E-MSD: an integrated data resource for bioinformatics

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

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

ABSTRACT

The Macromolecular Structure Database (MSD) group (http://www.ebi.ac.uk/msd/) continues to enhance the quality and consistency of macro-molecular structure data in the Protein Data Bank (PDB) and to work towards the integration of various bioinformatics data resources. We have implemented a simple form-based interface that allows users to query the MSD directly. The MSD 'atlas pages' show all of the information in the MSD for a particular PDB entry. The group has designed new search interfaces aimed at specific areas of interest, such as the environment of ligands and the secondary structures of proteins. We have also implemented a novel search interface that begins to integrate separate MSD search services in a single graphical tool. We have worked closely with collaborators to build a new visualization tool that can present both structure and sequence data in a unified interface, and this data viewer is now used throughout the MSD services for the visualization and presentation of search results. Examples show-casing the functionality and power of these tools are available from tutorial webpages (http://www. ebi.ac.uk/msd-srv/docs/roadshow_tutorial/).

INTRODUCTION

The Macromolecular Structure Database (MSD) is a relational database that aims to be a single access point for all types of data about protein and nucleic acid structures. The MSD group works closely with the Research Collaboratory for Structural Bioinformatics (RCSB) in the USA and PDBj in Japan, to maintain the Protein Data Bank (PDB) (1), the single worldwide repository of protein and nucleic acid structures. The three centres are actively involved with the clean-up and maintenance of the legacy archive and run deposition systems through which new macromolecular structures can be added to the PDB archive. There is a weekly exchange of data between the partner sites, which ensures the consistency and uniformity of the single archive. The approach of the MSD group to the problems of storing, maintaining and cleaning the large volume of PDB data has been to load it into a relational database, as described previously (2).

In addition to our close relationship with the PDB partner sites, the MSD group also maintains active collaborations with several other major bioinformatics projects. These partnerships have allowed us to enhance the usefulness and scope of the MSD by adding curated and actively maintained derived data to the existing structural data from the PDB. Our partners benefit from access to a consistent and readily accessible source of macromolecular structure data. Our many collaborations have helped us to add a large volume of additional data, such as cross-references to the SCOP (3) and CATH (4) structure classification databases, the sequence-centric databases Swiss-Prot (5), InterPro (6), Pfam (7) and ProSite (8), as well as the gene ontology (GO) database (9) and the PubMed literature archive (10). The MSD is now well on the way to achieving the goal of unifying a diverse range of biological data resources around macromolecular structure information from the PDB. Our emphasis is now on the creation of search systems that will make the wealth of data in the MSD accessible to the wider community.

MSD SEARCH TOOLS

We have developed search systems for the MSD database that allow users to access independent areas of the database, as well as interfaces that cover the database as a whole. Some of these systems have been described previously (2), and here we describe further enhancements and new developments to MSD services and search tools.

MSDlite

MSDlite (http://www.ebi.ac.uk/msd-srv/msdlite/) is an easy to use search tool that allows users to perform a wide range of different queries through a single web form (Fig. 1a). The
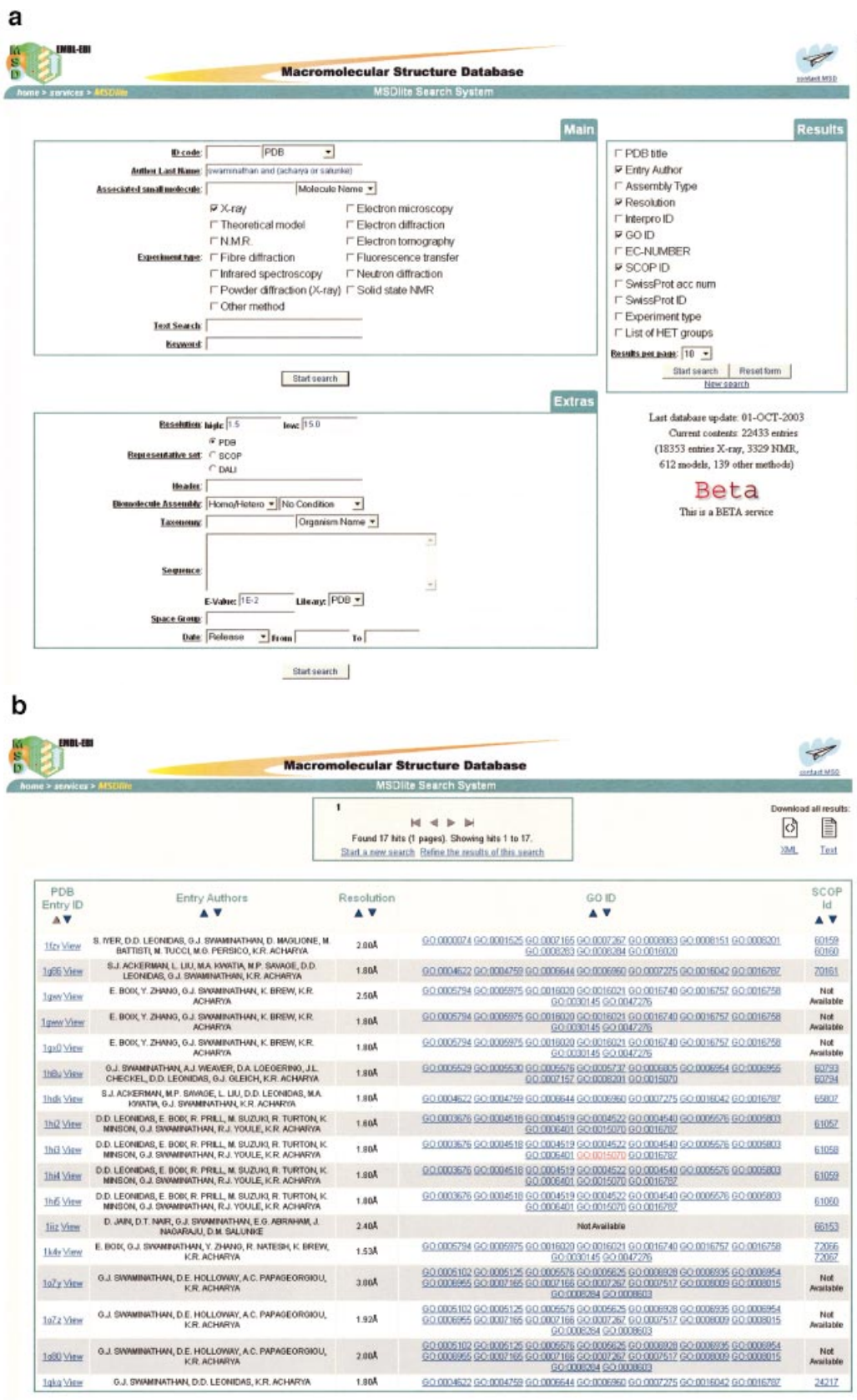
Figure 1. (**a**) The MSDlite query form. The form is divided into three panels. The two left-hand panels show search fields, with the most commonly used terms found in the uppermost panel. The right-hand panel allows the user to configure the results that will be returned by their query. (**b**) A sample MSDlite results page. Results are presented in a simple table, divided across several pages if necessary. Result items are hyperlinked to additional resources wherever possible. For every search hit, the PDB accession code is linked to the MSD atlas page for that entry, and the 'View' link allows the user to visualize the structure using AstexViewer@MSD-EBI.

Figure 2. The MSD atlas page for PDB entry 1hdk. The atlas pages present all available information for a particular PDB entry. Related information is grouped into separate tab panes, covering sequence, ligand data, journal citations, etc. Information in the atlas pages is hyperlinked wherever possible, directing the user to related MSD and external resources.

available search fields include commonly used terms such as author name, title, keywords, etc., but also include terms that allow the user to search the MSD through cross-references to other data resources: the interface includes search terms such as NCBI taxonomy, GO ID, InterPro ID, Swiss-Prot ID and accession number, and EC numbers (11). Furthermore, the interface allows users to search the MSD based on sequence similarity, using the FASTA algorithm.

MSDlite has been designed as a flexible and easy-to-use system that allows users with different fields of interest to extract information that is of direct relevance to their work. To facilitate this, MSDlite allows the user to select the result fields that will be returned by their query.

The previous MSD query system was implemented using mod_perl and Apache, but the current system is a series of Java servlets running in the Tomcat servlet container. The core of the system is a generic, meta-data-driven query engine that can form the basis for multiple query interfaces. MSDlite is one implementation of a query interface. When the user submits their search, a description of the user's search is sent to the query engine, which automatically constructs a valid SQL query on-the-fly and executes it on the MSD. The results are returned to the interface and presented as a table of hits (Fig. 1b). The system adds hyperlinks wherever possible, directing the user to additional resources on the web. The primary link for each hit leads the user to a set of 'atlas pages' (Fig. 2). These atlas pages show summary information, sequence data, references, including full PubMed abstracts where available, and further links to other data resources. The result set can also be downloaded for later use, as a simple XML file or a tab-delimited text file.

MSDsite

One of the most important uses of macromolecular structure data is the study of the interactions between macromolecules and bound ligands. Due to the local nature of the interactions between ligands and macromolecules, ligand binding sites are often more highly conserved across a functional family than the overall structure and fold of the macromolecule. As such, a catalogue of ligand–macromolecule interactions can be invaluable in the wider study of structure–function relationships. Furthermore, as structural genomics initiatives begin to produce large numbers of structures, an accurate classification of ligand interactions in molecules of known function will be increasingly important as an identification tool for molecules of unknown function.

The MSD database addresses the issues of classification of small molecules (ligands) and ligand binding sites in two ways. As new small molecules are found in newly deposited macromolecular structures, they are curated and added to our database of small molecule structures (accessible through the MSDchem search system at http://www.ebi.ac.uk/msd-srv/chempdb/). We have further catalogued the interactions between every ligand in the PDB and the structure with which each is associated, classifying each individual bond or non-bonded interaction between macromolecule and substrate. Interactions are categorized according to their type (covalent, ionic, van der Waals, etc.) and, for each one, details such as bond distances and angles are recorded.

As well as allowing a user to obtain highly detailed information about a specific macromolecule–ligand interaction, one of the major benefits of using a relational database to store this information is that it is possible to extract statistics describing similar interactions throughout the entire PDB. The MSDsite service (http://www.ebi.ac.uk/msd-srv/msdsite/) is a form-based interface to these data and statistics (Fig. 3).

MSDpro

MSDlite and MSDsite allow users to ask complex questions over a range of subjects. In order to answer a specific biological question, the researcher is compelled to combine and consolidate the results from several distinct search systems. MSDpro (http://www.ebi.ac.uk/msd-srv/msdpro/) provides a system that will assist a user in answering their
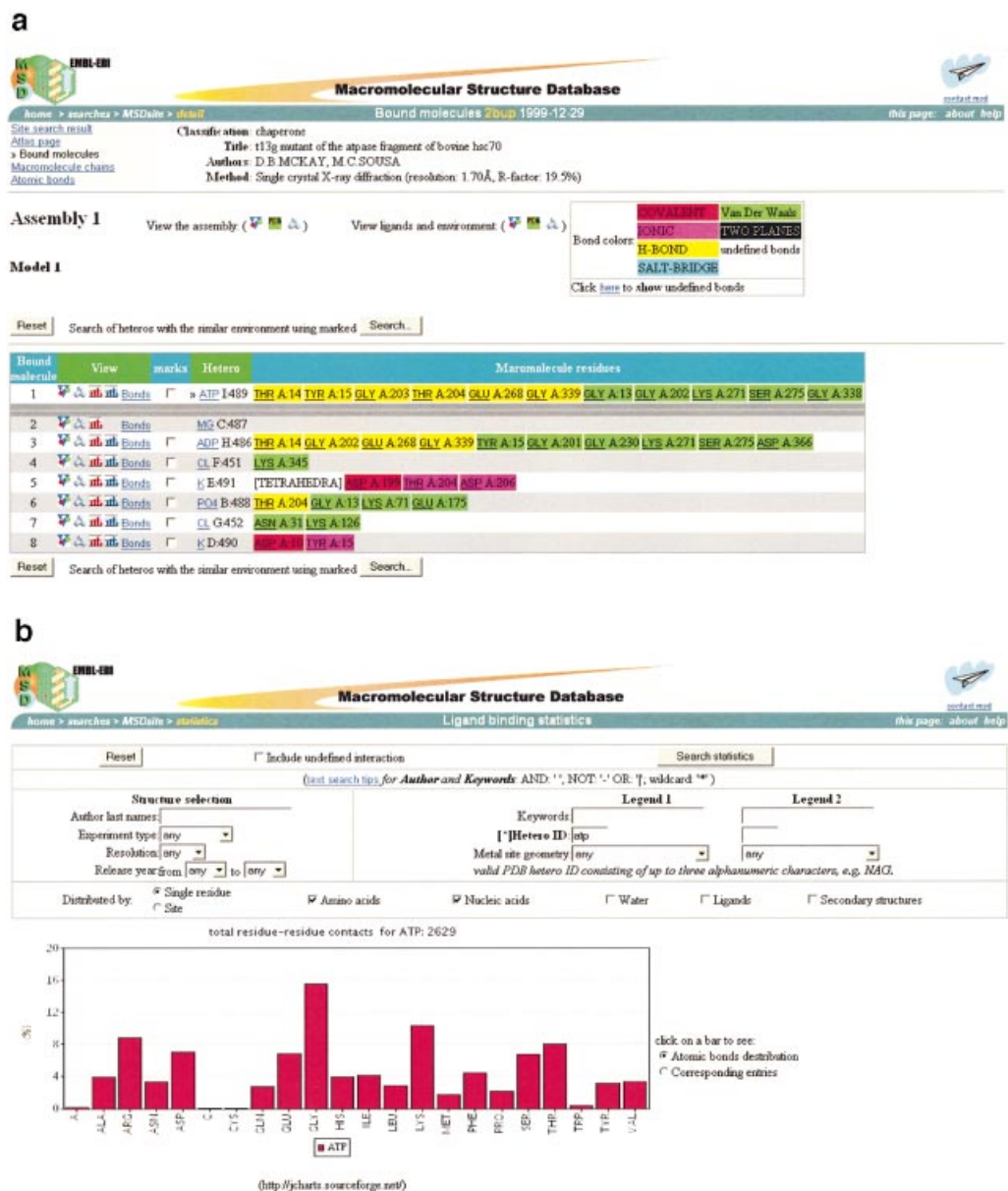
Figure 3. (**a**) The results of an MSDsite query. The results pages show the residues that interact with a given ligand, including the classification of each interaction and geometrical information in the case of metal coordination interactions. (**b**) The MSDsite statistics for a ligand binding environment. The environment is described in a simple chart, showing the most commonly found residues in the environment for a given ligand. Residues may be classified according to the type of secondary structure element in which they are found. Statistics can also be obtained for a complete ligand binding environment.

biological question without having to resort to a multitude of different resources, interfaces or bespoke applications.

The MSDpro interface is an interactive Java applet. The applet presents the user with a palette of query components, such as the name of the author of a PDB entry, a keyword, a Swiss-Prot accession number, etc. Crucially, these modules need not map directly to tables or columns in the database, but instead encapsulate concepts that will be familiar to the expert scientist. This abstraction is key to making the search system amenable to users who understand the data that are stored in the MSD, but whose knowledge of database systems may be limited. Furthermore, the modules may represent queries that are performed on a system that is entirely outside of the MSD. Examples of this are the MSDfold module, which uses the

secondary structure matching service (http://www.ebi.ac.uk/ msd-srv/ssm/) to find structurally similar proteins, and the FASTA module, which finds protein structures that have a similar sequence. The two services that underlie these 'external' modules are implemented as web services. The query server architecture is such that any web service can be similarly added as an MSDpro query component.

The user constructs a description of the query by first selecting the search components that they wish to use in the search, and, as in MSDlite, the specific fields that they wish to retrieve with their query. The most powerful feature of the interface is the way in which it allows users to combine components using the logical operations 'AND', 'OR' and 'NOT'. By grouping various components of a query
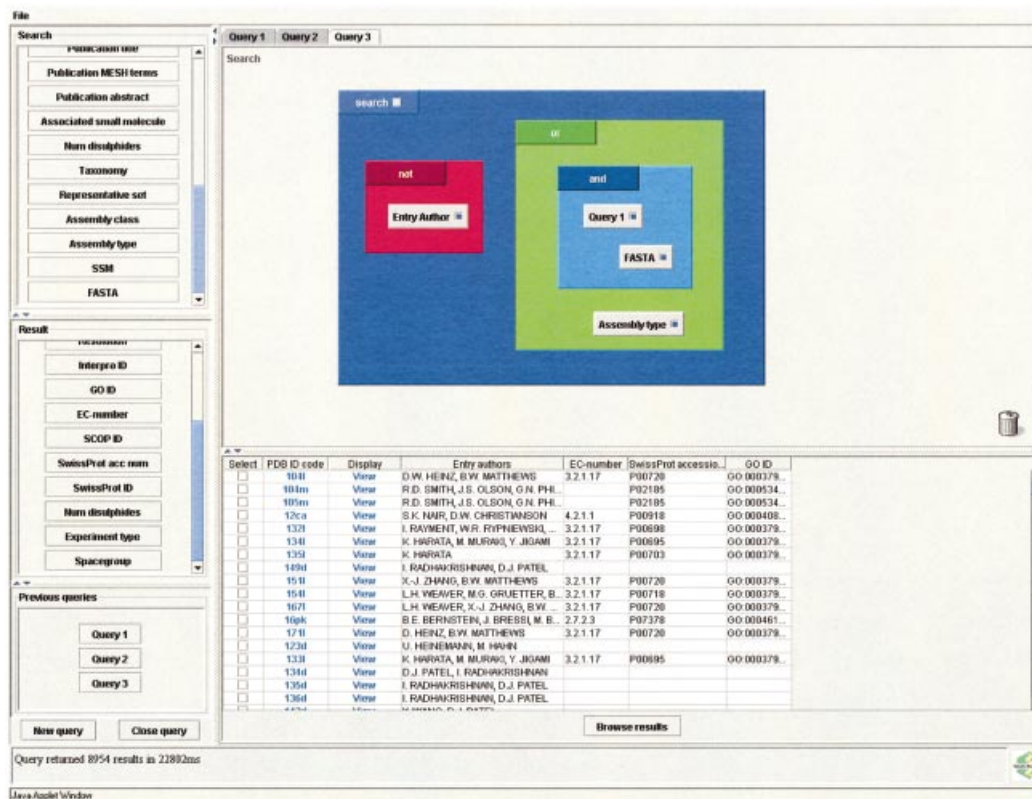
Figure 4. The MSDpro query interface. MSDpro allows the user to describe complex queries by dragging-and-dropping search and result terms from the lists on the left of the interface into the corresponding panels on the right. Search terms can be combined and using logical operations AND, OR and NOT. Results of a query are presented as a simple table within the interface itself, or they can be viewed through the MSDlite result pages.

appropriately, the user can graphically construct a description of a complex, highly specific question (see Fig. 4).

The results of the query are currently displayed as a simple table in the MSDpro applet, or as a more detailed table in the MSDlite system. Future development will add a suite of information visualization tools, allowing users to describe and execute queries and then analyse the results within an easy-to-use graphical environment.

AstexViewer@MSD-EBI

An essential component of a database query system is the provision for users to view and interpret the results of a search. We are now looking at new and novel ways of presenting the complex one-, two- and three-dimensional data that are available in the MSD.

In collaboration with Astex Technology (http://www.astex-technology.com/) we have developed AstexViewer@MSD-EBI, a macromolecular structure- and sequence-viewing program. This collaboration augments the functionality of the core structure viewer that has been developed by Astex (12). AstexViewer@MSD-EBI is capable of interactively displaying multiple macromolecular structures and sequence alignments (Fig. 5). The viewer further provides tools for querying and interrogating the views of these data using familiar tools and techniques. Regions of similarity or difference, in either the structure or sequence view, can be used to colour and/or highlight regions of the other view. Better integration of data from the structure and sequence

realms is achieved by allowing various properties of either view to be mapped onto the other.

To further enhance the integration of sequence and structural data, we are also experimenting with new and novel visualization tools, such as 'brushing' (13) and 'magic lens' (14). The brushing tool allows a user to interactively highlight a region of the data in one view, and see corresponding data points in every other view. By moving the brushing tool across the sequence view, the user simultaneously highlights a section of the polypeptide chain in the structure view. The magic lens is a semi-transparent area that can be moved across a structure view, allowing the user to overlay various types of information onto a structure without obscuring the view of the structure itself. For example, the lens could be used to display the sequence of a polypeptide chain as it is moved slowly across the structure, or the location of heterogens and their interaction with the macromolecule.

FUTURE

The MSD group continues to work closely with its partners, to further enhance the quality and consistency of the data in the database. Considerable efforts are being put into improving existing services and tools, and we are working towards a streamlined, unified interface to these services.

Advances have been made in the development of software to replicate the MSD search database at partner test sites. In the future we will support the download and installation of the
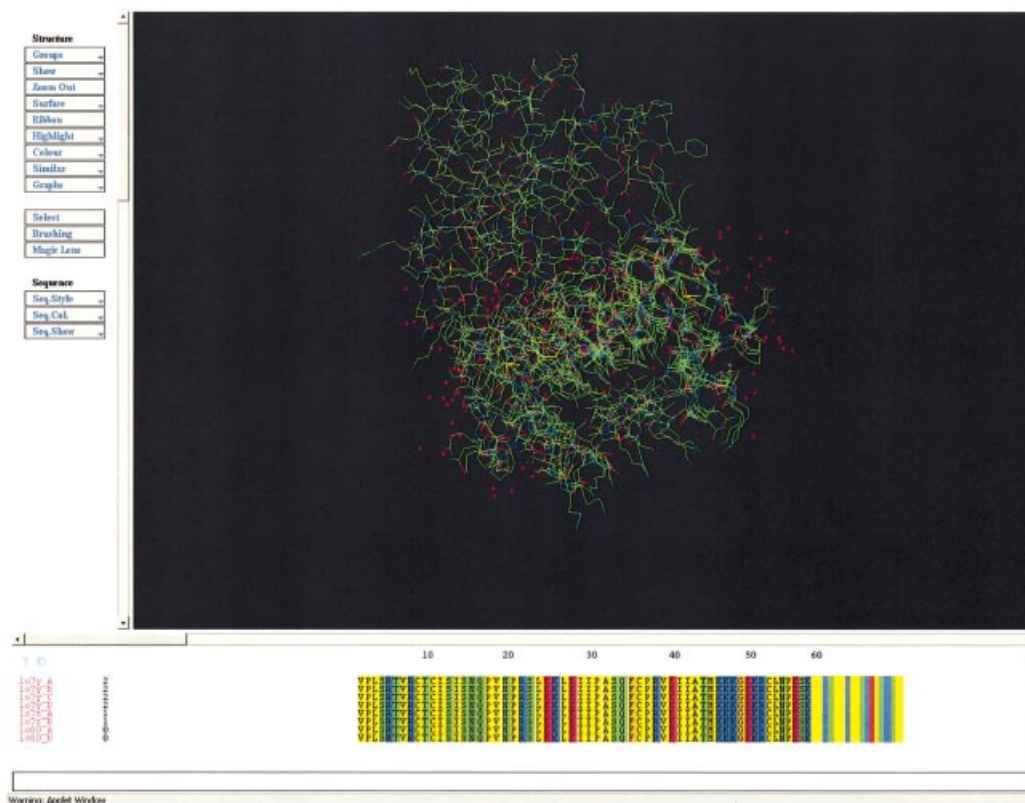
Figure 5. AstexViewer@MSD-EBI allows the user to view multiple, superposed structures and aligned sequences. The viewer provides a lightweight, easy-to-use interface for visualizing the results of various MSD searches.

database on third-party systems. As part of the distribution package a comprehensive application programming interface (API) is being developed, which will allow users to extract data from the database programmatically.

For the past year the MSD group has been presenting 'roadshows' at various labs across Europe (http://www.ebi.ac.uk/msd/roadshow.htm). These meetings provide us with an invaluable opportunity to present the current state of the MSD services, and, more importantly, to hear feedback and comments on the systems that we provide. These presentations and workshops will be continuing, as part of our ongoing commitment to meeting the needs and requirements of our users.

ACKNOWLEDGEMENTS

REFERENCES

1. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. D*, **58**, 899–907.
2. Boutselakis,H., Dimitropoulos,D., Fillon,J., Golovin,A., Henrick,K., Hussain,A., Ionides,J., John,M., Keller,P.A., Krissinel,E. *et al.* (2003) E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Res.*, **31**, 458–462.
3. Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
4. Orengo,C.A., Pearl,F.M. and Thornton,J.M. (2003) The CATH domain structure database. *Methods Biochem. Anal.*, **44**, 249–271.
5. Bairoch,A. and Boeckmann,B. (1994) The SWISS-PROT protein sequence data bank: current status. *Nucleic Acids Res.*, **22**, 3578–3580.
6. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
7. 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 protein families database. *Nucleic Acids Res.*, **30**, 276–280.
8. Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
9. Camon,E., Magrane,M., Barrell,D., Binns,D., Fleischmann,W., Kersey,P., Mulder,N., Oinn,T., Maslen,J., Cox,A. *et al.* (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL and InterPro. *Genome Res.*, **13**, 662–672.
10. Ebbert,J.O., Dupras,D.M. and Erwin,P.J. (2003) Searching the medical literature using PubMed: a tutorial. *Mayo Clin. Proc.*, **78**, 87–91.
11. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.
12. Hartshorn,M.J. (2002) AstexViewer: a visualisation aid for structure-based drug design. *J. Comput. Aided Mol. Des.*, **16**, 871–881.
13. Fua,Y.H., Ward,M. and Rudensteiner,E. (1999) Navigating hierarchies with structure-based brushes. *IEEE Proceedings Information Visualization '99*, pp. 58–64.
14. Stone,M., Fishkin,K. and Bier,E. (1994) The movable filter as a user interface tool. *ACM Proceedings CHI '94*, pp. 306–312.