D262–D265 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki058

E-MSD: an integrated data resource for bioinformatics

S. Velankar, P. McNeil, V. Mittard-Runte[1], A. Suarez, D. Barrell[1], R. Apweiler[1] and K. Henrick*

Macromolecular Structure Database Group (E-MSD) and [1]Sequence Database Group, EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 14, 2004; Revised and Accepted October 4, 2004

ABSTRACT

The Macromolecular Structure Database (MSD) group (http://www.ebi.ac.uk/msd/) continues to enhance the quality and consistency of macromolecular structure data in the worldwide Protein Data Bank (wwPDB) and to work towards the integration of various bioinformatics data resources. One of the major obstacles to the improved integration of structural databases such as MSD and sequence databases like UniProt is the absence of up to date and well-maintained mapping between corresponding entries. We have worked closely with the UniProt group at the EBI to clean up the taxonomy and sequence cross-reference information in the MSD and UniProt databases. This information is vital for the reliable integration of the sequence family databases such as Pfam and Interpro with the structure-oriented databases of SCOP and CATH. This information has been made available to the eFamily group (http://www.efamily.org.uk/) and now forms the basis of the regular interchange of information between the member databases (MSD, UniProt, Pfam, Interpro, SCOP and CATH). This exchange of annotation information has enriched the structural information in the MSD database with annotation from wider sequence-oriented resources. This work was carried out under the 'Structure Integration with Function, Taxonomy and Sequences (SIFTS)' initiative (http://www.ebi.ac.uk/msd-srv/docs/sifts) in the MSD group.

INTRODUCTION

The past few years have seen an explosion in the volume of bioinformatics data that is available to researchers. As the rate of discovery continues apace, it is becoming ever more difficult to make sense of these data. Although they may be categorized as sequence- or structure-oriented, the implications of a particular dataset often span the divide between the two realms, yet existing tools and techniques rarely achieve the same.

In order to exploit the information that is already available, and to cope with the ever-increasing volume of new data that is now being generated, it is essential that we develop a robust and maintainable mechanism for integrating data resources from different domains. It is important to note that most of the data resources devoted to derived data and annotation are linked back to the primary data resources on which they depend for raw data and our approach to the problem has therefore been to concentrate on forming tight links between primary resources. Three such primary resources are the EMBL nucleotide sequence database (1), the UniProt protein sequence database (2) and the single worldwide repository of macromolecular structures—the worldwide Protein Data Bank (wwPDB) (3). The Macromolecular Structure Database (MSD) is one of the three sites that together constitute the wwPDB, and therefore, we are ideally placed to work with our EBI colleagues in UniProt and EMBL, to maintain low-level linkages between these three primary data resources. Such close collaboration is of immediate benefit not only to these three separate projects, but also to the numerous other projects that use data from these sources.

One of the major achievements of the collaboration between MSD and UniProt has been the introduction of robust mechanisms for the exchange of data between these two databases. This has dramatically improved the quality of annotation in both databases and is aiding the continuing improvements in legacy data. In the longer term, this project will allow not only for better and closer integration of derived-data resources but will continue to improve the quality of all data in the primary resources. As we expand our collaborations to work more closely with the nucleotide data providers, such as EMBL nucleotide sequence database, we will be able to bring the same benefits to another broad section of the bioinformatics community.

METHODOLOGY

We have used sequence identity and taxonomy as the characteristics on which to link protein sequence data (from UniProt) and protein structure data (from MSD).

*To whom correspondence should be addressed. Tel: +44 1223 494629; Fax: +44 1223 494468; Email: henrick@ebi.ac.uk

Since the sequences of a structure in the MSD may represent either the native protein sequence or that of an engineered mutant or other variant, during the automatic procedure, the criterion for assessing sequence identity was that there should be 95% or higher agreement between the sequence of a protein structure and the corresponding sequence in UniProt. This was relaxed further down to 90% during the manual annotation.

In many cases, taxonomy information is entirely missing from the PDB entries or, where taxonomy information is supplied, it is given as the full scientific name of the source organism. This is inevitably prone to spelling or typographical errors but, more crucially, this does not provide the full and exact taxonomic classification for the organism. Furthermore, because protein structure is more conserved across evolutionary time than is protein sequence (4) and the structural differences between proteins with high-sequence identity are small, the rule for assessing taxonomy assignments can be even more relaxed.

Hence, the rules that determine the correct cross-reference between an MSD entry and a corresponding UniProt entry are (i) high-sequence identity (ideally 100% but not below 90%), and (ii) the taxonomy ID for the two entries, MSD and UniProt, must be the same or must have a common parent within one or two levels up the taxonomic tree, at the species level or below.

This approach required that we adopt the NCBI taxonomic identifiers (5) (http://www.ebi.ac.uk/newt/) as a standard way of representing the taxonomy information for all of the PDB entries within the MSD database. In the ideal case, every PDB entry should have a record of the organism from which each component of this particular structure derives, but in the legacy archive, the situation is far from ideal: many entries simply have no such record, while those records that are present have historically been prone to typographical or spelling errors. For entries with no taxonomy information, manual searches of the PDB file or accompanying literature were performed and for all entries we have put in place mechanisms that automatically check the user-supplied taxonomy information against the NCBI database, using the standard NCBI taxonomy identifier that we assign to each PDB entry. This allows us to correct spelling mistakes in legacy PDB files and to identify PDB entries where the taxonomy information is simply incorrect. Furthermore, by using a stable, curated taxonomy identifier throughout the database, we gain access to the wealth of annotation information in the NCBI database, such as synonyms and hierarchical relationships between different taxonomic nodes.

Simultaneously, we have also cleaned up the UniProt cross-references for every entry in the PDB and, in collaboration with the UniProt group, have put in place mechanisms to keep the cross-references up to date. In the cases where no cross-reference was available from the PDB archive, a semi-automatic process was used to correctly identify them. In cases where the PDB entry contained a chimeric protein (engineered proteins where different segments of a single polypeptide are derived from different proteins or different organisms), it was also important to identify the correct boundaries for the unique segments in each chain of the PDB entry. Once the correct taxonomic and cross-reference information had been obtained, these two sets of data were cross-checked, allowing us to identify entries with subtle problems that required manual intervention to correct them. Finally, after completing the clean up of archive, it was possible to map accurately the sequences from the PDB entries on to corresponding UniProt entries.

The main difficulty in determining this mapping is that many structures in the PDB have regions of unobserved residues in the middle of continuous polypeptide chains. This discontinuity in the sequence of the structure arises because it is often impossible to reliably construct a model for poorly defined regions of structure, such as flexible loops. Such gaps in the sequence are not taken into account by traditional sequence alignment algorithms, leading to incorrect alignments for regions flanking the unobserved regions.

To circumvent this problem, we modified the standard alignment protocol and developed software to use sequences of connected segments of a polypeptide chain from the PDB entry, corresponding to the observed regions of a protein structure. The separate alignments for these segments were then merged together to assemble the complete alignment between the sequence of the observed residues from the PDB entry and the complete sequence of the protein that was used in the experiment. This latter sequence is shown in the ‘SEQRES’ record in the PDB entry and does not have gaps reflecting unobserved residues. A similar procedure was carried out to obtain alignments between the sequences of observed residues and the corresponding UniProt entry. These two composite alignments were then merged to give the complete residue-level mapping between the sequence of the complete polypeptide from the experiment and its UniProt counterpart. This complex procedure also allows us to extract annotations from the PDB and UniProt entries to explain any differences that were detected between the two sequences, such as variants, isoforms, modified residues or engineered mutations. Unobserved residues and N- or C-terminal tags for the polypeptide chains in the PDB entry are also annotated. Regions from the UniProt entry that do not form part of the polypeptide under study and not included in the PDB entry are clearly marked. The program also copes with the more complex situation in chimeric structures, where sequences from two or more UniProt entries are involved.

The database schema supporting the residue-level mapping is shown in Figure 1 and the current status of the mapping procedure is shown in Table 1.

DATA DISTRIBUTION/FUTURE DEVELOPMENTS

The mapping data are available in the XML format from ftp://ftp.ebi.ac.uk/pub/databases/msd/sifts/xml/. The XML schema was developed under the auspices of the eFamily project, which is working to facilitate the distribution of domain-specific sequence data and improve the integration of sequence and structure data resources. The mapping data form the backbone of the eFamily project. The same collaboration has resulted in the development of a Perl interface to the data, which will be made available under the Bio-Perl project. We also plan to develop web-services to be integrated with other web-services that will be developed by the partners in the eFamily project, namely SCOP (6), CATH (7), Pfam (8) and Interpro (9). These web-services will, in future, allow clients to develop workflows that will assist in the integration

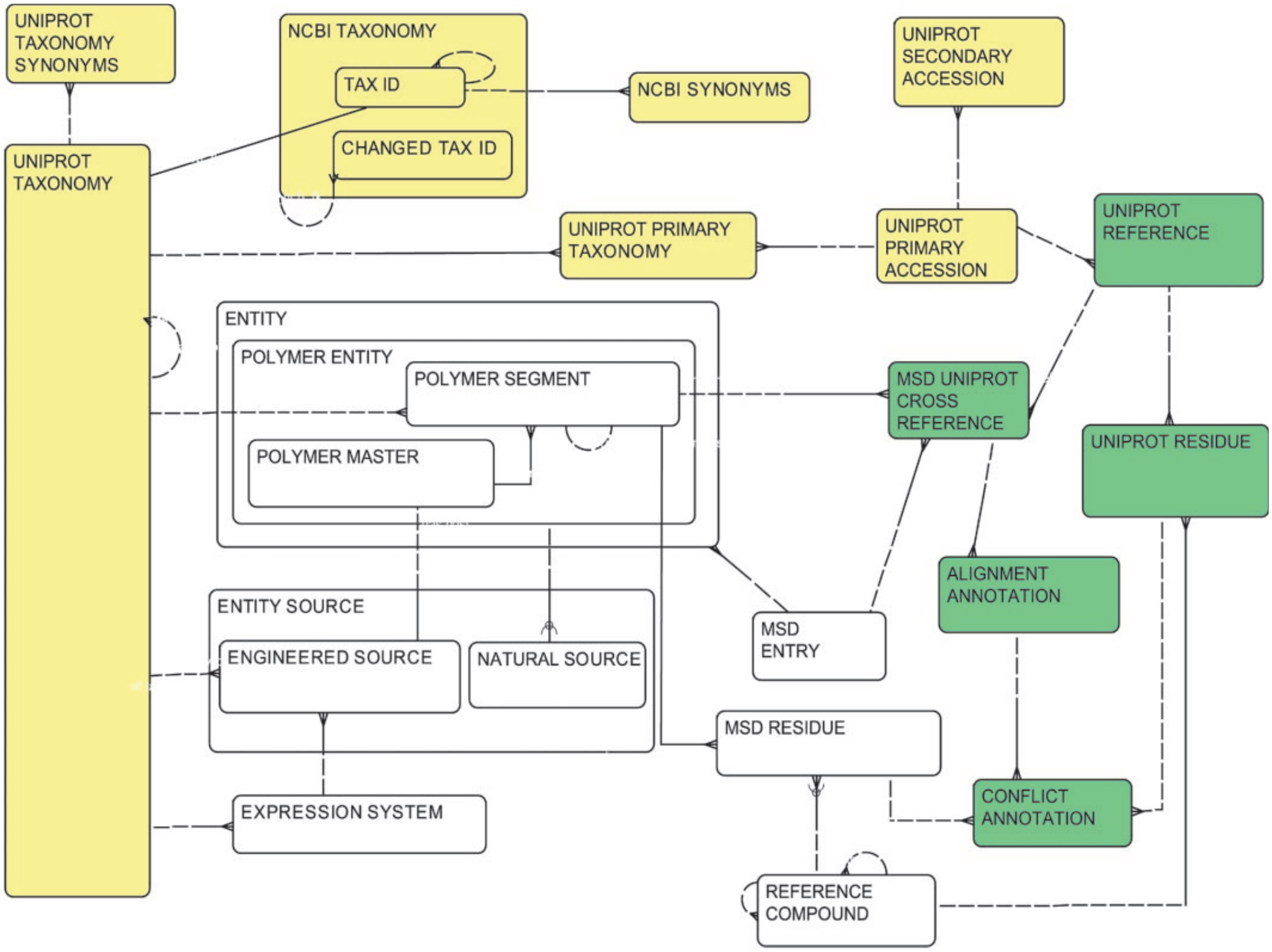

Figure 1. The database schema supporting the MSD to UniProt residue-level mapping. MSD components are in white, external database components in yellow and the cross-reference components in green.

Table 1. Current status of the MSD to Uniprot residue-level mappings

Total MSD entries	27 259
Entries with no possible Uniprot cross-reference	2 196
Entries with UniProt cross-reference	24 665 (98%)
Entries with residue-level mapping	24 218 (97%)
Entries awaiting mapping	845

of different bioinformatics resources based on the residue level mapping and annotation provided by the MSD.

OTHER MSD DEVELOPMENTS

Based on the UniProt cross-reference information, we have been able to drive forward the integration of structure information with not only the members of the eFamily group but also with other important biological resources such as GOA (10) and IntEnz (11). In the near future, we plan to enhance the structure information by integrating information from databases such as IntAct (12), ASD (13), KEGG (14) and MEROPS (15). These data have also benefited other bioinformatics groups who have built successful services based on UniProt cross-reference information (16). Other developments in the MSD group include the release of a completely new deposition system for the PDB data, which replaces the original AutoDep submission system. While forming the primary deposition service at the MSD, AutoDep can also be downloaded and used in-house by structural biology groups, providing a local archival and validation system. Furthermore, structures that have been deposited in a local AutoDep installation can be trivially uploaded into the MSD AutoDep system to form a complete PDB submission. The AutoDep system will become a part of the CCP4 (17) distribution.

The MSD search systems and the underlying relational database continue to improve, with new features and capabilities being added to many services, moving us ever closer to our ultimate goal of becoming a comprehensive, integrated resource for the research community.

ACKNOWLEDGEMENTS

Special thanks go to Dr John Tate for help during the preparation of this manuscript. E-MSD gratefully acknowledges the support of the Wellcome Trust (GR062025MA), the EU (TEMBLOR), the MRC and EMBL.

REFERENCES

1. Kulikova,T., Aldebert,P., Althorpe,N., Baker,W., Bates,K., Browne,P., van den Broek,A., Cochrane,G., Duggan,K., Eberhardt,R. *et al*. (2004) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res*., **32**, D27–D30.
2. 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.
3. Berman,H., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nature Struct. Biol*., **10**, 980.
4. Coulson,A.F. and Moult,J. (2002) A unifold, mesofold, and superfold model of protein fold use. *Proteins*, **46**, 61–71.
5. Wheeler,D.L., Church,D.M., Edgar,R., Federhen,S., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E. *et al*. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res*., **32**, D35–D40.
6. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*., **32**, D226–D229.
7. Pearl,F.M., Lee,D., Bray,J.E., Sillitoe,I., Todd,A.E., Harrison,A.P., Thornton,J.M. and Orengo,C.A. (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res*., **28**, 277–282.
8. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res*., **32**, D138–D141.
9. 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.
10. 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.
11. Fleischmann,A., Darsow,M., Degtyarenko,K., Fleischmann,W., Boyce,S., Axelsen,K.B., Bairoch,A., Schomburg,D., Tipton,K.F. and Apweiler,R. (2004) IntEnz, the integrated relational enzyme database. *Nucleic Acids Res*., **32**, D434–D437.
12. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P., Valencia,A. *et al*. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res*., **32**, D452–D455.
13. Thanaraj,T.A., Stamm,S., Clark,F., Riethoven,J.J., Le Texier,V. and Muilu,J. (2004) ASD: the Alternative Splicing Database. *Nucleic Acids Res*., **32**, D64–D69.
14. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
15. Rawlings,N.D., Tolle,D.P. and Barrett,A.J. (2004) MEROPS: the peptidase database. *Nucleic Acids Res*., **32**, D160–D164.
16. Martin,A.C. (2004) PDBSprotEC: a Web-accessible database linking PDB chains to EC numbers via SwissProt. *Bioinformatics*, **20**, 986–988.
17. CCP4 (1994) The CCP4 suite: programs for protein crystallography. *Acta Crystallogr*., **D50**, 760–763.