BioMANTA

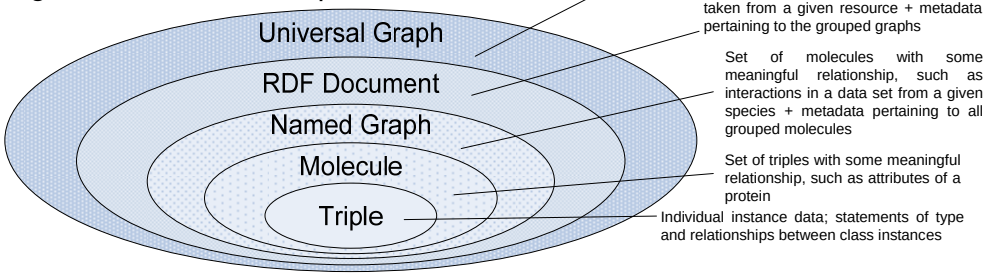
http://www.biomanta.org

Melissa J. Davis[1,3], Muhammad Shoaib Sehgal[1,3,] Kevin Burrage[1,3,4], Victor Farutin[5], Jane Hunter[2], Fred Jerva[5], Imran Khan[2], Yuan-Fang Li[2], Andrew Newman[2], Michael Schaffer[5], Christopher Bouton[5] and Mark A. Ragan[1,2,3]

[1]Institute for Molecular Bioscience; [2]Information Technology and Electrical Engineering Department, University of Queensland; [3]ARC Center of Excellence in Bioinformatics; [4]Advanced Computational Modeling Center; and [5]Computational Sciences Center of Emphasis, Pfizer Global Research & Development, Pfizer Inc.

The Modelling and Analysis of Biological Network Activity (BioMANTA) Project encompasses development of novel biological network analysis methods and infrastructure for querying biological data in a semantically-enabled format. Research within the BioMANTA project will focus on computational modelling and analysis, primarily using Semantic Web technologies, of large-scale protein-protein interaction and compound activity networks across a wide variety of species. A range of information such as kinetic activity, tissue expression, sub-cellular localization and disease state attributes will be included in the resulting data model. This project is a two- year scientific research collaboration between the Molecular Informatics group at the Pfizer Research Technology Centre (RTC), Cambridge, Massachusetts, USA and the Institute for Molecular Bioscience (IMB), The University of Queensland, Australia.

Organisation of RDF Representation



Storage and maintenance of meta-data are a basic requirement for presentation of semantic web resources. In order to efficiently maintain meta-data associated with instance data, a hierarchical ordering of RDF is adopted where provenance and meta-data is associated with sub-graphs ordered by some meaningful relationship.

Applying the Semantic Web

Resource Description Framework (RDF)

- A graph-based data structure consisting of triples
- For representation of relational data across the Web
- Open, extensible – INTEGRATION!

RDF Molecules

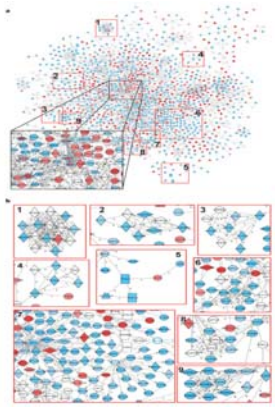
- A set of RDF triples within a graph
- Extension of RDF for easy merging

OWL

- Semantic modeling of biological data – an upper ontology to encompass all relevant datasets

Clustering

- Capitalizing on distributed processing over consumer-grade hardware
- Supports querying & inferencing

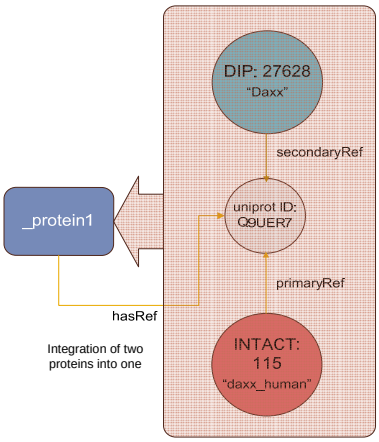


Knowledge Discovery and Network Analysis

The known biochemical interaction network is very small to answer the biological questions. The complete network is therefore inferred using new Machine learning and stochastic methods by fusing networks reconstructed from heterogeneous data sources.

The networks are first inferred using Gene Expression data, Protein Localization and Phylogenetic profiles datasets and then integrated together to form the final network.

The network produced through these “knowledge discovery” methods are made available by “knowledge representation”. The network can be evaluated for its usefulness in an array of already developed network analysis techniques such as, Markov Clustering and Gene Rank, as well as for development of new methods computationally utilizing information about relationships between genes (and their products) to analyze data and generate hypothesis relevant to biomedical research.



BioMANTA Queries

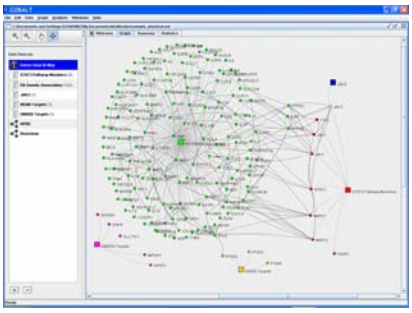
The aim of the BioMANTA project is to create a semantic web representation of biomedical and biomolecular data that can be queried to uncover links between biological systems, networks and disease. The following example illustrates the types of queries that the BioMANTA framework will address:

A recent study (Ray et al.,) demonstrated that blood plasma levels of 18 signaling proteins can be used to predict Alzheimer’s progression in presymptomatic patients.

- •How are these biomarkers linked to Alzheimer’s Disease?
- •What signaling pathways and biological functions are these proteins associated with?
- •Are these genes or pathways significantly linked to knockout mouse phenotypes related to memory or cognition?
- •Can a causal link between these pathways and Alzheimer’s be hypothesized?
- •Is this gene set significantly linked to other diseases?
- •Are there existing therapies for these diseases that could be viable Alzheimer’s treatments?

BioMANTA Output and Visualization

After executing queries across the BioMANTA resource, one significant challenge will be to visualize the returned results.



The COBALT (Connection of Biological Lists) prototype filters data from HPRD to create a list of genes annotated with specified properties, GO terms, or disease associations. The application takes two such gene lists and displays known protein-protein interaction networks associated with the listed genes, highlighting areas of overlap between networks (left).

Alternatively, taking a single node (gene) as a starting point, sets can be expanded to show the nearest neighbours (right).

One key challenge will be visualising data at a higher resolution than gene level. While network analysis is frequently carried out at this level of resolution, the recent discovery of extensive variability in the structure and function of the protein products of genes indicates that this diversity must be accommodated in future network analysis. This additional complexity, along with data from tissue specific and time-course expression profiling generates challenges for the task of visualising the output of queries over the semantic interactome.

The University Of Queensland AUSTRALIA
IMB Institute for Molecular Bioscience
Pfizer
Australian Government
Australian Research Council
ARC Centre of Excellence in Bioinformatics