Published online 8 October 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D367–D372***
doi:10.1093/nar/gkq906

ChemProt: a disease chemical biology database

Olivier Taboureau[1,*], Sonny Kim Nielsen[1], Karine Audouze[1], Nils Weinhold[1], Daniel Edsgärd[1], Francisco S. Roque[1], Irene Kouskoumvekaki[1], Alina Bora[2], Ramona Curpan[2], Thomas Skøt Jensen[1], Søren Brunak[1] and Tudor I. Oprea[1,3]

[1]Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, DK-2800 Denmark, [2]Department of Computational Chemistry, Institute of Chemistry, Romanian Academy, Timisoara 300223, Romania and [3]Department of Biochemistry and Molecular Biology, Division of Biocomputing, University of New Mexico School of Medicine, Albuquerque, New Mexico 87131, USA

Received August 12, 2010; Revised September 16, 2010; Accepted September 22, 2010

ABSTRACT

Systems pharmacology is an emergent area that studies drug action across multiple scales of complexity, from molecular and cellular to tissue and organism levels. There is a critical need to develop network-based approaches to integrate the growing body of chemical biology knowledge with network biology. Here, we report ChemProt, a disease chemical biology database, which is based on a compilation of multiple chemical–protein annotation resources, as well as disease-associated protein–protein interactions (PPIs). We assembled more than 700 000 unique chemicals with biological annotation for 30 578 proteins. We gathered over 2-million chemical–protein interactions, which were integrated in a quality scored human PPI network of 428 429 interactions. The PPI network layer allows for studying disease and tissue specificity through each protein complex. ChemProt can assist in the *in silico* evaluation of environmental chemicals, natural products and approved drugs, as well as the selection of new compounds based on their activity profile against most known biological targets, including those related to adverse drug events. Results from the disease chemical biology database associate citalopram, an antidepressant, with osteogenesis imperfect and leukemia and bisphenol A, an endocrine disruptor, with certain types of cancer, respectively. The server can be accessed at http://www.cbs.dtu.dk/services/ChemProt/.

INTRODUCTION

The old drug design paradigm, i.e. drugs interact selectively with one or two targets (proteins), resulting in treatment and prevention of disease, is now challenged by several studies that show most drugs interacting with multiple targets ('polypharmacology') (1,2). For example, celecoxib, often considered a selective cyclooxygenase-2 non-steroidal anti-inflammatory drug (NSAID), has been documented to be active on at least two additional targets, namely carbonic anhydrase II and 5-lipoxygenase (3). Rosiglitazone, which has been used for the treatment of type II diabetes mellitus, not only stimulates the peroxisome proliferator activated receptor γ, but also blocks interferon gamma-induced chemokine expression in Graves disease or ophthalmopathy (4). Polypharmacology is not always beneficial, as it often causes side effects: Cisapride, which acts as a serotonergic 5-HT4 receptor agonist, as well as astemizole, which blocks histamine H1 receptors (H1Rs), have both been withdrawn from all markets due to the risk of fatal cardiac arrhythmia associated with their blockade of the hERG potassium ion channel, an unanticipated and undesirable 'anti-target' associated to QT prolongation and 'torsades de pointes' (5). However, 'target' and 'anti-targets' are dynamic attributes, as exemplified by the case of H1R antagonists and their (in)ability to achieve clinically significant levels in the brain, influenced by the ATP-binding cassette transporter ABCB1 (also known as P-glycoprotein), which effluxes some of these drugs from the brain (6). Acquiring knowledge of the complete pharmacology profile has inspired new strategies to predict and to characterize drug-target associations in order to improve the success rates of current drug discovery paradigms, i.e. increase the efficacy and reduce toxicity and adverse effects (2).

*To whom correspondence should be addressed. Tel: +45 4525 2489; Fax: +45 4593 1585; Email: otab@cbs.dtu.dk
Correspondence may also be addressed to Tudor I. Oprea. Tel: +45 4525 2477; Fax: +45 4593 1585; Email: tuop@cbs.dtu.dk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

As large-scale chemical bioactivity databases are being assembled, the polypharmacology (i.e. high affinity bioactivity across related targets) and promiscuity (i.e. low affinity across multiple families) of chemicals are expanding the chemical space for druggable targets (7). These studies are often focused on specific protein families, such as G-protein coupled receptors (8), nuclear receptors (9) and kinases (10), but global pharmacology profiles of chemicals are considered as well (1,2). Recent chemoinformatics advances support the development of polypharmacology data mining, e.g. via iPHACE, an integrative web-based tool that enables pharmacological space navigation for small molecule drugs (11) or based on a Similarity Ensemble Approach (SEA) to relate protein pharmacology by ligand chemistry (12). Biological information can also be retrieved for a large set of chemical compounds through PubChem (13), CheBI and ChEMBL (14).

Two conceptual developments support polypharmacology: systems pharmacology, aimed at drug actions in the context of regulatory networks (15); and systems chemical biology (16), which introduces chemical awareness in systems biology. Since proteins rarely operate in isolation inside and outside cells, but rather function in highly interconnected cellular pathways, interactome networks have been developed by data integration. Yildirim *et al*. (17) combined FDA-approved drugs with a human protein–protein interaction (PPI) network (human interactome) in order to analyze the interrelationships between drug targets and disease–gene products i.e. disease–proteins. Similar work has been based on PubChem bioassays as source of polypharmacology (18). The use of side-effect similarity has been proposed on the assumption that drugs with similar side-effects are likely to interact with similar target proteins (19). Recent advances include a protein–protein association network based on the chemical toxicology of environmental chemicals (20) and a human disease network linking disorders and disease genes to various known phenotypes (21).

Our goal in the present work was to develop a disease chemical biology server, called ChemProt, based on the integration of chemical–protein annotation resources that are now accessible from large repositories, and curated disease-linked PPI data (22). ChemProt is designed to assist the elucidation of drug actions in the context of cellular and disease networks. Further to that, it allows the identification of additional genes that may play major roles in modulating chemical response i.e. to drugs, environmental chemicals and natural products, thus leading to new options in drug discovery and environmental chemical evaluation. Lastly, the ChemProt server could contribute to drug repurposing as well as to the investigation of chemicals related to anti-targets and adverse drug events.

IMPLEMENTATION

Data sources

We first gathered chemical–protein interaction data from different open source databases i.e. ChEMBL (version chembl_05) (14), BindingDB (23), PDSP Ki Database (24), DrugBank (version2.5) (25), PharmGKB (26) and two commercial databases, WOMBAT (version 2009) and WOMBAT-PK (version 2008) (7). Active compounds from the PubChem bioassay (2010) have been collected as well (13). We considered only active compounds from 'confirmatory' assays in order to capture high-confidence chemical–protein annotations from PubChem. These databases provide experimental evidence of chemical–protein interactions. Drug-target information was collected from DrugBank and PharmGKB. In addition, we integrated chemical–protein associations from CTD (version 2009) (27) and STITCH (version STITCH 2.0) (28). These last two databases consider the effect or modulation (positive or negative) of a chemical on proteins, other than that defined as binding activity. Examples include gene expression or pathway data, where the deregulation of a gene by a chemical may be not due to a physical interaction between the two entities but a response at a cellular level. Duplicate chemicals from the multiple databases were found by using InChI keys and were merged into a single ChemProt ID. However, the biological information associated to each chemical was conserved for users looking on selective databases. Overall, the final database contains 700 000 distinct molecules annotated for 30 578 proteins.

Descriptors and similarity measurement

The chemical structure of the molecules was encoded using two rather different types of fingerprints. The 166 MACCS keys, encode the presence or absence of predefined substructural or functional groups (29). On the other hand, a more complex 3-point pharmacophore fingerprint (GpiDAPH3) is based on an expansion of the PATTY pharmacophore feature recognition scheme of a 2D structure (30). This scheme assigns one or more pharmacophore feature types to all atoms in a molecule using a predefined list of SMART queries. The list of pharmacophore feature types comprises: hydrogen-bond donor (D), hydrogen-bond acceptor (A), polar (P) and hydrophobic (H). In addition, an extra label (p or pi) is added to each feature if the originating atom or group is sp^2-hybridized or planar for other reasons. The GpiDAPH3 pharmacophore feature scheme is expressed in 2D as triplet feature combinations with a graph based inter-atom distance binning scheme. Both fingerprints are implemented in the Molecular Operating Environment (MOE, version 2008.10) (31). The similarity between two molecules is measured using the Tanimoto coefficient (Tc), a method of choice for the computation of fingerprint-based similarity (32). The Tc is defined as the number of bits in common divided by the total number of used bits in both molecules. For any pair of chemicals, Tc assumes values between 0 and 1. A high Tc represents high similarity.

PPI network

The human interactome used is an in-house protein–protein interaction network inferred from experiments in both humans and model organisms (22). Using an

elaborate scoring scheme, all interactions have been validated against a gold standard (33). The current interactome contains 428 429 unique protein–proteins interactions derived from source databases such as BIND (34), GRID (35), MINT (36), dip_full (37), HPRD (38), intact (39), mppi (40), MPact (41), Reactome (42) and KEGG (43). Data are transferred between organisms by using the Inparanoid orthology database (44). In total the human interactome comprises 22 997 genes.

Human disease genes and complexes

Based on a previous study (45), disease-associated protein complexes were associated to the chemical–protein annotation by mining OMIM (46) and GeneCards (47), two data resources for genes association to diseases, we collected a list of 2227 unique disease-related proteins and mapped the complexes of genes to disease. Similarly, complexes of genes were mapped to Gene Ontology (GO) terms (48) and tissues by using the expression data from 73 non-disease tissues from the Novartis Research Foundation Gene Expression Database (GNF) (49) and Human Protein Atlas (50). Users of ChemProt can thus retrieve gene complexes that are related to a query chemical and visualize the annotations of each complex.

APPLICATIONS

Chemical–protein interactions

Chemicals can be searched using a common name, SMILES and by drawing the 2D structure, or retrieved through their annotation to a protein. Users can then choose the descriptor space and the Tc threshold to be used for similarity search. Following a successful query, hits grouped by species will be returned, together with computed physico-chemical properties such as Molecular Weight, LogP, the number of hydrogen bond donors and acceptors, the number of rigid bonds and the number of rings, based on the Marvin applet from Chemaxon (51). Hits are provided separately for known annotations, and for prediction of small molecule bioactivity, respectively. The biochemical and pharmacological effects of a chemical, e.g. substrate, inhibitor, agonist or antagonist, are provided if such information is available, together with hyperlinks to UniProt and Ensembl, which lead to more information on protein sequence and function, respectively.

From chemical–protein interactions to complex protein–disease associations

The unique feature of ChemProt is that it offers the user the possibility to get information at a cellular level, by linking chemically-induced biological perturbations to specific tissues and phenotypes.

Proteins that are both affected by a chemical and participate in one or more protein complexes are highlighted in the results table of the ChemProt server. By clicking on the protein, the user is redirected to the 'Disease complexes' server and has to choose which complex to visualize. On the 'Disease complexes' server, size and illustrations of the protein network are provided. Additionally, enrichment analysis results of the proteins in the complex are shown, with respect to disease association (OMIM, BioAlma), GO terms (biological process, cellular component) and tissue specificity (Human Protein Atlas, GNF). To ensure that the complexes were biologically relevant entities, the enrichment of the biological terms (OMIM, GO,...) was compared to randomly generated complexes (1.0e6). The significances were calculated using a hyper-geometric test and the *P*-value for the most significant enriched term for each of the data types was calculated as previously described (45). The table presenting the OMIM enrichment results is interactively linked with an illustration of the protein complex where proteins associated with the selected disease are colored yellow.

Output of the chemical–proteins interactions and disease complexes can be downloaded from the ChemProt website. In addition, the 'Reflect' service provides further information on chemicals and genes (52). 'Reflect' tags gene, protein and small molecule names in text and offers the opportunity to quickly view additional information on the ChemProt results, including synonyms, protein sequences, domains, 3D structures and subcellular location.

EXAMPLES

With the integration of several databases, ChemProt not only provides pharmacological information, but also includes biological data associated to environmental chemicals and natural products. As seen in the examples below, ChemProt can be queried for drugs as well as environmental chemicals. A search for citalopram, an antidepressant, illustrates the complementarity of the integrated databases within ChemProt (Figure 1). Marketed as a selective serotonin reuptake inhibitor (SSRI) (DrugBank), this drug displays bioactivity on seven human proteins (ChEMBL). Via ChemProt, four other proteins (DRD3, 5HT1B, 5HT3, ADRA2A) are retrieved from the Ki database. Additional information on drug-target associations is provided by STITCH and CTD. From the first annotation to the D4 dopamine receptor (DRD4), the disease term (under Disease Complexes) is highlighted, indicating that protein–protein interaction information for this protein is available. Using the link to the Disease Complexes server, one finds that DRD4 interacts with three proteins (SRC, GRB2 and NCK1). According to OMIM, this protein network is associated to osteogenesis imperfecta and leukemia and, according to BioAlma, to several psychotic disorders. GO enrichment indicates significant association of the protein complex to signal complex formation and vesicle membrane. Furthermore, tissue annotation suggests that this complex is mainly expressed in follicle and non-follicle cells (HPA) and dentritic cells (GNF). Although it might be surprising to see a connection between antidepressant and leukemia, it has been shown recently that antidepressants such as chlomipramine and

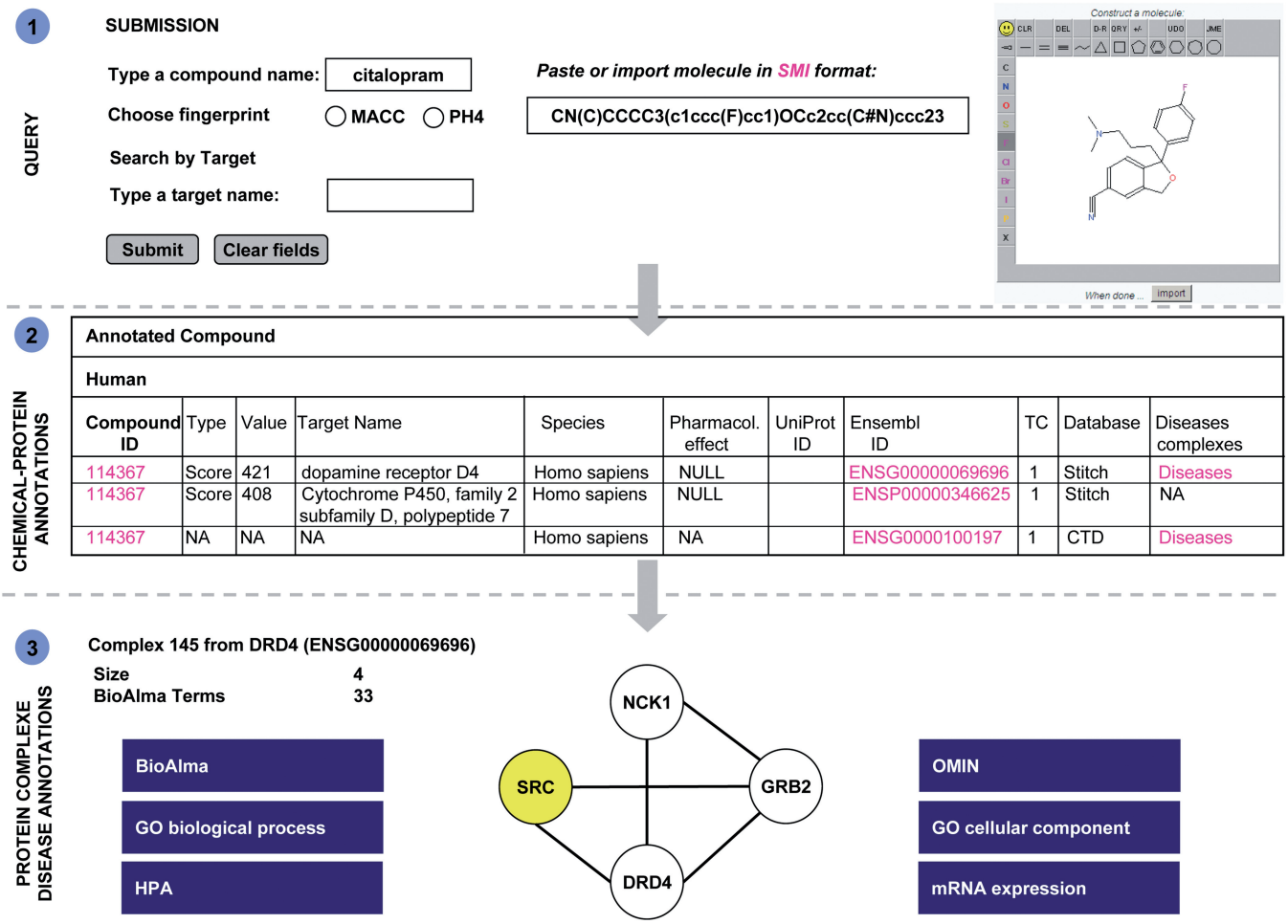

Figure 1. Chemical–protein annotation and disease associations retrieved from ChemProt for the compound citalopram. (**1**) The compound can be queried using different formats (name, SMILES and structure). (**2**) A query results in a table showing protein annotations and bioactivity predictions for the compound. (**3**) Finally, a protein–protein interaction network (protein–complex) for a target protein can be depicted and disease associations (OMIM and BioAlma) and other biological components (GO terms, HPA and mRNA expression) are displayed.

fluoxetine reduce the growth of B-cell malignancies in leukemia (53).

The second query, 'bisphenol A' (BPA), is an environmental pollutant used as plasticizer (54). BPA has biological activity on the estrogen receptor α (ESR1), the androgen receptor (AR) and the estrogen related receptor gamma (ERR3). However, several other proteins are retrieved from CTD and STITCH based on association data with this chemical. Looking at ESR1 in the Disease Complexes server, a complex of 17 proteins is depicted (complex 265) with significant associations to Li-FRAUMENI syndrome, breast cancer and neoplasms. Enrichment analysis indicates that the complex is found in the nucleus (GO cellular component), involved in the regulation of metabolic processes and transcriptionally regulated by the RNA polymerase II promoter (GO biological process). Furthermore, data from immunohistochemistry studies suggest that the complex is mainly located in the endometrium and the cerebral cortex (HPA). The disease chemical biology network for BPA indicates that, under certain conditions, this chemical may be associated with certain types of cancers.

We have illustrated that ChemProt integrates molecular, cellular and phenotypic data associated to small molecules, which can lead to novel links and suggest new avenues for research. We envisage that the ChemProt server will find applications within a variety of chemogenomics, polypharmacology and systems chemical biology studies. ChemProt will be updated once a year with new compounds, new interactions and more sophisticated descriptors.

ACKNOWLEDGEMENTS

Sunset Molecular Discovery LLC (www.sunsetmolecular.com) contributed with the WOMBAT databases.

FUNDING

EU (DEER); Innovative Medicines Initiative Joint Undertaking (eTOX); Danish Research Council for Technology and Production Sciences; Lundbeck

foundation and the Villum Rasmussen Foundation. Funding for open access charge: DEER.

Conflict of interest statement. None declared.

REFERENCES

1. Paolini,G.V., Shapland,R.H., van Hoorn,W.P., Mason,J.S. and Hopkins,A.L. (2006) Global mapping of pharmacological space. *Nat. Biothechnol.*, **24**, 805–815.
2. Keiser,M.J., Setola,V., Irwin,J.J., Laggner,C., Abbas,A.I., Hufeisen,S.J., Jensen,N.H., Kuijer,M.B., Matos,R.C., Tran,T.B. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature*, **462**, 175–181.
3. Mestres,J., Gregori-Puigjané,E., Valverde,S. and Solé,R.V. (2009) The topology of drug-target interaction networks: implicit dependence on drug properties and target families. *Mol. Biosyst.*, **5**, 1051–1057.
4. Antonelli,A., Ferrari,S.M., Fallahi,P., Piaggi,S., Paolicchi,A., Franceschini,S.S., Salvi,M. and Ferrannini,E. (2010) Cytokines (interferon-gamma and tumor necrosis factor-alpha)-induced nuclear factor-kappaB activation and chemokine (C-X-C motif) ligand 10 release in Graves disease and ophthalmopathy are modulated by pioglitazone. *Metabolism*, doi:10.1016/j.metabol .2010.02.002.
5. Vaz,R.J. and Klabunde,T. (2008) Antitargets: Prediction and prevention of drug side effects. In Mannhold,R., Kubinyi,H. and Folkers,G. (eds), *Methods and Principles in Medicinal Chemistry*. Wiley-VCH, Weinheim.
6. Broccatelli,F., Carosati,E., Cruciani,G. and Oprea,T.I. (2010) Transporter-mediated efflux influences CNS side effects: ABCB1, from antitarget to target. *Mol. Inf.*, **29**, 16–26.
7. Olah,M., Rad,R., Ostopovici,L., Bora,A., Hadaruga,N., Hadaruga,D., Moldovan,R., Fulias,A., Mracec,M. and Oprea,T.I. (2007) WOMBAT and WOMBAT-PK: bioactive databases for lead and drug discovery. In Schreiber,S.L., Kapoor,T.M. and Wess,G. (eds), *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*. Wiley-VCH, New York, pp. 760–786.
8. Weill,N. and Rognan,D. (2009) Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G-protein coupled receptors and their ligands. *J. Chem. Inf. Model.*, **49**, 1049–1062.
9. Mestres,J., Martin-Couce,L., Grgori-Puigjané,E., Cases,M. and Boyer,S. (2006) Ligand-based approach to in silico pharmacology: nuclear receptor profiling. *J. Chem. Inf. Model.*, **46**, 2725–2736.
10. Knight,Z.A., Lin,H. and Shokat,K.M. (2010) Targeting the cancer kinome through polypharmacology. *Nat. Rev. Cancer*, **10**, 130–137.
11. Garcia-Serna,R., Ursu,O., Oprea,T.I. and Mestres,J. (2010) iPHACE: integrative navigation in pharmacological space. *Bioinformatics*, **26**, 985–986.
12. Keiser,M.J., Roth,B.L., Armbruster,B.N., Ernsberger,P., Irwin,J.J. and Shoichet,B.K. (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, **25**, 197–206.
13. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccion,M., Edgar,R., Federhen,S. *et al.* (2007) Databases resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.
14. de Matos,P., Alcántara,R., Dekker,A., Ennis,M., Hastings,J., Haug,K., Spiteri,I., Turner,S. and Steinbeck,C. (2010) Chemical entities of biological interest: an update. *Nucleic Acids Res.*, **38**, D249–D254.
15. Berger,S.I. and Iyengar,R. (2009) Network analyses in systems pharmacology. *Bioinformatics*, **25**, 2466–2472.
16. Oprea,T.I., Tropsha,A., Faulon,J.L. and Rintoul,M.D. (2007) Systems chemical biology. *Nat. Chem. Biol.*, **3**, 447–450.
17. Yildirim,M.A., Goh,K.I., Cusick,M.E., Barabási,A.L. and Vidal,M. (2007) Drug-target network. *Nat. Biotechnol.*, **25**, 1119–1126.
18. Chen,B., Wild,D. and Guha,R. (2009) PubChem as a source of polypharmacology. *J. Chem. Inf. Model.*, **49**, 2044–2055.
19. Kuhn,M., Campillos,M., Letunic,I., Jensen,L.J. and Bork,P. (2010) A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.*, **6**, 343.
20. Audouze,K., Juncker,A.S., Roque,F.J., Krysiak-Baltyn,K., Weinhold,N., Taboureau,O., Jensen,T.S. and Brunak,S. (2010) Deciphering diseases and biological targets for environmental chemicals using toxicogenomics networks. *PLoS Comput. Biol.*, **6**, e10000788.
21. Goh,K.I., Cusick,M.E., Valle,D., Childs,B., Vidal,M. and Barabási,A.L. (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.
22. Lage,K., Karlberg,E.O., Størling,Z.M., Olason,O.I., Pedersen,A.G., Rigina,O., Hinsby,A.M., Tümer,Z., Pociot,F., Tommerup,N. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.
23. Liu,T., Lin,Y., Wen,X., Jorissen,R.N. and Gilson,M.K. (2007) Binding DB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.
24. Roth,B., Lopez,E., Beischel,S., Weskaemper,R.B. and Evans,J.M. (2004) Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for CNS drug discovery. *Pharmacol. Ther.*, **102**, 99–110.
25. 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.
26. 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.
27. Davis,A.P., Murphy,C.G., Saraceni-Richards,C.A., Rosentrein,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.
28. Kuhn,M., Szklarczyk,D., Franceschini,A., Campillos,M., von Mering,C., Jensen,L.J., Beyer,A. and Bork,P. (2010) STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res.*, **38**, D552–D556.
29. Durant,J.L., Leland,B.A., Henry,D.R. and Nourse,J.G. (2002) Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.*, **42**, 1273–1280.
30. Bush,B.L. and Sheridan,R.P. (1993) Patty: a programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.*, **33**, 756–762.
31. MOE (version 2007.09), Chemical Computing Group, Montreal, Canada. www.chemcomp.com (29 September 2010, date last accessed).
32. Willet,P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today*, **11**, 1046–1053.
33. Rual,J.F., Venkatesan,K., Hao,T., Dricot,A., Hirozane-Kishikawa,T., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M., Ayivi-Guedehoussou,N. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.
34. Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the biomolecular interaction network database. *Nucleic Acids Res.*, **31**, 248–250.
35. Stark,C., Breitkreutz,B.J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
36. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a molecular interaction database. *FEBS Lett.*, **513**, 135–140.
37. Salwinski,L., Miller,C., Smith,A., Pettit,F., Bowie,J. and Eisenberg,D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
38. Mishra,G., Suresh,M., Kumaran,K., Kannabiran,N., Suresh,S., Bala,P., Shivakumar,K., Anuradha,N., Reddy,R., Raghavan,T.M.

et al. (2006) Human protein reference database – 2006 update. *Nucleic Acids Res*., **34**, D411–D414.
39. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P., Valencia,A. *et al*. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res*., **32**, D452–D455.
40. Pagel,P., Kovac,S., Oesterheld,M., Braumer,B., Dunger-Kaltenbach,I., Frishman,G., Montrone,C., Mark,P., Stumpflen,V., Mewes,H.W. *et al*. (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics*, **21**, 832–834.
41. Guldener,U., Munsterkotter,M., Oesterheld,M., Pagel,P., Ruepp,A., Mewes,H.W. and Stumpflen,V. (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res*., **34**, D436–D441.
42. 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.
43. 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.
44. O'Brien,K.P., Remm,M. and Sonnhammer,E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*., **33**, D476–D480.
45. Lage,K., Hansen,N.T., Karlberg,E.O., Eklund,A.C., Roque,F.S., Donahoe,P.K., Szallasi,Z., Jensen,T.S. and Brunak,S. (2008) A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl Acad. Sci. USA*, **105**, 20870–20875.
46. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*., **33**, D514–D517.
47. Safran,M., Solomon,I., Shmueli,O., Lapidot,M., Shen-Orr,S., Adat,A., Ben-Dor,U., Esterman,N., Rosen,N., Peter,I. *et al*. (2002) GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics*, **18**, 1542–1543.
48. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The gene ontology annotation (GOA) database – sharing knowledge in Uniprot with gene ontology. *Nucleic Acids Res*., **32**, D262–D266.
49. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
50. Ponten,F., Jirström,K. and Uhlen,M. (2008) The human protein atlas – a tool for pathology. *J. Pathol*., **216**, 387–393.
51. Marvin, version5.3. http://www.chemaxon.com/ (29 September 2010, date last accessed).
52. Pafilis,E., O'Donoghue,S.I., Jensen,L.J., Horn,H., Kuhn,M., Brown,N.P. and Schneider,R. (2009) Reflect: augmented browsing for the life scientist. *Nat. Biotechnol*., **27**, 508–510.
53. Chamba,A., Holder,M.J., Jarrett,R.F., Shield,L., Toellner,K.M., Drayson,M.T., Barnes,N.M. and Gordon,J. (2010) SLC6A4 expression and anti-proliferative responses to serotonin transporter ligands fluoxetine in primary B-cell malignancies. *Leuk. Res*., **34**, 1103–1106.
54. Halden,R.U. (2010) Plastics and health risks. *Annu. Rev. Public Health*., **31**, 179–194.