Similarity between semantic description sets: addressing needs beyond data integration

Todd Vision^{1,3}, Judith Blake², Hilmar Lapp³, Paula Mabee⁴, and Monte Westerfield⁵

¹ University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
² Jackson Laboratory, Bar Harbor, ME, USA
³ National Evolutionary Synthesis Center, Durham, NC, USA
⁴ University of South Dakota, Vermillion, SD, USA

⁵ University of Oregon, Eugene, OR, USA

Abstract. Descriptive information is easy to understand and communicate in natural language. Examples in the biological realm include the cellular functions of proteins and the phenotypes exhibited by organisms. Large latent stores of such descriptive data are stored in databases that can be mined, but even more still reside only in the scientific literature. Although such information has traditionally been opaque to computers, in recent years significant efforts have gone into exposing descriptive information to computation through the development of ontologies and associated tools. A host of software applications now employ simple reasoning over Gene Ontology annotated data to help interpret experimental findings in genomics in terms of protein function. In the domain of biological phenotypes, the combination of entity terms from taxonspecific anatomy ontologies with quality terms from generic ontologies such as PATO have been used to construct semantically precise and contextualized descriptions. It is natural for multiple semantic descriptions to pertain to single instances in the real world, as in the case of both protein functions and organismal phenotypes. However, applications for ontology-based annotations that go beyond simple knowledge organization, and that exploit *sets* of semantic descriptions, are puzzlingly rare. In particular, we argue that there is wide applicability, and a sore need, for tools that can satisfy the simple, common use case of identifying statistically improbable similarity between sets of semantic descriptions. Several metrics have been proposed for this task in the literature, but not vet fully evaluated, explored, and adopted. The requirements for semantic similarity tools tailored to sets of semantic descriptions would include speed, scalability to large numbers of sets, demonstrated statistical and biological validity, and ease of use.

Ontologies are a foundational technology for a semantic web of linked data. As a key element for data discovery, reuse, and integration, they allow the standardization and relation of concepts across documents, databases, and communities of practice. Ontologies also allow the semantics of concepts to be exposed to formal machine reasoning, and thus provide the opportunity for the linked data web to be used for more sophisticated logical operations than are currently possible.

Here, we define *descriptive data* broadly as information about the qualitative features of objects in the world, *e.g.* "albatrosses have long wing". A large and diverse universe of descriptive data is known to science, but such statements are typically expressed in natural language, as above. These are then typically transformed to quantitative form (*e.g.* word frequency) for the purposes of computation, and loss of meaning accompanies this transformation. What if, instead, we had tools to exploit the semantic content of diverse collections of descriptive data directly [6]?

Biological discourse is particularly rich in descriptive data, though it is often expressed within text and not managed within data collections. The Gene Ontology (GO) [1] first introduced the biological community to the use of ontologies for standardization of descriptive data, in this case the function and location of gene products, across a broad community of practice. The years since have witnessed steady growth in the diversity of specialized knowledge domains within biology, particularly biomedicine, for which ontologies have been developed, as illustrated by the current breadth of the NCBO BioPortal [11] and the Open Biological and Biomedical Ontologies [15]. The popularity of ontologies among biologists is in large part due to their suitability as controlled vocabularies that aid in the harmonization and integration of terminology among different communities of practice [13]. Secondarily, they are increasingly used as a classification aid that enables richer navigation of information resources [14, 7], and to identify those concepts that are more frequently associated with documents in a corpus, either directly or by inference, than would be expected by chance alone [5].

Although tremendously useful, these relatively straightforward knowledge organization tasks fail to exploit the potential of ontologies as tools for scientific knowledge discovery. The effort required to produce an ontology for a knowledge domain can be significant, and engaging the community of domain experts to ensure its fitness for wide adoption is challenging. Knowledge discovery applications offer an additional route by which ontologies can deliver powerful scientific returns to bench biologists, and in so doing incentivize them to contribute to the building of more comprehensive and useful community ontologies.

One such application that has recently been shown to harbor great potential, particularly in biology [16] and drug discovery [2, 4], is semantic similarity search. Briefly, a semantic similarity search application takes as input a set of one or more ontological statements of descriptive data, compares the set against a database of such sets, and returns those sets from the database that have greater semantic similarity to the query set than would be expected by chance. Each set corresponds to a number of descriptive data statements, in the form of ontological classes, made about a common object. The semantic similarity between two classes may reflect the ontology graph distances between concepts and/or the information content of common subsuming ontology concepts (see [12] for an overview of metrics used in bioinformatics applications). A common application of semantic similarity search is to identify objects in a database that are semantically similar to a query object. For example, Washington *et al.* [16] used classes expressing the semantics of heritable human disease phenotypes to query a database of mutant phenotypes from genetically well-characterized model organisms. The query returned model organism genes that have mutant phenotypes with semantics similar to the phenotypes of human heritable diseases. The genes obtained in this way suggested testable hypotheses for the genetic causes of those diseases, which were previously unknown.

The semantic similarity metrics employed so far are relatively straightforward to calculate if they are applied to ontology classes that, aside from being placed in a subsumption hierarchy, are not axiomatically defined, for example when assessing the semantic similarity between genes based on the GO terms associated with each. However, the inferences possible from such simple hierarchies are limited. Conversely, axiomatically defining the classes by combining several orthogonal, modular domain ontologies as class expressions, such as using intersections of property restrictions in OWL, greatly increases expressivity of the ontological expressions, as well as the inferences a reasoner can make from them [9]. This in turn can increase sensitivity as well as specificity of finding semantically similar matches. However, as a consequence of the increased expressivity, enumerating all subsumers of a complex, possibly nested, class expression, which the currently best-performing metrics require, can quickly become time and memory consuming with large ontologies and large databases of class expressions. As an example, calculating the similarity statistics for the Washington et al. study took several days on a relatively small database with only several thousand sets of class expressions.

The use of ontologies to reason about qualitative phenotypes is now being explored by a number of different groups. Among these efforts is Phenoscape, a project with which we are all involved and which aims to enable computation across phenotypic information from different biological disciplines (e.g. genetics and biodiversity) [3]. Our application of semantic similarity is over sets of phenotypic data, where a *phenotype* is defined as the set of observable traits present in an individual organism as a result of the interaction of heredity, environmental influences, and the developmental process, *e.g.* the elongated wings of albatrosses. The qualitative phenotype descriptions are central to and investigated in meticulous detail in many different areas of biology, and such phenotypes have traditionally been reported and communicated in expressive, but highly discipline-specific natural language.

To express qualitative phenotypes with computable semantics, we use the emerging standard of an Entity-Quality (EQ) formalism [10], which decomposes phenotypes into three main components: a quality (e.g., "elongate" shape), the entity bearing that quality (e.g., a "wing"); and the class of organism expressing the phenotype (e.g., a taxon or a genotype). Each of the components of an EQ expression, and the relations between them, are expressed using terms and properties from appropriate ontologies [8]. When represented in OWL, EQ expressions are conjunctive class expressions, and thus axiomatically defined classes. Phenotypes may be somewhat complex, such as qualities borne only by

particular regions or parts of an anatomical entity, or described as spatial relationships between anatomical structures, and therefore such class expressions can consist of multiple and recursively nested property restrictions.

The Phenoscape Knowledgebase (http://kb.phenoscape.org) currently contains over half a million semantic phenotype descriptions in the form of EQ class expressions for more than 5,000 different biological taxa that are linked to more than 4,000 candidate genes through EQ phenotype descriptions from mutants of a single model organism. Thus, there are many sets of descriptions with a cardinality on the order of 100 descriptions. Although this is a large data store, it has been compiled from only a restricted branch of the tree of life (ostariophysan fishes), and thus represents only a small fraction of the amount of latent phenotype information for all organisms. Although restricted in taxonomic scope, the Phenoscape Knowledgebase already contains information that would be sufficient to generate thousands of hypotheses about the genetic basis of phenotypic transitions in evolution, were we in a position to compute the semantic similarity between the sets of semantic phenotype expressions associated with each of the thousands of taxa and genotypes!

Calculating the best performing of the currently available metrics for semantic similarity [12] requires the enumeration of all subsumers, or identifying the Least Common Subsumer, of the classes for which the similarity is being evaluated. Available DL reasoners can return only subsuming classes that are actually present in a knowledgebase. Although this is sufficient if the classes being evaluated are terms drawn from a subsumption hierarchy, in the case of class expressions there is typically a large number of possible combinations of property restrictions that subsume a given class expression, most of which will not be present in the knowledgebase. A custom-written algorithm could be used to enumerate all possible subsumers and add them to the knowledgebase so that a DL reasoner can subsequently return them, but the number of subsumers grows combinatorially with the number of property restrictions, and thus quickly becomes too large and too time-consuming to compute at query time. For example, for a conjunctive class expression with n property restrictions, where object c_i in restriction i has N_i asserted or inferred superclasses (in the knowledgebase), there are $(\prod N_i)/n!$ possible subsuming class expressions. Based on our results so far, semantic similarity search engines will require a speedup of two to three orders of magnitude to enable a user to launch multiple searches over even a few thousand sets of semantic descriptions (the current scale of Phenoscape), if the results are to be returned within a single user session. To be fit for large-scale reasoning over a web of linked data, algorithms will need to scale to substantially larger problem instances.

Additionally, although there have been some performance comparisons among metrics, and important biological demonstrations of the validity of some of the patterns found [12, 16], the available measures have not been statistically justified in terms of consistency and bias, nor benchmarked against error. It is likely that there is considerable room for improvement in the metrics that are available. Finally, to enable the broad use of semantic similarity search technology in bioinformatics and beyond, applications will need to be embedded within web applications that are accessible to the average information-seeking scientist. Complex dependencies on external software such as a database or reasoner, and the need to convert the format of source ontologies or data, will consign use of such tools to the realm of the specialist.

The paucity of tools available for computation over qualitative data in bioinformatics is a striking contrast to the vast array of tools that operate on other forms of non-numeric data, such as the character strings found in nucleotide and protein sequences. This is truly unfortunate given the importance and volume of descriptive data within biological discourse, and given the many applications that await to be developed for finding similar objects on the web of linked data.

References

- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25(1), 25–9 (May 2000)
- Campillos, M., Kuhn, M., Gavin, A.C., Jensen, L.J., Bork, P.: Drug target identification using side-effect similarity. Science (New York, N.Y.) 321(5886), 263–266 (Jul 2008)
- Dahdul, W.M., Balhoff, J.P., Engeman, J., Grande, T., Hilton, E.J., Kothari, C., Lapp, H., Lundberg, J.G., Midford, P.E., Vision, T.J., Westerfield, M., Mabee, P.M.: Evolutionary Characters, Phenotypes and Ontologies: Curating Data from the Systematic Biology Literature. PLoS ONE 5(5), e10708 (May 2010)
- Ferreira, J.a.D., Couto, F.M.: Semantic Similarity for Automatic Classification of Chemical Compounds. PLoS Computational Biology 6(9), e1000937 (Sep 2010)
- Huang, D.W., Sherman, B.T., Lempicki, R.A.: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research 37(1), 1–13 (Jan 2009)
- Jensen, L.J., Bork, P.: Ontologies in quantitative biology: a basis for comparison, integration, and discovery. PLoS Biology 8(5), e1000374 (May 2010)
- Kapushesky, M., Emam, I., Holloway, E., Kurnosov, P., Zorin, A., Malone, J., Rustici, G., Williams, E., Parkinson, H., Brazma, A.: Gene expression atlas at the European bioinformatics institute. Nucleic Acids Research 38(Database issue), D690–8 (Jan 2010)
- Mabee, P.M., Ashburner, M., Cronk, Q., Gkoutos, G.V., Haendel, M., Segerdell, E., Mungall, C., Westerfield, M.: Phenotype ontologies: the bridge between genomics and evolution. Trends in Ecology & Evolution 22(7), 345–50 (Jul 2007)
- Mungall, C.J., Bada, M., Berardini, T.Z., Deegan, J., Ireland, A., Harris, M.A., Hill, D.P., Lomax, J.: Cross-product extensions of the Gene Ontology. Journal of Biomedical Informatics (Feb 2010)
- Mungall, C.J., Gkoutos, G.V., Smith, C.L., Haendel, M.A., Lewis, S.E., Ashburner, M.: Integrating phenotype ontologies across multiple species. Genome Biology 11(1), R2 (Jan 2010)

- Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G., Musen, M.A.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research 37(Web Server issue), W170–3 (Jul 2009)
- Pesquita, C., Faria, D., Falcão, A.O., Lord, P., Couto, F.M.: Semantic similarity in biomedical ontologies. PLoS Computational Biology 5(7), e1000443 (Jul 2009)
- Schuurman, N., Leszczynski, A.: Ontologies for bioinformatics. Bioinformatics and Biology Insights 2, 187–200 (Jan 2008)
- Shah, N.H., Jonquet, C., Chiang, A.P., Butte, A.J., Chen, R., Musen, M.a.: Ontology-driven indexing of public datasets for translational bioinformatics. BMC Bioinformatics 10 Suppl 2, S1 (Jan 2009)
- Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L., Eilbeck, K., Ireland, A., Mungall, C., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R., Shah, N., Whetzel, P., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25(11), 1251–1255 (Nov 2007)
- Washington, N.L., Haendel, M.A., Mungall, C.J., Ashburner, M., Westerfield, M., Lewis, S.E.: Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biology 7(11), e1000247 (Nov 2009)