phyloXML

www.phyloxml.org

Christian Zmasek
Ethalinda Cannon

Overview

- Motivation
- Current string representations of phylogenies
- NHX (the father of phyloXML)
- XML for phylogenies?
- Basics of phyloXML
- PhyloXML example
- phylogeny, clade, and branch elements
- Noteworthy (?)
- Current status

Motivation I

- Phylogenetic trees used in phylogenomic studies are annotated with various fields, such as:
  - Nodes:
    - (sequence) name
    - taxonomy
    - gene duplication vs speciation
  - Branches:
    - branch length
    - support values (such as bootstrap values)

Motivation II

- Many other node data fields have been requested:
  - GO annotation
  - EC numbers
  - domain structure
  - sequence
  - ‘function’
  - URLs
  - …

Some current formats to describe phylogenies

- New Hampshire
- NEXUS
- New Hampshire X (NHX):
  - New Hampshire eXtended
  - Developed for phylogenomics
  - Input and output for tree view tool ATV, SDI implementation

Example of NHX (or …lots of tiny Anacondas)

gene name

branch length

```
ADH-3:0.12[&&NHX:S=Anaconda
:E=1.1.1.1:C=0.0.255:W=5
:XN=L=population=18434
:XN=D=weight=45.6=g]
```

species

branch width

EC number

color (red.green.blue)

Custom property 1: Population size

Custom property 2: Animal weight in grams

Existing XMLs for trees

- "Taxonomic Markup Language" by R. Gilmour (Bioinformatics Vol. 16 no. 4 2000): for taxonomic data.
- http://evolution.genetics.washington.edu/phylip/doc/retree.html mentions a "PhyloXML ".

Basics

- phyloXML is similar to the one mentioned by Joe Felsenstein in his book "Inferring Phylogenies ".
- Example:

```
<phylogeny>
  <clade>
    <clade>
      <clade><name>A</name></clade>
      <clade><name>B</name></clade>
      <name>ab</name>
    </clade>
    …
```

phyloXML example

```
<phyloxml>
  <phylogeny>
    <name>example</name>
    <description>this is in phyloXML</description>
    <clade>
      <clade branch_length="0.06">
        <clade branch_length="0.102">
          <name>A</name>
        </clade>
        <clade branch_length="0.23">
          <name>B</name>
        </clade>
      </clade>
      <clade branch_length="0.4">
        <name>C</name>
      </clade>
    </clade>
  </phylogeny>
</phyloxml>
```

phylogeny elements:

- name
- description
- support
- URI
- custom
- ##other
- clade

`clade` elements:

- `name`
- `taxonomy` (common name, scientific name, rank, distribution, …)
- `sequence` (name, GO term, EC number, location, support)
- `event` (name, support)
- `custom`
- `support`
- `URI`
- `branch`
- `##other`
- `clade`

`branch` elements:

- `length`
- `support`
- `custom (name, value, support)`

please note...

- most elements have `id_ref` and `id` attributes, thus:
  - description of networks is possible (`clade` can have a parent determined by hierarchical structure **and/or** parents set via `clade_ref` of its `branch`(es)
  - trees can be either described in a flat or hierarchical manner
- no agreement among users where to place branch length value, hence phyloXML allows:
  - a `branch_lenght` attribute of `clade` (easy)
  - or branch lengths can be set via `branch_length` element of `branch` (flexible)

Extensibility

- `custom` element (of phylogeny, clade, branch, sequence, taxonomy)
- `##other`

Current status

- Current version: 0.06 (level 1)
- phyloXML is enforced by a schema
- ATV (based on FORESTER package) is being updated to support phyloXML input, output and conversion
- Joe Felsenstein promised phyloXML output for PHYLIP

Acknowledgments

Joe Felsenstein , Arlin Stoltzfus, Todd Vision, Hilmar Lapp, Jason Stajich, Jim Leebens-Mack, Steven Cannon, Rod Page, Mark Miller, Marcus Daniels, Korbinian Strimmer, David Hillis, Wayne Maddison, Daniel Huson, Jim Pitman, Tim Hughes, Julian Humphries, Jonathan Eisen, Kimmen Sjölander, …
Sean Eddy, Adam Gozik, Burnham Institute for Medical Research