Proceedings of the 3rd German Conference on Bioinformatics (GCB 2003), Munich, vol. II, 55-57.

A Framework for Protein Classification

Anand Kumar, Barry Smith
[a]Laboratory of Medical Informatics, Department of Computer Science, University of Pavia, Italy
[b]Institute for Formal Ontology and Medical Information Science, University of Leipzig, Germany
[c]Department of Philosophy, University at Buffalo, NY, USA

ABSTRACT

It is widely understood that protein functions can be exhaustively described in terms of no single parameter, whether this be amino acid sequence or the three-dimensional structure of the underlying protein molecule. This means that a number of different attributes must be used to create an ontology of protein functions. Certainly much of the required information is already stored in databases such as Swiss-Prot, Protein Data Bank, SCOP and MIPS. But the latter have been developed for different purposes and the separate data-structures which they employ are not conducive to the needed data integration. When we attempt to classify the entities in the domain of proteins, we find ourselves faced with a number of cross-cutting principles of classification. Our question here is: how can we bring together these separate taxonomies in order to describe protein functions? Our proposed answer is: via a careful top-level ontological analysis of the relevant principles of classification, combined with a new framework for the simultaneous manipulation of classifications constructed for different purposes.

INTRODUCTION

The primary function of a protein is to regulate the functions of a cell. Protein functions have accordingly been classified in terms of the roles proteins play within the cell, and different classification systems have been proposed for this purpose. Recent reviews of classification discuss attributes related to protein structure and the July 2003 issue of *Nature Genetics* includes an overview of the various structural and functional classifications which have been proposed (Ouzounis *et al*., 2003). Its authors point to the need for a meta-classification as a foundation upon which more refined classifications – and ontologies – can be built. As (Liu & Rost, 2001) have argued, a wide range of additional factors needs to be taken into account for a complete description of a protein's function. These factors include not only cellular roles but also molecular functions and the involvement of proteins in physiology, pathology and evolution. (Benner & Gaucher, 2001) say that the prediction of a protein's function depends on the level of granularity (molecular, cellular, etc.) considered. Proteins with similar primary sequences can have entirely different three-dimensional structures and cellular roles, while proteins with similar cellular roles can have very different primary sequences.

But what *is* a protein function? And what do we need to know about a protein in order to be able to describe and predict its function? Probably we will never be able to provide an exhaustive list of such determinants. However, such a list would have to include at least the following items: protein sequence, protein structure, subcellular localization, biomolecular interactions, tissue specificity, and site of production.

ONTOLOGICAL DISTINCTIONS

Our proposal is that we can build the needed framework by drawing on certain distinctions made in ontology. When we talk about proteins in terms of their structure and functions, we can be talking about **continuants** and **occurrents**. Continuants are entities which continue to exist through time. Organisms, tissues, proteins are all examples of continuants: they preserve their identity continuously from one moment to the next even while undergoing a variety of different changes. Functions, too, are continuants; the function of a given protein to, say, bind oxygen exists identically from one moment to the next, and it exists even when it is not being exercised. Occurrents (also called processes, events, activities – for example *oxygen binding activity*) are entities which *occur*. They are marked by the fact that they never exist in full in any single instant of time. Rather, they unfold themselves through time: they have a beginning, a middle and an end. The oxygen transport function of hemoglobin molecules is a dependent continuant; the process of transporting oxygen performed by the same hemoglobin during some interval of time is an occurrent.

Orthogonal to the distinction between continuants and occurrents is that between **dependent** and **independent** entities. Dependent entities are entities which require support from other entities in order to be sustained in existence, for example the function of a protein is dependent on the protein; dependence relationships can obtain also between dependent continuants themselves, for example *regulation of alveolarcapillary protein gradient* is dependent on *alveolarcapillary protein gradient.* Independent entities, in contrast do not require a support of this kind in order to exist; examples are erythrocytes and organisms. Thus erythrocytes can exist in a suitable medium without any support from other entities, and so, too, can organisms.

THE THEORY OF GRANULAR PARTITIONS

When human beings classify the entities in some given domain, they partition it into cells and subcells of various types at different levels of granularity. The Theory of Granular Partitions (TGP) provides a framework for understanding and manipulating such partitions (Bittner and Smith, 2003). Partitions thus created can reflect in each case only one type of taxonomic structure. Hence in a complex domain like that of proteins, several partitions need to be connected together in order to provide the necessary complete picture.

A taxonomy is a partition consisting of cells and subcells organized via subsumption and projected onto entities. The cells and subcells are nested together in the form of a taxonomic hierarchy which can consist of many layers. The lowest layer of subcells corresponds to the finest grain of entities recognized by the partition in question. The conditions on a good taxonomy proposed by TGP include the following:

A1: Every partition has a unique maximal cell in which all other cells are included as subcells.

A2: The subcell relation is reflexive, antisymmetric, and transitive.

A3: If two cells within a partition overlap, then one is a subcell of the other.

Partitions which satisfy these simple principles have the graph-theoretical form of a tree.

PARTITIONS IN PROTEOMICS

The principal idea behind our framework for protein classification is to create distinct partitions reflecting the ontological distinctions mentioned above. The partitions so created will then constitute the required metaclassification built up on the basis of the parthood and dependence relationships among the entities classified. The underlying idea is that a protein is an independent continuant: it is a chemical substance which endures identically through in time and which can exist independently of other substances. The structures, attributes and functions of proteins are continuants also, but the processes in which proteins are involved are occurrents. Structures, attributes, functions and processes are all however dependent entities – they depend for their existence *inter alia* on the proteins which serve as their bearers or carriers.

Our goal to illustrate our projected framework by focusing on the human protein Hemoglobin A and taking into account the information present in different proteomic databases. Hemoglobin is an [a(2):b(2)] tetrameric hemeprotein found in erythrocytes, where it is responsible for binding oxygen in the lung and transporting the bound oxygen through the body, where it is used in aerobic metabolic pathways.

Protein Parts, Complexes and Structural Configurations: Human hemoglobin is represented in PDB (The Protein Data Bank, 2003) in a number of different structural configurations – for example as oxygenated and deoxygeneated hemoglobin – each of which are defined separately. When we create a partition of the protein domain, we can choose either to trace over the existence of such different configurations of the same protein, or we can acknowledge these differences and also the different functions and characteristics which inhere in them. PDB does the latter and also acknowledges the source of the different configurations as human beings. What it does not do – and the same holds of related systems like CATH, PDBSum and SCOP – is include the fact that they are all configurations *of* human hemoglobin. This suggests two requirements, which a needed metaclassification must satisfy:

R1. Different configurations and related functions of the same protein should be acknowledged.

R2. The sources of the proteins classified should be explicitly recorded.

Swiss-Prot (Boeckmann *et al*, 2003), in contrast to PDB, employs a partition which recognizes the different *parts* of hemoglobin – for example hemoglobin alpha chain, hemoglobin beta-1 chain – as distinct entities. Swiss-Prot does not unfortunately include any single cell in its partition which would allow us to recognize that these different molecular chains are parts of the same protein molecule. GO(The Gene Ontology) assigns separate representations both to the *parts* of human hemoglobin, and to the hemoglobin complex of which they form a part. This is an important improvement. Hemoglobin complex itself is then considered by GO as a *part_of Cytosol*. However, the pathways of hemoglobin production and destruction do not necessarily occur in the cytosol and not every cell has hemoglobin as a part of its cytosol. GO's reading of *part-of* as

sometimes part-of thus induces a loss of information. The fact that *hemoglobin complex* is not always a part of the cytosol should be taken into account in the representation, and this generates a third requirement:

R3. Time- and context-specificity of parthood and other relations should be explicitly recorded.

Function vs. Functioning: The function of hemoglobin is that of binding and transporting oxygen. However, it is not always exercising these functions since there are periods in its lifecycle during which the function is present merely as a power or disposition. Each token function, to repeat, is a dependent continuant. Each expression of a function, that is to say each actual performance of the function within a particular interval of time is a process, a token occurrent. As has been argued in (Smith, Williams and Schulze-Kremer, 2003), function and process are often run together in current bioinformatics ontologies, and only by keeping them separate can we do justice to the fact that the function of a protein may exist even when not being expressed. Swiss-Prot presents *Enzyme regulation* and *Function* as separate attributes to describe proteins in its database. It defines *Enzyme regulation* as "Description of an enzyme regulatory mechanism" and *Function* as "General description of the function(s) of a protein". (One incidental problem here turns on the use of the term "description" in such definitions. The definition of *Enzyme regulation* that is here proffered is strictly speaking a definition of *Description of Enzyme regulation*.) A major problem turns on the fact that, since each enzyme is a protein, *Enzyme regulation* is strictly speaking a protein function, and so should be represented as a continuant and as a daughter of the *Function* node. A further node, labelled *Enzyme regulation activity*, should then be introduced in order to do justice to the fact that activities are not continuants but occurrents. This generates a further requirement:

R4. Dependent continuants, independent continuants and occurrents should be differentiated.

GO on the other hand offers the following sequence of *is_a* relations: *Hemoglobin binding* is a *Protein Binding*, which is a *Binding*, which is a *Molecular Function*. Confusingly, all of these terms denote not functions but processes – a confusion that has been remedied only partially by the recent GO policy change (effective as of March 1, 2003) whereby "All GO molecular function term names [with the exception of the parent term *Molecular Function* and of the whole node binding] are to be appended with the word 'activity'."

Partition of Protein Processes: An adequate ontology of protein functions must deal also with the phenomenon of collective exercise of functions. Many functions performed within the cell are such that proteins depend for their functioning on interactions with other biomolecules. Thus, hemoglobin performs the function of oxygen transport only in conjunction with pH, pO2, pCO2 and with biomolecules such as cell membrane lipids, enzyme carbonic anhydrase and 2,3 Diphosphoglycerate which are present in the blood, cell membrane, alveolar epithelium and so on. This yields a fifth requirement:

R5. The dependence of biomolecular functions upon each other should be explicitly represented.

The interaction of proteins with other biomolecules in various biochemical pathways leads to the notion of collective processes, that is, to processes of different sorts which combine together to form a more complex process. The *transporting of oxygen* by blood at any given time involves a combination of associated processes including: *Deoxy hemoglobin binding activity to oxygen*, *Oxygen transport activity across the red blood cell membrane*, *Oxygen transport activity across the alveolar membrane*, *Blood flow activity of pulmonary circulation*, *Blood pH change activity*, *Oxygen carrying capacity change activity*, and so on. Note that even passive processes are called 'activity' here, in line with GO's recent decision to the effect that the nodes of its function hierarchy are not in fact functions but rather 'activities'. All of the mentioned processes are dependent on different organs, cells and biomolecules. Thus, the transporting of oxygen involves the heart, lungs and blood vessels at the organ level; it involves alveolar cells, red blood cells and vascular endothelial cells at the cellular level; and it involves hemoglobin, enzymes and cell membrane components at the molecular level. This yields a further requirement:

R6. The level of granularity of each entity should be recorded explicitly.

Partitions of the Protein Lifecycle: Each protein lifecycle starts from transcription in the nucleus, followed by translation in the cytosol on the rough endoplastic reticulum, which is followed by further steps leading to the final configuration of the protein. We can distinguish two partitions: one dealing with functions, the other with processes. As the Gene Ontology inadvertently recognizes, not all processes are the expressions of functions. Thus an ontology of processes must include more terms than an ontology of functions and expressions of functions. In the case of human hemoglobin, the partition of processes involved in the protein lifecycle would have to include: *porphyrin being synthesized*, *porphyrin converting to heme*, *globin being synthesized*, *heme being inserted into globin chains*, *pairing of globin chains*, *heme being metabolized into bilirubin*, *globin chains being broken down into component amino acids*, and so on. These processes are themselves regulated by other proteins and by reverse feedback from hemoglobin. Since almost all processes in nature involve regulation, as do the functions on which they are based, this leads to a further requirement:

R7. The different sorts of *regulation of* and *regulation by* need to be explicitly recorded.

Partition Protein Processes according to Location in the Human Organism: Human anatomy is involved in proteomics ontology at different levels. Proteins have a *site* of production (bone marrow in the case of hemoglobin); they exercise their functions in particular organs and subcellular locations; protein metabolism involves certain specific sites, and changes in protein structure occur in certain locations within the human body. Swiss-Prot mentions the factor of human anatomy in its classifications but in different contexts without drawing any connections between them. In the context of human hemoglobin, *Site of function*, *Site of induction*, *Pathway*, *Subcellular location*, *Tissue specificity*, etc., are defined in terms of human anatomy without being cross-related. This yields our final requirement:

R8: Representations of processes and functions should be associated with a framework for representing human anatomy at different levels of granularity.

IMPLEMENTATION FORMALISM

The list of partitions mentioned in this text is not exhaustive, and it can be further extended with partitions of Protein Attributes (for example, molecular weight, Nitric Oxide scavenging property, isoelectric point, and so on). But the creation of multiple partitions is of value only if we can understand and exploit the relationships which exist between them. A sample of this effort is shown in Figure 1, which includes partitions of hemoglobin structure, hemoglobin function, heme sysnthesis, oxygen transport activity and human anatomy and shows the dependent-independent, continuant-occurrent, boundary and parthood relationships which span these different partitions. The data sources used are PDB, Swiss-Prot, GO and text knowledge.

Our representation of the relations between the functions, for example the parthood relation between *Regulation by oxyHemoglobin concentration* and *Regulation of Oxygen Binding*, reflects the thesis that the human organism body consists of regulated systems, and thus consideration of its different processes from the regulation point of view provides a unique opportunity to put together the different body functions. This leads also to the conception of collective functions and processes (R5 and R7). The dependence of functions and processes on independent continuants – that is, on organs, tissues and cells – is then traced at different levels of granularity, for example in: *Capillary Endothelial Cell Membrane is-boundary-of Capillary Endothelial Cell*, which *is-part-of Capillary*, which *is-part-of Blood Vessels*, which *is-part-of Human Body*. This traces the relations between different protein configurations and the functions they are involved in (R6). The mereological relationships regarding the location of protein processes in the partition of human anatomy also indicate that hemoglobin is not a part of the cytosol of every cell but that it *is-located-in Red Blood Cell* (R3). Furthermore, the representation records that *Red Blood Cell is-located-in Blood*, which *is-located-in Blood Vessel*, which *is-part-of Human Body*, and thus this hemoglobin is a human hemoglobin and not the hemoglobin of any other organism (R2). Furthermore, it also shows the different granularities relating alveolar cell to lung or capillaries to blood vessels (R8). The links: *Regulation by deoxyHemoglobin concentration is-dependent-on deoxyHemoglobin concentration*, which *is-dependent-on deoxyHemoglobin*, which *is a deoxidised state of Hemoglobin*, provide the dependence relationship between two dependent continuants in the former case and between a dependent and an independent continuant in the latter case and thus provide a means to connect different functions with their different configurations and with the underlying substances (R1). Different partitions pertaining to functions and processes have been represented in addition in such a way that one can see how the corresponding entities are combined together via the dependence relations between the functions, processes and substances involved. (R4).

This implementation is a sample of how to satisfy the different requirements which arise when dealing with the schemas and structures of protein databases and taxonomies like Swiss-Prot and the Gene Ontology. It is a first step in the direction of a multi-partition ontology which would enable the integration of different classification systems which meets the criteria proposed in (Ouzounis *et al*., 2003)

REFERENCES

1 Benner SA and Gaucher EA. Evolution, language and analogy in functional genomics.Trends Genet. 2001 Jul;17(7):414-8.

2 Bittner T and Smith B. 'A theory of granular partitions, Foundations of Geographic Information Science, M. Duckham, M. F. Goodchild and M. F. Worboys, eds., London: Taylor & Francis, 117-151 (2003).

3 Boeckmann B, Bairoch A, Apweiler R, Blatter M, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, and Schneider M. The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003.Nucleic Acids Research 31: 365-370 (2003).

5 Gene Ontology http://www.geneontology.org/.

6 Liu J and Rost B. Comparing function and structure between entire proteomes. Protein Sci. Oct;10(10):1970-9 (2001).

7 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.
8 Smith B, Williams J and Schulze-Kremer S. The Ontology of the Gene Ontology. Proc AMIA Symp. 2003.
8 Bourne PE and Weissig H. The Protein Data Bank. Details of the history, function, development, and future goals of the PDB resource. Structural Bioinformatics Hoboken, NJ, John Wiley & Sons, Inc. pp. 181-198 (2003).

Acknowledgments: Work on this paper was supported by the Wolfgang Paul Program of the Alexander von Humboldt Foundation. We thank Riccardo Bellazzi and Bert Klagges for helpful comments on the manuscript.

Figure 1 Sample partitions related to Human Hemoglobin