

Application Note:

Mapping OMIM mutations to SwissProt

Andrew C. R. Martin^{a*}

^aDepartment of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT

ABSTRACT

Summary: OMIM is a manually curated resource containing detailed descriptions of inherited human disease. For some entries allelic variant (mutation) data are also available. However, there are no cross-references to a sequence database such as SwissProt and the numbering may not match the numbering of a SwissProt entry. Conversely, SwissProt provides cross references to OMIM, but the variant data within the downloadable SwissProt data do not indicate which mutations are disease-related (from OMIM) and which are polymorphisms. We have created an automatically updated resource which maintains a validated mapping between OMIM and SwissProt entries which may be queried interactively or downloaded for local processing.

Availability: The data may be queried on the basis of OMIM identifier, SwissProt accession or text from the OMIM entry title at <http://www.bioinf.org.uk/omim/>. The complete mapping (currently more than 7500 entries) may be downloaded as a comma-separated file or in XML.

Contact: andrew@bioinf.org.uk –or– martin@biochem.ucl.ac.uk

1 INTRODUCTION

Online Mendelian Inheritance in Man (OMIM) (McKusick, 2000) is a manually curated resource containing detailed descriptions of inherited human disease. The resource is largely maintained as free text and contains only limited computer parsable data. The resource may be queried interactively or downloaded as a compressed flat file. OMIM is described in detail in the FAQ available at <http://www.ncbi.nlm.nih.gov/Omim/omimfaq.html> and in the help document at <http://www.ncbi.nlm.nih.gov/Omim/omimhelp.html>.

OMIM entries are divided into six categories: 1) genes of unknown sequence — indicated with an asterisk, *; 2) descriptive entries, generally of a phenotype, without specific locus information — indicated with a hash, #; 3) entries with known gene and phenotype — indicated with a plus, +; 4) a confirmed mendelian phenotype but with unknown genetic/molecular basis — indicated with a percent-sign, %; 5) descriptive entries with a phenotype suspected of mendelian inheritance, but as yet unconfirmed — no symbol; 6) deprected entries — indicated with a caret, ^. For a large number of the entries (mostly in category 3), allelic variant (mutation) data are available. These data are relatively straightforward to parse from the OMIM flat file.

However, OMIM provides no cross-references to a sequence database such as SwissProt and the residue numbering used to indicate

mutations in OMIM may not match the numbering of residues in the corresponding SwissProt entry once it has been identified.

Conversely, SwissProt (Bairoch and Apweiler, 2000) provides VARIANT feature records which indicate observed mutations. However, within the downloadable SwissProt file, the level of annotation of these variants is variable. Specifically, it is not always clear whether variants are non-disease-related polymorphisms or are disease-related mutations. For example, P00480 (ornithine transcarbamylase) annotates mutations involved in disease as ‘in OTCD’ which is clear to the biologist reading the file (‘in OTC deficiency’), but is not computer parseable to distinguish the disease-related mutations from polymorphisms. Instead, this information appears in the separate SwissProt Variant database (Yip *et al.*, 2004) where each mutation is classified as ‘disease’, ‘polymorphism’, or ‘unclassified’. Unfortunately, these data are not downloadable (the only option is to ‘screen-scrape’ each HTML page) and the mutation data cannot easily be searched for disease-related text as is possible with OMIM.

We have therefore created a new database of OMIM mutation data linked to SwissProt entries. The data can be searched by OMIM entry number, SwissProt accession code or by allelic variant title text from OMIM. In addition the complete mapping may be downloaded as a comma-separated-value (CSV) file or as XML. This resource is invaluable for analysis of verified disease-related mutations in protein sequence and structure.

2 IMPLEMENTATION

The system has been implemented using Perl and PostgreSQL (<http://www.postgresql.org>). The OMIM flat file and SwissProt are mirrored locally. Data processing then proceeds as follows. First, the OMIM flat file is parsed to identify entries containing allelic variant records. Only missense mutations are recorded together with descriptive text from the title of the allelic variant. The OMIM identifier (a 6-digit number) together with the allelic variant sub-identifier (a 4-digit number) are used as keys into two database tables, one containing the descriptive text, the other containing the native and mutant residues and the residue number. Second, SwissProt is parsed to obtain OMIM identifiers linked to SwissProt accession codes. These data are stored in a second database table. Joining the two tables on the OMIM identifier now enables SwissProt accession codes to be linked to mutations extracted from OMIM.

However, there is still an important problem to be overcome. Very frequently the residue numbers for the mutations described in OMIM do not match the numbering in the SwissProt file: the ‘native’ residues extracted from OMIM do not match the residues

*to whom correspondence should be addressed

found at the respective locations in the PDB file. We therefore index the SwissProt FASTA file using a DBM hash to allow immediate access to any SwissProt entry. The native residues listed in the OMIM entry are then scanned against the sequence from SwissProt by applying an offset to the residue numbers. The offset at which the maximum number of matches is found is located. If the offset is zero or ± 1 (i.e. the numbering in OMIM matches the numbering in SwissProt, or is only out by one residue), and all residues match, then we accept this result irrespective of the number of residues involved. Otherwise, if the offset is zero and at least 2 residues match or if the offset is non-zero and at least 5 residues match, then we accept the result.

This offset is then applied to the residue numbers extracted from OMIM and the corrected residue numbers are inserted into the database table. In many cases, however, not all residues match when a non-zero offset has been applied. For a considerable number of these cases, the remaining residues match when an offset of zero is applied. In other words, it appears that some entries in the OMIM file have been numbered with respect to the SwissProt entry while others have been numbered differently. The residue numbers with zero-offset applied are then recorded in the database table.

We therefore have three categories of residues: A) those for which we are confident that the residue number with offset applied is correct, B) those which have been assigned with a zero offset, but where a majority of mutations have been assigned with confidence using a non-zero offset, C) those which cannot be assigned correctly in the SwissProt entry.

The current dataset contains 7588 mutations from 1356 OMIM entries, but of these only 1288 OMIM entries (95.0%, representing 7576 mutations, 99.8%) were cross-referenced from SwissProt. Of those OMIM entries which could be linked to SwissProt, 6989 (92.3%) of the mutations fall into Class A. A further 82 mutations (1.1%) from 23 OMIM entries (1.8%) fall into Class B. 505 residues (6.7%) from 235 OMIM entries (18.2%) have not been mapped and fall into Class C. In total, 250 OMIM entries which were cross-referenced in SwissProt (19.4%) contain errors or inconsistencies in the residue numbering which prevent at least some of the residues

being completely reliably mapped to SwissProt residues. Non-zero offsets had to be applied to 1751 (23.1%) of the mutations from 221 (17.2%) of the OMIM entries. The current version of these statistics can be viewed on the web site.

The procedure is completely automated and is run on a weekly basis using a Unix cron job.

3 AVAILABILITY

The web interface at <http://www.bioinf.org.uk/omim/> allows the data to be queried by SwissProt accession, OMIM identifier or text from the OMIM allelic variant titles. The results may be sorted by OMIM identifier or by SwissProt code and provide the OMIM identifier, SwissProt accession, native and mutant residues, the original residue number as it appeared in OMIM, the corrected residue number with reference to SwissProt and the title text from OMIM. The results are colour-coded with validated (Class A) residues appearing in white, Class B residues in yellow and Class C residues in red.

The complete mapping may be downloaded either as a comma-separated value (CSV) file or as XML. Details of the formats of these files, including a DTD for the XML file, are available on the web site.

4 ACKNOWLEDGEMENTS

This work was funded by The Wellcome Trust.

REFERENCES

- Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nuc. Ac. Res.*, **28**, 45–48.
- McKusick, V. A. (2000) Online Mendelian Inheritance in Man (OMIM) (TM), *McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD)*, <http://www.ncbi.nlm.nih.gov/omim/>.
- Yip, Y. L., Scheib, H., Diemand, A. V., Gattiker, A., Famiglietti, L. M., Gasteiger, E. and Bairoch, A. (2004) The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants, *Human Mut.*, **23**, 464–470.