Computational Genomics

Ron Shamir, Haim Wolfson & Irit Gat-Viks
Fall 2016-17
What’s in class this week

• Motivation
• Administration
• Some very basic biology & biotechnology, with examples of our type of computational problems
• Additional examples
Bioinformatics

- The information science of biology: organize, store, analyze and visualize biological data
- Responds to the explosion of biological data, and builds on the IT revolution
- Use computers to analyze A LOT of biological data.
Paradigm shift in biological research

**Classical biology:** focus on a single gene or sub-system. *Hypothesis driven*

- Large-scale data;
- Bioinformatics

**Systems biology:** measure (or model) the behavior of numerous parts of an entire biological system. *Hypothesis generating*
What do bioinformaticians study?

- Bioinformatics today is part of almost every molecular biological research.
Research in systems biology

**Reductionist approach**
Studying individual parts of the biological system

**Systems approach**: Unbiased analysis of numerous constituents of the biological system
Terminology

High throughput data  

Bioinformatics tools/algorithms/methods
The Bioinformatics Actors

- Biotechnology companies
- Academic biotechnology research

High throughput data
Bioinformatics tools

- Big Pharmas and Big Agriculture
- National and international research centers
Personalized medicine

Welcome to you

Find out what your DNA says about you and your family.

- Learn what percent of your DNA is from populations around the world
- Contact your DNA relatives across continents or across the street
- Build your family tree and enhance your experience with relatives

Order now $99

Watch Greta and Stacy's Story.

Find out how these two women discovered their connection as sisters.
Administration

• ~5 home assignments as part of a home exam, to be done independently (50%)
• Final exam (50%)
• Must pass the Final to pass the course (TAU rules)

• Classes: Tue 12:15-13:30; Thu 14:15-15:30
• TA: Ron Zeira (Thu 16-17).
Administration (cont.)

• Web page of the course: http://www.cs.tau.ac.il/~rshamir/cg/16/

• Includes slides and full lecture scribes of previous years on each of the classes.
Bibliography

• No single textbook covers the course :-(
• See the full bibliography list in the website (also for basic biology)
• Key sources:
  - Gusfield: Algorithms for strings, trees and sequences
  - Durbin et al.: Biological sequence analysis
  - Pevzner: Computational molecular biology
  - Pevzner and Shamir (eds.): Bioinformatics for Biologists
Introduction

1. Basic biology
2. Basic biotechnology

+ some computational challenges arising along the way

• Touches on Chapters 1-8 in “The Cell” by Alberts et al.
The Cell

- Basic unit of life.
- Carries complete characteristics of the species.
- All cells store hereditary information in DNA.
- All cells transform DNA to proteins, which determine cell’s structure and function.
- Two classes: eukaryotes (with nucleus) and prokaryotes (without).

http://regentsprep.org/Regents/biology/units/organization/cell.gif
Nucleotides/ Bases:
Adenine (A), Guanine (G), Cytosine (C), Thymine (T).

Weak hydrogen bonds between base pairs

phosphate
sugar
DNA (Deoxy-Ribonucleic acid)

- **Bases:**
  - Adenine (A)
  - Guanine (G)
  - Cytosine (C)
  - Thymine (T)

- **Bonds:**
  - G - C
  - A - T

- **Oriented from 5’ to 3’.
- **Located in the cell nucleus**
DNA and Chromosomes

- DNA is packaged
  *Chromatin*: complex of DNA and proteins that pack it (histones)

- **Chromosome**: contiguous stretch of DNA

- **Diploid**: two homologous chromosomes, one from each parent

- **Genome**: totality of DNA material
Replication

Replication fork
Proteins: The Cellular Machines
Proteins

- Build the cell and drive most of its functions.
- Polymers of amino-acids (20 total), linked by peptide bonds.
- Oriented (from amino to carboxyl group).
- Fold into 3D structure of lowest energy.
Protein structure

Primary protein structure is a sequence of a chain of amino acids.

Secondary protein structure occurs when the sequence of amino acids are linked by hydrogen bonds.

Tertiary protein structure occurs when certain attractions are present between alpha helices and pleated sheets.

Quaternary protein structure is a protein consisting of more than one amino acid chain.
The Protein Folding Problem

Ab-initio prediction – extremely difficult!
The Protein Folding Problem

- Given a sequence of amino acids, predict the 3D structure of the protein.
- Motivation: functionality of protein is determined by its 3D structure.
- Solution Approaches:
  - Homology
  - Threading
  - de novo (=from scratch)
Genes

- **Gene**: a segment of DNA that specifies a protein.
- The transformation of a gene into a protein is called **expression**.
- Genes are < 3% of human DNA
- The rest - non-coding (used to be called “junk DNA”)
  - RNA elements
  - Regulatory regions
  - Retrotransposons
  - Pseudogenes
  - and more...
RNA (Ribonucleic acid)

- Bases:
  - Adenine (A)
  - Guanine (G)
  - Cytosine (C)
  - Uracil (U); replaces T
- Oriented from 5’ to 3’.
- Single-stranded => flexible backbone => secondary structure => catalytic role.
Transcription of DNA into RNA
Transcription of DNA into RNA
The Genetic Code

- **Codon** - a triplet of bases, codes a specific amino acid (except the stop codons)
- **Stop codons** - signal termination of the protein synthesis process
- Different codons may code the same amino acid

The genetic code, written by convention in the form in which the Codons appear in mRNA. The three terminator codons, UAA, UAG, and UGA, are boxed in red; the AUG initiator codon is shown in green.

http://ntri.tamuk.edu/cell/ribosomes.html
DNA: DNA base sequence (triplets) of the gene codes for synthesis of a particular polypeptide chain

mRNA: Base sequence (codons) of the transcribed mRNA

tRNA: Consecutive base sequences of tRNA anticodons recognize the mRNA codons calling for the amino acids they transport

Polypeptide: Amino acid sequence of the polypeptide chain
The RNA Folding Problem

Given an RNA sequence, predict its folding = the one that creates a maximum number of matched pairs

GCCUUAUGCACAUGGGCAAGCCCACGUAGCUAGUCGCGACAGGACCAGUCCCCAAUUGUUCACCCAACUCGC
CUGACCUCGGCGACGUACUACUAACUCGACCUACGCCGUUGAAACUAAGACUUUCUAGGCAGCGCUGUCAUAGGAUGGUGACAGUCUUCUU AAUUUUGUAUUGGGCCAGGCCAGCAAAGCUUGGAGUAAGGCCCGCUUGACCAGAGAUGGAUACCGGGCCAGCCUGACCAUGAGAACUGAUUAGAUCUCGUGGUAGUGCUUGUCAAAUAGAAUGAGGCCAUUCACAGACAUAGCGUUUCCAAUGAGCUAGGCGGCCCAUGUCCAGGUCCCCUAAAUAAAAGAGUCUCAC

http://www.phys.ens.fr/~wiese/highlights/RNA-folding.html
The Gene Finding Problem

Given a DNA sequence, predict the location of genes (open reading frames) exons and introns.

- A simple solution: seeking stop codons.
- 6 ways of interpreting DNA sequence
- In most cases of eukaryotic DNA, a segment encodes only one gene.
- Difficulty in Eukaryotic DNA: introns & exons
Gene Structure

Figure 3  The Complexity of the Genome

Chromosomal DNA

Start of transcript  Initiator ATG  amino acids 1-30
amino acids 31-104
Stop codon  amino acids 105-146
Poly (A) addition site
Transcriptional terminator
Endonuclease cleavage

Promoter  Exon 1  Intron 1  Exon 2  Intron 2  Exon 3

Primary RNA transcript  5'  AAAAAAAAAAAAA  3'
Mature RNA transcript  5'  Poly(A) tail
Expression and Regulation

DNA → RNA → Protein

Transcription factors (TFs): proteins that control transcription by binding to specific DNA sequence motifs.

The Motif Discovery Problem
The Human Genome: numbers

- 23 pairs of chromosomes
- ~3,200,000,000 bases
- ~21,000 genes
- Gene length: 1000-3000 bases, spanning 30-40K bases
Sequencing the human genome

1990
Project initiation

2000
First draft

2006
Full sequence
The Sequence Assembly Problem

• Given a set of sub-strings, find the shortest (super)string containing all the members of the set.
The Rosetta stone

Writing: Ancient Egyptian hieroglyphs, Demotic script, and Greek script
Computational problems?
Model Organisms

- Eukaryotes; increasing complexity
- Easy to grow, manipulate.

**Budding yeast**
- 1 cell
- 6K genes

**Nematode worm**
- 959 cells
- 19K genes

**Fruit fly**
- vertebrate-like
- 14K genes

**mouse**
- mammal
- 30K genes

- Lots of common ground with humans: many / most genes are common - but with mutations

CG © 2016
Compare proteins with similar sequences and understand what the similarities and differences mean.
Sequence Alignment problems

- Given two sequences, find their best alignment: Match with insertion/deletion of min cost.
- Same for several sequences

- "Workhorse" of Bioinformatics!
- Key challenge: huge volume of data (more on this later)
Understanding difference

2 people: 99.9% similarity

Mouse: 98% (Chimp)

Yeast: 90% (Fly)

Bacteria: 23% (Mouse)

2 people: 99.9% similarity
Sequencing

• **Sequencing**: determining the sequence of bases in a given DNA or RNA molecule.

• **Classical approach**: gel electrophoresis

• **Next-generation sequencing**: Sequencing by synthesis (Illumina): MiniSeq/MiSeq: 1-25 Million; NextSeq: 130-00 Million, HiSeq 2500: 300 million - 2 billion, HiSeq 3/4000 2.5 billion, HiSeq X: 3 billion

• **Abilities** (Illumina): reconstructs sequences of 75-300 nucleotides.
Utilize RNA-sequencing and alignment to evaluate RNA levels
More problems in sequencing data

- Solve all the problems above (alignment, gene finding, rearrangements,...) on really huge datasets
- Need to handle practical problems of efficiency - time and space
- Need to overcome large noise (errors) due to data size
ADDITIONAL EXAMPLES
Example 1 - personalized medicine

Classification and clustering problems

Given the expression profiles of normal vs. disease:
- Build an algorithm to predict if a new sample is normal or disease (classifier)
- Cluster disease profiles into sub-classes
- Cluster genes into functional groups
Classifier construction

70 genes

- Classify to minimize incorrect assignments
Example 2 - Metagenomics

Sampling the human gut

- Metagenomic analysis
  - Noval genes
  - Antibiotic resistant genes
  - Functional dysbiosis
  - Microbial diversity
Human gut microbiome viewed across age and geography

Tanya Yatsunenko1, Federico E. Rey1, Mark J. Manary2,3, Indi Trehan2,4, Maria Gloria Dominguez-Bello5, Monica Contreras6, Magda Magris7, Glida Hidalgo7, Robert N. Baldassano8, Andrey P. Anokhin9, Andrew C. Heath9, Barbara Warner2, Jens Reeder10, Justin Kuczynski10, J. Gregory Caporaso11, Catherine A. Lozupone10, Christian Lauber10, Jose Carlos Clemente10, Dan Knights10, Rob Knight10,12 & Jeffrey I. Gordon1

Gut microbial communities represent one source of human genetic and metabolic diversity. To examine how gut microbiomes differ among human populations, here we characterize bacterial species in fecal samples from 531
Bacterial diversity increases with age (based on next-gen-sequencing of fecal samples from 531 individuals)
Example 3 - computational genetics

- DNA of two human beings is ~99.9% identical
- Phenotype and disease variation is due these 1/1000 mutations

Challenges:
- Associate mutations to specific disease
- Deal with huge datasets (noise and statistics)
Schizophrenia is one of the most prevalent, tragic, and frustrating of all human illnesses, affecting about 1% of the human population. Decades of research have failed to provide a clear cause in most cases, but family clustering has suggested that inheritance must play some role.
Searching for the genetic basis of Schizophrenia

Exome sequencing: 2K USD per patient.

Broad institute: 2000 patients per week!

Data here: 2500 healthy & 2500 Schizophrenia patients
• Most rice strains die within a week of complete submergence – a major constraint to rice production in south and southeast Asia.

• Some strains are highly tolerant and survive up to two weeks of complete submergence (no aerobic respiration, no photosynthesis) and renew growth when the water subsides

→The bioinformatics field of ‘computational genetics’
→ found a region near the centromere of chromosome 9, called sub1.
Confirming the submergence tolerance sub1 region

submergence-intolerant strain “Swarna”

submergence-tolerant strain, Sub1 donor

“Swarna”-sub1

Xu et al. 2006
Example 4 - cancer genomics

Network-based stratification of tumor mutations

Hofree et al. Nature methods 2013
Example 5 - Pathogenomics

revolutionizing HIV treatment
There are very efficient drugs for HIV

- Many viruses in blood
- A few viruses in blood
- Many viruses in blood

DRUG, +a few days
DRUG, +more days
Explanation: the virus mutates and some viruses become resistant to the drug.

Solution: combination of drugs (cocktail). But: do not to give drugs for which the virus is already resistant. For example, if one was infected from a person who receives a specific drug.

The question: how do one knows to which drugs the virus is already resistant?
Sequences of HIV-1 from patients who were treated with drug A:

AAGACGCATCGATCGATCGATCGTACG
ACGACGCATCGATCGATCGATCGTACG
AAGACACATCGATCGATCGATCGTACG

Sequences of HIV-1 from patients who were never treated with drug A:

AAGACGCATCGATCGATCGATCGTCTTTACG
AAGACGCATCGATCGATCGATCGATCGTCTTTACG
AAGACGCATCGATCGATCGATCGATCGATCGTCTTTACG
drug A+
AAGACGCA\text{C\text{A\text{C\text{T\text{G\text{A\text{T\text{C\text{G\text{A\text{T\text{C\text{G\text{T\text{A\text{C\text{G\text{G\text{A\text{T\text{C\text{G\text{T\text{A\text{C\text{G}}}}}}}}}}}}}}}}}}}}}}}}}}
AAGACGCA\text{C\text{A\text{C\text{T\text{G\text{A\text{T\text{C\text{G\text{A\text{T\text{C\text{G\text{T\text{A\text{C\text{G}}}}}}}}}}}}}}}}}}}}}}}}}}
AAGAC\text{A\text{C\text{A\text{C\text{T\text{G\text{A\text{T\text{C\text{G\text{T\text{T\text{T\text{T\text{T\text{T\text{T\text{T\text{T\text{T\text{T\text{T\text{T\text{T\text{T\text{T}}}}}}}}}}}}}}}}}}}}}}}}}

drug A-
AAGACGCA\text{C\text{A\text{C\text{T\text{G\text{A\text{T\text{C\text{G\text{A\text{T\text{C\text{G\text{T\text{A\text{C\text{G}}}}}}}}}}}}}}}}}}}}}}}
AAGACGCA\text{C\text{A\text{C\text{T\text{G\text{A\text{T\text{C\text{G\text{T\text{C\text{T\text{T\text{T\text{T\text{T\text{T\text{T\text{T\text{T}}}}}}}}}}}}}}}}}}}}}}}
AAGACGCA\text{C\text{A\text{C\text{T\text{G\text{A\text{T\text{C\text{G\text{T\text{C\text{T\text{T\text{T\text{T\text{T\text{T}}}}}}}}}}}}}}}}}}}}}

This is an easy example.
drug A+
AAGACGCATCGATCGATCGATCGTACG
ACGACGCATCGATCGATCGATCGTACG
AAGACACATCGATCATTCGATCATACG

drug A-
AAGACGCATCGATCTATCGATCTTACG
AAGACGCATCGATCTATCGATCTTACG
AAGACGCATCGATCAATCGATCGTACG

This is NOT an easy example. This is an example of a classification problem.
Genotypic predictors of human immunodeficiency virus type 1 drug resistance


Division of Infectious Diseases, Departments of *Medicine, †Statistics, and ‡Biochemistry, Stanford University, Stanford, CA 94305
Communicated by Bradley Efron, Stanford University, Stanford, CA, August 28, 2006 (received for review December 5, 2005)

Understanding the genetic basis of HIV-1 drug resistance is essential to developing new antiretroviral drugs and optimizing the use of existing drugs. This understanding, however, is hampered by the large numbers of mutation patterns associated with cross-resistance within each antiretroviral drug class. We used five

Results

Drug Susceptibility Results, Input Mutations, and Learning Methods. For each of the three drug classes, we created four mutation sets that included (i) a complete set of all mutations present in ≥2

(ii) a complete set of mutations at (ii) and (iii) a set of

sequences. We trained our models on the

classes.
Example 6 - Rearrangement

Rearrangement is a change in the order of complete segments along a chromosome.
Genome Rearrangements

Challenges:
- Reconstruct the evolutionary path of rearrangements
- Shortest sequence of rearrangements between two permutations
More Examples

- Sequencing cancer genomes
- Large scale proteomics studies
- Single-cell genomics

And much more!

The End