Grand Challenges in Computational Biology

Kimmen Sjölander
UC Berkeley

CITRIS-INRIA workshop
24 May, 2011

Supported in part by a grant from the DOE Systems Biology Knowledgebase
• Phylogenomic predictions of function and structure for microbial genomes and metagenomes.
• Simultaneous functional and taxonomic annotation of environmental sequences and human microbiome data.

Human microbiome and metagenome dataset analysis
Reconstructing the Tree of Life

Prediction of biological pathways and networks

Infectious disease: new drugs and diagnostics; pharmacogenomics
Interpreting genetic variation

This work is supported by:
The US DOE Systems Biology Knowledgebase,
the NSF Microbial Genome Sequencing Program,
a Presidential Early Career Award for Scientists and Engineers from the NSF,
and by an R01 from the NHGRI (NIH).
• The situation now: huge quantities of noisy, error-ridden and poorly connected data
  – Experimental data are sparse: \(~1\%\) of sequences have experimental support for their assigned functions
  – Errors abound: Up to 25\% of sequences are mis-annotated [1, 2]
  – The one-time static annotation protocol does not allow annotations to be modified in the light of new evidence [3]
  – Expert knowledge is critical to detecting and correcting annotation errors
    • But manual annotation is expensive and does not scale to the quantity of sequences being produced

3. “Genome re-annotation: a wiki solution?” Salzberg, Genome Biology 2007
Increasing the specificity of function prediction requires the integration of heterogeneous data & bioinformatics methods

Homology & orthology prediction
Genome neighbors
Expression data
Localization information
3D structure
Yeast-2-hybrid data
Phylogenetic profiles
Pull-down assays
Site-directed mutagenesis
Text-mining (co-occurrence in an abstract)
Etc.

Eisenberg et al, “Protein function in the post-genomic era” Nature 2000
Matthews et al, “Identification of Potential Interaction Networks Using Sequence-Based Searches for Conserved Protein-Protein Interactions or “Interologs”” Genome Research 2001
Myers et al, “Discovery of biological networks from diverse functional genomic data,” Genome Biology 2005
Data is not the same thing as information
Biologists who need to use bioinformatics tools are divided by a huge gulf from the computer scientists who are creating these tools.
Automatic protein function prediction using a hyper-dimensional network
Hyperdimensional information network
for data integration, navigation & community annotation

**Nodes:** Genes/proteins  
**Edges:** different types of connection between genes (e.g., orthology, similar structure, interaction, disease association, regulated by, adjacent in metabolic network, genome neighbor, etc.).

**Edges have weights** proportional to confidence

Experimental data can enter at any point in the graph, and be propagated to neighboring nodes based on learned rules:

- Biological process for one gene can be made available to genome neighbors
- A protein-protein interaction between two genes in one species can be used to infer corresponding interaction between their orthologs in another
- Roles in a pathway (e.g., EC number) known for one gene can be assigned to an ortholog
- Participation in a biological process can be inferred based on genome neighbors
- 3D structure information can be propagated to all homologs
- Protein structure information can be propagated to all homologs

**Biologists can:** subscribe to news feeds arriving at their selected nodes, upload data, attach links to their papers, manually curate biological “functions”

**Manual annotations** from biologists will need to be weighted according to estimated confidence
Phylogenomic tools for investigating and interpreting (meta)genome datasets
( DOE Systems Biology Knowledgebase grant)

Challenges in metagenome data analysis:
• Most tools designed for these data answer only “What species are present?” and do not answer the question, “What’s going on?” (what processes & pathways are represented)
• Sequences are fragmentary and noisy, presenting additional challenges to bioinformatics methods
• Huge datasets (in the millions of reads)

“Harnessing the power of the human microbiome”, Blaser, PNAS 2010
SNPs occurring in coding regions of the genome can be prioritized for investigation based on:

- Predicted biological process or function of gene containing SNP
- Predicted interactions (hubs of networks) of gene containing SNP
- Impact of mutation at that site (INTREPID and Discern methods)
PhyloFacts Pathogen Commons

- Drug target identification & prioritization
- Development of accurate diagnostics

TB collaborations
- UC Berkeley Center for Emerging and Neglected Diseases (Tom Alber, Lee Riley, others)
- Royal Institute of Tropical Diseases, Amsterdam, Netherlands (Richard Anthony)
- Institute of Bioinformatics, Bangalore, India (Akhilesh Pande)
- IISc, Bangalore, India (Nagasuma Chandra)
Was there really life before the web?

How can we bring this to biology?