Constructing a Lattice of Infectious Disease Ontologies from a *Staphylococcus aureus* Isolate Repository

Albert Goldfain[1,*] Barry Smith[2] and Lindsay G. Cowell[3]
[1] Blue Highway Inc., Syracuse, NY, USA
2 University at Buffalo, Buffalo, NY, USA
3 University of Texas Southwestern Medical Center, Dallas, TX, USA

ABSTRACT

A repository of clinically associated *Staphylococcus aureus* (Sa) isolates is used to semi-automatically generate a set of application ontologies for specific subfamilies of Sa-related disease. Each such application ontology is compatible with the Infectious Disease Ontology (IDO) and uses resources from the Open Biomedical Ontology (OBO) Foundry. The set of application ontologies forms a lattice structure beneath the IDO-Core and IDO-extension reference ontologies. We show how this lattice can be used to define a strategy for the construction of a new taxonomy of infectious disease incorporating genetic, molecular, and clinical data. We also outline how faceted browsing and query of annotated data is supported using a lattice application ontology.

1 INTRODUCTION

One of the more ambitious goals of current clinical and biomedical research is the personalization of medicine, in which treatments are selected on the basis of patient-specific as well as disease-specific information. Recent advances in high-throughput technologies have resulted in a push for the use of patient-specific information in care decisions, particularly genomic and functional genomic data, but also proteomic, metabolomic, and cytometry data. It is widely believed that the increased precision of personalized medicine will yield more effective treatments, with better outcomes and fewer adverse side effects.

Personalized medicine requires that genomic (and other) data be effectively classified and associated with known clinical phenotypes and disease types. Currently available taxonomies of disease do not support this, however, and are in general not well suited for integration and analysis of high-throughput molecular and cellular data with clinical data, such as the data found in electronic medical records. Current disease taxonomies were developed primarily to support diagnosis and reimbursement coding rather than as biological representations of disease. As a consequence, they are based on single, rigid hierarchies that do not reflect the complex interconnections between disease types; they lack links to molecular- and cellular-level data and information; and they lack the sort of formal structure that would support their use for the kinds of computational analyses applied in biological and clinical research. For example, the International Classification of Disease (ICD) version 9 includes catch-all codes such as "[041.19] Other Staphylococcus" and scattered exclusions such as "[041] Bacterial infection in conditions classified elsewhere and of unspecified site. Excludes: septicemia (038.0 – 038.9)".

The National Academies of Science have recently called for a new taxonomy of disease, along with informatics tools to support its construction (Committee on the Framework for Developing a New Taxonomy of Disease, 2011). In support of such a taxonomy, an information commons would be developed to store "bedside" clinical data collected during clinical encounters, effectively treating each patient as a participant in a clinical study, and integrate this information in a knowledge network that would formalize the relationships between different disease data sets. The long-term goal is to produce the new taxonomy of disease from a validated subset of the knowledge network.

We believe that biomedical ontologies will be essential to the construction of the envisioned taxonomy of disease, especially the ontologies in the Open Biomedical Ontology (OBO) Foundry (Smith *et al*., 2007). The OBO Foundry (OBOF) represents a coordinated effort to construct reference biomedical ontologies according to best practices and principles and to use these ontologies as the basis for OBOF-conformant application ontologies. The coordinated development of these ontologies and their use of a common formalism increases data interoperability and consistency for datasets annotated in their terms. The use of OBOF ontologies in construction of the new disease taxonomy can bring significant benefits. For example, the widespread use of OBOF ontologies for data annotation would link the disease taxonomy to many existing databases and information resources, and their underlying formalism allows the dynamic inference of different views and multiple interconnected hierarchies. In addition, many analysis algorithms for high-throughput data already utilize these ontologies.

The Infectious Disease Ontology (IDO) suite of ontologies is being developed within the OBO Foundry framework and includes a hub – the IDO-Core – consisting of terms and relations relevant to infectious diseases generally, together with a set of disease-specific extensions derived therefrom. The IDO ontologies are interoperable and jointly cover the infectious disease domain. Here we illustrate how the IDO ontologies can be used in the construction of a part of the new taxonomy of disease and to integrate clinically relevant phenotypic and genotypic data.

* To whom correspondence should be addressed: agoldfain@blue-highway.com

We take as our case study infectious diseases caused by *Staphylococcus aureus* (Sa) infection. We show how isolate data from the Network on Antimicrobial Resistance in *Staphylococcus aureus* (NARSA) can be annotated using IDO and its extensions. We then demonstrate a faceted browser in which both phenotypic and genotypic aspects of the IDO-annotated isolate data can be exposed and queried. Our goal is to provide a resource from which an IDO-conformant application ontology can be derived for a specific Sa infectious disease type. Such application ontologies can be generated in a semi-automated way and collectively form a lattice structure beneath IDO-Core (described below). While our example narrowly focuses on properties of infectious agents, this effort is part of a larger effort to create an ontological representation of Sa diseases, and we believe the same approach can be applied to host data and to the integration of host and pathogen data.

2 INFECTIOUS DISEASE ONTOLOGY

IDO-Core includes terms relevant for infectious diseases generally, terms such as 'host', 'infectious agent', 'fomite', and 'virulence factor', and the relations between the corresponding types. Disease- and pathogen-specific extensions are developed by extending the core to include terms and relations relevant to the corresponding infectious disease(s). For example, the IDO extension for Sa (IDO-Sa) includes terms such as 'Staphylococcus aureus bacteremia' and 'Staphylococcal cassette chromosome *mec*'.

IDO extensions are currently being developed for influenza, malaria, brucellosis, HIV, and Sa. Further extensions will involve the creation of specific application ontologies by IDO user groups. It will be necessary for these ontologies to import terms from several OBO Foundry ontologies, as well as from existing IDO extension ontologies. This will give rise to a lattice structure beneath IDO core and its extensions, as illustrated in Figure 1. At the bottom of the lattice is IDO-ALL, the (pre-inference) closure of possible the IDO ontologies.

When a new application ontology is needed, its position in the lattice will be determined by the terms it needs to import. IDO Core is agnostic to biological scale, host organism, and disciplinary perspective, but it will be desirable for some of the application ontologies in the lattice to hold some of these fixed (e.g., genetic aspects of influenza in birds), thus serving as granular partitions of the domain ontology they are extending. The lattice serves as a representation of some of the interdependencies in the existing IDO set of ontologies and the intended overall domain coverage.

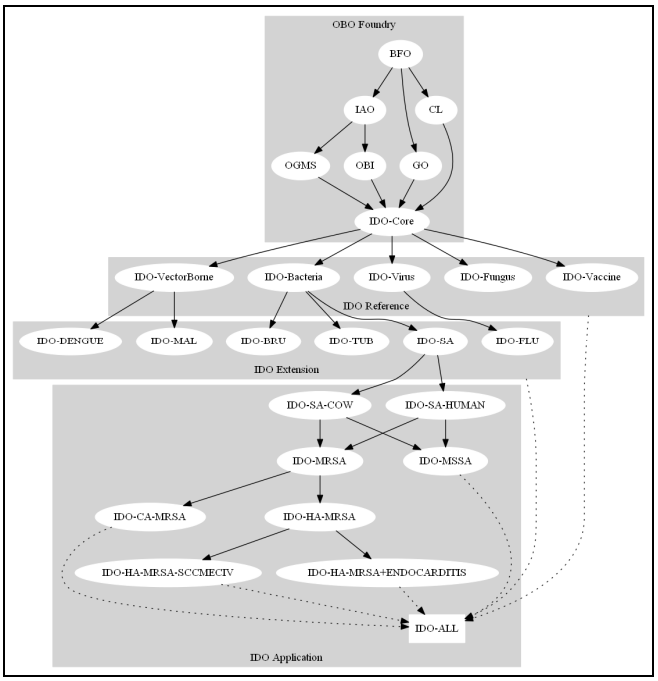

Fig 1. A possible lattice expansion of IDO

2.1 OGMS/IDO Disease Model

The IDO ontologies represent disease according to the *disorder – disease – disease course* framework provided by the Ontology for General Medical Science (OGMS), in which a disorder is the physical basis of a disease, which is itself a disposition to pathological processes realized in a disease course. For example, in IDO-Sa we assert the following in OWL-DL:

- Sa **subClassOf** obi:organism AND
  ido:'infectious agent'
- SaI =*def* ido:'infectious disorder' AND
  has_part SOME Sa
- SaID =*def* ido:'infectious disease' AND
  has_material_basis_in SOME SaI.
- SaID **realized_by** ONLY SaIDC

where, 'Staphylococcus aureus' = Sa, 'Sa Infectious Disorder' = SaI,'Sa Infectious Disease' = SaID, and 'Sa Infectious Disease Course'=SaIDC.

The primary classification of Sa is as an organism, but Sa bacteria are also infectious agents because they have a disposition to cause infectious disease in some hosts. Note we define Sa infectious disorder as an infectious disorder that has Sa as part, but we do not assert "Sa **part_of** SOME SaI" because Sa can be among a host's normal flora, for example on the skin or nasal mucosa.

We use the shortcut relation **has_material_basis** here to establish a link between the disease (disposition) and the disorder (material entity) (Goldfain, Smith and Cowell, *under review*). An infectious disorder is both an infection (a material entity composed of infectious agents) and a disorder (has reached the threshold of clinical significance to dispose a host to infectious disease).

2.2 Classifying *Staphylococcus aureus* diseases

Infectious diseases can usefully be classified in terms of a number of differentia, including: host type, (sub-)species of infectious agent, route of transmission, antibiotic resistance, and anatomical site of infection.

For many species of infectious agent, including Sa, a further classification into strain categories is useful. Many different typing systems are used, including: Pulse Field Gel Electrophoresis (into strains), Multi-Locus Sequence Typing (into sequence types), BURST Clustering (into clonal complexes), and gram staining (into gram positive and gram negative classes). Each of these typing systems is tied to a particular type of assay that can be described using the Ontology for Biomedical Investigations (OBI).

For our present purpose, we are interested in a typing system specifically created to differentiate Sa isolates, the Staphylococcal cassette chromosome *mec* (SCC*mec*) typing system. SCC*mec* is further differentiated by its subparts: (a) Cassette chromosome recombinases (*ccr*) and (b) *mec* gene complex (mec). The SCC*mec* is a mobile genetic element that carries the central determinant for broad-spectrum beta-lactam antibiotic resistance encoded by the *mecA* gene (Katayama, Ito and Hiramatsu, 2000). The genetic characteristics of SCCMec are of critical importance to the type of treatment and *Sa* disease course an infected host may undergo. The International Working Group on the Staphylococcal Cassette Chromosome elements[1] maintains a list with definitions of the latest known SCC*mec* types. At the time of this writing, there are 11 known SCC*mec* types. We include this information in IDO-Sa by leveraging the Sequence Ontology (SO) to assert the following:

- SCCMec **subClassOf** so:gene_cassette
- SCCMec **subClassOf** so:mobile_genetic_element
- 'mec gene complex' **subClassOf** so:gene_cassette_member
- 'ccr gene complex' **subClassOf** so:gene_cassette_member
- SCCMec **has_part** SOME 'mec gene complex'
- SCCMec **has_part** SOME 'ccr gene complex'

The classification of SCC*mec* as a gene cassette is to be preferred over its classification as a mobile genetic element because the former tells us what SCC*mec is*, while the latter tells us what SCC*mec* can *do*. However, we include both here, because most descriptions of SCCmec highlight its mobility. Description of a SCCMec subtype then proceeds as follows:

- SCCMecIV **subClassOf** SCCMec
- 'mec Class B' **subClassOf** 'mec gene complex'
- 'ccr Type 2' **subClassOf** 'ccr gene complex'
- SCCMecIV **has_part** SOME 'mec Class B'
- SCCMecIV **has_part** SOME 'ccr Type 2'

More fine grained sequence information about the *ccr* and *mec* complexes can be captured using SO terms and relations.

3 CASE STUDY

We will now show how a lattice of Sa isolates can be constructed using IDO-Sa and isolate metadata indicating properties such as the *mec* and *ccr* gene complex types. The isolate lattice is then used as the basis for our desired lattice of infectious disease application ontologies. Ontologically speaking, isolates are particulars that instantiate the organism type Sa and have been extracted from a host organism. Here we do not represented the distinctions between Sa as an 'isolate' or as part of a 'cell culture', however we believe these terms are general enough to infectious disease research to warrant inclusion in IDO-Core.

The ontology generated for this case study is stored across several OWL files. The full ontology, including external imports and automatically generated isolate information is currently available in OWL-DL format at http://www.awqbi.com/LATTICE/narsa-complete.owl. The ontology was developed using Protege 4.1 and was checked for inconsistency using the Hermit 1.3.5 and Fact++ reasoners.

3.1 Resources

Wherever possible, we import and reuse terms (and URIs) from OBO Foundry ontologies via the MIREOT technique (Courtot *et al.*, 2011) and use relations from the OBO relation ontology (RO) or proposed extensions thereto. The OBO Foundry ontologies we require for our case study are: Ontology for General Medical Science (OGMS[2]), Ontology for Biomedical Investigations (OBI[3]), Sequence Ontology (SO), Infectious Disease Ontology (IDO[4]), Information Artifact Ontology (IAO[5]), NCBI Taxonomy (NCBITaxon[6]), and Foundational Model of Anatomy (FMA[7]).

We also import drug file names from the National Drug File Reference Terminology (NDF-RT) to represent antibiotic resistance, and create links to two other resources: (1) Antibiotic Resistance Ontology[8] and Antibiotic Resistance Database Ontology[9]. Various other stakeholders (such as the DebugIT European Union initiative) have ontologies and databases of antimicrobial resistance, but we only to link to open resources for our case study.

[1] http://www.sccmec.org/Pages/SCC_ClassificationEN.html

[2] http://code.google.com/p/ogms/

[3] http://obi-ontology.org/page/Main_Page

[4] http://infectiousdiseaseontology.org/page/Main_Page

[5] http://code.google.com/p/information-artifact-ontology/

[6] http://www.ncbi.nlm.nih.gov/Taxonomy/

[7] http://sig.biostr.washington.edu/projects/fm/

[8] http://arpcard.mcmaster.ca

[9] http://ardb.cbcb.umd.edu/antibio_resis.obo

3.2 NARSA Isolate Repository

The Network on Antimicrobial Resistance in *Staphylococcus aureus*[10] maintains a repository of Sa isolates for clinical research which includes genetic, phenotypic, and demographic information on each isolate. For this example, we use a subset of 101 NARSA isolates, those listed in the "Known Clinically Associated Strains – ABCs Collection from CDC" repository. All of the isolates in this subset have an SCCMec type annotation in the NARSA repository and have diverse geographic origin in the United States.[11]

The NARSA subset was selected to demonstrate how a disease lattice could be constructed starting from only structured HTML content about isolates. NARSA maintains a database of extended information about such isolates; however we only used the information publicly available on the web.

A script was created to extract each isolate's NARSA id (NRS*nnn*), culture source, toxin profile, and antimicrobial profile. The script was implemented in Ruby and utilized the Hpricot HTML library and regular expressions to extract information. First, the NARSA id was used to assert the existence of a Sa instance type. Then the culture source data was extracted. The culture source was sometimes unspecified ('other') or underspecified ('blood' vs 'wound'). Only culture sources for which FMA types existed were asserted to exist as such, but IDO allows for an even more complete representation of host anatomical entities if such information is known. For example, the anatomical location from which the infectious organism is isolated may also be a portal of entry.

The toxin profile for NARSA subset isolates included the presence or absence of the Panton Valentine Leukocidin (PVL) and Toxic Shock Syndrome Toxin (TSST). These toxins are strong determinants of the virulence and clinical manifestation of Sa disease. We classify PVL and TSST as ido:exotoxin. The presence or absence of a toxin is not usually associated with drug resistance, but by representing both pieces of information we are able to query the application ontology for correlations between the presence of toxins and resistance to certain drug types.

The antimicrobial profile for the NARSA subset includes 15 drugs (see Figure 2 for a subset of these). For each drug, NARSA reports a minimum inhibitory concentration – a range or exact value – along with an interpretation of the antibiotic resistance indicated by this value following the Clinical and Laboratory Standards Institute guidelines.

S = Susceptible; **R** = Resistant; **NS** = Not Susceptible; **I** =Intermediate; **N/A** = Not Available

NARSA Antimicrobial Profile for Other Antimicrobial Agents

Drug	MIC (mg/ml)	CLSI Interptn
Chloramphenicol	= 8	S
Clindamycin	<= 0.25	S
Daptomycin	<= 0.5	S
Doxycycline	= 2	S
Erythromycin	= 0.5	S

Fig 2. Antimicrobial profile for an isolate in the NARSA subset

The NDF-RT was used to validate this profile by making sure that the set of drugs in the profile is a subset of:

{*d* | ndf-rt:'Staph Infection' **ndf-rt:may_be_treated_by** *d*}

For NARSA, or any other resource on antimicrobial resistance, there may be a good reason to restrict attention to a subset of antimicrobials. However, since new resistance evolves rapidly, a resource such as NDF-RT can be used to synchronize the latest antibiotics permissible in such a profile.

Minimum inhibitory concentration data (MIC) are represented using IAO and OBI as follows:

- 'MIC assay' **subclassOf** iao:assay
- 'MIC assay' **has_specified_output** SOME 'MIC data item'
- 'MIC scalar measurement datum' **is_about** SOME 'drug susceptibility of infectious agent'

Resistance is a disposition that an infectious agent bears towards some drugs and is realized in their presence. We have elsewhere modeled resistance in terms of pairwise complementary dispositions on the part of both the infectious agent and the drug (Goldfain, Smith & Cowell, 2011). Here we link resistance to MIC measurement data using the shortcut relation **has_qualitative_basis** as follows:

- ido:'resistance to drug' **has_qualitative_basis** SOME (**is_quality_measured_as** SOME 'MIC measurement datum')

Finally, for each drug *D* towards which the isolate Sa has a drug resistance we assert:

- 'resistance to *D*' **subclassOf** ido:'resistance to drug'
- Sa **has_disposition** SOME 'resistance to *D*'

3.3 From an Isolate Lattice to a Disease Lattice

The lattice of infectious diseases mirrors the isolate lattice by representing the types of infectious disease different isolates can give rise to. Infectious agents are parts of those infectious disorders which are the material basis for infec-

[10] See http://www.narsa.net/

[11] See http://www.cdc.gov/abcs/reports-findings/surv-reports.html

tious disease. Using the representation developed above, we can begin to make assertions about the specific types of disease the isolates give rise to and the profiles of the disease courses which realize these diseases. For example, the presence of the PVL toxin in Sa can lead to necrotic lesions (ogms:disorder) and necrotizing pneumonia (ogms:disease).

4 FACETED BROWSING OF THE LATTICE

A faceted browser of the ontologically annotated NARSA isolates was constructed using the MIT Exhibit 2.0 library (http://www.awqbi.com/LATTICE/narsa-complete.html). This tool allows the user to visualize and correlate isolate information across different dimensions (see Figure 3).

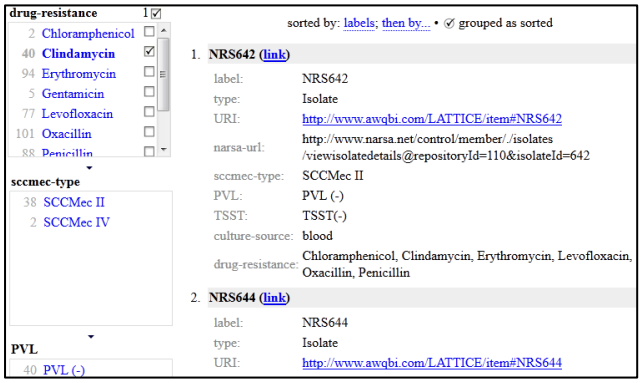

Fig 3. Faceted browsing illustrates that most isolates with a resistance to Clindamycin are of SCC*mec* type II and lack PVL

Linking to external resources is facilitated by the fact that such facets are assigned ontology types from the IDO lattice. These are exactly the kinds of links that will be needed for the knowledge network supporting a new taxonomy of disease.

5 CONCLUSION

A lattice of infectious disease ontologies can serve as a mechanism to integrate pathogen-specific typing systems such as SCCMec with phenotypic data such as drug resistance. Such genotype-phenotype relations will be the key to a more effective taxonomy of disease that enables truly personalized medicine. The lattice of infectious diseases is expected to grow along predictable dimensions (host organism, infectious agent organism, drug resistance), but can accommodate lightweight application ontologies that are created for very specific purposes. Each such application ontology will have a place in the lattice on the basis of what IDO terms it imports.

We have shown that IDO-conformant annotation of isolate data (such as that in the NARSA repository) is possible without the need to reassemble OBO Foundry resources for new applications. Other benefits of our approach include: exposing currently accepted SCC*mec* types in a computable format via an ontology and validating the NARSA antimicrobial profile using the NDF-RT.

We hope to reuse a similar technique to that outlined in this paper for isolate repositories across the infectious disease domain. In so doing, we hope to broaden the lattice and integrating organism specific typing systems with the IDO suite of ontologies. We believe that such an effort can be a powerful enabler for a new taxonomy of infectious disease and its supporting knowledge network.

ACKNOWLEDGEMENTS

This work was funded by the National Institutes of Health through Grant R01 AI 77706-01. Smith's contributions were funded through the NIH Roadmap for Medical Research, Grant U54 HG004028 (National Center for Biomedical Ontology).

REFERENCES

Committee on the Framework for Developing a New Taxonomy of Disease (2011). *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. The National Academies' Findings Report.

Courtot, M., Gibson F., Lister, A. L., Malone, J., Schober, D., Brinkman, R. R., and Ruttenberg, A. (2011). MIREOT: The minimum information to reference an external ontology term. *Applied Ontology*, 6(1), 23-33.

Goldfain, A., Smith, B., and Cowell, L. G. (under review). BFO Dispositions and their Bases: Two Shortcut Relations.

Goldfain, A., Smith, B., and Cowell, L. G. (2011). Towards an Ontological Representation of Resistance: The Case of MRSA. *Journal of Biomedical Informatics*, 44(1), 35-41.

Katayama, Y., Ito, T., and Hiramatsu, K. (2000). A New Class of Genetic Element, Staphylococcus Cassette Chromosome *mec*, Encodes Methicillin Resistance in *Staphylococcus aureus*. *Antimicrobial Agents and Chemotherapy*, 44(6), 1549-1555.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., The OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol, 25(11), 1251–1255.