Table S15 Modules identified from the BRR and ER network (Ye's data)

Table S16 Modules identified from the BRR network (Zhang's data)

Figure S1 Comparison of the scale of 13 interaction data subsets.

The supporting information is available online at life.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.