Acta Cryst. (1998). D**54**, 1178–1182

Databases for Protein–Ligand Complexes

Manfred Hendlich

Institute for Pharmaceutical Chemistry, University of Marburg, Marbacher Weg 6, 35032 Marburg, Germany. E-mail: hendlich@mailer.uni-marburg.de

(*Received 6 March 1998; accepted 18 May 1998*)

Abstract

Recent advances in experimental techniques have led to an enormous explosion of available data about protein–ligand complexes. To exploit the information that is hidden in these large data, collection tools for managing and accessing huge data collections are needed. This paper discusses databases for protein–ligand data which are accessible *via* the World Wide Web. A strong focus is placed on the ReLiBase database system which is a new three-dimensional database for storing and analysing structures of protein–ligand complexes currently deposited in the Brookhaven Protein Data Bank (PDB). ReLiBase contains efficient query tools for identifying and analysing ligands and protein–ligand complexes. Its application for structure-based drug design is illustrated.

1. Introduction

Recent advances in experimental techniques have led to a flood of data about protein–ligand complexes. An impressive example is the almost exponential growth of the Brookhaven Protein Data Bank (PDB) (Abola *et al.*, 1997) with now approximately 6500 structures and five new entries each day. This growing amount of sequence information, binding constants, thermodynamic and structural data gives us a chance to improve our understanding of the basic mechanisms of receptor–ligand interactions which is indispensable for a rational design of new drugs.

However, until very recently the large fraction of biochemical data was not accessible electronically. Getting this data *e.g.* binding constants for the parameterization of empirical scoring functions for docking programs, involved scanning literature. Even if information is archived in databases it is often difficult and time consuming to access because of very limited data-retrieval mechanisms. Analysing protein–ligand interactions in the PDB is difficult. Tools for locating ligands with specific chemical properties or complexes with specific intermolecular contacts, as implemented in the Cambridge Structure Database for small molecules (Allen *et al.*, 1991), are missing. Studies of protein ligands that search for differences in conformation between bound and free ligands (Nicklaus *et al.*, 1995) or studies of water structures in binding sites (Poornima & Dean, 1995) have always been confined to relatively small sets of compounds.

This situation has changed dramatically with the explosion of the World Wide Web (WWW). Driven by an acute need for high-quality and freely available data sets in many areas in biochemical research and by the simplicity of presenting information on the WWW we are now facing an explosion of WWW-based databases. Exploiting global expertise the process of data gathering and processing gets more and more decentralized and dispersed over the network.

In the following section a selection of databases which contain biochemical and structural information about protein ligands and protein–ligand complexes is discussed. The focus will be on the ReLiBase database system which has been developed recently by the author and co-workers (Hendlich *et al.*, 1998). ReLiBase is a novel database for three-dimensional structures of protein–ligand complexes. ReLiBase implements several tools for the efficient analysis of receptor–ligand complexes which are absent in the PDB.

2. Databases for protein ligands and protein–ligand complexes

A considerable number of network-accessible databases contain information about ligands and protein–ligand complexes to a varying degree. Most of these databases are organized as flat files or hyperlinked HTML pages with no, or very limited, query tools. However, many provide information of extreme value for modelling. Good examples include the databases for a G-protein-coupled receptor such as the GCRDb database (Kolakowski, 1994; at URL http://gcrdb.uthscsa.edu/GCRDBHOME.html), the GPCRDB database (Horn *et al.*, 1998; http://www.sander.embl-heidelberg.de/7tm/) the opioid database (http://www.opioid.umn.edu) and the Prolysis database (http://delphi.phys.univ-tours.fr/Prolysis/), a protease and protease-inhibitor database developed at the University of Tours. Besides extensive sequence, biochemical and bibliographic information these databases contain information about natural and synthetic ligands, physico-chemical properties of ligands, mutation data as well as, in some cases, three-dimensional structures of ligands. Especially interesting are

ISSN 0907-4449

Table 1. *Summary of query tools in ReLiBase*

Query algorithm in ReLiBase	Potential application
Two-dimensional substructure search	Identification of ligands with specific topological characteristics
Two-dimensional similarity search	Identification of protein ligands similar to a target molecule
Three-dimensional substructure search	Identification of protein–ligand complexes with specific contacts
Sequence similarity search (based on the *FASTA* program)	Determinations of ligand types in protein families

the large number of binding constants in the GPCRDB database.

The LIGAND database (Ligand Chemical Database for Enzyme reactions, http://www.genome.ad.jp/dbget/ligand.html; Suyama *et al.*, 1993) is a database with a focus on enzymes and metabolic compounds such as substrates, products and inhibitors. In addition LIGAND contains information about enzymatic reactions, metabolic pathways, diseases and crosslinks to various other databases. LIGAND provides simple query tools. Enzymes and metabolic compounds can be found by searching for chemical names or E.C. numbers. The LIGAND database is updated weekly.

A new database on metal ions and prosthetic groups in protein active sites is the PROMISE database (Degtyarenko *et al.*, 1997). PROMISE contains comprehensive sequence, structural, functional and bibliographic information on proteins containing prosthetic groups (http://bioinf.leeds.ac.uk/promise/). In the current release the authors have collected information for several classes of metal and porphyrin-binding proteins. Coordination geometries in bindings sites are explained in detail. PROMISE is a hierarchically organized collection of WWW documents with no query tools.

3. ReLiBase – a database for analysing protein–ligand complexes

ReLiBase (Hendlich *et al.*, 1998) is, in contrast to the data collections discussed in the preceding section, a complete data-management system with an object-

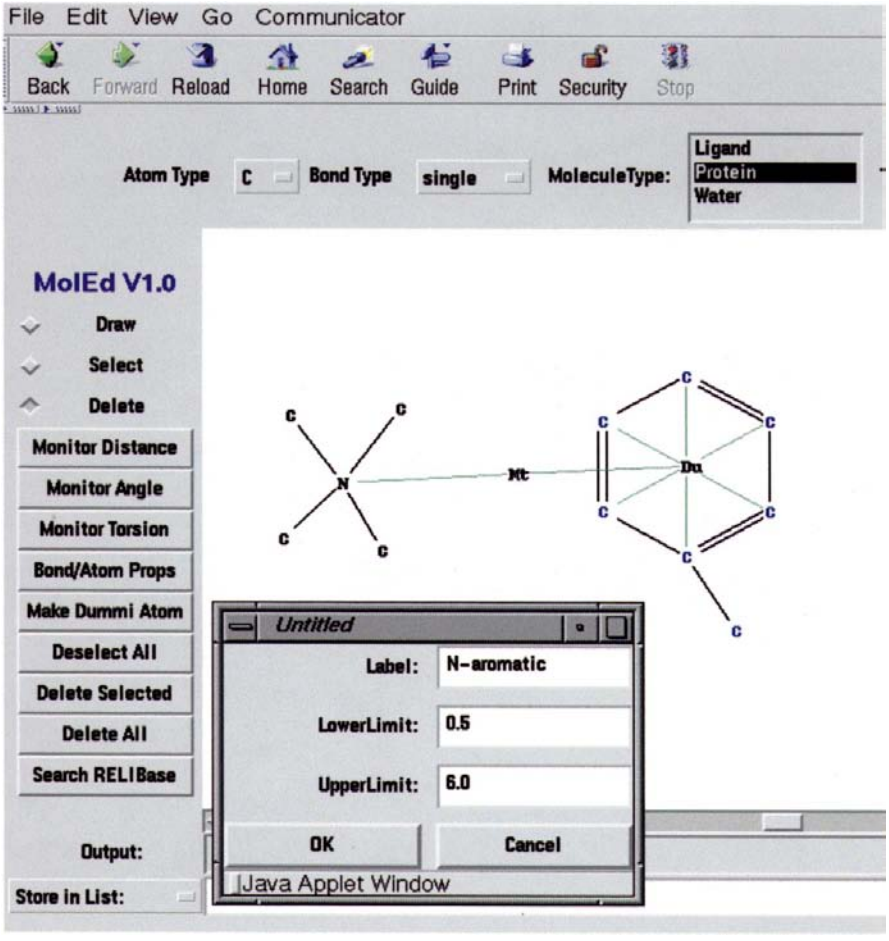

Fig. 1. Example of a ReLiBase search for ligands with quaternary N atoms in contact with phenyl rings using a distance cut-off of 6 Å.

oriented core optimized for the handling of protein–ligand structures, powerful query tools and a WWW-based interface for data visualization. The latest release of ReLiBase contains 6026 PDB protein structures. All entries have been derived from the PDB. Nucleic acids, models and entries with Cα coordinates only were not included.

ReLiBase was developed as a tool for studying binding preferences and binding geometries of protein ligands and as a practical help in modelling. The PDB has several severe deficiencies regarding the handling of non-protein molecules. First, in contrast to ReLiBase, the PDB does not include information about bond and atom types of non-protein molecules. The ability to discriminate between *e.g.* hydroxyl O atoms and carbonyl O atoms or between aromatic and saturated ring systems is essential for a detailed analysis of receptor–ligand interactions. ReLiBase contains bond and atom types for all entries according to the MOL2 convention as used in the *SYBYL* program of Tripos.

Secondly, ReLiBase implements query tools for identifying ligands and analysing receptor–ligand complexes which are not available in the PDB. In the PDB ligands can only be identified by searching for chemical names and keywords. Taking into account the complexity of chemical naming conventions this is highly ineffective *e.g.* for locating all ligands with a specific functional group. ReLiBase provides similar search tools as the Cambridge Structure Database but optimized for protein–ligand complexes. The main search tools and their applications are summarized in Table 1.

The query tools in ReLiBase are illustrated with an analysis of the preference of quaternary N atoms to be

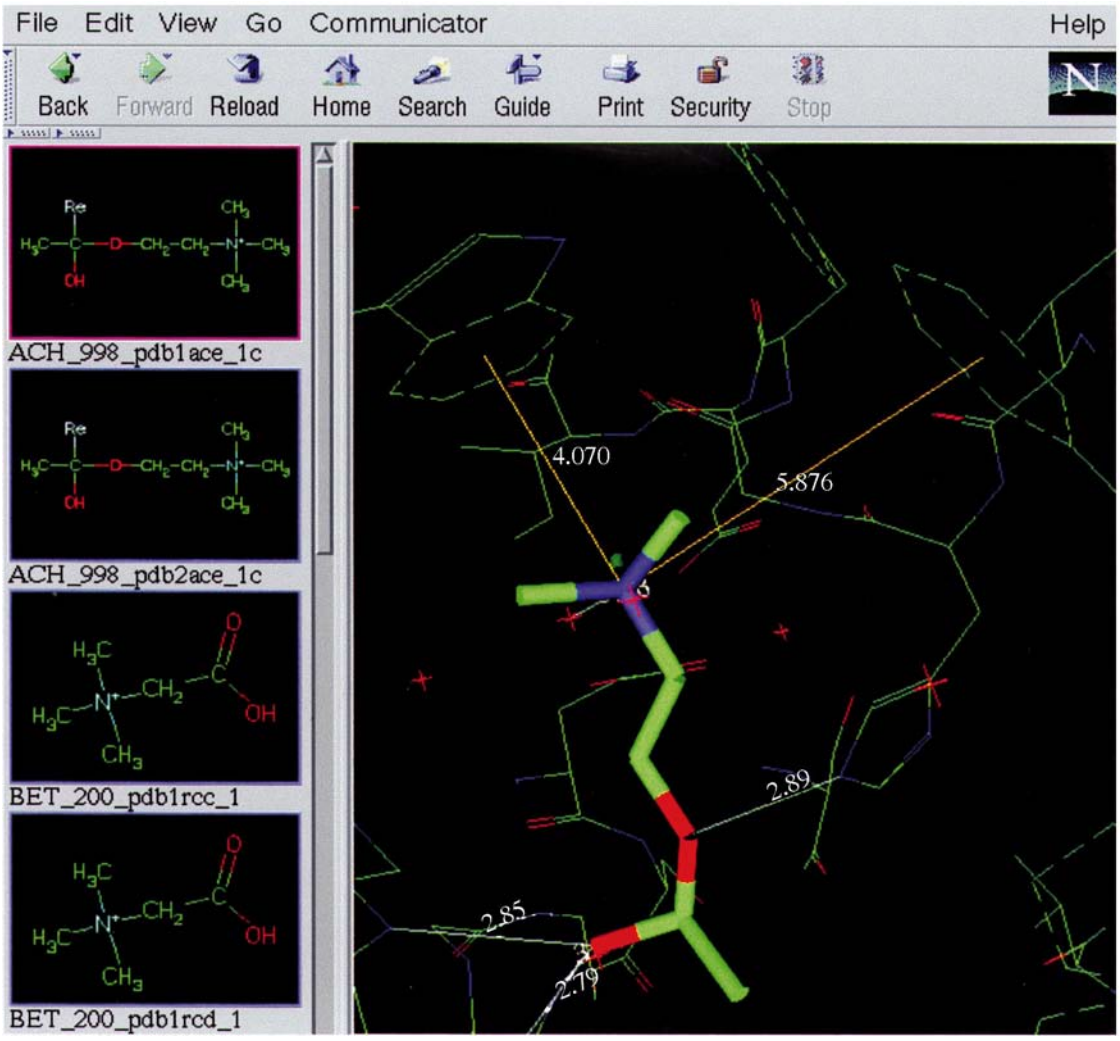

Fig. 2. Result of the ReLiBase search from Fig. 1. Four out of ten hits are displayed in the left frame. The three-dimensional model of acetylcholineesterase with acetylcholine (Sussman *et al.*, 1991) is displayed in the right frame. Distances between the N atom and the centres of the aromatic rings are indicated.

in spatial contact with the π-face of aromatic side chains in proteins. Cation–π interactions have been observed in small-molecule crystals (Verdonk *et al.*, 1993) and in several protein structures (Harel *et al.*, 1993) and have recently attracted much attention because of their unusual strength (Dougherty, 1996). Using the PDB to analyse such questions is, in principle, possible but inefficient and time consuming. With ReLiBase such an analysis can be performed within minutes. Substructures and distance restraints can be specified graphically with a JAVA-based molecule editor (Fig. 1) or with *SMILES* strings (Weininger, 1988) and submitted to ReLiBase across the WWW. ReLiBase tests all ligands for the presence of a quaternary N atom and checks all hits for aromatic side chains within the specified distance range. Of 13 ligands which contain a quaternary N atom, ten have one or more aromatic side chains within 6 Å (Fig. 2). In one complex, betain in bullfrog red-cell feritin (PDB code 1RCI), the next ring centroid is separated 6.036 Å from the quaternary N atom. In two other complexes (in both cases phosphatidylcholine lipids) the N atom points towards the solvent and makes no interaction with the protein.

The search time for this particular query was 11.0 s on a standard workstation. ReLiBase is, therefore, an extremely efficient tool for analysing structures of receptor–ligand complexes. Depending on the complexity of the query search times are normally in the range of several seconds for substructure searches to several minutes for complex three-dimensional queries.

However, the current release of ReLiBase has several limitations.

(i) ReLiBase stores structures as given in the corresponding PDB file. Interactions due to crystal contacts cannot be identified.

(ii) ReLiBase was developed for studying protein–ligand interactions. Queries regarding the packing of side chains in proteins and protein–protein contacts are in principle possible but slow and, therefore, not available on the WWW.

ReLiBase is still very much in development and driven by the acute needs of users. Several new enhancements are currently under development and will be made available within the next few months. This includes new components for analysing conformational flexibility in binding sites and tools for studying water structures in binding sites.

4. Availability of ReLiBase

A preliminary version of ReLiBase was made available to the public in July 1997 on servers at the Brookhaven Protein Data Bank (http://pdb.pdb.bnl.gov:8081/home.html) and the European Bioinformatics Institute (EBI) in Cambridge (http://www2.ebi.ac.uk:8081/home.html). An updated version of ReLiBase with improved performance and a new user interface (as shown in Fig. 1) will be installed at the PDB and the EBI in March 1998 and updated constantly.

5. Discussion and outlook

The recent explosion of experimental data about protein–ligand complexes will continue in the near future. New and improved experimental techniques will even accelerate the output of data. To exploit this rich body of data for studying and modelling of receptor–ligand complexes, databases are indispensable. Driven by this acute need for freely available data many highly specialized databases have been developed and made available on the WWW. Today information about ligands, mutations, binding and structural data is freely available in a variety of network-accessible databases which range from simple record-based flat files to complex database-management system such as ReLiBase with its powerful query tools. Many more databases will appear on the WWW within the next couple of years. However, a major drawback of this chaotic growth is the increasing dispersion of data over more and more databases with incompatible formats and missing interoperability. Methods in locating and retrieving heterogeneous information on the WWW currently rely heavily on human intervention as access to sites is by manual navigation along links. Finding all relevant information can entail reading large amounts of free text. To exploit the full potential of these databases the major goal in the next few years will be the integration of heterogeneous data and the interoperation of network-accessible databases (Kazic, 1995). To satisfy the growing needs in the next millennium the next generation of database tools in molecular and structural biology must be able to send queries to multiple network-accessible databases in a way transparent to the user. Data mining on the WWW will be orders of magnitude faster and more powerful than today.

References

Abola, E. E., Sussman, J. L., Prilusky J. & Manning, N. O. (1997). *Methods Enzymol.* **277**, 556–571.

Allen, F. A., Davies, J. E., Galloy, J. J., Johnson, J. M., Kennard, O., Macrae, C. F., Mitchell, E., Mitchel, G. F., Smith, J. M. & Watson, D. (1991). *J. Chem. Inf. Comput. Sci.* **31**, 187–204.

Degtyarenko, K. N., North, A. C. T. & Findlay, J. B. C. (1997). *Protein Eng.* pp. 183–186.

Dougherty, D. A. (1996). *Science*, **271**, 163–168.

Harel, M., Schalk, I., Ehret-Sabatier, L., Boulet, F., Goeldner, M., Hirth, C., Axelsen, P. H., Silman, I. & Sussman, J. L. (1993). *Proc. Natl Acad. Sci. USA*, **90**, 9031–9035.

Hendlich, M., Rippmann, F. & Barnickel, G. (1998). In preparation.

Horn, F., Weare, J., Beukers, M. W., Hörsch, S., Bairoch, A., Chen, W., Edvardsen, O., Campagne, F. & Vriend, G. (1998). *Nucleic Acids Res.* **26**(1), 277–281.

Kazic, T. (1995). *Proceedings of the 1995 International Chemical Information Conference*, Nimes, France, edited by H. Collier, pp. 48–61. Calne, UK: Infornotics Ltd.

Kolakowski, L. F. (1994). *Receptors Channels*, **2** 1–7.

Nicklaus, M. C., Wang, S., Driscoll, J. S. & Milne, G. W. (1995). *Bioorg. Med. Chem.* **4**, 411–428.

Poornima, C. S. & Dean, P. M. (1995). *J. Comput. Aided Mol. Des.* **6**, 500–512.

Sussman, J. L., Harel, M., Frolow, F., Oefener, C., Goldman, A., Toker, L. & Silman, I. (1991). *Science*, **253**, 872–879.

Suyama, M., Ogiwara, A., Nishioka, T. & Oda, J. (1993). *Comput. Appl. Biosci.* **9**, 9–15.

Verdonk, M. L., Boks, G. J., Kooijman, H., Kanters, J. A. & Kroon, J. (1993). *J. Comput. Aided Mol. Des.* **2**, 173–182.

Weininger, D. (1988). *J. Chem. Inf. Comput. Sci.* **28**, 31.