An Ontology for Carcinoma Classification for Clinical Bioinformatics

Anand Kumar[1], Yum Lina Yip[2], Barry Smith[1,3], Dirk Marwede[1,4], Daniel Novotny[1,3]

[1]IFOMIS, University of Saarland, Saarbruecken, Germany
[2]Swiss Institute of Bioinformatics, Geneva, Switzerland
[3]Department of Philosophy, University at Buffalo, Buffalo, New York, USA
[4]Department of Diagnostic Radiology, University of Leipzig, Leipzig, Germany

Abstract

There are a number of existing classifications and staging schemes for carcinomas, one of the most frequently used being the TNM classification. Such classifications represent classes of entities which exist at various anatomical levels of granularity. We argue that in order to apply such representations to the Electronic Health Records one needs sound ontologies which take into consideration the diversity of the domains which are involved in clinical bioinformatics. Here we outline a formal theory for addressing these issues in a way that the ontologies can be used to support inferences relating to entities which exist at different anatomical levels of granularity. Our case study is the colon carcinoma, one of the most common carcinomas prevalent within the European population.

Keywords:

Clinical bioinformatics, Ontology, Oncology

1. Introduction

Over 40 carcinomas have been analyzed in terms of the Tumor, Node and Metastasis (TNM) classification, which classifies a carcinoma on the basis of the extent of its spread, lymph node involvement and metastasis and which is used for staging the relevant guidelines for the management of patients [1,2]. An automated ontology-based system designed to establish for each specific pathology the appropriate TNM class would first need to register the anatomical structures (colon, lung, prostate, etc.) involved. Unfortunately, the Electronic Health Records (EHRs) often use anatomical terms different from those used in the TNM classification. Thus the systems need to provide a methodology by which different anatomical terms can be related to the corresponding TNM anatomical entity. In addition to anatomy, however, the system would need to take account also of pathologies, cellular characteristics, and other salient features. For an efficient integration with EHRs, the ontology system would need to represent such features explicitly and in such a way that it supports inferences drawing not only on EHR data but also on biological databases dealing with the mutations and protein variants which are involved in the different stages of carcinoma development.

Clinical bioinformatics is a field based on the integration and interpretation of life-science data at many levels of granularity, from the coarser (clinical) levels of signs, symptoms, radiological findings, to the finer levels of genotype-phenotype expression and associated molecular pathways. To achieve this end, an ontological theory is required that is

able to deal with such a diversity among the entities involved. Here, we present a theory of the needed sort, which is designed to do justice to the fact that the anatomical entities as represented within an anatomy reference ontology can be used to support the drawing of inferences relating to the anatomical and pathological data present within the EHRs of individual patients. We use the example of colon carcinoma as a case study and present an extension to our previous work, where using our own ontology of relations we integrated terms in Gene Ontology[1] with the database of protein mutations present within Swissprot[2] [3,4,5].

2. Formalization

Classes and individuals

As anatomy reference ontology we have selected the Foundational Model of Anatomy (FMA),[3] which is a representation of *canonical* anatomy. [6] Thus *colon*, as it appears within the FMA, represents a *normal colon* – where in the present context of course we need to deal also with abnormal colons. The FMA is however designed as a reference ontology that can serve as a basis also for non-canonical anatomy. In the correspondingly extended FMA (EFMA) we have both:

abnormal colon ***is_a*** *colon*
normal colon ***is_a*** *colon*

We distinguish between colon carcinoma disease and the associated pathological structure. Only the latter falls within the domain of EFMA. For the latter we have:

colon carcinoma pathological structure ***is_a*** *carcinomatous pathological structure*
colon carcinoma pathological structure ***part_of*** *abnormal colon*
colon carcinoma pathological structure ***part_of*** *colon*

On the reading of '*part_of*' adopted by us here with our own ontology for relations, *A part_of B* means that every instance of *A* is an instance-level part of *some* instance of *B* (this reading is used in the FMA in conjunction with the reading of '*part_of*' described in Smith Rosse Medinfo paper Every colon carcinoma is part of some colon, but not every colon has some colon carcinoma as part. We also have

abnormal colon ***transformation_of*** *normal colon*

where *A transformation of B* holds where *A* and *B* are distinct classes which are such that for any instance *a* of *A* and any time *t*, there is some earlier time t_1, at which *a* an instance of *B*. (We leave aside here those abnormalilities which are abnormal from the start because they are present abnormally within the fetus [7].) We further have:

carcinomatous colon ***is_a*** *abnormal colon*
carcinomatous colon ***transformation_of*** *normal colon*

Parthood and location relations

The FMA includes parthood relations such as:

ascending colon ***part_of*** *colon*
mucosa of colon ***part_of*** *wall of colon* ***part_of*** *colon*

Because the parthood relation implies a location relation, [8] we also have:

region of ascending colon ***located_in*** *region of colon*
region of mucosa of colon ***part_of*** *region of wall of colon* ***part_of*** *region of colon*

[1] www.geneontology.org
[2] http://au.expasy.org/sprot/
[3] http://sig.biostr.washington.edu/projects/fm/AboutFM.html

Bearer relation

A carcinomatous pathological process demands in every case some anatomical entity as its bearer. Thus

carcinomatous pathological function in ascending colon ***born_by*** *ascending colon*

T1 stage carcinomatous pathological function in ascending colon ***born_by*** *mucosa of ascending colon*

We also have:

carcinomatous pathological process in ascending colon ***has_participant*** *ascending colon*

T1 stage carcinomatous pathological processin ascending colon ***has_participant*** *mucosa of ascending colon*

Creation relation between process and anatomical entity

We have dealt with relations between anatomical entities and the associated biological processes elsewhere. [9] One of them is the creation relationship. A process creates a substance when the substance did not exist before the initiation of the process and exists after the process has started. It is not necessary that a single process brings a substance into being by itself. It is also not necessary that each process brings a substance into being. However, in case of carcinoma of colon, the pathology leads to a pathological structure, though not always. Thus,

carcinomatous pathological process in ascending colon ***creates*** *carcinomatous structure in ascending colon*

T1 stage carcinomatous pathological process in ascending colon ***creates*** *T1 stage carcinomatous structure in ascending colon*

The pathological structure is part of the organ affected by the pathology. There is also a parthood relation between a pathological structure and the pathological organ. Thus,

carcinomatous structure in ascending colon ***part_of*** *carcinomatous ascending colon*

region of carcinomatous structure in ascending colon ***part_of*** *region of carcinomatous ascending colon*

T1 stage carcinomatous structure in ascending colon ***part_of*** *carcinomatous mucosa of ascending colon*

Anatomical levels of granularity

Every anatomical entity in the FMA has a default level of granularity. For example, *colon* has an *organ* level of granularity; *ascending colon* has an *organ part* level of granularity. We then have a mapping 'gran' from FMA terms to levels of granularity, such that for example:

gran(*colon*) = *organ*

and

gran(*ascending colon*) = *organ part*

Finer levels of granularity – of atoms, molecules and subcellular organelles – are also needed for our carcinoma ontology in order that we can represent processes at the cellular level, markers like p53 and the various mutant proteins associated with colon carcinoma and represented in the modSNP database of Swissprot.

TNM Classification

TNM classification classifies carcinomas on the basis of the extent of tumor spread, lymph nodes involved and metastasis. There are certain levels of granularity in anatomy which apply to each of the TNM classes. We have dealt with these in more detail in [10]. We apply them here to the entities which are involved in a specific T, N or M class for colon carcinoma. However, anatomy is not the only partition which is salient to TNM classes. Other partitions, for example pathogenesis and number of lymph nodes involved are also taken into consideration. Our formalism represents the levels of granularity and mentions the other partitions involved.

Table 1. Anatomical levels of granularity for TNM classification

TNM	Explanation as stated in the TNM Classification by the respective Cancer Societies	Level of Granularity	Other partitions
TX	Primary tumor cannot be assessed	–	
T0	No evidence of primary tumor	–	
Tis	Carcinoma in situ: intraepithelial or invasion of the lamina propria	Intraepithelial cell (C); Epithelium & lamina propria (ORP/MPOT)	
T1	Tumor invades submucosa	Submucosa of colon (ORP/MPOT)	
T2	Tumor invades muscularis propria	Muscularis mucosa of colon (ORP/MPOT)	
T3	Tumor invades through the muscularis propria into the subserosa, or into nonperitonealized pericolic or perirectal tissues	Subserosa of colon (ORP/MPOT), nonperitonealized pericolic or perirectal tissues (ORP/MPOT-E)	path of invasion (pathogenesis)
T4	Tumor directly invades other organs or structures, and/or perforates visceral peritoneum	organs (OR-E), visceral peritoneum (ORP)	path of invasion (pathogenesis)
NX	Regional nodes cannot be assessed	–	
N0	No regional lymph node metastasis	–	
N1	Metastasis in 1 to 3 regional lymph nodes	lymph node (ORP), collection of lymph nodes (ORP-C)	number
N2	Metastasis in 4 or more regional lymph nodes	lymph node (ORP), collection of lymph nodes (ORP-C)	number
MX	Distant metastasis cannot be assessed	–	
M0	No distant metastasis	–	
M1	Distant metastasis	ORG	

ORG = Organism; OR = Organ; OR-E = Organ (external to the original organ); ORP = Organ Part; MPOT = Maximal Portion of Tissue (equivalent to organ part but in a different partition); C = Cell; SC = Subcellular

TNM Classification and Staging

Various cancer societies have proposed different staging criteria for carcinomas, which apply at two different levels. Among those carcinomas for which a TNM classification exists, stages are not classified other than by means of the relevant TNM classes. The corresponding management guidelines are then based on these classes, so that for example there is a specific management protocol for the T1N2M0 class of bladder carcinomas. However, in case of some carcinomas, including colon, an additional grouping is imposed upon the TNM classification by the American Joint Committee on Cancer (AJCC). For example, stage IIIa of colon carcinoma includes T1N1M0 and T2N1M0. The management protocols applicable to these TNM classes are similar and thus are usually specified in terms of stage IIIa. This means that the ontological representation of carcinomas like colon is incomplete if the stages of carcinoma defined over the TNM classes are not included, and therefore, we provide those representations in our ontology also.

Table 2: Definitions for AJCC staging for colon carcinoma

Stage	Clinical definition	Logical definition
0	Tis, N0, M0	Tis & N0 & M0
I	T1, N0, M0; T2, N0, M0	(T1 or T2) & N0 & M0
IIa	T3, N0, M0	T3 & N0 & M0
IIb	T3, N0, M0	T4 & N0 & M0
IIIa	T1, N1, M0; T2, N1, M0	(T1 or T2) & N1 & M0
IIIb	T3, N1, M0; T4, N1, M0	(T3 or T4) & N1 & M0
IIIc	Any T, N2, M0	N2 & M0
IV	Any T, Any N, M1	M1

Other staging systems

There are other staging systems for colon carcinoma, for example, the Modified Asler-Coller (MAC), Duke's and so on. These take into consideration entities similar to the ones for TNM, for example, invasion of mucosa or submucosa and so on, and they will not be considered here. According to the kinds of cells predominantly present within the pathology, the different classes are designated as follows: *adenomatous*, *mucinous*, *signet-ring*, *scirrhous* and *neuroendocrine*. This classification involves entities at the cellular and organ part levels.

Pathologic features

In addition to the features taken account of in the TNM classification, there are also several other features which are important in order to understand the prognostic factors and treatment planning for carcinomas. They are primarily related to finer levels of granularity.

Table 3: Pathologic features, their involvement in processes and their anatomical levels of granularity

Pathologic features	Process involved	Level of granularity
vascular endothelial growth factor	angiogenesis	M
reverse transcriptase	lymph node micrometastasis	M
radial margins of carcinomatous structure	carcinoma penetration	OR
degree of tumor differentiation	carcinoma penetration and metastasis	SC, C
mucin producing cancer cells	peritoneal seeding	SC, C
aneuploidy	karyotypic and phenotypic heterogeneity	SC
proliferation index (the sum of the percentage of cells in S phase plus those in G(2)/M phase)	karyotypic and phenotypic heterogeneity	C

Clinical Genomics Model in HL7 v3

The above formalization between various entities need to connected to EHRs in order to be applicable to individual patient attributes, Health Level 7 (HL7)[4] provides standards for the exchange, management and integration of data that supports clinical patient care. There are various domains represented within HL7 and each of those has a Domain Message Information Model (DMIM), based on which various Refined Message Information Models (RMIMs) are built.

The clinical genomics domain encapsulates various types of genomic data relating to a pair of alleles (a locus-genotype) including sequencing, expression and proteomics data. It also incorporates emerging standard formats like BSML[5] (Bioinformatic Sequence Markup Language) and MAGE-ML[6] (Microarray and Gene Expression Markup Language).

The main classes within the clinical genomics Genotype DMIM include genotype attributes, gene associated observation, Individual allele, sequence (recursive class), sequence variation property and expression. These classes represent entities present at the subcellular and molecular levels of granularities and are parts of the cell where they are present.

There are many pathological features which are present at these finer levels of granularity and usage of clinical genomics model would help connect to patient data at those levels of granularity for the purposes of inferences to be drawn on EHRs for carcinoma classifications.

[4] www.hl7.org
[5] http://www.bsml.org/
[6] http://www.mged.org/Workgroups/MAGE/mage-ml.html

3. Conclusion

The main advantages of assigning levels of granularity to the entities with parthood and dependence relations are:

a. Entities within clinical bioinformatics are present at different levels of granularity and various annotations related to genotypes and their expression as phenotypes are associated with them. The interpretation of those annotations are relevant only at the respective granular levels.
b. Inferences derived from data pertaining to the entities and their annotations depend on the granularity levels of the entities involved. For example, if we know that a mutation exists at the subcellular level within a collection of cells then this is not enough to be infer a diagnosis of carcinoma. The TNM classes involve entities clearly distinguished on the basis of anatomical entities existing at various granularity levels.
c. The formalism provides a basis of relations that obtain between substances, functions, processes and their attributes, which are useful to represent relations between pathological structures, anatomical structures and pathological processes applicable to carcinomas [10].

The formalism has been applied in the case of Colon carcinoma representing entities at coarser levels of granularity applicable principally for medical informatics and representing entities at finer levels of granularity applicable principally for bioinformatics. [3, 4] We believe that it is a valuable first step towards bridging the gaps in representing entities for clinical bioinformatics in the domain of oncology.

Acknowledgements

This paper was written under the auspices of the Wolfgang Paul Program of the Alexander von Humboldt Foundation, the European Union Network of Excellence on Medical Informatics and Semantic Data Mining, and the Volkswagen Foundation under the auspices of the project “Forms of Life”.

References

[1] Colon and rectum. In: American Joint Committee on Cancer: AJCC Cancer Staging Manual. 6th ed. New York, NY: Springer, 2002, pp 113-124.
[2] V.T. DeVita, S. Hellman and S.A. Rosenberg. Cancer: Principles and Practice of Oncology, 6th Edition, Lippincott Williams & Wilkins (2001).
[3] Kumar A, Yip L, Smith B, Grenon P. Bridging the Gap between Medical and Bioinformatics Using Formal Ontological Principles. Computers in Biology and Medicine. (submitted)
[4] Kumar A, Yip L, Jaremek M, Scheib H. Ontological Model for Colon Carcinoma: A Case Study for Knowledge Representation in Clinical Bioinformatics. GMDS 2004. 29 Sept. Innsbruck, Austria: 196-198.
[5] Kumar A, Smith B, Borgelt C. Dependence Relationships between Gene Ontology Terms based on TIGR Gene Product Annotations. CompuTerm Aug 29, 2004: 3rd International Workshop on Computational Terminology: 31-38.
[6] Rosse C, Mejino JL Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform. 2003 Dec;36(6):478-500.
[7] Smith B, Ceusters W, Koehler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector A, Rosse C. Relations in Biological Ontologies. (submitted) http://ontology.buffalo.edu/bio/OBORelations.doc
[8] Donnelly M. On parts and holes: the spatial structure of the human body. Medinfo. 2004;2004:351-5.
[9] Smith B, Grenon P. The Cornucopia of Formal-Ontological Relations, Dialectica, 58:3 (2004), 279-296.
[10] Kumar A, Smith B, Novotny DD. Biomedical Informatics and Granularity. Comparative and Functional Genomics. (In Press)

Address for Correspondence
Anand Kumar, IFOMIS, Universität des Saarlandes, Postfach 151150, D-66041 Saarbrücken, Germany. Phone: +49-172-8984640. Email: akumar@ifomis.uni-saarland.de URL: http://www.uni-leipzig.de/~akumar/