



Database tool

ChemProt-3.0: a global chemical biology diseases mapping

Jens Kringelum^{1,†}, Sonny Kim Kjaerulff^{1,†}, Søren Brunak², Ole Lund¹, Tudor I. Oprea^{1,3} and Olivier Taboureau^{1,4,*}

¹Department of Systems Biology, Center for Biological Sequence Analysis, buildin 208, kemitorvet, Technical University of Denmark, DK-2800 Lyngby, Denmark, ²Department of Disease Systems Biology, Faculty of Health and Medical Sciences, Novo Nordisk Foundation, Center for Protein Research, University of Copenhagen, Blegdamsvej 3A, DK-2200, Copenhagen, Denmark, ³Translational Informatics Division, Department of Internal Medicine, University of New Mexico Health Sciences Center, MSC09 5025, Albuquerque 87181, New Mexico, United States of America, ⁴INSERM, UMRS-973, MTi, Université Paris Diderot, 35 rue Helene Brion, 75205 Paris Cedex 13, Sorbonne Paris Cite, France

*Corresponding author: Tel: +45 4525 2489; Fax: +45 4593 1585; Email: olivier.taboureau@univ-paris-diderot.fr

[†]These authors contributed equally to this work.

Citation details: Kringelum, J., Kjaerulff, S.K., Brunak, S., *et al.* ChemProt-3.0: a global chemical biology diseases mapping. *Database* (2016) Vol. 2016: article ID bav123; doi:10.1093/database/bav123

Received 21 April 2015; Revised 7 December 2015; Accepted 8 December 2015

Abstract

ChemProt is a publicly available compilation of chemical-protein-disease annotation resources that enables the study of systems pharmacology for a small molecule across multiple layers of complexity from molecular to clinical levels. In this third version, ChemProt has been updated to more than 1.7 million compounds with 7.8 million bioactivity measurements for 19 504 proteins. Here, we report the implementation of global pharmacological heatmap, supporting a user-friendly navigation of chemogenomics space. This facilitates the visualization and selection of chemicals that share similar structural properties. In addition, the user has the possibility to search by compound, target, pathway, disease and clinical effect. Genetic variations associated to target proteins were integrated, making it possible to plan pharmacogenetic studies and to suggest human response variability to drug. Finally, Quantitative Structure–Activity Relationship models for 850 proteins having sufficient data were implemented, enabling secondary pharmacological profiling predictions from molecular structure.

Database URL: <http://potentia.cbs.dtu.dk/ChemProt/>

Introduction

Many chemical biology initiatives in Europe and the USA aim to screen large compound collections with dedicated

bioassays i.e. EU Lead Factory (1), EU-Openscreen (2) or BARD in the USA (3). Such large initiatives generate large amounts of data that support academic and industrial

research in the discovery of safer chemicals, with better efficacy. To make chemical biology information accessible to scientists, several repositories of bioactive small molecules have been developed: ChEMBL (4), PubChem (5), ChemSpider (6) and OpenPhacts (7) are the largest, more general databases available to the public. The National Institutes of Health's Molecular Libraries Program (MLP) funding developed the BioAssay Research Database (BARD), focusing on assay ontologies for PubChem bioassays (3).

Advances in chemical biology and systems biology have shown that most drugs interact with multiple targets and that the pharmacological profile of a drug is not as reductionist as once believed (8). Moreover, proteins rarely operate in isolation within and outside cells but function in interconnected pathways instead. Given the integration afforded by systems biology, it is now possible to consider a more general physiological environment for protein targets and biological processes. As massive amounts of data are generated and accumulated via new experimental technologies such as transcriptomic, proteomics and genomics (through next-generation sequencing), drug action can be explored across multiple scale of complexity, from molecular and cellular to tissue and organism levels (9–11).

Multi-target pharmacology exploration increases when information linking the relationship between chemical and target spaces is readily available. As archived data are processed and homogenized, our total knowledge on protein–ligand interactions is increasing at an amazing pace (12, 13). Scientists having access to these data, approaches such as chemogenomics, proteochemometrics and polypharmacology have started to emerge (14, 15). These help to mine evaluate and ultimately distil this vast amount of protein–ligand interactions data, enabling the predictions of single ligands against a set of heterogeneous targets (16).

This third version of ChemProt is not a simple update for disease chemical biology data. Rather, we provide a friendly platform to navigate through the various data sources, from global evaluations to a focused analysis. Several computational approaches are included: ligand-based similarity, target-based promiscuity, QSAR (Quantitative Structure–Activity Relationship) methodology and network biology-based enrichment analyses. These approaches support novel hypotheses generation for bioactivity of novel and already-annotated compounds, and the ability to identify additional genes that may play major roles in modulating chemical perturbations in man. The updates and new methods introduced in ChemProt-3.0 are presented below.

Data sources

We updated all the chemical protein interactions data from the open source databases ChEMBL (version 19) (4), BindingDB (17), PDSP Ki database (18), DrugBank (version 4) (19), PharmGKB (20), IUPHAR-DB database (21) and STITCH (version 4) (22). Clinical information from the Anatomical Therapeutic Chemical Classification System (23) developed by the World Health Organization, as well as side effect data from Sider 2 were also integrated (24).

From a biological perspective, we updated our internal human interactome platform to reach 14 421 genes interacting through 507 142 unique PPIs (25). OMIM (26), the human disease network (27) GeneCards (28), KEGG (29), Reactome (30), UniPathway (31) and Gene Ontology (32) databases were also downloaded, curated and included in our system. Overall, the integrated data sources were increased by over 60% compared to the earlier version.

As many different data types were aggregated in ChemProt, a 'zChemProt' value for each compound-bioactivity interactions was computed for visualization in the several heatmaps developed. Basically, for each of the 11 most prevalent data types (IC_{50} , EC_{50} , Potency, AC_{50} , pIC_{50} , $\log K_i$, pK_i , pEC_{50} , K_d , K_i), a zChemProt value was computed using the mean and standard deviation calculated from the distribution of the associated data types for each target in a similar way described in CARLSBAD database (33). IC_{50} , EC_{50} , Potency, AC_{50} , K_d and K_i were log transformed before computing the zChemProt values. Large values indicate strong chemical–protein interactions and are represented in orange. Low value (weak interactions) is depicted in blue.

Predictions methods

Daylight-like 1024 bit fingerprints was computed with RDkit (www.rdkit.org) and the chemical similarity between two compounds was quantitatively assessed using the Tanimoto coefficient (34). The Similarity Ensemble Approach (SEA) (35) has been re-compiled on the ChemProt server and updated according to the novel zChemProt data and integrated into ChemProt 3.0. Only proteins with >10 chemicals were included for SEA prediction, using the same protocol as described in the previous version (36). For sequence analyses, protein sequences were obtained from Uniprot (37). Sequences comparisons were computed using BLASTP and estimated to be similar when their E value was lower than 10^{-10} (38). All compounds were decomposed into ring scaffolds based on an internal implementation of the 'Scaffold Hunter'

hierarchical classification algorithm (39–40) with the addition of decomposition of non-ring molecule based on rules 7–10, as described by Schuffenhauer *et al.* (41). This hierarchical decomposition allows the generation of scaffold trees enabling an easy and interactive navigation of the chemical biology space in large datasets and the identification of potential new compound classes with desired bioactivity.

For this release, QSAR models were trained for each protein with >20 chemicals (in total 850 proteins). A Naive Bayes classifier was trained using 5-fold cross-validation for performance assessment. Features selection, five different computational fingerprints (Daylight and Morgan fingerprints) and three different cutoffs ($-\log_{10}$ value: 4, 5 and 6) for classifying active and non-active compounds, were used to produce overall 15 classification models for each target. To predict new compound, each model was weighted by the cross-validated performance measure resulting in a prediction value between 0 and 1 (where 1 is high predicted binding). To avoid bias toward negative or positive data, each of the three datasets used for training were balanced by including as many negative compounds as positive including random compounds from the ChemProt database. The performance of the developed QSAR ensemble approach was tested on a dataset of hERG binders and showed an improved performance compared to a previous reported study (42) (Aroc = 0.827, Matthews Correlation Coefficient (MCC) = 0.488 using 5-fold cross-validation). Furthermore, the method was benchmarked against the SEA implementation on a dataset consisting of 143 proteins with associated activity values from the ChemProt-3.0 dataset. In a 5-fold cross-validation scheme on each of the 143 proteins, the QSAR models outperform SEA ($P_{val} = 2.2e^{-16}$, paired *t*-test of Spearman correlation coefficients for each protein predictions). However, models developed from limited amounts of data might not provide reliable predictions. The user can consult the ‘prediction info’ tab (by clicking on the protein of interest on the heatmap) to obtain the number of molecules used for training the QSAR models. Details about the procedure are described in [Supplementary Information](#).

Visual interface

In ChemProt 3.0, the front page was modified to have all the functionality available on the page. The user has the possibility to search information about a compound, a protein and a clinical outcome, or he can choose to perform a QSAR prediction for a specific compound. A molecule can be imported as a SMILES code, or alternatively it can be drawn or uploaded from a compound structure file via the

SD file format. A new function called ‘Heatmap’ was integrated, which allows the user to have a global view of chemical-protein interactions (Figure 1). In this graphical interface, the user has the possibility to localize the bioactivity associated to a requested compound, a set of defined compounds or a set of similar compounds based on the chemical structure. Several layers of granularity have been implemented on the heatmap. The proteins have been categorized by families, using the protein classification tree implemented in ChEMBL and the compounds have been decomposed in scaffold and chemical groups based on the scaffold tree implementation similar to CARLSBAD. This gives the user the opportunity to visualize scaffold-protein activity relationships. A color spectrum from blue (low activity) to orange (strong activity) is used to indicate the activity. All compounds structures and protein IDs (based on Uniprot) are clickable, which gives access to more detailed information about physicochemical properties and protein function respectively.

An interesting feature with this graphical interface is the possibility to match other biological to chemical data. Instead of choosing ‘for drugs’, the user can select targets, pathways, diseases or side effects and see the association between chemicals and these endpoints.

From the protein ID, the user has access to the proteins complex (represented as a protein’s network from protein-protein interaction data). The complex of proteins is then mapped to biological terms such as diseases, GO terms and pathways with a corrected *P* value to evaluate the significance of these associations.

Finally, the user has the possibility to download the results in flat-file format to perform others analyses.

Methodology: Daylight-like fingerprints, defined by 1024 fragments were computed using RDKit (www.rdkit.org). The QSAR models were trained using scikit-learn software (<http://scikit-learn.org/stable/>). The visual interface was implemented using HTML 5 and JavaScript. The webserver is limited for an input file of 50 molecules (in SMILES or sdf file) per query to limit the time necessary to get the output. For larger queries, the user is advised to contact us.

Applications

Caffeine, a well-known natural product extracted from coffee beans or tea leaves, is often used as a central nervous system stimulant (43). Several outputs can be displayed like those shown in Figure 2. Typing the compound name in the ‘Compound’ field and clicking on the Submit button, the user is redirected to the global chemical-protein heatmap with the query compound showing up in the compound list as default. Any compound can be added to the

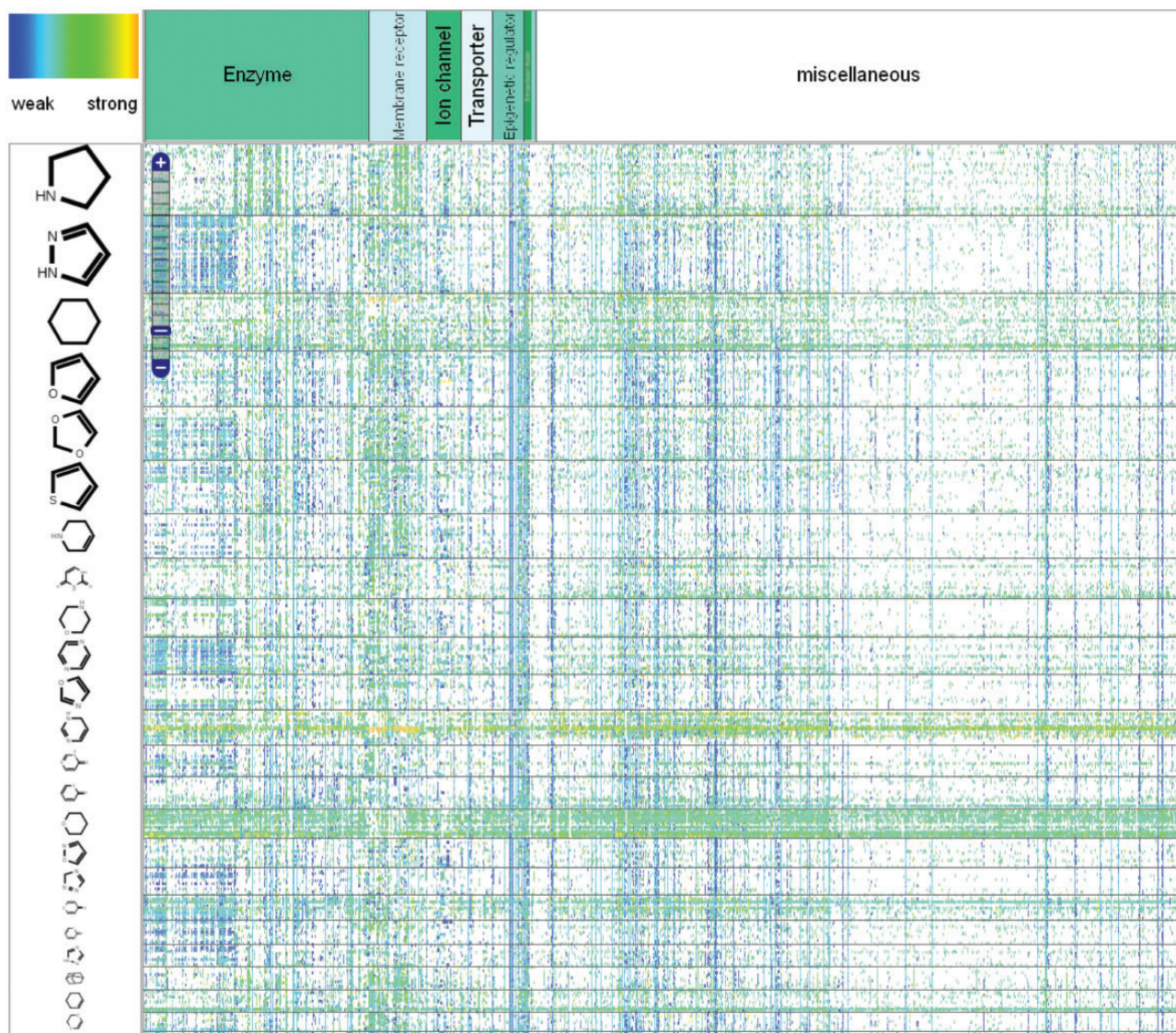




Figure 1. Global view of the chemical-protein interactions heatmap in ChemProt.

list by writing a new compound name in the ‘search’ box. Clicking on the ‘flag’  next to a compound name in the compound list prompts the heatmap to zoom in or out from that specific compound, enabling a fast way to visualize the proteins signature for the queried compound, as well as for compounds sharing scaffold similarities. Clicking on the ‘fingerprint’ logo  in the vicinity of the compound name, a chemical structure similarity profiling can be performed, enabling the user to visualize and to navigate within that pharmacological heatmap. For the Caffeine example, 105 similar compounds (with a Tanimoto coefficient > 0.85) were found, with bioactivities associated to 449 proteins (from weak in blue to strong in orange). The user is able to zoom in the heatmap and to narrow the information from the classification proteins tree to specific proteins (defined by uniprot ID). The user has also the possibility to navigate inside the heatmap. One

option is to fill out missing values by choosing ‘SEA’ or ‘QSAR’ under prediction in the top of the page.

By clicking on the compound structure, physicochemical features [such as the Lipinski rules (44)], number of proteins with bioactivities and the databases from which the information was gathered, are shown. Similarly, the user can click on a specific protein name and get more information on the function of the protein, diseases associated to this protein and predictions based on SEA and QSAR. For example, under ‘family A GPCR’, caffeine is shown to be potent (35.5 nM) on the rat muscarinic M1 acetylcholine receptor (P08482). It also shows a strong association with the dopamine D2 receptor (P14416) based on the STITCH system. By clicking on this protein, the user is presented with information on this dopamine receptor. Notably, disease associations are queried through the TCRD (Target Central Resource Database Application:

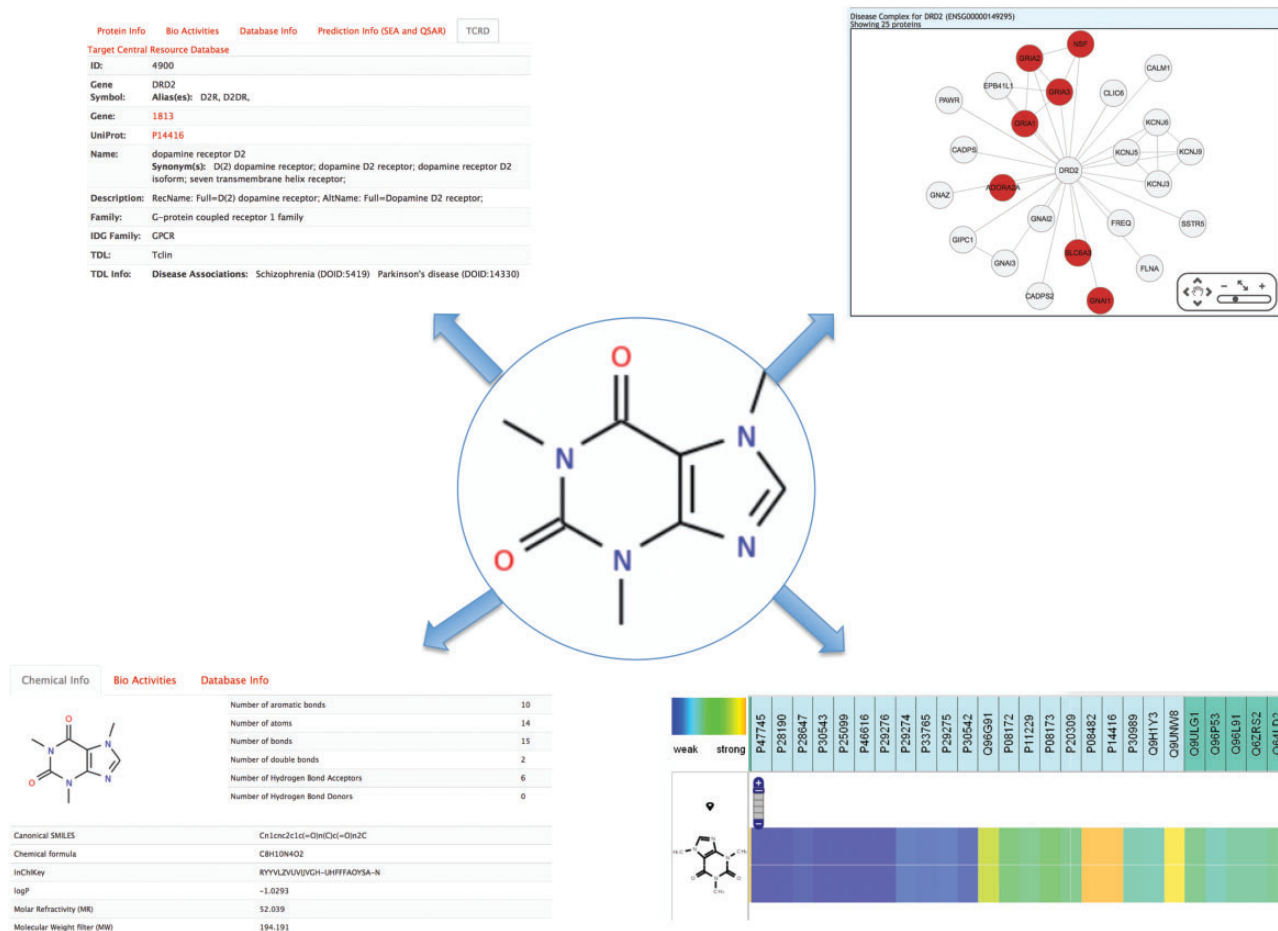


Figure 2. Information that can be collected from a search on caffeine. Top left, functional information on bioactive proteins for the query compound is depicted. Bottom left, chemical and physicochemical information is gathered. Top right, protein's complex associated to the chemical is shown and the bottom right is depicted the protein's annotation and prediction (through QSAR) for caffeine.

<http://juniper.health.unm.edu/tcrd/>) database and the genetic variation through the Ensembl database (45). A complex disease network is also associated to this protein. Clicking on this link, diseases (such as schizophrenia and epilepsy) and GO terms (plasma membrane, ligand-gated ion channel activity, etc.) are shown.

Instead of looking for 'functional' protein annotation, it is further possible to select 'pathway' or 'diseases' for caffeine (also for the set of 105 similar compounds). The heatmap will be depicted according to the query. Each protein annotation (functional, pathway, disease) is presented in a tree format. Proteins have been categorized from families to proteins using the protein target tree implemented in ChEMBL. It has been done similarly for the pathway using the unipathway (31) implementation tree and disease using the human disease network (27). For example, using the disease heatmap, strong associations are found between caffeine and ventricular tachycardia, slow acetylation, glycogen storage disease, dystonia and thyroid carcinoma.

Finally, from the ChemProt-3 front page, the user can write the caffeine's SMILES in the QSAR prediction box, click on 'Submit QSAR' and then get a prediction of positive and negative bioactivities for the ensemble of proteins in ChemProt where reliable QSAR models can be produced. This option allows the user to have a direct QSAR prediction for a new compound not present in the ChemProt database.

Conclusion

Given that access to many chemogenomics databases is possible, linking them to biological resources and using a number of machine learning tools, scientists can now estimate the bioactivity profile of molecules across a large number of targets, pathways, diseases and other clinical outcomes using ligand-based, target-based and network-based models. Such multi-target, multi-layer strategies are becoming more and more accepted by the scientific community. Within ChemProt, it is possible to

navigate the chemogenomics space and to link chemically induced target perturbations to diseases and other biological outcomes. Such tools might be of interest for drug discovery, drug safety and also chemical risk assessment. ChemProt 3.0 supports predicting bioactivities on targets and off-targets for new compounds and can assist in the associations to phenotypes and side effects relationships.

Supplementary Data

Supplementary data are available at *Database* Online.

Conflict of interest: None declared.

Funding

This work was supported by the Innovative Medicines Initiative Joint Undertaking, under grant agreement No 115002 (eTOX) for O. Taboureau and No 115191 OPENPHACTS (S.K., J.K., T.I.O.) the resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies in kind contribution.

References

- Mullard, A. (2013) European lead factory opens for business. *Nat. Rev. Drug. Discov.*, **12**, 173–175.
- Frank, R. (2014) EU-Openscreen—a European infrastructure of open screening platforms for chemical biology. *ACS Chem. Biol.*, **9**, 853–854.
- de Souza, A., Bittker, J.A., Lahr, D.L. *et al.* (2014) An overview of the challenges in designing, integrating and delivering BARD: a public chemical-biology resource and query portal for multiple organizations, locations and disciplines. *J. Biomol. Screen.*, **19**, 614–627.
- Gaulton, A., Bellis, L.J., Bento, A.P. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.
- Sayers, E.W., Barrett, T., Benson, D.A. *et al.* (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **40**, D13–D25.
- Williams, A. and Tkachenko, V. (2014) The Royal Society of Chemistry and the delivery of chemistry data repositories for the community. *J. Comput. Aided Mol. Des.*, **28**, 1023–1030.
- William, A.J., Harland, L., Groth, P. *et al.* (2012) Open PHACTS: semantic interoperability for drug discovery. *Drug Discov. Today*, **17**, 1188–1198.
- Wild, D.J., Ding, Y., Sheth, A.P. *et al.* (2012) Systems chemical biology and the semantic web: what they mean for the future of drug discovery research. *Drug Discov. Today*, **17**, 469–474.
- Keith, C.T., Borisy, A.A. and Stockwell, B.R. (2005) Multicomponent therapeutics for networked systems. *Nat. Rev. Drug Discov.*, **4**, 71–78.
- Zimmerman, G.R., Lehár, J. and Keith, C.T. (2007) Multi-target therapeutics: when the whole is greater than the sum of the parts. *Drug Discov. Today*, **12**, 34–42.
- Fitzgerald, J.B., Schoeberl, B., Nielsen, U.B. *et al.* (2006) Systems biology and combination therapy in the quest for clinical efficacy. *Nat. Chem. Biol.*, **2**, 458–466.
- Ekins, S. and Williams, A.J. (2010) When pharmaceutical companies publish large datasets: an abundance of riches or fool's gold? *Drug Discov. Today*, **15**, 812–815.
- Wang, J., Li, Z., Qiu, C. *et al.* (2012) The relationship between rational drug design and drug side effects. *Brief. Bioinform.*, **13**, 377–382.
- Rognan, D. (2007) Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.*, **152**, 38–52.
- Keiser, M.J., Setola, V., Irwin, J.J. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature*, **462**, 175–181.
- Schneider, G. (2010) Virtual screening: an endless staircase? *Nat. Rev. Drug Discov.*, **9**, 273–276.
- Liu, T., Lin, Y., Wen, X. *et al.* (2007) Binding DB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.
- Roth, B., Lopez, E., Beischel, S. *et al.* (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.
- Law, V., Knox, C., Djoumbou, Y. *et al.* (2011) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.
- McDonagh, E.M., Whirl-Carrillo, M., Garten, Y. *et al.* (2011) From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomark. Med.*, **5**, 795–806.
- Pawson, A.J., Sharman, J.L., Benson, J.L. *et al.* (2014) The IUPHAR/BPS guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. *Nucleic Acids Res.*, **42**, D1098–D1106.
- Kuhn, M., Szklarczyk, D., Pletscher-Frankild, S. *et al.* (2014) STITCH 4: integration of protein-chemical interactions with user data. *Nucleic Acids Res.*, **42**, D401–D407.
- De Smet, P.A.G.M. (1993) New applications of the ATC/DDD methodology in the Netherlands part 1. ATC/DDD principles and computerized medication surveillance. *Int. Pharm. J.*, **7**, 196–199.
- Kuhn, M., Campillos, M., Letunic, I. *et al.* (2010) A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.*, **6**, 343.
- Lage, K., Karlberg, E.O., Størling, Z.M. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.
- Amberger, J., Bocchini, C. and Hamosh, A. (2011) A new face and new challenges for Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.*, **32**, 564–567.
- Goh, K.I., Cusick, M.E., Valle, D. *et al.* (2007) The human disease network. *Proc. Natl. Acad. Sci. USA*, **104**, 8685–8690.
- Stelzer, G., Dalah, I., Stein, T.I. *et al.* (2011) In-silico human genomics with GeneCards. *Hum. Genomics*, **5**, 709–717.
- Kanehisa, M., Goto, S., Sato, Y. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
- Matthews, L., Gopinath, G., Gillespie, M. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.

31. Morgat, A., Coissac, E., Coudert, E. *et al.* (2012) Unipathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res.*, **40**, D761–D769.
32. Dimmer, E.C., Huntley, R.P., Alam-Faruque, Y. *et al.* (2012) The UniProt-GO Annotation database in 2011. *Nucleic Acids Res.*, **40**, D565–D570.
33. Mathias, S.L., Hines-Kay, J., Yang, J.J. *et al.* (2013) The CARLSBAD database: a confederated database of chemical bioactivities. *Database*, 2013, bat044.
34. Willet, P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today*, **11**, 1046–1053.
35. Keiser, M.J., Roth, B.L., Armbruster, B.N. *et al.* (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, **25**, 197–206.
36. Kjærulff, K.S., Wich, L., Kringelum, J. *et al.* (2013) ChemProt-2.0: visual navigation in a disease chemical biology database. *Nucleic Acids Res.*, **41**, D464–D469.
37. UniProt Consortium. (2011) Ongoing and future developments at the universal protein resource. *Nucleic Acids Res.*, **39**, D214–D219.
38. Altschul, S.F., Madden, T.L., Schäffer, A.A. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
39. Klein, K., Koch, O., Kriege, N. *et al.* (2013) Visual analysis of biological activity data with Scaffold Hunter. *Mol. Inf.*, **32**, 964–975.
40. Wetzel, S., Klein, K., Renner, S. *et al.* (2009) Interactive exploration of chemical space with Scaffold Hunter. *Nat. Chem. Biol.*, **5**, 581–583.
41. Schuffenhauer, A., Ertl, P., Roggo, S. *et al.* (2007) The scaffold tree—visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.*, **47**, 47–58.
42. Li, Q., Jorgensen, F.S., Oprea, T. *et al.* (2008) hERG classification model based on a combination of support vector machine method and GRIND descriptors. *Mol. Pharm.*, **5**, 117–127.
43. Nehlig, A., Daval, J.L. and Debry, G. (1992) Caffeine and the central nervous system: mechanisms of action, biochemical, metabolic and psychostimulant effects. *Brain Res. Brain Res. Rev.*, **17**, 139–170.
44. Lipinski, C.A., Lombardo, F., Dominy, B.W. *et al.* (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, **46**, 3–26.
45. Cunningham, F., Amode, M.R., Barrell, D. *et al.* (2015) Ensembl 2015. *Nucleic Acids Res.*, **42**(Database issue):D749–D755.