About OMICS Group

OMICS Group International is an amalgamation of Open Access publications and worldwide international science conferences and events. Established in the year 2007 with the sole aim of making the information on Sciences and technology 'Open Access', OMICS Group publishes 400 online open access scholarly journals in all aspects of Science, Engineering, Management and Technology journals. OMICS Group has been instrumental in taking the knowledge on Science & technology to the doorsteps of ordinary men and women. Research Scholars, Students, Libraries, Educational Institutions, Research centers and the industry are main stakeholders that benefitted greatly from this knowledge dissemination. OMICS Group also organizes 300 International conferences annually across the globe, where knowledge transfer takes place through debates, round table discussions, poster presentations, workshops, symposia and exhibitions.

UNO BIOINFORMATICS

About OMICS Group Conferences

OMICS Group International is a pioneer and leading science event organizer, which publishes around 400 open access journals and conducts over 300 Medical, Clinical, Engineering, Life Sciences, Pharma scientific conferences all over the globe annually with the support of more than 1000 scientific associations and 30,000 editorial board members and 3.5 million followers to its credit.

OMICS Group has organized 500 conferences, workshops and national symposiums across the major cities including San Francisco, Las Vegas, San Antonio, Omaha, Orlando, Raleigh, Santa Clara, Chicago, Philadelphia, Baltimore, United Kingdom, Valencia, Dubai, Beijing, Hyderabad, Bengaluru and Mumbai.

UNO BIOINFORMATICS

High Performance Computing and Private Clouds for Bioinformatics Applications



OMICS 4rd International Conference on Proteomics & Bioinformatics

August 2014

Hesham H. Ali
UNO Bioinformatics Core Facility
College of Information Science and Technology

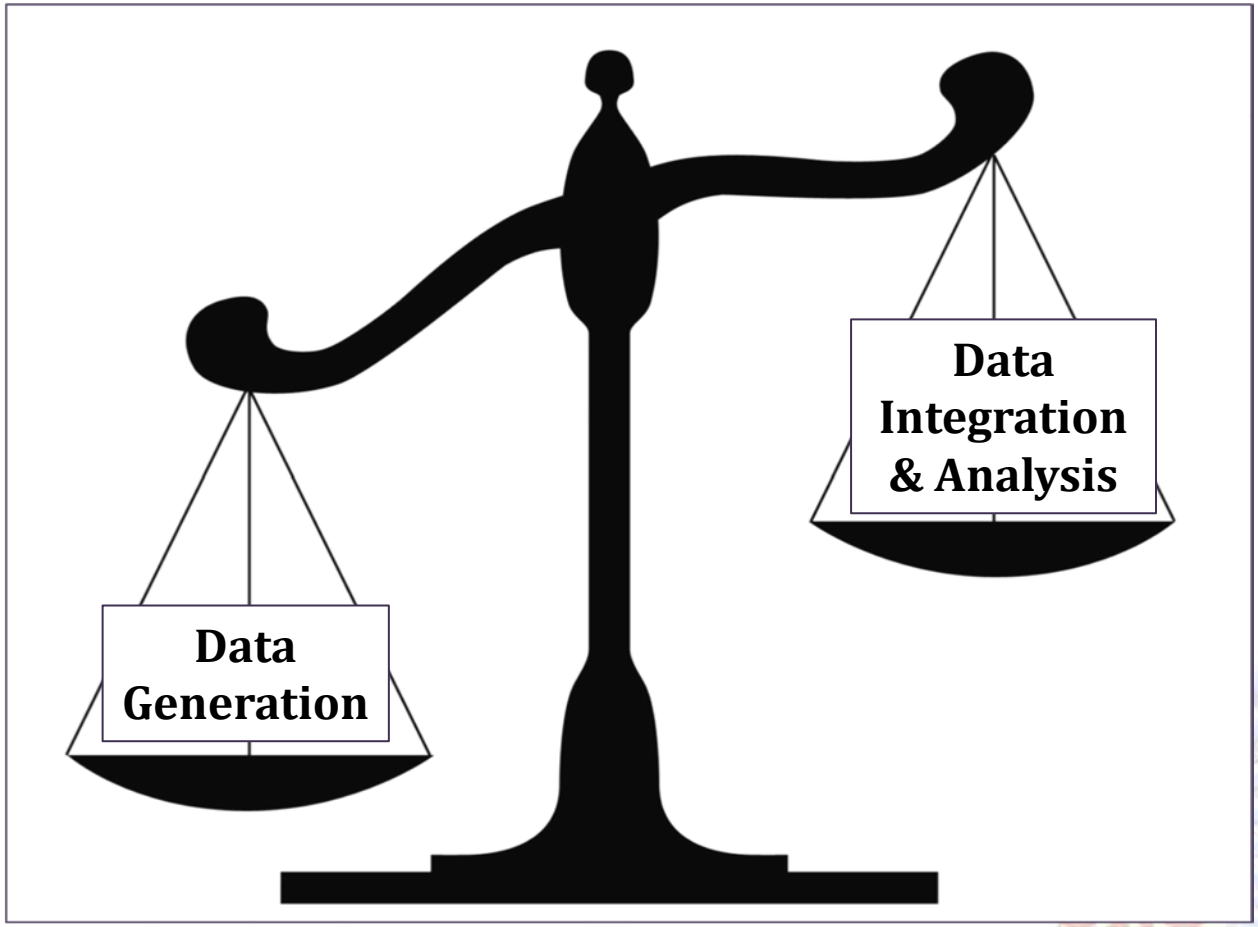
UNIVERSITY OF NEBRASKA AT OMAHA

UNO BIOINFORMATICS

Big Data Analytics and Biomedical Research

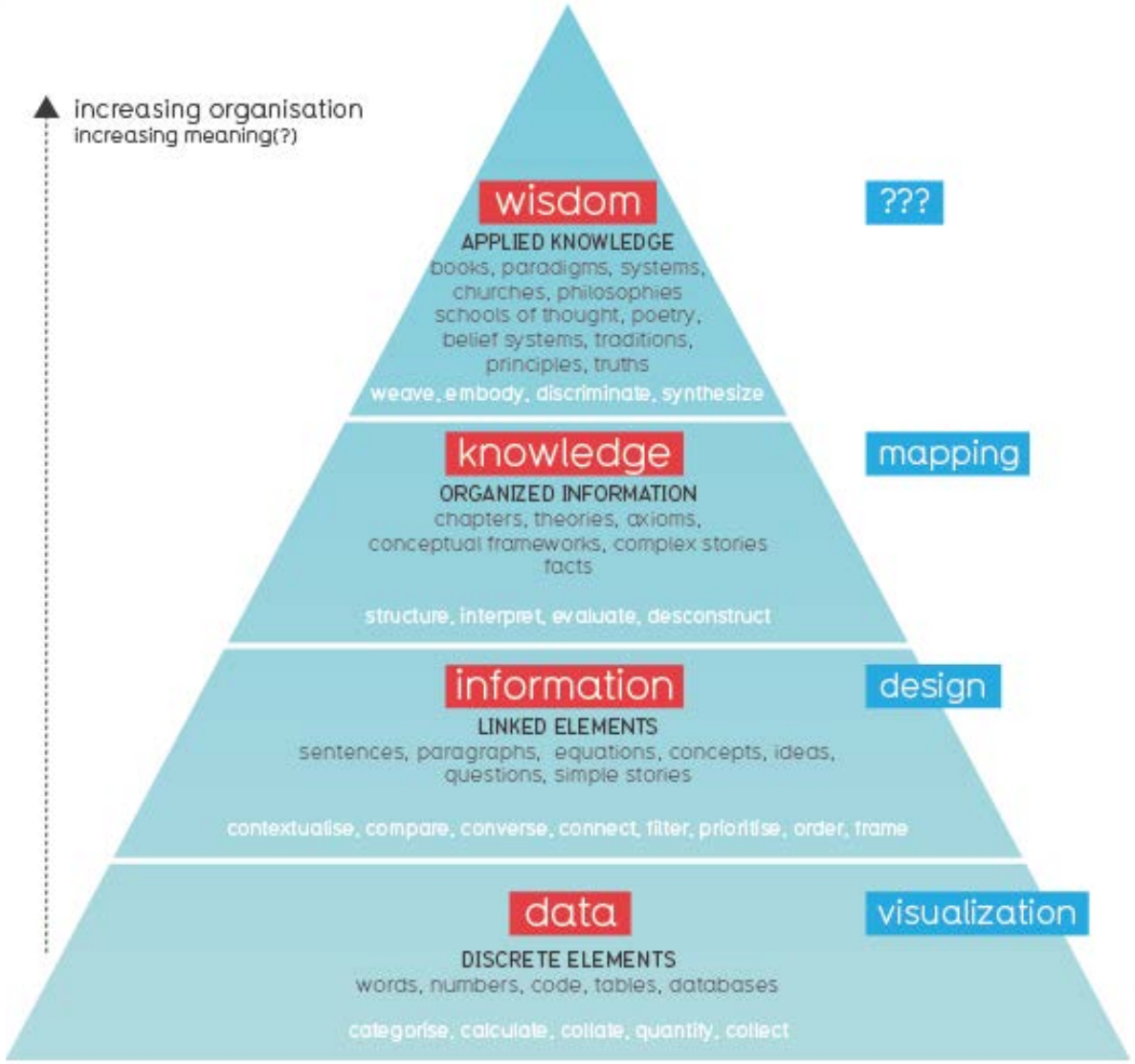
- Big data analytics and real time decision support/response are of prime importance in many critical applications
- Current decision support systems mostly work with passive/static information with limited context awareness and lack of system-wide holistic perspective
- There is a need for advanced tools that integrate various static and dynamic data and utilize network analysis algorithms to develop effective decision support systems

Data Generation vs. Data Integration/Analysis



BIOINFORMATICS

From Raw Data to Decision Support Systems



http://www.ritholtz.com/blog/wp-

What is driving Big Data?

- Volume



- Velocity



- Variety

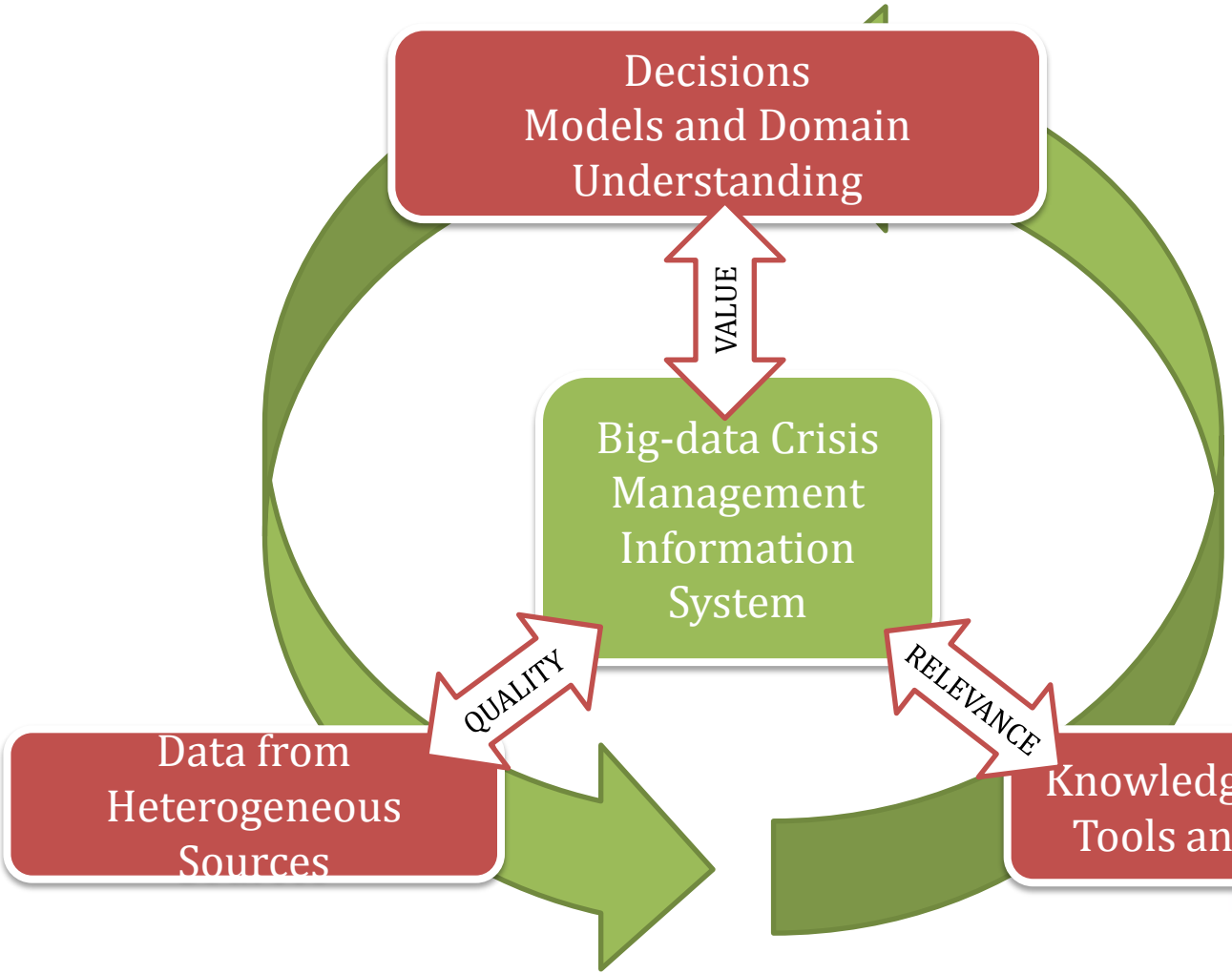


- Veracity



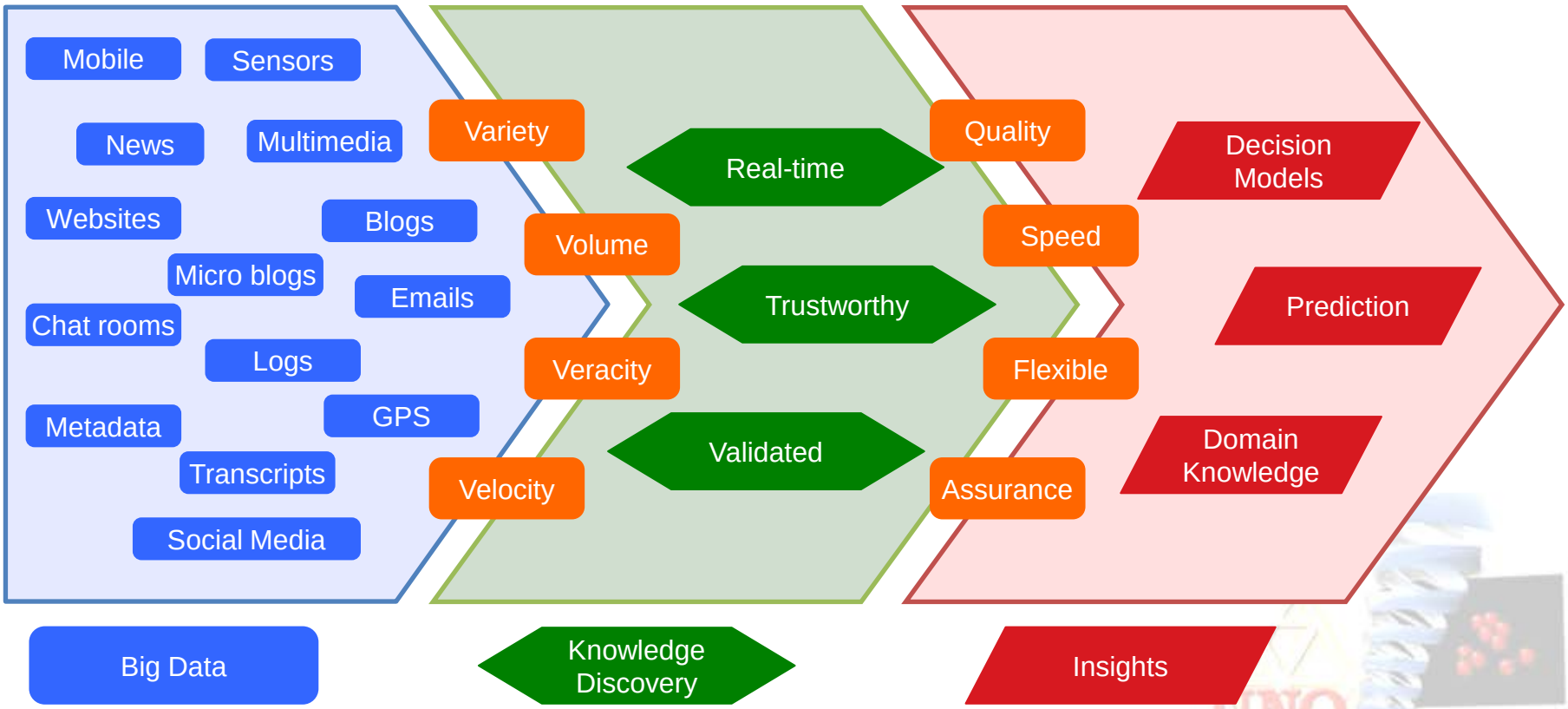
UNO BIOINFORMATICS

Big Data
Big Solution CMIS Blueprint



UNO
BIOINFORMATICS

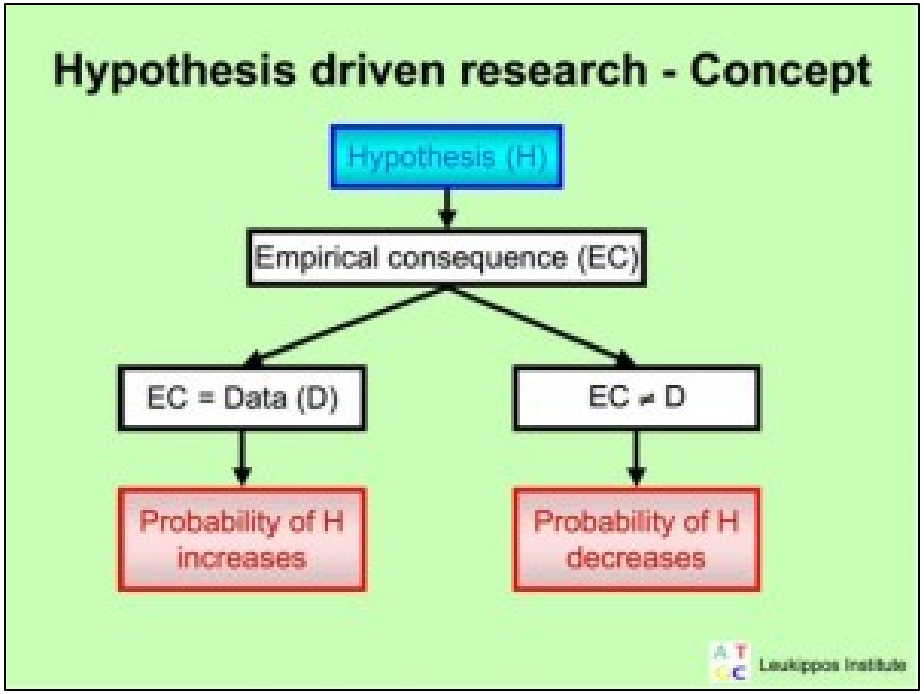
Big Data Enabled Real-time CMIS

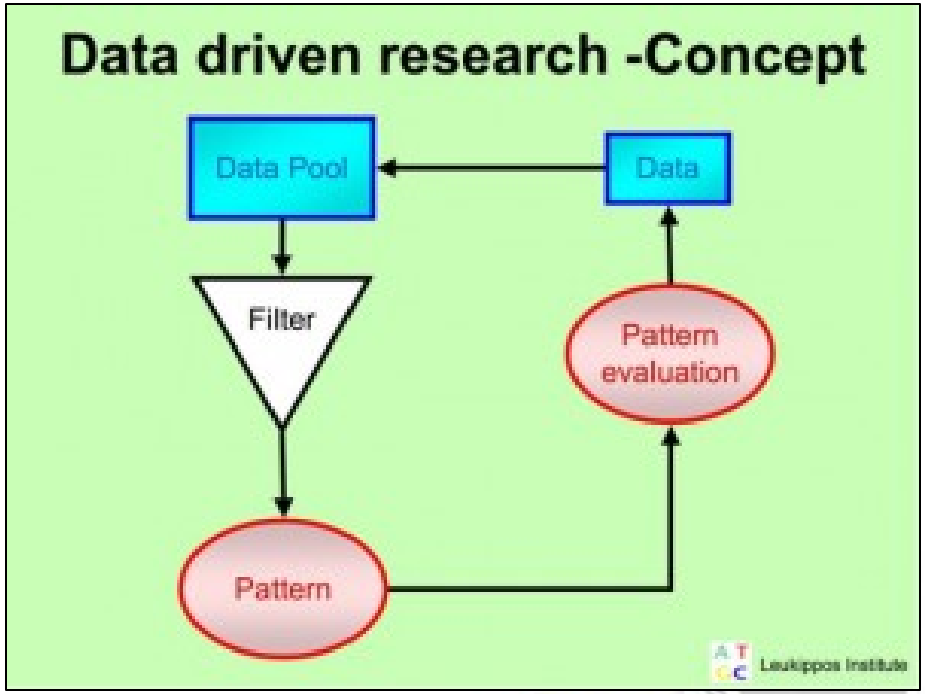


UNO BIOINFORMATICS

A New Game Changer

Data driven research vs. Hypothesis driven research





UNO BIOINFORMATICS

http://www.genome-engineering.com/from-hypothesis-to-data-driven-research.html

The "Systems Biology" Approach

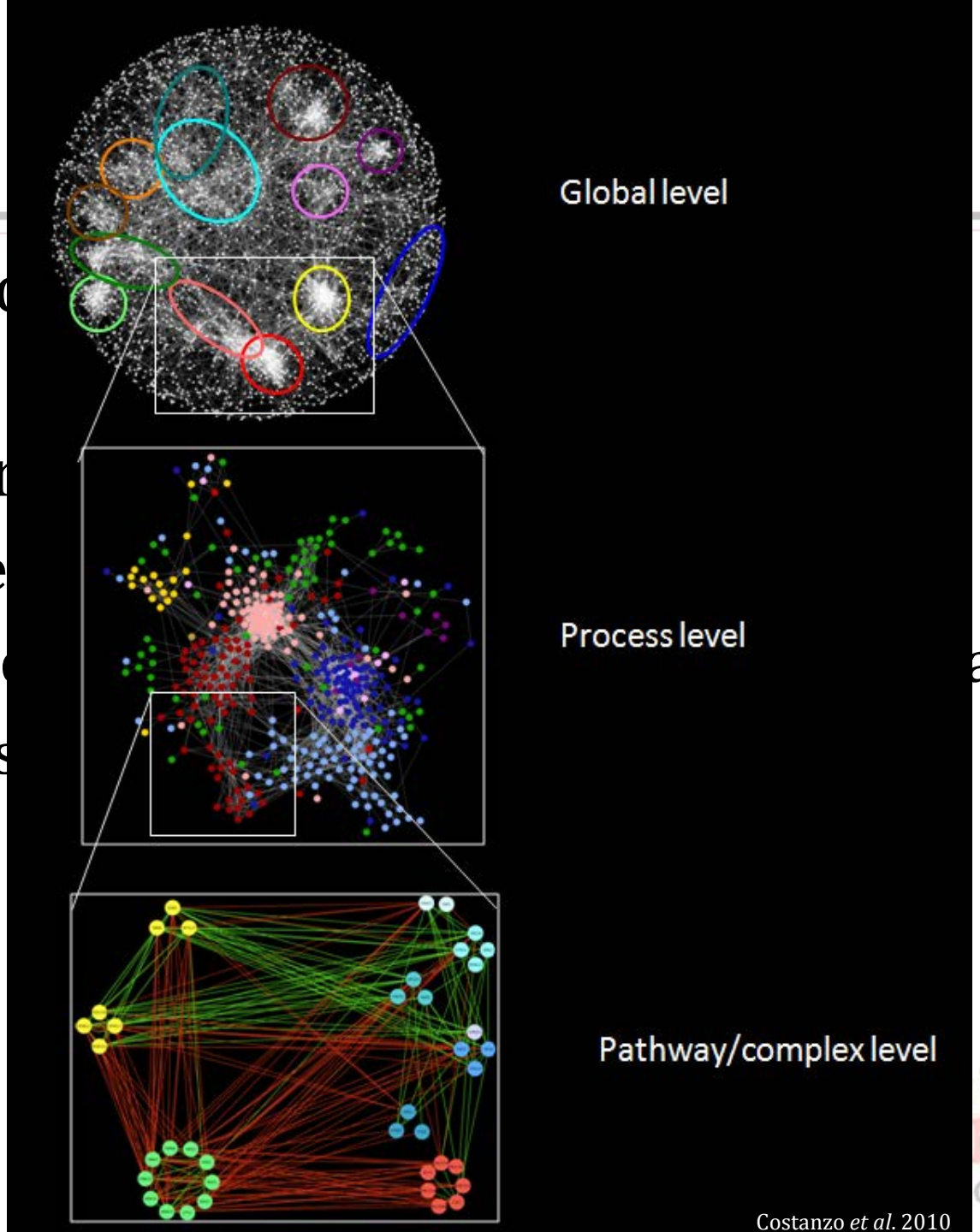
- Integrated Approach:
  - Networks model relationships, not just elements
  - Discover groups of relationships between genes

- Discovery
  - Examine changes in systems
    - Normal vs. diseased
    - Young vs. old
    - Stage I v. State II v. Stage III v. Stave IV



- Netwo
- Syster
  - Inte
  - Use
  - Pers

ase



Costanzo *et al.* 2010

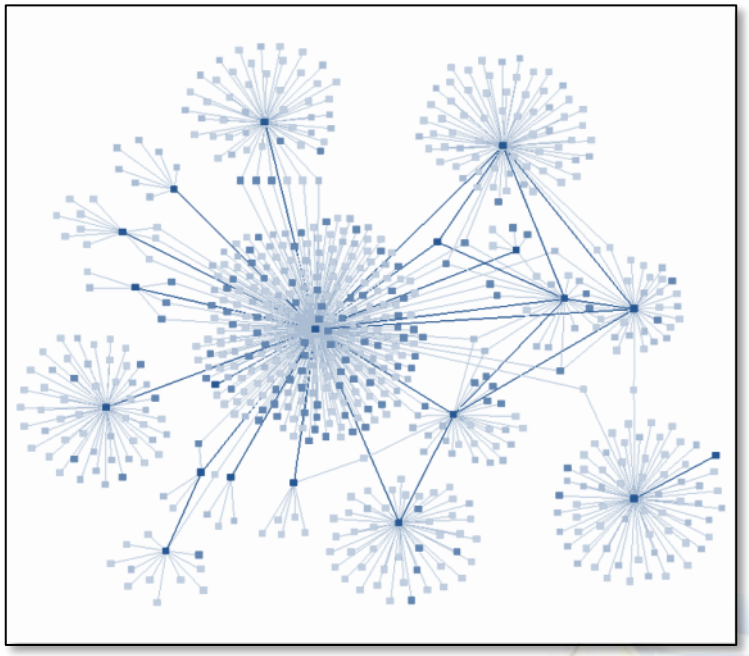
NO
OINFORMATICS

Biological Networks

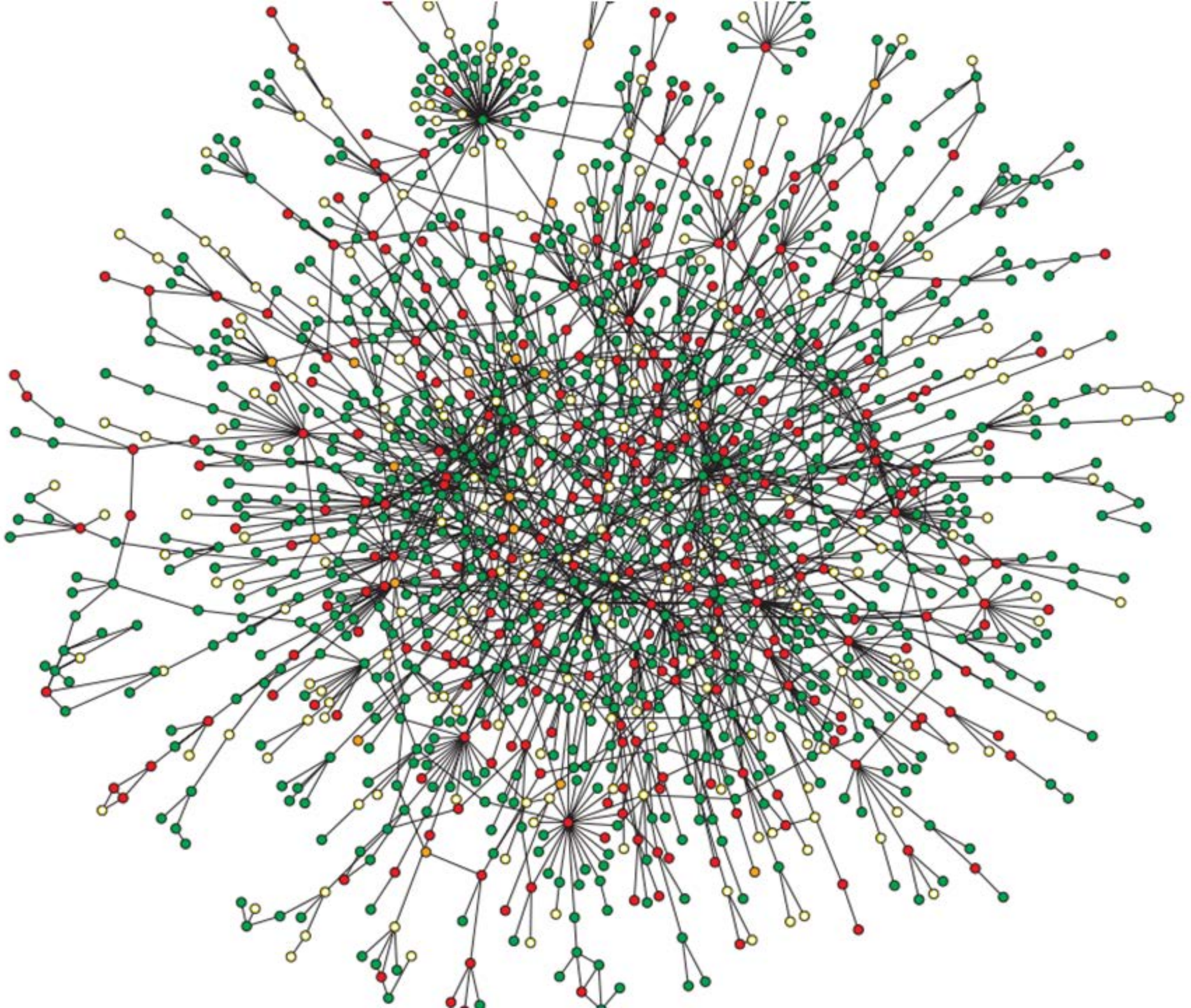
- A biological network represents elements and their interactions

- Nodes → elements
- Edges → interactions

- Can represent multiple types of elements and interactions



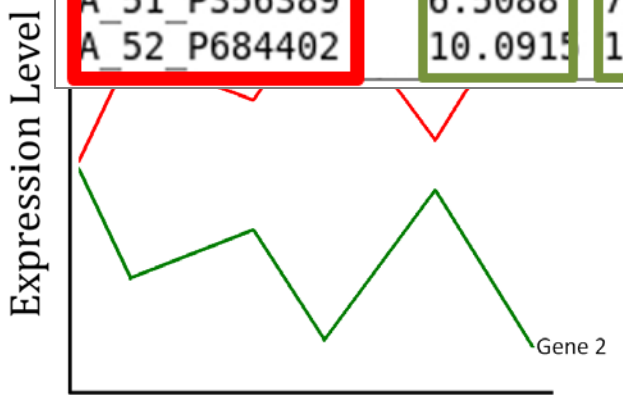
UNO
BIOINFORMATICS



Correlation Networks

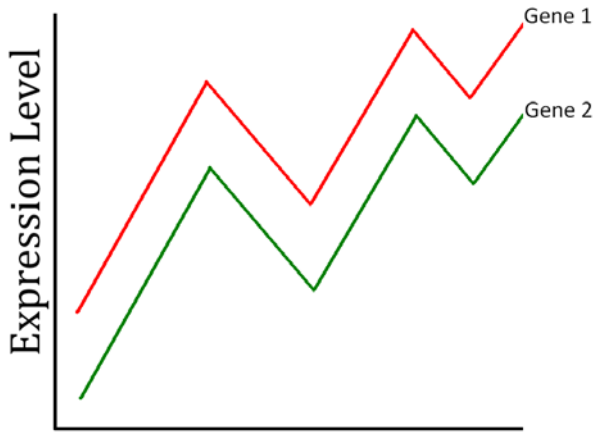
Gene ID	Sample1	Sample2	Sample3	Sample4	
A_52_P616356	5.9813	6.0079	5.9525	7.2753	6.2
A_52_P580582	7.7845	7.7512	8.0943	8.3608	8.1
A_52_P403405	5.9301	6.5153	6.0526	7.1707	6.1
A_52_P819156	6.5748	6.8645	6.981	7.4937	6.9
A_51_P331831	7.1732	7.8754	7.7632	8.1875	7.6
A_51_P430630	6.0661	6.4009	5.9525	7.1208	6.0
A_52_P502357	5.936	6.3206	5.9525	7.1819	6.2
A_52_P299964	6.3452	6.8025	6.6457	7.3445	6.7
A_51_P356389	6.5088	7.0545	7.2346	7.631	7.0
A_52_P684402	10.0915	10.7124	10.2245	10.377	10.

- 10,000-45,000+ probes
- UNO Blackforest cluster
- HCC Firefly





Correlation = 1

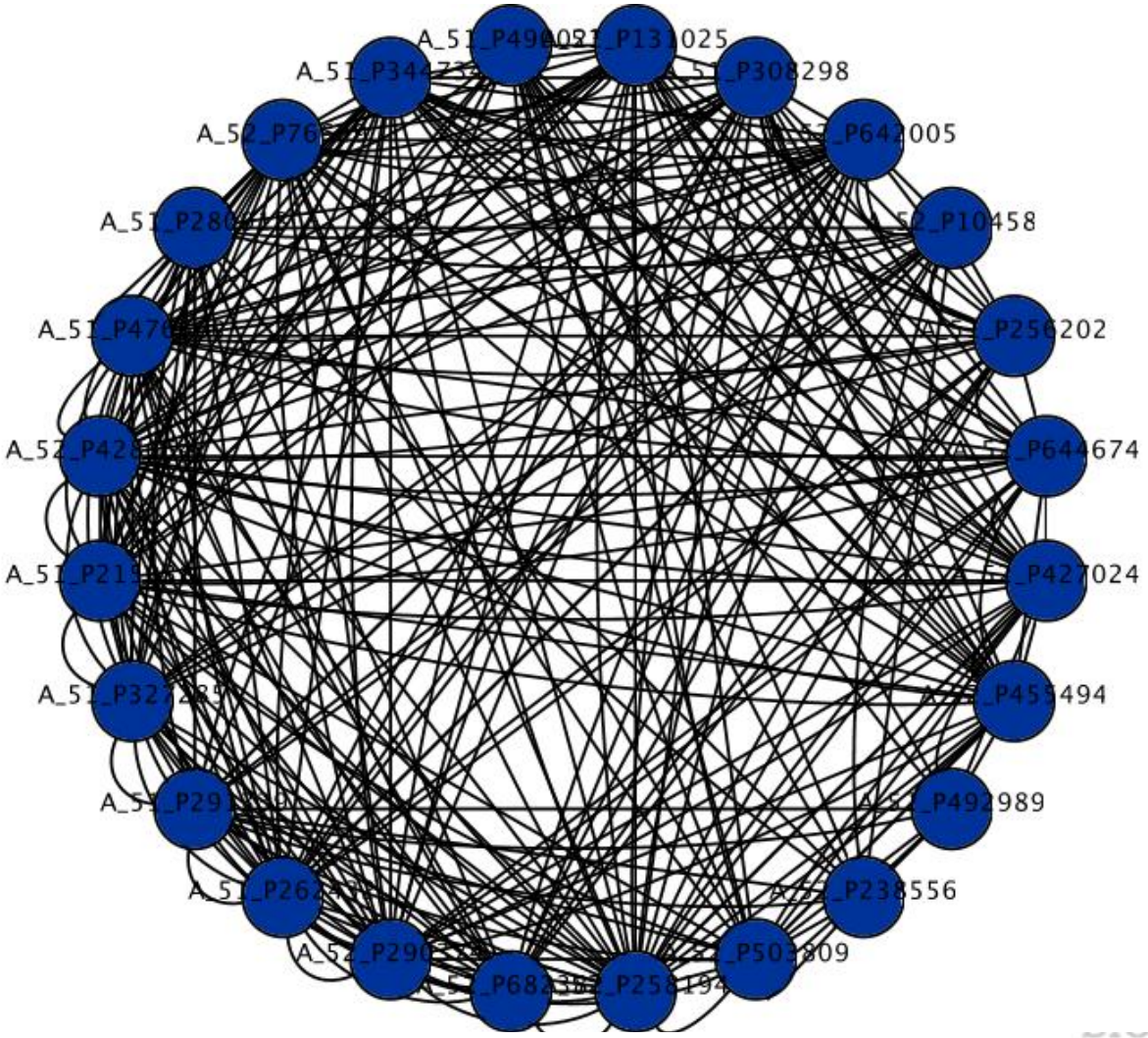


BIOINFORMATICS

Correlation Networks

24 node sample

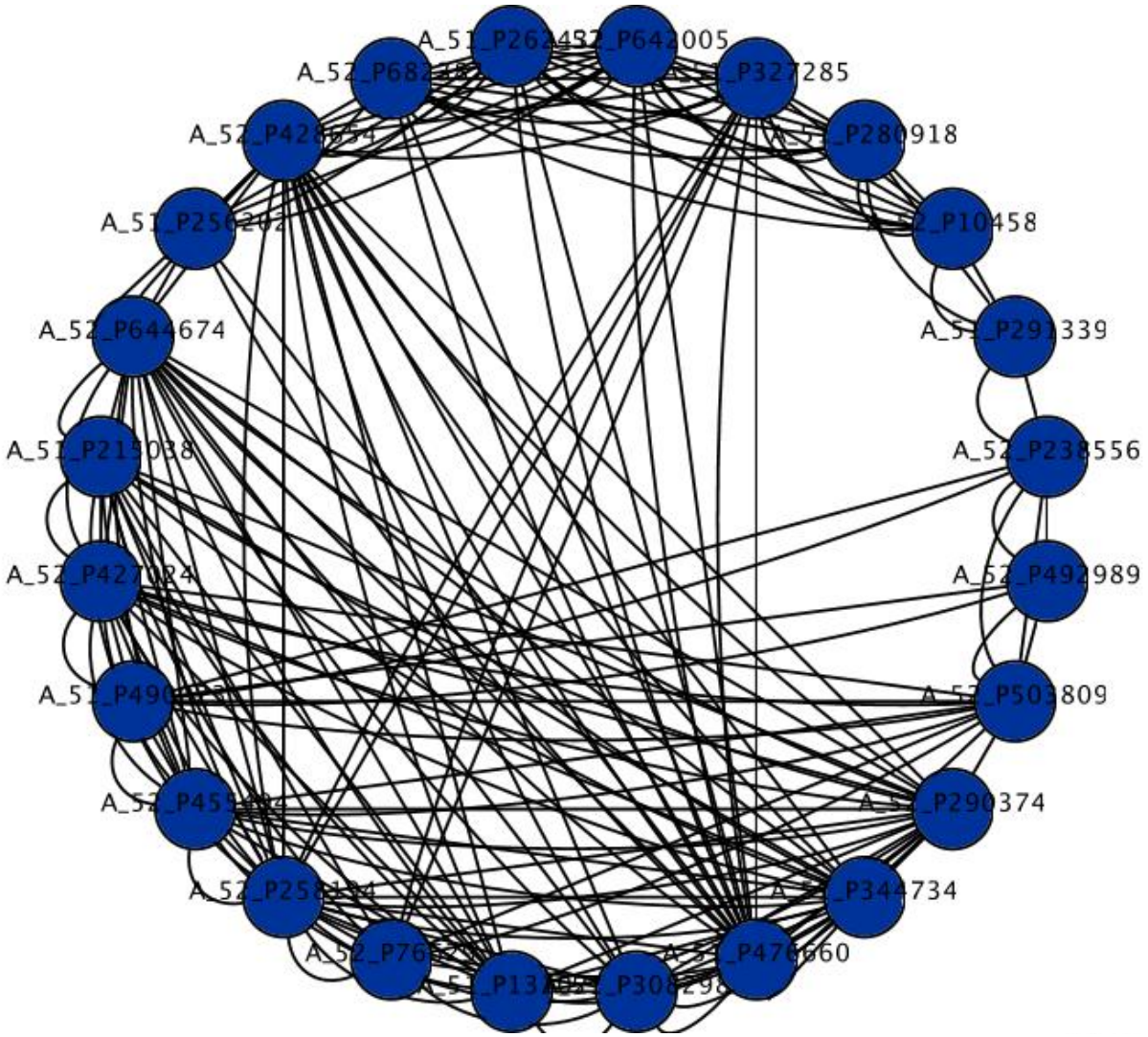
Threshold: 0.00-1.00



Correlation Networks

24 node sample

Threshold: 0.30-1.00

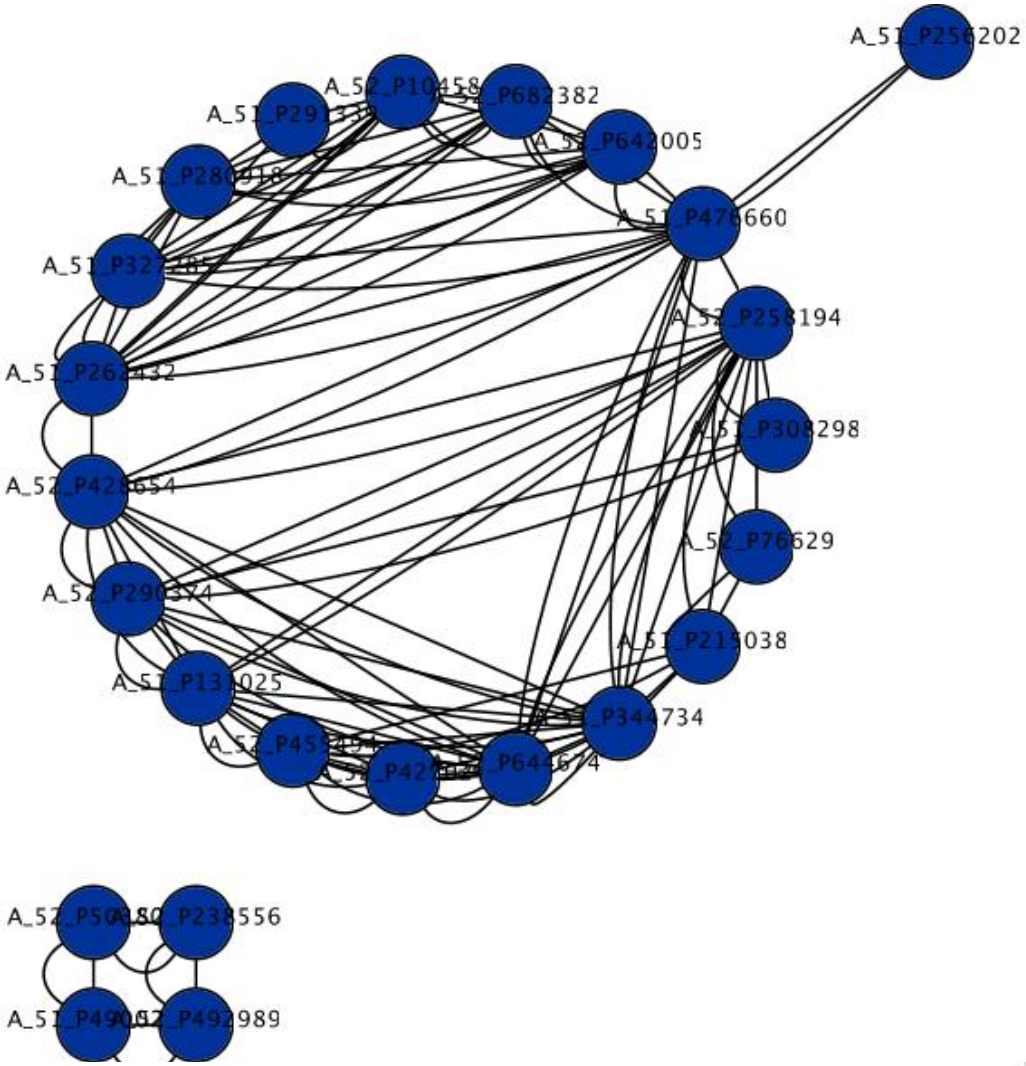


NFORMATICS

Correlation Networks

24 node sample

Threshold: 0.50-1.00

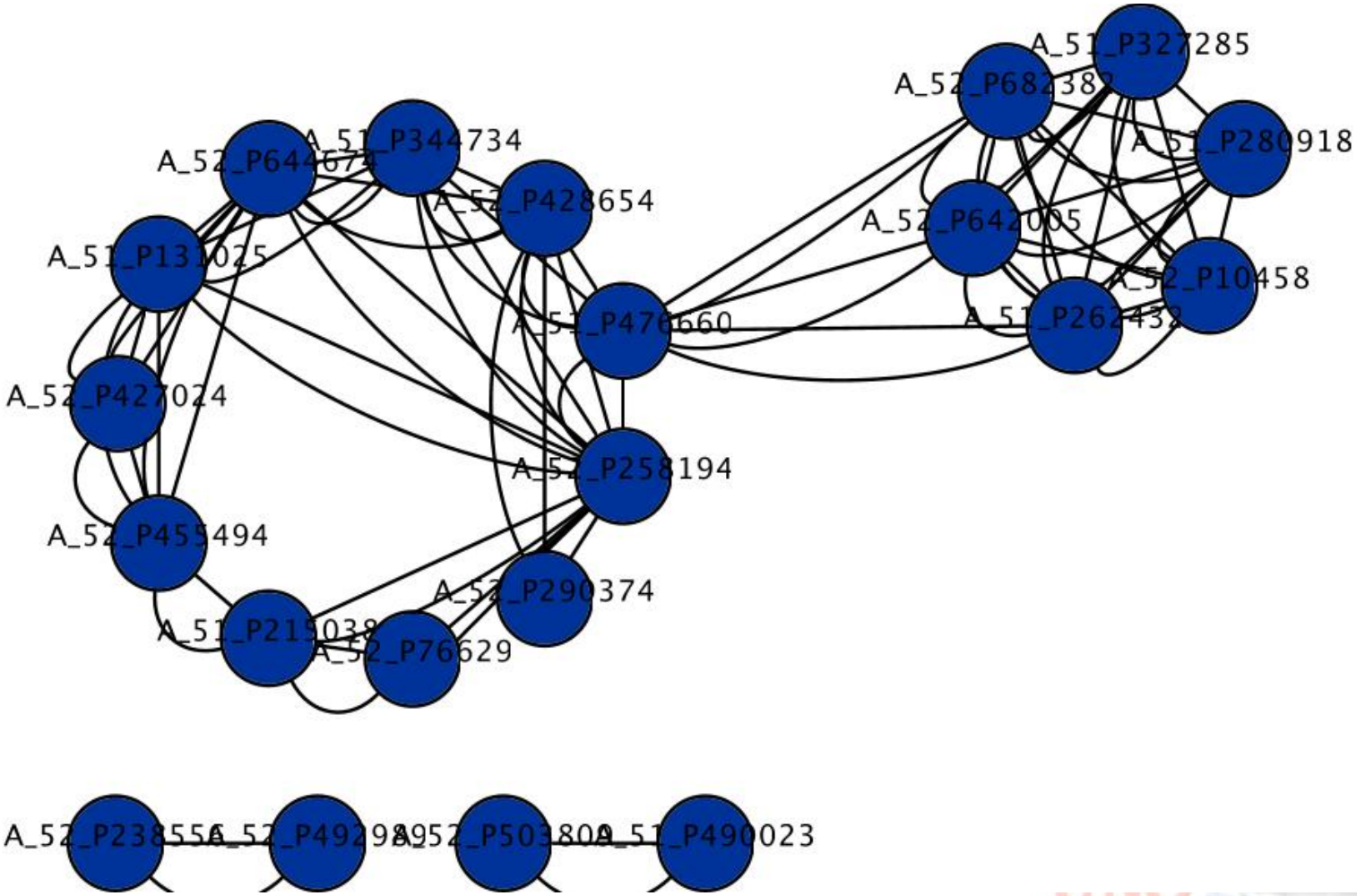


INO
BIOINFORMATICS

Correlation Networks

24 node sample

Threshold: 0.60-1.00

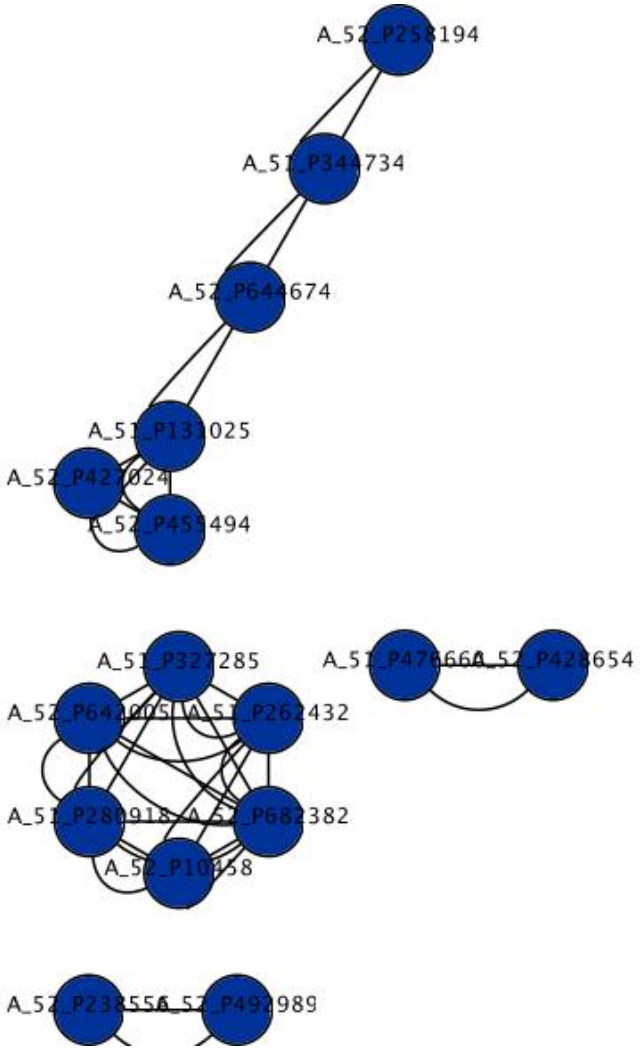


BIOINFORMATICS

Correlation Networks

24 node sample

Threshold: 0.80-1.00



UNO BIOINFORMATICS

Correlation Network Applications

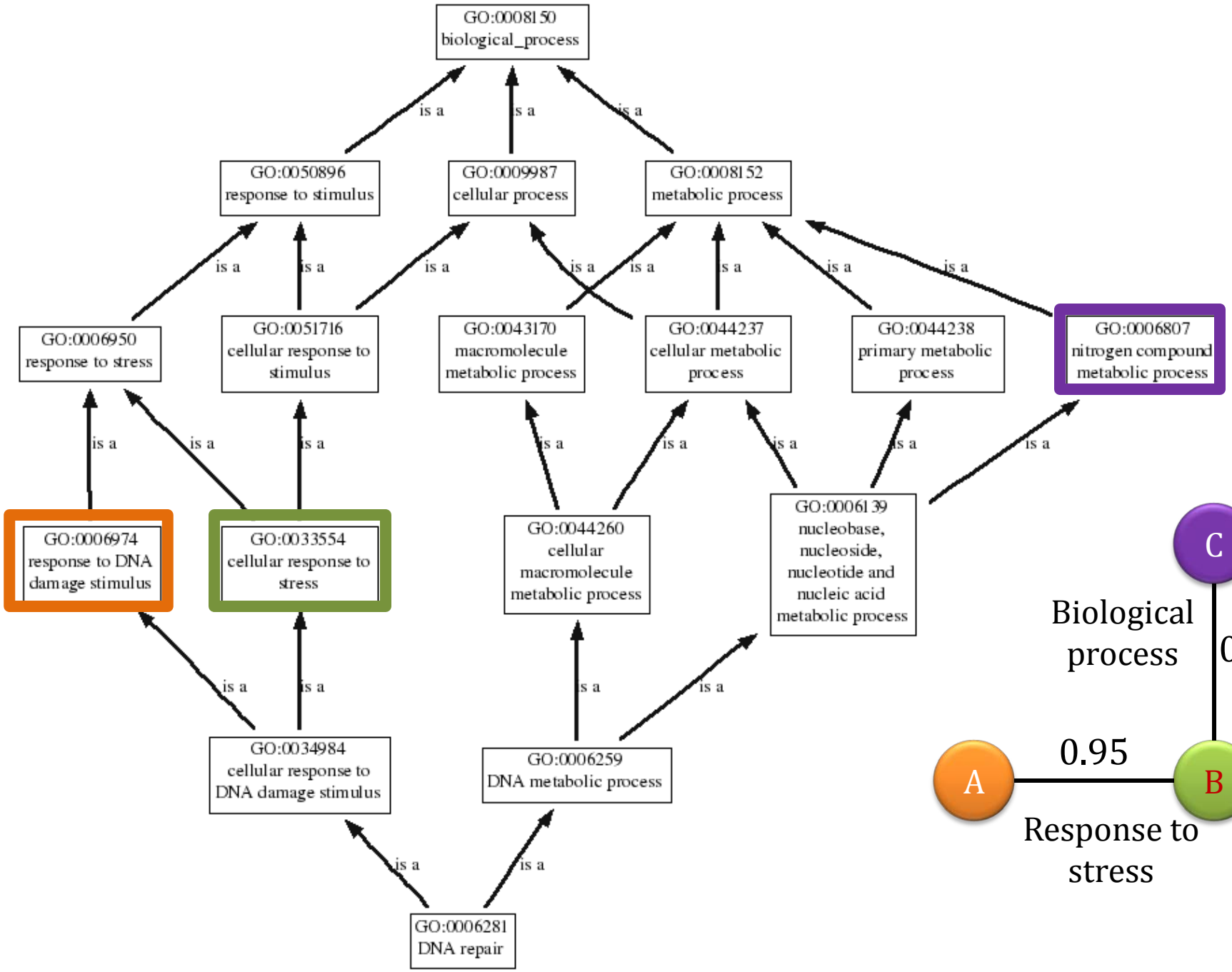
- "Versus" analysis
  - Normal vs. disease
  - Times/environments
- Model for high-throughput data
  - Especially useful in microarrays
- Identification of groups of causative genes
  - Ability to rank based on graph structure
  - Identify sets of co-regulated, co-expressed genes

UNO BIOINFORMATICS

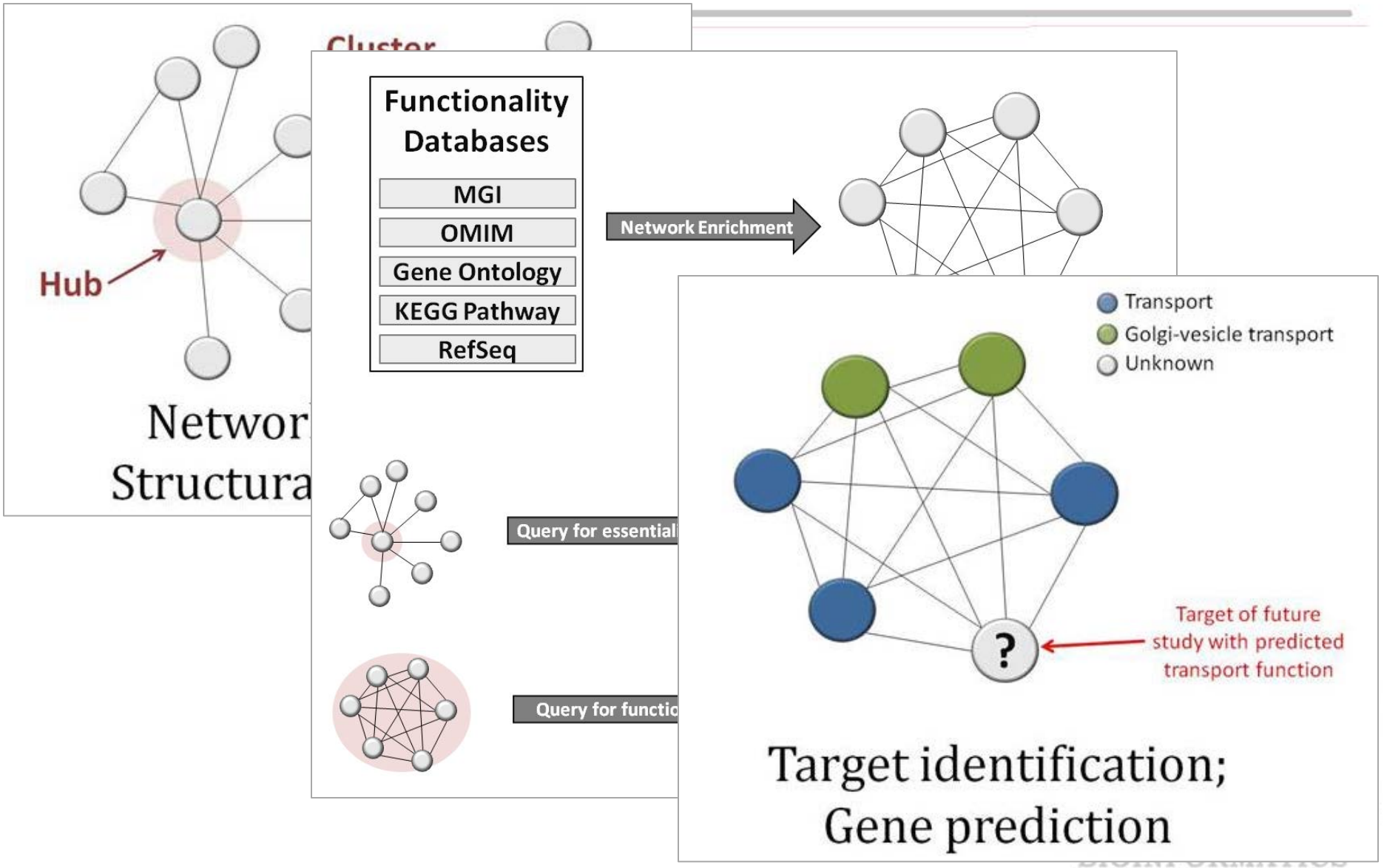
Power of Correlation Analysis

- Correlation versus Causation
- Correlation networks
- Casting the net wide – signal and noise
- The use of enrichment before obtaining information and after for validation

UNO
BIOINFORMATICS


GO:0008150
biological_process
GO:0050896
response to stimulus
GO:0009987
cellular process
GO:0008152
metabolic process
GO:0006950
response to stress
GO:0051716
cellular response to
stimulus
GO:0043170
macromolecule
metabolic process
GO:0044237
cellular metabolic
process
GO:0044238
primary metabolic
process
GO:0006807
nitrogen compound
metabolic process
GO:0006974
response to DNA
damage stimulus
GO:0033554
cellular response to
stress
GO:0044260
cellular
macromolecule
metabolic process
GO:0006139
nucleobase,
nucleoside,
nucleotide and
nucleic acid
metabolic process
GO:0034984
cellular response to
DNA damage stimulus
GO:0006259
DNA metabolic process
GO:0006281
DNA repair
is a
C
Biological
process
0.35
0.95
A
B
Response to
stress

Integrated Data Model



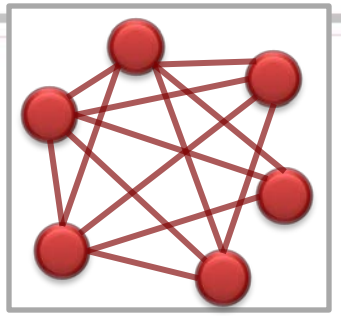
Structures & their Functions

Key Hypothesis:
Network structures correspond to key cellular structures

Local Network Structures

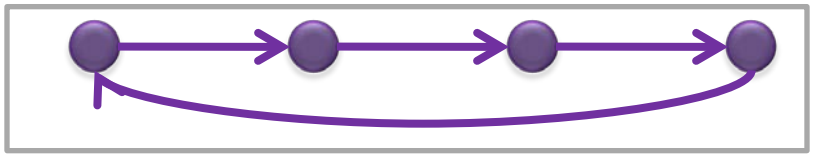
- **Cliques**

  Protein complexes, regulatory modules



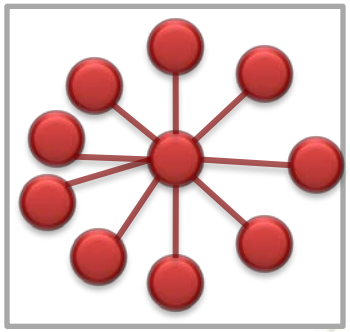
- **Pathways**

  Signaling cascades

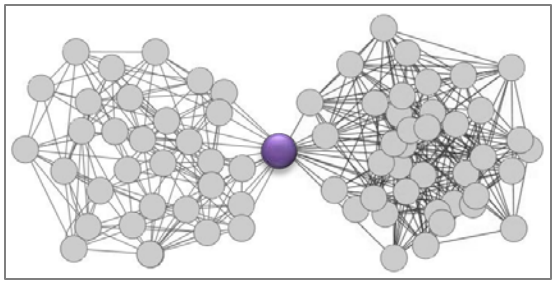


- **Hubs**

  Regulators, TFs, active proteins



- **Articulation points**



UNO BIOINFORMATICS

Case Study in Aging

- With aging, certain behaviors decrease
  - Eating, drinking, activity levels
- Observed gene expression changes in the hypothalamus
  - Can we capture these expression changes?
  - Can we correlate these changes to behavioral decreases?
- Goal: Identify temporal biological relationships
  - Progression of disease
  - Effect of pharmaceuticals on systems of the body
  - Aging



Case Study in Aging

- 5 sets of temporal gene expression data

Strain	Gender	Tissue Type	Ages
BalbC	Male	Hypothalamus	Young, mid-age, aged
CBA	Male	Hypothalamus	Young, mid-age, aged
C57_J20	Male	Hypothalamus	Young, aged
BalbC	Female	Hypothalamus	Young, aged
BalbC	Female	Frontal cortex	Young, aged

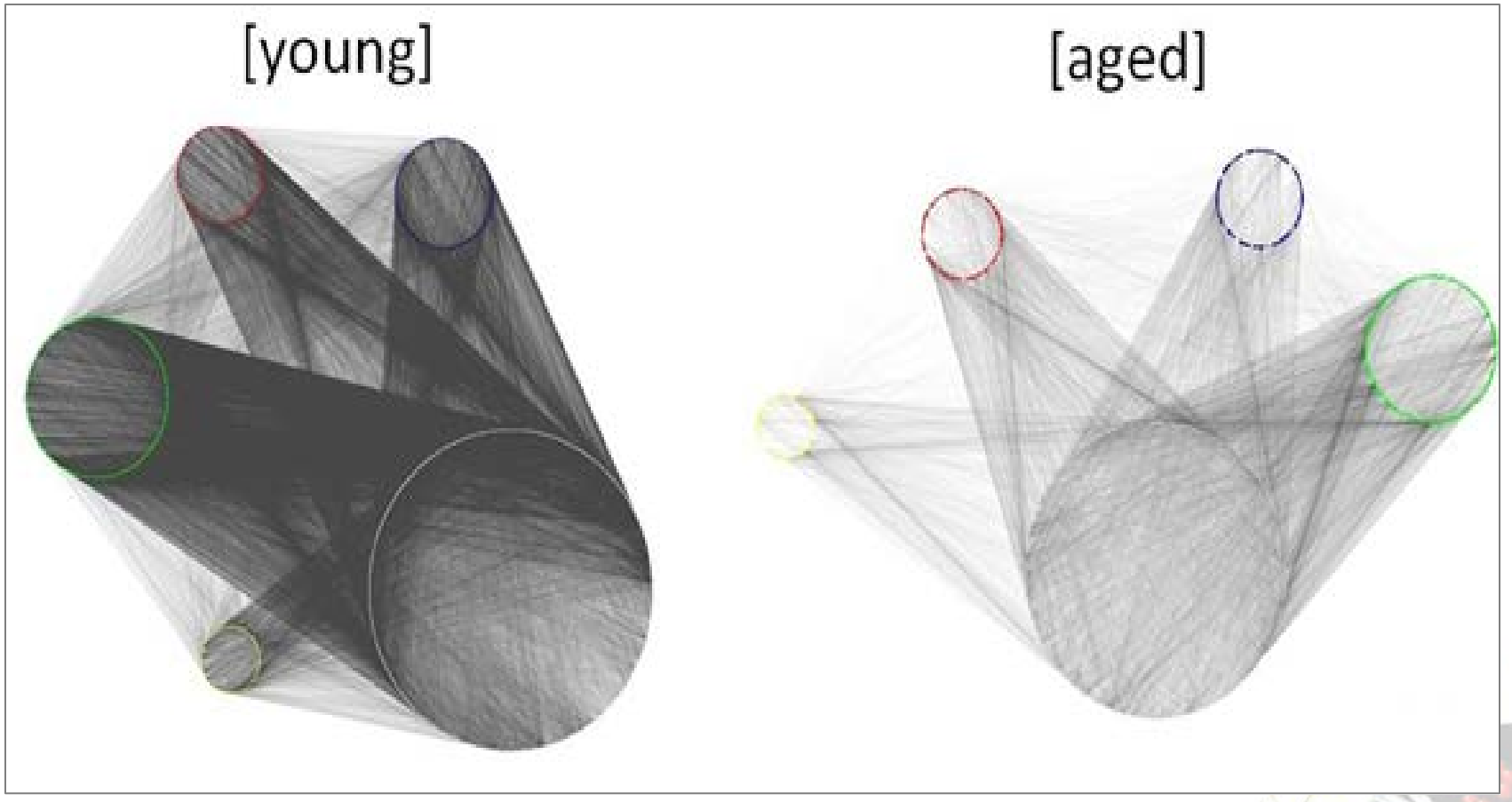
UNO BIOINFORMATICS

Hub Lethality

- Young Male BalbC Mouse
  - 12/20 hubs tested for *in vivo* knockout
    - 8/12 lethal phenotype
    - 4/12 non-lethal but system-affecting
    - 0/12 no observed phenotype
- Aged Male BalbC Mouse
  - 11/20 hubs tested for *in vivo* knockout
    - 7/11 lethal phenotype pre-/peri-natally
    - 3/11 non-lethal but system-affecting
    - 1/11 no observed phenotype (Aldh3a1)

UNO BIOINFORMATICS

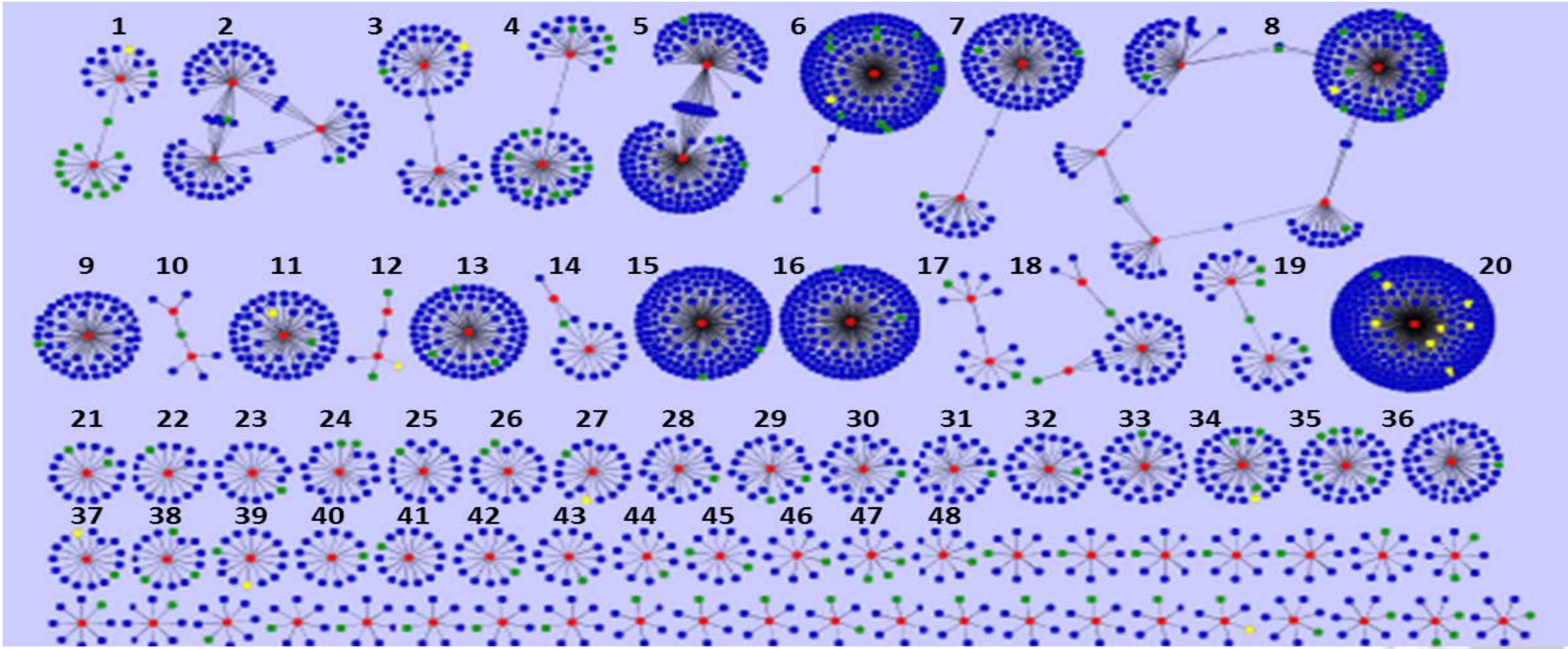
Correlation Networks and Aging



Network density changes over time for cellular functions

UNO BIOINFORMATICS

Discovery of Ageing Related Genes: Use of Protein Domains in Graph based analytics



(Blue)= Human Protein ; (Red)= Protein Domain; (Green)= Model Organism aging protein ; Yellow: Aging proteins discovered in humans

BIOINFORMATICS

Role of Drug Addition in HIV Treatment

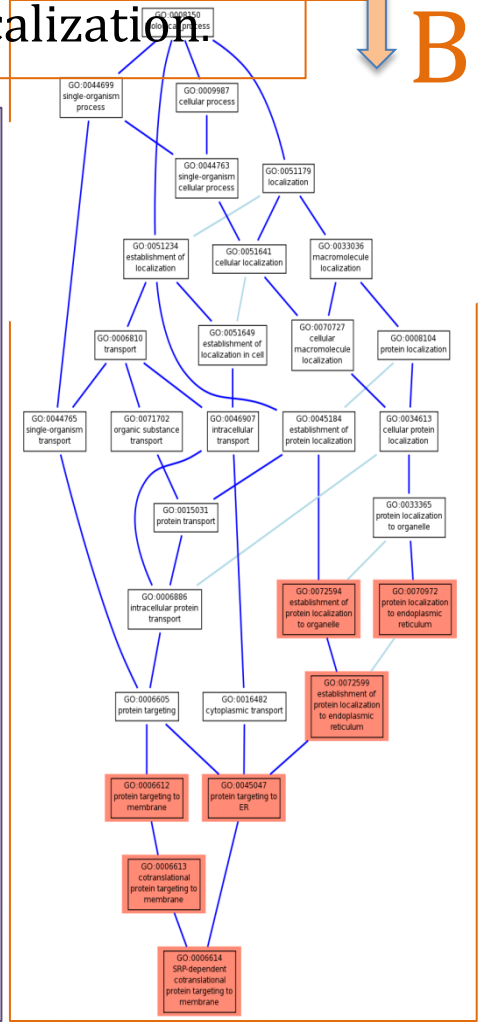
- Use correlation network to identify differential clusters between infected and control individuals
- Methamphetamine is a major drug of abuse with reported high use by HIV-infected groups with users have higher risk of infection
- Impact on nervous system is higher when Methamphetamine is used by HIV infected individual (neuronal injury)

UNO
BIOINFORMATICS

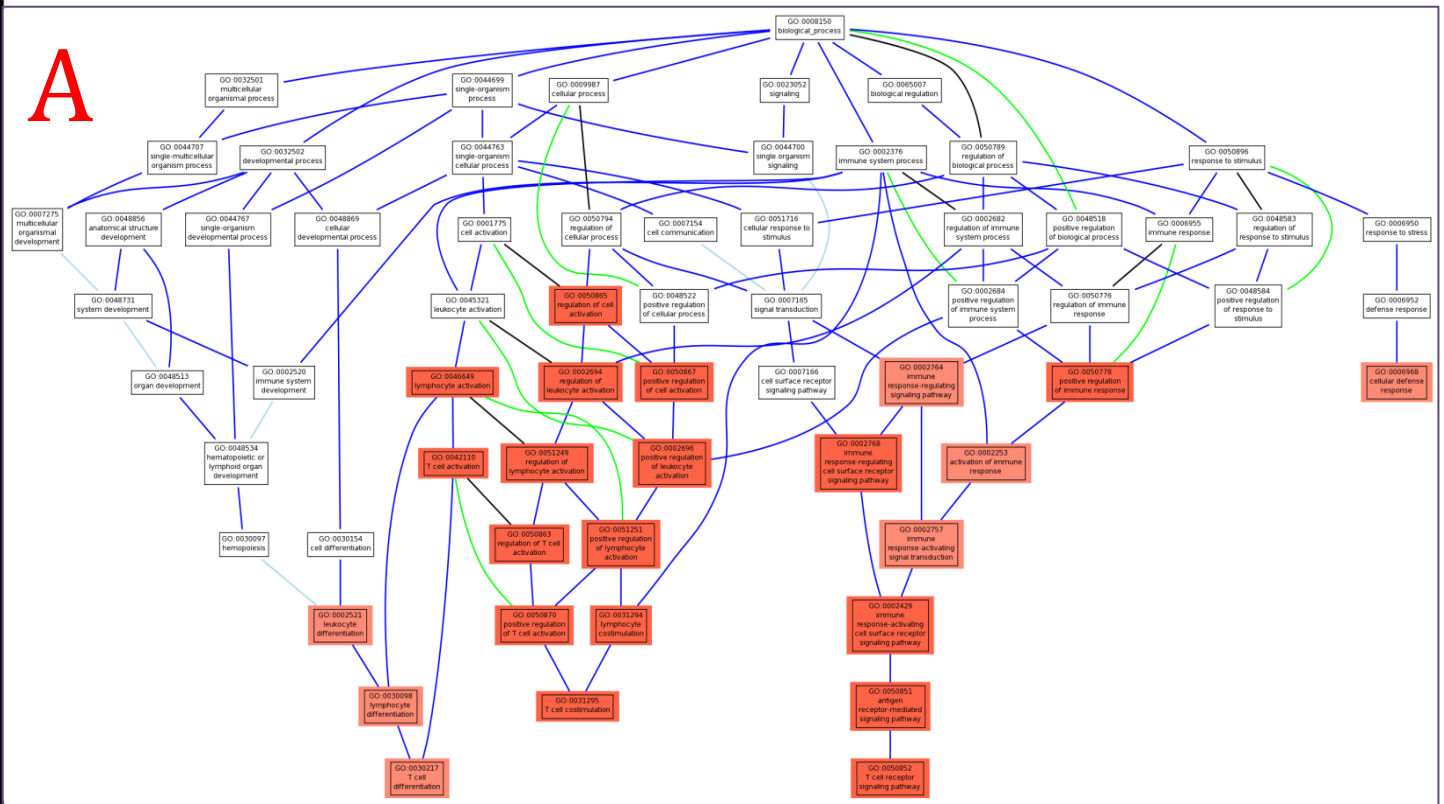
Clusters Enriched in Specific Functions

Functions enriched in clusterB are protein targeting and localization.

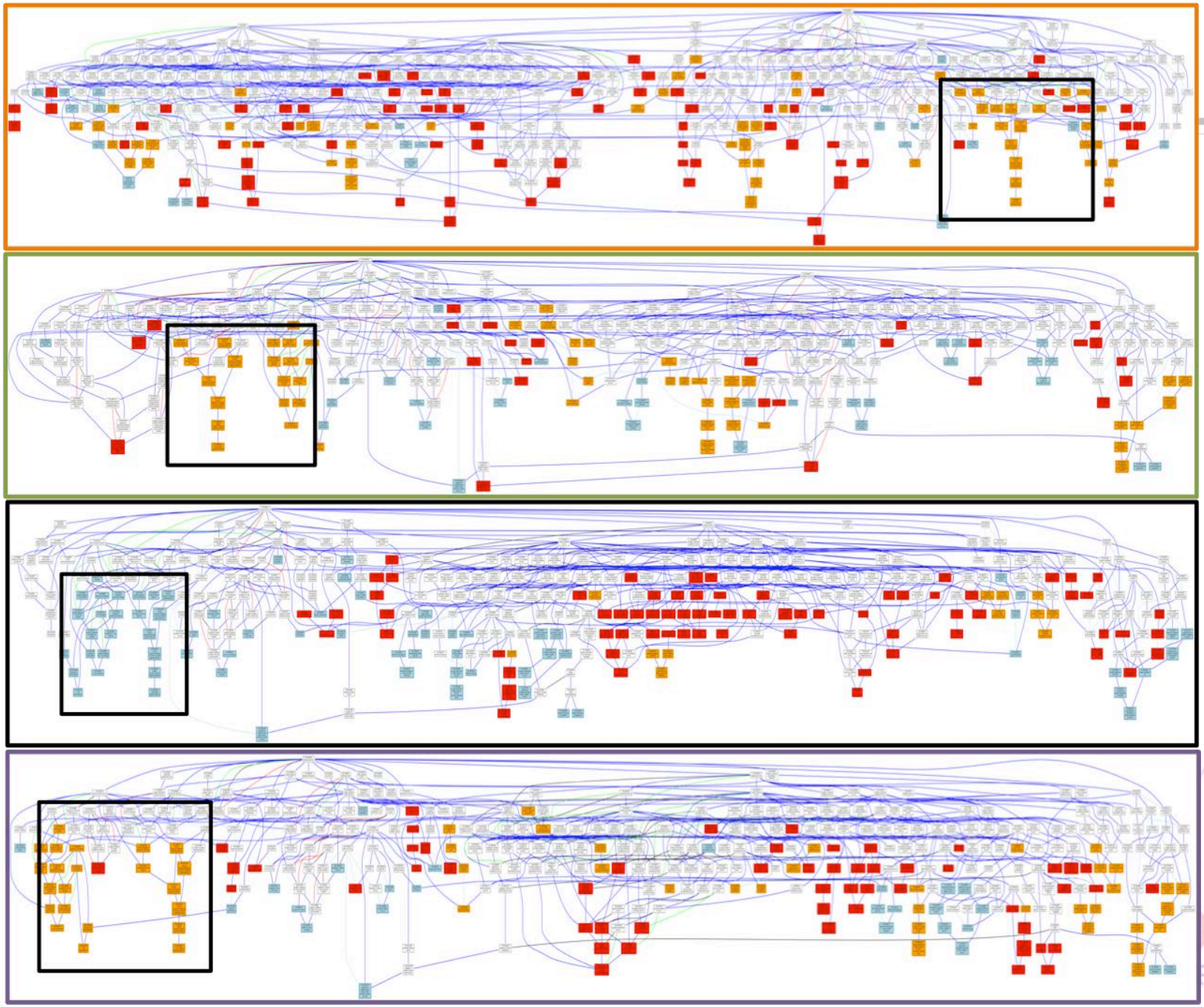
B



A



Functions enriched in clusterA are all immune response specific.



Infected vs. Uninfected

HIV treatment vs. Uninfected

Infected + Meth vs. Uninfected

Infected + Meth+ Treatment vs. Uninfected

How to implement this stuff? Computer Science Issues

- High Performance Computing
  - Beyond surface-level adaptation of previous algorithms
- Security and Privacy
  - Cloud Security
- Graph Algorithms and Network Analysis



HPC and Big Data in Biological Networks

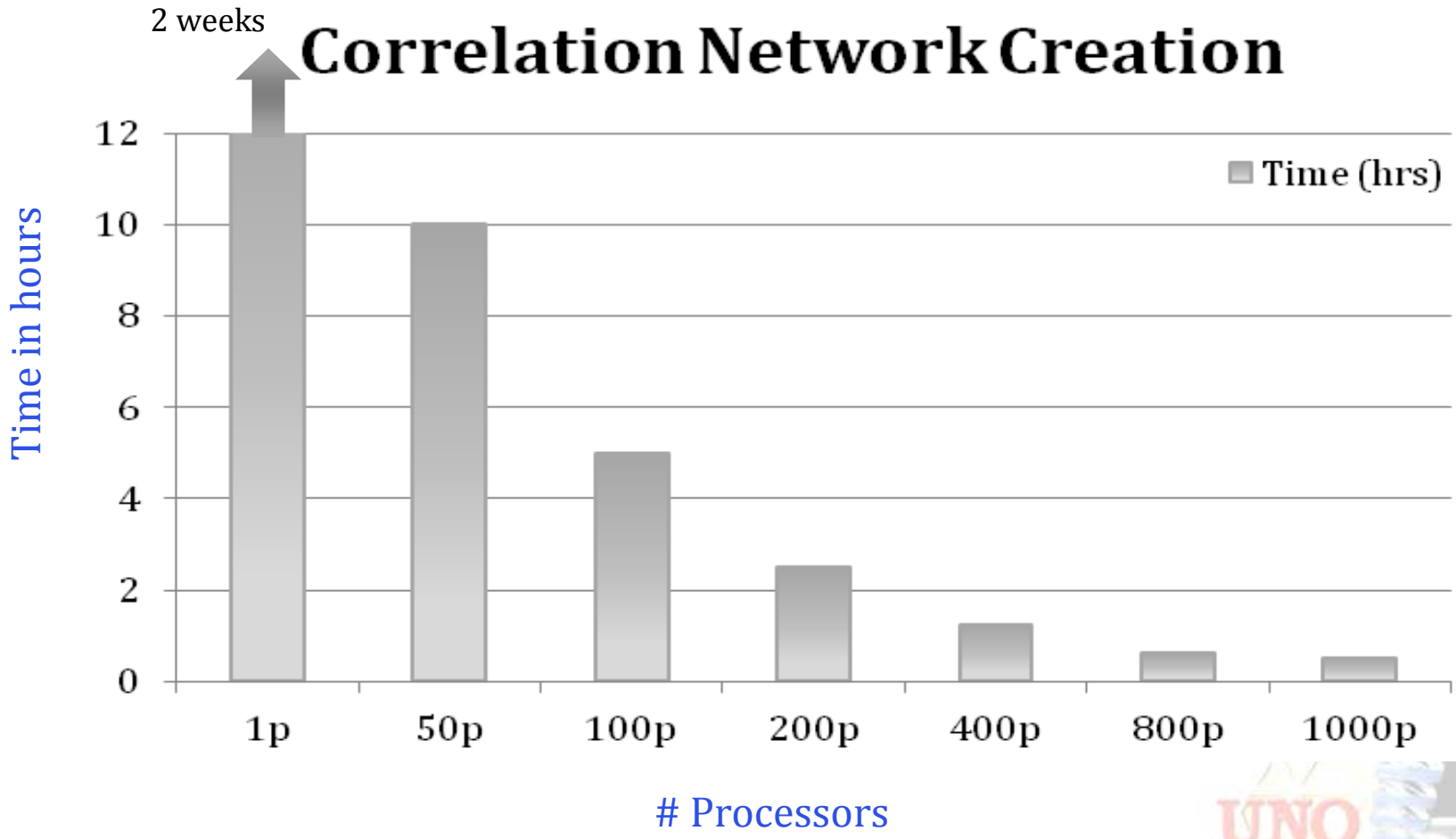
- Network creation: 2 weeks on PC
  - 10 hours in parallel, 50 nodes
  - 40,000 nodes = 800 million edges (pairwise)
  - 40,000 ! Potential relationships
  - Big data or big relationship domain

- Network analysis: Best in parallel
  - Only 3% of entire genome forms complexes

- Holland Computing Center: Firefly 1150 8-core cluster – from weeks to hours/minutes

UNO
BIOINFORMATICS

The Need for HPC



UNO BIOINFORMATICS

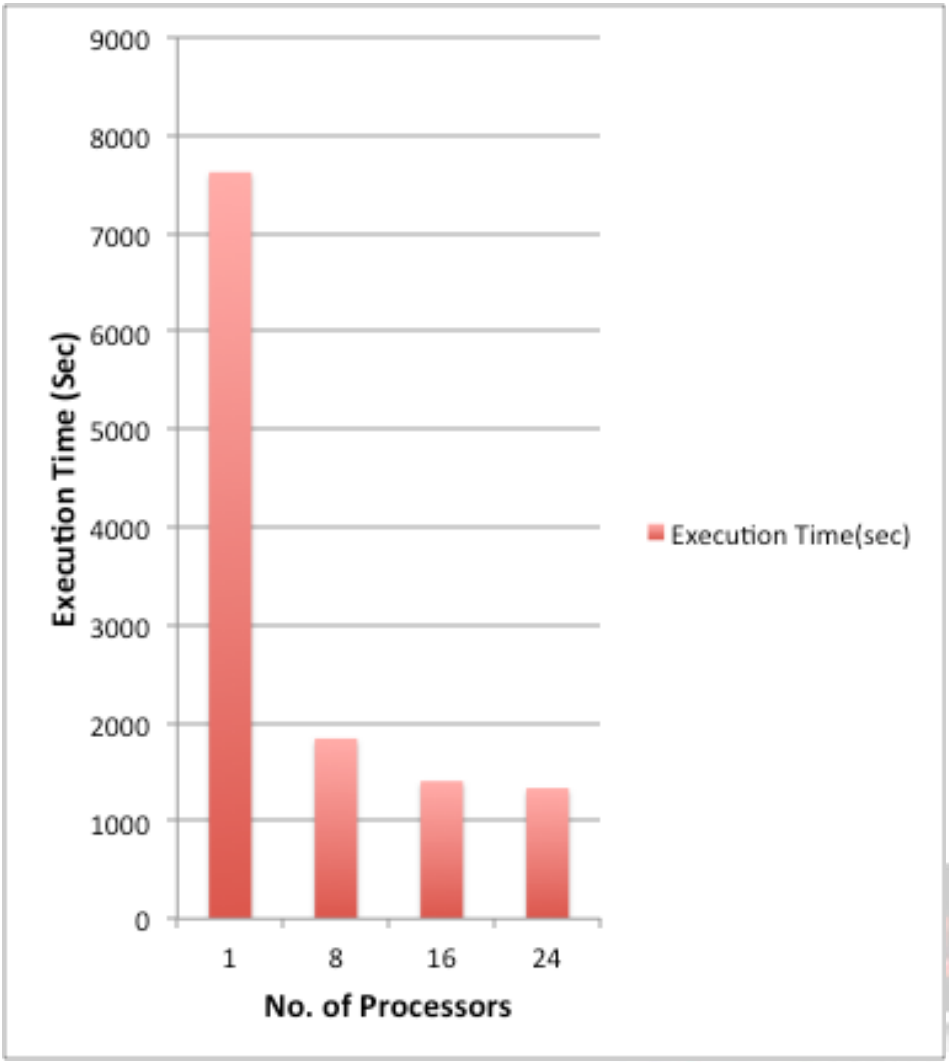
To Build or to Out-source?

- Utilize public resources or build a private infrastructure?
- What are the deciding factors?
  - Cost
  - Security/Privacy
  - Development ability
  - Customization
  - Nature of utilization
  - Systems administration
  - …

UNO
BIOINFORMATICS

Differentially Expressed Genes Analysis

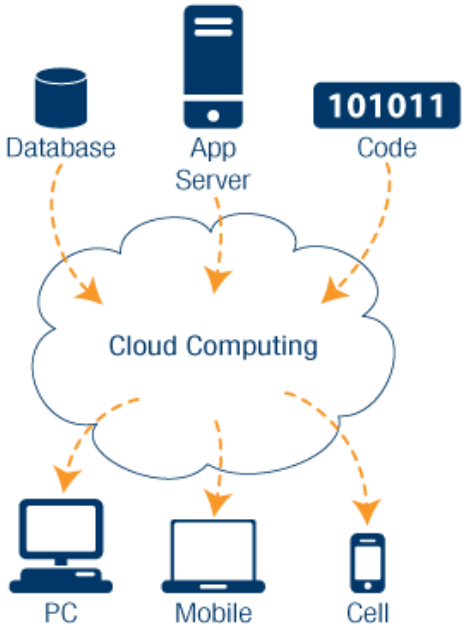
- Implement baySeq method to find differentially expressed genes
- Input: 19K genes, HIV Infected and Uninfected RNA seq counts
- Significantly faster when processors are added. After 16 processors, not much of improvement



BIOINFORMATICS

Working in the Cloud

- Cloud computing is Web-based processing and storage. Software and equipment are offered as a service over the Web.
  - Data and applications can be accessed from any location
  - Data and applications can easily be shared through a common platform
  - Clouds need not be public; companies can introduce private cloud computing solutions



UNO BIOINFORMATICS

Cost Reduction & Convenience



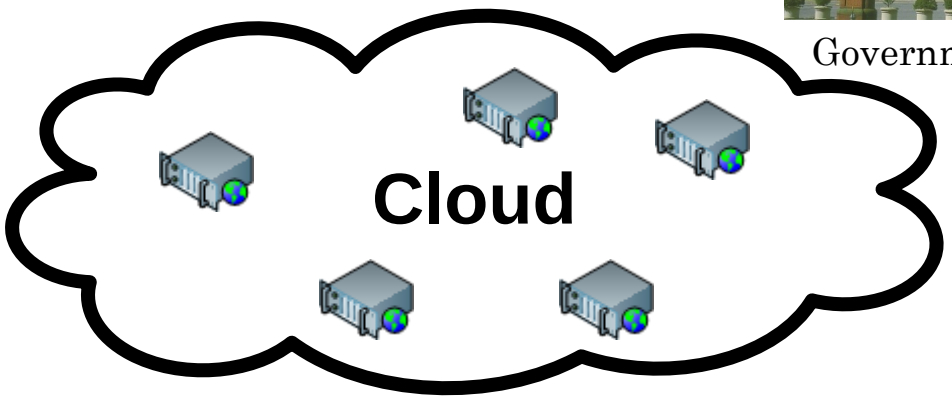
Small Business



Government Offices



Multinational Corporations



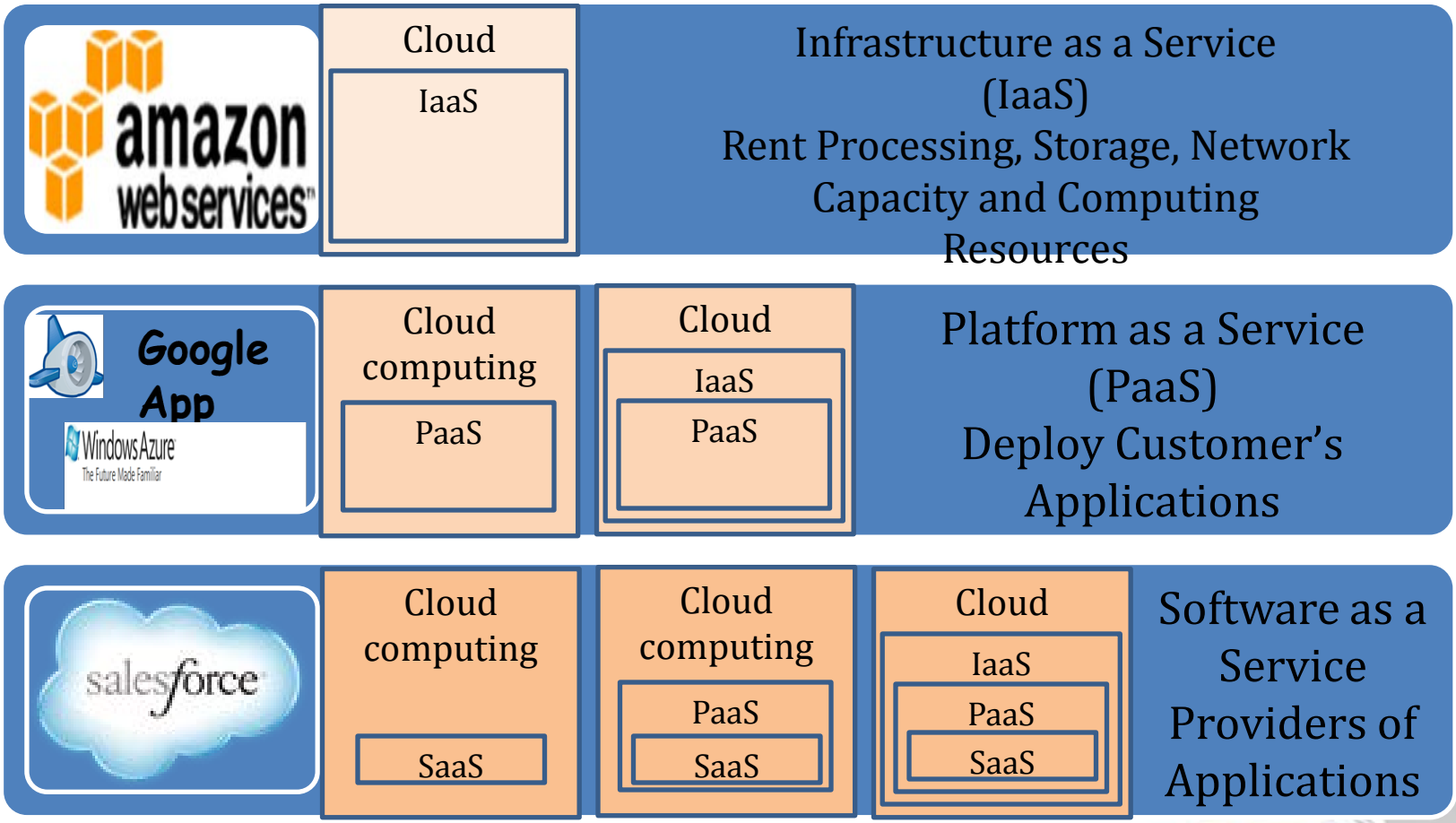


Homes

- Flexible availability of resources
- Opportunity for developers to easily push their applications
- Targeted advertising
- Easy Software Upgrades for customers
  - Example: Webmail

UNO BIOINFORMATICS

Cloud Computing - Models



UNO BIOINFORMATICS

HPC and Big Data in Biological Networks

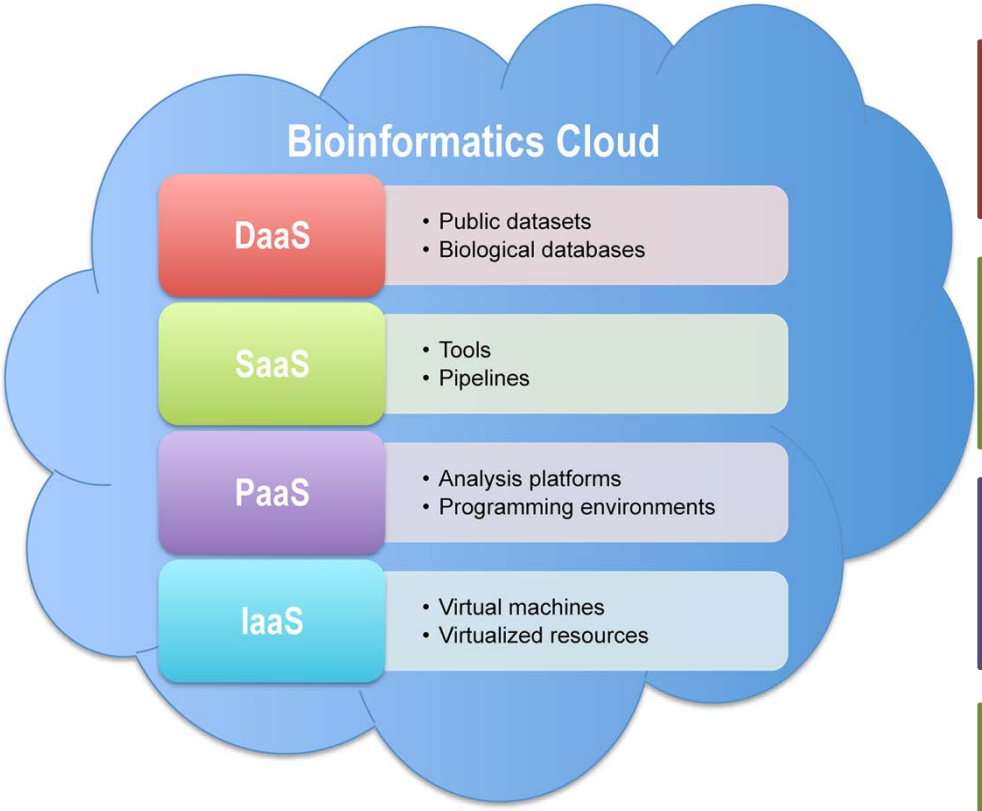
- Network creation: 2 weeks on PC
  - 10 hours in parallel, 50 nodes
  - 40,000 nodes = 800 million edges (pairwise)
  - 40,000 ! Potential relationships
  - Big data or big relationship domain

- Network analysis: Best in parallel
  - Only 3% of entire genome forms complexes

- Holland Computing Center: Firefly 1150 8-core cluster – from weeks to hours/minutes



Bioinformatics and Cloud Computing



DaaS (Data as a Service)
- Example: AWS public data sets include GenBank, Ensembl, 1000 Genomes, etc.

SaaS (Software as a Service)
- Sequence analysis tools (BGI easy genomics),
- Read mapping and SNP calling (Crossbow), etc.

PaaS (Platform as a Service)
- Galaxy Cloud (Workflow platform)
- Eoulsan (RNAseq pipeline for diff. analysis)

IaaS (Infrastructure as a Service)
- CloVR (Portable virtual machine),
- Cloud BioLinux

Dai et al. Bioinformatics clouds for big data manipulation, Biology Direct 2012

UNO BIOINFORMATICS

Cloud Resources in Bioinformatics

Resource	Description & availability
Data as a Service (DaaS):	
AWS Public Datasets	Cloud-based archives of GenBank, Ensembl, 1000 Genomes, Model Organism Encyclopedia of DNA Elements, Unigene, Influenza Virus, etc.; http://aws.amazon.com/publicdatasets
Software as a Service (SaaS):	
BGI Cloud (unpublished)	Cloud-based implementations of various genomic analysis applications; http://cloud.genomics.cn
CloudAligner [16]	Fast and full-featured MapReduce-based tool for sequence mapping; http://cloudaligner.sourceforge.net
CloudBLAST [19]	A cloud-based implementation of NCBI BLAST; http://ammatsun.acis.ufl.edu/amwiki/index.php/CloudBLAST_Project
CloudBurst [17]	Highly sensitive short read mapping with MapReduce; http://cloudburst-bio.sourceforge.net
Contrail (unpublished)	Cloud-based *de novo* assembly of large genomes; http://contrail-bio.sourceforge.net
Crossbow [18]	Read Mapping and SNP calling using cloud computing; http://bowtie-bio.sf.net/crossbow
EasyGenomics (unpublished)	Cloud-based NGS pipelines for whole genome resequencing, exome resequencing, RNA-Seq, small RNA and de novo assembly; http://www.easygenomics.org
eCEO [26]	Cloud-based identification of large-scale epistatic interactions in genome-wide association study (GWAS); http://www.comp.nus.edu.sg/~wangzk/eCEO.html
FX [20]	RNA-Seq analysis tool; http://fx.gmi.ac.kr
Gaea (unpublished)	Cloud-based genome re-sequencing assembly; http://bgiamericas.com/data-analysis/cloud-computing
Hecate (unpublished)	Cloud-based *de novo* assembly; http://bgiamericas.com/data-analysis/cloud-computing
Jnomics (unpublished)	Cloud-scale sequence analysis suite based on Apache Hadoop; http://sourceforge.net/apps/mediawiki/jnomics
Myrna [21]	Differential gene expression tool for RNA-Seq; http://bowtie-bio.sourceforge.net/myrna
PeakRanger [24]	Cloud-enabled peak caller for ChIP-seq data; http://www.modencode.org/software/ranger
RSD [23]	Reciprocal smallest distance algorithm for ortholog detection using Amazon's Elastic Computing Cloud; http://roundup.hms.harvard.edu
VAT [25]	Variant annotation tool to functionally annotate variants from multiple personal genomes at the transcript level; http://vat.gersteinlab.org
YunBe [22]	Pathway-based or gene set analysis of expression data; http://tinyurl.com/yunbedownload
Platform as a Service (PaaS):	
Eoulsan [27]	Cloud-based platform for high throughput sequencing analyses; http://transcriptome.ens.fr/eoulsan
Galaxy Cloud [28,29]	Cloud-scale Galaxy for large-scale data analysis; http://galaxy.psu.edu
Infrastructure as a Service (IaaS):	
Cloud BioLinux [30]	A publicly accessible virtual machine for high performance bioinformatics computing using cloud platforms; http://cloudbiolinux.org
CloVR [31]	A portable virtual machine for automated sequence analysis using cloud computing; http://clovr.org



Private Clouds

- Core facilities need to acquire private infrastructure-level virtual cloud technology. Best vendor for such technology is VMware. The Bioinformatics Core facility at UNO uses *VMware vSphere Enterprise.*
- Public Clouds like Amazon EC2, RackSpace cannot be used in all cases due to various restrictions put forth by regulations (e.g. HIPAA data locality requirement). Such public clouds could only be used as a scalable platform for already anonymized data.
- Private Virtual Cloud is on-premise solution allowing all the benefits of virtualization technology both from an administrative and end-user perspective.

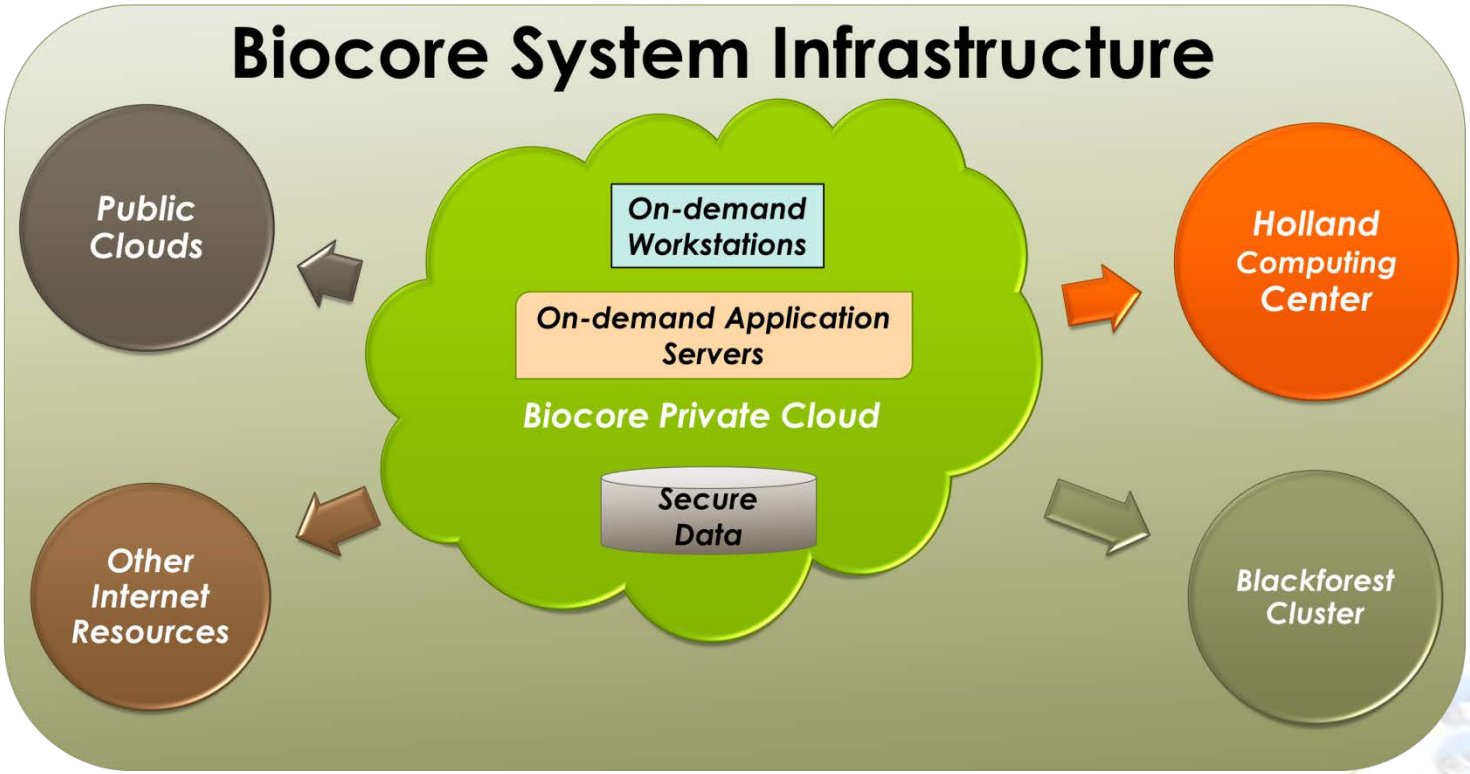
UNO
BIOINFORMATICS

Security and Private Cloud

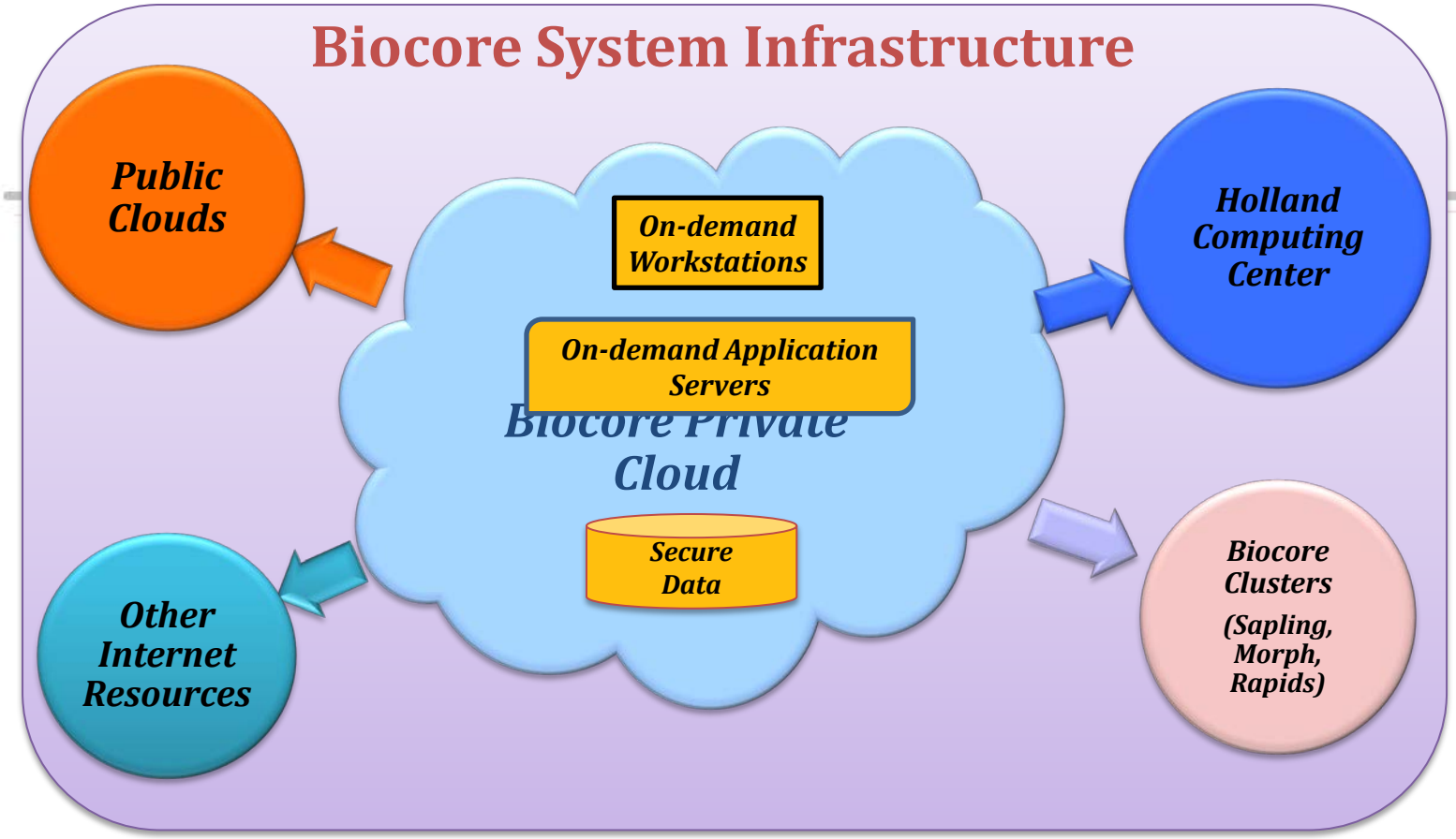
- Core facilities need to acquire private infrastructure-level virtual cloud technology. Best vendor for such technology is VMware. The Bioinformatics Core facility at UNO uses *VMware vSphere Enterprise.*
- Public Clouds like Amazon EC2, RackSpace cannot be used in all cases due to various restrictions put forth by regulations (e.g. HIPAA data locality requirement). Such public clouds could only be used as a scalable platform for already anonymized data.
- Private Virtual Cloud is on-premise solution allowing all the benefits of virtualization technology both from an administrative and end-user perspective.



Proposed Model



UNO BIOINFORMATICS



Infrastructure As A Service: Virtual workstations and servers provided to various core facilities from the Biocore private cloud

Data As A Service: Local mirrors of various public databases. (*GenBank*, *RefSeq* , *EMBL* and more...).Shared storage for various core facilities in the form of network drives

HIPAA compliance: (e.g. the **R**heumatoid **A**rthritis **I**nfrastructure **N**etwork website)

UNO BIOINFORMATICS

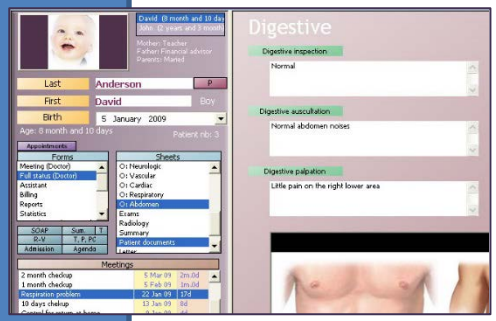
Acknowledgments

- UNO Bioinformatics Research Group
  Kiran Bastola
  Sanjukta Bhoomwick
  Kate Dempsey
  Jasjit Kaur
  Ramez Mena
  Sachin Pawaskar
  Oliver Bonham-Carter
  Ishwor Thapa
  Dhawal Verma
  Julia Warnke
- Former Members of the Group
  Alexander Churbanov
  Xutao Deng
  Huiming Geng
  Xiaolu Huang
  Daniel Quest
- Biomedical Researchers
  Steve Bonasera
  Richard Hallworth
  Steve Hinrichs
  Howard Fox
  Howard Gendelman
- Funding Sources
  NIH INBRE
  NIH NIA
  NSF EPSCoR
  NSF STEP
  Nebraska Research Initiative

Bioinformatics & the Future of Personalized Medicine

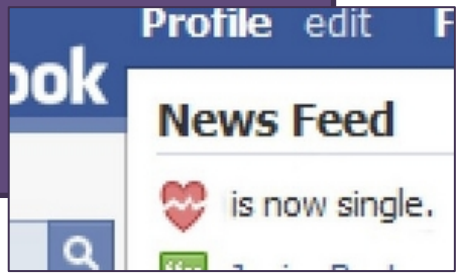
Health Data

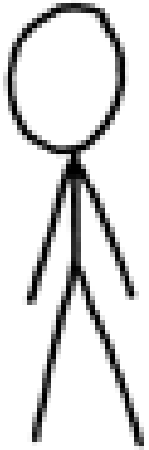
- Patient history
- Allergies
- Prescription history
- Family history
- Surgeries



Social Data?

- Relationship/friendship data
- Location
- Smog
- Light Exposure





Genetic Data

- Personalized genome
- Genome *over time?*
- Susceptibilities
- Preventative therapeutics



Wellness Data

- Sleep habits
- Eating habits
- Daily activity
- Stress levels



Size and Noise

- Network made from average gene expression experiment will have:
  - 40,000 nodes
  - 800 million edges
- Only 3% of genes in entire genome work together to form complexes
- Even with parallel computing resources, unfiltered networks are too noisy for biological discovery

UNO
BIOINFORMATICS

Network Filters

Design a network filter and obtain a sub-network of the original network such that:

- It maintains the important stuff – signal
- Remove unimportant stuff – noise
- Maintain network elements of biological relevance
- Uncover new ones

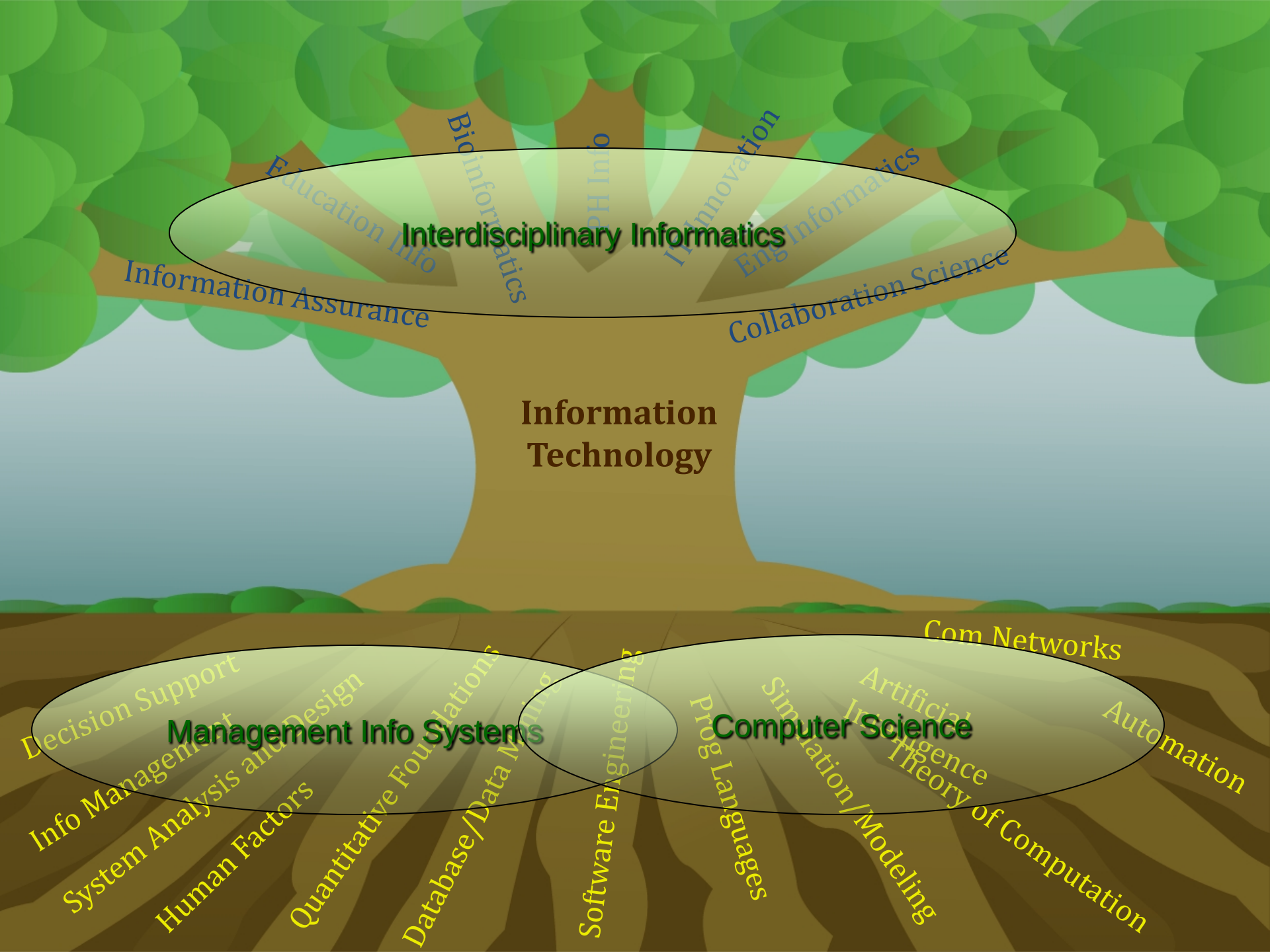
UNO
BIOINFORMATICS

Chordal Graph Sampling

Goal: Develop a parallel network sampling technique that *filters noise*, while *preserving the important characteristics of the network*.

- ✓Maximal Chordal Subgraph
  - ➢Spanning subgraph of the network w
    - ➢No cycles of length larger than three
- ✓Properties of Chordal Graph
  - ➢Preserves most cliques and highly connected regions of the network
  - ➢Most NP hard problems can be solved in polynomial time
  - ➢Complexity of finding maximal chordal subgraphs: **O(|E|*max_deg)**




Interdisciplinary Informatics
Education Info
Bioinformatics
IT Innovation
Eng Informatics
Information Assurance
Collaboration Science
Information Technology
Management Info Systems
Decision Support
Info Management
System Analysis and Design
Human Factors
Quantitative Foundations
Database/Data Mining
Software Engineering
Computer Science
Prog Languages
Simulation/Modeling
Artificial Intelligence
Theory of Computation
Com Networks
Automation

Let Us Meet Again

We welcome you all to our future conferences of OMICS Group International

Please Visit:

www.omicsgroup.com

www.conferenceseries.com

www.proteomicsconference.com

UNO BIOINFORMATICS