Computational Genomics

Ron Shamir, Roded Sharan
Fall 2017-18
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.

**Cost per Raw Megabase of DNA Sequence**

**Growth of GenBank (1982 - 2008)**

Base Pairs of DNA (billion)

Sequences (millions)
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.

• It is also essential to the new era of precision / personalized medicine: using computational methods for improving disease prevention, diagnosis and treatment.
Research in systems biology

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

High throughput data  
Big data  
Bioinformatics tools/algorithms/methods
The Bioinformatics Actors

- Biotechnology companies
- Academic biotechnology research
- Big Pharmas and Big Agriculture
- National and international research centers

High throughput data
Bioinformatics tools
Administration

• ~5 home assignments as part of a home exam, to be done independently (40% of grade)
• Final exam (60%)
• 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).
Adminisaration (cont.)

• Web page of the course: http://www.cs.tau.ac.il/~rshamir/cg/17/

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

• Revised slide presentations will be posted in the website prior to each class

• Utilize these resources - Avoid taking notes in class!
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
Deoxyribonucleic Acid (DNA)

Double helix

Chromosome
Nucleotides/Bases: Adenine (A), Guanine (G), Cytosine (C), Thymine (T).

Weak hydrogen bonds between base pairs.
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 types), linked by peptide bonds.
- Oriented (from amino to carboxyl group).
- Fold into 3D structure of lowest energy.
Protein structure

**Primary protein structure**
- sequence of a chain of amino acids
- Amino Acids

**Secondary protein structure**
- occurs when the sequence of amino acids are linked by hydrogen bonds
- Pleated sheet
- Alpha helix

**Tertiary protein structure**
- occurs when certain attractions are present between alpha helices and pleated sheets.
- Pleated sheet
- Alpha helix

**Quaternary protein structure**
- consists of more than one amino acid chain.
The Protein Folding Problem
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:
  • de novo / ab initio (= from scratch): extremely hard
  • Homology
  • Threading
Genes

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

The hard disk
One program
Its output

transcription
translation

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

Complementarity: A-U; C-G

Direction of synthesis

Template strand

RNA polymerase

Nontemplate strand

sense

antisense
Transcription of DNA into RNA
Given an RNA sequence, predict its folding = the one that creates a maximum number of matched pairs.

Motivation: RNA function is determined by its 2D structure.

http://www.phys.ens.fr/~wiese/highlights/RNA-folding.html
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

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.
Translation

Modified from Griffiths et al., AN INTRODUCTION TO GENETIC ANALYSIS, 6th Ed., W.H. Freeman & Co., 1996.
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

- 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'
- Mature RNA transcript 5'
- Poly(A) tail
Expression and Regulation

DNA \rightarrow RNA \rightarrow 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
- \(~3,000,000,000\) bases
- \(~20,000\) 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 sequences, find the shortest (super)string containing all of them.

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
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 differences

2 persons: 99.9% similarity

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

- **Sequencing**: reading the sequence of bases in a given DNA or RNA molecule.
- To be sequenced, long sequences must be broken into short segments called “reads”
- **Classical approach**: gel electrophoresis
- **Next-Generation Sequencing**: the modern sequencing techniques, producing many millions of short reads (100-300 nt) per run
One of Many NGS analysis problems

**READ MAPPING:** Given $10^8$ reads, each 100bp long, and a reference genome of length $10^7 - 10^9$ bp, quickly find all the matches of each read in the genome, with differences

- The simple alignment solution: way too slow
- Need better algorithms, sacrificing as little accuracy as possible for far higher speed and smaller space
- An ongoing challenge: By 2025 the amount of DNA sequences is expected to reach $10^{21}$ bp...
Utilize RNA-sequencing and alignment to evaluate RNA levels
Gene Expression analysis

• We can measure the amount of expression of every gene of a person quickly and cheaply, producing her expression profile.

• A working assumption: Expression ~ activity.

• => compare many profiles and infer biology from the commonalities and differences!
Clustering problem

Given the expression profiles of many individuals, partition the profiles into groups such that
- Within each group profiles are similar
- Between different groups profiles are dissimilar
Example: Clustering of B-cell lymphoma samples, no known subtypes

Output: Two molecularly distinct forms of B-cell lymphoma which had distinct gene expression patterns

Question: What is the clinical relevance of these distinct forms?
Evaluate clinical relevance
Kaplan-Meier plot

Fraction of surviving subjects
("survival probability")

The plot presents the fraction of subjects living for a certain amount of time
Kaplan-Meier plot of overall survival of B-cell lymphoma patients grouped on the basis of gene expression profiling.
ADDITIONAL EXAMPLES
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

Association score

Genomic position

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
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

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 NGS of fecal samples from 531 individuals)
Network-based analysis of tumor mutations

Hofree et al. Nature methods 2013
Example 4 - 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 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
AAGACACATCGATCGTTCGATCGTACG

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

AAGACGCATCGATCGATCGATCTTTACG
AAGACGCATCGATCGATCGATCTTTACG
AAGACGCATCGATCGATCGATCTTTACG
drug A+
AAGACGCATCGATCGATCGATCGATCGTACG
ACGACGCATCGATCGATCGATCGATCGTACG
AAGACACATCGATCGTTCGATCGTACG

drug A-
AAGACGCATCGATCGATCGATCGATCGTACG
AAGACGCATCGATCGATCGATCGATCGTACG
AAGACGCATCGATCGATCGATCTTACG

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

Soo-Yon Rhee*, Jonathan Taylor†, Gauhar Wadhera*, Asa Ben-Hur‡§, Douglas L. Brutlag‡, and Robert W. Shafer*∥

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

∥
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