Published online 22 May 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W55–W59***
doi:10.1093/nar/gkn307

SuperPred: drug classification and target prediction

Mathias Dunkel[1], Stefan Günther[1], Jessica Ahmed[1], Burghardt Wittig[1] and Robert Preissner[1,*]

[1]Institute of Molecular Biology and Bioinformatics, Charité – University Medicine Berlin, Arnimallee 22, 14195 Berlin, Germany

Received February 18, 2008; Revised April 8, 2008; Accepted April 30, 2008

ABSTRACT

The drug classification scheme of the World Health Organization (WHO) [Anatomical Therapeutic Chemical (ATC)-code] connects chemical classification and therapeutic approach. It is generally accepted that compounds with similar physicochemical properties exhibit similar biological activity. If this hypothesis holds true for drugs, then the ATC-code, the putative medical indication area and potentially the medical target should be predictable on the basis of structural similarity. We have validated that the prediction of the drug class is reliable for WHO-classified drugs. The reliability of the predicted medical effects of the compounds increases with a rising number of (physico-) chemical properties similar to a drug with known function. The web-server translates a user-defined molecule into a structural fingerprint that is compared to about 6300 drugs, which are enriched by 7300 links to molecular targets of the drugs, derived through text mining followed by manual curation. Links to the affected pathways are provided. The similarity to the medical compounds is expressed by the Tanimoto coefficient that gives the structural similarity of two compounds. A similarity score higher than 0.85 results in correct ATC prediction for 81% of all cases. As the biological effect is well predictable, if the structural similarity is sufficient, the web-server allows prognoses about the medical indication area of novel compounds and to find new leads for known targets.

Availability: the system is freely accessible at http://bioinformatics.charite.de/superpred. SuperPred can be obtained via a Creative Commons Attribution Noncommercial-Share Alike 3.0 License.

INTRODUCTION

The accessibility of large compound databases has changed from exclusive inhouse databases of large pharmaceutical companies to publicly available sources (1). At this time several million different compounds can be obtained from different vendors (2). About 7000 drugs currently exist and there are about 480 validated targets that are addressed (3). There are estimations about the number of medical targets between 2200 and 3000 that favour interactions with drug-like chemical compounds (4). To map these medical targets onto medical indication areas a classification scheme is needed.

Currently, the most commonly used classification system for drugs is the **A**natomical **T**herapeutic **C**hemical (ATC) classification system. This scheme is recommended by the World Health Organization (WHO) for all global drug utilization studies and categorizes drug substances at different levels according to application area, therapeutic properties, chemical and pharmacological properties (5).

A challenging aim is the mapping of the available compounds onto about 850 ATC-classes. The progress in understanding the mechanisms of action of a vast majority of drugs gives the opportunity to narrow down the gap between the medical indications and elucidation of drug effects at the molecular level. The relation between the structure of a compound and its biological activity was well investigated in some systematic analyses (6–8). It could be shown that a Tanimoto coefficient of >0.85 indicates that two molecules have similar activities (8). Based on this principle, it should be possible to predict medical indication areas for unclassified chemical compounds in case of sufficient structural similarity. A method based on the similar property principle (9) for predicting activity spectra of substances was described by Lagunin (10) and confirmed by several experiments (11,12). The PASS application is available at http://www.ibmc.msk.ru/PASS. Furthermore, new medical indication areas for approved drugs or drug candidates can be found by

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

applying this rule. Indeed recently, much efforts have been put into drug repositioning (13,14). To discriminate between drugs and nondrugs, the use of property distributions and (physico-) chemical descriptors is already used successfully (15,16). The increased knowledge about drug-target-pathway relations and the integration of molecular similarity with property distribution allow improved structure–function prediction. Here, we present a publicly available web-server to predict medical indication areas based on properties and similarity of chemical compounds.

METHODS

Data set for the web-server

The web-server called SuperPred was created for recognition of as many drug classes as possible. For this reason, the number of medical compounds was enlarged to about 6300. The calculated fingerprints from the 2500 compounds of the SuperDrug database were used for a further structural screening against the SuperTarget database (17). In this way, 3800 additional compounds were detected that are structurally very similar to drugs and resulted in Tanimoto coefficients of at least 0.85. These putative drugs are most likely candidates for having the same mode of action, binding to the same target/enzyme and being assigned to the same medical indication as the WHO-classified drugs. In order to allow the examination of the drug effect on a molecular level, information about the target proteins was extracted from literature and was provided for half of the drugs (17).

Reduced data set for prediction evaluation

For the purpose of statistical evaluation of the prediction accuracy, a subset consisting of 1035 drugs was utilized. The members of the subset were chosen according to the following rules: First, every drug having more than one indication was removed; then each included ATC-group had to consist of at least 3 molecules and last, to eliminate outliers, the drugs within one ATC-group were not allowed to deviate more than 1.5-fold from the average Tanimoto score of the group. Furthermore, ATC-codes with very similar indications were organized into ATC-classes. For instance, 'corticosteroids, moderately potent' (ATC: D07AB), 'corticosteroids, potent' (ATC: D07AC), 'corticosteroids, very potent' (ATC: D07AD) and 'corticosteroids, plain' (ATC: S01BA) were combined to form the ATC-class 'corticosteroids' (see website and Supplemental material).

Prediction

The prediction was carried out by the combination of physicochemical property analyses and similarity searching. The prediction of the ATC-class was performed by the assignment of the compound to the ATC-class of the most similar drug and property distribution. The prediction accuracy was determined by the leave-one-out cross-validation method.

Physicochemical properties

Lipinski's Rule of Five (18) is a general accepted standard for oral applicable drugs. The rule describes molecular properties important for a drug's absorption, distribution, metabolism and excretion in the human body. It is stated that an orally active drug has not more than 5 hydrogen-bond donors, not more than 10 hydrogen-bond acceptors, a molecular weight below 500 g/mol and a logP less than 5. These properties and several more were calculated for each drug in SuperPred. The distributions of the properties' values were saved in the database for each ATC-group and -class. In this way, the range of property values is comparable with the query.

Similarity searches

To calculate the similarity between two compounds, their structural fingerprints, generated by **C**hemistry **D**evelopment Tool**K**it (CDK) (http://almost.cubic.uni-koeln.de/cdk/), were used. Structural fingerprints are bit-vectors encoding for the chemical and topological features of small molecules. The similarity is determined by the Tanimoto coefficient (19):

$$T = \frac{N_{ab}}{N_a + N_b - N_{ab}}$$

where N_a is the number of bits set to 1 in compound a, N_b is the number of bits set to 1 in compound b and N_{ab} is the number of bits common to both, compounds a and b.

Input and output options

There are three ways to start a query with a molecule not included in the database:

- Enter SMILES (Simplified Molecular Input Line Entry System)
- Draw a molecule using Marvin Sketch
- Upload a MOL file using Marvin Sketch

Medical compounds can be retrieved through an expandable ATC-tree, by name, synonym, ATC-code or via known target (name, Uniprot-ID).

The output is a structured table, listing predictions (ATC-codes including confidence interval) containing similarity scores, compound-IDs, molecular structure visualized by Marvin View, target information and physicochemical property intervals of the ATC-group and of the query compound. The score and the color indicate the power of the prediction visualized.

RESULTS

Prediction results

The prediction accuracy is determined by the fraction of correct ATC-class predictions and amounts 67.6%. The distribution of the fractions of correctly predicted indications is shown in Table 1. For a Tanimoto coefficient >0.85 an accuracy of 80.6% is accomplished. A cumulative recall graph is shown in Figure 1. The graph shows the fraction of right predictions of ATC-classes in dependency

of the quantity of retrieved structures. By retrieving three molecules the recall gains to about 80% and with 20 retrieved molecules a recall of about 90% is achieved.

For the reduced data set of 1035 drugs, the recall is cumulated for one retrieved drug up to twenty retrieved drugs. With three retrieved molecules the recall gains to about 80% and with 20 retrieved molecules a cumulative recall of about 90% is achieved.

Table 1. Distribution of the fractions of correctly predicted indications

Range of Tanimoto coefficient	Numbers of hits/misses	Fraction of hits
0.4–0.5	5/18	21.7
0.5–0.6	18/27	40.0
0.6–0.7	40/60	40.0
0.7–0.8	93/84	52.5
0.8–0.9	171/58	74.7
0.9–1.0	367/79	82.3
0.0–1.0	700/335	67.6

For the reduced data set of 1035 drugs, 700 right and 335 wrong predictions are investigated. In detail: a similarity score of 90–100% specifies the correct ATC-class in about 82% (367 right and 79 wrong predictions). A hit/miss-rate of about 3/1 is achieved for similarity scores of 70% and higher.

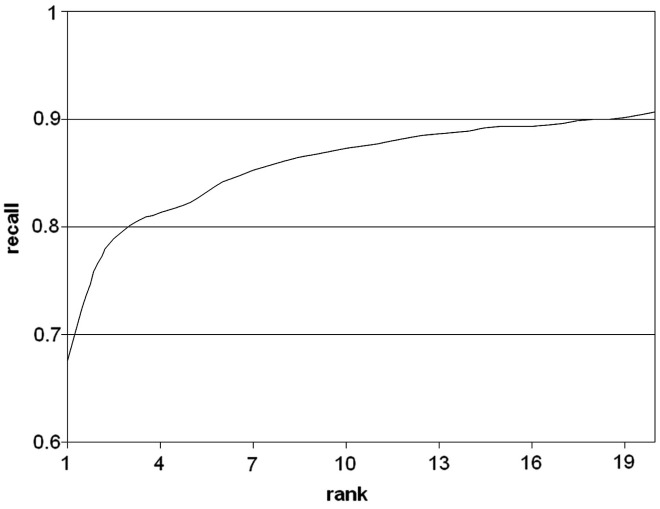

Figure 1. Cumulative recall for ATC-recognition relative to rank of retrieval.

Case study

Besides leave-one-out cross-validation statistics, the prediction method was proved by a number of compounds extracted from the SuperTarget database as well as compounds experimentally tested against tumor-cell line assays.

Starting point for the first screening was Enalapril, an ACE-inhibitor, that is used in treatment of hypertension and congestive heart failure. SuperPred identifies six putative drugs having a sufficient similarity to Enalapril indicated by a green color in the result table (Tanimoto coefficient >0.8). An inspection of the referenced literature via the Pubchem-database denoted a similar medical effect for all of them. Table 2 shows exemplarily the names of three of the six putative compounds and the associated reference that describes the medical effect of inhibiting the angiotensin-converting enzyme.

The National Cancer Institute Developmental Therapeutics Program (DTP) has screened about 100 000 compounds against a panel of 60 human tumor-cell lines. The results are available on the DTP web site (http://dtp.nci.nih.gov/). The growth inhibition (GI_{50}) and lethal dose (LD_{50}) of the compounds are also retrievable.
Application:

(i) NCI-compound (NSC: 600221) is a screening hit with unverified target. This compound shares a Tanimoto coefficient of 0.92 with the compound Paclitaxel and therefore, it is predicted to be an antineoplastic agent targeting Tubulin.
(ii) Enalapril is a well-known ACE-inhibitor. The compound Sch31846 has a similarity of 95% and is supposed to be an ACE-inhibitor, too.

Isolated compounds of the comprehensive tumor-related information resource of the NCI were extracted and screened against the approved drugs included in SuperPred. Many of the screening candidates were characterized by a high physicochemical similarity to well-annotated anti-cancer drugs. For instance, the compound NSC 600221 (Table 2) and the antineoplastic agent Paclitaxel hold a Tanimoto coefficient of 0.91. Both compounds are

Table 2. Compounds identified with SuperPred and similar to Enalapril and NSC 600221, respectively

Name of the compound	Tanimoto coefficient	Medical function	Target protein	Reference
Enalapril	100.00	ACE-inhibitor	Angiotensin-converting enzyme	(22)
Sch 31846[a]	94.57	ACE-inhibitor (predicted)	Angiotensin-converting enzyme (predicted)	(23)
Delapril hydrochloride	83.84	ACE-inhibitor (predicted)	Angiotensin-converting enzyme (predicted)	(24)
Hoe 065[b]	81.90	ACE-inhibitor/increasing central cholinergic activity (predicted)	Angiotensin-converting enzyme (predicted)	(25)
NSC 600221[c]	100.00	Antineoplastic agent	Tubulin (predicted)	http://dtp.nci.nih.gov
Paclitaxel	91.62	Antineoplastic agent	Tubulin beta-1chain	(26)

[a](2S,3aS,7aS)-1-((S)-N-((S)-1-Carboxy-3-phenylpropyl)alanyl) hexahydro-2 indolinecarboxylic acid, 1-ethyl ester, monohydrochloride.
[b]Cyclopenta(c)pyrrole-1-carboxylic acid, 2-(2-((1-(ethoxycarbonyl)-3-phenylpropyl)amino)-1-oxopropyl)octahydro-, octyl ester, (1S-(1-alpha,2-(R*(R*)),3a-beta,6a-alpha))-, (Z)-2-butenedioate (1:1).
[c]Beta-Phenylalanine, N-benzoyl-2-[[(2-carboxyethyl) carbonyl]oxy]-, 6,12b-diacetoxy-12-(benzoyloxy)-2a,3,3a,4,5,6,9, 10,11,12,12a,12b-dodecahydro-4,11- dihydroxy-4a,8,13, 13-tetramethyl-5-oxo-7,11-methano- 1H-cyclodeca[3,4]benz[1, 2-b]oxet-9-yl ester.

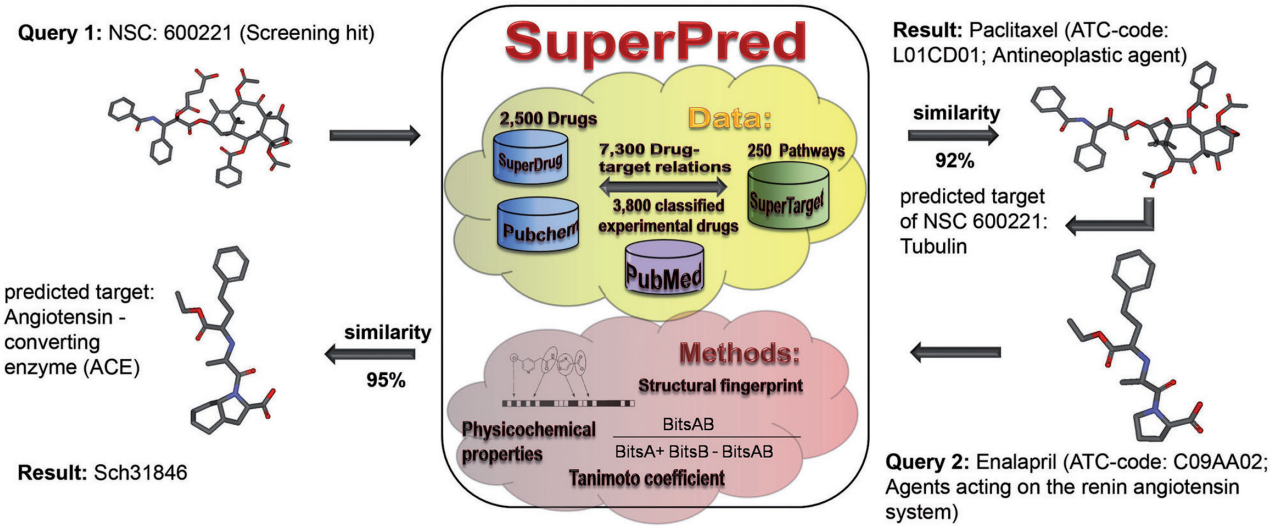

Figure 2. Assembly of the SuperPred server and possible requests for ATC-code prediction. Data: the SuperPred server now contains 2500 compounds of the SuperDrug database. Additionally, 3800 experimental drugs were classified and stored on the server. The drugs are annotated by 7300 links to targets. Methods: the structural properties of the compounds are stored in so-called structural fingerprints, where each bit encodes for an element of the compound structure. The similarity of two compounds is calculated by using the Tanimoto coefficient. Moreover, physicochemical properties are stored for each compound. SuperPred can be used to find new targets for ligands and vice versa to find new ligands for medical biological targets. There are two possibilities to use the SuperPred server. The figure shows two examples for querying the SuperPred server.

shown in Figure 2. To analyze the ability to inhibit the proliferation of cancer cells, the GI_{50}-values of both compounds were analyzed with COMPARE, a web accessible tool for investigating mechanisms of cell GI (20). The ability to inhibit the growth of the diverse set of cell lines was highly similar and was indicated by a correlation coefficient of 0.87 calculated by COMPARE. The high correlation coefficient even allowed predictions about the target protein of NSC 600221 (21). As Paclitaxel inhibits microtubule formation by binding to tubulin, the same target came into question for NSC 600221.

CONCLUSION

The SuperPred web-server was created for predicting medical indications for chemical compounds. The combination of physicochemical property and similarity searching provides the possibility to detect new biologically active compounds and novel targets for drug-like compounds. SuperPred can be applied for drug repositioning purposes, too. A further intention of SuperPred is to find side effects elicited by drugs caused through off-target hits. The use of the web-server is free for all academics.

ACKNOWLEDGEMENTS

The authors wish to thank A. Eckert for assistance during software testing and improvement and S. Struck for critical reading of the article. This work was supported by Deutsche Forschungsgemeinschaft: SFB 449, IRTG Berlin-Boston-Kyoto and Deutsche Krebshilfe. Funding to pay the Open Access publication charges for this article was provided by DFG-Sonderforschungsbereich 449.

Conflict of interest statement. None declared.

REFERENCES

1. Voigt,J.H., Bienfait,B., Wang,S. and Nicklaus,M.C. (2001) Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.*, **41**, 702–712.
2. Baurin,N., Baker,R., Richardson,C., Chen,I., Foloppe,N., Potter,A., Jordan,A., Roughley,S., Parratt,M., Greaney,P. *et al.* (2004) Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J. Chem. Inf. Comput. Sci.*, **44**, 643–651.
3. Drews,J. and Ryser,S. (1997) The role of innovation in drug development. *Nat. Biotechnol.*, **15**, 1318–1319.
4. Russ,A.P. and Lampel,S. (2005) The druggable genome: an update. *Drug Discov. Today*, **10**, 1607–1610.
5. WHO expert committee (2006) The selection and use of essential medicines. Report of the WHO expert committee, 2005 (including the 14th model list of essential medicines). *World Health Organ. Tech. Rep, Ser.*, **1–119**, back cover.
6. Basak,S.C., Gute,B.D. and Mills,D. (2002) Quantitative molecular similarity analysis (QMSA) methods for property estimation: a comparison of property-based, arbitrary, and tailored similarity spaces. *SAR QSAR Environ. Res.*, **13**, 727–742.
7. Liu,X., Yang,Z. and Wang,L. (2005) Three-dimensional, quantitative-structure-property-relationship study of aqueous solubility for phenylsulfonyl carboxylates using comparative-molecular-field analysis and comparative-molecular-similarity-indices analysis. *Water Environ. Res.*, **77**, 519–524.
8. 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.
9. Barbosa,F. and Horvath,D. (2004) Molecular similarity and property similarity. *Curr. Top. Med. Chem.*, **4**, 589–600.
10. Lagunin,A., Stepanchikova,A., Filimonov,D. and Poroikov,V. (2000) PASS: prediction of activity spectra for biologically active substances. *Bioinformatics*, **16**, 747–748.
11. Poroikov,V., Filimonov,D., Lagunin,A., Gloriozova,T. and Zakharov,A. (2007) PASS: identification of probable targets and mechanisms of toxicity. *SAR QSAR Environ. Res.*, **18**, 101–110.
12. Geronikaki,A.A., Lagunin,A.A., Hadjipavlou-Litina,D.I., Eleftheriou,P.T., Filimonov,D.A., Poroikov,V.V., Alam,I. and Saxena,A.K. (2008) Computer-aided discovery of anti-inflammatory

thiazolidinones with dual cyclooxygenase/lipoxygenase inhibition. *J. Med. Chem.*, **51**, 1601–1609.
13. O'Connor,K.A. and Roth,B.L. (2005) Finding new tricks for old drugs: an efficient route for public-sector drug discovery. *Nat. Rev. Drug Discov.*, **4**, 1005–1014.
14. Ashburn,T.T. and Thor,K.B. (2004) Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.*, **3**, 673–683.
15. Byvatov,E., Fechner,U., Sadowski,J. and Schneider,G. (2003) Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.*, **43**, 1882–1889.
16. Sadowski,J. and Kubinyi,H. (1998) A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.*, **41**, 3325–3329.
17. Gunther,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.
18. Lipinski,C.A., Lombardo,F., Dominy,B.W. and Feeney,P.J. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, **46**, 3–26.
19. Delaney,J.S. (1996) Assessing the ability of chemical similarity measures to discriminate between active and inactive compounds. *Mol. Divers.*, **1**, 217–222.
20. Zaharevitz,D.W., Holbeck,S.L., Bowerman,C. and Svetlik,P.A. (2002) COMPARE: a web accessible tool for investigating mechanisms of cell growth inhibition. *J. Mol. Graph. Model.*, **20**, 297–303.
21. Gunther,S., Neumann,S., Ahmed,J. and Preissner,R. (2007) Cellular fingerprints: a novel concept for the integration of experimental data and compound-target-pathway relations. *LNBI 4643*, 167–170.
22. Imanishi,T., Ikejima,H., Tsujioka,H., Kuroi,A., Kobayashi,K., Muragaki,Y., Mochizuki,S., Goto,M., Yoshida,K. and Akasaka,T. (2008) Addition of eplerenone to an angiotensin-converting enzyme inhibitor effectively improves nitric oxide bioavailability. *Hypertension.*, **51**, 734–741.
23. La Rocca,P.T., Squibb,R.E., Powell,M.L., Szot,R.J., Black,H.E. and Schwartz,E. (1986) Acute and subchronic toxicity of a nonsulfhydryl angiotensin-converting enzyme inhibitor. *Toxicol. Appl. Pharmacol.*, **82**, 104–111.
24. Fogari,R., Malamani,G., Zoppi,A., Mugellini,A., Rinaldi,A., Fogari,E. and Perrone,T. (2007) Effect on the development of ankle edema of adding delapril to manidipine in patients with mild to moderate essential hypertension: a three-way crossover study. *Clin. Ther.*, **29**, 413–418.
25. Grupp,L.A. and Chow,S.Y. (1991) Effects of the novel compound Hoe 065, a central enhancer of cholinergic activity, on voluntary alcohol consumption in rats. *Brain Res. Bull.*, **26**, 617–619.
26. Rao,S., He,L., Chakravarty,S., Ojima,I., Orr,G.A. and Horwitz,S.B. (1999) Characterization of the Taxol binding site on the microtubule. Identification of Arg(282) in beta-tubulin as the site of photoincorporation of a 7-benzophenone analogue of Taxol. *J. Biol. Chem.*, **274**, 37990–37994.