© 2003 Oxford University Press

Nucleic Acids Research, 2003, Vol. 31, No. 1 499–501
DOI: 10.1093/nar/gkg038

The RESID Database of Protein Modifications: 2003 developments

John S. Garavelli*

The EMBL Outstation—European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 30, 2002; Accepted October 2, 2002

ABSTRACT

The RESID Database is a comprehensive collection of annotations and structures for protein pre-, co- and post-translational modifications including amino-terminal, carboxyl-terminal and peptide chain cross-link modifications. The RESID Database includes: systematic and alternate names, atomic formulas and masses, enzyme activities generating the modifications, keywords, literature citations, Gene Ontology cross-references, Protein Information Resource (PIR) and SWISS-PROT protein sequence database feature table annotations, structure diagrams and molecular models. This database is freely accessible on the Internet through the European Bioinformatics Institute at http://srs.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-page+LibInfo+-lib+RESID, through the National Cancer Institute—Frederick Advanced Biomedical Computing Center at http://www.ncifcrf.gov/RESID, or through the Protein Information Resource at http://pir.georgetown.edu/pirwww/dbinfo/resid.html.

INTRODUCTION

Among other authors, Richard Dawkins has compared proteins to ‘exceedingly numerous machine tools for molecular mass production’ (1). If proteins are the machine tools of the cell, then protein modifications are the drill bits, locking pliers, light sensors and throw switches of the tool shop. The presence of modifications in particular sequences can be useful in predicting the function of proteins (2,3), their location and their fate (4). The development of high-throughput mass spectrometry methods for finding and identifying protein modifications complements the high-volume production of genomic information (5–7). These new approaches and methods depend on bioinformatics tools that include databases of protein modifications (8). The RESID Database of Protein structure modifications began in 1993 as a database of modified amino acid residues represented as standardized features in the Protein Information Resource (PIR) Protein Sequence Database (9). The RESID Database was originally designed:

(i) to document covalent binding sites, modified sites and cross-links with bibliographic, structural descriptions, keywords and other annotations;
(ii) to provide more detailed chemical information than is possible in a protein sequence database, and enable users to recognize when different authors may use synonymous descriptions of previously described features;
(iii) to provide an adaptable mechanism for calculating the molecular weights of modified proteins and their peptide fragments;
(iv) to be accessed through Internet and database access programs with search capabilities and display chemical structures.

The RESID Database was distributed on CD-ROM accompanying the PIR in 1995, and in 1998 it was made available on the Internet with graphical depictions of modification structures, and later with molecular models. In 2000, quarterly releases of the database were made available by FTP from the Advanced Biomedical Computing Center of the National Cancer Institute—Frederick, Frederick, MD, USA. In 2001, the database began maintaining concurrent cross-references with the Gene Ontology Database (10), and it became available through the SRS server at the European Bioinformatics Institute (11). The RESID Database is the only publicly available database that attempts the comprehensive documentation of more than 320 structural, regulatory and active site prosthetic modifications. It also documents current feature representations in the PIR (12), and the SWISS-PROT and TrEMBL (13) protein sequence databases, and provides both graphical display and molecular models for protein modifications.

DATABASE DESCRIPTION

The RESID Database includes entries for the 23 alpha-amino acids known to be genetically encoded, including N-formyl methionine, selenocysteine and pyrrolysine (14), for the four ambiguous ‘residues’ represented in the IUPAC standard single-letter code (B, J, X and Z), and for more than 300 other residues either predicted or observed in proteins arising

*Tel: +44 01223492529; Email: john.garavelli@ebi.ac.uk, jsgaravelli@earthlink.net

through natural, co- or post-translational modification of the encoded amino acids. Naturally produced modifications that have incompletely determined structures, or that have not yet been located in protein sequences, as well as a few modifications that are known not to exist but appeared in the literature at one time are included for comprehensiveness. The database format provides both for the inclusion of artificially produced modifications that are commonly encountered in mass-spectrographic analysis and for amino acids in small peptides that are not genetically encoded. Such modifications will be annotated appropriately in the 'Source' record of the entry.

Information in RESID Database entries includes: dates for database entry and modification of either text or structure; a systematic name and Chemical Abstracts Service registry number for either the free residue or the covalently bound moiety; frequently encountered alternate names; the atomic formula and mass; with difference formulas and masses from the original amino acids; enzyme activities producing the modification; indicators for amino-terminal, carboxyl-terminal or peptide chain cross-link modifications; literature citations, keywords and feature table representations for the modification in the PIR and SWISS-PROT protein sequence databases. The RESID Database maintains concurrent cross-references to the Gene Ontology, the PIR and SWISS-PROT protein sequence databases, the MEDLINE and PubMed citation databases, and the Protein Data Bank (PDB) (15). Structural diagrams and molecular models are provided suitable for use with widely available molecular display programs.

AVAILABILITY AND ACCESS

The RESID Database is released quarterly, and update versions are prepared three or four times each quarter. The database is distributed in an XML format file with a supporting DTD file. The structure diagrams are in GIF format files, and the models are in PDB format files, which are collected and compressed in ZIP files. These files are available for FTP through the National Cancer Institute—Frederick Advanced Biomedical Computing Center at ftp://ftp.ncifcrf.gov/pub/users/residues/ or through the European Bioinformatics Institute at ftp://ftp.ebi.ac.uk/pub/databases/RESID/. Index files for almost all records, and Perl scripts for producing them, are available upon request.

Several search tools are available on the Internet for finding and displaying information in the RESID Database. At the European Bioinformatics Institute, the SRS server at http://srs.ebi.ac.uk/ provides access to the RESID Database under the Protein3DStruct library collection. This server provides text searching through indexes of 24 different record types, along with numeric search in a range of either the molecular mass of the modified residue or the difference in the mass produced by the modification. These masses are calculated using either chemical-average, or monoisotopic molecular mass commonly employed in mass-spectroscopy. Using the SRS server, a mass search can be performed by selecting the 'Weightp' or 'CWeightp' index, respectively, then entering the low and high mass values of the range separated by a colon. For example, selecting 'CWeightp' and entering '177 : 179' will search for all entries with correction (or difference) masses between 177 and 179 daltons. At the PIR, the RESID Database search page at http://pir.georgetown.edu/cgi-bin/resid provides text searching through indexes of 19 different record types. Selectors also provide list of all entries, list of derivatives of each standard amino acid, and search in a range of formula or correction masses. To perform the same mass search as in the previous example, select the 'Weight C phys' option, enter '178', and in the '±' box enter '1'.

General information on the RESID Database and its current availability can be found at these URLs:
http://home.earthlink.net/~jsgaravelli/RESIDInfo.HTML;
http://www.ncifcrf.gov/RESID;
http://srs.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-page+LibInfo+-lib+RESID;
http://pir.georgetown.edu/pirwww/dbinfo/resid.html.
After an update, some of these web pages might be delayed in reflecting the information about the latest version of the database. However, the version number is indicated in the 'release' and 'date' records of the XML file, and in the README or the Release Notes file available at the FTP sites.

The RESID Database is copyrighted and is distributed free with no license required. It is appreciated if users cite this article or the introductory announcements of the RESID Database (9).

SUBMISSIONS AND REVISIONS

Researchers are invited to submit information for new entries or for the revision of existing entries in the RESID Database. Those wishing to submit material may do so by E-mail directed to the author's attention at either jsgaravelli@earthlink.net or john.garavelli@ebi.ac.uk. New database entries are assigned unique identifiers, which may be cited in publications.

ACKNOWLEDGEMENTS

Hosting and support for the RESID Database that is kindly provided by the European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom, by the Advanced Biomedical Computing Center, SAIC, National Cancer Institute—Frederick, Frederick, MD, USA, and by the Protein Information Resource, Georgetown University, Washington, DC, USA. The RESID Database of Protein Modifications is the service mark of J. S. Garavelli.

REFERENCES

1. Dawkins,R. (1996) *Climbing Mount Improbable.* Penguin Books Ltd, London, UK, p. 66.
2. Degtyarenko,K. (2000) Bioinorganic motifs: towards functional classification of metalloproteins. *Bioinformatics*, **16**, 851–864.
3. Jensen,L.J., Gupta,R., Blom,N., Devos,D., Tamames,J., Kesmir,C., Nielsen,H., Staerfeldt,H.H., Rapacki,K., Workman,C., Andersen,C.A., Knudsen,S., Krogh,A., Valencia,A. and Brunak,S. (2002) Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol.*, **319**, 1257–1265.
4. Nakai,K. (2001) Review: prediction of *in vivo* fates of proteins in the era of genomics and proteomics. *J. Struct. Biol.*, **134**, 103–116.
5. Yates,J.R.,III, Eng,J.K., McCormack,A.L., and Schieltz,D. (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.*, **67**, 1426–1436.

6. Annan,R.S. and Carr,S.A. (1997) The essential role of mass spectrometry in characterizing protein structure: mapping posttranslational modifications. *J. Protein Chem.*, **16**, 391–402.
7. Wilkins,M.R., Gasteiger,E., Gooley,A.A., Herbert,B.R., Molloy,M.P., Binz,P.A., Ou,K., Sanchez,J.C., Bairoch,A., Williams,K.L. and Hochstrasser,D.F. (1999) High-throughput mass spectrometric discovery of protein post-translational modifications. *J. Mol. Biol.*, **289**, 645–657.
8. Meng,F., Cargile,B.J., Patrie,S.M., Johnson,J.R., McLoughlin,S.M. and Kelleher,N.L. (2002) Processing complex mixtures of intact proteins for direct analysis by mass spectrometry. *Anal. Chem.*, **74**, 2923–2929.
9. Garavelli,J.S. (1993) A database of protein structure modifications. *Protein Sci.*, **2** (Suppl. 1), 133.
10. Ashburner,M. *et al.* (2001) Creating the Gene Ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
11. Zdobnov,E.M., Lopez,R., Apweiler,R. and Etzold,T. (2002) The EBI SRS server-new features. *Bioinformatics*, **18**, 1149–1150.
12. Wu,C.H., Yeh,L.-S.L., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E., Vinayata,C.R., Zhang,J. and Barker,W.C. (2003) Protein Information Resource. *Nucleic Acids Res.*, **31**, 345–347.
13. Boekmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreocjer,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledge base and its supplement TrEMBL. *Nucleic Acids Res.* **31**, 365–370.
14. Srinivasan,G., James,C.M. and Krzycki,J.A. (2002) Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. *Science*, **296**, 1459–1462.
15. Westbrook,J., Feng,Z., Chen,L., Yang,H. and Berman,H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.* **31**, 489–491.