Databases and Ontologies

The Protein Feature Ontology: A Tool for the Unification of Protein Feature Annotations

Gabrielle A. Reeves1*, Karen Eilbeck2, Michele Magrane1, Claire O’Donovan1, Luisa Montecchi-Palazzi1, Midori A. Harris1, Sandra Orchard1, Rafael C. Jimenez3, Andreas Prlic4, Tim J.P. Hubbard4, Henning Hermjakob1 and Janet M. Thornton1

1EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom.
2Eccles Institute of Human Genetics, 15 North, 2030 East, University of Utah, Salt Lake City, 84112
3Bioinformatics and Genomics Department, Centro de Investigacion Principe Felipe (CIPF), Valencia, Spain
4The Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK.

ABSTRACT

Motivation: The advent of sequencing and structural genomics projects has provided a dramatic boost in the number of uncharacterised protein structures and sequences. Consequently, many computational tools have been developed to help elucidate protein function. However such services are spread throughout the world, often with standalone web pages. Integration of these methods is needed and so far this has not been possible as there was no common vocabulary available that could be used as a standard language.

Results: The Protein Feature Ontology has been developed to provide a structured controlled vocabulary for features on a protein sequence or structure and comprises approximately 100 positional terms, now integrated into the Sequence Ontology (SO) and 40 non-positional terms which describe features relating to the whole protein sequence. In addition, post-translational modifications are described by using an existing ontology, the Protein Modification Ontology (MOD). This ontology is being used to integrate over 150 distinct annotations provided by the BioSapiens Network of Excellence, a consortium comprising 19 partner sites in Europe.

Availability: The protein feature ontology can be browsed by accessing the ontology lookup service at the European Bioinformatics Institute (http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=BS).

1 INTRODUCTION

Genome sequencing has elucidated the locations of many genes on more than 700 genomes (Liolios, et al., 2008). Understanding human variation and disease requires knowledge of the role of each amino acid in a protein and how mutations or alternative splicing events can change function and phenotype.

1.1 Prediction and Computational Methods

The development of tools for comparisons between structure and sequence has increased rapidly, since these automatic methods are crucial in order to fill in the functional space between characterised and uncharacterised protein sequences and structures. Any knowledge we have of a protein is attached to the sequence and structure through annotation. Tools for the annotation of these data by automatic methods or the manual annotation of experimental data from the literature has become increasingly important (Reeves and Thornton, 2006). Many computational biology laboratories specialise in different aspects of proteome annotation for a range of features and processes: modifications (phosphorylation, lipidation etc), secondary structure prediction methods, fold recognition, effects of alternative splicing, single nucleotide polymorphisms (SNPs), domain and function assignment for catalytic residues, metal and ligand binding residues, protein-protein interactions and protein-DNA interfaces.

1.2 Challenges in Accessing Annotations

We now have a host of computational methods and tools available to help us learn more about sequences and structures. However these tools and databases are spread throughout the world and are numerous, often with more than one method annotating a similar feature. The potential user finds it hard to query them all and compare the results across different methods. At best, the user must traverse multiple websites, using a click and drag approach. The nature of bioinformatics also means that these tools can change rapidly as the software is developed to include cutting edge research findings and favoured web servers may even change location. More knowledge can be gained by combining and comparing...
annotations from these many sources. This inherently relies on the organisation and presentation of the data displayed, and consistency is crucial. An effective ‘first step’ for dealing with such a problem is to develop an ontology of protein features. An ontology is a standardised set of structured and precisely defined terms and relationships, providing a platform for both manual and automated reasoning in a dynamic environment so that changes can occur as different uses arise and new terms added. From this, annotations which adopt the ontology can be integrated into a single location, aiding comparisons between them.

The Gene Ontology resource (The Gene Ontology Consortium, 2008; Ashburner, et al., 2000) has been developed to describe gene and gene product attributes. This collaborative resource was set up to provide consistent descriptions of gene products that can be used by different databases annotating different species. There are three categories; biological process describing the biological objective of the gene or the gene product, molecular function, describing the biochemical activity and cellular component referring to the place in the cell where the gene product is active. The Sequence Ontology (Eilbeck, et al., 2005) has been developed to facilitate exchange, analysis and management of genomic annotation data. This standard has been used to underpin the features stored in the sequence databases of model organisms (Mungall and Emmert, 2007) and to standardize the annotation exchange formats (www.sequenceontology.org/gff3.shtml). It is used by many of the model organism communities to annotate their sequence features, such as FlyBase (Grumbling and Strelets, 2006), WormBase (Rogers, et al., 2008), DictyBase (Chisholm, et al., 2006) and SGD (Christie, et al., 2004).

Ontologies have also been created for other aspects of biology such as the MOD ontology for post-translational modifications (L. Montecchi-Palazzi, 2008), PSI-MI for molecular interactions (Kerrien, et al., 2007) and the Pathway ontology (Twigger, et al., 2007) as well as a number of specialist database ontologies (Avraham, et al., 2008; Drysdale and Crosby, 2005; Grumbling and Strelets, 2006; Rhee, et al., 2006; Sprague, et al., 2008). There is also an initiative to create an all-encompassing ‘protein ontology’ by the Protein Ontology (PRO) Consortium (Natale, et al., 2007). This ontology will model all aspects of proteins from their evolution to their form and function. This project is still in its infancy and so far provides an ontological description (from bottom up) of protein modifications, sequence forms (including alternative splicing, mutant forms, cleaved and post-translationally modified products), the whole protein unit, detected sequence domains, structural domains, and the evolutionary unit. This top level ontology is created by linking many resources which already exist, including the Gene Ontology, the protein modification ontology PSI-MOD, and the human disease ontology.

The Protein Feature Ontology was created to facilitate comparison of protein annotations. In contrast to the PRO, the Protein Feature Ontology comprises two types of terms: Those related to regions on the protein such as residues which form an alpha helix or are linked together by a disulphide bond, the positional terms. These terms are commonly used need to be standardised so that similar terms can be identified and compared. Such terms did not exist before this ontology was created and are now being used to cover this aspect within the PRO. In addition to the positional, the non positional terms describe aspects of a protein as a whole, comprising terms and descriptions from the UniProtKB comment lines and keywords. Ontological relationships between these particular terms did not exist before. In its initial application within the BioSapiens NoE inconsistencies such as different spellings, casings and synonymous names were rife. For example, a predicted transmembrane segment of peptide is annotated as both TRANSMEM (http://phobius.blnf.ku.dk/) and Membrane (http://www.cbs.dtu.dk/services/) by different servers. A domain annotation is traditionally indicated by the name of the method which provided it, for example, InterPro (Mulder and Apweiler, 2007) provides domain annotations such as SMART, ProDom, SCOP rather than the annotation domain. These annotations clearly make sense in their own contexts but when annotations are brought together from a number of sources, a more uniform approach needs to be adopted to provide effective comparison.

Such a project of integration has been undertaken by the BioSapiens Network of Excellence (The BioSapiens Network of Excellence, 2005), a consortium comprising 19 participating partners from bioinformatics laboratories in 14 European countries. A main goal of their work is to bring together annotations created by their in-house methods and algorithms in order to create a European ‘virtual institute of annotations’. These annotations are derived from methods of manual annotation and informatics tools, resulting in a set of protein/nucleic acid sequence and structural annotations from some of the leading bioinformatics laboratories in the world. Access to this information has been achieved technically through the implementation of a distributed annotation system (DAS (Dowell, et al., 2001)). This system comprises both a central reference server (in the case of protein annotations serving UniProt Knowledgebase (UniProtKB) sequences) and individual annotation servers which provide the annotations for sequences held in the central reference server. This information is then interpreted by a DAS client which reads the sequence and the annotations and displays the information in a human readable format. This allows the annotations from each partner site to remain under the control of the partner but to be co-ordinated and instantly brought together at a central point. A number of clients have been created including; the Ensembl genome browser (Spudich, et al., 2007) for both genomic and proteomic annotations, Dasy2 (Jimenez, R. personal communication) for protein sequence annotations, the Pfam DAS alignment viewer (Finn, et al., 2008) and Spice (Prlic, et al., 2005) for both protein sequence and structural annotations. The beauty of this method is that individual sites control their own DAS sources and therefore it is open to all to participate regardless of location or agreement. At present, the BioSapiens Network of Excellence has collected 40 different distributed annotation sources for protein sequence and structure providing over 150 annotation types. This method of collecting and centralising the display of annotations from disparate laboratories across Europe is unprecedented and as such, provides a unique data source as well as a central platform for participating laboratories to display their data. However, with the control of the information lying with each partner site, a lack of consistency in the annotation terms provided by each source has evolved. The adoption of a controlled vocabulary of protein features would not only allow ‘like’ annotations from different servers
to be identified and viewed together on the clients, but also enable complex manipulation of these data both manually and automatically, deriving relationships between annotation methods and enabling the creation of intra-method consensus tracks.

2 METHODS

The Protein Feature Ontology is a set of terms which describe the features which make up protein function and form (Fig. 1.), from features describing the local structure such as residues involved in disulfide bonds, helices, strands and motifs (such as the helix-turn-helix motif) to overall tertiary structure marking a globular domain. Functional residues such as those which are important for signalling or catalysis can be annotated as well as those in contact with ligands or involved in protein interactions. Figure 1 shows selected DAS tracks from the BioSapiens NoE annotating features on the alpha and beta subunits of the insulin receptor and illustrates how this ontology can be used to integrate annotations from different methods. The origin of each annotation is listed under the heading “server” showing the variety of information sources. This diagram illustrates how the ontology has enabled the integration of these different information sources by the standardization of the language they use. The protein feature ontology serves to provide a uniform description of the features and properties of a protein. In addition an ontology provides not only a set of terms and definitions but also describes relationships between these terms. This not only allows these terms to be browsed and located easily by eye but also allows these terms to be automatically linked. In the context of this particular application of the ontology once annotations are displayed on the DAS client the terms can be grouped, for example, in Figure 1 annotations which illustrate membrane structure (cytoplasmic, non-cytoplasmic, extramembrane and transmembrane regions) are able to be clustered together. In addition to this, in this particular application of the ontology, each track is provided with an evidence code (http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=ECO) to display provenance (also shown on Fig. 1). The implementation of the protein feature ontology has been fully supported by the DAS registry (Prlic, et al., 2007) which provides the facility for users to validate server output in order to make sure it is fully ontology compliant.

The ontology is divided into two parts: Non-positional terms which refer to the whole protein sequence or structure and Positional terms which refer to a specific residue or range of residues in the protein. For positional terms, features are located using sequence residue numbers and the properties of these features describe an attribute of the feature, for example, residues 5-130 are described as a domain. Figure 1 provides an illustration of these types of terms. Other terms are not associated with a particular region of the sequence (non-positional), but instead provide a description of the properties of the whole protein such as an associated publication or links to related sources of information such as GO term annotation or EC annotation. The protein feature ontology currently comprises approximately 140 terms: 100 positional terms and 40 non-positional terms.

3 RESULTS

3.1 Positional Annotations

Positional terms (Fig. 2.) describe features that can be “associated” with a particular region in the peptide. These annotations are derived from programs or methods which detect features on the protein sequence or structure. Examples of such features include the specific role of an amino acid residue such as its catalytic activity, involvement in the binding of a metal ion, an indication of the location within the cell such as intramembrane, or the structural conformation of the residue range such as the alpha-helix. These terms fall within the scope of the Sequence Ontology (SO) and as a result, these terms have been integrated into the SO. The Protein Feature Ontology was created in collaboration with UniProtKB and the UniProtKB Feature Types were used as the starting point. All UniProtKB feature types exist in the ontology, but in order to fit in with the Sequence Ontology naming schemes, some term names have been modified. For example, the word polypeptide has been added to ontology terms to disambiguate them from more general terms in SO, which includes a wider selection of features. Motif becomes polypeptide motif. A mapping between the ontology and the UniProtKB feature types is being maintained so that it is possible to automatically map between them. The mapping is also maintained in the SO synonyms.

The 3 sections in SO that are populated by positional protein feature terms are:

(1) **Polypeptide region**: Biological sequence region that can be assigned to a specific subsequence of a polypeptide. Within this category lie a number of terms:

- **Biochemical region**, amino acids involved in binding, interactions, catalysis or peptide bonds (which are represented by the positions of the two flanking amino acids).
• Polypeptide_domain, describing a structurally or functionally defined protein region which has been shown to recur throughout evolution. In order to distinguish further, the term polypeptide_domain has been further categorised into three child terms. Two UniProtKB feature types have been classified here: polypeptide_motif indicating a short (up to 20 amino acids) region which is conserved in different proteins and polypeptide_repeat which indicates internal sequence repetition. In addition, the term polypeptide_structural_domain has also been created, allowing the difference between a structural domain (a structure which is self-stabilising and folds independently from the rest of the protein chain) and the parent term (polypeptide_domain) to be distinguished. This allows for the term polypeptide_structural_domain to also exist within the structural_region branch of the ontology so that annotations can be clustered and potentially viewed with secondary structural and membrane structure features.

• Also within this category is the term immature_peptide_region - the extent of the peptide after it has been translated and before any processing occurs. This is then divided into mature_protein_region, the extent of a polypeptide chain in the mature protein, and cleaved_peptide_region for regions which are cleaved during maturation (including signal_peptide and transit_peptide).

• Structural_region describes the backbone conformation of the polypeptide and includes child terms to describe both secondary structure and the structure of the protein in the membrane.

• Polypeptide_variation_site, indicates alternative sequences due to naturally occurring events such as polymorphisms and alternative splicing or experimental methods such as site-directed mutagenesis.

3.2 Non-positional annotations
This section classifies annotations which do not refer directly to a particular feature on the protein sequence or structure but instead refer to the full length of the protein. They typically describe those attributes which would be included in a database entry. For example, a number of methods provide a GO term as output. Each partner would classify their output with the ontology term GO_annotation, allowing the client to cluster these together in the table. Terms within this non-positional category are mainly derived from categories within a UniProtKB entry. Two main areas are currently covered, Uniprot comments (CC field), keyword categories (KW field) with additional entries to describe taxonomy and publication.

3.3 Format/Rules/Naming-conventions/Relationships
Where applicable, UniProtKB feature types have been used as term names. Naming conventions have been used in line with the Sequence Ontology regulations. Terms must be computer readable. Therefore, underscores are used instead of spaces, numbers spelt out and common abbreviations used. No full stops, points, slashes, hyphens or brackets are allowed but common abbreviations are used. All entries are in lower case except for common abbreviations and where there are differences, the US form of spelling is chosen. Synonyms aid ontology searching and there is no limit to the number of synonyms. Here, normal term rules do not apply so that, for example, common abbreviations are spelt out and English spellings can be stated.

The use of the controlled vocabulary allows annotations of the same feature to be viewed side by side, for example, residues in contact with metal identified from 3D data by PDBsum (Laskowski, 2007) will be viewed alongside metal contact residues identified by the UniProtKB curators, as both have the ontology term metal_contact. However, the structure of the ontology also provides information on how this term relates to other terms. With this information, the metal contact residues can be viewed alongside related annotations such as those which describe ligand_contact and catalytic_residues. The protein feature ontology includes two relationship types; ‘is_a’ (e.g. a helix is_a polypeptide_secondary_structure) and ‘part_of’, which indicates when a term forms only a portion of its parent: the extramembrane region is part_of the whole membrane_structure.

3.4 Additional terms to describe post-translational modifications

(2) Polypeptide_sequencing_information: This category clusters annotations which report incompatibility in the sequence due to some experimental uncertainty.

(3) No_output: which allows annotators to report where an analysis has been run and not produced any annotation.

All terms under the parent ‘positional’ are located primarily in SO. It is possible to view and use these terms within SO along with other SO terms or within the Protein Feature Ontology. In addition, the terms in SO can be automatically extracted to create a stand-alone ontology by filtering in OBO-Edit (Day-Richter, et al., 2007) for the category ‘biosapiens protein feature ontology’.
Also falling within the scope of the protein feature ontology are the terms and definitions describing post-translational modifications. An ontology describing this area has already been created (The Protein Modification (MOD) Ontology http://www.ebi.ac.uk/ontologylookup/browse.do?ontName=MOD) comprising approximately 1050 terms in more than 45 top-level nodes. The ontology describes alternative hierarchical paths for the classification of protein modifications. These paths describe either the molecular structure of the modification, for example, phosphorylated residue (MOD:00696, a protein modification that effectively substitutes a phosphoryl group for a hydrogen atom) or a description of the amino acid residue that is modified, for example, modified L-tyrosine residue (MOD:00919, a protein modification that modifies an L-tyrosine residue). These terms are inserted into the anchor term post translational modification (SO:0001089) in the composite Protein Feature Ontology. The MOD ontology contains many terms to describe the artefactual modifications made as part of the process of mass spectrometry. These have been retained in the merged ontologies to allow the accurate visualisation of data obtained using this technique.

Fig. 2. An overview of the child terms in the parent category polypeptide_region. This category contains both structural and functional feature annotations. A large number of tools and databases provide annotations which fall into this parent category and as a result will be able to be viewed and compared together. Dark grey illustrates the top level term, light grey, the children of polypeptide_region and white, a summary of the terms within these categories.

3.5 Applications of the Ontology

One of the great strengths of the ontology can be seen where annotations from a number of different sources are brought together. This can be seen by looking at the functional residues on the 11β-HSD-1 which is a NADPH-dependent protein (PDB code P28845). Above, the DAS tracks for servers UniProt (inferred by curator), Pfam (from sequence similarity), the Catalytic Site Atlas (inferred from sequence similarity) and PROSITE (from sequence similarity). Below depicts the residues as they lie in the protein. Dark black ball and stick show the catalytic tetrad, dark grey, the substrate cortisone and co-factor NADPH and white are those residues annotated by PDBsum as being in contact with substrate and co-factor.

3.6 The Open Biomedical Ontologies Community

This project has been undertaken as part of the Open Biomedical Ontologies (OBO) (Smith, et al., 2007) and thus delivers the ontology in OBO format (http://www.geneontology.org/GO.format.obo-1_2.shtml). The protein feature ontology is a directed acyclic graph (DAG) which has been edited using the OBO-Edit ontology manager (Day-Richter, et al., 2007) and can be viewed using the Ontology Lookup Service at the EBI (http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=BS). As an opensource ontology, new terms are currently suggested and reviewed by BioSapiens members as well as the Sequence Ontology community. The current version, including BioSapiens terms, Sequence Ontology terms and Protein Modification terms can be downloaded from the ontology website (http://www.biosapiens.info/ontology). The BioSapiens terminology integrated into SO can also be downloaded from www.sequenceontology.org. Users should note that the BS identifier is given as an alternate ID. Suggestions and comments are
welcomed on the tracker http://www.ebi.ac.uk/seqdb/jira/secure/Dashboard.jspa and the SO term tracker http://sourceforge.net/tracker/?group_id=72703. The SO community also provides a mailing list for debate of new terminology at song_dev@lists.sourceforge.net.

4 CONCLUSIONS

We have created an ontology for protein features in order to facilitate integration of protein feature annotations provided by a growing number of methods from around the world. An ontology is a controlled vocabulary composed of types (terms with synonyms) and the relations that hold between them. This allows two things: firstly, a standardisation of the terms that are used, allowing ‘like’ annotations to be identified and, secondly, the relationships allow automatic inferences to be drawn between annotation types. Computer programs will know that the extramembrane region and the intramembrane region are both part of the membrane structure and in turn, they are all structural regions alongside polypeptide secondary structures such as helices and beta strands, allowing inferences on the exact structure and cellular location to be drawn. An initial use of the protein feature ontology is illustrated by the BioSapiens Network of Excellence. A major goal of the consortium is to provide a ‘virtual centre for annotation’ and as part of this, an ontology is needed on which to base the annotations. The implementation of this ontology has allowed the annotations collected by the BioSapiens partners to be clustered and manipulated to provide greater biological meaning. The creation of this resource provides biologists, biochemists and bioinformaticists with a united view of all available annotations; so that reliability of data/annotations can be better assessed. We have advertised the ontology by showing its use in one particular project, the integration of terms within the BioSapiens Network of Excellence however outside the scope of this project this ontology can provide a common language for any method or database providing protein feature annotations.

ACKNOWLEDGEMENTS

With grateful thanks for the input from Eugene Kulesha, Andy Jenkinson, all participants of the ontology workshop held in February 2007 and all participants of the BioSapiens Network of Excellence.

REFERENCES


G. Reeves et al.

REFERENCES

This work was completed as part of the BioSapiens Network of Excellence, funded by the European Commission within its FP6 Programme, under the thematic area ‘Life sciences, genomics and biotechnology for health,’ contract number LHSCT-2003-503265’.

ACKNOWLEDGEMENTS

With grateful thanks for the input from Eugene Kulesha, Andy Jenkinson, all participants of the ontology workshop held in February 2007 and all participants of the BioSapiens Network of Excellence.

503265'.
The Protein Feature Ontology: A Tool for the Unification of Protein Feature Annotations


