Towards a Logical Foundation for Bioinformatics Based on Semantic Web Based Scientific Reasoning

Ross D. King
Department of Computer Science
Aberystwyth University
UK
rdk@aber.ac.uk

1. Abstract

The advantages of using logic to represent scientific knowledge have long been understood. Despite this very little scientific knowledge has ever been represented using logic. This is now changing, and the application of the Semantic Web to science is developing a logic-based distributed infrastructure that is integrating large amounts of scientific knowledge. This advance opens up the possibility of utilising the Semantic Web to provide a logical foundation for bioinformatics, and then using this foundation to develop new bioinformatic tools and services. In this paper I sketch out the form of a logical foundation for bioinformatics. Such a foundation is necessary to semantically integrate the existing bioinformatic service infrastructure with the growing amount of bioinformatic knowledge available on the Semantic Web. I describe the need to develop general scientific reasoning tools for the bioinformatic Semantic Web, as existing Semantic Web inference methods are insufficient for scientific reasoning. Finally I suggest the development of novel bioinformatic tools and applications based on logical and general Semantic Web scientific reasoning methods. The goal is to provide biologists with the tools and services needed to meet the demands of 21st century biology.

2. Background

2.1. Logic and Science

With a two and half thousand year tradition logic is the best understood way of representing scientific knowledge. Only logic provides the semantic clarity necessary to ensure the comprehensibility, reproducibility, and free exchange of knowledge [*Tol*]. Use of logic is also necessary to enable computers to play a full part in science: it removes the intractable difficulties with understanding natural language, and enables computational reasoning. Although the advantages of logic for science have long been understood [*Dav*], very little scientific knowledge has ever been represented using logic.

2.2. The Semantic Web

The Semantic Web was born out of a confluence of ideas from computer science, logic, and library science [*Ber*]. The best way to understand the Semantic Web, is not as the standard Web with an extra semantic layer, but rather as a world-wide knowledge base represented in logic. The Semantic Web is becoming a universal publishing platform for scientific knowledge [*Sha*]. *T*he focus of Semantic Web development is now on the logical layer and developing applications.

2.3. Reasoning and the Semantic Web

Like the standard Web, the Semantic Web it can be used to search for information [*Ber*]. The advantage of the Semantic Web is that its information has clearer semantics, enabling information to be found easier. For example, if a human user or a computer are searching for information on "RIF" (the rule interchange format), using the Semantic Web both should be able to easily avoid getting information on the Rif region in Morocco, the company RIF Worldwide, etc. For science the Semantic Web can also provide facilities such as integrating metadata, providing provenance information, integrating publications with original data and analysis methods, etc. Important as these advantages of the Semantic Web will be for science, the real benefits will be in enabling new inferences to be made from the knowledge available on the Semantic Web. *This is because it is these inferences that will enable new types of tools and services* .

There are three basic form of logical inference: deduction, abductions, and induction, and these along with probabilistic reasoning are the basis of all scientific inference. Deduction is the basis of traditional logic, mathematics, and computer science. It is a valid form of reasoning, so if a knowledge base is consistent then only new truths can be inferred. An example of a bioinformatic deduction is the following: rule) if a cell grows it can synthesise tryptophan (P → Q); fact) cell cannot synthesise tryptophan (¬Q); then infer) cell cannot grow (P). Research on deduction has until recently dominated research on inference for the Semantic Web (e.g. [*Hor*] is typical). There are now stable open source and commercial reasoning engines .

Deductive reasoning is insufficient for science as it cannot infer any knowledge that isn't already implicit in a knowledge base. This means that abductive and inductive inference are required to advance science. The easiest way to think about abduction is as deduction in reverse. An example of abduction is: rule) if a cell grows it can synthesise tryptophan (P → Q); fact) cell cannot grow (¬P); then infer) cell cannot synthesise tryptophan (Q). Abductive reasoning is not valid, and therefore new empirical observations are required to ensure the truth of abductive inferences. Very little research has been done on developing abduction for the Semantic Web, but see e.g. [*Col*].

More work has been done on developing induction for the Semantic Web (e.g. [*Ian*]), but it is still an under researched area. In relational learning (RL) there exists a technology which is "pre-adapted" for inductive reasoning over the Semantic Web [*Lis*]. The main technical challenge of adopting RL for the the Semantic Web are: the large amounts of data involved, engineering the inference methods to work over an open, and distributed environment of the Web, and the previous focus of RL on Datalog [*Ull*] rather than description logics [Baa]. Within machine learning RL's position is unusual. It is generally agreed to be theoretically important, yet its practical impact has been low. *The main reason for this is that very little data has been natively represented using logic, this is now changing with the Semantic Web, and RL is becoming a central technology*.

Logical inference and the Semantic Web fit well together. However, as James Clerk Maxwell stated "the true logic of this world is in the calculus of probabilities". By this he meant that all scientific knowledge is essentially probabilistic. The integration of relational learning with probability theory is one of the most exciting areas in machine learning [Get]. The main theoretical issue is that the traditional foundation of probability theory is propositional logic, while some variety of 1st-order predicate logic is required for RL and the Semantic Web.

2.3. Bioinformatics and the Semantic Web

The use of bioinformatic software is essential to modern biology. Typical bioinformatic tasks are: genome annotation, analysing gene expression, protein structure prediction, phylogenetics, metabolomic analysis, etc. The state-of-the-art in bioinformatics is to use sophisticated scripting languages and Web services. This enables the zoo of existing bioinformatic programs to be integrated together, and enables some form of reproducibility.

Bioinformatics is one of the undoubted successes stories of applying the Semantic Web to science. Bioinformatic knowledge makes up a large percentage of the scientific Semantic Web, and many of the problems that makes general Semantic Web reasoning difficult don't apply to bioinformatics:

- A ground truth of scientific knowledge exists.
- A top level ontology have been agreed - the Basic Formal Ontology (BFO). This ensures that specific bioinformatic ontologies are logically compatible, and promotes cross-domain reasoning.
- The bioinformatic Semantic Web is large, but not as large as many other areas of the Semantic Web. It is therefore more computationally tractable.

These advantages have enabled work to proceed on describing biological knowledge using logic, and the European Bioinformatics Institute (EBI), and other large providers of bioinformatic data are now routinely publishing bioinformatics knowledge on the Semantic Web.

However, there is a mismatch between the growing use of the Semantic Web to represent biological knowledge, and the tools and scripts currently used for bioinformatic inference . Traditional bioinformatics software uses *ad hoc* inference, and the assumptions (logical and biological) they make are rarely explicit. This is unsatisfactory, as the hard-coding of scientific assumptions makes them obscure, difficult to understand, and difficult to change. It also precludes biologists checking these assumptions. From a formal point of view bioinformatic programs are invariably making logical inferences: deductions, abductions, inductions, with perhaps a probabilistic element. The form of these inferences need to be clarified if bioinformatics is ever to have a solid scientific foundation.

3. The Vision

3.1. A logical foundation for bioinformatics.

The goal is to semantically integrate the existing bioinformatic service infrastructure with the growing amount of bioinformatic knowledge available on the Semantic Web. This will have two parts:

- Clarification of the semantics of existing bioinformatic software. The assumptions and inference mechanisms used by most existing bioinformatic software are not explicit. The aim is to make them explicit for the main classes of bioinformatic software.
- Formation of general purpose implementations of existing bioinformatic software. Given known assumptions and using general purpose Semantic Web inference tools we will implement standard bioinformatic tools. We expect these tools to be less efficient than the current *ad hoc* ones, but the implementations will demonstrate that the bioinformatic task has been logically defined and understood.

To illustrate what we mean below I will sketch what this would mean for two separate problem classes of bioinformatic software:

1. Predicting the structure of a protein domain based on sequence homology. This is typically the first step in a structural bioinformatics investigation. The computation is as follows: the distance between the target domain's sequence and all the domain sequences in the database of known structure is first calculated, then the target's structure is predicted to be the same as that of the closest sequence in the database. The biological rationale for this is based on the conservation of domain structure by evolution. Logical analysis reveals that many assumptions are made concerning the conservation of structure during evolution. It also reveals that the inference method is abductive. What is being abduced is the existence of a common ancestral domain shared by both the target domain and the domain with the closest sequence in the database, but by no other domains in the database
2. Predicting protein function from a micro-array profile. This is a common task in functional genomics. The goal is to predict the function of a gene by generalising patterns observed in transcriptomic experiments. The problem is technically interesting for machine learning as protein functions are organised in class hierarchy using gene ontology, and proteins may have more than one function. Logical analysis reveals that a number of implicit assumptions are made when applying machine learning to this problem. The most important of these is the closed-world assumption: if a protein is not known to have a specific function then it doesn't have that function. This assumption makes learning much more efficient as it generates large amounts of negative examples. However, it is in general a false assumption, as proteins may have functions which we do not yet know. This closed-world assumption clashes with the use of the semantic web language OWL. For the prediction task the inference mechanism is induction, as transcriptomic patterns associated with gene ontology classes are generalised.

3.2. Scientific Reasoning for the Semantic Web

There is a need to develop new inference mechanisms designed that takes full advantage of the logical infrastructure of the Semantic Web. These non-deductive reasoning methods will necessarily be based on Relational Learning [*De1*] to be powerful enough to be able to reason using the logics used to represent scientific knowledge in the Semantic Web. The Relational Learning methods will include: Abductive Logic Programming, Relational Machine Learning, and Probabilistic methods.

3.3. Novel bioinformatic tools

The key motivation for providing a logical foundation for bioinformatics and developing general purpose scientific inference mechanisms is not simply to improve our understanding of bioinformatic software, important as that is, but rather to use this understanding to develop new improved bioinformatic tools and services. Taking the same two examples from above, logical analysis will enable new variants to be envisaged, and these can be implemented using the general purpose scientific reasoning methods developed in the following ways:

1. Predicting protein domain structure on sequence homology. When it is realised that the basic logical inference involved is abduction of a common ancestral sequence, plus an assumption of conservation of function, it is possible to envisage variants of the basic bioinformatic method which are biologically more realistic, and which will result in more accurate predictions. For example it is clear that it should not be just a single ancestral sequence should be abduced, but rather a population of ancestors; and that the use of this population for prediction should be weighted by their evolutionary distance as estimated by the sequence metric. This produces a complex probabilistic relational graph similar to that generated in probabilistic relational learning [*De2*]. Logical analysis of the problem can therefore be used to develop a representation then be solved using general purpose statistical relational learning methods.
2. The problem of predicting protein function from a micro-array profile is currently normally tackled using propositional learning methods, and these methods are generally limited to using only a limited set of attributes for prediction [*De1*]. Logical analysis reveals that there is a large number of important sources of information that should be used in prediction: the gene ontology hierarchical class structure, the existence of multiple functions for the same protein (multiple-labels), that the micro-array data comes from multiple experiments often consisting of small time-series, the metabolic network that integrates the enzymes, the signalling networks that integrate the signalling pathways, the genome structure, etc. The bioinformatic semantic web will make collection and logical and biological integration of these sources simple to do. Then general purpose relational learning algorithms, plus the closed-world assumption, can be used to exploit all available sources of information for prediction – a basic law of reasoning is to use all available relevant information [*Jay*].

4. References

[*Baa*] Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P., eds.: The Description Logic Handbook. Cambridge University Press (2003)

[*Ber*] Berners-Lee, T., Hendler, J., Lassila, O. (2001) The Semantic Web. *Sci. Am*. 284, 34-43.

[*Col*] Colucci, S., Di Noia, T., Di Sciascio, E., Donini, M.F., & Mongiello, M. (2005) Concept abduction and contraction for semantic-based discovery of matches and negotiation spaces in an e-marketplace. *Electronic Commerce Research and Applications* **4**, 345–361

[*Dav*] Davis, M. (2000) *The Universal Computer: The Road from Leibniz to Turing*. WW Norton.

[*De1*] De Raedt, L. (2008) *Logical and Relational Learning*. Springer-Verlag

[*De2*] De Raedt, L., Frasconi, P., Kersting, K.,Stephen Muggleton, S. (2008) *Probabilistic Inductive Logic Programming*. Springer-Verlag

[*Hor*] Horrocks I., and Sattler U. A Tableau Decision Procedure for SHOIQ. *J. of Automated Reasoning*, 39(3):249-276, 2007.

[*Ian*] Iannone, L., Palmisano, I., Fanizzi, N. (2007) DL-FOIL Concept Learning in Description Logics. *Applied Intelligence.* 26, 139-159.

[*Jay*] Jaynes, E.T. (2003) Probability theory: The logic of science. Cambridge

[*Lis*] Lisi, F. A. and Esposito, F. 2008. Foundations of onto-relational learning. In Proc. of ILP'2008, F. Zelezny and N. Lavrac, Eds. Lecture Notes in Computer Science, vol. 5194. Springer, 158-175.

[*Sha*] Shadbolt, N., Hall, W., Berners-Lee, T., (2006). The semantic web revisited. IEEE Intelligent Systems.

[*Tol*] Toulmin, S. (2003) The Philosophy of Science. In *Encyclopaedia Britannica Deluxe Edition 2004 CD* (Encyclopaedia Britannica UK, London).

[*Ull*] J. D. Ullman. *Principles of Database and Knowledge-Base Systems*. Computer Science Press, 1989.