198–199 *Nucleic Acids Research, 1999, Vol. 27, No. 1*

The RESID Database of protein structure modifications

John S. Garavelli*

National Biomedical Research Foundation, Protein Information Resource, Washington, DC 20007, USA

Received September 1, 1998; Revised September 22, 1998; Accepted September 29, 1998

ABSTRACT

Because the number of post-translational modifications requiring standardized annotation in the PIR-International Protein Sequence Database was large and steadily increasing, a database of protein structure modifications was constructed in 1993 to assist in producing appropriate feature annotations for covalent binding sites, modified sites and cross-links. In 1995 RESID was publicly released as a PIR-International text database distributed on CD-ROM and accessible through the ATLAS program. In 1998 it was made available on the PIR Web site at http://www-nbrf.georgetown.edu/pir/searchdb.html . The RESID Database includes such information as: systematic and frequently observed alternate names; Chemical Abstracts Service registry numbers; atomic formulas and weights; enzyme activities; indicators for N-terminal, C-terminal or peptide chain cross-link modifications; keywords; and literature citations with database cross-references. The RESID Database can be used to predict atomic masses for peptides, and is being enhanced to provide molecular structures for graphical presentation on the PIR Web site using widely available molecular viewing programs.

INTRODUCTION

The PIR-International Protein Sequence Database (1) strives for comprehensiveness, accuracy, precision and consistency in annotation. Unfortunately, chemical information beyond the biopolymer sequence was generally neglected in sequence databases and, when included, occurred in free-text comments that were human- but not computer-readable. The representation of chemical information for modifications was a particular problem. For example, a residue represented by an 'X' in a published sequence might be accompanied by an explanation that a non-standard residue such as hypusine had been found. Unless the reader was familiar with that compound and knew it was a post-translational modification of lysine, the annotation would not be very helpful and the sequence data could be misrepresented. Furthermore, when such annotations existed only in free-text, sequences containing particular modified residues could not easily be found and checked for consistency. To deal with this problem, the PIR became the first sequence database to augment sequence information with position-dependent structural information in feature records that are both human- and computer-readable. PIR feature records also assist users by indicating whether features were experimentally observed or predicted.

Without documentation to assist in preparing featuures, annotators can only rely on authors' comments. What one author reports as a 'meso-lanthionine residue', another may report as 'cysteine thioether-linked to D-alanine'. If these were annotated only as reported, one simple query could not find both entries, and annotators and users not familiar with this particular compound would not realize that these authors are reporting the same modification. The feature records could not be adequately or consistently documented within the database and were difficult to update to reflect current states of knowledge concerning post-translational modifications. Until 1992, when the PIR began standardizing feature records, errors and misrepresentations were common in protein sequence databases. For example, prior to that date 3-methyl-lanthionine (a cross-link between threonine and cysteine found in some antibiotics) appeared as one of at least five different features, including misspellings and misrepresentations. After standardization, that feature (which includes completely specified stereochemistry) appeared in PIR entries only as 'Cross-link: (2S,3S,6R)-3-methyl-lanthionine (Cys-Thr)'.

Because of these difficulties and because the number of different protein structure modifications had been steadily increasing, we imposed a restricted vocabulary and standardized syntax for feature annotations in the PIR Protein Sequence Database. In 1993 we constructed the RESID Database of protein modifications to assist annotators in using the standard syntax and vocabulary.

This database was designed: (i) to document standardized features annotations for covalent binding sites, modified sites and cross-links, and to provide appropriate keywords and other annotations to accompany those features; (ii) to convey chemical information in more detail than is possible in a protein sequence database, and to enable annotators to recognize when authors are using synonymous descriptions of previously described features; (iii) to provide an adaptable mechanism for calculating the molecular weights of modified proteins and their peptide fragments; and (iv) to be accessible through the ATLAS multi-database access program, and the Internet database access programs.

During each update, all feature records in the PIR Protein Sequence Database are automatically checked for syntax and vocabulary using the standard records in the latest version of the RESID Database. These procedures ensure accuracy and consistency in annotation and help propagate annotation revisions quickly throughout PIR Protein Sequence Database entries.

The RESID Database was publicly released in 1995 with 181 entries. Within a year it was recognized by a reviewer in the field of mass-spectroscopic analysis of peptides as a resource useful in identifying post-translationally modified peptides (2).

*Tel: +1 202 687 2121; Fax: +1 202 687 1662; Email: garavelli@nbrf.georgetown.edu

DATABASE DESCRIPTION

The database includes entries for the 22 α-amino acids currently known to be genetically encoded, including *N*-formyl methionine and selenocysteine, three ambiguous 'residues' represented in the standard single-letter code, and more than 225 other residues observed in proteins and known to arise through natural, post-translational modification of encoded amino acids.

Information in RESID Database entries includes: dates for database entry and modification of text and structure; a systematic chemical name and Chemical Abstracts Service registry number for the free residue; frequently observed alternate names; the atomic formula and weight; original amino acids with difference formulas and weights; enzyme activities producing the modification; indicators for N-terminal, C-terminal or peptide chain cross-link modifications; how the modification is presented in feature tables in the Protein Sequence Database and the keywords associated with it; and literature citations with database cross-references. The RESID Database maintains dynamic links with the PIR-International Protein Sequence and NRL_3D databases, and concurrent cross-references to Chemical Abstracts (CAS), the MedLine citation database, and the Protein Data Bank (PDB). (CAS Registry Numbers are copyrighted by the American Chemical Society and used with permission of the Chemical Abstracts Service of the American Chemical Society.) The RESID Database provides a means for calculating both chemical-average and monoisotopic molecular weights for modified peptides in order to facilitate their identification by mass-spectroscopy (3,4). Release 16.00 contains over 250 entries describing features annotated in the Protein Sequence Database.

Work is underway to provide molecular structures for graphical presentation on the PIR Web site using widely available molecular viewing programs such as RasMol and Kinemage (5,6).

A sample entry in the RESID Database is presented in Table 1.

Table 1. RESID Database sample entry

RESID:AA0252
dipyrrolylmethyl-L-cysteine
Alternate names: 3-[5-(3-acetic acid-4-propanoic acid-1-pyrrol-2-yl)methyl-3-acetic acid-4-propanoic acid-1-pyrrol-2-yl]methylthio-2-aminopropanoic acid; dipyrrole cofactor; dipyrrolylmethanemethyl-L-cysteine; dipyrromethane cofactor; pyrromethane cofactor
Systematic name: 3-[5-[4-(2-carboxy)ethyl-3-carboxymethyl-1-pyrrol-2-yl] methyl-4-(2-carboxy)ethyl-3-carboxymethyl-1-pyrrol-2-yl]methylthio-2-aminopropanoic acid
Cross-references: CAS:29261-13-0
Formula: C 23 H 27 N 3 O 9 S 1
Formula weight: #chem 521.55 #phys 521.1468
Correction formula: C 20 H 22 N 2 O 8
Correction weight: #chem 418.41 #phys 418.1376
Date: 12-Dec-1997 #structure_revision 12-Dec-1997 #text_change 12-Dec-1997
Jordan, P.M.; Warren, M.J.; Williams, H.J.; Stolowich, N.J.; Roessner, C.A.; Grant, S.K.; Scott, A.I.
FEBS Lett. 235, 189-193, 1988
Title: Identification of a cysteine residue as the binding site for the dipyrromethane cofactor at the active site of Escherichia coli porphobilinogen deaminase
Reference number: A58694
Note: radioisotope labeling; (13)C-NMR characterization
Miller, A.D.; Hart, G.J.; Packman, L.C.; Battersby, A.R.
Biochem. J. 254, 915-918, 1988
Title: Evidence that the pyrromethane cofactor of hydroxymethylbilane synthase (porphobilinogen deaminase) is bound to the protein through the sulphur atom of cysteine-242.
Reference number: A58695; MUID:89061636
Note: chemical characterization
Hart, G.J.; Miller, A.D.; Battersby, A.R.
Biochem. J. 252, 909-912, 1988
Title: Evidence that the pyrromethane cofactor of hydroxymethylbilane synthase (porphobilinogen deaminase) is bound through the sulphur atom of a cysteine residue.
Reference number: A58696
Note: chemical characterization; (13)C-NMR identification
Louie, G.V.; Brownlie, P.D.; Lambert, R.; Cooper, J.B.; Blundell, T.L.; Wood, S.P.; Malashkevich, V.N.;
Haedener, A.; Warren, M.J.; Shoolingin-Jordan, P.M.
Proteins 25, 48-78, 1996
Title: The three-dimensional structure of Escherichia coli porphobilinogen deaminase at 1.76-anstroms resolution.
Reference number: A58699; MUID:96323958
Note: X-ray crystallography, 1.76 angstroms
Louie, G.V.; Brownlie, P.D.; Lambert, R.; Cooper, J.B.; Blundell, T.L.; Wood, S.P.; Warren, M.J.; Woodcock, S.C.; Jordan, P.M.
submitted to the Brookhaven Protein Data Bank, November 1992
Reference number: A51329; PDB:1PDA
Note: X-ray crystallography, 1.76 angstroms
Sequence code: C

Residues	Feature
	Modified site: dipyrrolylmethanemethyl (Cys) (covalent)

AVAILABILITY AND ACCESS

The RESID Database is updated weekly and distributed quarterly with the PIR-International Protein Sequence Database. It is available on CD-ROM and through the Internet. Entries may be retrieved by entry code at http://www-nbrf.georgetown.edu/pir/resid/get.html and may be located by text search at http://www-nbrf.georgetown.edu/pir/resid/find.html

Using the ATLAS multidatabase access program on CD-ROM, the RESID database is automatically available with all other PIR-International databases, or it can be searched exclusively by using the 'BASES RESID' command or the define bases menu selection. The search commands usable with the RESID Database are FIND for systematic names and synonyms, and KEYWORD, AUTHOR, FEATURE, JOURNAL, REFERENCE and CROSS_REF for the corresponding records. Detailed instructions are provided in the ATLAS program documentation on the CD-ROM and at http://www-nbrf.georgetown.edu/pir/atcd.html . Information about the ATLAS CD-ROM is available in the PIR-International Protein Sequence Database article elsewhere in this issue.

SUBMISSIONS AND REVISIONS

The indexing of novel entities in abstract sources such as MedLine is inherently inconsistent and unreliable, and most new entries for RESID are generated as a result of sequence annotation and re-evaluation. The author invites the submission of information for new entries or for the revision of existing entries. Those wishing to submit material may do so using the electronic submission form at http://www-nbrf.georgetown.edu/pir/pirsub.html or by Email directed to the author's attention at PIRMAIL@nbrf.georgetown.edu. New database entries are assigned unique access codes which may be cited in publications. It would be appreciated if references to the RESID Database would cite this article or the introductory announcement (7).

ACKNOWLEDGEMENTS

Enhancement of the RESID Database is supported by NSF grant DBI-9808414. The RESID Database is copyrighted by the National Biomedical Research Foundation.

REFERENCES

1 Barker,W.C., Garavelli,J.S., McGarvey,P.B., Marzec,C.R., Orcutt,B.C., Srinivasarao,G.Y., Yeh,L.S., Ledley,R.S., Mewes,H.W., Pfeiffer,F., Tsugita,A. and Wu,C. (1999) *Nucleic Acids Res.*, **27**, 39–43.
2 Yates,J.R. (1996) *Methods Enzymol.*, **271**, 351–377.
3 Biemann,K. and Scoble,H. (1987) *Science*, **237**, 992–998.
4 Takao,T., Yoshino,K., Suzuki,N. and Shimonishi,Y. (1990) *Biomed. Environ. Mass Spectrom.*, **19**, 705–712.
5 Sayle,R.A. and White,E.J.M. (1995) *Trends Biochem. Sci.*, **20**, 374–376.
6 Richardson,D.C. and Richardson,J.S. (1994) *Trends Biochem. Sci.*, **19**, 135–138.
7 Garavelli,J.S. (1993) *Protein Sci.*, **2** (Suppl. 1), abstract 450.