

BioWorks: a Platform for Integrated Genomics

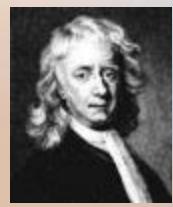
Andrea Califano
Columbia University

Scientific Progress

- 1600-1700
 - Mathematics
- 1700-1800
 - Chemistry
- 1800-1900
 - Physics
- 1900-2000
 - Biology



Leibniz



Newton



Euler



Avogadro



Dalton



Lavoisier



Dalton



Einstein



Bohr



Hartwell

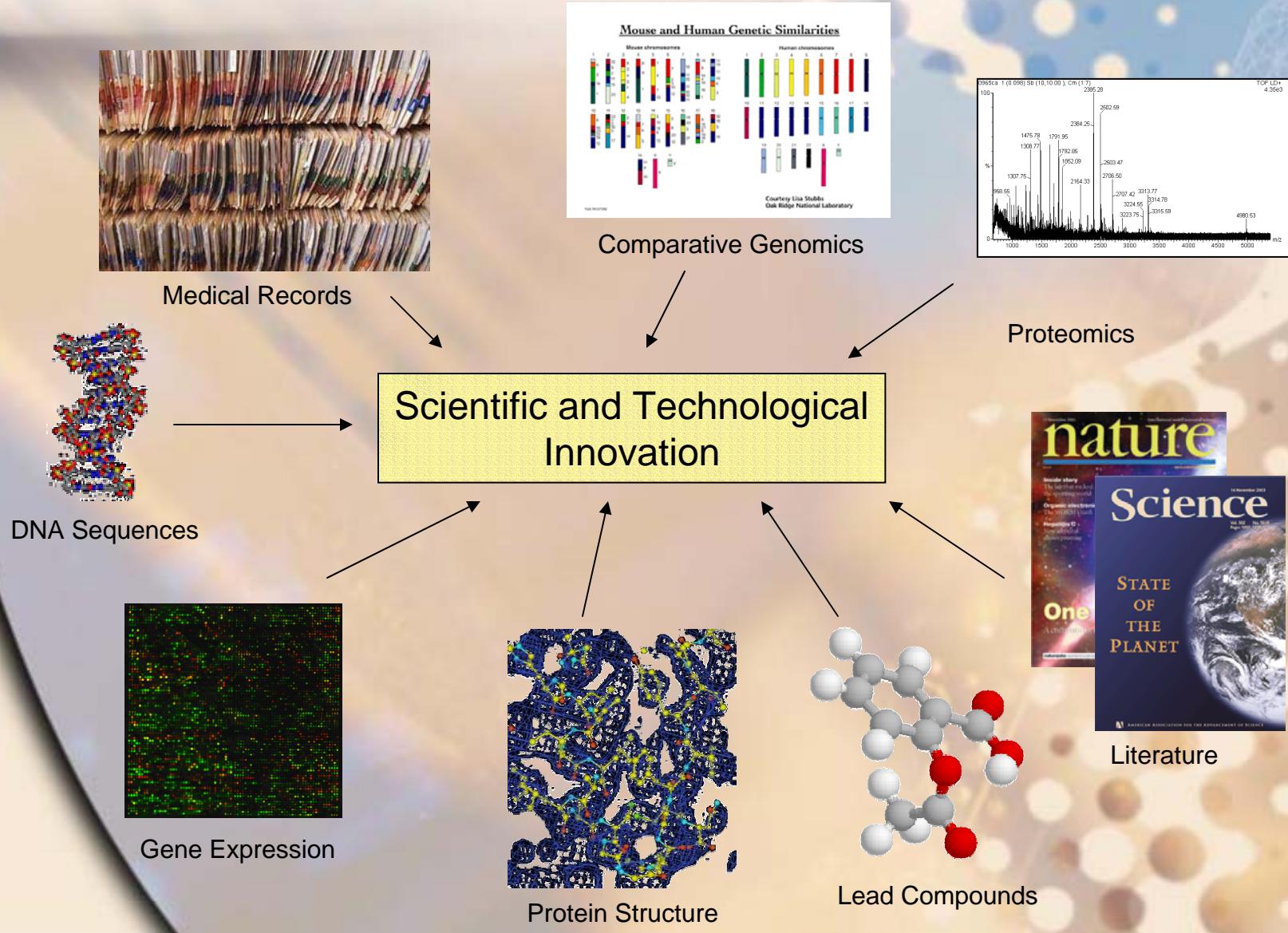


Brenner



Horvitz

Integrated Genomics (1 + 1 = 3)



BioWorks

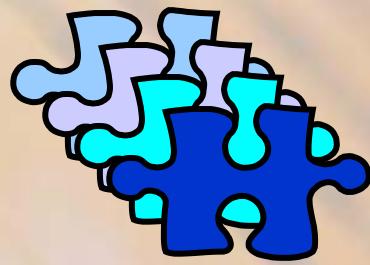
- Repository Based Server-Grid for the integration of distributed interoperable components
 - Open Source
 - Plug&Play
 - Grid-Enabled and Scaleable
 - Ontology-Anchored
 - Easy to extend
 - Community based and driven

Platforms

- caBIO/caWorkbench
 - NCI sponsored Open source data analysis platform
http://caarray.nci.nih.gov/data_analysis
- AMDeC Cores:
 - AMDeC Bioinformatics Core at Columbia
 - AMDeC Microarray Resource Core
- AMDeC Integrated Genomics Core

BioWorks

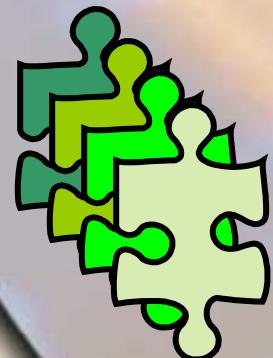
Visualization



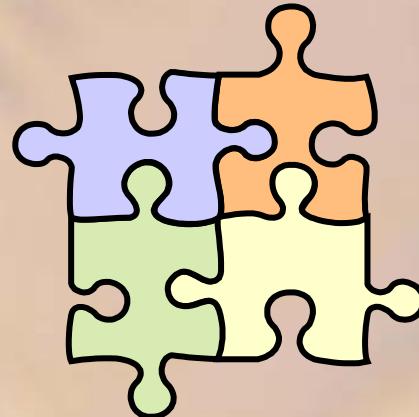
Data Management



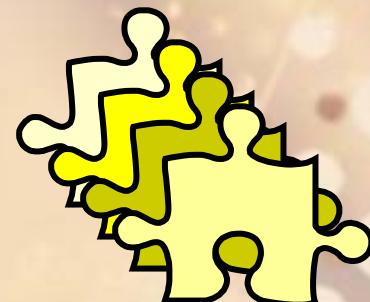
Algorithms



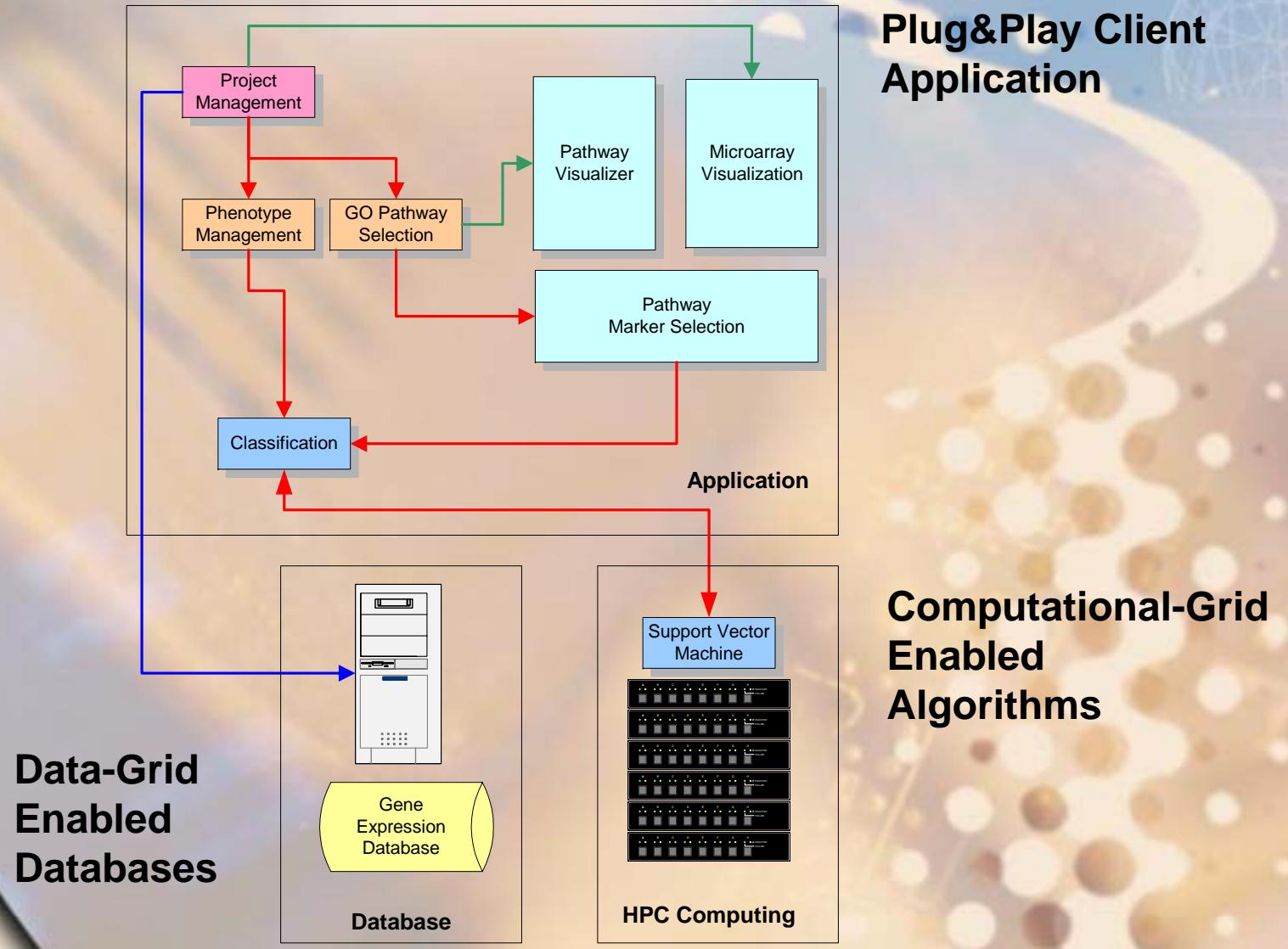
Biomedical
Application



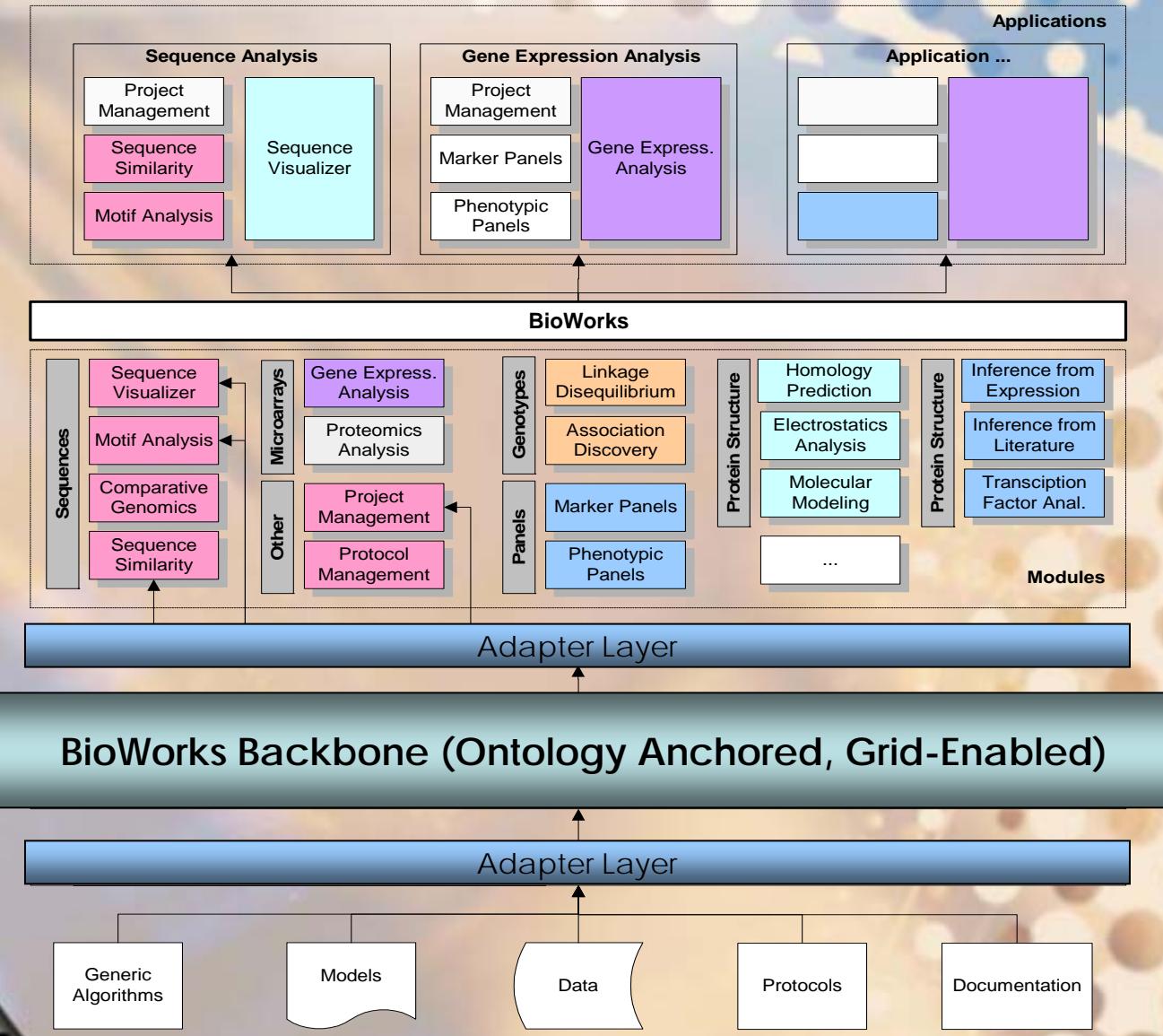
Databases



BioWorks: Plugin Model



BioWorks: Architecture



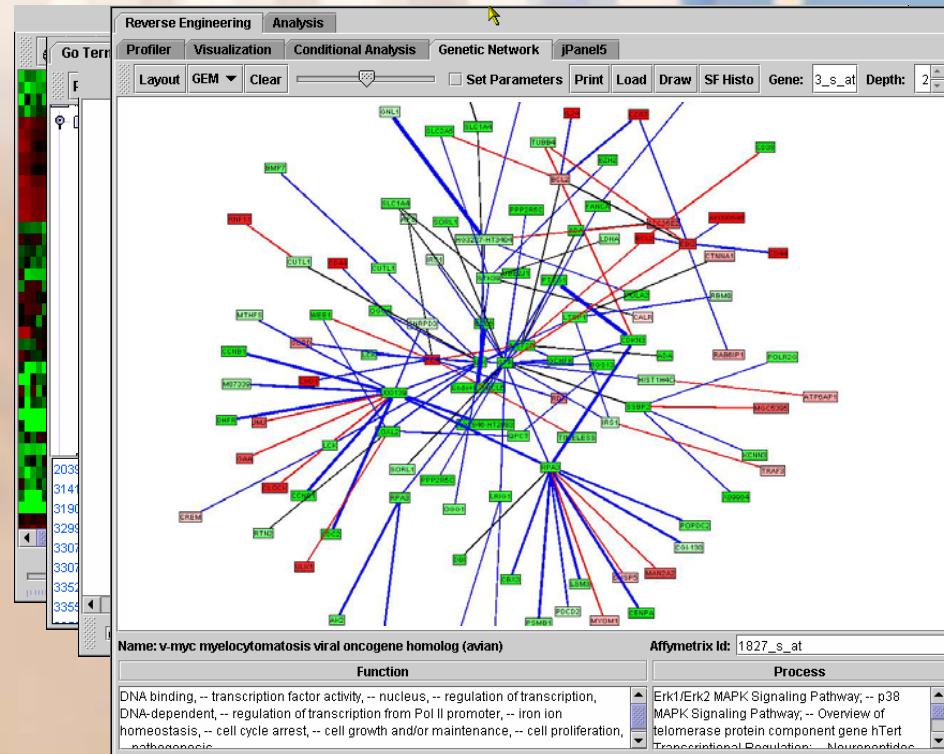
Databases

The screenshot shows the Project Panel with the title 'Project Panel' at the top. Below it, a tab labeled 'Project Folders' is selected, indicated by a grey background. The main area displays a tree view of project folders. At the top level, there is a folder icon followed by the text 'Workspace'. Below it, there are two project icons, each followed by the text 'Project'. Under the second 'Project' node, there is a blue folder icon followed by the text 'TheMatrix75_annotated.exp'. The entire interface has a light grey background with dark grey horizontal bars separating sections.

Data Management

Gene Panel		Phenotype Panel
19	86cl1EV.CHP	
200	85cl1EV.CHP	
200	72D40L30.CHP	
200	71D40L30.CHP	
200	70D40L30.CHP	
200	69DEV30.CHP	
201	68DEV30.CHP	
201	67DEV30.CHP	
201	66D40L27.CHP	
201	65D40L27.CHP	
198	64D40L27.CHP	
199	63DEV27.CHP	
199	62DEV27.CHP	
199	61DEV27.CHP	
197	58B40L219.CHP	
197	54EV215.CHP	
Phenotype		
detailed designation		
<ul style="list-style-type: none">   BL cell line   BL cell line expressing BCL6DPEST (Treated)   BL cell line expressing BCL6DPEST   BL cell line (Treated) 		

Visualization



Algorithms

Analysis			
Hierarchical Clustering Analysis SOM Analysis			
Clustering Method	Single Linkage ▼		
Clustering dimension	Microarray ▼		
Distance Metric	Euclidean ▼		
		Save Settings	Analyze

Dataset Management		
Project management	S	Allows heterogeneous data to be loaded into multiple projects and selected for further analysis or visualization. Furthermore, allows ancillary data, such as results of algorithmic runs, parameters, and images to be stored in the project with the originating dataset and saved.
Gene Selection	S	A GUI that allows collection of genes to be selected into a hierarchical panel structure and activated/deactivated. This allows other components to display or process information based on the selection.
Phenotype Selection	S	Allows specific phenotypic criteria to be selected. This allows other components to display or process information based on the selection
Gene Ontology GUI	N	Allows selection of genes using GO function, process, or compartment-based hierarchies. Allows individual genes or entire functional, process, or compartment-based set of genes to be selected as panels. (See Gene Selection)

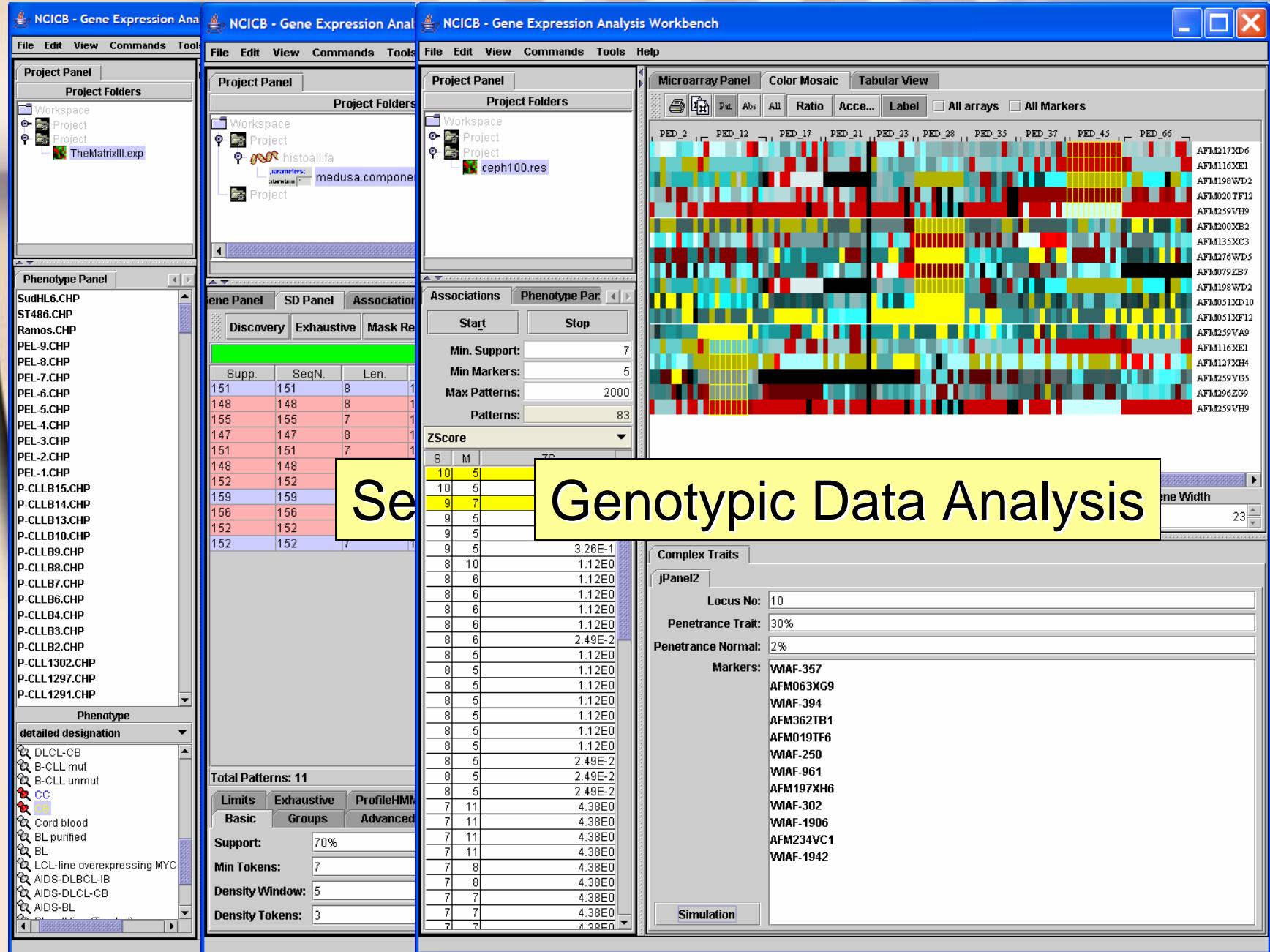
Sequence Analysis and Visualization		
Sequence Visualization	N	Allows the display of DNA, Protein, and other 1D sequences and their annotation using patterns or other types of annotation.
Sequence Pattern Disc. GUI	N	This allows the discovery of patterns in DNA, Protein, or other 1D data using plugin client-server pattern discovery algorithm. It supports normal, exhaustive, and hierarchical discovery.
1D Pattern Visualization	N	Allows the visualization of sequence-based patterns.
Motif Histogram	N	Allows visualization and positional analysis of 1D patterns on sequences, including the discovery of patterns of patters (Flexible patterns or components)

Pathways Analysis and Visualization		
Pathway GUI	W	Allows pathways to be visualized and their genes to be selected for visualization in other contexts
Pathway Analysis	N	GUI for the reverse engineering, analysis, and visualization of genetic pathways from gene expression data. Includes an interactive graph visualization tool and coupling with GO annotations.

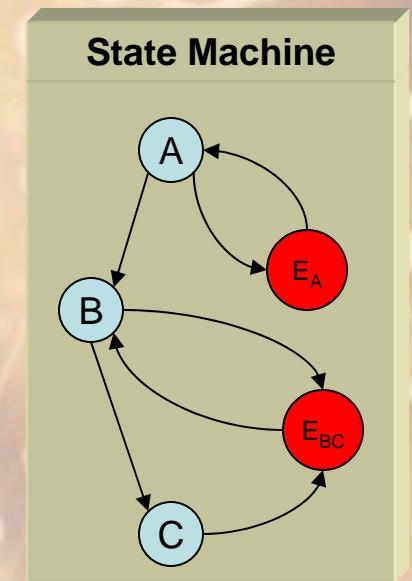
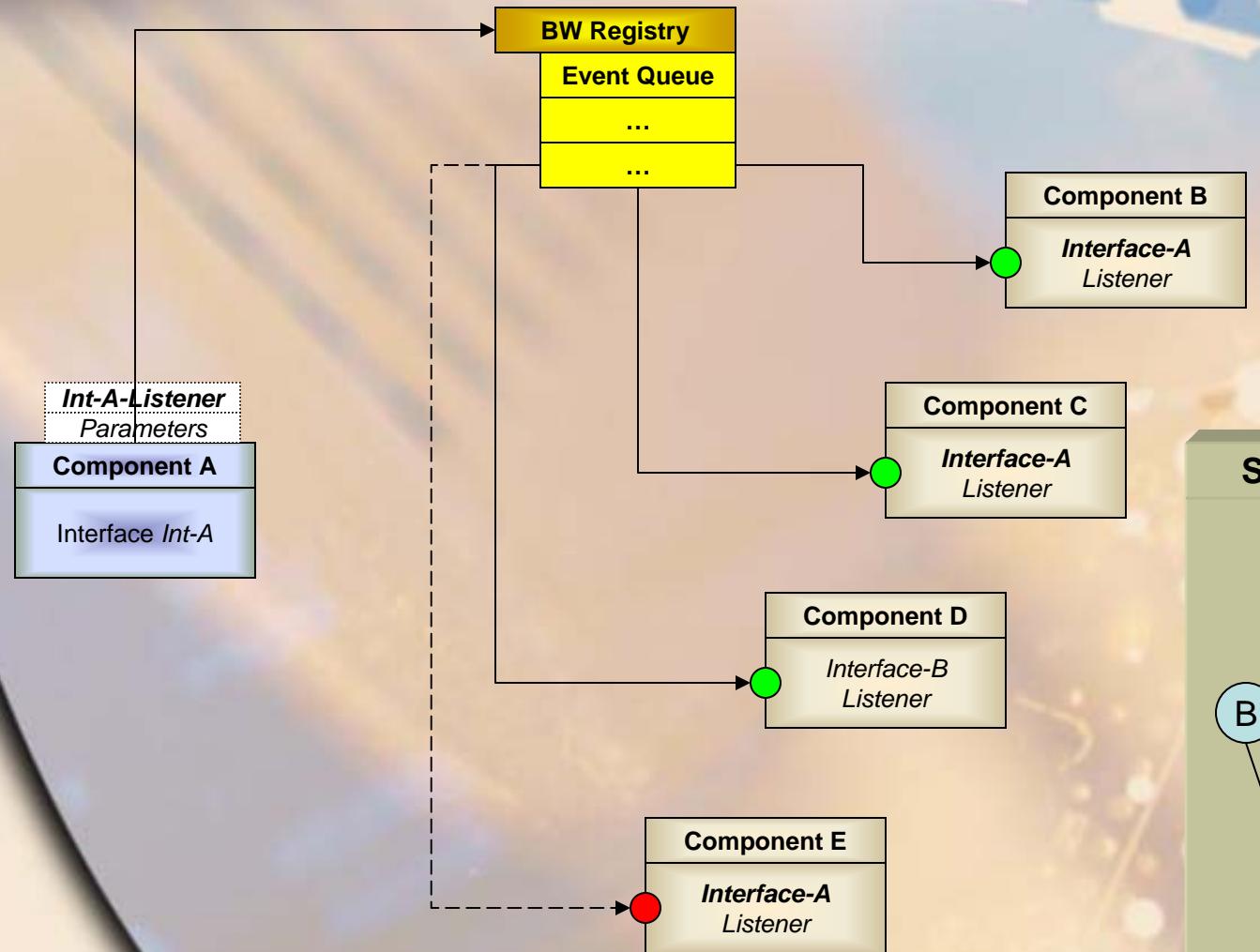
Microarray Data Analysis and Visualization		
Microarray Visualization	S	Allows interactive microarray data visualization, including gene expression, proteomics, and genotypic data.
Color Mosaic GUI	S	Allows microarray data to be displayed using the standard 2D matrices, with mRNA probes in the rows and microarrays in the columns
Microarray Table GUI	S	Allows visualization and editing of microarray data, including Gene Expression, Proteomics, and Genotypic data.
2D Pattern Disc. GUI	N	This allows the discovery of patterns in gene expression, genotypic, or other 2D data using plugin client-server pattern discovery algorithm.
Analysis Plugin GUI	W	Allows several plugin algorithms to be selected for the analysis of microarray data and their parameters to be set. These include SOM and hierarchical clustering.
Filtering Plugin GUI	W	Allows several plugin algorithms to be selected for the filtering or preprocessing of microarray data and their parameters to be set. These include missing value filtering, Deviation filtering, Expression-threshold filtering, Affy detection-call filter, 2-channel threshold filter.
Normalization Plugin GUI	W	Allows several plugin algorithms to be selected for the normalization of microarray data and their parameters to be set. These include: missing values analysis, Log transform, threshold normalization, Marker-based normalization, Array-based normalization, mean and variance based normalization.
Annotation GUI	W	Allows DAS annotation to be retrieved and displayed relative to microarray genes.
Comments GUI	W	Allows comments about specific genes or microarrays to be entered and saved
Experiment GUI	W	Allows experiment related annotations to be entered and saved
History GUI	W	Shows all analyses (and related parameters) run until this point on the data set.
Hierarchical Clustering	W	Visualizes the dendograms resulting from hierarchical clustering
Expression Profile	W	Visualizes the expression profile clusters resulting from SOMs
Microarray Table GUI	W	Allows visualization and editing of microarray data, including Gene Expression, Proteomics, and Genotypic data.

Lightweight Client-side Algorithms		
SOM	W	Clusters gene expression data with the Self Organizing Map algorithm
Hierarchical Clustering	W	Builds a dendrogram using the hierarchical clustering algorithm
Missing Value Filter	W	Removes missing data to produce a fully instantiated data matrix
Deviation Filtering	W	Removes outliers
Threshold Filtering	W	Removes data based on an expression threshold parameter
Affy Detection call Filter	W	Filters Affy data based on their assessment
2-channel threshold Filter	W	Filters the data based on threshold parameters
Missing Value Analysis	W	Interpolates missing data to produce a fully instantiated data matrix
Log-transform	W	transforms the data using a log-transformation
Threshold Normalization	W	Normalizes the expression based on a threshold value
Marker-based Normalization	W	Normalizes the expression based on the value of a marker
Array-based Normalization	W	Normalizes the expression based on a set of microarrays
Mean+Var. Normalization	W	Normalizes the expression based on a the mean and the variance
Association Discovery	N	Dynamic Library to perform 2D pattern discovery over Gene Expression or Genotypic Data. (Genes@Work)

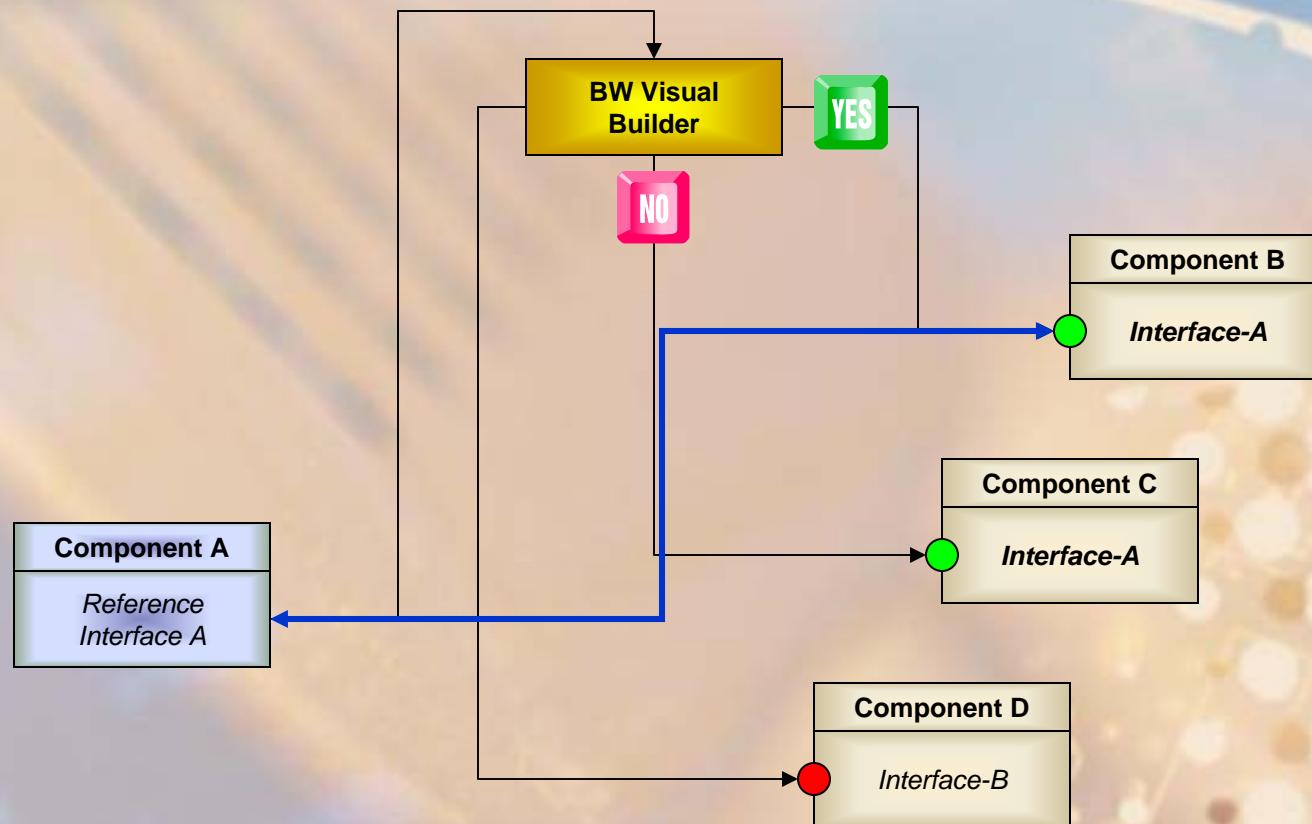
Computational Adapters		
BLAST Suite	N	Java adapter for Paracel Server-based BLAST Suite
1D-SPLASH Pattern discovery	N	Java adapter for C++ Server-based sequence pattern discovery algorithm
2D-SPLASH Assoc. Discovery	N	Java adapter for C++ Client-based Gene Expression and Genotype pattern discovery algorithm. Must be ported to server-based architecture
Data Adapters		
Several	S	components for reading Affymetrix, Genpix, FASTA, and other data formats



Communication Models: Asynchronous



Communication Models: Synchronous

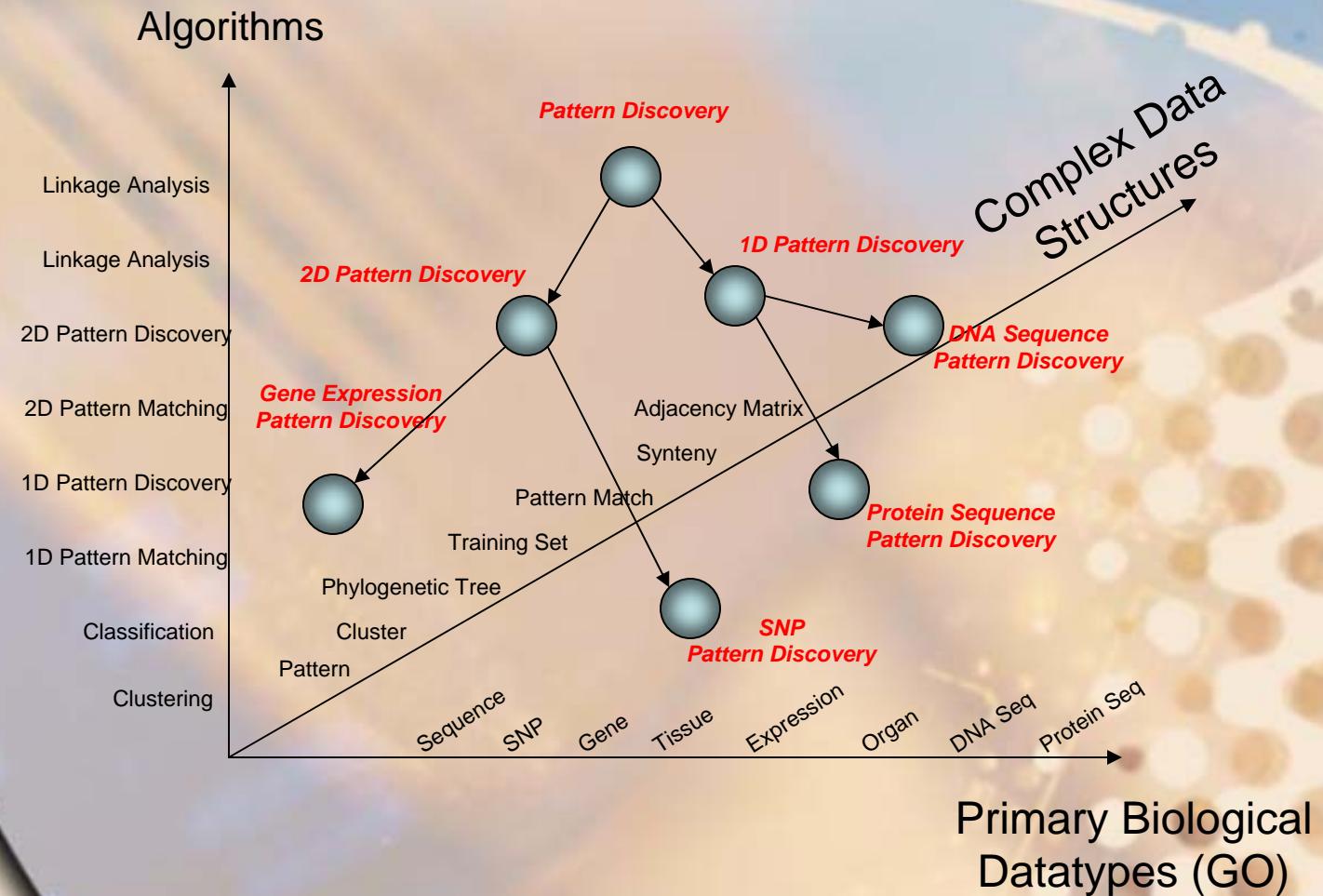


Ontology-Anchored Interfaces

- Scalable mechanisms for the definition of community based standards
 - Vocabulary
 - Grammar
 - Semantics
- Ontology Anchored Interface Design
 - Available only for Fundamental Biomedical Types
 - DNA Sequence
 - Gene
 - Gene Expression
 - ...
- Extend Existing Ontologies with Two New Axes
 - Complex Data Structures
 - Algorithms

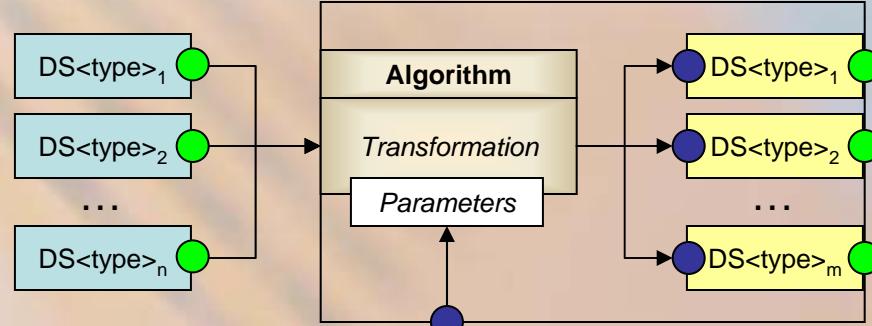
BISON: Biomedical Informatics Structured Ontology Notation

Interface Design: Algorithms



Algorithms and DB Adapters

Algorithmic Component



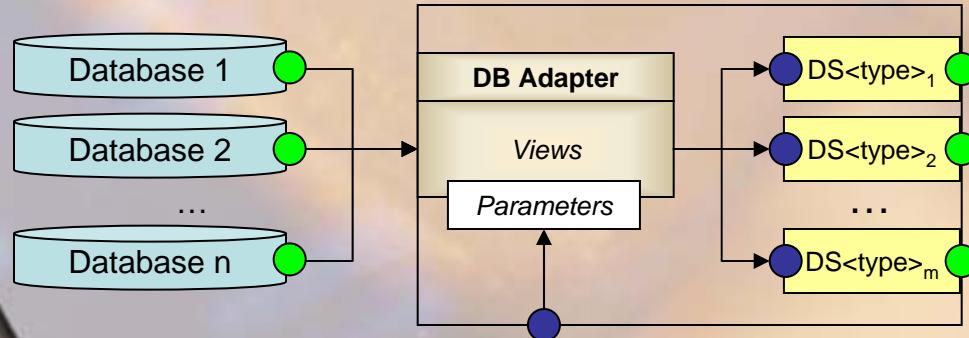
Optimization

- Architecture
- Persistence

● Interface (GET)

● Interface (SET)

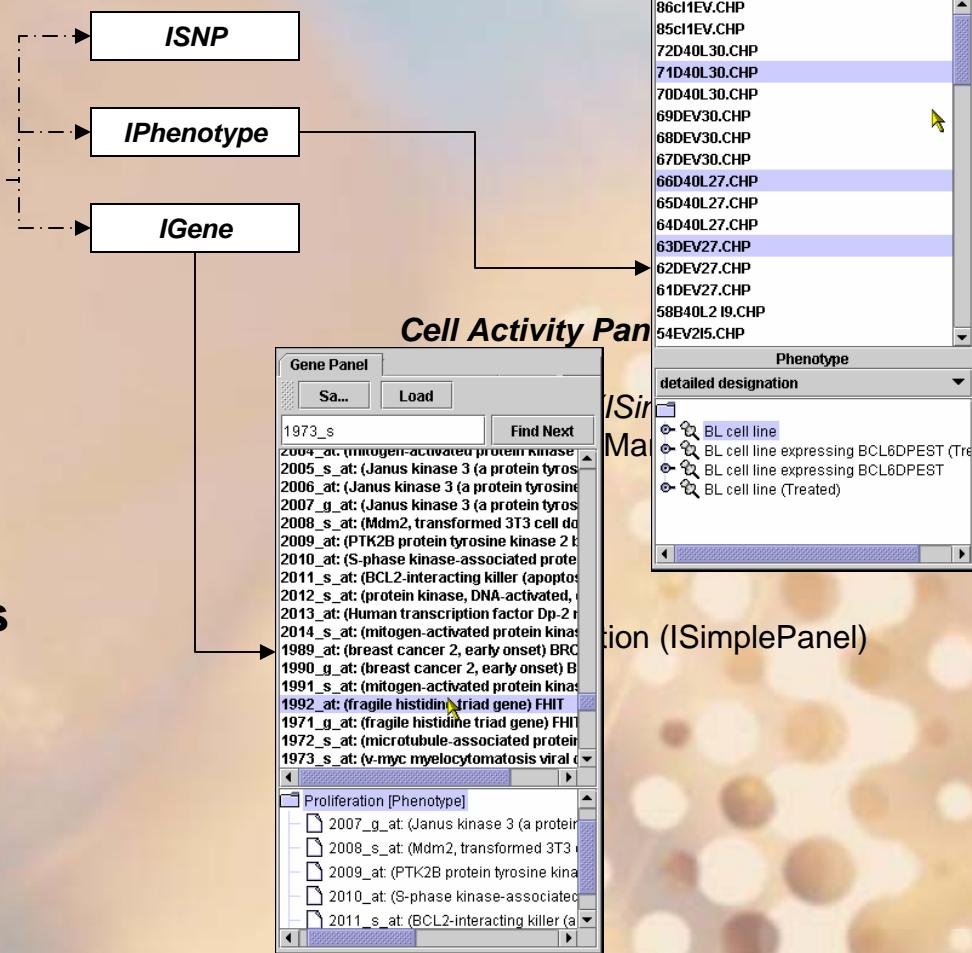
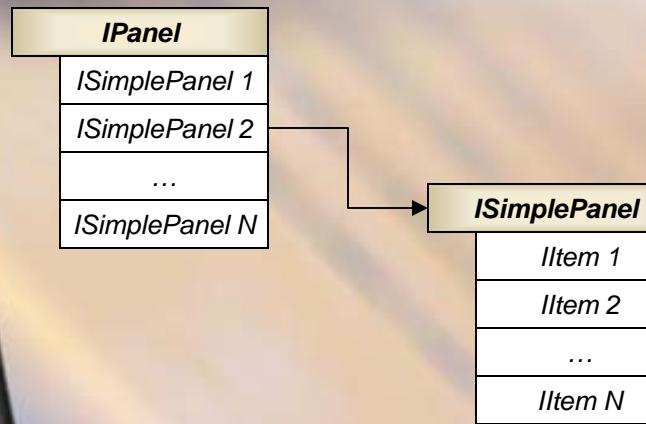
DB Adapter Component



Optimization

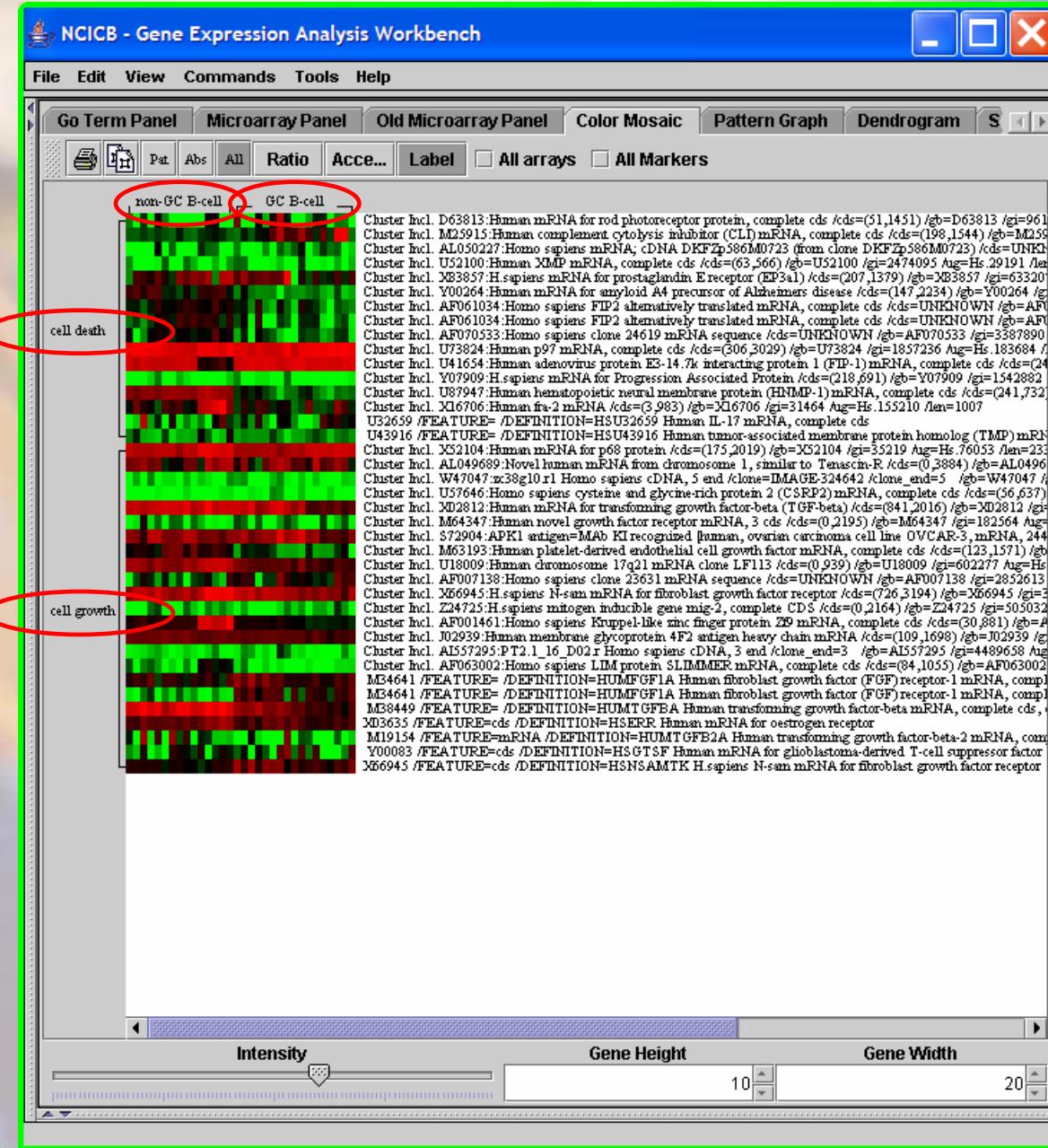
- Local Caching
- DB Updates

Interface Design: Data Structures

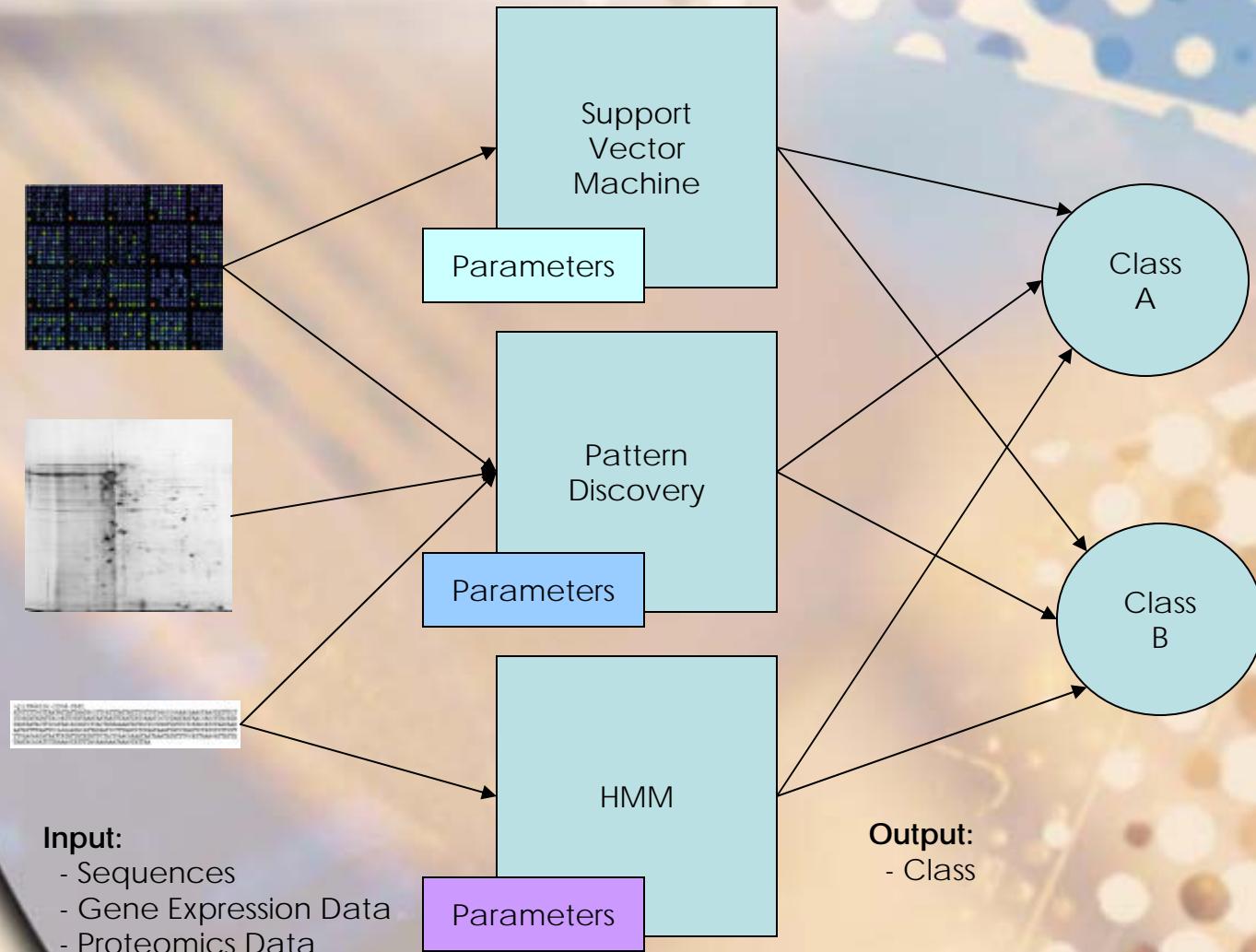


Hierarchical Grouping
Of Primary Biodata Types

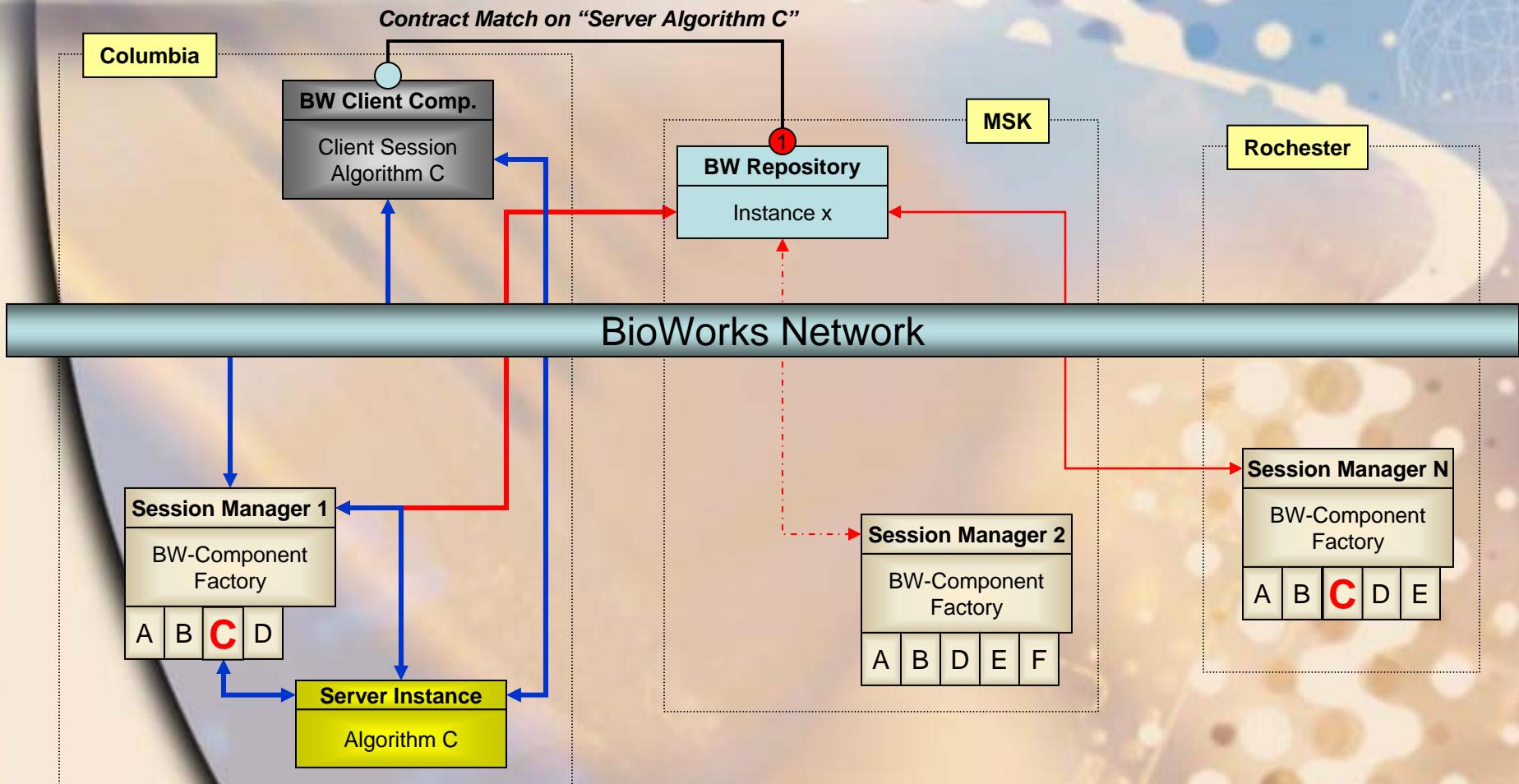
Selection (ISimplePanel)



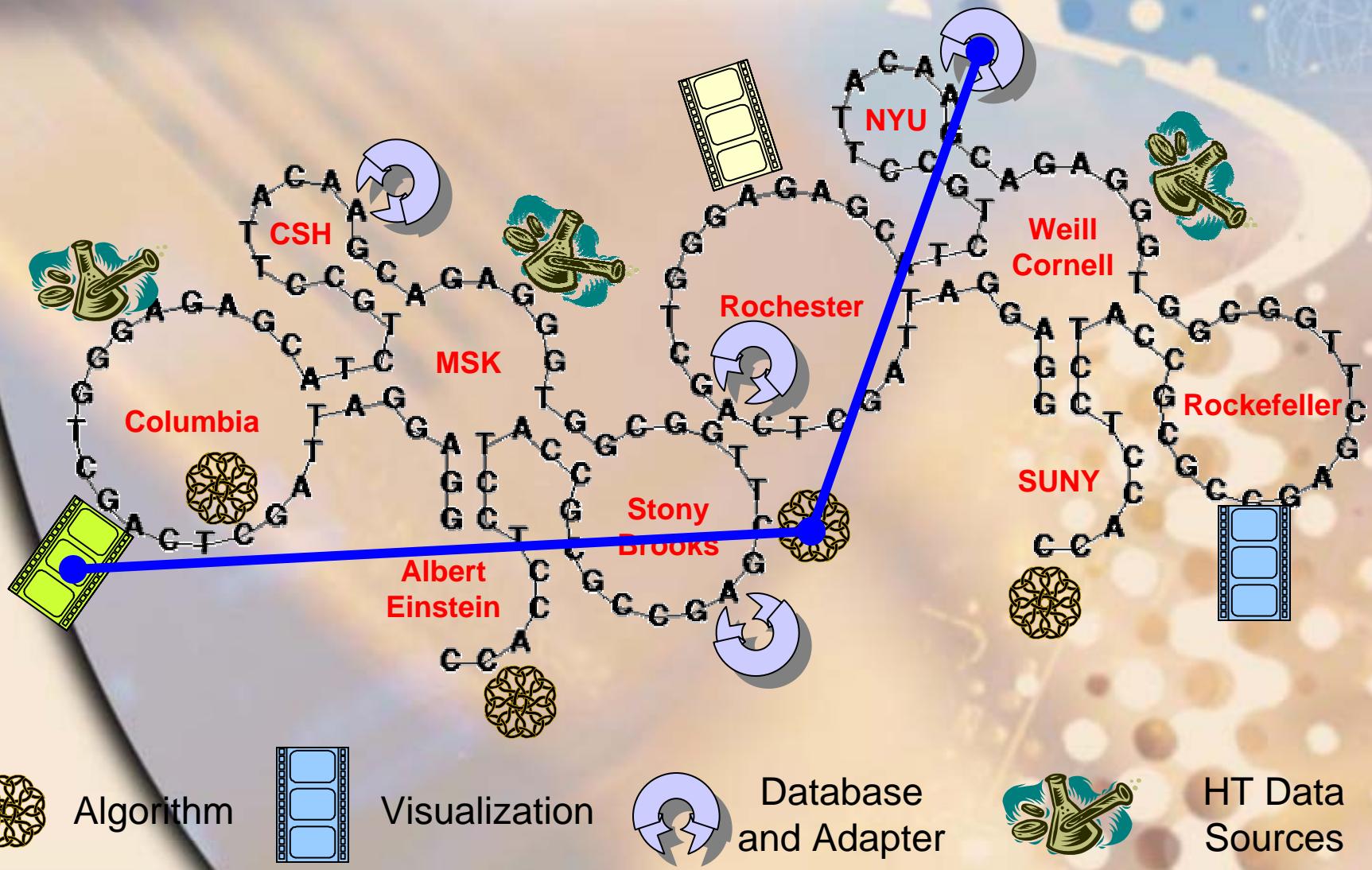
Universal Algorithms:



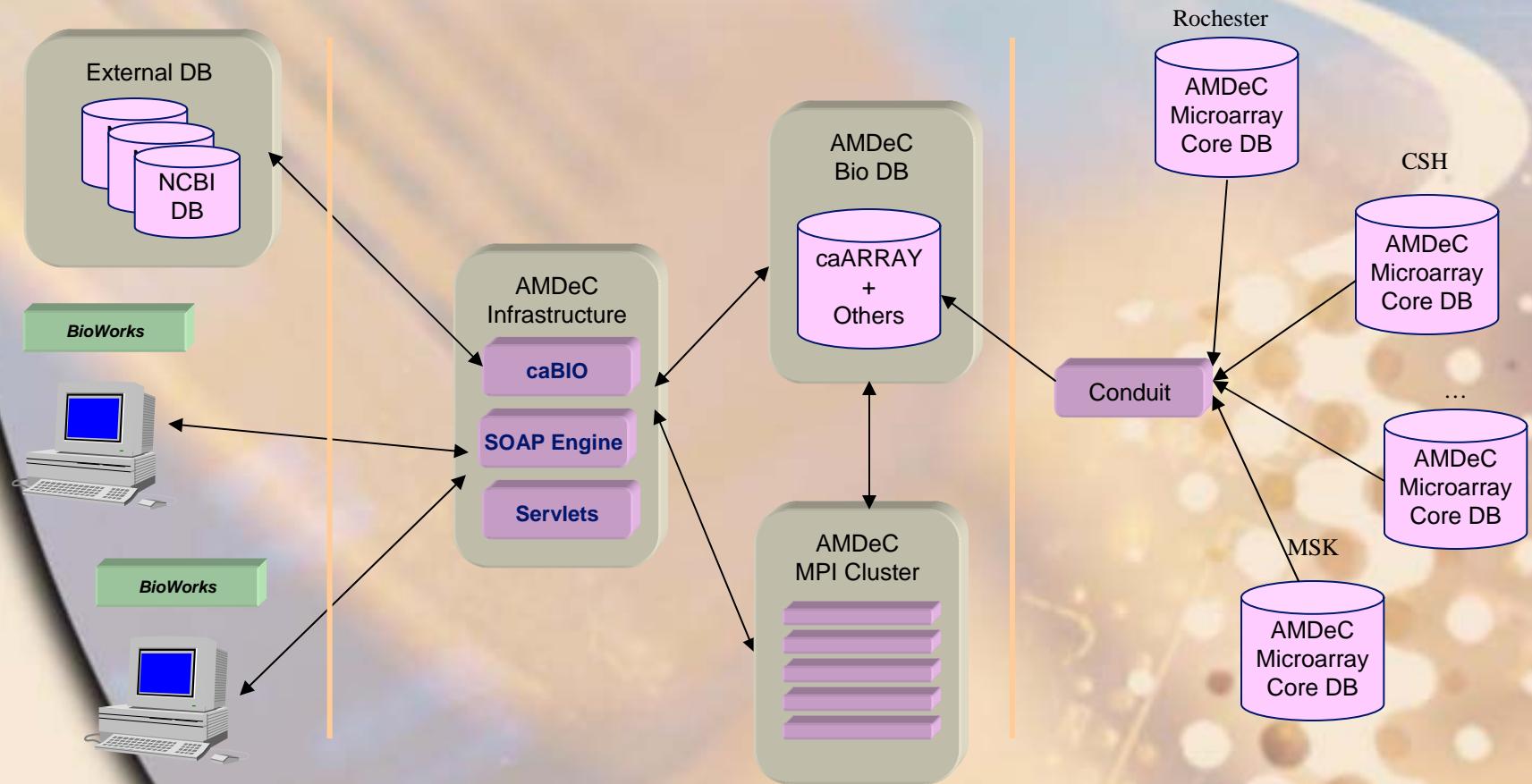
Grid-Enabled Computational Modules



NY Integrated Genomics Grid



Integrated Genomics Core: Pilot



Timeline:

- caWorkbench: Available now
- BioWorks applications: Summer 2004
- BioWorks Compu-Grid: Summer 2005
- BioWorks Data-Grid: Fall 2005