

A Framework for Protein Classification

Anand Kumar^a, Barry Smith^{b,c}

^aLaboratory of Medical Informatics, Department of Computer Science, University of Pavia, Italy

^bInstitute for Formal Ontology and Medical Information Science, University of Leipzig, Germany

^cDepartment of Philosophy, University at Buffalo, NY, USA

INTRODUCTION

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. Our proposal is that we can build the needed framework by drawing on certain distinctions made in ontology.

ONTOLOGICAL DISTINCTIONS

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. 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, for example, 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). The conditions on a good taxonomy proposed by TGP include – every partition has a unique maximal cell in which all other cells are included as subcells; the subcell relation is reflexive, antisymmetric, and transitive and if two cells within a partition overlap, then one is a subcell of the other.

PARTITIONS IN PROTEOMICS

The principal idea behind our framework for protein classification is to create distinct partitions reflecting the ontological distinctions mentioned above. Our goal to create an example of how our projected framework will work by focusing on the human protein Hemoglobin A and taking into account the information present in different proteomic databases.

Protein Parts, Complexes and Structural Configurations: Human hemoglobin is represented in PDB (The Protein Data Bank) in a number of different structural configurations – for example as oxygenated and deoxygenated hemoglobin – each of which are defined separately. 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.

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.

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

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. This yields a fifth requirement:

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

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. All of the mentioned processes are dependent on different organs, cells and biomolecules. 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 endoplasmic 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. 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. 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

Our representation in Figure 1 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 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 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).
- 2 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).
- 3 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.
- 4 Smith B, Williams J and Schulze-Kremer S. The Ontology of the Gene Ontology. *Proc AMIA Symp.* 2003.

Acknowledgment: 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

