An Introductory Course on BIOINFORMATICS

Liviu Ciortuz
Plan

1. What is bioinformatics? 
   Why should we study it?
2. Bibliography
3. A molecular biology primer 
   3.1 The cell 
   3.2 The DNA 
   3.3 The Central Dogma of molecular biology 
   3.4 Model organisms
4. Exemplifying genetic diseases:
   4.1 Thalassemia 
   4.2 Cystic Fibrosis
5. What you should know; Discovery question
6. Special thanks
What is Bioinformatics?

Bioinformatics is a pluri-disciplinary science focussing on the applications of computational methods and mathematical statistics to molecular biology.

Bioinformatics is also called Computational Biology (USA), Computational Molecular Biology, Computational Genomics.

The related ...ics family of subdomains: Genomics, Proteomics, Phylogenetics, Pharmacogenetics, ...
Why should I teach/study bioinformatics?

Because bioinformatics is an **opportunity** to use some of the most interesting computational techniques...

to **understand** some of the deep mysteries of life and diseases and hopefully to contribute to cure some of the diseases that affect people.

*Note: The next 3 slides are from Thomas Nordahl Petersen, University of Copenhagen*
Example: Parkinson’s disease

a degenerative central nervous disorder
due to the loss of brain cells which produce dopamine, a protein important for the initiation of movement.

Muhammed Ali, Pope John-Paul II died from Parkinson..., my father too
Dopamine produced by cells in **Substantia nigra** activates neurons in **Striatum/Basal ganglia**
Is there a cure for Parkinson’s disease?

Parkinson disease may be cured provided that new dopamine producing cells replace the dead ones. As a medical experiment, dopamine producing brain cells from aborted foetuses have been operated into the brain of Parkinson patients and in some cases cured the disease. Brain tissue from approx. 6 foetuses were needed. Major ethical problems!

Search for a protein drug is the only valid option. The genes producing dopamine are still unknown. Until now, only genes involved in the dopamine transport were identified.
Bibliography for this course

- **Essential Cell Biology**, ch. 1, and 5–7

- **Biological sequence analysis:**
  Probabilistic models of proteins and nucleic acids
  R. Durbin, S. Eddy, A. Krogh, G. Mitchison,
  Cambridge University Press, 1998

- **Problems and solutions in Biological sequence analysis**
  Mark Borodovsky, Svetlana Eksheva
  Cambridge University Press, 2006
“Biological Sequence Analysis” Contents

1. Introduction
3. Hidden Markov Models
2. Alignment of pairs of DNA/protein sequences
4. Alignment of pairs of DNA/protein seq. using HMMs
5. Multiple alignment of DNA/protein sequences
6. Multiple alignment of DNA/protein seq. using HMMs
7–8. Phylogenetics; probabilistic models
9. Probabilistic CFGs
10. Alignment of RNA sequences using PCFGs
11. Background on probability
A Molecular Biology Primer

3.1 The Cell

The cell is the fundamental working unit of every organism.

Instead of having brains, cells make decisions through complex networks of chemical reactions called pathways:

- synthesize new materials
- break other materials down for spare parts
- signal to eat, replicate or die

There are two different types of cells/organisms: Prokariotes and Eukariotes.
Life depends on 3 critical molecules

DNAs — made of A,C,G,T nucleotides (“bases”)
hold information on how a cell works

RNAs — made of A,C,G,U nucleotides
provide templates to synthesize amino-acids into proteins
transfer short pieces of information to different parts of the cell

Proteins — made of (20) amino acids
form enzymes that send signals to other cells and regulate gene activity
make up the cellular structure
form body’s major components (e.g. hair, skin, etc.)
Some basic terminology

**Genome:** the complete set of one organism’s DNA

- a bacteria contains approx. 600,000 base pairs
- human: approx. 3 billion, on 23 pairs of **chromosomes**
- each chromosome contains many genes

**Gene:** the basic functional and physical unit of heredity,
a specific sequence of bases that encode instructions on how to make proteins
Human chromosomes!
3.2 The DNA Helix

Discovered in 1953
(following hints by Erwin Chargaff and Rosalind Franklin) by
James Watson (biologist), and Francis Crick (physicist, PhD std.)
James Watson (1928-),
and
Francis Crick (1916-2005)
Nobel Prize 1962
Rosalind Franklin
1920-1958

The X-ray image of a DNA molecule
DNA copied/“replicated”
3.3 The Central Dogma of Molecular Biology

DNA → RNA → proteins
The Central Dogma of Molecular Biology
Prokariotes vs. Eukariotes
The Central Dogma of Molecular Biology

DNA → RNA → proteins

in Eukariotes
RNA to Amino Acid Coding Table

Each **codon** (triplet of DNA nucleotides) corresponds to one of the 20 amino acids. Among the 64 codons there are a **start** codon and three **stop** codons.

The **redundancy** in the table — one amino acid may be encoded by several different codons — is a kind of **defence against mutations**...

<table>
<thead>
<tr>
<th>First letter</th>
<th>Second letter</th>
<th>Third letter</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>U</td>
<td>C</td>
<td>Phenylalanine</td>
</tr>
<tr>
<td>U</td>
<td>U</td>
<td>C</td>
<td>Leucine</td>
</tr>
<tr>
<td>U</td>
<td>A</td>
<td>G</td>
<td>Glycine</td>
</tr>
<tr>
<td>C</td>
<td>U</td>
<td>C</td>
<td>Serine</td>
</tr>
<tr>
<td>C</td>
<td>U</td>
<td>C</td>
<td>Proline</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>C</td>
<td>Aspartic acid</td>
</tr>
<tr>
<td>A</td>
<td>U</td>
<td>C</td>
<td>Methionine: START codon</td>
</tr>
<tr>
<td>A</td>
<td>U</td>
<td>A</td>
<td>Valine</td>
</tr>
<tr>
<td>G</td>
<td>U</td>
<td>C</td>
<td>Alanine</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenylalanine</td>
</tr>
<tr>
<td>Leucine</td>
</tr>
<tr>
<td>Glycine</td>
</tr>
<tr>
<td>Serine</td>
</tr>
<tr>
<td>Proline</td>
</tr>
<tr>
<td>Aspartic acid</td>
</tr>
<tr>
<td>Methionine: START codon</td>
</tr>
<tr>
<td>Valine</td>
</tr>
<tr>
<td>Alanine</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>A</td>
<td>U</td>
<td>Serine</td>
</tr>
<tr>
<td>U</td>
<td>C</td>
<td>C</td>
<td>Proline</td>
</tr>
<tr>
<td>A</td>
<td>U</td>
<td>C</td>
<td>Aspartic acid</td>
</tr>
<tr>
<td>C</td>
<td>U</td>
<td>A</td>
<td>Valine</td>
</tr>
<tr>
<td>G</td>
<td>U</td>
<td>C</td>
<td>Alanine</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serine</td>
</tr>
<tr>
<td>Proline</td>
</tr>
<tr>
<td>Aspartic acid</td>
</tr>
<tr>
<td>Valine</td>
</tr>
<tr>
<td>Alanine</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>A</td>
<td>C</td>
<td>Tyrosine</td>
</tr>
<tr>
<td>U</td>
<td>U</td>
<td>A</td>
<td>STOP codon</td>
</tr>
<tr>
<td>U</td>
<td>G</td>
<td>C</td>
<td>Cysteine</td>
</tr>
<tr>
<td>A</td>
<td>U</td>
<td>C</td>
<td>Histidine</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>A</td>
<td>Glutamine</td>
</tr>
<tr>
<td>G</td>
<td>U</td>
<td>C</td>
<td>Arginine</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>Asparagine</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>C</td>
<td>Lysine</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>G</td>
<td>Serine</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>C</td>
<td>Arginine</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tyrosine</td>
</tr>
<tr>
<td>STOP codon</td>
</tr>
<tr>
<td>Cysteine</td>
</tr>
<tr>
<td>Histidine</td>
</tr>
<tr>
<td>Glutamine</td>
</tr>
<tr>
<td>Arginine</td>
</tr>
<tr>
<td>Asparagine</td>
</tr>
<tr>
<td>Lysine</td>
</tr>
<tr>
<td>Serine</td>
</tr>
<tr>
<td>Arginine</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>A</td>
<td>U</td>
<td>Threonine</td>
</tr>
<tr>
<td>U</td>
<td>U</td>
<td>A</td>
<td>Alanine</td>
</tr>
<tr>
<td>U</td>
<td>G</td>
<td>C</td>
<td>Glutamic acid</td>
</tr>
<tr>
<td>A</td>
<td>U</td>
<td>A</td>
<td>Aspartic acid</td>
</tr>
<tr>
<td>G</td>
<td>U</td>
<td>C</td>
<td>Glycine</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threonine</td>
</tr>
<tr>
<td>Alanine</td>
</tr>
<tr>
<td>Glutamic acid</td>
</tr>
<tr>
<td>Aspartic acid</td>
</tr>
<tr>
<td>Glycine</td>
</tr>
</tbody>
</table>
A Romanian won the Nobel Prize in molecular biology

George Emil Palade (1912–2008) showed in 1956 that the site of protein manufacturing in the cytoplasm is made of RNA organelles called ribozomes.
3.4 Model organisms

- *Escherichia coli*
- *Saccharomyces cerevisiae*
- *Arabidopsis thaliana*
- *Caenorhabditis elegans*
- *Drosophila melanogaster*
- *Mus musculus*
4 Examples of genetic diseases

4.1 Thalassemia — a genetic disease due to faulty DNA replication

A *mutation* in a gene is a change in the DNA’s sequence of nucleotides.

Sometimes even a mistake of *just one position* can have a profound effect.

Here is a small but devastating mutation in the gene for *hemoglobin*, the protein which carries oxygen in the blood.

*good gene:* AACCAG  
*mutant gene:* AACTAG
The reason, of course, is that the change is reflected in the protein which the gene encodes... First the mRNA comes out. Wrong, and then the protein...

RIGHT

\[
\begin{array}{c}
\text{AAC} \\
\downarrow \\
\text{GLN}
\end{array}
\]

WRONG

\[
\begin{array}{c}
\text{AAC} \\
\downarrow \\
\text{STOP}
\end{array}
\]

This especially disastrous mutation, which interrupts the protein in the middle, causes a serious condition called thalassemia, an inability to make hemoglobin. The victim suffers from a painful lack of oxygen.
Note

In Cyprus, a screening policy — including pre-natal screening and abortion — introduced since 1970s to reduce the incidence of thalassemia, has reduced the number of children born with the hereditary blood disease from 1 out of every 158 births to almost 0.
4.2 Cystic Fibrosis — a genetic disease due to deletion of a triplet in the CFTR gene

The cystic fibrosis disease is characterised by an abnormally high content of sodium in the mucus in lungs, that is life threatening for children.

The cystic fibrosis transport regulator (CFTR) gene adjusts the “waterness” of fluids secreted by the cell.

Due to the deletion of a single triplet in the CFTR gene, the mucus ends up being too thick.
Cystic Fibrosis Transport Regulator (CFTR)

Acknowledgement: this and the next two slides are from Jones & Pevzner

Francis Collins
A fatal mutation in the Cystic Fibrosis Transport Regulator (CFTR) gene
The Cystic Fibrosis Transport Regulator (CFTR) Protein
What you should know

• What is the “Central Dogma” of molecular biology?
• What is the difference between transcription and translation of the DNA message?
• What is a codon?
• Why it is necessary to have a three-letter code?
• How would you define a gene?
• Why can there be more than one possible mRNA sequence for a DNA sequence?
• What is the difference between an intron and an exon?
• What is DNA sequencing?
• What are the positive results of DNA mutations?
Discovery Question:

How do we read DNA sequences?

Knowing how DNA replication works, and assuming that you can get the molecular mass of any given DNA fragment,

design a strategy to get the “reading” of the base composition of an unknown DNA sequence (i.e. the output should be a string over the alphabet $\{A, C, G, T\}$).

What if, due to physical limitations, only fragments of relatively short length (500-700 bases) can be treated in the above way, but the genome that you want to “read” is much larger ($10^6$ or more)?
Short answer:
Fred Sanger’s Method, Nobel Prize, 1980

In 1977 Sanger sequenced the DNA of the FX 174 Phage virus (5386 nucleotides).

From *Discovering Genomics, Proteomics, and Bioinformatics*, Campbell and Hayer, 2006
Scaling up Sanger’s method to whole genome sequencing

Problems:

- limited size of the reads: 500–700 nucleotides
- genomes are much larger (human: $3 \times 10^9$), and contain lots of repeats (human: more than 50%)
- sequencing errors: 1-3%

Solutions:

- use overlapping reads, then assemble them
- BAC-by-BAC sequencing
- using tandem reads to cope with repeats

Recommended reading:

*Bioinformatic Algorithms*, Jones & Pevzner, Ch. 8.
This bioinformatics course would not have been possible without the help of

- the BSc students who took my AI labs on bioinformatics, during the spring 2004 semester:
  Ioana Brudaru, Cristian Prisecariu, Lăcrămioara Aștefănoaiei, ...
- the MSc students, the fall 2005 semester:
  Marta Gîrdea, Oana Rățoi, ...
- MSc students, the fall 2006 semester:
  Sergiu Dumitriu, Diana Popovici, ...
- the BSc students, who took my Bioinformatics course during the spring 2007 semester:
  Ioana Boureanu, Anca Luca, Ștefana Munteanu, Irina Ghiorghiță, Cristian Rotaru, ...
- a former student and colleague of mine who provided me copies of some very good bioinformatics books: Dr. Liliana Ibănescu.
Former students of ours who did or are currently doing PhD’s in bioinformatics

- Raluca Gordân, 2005, Duke University, USA
- Raluca Uricaru, 2005, Université de Monpellier, France
- Marta Gîrdea, 2005, Université de Lille, France
- Luminiţa Moruz, 2005, University of Stockholm, Sweden
- Irina Mohorianu, 2008, University of East Anglia, UK
- Alina Sîrbu, 2008, University of Dublin, UK
- Irina Roznovăţ, 2008, University of Dublin, UK
- Florin Chelaru, 2008, University of Maryland, USA
- [Călin-Rareş Turliuc, 2010, Imperial College of London, UK]
- Alina Munteanu, 2011, University of Iaşi, Romania
- Bogdan Luca, 2012, University of East Anglia, UK
- Claudia Păuleţ (Paicu), 2013, University of East Anglia, UK
Published Papers


Published Papers (cont’d)


Additional Bibliography (I)

- Algorithms on Strings, Trees, and Sequences
  Computer Science and Computational Biology
  Dan Gusfield
  Cambridge University Press, 1997

- Computational Molecular Biology: An Algorithmic Approach
  Pavel Pevzner
  MIT Press, 2000

- Statistical Methods in Bioinformatics: An Introduction
  Warren Ewens, Gregory Grant
  Springer, 2001
  - Introduction to Computational Genomics: A Case Studies Approach
    Nello Cristianini, Matthew Hahn
    Cambridge University Press, 2006

- An Introduction to Bioinformatics Algorithms
  Neil Jones, Pavel Pevzner
  MIT Press, 2004
Additional Bibliography (II), more “Bio...”

- **Essential Cell Biology**, (2nd ed.)
  B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J. Watson
  Garlands, 2005

- **Discovering Genomics, Proteomics, and Bioinformatics**, (2nd ed.)
  Malcolm Campbell, Laurie Hayer
  Benjamin Cummings, 2006

- **Introduction to Bioinformatics**
  Arthur Lesk
  Oxford University Press, 2002

- **Bioinformatics**
  David Mount
  Cold Spring Harbor Laboratory Press, 2001

- **Fundamental Concepts of Bioinformatics**
  Dan Krane, Michael Raymer
  Benjamin Cummings, 2003
Additional Bibliography (III), more “...informatics”

- **Machine Learning Approaches to Bioinformatics**
  Zheng Rong Yang
  MIT Press, 2010

- **Bioinformatics: The Machine Learning Approach**
  Pierre Baldi, Søren Brunak
  MIT Press, 2001

- **Flexible Pattern Matching in Strings:**
  Practical on-line search algorithms for texts and biological sequences
  Gonzalo Navarro, Mathieu Raffinot
  Cambridge University Press, 2002

- **Jewels of Stringology**
  M. Crochemore and W. Rytter
  World Scientific Press, 2002

- **Parallel Computing for Bioinformatics and Computational Biology**
  Alber Zomaya (ed.); Wiley, 2006
Recommended bibliography for laboratory

- **Bioinformatics and Computational Biology Solutions using R and Bioconductor**
  Robert Gentleman, Vincent Carey, Wolfgang Huber, Rafael Irizarry, Sandrine Dudoit
  Springer, 2005

- **Beginning Perl for Bioinformatics**
  James Tisdall
  O’Reilly, 2001

- **Mastering Perl for Bioinformatics**
  James Tisdall
  O’Reilly, 2003