Biosciences will never be the same

- IT changed the world 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 not easy
- The answer is Bioinformatics

Scientific Advances in 20th Century

In the second half of the 20th century there have been two fundamental major developments in the field of science:

- Integrated circuits and the digital revolution
- Discovery of DNA structure and the genetic code
**Introduction**

Bioinformatics is an emerging "exciting" interdisciplinary science that deals with the development and use of mathematical and computational approaches to solve various problems in biosciences.

Hence, Bioinformatics Algorithms is the central component of Bioinformatics education and research.

**What is not Bioinformatics?**

- The process of speeding up Biosciences processes by automating what otherwise Biologists have to do manually.
- A code word for Biologists having to deal with very large amounts of Information.
- Another application domain for Computer Scientists to apply or employ their currently available algorithms or theories.

**What is Bioinformatics?**

NCBI: "Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned."
Computational Sciences and Biosciences: A Perfect Fit

- The digital nature of the biological data
- The massive size of the available databases and the high degree of non-determinism associated with them
- The potential of using several well-researched algorithms and problem solving techniques in solving bioinformatics-related problems
- The importance of sequence comparison, database search and string matching algorithms

Impact on Industry

- Increasing number of Biotech companies
- Increasing sales of Biotech drugs
- The emergence of genetic tests
- Emergence of a new paradigm for drugs: right dose of the right drug for the right patient (Pharmacogenomics)
Why Bioinformatics is hot?

• Supply/demand: few people adequately trained in both biology and computer science
• Genome sequencing, microarrays, next generation sequencing lead to large amounts of data to be analyzed
• The ability of Bioinformatics researcher to innovate and discover beyond the narrow focused traditional way
• The phase of basic tools is almost over
• Bioinformatics role: from providing support to asking questions
• The availability of more NIH and NSF funds
• Translational Bioinformatics
• Data and more data

Opportunities?

• SmartMoney ranks Bioinformatics as #1 among next Hot Jobs
• Business Week 50 Masters of Innovation
• Jobs available, exciting research potential
• Important information waiting to be decoded!

Opportunities

As of July 31, 2012:

[Image of naturejobs.com]
Typical Projects

What is Translational Bioinformatics?

- Translational bioinformatics
  - Development of analytic, storage, and interpretive methods to optimize the transformation of increasingly voluminous genomic and biological data into diagnostics and therapeutics for the clinician

- Includes
  - Research on the development of novel techniques for the integration of biological and clinical data
  - Evolution of clinical informatics methodology to encompass biological observations
  - End product of translational bioinformatics: Newly found knowledge from these integrative efforts that can be disseminated to a variety of stakeholders, including biomedical scientists, clinicians, and patients
Bioinformatics Challenges

- It is a multidisciplinary field involving different areas such as Biology, CS and MIS
- It requires different types of skills
- The lack of well established infrastructures
- The lack of having a large group of specialized researchers
- Simplified use of existing methods and algorithms
- The focus on recent trends

Challenges Facing Bioinformatics Education and Research

- Significant gaps between tool developers and tool users
  - Different objectives
  - Different funding agencies
  - Different academic cultures
- Significant problems with available Biological Data
  - Archival based
  - Lack of structure

Current Steps in Nebraska

- Great interest from researchers/educators/students
- Support from the administration
- Infrastructure Supported by NRI, NICLS and INBRE
- Various grants funded by NIH and NSF
- High degree of communication
- New innovative programs in Bioinformatics
- Various interesting research projects
- Student organization/society in Bioinformatics

Focus on an interdisciplinary approach
State of the Field

- Availability of many large useful database systems; private and public
- Availability of numerous helpful software packages
- Fragmented efforts by computational scientists and biologists
- Advances in new technologies as high throughput next generation sequencing
- Trendiness of the discipline
- Increasing interest among researchers and educators
- Huge interest from industry and the public

Data in Bioinformatics

- 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 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

Data versus Knowledge

- With high throughput data collection, biology needs 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

Data Generation vs. Data Analysis/Integration

- 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

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 Storage Infrastructure
- Innovative Methods Integrated Data Interdisciplinary Approach

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
Central Concept in Computer Science

ALGORITHMS

• Algorithms lead to automation

The question of solving a problem becomes the question of describing specifically how to solve the problem.

Focus on Algorithms

• It is all about solving problems – lead to new knowledge
• Integration problem solving techniques with hypothesis based research
• Bioinformatics Algorithms, along with innovative data models, are the central component of Bioinformatics discoveries

Biological vs. Computer Algorithms

• Instructions and data versus coding and non-coding regions
• Parsing of instructions and data versus recognition of sites in biological sequences
• Termination conditions versus termination of duplications
• Problem: The environmental and the time dependent factors
Course Objectives

• Introduction to Bioinformatics and current state of the discipline
• Overview of key Current Bioinformatics algorithms and tools – examples using sequence comparison and phylogenetic analysis
• Next Generation Bioinformatics Tools: Intelligent, Collaborative and Dynamic (ICD) tools – A focus on biological networks, systems biology and genome assembly
• Opportunities and Challenging in Bioinformatics: The next steps – A focus on systems biology, high performance computing, computing facilities and the cloud, genome wide association studies, and security/privacy issues
• Present several case studies to highlight the value of the ICS model with a focus on aging research

Course Objectives by Days

• Day One: General introduction to Bioinformatics
• Day Two: Basic Bioinformatics algorithms - sequence comparison and phylogenetic analysis
• Day Three: From data generation to data integration/analysis Systems biology and network analysis approaches
• Day Four: The impact of sequencing technology on Bioinformatics: a focus on sequence assembly Genome wide association studies
• Day Five: Opportunities and Challenging in Bioinformatics: The next steps – a focus on biomedical informatics, high performance computing, computing facilities and the cloud, genome wide association studies, and security/privacy issues

Overview of Molecular Biology
“Biological” Database

- Bioinformatics is a DATA-DRIVEN scientific discipline
- 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
- Results of many experiments like Microarray Data

Data and more Data

- Incredible amounts of publicly-available data
  - GenBank: Hundreds of organisms have been completely sequenced, including man and mouse; 260,000 species have had SOME sequence measured
  - GEO has 236,000 samples today from 9080+ experiments
  - ArrayExpress has 118,000 samples today from 6650+ experiments
  - EBI Prider: 7900+ samples yielding over 7.4+ million mass spectra
  - NCBI dbGAP (genotype and phenotype): 25 genetic studies with 50,000+ human samples
- Doubling or tripling in size each year

Growth of GenBank

(1982 - 2006)
Growth of Biological Databases

“Big Data”

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

Problems: Current Biological Data

- The availability of large biological data and the increasing rate in producing new data, available in public data banks or via microarray data
- The increasing pressure to maximize the use of the available data, particularly to impact key related industries (biotech companies, biotech drugs)
- The large degree of heterogeneity of the available data in terms of quality
**Issues: Current Biological Databases**

- 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

**Bioinformatics Solutions**

- Develop new inventive database models
  - Custom database for specific domains
  - Centralized Structured integrated data
- Develop innovative Bioinformatics tools
  - Clustering/classification algorithms
  - Advanced motif finding approaches
- Systems Biology Approach

**UNO Bioinformatics Research Group**

- Real Biological Problems
- Advanced Tools
- Innovative Database Models
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 microorganism. And they’ve named it “Mycobacterium nebraskense,” after the Cornhusker state.
- Discovered using Mycoalign: A Bioinformatics program developed at PKI

Global Network Structures

Network density changes over time for specific cellular functions

Gateway Nodes in Aging Mice
First 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
Examples of First Generation of Bioinformatics Tools

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

Sequence Comparison

- 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 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.
Example

- Compare the two DNA sequences: GACGATTAG and GATCGAATAG
- GACGATTAG
  GATCGAATAG
  20% similar
- GA--CGATTAG
  GATCGAATAG
  80% similar

Score of Alignment

- GA--CGATTAG
  GATCGAATAG
- % of similarity 80%
- Alignment score = 8 * 1 + 1 * (-1) + 1 * (-2) = 5
  +1: match
  -1: mismatch
  -2: a space

Evolutionary Trees

- Science of estimating the evolutionary past
- Phylogenetic analysis is the means used to estimate evolutionary relationships based on observable evidence
- Evidence can include morphology, physiology, and other properties of organisms. Paleontological and geological evidence is also used.
- Linnaeus’s system of grouping and naming organisms to reflect evolutionary relationship- Phenotype Phylogenetic
Molecular Phylogenetics

- The molecular biology of an organism can also provide evidence for phylogenetic analysis (DNA, RNA, Protein Sequence)
- Accumulated mutational changes in DNA and protein sequence over time constitutes evidence
- Sequence-based phylogenetic analysis can be automated or semi-automated using computers

Phylogenetic Analysis

Relationships among species

DNA Sequence Evolution

-3 mil yrs
-2 mil yrs
-1 mil yrs
today
**Phylogeny Problem**

```
<table>
<thead>
<tr>
<th>U</th>
<th>AGGCAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>TAGCCA</td>
</tr>
<tr>
<td>W</td>
<td>TAGCTT</td>
</tr>
<tr>
<td>X</td>
<td>TGCACA</td>
</tr>
<tr>
<td>Y</td>
<td>TGGCCTT</td>
</tr>
</tbody>
</table>
```

**Limitations of First Generation Bioinformatics Tools**

- Lack of focus on data integration
- Mostly data independent
- Lack on new computational innovation
- Has little room for incorporating biological knowledge
- Developed in isolation

**Next Generation Bioinformatics 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

Systems Biology Approach

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

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

- 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

Background: Basic Biological Concepts

- DNA → RNA → Protein
  - Occurs millions of times within a cell
  - Results in different cell/tissue types
    - Measurable relationships
- High-throughput analyses: measuring thousands of relationships in one experiment
**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

**Cells**

- Complex system enclosed in a membrane
- Organisms are unicellular (bacteria, baker’s yeast) or multicellular
- Humans:
  - 60 trillion cells
  - 320 cell types
Basic Functions of Cells

- Maintain Low entropy – Living organisms must expend energy to keep things orderly
  - Maintain homeostasis: Cells want to stay alive
  - Real World Similarity: If Dodge Street is blocked, you do not quit your travel, you take Pacific Street or Center St.

- Store and save the information about how and when to build these molecules

- Be able to build molecules that assist in the critical functions of life

How to maintain low entropy

LIPIDS

- Hydrophilic fragments connected to hydro‐phobic fragments.

- They form sheets (Lipid bi‐layers) with hydrophilic ends on the outside and hydro‐phobic ends on the inside.

- The result creates a very stale separation, not easy to pass thought except for water and a few other small atoms/molecules

How to maintain low entropy

The lipid bi‐layer:

![Image Source: http://images.tutorvista.com/content/cellular‐micromolecules/lipid‐bilayer‐structure.jpeg]
Organisms

- Classified into two types:
  - **Eukaryotes**: contain a membrane-bound nucleus and organelles (plants, animals, fungi,...)
  - **Prokaryotes**: lack a true membrane-bound nucleus and organelles (single-celled, includes bacteria)

- Not all single celled organisms are prokaryotes!

Chromosomes

- In eukaryotes, nucleus contains one or several double stranded DNA molecules organized as chromosomes
- Humans:
  - 22 Pairs of autosomes
  - 1 pair sex chromosomes

Human Karyotype

[Image Source: http://avery.rutgers.edu/WSSP/StudentScholars/Session8/Session8.html]
What is DNA?

- DNA: Deoxyribonucleic Acid
- Single stranded molecule (oligomer, polynucleotide) chain of nucleotides
- 4 different nucleotides:
  - Adenosine (A)
  - Cytosine (C)
  - Guanine (G)
  - Thymine (T)

The Main Code

- The Deoxyribonucleic Acid (DNA) sequence carries the main code of living organisms.
- It has a common backbone and specialized side-chains: Adenine, Cytosine, Guanine and Thymidine or A, C, G and T.
- It can be uniquely represented by the 4 letter alphabets (AGT) called bases
- Long sequences of the 4 letters are linked together to create genes and control information

Double Helix

- Two complementary DNA strands form a stable DNA double helix
- Discovered by Watson & Crick in April 1953

Image source: www.uncio.ac.uk/biology/intro.htm
**RNA**

- Ribonucleic Acid
- Similar to DNA
- Thymine (T) is replaced by uracil (U)
- RNA can be:
  - Single stranded
  - Double stranded
  - Hybridized with DNA

![RNA secondary structure](image_source)

**RNA secondary structure**

- E. coli Rnase P
- RNA secondary structure

**mRNA**

- Messenger RNA
- Linear molecule encoding genetic information copied from DNA molecules
- Transcription: process in which DNA is copied into an mRNA molecule
mRNA processing

- Eukaryotic genes can be pieced together
  - Exons: coding regions
  - Introns: non-coding regions

- mRNA processing removes introns, splices exons together

- Processed mRNA can be translated into a protein sequence

mRNA Processing

Genetic Code

- 4 possible bases (A, C, G, U)
- 3 bases in the codon
- $4 \times 4 \times 4 = 64$ possible codon sequences
- Start codon: AUG
- Stop codons: UAA, UAG, UGA
- 61 codons to code for amino acids (AUG as well)
- 20 amino acids – redundancy in genetic code
Proteins

- Polypeptides having a three dimensional structure.
- **Primary**—sequence of amino acids constituting the polypeptide chain
- **Secondary**—local organization into secondary structures such as α helices and β sheets
- **Tertiary**—three dimensional arrangements of the amino acids as they react to one another due to the polarity and resulting interactions between their side chains
- **Quaternary**—number and relative positions of the protein subunits

Protein Molecules

- Proteins are defined by a sequence of linked subunits called amino acids (protein residues).
- There are 20 different amino acids with different physical and chemical properties.
- The interaction of these properties allows a chain of the amino acids to fold into a unique reproducible 3D structure.

Protein Representation

- The protein sequence has a common repeating backbone and unique side chains
- Since the backbone is constant, it can be uniquely represented by the amino acid sequence
- Protein and Ribonucleic (RNA) are encoded by DNA
**DNA codes Protein and RNA**

- Each of the 20 amino acids can be specified by a 3 consecutive DNA bases
- The sequence of DNA is read, 3 at a time, and creates the corresponding sequence of protein chain
- The chain folds itself based on the amino acid properties
- The 64 mappings of the 3 bases to one amino acid is called the Genetic Code and is universal.

---

**20 Amino Acids**

<table>
<thead>
<tr>
<th>Amino Acids</th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine</td>
<td>Ser</td>
<td>Tyr</td>
<td>Cys</td>
<td>Glu</td>
</tr>
<tr>
<td>Alanine</td>
<td>Ala</td>
<td>His</td>
<td>Arg</td>
<td>Gln</td>
</tr>
<tr>
<td>Valine</td>
<td>Val</td>
<td>His</td>
<td>Arg</td>
<td>Gln</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu</td>
<td>His</td>
<td>Arg</td>
<td>Gln</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
<td>His</td>
<td>Arg</td>
<td>Gln</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe</td>
<td>His</td>
<td>Arg</td>
<td>Gln</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro</td>
<td>His</td>
<td>Arg</td>
<td>Gln</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser</td>
<td>Ala</td>
<td>Lys</td>
<td>Arg</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr</td>
<td>Ala</td>
<td>Lys</td>
<td>Arg</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
<td>Ala</td>
<td>Lys</td>
<td>Arg</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met</td>
<td>Ala</td>
<td>Lys</td>
<td>Arg</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Trp</td>
<td>Ala</td>
<td>Lys</td>
<td>Arg</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr</td>
<td>Ala</td>
<td>Lys</td>
<td>Arg</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
<td>Asp</td>
<td>Glu</td>
<td>Glu</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln</td>
<td>Glu</td>
<td>Asp</td>
<td>Asp</td>
</tr>
<tr>
<td>Aspartic Acid</td>
<td>Asp</td>
<td>Glu</td>
<td>Asp</td>
<td>Asp</td>
</tr>
<tr>
<td>Glutamic Acid</td>
<td>Glt</td>
<td>Glu</td>
<td>Asp</td>
<td>Asp</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys</td>
<td>Lys</td>
<td>Lys</td>
<td>Lys</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
<td>Arg</td>
<td>Arg</td>
<td>Arg</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
<td>His</td>
<td>His</td>
<td>His</td>
</tr>
<tr>
<td>START:</td>
<td>AUG</td>
<td>STOP: UAA, UAG, UGA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Amino Acids**

- building blocks for proteins (20 different)
- vary by side chain groups
- **Hydrophilic**: amino acids are water soluble
- **Hydrophobic**: are not
- Linked via a single chemical bond (**peptide bond**)
- **Peptide**: Short linear chain of amino acids (< 30) **polypeptide**: long chain of amino acids (which can be upwards of 4000 residues long).
What is a Gene?

- The physical and functional unit of heredity that carries information from one generation to the next
- DNA sequence necessary for the synthesis of a functional protein or RNA molecule

Genome

- Chromosomal DNA of an organism
- Number of chromosomes and genome size varies quite significantly from one organism to another
- Genome size and number of genes does not necessarily determine organism complexity
## Genome Comparison

<table>
<thead>
<tr>
<th>ORGANISM</th>
<th>CHROMOSOMES</th>
<th>GENOME SIZE</th>
<th>GENES</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em> (Humans)</td>
<td>23</td>
<td>3,200,000,000</td>
<td>~30,000</td>
</tr>
<tr>
<td><em>Mus musculus</em> (Mouse)</td>
<td>20</td>
<td>2,600,000,000</td>
<td>~30,000</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em> (Fruit Fly)</td>
<td>4</td>
<td>180,000,000</td>
<td>~18,000</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em> (Yeast)</td>
<td>16</td>
<td>14,000,000</td>
<td>~6,000</td>
</tr>
<tr>
<td><em>Zea mays</em> (Corn)</td>
<td>10</td>
<td>2,400,000,000</td>
<td>??</td>
</tr>
</tbody>
</table>

## Proteome

- The complete collection of proteins that can be produced by an organism.
- Can be studied either as static (sum of all proteins possible) or dynamic (all proteins found at a specific time point) entity.

## Genes and Control Info

- The set of all genes required for a living organism is called the organism’s Genome.
- The Human Genome has about 3 billion bases divided into 23 linear segments called chromosomes.
- A gene has an average of 1340 DNA bases specifying a protein of about 447 amino acids.
- Humans have about 35K genes = 40M DNA bases = about 3% of total DNA in the genome.
- The rest, about 2,960M bases are believed to include control information (when, where, how long, etc.)
Genes and Control Info

• The behavior, development, disease, body structure, microscopic structure of an organism are determined by its genes and their control – the Genotype.
• To understand biology, we need to understand how the gene products (protein) interact and produce observed features – the Phenotype.
• Many genes are shared among organisms, but each has made changes in the detailed DNA sequence.

Functions of Genes

• Signal transduction: sensing a physical signal and turning into a chemical signal
• Structural support: creating the shape and pliability of a cell or a set of cells
• Enzymatic catalysis: accelerating chemical transformation otherwise too slow
• Transport: getting things into and out of separated compartments

Functions of Genes

• Movement: contracting in order to pull things together or push things apart
• Transcription control: deciding when other genes should be turned on/off
• Trafficking: affecting where different elements end up inside the cell
Genomes and Algorithms

Genomes and Computer Programs

- Is it a genetic program metaphor?
  - Genome ➔ computer program (algorithm) that completely specifies the organism
  - Cell Machinery ➔ interpreter of the program
  - Biological functions performed by proteins ➔ execution of the program

Genomes and Computer Algorithms

Tempting analogy – but...

- DNA program undergoes changes during the transcription and translation, hence the genetic code can't be simply applied to a stretch of DNA known to contain a gene to find the protein corresponds to that gene
- Gene expression is a complex process that may depend on spatial and temporal context

Is it a very complex algorithm?
The Need for Bioinformatics algorithms

"Biology has rapidly become a large source for new algorithmic and statistical problems, and has arguably been the target for more algorithms than any of the other fundamental sciences. This link between computer science and biology has important educational implications that change the way we teach computational ideas to biologists, as well as how applied algorithms are taught to computer scientists."

So Many Techniques Available

- Optimization (Expectation maximization, Monte Carlo, Simulated Annealing, gradient-based methods)
- Dynamic programming
- Bounded search algorithms
- Cluster analysis
- Classification
- Neural networks
- Genetic algorithms
- Bayesian inference
- Stochastic context tree grammars

Systems Complexity:
The Need for Bioinformatics, High Performance Computing, and Data Analysis Tools
Motivation

Why don’t we have personalized medicine?  
Where is the cure for cancer?  
Why is AIDS still misunderstood?

Can effectively be boiled down to:
Why hasn’t high-throughput data been effectively harnessed yet?

Motivation

Answer:
– Complexity of the system
– Complexity of the organism
– Size of the data (“big data”)
– Heterogeneity of the data
– Computing power
– …..

Motivation
But....

- Her2+ BC
  - 20-25% of breast cancers
  - Normal treatment: Herceptin
    - Eventually stops working if cancer comes back
- T-DM1 Drug
  - Trojan horse drug
  - 3 extra months of improvement
  - Lack of usual side effects

**Personalized Medicine**

**Resource**

**Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes**

- [Link](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3219548/)

[Image of a table or chart related to personalized medicine]
Methods

- 54yr old male volunteer
- Plasma and serum used for testing
- 14 month time course
- Complete medical exams and labs at each meeting (20 time points total)
- Extensive sampling at 2 periods of viral infection:
  - HRV (human rhinovirus) – common cold
  - RSV (respiratory synticial) - bronchitis

Time-course summary

226 days total
HRV - red
RSV - green
Fasting - blue
Lifestyle changes: ↑ exercise, took ibuprofen daily, ↓ sugar intake

Methods

- Whole-genome sequencing
  - Complete Genomics Deep WGS and Illumina
  - CG 35nt paired end
  - Illumina 100 nt paired end
  - 150 and 120-fold coverage
- 91% mapped to human RefSeq
- Remaining 9%:
  - 1,425 contigs mapped to non-RefSeq data
  - Remaining were unique
  - 2919 exons expressed using RNA-Seq

"A large number of undocumented genetic regions exist in the personal human genome"
Sequencing Data

<table>
<thead>
<tr>
<th>Type</th>
<th>Total Variants</th>
<th>Total/Homozygous</th>
<th>Heterozygous/Homozygous</th>
<th>Heterozygous/Homozygous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total SNVs</td>
<td>3,709,347</td>
<td>3,709,347</td>
<td>5,652,853</td>
<td>5,652,853</td>
</tr>
<tr>
<td>Total indels</td>
<td>46,467</td>
<td>46,467</td>
<td>77,586</td>
<td>77,586</td>
</tr>
<tr>
<td>Introns</td>
<td>30,925</td>
<td>34,603</td>
<td>72,045</td>
<td>72,045</td>
</tr>
<tr>
<td>Exons</td>
<td>15,542</td>
<td>3,054</td>
<td>2,386</td>
<td>2,386</td>
</tr>
<tr>
<td>Introns</td>
<td>80,798</td>
<td>73</td>
<td>49</td>
<td>24</td>
</tr>
<tr>
<td>Exons</td>
<td>11,809</td>
<td>15,601</td>
<td>6,747</td>
<td>4,117</td>
</tr>
<tr>
<td>SNI</td>
<td>4,094</td>
<td>2,070</td>
<td>1,992</td>
<td>1,156</td>
</tr>
<tr>
<td>INDEL</td>
<td>35,798</td>
<td>55,804</td>
<td>10,561</td>
<td>8,105</td>
</tr>
<tr>
<td>Others</td>
<td>1,202,745</td>
<td>1,192,596</td>
<td>165,702</td>
<td>146,232</td>
</tr>
<tr>
<td>Total</td>
<td>4,492,754</td>
<td>5,147,140</td>
<td>7,204,305</td>
<td>7,363,009</td>
</tr>
<tr>
<td>Average per sample</td>
<td>8,985</td>
<td>10,294</td>
<td>14,410</td>
<td>14,727</td>
</tr>
<tr>
<td>SD</td>
<td>4.283</td>
<td>0.325</td>
<td>0.22</td>
<td>0.17</td>
</tr>
<tr>
<td>Median</td>
<td>4.154</td>
<td>0.317</td>
<td>0.21</td>
<td>0.16</td>
</tr>
</tbody>
</table>

WBS Based Disease Risk Eval

- 51 SNVs and 4 indels found in OMIM
  - LD mutations
  - Validated by Sanger sequencing

- High interest genes:
  - SerpinA1
  - TERT (acquired aplastic anemia)
  - Diabetes and hypertension: GKCR, KCNJ11, TCF7

- BASED ON THESE MUTATIONS:
  - Medical phenotype monitoring
    - Monitor blood glucose levels

---

**Diagram:**

- Glucose level changes over time
-高峰点：(Day 3-4)
-低谷点：(Day 5-6)
Dynamic Omics Analysis

- Transcriptome: RNA-Seq of 20 time points
  - 2.67 billion paired end reads
  - 19,714 isoforms for 12,659 genes tracked

- Proteome: Quantification of 6,280 proteins
  - 14 time points via TMT and LC/MS

- Metabolome: 1,020 metabolites tracked during viral infections
  - miRNA analysis also during HRV infection

Techniques Used

- Summary of techniques used:
  - Sample collection
  - HRV and RSV detection
  - Whole-genome sequencing
  - Whole-exome sequencing
  - Sanger-DNA sequencing
  - Whole-transcriptome sequencing: mRNA-Seq
  - Small RNA sequencing: microRNA-Seq
  - Serum Shotgun Proteome Profiling
  - Serum Metabolome Profiling
  - Serum Cytokine Profiling
  - Autoantibodyome Profiling
  - Telomere Length Assay
  - Genome Phasing
  - Omics Data Analysis

Team Count: 41
Complexity of the Cellular System

Structural Complexity

DNA:
Made of elements in structures
Forms a helix
4 Types of DNA
Stored on chromosomes
1 link = 1 base pair/bp

Chromosomes:
249 million bp of DNA on human Chromosome 1
23 pairs of chromosomes in human genome
6 billion bp in entire genome
Stored in nucleus
Structural Complexity

Cell:
DNA stored in nucleus

411 different types of cells in the human body

Storage
Barrier
Blood
Pigment
Contractile
Germ
Neural
Epithelial etc.

Stored together as tissues

Structural Complexity

Tissues:
Few major types in the body

Some pathology studies how to determine cell types with microscopy

Tissue forms the body

Systems Complexity of the Cellular System
System Level Complexity

DNA:
- Made of elements in structures
- Forms a helix
- 4 Types of DNA
- Stored on chromosomes
  - 1 link = 1 base pair/bp

System Level Complexity

DNA:
- Divided into genes, promoters, enhancers, repressors
- 30,000+ genes in human genome
- Genes become proteins/RNA
  - Promoters/enhancers/repressors regulate DNA transcription

System Level Complexity

Human genome
- Nucleus genome: 30,000+ genes
- Mitochondrial genome: 37 genes

- Genes and gene-related sequences
- Unique or mostly unique: 10%
  - Exons
  - Introns
  - Tandemly repeated or dispersed repeats
  - Interspersed repeats

- Coding DNA
  - Peptidoglycan
  - Gene neighborhood
- Noncoding DNA
  - Unique or mostly nonrepetitive
  - Moderate to high repetitiveness

- Transcribed
  - Unique or mostly nonrepetitive
  - Moderately repeated
  - Highly repeated
System Level Complexity

Pre-mRNA:
- Genes are coded to pre-mRNA
- Introns are DNA with unknown/junk function (non-protein coding)
- Exons are regions of genes that are used in the final product

mRNA:
- Messenger RNA
  - Surrounded by identification regions
  - Multiple RNAs can come from one gene via alternate splicing
  - Folded into protein

Alternative splicing:
- Exons connected in a different way than transcribed
- Result in different protein
- Different function? COF?
- No function?
System Level Complexity

Proteins:
- Final “stop” of DNA
- Can function alone or in complexes
- Protein complexes are composed of 2 or more proteins bonded together
- Can have one or many functions

Key Issues

- Complexity: structural, system, assay, data

Structural:
- 6 billion bp stored as DNA
- 411 types of cells in 6 major tissue types

System:
- 30,000+ genes (Human genome)
- Each gene can have many protein isoforms (alternative splicing)
- Proteins in final form can complex or stand alone
  - Can have multiple functions
- Many of these forms are unknown!

Data Level Complexity of the Cellular System
**Data Level Complexity**

- High-throughput assays
  - Survey entire cellular landscape at a time
  - Measure quantities of entire “-omics”

- Genome – which genes are expressed
- Transcriptome – which RNAs are being made
- Proteome – which proteins are made and interacting in the cell at a time
- Metabolome – how are metabolites reacting together
Data Level Complexity

• Types of high throughput assays:
  – DNA Microarray
  – RNAi HTS
  – Sequencing
  – Yeast 2-Hybrid/Co-IP
  – Synthetic genetic array
  – Chemical screens
  – SAGE
  – Antibody microarray
  – Phage display
  – SNP screening

Data Level Complexity

• DNA Microarray
  – Measures the level of expression of genes for any given state/experiment

• Methodology:
  http://www.bio.davidson.edu/courses/genomics/chip/chip.html
Data Level Complexity

• High throughput sequencing
  – What is the sequence of a genome?

• More on this later….BUT....

Data Level Complexity

• High throughput sequencing
  – THE HUMAN GENOME PROJECT
  – Started in 2000, completed in 2003
  – $3.7 billion dollars cost (then)
  – Now: Cheap! But can we make it cheaper?
    • X Prize

Data Level Complexity
Data Level Complexity

- Genomics.xprize.org
  - $10 million to the group who sequences
    - 100 genomes of 100 centenarians
    - 30 days or less
    - Scored for accuracy, completeness, and haplotype phasing
      - 1 error per 1,000,000 bases
      - 98% completeness
      - Order, orientation, copy number, indels
    - Max cost of $1000 per genome

Data Level Complexity

- Yeast 2-hybrid
  - Method for detecting protein-protein interactions

Data Level Complexity

- Activate a reporter gene (LacZ) using transcription factor (Gal4) on an upstream activating sequence
  - This is regular transcription (gene gets made)

A. Regular transcription of the reporter gene
Data Level Complexity

– Gal4 is split into binding domain (BD) and activating domain (AD)
– BD and AD both needed for transcription
– Add a bait (protein 1) to the TF BD
– No transcription because no AD

Data Level Complexity

– Gal4 is split into binding domain (BD) and activating domain (AD)
– Add a prey (protein 1) to the TF AD
– No transcription because no BD

Data Level Complexity

– Gal4 is split into binding domain (BD) and activating domain (AD)
– If Bait and Prey interact, BD and AD are both present
– Reporter gene gets transcribed

Data Level Complexity

– Gal4 is split into binding domain (BD) and activating domain (AD)
– BD and AD both needed for transcription
– Add a bait (protein 1) to the TF BD
– No transcription because no AD

B. One fusion protein only (Gal4-BD + Bait) - no transcription

C. One fusion protein only (Gal4-AD + Prey) - no transcription

D. Two fusion proteins with interacting Bait and Prey
Data Level Complexity

• Gene knockouts: Testing one gene individually

Data Level Complexity

• Synthetic Genetic Array
  – Generate a set of query genes
  – Use a gene library (entire yeast genome = 6k genes)
  – Systematic knockouts of gene_{n} vs. gene_{n+1}
  – If yeast grows normally – no interaction
  – If yeast grows larger – positive interaction
  – If yeast grows smaller/not at all – negative interaction
**Data Level Complexity**

- Synthetic Genetic Array
  - Currently only developed for yeast

**Data Level Complexity**

- Chemical screens:
  - Start with a library of drugs/chemicals
  - OR Start with a library of genes/proteins
  - Combine libraries
  - Analysis by microscopy+detectors
  - Identify which drugs have most impact on quantity/function

**Conclusions**

- Multiple levels of complexity in biological systems
- Noise to signal ratio imminent with amounts of data recorded
- How do we sort through noise to signal ratio?
  - Future lectures: high performance computing, network analysis
Conclusions

- Big Data in other industries?

- As for cancer...
  - Results from a series of events
  - Not one single event

- Multiple "system failures" must occur

Next Step?

- William Li:
  - Can we eat to starve cancer? TED Talk
    - Focus on how angiogenesis is a major component for cancer
    - "Cancer without disease"
    - Lifestyle changes and diet changes that can help prevent

Conclusions

• Need algorithms to sort through the data
• Need systems that can manage multiple types of heterogeneous data
• Need shift from data generation to data analysis
• This includes:
  – High performance computing
  – Network analysis
  – Integrated softwares and data warehousing

Exercise One

• TED Talks: See Exercise One Sheet
  – Choose at least three out of the six talks
  – For each talk, provide a brief summary paragraph and identify two things as main take away points
• Get familiar with Bioinformatics Server
  – Needed for next three exercises

Course Objectives by Days

• Day One: General introduction to Bioinformatics
• Day Two: Basic Bioinformatics algorithms - sequence comparison and phylogenetic analysis
• Day Three: From data generation to data integration/analysis
  Systems biology and network analysis approaches
• Day Four: The impact of sequencing technology on Bioinformatics: a focus on sequence assembly Genome wide association studies
• Day Five: Opportunities and Challenging in Bioinformatics: The next steps – a focus on biomedical informatics, high performance computing, computing facilities and the cloud, genome wide association studies, and security/privacy issues
Acknowledgments

• UNO Bioinformatics Research Group
  Kiran Bastola
  Sanjukta Bhoomwiclk
  Kate Dempsey
  Jasjit Kaur
  Ramez Mena
  Sachin Pawaskar
  Oliver Bonham-Carter
  Ishwor Thapa
  Dhawal Verma
  Julia Warnke

• Former Members of the Group
  Alexander Churbanov
  Xutao Deng
  Huiming Geng
  Xiaolu Huang
  Daniel Quest

• Biomedical Researchers
  Steve Bonasera
  Richard Hallworth
  Steve Hinrichs
  Howard Fox
  Howard Gendelman

• Funding Sources
  NIH INBRE
  NIH NIA
  NSF EPSCoR
  NSF STEP
  Nebraska Research Initiative