8 Annotation of Protein Sequences

W. C. Barker and C. H. Wu

Abstract
Protein sequence databases with accurate, comprehensive and up-to-date information on proteins are fundamental resources for a wide range of researchers, educators and students. Inaccurate annotations in these databases have led to erroneous conclusions and to propagation of errors into other databases and into the literature. The relatively few annotations in these databases that are based on direct experimental data are the most reliable and should be well identified, with citations to the sources of the information. Conversely, it must also be very clear which annotations are based, implicitly or explicitly, on sequence similarity or predictive algorithms. Such predictions are more reliable when applied within groups of closely related proteins in which one or more members have been experimentally characterized. UniProt, the Universal Protein Resource, strives to assist users to judge the quality of the data by clearly distinguishing experimental from predicted annotations, by using reliable and multiple prediction methods, by providing clear evidence attribution and by using standardized terminology. PIR protein family and superfamily classification serve as one basis for rule-based procedures that perform integrity checks and provide rich automatic functional annotation among homologous sequences.

Keywords
database annotation, protein sequence databases, UniProt, Protein Information Resource (PIR), protein family classification, protein superfamilies, rule-based annotation, automated annotation, annotation errors, evidence attribution
8.1 Introduction

With the recent accumulation of genome sequences of many organisms, notably including the human sequence, attention has turned to the identification of all of the different proteins in cells and to the comprehensive study of protein interactions, modifications and functions, defining the field of proteomics. In addition, structural genomics initiatives are generating new protein structures that will elucidate the mechanistic details of protein function. Inevitably, the experimental determination of protein functions and properties lags far behind the current avalanche of sequence data.

Identification of all of the genes encoding functional proteins in the genome of an organism is essential but not sufficient for understanding how these proteins function in making up a living cell. The number of different functional proteins in an organism often substantially exceeds the number of genes due to the generation of protein isoforms by alternative RNA processing as well as by covalent modifications of the precursor proteins. To cope with the complexity of protein sequence and functional information, annotated databases of protein sequence and function, with high interoperability to the multitude of related biomolecular databases, provide a cornerstone for a wide range of scientists active in modern biological research, especially in the field of proteomics. The major protein sequence databases are the most comprehensive sources of information on proteins. In addition to these universal databases that cover proteins from all species, there are collections storing information about specific families or groups of proteins, or about the proteins of a specific organism. Regardless of the scope and size of the database or whether it is intended for public, commercial or in-house use, the extent and quality of the annotation that accompanies the sequence data will be critical in determining the utility of the database for scientific investigations.
• Classifications
• Genetics
• Function and activities
• Secondary, tertiary and quaternary structure
• Covalent modifications and other position specific features
• Sequence variants
• Cross-references to other databases
• Keywords.

To make the data readable, searchable and retrievable, annotations are structured into labelled records and fields, which are presented in a database-defined order in each entry. They also usually follow a defined syntax and often employ a controlled vocabulary. Although data structure, syntax and semantics vary significantly among databases, there are ways for database designers to maximize interoperability with related databases. These include, for example, adopting widely used ontologies and classifications and providing PubMed cross-references for citations. Here we will discuss some practical aspects of protein sequence annotation, drawing examples
EMBL/GenBank nucleotide sequence database, sequences extracted from the literature and sequences submitted to Swiss-Prot. Together, PIR, EBI and SIB maintain and provide UniProt, a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces as the central hub for the collection of functional information on proteins (http://www.uniprot.org). The three groups continue to maintain web sites (http://pir.georgetown.edu, http://www.ebi.ac.uk and http://expasy.org) that provide many useful sequence analysis tools and databases.

8.4 Protein Family Classification

Classification of proteins provides valuable clues to structure, activity and metabolic role. Protein family classification has several advantages as a basic approach for large-scale genomic annotation: (i) it improves the identification of proteins that are difficult to characterize based on pairwise alignments; (ii) it assists database maintenance by promoting family-based propagation of annotation and making annotation errors apparent; (iii) it provides an effective means to retrieve relevant biological information from vast numbers of data and (iv) it reflects the underlying gene families, the analysis of which is essential for comparative genomics and phylogenetics.

A number of different classification systems, some highly automated and others curated, have been developed in recent years to organize proteins. Among the variety of classification schemes are (i) hierarchical families of proteins, such as the superfamilies/families (Barker, Pfeiffer and George, 1996) in the PIR-PSD, and protein groups in ProtoNet (Sasson et al., 2003), (ii) families of protein domains, such as those in Pfam (Bateman et al., 2002) and ProDom (Corpet et al., 2000), (iii) sequence motifs or conserved regions, such as in PROSITE (Falquet et al., 2002) and PRINTS (Attwood et al., 2003), (iv) structural classes, such as in SCOP (Lo Conte et al., 2002) and CATH (Pearl et al., 2003) and (v) integrations of various
SMART (Letunic et al., 2002) and TIGRFAMs (Haft et al., 2001), collaborate to produce InterPro (Mulder et al., 2003), an integrated documentation resource for protein families, domains and functional sites. Each InterPro entry contains signatures from one or more of the member databases describing the same group of proteins. Each entry includes a unique name and short name; an abstract, which provides annotation about the proteins matching the entry; literature references and links back to the relevant member database(s) and a list of precomputed matches against the whole of Swiss-Prot and TrEMBL. PIR protein superfamilies have been included in InterPro starting in 2003.

The InterPro database is maintained at the EBI with close collaboration with the member databases. It is updated regularly to coincide with new releases of the member databases. All data are stored in a relational database and direct web access via Java servlets is provided. The InterPro database is also distributed as flat files in XML format via FTP. On the web server (http://www.ebi.ac.uk/interpro/) the data are available for text or sequence searches. The sequence search package, InterProScan (Zdobnov and Apweiler, 2001), combines the search methods from each of the databases into a single package and provides an output with all results in a single format. In this way, researchers can submit their sequences using a web interface and obtain results of hits in InterPro in both a graphical and tabular view.

8.6 PIR Protein Families and Superfamilies

PIR defines 'closely related' proteins as having at least 50 per cent sequence identity; such sequences are automatically assigned to the same family. The families produced by automatic clustering can be refined during curation to produce groups that make biological sense, for example to include somewhat more distantly related members that are clearly orthologous and functionally equivalent. A PIR superfamily is a
among multidomain proteins where functional differences are associated with the presence or absence of one or more domains.

Family and superfamily classification frequently allow identification or probable function assignment for uncharacterized ('hypothetical') sequences. To assure correct functional assignments, protein identifications must be based on both global (whole protein, e.g. PIR superfamily) and local (domain and motif) sequence similarities.

8.7 Ontologies

The use of non-standardized vocabularies to name proteins, genes or organisms or to describe protein function can hinder searching across multiple proteins and species in different databases for common characteristics. Compiling a dictionary or thesaurus of biological terms is a major project that is most effectively done by a consortium of interested and expert parties, with the visibility needed to assure widespread adoption. In order to maximize interoperability and take advantage of the work of others, databases are well advised to minimize the use of in-house controlled vocabularies and instead rely on established external ontologies. Useful examples of these are the Enzyme Commission categorization, naming and description of enzyme functions (http://www.chem.qmul.ac.uk/iubmb/enzyme/), the taxons and their hierarchical arrangement from the NCBI Taxonomy Browser (Wheeler et al., 2003), the RESID
8.8 Protein Names, Source Information and Unique Identifiers

Databases typically assign a name to each protein. These names are characterized by their great variety. Like people, proteins can have the equivalent of nicknames as well as official or formal or full names. The same protein can be called by different names. It is extremely common to have a plethora of variations of spelling, capitalization, punctuation and spacing, especially for the nicknames. Some protein names are single words: trypsin, myosin, tropomyosin, insulin, haemoglobin, collagen. Even these, however, are more properly understood to refer to a fairly specific class or type of protein that may be further differentiated by additional modifiers. More often protein names contain several terms and mix uppercase and lowercase letters, numerical figures and non-alphabetical characters. Common examples are enzyme names, well established or ad hoc abbreviations (the equivalent of nicknames), gene symbols and arbitrary designations. These names can include simple protein names, nicknames, common English words (even including ‘and’ and ‘of’), words that describe some general or specific property or activity of the protein and indications of the source of the protein (organism, tissue, organelle). Curated databases typically impose some conventions of syntax and semantics and provide a list of other names (including misnomers as well as synonyms) that have been used so that users can search using terminology with which they are familiar.
recording the original identifications as submitted by the sequencers unless a revision is submitted by the same group. Therefore, the protein identifications in GenPept, which are taken directly from GenBank annotations, may never be updated in light of more recent knowledge. In many databases, it may not be evident when identification is based on solid experimental evidence, inference from properties of a closely related protein or simply the best partial match in a database search.

Over-identification is a common error when similarity is not strong over the entire lengths of the query and target sequences. Sequences in different families in the same superfamily may have different activities, though often falling within the same general class. For example, the long chain alcohol dehydrogenase superfamily contains alcohol dehydrogenase (EC 1.1.1.1), L-threonine 3-dehydrogenase (EC 1.1.1.103), L-iditol 2-dehydrogenase (EC 1.1.1.14), D-xylulose reductase (EC 1.1.1.9), galactitol-1-phosphate 5-dehydrogenase (EC 1.1.1.251) and others. Of five sequences from the recently sequenced genome of Brucella melitensis (DelVecchio et al., 2002) that were identified specifically as alcohol dehydrogenase (EC 1.1.1.1), only two are closely related (60 per cent identity) to well characterized alcohol dehydrogenases. For the others, the functional assignment may be overly specific, as they are more distantly related (less than 40 per cent identity). For the most part, users will need to inspect database entries and read at least the abstracts of published reports to ascertain whether a functional assignment is based on experimental evidence or only on sequence similarity. Users should also ascertain that any residues critical for the ascribed activity (e.g. active site residues) are conserved.

In many cases a more thorough and time-consuming analysis is needed to reveal the most probable functional assignments. Factors that may be relevant, in addition to presence or absence of domains, motifs or functional residues, include similarity or potential similarity of three-dimensional structures (when known), proximity of genes (may indicate that their products are involved in the same pathway), metabolic capacities of the organisms and evolutionary history of the protein as deduced from aligned sequences. Bork and Koonin (1998) discuss additional effective strategies. Iyer et al. (2001) analyse several additional examples of misidentifications and their subsequent correction.

8.10 Evidence Attribution

To prevent propagation of annotation errors and improve information content, protein annotations ideally should include, for both experimental and computational data, the types of evidence and methods for the annotation along with attribution of their sources. Evidence attribution is of growing importance since the comprehensive biomolecular databases combine data from a broad variety of sources. Data in TrEMBL, for example, may be automatically imported from the underlying DDBJ/EMBL/GenBank coding sequences, imported from other databases, partially curated
by EBI staff, generated from specific programs, or added by automatic annotation systems. Although every effort is made to ensure correct and consistent data, data quality is limited by the quality of the input data. Currently, it is difficult for database users to recognize where individual data items come from and whether they are well founded, reasonably inferred or highly speculative. Too often users assume that all data in the major databases are accurate and authoritative.

Attribution of protein annotations to validated experimental sources provides effective means to avoid propagation of errors that may have resulted from large scale genome annotation. This is already possible, in part, both in Swiss-Prot/TrEMBL (http://expasy.org/cgi-bin/lists?annbioch.txt) (Junker, Apweiler and Bairoch, 1999) and in PIR, through the use of a number of qualifiers or status tags which differentiate between experimental and non-experimental data. To distinguish experimentally verified from computationally predicted data, PIR entries are labelled with status tags of 'validated', 'similarity' or 'imported' in protein title, function and complex annotations (Figure 8.1(a)). The validated function or complex annotation includes hypertext-linked PubMed unique identifiers for the articles in which the experimental determinations are reported. The entries are also tagged with 'experimental', 'absent', 'atypical' or 'predicted' in feature annotations (Figure 8.1(b)). The first two tags are used to indicate the experimentally determined
features. Subsequently, the filtered papers are manually curated and added to the feature lines as literature attributions.

Evidence tags will be added to UniProt records by computer programs and during the literature-based curation process, allowing users to view the source of all data items in each record and to distinguish between experimental and computationally derived data. During the annotation process, curators will assess any information that has been added by a program, and if it is correct they will confirm it using the appropriate evidence tag. This will increase the reliability of all program-added information and will also allow for improvements to programs through feedback from curators to programmers. It will also prevent the overwriting by a program of any data that have been edited by a curator so that it will be possible to use programs to add information to a curated entry without touching manually curated data items. As each piece of data may have more than one evidence tag, this system is appropriate for data items that are derived from multiple sources. This system has been used internally at EBI for some time and a large number of data items in the TrEMBL database have already been tagged. The UniProt consortium will build on this method and extend it to allow tagging of known false information, e.g. if a keyword generated by an automated annotation system is known to be wrong by curator judgment. This will allow the incorporation of feedback from users and curators to improve the rules for automatic annotation. All data items tagged with a ‘negative’ evidence tag will be removed from the entry before publication.

8.11 Position-Specific Annotations

Position specific annotations appear in the ‘feature table’ sections of database entries and are of several types, including amino acids that are post-translationally modified, unusual encoded amino acids (selenocysteine being the most common), active or inhibitory sites and binding sites. These are annotated with reference to a specific residue (position) or to a short list of residues (e.g. 34, 52). Properly speaking, a ‘site’ should not be designated by a range (e.g. 32–38) of residues.

An important class of post-translational modifications comprises those amino acids that are chemically changed in such a way that they could not be restored by physiological processes of hydrolysis, ammonolysis or simple reduction, including chemical changes involving the alpha amino group (e.g. N-formylmethionine), or the alpha-carboxyl group of the carboxyl terminal residue, and selenocysteine and other rare amino acids that are translationally incorporated but for historical reasons are represented as modified residues. Active site residues are those known or thought to function in the actual catalytic reaction of an enzyme. Similarly, inhibitory site residues are those residues in enzyme inhibitors that serve as pseudo-substrates to block enzymatic activity.

Binding sites may be either covalent or non-covalent. Covalent changes may occur transiently or more or less permanently, but the original amino acid could in principle
Table 8.1. Classification-driven and rule-based approach for automated and quality annotation
'glycosyltransferase' and 'hexosyltransferase' are added (Action 5). Superfamily
members from animals are tested for the phosphorylase kinase phosphorylation site
and the appropriate feature is added (Action 3). The two new features trigger the
addition of keywords 'pyridoxal phosphate' (Action 6), and 'phosphoprotein' and
'allosteric regulation' (Action 7), respectively.

Family specific patterns for such features are derived from alignments of closely
related sequences for which some of the sequences have experimentally determined
properties. The rule may further specify other topological constraints for the pattern,
such as restricting the annotation of the P-loop feature to the ABC transporter domain
regions for the excinuclease ABC chain A superfamily. Looking for expected active
site and binding site sequence motifs and disulfide bonds only by homology within
the family or superfamily prevents errors such as annotation of signal sequences for
nuclear proteins, pyridoxal sites, internally in sequences, phosphorylation sites.
Acknowledgements

The PIR is supported by grant U01 HG02712 from the National Institutes of Health. The iProClass database is supported by DBI-9974855 and DBI-0138188 from the National Science Foundation.

References


