BioXSD

An XML exchange format for basic bioinformatics data

Matúš Kalaš
Jan Christian Bryne*
Armin Töpfer**
Pål Puntervoll (PI)
Inge Jonassen (PI)
CBU, Uni Bergen

Edita Karosiene
Kristoffer Rapacki (PI)
CBS, DTU, Greater Copenhagen

Jon Ison
EBI, EMBL, Hinxton

Alexandre Joseph
Christophe Blanchet (PI)
IBCP, CNRS, Lyon

Steve Pettifer
University of Manchester

* now Oslo University Hospital
** now ETH Zürich in Basel
call Web service 1

output of service 1

parse output

intermediate data

print into input format

input for service 2

call Web service 2

call Web service 1

data in standard format

call Web service 2
Goals of **BioXSD:**

- Being an XSD-based XML format
to complement plain-text/TSV and RDF formats

- Filling the gap between specialised XSD-based exchange formats
  (such as SBML, MAGE-ML, PDBML, phyloXML, PSI-MI MIF, GCDML, GLYDE-II, ...)

- Compatible with data binding libraries for all main programming languages

- As lightweight as possible, but fitting everyone

- Developed and maintained in an open but organised collaboration
  welcoming requests from the community

- Detailed structure
  in-depth validation, semantic annotation (EDAM & more),
efficient compression (EXI)
BioXSD is an exchange format for basic bioinformatics data. It supports:

- Biomolecular sequences
- Sequence alignments
- Sequence and genome features (annotation)
- References to data, identifiers/accessions, ...
Example data in BioXSD format:

A sequence record:

```xml
<exampleSequenceRecord xsi:type="bx:NucleotideSequenceRecord">
  <bx:sequence>gtgcgagaggcccgtgccgctgcctacgaggctttctgccgctggagggaggtc</bx:sequence>
  <bx:species>
    dbName="NCBI Taxonomy"
    accession="9598"
    entryUri="http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=9598"
    speciesName="Chimp"
  </bx:species>
  <bx:reference>
    dbName="GenBank/Nucleotide"
    accession="NM_001008991"
    sequenceVersion="1"
  </bx:reference>
  <bx:subsequencePosition>
    <bx:segment min="282" max="345"/>
  </bx:subsequencePosition>
</exampleSequenceRecord>
```

Snippet of aldehyde dehydrogenase 5 family, member A1 (ALDH5A1)

Nuclear gene encoding mitochondrial protein, mRNA (GI:57113868)
biomolecular sequences

sequence alignments

sequence and genome features (annotation)

references to data, identifiers/accessions, ...
find similar sequences

BLAST Swiss-Prot

list of similar sequences

align sequences

ClustalW

multiple seq. alignment

optimise alignment

MaxAlign

**BioXSD**: BiosequenceRecord

**BioXSD**: BiosequenceAlignment
BioXSD 1.1 types:

**SimpleTypes:**

NucleotideSequence
AminoacidSequence
GeneralNucleotideSequence
GeneralAminoacidSequence
Biosequence

Accession(s)

**helper types:**

Name, Text
Uri
Integer(s), Decimal(s)
... and a few more

**ComplexTypes:**

NucleotideSequenceRecord
AminoacidSequenceRecord
GeneralNucleotideSequenceRecord
GeneralAminoacidSequenceRecord
BiosequenceRecord

..SequenceAlignment(s)

FeatureRecord

DatabaseReference, EntryReference
OntologyReference, SemanticConcept
Species, SequenceReference, Method

**helper types:**

Score, SequencePosition(s)
... a few more
BioXSD can be used:

- Directly as an input/output format of tools

- BioXSD can be extended, restricted, or included within other formats

- BioXSD can serve as the intermediate canonical format