The RESID Database of Protein Modifications as a resource and annotation tool

John S. Garavelli

The EMBL Outstation – European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambs., UK

The RESID Database of Protein Modifications is a comprehensive collection of annotations and structures for protein modifications and cross-links including pre-, co-, and post-translational modifications. The database provides: systematic and alternate names, atomic formulas and masses, enzymatic activities that generate the modifications, keywords, literature citations, Gene Ontology (GO) cross-references, protein sequence database feature table annotations, structure diagrams, and molecular models. This database is freely accessible on the Internet through resources provided by the European Bioinformatics Institute (http://www.ebi.ac.uk/RESID), and by the National Cancer Institute – Frederick Advanced Biomedical Computing Center (http://www.ncifcrf.gov/RESID). Each RESID Database entry presents a chemically unique modification and shows how that modification is currently annotated in the protein sequence databases, Swiss-Prot and the Protein Information Resource (PIR). The RESID Database provides a table of corresponding equivalent feature annotations that is used in the UniProt project, an international effort to combine the resources of the Swiss-Prot, TrEMBL and PIR. As an annotation tool, the RESID Database is used in standardizing and enhancing modification descriptions in the feature tables of Swiss-Prot entries. As an Internet resource, the RESID Database assists researchers in high-throughput proteomics to search monoisotopic masses and mass differences and identify known and predicted protein modifications.

Keywords: Bioinformatics / Database annotation / Protein cross-links / Protein modifications

Received 24/10/03
Revised 4/2/04
Accepted 6/2/04

1 Introduction

With continual advances in proteomics, not only must protein sequence databases include more sequence information, they must also strive to present all the new proteomics discoveries in a way that is as meaningful and as easily accessible as possible. While improving computer and algorithm performances in sequence analysis have managed to keep pace with database growth, the increasing volume of genomics and proteomics data requires that annotations include more detail in order to improve discrimination in search and query operations. However, the inclusion of more detail requires a continual re-evaluation of the annotation data with regard to its relevance and significance in light of later findings. Data deposited in an archive with no later indications of its accuracy and significance rapidly loses its value. In order to provide reliability in search and query operations, consistency and standardization of annotation are required.

In biological sequence databases, feature tables are used, among other things, to describe chemical, structural or functional properties of the nucleic acid or protein sequence, and associate those properties with specific locations or regions of the sequence. It is also common for protein sequence databases to indicate the state of knowledge about the features, whether they are predicted, or experimentally verified to be present or absent. Protein modifications and cross-links are particularly important features that must be represented. The occurrence of modifications in proteins has been used to pre-

Correspondence: John S. Garavelli, The EMBL Outstation – European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambs. CB10 1SD, UK
E-mail: john.garavelli@ebi.ac.uk, jsgaravelli@earthlink.net
Fax: +44-(0)1223-494468

Abbreviations: EBI, European Bioinformatics Institute; **GO**, Gene Ontology; **IUPAC**, International Union of Pure and Applied Chemistry; **PDB**, Protein Data Bank; **PIR**, Protein Information Resource; **UniProt**, Universal Protein Resource dict the function of proteins [1, 2], their location and their fate [3]. With the development of high-throughput mass spectrometry, methods for finding and identifying protein modifications are beginning to complement the high-volume production of genomic information [4–7]. These new approaches and methods in high-throughput proteomics critically depend on bioinformatics tools that include databases of protein modifications. For example, the "top-down" approach to mass spectrometric proteome analysis [8] starts with genome sequences to generate potential protein translations. Those translations are used to generate tables of masses expected for whole proteins and their fragments. Those tables are searched to identify the peptide masses, and patterns of masses, observed in mass spectrometry of peptide mixtures. Finally, discrepancies between observed and expected masses of predicted peptides are resolved by fitting mass differences from the table of potential modifications.

The RESID Database of Protein Modifications [9, 10] began in 1993 as a database of the modified amino acid residues that were represented as standardized features in the Protein Information Resource (PIR) Protein Sequence Database. The RESID Database was 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 feasible or desirable in a protein sequence database, and enable users to recognize when different authors might be 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 accessed through Internet and database access programs with search capabilities and display chemical structures.

The RESID Database was distributed on hard-media 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 (GO) of the Gene Ontology Consortium [11]. Later that year, it became available through the SRS server at the European Bioinformatics Institute (EBI) [12].

The RESID Database is the only publicly available database that attempts the comprehensive documentation of more than 330 structural, regulatory, and active site modifications, covalent binding sites, sites for the attachment of active site prosthetic groups such as phosphopantetheine, and protein cross-links such as the *N*6-glycyllysine isopeptide cross-link important in the attachment of ubiquitin, SUMO or Nedd8. It also documents how those modifications are annotated as features in the Swiss-Prot and TrEMBL [13], and the PIR [14] protein sequence databases, and provides both graphical display and molecular models for protein modifications.

2 Database description

The RESID Database includes entries for the 23 α-amino acids currently known to be genetically encoded, including *N*-formyl methionine, selenocysteine, and pyrrolysine [15], for the four ambiguous "residues" represented in the International Union of Pure and Applied Chemistry (IUPAC) standard single-letter code (B, J, X, and Z), and for more than 330 other residues either predicted or observed in proteins arising through natural modification of encoded amino acids. Also included for comprehensiveness are naturally produced modifications that have incompletely determined structures, modifications that have not yet been located in protein sequences, and a few modifications that were predicted at one time, but are now deprecated, or known not to exist. New entries are being added at a rate of between 10 and 20 per year.

The RESID Database consists of text entries in XML format, and associated files containing DTD for the XML format, graphic images and molecular models. XML records are defined for: dates for database entry, and last modification of text and structure; a systematic name according to IUBMB IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN) rules [16] (available at http://www.chem.qmul.ac.uk/iupac/AminoAcid/); a Chemical Abstracts Service registry number for either the free modified residue, or for the covalently bound moiety; frequently encountered alternate names; the atomic formula and mass; difference formulas from the original amino acids and the corresponding masses; enzyme activities producing the modification; indicators for amino-terminal, carboxyl-terminal or peptide chain cross-link modifications; literature citations; keywords; and representation of the modification in the feature annotations of the PIR and Swiss-Prot protein sequence databases. Concurrent cross-references are maintained to GO terms [11], the COMe the bioinorganic motif database [17], the PIR and Swiss-Prot protein sequence databases, the PubMed citation databases, and the Protein Data Bank (PDB) [18]. The associated molecular model files are in PDB format, suitable for use by most molecular display programs and for library construction with molecular modeling programs. The models are based on structures found in the PDB, when these are available, or on generally accepted molecular geometry parameters when PDB structures are not available. The associated graphic image files are in GIF format. To illustrate the information available in a typical entry, the entry for RESID:AA0106, *S*-palmitoyl-L-cysteine, transformed from XML to HTML, is shown in Table 1. This table shows the entry consisting of the permanently assigned entry identifier, followed by the entry name, an alternative name, a systematic name, and an identifier for the corresponding entity in the Chemical Abstracts Service Registry. Four references are presented for detection and structural determination of this modification. There are two comments, one of which is a warning pointing to a different modification, *N*-palmitoylcysteine, which may be confused with *S*-palmitoylcysteine.

The database fits a relational model, so the entries are not arranged in any predetermined order. The restricted vocabulary used in the keyword, source and condition records is intended to enable searching for many of the categories commonly encountered in classifying modifications. The mass and mass difference records can be used to order the entries to assist in mass spectrometric analysis. With the reciprocal cross-references between RESID and GO, the GO can be used to generate ontological views of protein modifications.

Table 1. Database Entry AA0106

RESID:AA0106
_S_-palmitoyl-L-cysteine

Alternate names: hexadecanoate cysteine thioester

Systematic names: (*R*)-2-amino-3-(hexadecanoylthio)propanoic acid
Cross-references: CAS:114507–35–6

Dates:
Created: 31-Mar-1995
Structure revised: 31-Mar-1995
Text changed: 30-May-2003

Formula: C 19 H 35 N 1 O 2 S 1
Formula weight: #chem 341.55 #phys 341.2388

Correction formula: C 16 H 30 N 0 O 1 S 0
Correction weight: #chem 238.41 #phys 238.2297

Reference 1:
Authors: Bach, R.; Konigsberg, W. H.; Nemerson, Y.
Journal: Biochemistry 27, 4227–4231, 1988
Title: Human tissue factor contains thioester-linked palmitate and stearate on the cytoplasmic half-cystine.
Cross-references: PIR:A37422; PMID:3166978
Note: chemical characterization

Reference 2:
Authors: Stults, J. T.; Griffin, P. R.; Lesikar, D. D.; Naidu, A.; Moffat, B.; Benson, B. J.
Journal: Am. J. Physiol. 261, L118-L125, 1991
Title: Lung surfactant protein SP-C from human, bovine, and canine sources contains palmityl cysteine thioester linkages.
Cross-references: PIR:A61249; PMID:1872406
Note: mass spectrographic identification

Reference 3:
Authors: Johansson, J.; Szyperski, T.; Curstedt, T.; Wuthrich, K.
Journal: Biochemistry 33, 6015–6023, 1994
Title: The NMR structure of the pulmonary surfactant-associated polypeptide sp-C in an apolar solvent contains a valyl-rich alpha-helix.
Cross-references: PIR:A58575; PMID:8180229
Note: (1)H-NMR identification

Table 1. Continued

Reference 4:
  Authors: Steinert, P. M.; Kim, S. Y.; Chung, S. I.; Marekov, L. N.
  Journal: J. Biol. Chem. 271, 26242–26250, 1996
  Title: The transglutaminase 1 enzyme is variably acylated by myristate and palmitate during differentiation in epidermal keratinocytes.
  Cross-references: PMID:8824274
  Note: mass spectrographic identification; differential S-palmitoylation and S-myristoylation (see RESID:AA0307)

Comment: Although the predominant palmitoyl transferase in mammalian systems appears to utilize a mixture of saturated and unsaturated fatty acids, some systems may be more specific in their incorporation of other fatty acids. See RESID:AA0307 and RESID:AA0308.

Comment: This modification should not be confused with N-palmitoyl-cysteine (see RESID:AA0060.

Generating Enzyme: protein-cysteine *S*-palmitoyltransferase (EC 2.3.1.-)

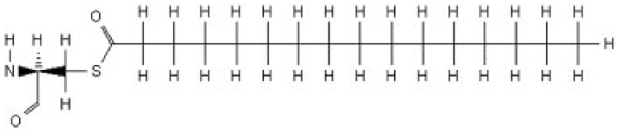
Sequence code: C
  Conditions: combinable
  Cross-references: GO:0018230; GO:0042050

Source: natural

Keywords: lipoprotein; palmitoylation; thioester bond

PIR Features
  Binding site: palmitate (Cys) (covalent)

SWISS-PROT Features
  LIPID S-palmitoyl cysteine



A sample entry from the RESID Database after a transformation from XML to HTML is presented.

3 The RESID Database as an annotation tool

In 2002 the European Bioinformatics Institute, the Swiss Institute of Bioinformatics, and the Protein Information Resource group at the Georgetown University Medical Center and the National Biomedical Research Foundation formed a consortium. The objective of this international effort, the Universal Protein Resource (UniProt) Project [19], was to combine the resources of the Swiss-Prot, TrEMBL, and PIR protein sequence databases and produce a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase. As a part of this project, the feature annotations of the Swiss-Prot and TrEMBL databases are being standardized and enhanced, and merged with information in the standardized feature annotations of the PIR. Briefly, these annotations are being merged and standardized by the following procedure. On the basis of sequence, taxonomic classification and literature cited, entries in the PIR are matched with entries in Swiss-Prot/TrEMBL. Any information that is unique in PIR is either added to the matched Swiss-Prot/TrEMBL entry, or used to produce a new UniProt entry. In particular, the feature table information must be mapped through a sequence alignment of the matched entries, and then corresponding features matched through both sequence location and the chemical description of the modification. Because the RESID Database tracks how each chemically unique modification is annotated in the feature table descriptions of both Swiss-Prot and the PIR, it provides part of the table of corresponding feature annotations that is required by the UniProt project. Each unique modification feature in the PIR has a corresponding unique RESID entry, which also documents the current Swiss-Prot feature annotation, or annotations, for that modification.

Another way in which the RESID Database is being used as an annotation tool is in the enhancement of feature descriptions. The enhancement of the feature descriptions in Swiss-Prot was important, because many of those feature descriptions were simply names of processes producing a modification, that is names like "PHOSPHORYLATION" or "METHYLATION". These process names did not unambiguously identify which of the different chemical modifications that could arise through that process actually occurred in a peptide sequence, nor did they specify which amino acid was supposedly modified, so that mislocation errors could not be detected during annotation preparation. For example, before this revision was undertaken the feature description "METHYLATION" was used for at least 23 different chemical modifications, including two different leucine modifications, *N*-methyl leucine and leucine methyl ester, and five different arginine modifications, ω*N*-methylarginine, symmetrical ω*N*,ω'*N*-dimethylarginine, unsymmetrical ω*N*,ω*N*-dimethylarginine, 5-methylarginine, and *N*5-methylarginine. Unless the chemical identity of the modification is uniquely identified in the feature description, it is impossible to compute the correct molecular mass or construct a complete molecular model for the modified peptide. Further, when chemically unique modifications could not be collected and analyzed for the sequences patterns they occur with, it is difficult to discover the biological and chemical rules governing protein modifications.

In constructing the procedures that perform the revision of Swiss-Prot feature annotations, information in the RESID Database is used to build decision trees for analyzing the feature annotations. In these procedures each feature in a Swiss-Prot entry is automatically analyzed to determine the RESID entry that it corresponds with based on the modification description, the sequence, and other information in the Swiss-Prot entry, such as taxonomic classification. A revised feature is then prepared using the standardized, restricted vocabulary designated for that RESID entry. A slightly simplified portion of this decision tree analysis is presented in Fig. 1. In this example, entries that had a feature with the description "PALMITATE" were analyzed. By considering the sequence at the feature position, as well as the relative position in the sequence, and the position relative to other features, this feature could be assigned as one of five different modifications in the RESID Database. In this case, the taxonomic classification for the source species is also checked for quality assurance purposes – these particular modifications have not been observed in Archaebacteria. During 2003 and 2004, new and revised descriptions, based on the modifications collected in the RESID Database, were created for 30 CROSSLNK, 24 LIPID, and 130 MOD_RES modifications annotated in more than 18 600 Swiss-Prot features.

4 The RESID Database as a resource

The RESID Database is available for anonymous 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 (EBI) at ftp://ftp.ebi.ac.uk/pub/databases/RESID/. An Internet search engine makes the RESID Database available for search and query operations. At the EBI, the SRS server at http://srs.ebi.ac.uk/ provides access to the RESID Database under the "Protein structure database" collection. This search engine permits text searching through indexes of 24 different record types, along with numeric search within a range for either the molecular mass of the modification, or the mass difference produced by the modification. These masses are provided as chemical, or isotopic average, molecular masses (Weightc), and as physical, or most common isotope, molecular masses (Weightp). The latter is commonly used in mass spectrometry. The corrections, or differences, between the mass of the modified residues and the mass of the encoded amino acid are also provided as chemical masses (CWeightc), and as physical masses (CWeightp). Using the SRS server a monoisotopic mass search can be performed by selecting the "Weightp" or "CWeightp" index, for modification mass or the difference mass, 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 (difference) masses between 177 and 179 Da.

Natural modifications that have been chemically characterized, but not yet located in any sequence, are included in the RESID Database (Table 2). These are included to

Table 2. Some observed or predicted modifications not yet located in a peptide sequence

RESID ID	Name
AA0122	L-2-Aminoadipic acid
AA0175	L-3'-Bromophenylalanine
AA0296	*O*-(*N*-Acetylglucosamine-1-phosphoryl)-L-serine
AA0297	*O*-(Phosphoglycosyl-D-mannose-1-phosphoryl)-L-serine
AA0302	L-Aspartimide
AA0303	L-Glutamimide
AA0304	L-β-Carboxyaspartic acid
AA0308	*S*-Palmitoleyl-L-cysteine
AA0312	*N*6-3,4-Didehydroretinylidene-L-lysine
AA0320	L-β-Methylthioasparagine

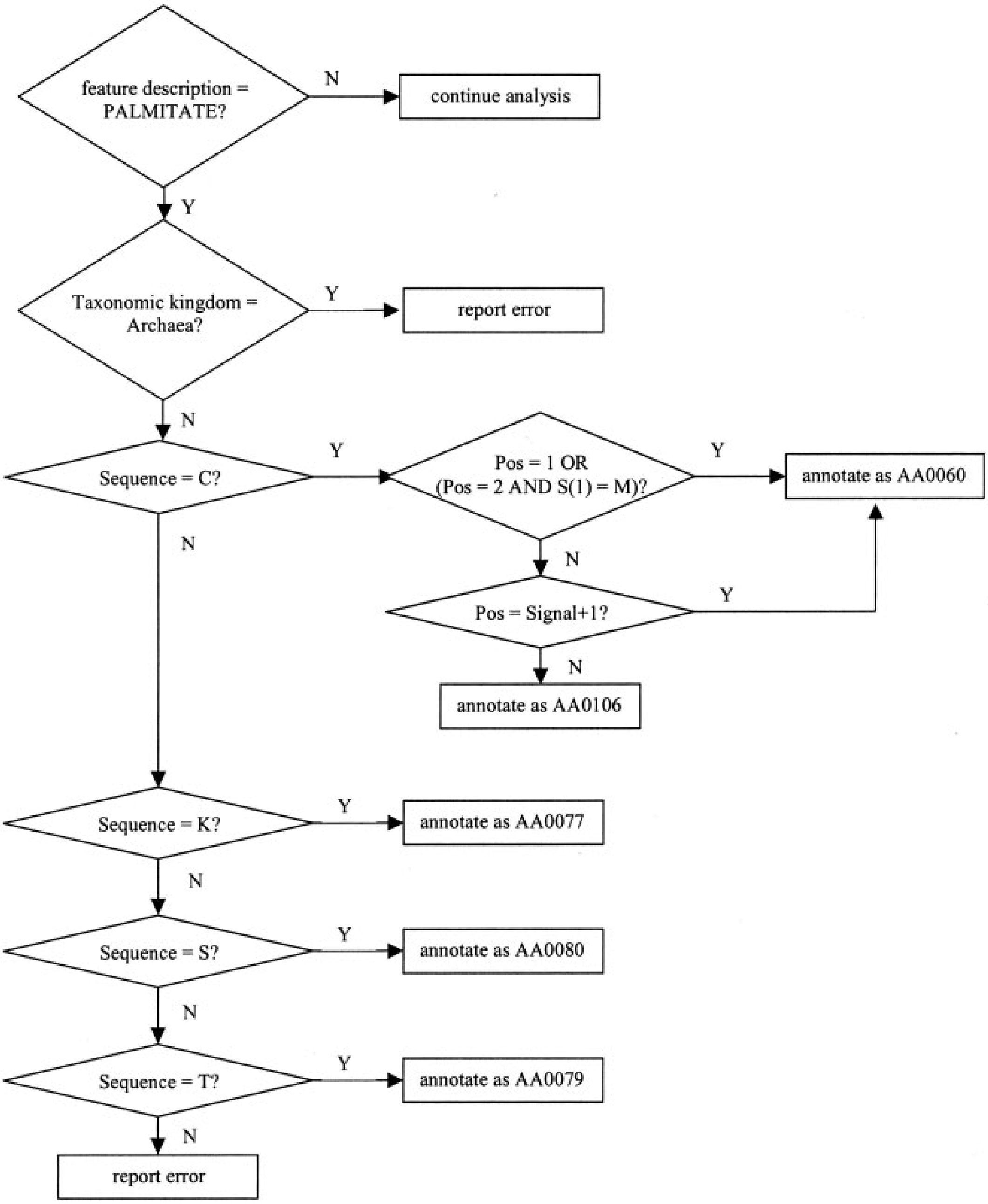


Figure 1. Decision tree analysis of Swiss-Prot "PALMITATE" feature. Each feature in a Swiss-Prot entry has a description field. In 2003 and 2004, these descriptions fields were standardized and enhanced. The RESID Database was used in identifying the ambiguous descriptions. This decision tree begins with the analysis of feature descriptions containing the text "PALMITATE". At step 2, the taxonomic classification for the source species in the entry is checked for quality assurance purposes. No modification feature with the description "PALMITATE" should occur in a protein from *Archaebacteria*. At step 3, if the sequence at the feature position is a "C", a cysteine residue, then depending on whether that feature position is amino terminal (at position 1, at position 2 after an initial methionine, or at a position after a signal peptide cleavage), it is identified as the modification for RESID:AA0060, otherwise it is identified as the modification for RESID:AA0106. At steps 4, 5 and 6, the sequence at the feature position is successively checked for "K", lysine, "S", serine, or "T", threonine, and identified as the modification for RESID:AA0077, RESID:AA0080, or RESID:AA0079, respectively. Otherwise, if the sequence at the feature position is some other residue, an error is reported for manual correction.

assist researchers, especially in high-throughput proteomics, in recognizing these potential modifications if they should encounter these mass signatures. These modifications that are known to exist but not yet been located in any sequence are typically omitted from lists of masses compiled for identification of peptide modifications. Some examples of modifications that may eventually be identified through proteome analysis are β-carboxyaspartic acid, β-methylthioaspartic acid and β-methylthioasparagine. β-Carboxyaspartic is known to occur in some ribosomal proteins, but is so labile it has not been possible to locate in a peptide sequence through the usual methods of analysis [20]. β-Methylthioaspartic acid has been observed in the ribosomal S12 protein, but thus far only from *Escherichia coli* [21], while β-methylthioasparagine could be a modification predicted to occur in the ribosomal S12 proteins of species where aspartic acid is conservatively replaced by asparagine at the corresponding position.

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://www.ebi.ac.uk/RESID; http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+LibInfo+-lib+RESID. 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.

Researchers are encouraged 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 at either jsgaravelli@earthlink.net or john.garavelli@ebi.ac.uk. New database entries are assigned unique identifiers, which may be cited in publications. The RESID Database of Protein Modifications is copyrighted, and is distributed without charge and with no license required. Users should cite either the introductory announcement [10], the most recent descriptive article on the RESID Database in the database issue of *Nucleic Acids Research* [9], or this article.

The project for standardizing Swiss-Prot and TrEMBL, and merging PIR information annotations is supported by the National Institutes of Health National Human Genome Research Institute grant 1 U01 HG02712-01 for the UniProt Sequence and Function Knowledgebase, and by the Swiss Federal Government for maintenance of Swiss-Prot. Hosting and support for the RESID Database is kindly provided by the EBI, Hinxton, Cambridge, United Kingdom, and by the Advanced Biomedical Computing Center, SAIC, National Cancer Institute – Frederick, Frederick, MD, USA. The RESID Database of Protein Modifications is the service mark of John S. Garavelli.

5 References

[1] Degtyarenko, K., *Bioinformatics* 2000, *16*, 851–864.

[2] Jensen, L. J., Gupta, R., Blom, N., Devos, D. *et al.*, *J. Mol. Biol.* 2002, *319*, 1257–1265.

[3] Nakai, K., *J. Struct. Biol.* 2001, *134*, 103–116.

[4] Wilkins, M. R., Gasteiger, E., Gooley, A. A., Herbert, B. R. *et al.*, *J. Mol. Biol.* 1999, *289*, 645–657.

[5] Taylor, G. K., Kim, Y. B., Forbes, A. J., Meng, F. *et al.*, *Anal. Chem.* 2003, *75*, 4081–4086.

[6] Lin, D., Tabb, D. L., Yates III, J. R., *Biochim. Biophys. Acta* 2003, *1646*, 1–10.

[7] Mann, M., Jensen, O. N., *Nat. Biotechnol.* 2003, *21*, 255–261

[8] Meng, F., Cargile, B. J., Patrie, S. M., Johnson, J. R. *et al.*, *Anal. Chem.* 2002, *74*, 2923–2929.

[9] Garavelli, J. S., *Nucleic Acids Res.* 2003, *31*, 499–501.

[10] Garavelli, J. S., *Prot. Sci.* 1993, *2* (Suppl. 1), 133.

[11] *Gene Ontology Consortium, Genome Res.* 2001, *11*, 1425–1433.

[12] Zdobnov, E. M., Lopez, R., Apweiler, R., Etzold, T., *Bioinformatics* 2002, *18*, 1149–1150.

[13] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C. *et al.*, *Nucleic Acids Res.* 2003, *31*, 365–370.

[14] Wu, C. H., Yeh, L. S., Huang, H., Arminski, L. *et al.*, *Nucleic Acids Res.* 2003, *31*, 345–347.

[15] Srinivasan, G., James, C. M., Krzycki, J. A., *Science* 2002, *296*, 1459–1462.

[16] *IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN), Eur. J. Biochem.* 1984, *138*, 9–37.

[17] Degtyarenko, K. N., Contrino, S., *BMC Structural Biology* 2004, *4*, 3.

[18] Westbrook, J., Feng, Z., Chen, L., Yang, H., Berman, H. M., *Nucleic Acids Res.* 2003, *31*, 489–491.

[19] Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C. *et al.*, *Nucleic Acids Res.* 2004, *32*, D115–D119.

[20] Koch, T. H., Christy, M. R., Barkley, R. M., Sluski, R. *et al.*, *Methods Enzymol.* 1984, *107*, 563–575.

[21] Kowalak, J. A., Walsh, K. A., *Prot. Sci.* 1996, *5*, 1625–1632.