Chapter 5

Manual Annotation of Protein Interactions

Svetlana Bureeva, Svetlana Zvereva, Valentin Romanov, and Tatiana Serebryiskaya

Abstract

Protein interactions are the basic building blocks for assembly of pathways and networks. Almost any biologically meaningful functionality (for instance, linear signaling pathways, chains of metabolic reactions, transcription factor dimmers, protein complexes of transcriptosome, gene–disease associations) can be represented as a combination of binary relationships between "network objects" (genes, proteins, RNA species, bioactive compounds). Naturally, the assembled pathways and networks are only as good as their "weakest" link (i.e., a wrongly assigned interaction), and the errors multiply in multi-step pathways. Therefore, the utility of "systems biology" is fundamentally dependent on quality and relevance of protein interactions. The second important parameter is the sheer number of interactions assembled in the database. One needs a "critical mass" of species-specific interactions in order to build cohesive networks for a gene list, not a constellation of non-connected proteins and protein pairs. The third issue is semantic consistency between interactions of different types. Transient physical signal transduction interactions, reactions of endogenous metabolism, transcription factor–promoter binding, and kinetic drug–target interactions are all very different in nature. Yet, they have to fit well into one database format and be consistent in order to be useful in reconstruction of cellular processes.

High-quality protein interactions are available in peer-reviewed "small experiment" literature and, to a much smaller extent, patents. However, it is very challenging to find the interactions, annotate with searchable (and computable) parameters, catalogue in the database format in computer readable form, and assemble into a database. There are hundreds of thousands of mammalian interactions scattered in tens of thousands of papers in a few thousands of scientific journals. There are no widely used standards for reporting the interactions in scientific texts and, therefore, text-mining tools have only limited applicability. In order to generate a meaningful database of protein interactions, one needs a well-developed technology of manual curation, equipped with computational solutions, managerial procedures, quality control, and users' feedback. Here we describe our ever-evolving annotation approach, the important annotation issues and our solutions, and the mammalian protein interactions database MetaBase^{TM} which we have been working on for over 8 years.

Key words: Protein interactions, manual annotation, interaction database, literature curation, biological pathways, networks.

Yuri Nikolsky, Julie Bryant (eds.), *Protein Networks and Pathway Analysis, vol. 563*

DOI 10.1007/978-1-60761-175-2_5 Springerprotocols.com

1. Introduction

The goal of systems biology is to understand how genetic information is realized at the level of macromolecular (protein) networks and how the networks respond to stimuli in a cell type-specific, organ-specific manner (1). The first step in this process is to build an inventory of "building blocks" that the networks consist of – the nodes and links (edges) between them. The *nodes* are represented by molecular entities – proteins, genes, RNA species, endogenous metabolites, and xenobiotics. By now, the majority of "molecular entities" for human and model organisms are fairly well studied, standardized, and catalogued in multiple databases and linked to genome browsers in a generally non-contradictory way. The situation with *edges* is different. Proteins, genes, RNA species, and compounds are connected by interactions and associations of different types, and there is much less consistency between the sources of interactions, the ways they are generated and assembled into these databases.

There are several types of interactions traditionally studied by different fields, each with a different purpose. Thus, metabolic "interactions", i.e., reactions between endogenous substrates and products mediated by enzymes, were mostly collected decades ago by traditional biochemical in vitro experiments and assembled into reaction and metabolic pathways databases such as BRENDA (2), EMP/MPW (3), and, later, KEGG (4). Metabolic databases were assembled mainly as repositories of biochemical knowledge with potential applications such as stoichiometric modeling and strain development. Signaling interactions were mostly studied by geneticists and "systems biology" researchers over the last decade in yeast, fly, and worm using yeast two-hybrid assays, co-immunoprecipitation, or manual literature curation. These protein–protein interactions, or PPIs, became the bulk of the multi-species "PPI" databases such as BIND (5), MIPS (6), DIP (7), STRING (8), and IntAct (9). Later, a database of human specific interactions HPRD was launched by Johns Hopkins University and the Indian Bioinformatics Institute (10). PPI databases were mostly used for network analysis and simulation in model organisms. The third type of interaction is between bioactive xenobiotic "drug-like" compounds and protein targets, mostly known from screening assays carried out by medicinal chemists, mainly at pharmaceutical companies. The bulk of this knowledge has never been published, but probably several million of ligand–target data points are available in medicinal chemistry publications and patents. Over the years, these data were assembled in well-curated databases such as Prous' Integrity (now part of Thomson Scientific), MDL's Discovery Gate and, later, GVK (*see* chapter 3 in this

book). The ligand–target interaction data are supported by extensive chemistry and pharmacology annotations such as synthesis schemas, indication, and side effects. These databases are mostly used as reference by medicinal chemists at different walks in preclinical drug discovery.

Since three domains of interactions relevant for biological networks were assembled largely independently from each other for different purposes, there is little semantic consistency between them. For instance, metabolic pathways and reactions connect metabolites with EC numbers; functions assigned to enzymatic activity originally by Enzymatic Commission (EC). However, EC numbers are redundant, i.e., there are several individual proteins encoded by different genes which belong to the same EC function. However, signaling PPIs are gene specific (for instance, transcription factors bind to promoters of individual genes), and therefore, it is not possible to link PPIs with "traditional" enzymatic reactions. There are some conventions such as SBML (systems biology markup language (11)) and Bio-Pax for standardization and better consistency between interactions domains and databases, but most interactions annotation data are "lost in translation" due to poor connectivity.

There are many other critical issues regarding annotation of interactions that are not yet resolved. For instance,

- There is an ongoing discussion of the *quality* of interactions assembled by "automated" methods such as natural language processing (NLP) vs. that of manual curation of full text experimental articles.
- Underlying *experimental evidence* behind interactions. "High throughput" experiments such as co-immunoprecipitation or yeast two-hybrid allow for identification of a large number of interactions, which are usually inconsistent with "small experiments" interactions.
- *"Critical mass" of interactions.* It is still an open question of how many experimentally confirmed interactions are published (as scientific papers are not standardized it is not trivial to calculate it), as well as how many interactions are actually needed for generation of meaningful organism-specific networks.
- *Species specificity.* It is not clear what interactions are species specific and whether inter-species interactions help or hurt in "high resolution" network analysis of complex conditions such as human diseases.
- The *level of abstraction* in annotation of interactions and their attributes.
- *Protein families and complexes.* The composition of protein groups and interactions division between groups and within groups.

– *Software solutions* for manual annotation.
– *Ontologies and controlled dictionaries* used for hierarchical classification of interactions and "nodes".

In this chapter, we summarize our experience and philosophy, cultivated over 7 years, during which we have developed a methodology of manual annotation of interactions of different domains, then assembled them into a relational knowledgebase known as MetaBase™ (available from GeneGo, Inc.). We explain the details of the annotation process, comment on the issues raised above, and briefly describe the software suite called Pathway Editor, which we use for the annotation process and populating the database.

2. Molecular Entities and Their Database Attributes

2.1. Gene

Gene in MetaBase™ corresponds to that in the Entrez Gene database. It can be defined as a gene or a sequence from the RefSeq database. Every gene in the database has the following attributes:

1. Name and synonyms
2. Identifiers (EntrezGene, Affy, OMIM, Unigene, HGMD, GeneBank, RefSeq, SNP, etc.)
3. Location on chromosome
4. Orthologs (according to HomoloGene)
5. Tissue expression

2.2. Protein

Protein is a product of a gene translation. Two types of proteins exist in MetaBase™: (1) real proteins described in Swiss-Prot database (12) and (2) "fake" proteins that are products of translation of genes listed in Entrez Gene, but not having a Swiss-Prot ID.

1. Name and synonyms
2. Identifiers (Swiss-Prot ids, PIR, EMBL, KEGG, OMIM, etc.)
3. Tissue distribution
4. Cellular localization
5. Organism

Most proteins in higher eukaryotes are classified in groups (protein families of hierarchical structure) and work in multi-protein aggregations (protein complexes). The relations between proteins within groups have to be reconstructed manually.

2.2.1. Protein Groups

There are many ways to define protein groups (families) based on sequence homology, evolutionary conservation, function, etc. We define groups based on their networking properties. In

MetaBase™, if all proteins united in a group can participate in the interaction with the same protein, the interaction is assigned to the group, and the group can be "collapsed" into one object on the network. Protein groups are used when there is no direct data on which isoform in the group participates in the "upstream" and "downstream" signaling interaction, which creates an ambiguity on the network, as the pathway is allowed to proceed through the group.

2.2.2. Protein Complexes

In MetaBase™, protein complexes are considered as multiple proteins which are physically connected to each other and function as a whole. These are stable species-specific interactions which keep complex subunits together and co-operate for executing the common function of a complex. For instance, many transcriptional factors function as homo or heterodimers. Importantly, different dimmers are formed in different cell types, and also depend on conditions. In MetaBase™, two types of complexes are defined: (1) complexes composed of equally participating subunits and (2) heterocomplexes composed of regulatory and catalytic subunits.

2.3. RNA

RNA is an object, which is not connected with any external database. It represents the end product of gene expression, for example *microRNA*, *tRNA*, *rRNA*.

2.4. Compounds

In MetaBase™ the following molecular entities are ascribed to the category Compound:

1. Xenobiotic compounds (including drugs)
2. Endogenous compounds (including metabolites of xenobiotics) and nutrients
3. Endogenous (including DNA and RNA), natural, and artificial polymers
4. Modified and recombinant proteins (including monoclonal antibodies)

Every compound in the database has the following attributes:

1. Name and synonyms
2. Chemical structure. In MetaBase™, most of the organic compounds are characterized by its neutral structures. Charged structures exist for endogenous and essential nutrimental inorganic ions and for representation of drugs salt form, provided that neutral structure for the drug is present. All three types of isomers (structural, geometrical, and enantiomers) are depicted isomers specifically except the cases when there are no experiments for separate isomers, and isomer mixtures and racemates have to be created for their activity annotation.

3. Identifiers of other databases (CAS-numbers, KEGG, PubChem SID and CID, MeSH, DrugBank, ChEBI, etc.)
4. Compound category (xenobiotics, metabolite of xenobiotic, endogenous compound, nutrient, environmental compound, modified and recombinant proteins)

2.4.1. Compound Groups

In accordance with proteins, if some compounds can participate in an interaction with the same protein, all the compounds are united into a group and the group is "collapsed" into one object on the network. The interaction with protein is assigned to the group. Isomer mixtures represent another kind of compound group and are created for activity annotation when there are no experimental data for separate isomers.

3. Data Connection Types

3.1. Ontologies

Use of ontologies, or structurally controlled vocabularies, is a basis for effective annotation process and compatibility of the resulting database with other databases. In addition to creating our own ontologies (*see* below), we adapt public domain ontologies developed by The Gene Ontology Project: Molecular Function Ontology and Biological Process Ontology (13).

Some uniquely controlled vocabularies are used for annotation of other entities, attributes, and experimental details. Being based on some generally accepted ontologies or terms collections or being developed from scratch, these ontologies are actively evolved further by GeneGo annotators to allow controlled and consistent annotation of uninvolved terms or entities. Examples are controlled dictionaries for

1. diseases, originally based on Medical Subject Headings (MeSH) classification (14), which have been considerably expanded by chemical toxicity terms (histopathology, clinical pathology and other adverse reactions).
2. tissues
3. cellular localizations
4. experimental methods
5. units

3.2. Classifications

The placement of a compound in the general classification of compounds is one of its important characteristics reflecting its structural features and functional role. It facilitates compound annotation and rational grouping into network objects. There are four compound trees in MetaBase™:

1. Classification of endogenous compounds.
2. Classification of nutrients.
3. Classification of drugs by pharmacologic action according to the Medical Subject Headings (MeSH) classification.
4. Classification of drugs according to the Anatomical Therapeutic Chemical (ATC) classification system.

3.3. Associations

An example of associations is gene/protein–disease associations. These are "causative" genetic links describing the connection between a certain form of gene/protein (SNP, mutation, isoform, RNA overexpression, etc.) and a certain onset of the disease. In most cases, causative associations are established statistically and the direct mechanism of gene/protein involvement in the disease is not known.

Compounds are also connected with diseases. Compound can treat disease, being a drug, or can cause disease (or any toxic endpoint), being a toxicant, and it can be a biomarker of a disease state. All these cases are captured by annotations accompanied by necessary attributes.

3.4. Interactions

In general, "interaction" represents an influence an object has on another. There are a few main types of interactions between molecular entities (*see* more below):

1. Protein–protein interactions. This type of interactions constitute the majority of signaling networks. Most protein interaction databases focus on signaling protein–protein interactions, including Swiss-Prot (12), MINT (15), DIP (7), BIND (5), and HPRD (10).
2. RNA–protein interactions. A variety of interactions mostly describing protein translation and RNA degradation.
3. RNA–RNA interactions. Interactions between different types of RNA (for instance, in the ribosome). Also includes micro RNA–target mRNA interactions. Some of the latter interactions are collected at miRBase (16).
4. Protein–DNA interactions. Mostly, these interactions represent interactions between transcription factors and target gene promoters. One of the best known databases on protein–DNA interactions is DBTSS (17).
5. Compound–protein interactions. Interactions of low-molecular weight endogenous and xenobiotic ligands with proteins leading to modulation of proteins' activity.
6. Compound–DNA, RNA interactions. Mostly unspecific interactions of planar or highly reactive compounds leading to interruptions in genes expression.

7. Compound–compound interactions. These are indirect interactions between bioactive compounds: drug–drug interactions or endogenous ligand–synthetic ligand interactions.

3.5. Reactions

Reactions represent a distinct type of relationships between the database objects, showing either transformation of an object to another one or changing of the object localization. Metabolic reactions (which further assemble into metabolic pathways) represent one-step biochemical transformations of a compound occurring either spontaneously or under the action of an enzyme. Steps of protein proteolysis as part of protein maturation process are also represented as reactions. Another type of reactions, transport reactions, depicts transportation of a compound or an ion by an organic transporter or an ion channel, for example, through a cellular membrane.

4. Interactions in the Network

4.1. Attributes for Interactions Stored in the Database

In MetaBase™, all interactions are attributed with a (1) direction, indicating signal transduction; (2) effect, depicting character of influence; (3) mechanism, showing how the effect has been reached; (4) experimental details from literature source, confirming the interaction; and (5) trust, giving by an expert and indicating the probability of the interaction existence. Based on the mechanism, interactions are divided as direct interactions meaning that physical contact between interacting objects occurs and indirect interactions, when observing effect between the objects is mediated by omitted interactions or a whole pathway.

4.1.1. Direction

In MetaBase™, the vast majority of interactions are directional, i.e., depict "from–to" relations. This is characteristic for individual interactions such as microRNA–target inhibition, as well as interactions linked into multi-step linear signaling or metabolic pathways. A classical schema of a signaling pathway is shown in **Fig. 5.1A**. It initiates with a ligand–receptor interaction on cellular membrane and is transmitted via several signal transduction interactions to the transcription factor, followed by its binding to the promoter of a target "effector" gene such as endogenous metabolic enzymes. Information on the direction of interaction is not always available from experimental literature. For instance, yeast two-hybrid or co-immunoprecipitation assays can only establish the fact of binding but not the direction of interaction. In such cases, the direction can be established by its relative position in the signaling pathway. When no additional data on the pathways is

available (typical case for recently annotated proteins), the direction is not marked until more data establishes it.

4.1.2. Effect

The "effect" depicts the result of the interaction on activity or/and abundance of one of the objects. In MetaBase™, interactions have three effects: activation or increase in quantity inhibition or decrease in quantity, and uncertain or not specified in the article.

As in the following example, *We propose that the transcription factor NF-kappaB acts as a repressor in neurons but as an activator of BACE1 transcription in activated astrocytes present in the CNS under chronic stress, a feature present in the AD brain.* (18), an object can influence another object causing dual or multiple effects. Usually, the difference is attributed to different cell types and conditions (for instance, hypoxia in solid tumors). In these cases (about 2% object pairs), both interaction effects are annotated and can be visualized on the network (**Fig. 5.1**).

4.1.3. Mechanism

Every interaction in MetaBase™ is attributed with a "mechanism" which is deduced from the experimental assays used for the interaction's evidence. We distinguish direct (physical) and indirect (functional) interactions. Main mechanisms used in MetaBase™ are listed in **Table 5.1** and illustrated in **Fig. 5.1A**.

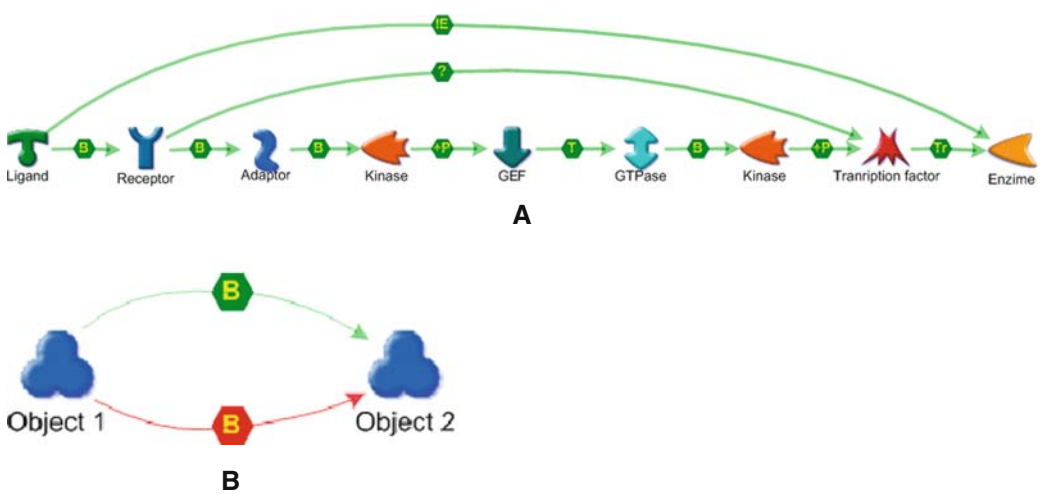

Fig. 5.1. Protein–protein interactions in network. **(A)** General schema of signaling pathway consisting of direct interactions. Possible indirect interactions are also shown. **(B)** Interaction with different effects between interacting objects.

4.1.4. Experimental Details from Literature Sources

Every interaction is accompanied by the literature source(s) where the information about the interaction or reaction was found. Additional information containing experimental details is taken from the literature source and inserted by annotators into about 20 fields, including organism, tissue, cell line, method, text description, concluding the main statement from the article confirming the interaction, numerical data, etc. (**Fig. 5.2**).

5. Reactions in the Network

Traditional differentiation of all cellular processes into metabolic and signaling components is largely rooted in the history of modern biology and the specific training and background of the scientists that study such processes. In reality, in a living cell both signaling and metabolic cascades work together and are very closely depend on each other, as many endogenous compounds act as ligands for signaling cascades and almost all transcription factors in turn regulate metabolic enzymes. Although one can manually trace this chain of signaling and metabolic transformations on static pathway maps, the ability to combine protein–protein interactions and metabolic reactions into one self-assembling network is invaluable

- for the discovery of missing information;
- for the formulation of *in silico* models; and
- as a tool for integrated analysis of genomic, proteomic and metabolic data sets.

The integration of proteins interactions and metabolic reactions into one dynamic interactome through the merging of signaling, metabolic, and transport networks has two main challenges.

The *first* one is that signaling can be effectively described as a network of binary physical proteins and compounds interactions, with billions of possible multi-step combinations. Representation of metabolic transformation through binary interactions (**Fig. 5.3A**) has serious limitations. The following example represents this: phosphatidylinositol phosphatase PTEN inhibits AKT by reducing concentration of its substrate, phosphatidylinositol 3,4,5-triphosphate, a small molecule ligand that activates AKT. PTEN dephosphorylates phosphatidylinositol 3,4,5-triphosphate converting the former to phosphatidylinositol 4,5-bisphosphate.

Table 5.1
Main interaction mechanisms in MetaBase™

Direct interactions	
Binding	Two proteins or a protein and compound physically bind through non-covalent binds. Most direct signaling interactions are assigned as "binding"
Cleavage	Most often a proteolytic cleavage of a protein at a specific site yielding distinct peptides fragments. Interactions with the mechanism are now being transformed from binary interactions to reactions
Covalent modification	Covalent binding of a reactive chemical group of a compound to protein or nucleic acid
Phosphorylation	Protein activity is altered by an addition of a phosphate group
Dephosphorylation	Protein activity is altered by a removal of a phosphate group
Transcription regulation	Direct binding of a transcription factor to the promoter of its target gene
Competition	The mechanism is given to interactions between endogenous and synthetic ligands binding to a receptor and competing for the same binding site
Transport	Small molecule or protein is transported by another protein, transporter, or a channel. Interactions with the mechanism are now being transformed from binary interactions to transport reactions (*see* below)
Transport catalysis	The mechanism is given to interactions between a transporter or a channel and transport reaction
Catalysis	The mechanism is given to interactions between an enzyme and reaction catalyzed by it (*see* below)
Indirect interactions	
Influence on expression	A protein/peptide ligand or a small molecule changes the expression level of the target gene indirectly; for instance via binding to upstream receptors
Drug–drug interactions: pharmacological effect	A small compound changes pharmacological effects of other drugs indirectly, for instance, influencing metabolic enzymes or organic transporters
Drug–drug interactions: toxic effect	Drugs change toxic effects of other drugs, for instance, influencing metabolic enzymes or organic transporters
Unspecified interactions	Mechanism is unknown
Groups and complex connections	
Class relations	Relationships within protein or compound groups are considered as an independent mechanism
Complex subunit	Relationships within protein complexes

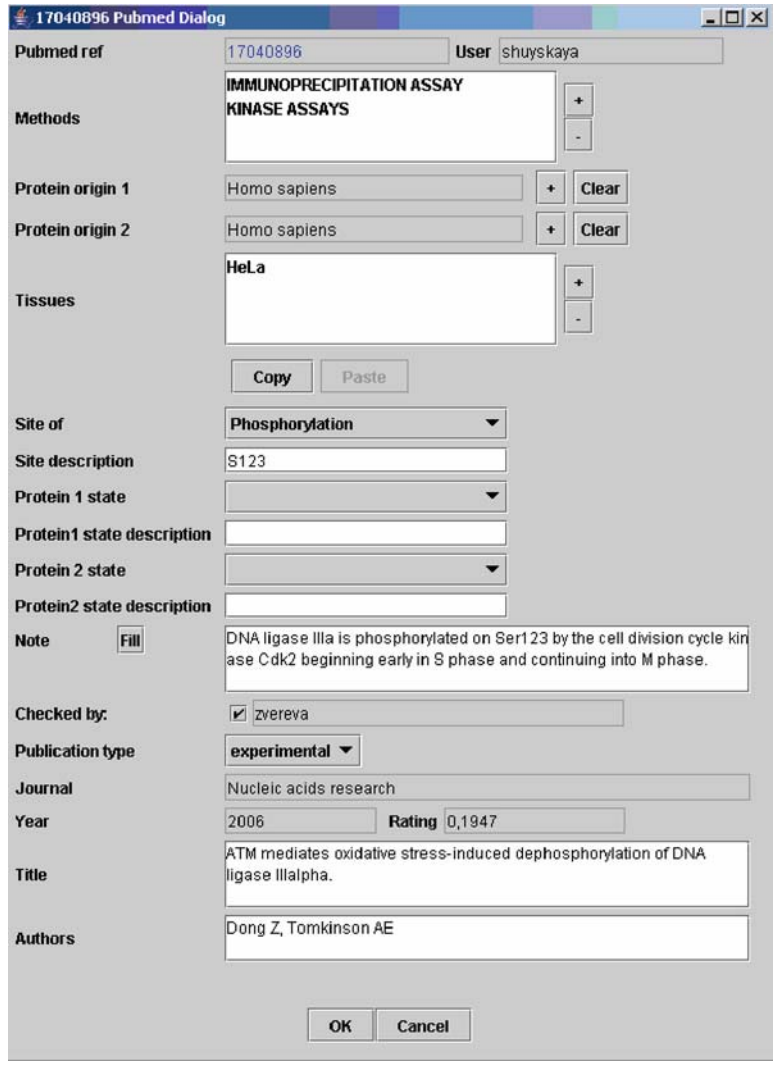

Fig. 5.2. Annotation form for experimental details of protein–protein interaction.

- This representation (**Fig. 5.3A**) contradicts the classic representation for the metabolic reaction, where the direction of the arrow should be reversed from the substrate, phosphatidylinositol 3,4,5-triphosphate to its enzyme, PTEN. If we had reversed the arrow, the important reaction of AKT inhibition would be discarded from the network, and network-building algorithms will have interrupted the signaling cascade from PTEN.
- Mechanism of the reaction (dephosphorylation in this case) cannot be shown.
- If a protein catalyses several reactions, it is not possible to associate substrates and products in pairs.

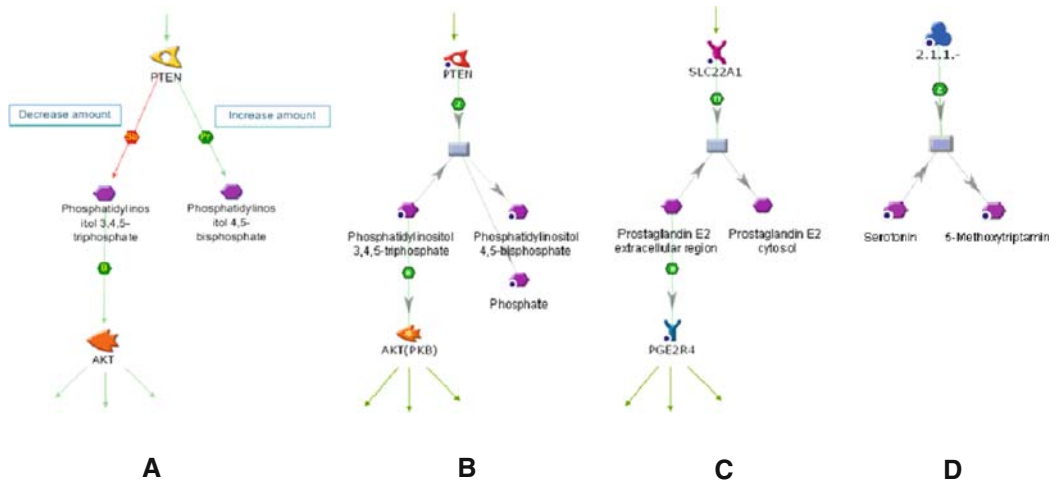

Fig. 5.3. Graphic representation of metabolic and transport reactions. **(A)** Representation through binary interactions. The reaction is indicated by two binary interactions characterized by effect, mechanism, and direction: in the far upper left – inhibition and the mechanism "Decrease amount of substrate" and/or in the other cases – activation and the mechanism "Increase amount of product". **(B)** Example of metabolic reaction representation in MetaCore/MetaDrug network. **(C)** Example of transport reaction representation in MetaCore/MetaDrug network. **(D)** Graphic representation of metabolic reaction with EC number.

- This representation does not allow for visualization of reversible reactions catalyzed by the same protein.

The limitations prompted us to create a new approach for metabolic reaction representation, which was developed to overcome the above described limitations, and was necessary for graphically aligned protein–protein interactions in the signaling network. We offer a novel way of metabolic reaction representation (**Fig. 5.3B**), where essential parts of metabolic reactions are introduced into signaling networks not as the typical node–node connections, but as nodes of a different shape (gray boxes) hyperlinked to the particular transformations in the underlying metabolic database. Such representation allowed us to preserve the integrity of signaling networks by avoiding the building of nonsensical cascades, and yet is simple and intuitive as the arrows point from the substrate to reaction and from reaction to the product (compare with **Fig. 5.3A**). A user is able to retrieve the complete information about a particular metabolic reaction when needed in a classic representation A + B = C + D by mouse-over or clicking the gray box representing a reaction on network. Such a system enables us to unambiguously ascribe substrates and products to a definite reaction and represent reversible reactions catalyzed by the same protein.

Another advantage of this representation is the possibility of unification, namely application of the same approach for representation of transport reactions (**Fig. 5.3C**). In the case of network

object for a transport reaction (represented by a gray box), this is connected by two direction-specified edges with nodes representing transported compound in different localizations and a protein (transporter or a channel) linked to the transport reaction (**Table 5.1**).

The *second* issue that comes up when representing metabolic reactions is the lack of fine mapping, i.e., direct connectivity between individual metabolic transformations, specific enzymes, and the encoding genes. Metabolic reactions in most databases (KEGG (4), MetaCyc (19), The Structure–Function Linkage Database (SFLD) (20)) traditionally include all EC numbers (generic enzymatic functions) capable of performing a particular reaction.

The reactions are being re-annotated at GeneGo to assign a particular protein to each metabolic reaction. For the reactions, for which it is impossible to identify the concrete enzymes (isoforms) responsible for unique substrate–product transformation in human tissues, we incorporated a new network object: EC number (**Fig. 5.3D**), which is connected with a metabolic reaction in the same way as concrete enzymes (**Fig. 5.3B**). Interactions between enzyme and reaction or EC number and reaction have the same attributes as a binary interaction between any other objects in database (**Table 5.1**).

6. Level of Abstraction

Proper formalization of interactions is critical for creation of a semantically consistent interactions database. With billions of interactions between millions of active proteins in the cell, a rule-based simplification is necessary for generation of meaningful yet comprehensible networks. For instance, the whole protein biosynthesis machinery should be collapsed to a single interaction TR → T (transcription factor binding to the promoter region of the target protein) for the continuity of the network downstream of T. If the whole protein biosynthesis machinery was reconstructed de novo, it would add over 100 proteins and RNAs to every network, and calculations and network visualization would become impossible.

Also, narrowing down the set of interactions to "protein–protein" type alone would make the network interrupted. For instance, the key interaction between RelA and MMP9 would be omitted because RelA as a transcriptional factor binds to the promoter region of MMP9 gene, not the MMP9 protein.

Optimizing the abstraction level is necessary for depicting a biological role of small compounds representing the most populous class of molecular entities in MetaBase™. Compounds are grouped based on structural and functional principles. For example, the group "Amino acids" on the "Glutathione metabolism" pathway map shows the role for every amino acid, without loss of sense. Presence of two tens of network objects for every amino acid would simply overload the map.

The speed and performance of network-building algorithms are directly dependent on the number of network objects, and, therefore, the issue of objects grouping is of key importance. The number of synthetic compounds that have been published or patented is, probably, greater than 10-fold higher than the number of published protein–protein interactions. Some public (PubChem Compounds, PubChem Substances) and commercial (GVK BIO) databases contain millions of compounds and this number grows by over a million per year. No network algorithm is capable of processing such amounts "on the fly" which is required by the database users. In MetaBase™, compounds described in one article or patent, with activity associated with the same target protein, and with the same attributes for compound–protein interaction are grouped into one complex network object and interaction with the protein is shown by one node. After incorporation of MedChem database (GVK Biosciences) to MetaBase™, this approach allowed us to reduce the number of compound network objects up to 12-fold and the number of nodes up to 5-fold.

7. Species Specificity

Species specificity of protein interactions is an important and controversial topic in systems biology (**Table 5.2**). There is a large literature on using interactions in one species (usually, lower eukaryotes model organisms such as yeast, fly, and worm) for extrapolation of interactions in other organisms based on matching orthologs for node proteins and assigning higher confidence for inter-species interactions (21). Using such deduced interactions for analysis of human disease data is, arguably, a stretch. On the other hand, some interactions are highly species specific and assigned only to human, mouse, or rat. An example of such a species-specific database is HPRD developed by Johns Hopkins University and the Indian Institute of Bioinformatics (10). The middle ground is covered in taxon-specific interaction databases, such as MetaBase™ which is focused on human-, mouse-, and rat-specific interactions and matching mammalian orthologs.

Table 5.2
Comparison of strict and multi-organism annotations of interactions

	Strict organism-specific interactions annotation	Multi-organism interactions annotation
Pros	• Accuracy, the highest fidelity approach	• Information is not lost. Critical mass of interactions is assembled faster, at lower cost
		• Multi-organism interactions help to reconstruct biological networks when there are gaps in knowledge for individual species
Cons	• The body of experimental data is insufficient for collecting a "critical mass" of organism-specific interactions, at least for higher eukaryotes (**Fig. 5.4A**) • Interactions between protein orthologs may be useful for understanding the core mechanisms of biological processes and evolutionary aspects	• Loss of accuracy. It is very difficult and time consuming to confirm protein interactions deduced by orthologs comparison. Biological networks are very sensitive to inaccurate links information, as every false-positive and false-negative link is multiplied and magnified by network algorithms

8. Manual vs. Automatic Annotations

MetaBase™ is based on a comprehensive collection of associations, interactions, and reactions that have been manually curated from "small experiments" literature and patents. There are several approaches to extraction of object connections from the literature, which in general can be divided into automated, manual, or a combination of the two. Manual annotation means that the initial processing of information is carried out by trained specialists – biologists and chemists, with the complete analysis of experimental procedures and data. Manual annotation is fundamentally based on establishing a set of rules which are executed in annotation forms with parsable fields and tables. Establishing rules is a necessity, because annotation is a team process, and the collected information has to be consistent with minimal possible subjective bias. The main advantages of manual annotation are accuracy, completeness, and consistency within the annotation database. The drawbacks are high cost and a necessity of a well-trained annotation team.

Natural language processing (NLP) technologies represent automated approaches for text analysis and harvesting object connections from scientific literature. NLP software solutions, such as

I2E (Linguamatics, *see* chapter 1 in this book) and MedScan (Ariadne), perform keyword searches in PubMed lists in order to collect all the articles of interest, then scan the text in the abstracts searching for "connections" between keywords, highlight these connections in the articles, and then list these "findings" connections in separate tables. NLP associations are invaluable in many applications, but high false-positive and false-negative errors limit "automatic" interactions extraction in a curated database. For instance, in one of our studies, an NLP algorithm retrieved an interaction between CXCL9 (chemokine (C-X-C motif) ligand 9) and CXCR4 (chemokine (C-X-C motif) receptor 4), although the proteins were not studied in the article. Even few of such "false-positive" assignments may substantially skew the network topology (**Fig. 5.4B**).

However, despite the fact that NLP technologies require manual curation of the resulted interactions, they can be very helpful for initial collection of information.

9. Execution in Software. Annotation Process: Forms and Tables

MetaBase™ is an Oracle database that stores over a million objects of different types and over 300,000 of their interactions and associations (**Fig. 5.5**). The Accord Chemistry Cartridge (Accelrys, USA) enables Oracle database to store, index, and search chemical structures.

Annotation is performed using Pathway Editor, a commercially available software module (GeneGo, Inc.). Pathway Editor is a cross-platform client/server application written in Java and implemented using JDK 1.4. Java Web Start technology enables Pathway Editor to be deployed with a single click at any computer with an Internet connection.

Adding new interactions in Pathway Editor includes four main steps: (1) search of the entities in the database and creation of new ones in case of absence, (2) selection or creation of network objects for the entities, (3) creation of an interaction or association (**Fig. 5.6**), and (4) ascribing necessary attributes for the object connection. The objects search can be done using keywords, names, abbreviations, and identifiers of external databases as queries. For chemical structures, the exact structure search, salt search, isomers search, substructure, and similarity searches are additionally available. The structures can be created using drawing tool or inserted as MOL file or SMILES. At the instance of creation of new entities or connections, checking for duplicates is performed using the embedded synonyms dictionary. Typeable fields are available

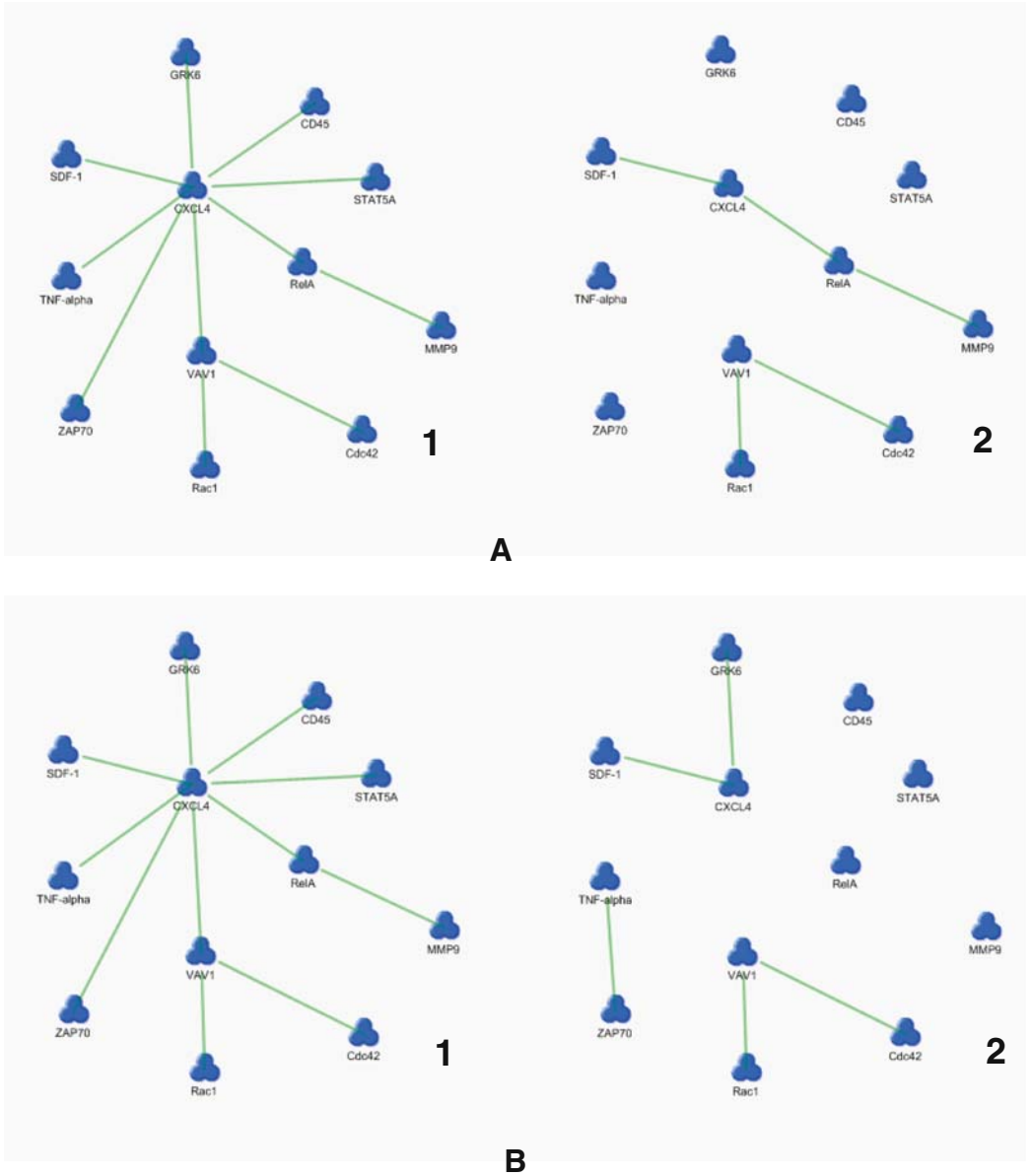

Fig. 5.4. Comparison of protein–protein interaction networks. **(A)** Comparison of strict and multi-organism annotations of interactions. **1:** Network of interactions based on interactions shown for human and murine protein; **2:** network of interactions shown only for humans. **(B)** Comparison of interactions annotated **1:** manually; **2:** using NLP interactions.

only for text notes and descriptive fields. All other fields are provided with vocabularies for terms selection (*see* above), which excludes possibility of typos while filling in the forms.

Pathway Editor is not only an annotation tool but also a powerful enterprise solution for managing a large team of annotators. The access is controlled by hierarchical licensing, which

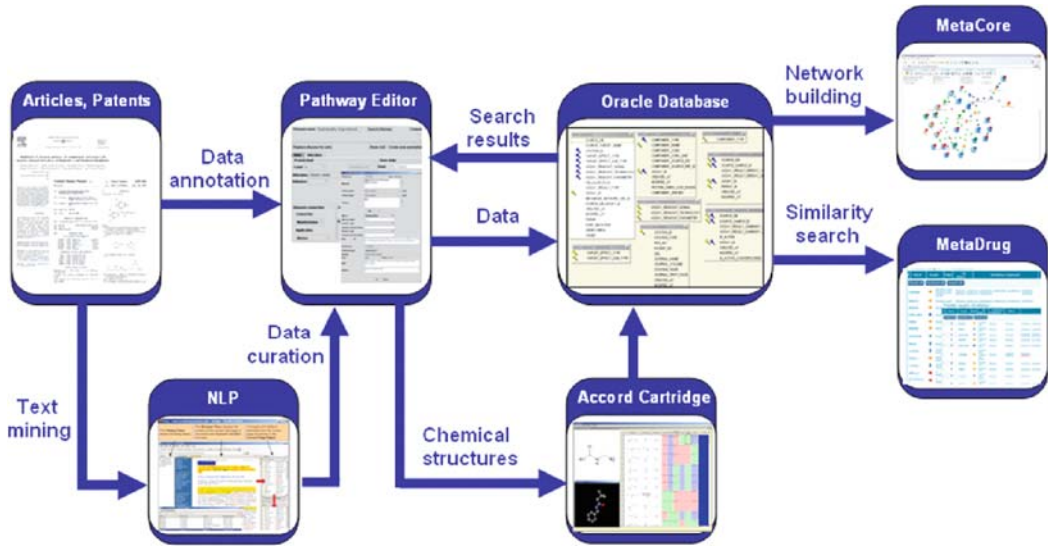

Fig. 5.5. General scheme of data flow from the annotation server to MetaBase™.

limits the annotators' access according to their expertise level and role in the project. The field-centered licensing distinguishes "chemistry", "biology", "ontology", and some other access

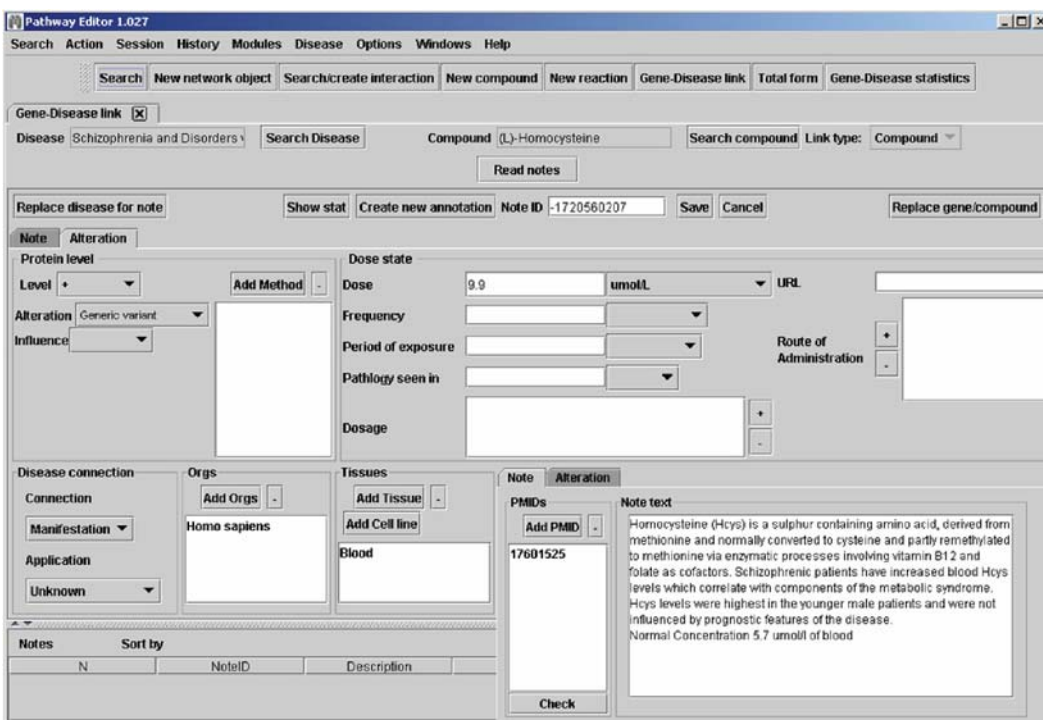

Fig. 5.6. Pathway Editor form for compound–disease association annotation.

types in order to enable easier management and quality control. Accounting, analysis, and editing of entered annotations by seniors experts are also performed within Pathway Editor.

References

1. Kitano H. (2007) Towards a theory of biological robustness. *Mol Syst Biol.* **3**, 137.
2. Barthelmes J, Ebeling C, Chang A, Schomburg I, Schomburg D. (2007) BRENDA, AMENDA and FRENDA: The enzyme information system in 2007. *Nucleic Acids Res.* **35**, D511–D514.
3. Selkov JE, Grechkin Y, Mikhailova N, Selkov E. (1998) MPW: The Metabolic Pathways database. *Nucleic Acids Res.* **26**, 43–45.
4. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480–D484.
5. Willis RC, Hogue CW. (2006) Searching, viewing, and visualizing data in the Biomolecular Interaction Network Database (BIND). *Curr Protoc Bioinformatics.* Chapter 8, Unit 8.9.
6. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stümpflen V, Mewes HW, Ruepp A, Frishman D. (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics.* **21(6)**, 832–834.
7. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451.
8. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, Mering CV. (2008) STRING 8 – A global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **37,** D412–416.
9. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H. (2007) IntAct – Open source resource for molecular interaction data. *Nucleic Acids Res.* **35**, D561–D565.
10. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavnath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, Kumar HG, Nagini M, Kumar GS, Jose R, Deepthi P, Mohan SS, Gandhi TK, Harsha HC, Deshpande KS, Sarker M, Prasad TS, Pandey A. (2006) Human protein reference database – 2006 update. *Nucleic Acids Res.* **34**, D411–D414.
11. Sauro HM, Bergmann FT. (2008) Standards and ontologies in computational systems biology. *Essays Biochem.* **45**, 211–222.
12. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. (2007) UniProtKB/Swiss-Prot: The manually annotated section of the UniProt KnowledgeBase. *Methods Mol Biol.* **406**, 89–112.
13. Blake JA, Harris MA. (2008) The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis. *Curr Protoc Bioinformatics.* Chapter 7, Unit 7.2.
14. Mottaz A, Yip YL, Ruch P, Veuthey AL. (2008) Mapping proteins to disease terminologies: from UniProt to MeSH. *BMC Bioinformatics.* **9**, Suppl 5, S3.
15. Chatr-Aryamontri A, Zanzoni A, Ceol A, Cesareni G. (2008) Searching the protein interaction space through the MINT database. *Methods Mol Biol.* **484**, 305–317.
16. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. (2008) miRBase: Tools for microRNA genomics. *Nucleic Acids Res.* **36**, D154–D158.
17. Wakaguri H, Yamashita R, Suzuki Y, Sugano S, Nakai K. (2008) DBTSS: Database of transcription start sites, progress report 2008. *Nucleic Acids Res.* **36**, D97–D101.
18. Bourne KZ, Ferrari DC, Lange-Dohna C, Rossner S, Wood TG, Perez-Polo JR. (2007) Differential regulation of BACE1 promoter activity by nuclear factor-kappaB

in neurons and glia upon exposure to beta-amyloid peptides. *J Neurosci Res.* **85(6)**, 1194–1204.

19. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, Walk TC, Zhang P, Karp PD. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **36**, D623–D631.
20. Pegg SC, Brown SD, Ojha S, Seffernick J, Meng EC, Morris JH, Chang PJ, Huang CC, Ferrin TE, Babbitt PC. (2006) Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry.* **45(8)**, 2545–2555.
21. Suthram S, Shlomi T, Ruppin E, Sharan R, Ideker T. (2006) A direct comparison of protein interaction confidence schemes. *BMC Bioinformatics.* **7**, 360.