Applying partitions to infectious diseases

C. Maria Keet[a], Anand Kumar[b]

[a] KRDB Research Centre, Department of Computer Science, Free University of Bozen-Bolzano, Italy
[b] IFOMIS, University of Saarland, Saarbrücken, Germany

Abstract

Due to exponential growth in biological data, partitioning is a necessity to manage data and to use it for reasoning mechanisms to deduce and infer new information automatically. Starting from a categorisation of infectious diseases, we formalise the relations of relevant partitions, such as mode of transmission, pathological process, and infectious organism, with examples for pneumonia and cholera to illustrate usage of partitioning and the relations between the granular levels.

Keywords: Infectious diseases; Pneumonia; Cholera; Medical Informatics

1. Introduction

Existing databases and information repositories about infectious disease, such as PathInfo [1], NCID [2] and the Encyclopedic Reference of Parasitology [3], do provide more or less comprehensive access, but it is still up to the researcher to manually combine data to discover commonalities and patterns among pathogens. Understanding of the mode of action of pathogens can lead to better understanding of human physiology, as e.g. with *Listeria* infections and nucleation of actin filament polymerization [4, 5], that in turn supports advances in medicine. From an informatics perspective, this requires an emphasis on creating ontologies and applying biological partitions and granularity to add additional formalised structure to this information to enable development of more useful software applications. With the exponential growth in biological information, dealing with partitions and their levels of granularity it is a necessity to be able to manage this data and to use it for reasoning mechanisms to verify existing information, deduce new information automatically, and generate research hypotheses. Here, we address the aspect of granularity and apply it to the subject domain of infectious diseases. This partitioning allows us to abstract away details or more encompassing situations in certain contexts and to convert this into a manner usable for computer applications.

The methodological approach taken is that of a bottom-up structuring of the domain in section 2, supplemented with top-down advances by building upon previous work on ontologies and theory of granularity (section 3).

2. A partitioning of infectious diseases

We reviewed the infectious disease axis of Snomed CT and ICD10, and biology of infectious organisms [2, 3, 6-10]. One can partition infectious diseases along many different di-

mensions, or granular perspectives, of which some examples are included in *Table 1*. Here, we discuss informal decisions on partitioning before addressing the formalisations in §3.

Table 1. Partitions with some examples for each level.

Dimensions		Level 1	Level 2-3	
Source	Mode of Transmission	Air-borne, Food-borne, Water-borne, Direct contact	Direct Contact: Person-to-person, Animal-to-person (zoonoses)	Person-to-person: Sexual intercourse, Skin, Blood
			Food-borne: Production, Preservation, Preparation	
Site	Site of entry	Respiratory system, Digestive system	Digestive system: Stomach, Duodenum, Colon	
	Site of effect	Respiratory system, Digestive system	Digestive system: Stomach, Duodenum, Colon	
Infectious organism	Common name	Multi-cellular animal Worms and flukes Arthropods Micro-organism Protozoa Fungi and moulds Bacteria	Worms and flukes: Roundworms, Hookworms, Tapeworms, Threadworms	
			Fungi and moulds: Amoebae, Fungi	
			Bacteria: Gram-negative, Gram-positive, Cocci, Rod, Flagellate	Cocci: Mono, Di, Strepto, Staphylo
	Phylogeny	*Prokaryote* *Eubacteria* *Eukaryote* *Mycota* *Protozoa* *Metazoa* *Trypanosomatidae* *Ancylostomatidae*	*Eubacteria: Salmonella* spp., *Aeromonas* spp.	*Salmonella* spp.: *S. enteritidis, S. typhi*
			Mycota: Myxomycetes, Phycomycetes, Eumycetes	
			Trypanosomatidae: Leishmania spp., *Trypanosoma* spp.	*Leishmania* spp.: *L. braziliensis, L. donovani, L. tropica*
			Ancylostomatidae: Ancylostoma duodenale, Necator americanus	
Disease classification	Infectious disease	Infectious disease: Dysentery, Pneumonia, Meningitis	Pneumonia: Lobar pneumonia, Segmental or lobular pneumonia, Bronchopneumonia, Interstitial pneumonia	
Pathology	Mode of action	Toxin-producer, Genetic interference	Toxin-producer: Stimulator, Inhibitor	Inhibitor: Covalent binding of the small subunit of the cholera toxin to the G-protein of the Second Messenger System, Covalent modification by pertussis toxin of inhibitory G_i protein that blocks inhibition of adenylate cyclase of the Second Messenger System
	Path. structure	Marbled parenchyma, White cicatricial tissue	White cicatricial tissue: Dense collagen connective tissue with reduced cell density	
	Path. process	Inflammatory process, Proliferative process	Inflammatory process: Congestion, Red hepatisation, Grey hepatisation, Resolution	Congestion: Serous exudation, Vascular engorgement, Rapid bacterial proliferation
Predisposing factors		Living habits, Hereditary, Environment, Age	Living habits: Diet, Smoking, Stress, Personal hygiene	

Note: the *Congestion* example for the *Inflammatory process* in *Pathological process* applies to lobar pneumococcal pneumonia [24].

Analysing the site of entry and effect of an infectious agent requires deploying a combination of human anatomy with levels of granularity dividing the human organism, thereby allowing distinguishing between e.g. the respiratory system and alveoli. For infectious diseases it is important to represent the kind of infectious agent, e.g. to avoid incorrectly administering an antibiotic treatment for a viral infection. Taking a speciesist approach, a complete or condensed version of the phylogenetic tree is preferable, although in a medical setting it may be less relevant where more intuitive 'common names' suffice. Informing a patient s/he is infected with a hookworm results in more effective communication than

"you are hosting some *Ancylostomatidae*", while for treatment purposes it is relevant to know if the hookworm is of the kind *Ancylostoma duodenale* or *Necator americanus*. A disadvantage of using common distinctions is that for example protozoa belong to both the group of micro-organisms and animals. A drawback of using the phylogenetic tree is that viruses are not organisms; hence do not occur in the tree (see e.g. [11]). Partitioning the mode of action faces the problem of lack of ample scientific knowledge for many infections to model general principles. However, this is also an interesting challenge, where ontology and biomedicine can work together to take advantage of the power of reasoning using formal ontology and a theory of granularity to make inferences and generate new research hypotheses. At its finer-grained levels, one can reuse existing biochemical, metabolome and other cell-level facets categorised in for example the Gene Ontology [12]. [13] Describes informal levels of granularity of treatments and [14] levels of canonicity, which are not addressed further in the remainder of this article. Causality is subject to much philosophical debate, both from an ontological as from a biological perspective (e.g. [15]) and therefore not further discussed here.

3. Formalising partitions

3.1 Granularity

Granularity deals with dividing something (that may comprise multiple entities) hierarchically according to certain criteria, called the *granular perspective* or *dimension*, where a lower level contains entities or knowledge that is more detailed than the adjacent higher-level grain. A *granular level*, also referred to as *granular partition*, contains one or more entities, which are indistinguishable [16, 17] at a higher, more coarse-grained, level. Each level is partitioned alike a grid with cells or list of items, which may, or may not, be exhaustive for the kinds of stuff found at that level. To be able to specify completely all levels and, more importantly, the entities at a given level, i.e. fullness and cumulativeness [18], is realistically not feasible in the biological sciences. Although using partitioning increases complexity initially, requires training of staff collaborating in creating such partitions and the current state of available software to create such partitions and levels of granularity is limited. Nevertheless, the benefits eventually will outweigh the initial efforts. Advantages of partitioning include primarily tighter data integration in life-sciences, both within the biological aspects of infectious disease and between biology and medicine, and improved inferencing because of its formal foundations that can highlight lacunas and inconsistencies in biological and biomedical knowledge, which in turn can aid in hypotheses formulation. From the perspective of ontology research, it will be easier to maintain ontologies and find inconsistencies because of the addition layer of knowledge about the subject domain that is not captured in e.g. structured controlled vocabularies.

When partitioning a granular level, it may consist only of a list of entities where each level is either fully divided (disjoint exhaustive) or with an (ontologically incorrect) entity *Everything else*. Thus, there is a hierarchical partitioning resulting in levels of granularity, and per level of granularity, there is 'horizontal' partitioning resulting in a grid or list; this is visualised in *Figure 1*. It is beyond the scope of the article to provide a full discussion on formal representations of granularity. We will limit ourselves to the definitions as provided by Kumar *et al.* [19, 20], where the annotation is summarised here. Let **GR** be the ordered set of levels of granularity applicable to a domain and dimension and **U** denotes the set of biological universals, then **gran** is the function of **U** onto **GR** (1). Taking anatomical entities as example, an instance x (and by implication its class that is in U) belongs to a level of granularity (2), illustrated for neurons (3) at the cell-level of granularity (4).

$\textbf{gran}:u \rightarrow \textbf{gran}(u)$, for $u \in U$ and $\textbf{gran}(u) \in GR$ (1)

$x \rightarrow \textbf{gran}(x)$ (2)

$$neuron \rightarrow \mathbf{gran}(neuron) \ (3)$$
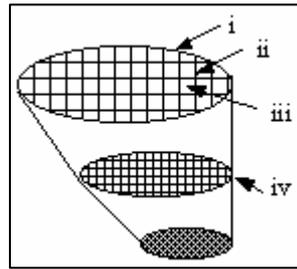$$\mathbf{gran}(neuron) = cell \ (4)$$

Figure 1. Granularity: top ellipse (i) is a coarse-grained granular level that is partitioned (ii) with less detail in larger cells (iii) than at the finer-grained level (iv).

3.2 Relations

Revisiting *Table 1*, there are two basic relations: *partOf* and *isA*, where the former, when used for substances, implies relations such as *locatedIn* or *contains*. Obviously, the phylogenetic tree is characterised by having the *isA* relation throughout the taxonomic tree, e.g. (5). However, structuring infectious agents according to their common names brings afore its inconsistencies. Hookworms are actually a subtype of roundworms – they are of the family *Ancylostomatidae*, which are of class *Secernentea* of phylum *Nematoda* (6a) – and threadworm and roundworm are synonyms for nematode (6b).

$$Metazoa \ \mathbf{isA} \ Eukaryote \ (5)$$
$$Ancylostomatidae \ \mathbf{isA} \ \dots \ \mathbf{isA} \ Secernentea \ \mathbf{isA} \ Nematoda \ (6a)$$
$$Hookworm \ \mathbf{isA} \ (Nematode = Roundworm = Threadworm) \ (6b)$$

More serious is the inconsistency of listing Gram-negative and Gram-positive bacteria: although it is a widely-used subdivision, it is based on the difference in murein content of the cell wall of bacteria and many more kinds of groups of bacteria exist, which are categorised by varying criteria [21] that are inconsistent ontologically. Therefore, we opt for the phylogenetic tree and a separate taxonomic (*isA* relation) categorisation for viruses. The nature of the *partOf* and its implied location relation concerning human anatomy, of relevance for formalising the site of entrance and effect of the infectious agent, has been discussed elsewhere [20, 22, 23]. Concerning formalising anatomical aspects of infectious diseases, we adhere to the nomenclature of the Foundational Model of Anatomy [22] throughout. From the relation concerning body parts (7-8), we can derive parthood relations for diseases, with corresponding taxonomic classification (9-10):

$$Upper \ lobe \ of \ right \ lung \ \mathbf{partOf} \ Whole \ right \ lung \ (7)$$
$$Right \ upper \ lobe \ pneumonia \ \mathbf{partOf} \ Whole \ right \ lung \ pneumonia \ (8)$$
$$Right \ upper \ lobe \ pneumonia \ \mathbf{isA} \ Pneumonia \ \ (9)$$
$$Right \ lung \ pneumonia \ \mathbf{isA} \ Pneumonia \ (10)$$

The pathological process and mode of action are closely related, but the former contains partitions of processes and sub-processes that occur in the human body, whereas the latter uses a functional categorisation from the perspective of the infective agent. Thus, the mode of action of *Vibrio cholerae* is of the level *ToxinProducer* (11-13) with one level of granularity lower the *CholeraToxin* with the function of inhibitor (14-15) and so forth (16-17). Note that *ToxinProducer* does not have always as part *Inhibitor*. In addition, (1-2) provides a generic mechanism for levels in granularity, but without further specification of the **gran** function, at the implementation level one has to distinguish between different granular perspectives to prevent incorrectly mixing different hierarchies. Thus, while e.g. **gran**(*CholeraToxin*) is logically correct, the computational implementation will refer to this as **gran-moa**(*CholeraToxin*) to distinguish it from **gran-anat**(*CholeraToxin*) for anatomi-

cal granularity, where the toxin is at the level of *Molecule*.

> *Vibrio cholerae* → **gran**(*Vibrio cholerae)* (11)
> **gran-moa**(*Vibrio cholerae*) = *ToxinProducer* (12)
> *Inhibitor* **partOf** *ToxinProducer* (13a)
> *Stimulator* **partOf** *ToxinProducer* (13b)
> *CholeraToxin* → **gran-moa**(*CholeraToxin)* (14)
> **gran-moa**(*CholeraToxin*) = *Inhibitor* (15)
> *Covalent binding of the small subunit of the cholera toxin to the G-protein of the Second Messenger System* **partOf** *Blocking G-protein to return to its inactive state* (16)
> *Blocking G-protein to return to its inactive state* **partOf** *Inhibitor* (17)

The *partOf* relation relates the pathological processes with their sub-processes. For example, in case of lobar pneumococcal pneumonia, one can think of (18). However, a medical condition involving *Congestive process* does not imply it is part of an inflammatory process, but it is the first stage of inflammation for pneumococcal pneumonia that does involve (a subtype of) congestion [24], therefore (19) represents (18) biologically more accurately (refer to [19] for details on the **involvedIn** relation).

> *Congestive process* **partOf** *Inflammatory process* (18)
> *Congestive process during pneumonia* **involvedIn** *Inflammatory process* (19)

Regarding the mode of transmission as process, the levels are more loosely related than *partOf*, therefore we can use *involvedIn* to relate processes with their sub-processes within the perspective of transmission (20a) or remodel it using *isA* in case of *Direct contact*-related levels (20b). It follows that if a human has pneumococcal pneumonia spread to the blood (bacteraemia), the patient can transmit the infection through blood to another person. While this is unlikely, it is theoretically not impossible, and common for several other infectious diseases. Because *Blood* is also defined in anatomy (part of the *Hemolymphoid system*) and at the finer-grained level involved in transmission of infectious agents through *Direct contact*, *Blood* is positioned on the intersection of these two distinct axes of granular perspectives, therefore one may derive (22) by traversing each axis upwards. In contrast, traversing "down" from *Hemolymphoid system* through another path in the partonomy (23), one cannot conclude that *Skin-associated lymphoid tissue* is related to *Direct contact* transmission, but does pose hypotheses on some connection with infections. Actually, *Skin-associated lymphoid tissue* (SALT)'s primary function is to *prevent* infectious agents to enter the vascular system, thereby preventing infectious agents to enter the blood.

> *Blood* **involvedIn** *Person-to-person* **involvedIn** *Direct contact* (20a)
> *Transmission via Blood* **isA** *Person-to-person transmission* **isA** *Transmission with direct contact* (20b)
> *Blood* **partOf** *Hematopoietic system* **partOf** *Hemolymphoid system* (21)
> *Hemolymphoid system* **involvedIn** *Direct contact* (22)
> *Hemolymphoid system* **hasPart** … **hasPart** *Skin-associated lymphoid tissue* (23)

The predisposing factors are formalised analogous to mode of transmission, where e.g. the higher-level partition *Living habit* of *Smoking* impairs the *Respiratory system*, which increases the likelihood of contracting pneumonia. Disease classification follows the standard taxonomic classification. Partitioning pathological structures are a topic of further investigation because the common categorisations are either canonical [14, 22] or based on an intuitive mixture of taxonomy and partonomy, hence requires considerable further ontological analysis before it will be included in the implementation.

4. Conclusions

Starting from an informal categorisation of infectious diseases, we identified and formalised the relations of its main partitions, illustrated with several prevalent infectious dis-

eases. The advantages of this approach will be better data integration in life-sciences, better inferencing, and easier correction and maintenance of ontologies, which will outweigh initial additional work and training of domain experts and increased complexity of the represented semantics. We will integrate this case study of granularity for infectious diseases computationally with existing generalised granular perspectives, such as human anatomy and function, to create a comprehensive system for automatic reasoning with biomedical data.

Acknowledgements

References

[1] He Y, Vines RE, Wattam AR, Abramochkin GV, Dickerman AW, Eckart JD, Sobral BWS. PIML: the Pathogen Information Markup Language. *Bioinformatics* 2005: 21(1): 116-121.
[2] National Center for Infectious Disease. http://www.cdc.gov/ncidod/.
[3] Melhorn H (ed.). *Encyclopedic reference of parasitology*. 2nd ed. Springer-Verlag Heidelberg, 2004.
[4] Rodal AA, Sokolova O, Robins DB, Daugherty KM, Hippenmeyer S, Riezman H, Grigorieff N, Goode BL. Conformational changes in the Arp2/3 complex leading to actin nucleation. *Nat Struc & Mol Bio* 2005: 12: 26-31 (Published online: 12-12-2004).
[5] Editorial. Lessons from Listeria. *Nat Struct & Mol Bio,* 2005: 12: 1.
[6] Snomed CT: http://www.snomed.org/snomedct/.
[7] International Classification of Diseases, ICD-10 (2003): http://www.who.int/classifications/icd/en/.
[8] Schlegel HG. *General Microbiology*. 7th ed. Cambridge: Cambridge University Press, 1995.
[9] Stryer L. *Biochemistry*. 3rd ed. New York: WH Freeman and Co, 1988.
[10] Pathologie Online: http://www.pathologie-online.de/.
[11] Maddison DR, Schulz K-S (ed.). The Tree of Life Web Project. 2004: http://tolweb.org.
[12] Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucl Ac Res* 2004: **32**: D258-D261.
[13] Tange HJ, Schouten HC, Kester ADM, Hasman, A. The Granularity of Medical Narratives and Its Effect on the Speed and Completeness of Information Retrieval. *J Am Med Inf As* 1998: 5(6): 571-582.
[14] Schultz S, Hahn U. Ontological foundations of biological continuants. In: Varzi AC Vieu L eds. *Proceedings of Formal Ontology in Information Systems 2004*. Amsterdam: IOS press, 2004: pp319-330.
[15] Johnson DM. Can Abstractions be Causes?. *Biology and Philosophy* 1990: 5: 63-77.
[16] Hobbs JR. Granularity. *Intl Joint Conference on Artificial Intelligence 1985*. 1985: 432-435.
[17] Mani I. A theory of granularity and its application to problems of polysemy and underspecification of meaning. In: Cohn AG Schubert LK Shapiro SC, eds. *Proceedings of the 6th International Conference on Principles of Knowledge Representation and Reasoning*. San Mateo: Morgan Kaufmann, 1998: pp245-255.
[18] Bittner T, Smith B. A Theory of Granular Partitions. In: Duckham M Goodchild MF Worboys MF, eds. *Foundations of Geographic Information Science*. London: Taylor & Francis Books, 2003; pp117-151.
[19] Kumar A, Yip L, Smith B, Grenon P. Bridging the Gap between Medical and Bioinformatics Using Formal Ontological Principles. Computers in Biology and Medicine. (In press).
[20] Kumar A, Smith B, Novotny DD. Biomedical Informatics and Granularity. *Comp & Func Gen* (In Press).
[21] Keet CM. *Conceptual Modelling for Applied Bioscience: the Bacteriocin Database*. CSPS/computational intelligence/0310001. 2003.
[22] Rosse C, Mejino JLV. A reference ontology for biomedical informatics: the foundational model of anatomy. *J Biomed Inf* 2003: 36: 478-500.
[23] Smith B, Rosse C. The Role of Foundational Relations in the Alignment of Biomedical Ontologies. *Proceedings of MedInfo 2004*, San Francisco.
[24] Merck. *Pneumonia*. http://www.merck.com/mrkshared/mmanual/section6/chapter73/73a.jsp

Address for Correspondence
C. Maria Keet, KRDB Research Centre, Department of Computer Science, Free University of Bozen-Bolzano, Piazza Domenicani 3, 39100 Bozen-Bolzano, Italy. Phone: +39 04710 16128. Email: keet@inf.unibz.it. URL: http://www.inf.unibz.it/krdb.