Formalising Phylogenetic Experiments: Ontologies and Logical Inference

Ross D. King & Larisa N. Soldatova
Department of Computer Science,
University of Wales, Aberystwyth, UK.
{rdk, lss}@aber.ac.uk

Abstract

The ontology of scientific experiments EXPO formalises the generic concepts of experimental design, methodology, and results representation. We describe an application of EXPO to describe phylogenetic experiments, focusing on a case study involving Solonedons. We explain how the details of the experiment were formalised using EXPO We argue that abductive inference is the basis of evolutionary phylogenetics, that inductive inference is necessary to generalise phylogenetic conclusions from sequences to genomes, and that deductive inference is also often required. This is novel because phylogenetic experiments are generally thought to be based on purely probabilistic methods. The recognition that different forms of logical inferences are taking place may enable novel techniques from logic to be applied.

1. Introduction

The central interests of the Computational Biology group of the University of Wales, Aberystwyth are in formalising and automating scientific experiments [Soldatova & King, 2005; King *et al*., 2004]. We have developed a generic ontology of experiments, EXPO, was to support our research [Soldatova & King, 2006]. EXPO links a relevant subset of the upper ontology SUMO with subject-specific ontologies of experiments by formalising the generic concepts of experimental design, methodology, and results representation. EXPO is expressed in the W3C standard ontology language OWL-DL.

Along with our interests in automating and formalizing science, we are also interested in both the application of phylogenetic methods to biological problems and in developing new phylogenetic methods. As it is important that work on formalising science is done in contact with actual science, it has been natural for us to use phylogenetics as a test-bed.

Phylogenetics is the reconstruction of the evolutionary relationships (that is, the phylogeny) of a group of taxa, such as species [Nature]. Work in phylogenetics used to be generally done based on phenotypic features of organisms considered to be evolutionary stable. It is now generally done using genetic techniques; except when this is impossible, e.g. in fossils. Many important phylogenetic questions remain unanswered, e.g. the relationship between the main animal phyla [Valentine, 2004].

In this paper we argue that formalising the details of phylogenetic experiments makes their results more explicit, the knowledge generated more reusable, and the experiments more repeatable. We also argue that teasing out the different forms of logical inference involved in phylogenetic experiments opens up new methodological opportunities missed by the assumption that the existing probabilistic methods are sufficient

2. A case study: Solonedons

In Soldatova & King (2006) we used EXPO to annotate an experiment investigating the phylogenetic status of the mammalian species *Solenodon cubanus* and *Solenodon paradoxus*. Solenodons are endangered insectivores that inhabit the forests of Hispaniola and Cuba. Their phylogenetic relationship with other mammals has long been a matter of controversy [Roca, et al. 2004]. Here we briefly sketch the use of the ontology EXPO to annotate this phylogenetic experiment. This work differs from (Soldatova & King, 2006) in emphasizing the various forms of logical inference that are implicitly involved in phylogenetic experiments.

2.1. EXPO

A small part of the EXPO annotation of the Solonedon experiment is given in Figure 1.

One advantage of the EXPO formalism is that it forces the explicit expression of research hypotheses, negative hypotheses, alternative hypotheses and all the available evidences to support or reject them. EXPO can also be used to make explicit the argumentation used to make research assumptions and conclusions. In addition EXPO serves as a basis for formalising background knowledge about a research domain - in cases where it has not previously been formalised in an ontology. Once information about an experiment has been formalised inference methods can be used to reason about the validity of conclusions, and EXPO helps to define predicates and rules required for such reasoning.

Below we discuss how different types of logic inference are used in phylogenetics.

<scientific investigation>: Discovery of the phylogeny of *Solenodon paradoxus* and *Solenodon cubanus*
<investigation metadata>:
<DC: title> Mesozoic Origin of West Indian Insectivores
<DC: author>: Roca, A.L., Bar-Gal, G.K., Eizirik, E., Helgen, M.K. *et al.*
<DC: reference>: Nature, 429: 649-651 (2004)
<DC: subject>: Zoology <DDC(Dewey) classification>: 599: mammalology
<motivation>: It is important to produce more experiment data and analyse the phylogeny of Solenodon paradoxus and Solenodon cubanus because of the threat of their extinction.
<problem analysis>: The phylogeny of the Solenodons has long been ambiguous
<null hypothesis> H01: <representation>:
<linguistic expression>: <natural language>:
"Some have suggested a close relationship to soricids (shrews) but not to talpids"
<linguistic expression>: <arificial language>:
So, Sh, T, An ∈ mammalian.
∀So . ∀Sh . ∀T . ∃An . solenodon(So) ∧ soricoidea(Sh) ∧ talpoidea(T) ∧ ancestor(An, So) ∧ ancestor(An, Sh) ∧ ¬ancestor(An, T).
Comment: Solenodons and the soricoisea share a common ancestor that the talpoidea do not have
<research method>: <experiment method>
<scientific experiment>: <physical experiment>: <hypothesis forming>
<object of experiment>: Living and dead specimens of the species *Solenodon paradoxus* and *Solenodon cubanus.*
<experimental equipment>:
<hardware>: Qiagen column-based DNA cleanup kit
PCR primers supplier "high-fidelity Taq-Gold (ABI)" sequences
Microcoson-50 for PCR product purification
ABI 3700 automated sequencer
<software>: PAUP*4.0b10 (Altivec)
<experiment conclusion> C1:
<logic of inference >: deduction
<representation>: <linguistic expression>: <natural language>:
There existed a mammal that is the ancestor of: *Solenodons*, *Soricoidea*, *Talpoidea*, *Erinaceidea*, and which is not the ancestor of any other mammal.
. . . . . .
<experiment conclusion> C5:
<logic of inference >: non-monotonic logic
<representation>: <linguistic expression>: <natural language>:
"our results lend support to an alternative proposal that Cuban solenodons be classified as a distinct genus *Atopagale*."
<linguistic expression>: <artificial language>:
retract(species(solenodon, cubanus)) ∧ assert(species(atopagale, cubanus)) ∧ assert(taxon(genus, atopagale)) ∧ retract(is_a(solenodon_cubanus, solenodon)) ∧ assert(is_a(solenodon_cubanus, atopagale)) ∧ assert(is_a(atopagale, solenodontidae).

Figure 1

2.2. Abduction

We argue that abductive inference is central to modern evolutionary based phylogenetics. This can be seen in evolutionary definition of a taxon (grouping of organisms): "that all members of a taxon are descendants of the nearest common ancestor (monophyly sensu stricto)" [Mayer, 1982]. We express this in logic as:

$\forall A \,.\, A \in taxon1 \Rightarrow (\exists Ancestor \,.\, \forall B \,.\, B \notin taxon1 \wedge ancestor(Ancestor, A) \wedge \neg ancestor(Ancestor, B))$.

This definition is based on the abductive inference of the existence of an ancestor organism not shared by any other taxon. An applied example of this from the Solenodon work is:

$So, Sh, T, E, An, X \in mammalia$

$\forall So \,.\, \forall Sh \,.\, \forall T \,.\, \forall E \,.\forall X \,.\, \exists An \,.\, solenodon(So) \wedge soricoidea(Sh) \wedge talpoidea(T) \wedge erinaceidea(E) \wedge \neg solenodon(X) \wedge \neg soricoidea(X) \wedge \neg talpoidea(X) \wedge \neg erinaceidea(X) \wedge ancestor(An, So) \wedge ancestor(An, Sh) \wedge ancestor(An, T) \wedge ancestor(An, E) \wedge \neg ancestor(An, X)$

Which states that there existed a mammal that was the ancestor of: Solenodons, Soricoidea, Talpoidea, Erinaceidea, and which is not the ancestor of any other mammal (see Fig 1).

N.B. the science of Cladistics [Valentine, 2004] predates the rise of molecular phylogenetics and is also based on the abduction of ancestral organisms. Cladistics was used in the Solonedon paper to analyse fossil evidence [Roca, et al. 2004]

2.3. Induction

We also argue that Phylogenetics requires inductive inferences. This is because general conclusions about the relationship of organism are generally based on one (or at most a few) sequences from each organism - not from the full genome. For example in a distance based phylogenetic method it is inductively inferred that:

$distance(seq\_a\_species\_s1, seq\_a\_species\_s2) = distance(species\_s1, species\_s2)$.

In the Solenodon work we studied, the phylogenetic relationships between the two Solenodon species and other mammals were inductively inferred by use a small set of mitochondrial and ribosomal gene sequences.

Inductive inference is also important in phylogenetics because the older, Linnaean non-evolutionary based, definition of a taxon is inductive. This definition is based on similarity: "that the members of each taxon are each other's nearest 'relatives' (that is, most similar to each other)" [Mayer, 1982]. This definition leads to use of clustering ("classification" in statistics, "unsupervised learning" in machine learning) methods to define taxa. Given an induced cluster the most natural way to define a taxon is to define a set of features that must be present in an organism to place it in a specified taxon. This is what was traditionally done for higher level taxa. Interestingly, however, it is not what was done for the taxa species and genera. For these a cluster was defined by similarity to a "type specimen". This is an example organism (usually preserved in a museum) that is asserted to be of the specified taxon. Once such type specimens exist, and there is some way to measure organism similarity, then the correct taxon for an organism can be computed. The use of type specimens has the feel of case based reasoning.

Use of a type specimen and a similarity based measure of a taxon can be expressed as:

$A \in taxon2 \Rightarrow (\exists Type \,.\forall B \,.\, Type \in taxon2 \wedge B \notin taxon2 \wedge (distance(A, Type) < distance(A, B)))$

Over and above the above uses of induction, the central induction in evolutionary science is that evolution took place: "that the Linnaean hierarchy of taxa is consistent with the inferred phylogeny" [Mayer, 1982]. Therefore, the evolutionary definition of taxon1 is identical to the clustering one.

$induction(taxon1 == taxon2)$

2.4. Deductions

Phylogenetic experiments may also involve deductive inference. In our annotation of the Solonedon experiment we interpreted the text to be using the following definition of how long two taxa have to have diverged to be separate families:

$diverged(X, Y, Date) \wedge Date > 20000000 \wedge Date \leq 30000000 \Rightarrow different\_family(X, Y)$.

It also inferred inductively the fact:
$diverged(solonodon\_cubanus, solonodon\_paradoxus, 25000000)$.

The authors did not however deductively infer:
$different\_family(solonodon\_cubanus, solonodon\_paradoxus)$.

They instead inferred:

$different\_genus(solonodon\_cubanus, solonodon\_paradoxus)$.

This illustrates that one advantage of formalisation - the identification of errors

3. Discussion

One aspect of the Solenedon paper that could not be represented using traditional logic is the conclusion described above that:

different_genus(solonodon_cubanus, solonodon_paradoxus).

The paper concludes that Cuban Solenodons be classified in a new genus Atopagale. In the standard Linnean classification Cuban Solenodons belong to the species *Solenodon cubanus* i.e. are in the genus Solenedon. The conclusions of the paper are that a new genus Atopagale should be created, that the Cuban Solenodon species be renamed *Atopagale cubanus*, that *Atopagale cubanus* beplaced in Atopagale, and that the species *Solenodon cubanus* be removed from the genus Solenodon These inferences require non-monotonic logic.

A probabilist might argue that the only form of inference that is required in phylogenetic experiments is probabilistic inference. This argument certainly has some merit and is the traditional view. However, we argue that very generality of probabilistic inference obscures the different aspects of the inferences that are taking place. In addition, recognition that abductive, inductive, and deductive inferences are taking place enables novel techniques from logic to be applied. We also argue that the logical view also meshes much more cleanly with the use of ontologies, and that ontologies are becoming increasingly important in phylogenetics. Finally, the convergence of description logics with probabilities is one possible approach that may enable the best of both logical and probabilistic reasoning [Lukasiwicz, 2007].

References

King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G.K., Bryant, C.H., Muggleton, S.H., Kell, D.B., & Oliver, S.G. (2004) Functional genomic hypothesis generation and experimentation by a robot scientist. Nature, 427, 247-252.

Lukasiwicz, T. (2007) Probabilistic description logics for the semantic web. INFSYS Research Report 1843-06-05

Mayer, E. (1982) The growth of biological thought. Harvard University Press.

Nature : http://www.nature.com/nrg/journal/v4/n2/glossary/nrg999_glossary.html

Roca, A.L., Bar-Gal, G.K., Eizirik, E., Helgen, M.K., Maria, R. at all. (2004) Mesozoic origin for West Indian insectivores. Nature, 429, 649-651

Soldatova, L.N. & King, R.D. (2005) Are the current ontologies used in biology good ontologies? Nature Biotechnology 23, 1095-1098.

Soldatova, L.N. & King, R.D. (2006) An Ontology of Scientific Experiments. Journal of the Royal Society Interface 3, 795-803.

Valentine, J.W. (2004*) On the origin of Phyla.* Chicago