458–462 Nucleic Acids Research, 2003, Vol. 31, No. 1
DOI: 10.1093/nar/gkg065

E-MSD: the European Bioinformatics Institute Macromolecular Structure Database

H. Boutselakis, D. Dimitropoulos, J. Fillon, A. Golovin, K. Henrick*, A. Hussain, J. Ionides, M. John, P. A. Keller, E. Krissinel, P. McNeil, A. Naim, R. Newman, T. Oldfield, J. Pineda, A. Rachedi, J. Copeland, A. Sitnov, S. Sobhany, A. Suarez-Uruena, J. Swaminathan, M. Tagari, J. Tate, S. Tromm, S. Velankar and W. Vranken

EMBL Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received August 23, 2002; Accepted October 3, 2002

ABSTRACT

The E-MSD macromolecular structure relational database (http://www.ebi.ac.uk/msd) is designed to be a single access point for protein and nucleic acid structures and related information. The database is derived from Protein Data Bank (PDB) entries. Relational database technologies are used in a comprehensive cleaning procedure to ensure data uniformity across the whole archive. The search database contains an extensive set of derived properties, goodness-of-fit indicators, and links to other EBI databases including InterPro, GO, and SWISS-PROT, together with links to SCOP, CATH, PFAM and PROSITE. A generic search interface is available, coupled with a fast secondary structure domain search tool.

INTRODUCTION

The European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk) was established in 1995 as a centre for biological databases covering a broad range of topics from nucleotide sequence through to protein function. From its inception, the EBI has hosted the EMBL nucleotide sequence database (1), and the protein sequence database SWISS-PROT/TrEMBL (2,3). The E-MSD (http://www.ebi.ac.uk/msd) project was set up in 1996, initially as a pilot study, to create the infrastructure based on emerging relational database technologies to provide clean macromolecular structure data. The challenge of presenting the available information in an intuitive way to users from various backgrounds and expertise demands that the data are archived in a meaningful and flexible way that represents the hierarchy and constraints within the data. Relational database technology offers both the flexibility and the framework to achieve this goal. The E-MSD has applied these database technologies for the extremely complex processes of importing legacy data from the Protein Data Bank (PDB, 4), creation of a deposition system for new depositions to the PDB with automated annotation procedures, achieving data conformity and the integration of relevant information from other biological databases. A generic query system has been developed to allow access to the database. The overall system has been designed from the outset to cope with the expected exponential growth in structure data through the structural genomics initiatives (5).

The PDB search database (E-MSD)

Database framework. The search database is implemented using relational database technology, in a generic form that can be used on a variety of database engines (e.g. MySQL (6) http://www.mysql.com, Oracle, http://www.oracle.com). The organization of the structural information is hierarchical, with the topmost level corresponding to potential biological assemblies [based on the PQS (7) service, http://pqs.ebi.ac.uk], followed by the constituent polymer chains (protein and nucleic acid) and associated bound molecules. The chains are decomposed into residues and finally the constituent atoms. Derived data are added at each level of the hierarchy (accessible surface area, torsion angles etc) see Figure 1. Other data are also represented, for example, experimental and bibliographical information. Another level of organization divides the data into entry specific data (e.g. coordinates, experimental details) and reference data (data that is not specific to any particular entry, such as the chemical description of ligands and amino acids).

The search database is designed to support efficient querying and data retrieval, and therefore, contains considerable data redundancy. Its contents are derived from another database (the 'deposition database') which has a much more complex structure and lower redundancy, making it unsuitable for performing complex queries in real time. The deposition database was designed using the Oracle Designer CASE tool,

*To whom correspondence should be addressed. Email: henrick@ebi.ac.uk

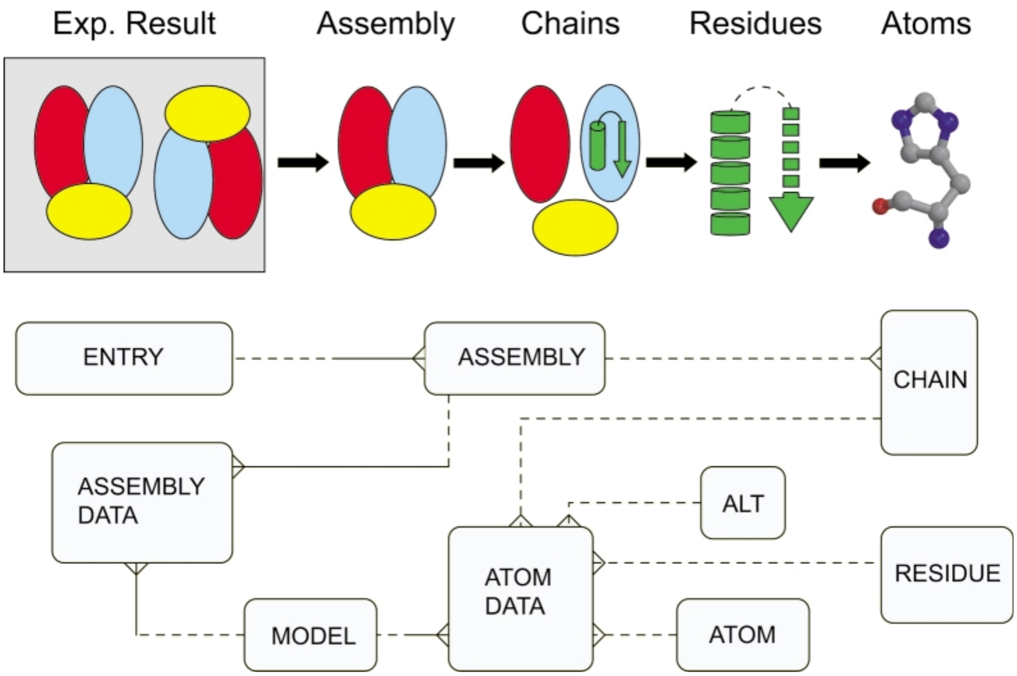

Figure 1. The E-MSD database core entity relationships. Each level of the hierarchy can have associated properties, e.g. Bound molecules, Domains definitions, Site residues, Derived properties (e.g. Accessible Surface Area), Reference information (e.g. standard geometry).

which has been invaluable for tracking the development of such a complex data model (around 400 tables linked by 1000 foreign key relationships). The maintenance of the integrity of relationships within the data is one of the guiding principles of its design.

The deposition database performs two key functions. It provides a filter that forces the legacy PDB data into a consistent framework, thus forming the basis for development of search services described below. Secondly, it is coupled to a deposition service for structural data to the PDB through AutoDep (8) (http://www.ebi.ac.uk/msd-srv/autodep), providing a versatile way of handling the depositions.

Biologically relevant organization. The quaternary structure of a protein molecule is the arrangement of its subunits in space and the ensemble of its intersubunit contacts and interactions, without regard to the internal geometry of the subunits. The quaternary state of a protein is important in understanding its biological function. For a protein structure determined using X-ray crystallography, the PDB entry describes the contents of the asymmetric unit (ASU) of the crystal. The PDB entry may, therefore, partially describe the quaternary state of the protein. The complete description of the quaternary state requires crystallographic symmetry operations to be applied to the contents of the ASU. We have developed algorithms (7,9) to determine the most likely oligomeric state, taking into account the symmetry related chains, that are used to determine the assemblies for each PDB entry and are then loaded into the database.

Inter-database consistency. To maintain consistency between the structure (E-MSD) and sequence (SWISS-PROT) databases, it is important to determine the correct sequence database cross-reference. The subsequent derived data pertaining to protein families, domains, functional sites and sequences from other databases (InterPro 10, GO 11, SCOP 12, CATH 13, PFAM 14 and PROSITE 15) are dependent on the correct SWISS-PROT database cross-reference. These data are integrated into the E-MSD search database and are made available to users via various interfaces. For new depositions, steps are taken to ensure that the SEQRES record in the PDB entry represents the correct amino acid sequence of the sample. Since many of the legacy PDB entries contain only the coordinates of the observed atom positions, it is difficult to obtain the complete sequence of the protein(s) studied. Procedures developed in the group are implemented, in collaboration with SWISS-PROT, to ensure correct mapping of the SEQRES records in a PDB entry to the sequence database entry at the residue level. Exchange of information between the E-MSD and SWISS-PROT further helps to maintain consistent information between structure and protein sequence databases.

Secondary structure. An in-house implementation of DSSP (16) and PROMOTIF (17) is used to derive secondary structure information based on the clean PDB data held in the search database. Information about strands, helices, sheets, beta bulges, turns, (beta turns and gamma turns), helix-to-helix interactions, and motifs (beta–alpha–beta units, beta hairpins and psi-loops) are calculated for a complete assembly to take

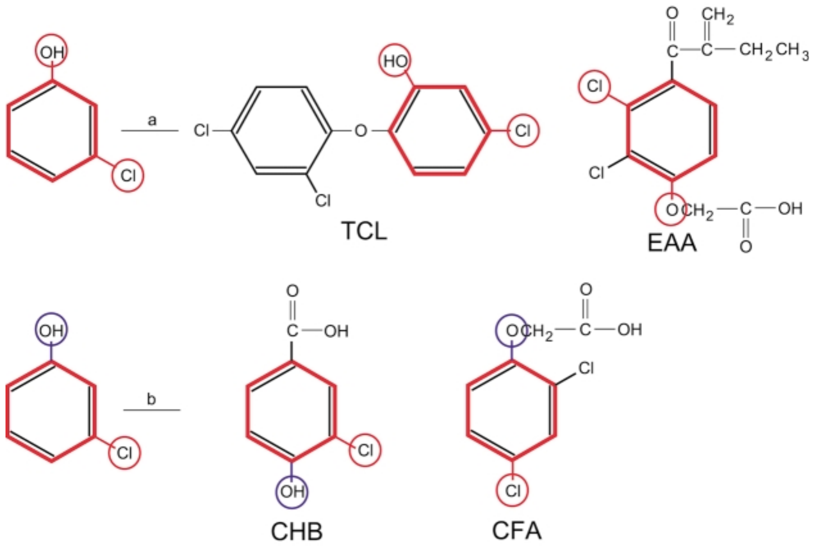

Figure 2. Sample SMILES based search using chempdb and starting from 3-chorophenol, (**a**) selected search results using the 'has substructure' option wherein the results have the connected fragment, and (**b**) selected search results using the fingerprint option where the matching ligands contain the chemical constituents of the query structure. The matched compounds shown are: TCL 5-chloro-2-(2,4-dichlorophenoxy)phenol, EAA [2,3-dichloro-4-(2-ethylacryloyl)phenoxy]acetic acid, CHB 3-chloro-4-hydroxybenzoic acid, and CFA 2,4-dichlorophenoxy acetic acid.

into account the cases where strands from symmetry related chains contribute to the sheet structure. This results in a uniform assignment of secondary structure information across the whole of the archive for the complete biological assemblies.

Validation and confidence levels. The validation processes include a range of issues including authentication of source, authentication of versions, validation of correct methodology, conformity to standards and error checks for consistency and outlier detection. The database records goodness-of-fit, and quality indicators for entries, chains and residues allowing the coordinate information to be studied in conjunction with reference data. For those crystallographic entries, where structure factors have been deposited, the map correlation indicators are included at the residue level. Accurate, consistent data along with clear indicators of quality are key aspects that make the E-MSD useful to a diverse community of researchers. Work is in progress to further associate electron density confidence criteria in a meaningful manner.

Ligand data. Of the ca. 18 000 entries in the PDB, many contain metal ions or other non-polymer groups. The E-MSD has put in place a system for matching all the chemical constituents of an entry against the PDB ligand dictionary derived from the clean data in the search database. Detailed descriptions of the ligands have been built in a consistent manner across the whole archive. These include:

(i) *General information*: Name, formula, GIF images, IUPAC names, overall charge, stereo SMILE (18), non stereo SMILE, classification, CAS Registry number, therapeutic category, Merck Index Identifier, topological classification, parent, ring and plane details.

(ii) *Atom specific information*: Atom names (PDB and IUPAC, in PDB order), energy types, the R and S stereochemistry and the Vega atom types (19) together with an E-MSD extended Vega atom type.

(iii) *Bond specific information*: Bond order, aromaticity and stereochemistry [*cis*/*trans* (Z/E)].

(iv) Coordinates from a selected PDB entry and idealized coordinate set (CORINA, 20).

The additional information, generated by these procedures, along with graph isomorphism approaches are used to match the ligands in a database entry against a graph derived from the reference information recorded within the search database. For new ligands, procedures are in place to automatically generate the dictionary descriptors. Many of the procedures are built around the CACTVS software suite (21). In addition, algorithms have been developed for matching fragments of molecules against fragments in the ligand dictionary (subgraph–subgraph isomorphism).

Ligand binding site environments. The ligand/protein interactions within assemblies are derived automatically during the transformation step from the deposition database to the search database. The binding site environments are stored within a simple relational structure that allows for fast search and retrieval. The database contains information on the ligand-polymer (protein and nucleic acid) contacts for the types; covalent, ionic, van der Waals, hydrogen bonding, salt bridges and ligand plane to polymer side chain plane interactions. In addition, the actual distances, angles and the angles between the normals for two planar groups are stored. A comprehensive application program interface (API) and web services are being developed for searching the binding site and

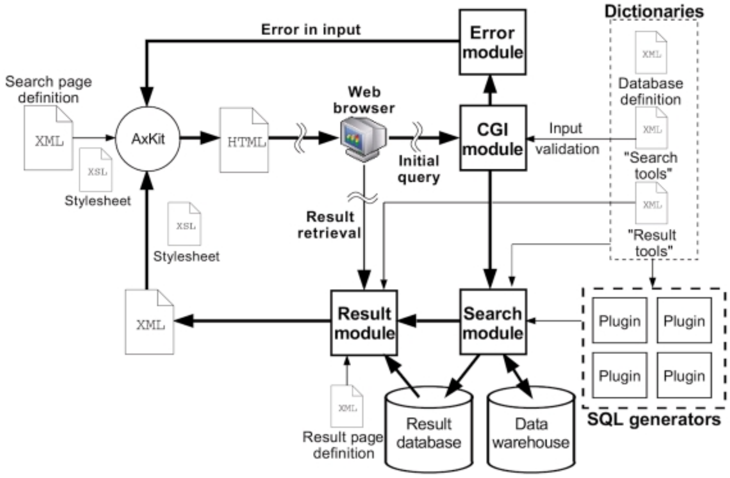

Figure 3. The process is driven by a number of dictionaries describing the database-model (Database Definition), interface contents and layout (Search page definition, Result page definition) or the description useful in construction of the SQL query (Search tools, Result tools). The system uses the XML-XSL technology to generate HTML pages using AxKit module.

active site information. Three-dimensional graph-matching algorithms similar to those used for the protein structure matching service (see below) are being applied.

SERVICES AND TOOLS

Protein structure matching (http://www.ebi.ac.uk/msd-srv/ssm). The protein structure-matching tool allows for alignment of a query structure against individual protein chains, domains or the PDB and SCOP archives. A modified Ullman's algorithm (22) developed in the group (23) allows for a fast and efficient method for returning substructure or exact structure alignments. The individual results are annotated with detailed secondary structure alignment and the CATH and SCOP domain classifications. Aligned structures may be viewed separately or superimposed using RASMOL. The individual results may also be downloaded as PDB formatted files.

The chempdb search system (http://www.ebi.ac.uk/ msd-srv/chempdb). The 'chempdb' search service provides web access to the ligands and small molecule dictionary of the E-MSD database. The search facility provides a simple to use functionality for queries based on chemical equivalence, similarity, substructure and superstructure. In addition, the database associates the atoms with energy types from reference libraries such as CNS (24) and CCP4/Refmac5 (25). The service provides a range of methods for exploring the database contents based on the additional information. For example the query molecule can either be drawn using the JME Molecular editor (26), or uploaded as a file in Mol2, SDF or PDB format. In addition to graph isomorphism methods, a fingerprint method has been developed. This is a fuzzy similarity search operation where the result lists molecules that have at least 99% of chemical fragments in common (see Fig. 2). The fragmentation patterns are derived from a predefined set of 500 chemical groups defined by CACTVS (21). The results page allows for a selection of export formats (PDB, SDF, mmCIf 27 or XYZ) and interactive viewers (JMOL, http://jmol.sourceforge.net/ or RASMOL, 28) for idealized coordinates (including calculated hydrogen atom positions) or coordinates from a pre-selected PDB entry.

E-MSD query system. The search interface is a generic system designed using a flexible meta-data driven mechanism and is implemented in the Apache mod_perl environment (http://perl.apache.org). This approach ensures that the software system is not tied to a specific database model. The system provides access to the search database by generating SQL queries based on the input information on the form. At present, the interface provides facilities to search on IDs [PDB, SWISS-PROT, Medline (29), GO, Interpro and EC-numbers], authors, keywords, experiment type, resolution, date of deposition or release, taxonomy (using NCBI tax-id's, (30), or organism names), ligand names or three letter codes. The searches can also include amino acid sequence in FASTA format or information about assemblies. The query can be restricted to search on the SCOP or DALI (31) representative sets. The interface also provides options to select the contents of the result pages. Figure 3 describes the typical life cycle of a request.

For secondary structure searches, a separate interface is available via (http://www.ebi.ac.uk/msd-srv/chempdb/cgi-bin/structure.pl). This interface allows for queries against any secondary structural element or property. For each entry, an overview of the secondary structure based at the assembly level is provided. Users may search for entries based on the existence and number of secondary elements e.g. find all entries with 8–10 helices with more than 1 sheet. Searches may also be based on individual secondary structure element attributes, for example, find all helices with linearity greater than 80 degrees with a unit rise between 1.2 and 1.3 Å.

FUTURE

Our aim is to deploy a fully integrated system that consists of protein data, ligands, sequence data, domain families, textual data MEDLINE (29), functional data BRENDA (32), with helper applications (FASTA, structure alignment). The user will see a single context sensitive interface to the E-MSD where complex relationships can be proposed in a natural way. Results will be returned based on the search context in an intuitive manner. Experience suggest that it is necessary to present no more than 6 pieces of information at one time to avoid data overload. A major problem is that each information item that is useful to a user depends on the background of the user. A protein sequence as a string of characters is a single datum to a biologist but represents many data to a chemist. A first aim is to provide viewers to present the structural, sequential, graphical and relational results to the user. These should all work together, be context sensitive and be fully browser complaint. A second aim is to provide a graphical drawing system to allow the expert user to set up graphical relationship queries.

In addition, a versatile API to the E-MSD database is under development. The API will consist of a series of functions that external third party software can use to allow their systems to access the E-MSD database independently. The API will be derived by defining both queries and event requests. In order to allow a more general and database independent interoperation, alternative integration strategies will also be explored.

ACKNOWLEDGEMENTS

E-MSD gratefully acknowledge the support from the Wellcome Trust (GR062025MA), the EU (TEMBLOR, NMRQUAL and IIMS), CCP4, the BBSRC, the MRC and EMBL.

REFERENCES

1. Hamm,G.H. and Cameron,G.N. (1986) The EMBL data library. *Nucleic Acids Res.*, **14**, 5–10.
2. Bairoch,A. and Boeckmann,B. (1994) The SWISS-PROT protein sequence databank: current status. *Nucleic Acids Res.*, **22**, 3578–3580.
3. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acid Res.*, **28**, 45–48.
4. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
5. Service,R.F. (2000) Structural genomics offers high-speed look at proteins. *Science*, **287**, 194–196.
6. Dubois,P. (1999) *MySQL*. New Riders Press, ISBN 0735709211.
7. Henrick,K. and Thornton,J.M. (1998) PQS: a Protein Quarternary File Server. *Trends Biochem. Sci.*, **23**, 358–361.
8. Lin,D., Manning,N.O., Jiang,J.S., Abola,E.E., Stampf,D., Prilusky,J. and Sussman,J.L. (2000) Autodep: a web based system for deposition and validation of macromolecular structural information. *Acta Crystallogr.*, **D56**, 828–841.
9. Ponstingl,H., Henrick,K. and Thornton,J.M. (2000) Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins*, **41**, 47–57.
10. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D.R., Durbin,R., Falquet,L., Fleischmann,W., Gouzy,J., Hermjakob,H., Hulo,N., Jonassen,I., Kahn,D., Kanapin,A., Karavidopoulou,Y., Lopez,R., Marx,B., Mulder,N.J., Oinn,T.M., Pagni,M., Servant,F., Sigrist,C.J.A. and Zdobnov,E.M. (2001) InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
11. The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
12. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
13. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH-a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
14. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam families database. *Nucleic Acids Res.*, **30**, 276–280.
15. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
16. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
17. Hutchinson,E.G. and Thornton,J.M. (1996) PROMOTIF—a program to identify and analyze structural motifs in proteins. *Protein Sci.*, **5**, 212–220.
18. Weininger,D., Weininger,A. and Weininger,J.L. (1989) SMILES2 algorithm for generation of unique smiles notation. *J. Chem. Inf. Comput. Sci*, **29**, 97–101. For a more recent exposition on the SMILES language, see Daylight Theory Manual, Chapter 3, Daylight Chemical Information Systems, Inc., 18500 Von Karman Ave. 450, Irvine, CA 92715, USA.
19. Pedretti,A., Villa,L. and Vistoli,G. (2002) VEGA: a versatile program to convert, handle and vusualize molecular structure on winows-based PCs. *J. Mol. Graph.*, **21**, 47–49.
20. Sadowski,J., Gasteiger,J. and Klebe,G. (1994) Comparison of automatic three-dimensional model builders using 639 X-Ray structures. *J. Chem. Inf. Comput. Sci.*, **34**, 1000–1008.
21. Ihlenfeldt,W.D., Takahashi,Y., Abe,H. and Sasaki,S. (1994) Computation and Management of Chemical Properties in CACTVS: An extensible Networked Approach toward Modularity and Flexibility. *J. Chem. Inf. Comput. Sci.*, **34**, 109–116.
22. Ullman,J.R. (1976) An algorithm for subgraph isomorphism. *J. Assoc. Comput. Mach.*, **23**, 31–42.
23. Krissinel,E., Velankar,S. and Henrick,K. (in press). Protein Structure comparison based on a new algorithm for common subgraph isomorphism.
24. Brunger,A.T., Adams,P.D., Clore,G.M., Delano,W.L., Gros,P., Grosse-Kunstleve,R.W., Jiang,J.-S., Kuszewski,J., Nilges,M., Pannu,N.S., Read,R.J., Rice,L.M., Simonson,T. and Warren,G.L. (1998) Crystallography and NMR System. *Acta Crystallogr.*, **D54**, 905–921.
25. Collaborative Computational Project, Number 4 (1994) The CCP4 Suite Programs for Protein Crystallography. *Acta Cryst.*, **D50**, 760–763. See also Murshudov,G.N., Vagin,A.A. and Dodson,E.J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr.*, **D53**, 240–255.
26. Ertl,P. and Jacob,O. (1997) WWW-based Chemical Information System. *Theochem*, **419**, 113–120.
27. Bourne,P., Berman,H.M., Watenpaugh,K., Westbrook,J.D. and Fitzgerald,P.M.D. (1997) The macromolecular Crystallographic Information File (mmCIF). *Methods Enzymol.*, **277**, 571–590.
28. Bernstein,H.J. (2000) Recent changes to RasMol, recombining the variants. *Trends Biochem. Sci.*, **25**, 453–455; Sayle,R. and Milner-White,E.J. (1995) RasMol: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.
29. National Library of Medicine (1989) MEDLINE [database online]. Bethesda (MD). Updated weekly. Available from: National Library of Medicine; OVID, Murray, UT; The Dialog Corporation, Palo Alto, CA.
30. Wheeler,D.L., Chappey,C., Lash,A.E., Leipe,D.D., Madden,T.L., Schuler,G.D., Tatusova,T.A. and Rapp,B.A. (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28**, 10–14.
31. Holm,L. and Sander,C. (1999) Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.*, **27**, 244–247.
32. Schomburg,I., Chang,A. and Schomburg,D. (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.*, **30**, 147–149.