

# A Life Scientist's Road to Interoperability of Data and Tools

*Information Science Standards to Enable  
Biomedical Research, November 4th 2003*

**Bruno Sobral (sobral@vt.edu)**

**Virginia Bioinformatics Institute**

VIRGINIA  
BIOINFORMATICS  
INSTITUTE  
AT VIRGINIA TECH



**<http://www.vbi.vt.edu>**

# NIH Roadmap

- **“The scale and complexity of today's biomedical research problems increasingly demand that scientists move beyond the confines of their own disciplines and explore new organizational models for team science....Many sciences will still continue to pursue individual research projects, but they too will be encouraged to make changes in the way they approach the scientific enterprise.”**
- 
- **“This demands that we break down barriers among disciplines, as well as among our own institutes and centers. We need to challenge ourselves to find even more innovative and effective ways of doing biomedical research and converting that into cures.”**

# NSF's

# Cyberinfrastructure



## Cyberinfrastructure Promise

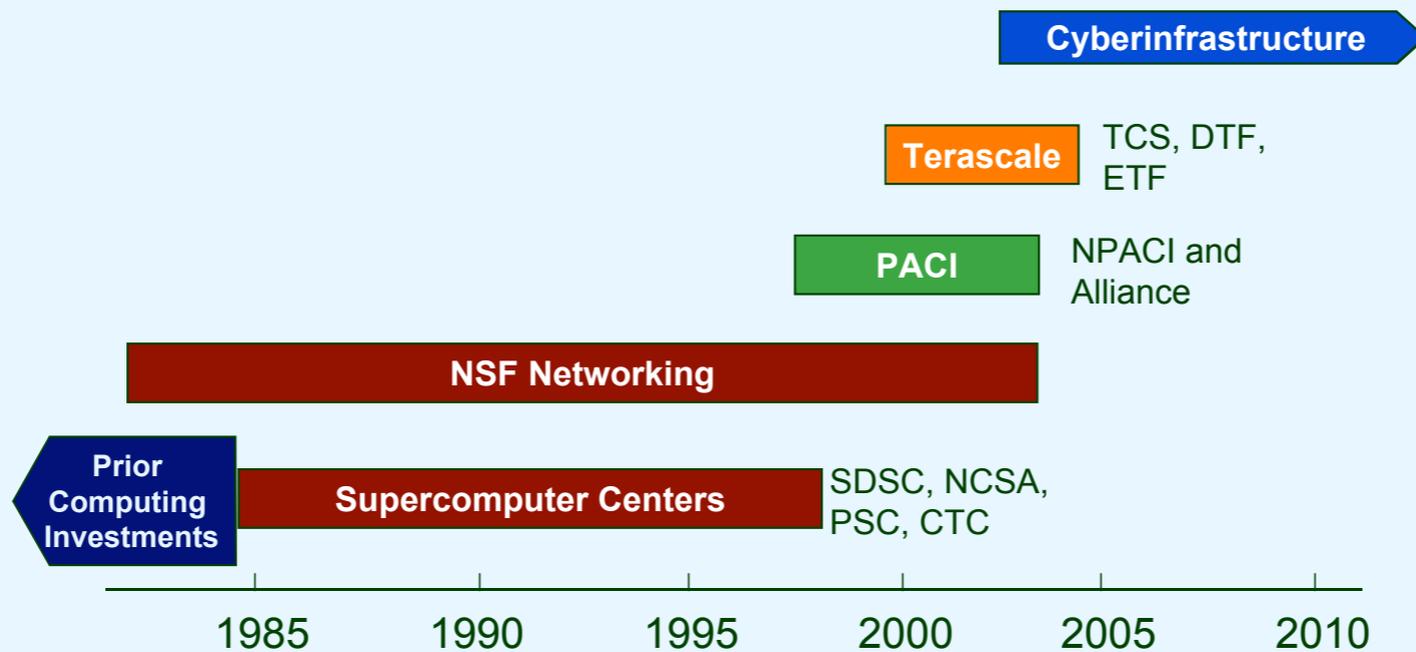
---

- Ubiquitous, digital knowledge environments that are both interactive and functionally complete..... (Atkins report)
- revolutionize the processes of discovery, learning and innovation across the science and engineering frontier.

# CyberInfrastructure

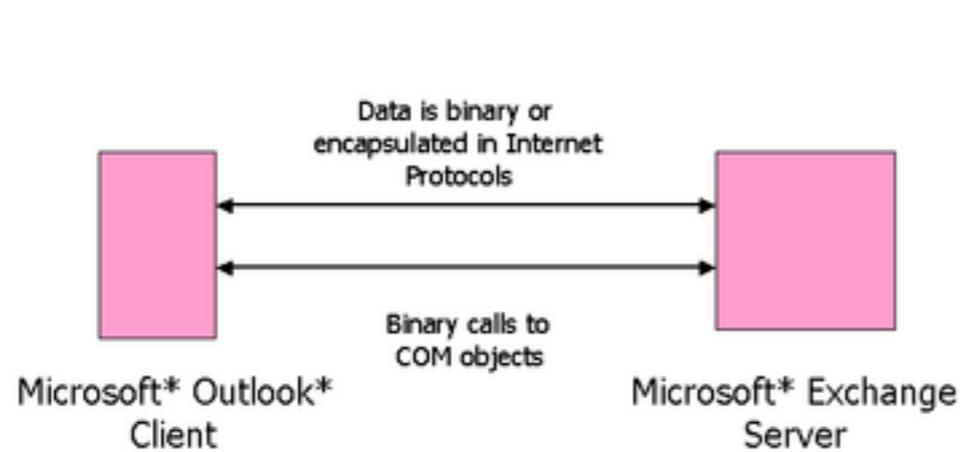


## Evolution of the Computational Infrastructure

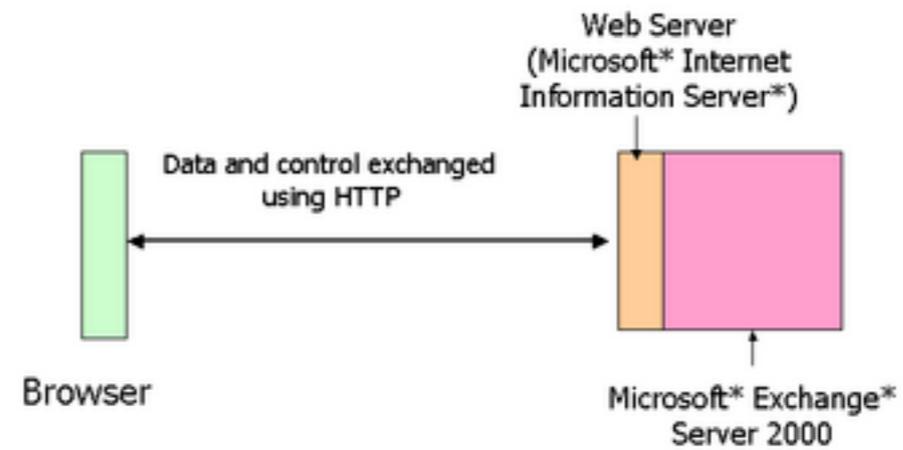


2

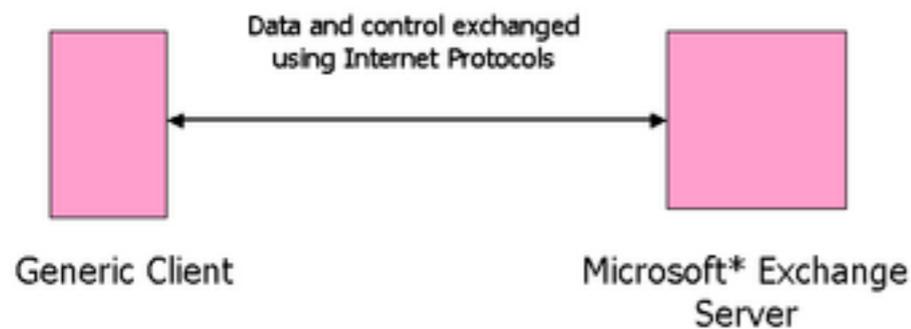
# From Client-Server to Web Services



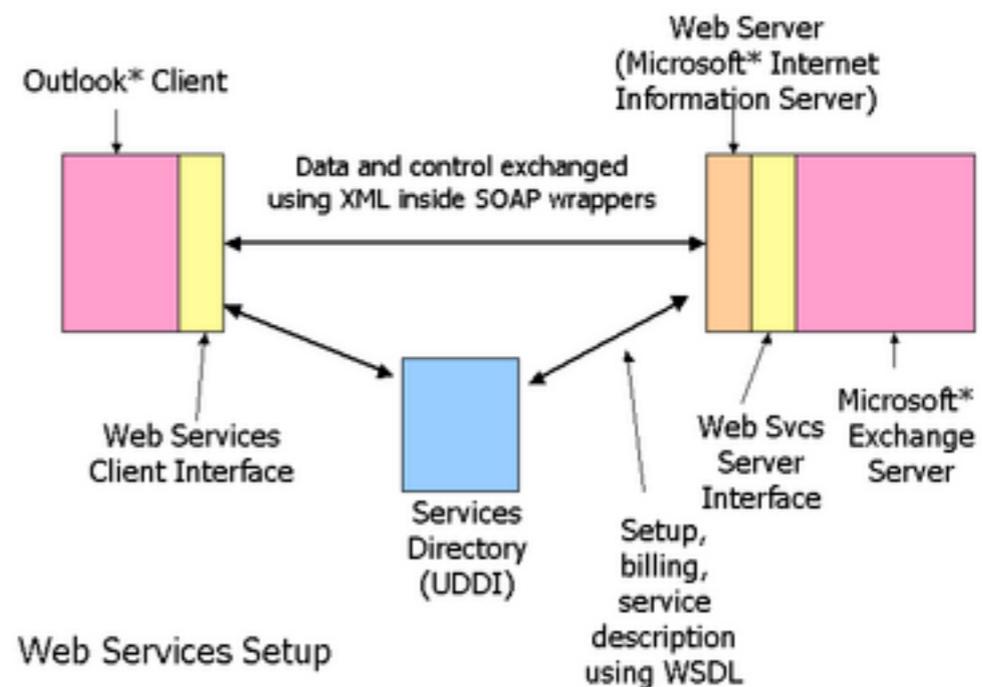
Example of Traditional Client/Server Setup. Not a Web Service



Client/Server Browser Interface Not a Web Service



Client/Server setup using Internet-standard client. Not a Web service



Web Services Setup

# Definition

- Web services are *loosely coupled, self-describing* services that are *accessed programmatically* across a distributed network, and exchange data using vendor, platform, and *language-neutral protocols*
- Fundamentally enabled by *agreement on standards* across a broad group of hardware and software organizations

# Standards of Web Services Stack



# Supporting Collaboration

- **Collaboration - cooperation to achieve goal(s)**
  - Much more than static exchange of e-mail or spreadsheets
    - Interactive, live, real-time (as required)
- **Non-traditional IT architecture - not internally focused**
  - Must support/facilitate interactions
  - Collaboration is rapidly becoming the rule rather than the exception
- **Web services allow support for collaboration at the process level**

# Initial Phases of Web Services

- Integration/Interoperation
- Collaboration
- Innovation

# Integration/ Interoperation

- Building wrappers around legacy applications and systems
- Fast cycles of learning
  - Deploy early and often approach
- Increase in shared information across collaborators
- Reach limits with immature standards and unprepared IT architectures

# Collaboration

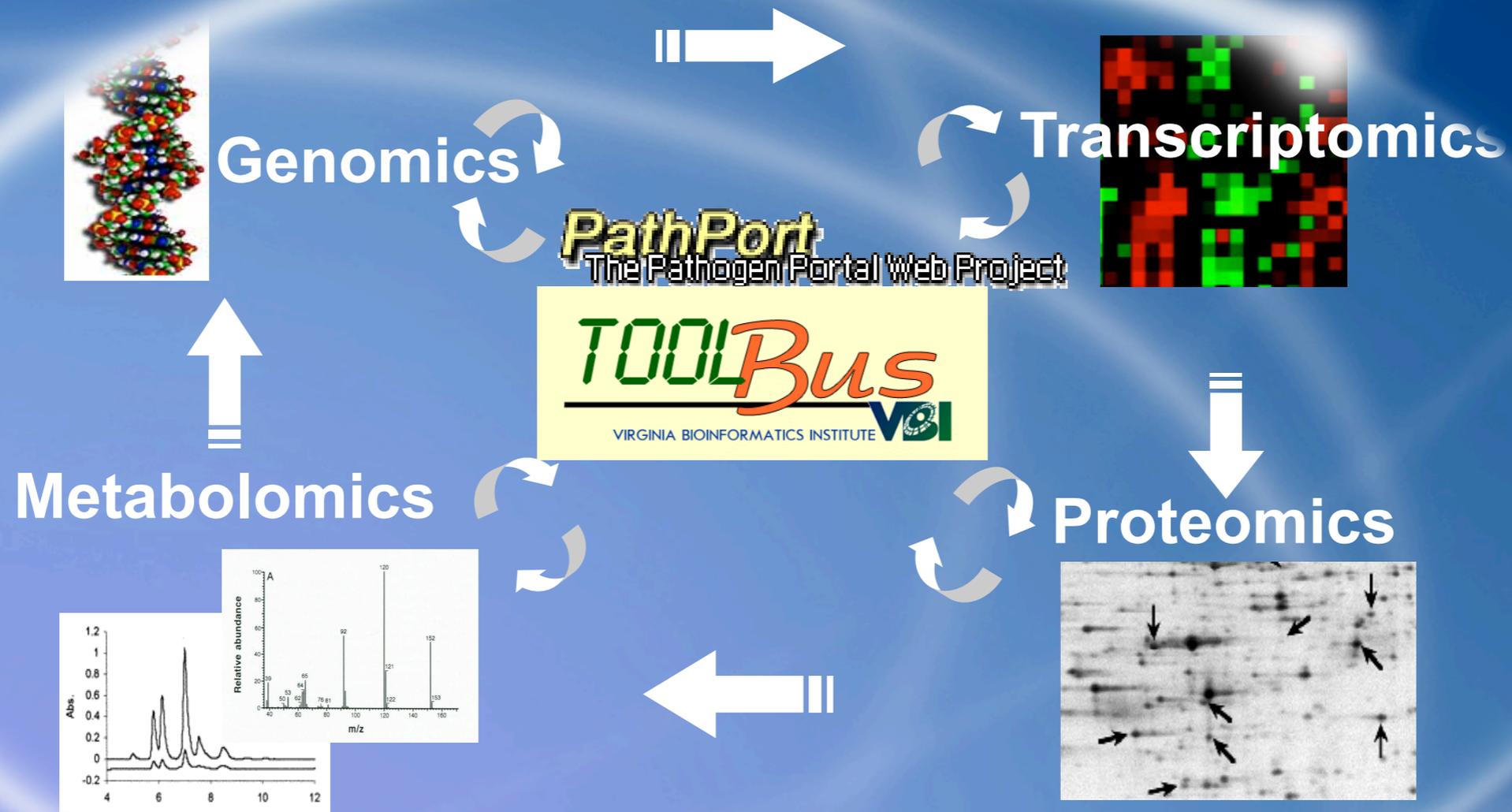
- **Web services reduce the level of human intervention in collaborative process**
  - **Increased experimentation outside firewalls**
  - **Increased interactions with collaborators and partners**
  - **Closer partners share standards and implement them to drive open architecture**
  - **“External” partners start to share and collaborate further driving the chain**

- **Goal of collaboration is to establish/maintain/strengthen connections**
  - People to people
  - People to content
  - People to applications
  - Applications to applications to content to applications
- **Main driver: improving connections**
- **Importance of understanding and analyzing social networks**

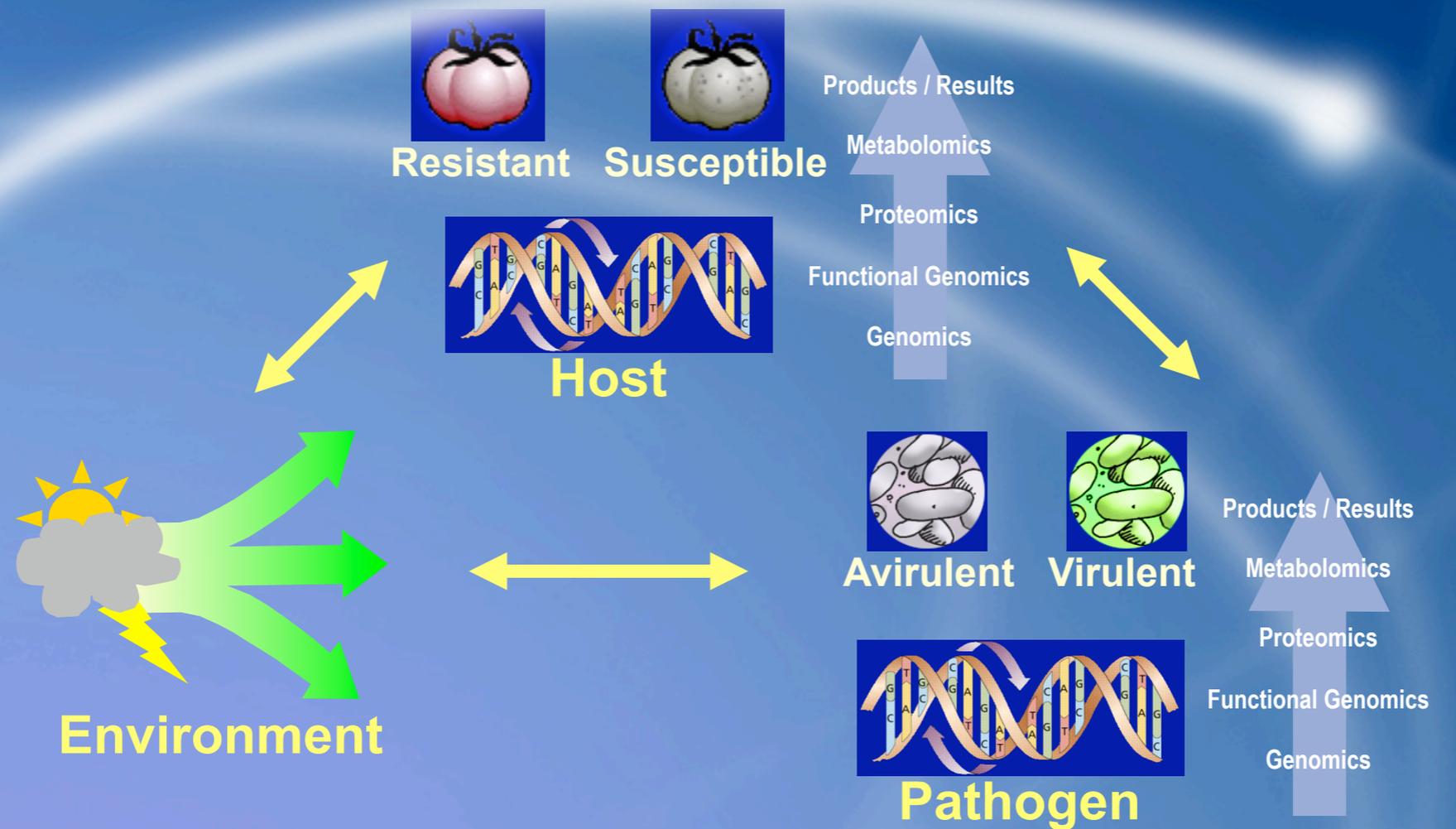
# Innovation

- **Lessons from integration and collaboration applied to drive new processes and models**
  - New, distributed web service processes and applications drive change
- **Redefinition of how research is conducted across boundaries of organization**
  - Exposing specific operational elements for dynamic linking to processes of partners
  - Organizations operating as part of truly interconnected ecosystem

# Systems Interoperation

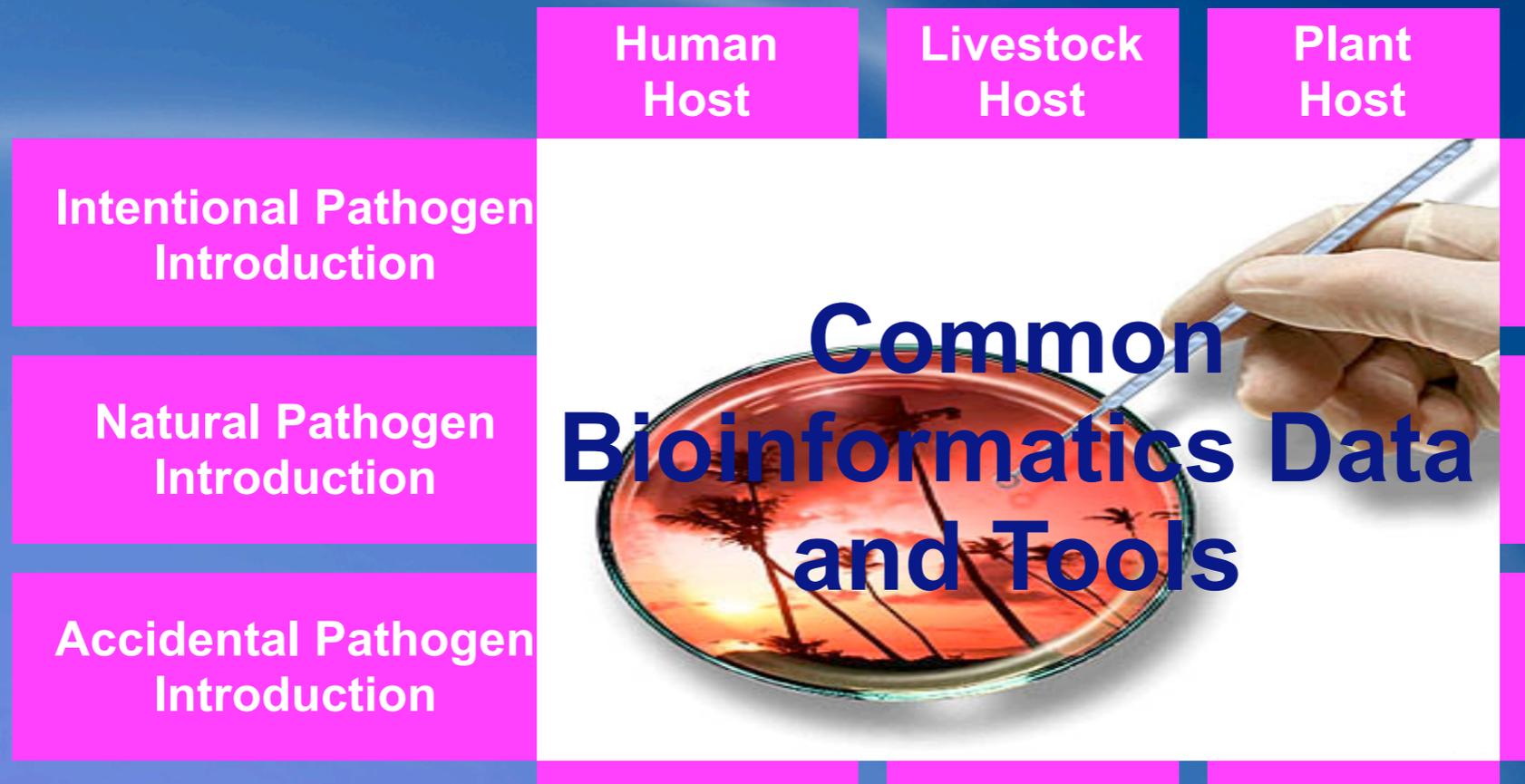


# PathoSystems Biology



Reverse Engineering the “Disease Triangle”

# Global View of Infectious Diseases



**Data/Tool  
Interoperation**

# Role of IT

## BIOTERRORISM

### Information Technology Strategy Could Strengthen Federal Agencies' Abilities to Respond to Public Health Emergencies

GAO Report, May 2003, 03-139

#### What GAO Found

The six key federal agencies involved in bioterrorism preparedness and response identified about 70 planned and operational information systems in several IT categories associated with supporting a public health emergency. These encompass detection (systems that collect and identify potential biological agents from environmental samples), surveillance (systems that facilitate ongoing data collection, analysis, and interpretation of disease-related data), communications (systems that facilitate the secure and timely delivery of information to the relevant responders and decision makers), and supporting technologies (tools or systems that provide information for the other categories of systems)—see table below.

### Summary of the Systems Inventory by Agency

IT Categories	HHS	Defense	Energy	Agriculture	EPA	VA	Total
Detection	0	4	6	0	0	0	10
Surveillance	18	7	2	6	0	1	34
Communications	5	2	0	3	0	0	10
Supporting Tech	5	1	6	1	5	0	18
Total	28	14	14	10	5	1	72

Source: GAO.

# Required Components To Achieve Synthesis

- Large high-quality data sets (DNA, mRNA, proteins, metabolites, moving from molecular to higher levels of organization)
- Integrated wet chemistry and *in silico* experimentation, modeling, simulation, and theory development - with goals of prediction and mechanistic understanding
- IT infrastructure (cyberinfrastructure) - software, hardware, bandwidth, personnel

# PathPort - The Pathogen Portal Project

- Facilitate knowledge extraction from diverse data types
  - Interoperable access to diverse (molecular) data types
  - Interoperable access to analysis tools
  - Multiple domain-specific viewers
- Easily extensible - planned from molecules to higher levels of organization
- Ability to save and load work sessions
- Feedback loop from viewers to tools
- Allow association of different data models

# PathPort Is Built on Open Standards

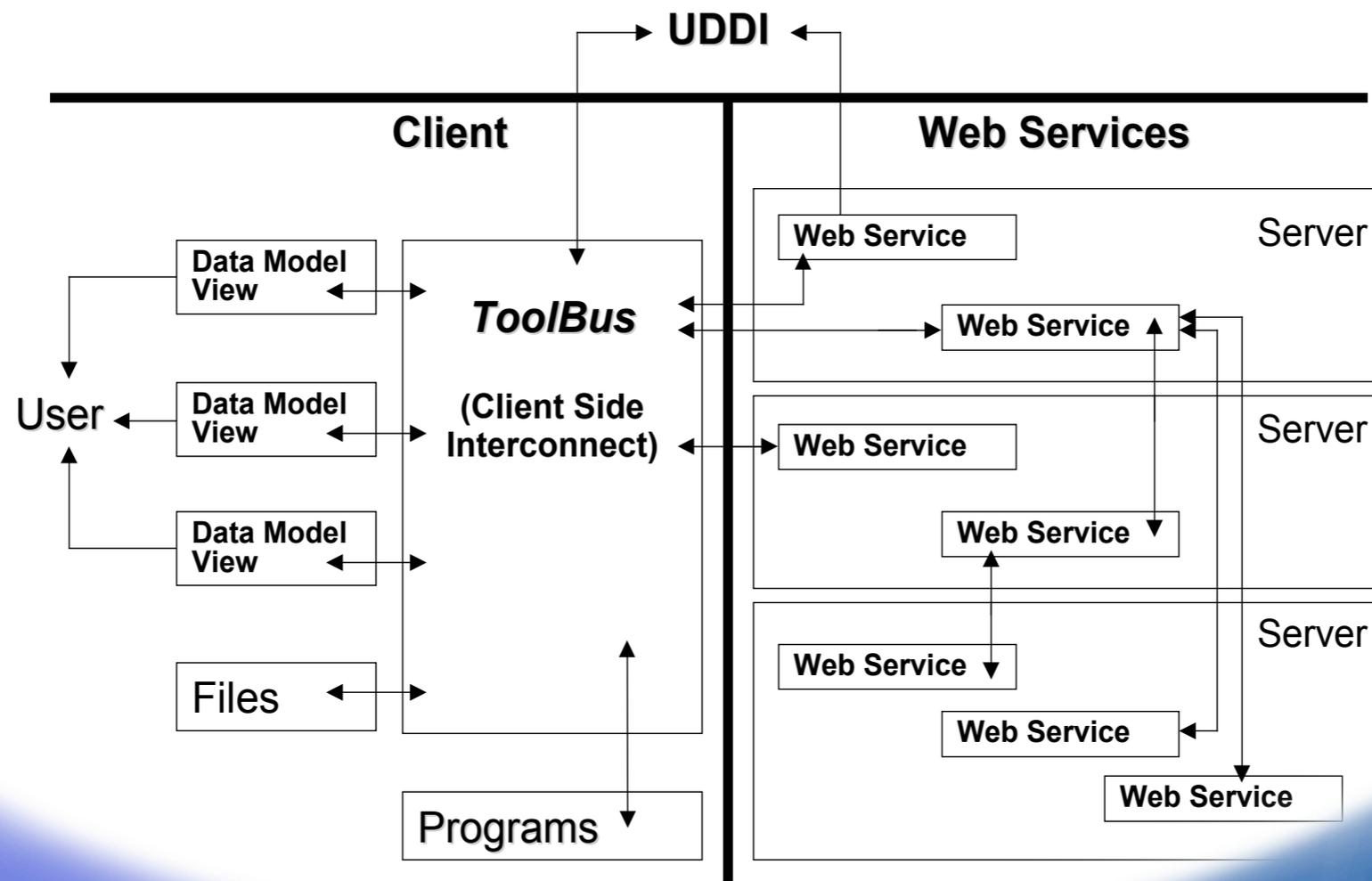
- Common vocabulary: Gene Ontology (GO)
- Transport format: XML
- Data definition language: XSD
- Wire protocol: SOAP
- Service definition language: WSDL
- Service registry: UDDI [OGSA]

# PathPort XML

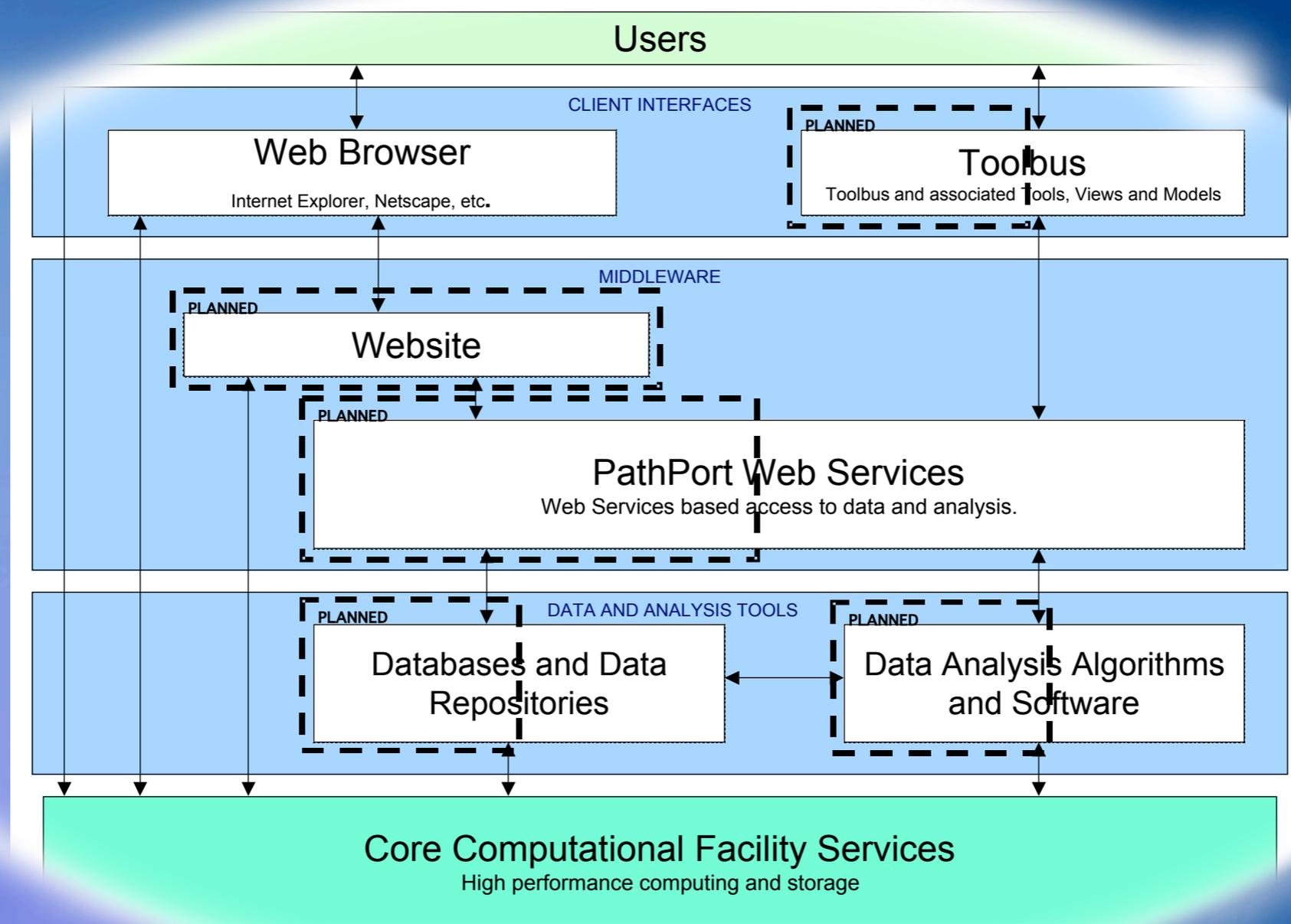
- Utilizes established, open community standards
  - *DAS-ML, BSML, MSA-ML (DNA) - Year 1*
  - *MAGE-ML (mRNA profiling) - Year 2*
  - *PEDRo (protein profiling) - Year 3*
  - *SBML (molecular models) - Year 3*
  - *CellML (cellular levels, including metabolism and signal transduction) - Year 4*
  - *AnatML (organ level) - Year 4*
  - *FieldML (spatially and temporally varying field information using finite elements) - Year 5*

# PathPort Architecture

## Data Integration: *PathPort* Architecture



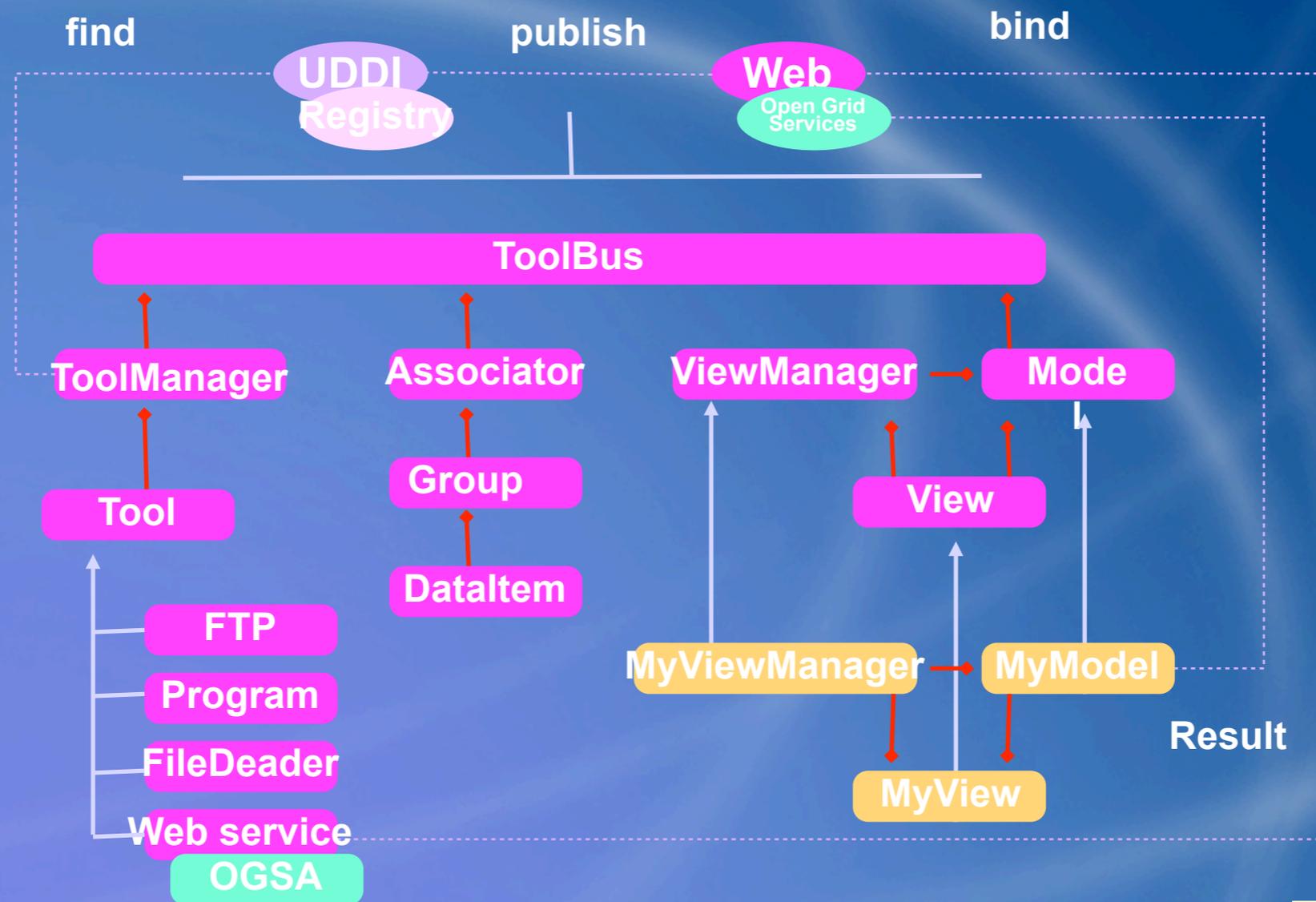
# CyberInfrastructure to Support Analysis



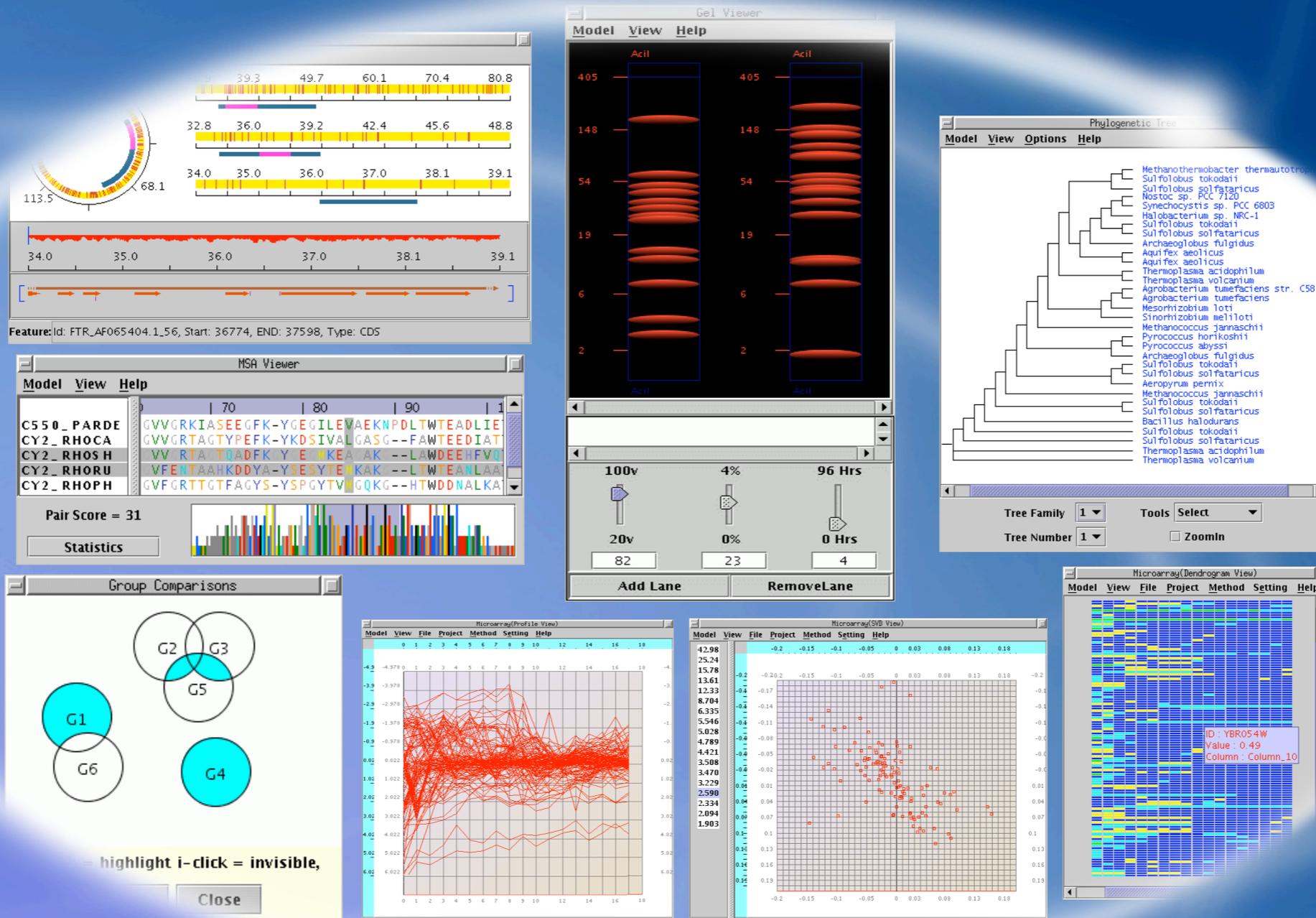
# ToolBus

- A client-side interconnect with the following goals:
  - Platform independent
  - Easily extensible
  - Allow user-defined associations
  - Easy to use

# ToolBus Architecture

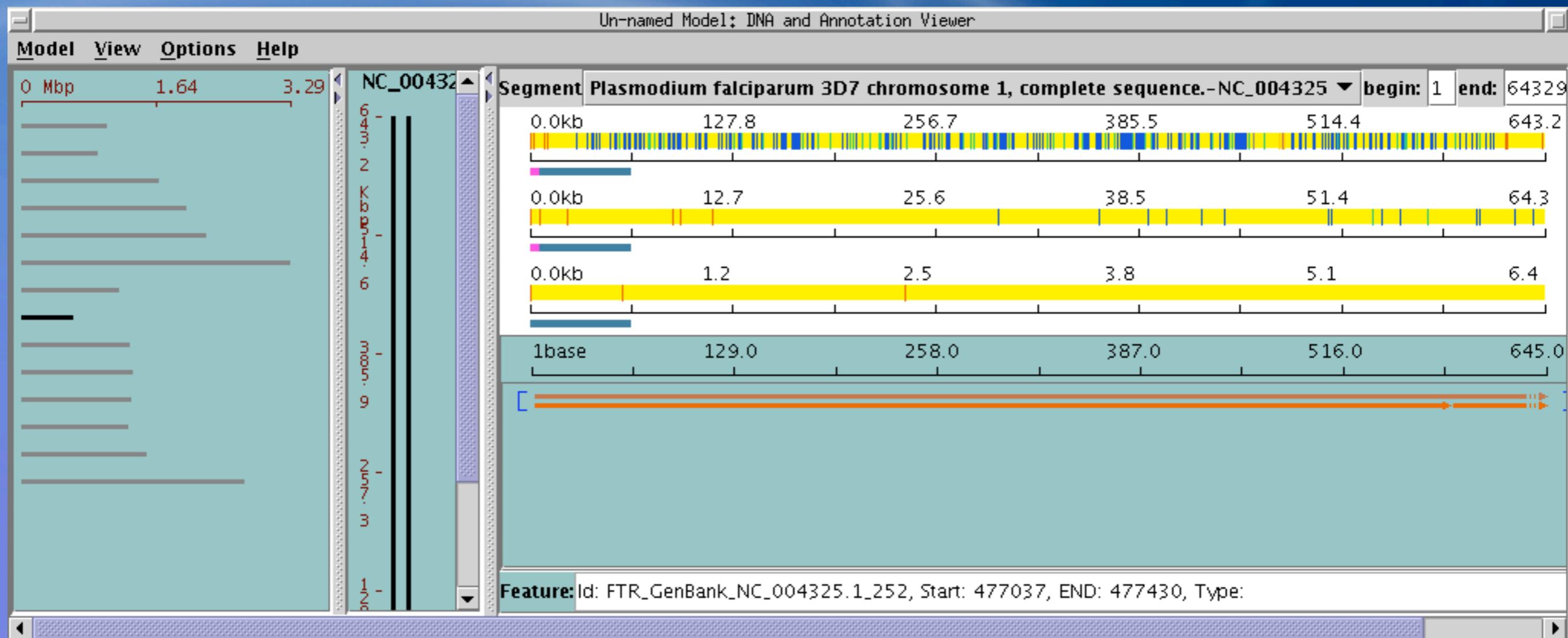


# An Interoperable Work Environment for Discovery



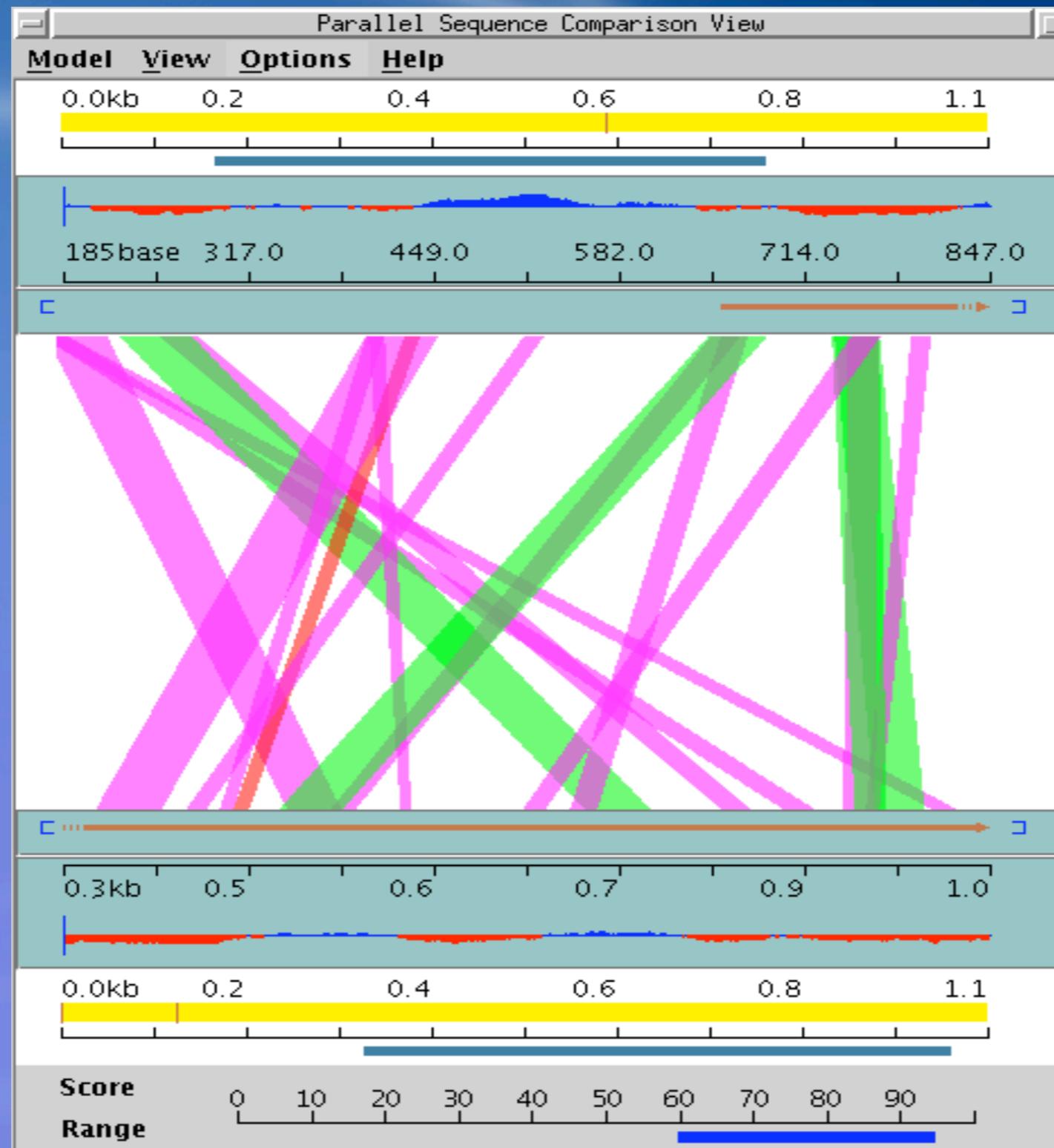
# Annotation

## Viewing/Editing





# Comparative Genomics



# Analysis of Transcriptional Profiles

Microarray(Table View)

Model View File Data Method Setting Help

Filter

- Max/Min >=
- Max - Min >=
- Percent Allowed Missi
- Selected Row(s)
- Selected Column(s)
- Column\_0
- Column\_0

Transform

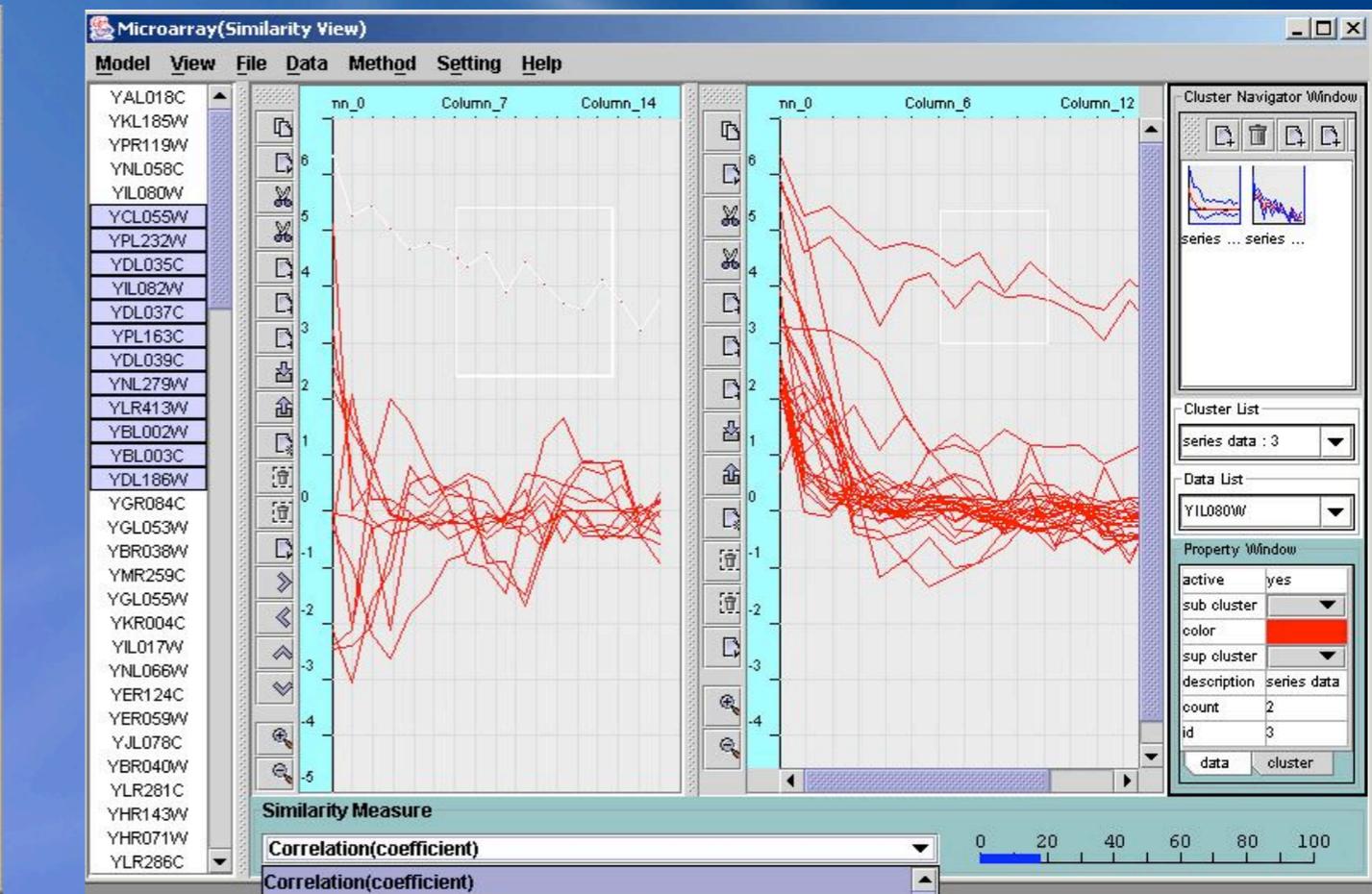
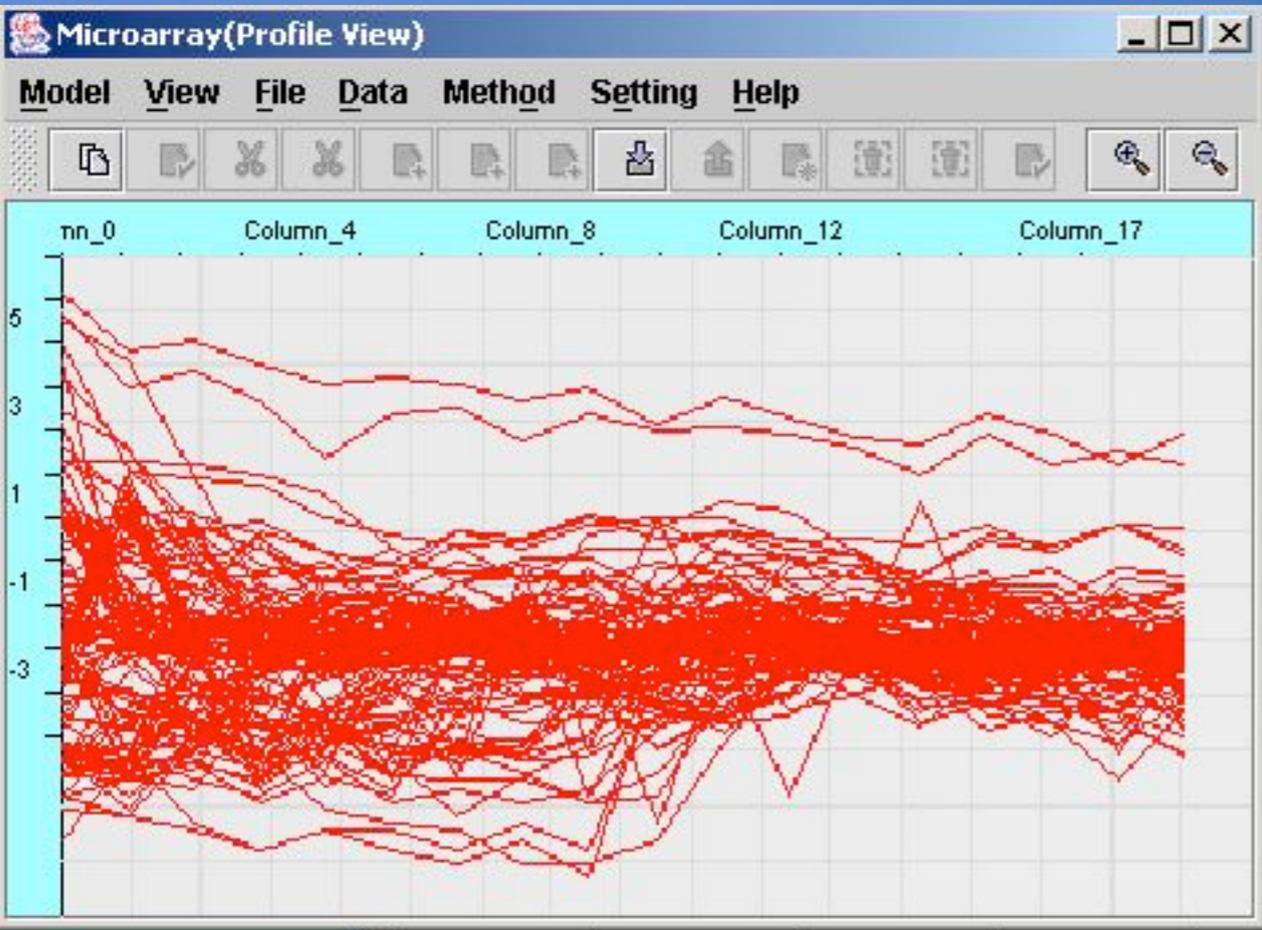
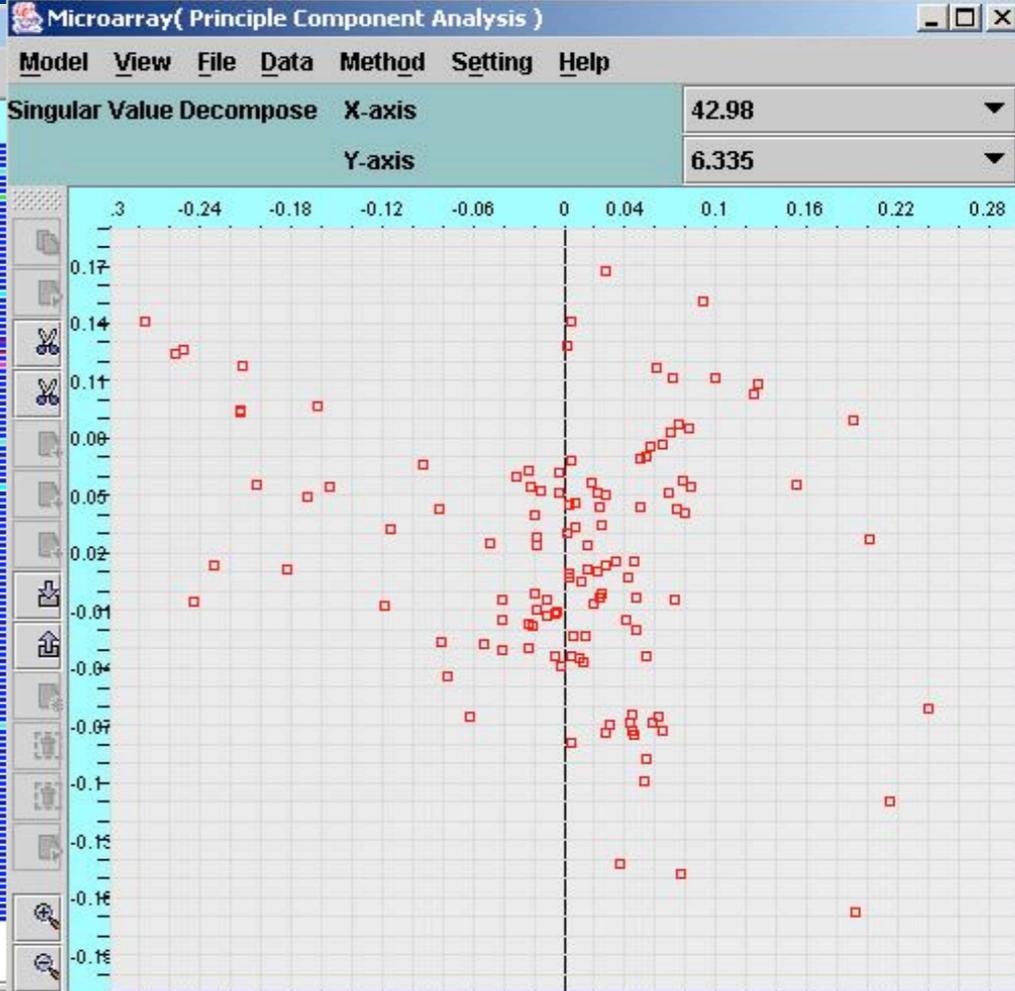
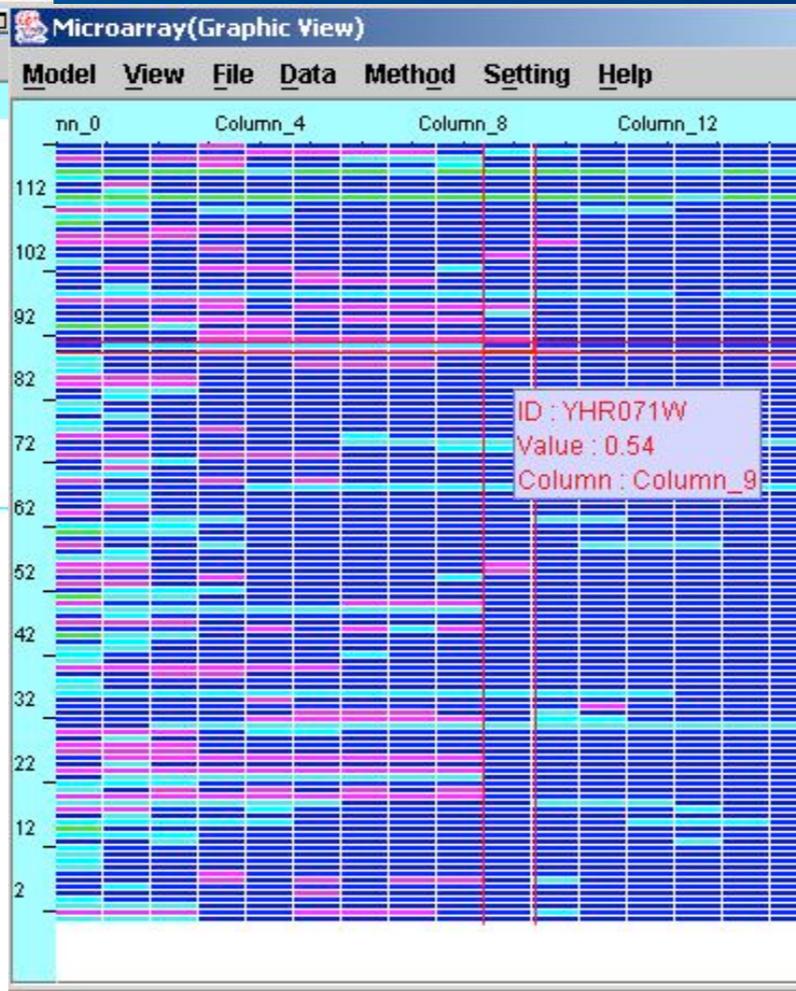
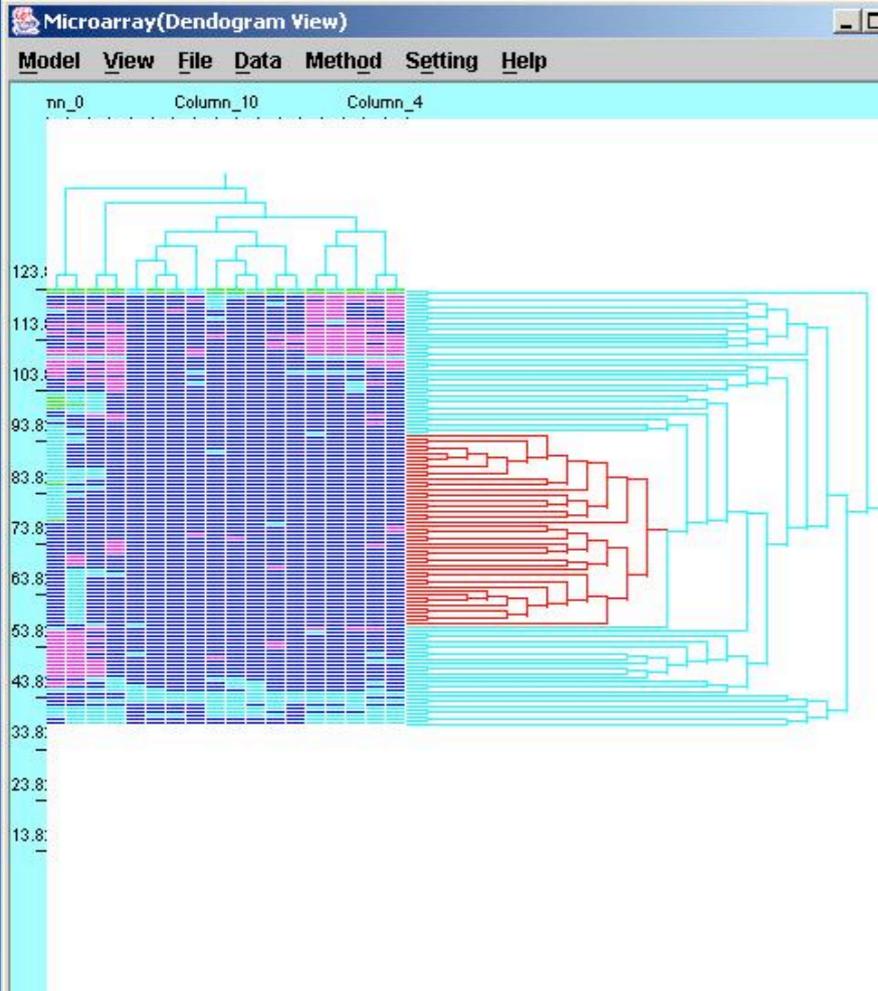
- Transpose DataSet
- Handle Missing Value (Average):  by row  by column
- Clip Value to Min =
- Clip Value to Max =
- Normalize to Mean =
- Normalize to Variance =
- x -> Log10(x)

Method Setting Help

- SOM clustering Ctrl-O
- Agglomerative Hierarchical Clustering Ctrl-H
- K-mean clustering Ctrl-N
- SVM data classification Ctrl-F
- Linear Discriminant Analysis Ctrl-I
- K-nearest neighbors Analysis Ctrl-G
- Principle Component Analysis Ctrl-J
- Statistic Analysis (T-test or F-test) Ctrl-Y

	Column_0	Column_1	Column_2	Column_3	Column_4
2	2.13	0.19	0.99	0.62	
4	-2.56	-2.32	-0.32	0.68	
5	-1.89	-2.0	-0.43	0.4	
6	-2.06	-2.56	-2.84	-2.18	
6	2.98	0.16	-0.03	0.06	
6	-3.18	-3.47	-3.84	-3.47	
	3.1	2.97	2.83	2.19	
	1.04	1.45	0.46	0.69	
	0.91	-2.0	-0.03	-2.18	
4	-2.12	-3.32	-3.84	-3.47	
YMR011W	1.1	1.84	2.12	1.65	1.1
YBR010W	-2.12	-2.25	-1.64	-0.3	1.2
YDR193W	-0.15	2.34	0.2	0.18	0.03
YMR305C	2.49	1.41	1.53	1.24	1.44
YMR232W	3.91	0.91	0.21	0.21	0.0
YGL032C	3.53	2.6	1.74	0.5	-1.18
YBR230C	-0.1	0.67	1.02	0.45	-0.06
YCLX07W	2.28	0.7	-0.27	-0.32	-0.62
YOR343C	2.45	0.7	0.54	0.33	0.04
YDR124W	2.16	0.03	-0.23	-0.17	-0.45
YJR153W	2.21	0.36	-0.03	0.53	0.04
YBL065W	0.19	0.45	0.26	-2.74	0.31
YDR055W	-1.15	-0.86	-1.6	-2.47	-1.51
YDL024C	0.0	2.35	0.0	-0.23	-0.12
YNL196C	0.08	0.28	-0.51	-0.45	-0.14
YPL156C	2.34	-0.3	-1.36	-0.58	-0.49
YIL079C	2.3	1.6	1.26	0.85	0.5
YPL158C	-2.25	-2.47	-2.74	-1.09	-1.64
YPL088W	2.06	2.28	1.62	0.82	-0.09

Run Undo Clear ExportData Reset



#### Cluster Navigator Window

series ... series ...

Cluster List  
series data : 3

Data List  
YIL080W

Property Window

- active: yes
- sub cluster: [dropdown]
- color: [red]
- sup cluster: [dropdown]
- description: series data
- count: 2
- id: 3

[data] [cluster]



# Tabular Data Viewer

The screenshot displays the 'Microarray(Table View)' application window. The main window contains a menu bar (Model, View, File, Data, Method, Setting, Help) and a data table with columns labeled ID, Column\_0, Column\_1, Column\_2, Column\_3, and Column\_4. The data rows include various gene identifiers such as YAL018C, YKL185W, YPR119W, YNL058C, YIL080W, YCL055W, YPL232W, YDL035C, YIL082W, YDL037C, YPL163C, YDL039C, YNL279W, YLR413W, YBL002W, YBL003C, YDL186W, YGR084C, YGL053W, YBR038W, YMR259C, YGL055W, YKR004C, YIL017W, YNL066W, YER124C, YER059W, YJL078C, and YBR040W. The values in the table are color-coded based on their range. A 'Table Config Panel' dialog box is open, titled 'Color Coding For Table Value', showing five color-coded ranges: magenta for (-4.3, -1.6), blue for (-1.6, 1.0), cyan for (1.0, 3.66), green for (3.66, 6.32), and red for missing values. The dialog has 'Ok', 'Cancel', and 'Set Default' buttons. On the left side of the main window, there are 'Filter' and 'Transform' sections. The 'Filter' section includes checkboxes for 'Max/Min >=', 'Max - Min >=', 'Percent Allowed Missing Value', 'Selected Row(s)', and 'Selected Column(s)', along with dropdown menus for 'Column\_0'. The 'Transform' section includes checkboxes for 'Transpose DataSet', 'Clip Value to Min =', 'Clip Value to Max =', 'Normalize to Mean =', 'Normalize to Variance =', and 'x-> Log10(x)', along with radio buttons for 'Handle Missing Value (Average): by row' and 'by column'. At the bottom of the main window are buttons for 'Run', 'Undo', 'Clear', 'Export Data', and 'Reset'.

ID	Column_0	Column_1	Column_2	Column_3	Column_4
YAL018C	0.12	-0.25	-0.18	-2.06	-0.01
YKL185W	-0.56	-1.36	-0.62	-1.6	-1.74
YPR119W	-2.4	-1.43	-2.0	-2.32	-1.4
YNL058C					
YIL080W					
YCL055W					
YPL232W					
YDL035C					
YIL082W					
YDL037C					
YPL163C					
YDL039C					
YNL279W					
YLR413W					
YBL002W					
YBL003C					
YDL186W					
YGR084C					
YGL053W	2.11	0.72	0.42	-0.09	0.31
YBR038W	-1.51	-2.18	-1.03	-2.56	-2.0
YMR259C	-0.09	0.21	-0.36	-0.17	-0.12
YGL055W	0.54	0.36	0.14	-0.94	-1.36
YKR004C	-0.38	2.85	-0.45	-0.23	-0.03
YIL017W	1.79	2.1	1.34	1.88	1.08
YNL066W	-1.89	-2.56	-2.25	-2.56	-1.43
YER124C	-0.4	-1.12	-1.43	-2.56	-1.6
YER059W	-0.4	0.04	-0.29	-0.3	-0.25
YJL078C	-1.4	-1.56	-1.79	-2.12	-1.94
YBR040W	5.79	5.07	3.01	0.86	0.16

# PathInfo Viewer

HTML w/ TOC

Model View Edit Help

Brucella spp.

- I. Taxonomy Information
  - A. Species
  - B. Variant(s)
  - C. Carrier(s)
- II. Organism Information
  - A. Life Cycle
  - B. Physical Characteristics
  - C. Genome Information
- III. Epidemiology Information
  - A. Location
  - B. Transmission
  - C. Environment
  - D. Intentional Release
- IV. Host Interaction
  - A. Human
  - B. Cow
  - C. Goats and Sheep
- V. Labwork Information
  - A. Biosafety Information
  - B. Culturing Information
  - C. Diagnostic Tests
- VI. Bibliography
  - A. Journal References
  - B. Book References
  - C. Website References
  - D. Thesis References
  - E. Curation Information

7. **Brucella melitensis biovar 3 Ether strain**

- a. **Scientific Name:** Brucella melitensis biovar 3 Ether strain
- b. **Parent Name:** B. melitensis
- c. **Description:** Ether strain, corresB. melip ([Gandara et al., 2001](#) ; [Alton et al., 1988](#) ) or this biovar.

8. **Brucella melitensis biovar 3 isolates 254, 255, 257, 258, 259 and 306**

- a. **Scientific Name:** Brucella melitensis biovar 3 isolates 254, 255, 257, 258, 259 and 306
- b. **Parent Name:** B. melitensis
- c. **Description:** These isolates were B. melip ([Gandara et al., 2001](#) ; [Alton et al., 1988](#) ) d bone marrow samples.

9. **Brucella melitensis biovar 3 isolates G914, G1024 and T64/40**

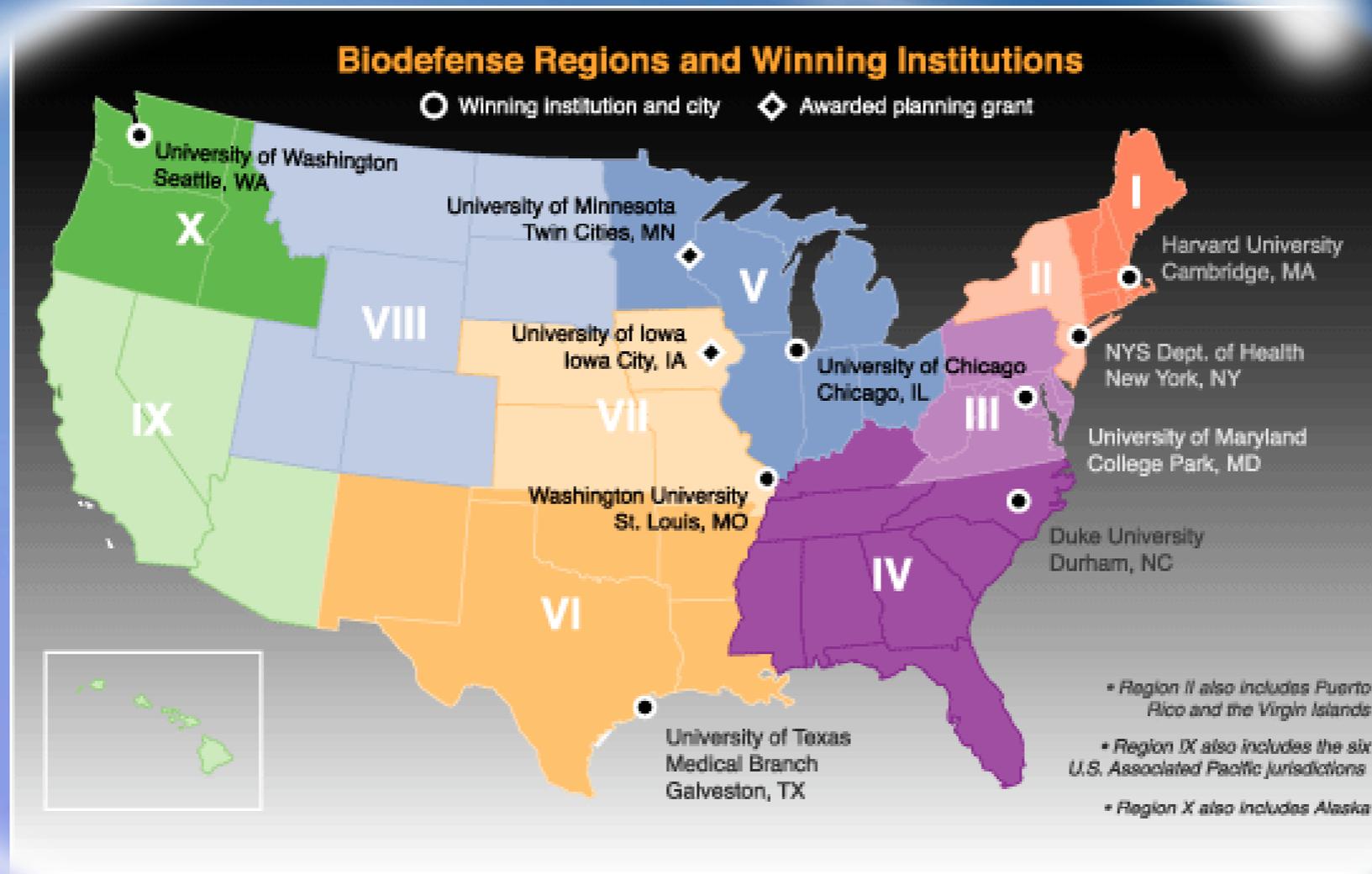
- a. **Scientific Name:** Brucella melitensis biovar 3 isolates G914, G1024 and T64/40
- b. **Parent Name:** B. melitensis

10. **Brucella abortus biovar 1 strain 544**

- a. **Scientific Name:** Brucella abortus biovar 1 strain 544
- b. **Parent Name:** B. abortus
- c. **Description:** Strain 544, corresponding with ATCC 23448, is the type strain for this biovar.

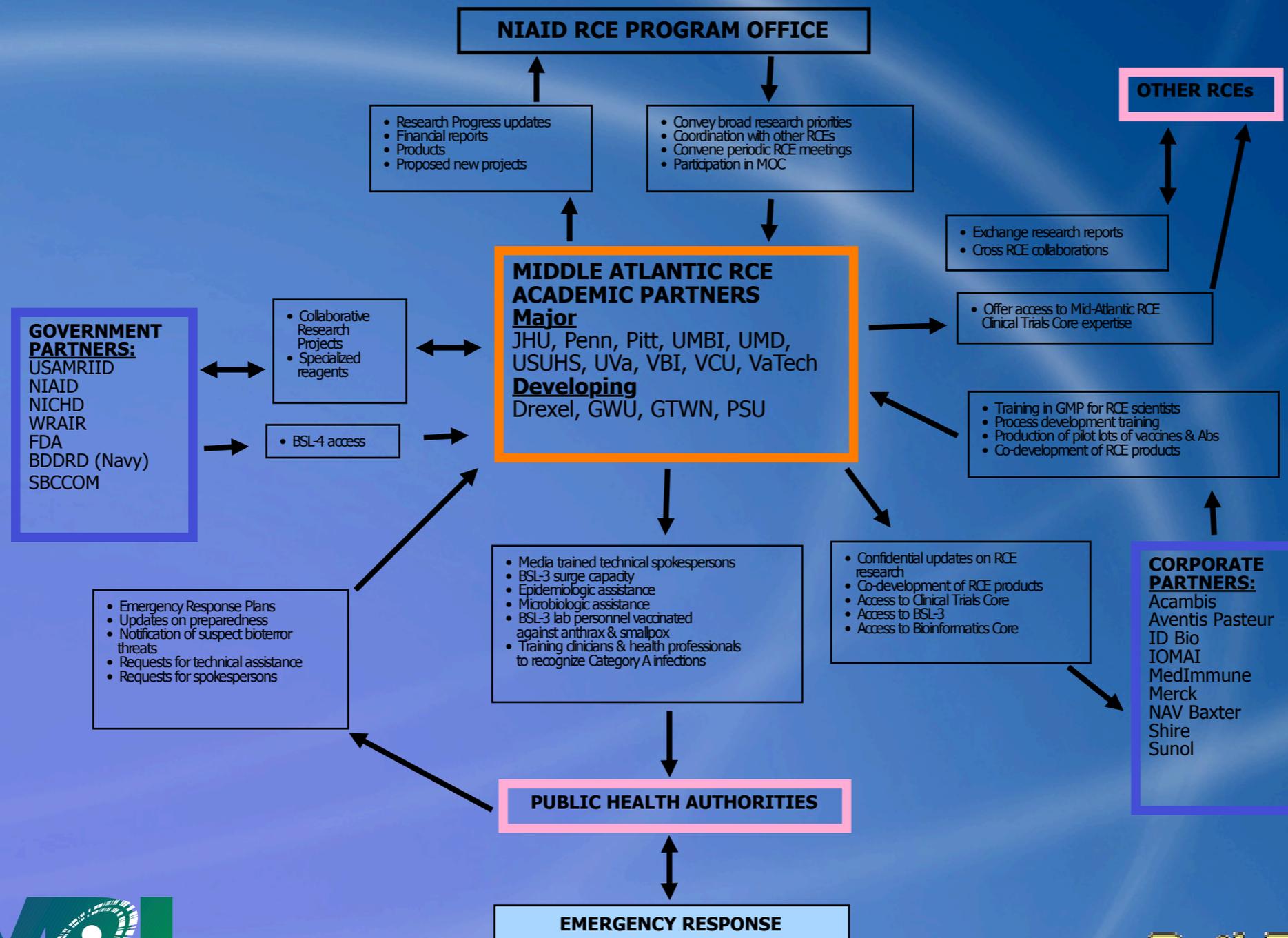
Alton GG, Jones LM, Angus RD, dummy title. 13148 - 13153. In: Elliot REW, Christiansen KH, *techniques for the brucellosis laboratory*. 1988; Institut National De La Recherche Agronomique, Paris. [ISBN:2-7380-0042-8]

# NIAID's RCEs

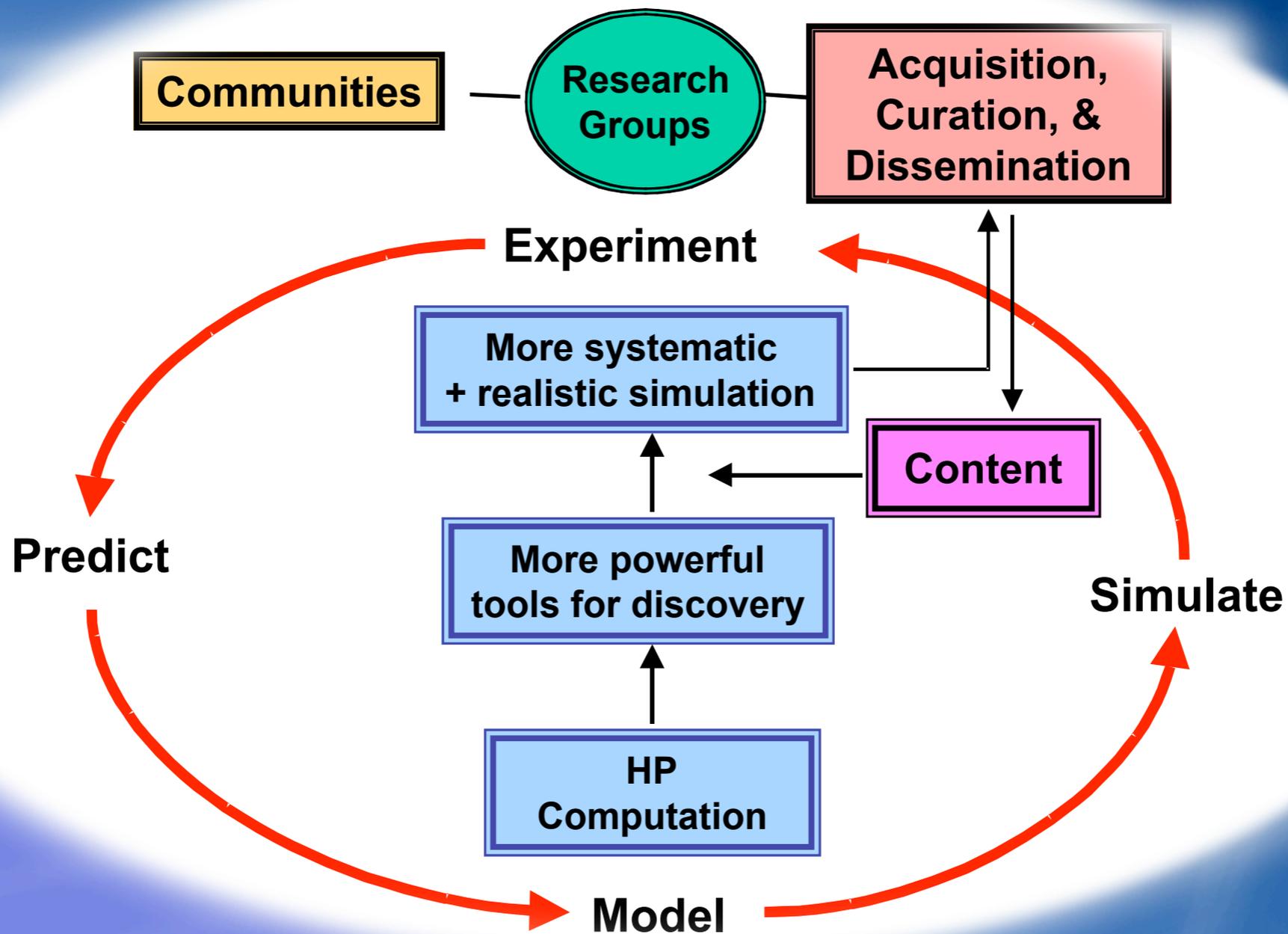


# Cyberinfrastructure

## Enables MARCE



# 21st Century Pathosystems Biology



# Financial Support

- **Acknowledging funding from:**
  - **Corporations**
    - IBM Corporation
    - Sun Microsystems
    - TimeLogic Inc.
  - **State and Federal Grant Agencies**
    - Virginia's Commonwealth Technology Research Fund - CTRF
    - NIH, NSF, USDA, DOE and DoD

# Questions?



VIRGINIA  
BIOINFORMATICS  
INSTITUTE  
AT VIRGINIA TECH

