A structural keystone for drug design

Kristian Rother[1*], Mathias Dunkel[1], Elke Michalsky[1], Silke Trissl[2], Andrean Goede[1], Ulf Leser[2], Robert Preissner[1]

[1]Institute of Biochemistry, Charité Universitätsmedizin Berlin, Monbijoustraße 2, 10117 Berlin, Germany

[2]Institute of Computer Science, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

Abstract

3D-structures of proteins and potential ligands are the cornerstones of rational drug design. The first brick to build upon is selecting a protein target and finding out whether biologically active compounds are known. Both tasks require more information than the structures themselves provide. For this purpose we have built a web resource bridging protein and ligand databases. It consists of three parts: i) A data warehouse on annotation of protein structures that integrates many well-known databases such as Swiss-Prot, SCOP, ENZYME and others. ii) A conformational library of structures of approved drugs. iii) A conformational library of ligands from the PDB, linking the realms of proteins and small molecules.

The data collection contains structures of 30,000 proteins, 5,000 different ligands from 70,000 ligand-protein complexes, and 2,500 known drugs. Sets of protein structures can be refined by criteria like protein fold, family, metabolic pathway, resolution and textual annotation. The structures of organic compounds (drugs and ligands) can be searched considering chemical formula, trivial and trade names as well as medical classification codes for drugs (ATC). Retrieving structures by 2D-similarity has been implemented for all small molecules using Tanimoto coefficients. For the drug structures, 110,000 structural conformers have been calculated to account for structural flexibility. Two substances can be compared online by 3D-superimposition, where the pair of conformers that fits best is detected. Together, these web-accessible resources can be used to identify promising drug candidates. They have been used in-house to find alternatives to substances with a known binding activity but adverse side effects.

Availability: http://bioinformatics.charite.de

1 Introduction

In drug design, one faces the problem to predict reasonably which molecules from a pool of millions of possible compounds will interact with a target of medical or biological interest. One approach is to utilize 3D-models of target proteins and ligands. In order to build reasonable and useful models, as much information as possible has to be incorporated into the modelling process. Target proteins are found in overwhelming numbers in sequence and structure databases. For small organic compounds, a multitude of databases containing formulae and structures

[*] corresponding author: kristian.rother@charite.de

exists in equally impressive numbers. However, the data are often scattered over multiple resources and often available in a non-uniform manner. To clarify the matter, five different types of databases are sketched here:

The first group contains resources of structures of low molecular weight compounds that are potential ligands. These can be experimentally determined 3D-structures of low molecular weight structures, like the more than 250,000 molecules found in the CSD [1], or computed structures of chemical compounds. The Asinex and NCI [16] databases hold such computed structures. The Chembridge database [6] contains one of the highest numbers of compounds (700,000).

The second group provides structures of proteins, and here the Protein Data Bank (PDB) [5] is the only player in the field. Information in the PDB has often been enriched by so-called 'secondary databases' that belong to the third group. These databases provide fold classifications [2], enzymatic functions [3], links to sequence data [4], non-redundant subsets [39], and the list could continue for some paragraphs.

The fourth group focuses on databases delivering structures and additional annotation about ligand molecules from the PDB. The chemical and spatial information within ligand structures can be used to refine protein models, specifically to optimise side-chain conformations around binding-sites. HIC-Up [19] comprises chemical and structural information for small molecules found in the PDB. Ligand Depot [11] and Relibase [15] provide a graphical interface search among the ligands by two-dimensional similarity and chemical substructure as well as for sequence similarity search among the corresponding proteins. LigBase [35] is a database of ligand binding sites aligned with related protein structures and sequences.

In the fifth and last group, several of the services described above have been combined. Some of them digest PDB entries and attach a wealth of information to them, such as PDBSum [20], the Macromolecular Structure Database (MSD) [38], and the IMB Jena Image Library of Biological Macromolecules [32]. The Jena Image Library can also be searched by geometrical properties of the ligand binding sites.

None of the databases listed here directly focuses on drug design. There exists a number of commercial databases that cover a broad range of bioactive compounds, developmental drugs or patented compounds (WDI: 58,000; CMC: 7,500; MDDR: 106,000). Computed structures of drugs are also commercially available, but no publicly available source providing conformers exists.

To overcome this lack, we have created structure-based databases that cover three different aspects important in the drug design process. As stated above, protein targets are often a starting point for this. First, the Columba database [37] (www.columba-db.de) provides easy access to annotation on protein structures. For modelling and simulation purposes, chemical and spatial information about protein ligands is vitally important. Second, SuperLigands [27] (bioinformatics.charite.de/superligands) - a collection of small molecule structures contained in the PDB - addressing this fact by facilitating comparison of the molecules regarding their two- and three-dimensional similarity. Finally, structures of well-characterized and approved drugs have been compiled in the SuperDrug database [13] (bioinformatics.charite.de/superdrug). In this paper, we describe the content and query options of these databases, and how they interact. We outline by what methods this data can be used efficiently in order to find promising drug candidates.

2 Methods

2.1 Protein structure annotation

As the number of protein structures deposited in the PDB is growing rapidly, it becomes more and more important to have efficient ways to find structures of interest. Columba is a data warehouse of information on protein structures that physically integrates information about structural and sequence-based classification schemes, functional annotation, secondary structural elements, and participation in metabolic pathways [37].

The Columba database integrates twelve databases related to annotation of structures from the PDB [5]. This is reflectted in the data model, where data from the PDB is at the center, and data from other sources are grouped around it in a star-shaped manner. General information from the PDB entries is accompanied by a description of all compounds or biological units of that entry, and the polypeptide chains a particular compound consists of. Compounds having an enzyme classification (E.C.) number are annotated with functional information from ENZYME [3], and with the participation of that enzyme in metabolic pathways from the Kyoto Encyclopedia of Genes and Genomes, KEGG [18]. Columba also integrates data from the Roche Biochemical Pathway Map [26].

The protein chains are linked to the two hierarchical fold classification schemes, i.e. SCOP [2] and CATH [28]. Furthermore, each chain is assigned to a cluster based on sequence identity offered by the PDB itself [21]. Also, culled subsets of protein chains according to sequence identity and experimental properties from PISCES [39] are included. For each chain, the secondary structure is computed using the DSSP program [17]. Links to Swiss-Prot entries [4] were retrieved from the PDBSprotEC database [23]. Exploiting the links from Swiss-Prot to other databases, PDB chains are directly connected to the NCBI Taxonomy database [40] and functional annotation from Gene Ontology [7].

2.2 Ligands from the PDB

Currently, the PDB contains more than 30,000 protein structures most of which have one or many low molecular weight compounds attached to protein chains. The native conformations of these small molecules and additional information were collected from the PDB [5], Ligand Depot [11] and MSD [38] databases. The SuperLigands database [27] delivers these PDB ligands in the MDL Mol file format which, in contrast to the PDB format, includes information about bond types.

To enable fast similarity-based screenings against the whole database, the following 2D- matching procedure was established: From the two-dimensional structural formulae of all ligands, 960 bit binary fingerprints (MDL MACCS Keys [10]) representing the occurrence of most chemical groups and their topology were calculated and stored in the database. Two fingerprints can be compared using the Tanimoto coefficient [9, 24], which is defined as

$$T(a,b) = \frac{N_{ab}}{N_a + N_b - N_{ab}}, \tag{1}$$

where N_a and N_b are the numbers of bits set in the fingerprint of structure a and b, respectively, and N_{ab} is the number of bits which are common to both fingerprints. This way, the whole database can be screened rapidly with a given ligand resulting in a list of similar ones sorted by $T(a, b)$.

Additionally, a procedure for three-dimensional superposition of two PDB ligands has been implemented, as 3D-fragment based comparisons were shown to supplement 2D-screening results [8, 36]. For comparing the structures of two ligands from SuperLigands, they are superimposed with each other using the algorithm of [30]. The resulting superpositions are ranked by a score defined by

$$score = \frac{Number\ of\ superposed\, atoms}{Number\ of\ atoms\ in\ the\ smaller\ molecule} e^{-RMSD}, \tag{2}$$

where RMSD is the Root Mean Square Deviation of the superposed atoms. This procedure allows to match two molecules more reliably than the 2D-comparison alone: The 2D-algorithm sorts out relatively similar molecules by chemical criteria, while the 3D-superposition can distinguish these database hits further by taking their structural properties into account. The freely available MDL®Chime plug-in is used to display molecules and allows the user some manipulations of the view and to store the displayed molecule in the MDL Mol file format.

2.3 Structures of known drugs

Well-characterized, known drugs approved by the WHO have been collected in the SuperDrug database [13]. The Chemical Abstracts (CA) provide information on drugs including the CAS-number, useful as cross-reference to other databases and the chemical 2D-structure. The latter was used to generate 3D-structures using Discovery Studio from Accelrys. As most low molecular weight compounds have many degrees of freedom in their rotatable bonds, the low- energy conformational space of each molecule was sampled using the algorithm of [34] as implemented in MedChem Explorer from Accelrys. This resulted in at most 105 structures of conformers per ligand, with an average of 47 conformers. This addtional information can be used for comparing two drug structures: each conformer of one molecule is superimposed with each conformation of the other using the same algorithm as for PDB ligands. The pair of conformations with the highest score is displayed.

Recently, the recommendations of the WHO Expert Committee responsible for updating the WHO Model List of Essential Medicines were published [41]. For the first time, a list was given that sorts all compounds on the Model List according to their 5-level Anatomical Therapeutic Chemical (ATC) classification codes was given. The ATC classification was also included in the database.

3 Results and Discussion

3.1 Database content

The database of structures consisting of three parts was set up as described in the methods section. 31,926 protein structures from the PDB and annotation from 11 other sources were compiled into the Columba database. Structures of 72,951 ligands of 5,040 different low molecular weight compounds were identifed for the SuperLigands database. Structures of 2,396 WHO-approved drugs have been calculated for the SuperDrug database. In Table 1, the exact numbers of elements from each source database are shown.

To allow comparison of small molecules by superposition, their flexibility needs to be considered. For that, an average of 47 conformers per drug have been calculated resulting in a total of 110,000 conformers for the drug molecules. The PDB is known to be highly redundant; already 10% of the database are composed by the five most frequently occuring proteins, namely Lysozyme, Ribonuclease, Hemoglobin, Immunoglobulin and Cytochrome [33].

In contrast, each of the small molecules is unique, although many small molecules have been deposited in the PDB one would not regard as typical ligands. These are metal ions and groups covalently bound to proteins, like many phosphate and saccharide residues. Looking at the ligands occuring most frequently in the PDB, one finds sulfate groups (SO4), N-acetyl- glucosamine (NAG), glycerol (GOL), N-dimethyl-lysine (MLY) and heme (HEM), together making up 30% of the PDB ligands. Also, well-known coenzymes like NADH (NAD) and FADH_2 occur very often.

The PDB contains hetero groups that are drugs, that are drug-like ligands, and others. All of them have been included in SuperLigands for two reasons: it is known that many drugs act as antagonists by having a structure similar to a coenzyme or messenger, e.g. coffeine and adenine. A user is thereby able to find the names of the three-letter abbreviations that often remain obscure in PDB files. Each molecule contained in the drug database is being used as a drug. The most frequent drug class are antibacterials for systemic use. The classes mostly prescribed are antidepressants and antihistamines.

3.2 Comparison of PDB ligands with drugs

To assess the usability of the databases presented above, it is important how many of the ligands are drugs or have at least similar properties. Thus, all structures from the SuperLigands and SuperDrug databases were screened against each other using the Tanimoto coefficient as described in the Methods section. It is generally accepted that similar compounds having Tanimoto coefficients larger than 0.85 tend to exhibit similar biological activity [25]. The cross-comparison resulted in a total of twelve million 2D-scores. Of the 5,040 different PDB ligands, 413 could be matched to a drug molecule with a Tanimoto coefficient of 100%. Extending the analysis to 90% similarity, already 1,475 of the PDB ligands had a drug counterpart and were considered as drug-like. It has to be noted, that the 960-bit fingerprints are not unfailable, because their information is degenerate. Theoretically, molecules with a different chemical configuration of atoms can have the same fingerprint.

To characterize the drug-likeness of the PDB ligands from a more general point of view, the Lipinski ’Rule of five’ [22] was checked for both sets of compounds. These are rough guidelines what properties a molecule needs in order to be likely suitable as a drug in terms of transportability and toxicity (molecular weight< 500, logP< 5, hydrogen bond donors< 5 and acceptors < 10). As can be seen from table 2, almost 90% of the drugs violate none or one of these rules. But only 81% of the PDB ligands are in this region. Obviously, many of the coenzymes and saccharides in the PDB violate one or more of the rules. Compounds violating more than one of the Lipinski Rules are assumed to have problems with bioavailability. However, metal ions will not be excluded by these measures, but their proportion on the 5,040 ligands is small while they occur in the PDB relatively often. Analyzing the molecules in more detail, the PDB ligands tend to have more hydrogen bond donors and acceptors, lower logP values, and are heavier [27]. This analysis reveals that despite these differences a significant amount of the ligands in the PDB either are drugs themselves, are structurally similar to drugs, or have at least similar chemical properties as drugs making them suitable as potential templates for drug design.

3.3 Web interface

Columba can be searched through a web interface available at http://www.columba-db.de. The interface allows two types of queries: Full text search and queries specific for data sources and attributes. In both cases, the query results in a list of PDB entries. Queries can be combined, allowing to narrow down the desired set of entries iteratively. The resulting data sets and reports for individual structures can be viewed on the web or downloaded as XML data. Online molecular visualisation via the Java based JMol application is also included. By these means, the Columba web interface greatly reduces the required time to collect information for any list of PDB entries.

SuperLigands can be searched by chemical name, three-letter-abbreviation, formula and PDB code. The interface displays a list of the molecules found and provides links to a detailed description of them. Besides that, a user may search for compounds by chemical 2D-similarity using the Tanimoto coefficient or assess the three-dimensional similarity of two compounds by superposing them with each other.

The SuperDrug database can be queried by ATC codes, scientific and trivial names of drugs and chemical formulae. Methods for calculating 2D- and 3D-similarity measures to other drug molecules have been implemented in the same way as for PDB ligands. There are commercial databases which also provide drug structures, but the SuperDrug database is the first exhaustive free resource for WHO-classified drugs.

As a bridge between the SuperLigands and SuperDrug databases, the similarity of PDB ligands to known drugs can be assessed in a comfortable manner. Starting with a ligand, a two-dimensional similarity search will detect the 30 drug structures having the best Tanimoto coefficients. The drug structures found can be superposed in 3D, which is not easily possible in similar databases.

3.4 Opportunities for drug design

Together, these three databases provide a unique combination of resources that support the drug design process. In our group, several approaches are pursued based on structural 3D-similarity. Obviously, the knowledge about known ligands can be used to find similar substances that bind to the same target protein and, in contrast to the native ligands, act as inhibitors. The database of known drugs is a resource containing such potential agents, but it is not restricted to it. Moreover, known ligands can serve as a starting point for drug screening themselves. By integrating libraries of other low molecular weight compounds such as NCI containing over 250,000 substances, molecules that have a high structural similarity to known drugs can be examined. This opens possibilities to characterize the effect of well-known substances such as alkaloids in more detail. We have also established a procedure to design ligands based on protein structures. Here, the surface of a protein is decomposed into patches ranging from 30 to 150 atoms [30]. Such patches on the protein surface have been used to design peptidic inhibitors [14], or can be processed further by looking for organic molecules that can replace parts of a peptide [31]. In both cases, the surface of a protein is mimicked by a smaller molecule, enabling a smaller compound to activate or inhibit processes in the same way as the native protein does in vivo. It is also thinkable to find target proteins for a certain ligand, like was done by Paul et al. in [29]. There, a collection of protein active sites was extracted from the PDB and scanned with a docking algorithm.

It has recently been demonstrated that the combination of quick 2D and more accurate 3D-screening methods is able to find inhibitors of the COP9 signalosome associated kinases (CSN). Seven compounds out of thirty-five candidates were verified experimentally to inhibit the CSN kinase [12]. By the resources described above, this approach can be supplemented by taking into account structural information of the target proteins and of drugs already known. This way, it is also thinkable to devise new applications for old substances.

4 Conclusions

The databases presented here supplement the existing resources of information about small molecules and protein structures. As novel features, the SuperLigands and SuperDrug databases provide three-dimensional comparison of small molecules, moreover topology comparison of PDB ligands with known drugs is made possible. Columba has proven to be very useful for a number of tasks in our own structural research. Generating sets of structures, which previously required days of manual browsing or writing of parsers, now only takes a few mouse clicks, or an SQL query. These databases will be extended and unified further, leading to a single structure-based database focusing on ligands and drug-like substances in the long run.

Rational drug design is a complicated process that consists of many steps. We have described, how to obtain molecules that can be used to find interesting drug candidates e.g. by in-silico similarity screening. Knowing that the substances in the SuperDrug database are suitable in terms of bioavailability and toxicity, is important to compare them virtually to find drug candidates that also have these crucial properties. In any case, the resulting candidate substances need to be tested experimentally to verify that they really have the desired function and are free from adverse effects. Selecting candidates via the described in-silico screening should help to reduce the list of candidates to some promising molecules, and thereby reduce time and costs significantly. The databases described in this paper are available via http://bioinformatics.charite.de.

5 Acknowledgements

We like to thank the following persons that have participated in building and maintaining the various resources: Andrean Goede, Stefan Günther, Raphael Bauer, Thomas Steinke, Heiko Müller and Nina Mester. We also thank Cornelius Frömmel for various inputs. The project was funded by BMBF Grant No. O312705B.

Table 1: Types of data integrated into the databases presented here.

Source database	Number of entries included	type of entry
PDB	31,926	protein structures
DSSP	65,809	secondary structures of protein chains
SCOP	70,859	fold classifications of protein chains
CATH	67,054	fold classifications of protein chains
KEGG	156	metabolic pathways
ENZYME	4,290	enzyme functions
Swiss-Prot	189,543	protein sequences
PDBSprotEC	30,205	references to Swiss-Prot
GO	19,015	biological terms
GOA	23,7751	references to GO terms
PDB	63,551	cluster entries of PDB structures
PISCES	34,3582	assignment of PDB structures to non-redundant lists
NCBI	27,8791	taxonomic tree of species
SuperLigands	72,951	protein ligands
SuperDrug	2396	approved drugs

Table 2: Percentage of PDB ligands and drugs violating certain numbers of Lipinski Rules

Number of violated Lipinski Rules	PDB ligands	Drugs
0	64.44%	75.65%
1	16.87%	13.96%
2	10.32%	5.69%
3	8.29%	4.69%
4	0.08%	0.00%

References

[1] F. H. Allen, J. E. Davies, J. J. Gallo, O. Johnson, O. Kennard, C. F. Macrae, E. M. Mitchell, G. F. Mitchell, J. M. Smith, and D. Watson. The development of versions 3 and 4 of the cambridge structural database system. *J Chem Inf Comput Sci*, 31:187–204, 1991.

[2] A. Andreeva, D. Howorth, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*, 32:D226 – D229, 2004. Database issue.

[3] A. Bairoch. The enzyme database in 2000. *Nucleic Acids Res*, 28:304–305, 2000.

[4] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L.-S. L. Yeh. The Universal Protein Resource (UniProt). *Nucleic Acids Res*, 33(Database issue):D154–159, 2005.

[5] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–242, 2000.

[6] M. P. Bradley. An overview of the diversity represented in commercially-available databases. *Mol Divers*, 5(4):175–183, 2002.

[7] G. O. Consortium. The Gene Ontology (GO) database and inforamtics resource. *Nucleic Acids Res*, 32:D258 – D261, 2004. Database issue.

[8] R. D. Cramer, R. J. Jilek, and K. M. Andrews. Dbtop: topomer similarity searching of conventional structure databases. *J Mol Graph Model*, 20(6):447–462, Jun 2002.

[9] J. Delaney. Assessing the ability of chemical similarity measures to discriminate between active and inactive compounds. *Mol Divers*, 1:217–222, 1996.

[10] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci*, 42(6):1273–1280, 2002.

[11] Z. Feng, L. Chen, H. Maddula, O. Akcan, R. Oughtred, H. M. Berman, and J. Westbrook. Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics*, 20(13):2153–2155, 2004.

[12] M. Füllbeck, X. Huang, R. Dumdey, C. Frömmel, W. Dubiel, and R. Preissner. Novel curcumin- and emodin-related compounds identified by in silico 2D/3D conformer screening induce apoptosis in tumor cells. *BMC Cancer*, 5:97, Aug 2005.

[13] A. Goede, M. Dunkel, N. Mester, C. Frommel, and R. Preissner. SuperDrug: a conformational drug database. *Bioinformatics*, 21(9):1751–1753, 2005.

[14] A. Goede, I. Jaeger, and R. Preissner. Superficial–surface mapping of proteins via structure-based peptide library design. *BMC Bioinformatics*, 6:223, 2005.

[15] M. Hendlich, A. Bergner, J. Günther, and G. Klebe. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J Mol Biol*, 326(2):607–620, 2003.

[16] W.-D. Ihlenfeldt, J. H. Voigt, B. Bienfait, F. Oellien, and M. C. Nicklaus. Enhanced CACTVS browser of the Open NCI Database. *J Chem Inf Comput Sci*, 42(1):46–57, 2002.

[17] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[18] M. Kanehisa, S. Goto, S. Kavashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Bioinformatics*, 32:D277 – D280, 2004. Database issue.

[19] G. Kleywegt and T. Jones. Databases in protein crystallography. *Acta Crystallogr D Biol Crystallogr*, 54:1119–1131, Nov 1998.

[20] R. A. Laskowski, V. V. Chistyakov, and J. M. Thornton. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res*, 33(atabase issue):D266–268, 2005.

[21] W. Li, L. Jaroszewski, and A. Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3):282–283, 2001.

[22] C. Lipinski, F. Lombardo, B. Dominy, and P. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev*, 46(1-3):3–26, Mar 2001.

[23] A. C. R. Martin. PDBSprotEC: a Web-accessible database linking PDB chains to EC numbers via SwissProt. *Bioinformatics*, 20(6):986–988, Apr 2004.

[24] Y. Martin, J. Kofron, and L. Traphagen. Do structurally similar molecules have similar biological activity? *2002*, 45:4350–4358, J Med Chem.

[25] H. Matter. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J Med Chem*, 40(8):1219–1229, Apr 1997.

[26] G. Michal. Biochemical Pathways. Boehringer Mannheim GmbH, 1993.

[27] E. Michalsky, M. Dunkel, A. Goede, and R. Preissner. SuperLigands - a database of ligand structures derived from the Protein Data Bank. *BMC Bioinformatics*, 6(1):122, 2005.

[28] C. A. Orengo, F. M. Pearl, and J. M. Thornton. The cath domain structure database. *Methods Biochem Anal*, 44:249–271, 2003.

[29] N. Paul, E. Kellenberger, G. Bret, P. Müller, and D. Rognan. Recovering the true targets of specific ligands by virtual screening of the protein data bank. *Proteins*, 54(4):671–680, Mar 2004.

[30] R. Preissner, A. Goede, and C. Frömmel. Dictionary of interfaces in proteins (dip). data bank of complementary molecular surface patches. *J Mol Biol*, 280:535–550, 1998.

[31] R. Preissner, A. Goede, K. Rother, F. Osterkamp, U. Koert, and C. Frömmel. Matching organic libraries with protein-substructures. *J Comput Aided Mol Des*, 15(9):811–817, Sep 2001.

[32] J. Reichert and J. Sühnel. The imb jena image library of biological macromolecules: 2002 update. *Nucleic Acids Res*, 30:253–254, 2002.

[33] K. Rother, E. Michalsky, and U. Leser. How well are protein structures annotated in secondary databases? *Proteins*, 60(4):571–576, 2005.

[34] A. Smellie, R. Stanton, R. Henne, and S. Teig. Conformational analysis by intersection: CONAN. *J Comput Chem*, 24(1):10–20, Jan 2003.

[35] A. C. Stuart, V. A. Ilyin, and A. Sali. LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics*, 18(1):200–201, 2002.

[36] M. Thimm, A. Goede, S. Hougardy, and R. Preissner. Comparison of 2D similarity and 3D superposition. Application to searching a conformational drug database. *J Chem Inf Comput Sci*, 44(5):1816–1822, 2004.

[37] S. Trissl, K. Rother, H. Müller, T. Steinke, I. Koch, R. Preissner, C. Frömmel, and U. Leser. Columba: an integrated database of proteins, structures, and annotations. *BMC Bioinformatics*, 6(1):81, 2005.

[38] S. Velankar, P. McNeil, V. Mittard-Runte, A. Suarez, D. Barrell, R. Apweiler, and K. Henrick. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res*, 33(Database issue):D262–265, 2005.

[39] G. Wang and R. L. Dunbrack Jr. PISCES: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, 2003.

[40] D. Wheeler, C. Chappey, A. Lash, D. Leipe, T. Madden, G. Schuler, T. Tatusova, and B. Rapp. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 28(1):10–14, 2000.

[41] WHO. The selection and use of essential medicines. *World Health Organ Tech Rep Ser*, 920:1–127, back cover, 2004.