Enhancing GO for the Sake of Clinical Bioinformatics

Anand Kumar[1], Barry Smith[1,2]
[1]IFOMIS, University of Leipzig, Germany.
[2]Department of Philosophy, SUNY at Buffalo, USA.

Recent work on the quality assurance of the Gene Ontology[1] (GO, Gene Ontology Consortium 2004) from the perspective of both linguistic and ontological organization has made it clear that GO lacks the kind of formalism needed to support logic-based reasoning. At the same time it is no less clear that GO has proven itself to be an excellent terminological resource that can serve to combine together a variety of biomedical database and information systems. Given the strengths of GO, it is worth investigating whether, by overcoming some of its weaknesses from the point of view of formal-ontological principles, we might not be able to enhance a version of GO which can come even closer to serving the needs of the various communities of biomedical researchers and practitioners.

It is accepted that clinical and bioinformatics need to find common ground if the results of data-intensive biomedical research are to be harvested to the full. It is also widely accepted that no single method will be sufficient to create the needed common framework. We believe that the principles-based approach to life-science data integration and knowledge representation must be one of the methods applied. Indeed in dealing with the ontological representation of carcinomas, and specifically of colon carcinomas, we have established that, had GO (and related biomedical ontologies) followed some of the basic formal-ontological principles we have identified (Smith et al. 2004, Ceusters et al. 2004), then the effort required to navigate successfully between clinical and bioinformatics systems would have been reduced. We point here to the sources of ontologically-related errors in GO, and also provide arguments as to why and how such errors need to be resolved.

The State of Affairs with Biomedical Terminologies

GO is of course not alone among the terminologies created by researchers in biomedical informatics in manifesting problems of an ontological sort. The depth of these problems becomes clear when we find GO and other terminologies (SNOMED CT[2], UMLS[3], etc.) breaching principles at the very foundations of ontology, to a degree which might lead one to question whether they are ontologies at all. Some of these breaches include:

1. The failure to differentiate between universals and instances (*red blood cell* and *this red blood cell*)
2. The failure to differentiate between continuants and occurrents (*enzyme* and *enzymatic activity*)
3. The failure to differentiate between dependent and independent entities (*respiration* and *lungs*)
4. The failure to adhere to the principle according to which the classes on any given level of a classification should be jointly exhaustive and pairwise disjoint (as in: *endodeoxyribonuclease activity* and *exodeoxyribonuclease activity*)
5. The failure to differentiate in definitions between parents and children
6. Inaccuracies in treatment of the parthood relationships (*nucleus* part of *cell*)
7. The failure to keep track in a consistent way of levels of granularity
8. The failure to respect standard principles in the formulation of definitions (*hemolysis* is defined in GO as: The processes that cause hemolysis, the lytic destruction of red blood cells with the release of intracellular hemoglobin, in another organism.)
9. Loose principles for synonymy (*antibody* and *antigen binding*)
10. No concern for consistent treatment of the compositional structure of terms (*activator of the establishment of competence for transformation activity*)
11. The failure to deal adequately with time.

[1] www.geneontology.org
[2] http://www.snomed.org/
[3] http://www.nlm.nih.gov/research/umls/

It is these problems which provide a hindrance to the kind of integration of data and knowledge we need within clinical bioinformatics.

THE NEEDS OF CLINICAL BIOINFORMATICS

Clinical bioinformatics and its various associated disciplines (including clinical pathology, biochemistry, biology, and pharmacy) involve the representation of entities which exist at various levels of granularity and which stand to each other in a variety of complex relations. To make a sound clinical bioinformatics ontology, we need therefore not only to respect the formal principles outlined above, but also to provide a good theory of the relevant levels of granularity, namely:

organism:	*human body*
organ system:	*digestive system*, *respiratory system*, *nervous system*, …
organ:	*pharynx*, *esophagus*, *stomach*, *colon*, …
tissue$_1$:	*mucosa of colon*, *submucosa of colon*, …
tissue$_2$:	*epithelium of mucosa of colon*, …
cell:	*colon epithelial cell*, *fibrocytes*, …
subcellular:	*colon epithelial cell nucleus*, *colon epithelial cell membrane*, …

of the specific material structures represented in each. We then need to localize each pathological structure and each pathophysiological process by associating it with the anatomical entity to which it belongs. Thus colon carcinoma of Stage T1 is associated at the tissue level of granularity with the *mucosa and submucosa of the wall of colon*. The co-occurring abnormal p53 activity (on the subcellular level of granularity) is associated, within the enterocytes (on the cellular level of granularity), with the epithelium of the mucosa of the colon (at the tissue level of granularity).

GO contains for purposes of locating entities anatomically only the cellular component ontology, which deals with anatomical entities only at the cellular and subcellular levels of granularity. To deal with entities situated at higher levels we need to use external ontologies, for example, the Foundational Model of Anatomy[4] (FMA). GO also contains ontologies for molecular functions and biological processes, which are again localized to specific levels of granularity; molecular functions (usually) to the subcellular levels and biological processes to granularities anywhere from the cellular to the organism level.

THE METHODS TO ENHANCE GO AND OTHER BIOMEDICAL ONTOLOGIES

To deal with such matters we need, in addition to the resources of GO, also textbook knowledge and knowledge deduced from statistical, probabilistic and linguistic methods which can help to establish links within the three orthogonal axes of GO and also to locate the biological processes to specific cells and tissues related in turn to specific organs.

REFERENCES

1. The Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32: D258-D261.
2. Smith B, Köhler J, Kumar A. On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology. DILS 2004, Leipzig. Lecture Notes in Bioinformatics. 2994; 79-94, 2004.
3. Ceusters W, Smith B, Kumar A, Dhaen C. Mistakes in medical ontologies: Where do they come from and how can they be detected? in DM Pisanelli (ed.), Ontologies in Medicine: Proceedings of the Workshop on Medical Ontologies, Rome October 2003, Amsterdam: IOS Press, 2004.

[4] http://sig.biostr.washington.edu/projects/fm/AboutFM.html