D552–D556 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 6 November 2009*
doi:10.1093/nar/gkp937

STITCH 2: an interaction network database for small molecules and proteins

Michael Kuhn[1], Damian Szklarczyk[2], Andrea Franceschini[3], Monica Campillos[4], Christian von Mering[3], Lars Juhl Jensen[2], Andreas Beyer[1] and Peer Bork[4,5,*]

[1]Biotechnology Center, TU Dresden, 01062 Dresden, Germany, [2]Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Blegdamsvej 3b, 2200 Copenhagen, Denmark, [3]Institute of Molecular Biology and Swiss Institute of Bioinformatics, University of Zurich, Switzerland, [4]European Molecular Biology Laboratory, Meyerhof street 1, 69117 Heidelberg and [5]Max-Delbrück-Centre for Molecular Medicine, Robert-Rössle-Strasse 10, 13092 Berlin, Germany

Received September 15, 2009; Revised October 8, 2009; Accepted October 9, 2009

ABSTRACT

Over the last years, the publicly available knowledge on interactions between small molecules and proteins has been steadily increasing. To create a network of interactions, STITCH aims to integrate the data dispersed over the literature and various databases of biological pathways, drug–target relationships and binding affinities. In STITCH 2, the number of relevant interactions is increased by incorporation of BindingDB, PharmGKB and the Comparative Toxicogenomics Database. The resulting network can be explored interactively or used as the basis for large-scale analyses. To facilitate links to other chemical databases, we adopt InChIKeys that allow identification of chemicals with a short, checksum-like string. STITCH 2.0 connects proteins from 630 organisms to over 74 000 different chemicals, including 2200 drugs. STITCH can be accessed at http://stitch.embl.de/.

INTRODUCTION

The effects of small molecules on organisms have long been the focus of biochemistry and pharmacology. Over the last years there has been a considerable increase in the number of high-throughput screens that have been performed using chemical libraries (1–3). At the same time, the molecular targets of individual chemicals are being studied in ever greater detail (4,5). There also is a great interest in chemical biology approaches, using small molecules to perturb cellular functions (6). For the design and interpretation of these studies, the context of the chemicals and proteins needs to be considered. For example, in the case of high-content screening for specific cellular effects, it is important to know whether the active chemicals already have known activities that can explain the observed effects, or whether novel mechanisms of actions might be present. Therefore, we have developed a 'search tool for interactions of chemicals' (STITCH) both as a large-scale, downloadable database of interaction data and as an interactive web tool for the exploration of interaction networks (Figure 1). Since its first release (7), STITCH is being accessed by over one hundred scientists each week and has been used as a source of protein–chemical associations e.g. by Prathipati *et al.* (8), who used the STITCH network to automatically extract the targets of anti-tuberculosis compounds in *Mycobacterium tuberculosis*.

Here, we present the second version of STITCH. In addition to the sources of protein–chemical interactions included in the previous version—PDSP K_i Database (9), Protein Data Bank (PDB) (10), KEGG (11), Reactome (12), NCI-Nature Pathway Interaction Database (http://pid.nci.nih.gov), DrugBank (13) and MATADOR (14)—we now further include interactions imported from GLIDA (15), PharmGKB (16,17), Comparative Toxicogenomics Database (CTD) (18) and BindingDB (19). These added databases mainly provide information on interactions between human proteins and drugs or drug-like molecules.

The imported sources of information are scored separately and then combined with information from text-mining (7). Databases which contain manually annotated interactions receive high scores, while interactions based on experimental information are scored by the confidence or relevance of the reported interaction. The number of high-confidence (score ≥ 0.7) human chemical–protein interactions increased from 51 000 to 85 000. For these high-confidence interactions, the number of interacting

*To whom correspondence should be addressed. Tel: +49 6221 387 8526; Fax: +49 6221 387 517; Email: bork@embl.de

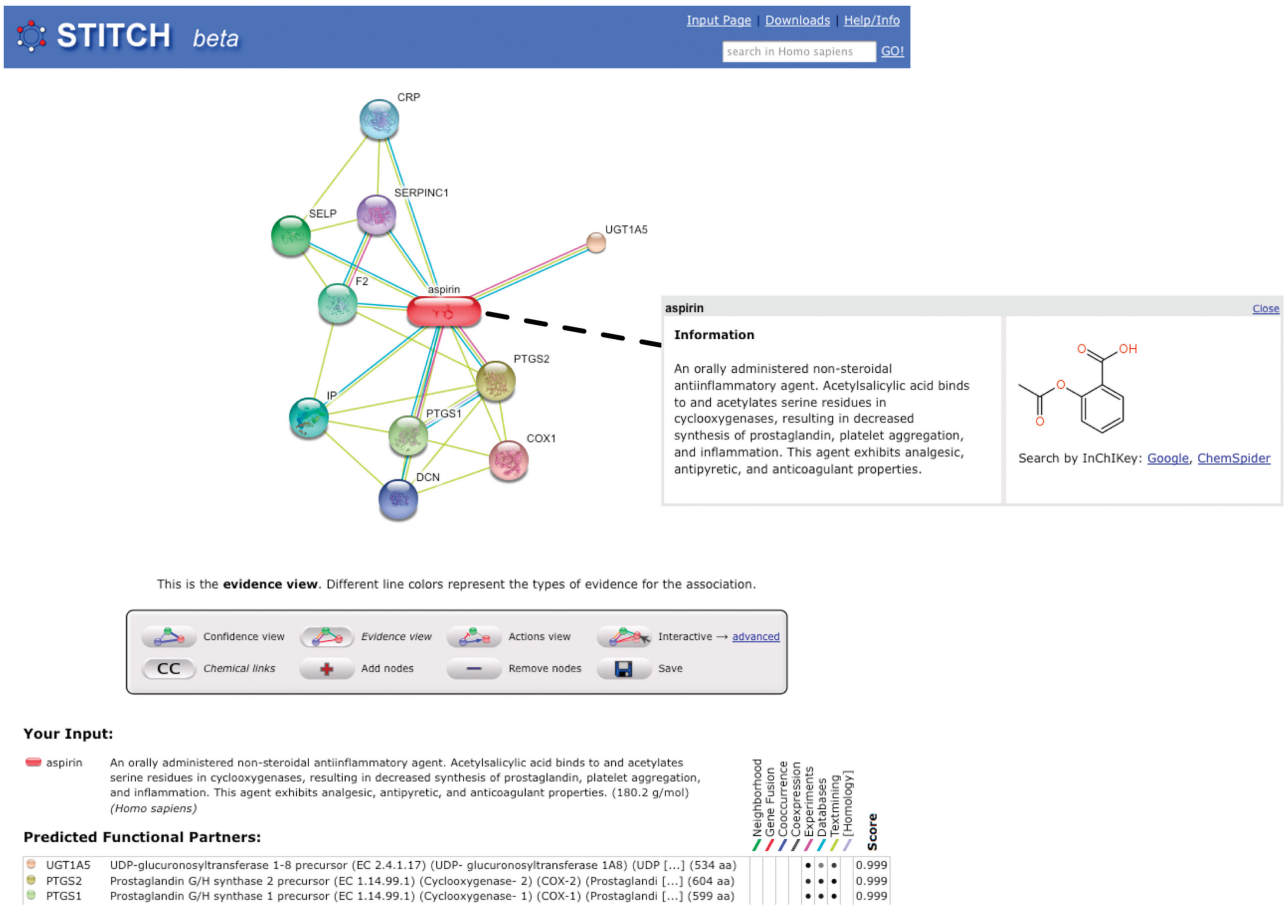

Figure 1. Interaction network around aspirin. Human proteins predicted to interact with aspirin according to different sources of evidence are shown. Edges are colored according to the source of evidence (magenta: experimental information, cyan: manually curated databases, yellow: text-mining). Clicking on the node 'aspirin' will display a pop-up showing the structure and description.

human proteins increases from 5300 to 7400 (as STITCH is locus-based, only one gene product is counted per gene).

INCREASING THE NUMBER OF SPECIFIED ACTIONS

The STITCH network is created by mapping interactions from the sources mentioned above and from text-mining onto a consolidated set of chemicals that has been derived from PubChem, assigning a confidence score for each interaction (7). The newly-derived protein–chemical and chemical–chemical associations are then complemented with protein–protein interactions from the STRING database (20). In the previous version of STITCH (7), we began to import 'actions' derived from natural language processing (NLP), pathway and interaction databases. These actions specify the nature of the interaction independent of the source of interaction information. For example, a 'binding' action could be derived from a binding affinity database and an 'inhibition' action could be imported from NLP. We have greatly extended the set of available actions by further importing action types from GLIDA (15), PharmGKB (16,17), CTD (18), BindingDB (19) and a manually annotated set of interactions. This set of interactions has been curated from DrugBank (13) records, results from NLP analysis of PubMed abstracts, Medical Subject Headings (MeSH) pharmacological actions, Anatomical Therapeutic Chemical classification (ATC) entries and a review paper on drugs and their targets (21). An action has been assigned to 81% of the high-confidence human chemical–protein interactions. The number of available edges with a high-confidence action annotation increased from 44 000 to 65 000 human chemical–protein interactions.

HANDLING OF CHEMICAL STRUCTURES

As described previously (7), STITCH creates a consolidated set of chemicals from PubChem (22) by merging stereo isomers and salt forms of the same molecule into one compound. This is done to ensure that all information about the same biologically active entity is merged. While this works very well for drugs that can be supplied in different formulations (e.g. different salt forms), it also has limitations, especially regarding carbohydrates. It is our long-term goal to associate interactions both with the individual isomer and the

merged structure. For now, we have taken the step to explicitly display all the different compounds that have been deemed biologically equivalent (Figure 2).

Recently, the International Union of Pure and Applied Chemistry (IUPAC) has standardized an open format for chemical structures, namely the IUPAC International Chemical Identifier (InChI). In addition to the existing capability to search chemical structures using SMILES string, we have now also implemented a search for InChIs. We use the tool Open Babel to convert InChIs to SMILES strings, which are in turn searched against our chemical database by using hashed fingerprints as implemented in the open-source Chemical Development Kit (23). Furthermore, we have implemented a search for InChIKeys, which are short strings that represent an encoded (hashed) form of the chemical structure. InChIKeys consist of two parts, the first of which is based on the chemical connectivity, whereas the second part contains information about stereochemistry, tautomers and other structural variations. As STITCH currently considers structures with the same connectivity to be equivalent (thus merging stereo isomers), only the first part of the InChIKey is queried against our chemical database. We also use this part of the InChIKey to provide links to Google and ChemSpider.

USER INTERFACE IMPROVEMENTS

Many proteins, especially drug targets, have a large number of high-scoring interactions with small molecules in the STITCH network. In this case, a network centered on such a protein will only show chemicals unless very many interaction partners are requested to be shown (Figure 3a). In order to allow the user to see more of the context of the query protein, we now offer the option to show a network in which proteins and chemicals each make up more than a third of the nodes (Figure 3b). When this option is selected, only a limited number of the highest-scoring chemicals are displayed. Further chemicals are omitted in favor of proteins (or vice versa) and their number is shown to the user. If the network consists of only chemicals, but no proteins are available at the current

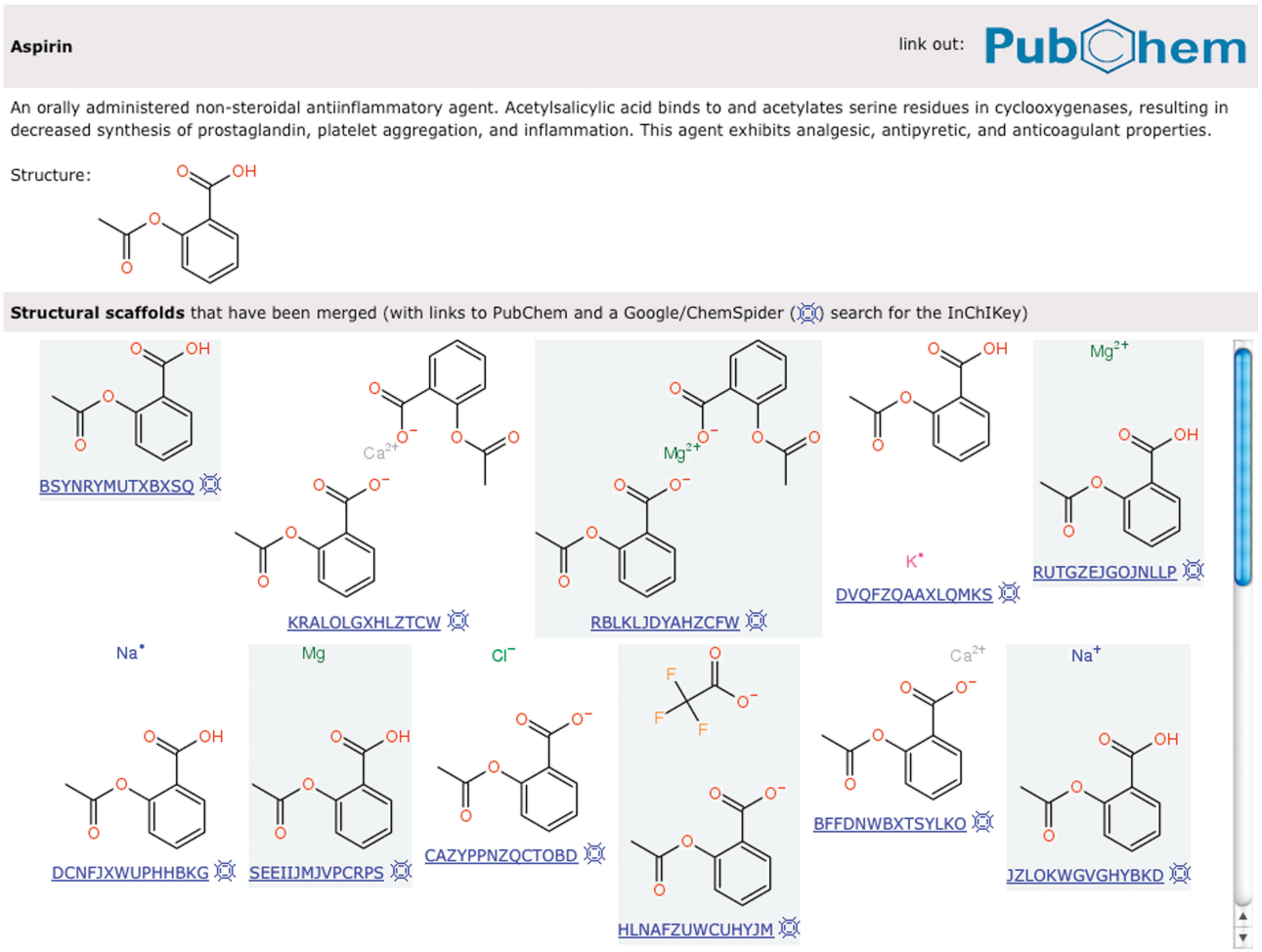

Figure 2. Different structural scaffolds corresponding to aspirin. For the drug aspirin, a link to PubChem and a short description is shown. Different salts of aspirin that will have the same bioactivity have been consolidated and merged with the main, uncharged form. Below each chemical structure, the first part of the InChIKey is shown, corresponding to an encoded (hashed) description of the structure. This short string can be used to search for more information about the compound on the Internet.

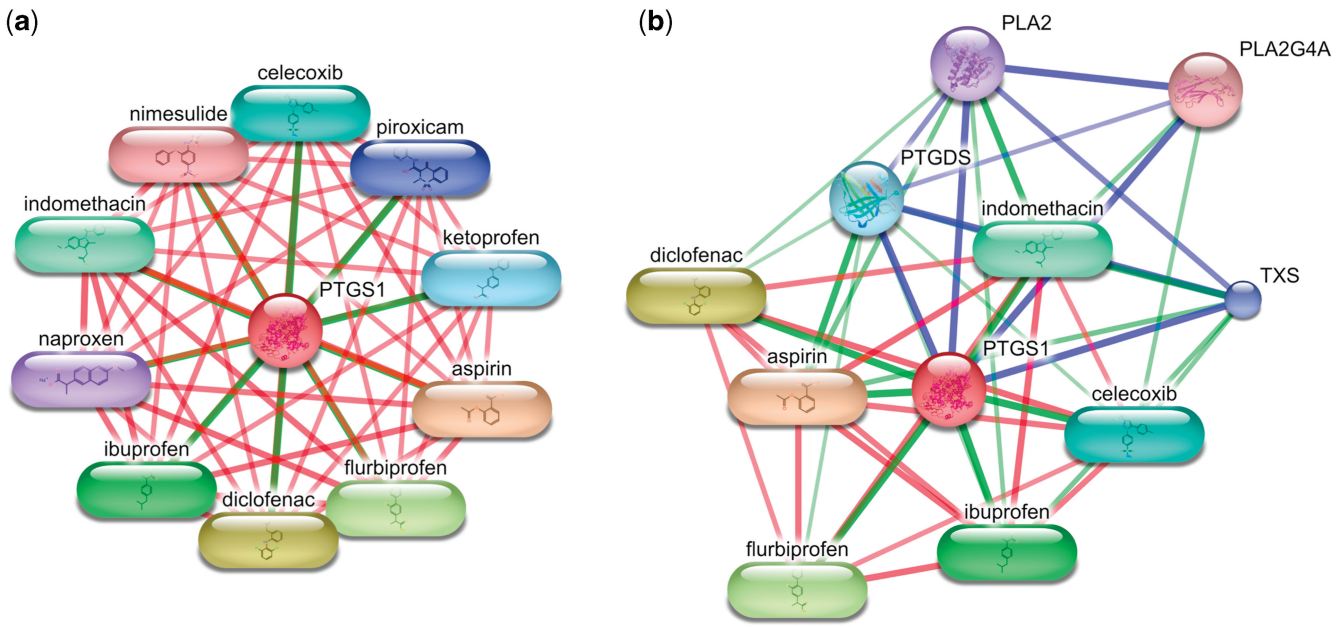

Figure 3. Interactions of prostaglandin-endoperoxide synthase 1 (PTGS1). (**a**) The highest-scoring interaction partners of PTGS1 are non-steroidal anti-inflammatory drugs (NSAIDs). As the confidence scores for these interactions are very high, no interacting proteins are shown. (**b**) The user may ask STITCH to display more of the interaction context and to let at least one-third of the interaction partners be proteins. In this case, STITCH is skipping 19 high-scoring chemicals in order to include four interacting proteins. In both networks, the color of the edge corresponds to the type of connected nodes (e.g. green: chemical–protein interaction) and the width of the edge correlates with the confidence score.

settings (e.g. due to a minimum score limit), then the option to show more context is not shown.

Previously, STITCH required the user to select an organism when searching for interactions with a chemical. Now, this is not required anymore. When no organism is selected, the organism with the highest-scoring interaction partners is selected. In case of multiple organisms with equal scores, human and several model organisms are preferentially selected. (Human is one of the highest-ranking species for 60% of the chemicals with protein–chemical interactions.) For example, the binding between the antipsychotic agent fluspiperone and the 5–hydroxytryptamine (serotonin) receptor 7 has only been studied in mouse and rat. Consequently, a user searching for this compound would be directed to the protein–chemical interaction network in mouse. It is also possible to restrict the search to different levels of the NCBI taxonomy (24), e.g. bacteria, fungi or rodents.

While central repositories of gene annotations exist, no such information is available in a centralized manner for chemicals. To be able to display text annotation for chemicals, we have imported information from the following databases: DrugBank (13), National Cancer Institute (NCI) thesaurus (25), MeSH descriptors and qualifiers. Using STITCH's dictionary of chemical synonyms we mapped compounds from these databases to STITCH identifiers. In case where descriptions are available for different forms of the same compound (e.g. different salt forms, which have been merged in STITCH), we have automatically assigned the description of the main compound. Any remaining inconsistencies were manually resolved. For each chemical we have assigned the text annotation from only one source, prioritizing sources as follows: NCI (descriptions), DrugBank (descriptions), DrugBank (pharmacology), DrugBank (drug category), MeSH (pharmacological action), NCI (tags) and MeSH (scope note). Descriptions are available for 33 352 chemicals, covering 33% of the chemicals with interactions.

USE CASES

The STITCH homepage offers several short tutorials to introduce the different query options (e.g. searching for a single identifier or multiple chemical structures). A search for 'aspirin' on the homepage will lead to the interaction network shown in Figure 1. Here, the main interactors of the drug are shown in human (which is selected automatically as described above). The known main targets, PTGS1 and PTGS2, are connected by very high scores. While most interaction partners are backed up by evidence from manually curated databases and are therefore very reliable, one interaction is derived only from text-mining: COX1 is actually a false positive arising from an ambiguous synonym.

Taken together, STITCH 2 offers an enlarged set of protein–chemical interactions, extended inter-database operability, increased query options and an improved user interface. STITCH can be accessed at http://stitch.embl.de/. Users can explore the interaction network interactively or download the complete set of interactions. In addition, we provide an application programming interface (API) to let scripts resolve identifiers and

retrieve interaction networks either as an image or in standard network formats (20).

FUNDING

Klaus Tschira Foundation (to M.K. and A.B.). Novo Nordisk Foundation Center for Protein Research (partial). Funding for open access charge: European Molecular Biology Laboratory.

Conflict of interest statement. None declared.

REFERENCES

1. Han,L., Wang,Y. and Bryant,S.H. (2009) A survey of across-target bioactivity results of small molecules in PubChem. *Bioinformatics*, **25**, 2251–2255.
2. Zanzoni,A., Soler-López,M. and Aloy,P. (2009) A network medicine approach to human disease. *FEBS lett*., **583**, 1759–1765.
3. Peterson,R.T. (2008) Chemical biology and the limits of reductionism. *Nature Chem. Biol*., **4**, 635–638.
4. Ovaa,H. and van Leeuwen,F. (2008) Chemical biology approaches to probe the proteome. *Chembiochem : Eur. J. Chem. Biol*., **9**, 2913–2919.
5. Rix,U. and Superti-Furga,G. (2009) Target profiling of small molecules by chemical proteomics. *Nature Chem. Biol*., **5**, 616–624.
6. Edwards,A.M., Bountra,C., Kerr,D.J. and Willson,T.M. (2009) Open access chemical and clinical probes to support drug discovery. *Nature Chem. Biol*., **5**, 436–440.
7. Kuhn,M., von Mering,C., Campillos,M., Jensen,L.J. and Bork,P. (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res*., **36**, D684–D688.
8. Prathipati,P., Ma,N.L., Manjunatha,U.H. and Bender,A. (2009) Fishing the target of antitubercular compounds: in silico target deconvolution model development and validation. *J. Proteome Res*., **8**, 2788–2798.
9. 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.
10. 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.
11. 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.
12. Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G.R., Wu,G.R., Matthews,L. *et al*. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*., **33**, D428–D432.
13. 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.
14. Günther,S., Kuhn,M., Dunkel,M., Campillos,M., Senger,C., Petsalaki,E., Ahmed,J., Urdiales,E.G., Gewiess,A., Jensen,L.J. *et al*. (2008) SuperTarget and Matador: Resources for exploring drug-target relationships. *Nucleic Acids Res*., **36**, D919–D922.
15. Okuno,Y., Yang,J., Taneishi,K., Yabuuchi,H. and Tsujimoto,G. (2006) GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Res*., **34**, D673–D677.
16. Gong,L., Owen,R.P., Gor,W., Altman,R.B. and Klein,T.E. (2008) PharmGKB: an integrated resource of pharmacogenomic data and knowledge. *Curr. Protoc. Bioinformatics*, Chapter 14(Unit 14):17.
17. Hewett,M., Oliver,D.E., Rubin,D.L., Easton,K.L., Stuart,J.M., Altman,R.B. and Klein,T.E. (2002) PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res*., **30**, 163–165.
18. Davis,A.P., Murphy,C.G., Saraceni-Richards,C.A., Rosenstein,M.C., Wiegers,T.C. and Mattingly,C.J. (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res*., **37**, D786–D792.
19. Liu,T., Lin,Y., Wen,X., Jorissen,R.N. and Gilson,M.K. (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res*., **35**, D198–D201.
20. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M. *et al*. (2009) STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*., **37**, D412–D416.
21. Imming,P., Sinning,C. and Meyer,A. (2006) Drugs, their targets and the nature and number of drug targets. *Nature Rev. Drug Disc*., **5**, 821–834.
22. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al*. (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **35**, D5–D12.
23. Steinbeck,C., Hoppe,C., Kuhn,S., Floris,M., Guha,R. and Willighagen,E. (2006) Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des*., **12**, 2111–2120.
24. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al*. (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **35**, D5–D12.
25. Sioutos,N., de Coronado,S., Haber,M.W., Hartel,F.W., Shaiu,W.L. and Wright,L.W. (2007) NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Inform*., **40**, 30–43.