BMC Medical Genomics

RESEARCH Open Access

Advancing data reuse in phyloinformatics using an ontology-driven Semantic Web approach

Maryam Panahiazar[1,4*], Amit P Sheth[1], Ajith Ranabahu[1], Rutger A Vos[3], Jim Leebens-Mack[2,4]

From IEEE International Conference on Bioinformatics and Biomedicine 2012
Philadelphia, PA, USA. 4-7 October 2012

Abstract

Phylogenetic analyses can resolve historical relationships among genes, organisms or higher taxa. Understanding such relationships can elucidate a wide range of biological phenomena, including, for example, the importance of gene and genome duplications in the evolution of gene function, the role of adaptation as a driver of diversification, or the evolutionary consequences of biogeographic shifts. Phyloinformaticists are developing data standards, databases and communication protocols (e.g. Application Programming Interfaces, APIs) to extend the accessibility of gene trees, species trees, and the metadata necessary to interpret these trees, thus enabling researchers across the life sciences to reuse phylogenetic knowledge. Specifically, Semantic Web technologies are being developed to make phylogenetic knowledge interpretable by web agents, thereby enabling intelligently automated, high-throughput reuse of results generated by phylogenetic research. This manuscript describes an ontology-driven, semantic problem-solving environment for phylogenetic analyses and introduces artefacts that can promote phyloinformatic efforts to promote accessibility of trees and underlying metadata. *PhylOnt* is an extensible ontology with concepts describing tree types and tree building methodologies including estimation methods, models and programs. In addition we present the *PhylAnt* platform for annotating scientific articles and NeXML files with *PhylOnt* concepts. The novelty of this work is the annotation of NeXML files and phylogenetic related documents with PhylOnt Ontology. This approach advances data reuse in phyloinformatics.

Background

Forty years ago, Theodosius Dobzhansky asserted “Nothing in biology makes sense except in the light of evolution” [1], and phylogenetic trees offer a historical representation of the evolutionary process. Since Darwin and Haeckel published their iconic tree figures some 150 years ago [2,3] phylogenies have provided the historical framework for elucidating the evolution of form and function [4]. In addition to estimating organismal relationships and the timing of gene duplications [5,6], phylogenies can be applied to many more research questions. For example, they can be used to inform prediction of protein function [7] and investigations of disease transmission [8]. More generally, phylogenies provide a unifying context across the life sciences for investigating the diversification of biological form and function from genotype to phenotype.

The increased interest in using and reusing phylogenies has exposed major limitations in the accessibility and reusability of published phylogenetic trees and the data used to estimate these trees. Most published phylogenetic trees can only be found in text and graphical format embedded in printed or electronic research publications [9,10]. As a consequence, these trees are typically inaccessible for semantic processes, including web-based identification and acquisition of trees, analytical methods, or the data on which phylogenetic inferences are based. This greatly limits the ability of biologists to reuse gene and species trees in meta-analyses with other structured sources.

There is a wealth of information that surrounds each phylogenetic study, including comparative data such as morphological character state matrices and nucleotide or amino acid sequence alignments, methodological descriptions such as substitution model and provenance

* Correspondence: mary@knoesis.org
[1]Ohio Center for Excellence in Knowledge-enabled Computing (kno.e.sis) College of Computer Science and Engineering, Wright State University, Dayton, OH, USA
Full list of author information is available at the end of the article

information. All of this information is represented in a variety of different formats ranging from unstructured data such as texts and images in published technical reports and academic articles to semi-structured data such as tables and key delimited records and structured data such as database entries, and XML files. This variation of formats poses informatics challenges to the integration of diverse data and the generation of federated queries to answer specific research questions.

Here we present results to promote an ontology-driven, semantic problem-solving solution for phylogenetic analyses and downstream use of phylogenetic trees. We have constructed a network of concepts and defined them in an ontology, *PhylOnt*, and provide examples for how these concepts can be used to annotate publications and data files. *PhylOnt* is an extensible ontology that describes the methods employed to estimate trees given a data matrix, models and programs used for phylogenetic analysis and descriptions of phylogenetic trees as well as provenance information.

The common vocabulary included in *PhylOnt* will facilitate the integration of heterogeneous data types derived from both structured and unstructured data sources. Annotation tools for tagging *PhylOnt* terms in scientific literature and NeXML formatted data files are also presented. NeXML is an exchange standard for representing taxa, phylogenetic trees, character matrices (e.g. sequence alignments) and associated metadata [11]. As such, well annotated NeXML files could contain the minimum information about a phylogenetic analysis (MIAPA) [4] necessary to enable reproducibility and reuse of phylogenetic inferences.

In addition, we evaluate *PhylOnt* using formal metric-based and annotation-based approaches. This assessment indicates that more than half of the connections between *PhylOnt* classes are information-rich. Further, an analysis of exemplar publications indicates that for phylogenetic operations, methods, models and programs the majority of phylogenetic concepts can be accurately annotated using *PhylOnt*.

Related work

The work described here builds on the needs assessment described by Stoltzfus et al [10] and our research previously presented at the IEEE ICSC 2011 [12], iEvoBio 2011 [13], the W3C Workshop on Data and Services Integration [14], Translational Medicine Conference at AMIA 2012 [15,16] and IEEE International Conference on Bioinformatics and Biomedicine [17]. Recent "PhyloTastic" hackathons [18] have also developed resources to promote the reuse of published trees and underlying metadata.

Other prior art with regard to the Semantic Web-ready definition of phylogeny-related concepts exists in the form of previously published ontologies, most notably the Comparative Data Analysis Ontology (CDAO) [19] and the Embrace Data And Methods (EDAM) ontology [20]. CDAO is an ontology that describes fundamental data and transformations commonly found in the domain of evolutionary analyses. CDAO [19] includes concepts relevant to phylogenies such as nodes, edges, branches, and networks, but concepts relating to phylogenetic analysis methods or provenance are omitted. EDAM [20] is an ontology developed for general bioinformatics concepts including operations, topics, types and formats. EDAM includes phylogeny-related concepts but phylogenetic analysis terms relating to methods, models and programs are either not reported in EDAM or have not been explicitly defined under a correct hierarchy for phylogenetic analysis purposes. *PhylOnt* aims to cover the general concepts necessary to describe phylogenetic analyses. These ontologies are explained and compared with *PhylOnt* in [17].

Methods

PhylOnt aims to characterize selected "phylogenetic resource" concepts and the relationships among these concepts. In this context, we define a "phylogenetic resource" as any uniquely identifiable object or procedure from the domain of phylogenetic research, ranging from the granular, e.g. a specific node in a tree, to the holistic, e.g. a study, or a step in an analysis workflow. *PhylOnt* includes concepts for estimation programs, models of evolution, methods of analysis, search algorithms, support assessments, and relevant provenance information. *PhylOnt* will grow as new tree estimation technologies are developed and used in published phylogenetic studies. Developing an ontology and using it to annotate the data and services in analysis workflows can provide a foundation for other semantic technologies, such as concept-based searches and comprehensive federated queries over data sources.

Systematic approach for ontology development

In developing *PhylOnt* we worked closely with phylogeneticists and computer scientists to iteratively validate the ontology based on community feedback. As shown in Figure 1, development of *PhylOnt* started with data collection and organization of concepts in relational diagrams. Concept maps drawn from the primary literature included descriptions, properties, metadata, usage of concepts and relations between them. Subsequently, the version of *PhylOnt* presented here was developed in Protege 4.1.0 (Figure 2), which supports the Web Ontology Language (OWL). *PhylOnt* is accessible at NCBO through BioPortal [21]. A phylogentics domain specific extension of the Kino annotation package [12,13] was used in *PhylAnt* platform to facilitate annotation and faceted search over the annotated resources including scientific literature and NeXML format data files.

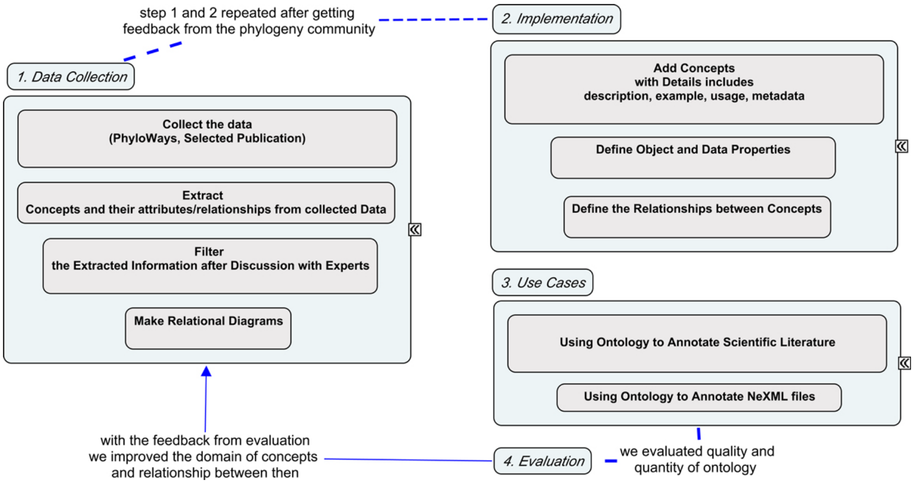
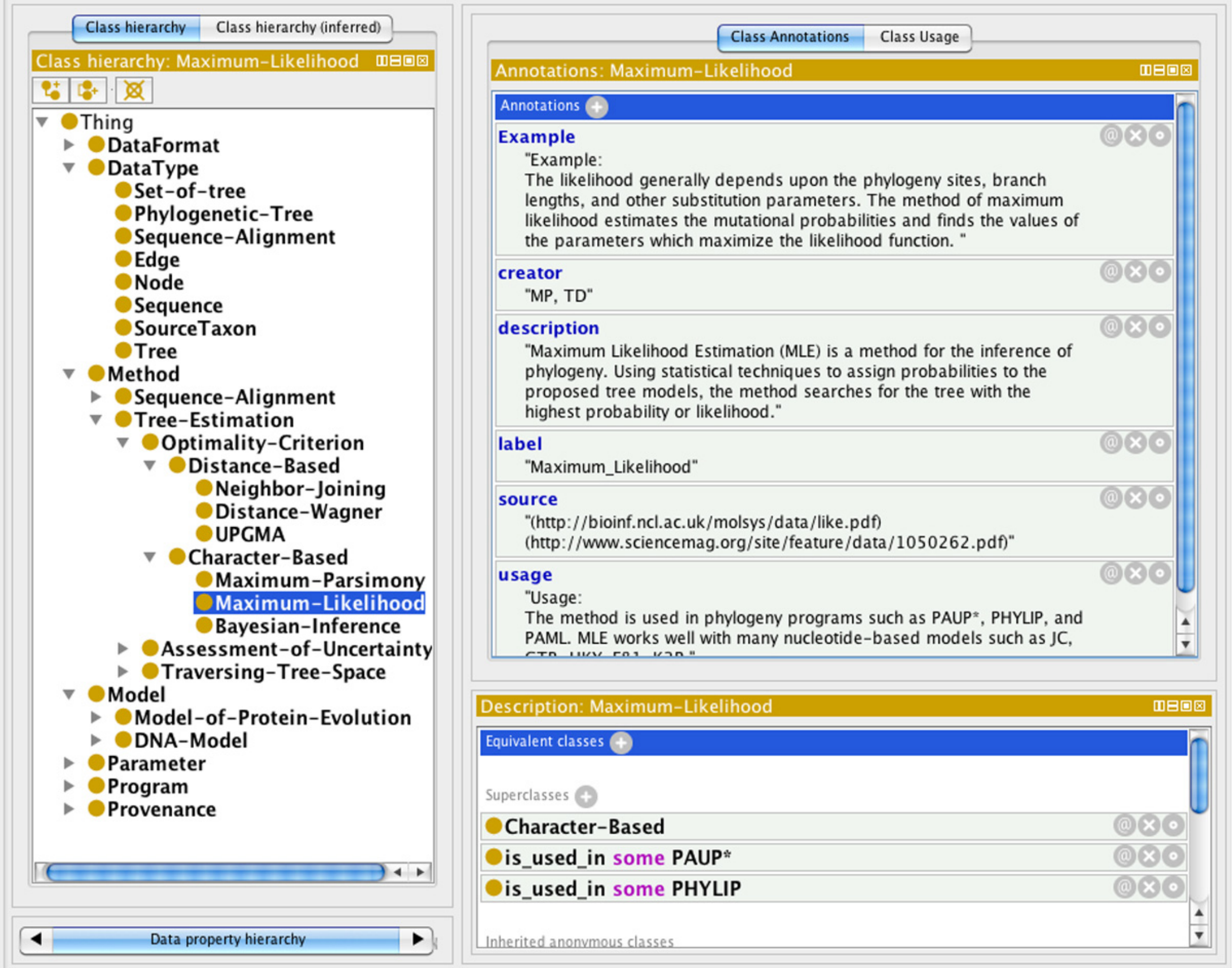

step 1 and 2 repeated after getting
feedback from the phylogeny community
2. Implementation
1. Data Collection
Collect the data
(PhyloWays, Selected Publication)
Extract
Concepts and their attributes/relationships from collected Data
Filter
the Extracted Information after Discussion with Experts
Make Relational Diagrams
Add Concepts
with Details includes
description, example, usage, metadata
Define Object and Data Properties
Define the Relationships between Concepts
3. Use Cases
Using Ontology to Annotate Scientific Literature
Using Ontology to Annotate NeXML files
with the feedback from evaluation
we improved the domain of concepts
and relationship between then
4. Evaluation
we evaluated quality and
quantity of ontology

Figure 1 Systematic approach for ontology development. Our phylogenetics ontology development efforts started with data collection with experts in the field, making analytical diagram for the key concepts such as methods, models and programs. The ontology was formally constructed using Protege. Use cases were employed to evaluate the benefit of *PhylOnt* concepts for annotation of published phylogenetic trees, their estimation and underlying data.


Class hierarchy
Class hierarchy (inferred)
Class hierarchy: Maximum-Likelihood
Data property hierarchy
Class Annotations
Class Usage
Annotations: Maximum-Likelihood
Annotations
Description: Maximum-Likelihood
Equivalent classes
Superclasses
Inherited anonymous classes

- Thing
  - DataFormat
  - DataType
    - Set-of-tree
    - Phylogenetic-Tree
    - Sequence-Alignment
    - Edge
    - Node
    - Sequence
    - SourceTaxon
    - Tree
  - Method
    - Sequence-Alignment
    - Tree-Estimation
      - Optimality-Criterion
        - Distance-Based
          - Neighbor-Joining
          - Distance-Wagner
          - UPGMA
        - Character-Based
          - Maximum-Parsimony
          - Maximum-Likelihood
          - Bayesian-Inference
      - Assessment-of-Uncertainty
      - Traversing-Tree-Space
  - Model
    - Model-of-Protein-Evolution
    - DNA-Model
  - Parameter
  - Program
  - Provenance

Example
"Example:
The likelihood generally depends upon the phylogeny sites, branch lengths, and other substitution parameters. The method of maximum likelihood estimates the mutational probabilities and finds the values of the parameters which maximize the likelihood function. "

creator
"MP, TD"

description
"Maximum Likelihood Estimation (MLE) is a method for the inference of phylogeny. Using statistical techniques to assign probabilities to the proposed tree models, the method searches for the tree with the highest probability or likelihood."

label
"Maximum_Likelihood"

source
"(http://bioinf.ncl.ac.uk/molsys/data/like.pdf)
(http://www.sciencemag.org/site/feature/data/1050262.pdf)"

usage
"Usage:
The method is used in phylogeny programs such as PAUP*, PHYLIP, and PAML. MLE works well with many nucleotide-based models such as JC, GTR, HKY, F81, K2P."

Superclasses
- Character-Based
- is_used_in some PAUP*
- is_used_in some PHYLIP

Figure 2 PhylOnt implemented with Protégé. The *PhylOnt* ontology as represented in Protege. The ontology includes descriptions of classes, definitions, properties, metadata, usage of classes with an example and relations among them.

Data collection

Data resources in phylogenetic studies can be classified into primary and metadata categories. Primary data exist as published data files, literature with text, images, Excel files, and other supplementary materials. Primary data can also refer to methods, models, programs and even parameters used in applications and web services. Metadata includes information such as when and where the primary data were created. This information plays a very important role in enabling reusability.

To perform data extraction, a well-framed approach was required to identify and capture steps in phylogenetic workflows described in published phylogenetic studies [17]. We used *PhyloWays* [22], as a set of interpreted phyloinformatic workflows described in the primary phylogenetics literature. We identified all the information required to repeat the analysis presented in the *PhyloWays* papers, including the phylogeny estimation programs used in each paper, methods of analysis, evolutionary models and provenance information. These descriptions paved the way for classification of concepts associated with phylogenetic data (including provenance information), phylogenetic workflows, and the results of phylogenetic analysis.

Based on discussions with domain experts, literature reviews and the data in *PhyloWays* we then created concept maps describing methods of phylogenetic analysis, evolutionary models used in applications of these methods, and widely used phylogenetic software.

Methods of phylogenetic analysis

Phylogenetic methods vary considerably in approaches for assessing alternative hypotheses (i.e. trees), traversing through the complex universe of alternative hypotheses (i.e. tree and model parameter landscapes) and characterizing the degree of support for an optimized solution. As shown in Figure 3, a hierarchical classification of optimality criteria, search algorithms and uncertainty assessment concepts is implemented in *PhylOnt*. For example, tree inference methods based on maximum parsimony, maximum likelihood or Bayesian statistics rely on the analysis of a homologized character state matrix, whereas UPGMA, neighbor joining and distance-Wagner are based on sets of pairwise distances that may be estimated from a character state matrix or computed in some other way.

The universe of possible trees is extremely complex and identifying the optimal tree in this tree landscape is an NP-hard computational problem. Therefore, there is a variety of heuristic approaches for traversing the tree space in search of the optimal tree. Most maximum parsimony and maximum likelihood analysis methods build an initial tree and then iteratively test for improvement by rearranging the tree topology using branch-swapping algorithms such as nearest neighbor interchange (NNI), subtree pruning and regrafting (SPR), tree bisection and reconnection (TBR), or combinations thereof. Bayesian inference methods also include a branch-swapping process within a Markov Chain Monte Carlo (MCMC) strategy for sampling tree space.

Assessment of support for a phylogenetic inference is key in deciding whether an optimized solution is acceptable [23]. Bayesian inference methods provide posterior probabilities for the relationships conveyed in a phylogenetic tree, whereas other methods typically use bootstrap or jackknife resampling to assess the degree of support for hypothesized relationships. Resampling

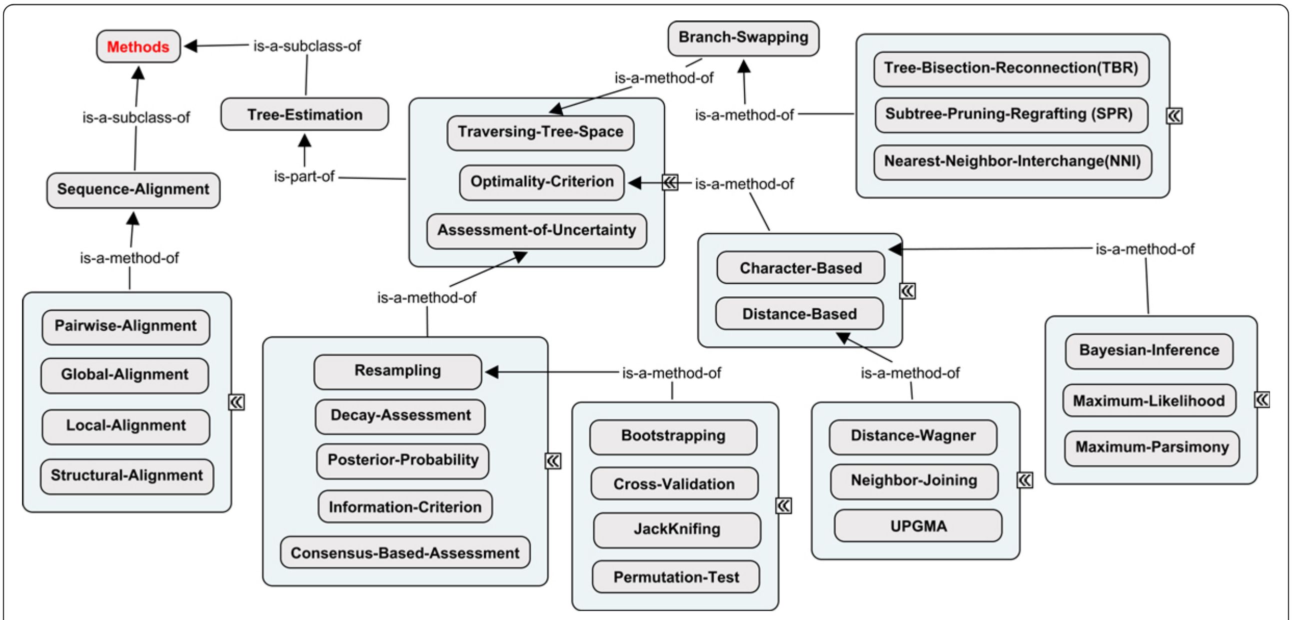


Figure 3 Methods in phylogenetic studies. A hierarchy of concepts used to describe methods commonly used in phylogenetic studies.

approaches can be combined with MCMC sampling in Bayesian analyses and the process of randomly resampling the original data matrix typically reduces posterior probabilities relative to those reported for MCMC searches without resampling [4].

Models in phylogenetic analysis

All phylogenetic analyses are performed with an explicit or implicit model of character evolution. Maximum likelihood, Bayesian inference and most distance-based methods rely on nucleotide or amino acid substitution models. Branch lengths for phylogenetic trees often represent time or evolutionary change. Correct interpretation of branch lengths requires an understanding of the models used to estimate time or evolutionary change. Separate substitution models are used for analyses of DNA and protein sequence alignments. Nucleotide substitution models include JC69, K80, HKY85, SYM, F81, and GTR [24,25]. Commonly used amino acid substitution models include PAM [26], JTT [27] and WAG [28]. Gene sequences typically include conserved domains and less conserved regions. The resulting among-site variation in substitution rates is often modeled in phylogenetic analysis of either nucleotide or amino acid alignments using a discrete approximation of the gamma distribution [29], a fraction of invariant sites [30], or a combination thereof. Both of these forms of rate variation can be layered upon the nucleotide and amino acid substitution models described above. Figure 4 shows a hierarchy of concepts used to describe evolutionary models most commonly used in phylogenetic studies.

Phylogenetic methods and the models they use are constantly changing as the phylogenetics community works to make more accurate and precise inferences about relationships and evolutionary processes. Therefore *PhylOnt* is necessarily incomplete, but easily extended to include additional models.

Programs in phylogenetic analysis

At time of writing, there are approximately 400 phylogeny packages and more than 50 free web servers for phylogenetic analysis [17]. *PhylOnt* currently identifies the most commonly used phylogenetic inference programs such as MrBayes [31], and PAUP* [32]. Programs can be categorized based on the methods they use. For example, PAUP* can be used to perform most major methods of analysis such as maximum parsimony and maximum likelihood. For more details about the programs, such as description for each and relation between programs, models and methods readers are referred to the *PhylOnt* project page on BioPortal [21].

PhylAnt, a platform for semantic annotation, indexing and searching of phylogenetic resources

Semantic annotation maps target data resources to concepts in ontologies. In the process of annotation, extra information is added to the resource to connect it to its corresponding concept(s) in the ontology. *PhylAnt* offers a semi-automatic approach for such annotation of phylogenetic resources with the help of a suite of tools called *Kino-Phylo*. The complete suite of tools and instructions can be found at [33].

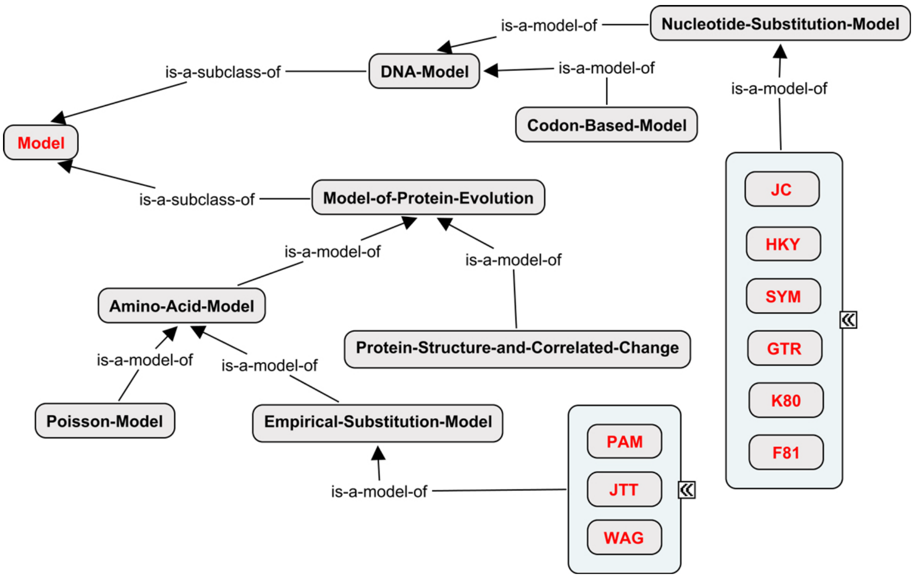


Figure 4 Models in phylogenetic studies. A hierarchy of concepts used to describep evolutionary models most commonly used in phylogenetic studies.

Annotating phylogenetic documents with Kino-Phylo

Kino for phylogenetics, also known as *Kino-Phylo* [13,17] is built on top of the Kino platform [12,33]. It is an integrated suite of tools that enables scientists to annotate phylogeny related documents in the *PhylAnt* platform. *Kino-Phylo* can annotate documents by accessing *PhylOnt* and other NCBO ontologies, via the NCBO Web API.

Kino-Phylo presents a comprehensive architecture for annotating and indexing phylogenetic oriented resources that should be of great use for the phylogenetic community. This system includes two main components, a browser-based annotation front-end, integrated with NCBO and an annotation-aware backend index to provides faceted search capabilities. It is designed around a basic workflow consisting of three steps, annotation, indexing, and searching[17,12]:

1. Annotation: In the annotation step, users provide annotations via a browser plug-in. After the annotations are added, the augmented document can be directly submitted to the indexing engine.
2. Indexing: Indexing is performed using Apache SOLR. It can be installed as an independent application and exposes multiple interfaces for client programs. SOLR provides the isolation for the index as well as support for faceting. Note that the SOLR interfaces are not directly exposed. They are wrapped by the *Kino-Phylo* submission API, described later in this paper. The annotation-aware back-end index is exposed via a RESTful API. It is designed such that the browser plug-in can directly submit the annotated web pages to the indexing engine.
3. Search: The search is performed via a Web interface. It presents the notions of a typical search engine and additionally gives the ability to filter the results via the facets. The current UI is built upon the JSON based Kino search API, which can be used to integrate other tools as well.

Browser plug-in for phylogenetic annotation

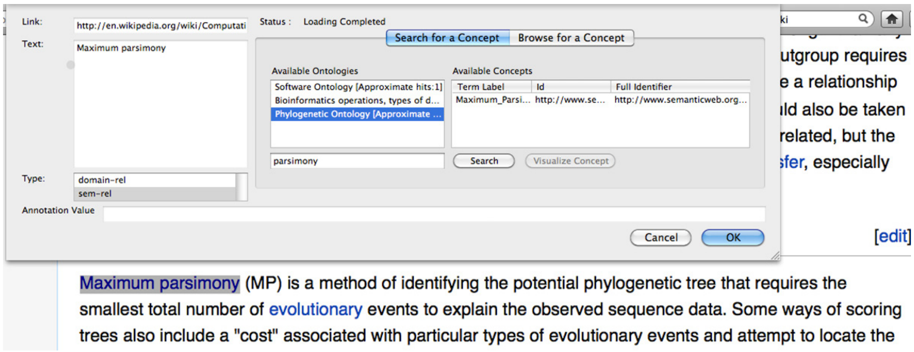
To use the browser plug-in, the user opens a topical web document in her browser and highlights words and phrases of interest. The plug-in provides hints on matching concepts fetched from NCBO. The user can also opt to browse for a concept in any ontology in NCBO (Figure 5). Once the annotations are added, the user can submit the annotations to a predefined *Kino-Phylo* instance (configured via the plug-in configuration page), by selecting the "publish annotations" menu item.

The plug-in modifies the HTML source of the document and embeds annotations using the SA-REST specification[12]. At submission, the augmented document tree in the browser is serialized and submitted as XML to the back end index via the document submission API (See next section).

Kino-Phylo index and search manager

The *Kino-Phylo* index manager is based on the Java JSP/Servlets technology and includes two major components, Document Submission API and Search API. The submission API acts as the receiver for the submitted documents. After receiving a document via the Document Submission API, the document will be filtered for embedded annotations and indexed. The index runs full-text indexing and special indexing for the filtered-out concepts. Additionally, the indexing process extracts extra information (such as synonyms) via NCBO and inserts this information in the index as well.

The *Kino-Phylo* search API includes a selection window that helps users to filter search results. For example, a user can search for parsimony as a concept or as a word. Once she finds a set of documents, they can be further filtered by co-locating concepts. For example, she can filter out the documents that have annotations on parsimony only



Link: http://en.wikipedia.org/wiki/Computati

Status : Loading Completed

Text: Maximum parsimony

Software Ontology [Approximate hits:1]
Bioinformatics operations, types of d...
Phylogenetic Ontology [Approximate ...

Available Concepts

Term Label	Id	Full Identifier
Maximum_Parsi...	http://www.se...	http://www.semanticweb.org...

parsimony

Type: domain-rel
sem-rel

Annotation Value

Maximum parsimony (MP) is a method of identifying the potential phylogenetic tree that requires the smallest total number of evolutionary events to explain the observed sequence data. Some ways of scoring trees also include a "cost" associated with particular types of evolutionary events and attempt to locate the

Figure 5 Annotation of phylogeny literature with Kino-Phylo Tools. The *Kino-Phylo* for the literature annotation plug-in shows how papers may be annotated. When the user highlights and right clicks in a term or phrase for annotation, the browser's context menu includes the annotation as a phylogenetical concept menu item. Selecting this menu item brings the annotations window where the highlighted term is searched using the NCBO RESTful API and a detailed view of the accessible ontological terms is shown to the user to choose the term from ontology and assign for annotation.

across the documents that contain parsimony as an annotation for the methods used in phylogeny study. The User Interface includes an intuitive facet selection tool that helps the user to filter the results.

Annotation of NeXML files with Kino-Phylo

Vos, et al. [11] proposed NeXML as an exchange standard for representing phylogenetic data, inspired by the commonly used NEXUS format [34], but more robust and easier to process. XML formats such as NeXML play an essential role in promoting the accessibility and reuse of data on the web. Using this technology can simplify and improve robustness in the processing of rich phylogenetic data and enable their reuse.

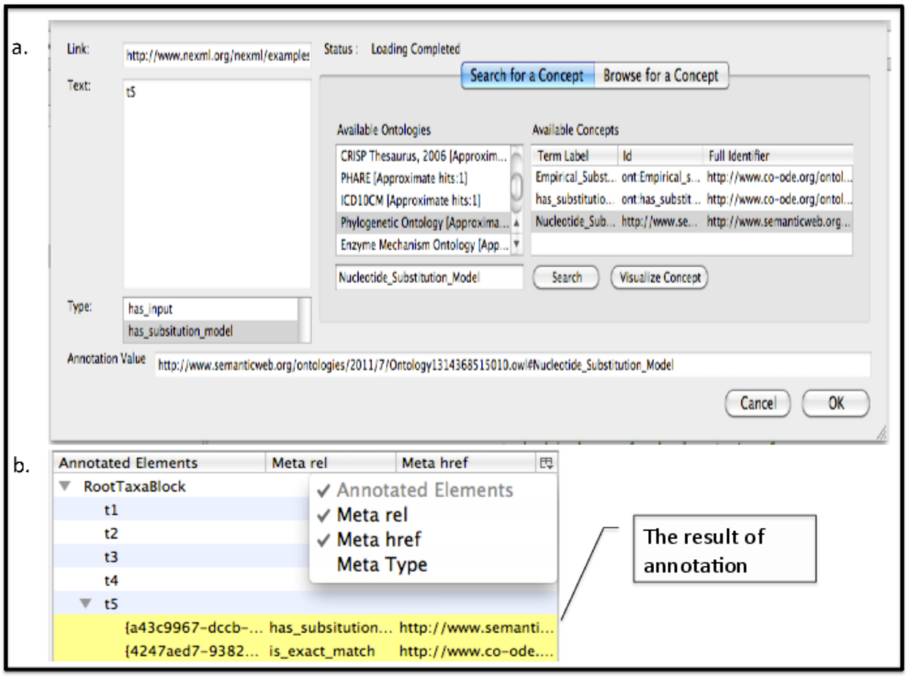
Annotations in NeXML are expressed using recursively nested “meta” elements that conform to RDFa syntax. The annotations thus form triples of subject, predicate, and object, where the subject is either a fundamental data resource from the NeXML document such as a tree, character state matrix, or taxon; or, transitively, the object of another triple. Instead of trying to provide vocabulary for all metadata types within the NeXML standard, users can thus use vocabularies or ontologies in common usage in the phyloinformatics community to annotate NeXML fles. To demonstrate this facility, we annotated NeXML documents using *Kino-Phylo*. With this approach, users can identify concepts from any NCBO ontology using exact or approximate searches to annotate selected element in a NeXML file (Figure 6). Users can then annotate a NeXML element to the desired triple, so that a statement can be made such as (subject NeXML element) “tree” (predicate) *has* – *substitution* – *model* (object) *nucleotide* – *substitution* – *model.*

Results and evaluations

PhylOnt is publicly available. As shown in Figure 2, this ontology includes descriptions of classes, definitions, properties, metadata and usage of classes with an example for each one and relations between them. With the help of NCBO researchers, *PhylOnt* has been deployed on BioPortal, a web-based portal designed and hosted by NCBO to enable accessibility to biological knowledge on the Semantic Web. In the comparison of PhlOnt at [17] In addition, we introduced and implemented *PhylAnt* platform for annotating, indexing and searching phylogenetic resources such as scientific articles and NeXML files.

Evaluation

Ontology evaluation is needed to guarantee that what has been built meets application requirements. There are different approaches for ontology evaluation, such as metric-based and application-based [35]. In the following sections we present results from both approaches.



Link: http://www.nexml.org/nexml/examples

Status : Loading Completed

Text: t5

Available Ontologies
- CRISP Thesaurus, 2006 [Approxim...
- PHARE [Approximate hits:1]
- ICD10CM [Approximate hits:1]
- Phylogenetic Ontology [Approxima...
- Enzyme Mechanism Ontology [App...

Available Concepts

Term Label	Id	Full Identifier
Empirical_Subst...	ont:Empirical_s...	http://www.co-ode.org/ontol...
has_substitutio...	ont:has_substit...	http://www.co-ode.org/ontol...
Nucleotide_Sub...	http://www.se...	http://www.semanticweb.org...

Nucleotide_Substitution_Model

Type: has_input
has_subsitution_model

Annotation Value http://www.semanticweb.org/ontologies/2011/7/Ontology1314368515010.owl#Nucleotide_Substitution_Model

Annotated Elements	Meta rel	Meta href
RootTaxaBlock		
t1		
t2		
t3		
t4		
t5		
{a43c9967-dccb-...	has_subsitution...	http://www.semanti...
{4247aed7-9382...	is_exact_match	http://www.co-ode....

✓ Annotated Elements
✓ Meta rel
✓ Meta href
Meta Type

Figure 6 Annotation of NeXML file with Kino-Phylo Tools. The *Kino-Phylo* user interface for NeXML annotation plug-in shows how users can identify concepts from any NCBO ontology including *PhylOnt* using exact or approximate searches to annotate selected element in a NeXML file with the triples from ontology.

Metric-based approach

Metric-based evaluations scan through an ontology to gather different types of statistical criteria about structural knowledge represented in an ontology. We used schema metrics [35], which evaluate ontology design and its potential for rich knowledge representation. Table 1 shows the results of these evaluations. “Relationship richness” reflects the diversity of the types of relations in the ontology, which is now higher in the comparison of our previous work [17] “Attribute richness” indicates both quality of ontology design and the amount of information pertaining to instance data. The results of the relationship richness assessment show that 74 percent of the connections between classes are “rich” relationships compared to all of the possible connections.

We also compared the *PhylOnt* Ontology with the Comparative Data Analysis Ontology(CDAO) [36,19] and the Embrace Data And Methods(EDAM) [20] ontology for these metrics. Table 2 shows the result of this comparison. *PhylOnt* includes deeper resolution of concepts related to phylogenetic inference methods, substitution models, tree estimation programs and provenance.

Annotation-based approach

A fundamental driving principle for the development of ontologies is their utility for data annotation and management. Therefore, as we developed *PhylOnt*, we evaluated it by annotating resources in phylogenetic documents using *Kino-Phylo* tools.

Collaborating domain experts selected exemplar publications that we used to investigate which concepts are missing in *PhylOnt* by trying to annotate the exemplars with it. The rationale is that we could determine the quality of *PhylOnt* by counting the relevant concepts encountered in a paper that are not present in *PhylOnt*, but are present in other relevant ontologies. This approach is used to compute Precision, Recall, and F-measure [37]. Suppose that $C_{\{P \cap O\}}$ is the set of concepts from the papers that have been annotated using *PhylOnt.* Then Precision and Recall can be calculated by the following equations:

$$Precision = \frac{|C_{\{P \cap O\}}|}{|C_P|} \quad (1)$$

$$Recall = \frac{|C_{\{P \cap O\}}|}{|C_O|} \quad (2)$$

Table 1 Metric-Based Approach for Ontology Evaluation

Metric name	Metric formula1	Metric value
Relationship Richness	$RR = \frac{\lvert P\rvert}{\lvert H\rvert + \lvert P\rvert}$	0.74
Attribute Richness	$AR = \frac{\lvert T\rvert}{\lvert C\rvert}$	0.30

[1] |*P*|: Number of non-inheritance relationships, |*C*|:Number of classes, |*T* |: Number of attributes

Table 2 Numerical Comparison of Ontologies EDAM, CDAO, PhylOnt

Parameters	EDAM	CDAO	PhylOnt
Number of classes	2746	143	147
Phylogeny analysis terms	26	129	138
Phylogeny methods	9	8	41
Substitution models	2	NA	31
Phylogeny programs	8	14	33
Provenance	NA	NA	21
Phylogeny data and Types	5	NA	12

C_P and C_O refer to the concepts of the paper and concepts in ontology respectively. The F-measure is the harmonic mean of precision and recall and it is calculated as:

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

For this experiment, we annotated selected papers by experts using *PhylOnt*, EDAM and CDAO. We increased the number of concepts in PhylOnt in vers8 as of Dec 2012 after getting feedback from community. After annotation the expert selected papers with ontologies, the precision is higher than our previous published results [17]. As it is shown in Table 3. the precision score of *PhylOnt* indicates that 85% of the phylogenetics concepts in the papers are covered by the ontology, while 57% of all concepts in the ontology are recalled in the selected papers. In combination, these scores were used to estimate an F-measure for *PhylOnt* that is higher than that of CDAO and EDAM (Figure 7).

Discussion

A big challenge in phylogenetic studies is the complexity of data being used in phylogenetic reconstruction and the diversity of analysis methods. Some of the barriers to reuse of this data are incomplete and non-tractable provenance data; insufficient method descriptions to reproduce the results; and the lack of semantic annotations of resources. Our focus in this study was on formally characterizing phylogenetic resources and identifying the relationships among key concepts. To the best of our knowledge and the feedback from the phylogenetics community [13], *PhylOnt* is the first ontology specifically created for phylogenetic analysis operations and related metadata.

Table 3 Annotation-based approach for ontology evaluation

Ontology	Precision	Recall	F-measure
PhylOnt	0.85	0.57	0.68
EDAM	0.17	0.013	0.024
CDAO	0.07	0.15	0.095

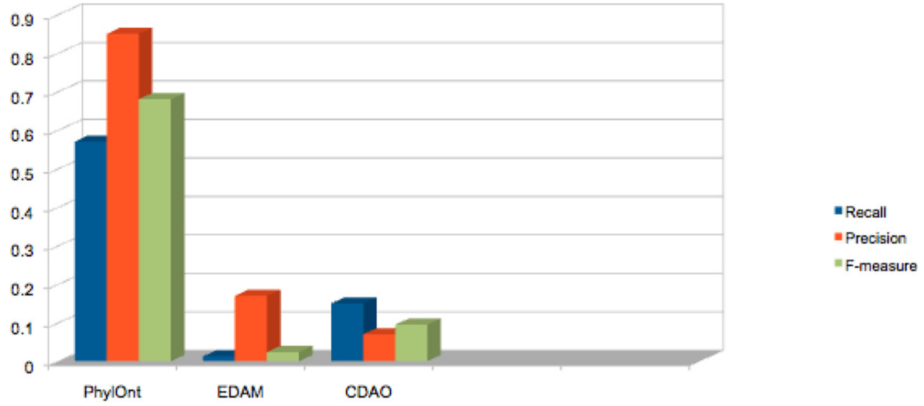


Figure 7 Annotation-based approach for ontology evaluation. Precision, recall and f-mature calculated for PhyLOnt in the comparison of EDAM and CDAO Ontology. We increased the number of concepts in PhylOnt in vers8 of ontology [21]. After annotation the expert selected papers with ontology, the precision is higher than our previous published results [17].

As of March 2013, the 8th version of *PhylOnt* has been submitted to NCBO. Our results show that *PhylOnt* is a rich ontology for the concepts in phylogeny applications compared to putative alternatives such as EDAM and CDAO [17]. Note, however, that the EDAM ontology is much more broadly scoped to the entire bioinformatics domain, whereas CDAO is scoped to defining the relationships among fundamental data concepts (e.g. nodes, trees, character state matrices), not methods of phylogenetic analysis or provenance metadata. As real-world use cases of richly annotated phylogenetic data develop it is likely that these three artefacts will therefore be complementary rather than in competition.

We introduced the *PhylAnt* platform, which enables semantic annotation of phylogenetic resources. Annotating phylogenetic documents using ontologies is the foundation for the use of other semantic technologies in this domain and it is a preliminary step to semantic search, information retrieval, and heterogeneous data integration that can support phylogenetic workflows. These annotations have a variety of uses, ranging from extended search capabilities to advanced data mining. Annotated documents are indexed using a faceted indexing and search engine that provides fine-grained search capabilities.

PhylOnt does not currently cover all concepts included in phylogenetic analyses, but rather forms a foundation for an extensible ontology that will grow as researchers develop and apply new analysis methods. Further, the ontology does not currently include all method or model specific parameter definitions. Again, these can be added to the ontology as needs are defined by the phyloinformatics community.

Conclusion

The research presented in this manuscript is aimed at applying semantic web technologies to phyloinformatics. We addressed these objectives from both a phylogenetics and a computer science perspective. From the phylogenetics community perspectives, reusability and the ability to search for phylogenetic information are improved with the help of semantic web technology. From a computer science perspective, semi-automatic annotation of different resources with the concepts defined in *PhylOnt*, indexing and searching through resources will facilitate interoperability among phylogenetic resources. These advances allow researchers to access, explore and reuse the resources and products of phylogenetic studies.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MP conceived of the study, ontology development, annotation and drafted the manuscript. AS participated in semantic approaches for data reuse. AR participated in annotation and searching framework for both literature and NeXML documents. RV participated in NeXML annotation process. JL participated in study design in phyloinformatics, ontology development, annotation, validation and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This article is based on "PhylOnt : A domain-specific ontology for phylogeny analysis", by Maryam Panahiazar, Ajith Ranabahu, Vahid Taslimi, Hima Yalamanchili, Arlin Stoltzfus, Jim Leebens-Mack, and Amit Sheth, which appeared in Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on.
© 2012 IEEE, http://dx.doi.org/10.1109/BIBM.2012.6392677.

Declarations

The publication costs for this article were funded by Jim Leebens-Mack.
This article has been published as part of *BMC Medical Genomics* Volume 6 Supplement 3, 2013: Selected articles from the IEEE International Conference on Bioinformatics and Biomedicine 2012: Medical Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcmedgenomics/supplements/6/S3.

Authors' details

[1]Ohio Center for Excellence in Knowledge-enabled Computing (kno.e.sis) College of Computer Science and Engineering, Wright State University,

Dayton, OH, USA. [2]Department of Plant Biology, University of Georgia, Athens, GA, USA. [3]Naturalis Biodiversity Center, Leiden, the Netherlands. [4]Bioinformatics Institute, University of Georgia, Athens, GA, USA.

Published: 11 November 2013

References

1. Dobzhansky T: **Nothing in Biology Make Sences Except in the Light of Evolution.** *The Americsn Biology Teacher* 1973, **35**:125-129.
2. Darwin C: **On the Origin of Species by means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.** London. (J.Murray); 1859.
3. Haeckel E: **Generelle Morphologie der Organismen.** *Reimer, Berlin* 1866.
4. Leebens-Mack J, Vision T, Brenner E, Bowers JE, Doyle JJ, Eisen JA, Gu XUN, Harshman J: **Taking the First Step towards a Standard for Reporting on Phylogenetics: Minimal Information about a Phylogenetic Analysis (MIAPA).** *OMICS* 2006, **10(2)**:231-237.
5. Barker D, Pagel M: **Predicting functional gene links from phylogenetic-statistical analyses of whole genomes.** *PLoS computational biology* 2005, **1**:e3.
6. Gaudet P, Livstone MS, Lewis SE, Thomas PD: **Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium.** *Briefings in bioinformatics* 2011, **12(5)**:449-62.
7. Eisen A: **Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis.** *Genome Research* 1998, **8**:163-167.
8. Holmes E, Nee S, Rambaut A, Garnett G, Harvey P: **Revealing the history of infectious disease epidemics through phylogenetic trees.** *Philos Trans R Soc Lond B Biol Sci* 1995, **349(1327)**:33-40.
9. Piel WH, Sanderson MJ, Donoghue MJ: **The small-world dynamics of tree networks and data mining in phyloinformatics.** *Bioinformatics* 2003, **19(9)**:1162-1168.
10. Stoltzfus A, O'Meara B, Whitacre J, Mounce R, Gillespie EL, Kumar S, Rosauer DF, Vos RA: **Sharing and re-use of phylogenetic trees (and associated data) to facilitate synthesis.** *BMC Res Notes* 2012, **5**:574.
11. Vos RA, Balhoff JP, Caravas JA, Holder MT, Lapp H, Madison WP, Midford PE, Priyam A, Sukumaran J, Xia X, Stoltzfus A: **NeXML: Rich, Extensible, and Verifiable Representation of Comparative Data and Metadata.** *Systematic biology* 2012, **61(4)**:675-89.
12. Ranabahu A, Parikh Pp, Panahiazar M, Sheth AP, Logan-Klumpler F: **Kino: A Generic Document Management System for Biologists Using SA-REST and Faceted SearcRanabahu, A., Parikh, P., Panahiazar, M., Sheth, A., & Logan-Klumpler, F. (2011). Kino: A Generic Document Management System for Biologists Using SA-REST and Faceted.** *2011 IEEE Fifth International Conference on Semantic Computing* 2011, 205-208.
13. Panahiazar M, Vos RA, Enrico P, Todd V, Leebens-Mack J: **Building a Foundation to Enable Semantic Technologies for phylogenetically based Comparitive Analysis.** *2011 Informatics for Phylogenetics, Evolution, and Biodiversity (iEvoBio 2011)*, **2011.**
14. Ranabahu A, Sheth AP, Panahiazar M, Wijeratne S: **Semantic Annotation and Search for resources in the next Generation Web with SA-REST SA-REST for Service Annotation.** *W3C Workshop on Data and Services Integration* Bedford, MA; 2011.
15. Panahiazar M, Leebens-Mack J, Ranabahu A, Sheth AP: **Using semantic technology for Phylogeny.** *AMIA.Annual Symposium proceedings, TBI, iEvoBio* 2012, 175.
16. Panahiazar M, Sheth AP, Ranabahu A, Leebens-Mack J: **Semantic Technology and Translational Genomic Research.** *AMIA Annual Symposium proceedings, TBI* , **2012:** 176.
17. Panahiazar M, Ranabahu A, Taslimi V, Yalamanchili H, Stoltzfus A, Leebens-Mack J, Sheth AP: **PhylOnt: A domain-specific ontology for phylogeny analysis.** *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on: 4-7 October 2012* , **2012:** 1-6.
18. **PhyloTastic.** [http://phylotastic.org].
19. Chisham B, Wright B, Le T, Son TC, Pontelli E: **CDAO-Store: Ontology-Driven Data Integration for Phylogenetic Analysis.** *BMC bioinformatics* 2011, **12**:98.
20. Lamprecht AL, Naujokat S, Steffen B, Margaria T: **Constraint-Guided Workflow Composition Based on the EDAM Ontology.** *Nature Proceedings* 2010.
21. Panahiazar M: *PhylOnt: An Ontology for Phylogeny analyses* , **2011 [http://bioportal.bioontology.org/ontologies/1616].**
22. MIAPA/PhyloWays: *A list of interpreted phyloinformatics workflows* , **2011 [http://www.evoio.org/wiki/MIAPA/PhyloWays].**
23. Harrison CJ, Langdale Ja: **A step by step guide to phylogeny reconstruction.** *The Plant journal : for cell and molecular biology* 2006, **45(4)**:561-72.
24. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM: **Phylogenetic inference.** *Molecular systematics* 1996, 407-514.
25. Posada D, Buckley TR: **Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests.** *Systematic biology* 2004, **53(5)**:793-808.
26. Dayhoff M, Schwartz R, Orcutt B: **A Model of Evolutionary Change in Proteins.** *Atlas of protein sequence and structure* 1978.
27. Jones D, Taylor W, Thornton J: **The rapid generation of mutation data matrices from protein sequences.** *Computer Applications in the Biosciences* 1992, **8**:275-282.
28. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution.** *Bioinformatics* 2005, **21(9)**:2104-5.
29. Yang Z: **Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods.** *Journal of Molecular Evolution* 1994, **39(3)**:306-314.
30. Waddell P, Steel M: **General time-reversible distances with unequal rates across sites: mixing and inverse Gaussian distributions with invariant sites.** *Mol Phylogenet Evol* 1997, **8**:398-414.
31. Ronquist F, Teslenko M, Van der Mark P, Ayres D, Darling A, Höhna S, Larget B, Liu L, Suchard M, Huelsenbeck J: **MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space.** *Syst Biol* 2012, **61(3)**:539-42.
32. Wilgenbusch JC, Swofford DL: **Inferring evolutionary trees with PAUP*.** *Curr Protoc Bioinformatics* 2003, **Chapter 6**:Unit 6.4.
33. Ranabahu A, Panahiazar M: *kino platform* [http://wiki.knoesis.org/index.php/Kino].
34. Maddison DR, Swofford DL, Maddison WP: **NEXUS: an extensible file format for systematic information.** *Systematic Biology* 1997, **46(4)**:590-621.
35. Vrandečić D, York S: **How to Design Better Ontology Metrics.** *ESWC '07 Proceedings of the 4th European conference on The Semantic Web: Research and Applications* 2007, 311-325.
36. Prosdocimi F, Chisham B, Pontelli E, Thompson J, Stoltzfus A: **Initial implementation of a comparative data analysis ontology.** *Evol Bioinform Online* 2009, **5**:47-66.
37. Cross V, Parikh PP, Panahiazar M: **Aligning the Parasite Experiment Ontology and the Ontology for Biomedical Investigations Using AgreementMaker.** *ICBO: Internationla Conference on Biomedical Ontology* 2011, 2-8.

doi:10.1186/1755-8794-6-S3-S5
Cite this article as: Panahiazar *et al.*: Advancing data reuse in phyloinformatics using an ontology-driven Semantic Web approach. *BMC Medical Genomics* 2013 **6**(Suppl 3):S5.