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

Paul Breimyer

Matt Shmidt

William Hendrix

Theresa-Marie Rhyne
Outline

• Biology 101
• Motivational Problems
• Cross-Cutting Ultrascale Visualization Challenges
  – Top 6 Quests for Next Generation UV of Biological Data
Genome (ROM): assembly code on how to build proteins
Instructions: A, C, T, G
3 variables ⇔ amino acid
Genome consists of genes

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

Guilt-by-Association goes global

Function assigned based on sequence similarity to another sequence with a

Function assigned based on sequence similarity to another sequence with a
Proteins Function Interactively

What we observe

- Genomics
- NMR
- X-ray
- Neutron Scattering
- Imaging

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

Genetic Manipulations
Cells & Tissues
Phenotypes
Populations
Environments
Time

Noisy

Large-Scale Data

++ and -- feedbacks

Non-linear correlations

Co-activator
Acetylation
Methylation
Phosphorylation

Activator
Protein Lysine
Data Describes Complex Natural Phenomena

How to untangle riddles of BioComplexity?

Complexity of biological systems comes from interconnections.

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

50 trans elements control single gene expression

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
A **phylogeny** is a tree representation for the evolutionary history relating the species of interest.
The Tree of Life for 10-100 Million Organisms
Computing the optimal phylogenetic tree based on the entire genome of 10 species will remain intractable even with peta-scale computers.

**Complexity**

Maximum likelihood

\[ C \sim n^2 m \]

where

- \( n \) is the number of contemporary species
- \( m \) is the 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.
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**

- Proteomics-based Predictions
  - Bait-prey background p-score
  - Prey-prey Jaccard, Dice, or Cosine score
  - Thresholds for optimal score

- Bait-Prey & Prey-Prey Affinities
  - Specific?
  - YES: Protein Affinity Network
  - NO: Experimental Pull-Down Data

- Genomic-Context Predictions
  - Bait-prey operon
  - Prey-prey operon
  - Rosetta Stone
  - Gene Neighborhood

- Specific?
  - NO: Protein Affinity Network
  - YES: Parallel MCE

- Protein Affinity Network
  - Perturbed?
  - NO: Database-Assisted MCE
  - YES: Database

- Merge Cliques
- Protein Complexes

- Parallel MCE
- Merge Cliques
- Protein Complexes

- Stop?
Visual Exploration of Genome-scale Interactome?

Find cliques

2,109 vertices
16,169 edges

Merge cliques

4123 modules

851 “common-target” associations

Module

Association

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.
Are Network Motifs Evolutionarily Conserved?

Protein Interaction Networks

Evolutionary History

Sequence alignment

Network Alignment

Drosophila

C.elegans

S.cerevisiae

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:**

<table>
<thead>
<tr>
<th>pH</th>
<th>T(°C)</th>
<th>Mannose metabolism</th>
<th>Xylose metabolism</th>
<th>Ethanol yield (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.5</td>
<td>35</td>
<td>+</td>
<td>+</td>
<td>89</td>
</tr>
<tr>
<td>5.5</td>
<td>30</td>
<td>+</td>
<td>+</td>
<td>95</td>
</tr>
<tr>
<td>5.5</td>
<td>30</td>
<td>—</td>
<td>+</td>
<td>84</td>
</tr>
</tbody>
</table>
Identification of Phenotype-Related Genes

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

Ultrascale Visualization Problem
Mathematical Laws of Metabolism

Chemical reaction:

\[ aA + cC \xrightarrow{v_i} eE + hH \]

Reaction Network:

Steady state flux vectors are in the Null Space of \( S \)

\[ S \cdot v = 0 \]
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, $C$

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

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 | Transcriptome | Physiome | Metabolome | Phenome | Morphome | Interactome | Glycome | Secretome | Ribonome | Orfeome | Regulome | Cellome | Operome | Transportome | Functome | Translatome | Pseudome | Foldome |
| **Comparative Metagenomics** | **Comparative Transcriptomics** | **Comparative Proteomics** |

![Graph showing Comparative Metagenomics, Transcriptomics, and Proteomics]

**Comparative Interactomics**

![Graph showing Comparative Interactomics]
Q2: Integrative Ultrascale "Visual-omics"

A collection of Parts
From Parts to Parts Assembly

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

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

Where are the parts located? Which parts interact?
Towards dynamic systems through Integrative “visual-omics”

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

Genomics

Transcriptomics

Proteomics

Metabolomics

Interactomics
Q3: “Fuzzy” Visualization w/ Knowledge Priors

Mass spec pull-down experiments

Prior-knowledge

- “Proactive” visualization
- Analogy: Query by example
- Analogy: Active learning

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**

- **Known structures**
  - 100 GB
  - Finding Common Motifs
  - Search, Optimization, Enumeration

- **Decoy structures**
  - $10^3$-$10^5$ (10^4~50TB)
  - Finding Quasi Cliques
  - Search, Optimization, Enumeration

- **Knowledge-based Energy Tables**
  - 3 GB – 5 TB
  - Clusters of structures
  - 10 GB – 500 TB

- **ROSETTA Monte Carlo protein folding**
  - Energy Optimization

- **Merging & Scoring**
  - Search

- **Native Structures**
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