Facilitating Efficient Knowledge Management and Discovery in the Agronomic Domain

Aravind Venkatesan
Post-doctoral researcher, Institut de Biologie Computationnelle (IBC)
aravind.venkatesan@lirmm.fr

Outline
• Data Landscape in the Life Sciences
• Data integration challenges in the Life Sciences
• Semantic Web Technologies
• Agronomic Linked Data (AgroLD) project
• Conclusions

Data landscape in the Life Sciences
• Amount of data generated in biological domain continues to grow exponentially.
• Trend in biological data growth:
  • Volume - empirical technologies
  • Velocity
  • Variety - diversity in biological system

Data integration challenges
• Lack of effective approaches to integrate data that has created a gap between data and knowledge.
• Need for an effective method to bridge gap between data and underlying meaning.
• Harvest the power of overlaying different data sets.

Today’s Web
• Today’s Web content is suitable for human consumption
• Collection of documents
  • the existence of links that establish connections between documents
• Low on data interoperability and lacks semantics.

Semantic Web Technology
• An extension of the current Web technologies.
• Enables navigation and meaningful use of digital resources.
• Support aggregation and integration of information from diverse sources.
• Based on common and standard formats.
**Resource Description Framework (RDF)**

- Framework for representing information about resources on the Web
- Provides a labeled connection between two resources
- Uses Unique Resource Identifiers (URI)
- Statements take the form of triples:

```
<Gene_A> <codes_for> <Protein_A>
```

**SPARQL**

- Language which allows querying RDF models (graphs)
- Powerful, flexible.
- Its syntax is similar to SQL.

**Semantic Web meets Systems Biology**

- The Semantic Web has gained steady acceptance among the life science community.
- The RDF data model complements systems biology – linking varied information sources
- SPARQL - address questions that were unapproachable using traditional methods.
- Number of initiatives undertaken to apply Semantic Web technology to assist Systems Biology approaches:
  - Semantic Systems Biology (SSB) platform
  - Hypothesis query and evaluation system (HyQue)
  - SEEK project

**Agronomic Linked Data (AgroLD)**

- Semantic web based system that captures knowledge pertaining to Agronomic data
- Aim:
  - Capability to answer complex real life questions
  - Efficient information integration / retrieval.
  - Easy extensibility.
  - Aid in holistic understanding of domain

**AgroLD – Phase I**

- AgroLD will be developed in phases –
  - **Phase I** includes data on Oryza spp. and Arabidopsis thaliana
- SPARQL endpoint: http://bioportal.lirmm.fr:8081/test
Information in AgroLD

- Integrates information from:
  - **Ontologies:** Gene Ontology, Plant Ontology, Plant Trait Ontology, Plant Environment Ontology, Crop Ontology
  - **Other information sources:**
    - Gene/Protein information: GOA, Gramene, Tair, OryzaBase, UniPort
    - QTL information: Gramene
    - Pathway information: RiceCyc, AraCyc
    - Germplasm information: Oryza Tag Line
    - Homology prediction: GreenPhyl

Future directions

- Phase II: having both wider and deeper coverage to promote comparative analysis
  - Include more crop species – *Sorghum* spp., *Zea mays*, *Triticum* spp.
  - Include varied data types – gene expression data, protein-protein interaction, Transcription factor-target gene
  - Developing methods to aid the process of hypotheses generation - e.g. inference rules.
  - Pluggable with workflow platforms e.g.: Galaxy, OpenAlea (VirtualPlants).
  - Engage with biologists to mobilise ‘user-pull’:
    - Develop real world use cases – studying the molecular mechanism of panicle differentiation in rice

Conclusions

- Biologists are beginning to adopt Systems Biology approaches to study biological systems in their entirety.
- Information shows its true value when linked with other pieces of information – the Semantic Web creates an opportunity.
- AgroLD - we aim to provide semantic web-based resource to manage agronomic data and provide a knowledge discovery platform.

Acknowledgements

Dr. Patrick Valduriez  Dr. Pierre Larmande  Dr. Manuel Ruiz  South Green team