ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/23443481

The Protein Data Bank (PDB), Its Related Services and Software Tools as Key Components for In Silico Guided Drug Discovery

ARTICLE *in* JOURNAL OF MEDICINAL CHEMISTRY · DECEMBER 2008

Impact Factor: 5.45 · DOI: 10.1021/jm8005977 · Source: PubMed

CITATIONS
59

READS
79

8 AUTHORS, INCLUDING:

Simona Distinto
Università degli studi di Cagliari
46 PUBLICATIONS **1,002** CITATIONS
SEE PROFILE

Daniela Schuster
University of Innsbruck
141 PUBLICATIONS **2,017** CITATIONS
SEE PROFILE

Gudrun M Spitzer
University of Innsbruck
20 PUBLICATIONS **482** CITATIONS
SEE PROFILE

Gerhard Wolber
Freie Universität Berlin
148 PUBLICATIONS **2,863** CITATIONS
SEE PROFILE

Available from: Simona Distinto
Retrieved on: 02 February 2016

Journal of Medicinal Chemistry

© Copyright 2008 by the American Chemical Society

Volume 51, Number 22 November 27, 2008

Perspective

The Protein Data Bank (PDB), Its Related Services and Software Tools as Key Components for In Silico Guided Drug Discovery

Johannes Kirchmair,[†,‡] Patrick Markt,[†] Simona Distinto,[†,§] Daniela Schuster,[†] Gudrun M. Spitzer,[||] Klaus R. Liedl,[||] Thierry Langer,[⊥] and Gerhard Wolber*[,†,‡]

Department of Pharmaceutical Chemistry, Faculty of Chemistry and Pharmacy and Center for Molecular Biosciences (CMBI), University of Innsbruck, Innrain 52, A-6020 Innsbruck, Austria, Inte:Ligand Software-Entwicklungs and Consulting GmbH, Clemens Maria Hofbauer-Gasse 6, A-2344 Maria Enzersdorf, Austria, Dipartimento Farmaco Chimico Tecnologico, Università degli Studi di Cagliari, 09124 Cagliari, Italy, Department of Theoretical Chemistry, Faculty of Chemistry and Pharmacy, University of Innsbruck, Innrain 52a, A-6020, Austria, and Prestwick Chemical Inc., Boulevard Gonthier d'Andernach, 67100 Illkirch, France

Received May 20, 2008

1. Introduction

In 1970, at a time when structural information of one atom was stored on a single punched card, the crystallographers Helen M. Berman, Edgar Meyer, and Gerson Cohen began forming the idea of a comprehensive, public structural data repository for protein structures. Only very few research sites had been exchanging structural data, which was evidently linked with enormous logistics efforts. Thousands of punched cards, each representing data of only a single atom of a protein, were shipped with postal services.[1]

In the autumn of 1971, the Protein Data Bank (PDB[a]) was established as a joint effort of the Brookhaven National Laboratory (BNL) and the Cambridge Crystallographic Data Centre (CCDC). Starting from seven available structures, the data growth rate was very low in its early stage, with only 13 structures available in 1974. In the following years, however, dramatic technological inventions made protein engineering and gene cloning possible. Computer power increased considerably, remote access became established, and software tools for computerized electron-density fitting became available. Therefore, the number of entries in the PDB started to rise significantly in the late 1980s. Not only the number of structures but also the complexity of the determined structures increased. In 1999, the Research Collaboratory of Structural Bioinformatics (RCSB) was established as the consortium managing the PDB. The institution is formed by Rutgers (The State University of New Jersey), the San Diego Supercomputer Center (University of California San Diego), and the National Institute of Standards and Technology.[2] By the turn of the millennium, structural knowledge had reached a point where understanding biology and medical targets on a molecular scale started to become tangible. At this time, the three major data sources, the RCSB PDB (www.pdb.org), the Macromolecular Structure Database at the European Bioinformatics Institute[3,4] (MSD-EBI, www.ebi.ac.uk/msd), and PDB Japan[5] (PDBj, www.pdbj.org) were joined in the worldwide PDB (wwPDB).[6] The latest member in this consortium is the BioMagResBank of the University of Wisconsin—Madison (BMRB, www.bmrb.wisc.edu).[7]

Today, the RCSB PDB Web site is the most important Web portal of the wwPDB. This site itself registers about 100 000 unique visitors per month from all over the world. Data traffic counts around 500 GB per month.[1] While the PDB was used almost exclusively by crystallographers in the beginning, nowadays the community is much more diverse, including scientists in biology, chemistry, cheminformatics, bioinformatics, molecular modeling, and many more. The well-designed Web interfaces and the variety of user-friendly and mostly free 3D viewers enabled this information pool to also find its way to education courses in middle and undergraduate schools. An

* To whom correspondence should be addressed. Phone: +43-512-507-5252. Fax: +43-512-507-5269. E-mail: gerhard.wolber@uibk.ac.at.
[†] Department of Pharmaceutical Chemistry, University of Innsbruck.
[‡] Inte:Ligand Software-Entwicklungs and Consulting GmbH.
[§] Università degli Studi di Cagliari.
[||] Department of Theoretical Chemistry, University of Innsbruck.
[⊥] Prestwick Chemical Inc.

10.1021/jm8005977 CCC: $40.75 © 2008 American Chemical Society
Published on Web 11/01/2008

example of the educational efforts is the "molecule of the month" series where short accounts of over 100 selected molecules, e.g., adrenergic receptors, multidrug resistance transporters, and potassium channels from the PDB, are presented.

In this work we review scope, limits, and latest developments of the world's largest public collection of protein structures, regarding its relevance to drug development in particular. First, we describe composition, coverage, and management of the PDB. Second, we analyze data quality of the PDB, characterizing issues and errors observed, and point out quality benchmarks. Subsequently, we focus on relevant PDB Web services like data mining tools and PDB-derived data sources. For obvious reasons this work does not claim to provide an exhaustive list of all PDB related facilities and databases available (there are approximately 1000 free online databases of interest for structural biologists available; see Galperin[8] and Carugo and Pongor[9] for more information). We think that it is beneficial to render a global view of state-of-the-art technology and to present recent drug discovery success stories based on PDB data. An overview of all tools discussed in this Perspective is given in Table 1. As locations of Web services and databases (URLs) are changing frequently, we provide up-to-date URLs to the tools discussed in this work on http://www.uibk.ac.at/pharmazie/phchem/camd/pdbtools.html.

2. Data Uniformity in the wwPDB

One of the biggest challenges of the wwPDB so far, and also in the foreseeable future, is data uniformity and structure validation. For several years, the RCSB PDB, PDBj, MSD-EBI, and BMRB have been developing tools for data curing and organization individually. With the formation of the wwPDB consortium, these efforts have been pooled in order to ensure data uniformity and also to improve data quality.[6,10] The wwPDB remediation project has been presented recently and includes the following major aspects: (i) improvement of chemical description and nomenclature, (ii) remediation of atom names in polymer chains and distinct chemical definitions of DNA and RNA nucleotides, (iii) revision of sequence and taxonomy inconsistencies, (iv) improvement of virus representation, (v) recheck and update of primary citations, and (vi) further development of file formats.[11] RCSB is the only institution of the wwPDB consortium with write access to the PDB archive and controls the directory structure and contents. All partner sites send the data processed to the RCSB PDB on a weekly basis to be included in the archive. Recently, a new archive accessible via ftp.wwpdb.org has been introduced.

Within the past few years, these strong efforts to increase data uniformity have pushed the quality of PDB structural data files and related metadata considerably. Comments as well as chemical and experimental descriptions have been standardized, the level of metadata detail has been assimilated, and the linking between different data sources has been upgraded.

3. PDB Statistics and Coverage of Structural Classes

Dramatic technological advances in the field of structural biology are causing the number of structures stored in the PDB archive to grow rapidly. Today, the data repository has reached the 50 000 entries mark, about 15% of the stored data originates from NMR analyses.[12] While 2983 structure depositions were reported in 2000, 8128 PDB entries were filed in 2007. Despite this quantitative increase of data, however, the fraction of novel protein structures stagnated in 1995 and was stopped from dropping not before 2003 by the establishment of structural genomics initiatives (see below).

Overington et al.[13] have analyzed the gene-family of targets of all FDA drugs currently approved, finding that rhodopsin-like GPCRs represent by far the most important family targeted by small organic molecules, followed by nuclear receptors and ion channels. Overall, 60% of all current drug targets are located at the cell surface. As a matter of fact, the PDB is biased toward targets that are less-challenging in obtaining crystals, and unfortunately, membrane-bound proteins (including GPCRs, ion channels, multidrug efflux transporters, etc.), which are tremendously important as drug targets and for rational ligand design, are particularly difficult to crystallize.

For this reason, the PDB is struggling with a heavy bias toward globular crystallizable proteins, limiting the field of application and development of structure-based methods for membrane proteins (see below). However, enormous efforts are going on in order to overcome these difficulties, and recently structural biology celebrated a new breakthrough with the structural determination of an engineered human β_2-adrenergic receptor.[14] The Database of Membrane Proteins of Known 3D Structure[15] provides an up-to-date collection of all related structures available in the PDB. As of September 9, 2008, 169 unique proteins (including proteins of same type from different species) are listed in this database.

There is also a lot of space for improvement for nuclear receptor structures (e.g., retinoid X receptor-like, thyroid hormone receptor-like, and estrogen receptor-like proteins), considering their tremendous therapeutic potential. As of September 9, 2008, 214 structures of nuclear receptor binding domains are available in the PDB according to a SCOP classification search.

Enzymes are the most prominent structural family in the PDB. As of July 22, 2008, the Enzyme Structures Database[16] counts

[a] Abbreviations: PDB, Protein Data Bank; BNL, Brookhaven National Laboratory; CCDC, Cambridge Crystallographic Data Centre; RCSB, Research Collaboratory of Structural Bioinformatics; MSD, Macromolecular Structure Database; EBI, European Bioinformatics Institute; PDBj, PDB Japan; wwPDB, worldwide PDB; BMRB, Biological Magnetic Resonance Data Bank; EC, Enzyme Commission; PSI, protein structure initiative; GO, gene ontology; OMIM, online Mendelian inheritance in man; NCBI, National Center for Biotechnology Information; GPCR, G-protein-coupled receptor; TargetDB, Target Database; PepcDB, Protein Expression Purification and Crystallization Database; MSDpisa, MSD protein interfaces, surfaces, and assemblies; SSM, secondary structure matching; ADIT, AutoDep input tool; mmCIF, macromolecular crystallographic information file; PDBML, Protein Data Bank markup language; DUD, Directory of Useful Decoys; SCOP, structural classification of proteins; JCSG, Joint Center for Structural Genomics; CESG, Center for Eukaryotic Structural Genomics; SGC, Structural Genomics Consortium; HPUB, hold for publication; rmsd, root mean square deviation; PQS, protein quaternary structure; EDS, electron density server; RSR, real-space *R*-factor; VADAR, volume, area, dihedral angle reporter; PSVS, protein structure validation software; ED, electron density; sc-PDB, screening PDB; PLD, Protein−Ligand Database; PASS, putative active sites with spheres; PSIbase, Protein Structural Interactome map Database; MOAD, Mother of All Databases; HIC-up, Heterocompound Information Centre−Uppsala; RECOORD, Recalculated Coordinates Database; PiQSi, Protein Quaternary Structure Investigation; PSAP, Protein Structure Analysis Package; PMG, Protein Movie Generator; EzCatDB, Enzyme Catalytic Mechanism Database; PDB-UF, Protein Data Bank Unknown Function; CATH, class, architecture, topology, homologous (superfamily); HSSP, homology derived secondary structure of proteins; BRENDA, Braunschweig Enzyme Database; GOA, gene ontology annotation; SuMo, surfing the molecules; DMAPS, database of multiple alignments for protein structures; SURFACE, surface residues and functions annotated, compared, and evaluated; PAST, Polypeptide Angle Suffix Tree; PLASS, protein−ligand affinity statistical score; CSD, Cambridge Structural Database; SABBAC, structural alphabet based protein backbone builder from α-carbon trace; VS, virtual screening; 17β-HSD1, 17β-hydroxysteroid dehydrogenase 1; PPAR, peroxisome proliferator activated receptor.

24000+ enzyme PDB files: 4000 oxidoreductases, 7000+ transferases, 9500+ hydrolases, 1600+ lyases, 1000+ isomerases, and 700+ ligases. However, Mestres[17] reported in 2005 that the coverage of enzyme families within the PDB archive is not balanced at all: 34.5% of all enzyme entries in the PDB represented the structures of only 34 enzymes (i.e., about 1% of all enzymes characterized with an EC number; see below).

Large molecular complexes (i.e., complexes exceeding 500 kDa) represent a small, yet rapidly growing portion of the PDB (approaching 10% of all PDB entries). The largest groups include viruses, ribosome and ribosomal complexes, large enzyme complexes, chaperonins, and structural protein assemblies.[12]

The most prominent structure families in the PDB archive are structures of potential pharmaceutical relevance. However, novel insights on biochemistry and pathology can quickly turn a so far irrelevant protein into a hot target for drug therapy. Structure-based virtual screening methods in particular suffer from the lack of structural knowledge. Structural biologists became well aware of this problem and started the Protein Structure Initiative (PSI) in 1999. Since this time, structural genomics projects help to lower the redundancy level of the PDB archive by focusing on the determination of novel structural entities.[1,18] PSI aims at lowering costs and time needed for protein structure elucidation, increasing throughput, and improving structural diversity. The high efficiency of these research facilities will allow the Protein Structure Initiative to become the most important source for novel structural knowledge on proteins in future. TargetDB (and its extension, the Protein Expression Purification and Crystallization Database, PepcDB),[19] hosted by the RCSB PDB, provides information on the current status of solutions of structures in production of a multitude of structural genomics projects. As of September 9, 2008, TargetDB reports the deposition of 6098 PDB structures by worldwide structural genomics projects (half of them originating from PSI centers). Latest information on structural genomics projects is provided at the RCSB PDB information portal for structural genomics[20] and the PSI Web site.[21]

One issue of these protein structure initiatives is that frequently not very well understood proteins are solved and deposited in the PDB without exact annotation. For example, one can find the "crystal structure of an unknown protein from Galdieria sulfuraria" (PDB entry 2nyi) or the "protein of unknown function (DUF946) from *Bacillus stearothermophilus*" (PDB entry 2oeq), and many more. The novelty of folds can be automatically predicted using the ProTarget[22] Web service by providing an amino acid sequence query: Reference structures pooled from the Swiss-Prot[23] and the PDB databases are automatically assessed in terms of similarity to the query.

While the current coverage of targets in the PDB may be adequate for classic molecular modeling techniques of a particular target of interest, the lacking global image of all structures of biological relevance is problematic for the development and validation of structure-based methods that attempt to assess such activity spectra. Several structure-based approaches and algorithms for multitarget screening have been proposed in recent years; however, so far the performance of these methods has been examined only on structurally well-characterized targets and target families most of the time. Examples include protein−ligand docking approaches[24−31] and pharmacophore-based parallel screening.[32−34]

4. Services Provided by the wwPDB Partner Sites

The individual Web sites of the wwPDB partners offer different online platforms for basic and advanced searching and data browsing. Besides a plurality of search functions and browsing features, the RCSB PDB offers interconnection to external data pools (e.g., Gene Ontology (GO), Enzyme Commission (EC), Online Mendelian Inheritance in Man (OMIM), National Center for Biotechnology Information (NCBI) repositories, etc.) as well as a user-friendly overview and summary page for all PDB entries and yearly snapshots of the archive. Furthermore, the RCSB PDB supports inspection of PDB structures using several different 3D viewers. PDBj services include the alignment of structural homologues,[35] visualization tools (e.g., visualization of the protein surface), and subsets derived from PDB data. The BMRB provides the NMR restraints grid, which contains the original NMR data collected for over 2500 protein and nucleic acid structures with corresponding PDB entries.

The MSD-EBI offers several search tools: MSDlite is an easy to use Web interface and supports basic queries such as author name, ligand name, and keywords but also to search for cross-references, e.g., EC numbers, NCBI taxonomy, and Swiss-Prot IDs. Moreover, FASTA[36] similarity searches can be processed. MSDpro is a Java-based query editor for complex searches on a range of subjects. The user defines questions with a drag and drop system; the search results can be directly exported in text and XML file formats. MSDsite[37] represents an advanced search and statistical analysis tool for small 3D motifs. Results can be provided as charts, tables, sequence multiple alignment and 3D multiple alignment of fragments, motifs, and protein chains. MSDtemplate supports the investigation of local residue interactions in the PDB archive. Again, all results can be inspected as a 3D multiple alignment. MSDpisa[38,39] (protein interfaces, surfaces, and assemblies) is an interactive service for the investigation of macromolecular (protein−protein, DNA/RNA, and protein−ligand) interfaces. The software supports similarity-based searches and allows the prediction of the most probable multimeric state of the protein. MSDchem is a mining tool focusing on small organic molecules that are stored in complex with proteins in the PDB archive. Search options include different molecular identifiers and SMILES input. MSDfold or secondary structure matching (SSM) is an interactive service for processing 3D similarity analyses. It supports pairwise as well as multiple comparison, 3D alignment of protein structures, and the download and visualization of the best-superimposed structures.[40] MSDanalysis provides validation and analysis tools for PDB data. Thereby, structures can be analyzed directly from the PDB archive as well as via upload of PDB files. Results are depicted as interactive histograms. The user picks residues of interest and examines the corresponding 3D structure with the AstexViewer.[41] MSDmine[42] is a data-mining tool for advanced users to generate complex PDB queries.

5. Structure Deposition and File Formats

PDB deposition tools are probably the most important pillars of the PDB data repository system. These tools are developed to fulfill two highly important aspects of data acquisition: (i) offering a user-friendly and well-defined upload interface and (ii) data processing and checking. The deposition portals underlie a continuous development process. Therefore, focus is on the design of powerful algorithms for error detection and data curation. Structural information can be deposited on all platforms of the wwPDB partners. RCSB PDB and PDBj use the AutoDep Input Tool (ADIT[2,43]), MSD-EBI AutoDep, and BMRB uses

Table 1. PDB Data Repositories, Second-Party Resources, Web Services, and Software Tools Discussed in This Work[a]

wwPDB Portals and Partner Sites	
RCSB PDB	RCSB PDB portal
PDBj	PDBj portal
BMRB	BMRB portal
MSD-EBI	MSD-EBI portal
TargetDB	target search for structural genomics
PepcDB	current status of solutions of structures in production
Cross-Linked Datasources	
Gene Ontology	controlled vocabulary to describe gene and gene product attributes in any organism
ENZYME	repository of information on the nomenclature of enzymes
OMIM	compendium of human genes and genetic phenotypes
SCOP	hierarchical classification of proteins
UNIPROT	comprehensive protein sequence repository
Protein Visualization Tools	
AstexViewer	protein structure viewer
Chimera	protein visualization, investigation of electron density maps
Coot	investigation of electron density maps, model building
LIGPLOT	2D diagrams of protein−ligand interactions
RASMOL	protein structure viewer
Tools for PDB Data Validation and Manipulation	
PQS	analyze and predict quaternary protein structures
WHAT_CHECK	tool for protein structure checks
PDBREPORT	repository of WHAT_CHECK reports
PROCHECK	tool for protein structure checks
PROCHECK-NMR	tool for protein structure checks
MolProbity	tool for protein structure checks
NQ Flipper	erroneous Asn and Gln rotamer detection
PSVS	protein structure validation software suite
Modeling Suites Featuring Ligand Structure Interpretation and Correction Algorithms	
MOE	comprehensive modeling package
SYBYL	comprehensive modeling package
LigandScout	modeling suite for pharmacophore modeling and virtual screening
Tools for Visualizing and Manipulating Electron Density Maps	
SFCHECK	tool for structure factor file checks
EDS	Electron Density Server, repository of electron density maps and collection of structure validation tools
Chimera	protein visualization, investigation of electron density maps
Coot	investigation of electron density maps, model building
AFITT	real-space fitting of ligands in protein−ligand complexes
Protein Interface, Binding Site, and Cavity Analyses	
PASS	identification of protein cavities
SURFNET	identification of protein cavities
Q-SiteFinder	identification of protein cavities
InSite	identification of protein−protein interaction sites
Protemot	prediction of protein binding sites with automatically extracted geometrical templates
SitesBase	comparative investigation of protein−ligand binding sites
PDBSITE	protein binding site analysis
PSIbase	repository of PDB-derived protein interface information
PROTCOM	data pool of protein interface structures
Molsurfer	analysis and visualization of protein interfaces
iPFAM	visualization and browsing of protein interfaces
DMAPS	database of multiple alignments for protein structures
Binding Data Collections	
Binding DB	affinity data collection
PDBbind	affinity data collection
AffinDB	affinity data collection
KiBank	affinity data collection
Binding MOAD	affinity data collection
Ligand Collections Based on PDB Structures	
SuperLigands	ligand structure repository
PDB-Ligand	ligand structure repository
Ligand Depot	ligand structure repository
HIC-up	ligand structure repository
PDB-Derived Subsets and Recalculated Data Repositores	
sc-PDB	PDB subset of structures prepared for virtual screening
LigBase	ligand-binding proteins aligned to structural templates
PLD	protein−ligand database

Table 1. Continued

PDB-REPRDB	representative protein chains from PDB
RECOORD	database of 500+ recalculated NMR structures
Database of Membrane Proteins of Known 3D Structure	collection of PDB data on membrane proteins
PiQSi	database for comparing the quarternary structure of protein complexes
iMolTalk	Web toolkit for analyzing and searching PDB complexes
PSAP	protein structure analysis package
PDBsum	at-a-glance overviews on PDB structures, cross-linking to second-party databases
EZCatDB	database of enzyme catalytic mechanisms
Data Mining and Analysis Tools	
PDB-UF	prediction of enzymatic functions of not-annotated PDB entries
pKNOT	analyses tools and structural data on knotted proteins
PDBSprotEC	links PDB chains with Swiss-Prot codes and EC numbers
SuMo	analysis and comparison of protein binding sites
Relibase+	analysis and data mining interface for protein structures
pdbFun	collection of PDB data mining tools
SURFACE	surface-based structural comparison and similarity assessment of protein structures
PAST	fast protein structure search
3dLogo	identification of conserved residues in a set of structurally superimposed proteins
FeatureMap3D	align query sequences to PDB structures
PISCES	PDB sequence culling
MMsINC	ligand structure and substructure search
DBAli	database of structure alignments
MaxSprout	generating protein backbone and side chain coordinates from a C(alpha) trace
SABBAC	reconstruction of protein backbones and amino acid side chains based on a C(alpha) trace
Miscellaneous Tools	
Columba	online service for the selection of PDB subsets
PDB Goodies	Web-based manipulation of PDB files
PMG	online movie generator for PDB structures
ZINC	free database of commercially available compounds for virtual screening
DUD	directory of useful decoys and known active compounds

[a] An up-to-date link list to all resources is available from http://www.uibk.ac.at/pharmazie/phchem/camd/pdbtools.html.

ADIT-NMR, respectively. All uploaded data are checked for geometric accuracy, chemistry of both proteins and ligands, nomenclature, and the likely biological assembly.[1] A short overview on the importance of well-defined structure deposition procedures and structure validation tools is provided by Josten and Vriend.[44]

The PDB data file format has been developed in 1976, based on 80 column punched cards. PDB files contain a header section including details on the PDB ID number, target, resolution, authors, citation, sequences of molecules in the crystal, secondary structure information. This section is followed by the atom site records, which contain information on the atom coordinates, atom names and numbers, and several more identifiers. PDB files can be easily edited with any text editor (e.g., EMACS[45]); there is also a large number of Python/Perl scripts and programs available for data manipulation.[46] PDB Goodies provides a Web-based manipulation interface for PDB files.[47]

The PDB file format has been updated continuously in order to fulfill the needs of novel applications. However, with the dramatic increase in size and complexity of resolved protein structures the PDB file format is approaching its limits in terms of storage capacity and data handling.[1,12] In 1997, Bourne et al.[48] introduced a new file format in order to overcome these limitations. The so-called macromolecular crystallographic information file (mmCIF) is an advanced data file format that considers all aspects of structural characteristics and metadata.[49] In 2005, Westbrook et al.[50] introduced the PDBML (PDB markup language) format, an XML format derivative. Both the mmCIF and PDBML file formats are advantageous in particular for data mining campaigns. For more information on deposition tools, methods (including validation), and policies see the recent publication by Dutta et al.[51] Issues relating to NMR depositions are discussed by Markley et al.[7]

6. Data Quality

6.1. Data Annotation, PDB Subsets, and Bias Induced by Data Sets. The interconnection of the PDB with second party databases raises the importance and data mining possibilities considerably, as it allows for elucidating PDB subsets considering different aspects. An example of a PDB subset is a collection of high-resolution protein structures of targets that are responsible for drug side effects of human and closely related species. The Columba[52] server is one of several services that allow users to quickly build up PDB subsets considering various parameters. Columba consists of PDB data connected to 12 second-party databases.

In today's literature introducing novel computational methods for drug discovery, besides the methodical part, in general direct comparisons with related methods are presented. Issues arise if the performance of a novel method in combination with an up-to-date test set (e.g., PDB subsets) is directly compared to well-established methods that have been developed and evaluated using seasoned data sets. Conclusions about a better performance of novel methods in such studies should be considered with care, as a good rating of the presented novel method does not necessarily indicate better global performance. The *Journal of Computer-Aided Molecular* design has published a special issue on the evaluation of computational methods recently.[53]

The DUD database is a collection of known active compounds and decoys for targets included in the PDB for the evaluation of docking programs.[54] Currently the DUD is considered the

de facto industry standard for the evaluation of docking algorithms. As a matter of fact, this data set is biased on the availability of known active compounds and PDB data on the targets. As docking algorithms and knowledge-based scoring functions in particular are commonly derived and tuned using structure-based data of prominent targets, the validation and drawing of conclusions on the global performance of these methods based on closely related test sets are critical.

Rother et al.[55] have investigated the extent to which protein structures are annotated in 15 secondary databases. They found a high overlap between PDB entries that have been deposited before 1997. However, during the past decade the level of annotation decreased for several secondary databases, especially for recently released structures. This annotation gap could easily lead to heavily biased results of data mining campaigns, as insufficiently annotated PDB entries may be disregarded. It is likely that this bias suppresses peptides, non-proteins, nonstandard structures, and also recently published structures.

Such shortcomings in annotation can be overcome by manual intervention for a small selection of targets and structures; however, problems arise for multitarget data mining approaches, which attempt to characterize the activity profile of compounds against a plethora of targets. While a classical virtual screening approach reports a simple rank-ordered list of putative active compounds, in the case of multitarget screening, multidimensional data matrices are obtained. For a sensible and meaningful elaboration of data, relationships and dependencies between hits, models, and targets need to be considered: for example, structural relationships represented by SCOP classification, Swiss-Prot, and UniProt IDs and functional relationships considered by GO and EC classification, pathway affiliations, relationship of diseases, and pharmacological target types (such as ADME or toxicity related targets). PDB entries lacking certain annotation data may therefore be missed or rejected during such multitarget screening campaigns, leading to a significant bias. Moreover, the validation of multitarget screening methods is likely to be affected by such annotation issues, since the selection of structures considered for profiling (based on the availability of metadata) determines the outcome.

6.2. Structural Data Quality and Accuracy. State-of-the-art technology for structure determination and refinement allows structural biologists to analyze structures that have not been accessible until recently. Several structural genomics projects have been introduced in order to enrich structural biology data. However, the data quality of recently published structures is not necessarily superior in terms of structure quality compared to older data. Brown et al.[56] performed an in-depth analysis of PDB structures and compared the most prominent structural proteomics projects to each other. The Joint Center for Structural Genomics (JCSG), the Center for Eukaryotic Structural Genomics (CESG), and the Structural Genomics Consortium (SGC) were determined to produce the highest quality among all projects. Not a long time ago, the publication of a protein structure was a major event for the scientific community and considerable efforts were taken to check the structure during peer review. Nowadays, "mass production" of protein structures inevitably drops awareness and thereby also structure validation. Now, publishing of protein structures in the PDB is commonly obligatory for peer-review journals. Statistics on the number of publications per journal are available at the PDBsum Web site (see below).

The PDB itself is taking and will take also in future considerable efforts to increase data accuracy and transparency.[1] Since 2006 there are no theoretical models accepted anymore,[10] and as of February 2008, structure factor amplitudes/intensities for crystal structures and restraints for NMR structures are required for structure deposition in the PDB. The hold for publication (HPUB) policy includes that the citation for a structure deposited is to be published within 1 year after deposition.

There is high interest in the scientific community about the advantages and disadvantages of structural data derived from X-ray crystallography and NMR. Generally speaking, X-ray data seem to be preferred as a starting point for molecular modeling. The benefit of both methods, however, originates from their complementarity, supplementing gaps of the partner method. NMR structures are usually published and deposited at the PDB as an ensemble of models with different conformations (in general about two dozen but sometimes significantly more). To solve the problem of the availability of several protein conformations as a starting material, a single, representative, average conformation (energy-minimized) is chosen as a starting point for molecular modeling studies. While the resolving power and accuracy of NMR are considered inferior to those of X-ray (though there is rapid technological progress in NMR technologies to increase precision and there are no well-established measures for NMR structures available that allow for defining and comparing accuracy), solubility effects can be explored using NMR. X-ray allows for structural determination of very high molecular mass proteins; NMR is favored for peptides and protein segments. On the other hand, NMR allows for motion analyses of domains and investigations on chemical kinetics and is not dependent on the availability of the crystallized protein.

Andrec et al.[57] recently published a large-scale study on differences between structures determined by X-ray and NMR. Therefore, they selected a set of 148 structure pairs of which structural models derived from both X-ray and NMR experimental data are available and statistically analyzed the differences between these pairs of models using the FindCore[58] method for structural superposition. The authors found that the root-mean-square deviation (rmsd) between the crystal structure and the average NMR structure exceeds the rmsd within the NMR ensemble. Moreover, in 73 of all 148 structure pairs the core heavy atoms were diagnosed to be located at significantly different positions. The authors point out several possible reasons for this problematic observation. Steric interactions, salt bridges, hydrogen bonds, and other interactions formed in the crystalline state can occasionally alternate protein structures observed by X-ray crystallography. In this way it seems possible that the crystalline environment might stabilize a certain protein conformation that is unfavorable in solution. Flexible protein residues may be buried or relatively rigid due to crystal packing, which can only be checked if crystal symmetry parameters are available. Another way to explain the structural differences would be problematic statistical measures of similarity and different, continuously further developed refinement methods of both approaches. Several studies reflect the high interest on determining the reasons for structural differences between X-ray and NMR data,[59−65] and there is a strong need for further investigation of these structural discrepancies and their offspring(s).

The nature of errors found in PDB records is quite manifold; it reaches from administrative errors like wrong residue or chain names and wrong atom nomenclature to complex but less common structural errors like wrong protein topology. Wrong bond lengths, wrong bond angle values, out-of-plane issues (e.g., in aromatic ring systems), and chirality mismatches are among

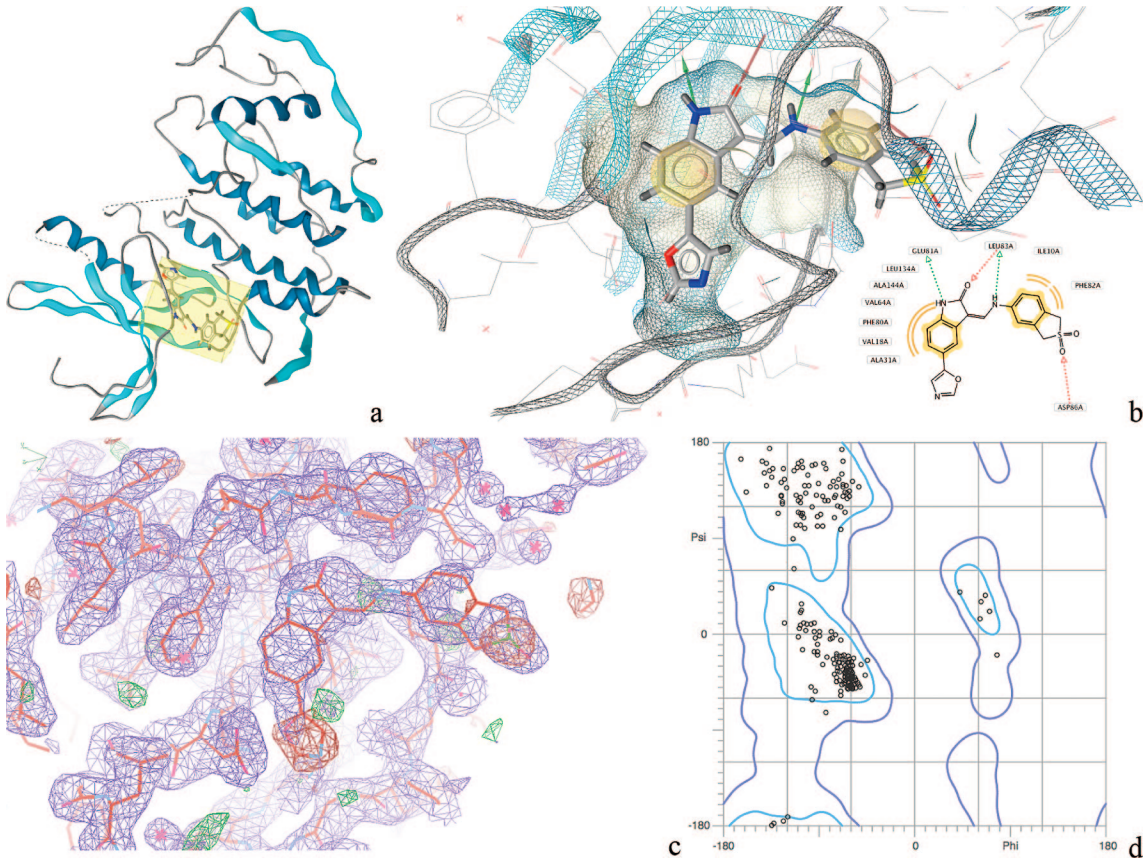

Figure 1. Analysis of PDB entry 1ke7, CDK2 in complex with an oxindole-based inhibitor: (a) LigandScout visualization of the protein structure with the binding site and the ligand; (b) 3D and 2D depiction of the protein−ligand interactions in LigandScout, with hydrogen bond donors (green vectors), hydrogen bond acceptors (red vectors), and hydrophobic areas (yellow spheres); (c) electron density maps visualized in Coot, with the $2F_o - F_c$ map (magenta), the positive density of the $F_o - F_c$ map (i.e., parts of the electron density not represented in the model, in green color), and the negative density (i.e., parts of the model that are not backed up by electron density, in red color); (d) Ramachandran plot of 1ke7 generated with MolProbity, with 98% of all φ/ψ values located in favorable areas; Pro254 is the only amino acid reported as outlier.

the errors that are easily detectable and curable. Steric clashes (bumps) are quite frequently observed in PDB structures, but these can also be detected easily. There is also a variety of tools available for highlighting and curing wrong atom types and omitted atoms or side chains.

Major issues in the PDB are entries with incomplete information about the quaternary structure of the target. Despite the problem that the biological quaternary structure may be unknown, PDB structure showing a protein monomer may in reality be a dimer, trimer, tetramer, or multimer. The reason for missing information on the quaternary structure may be the nonexistence of the biological quaternary structure in the crystal. As X-ray structures represent only information on the asymmetric unit of the crystal, however, the complete quaternary structure may not be published even if this structure exists in the crystal. Symmetry records, such as CRYST1, MTRIX, and SCALE[66] characterize crystallographic properties and are particularly important for drug design, as crystal packing may influence symmetry-derived structures. Computational chemists should be well aware of structural artifacts of symmetry-derived structures that influence, for example, the conformation of the protein binding site. There are algorithms available for the calculation of symmetry records, as they may be unavailable. Web facilities like the PQS[67] (protein quaternary structure) server perform analyses of PDB structures and attempt to predict the most probable quaternary structure of proteins.

Torsion angle evaluation compares the torsion angles of each residue to "normal" values and identifies issues on a statistical basis. Special attention is thereby taken on the protein backbone torsion angles, which can be assessed using Ramachandran plots (φ/ψ plots).[68] This protein verification approach is one of the earliest applied to protein structures, based on the observation that the degree of freedom of rotatable bonds of the protein backbone is rather small, as the side chains consume most of the possible torsions. Two major areas of favored φ/ψ values are found in Ramachandran plots: one for α helix like torsion angles and one for β-strand like torsion angles. Besides, rather few residues show individual, outlying φ/ψ values (e.g., residues in loop regions). Therefore, protein Ramachandran plots should detect only a few torsion values that are separated from the favorable areas (Figure 1d). An accumulation of such outliers is a strong indication for structural issues.

Nine types of amino acid side chains have a planar moiety: Asp, Glu, Phe, His, Asn, Gln, Arg, Trp, and Tyr.[69] Planarity issues are not uncommon in PDB in particular not for certain ring systems on the ligand side. However, they are easy to detect and cure.

The protein sequence of the PDB data may deviate from the respective sequence provided by the universal protein resource (UniProt):[70] Residues that cannot be defined correctly by the electron density (ED) map are sometimes renamed to Ala during model generation. Therefore, it is highly recommended to directly compare the sequence reported by UniProt with the PDB data. This can be quickly checked on the RCSB PDB site (Figure 2a). Sequence register errors occur if residues are placed into the ED of the consecutive residue. These issues can be found

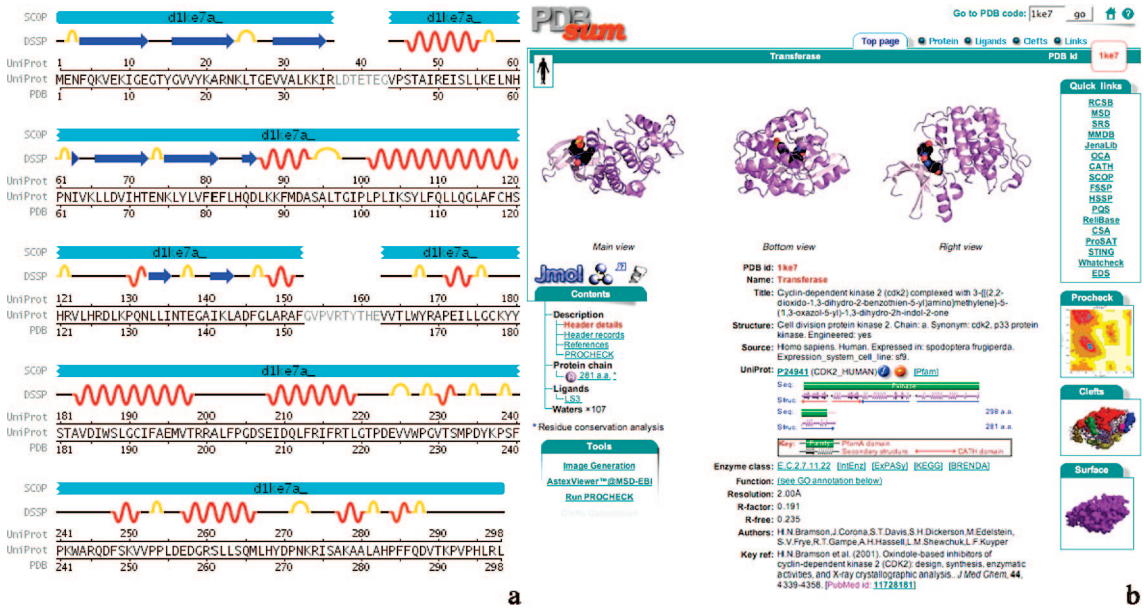

Figure 2. (a) Sequence details for 1ke7 showing the resolved parts of the protein in direct comparison to the UniProt sequence. The diagram offers a quick overview of the agreement of the structural data with the UniProt protein sequence as well as of the secondary structure elements of the protein. (b) PDBsum top page showing an overview of characteristics of 1ke7. The site is especially useful as a starting point for investigations, as there are several Web services directly connected via the quick links panel on the right.

particularly at loops of low-resolution models. Detection is feasible but only if the experimental data are available.[71]

The majority of water molecules can be correctly detected by ED analyses (see section below). However, the placement of waters can be considered as being subjective to certain extend.[72] Adding waters extensively to protein structures lowers the *R*-factor and may therefore bias this quality benchmark. Moreover, the discrimination between water molecules and ions like Na^+ is nontrivial. Therefore, ions are frequently exchanged by water molecules. For protein−ligand docking it is common to delete all water molecules before docking except for waters that are known to be tightly bound to the protein and that are important for mediating the ligand binding.

Hydrogens are only visible in ultrahigh-resolution X-ray structures. However, they are of exceptional importance for the formation of hydrogen bonds and thus also for the proper characterization of protein−ligand complexes. High-resolution structures show only a very few unsaturated hydrogen bonds, while the degree of unsaturated hydrogen bonds is much higher in models of low resolution. Moreover, correct hydrogen placement is essential for the appropriate structure allocation of Gln, Asn, and His. WHAT_CHECK[73] supports checking the hydrogen bond network health (see below). Adjusting and optimizing the hydrogen bond network of macromolecular structures are computationally demanding, yet highly recommended before starting any modeling activity.[74]

Ligands frequently suffer from lacking attention of crystallographers on small organic molecules, as quality benchmarks are usually for the global structure and the global model is minimized according to these values, which is a severe problem for modelers interested in the analysis of protein−ligand interactions. The lacking focus on the ligand structure is also reflected by the insufficient atom type definitions in PDB file format: PDB data do not contain any information on the hybridization states and connectivity of ligand atoms but only about atom coordinates. Therefore, several software tools have been developed in order to overcome this data lack. MOE[75] and SYBYL[76] are examples of two well-established modeling suites featuring ligand structure interpretation and correction algorithms. LigandScout[77,78] is modeling suite for both ligand- and structure-based pharmacophore modeling and virtual screening that supports extracting relevant information on the respective binding mode, pharmacophore modeling, and virtual screening. Existing ligand interpretation algorithms were adopted and new strategies were developed to deduce ligand topology adequately in LigandScout. A step-by-step interpretation is performed on the PDB ligand entries: planar ring detection, assignment of functional group patterns, hybridization state determination, and Kekulé pattern assignment (Figure 1a,b).

Over the years, severe failures have become uncommon in the PDB (yet there are still some problematic issues[44]), and also the occurrence of geometry-related errors has been decreasing since the early 1990s. Despite this, Badger and Hendle[79] found that about 3% of all amino acids of structures deposited in the PDB are modeled incorrectly. Asn, Gln, and His side chain flips were reported as the most frequent errors. Another study reported an estimated percentage of severe structure issues in protein chains of up to 1%.[80] However, both values need to be interpreted in context of the authors' definitions of structural errors of course, since computational approaches differ considerably in their sensibility for structural insufficiencies.

6.3. Electron Density Maps and Their Utility in Molecular Modeling. Inspection of ED maps is the best way for both experts and novice users to get to the bottom of X-ray crystallographic models, to learn about their quality and characteristics, and to understand and critically evaluate literature. The models published in the PDB are heavily influenced by modeling procedures and the individual experience, expertise, knowledge, and possible mistakes of a crystallographer. ED maps represent crystallographic experiments without this expert bias that is encountered in atomic models and therefore retain information that cannot be included in a model. In this way, ED maps are crucial for the comprehension, retracing, and validation of models. Even more, ED maps are highly useful for selecting a high quality structure out of a set of PDB structures as a starting point for molecular modeling (Figure 3).

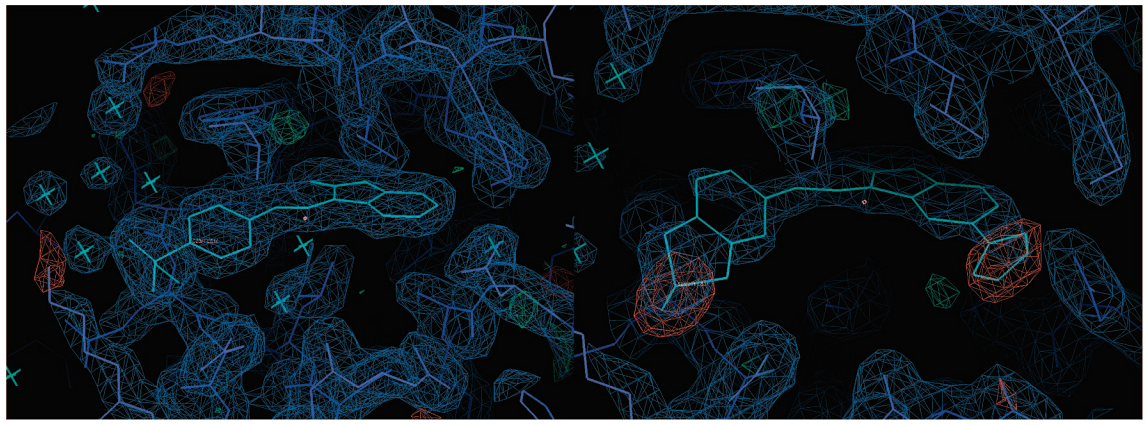

Figure 3. Comparison of two PDB structures of inhibitors cocrystallized with CDK2: 1ke5 (left) and 1ke7 (right); $2F_o - F_c$ map (blue color); positive density of the $F_o - F_c$ map (green color); negative density of the $F_o - F_c$ map (red color). The ED plot demonstrates the well-defined location of the inhibitor in 1ke5, as all parts of the inhibitor model do fit very well into the electron density. In the case of 1ke7 some uncertainties can be detected, as both terminal areas of the ligand do have negative density, indicating parts of the molecule that could not be observed in the experimental data. These deviations can be identified by calculation of the RSR value, which is lower for the ligand in 1ke5 (0.099) than for the ligand in 1ke7 (0.345). In this perspective both electron density maps and RSR values are of particular importance for molecular modeling studies, as these data allow estimation of the quality of a model at the residue level.

Densities can only be calculated if structure factors and the model are available. The portion of structures published in the PDB including structure factor information has been increased steadily, and since February 2008 the submission of structure factors is a mandatory requirement for PDB deposition, as already mentioned above. However, this does not solve the problem of missing structure factors for earlier submitted PDB structures. And even though the deposition of structure factors is now required, some issues remain, as the standards used upon publication are not well defined and there are different file formats for storing structure factors available.[81] The PDB aims at the standardization of structure factor files and hosts SF-Tool, a program for the validation and conversion of structure factor files for deposition. The tool is based on SFCHECK[82] and SFCONVERT.

The EDS[81] (electron density server) of the University of Uppsala calculates the ED maps from the coordinate and structure factor files deposited at the PDB. The EDS archive is updated on a regular basis. ED maps can be visualized directly at the EDS site with the AstexViewer, and the Uppsala viewer; ED maps can be downloaded directly from the server. Moreover, the facility also offers several statistical measures for PDB data, such as Ramachandran plots. Real-space *R* values (see section 6.4) plots as a function of residue numbers can be explored interactively in combination with the AstexViewer.

Chimera[83] is a sophisticated tool for the visualization of proteins and large-scale molecular assemblies. Besides modules for multiple protein alignment functions, investigation of ligand docking poses and molecular volumes, and tools for the generation of movies from conformational changes and molecular dynamics trajectories, the software allows for visualizing ED maps in context of the model. Coot[84] (Figure 1c) is one of the most prominent tools for X-ray crystallographic model building and ED maps visualization. ED maps can be directly downloaded from the EDS using the Coot graphical user interface. AFITT[85] is a software tool for real-space fitting of ligands in protein−ligand complexes. The program allows for placing the ligand in an optimized position into the ED while considering conformational strain energy. The GUI enables users to interactively match the ligand to the density and to refine results, and the command line interface offers integration to automated workflows.

6.4. Benchmarks for X-ray Crystallographic and NMR Data. While there are several well-defined benchmarks for the definition of data precision available for X-ray crystallography, standardized precision values for NMR structures are still under development because of a lack of widely accepted algorithms and protocols.[58] Moreover, NMR data preparation procedures offer more options for differing interpretation.

The goodness of models derived from X-ray crystallographic and NMR data is limited to the resolution: The higher the resolution, the more structural characteristics can be deduced. Electron densities of less than 5 Å resolution provide sufficient information on the overall shape properties of a protein; at a resolution of 3 Å, side chains are detectable. Structures with 2.5 Å resolution show defined conformations of amino acid side chains; the characteristic ED gap in aromatic ring systems can be observed at about 1.5 Å. Highest resolutions currently reported are located around 0.6 Å.

One of the most important benchmarks for X-ray crystallographic data quality is the *R*-factor, which is defined as the relative deviation of the calculated structure factors from those experimentally observed. In general, a lower *R*-factor stands for a better model. The expressiveness of this measure suffers from possible overfitting, which cannot be detected by the *R*-factor. R_{free}[86] is a measure for the deviations of the calculated model, as it applies to a smaller testing data set that has not been used for the generation of the structural model. The difference between R_{free} and *R*-factor is useful for the detection of overfitting of the 3D structure; the smaller the difference is between both measures, the better. However, both values are global measures and do not provide information on residue or atom level.

The real-space *R*-factor (RSR)[87] represents a local benchmark for the fit of data at residue level (i.e., for one amino acid residue, nucleotide, or small molecule at a time); it compares the calculated density of a residue with the experimentally obtained density data. Lower RSR values indicate better fit. The RSR can also be represented as a correlation coefficient, where values approaching 1 imply better fit. Both quality measures can be directly investigated on the EDS Web site (Figures 3 and 4).

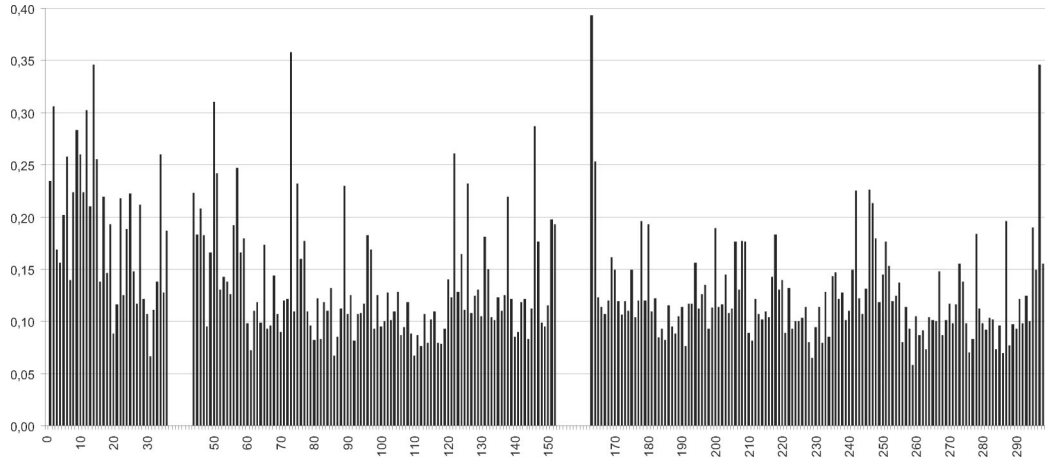

Figure 4. Representation of the RSR values of 1ke7 on a residue basis (residue number on *x*-axis, RSR value on *y*-axis) for all resolved amino acids. High RSR values indicate low fit of the model and the electron density map. RSR values are of particular importance for computational studies, as they allow for analysis of the goodness of models and for identification of possible insufficiencies efficiently. Looking at proximate RSR values of a certain residue of interest allows the deduction of conclusions about the reliability of the model in areas of particular interest. Also, the quality of the fit of a bound ligand can be characterized by a RSR value and therefore provides an estimation of the suitability of a ligand conformation and location for structure-based molecular modeling.

The *B*-factors reflect the mobility or flexibility of various parts of the protein: The higher the *B*-factors are, the greater is the uncertainty about the actual atom position. As a guideline, disorders are likely if the *B*-factors are 60 or larger; *B*-factors larger than 40 indicate possible structural uncertainties. For more details on these benchmarks the reader is referred to two excellent book chapters.[71,74]

6.5. Software Tools and Web Resources for Structure Checking and Data Curation. WHAT_CHECK[73] is one of the most prominent PDB data validation programs. The program checks for administrative problems such as atom naming issues and ambiguous residue numbering and generates Ramachandran plots. Moreover, WHAT_CHECK investigates protein structures for missing or unexpected atoms and coordinate problems. *B*-factor plots allow for investigating the certainty of atom positions. Geometric checks include bond lengths and angles, chirality mismatches, planarity problems, torsion angle issues, steric bumps, and structural issues based on crystal packing effects. It is important to state that the structural issues reported here are not necessarily errors. The issues detected by WHAT_CHECK are based on statistical means, and some of the anomalies may be caused by genuine influences. Users should check putative errors carefully. The WHAT_CHECK reports of PDB structures can be directly accessed from the PDBREPORT database.[73] PROCHECK[88] enjoys great popularity in the scientific community. The program supports an intuitive and user-adaptable visualization of the results (e.g., Ramachandran plots, histograms, and frequency distributions). PROCHECK-NMR[89] is the NMR counterpart of PROCHECK, providing valuable checking tools for NMR structures. MolProbity[90] checks X-ray and NMR data of PDB entries and uploaded user data. The Web service supports the addition of hydrogens to PDB data featuring H-bond network optimization and Asn, Gln, or His flipping. Structure validation tools include bump detection, Ramachandran plots, rotamer and geometry evaluations, etc. Results can be analyzed with interactive diagrams and interactive 3D views. Also MSDanalysis offers several tools for structure checking, as already mentioned above. Recently, NQ Flipper[91] has been published by Eichenberger and Sippl. This Web facility automatically detects erroneous Asn and Gln rotamers based on mean force potentials. Users can directly upload PDB files for online data curing. Further examples include VADAR[92] (volume, area, dihedral angle reporter), and Verify3D.[93,94]

The JCSG Structure Validation Central[21] offers a structure validation system that integrates seven structure checking tools, including PROCHECK, SFCHECK, and WHAT_CHECK. In a similar way, Bhattacharya[95] et al. have incorporated several protein structure evaluation tools, including PROCHECK and MolProbity into an integrated benchmarking environment to the Protein Structure Validation Software (PSVS) suite. The program collection aims at the automated and standardized validation of protein structures determined by structural genomics consortia.

Chimera, Coot, and AFITT, all free for academics, have been introduced in section 6.3 already.

6.6. General Guidelines for Molecular Modeling Using PDB Structural Data. By the consideration of a few guidelines, most typical issues arising from the usage of PDB structural information for molecular modeling can be avoided. The investigation of the primary literature of a PDB structure is of fundamental importance in order to be able to correctly interpret structural data. Besides a detailed description on the methods used for structure determination, literature provides crucial information about the characteristics of the target structure, comments on the quality of the ED representation, the fit of residues into the ED, residue flexibility, crystal packing effects, and related issues. The validity of the ligand structure needs to be confirmed manually, as correct bond orders and hybridization states of the ligand are not stored in PDB files. From our own experience we find that in approximately 80% of all cases the correct ligand structure can be derived from the atom coordinates stored in the PDB files in an automated way; however, in about one-fifth of all ligand structures issues (e.g., wrong bond types) need to be cleared manually. Examination of published errata and studies citing the primary literature should be included into a standard modeling workflow to ensure that the structure available is state-of-the-art and that no problems with the PDB structures have been encountered so far. Details on the sequence and portion of protein residues that have been structurally determined are decisive for the suitability of structural data for modeling. Targets with a particular site of interest (usually interfaces) may have been structurally resolved only in the proximity of this area (e.g., the active domain); thereby spacious

conformational changes that may be induced by distant protein segments missing in the structural model are often neglected. Mutations may have been applied in order to enable crystallization or to stabilize or force the protein into a certain conformation of interest. Again, this influence also needs to be considered carefully, including also the impact of distant mutations on the local protein conformation.

Once the global suitability of a certain PDB structure is confirmed, quality indicators need to be checked in order to estimate the goodness of a model. Resolution, *R*-factor, and R_{free} are directly available via the PDB. Moreover, most modeling programs allow the visualization of *B*-factors, if available in the structural data. WHAT_CHECK or related programs quickly provide an overview of possible structural issues. Supposedly the most powerful approach to grasp and investigate the quality (i.e., the accuracy of the data but even more, the certainty) of a model is to investigate the ED with tools like Chimera or Coot. Another way to improve confidence in a model is to overlay closely related crystal structures and to evaluate discrepancies observed between those models. Considering this workflow, a well-characterized starting point for molecular modeling can be defined. In a similar way, the reliability of results published can be estimated by following these procedures, whereby the critical analysis of the model in combination with ED maps is most promising.

7. PDB Selections, Data Mining Tools, and Cross-Linked Databases

A multitude of PDB subsets, data mining tools, and databases cross-linked with the PDB have been reported during the past few years. In this section we summarize valuable PDB-derived libraries, Web tools for data mining, and second-party databases. The repositories and tools are categorized in several classes, although there is obviously a smooth transition between these categories.

7.1. Repositories and Services Focusing Interfaces and Interactions. 7.1.1. Protein−Ligand Interfaces. The Screening Protein Data Bank (sc-PDB)[96] is a PDB subset that offers 6000+ 3D structures of druggable binding sites prepared for virtual screening. The annotation of sc-PDB entries with second party databases has been revised and extended. Solvents, detergents, and most metal ions have been removed from sc-PDB. sc-PDB is particularly valuable for inverse docking campaigns,[26] target fishing, analyzing molecular similarities between binding pockets, and examining pharmacophoric relations between targets. LigBase[97] is a repository of ligand-binding proteins aligned to structural templates taken from the PDB. The Web service offers multiple alignments and schematic LIGPLOT[98] diagrams of protein−ligand interactions. Biomolecular data such as calculated binding energies, ligand similarities based on Tanimoto coefficients, and protein sequence similarity percentages on 450+ complexes are provided by the PLD[99] (Protein−Ligand Database). SitesBase[100,101] is a Web facility offering tools for the investigation of protein−ligand binding site similarities and comparison of the spatial location of ligands in the binding site. The repository supports keyword searches, browsing, atomic multiple alignment, and superposition. Further examples of programs for the identification of protein cavities include PASS[102] (putative active sites with spheres), SURFNET,[103] and Q-SiteFinder.[104] MOE features a novel protein−ligand interaction diagram generator for intuitive inspection of polar, hydrophobic, acidic, and basic interactions. Recently, Stierand and Rarey[105] introduced a novel algorithm for 2D depiction of protein−ligand complexes. Also, LigandScout supports a 2D protein−ligand interaction maps (Figure 1b).

7.1.2. Domain−Domain, Protein−Protein, and Protein−DNA/RNA Interfaces. The investigation of protein−ligand interfaces can be extremely helpful for protein classification. Several different kinds of approaches are available for the classification of protein 3D structures according to secondary structure elements and their fold. Recently, Nebel et al.[106] presented an automated classification approach based on the 3D motifs of protein binding sites. Again, there are in silico methods available that aim at the identification of protein interfaces in order to gain knowledge on the drugability of such protein regions (e.g., InSite,[107] Protemot[108]).

PDBSITE[109] is a Web service offering structural and functional information on various protein site categories of PDB complexes, including protein−ligand and protein−protein interaction sites. It accumulates amino acid content and physicochemical properties of the protein sites and the related vicinities. Focusing on domain−domain and protein−protein interaction information of PDB structures, PSIbase[110] (protein structural interactome map) provides a data repository for domain−domain and protein−protein interaction information of PDB structures based on the PSIMAP algorithm. Interchain and intrachain interfaces can be displayed and statistically analyzed with InterPare.[111] PROTCOM[112] is a data pool offering a set of protein−protein and domain−domain structures. This information is useful for protein−protein docking benchmarks and as a template collection for homology modeling. Molsurfer[113] is a powerful Java-based Web application for analyzing and visualizing protein interfaces and their physicochemical properties, such as hydrophobicity and electrostatic potential. Because of its direct coupling of 2D and 3D views, this application allows the user to gain insight to the characteristics of the protein site quickly. Thereby, Molsurfer is not restricted to protein−protein interfaces but also allows for investigating protein−DNA and protein−RNA complexes. In a similar way, iPfam[114] supports visualizing and browsing of domain−domain interfaces and interactions.

7.2. Repositories and Services Focusing on Affinity Data. BindingDB[115] is a Web service providing about 20 000 biologically tested binding affinities of protein−ligand complexes, with special regard to drug-relevant targets. The binding data have been gained from literature research, and powerful search features allow the user to filter data by target name, sequence, ligand name, affinity data, chemical structure, etc. Similar to Binding DB, Wang et al. have introduced the PDBbind[116,117] database. This online service provides comprehensive and easily accessible affinity data of 4300 PDB protein−ligand complexes. AffinDB[118] is another Web platform for affinity data. The facility currently provides 700+ affinities of 450+ PDB complexes. Recently, KiBank[119] has been updated and now contains 16000+ K_i values, 50 revised target protein structures, and 5900 chemical structures. Binding MOAD[120] (Mother of All Databases) is a comprehensive PDB subset of high-quality crystal structures. In its latest release, this database contains 9500+ protein−ligand complex structures with a resolution of 2.5 Å or better, biologically relevant ligands, and affinity data extracted from literature.

7.3. Ligand Repositories. SuperLigands[121] features searching PDB ligands by ID, compound name, molecular formula, and PDB code. Results can be depicted in 2D and 3D, superimposed, and assessed in terms of drug similarity. The PDB-Ligand[122] Web service supports browsing, classifying, and

superimposing interactions of ligands with proteins; results are visualized using Chime.[123] Ligand Depot[124] is a repository of ligands bound to proteins, providing structural and chemical data of small organic molecules excited from PDB data. The database can be searched for keywords and chemical (sub)structures. Further examples include a complete small molecule data set from the PDB,[125] the HIC-up[126] (Heterocompound Information Centre−Uppsala) Web facility, providing information on 7500+ ligands, and a large set of protein−ligand complexes including chemical properties for the development and performance assessment of knowledge-based docking algorithms.[127]

7.4. PDB-Derived Subsets and Recalculated Data. A 80000+ chains counting PDB subset of representative protein chains is offered with the PDB-REPRDB[128] Web service. The service features list sorting and entry selection based on several parameters such as resolution, *R*-factor, method, etc. The Recalculated Coordinates Database (RECOORD)[129] is a database containing 500+ recalculated NMR protein structures from the PDB. The authors report an improvement of packing benchmarks. In addition, Ramachandran appearance moved 1 standard deviation closer to the mean of the reference database. A subset of the PDB comprising data for transmembrane proteins with known structures is provided by the Database of Membrane Proteins of Known 3D Structure, as discussed in section 3.

There is a strong request for evaluative studies that provide insight on the performance of different virtual screening tools, such as protein−ligand docking or pharmacophore-based screening. Generally, the power to discriminate active and inactive molecules is taken as a benchmark for the potency of the method. Thereby, one essential precondition for such tests is the availability of active and inactive compounds for the targets to be investigated. On the basis of 40 PDB protein−ligand complexes (all of individual targets), Irwin et al. have generated the Directory of Useful Decoys (DUD),[54] the largest public available database of decoys. The DUD is a collection of 36 decoys for each of the 2950 collected actives of 40 different targets (95 316 in total, after duplicate removal). The compounds represent a subset of the ZINC database[130] and have similar physicochemical properties (e.g., molecular weight, calculated log *P*). In this way, the DUD provides a huge collection of molecules that supports testing the predictive power of structure-based virtual screening methods. For more details on this, the reader is referred to a recent review of 3D virtual screening protocols.[131] Protein Quaternary Structure Investigation[132] (PiQSi) is a community-based Web service that facilitates the investigation and curation of protein quaternary structures. Currently, about 15000 manually cured structures are available and allow for direct comparison of structures to homologous proteins.

7.5. PDB Tools and Online Services. iMolTalk[133] is an easy-to-use Web toolkit for analyzing and searching PDB complexes. The services include data analysis, computation, and visualization of Ramachandran plots, distance matrices, analysis of chain interactions within a protein structure, secondary structure assignment, and computation of interactions between a residue or a pair of residues. Furthermore, iMolTalk features multiple sequence alignment and energy minimization. The Protein Structure Analysis Package[134] (PSAP) offers a variety of tools for the investigation of 3D protein structures. Users can upload in-house PDB files or directly access PDB data and receive a comprehensive overview on the characteristics of the investigated protein structure, including information on the protein sequence, water bridges, intra- and interactions, Ramachandran plots, etc. The PDBsum[16,135] is a powerful Web facility that allows for browsing, visualizing, and summing up PDB data (Figure 2b). Furthermore, PDBsum includes a wide range of structure images,[136] annotated plots of protein secondary structures, diagrams of protein−ligand as well as protein−DNA interactions, 3D viewing, surface depiction, and Ramachandran plots. Moreover, browsing by species and ligand is supported. The "highlights" section provides information on the oldest (1b5c, August 10, 1972), latest, and largest (1vri, 150 720 atoms) structures. Depictions of enzyme reactions are available for enzyme structural data files, and several list files of PDB summaries are offered for download. At this point we recommend the Protein Movie Generator[137] (PMG), a Web-based service for the generation of protein structure pictures and animations based on a POV-Ray render engine. Besides simple illustration, complex animations describing molecular dynamics and ligand animations are also supported. The PMG service is an excellent tool for inspiring students and scientific audience.

The Enzyme Structures Database is a repository related to the PDBsum, which supports browsing the PDB by enzyme classification and is accessible from the PDBsum Web site. Enzyme Catalytic Mechanism Database (EzCatDB)[138] is a data pool similar to the Enzyme Structures Database and allows for browsing and searching PDB files considering enzyme classification schemes. The PDB-UF (Protein Data Bank Unknown Function)[139] service features predictions of enzymatic functions of not-annotated PDB entries based on 3D structures. The authors report the identification of probable enzyme functions in cases where standard BLAST tools fail to assign any function. The pKNOT[140] Web server provides analyses tools and structural data on knotted proteins.

7.6. Cross-Linked Data Sources. As already pointed out before, PDB data are cross-linked with several second-party databases. These data repositories can be classified considering four major aspects: (i) protein folds and protein families (e.g., SCOP, structural classification of proteins;[141] CATH, an acronym of the four main levels of this classification scheme, class, architecture, topology, homologous superfamily;[142] and HSSP, homology derived secondary structure of proteins),[143] (ii) protein sequences (e.g., UniProt), (iii) enzyme function and pathways (e.g., BRENDA, Braunschweig Enzyme Database),[144] and (iv) functional annotations (e.g., GO, gene ontology[145]) and taxonomic classifications (e.g., NCBI[146]). PDBSprotEC[147] represents the linking interface between PDB structures and the EC numbers via Swiss-Prot. When a PDB ID or EC number is entered, the Web tool returns the according partner value. A similar cross-linking interface between PDB residues and residues of the UniProtKB/Swiss-Prot and UniProtKB/trEMBL is described by Martin.[148]

7.7. Data Mining Applications. SuMo (initially an acronym for surfing the molecules)[149] allows for screening the PDB for similar protein structures and substructures. The service is valuable for finding binding sites similar to a certain protein structure and also for finding protein structures fitting to a certain binding site. Results can be browsed in list view and detail on every hit is provided on a dedicated page. SuMo can be used for estimating and investigating drug side effects and target similarities. The Database of Multiple Alignments for Protein Structures (DMAPS)[150] offers direct access on precomputed multiple structure alignments of protein structures. The PDB chain ID is required as input; results can be visualized or downloaded in several different formats, including FASTA and

superimposed coordinates in PDB file format. Relibase+ is a popular data mining tool for the investigation of protein−ligand complexes.[151,152] The software suite supports searches based on keywords, SMILES, SMARTS, 2D substructures, 3D protein−ligand as well as 3D protein−protein interaction queries. The comprehensive statistical analyses and filtering tools make Relibase+ a powerful software suite for the large-scale inspection of protein−ligand complexes. pdbFun[153] is a data mining tool for mass selection and fast comparison of annotated PDB residues. Residues can be selected on the basis of the whole PDB structure, protein domains, chains, 2D structure features, residue types, etc. Precalculated selections based on maximum chain dissimilarity are available. Furthermore, pdbFun supports selecting residues, considering surface exposition, clefts, and binding sites. SURFACE[154] (surface residues and functions annotated, compared and evaluated) is a data pool offering surface-based structural comparison and similarity assessment of protein structures and substructures. Results are presented as lists and can be visualized using Chime or RASMOL.[155] The Polypeptide Angle Suffix Tree (PAST) search engine[156] is a powerful and fast Web facility for protein structure search that is based on protein substructures such as functional motifs. Results are depicted in list style and are directly cross-linked to the PDB. Output parameters include the amino acid positions of the hitting structures and several similarity measures. Three-dimensional locally conserved residues in an ensemble of protein structures can be identified using 3dLOGO.[157] Conserved amino acids are identified after superimposition of all input structures. In the next step, a consensus sequence is calculated that can be used as input for sequence database searches. FeatureMap3D[158] supports aligning query sequences to PDB structures. Hits are visualized by colorizing mapped sequences on the protein structure. PISCES[159] is a PDB sequence culling service that allows users to generate subsets of PDB sequences according to quality (*R*-factor, resolution) and maximum mutual sequence identity. PISCES can search the whole PDB or user-defined subsets. Precompiled lists of PDB subsets are available for defined *R*-factor, resolution, and similarity cutoffs. MMsINC[160] (an acronym for the molecular modeling section of the University of Padova) is a free Web facility providing more than four million chemical entities, integrating the PubChem and PDB databases. The Web interface allows for searching for substructures and structurally similar compounds and is especially useful for the elucidation of interesting PDB complexes based on the ligand structure.[160] DBAli[161] tools represents a broad suite of software tools for the examination of PDB data. The services include DBAlit, a program for the comparison of protein structures with PDB data, AnnoLite and AnnoLyze for protein chain annotation, ModClus for chain clustering, ModDom for domain assignment, and SALIGN for multiple chain alignment.

8. Techniques and Applications Relying on PDB-Derived Data

Today, structural protein data provided by the PDB have become of indispensable value for various applications in structural biology, bioinformatics, cheminformatics, and molecular modeling. PDB data mining allows for fast and efficient statistical analysis of protein−ligand interactions. Recently, Cotesta and Stahl[162] presented an investigation on the environment of amide groups (NH and C═O groups) in 3200+ protein−ligand complexes. They found that the vast majority of these amide functions at protein−ligand interfaces are buried deeply within the binding site and are crucial for the formation of a hydrogen bond network. These results are especially valuable for the refinement and design of scoring functions. Examples of scoring functions derived from protein−ligand complexes include DrugScore[163] and the protein−ligand affinity statistical score (PLASS).[164]

The PDB is also a valuable data pool for bioactive conformations of ligands. Brameld et al.[165] have published an analysis of conformational preferences of small, organic druglike molecules based on Cambridge Structural Database (CSD)[166] and PDB data. They report preferred conformations for acyclic moieties and sulfonamides, as these are of particular importance for drug design.

Most 3D virtual screening techniques rely on the accurate representation of the bioactive conformation. However, the bioactive conformation is not necessarily located at the global energetic minimum, as it is considerably influenced by the protein environment and the resulting protein−ligand interactions. Therefore, the bioactive conformation is predicted in silico by covering the low energy conformational space smoothly with only a few calculated conformers. Today, high-throughput screening technologies usually use precalculated 3D databases for fast querying. These databases consist of conformational ensembles for each molecule. Conformational model generators are used for the calculation of such conformational models. In order to investigate the performance of the conformational model generators CATALYST[167] and OMEGA,[123] we have used large samples of PDB ligands in their experimentally determined conformation, extracted these from the protein environment, and calculated the conformational models. The best fitting calculated conformer was compared to the bioactive conformation and measured in terms of rmsd. Our results show that in the vast majority of cases conformational model generators are able to represent the bioactive conformation in a quality that is suitable for virtual screening. In only a minority of cases chemical features drift too far from the reference and insufficient solutions are calculated.[168,169] Follow-up investigations on the impact of conformational model quality on pharmacophore-based and shape-based screening confirmed that the quality of conformational models generated with CATALYST, CAESAR,[170] or OMEGA is downright suitable for virtual screening.[171]

Also on the macromolecular side, PDB data are investigated for the optimization of force field parameters for proteins. Sakae and Okamoto[172] report the refinement of parameters of the AMBER parm94 force field using PDB structures. Wu et al.[173] demonstrate the refinement of NMR-determined protein structures based on knowledge-based potentials derived from PDB structures.

The PDB data repository features protein classification based on structural relations. Prasad et al.[174] reported a method for assessing protein similarities using signature patterns of intramolecular interaction networks. Five different classes of protein structures have been investigated, and the whole PDB data pool was searched for similarities. DALI[175] supports comparing protein contacts based on distance matrices. A Monte Carlo based algorithm is used for the optimization of the similarity scoring function. Other similarity analysis methods include geometric hashing[176] and incremental combinatorial extension.[177]

Janin et al.[178] investigated the characteristics of protein−protein interfaces in a large scale study. They report that about 45 atoms of each side forming a molecular contact surface of about 900

Å^2 are needed in order to form biologically relevant protein−protein or protein−DNA interactions.

PDB data are inevitable for developing and validating algorithms for the identification of putative ligand binding sites. An et al.[179] report a method for the efficient detection of "druggable" protein sites based on the grid potential map of the van der Waals interaction of the receptor. Moreover, PDB structures can be used as templates for the calculation of 3D models of protein−ligand complexes. Hare et al.[180] describe an automated approach for predicting the 3D structure of such complexes and report that their method is able to predict retrospectively the binding of 70% of small-molecule protein kinase inhibitors published in the *Journal of Medicinal Chemistry* since 1993 with an rmsd from the X-ray structure smaller than 2 Å.

In cases where a protein structure is unavailable, homology modeling based on related structures is a prominent approach for the generation of a theoretical structure. Thereby, the putative protein structure is deduced from template structures[181] of experimentally determined proteins. An example for the successful application of homology modeling techniques is the identification of human histamine H3 receptor inhibitors using homology modeling in combination with pharmacophore modeling and docking.[182] One of the most-established Web services for homology modeling is the Swiss-Model server.[183] The server offers an integrated homology modeling platform providing regularly updated databases, tools, and programs. Users can create their own workspace and are informed about the current status of their Web jobs via e-mail. Structural genomics projects are of particular benefit for homology modeling, as these initiatives focus on structural diversity and therefore provide collections of novel protein structure templates to be used for homology-based modeling.

MaxSprout[184] and SABBAC[185] (structural alphabet based protein backbone builder from α-carbon trace) allow for reconstructing the backbone of proteins and amino acid side chains based on α-carbon trace.

Despite the popular assumption that proteins similar in terms of their 2D sequence are also similar in their 3D structure, Kosloff and Kolodny[186] prove that sequence similarity is not necessarily related to 3D structure homology. The authors point out that this fact can lead to structural and functional information loss upon diversity-based PDB culling. Moreover, they highlight possible issues arising from template selection for homology modeling if based on 2D sequence similarity.

The increase of available protein structural data, in particular from the PDB, has accelerated the development of structure-based methods considerably. Today, protein−ligand docking is considered the most important approach for structure-based virtual screening. There is a plentitude of very different approaches available that aim at distinct fields of application, and there is an even higher number of scoring functions available. However, so far there is no universal tool available that offers reliable scoring for all pharmaceutically relevant targets. Warren et al.[187] provide a comprehensive survey, investigating the performance of 10 docking programs and 37 scoring functions. They found that docking programs are in general able to generate ligand poses that are similar to the experimentally determined ligand pose bound to the protein. The authors report no statistically significant correlation between docking scores and ligand affinity. Nevertheless, docking is a powerful approach to rationalize drug action.[188−191] For more detail we refer to the review by Coupez et al.,[192] which provides an overview on currently available docking techniques and their reliability.

When structure-based pharmacophore models are used as screening filters instead of 3D coordinates of protein atoms, affinity estimation is based on the geometric fit of structures to the model. In this case, the values calculated are often far from reality; however, they are useful for filtering possible hits from nonbinding molecules. LigandScout generates pharmacophore models based on a given 3D structure of a PDB protein−ligand complex. The fully automated creation of pharmacophore is based on a set of rules that automatically detects and classifies protein−ligand interactions into hydrogen bonds, charge transfer interactions, and lipophilic regions (Figure 1a,b). The entire set of interactions forms a pharmacophore model, which can be used for VS in LigandScout but also external screening platforms such as CATALYST,[171] MOE, and PHASE.[193]

Publications of structure-based pharmacophore screening include the successful identification of novel 17β-HSD1 inhibitors by Schuster et al.[194] Steindl et al.[195] used structure-based pharmacophore modeling together with statistical analysis of molecular descriptors to identify new inhibitors of the human rhinovirus coat protein. Another work group discovered new human 5-lipoxygenase inhibitors active at concentrations in the nanomolar range by a combined ligand- and target-based approach.[196] A comparison of results from pharmacophore modeling and docking led to structural insights into the mode of action of such compounds. Rella et al.[197] were able to find novel chemical scaffolds for the development of selective angiotensin-converting enzyme 2 using structure-based pharmacophore modeling. Spitzer et al.[198] used pharmacophore models to describe interactions of ligands binding to the minor groove of the DNA. The study focused on the implementation of sequence-specific properties encoded by the minor groove. The pharmacophore models were created by using DNA structure information exclusively, as provided by the PDB.

In order to lower the risk of failure of promising clinical candidates, pharmaceutical industry puts considerable effort in the early detection of so-called antitarget interactions. Pharmacological profiling of compounds in an early stage of drug discovery would lower experimental costs and the risk of failure significantly. While fast in vitro assays in this early stage of drug discovery have been established, recent advances in technology (in particular, in terms of screening speed but also in data visualization and data handling) and the availability of comprehensive collections of QSAR data are currently boosting the development and application of so-called parallel screening approaches for activity profiling.[33,34,199] Moreover, parallel screening techniques are also of great value for revealing unknown binding modes by target fishing as well as for scanning approved drugs and off-patent medications for so far unknown (inter)actions. Such compounds could be approved for new indications with considerably lower financial and experimental costs and offer in particular academia an interesting alternative.

The fully automated concept of pharmacophore-based parallel screening was realized in the latest version of Discovery Studio.[13] The program allows screening of one or more single- or multiconformer compounds or even whole databases against a series of pharmacophore models using CATALYST components. The results can be displayed as a heat map, where a color-coded matrix presents all compounds. Along with the parallel screening technology, the Inte:Ligand pharmacophore database[77] is available, which currently contains 1846 structure-based pharmacophore models covering 195 unique pharmacological

targets. All models have been developed with LigandScout and are based on PDB data, which is another evidence for the PDB's global significance. For each model, metadata are available describing the PDB entry and the ligand it was derived from, a selectivity index (hits from a random, druglike database), the pharmacological target, and the mechanism of action.

Steindl et al.[200] published the results of a parallel screening campaign of 100 antiviral compounds against 50 models belonging to five different targets. A correct activity profile was retrieved in 89% of the cases. In a second experiment,[33] they determined the selectivity of HIV protease inhibitor models against other protease inhibitors and inactive compounds. The results showed a clear trend toward most extensive retrieval of known actives followed by general protease inhibitors and lowest recovery of inactive compounds. Markt et al.[171] performed a validation study of the target fishing approach using 357 compounds with known activity on the peroxisome proliferator activated receptor (PPAR). They screened all compounds against all models from the Inte:Ligand pharmacophore database. The PPAR target was ranked first more often than any other of the 181 targets.

9. Future Directions of the PDB

After more than 3 decades of constant development and maturation, the PDB has been established undoubtedly as the most important public data source for structure-based drug design. Major challenges in the future will be the standardization, annotation, and linking of PDB content with second-party databases as well as the further development of file formats. Structural biologists locate major spaces for improvement in the characterization of disordered structures and very large macromolecules. From the perspective of computational chemistry and molecular modeling we perceive a strong need for more attention to protein interaction sites and ligands bound to the target. Structure-based modeling methods suffer from lack of definitions of the hybridization states of ligand atoms. Moreover, crystallographers tend to generate globally optimized models that may prevent highly precise data on protein areas that are of particular interest for molecular modeling. The still very unbalanced representation of targets in terms of PDB entries and the little information available on certain target families (e.g., GPCRs) do not allow global, structure-based multitarget screening. We hope (and are confident) that further technological advances and the strong efforts taken by structural genomics projects will allow drastic increase of the spectrum and diversity of known protein structures, and we want to encourage structural biologists to increase attention on target areas of special interest to molecular modeling. However, the determination of only one structure per target is certainly not enough. In order to understand the conformational flexibility of a protein (interface), an ensemble of protein structures in complex with chemically diverse ligands is required. Robotics for high-throughput structural determination, as they are being developed by structural genomics centers, will also help to increase mass production of proteins cocrystallized with small organic molecules. Even more, high-resolution X-ray crystallography is able to detect alternate conformations of amino acid residues and ligands and provides insight on protein mobility at atom level. It allows us to understand the function and biology of proteins at the atom level. These increased efforts to characterize and reveal the impact of protein mobility are of fundamental importance to understanding the biology of proteins and may allow us to predict 3D structures of proteins reliably and, even more, to predict the function of proteins in the future.

10. Conclusions

The PDB and its partners have become of indispensable value for rational ligand design and will gain, with growing coverage and diversity, even more importance in future. A plethora of PDB-related data mining and validation tools, Web services, and subsets have been developed within the past few years. The vast majority of these tools not only is of interest to scientists focused on structure determination but is of great value for medicinal chemists. We have highlighted the current coverage and features of the PDB and related services, their scope, and their limits and provide an overview on available software tools and online services. We hope that this Perspective and the guidelines provided will encourage medicinal chemists to start and intensify the usage of PDB-related resources to develop new ideas and to boost the progress of drug discovery projects.

Biographies

Johannes Kirchmair studied pharmacy at the Leopold-Franzens University of Innsbruck, Austria, from 1999 to 2004 and received his Ph.D. degree under the guidance of Professor Thierry Langer in 2007. For his Ph.D. he specialized in structure-based virtual high-throughput screening and parallel screening techniques. He began his career at Inte:Ligand GmbH, Vienna, Austria, working on computer-guided drug development using 3D virtual high-throughput screening techniques for the identification and optimization of novel anticancer agents. His research interests cover medicinal chemistry, computational chemistry, and drug design as well as QSAR and 3D-QSAR molecular modeling techniques. Since April 2008, Dr. Kirchmair has been holding an assistant lecturer position at the University of Innsbruck and is working for Inte:Ligand GmbH.

Patrick Markt studied pharmacy at the Leopold-Franzens University of Innsbruck, Austria (1999−2005). In 2006, after finishing his training as pharmacists, he started his Ph.D. thesis at the University of Innsbruck under the guidance of Professor Thierry Langer in collaboration with Inte:Ligand GmbH. Dr. Markt finished his Ph.D. thesis on parallel screening and drug discovery for targets involved in the metabolic syndrome in 2008. His main research fields are ligand-based and structure-based virtual screening techniques including pharmacophore modeling, molecular docking, fingerprints, shape and electrostatic similarity search techniques, and data mining. He is holding a research assistant position at the University of Innsbruck.

Simona Distinto received her degree in Chemistry and Pharmaceutical Technology from the University of Cagliari in 2002. After receiving her Ph.D. in Pharmaceutical Chemistry under the supervision of Prof. Elias Maccioni, she was a postdoctoral fellow at the laboratory of the Computational Pharmaceutical Chemistry of the University of Catanzaro under supervision of Professor Stefano Alcaro. She obtained the Master-and-Back scholarship of the Region of Sardinia in 2006 and joined the CAMD group of Professor Thierry Langer at the University of Innsbruck. Her research interests include molecular modeling and simulations for rational ligand design.

Daniela Schuster performed her Masters and Ph.D. studies at the University of Innsbruck at the molecular modeling group of Prof. Thierry Langer. For her Ph.D. thesis, which focused on pharmacophore modeling for targets involved in the metabolic pathway, she received the Sosnowsky Award 2007 and the Dr. Maria Schaumayer Award 2006. Apart from the application of pharmacophore-based parallel screening for the discovery of bioactive natural products, she is involved in e-learning projects. Currently, she is holding a position as a project researcher at the University of Innsbruck.

Gudrun M. Spitzer finished her chemistry studies in 2005. After an internship at the Department of Lead Discovery at Boehringer Ingelheim Pharma GmbH & Co. KG in Biberach, Germany, she is now heading a research project at the Center for Molecular

Biosciences at Innsbruck. She obtained the Dr. Otto Seibert Research Award 2007. Her research is focused on the development of lead discovery methods aiming at the minor groove of DNA as target for drug design. Gudrun M. Spitzer is teaching molecular modeling, computer simulations, chemoinformatics, and bioinformatics at the Department of Theoretical Chemistry, University of Innsbruck, Austria.

Klaus R. Liedl finished his chemistry and mathematics studies at University of Innsbruck in 1989 and 1992, and he studied theoretical physics and obtained a Ph.D. in chemistry in 1995. Since 1998 he has been a tenured Associate Professor of theoretical chemistry. In 1998 he obtained the Novartis research award, and in 2002 he obtained the Hellmann award of AG Theoretical Chemistry (German Bunsen Society, DPG, GDCh). In 2006 he finished his doctorate in juridical science focusing on intellectual property rights, complementing his chemical research as proven by his more than 100 publications in the field. His main research interests are focused on drug development and the prediction of kinetics and thermodynamics in the course of biomolecular binding processes and reactions.

Thierry Langer began his career at Leopold-Franzens University of Innsbruck after completing a postdoctoral fellowship at the Université Louis Pasteur, Strasbourg, France. He holds an M.S. degree in Pharmacy and a Ph.D. from the University of Vienna, Austria. His research interests range from medicinal chemistry, including theoretical pharmaceutical chemistry, drug design, and pharmacophore modeling as well as QSAR and 3D-QSAR molecular modeling techniques. His scientific work, which has culminated in more than 160 original articles, book chapters, and invited reviews, together with more than 200 lectures and presentations at scientific meetings, led to the founding of the spin-off company Inte:Ligand GmbH, in which he also was CEO from 2003 to 2008. In April 2008, Prof. Langer left University of Innsbruck to be appointed CEO of Prestwick Chemical Inc., Strasbourg-Illkirch, France.

Gerhard Wolber received his Ph.D. in pharmaceutical chemistry at the University of Innsbruck in 2003 after his studies of computer science and pharmacy at the University of Innsbruck and University of Vienna, Austria, respectively. As one of the founders of the drug design company Inte:Ligand, he has been working as head of cheminformatics and research since 2003, where he has been developing the two programs "ilib diverse" and "LigandScout". In 2008 he took a position as a lecturer in pharmaceutical chemistry at the Institute of Pharmacy at the University of Innsbruck, where he is now heading his own research group. His research interests include computational drug design and the development of algorithms for drug−receptor interaction, virtual screening, 2D and 3D visualization techniques, and 3D pharmacophore modeling.

References

(1) Berman, H. M. The Protein Data Bank: A historical perspective. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2008**, *A64*, 88–95.

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

(3) Boutselakis, H.; Dimitropoulos, D.; Fillon, J.; Golovin, A.; Henrick, K.; Hussain, A.; Ionides, J.; John, M.; Keller, P. A.; Krissinel, E.; McNeil, P.; Naim, A.; Newman, R.; Oldfield, T.; Pineda, J.; Rachedi, A.; Copeland, J.; Sitnov, A.; Sobhany, S.; Suarez-Uruena, A.; Swaminathan, J.; Tagari, M.; Tate, J.; Tromm, S.; Velankar, S.; Vranken, W. E-MSD: The European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Res.* **2003**, *31*, 458–462.

(4) Golovin, A.; Oldfield, T. J.; Tate, J. G.; Velankar, S.; Barton, G. J.; Boutselakis, H.; Dimitropoulos, D.; Fillon, J.; Hussain, A.; Ionides, J. M. C.; John, M.; Keller, P. A.; Krissinel, E.; McNeil, P.; Naim, A.; Newman, R.; Pajon, A.; Pineda, J.; Rachedi, A.; Copeland, J.; Sitnov, A.; Sobhany, S.; Suarez-Uruena, A.; Swaminathan, G. J.; Tagari, M.; Tromm, S.; Vranken, W.; Henrick, K. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.* **2004**, *32*, D211–216.

(5) Nakamura, H. Data curation, quality control, and user services at Protein Data Bank Japan. *Nippon Kessho Gakkaishi* **2005**, *47*, 334–340.

(6) Berman, H.; Henrick, K.; Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **2003**, *10*, 980.

(7) Markley, J. L.; Ulrich, E. L.; Berman, H. M.; Henrick, K.; Nakamura, H.; Akutsu, H. BioMagResBank (BMRB) as a partner in the worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *J. Biomol. NMR* **2008**, *40*, 153–155.

(8) Galperin, M. Y. The Molecular Biology Database Collection: 2007 update. *Nucleic Acids Res.* **2007**, *35*, D3–4.

(9) Carugo, O.; Pongor, S. The evolution of structural databases. *Trends Biotechnol.* **2002**, *20*, 498–501.

(10) Berman, H.; Henrick, K.; Nakamura, H.; Markley, J. L. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* **2007**, *35*, D301–D303.

(11) Henrick, K.; Feng, Z.; Bluhm, W. F.; Dimitropoulos, D.; Doreleijers, J. F.; Dutta, S.; Flippen-Anderson, J. L.; Ionides, J.; Kamada, C.; Krissinel, E.; Lawson, C. L.; Markley, J. L.; Nakamura, H.; Newman, R.; Shimizu, Y.; Swaminathan, J.; Velankar, S.; Ory, J.; Ulrich, E. L.; Vranken, W.; Westbrook, J.; Yamashita, R.; Yang, H.; Young, J.; Yousufuddin, M.; Berman, H. M. Remediation of the Protein Data Bank archive. *Nucleic Acids Res.* **2008**, *36*, D426–D433.

(12) Dutta, S.; Berman, H. M. Large macromolecular complexes in the Protein Data Bank: a status report. *Structure (Cambridge, MA)* **2005**, *13*, 381–388.

(13) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discovery* **2006**, *5*, 993–996.

(14) Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G. F.; Thian, F. S.; Kobilka, T. S.; Choi, H.-J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High-resolution crystal structure of an engineered human 2-adrenergic G protein coupled receptor. *Science (Washington, D.C.)* **2007**, *318*, 1258–1265.

(15) White, S. H. The progress of membrane protein structure determination. *Protein Sci.* **2004**, *13*, 1948–1949.

(16) Laskowski, R. A.; Chistyakov, V. V.; Thornton, J. M. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res.* **2005**, *33*, D266–D268.

(17) Mestres, J. Representativity of target families in the Protein Data Bank: impact for family-directed structure-based drug discovery. *Drug Discovery Today* **2005**, *10*, 1629–1637.

(18) Levitt, M. Growth of novel protein structural data. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 3183–3188.

(19) Chen, L.; Oughtred, R.; Berman, H. M.; Westbrook, J. TargetDB: a target registration database for structural genomics projects. *Bioinformatics* **2004**, *20*, 2860–2862.

(20) Kouranov, A.; Xie, L.; de la Cruz, J.; Chen, L.; Westbrook, J.; Bourne, P. E.; Berman, H. M. The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.* **2006**, *34*, D302–D305.

(21) Protein Structure Initiative. http://www.nigms.nih.gov/Initiatives/PSI (accessed September 9, 2008).

(22) Sasson, O.; Linial, M. ProTarget: automatic prediction of protein structure novelty. *Nucleic Acids Res.* **2005**, *33*, W81–W84.

(23) Gasteiger, E.; Gattiker, A.; Hoogland, C.; Ivanyi, I.; Appel, R. D.; Bairoch, A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **2003**, *31*, 3784–3788.

(24) Lamb, M. L.; Burdick, K. W.; Toba, S.; Young, M. M.; Skillman, A. G.; Zou, X.; Arnold, J. R.; Kuntz, I. D. Design, docking, and evaluation of multiple libraries against multiple targets. *Proteins: Struct., Funct., Genet.* **2001**, *42*, 296–318.

(25) Aronov, A. M.; Munagala, N. R.; Kuntz, I. D.; Wang, C. C. Virtual screening of combinatorial libraries across a gene family: in search of inhibitors of giardia lamblia guanine phosphoribosyltransferase. *Antimicrob. Agents Chemother.* **2001**, *45*, 2571–2576.

(26) Chen, Y. Z.; Zhi, D. G. Ligand−protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins: Struct., Funct., Genet.* **2001**, *43*, 217–226.

(27) Schapira, M.; Abagyan, R.; Totrov, M. Nuclear hormone receptor targeted virtual screening. *J. Med. Chem.* **2003**, *46*, 3045–3059.

(28) Muller, P.; Lena, G.; Boilard, E.; Bezzine, S.; Lambeau, G.; Guichard, G.; Rognan, D. In silico-guided target identification of a scaffold-focused library: 1,3,5-triazepan-2,6-diones as novel phospholipase A2 inhibitors. *J. Med. Chem.* **2006**, *49*, 6768–6778.

(29) Zahler, S.; Tietze, S.; Totzke, F.; Kubbutat, M.; Meijer, L.; Vollmar, A. M.; Apostolakis, J. Inverse in silico screening for identification of kinase inhibitor targets. *Chem. Biol.* **2007**, *14*, 1207–1214.

(30) Kellenberger, E.; Foata, N.; Rognan, D. Ranking targets in structure-based virtual screening of three-dimensional protein libraries: methods and problems. *J. Chem. Inf. Model.* **2008**, *48*, 1014–1025.

(31) Paul, N.; Kellenberger, E.; Bret, G.; Müller, P.; Rognan, D. Recovering the true targets of specific ligands by virtual screening

of the Protein Data Bank. *Proteins: Struct., Funct., Bioinf.* **2004**, *54*, 671–680.
(32) Markt, P.; Schuster, D.; Kirchmair, J.; Laggner, C.; Langer, T. Pharmacophore modeling and parallel screening for PPAR ligands. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 575–590.
(33) Steindl, T. M.; Schuster, D.; Laggner, C.; Chuang, K.; Hoffmann, R. D.; Langer, T. Parallel screening and activity profiling with HIV protease inhibitor pharmacophore models. *J. Chem. Inf. Model.* **2007**, *47*, 563–571.
(34) Steindl, T. M.; Schuster, D.; Laggner, C.; Langer, T. Parallel screening: a novel concept in pharmacophore modeling and virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 2146–2157.
(35) Standley, D.; Toh, H.; Nakamura, H. ASH structure alignment package: sensitivity and selectivity in domain classification. *BMC Bioinf.* **2007**, *8*, 116.
(36) Pearson, W. R. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **1990**, *183*, 63–98.
(37) Golovin, A.; Dimitropoulos, D.; Oldfield, T.; Rachedi, A.; Henrick, K. Msdsite: a database search and retrieval system for the analysis and viewing of bound ligands and active sites. *Proteins: Struct., Funct., Bioinf.* **2005**, *58*, 190–199.
(38) Krissinel, E.; Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **2007**, *372*, 774–797.
(39) Krissinel, E.; Henrick, K. Detection of Protein Assemblies in Crystals. In *Computational Life Sciences*; Springer: Berlin/Heidelberg, Germany, 2005; pp 163−174.
(40) Krissinel, E.; Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 2256–2268.
(41) Hartshorn, M. J. AstexViewer: a visualisation aid for structure-based drug design. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 871–881.
(42) Dimitropoulos, D.; Henrick, K.; Swaminathan, J.; Golovin, A. Analytical processing of the PDB on the Web: examining ligand fragments and their environment. *WSEAS Trans. Biol. Biomed.* **2006**, *3*, 414–420.
(43) Westbrook, J.; Feng, Z.; Burkhardt, K.; Berman, H. M. Validation of protein structures for Protein Data Bank. *Methods Enzymol.* **2003**, *374*, 370–385.
(44) Joosten, R. P.; Vriend, G. PDB improvement starts with data deposition. *Science (Washington, D.C.)* **2007**, *317*, 195–196.
(45) Bond, C. S. Easy editing of Protein Data Bank formatted files with EMACS. *J. Appl. Crystallogr.* **2003**, *36*, 350–351.
(46) Hamelryck, T.; Manderick, B. PDB file parser and structure class implemented in python. *Bioinformatics* **2003**, *19*, 2308–2310.
(47) Hussain, A. S. Z.; Shanthi, V.; Sheik, S. S.; Jeyakanthan, J.; Selvarani, P.; Sekar, K. PDB goodies, a Web-based GUI to manipulate the Protein Data Bank file. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *D58*, 1385–1386.
(48) Bourne, P. E.; Berman, H. M.; McMahon, B.; Watenpaugh, K. D.; Westbrook, J. D.; Fitzgerald, P. M. D. Macromolecular crystallographic information file. *Methods Enzymol.* **1997**, *277*, 571–590.
(49) Deshpande, N.; Addess, K. J.; Bluhm, W. F.; Merino-Ott, J. C.; Townsend-Merino, W.; Zhang, Q.; Knezevich, C.; Xie, L.; Chen, L.; Feng, Z.; Kramer Green, R.; Flippen-Anderson, J. L.; Westbrook, J.; Berman, H. M.; Bourne, P. E. The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.* **2005**, *33*, D233–D237.
(50) Westbrook, J.; Ito, N.; Nakamura, H.; Henrick, K.; Berman, H. M. PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics* **2005**, *21*, 988–992.
(51) Dutta, S.; Burkhardt, K.; Swaminathan, G. J.; Kosada, T.; Henrick, K.; Nakamura, H.; Berman, H. M. Data deposition and annotation at the worldwide Protein Data Bank. *Methods Mol. Biol.* **2008**, *426*, 81–101.
(52) Rother, K.; Müller, H.; Trissl, S.; Koch, I.; Steinke, T.; Preissner, R.; Frömmel, C.; Leser, U. *Columba: Multidimensional Data Integration of Protein Annotations*; Data Integration in the Life Sciences: Leipzig, Germany, 2004; pp 156−171.
(53) Jain, A. N.; Nicholls, A. Special issue on "evaluation of computational methods". *J. Comput.-Aided Mol. Des.* **2008**, *22*, 131–265.
(54) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
(55) Rother, K.; Michalsky, E.; Leser, U. How well are protein structures annotated in secondary databases? *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 571–576.
(56) Brown, E. N.; Ramaswamy, S. Quality of protein crystal structures. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2007**, *D63*, 941–950.
(57) Andrec, M.; Snyder, D. A.; Zhou, Z.; Young, J.; Montelione, G. T.; Levy, R. M. A large data set comparison of protein structures determined by crystallography and NMR: statistical test for structural differences and the effect of crystal packing. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 449–465.
(58) Snyder, D. A.; Bhattacharya, A.; Huang, Y. J.; Montelione, G. T. Assessing precision and accuracy of protein structures derived from NMR data. *Proteins: Struct., Funct., Bioinf.* **2005**, *59*, 655–661.
(59) Billeter, M. Comparison of protein structures determined by NMR in solution and by X-ray diffraction in single crystals. *Q. Rev. Biophys.* **1992**, *25*, 325–377.
(60) Garbuzynskiy, S. O.; Melnik, B. S.; Lobanov, M. Y.; Finkelstein, A. V.; Galzitskaya, O. V. Comparison of X-ray and NMR structures: Is there a systematic difference in residue contacts between X-ray- and NMR-resolved protein structures? *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 139–147.
(61) Malcolm, W.; MacArthur, J. M. T. Conformational analysis of protein structures derived from NMR data. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 232–251.
(62) Abagyan, R. A.; Totrov, M. M. Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. *J. Mol. Biol.* **1997**, *268*, 678–685.
(63) Ratnaparkhi, G. S.; Ramachandran, S.; Udgaonkar, J. B.; Varadarajan, R. Discrepancies between the NMR and X-ray structures of uncomplexed barstar: analysis suggests that packing densities of protein structures determined by NMR are unreliable. *Biochemistry* **1998**, *37*, 6958–6966.
(64) Doreleijers, J. F.; Rullmann, J. A. C.; Kaptein, R. Quality assessment of NMR structures: a statistical survey. *J. Mol. Biol.* **1998**, *281*, 149–164.
(65) Spronk, C. A. E. M.; Linge, J. P.; Hilbers, C. W.; Vuister, G. W. Improving the quality of protein structures derived by NMR spectroscopy. *J. Biomol. NMR* **2002**, *22*, 281–289.
(66) Hooft, R. W. W.; Sander, C.; Vriend, G. Reconstruction of symmetry-related molecules from Protein Data Bank (PDB) files. *J. Appl. Crystallogr.* **1994**, *27*, 1006–1009.
(67) Henrick, K.; Thornton, J. M. PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **1998**, *23*, 358–361.
(68) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **1963**, *7*, 95–99.
(69) Hooft, R. W. W.; Sander, C.; Vriend, G. Verification of protein structures: side-chain planarity. *J. Appl. Crystallogr.* **1996**, *29*, 714–716.
(70) Bairoch, A.; Bougueleret, L.; Altairac, S.; Amendolia, V.; Auchincloss, A.; Puy, G. A.; Axelsen, K.; Baratin, D.; Blatter, M.-C.; Boeckmann, B.; Bollondi, L.; Boutet, E.; Quintaje, S. B.; Breuza, L.; Bridge, A.; deCastro, E.; Coral, D.; Coudert, E.; Cusin, I.; Dobrokhotov, P.; Dornevil, D.; Duvaud, S.; Estreicher, A.; Famiglietti, L.; Feuermann, M.; Gehant, S.; Farriol-Mathis, N.; Ferro, S.; Gasteiger, E.; Gateau, A.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Hulo, N.; Ioannidis, V.; Ivanyi, I.; James, J.; Jain, E.; Jimenez, S.; Jungo, F.; Junker, V.; Keller, G.; Lachaize, C.; Lane-Guermonprez, L.; Langendijk-Genevaux, P.; Lara, V.; Lemercier, P.; Saux, V. L.; Lieberherr, D.; Lima, T. d. O.; Mangold, V.; Martin, X.; Michoud, K.; Moinat, M.; Moreira, C.; Morgat, A.; Nicolas, M.; Ohji, S.; Paesano, S.; Pedruzzi, I.; Perret, D.; Phan, I.; Pilbout, S.; Pillet, V.; Poux, S.; Redaschi, N.; Reynaud, S.; Rivoire, C.; Roechert, B.; Sapsezian, C.; Schneider, M.; Sigrist, C.; Silva, M. d.; Sonesson, K.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Veuthey, A.-L.; Vitorello, C.; Yip, L.; Apweiler, R.; Alam-Faruque, Y.; Barrell, D.; Bower, L.; Browne, P.; Chan, W. M.; Daugherty, L.; Donate, E. S.; Eberhardt, R.; Fedotov, A.; Foulger, R.; Fraser, G.; Frigerio, G.; Garavelli, J.; Golin, R.; Horne, A.; Jacobsen, J.; Kleen, M.; Kersey, P.; Kretschmann, E.; Laiho, K.; Leinonen, R.; Legge, D.; Magrane, M.; Martin, M. J.; Monteiro, P.; O'Donovan, C.; Orchard, S.; O'Rourke, J.; Patient, S.; Pruess, M.; Sitnov, A.; Sklyar, N.; Whitfield, E.; Wieser, D.; Lin, Q.; Rynbeek, M.; Martino, G. d.; Donnelly, M.; Rensburg, P. v.; Wu, C.; Arighi, C.; Arminski, L.; Barker, W.; Chen, Y.; Chung, S.; Fang, C.; Hermoso, V.; Hu, Z.-Z.; Hua, H.-K.; Huang, H.; Kahsay, R.; Mazumder, R.; McGarvey, P.; Natale, D.; Ni, A. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2008**, *36*, D190–D195.
(71) Laskowski, R. A.; Swaminathan, G. J. Problems of Protein Three-Dimensional Structures. In *Comprehensive Medicinal Chemistry II*; Triggle, D., Taylor, J., Eds.; Elsevier: Amsterdam, The Netherlands, 2006; Vol. 3, pp 531−550.
(72) Fields, B. A.; Bartsch, H. H.; Bartunik, H. D.; Cordes, F.; Guss, J. M.; Freeman, H. C. Accuracy and precision in protein crystal structure analysis: two independent refinements of the structure of poplar plastocyanin at 173 K. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1994**, *D50*, 709–730.
(73) Hooft, R. W. W.; Vriend, G.; Sander, C.; Abola, E. E. Errors in protein structures. *Nature (London)* **1996**, *381*, 272.
(74) Joosten, R. P.; Chinea, G.; Kleywegt, G. J.; Vriend, G. Protein Three-Dimensional Structure Validation. In *Comprehensive Medicinal*

Chemistry II; Triggle, D., Taylor, J., Eds.; Elsevier: Amsterdam, The Netherlands, 2006; Vol. 3, pp 507−530.
(75) *Moe*; Chemical Computing Group (CCG): Montreal, QC.
(76) *Sybyl*; Tripos: St. Louis, MO.
(77) Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45*, 160–169.
(78) Wolber, G.; Dornhofer, A.; Langer, T. Efficient overlay of small organic molecules using 3D pharmacophores. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 773–788.
(79) Badger, J.; Hendle, J. Reliable quality-control methods for protein crystal structures. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 284–291.
(80) Venclovas, C.; Ginalski, K.; Kang, C. Sequence-structure mapping errors in the PDB: OB-fold domains. *Protein Sci.* **2004**, *13*, 1594–1602.
(81) Kleywegt, G. J.; Harris, M. R.; Zou, J.-y.; Taylor, T. C.; Wahlby, A.; Jones, T. A. The Uppsala Electron-Density Server. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 2240–2249.
(82) Vaguine, A. A.; Richelle, J.; Wodak, S. J. SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1999**, *D55*, 191–205.
(83) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF chimera, a visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612.
(84) Emsley, P.; Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *D60*, 2126–2132.
(85) Wlodek, S.; Skillman, A. G.; Nicholls, A. Automated ligand placement and refinement with a combined force field and shape potential. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2006**, *D62*, 741–749.
(86) Brunger, A. T. Free *R*-value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature (London)* **1992**, *355*, 472–475.
(87) Jones, T. A.; Zou, J. Y.; Cowan, S. W.; Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1991**, *47*, 110–119.
(88) Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **1993**, *26*, 283–291.
(89) Laskowski, R. A.; Antoon, J.; Rullmann, C.; Macarthur, M. W.; Kaptein, R.; Thornton, J. M. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **1996**, *8*, 477–486.
(90) Davis, I. W.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. MolProbity: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res.* **2004**, *32*, W615–W619.
(91) Weichenberger, C. X.; Sippl, M. J. NQ-Flipper: recognition and correction of erroneous asparagine and glutamine side-chain rotamers in protein structures. *Nucleic Acids Res.* **2007**, *35*, W403–W406.
(92) Willard, L.; Ranjan, A.; Zhang, H.; Monzavi, H.; Boyko, R. F.; Sykes, B. D.; Wishart, D. S. Vadar: a Web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res.* **2003**, *31*, 3316–3319.
(93) Bowie, J. U.; Luthy, R.; Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science (Washington, D.C.)* **1991**, *253*, 164–170.
(94) Eisenberg, D.; Luthy, R.; Bowie, J. U. VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol.* **1997**, *277*, 396–404.
(95) Bhattacharya, A.; Tejero, R.; Montelione, G. T. Evaluating protein structures determined by structural genomics consortia. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 778–795.
(96) Kellenberger, E.; Muller, P.; Schalon, C.; Bret, G.; Foata, N.; Rognan, D. sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *J. Chem. Inf. Model.* **2006**, *46*, 717–727.
(97) Stuart, A. C.; Ilyin, V. A.; Sali, A. LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics* **2002**, *18*, 200–201.
(98) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein−ligand interactions. *Protein Eng.* **1995**, *8*, 127–134.
(99) Puvanendrampillai, D.; Mitchell, J. B. O. L/D protein ligand database (PLD): additional understanding of the nature and specificity of protein−ligand complexes. *Bioinformatics* **2003**, *19*, 1856–1857.
(100) Gold, N. D.; Jackson, R. M. Sitesbase: a database for structure-based protein−ligand binding site comparisons. *Nucleic Acids Res.* **2006**, *34*, D231–D234.
(101) Gold, N. D.; Jackson, R. M. A searchable database for comparing protein−ligand binding sites for the analysis of structure−function relationships. *J. Chem. Inf. Model.* **2006**, *46*, 736–742.
(102) Brady, G. P.; Stouten, P. F. W. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 383–401.
(103) Laskowski, R. A. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graphics* **1995**, *13*, 323–330.
(104) Laurie, A. T. R.; Jackson, R. M. Q-SiteFinder: an energy-based method for the prediction of protein−ligand binding sites. *Bioinformatics* **2005**, *21*, 1908–1916.
(105) Stierand, K.; Rarey, M. From modeling to medicinal chemistry: automatic generation of two-dimensional complex diagrams. *ChemMedChem* **2007**, *2*, 853–860.
(106) Nebel, J.-C.; Herzyk, P.; Gilbert, D. R. Automatic generation of 3D motifs for classification of protein binding sites. *BMC Bioinf.* **2007**, *8*, 321.
(107) Wang, H.; Segal, E.; Ben-Hur, A.; Li, Q.-R.; Vidal, M.; Koller, D. InSite: a computational method for identifying protein−protein interaction binding sites on a proteome-wide scale. *Genome Biol.* **2007**, 8.
(108) Chang, D. T.-H.; Weng, Y.-Z.; Lin, J.-H.; Hwang, M.-J.; Oyang, Y.-J. Protemot: prediction of protein binding sites with automatically extracted geometrical templates. *Nucleic Acids Res.* **2006**, *34*, W303–W309.
(109) Ivanisenko, V. A.; Pintus, S. S.; Grigorovich, D. A.; Kolchanov, N. A. PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Res.* **2005**, *33*, D183–D187.
(110) Gong, S.; Yoon, G.; Jang, I.; Bolser, D.; Dafas, P.; Schroeder, M.; Choi, H.; Cho, Y.; Han, K.; Lee, S.; Choi, H.; Lappe, M.; Holm, L.; Kim, S.; Oh, D.; Bhak, J. PSIbase: a database of protein structural interactome map (PSIMAP). *Bioinformatics* **2005**, *21*, 2541–2543.
(111) Gong, S.; Park, C.; Choi, H.; Ko, J.; Jang, I.; Lee, J.; Bolser, D.; Oh, D.; Kim, D.-S.; Bhak, J. A protein domain interaction interface database: InterPare. *BMC Bioinf.* **2005**, *6*, 207.
(112) Kundrotas, P. J.; Alexov, E. PROTCOM: searchable database of protein complexes enhanced with domain−domain structures. *Nucleic Acids Res.* **2007**, *35*, D575–D579.
(113) Gabdoulline, R. R.; Wade, R. C.; Walther, D. MolSurfer: a macromolecular interface navigator. *Nucleic Acids Res.* **2003**, *31*, 3349–3351.
(114) Finn, R. D.; Marshall, M.; Bateman, A. iPfam: Visualization of protein−protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* **2005**, *21*, 410–412.
(115) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a Web-accessible database of experimentally determined protein−ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.
(116) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.
(117) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: collection of binding affinities for protein−ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977–2980.
(118) Block, P.; Sotriffer, C. A.; Dramburg, I.; Klebe, G. AffinDB: a freely accessible database of affinities for protein−ligand complexes from the PDB. *Nucleic Acids Res.* **2006**, *34*, D522–D526.
(119) Zhang, J.; Aizawa, M.; Amari, S.; Iwasawa, Y.; Nakano, T.; Nakata, K. Development of KiBank, a database supporting structure-based drug design. *Comput. Biol. Chem.* **2004**, *28*, 401–407.
(120) Hu, L.; Benson, M. L.; Smitch, R. D.; Lener, M. G.; Carlson, H. A. Binding MOAD (Mother of All Databases). *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 333–340.
(121) Michalsky, E.; Dunkel, M.; Goede, A.; Preissner, R. Superligands, a database of ligand structures derived from the Protein Data Bank. *BMC Bioinf.* **2005**, *6*, 122.
(122) Shin, J.-M.; Cho, D.-H. PDB-ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures. *Nucleic Acids Res.* **2005**, *33*, D238–D241.
(123) *Omega*; OpenEye: Santa Fe, NM.
(124) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H. M.; Westbrook, J. Ligand depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* **2004**, *20*, 2153–2155.
(125) Feldman, H. J.; Snyder, K. A.; Ticoll, A.; Pintilie, G.; Hogue, C. W. V. A complete small molecule dataset from the Protein Data Bank. *FEBS Lett.* **2006**, *580*, 1649–1653.
(126) Kleywegt, G. Crystallographic refinement of ligand complexes. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2007**, *63*, 94–100.
(127) Diago, L. A.; Morell, P.; Aguilera, L.; Moreno, E. Setting up a large set of protein−ligand PDB complexes for the development and validation of knowledge-based docking algorithms. *BMC Bioinf.* **2007**, *8*, 310.

(128) Noguchi, T.; Akiyama, Y. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. *Nucleic Acids Res.* **2003**, *31*, 492–493.
(129) Nederveen, A. J.; Doreleijers, J. F.; Vranken, W.; Miller, Z.; Spronk, C. A. E. M.; Nabuurs, S. B.; Guentert, P.; Livny, M.; Markley, J. L.; Nilges, M.; Ulrich, E. L.; Kaptein, R.; Bonvin, A. M. J. J. RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins: Struct., Funct., Bioinf.* **2005**, *59*, 662–672.
(130) Irwin, J. J.; Shoichet, B. K. ZINC, a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(131) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—what can we learn from earlier mistakes? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 213–228.
(132) Levy, E. D. PiQSi: protein quaternary structure investigation. *Structure (Cambridge, MA)* **2007**, *15*, 1364–1367.
(133) Diemand, A. V.; Scheib, H. iMolTalk: an interactive, internet-based protein structure analysis server. *Nucleic Acids Res.* **2004**, *32*, W512–W516.
(134) Balamurugan, B.; Roshan, M. N.; Shaahul Hameed, B.; Sumathi, K.; Senthilkumar, R.; Udayakumar, A.; Venkatesh Babu, K. H.; Kalaivani, M.; Sowmiya, G.; Sivasankari, P.; Saravanan, S.; Vasuki Ranjani, C.; Gopalakrishnan, K.; Selvakumar, K. N.; Jaikumar, M.; Brindha, T.; Michael, D.; Sekar, K. PSAP: protein structure analysis package. *J. Appl. Crystallogr.* **2007**, *40*, 773–777.
(135) Laskowski, R. A. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.* **2001**, *29*, 221–222.
(136) Laskowski, R. A. Enhancing the functional annotation of PDB structures in PDBsum using key figures extracted from the literature. *Bioinformatics* **2007**, *23*, 1824–1827.
(137) Autin, L.; Tuffery, P. PMG: online generation of high-quality molecular pictures and storyboarded animations. *Nucleic Acids Res.* **2007**, *35*, W483–W488.
(138) Nagano, N. EzCatDB: the enzyme catalytic-mechanism database. *Nucleic Acids Res.* **2005**, *33*, D407–D412.
(139) von Grotthuss, M.; Plewczynski, D.; Ginalski, K.; Rychlewski, L.; Shakhnovich Eugene, I. PDB-UF: database of predicted enzymatic functions for unannotated protein structures from structural genomics. *BMC Bioinf.* **2006**, *7*, 53.
(140) Lai, Y.-L.; Yen, S.-C.; Yu, S.-H.; Hwang, J.-K. pKNOT: the protein knot Web server. *Nucleic Acids Res.* **2007**, *35*, W420–424.
(141) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536–540.
(142) Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH, a hierarchic classification of protein domain structures. *Structure (Cambridge, MA)* **1997**, *5*, 1093–1108.
(143) Schneider, R.; de Daruvar, A.; Sander, C. The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.* **1997**, *25*, 226–230.
(144) Schomburg, I.; Chang, A.; Schomburg, D. BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.* **2002**, *30*, 47–49.
(145) Camon, E.; Magrane, M.; Barrell, D.; Lee, V.; Dimmer, E.; Maslen, J.; Binns, D.; Harte, N.; Lopez, R.; Apweiler, R. The Gene Ontology Annotation (GOA) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Res.* **2004**, *32*, D262–266.
(146) Wheeler, D. L.; Barrett, T.; Benson, D. A.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; DiCuccio, M.; Edgar, R.; Federhen, S.; Geer, L. Y.; Kapustin, Y.; Khovayko, O.; Landsman, D.; Lipman, D. J.; Madden, T. L.; Maglott, D. R.; Ostell, J.; Miller, V.; Pruitt, K. D.; Schuler, G. D.; Sequeira, E.; Sherry, S. T.; Sirotkin, K.; Souvorov, A.; Starchenko, G.; Tatusov, R. L.; Tatusova, T. A.; Wagner, L.; Yaschenko, E. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **2007**, *35*, D5–12.
(147) Martin, A. C. R. PDBSprotEC: a Web-accessible database linking PDB chains to ec numbers via SwissProt. *Bioinformatics* **2004**, *20*, 986–988.
(148) Martin, A. C. R. Mapping PDB chains to UniProtKB entries. *Bioinformatics* **2005**, *21*, 4297–4301.
(149) Jambon, M.; Andrieu, O.; Combet, C.; Deleage, G.; Delfaud, F.; Geourjon, C. The SuMo server: 3D search for protein functional sites. *Bioinformatics* **2005**, *21*, 3929–3930.
(150) Guda, C.; Pal, L. R.; Shindyalov, I. N. DMAPS: a database of multiple alignments for protein structures. *Nucleic Acids Res.* **2006**, *34*, D273–D276.
(151) Hendlich, M. Databases for protein−ligand complexes. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1998**, *D54*, 1178–1182.
(152) Hendlich, M.; Bergner, A.; Gunther, J.; Klebe, G. Relibase: design and development of a database for comprehensive analysis of protein−ligand interactions. *J. Mol. Biol.* **2003**, *326*, 607–620.
(153) Ausiello, G.; Zanzoni, A.; Peluso, D.; Via, A.; Helmer-Citterich, M. pdbFun: mass selection and fast comparison of annotated PDB residues. *Nucleic Acids Res.* **2005**, *33*, W133–W137.
(154) Ferre, F.; Ausiello, G.; Zanzoni, A.; Helmer-Citterich, M. SURFACE: a database of protein surface regions for functional annotation. *Nucleic Acids Res.* **2004**, *32*, D240–D244.
(155) Sayle, R. A.; Milner-White, E. J. RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.* **1995**, *20*, 374–376.
(156) Taeubig, H.; Buchner, A.; Griebsch, J. PAST: fast structure-based searching in the PDB. *Nucleic Acids Res.* **2006**, *34*, W20–W23.
(157) Via, A.; Peluso, D.; Gherardini Pier, F.; de Rinaldis, E.; Colombo, T.; Ausiello, G.; Helmer-Citterich, M. 3dLOGO: a Web server for the identification, analysis and use of conserved protein substructures. *Nucleic Acids Res.* **2007**, *35*, W416–W419.
(158) Wernersson, R.; Rapacki, K.; Staerfeldt, H.-H.; Sackett, P. W.; Molgaard, A. Featuremap3d, a tool to map protein features and sequence conservation onto homologous structures in the PDB. *Nucleic Acids Res.* **2006**, *34*, W84–W88.
(159) Wang, G.; Dunbrack, R. L., Jr. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res.* **2005**, *33*, W94–W98.
(160) Fanton, M.; Floris, M.; Frau, G.; Sturlese, M.; Masciocchi, J.; Palla, P.; Cedrati, F.; Rodriguez-Tomé, R.; Moro, S. In *MMsINC: A New Public Large-Scale Chemoinformatics Database System*, Biotechno 2008 IEEE Proceedings, Bucharest, Romania, 2008; IEEE Computer Society Press: Washington, DC, 2008; pp 64−69.
(161) Marti-Renom Marc, A.; Pieper, U.; Madhusudhan, M. S.; Rossi, A.; Eswar, N.; Davis Fred, P.; Al-Shahrour, F.; Dopazo, J.; Sali, A. DBAli tools: mining the protein structure space. *Nucleic Acids Res.* **2007**, *35*, W393−7.
(162) Cotesta, S.; Stahl, M. The environment of amide groups in protein−ligand complexes: H-bonds and beyond. *J. Mol. Model.* **2006**, *12*, 436–444.
(163) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337–356.
(164) Ozrin, V. D.; Subbotin, M. V.; Nikitin, S. M. PLASS: protein−ligand affinity statistical score, a knowledge-based force-field model of interaction derived from the PDB. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 261–270.
(165) Brameld, K. A.; Kuhn, B.; Reuter, D. C.; Stahl, M. Small molecule conformational preferences derived from crystal structure data. A medicinal chemistry focused analysis. *J. Chem. Inf. Model.* **2008**, *48*, 1–24.
(166) Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *B58*, 380–388.
(167) *Catalyst*,; Accelrys: San Diego, CA.
(168) Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. Comparative analysis of protein-bound ligand conformations with respect to Catalyst's conformational space subsampling algorithms. *J. Chem. Inf. Model.* **2005**, *45*, 422–430.
(169) Kirchmair, J.; Wolber, G.; Laggner, C.; Langer, T. Comparative performance assessment of the conformational model generators Omega and Catalyst: a large-scale survey on the retrieval of protein-bound ligand conformations. *J. Chem. Inf. Model.* **2006**, *46*, 1848–1861.
(170) Li, J.; Ehlers, T.; Sutter, J.; Varma-O'Brien, S.; Kirchmair, J. CAESAR: a new conformer generation algorithm based on recursive buildup and local rotational symmetry consideration. *J. Chem. Inf. Model.* **2007**, *47*, 1923–1932.
(171) Kirchmair, J.; Ristic, S.; Eder, K.; Markt, P.; Wolber, G.; Laggner, C.; Langer, T. Fast and efficient in silico 3D screening: toward maximum computational efficiency of pharmacophore-based and shape-based approaches. *J. Chem. Inf. Model.* **2007**, *47*, 2182–2196.
(172) Sakae, Y.; Okamoto, Y. Optimization of protein force-field parameters with the Protein Data Bank. *Chem. Phys. Lett.* **2003**, *382*, 626–636.
(173) Wu, D.; Jernigan, R.; Wu, Z. Refinement of NMR-determined protein structures with database derived mean-force potentials. *Proteins: Struct., Funct., Bioinf.* **2007**, *68*, 232–242.
(174) Prasad, T.; Subramanian, T.; Hariharaputran, S.; Chaitra, H. S.; Chandra, N. Extracting hydrogen-bond signature patterns from protein structure data. *Appl. Bioinf.* **2004**, *3*, 125–135.
(175) Holm, L.; Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **1993**, *233*, 123–138.
(176) Bachar, O.; Fischer, D.; Nussinov, R.; Wolfson, H. A computer vision based technique for 3-D sequence-independent structural comparison of proteins. *Protein Eng.* **1993**, *6*, 279–287.
(177) Shindyalov, I. N.; Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **1998**, *11*, 739–747.

(178) Janin, J.; Rodier, F.; Chakrabarti, P.; Bahadur, R. P. Macromolecular recognition in the Protein Data Bank. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2007**, *D63*, 1–8.
(179) An, J.; Totrov, M.; Abagyan, R. Comprehensive identification of "druggable" protein ligand binding sites. *Genome Inf. Ser.* **2004**, *15*, 31–41.
(180) Hare, B. J.; Walters, W. P.; Caron, P. R.; Bemis, G. W. Cores: an automated method for generating three-dimensional models of protein/ligand complexes. *J. Med. Chem.* **2004**, *47*, 4731–4740.
(181) Vallat, B. K.; Pillardy, J.; Elber, R. A template-finding algorithm and a comprehensive benchmark for homology modeling of proteins. *Proteins: Struct., Funct., Bioinf.* **2008**, *72*, 910–928.
(182) Schlegel, B.; Laggner, C.; Meier, R.; Langer, T.; Schnell, D.; Seifert, R.; Stark, H.; Höltje, H.-D.; Sippl, W. Generation of a homology model of the human histamine h3 receptor for ligand docking and pharmacophore-based screening. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 437–453.
(183) Arnold, K.; Bordoli, L.; Kopp, J.; Schwede, T. The SWISS-MODEL workspace: a Web-based environment for protein structure homology modelling. *Bioinformatics* **2006**, *22*, 195–201.
(184) Holm, L.; Sander, C. Database algorithm for generating protein backbone and side-chain co-ordinates from a C[alpha] trace: application to model building and detection of co-ordinate errors. *J. Mol. Biol.* **1991**, *218*, 183–194.
(185) Maupetit, J.; Gautier, R.; Tuffery, P. SABBAC: online structural alphabet-based protein backbone reconstruction from alpha-carbon trace. *Nucleic Acids Res.* **2006**, *34*, W147–W151.
(186) Kosloff, M.; Kolodny, R. Sequence-similar, structure-dissimilar protein pairs in the PDB. *Proteins: Struct., Funct., Bioinf.* **2008**, *71*, 891–902.
(187) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
(188) Fousteris, M. A.; Papakyriakou, A.; Koutsourea, A.; Manioudaki, M.; Lampropoulou, E.; Papadimitriou, E.; Spyroulias, G. A.; Nikolaropoulos, S. S. Pyrrolo[2,3-*a*]carbazoles as potential cyclin dependent kinase 1 (CDK1) inhibitors. Synthesis, biological evaluation, and binding mode through docking simulations. *J. Med. Chem.* **2008**, *51*, 1048–1052.
(189) Rollinger, J. M.; Schuster, D.; Baier, E.; Ellmerer, E. P.; Langer, T.; Stuppner, H. Taspine: bioactivity-guided isolation and molecular ligand−target insight of a potent acetylcholinesterase inhibitor from magnolia × soulangiana. *J. Nat. Prod.* **2006**, *69*, 1341–1346.
(190) Frederick, R.; Robert, S.; Charlier, C.; Wouters, J.; Masereel, B.; Pochet, L. Mechanism-based thrombin inhibitors: design, synthesis, and molecular docking of a new selective 2-oxo-2*H*-1-benzopyran derivative. *J. Med. Chem.* **2007**, *50*, 3645–3650.
(191) Jozwiak, K.; Ravichandran, S.; Collins, J. R.; Moaddel, R.; Wainer, I. W. Interaction of noncompetitive inhibitors with the $\alpha 3\beta 2$ nicotinic acetylcholine receptor investigated by affinity chromatography and molecular docking. *J. Med. Chem.* **2007**, *50*, 6279–6283.
(192) Coupez, B.; Lewis, R. A. Docking and scoring theoretically easy, practically impossible. *Curr. Med. Chem.* **2006**, *13*, 2995–3003.
(193) Wolber, G.; Seidel, T.; Bendix, F.; Langer, T. Molecule−pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discovery Today* **2008**, *13*, 23–29.
(194) Schuster, D.; Nashev, L. G.; Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T.; Odermatt, A. Discovery of nonsteroidal 17β-hydroxysteroid dehydrogenase 1 inhibitors by pharmacophore-based screening of virtual compound libraries. *J. Med. Chem.* **2008**, *51*, 4188–4199.
(195) Steindl, T. M.; Crump, C. E.; Hayden, F. G.; Langer, T. Pharmacophore modeling, docking, and principal component analysis based clustering: combined computer-assisted approaches to identify new inhibitors of the human rhinovirus coat protein. *J. Med. Chem.* **2005**, *48*, 6250–6260.
(196) Charlier, C.; Henichart, J.-P.; Durant, F.; Wouters, J. Structural insights into human 5-lipoxygenase inhibition: combined ligand-based and target-based approach. *J. Med. Chem.* **2006**, *49*, 186–195.
(197) Rella, M.; Rushworth, C. A.; Guy, J. L.; Turner, A. J.; Langer, T.; Jackson, R. M. Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors. *J. Chem. Inf. Model.* **2006**, *46*, 708–716.
(198) Spitzer, G. M.; Wellenzohn, B.; Laggner, C.; Langer, T.; Liedl, K. R. DNA minor groove pharmacophores describing sequence specific properties. *J. Chem. Inf. Model.* **2007**, *47*, 1580–1589.
(199) Steindl, T. M.; Schuster, D.; Wolber, G.; Laggner, C.; Langer, T. High-throughput structure-based pharmacophore modelling as a basis for successful parallel virtual screening. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 703–715.
(200) Steindl, T. M.; Schuster, D.; Laggner, C.; Langer, T. Parallel screening: a novel concept in pharmacophore modeling and virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 2146–2157.

JM8005977