Using DAML format for representation and integration of complex gene networks: implications in novel drug discovery

K. Baclawski Northeastern University

E. Neumann Beyond Genomics

T. Niu Harvard School of Public Health

Abstract

Great strides have been made in understanding the complex molecular networks underlying biological systems. Advances in high-throughput microchip-based assays are providing us with global gene activity profiles characterizing the output of the gene regulatory network. The eXtensible Markup Language (XML) is becoming a key component in data exchange in biology and currently there are over a dozen XML DTDs and schemas for bioinformatic applications, including BioML, CellML, RNAML, DDBJ-ML and BSML. However, there are intrinsic limitations of using XML in representing biopathways because of its hierarchical nature. To address this problem, we introduced a new way of knowledge representation using DARPA Agent Markup Language (DAML) schemas through ontology-based models for integration of complex genetic networks. This new approach has overcome the restriction of the use of hierarchical data structure, and can facilitate automatic retrieval and update of gene network-related data and can be transformed into other XML dialects. Our prototype is expected to shed novel insights on genetic networks that will help our efforts towards better understanding of complex disease etiology, which in turn, can have significant expectations in clinical pharmacogenomics.

Introduction
The Biological Knowledge Onslaught

- Advances in gene sequences and microchip-based assays are producing large amounts of complex information about biological systems.
- The many incompatible formats make it difficult for applications to interoperate.
- Complex linkages between information from multiple sources/formats is essential for understanding complex disease etiology.

XML for Data Interchange

- Commonly recognized format for data interchange supports interoperation.
- Many tools are available including XSLT, a powerful transformation language for XML.
- XML supports deep, semantically rich data.
- Numerical data, text and sequences can be specified in the same document.
- Document Type Definition (DTD) specifies the meaning of tags and content of elements.

Example of BioML

```
<chromosome number="11">
  <locus label="HUMINS locus">
  <gene label="Insulin gene">
    <dna start="1" end="4992" label="Complete HUMINS sequence">
      <ddomain start="1" end="2185" label="flanking domain"/>
      <ddomain start="1340" end="1823" label="polymorphic domain"/>
      <ddomain start="2424" end="2495" label="Signal peptide"/>
      <ddomain start="2496" end="2585" label="Chain B"/>
      <ddomain start="2586" end="2610" label="Chain C(1)"/>
      <ddomain start="3397" end="3476" label="Chain C(2)"/>
      <ddomain start="3477" end="3539" label="Chain A"/>
      <exon start="2186" end="2227" label="Exon 1"/>
      <intron start="2228" end="2406" label="Intron 1"/>
      <exon start="2407" end="2610" label="Exon 2"/>
      <intron start="2611" end="3396" label="Intron 2"/>
      <exon start="3397" end="3615" label="Exon 3"/>
      <ddomain  start="3615" end="4992" label="flanking domain"/>
<comment>
  The browser will ignore any symbol that cannot be a nucleotide
  residue, so the numbers can remain in place to aid the author.
</comment>
          1 ctcgaggggc ctagacattg ccctccagag agagcaccca acacccctcca
```

Limitations of XML

- Extending XML DTDs is very difficult.
  - Existing tools will only recognize extensions already provided by the original design.
- Linking information in different documents is difficult even for the same DTD.
- Merging information from different markup languages is usually impossible.
- XML does not support concept taxonomies.
- Limited support for general relationships.

Methods

Ontology based models

- Formal, declarative semantic models.
- Includes vocabulary terms, general relationships, constraints and rules.
- Supports concept taxonomies.
- Supports merging of information from disparate sources in many formats.
- Rules permit reasoning by autonomous agents at run-time.

RDF and DAML

- The Resource Description Framework (RDF) and DARPA Agent Markup Language (DAML) are the basis for the “Semantic Web.”
- DAML is an ontology language for semantic interoperability between Web pages, databases, programs and sensors.
- DAML is based on RDF, and RDF is based on XML.

Results

PathML Ontology and Language

- PathML is a DAML ontology for biopathways as well as a markup language for representing specific examples of complex biological pathways.
- Facilitates automatic retrieval and update of gene network-related data.
- Relevant PathML information can be easily transformed into XML markup languages using the XSLT transformation language.

The PathML ontology

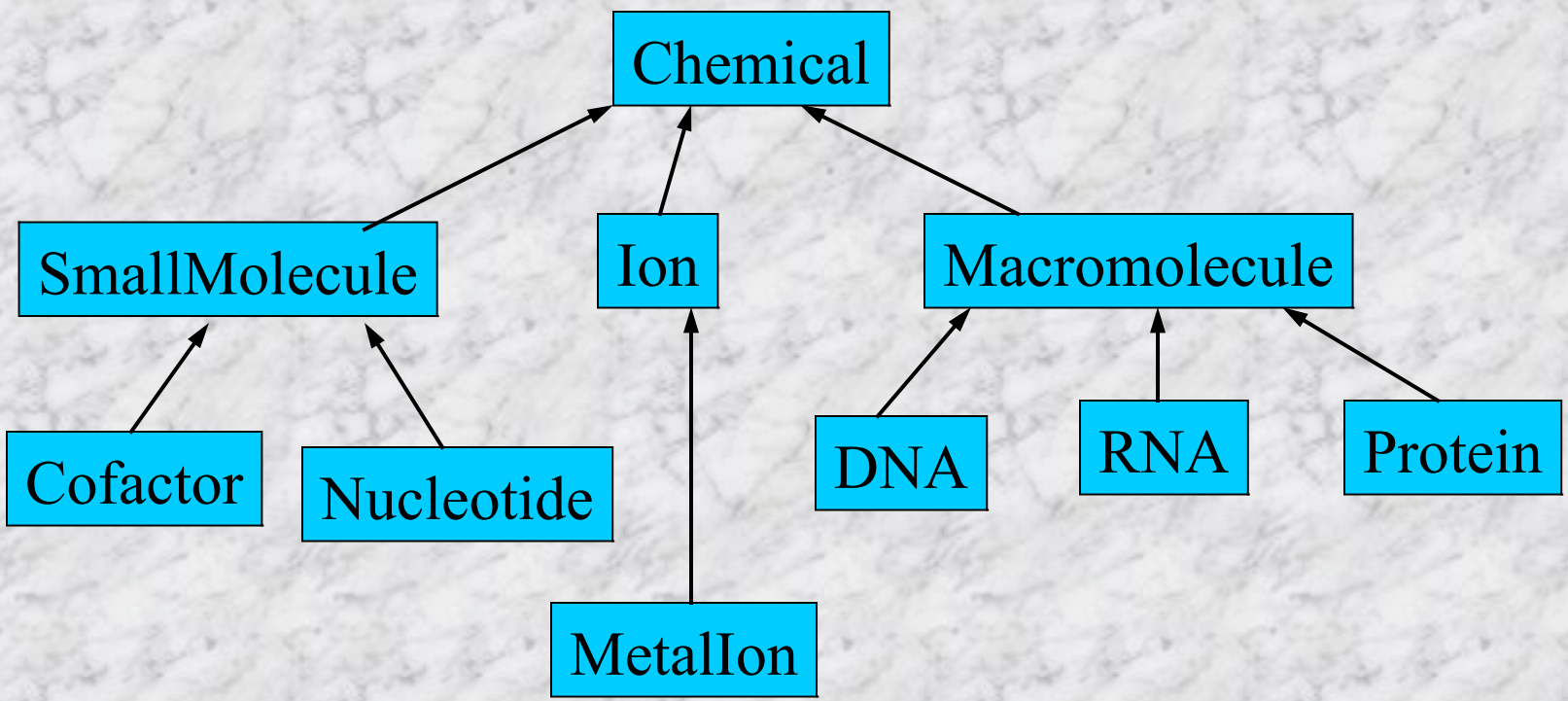
Examples of Classes

```
ReceptorSignallingProtein
Deoxynucleotide
Species
DefenseimmunityProtein
DNAStructure
TwoComponentSensorMolecule
ComplementDNA
Transporter
LigandBindingOrCarrier
ProteinTagging
Ribozyme
Chromosome
CytoskeletalRegulator
```

Examples of Properties

```
boundTo
hasStrandedness
stronglyBinds
translatedFrom
hasSpecies
polymerOf
formsPolymer
stronglyBoundTo
partOf
catalysedBy
```

Part of the PathML taxonomy



Discussion and Conclusion

- The ontology and markup language PathML has been developed for representing complex biological pathways.
- PathML is compatible with the many XML-based markup languages being developed biological data interchange.
- PathML supports automatic retrieval and update of gene network-related data from a variety of sources.

MotifML

- MotifML is a DAML ontology for gene regulation as well as a markup language for gene regulation motifs.
- Facilitates automatic retrieval and update of gene network-related data.
- Relevant MotifML information can be easily transformed into XML markup languages.

Part of the MotifML Ontology