Post-Genomic Challenges For the Open Source Bioinformatics Community:

Systems Biology as a Target Source

Dan Kilburn, PhD

Beyond Genomics, Inc.

Bioinformatics in 2002

- Transition from a gene-centric to systems-oriented approach to knowledge discovery
- Introduction to Systems Biology
- Overview of approaches and methods in need of "Open Source Intervention"

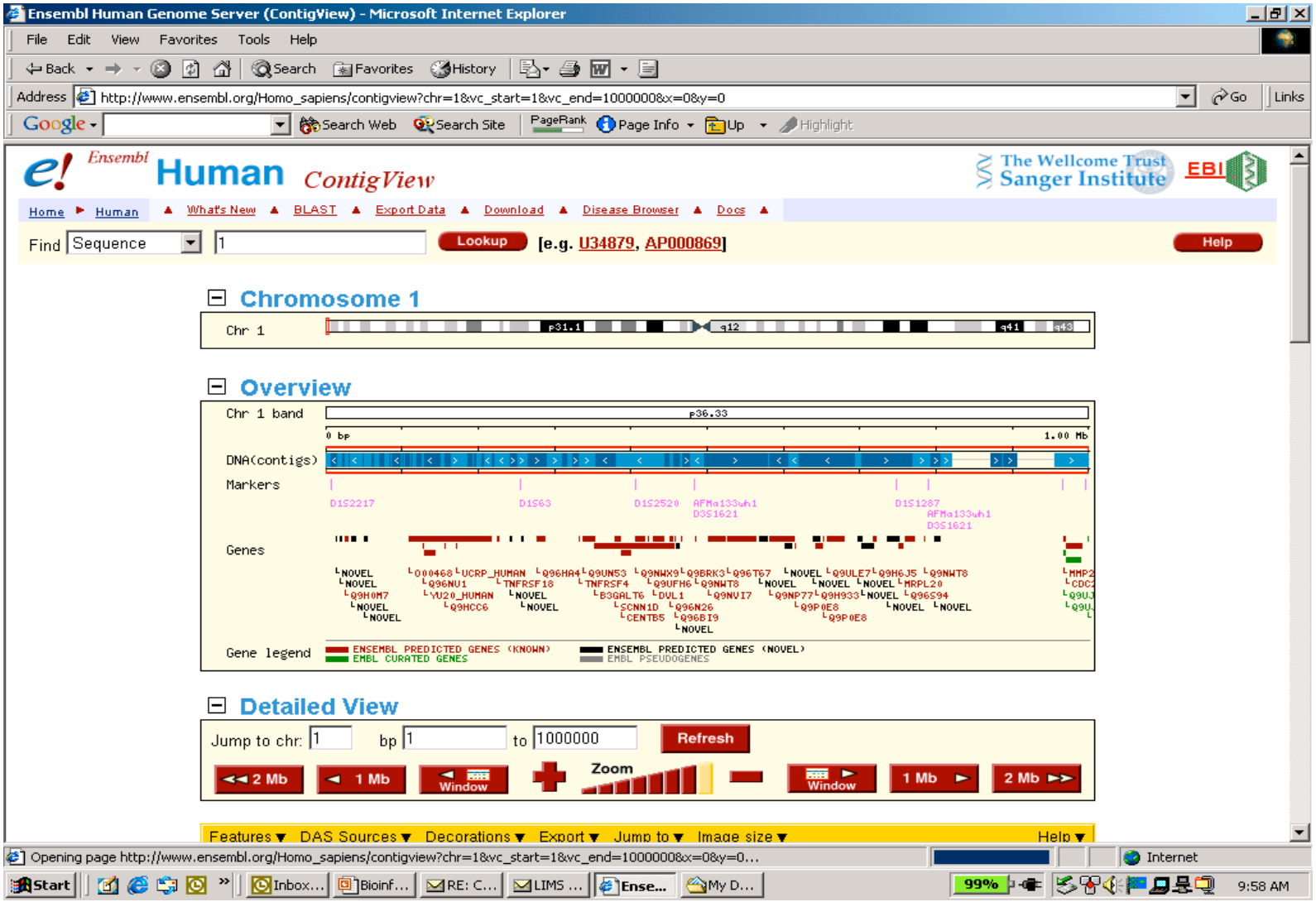
The Hegelian Dialectic and the Race to Sequence the Human Genome

- **Thesis:** Gene Myers, Craig Venter and the Apocryphal Cocktail Napkin
- **Antithesis:** Ensembl and the OpenBio Movement
- **Synthesis:** Demise of “Genomes For Sale”
- *What’s the new thesis?*

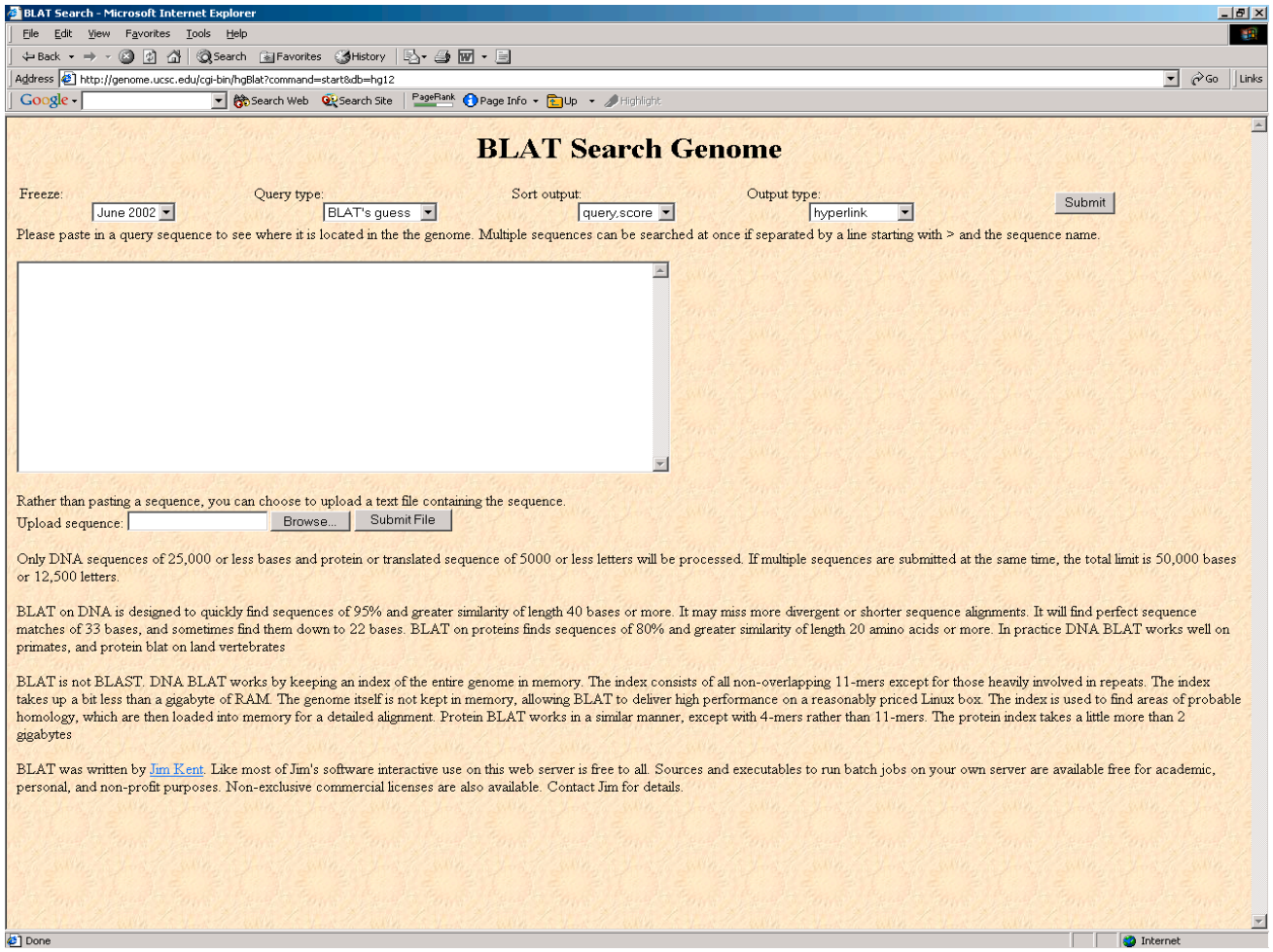
The Dialectic, Part II

- Pre-1998 Bioinformaticists often worked in labs headed by biologists. Problem: it's always difficult to work for someone who doesn't understand what you do
- 1998-2001 The Genome Utopia: Genomic data was available over the web at a rate of ~100 megabases a week, freeing us from the tyranny of biological rule and worrying about the messiness of where/how data was generated
- Post 2001 The Dream is over: now we have to work with biologists and raw data again

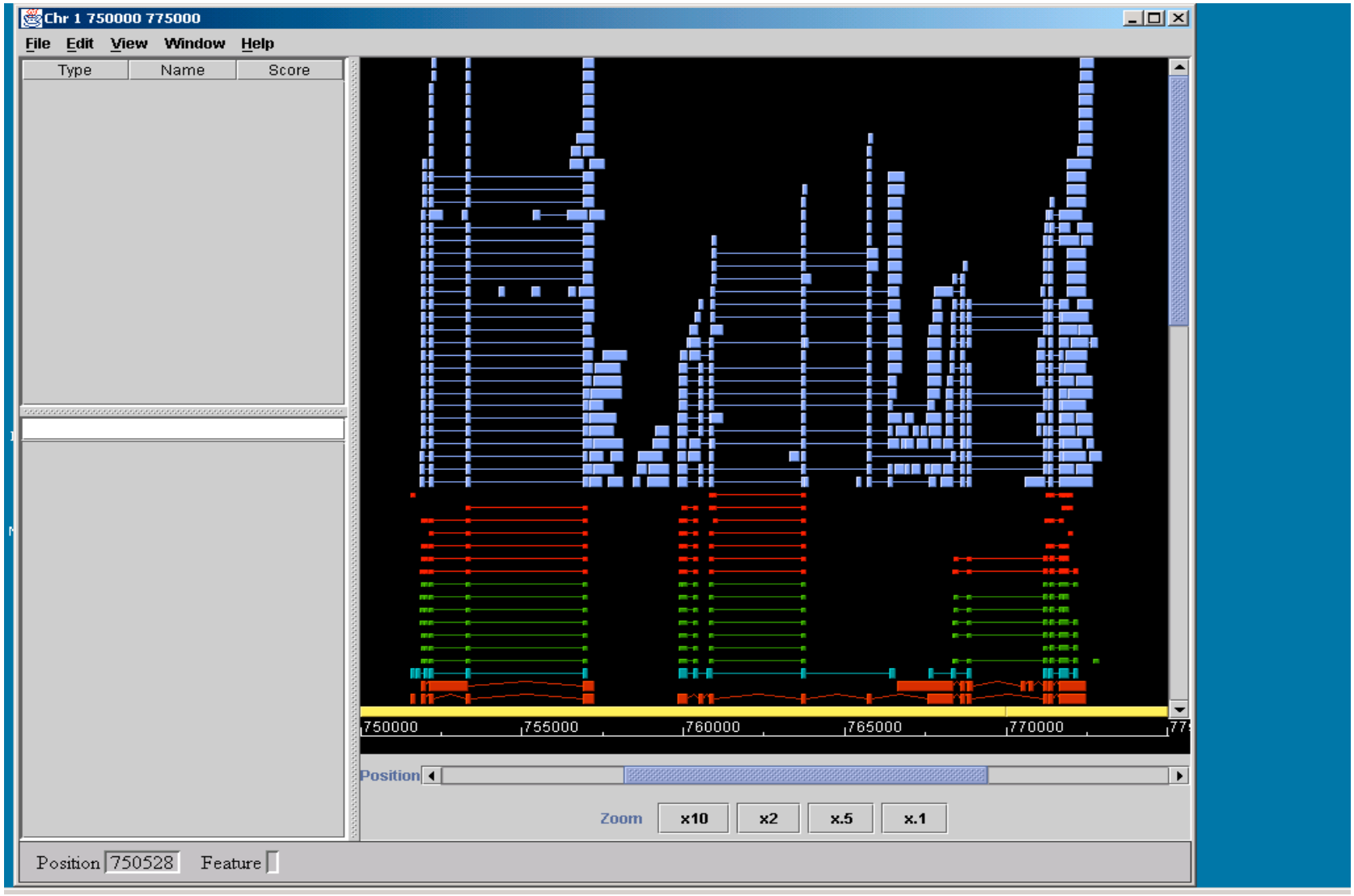
Ensembl: A Gene-Centric Genome Viewer



Jim Kent's BLAT Search Engine



The Apollo Genome Browser



Genome Browsers:
Challenges to Overcome / A Wish List

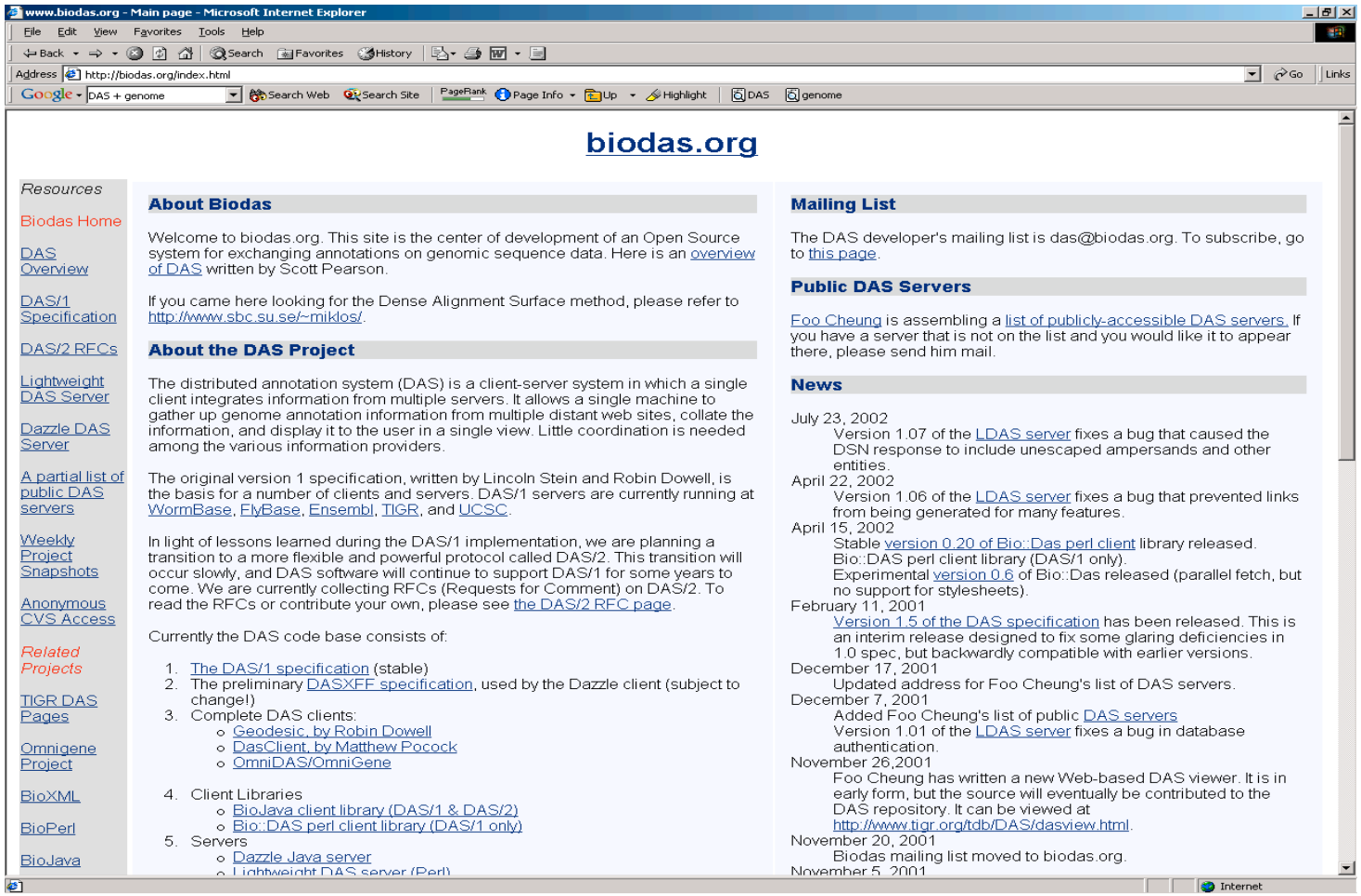
Static: underlying data is updated sporadically

Noninteractive I: users have little or no capacity to integrate proprietary data

Noninteractive II: browsers have little or no capacity to respond to or record users' path through data

Orthogonal: browsers depict physical relationships and gene family relationships, but not biologically-relevant ones like receptor-ligand and other partner relationships, biological processes, cellular and tissue localization.

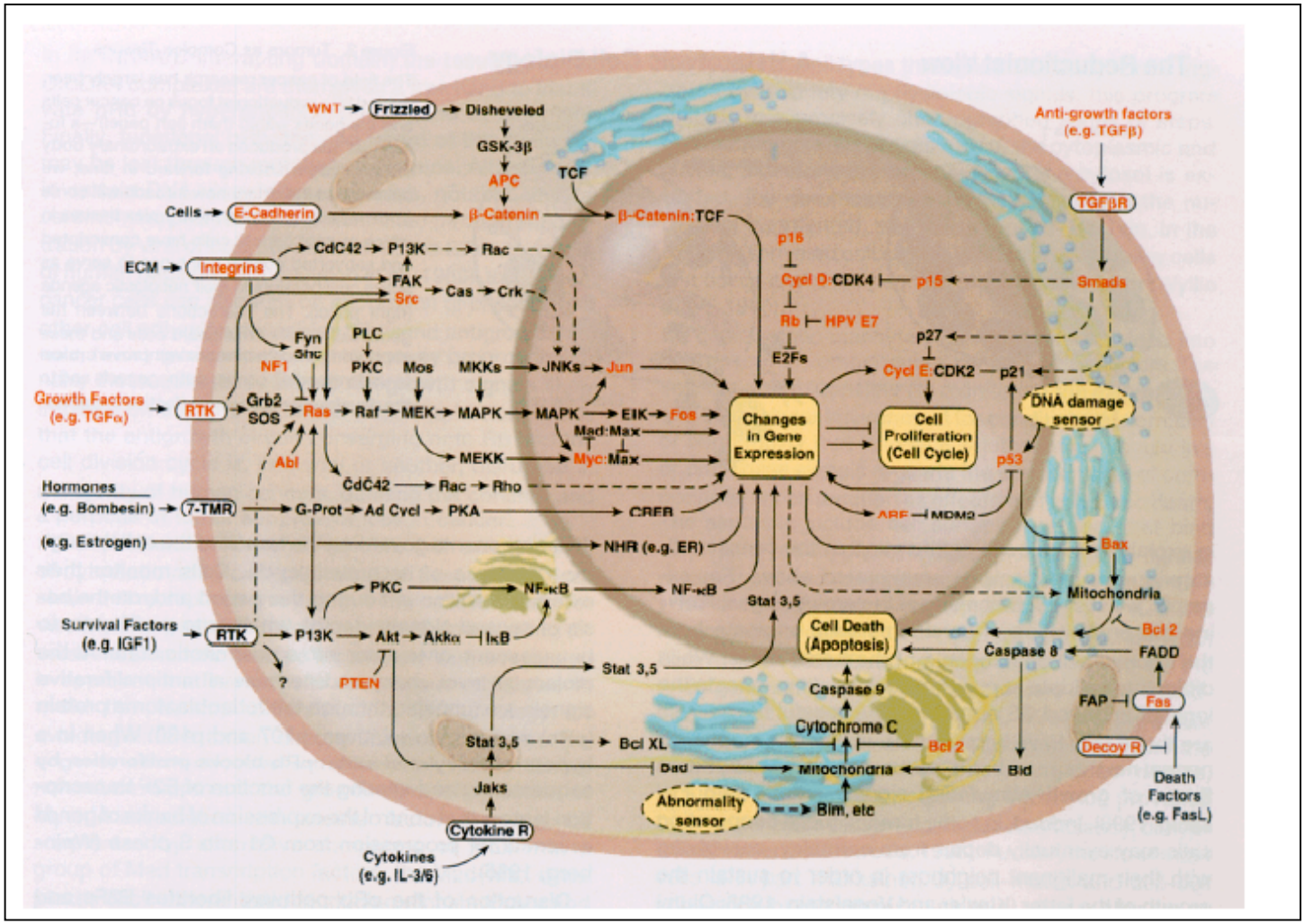
DAS: Integrating Distributed Data



Foreshadowing

- Moving Beyond the Genome Utopia
- Standards to Cope with Proliferation of Platforms and Vendors
- Lims Systems and Pipeline Management
- Data Analysis Standards
- Visualization Tools
- Working with Clusters while Avoiding Bankruptcy

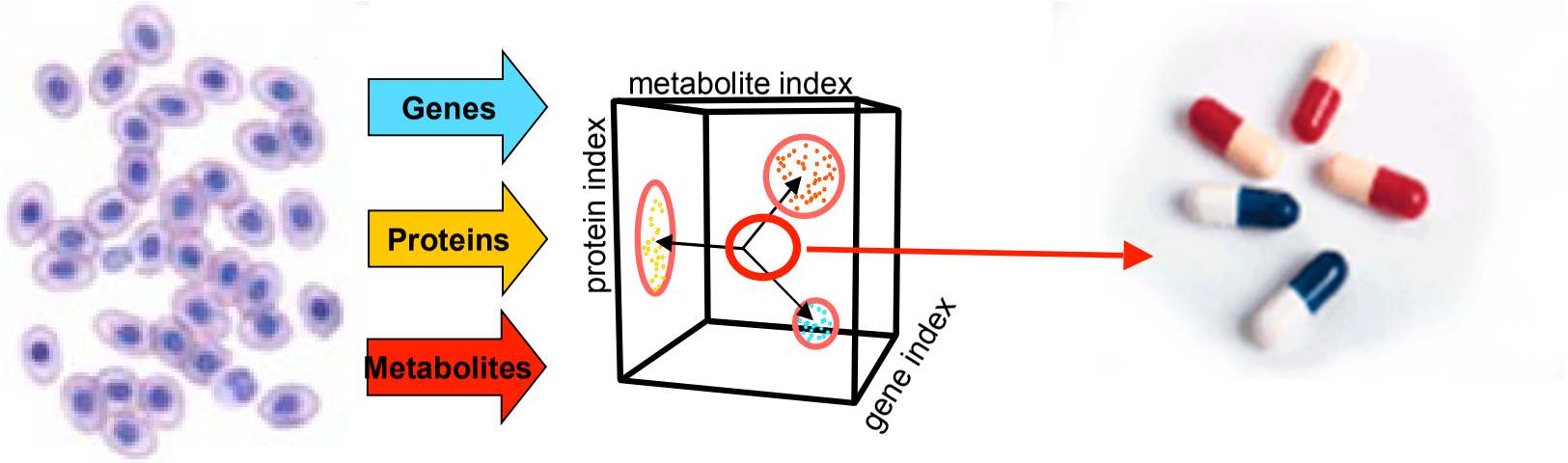
Biological Systems



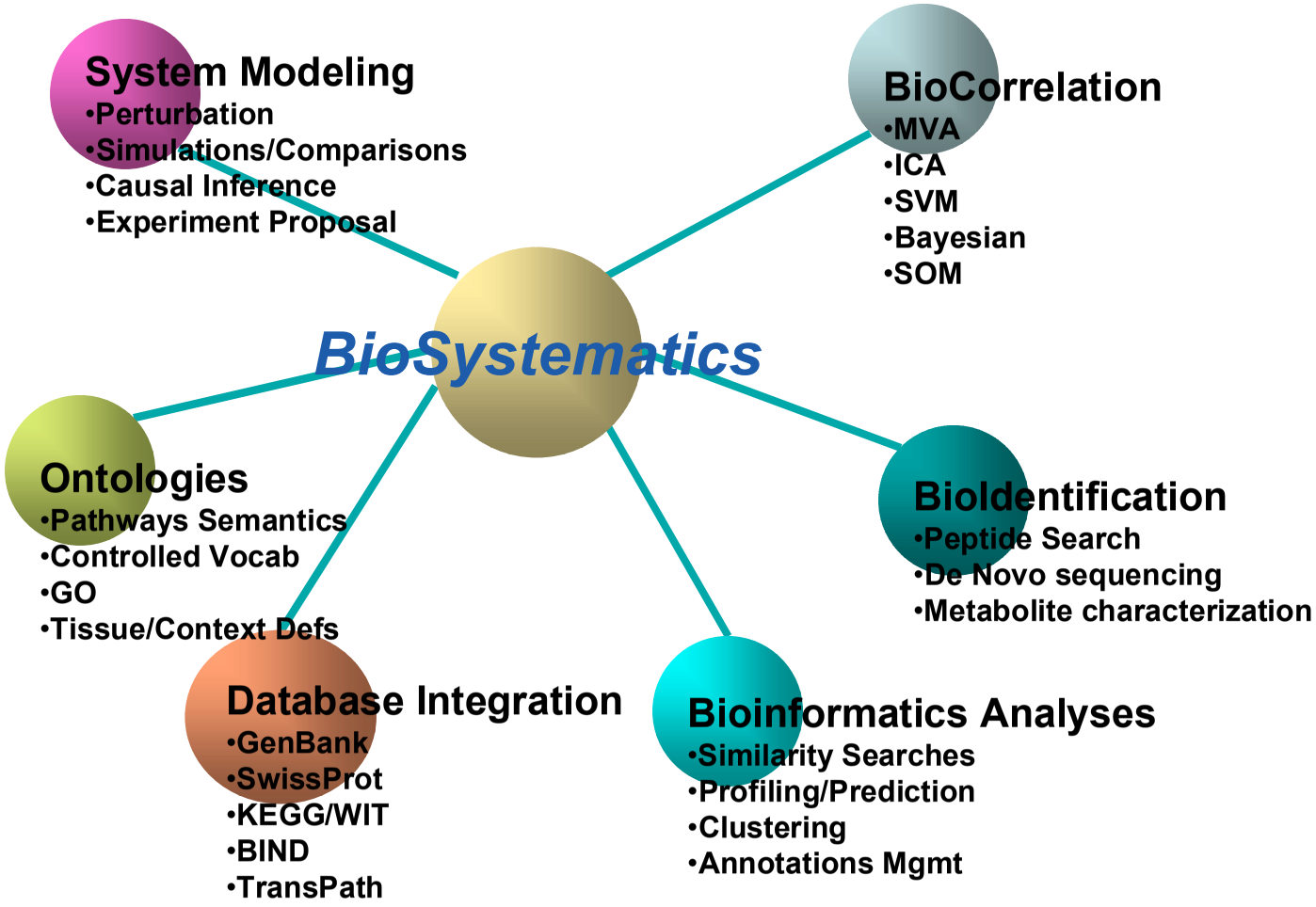
D. Hanahan and R. A. Weinberg. Cell., 100(1):57–70 Review, 2000.

What is Systems Biology?

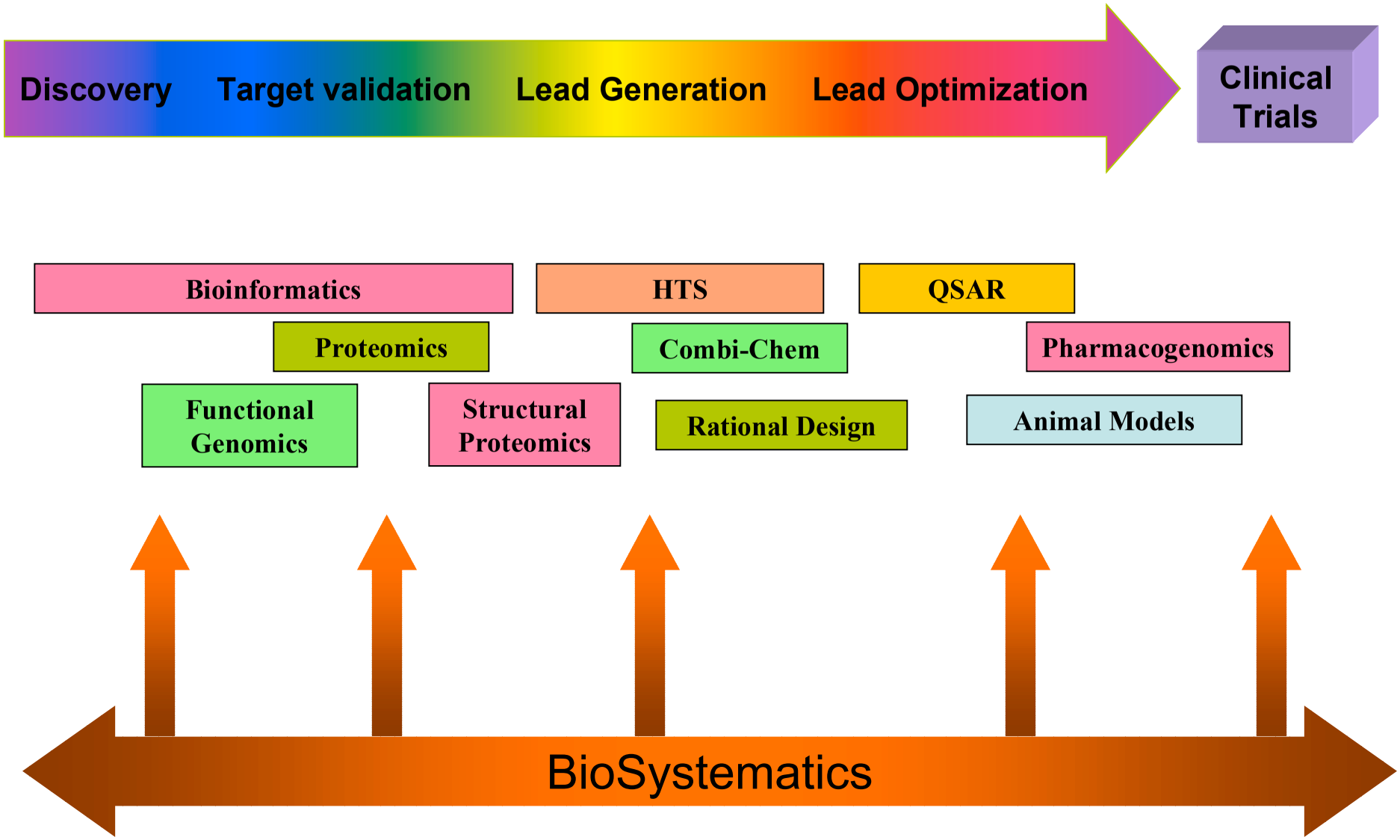
- Parallel analyses of proteins, metabolites and mRNA from complex samples
- Perturbations using mutations, drugs, and experimental conditions
- Determination of molecular function and elucidation of cellular mechanisms
- Informatics tools to link gene response, protein activity and metabolite dynamics
- *BioSystematics* translates covariant sets of genes, proteins, metabolites into *biochemical interaction and target information*



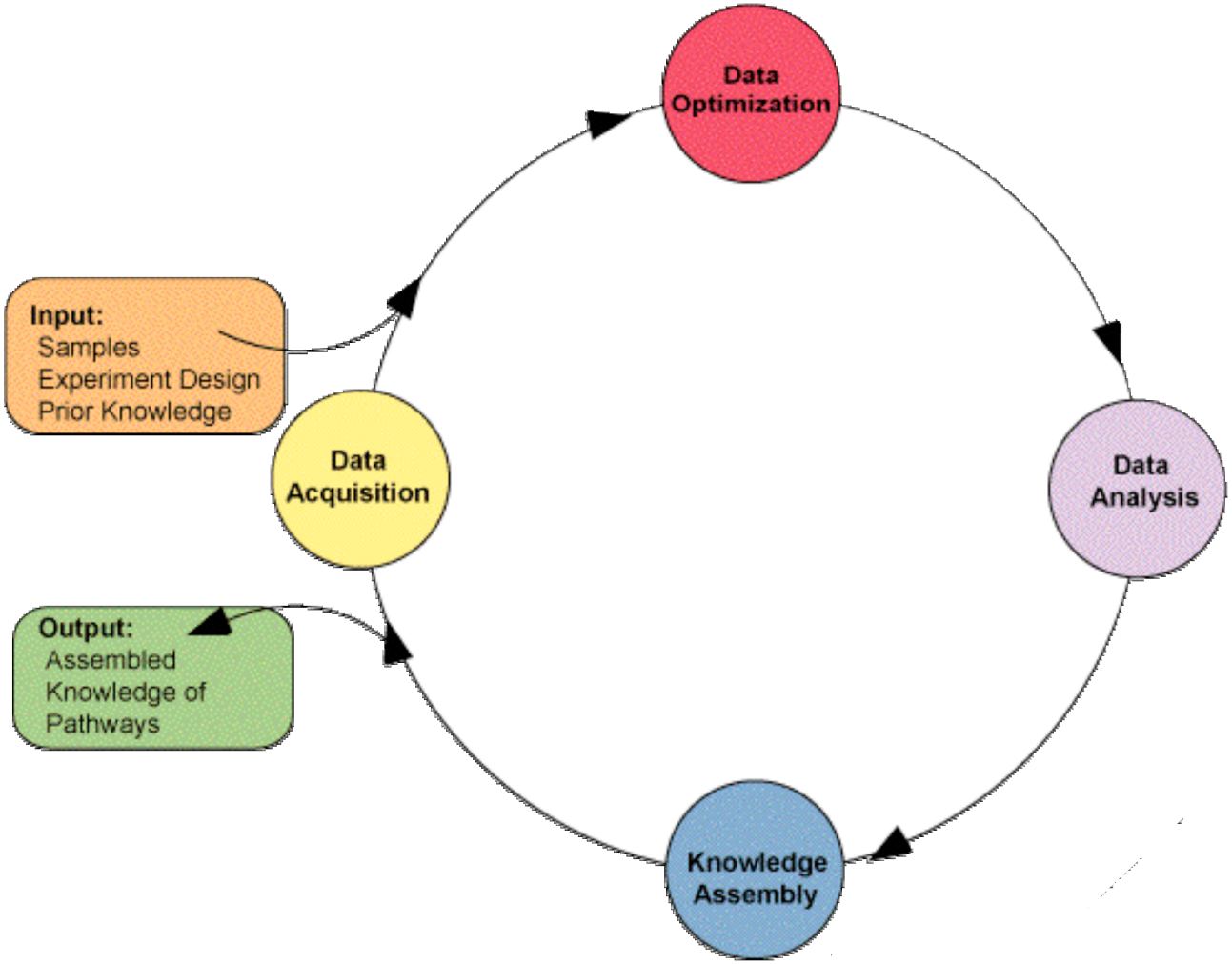
Systems Biology + Bioinformatics =



Role of BioSystematics in Drug Discovery



BioSystematics and the Scientific Method

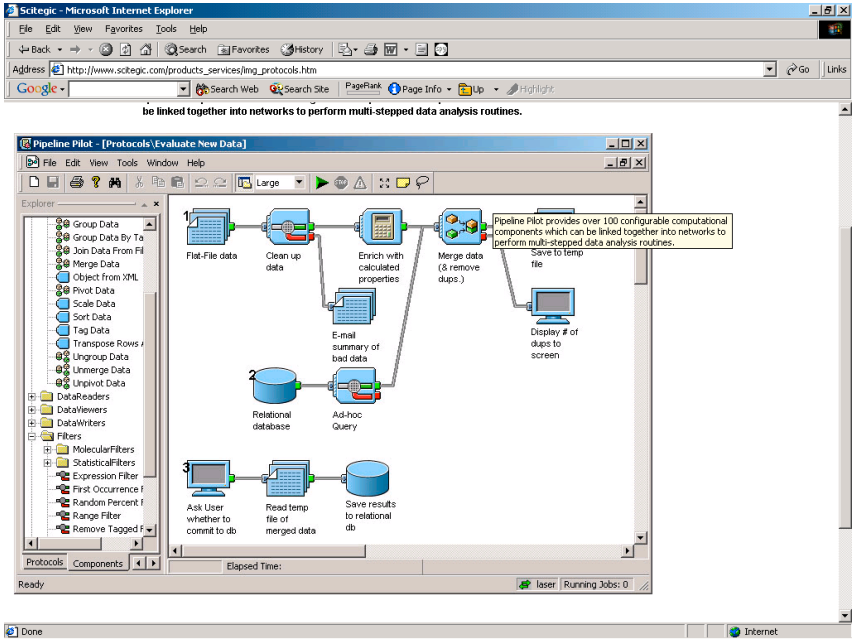


Analysis Pipelines

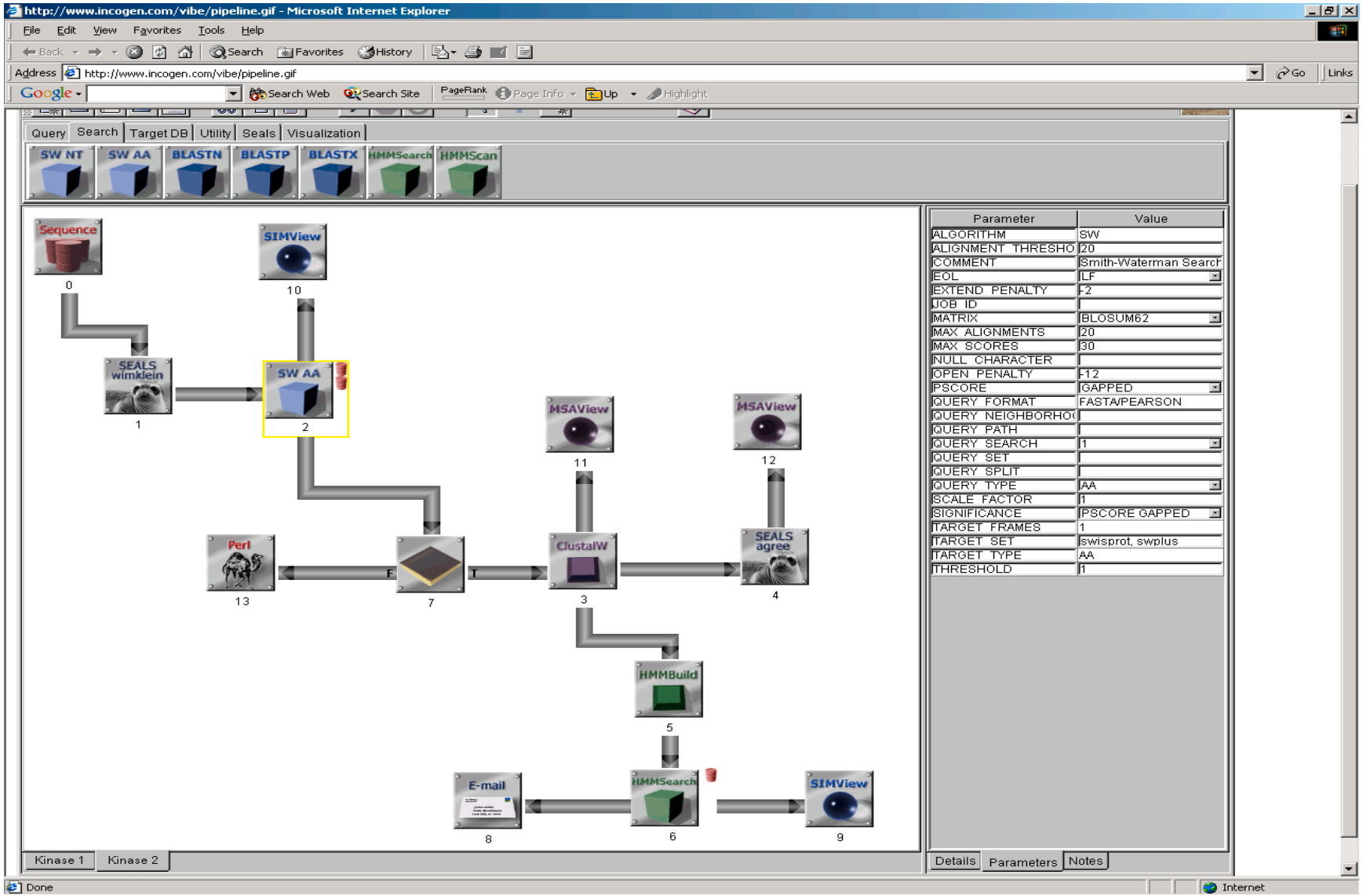
- A real world example: TF identification
- Two private sector approaches:
  - Scitegic
  - Incogen
- Open Source Initiatives

Scitegic

- **What Is Data Pipelining?**
  - The processing, analysis, and mining of large volumes of data through a user-defined computational protocol.
  - Data Pipelining: Data is progressively read into a protocol and analyzed or modified by each component as it flows though the network toward the output endpoints.

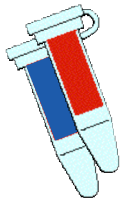


Incogen


http://www.incogen.com/vibe/pipeline.gif - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.incogen.com/vibe/pipeline.gif
Query Search Target DB Utility Seals Visualization
SW NT SW AA BLASTN BLASTP BLASTX HMMSearch HMMScan
Sequence 0
SIMView 10
SEALS wimklein 1
SW AA 2
MSAView 11
MSAView 12
Perl 13
F
T
7
ClustalW 3
SEALS agree 4
HMMBuild 5
E-mail 8
HMMSearch 6
SIMView 9
Kinase 1 Kinase 2
Parameter Value
ALGORITHM SW
ALIGNMENT_THRESHO 20
COMMENT Smith-Waterman Search
EOL LF
EXTEND_PENALTY 2
JOB_ID
MATRIX BLOSUM62
MAX_ALIGNMENTS 20
MAX_SCORES 30
NULL_CHARACTER
OPEN_PENALTY 12
PSCORE GAPPED
QUERY_FORMAT FASTA/PEARSON
QUERY_NEIGHBORHOO
QUERY_PATH
QUERY_SEARCH 1
QUERY_SET
QUERY_SPLIT
QUERY_TYPE AA
SCALE_FACTOR 1
SIGNIFICANCE PSCORE GAPPED
TARGET_FRAMES 1
TARGET_SET swisprot, swplus
TARGET_TYPE AA
THRESHOLD 1
Details Parameters Notes
Done
Internet

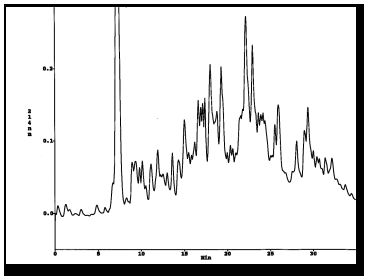
Proteomics Analysis

- control sample
- experimental sample

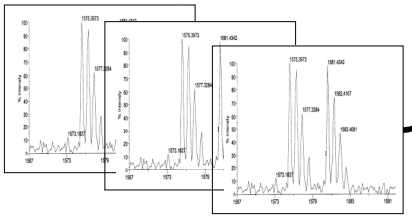


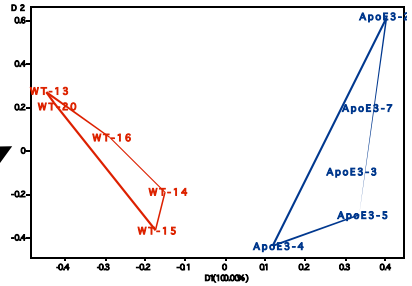
sample preparation chemistry / isotopic labeling

multi-dimensional chromatographic separation

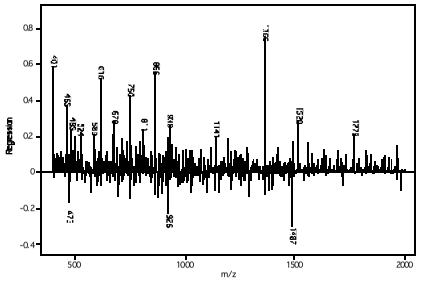


collection and mass spectrometry

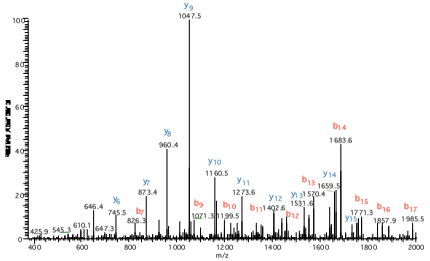




clustering of datasets

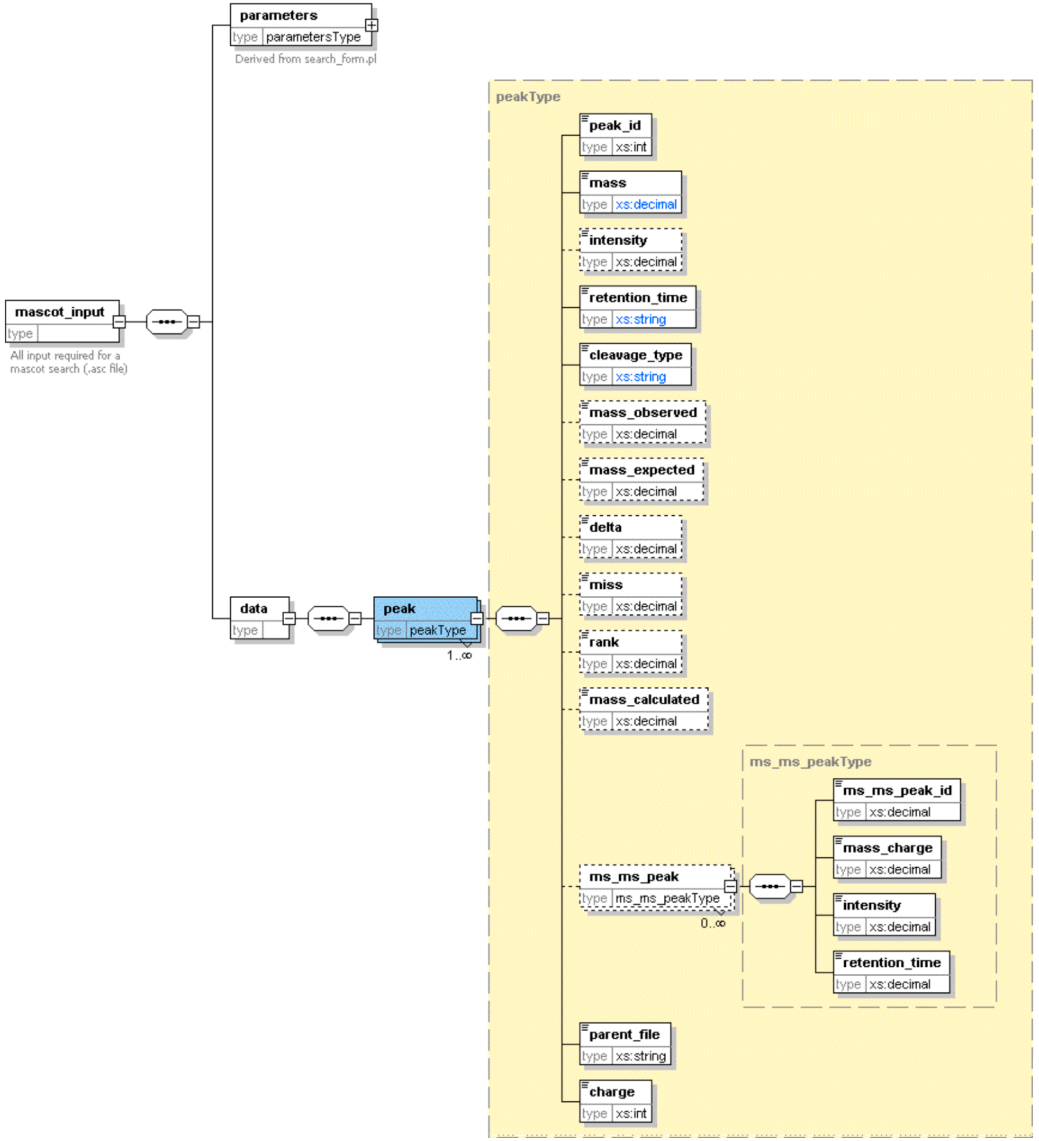


pattern recognition of significant peptide differences



MS/MS identification of relevant peptides

MS Peaks and Proteins XML



Metabolomics

NMR

MALDI

LC-MS

GC-MS

2-D LCMS

2-D GCMS

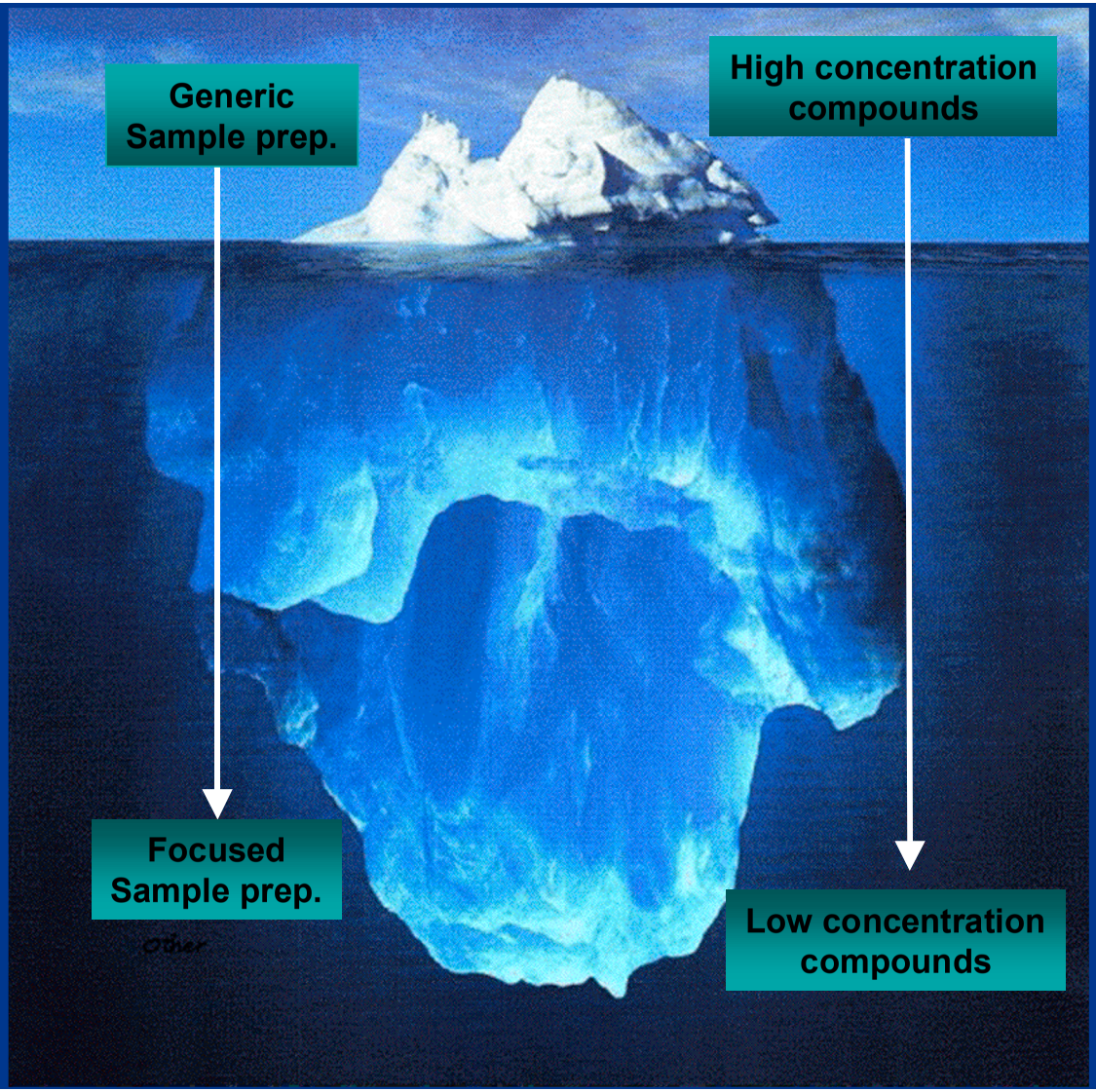
LC-MS/MS

LC-ECD

LC-LIF

ICP-MS

….



Homeostasis 'Housekeeping'

Organic Acids

Lipids

Amino Acids

Nucleotides

Steroids

Eicosanoids

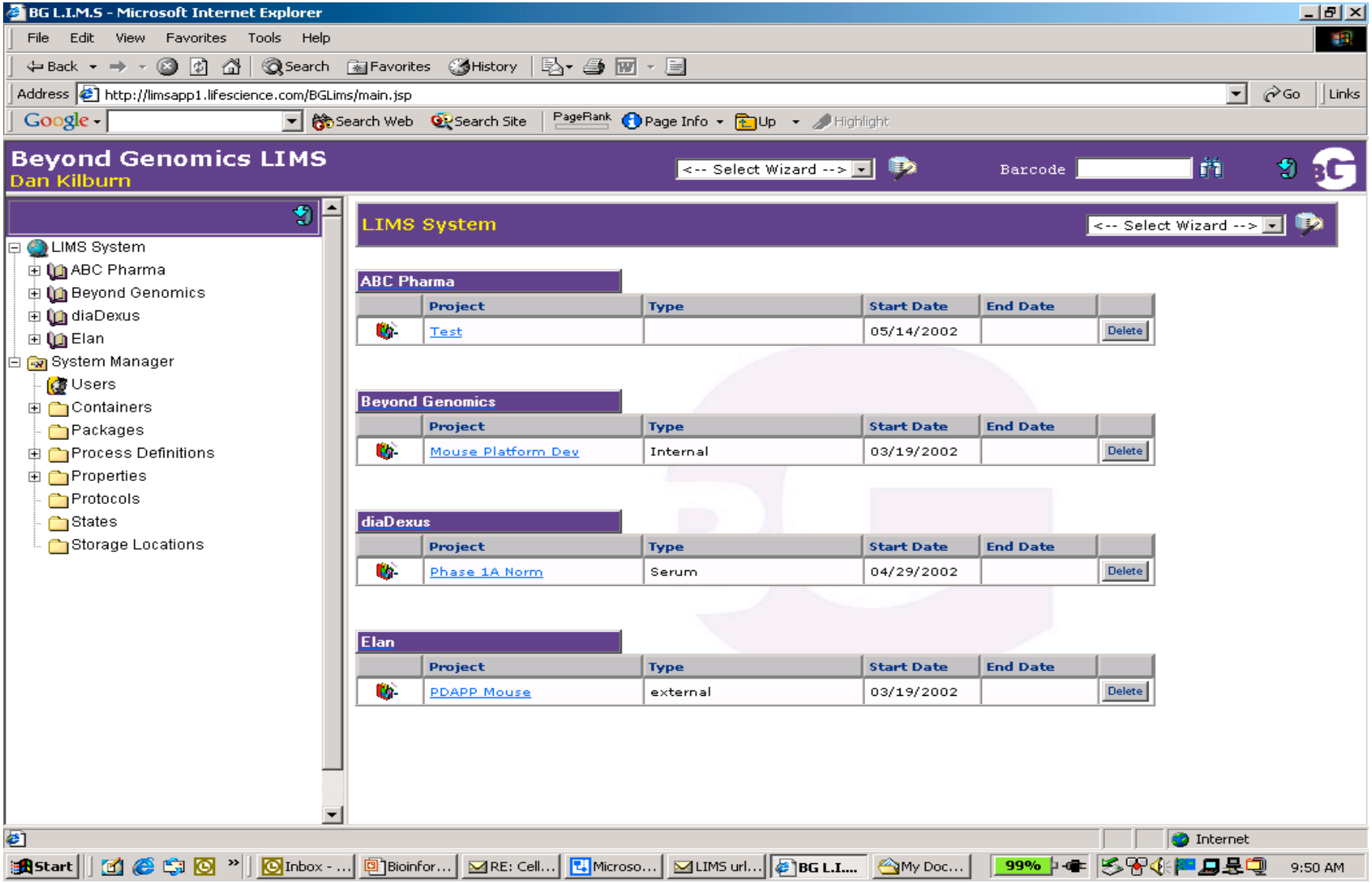
Neurotransmitters

Peptides

Trace elements

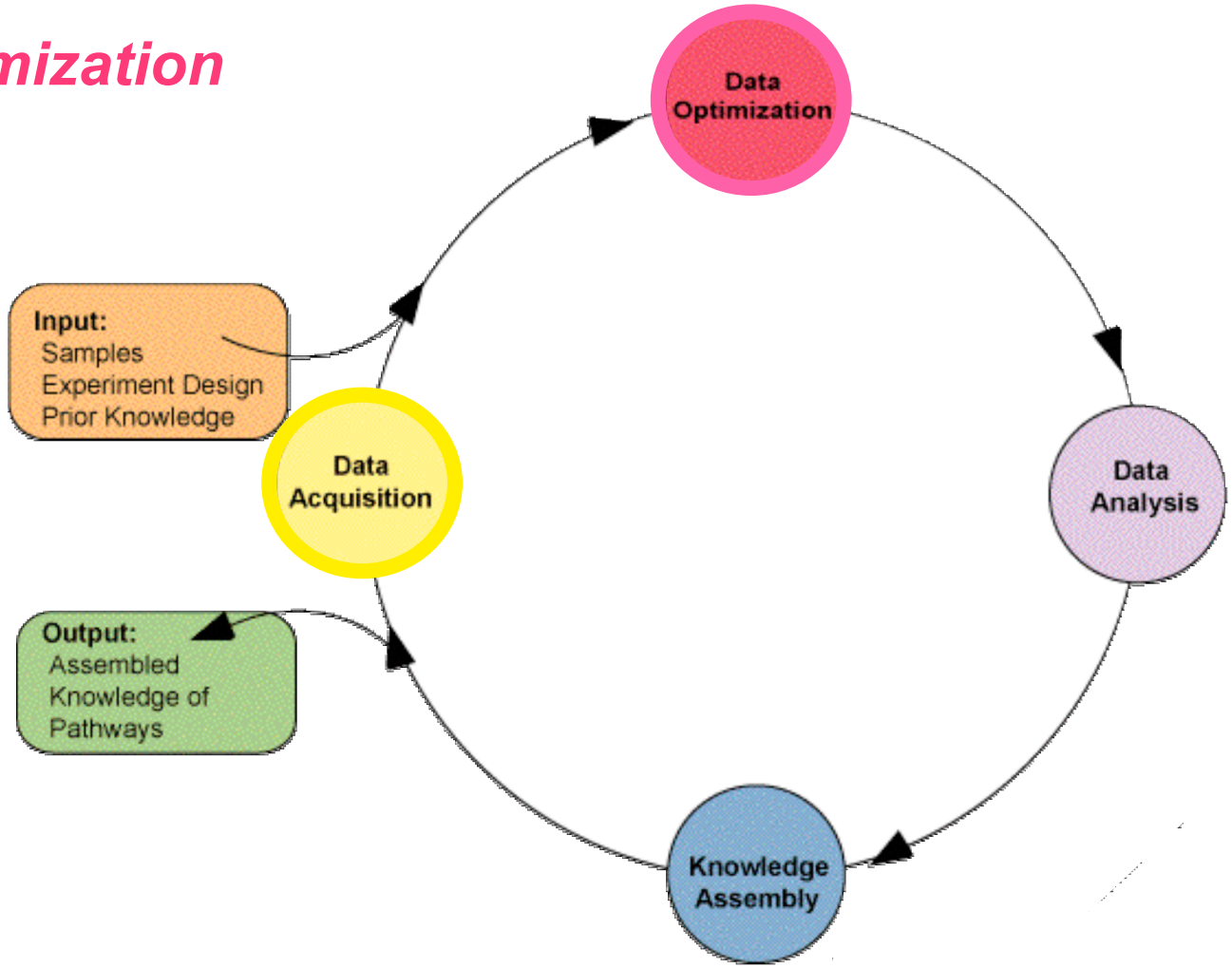
….

Data Acquisition: LIMS *Expt Design*



BioSystematics™ at Beyond Genomics:

Data Optimization

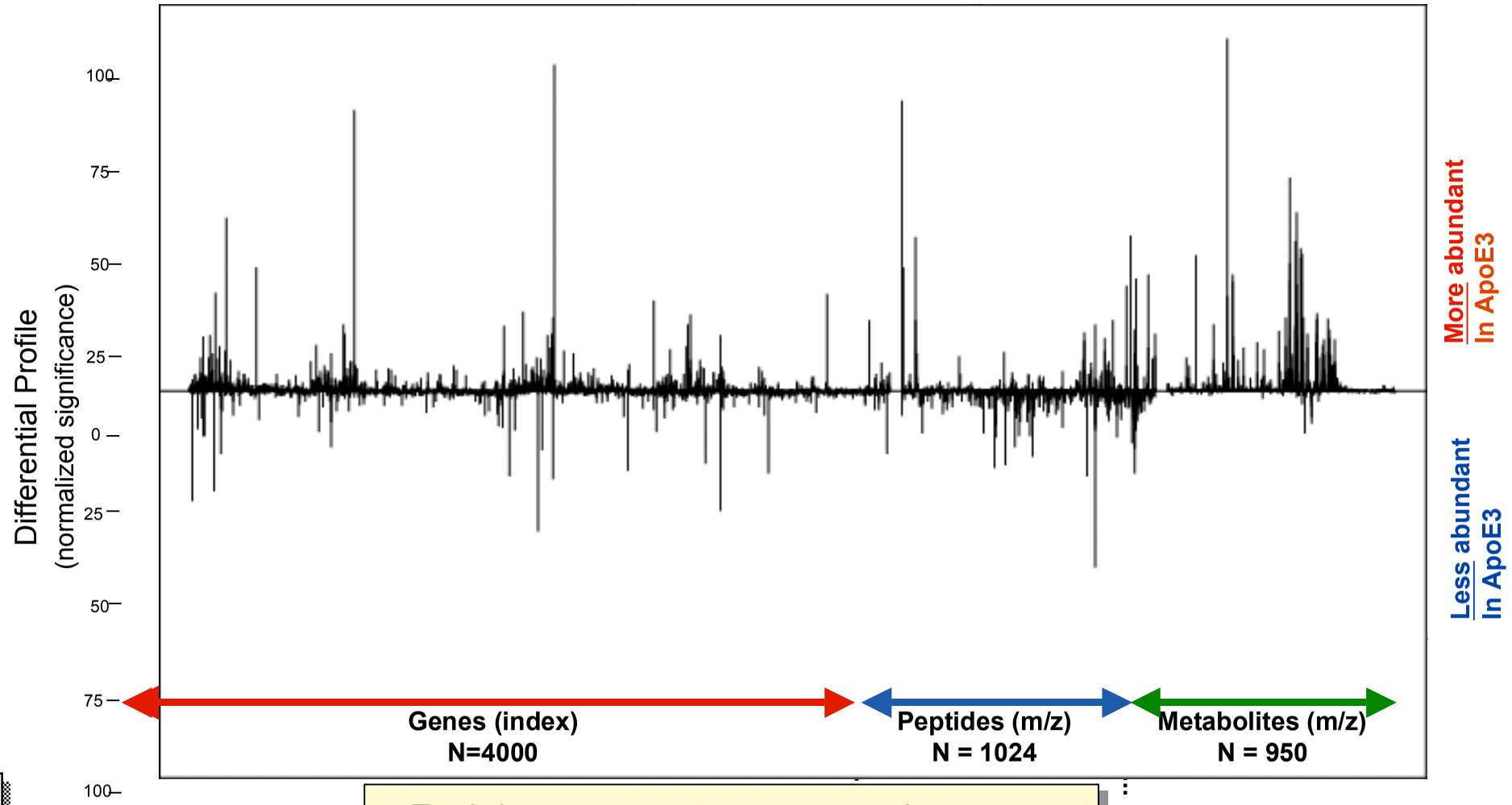


Data Optimzation: Quality Control

Data *always* includes:

- Error1 *Measurement*
- Error2 *Integrity*
- Noise *Thermal*
- Randomness *Markov Process*
- Variance **Inherent Structure**

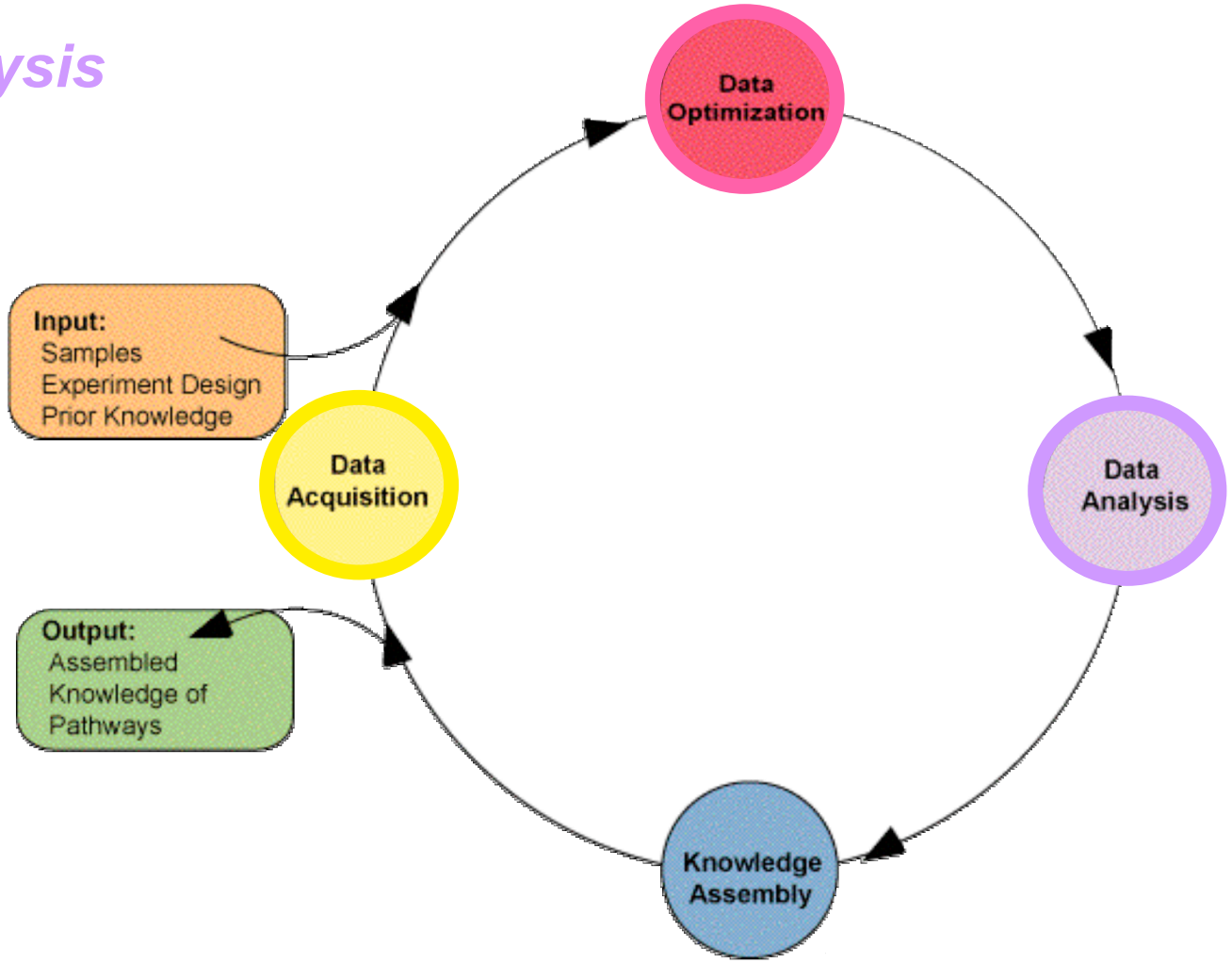
Proteomic, Metabolomic, Genomic— Integrated Differential Expression Profile



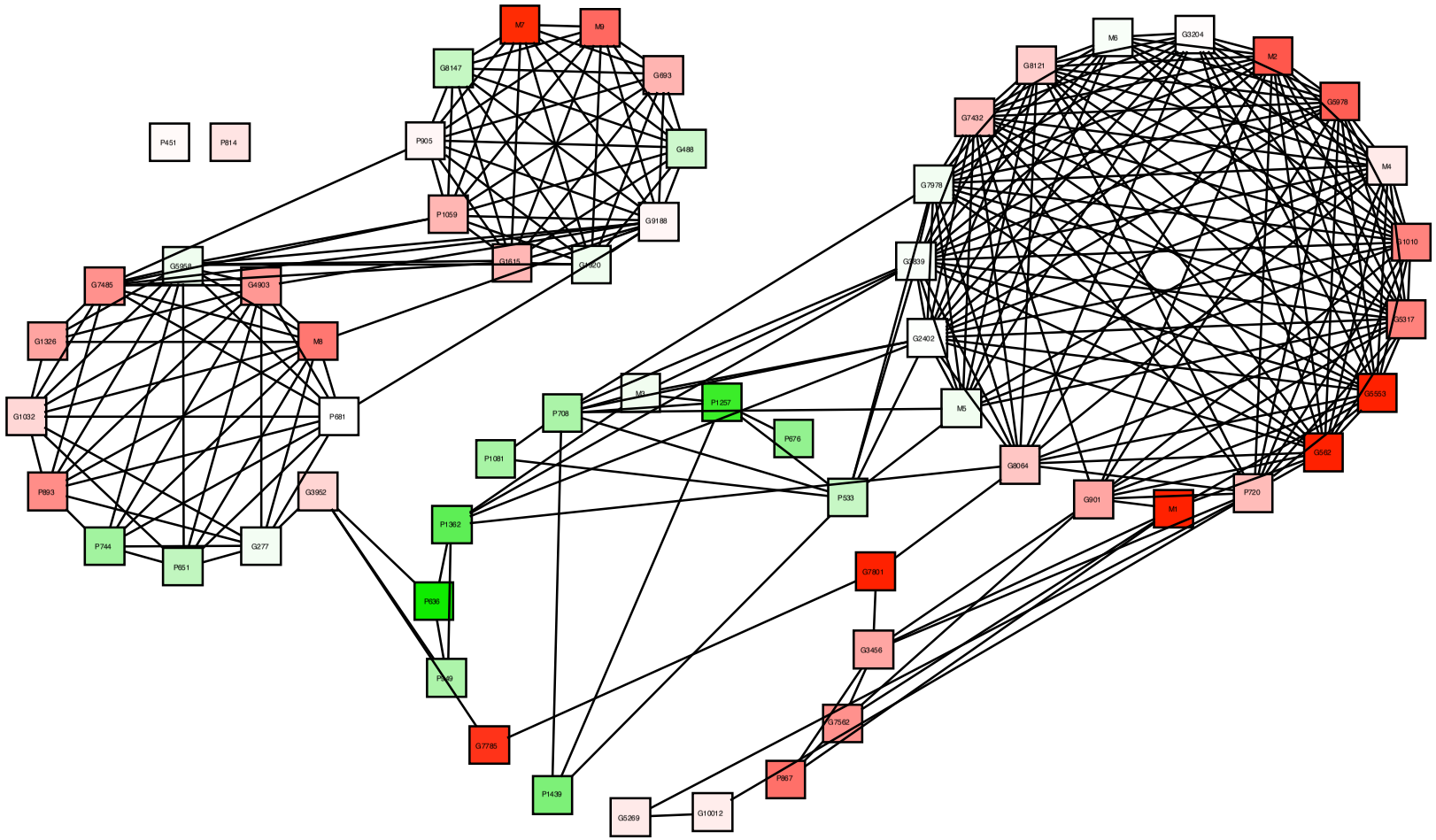
- Each bar represents a measured component
  1. **fold-change / variance value**
  2. **confidence value p < 0.05**

BioSystematics™ at Beyond Genomics:

Data Analysis

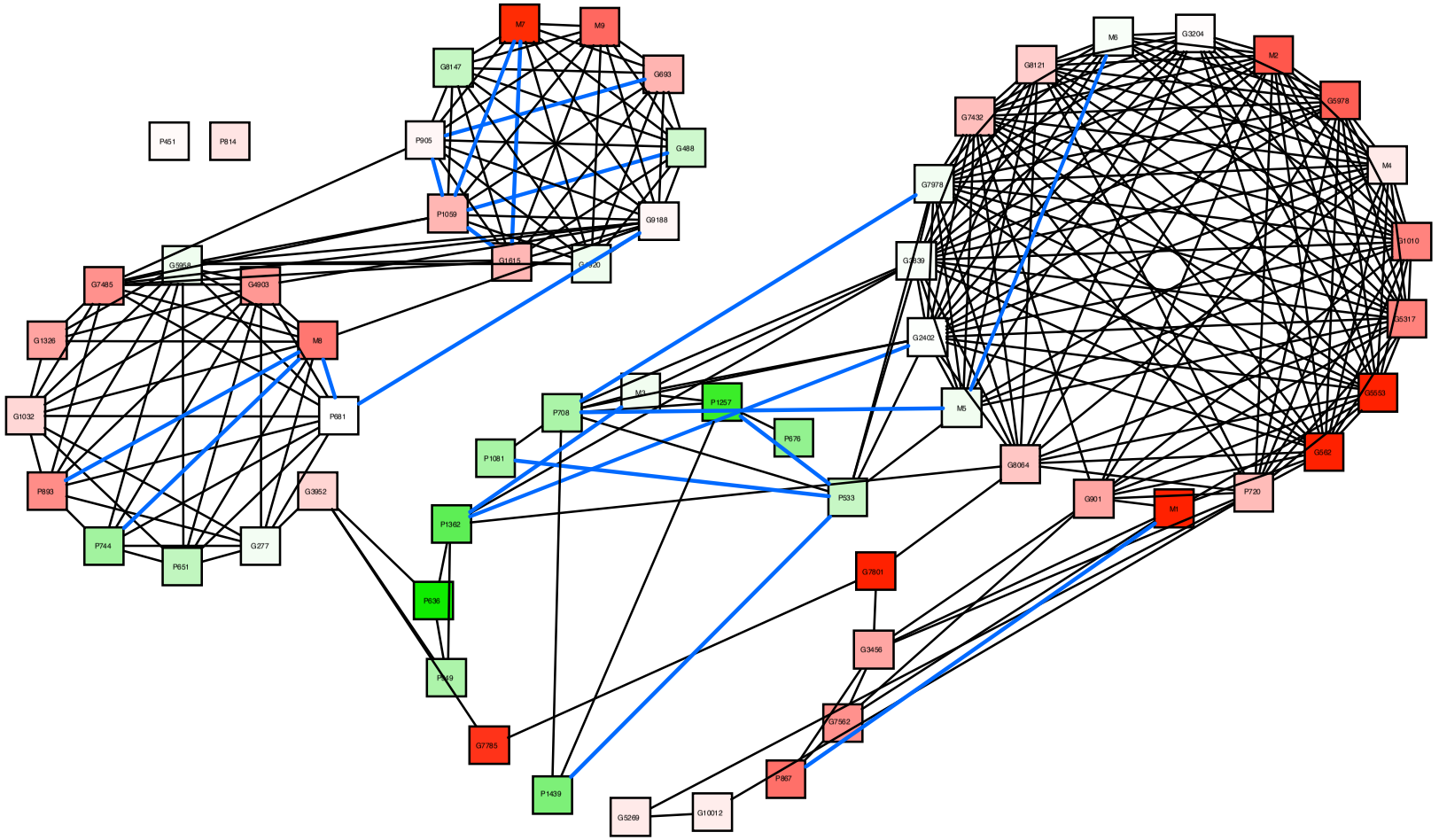


Association Network Analysis



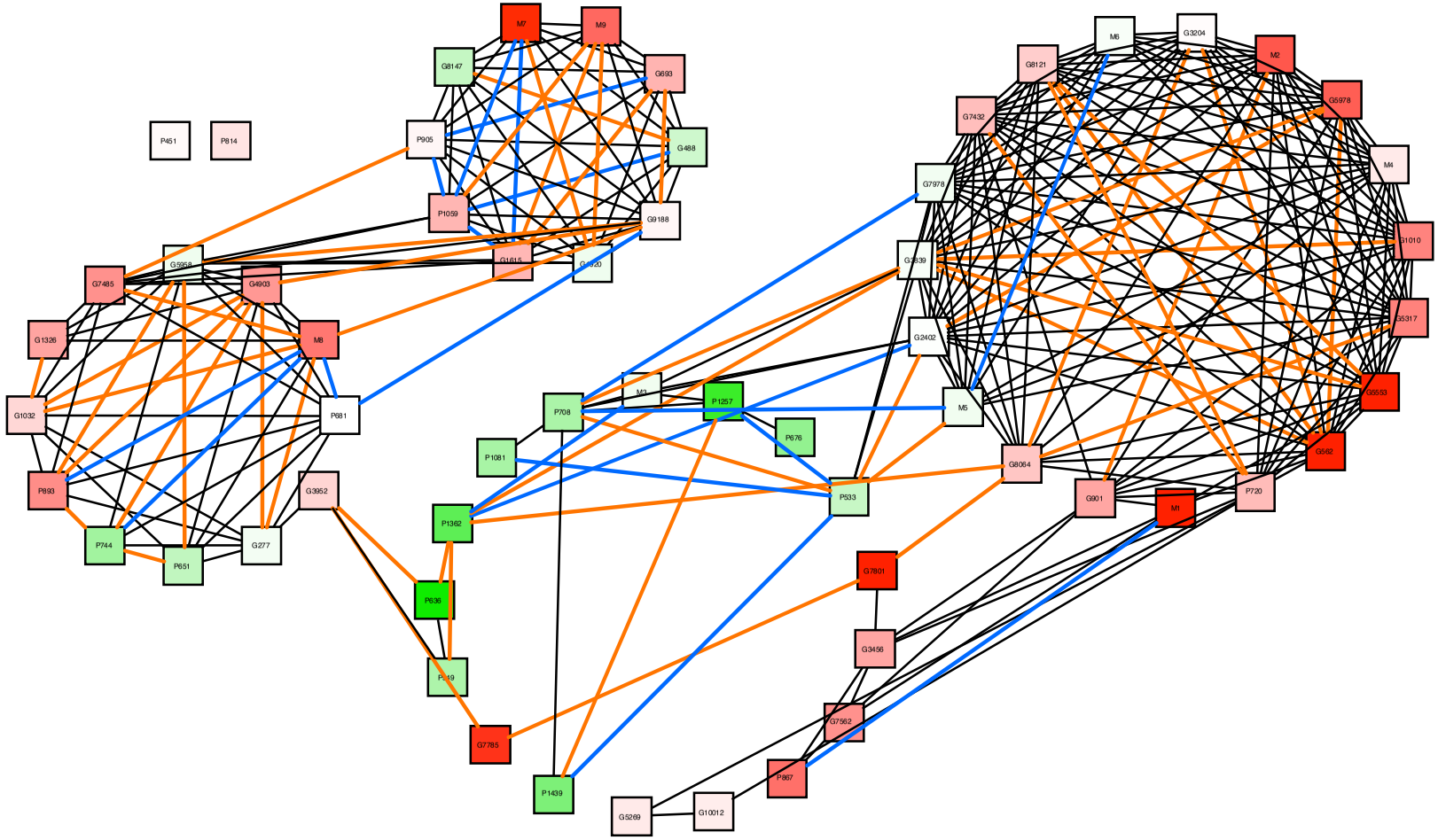
Nodes connected based on high association using linear correlations, kernel PCA, or mutual information.

Association Network Analysis



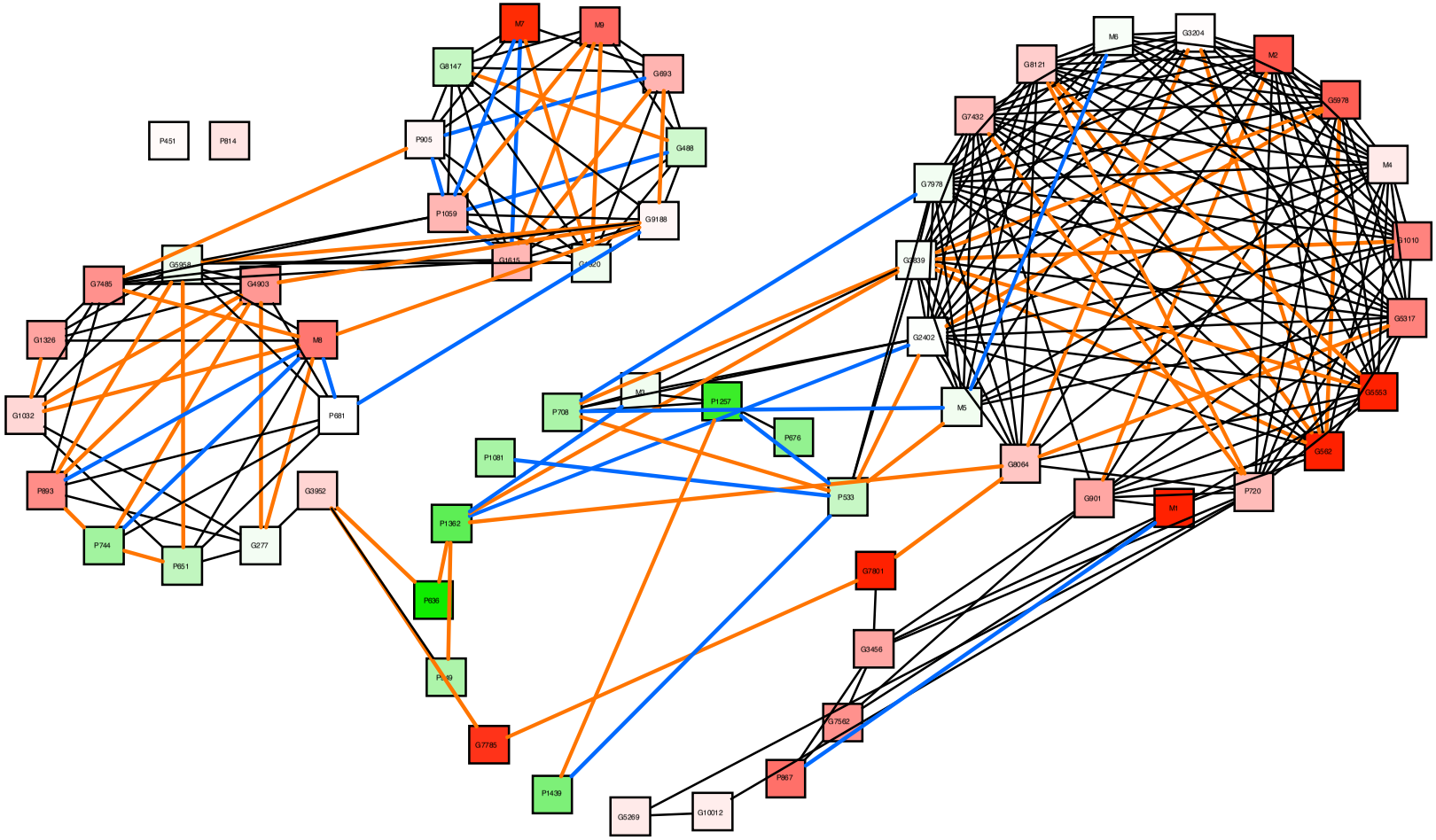
Annotation of edges using information from databases

Association Network Analysis



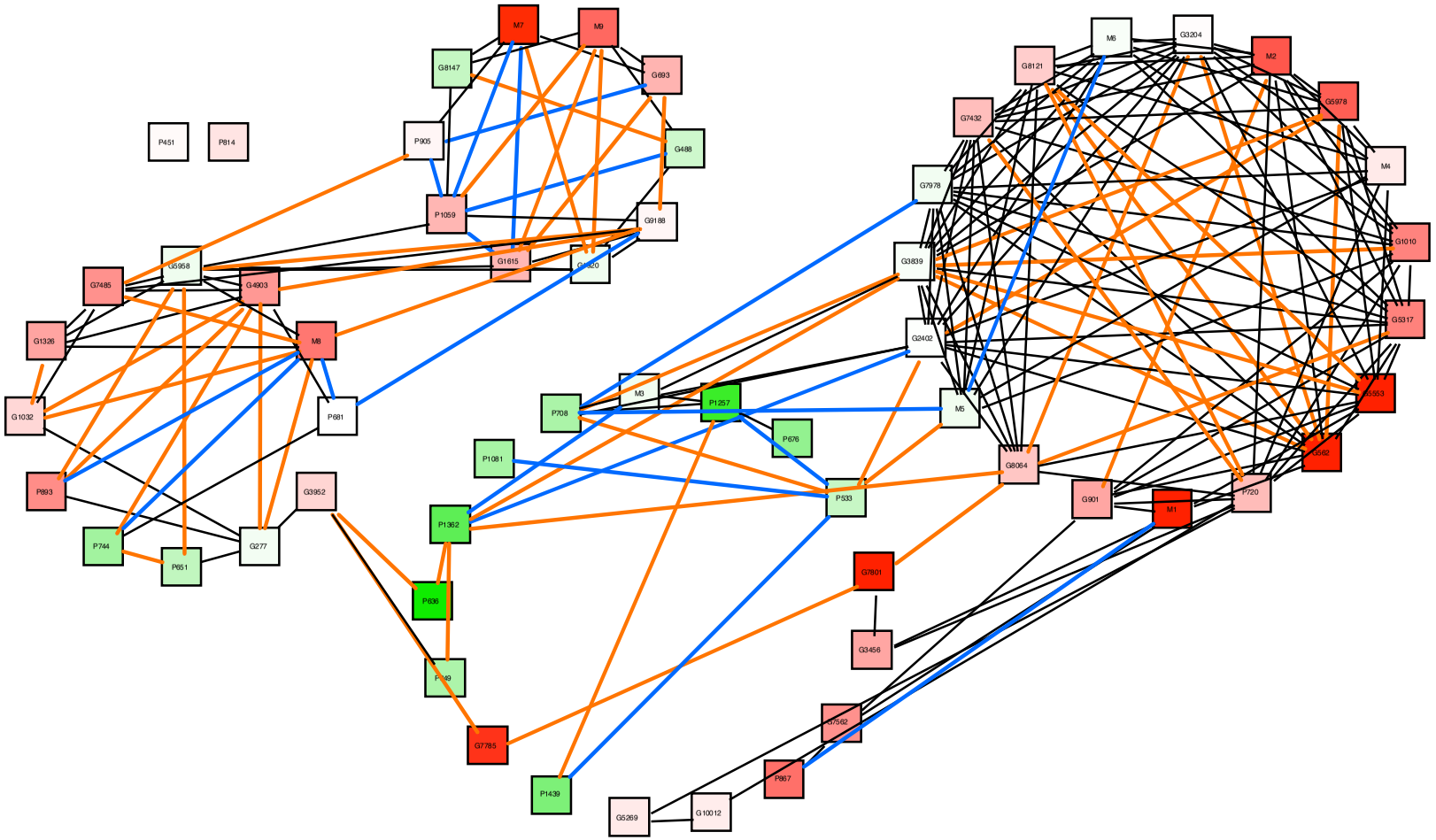
Annotation of edges using information from literature:
Co-occurences of gene/protein/metabolite names in literature

Association Network Analysis



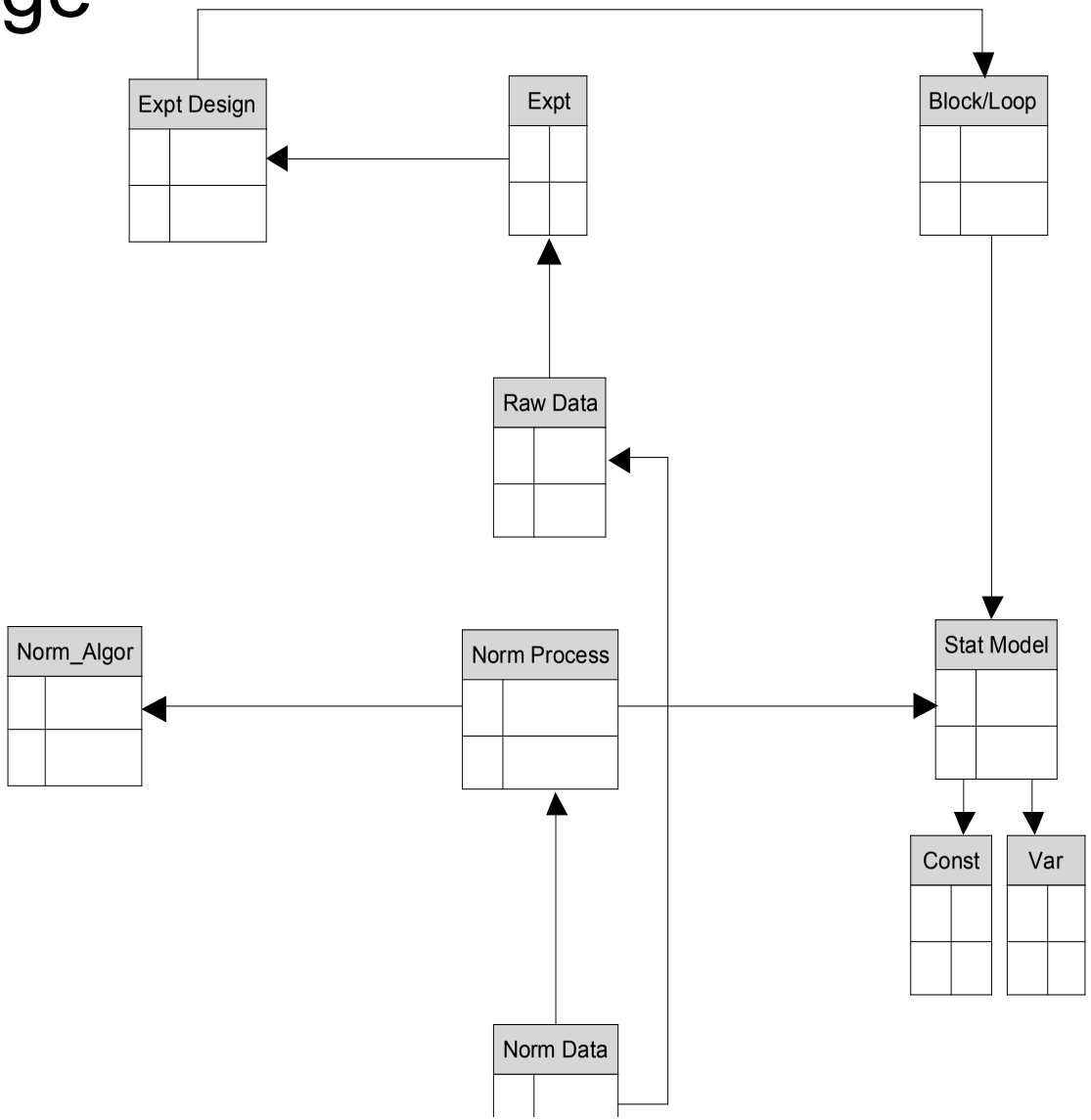
Elimination of redundant edges based on observational data

Association Network Analysis



We are left with known and unknown associations. If no further "cleanup" is possible, need to start model building and discovery from here (Network analysis, new hypothesis building, further experiments)

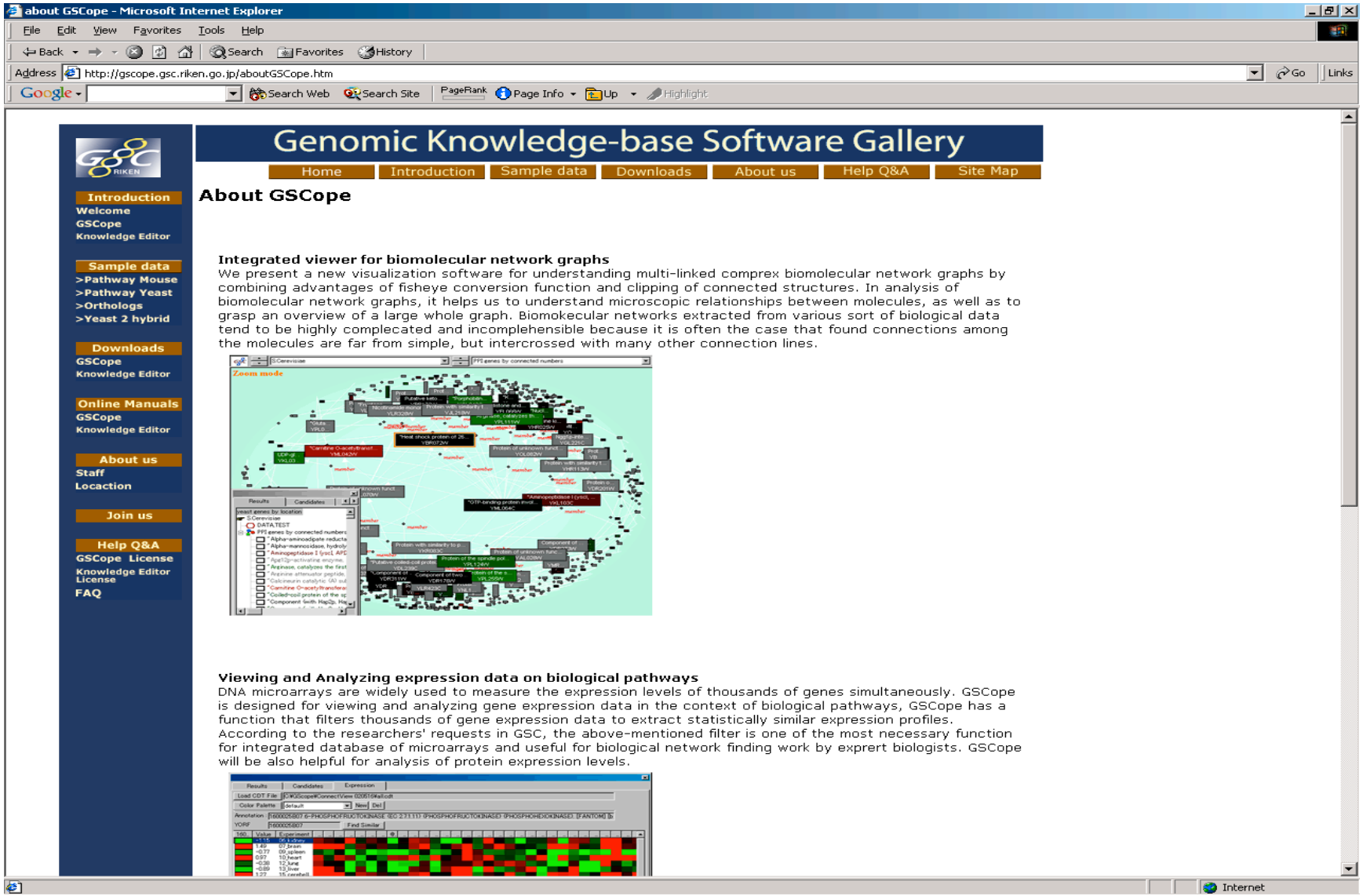
Statistical Model and Markup Language



Pathway and Connection Visualization

- Bioinformatics + Pathways
  - Incellico
  - Genstruct
- Riken
- LSID – Common unique indexing system for mixing and merging heteroegenous data in both views and knowledge … ala DAS!
  - Ex: KEGG with LSID: Cmp(CAS), Rxn (EC), Enz (GB)

Riken


about GSCope - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://gscope.gsc.riken.go.jp/aboutGSCope.htm
Go Links
Google Search Web Search Site PageRank Page Info Up Highlight
Genomic Knowledge-base Software Gallery
Home Introduction Sample data Downloads About us Help Q&A Site Map
Introduction
Welcome
GSCope
Knowledge Editor
Sample data
>Pathway Mouse
>Pathway Yeast
>Orthologs
>Yeast 2 hybrid
Downloads
GSCope
Knowledge Editor
Online Manuals
GSCope
Knowledge Editor
About us
Staff
Locaction
Join us
Help Q&A
GSCope License
Knowledge Editor License
FAQ
About GSCope
Integrated viewer for biomolecular network graphs
We present a new visualization software for understanding multi-linked comprex biomolecular network graphs by combining advantages of fisheye conversion function and clipping of connected structures. In analysis of biomolecular network graphs, it helps us to understand microscopic relationships between molecules, as well as to grasp an overview of a large whole graph. Biomokecular networks extracted from various sort of biological data tend to be highly complecated and incomplehensible because it is often the case that found connections among the molecules are far from simple, but intercrossed with many other connection lines.
Zoom mode
Viewing and Analyzing expression data on biological pathways
DNA microarrays are widely used to measure the expression levels of thousands of genes simultaneously. GSCope is designed for viewing and analyzing gene expression data in the context of biological pathways, GSCope has a function that filters thousands of gene expression data to extract statistically similar expression profiles. According to the researchers' requests in GSC, the above-mentioned filter is one of the most necessary function for integrated database of microarrays and useful for biological network finding work by exprert biologists. GSCope will be also helpful for analysis of protein expression levels.
Internet

Data Analysis: CELL from Incellico

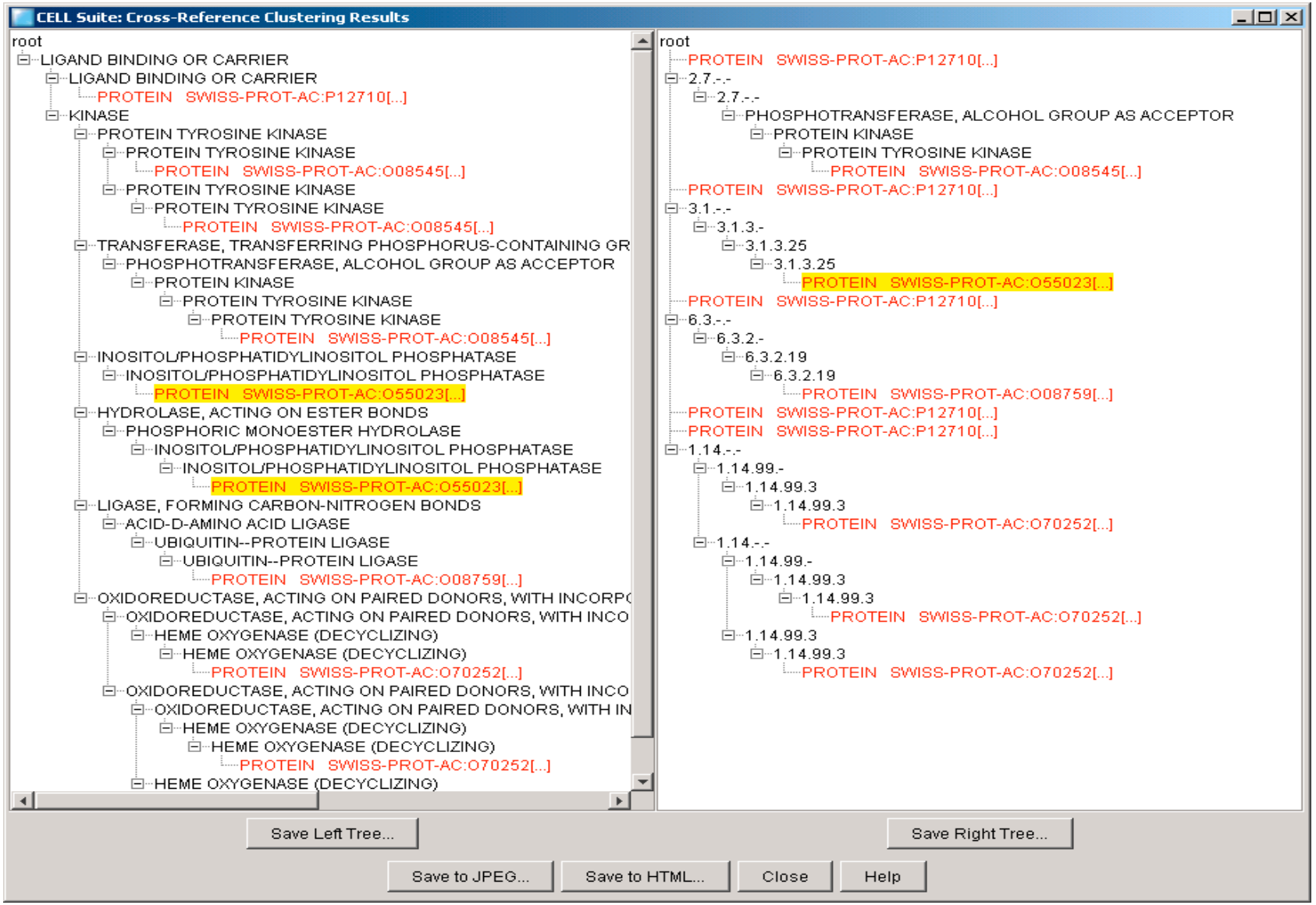
- A principal technology for knowledge integration is Incellico's CELL, a database of relationships between databases
- CELL allows experimental findings to be constrained and enhanced by prior knowledge
- CELL supports data navigation and high-throughput data analysis.

Incellico
Analytic software for the life sciences

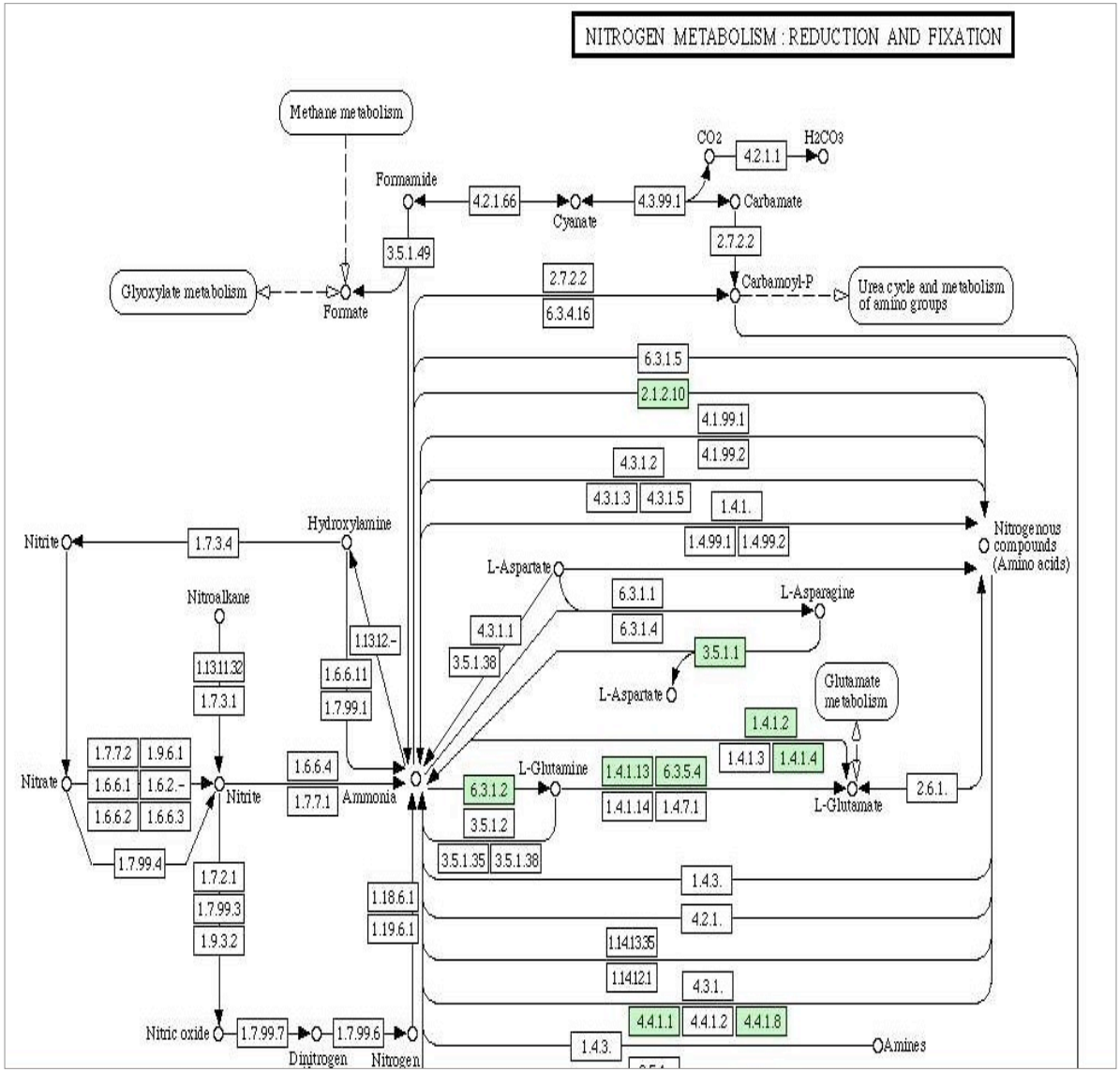
Databases Linked:

- OMIM
- SWISS-PROT-AC
- FLYBASE-ID
- PDB
- SWISS-PROT-ID
- GDB-ID
- PFAM-DESC
- SYMBOL
- GENE-Name
- PFAM-ID
- TAX-SCIENTIFIC-NAME
- GENPEPT
- PIR-ID, TAXID
- GENPEPT-ID
- PRINTS-DESC
- UNIGENE-ID
- GO
- PRINTS-ID
- CHROMOSOME
- GO-COMPONENT-DESC
- PRODOM-DESC
- CHROMOSOME-ARM
- GO-FUNCTION-DESC
- PRODOM-ID
- CHROMOSOME-BAND
- INTERPRO-DESC
- PROSITE-DESC
- CHROMOSOME-SUB-BAND
- INTERPRO-ID,
- PROSITE-ID
- EMBL/GENBANK
- KEYWORD,
- PUBMED
- EMBL/GENBANK-GI
- LOCUSLINK
- RATMAP-ID
- EMBL/GENBANK-LOCUS
- MEDLINE,
- REFSEQ
- EMBL/GENBANK
- VERSION
- MGI-ID
- SGD-ID
- EXPRESS-IN

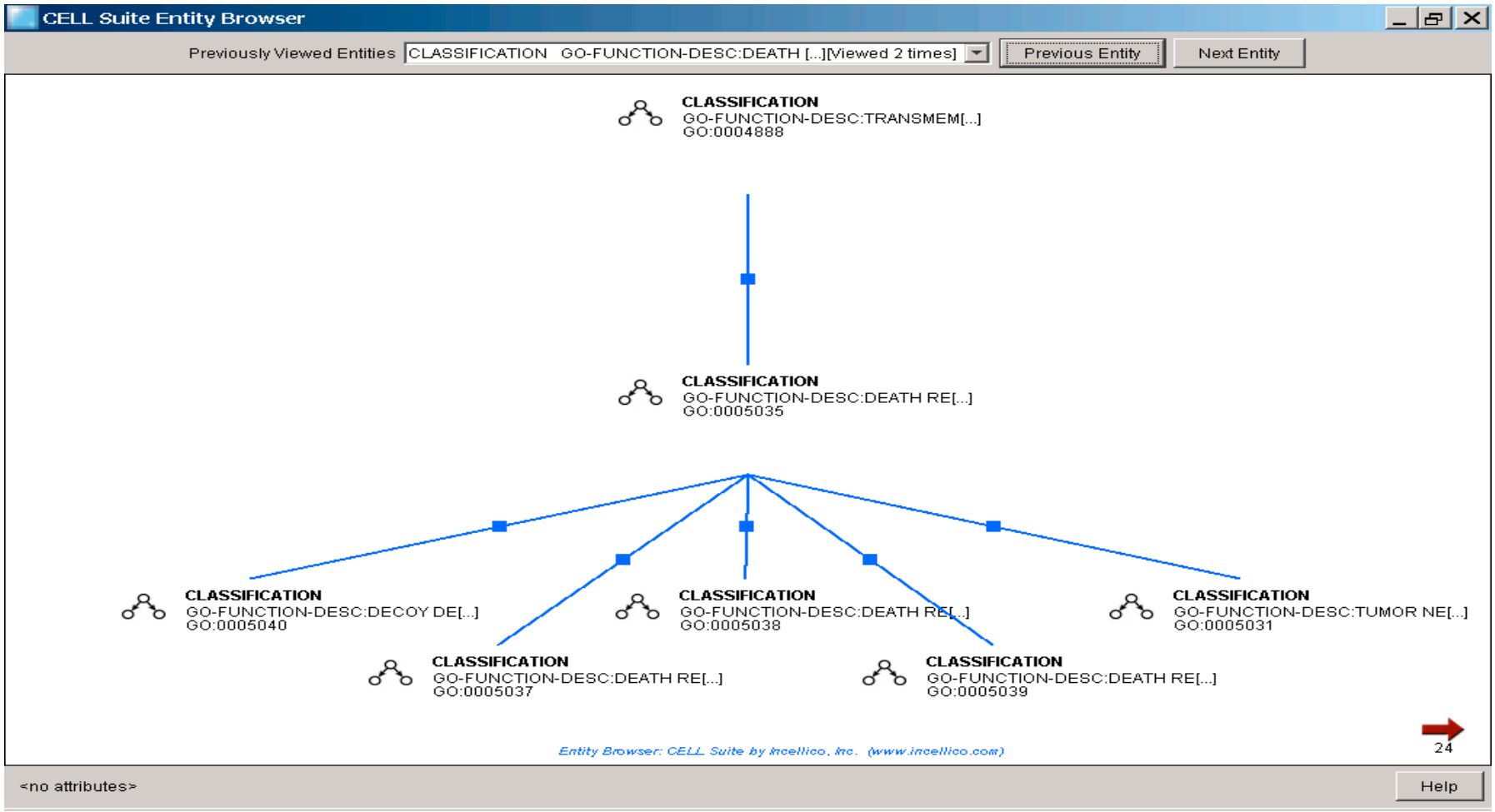
Incellico links GO Data to…



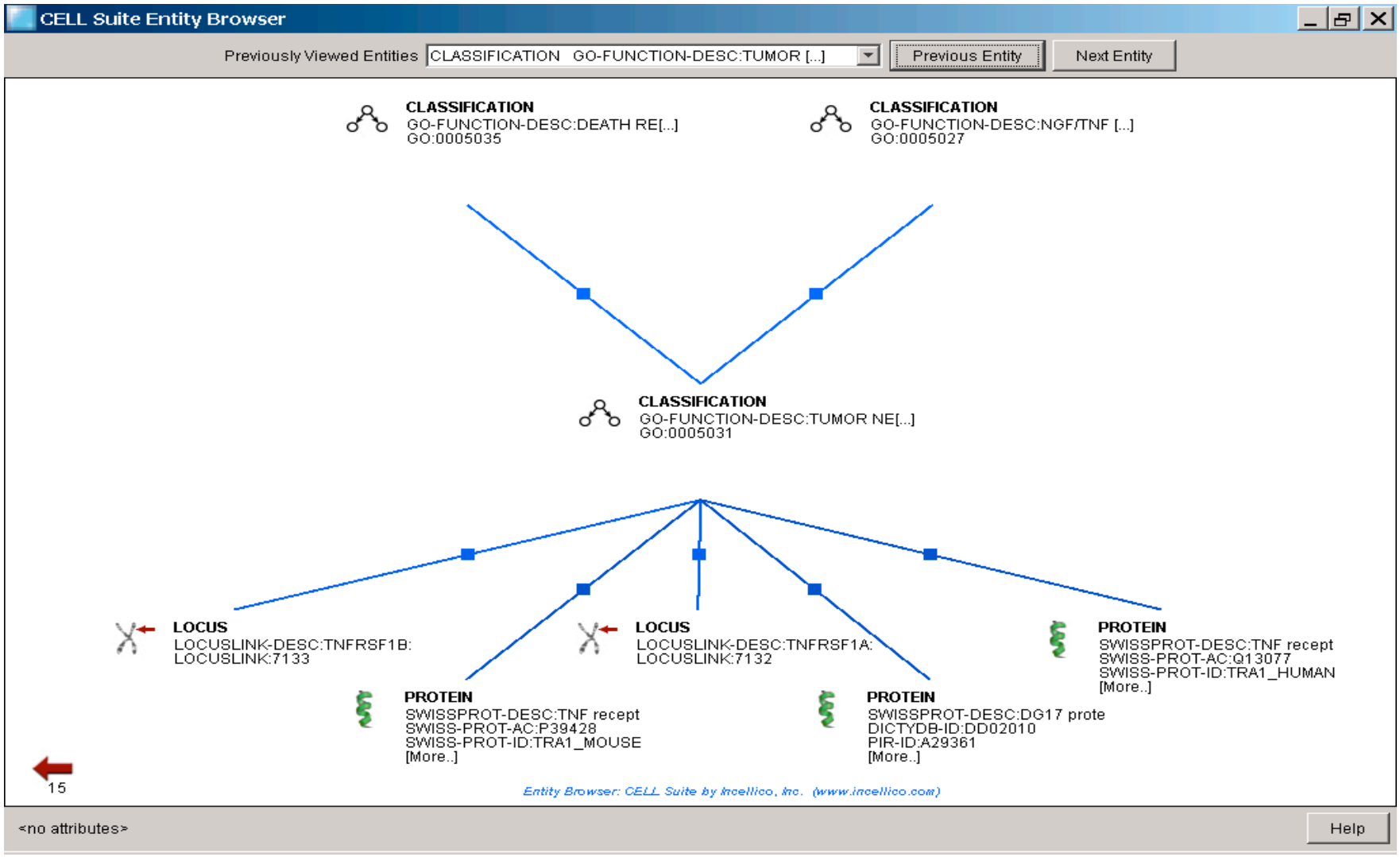
KEGG Metabolic Pathways



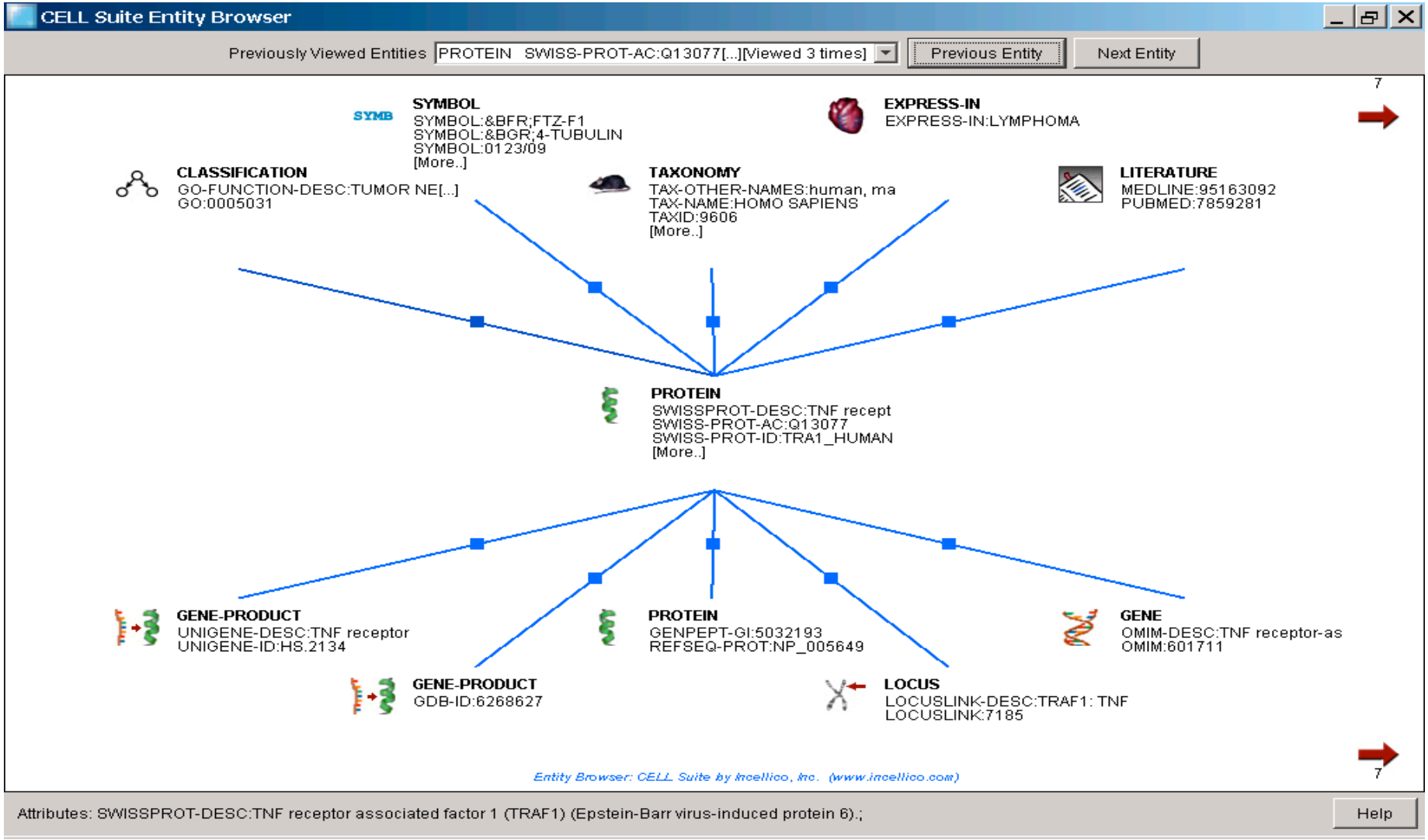
Cell Search: Go Molecular Function: Death Receptor



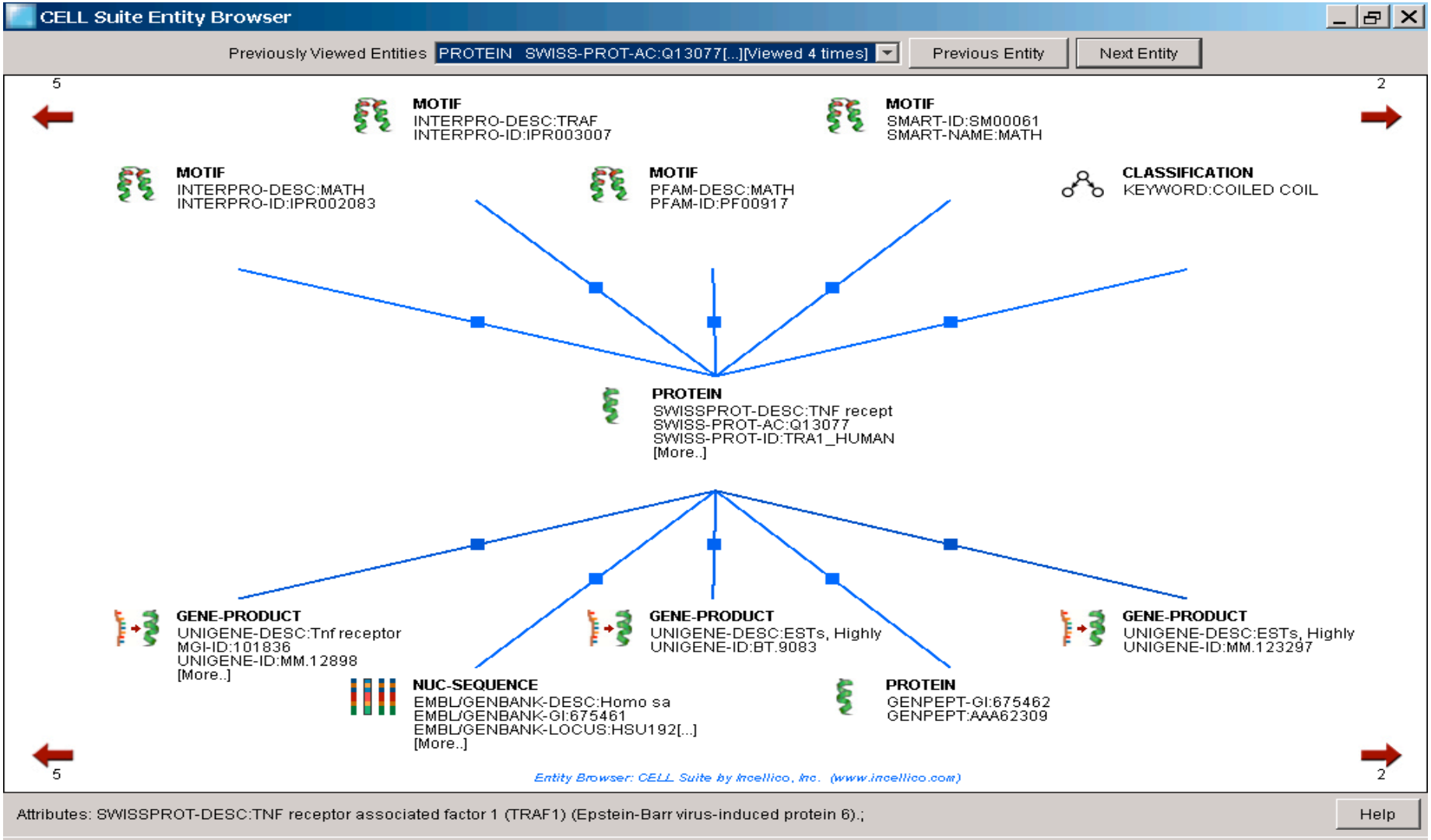
Drill for Tumor Necrosis Factor, a specific Death Receptor



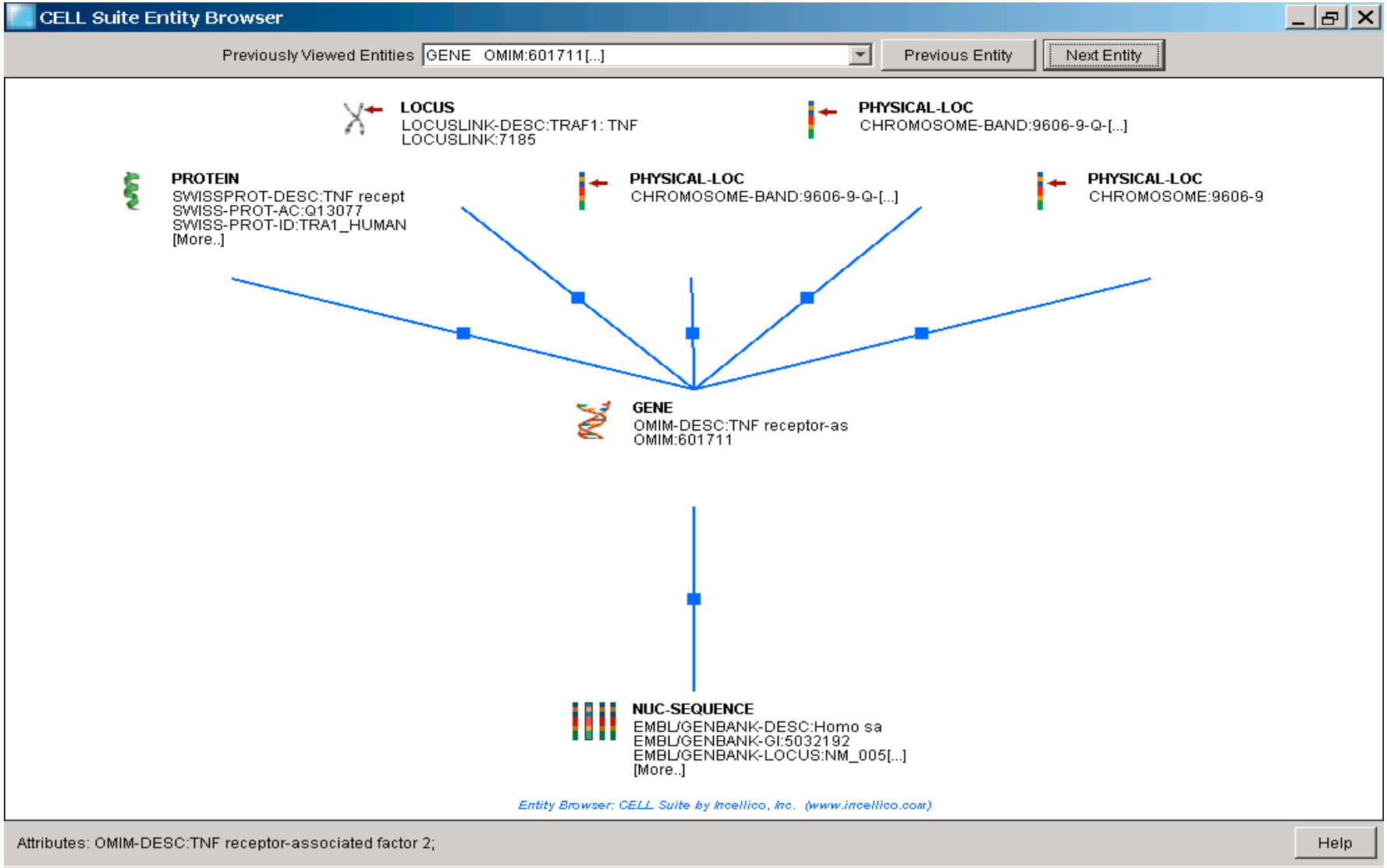
Choose the Human Protein and view the full range of associated data



Horizontal Scrolling Allows Searching the Breadth of Data

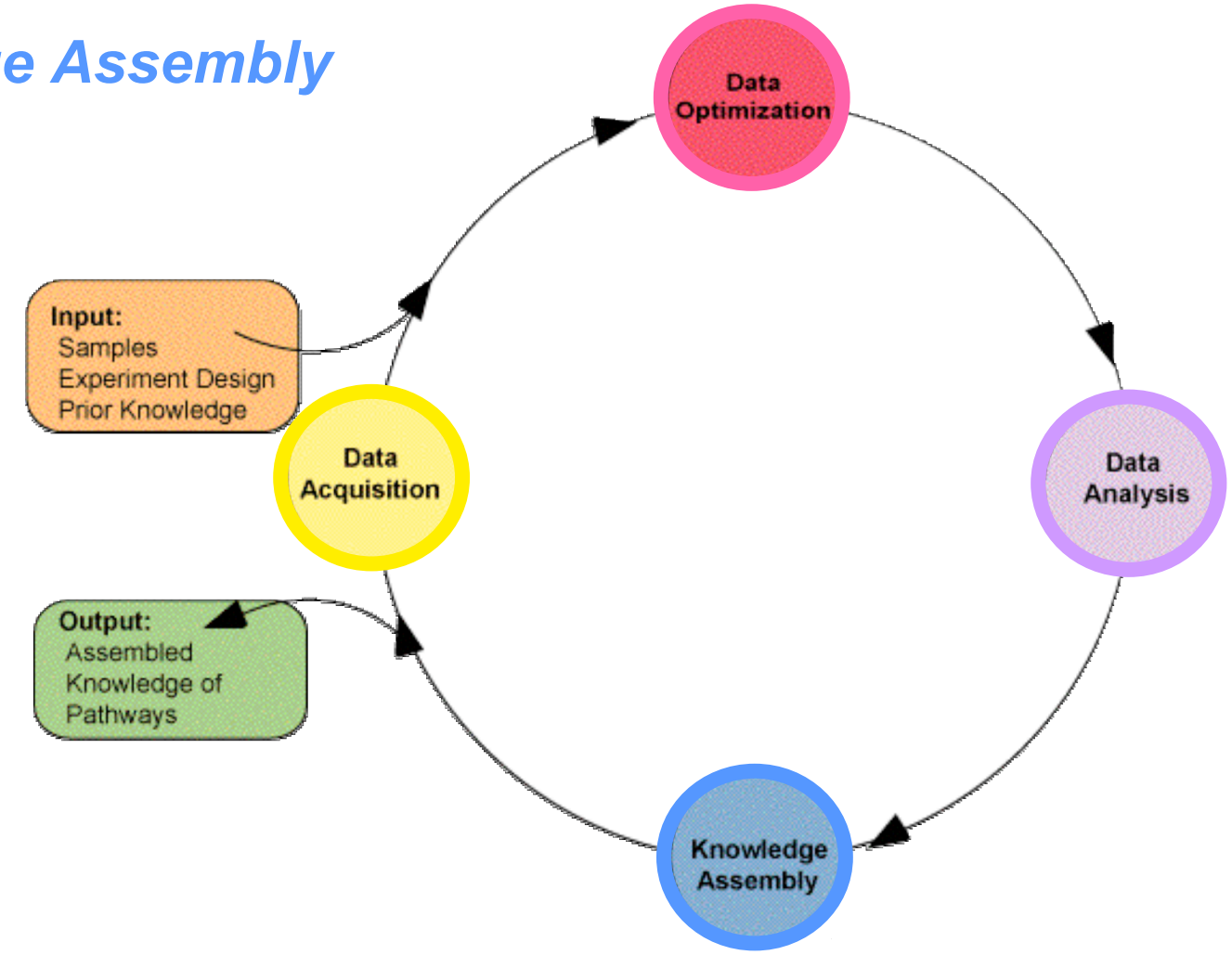


Locus Link accessible via Nucleic Sequence



BioSystematics™ at Beyond Genomics:

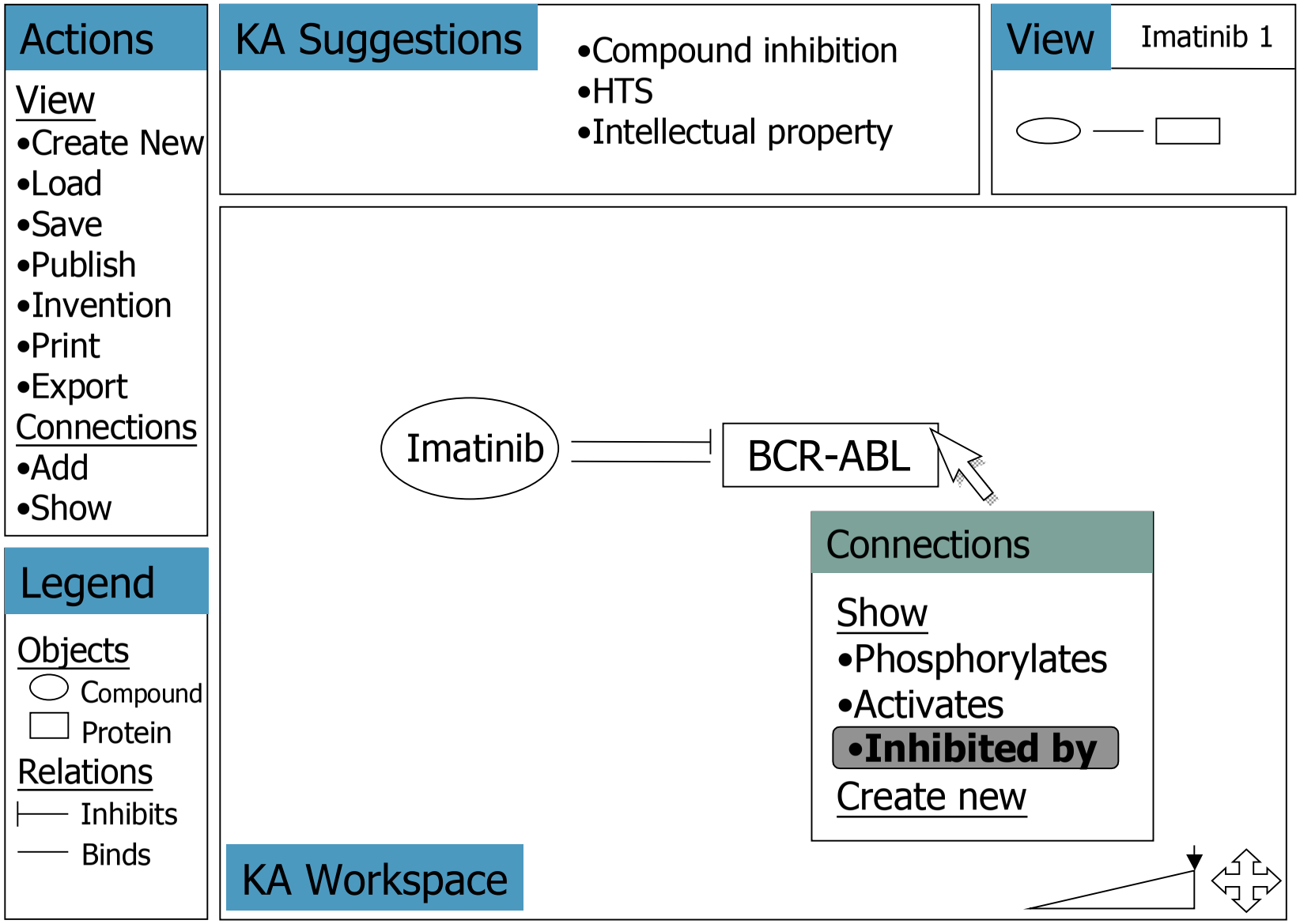
Knowledge Assembly



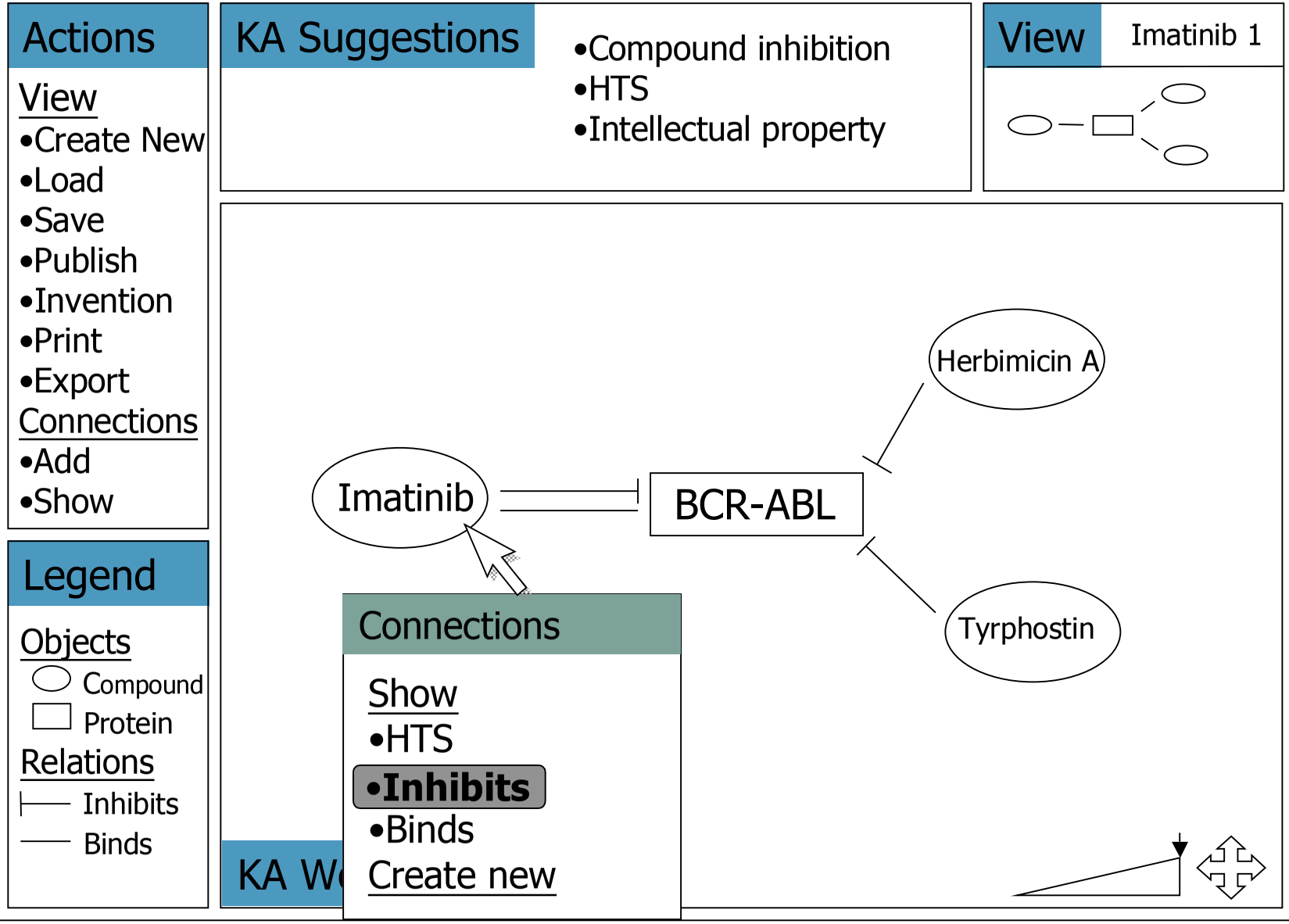
Knowledge Representation / Assembly: (Genstruct)

- Provides Expressivity to Biology; many complex semantics that WILL change in time, that need to be captured!
- Provides a software environment for revealing the logic of biology
- Uses technology that renders connections of any type (logic bytes) in an environment in which scientists can:
  - Navigate connections between biological objects and concepts
  - Query the connections
  - Reason computationally

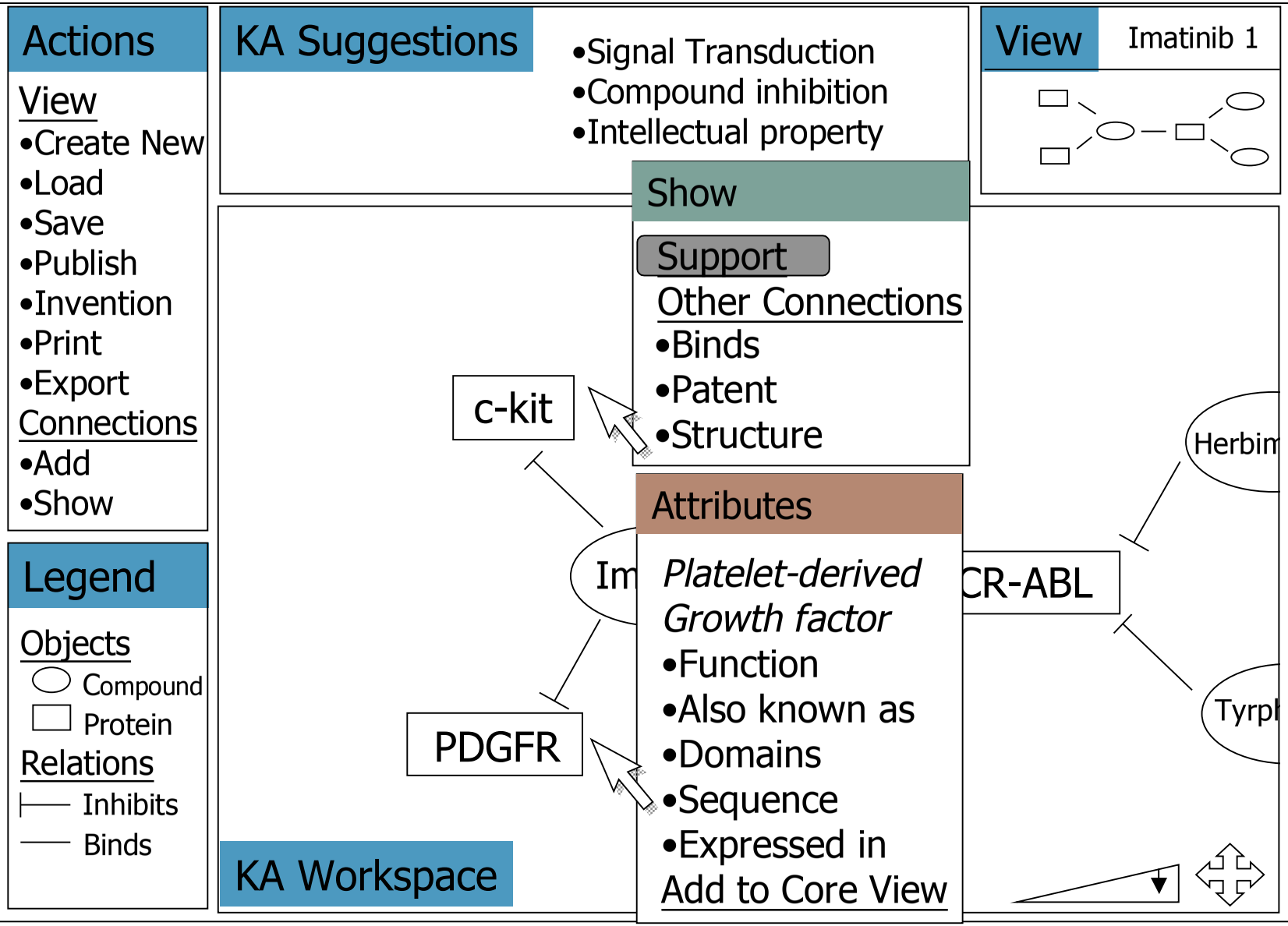
Discovery Environment



Discovery Environment

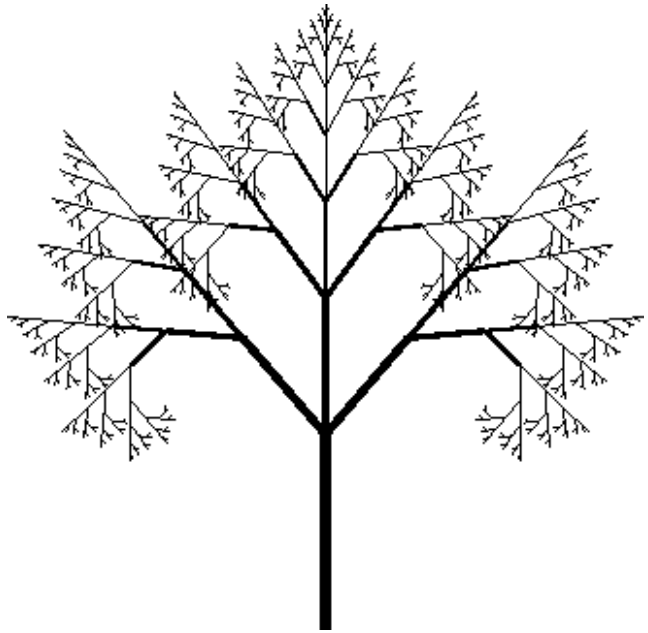


Discovery Environment



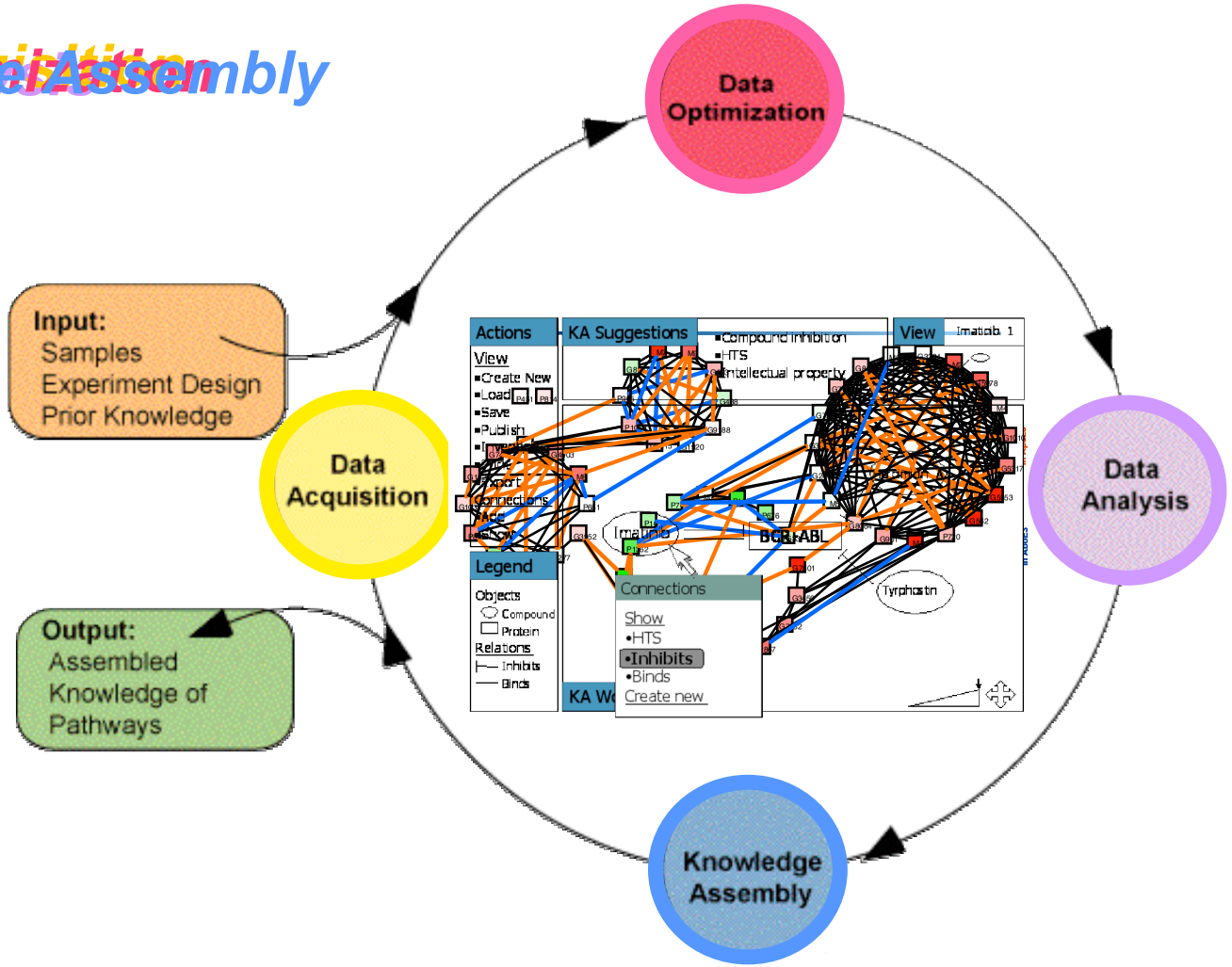
Computational Reasoning in the Knowledge Base

- Ontologies: biologic axioms, normalized nomenclature
- Algorithms: Relevance
  network analysis,
  Bayesian nets
- Knowledge mining: Graph equivalence, Arborescence, Hamiltonian paths and subgraph matching
- Intelligent Agents: User-defined parameters/algorithms search assembled knowledge



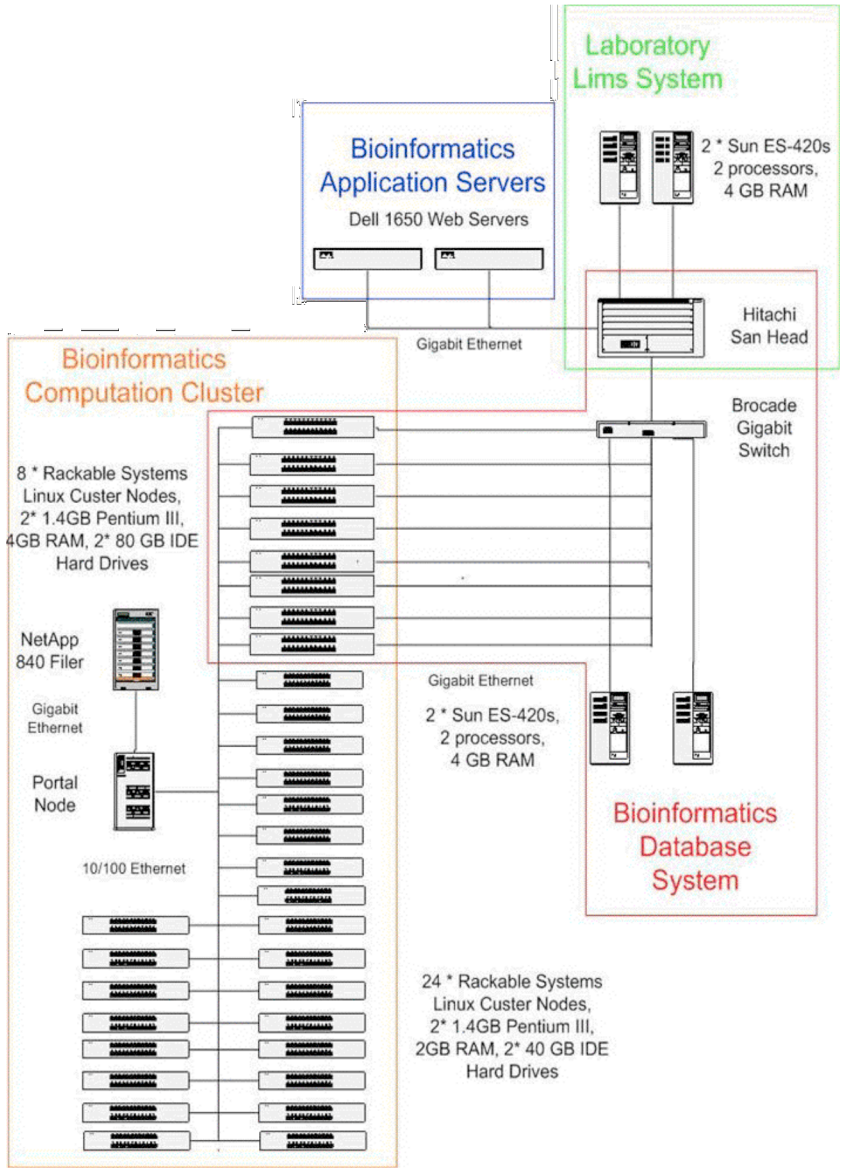
arborescence

The Integrated Flow:



BioSystematics™ at Beyond Genomics:

Hardware Platform



Issues raised by Use of Clusters

- Clustered computing solutions are often a good way to cope with the need for increased computational power, but . . .
- What a world needs is a good cluster management tool
- The problem of “by processor licensing”

Reiteration

- Moving Beyond the Genome Utopia: Systems Biology requires Bioinformatics to take its tools, expertise and agenda to a higher level
- Standards to Cope with Proliferation of Platforms and Vendors
- Lims Systems and Pipeline Management
- Data Analysis Standards
- Visualization Tools
- Working with Clusters, Avoiding Bankruptcy

Acknowledgements

- Eric Neumann
- Matej Oresic
- Tom Plasterer
- Carl Ruel
- Stacey Horrigan
- Genstruct: Jeffrey Thomas
- Incellico: Tre-Tin Lee