University of Nebraska Omaha

Innovative Population-based Approaches for Analyzing Mobility Data in Continuous Health Applications

NexTech 2019

Hesham H. Ali
UNO Bioinformatics Core Facility
College of Information Science and Technology
September 2019

Population Analysis in Biomedical Informatics

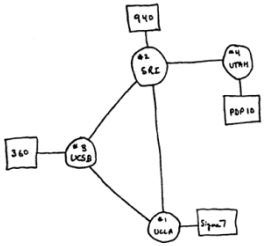
- Background and general introduction
- The Health Informatics Angle – connecting mobility and health
- The Bioinformatics Angle – Systems Biology and Network Models
- The Computing Angle – How to implement the proposed models

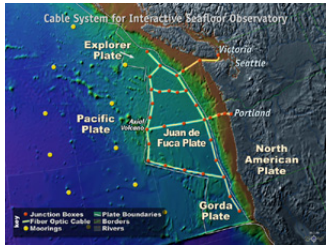
Background and General Introduction

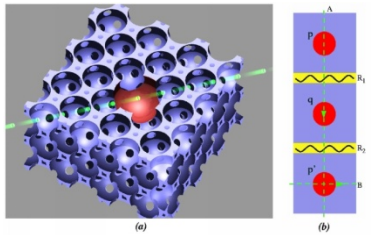
IT and Future Opportunities

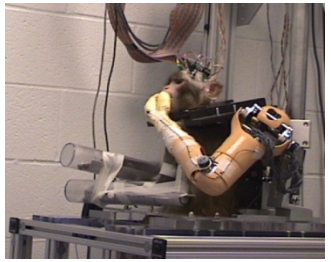
UNO BIOINFORMATICS

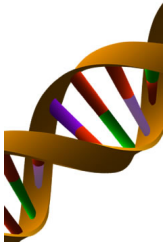
- Creating the future of networking
- Driving advances in all fields of science and engineering
- Revolutionizing transportation
- Personalized education
- The smart grid
- Predictive, preventive, personal medicine
- Quantum computing
- Empowerment for the developing world
- Personalized health monitoring => quality of life
- Harnessing parallelism
- Neurobotics
- Synthetic biology

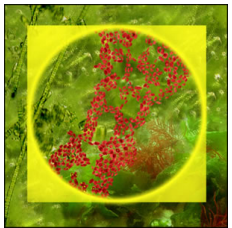












Biomedical Informatics: The new major revolution in Sciences!

- Each generation, a scientific discipline emerges with a bang and promises to change the way we do things – a game changer.
- The last major new discipline was Computer Science over 50 years ago.
- Is it Biomedical Informatics (BMI) for this generation?
- The connection to Human Health add another layer of significance to BMI

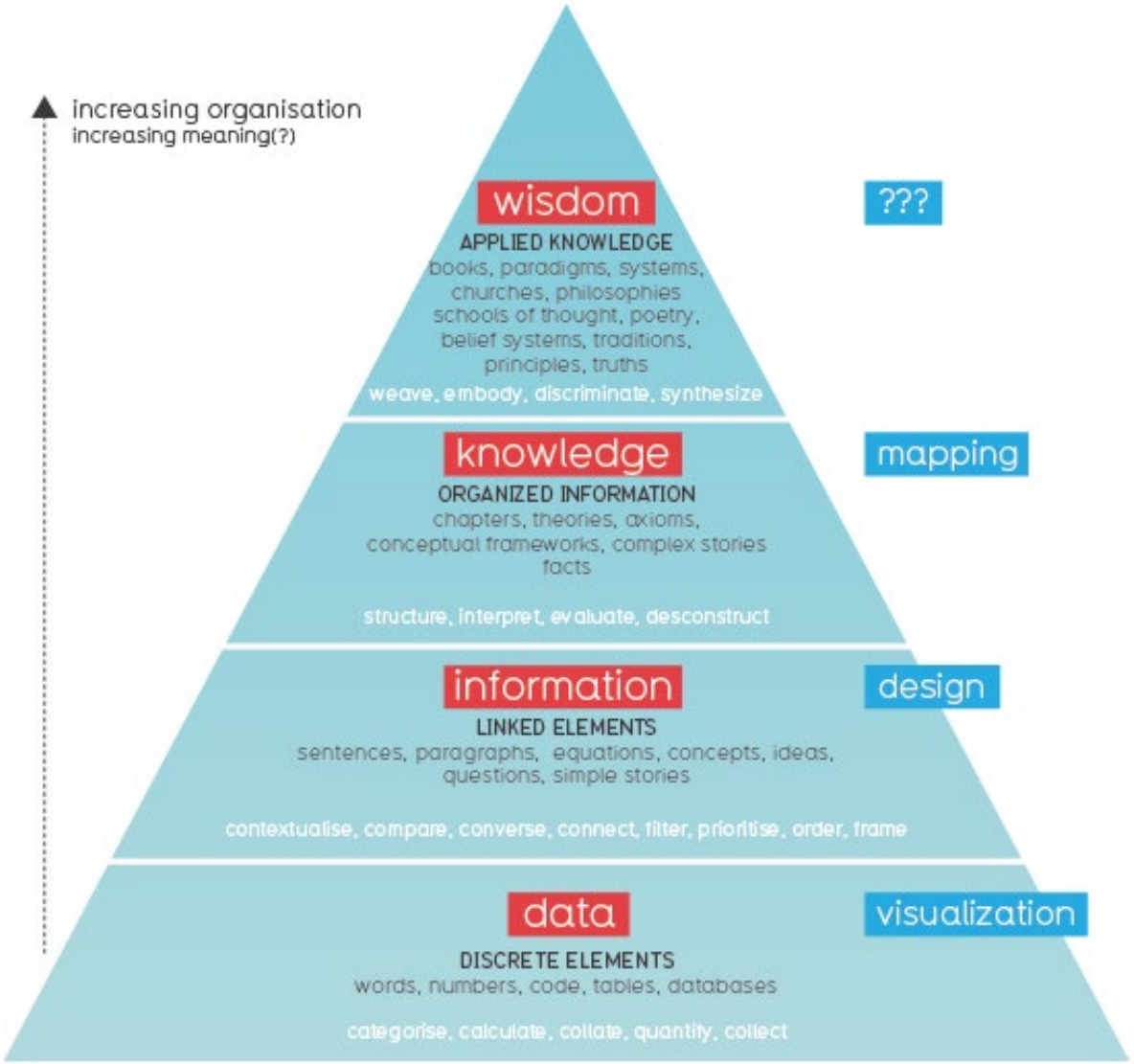
Revolution in Healthcare and Biomedical Research

- Revolutionize Biomedical Research and change healthcare models
- So much relevant data is currently available:
  - Remove the guessing aspect in conducting scientific research and practicing medicine
  - Proactive treatment and personalized medicine
- The availability of data shifted biomedical sciences sciences from pure experimental disciplines to hybrid knowledge-experimental based disciplines
- Incorporating Computational Sciences and Biosciences remains challenging - Interdisciplinary Research? Translational Research? Big Data Analytics?

It's all about the Data!

- How it all began:
  - Advances in medical instruments and computational technologies led to new new research directions
  - Massive accumulation of Biomedical data led to investigating new potential discoveries
  - The availability of enormous various types of public/private Biomedical data
  - How to take advantage of the available data
- Bioinformatics - Health Informatics - Biomedical Imaging - Public Health Informatics Biomedical Devices
- A new direction is now possible

Data-Information-Knowledge-Wisdom



http://www.ritholtz.com/blog/wp-content/uploads/2010/11/data_info_knowledge_wisdom.png

Smart Data Data-Driven Decisions

- Data: Physical entities at lowest abstraction level; contain little/no meaning – Measured data
- Information: Derived from data via interpretation – Processed data
- Knowledge: Obtained by inductive reasoning, typically through automated analysis and iterative collaboration – data + relationships
- Decision Support

Biomedical Informatics in 2019

- Bioinformatics/BMI is well recognized by researchers and practitioners
- Many believe that Bioinformatics or BMI to be that special discipline for this generation
- Back at mid nineties, one would have expected Bioinformatics to be further along after over 20 years.

Current Barriers in BMI

- On the biomedical side:
  - Too much focus on data collection
  - Competition to own the latest technology
  - Excitement associated with New technologies – which leads to more raw data
  - The black box syndrome
- On the computational side:
  - Certain level of casualness remain a major concern – just another application domain
  - Inconsistent results – lack of robustness and reproducibility
  - Heuristics and thresholds
  - Lack of Biomedical-rich integration

Data Generation vs. Analysis/Integration

- New technologies lead to new data:
  - Competition to have the latest technology
  - Focus on storage needs to store yet more data
- Biomedical community needs to move from a total focus on data generation to a blended focus of measured data generation and data analysis/interpretation/visualization
- How do we leverage data? Integratable? Scalable?
- From Data to Information to Knowledge to Advanced Decision Making

Biomedical Informatics and Big Data

- All the features of Big Data are represented:
  - Volume: New levels of massive data
  - Variety: Only one type of data is not enough
  - Veracity: Not always fully compete or fully trusted
  - Velocity: Data is collected continuously

- Multiple levels of Big Data analysis:
  - Populations
  - Individuals
  - Many granularities in between

Network Modeling and Population Analysis

- It is difficult to provide useful analysis or assessment elements in isolation
- Almost all analysis-related studies are conducted by comparing elements to its population or a group of similar features
- We approach big data analytics by building networks (graphs) of elements under study using different types of inter-relationships among the elements – such as correlations
- We then use graph theoretic properties of constructed networks to mine useful knowledge associated with big data

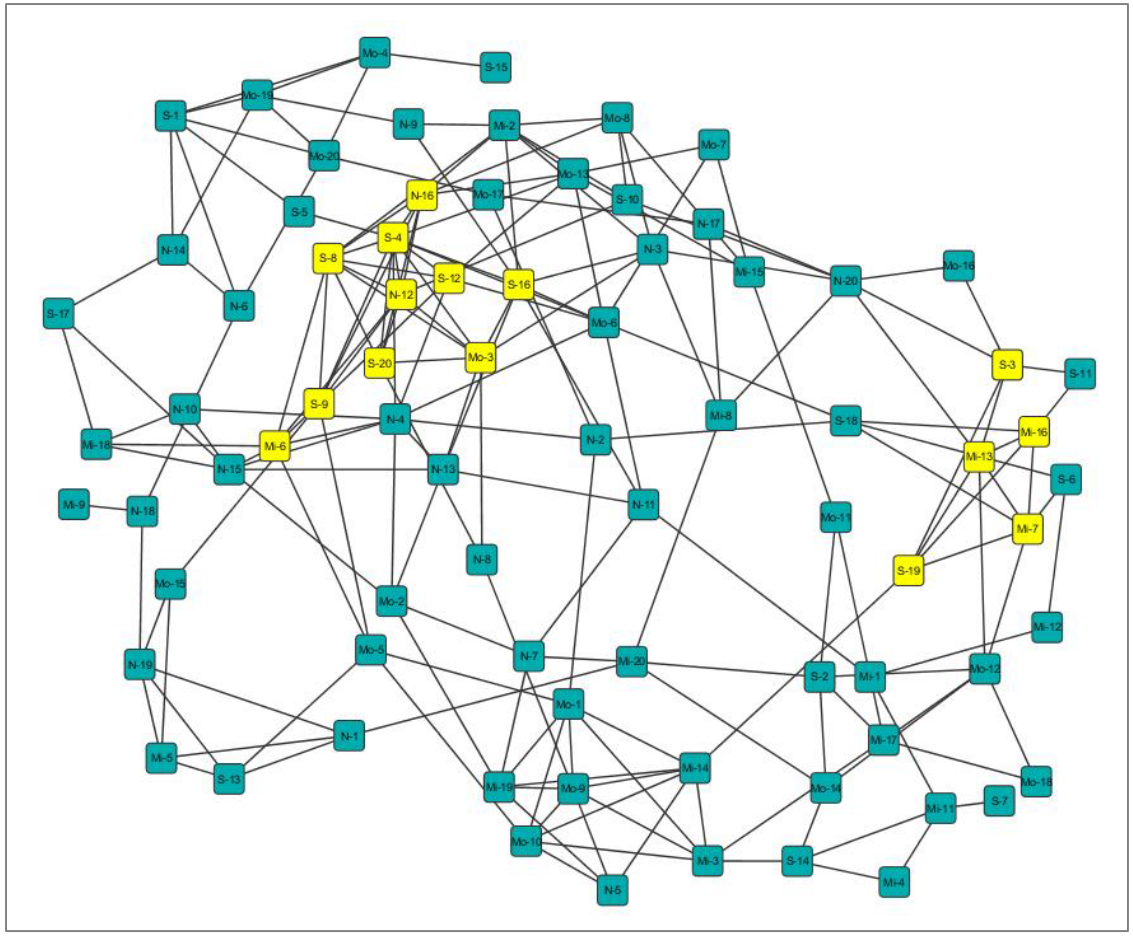
Examples of Population Analysis

- Biological Networks in Bioinformatics
  - Correlations among genes or interactions among proteins
  - Analysis of microbiomes in soil, water, human guts
- Correlation between mobility and health level
  - Monitoring mobility levels
  - Aging of cells and aging of systems
- Similarity networks and population analysis to study safety issues in bridges
  - Identify bridges that are not safe
  - Propose different maintenance schedules of bridges based on their sufficiency rating
- Analysis of financial markets using behavior networks
  - Analysis of stocks
  - Analysis of financial sectors

Population Analysis

UNO BIOINFORMATICS

- Correlation graph using mobility parameters



Data Analysis: Systems

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

- Discovery
  - Examine changes in systems
    - Control Group vs. Patient Group
    - Young vs. old
    - Stage x versus Stage y in disease progression

The Health Informatics Angle Connecting Mobility and Health

Wireless Networks in Aging

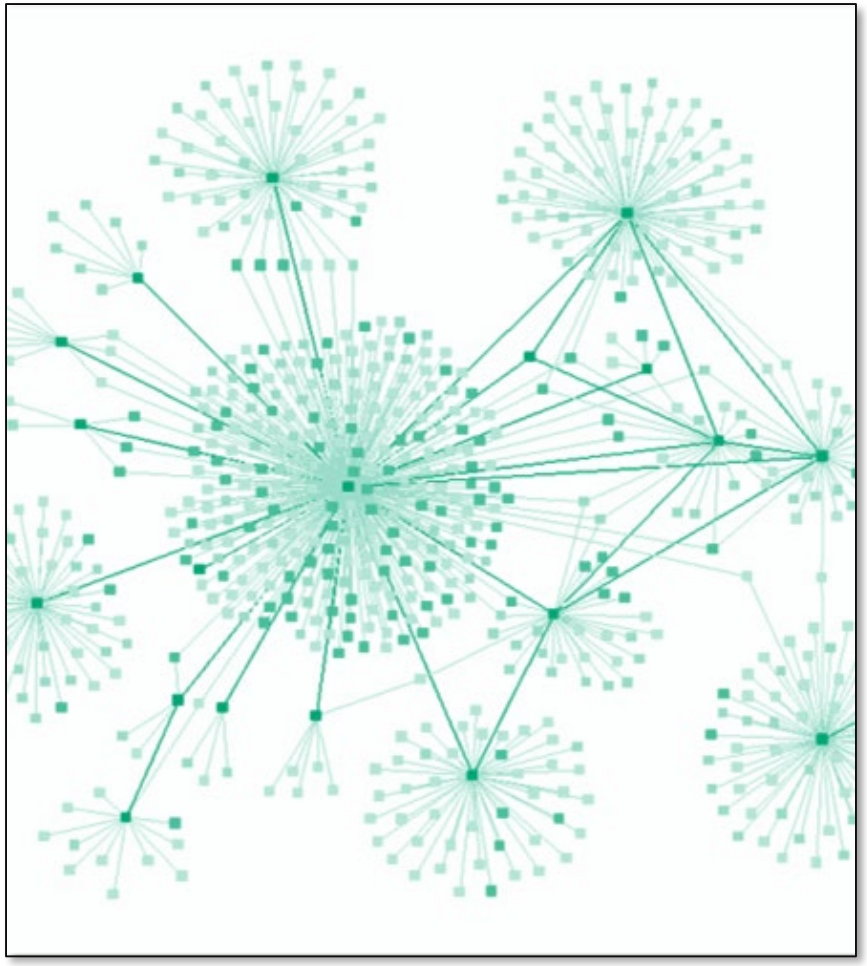
- Correlation between mobility and health level
- Monitoring mobility levels
- Aging of cells and aging of systems
- Collaboration between Bioinformatics group, Wireless Networks group and Decision Support Systems group

BMI Networks

- A BMI network represents elements and their interactions

- Nodes → elements
- Edges → relationships

- Can represent multiple types of elements and relationships



Correlation and Co-occurance 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

Power of Correlation Analysis

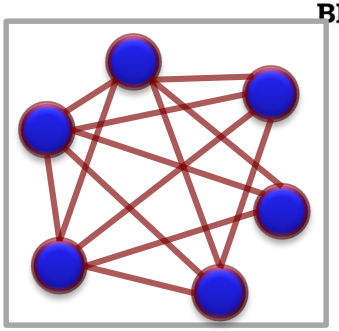
UNO BIOINFORMATICS

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

Local Structures

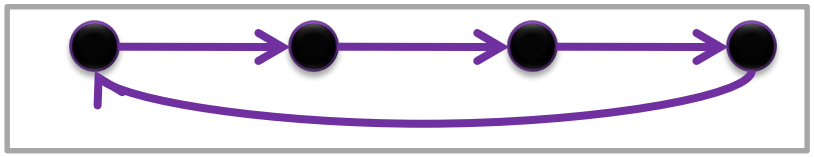
- **Cliques**

  Protein complexes, regulatory modules



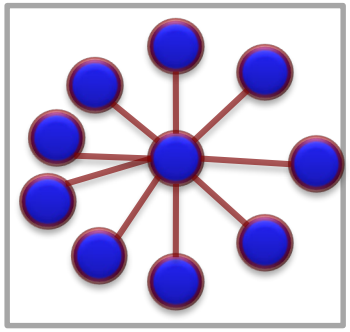
- **Pathways**

  Signaling cascades

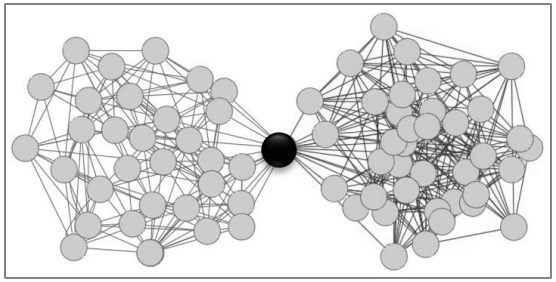


- **Hubs**

  Regulators, TFs, active proteins



- **Articulation points – Gateways**



Hypothesis

Correlation networks are an excellent tool for mining relationship rich knowledge from high-throughput data

Using systems biology approach, CN can help identify:

- *Critical Genes* that are essential for survival
- *Subsets of genes* that are responsible for biological functions

Measures of centrality to identify key elements:

Proves existence of structure/function relationship in correlation networks

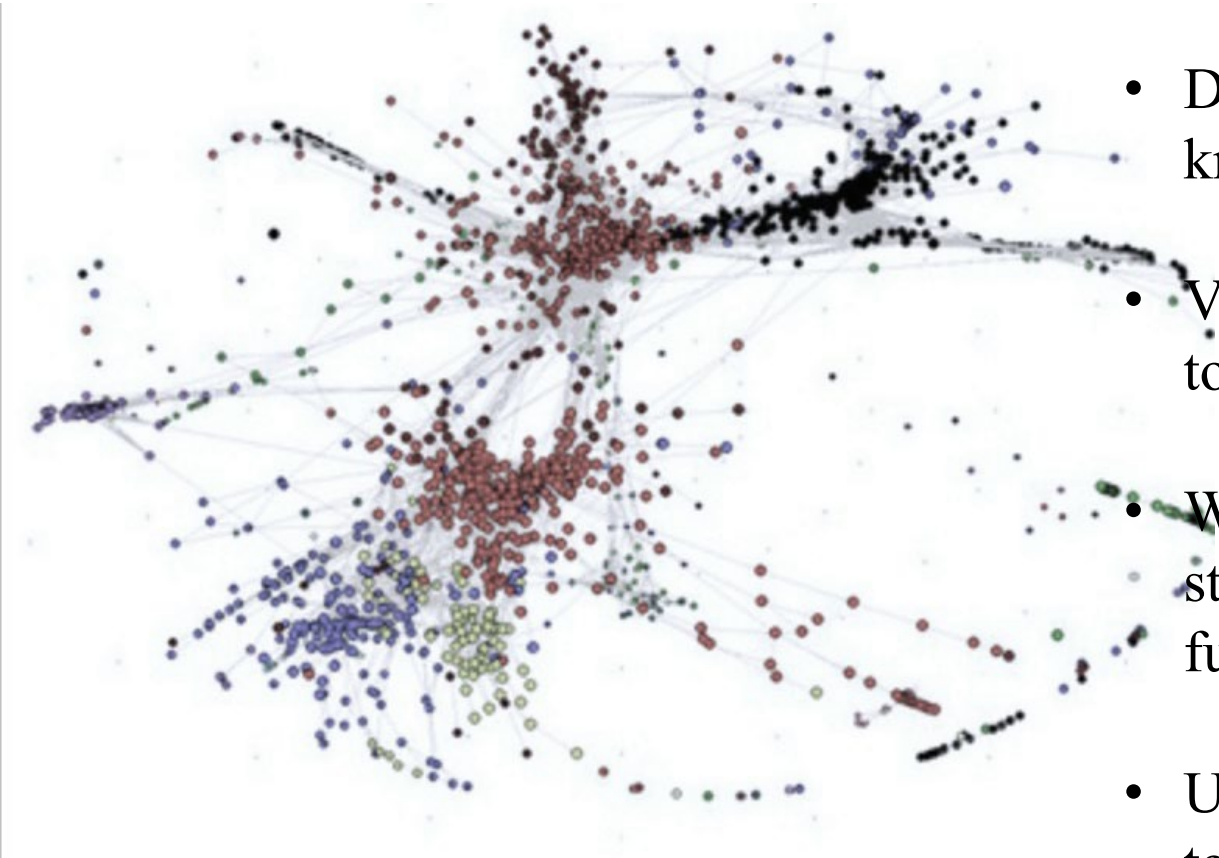
Health Monitoring

- Availability of many large useful devices – focus on collecting relevant data
- Availability of numerous helpful software packages
- Lack of data integration and trendiness of the discipline
- Fragmented efforts by computational scientists and biomedical scientists
- Lack of translational work – from the research domain to health care applications
- Increasing interest among researchers, industry and educators

How to collect mobility data?

- Laboratory setting
- Real-world setting
- Self-reported data collection method
- Using monitoring devices, sensors and accelerometers or using Internet of Things (IoT) devices

Correlation Networks and Population Analysis



- Able to handle ‘big’ data
- Draws from centuries of knowledge in graph theory
- Visually appealing and easy to understand
- When built correctly, structures can be tied to function
- Used in social, biological, technical applications

Goals of the Project

- Mobility Profile
  - Patient wearing a 3D-accelerometer will be monitored 24/7.
  - A complete mobility profile will be available for patients and care providers.
- Health hazards Prediction using Mobility Profiles
  - The system will identify anomalous movement and patterns that usually result in a fall or injury,
  - We would be able to take preemptive measures when such a pattern is detected, in order to reduce the occurrence of falls and prevent fall-related injuries.
  - We will develop an index that enables health care providers to determine how likely people are to fall.

Earlier Mobility Models

Experimental Studies

- Simulation Study
  - Mobility of nurses in a hospital – 8 hour shifts versus 12 hour shifts
  - Monitoring mobility pattern changes at different times during the shift
- Experimental Study
  - Mobility of mice in a cage
  - Identifying/classification of various groups based on mobility characteristics

Nursing Study

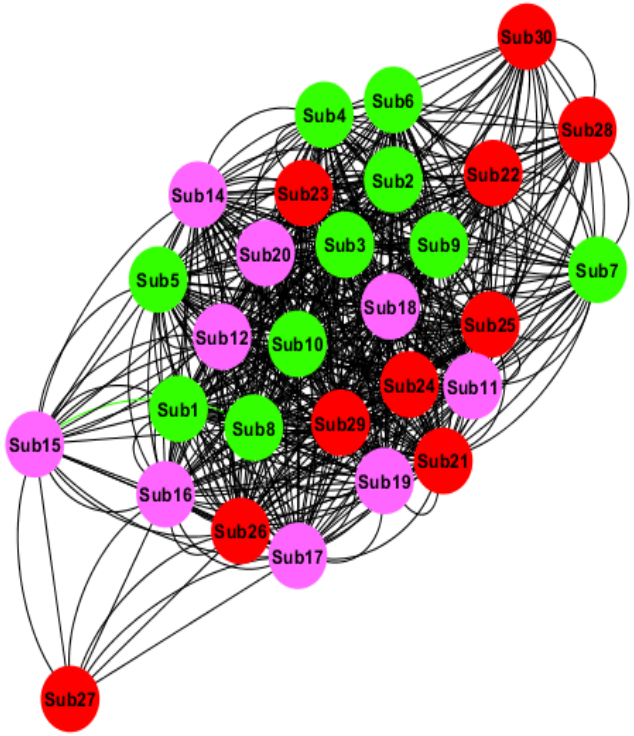
- Sample Generation Setting
  - Weighted activity level value
  - Each group has different mobility decline rate per hour
  - Group1 - 10%/hour, group2 - 20%/hour and group 3 – 30%/hour (shown in different colors)

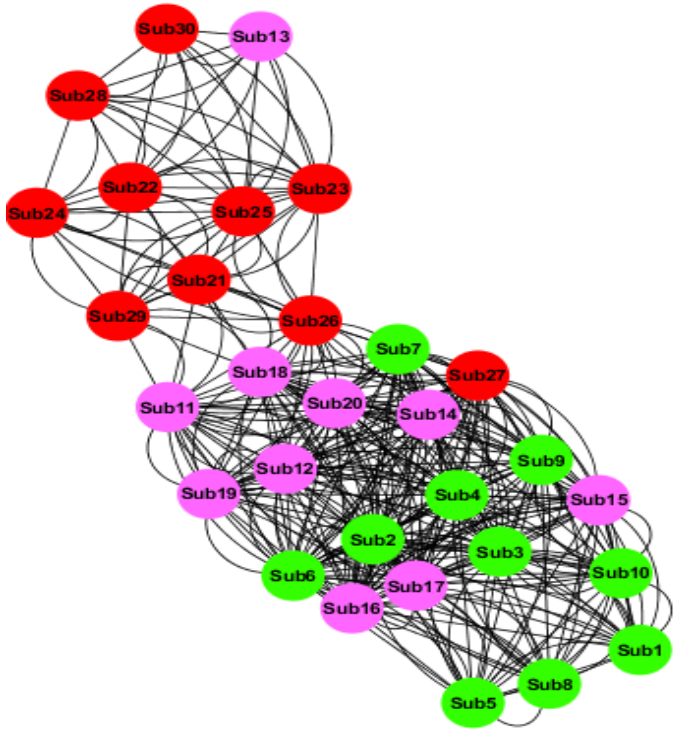
	Sub1	-	-	Sub30
Work start	553.78	-	-	384.85
2nd hours	498.40	-	-	269.40
4th hours	448.56	-	-	188.58
6th hours	403.71	-	-	132.01
Work end	363.34	-	-	92.40

Scenario Description

- Analyzing clusters from correlation networks
- Networks are constructed for every mobility samples captured from nurses at 4 different times as the day progresses.
- Magnitude based analysis applied
- Different levels of mobility decline in nurses identified from networks.

Networks Formed



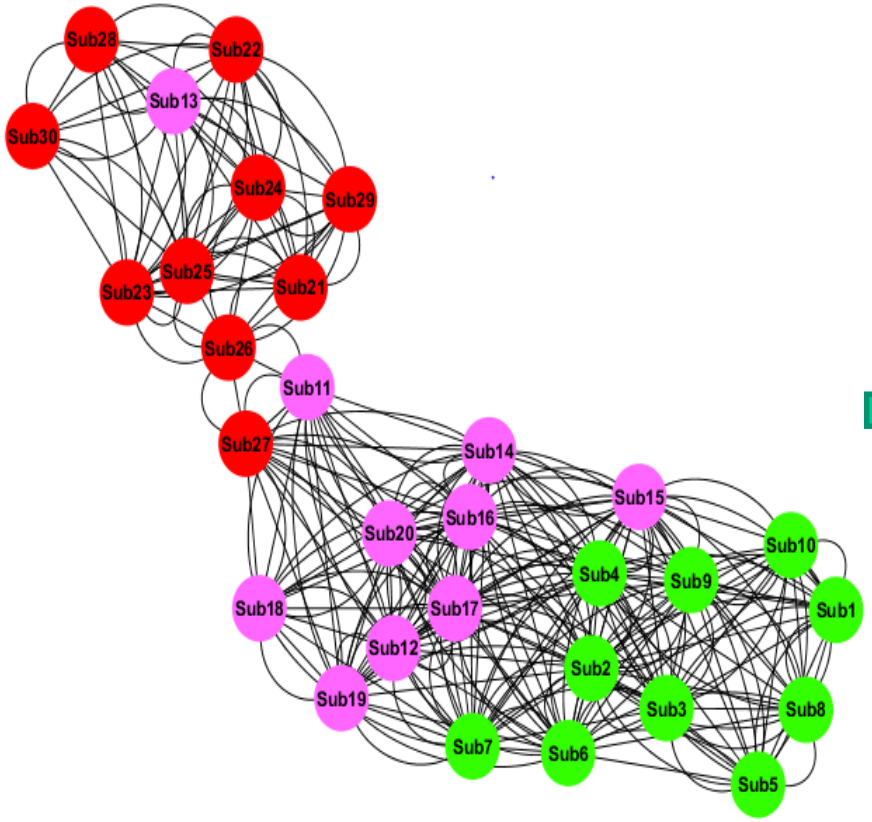


Sample 1

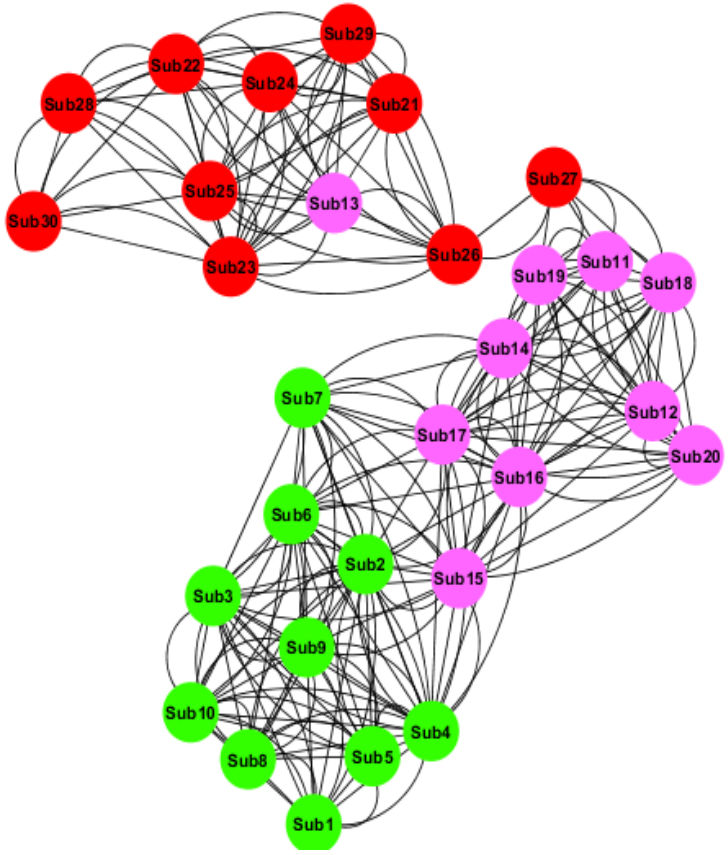
Sample 2

Green nodes – Group 1
Pink nodes – Group 2
Red nodes – Group 3

Networks Formed



Sample 3



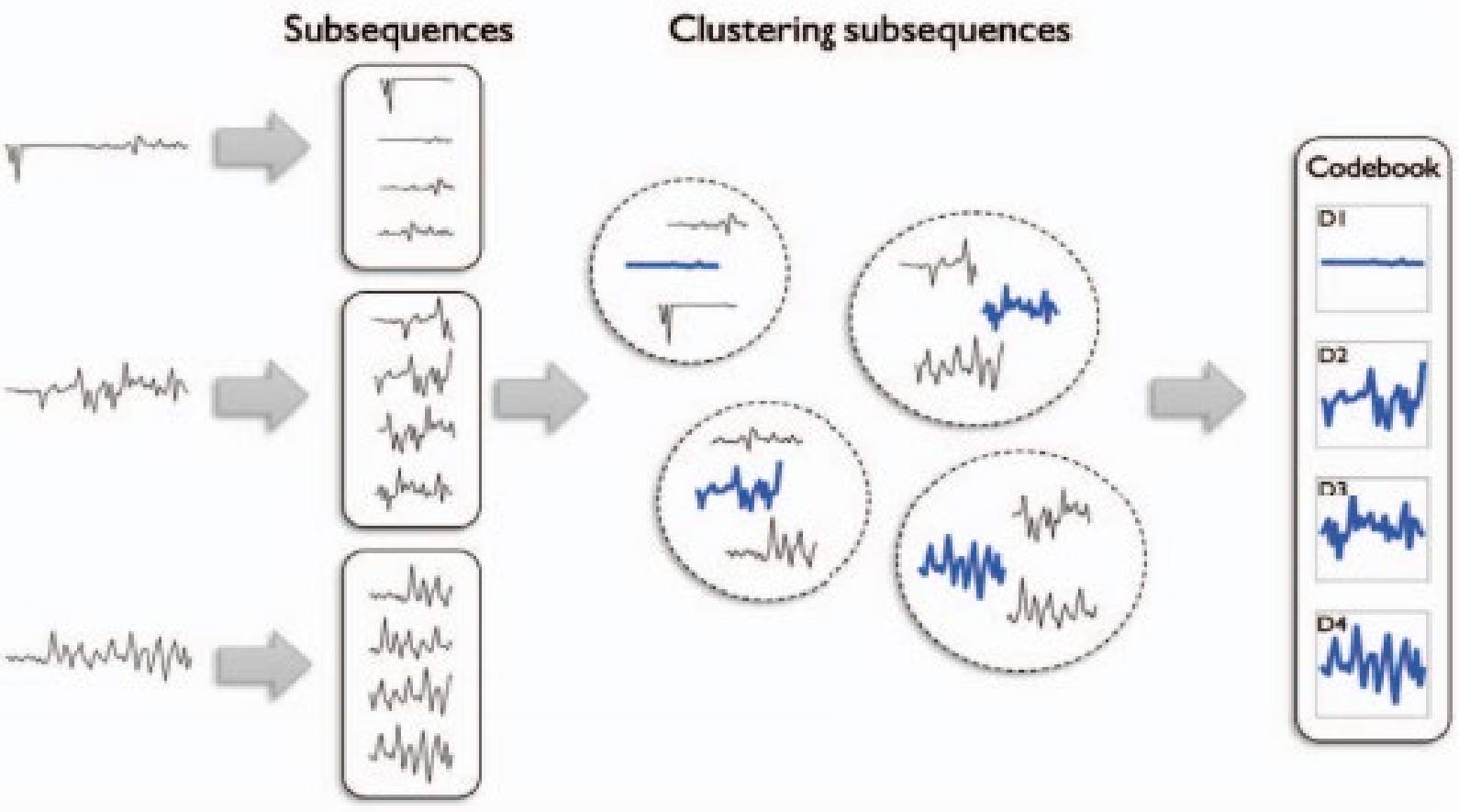
Sample 4

Green nodes – Group 1
Pink nodes – Group 2
Red nodes – Group 3

Results

- In the beginning of working days, clusters have mixed groups
- As the time progressed in the day, the same colored nodes (nurses) formed a separate clusters according to their raw mobility meausres.
- This network shows different clusters each of low (red), medium (pink) and high (green) mobilities.
- Application to predict medical hazards
- Practical to free-living environment

Feature Engineering- Movement Words Coding Scheme



Dataset: Participants and Protocol (Ankle Data)

- Protocol:
  - 40 Meter Walking (10-meter walkway back and forth)
  - Sampling frequency:100
  - Mild PD

	Control	PD	Geriatrics
Number of subjects	10	10	10
Gender (M/F)	5:5	6:4	5:5
Age	64 ± 8.4	63.8 ± 9.3	81 ± 4.1
UPDRS III		12.7 ± 6.0	
H & Y		1.7 ± 0.9	

Dataset: Participants and Protocol (Ankle Data)

- Protocol:
  - 4 minute Walking (around the hospital)
  - Sampling frequency:100
  - Moderate PD

	Control	PD	Geriatrics
Number of subjects	5	5	5
Gender (M/F)	3:2	3:2	2:3
Age	64 ± 10	72 ± 6.3	81 ± 5.9
UPDRS III		20.8 ± 6.1	
H & Y		2.6 ± 0.5	

Dataset: Participants and Protocol (Wrist Data)
First Phase

- Three phases of data collection (6-months period between each two phases)- One week of data per individual-
- Sampling frequency:100
- Mild, moderate, and sever PD (overall mild PD)

	Healthy young	Healthy elderlies	PD
Number of subjects	24	32	25
Gender (M/F)	14/10	10/22	20/5
Age	24 ± 3.6	64.2 ± 7	71 ± 6.2
UPDRS III		------	
H & Y			1.73 ± 0.83

Modeling: Machine Learning

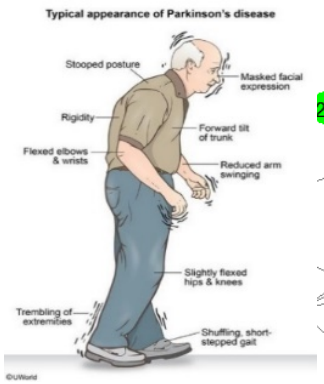
- **Standard Features:**
  - All features (32)
  - First reduced set of features (22)
    - Using Information Gain and Ranker methods
  - Second reduced set of features (8)
    - Using Pearson Correlation coefficient and ANOVA table
  - Third reduced set of features (7)
    - feature sets with one feature less than the optimal number of features
- **Document-of-Words Features:**
  - 10 Features for wrist data and 4 features for ankle data
- **Various Machine Learning Techniques:**
  - SVM, Random Forest, Naïve Bayes, AdaBoost, and bagging
- **Validation:**
  - K-Fold Cross validation
- **Accuracy measures:**
  - F-measure, Precision, Recall

Population Analysis and Similarity Network Models

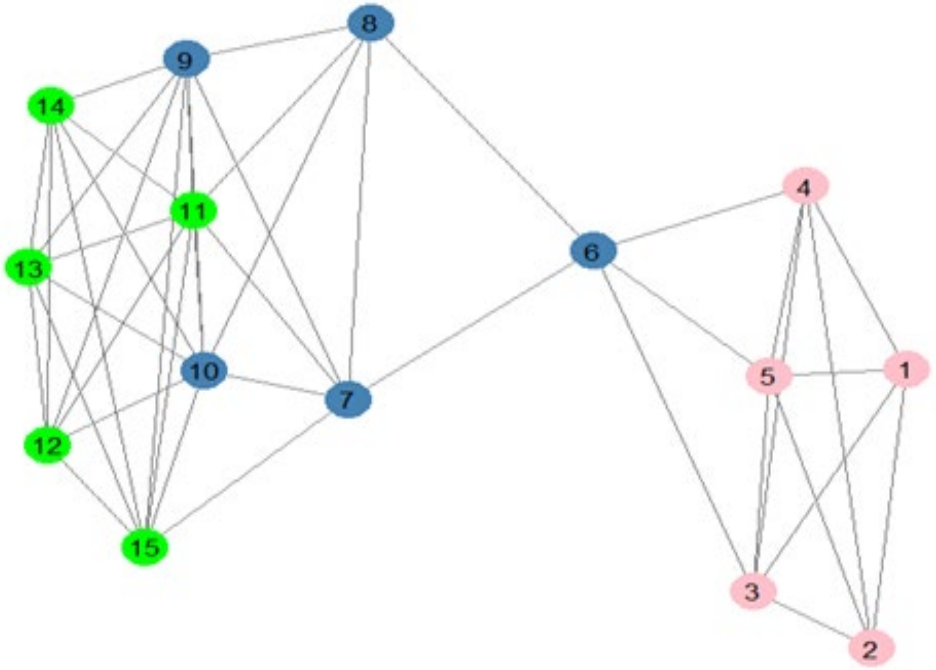
- Pairwise Correlation
  - A pairwise Pearson correlation analysis between subjects, using gait parameters
  - Threshold → 90%
  - Significance → 0.05

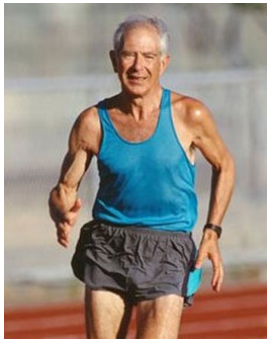
- Creating Network Model
  - Vertices represent subjects
  - If two subjects are highly correlated, there is an edge

Similarity Network Model – Wrist Data-Word Features

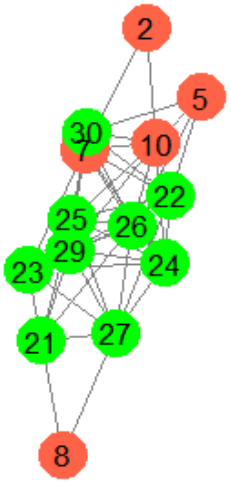


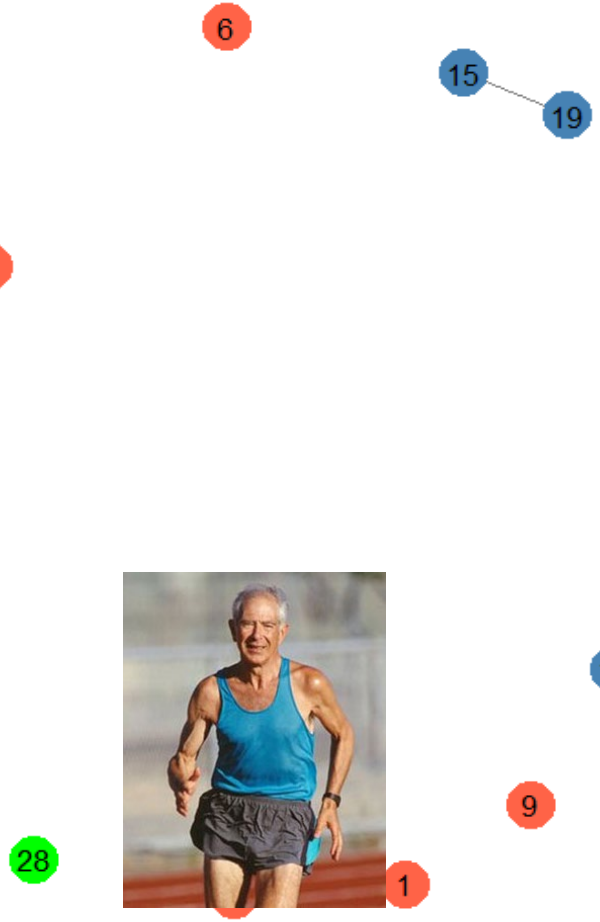
Similarity Network Model- Ankle Data (Moderate PD)-All Features



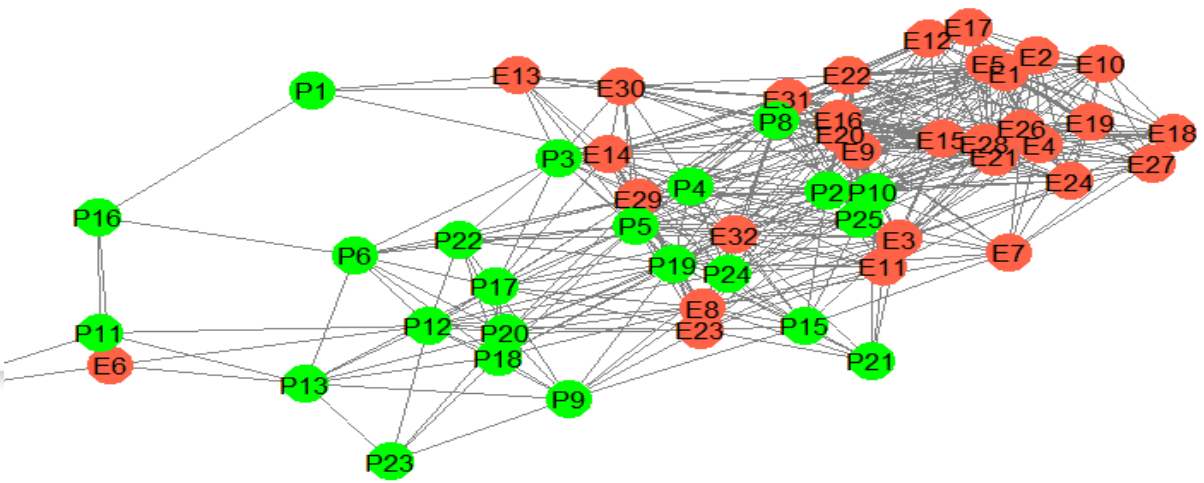


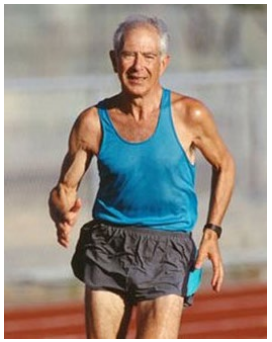
Similarity Network Model- Ankle Data (Mild PD) All Features





Similarity Network Model for the data from the first phase of wrist dataset- Threshold at **90%-** PD and HE





Subject	Gender	Age	MoCA	FoG	FAB	TUG	GDS	H&Y	MFES	Lawton
PD8	Male	69	28	2	39	6.7	0	1	10	8
PD10	Male	71	28	1	39	6.7	1	1	9.3	8
PD1	Male	83	26	8	39	11.2	0	1	8.6	8
PD21	Male	54	25	0	39	9.0	0	1	10	8

Summary of Case Study

- Correlation Network model worked beautifully when we applied it to both dataset.
- MIGMC provided us with the best set of features
- Accelerometers at ankles and Wrist can capture gait parameters that are useful in early diagnosis of disease.
- The performance of Bag-of-Words model is ,if not higher than, equal to the Standard Model
- Ankle data are more precise in identification of patients with PD compared to Wrist Data
- Still wrist could be argued as a better body location (87.5% accuracy is good enough)

Applications in hospital: Post-operative Nursing Care

- A *post-operative* assessment is very important to a full and speedy *recovery from* any type of *surgery*.
  - a full assessment and an individualized treatment plan based upon the patient's needs and level of function, coupled with clinician expectations

Applications for health subject: Physical therapy / Rehabilitation

- help a patient perform rehabilitation exercises to improve their balance and mobility, and
- find exercises that meet patient's specific needs and abilities.

The Bioinformatics Angle
Systems Biology and Network Models

Data Analysis: Systems Biology

UNO BIOINFORMATICS

- 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 x versus Stage y

Case Study in Aging

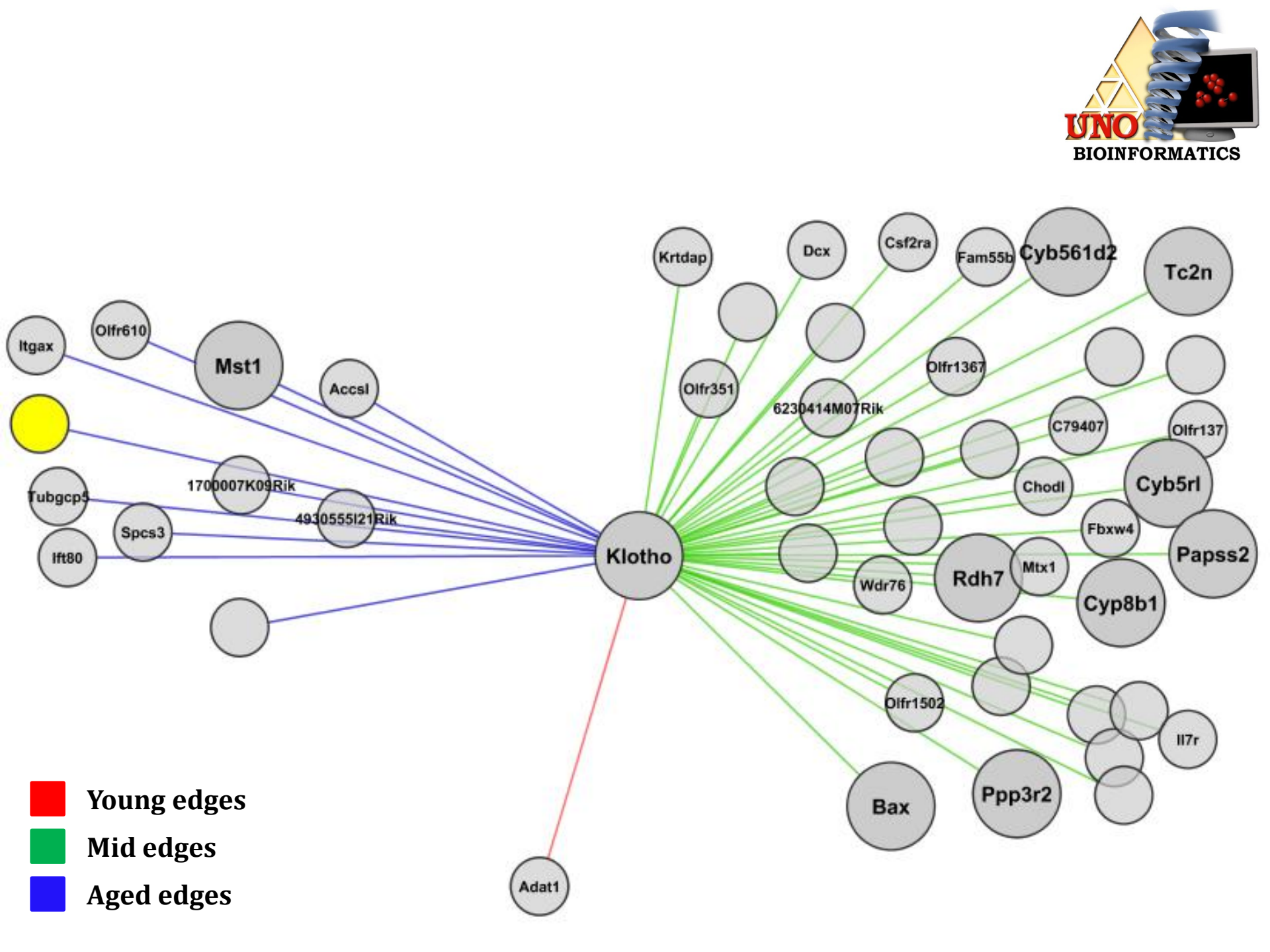
- 5 sets of temporal gene expression data

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

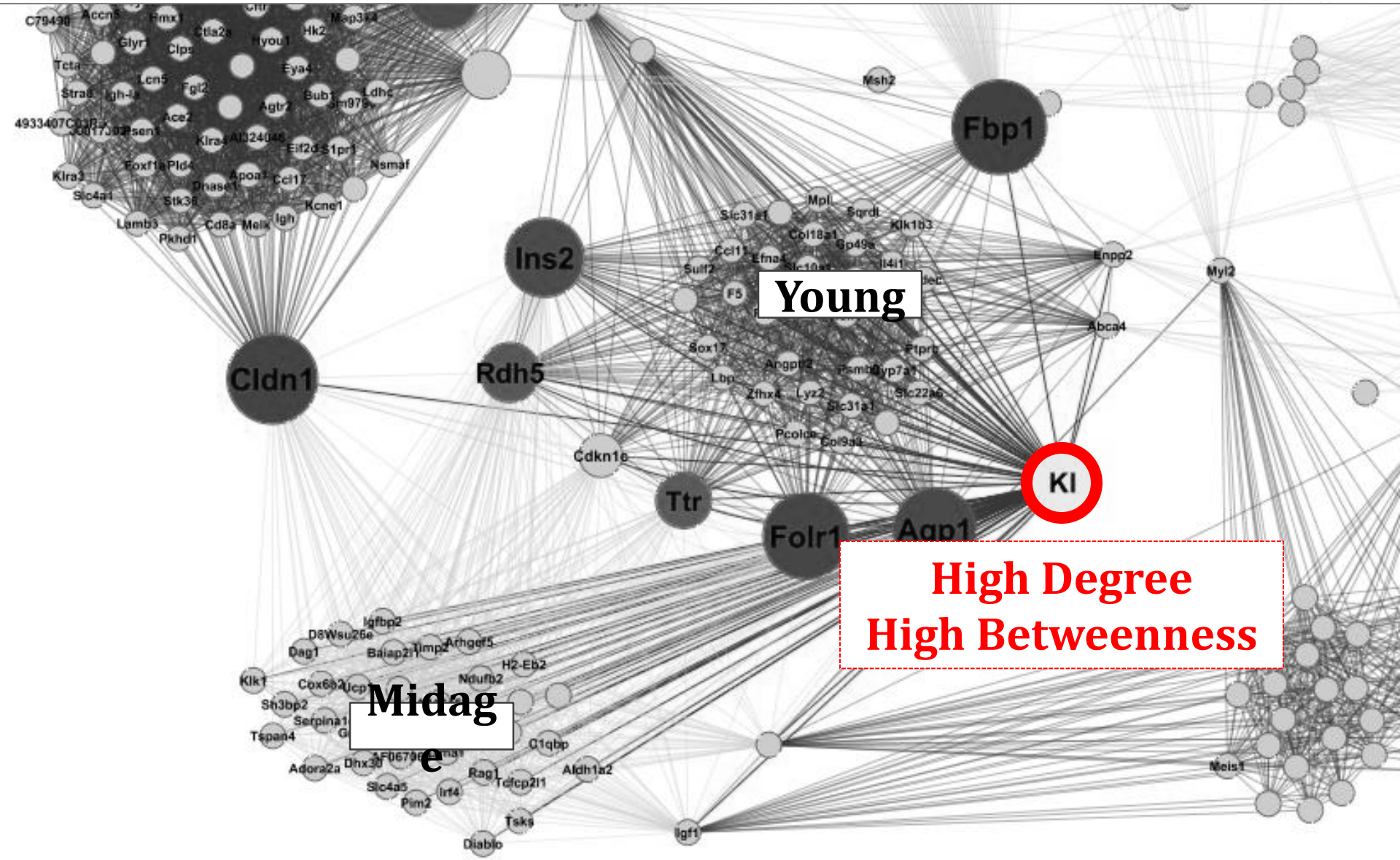
Hub Lethality

- Young Male BalbC Mouse
  - 12/20 hubs tested for *in vivo* knockout
    - 8/12 lethal phenotype pre-/peri-natally
    - 4/12 non-lethal but system-affecting
    - 0/12 no observed phenotype

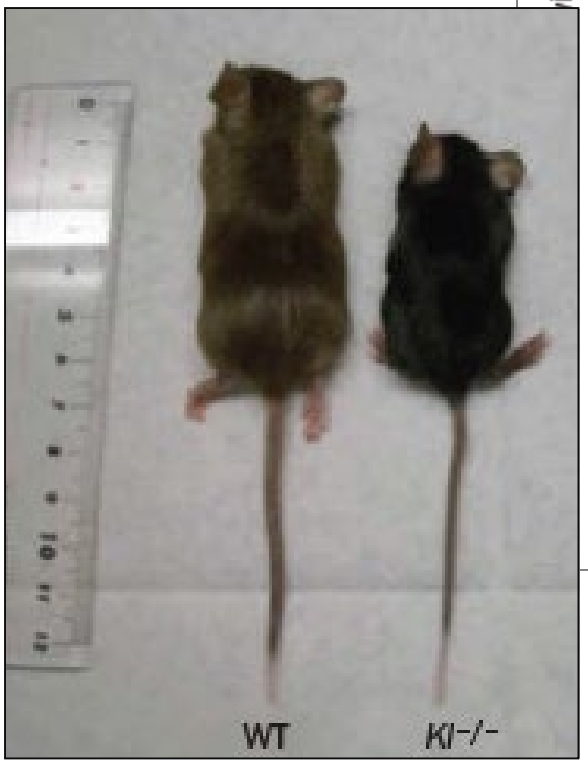
- Aged Male BalbC Mouse
  - 11/20 hubs tested for *in vivo* knockout
    - 7/11 lethal phenotype pre-/peri-natally
    - 3/11 non-lethal but system-affecting
    - 1/11 no observed phenotype (Aldh3a1)

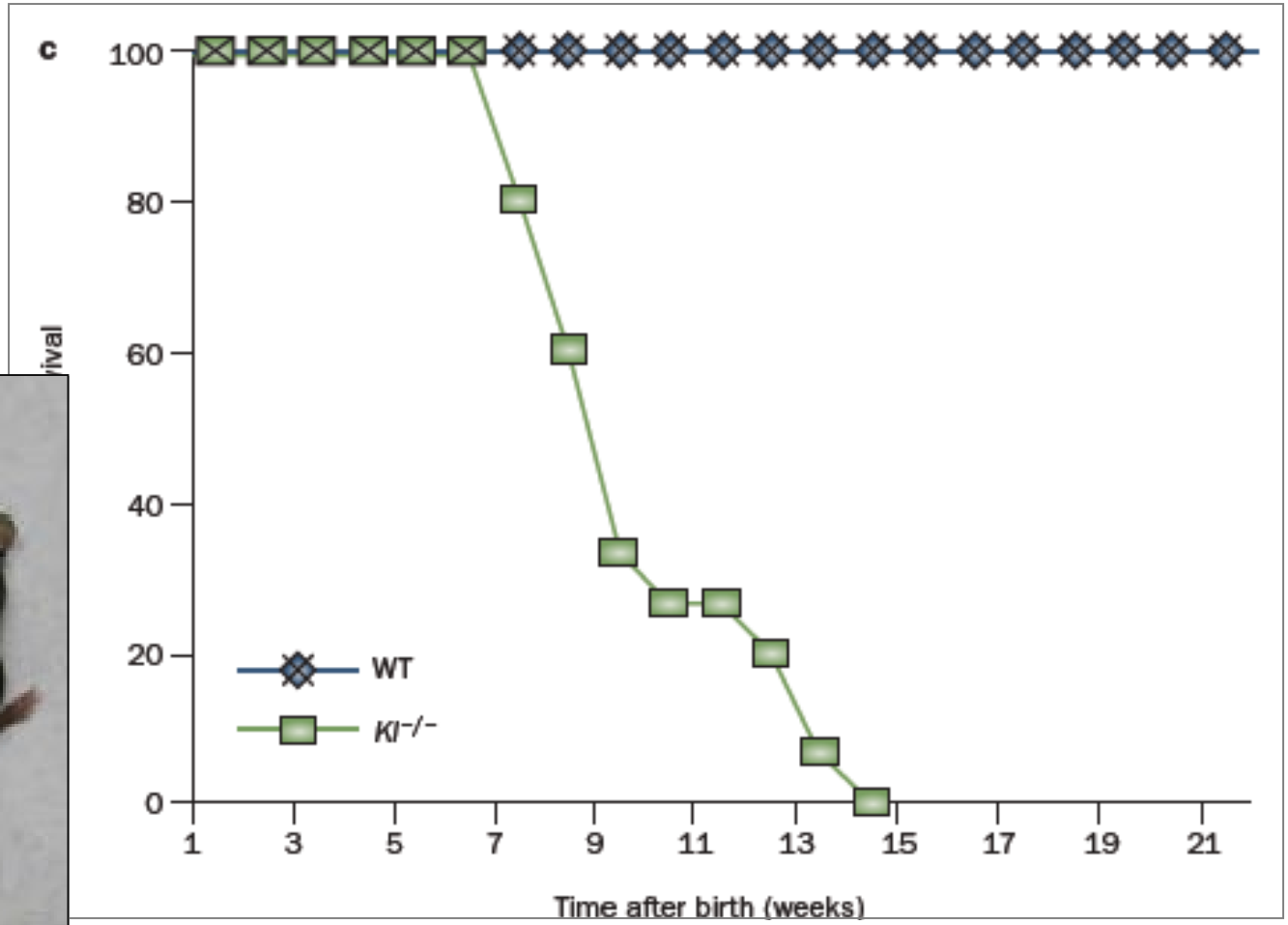

UNO
BIOINFORMATICS
Krtdap
Dcx
Csf2ra
Fam55b
Cyb561d2
Tc2n
Itgax
Olfr610
Mst1
Accsl
Olfr351
Olfr1367
6230414M07Rik
C79407
Olfr137
Tubgcp5
1700007K09Rik
Chodl
Cyb5rl
Spcs3
4930555I21Rik
Fbxw4
Ift80
Klotho
Papss2
Wdr76
Rdh7
Mtx1
Cyp8b1
Olfr1502
Il7r
Bax
Ppp3r2
Adat1
Young edges
Mid edges
Aged edges

Critical Node: Klotho



Validation





Razzaque,2009. Nat. Rev. Endocrin. 5:611-619

HIV and Drug Addiction

- Methamphetamine is a major drug of abuse with reported high use by HIV-infected groups
- Methamphetamine users have higher risk of getting HIV infection
- Impact on nervous system is higher when Methamphetamine is used by HIV infected individual (neuronal injury)

Role of Methamphetamine

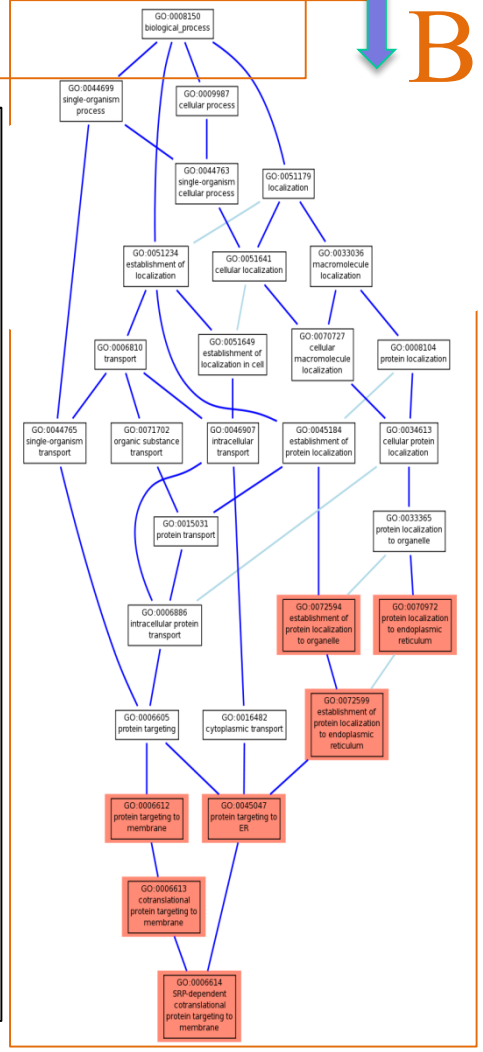
Infected	Not Infected
Infected + Combinatorial Drugs	Not Infected + Combinatorial Drugs
Infected + Meth	Not Infected + Meth
Infected + Meth + Combinatorial Drugs	Not Infected + Meth + Combinatorial Drugs

Clusters Enriched in Specific Functions

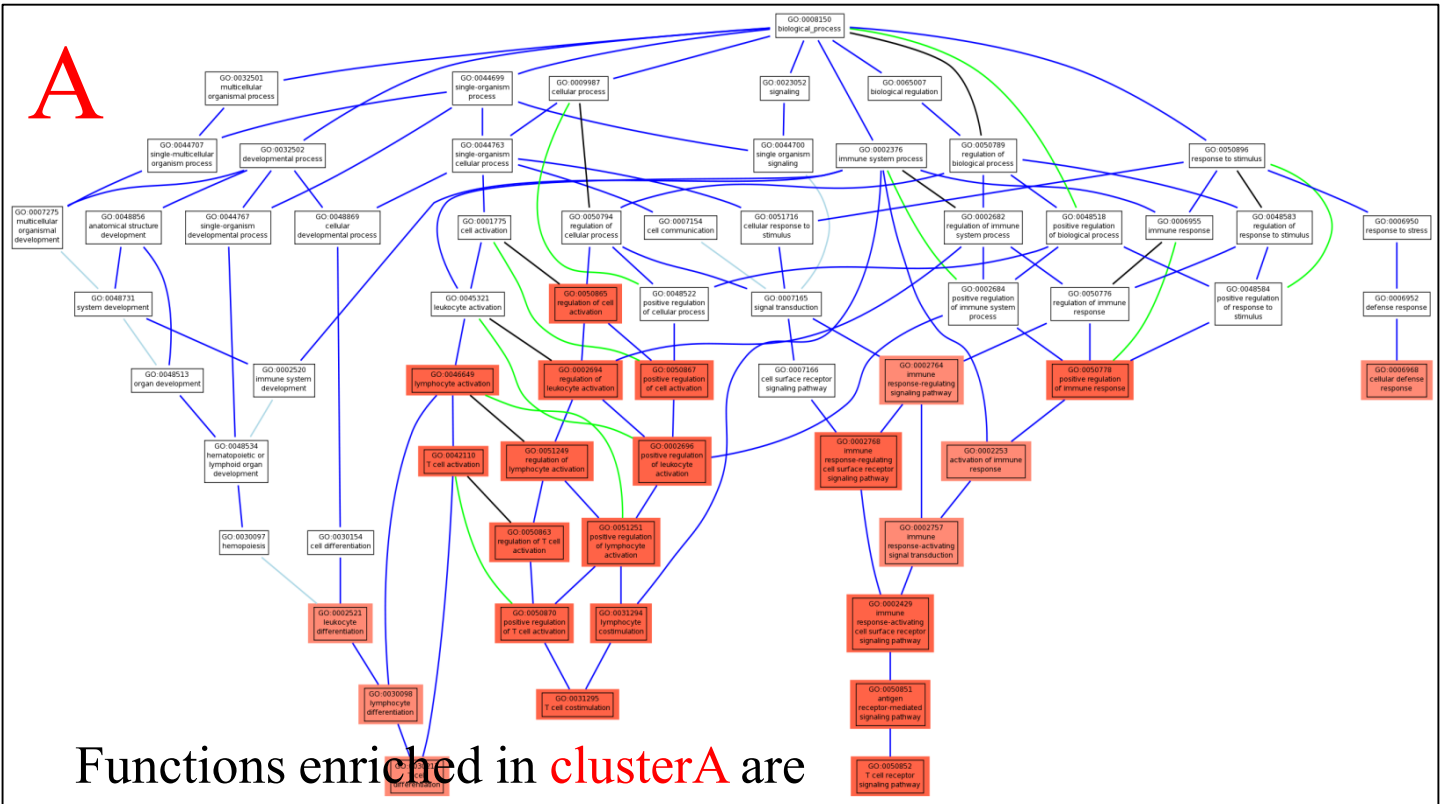
UNO BIOINFORMATICS

Functions enriched in clusterB are protein targeting and localization.

B

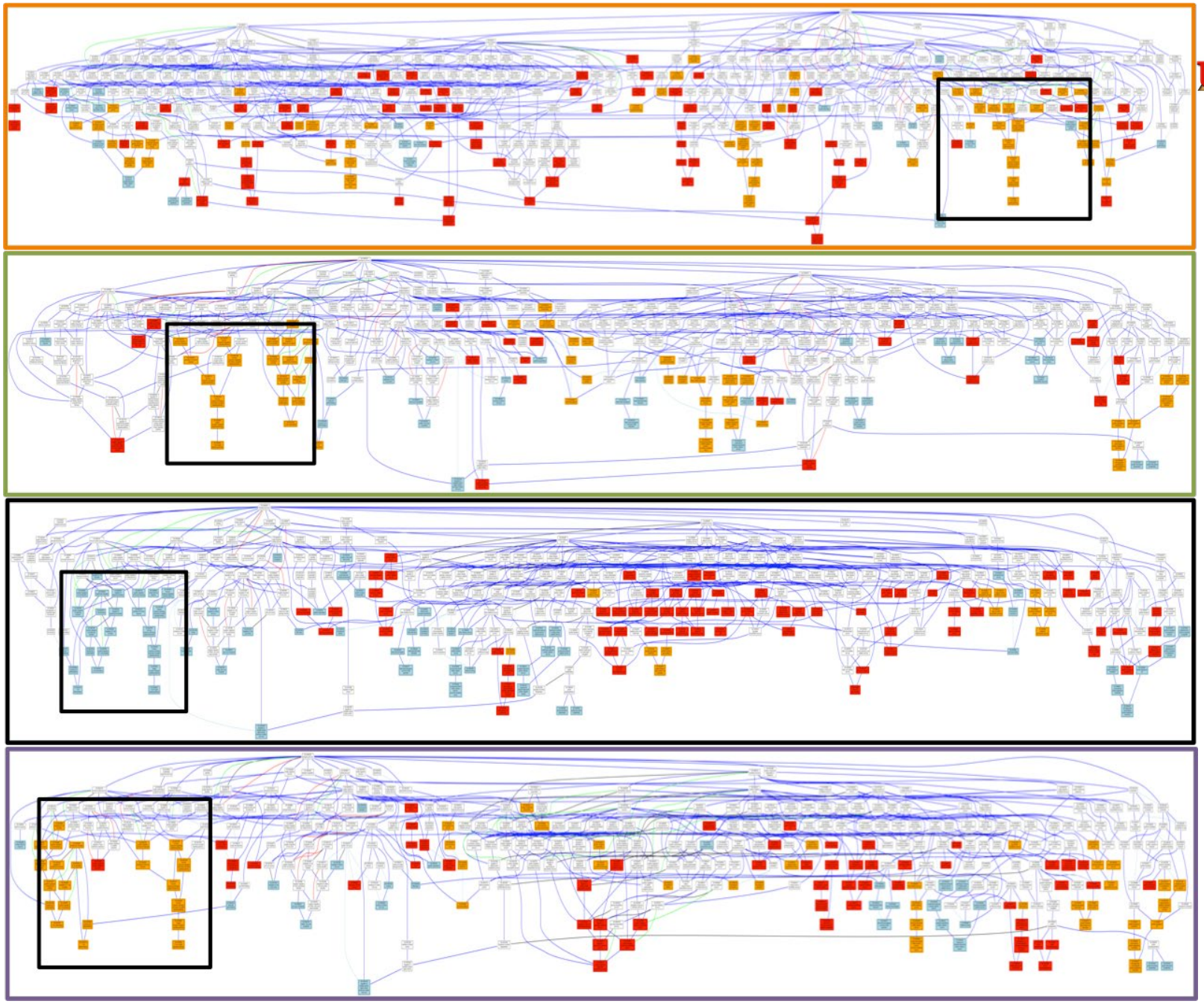


A



Functions enriched in clusterA are all immune response specific.

Orange nodes = enriched in both sets; Blue nodes = enriched only in Uninfected



UNO BIOINFORMATICS

Infected vs. Uninfected

HIV treatment vs. Uninfected

Infected + Meth vs. Uninfected

Infected + Meth+ Treatment vs. Uninfected

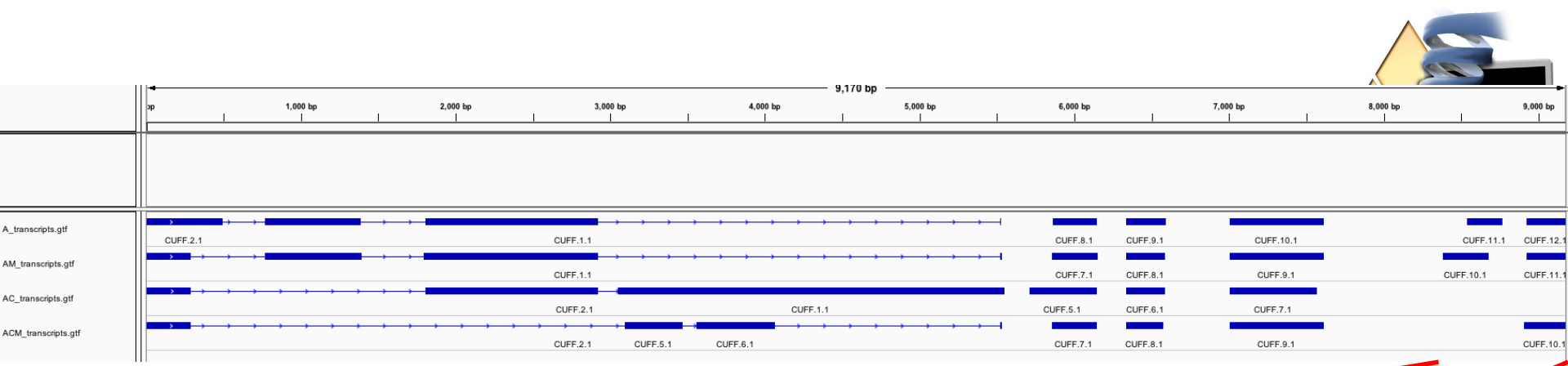
Obtained Results

UNO BIOINFORMATICS

- Large number of nodes are enriched in only one network in Infected + Meth network.
  - Many functions enriched in other conditions have been dropped out in Infected + Meth network.
- Most of the lost functions reappear in Infected + Treated
- Some of these lost functions reappear in Infected + Meth + Treatment

Validation: *nef* gene (GenBank)

- viral accessory protein
- important for virus replication in vivo
- determinant of HIV-1 pathogenesis
- down-regulates cell surface CD4 and MHC class I molecules; enhances virus infectivity through interactions with multiple cellular signaling proteins



Region of *nef* gene

A

AM

AC: disappeared

ACM: comes back

The Computing Angle
How to implement the proposed models

How to implement this stuff? Computational Issues

- Graph/Network Modeling
- Graph Algorithms
- High Performance Computing
  - Beyond surface-level adaptation of known algorithms
- Wireless Networks
- Statistical Analysis
- Storage/processing models - Security and Privacy

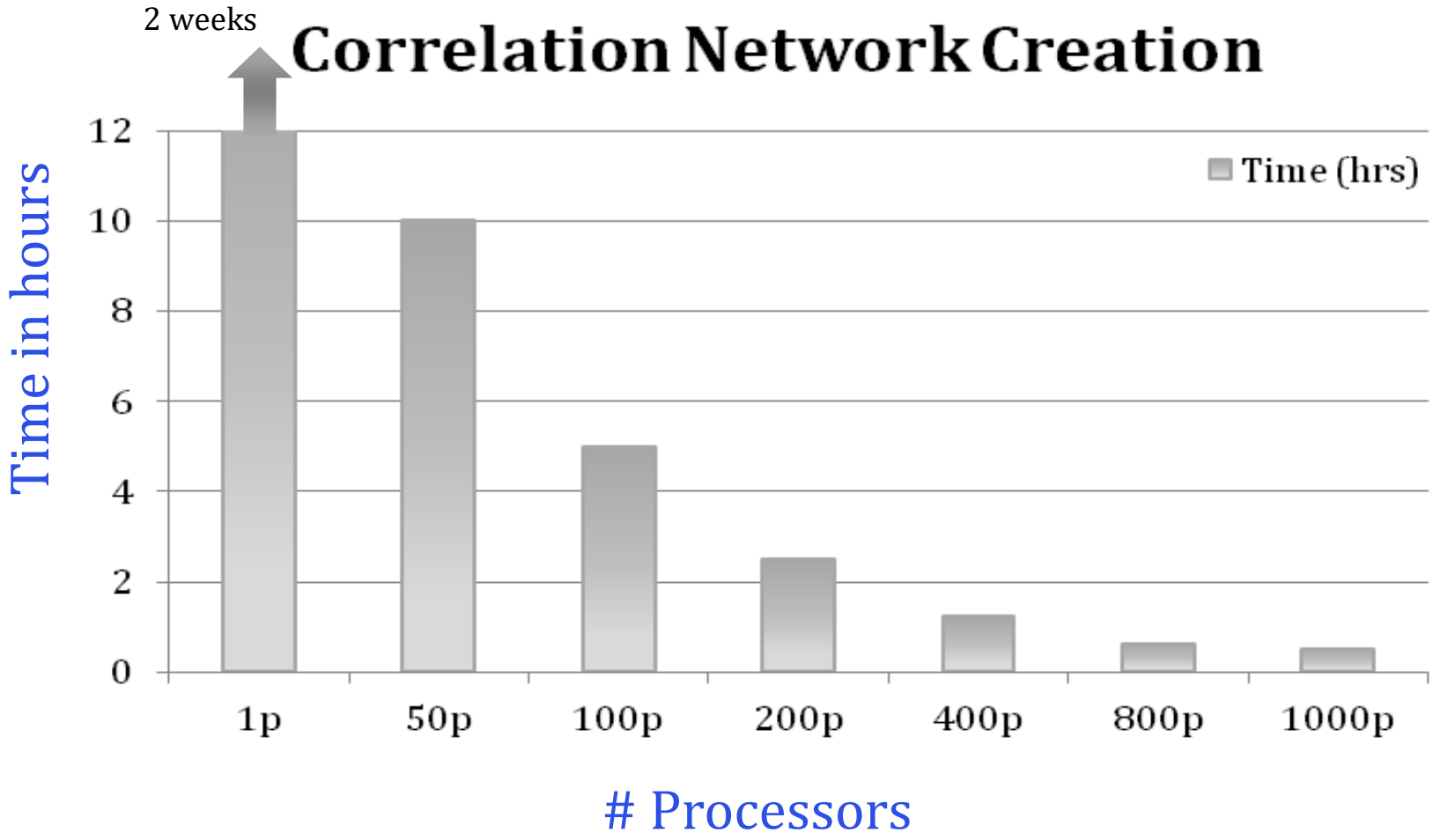
HPC and Big Data

- 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

The Need for HPC



Technical/Innovative Solutions

- Smarter input data: better user level utilization plus integrated domain knowledge with computational tools
- New data reduction models to deal with large data sizes and allow for better and faster data mining
- Better application specific parallelization – custom solutions lead to better performance
- Better scheduling model: multi-layer dynamic scheduling solution

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

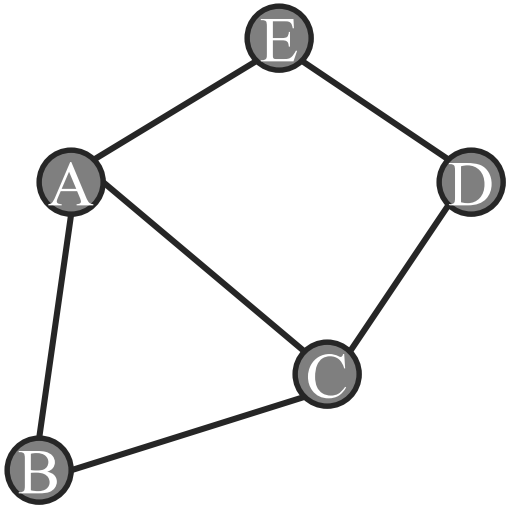
Chordal Graph Sampling

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)**

Why chordal graphs?

- Chordal graphs are triangulated
  - We want to preserve K_3 subgraphs (triangle)
  - K_3 graphs/motifs are known to represent co-regulated genes
  - Use chordal graphs as a filter for finding co-regulated structures



Subgraph formed by A,B,C is more likely to be biologically relevant.

If gene A and gene B are co-regulated, and if gene A and gene C are co-regulated, then genes B and C will be co-regulated.

Biosciences at Crossroads

- Many Scientific disciplines are now at crossroads
- The proper penetration of IT represent tremendous challenges and great opportunities
- The importance of interdisciplinary approach and knowledge integration to problem solving
- The need for in-depth analysis and problem solving rather than the surface-level approaches
- This may lead to scientific revolution

Acknowledgments

- UNO Bioinformatics Group

  Kiran Bastola

  Sanjukta Bhoomwick
  Kate Cooper
  Dario Ghersi
  Ishwor Thapa
  Ling Zhang

  Sean West
  Vi Dam
  Suyeon Kim
  Donovan Orn
  Elham Rastegari

- Biomedical Researchers

  Richard Hallworth
  Vivian Marmalat

- Funding Sources

  NIH
  NSF

  Nebraska Research Initiative

- Former Group Members

  Alexander Churbanov
  Xutao Deng
  Huiming Geng
  Xiaolu Huang
  Daniel Quest
  Julia Warnke-Sommer