doi:10.1016/S0022-2836(02)01408-0 J. Mol. Biol. (2003) 326, 607–620

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Relibase: Design and Development of a Database for Comprehensive Analysis of Protein–Ligand Interactions

Manfred Hendlich[1,2], Andreas Bergner[3], Judith Günther[1] and Gerhard Klebe[1]*

[1]*Institute for Pharmaceutical Chemistry, Philipps-University of Marburg, Marbacher Weg 6 35032 Marburg, Germany*

[2]*Lion Bioscience ChemInformatics, Waldhoferstr. 98, 69123 Heidelberg, Germany*

[3]*Cambridge Crystallographic Data Centre, 12 Union Road Cambridge CB2 1EZ, UK*

**Corresponding author*

Knowledge discovery from the exponentially growing body of structurally characterised protein–ligand complexes as a source of information in structure-based drug design is a major challenge in contemporary drug research. Given the need for powerful data retrieval, integration and analysis tools, Relibase was developed as a database system particularly designed to handle protein–ligand related problems and tasks. Here, we describe the design and functionality of the Relibase core database system. Features of Relibase include, e.g. the detailed analysis of superimposed ligand binding sites, ligand similarity and substructure searches, and 3D searches for protein–ligand and protein–protein interaction patterns. The broad range of functions provided in Relibase and its high level of data integration, along with its flexible and intuitive interface, makes Relibase an invaluable data mining tool which can significantly enhance the drug development process. An example, illustrating a 3D query for quarternary ligand nitrogen atoms interacting with aromatic ring systems in proteins, a pattern found in pharmaceutically relevant target proteins such as, e.g. acetylcholine-esterase, is discussed.

Keywords: structure-based drug design; protein–ligand interaction; 3D database; data mining

Introduction

The Protein Data Bank (PDB)[1,2] currently contains more than 18,000 structures, and is growing exponentially at an average rate of more than 45 new entries per week. Due to the expected influx of new structures as a result of the increasing interest in structural proteomics, it is estimated that the PDB could grow to more than 35,000 structures by 2005. Much effort is going into unifying and cleaning PDB data to improve it and to eliminate inconsistencies.[3,4] This rapid increase in data, accompanied by the correction of erroneous data will significantly extend the value of 3D protein structures for drug development. However, the benefit of 3D structural data increasingly depends upon effective data management systems capable of handling the data in an integrated and flexible way. In particular, the retrieval of chemically relevant information such as ligand substructures, 3D interaction patterns, and similar ligand binding sites from the PDB can still be difficult.

Relibase is an object-oriented database system, which has been designed to facilitate data mining for protein–ligand related information. It is essentially a data retrieval system built up of several software components and modules, consisting of functionality such as data enhancement and validation, generation of derived knowledge-based data, and a graphical user interface (GUI). Here, the Relibase core module and its GUI are described, and the functionality is illustrated with

We dedicate this paper to Professor J. D. Dunitz.

Present address: M. Hendlich, Lion Bioscience, ChemInformatics, Waldhoferstr. 98, 69123 Heidelberg, Germany.

Abbreviations used: CSD, Cambridge Structural Database; GUI, graphical user interface; PDB, Protein Data Bank; RMS, root-mean-square; XML, extended markup language.

E-mail address of the corresponding author: klebe@mailer.uni-marburg.de

selected examples. The current approach to pre-processing PDB data will also be addressed.

In addition to the Relibase core functionality, several modules are currently under development. These modules will enable handling and comparison of ligand-binding sites (CavBase module),[5] the description, analysis and validation of water molecules (WaterBase module), and the assignment and validation of protonation states and hydrogen-bond networks (HydroBase module). A module handling crystallographic packing effects around ligand-binding sites has been developed recently.[6] The GUI is complemented by a Python-based toolkit, which allows interactive querying *via* a simple scripting language. This will be described elsewhere. Relibase has also been used in the development of new knowledge-based methods such as the scoring function DrugScore.[7]

Database design

The tremendous growth of experimental data in all areas of molecular biology has stimulated the development and application of a broad range of databases. Many data archives in biology, e.g. sequence databases, are organised as a flat collection of structured plain ASCII files. This appears to be adequate for small databases or for archives with simple underlying data schemes. With increasing complexity of the stored information, however, such an approach becomes ponderous, as the query performance strongly depends upon the complexity of the parsing process. Analysing structures of proteins and protein–ligand complexes by parsing thousands of PDB files is extremely inefficient. Therefore several attempts based on relational[3,8] and object oriented[9,10] database technologies have been made to define a data representation scheme for 3D structures suitable for efficient data access and retrieval.

For Relibase we decided to adopt an object oriented approach and to develop a new object oriented kernel for storing and managing 3D structures. Relibase was completely written in C++. The main components are a fast and effective search engine, which has been optimised to handle spatial data, and the object oriented persistence manager. Objects are stored in multiple archives in a B-tree data structure.[11] Efficient object access and navigation has been implemented *via* pointer swizzling.[12]

The main reason for adopting an object oriented approach was the inability of relational databases to support complex data types (e.g. topological representation of molecules and spatial data). Within Relibase many query operations such as the identification of topological and spatial patterns are performed on the fly on the topological and three-dimensional representations of molecules. The immense number of possible chemical building blocks considered in drug design makes a pre-compilation of all possible patterns impossible. A pre-calculation would significantly increase the amount of data and would impose severe restriction on the search capabilities. Full flexibility when searching for any substructure or interaction patterns between ligands and receptors is therefore essential.

In 1996, when the development of Relibase started, relational databases had no or only a very limited support for technologies to handle such complex data types (object oriented extensions) or data cartridges, which allow complex computations within the database system. A hybrid approach, mapping of complex objects to relational tables, is often very inefficient and would not allow the query performance needed for Relibase. Even with recent developments of relational database systems such as object oriented extensions, data cartridge technologies and interfaces to JAVA programming languages we doubt that the query capabilities of Relibase could be implemented in a system such as ORACLE with reasonable effort and acceptable performance.

Data processing and curation

The building of the Relibase database involves two data processing steps. In the first, all relevant information is extracted from the PDB files, and is enhanced with additional information such as atom and bond types. At present, these enhancements are carried out using the program BALI,[13] which stores all the data in a uniform XML file format. In the second step, the XML files are transformed to a Relibase-specific binary format.

Currently, BALI extracts the title, compound names, structure determination method, biological source, author, release date, resolution, *R*-factors, cell parameters and space group of the crystal, crystallisation conditions (pH), literature references, secondary structure assignments, etc. from the PDB file header section. Information that is difficult to classify, such as remarks, is retained in ASCII files representing the complete header section. Sequence information, which is also extracted from the header, is stored to provide the basis for subsequent sequence alignments.

The building of a receptor–ligand database obviously requires discrimination between protein and ligand molecules for any database entry. Although the HETATM annotations provide a simple filter for ligand molecules, they do not represent a sufficient criterion for differentiation. For example, small peptides can be annotated as ATOM or as HETATM. Apart from these inconsistencies, it remains a conceptual question, as to what type of molecules should be considered as ligand. Currently, Relibase handles three classes of molecules: protein chains, ligands, and water molecules. The obvious shortcomings of these partially arbitrary definitions will be addressed in the ongoing Relibase development.

Apart from the Relibase interface to PDB format, an interface to mmCIF format[14,15] is currently under development, which will allow utilization of the curator version and ligand chemical annotation of the PDB. This will facilitate future processing of PDB structures considerably and supersede most of BALI's functionality. However, some functionality such as the assignment of atom types will remain in the pre-processing software of Relibase.

Protein chain definition in Relibase

Any polypeptide chain consisting of 21 or more amino acid residues is by definition a molecule of type protein. Protein molecules are built up of the 20 standard amino acids only. This implies that covalent modifications of amino acids and non-standard amino acids (post-translational modifications (e.g. Tyr-phosphate, Ser-phosphate, glycosylations), Cys-oxide, selenomethionine, etc.) are separated into a standard amino acid moiety and a ligand moiety (see below).

Ligand definition in Relibase

Ligands, by definition, consist of all non-protein and non-water molecules. Thus, nucleic acids, metal ions and small organic/inorganic ions (e.g. sulphate, acetate) and organic solvent molecules (e.g. dimethyl sulphoxide, ethanol) are treated as ligands in Relibase. Peptides of up to 20 residues are also classified as ligands. Ligands can also be covalently bound to the protein (e.g. suicide inhibitors, phosphorylated or glycosylated protein-hydroxyl groups).

Water molecules in Relibase

Water molecules in the PDB can be represented as either H_2O or ^2H_2O. Since information on hydrogen atoms is usually not available for protein structures, all single oxygen atoms are considered as water molecules.

Correctly assigned atom and bond-type information is an absolute prerequisite for a tool designed to analyse protein–ligand interactions. The assignment of an element type to an atom, as provided by PDB files, is not sufficient on its own to enable meaningful analysis of intramolecular structure or the geometry and energetics of non-bonded interactions.

In addition to unification of the format, the program BALI was designed to assign bond orders and atom types for small molecules. BALI calculates the atomic connectivity of a ligand, and assigns atom and bond types according to the notation given in Sybyl (Tripos Inc.). The approach is based on a series of heuristic rules for analysing the local environment and topology of any atom (bond distances, valence and torsion angles, ring planarity, etc.). This way of characterising a ligand is independent of any names or three-letter codes given in the PDB file. It therefore allows for an unambiguous recognition of identical chemical compounds. Since the success rate of BALI is ~85%, a template library of unique ligand building blocks is used to assign atom and bond types to new ligands before they are added to the database. All new building blocks undergo a manual check before being added to the template library. This mechanism ensures a high quality of data for the ligand descriptions in Relibase.

However, as mentioned earlier, a HET-atom dictionary for all available PDB structures has been released recently. This will render much of the BALI functionality obsolete in due course.

Relibase data structure

All PDB data pre-processed by BALI is stored in the form of XML files. As XML is a standard database exchange format, these files could also serve as input to other database applications. The Relibase database is built up from XML files, which are transformed to binary B-tree files. Although hierarchical relations (e.g. PDB → protein-chain → residue → atom → coordinates) are already apparent in the XML files, all cross-links between objects and global object containers are generated in this step, according to the internal organisation of the Relibase object class library.

In order to ensure efficient data access and query performance, 3D information (coordinates) is stored in independent data blocks separated from any other information such as chemical and textual annotations. This implies the discrimination between the chemical composition of a ligand, i.e. its template, and all individual ligand molecules appearing in different binding sites or entries. Although ligand templates share the same connectivity, chemical formula, 2D diagram representation and molecular weight, they usually have of course different 3D coordinates, *B*-factors and solvent accessibilities. Relibase handles the different information types by using different classes and database objects, inter-related by pointers only. This concept avoids storage of redundant information and allows faster querying. For example, Relibase conducts a two-step process to carry out ligand substructure searches. A fast pre-filtering step is used to pre-select potential hits. It performs a bitvector match based on topological fingerprints, for the ligand templates only. Fingerprints have been computed by hashing all non-overlapping paths through the ligand molecules up to length of eight bonds (Daylight User Manual), and are pre-calculated and stored in the database. All pre-selected hits are then subjected to a sub-graph matching algorithm, which performs the actual substructure identification.

Another approach for increasing the speed of queries is the storage as a separate database object of the binding site environment within a 7 Å radius around every ligand. Although these coordinates

could be calculated on request, the high percentage of potential Relibase queries that would include this calculation step makes their pre-calculation worthwhile. Sequence alignments are another pre-calculated feature in Relibase. Analogous to the ligand templates, identical sequences refer to a unique reference sequence. The alignment of sequences, contained in the sequence archive, is performed using FastA.[16]

The graphical user interface (GUI)

The Relibase database search engine is front-ended by an intuitive web-based graphical user interface (GUI), thus allowing decentralised access to the Relibase database search engine *via* any web browser supporting Java (JDK 1.1 and higher) and JavaScript. All GUI frames are HTML pages built up dynamically on request by a series of CGI scripts written in Perl. Further GUI components comprise Java-based applets, particularly the query sketcher MolEd. Internally, the GUI translates all interactively built queries into a Relibase specific query language, which is then communicated to the Relibase server.

Relibase features simple queries such as keyword searches, ligand substructure and similarity searches, and sequence based searches. More complex operations use a tool for superimposition and analysis of binding sites across a series of proteins sharing a significant level of sequence similarity. A Java-based query sketcher (MolEd) allows 3D substructure-based queries to be set up, providing a powerful tool to search for particular interaction patterns between protein, ligand and/or water moieties. Basic geometric constraints can be applied easily. Depending on the type of query, different unified frame layouts are used to represent the query results. Single frame pages are designed to show detailed information about particular protein entries or ligands. These include tools for visualisation, chemical and textual annotation, a list of ligands contained in the entry, and links to related information in the database. In addition to textual annotation, all ligands are depicted as 2D diagrams which have been generated by Cactvs.[17] Lists of protein entry codes or ligands are used to represent results consisting of multiple hits. All elements in these lists point to detailed protein or ligand information pages *via* hyperlinking. Query results can be stored permanently as sets of ligands or PDB entry codes (hit lists), and a hit list manager allows for cutting down and combining sets of results. At present, Rasmol[18] is used as a visualisation tool running locally on the client site. Interaction with Rasmol is achieved by using Java components in the GUI (visualiser toolboxes), and a Perl (UNIX) or Python (PC) control program operating as a helper application assigned to the browser.

The GUI was designed to allow maximum flexibility in setting up and combining queries. Relibase achieves its flexibility by hyperlinking query results to a series of potential new queries, thus providing the user with a virtual network of query paths through the entire database. Database access and triggering of queries are conducted on-request when activating hyperlinks. This concept allows the user to navigate through the PDB in a very flexible way.

Relibase query types

Keyword and text-based searches

The most basic search in Relibase is on PDB entry code, which results in a single-frame protein information page. The standard keyword search (search on HEADER, COMPND and SOURCE records of the PDB file) and author search (AUTHOR record only) performs a case insensitive (sub)string match, resulting in a list of PDB entry codes. Searches for ligand compound names (HET records) and ligand three-letter codes results in a browsable list of ligands. Searches for ligand three-letter codes retrieve the ligand itself or composed ligands containing the ligand three-letter code, e.g. "MQI" would also find "MQI-ARG-MCP".

Sequence search

Searches for similar sequences in Relibase are carried out using FastA.[16] Query sequences are defined as one-letter code strings. Any protein information page provides links to similar sequence searches for all protein chains contained in an entry. This allows the generation of PDB entry sets containing all related protein chains exhibiting a given range of sequence identity. Relibase features a pre-calculated sequence alignment database comprising all entries stored in the PDB, allowing fast querying for related protein chains. Since annotations in many PDB files are inconsistent, sequence-based searches, starting from a protein information frame, provide the best way of retrieving comprehensive sets of proteins belonging to a particular family, rather than using text-based query tools. However, these problems should lessen with utilisation of the PDB curator version.

SMILES based substructure searches

Ligand substructure searches can be carried out using SMILES strings.[19] Due to certain requirements in Relibase, some modifications have been made to the standard Daylight SMILES† strings.

- Relibase does not support SMILES syntax for defining isotopes, charges, and

† www.daylight.com

stereochemistry (chiral atoms and *cis–trans* isomerism).
- Relibase SMILES supports wild card atoms (A, any atom; R, any atom except hydrogen; X, any atom except hydrogen or carbon) and bonds (~, bond type "any").
- Aromatic bonds are only supported for six-membered rings. Five-membered rings must be defined by using single and double bonds, or bonds of type "any".
- Hydrogen atoms can be added to fill up valencies, rather than actually defining hydrogen atoms. This allows terminating chains in the query substructure, and is particularly useful when searching for distinct structures rather than substructures.

Similar ligand searches

Any ligand information page can be used as starting point to retrieve a list of similar ligands. Relibase calculates a 2D similarity index (Tanimoto coefficient), and compares all ligands stored in the database with the reference ligand and generates a sorted list. Since 2D fingerprints are pre-calculated for all ligands, this approach is fast. Ligands occurring only once in the database are hyperlinked directly to the corresponding ligand information frame. However, ligands present in multiple PDB entries are hyperlinked to a list of ligands, which are then linked individually to the corresponding ligand information pages.

MolEd-based searches

The Relibase GUI comprises a Java applet (MolEd), which can be used to search for structural motifs based on user-defined substructures and their geometric relationship. MolEd allows queries to be set up by drawing a substructure or interaction pattern in an interactive and flexible way. Any valid MolEd query is composed of one or more substructures, each defined as an ensemble of interconnected atoms and bonds. In accordance with the basic molecule types known to Relibase, any substructure represents a protein or ligand moiety, or a water molecule. If more than one substructure is present, constraints must be applied to describe the geometric inter-relationship between the substructures. Substructure drawings can also be built using template structures stored in a library, or by drag and drop sketching using a mouse. GUI components to define, modify, select and delete atoms, bonds and constraints are available. MolEd supports wild card type atoms (X, not carbon, hydrogen; R, not hydrogen), and a bond type "any". This allows the user to search for a broader range of substructures and avoids problems with tautomeric states, which cannot be handled by Relibase. Geometric constraints correlating different query substructures comprise distances, angles and torsion angles. Centroids generated from a selected set of atoms and normal vectors of planes can be used to set up more complex query motifs and geometric constraints. An atom selection mechanism allows the user to superimpose hits subsequent to data retrieval; at present, only three atoms belonging to one of the query substructures can be selected and subsequently used for superimposition. A hyperlink is provided to visualise the superimposed structures. This type of scatter plot[20] shows the spatial distribution of the query fragments, and provides insights into the preferred protein–ligand binding geometry. Queries involving any distance, angle or torsion angle constraints will automatically produce a histogram showing the distribution of all geometric properties.

Typical MolEd queries

Search for a particular ligand substructure

This type of query is very similar to SMILES-based searches for ligand substructures. It consists of a single query substructure of molecule type ligand. However, in contrast to SMILES-based searches, internal constraints can be defined. For example, this can be used to restrict the conformational space of a rotatable bond by applying a torsion angle constraint to four atoms.

Search for a particular protein–ligand interaction motif

MolEd enables the user to set up queries for specific protein–ligand binding motifs consisting of any number of protein and ligand substructure fragments inter-related by geometric constraints. Relibase uses a hierarchical search strategy to speed up the retrieval of hits. Fast pre-filtering techniques are applied before carrying out more CPU intensive steps such as exact substructure matching and computation of geometric properties.

Primarily, the substructure of the ligand, which is drawn first, is used to pre-select potential hits using the substructure match mechanism applied to ligand substructure searches. Secondly, the pre-processed information about this ligand-binding site environment is retrieved from the database, and used for pre-filtering the remaining substructures. Further ligand-type substructures (number of ligands = 2, 3…) are investigated by substructure matching to all additional ligands in the binding site. Protein-type substructures are pre-filtered by analysing the occurrence of amino acids in the ligand-binding site which match with the query substructure. Exact substructure matching is only performed after successful pre-filtering. At the moment, only single amino acid substructures including N and C-terminal one-atom extensions to the adjacent amino acid can be processed. However, queries for contiguous protein motifs can be

defined by multiple definition of single amino acid residues in conjunction with distance constraints interrelating the C and N-terminal tails of the amino acid substructures. More complex constraints such as centroids and normal vectors are evaluated after substructure matching of the query components. At this intermediate stage, potential hits are represented by spatial positions linked to the constraint descriptors. Finally, these constraints are evaluated using a recursive algorithm testing compliance with the numerical values of the constraints. Protein–ligand query motifs can be extended by defining water molecules, which are involved in the binding pattern, e.g. by bridging the protein–ligand contact. Depending on the complexity of a query and the number of initial hits obtained from the ligand screen, such queries can usually be accomplished within a few seconds to ca ten minutes. Even generic queries (wild card atoms, very small ligand fragments) involving most of the complexes stored in the database can be processed in a reasonable time (usually <30 minutes on an SGI O2 R12000). Queries can be restricted to pre-processed hit lists, thus shortening the processing time.

Search for protein–protein interactions

MolEd also allows the user to set up queries for particular protein–protein or protein–water interactions. However, these queries are time consuming, since most of the pre-filtering and speed up techniques cannot be applied.

Superimposition and analysis of similar binding sites

Comparison of structures after 3D superimposition is a standard task in molecular modelling. Almost all molecular modelling programs feature tools to superimpose molecules by minimising the RMS deviation between pairs of atoms. However, most programs require some manual intervention to select topologically equivalent atoms for superimposition, particularly when treating homologous protein chains of different length and composition. A tool capable of automated superimposition of protein chains is FSSP†.

Relibase also provides an interactive tool for the superimposition of ligand binding sites onto a reference binding site by overlaying homologous protein chains. Moreover, a detailed analysis classifying differences and reporting invariances amongst these binding sites is carried out after superimposition. The approach can be initiated from any ligand information page, which defines the reference binding site structure. Initially, a protein chain in close proximity (<7 Å) to the ligand has to be selected as a reference protein chain. A list of homologous protein chains is retrieved from the pre-processed sequence alignment database. Relibase allows the user to interactively select protein chains of interest for further processing. Any of the selected chains is aligned with the reference chain using the program ALIGN.[21] Aligned positions not exhibiting any insertions and deletions are extracted from the ALIGN result. The C^{α} atom positions of the corresponding residues are used for a first rough superimposition (overall superimposition). For a second, refined superimposition, only 60% of the C^{α} atoms showing the lowest RMS deviations of superimposed C^{α} atom pairs are used (core superimposition). The transformation matrix resulting from this overlay is then applied to the entire structure. Optionally, the superimposition can be restricted to the binding site C^{α} atoms only. In this case, the second superimposition of all complexes is used to detect conserved residues amongst the binding sites, which are subsequently used for a final third superimposition (binding site superimposition). All superimposition steps are carried out automatically, independent of any non-automated interventions.

Superimposed complexes can be inspected visually using a Java-based visualiser toolbox. Global structural differences between all superimposed binding sites are summarised in a table. The table reports the RMS deviation values of the overall, core and binding site superimpositions. Occurrences of protein flexibility, mutations and insertions, ligand overlaps, clashes, and conserved water positions are counted and reported in the cells of the table. The threshold values can be adjusted interactively, allowing the user to redraw the overview table and to trim the layout according to the particular case. Any table cell is hyperlinked to an information page providing more details of the analysis.

Protein flexibility is classified in terms of backbone and side-chain movements. Backbone

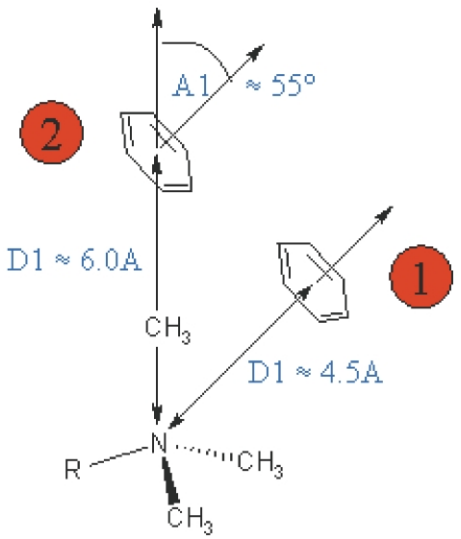


Figure 1. Preferred interaction geometries for trimethyl-ammonium groups and phenyl rings in small molecule crystal data, as observed by Verdonk *et al.*[27]

† http://www.cmbi.kun.nl/swift/fssp/paper.html

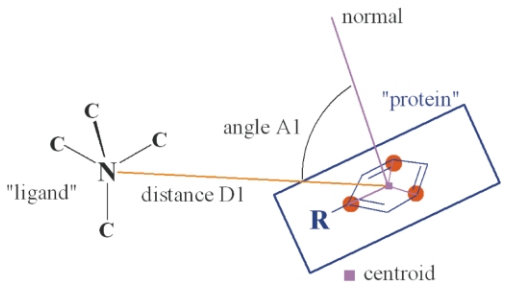


Figure 2. Query set-up to search for ligands with tetra alkyl-ammonium (onium) groups interacting with phenyl rings in protein side-chains. A nitrogen atom bound to four carbon atoms was sketched to define a tetramethyl-ammonium substructure (black atoms indicate molecule of type ligand). The interaction partner, a phenyl ring of type protein, is shown on the right-hand-side. A non-hydrogen atom (atom type R) has been added to the ring to reduce the number of substructure matches, and to define a reference frame for superimpositions. The centroid of the ring and the normal vector of the ring were defined to assist with the geometrical description of the interaction pattern. A distance constraint *D*1 between the ring centroid and the ligand nitrogen atom was defined (upper limit 7 Å). The angle *A*1 spanned by the normal vector and the vector pointing from the ring centroid to the ligand nitrogen atom was used to characterise the preferred geometries of the interaction patterns. Atoms depicted in red refer to atoms used as a reference for superimposition of the hits.

movements are reported depending on the RMS deviations from within C^{α} atom pairs (default threshold: 0.5 Å). Two criteria are used to detect and analyse side-chain movements: the RMSD value of the side-chain atoms centroid has to exceed a threshold (default: 1.0 Å), or any torsion angle defining the side-chain conformation must deviate by more than a threshold value (default: 10°) from the reference. Ligand overlaps are computed by calculating the intersecting ligand volumes and approximated using Gaussian functions. Ligands in close proximity to any moiety in the reference are reported as clashes. Water molecules are considered to be located in conserved positions if the deviation from the corresponding water in the reference structure is $\leq$1.2 Å.

By default, all amino acid residues within a 4 Å radius around the superimposed ligand binding sites are considered in the analysis described above. In addition, Relibase reports significant C^{α} atom positional shifts of larger protein segments beyond this sphere. By increasing the radius, larger regions around binding sites can be surveyed and the relevant aspects of protein flexibility can be examined. This is particularly useful when seeking or investigating large-scale domain movements amongst a series of related protein structures.

Permanent storage of query results

Results from most queries can be stored in hit lists. Hit lists are B-tree archive files permanently stored on the Relibase server site. At present, Relibase supports two types of objects handled by hit lists: PDB entry codes (protein type) and individual ligand molecules (ligand type). Keyword and author searches, sequence searches and 3D-MolEd queries, not containing any ligand moiety, result in protein-type hit lists. Queries for ligand names, three-letter codes, SMILES substructure searches and all MolEd queries, in which at least one ligand substructure is defined, result in ligand-type hit lists. A hit list manager allows the user to combine hit lists of one type by applying boolean operators AND, OR and NOT. In addition, protein-type hit lists can be translated into ligand-type hit lists and *vice versa*. Any object in a hit list can be individually selected and removed, merged into an already existing hit list, or used to generate a new hit list. Most queries can be restricted to search a given hit list only, providing the user with a flexible and effective tool to limit results, combine queries, and restrict the search space.

Database content

According to the ligand definition given above, and neglecting DNA and RNA type ligands, Relibase (version 1.0, release date September 2001) consists of 4443 different unique ligands present in one or more entries. The most frequently occurring ligands are inorganic ions such as sulphate (3597), calcium (3318), zinc (2412), magnesium (1791) and chloride (1468). Any single binding site of X-ray structures was considered, and NMR structures are considered to comprise one ligand only. There are 3959 purely organic ligands (CHNOSP atoms only), 169 inorganic compounds (e.g. sulphate, phosphate, ammonium, Fe–S cluster), 45 different metal cations (Ca, Na), 13 non-metal single atom (chloride, bromide, xenon), and 180 organo-metallic compounds (heme, co-cyanocobalamine, cacodylate, boron-organic compounds, etc.). The most commonly occurring organic ligands comprise, e.g. hem (1405 binding sites), NAG (1223), glycerol (795), NAD (416), FAD (348), acetate (344), ADP (299) and pyridoxal-5′-phosphate (264). Analysing the ligand set according to Lipinski's rule of five[22] (and neglecting the ClogP criterion), 2097 different ligands bound to 14,268 ligand binding sites can be considered as being drug-like. Due to MIR (multiple isomorphous replacement) and MAD (multiple anomalous dispersion) phasing techniques applied in protein crystallography, a large variety of chemical elements not naturally occurring in biomolecules are also present in the PDB.

Analysis of protein–ligand interaction patterns

A comprehensive overview of application examples covering most of the functionality implemented in Relibase, will be given in a

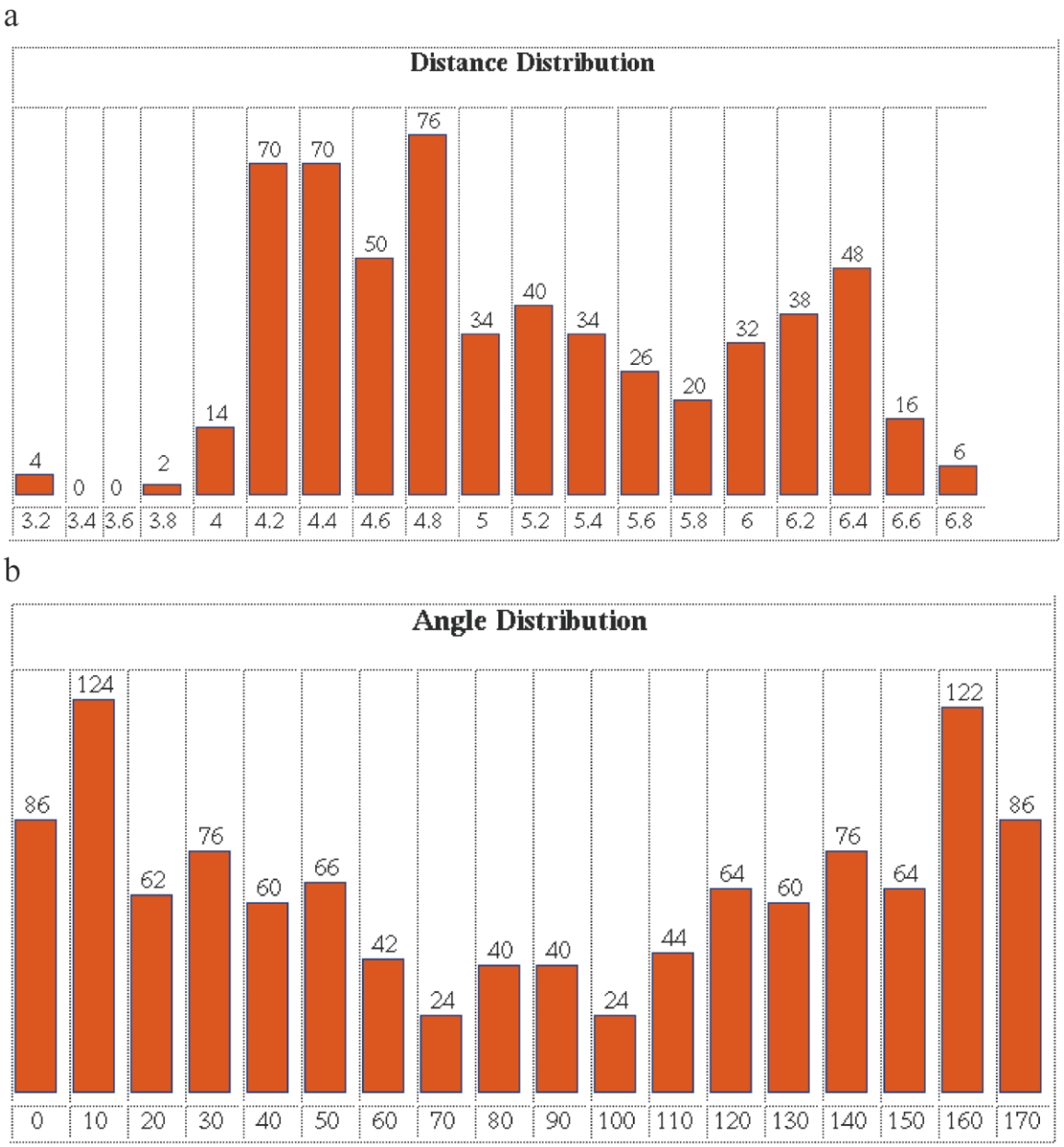


Figure 3 (*legend opposite*)

subsequent paper.[23] Here, we demonstrate the Relibase 3D-query tools used to investigate protein–ligand recognition patterns. Our example examines the interactions between charged nitrogen atoms (onium groups) of ligands and the aromatic rings of protein side-chains. The motifs found are compared with the closely related protein–protein interaction patterns comprising aromatic rings and Lys or Arg side-chains.

Charged nitrogen–aromatic interactions

Analyses of acetylcholine binding to a synthetic receptor[24] or the binding mode of quaternary ammonium ligands to acetylcholinesterase[25] revealed that quaternary nitrogen atoms could interact favourably with the electron-rich π-systems of aromatic rings. Searching the CSD[26] for such interaction patterns provided strong evidence for this.[27] A related analysis on PDB data was performed by Burley & Petsko.[28] In this study, 33 high-resolution crystal structures were searched for, inter alia, charged Lys or Arg residues interacting with the aromatic rings of Tyr, Trp or Phe side-chains.

To our knowledge, a systematic survey on interactions between onium groups and aromatic rings in protein–ligand complexes has not yet been performed. We used the Relibase 3D query tool, MolEd, to search for these patterns in PDB protein–ligand

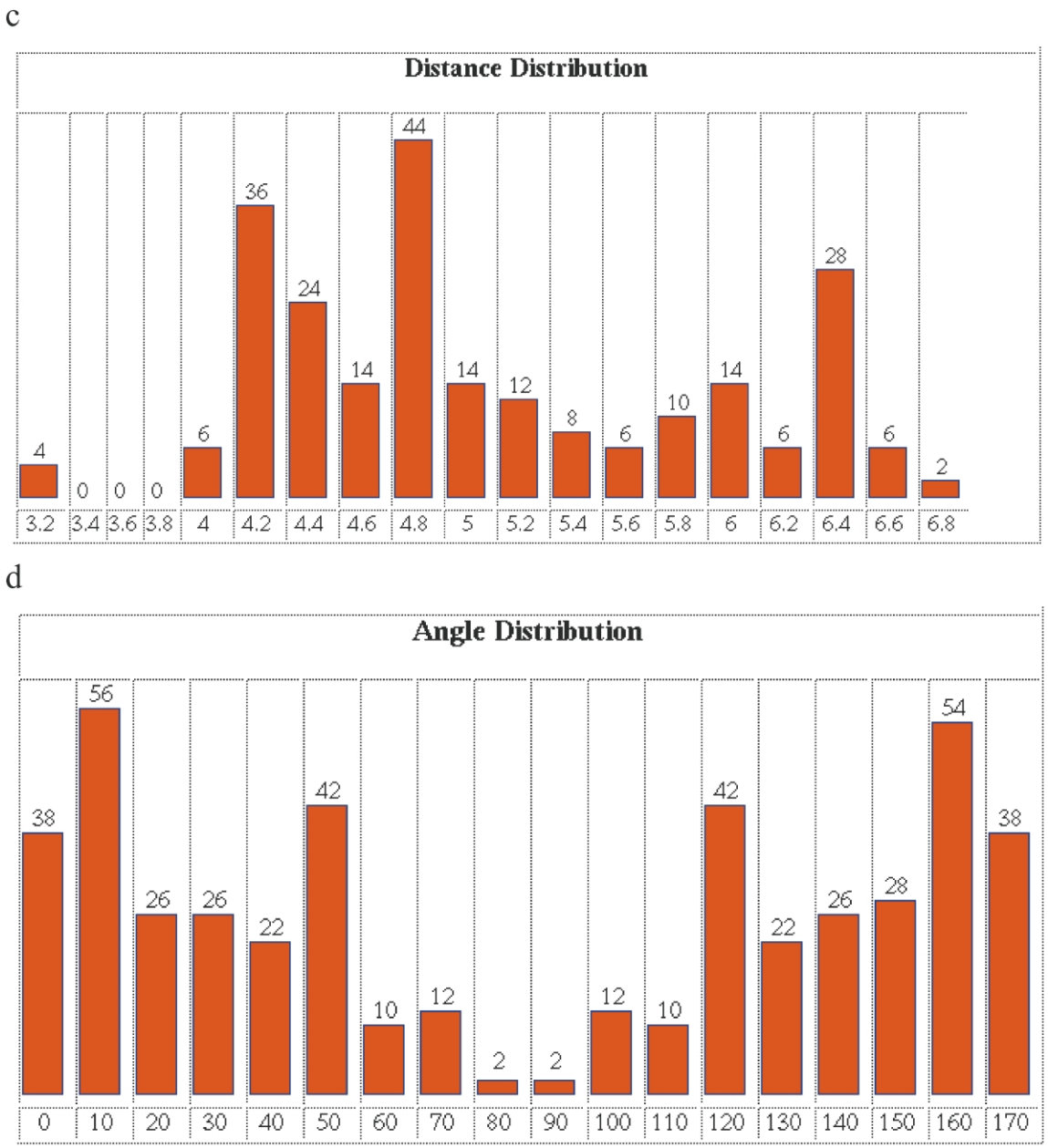
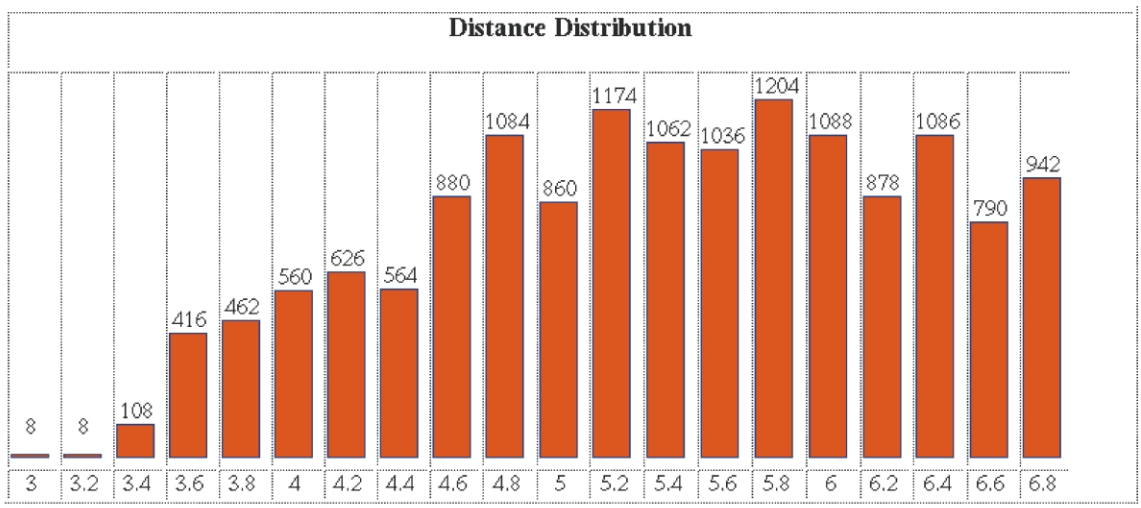


Figure 3. Histograms of the distance (*D*1) and angular (*A*1) distributions as defined in the query set-up (Figure 2). The first two diagrams (a and b) result from a query including all structures with resolution ≤2.5 Å, while for the latter two diagrams (c and d) onium ligands containing alkyl side-chains with ten or more methylene units were excluded.

data, and compared the results with the observations found in small molecule data.[27] The study by Verdonk *et al*.[27] revealed two preferred interaction patterns represented by two distinctive clusters (see Figure 1). In the first and preferred orientation (1), the ligand nitrogen atoms are located above the planes of the phenyl ring about 4.5 Å away from the ring centroid, equidistant to the atoms of the ring. In the second, tilted orientation (2), the plane of the phenyl ring is arranged parallel to one of the N–C bonds. In this case, the preferred distance between the ring-centroid and the nitrogen atom of the onium group is about 6 Å.

Our query (see Figure 2) consisted of a nitrogen atom bound to four carbon atoms (a tetramethylammonium substructure, all atoms of type "ligand") and a phenyl ring of type "protein". A non-hydrogen atom (atom type "R") has been attached to this ring substructure to reduce the number of substructure matches, and to define a reference frame for superimpositions. The distance (*D*1) between the ring centroid and the nitrogen atom of the ligand was restricted to 7 Å. The normal vector of the plane of the ring was defined, and the angle (*A*1) spanned by this normal vector and the vector pointing from the ring centroid to the ligand nitrogen atom was used to characterize the preferred geometries of the interaction patterns. In this query set up, Verdonk's orientation (1) corresponds to a distance $D1 = 4.5$ Å

a



b

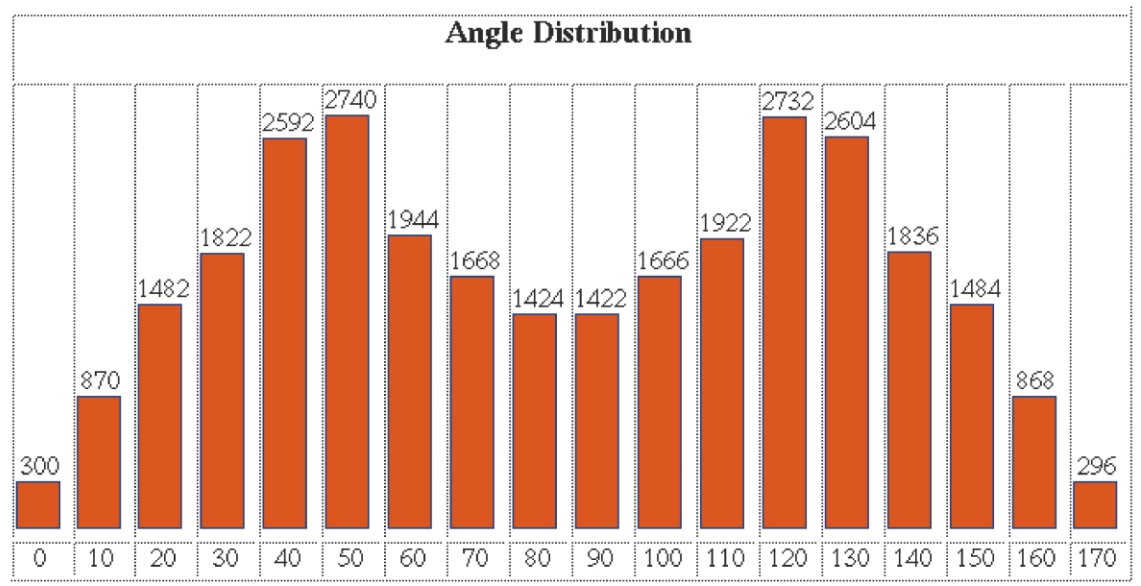


Figure 4 (*legend opposite*)

and an angle $A1 = 0°/180°$, whereas orientation (2) is represented by distance $D1 = 6.0$ Å and an angle $A1 = 50°/120°$, respectively.

Currently, Relibase contains 161 ligands comprising quaternary ammonium groups, 62 of which are located within a distance of 7.0 Å† from the centroids of six-membered aromatic rings (Phe, Tyr and Trp residues). Fortyfive ligands were found within 7.0 Å around five-membered rings (His and Trp residues). The distance ($D1$) distribution shows a broad maximum ranging from 4.2 Å to 4.8 Å, and a fairly flat maximum around 6.4 Å. However, values were widely distributed in the range between 4.0 Å and 7.0 Å (see Figure 3). Furthermore, the angle ($A1$) distribution showed no clear preference for the expected orientations, and, in particular, an additional maximum around $A1 = 90°$ was found. Thus, at a first glance, the orientations reported in Verdonk's study could not be clearly identified amongst the hits. A detailed analysis, restricted to structures with resolution ≤ 2.5 Å only (39 hits found amongst 131 ligands), revealed a strong bias by structures consisting of ligands with one or more long alkyl-tails bound to the charged nitrogen atom. The 39 hits were split up into a set of ligands exhibiting a decyl-group tail or longer (ten hits), and all others (29 hits). The long alkyl-tail ligands are found in, for example, mycolic acid cyclo-propane synthase

† From CSD data it was found that the phenyl rings have a random position and orientation at distances ≥ 7.3 Å.

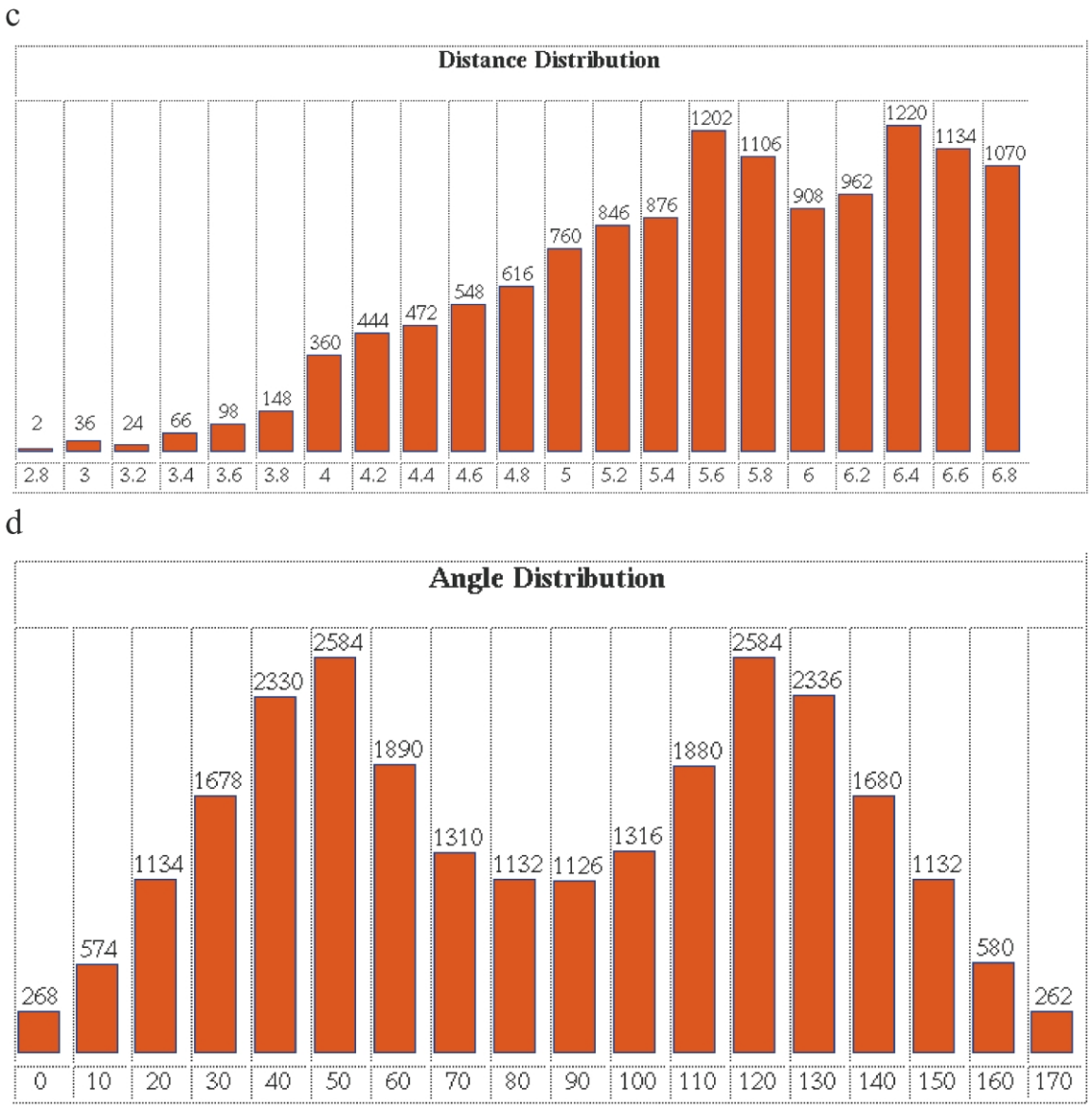


Figure 4. Histograms of the distances ($D1$) and angular ($A1$) distributions between charged nitrogen atoms and phenyl rings in protein–protein interactions, defined in analogous fashion to the query set-up in Figure 2. The first two diagrams (a and b) depict preferred interaction geometries between phenyl rings and the centres of guanidino groups in Arg side-chains, the latter two (c and d) represent the interaction geometries observed between phenyl rings and Lys N^{ε} atoms.

structures (1kph, 1kpg) and tyrosine phosphatase (2shp). The ligand binding sites of these structures are lined by a number of aromatic side-chains, and up to four aromatic rings interacted with the ligand onium groups in different orientations. All structures representing the unexpected situation with $A1 = 90°$ belong to this group.

Analysing the interaction motifs of the remaining 29 hits revealed two preferred interaction patterns in accordance with the study by Verdonk *et al*. The maxima at $D1 = 4.2/4.8$ Å (d_{max1}) and 6.4 Å (d_{max2}) appear slightly sharper in the distance distribution (see Figure 3). The angle $A1$ distribution exhibited two major maxima at 0°–10°/170°–180° (a_{max1}), two smaller but significant maxima around $A1 = 50°/120°$ (a_{max2}), and no observations around 90°. Further analysis clearly shows d_{max1} cohering with a_{max1}, and d_{max2} with a_{max2}, respectively. Such an analysis can be easily carried out by constraining $D1$ to values around a given maxima, thereby generating the respective angle distribution. The slight shift of a_{max1} from 0°/180° towards 10°/170° can be attributed to interactions with the Trp indole ring systems, for which the centre of the π-electron density is shifted towards the five-membered ring (e.g. ACHE 1maa). Other deviations can be explained by simultaneous interactions with several phenyl rings, e.g. ferritin 1rcd. The superimposition of the hits (see Figure 5) showed a stacking of the phenyl rings

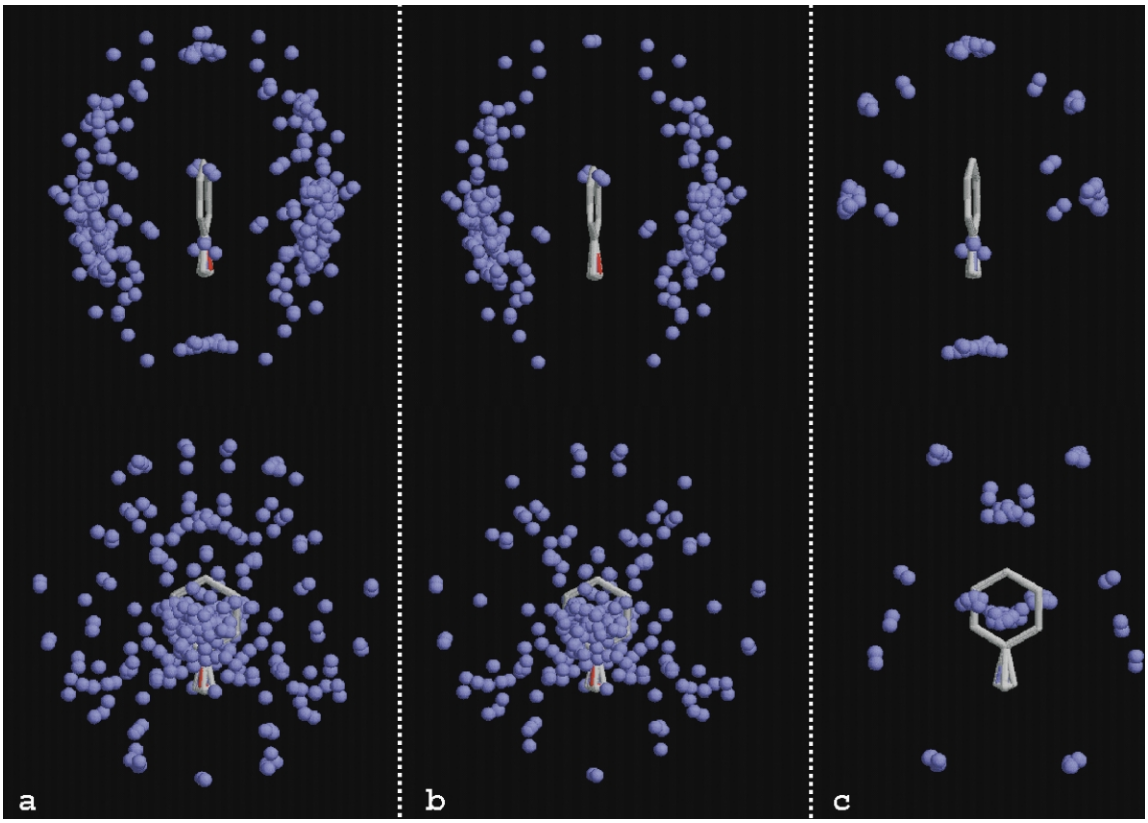


Figure 5. Preferred interaction pattern between ligands possessing a tetra alkyl-ammonium (onium) group and phenyl rings of Tyr, Phe or Trp side-chains. The rings were superimposed upon retrieval of the hits, resulting in the scatter plots shown. Only the quaternary nitrogen atoms of the ligands are depicted as blue balls. (a) Scatter plot comprising all hits found within a distance $D1 = 7.0$ Å (resolution ≤2.5 Å, 39 hits). (b) Scatter plot as described in (a) excluding all hits comprising long alkyl-tails (decyl-groups or longer). The preferred clustering of the 29 hits as found in the CSD can be easily detected. (c) Scatter plot comprising only long alkyl-tail hits (10). These hits form a cluster representing a preferred orientation not previously reported. The combination of scatter plots (b) and (c) would result in (a). The Figure was generated using Rasmol.[18]

with some variance in the contact distance to the ligand.

As the majority of the ligands interact with at least two aromatic rings simultaneously, hybrids of both orientations are very common. Ligands within *D*1 contact distances of 5.0 Å–5.5 Å seem to belong to orientation (2), however, a clear separation was not possible.

Generally, the preferred interaction patterns described by Verdonk *et al.* could be confirmed by this study, and, in addition to the patterns previously reported, a series of onium ligands are found adopting binding modes not yet detected. Thus, onium ligands apparently adopt a broader variety of binding modes than expected. Including further structures with a resolution >2.5 Å results in a more blurred distribution. Protein–protein interaction motifs, relating to the protein–ligand patterns discussed above, comprise Lys or Arg side-chains in contact with the aromatic rings of Phe, Tyr or Trp residues. The searches for these motifs were set up in analogous fashion to the above protein–ligand query (see Figure 2), using the Lys N^{ε} atom and the centroid of the Arg guanidinium group as reference points for defining the distance constraint *D*1. The distance between the ring centroid and the reference points was again restricted to 7.0 Å. Due to the enormous number of protein structures elucidated since the pioneering study of Burley & Petsko,[28] the search was restricted to high-resolution structures (≤1.5 Å, 954 entries) only. A total of 625 contacts with ≤7.0 Å were found for Arg side-chains, and 645 for Lys interactions. The distance distribution is represented by an almost monotonically ascending curve between 3.0 Å and 7.0 Å (see Figure 4) not showing any significant maxima. While the distributions are fairly similar for both Arg and Lys, Arg side-chains form significantly more short contacts, ≤4.0 Å, to phenyl rings. Many of these represent stacking interactions that can be formed between the π-face of a guanidinium group of Arg and the aromatic ring system, but are clearly impossible with Lys side-chains. The importance of this type of interaction has been pointed out in previous studies.[29,30] The angle (*A*1) distribution for both Arg and Lys shows distinct maxima around 50°/120°, a local minimum around 90°,

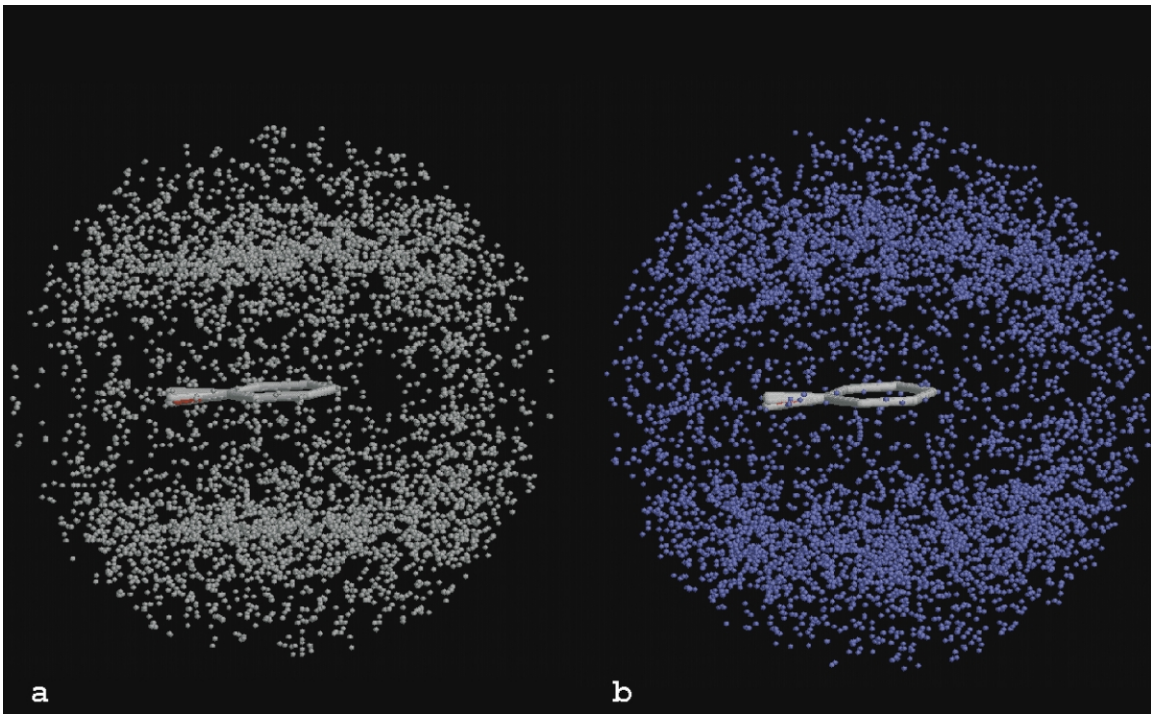


Figure 6. Preferred geometrical pattern of Arg (a) and Lys (b) side-chains interacting with the phenyl rings of Tyr, Phe or Trp residues. The aromatic rings were superimposed after retrieval of the hits. Only the C^{x} atoms of Arg (a) and the N^{ε} atoms of Lys (b) are shown as white and blue balls, respectively. A total of 954 structures with a resolution $\leq$1.5 Å were included in the study, resulting in 625 (a) and 645 (b) hits. No clustering as distinct as in the protein–ligand search (Figure 5) was observed, however, the hits cluster in a distance of 4.5 Å − 6.0 Å away from the phenyl ring, forming a layer representing preferred contact distances. The Figure was generated using Rasmol.[18]

and a deep minimum around 0°. The few hits around 0° show a clear preference for a distance *D*1 around 4.2 Å, in accordance with Verdonk's orientation (1). However, no clear distance preference could be detected for angle values of *A*1 above 20° (40° for Arg). This shows, that in frontal arrangements of Lys or Arg side-chains pointing towards phenyl rings, small contact distances cumulate, whereas tilted orientations are allowed for a broader range of contact distances. Motifs as described by Verdonk, in particular orientation (1), are rare amongst protein–protein contacts, but could be observed. Generally, the position of Lys and Arg side-chains relative to phenyl rings appear to be less distinct with respect to the angle parameter *A*1, compared with the protein–ligand interaction pattern. However, a preferred distance from the ring plane (4.5 Å–6.0 Å) could be clearly observed, appearing as a bulky layer of interaction points in the scatter plot (see Figures 5 and 6).

Conclusion

We have presented the design and development of Relibase, a database system designated for easy handling and analysing protein–ligand data. A comprehensive overview on the features and analysis tools of Relibase was given, which, alongside with two selected 3D query examples, illustrates the power of Relibase as a state-of-the-art data mining tool in structural biology and drug discovery software.

Availability

Relibase is accessible on the web from http://relibase.ccdc.cam.ac.uk, http://relibase.ebi.ac.uk, or http://relibase.rutgers.edu

Acknowledgements

We gratefully thank Karen Lipscomb and Dr Robin Taylor for valuable comments and carefully reading the manuscript. The present project has been supported by the German Minister of Science and Education (bmb + f) in the framework of the ReLiMo project (grant no. 0311619). We thank all partners in this project for a fruitful and successful collaboration.

References

1. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr, Brice, M. D. *et al.* (1977). The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur. J. Biochem.* **80**, 319–324.

2. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
3. Bhat, T. N., Bourne, P., Feng, Z., Gilliland, G., Jain, S., Ravichandran, V. *et al.* (2001). The PDB data uniformity project. *Nucl. Acids Res.* **29**, 214–218.
4. Westbrook, J., Feng, Z., Jain, S., Bhat, T. N., Thanki, N., Ravichandran, V. *et al.* (2002). The Protein Data Bank: unifying the archive. *Nucl. Acids Res.* **30**, 245–248.
5. Schmitt, S., Hendlich, M. & Klebe, G. (2001). From structure to function: a new approach to detect functional similarity among proteins independent from sequence and fold homology. *Angew. Chem. Int. Ed. Engl.* **40**, 3141–3144.
6. Bergner, A., Günther, J., Hendlich, M., Klebe, G. & Verdonk, M. L. (2002). Use of Relibase for retrieving complex 3D interaction patterns including crystallographic packing effects. *Biopol. (Nucl. Acid Sci.)*, **61**, 99–110.
7. Gohlke, H., Hendlich, M. & Klebe, G. (2000). Knowledge-based scoring function to predict protein–ligand interactions. *J. Mol. Biol.* **295**, 337–356.
8. Thornton, J. M. & Gardner, S. P. (1989). Protein motifs and data base searching. *Trends Biochem. Sci.* **14**, 300–304.
9. Gray, P. M., Paton, N. W., Kemp, G. J. & Fothergill, J. E. (1990). An object oriented database for protein structure analys. *Protein Eng.* **3**, 235–243.
10. Shindyalov, I. N., Chang, W., Pu, C. & Bourne, P. E. (1994). Macromolecular query language (MMQL): prototype data model and implementation. *Protein Eng.* **7**, 1311–1322.
11. Ladd, S. R. (1994). Hungry Minds.
12. Alger, J. (1995). *Secrets of the C++ Masters*, Morgan Kaufmann Publishers, Los Altos, CA.
13. Hendlich, M., Rippman, F. & Barnickel, G. (1997). BALI: automatic assignment of bond and atom types for protein ligands in the Brookhaven Protein Databank. *J. Chem. Inf. Comput. Sci.* **37**, 774–778.
14. Westbrook, J. D. & Bourne, P. E. (2000). STAR/mmCIF: an ontology for macromolecular structure. *Bioinformatics*, **16**, 159–168.
15. Bourne, P. E., Berman, H. M., McMahon, B., Watenpaugh, K. D., Westbrook, J. & Fitzgerald, P. M. D. (1997). The macromolecular CIF dictionary (mmCIF). *Methods Enzymol.* **277**, 571–590.
16. Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
17. Ihlenfeld, W., Takahashi, Y., Abe, H. & Sasaki, S. (1994). Computation and management of chemical properties in CACTVS: an extensible networked approach toward modularity and compatibility. *J. Chem. Inf. Comput. Sci.* **34**, 109–116.
18. Sayle, R. A. & Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.* **20**, 374.
19. Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36.
20. Bruno, I. J., Cole, J. C., Lommerse, J. P., Rowland, R. S., Taylor, R. & Verdonk, M. L. (1997). IsoStar: a library of information about nonbonded interactions. *J. Comput. Aided Mol. Des.* **11**, 525–537.
21. Myers, E. & Miller, W. (1988). Optimal alignements in linear space. *CABIOS*, **4**, 11–17.
22. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23**, 3–25.
23. Günther, J., Bergner, A., Hendlich, M. & Klebe, G. (2003). Utilising structural knowledge in drug design strategies—applications using Relibase. *J. Mol. Biol.*
24. Dougherty, D. A. & Stauffer, D. A. (1990). Acetylcholine binding by a synthetic receptor: implications for biological recognition. *Science*, **250**, 1558–1560.
25. Harel, M., Schalk, I., Ehret-Sabatier, L., Bouet, F., Goeldner, M., Hirth, C. *et al.* (1993). Quaternary ligand binding to aromatic residues in the active-site gorge of acetylcholinesterase. *Proc. Natl Acad. Sci. USA*, **90**, 9031–9035.
26. Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F. *et al.* (1991). The development of versions 3 and 4 of the Cambridge structural database system. *J. Chem. Inf. Comput. Sci.* **31**, 535.
27. Verdonk, M. L., Boks, G. J., Kooijman, H., Kanters, J. A. & Kroon, J. (1993). Stereochemistry of charged nitrogen-aromatic interactions and its involvement in ligand–receptor binding. *J. Comput. Aided Mol. Des.* **7**, 173–182.
28. Burley, S. K. & Petsko, G. A. (1986). Amino–aromatic interactions in proteins. *FEBS Letters*, **203**, 139–143.
29. Mitchell, J. B., Nandi, C. L., McDonald, I. K., Thornton, J. M. & Price, S. L. (1994). Amino/aromatic interactions in proteins: is the evidence stacked against hydrogen bonding? *J. Mol. Biol.* **239**, 315–331.
30. Mitchell, J. B., Nandi, C. L., Ali, S., McDonald, I. K., Thornton, J. M., Price, S. L. & Singh, J. (1993). Amino–aromatic interactions. *Nature*, **366**, 413.

Edited by R. Huber

(Received 1 August 2002; received in revised form 19 November 2002; accepted 25 November 2002)