MSDsite: A Database Search and Retrieval System for the Analysis and Viewing of Bound Ligands and Active Sites

Adel Golovin, Dimitris Dimitropoulos, Tom Oldfield, Abdelkrim Rachedi, and Kim Henrick
EMBL Outstation, The European Bioinformatics Institute, Welcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

ABSTRACT **The three-dimensional environments of ligand binding sites have been derived from the parsing and loading of the PDB entries into a relational database. For each bound molecule the biological assembly of the quaternary structure has been used to determine all contact residues and a fast interactive search and retrieval system has been developed. Prosite pattern and short sequence search options are available together with a novel graphical query generator for inter-residue contacts. The database and its query interface are accessible from the Internet through a web server located at: http://www.ebi.ac.uk/msd-srv/msdsite. Proteins 2005;58:190–199.** © 2004 Wiley-Liss, Inc.

Key words: binding-site; conformational superposition; function; pattern; short sequence; statistics; structure

INTRODUCTION

One of the most important uses of macromolecular structure data is the study of interactions between macromolecules and bound ligands. The three-dimensional (3D) environments of ligand-binding for the structures held in the Protein Data Bank (PDB)[1,2] are often more highly conserved across a functional family than the overall structure and fold of the macromolecule. In the wider study of structure–function relationships, few of the public systems for searching the PDB provide users with the ability to make queries related to active site environments and to visualize results or carry out analyses of the chemical environment.

The related public database and search system for ligand data in the PDB is the powerful Relibase[3,4] system. Relibase is both an object-oriented comprehensive receptor–ligand database and a program for searching protein–ligand databases. Relevant information is identified and extracted from PDB entries, and detailed associated chemical analysis of receptor–ligand interactions is based on bond orders and hybridization information derived for each PDB ligand. Relibase allows complex queries (including substructure searches, similarity searches and searches for specific interactions) regarding both small molecule and protein aspects. The IMB Jena Image Library[5,6] offers tools for their Hetero Components Database and their Site Database. These systems allow searches for the occurrence of any components in the environment of hetero components. Site data is merged with environmental components having at least one atom located within a distance of 4.2 Å from any of the site atoms. A problem with these data is that the database is populated from PDB entry records that contain sparse and unreliable SITE record information. Other public services include the PDBSite[7] database on protein active sites and their spatial environments, including structural features calculated by spatial protein structures and the physicochemical properties of sites. The system is accessed via the Sequence Retrieval System (SRS)[8] using textual and accessibility fields, and the authors offer an automatic best superposition of sites with the 3D structure of the protein under study via their software PDBSiteScan. The PROMISE database[9] of prosthetic groups and the Metalloprotein database[10] offer limited search facilities, while the PROCAT facility[11,12] provides views of 3D enzyme active site templates and allows search by PDB code. ProBiSite,[13] a protein binding site database that contains PDB data integrated with the CATH database,[14] along with small ligands and protein binding sites in the PDB, also has only limited search abilities by text and identifiers.

We have developed tools to extract ligand binding information and to characterize their chemical and geometric environments using the Macromolecular Structure Database (MSD).[15] Within the MSD database, we have addressed issues of classification of small molecules (ligands) and ligand binding sites in two ways. For all the legacy and new small molecules in the PDB, we catalogued topology, bond orders and hybridization information to give a self-consistent set and added it to our database of small molecule structures (accessible through the MSDchem search system at http://www.ebi.ac.uk/msd-srv/chempdb/). MSDchem ensures that the data are continually kept up-to-date with all PDB releases. For each instance of a ligand present in the PDB, the MDSsite data mart tables (a subset of the MSD), was populated with derived interaction and ligand environment details. Individual bonded and non-bonded interactions between macromolecules and substrates were classified and stored. The interactions were categorized according to their type (covalent, ionic, van der Waals, etc.), and, for each one, details such as bond

Correspondence to: Kim Henrick, EMBL Outstation, the European Bioinformatics Institute, Welcome Trust Genome Campus, Hinxton, Cambridge UK. Email: henrick@ebi.ac.uk.

Received 20 February 2004; Revised 30 June 2004; Accepted 9 July 2004

Published online 1 October 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20288

distances and angles were 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 is freely accessible from the Internet through a web server located at: http://www.ebi.ac.uk/msd-srv/msdsite/ and is a form-based interface to these data and statistics.

MATERIALS AND METHODS

MSDsite Data

The database contains the following information: (i) all PDB ligands and their interactions with macromolecules (proteins/RNA/DNA), with other ligands and with solvents (water); (ii) geometries of metal coordination and (iii) protein sequences and ProSite patterns[16] with their ligand interactions. A ligand was considered to be any bound molecule or modified residue, where a modified residue is a residue of a protein chain that is not one of the twenty standard amino acids. The spatial surroundings of each bound molecule were calculated using algorithms we have made available as extensions of the CCP4 library.[17] Information about specific hetero groups was taken from MSDchem and stored in the reference database tables as part of the MSD database. Interactions between coordinates were, where appropriate, searched over all crystal symmetry operations. If an interaction was found, then coordinates of the symmetry transformation were stored. All water interactions were treated similarly.

Ligand interactions were classified using the following criteria.

Covalent and Ionic Interactions

Inter atomic interactions for both covalent bonds and ionic interactions were derived, including those for disulfides. For metal atoms, covalent and ionic interactions were assigned using these guidelines: (1) if the distance between two atoms is within covalent bond length, then the bond has covalent coordination and (2) If the distance between two atoms is within (ionic + covalent) / 2, then the bond has covalent coordination. Otherwise, if the distance is within ionic bond length, the bond is an ionic coordination bond. For some metals, we also applied an empirical set of rules to assign ionic or covalent interactions.

Hydrogen Bonding

We applied the HBExplore 2.0[18] hydrogen bond criteria calculated after taking into account a local implementation of the Vega[19] topology descriptors, hybridization vectors and the binding geometry of potential donor–acceptor pairs. The original software used a restricted set of templates to provide the necessary information for hybridization vector calculations. The PDB lists some 5,000 ligands, and we extended the templates by deriving Vega atom-type descriptors to describe and recognize the atom types from chemical connectivity from idealized coordinates generated with CORINA.[20,21]

Van der Waals Bonding

When two atoms are within van der Waals distances, they were listed as van der Waals interactions.

Non-Bonding Type

Interactions between atoms within 4 Å of each other for which no bond type had been classified were assigned as non-bonding type interactions.

Rings

We defined rings using the merged full set of smallest rings from the spanning tree definition,[22] with some local rules to eliminate cyclic peptides and cases in which covalent bonds in a macromolecular structure would lead to chemically unreasonable results. Only those rings that have fewer than eight atoms and only contain the elements C, N and O were kept in the database, while rings containing B, P and S were eliminated. We also eliminated bridgehead atoms. If the edited ring had more than three atoms, then its geometric center and planar equations were stored.

Planar Groups

Chemical fragments, such as guanidinium groups with four or more atoms, were recognized by applying an extended Vega algorithm[19] and using a lookup table of topology typing fragments. A set of atoms defining a plane was accepted where the deviations from the equation of the least squares plane were less than 0.1 Å.

Ring–ring, plane–plane and ring–plane interactions were accepted when the two centers were within six Å; the distance and normal vectors were stored.

Coordination Geometry of Metals

Metal site geometries were stored in the MSD and classified as tetrahedral, trigonal-bi-pyramidal, square-based pyramidal, octahedral and undefined, based on all valid symmetry-related donor atom positions. The metal site geometry angle and distance data were taken from published material.[23–26]

Sequences and ProSite Patterns

ProSite patterns[16] were stored, and a matching program was run across all PDB entries. For any patterns, either those defined by ProSite or those that comply with the ProSite pattern format can be aligned by a fast search algorithm developed and embedded into the Oracle database.

The limited information available in the PDB from SITE records was not used. However, we loaded sequence data for catalytic residues in enzyme active sites from two sources: (i) CATRES[27,28] and (ii) MEROPS.[29,30] In addition, a set of similar spatial arrangements of amino acids in protein structures that are close in space but not necessarily adjacent was derived using a data mining method[31] and loaded into the database. These structural motifs have been shown not only to match many known configurations of residues such as the catalytic triad and

metal binding sites but also to contain many other multiple-residue interactions not previously categorized.

MSDsite Database and Server Structure

This service is designed to return information regarding the interaction of ligands with macromolecules. Therefore Oracle tables were created to store the details of ligand atoms, of the atoms/residues that interact with the ligands, and of the interaction types (Appendix A). A key aspect of the service is the responsiveness of the interface, and this is defined by good design of the underlying query system. Structural Query Language (SQL) is used as the basic tool for searching the MSD and the Oracle Relational Database Management System (RDBMS). It has features that dramatically improve query performance, such as Bitmap indexes, packages and functions, as well as optimizer hints.[32] Relational algebra set operations (union, exclude, intersect) are used with a fast parallel implementation in the RDBMS. These optimization techniques result in fast interactive queries ($l<3$ s) for statistical analysis of the current 23,000 PDB entries in the MSD and any size increase due to several years of structure submissions.

Short Sequence and ProSite Pattern Search

Sequence searching on a structural database is a major challenge. There are many tools that perform fast sequence matching; however, few support ProSite patterns because their integration into a database demands extensive resources and most are optimized for long sequences and not effective for short ones. For short sequence alignment and pattern matching, we tried two approaches. The first was to write an Oracle package implemented in C that performs a fast sequence scan and stores the results in temporary tables, which are then used in a query that reflects all other desired features, such as ligand and/or environment characteristics and structure filters. The second approach was to use only SQL by building an effective query to a special table. This approach has the benefit of ease of integration with other queries. We have found that both methods give approximately the same response time, but the SQL method was preferred.

Details of the SQL design to enable rapid short sequence queries are presented in Appendix B.

Search Engine and Web Application Server

The MSDsite architecture is called n-tier; it uses Oracle server 9i for storing, querying and managing the data. The web application server is Tomcat 4.x. The web application was developed using the J2EE application development platform by Java Sun and is based on XML web service technology with the advantages of openness and independence of data source. XML provides the layer between data storage and interface, thus providing a transparent layer for storage and any application that uses this data. For example, it allows the same mechanism to be used regardless of whether the data origin is the database or an uploaded file. The same approach is used for output in different formats to support 3D visualization tools. Applying style sheets to the same XML generates HTML pages, tabulated files and PDB files. An additional benefit is the ability to query the resulting XML using standard tools. This is a very powerful mechanism for highlighting searched sites and other search targets on results pages and 3D visualization tools.

The web interface is a search form that can initiate a search and generate results pages that allow additional searches to be initiated, based on the results of a previous search. For example, each detailed results page about a particular PDB entry is a search form. Statistics pages can also be used as search objects through interactive charts. Search and statistics forms consist of two parts, ligand search fields and a structure filter. The filter is limited to search by PDB header information: TITLE, KEYWORDS, HEADER CLASS, RESOLUTION, AUTHOR and DATE RELEASED.

3D Multi-Visualization

Three-dimensional visualization tools are supported with Rasmol script output[33] and output for the AstexViewer™@MSD-EBI[34] applet. In addition, a provision for site superposition is given with options for alignment: by ligand, by ProSite pattern, by active site residues of a pattern or by environment. The first three options use the algorithm for aligning structures by three axes of minimal moment of inertia. In the case of alignment by ligand, we align atoms with the same names; however, this is limited to similar ligands and will fail when the ligands are quite different. In the second option, alignment is only for the same or similar patterns and is carried out using the main-chain atoms C, C_α and N. Alignment of active sites uses all atoms (except hydrogen) and can be sensitive to pattern differences. Alignment by environment is a Non-Polynomial (NP) complete problem with no known polynomial time algorithm to solve it. We adopted a global optimization method that allows us to stop the optimization process either when the solution meets a given precision or when a predefined number of steps have been applied.

Uploading PDB Files

An option to upload a PDB file from the client computer is available. The uploaded file is parsed for ligands and active sites, and these are used as search objects. The server applies uploaded symmetry operations to the ligand and water molecules.

Figure 1 shows a conceptual view of the web pages that make up the MSDsite service. The main search page at the top of the figure represents the entry point for all different information that can be queried from the service. The three main pages – statistics, details and links – are all subdivided depending on the user-requested information that is to be returned. Since each web page acts as a search page as well as a results page the linkage between the service components is bidirectional in many cases.

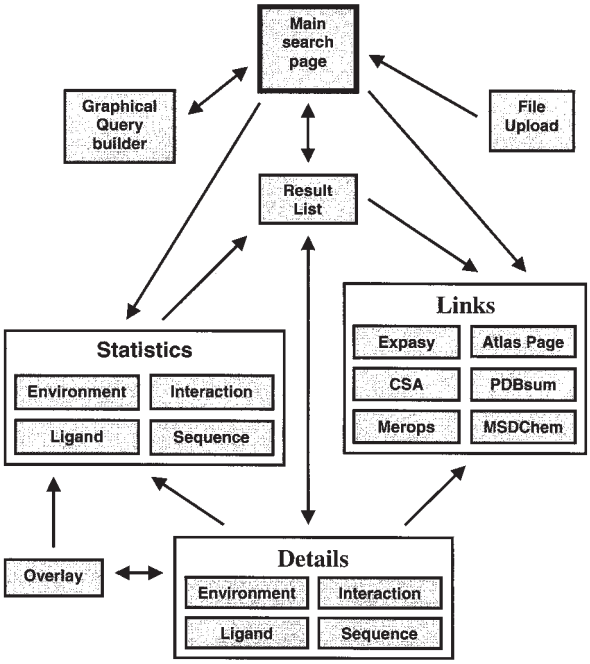

Fig. 1. Conceptual view of the MSDsite service. The main search page at the top of the figure represents the entry point and query page for the service. There are three main results pages: statistics, details and links and these are further subdivided to improve clarity of the results. A number of the links between web pages are bi-directional because the results pages are also query pages to allow comprehensive mining of the data. The LINKS module shown above references the following resources: EXPASY—http://www.expasy.org; PDBsum—http://www.biochem.ucl.ac.uk/bsm/pdbsum/; CSA—http://www.ebi.ac.uk/thornton-srv/databases/CSA/; Merops—http://merops.sanger.ac.uk/; MSDchem—http://www.ebi.ac.uk/msd-srv/chempdb/.

RESULTS AND DISCUSSION

Query Examples

Ligand and Environment Search

This search is intended to perform the following functions: (1) list PDB entries that contain a given ligand, (2) list PDB entries that contain a given ligand interacting with a specified site, (3) list PDB entries that contain small molecules with the given ligand environment and (4) list PDB entries that contain a metal with a given site geometry.

Any combination of the above functions is allowed and is defined from the main search page. A ligand interaction with its environment can be described at the residue and atomic levels. Search options also allow specification of the secondary structure for the residue, helix, strand or turn, and atoms that have the search attribute of main chain or side chain. Residues can be described as "ATP|ADP|GTP|GDP," which means that the small molecule may be ATP, ADP, GTP or GDP combined by logical OR. For example, a particular atom of a ligand can be specified as "HEM|HEC.FE," which means that the small molecule is HEM or HEC and contains an interacting iron atom FE. Finally, interactions of just the adenosine group of ATP or ADP may be specified using logical OR on the atoms as "ADP|ATP.N6|N1|N3|N7|N9|C2|C8." Ligand descriptions are separate from the ligand environment description, where the environment is a set of residues. In addition to residue and atom specifications, the ligand environment description can have secondary structure, while interactions can be specified in terms of distances between atoms and/or bond type. In the case of plane/ring interactions, the angle between the normals to the planes can also be specified. Complex queries can be generated by either textual shorthand or graphically using a java applet. A search for a ligand environment containing Cys or His, Cys or His, Gln or Asn and Arg or Lys can be expressed as the input query "CYS|HIS CYS|HIS GLN|ASN ARG|LYS." Requesting all ligand environments containing stacked His or Trp side chains can also be specified as the text, "HIS|TRP>P3.5^0," wherein ">P3.5" means that the distance between ligand and amino-acid planes is less than 3.5 Å, "^0" means that the angle between the normals to the planes is 0° with deviation of ±5 degrees. Residue secondary structure and main chain–side chain atom interactions may also be specified using a notation such as "HIS/H:M HIS/H:M HIS/S:S," which searches all ligands that have no fewer than three residues in their environment and the first residue is HIS in a helix and interacts with a ligand by its main chain atoms, the second residue has the same specification as the first and the third residue is His in a strand and interacts with a ligand by its side-chain atoms. For metal sites, in addition to the environment description, a site geometry can be specified, such as tetrahedral, trigonal bi-pyramidal, square-based pyramidal or octahedral.

The above textual notation is non-trivial, and to facilitate easy query generation a java applet has been developed. This graphical query interface is provided where all of the above search attributes can be specified with dialog boxes to build complex queries. The interface was built with a biological content in mind and reflects specific features that are well-known to structural biologists. A query to find all ligands that have a planar group that is stacked in a parallel fashion between two residues of any combination of Phe, Tyr, Trp or His side chains is shown in Figure 2. The hit list of PDB entries includes the ligand 3MA, 6-amino-3-methylpurine, in PDB entry 3mag.[35] Here the ligand's purine ring is bound between the Tyr22 and Phe180 residues. The main finding in the article[35] describing the structure of 3mag is the dominant role of enhanced stacking interactions between methylated bases and protein aromatic side chains.

ProSite Pattern/Code Search

The pattern search is intended to perform the following functions: (1) list PDB entries whose sequence matches the given signature and (2) list PDB entries that interact with a small molecule by residues that are in the pattern. A pattern is specified in terms of a ProSite identifier code or as text, and any pattern that is compliant with ProSite may be used, such as P-x-[STA]-x-[LIV]-[IVT]-x-[GS]-G-Y-S-[QL]-G. A graphical interface to construct these patterns similar to that used for active sites is being developed, although a ProSite signature does not represent such a problematic format. It is not necessary to define a known

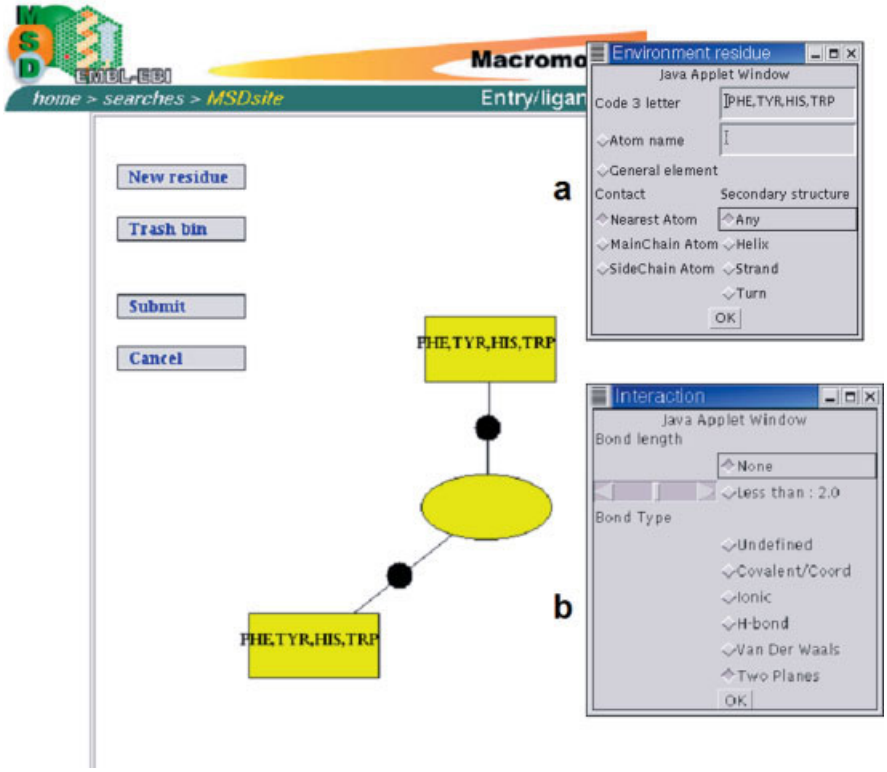

Fig. 2. The MSDsite Java Applet for biological query generation. The query "PHE|TYR|HIS|TRP>P^0 PHE|TYR|HIS|TRP>P^0" was generated graphically using (a) the residue option menu and (b) the interaction menu (shown as inserts). Twenty hits were found for ligands having a ring/plane group stacked between two planar amino-acid side chains (Using the default angle deviation of 5°).

ProSite signature due the implementation of the short sequence alignment method within MSDsite allowing research into new signatures by any user.

An example of a pattern search could use the Prosite identifier PS00075 (dihydrofolate reductase signature). One may restrict the search to only those proteins that also have bound methotrexate inhibitor (MTX), PDB hetgroup code MTX, by checking the box marked "Interacts" and entering MTX in the "Hetero" field [Fig. 3(a)]. MSDsite has options to further refine patterns associated with a ligand, and this can be illustrated using the information returned for PDB entry 1DF7[36] (returned from the first search). An excerpt of the web page details about its sequence interaction with ligands is shown in Figure 2(b) and is a view of the sequence details page shown in Figure 1.

The region of sequence, LPAKVRPLPGR illustrated in Figure 3b, may be used as the basis of a pattern search and by adding the restriction that Arg and Pro are part of the ligand binding environment then a set of PDB entries is returned that has the sequences in this region of:

MTX * * * *

1df7: L-P-A-K-V-R-P-L-P-G-R

1dra: W-E-S-I-G-R-P-L-P-G-R

3cd2: I-P-L-Q-F-R-P-L-K-G-R

3dfr: E-S-F-P-K-R-P-L-P-E-R

3drc: W-E-S-I-G-R-P-L-P-G-R

4dfr: W-E-S-I-G-R-P-L-P-G-R

Generalizing the pattern, we have [LIWE]-[PES]-x(3)-R-P-L-[PK]-[GE]-R, which can be used in the "ProSite pattern statistics" page of MSDsite (Fig. 1: statistics of sequence) where we search the pattern statistics [Fig. 4(a)]. This page provides the means to search for statistical information regarding an interaction pattern with respect to which ligands bind to the pattern. The service creates two charts, both of which are search interfaces; the first shows ordinary statistics as a bar chart, while the second presents normalized results as a percentage over all the ligands, so non-specific binding sites information is truncated [Fig. 4(b)]. This method is biased because it does not take account of sequence homology, and a new method to normalize the result is being developed. It should be noted that generating statistics on ligand binding from representative sets of PDB entries either by fold or by sequence is not valid. Often site-directed mutagenesis has been directed at active site residues, and in these cases (such as the studies on mannose binding protein) the specificity has been changed, as in PDB entry 1FIF.[36] The statistics formula currently used is as follows:

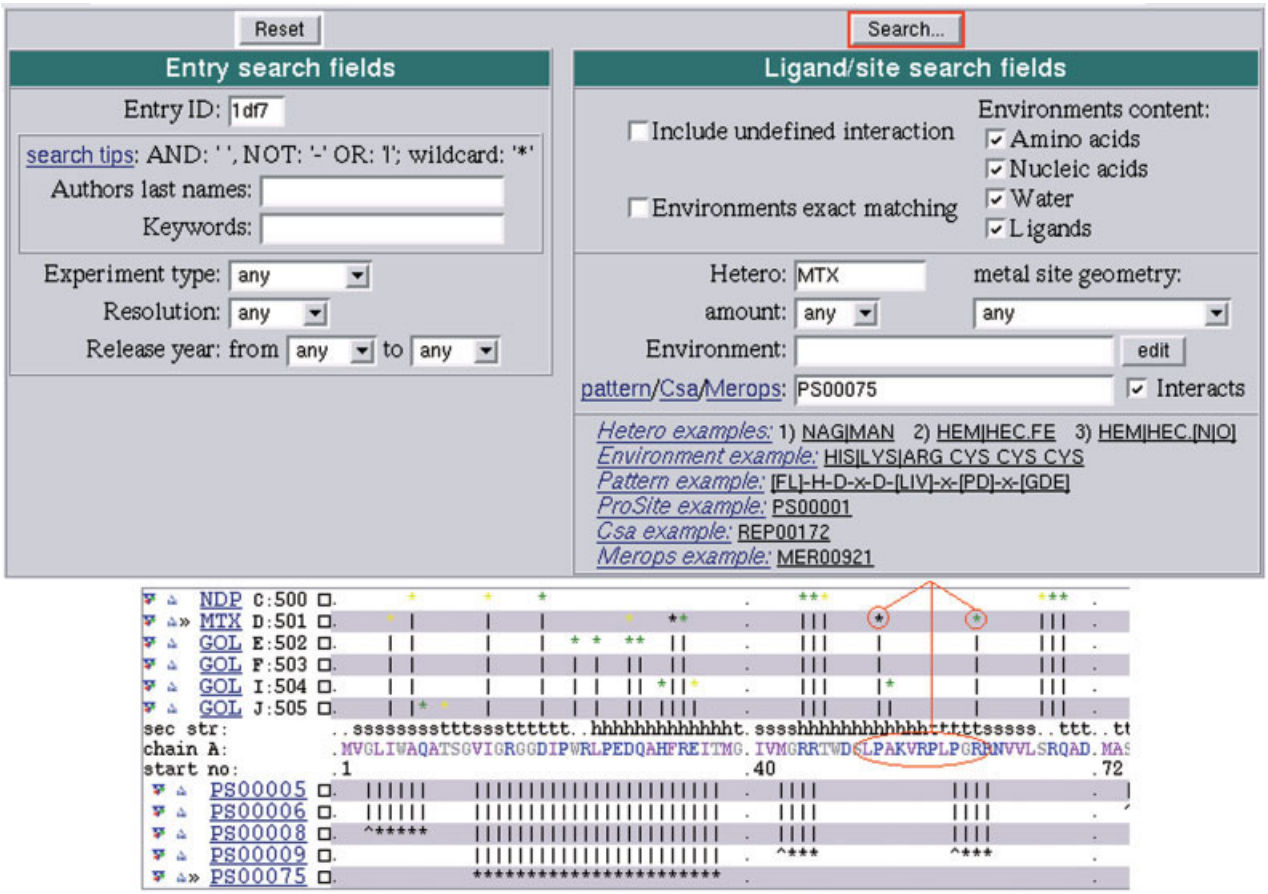

Fig. 3. (a) Main query page showing filter values to search for entry ID 1df7 with ligand MTX that interacts within ProSite pattern PS000075. (b) Extract from web page output for PDB entry 1df7 showing the residues matched to the Prosite patterns and those residues in contact with each ligand in the entry. For the MTX, the residues Pro51 and Arg60 are highlighted as residues that lie outside of the ProSite pattern PS00075.

$$f(\text{ligand}) = (\text{amount of ligand interacting with pattern}) / (\text{total instances of ligand})$$

$$\text{chart}\ (\%) = 100 \times f(\text{ligand}) / \{\text{summ}\ [f(\text{ligand})]\ \text{for all ligands}\}$$

Ligand Statistics Search

The interactive statistics page (Fig. 1: statistics) provides a means to view statistics on the interaction of a ligand with respect to residues that define an environment. The environment is a set of neighbor residues that interact with the ligand through a number of interaction types. This option may be used to obtain the following information: (1) which residue types preferentially interact with particular ligands, (2) which residue set preferentially interacts with a ligand in the PDB archive and how often, (3) which pair/triplet/etc. of residues are preferable for a specified ligand, (4) as above, with respect to a secondary structure or combination of a secondary structure and a residue type and (5) as above, with respect to a part of a ligand (group of atoms) in which an atom can be specified by name or general element.

Comparison of Adenine and Guanine Binding to Amino Acid Properties

MSDsite can be used to compare the binding of different ligands, as to display the statistics about which amino acids interact with the base rings of ATP and GTP (Fig. 1: statistics for ligands). This can be carried out by entering "ATP|ADP|A.N6|N1|N3|N7|N9|C2|C8" into the first "Hetero" field of the form and "GTP|GDP|G.N2|N1|N3|-N7|N9|O6|C8" into the second and restricting distribution to amino acids only, as shown in Figure 5(a). The search provides the chart shown in Figure 5(b).

Multiple Ligand Superposition

Selecting a list of desired ligands with their environment to the multi-view page and choosing the option "align by ligand" can superpose a set of results (Fig. 1: Overlay). A sample multi-view page is shown in Figure 6(a); the resultant superposition by ligand atomic coordinates is shown in Figure 5(b).

Discussion

MSDsite implements a research concept wherein every results page is a search form that creates a well-connected

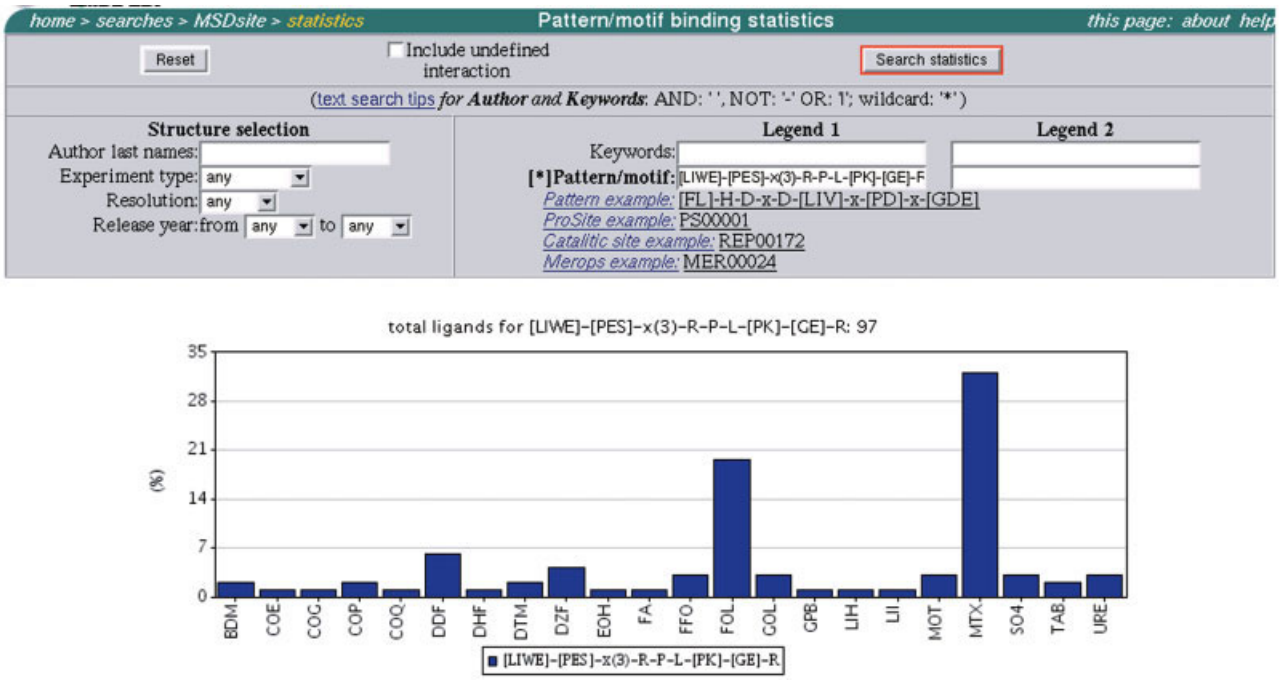

Fig. 4. (a) Sequence statistics query page to return the ligand binding statistics for the sequence environment [LIWE]-[PES]-x(3)-R-P-L-[PK]-[GE]-R. (b) Statistics for ligands interacting with the pattern [LIWE]-[PES]-x(3)-R-P-L-[PK]-[GE]-R.

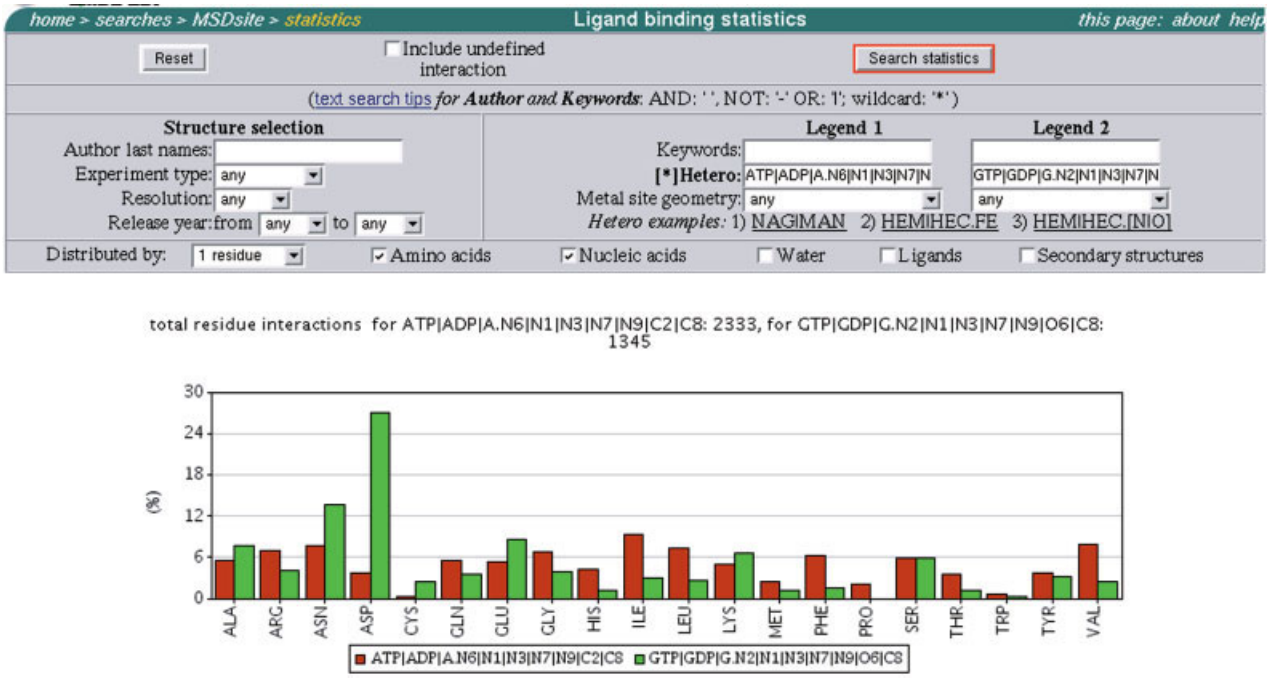

Fig. 5. (a) Ligand binding statistics query page used to search for the interactions with the base rings of ATP/ADP compared to GTP/GDP. The results of this search are shown in (b). (b) The distribution of amino acids interacting with adenine of ATP, ADP, A (red) and guanine of GTP, GDP, G (green). Guanine has a higher preference for ASP, CYS, GLU than does adenine. This agrees with the results found in an extensive study on the molecular discrimination between adenine and guanine by proteins.[37]

graph in which nodes are web pages and edges are links between them (see Fig. 1). Using this approach, a researcher can go from the start page to the hit list to the details pages and then form a query based on these, for which the database returns another hit list, which can be used to amend the search, get a statistics chart whose bars can be selected to get corresponding hit lists and so on. The functionality provided is based on chemical and biological concepts implemented using software integrated with relational database technology and the use of J2EE as a

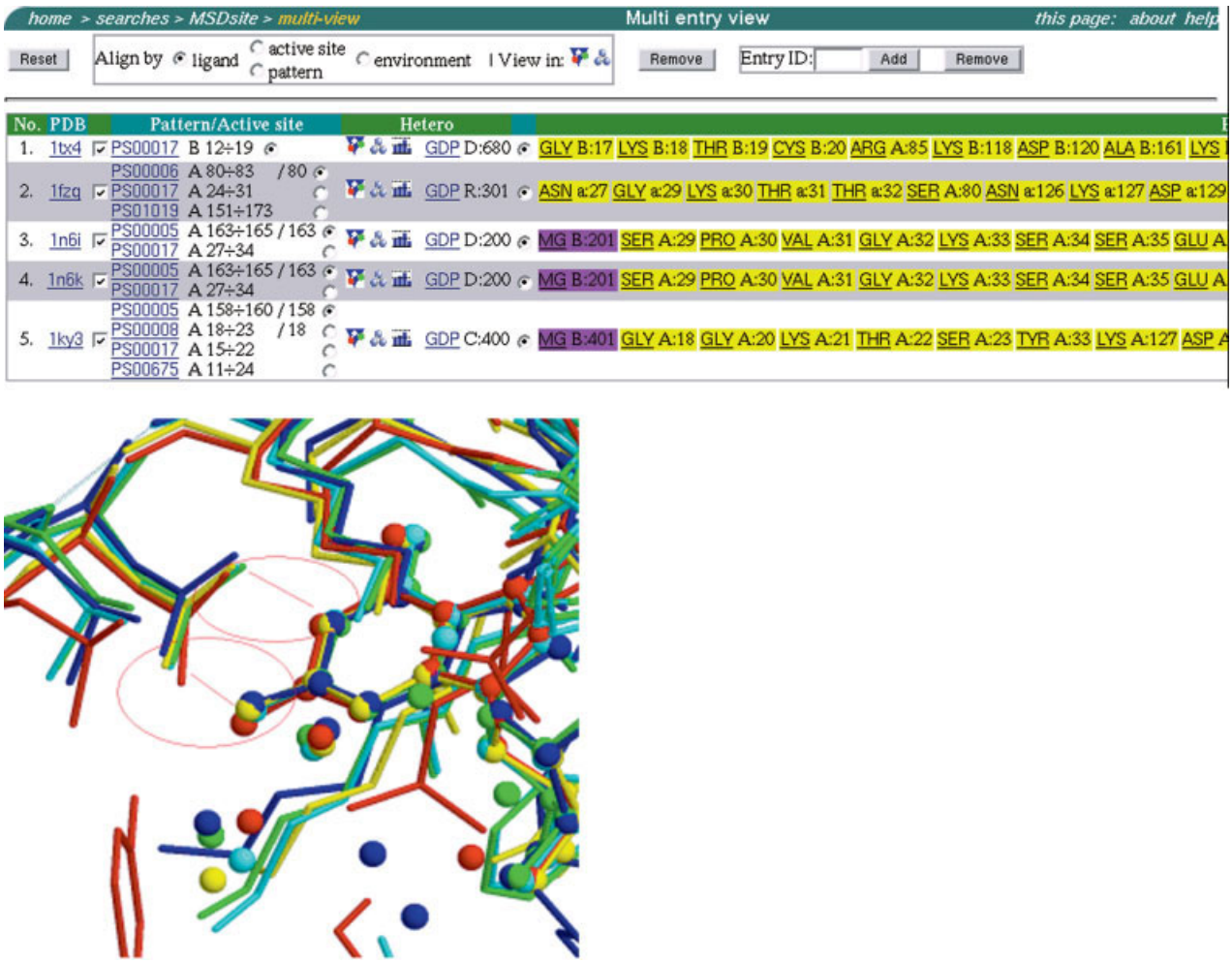

Fig. 6. (a) Multi-view page showing five proteins that interact with GDP, which are aligned by ligand to produce (b). The interacting residue lists are truncated within the figure. (b) Superposition display option. A selected set of guanine binding showing the guanine N1 and N2 hydrogen bonding to OD1 and OD2 of ASP in a series of homology structures around the ligand [1fzq,[38] 1n6k,[39] 1n6i,[39] 1ky3,[40] 1tx4[41]].

platform for the web services. Current developments include the incorporation of a sub-structure chemical search interface and 3D active site search and retrieval methods based on the annotated site details from CATRES[27,28] and MEROPS.[29,30]

Each web page of MSDsite has an "about" link that defines the aim and overall functionality of the web page and a "help" link with detailed descriptions of each active component. This style of documentation conforms to a standard used throughout the MSD services. Two MSDsite tutorials are available on the MSD education pages (http://www.ebi.ac.uk/msd/education/Tutorial.html); the first works through the basic functionality and the second represents a scientific case study. A full description of the Oracle MSD schema can be found at (http://www.ebi.ac.uk/msd-srv/docs/dbdoc/), and details specific to this article can be found under the "Active Site" mart.

The MSDsite is under active development as part of the overall design and function of the MSD services. We are aware of a number of shortcomings in the service, and these will be addressed during the normal development cycles. As to the data itself, there are ambiguities within the classification of the interaction types, such as the division between a hydrogen bond and an electrostatic interaction. We note that there is a continuous spectrum of electronic structure for many bonding types, which can result in classification problems where the chemistry is not obvious. There are limitations in the service (e.g. ring planarity description) that result because MSDsite is a portal to an extremely complex database, and these data are stored but do not have references within the web interface. The reason for this is that the inclusion of user-defined filters for every stored attribute would result in an overly complex interface. We are therefore actively designing more intuitive methods for query building to open up more functionality; the active site sketcher described in this article represents one such example.

ACKNOWLEDGEMENTS

The project was funded by the European Commission as the TEMBLOR, contract-no. QLRI-CT-2001-00015 under

the RTD programme "Quality of Life and Management of Living Resources."

REFERENCES

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
2. Bernstein FC, Koetzle TF, Williams GJ, Meyer EE, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. Protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
3. Hendlich M. ReLiBase Databases for protein-ligand complexes. Acta Crystallographica 1998;D54:1178–1182 (see http://relibase.ebi.ac.uk).
4. Bergner A, Günther J, Hendlich M, Klebe G, VerdonkM. Use of relibase for retrieving complex 3D interaction patterns including crystallographic packing effects. Biopolym Nuc Acid Sci 2002;61:99–110.
5. Reichert J, Jabs A, Slickers P and Sühnel J. The IMB jena image library of biological macromolecules. Nucleic Acids Res 2000;28:246–249 (see http://www.imb-jena.de/IMAGE.html).
6. Reichert J and Sühnel J. The IMB jena image library of biological macromolecules: 2002 update. Nuc Acid Res 2002;30:253-254.
7. Ivanisenko VA, Grigorovich DA, Kolchanov NA. PDBSite: a database on biologically active sites and their spatial surroundings in proteins with known tertiary structure. The Second International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2000). Novosibirsk, Russia, August, 2000;2:171–174 (see http://wwwmgs.bionet.nsc.ru/mgs/gnw/pdbsite/).
8. Etzold T, Ulyanov A, Argos P. SRS: information retrieval system for molecular biology data banks. Meth Enzymol 1996;266:114–128.
9. Degtyarenko KN, North ACT and Findlay JBC. PROMISE: a database of bioinorganic motifs. Nucleic Acids Res 1999;27:233–236.
10. Castagnetto JM, Hennessy SW, Roberts VA, Getzoff ED, Tainer JA, Pique ME. MDB: the metalloprotein database and browser at the Scripps Research Institute. Nucleic Acids Res 2002;30:379–382.
11. Wallace AC, Laskowski RA and Thornton JM. Derivation of 3D coordinate templates for searching structural databases: application to the Ser-His-Asp catalytic triads of the serine proteinases and lipases. Prot Sci 1996;5:1001–1013.
12. Wallace AC, Borkakoti N, Thornton JM.TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases: application to enzyme active sites. Prot Sci 1997;6:2308–2323.
13. Santander V, Portales MA and Melo F. A tool to assist the study of specific features at protein binding sites. Bioinformatics 2003;19:250-251.
14. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB and Thornton JM. CATH- a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
15. Boutselakis H, Dimitropoulos D, Fillon J, Golovin A, Henrick K, Hussain A, Ionides J, John M, Keller PA, Krissinel E, McNeil P, Naim A, Newman R, Oldfield T, Pineda J, Rachedi A, Copeland J, Sitnov A, Sobhany S, Suarez-Uruena A, Swaminathan J, Tagari M, Tate J, Tromm S, Velankar S, Vranken W. E-MSD: the European Bioinformatics Institute macromolecular structure database. Nuc Acid Res 2003;31:458–462.
16. Sigrist CJA, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 2002;3:265–274.
17. Krissinel EB, Winn MD, Ballard CC, Ashton AW, Patel P, Potterton EA, McNicholas SJ, Cowtan KD, Emsley P. New CCP4 coordinate library as a toolkit for designing the coordinate-related applications in protein crystallography. Acta Cryst D 2004, In press.
18. Lindauer K, Bendic C, Suhnel J. HBExplore – a new tool for identifying hydrogen bonding patterns in biological molecules. Comput Appl Biosci 1996;12:281–289.
19. Pedretti A, Villa L, Vistoli G. Atom-type description language: a universal language to recognize atom types implemented in the vega program. Theor Chem Acc 2003;109:229–32.
20. Gasteiger J, Rudolph C, Sadowski J. Automatic generation of 3D-atomic coordinates for organic molecules. Tetrahedron Comp Method 1990;3:537–547.
21. Sadowski J, Gasteiger J, Klebe G. Comparison of automatic three-dimensional model builders using 639 X-ray structures. J Chem Inf Comput Sci 1994;34:1000–1008.
22. Balducci R, Pearlman RS. Efficient exact solution of the ring perception problem. J Chem Info and Comp Sci 1994;34:822–831.
23. Harding MM. Geometry of metal-metal interactions in proteins. Acta Crystallogr 2001;D57:401–411.
24. Harding MM. Metal-ligand geometry relevant to proteins and in proteins: sodium and potassium. Acta Crystallogr 2002;D58:872–874.
25. Harding MM. The geometry of metal-ligand interactions relevant to proteins. Acta Crystallogr 1999;D55:1432–43.
26. Alberts IL, Nadassy K, Wodak WJ. Analysis of zinc binding sites in protein crystal structures. Prot Sci 1998;7:1700–1716.
27. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. J Mol Biol 2002;324:105–121.
28. Porter CT, Bartlett GJ, Thornton JM. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nuc Acid Res 2004;32:D129–D133.
29. Rawlings ND, Tolle DP, Barrett AJ. MEROPS: the peptidase database. Nuc Acid Res 2004;32:160–4.
30. Rawlings ND, Tolle DP, Barrett AJ. Evolutionary families of peptidase inhibitors. J Biochem 2004, In press.
31. Oldfield TJ. Data mining the protein data bank: residue interactions. Proteins SFG 2002;49:510–528.
32. Hobbs L, Hillson S. Oracle8i data warehousing. Digital Press; 1999:400.
33. Sayle RA and Milner-White EJ. RASMOL: biomolecular graphics for all. Trends Biochem Sci 1995;20:374.
34. Oldfield MJ. A Java applet for multiple linked visualization of protein structure and sequence. J Comp Chem, In press.
35. Hu G, Hodel AE, Gershon PD, Quiocho FA. mRNA cap recognition: dominant role of enhanced stacking interactions between methylated bases and protein aromatic sidechains. Proc Nat Acad Sci USA 1999;96:7149–7154.
36. Feinberg H, Torgersen D, Drickamer K, Weis WI. Mechanism of pH-dependent N-acetylgalactosamine binding by a functional mimic of the hepatocyte asialoglycoprotein receptor. J Biol Chem 2000;275:35176–35184.
37. Nobeli I, Laskowski RA, Valdar WS, Thornton JM. On the molecular discrimination between adenine and guanine by proteins. Nuc Acid Res 2001;29:4294–309.
38. Hillig RC, Hanzal-Bayer M, Linari M, Becker J, Wittinghofer A, Renault L. Structural and biochemical properties show ARL3-GDP as a distinct GTP binding protein. Structure Fold Des 2000;8:1239–45.
39. Zhu G, Liu J, Terzyan S, Zhai P, Li G, Zhang XC. High resolution crystal structures of human Rab5a and five mutants with substitutions in the catalytically important phosphate-binding loop. J Biol Chem 2003;278:2452–60.
40. Constantinescu AT, Rak A, Alexandrov K, Esters H, Goody RS, Scheidig AJ. Rab-subfamily-specific regions of Ypt7p are structurally different from other RabGTPases. Structure (Camb) 2002;10:569–79.
41. Rittinger K, Walker PA, Eccleston JF, Smerdon SJ, Gamblin SJ. Structure at 1.65 Å of RhoA and its GTPase-activating protein in complex with a transition-state analogue. Nature 1997;389:758–62.

APPENDIX A

The following tables were created in Oracle, with bitmap indexes built on selected columns: (1) RESIDUE_INTERACTIONS for ligand—macromolecule interactions at the residue level, (2) ATOM_RESIDUE_INTERACTIONS for ligand atom–macromolecule residue interactions, (3) ATOM_INTERACTIONS for ligand–macromolecule interactions at the atomic level.

Ligand searching requires specific tasks that cannot be resolved using simple basic SQL relational operators like join and merge. For fast query generation, a web applica-

tion server has been developed that contains java classes that compose complex SQL queries and perform calculations on the dataset. In addition, the java classes regulate the SQL to enforce the correct use of indexes and apply query hints. The SQL queries are designed to use the relational algebra operation 'INTERSECT,' which allows execution within the Oracle RDBMS in parallel without nesting and is faster than self-joins for those cases in which the result set is many times less than the size of the table queried.

The approach used is based on a star architecture query in which the ligand is central, and interactions to the environment residues fan out. This design is used because it results in an algorithm of order 'N' with regard to the number of interactions and is therefore scalable for complex active sites. The following example defines a zinc finger site query, in which the environment consists of three histidine residues and one cystine residue: select entry_id, ligand_id from residue_interactions where ligand_code_3_letter = 'ZN' and neighbor_code_3_letter = 'HIS' group by entry_id, ligand_id having count(*) = 3 intersect select entry_id, ligand_id from residue_interactions where ligand_code_3_letter = 'ZN' and neighbor_code_3_letter = 'CYS'

There are two levels of optimization. First, the query is split into two separate parts, which can be executed in parallel and then combined by intersection. Second, the largest table of atom interactions is used only when a user performs a search at the atomic level, as when atom names or general elements are specified. The table hierarchy is therefore reflected within the query design and is dependent on the level of detail the user requests.

APPENDIX B

Short sequence alignment is performed within the MSDsite service using SQL queries based on a database table that holds the unique protein chain sequences for the PDB. The primary key ID for this table is derived from the sequence number with gaps between chains flagged with the numerical insert of 100. To demonstrate the advantages of this, consider the following method to select all ligands that interact with the part of a protein chain that matches a specified pattern. The SQL for this pattern N-{P}-[S,T]-{P} is as follows: with (select /*+ LEADING(r0) USE_NL(r0) INDEX_COMBINE(r0) */ r0.id - 0 as start_id, 4 as length from seqs r0 where r0.CODE= 'N' and r0.id +1 in (select /*+ USE_NL(seqs) */ id from seqs where CODE != 'P') and r0.id +2 in (select /*+ USE_NL(seqs) */ id from seqs where CODE in ('S','T')) and r0.id +3 in (select /*+ USE_NL(seqs) */ id from seqs where CODE !='P')) select /*+ LEADING(m) INDEX(ls ligand_seqs_seqs_ind) */ ls.ligand_id from m, ligand_seqs ls where ls.seqs_id between m.start_id and m.start_id + m.length – 1;

The query uses self joins of the table SEQS as many times as the number of items in a search pattern and these joins are made by the SEQS primary key. The table with alias r0 is accessed by its bitmap index built on the CODE column, and the Oracle hint 'LEADING' is specified to define this as the leading table in the query. All other joined SEQS are accessed by ID, their primary key. Use of the Oracle hint mechanism 'USE_NL' implicitly forces access to nested tables by the primary key. Access is fast because primary keys of a searched sequence lie mostly on the same database page and are read when the first table is accessed using the alias r0. The primary keys were specifically designed to lie off of the page boundary as part of the design philosophy. It can be seen from the example that all nested queries are relative to the leading one, where we chose the leading residue as the most rare within the MSD. A second table in the query is LIGAND_SEQS that integrates the sequence query and ligand interactions. The table idenfities the relationships between the ligand table and its interacting residues from the sequence table (SEQS). This query design provides rapid answers (in less than 3 s) for the current known macromolecular structures and will also service the expected exponential increase in structure submissions for several years with the current hardware.