BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 5. 485–488
Advance Access published on 11 June 2011

doi:10.1093/bib/bbr025

Letter to the Editor: SeqXML and OrthoXML: standards for sequence and orthology information

Thomas Schmitt, David N. Messina, Fabian Schreiber and Erik L.L. Sonnhammer

Submitted: 28th February 2011; Received (in revised form): 30th March 2011

Abstract

There is a great need for standards in the orthology field. Users must contend with different ortholog data representations from each provider, and the providers themselves must independently gather and parse the input sequence data. These burdensome and redundant procedures make data comparison and integration difficult. We have designed two XML-based formats, SeqXML and OrthoXML, to solve these problems. SeqXML is a lightweight format for sequence records—the input for orthology prediction. It stores the same sequence and metadata as typical FASTA format records, but overcomes common problems such as unstructured metadata in the header and erroneous sequence content. XML provides validation to prevent data integrity problems that are frequent in FASTA files. The range of applications for SeqXML is broad and not limited to ortholog prediction. We provide read/write functions for BioJava, BioPerl, and Biopython. OrthoXML was designed to represent ortholog assignments from any source in a consistent and structured way, yet cater to specific needs such as scoring schemes or meta-information. A unified format is particularly valuable for ortholog consumers that want to integrate data from numerous resources, e.g. for gene annotation projects. Reference proteomes for 61 organisms are already available in SeqXML, and 10 orthology databases have signed on to OrthoXML. Adoption by the entire field would substantially facilitate exchange and quality control of sequence and orthology information.

Keywords: *OrthoXML; SeqXML; 'sequence format'; 'orthology format'; FASTA format, XML*

INTRODUCTION

Orthologs are defined as genes in different species directly descending from a single gene in the last common ancestor [1]. In recent years the development of new orthology prediction methods has flourished, increasing the need for comparisons between methods [2]. In order to facilitate this task, we here present two standards: SeqXML for the input and OrthoXML for the output of orthology prediction. The standards were first mentioned in [3], but have since then been greatly improved. In addition to being useful for comparative benchmarks, the standards will make data exchange between orthology producers and consumers easier and will simplify the development of new tools and resources.

Because of its compactness and simplicity, FASTA is one of the most widely used formats to store sequence information, and it is the typical input format for orthology prediction. However, there is no standardized way to store metadata such as descriptions, database cross-references, or species of origin in FASTA format. As the need has grown for storing

Corresponding author. Erik L.L. Sonnhammer, Science for Life Laboratory, Box 1031, SE-17121 Solna, Sweden. Tel: +46 (0)8 52481184; Fax: +46 (0)8 52481425; E-mail: erik.sonnhammer@sbc.su.se

Thomas Schmitt is a PhD student at the Stockholm Bioinformatics Centre. He is working on protein interaction networks and methods for transferring functional coupling between orthologs.

David N. Messina works on finding new gene families in metagenomics samples and improving biological data exchange for his PhD at the Stockholm Bioinformatics Centre.

Fabian Schreiber is a postdoc at the Stockholm Bioinformatics Centre. His research focuses on orthology detection, taxonomic profiling in metagenomics and phylogenomics of early animals.

Erik L.L. Sonnhammer is Professor of Bioinformatics at Stockholm University and director of the Stockholm Bioinformatics Centre. He heads a bioinformatics research lab focusing on protein function prediction, protein domain evolution, orthology analysis and protein networks.

metadata along with sequence data, many have put metadata in the FASTA header line in idiosyncratic ways, which makes parsing laborious and often unreliable. In contrast, SeqXML stores metadata in a structured and standardized way, and the sequence itself as well as the general data structure of a SeqXML entry can be validated to ensure data integrity. We present SeqXML together with OrthoXML in the context of orthology prediction and benchmarking because it was developed for this purpose, but the range of applications for SeqXML is as broad as for the FASTA format.

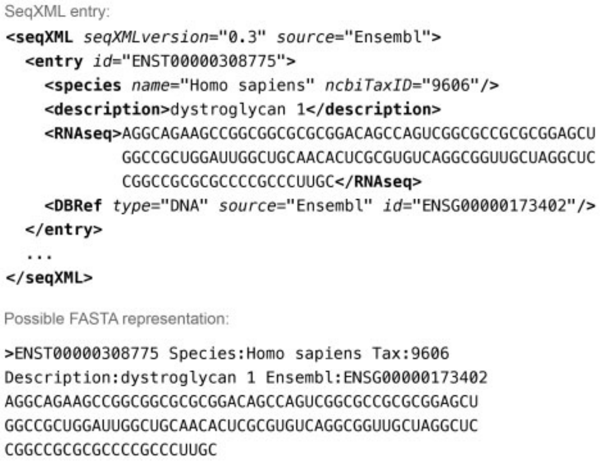
For orthology information the situation is even worse. There is no standard, and every database provides assignments in a different format. This makes an automated integration impossible and enforces tedious and error-prone manual procedures.

One can distinguish between two different types of orthology prediction methods: graph- and tree based. The output of graph-based methods is in most cases ortholog groups, i.e. groups of genes that stem from the same gene in their last common ancestor. Ortholog groups are typically encoded in self-defined delimiter-separated formats. These formats are database specific and often lack important information such as input sequence sources, ortholog group identifiers or scores for the assignments. OrthoXML resolves these problems while retaining enough flexibility to fulfill the needs of different databases.

Tree-based methods predict orthology relationships by reconciling gene trees with a species tree. The resulting gene trees are typically stored in standard tree formats like Newick or Nexus, which also lack orthology-related information. PhyloXML [4] supports orthology information in trees, but is not applicable to nonhierarchical ortholog groups. Conversely, OrthoXML was developed to store both graph- and tree-based orthology assignments.

SeqXML

Figure 1 shows an example SeqXML record. The information that can be encoded corresponds to what can typically be found in a FASTA format entry, including the sequence, an identifier, the species of origin and, cross-references to other databases. Like FASTA, only the first two are strictly required. Unlike FASTA, however, all datatypes are kept in a clearly defined and structured way so that there is no need to construct parsing code for each FASTA format source. Additionally, custom annotations can be stored in the form of key-value pairs. One further advantage of SeqXML is that the schema includes definitions for RNA, DNA and amino acid alphabets; hence, all of the sequences in a SeqXML file can be automatically validated by standard tools such as XMLlint [5]. Detailed documentation with examples can be found at http://SeqXML.org.



SeqXML entry:

```
<seqXML seqXMLversion="0.3" source="Ensembl">
  <entry id="ENST00000308775">
    <species name="Homo sapiens" ncbiTaxID="9606"/>
    <description>dystroglycan 1</description>
    <RNAseq>AGGCAGAAGCCGGCGGCGCGGACAGCCAGUCGGCGCCGCGCGGAGCU
            GGCCGCUGGAUUGGCUGCAACACUCGCGUGUCAGGCGGUUGCUAGGCUC
            CGGCCGCGCGCCCCGCCCUUGC</RNAseq>
    <DBRef type="DNA" source="Ensembl" id="ENSG00000173402"/>
  </entry>
  ...
</seqXML>
```

Possible FASTA representation:

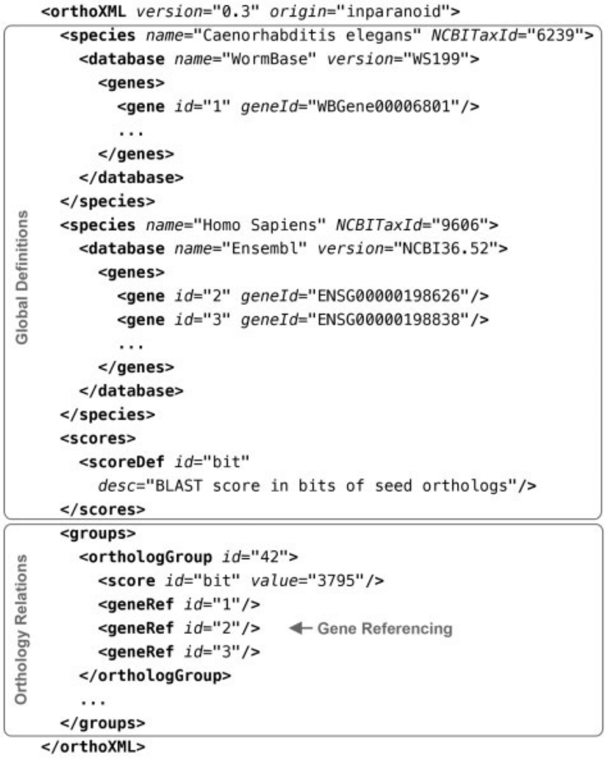
```
>ENST00000308775 Species:Homo sapiens Tax:9606
Description:dystroglycan 1 Ensembl:ENSG00000173402
AGGCAGAAGCCGGCGGCGCGGACAGCCAGUCGGCGCCGCGCGGAGCU
GGCCGCUGGAUUGGCUGCAACACUCGCGUGUCAGGCGGUUGCUAGGCUC
CGGCCGCGCGCCCCGCCCUUGC
```

Figure 1: A SeqXML entry can store the same information as a typical FASTA format record, but in a structured and standardized way.

OrthoXML

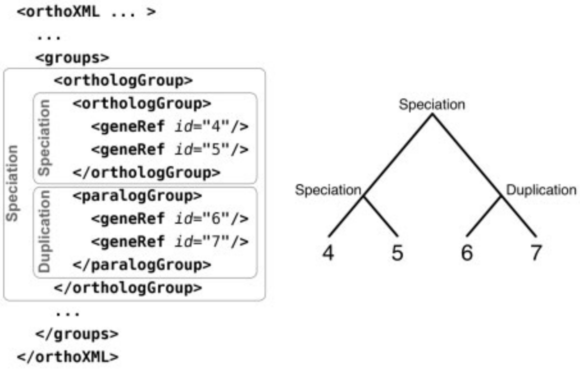
Figure 2 shows an example OrthoXML record. At the beginning of an OrthoXML file, the genes are defined globally by the species of origin, the database they come from, and identifiers for the gene and gene products in the database. This way OrthoXML files can be fully interpreted without the need for extra information. After the gene definitions follows a list of groups that describe the orthology relations between the genes. In these groups, genes are simply references to the global definition. This two-part structure avoids redundancy.

To represent trees, there are two types of groups, and these can be nested as shown in Figure 3. 'Ortholog groups' represent speciation nodes of the gene tree, while duplication nodes are represented by 'paralog groups'. These nodes can be defined at various levels of detail. It is for instance possible to omit all duplication nodes that have no preceding speciation nodes. A tree reduces to a single ortholog group if all internal nodes are omitted, i.e. if no information is given about how the individual genes have evolved.



```
<orthoXML version="0.3" origin="inparanoid">
  <species name="Caenorhabditis elegans" NCBITaxId="6239">
    <database name="WormBase" version="WS199">
      <genes>
        <gene id="1" geneId="WBGene00006801"/>
        ...
      </genes>
    </database>
  </species>
  <species name="Homo Sapiens" NCBITaxId="9606">
    <database name="Ensembl" version="NCBI36.52">
      <genes>
        <gene id="2" geneId="ENSG00000198626"/>
        <gene id="3" geneId="ENSG00000198838"/>
        ...
      </genes>
    </database>
  </species>
  <scores>
    <scoreDef id="bit"
      desc="BLAST score in bits of seed orthologs"/>
  </scores>
  <groups>
    <orthologGroup id="42">
      <score id="bit" value="3795"/>
      <geneRef id="1"/>
      <geneRef id="2"/>
      <geneRef id="3"/>
    </orthologGroup>
    ...
  </groups>
</orthoXML>
```

Figure 2: In the first section of an OrthoXML file, the genes are defined globally. The second section describes the orthology relations between the referenced genes. The example above shows a *Caenorhabditis elegans* gene with two orthologs in human. OrthoXML supports the definition of scores for groups or individual group members. In this example, a bit score (of the seed orthologs' match) of 3742 is assigned to an ortholog group.



```
<orthoXML ... >
  ...
  <groups>
    <orthologGroup>
      <orthologGroup>
        <geneRef id="4"/>
        <geneRef id="5"/>
      </orthologGroup>
      <paralogGroup>
        <geneRef id="6"/>
        <geneRef id="7"/>
      </paralogGroup>
    </orthologGroup>
    ...
  </groups>
</orthoXML>
```

Figure 3: OrthoXML has two types of groups which can be nested to represent gene trees. 'Ortholog groups' represent speciation nodes and 'paralog groups' duplication nodes.

Some orthology databases provide scores for their predictions. To incorporate these, OrthoXML allows for the global definition of scoring schemes, which can later be used at different levels of the assignments. Additional annotations for groups can be stored as user-defined key-value pairs. The XML schema for OrthoXML and detailed documentation with examples can be found at http://OrthoXML.org.

COMMUNITY UPTAKE AND TOOLS

A number of the major ortholog databases [3, 6–14] provide their assignments in OrthoXML or are in the process of adopting the standard. The Reference Proteome Project [15] aims to create a database with a single representative protein per gene for all major sequenced eukaryotes. This standardization is especially useful for comparing orthology algorithms, and to this end we have made the reference proteomes available in SeqXML format at http://SeqXML.org.

The official distribution of the BioPerl [16] library includes parsing and writing functionality for SeqXML. APIs for BioJava [17] and Biopython [18] are available at http://SeqXML.org. A Java library for parsing and writing OrthoXML is provided at http://OrthoXML.org; Python and Perl implementations are in development and will be provided at the same site.

We are continuously working on improving the standards and are happy for any feedback.

Key Points

- OrthoXML is the first standard for representing orthology information.
- OrthoXML supports groups and trees in a consistent way.
- SeqXML is a robust alternative to the error-prone FASTA format.
- SeqXML input–output tools have been implemented for the widely used biological software libraries BioPerl, BioJava and Biopython.
- The standards have been adopted by the reference genome project and several leading orthology resources.

Acknowledgements

We thank S. Roopra for help with the XML schemas and library implementations.

FUNDING

This work was supported by the Department of Biochemistry and Biophysics at Stockholm

University and grants from The Wenner-Gren Foundations and the Swedish Research Council.

References

1. Fitch WM. Distinguishing homologous from analogous proteins. *System Zool* 1970;**19**:99.
2. Gabaldón T, Dessimoz C, Huxley-Jones J, *et al*. Joining forces in the quest for orthologs. *Genome Biol* 2009;**10**:403.
3. Östlund G, Schmitt T, Forslund K, *et al*. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* 2010;**38**:D196–203.
4. Han MV, Zmasek CM. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* 2009;**10**:356.
5. The XML C parser and toolkit of Gnome. http://xmlsoft.org (26 February 2011, date last accessed).
6. Uchiyama I, Higuchi T, Kawai M. MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucleic Acids Res* 2010;**38**:D361–5.
7. Kuzniar A, Lin K, He Y, *et al*. ProGMap: an integrated annotation resource for protein orthology. *Nucleic Acids Research* 2009;**37**:W428–34.
8. Blake JA, Bult CJ, Kadin JA, *et al*. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic acids research* 2011;**39**:D842–8.
9. Linard B, Thompson JD, Poch O, *et al*. OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC bioinformatics* 2011;**12**:11.
10. Eyre TA, Wright MW, Lush MJ, *et al*. HCOP: a searchable database of human orthology predictions. *Briefings in bioinformatics* 2007;**8**:2–5.
11. Vilella AJ, Severin J, Ureta-Vidal A, *et al*. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research* 2009;**19**:327–35.
12. Chen F, Mackey AJ, Stoeckert CJ, *et al*. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic acids research* 2006;**34**:D363–8.
13. Datta RS, Meacham C, Samad B, *et al*. Berkeley PHOG: PhyloFacts orthology group prediction web server. *Nucleic acids research* 2009;**37**:W84–9.
14. Altenhoff AM, Schneider A, Gonnet GH, *et al*. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic acids research* 2010;**39**:D289–294.
15. Reference proteomes - Primary proteome sets for the Quest For Orthologs. http://www.ebi.ac.uk/reference_proteomes/ (26 February 2011, date last accessed).
16. Stajich JE, Block D, Boulez K, *et al*. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* 2002;**12**:1611–18.
17. Holland RCG, Down TA, Pocock M, *et al*. BioJava: an open-source framework for bioinformatics. *Bioinformatics* 2008;**24**:2096–7.
18. Cock PJA, Antao T, Chang JT, *et al*. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;**25**:1422–3.