Analytical Processing of PDB Ligand Fragments and Protein Environment directly on the Web

Dimitris Dimitropoulos *, Kim Henrick, Jawahar Swaminathan, Adel Golovin
Macromolecular Structure Database
EMBL Outstation, The European Bioinformatics Institute
Welcome Trust Genome Campus, Hinxton, Cambridge, UK
http://www.ebi.ac.uk/msd-srv/msdmine

Abstract: - The PDB protein-ligand complexes are the main public resource that provides detailed chemical environment information for atomic level interactions between the ligand and the protein. In order to investigate relevant research questions PDB users have often to extract this data manually or semi-automatically. The public MSD search relational database hosted by the EBI has been designed to address exactly this type of research providing data on ligand chemistry, secondary structure assignment, binding site interactions and cross-references to sequence and protein classification databases. Furthermore, the MSD-mine www service offers the capability to perform ad-hoc flexible queries, interactive statistical chart generation and analytical processing, directly on the web.

Key-Words: - Protein Structure Databases; Ligands; Binding Sites; Decision Support Systems, Management; Data Analysis; Statistical Study; Tables and Charts

1 Introduction
The Protein Data Bank [1] has been a valuable resource for many years, in understanding the biological significance of ligands and the way they interact with macromolecule function, in 3D detail. While the archive was evolving and increasing in size, researchers started to treat it as a collection, trying to analyse it systematically looking for hidden patterns, for use in various areas like structure prediction, active site identification [2] and drug discovery. As the PDB is a complex and rich information source that covers areas of chemistry, 3D structure and protein sequence, such work can also combine and correlate data and investigate information from different viewpoints.

Relational databases and data warehouses [3], by offering flexible ad-hoc queries and data aggregation, are excellent tools for this type of research, by offering online analytical processing and data-mining capabilities.

Several previous initiatives have been undertaken with the aim of loading contents of the PDB into a relational database [4], [5] and while these databases are valuable tools for the search and retrieval of individual entries, they are of much less value in performing data analysis and mining tasks because most statistical analysis resources of the PDB are based on static pre-calculated data [6].

The Macromolecular Structure Group (MSD) at the European Bioinformatics Institute (EBI), one of the three partners that constitute the wwPDB (worldwide Protein Data Bank) consortium [7] adopted relational database technology as the main infrastructure for its activities at an early stage. In addition to acting as a deposition centre, the MSD produces and maintains the MSD Search Database (MSDSD), which forms the core of its public systems and services. The database is publicly available [8] and is distributed in a relational format on Oracle and mySQL to collaborators and research groups.

MSDSD incorporates data derived consistently from the original PDB archive and serves as a repository for extracted, integrated and homogenized scientific knowledge. For example, secondary structure information has been consistently re-assigned using DSSP [9] and cross-references have been established and stored for SCOP (structural classification of proteins), GO (gene ontology), PFAM (protein family), EC (Enzyme), Pub-Med and UniProt external databases [10].

MSDSD database design follows the data-warehouse paradigm and can be used for ad-hoc querying, analytical processing and data mining.

However, this is not always sufficient since most biologists are not familiar with SQL and data warehousing concepts or tools. They may also not be ready to invest resources involved in downloading and using MSDSD locally, especially when data analysis is only an occasional need in the overall scheme of their research.

The MSD-mine web application has been developed as part of the MSD toolset in order to provide a public on-line service that allows the exploitation of the MSDSD for data analysis and knowledge mining. The analysis capabilities of this tool are described in this paper with examples illustrating potential links
between ligand chemistry and the protein environment characteristics.

2 Chemical Data Modelling

One of the big limitations of the PDB archive is the lack of an underlying data model to allow for exact chemical ligand identification. This has been addressed in the wwPDB with the exchange of a reference ligand dictionary [11] that incorporates work done at the time of curation of new PDB entries where ligand chemistry issues are carefully resolved in cooperation with the depositors. The ligand dictionary provides the chemical definition of the conventional 3-letter-codes in the PDB, but what is missing is a more abstract view of common chemical functional areas. For example, the biotin functional group, which plays an important role in biomolecule immobilization, is present in more than 5 different ligands. It is reasonable to assume that researchers would be interested in doing analysis on the common functional fragment of biotin in all these ligands instead of using the ligands themselves, particularly when looking at active site patterns or drug discovery targets.

In the MSDSD we have compiled a list of about 90 common functional fragments, automatically identified their occurrences in the reference ligands, and have added associations on the molecular and atomic level. Compared to other collections of functional groups [12], these fragments (http://www.ebi.ac.uk/msd-srv/msdchem/ligand/fragform.htm), were chosen to be characteristic enough in order to locate real pharmacophores and they were referenced on the atom level. For example, due to molecule symmetry, the 10 atoms of a naphthalene fragment (a pair of conjugated aromatic benzene rings) can be divided into 3 types (Fig 1.):

- The 2 bridgehead atoms (on the common bond of the 2 rings) named as C10
- The 4 ortho-to-bridgehead atoms (next to the atoms of the common bond) named as C4
- The 4 meta-to-bridgehead (outer atoms) named as C3

The atom type identifiers like C10, C4 and C3 are neither significant nor based on a systematic naming scheme. Instead they are chosen from the most frequent corresponding atom names in occurrences of fragments in PDB, with the idea that will be more familiar to PDB users in certain types of chemistry like peptides and sugars.

As an example for the PDB ligand, “FR3” (Fig. 2), its naphthalene matching atoms associated with the naphthalene fragment atom types atoms are:

- C28,C29 with C10
- C2,C4,C27,C22 to C4
- C6,C7,C25,C26 to C3.

Fig 2. FR3 ligand including a naphthalene fragment

2.1 On-line Analysis

Providing data analysis and mining services based on relational database infrastructure on the web has been traditionally considered as technically difficult. Data analysis operations require many resources and typically only a limited number of users can be supported. So most on-line web services provide access to a small number of predefined optimised searches and do not exploit in full the capabilities of relational technology. However, with advances in technology and the reduction of cost, the on-line exploration and data mining of a complex public scientific data warehouse is becoming technically practical, especially since modern databases may be instructed to impose limits on over-demanding queries.

Exposing the database infrastructure to the public is not always enough since even if the end users are experienced with SQL querying - which is not often the case - they will certainly not be familiar with the particular database design. A challenge is to provide an environment easy to follow and understand that allows building data analysis queries, offering handy documentation details on the contents of tables and columns and a fast visual demonstration of the analysis results.
3 The MSD-mine system

The MSD-mine pages (in production since 2003 at http://www.ebi.ac.uk/msd-srv/msdmine) allow web browser users to build complex ad-hoc queries, execute them and visually analyse their results in the form of interactive charts and histograms, providing a framework for on line analytical processing. Additionally, it includes a documentation component that ensures that detailed context sensitive descriptions of entities, attributes and relations of the database are conveniently available when needed. The aim of MSD-mine is to allow researchers to explore and understand the structure of the database, learn how to combine and filter information and provide a flexible way to perform any ad-hoc queries and view charts and statistics on MSDSD.

3.1 Analysing Fragments

We will start using this tool in a first example to analyse naphthalene fragment data by joining information from the tables “Ligand” (reference ligand), “Ligand Fragment”, “Residue” (bound molecule instance) and “Chain” (bound molecule group) of the corresponding sections of the database.

This is illustrated in Fig. 3, which gives a part of the web page to carry out this set of table joins. The operation of the “Query” page is based on the node of the query tree that is currently selected. As soon as users click and select nodes of the query tree, they may add restrictions, join relationships or refine the result attributes of each node and view relevant context sensitive description text.

Now we have built a query that will encounter all fragment occurrences in PDB bound molecules and by executing it we get results that show the corresponding PDB accession code, chain code, serial and 3-letter code together with the name of the fragment that it includes (see Fig. 4).

In the result page we get the chemical diagram of the corresponding ligand and links to external pages with more specialised information. Going back in the “Query” page we will need to add certain restrictions (Fig. 5) in order to use bound molecules (ignore polymers and water groups) that include naphthalene.

We can generate statistical charts from the results of the new modified query. For example in order to see the distribution of molecules sizes we just have to click on the corresponding icon in the header of the relevant column (Fig. 6).

Fig. 3. The MSD-mine query tree that joins all relevant information about ligand fragments

Fig. 4. MSD-mine fragment results that display the PDB bound molecules and the fragment they contain. On the header of each data column there are icons for: adding restrictions based on result data, analyse the column, use it to group the analysis of another column, and sort results.
Fig. 6. Histogram of size (in terms of the number of non-hydrogen atoms), for ligands that include naphthalene as generated by MSD-mine

3.2 Using Protein Classification
We will extend our query to include the proteins that these ligands interact with, and their SCOP classification [13] (Fig. 7).

Fig 7. The MSD-mine query tree that joins information from the fragment classification of reference ligands, up to the SCOP classification of their associated protein environment.

Following the same steps as above, it is possible to analyse and find the most common SCOP domains that interact with ligands that contain the naphthalene fragment. We see that some of the most common ones in the result chart are (Fig. 8)

- b.47.1.2, Eukaryotic proteases
- d.117.1.1, Thymidylate synthase/dCMP hydroxymethylase
- d.17.4.3, Delta-5-3-ketosteroid isomerase, steroid delta-isomerase

The “Analysis” page enables us to “drill-down” by clicking on the corresponding areas in the interactive chart, collecting all the values of interest and choosing to “Add” the selected values in the filter.

**MSD-mine SCOP per chain**

![MSD-mine SCOP per chain chart](chart.png)

Now that we have drilled-down on an analysis area of interest, we can roll-up and analyse the new result set from a different viewpoint by removing the naphthalene constraint and analyse the fragment name attribute. This will give us other chemical fragments that associate with these three domains (Fig. 9).

**Other fragments that often associate with proteins classes that interact with naphthalene given via an MSD-mine roll-up operation**

![Other fragments chart](chart2.png)

3.3 Fragment Atom Types
In another example we will investigate the binding site characteristics [14] of naphthalene atom types.
We need to join information from some other entities on the atomic level like binding sites and their pair of contacts. Now the restrictions that we add is that we are interested in naphthalene, and for simplicity use interactions only with alpha carbons from the polymer side, by restricting the atom name of the second binding site to “CA” as shown in Fig. 10.

![Fig. 10. MSD-mine query to retrieve data for binding site characteristics of naphthalene atoms with amino acid alpha-carbons](image)

From the result data we can get information about the most common amino acids where the alpha carbon interacts with naphthalene (Fig. 11).

![Fig. 11. MSD-mine analysis chart of the 10 most common amino acids with alpha carbon interactions to naphthalene](image)

In order to get detailed statistics about the strength of the interaction for different atom types of naphthalene, we go back into the results page and use the relevant analyse icon above the “Atom name” attribute of the “Fragment Atom” entity so that it becomes the basis of our analysis. Then we can analyse the “Distance” attribute of the contacts for different fragment atom types, specifying that we would like to get the minimum contact distance for each fragment atom (Fig. 12).

![Fig. 12. MSD-mine fragment atom type based analysis of the minimum contact distances of naphthalene with alpha carbons](image)

We would like then to investigate the effect of outliers in these results, so we need to analyse contact instances in general, and work in a region of strong contacts but avoiding outliers. By following steps described above, we do an analysis of contact distances (removing the analysis base first), in order to generate an interactive histogram that will enable us to drill down our research in the region of interest (Fig. 13).

![Fig. 13. Drilling down naphthalene – alpha carbon contacts in a region of strong contacts, avoiding outliers marking the beginning (minimum value) and end (maximum value) of the region of interest.](image)

Now we can roll-up back on our new results and examine the most common amino acids with strong...
contacts and the contribution of the three different naphthalene atom types in them (Fig. 14).

Fig. 14. Rolling up to the contribution of naphthalene atom types in naphthalene – alpha carbon contacts in the strong contacts region of interest

4 Conclusion

Further discussion of the implications of the actual results of this analysis is out of the context of this demonstration. What we have shown here is the concept of working with ad-hoc querying and online analytical processing directly on the web by using a pre-built data warehouse like MSDSD. It may have already become obvious that these are just a couple of random examples from the countless research paths that MSD-mine exposes to the researcher. In any analysis, bias introduced by the data set can be avoided by using well-known predefined representative sets of PDB entries, an option that MSD-mine already offers. However, data analysis and drawing conclusions based on such analysis are two distinctly different processes. In order to authenticate any conclusions, it is important for the researcher to be certain that there were no mistakes in their approach and cross-validate the MSD-mine results using other methods, wherever possible. Our experience based on user feedback suggests that most users find MSD-mine to be a very good environment for rapid evaluation of new ideas and getting introduced to the data warehouse technologies.

References: