Ontological Model for Colon Carcinoma: A Case Study for Knowledge Representation in Clinical Bioinformatics

Kumar A[1], Yip L[2], Jaremek M[3], Scheib H[2]

[1]*Institute for Formal Ontology and Medical Information Science, Universität Leipzig, Deutschland*
[2]*Swiss Institute for Bioinformatics, Geneva, Switzerland*
[3]*Ludwig-Maximilians Universität, München, Deutschland*
anand.kumar@ifomis.uni-leipzig.de

Introduction

Clinical bioinformatics lacks a formal representation for diseases which can take care of the levels of granularity which exist from the clinical level to the genetic level. Such a representation needs integration across various databases and creation of parts of the representation manually due to missing information within the existing databases. [1] Such missing information can be partly supplemented by knowledge discovery using statistical and probabilistic methods. We have considered colon carcinoma and related disorders in order to create a formal ontological representation as a case study for the feasibility of such an approach in clinical bioinformatics.

Methods

The clinical level of knowledge representation includes disease classification, its symptoms, predisposing factors, investigations and screening. Snomed CT classification is used for the diseases and their predisposing factors. [2,3] The problems with Snomed CT include those regarding its child-parent representation, siblings and multiple inheritance and we have addressed some of these issues in [4]. The existing Snomed CT problems meant that we modify the Snomed CT hierarchy to some extend suited for our purposes. Various relations like “predisposing factor for”, “investigation for” etc. are introduced to make the ontology specific to the disease, in this case colon cancer and related diseases.

We use the Online Mendelian Inheritance in Man (OMIM) database to find the gene products related to the diseases in question. Various terms are used to query OMIM like “colon carcinoma”, “colon cancer”, “colon polyps” etc. and we extract the set of genes for each of them. [5] These genes are mapped to the LocusLink and the SwissProt database, which provide information for the respective genes and proteins with structures. [6,7] Such structural information is useful for the representation of mutations. The SwissProt database provides information about the function of the respective gene product and its subcellular locations.

The Gene Ontology Annotations for the gene products are used to supplement the information on the location of the gene products and their functions. [8] There are problems with the formal structure of Gene Ontology, which we have pointed them out in various papers, a major drawback being that the three axes (cellular component, biological function and molecular process) are not connected and are

considered orthogonal. [9] The issue of granularity has not be addressed in GO since biological process actually consist of molecular functions in a lower granularity. We use the weights of database occurrence and apriori algorithm to find association rules between the entities belonging to the three GO axes. [10] Using the two approaches together helps us to find the answers related to the location where a biological function is executed, and also regarding the granular components of a larger biological process. We had to extend the cellular component axis manually as GO does not deal fully with the mereotopological relations, which is better dealt in Foundational Model of Anatomy (FMA). [10] We situate GO terms within FMA wherever they can be suitably placed and then take the complete FMA axis above it.
The protein interaction and KEGG pathway database are used for deciphering the pathways which the gene products belong to. [12,13,14] However, this is where we face largest problems with integration. The issue of non-metabolic pathways are not well addressed within KEGG especially those involved in carcinomas. Moreover, the pathways are at a subcellular level and the interaction of cells with each other during a carcinoma can only be found in textbooks. The same applies for various staging which are used to determine the extent of the carcinoma. These aspects are modeled manually based on the knowledge present within textbooks and take the maximum time and effort.
The complete representation is done within Protégé environment. [15] It provides a frame-based system which is compatible with the OWL standards, provides an extensive framework to represent various relations and also has plug-ins for providing reasoning and web-based representation.

Discussion

We approach an evolving topic – that of representation of various aspects of knowledge related to a disease or a set of diseases from clinical to the molecular level, taking into consideration the granularity at various levels very formally.
To the extent possible, we reuse the knowledge and information present within various databases and ontologies, which provide a relatively complete knowledge at the gene-to-gene product level but provide little or no knowledge as one moves to the cellular and tissue level. Moreover, the databases and ontologies at molecular level are not well-connected with those which exist at the clinical level. Thus, while a large part of our work is automated, there is a big manual component of the work because the relevant data does not exist.
Protégé provides us with an extensive framework and environment for such a representation but it is difficult to automate the representation in Protégé and one tends to lose a lot of information within the structure of various databases and ontologies if one tries to import them within Protégé using automated database connections. So a large part of the work dealing with the representation within Protégé is manual too.
The reason why we considered colon carcinoma to begin with, is also formal. We had previously mapped all the diseases present within SwissProt database, dealing with structural mutations of gene products into Snomed CT and based on graph theory, found out which set of diseases are the most specific and still have the maximum number of associated mutated gene products. This will help within our

final representation where we could associate the structural mutations at the molecular level, to the functions and processes affected at subcellular, cellular and tissue level. As a side argument, this also means that we probably will be more successful with automation of the representation and related data integration if we consider diseases which deal with metabolism, for example, diabetes mellitus, since the metabolic pathways are represented with KEGG with reasonable accuracy and details. However, the integration of such pathways at the subcellular level to those at cellular and tissue level will still be manual for a long time to come.

We hope that a detailed granular representation of the sorts, will provide a large framework to locate the work done at various levels by different experts, including clinicians, biologists, pharmacists, nurses and so on. This will, in turn, lead to integration of data required for drug development, education of biomedical and medical students about the relations between diseases and pathways and data integration to the clinical level which will provide individualized decision support.

Acknowledgement

This paper was written under the auspices of the Wolfgang Paul Program of the Alexander von Humboldt Foundation and partly by the Network for Excellence project "SemanticMining" funded by the European Union in the Sixth Framework.

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] Spackman KA, Campbell KE, Cote RA. SNOMED RT: A reference terminology for health care. Proceedings/AMIA Annual Symposium. :640-4, 1997.

[3] College of American Pathologists. SNOMED Clinical Terms Requirements Document. http://www.snomed.org/snomedct1/SNOMEDCT_Objective_V05.pdf

[4] Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in DL-based terminologies: A case study in SNOMED CT. KR-MED 2004. [In press]

[5] McKusick, V.A.: Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders. Baltimore: Johns Hopkins University Press, 1998 (12th edition).

[6] Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources.Nucleic Acids Res 2001 Jan 1;29(1):137-140

[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] Camon E, Magrane M, Barrell D, Binns D, Fleischmann W, Kersey P, Mulder N, Oinn T, Maslen J, Cox A and Apweiler R. The Gene Ontology Annotation (GOA) project: implementation of GO in Swiss-Prot, TrEMBL and InterPro. Genome Research Apr; 13 (4): 662-672 (2003).

[9] Kumar A, Smith B. Towards a Proteomics Metaclassification. IEEE Fourth Symposium on Bioinformatics and Bioengineering 2004. [In Press]

[10] Borgelt C, Kruse R. Induction of Association Rules: Apriori Implementation. in: 15th Conference on Computational Statistics (Compstat 2002, Berlin, Germany) Physica Verlag, Heidelberg, Germany 2002.

[11] Rosse, C. and Mejino, J. L. V. (2003) A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy. Journal of Biomedical Informatics 36:478-500.

[12] Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D (2000) DIP: The Database of Interacting Proteins. Nucleic Acids Res. 28:289-91

[13] Bader GD, Betel D, Hogue CW. (2003) BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. 31(1):248-50

[14] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 27, 29-34 (1999).

[15] Noy NF, Sintek M, Decker S, Crubezy M, Fergerson RW, Musen MA. Creating Semantic Web Contents with Protege-2000. IEEE Intelligent Systems 16(2):60-71, 2001.