Proceedings of ANAS (Biological Sciences), 65(5-6): 168-176 (2010)

A Redesigned Database of Three-dimensional Protein Structures for Reuse in Macromolecular Crystallography

Alexei A. Vagin, Fei Long, Garib N. Murshudov*
York Structural Biology Laboratory, Chemistry Department, University of York, Heslington, York, YO10 5DD, UK

The number of macromolecular structures solved and deposited to the Protein Data Bank (PDB) is more than 66000. Exploiting this information for macromolecular crystallography (MX) dramatically increases the efficiency of MX structure solution. This paper describes a database of "unique" PDB entries with their domains and potential oligemeric states. The database is used, as a proof of principle, in one of the automatic MX structure solution pipelines. Tests with the new entries from the PDB shows that the pipeline based on this database is able to solve around 80% of all solvable structures automatically.

Keywords: *macromolecular crystallography, protein structures, database, domains*

INTRODUCTION

The number of macromolecular structures deposited in the Protein Data Bank - PDB (Berman et al., 2000) is increasing rapidly every year. As of May 2010, this bank contains more than 66000 entries; the number of newly deposited structures in 2009 was more than 6700 (http://www.ebi.ac.uk/pdbe/). Models of macromolecular structures deposited to the PDB were derived using one of the several experimental or theoretical modelling techniques. X-ray crystal structure analysis (MX) is by far the most common technique used for the determination of 3D structures (approximately 84%), followed by NMR with around 14% (Figure 1). Popularity of MX can be explained by the fact it is the only technique enabling to give detailed 3D structures of macromolecules at the atomic level and this ability of MX has been inspiring rapid development of the technique in the last two decades. It is interesting to note that the number of theoretical models derived using by similarity search and energy minimisation (Schwede et al., 2003) is now around 1400. Although as the number of the entries in the PDB increases, homology modelling should be able to give more and more accurate structures, unfortunately currently this technique is far from ideal and does not always produce the most reliable structures and it is yet to replace structure elucidation using one of the experimental techniques (MX, NMR, EM). Up to the present time, the problem of theoretical macromolecular structure modelling remains an important and intriguing problem. A potential solution of the problem may be found by using modern statistical and knowledge-based approaches: 1) efficient use of information contained in the databases of macromolecules, their domains and/or smaller structural units and 2) application of such proven statistical techniques as Bayesian statistics, pattern recognition and machine learning methods.

The PDB is a treasure of the structural biology community, the implications of which have yet to be fully appreciated. One can imagine the amount of information contained in this repository. How do we extract and analyse this information and use it to understand fundamental biological problems such as protein folding and protein evolution? This and other questions are the subjects of many research disciplines including bioinformatics. There have already been huge amounts of work carried out in this area. Two areas relevant to this paper are the classification of domains (CATH – Pearl et al., 2005; SCOP - Murzin et al., 1995) and the extraction of biological oligomers from crystal structures (Krissinel and Henrick, 2005). While the domains defined by both CATH and SCOP are extremely useful for the biological community in general, our attempts to use them for MX structure solution did not produce consistent results. Therefore we undertook to redefine the domains so that they could be used for MX structure solution routinely and consistently.

One of the obvious applications of the PDB is its reuse for MX analysis. Some applications of information derived from the PDB at the various stages of MX analysis (Jones et al., 1991; Terwilliger and Berendzen, 1999; Emsley and Cowtan, 2004) have been a subject of study for some time

*E-mail: garib@ysbl.york.ac.uk

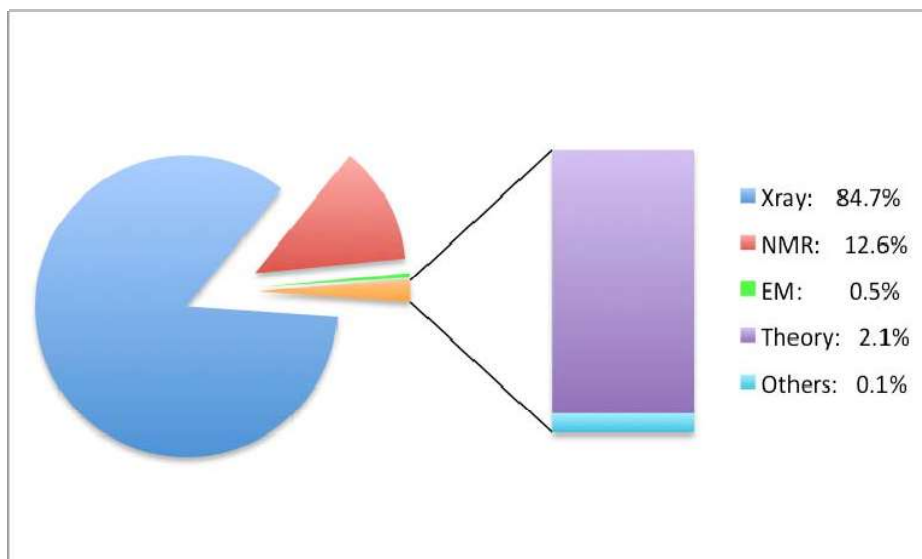

Figure 1. A pie chart showing the distribution of the methods used to derive models in the PDB released before May 2010. As it can be seen X-ray is the most popular technique. Single particle EM is an emerging experimental technique of structural biology. There are few but important structures that have been determined with this method. NMR is applied to relatively small macromolecules but has advantage that it can give indication about dynamics of the molecule. Theoretical models were popular in the late 90s early 2000, however in the last five years no theoretical models have been produced owing to recognition that this technique is not able to replace experimental results yet.

now. In the near future, one can envisage that the information invariant for all entries in the PDB (e.g. secondary structure elements, interatomic distance patterns) or classes of proteins (e.g. domains, oligomer formation) will be used during all stages of structure analysis, thereby transferring information from high-resolution structures to newly derived structures, thus increasing the reliability of the derived models. However, one should be careful in reusing the derived information for treatment of the future experimental data since if the extracted and used information is erroneous then this may persist throughout future model derivations and errors may infest the data bank. To avoid this, only (at least statistically) verifiable invariant information should be used.

One can also speculate that increased power of the PDB combined with dramatic advances in computer technology and computational techniques can help us to tackle such long standing problems as the celebrated phase problem; for instance by using substructure classes (e.g. domains) from the PDB then by applying such well-established ideas as the multi-solution techniques (Germain et al., 1970) used in the small molecular crystallographic world.

In recent years there has been an explosion of the developments of automatic procedures for MX analysis. These approaches have already produced several highly automated and very popular software packages, including automatic model building and refinement - ARP/wARP (Perrakis et al., 1999), automatic phasing and model building – solve/resolve (Terwiliger and Berendzen, 1999), CRANK (Ness et al., 2005) and Autorickshaw (Panjikar et al., 2005), automatic molecular replacement pipelines – BALBES (Long et al., 2008), NORMA (Delarue, 2008), MrBump (Keegan and Winn, 2008) and automation component of the JSCS structure solution pipeline (Schwarzenbacher et al., 2008).

The current contribution describes a redesigned database derived from the PDB and application of the database to the MX analysis. More detailed description of this application can be found in Long et al. (2008).

1. Database of macromolecular structures

Motivation

Although the PDB is an immensely rich source of information, its direct use for various purposes including MX analysis poses certain problems. For example, since in many cases protein structures and ligand-protein complexes are studied together as a part of larger biological projects, it can be expected that there are many redundant structures in the PDB. These redundant structures are important for investigation of the reaction mechanism or the binding properties of a

certain protein. However, statistical analysis of such redundant data without proper classification and within/between class analysis may result in extremely biased results. Therefore, if reliable information from the PDB to be extracted, redundancy analysis should be carefully carried out. Moreover, since the number of entries in the PDB is very large and increasing very rapidly, for a given amino acid sequence it is not obvious how to perform efficient search and find all homologous protein structures with immediate access to such properties as domains, oligomeric states. Moreover, PDB does not contain information about conformational variability of proteins and extracting oligomeric information routinely is not trivial. For efficient extraction of all relevant information from the PDB we have analysed all entries in this bank and organised a database so that it can be efficiently used for the treatment of structural biology experimental data in general and MX in particular. Although the current focus of our research is optimal treatment of MX experimental data, application of the designed database is not limited to this field. With further reclassification of domains, proper links between them the database potentially can be applied to study such fundamental biological problems as protein folding.

Organisation of the database

The database consists of non-redundant raw data of protein structures, several tables and relationships between them (Figure 2). Essential components of the database are:

1. A coordinate set for each "unique" chain (see below).
2. A table of all unique chains with direct references to the table of domains and indirect references to the table of oligomers. For each chain there is the number of domains this chain contains and access instructions to the table of domains. Each chain also contains information whether this chain can be the part of an oligomer formation. If yes then it gives a signal to the search engine to search in the oligomer list also (see below). The chains have been organised in hierarchical manner according to their similarity using agglomerative clustering algorithm (Krzanowski and Marriott, 1995). Each chain also has a reference to the corresponding coordinate set.
3. A table of domains with the references to the table of chains and instructions how to form domains using the coordinates of the chains. The domains have also been organised hierarchically using pair-wise similarities based on the scores of the sequence alignment corrected by 3D comparison (see below).
4. A list of oligomers with the transformation operators to form them and references to the chain(s) that operations are applied. Since the table of chains contains only unique set (see below), some components of an oligomer may not be present in this table. To complete oligomer formation for these cases, there is a reference to the list of additional chains and their coordinates.
5. A list of additional chains and their coordinates to complete oligomer formation.
6. Database also contains a table of all PDB entries solved by MX, their cell dimensions and space groups. It is a useful table that allows quick search in the PDB if only basic information (cell parameters and space group) about MX experimental data is available.

Selection of entries

All protein entries from the PDB with a length greater than 15 amino acid residues that had been solved using MX and had been refined against data better than 3.5E resolution were selected to build the current database. Basic entries in the database were macromolecular subunits. If two subunits had a sequence identity greater than 80% and a root mean square deviation (rmsd) between corresponding Ca atoms of less than 1E, then the one that had been refined against higher resolution data was retained. This approach, while reducing substantially the number of entries kept in the database, retained the conformational variability of the molecules. For example, if there were two copies of a molecule and there was a domain motion during binding of a substrate (Figure 3), then even if the sequence identity was 100%, both representatives were kept in the database.

The sequence, information about the secondary structure, domain information (see below) and potential oligomer formation(s) for each entry were also stored. Therefore, when an entry is extracted, all necessary information is immediately available. In the early stages of the design of this database all chains (7000) were checked and domains were identified manually. These domains formed the bases for further updates.

Organisation of chains

All entries in the database are aligned against each other. The alignment is carried out in two steps: 1) Sequence alignment using a modified version of the Needleman and Wunsch (1970) dynamic alignment algorithm using BLOSUM62 (Henkoff and Henkoff, 1992) similarity matrix is carried out; 2) Alignment is corrected using domain information (see below). The result of the alignment is considered as a measure of similarity. Using this, a hierarchical database is organised with agglomerative clustering (refer). The results are kept as a search tree.

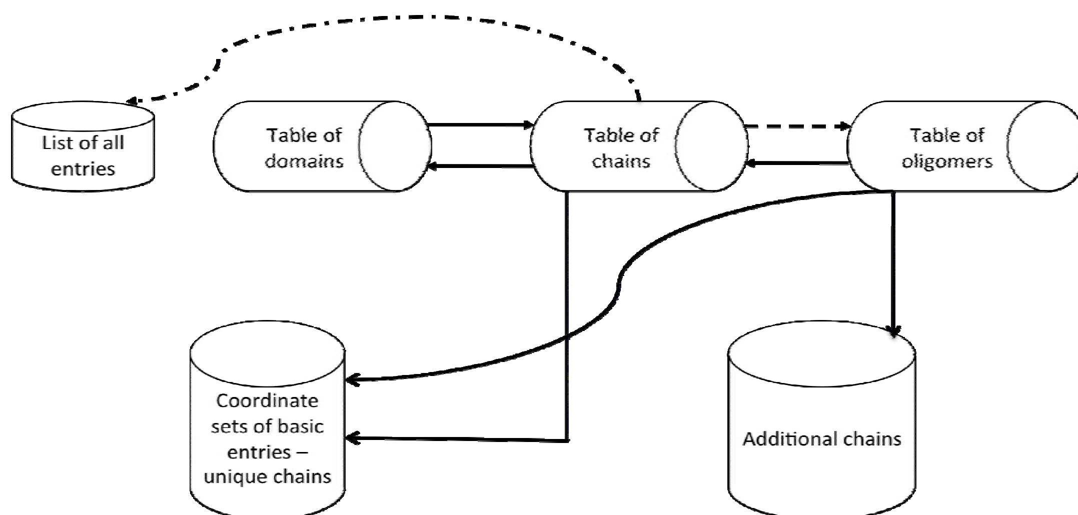

Figure 2. Overall organization of the database: All coordinates are stored in a set of files and they are accessed from the table of chains or oligomers. The table of chains has information about existence of domains, instructions to form pointers to the table of domains, information about existence of oligomer for current chain. The table of domains has information how to form domains using the chains. Oligomer list has instructions (pointers to the constituent chains, transformation operators to be applied to the chains in the PDB) how to form them. The list of additional chains contains those that are not in the list of "unique" chains but are essential for oligomer formation. Unique chains and domains also contain hierarchical organization. Solid arrows indicate that the table has instructions to access to another table or list; dashed arrows show that the table has information about existence of an element in another table without explicit instructions.

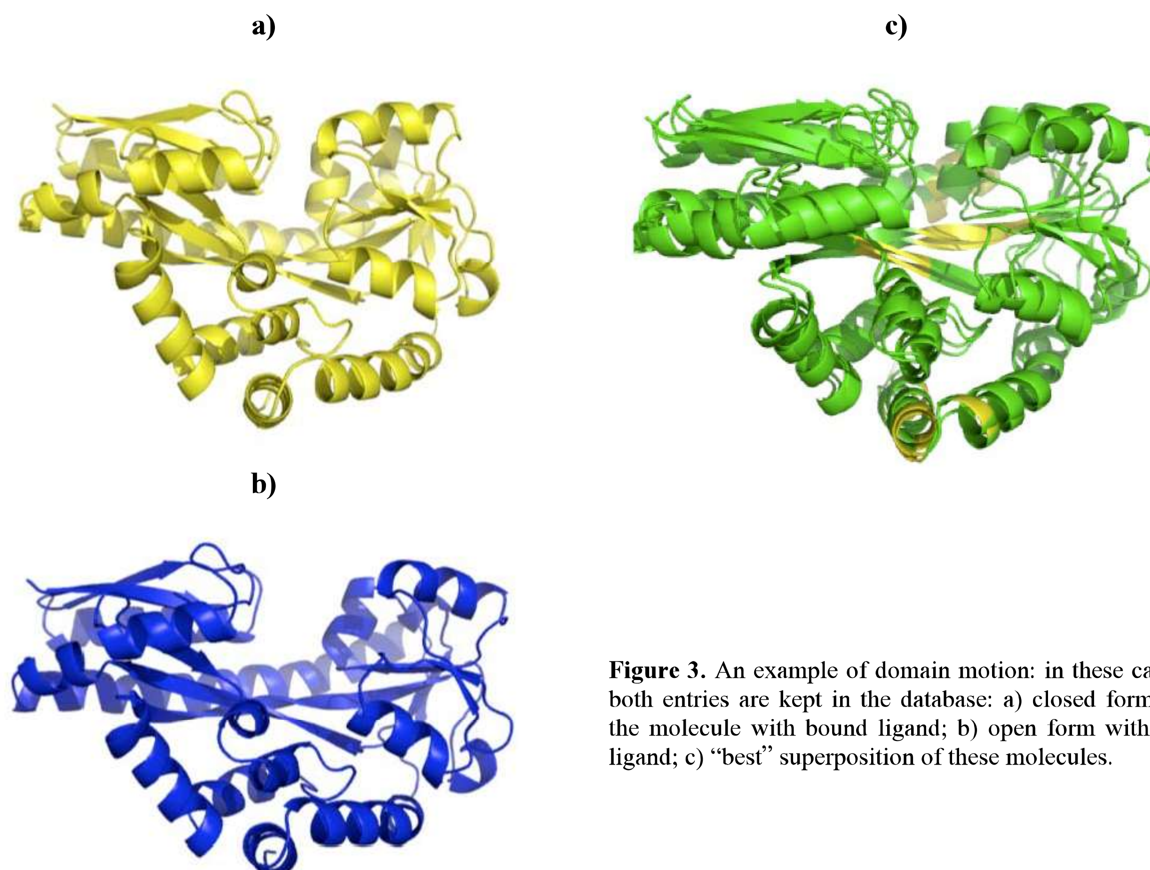

Figure 3. An example of domain motion: in these cases both entries are kept in the database: a) closed form of the molecule with bound ligand; b) open form with no ligand; c) "best" superposition of these molecules.

Domains

As it was stated above, in the beginning all domains were identified manually. The main criteria for domain definition were 3D compactness and separability from other parts of the subunit. However, if there was no well-defined domain in a molecule, then the whole molecule was considered as one domain. If a tentative domain contained completely exposed loops, and N or C terminal stretches, they were considered as flexible parts and were removed from the domain definition. The result of this analysis gave approximately 40 000 domains.

Organisation of domains

All residues in the domains that were aligned at the stage of analysis of chains were used for further refinement of domain alignment. The aligned residues were used to superimpose the domains using the 3D fitting algorithms (Kabsch, 1976). Quality factors (Q-factor) were calculated using the iterative procedure described by Krissinel and Henrick (2004). The Q-factors were used as similarity measures in the hierarchical agglomerative clustering of the domains. Once clusterisation of the domains was finished, 3D aligned stretches of the corresponding domains were used to check and correct the clustering of chains. This procedure ensured that subunits and domains belonging to the same class were similar in 3D and not merely in sequence. It should be noted that information about domains were kept in the database as a set of instructions, which were necessary to generate them from the basic entries - chains. It reduced the amount of information stored in the database. When domains are needed then special software using these instructions generates domains from the coordinates of the chains.

Oligomers

Oligomers for each entry are generated using the procedure described by Krissinel and Henrick (2005). Information about oligomer formation is stored as operations to generate them from the chains. This again substantially reduces the amount of information stored in the database. In the early stages of the development of this database, we obtained the oligomer information for each entry using the program PISA at EBI MSD server (Krissinel and Henrick 2005, www.ebi.ac.uk/pdbe). However, in the middle of 2009 EBI stopped updating the database component supporting PISA, we had to develop our own PISA based oligomer generation software locally at York. Now oligomers are generated every 15 days as the system is updated.

Update

Every 15 days the database is updated using newly deposited structures. If the sequence and the 3D structure of the newly deposited structures are similar to an entry in the existing database (criteria for selection were the same as described above) then this entry is rejected unless the new entry has been analysed using higher resolution data, in the latter case the new entry replaces the old one. All accepted entries are aligned with all entries in the database and with each other. This alignment again is corrected using information about domains. Newly generated similarity matrix is used to update hierarchical database of chains. At this stage, information about oligomer formation is also updated. The domain list is updated in the next stage. It is done in two steps: 1) the domain information from the database is transferred to the accepted chains. If there is no similar structure and/or domain information then this chain is checked and domains are identified manually. Even automatically defined domains are checked to ensure that no error is introduced to the definitions of domains, as they are used in the next update stages. Errors introduced to the database may propagate to the future and it may be difficult or even impossible to correct them in the later stages. 2) Sequence alignment of the domains is performed and the Q-factor is calculated (Krissinel and Henrick, 2004). The result of this alignment is used to update similarity matrix for the domains. Hierarchical database of domains is updated based on the derived similarity matrix.

2. Use of the database

As an example of applications, the database has been integrated into a molecular replacement pipeline software suite – BALBES (Long et al., 2008). A schematic workflow of the pipeline is shown in Figure 4 and its York hosted web interface is shown in Figure 5. It can be seen from the Figure 4, once the input data (experimental data and sequence information) are validated, BALBES searches the database to find the template models for MX calculations. Depending on whether the input sequence file contains one or multiple sequences, BALBES carries out one or all of the search algorithms described below.

Search using a single sequence

Search is started using domains. For a given sequence if a similar domain from the database by sequence alignment is found the information about it is stored, the corresponding portion of the sequ-

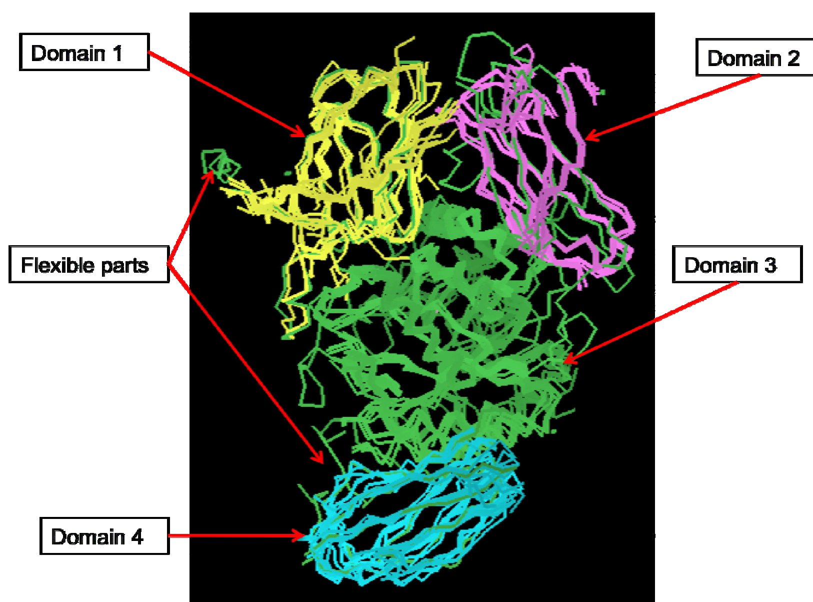

Figure 4. An example of ensembles for a molecule with four domains (pdb entry 3fna): for each domain there are eight representatives and all they have been superimposed with each other. This ensemble represents variations within domains. Note that flexible parts of the molecule have only one representative.

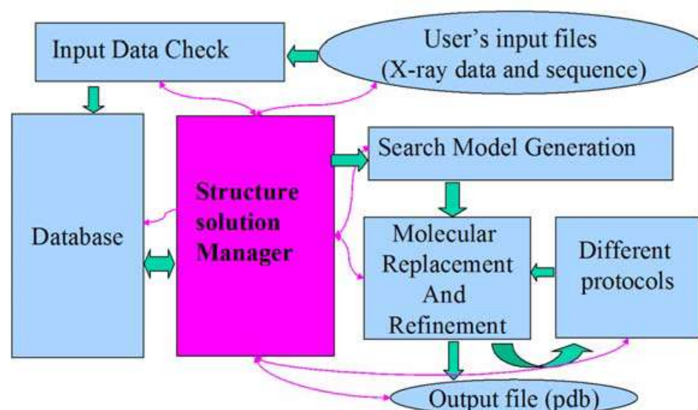

Figure 5. A schematic view of the BALBES workflow: All decisions are made internally according to the amount of the data (reflections and sequence) and the stage of structure solution. The pink arrows show that the manager controls all the activities involved and the block arrows show the directions of the workflow.

ence is removed and procedure is repeated. When no more domains can be found, the procedure is stopped. At the end of this search, domain composition for the new protein or its subunits is known. If the database contains a chain with similar domain composition then that chain is considered similar to the search sequence and the full chain is built. If no domains can be found then the chain database is searched. At the end of the search the system has access to the coordinates of the chain, domains as well as to potential oligomers. For most of the searches domain database gives reliable homologues. In very rare cases, when domain search does not produce homologues structures then the tables of chains is also searched.

a)

b)

Figure 6. A web interface for the pipeline using the database: A user needs to give only experimental data and click few basic options. The system makes all decisions at all stages automatically. At the end of the process the results are presented and they either can be viewed using jmol (jmol.sourceforge.net) or downloaded.

Search for assemblies

If an input file contains more than one sequence, then the system assumes that it is a complex of proteins. In this case it searches for assemblies containing all or subsets of these sequences. The search is done using the oligomer list. If they are found then they are used as template models for structure solution. If no such assembles are found, then each sequence is searched in turn using the single sequence algorithm described above and for each sequence a set of template models is generated.

Ensemble generation

Once one representative of a chain and/or domains is found, it is used to generate ensembles. Domain composition of this representative is extracted from the database during search for similar structures. For each domain all similar domains from hierarchical domain database is extracted. The number of such domains could be more than 1000. Among these domains *k* (default values is 8) most similar domains according to the similarity by *rmsd* are selected. All these domains are aligned in 3D using dynamic alignment. The elements of the similarity matrix for this alignment are distances between all pairs of residues formed one from each domain. Iterative improvement of the alignment is again performed (Krissinel and Henrick, 2004) and Q-factor is calculated. This procedure is applied to all domains. Using the aligned domains sequence profile is generated. This profile is used to improve alignment of the sequences with the found structure. Using the results of the alignment, further model improvements (for details see Lebedev et al., 2008) are performed to increase probability of solving the structure. An example of such ensembles is shown in Figure 6.

3. Conclusions and future perspectives

A database encapsulating the existing knowledge with fast access to variety of information is an important ingredient of automation and optimal information extraction from experimental data. Such database can also be used for an efficient analysis of wealth of information available in the data bank. We have designed such a database of protein structures using existing Protein Data Bank and as a proof of principle we have integrated it into a crystallographic macromolecular structure solution pipeline. Tests have shown that, with relatively simple protocols, MX pipeline based on this database can solve around 80% of all structures available in the PDB automatically. It is substantially higher than that is reported in the PDB (around 67%). Although this number looks impressive it should be considered with care, as all tests are done using structures from PDB and we know that these structures can be solved. Therefore, these results are necessarily upwards biased.

The system is currently under intensive development. For example, procedures described by Isupov (2008) and Lebedev et al. (2008) are currently being implemented. A future version of the system will also include decisions based on supervised and unsupervised machine learning algorithms (Haykin, 1999).

One of the advantages of an automatic pipeline is that information can easily be extracted during structure solution and used when it is necessary. For example, if a structure is solved using one of the known structures, then information about the used model can be utilised in refinement (Schroder et al., 2010; Murshudov et al., 1997) as a prior knowledge to increase reliability of the derived model.

The system can be downloaded from www.ysbl.york.ac.uk/~fei/balbes/download. The web server using the developed system based on the described database is on: www.ysbl.york.ac.uk/YSBLPrograms/index.jsp. It is also distributed to the user community via UK based software initiative - CCP4 at www.ccp4.ac.uk (CCP4, 1994).

ACKNOWLEDGEMENTS

We thank Andrey Lebedev for discussions and useful suggestions, Misha Isupov, Gleb Bourunkev and Victor Lamzin for testing and useful feedback. This work was supported by the Wellcome Trust (FL and GNM grant number: 064405/Z/01/A), BBSRC (AAV, grant number: 1 RO1 GM069758-03) and BIOXHIT (FL: grant number: LSHG-CT-2003-503420). The computers used for testing the system were acquired using funds from NIH (grant number: 1 RO1 GM069758-03) and Wellcome Trust grants.

REFERENCES

Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. (2000) The Protein Data Bank. Nucleic Acid Res. **28:** 235-242.

Delarue M. (2008) Dealing with structural variability in molecular replacement and crystallographic refinement through normal-mode analysis. Acta Cryst. **D64:** 40-48.

Emsley P., Cowtan K. (2004) Coot: model-building tools for molecular graphics. Acta Cryst. **D60:** 2126-2132.

Collaborative Computational Project: Number 4. (1994) The CCP4 suite: programs for protein crystallography. Acta Cryst. **D50:** 760-763.

Germain G., Main P., Woolfson M.M. (1970) On application of phase relationships to complex

structures. 2. Getting a Good Start. Acta Cryst. **B26:** 274-285.

Haykin S. (1999) Neural networks: a comprehensive foundation. Prentice Hall, Inc., Upper Saddle River, N.Y.

Henikoff S., Henikoff J.G. (1992) Amino acid substitution matrices from protein blocks. Proc. Nat. Acad. Sci. USA **89:** 10915-10919.

Isupov M., Lebedev A.A. (2008) NCS-constrained exhaustive search using oligomeric models. Acta Cryst. **D64:** 90-98.

Jones T.A., Zou J.Y., Cowan S.W., Kjeldgaard M. (1991) Improved methods for building protein models in electron-density maps and the location of errors in these models. Acta Cryst. **A47:** 110-119.

JMOL jmol.sourceforge.net

Keegan R., Winn M. (2008) MrBUMP: an automated pipeline for molecular replacement. Acta Cryst. **D64:** 119-124.

Kabcsh W. (1976) Solution for best rotation to repate 2 sets of vectors. Acta Cryst. **A32:** 922-923.

Krissinel E., Henrick K. (2005) Detection of protein assemblies in crystals. In: Lecture notes in computer science. Comp. Life Sci. **3695:** 163-174.

Krissinel E., Henrick K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Cryst. **D60:** 2256-2268.

Krzanowski W.J., Marriot F.H.C. (1994) Multivatiate analysis. Kendall's library of statistics. A Hodder Arnold Publications.

Lebedev A., Vagin A.A., Murshudov G.N. (2008) Model preparation in MOLREP and examples of model improvement using X-ray data. Acta Cryst. **D64:** 33-39.

Long F., Vagin A.A., Young P., Murshudov G.N. (2008) BALBES: a molecular replacement pipeline. Acta Cryst. **D64:** 125-132.

Murshudov G.N., Vagin A.A., Dodson E.J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Cryst. **D53:** 240-255.

Murzin A.G., Brenner S.E., Hubbard T., Chothia C. (1995) SCOP - A structural classification of proteins database for the investigation of sequence and structure. J. Mol. Biol. **147:** 536-540.

Needleman S., Wuncsh C. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48:** 443-453.

Ness S.R., de Graff R.A.G., Abrahams J.P., Pannu N.S. (2004) CRANK: New methods for automated macromolecular crystal structure solution. Structure **12:** 1753-1761.

Panjikar S., Parthasarathy V., Lamzin V.S., Weiss M.S., Tucker P.A. (2005) Auto-Rickshaw: an automated crystal structure determination platform as an efficient tool for the validation of an X-ray diffraction experiment. Acta Cryst. **D61:** 449-457.

Pearl F., Todd A., Sillitoe I., Dibley M., Redfern O., Lews T., Bernett C., Marsden R., Grant A., Lee D., Akpor A., Maibaum M., Harrison A., Dallman T., Reeves G., Diboun I., Addou S., Lise S., Johnston C., Sillero A., Thornton J., Orengo C. (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acid Res. **33:** 247-251.

Perrakis A., Morris R., Lamzin V.S. (1999) Automated protein model building combined with iterative structure refinement. Nat. Struc. Biol. **6:** 458-463.

Schwarzenbacher R., Godzik A., Jaroszewski L. (2008) The JCSG MR pipeline: optimized alignments, multiple models and parallel searches. Acta Cryst. **D64:** 133-140.

Schwede T., Kopp J., Guex N., Peitsch M.C. (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res. **31:** 3381-3385.

Schroder G.F., Brunger A.T., Levitt M. (2010) Super-resolution biomolecular crystallography with low-resolution data. Nature **464:** 1218-1222.

Terwilliger T.C., Berendzen J. (1999) Structural genomics and automated X-ray crystal structure solution. Acta Cryst. **D55:** 849-861.