



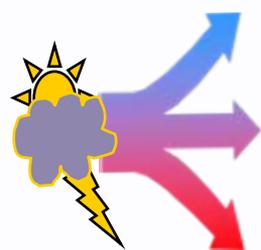
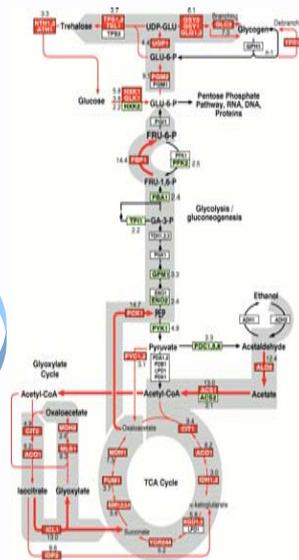
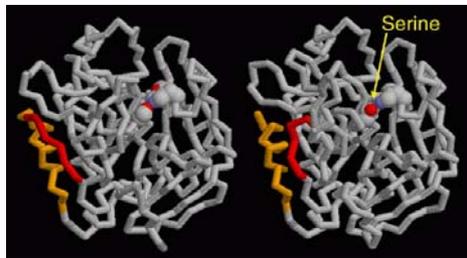
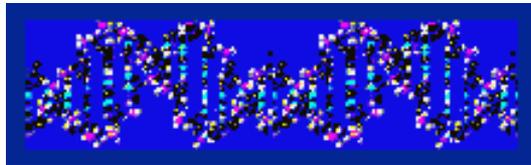
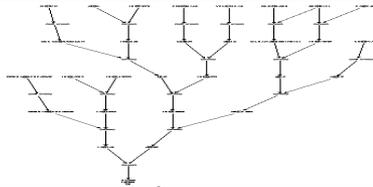
Enabling Biology to Become a Predictive Science

William D. Beavis, Geert Wenes, John Ambrosiano

Computational Infrastructure for GTL Workshop
22 January 2002

NATIONAL CENTER FOR GENOME

Goal: $P(\text{Phenotype}|\text{Data},\text{Model})$





Discovery-Driven Data, Models and Computing Resources for Predicting the Phenotype.

Data and Quality

Models and Lack of Fit

Compute Resources

Ancestry &
Environment

DNA

*m*RNA

Proteins

Biochemical
Networks





Ancestry and Environment

Data:

- ⌘ Pedigrees, Breeding Records, Progeny Performance Trials
- ⌘ Family Histories, Clinical Records, Environmental Exposure
- ⌘ $10^5 - 10^7$ records per trait in geographically distributed, heterogenous repositories.
- ⌘ $\varepsilon \sim .05 - .5$

Models:

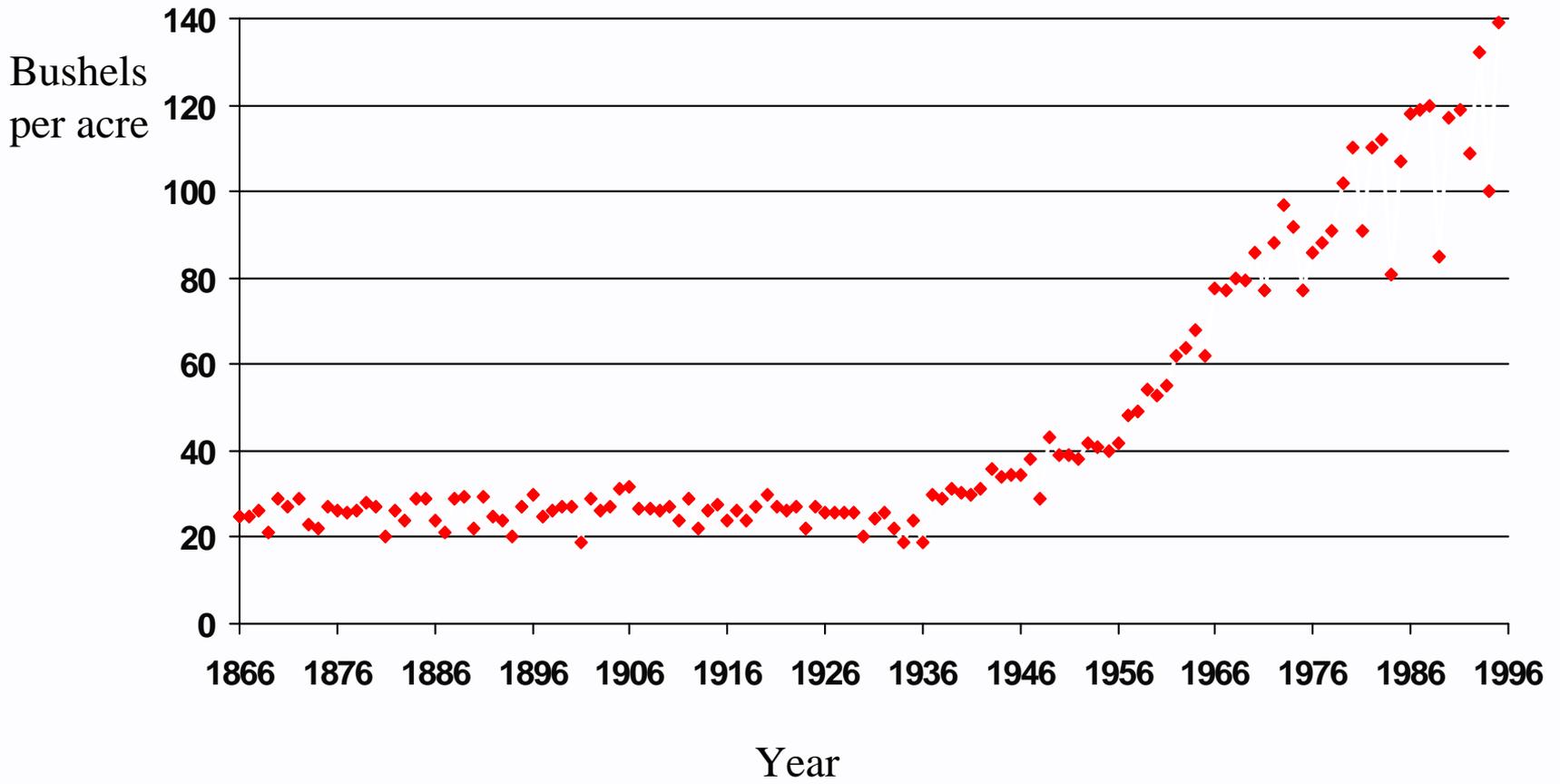
- ⌘ $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mu + \varepsilon$

Compute Resources:

- ⌘ For brute-force ML, RAM needs to accommodate data from multiple sources and a few million records.



Average U.S. Corn Yields



Predictions from DNA information

Data:

- ⌘ Physical Maps, Linkage Maps and SNPs
- ⌘ Nucleotide Sequences
- ⌘ 1.6×10^{10} sequences in a few geographically distributed repositories.
- ⌘ Sequencing errors, $\varepsilon \sim .01$; annotation errors ?

Models:

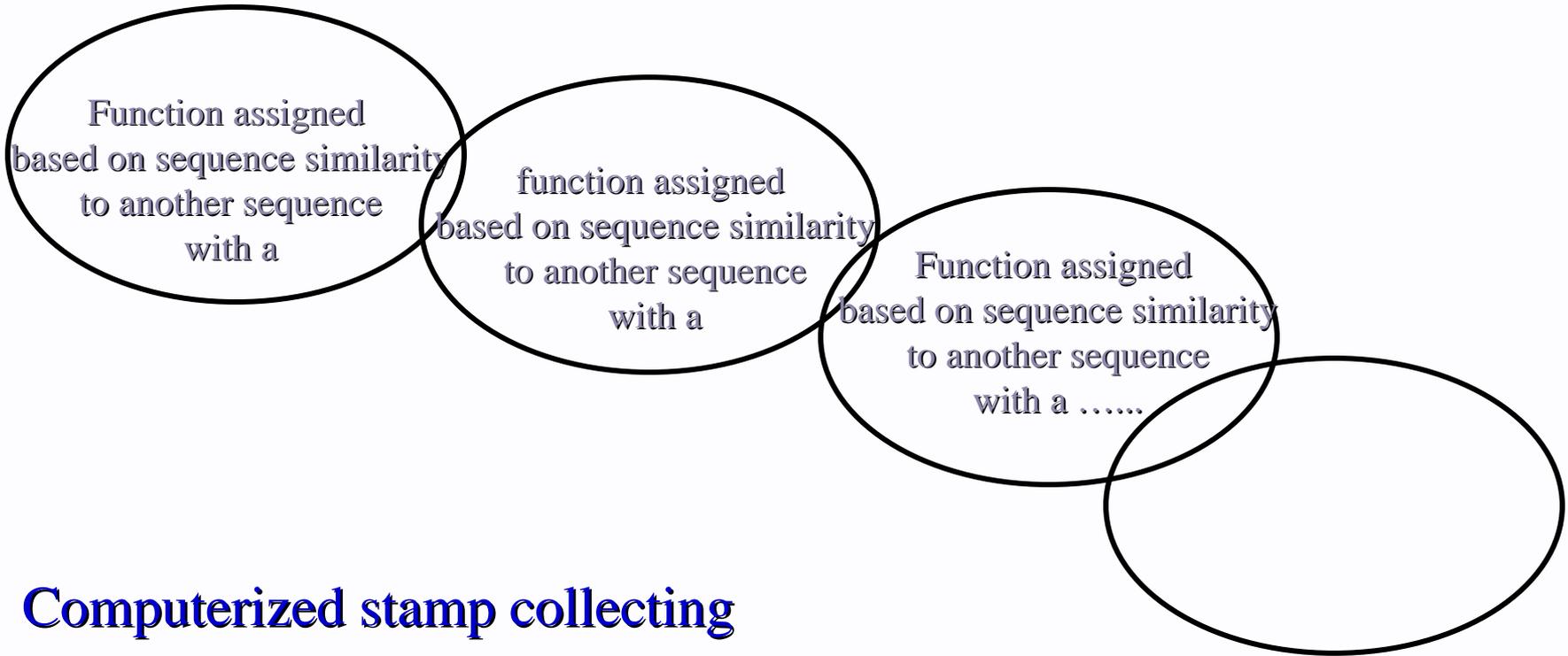
- ⌘ $y = X\beta + Z\mu + Wt + \varepsilon$
- ⌘ Blast ($\text{seq}_i \sim \text{seq}_j \Rightarrow \text{function}$) ; $s_{ij} = \ln(q_{ij}/P_i P_j) / \lambda_u$

Compute Resources:

- ⌘ 200K ESTs: 2 weeks, 16-node Linux cluster with 16 Gb RAM



Weakness of sequence annotations:





Predictions from DNA information

Data:

- ⌘ Physical Maps, Linkage Maps and SNPs
- ⌘ Nucleotide Sequences
- ⌘ 1.6×10^{10} sequences in a few geographically distributed repositories.
- ⌘ Sequencing errors, $\varepsilon \sim .01$; annotation errors ?

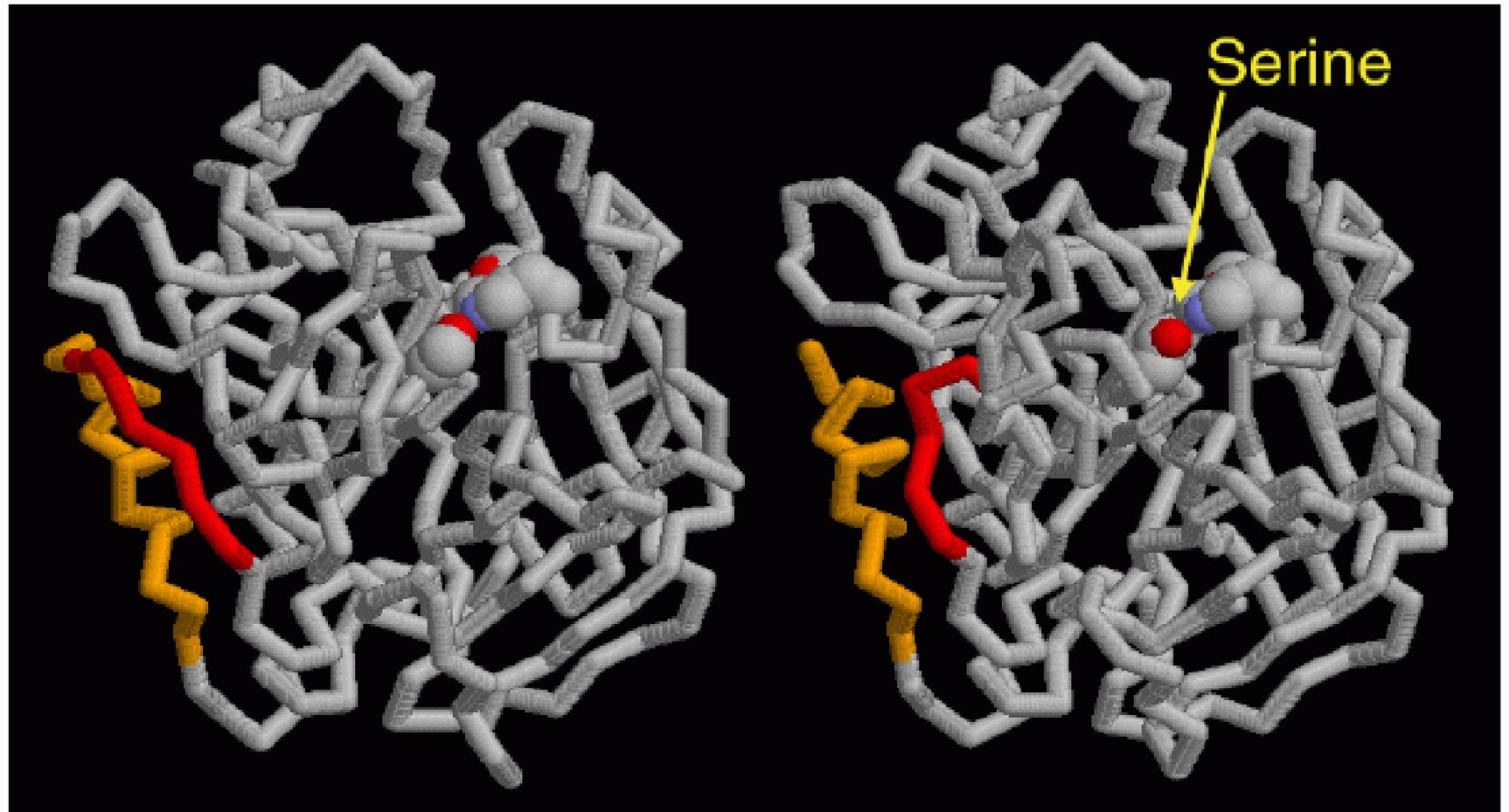
Models:

- ⌘ $y = X\beta + Z\mu + W\tau + \varepsilon$
- ⌘ Blast ($\text{seq}_i \sim \text{seq}_j \Rightarrow \text{function}$) ; $s_{ij} = \ln(q_{ij}/P_i P_j) / \lambda_u$
- ⌘ **CASP (seq \Rightarrow structure \Rightarrow function)**

Compute Resources:

- ⌘ Moderate sized clusters of commodity-grade processors.





Understanding the Mechanisms of Protein Folding

⌘ Will lead to an understanding of protein function in various cellular environments

(Misfolding is known to occur and be responsible for serious diseases)

⌘ Resulting in development of predictable diagnostics, novel protein-based therapeutics, novel proteins for sequestering C, heavy metals, and radioactive isotopes.



Understanding the Mechanisms of Protein Folding

Experimental techniques are limited for relevant time scales, thus there is a need for simulation of

- ⌘ **folding kinetics**
- ⌘ **folding pathways**
- ⌘ **force-field assessments**

Allen et al. 2001. “Blue Gene: A vision for protein science using a petaflop supercomputer”. IBM Systems Journal 40:310-327.

Levitt et al. 2002. “Modeling Across the Scales - Atoms to Organisms” Mathematics and Molecular Biology VII. Program in Mathematics and Molecular Biology. Santa Fe, N. Mex.





Computing Needs for Mechanisms of Protein Folding

Time frame to simulate	10^{-4} sec
Time-step size	10^{-15} sec
Number of MD time steps	10^{11}
Atoms in a typical protein/water simu	3.2×10^5
Number of interactions in a force calc	10^9
Instructions per force calc	10^3
Total number of instructions	10^{23}

Allen et al. 2001. "Blue Gene: A vision for protein science using a petaflop supercomputer"
IBM Systems Journal 40:310-327.





Predictions from mRNA and proteomic arrays

Data:

- ⌘ gene-chips, micro-arrays, bead-arrays (MPSS), proteomic arrays
- ⌘ very sparse data, relative to biological time scales, snap shots of average transcription from pools of similarly treated cells
- ⌘ $10^4 - 10^5$ genes per experiment; 10^4 experiments in geographically distributed, heterogenous repositories.
- ⌘ $\varepsilon \sim .1 - .5$

Models:

- ⌘ $y = X\beta + Z\mu + W\iota + T\nu + \varepsilon$
- ⌘ Support Vector Machines

Compute Resources:

- ⌘ RAM needs to accommodate data from multiple arrays, multiple experiments, GenBank, SwissProt, Kegg, Pfam, PathDB



Examples of data from micro-arrays and 2 D gel protein arrays

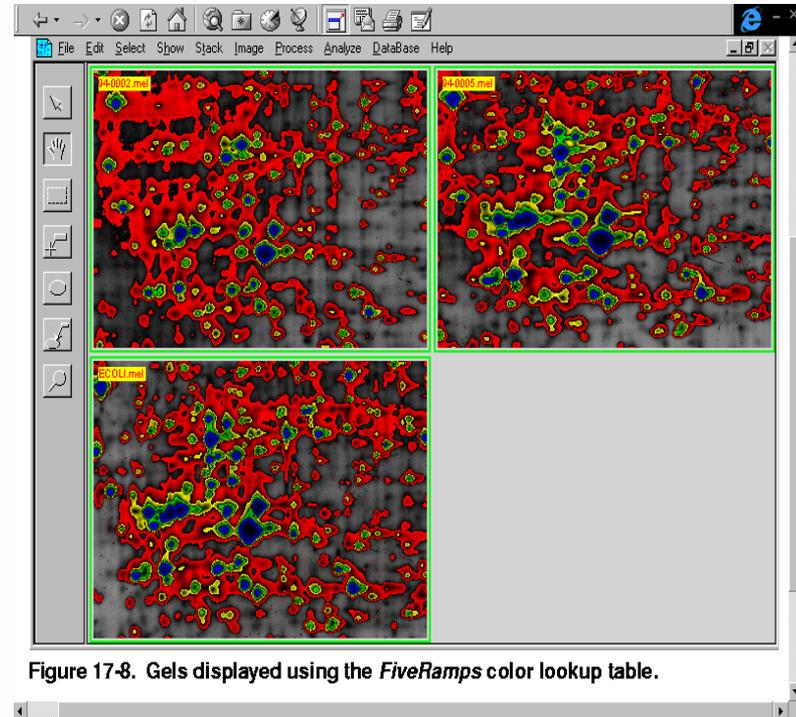
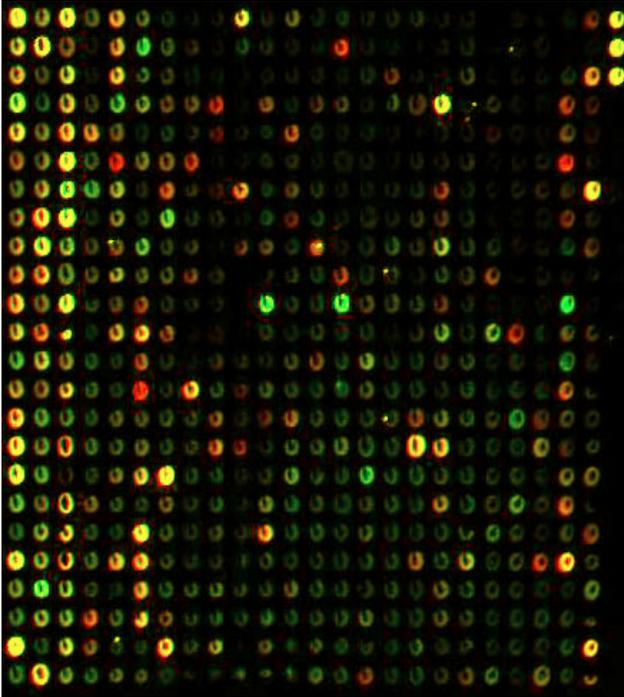
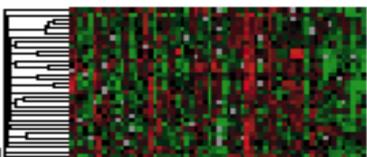
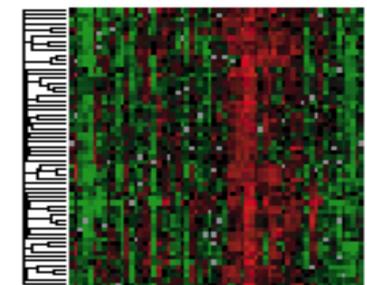
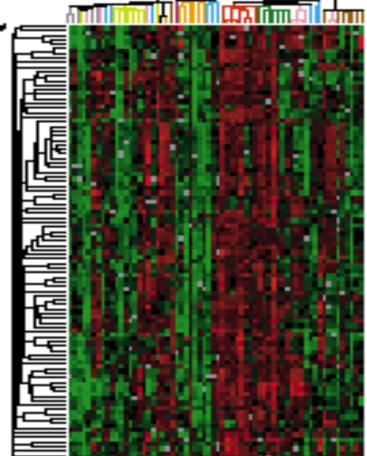
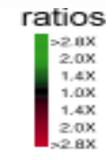
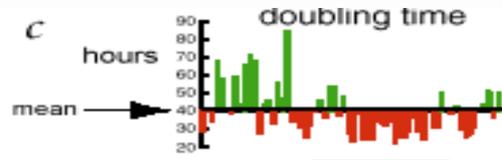
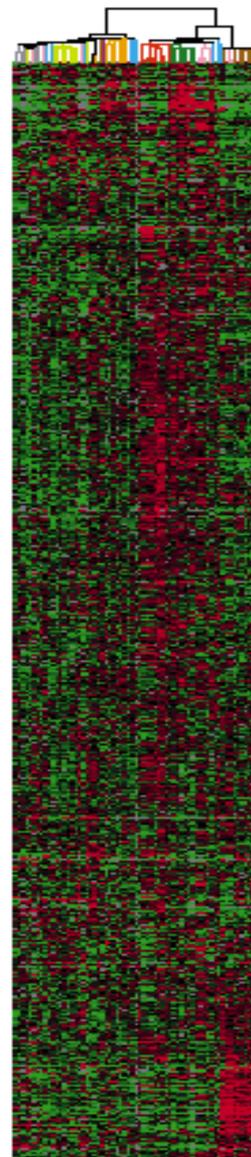


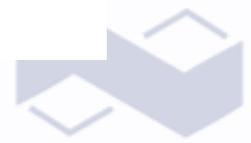
Figure 17-8. Gels displayed using the *FiveRamps* color lookup table.

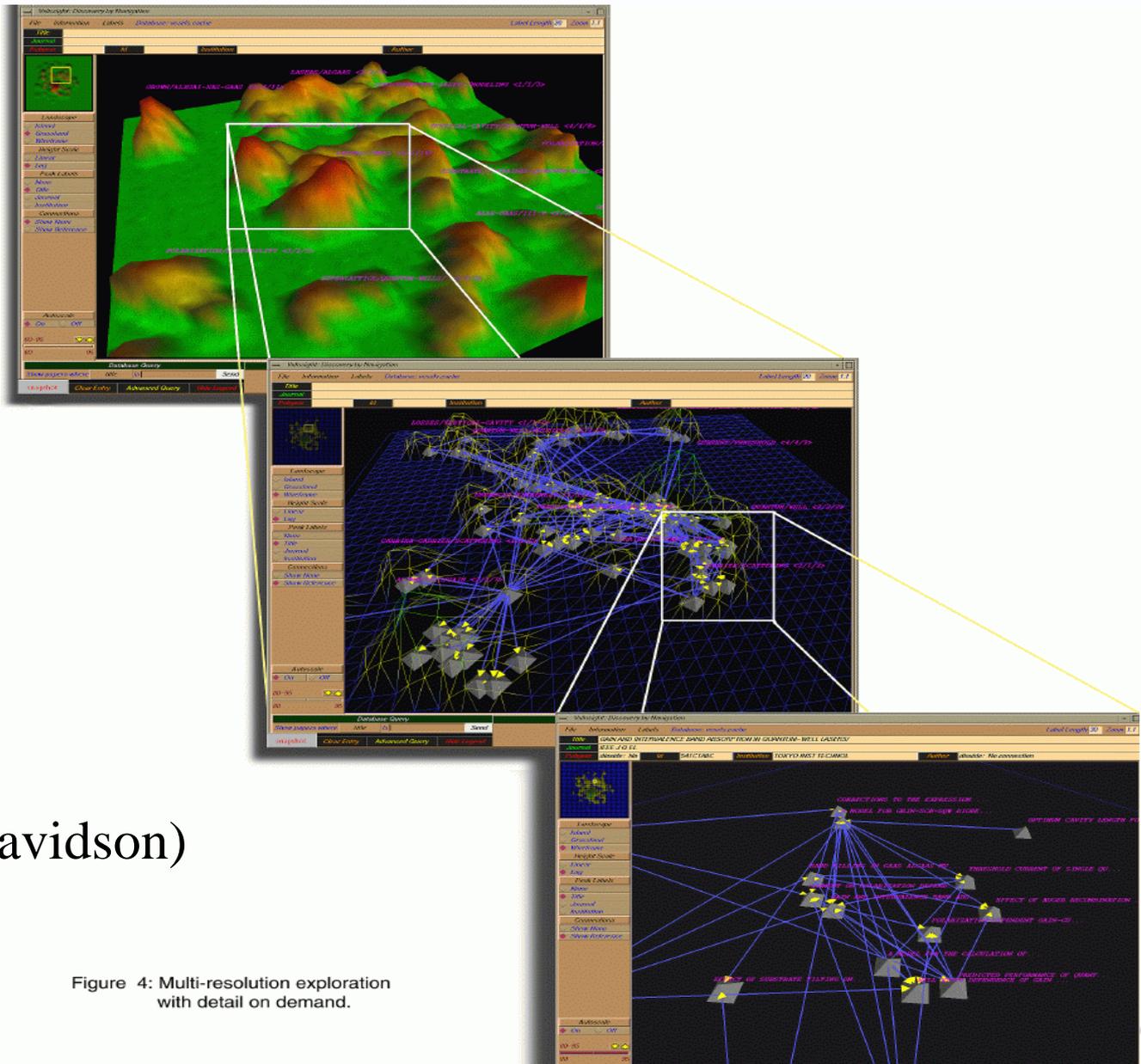


b
6831 genes



d proliferation cluster





VxInsight (SNL: George Davidson)

Figure 4: Multi-resolution exploration with detail on demand.



Predictions from Expression Arrays

Data:

- ⌘ gene-chips, micro-arrays, bead-arrays (MPSS), proteomic arrays
- ⌘ very sparse data, relative to biological time scales, snap shots of average transcription from pools of similarly treated cells
- ⌘ $10^4 - 10^5$ genes per experiment; 10^4 experiments in geographically distributed, heterogenous repositories.
- ⌘ $\varepsilon \sim .1 - .5$

Models:

- ⌘ $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mu + \mathbf{W}\iota + \mathbf{T}\nu + \varepsilon$
- ⌘ **Support Vector Machines**

Compute Resources:

- ⌘ RAM needs to accommodate data from multiple arrays, multiple experiments, GenBank, SwissProt, Kegg, Pfam, PathDB



Pathways and Networks

Data:

- ⌘ expression arrays
- ⌘ cross-link arrays of protein-protein, protein-DNA and protein-RNA interactions.
- ⌘ Very sparse data, i.e., snap shots at discrete time/treatment/tissue samples
- ⌘ geographically distributed, heterogeneous repositories

Models:

- ⌘ Hypothetical networks of nodes and edges, e.g., Bayesian Networks

Compute Resources:

- ⌘ Similar to “radiation transport” simulations?





Data Needed for Inferring Genetic Networks from N Genes

Boolean, Fully Connected

$$2^N$$

Boolean, K Connections per Gene

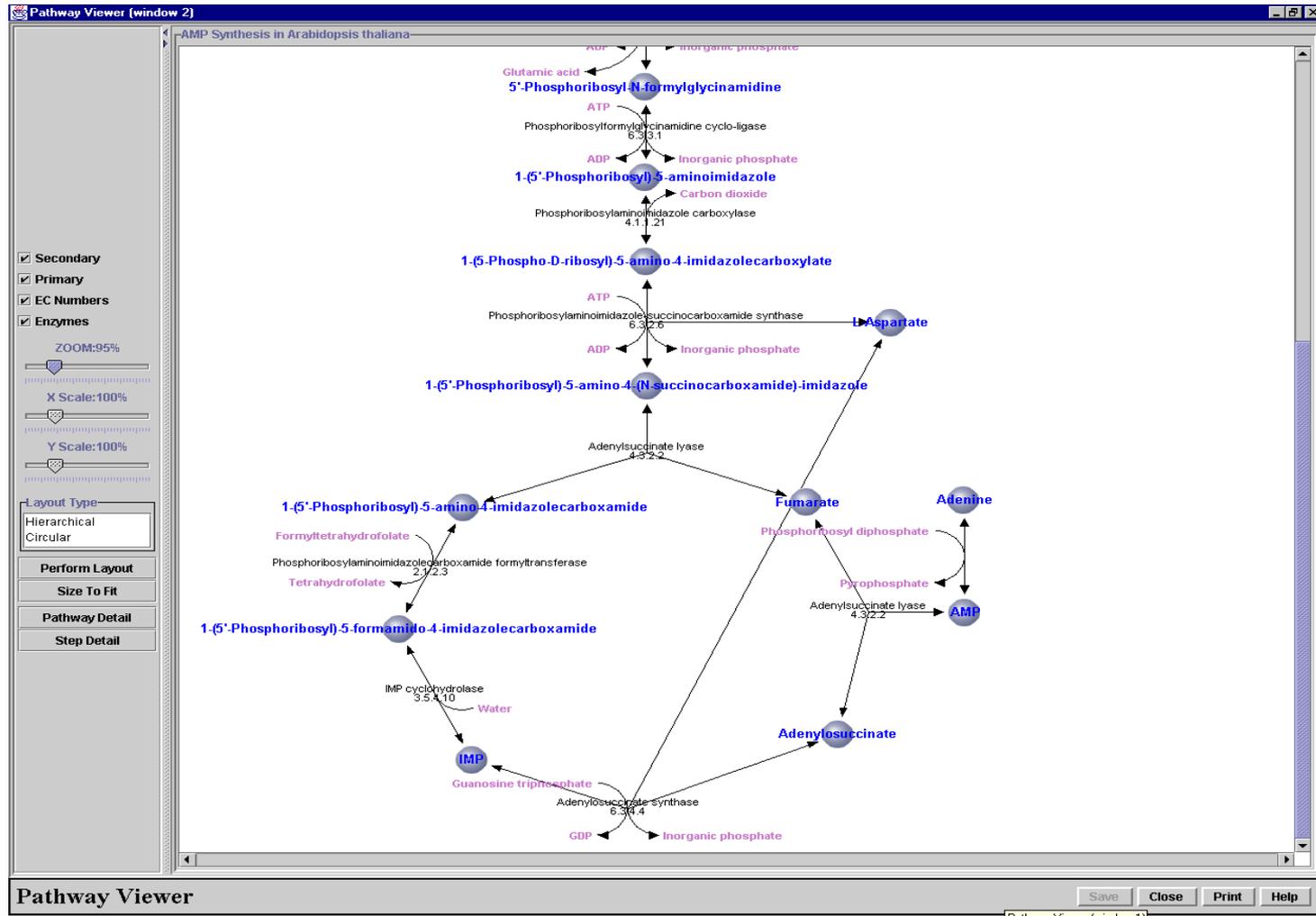
$$2^K[\log(N)]$$

Boolean, K Connections per gene based on linearly separable functions

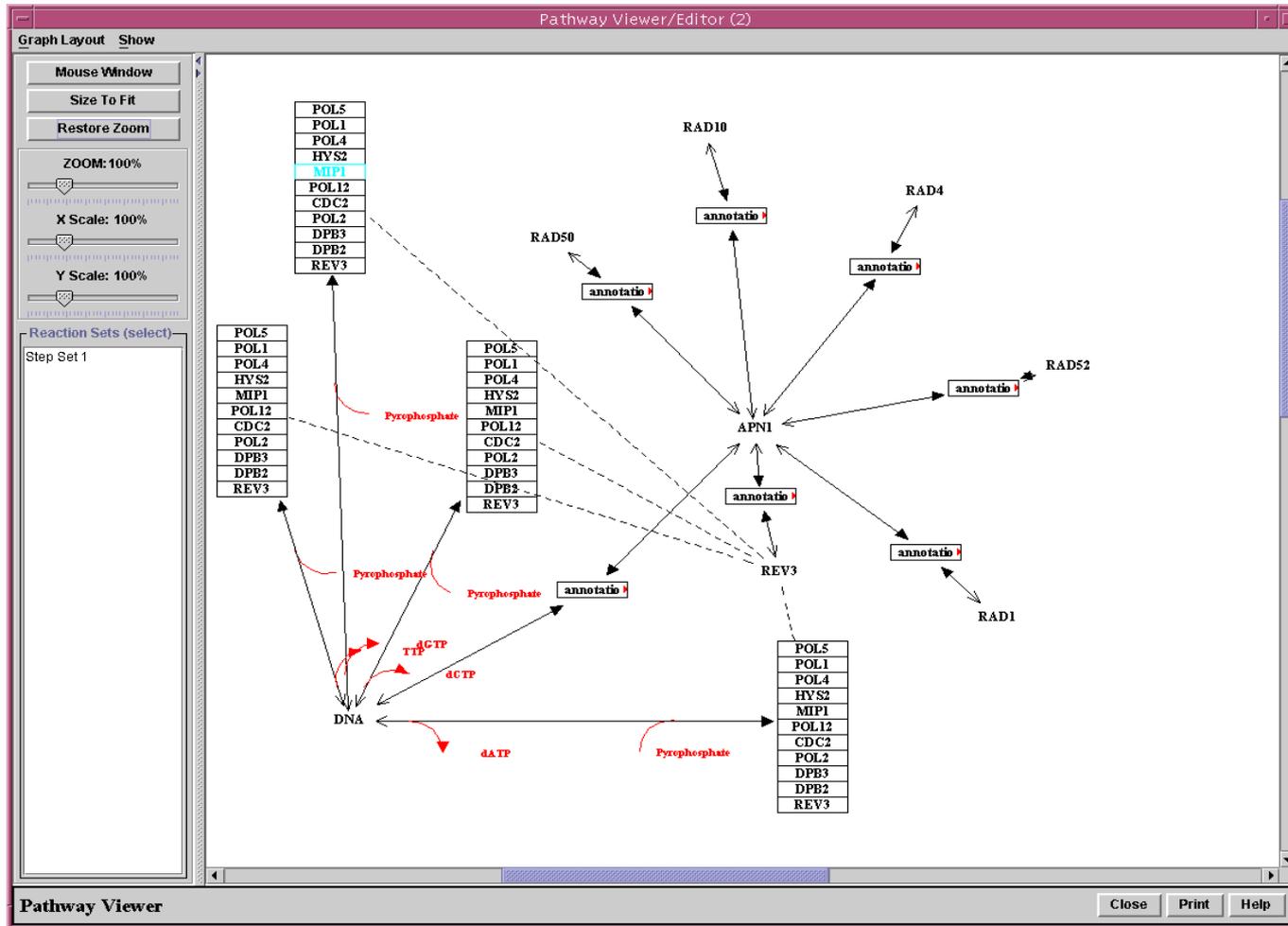
$$K[\log(N)]$$

D'haeseleer, 1997. Data Requirements for Gene Network Inference
<http://www.cs.unm.edu/~patrik/networks/>





Simple network diagram including protein complexes, metabolic reactions and protein-protein interactions



Pathway Viewer/E editor

Graph Layout Show

Mouse Window

Size To Fit

Restore Zoom

ZOOM:46%

X Scale:100%

Y Scale:100%

Reaction Sets (select)

- New PathDB Look
- Alkaloid biosynthesis II - F
- Alkaloid biosynthesis I - F
- Flavonoids, stilbene and
- Xenobiotics metabolism -
- Sulfur metabolism - Refe
- Nitrogen metabolism - R
- Terpenoid biosynthesis -
- Porphyryn and chlorophyll
- Retinol metabolism - Ref
- Folate biosynthesis - Ref
- Biotin metabolism - Refe
- Pantothenate and CoA bi
- Nicotinate and nicotinam
- Vitamin B6 metabolism -
- Riboflavin metabolism - F
- Thiamine metabolism - R
- Reductive carboxylate cyc
- Carbon fixation - Referen
- Methane metabolism - R
- One carbon pool by folate
- C5-Branched dibasic ac

Choose a Color

Swatches HSB RGB

Recent:

Preview

OK Cancel Reset

Record Navigator

Record: 85 of 107 85_GD_21689.class

Pathway Viewer

Close Print Help

Microsoft

Office

Start Meeting Maker® ... Inbox - Outlook Ex... Microsoft Word ... Microsoft PowerP... My Computer Find: Files named ... C:\program file\Pa... C:\WINNT\Syste... Pathway View... 1:05 PM





Pathway Viewer/Editor (1)

Graph Layout Show

Mouse Window

Size To Fit

Restore Zoom

ZOOM: 46%

X Scale: 100%

Y Scale: 100%

Reaction Sets (select)

Glycolysis / Gluconeogen

Step Set 0

Pathway Viewer

Close Print Help

Detailed description: The image shows a screenshot of a software application titled 'Pathway Viewer/Editor (1)'. The main window displays a complex metabolic pathway diagram. The diagram consists of numerous rectangular nodes representing metabolites, interconnected by a dense network of arrows representing enzymatic reactions. The nodes are arranged in a somewhat circular or radial pattern, with many arrows pointing towards a central region. A specific cluster of nodes at the top center is highlighted with red arrows pointing towards it, indicating a point of interest or a specific reaction set. The interface includes a left-hand sidebar with various controls: 'Graph Layout Show', 'Mouse Window', 'Size To Fit', 'Restore Zoom', zoom and scale sliders (set to 46% zoom, 100% X and Y scale), and a 'Reaction Sets (select)' dropdown menu currently showing 'Glycolysis / Gluconeogen' and 'Step Set 0'. At the bottom of the window, there are 'Close', 'Print', and 'Help' buttons.





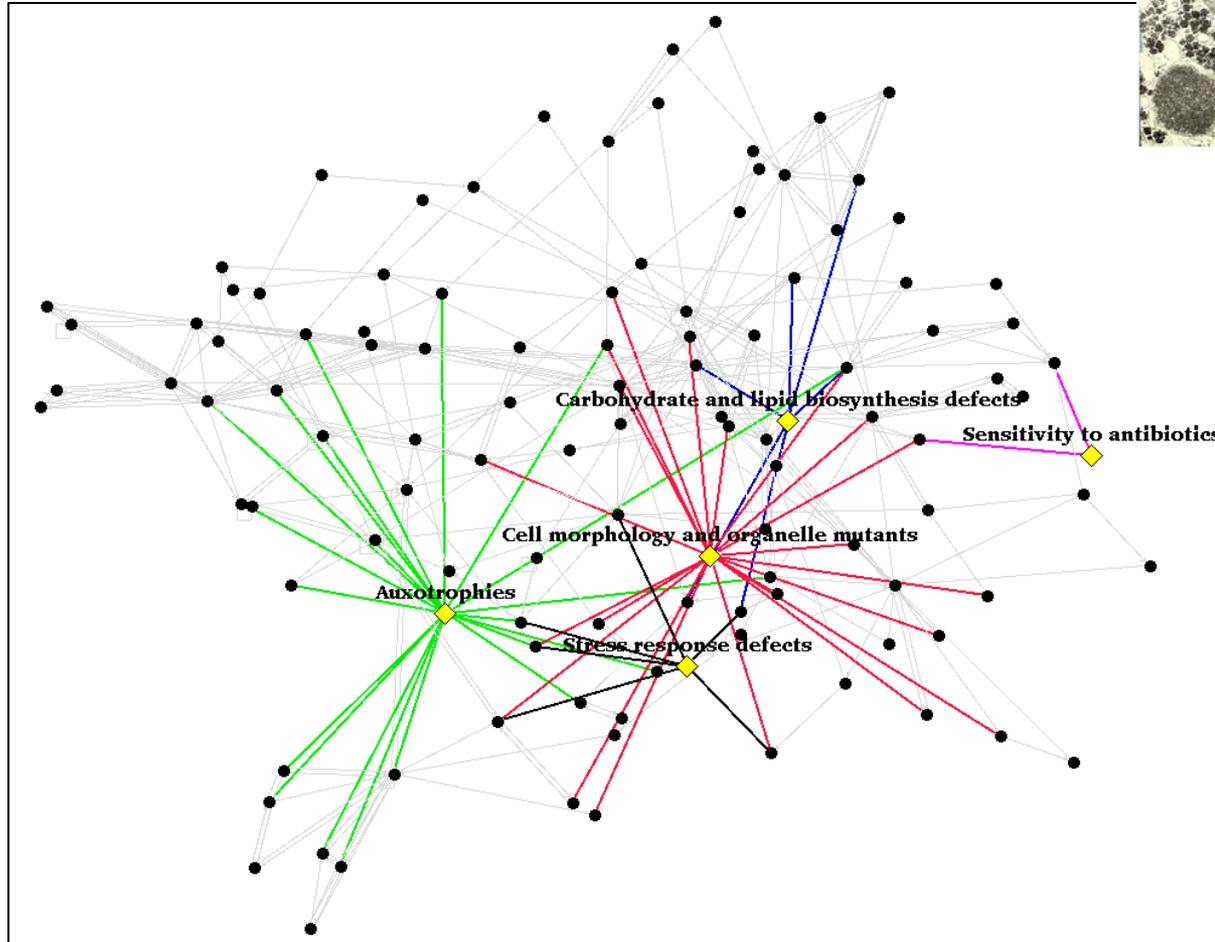
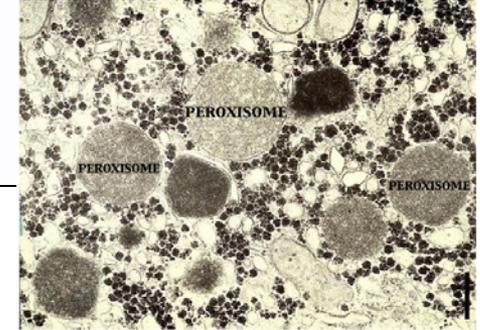
A Need for Integration

- ⌘ The data are located in hundreds of geographically distributed, heterogenous repositories.
- ⌘ Analysis tools (implemented algorithms) are likewise widely dispersed.
- ⌘ Novel algorithms will be developed redundantly by independent PI.
- ⌘ High performance computing resources are limited, but geographically distributed.



Functional information highlighted on a network of yeast peroxisomal proteins

PEROXISOMES AS SEEN IN A HUMAN LIVER SECTION



Integration Taxonomy

Data Oriented



- Data Warehouse



- Federated Databases



- Multiple Databases

Software Oriented

- Web based (hyperlinks)

- Component based

- proprietary
- enterprise



Complex database and software

⌘ **Rigidity**

⌘ **Fragility**

⌘ **Expense**

Software invariably lasts longer than you think (Y2K)



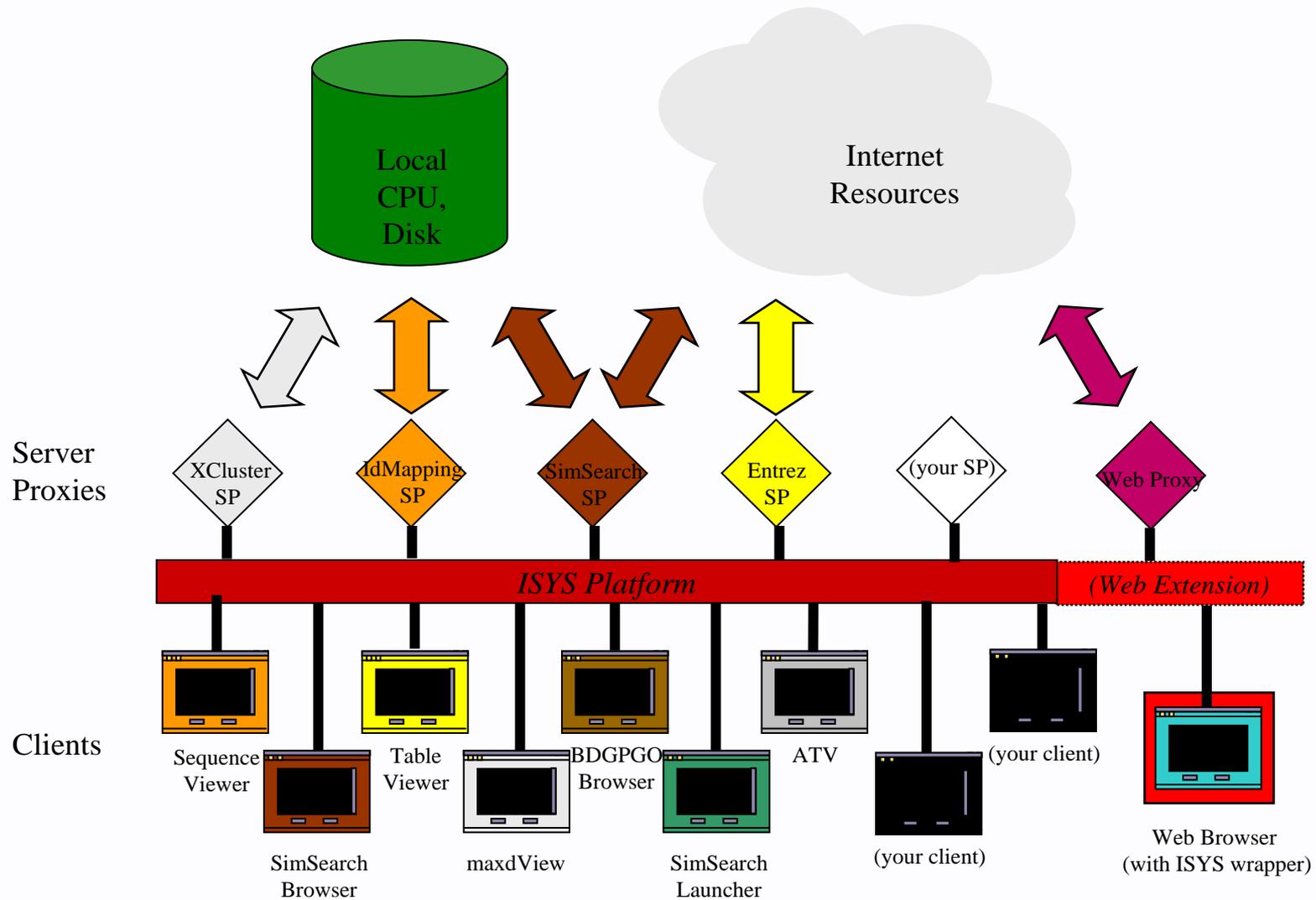
Loose Coupling Architecture

⌘ **Tight packaging of data and application allows context sensitive display of information.**

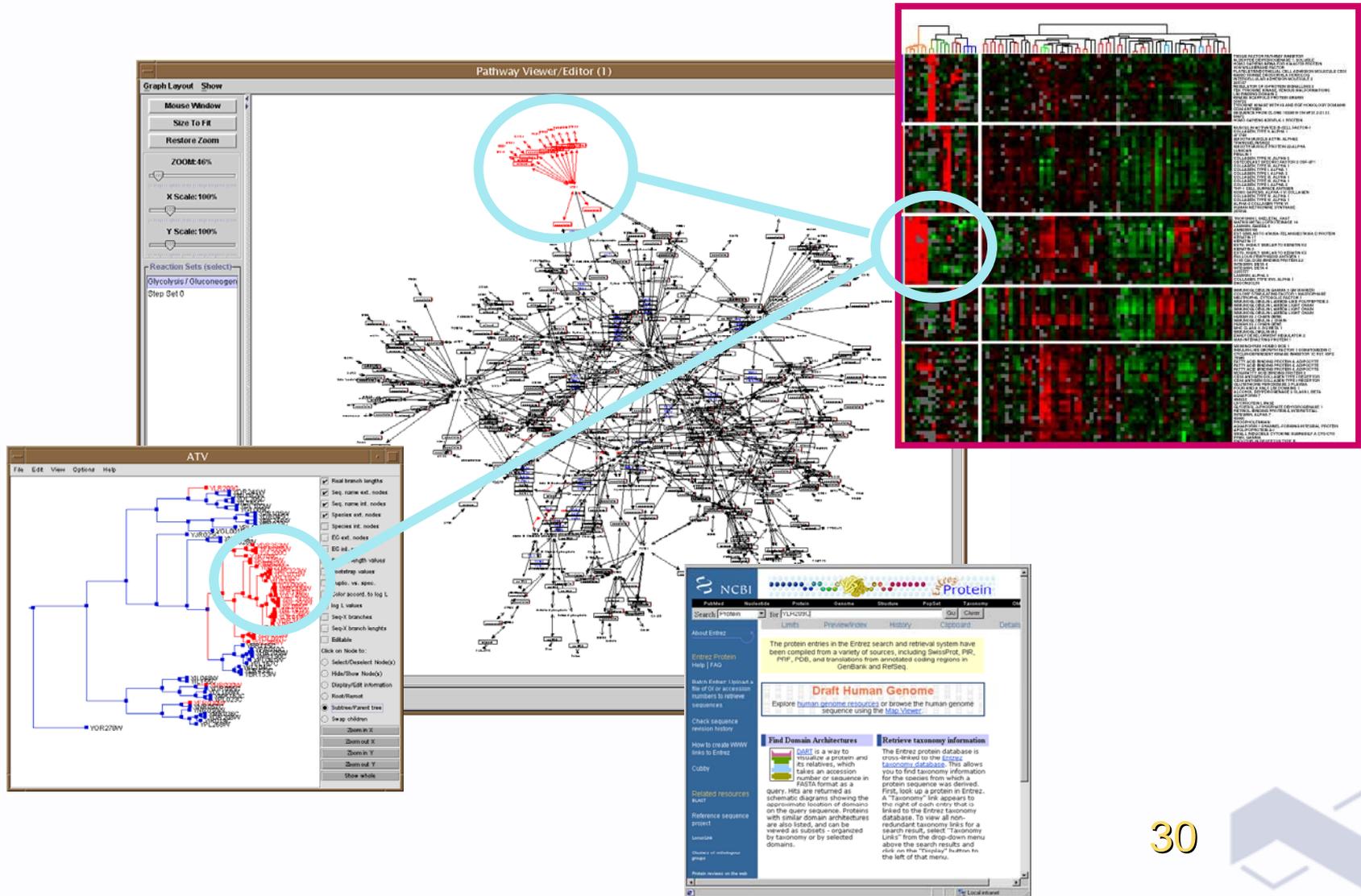
⌘ **Plug and play architecture emphasizes communication between components via fundamental biological concepts.**

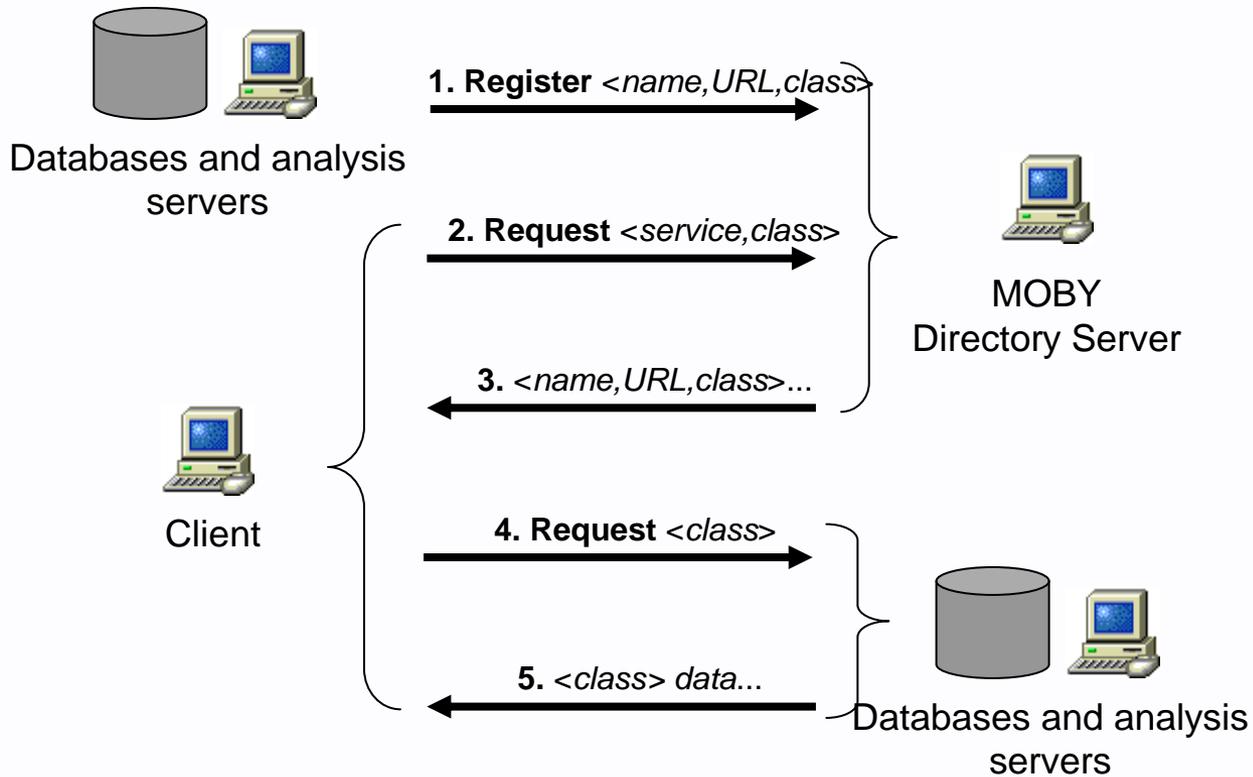
⌘ **Separate components allow parallel and independent scientific advancements.**





Integrated biological data types with an Integration Platform.





Discovery-Driven Data, Models and Computing Resources for Predicting the Phenotype.

	Data and Quality	Models and Lack of Fit	Compute Resources
Ancestry & Environment	Large volume Highly distributed low error	Linear Models Very large LOF	Commodity Clusters
DNA	Huge volume Few repositories very low error	Linear Models with a large LOF Annotation with potentially huge LOF	Commodity Clusters
mRNA	Huge volume Thousands of repositories Multiple Platforms high error	Linear Models with a LOF SVM ?	Large RAM
Proteins	Huge volume Hundreds of repositories Multiple Platforms high error	Protein Structure Predictions Protein Folding Mechanisms via simulation modeling?	Commercial SPM Blue-Genes +
Biochemical Networks	Very little experimental data	Baysian Networks Radiation Transport and other simulation models?	Commercial SPM Blue-Genes +



Enabling Biology to Become a Predictive Science

NCGR

Damian Gessler
Robert Kueffner
Honghui Wan
Jeff Blanchard
Mark Waugh

Sandia National Labs

Grant Hefflefinger
George Davidson

Joint Genome Institute

Dan Rokshar

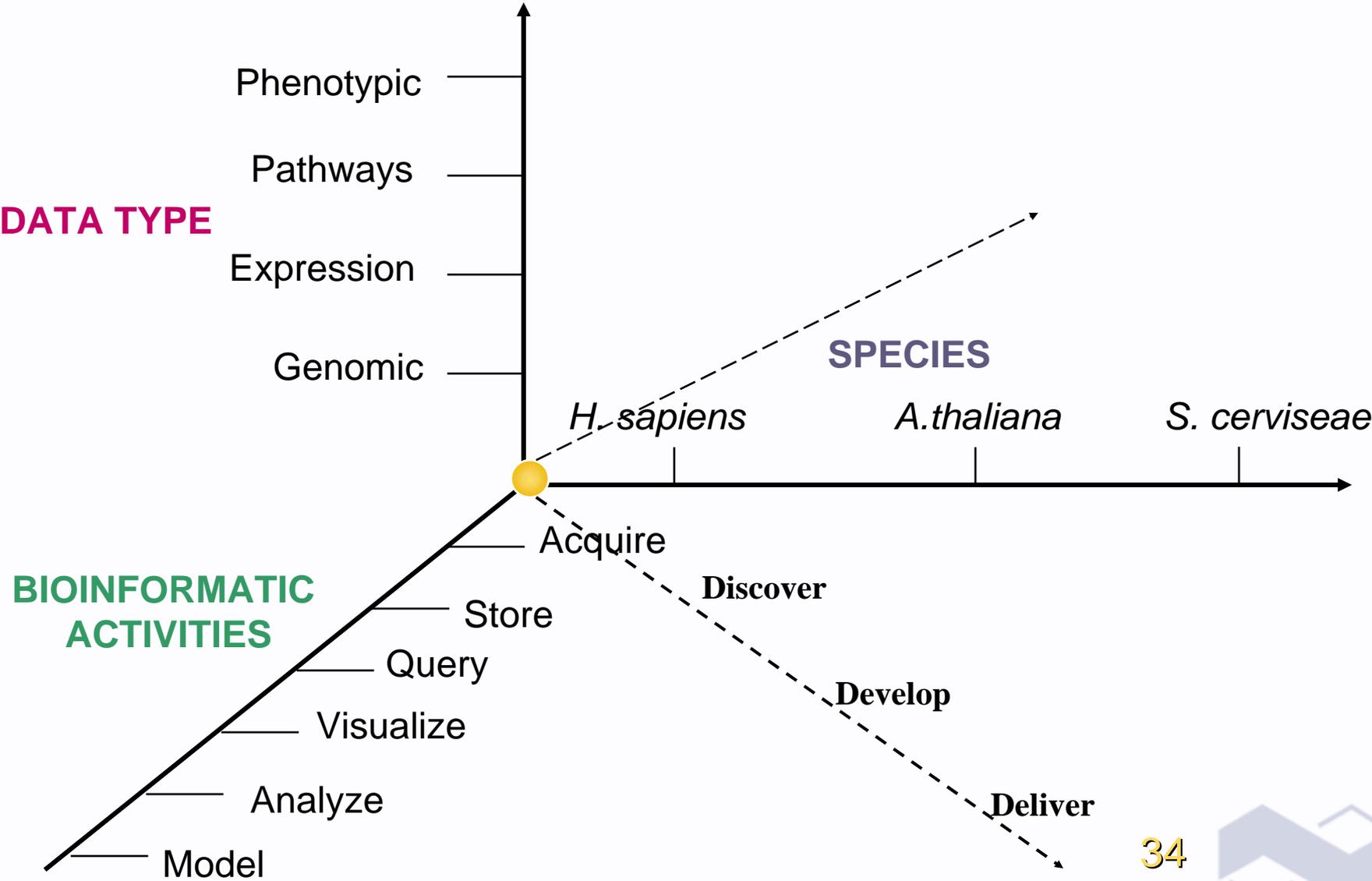
Oak Ridge National Labs

Jerry Tuskan
Frank Larimer

Los Alamos National Labs

Mark Wall

Bioinformatics



DISCOVERY-DRIVEN BIOLOGICAL RESEARCH

