Bridging the gap between medical and bioinformatics: An ontological case study in colon carcinoma

Anand Kumar[a,*], Yum Lina Yip[b], Barry Smith[a,c], Pierre Grenon[a]

[a]*IFOMIS, University of Saarland, Saarbruecken, Germany*
[b]*University of Geneva, Department of structural biology and bioinformatics, Geneva, Switzerland Swiss Institute of Bioinformatics, Geneva, Switzerland*
[c]*Department of Philosophy, SUNY at Buffalo, USA*

Abstract

Ontological principles are needed in order to bridge the gap between medical and biological information in a robust and computable fashion. This is essential in order to draw inferences across the levels of granularity which span medicine and biology, an example of which include the understanding of the roles of tumor markers in the development and progress of carcinoma. Such information integration is also important for the integration of genomics information with the information contained in the electronic patient records in such a way that real time conclusions can be drawn. In this paper, we describe a large multi-granular datasource built by using ontological principles and focusing on the case of colon carcinoma.
© 2005 Elsevier Ltd. All rights reserved.

1. Introduction

Medicine and biology have become ever more closely related in the last decades. While medicine benefits from advances in molecular biology, biology benefits from the use of clinical data to derive and test hypotheses. In light of the many studies pointing to an information gulf between those who discover new gene products and those who can exploit these discoveries in clinical contexts [1–7], however, it is important that researchers are able to understand the interrelations between the two fields. Thus, while it

* Corresponding author. Tel.: +49 172 8984640.
  E-mail addresses: akumar@ifomis.uni-saarland.de (A. Kumar), lina.yip@isb-sib.ch (Y.L. Yip), phismith@buffalo.edu (B. Smith), pierre.grenon@ifomis.uni-saarland.de (P. Grenon).

is certainly the case that the rapid development of bioinformatics has brought significant data integration at the level of genes and proteins, comprehensive integration of data at this molecular level of granularity with data at the clinico-pathological level is still lacking.

We have here chosen colorectal cancer as a case study and we seek to establish links connecting data pertaining to this disease to the molecular level via the integration of three ontologies—SNOMED-CT [8], foundational model of anatomy (FMA) [9] and the gene ontology (GO) [10]—with the UniProt [11] and OMIM [12] databases. None of these ontologies and terminologies is free of mistakes [13–17]; however, we here consider them as they stand in order to take the first step in building the needed integrated framework.

2. Methods

2.1. Disease representation and classification

Our treatment of colorectal carcinoma on the clinical level is based on SNOMED together with information taken over manually from deVita (*Principles of Oncology*) [18] and *Harrison's Principles of Internal Medicine* [19]. A range of aspects were incorporated into the classification, including the staging of diseases according to the TNM, Duke's, and Modified Asler–Coller criteria, patient screening information, risk factors such as pathological predisposing factors, chemical exposure, family history and age, localisation of the disease based on organ and organ system affected and on the tissues involved, pathology aspects such as gross and microscopic pathology, size, shape, extent, vessels or nerves involved, cell type, nuclear characteristics and staining.

Gross pathology was first mapped to carcinoma location on the basis of a representation of the anatomy of the colon at the organ system, organ, tissue, cell and subcellular levels of granularity using the FMA. Information regarding clinical procedures, extent of carcinoma, vascular invasion, and histological pathology was also added. In order to make the anatomical representation accessible to information retrieval software, relations such as *is-located-in*, *is-surrounded-by* were added to the *is-a* and *part-of* relations present in the FMA and GO's Cellular Component ontology, the latter was situated within the corresponding axis of the former.

2.2. Representation of biological processes

Representations of biological processes were taken over from GO and UniProt at the molecular level. GO was chosen as the most widely used ontology in the biosciences, UniProt (which includes Swiss-Prot, TrEMBL and PIR-PSD) [11] because it is the world's most highly annotated protein sequence database and because it contains high quality disease-related annotations [20]. We supplemented the annotations within UniProt with those present with the gene ontology annotation (GOA) project [21]. Two protein datasets were retrieved from UniProt:

1. A comprehensive manually curated human protein dataset consisting of over 10,000 proteins and 24,500 related GO terms.
2. A subset of 233 human proteins involved in colorectal carcinoma (with over 800 associated GO terms) obtained by querying the OMIM database.

Table 1
Relations between terms from GO's *molecular function* and *cellular component* axes

Molecular function	Cellular component	Weight
Protein binding	Cytoplasm	36
Zinc ion binding	Nucleus	43
Protein binding	Nucleus	45
Receptor activity	Integral to plasma membrane	56
G-protein coupled receptor activity	Integral to plasma membrane	70
DNA binding	Nucleus	100
Antigen binding	Extracellular	123
Transcription factor activity	Nucleus	171

3. A dataset of correspondences between genes in the OMIM database and GO terms established via Locuslink (http://www.ncbi.nih.gov/LocusLink).

GO consists of three separate ontologies—for biological processes (bp), molecular functions (mf), and cellular components (cc), respectively—between the nodes of which no ontological relations are defined. Thus GO does not record the fact that a given process is the *exercise of* a given function or that it *takes place at* a particular cellular location. In order to find such missing associations between GO terms, in particular those associated with colorectal carcinoma, we employed our human protein dataset and analyzed the GO terms used in annotations associated therewith. We used both a statistical approach and a probabilistic approach to address this issue.

2.2.1. Statistical approach

GO terms associated with all human proteins in the SwissProt database were separated into three tables, labelled bp, mf, and cc according to the GO axes to which they belonged. Wherever a single protein was annotated by GO terms belonging to two distinct axes, the corresponding terms were paired together and three new tables were created for the resulting cc-mf, mf-bp, and cc-bp pairs. Identical tuples present within these tables were then grouped together and associated with an occurrence count in such a way as to provide a measure of how often GO terms from two distinct axes are annotated together. Examples are given in Table 1.

2.2.2. The probabilistic approach to association rule induction using the a priori algorithm

The a priori algorithm [22] is the most often used approach to the mining of associations. It consists of two steps. In the first, *frequent item sets* are determined by searching the lattice formed by all the subsets of the total set of available items. In a second step, the strongest association rules were determined from the frequent item sets. As there are more than 2500 GO terms that may be used to annotate relevant gene products, only association rules which satisfy predefined minimal *support* and *confidence* level are taken into consideration.

The *support* of an association rule is defined as the fraction of all transactions to which the rule is applicable.

The *confidence* of an association rule is defined as that fraction of those transactions to which the rule is applicable and in which the rule is also correct.

In the case of GO, we are interested in rules that predict cellular component given molecular function and biological process, or rules that predict molecular function given cellular component and biological process, and so on.

Frequent item sets were determined by searching the subset lattice of all GO terms which occur together as associations of some selected gene product. Association rules were constructed from the frequent item sets and filtered with respect to quality criteria.

An example of a rule is:

ribosome ← *ribosome biogenesis*; *protein biosynthesis* (0.2%, 93.2%).

This states that 0.2% of the total annotations (to GO terms originating from a specific database) annotate *ribosome biogenesis* and *protein biosynthesis* together, and that within those cases where they occur together, 93.2% are also annotated with the term *ribosome*.

3. Formal ontological principles

Our goal is to perform information integration across the disciplinary divide between medicine and molecular biology by developing an ontological framework which can allow multiple perspectives upon complex phenomena (such as diseases, associated risk factors, symptoms, pathological hallmarks, and so forth). For a given target disease, this ontological framework will allow us to search, for example, for other related diseases that have been identified as risk factors for the target.

A series of consensus ontological principles form the basis of our work, principles which are incorporated in different forms in a variety of top-level ontologies, including the DOLCE ontology [23]. Here, we adopt the version of these principles that is present in the basic formal ontology (BFO) [24,25], in which entities are divided into two non- overlapping categories of continuants and occurrents. To say that an entity is a continuant is to say that it enjoys continuous existence in time and is such as to preserve its identity through change. Occurrent entities, in contrast, have temporal parts or phases. They unfold themselves phase by phase, and exist only in these successive phases. Examples of continuant entities are: molecules, cells, organisms. Examples of occurrent entities are the processes in which such entities participate.

For a universal or class to be a subclass of (stand in as an *is-a* relation to) another universal or class, all instances of the subclass existing at any given time should be instances of the superclass existing at the same time. Thus we have:

lung **is-a** *lobular organ*,

because all instances of lung are instances of lobular organ. We can express part of what is involved in this relationship in the language of first-order logic, as follows:

R1: A **is-a** $B \rightarrow \forall x(\mathbf{inst}(x, A) \rightarrow \mathbf{inst}(x, B))$

Here '**inst**' abbreviates the relation of instantiation between an instance and a corresponding universal; $\forall$ is the standard universal quantifier signifying 'for all' or 'given any value of'; variables $x, y, z \ldots$ range over individuals and variables $A, B, C, \ldots$ range over universals.

R1 states that if A is-a B, then every instance of A is an instance of B. Some relations between classes need to be expressed not with the universal quantifier $\forall$ but rather with the existential quantifier $\exists$,

signifying 'there is a' or 'for some value of'. For example, in *lung tumor* in *lung*, we have a relationship which satisfies the following:

> R2: *A* **in** *B* → ∀*x*(**inst**(*x*, *A*) → ∃*y*(**inst**(*y*, *B*) & **in**(*x*, *y*)))

where '**in**' symbolizes the instance-level relation which obtains when one individual entity is located in another.

The relation of parthood between universals calls for yet another sort of treatment. For universals *A* and *B* the relation *A* part-of *B* holds (on our reading here) only when each instance of *A* is part of some instance of *B* and each instance of *B* contains an instance of *A* as part. For example, an instance of *ascending colon* is always part of an instance of *large intestine* and an instance of *large intestine* always contains an instance of *ascending colon* as part. This relation belongs to what is called *canonical anatomy* within FMA.

> R3: *A* **part-of** *B* = def ∀*x*(**inst**(*x*, *A*) → ∃*y*((**inst**(*y*, *B*) & **part**(*x*, *y*)))
> & ∀*y*((**inst**(*y*, *B*) → ∃*x*((**inst**(*x*, *A*) & **part**(*x*, *y*))

3.1. Introducing the factor of time and granularity

An instance of *lung* at the granularity of whole *organ* is a continuant. Thus it remains the same individual entity from one moment to the next, even though at lower levels of granularity it gains and loses parts (for example cells and molecules). To do justice to these facts we need to take time into account. Thus we write '**inst**(*x*, *A*, *t*)' to signify that *x* instantiates *A* at time *t*, and similarly '**part**(*x*, *y*, *t*)' to signify that *x* is an individual level part of *y* at time *t*.

We then have for example:

> R4: (**inst**(*x*, *colon*, *t*1) & **inst**(*x*, *colon*, *t*2)) → ∀*t*(*t*1 ⩽ *t* ⩽ *t*2 → **inst**(*x*, *colon*, *t*))

which asserts that if *x* is a *colon* at two distinct instants of time, then *x* is also a *colon* at all intervening points of time.

To capture the part relation:

> *mucosa of colon part-of colon*

We write:

> R5: ∀*x*∀*y*(**inst**(*x*, *mucosa of colon*, *t*) → ∃*y*(**inst**(*y*, *colon*, *t*) & **part**(*y*, *x*, *t*)))

R5 captures the fact that each instance of *mucosa of colon* is a part of some instance of *colon*. Note that this part relation spans two distinct levels of granularity, of *organ* and *tissue*, respectively.

Table 2
Tumor axis of the TNM classification for Colorectal carcinoma by the AJCC

TX: The primary tumor cannot be evaluated.
T0: The existence of a primary tumor cannot be ascertained.
Tis: Carcinoma in situ (tumor in place): an intraepithelial tumor or an invasion of the lamina propria. Tis includes cancer cells entirely contained within the glandular basement membrane (intraepithelial) or lamina propria (intramucosal) with no breach through the muscularis mucosae into the submucosa.
T1: The tumor invades the submucosa, the second layer of the large intestine.
T2: The tumor invades the muscularis propria.
T3: The tumor invades through the muscularis propria into the subserosa, or into nonperitonealized pericolic or perirectal tissues.
T4: The tumor directly invades other organs or structures, and/or penetrates the visceral peritoneum. In T4, the term 'direct invasion' includes invasion of other sections of colon or rectum by way of the serosa; for instance, invasion of the sigmoid colon by a carcinoma of the cecum.

Our scale of granularity recognizes six levels, for which we provide our preferred names and examples as follows:

organ system:	digestive system, respiratory system, nervous system
organ:	pharynx, esophagus, stomach, colon
organ part:	mucosa of colon, submucosa of colon
tissue:	maximal portion of epithelial tissue of colon
cell:	colon epithelial cell, fibrocytes
subcellular:	colon epithelial cell nucleus, colon epithelial cell membrane

We now introduce a function which associates with each universal or class the granularity level of the corresponding instances, for example, in

$$\mathbf{gr}(\textit{digestive system}) = \textit{organ system},$$
$$\mathbf{gr}(\textit{colon}) = \textit{organ},$$
$$\mathbf{gr}(\textit{mucosa of colon}) = \textit{tissue}.$$

For present purposes we assume that **gr** does indeed designate a genuine function (i.e. that each universal is associated with exactly one level of granularity).

We can now project the representation of structural pathologies onto the corresponding granular levels. The American Joint Committee on Cancer (AJCC) has designated staging by, means of its TNM classification [26,27], which classifies carcinomas according to: *tumor extent*, *lymph node involvement* and *metastasis* (Table 2).

Thus, while the T is stage deals with entities at the tissue, cell and subcellular levels of granularity, from T1 on the granularity is in every case at the tissue level.

We introduce the following abbreviation to indicate the granularity involved in a given case of instantiation:

R6: $\mathbf{inst}(x, A) =_{\mathrm{def}} \mathbf{inst}(x, A)$ & $\mathbf{gr}(x) = \mathrm{level1}$.

We also define a transgranular parthood relation, considering entity *A* belongs to *level* 1 of granularity and *B* to *level* 2:

R7: *A* **part-of** *B*$=_{\text{def}}$(**inst**(x, A) & **inst**(y, B) & **part**(x, y) & **gr**$(x) =$ level1 & **gr**(y)
$=$ level 2).

We can now define what it is for an entity to be a T1-stage carcinoma of colon structure as follows:

R8: **inst**$(x, T1$-*stage carcinoma of colon structure*) $=_{\text{def}} \exists y \exists z \exists v \exists w($**inst** $(y,$ *carcinoma of mucosa of colon structure*) & **inst**$(z,$ *carcinoma of submucosa of colon structure*) & **inst**$(v,$ *carcinoma of muscularis layer of colon structure*) & **inst**$(w,$ *carcinoma of serosa of colon structure*) & **part**(y, x) & **part**(z, x) & not **part**(v, x) & not **part**(w, x) & **gr**(*mucosa of colon structure*)=*organ part* & **gr**(*submucosa of colon structure*)= *organ part* & **gr**(*muscularis layer of colon structure*)=*organ part* & **gr**(*serosa of colon structure*)=*organ part*)

In order to make the formulas shorter we will not represent the granularity information in them from here on. We can create such formalizations for each of the stages listed in Table 3. Going one step further, we can relate the (normal) *mucosa of colon* with the carcinoma of mucosa of colon structure, since the latter is still the *mucosa of colon* but with an abnormal trait

R9: **inst**$(x,$ *carcinoma of mucosa of colon structure*)$=_{\text{def}} \exists y \exists z($**inst**$(x,$ *carcinoma*) & **inst**$(y,$ *mucosa of colon*) & **inst**$(z,$ *carcinomatous pathology of mucosa*) & **has-anatomical-extent**(x, y) & **has-pathology**(y, z)).

Here, **has-anatomical-extent** is the relation between a given carcinoma and the anatomical entity at which it is localized and **has-pathology** is the relation between a given anatomical entity and the type of pathology present therein. *Carcinomatous pathology of mucosa* can then be further specified at the tissue, *cellular* and *subcellular levels*. If it is an ulcerative adenocarcinoma we have

R12: **inst**$(x,$ *carcinomatous pathology of mucosa*)$=_{\text{def}} \exists y($**inst**$(x,$ *carcinomatous pathology*) & **inst**$(y,$ *ulceration with raised edges of mucosa*) & **has- pathological-feature**(x, y)).

The relation **has-pathological-feature** specifies the pathological characteristic that is at issue in a given pathology.

While *staging* exists at the *tissue*, *organ* and *organ system* levels, at the cellular and subcellular levels there exist *grades* for a carcinoma. A *grade* 2 *carcinomatous pathology* can be formalized by means of

R13: **inst**$(x,$ *Grade* 2 *carcinomatous pathology of mucosa structure*)$=_{\text{def}} \exists y($**inst** $(x,$ *carcinomatous pathology*) & **inst**$(y,$ *non-polar nucleus of mucosal epithelium*) & **has-pathological-feature**(x, y) & **gr**(*Grade* 2 *carcinomatous pathology of mucosa structure*)=*subcellular*).

In this way, we reach the *subcellular level* for normal and pathological structures at various levels of granularity. Since subcellular locations are more elaborately (albeit less formally) represented in GO than in the FMA, we map GO terms to the corresponding loci within the FMA in order to make a transgranular representation possible.

4. Formal-ontological relations

4.1. Transgranular part–whole relations

Various sorts of associations were found from our work on GO using statistical or probabilistic analyses, in addition to the associations discussed above between process and function terms and between location and process terms.

Relations crossing granularity levels are pre-eminently relations of part and whole. Using our method for association rule induction, we were able to find the following transgranular part–whole relations not currently present in GO:

> *transposase activity* **part**-**of** *DNA transposition*
>
> *development* **part**-**of** *nucleus function* (from GO's nucleus)
>
> *apoptosis* **part**-**of** *cytoplasmic function* (from GO's cytoplasm)

GO does not clearly distinguish between functions (continuants) and processes (occurrents). When this distinction is taken into account, however, then it becomes possible to distinguish two sorts of transgranular relations:

> *transposase function* **part**-**of** *DNA transposition function*
>
> *transposase activity* **part**-**of** *DNA transposition activity*

Where entities are related across granularities, problems can arise when one attempts to use these relations to derive inferences without taking granularity into account. This is especially the case when a process of coarser granularity, for example *development*, is associated with an anatomical entity such as *nucleus*, which is of finer grain. One does not know which is the development considered here, that of a cell in a mammal or that of a cell in a unicellular organism or psychological development. In order to use our formalism to bridge the granularity gulf between these entities, we will need to distinguish those relations which are relevant in a given context and then find association rules involving the nucleus together with these smaller parts.

The approach we take appeals to the notion of *involvement*, defining an operator *involved-in*, which allows us to create general terms of the form *A involved-in B*, which refer precisely to those instances of *A* which are involved in some instance of *B*. Here *A* is a smaller process that can (but need not) occur as part of an instance of the larger process *B*. *A involved-in B* then designates the class of all and only those instances of *A* which are part of instances of *B*.

> R14: **inst**$(x, A \text{ involved-in } B) =_{\text{def}}$ **inst**(x, A) & $\exists y($**inst**(y, B) & **part**$(x, y))$

Similarly, the larger process *B* may or may not have an instance of the smaller process *A* occurring as part. We again declare a universal *B involving A*, which is a specification of *B* containing only those instances of *B* which have instances of *A* as part:

> R15: **inst**$(x, B \text{ involving } A) =_{\text{def}}$ **inst**(x, B) & $\exists y($**inst**(y, A) & **part**$(y, x))$

We can then assert that in every case:

> R16 : *A involved-in B* **part**-**of** *B involving A*

Example. *hexokinase* 1 (KEGG K00844, EC 2.7.1.1) *activity* is involved in glycolysis, fructose and mannose metabolism, galactose metabolism, starch and sucrose metabolism and in aminosugar metabolism. Thus we can create universals such as

> *hexokinase* 1 *activity involved in glycolytic pathway*
>
> *hexokinase* 1 *activity involved in galactose metabolism pathway*

and so on.

We next sought to establish transgranular associations between processes and anatomical structures, for example between fine-grained processes like *p53 activity* with a coarse-grained anatomical entity like the *colon*. Once those associations are established, relations can be more easily formalized by appealing to the fact that a robust anatomical ontology already exists in the form of the FMA. For example, in the *T*1 *stage of colon carcinoma* at the *organ level* of granularity, a *carcinomatous process* projected onto the *colon*. Since we cannot observe the detailed tissue structure of the colon at this level of granularity, our formula reads as follows:

> R17: **inst**(x, *T*1-*stage carcinomatous process* of *carcinoma of colon*)$=_{\text{def}}\exists y$(**inst**(x, *carcinomatous process*) & **inst**(y, *colon*) & **has-anatomical-projection**(x, y))

When we move to the lower level of granularity of the tissue, however, we can project the *carcinomatous process of T*1-*stage of colon* onto its *mucosa* and *submucosa*

> R18: **inst**(x, *T*1-*stage carcinomatous process of carcinoma of colon*)$=_{\text{def}}\exists y\exists z\exists w\exists w$ (**inst**(x, *carcinomatous process*) & **inst**(y, *mucosa of colon*) & **inst**(z, *submucosa of colon*) & **inst**(v, *muscularis layer of colon*) & **inst**(w, *serosa of colon*)) & **has-anatomical-projection**(x, y, z)

If we then put R12 and R13 together

> R19: **inst**(x, *T*1-*stage carcinomatous process of carcinoma of colon*)$=_{\text{def}}\exists y\exists z\exists v\exists w\exists u$ (**inst**(x, *carcinomatous process*) & **inst**(y, *mucosa of colon*) & **inst**(z, *submucosa of colon*) & **inst**(v, *muscularis layer of colon*) & **inst**(w, *serosa of colon*) & **inst**(u, *colon*) & **has-anatomical-projection**(x, y, z) & **part**(y, u) & **part**(z, u) & **part**(v, u) & **part**(w, u))

This means that within the *T*1 *stage of the carcinomatous process* of *carcinoma of colon*, and at the *tissue level* of granularity, the *carcinomatous process* can be projected onto the *mucosa* and *submucosa* of *colon*; and that the *mucosa*, *submucosa*, *muscularis layer* and *serosa* of the *colon* existing at the *tissue level* of granularity are parts of the *colon*, an *organ* whose instances are observed (from the perspective of this staging) at both the *tissue* and the *organ level* of granularity. In this way, we are able to represent a transgranular part-whole relation which zooms from the organ level to the tissue level.

4.2. Relations from substances to processes

The relation of participation is a species of dependence; it is instantiated wherever an independent entity such as an organism is involved as agent or patient in a process. There are different kinds of participation, which we can order along the dimensions as shown in Fig. 1.
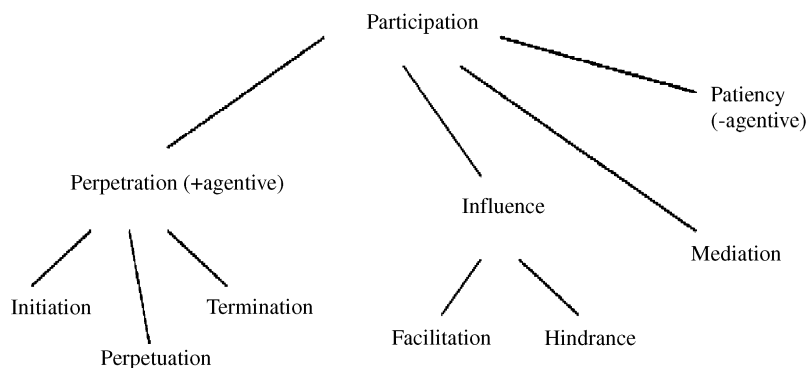
Fig. 1. Modes of participation [25].

4.2.1. Perpetration

The most important subtype of participation is that of perpetration. A substance perpetrates (does, performs) an action (direct and agentive participation in a process). Perpetration can be of three subtypes: Initiation, Perpetuation and Termination. To represent perpetration formally we need to distinguish the substance which performs the relevant action, the action itself—in other words the process in which the substance features as agent—and the time interval during which the action takes place (z).

4.2.2. Initiation

When it is stated that a substance, for example a cellular component, initiates a process, this means that the process was not occurring before the initiation took place. Many of the association links induced in the course of our experiment include such situations.

Example. *Electron transporter* initiates the process of *electron transfer activity*. However, since GO does not include the relevant substance/component term (e.g. *electron transportor*) but only a term for the related activity (e.g. *electron transporter activity*). Thus we cannot directly define the corresponding relations—for example the relations between *enzymes* (substance) and functions within GO. In order to compensate for the missing GO terms, we supplement a relation such as

> *electron transporter activity* **initiates** *electron transport*

by including also the association rule

> *electron transport* ← *electron transporter*

and similarly for other cases in what follows. A *transporter* is a continuant, which exists even before it exerts the *function of electron transport* and thereby initiates a *process* of *electron transport*. We need to add this information into the association rules manually.

Initiation is a relation between universals satisfying for example

R21: *electron transporter* **initiates** *electron transport* $\rightarrow$ $\exists x, y$ (**inst**(x, *electron transporter*, t) & (**inst**(y, *electron transport*, t) & **initiates**(x, y, t))

Here *x* and *y* are substance and process instances and the initiation process takes place at the time instant *t*. Initiation is a kind of perpetration, which satisfies:

R22: **initiates**$(x, y, t) \rightarrow ($**perpetrates**(x, y, t) & $\forall t'(t' < t \rightarrow$ not **occurs**$(y, t'))$

Relevance to colon carcinoma: Many *compounds* (for example, *thiacarbocyanine compounds*) act against the initiation of *electron transporter activity* in such a way as to have cytotoxic effects on *human carcinoma cells*. These are the building blocks which can be used to represent the flow of pathophysiological processes which occur in carcinomas like that of colon.

4.2.3. Perpetuation

This relation obtains when substance perpetuates (sustains) a process. Perpetuation normally presupposes that an entity existed at some earlier time which entered in the relation of initiation with the process in question. However, perpetuators are of course not of necessity themselves initiators.

Example. The *subtilase enzyme* perpetuates the process of *proteolysis* and *peptidolysis*. It helps the process to continue once it has been started. In some cases, *subtilase* can also help in the initiation of *peptidolysis*. The same instance of subtilase can initiate a *peptidolysis* at a particular instance of time and perpetuate it at the other.

The association algorithm relates the entities below

proteolysis and peptidolysis $\leftarrow$ subtilase activity,

Since GO does not contain *enzymes* but only their *activities*, this must be supplemented with

proteolysis and peptidolysis **perpetuates** *subtilase*

We can then write

R23: *subtilase* **perpetuates** *proteolysis and peptidolysis* $\rightarrow \exists x, y($**inst**$(x,$ *subtilase*$, t)$ & **inst**$(y,$ *proteolysis and peptidolysis*$, t)$ & **perpetuates**$(x, y, t))$

Here **perpetuates** is a relation on the instance level which satisfies for example axioms to the effect that

R24: **perpetuates**$(x, y, t) \rightarrow ($**perpetrates**(x, y, t) & not $\forall t'(t' > t \rightarrow$ not **occurs**$(y, t'))$ & not $\forall t''(t'' < t \rightarrow$ not **occurs**$(y, t''))$

Relevance to colon carcinoma: *Subtilase SKI*-1 cleaves *proproteins* at non-basic residues. A *proteinase K-like subtilase*, *neural apoptosis-regulated convertase* 1 (NARC-1) has one of the highest expressions within *colon carcinoma LoVo-C*5 *cell lines*. Such perpetuations are present throughout the colon carcinoma's pathophysiological pathways.

4.2.4. Termination

A substance terminates (brings to an end) a process.

Example. *Carbon-monoxide oxygenase* is an *enzyme* which acts on an occurring *electron transport process* and terminates this process. The relevant induced association rule is:

> *electron transport* ← *carbon-monoxide oxygenase activity*

which we supplement with:

> *electron transport terminates carbon-monoxide oxygenase*

Terminates is a relation on the instance level which satisfies axioms such as:

> R26: **terminates**$(x, y, t) \rightarrow ($**perpetrates**$(x, y, t)) \,\&\, \forall t'(t' > t \rightarrow$ not **occurs**$(y, t'))$

The above formula asserts that to terminate a process at t is to bring it about that the process does not occur at any time later than t.

Relevance to colon carcinoma: The termination of *electron transport* has cytotoxic effects on *human carcinoma cells*.

4.2.5. Influence
A substance has an effect on a process.

Example. The peptide actin cytoskeleton helps in protein binding.

> *protein binding* **influences** *actin cytoskeleton*

In a similar manner to the cases treated above, this can be formulated as:

> R27: *actin cytoskeleton* **influences** *protein binding* $=_{\text{def}} \forall x \,($**inst**$(x, actin\ cytoskeleton) \rightarrow \exists y \,($**inst**$(y, protein\ binding) \,\&\,$ **influences**$(x, y, t))$

Relevance to colon carcinoma: By binding to ATP and its hydrolysis, *actin cytoskeleton* influences binding to *antibodies* (*proteins*) *like TCP* 22, a monoclonal marker for colon carcinoma.

4.2.6. Facilitation
Facilitation occurs where a substance plays a secondary role in a process.

Example. The *large and small ribosomal units* facilitate protein synthesis by helping to bring together *mRNA* with *tRNAs*. The mRNA is sandwiched between the two subunits and then the codon-anticodon match takes place between mRNA and tRNA. The association rule induced is

> *large ribosomal subunit* **facilitates** *protein biosynthesis*

Relevance to colon carcinoma: Protein biosyntheses and expression involving ribosomes play a major role in the pathophysiology of colon carcinoma.

4.2.7. Hindrance, prevention

This relation obtains where a substance has a negative effect on the unfolding of a process.

Example. *Transposase* not only initiates *DNA transposition*, it also causes an auto-inhibition at a later stage. Auto-inhibition takes place where the substance which initiates a reaction or process or which was produced as a product of this process itself prevents the process from continuing further. The association rule induced is

>   *DNA transposition ← transposase activity*

Relevance to colon carcinoma: *Transposase activities* are involved in *inflammatory bowel disease*, one of the predisposing factors to colon carcinoma and provide some evidence that the carcinoma could have an infective component.

4.3. Relations between processes and substances

The above relations were exposed via our association rule induction as obtaining *from* substances *to* processes, but relations were also revealed which point in the converse direction, from processes to substances. Involvement is the most general form of relations of this type between a process and its bearers. Its subtypes are creation, substaining in being, and degradation.

4.3.1. Creation

A process brings into being a substance.

Example. *Beta-hydroxysteroid dehydrogenase activity* creates *steroids*. The relation of **initiation** holds between a substance and a process, the relation of creation holds in the converse direction: here a substance (a *steroid*) is brought into existence by a process and continues to exist for at least a small time interval thereafter. We supplement

>   *steroid biosynthesis ←* 3(*or* 17)*beta-hydroxysteroid dehydrogenase activity*

with

>   3(*or*17)*beta-hydroxysteroid dehydrogenase activity* **creates** *steroid*

4.3.2. Sustaining in being

A process sustains a substance in existence for a certain period of time.

Example. Protease inhibitor activity maintains a membrane structure.

>   *protease inhibitor activity* **sustains in being** *membrane*

Relevance to colon carcinoma: *Protease inhibitor activity* protects the colon cells from the actions of *protein kinase C inhibitors like* 7-*hydroxystaurosporine*; it contributes to the resistance to apoptosis of the colon cells.

4.3.3. Degradation

This relation obtains when a process has negative effects upon a substance.

Example. When the activity of lyase enzyme causes a degradation of the D-amino-acid dehydrogenase complex.

> *lyase activity* **degrades** *D-amino-acid dehydrogenase complex*

4.4. Transgranular relations revisited

Once the associated relations have induced and formalized in the way described above, our strategy is to project the processes onto the locations (usually subcellular) where they take place and then to zoom up from the subcellular organelles thus identified to the relevant overarching entities on the cellular, tissue, organ, organ system and organism levels. In this way the entities at the finest grains represented within bioinformatics will become formally linked to the entities at the coarser grains standardly represented within medical informatics. An example of such a derivation is

R28: **inst**(x, T1-*stage carcinomatous process of carcinoma of colon*)$=_{\text{def}}$ $\exists y \exists z \exists v \exists w \exists u \exists p \exists q \exists r \exists s$ (**inst**(x, *carcinomatous process*) & **inst**(y, *mucosa of colon*) & **inst**(v, *submucosa of colon*) & **inst**(v, *muscularis layer of colon*) & **inst**(w, *serosa of colon*) & **inst**(u, *colon*) & **has anatomical projection**(x, y, z, v, w) & **part**(y, u) & **part**(z, u) & **part**(v, y) & **part**(w, u) & **inst**(p, *epithelium of mucosa of colon*) & **part**(p, y) & **inst**(q, *epithelial cell of mucosa of colon*) & **part**(q, p) & **inst**(r, *nucleus of epithelial cell of mucosa of colon*) & **part**(r, q) & **inst**(s, p53 *activity in nucleus of epithelial cell of mucosa of colon*) & **has spatial projection**(s, r))

The above formula represents four levels of granularity: that of subcellular, cell, tissue and organ levels. It represents the fact that the mucosa, submucosa, muscularis layer and serosa of colon exist within the colon at the tissue level of granularity and are parts of the colon, which itself is an organ. The representation thus allows us to move from the lower granularity of organ to the higher granularity of tissue in order to observe the transgranular part-relations involved within the entire colon structure. The formula represents more precisely the fact that T1 stage carcinomatous processes are projected onto the mucosa and submucosa of colon. It further represents the parts of the mucosa of the colon at the cellular level of granularity (epithelial cell) and at the subcellular level (nucleus of the epithelial cell). It then projects the p53 activity onto the nucleus of the epithelial cell, again at the subcellular level of granularity. Thus the formula not only represents where the T1 carcinomatous process takes place, but also how one of the key processes at the subcellular levels can be related to it.

5. Ontological representation

We employed the Protégé 2.0 framework (http://protégé.stanford.edu/), a knowledge-base-creating and editing tool, to create our ontologies. Protégé 2.0 was chosen because its frame-based architecture provides a tractable framework for the representation of the sorts of complex relationships which fall within our area of concern. While there are problems in representing some of the predicate-logic-based constructs in Protégé, some of the features it contains, including the Description Logic-based OWL editor

(http:77www.w3.org/TR/owl-ref/) make it useful for certain forms of reasoning. The relations between entities at various levels of granularity are represented in the ontologies, together also with relations which exist between the various axes.

6. Conclusion

Our approach still has several limitations. By connecting proteins to biological processes and establishing relationships of a type not represented in GO between biological processes, molecular functions and cellular components, we aim to provide a unified representation which can help in answering the difficult questions which arise at the borders of medical and bioinformatics. This does not mean that we are aiming to simulate the actual dynamic behavior within and between the entities here surveyed. This is because we are interested in the first place in the ways in which particulars instantiate specific *qualitative* categories rather than in the *quantitative* correlations which can be associated with such instantiations. Such qualitative categories must in any case be specified, together with the kinds of (qualitative) relations which can obtain between them, before measures can be assigned to the terms of these relations. Ontologies by themselves cannot represent the details of dynamic behavior. What they can do is to provide a robust representation which can form a framework in whose terms quantitative models of dynamics can be built - ideally in such a way as to avoid the characteristic defects of such models, namely that they focus on a narrow range of levels of granularity, primarily the micro-levels, where numerical values can most easily be assigned.

Several challenges still need to be overcome.

First, how are we going to maintain and align (and in some cases build) the ontologies for anatomy, physiology, pathology, molecular functions, and pathways in an environment where we are witnessing a constant increase in the amount of information, both in the basic biomedical sciences and clinical medicine? A collaborative effort is certainly needed. We believe that a formal ontology based on sound principles can provide valuable support for such an effort.

Second, how can we derive accurate and reliable association rules? Currently, manual inspection is required. However, it should, in the future, be possible to develop software tools with the capacity to estimate automatically the confidence level of each association.

Finally, if robust ontologies can ease data integration and information retrieval, can they be made powerful enough to help derive and test new hypotheses? The answer to this question depends on the extent to which ontologies can be used to represent in adequate fashion highly complex biological phenomena.

We are attempting to tackle each of these problems in order to make the ontology capable of serving as a real bridge between medicine and molecular biology. The recently created Human Interaction Database [28], together with IntAct [29] and IntEnz [30], combine information pertaining to interactions between human proteins (taken from the Database of Interacting Proteins [31] and many other sources) with the associated structural and functional information about the proteins themselves (taken from the Structural Classification of Proteins [32] and the Protein Data Bank [33]). Integration of information from such databases will provide further insights in bridging the gap between medical informatics and bioinformatics based on formal ontological principles.

Acknowledgements

References

[1] F. Martin-Sanchez, I. Iakovidis, S. Norager, V. Maojo, P. de Groen, J. Van der Lei, T. Jones, K. Abraham-Fuchs, R. Apweiler, A. Babic, R. Baud, V. Breton, P. Cinquin, P. Doupi, M. Dugas, R. Eils, R. Engelbrecht, P. Ghazal, P. Jehenson, C. Kulikowski, K. Lampe, G. De Moor, S. Orphanoudakis, N. Rossing, B. Sarachan, A. Sousa, G. Spekowius, G. Thireos, G. Zahlmann, J. Zvarova, I. Hermosilla, F.J. Vicente, Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care, J. Biomed. Inform. 37 (1) (2004) 30–42.

[2] V. Maojo, C.A. Kulikowski, Bioinformatics and medical informatics: collaborations on the road to genomic medicine?, J. Am. Med. Inform. Assoc. 10 (6) (2003) 515–522.

[3] A.M. Grant, A.M. Moshyk, A. Kushniruk, J.R. Moehr, Reflections on an arranged marriage between bioinformatics and health informatics, Methods Inf. Med. 42 (2) (2003) 116–120.

[4] J. Wiemer, F. Schubert, M. Granzow, T. Ragg, J. Fieres, J. Mattes, R. Eils, Informatics united: exemplary studies combining medical informatics, neuroinformatics and bioinformatics, Methods Inf. Med. 42 (2) (2003) 126–133.

[5] F. Martin-Sanchez, V. Maojo, G. Lopez-Campos, Integrating genomics into health information systems, Methods Inf. Med. 41 (1) (2002) 25–30.

[6] I.S. Kohane, Bioinformatics and clinical informatics: the imperative to collaborate, J. Am. Med. Inform. Assoc. 7 (5) (2000) 512–516.

[7] R.B. Altman, The interactions between clinical informatics and bioinformatics: a case study, J. Am. Med. Inform. Assoc. 7 (5) (2000) 439–443.

[8] K.A. Spackman, SNOMED-CT milestones: endorsements are added to already-impressive standards credentials, Health Inform. 21 (9) (2004) 54, 56.

[9] C. Rosse, J.L.V. Mejino, A reference ontology for bioinformatics: the foundational model of anatomy, J. Biomed. Inform. 36 (2003) 478–500.

[10] The Gene Ontology Consortium. The gene ontology (GO) database and informatics resource, Nucleic Acids Res. 32 (2004) D258–D261.

[11] R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, L.S. Yeh, UniProt: the Universal Protein knowledgebase, Nucleic Acids Res. 32 (2004) D115–D119.

[12] Online Mendelian Inheritance in Man, OMIM (TM). McKusick–Nathans Institute for Genetic Medicine, Johns Hopkins University, Baltimore, MD, and National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, 2000. World Wide Web URL: (http://www.ncbi.nlm.nih.gov/omim/)

[13] A. Kumar, B. Smith, Towards a proteomics metaclassification, IEEE Fourth Symposium on Bioinformatics and Bioengineering, Taichung, Taiwan. IEEE Press, New York, 2004, pp. 419–427.

[14] B. Smith, J. Koehler, A. Kumar, On the application of formal principles to life science data: a case study in the gene ontology, in: DILS 2004, Leipzig, Lecture Notes in Bioinformatics, vol. 2994, 2004, pp.79–94.

[15] A. Kumar, B. Smith, The Unified Medical Language System and the Gene Ontology: Some Critical Reflections, Lecture Notes in Computer Science, vol. 2821/2003, 2003, pp. 135–148.

[16] B. Smith, J. Williams, S. Schulze-Kremer, The ontology of the gene ontology, in: Proceedings of the AMIA, 2003, pp. 609–613.

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

[18] V.T. DeVita, S. Hellman, S.A. Rosenberg, Cancer: Principles and Practice of Oncology, sixth ed., Lippincott, Williams & Wilkins, Philadelphia, PA, Baltimore, MD, 2001.

[19] E. Braunwald, A.S. Fauci, D.L. Kaspere, S.L. Hauser, D.L. Longo, J.L. James, Harrison's Principles of Internal Medicine, 15th ed., McGraw-Hill Professional Publishing, New York, 2001.

[20] Y.L. Yip, H. Scheib, A.V. Diemand, A. Gattiker, L.M. Famiglietti, E. Gasteiger, A. Bairoch, The Swiss–Prot variant page and the ModSNP database, Hum Mutat. 23 (5) (2004) 464–470.

[21] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, R. Apweiler, The gene ontology annotation (GOA) database: sharing knowledge in uniprot with gene ontology, Nucleic Acids Res. 32 (2004) D262–D266.

[22] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: Proceedings of the International Conference on Very Large Databases, Santiage, Chile, Morgan Kaufmann, Los Altos, CA, 1994, pp. 478–499.

[23] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, Ontology Library (final), WonderWeb deliverable D18, December 2003.

[24] P. Grenon, B. Smith, SNAP and SPAN: towards dynamic spatial ontology, Spatial Cognition and Computation 4 (1) (2004), forthcoming.

[25] B. Smith, P. Grenon, The Cornucopia of formal-ontological relations, Dialectica, 58(3) (2004) 279–296.

[26] Anon, TNM system, J. Am. College Surg. 181 (1995) 182–188.

[27] J.W. Yarbro, D.L. Page, L.P. Fielding, et al., American joint committee on cancer prognostic factors consensus conference, Cancer 86 (11) (1999) 2436–2446.

[28] K. Han, B. Park, H. Kim, J. Hong, J. Park, HPID: the human protein interaction database, Bioinformatics 20 (15) (2004) 2466–2470.

[29] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, R. Apweiler, IntAct: an open source molecular interaction database, Nucleic Acids Res. 32 (2004) D452–D455.

[30] A. Fleischmann, M. Darsow, K. Degtyarenko, W. Fleischmann, S. Boyce, K.B. Axelsen, A. Bairoch, D. Schomburg, K.F. Tipton, R. Apweiler, IntEnz, the integrated relational enzyme database, Nucleic Acids Res. 32 (2004) D434–D437.

[31] L. Salwinski, C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie, D. Eisenberg, The database of interacting proteins: 2004 update, Nucleic Acids Res. 32 (2004) D449–D451.

[32] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, J. Mol. Biol. 247 (1995) 536–540.

[33] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.

Anand Kumar is Alexander von Humboldt researcher at the Institute for Formal Ontology and Medical Information Science in Saarbrücken, Germany. He has a medical background with a Ph.D. in medical informatics. He has worked in various European Union projects and has authored over 40 publications over the last four years. Anand's main interest lies in the usage of formal ontological principles in biomedical ontologies and electronic health records. For further information see: http://www.uni-leipzig.de/~akumar/

Yum Lina Yip has a Ph.D. in Medicine and a MS in Bioinformatics. Before joining the Swiss-Prot group in 2002, she worked as a research officer in the Molecular and Cellular Oncology group in St. Vincent's Hospital (Sydney, Australia) where she participated in several cancer researh projects as well as in a Phase I cancer vaccine clinical trial. In Swiss-Prot, she has developed the ModSNP database and is conducting a research project aiming to understand the effect of single amino acid polymorphisms on protein structure and function. Since December 2003, she has collaborated with Dr. Anand Kumar and Prof. Barry Smith from IFOMIS to investigate the use of ontology for clinical and biomedical information integration.

Barry Smith is Julian Park Distinguished Professor of Philosophy in the University at Buffalo (New York, USA) and Director of the Institute for Formal Ontology and Medical Information Science in Saarbrücken, Germany. He is the author of some 450 scientific publications, including 15 authored or edited books, and editor of The Monist: An International Quarterly Journal of General Philosophical Inquiry. His research has been funded by the National Institutes of Health, the US, Swiss and Austrian National Science Foundations, the Volkswagen Foundation, and the European Union. In 2002 he received in recognition of his scientific achievements the Wolfgang Paul Award of the Alexander von Humboldt Foundation. His current research focus is

ontology and its applications in biomedicine and biomedical informatics, where he is working on a variety of projects relating to biomedical terminologies and electronic health records. For further information see: http://www.ontology.buffalo.edu/smith

Pierre Grenon has degrees in logic and philosophy from the Univeristy of Paris 1 and has worked professionally as an Ontologist for Cycorp (Texas). Currently, he is a researcher at the Institute for Formal Ontology and Medical Information Science (Saarbruecken, Germany) and is completing a Ph.D. in analytic philosophy at the University of Geneva. His research interests include formal and spatiotemporal ontology and the metaphysics of relations.