Supercomputing 2008

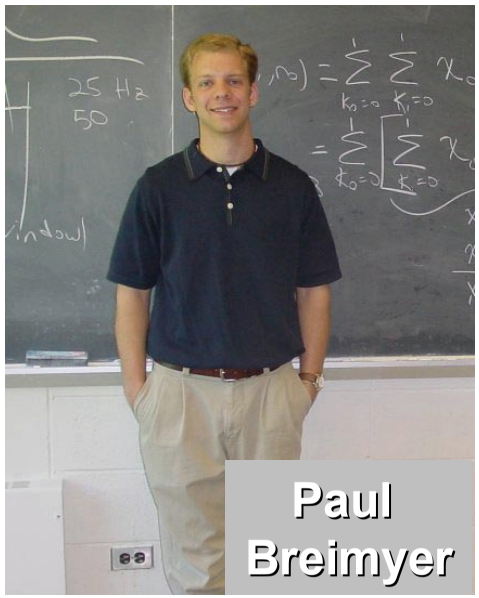
An Outlook into Ultrascale Visualization of Large-Scale Biological Data

A Customer's Perspective

Dr. Nagiza F. Samatova
Department of Computer Science
North Carolina State University
and
Computer Science and Mathematics Division
Oak Ridge National Laboratory

Ultrascale Visualization Workshop

Contributors



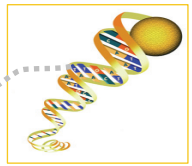
Outline

- **Biology 101**
- **Motivational Problems**
- **Cross-Cutting Ultrascale Visualization Challenges**
  - Top 6 Quests for Next Generation UV of Biological Data

Biology in a Nutshell – CS Perspective

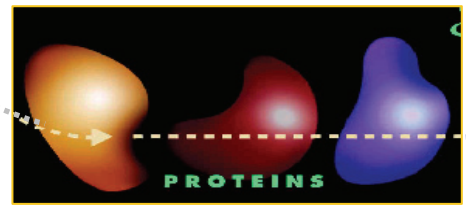
For people with little knowledge but infinite intelligence

Genomes

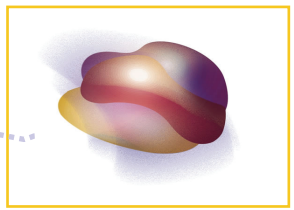


Gene Products

Genome (ROM): assembly code on how to build proteins
Instructions: A, C, T, G
3 variables ⇔ amino acid
Genome consists of genes

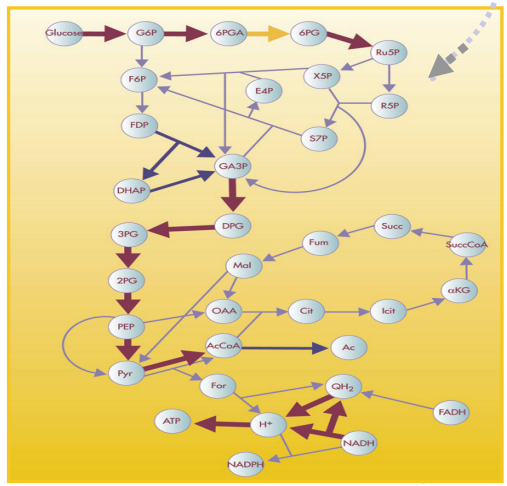


Structure & Function



Pathways & Physiology

Gene ⇔ Protein:
Object description ⇔Object instantiation
Protein: a sequence of amino acids
Protein ⇔Functions
Enzymes: proteins that catalyze biochemical reactions



Pathway: a sequence of reactions
Network (directed graph): a set of pathways
(metabolites: nodes and enzymes: edges)

From Genes to Protein Functions

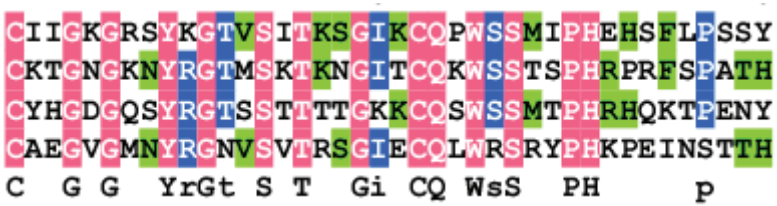
The first and most crucial step in systems biology

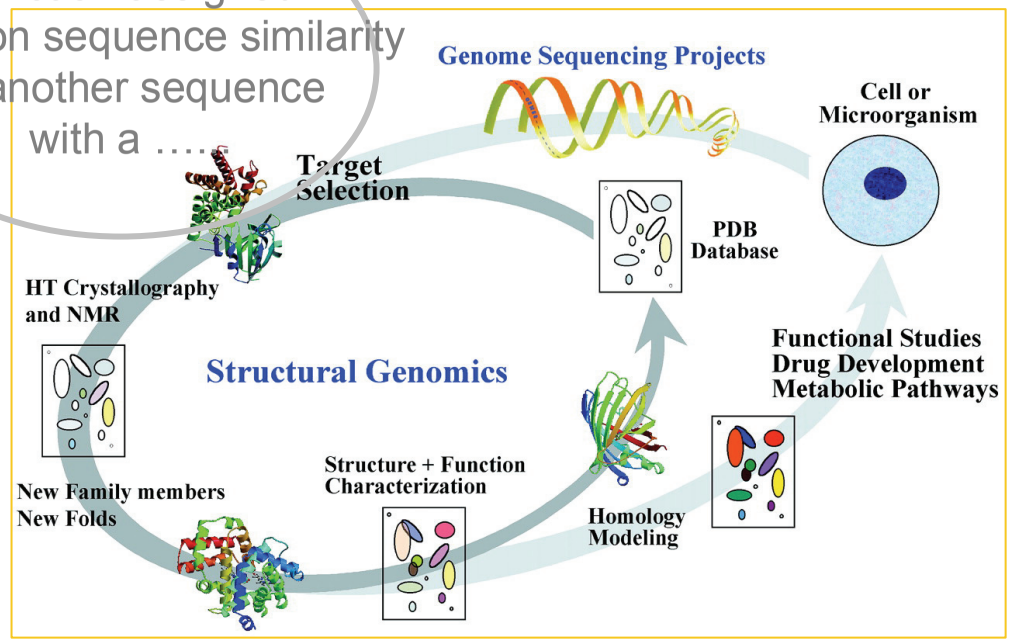
Function assigned based on sequence similarity to another sequence with a

Function assigned based on sequence similarity to another sequence with a

Function assigned based on sequence similarity to another sequence with a ….

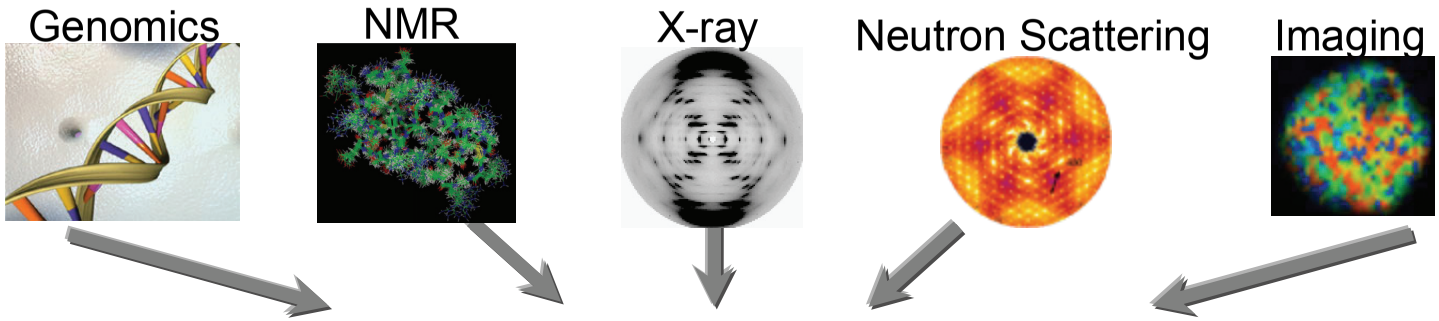
Guilt-by-Association goes global



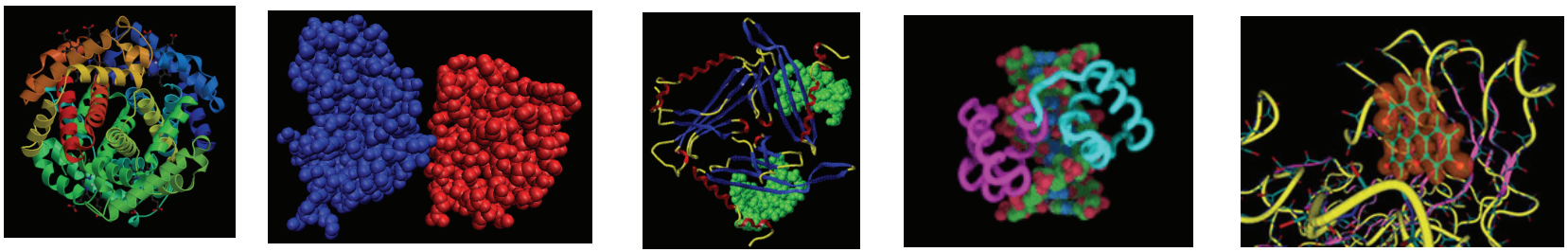


Proteins Function Interactively

What we observe



Protein Machines



3-d Structure **Protein-Protein** **Protein-RNA** **Protein-DNA** **Protein-Ligand**

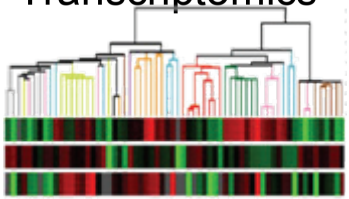
What we want to derive

Proteins Function in Pathways/Networks

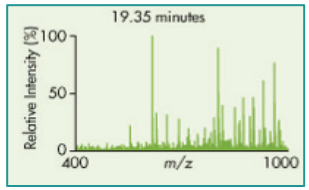
What we observe

Genomics

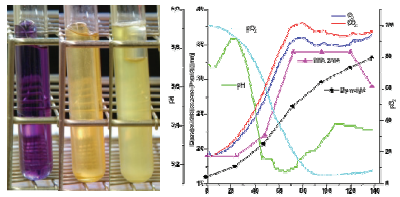
Transcriptomics



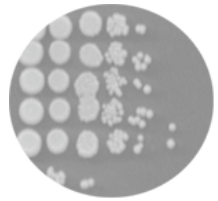
Quantitative Proteomics



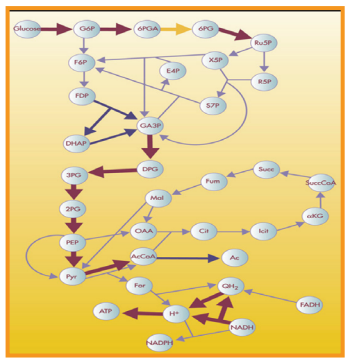
Metabolomics

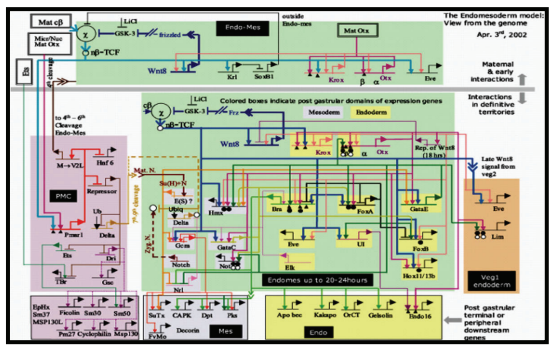


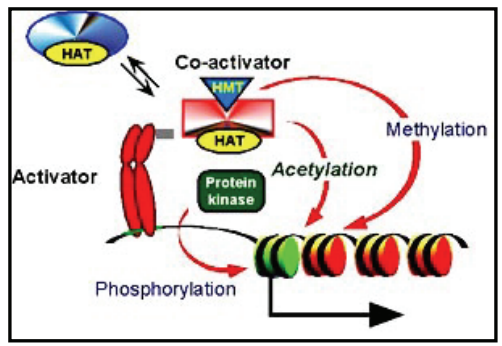
Interactomics

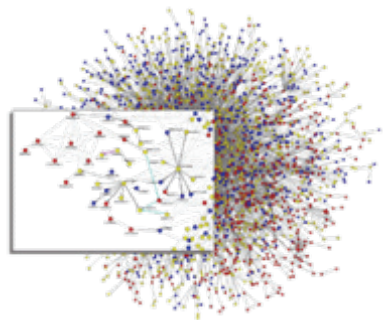


Networks/Pathways









Metabolic **Regulatory** **Signaling** **Protein Interaction**

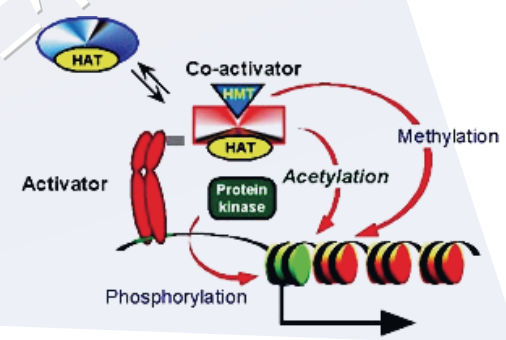
What we want to derive

It is not just the Size – but the Complexity

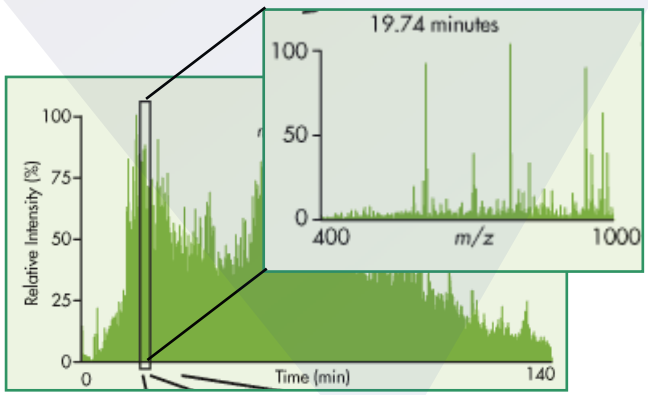
High-dimensional

Large-Scale Data

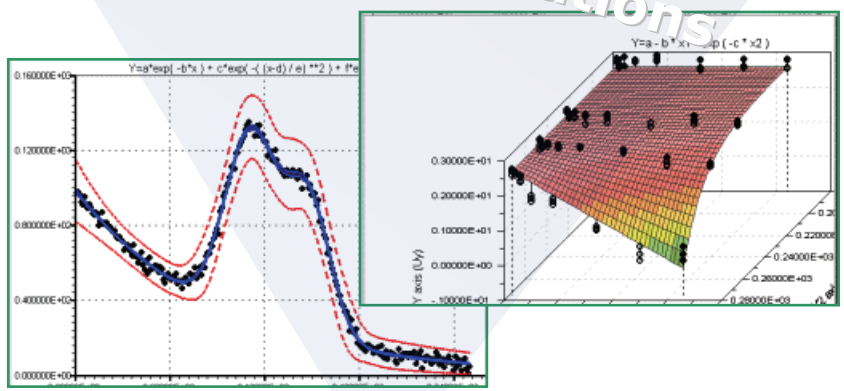
'+' and '–' feedbacks



Noisy



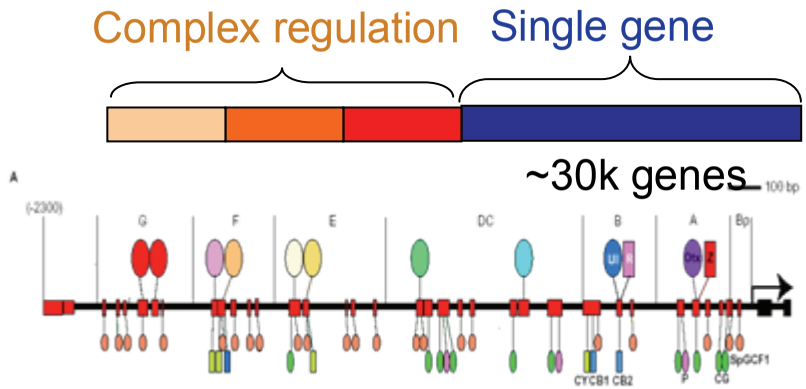
Non-linear correlations



Data Describes Complex Natural Phenomena

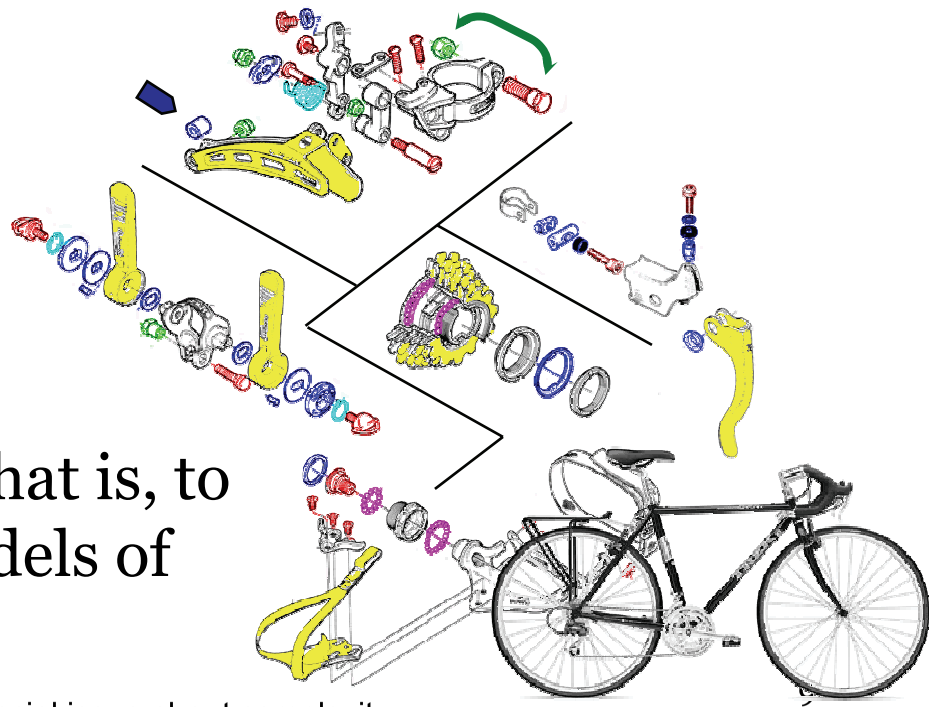
How to untangle riddles of BioComplexity

Complexity of biological systems comes from interconnections.



50 trans elements control single gene expression

Biological systems are complex because of non-linear coupling of their structural, genotypic, & phenotypic properties.



Challenge:

How to "connect the dots", that is, to construct predictive *in silico* models of these biological systems.

Science Vol. 284. No. 5411 (1999), Special issue about complexity

How Can the Viz Community Help?

Outline

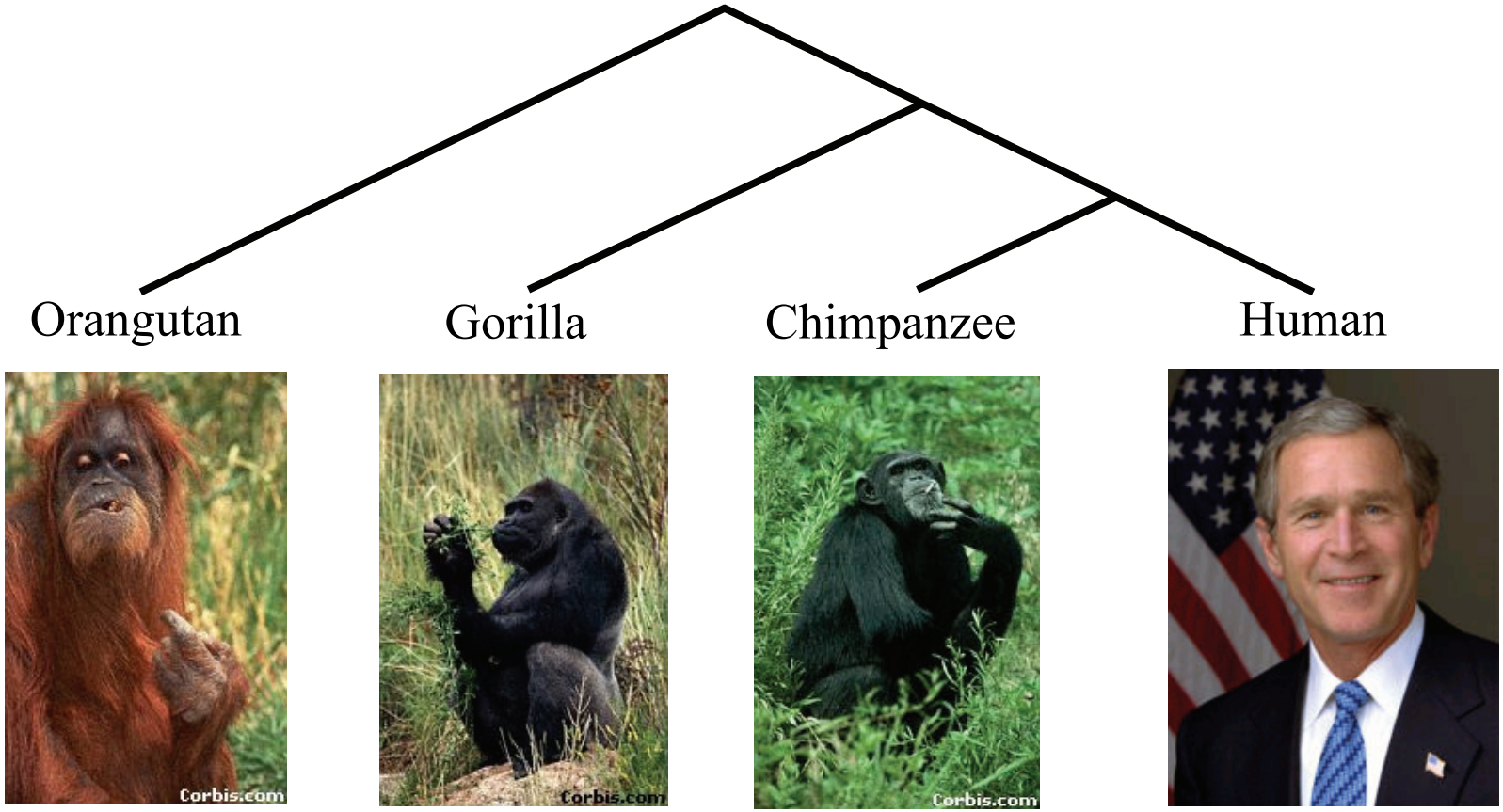
- **Biology 101**
- **Motivational Problems**
- **Cross-Cutting Ultrascale Visualization Challenges**
  - Top 6 Quests for Next Generation UV of Biological Data

Reconstruction of the Tree of Life

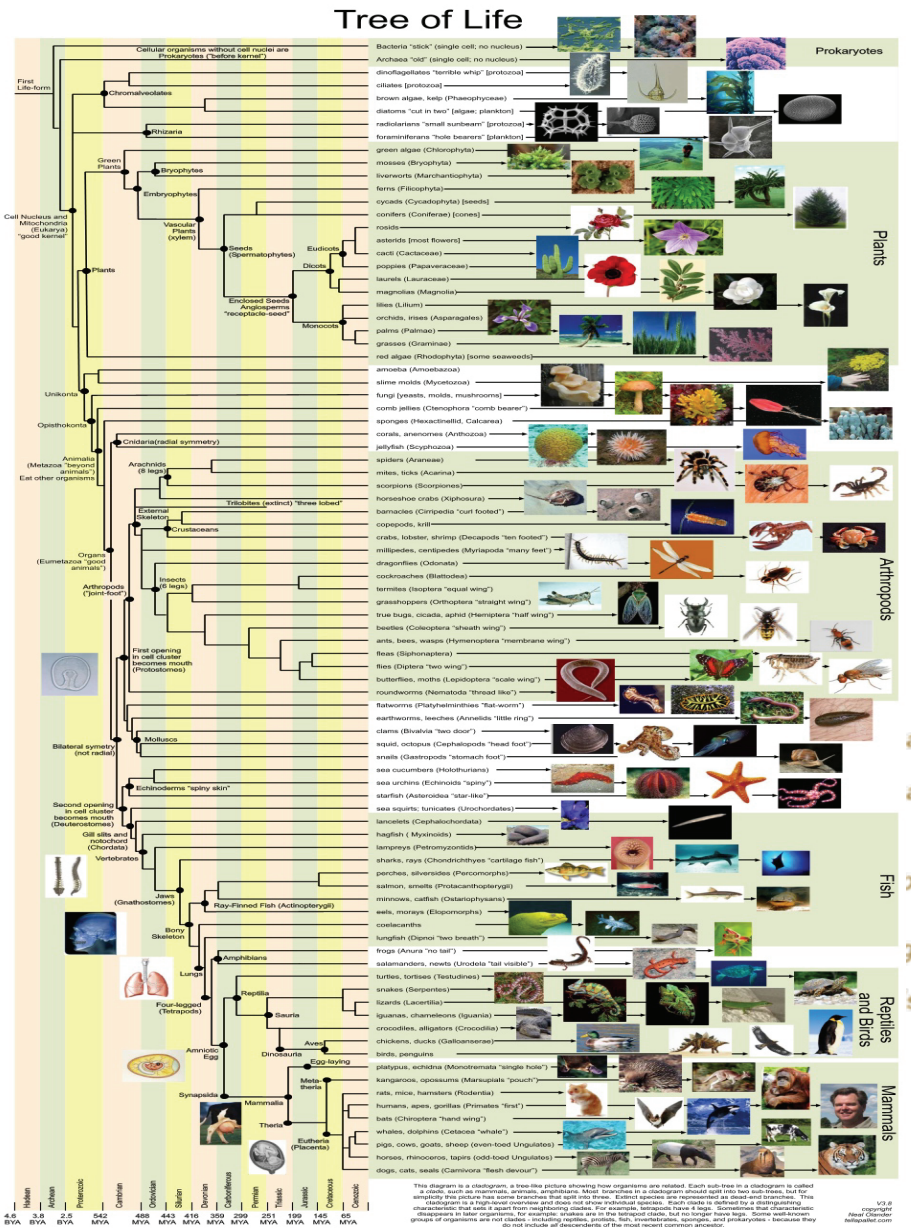
Ultrascale Visualization Problem

Phylogeny

A **phylogeny** is a tree representation for the evolutionary history relating the species of interest.



The Tree of Life for 10-100 Million Organisms



Ultrascale Challenge

Computing the optimal phylogenetic tree based on the entire genome of 10 species will remain intractable even with peta-scale computers

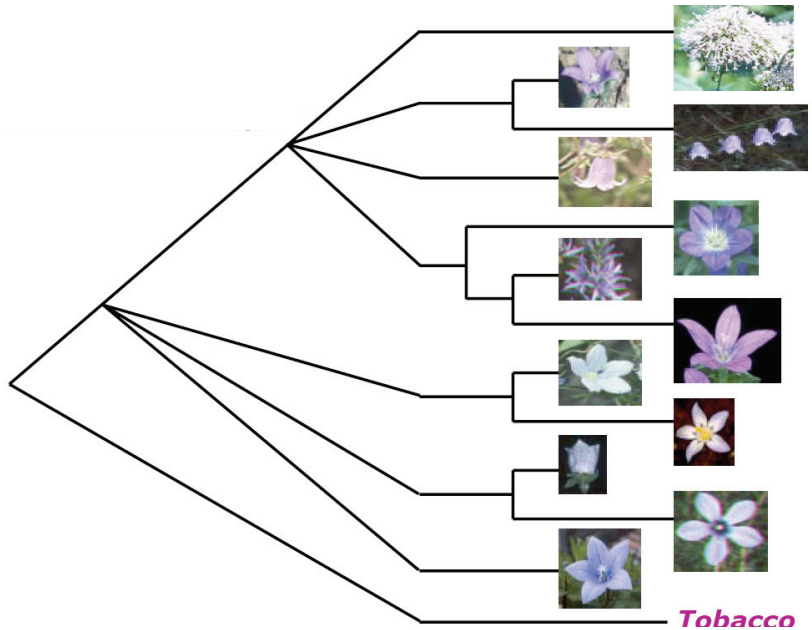
Complexity

Maximum likelihood

C~=n^2*m where

n is number of contemporary species

m is size of the genome in question



Requirements

For 10^6 species on a gene sequence of about 1000 pairs or 1000 species on the entire genome:

Runtime: 1-3h at 1Pflop

Memory: ~3 TB

Dealing with Computational Intractability

Computational intractability drives for various search heuristics to navigate a small fraction of exponentially-sized tree space in practical time.

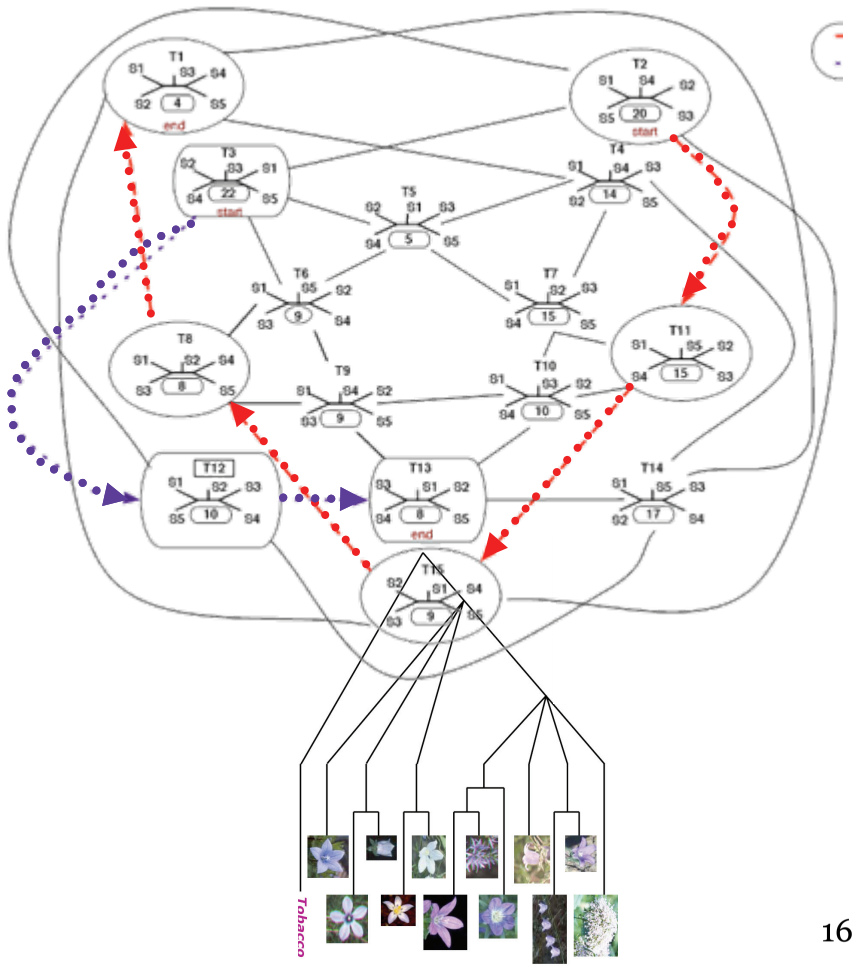
Sequence Alignment

```
GAATTCAG
|  |  ||  |
GGA-TC-G
```

Search Heuristics

- PAUP
- RAxML
- Phylip
- MrBayes
- Parsimony

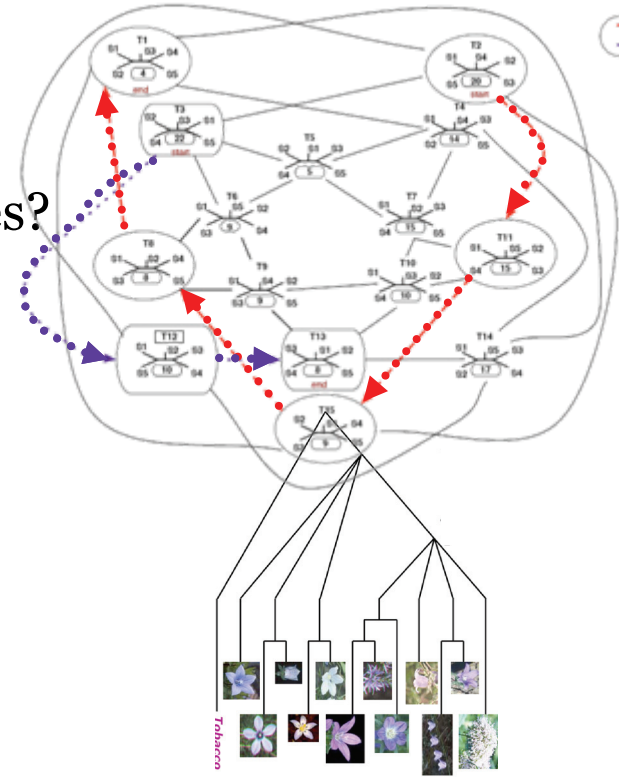
Search Histories



Visual Exploration of the Search History

Ultrascale Visualization Questions:

- How to visualize the **landscape of local optimalities** (search histories)?
- How to visually **compare** the search histories from different heuristics?
- How to visually **align** two/many trees?
- How to visualize the **hierarchical clusters** of trees?



Impact

Design of better heuristics
More accurate reconstruction of phylogenies

Comparative Analysis of Networks

Ultrascale Visualization Problem

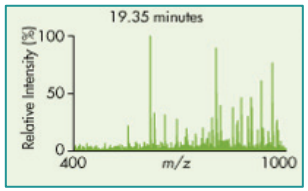
What is Genome-scale Interactome?

Integrated structural, genomic, and MS pull-down experimental clues for inference of genome-scale Interactome.

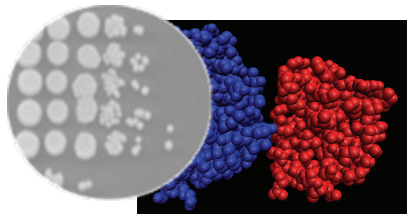
Input Data

Genomics

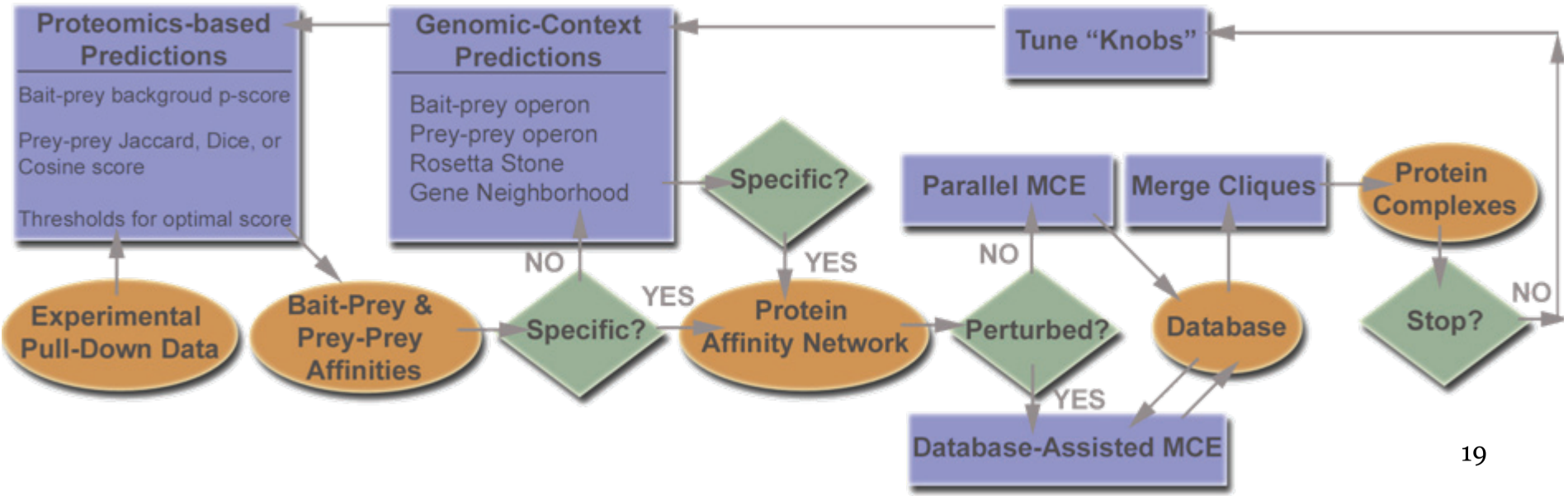
MS Pull-down



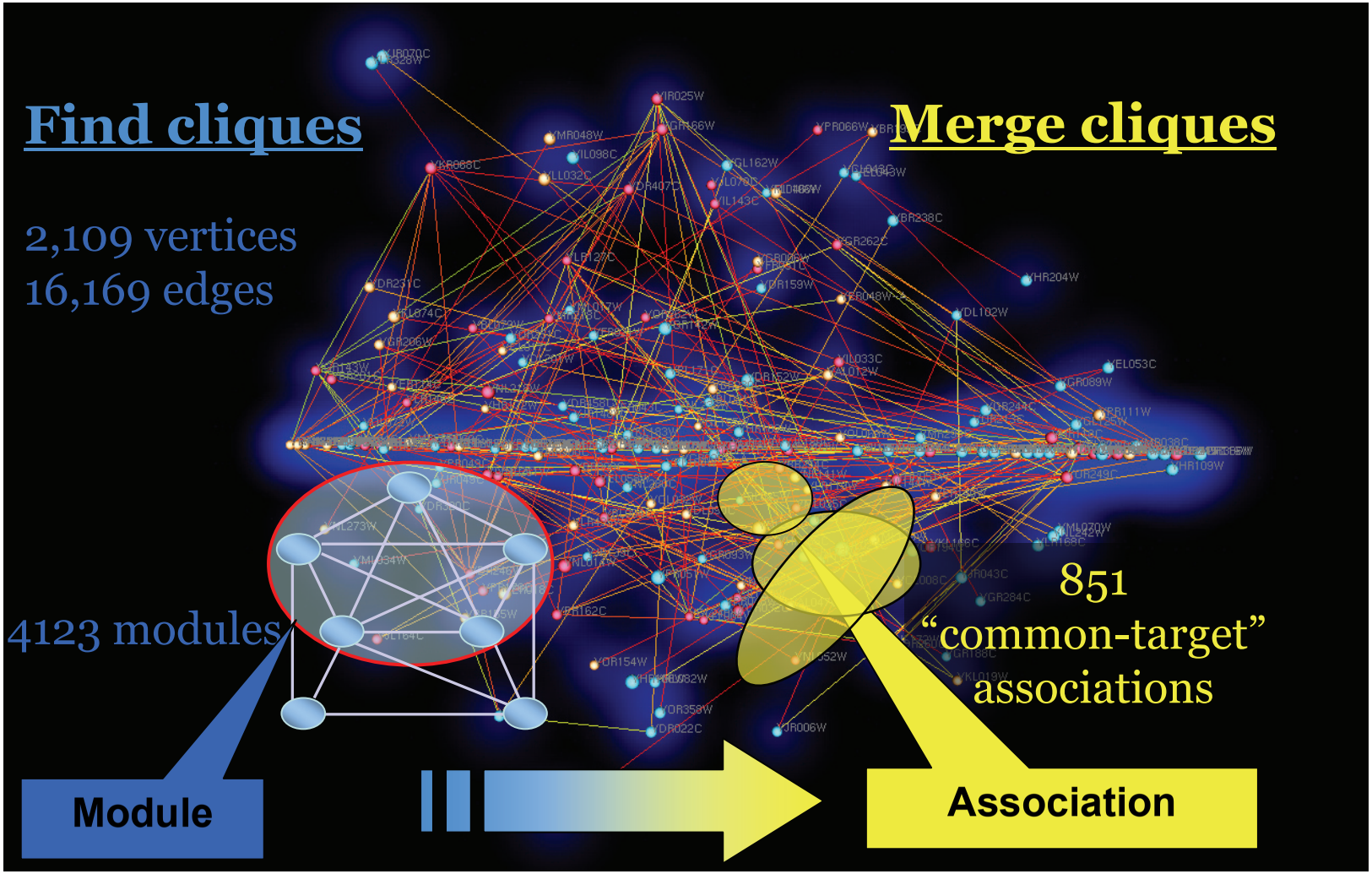
Interactomics



Network Inference Pipeline



Visual Exploration of Genome-scale Interactome?

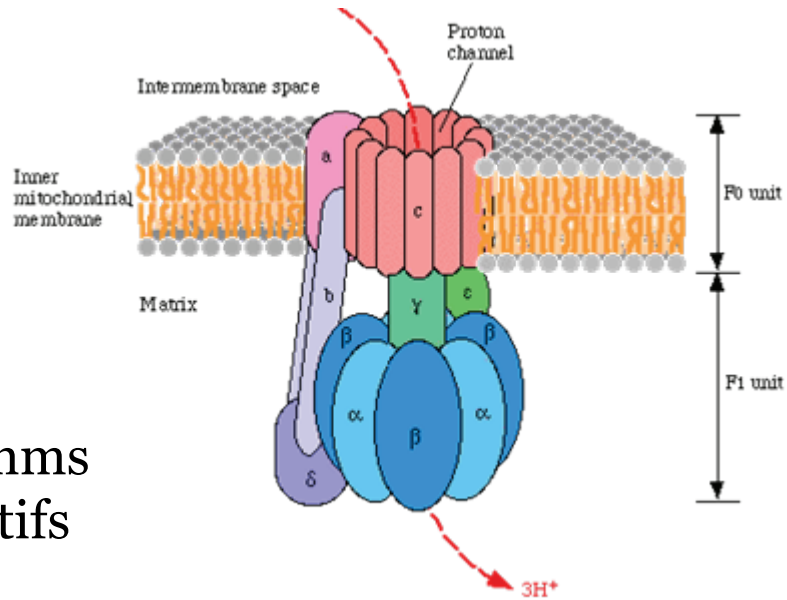


Visualization provided by Dr. Jian Huang and Mr. Joshe New, SciDAC Ultrascale Viz. Institute

Visual Exploration of Networks

Ultrascale Visualization Questions:

- What are highly connected (dense) **network motifs**?
- Are the network motifs **statistically significant/biologically relevant**?
- How do network motifs change for different **parameters** (knobs)?
- Is a network **motif of interest** (e.g. ATP synthase) present and what is it connected to?



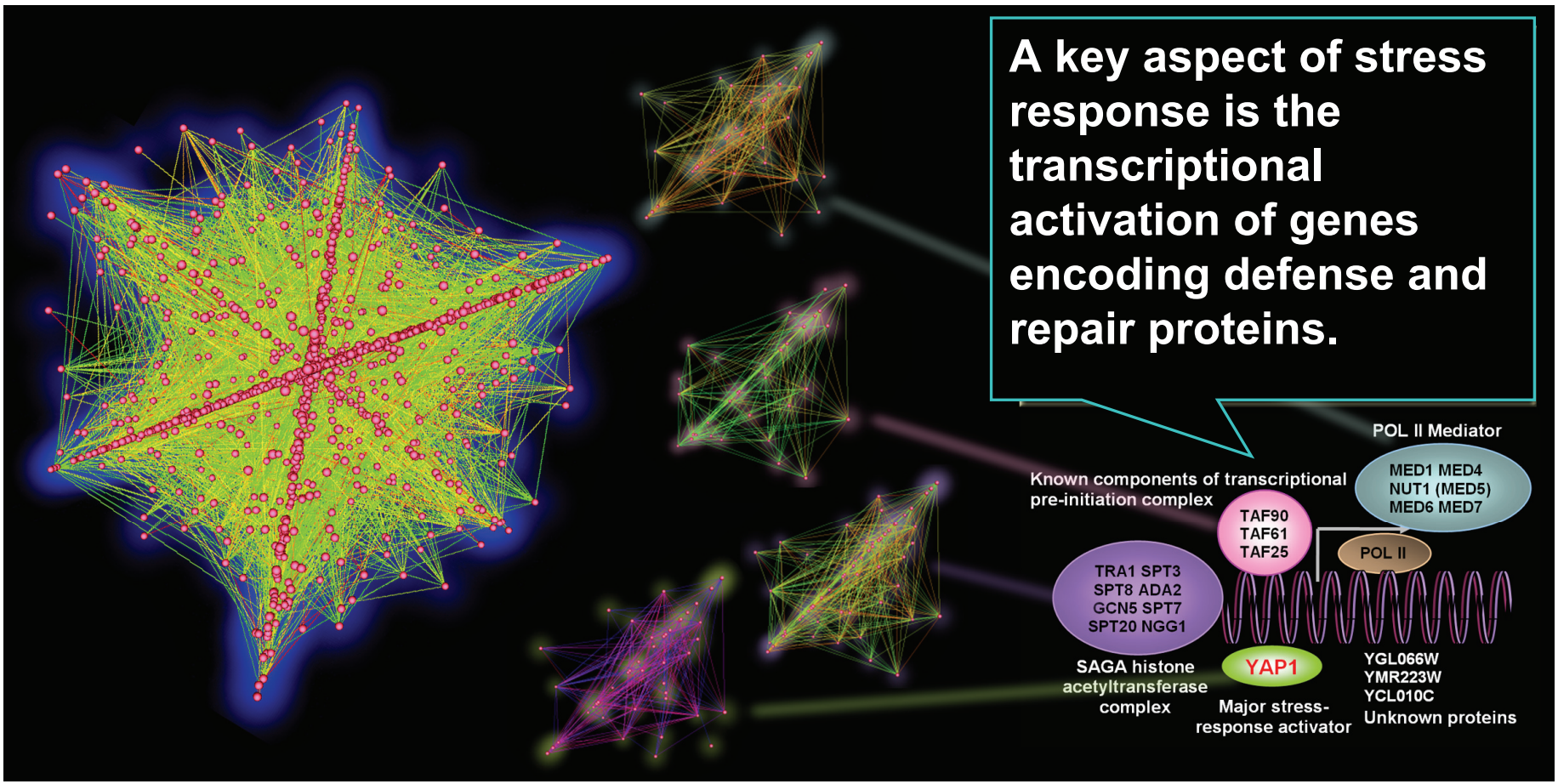
Impact

Design of better network inference algorithms
Discover biologically relevant network motifs

Prior Knowledge: ATP synthase

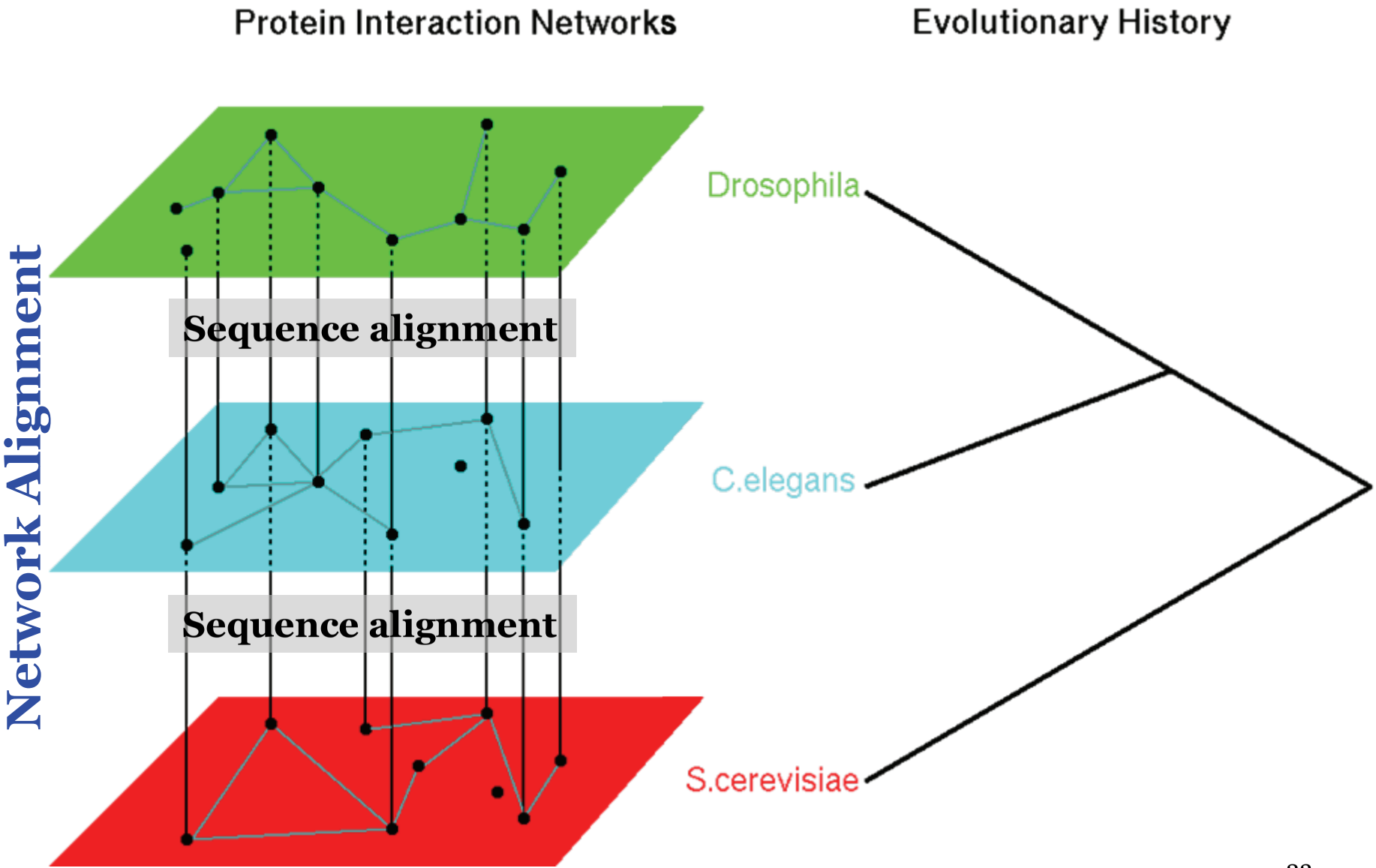
What cellular machineries are responsible for organismal resistance to stress?

Thermochemical pre-treatment of biomass and acidic compounds of sugar fermentation to ethanol require organisms resistant to stress.



Visualization provided by Dr. Jian Huang and Mr. Joshe New, SciDAC Ultrascale Viz. Institute

Are Network Motifs Evolutionarily Conserved?

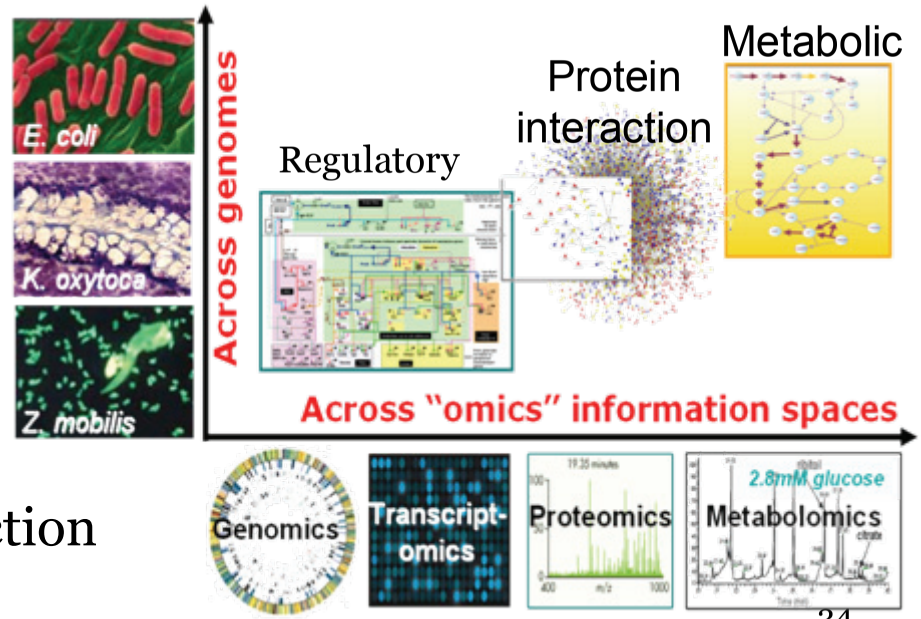


http://www3.imperial.ac.uk/pls/portallive/docs/1/50690.PNG

Visual Exploration of Networks Evolution

Ultrascale Visualization Questions:

- What network motifs are **evolutionary conserved**?
- Is the conservation **statistically significant** (compared to **random networks**)?
- Is a network **motif of interest** evolutionary conserved? Across what organisms? Are these organisms evolutionary close or distant?
- How to visually compare networks **across organisms** and **"omics" information spaces**?



Impact

Design better network analysis
Discover novel network motifs
Annotate proteins with unknown function

Identification of Phenotype-Specific Genes

Ultrascale Visualization Problem

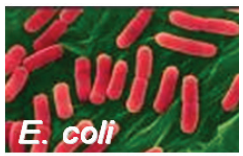
What Genes are Responsible for a Phenotype?

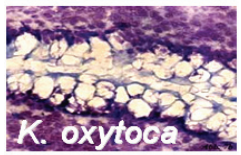
Phenotypes:

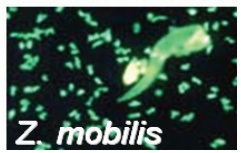
Resistance to:
- High/low temperature
- Low/high pH
- High EtH concentration

Growth:
- Aerobic
- Anaerobic

Metabolism & Productivity:
- Ferment multiple sugars
- High EtH yield





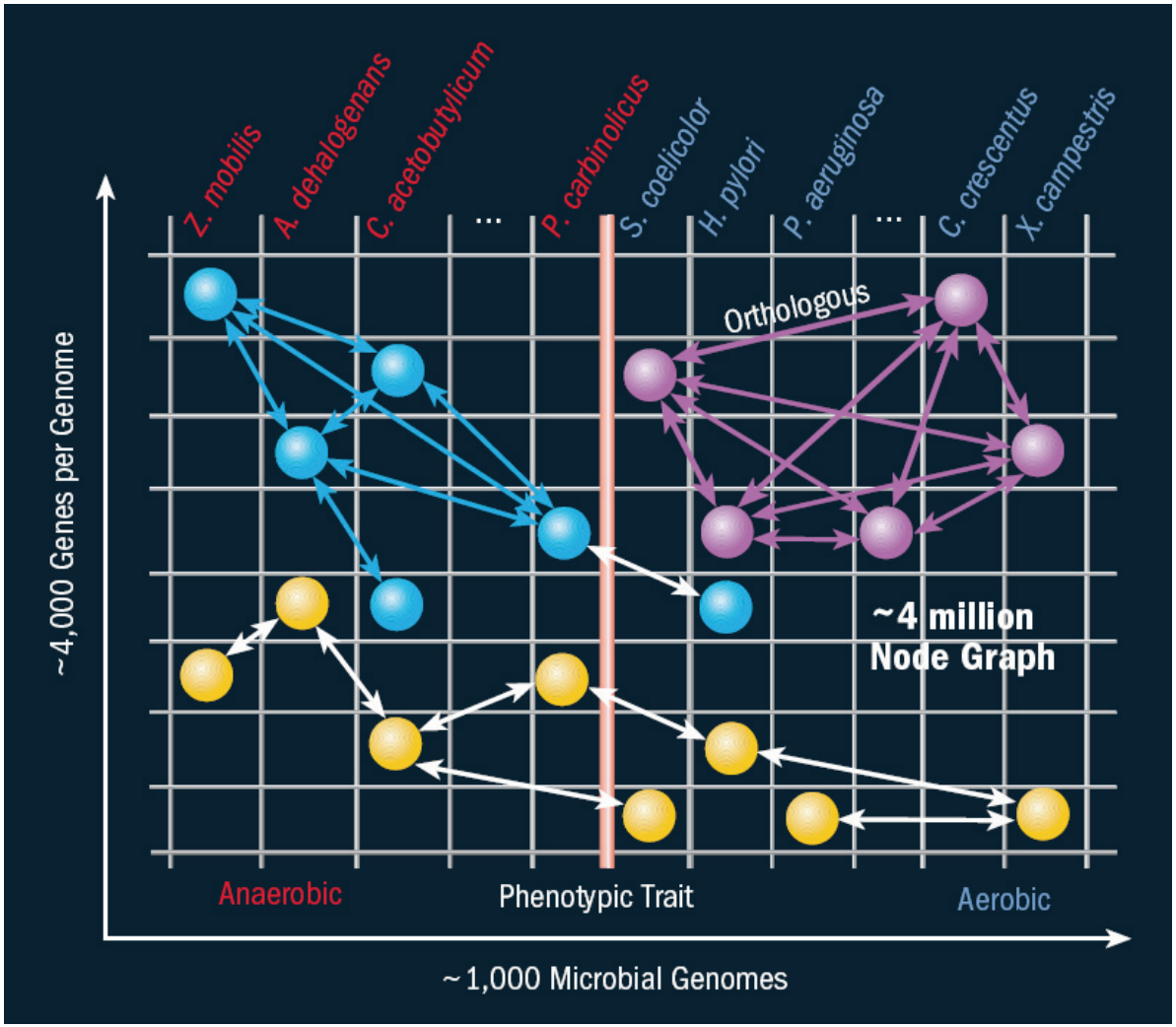


Phenotypic Traits:

pH	T(°C)	Mannose metabolism	Xylose metabolism	Ethanol yield (%)
6.5	35	+	+	89
5.5	30	+	+	95
5.5	30	—	+	84

Identification of Phenotype-Related Genes

If a gene is responsible for a phenotypic trait, then it is evolutionarily conserved across several organisms.



4 million nodes

Multi-partite graph

Genotype-Phenotype Relationships

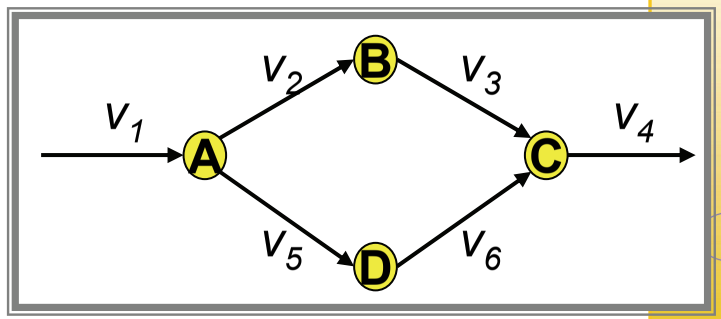
Ultrascale Visualization Problem

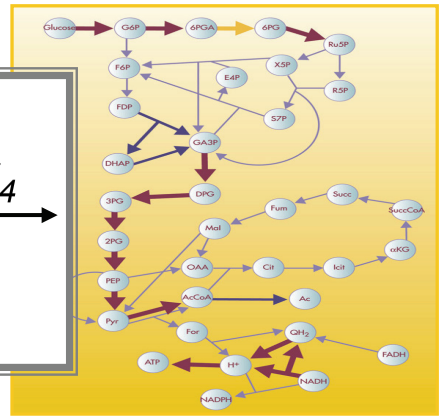
Mathematical Laws of Metabolism

Chemical reaction:

$$aA + cC \xrightarrow{v_i} eE + hH$$

Reaction Network:





Stoichiometry Matrix, S:

reactions

metabolites

A	-a
B	0
C	-c
D	0
E	+e
H	+h

Flux Vector, V:

$$\begin{bmatrix} -a & \cdots \\ 0 & \cdots \\ -c & \cdots \\ 0 & \cdots \\ +e & \cdots \\ +h & \cdots \end{bmatrix} \times \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Steady State/ Mass Balances:

$$A: v_1 - v_2 - v_5 = 0$$
$$B: v_2 - v_3 = 0$$
$$C: v_3 - v_4 + v_6 = 0$$
$$D: v_5 - v_6 = 0$$

Steady state flux vectors are in the Null Space of *S*

$$S \bullet v = 0$$

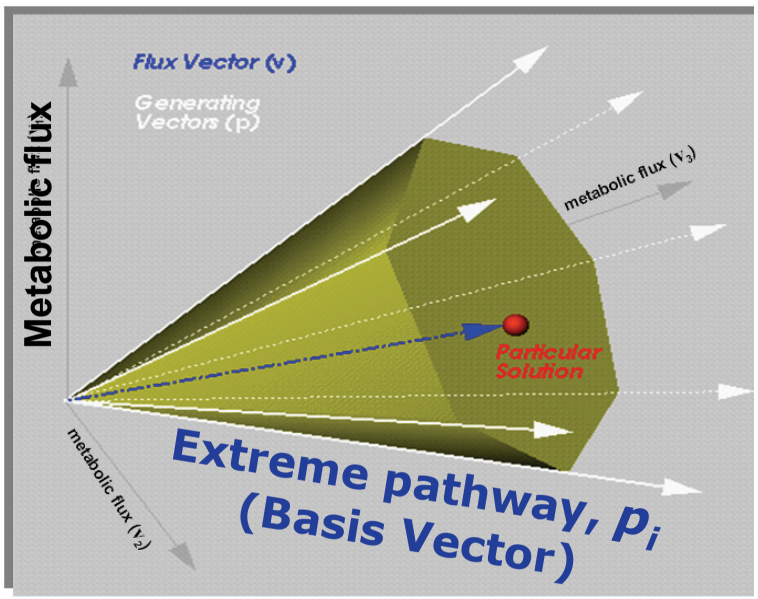
High Dimensional Space of System Phenotypes

Metabolism could be studied within convex analysis context. Extreme pathways are a basis for a metabolic genotype. Their regulation defines all possible metabolic phenotypes.

Convex polyhedral cone, *C*

$$S \bullet v = 0,\ v_i \geq 0,\ i = 1,\ldots,n$$

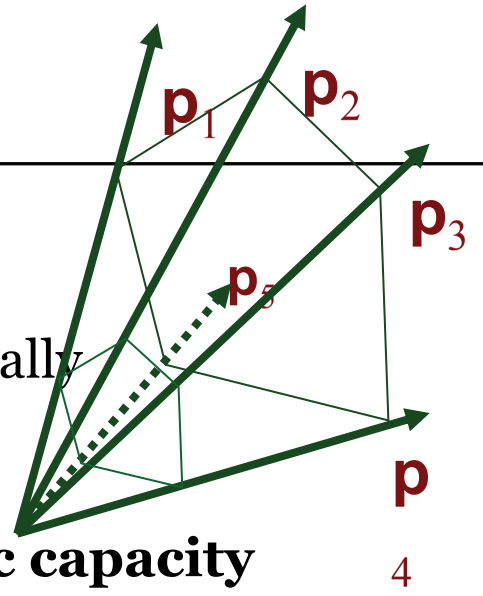
$$C = \left\{ v \in R^n \;\middle|\; v = \sum_{i=1}^{k} \alpha_i p_i,\quad \alpha_i \geq 0 \right\}$$



Convex Analysis	Cellular Biology
Convex Hull	Capabilities of a Metabolic Genotype
Unique Generating Vectors	Independent Extreme Pathways
Particular Solution	Metabolic Phenotype
Flux Vector	Positive Combination of Extreme Pathways

Every phenotype that the system can exhibit is a combination of these extreme pathways, which are then turned on or off.

The UV Challenges



- **The number of extreme pathways can exceed the dimensions of the cone (i.e. linearly dependent):**
  - How to visualize the pathways (potentially exponentially many)?
- **The volume of the cone is a measure of metabolic capacity of the organism:**
  - Volume computation for a *d*-polytopes is an exponential problems, ($O(n^d)$)
  - How to visually represent the volume?
  - Is metabolic potential of an organism is rich or poor?
- **If the objective function is known, then it is a linear or non-linear programming problem:**
  - Organism may pursue multiple objective functions
  - How to solve the inverse problem? – How to find the objective functions based on the observed phenotypes?
  - Can visualization help with this problem?

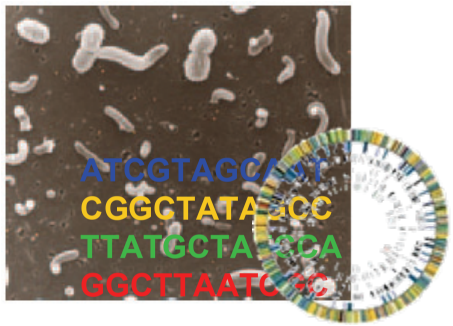
Outline

- **Biology 101**
- **Motivational Problems**
- **Cross-Cutting Ultrascale Visualization Challenges**
  - Top 6 Quests for Next Generation UV of Biological Data

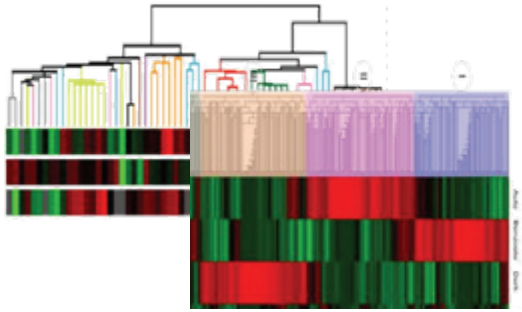
Q1: Comparative Ultrascale "Visual-omics"

- Genome
- Proteome
- Transcriptom e
- Physiome
- Metabolome
- Phenome
- Morphome
- Interactome
- Glycome
- Secretome
- Ribonome
- Orfeome
- Regulome
- Cellome
- Operome
- Transportome
- Functome
- Translatome
- Pseudome
- Foldome

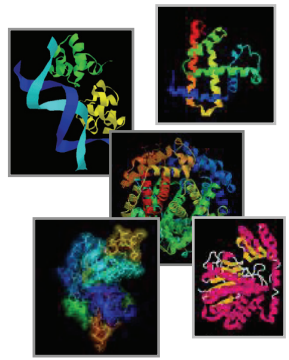
Comparative Metagenomics



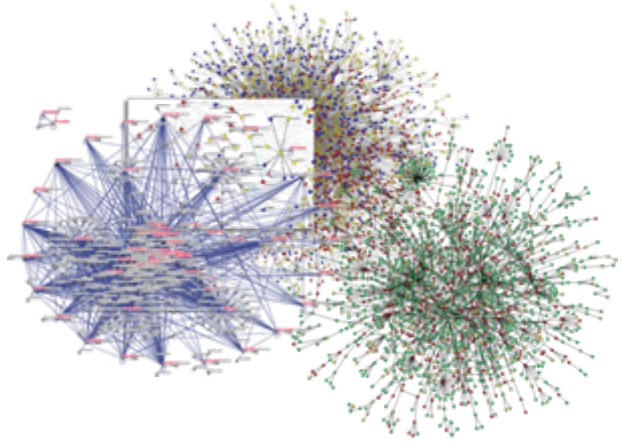
Comparative Transcriptomics



Comparative Proteomics

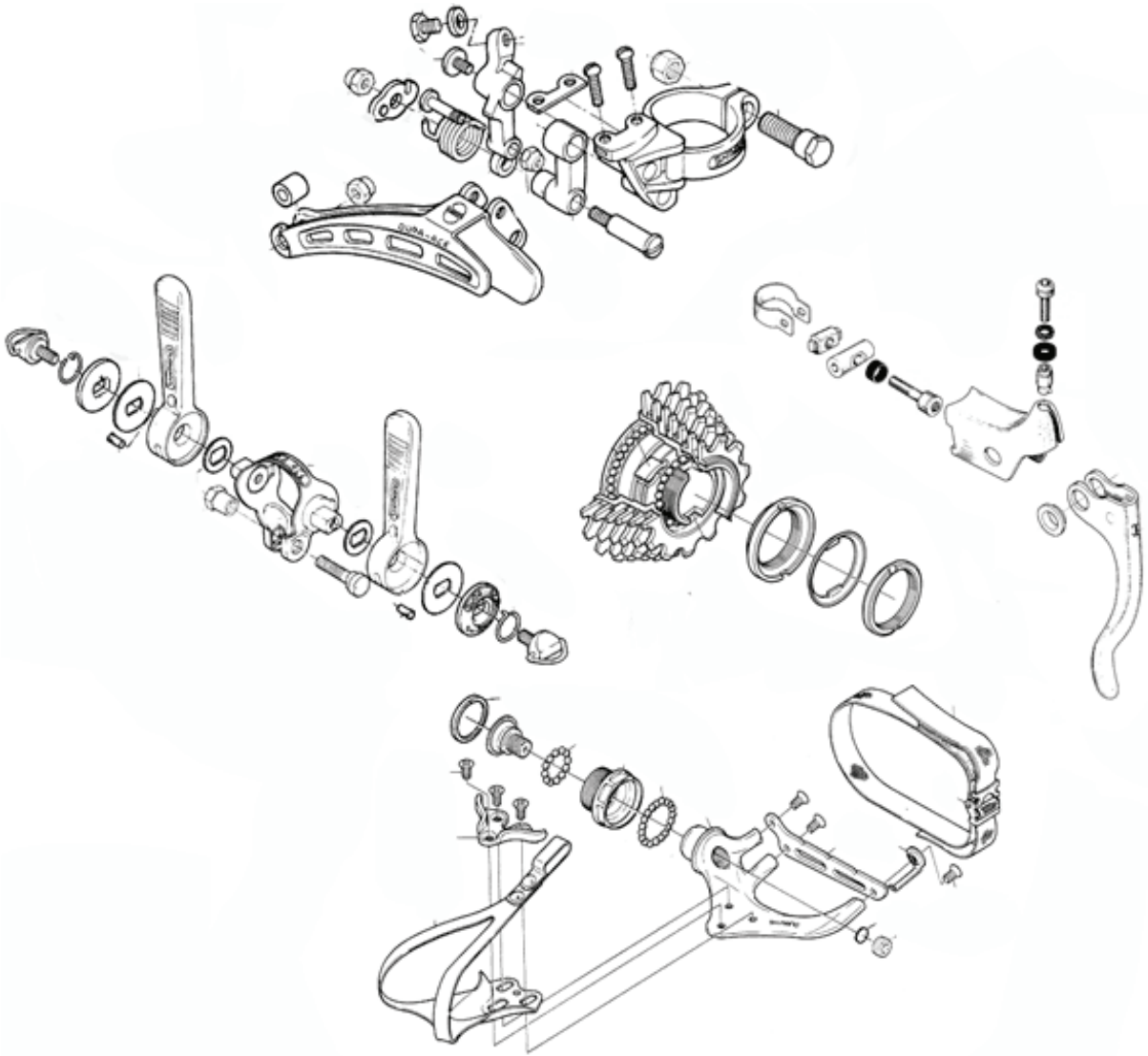


Comparative Interactomics

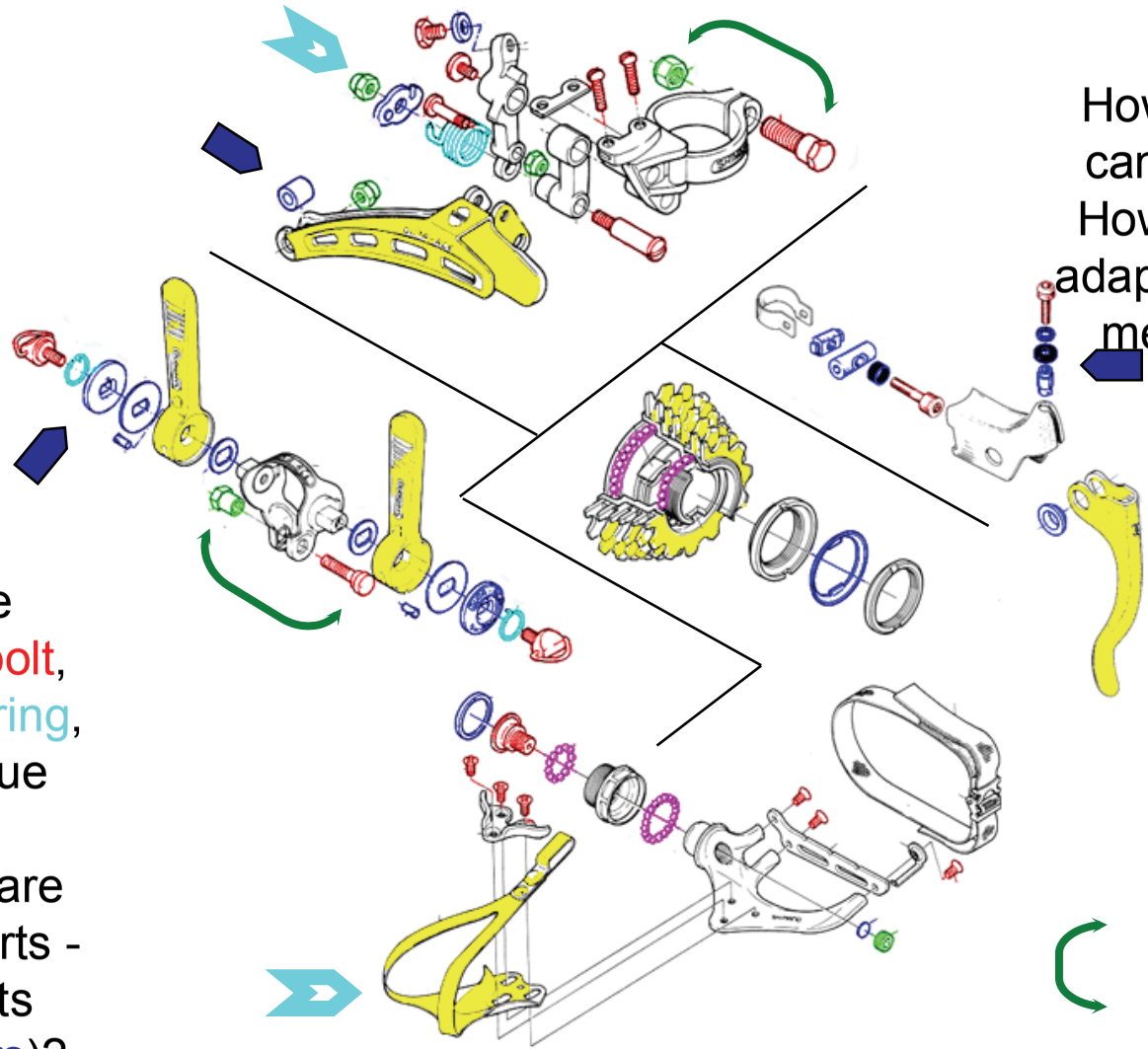


Q2: Integrative Ultrascale “Visual-omics”

A collection of Parts



From Parts to Parts Assembly



How many roles can these play? How flexible and adaptable are they mechanically?

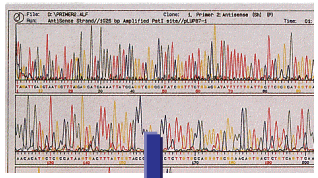
What are the shared parts (bolt, nut, washer, spring, bearing), unique parts (cogs, levers)? What are the common parts -- types of parts (nuts & washers)?

Where are the parts located? Which parts interact?

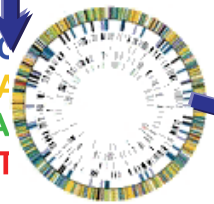
Towards dynamic systems
through Integrative "visual-omics"

- Genome
- Proteome
- Transcriptom
- e
- Physiome
- Metabolome
- Phenome
- Morphome
- Interactome
- Glycome
- Secretome
- Ribonome
- Orfeome
- Regulome
- Cellome
- Operome
- Transportome
- Functome
- Translatome
- Pseudome
- Foldome

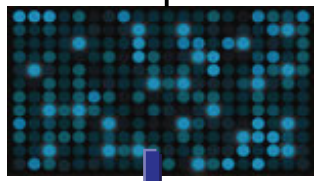
Genomics

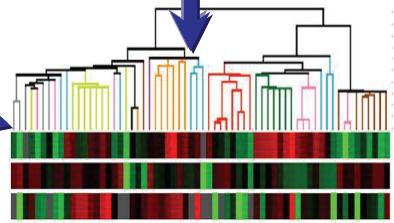


ATCGTAGC
CGGCTATA
TTATGCTA
GGCTTAAT

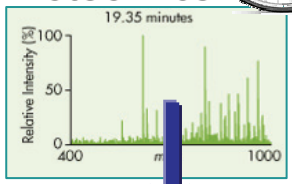


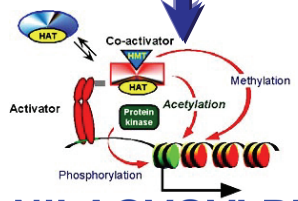
Transcriptomics





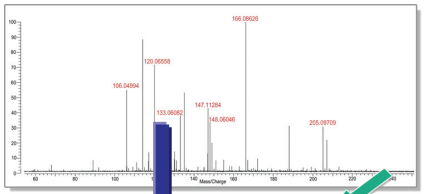
Proteomics





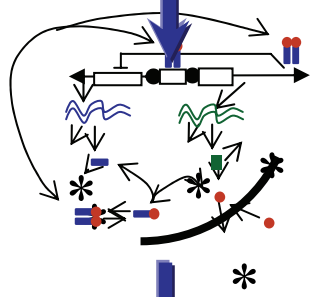
R.NILAQVGVLPR.T

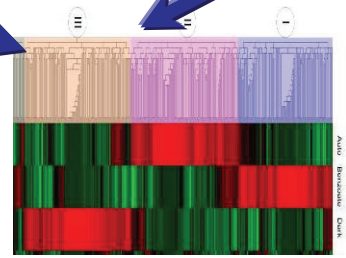
Metabolomics

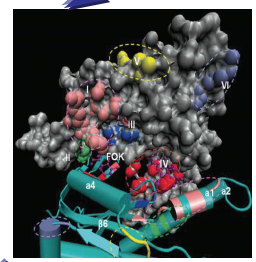


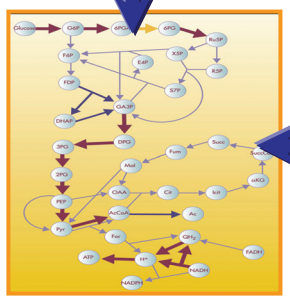
Protonated Parent Ions

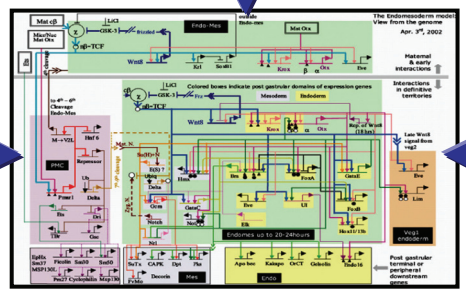
Observed Mass	Exact Mass	Error (ppm)	Mol. Formula	Amino Acid
106.04994	106.04987	-0.68	$(C_3H_7NO_2)H^+$	L-serine
120.06558	120.06552	-0.50	$(C_4H_9NO_2)H^+$	L-threonine
133.06082	133.06077	-0.38	$(C_4H_8N_2O_3)H^+$	L-asparagine
147.11284	147.11280	-0.27	$(C_6H_{14}N_2O_2)^+$	L-lysine
148.06046	148.06043	-0.20	$(C_5H_9NO_4)H^+$	L-glutamic acid
166.08626	166.08626	0.00	$(C_9H_{11}NO_2)H_+$	L-phenylalanine
205.09709	205.09715	0.29	$(C_{11}H_{12}N_2O_2)H^+$	L-tryptophan

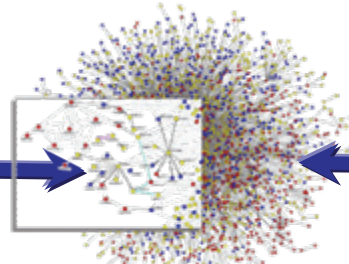




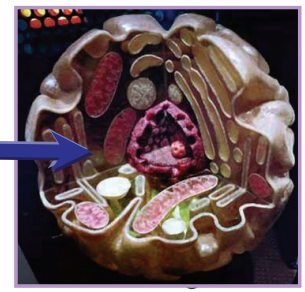






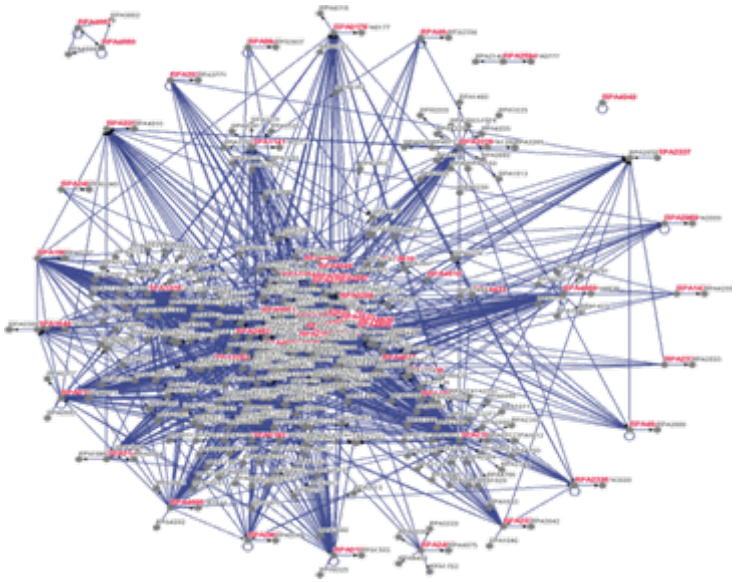


Interactomics



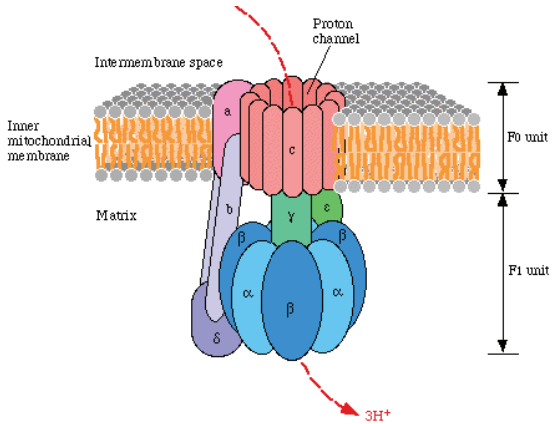
Q3: "Fuzzy" Visualization w/ Knowledge Priors

Mass spec pull-down experiments



- "Proactive" visualization
- Analogy: Query by example
- Analogy: Active learning

Prior-knowledge



ATP synthase complex

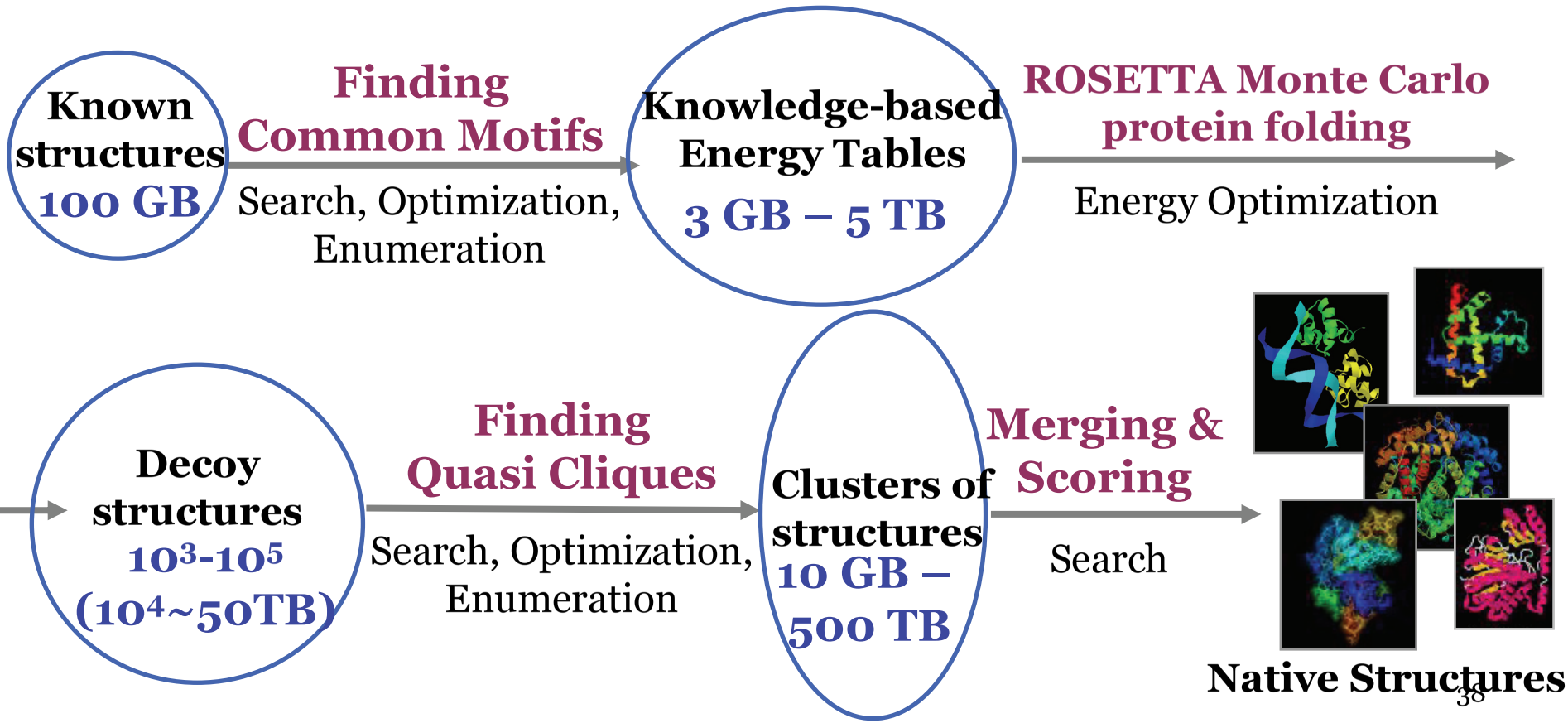
or Constrains

- Forms a (quasi) clique
- Contains at least 3 nodes
- Enriched by GO/KEGG terms
- Statistically significant

Q4: Visual Landscapes of Optimalities

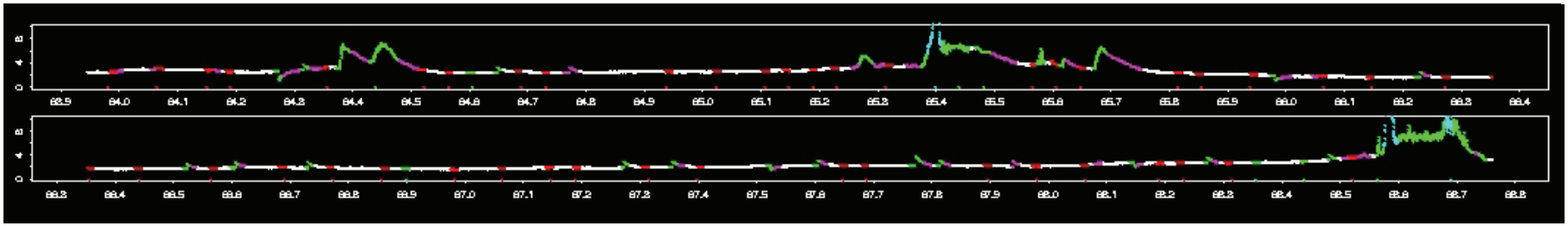
Each step is an *NP*-hard *combinatorial optimization problem* with different search heuristics.

Example: *Ab Initio* Prediction of Protein 3-d Structure

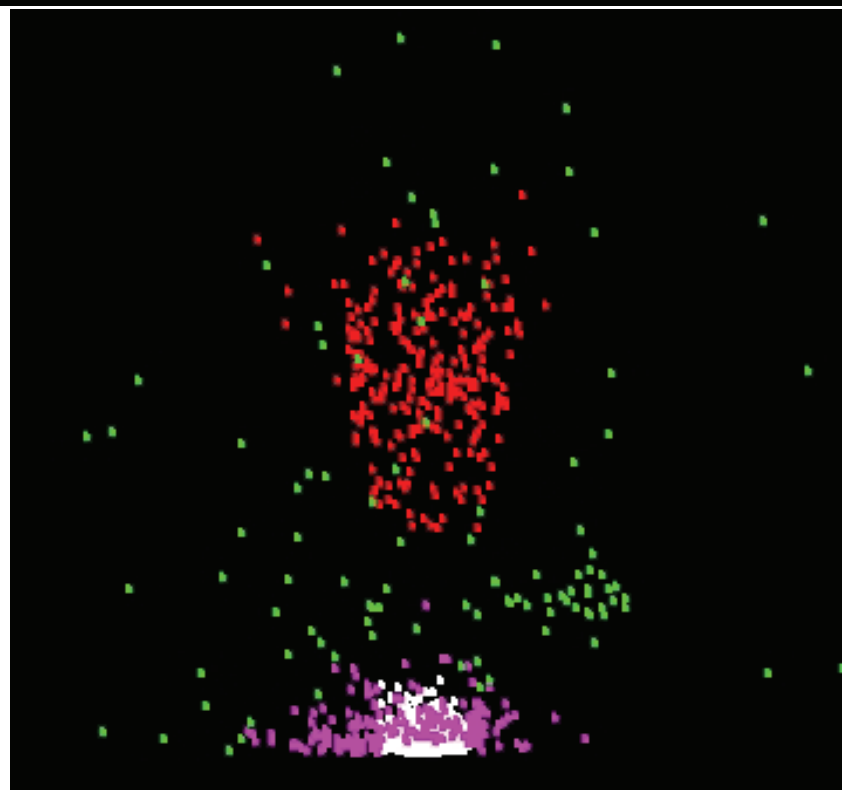


Q5: Modeling the Usual to Discover the Unusual

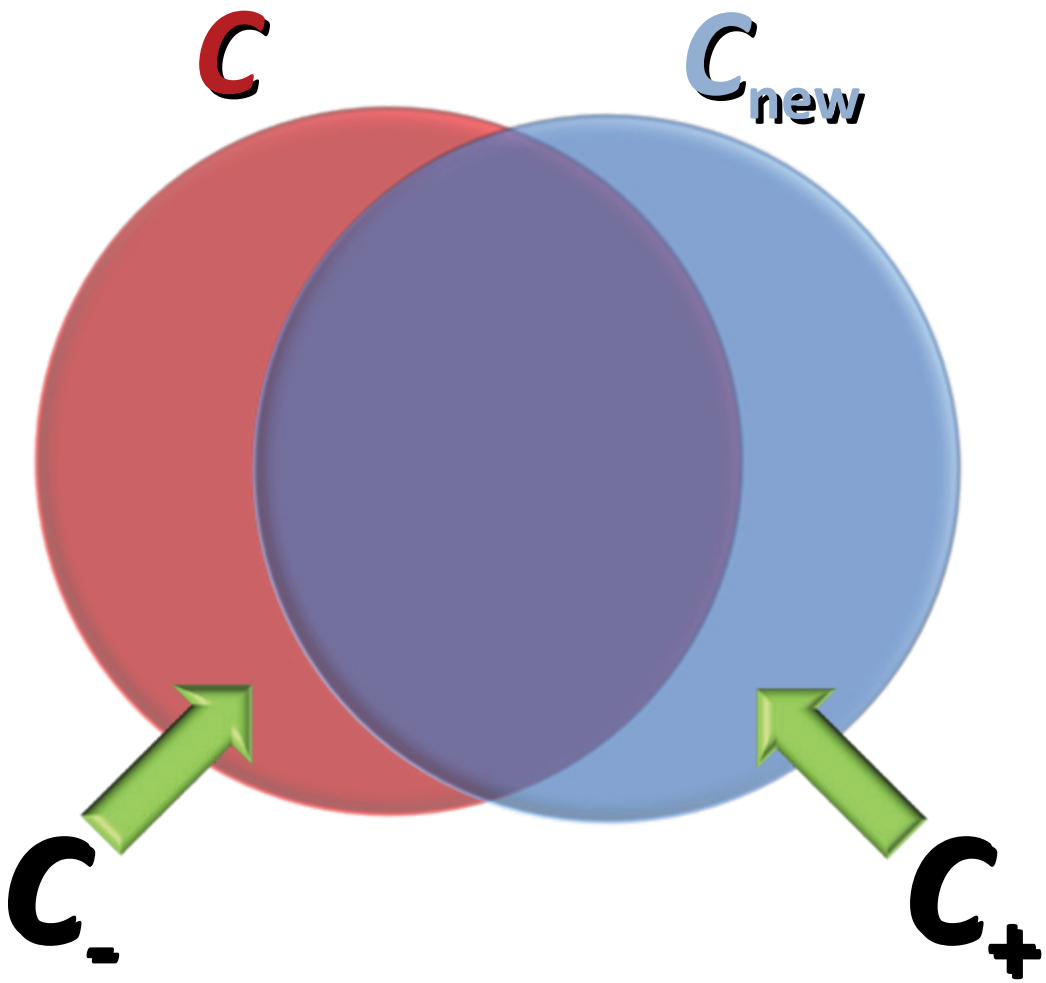
To reduce data & detect extreme events in global context.



1. Segment series (100 obs)
2. Fit simple local models to series
   (c_0, c_1, c_2, $||e||_\infty$, $||e||_2$)
3. Reduce data to model parameters
4. Select extremes for global analysis
5. Cluster the extremes (4)
6. Map back to series



Q6: “Perturbation” Visualization



Summary – Top 6 Quests for UV of Bio Data

1. **Comparative "Visual-omics"**
2. **Integrative "Visual-omics"**
3. **"Fuzzy" Visualization with Knowledge Priors**
4. **Visual Landascapes of Optimalities**
5. **Modeling the Usual to Discover the Unusual**
6. **"Perturbation" Visualization**

Acknowledgements

- **Juan Huang, UTK**
- **DOE ASCR**
- **DOE BER**