OMICS A Journal of Integrative Biology
Volume 7, Number 1, 2003
© Mary Ann Liebert, Inc.

Explaining Biology to Computers

MARK S. TUTTLE

THE KEY TO INTEROPERATION and therefore to computer-empowerment in biology will be a technology infrastructure that supports the development and use of a repertoire of pragmatically driven, longitudinally maintained, terminology models that "explain biology to computers." Called "Reference Terminology Models," these resources name and relate concepts such as sequences, structures, and functions, in a way that computers can exploit on behalf of biologists. Thus, a computer can use an appropriate reference terminology to help it determine if a given represented concept in one computer is equivalent to or related to a represented concept in another computer. The computer science challenge is the creation of an infrastructure that supports the productive creation, deployment, maintenance, evolution and use of Reference Terminology Models. In short, such models will be successful only if they are part of a supported, scalable, longitudinal process, and the process needs to be supported technologically. Such an infrastructure would create much of what would be required for a Semantic Web for biology.

EMPOWERING COMPUTERS

In a deep sense this meeting is about empowering computers to help biologists, and those served by biologists. Today, thanks to the World Wide Web, computers can help humans find information in other computers, because, for example, Google can see that information. But, as powerful as Google is, it still can't find things by MEANING, and it is generally of little if any help with biologic data. Put differently, Google is remarkably effective at finding Web information about instances, for example, a named individual "Mark S. Tuttle." It is less effective at finding Web information about categories or classes, for example "The class of all molecules known to result from the organized action of genes." The good news is that Web search is becoming a commodity; the bad news is that the techniques employed by today's Web search engines do not offer opportunities for incremental improvement. In a phrase, "they're stuck"—for the most part, and they will not provide a foundation for interoperation.

For the foreseeable future, the means of computer empowerment in biology will have little to do with Web search on the one hand, or artificial intelligence, agents, or mediation on the other. It will have everything to do with scalable semantic infrastructure—ways to help computers deal with meaning productively—the kind of infrastructure that will make it easier for databases of biological information to interoperate. It has to do with developing and deploying reusable semantics in selected parts of biology in ways already partly contemplated by the Semantic Web.

Since the Semantic Web will not be available in a predictable amount of time, biology, like biomedicine, must pursue interoperation by pursuing the "explanation" of a collection of semi-independent, tractable domains. In this context, EXPLANATION is the means by which one computer can tell if there is something usefully equivalent or usefully related in another computer. One result of such a capability is that one computer may be able to aggregate data found in two other computers. As we're discovering in biomedicine this is a tall order, but one against which progress is being made. However, most progress to date in biomedicine has resulted from experiential learning stemming from the creation and use of reference terminologies; little progress has been made on overcoming the terminology scalable infrastructure challenge.

Apelon, Alameda, California.

THE CHALLENGES

1. Develop terminology-enabled applications: These applications would operate at Web scale on data that was linked to reference terminologies. In effect, these applications would access virtual semantic databases, in which the semantics evolved over time. Successful use of any such application would "pull" development on the remaining challenges, and serve an educational function.
2. Develop longitudinal terminology creation, maintenance and evolution tools and services: These tools and services would help biologists name and relate concepts that "explained" what they were doing—a kind of advanced Web "publication."
3. Understand how representing biology in terminology models is the same and different from representing biomedicine and other domains that develop on the Semantic Web: For example, because they are artificial some potential Semantic Web domains are "arbitrary"; in contrast we like to believe that biology is not arbitrary and is predictable in proportion to our understanding of it. To the degree that this is true, how should this predictability be represented usefully?
4. Create a technology infrastructure that allows biology to proceed on its own while at the same time benefiting from semantic infrastructure in other domains: The potential for reuse of semantics and semantic infrastructure in biomedicine and chemistry is obvious; less obvious is any domain that could be included under the broad notion of phenotype.

TRACTABILITY

Reference terminology development and maintenance can be made tractable by introducing pragmatics, or utility. Typically, the first priority of terminology models is to support scientifically driven aggregation, by defining concept-class membership, or genera; the second priority is to define how otherwise identical things are different, or differentia. The point of genera and differentia is to define similarity and relatedness in scientifically useful ways. While biomedicine is making important progress toward the development and deployment of such terminology models in healthcare, biology will require an information technology infrastructure that supports continuous, scalable evolution of pragmatically driven concepts, their names, and inter-concept relationships. But even with careful articulation of priorities the number of named and formally defined concepts may grow steadily; in biomedicine we are approaching one million authoritative concepts, with several million names.

INCREMENTAL INTEROPERATION

An example of a scalable, tractable domain in biomedicine is laboratory test results. Typically, such results are sent from machines, or human technicians, to the ordering computer as a message with standard syntax as specified by HL7 v. 2 (www.hl7.org); increasingly the contents of the message is becoming standardized as well using LOINC (Logical Observations Identifiers, Names, and Codes), thereby making lab test results more comparable and potentially aggregatable. LOINC is successful because (1) much of laboratory testing is repetitious, (2) it is quickly updated to reflect the ongoing introduction of new tests, thereby preventing "forking," and (3) it is "free" because of its support by the U.S. Government. An additional fact that makes LOINC interesting in the context of biology is that it started out being mostly about "chemistry"—especially "physical chemistry" (the way a clinical sample was analyzed), but it is becoming increasingly about molecular and cell biology, in ways that this audience can anticipate.

The good news is that LOINC is an example of a public domain reference terminology with an ongoing maintenance trajectory. However, it is not deployed on the Web, as a service; instead enterprises make their own use of it in a way that inhibits economies of scale and national data comparability.

The next scalable, tractable domain to be at least partly standardized with Government support will be medications—"drugs," and here part of the computer explanation of medications makes use of the "molecular and cellular biology" of humans and microorganisms. Here again, there is a stable core of medications worth "formalizing" for computers, with new medication attributes and new medications being discovered every day.

In each of these domains, reference terminologies act as reference models that can be used by computers to compare and aggregate collections of lab results and medications. Each terminology model is a mixture of about equal parts of science and "clerical" attributes, both being necessary to support interoperation. But, as with LOINC, medication reference terminology models are not yet available as Web services.

A MODEL OF BIOLOGY

Richard Klausner, then Director of the National Cancer Institute, decried the current trend toward the even finer "splitting" of cancer diagnoses; instead, he argued, researchers should represent what is the "same" in what they are doing relative to their research community at large. He noted, for example, that one had to "go back to" single-celled organisms to find cells with metabolic activity fundamentally different than that found in human cells.

While personal advancement in biology usually results from assertions regarding novel "differences," interoperation can most easily exploit what is today under-represented, namely that degree to which many biological things are the same. Computer science understands how to scalably exploit "sameness" as long as that sameness is clearly separated from "variation." One way to achieve this partition is create a semantic infrastructure that represents the relatively stable parts of biology. For example, excluding microorganisms, our view of species is stable and it should be represented once or a few times and then be maintained in ways that achieve economies of scale. The computer science challenge is how to achieve this, and at the same time relate to the fluid world of microorganism genotyping.

Similarly, notions of biological structure are profound organizers of what would otherwise be seemingly chaotic. Thus, mammals have profound diversity in the appearance of "feet"—think of the human foot, paws, hoofs, claws—yet it is argued that all mammalian feet share the same number of bones.

In this writer's opinion, the problem of representating the "sameness" in biology alongside "variation" is inextricably linked to technology; without an appropriate infrastructure the utility driven semantics cannot be accumulated, maintained and evolved. Without the natural process of that evolution, computer science will not know what to build.

REPRESENTATION

Enough experience exists with the use of DAGs (Directed Acyclic Graphs) and Description Logics to represent formal terminologies to get started by selecting a small repertoire of representations on which to standardize so as to focus attention on the remaining technology challenges. NSF could lead by sponsoring the appropriate consensus conferences; there is ample precedent in computer science for accelerating research through standardization of tools and representations.

CONCLUSION

Computer science has progressed by ignoring semantics and focusing on syntax. Scalable semantics, large reference terminologies, cannot be created and maintained without tools, and their deployment, use and productive evolution will require a technology infrastructure barely contemplated at present. The biggest challenge of all will be coupling the development of that technology with the development of the accompanying semantics; one will not be useful without the other, requiring a shift in research paradigms.

Address reprint requests to:
Mark S. Tuttle
Apelon
151 West Atlantic Avenue
Alameda, CA 94501

E-mail: mtuttle@apelon.com