Nucleic Acids Research, 2006, Vol. 34, Database issue ***D287–D290***
doi:10.1093/nar/gkj163

E-MSD: improving data deposition and structure quality

M. Tagari, J. Tate, G. J. Swaminathan, R. Newman, A. Naim, W. Vranken, A. Kapopoulou, A. Hussain, J. Fillon, K. Henrick and S. Velankar*

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

Received September 15, 2005; Revised and Accepted November 1, 2005

ABSTRACT

The Macromolecular Structure Database (MSD) (http://www.ebi.ac.uk/msd/) [H. Boutselakis, D. Dimitropoulos, J. Fillon, A. Golovin, K. Henrick, A. Hussain, J. Ionides, M. John, P. A. Keller, E. Krissinel *et al.* (2003) E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Res.*, 31, 458–462.] group is one of the three partners in the worldwide Protein DataBank (wwPDB), the consortium entrusted with the collation, maintenance and distribution of the global repository of macromolecular structure data [H. Berman, K. Henrick and H. Nakamura (2003) Announcing the worldwide Protein Data Bank. *Nature Struct. Biol.*, 10, 980.]. Since its inception, the MSD group has worked with partners around the world to improve the quality of PDB data, through a clean up programme that addresses inconsistencies and inaccuracies in the legacy archive. The improvements in data quality in the legacy archive have been achieved largely through the creation of a unified data archive, in the form of a relational database that stores all of the data in the wwPDB. The three partners are working towards improving the tools and methods for the deposition of new data by the community at large. The implementation of the MSD database, together with the parallel development of improved tools and methodologies for data harvesting, validation and archival, has lead to significant improvements in the quality of data that enters the archive. Through this and related projects in the NMR and EM realms the MSD continues to improve the quality of publicly available structural data.

INTRODUCTION

Streamlining of protein cloning, expression, purification through to crystallization and structure determination has resulted in rapid increase in the rate at which new structures are determined. It is becoming ever more important that the associated deposition tools keep pace with this exponential growth of data. The process of structure deposition is being adapted to cope with these rapid changes and remains an important step in the structure determination and publication pipeline. We are addressing the real problems that are involved with quickly and accurately annotating and archiving new structures to avoid the risk of disillusionment and disaffection amongst the structural biologists whose work populates data resources such as the Macromolecular Structure Database (MSD). To ensure that the core data is represented uniformly at all the worldwide Protein DataBank (wwPDB) partner sites, the three sites have active collaboration involving exchange of the core reference information (e.g. the dictionary description for ligands) and regular discussions to apply uniform standards for deposition and annotation of the core data.

The MSD is approaching these problems in several ways. As a partner in the CCP4 project (1) we have contributed to the definition of standards for data harvesting and are working with CCP4 software developers to integrate the MSD deposition tool AutoDep into the CCP4 structure determination flow. The CCP4 framework performs data harvesting, extracting and recording relevant statistics from the component programs, reducing the occurrence of errors and inconsistencies being introduced by manual record keeping. MSD has implemented an interface in AutoDep that accepts uploads of harvest data files and uses them to pre-fill forms that would otherwise have to be completed manually by the depositor. We are also using AutoDep as a delivery platform for extra data and annotations that are generated during curation, making the deposition process richer and more useful to the depositor. Examples of the generated data that are returned to the depositor include quaternary structure assessments, structure

*To whom correspondence should be addressed. Tel: +44 1223 494 646; Fax: +44 1223 494 468; Email: sameer@ebi.ac.uk

quality metrics using electron density calculations (where structure factors are provided for X-ray entries) and detailed heterogen dictionaries.

The UK-based eHTPX initiative (2) aims to unify the procedures of protein structure determination into a single interface from which users can initiate, plan, direct and document their experiment either locally or remotely. As a partner in this initiative, the MSD is co-ordinating the development of common data models that cover all stages from protein production through to structure solution and data deposition (http://www.ebi.ac.uk/msd-srv/docs/ehtpx/lims/) (3). The above data model is now being implemented as a fully functional Laboratory Information Management System (LIMS) designated PIMS (http://www.pims-lims.org/). This forms part of a UK BBSRC-funded initiative that is working with EU projects and is based on the earlier HALX (4) and MOLE (5) applications. The BIOXHIT (http://icarus.embl-hamburg.de/bioxhit/) initiative forms a complimentary effort involving the MSD and CCP4 to define meta-data dictionaries for data exchange that will be implemented as XML schema to automate structure solution software pipelines. These initiatives will develop a framework to facilitate the tracking of data from target selection through to deposition of the coordinates in the wwPDB.

Similar mechanisms for data harvesting and deposition have been developed by the CCPN project (6), in which MSD is a consultative partner. The MSD group is also working with the EM community to develop a database for EM volume data and a deposition system for this data. The EM work is progressing with our involvement in the '3D-EM' Network of Excellence funded within Research Framework Programme 6 of the European Commission (http://www.3dem-noe.org/). These close ties between the MSD and these two communities mean that we are well placed to bring together the three parallel efforts in structural biology, leveraging expertise from all three realms in the development of a new, unified data deposition architecture. Such a framework will allow us to perform rigorous checks and validations of deposited data, thereby improving the quality of the final archives and ensuring accuracy for users of the data.

AutoDep

AutoDep version 4 (http://www.ebi.ac.uk/msd-srv/autodep4/) is a complete rewrite of the original AutoDep system, which was inherited from the first deposition system, written at BNL during the early days of the PDB. The new application maintains a superficial similarity to the old AutoDep tool, in order to ensure an easy transition for users from the legacy system, but is a completely new implementation, using a Java servlet framework and XML as a data storage and interchange format. The system is highly customizable, owing to the system of flexible XML-based dictionaries that define all aspects of the user interface and internal data formats.

The new system includes various features that improve the quality of the final data which are deposited into the PDB. The interface definitions include validation patterns that are automatically applied to user-supplied data as they are entered, allowing a first round of data validation to be applied even as data are entered into the system. Once deposition is complete, but before submission, a suite of analysis programs are run, providing a further layer of validation for structures. Major problems in the uploaded coordinates are often identified at this stage of the deposition process, allowing the depositor to correct their data themselves, usually without intervention from AutoDep maintainers.

Once the deposition is complete the entry is curated by the EBI staff. Allowing more rigorous and more detailed validations to be performed on the deposited coordinates, the curation process also allows us to return additional, value-added data to the depositor. Our hope is to encourage depositors to submit PDB entries before completion of the accompanying journal article, so that the deposition process becomes an additional analysis step, data from which can be included in the structure description. An example of the extra data that are derived, rather than being added by the depositor, is secondary structure information: the PDB uses standard programs to generate secondary structure records, ensuring consistent and uniform definitions are used throughout the archive. We also run a complex structural analysis that suggests the putative quaternary structure of the molecule in the crystal, the so-called PQS assembly (7). For X-ray entries where experimental data are deposited along with coordinates, we run an in-house version of the Uppsala Electron Density Server (8) and the CCP4 program SFCHECK (9). These allow us to verify that the supplied structure factors are those used to determine the structure as well as providing an additional measure of the quality of the structure in the form of the real-space *R*-factor and correlation coefficients. Further checks on the structure are devoted to verifying the sequence of the PDB entry, through consideration of a residue-by-residue mapping to the UniProt sequence database (10), similar in nature to that available through the Structure Integration with Function, Taxonomy and Sequence (SIFTS) initiative (11). The sequence mapping allows us to validate several values in the PDB entry, such as the source organism for example, as well as the sequence itself. Depositors are notified of differences between the sequence of their structure and the corresponding UniProt entry, allowing them to correct sequence problems before the entry is released. Future developments in AutoDep will include data from other MSD services, such as MSDfold (12), which identifies released PDB entries which exhibit a similar fold. An MSDfold query would permit us to show SCOP and CATH domain information for the new structure, further enhancing the range of information that is returned to the depositor during deposition.

ELECTRON MICROSCOPY DATABASE

The Electron Microscopy Database (EMDB) contains 3D volumes of biological macromolecules, determined by cryo-electron microscopy (13,14). Volumes are deposited by the cryo-EM community using the EMDep tool (http://www.ebi.ac.uk/msd-srv/emdep/), a system equivalent in scope to AutoDep. Since EMDep was launched in June 2002 we have received depositions of EM data from a wide variety of experiments, including tomography, icosahedral viruses, ribosomes and chaperonins.

EMDep is a flexible system for users to input complex macromolecular volume information into a databank in an interactive manner. The system leverages the knowledge

and expertise of the experimenters in order to obtain a complete and accurate description of the structural experiment. For example, the author can upload an XML file containing a Fourier Shell Correlation graph, used in determination of the resolution of the uploaded volume. In order to ensure an accurate representation of the final volume on the publicly available summary pages for a given entry, depositors are encouraged to supply images to accompany their entries. Curation of the deposition involves several checking steps, such as verification of the integrity of the uploaded map and its format (CCP4) by visualizing it using a display program such as CHIMERA (http://www.cgl.ucsf.edu/chimera/). The accuracy of information relating to the experimental details given in the 'header' XML file is checked by suitable cross-referencing to reference data, such as hardware manufacturers, taxonomy, processing program names, etc. Finally, external cross-references to related databases are checked, including GO (15) (http://www.ebi.ac.uk/ego/), INTERPRO (16) (http://www.ebi.ac.uk/interpro/), ICTVdb VIRUS database (http://www.ncbi.nlm.nih.gov/ICTVdb) and PUBMED (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi). As in the AutoDep tool, the results of curation are presented to the depositor through the EMDep web interface, enabling the depositor to view the finally annotated deposition in the same form as the original deposition.

Once the author of the deposition has granted approval for any changes made during curation then the 'header' information for each entry is made available. Release options for the volume, structure factors or layer-line data are 'hold for publication' or for periods of 1, 2 or 4 years. The current status of a particular deposition can be found on the EMDB area of the EBI FTP site (ftp://ftp.ebi.ac.uk/pub/databases/emdb/) or through the EMDB search tool, EMSearch (http://www.ebi.ac.uk/msd-srv/emsearch/). Summaries of data from EMDB depositions are presented in the form of atlas pages, which can also be accessed via the EMSearch tool.

The rapid pace of development in the field of cryo-EM suggests that comprehensive descriptions of the structures of whole cells and organelles in terms of the spatial arrangements of their molecular components may soon become routine. In this event the storage of data from these techniques will be of growing importance in terms of allowing public access to the resulting 3D reconstructions and the experimental details of the investigation (17) and EMDep will be an important central repository for these data. It is encouraging that many scientific journals now require an EMD ID before they publish the manuscript.

NMR

The MSD group is involved in the Collaborative Computing Project for the NMR community (CCPN) (18), which aims to provide a standard model, related libraries (19) and applications (6), for storage and handling of NMR data (http://www.ebi.ac.uk/msd-srv/docs/NMR/main.html). Essentially, an NMR project stored within the CCPN framework has clearly defined data elements that are consistent with each other. The MSD group has converted its internal library of ligands, small molecules and monomers to the CCPN framework, and this information is now the prime reference data for creating molecules when using CCPN. During the conversion process, the data were rearranged to handle different protonation variants of the same chemical group, and extra information was added (e.g. atom sets that are relevant for NMR, and protons that are likely to exchange with water, commonly used atom and residue naming systems). The information on atom and residue naming systems that are commonly used in NMR were gathered and incorporated, mainly from the AQUA (20) library files and the BioMagResBank (http://www.bmrb.wisc.edu/). We have further extended and cleaned up this naming system data, and although it is an integral part of the CCPN compound library (http://www.ebi.ac.uk/msd-srv/docs/NMR/chemCompXml/main.html) we have now made separate reference files available in tabulated and XML formats (http://www.ebi.ac.uk/msd-srv/docs/NMR/refData/main.html). These files are free to download and use, although we request that any changes or mistakes are reported so they can be fed back to the community. This library of atom naming systems was already extensively and successfully used within the CCPN framework during the RECOORD project (21). In this project, collaboration with the BioMagResBank, deposited NMR constraint files are made consistent with the deposited coordinate atoms (22). The structures for these molecules are then recalculated with the same protocol, allowing a better comparison between them. Overall, the involvement of the MSD in the CCPN project means that deposited CCPN files can be directly mined for consistent information on the molecular system. Based on this effort we will also develop an NMR component of the E-MSD database to store NMR data and relate it directly to the relevant PDB entries.

OTHER MSD DEVELOPMENTS

The SIFTS initiative (11) is intended to bring together various disparate bioinformatics resources, from the structural, sequence and related realms. The SIFTS data have been updated recently to include version information from UniProt, to ensure that the various residue-level mappings are properly assigned and tracked through different updates of the constituent data resources.

The MSD group is currently rolling out several new services or resources. MSDtemplate is a search system for a set of data on the local interactions of residues. MSDpisa is a new MSD service that is similar in scope to PQS (7). It aims to apply unit cell symmetry (for X-ray determined structures) to obtain possible quaternary structure. The service provides information on putative assemblies for PDB entries, as well as detailed data on the interactions at inter-molecule interfaces. The MSDmotif service provides information on known 3D hydrogen-bonded motifs.

The MSD deposition systems and related tools continue to ensure that deposited data are consistent and of high quality, thereby improving the quality of data in a range of inter-related publicly accessible archives.

ACKNOWLEDGEMENTS

E-MSD gratefully acknowledges the support of the Wellcome Trust (GR062025MA), the EU (TEMBLOR, NMRQUAL and 3D-EM NoE), CCP4, the BBSRC, the MRC and EMBL.

Funding to pay the Open Access publication charges for this article was provided by The Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Collaborative Computational Project Number 4 (1994) The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.*, **50**, 760–763.
2. Allan,R., Diakun,G., Guest,M., Keegan,R., Nave,C., Papiz,M., Winn,M., Winter,G., Diprose,J., Esnouf,R. *et al.* (2003) Science resource for high throughput protein crystallography. *Proceedings of the UK e-science All Hands Meeting*. EPSRC, September 2–4, Nottingham, UK , pp. 230–234.
3. Pajon,A., Ionides,J., Diprose,J., Fillon,J., Fogh,R., Ashton,A.W., Berman,H., Boucher,W., Cygler,M., Deleury,E. *et al.* (2005) Design of a data model for developing laboratory information management and analysis systems for protein production. *Proteins*, **58**, 278–284.
4. Prilusky,J., Oueillet,E., Ulryck,N., Pajon,A., Bernauer,J., Krimm,I., Quevillon-Cheruel,S., Leulliot,N., Graille,M., Liger,D. *et al.* (2005) HalX: an open-source LIMS (Laboratory Information Management System) for small- to large-scale laboratories. *Acta Crystallogr. D Biol. Crystallogr.*, **61**, 671–678.
5. Morris,C., Wood,P., Griffiths,S.L., Wilson,K.S. and Ashton,A.W. (2005) MOLE: a data management application based on a protein production data model. *Proteins*, **58**, 285–289.
6. Vranken,W.F., Boucher,W., Stevens,T.J., Fogh,R.H., Pajon,A., Llinas,M., Ulrich,E.L., Markley,J.L., Ionides,J. and Laue,E.D. (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins*, **59**, 687–696.
7. Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.
8. Kleywegt,G.J., Harris,M.R., Zou,J.Y., Taylor,T.C., Wahlby,A. and Jones,T.A. (2004) The Uppsala Electron-Density Server. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 2240–2249.
9. Vaguine,A.A., Richelle,J. and Wodak,S.J. (1999) SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr. D Biol. Crystallogr.*, **55**, 191–205.
10. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
11. Velankar,S., McNeil,P., Mittard-Runte,V., Suarez,A., Barrell,D., Apweiler,R. and Henrick,K. (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **33**, D262–D265.
12. Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 2256–2268.
13. Tagari,M., Newman,R., Chagoyen,M., Carazo,J.M. and Henrick,K. (2002) New electron microscopy database and deposition system. *Trends Biochem. Sci.*, **27**, 589.
14. Henrick,K., Newman,R., Tagari,M. and Chagoyen,M. (2003) EMDep: a web-based system for the deposition and validation of high-resolution electron microscopy macromolecular structural information. *J. Struct. Biol.*, **144**, 228–237.
15. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.
16. 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.
17. Fuller,S.D. (2003) Depositing electron microscopy maps. *Structure (Camb.)*, **11**, 11–12.
18. Fogh,R., Ionides,J., Ulrich,E., Boucher,W., Vranken,W., Linge,J.P., Habeck,M., Rieping,W., Bhat,T.N., Westbrook,J. *et al.* (2002) The CCPN project: an interim report on a data model for the NMR community. *Nature Struct. Biol.*, **9**, 416–418.
19. Fogh,R.H., Boucher,W., Vranken,W.F., Pajon,A., Stevens,T.J., Bhat,T.N., Westbrook,J., Ionides,J.M. and Laue,E.D. (2005) A framework for scientific data modeling and automated software development. *Bioinformatics*, **21**, 1678–1684.
20. Doreleijers,J.F., Rullmann,J.A. and Kaptein,R. (1998) Quality assessment of NMR structures: a statistical survey. *J. Mol. Biol.*, **281**, 149–164.
21. Nederveen,A.J., Doreleijers,J.F., Vranken,W., Miller,Z., Spronk,C.A., Nabuurs,S.B., Guntert,P., Livny,M., Markley,J.L., Nilges,M. *et al.* (2005) RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins*, **59**, 662–672.
22. Doreleijers,J.F., Nederveen,A.J., Vranken,W., Lin,J., Bonvin,A.M., Kaptein,R., Markley,J.L. and Ulrich,E.L. (2005) BioMagResBank databases DOCR and FRED containing converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures. *J. Biomol. NMR*, **32**, 1–12.