D684–D688 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 15 December 2007*
doi:10.1093/nar/gkm795

STITCH: interaction networks of chemicals and proteins

Michael Kuhn[1], Christian von Mering[2], Monica Campillos[1], Lars Juhl Jensen[1,*] and Peer Bork[1,3]

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany, [2]University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland and [3]Max-Delbrück-Centre for Molecular Medicine, Robert-Rössle-Strasse 10, 13092 Berlin, Germany

Received August 14, 2007; Revised September 14, 2007; Accepted September 17, 2007

ABSTRACT

The knowledge about interactions between proteins and small molecules is essential for the understanding of molecular and cellular functions. However, information on such interactions is widely dispersed across numerous databases and the literature. To facilitate access to this data, STITCH ('search tool for interactions of chemicals') integrates information about interactions from metabolic pathways, crystal structures, binding experiments and drug–target relationships. Inferred information from phenotypic effects, text mining and chemical structure similarity is used to predict relations between chemicals. STITCH further allows exploring the network of chemical relations, also in the context of associated binding proteins. Each proposed interaction can be traced back to the original data sources. Our database contains interaction information for over 68 000 different chemicals, including 2200 drugs, and connects them to 1.5 million genes across 373 genomes and their interactions contained in the STRING database. STITCH is available at http://stitch.embl.de/

INTRODUCTION

In pharmacology and biochemistry the interplay of chemicals and proteins has been studied over many years, but much of the existing data on chemicals is either hidden in a vast amount of dispersed literature or is locked away in commercial databases such as the Chemical Abstracts Service Registry. Recently, however, several projects have begun to provide easy public access to chemical information. Resources such as PubChem (1), ChEBI (2) and ChemDB (3) provide an ever-growing inventory of the chemical space that can be used as the basis for the integration of knowledge about chemicals themselves, their biological interactions and their phenotypic effects. Thus, many problems in Chemical Biology are now becoming approachable by the academic research community.

Valuable information about the biological activity of chemicals is provided by large-scale experiments. Phenotypic effects of chemicals were first made available on a large scale by the US National Cancer Institute (NCI), which conducts anti-cancer drug screens on 60 human tumour cell lines (NCI60) (4). The patterns of growth inhibition in the different cell lines by small molecules can not only be used to judge the efficacy of individual compounds, but also to relate compounds by their mechanism of action (5,6). Other unexpected relationships between compounds can be found using the PubChem BioAssay resource, where NCI60 data and many other assays are aggregated. As of July 2007, it contains 587 highly diverse assays, ranging from studies of single molecules to high-throughput screens with over 100 000 tested substances. Recently, the Connectivity Map project (7) set out to catalogue the perturbations in gene expression upon chemical treatment. As the Connectivity Map increases in coverage of small molecules, it will develop into a very useful resource for both the similarities of chemicals and the relations of chemicals with proteins.

Information about interactions between proteins and small molecules is essential for understanding metabolism, signalling and drug treatment. Part of this information is stored in different types of databases. There are some databases that focus on the biological actions of drugs, for example DrugBank (8), TTD (9), SuperTarget and MATADOR (10). Another database, the PDSP K_i Database (11) provides protein binding constants (K_i) for compounds, combining data from the literature and internal screens. Finally, there are many pathway databases most prominently KEGG (12), MetaCyc (13) and Reactome (14). These diverse

*To whom correspondence should be addressed. Tel: +49 6221 387 296; Fax: +49 6221 387 519; Email: jensen@embl.de
Correspondence may also be addressed to Peer Bork. Tel: +49 6221 387 526; Fax: +49 6221 387 519; Email: bork@embl.de

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

knowledge sources can be integrated to provide links between chemical space and the protein universe through genome databases such as Ensembl (15) and RefSeq (16).

In order to get a complete picture of the biochemistry of metabolites, drugs and other compounds, the current challenge is to integrate the various sources of chemical knowledge into a single resource and to link it with the knowledge about proteins. For example, to be able to link phenotypic observations from cell line screens to molecular events, their interactions with proteins in the context of cellular networks are essential.

Here, we present search tool for interactions of chemical (STITCH), a search tool and resource for the interactions of chemicals and proteins. A consolidated set of chemicals is derived from PubChem. Relations between chemicals are derived from similar activity profiles in the NCI60 cell lines, from pharmacological actions assigned to chemicals in the Medical Subject Headings (MeSH) and from the literature. Chemical–chemical and chemical–protein associations are integrated from pathway and experimental databases, as well as from the literature. Lastly, as many associations as possible are annotated with interaction types.

CONSOLIDATED SET OF CHEMICALS

Chemicals are the basis of STITCH and are currently imported from PubChem. All stereoisomers and charge forms of a compound are merged into one record via the canonical SMILES string. While this might be an oversimplification, it is necessary and valid for three reasons: Stereoisomers often share names, for example the name ‘valine’ is assigned to L-valine, D-valine and a third compound without stereochemistry and therefore, it is not possible to automatically assign the synonym. Second, external databases may link to the compound with or without stereochemistry. Lastly, enantiomers with different biological activity may interconvert *in vivo*. This is the case for the drug thalidomide, where one enantiomer can be used to treat morning sickness, but the other enantiomer causes birth defects (17).

Drugs are often marketed as different salts and mixtures of the same active substance, which are represented as distinct entries in the chemical databases. As the different formulations will have the same biological effect, they are joined into one entry in STITCH. In order to do this, a list of 30 additive compounds that can be discarded was created by manual inspection of mixtures. This list includes water, ions such as calcium and organic acids such as methanesulfonic acid (which forms mesylates). When these compounds are present together with an organic compound of at least 100 g/mol, the additive compounds are discarded and the database entry of the mixture is joined to the entry of the base compound. For example, imatinib mesylate is reduced to its main form imatinib.

The PubChem database is an aggregation of many databases. Often, different databases contradict each other in the assignment of synonyms to compounds. For this reason, some of the source databases have been identified as either trusted or dubious sources. This information is used to arbitrate conflicts in the assignment of synonyms.

ASSOCIATIONS BETWEEN CHEMICALS

Building upon the set of chemicals, associations of chemicals and proteins can be imported from various sources. Taken together, these associations form a network that can be used to explore the context of chemicals and proteins. The associations derived here are combined with protein–protein interactions stored in the STRING database (18) to form one large network.

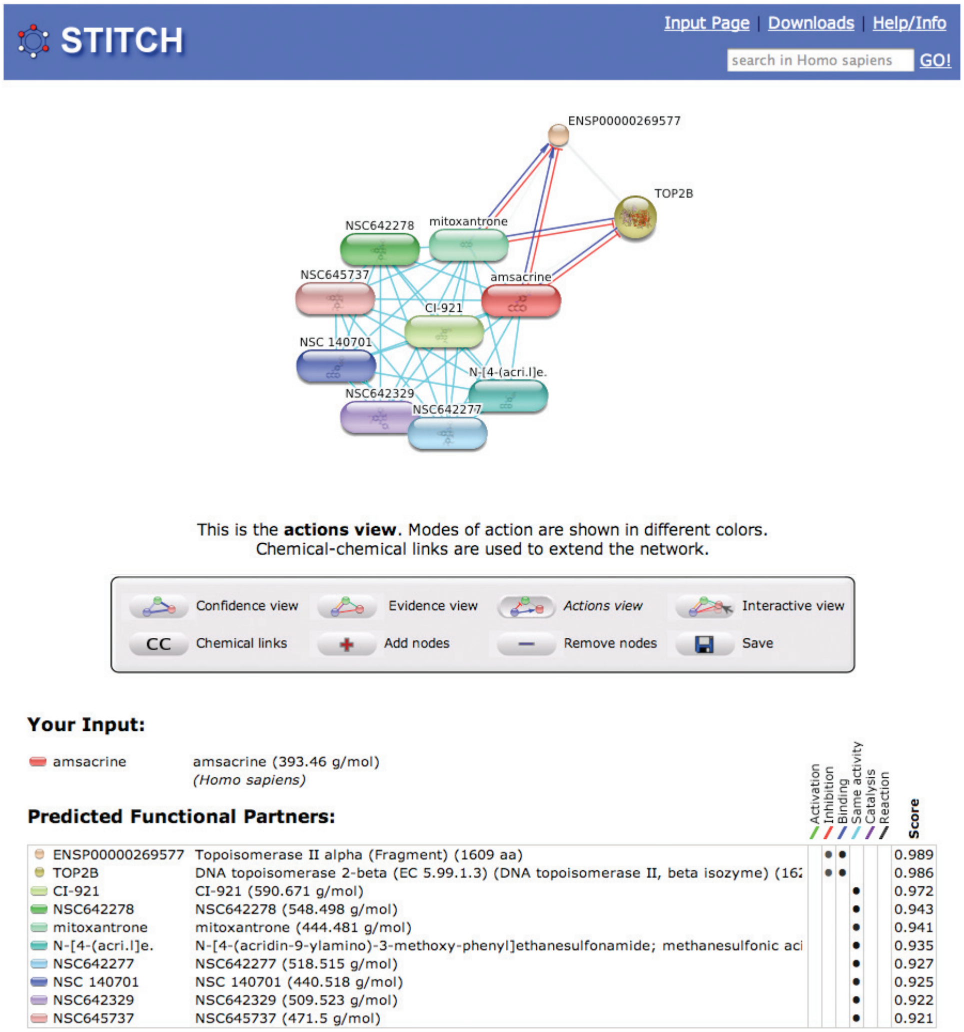
Four types of edges link chemicals in the chemical–chemical network: reactions from pathway databases, literature associations, similar structures and similar activities. Pathway databases contain records about chemical reactions that are used to derive associations. The open-source Chemistry Development Kit (19) was used to calculate chemical fingerprints and the commonly used Tanimoto 2D chemical similarity scores (20,21). Literature associations were derived in the same manner as chemical–protein associations (see subsequently).

To predict whether two chemicals have similar molecular activities, data from MeSH pharmacological actions and activities in the NCI60 screens were used. Compounds from MeSH were mapped by their name to the database of chemicals. To assess the relevance of the association by a shared pharmacological action, the number of chemicals annotated with each pharmacological action is determined and benchmarked against the probability of sharing a drug target in the MATADOR database.

An activity pattern was calculated for each compound based on the NCI60 screens, by converting its log(GI_{50}) values (the concentrations required to inhibit growth by 50%) to Z-scores in each cell line (6). Compounds with uninformative activity patterns were excluded using a threshold for the standard deviation of the Z-scores. Then, the Pearson correlation of the activity patterns was calculated for all compound pairs and benchmarked against known mechanisms of action (6,22). This allows the user to link compounds with unknown mechanism of action to well-studied compounds (Figure 1).

SOURCES OF CHEMICAL–PROTEIN INTERACTIONS

In order to link the derived chemical–chemical associations to the protein world, a variety of databases of chemical–protein interactions are imported. Experimental evidence of direct chemical–protein binding is derived from the PDSP K_i Database (11) and the protein data bank (PDB) (23). Additional interactions between metabolites and proteins are extracted from the manually annotated pathway databases such as KEGG (12), Reactome (14) and the NCI-Nature Pathway Interaction Database (http://pid.nci.nih.gov), and drug–target relations are imported from DrugBank (8) and MATADOR (10).



This is the **actions view**. Modes of action are shown in different colors.
Chemical-chemical links are used to extend the network.

Your Input:

🟥 amsacrine	amsacrine (393.46 g/mol) *(Homo sapiens)*

Predicted Functional Partners:

		Activation	Inhibition	Binding	Same activity	Catalysis	Reaction	Score
ENSP00000269577	Topoisomerase II alpha (Fragment) (1609 aa)		●	●				0.989
TOP2B	DNA topoisomerase 2-beta (EC 5.99.1.3) (DNA topoisomerase II, beta isozyme) (16:		●	●				0.986
CI-921	CI-921 (590.671 g/mol)				●			0.972
NSC642278	NSC642278 (548.498 g/mol)				●			0.943
mitoxantrone	mitoxantrone (444.481 g/mol)				●			0.941
N-[4-(acri.l]e.	N-[4-(acridin-9-ylamino)-3-methoxy-phenyl]ethanesulfonamide; methanesulfonic aci				●			0.935
NSC642277	NSC642277 (518.515 g/mol)				●			0.927
NSC 140701	NSC 140701 (440.518 g/mol)				●			0.925
NSC642329	NSC642329 (509.523 g/mol)				●			0.922
NSC645737	NSC645737 (471.5 g/mol)				●			0.921

Figure 1. Interactions of the topoisomerase II inhibitor amsacrine. Chemicals are represented as pill-shaped nodes, while proteins are shown as spheres. Nodes that are associated to each other are linked by an edge: like mitoxantrone, amsacrine is known to bind (blue lines) and inhibit (red arrows) type II topoisomerases. Other compounds have similar activities as amsacrine or mitoxantrone in the NCI60 anti-cancer drug screens and are thus predicted to have the same mechanism of action (cyan lines).

Text mining of MEDLINE and OMIM yields additional evidence, based both on a simple co-occurrence scheme and a more complex natural language processing (NLP) approach (24,25). In order to increase the coverage of the text-mining approach, groups of proteins that are described in MeSH terms are also used as entities during text mining. This allows us to capture interactions such as the binding of memantine to multimeric NMDA (*N*-methyl D-aspartate) receptors. The MeSH terms are mapped to proteins by the identifiers and synonyms provided by MeSH. If this automatic mapping is not possible, common proteins were manually assigned to candidate MeSH terms that were determined by using the annotations of abstracts with MeSH terms in MEDLINE. To reflect the decreased confidence associated with an interaction of a compound with a group of proteins rather than a single protein, the interaction score was scaled down as a function of the sequence diversity within the group.

For each individual evidence type, likelihood or relevance scores have been developed. The individual scores for a given chemical–protein or chemical–chemical interaction are then combined into one overall score (26). Chemical–protein interactions are transferred between species based on the sequence similarity of the proteins (26).

In order to be as comprehensive as possible, STITCH derives data from a variety of sources. While the sources have different focuses, it is well possible that certain areas of knowledge are not covered. In the future, more data sources will be added to further increase the coverage of STITCH.

ANNOTATION OF SMALL MOLECULE ACTIONS

Many data sources contain information about the biological or biochemical action associated with a certain interaction. This information can be stated explicitly, like in databases using the BioPAX ontology (27), or implicitly, like in crystal structures. As one of the display modes, STITCH allows the user to view a network of interactions augmented by the types of actions (Figure 2). Possible actions are: activation, inhibition, direct binding, catalysis, (bio)chemical reaction and similar activity. To avoid overloading the visual network representation with a great number of edge types, this representation is intentionally not as detailed as provided by the BioPAX ontology or recently suggested by Lu *et al.* (28).

Taken together, STITCH links molecular, cellular and phenotypic data related to small molecules and allows easy navigation in and visualization of networks of large collections of associations between chemicals as well as interactions between chemicals and proteins. It thus represents a useful resource for both in-depth and large-scale projects in Chemical Biology.

DATABASE ACCESS AND NETWORK VIEWS

Users can query the STITCH database (http://stitch.embl.de) in several ways. A full-text search is available for identifiers and common names of chemicals and proteins. Chemical structures may be entered as SMILES strings to search for similar chemicals that are stored in the database. Finally, protein sequences can be submitted to find similar proteins in the database.

When searching STITCH with a chemical as entry point, the user is presented with a network of related proteins that places the chemical into a biological context. The network can be extended to also show related chemicals, which is useful for highlighting, for example, compounds with similar pharmacological activity or metabolized forms. Querying STITCH for a protein will provide the user with a network that places the protein into its chemical and biological context. The network viewer displays chemical and protein structures and provides the user with easy access to information from resources such PubChem (1), PDB (23) and SMART (29).

To aid the user in exploring and interpreting the networks, we provide four different views. By default, the confidence view is shown (Figures 1 and 2a). There the thickness of the lines represents the confidence score of the association. In the evidence view, separate lines with different colours are used to show the type of evidence that support each interaction, for example experimental evidence or text mining. Where possible, the actions view uses different colours to visualize the types of interaction between chemicals and proteins, for example activation, inhibition or metabolization (Figure 2b). Finally, an interactive view allows the user to modify the network layout.

For large-scale analyses, the interaction network of chemicals and proteins can be downloaded from the STITCH website. We also make available synonyms lists for chemicals and proteins as well as a set of database cross-references to other chemical databases. These files are available under a Creative Commons Attribution 3.0 License.

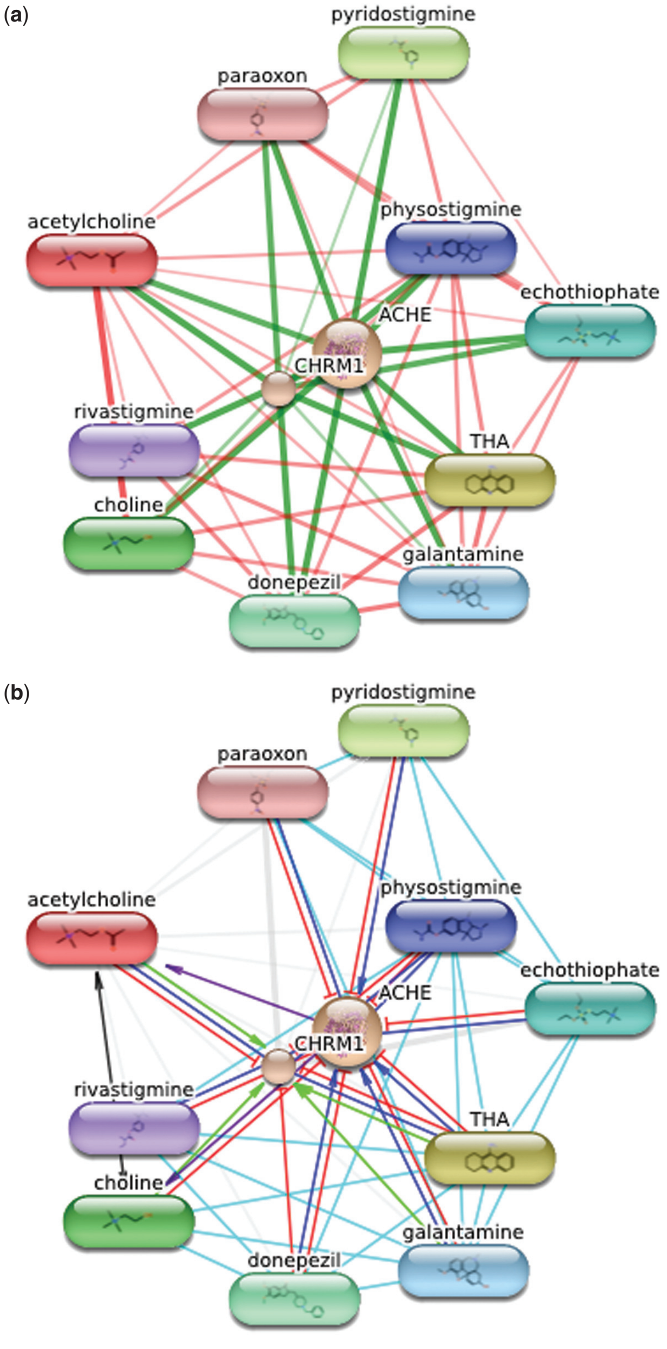


Figure 2. Network around acetylcholine and acetylcholinesterase (ACHE). (**a**) In confidence view, thicker lines represent stronger associations. (**b**) Lines and, for directed edges, arrows of different colours stand for different edge types in the actions view: binding (blue), activation (green), inhibition (red), catalysis (magenta), same activity (cyan) and reaction (black). The network shows the hydrolysis from acetylcholine to choline that is catalysed by ACHE. Several drugs, for example, the nootropic drug donepezil, inhibit ACHE.

ACKNOWLEDGEMENTS

The authors thank members of the Bork and von Mering groups for comments on the STITCH web interface. This work was funded by the BMBF QuantPro grant (0313831D). Funding to pay the Open Access publication charges for this article was provided by the European Molecular Biology Laboratory.

Conflict of interest statement. None declared.

REFERENCES

1. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.
2. Brooksbank,C., Cameron,G. and Thornton,J. (2005) The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.*, **33**, D46–D53.
3. Chen,J.H., Linstead,E., Swamidass,S.J., Wang,D. and Baldi,P. (2007) ChemDB update—full-text search and virtual chemical space. *Bioinformatics*, doi: 10.1093/bioinformatics/btm1341.
4. Weinstein,J.N., Myers,T.G., O'Connor,P.M., Friend,S.H., Fornace,A.J., Kohn,K.W., Fojo,T., Bates,S.E., Rubinstein,L.V. *et al.* (1997) An information-intensive approach to the molecular pharmacology of cancer. *Science*, **275**, 349.
5. Huang,R., Wallqvist,A., Thanki,N. and Covell,D.G. (2005) Linking pathway gene expressions to the growth inhibition response from the National Cancer Institute's anticancer screen and drug mechanism of action. *Pharmacogenomics J.*, **5**, 399.
6. Rabow,A.A., Shoemaker,R.H., Sausville,E.A. and Covell,D.G. (2002) Mining the National Cancer Institute's tumor-screening database: identification of compounds with similar cellular activities. *J. Med. Chem.*, **45**, 840.
7. Lamb,J., Crawford,E.D., Peck,D., Modell,J.W., Blat,I.C., Wrobel,M.J., Lerner,J., Brunet,J.P., Subramanian,A. *et al.* (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1935.
8. Wishart,D.S., Knox,C., Guo,A.C., Shrivastava,S., Hassanali,M., Stothard,P., Chang,Z. and Woolsey,J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–D672.
9. Chen,X., Ji,Z.L. and Chen,Y.Z. (2002) TTD: therapeutic target database. *Nucleic Acids Res.*, **30**, 415.
10. Günther,S., Kuhn,M., Dunkel,M., Campillos,M., Senger,C., Petsalaki,E., Ahmed,J., Urdiales,E.G., Gewiess,A. *et al.* (2008) SuperTarget and matador: resources for exploring drug-target relationships. *Nucleic Acids Res.*, this issue.
11. Roth,B., Lopez,E., Patel,S. and Kroeze,W. (2000) The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches? *Neuroscientist*, **6**, 262.
12. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
13. Caspi,R., Foerster,H., Fulcher,C.A., Hopkinson,R., Ingraham,J., Kaipa,P., Krummenacker,M., Paley,S., Pick,J. *et al.* (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **34**, D511–D516.
14. Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G.R., Wu,G.R. *et al.* (2005) Reactome: a knowledge base of biological pathways. *Nucleic Acids Res.*, **33**, D428–D432.
15. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
16. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
17. Eriksson,T., Björkman,S., Roth,B., Fyge,A. and Höglund,P. (1995) Stereospecific determination, chiral inversion in vitro and pharmacokinetics in humans of the enantiomers of thalidomide. *Chirality*, **7**, 52.
18. von Mering,C., Jensen,L.J., Kuhn,M., Chaffron,S., Doerks,T., Krüger,B., Snel,B. and Bork,P. (2007) STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.
19. Steinbeck,C., Hoppe,C., Kuhn,S., Floris,M., Guha,R. and Willighagen,E.L. (2006) Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.*, **12**, 2120.
20. Martin,Y.C., Kofron,J.L. and Traphagen,L.M. (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, **45**, 4358.
21. Willett,P., Barnard,J.M. and Downs,G.M. (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.*, **38**, 983–996.
22. Weinstein,J.N., Kohn,K.W., Grever,M.R., Viswanadhan,V.N., Rubinstein,L.V., Monks,A.P., Scudiero,D.A., Welch,L., Koutsoukos,A.D. *et al.* (1992) Neural computing in cancer drug development: predicting mechanism of action. *Science*, **258**, 447–451.
23. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 242.
24. Jensen,L., Saric,J. and Bork,P. (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.*, **7**, 129.
25. Saric,J., Jensen,L.J., Ouzounova,R., Rojas,I. and Bork,P. (2006) Extraction of regulatory gene/protein networks from Medline. *Bioinformatics*, **22**, 645–650.
26. von Mering,C., Jensen,L.J., Snel,B., Hooper,S.D., Krupp,M., Foglierini,M., Jouffre,N., Huynen,M.A. and Bork,P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, **33**, D433–D437.
27. Stromback,L. and Lambrix,P. (2005) Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics*, **21**, 4407.
28. Lu,L.J., Sboner,A., Huang,Y.J., Lu,H.X., Gianoulis,T.A., Yip,K.Y., Kim,P.M., Montelione,G.T. and Gerstein,M.B. (2007) Comparing classical pathways and modern networks: towards the development of an edge ontology. *Trends Biochem. Sci.*, **32**, 310–321.
29. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.