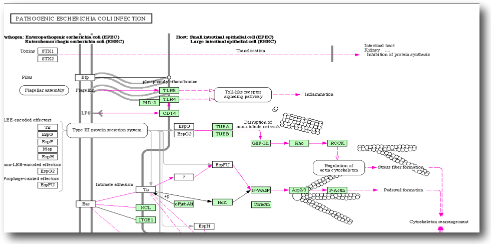
Grand Challenges in Computational Biology

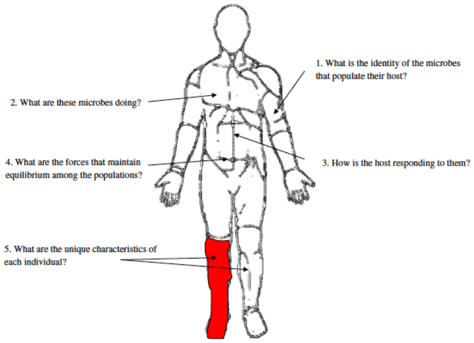
Kimmen Sjölander
UC Berkeley

CITRIS-INRIA workshop
24 May, 2011

Reconstructing the Tree of Life

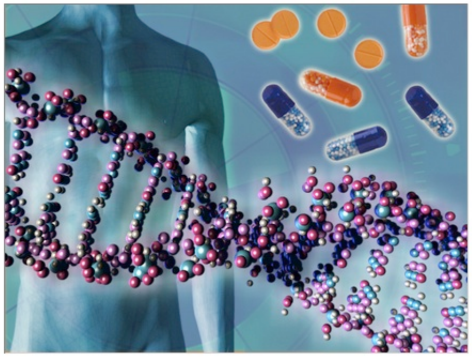

PATHOGENIC ESCHERICHIA COLI INFECTION
Pathogen: Enteropathogenic escherichia coli (EPEC) Enterohemorrhagic escherichia coli (EHEC)
Host: Small intestinal epithelial cell (EPEC) Large intestinal epithelial cell (EHEC)
Toxins
STX1
STX2
Translocation
Intestinal tract
Kidney ....
Inhibition of protein synthesis
Pilus
Bfp
phosphatidylethanolamine
Flagellar assembly
Flagellin
TLR5
TLR4
MD-2
Toll-like receptor signaling pathway
Inflammation
LPS
CD14
LEE-encoded effectors
Tir
EspG
EspF
Map
EspH
non-LEE-encoded effectors
EspG2
Prophage-carried effectors
EspFU
Type III protein secretion system
EspG
EspG2
TUBA
TUBB
Disruption of microtubule network
GEF-H1
Rho
ROCK
Regulation of actin cytoskeleton
Stress fiber formation
EspFU
?
Intimate adhesion
Tir
+p
c-Yes
NcK
N-WASP
Arp2/3
F-Actin
Pedestal formation
Cortactin
Eae
NCL
ITGB1
EspH
Cytoskeleton rearrangement

Prediction of biological pathways and networks


1. What is the identity of the microbes that populate their host?
2. What are these microbes doing?
3. How is the host responding to them?
4. What are the forces that maintain equilibrium among the populations?
5. What are the unique characteristics of each individual?

Human microbiome and metagenome dataset analysis

Infectious disease: new drugs and diagnostics; pharmacogenomics



Interpreting genetic variation

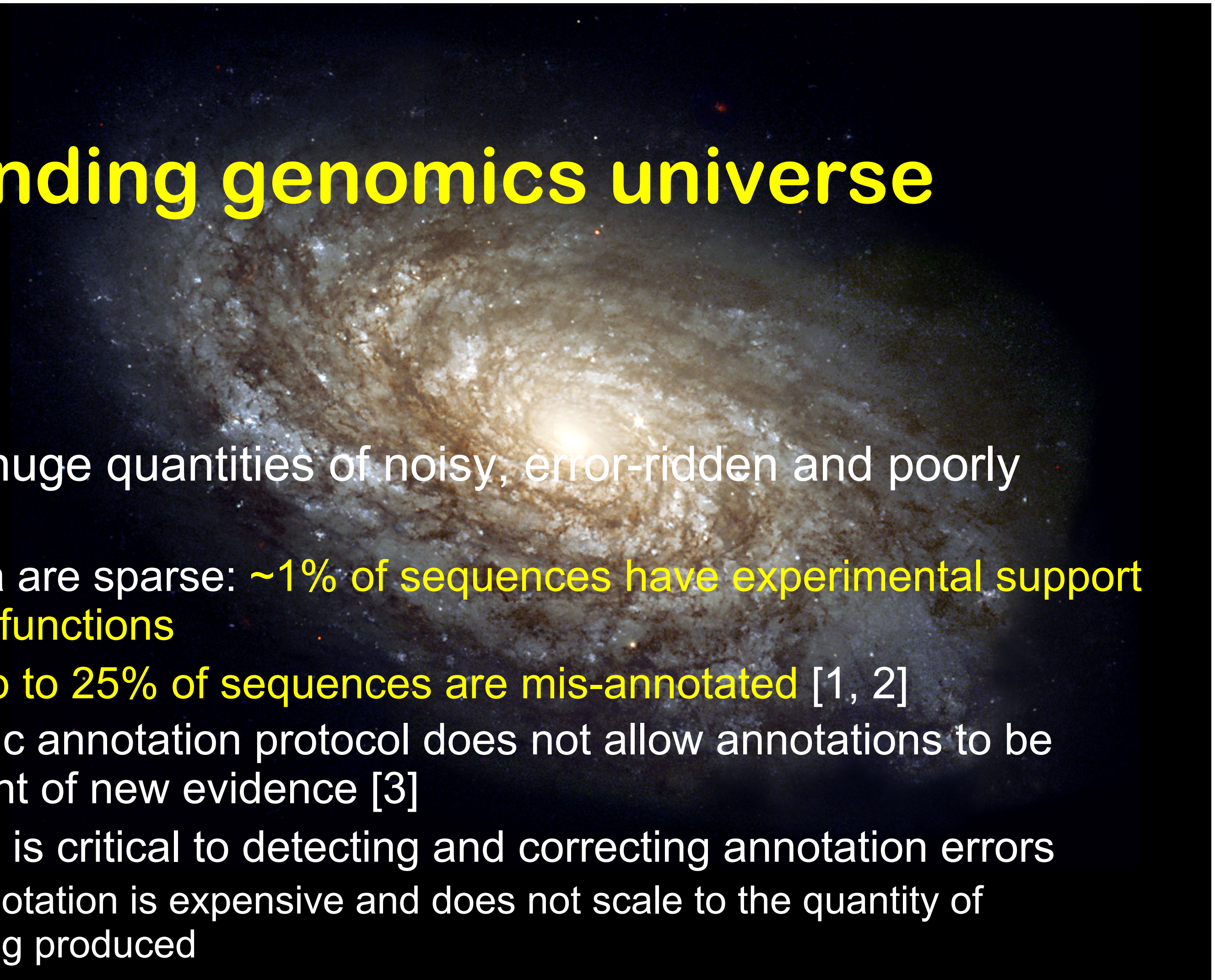
Supported in part by a grant from the DOE Systems Biology Knowledgebase
•Phylogenomic predictions of function and structure for microbial genomes and metagenomes.
•Simultaneous functional and taxonomic annotation of environmental sequences and human microbiome data.

U.S. DEPARTMENT OF ENERGY

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).

NSF

NATIONAL INSTITUTES OF HEALTH

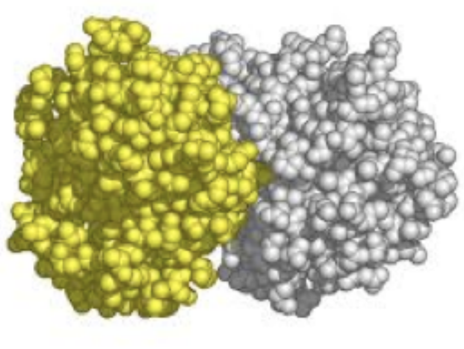


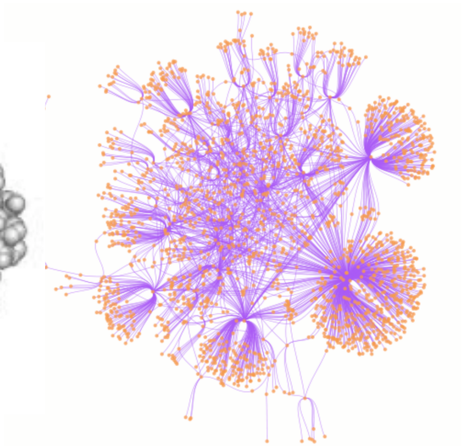
The expanding genomics universe

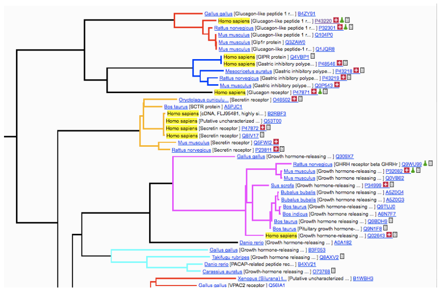
- The situation now: huge quantities of noisy, error-ridden and poorly connected data
  - Experimental data are sparse: ~1% of sequences have experimental support for their assigned functions
  - Errors abound: Up to 25% of sequences are mis-annotated [1, 2]
  - The one-time static annotation protocol 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

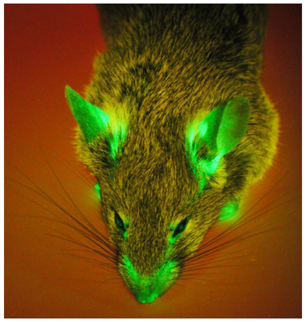
1. "Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies," Schnoes et al, PLoS Computational Biology 2009
2. "Phylogenomic inference of protein molecular function: advances and challenges," Sjolander, Bioinformatics 2004
3. "Genome re-annotation: a wiki solution?" Salzberg, Genome Biology 2007

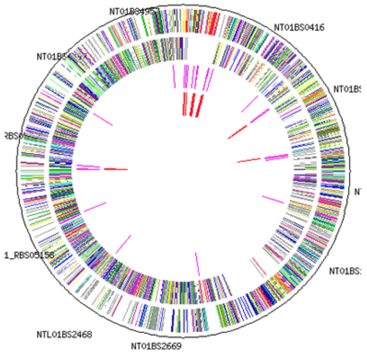
Increasing the specificity of function prediction requires the integration of heterogeneous data & bioinformatics methods



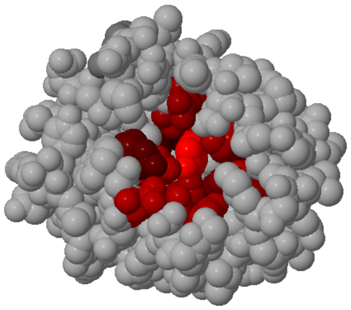


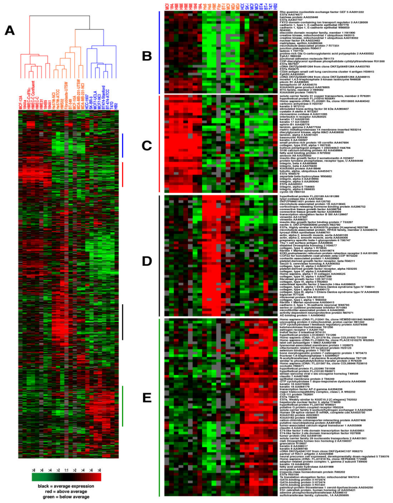






Homology & orthology prediction
Genome neighbors
Expression data
Localization information
3D structure
Yeast-2-hybrid data
Phylogenetic profiles
Pull-down assays
Site-directed mutagenesis
Text-mining (co-occurrence in an abstract)
Etc.

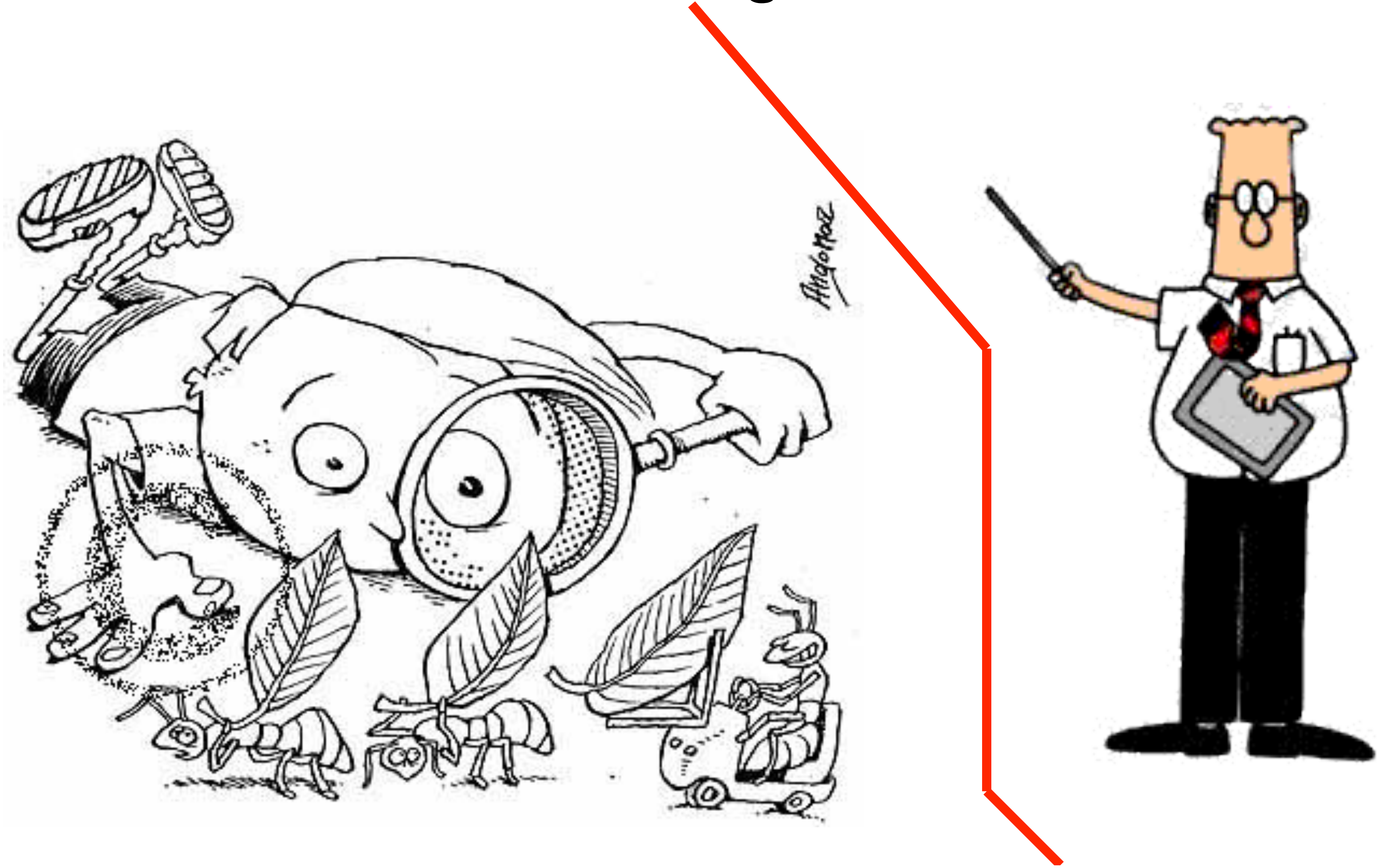




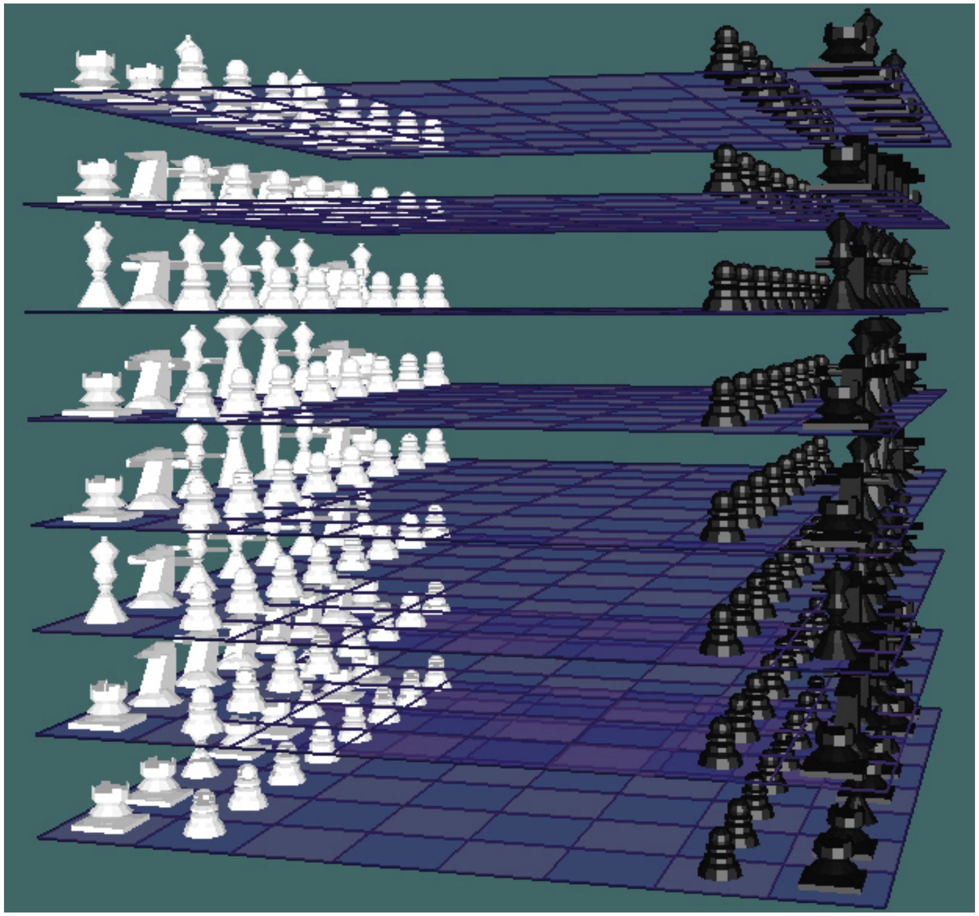
Eisenberg et al, “Protein function in the post-genomic era” *Nature* 2000
Sjölander, K., "Phylogenomic inference of protein molecular function: advances and challenges," Bioinformatics 2004
Matthews et al, “Identification of Potential Interaction Networks Using Sequence-Based Searches for Conserved Protein-Protein Interactions or “Interologs”” *Genome Research* 2001
Troyanskaya et al, “A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae),” PNAS, 2003
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

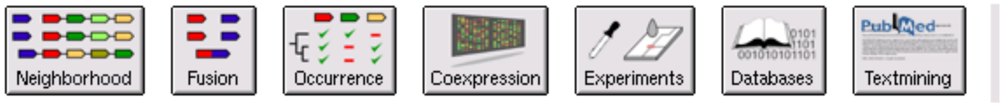


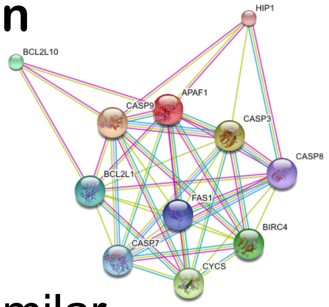
Automatic protein function prediction using a hyper-dimensional network



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:

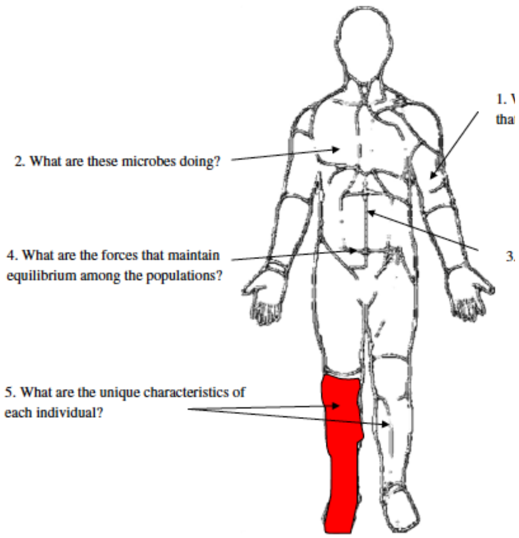
- •Biological process for one gene can be made available to genome neighbors
- •A protein-protein interaction between two genes in one species can be used to infer corresponding interaction between their orthologs in another
- •Roles in a pathway (e.g., EC number) known for one gene can be assigned to 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

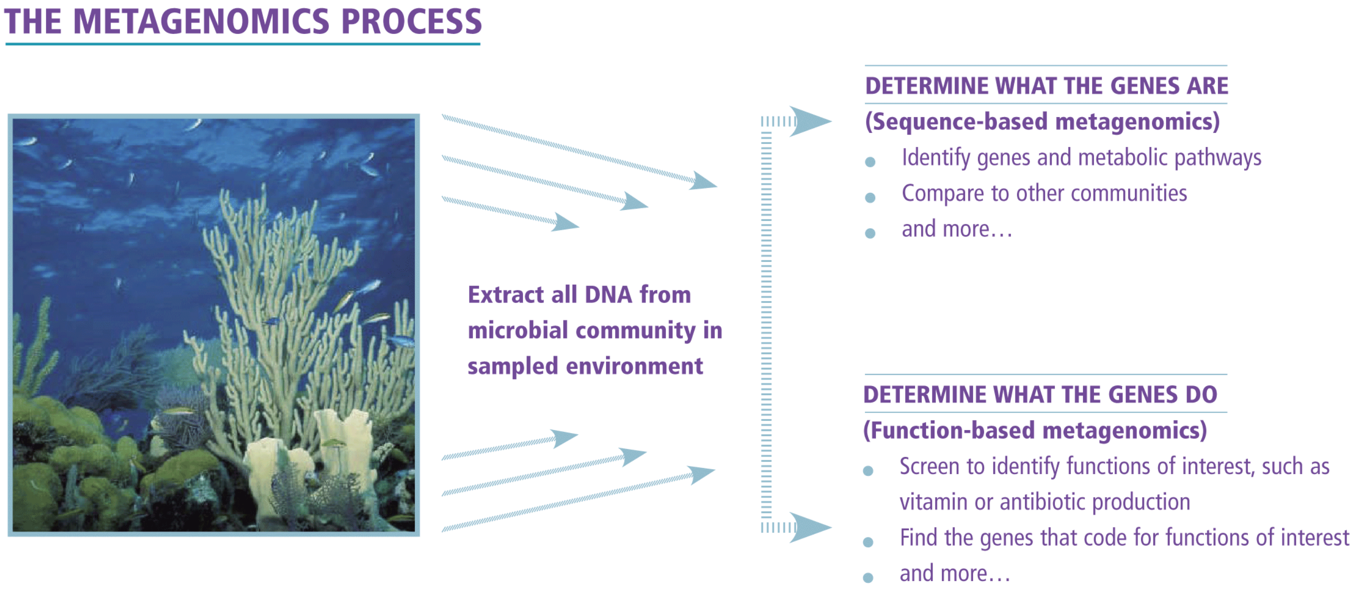
Biologists can: subscribe to news feeds arriving at their selected nodes, upload data, attach links to their papers, manually curate biological “functions”

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)


2. What are these microbes doing?
4. What are the forces that maintain equilibrium among the populations?
5. What are the unique characteristics of each individual?
1. V
tha
3.

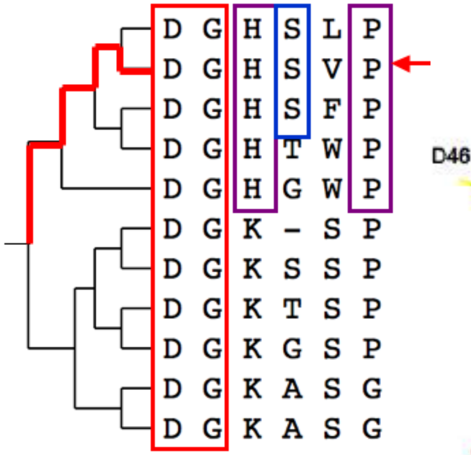

THE METAGENOMICS PROCESS
Extract all DNA from microbial community in sampled environment
DETERMINE WHAT THE GENES ARE
(Sequence-based metagenomics)
Identify genes and metabolic pathways
Compare to other communities
and more…
DETERMINE WHAT THE GENES DO
(Function-based metagenomics)
Screen to identify functions of interest, such as vitamin or antibiotic production
Find the genes that code for functions of interest
and more…

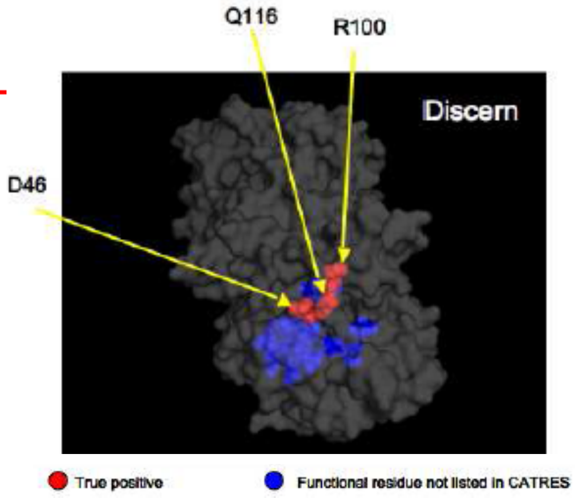
Challenges in metagenome data analysis:

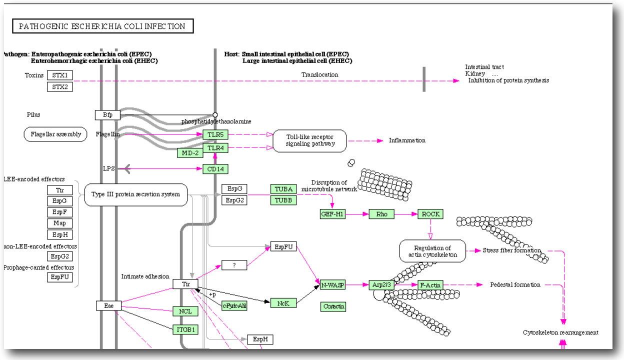
- Most tools designed for these data answer only “What species are present?” and do not answer the question, “What’s going on?” (what processes & pathways are represented)
- Sequences are fragmentary and noisy, presenting additional challenges to bioinformatics methods
- **Huge** datasets (in the millions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

“Harnessing the power of the human microbiome”, Blaser, PNAS 2010
“The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet” Committee on Metagenomics: Challenges and Functional Applications, National Research Council. 2007.

SNP prioritization and interpreting human genetic variation







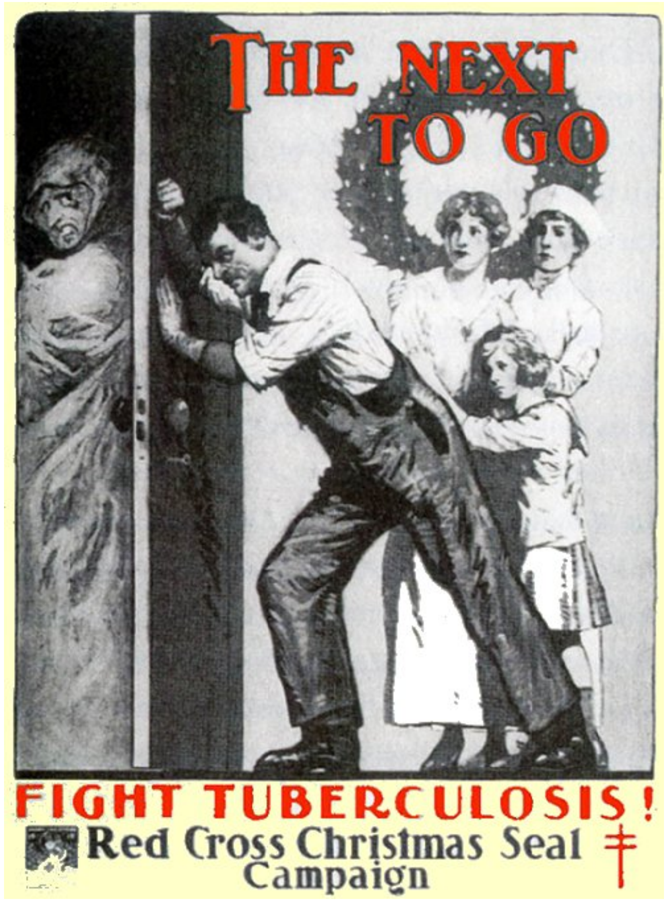
PATHOGENIC ESCHERICHIA COLI INFECTION
Translocation
Inflammation
Toll-like receptor signaling pathway
Type III protein secretion system
Intimate adhesion
Regulation of actin cytoskeleton
Stress fiber formation
Pedestal formation
Cytoskeleton rearrangement

Prediction of biological pathways and network alignment

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

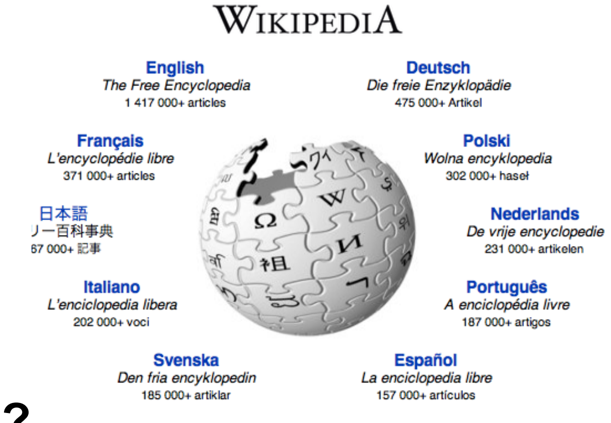


- **Drug target identification & prioritization**
- **Development of accurate diagnostics**

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)

Google™



WIKIPEDIA

English
The Free Encyclopedia
1 417 000+ articles

Deutsch
Die freie Enzyklopädie
475 000+ Artikel

Français
L'encyclopédie libre
371 000+ articles

Polski
Wolna encyklopedia
302 000+ haseł

日本語
フリー百科事典
67 000+ 記事

Nederlands
De vrije encyclopedie
231 000+ artikelen

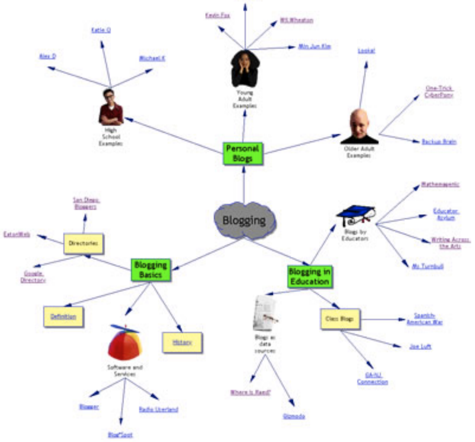
Italiano
L'enciclopedia libera
202 000+ voci

Português
A enciclopédia livre
187 000+ artigos

Svenska
Den fria encyklopedin
185 000+ artiklar

Español
La enciclopedia libre
157 000+ artículos

Was there really life before the web?



facebook

Welcome to Facebook!

How can we bring this to biology?

tripadvisor®
get the truth. then go.™

5,000,000+ traveler reviews & opinions of hotels, vacations & more

Sign in • Sign up