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. Stage 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

• Average microarray experiment: 1200 pages of data*

• How can we extract information from data?

Each sample can have over 40,000 genes

* http://www.mc.vanderbilt.edu/peerreview/fall026.html
Interactive Aspect of BMI

Lead to new data analysis problems

Lead to new experiments

Useful Knowledge

• Biology
• Genomics
• Microbiology
• Molecular Biology
• Biochemistry

• Computer Science
• Mathematics
• Statistics
• Data Mining
• Parallel Computing
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.
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
“Big Data”

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

Data Velocity

Data Volume

Data Variety

- Social
- Video
- unstructured
- Mobile
- Audio
- Web
- Batch
- Terabyte (TB)
- Petabyte (PB)
- Gigabyte (GB)
- Megabyte (MB)

- Real-Time
- near Real-Time
- Periodic

Image Source: http://api.ning.com/files/O6-JQcfS6sxRzu/8f22/5a/Vx59d-kT/a6UpnRqMvC/0w/R0jy2aDPYsL-AboHiZe-Vr8gPv5jFBlhDIH/BigData.001.jpg
A Potential Major Change

- Data driven research vs. Hypothesis driven research

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
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

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.
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

Simple Tools – Infrastructure

Innovative Approaches – Integrative Skills - Culture
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
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. Stage 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

<table>
<thead>
<tr>
<th>Site contents</th>
<th>Public data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platforms</td>
<td>9,267</td>
</tr>
<tr>
<td>Samples</td>
<td>611,215</td>
</tr>
<tr>
<td>Series</td>
<td>24,571</td>
</tr>
<tr>
<td>DataSets</td>
<td>2,720</td>
</tr>
</tbody>
</table>

- Average microarray experiment: 1200 pages of data*

- How can we extract information from data?

* http://www.mc.vanderbilt.edu/peerreview/fall026.html
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.

Source: Omaha World Herald,
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

*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

ABI PRISM 3100 Genetic Analyzer
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

Infectious Diseases

Various strains of HIV are now aggressively targeted by specifically tailored treatments.

In the near future, medicine will be able to be personalized according to an individual’s genetic makeup, resulting in better treatment response and fewer treatment side effects.

Environmental Genomics

The ocean revealed some of its secrets as novel microorganisms were collected and sequenced in the global ocean sampling expedition (Craig Venter Institute).

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

GGATCCATGGGATAGGATAATGGA
GATAATGGGATAGAGGATCCATGG
ATGGGATAGAGGATCCATGGCTAG
ATGGGATAGGGATTATGGGATAGAGG
GATAGAGGATCCATGGCTAGATC

Overlapping

GGATCCATGGGATAGGATAATGGA
GATAATGGGATAGAGGATCCATGG
ATGGGATAGAGGATCCATGGCTAG
ATGGGATAGGGATTATGGGATAGAGG
GATAGAGGATCCATGGCTAGATC

Contigs

GGATCCATGGGATAGGATAATGGGATAGAGGATCCATGGCTAGATC
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\textcolor{red}{T}AGGGTTTGAT
  TTACATTAGG\textcolor{red}{C}CG

  - Is this a large enough fragment overlap?
    ATTAGTTAGTTAGATTAC
    \ldots\textcolor{red}{G}GCATTA

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

- Biological expertise
- Mathematical modeling
- Computational algorithms
- Data Management
A Graph Theoretic Model for Assembly

• The algorithm maps each read to a node (vertex) and every overlap relationship to edges.
Sequencing and Interval Graphs

- 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, \( t_i = 2 \)**

- 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

Fragments

1. AGTAGCGGGC
2. CGGAATCGATAGC
3.
4. AATCGATAGCAA
5.

Genomic Sequence

1. AGTAGCGGGGCTAGATCGGA
2. AGTAGCGGGCTAGATCGGAATCGATAGCAA
3.
4. CGGAATCGATAGC
5.

2. 4.

1. 3.

4

2. 4.

5

5
Assembly Algorithm

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.

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.

Read_File.txt

TGGAC
ACCAA
AACTG
ATCAA
TTCAA
GTTCA
TCAAT

Graph Traversal

1. GTTCAAT
2. ATCAA
3. TGGAC
4. ACCAA
5. AACTG

Node Merging

TTCAA
GTTCA
TCAAT

Final Output

>contig 1
GTCAATCAACTGACCACTG
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

\[ Z_{\text{score\_degree}}(u) = \frac{(\text{node\_degree}(u) - \text{average\_node\_degree})}{\text{node\_degree\_standard\_deviation}} \]

\[ Z_{\text{score\_read\_length}}(u) = \frac{(\text{read\_length}(u) - \text{average\_read\_length})}{\text{read\_length\_standard\_distribution}} \]

\[ Z_{\text{score\_minoverlap}}(u) = \frac{(\text{minoverlap\_length} - \text{average\_overlap\_length})}{\text{overlap\_length\_standard\_deviation}} \]
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?

<table>
<thead>
<tr>
<th>N50 (bps)</th>
<th>Read Length Weight</th>
<th>Degree Weight</th>
<th>Domain Specific Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>12978</td>
<td>0.1</td>
<td>0.1</td>
<td>0.8</td>
</tr>
<tr>
<td>13677</td>
<td>0.1</td>
<td>0.3</td>
<td>0.6</td>
</tr>
<tr>
<td>13406</td>
<td>0.1</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>12549</td>
<td>0.3</td>
<td>0.1</td>
<td>0.6</td>
</tr>
<tr>
<td>16761</td>
<td>0.3</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>16464</td>
<td>0.3</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>12095</td>
<td>0.5</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>15988</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>15427</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
</tr>
</tbody>
</table>
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

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.

<table>
<thead>
<tr>
<th>Sequence Source</th>
<th>Length (bp)</th>
<th>GC Content</th>
<th>Repeat Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escherichia coli</td>
<td>79237 bp</td>
<td>47.27%</td>
<td>0.09%</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>79745 bp</td>
<td>42.47%</td>
<td>3.21%</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>79590 bp</td>
<td>37.71%</td>
<td>9.75%</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>80914 bp</td>
<td>43.08%</td>
<td>49.35%</td>
</tr>
</tbody>
</table>

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

*Escherichia coli* Assembly

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

A T C G 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)

Sequence

GenBank

Alignment Sequence

Proposed Method

Other Methods
(word, Suffix/sub-sequence)

Identification
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

\[ S = AACGTTACCATTG \quad R = CTAGGGACTTTAT \]
\[ Q = ACGGTCACCAA \]

\[ H_E(S) = A/AC/G/T/ACC/AT/TG \]
\[ H_E(R) = C/T/A/G/GGA/CTT/AT \]
\[ H_E(Q) = A/C/G/GT/CA/CC/AA \]

- \[ H_E(SQ) = A/AC/G/T/ACC/AT/TG/ACGG/TC/ACCAA \]
- \[ H_E(RQ) = C/T/A/G/GGA/CTT/AT/ACG/GT/CA/CC/AA \]

\[ c(SQ) - c(S) = 3 \quad \quad c(RQ) - c(R) = 5 \]

\( \Rightarrow \) 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

- Gold standard tree
- MSA tree
- Algorithm tree

Distance

- Problems with visual inspection!

- Computational ways?
  - Accurate
  - Fast

The distance between trees
Path-Difference Length
Distance Between Trees

![Distance Between Trees Diagram]

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A'</th>
<th>B'</th>
<th>C'</th>
<th>D'</th>
</tr>
</thead>
<tbody>
<tr>
<td>A'</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>B'</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>C'</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>D'</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
## Results

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

<table>
<thead>
<tr>
<th>Test Algorithm</th>
<th>Variant</th>
<th>Neighbor-Joining</th>
<th>UPGMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kolmogorov using Huffman coding</td>
<td>CD</td>
<td>2.395244</td>
<td>3.169468</td>
</tr>
<tr>
<td></td>
<td>NCD</td>
<td>2.328382</td>
<td>2.264505</td>
</tr>
<tr>
<td></td>
<td>UCD</td>
<td>2.328382</td>
<td>2.264505</td>
</tr>
<tr>
<td>Kolmogorov using LZW compression</td>
<td>CD</td>
<td>2.176959</td>
<td>2.165911</td>
</tr>
<tr>
<td></td>
<td>NCD</td>
<td>2.210704</td>
<td>2.215544</td>
</tr>
<tr>
<td></td>
<td>UCD</td>
<td>2.305268</td>
<td>2.238781</td>
</tr>
<tr>
<td>Lempel - Ziv complexity</td>
<td>Distance 1</td>
<td>2.337454</td>
<td>2.26598</td>
</tr>
<tr>
<td></td>
<td>Distance 2</td>
<td>2.248862</td>
<td>2.192803</td>
</tr>
<tr>
<td></td>
<td>Distance 3</td>
<td>2.244591</td>
<td>2.284809</td>
</tr>
<tr>
<td></td>
<td>Distance 4</td>
<td>2.222918</td>
<td>2.371806</td>
</tr>
<tr>
<td>Multiple Sequence Alignment</td>
<td></td>
<td>2.182934</td>
<td>2.371806</td>
</tr>
</tbody>
</table>
Comparisons of the compression algorithms and multiple sequence alignment for the Mitochondrial genome dataset

<table>
<thead>
<tr>
<th>Test Algorithm</th>
<th>Variant</th>
<th>Neighbor-Joining</th>
<th>UPGMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kolmogorov using Huffman coding</td>
<td>CD</td>
<td>7.871585</td>
<td>7.871585</td>
</tr>
<tr>
<td></td>
<td>NCD</td>
<td>7.871582</td>
<td>7.871582</td>
</tr>
<tr>
<td></td>
<td>UCD</td>
<td>7.871582</td>
<td>7.871582</td>
</tr>
<tr>
<td>Kolmogorov using LZW compression</td>
<td>CD</td>
<td>3.034474</td>
<td>3.034474</td>
</tr>
<tr>
<td></td>
<td>NCD</td>
<td>2.797647</td>
<td>2.797647</td>
</tr>
<tr>
<td></td>
<td>UCD</td>
<td>2.878755</td>
<td>2.878755</td>
</tr>
<tr>
<td>Lempel Ziv complexity</td>
<td>Distance 1</td>
<td>1.357058</td>
<td>1.357058</td>
</tr>
<tr>
<td></td>
<td>Distance 2</td>
<td>1.357058</td>
<td>1.357058</td>
</tr>
<tr>
<td></td>
<td>Distance 3</td>
<td>1.357058</td>
<td>1.357058</td>
</tr>
<tr>
<td></td>
<td>Distance4</td>
<td>1.357058</td>
<td>1.357058</td>
</tr>
<tr>
<td>Multiple Sequence Alignment</td>
<td></td>
<td>1.5547053</td>
<td>1.878762</td>
</tr>
</tbody>
</table>
Comparisons of the compression algorithms and multiple sequence alignment for the Mitochondrial genomes

<table>
<thead>
<tr>
<th>Test Algorithm</th>
<th>Variant</th>
<th>Neighbor-Joining</th>
<th>UPGMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kolmogorov using Huffman coding</td>
<td>CD</td>
<td>12.3431994</td>
<td>17.8482358</td>
</tr>
<tr>
<td></td>
<td>NCD</td>
<td>12.3431994</td>
<td>17.8482337</td>
</tr>
<tr>
<td></td>
<td>UCD</td>
<td>12.3431994</td>
<td>17.8482337</td>
</tr>
<tr>
<td>Kolmogorov using LZW compression</td>
<td>CD</td>
<td>10.5262895</td>
<td>10.5263158</td>
</tr>
<tr>
<td></td>
<td>NCD</td>
<td>6.4460256</td>
<td>10.5263158</td>
</tr>
<tr>
<td></td>
<td>UCD</td>
<td>9.8464668</td>
<td>10.5263158</td>
</tr>
<tr>
<td>Lempel Ziv complexity</td>
<td>Distance 1</td>
<td>0.0000000</td>
<td>0.0000000</td>
</tr>
<tr>
<td></td>
<td>Distance 2</td>
<td>0.0000000</td>
<td>0.0000000</td>
</tr>
<tr>
<td></td>
<td>Distance 3</td>
<td>0.0000000</td>
<td>0.0000000</td>
</tr>
<tr>
<td></td>
<td>Distance 4</td>
<td>0.0000000</td>
<td>0.0000000</td>
</tr>
<tr>
<td>Multiple Sequence Alignment</td>
<td></td>
<td>6.4460256</td>
<td>0.0000000</td>
</tr>
</tbody>
</table>
Using Compression to Compare Incomplete Fragments of Genomes

- Whole Genome
  - Incomplete one fragment 10%
  - Incomplete one fragment 50%
  - Incomplete one fragment 90%

- Whole Genome
  - Incomplete two fragments (10% of the whole genome)
  - Incomplete three fragments (50% of the whole genome)
  - Incomplete four fragments (90% of the whole genome)

- Incomplete four fragments in order (90% of the whole genome)
- Incomplete four fragments NOT in order (90% of the whole genome)
Results of Experiment

fragments of genomes, not continuous, not ordered

![Graph showing the results of various experiments.](image)
Results of Experiment

Fragments of genomes, not continuous

Tree distance

Percentages of genomes
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.

Source: Omaha World Herald,
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

- Explosion of data from high-throughput 'omics technologies
  - 1 gene expression experiment = 1200 pages of data

Pathway Commons Quick Stats:
- Number of Pathways: 1,623
- Number of Interactions: 585,237
- Number of Physical Entities: 105,949
- Number of Organisms: 564
Motivation:

• Provide interactive view of cellular systems
• Target mechanisms behind disease and aging
• Re-define personalized medicine

Costanzo et al. 2010
Correlation Networks

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

0

Gene ID
A_52_P616356
A_52_P580582
A_52_P403405
A_52_P819156
A_51_P331831
A_51_P430630
A_52_P502357
A_52_P299964
A_51_P356389
A_52_P684402

Sample 1  Sample 2  Sample 3  Sample 4
5.9813   6.0079   5.9525   7.2753   6.7275
7.7845   7.7512   8.0943   8.3608   8.1707
5.9301   6.5153   6.0526   7.1707   6.1707
7.1732   7.8754   7.7632   8.1875   7.6875
6.0661   6.4009   5.9525   7.1208   6.6208
5.936    6.3206   5.9525   7.1819   6.6819
6.3452   6.8025   6.6457   7.3445   6.7445
6.5088   7.0545   7.2346   7.631    7.131

Gene 1
Gene 2
Gene 1
Gene 2

Correlation = 1
Correlation Network

Correlation = -1

Correlation = 0

Correlation = 1
Correlation Networks

A graph model that examines the degree of correlation over some biological entity

<table>
<thead>
<tr>
<th></th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>10.5</td>
<td>11.0</td>
<td>12.1</td>
</tr>
<tr>
<td>Gene 2</td>
<td>3.2</td>
<td>3.3</td>
<td>2.9</td>
</tr>
<tr>
<td>Gene 3</td>
<td>1.4</td>
<td>1.5</td>
<td>0.9</td>
</tr>
<tr>
<td>Gene 4</td>
<td>7.8</td>
<td>7.1</td>
<td>8.2</td>
</tr>
</tbody>
</table>
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
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
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

<table>
<thead>
<tr>
<th>Strain</th>
<th>Gender</th>
<th>Tissue Type</th>
<th>Ages</th>
</tr>
</thead>
<tbody>
<tr>
<td>BalbC</td>
<td>Male</td>
<td>Hypothalamus</td>
<td>Young, mid-age, aged</td>
</tr>
<tr>
<td>CBA</td>
<td>Male</td>
<td>Hypothalamus</td>
<td>Young, mid-age, aged</td>
</tr>
<tr>
<td>C57_J20</td>
<td>Male</td>
<td>Hypothalamus</td>
<td>Young, aged</td>
</tr>
<tr>
<td>BalbC</td>
<td>Female</td>
<td>Hypothalamus</td>
<td>Young, aged</td>
</tr>
<tr>
<td>BalbC</td>
<td>Female</td>
<td>Frontal cortex</td>
<td>Young, aged</td>
</tr>
</tbody>
</table>
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 $\rightarrow$ 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
Integrated Data Model

Network Structure

Functionality Databases
- MGI
- OMIM
- Gene Ontology
- KEGG Pathway
- RefSeq

Network Enrichment

Target identification; Gene prediction

Query for essential

Query for function

Target of future study with predicted transport function
Correlation Networks

- Young edges
- Mid edges
- Aged edges
Control

Treated mice

Case Study: Aging

Young

Midage
## Results Validation

### Table 1
Comparison of phenotypes between klotho deficient and klotho overexpression mice

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Klotho deficient mice</th>
<th>Klotho overexpression mice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Body weight</td>
<td>Showing growth retardation and becoming inactive and marantic at 3 to 4 weeks of age (Kuro-o et al., 1997).</td>
<td>Normal (Kurosu et al., 2005)</td>
</tr>
<tr>
<td>Average lifespan</td>
<td>About 2 months (vs 2.5 to 3 years for wild-type mice) (Kuro-o et al., 1997).</td>
<td>About 20–30% longer than wild-type mice (Kurosu et al., 2005).</td>
</tr>
<tr>
<td>Maximal lifespan</td>
<td>Less than 100 days (Kuro-o et al., 1997).</td>
<td>More than 936 days (Kurosu et al., 2005).</td>
</tr>
<tr>
<td>Insulin</td>
<td>Decreased insulin secretion and enhanced insulin sensitivity (Kuro-o et al., 1997).</td>
<td>Increased resistance to insulin and IGF-1 signaling (Kurosu et al., 2005).</td>
</tr>
<tr>
<td>Phosphorus homeostasis</td>
<td>Hyperphosphatemia (Kuro-o et al., 1997).</td>
<td>Normal (Kurosu et al., 2005).</td>
</tr>
<tr>
<td>Calcium homeostasis</td>
<td>Ectopic calcification in various organs (Kuro-o et al., 1997).</td>
<td>Normal (Kurosu et al., 2005).</td>
</tr>
<tr>
<td>Diseases</td>
<td>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).</td>
<td>Protection of the angiotensin II-induced renal damage (Mitani et al., 2002). Suppression of H₂O₂-induced apoptosis and cellular senescence in vascular cells (Ikushima et al., 2006). Reduction of risk factors for atherosclerosis. Enhanced hearing ability (Bektas et al., 2004).</td>
</tr>
</tbody>
</table>
High BD Node: Validation

Validation

Chateau et al. 2010. Aging
Subsystem Validation

Discovery

Function?
Summary

• Networks $\Rightarrow$ 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

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

Slide: Courtesy of Dr. Nick Stergiou
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

http://fiji.eecs.harvard.edu/Mercury
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 are 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)
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

Phase I: Fall Detection
Accelerometer-based fall detection

- Determine an acceleration threshold.
- Detect fall.

![Falls Backward Chart]
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.
Phase I: Fall Detection
Accelerometer-based fall detection

- Measure acceleration in three orthogonal directions.
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 sideway  
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

$$\Phi = \arctan \frac{A_y}{A_x}$$

- $A_x$: 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

![Graph showing SD of Standing, Walking, and Running](image)

- Static Threshold: 0.075
- Dynamic Threshold 1: 0.12
- Dynamic Threshold 2: 0.3
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

Activity Classification

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

<table>
<thead>
<tr>
<th>Activity</th>
<th>Place</th>
<th>Time (second)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Subject I</td>
<td>Subject II</td>
<td></td>
</tr>
<tr>
<td>Walking</td>
<td>Outdoor</td>
<td>360</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>Standing</td>
<td>Outdoor</td>
<td>20</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>Running</td>
<td>Outdoor</td>
<td>30</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Standing</td>
<td>Outdoor</td>
<td>30</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>Walking</td>
<td>Outdoor</td>
<td>300</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>Sitting</td>
<td>Indoor</td>
<td>360</td>
<td>340</td>
<td></td>
</tr>
<tr>
<td>Walking</td>
<td>Indoor</td>
<td>30</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>Lying</td>
<td>Indoor</td>
<td>72</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td><strong>Total time</strong></td>
<td></td>
<td><strong>1200</strong></td>
<td><strong>1200</strong></td>
<td></td>
</tr>
</tbody>
</table>
Profiles: Experiment Results

**Test**
- Sitting: 32%
- Walking: 54%
- Standing: 5%
- Lying: 7%
- Running: 2%

**Goal**
- Sitting: 30%
- Walking: 57%
- Standing: 4%
- Lying: 6%
- Running: 3%
Profile: Experiment Results

Test:
- running: 1%
- lying: 6%
- sitting: 29%
- standing: 12%
- walking: 52%

Goal:
- running: 1%
- lying: 5%
- sitting: 28%
- standing: 12%
- walking: 54%
Phase IV: Fall Prediction

User’s Mobility Profiles

Fall-Prediction Module (Hidden Markov model)

User’s Mobility Health Index

Training Mobility Profiles
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
    \[ a_{kl} = P(\pi_i = l | \pi_{i-1} = k) \], which is probability from state \( k \) to state \( l \)
  
  - emission probabilities
    \[ e_l(b) = P(x_i = b | \pi_i = l) \], 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

• $a_{kl} = \frac{A_{kl}}{\sum_l A_{kl}}$

• $e_k(b) = \frac{E_k(b)}{\sum_b E_k(b)}$
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

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

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

\[
accuracy = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{numbers of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}
\]
Table: Showing True positives (TP), true negatives (TN), false positives (FP), false negatives (FN) and accuracy (ACC) values for predicting each state

<table>
<thead>
<tr>
<th>State</th>
<th>#of states predicted</th>
<th>TP</th>
<th>FN</th>
<th>FP</th>
<th>TN</th>
<th>ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unhealthy</td>
<td>11</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>8</td>
<td>100.00</td>
</tr>
<tr>
<td>Mildly Unhealthy</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>11</td>
<td>85.71</td>
</tr>
<tr>
<td>Healthy</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>11</td>
<td>88.89</td>
</tr>
</tbody>
</table>
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
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

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

- Small Business
- Multinational Corporations
- Government Offices
- Homes
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

Biocore System Infrastructure

- Public Clouds
- On-demand Workstations
- On-demand Application Servers
- Blackforest Cluster
- Other Internet Resources
- Secure Data
- Holland Computing Center
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