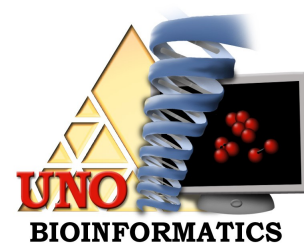


University of Nebraska at Omaha



Data Analysis and Integration Tools in  
Biomedical Informatics – Case Study in Aging  
Research



BIOTECHNO 2013

Hesham H. Ali

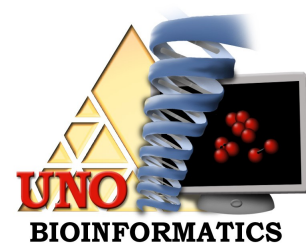
UNO Bioinformatics Core Facility  
College of Information Science and Technology

# Tutorial Outlines

- Introduction to Biomedical Informatics
  - State of the discipline - Challenges and Opportunities
  - Data-driven biomedical research
- Next Generation Bioinformatics Tools
  - Intelligent Collaborative Dynamic (ICD) Tools
- Case Study: Aging Research
  - The genomic study: Correlation Networks
  - Mobility and aging: Wireless monitoring
  - Data collection and Virtual Environments
- Next Steps: Where do we go from here?
  - HPC and Cloud Computing

# Tutorial Outlines

- *Introduction to Biomedical Informatics*
  - State of the discipline - Challenges and Opportunities*
  - Data-driven biomedical research
- Next Generation Bioinformatics Tools
  - Intelligent Collaborative Dynamic (ICD) Tools
- Case Study: Aging Research
  - The genomic study: Correlation Networks
  - Mobility and aging: Wireless monitoring
  - Data collection and Virtual Environments
- Next Steps: Where do we go from here?
  - HPC and Cloud Computing



# Biosciences will never be the same

- IT advances changed Biosciences forever
- So much biological data is currently available
- The availability of data shifted many branches in Biosciences from pure experimental disciplines to knowledge based disciplines
- Integrating Computational Sciences and Biosciences is critical but not easy
- Would Bioinformatics be the answer?





# Biomedical Informatics (BMI)

- Bioinformatics
- Biomedical Imaging
- Health Informatics
- Medical Informatics
- Public Health Informatics

# Biomedical Informatics – Where are we?

- High throughput data
- Next generation sequencing
- Personalized medicine
- Sensor based monitoring systems
- Biomarkers
- Genome-wide association study
- Differentially expressed genes
- Single position variants and copy number variants
- ...

# State of the Field - BMI

- Availability of many large useful database systems; private and public
- Availability of numerous helpful software packages
- Fragmented, in some case isolated, efforts by computational scientists and bioscientists
- Advances in new technologies as high throughput next generation sequencing
- The trendiness of the discipline
- Increasing use of sensors in monitoring applications
- Huge interest from Industry, researchers and the public

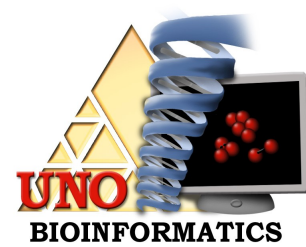
# Data Generation vs. Integration/Analysis

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

# Systems Biology Approach

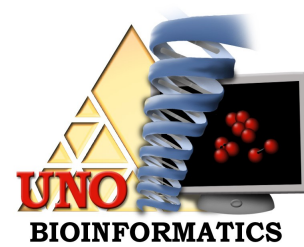
- Realistic and Innovative:
  - Networks model relationships, not just elements
  - Discover groups of relationships between genes and gene products
- Validation and Discovery Aspects
  - Examine changes in systems
    - Normal vs. diseased
    - Young vs. old
    - Stage I v. State II v. Stage III v. Stave IV

# Systems Biology




- Holist view of the system
  - Ability to zoom in/out to view critical system components
- Past: Reductionist biology
  - Find a gene/protein of interest
  - Examine under different conditions
- Systems biology: examine an entire system at different conditions

# Why Networks/Graphs?



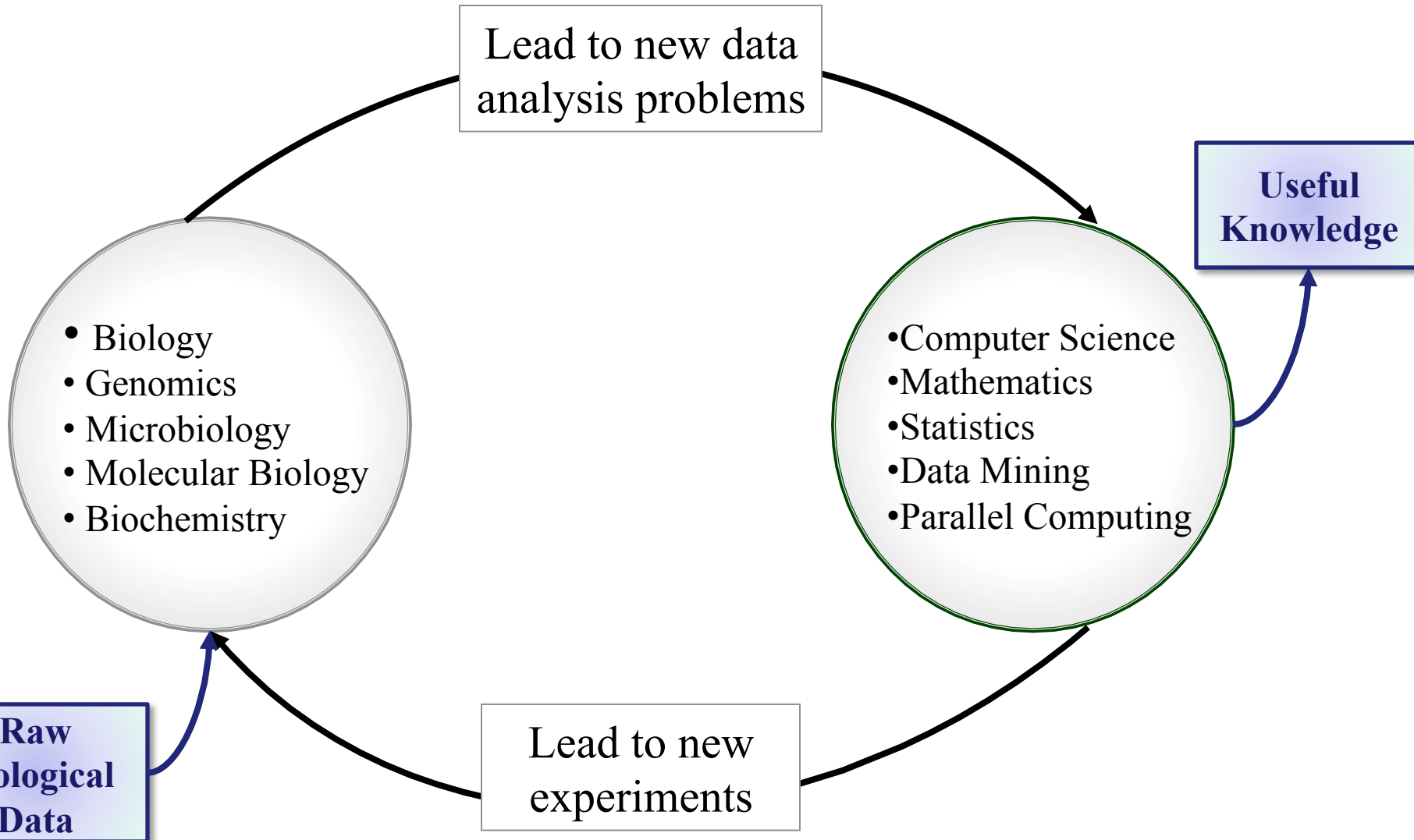
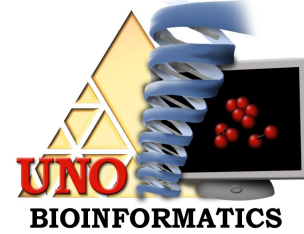
- Explosion of biological data

Site contents	
Public data	
Platforms	9,267
Samples	611,215
Series 	24,571
DataSets	2,720

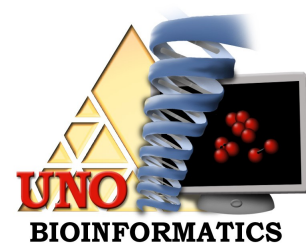
**Each sample can have over 40,000 genes**

- Average microarray experiment: 1200 pages of data\*
- How can we extract information from data?

# Interactive Aspect of BMI







# Current Steps in Nebraska

- Great interest from researchers/educators/students
- Support from the administration
- Infrastructure Supported by NRI, NICLS and INBRE
- A number of ground breaking research projects
- Various grants funded by NIH and NSF
- High degree of communication
- New innovative programs in Biomedical Informatics at all levels: undergraduate, masters, and doctoral

Focus on an interdisciplinary approach

# Informatics versus Computing

- The information age: taking full advantage from available information
  - From data to information to knowledge to *Wisdom*
  - Data driven decision making
- IT is a super scientific discipline that includes the disciplines that address issues related to collecting, storing, managing, processing information, and employing information and algorithmic techniques to solve problems in various application domains.

**Interdisciplinary Informatics**

**Information  
Technology**

**Management Info Systems**

**Computer Science**

Decision Support  
Info Management  
System Analysis and Design  
Human Factors

Quantitative Foundations  
Database/Data Mining

Software Engineering

Prog Languages  
Simulation/Modeling  
Theory of Computation  
Artificial Intelligence  
Automation  
Com Networks

Information Assurance

Collaboration Science

Education Info

Bioinformatics

PH Info

Innovation

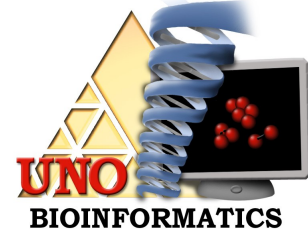
Eng Informatics

# Tutorial Outlines

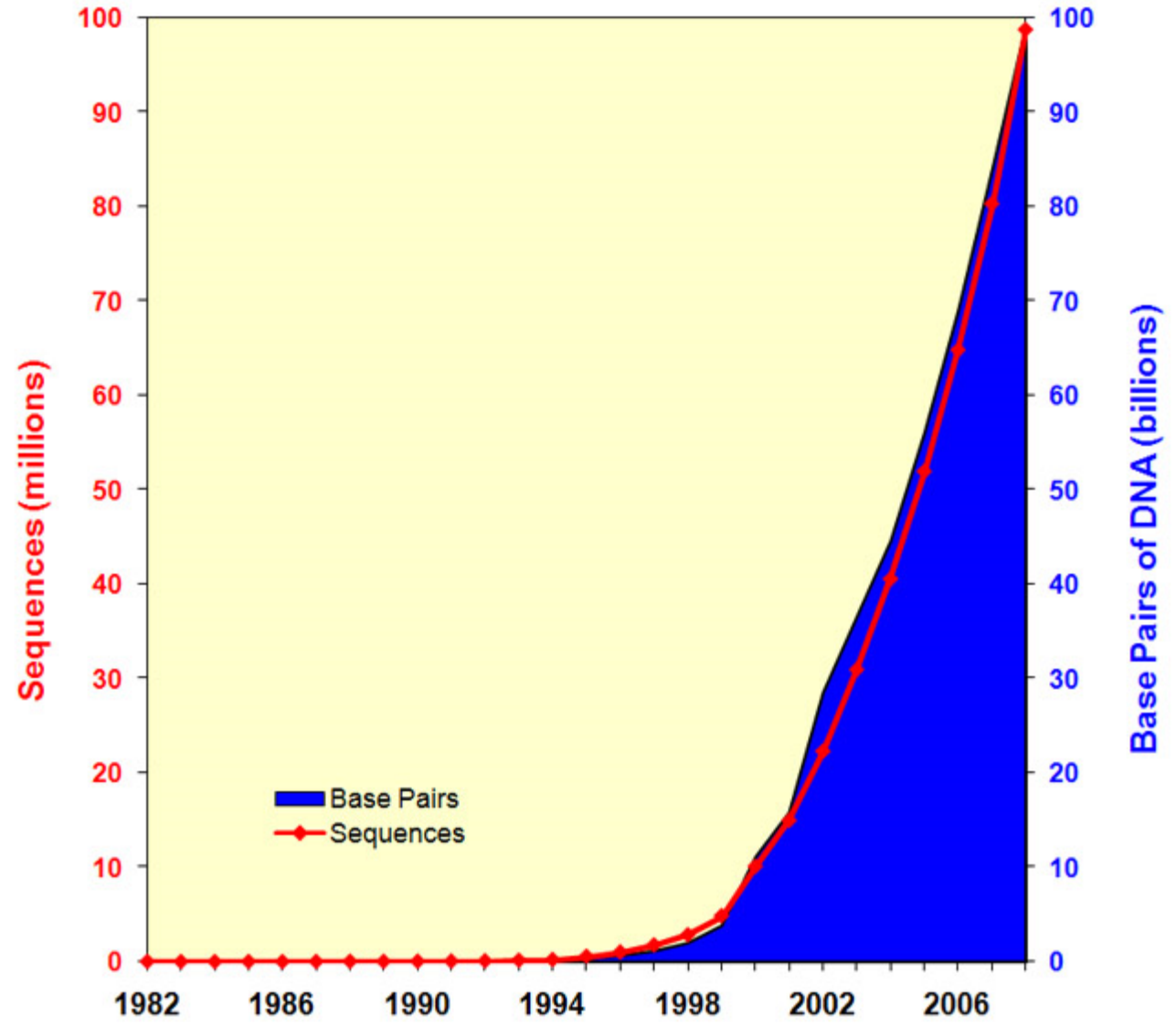
- *Introduction to Biomedical Informatics*
  - State of the discipline - Challenges and Opportunities
  - Data-driven biomedical research*
- Next Generation Bioinformatics Tools
  - Intelligent Collaborative Dynamic (ICD) Tools
- Case Study: Aging Research
  - The genomic study: Correlation Networks
  - Mobility and aging: Wireless monitoring
  - Data collection and Virtual Environments
- Next Steps: Where do we go from here?
  - HPC and Cloud Computing

# A Focus on Biological Database

- Mainly large set of catalogues sequences.
- No extra capabilities of fast access, data sharing or other features found in standard database management systems.
- Collection of sequences complemented with additional information such as origin of the data, bibliographic references, sequences function (if known) and others.
- The trendy factor and the lack of integration



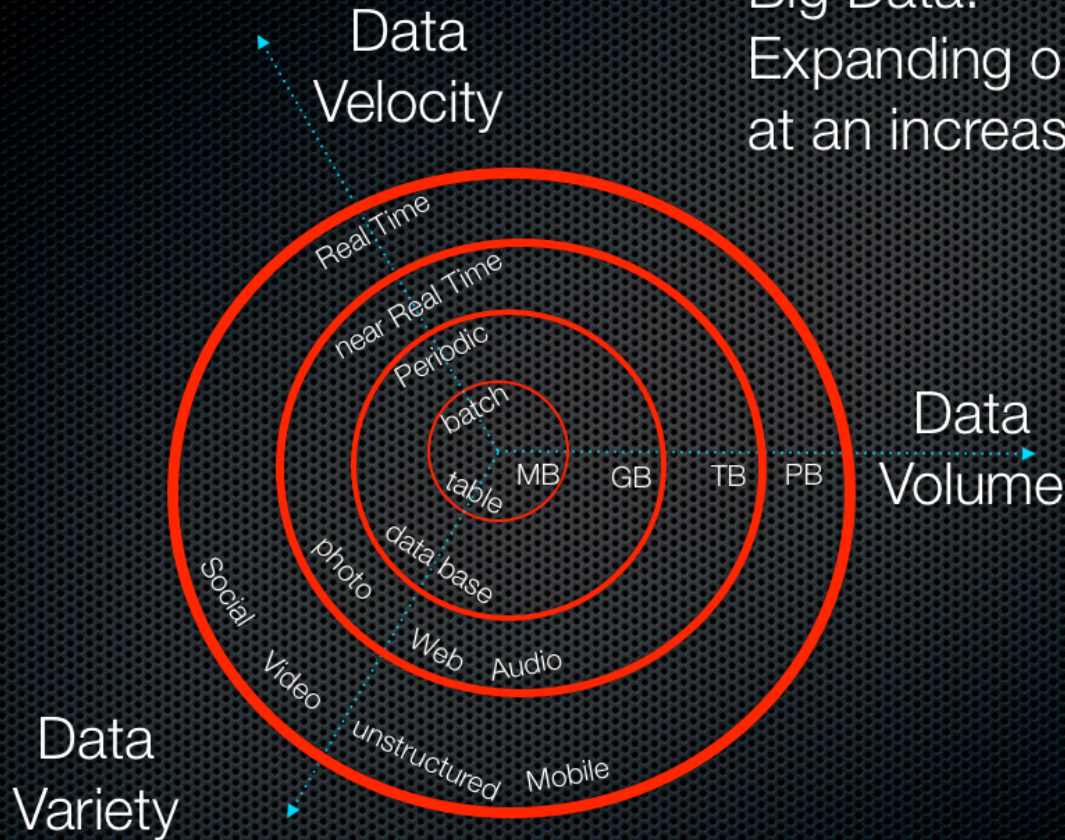
# Growth of GenBank (1982 - 2008)





# “Big Data”

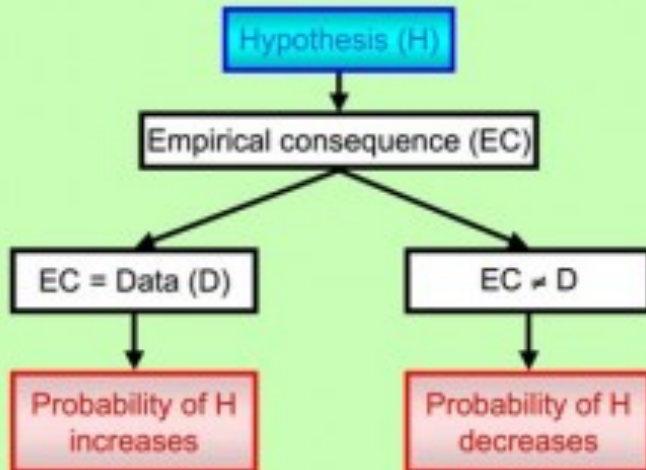
Big Data:  
Expanding on 3 fronts  
at an increasing rate.



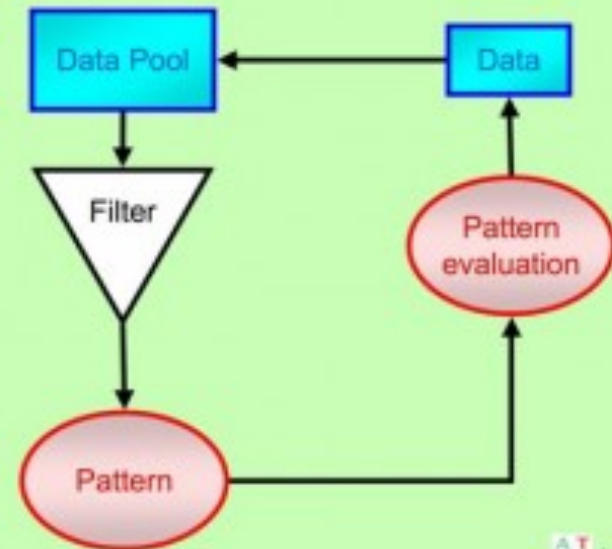
# A Potential Major Change

- Data driven research vs. Hypothesis driven research

## Hypothesis driven research - Concept



## Data driven research - Concept





# Impact of New Technology

- Next Generation Sequencing
  - Push towards “personal sequencers”
  - Higher error rates due to mobility, desire for affordable cost
  - Creates a need for change in sequence analysis algorithms

## Ion Proton™ Sequencer



## Benchtop Genome Center

Powered by Ion Torrent™ semiconductor chip technology, the Ion Proton™ Sequencer is the first benchtop next-generation sequencer to offer fast, affordable human genome and human exome sequencing.

Have a rep  
contact me >



Proton Technology

ion torrent  
by life technologies™

# Impact of New Technology

- High Performance Computing
  - Need for algorithms that are *fast, effective*
  - Need for systems that can hold models in memory at once
  - Need for new ways to compute quickly

CNET › News › Nanotech - The Circuits Blog

## Report: IBM researcher says Moore's Law at end

IBM Fellow Carl Anderson says at a conference this week that Moore's Law is hitting a ceiling, according to a report.



by Brooke Crothers | April 9, 2009 9:00 PM PDT



# The Impact of Sequencing Technology

- Third Generation Sequencing technology
  - Higher throughput
  - More accurate
  - Longer reads
- Personal Sequencers?
  - Less expensive
  - Shorter reads
  - Errors versus variants
- The analogy of sequencing technologies to computing advancements

## Issue with Current Biological Data Bases

- The large degree of heterogeneity of the available data in terms of quality, completeness and format
- The available data are mostly in raw format and significant amount of processing is needed to take advantage of it
- Mostly in semi flat files – hence the lack of structure that support advanced searching and data mining

# Data versus Knowledge

- With high throughput data collection, Biology needs ways not only to store data but also to store knowledge (Smart data)
- Data: Things that are measured
- Information: Processed data
- Knowledge: Processed data plus meaningful relationships between measured entities

Power of graph modeling

# The Industry Perspective

- Deliver the right treatment to the right patient with the right dosage at the right time (the first time)
- How to leverage data?
  - Integratable?
  - Scalable?
- Hybrid research needs to be developed in non-linear fashion
  - Example: Pancreatic Cancer research at CMU – Boolean networks and hybrid automation that produced 12 candidate genes for further study
- Achieving the balance
  - **Eliminate division** between theory and experimental work
  - **Guide** the experimental design and theory design
  - **Understand** the generated and processed data

# Bioinformatics Data Cycle

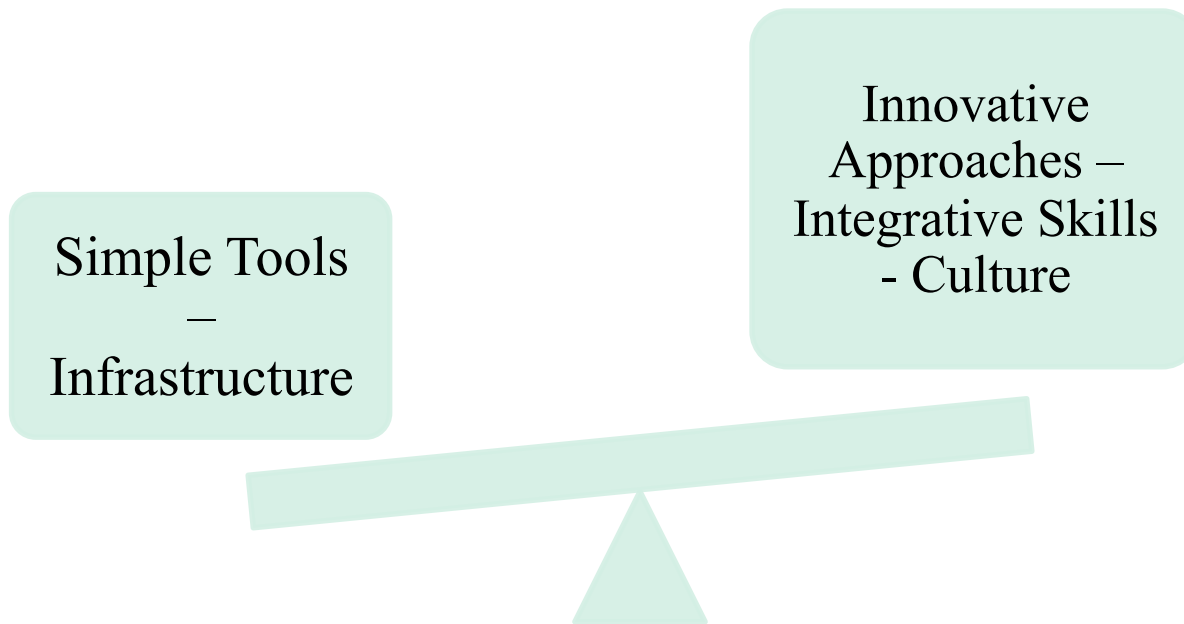
- Data Generation and Collection
- Data Access, Storage and Retrieval
- Data Integration
- Data Visualization
- Analysis and Data Mining
- Decision Support
- Validation and Discovery

# Data-Driven Decisions

- With high throughput data collection, Biology needs ways not only to store data but also to store knowledge (Smart data)
- Data: Things that are measured
- Information: Processed data
- Knowledge: Processed data plus meaningful relationships between measured entities
- Decision Support

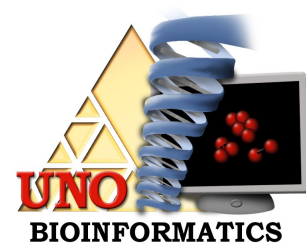


# Tipping the balance



# Tutorial Outlines

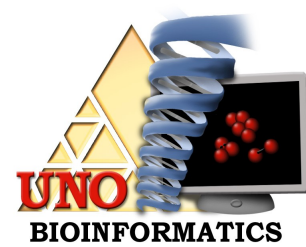
- Introduction to Biomedical Informatics
  - State of the discipline - Challenges and Opportunities
- *Next Generation Bioinformatics Tools*
  - Intelligent Collaborative Dynamic (ICD) Tools*
- Case Study: Aging Research
  - The genomic study: Correlation Networks
  - Mobility and aging: Wireless monitoring
  - Data collection and Virtual Environments
- Next Steps: Where do we go from here?
  - HPC and Cloud Computing



# Generations of Bioinformatics Tools

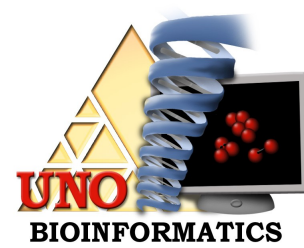
- Simple computational tools for manipulation of available data
- Complex algorithmic tools
- Data-focus tools: curated data, clean data, managed access to data

# Early Generation Bioinformatics Tools



- Filled an important gap
- Mostly data independent
- Based on standard computational techniques
- Has little room for incorporating biological knowledge
- Developed in isolation
- Focus on trendy technologies
- Lack of data integration
- Lack of embedded assessment

# Examples of First Generation Bioinformatics Tools

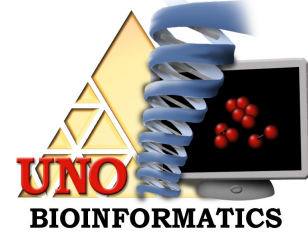


- Sequence comparison (alignment) tools
- Phylogenetic trees generation tools
- Microarray data statistical tools
- Clustering tools
- Hidden Markov Model (HMM) Based Tools

# Next Generation Tools

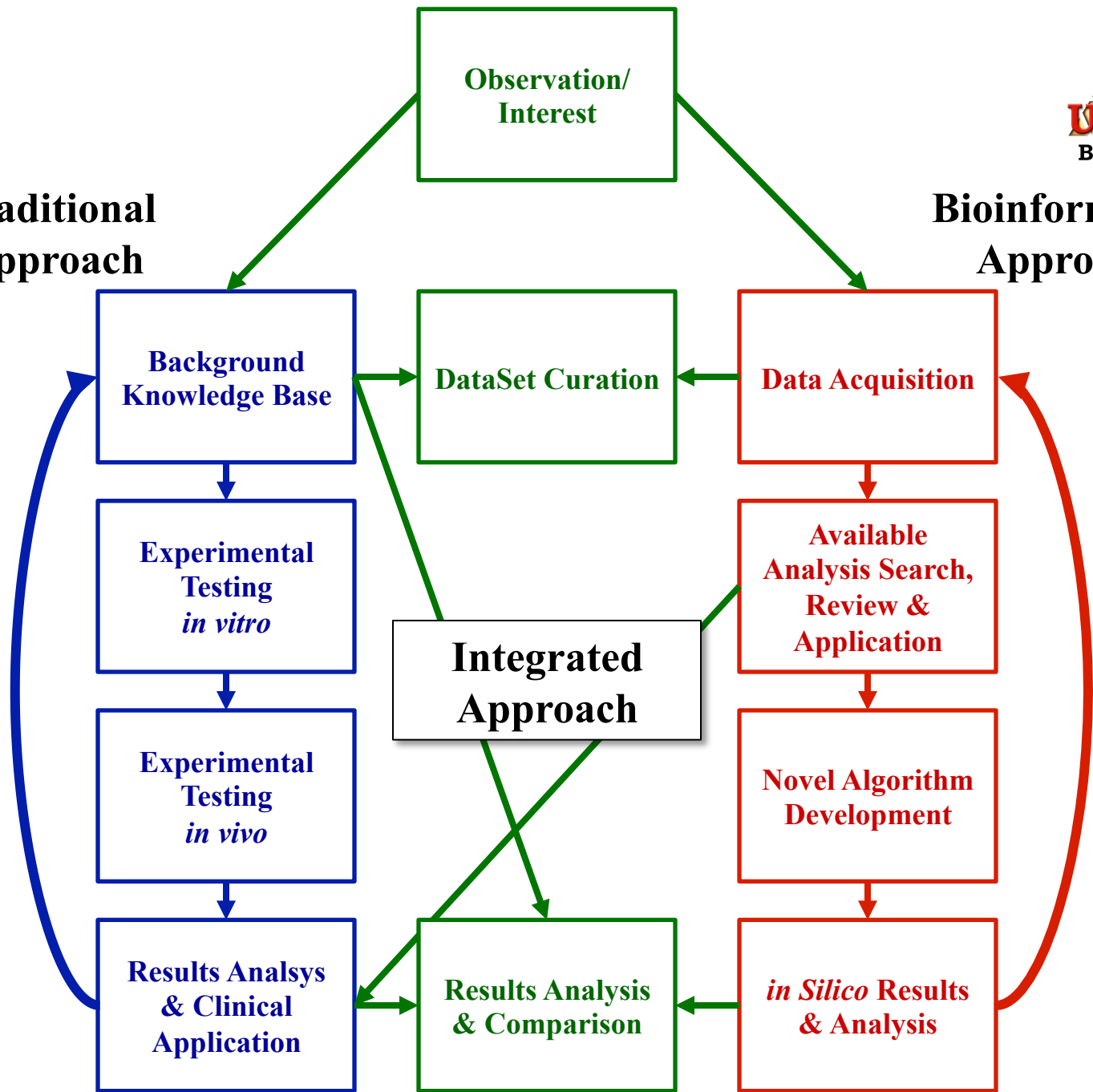
- **Dynamic:** Custom built and domain dependent
- **Collaborative:** Incorporate biological knowledge and expertise
- **Intelligent:** based on a learning model that gets better with additional data/information

Intelligent Collaborative Dynamic (ICD) Tools highlight the need for data integration and explore interrelationships of data elements



### Traditional Approach

### Bioinformatics Approach



**Background Knowledge Base**

**Experimental Testing *in vitro***

**Experimental Testing *in vivo***

**Results Analysis & Clinical Application**

**DataSet Curation**

**Integrated Approach**

**Results Analysis & Comparison**

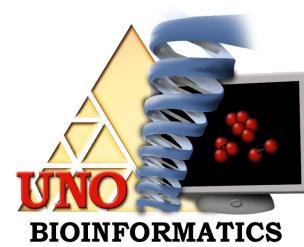
**Data Acquisition**

**Available Analysis Search, Review & Application**

**Novel Algorithm Development**

***in Silico* Results & Analysis**

# Examples of Next Generation Tools



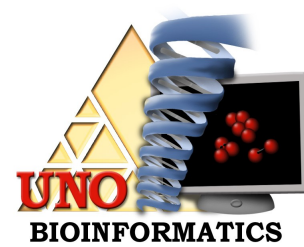
- ICD approaches for sequence comparison and recognition/classifications of microorganisms
- Network analysts approaches for integrating and analysis of heterogeneous biological data
- Domain-specific approaches for genome assembly
- Translational bioinformatics approach for aging research – integrating bioinformatics, health informatics and public health informatics
- ICD approaches for genome wide studies



# Systems Biology Approach

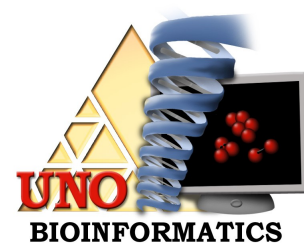
- Realistic and Innovative:
  - Networks model relationships, not just elements
  - Discover groups of relationships between genes and gene products
- Validation and Discovery Aspects
  - Examine changes in systems
    - Normal vs. diseased
    - Young vs. old
    - Stage I v. State II v. Stage III v. Stave IV

# Systems Biology




- Holist view of the system
  - Ability to zoom in/out to view critical system components
- Past: Reductionist biology
  - Find a gene/protein of interest
  - Examine under different conditions
- Systems biology: examine an entire system at different conditions

# Why Networks/Graphs?



- Explosion of biological data

Site contents	
Public data	
Platforms	9,267
Samples	611,215
Series 	24,571
DataSets	2,720

**Each sample can have over 40,000 genes**

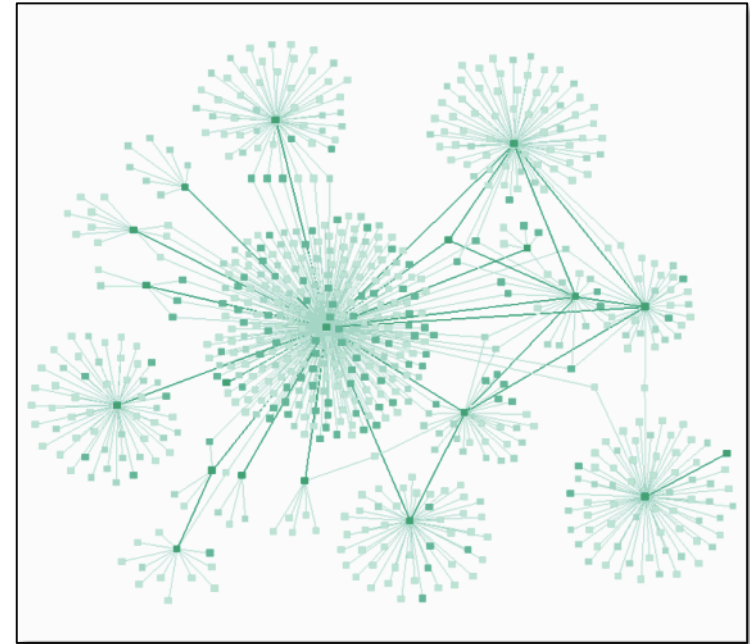
- Average microarray experiment: 1200 pages of data\*
- How can we extract information from data?

# ICD Tools and HPC/Clouds

- How does the network allow us to achieve these ICD goals?
  - Layers of information
  - Integration of different types of knowledge
  - High performance computing
    - Key to analysis of large, complex sets of data with multiple layers

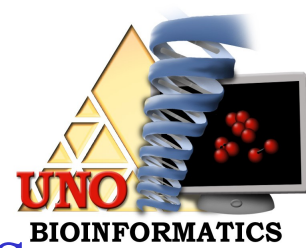
# Biological Networks

- A biological network represents elements and their interactions
- Nodes → elements
- Edges → interactions
- Can represent multiple types of elements and interactions



# State of the Discipline

- Biological Data is a Tsunami that is sweeping the society
- New Generated data from Biomedical instruments plus the availability through the web and data banks
- Data generation is no longer as critical as it is used to be
- Problems related to data integration and data analysis continue to escalate
- Broad impact and applications in many facets of society such as healthcare, environmental studies and energy issues



# Challenges in Biomedical Informatics

- Data Integration models
- Knowledge representation
- Visualization
- Personalization
- Cost

# Opportunities

- Data analysis and integration:
  - Collaboration
  - Multiple angles to approaching Bioinformatics problems
  - Validation and assessment
- Adaptive algorithms and tools:
  - New technologies
  - Various domains
- Short research cycles versus long research cycles



# Next Generation Tools

- **Dynamic:** Custom built and domain dependent
- **Collaborative:** Incorporate biological knowledge and expertise plus facilitate the integration of various, potentially heterogeneous biological data
- **Intelligent:** based on a learning model that gets better with additional data/information

*Intelligent Collaborative Dynamic (ICD) Tools  
with a focus on assessment*

# A Sample of ICD Tools

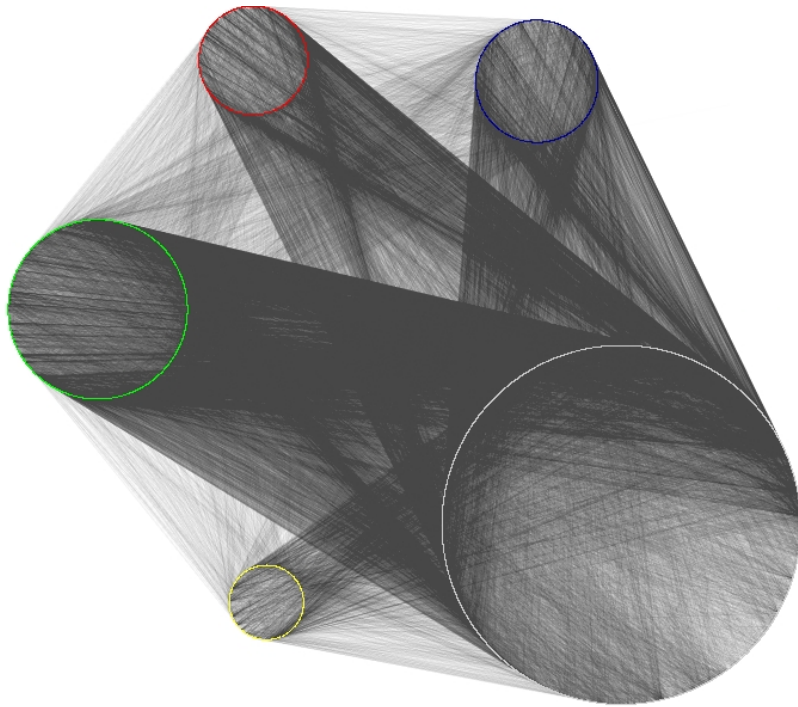
- Grammar Based Identification and Classification Tool
- Using Data Compression to Compare Sequences
- Using Cut Orders in the Recognition and Classification of Biological Sequences
- Next Generation Sequencing: A Graph-Theoretic Assembly Tool of Short Reads
- ICD Tools for the Identification of Similarities and Differences in Correlation Networks
- ICD Bioinformatics Tool for Finding Structural Motifs in Proteins

# Nebraska gets its very own organism

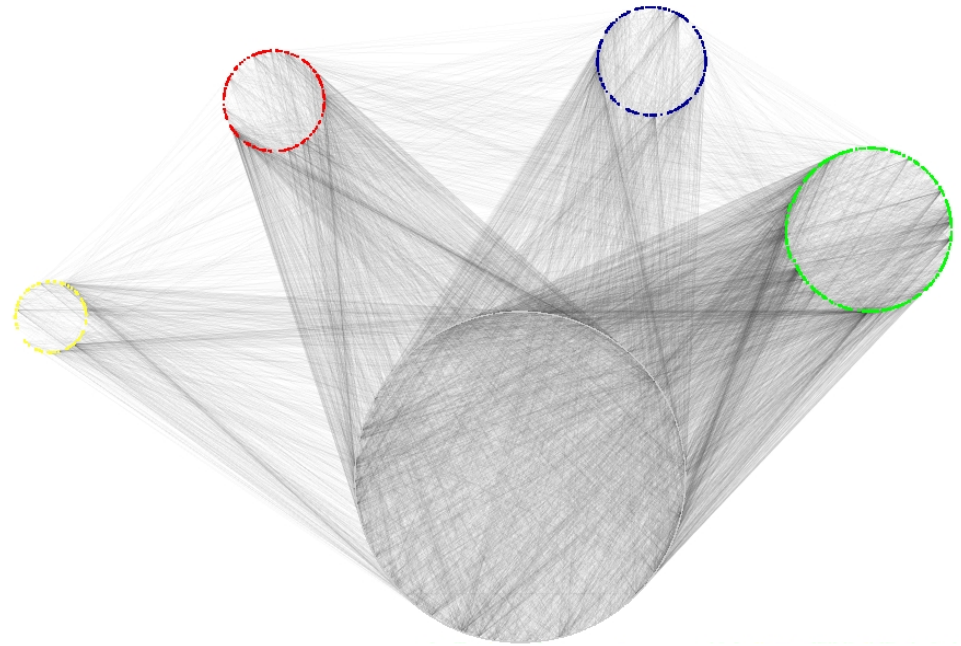
- ✚ While trying to pinpoint the cause of a lung infection in local cancer patients, they discovered a previously unknown micro-organism. And they've named it "mycobacterium nebraskense," after the Cornhusker state.
- ✚ It was discovered few weeks ago using Mycoalign: A Bioinformatics program developed at PKI



# Aging and Biological Networks



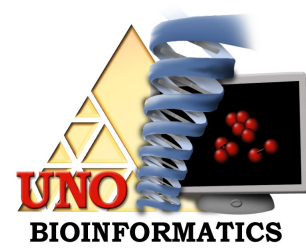
[young]



[aged]

# From Fragments to Sequences to Networks

- Fragments assembly
  - Constructing genomic sequences from short reads
  - Take advantage of high throughput sequencing instruments
- Alignment-free sequence comparisons
  - Alignment of imperfect sequences
  - Alignment of genomic fragments
- Correlation Networks
  - Systems biology approach
  - Data integration tools



# Next Generation Sequencing: ICD Assembly Tool of Short Reads

# Whole genome sequencing



ABI PRISM 3100 Genetic Analyzer

## *Sanger sequencing*

Old platform for DNA sequence determination

~500 bp Fragments

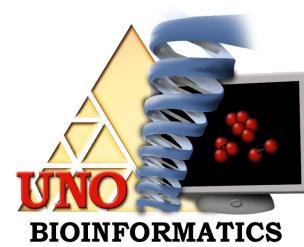
- 6 to 10x Coverage of the Original Sequence
- Place great emphasis on the optimal exploration of all reads

## *Next generation sequencing*

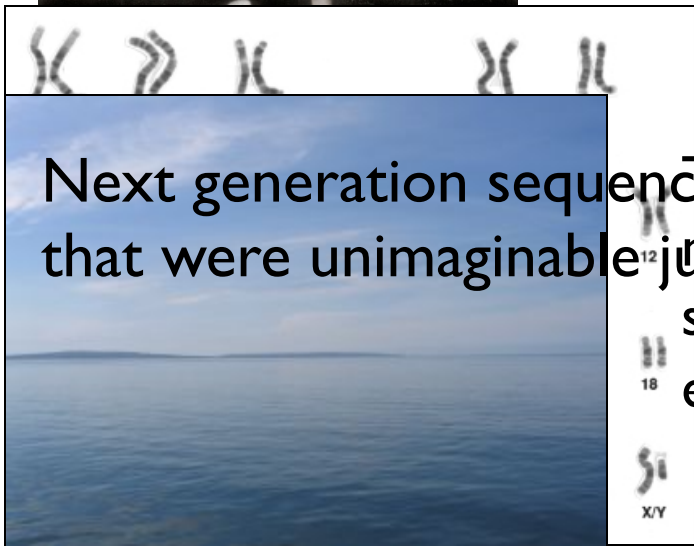
New platform for DNA sequence determination

- 100 - 400 bp for Roche systems
- 35 – 75 bp for Illumina systems
- 25 – 35 bp for ABI systems
- Much shorter read length, although read length is improving.
- High coverage of the original sequence

# Whole genome sequencing



- Landmarks in whole genome sequencing:
  - 1970's : First genome sequenced by Frederick Sanger
  - 1995 : First free living organism sequenced by J. Craig Venter
  - 2003: Human genome project is completed



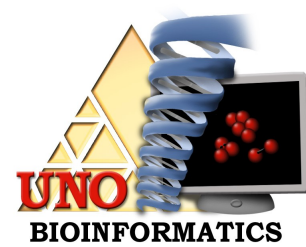
Next generation sequencing has advanced to the point that what was unimaginable just a decade ago is now routine.

## Enhanced Personalized Medicine

This strains of HIV are now aggressively being tested by specifically tailored treatments and sequences in the global ocean sampling expedition (Craig Venter Institute) side affects.



# Next generation sequencing



Since its inception in the mid 2000's, next generation sequencing has produced massive amounts of genetic information.

-Massively parallel sequencing has become the cornerstone of many diverse research endeavors.

- Personalized medicine
- SNP association studies
- Cancer research
- Metagenomics

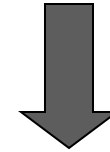
Next generation sequencing has allowed us to do many things that were unimaginable just a decade ago!

# Next generation sequencing

## DNA Fragments

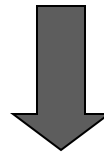
GGATCCATGGATAGGATAATGGA  
GATAATGGATAGAGGATCCATGG  
ATGGATAGAGGATCCATGGCTAG  
ATGGATAGGATTATGGATAGAGG  
GATAGAGGATCCATGGCTAGATC

**Overlapping**



GGATCCATGGATAGGATAATGGA  
                                  GATAATGGATAGAGGATCCATGG  
  ATGGATAGAGGATCCATGGCTAG  
  ATGGATAGGATTATGGATAGAGG  
  GATAGAGGATCCATGGCTAGATC

**Contigs**



GGATCCATGGATAGGATAATGGATAGAGGATCCATGGCTAGATC

# Next generation sequencing

The characteristics of sequencing fragments depend on:

## 1) Sequencing technology

- Illumina sequencing
- 454 sequencing

## 2) Domain characteristics

- GC content
- Repeat content

## 3) Project characteristics

- Sequencing depth
- Species abundance level

# Assembly challenges

## Errors and small overlaps:

- Is it a sequencing error?

ATTAGG**T**AGGTTTGAT  
TTACATTAGG**CCG**

- Is this a large enough fragment overlap?

ATTAGTTAGTTAGATTAC  
...GGCATT

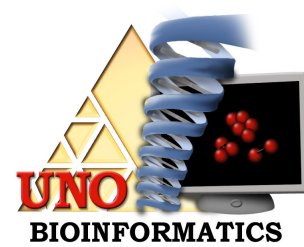
## Repetitive regions of the genome:

- Ambiguous fragments that lead to false overlaps

## Gaps:

- Some areas of the genome may not be able to be fully assembled.

# ICD Tool for Genome Assembly



## **Intelligent:**

Based on a learning model that gets better with additional data/information

## **Collaborative:**

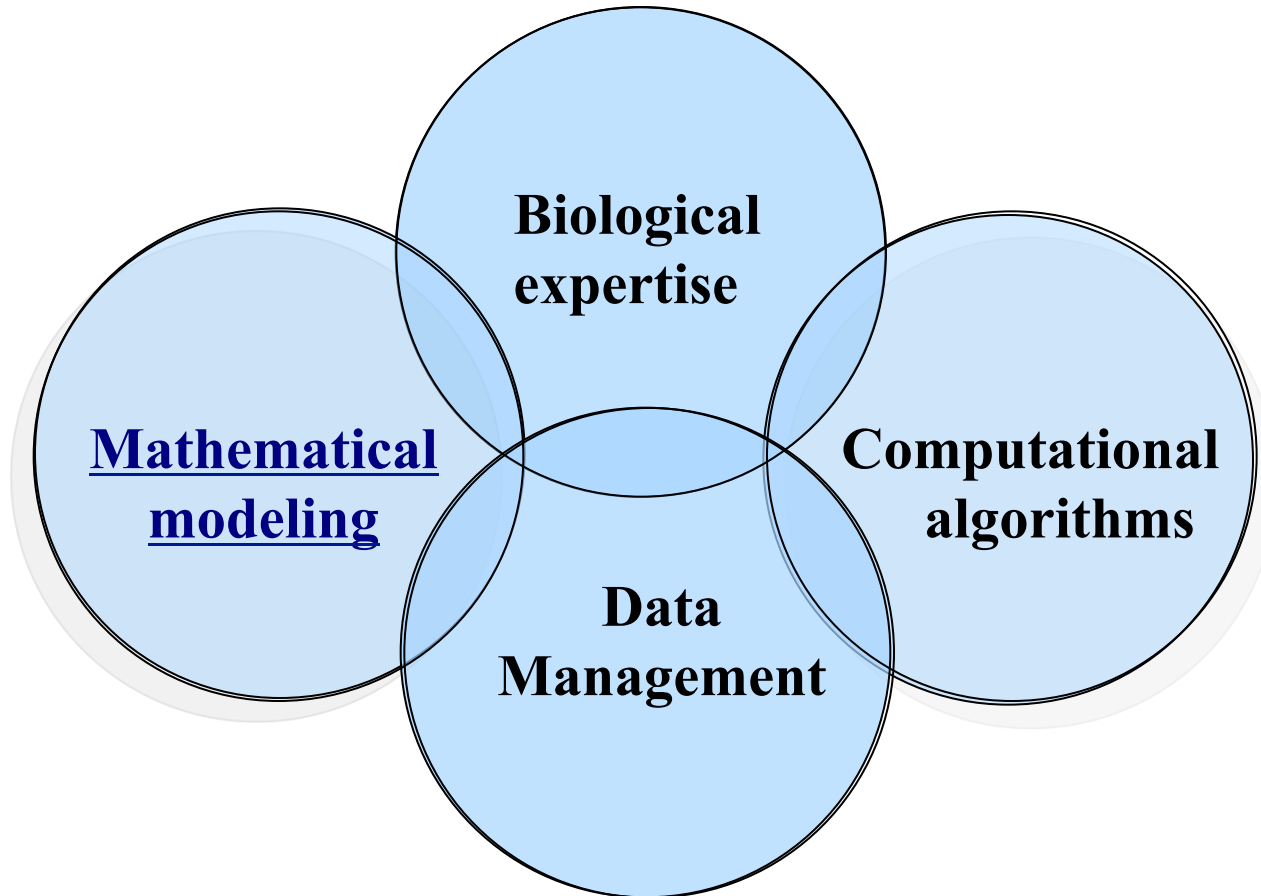
Integrate knowledge and expertise from many diverse research fields, particularly Biosciences and Computational sciences

## **Dynamic:**

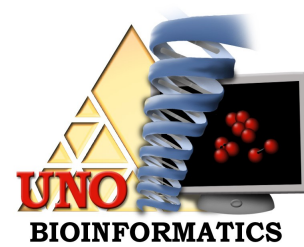
Custom built and domain specific – avoid one size fits all

# ICD Tool for Genome Assembly

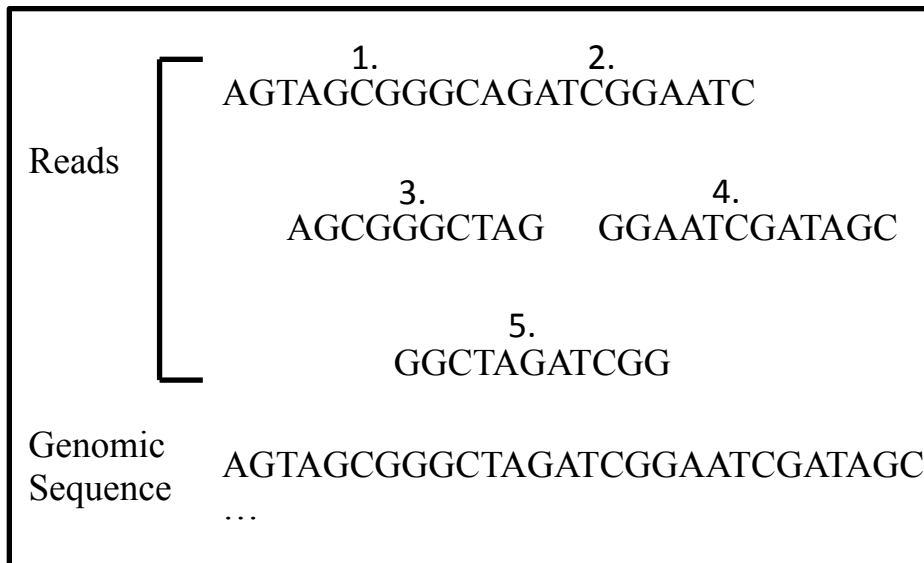
## Collaboration



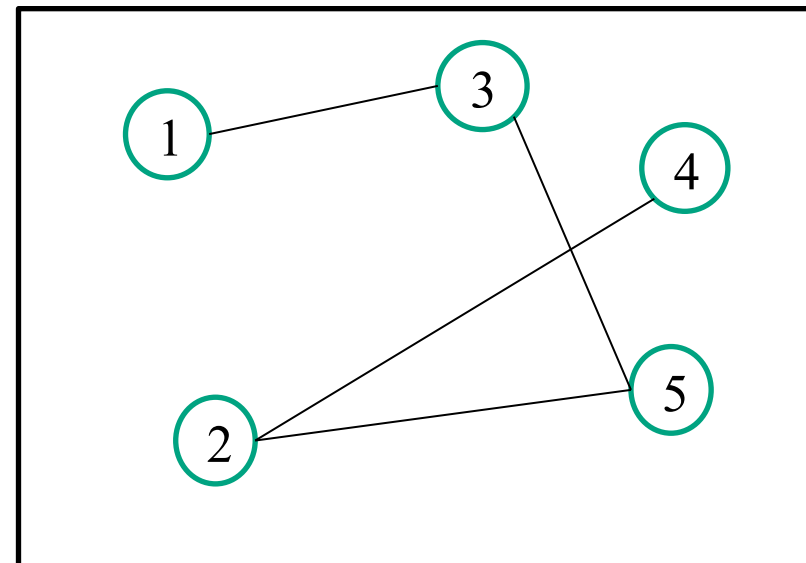
# A Graph Theoretic Model for Assembly



## Reads



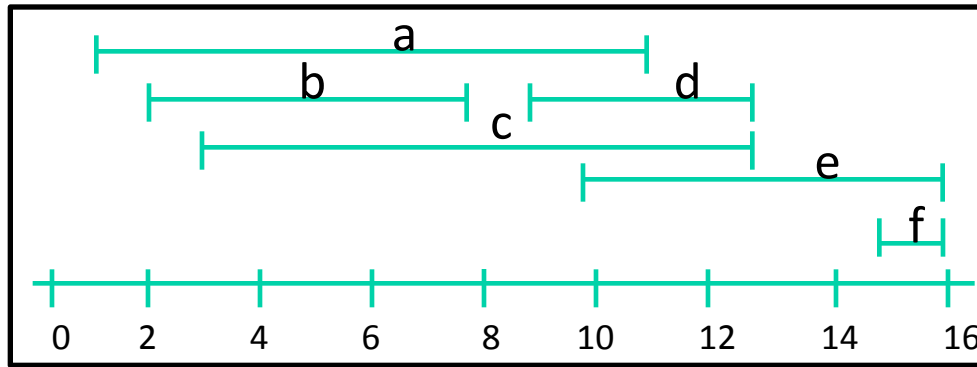
## Overlap Graph



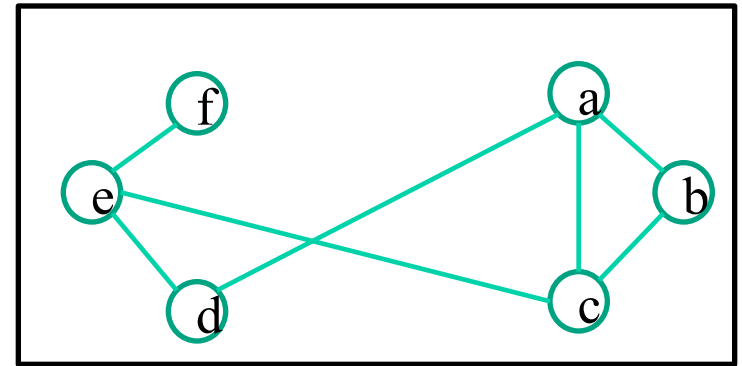
- The algorithm maps each read to a node (vertex) and every overlap relationship to edges.

# Sequencing and Interval Graphs

## Interval Representation



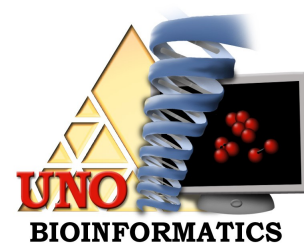
## Interval Graph



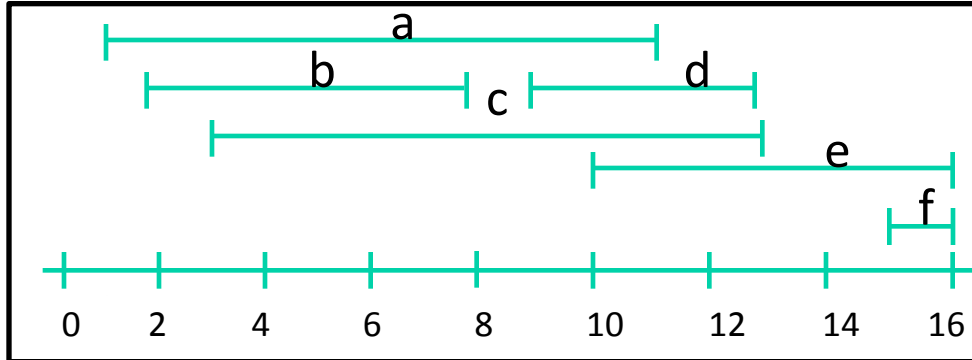
- In graph theory, there is a special class of graphs called perfect graphs.
  - Some NP hard problems are polynomial on perfect graphs
  - Known recognition algorithms for the subclasses of perfect graphs
- Ideally, the overlap graph should form an interval graph
- This is rarely the case due to:
  - Low quality or ambiguous overlaps
  - Sequencing errors
  - Repeat regions



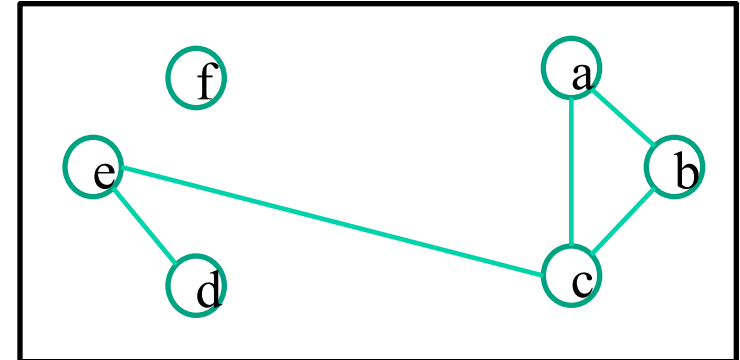
# Sequencing and Tolerance Graphs



Tolerance Representation,  $t_i = 2$

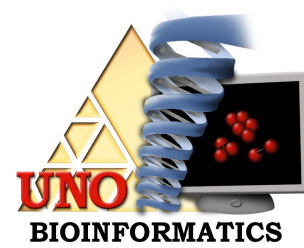


Tolerance Graph



- To address false-positive edges, a user-input parameter was established
  - This parameter specifies minimum overlap length
  - If reads overlap more than this minimum, an edge is added between the nodes representing the reads
- This added parameter shifts the graph model to a tolerance graph model
  - The tolerance graph is also a perfect graph
  - Known properties and recognition algorithms are established
  - Provides a well-defined foundation to build upon

# Two step algorithm



## **Node Merging:**

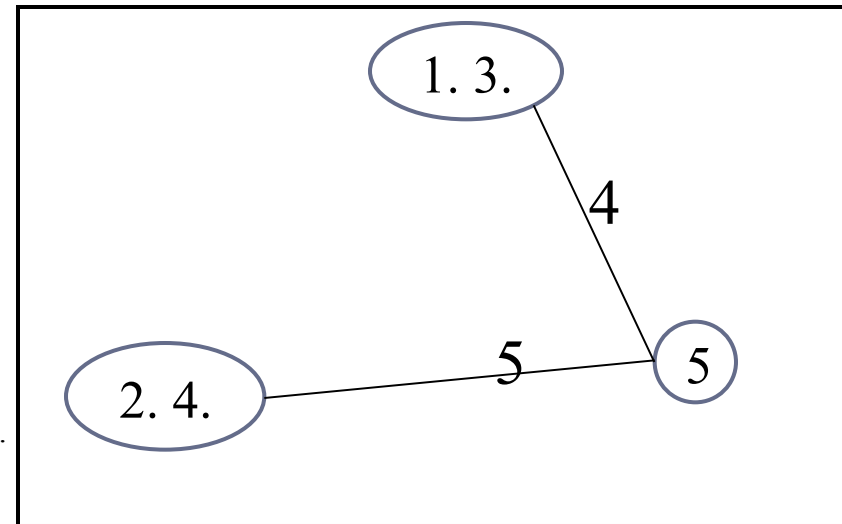
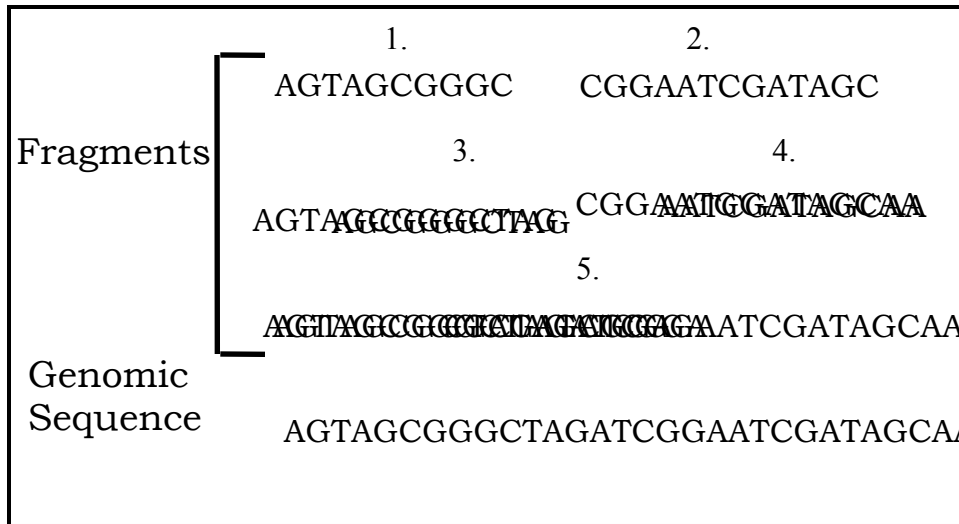
- Longer overlaps shared between reads increases the confidence that they are consecutive in the genome
- Reads with high-confidence overlap relationships are merged into “super-nodes” in the overlap graph
- Reduces graph complexity and ambiguity

## **Graph Traversal:**

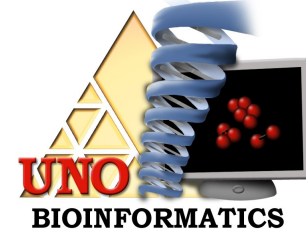
- Contigs are constructed by proper traversal of the overlap graph
- An Euler path algorithm is used to traverse the overlap graph
- The sequences represented by the nodes are merged in the order of the Euler path to produce contigs

# Assembly method

## Assembly through fragment merging and graph traversal



# Assembly Algorithm



Read\_File.txt

```
TGGAC
ACCAA
AACTG
ATCAA
TTCAA
GTTCA
TCAAT
```

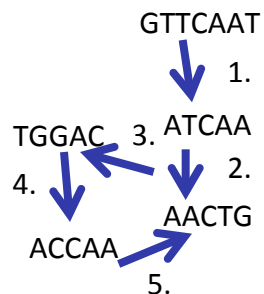
i) The algorithm accepts a text file containing short read sequences; A minimum overlap size for node merging; A minimum overlap size for edges. From this information, the algorithm creates an overlap graph.

ii) If two reads, have a high degree of overlap and all of their neighbors overlap, the more confidence there is that they are consecutive in the genome. High confidence reads are merged to reduce the overlap graph's size.

## Node Merging

```
TTCAA
GTTCA
TCAAT
↓
GTTCAAT
```

## Graph Traversal



iii) For the purpose of constructing contigs from the read data, the algorithm attempts to find an Euler path in the overlap graph. An Euler path is labeled on the graph to the left.

iv) The sequences represented by each node are merged according to the order of the Euler path to produce contigs.

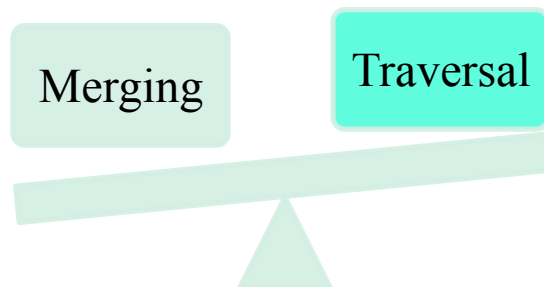
## Final Output

```
>contig 1
GTTCAATCAACTGGACCAACTG
```

# ICD Assembler Tool

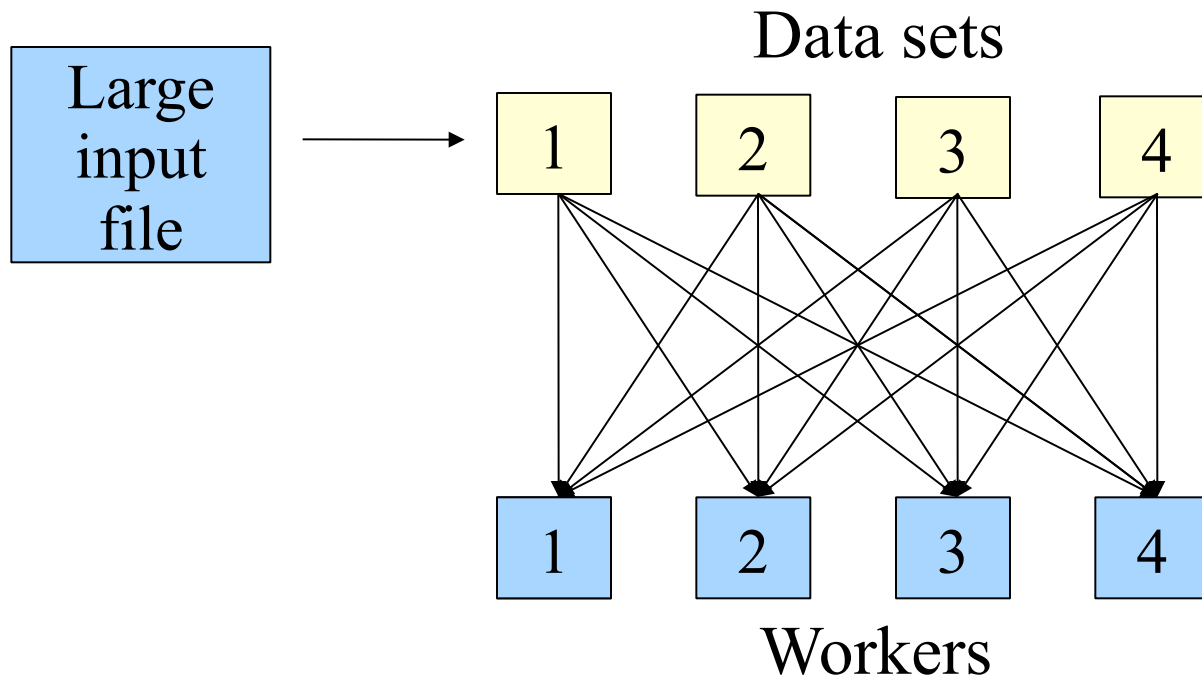
- The greedy node merging process may merge reads that have false-positive relationships
- A pure graph traversal approach may not be feasible in a highly complex and large overlap graph
- The proportions of graph traversal and node merging can be adjusted by changing the stringency of the merging parameter

What balance of graph traversal and node merging will produce the best assembly?

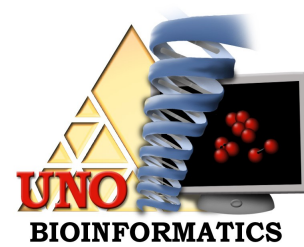


# Parallel computing

- Finding overlaps in a large dataset of fragments can be a time consuming task
- Solution: Parallel computing



# Customize Assembler Parameters



## **Minimum overlap length:**

- If reads overlap more than this minimum, an edge is added between the nodes representing the reads
- Reduces false positive edges

## **Minimum overlap length for merging:**

- If reads overlap more than this minimum, they are merged into super nodes
- Reduces graph complexity

## **Iterative Steps:**

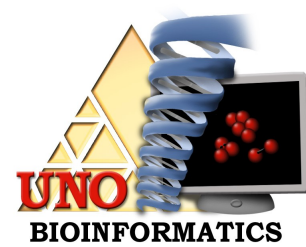
- Number of iterations and how much work in each one

# ICD Assembly Model

- Node Merging
  - Longer overlaps shared between reads increases the confidence that they are consecutive in the genome
  - Reads with high-confidence overlap relationships are merged into “super-nodes” in the overlap graph
- Dynamic Approach
  - Should all the reads/contigs have the same tolerance?
  - Should we also consider the global vs local graph properties?
  - How to incorporate known knowledge?



# Modeling Read Overlaps

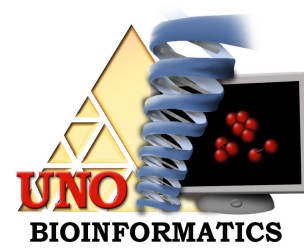


- Overlap Tolerances
- False-positive overlaps
  - Repeats
  - Sequencing errors
  - Random alignments
- Overlap Tolerance
  - Threshold assigned to each read
  - Minimum overlap length

# Modeling Reads Overlaps

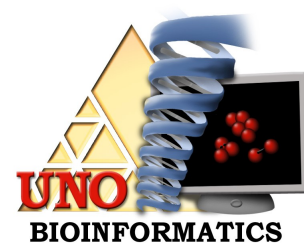
- One size fits all, static tolerance thresholds
  - Inflexible
  - Not data-centric
- Dynamic data
  - Read lengths
  - Genome repeats
- Dynamic overlap tolerance threshold adjustment
  - Flexible
  - Individualized
  - Incorporates dataset specific knowledge

# Dynamic Tolerance threshold adjustment



- Knowledge driven threshold fine-tuning
  - Graph properties
  - Read properties
  - Domain specific knowledge
- Graph properties
  - Node degree: How many neighbors
- Read properties
  - Read length
- Domain specific knowledge
  - Minimum overlap length

# Dynamic Tolerance Adjustments



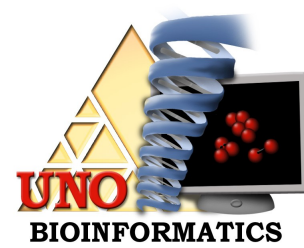
Normalizing graph and read information

$$Zscore\_degree(u) = \frac{(\text{node\_degree}(u) - \text{average\_node\_degree})}{\text{node\_degree\_standard\_deviation}}$$

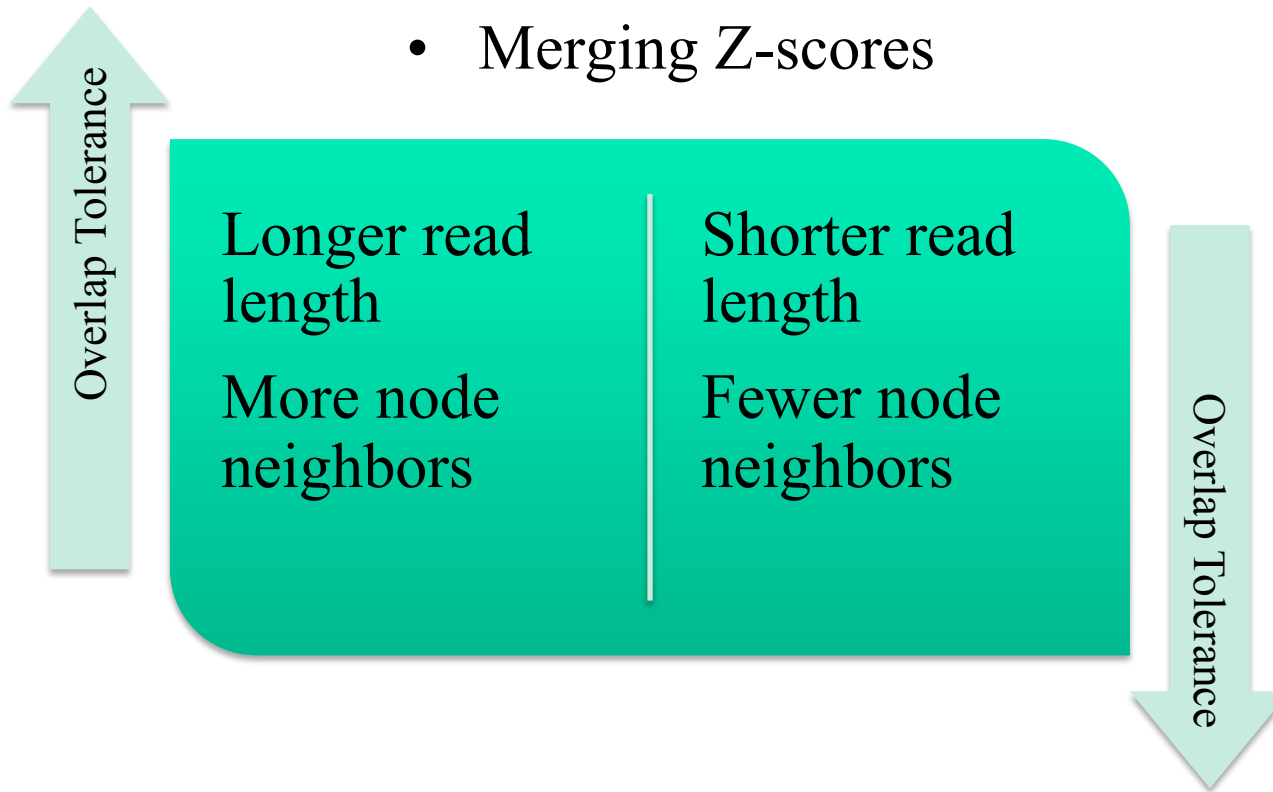
$$Zscore\_read\_length(u) = \frac{(\text{read\_length}(u) - \text{average\_read\_length})}{\text{read\_length\_standard\_distribution}}$$

$$Zscore\_minoverlap(u) = \frac{(\text{minoverlap\_length} - \text{average\_overlap\_length})}{\text{overlap\_length\_standard\_deviation}}$$

# Dynamic Tolerance threshold adjustment

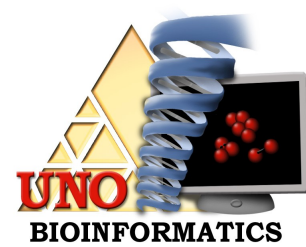


- Merging Z-scores



$$Tolerance(u) = a(Zscore\_degree(u)) + b(Zscore\_read\_length(u)) + c(Zscore\_minoverlap(u))$$

# Preliminary Results



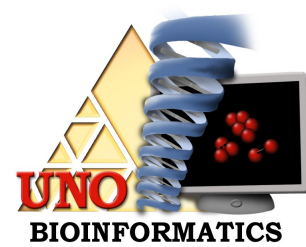
- How much influence should each parameter have?
- Escherichia coli
- 300 bp reads at 20x coverage
- Uncorrected N50 lengths

# Preliminary Results

- How much influence should each parameter have?

<b>N50 (bps)</b>	<b>Read Length Weight</b>	<b>Degree Weight</b>	<b>Domain Specific Weight</b>
<b>12978</b>	<b>0.1</b>	<b>0.1</b>	<b>0.8</b>
<b>13677</b>	<b>0.1</b>	<b>0.3</b>	<b>0.6</b>
<b>13406</b>	<b>0.1</b>	<b>0.5</b>	<b>0.4</b>
<b>12549</b>	<b>0.3</b>	<b>0.1</b>	<b>0.6</b>
<b>16761</b>	<b>0.3</b>	<b>0.3</b>	<b>0.4</b>
<b>16464</b>	<b>0.3</b>	<b>0.5</b>	<b>0.2</b>
<b>12095</b>	<b>0.5</b>	<b>0.1</b>	<b>0.4</b>
<b>15988</b>	<b>0.5</b>	<b>0.3</b>	<b>0.2</b>
<b>15427</b>	<b>0.5</b>	<b>0.5</b>	<b>0</b>

# Assembly of HIV 454 data



Nebraska Center for Virology

Ten HIV virus data sets (pol gene) 1,200 bp region

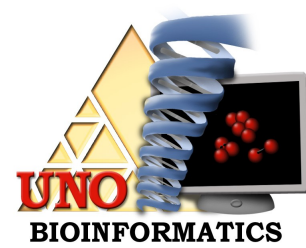
3000 – 4000 Fragments per dataset

200 bp reads

Genome wide association and drug resistance



# Assembly assessment



## **Max contig:**

- The longest stretch of contiguous sequence produced by each assembler

## **N50:**

- A statistical measure of average length of a set of contigs.

## **Number of contigs produced**

## **Coverage:**

- The percentage of the reference sequence that was assembled

## **Identity:**

- The percentage of bases in the assembly that are identical to the reference sequence

# Assembly Testing

## Data set 1:

- 79,237 bp *Escherichia coli* plasmid [NC\_009786.1]
- 45x coverage
- 36 bp reads with 0% error rate
- Final results were validated with the MUMer package

## Results are compared to Velvet:

- Complete assembler

## Tested range:

- Node merging
- Graph traversal

# Assembler Testing

Sequence Source	Length (bp)	GC Content	Repeat Content
Escherichia coli	79237 bp	47.27%	0.09%
Drosophila melanogaster	79745 bp	42.47 %	3.21 %
Arabidopsis thaliana	79590 bp	37.71 %	9.75 %
Homo sapiens	80914 bp	43.08 %	49.35 %

Table 1. Table showing the sequence information for the four datasets used to test the algorithm. RepeatMasker was used to determine the repeat content of each sequence

- 30 bp reads
- 80, 000 reads in each dataset
- Error free reads
- 30x coverage

# Testing Results

## *Escherichia coli* Assembly

Mostly Traversal Mostly Merging

Merging Parameter (bps)	# of Contigs > 100 bp	N50 (bps)	Expected Sequence Identity %	Max Length (bps)	Est. Coverage
15	93	1993	98.868%	7461	94.6%
17	76	3784	99.383%	9137	95.4%
19	64	4405	99.762%	9055	94.7%
21	69	3786	99.856%	9063	96.4%
23	84	1741	99.881%	6183	96.0%
25	241	453	99.886%	1578	95.7%
27	94	9763	99.441%	20438	61.8%
29	15	9768	98.384%	13475	42.1%
<b>Velvet</b>	47	4513	99.993%	9990	88.2%

# Summary

- The assembly of short read sequences is still very difficult. To provide a strong basic platform to build upon, we have introduced a new graph theoretic approach for the assembly of short read sequences.
- The use of a tolerance graph model and an algorithm consisting of greedy node merging and graph traversal is novel from previous approaches.
- The proportions of merging and traversal affect the quality of the assembly.
- The Proposed method, while still a basic platform, produced results that are comparable to the Velvet assembler.

# ICD Identification and Classification Tool using Compression- Based Sequence Comparison

# Sequence Comparison

- Sequence comparison is one of the central and well studied problem in Bioinformatics
- Biology has a long tradition of comparative analysis leading to discovery
- The number of sequences available for comparison has been growing explosively
- Efficient algorithms already exist for solving many sequence comparison related problems.
- Sequence Comparison is used for identification, classification, structure, and function related problems

# Problem Definition

- Biological sequences
  - DNA sequences, base of ACTG or ACUG
    - ACTGAGGGTAAG
  - Protein sequences, base of 20 amino acids
    - MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYD
    - Protein sequences are generated from DNA sequences.
- Comparing different sequences:
  - Identify similar structures
  - Identify similar functions
  - Identify evolutionary relationships



# Sequence Alignment

- Goal: To enable researchers to determine whether two sequences display sufficient similarity to justify the inference of homology.
- Definition: Given two sequences of sizes  $m$  and  $n$ , an alignment is the insertion of spaces in arbitrary locations along the sequences so that they end up with the same size. Possible restriction: No space in one sequence is aligned with a space in the other.

# Alignments

- Which alignment is best?

```

A - C - G G - A C T
|   |   |           | |
A T C G G A T _ C T

```

```

A T C G G A T C T
|   | | |           | |
A - C G G - A C T

```

# Pluses and Minuses with Alignment

- Extremely viable way to compare biological sequences
- Based on a solid computational technique and has an acceptable biological model
- The granularity is too fine, every position counts
- All positions are treated equally and no room for incorporating evolutionary clocks
- Can we integrate it with alignment free methods?

# Problems with Alignment

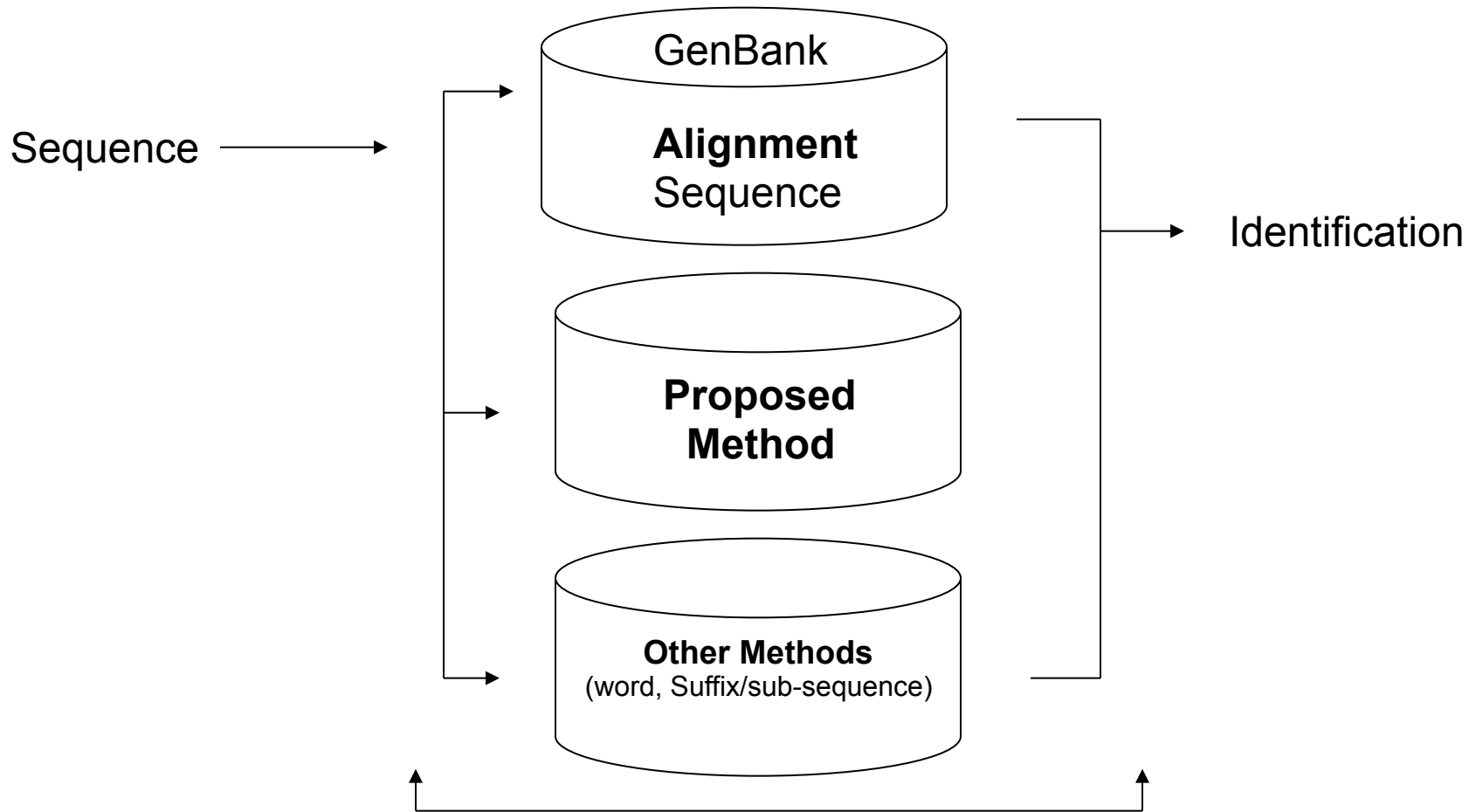
- Optimal alignment is expensive  $O(n^2)$
- Alignment based approach use a very fine grain perspective which may not be suitable for all applications
- Fails to compare long sequences, easy to fool by repetitions and translocations
- It is independent of the input domain
- Inaccurate with incomplete genomes

Alignment-free methods?

# One Method is not Enough

- Would different objectives of sequence comparison demand different comparison approaches
- Recently, alignment free methods have been approached:
  - Data Compression based approaches
  - Motifs based approaches
  - Statistics based approaches

# The Integrated Approach



Integrated Advanced Identification System  
(Development of Algorithm to make route-decision)

# ICD Compression Based Techniques

- Each sequence is scanned and linearly independent strings are obtained and form a dictionary
- Weighted differences among dictionaries reflects dissimilarity among input sequences
- Repetitions and translocation don't impact the dictionaries as compared to alignment
- For any two sequences  $x$  and  $y$ , we need
  - $C(x)$ ,  $C(y)$ ,  $C(xy)$  and  $C(yx)$ .

# Example

$S = AACGTTACCATTG \quad R = CTAGGGACTTAT$

$Q = ACGGTCACCAA$

$H_E(S) = A/AC/G/T/ACC/AT/TG$

$H_E(R) = C/T/A/G/GGA/CTT/AT$

$H_E(Q) = A/C/G/GT/CA/CC/AA$

•  $H_E(SQ) = A/AC/G/T/ACC/AT/TG/ACGG/TC/ACCAA$

•  $H_E(RQ) = C/T/A/G/GGA/CTT/AT/ACG/GT/CA/CC/AA$

$c(SQ) - c(S) = 3$

$c(RQ) - c(R) = 5$

→  $Q$  is “closer” to  $S$  than  $R$

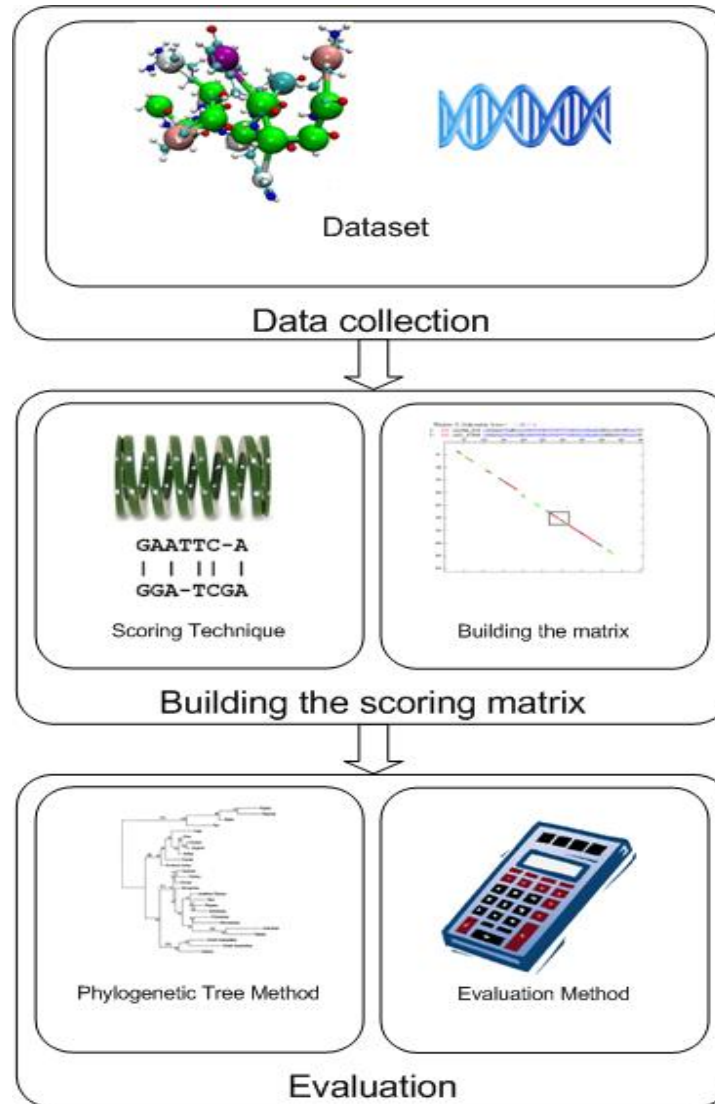
$Distance(S, Q) = c(SQ) - c(S) + c(QS) - c(Q)$



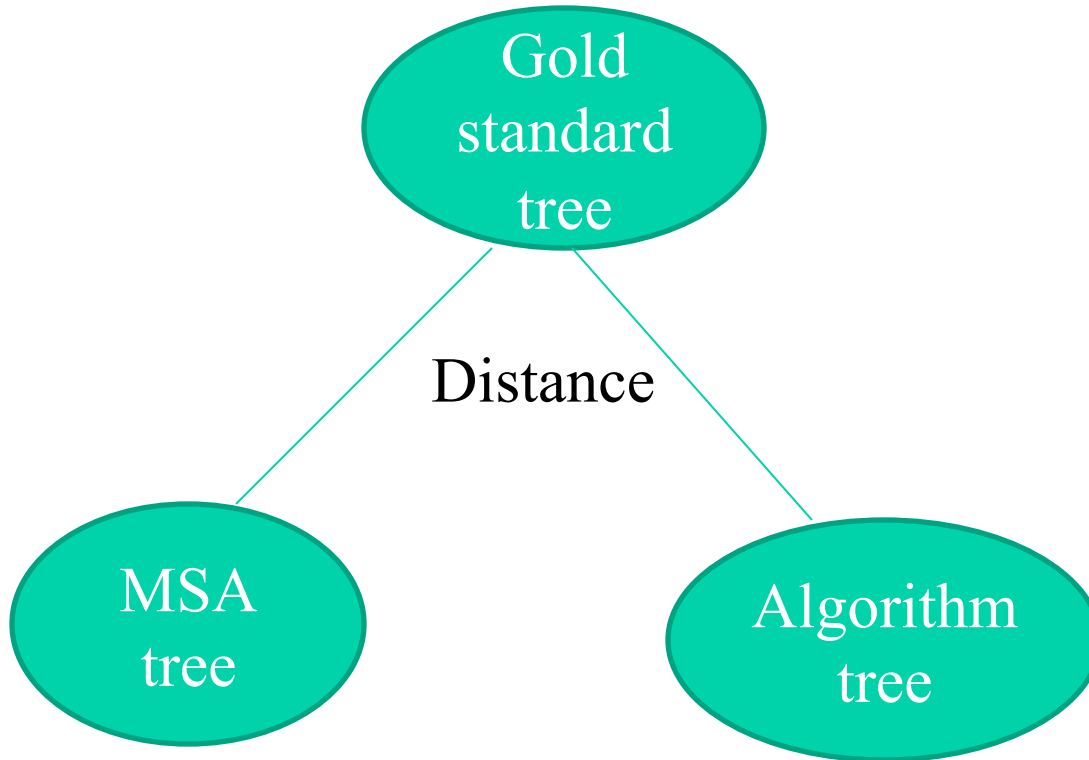
# Compression-Based Methods

- We used
  - Kolomogrov complexity (3 distance measures)
  - Lempel-Ziv complexity (4 distance measures)
  - Clustering
    - UPGMA
    - NJ
  - Gold standard tree
  - Path-length difference

# Evaluation and Assessment



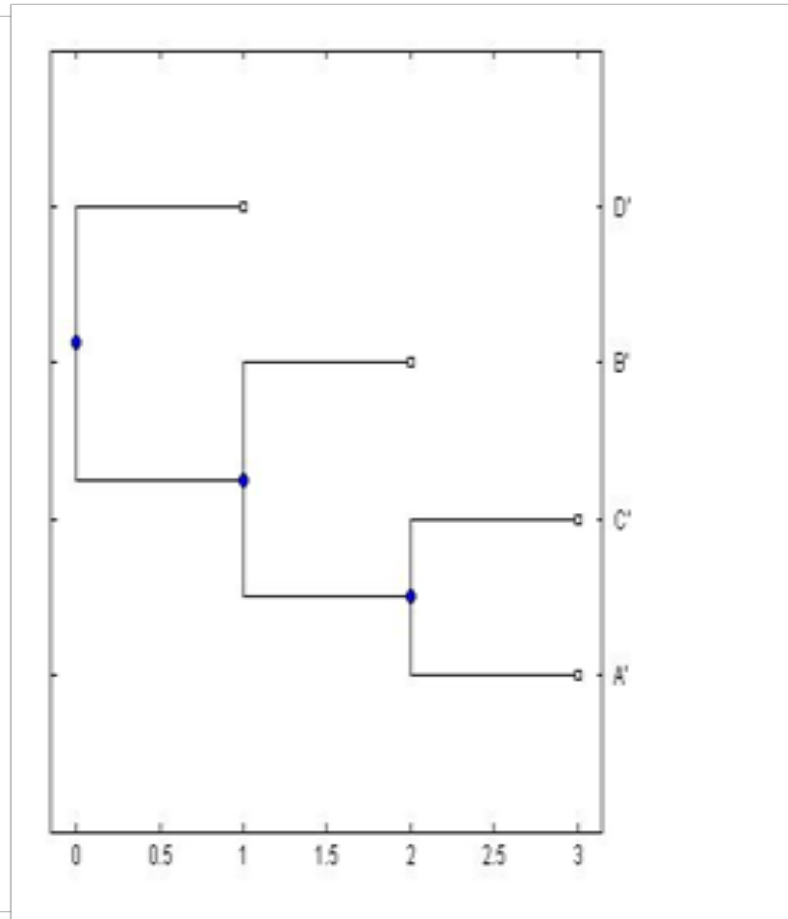
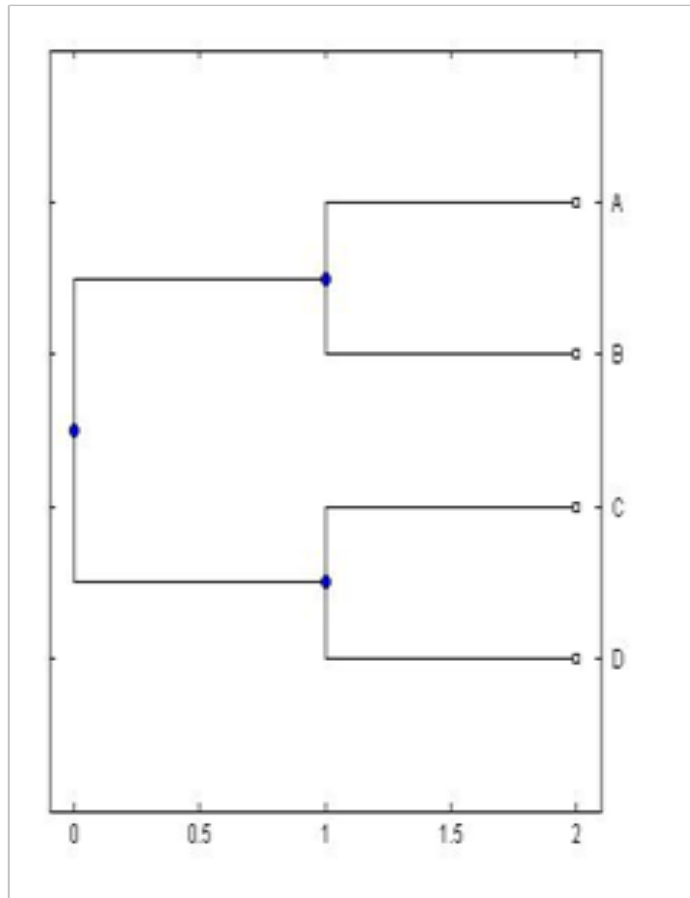
# Evaluation of Comparison Approaches



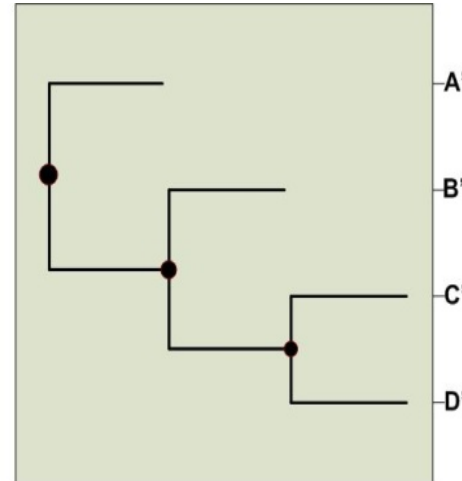
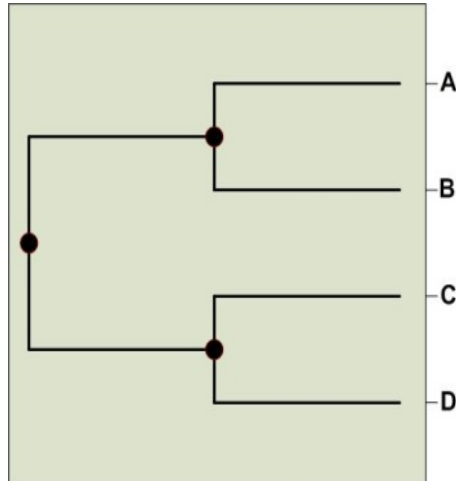
The distance between trees

- Problems with visual inspection!
- Computational ways?
  - Accurate
  - Fast

# Path-Difference Length



# Distance Between Trees



	A	B	C	D
A	0	2	4	4
B	2	0	4	4
C	4	4	0	2
D	4	4	2	0

	A'	B'	C'	D'
A'	0	3	4	4
B'	3	0	3	3
C'	2	3	0	2
D'	4	3	4	0

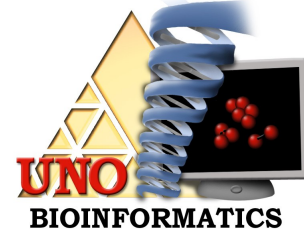
# Results



## Comparisons of the compression algorithms and multiple sequence alignment for the protein dataset CK-36-PDB

		Protein dataset <u>CK-36-PDB</u>	
Test Algorithm	Variant	Neighbor-Joining	UPGMA
Kolmogorov using Huffman coding	CD	2.395244	3.169468
	NCD	2.328382	2.264505
	UCD	2.328382	2.264505
Kolmogorov using LZW compression	CD	2.176959	2.165911
	NCD	2.210704	2.215544
	UCD	2.305268	2.238781
Lempel - Ziv complexity	Distance 1	2.337454	2.26598
	Distance 2	2.248862	2.192803
	Distance 3	2.244591	2.284809
	Distance 4	2.222918	2.371806
Multiple Sequence Alignment		2.182934	2.371806

# Comparisons of the compression algorithms and multiple sequence alignment for the Mitochondrial genome dataset



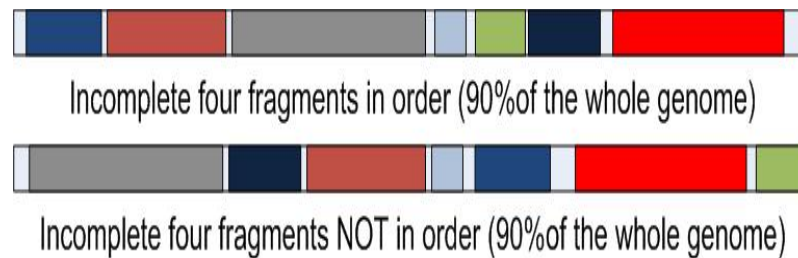
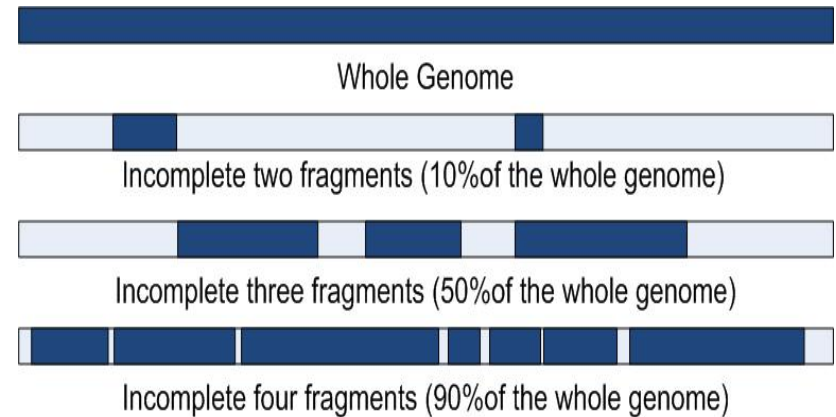
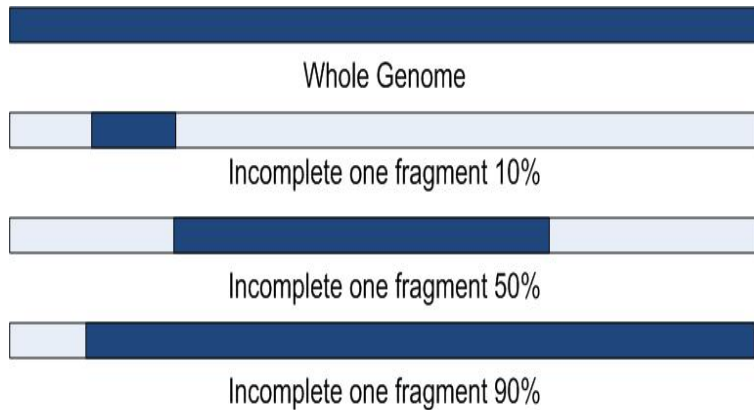
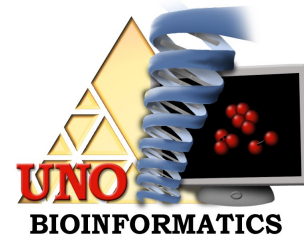
		Mitochondrial Genome dataset <u>AA-15-DNA</u>	
Test Algorithm	Variant	Neighbor-Joining	UPGMA
Kolmogorov using Huffman coding	CD	7.871585	7.871585
	NCD	7.871582	7.871582
	UCD	7.871582	7.871582
Kolmogorov using LZW compression	CD	3.034474	3.034474
	NCD	2.797647	2.797647
	UCD	2.878755	2.878755
Lempel Ziv complexity	Distance 1	1.357058	1.357058
	Distance 2	1.357058	1.357058
	Distance 3	1.357058	1.357058
	Distance4	1.357058	1.357058
Multiple Sequence Alignment		1.5547053	1.878762

# Comparisons of the compression algorithms and multiple sequence alignment for the Mitochondrial genomes

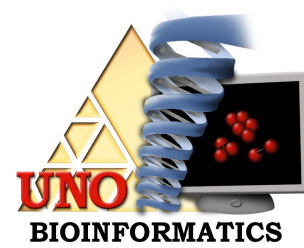
		Mitochondrial Genome dataset of sequences with high repetitions of subsequences.	
Test Algorithm	Variant	Neighbor-Joining	UPGMA
Kolmogorov using Huffman coding	CD	12.3431994	17.8482358
	NCD	12.3431994	17.8482337
	UCD	12.3431994	17.8482337
Kolmogorov using LZW compression	CD	10.5262895	10.5263158
	NCD	6.4460256	10.5263158
	UCD	9.8464668	10.5263158
Lempel Ziv complexity	Distance 1	0.0000000	0.0000000
	Distance 2	0.0000000	0.0000000
	Distance 3	0.0000000	0.0000000
	Distance4	0.0000000	0.0000000
Multiple Sequence Alignment		6.4460256	0.0000000



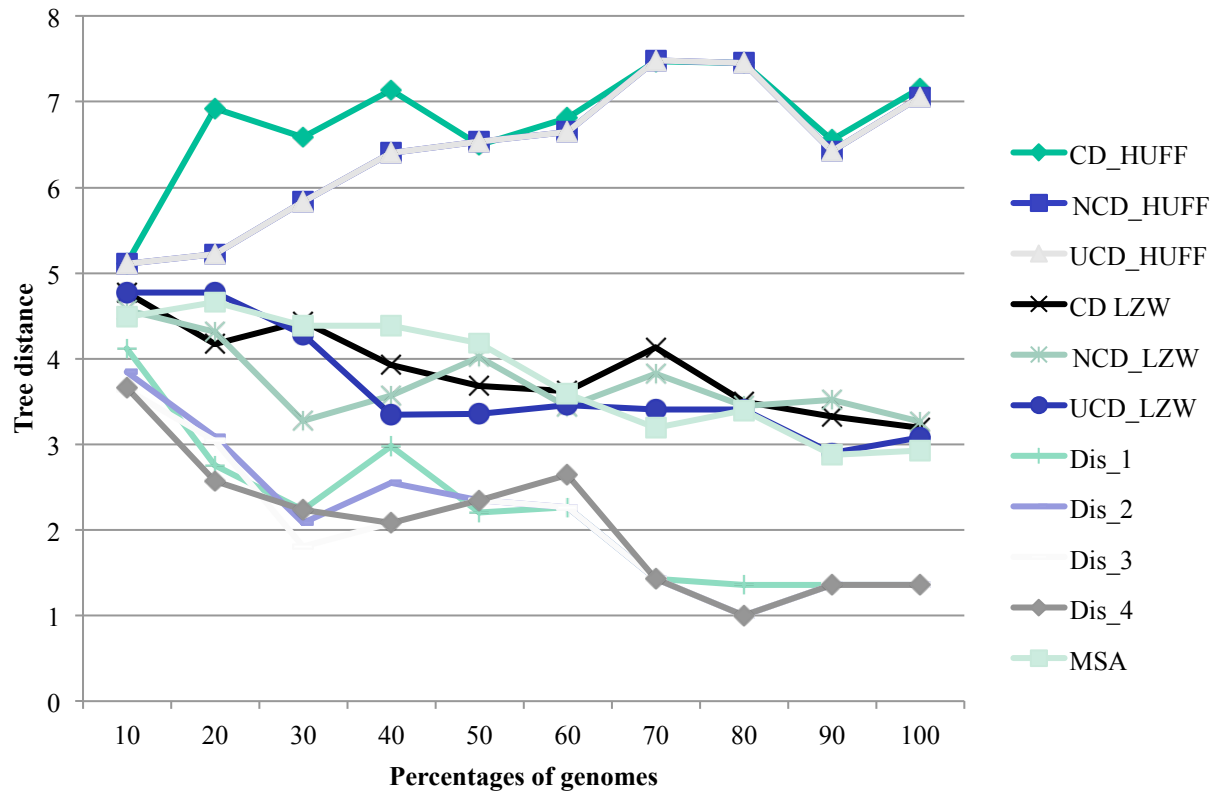
# Using Compression to Compare Incomplete Fragments of Genomes



# Results of Experiment

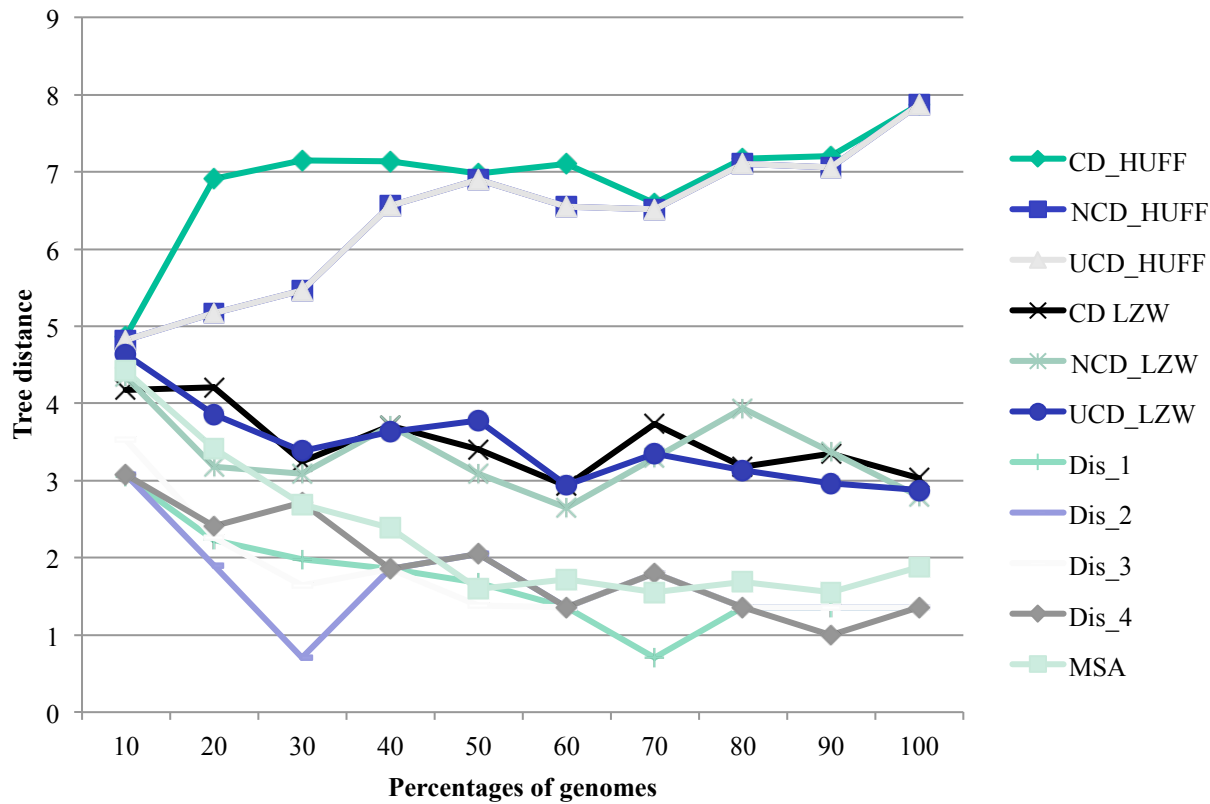


fragments of genomes, not continuous, not ordered



# Results of Experiment

## Fregments of genomes, not continous



# Analysis of Results

- Compression techniques are more like to cluster genomes with errors, as compression look at these data in a linear fashion rather than in a parallel fashion
- Multiple sequence alignment does not consider the input domain in obtaining similarity measures which limits its use for a diverse input set
- Compression methods identify important signals/ motifs in the input sequences and use them in the process

# Nebraska gets its very own organism

- ✚ While trying to pinpoint the cause of a lung infection in local cancer patients, they discovered a previously unknown micro-organism. And they've named it "mycobacterium nebraskense," after the Cornhusker state.
- ✚ It was discovered few weeks ago using Mycoalign: A Bioinformatics program developed at PKI



# Tutorial Outlines

- Introduction to Biomedical Informatics
  - State of the discipline - Challenges and Opportunities
  - Data-driven biomedical research
- Next Generation Bioinformatics Tools
  - Intelligent Collaborative Dynamic (ICD) Tools
- *Case Study: Aging Research*
  - The genomic study: Correlation Networks
  - Mobility and aging: Wireless monitoring
  - Data collection and Virtual Environments
- Next Steps: Where do we go from here?
  - HPC and Cloud Computing

# IT Aging Research Projects

- Bioinformatics – Correlation Networks
- Wireless Networks - IT for Assisted Living
- Public Health Informatics
- Data Collection and Virtual Environments

# Tutorial Outlines

- Introduction to Biomedical Informatics
  - State of the discipline - Challenges and Opportunities
  - Data-driven biomedical research
- Next Generation Bioinformatics Tools
  - Intelligent Collaborative Dynamic (ICD) Tools
- *Case Study: Aging Research*
  - *The genomic study: Correlation Networks*
  - Mobility and aging: Wireless monitoring
  - Data collection and Virtual Environments
- Next Steps: Where do we go from here?
  - HPC and Cloud Computing



# Motivation: Data Explosion

- E
- t

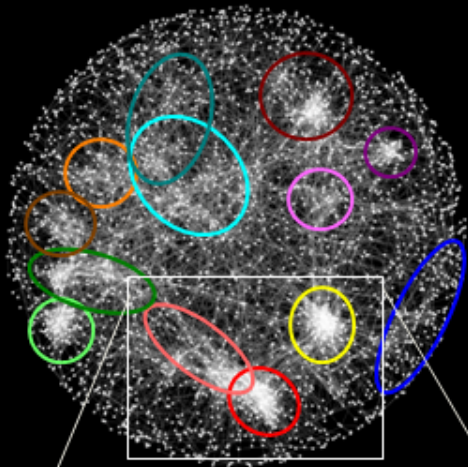
		
		
		

## Pathway Commons Quick Stats:

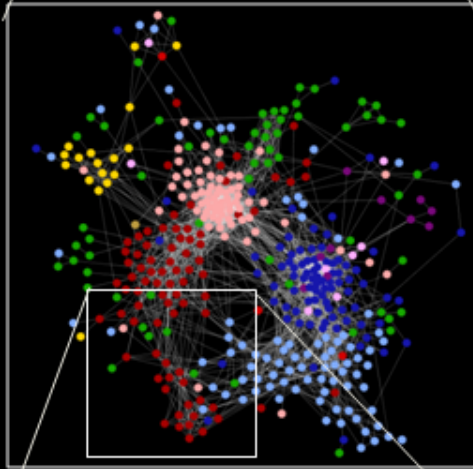
Number of Pathways:	1,623
Number of Interactions:	585,237
Number of Physical Entities:	105,949
Number of Organisms:	564

# Motifs

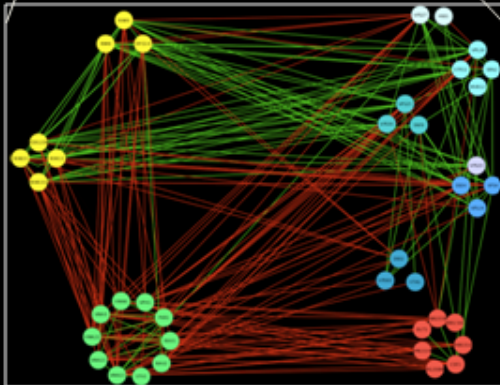
- *Provi*
- *Target*
- *Re-de*



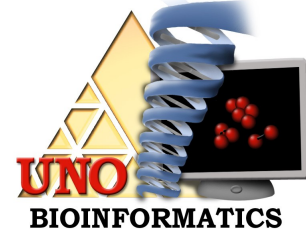
Global level



Process level



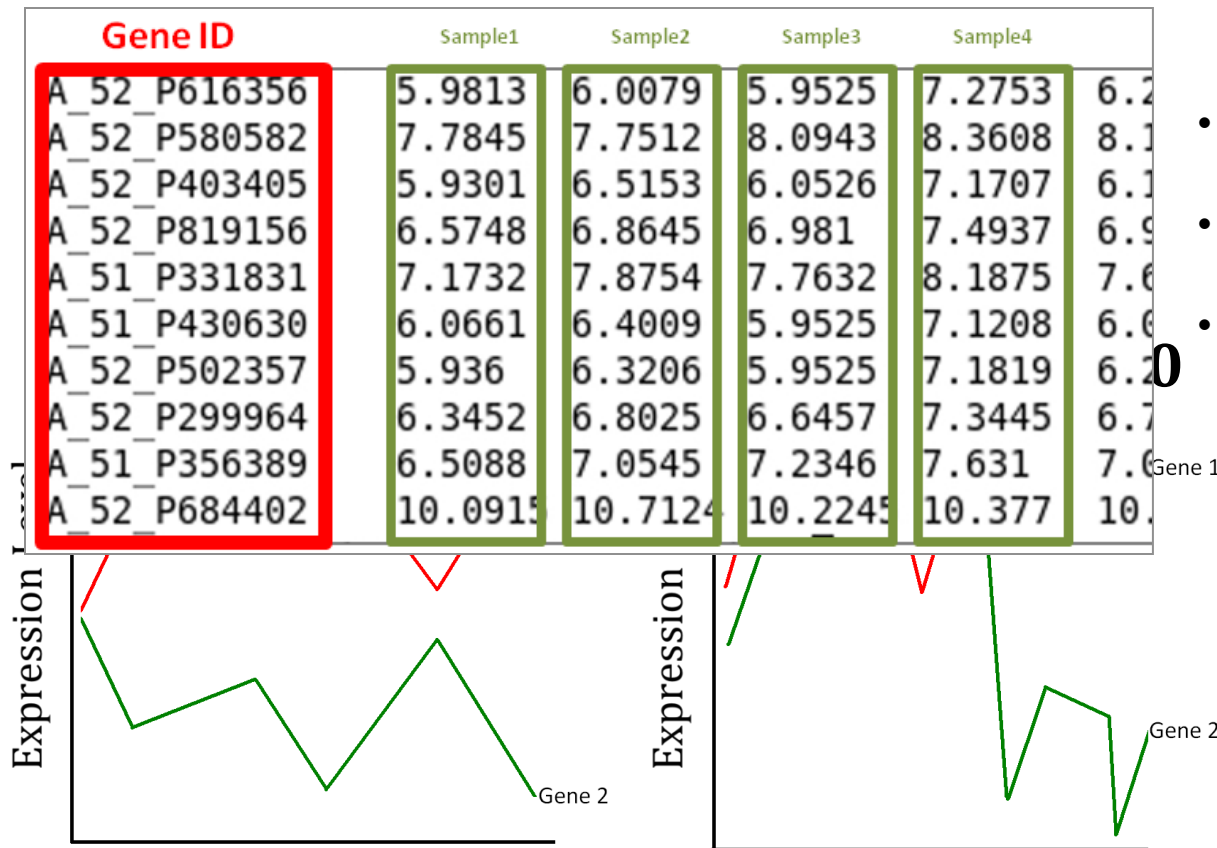
Pathway/complex level



MS

ing

# Correlation Networks

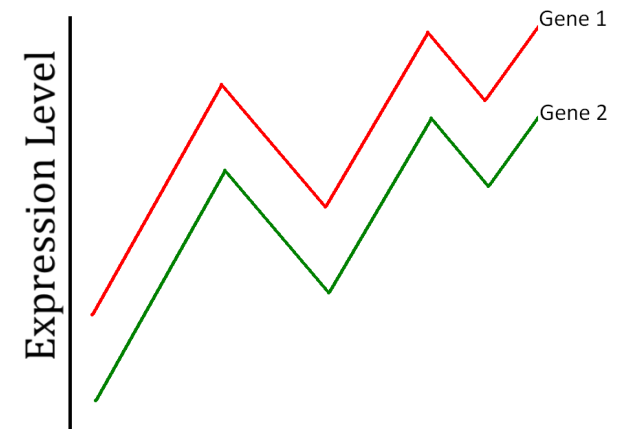


- 10,000-45,000+ probes

- UNO Blackforest cluster

- HCC Firefly

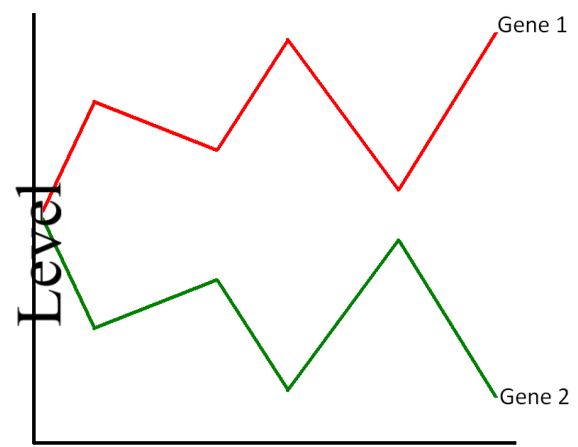
**Correlation = 1**



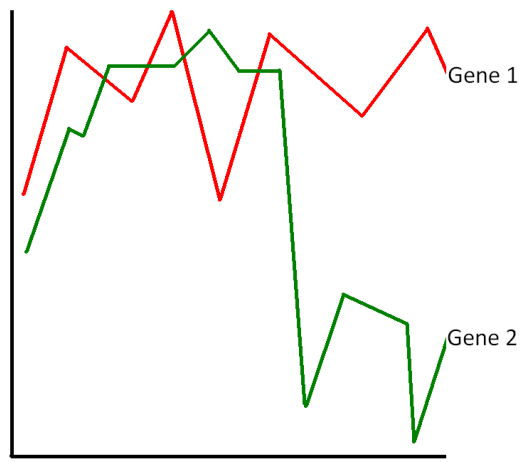
Sample

# Correlation Network

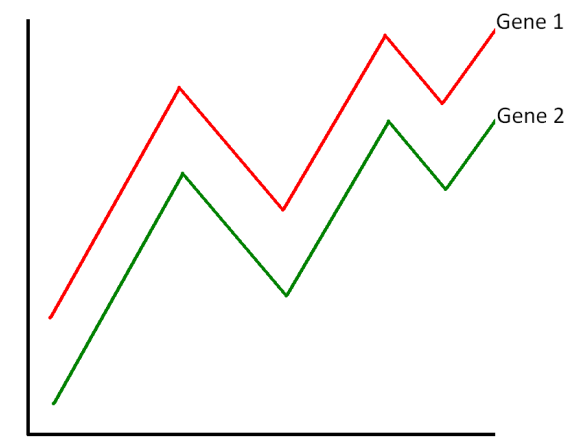
**Correlation = -1**



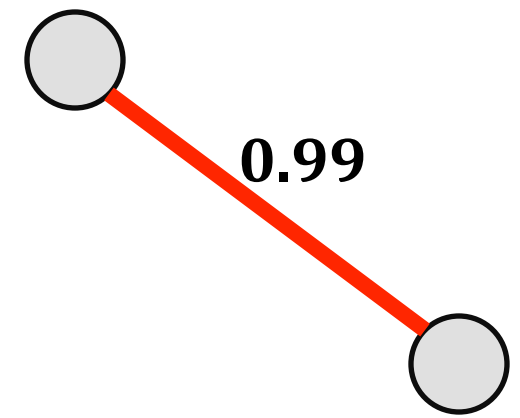
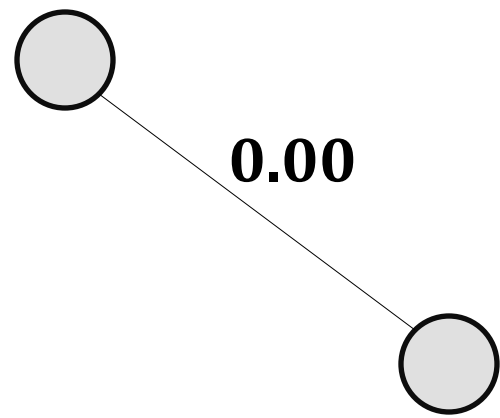
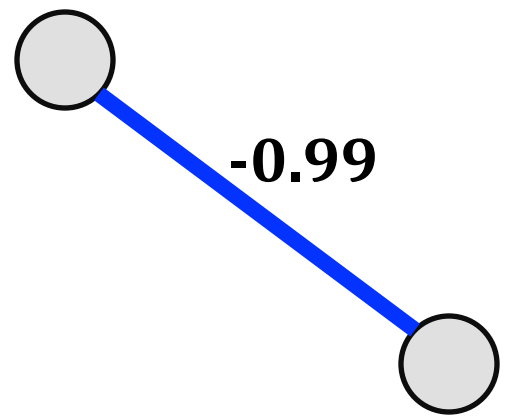
**Correlation = 0**



**Correlation = 1**

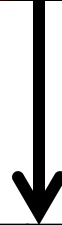
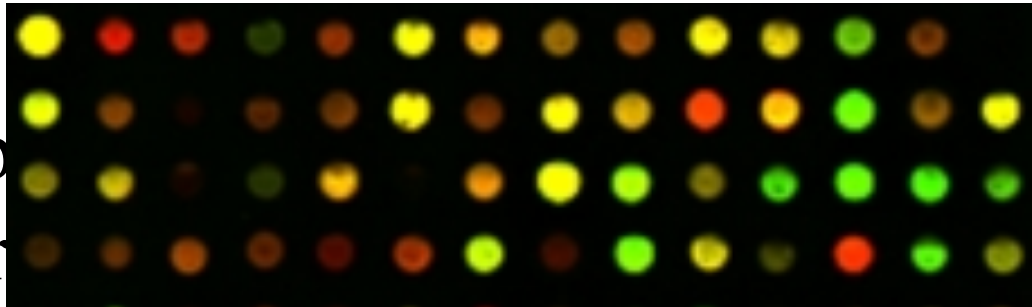


Sample

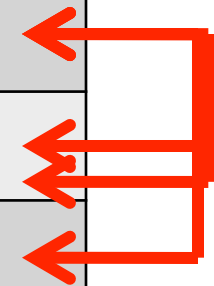


# Correlation Networks

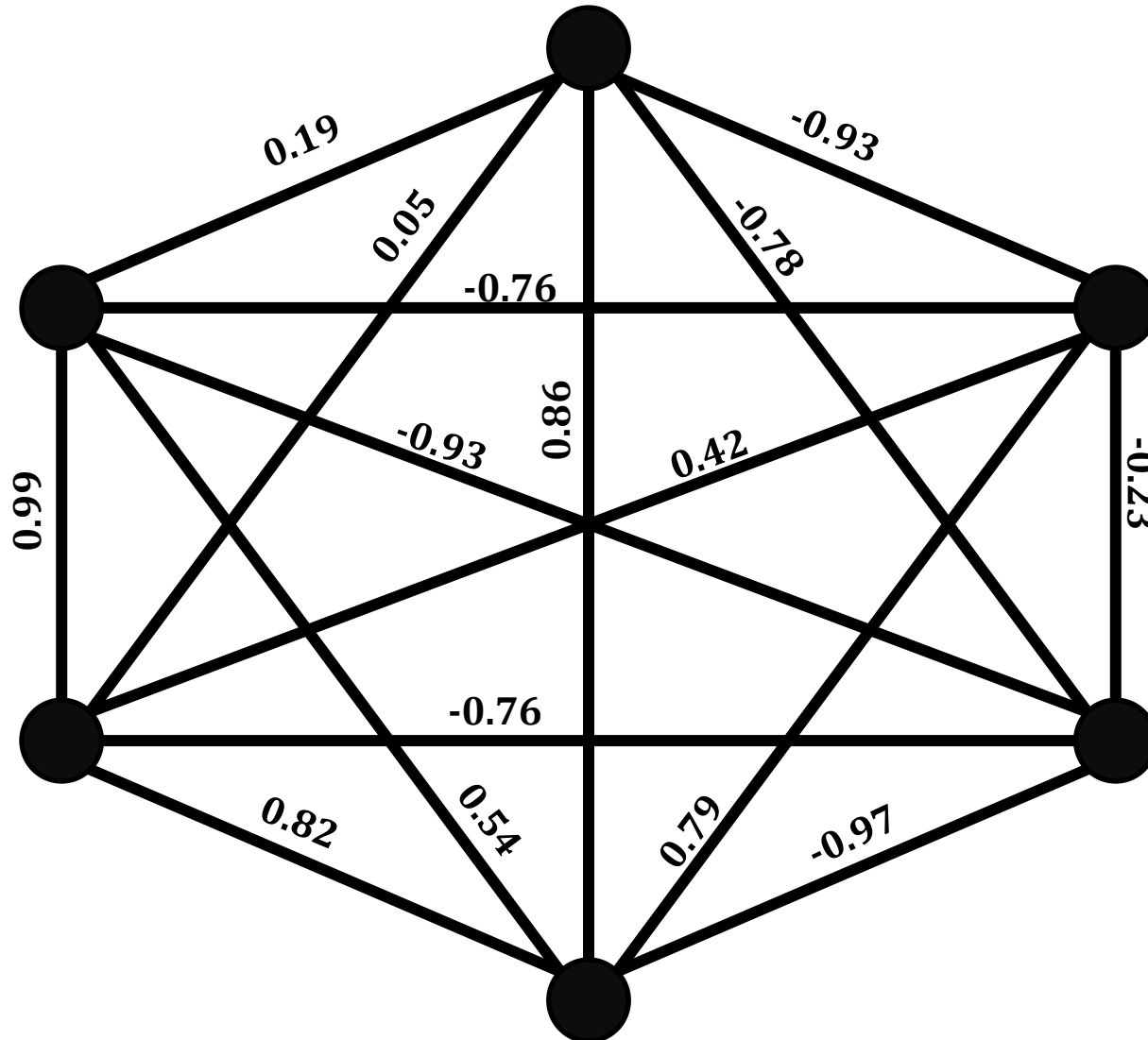
A graph of correlation coefficients between genes of different samples



	Sample 1	Sample 2	Sample 3
Gene 1	10.5	11.0	12.1
Gene 2	3.2	3.3	2.9
Gene 3	1.4	1.5	0.9
Gene 4	7.8	7.1	8.2

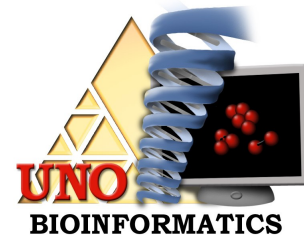


# Correlation Networks

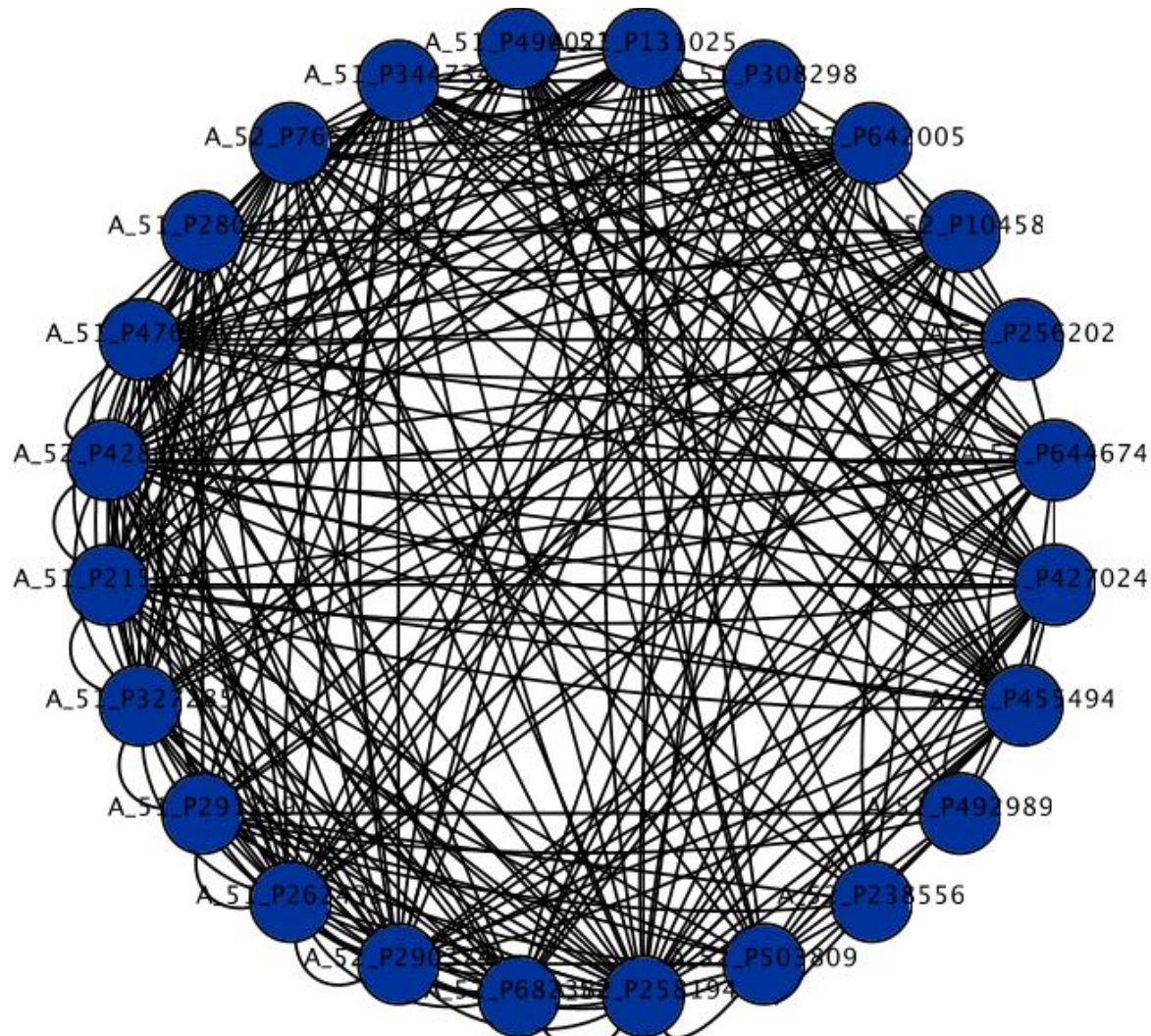




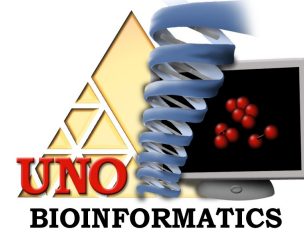
# Correlation Networks



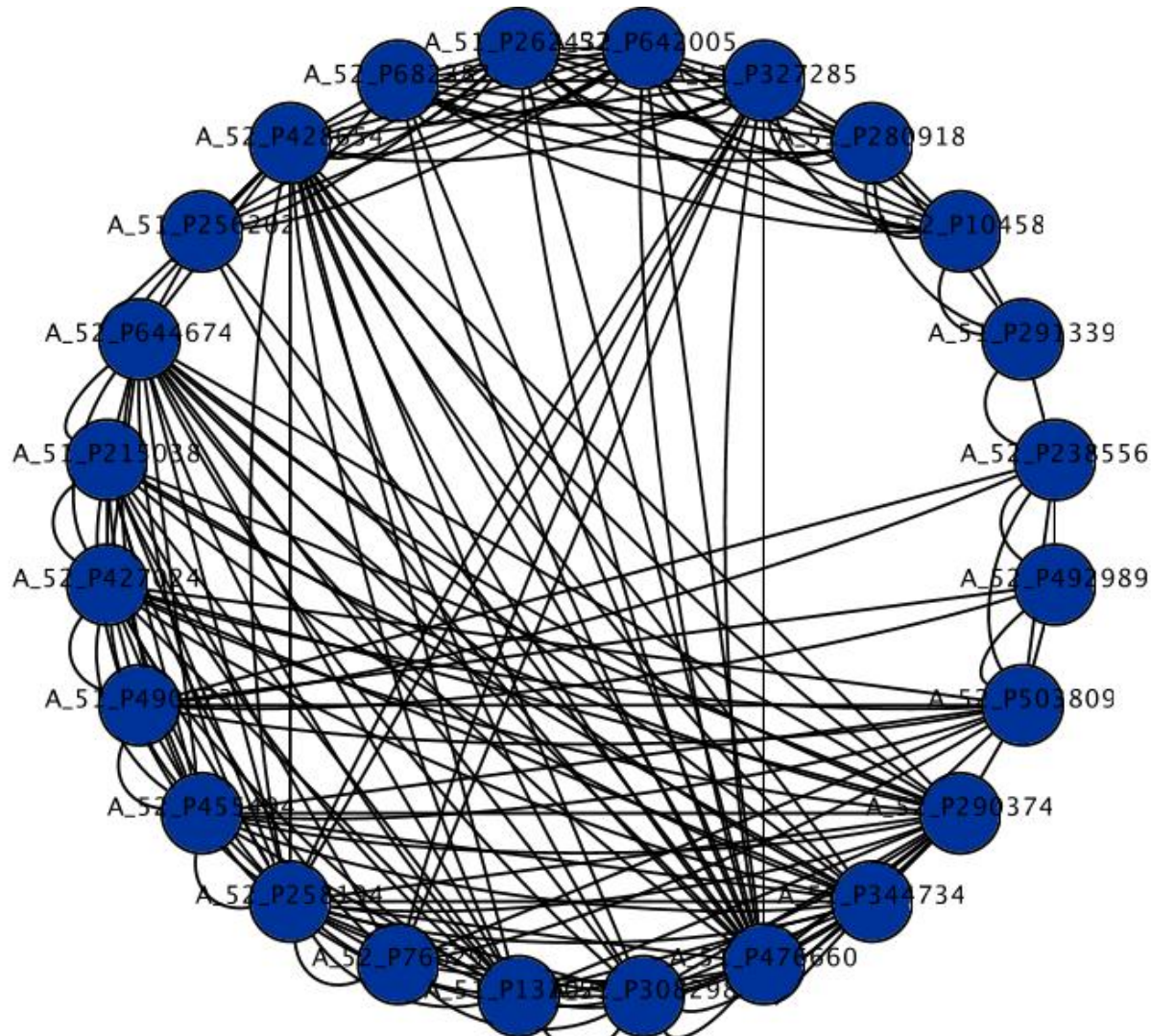
24 node sample  
Threshold: 0.00-1.00



# Correlation Networks



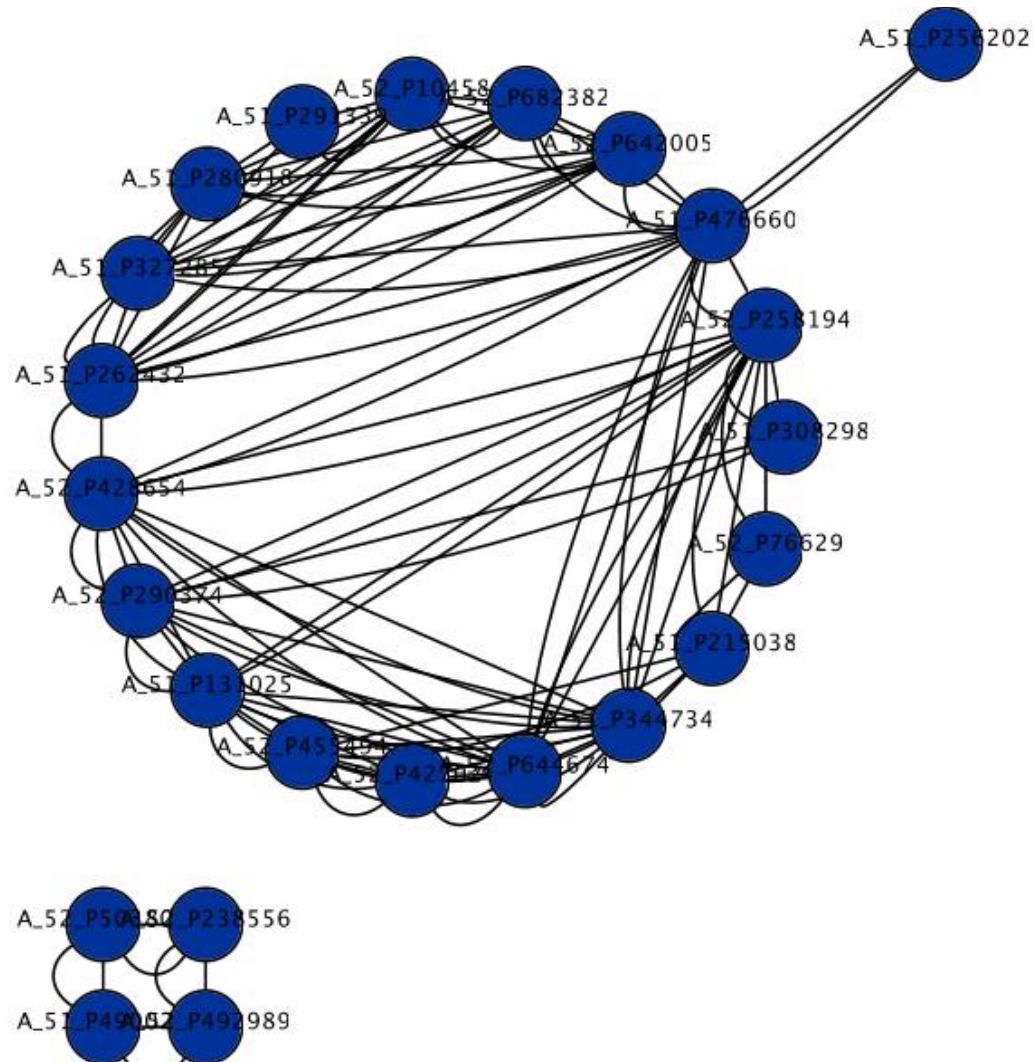
24 node sample  
Threshold: 0.30-1.00



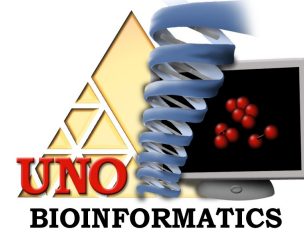


# Correlation Networks

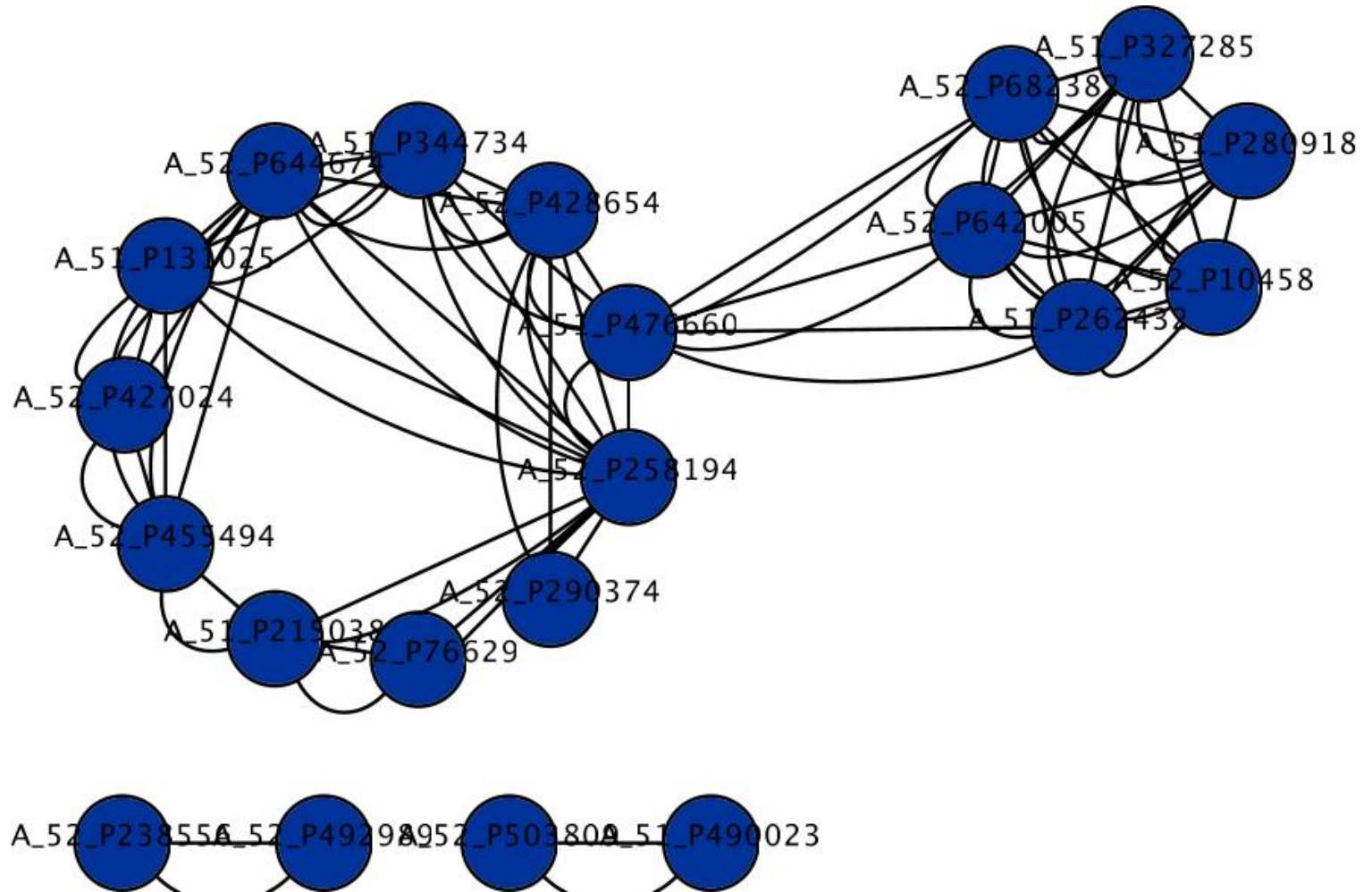
24 node sample  
Threshold: 0.50-1.00



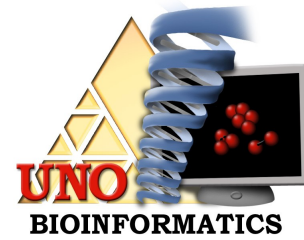
# Correlation Networks



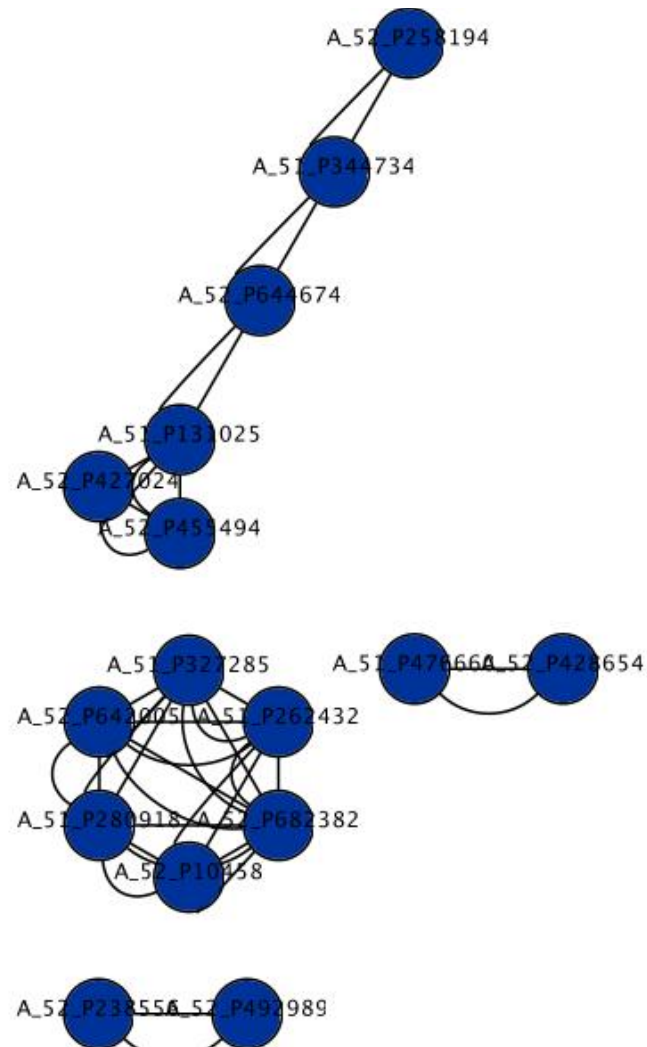
24 node sample  
Threshold: 0.60-1.00



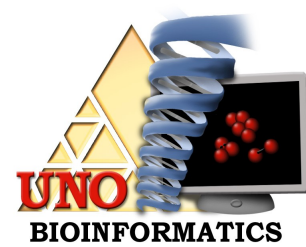
# Correlation Networks



24 node sample  
Threshold: 0.80-1.00

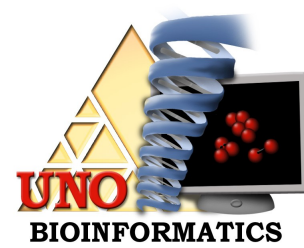


# Correlation Network Applications



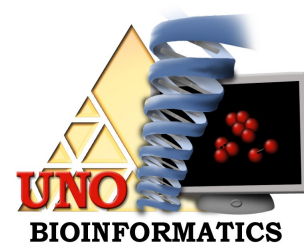
- “Versus” analysis
  - Normal vs. disease
  - Times/environments
- Model for high-throughput data
  - Especially useful in microarrays
- Identification of groups of causative genes
  - Ability to rank based on graph structure
  - Identify sets of co-regulated, co-expressed genes

# Network Concepts



- **Biological networks have structural properties**
  - Can differ from one network to another
- **Specific structures/characteristics have biological meaning**
  - Degree can indicate essentiality
  - Cluster density can indicate relevance
- **Networks do not have to be static**
  - Most interesting discoveries coming from temporal or state-change network alignment & comparison

# 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

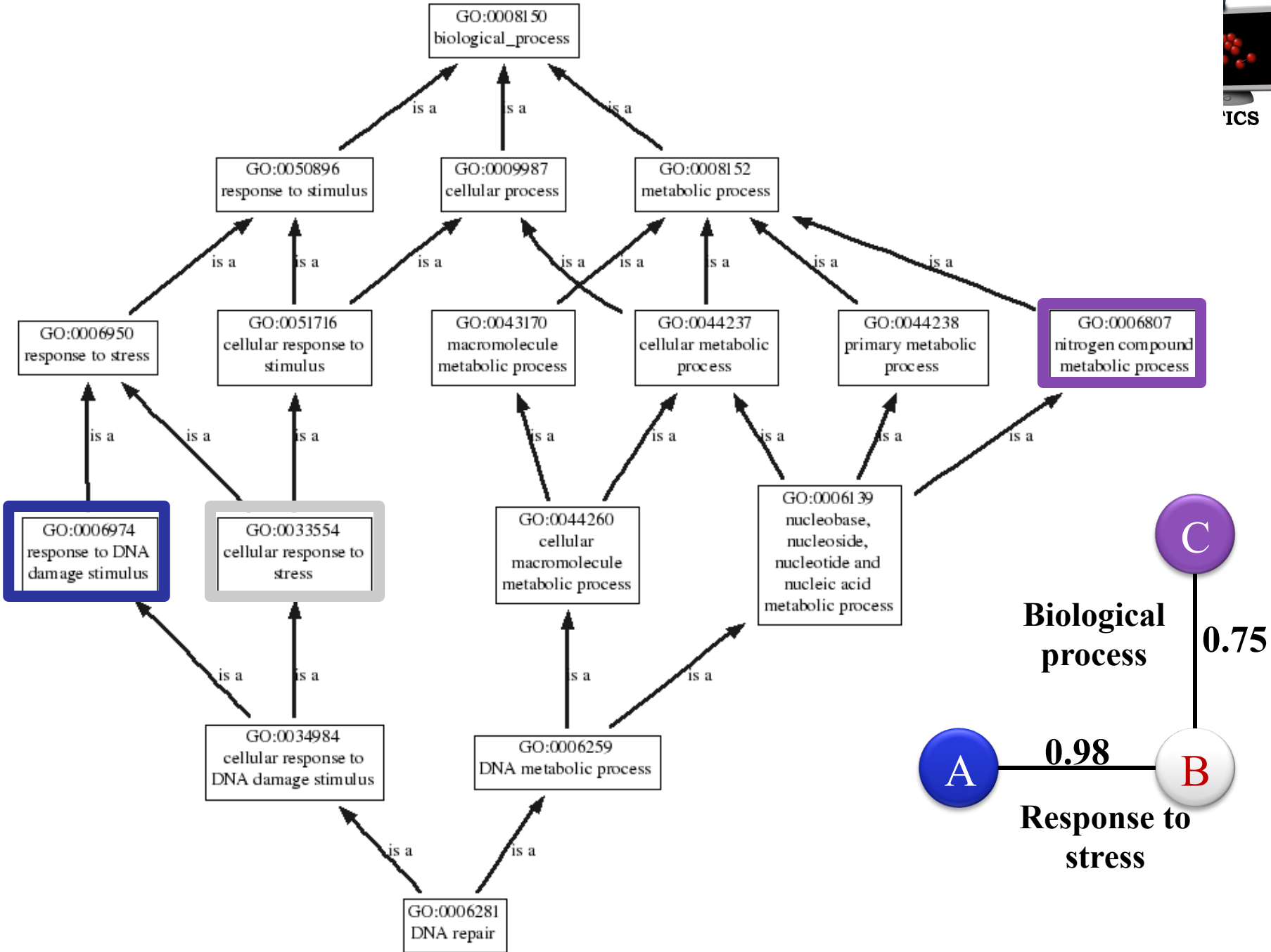
**Measures of centrality to identify key elements:  
Proves existence of structure/function  
relationship in correlation networks**



# Structures & their Functions

Network structures correspond to key cellular structures



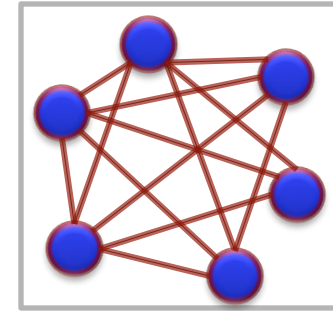




# Local Network Structures

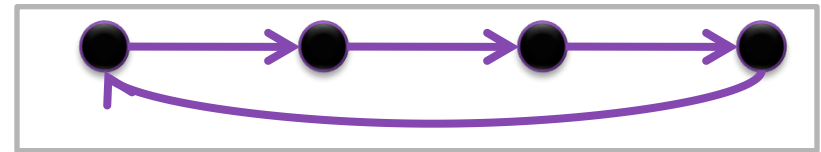
- **Cliques**

Protein complexes, regulatory modules



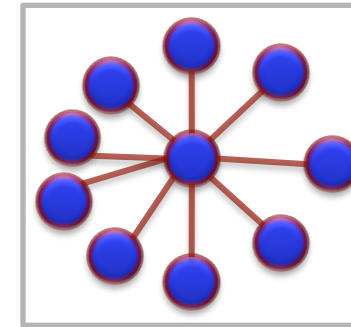
- **Pathways**

Signaling cascades

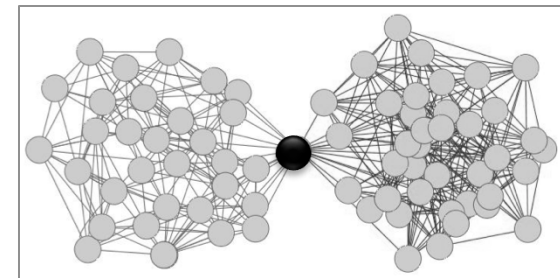


- **Hubs**

Regulators, TFs, active proteins



- **Articulation points**



# 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

# Hubs

- **Hub:** a high-degree node in a network
- Node degree in filtered correlation networks follows power-law relationship
- Few nodes with high degree

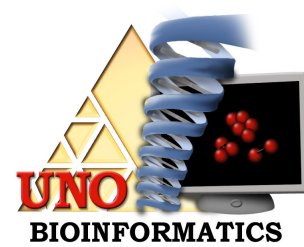
Albert et al 2005

- High degree nodes → highly essential

Bergmann et al 2004

Carlson et al 2006

# 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)

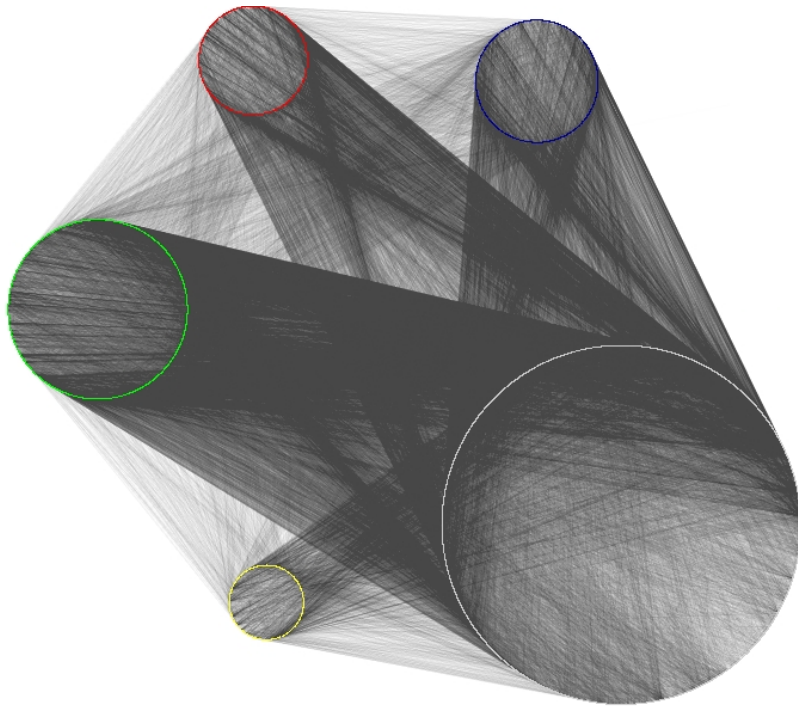
# Hub Lethality

- 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-affected:
      - Hspa1a: cellular, growth/size, homeostasis
      - Dapk1: cellular, renal/urinary
      - Ffar2: Increased susceptibility to colitis, asthma, arthritis

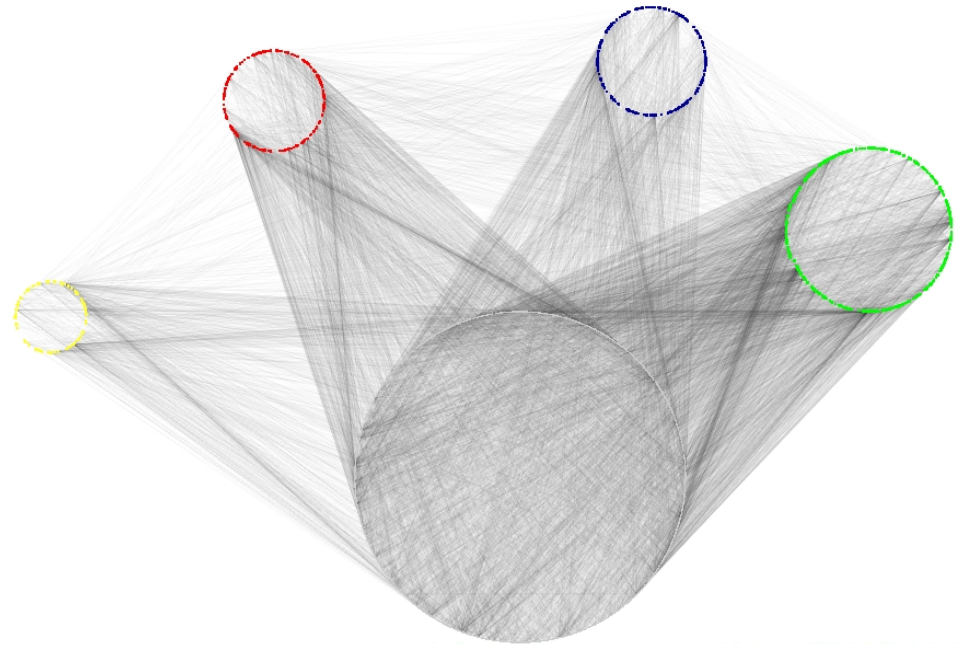
# Hub Lethality

- 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-affected:
      - Btn1a1: impaired lactation, impaired lipid accumulation in mammary gland
      - Bcl2l11: die later in life from auto-immune kidney disease
      - Rag2: arrested development of T and B cell maturation
    - 1/11 no observed phenotype (Aldh3a1)

# Aging and Biological Networks



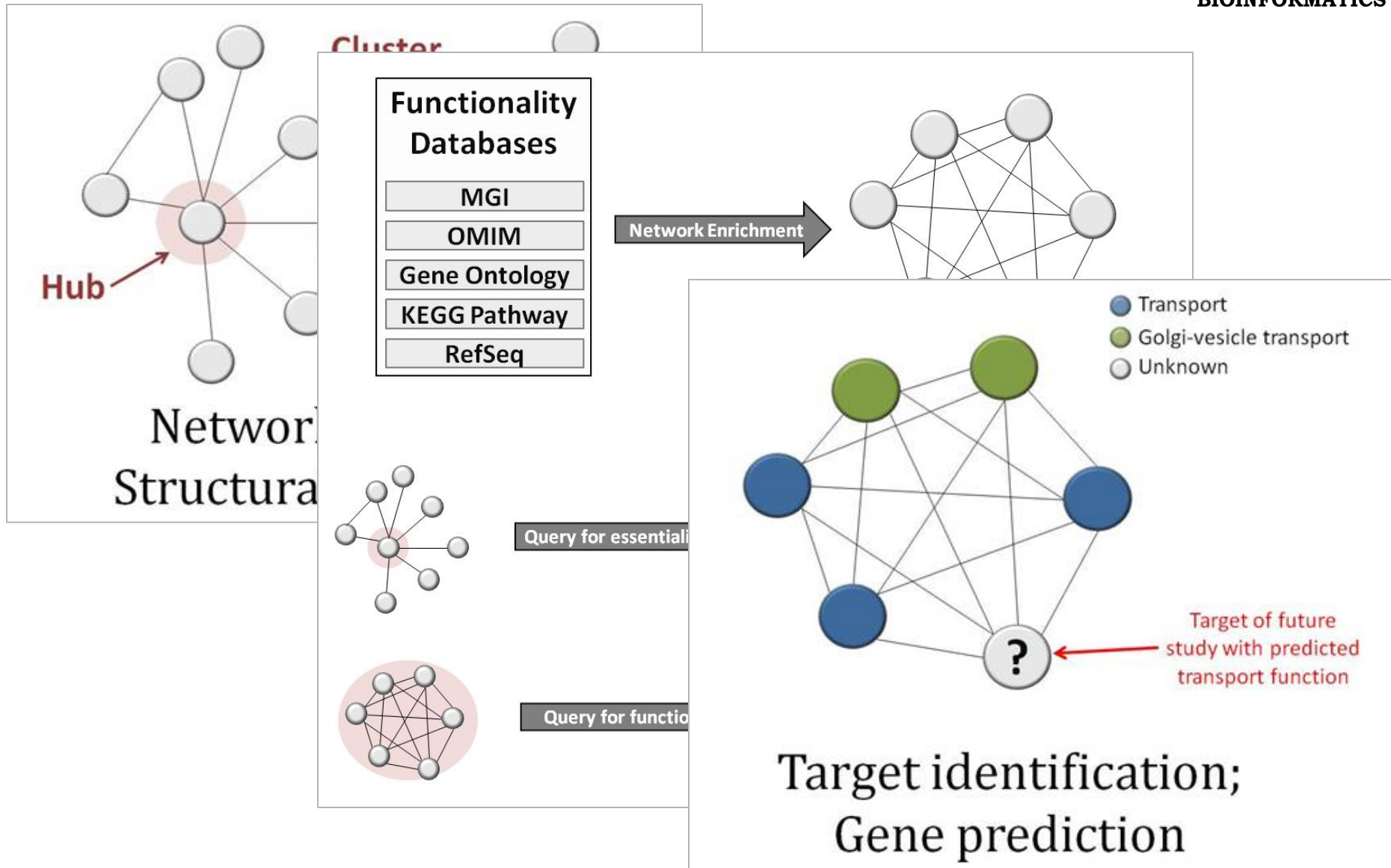
[young]



[aged]



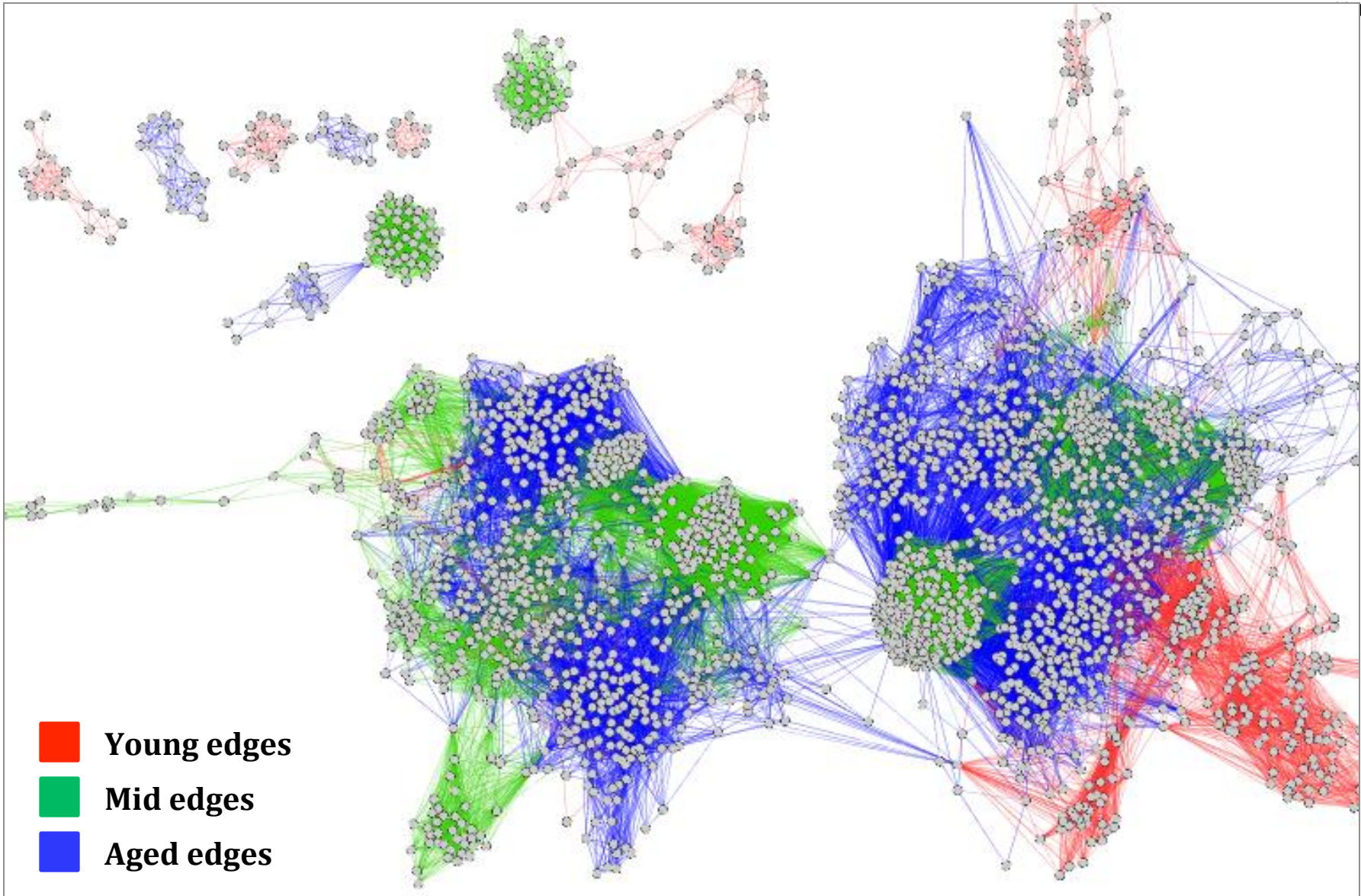
# Integrated Data Model



# Correlation Networks

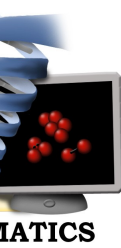
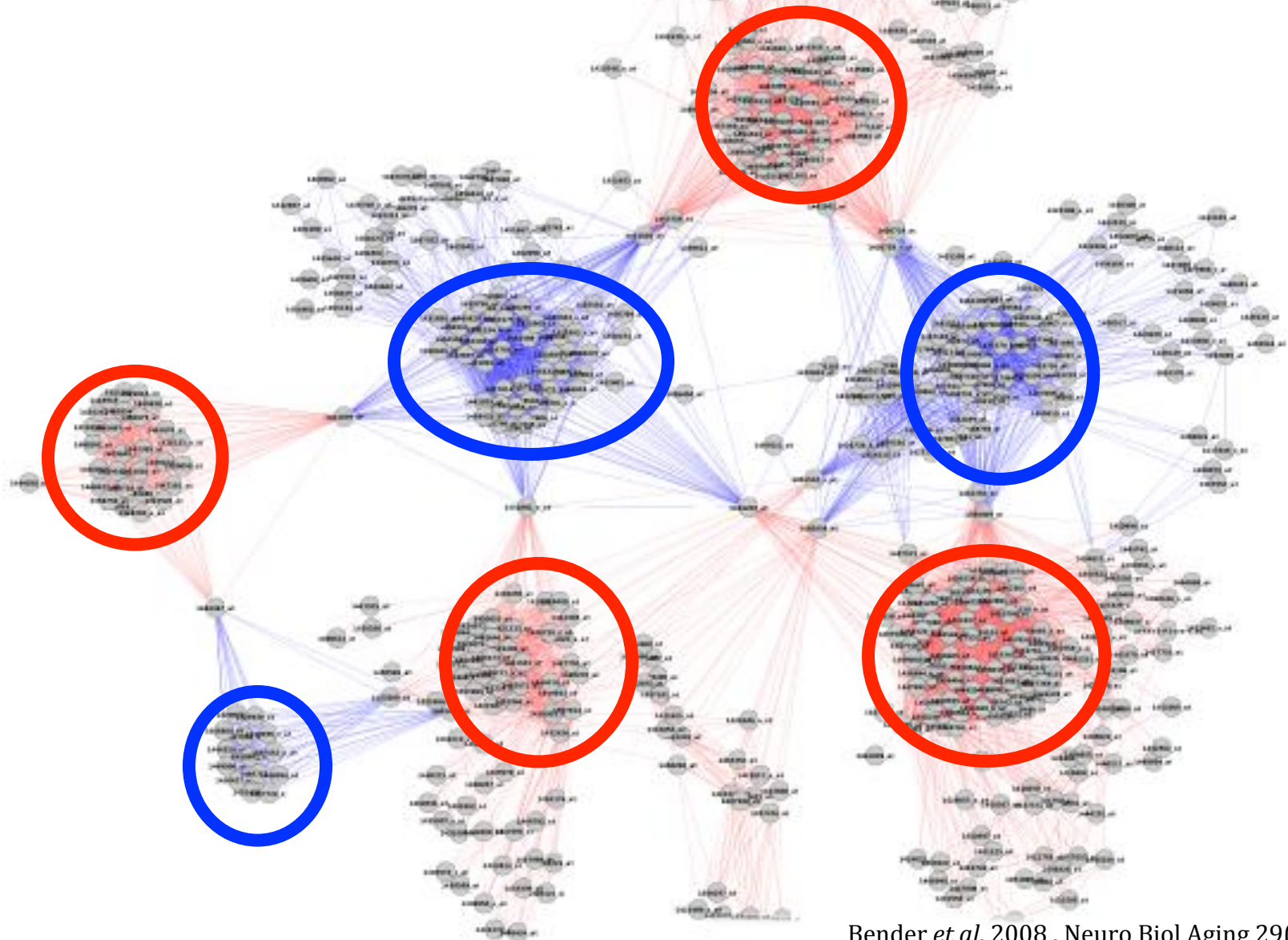


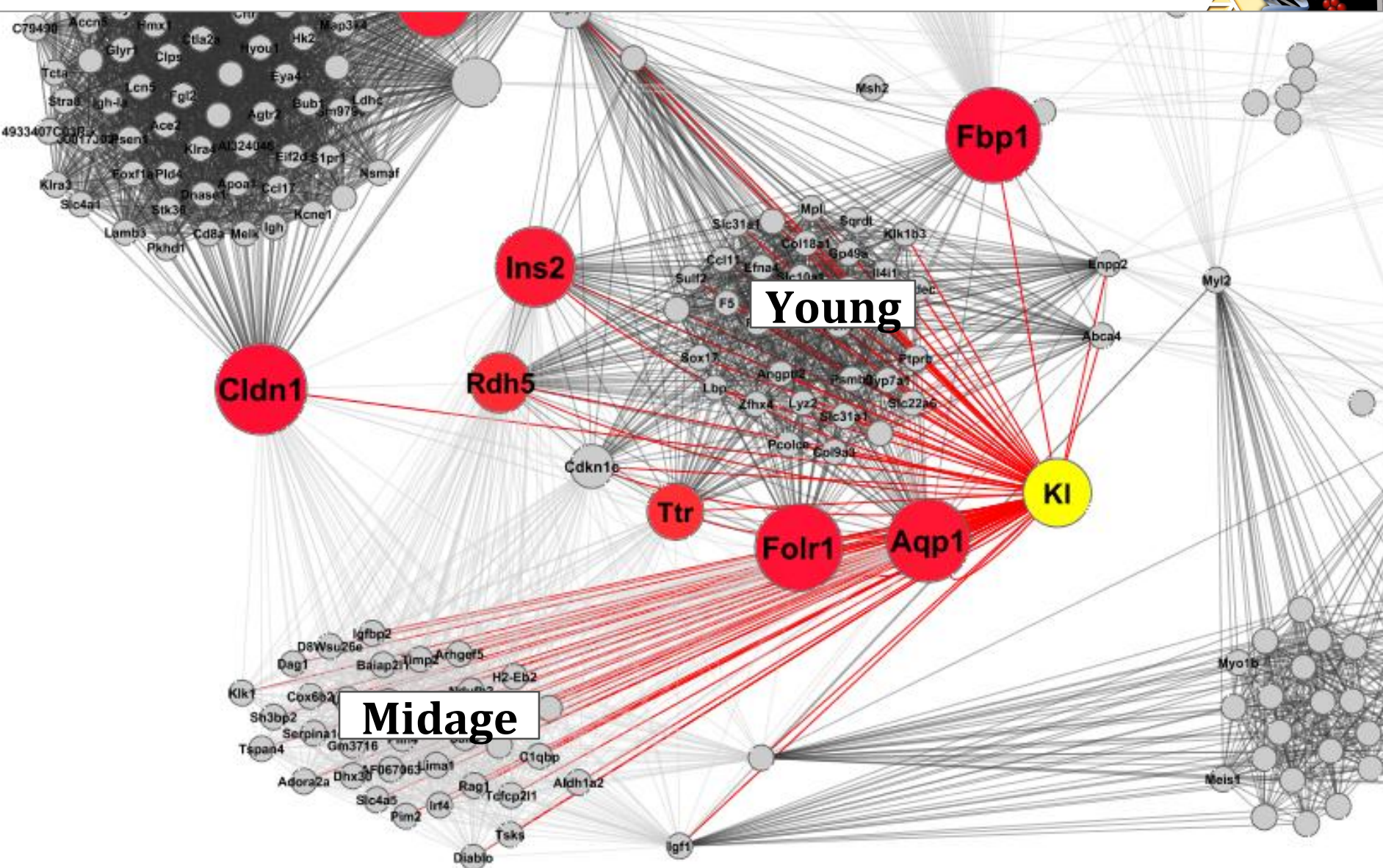
CS



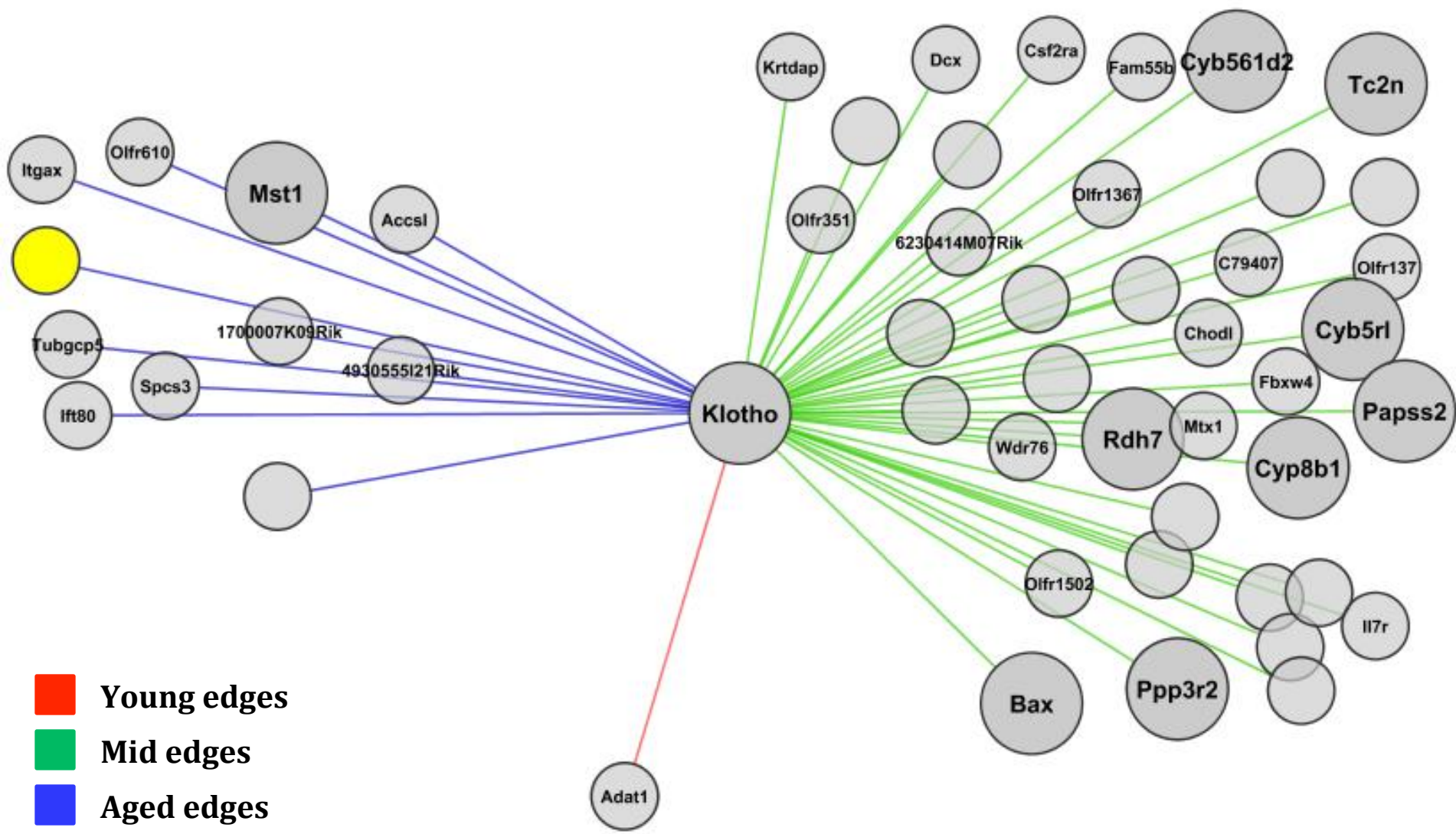


**Control**  
**Treated mice**









# Results Validation

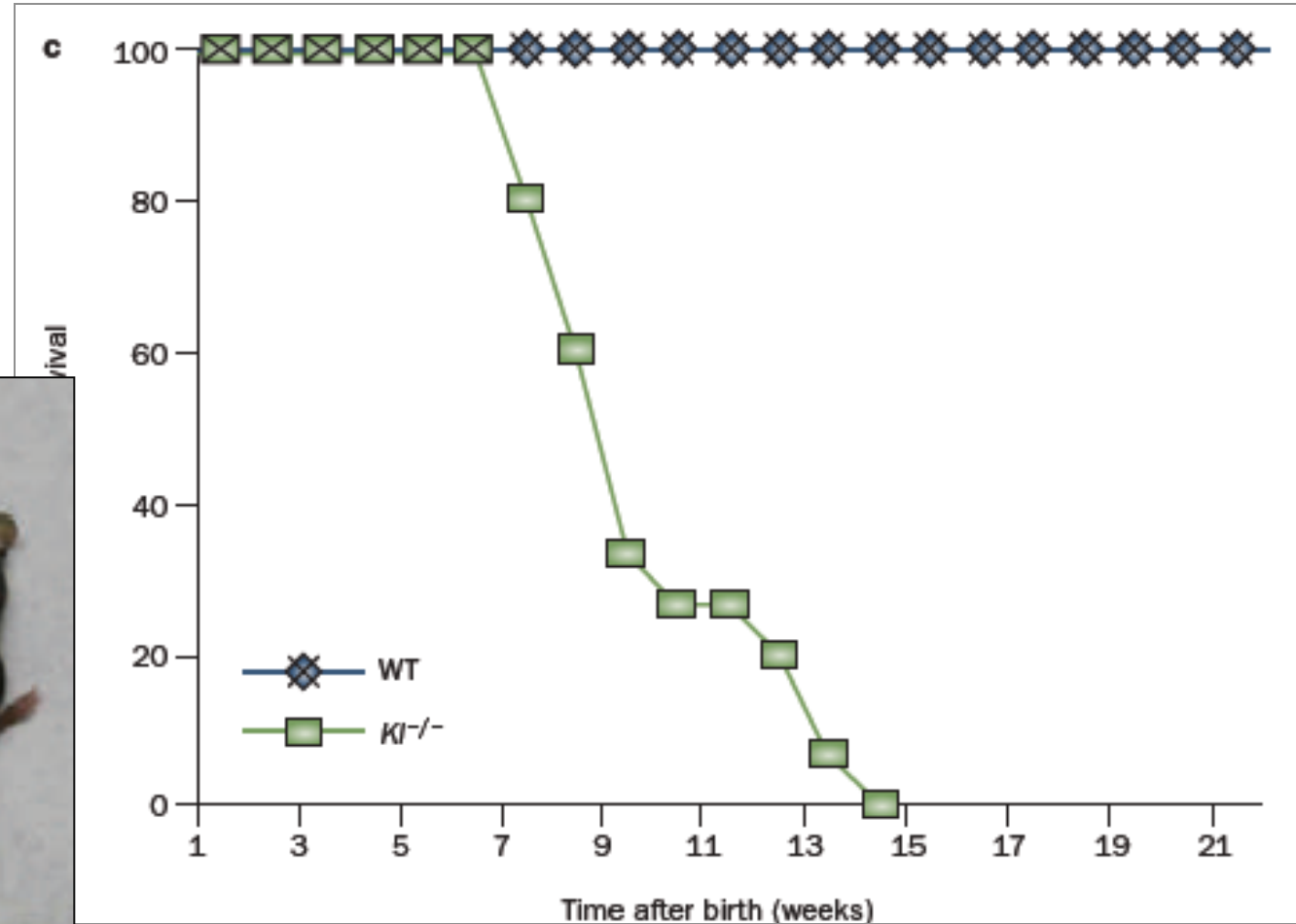
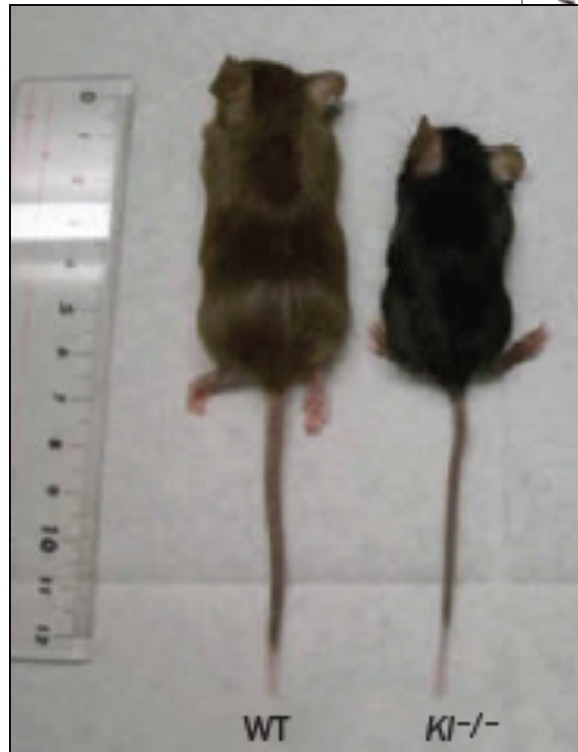


**Table 1**

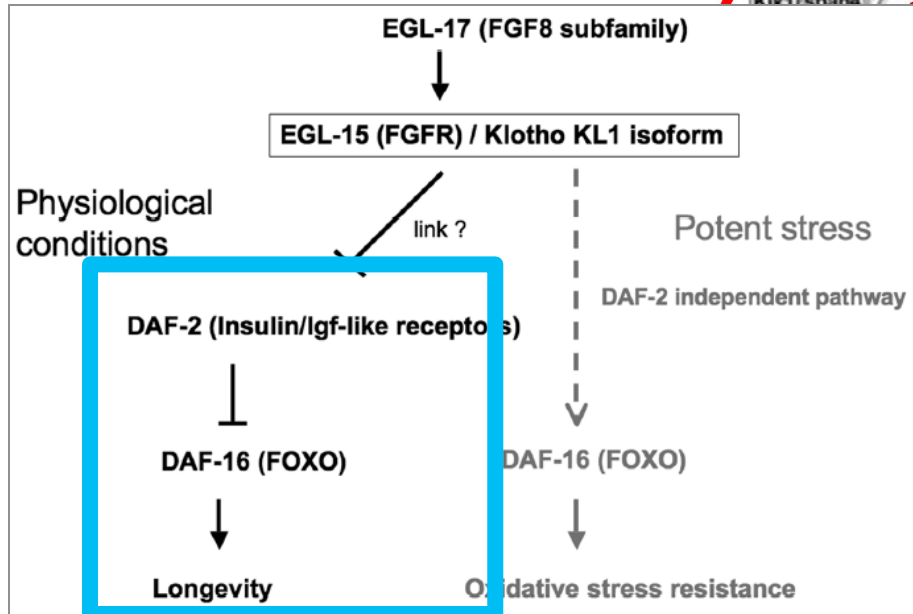
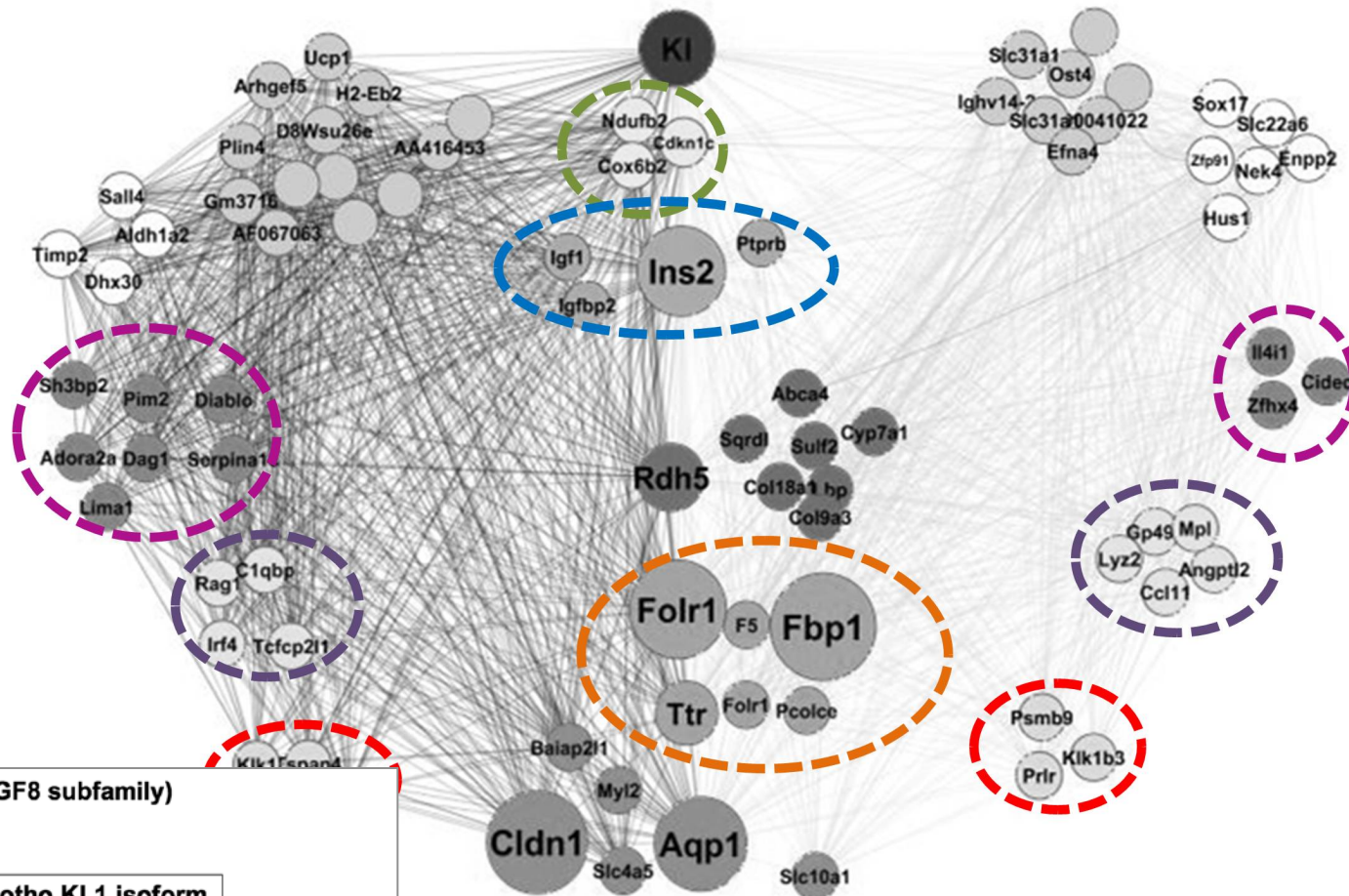
Comparison of phenotypes between klotho deficient and klotho overexpression mice

Parameters	Klotho deficient mice	Klotho overexpression mice
Body weight	Showing growth retardation and becoming inactive and marantic at 3 to 4 weeks of age (Kuro-o et al., 1997).	Normal (Kurosu et al., 2005)
Average lifespan	About 2 months (vs 2.5 to 3 years for wild-type mice) (Kuro-o et al., 1997).	About 20–30% longer than wild-type mice (Kurosu et al., 2005).
Maximal lifespan	Less than 100 days (Kuro-o et al., 1997).	More than 936 days (Kurosu et al., 2005).
Insulin	Decreased insulin secretion and enhanced insulin sensitivity (Kuro-o et al., 1997).	Increased resistance to insulin and IGF-1 signaling (Kurosu et al., 2005).
Phosphorus homeostasis	Hyperphosphatemia (Kuro-o et al., 1997).	Normal (Kurosu et al., 2005).
Calcium homeostasis	Ectopic calcification in various organs (Kuro-o et al., 1997).	Normal (Kurosu et al., 2005).
Diseases	Hypogonadism, infertility, premature thymic involution, ectopic calcification, decreased bone mineral density, skin and muscle atrophy, ataxia, emphysema, cognitive impairment, hearing loss, vascular calcification (Kuro-o et al., 1997). Reduction of NO synthesis in vascular endothelial cells (Saito et al., 1998).	Protection of the angiotensin II-induced renal damage (Mitani et al., 2002). Suppression of H <sub>2</sub> O <sub>2</sub> -induced apoptosis and cellular senescence in vascular cells (Ikushima et al., and 2006). Reduction of risk factors for atherosclerosis. Enhanced hearing ability (Bektas et al., 2004)

# High BD Node: Validation

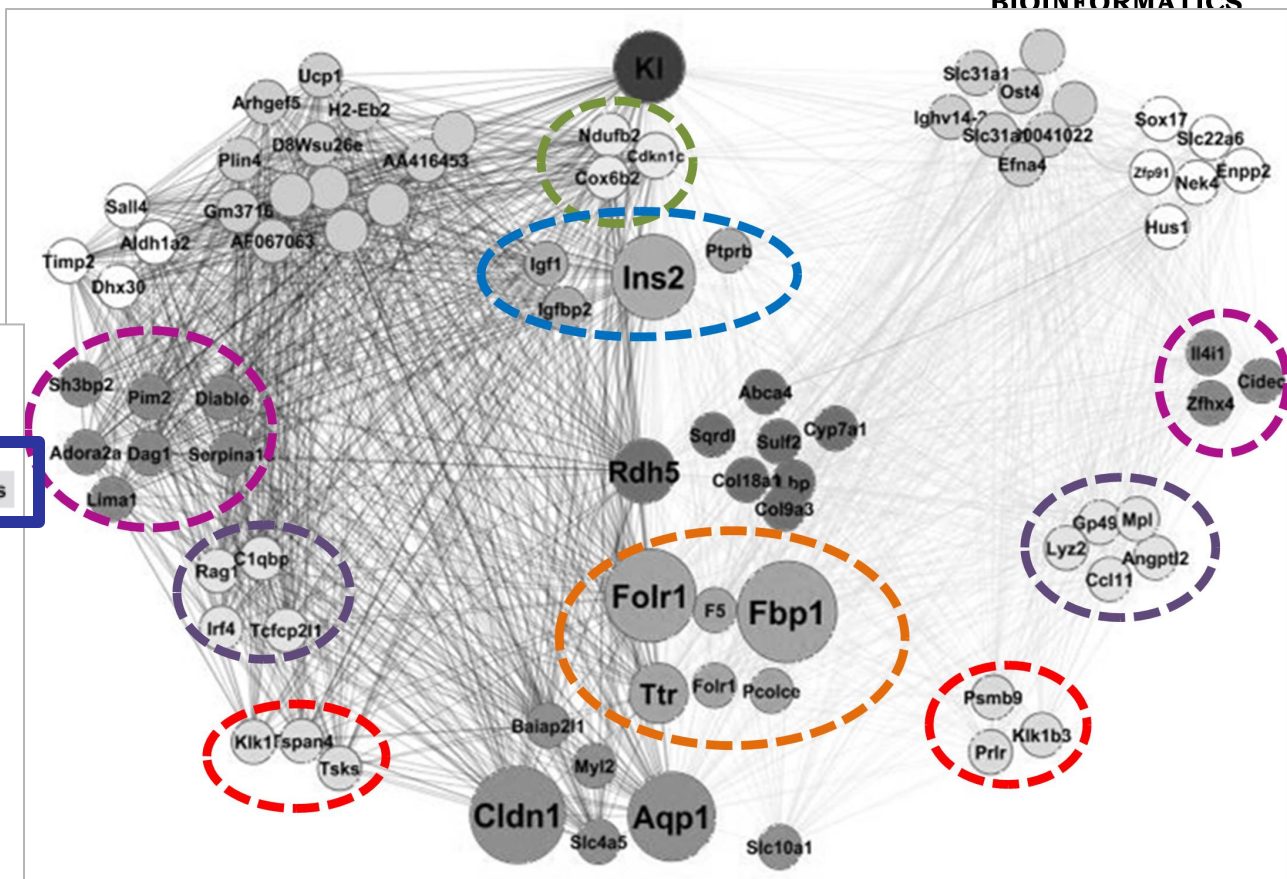
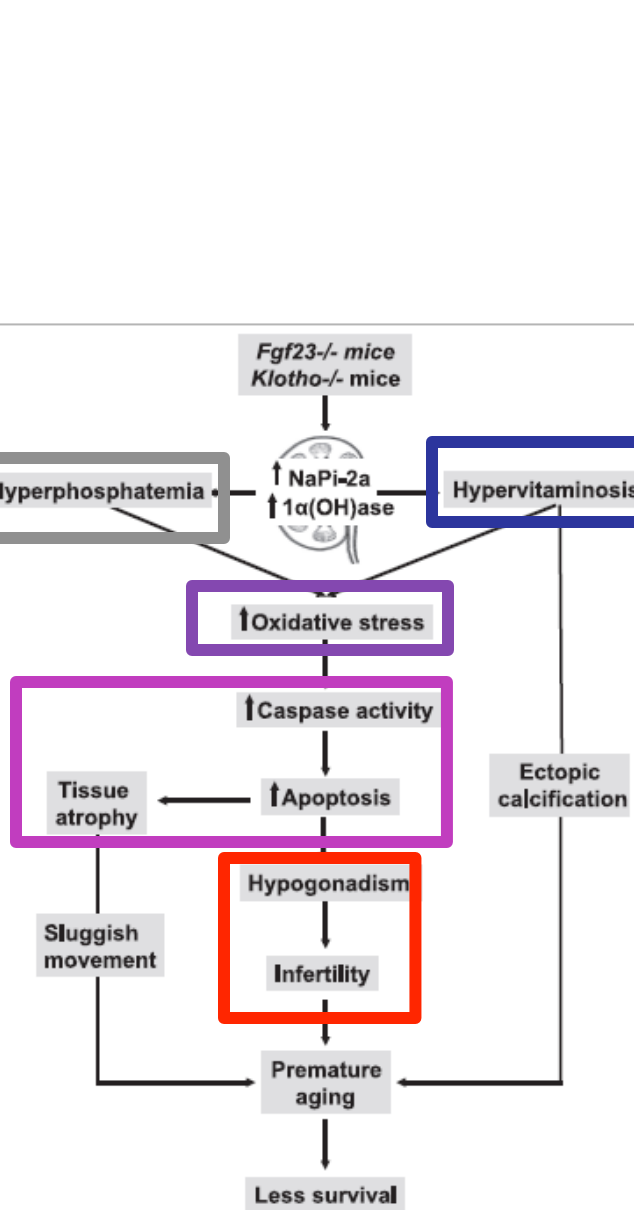


# Validation

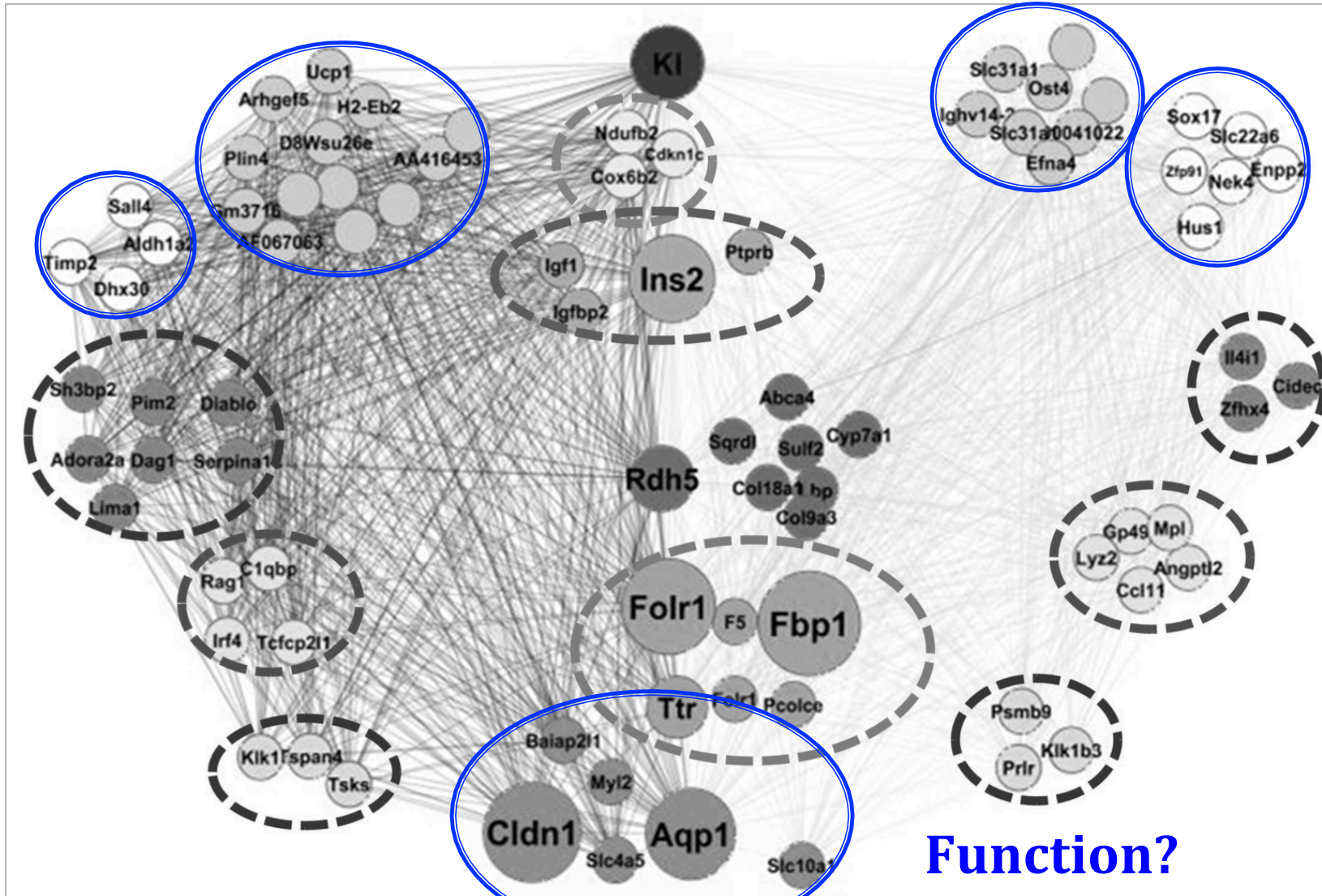




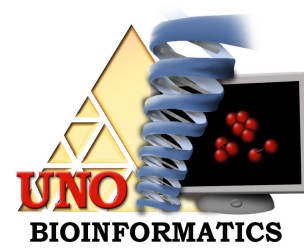
# Subsystem Validation



# Discovery



# Summary

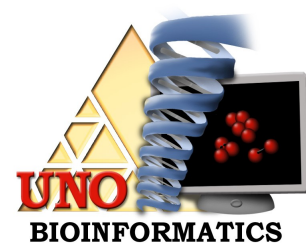


- Networks → very efficient modeling system
  - Basis of next generation data analysis tools in systems biology
- Structure/function relationship exists
  - Integrated networks to identify gene drivers
- Future: Model will play a role in aging/  
disease *prevention, diagnosis, and treatment*

# Tutorial Outlines

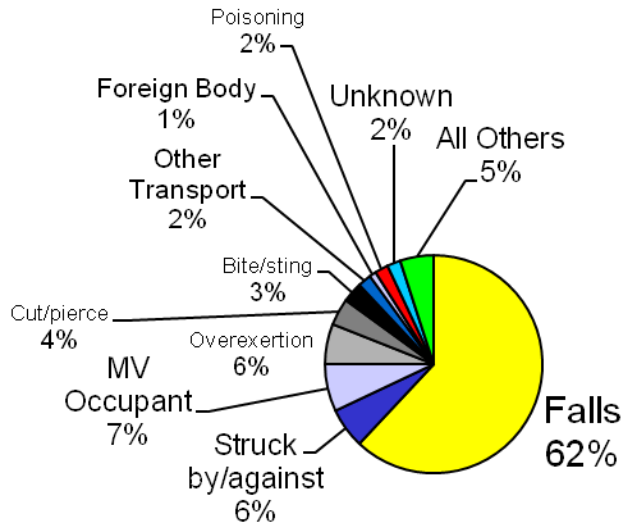
- Introduction to Biomedical Informatics
  - State of the discipline - Challenges and Opportunities
  - Data-driven biomedical research
- Next Generation Bioinformatics Tools
  - Intelligent Collaborative Dynamic (ICD) Tools
- *Case Study: Aging Research*
  - The genomic study: Correlation Networks
  - *Mobility and aging: Wireless monitoring*
  - Data collection and Virtual Environments
- Next Steps: Where do we go from here?
  - HPC and Cloud Computing

# Falls and Associated Problems

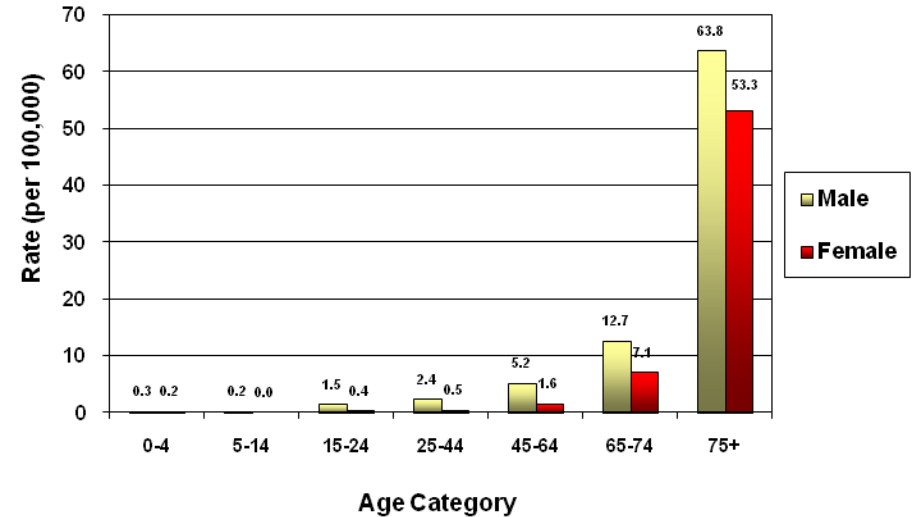


- Falls are the leading cause of accidental deaths in the United States among people over the age of 75
  - the number of fatalities due to falls increased steadily from 14,900 in the year 2000 to 17,700 in 2005.
- Nebraska's over age 65 population is 13.3% versus 12.4% for the national average.
  - Generally speaking, the more rural the area, the higher the percentage of older adults.
  - In Nebraska, approximately 78% of those hospitalized for fall related injuries were 65 years and older.

# Falling Problems



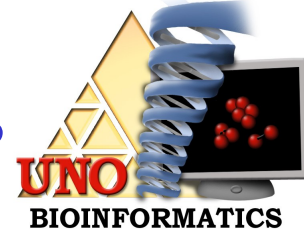
Incidence Rate of Fatal Fall Injuries



- Approximately 78% → 65 years and older.
- falls – leading cause of
  - injury deaths
  - injuries and trauma
- The risk of falling increases with age.



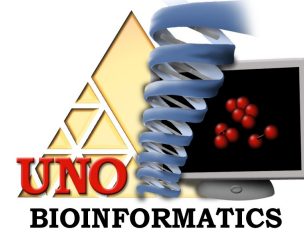
# Traditional Gait Monitoring Methods



- Expensive
- Uncomfortable
- Limited mobility
- Complicated



# Laboratory-based Gait Monitoring



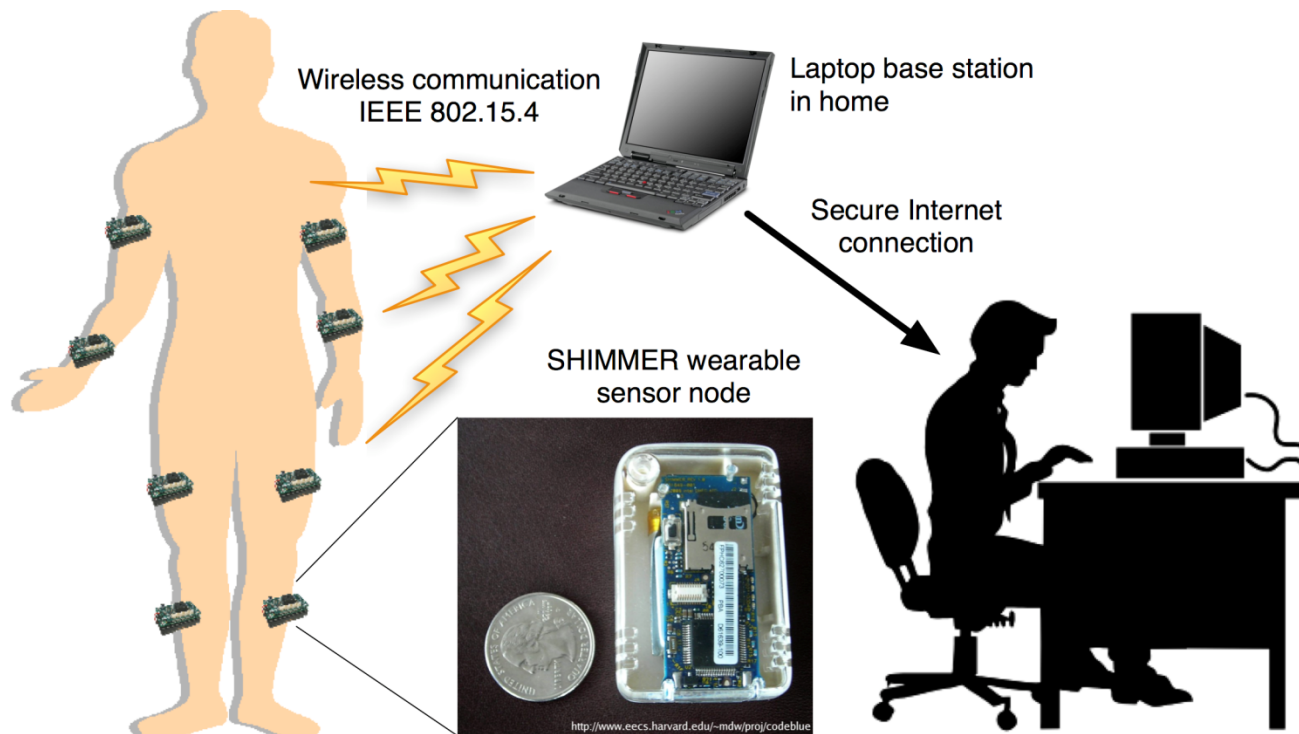
- Expensive
- Uncomfortable
- Limited mobility
- Complicated





# Wireless Sensor Based Mobility Monitoring

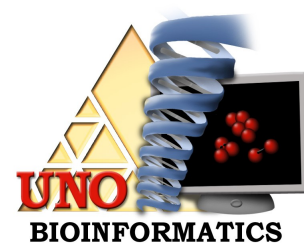
- Inexpensive
- Comfortable
- High mobility
- Simple



# Wireless Sensors and Monitoring

- Human health can be significantly improved by monitoring the mobility patterns of individuals
- It is not easy to measure human activities because they vary from person to person
- We developed a physical activity monitoring system using one wireless sensor
- Wireless sensor platform for this study is small, lightweight, and user-friendly
- We considered wireless communication and computation on this platform to minimize energy consumption

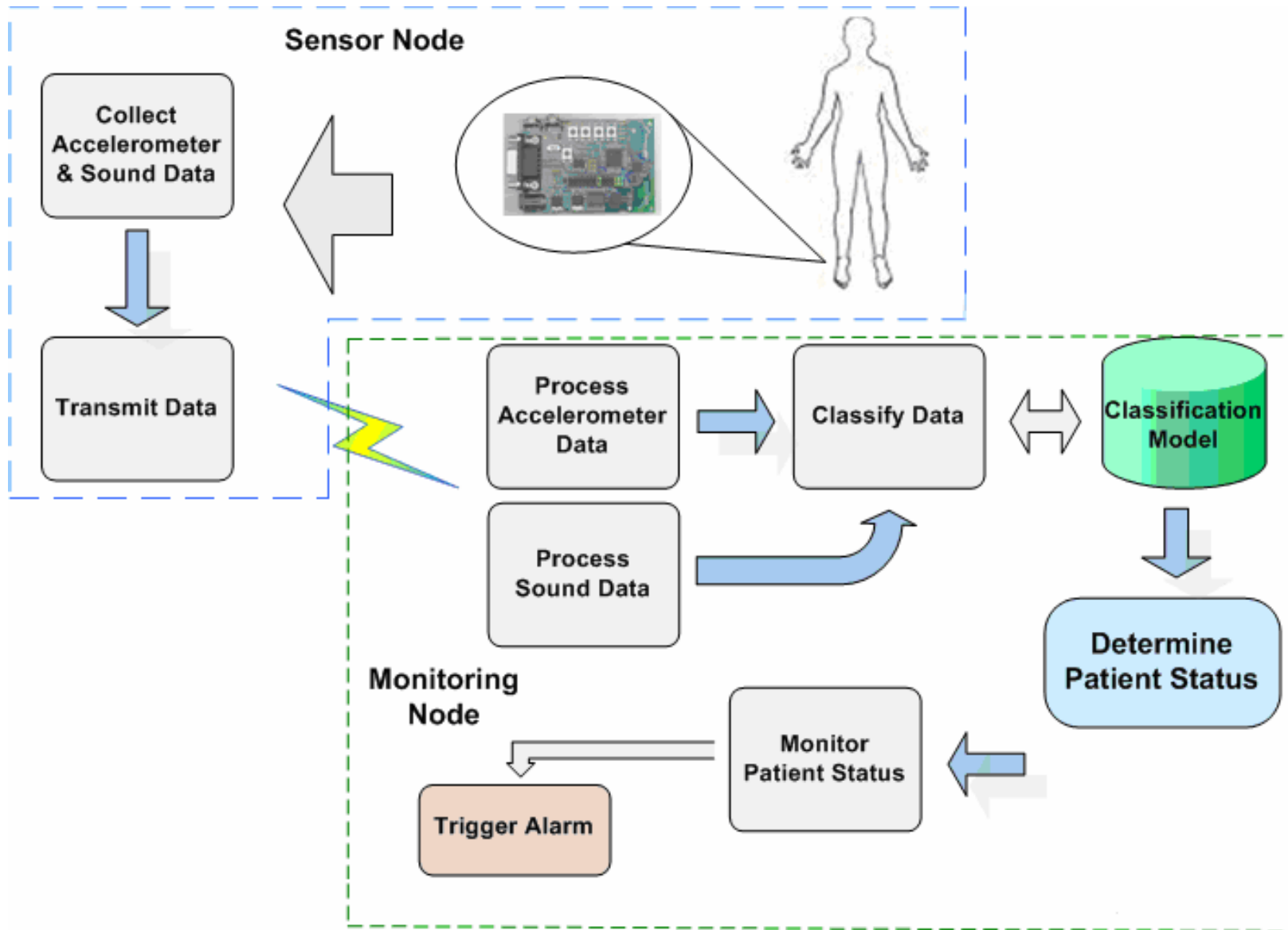
# Goals of the Project



- Mobility Profile
  - Patients wearing the accelerometers will be monitored 24/7.
  - A complete mobility profile will be available for patients and care providers.
- Fall 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,

# Four Phases of the Project

- Phase I: Fall Detection
  - achieved over 95% of fall detection rate
- Phase II: Classification of ADLs (Activities of Daily Living)
  - Running, Walking, Stair Climbing, Jumping, ...
- Phase III: Construction of Mobility Profiles
- Phase IV: Fall (major health hazards) Prediction based on mobility profiles



Fall Detection Algorithm using Accelerometer and sound data

# Mobility Sensors

- Accelerometer
  - Impact detection
  - (unit: gravity)
- Gyroscope
  - Measure rate of rotation
  - (unit: degrees per second)



(shimmer-research.com)

# Shimmers

- A wireless sensor platform for various types of wearable applications
- It consists of a number of integrated and extended sensors, a central processing unit, wireless communication module, and storage devices
- It has a low-power 8MHz MSP430 CPU, 10 KB RAM, 48 KB Flash memory, and 2 GB MicroSD card
- A 3-axis MMA 7361 accelerometer is integrated into Shimmer



# Shimmers



Image source: [http://www.shimmer-research.com/wp-content/uploads/wpsc/product\\_images/Shimmer-Size-Illustration.jpg](http://www.shimmer-research.com/wp-content/uploads/wpsc/product_images/Shimmer-Size-Illustration.jpg)

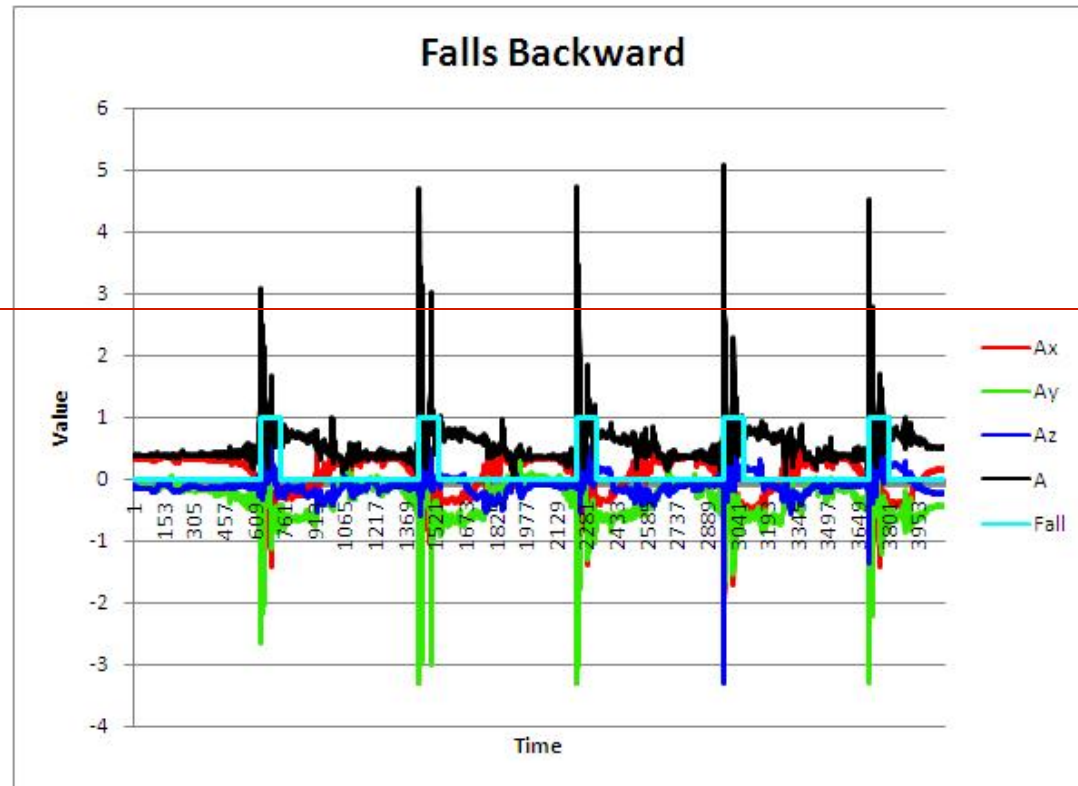


# Phase I: Fall Detection

## Accelerometer-based fall detection

- Determine an acceleration threshold.
- Detect fall.

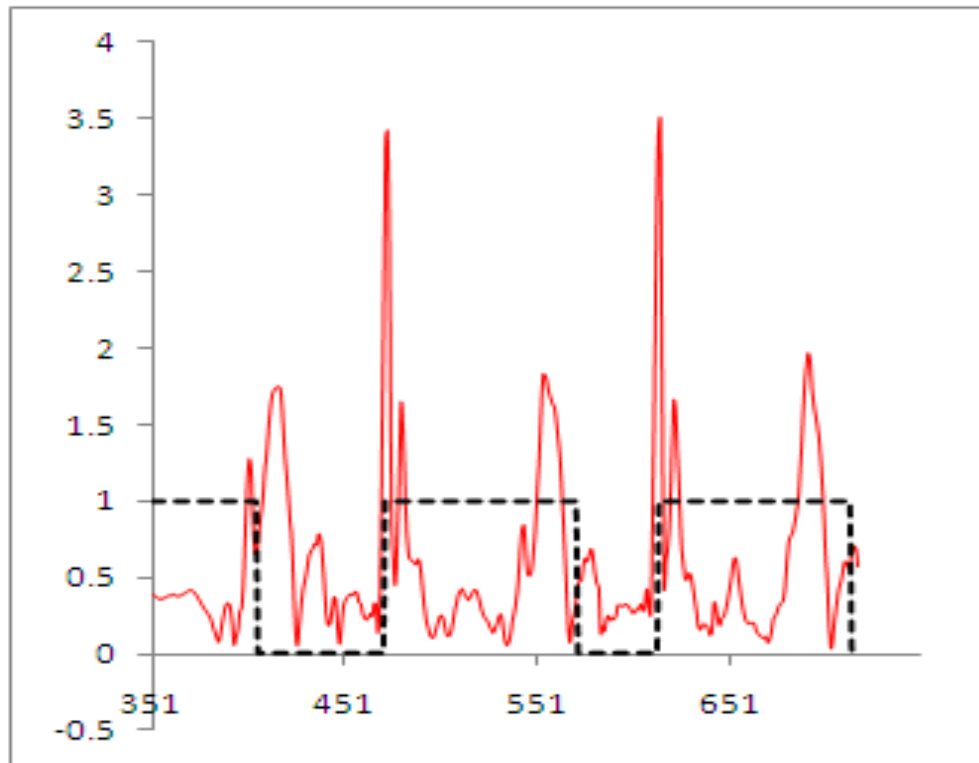
Threshold



# Phase 1: Fall Detection

## Accelerometer-based fall detection

- While the accelerometer-based algorithm is able to accurately detect major fall events, it also produced false positives for some events such as Jumping.♪



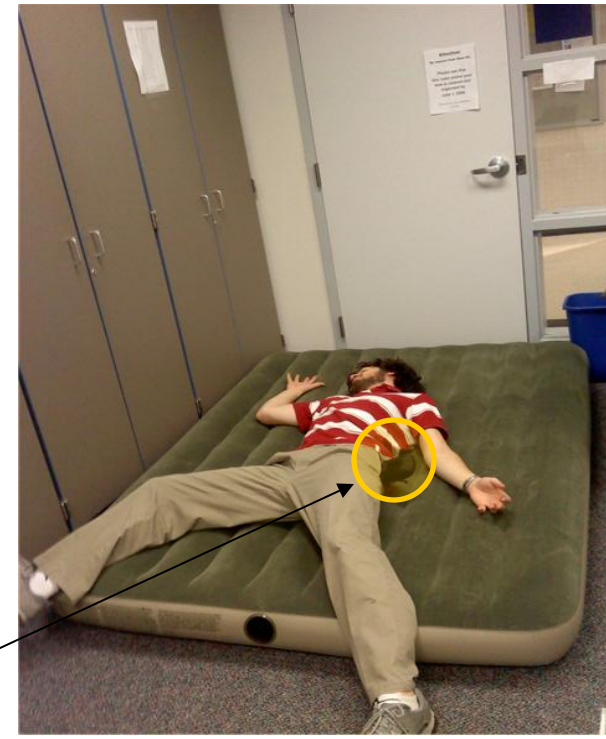
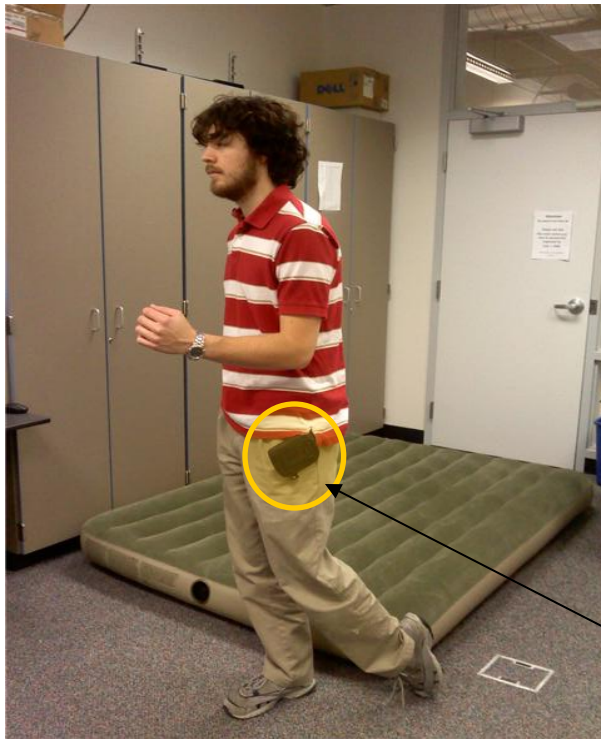
— A      - - - - - Fall Detected

Jumping

# Phase I: Fall Detection

## Accelerometer-based fall detection

- Measure acceleration in three orthogonal directions.



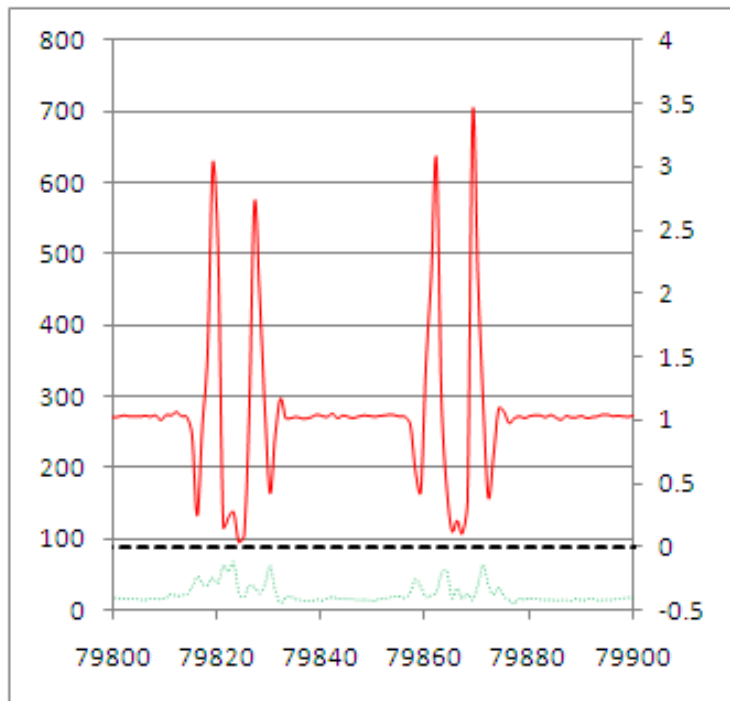
Shimmer Device

# Phase I: Fall Detection

## Adding Additional Sensors

- Using 3-D accelerometer and gyroscope sensors

Legends: — A    ..... G    - - - - - Fall Detected



Jumping



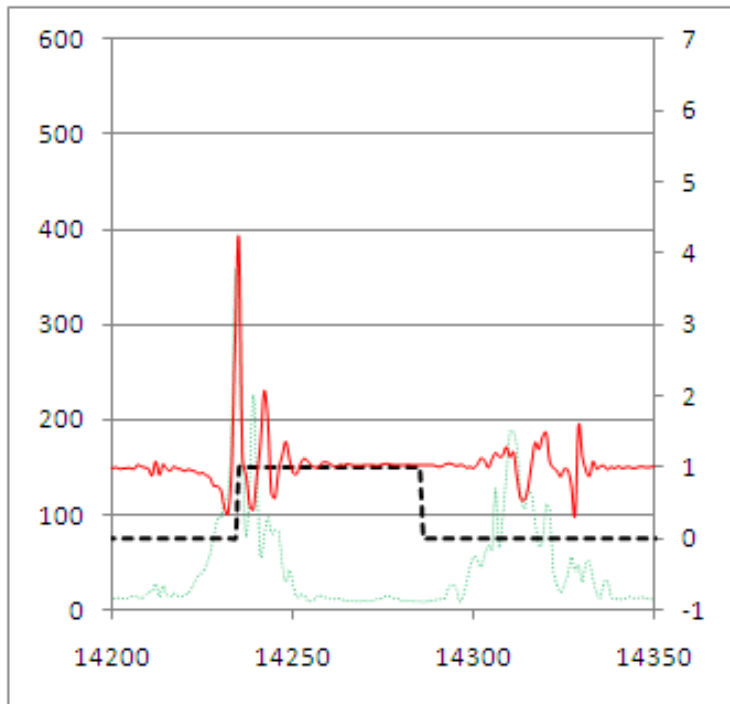
Fall backward

# Phase I: Fall Detection

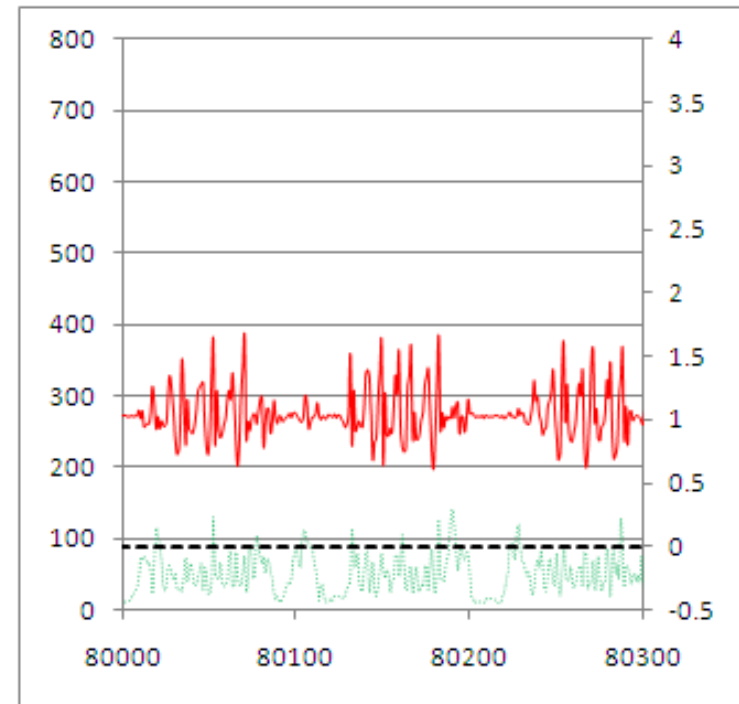
## Adding Additional Sensors

- Using 3-D accelerometer and gyroscope sensors

Legends: — A    ..... G    - - - - - Fall Detected



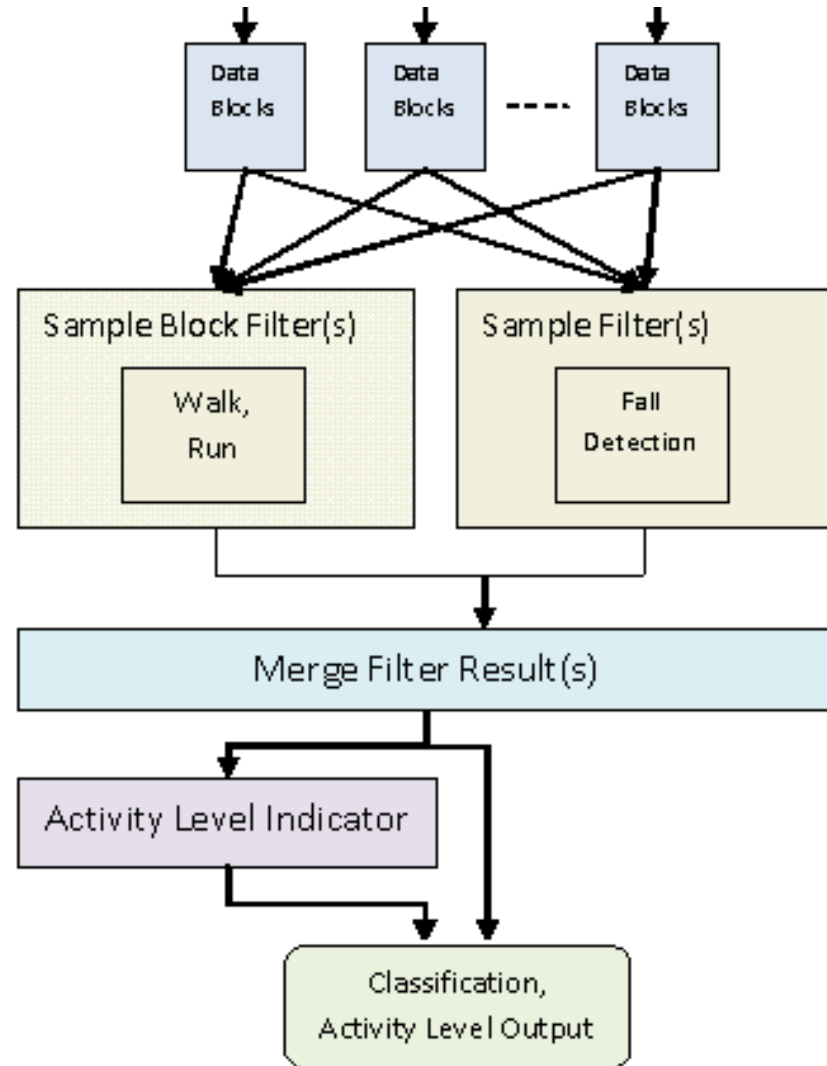
Fall sideways



Walking

## Phase II: Classification of ADLs

- An activity can be classified as walking or running based on the magnitude and frequency of a peak in Fourier transform.



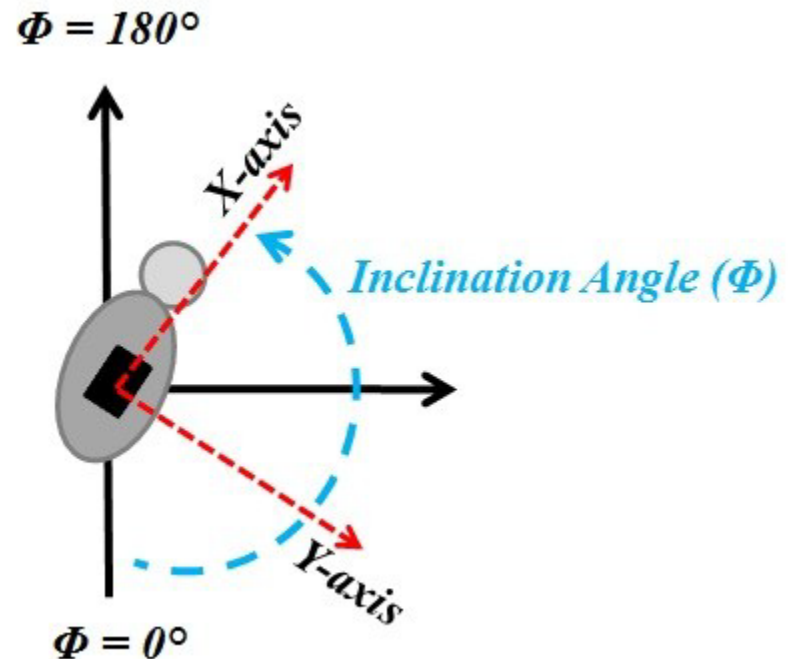
# Activity Classification

- Inclination angle has been selected to classify static activity like standing, sitting, and lying
- Calculates the angle with the x- and y-axis of accelerometer sensor

$$\Phi = \arctan A_{\downarrow y} / A_{\downarrow x}$$

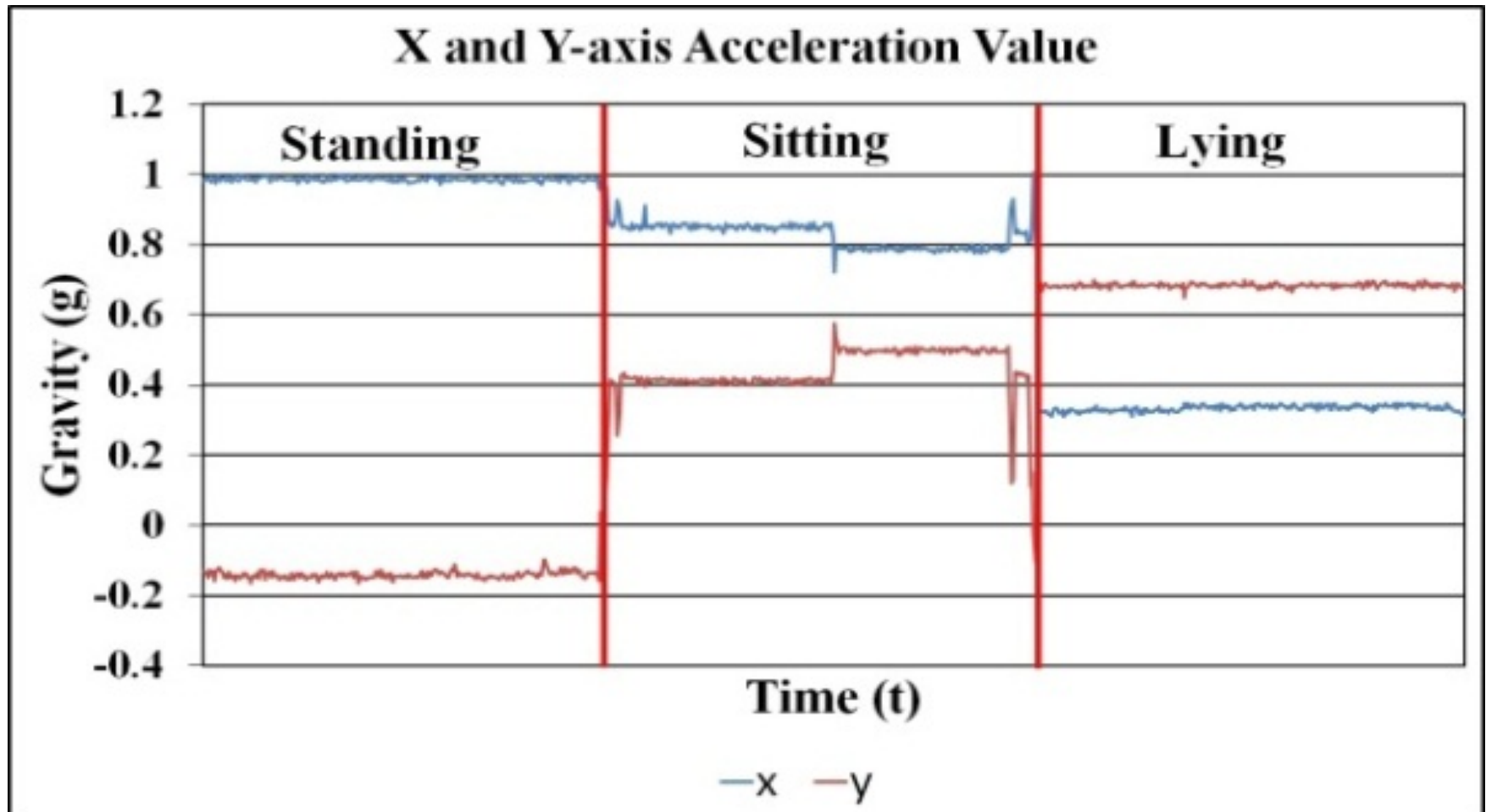
$A_x$ : acceleration value of x-axis

$A_y$ : acceleration value of y-axis



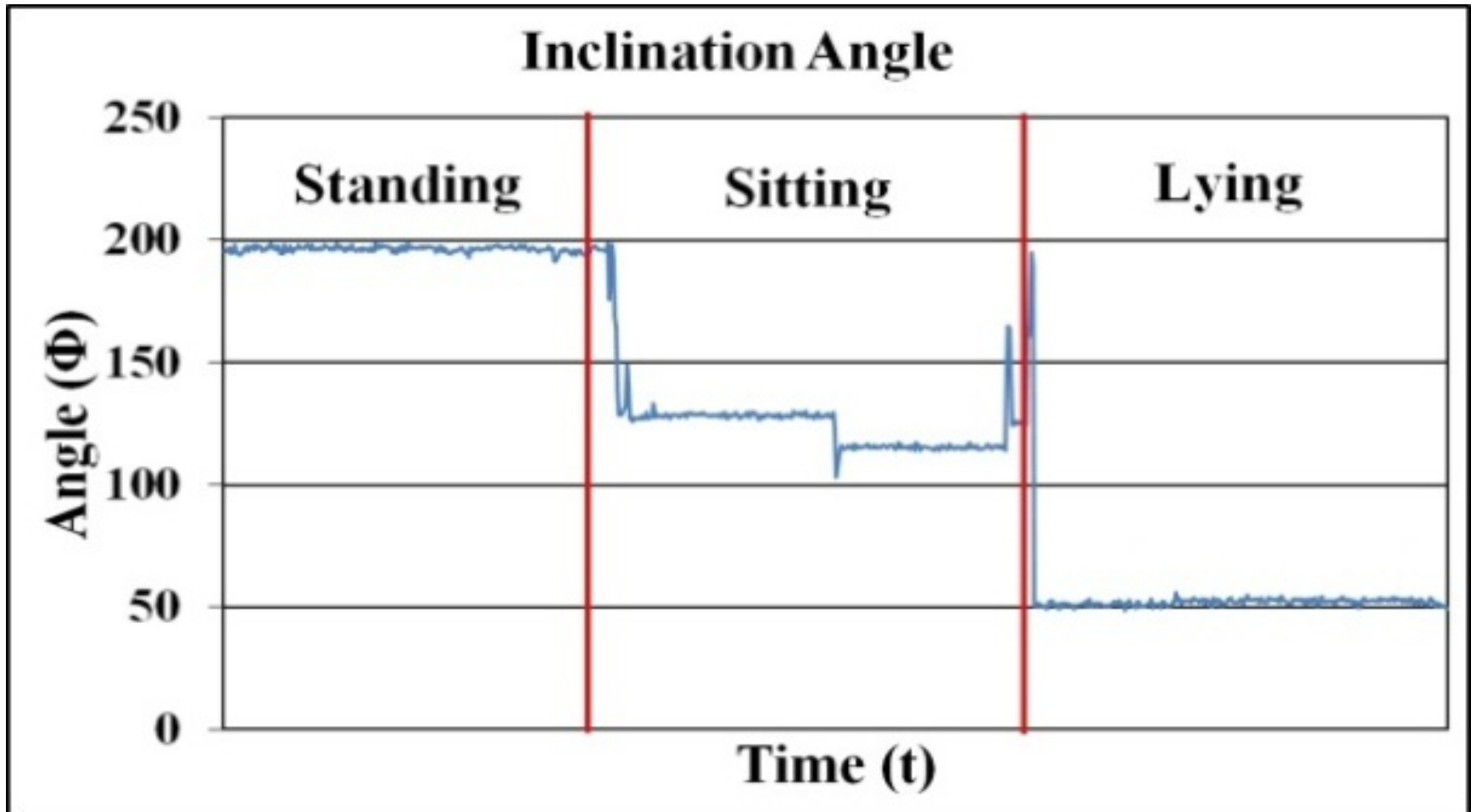


# Acceleration value of x- and y-axis

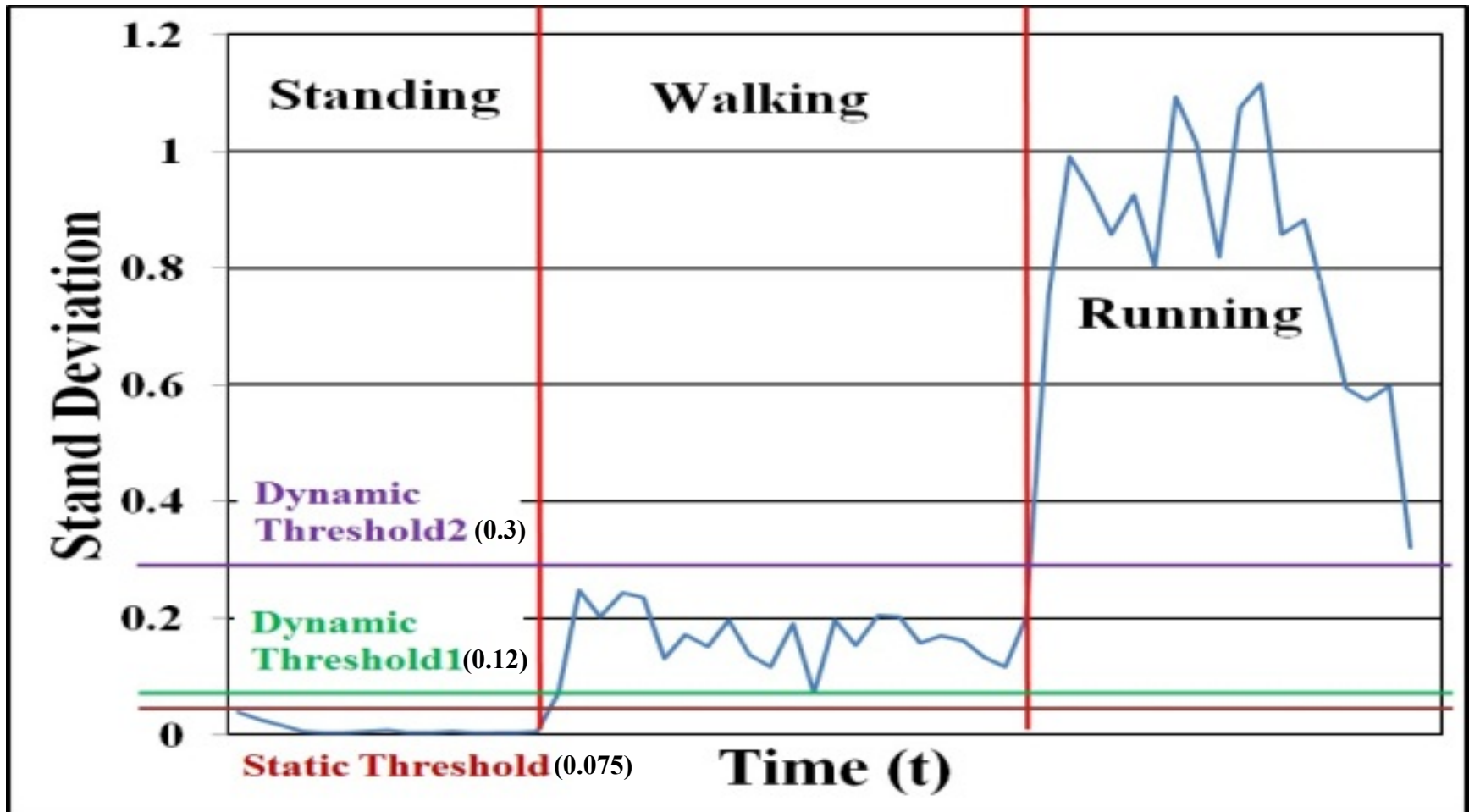




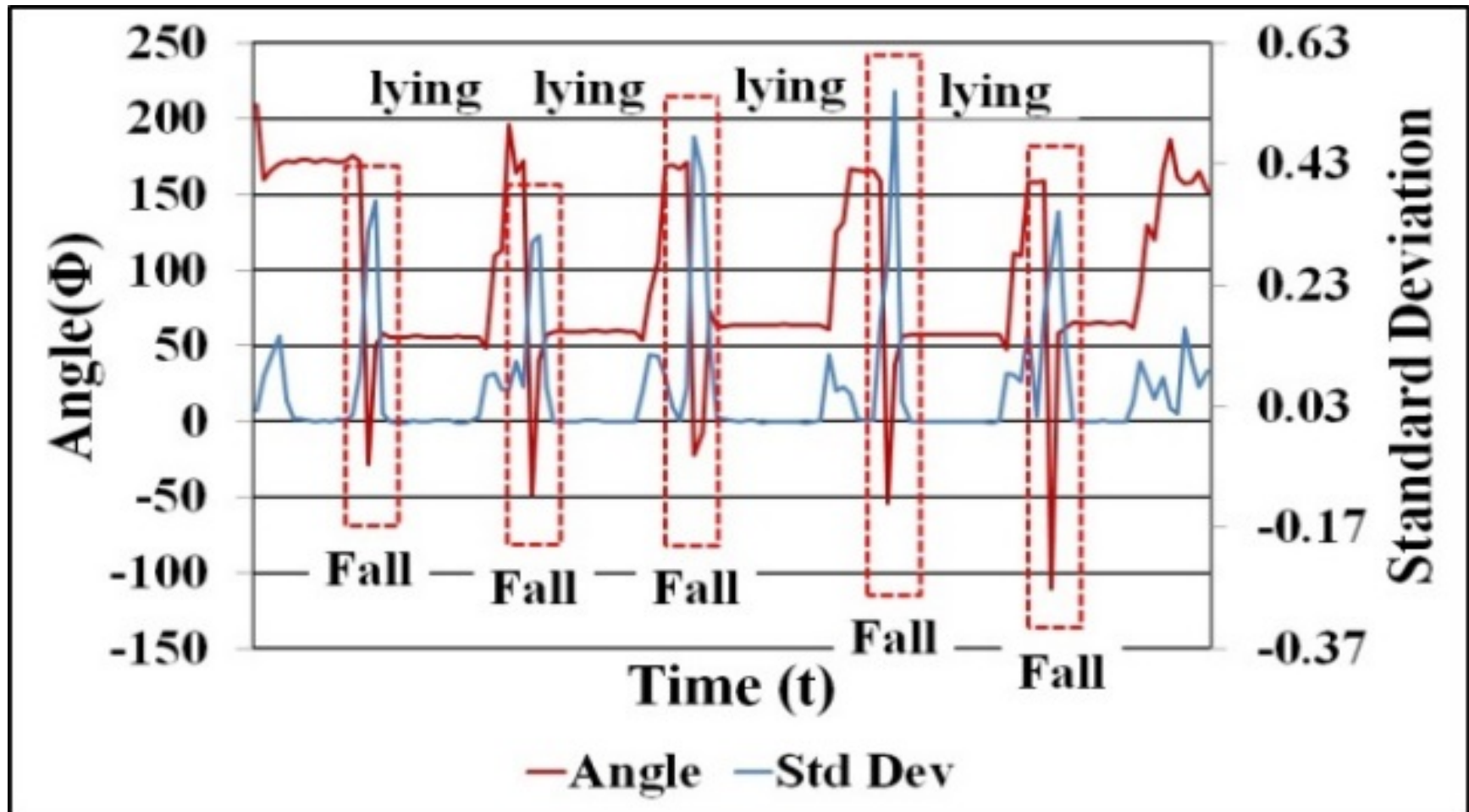
# Inclination Angle Value



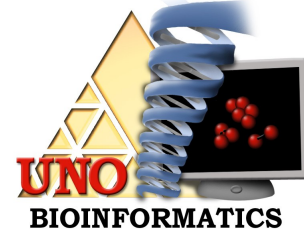
# SD of Standing, Walking, and Running



# Angle and SD of Fall and Lying

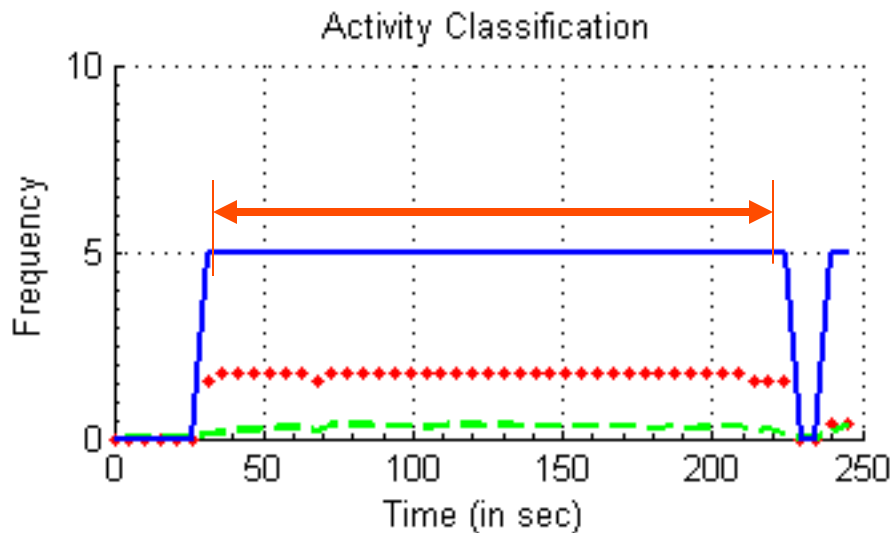


# Phase II: Classification of ADLs

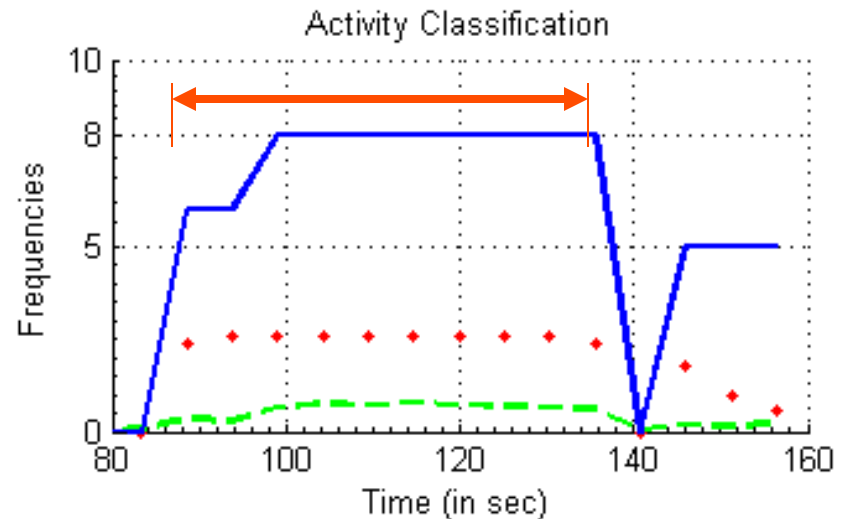


Walking vs. Running based on the magnitude and frequency of a peak in Fourier transform.

Legend: — Activity Index, ◆◆◆ Peak Frequency, - - - Peak Frequency Magnitude



walking



running

# Energy Conservation Issues

- Instead of sending acceleration values of x, y, and z axis to the base station, data collection and activity classification are done through on-board processing on Shimmer platform
- Sends only classified activity data to the base station
- Our data packet size is just one byte to represent classified data

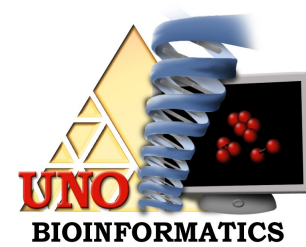
## Phase III: Mobility Profiles

- Mobility Profile (between the given start-time and end-time)
  - Total # of steps
  - Average # of steps per second
  - Activity level with different precisions
  - # of rooms traveled (will be added)

## Phase IV: Fall Prediction

- Fall can injure the elderly in large scale.
- 10-15% falls cause some serious physical injury in older people.
- The early prediction of fall is an important step to alert and protect the subject to avoid injury.
- We employ Hidden Markov Models for detection and prediction of anomalous movement patterns among the human subjects.

# Experiment Schedule



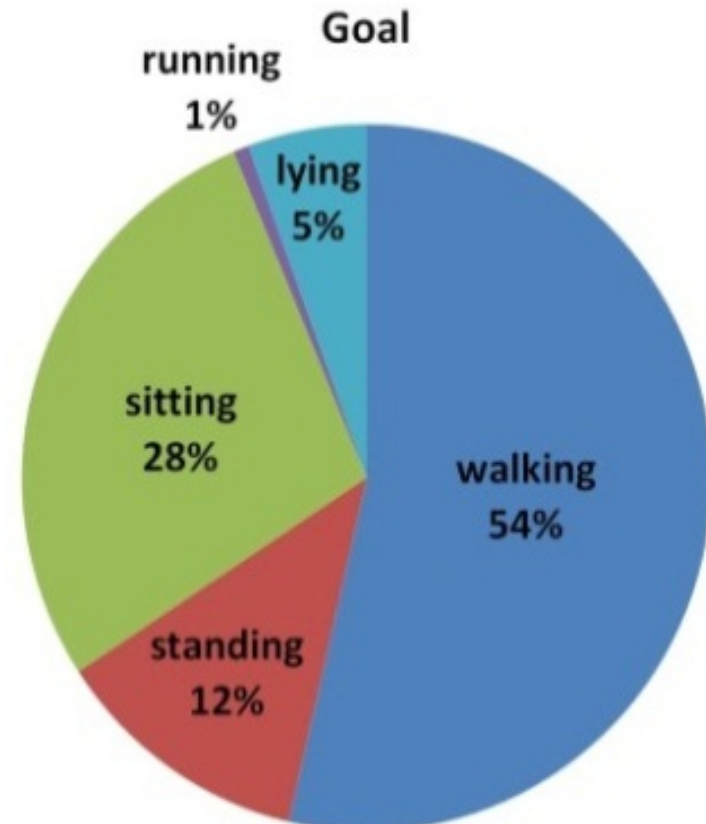
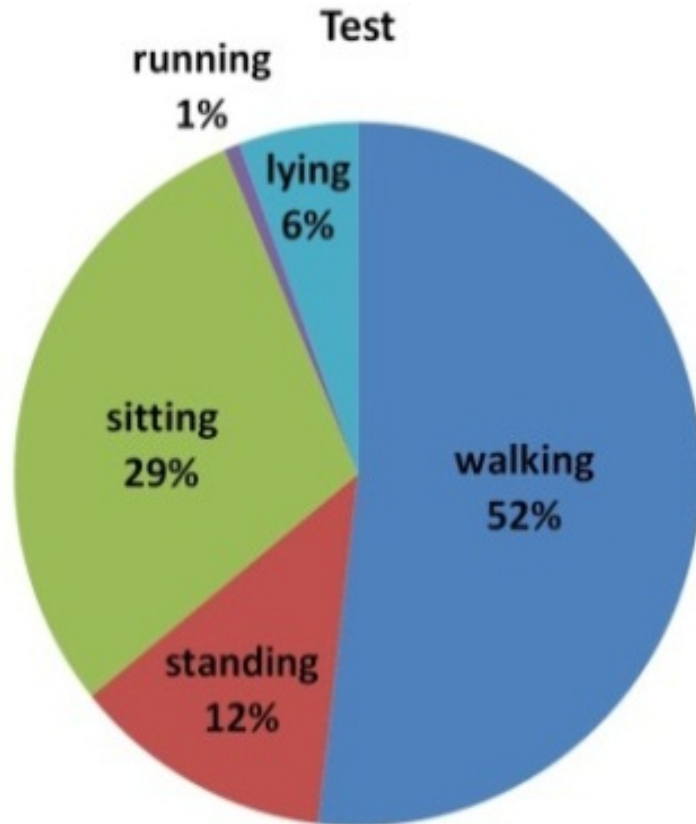
Activity	Place	Time (second)	
		Subject I	Subject II
Walking	Outdoor	360	300
Standing	Outdoor	20	20
Running	Outdoor	30	15
Standing	Outdoor	30	120
Walking	Outdoor	300	300
Sitting	Indoor	360	340
Walking	Indoor	30	45
Lying	Indoor	72	60
Total time		1200	1200



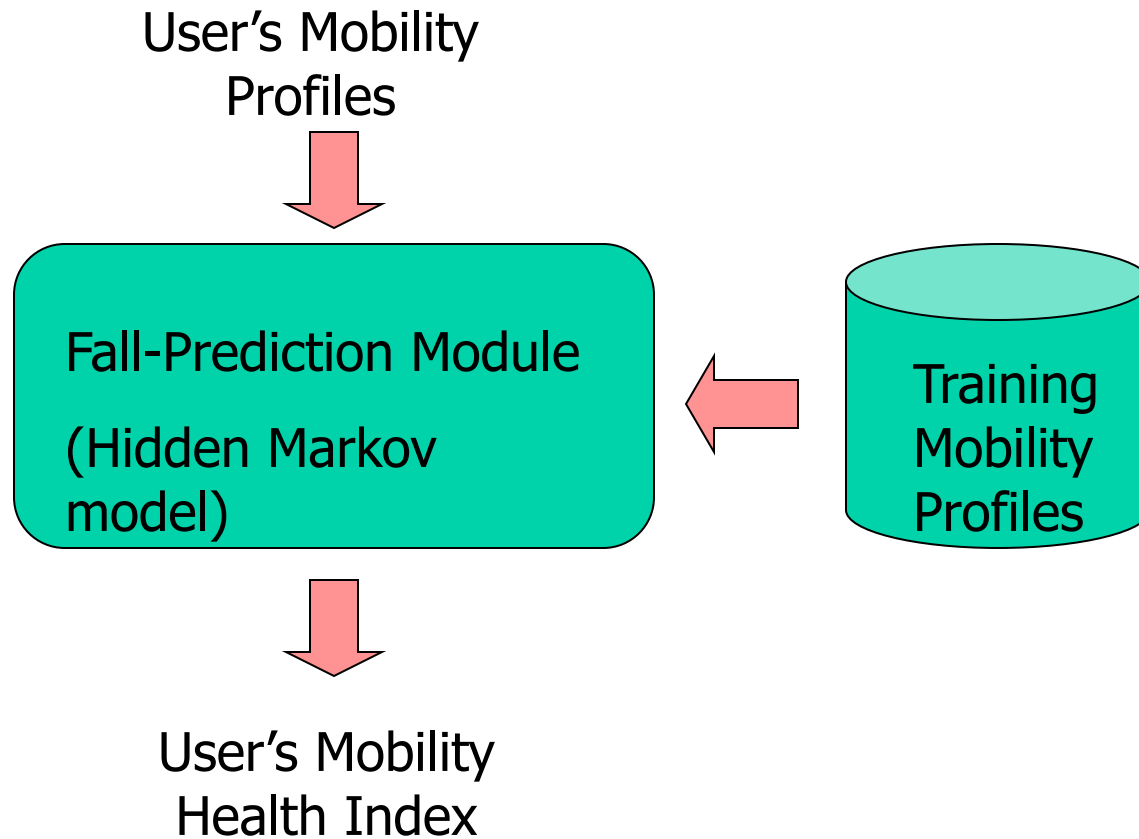
# Profiles: Experiment Results



# Profile: Experiment Results

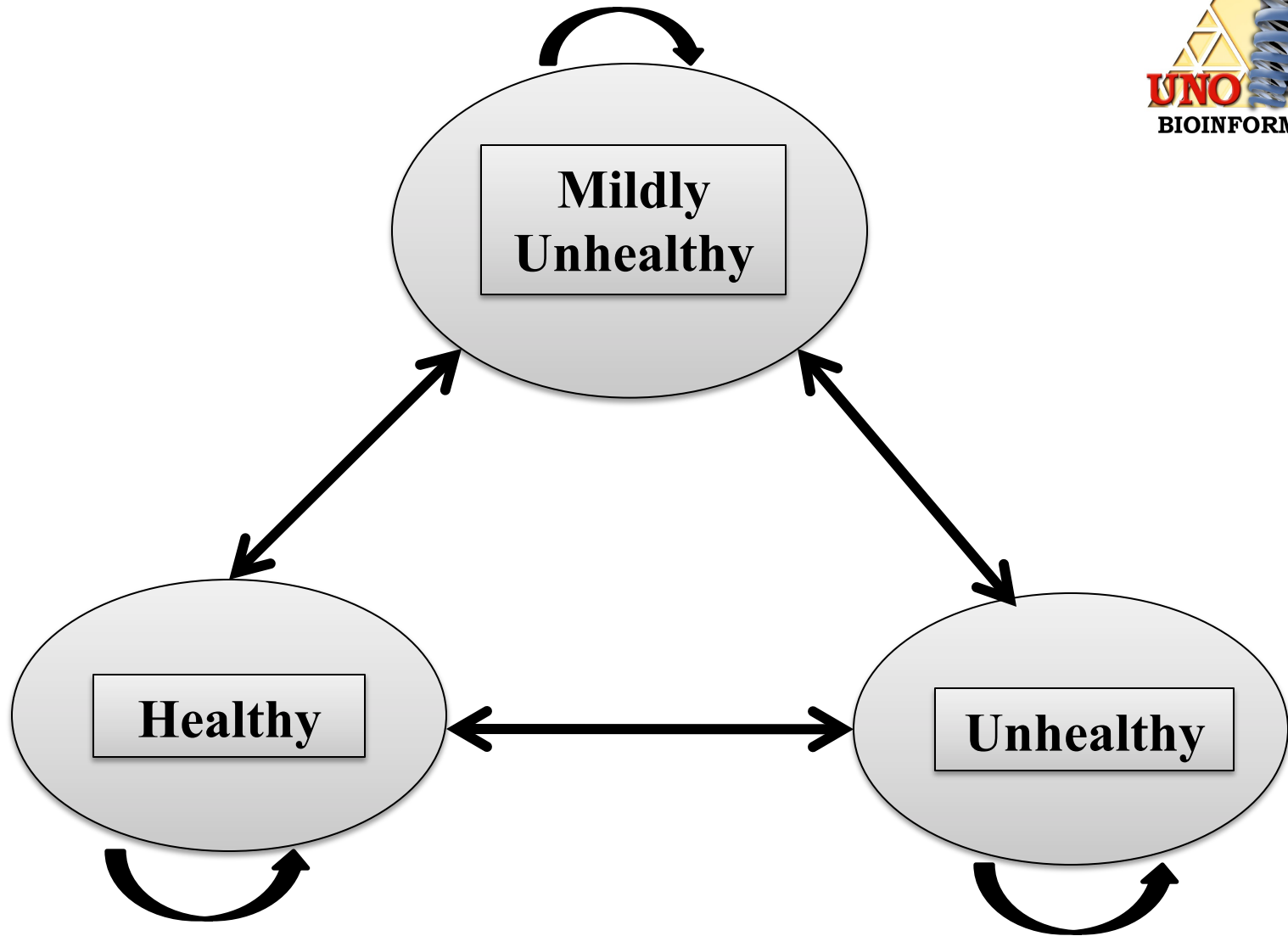


# Phase IV: Fall Prediction



# Prediction Model

- Develop a HMM model with
  - 3 states ( Healthy, Unhealthy, Mildly Unhealthy)
  - 3 parameters (#steps moved, #rooms visited, # movement in arms)
- Entire dataset was split into
  - Train data
  - Test data



THREE- STATE TRANSITION DIAGRAM

# Probability Calculations

- The model parameters for a HMM are generated from:
  - state transition probabilities  
 $a_{kl} = P(\pi_i = l \mid \pi_{i-1} = k)$ , which is probability from state  $k$  to state  $l$
  - emission probabilities  
 $e_l(b) = P(x_i = b \mid \pi_i = l)$ , which is probability distribution over all the possible output symbols  $b$  for each state  $l$ .

# Training

- For each observation the trained model parameters were calculated using maximum likelihood approach
- $a_{kl} = \frac{A_{kl}}{\sum_t A_{kl}}$
- $e_l(b) = \frac{E_l(b)}{\sum_b E_l(b)}$

# Predicting Hidden State

- With the trained model parameters
  - predict hidden state path for a new sequence of observations using Forward-Backward (Posterior) algorithm.
  - predicts the hidden, probable state path



# Assessment

- Performance evaluation
  - Accuracy
  - Sensitivity
  - Specificity

$$\text{specificity} = \frac{\text{number of True Negatives}}{\text{number of True Negatives} + \text{number of False Positives}}$$

$$\text{sensitivity} = \frac{\text{number of True Positives}}{\text{number of True Positives} + \text{number of False Negatives}}$$

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{numbers of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}$$

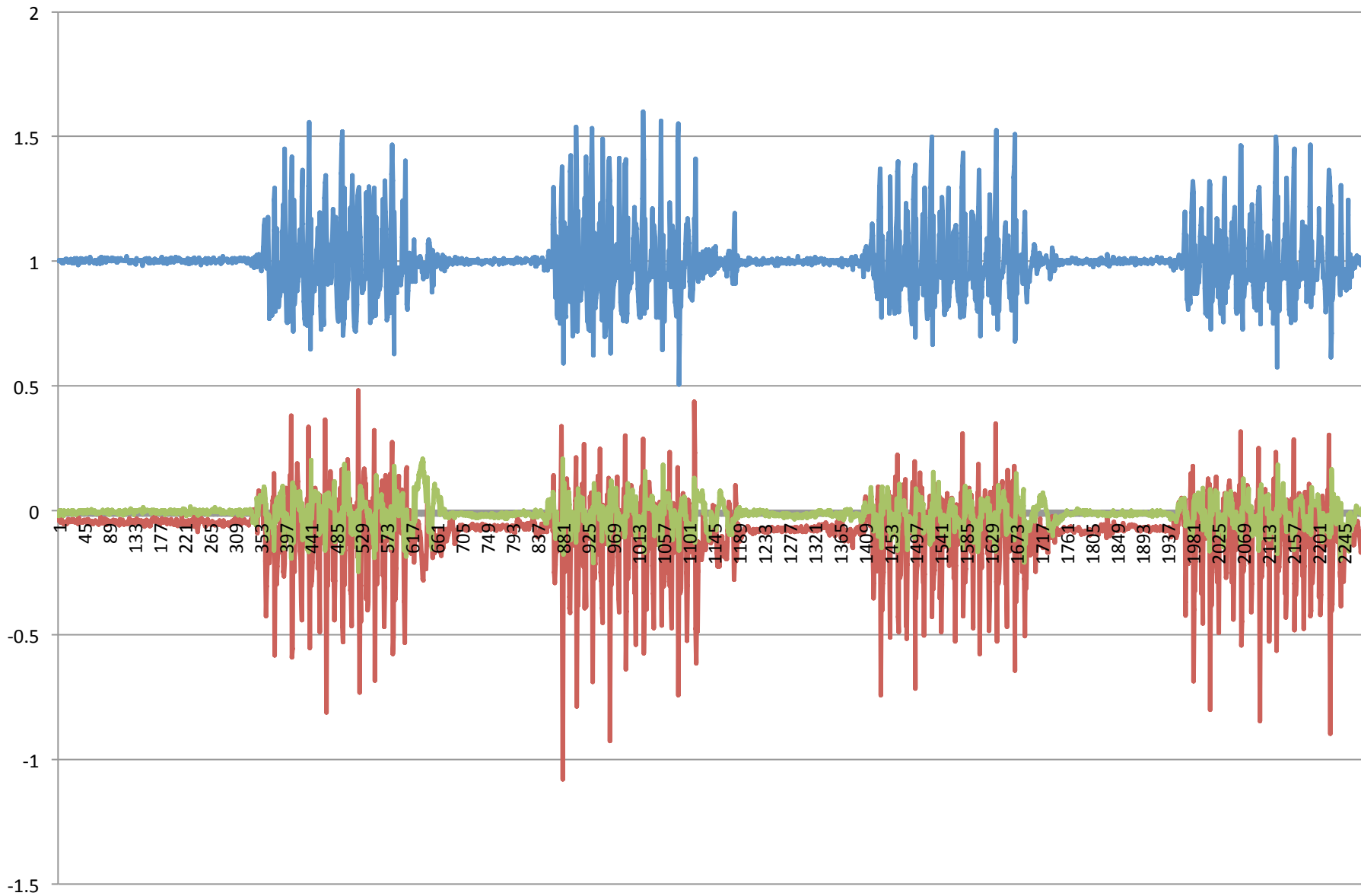
**Table: Showing True positives(TP), true negatives(TN), false positives(FP), false negatives(FN) and accuracy(ACC) values for predicting each state**

<b>State</b>	<b>#of states predicted</b>	<b>TP</b>	<b>FN</b>	<b>FP</b>	<b>TN</b>	<b>ACC</b>
<b>Unhealthy</b>	11	11	0	1	8	100.00
<b>Mildly Unhealthy</b>	7	6	1	2	11	85.71
<b>Healthy</b>	9	8	1	0	11	88.89

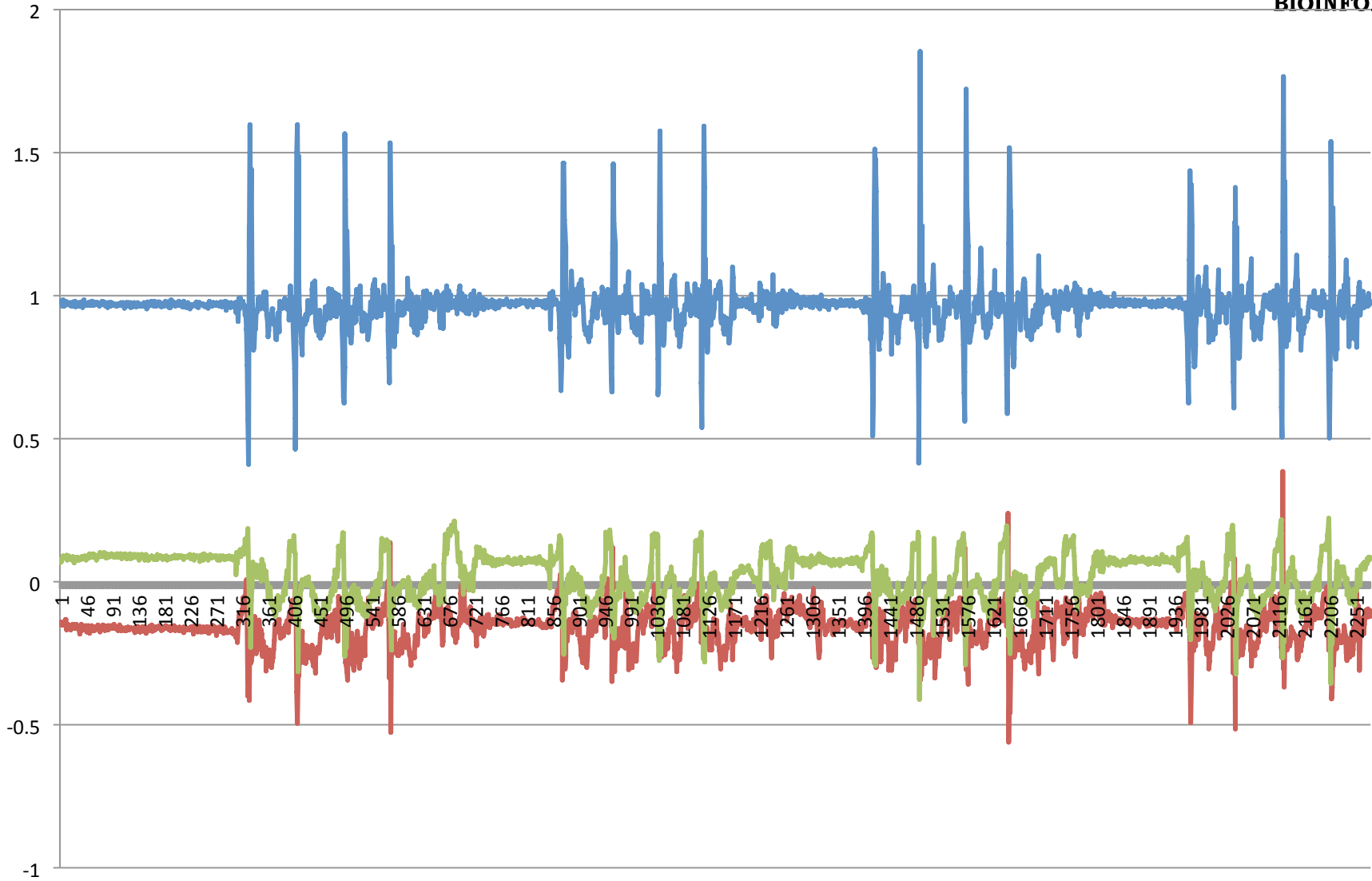
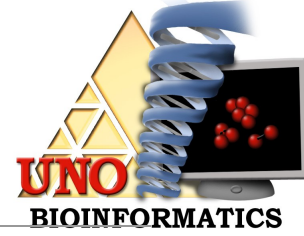
# Mobility Signatures

- Can we detect specific mobility signatures for specific stages associated with certain diseases or phases of recovery?
- Need to involve more parameters for training HMM including medications, psychological state, clinical observations, etc.

# Balanced Walking



# Unbalanced Walking



# Summary

- Proposed an on-board processing approach for classifying Activity of Daily Living using a triaxial acceleration sensor
- Implemented this mechanism on a tiny wireless sensor which is easy to wear and user-friendly
- Integrated all core functionalities such as an activity classification, wireless communication module, and data storage function into a single wireless sensor platform
- Signatures for disease or recovery from operations

# Tutorial Outlines

- Introduction to Biomedical Informatics
  - State of the discipline - Challenges and Opportunities
  - Data-driven biomedical research
- Next Generation Bioinformatics Tools
  - Intelligent Collaborative Dynamic (ICD) Tools
- *Case Study: Aging Research*
  - The genomic study: Correlation Networks
  - Mobility and aging: Wireless monitoring
  - *Data collection and Virtual Environments*
- Next Steps: Where do we go from here?
  - HPC and Cloud Computing

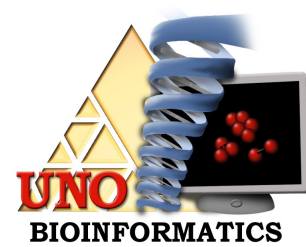






**OMAHA**  
TECHNOLOGY

# Aging, Mobility and Wireless Networks

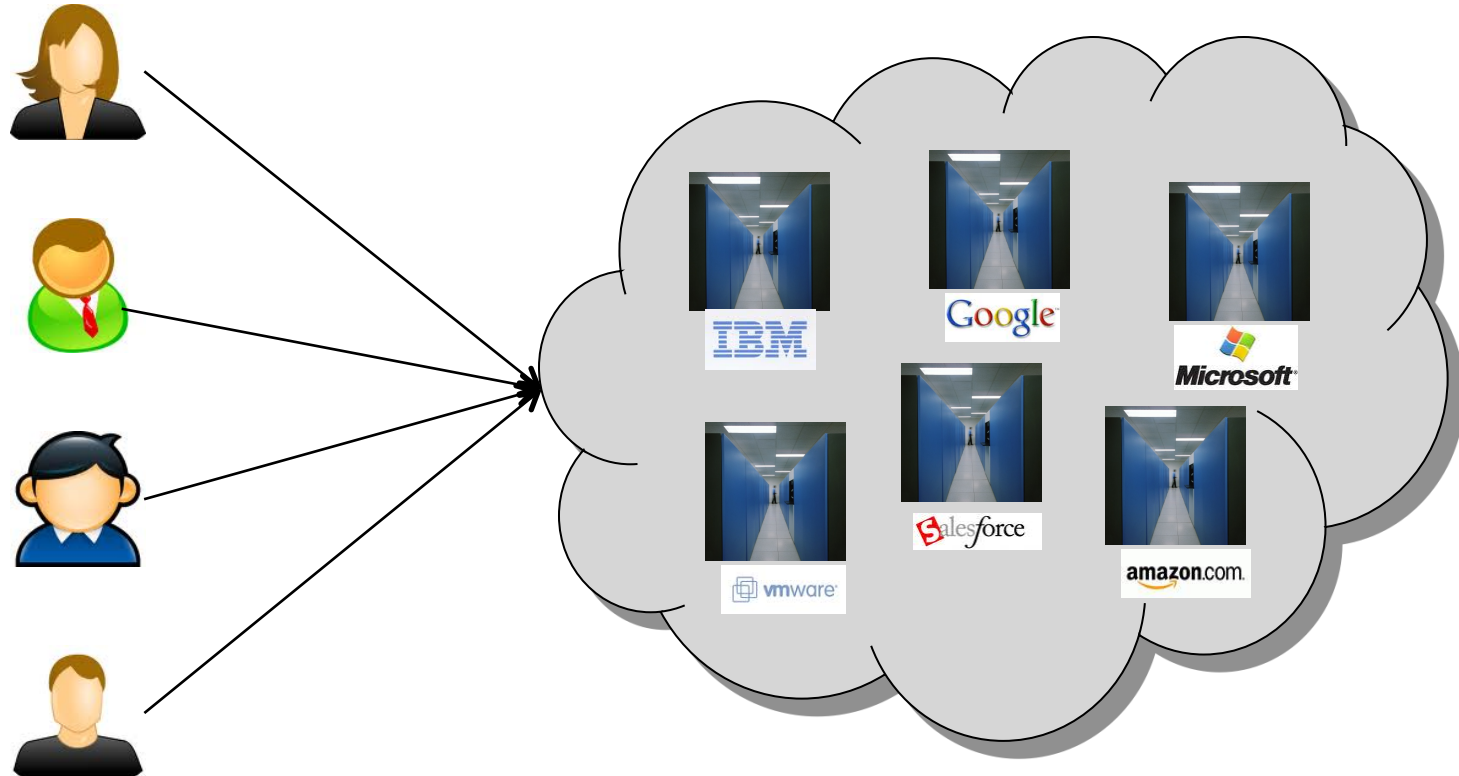


- Correlation between mobility and health level
- Monitoring mobility levels
- Aging of cells and aging of systems
- Collaboration between Bioinformatics group, Wireless Networks group and Decision Support Systems group

# Tutorial Outlines

- Introduction to Biomedical Informatics
  - State of the discipline - Challenges and Opportunities
  - Data-driven biomedical research
- Next Generation Bioinformatics Tools
  - Intelligent Collaborative Dynamic (ICD) Tools
- Case Study: Aging Research
  - The genomic study: Correlation Networks
  - Mobility and aging: Wireless monitoring
  - Data collection and Virtual Environments
- *Next Steps: Where do we go from here?*
  - *HPC and Cloud Computing*

# Next Step: Cloud Computing – Lifting the Veil



# Energy Management – Research Vision



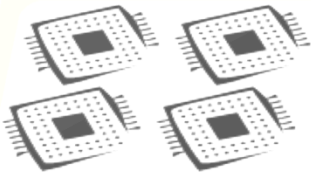
## Energy Adjustment @ Cloud Level

- Using Energy Index for Datacenters



## Energy Adjustment @ HPC Level

- Using Adjustments to # of Nodes

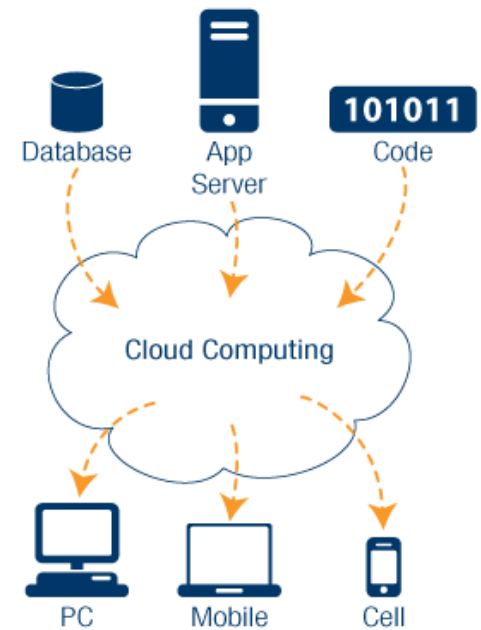


## Energy Adjustment @ Node Level

- Using Dynamic Voltage Scaling (DVS)

# 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





# Cost Reduction & Convenience



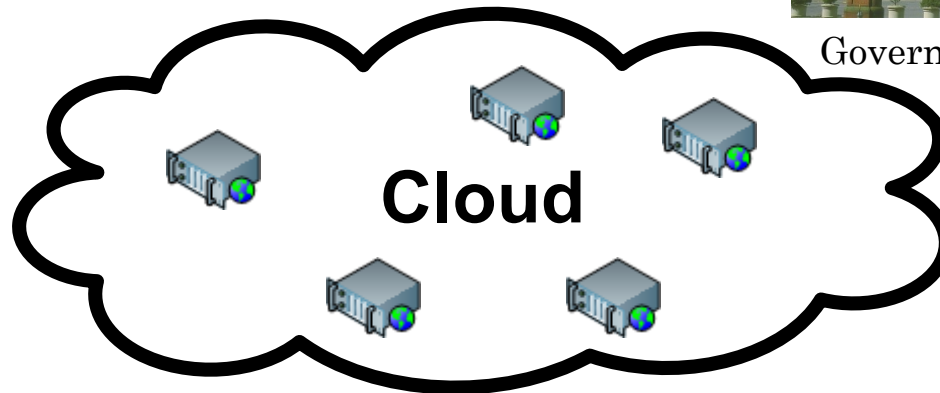
Small Business



Government Offices



Multinational  
Corporations



Homes

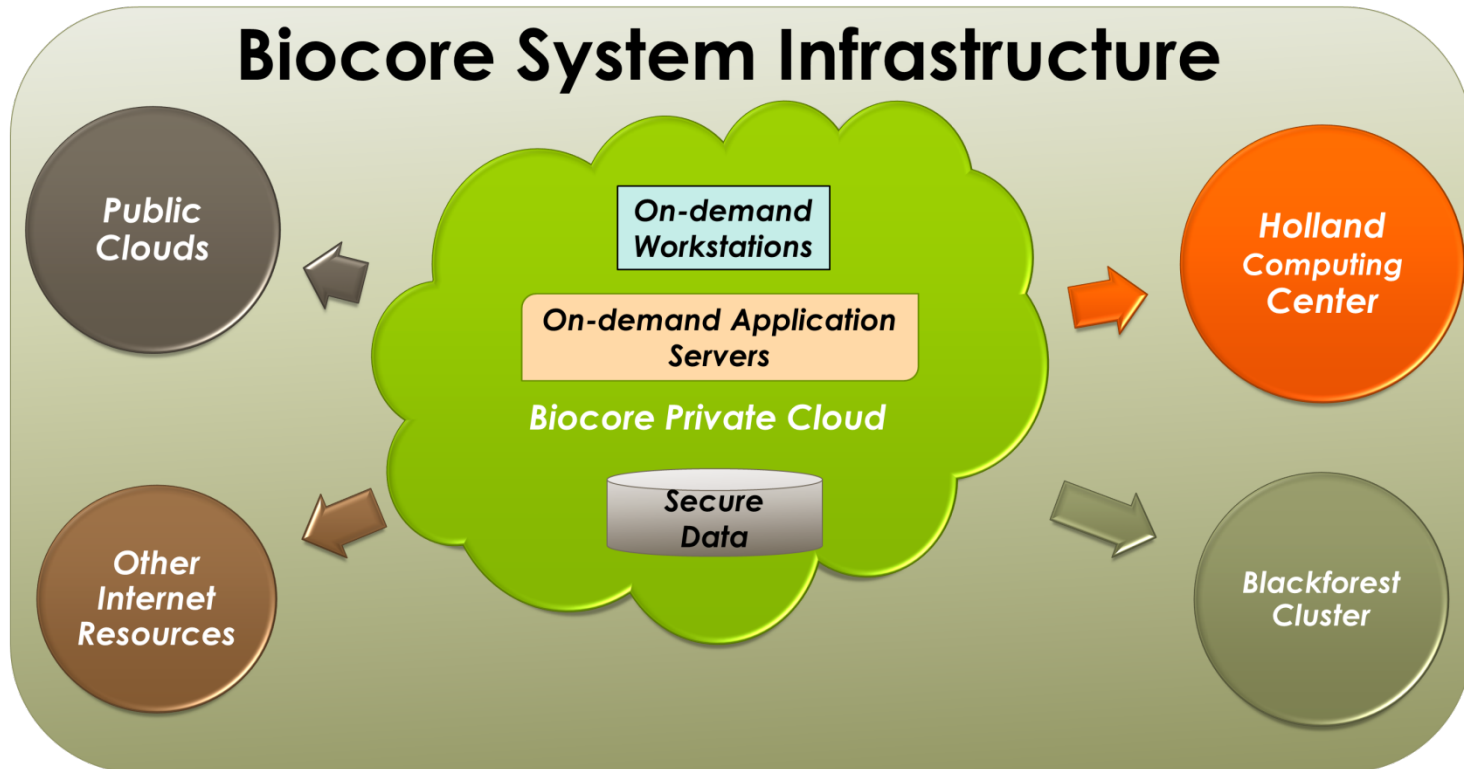
- Flexible availability of resources
- Opportunity for developers to easily push their applications
- Targeted advertising
- Easy Software Upgrades for customers
  - Example: Webmail

# Private Clouds

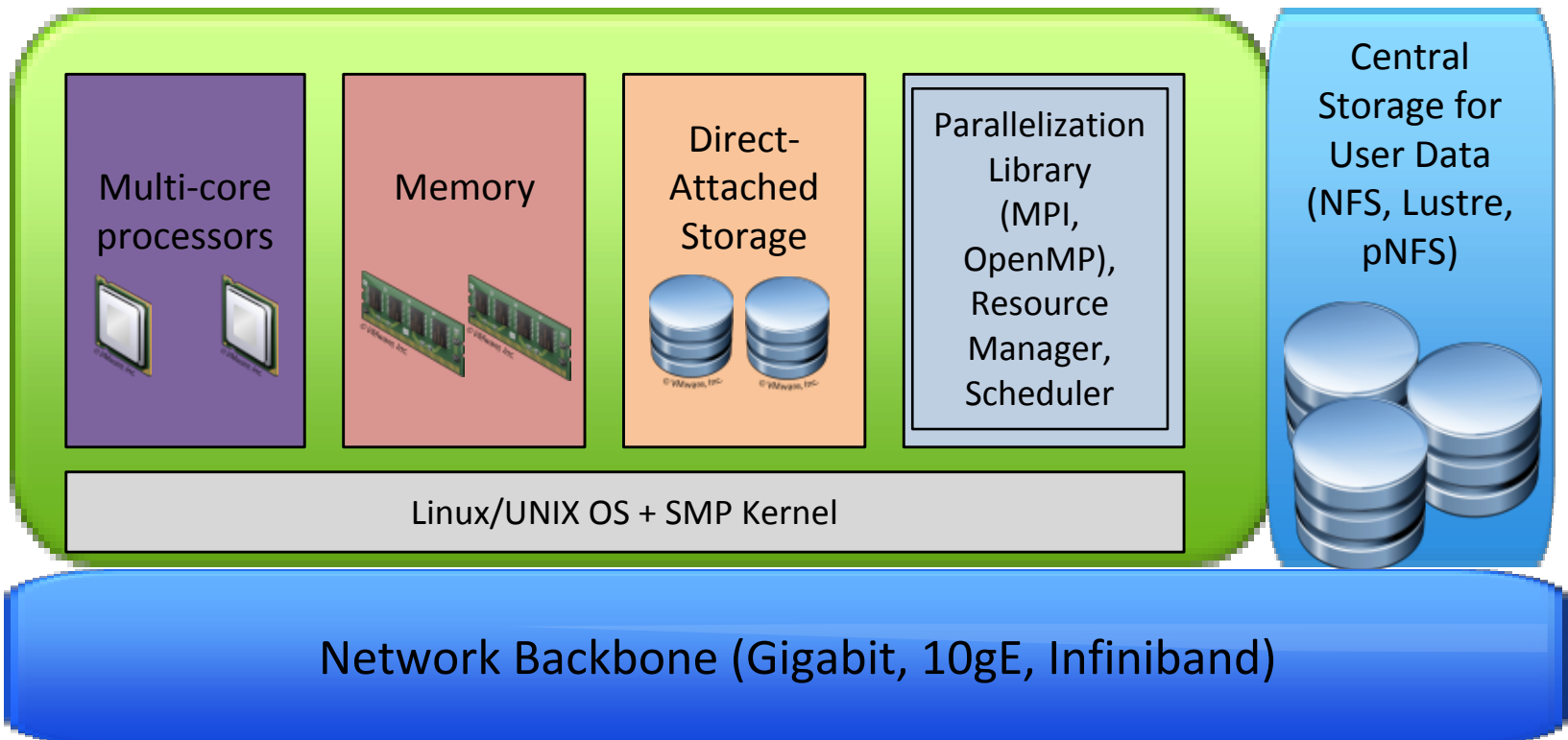
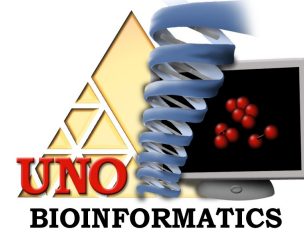
- Core facilities need to acquire private infrastructure-level virtual cloud technology. Best vendor for such technology is VMware. The Bioinformatics Core facility at UNO uses *VMware vSphere Enterprise*.
- Public Clouds like Amazon EC2, RackSpace cannot be used in all cases due to various restrictions 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.




# Proposed Model




# HPC at UNO




Multi-core processors



Memory



Direct-Attached Storage



Parallelization Library (MPI, OpenMP), Resource Manager, Scheduler

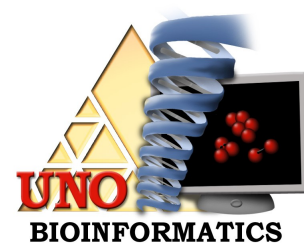
Linux/UNIX OS + SMP Kernel

Central Storage for User Data (NFS, Lustre, pNFS)

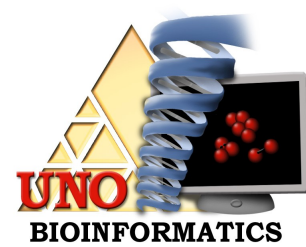


Network Backbone (Gigabit, 10gE, Infiniband)

# Conclusions



- Next Generation Bioinformatics Tools need to be Intelligent, Collaborative, and Dynamic
- Biomedical scientists, Bioinformatics researchers and computer scientists need to work together to best utilize the combination of tools development and domain expertise
- HPC is critical to the success of the next phase of Biomedical research but again the integration needs to happen at a deeper level
- The outcome of collaboration has the potential of achieving explosive results with significant impact on human health and overall understanding of biological mysteries



# Acknowledgments

- UNO Bioinformatics Research Group
  - Kiran Bastola
  - Sanjukta Bhoomwick
  - Kate Dempsey
  - Ramez Mena
  - Sachin Pawaskar
  - Joe Steele
  - 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