Proceedings of Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE2004), Taiwan, 19-21 May 2004, 419–427

Towards a Proteomics Meta-Classification

Anand Kumar
Institute for Formal Ontology and Medical Science, University of Leipzig, Germany.
Laboratory of Medical Informatics, University of Pavia, Italy.
E-mail: anand.kumar@ifomis.uni-leipzig.de

Barry Smith
Institute for Formal Ontology and Medical Science, University of Leipzig, Germany.
Department of Philosophy, University at Buffalo, USA.
E-mail: phismith@buffalo.edu

Abstract

There is a recognized need for a meta-classification that can serve as a foundation for more refined ontologies in the field of proteomics. Standard data sources classify proteins in terms of just one or two specific aspects. Thus SCOP (Structural Classification of Proteins) is described as classifying proteins on the basis of structural features; SWISS-PROT annotates proteins on the basis of their structure and of parameters like post-translational modifications. Such data sources are connected to each other by pairwise term-to-term mappings. However, there are obstacles which stand in the way of combining them together to form a robust meta-classification of the needed sort. We discuss some formal ontological principles which should be taken into account within the existing datasources in order to make such a meta-classification possible, taking into account also the Gene Ontology (GO) and its application to the annotation of proteins.

1. Introduction

There are a large number of existing databases and ontologies which classify proteins on the basis of their structure, function, location, evolution and so on. [1] When it comes to providing a complete description of a protein's function, however, a wide range of factors needs to be taken into account, including cellular roles, molecular functions and the involvement of proteins in physiology and pathology. [2] This creates a need for data integration between the different proteomics databases, which at the moment is primarily achieved via pairwise term-to-term mapping between existing databases. This approach presumes that the databases to be mapped have classified proteins on the basis of just one or a small number of aspects. Thus for example the Structural Classification of Proteins (SCOP) is described as providing a protein classification on the basis of protein structure [3,4]. Protein Data Bank (PDB) provides a protein classification based on 3-D macromolecular structure [5,6]. In Swiss-Prot the core data consists primarily of sequence information, with annotations describing factors such as functions, post-translational modifications, domains and sites, and diseases associated with deficiencies. [7,8] At a somewhat different level, the Gene Ontology (GO) provides a large ontology of cellular components, biological processes and molecular functions, and serves as a de facto standard for annotation of gene products, including proteins. [9,10] However, GO has problems with its structure and with many of its terms, [11] and a robust proteomics meta-classification will require that such problems be resolved, not only in GO but also as they arise within each of the other existing systems.

2. Structural Classification of Proteins

SCOP divides proteins into the following top-level classes: 46456: **All alpha proteins**; 48274: **All beta proteins**, 51349: **Alpha and beta proteins (a/b)**; 53931: **Alpha and beta proteins (a+b)**; 56572: **Multi-domain proteins (alpha and beta)**; 56835: **Membrane and cell surface proteins and peptides**; 56992 **Small proteins**; 57942 **Coiled coil proteins**; 58117: **Low resolution protein structures**; 58231: **Peptides**; and 58788: **Designed proteins**. The SCOP database provides a detailed and comprehensive description of the structural and evolutionary relationships between the proteins whose structure is known, including all entries in the PDB, combining together the classifications made on the basis of structure and evolution.

The major levels in the hierarchy of SCOP beneath **class** are:

Fold: "*Major structural similarity.* Proteins are defined as having a common fold if they have the same

major secondary structures in the same arrangement and with the same topological connections. Different proteins with the same fold often have peripheral elements of secondary structure and turn regions that differ in size and conformation. In some cases, these differing peripheral regions may comprise half the structure. Proteins placed together in the same fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins favoring certain packing arrangements and chain topologies."

Superfamily: "*Probable common evolutionary origin* Proteins that have low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable are placed together in superfamilies. For example, actin, the ATPase domain of the heat shock protein, and hexakinase together form a superfamily."

Family: "*Clear evolutionarily relationship:* Proteins clustered together into families are clearly evolutionarily related. Generally, this means that pairwise residue identities between the proteins are 30% and greater. However, in some cases similar functions and structures provide definitive evidence of common descent in the absence of high sequence identity; for example, many globins form a family though some members have sequence identities of only 15%."

Below family we have in addition protein, species, and domain.

To test our methods we studied the the different membrane protein axes within SCOP and found problems related to the **membrane and cell surface proteins and peptides** axis.

2.1. Problem of exclusion

The first problem is that SCOP classifies many membrane proteins not within the class **membrane and cell surface proteins and peptides**, but rather within other classes which comprehend protein groups classified on the basis of their structural features. Examples from the version of December 2003 (SCOP 1.65) are: 82055: **Peroxisomal membrane protein,** classified under 48724: **All-beta proteins** and 69907: **Membrane penetration protein mu1**, classified under 69907: **Multi-domain proteins (alpha and beta)**.

While these classifications may in themselves be correct, given the presence of a **membrane and cell surface proteins and peptides** class, they point to an incoherence in the underlying classificatory order. For they imply that we cannot infer, from: *x* is not a **membrane and cell surface protein**, to: *x* is not a membrane protein. When we move down the SCOP hierarchy through folds, superfamilies and families we find 219 problematic membrane protein designations at the level of folds, 140 and 111 further problematic designations at the level of superfamilies and families, making 470 problematic cases in all.

2.2. Structure vs. location

Membrane proteins constitute 36 of the 800 folds, 66 of the 95 families and 150 of the 2327 superfamilies present within SCOP. The problem, for a purportedly *structural* classification of proteins like SCOP, is that not all membrane proteins have a common structure. For example, while 931 of the proteins classified within the axis **Membrane and cell surface proteins and peptides** have transmembrane helices, 158 of them do not. Indeed the question arises whether a class like **Membrane and cell surface proteins and peptides** should be present at all within a properly structural classification, or whether it would not be better to separate out a class labeled **Proteins with transmembrane helices** and relocate the other proteins classified as **Membrane and cell surface proteins and peptides** (some 14.5%) elsewhere. For the latter is a class that is based not on protein structure but rather on protein location, and it is this mixing of two aspects within a single classification which leads to the problems mentioned above. Another argument for a reform along these lines is that SCOP does not contain protein groups like 'blood proteins' or 'neuronal proteins' and thus a class that depicts 'membrane proteins' should also not be included.

3. Gene Ontology and Its Annotations

GO is an important tool for the representation and processing of gene- and gene-product-related information. It provides a 'controlled vocabulary,' designed to ensure that researchers in biomedicine should use a common terminology in reporting the results of their work. Gene products have been annotated to GO and the December 2003 version of GO contains over 4 million such annotations. This gives GO a very important position with the field of proteomics meta-classification.

GO's considerable success is testimony to the correctness of a number of crucial choices which were made by the GO Consortium in the early stages of its development. Above all, the adoption of a relatively simple architecture meant that work on populating GO could start immediately. Populating GO does not

require the completion of complex protocols but can be done intuitively by the expert biologist, who is subject to few formal constraints when incorporating new terms.

However, the authors of the Gene Ontology have at the same time ignored certain benefits of formal rigor, both at the level of logical relations and at the level of syntactic regimentation, and both in the structure of GO's term hierarchies and in its definitions. The upshot is that there are aspects of the current architecture of the Gene Ontology that are predestined to cause ever more significant problems as GO increases in size. We have discussed these issues in detail elsewhere [11,20,32] and here we consider only those which are directly relevant to the issue of proteomics meta-classification.

The Gene Ontology contains three orthogonal axes, corresponding to the three root classes **Cellular component**, **Molecular function** and **Biological process**.

The cellular component ontology is GO's counterpart of anatomy. It consists of terms such as *flagellum*, *chromosome*, *ferritin*, *extracellular matrix* and *virion* and is intended to allow biologists to register the physical structure with which a gene or gene product is associated. It includes both the extracellular environment of cells and the cells themselves (that is, *cell* in GO is a subclass of *cellular component*).

GO's molecular function ontology deals with the actions characteristic of gene product *Molecular function* accordingly subsumes terms describing actions, for example *ice nucleation*, *binding*, or *protein stabilization*, entities which do not *endure* but rather *occur*.

GO's biological process ontology comprehends Biological process terms can be quite specific (*glycolysis*) or very general (*death*).

3.1 Problems with the tripartite classification

Molecular function and biological process terms are clearly closely interrelated. The process of anti-apoptosis, for example, certainly involves the molecular function now labeled apoptosis inhibitor activity. But GO's curators give us too little information is to what this relationship might be, suggesting only that it might be a matter of differing granularity: 'A biological process is accomplished via one or more ordered assemblies of molecular functions.' This would suggest that molecular functions are constituents of biological processes, but this in turn would suggest also that they stand to such processes in a part-of relation. At the same time, however, GO's authors insist that the relation part-of holds only within a single ontology and never between entities from distinct vocabulary sets.

This leads to difficulties in automated interpretation of the gene product annotations. Thus we have been working on disease-related gene products focusing on annotations to GO of the genes present within

Locus_id	GO_id	GO_term	Term_type
1356	GO:0006811	ion transport	biological_process
1356	GO:0006825	copper ion transport	biological_process
1356	GO:0006878	copper ion homeostasis	biological_process
1356	GO:0006879	iron ion homeostasis	biological_process
1356	GO:0005507	copper ion binding	molecular_function
1356	GO:0016491	oxidoreductase activity	molecular_function
1356	GO:0004322	ferroxidase activity	molecular_function
1356	GO:0005375	copper ion transporter activity	molecular_function
1356	GO:0005615	extracellular space	cellular_component

Table 1. GO annotations for LocusID 1356: ceruloplasmin (ferroxidase)

LocusLink. [13] We encounter problems in the interpretation of such annotations if we proceed on the basis of the assumption that GO's molecular function and biological process axes are orthogonal. The LocusID 1356: **ceruloplasmin (ferroxidase)** annotations, for example, are listed in our database for clinical genomics models as shown in Table 1. Accepting GO's assumption of orthogonality in this case would mean that there is no relationship between

GO 006825**: copper ion transport** and GO 0005375: **copper ion transporter activity**

or between:

GO 006878: **copper ion homeostasis** and GO 0005507: **copper ion binding.**

Thus if automated deductions are made on the basis of the orthogonality principle then this leads to loss of information. (For a complete list of such annotations related to loci annotated for colon cancer and colorectal cancer, see [14,15].)

3.2 Problems with undefined terms

Many GO terms are compositional in nature. [16] That is, they are built out of other terms as proper parts. The latter however are not always GO terms in their own right, and many of them are not treated, either, within the framework of other large terminology systems such as the Unified Medical Language System [17]. Thus, terms like 'adult' or 'evasion', as well as operators like 'sensu"', ':' and '/' etc., which have not been defined explicitly in the GO documentation, making the interpretation of terms which contain the corresponding expressions difficult and leading to knowledge which is inaccessible to automatic tools for information retrieval. One particularly interesting example is the use of the '/' operator, which is used within GO to represent

- '**and**', as in:

GO 0008608: **microtubule/kinetochore interaction**

defined as: Physical interaction between microtubules and chromatin via proteins making up the kinetochore complex;

- '**or**', as in:

GO 0001539: **ciliary/flagellar motility**

defined as: Locomotion due to movement of cilia or flagella;

- '**and/or**', as in:

GO:0045798 **negative regulation of chromatin assembly/disassembly**

defined as: Any process that stops, prevents or reduces the rate of chromatin assembly and/or disassembly;

- '**reaction substrates and products**', as in:

GO:0015539 **hexuronate (glucuronate / galacturonate) porter activity**

defined as: Catalysis of the reaction: hexuronate(out) + cation(out) = hexuronate(in) + cation(in);

- '**successive reaction steps**', as in:

GO 0000082: **G1/S transition of mitotic cell cycle**

defined as: Progression from G1 phase to S phase of the standard mitotic cell cycle; and

- '**ratio**', as in:

GO:0001559 **interpretation of nuclear/cytoplasmic to regulate cell growth**

For a complete list of GO's undefined terms and operators, see [18]. This lack of definitions and of rules for use of syntactic operators in the creation of compound terms leads to problems when GO is used for annotations of gene products, since such annotations need to be carried out by human experts who need to understand the meanings of the terms they are using. If we put together all such terms which do not have well-defined words or operators, they make 2323 out of 16660 terms within the December 2003 version of GO (13.94%). There are currently 263,632 annotations present within GO involving these terms, which signifies a fraction of 6.28% of the total 4,197,659 annotations. This is less than half of what one would expect if these terms were used at the same rate as other terms in GO, which lends support to our thesis that these terms not only cause difficulties for automatic tools but are also understood less well by human annotators.

3.3 Problems related to part-of relations

GO is structured by means of two relations, called 'is-a' and 'part-of'. We here concentrate on problems with the latter. On the one hand this must sometimes be interpreted as signifying 'can-be-part-of', meaning that *A* part-of *B* does not signify that an *A* term is a part of a *B* term in every case, as reflected in propositions such as:

GO 0005833: **hemoglobin complex** part-of GO 0005829: **cytosol**

GO 0005829: **cytosol** part-of GO 0005737: **cytoplasm**

GO 0005737: **cytoplasm** part-of GO 0005622: **intracellular**

GO 0005634: **nucleus** part-of GO 0005622: **cell**

GO's new definition of 'part-of', however, reads in the terminology proposed in [19,20] as follows:

A part-of *B* =: all instances of *A* are necessarily parts of instances of *B*. [21]

According to this definition, hemoglobin complex would necessarily be part of all cells, which is of course not the case. GO's inconsistent treatment of 'part-of' is reflected in many errors in annotations. For example, 21 gene products are annotated to GO 0005833: **hemoglobin complex**. These annotations include various chains which constitute different hemoglobin forms, for example hemoglobin alpha and beta chains. These chains, are themselves annotated for various hemoglobinopathies on LocusLink. By reasoning from GO's treatment of hemoglobin on the basis of the new definition of part-of deduction, such hemoglobinopathies can be inferred to affect every cell, which would be a major error.

Such errors within the cellular component axis are common. There are 36 subparts of the term GO 0005622: **intracellular** none of which is present within all cells. One can imagine the cumulative effect of such 'small errors' if such relations are used within large database integrations such as those involving LocusLink.

Some of these relations have been better dealt with in other anatomical ontologies, one of the more important examples being the Foundational Model of Anatomy [22]. For example, GO has the following relation:

GO 0005634: **nucleus** is-part-of GO 0005622: **intracellular**

which is dealt with in FMA as follows:

FMA 155394: **cell nucleus** is-part-of FMA 155292: **protoplasm**

FMA 155292: **protoplasm** is-part-of FMA 155394: **nucleated cell**

FMA 154635: **cytoplasm** is-part-of FMA 164388: **non-nucleated cell**

Drawing differentiations of this sort adds clarity to the treatment of the relation in qeustion, which in turn contributes to the correctness of subsequent deductions. Moreover, the terms which are represented as standing in part-of relations to other terms in GO usually do not have any is-a relations associated with them. FMA provides separate is-a and part-of relations for the entities within its scope. This brings further clarity to the FMA ontology, as illustrated for example by:

FMA 155394: **cell nucleus** is-a FMA 146889: **organelle,**

which means that cell nucleus is assigned in FMA to a type which is not recognized in GO by means of a separate term.

3.4 Problems with 'sensu'

A related set of problems can be illustrated by examining GO's use of its 'sensu' operator, which is introduced to cope with those cases where a word or phrase has different meanings when applied to different organisms. Consider for example the case of GO 0005618: **cell wall**, whose subclasses include GO 0009274: **cell wall (sensu bacteria)** and GO 0009277: **cell wall (sensu Fungi)**, the latter being introduced in reflection of the fact that cell walls in bacteria and fungi have a completely different composition from cell walls in most other types of organisms. As the GO documentation has it: 'Using the sensu reference makes the node available to other species that use the same process/function/component'. [23] The problem is however that, if 'sensu' is indeed designed to indicate that the modified term refers to a *different* class from that to which the unmodified term refers, then in what sense are we still dealing with 'the same process/function/component'?

Since the primary goal of the GO Consortium is to provide an ontology of gene products applicable to all species, they insist that sensu terms be introduced sparingly. In consequence, sensu terms are allowed to have non-sensu terms as children, as in

GO 0000326: **protein storage vacuole** is-a GO 0000325: **vacuole (sensu Streptophyta)**

This, however, is to imply that protein storage vacuoles occur only in Streptophyta, which is to ignore for example the existence of fungal protein storage vacuoles. (This case has been reported as an error to GO's SourceForge tracker.)

In all 79 out of 469 (16.84%) of GO *sensu* terms are subject to errors of this kind. A particularly intriguing example, which also illustrates GO's inconsistent handling of the relation of localization, is GO's postulation of:

GO 0005934: **bud tip** is-a GO 0000134: **site of polarized growth (sensu Saccharomyces)**

What this means is that every instance of bud tip in every organism has an instance of Saccharomyces polarized growth located therein.

Another problematic example pertains to GO 0045500: **R7 differentiation**, for which GO asserts:

GO 454666: **R7 differentiation** is a GO 0001751: **eye photoreceptor differentiation (sensu Drosophilia)**

For again, there is R7 differentiation in species other than Drosophilia, for example in crustaceans.

A further problem is caused by GO's use of 'sensu Invertebrata'. Whereas vertebrate is a well-defined biological taxon, biologists tend to disagree on what the definition of invertebrate should be, and thus apply the 'sensu Invertebrata' modifier to different taxa. The resultant errors are illustrated for example in the genes annotated to

> GO 0006960: **antimicrobial humoral response (sensu Invertebrata)**

many of which are not invertebrate genes but rather human genes (for example COPE HUMAN, PTGE HUMAN, PTE1 HUMAN, and so on). It is surely obvious that a gene with suffix 'HUMAN' should not be annotated to a biological process which is assigned to invertebrates.

In addition, there are some 25 cases where sensu terms are listed by GO as synonyms of non-sensu terms, which seems to contravene GO's own stated rationale for the introduction of the sensu operator. Note that a term of the form '**X (**sensu Y**)**' should not be taken to refer only to those instances of the class X which occur only in species Y. Rather it takes the form of an instruction: use the term 'X' in the way this term is used by people working on species Y.

Jane Lomax (personal communication) has pointed out that sensu was originally created to distinguish identical text strings with different meanings: an early example was 'mating' in mouse and in yeast. By adding 'sensu', the idea was not to exclude certain taxa from using a sensu term, but rather to give a user an idea of what sense a term should be used in. For example, if another flying insect were to be annotated to GO, we would hope that the 'sensu Drosophila' terms could be used for this new species.

3.4 Problems with synonyms

GO has within its December 2003 edition some 4740 synonyms for its 16660 terms, many terms having more than one synonyms. Such a large percentage of terms (3570 distinct terms, 21.43%) with synonyms implies that one needs to be rigorous in determining if the synonyms of the terms are correct. There are many different errors which one finds within GO's synonyms, some of them are as follows:

a. Cases where a term and a subordinate term are treated as synonyms, for example:

> GO 0004601: **peroxidase activity** has synonyms **lactoperoxidase activity** and **myeloperoxidase activity**

b. Cases where terms for an activity and its bearer are made synonyms (this seems to be the commonest type of error):

> GO 0005344: **oxygen transport activity** has synonyms **hemoerythrin and hemocyanin**
>
> GO 0004907: **interleukin receptor activity** has **synonym IL receptor**
>
> GO 0003823: **antigen binding** has synonym **antibody**

c. Cases where terms for events occurring in sequence are treated as synonyms:

> GO **associative learning** has synonyms **conditional learning** and **conditional response**

d. Cases where terms for location and content are treated as synonyms:

> GO 0042597: **periplasmic space** has synonym **periplasm**

e. Cases where terms can be treated as synonyms only in relation to a specific group of cases:

> GO 0005803: **secretory vescicle** has synonym **transport vescicle**

f. Cases where a 'sensu' term has a synonym which is not specific for that particular species

> GO 0012511: **lipid storage body (sensu Viridiplantae)** has synonyms **oil body**, **oleosome**, **spherosome** and **GO: 0009520**

g. Cases where synonyms are term IDs when those IDs do not map to any further term. (For example under f., **GO: 0009520** is considered a synonym of the other terms mentioned.)

Some, at least, of these problems are to a degree ameliorated if one recognizes that GO means something very special by 'synonym', which despite their name, "are not always exactly synonymous to the term they are attached to. This is because it is often useful to search on a string related to the term of interest, for example, if I search GO for 'respiration' I retrieve two terms, 'respiratory gaseous exchange; GO:0007585' and 'cellular respiration; GO:0045333' which I can choose between, although respiration is not directly synonymous with either term. Equally, it is also useful to include exactly synonymous terms and very loosely related terms such as individual gene products, so there is actually a spectrum of relationships between GO terms and their 'synonyms'." [24]

4. SWISS-PROT

SWISS-PROT is a curated protein sequence database which provides descriptions of proteins and annotations to a variety of further types of data and is integrated with over 60 proteomics and protein-related databases.

There are two types of data present within SWISS-PROT – core data, which consists of sequence data together with citation information and taxonomic data (descriptions of the biological source of the protein); and annotation data, which pertains to the functions of the protein and post-translational modifications (for example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc.), domains and sites (for example calcium binding regions, ATP-binding sites, zinc fingers, homeobox, kringle, etc.); secondary structure, quaternary structure (for example homodimer, heterotrimer, etc.), similarities to other proteins, disease(s) associated with deficiencie(s) in the protein, sequence conflicts, variants, etc.

SWISS-PROT affirms that, "the core data is well mapped, [but] there are problems with some of the annotation data, especially those which are put within 'Comments' and are usually free text."

The headings used to organize annotation data consist of: **allergen**, **alternative products**, **biotechnology**, **catalytic activity**, **caution**, **cofactor**, **database**, **developmental stage**, **disease, domain**, **enzyme regulation**, **function**, **induction, mass spectrometry**, **miscellaneous**, **pathway**, **pharmaceutical**, **polymorphism**, **posttranslational modification**, **RNA editing**, **similarity**, **subcellular location** and **tissue specificity**.

4.1 Problems with comments

In order to determine the sorts of problems which might associated with the task of integrating its annotation information, we investigated the comments included in SWISS-PROT with respect to GO:

SWISS-PROT: **induction** has 59 non-obsolete GO counterparts including GO 0001759: **induction of an organ**, GO 0006917: **induction of apoptosis**. Since the annotation component of SWISS-PROT and GO seem to be complementary, and since GO's has the more elaborate structure, it is interesting to investigate how far we can map from SWISS-PROT to the terminology and classification of GO.

Considering SWISS-PROT: **catalytic activity**, SWISS-PROT: **enzyme regulation**, SWISS-PROT: **function**, SWISS-PROT: **induction** and SWISS-PROT **pathway**, almost all of them fall within GO's **biological process** or **molecular function** axes.

GO 003824: **catalytic activity** is classified within the **molecular function** axis. 36 other non-obsolete GO terms containing 'catalysis' or related grammatical forms are divided between this axis (16) and the **biological process** (13) and **cellular component** (7) axes. There are 1305 non-obsolete GO terms which contain the word 'regulation' or related grammatical forms, all of which fall within the **biological process** axis. There 4 are non-obsolete GO terms which contain the word 'function', all of which fall within the **molecular function** axis. Among the 59 terms containing 'induction' or related grammatical forms mentioned above, 54 belong to **biological process**, 4 to **molecular function** and 1 to the **cellular component** axis. SWISS-PROT **pathway** can be related to GO terms containing 'pathway' or related forms (170 within **biological process** only for 'pathway' and 192 including related forms, 6 within **molecular function** and 1 within **cellular component** axes), or to terms containing 'cycle' (86 within **biological process** and 3 within **cellular component** axes).

In any case, therefore, a mapping from SWISS-PROT annotations to the corresponding GO terms will be tangled indeed. This also adds weight to our thesis that GO's **biological process** and **molecular function** axes are not orthogonal. [11]

4.2 Problems related to subcellular location annotations

SWISS-PROT annotates subcellular location for its proteins, while GO's cellular component axis has its own annotations for subcellular locations. These two systems of annotations have not been synchronized, and it seems that GO's cellular component axis has more annotations than are to be found within SWISS-PROT. This might be due to the Gene Ontology Annotation (GOA) project, which aims to apply GO's vocabulary to a non-redundant set of proteins described in the UniProt Resource (Swiss-Prot/TrEMBL/PIR-PSD) and Ensembl databases that collectively provide complete proteomes for Homo sapiens and other organisms. [25] It would be useful if SWISS-PROT's data were synchronized with this GOA data. In light of the problems with GO's treatment of location, however, such a synchronization will only be a first step towards representation of location for proteins. For example,

SWISS-PROT annotates the subcellular location of SWISS-PROT CCHL_HUMAN P53701: **cytochrome c-type heme lyase** to: **mitochondrial inner membrane**. GO, on the other hand, links the same protein with GO 0005758: **mitochondrial**

intermembrane space (defined as 'The space between the mitochondrial outer membrane and inner membrane') and also with GO 0005743: **mitochondrial inner membrane**, both of which are described as part-of GO 0005735: **mitochondrion**. Comparing the two annotations, we can find SWISS-PROT as having a less compete coverage than GOA.

To take matters further, Lill *et al* have specified the protein to be 'bound to mitochondrial inner membrane and located within the intermembrane space'. [26] GO's cellular component axis does not have relations of the type 'is-bound-to' or 'is-located-in'. This means that there is still unclaimed knowledge within the medical text which can not be accounted by GO's annotations, especially when 'boundedness' or 'location' to a subcellular location has implications regarding protein's 3D structure and posttranslational modifications and is well-documented within SWISS-PROT itself.

5. Conclusion

There are a number of reasons why the first steps towards proteomics meta-classification must be taken manually:

1. Use of different classificatory aspects rests on tacit specialist knowledge, and there is no automated method to derive the latter.
2. Is-a and part-of relations have been employed in inconsistent ways.
3. Some relationships can be verified automatically; but where (for example in SCOP's treatment of membrane proteins) entities have been classified in a way which hinders meta-classification, one needs to modify the relations manually.

Broadly speaking, such problems exist with most, if not all biomedical ontologies and classifications, and a fundamental review is needed in order to address the pertinent issues. [11,16,19,20,27] Pragmatic principles have led to the creation of relatively simple models and this has facilitated the storage of large amounts of biomedical data within a reasonably short period of time by biologists and biomedical researchers. In the next era, adherence to formal principles of representation would help make such data better manageable, and this in turn would contribute to more robust bioinformatic science.

Acknowledgements: Our thanks go to the Wolfgang Paul Program of the Alexander von Humboldt Foundation.

References

[1] Ouzounis CA, Coulson RM, Enright AJ, Kunin V, Pereira-Leal JB. Classification schemes for protein structure and function. Nat Rev Genet. 2003 Jul; 4(7):508-19.
[2] Liu J, Rost B. Comparing function and structure between entire proteomes. Prote Sci. Oct;10(10):1970-9, 2001.
[3] Lo Conte L, Brenner SE, Hubbard TJP, Chothia C, Murzin A. (2002). SCOP database in 2002: refinements accommodate structural genomics. Nucl. Acid Res. 30(1), 264-267.
[4] http://scop.mrc-lmb.cam.ac.uk/scop/
[5] Bourne PE, Weissig H. The Protein Data Bank. Structural Bioinformatics. Hoboken NJ, John Wiley. 2003; 181-198.
[6] http://www.rcsb.org/pdb/
[7] Gasteiger E, Jung E, Bairoch A. Swiss-Prot: connecting biomolecular knowledge via a protein database. Curr Issues Mol Biol. 2001 Jul;3(3):47-55.
[8] http://us.expasy.org/sprot/
[9] Gene Ontology Consortium. Creating the Gene Ontology Genome Res. 2001. 11: 1425-1433.
[10] http://www.geneontology.org/
[11] Smith B, Williams J and Schulze-Kremer S. The Ontology of the Gene Ontology. Proc AMIA Symp. 2003.
[12] http://www.geneontology.org/doc/GO.doc.html
[13] http://www.ncbi.nlm.nih.gov/LocusLink/
[14] http://www.uni-leipzig.de/~akumar/colon_cancer.zip
[15] http://www.uni-leipzig.de/~akumar/colorectal_cancer.-zip
[16] Ogren PV, Cohen KB, Acquaah-Mensah GK, Eberlein J, Hunter L. The compositional structure of Gene Ontology terms. Pacific Symp Biocomputing 2004. http://www-smi.stanford.edu/projects/helix/psb04/ogren.pdf
[17] Achour LS, Dojat M, Rieux C, Bierling P, Lepage E. A UMLS-based Knowledge Acquisition Tool for Rule-based Clinical Decision Support System Development. J Am Med Inform Assoc 2001;8(4):351-360.
[18] http://www.uni-leipzig.de/~akumar/terms_GO.htm
[19] Smith B, Rosse C. The role of foundational relations in the alignment of biomedical ontologies. Medinfo 2004.
[20] Smith B and Mulligan K, Framework for formal ontology, Topoi, 1983; 3: 73-85.
[21] http://www.geneontology.org/GO.usage.html#partof.
[22] Mejino JLV, Agoncillo AV, Rickard KL, Rosse C. Representing complexity in part-whole relationships within the Foundational Model of Anatomy. Proc AMIA Symp. 2003.
[23]http://www.geneontology.org/doc-/GO.usage.html#sensu
[24] http://www.geneontology.org/GO.synonyms.html
[25] http://www.ebi.ac.uk/GOA/
[26] Lill R, Stuart RA, Drygas ME, Nargang FE, Neupert W. Import of cytochrome c heme lyase into mitochondria: a novel pathway into the intermembrane space. EMBO J. 1992 Feb;11 (2):449-56.
[27] Kumar A, Smith B. The Unified Medical Language System and the Gene Ontology: Some Critical Reflections. (Lecture Notes in Computer Science. 2821) 2003: 135-148.