Published online 16 October 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D919–D922***
doi:10.1093/nar/gkm862

SuperTarget and Matador: resources for exploring drug-target relationships

Stefan Günther[1], Michael Kuhn[2], Mathias Dunkel[1], Monica Campillos[2], Christian Senger[1], Evangelia Petsalaki[2], Jessica Ahmed[1], Eduardo Garcia Urdiales[2], Andreas Gewiess[3], Lars Juhl Jensen[2], Reinhard Schneider[2], Roman Skoblo[3], Robert B. Russell[2], Philip E. Bourne[4], Peer Bork[2,5] and Robert Preissner[1,*]

[1]Structural Bioinformatics Group, Institute of Molecular Biology and Bioinformatics, Charité—University Medicine Berlin, Arnimallee 22, 14195 Berlin, [2]EMBL—Biocomputing, Meyerhofstraße 1, 69117 Heidelberg, [3]Institute for Laboratory Medicine, Windscheidstr, 18, 10627 Berlin, Germany, [4]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, 9500 Gilman Drive, La Jolla CA 92093, USA and [5]Max-Delbrück-Center for MolecularMedicine (MDC), 13092 Berlin-Buch, Germany

Received August 15, 2007; Revised September 26, 2007; Accepted September 27, 2007

ABSTRACT

The molecular basis of drug action is often not well understood. This is partly because the very abundant and diverse information generated in the past decades on drugs is hidden in millions of medical articles or textbooks. Therefore, we developed a one-stop data warehouse, *SuperTarget* that integrates drug-related information about medical indication areas, adverse drug effects, drug metabolization, pathways and Gene Ontology terms of the target proteins. An easy-to-use query interface enables the user to pose complex queries, for example to find drugs that target a certain pathway, interacting drugs that are metabolized by the same cytochrome P450 or drugs that target the same protein but are metabolized by different enzymes. Furthermore, we provide tools for 2D drug screening and sequence comparison of the targets. The database contains more than 2500 target proteins, which are annotated with about 7300 relations to 1500 drugs; the vast majority of entries have pointers to the respective literature source. A subset of these drugs has been annotated with additional binding information and indirect interactions and is available as a separate resource called *Matador*. *SuperTarget* and *Matador* are available at http://insilico.charite.de/supertarget and http://matador.embl.de

INTRODUCTION

Within the past two decades our knowledge about drugs, their mechanisms of action and target proteins has increased rapidly. Nevertheless, knowledge on their molecular effects is far from complete. For some drugs even the primary targets are still unknown, for example, Diloxanide, Niclosamide and Ambroxol are administered successfully although their effect on human metabolism is still not clarified at a molecular level (1). Even if the medical effect has been explained by a certain molecular interaction, most drugs interact with several additional targets, which may either strengthen the therapeutic effect or cause unwanted adverse drug effects (2). Moreover, our knowledge on drugs and their targets is highly fragmented, most of it residing in millions of medical articles and textbooks, which precludes systematic studies.

Several databases exist that collect binding data on small molecules, in particular drugs and proteins. The largest such resource is DrugBank (3), which contains 2600 drug-target relations for 900 FDA-approved drugs and additional annotations for 3200 experimental drugs. Another notable database is the Therapeutic Target Database (TTD) (4), which holds target information on about 1000 small molecule drugs. Unfortunately, DrugBank only provides references on the target, although generally not on the interactions, which makes it difficult to obtain information on the experimental context under which an interaction was observed. Moreover, the drugs in the TTD are not cross-linked

*To whom correspondence should be addressed. Tel: +49 30 8445 1649; Fax: +49 30 8445 1551; Email: robert.preissner@charite.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.

with compound databases such as PubChem, ChemDB or the commercial CAS® Registry, and the targets are not linked to protein databases such as UniProt or PDB. This makes it difficult to retrieve information such as the chemical structure of the drug, its physiochemical properties, the sequence or 3D structure of its target or the biological pathways that it affects.

In order to be able to derive further information about drug-target relations, we have developed a one-stop data warehouse, *SuperTarget* that provides this functionality and integrates drug-target relations from different resources using heterogeneous retrieval methods. We consider a drug-target relation as a specific interaction of a small chemical compound administered to treat or diagnose a disease and a macromolecule, namely protein, DNA or RNA. The first release of *SuperTarget* contains a core dataset of about 7300 drug-target relations of which 4900 interactions have been subjected to a more extensive manual annotation effort to incorporate additional binding information as well as indirect interactions. The resulting data on 775 drugs is provided separately as *Matador* (Manually Annotated Targets And Drugs Online Resource).

DRUG-TARGET RELATIONSHIPS

Drug-target relationships described in *SuperTarget* were obtained in three different ways. Starting with 2400 drugs and their synonyms from the SuperDrug Database (5), the text mining tool EbiMed (6) was used to extract relevant text passages containing potential drug-target relations from about 15 millions public abstracts listed in PubMed. Many thousands of false positive or irrelevant relations were eliminated by manual curation.

In parallel, potential drug-target relations were automatically extracted from Medline by searching for synonyms of drugs, proteins and Medical Subject Headings (MeSH terms) describing groups of proteins (7). MeSH terms were used to capture and down-weight interactions that are not explicitly described in the abstracts e.g. for protein families or protein complexes. In the case of families, the specific interacting family member might not be known yet, whereas in the case of complexes, the drug might interact with more than one subunit. Proteins associated to MeSH terms were assigned by a semi-automated procedure relying on mappings provided by MeSH and synonyms of proteins that are aggregated in the STRING resource (8). Proteins that were often mentioned in abstracts, but could not be automatically assigned to families, were manually assigned. Depending on the size and nature of the families, the confidence of an interaction between drugs and individual proteins was decreased. More heterogeneous families are assigned a lower confidence. The most probable candidates were identified using a benchmarking scheme (8) and manually curated.

In a last step, relations from other databases, namely DrugBank (3), KEGG (9), PDB (10), SuperLigands (11) and TTD (4), were checked for drug-target interactions not identified with the preceding steps.

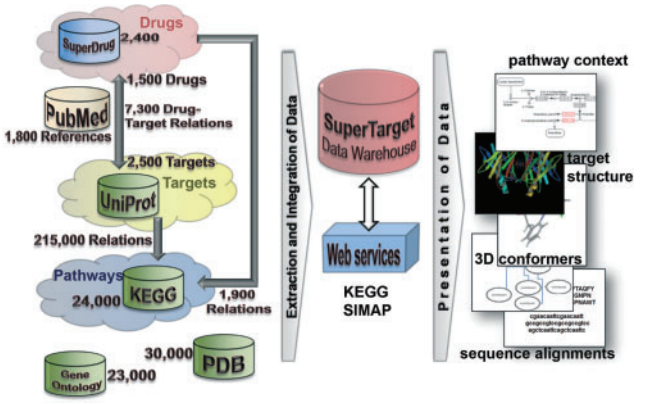

Figure 1. System architecture and number of database entries of *SuperTarget*. The database contains the complete Uniprot with more than 3 million entries. Beside the targets, drugs and pathways the database provides 23 000 different GO-terms and 30 000 links to protein structures.

If those interactions could be confirmed by literature listed in PubMed, the references were included in *SuperTarget* otherwise the describing database is referenced.

In consideration of the large number of entries we cannot rule out that some of the data is erroneous, change over time or is too unspecific. In the case of doubt we refer to the referenced relation source.

To be able to obtain more information on the drug-target relations, *SuperTarget* provides links to physicochemical properties and further structural information of drugs. Proven or potential target proteins are represented by sequences as stored in UniProt (12), by functional annotations extracted from GOA (13) and by related pathway information provided by KEGG (9) (compare Figure 1). Adverse drug reactions were extracted from the free accessible Canadian Adverse Reaction Monitoring Program (CADRMP, http://www.hc-sc.gc.ca/).

For a subset of the drug-target relations, namely those where our text-mining approach indicated a wealth of additional information, the type of binding was further analyzed and direct and indirect interactions were manually distinguished. Indirect interactions can, for example, be caused by active metabolites of the drug or by changes in the expression of a protein. The extensively annotated subset, which is contained in *Matador* should be well-suited as training set for various large-scale discovery approaches.

ANALYSIS OPTIONS IN *SUPERTARGET*

The integration of drug related information provides numerous different query entry points as well as global surveys on complex topics using heterogeneous data types (for an illustration of query capabilities see Figure 2).

Since compounds with high structural similarity frequently exhibit similar activities (14), all drugs with comparable structural fingerprints are made accessible.

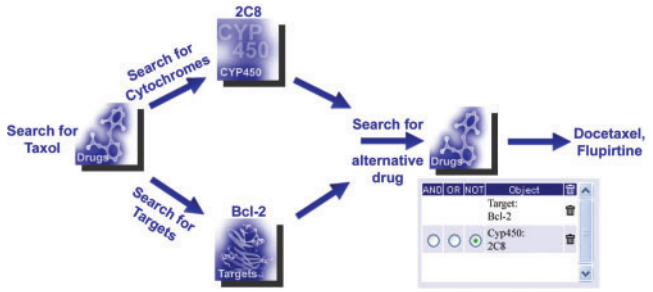

Figure 2. Example of a complex query: search for an alternative drug to Taxol, which adresses the same target Bcl-2, but is not metabolized by the cytochrome 2C8. Procedure: start to search for the targets and the metabolizing cytochromes of Taxol in the associated categories. The resulting lists contain among others the target Bcl-2 as well as the P450 cytochrome 2C8. Second, forward 2C8 and Bcl-2 into the query term box, combine them using the 'NOT' operator and submit the query. The resulting drug list contains two alternative drug candidates for Taxol (Docetaxel and Flupirtine).

Fingerprints allow a fast identification of drugs with very similar physiochemical properties that may interact with the same target molecules. Structural similarity scores and fingerprints are computed with the open-source Chemistry Development Kit (15).

Analogously, similar proteins are quickly identifiable by precomputed sequence alignments provided by SIMAP (16). Proteins homologous to annotated target proteins are candidates for interacting with the drug and may explain adverse effects of drugs.

Drug metabolizing enzymes come into question to explain adverse drug responses. Genetic polymorphism of cytochrome P450 genes or associated regulatory factors may lead to diverse ability of drug degradation (17). The web interface provides an extra section to retrieve the cytochrome interaction data for most of the annotated drugs.

Drugs are classified in the Anatomical Therapeutic Chemical Classification System (ATC). This hierarchical classification system was introduced by the World Health Organization in 1990, classifying drugs according to their medical indications and chemical scaffold. Groups of drugs can be selected by their ATC code. A searchable, hierarchical ATC-tree enables a fast detection of drug classes and medical indications like, for example, 'anti-Parkinson drugs' by expanding the branch 'nervous system'.

The integration of Gene Ontology (GO) terms in *SuperTarget* enables complex queries like 'Are there anti-cancer drugs that induce apoptosis and are associated with transcription factors?' which can be answered by combined selection of apoptosis in the pathway tab and transcriptional activator activity in the ontology tree.

To facilitate the analysis, targets or drugs can be collected and stored in a clipboard. A session can be saved on the server and restored up to 30 days later. Each object of the clipboard links to a window with object-related information. Depending of the object type, the window contains additional information relating to drugs, targets, pathways, GO terms or metabolization. A further hyperlink leads to the search history.

CONCLUSIONS AD FUTURE DIRECTIONS

Although the first version of *SuperTarget* with all the search and discovery tools around drug-target relations is already an extensive resource for both large-scale research and in depth analysis, the captured knowledge is still far from complete and we would like to invite the community to help in increasing quality and quantity of the records. *SuperTarget* offers an option to upload and incorporate drug-target relations into a working queue. Uploaded entries will be reviewed and approved by an annotation team comprised of graduated scientists. Both *SuperTarget* and *Matador* can be used as knowledge sources, discovery tools or training sets for various applications in chemical biology and elsewhere.

AVAILABILITY

SuperTarget and *Matador* are available via their web sites, http://insilico.charite.de/supertarget and http://matador.embl.de. They can be obtained via a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License.

ACKNOWLEDGEMENTS

The authors wish to thank B. Grüning, P. Pyl, D. Jansen, J. Tschörtner, M. Füllbeck, E. Michalsky, J. Hossbach and I. Jaeger for assistance during literature screening, software testing and improvement of the web interface. Furthermore, we thank M. Kanehisa for providing the KEGG web service as well as R. Arnold for the support using the SIMAP web service. This work was supported by BMBF (Quantpro), Deutsche Forschungsgemeinschaft (DFG: SFB 449), IRTG Berlin-Boston-Kyoto, Investitionsbank Berlin (IBB) and Deutsche Krebshilfe. Funding to pay the Open Access publication charges for this article was provided by EMBL.

Conflict of interest statement. None declared.

REFERENCES

1. Imming,P., Sinning,C. and Meyer,A. (2006) Drugs, their targets and the nature and number of drug targets. *Nat. Rev. Drug Discov.*, **5**, 821–834.
2. Frantz,S. (2005) Drug discovery: playing dirty. *Nature*, **437**, 942–943.
3. 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.
4. Chen,X., Ji,Z.L. and Chen,Y.Z. (2002) TTD: Therapeutic Target Database. *Nucleic Acids Res.*, **30**, 412–415.
5. Goede,A., Dunkel,M., Mester,N., Frommel,C. and Preissner,R. (2005) SuperDrug: a conformational drug database. *Bioinformatics*, **21**, 1751–1753.
6. Rebholz-Schuhmann,D., Kirsch,H., Arregui,M., Gaudan,S., Riethoven,M. and Stoehr,P. (2007) EBIMed–text crunching to gather facts for proteins from Medline. *Bioinformatics*, **23**, e237–e244.

7. Jensen,L.J., Saric,J. and Bork,P. (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.*, **7**, 119–129.
8. 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.
9. 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.
10. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.
11. Michalsky,E., Dunkel,M., Goede,A. and Preissner,R. (2005) SuperLigands—a database of ligand structures derived from the Protein Data Bank. *BMC Bioinform.*, **6**, 122.
12. niProt-Consortium1 (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.
13. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. *et al.* (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.
14. Martin,Y.C., Kofron,J.L. and Traphagen,L.M. (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, **45**, 4350–4358.
15. 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**, 2111–2120.
16. Rattei,T., Arnold,R., Tischler,P., Lindner,D., Stumpflen,V. and Mewes,H.W. (2006) SIMAP: the similarity matrix of proteins. *Nucleic Acids Res.*, **34**, D252–D256.
17. Fujita,K. (2006) Cytochrome P450 and anticancer drugs. *Curr. Drug Metab.*, **7**, 23–37.