How to enrol for the course?

- Use the registration system of the Computer Science department: [https://ilmo.cs.helsinki.fi](https://ilmo.cs.helsinki.fi)
  - You need your user account at the IT department ("cc account")
  - If you cannot register yet, don’t worry: attend the lectures and exercises; just register when you are able to do so

Teachers

- Esa Pitkänen, Department of Computer Science, University of Helsinki
- Elja Arjas, Department of Mathematics and Statistics, University of Helsinki
- Sami Kaski, Department of Information and Computer Science, Helsinki University of Technology
- Lauri Eronen, Department of Computer Science, University of Helsinki (exercises)

Lectures and exercises

- Lectures: Tuesday and Friday 14.15-16.00 Exactum C221
- Exercises: Tuesday 16.15-18.00 Exactum C221
  - First exercise session on Tue 9 September

Status & Prerequisites

- Advanced level course at the Department of Computer Science, U. Helsinki
- 4 credits
- Prerequisites:
  - Basic mathematics skills (probability calculus, basic statistics)
  - Familiarity with computers
  - Basic programming skills recommended
  - No biology background required
Course contents
- What is bioinformatics?
- Molecular biology primer
- Biological words
- Sequence assembly
- Sequence alignment
- Fast sequence alignment using FASTA and BLAST
- Genome rearrangements
- Motif finding (tentative)
- Phylogenetic trees
- Gene expression analysis

How to pass the course?
- Recommended method:
  - Attend the lectures (not obligatory though)
  - Do the exercises
  - Take the course exam
- Or:
  - Take a separate exam

How to pass the course?
- Exercises give you max. 12 points
  - 0% completed assignments gives you 0 points,
    80% gives 12 points, the rest by linear interpolation
  - "A completed assignment" means that
    - You are willing to present your solution in the
      exercise session and
    - You return notes by e-mail to Lauri Eronen (see
      course web page for contact info) describing the main
      phases you took to solve the assignment
  - Return notes at latest on Tuesdays 16.15
- Course exam gives you max. 48 points

How to pass the course?
- Grading: on the scale 0-5
  - To get the lowest passing grade 1, you need to get at
    least 30 points out of 60 maximum
- Course exam: Wed 15 October 16.00-19.00
  Exactum A111
- See course web page for separate exams
- Note: if you take the first separate exam, the
  best of the following options will be considered:
    - Exam gives you 48 points, exercises 12 points
    - Exam gives you 60 points
- In second and subsequent separate exams, only
  the 60 point option is in use

Literature
- Deonier, Tavaré, Waterman: Computational Genome Analysis, an
  Introduction. Springer, 2005
- Jones, Pevzner: An Introduction to Bioinformatics Algorithms.
  MIT Press, 2004
- Slides for some lectures will be available on the
  course web page

Additional literature
- Gusfield: Algorithms on strings, trees and sequences
- Griffiths et al: Introduction to genetic analysis
- Alberts et al.: Molecular biology of the cell
- Lodish et al.: Molecular cell biology
- Check the course web site
Questions about administrative & practical stuff?

Master's Degree Programme in Bioinformatics (MBI)

- Two-year MSc programme
- Admission for 2009-2010 in January 2009
- You need to have your Bachelor's degree ready by August 2009

MBI programme organizers

- Department of Computer Science, Department of Mathematics and Statistics, Faculty of Science, Kumpula Campus, HY
- Laboratory of Computer and Information Science, Laboratory of CS and Engineering, TKK
- Faculty of Biosciences
- Faculty of Agriculture and Forestry, Viikki Campus, HY
- Faculty of Medicine, Meilahti Campus, HY
- Faculty of Medicine, Meilahti Campus, HY

Four MBI campuses

Admission

- Admission requirements
  - Bachelor's degree in a suitable field (e.g., computer science, mathematics, statistics, biology or medicine)
  - At least 60 ECTS credits in total in computer science, mathematics and statistics
  - Proficiency in English (standardized language test: TOEFL, IELTS)
- Admission period opens in late Autumn 2009 and closes in 2 February 2009
- Details on admission will be posted in www.cs.helsinki.fi/mbi during this autumn

MBI highlights

- You can take courses from both HY and TKK
- Two biology courses tailored specifically for MBI
- Bioinformatics is a new exciting field, with a high demand for experts in job market
- Go to www.cs.helsinki.fi/mbi/careers to find out what a bioinformatician could do for living
Bioinformatics courses in Helsinki region: 1st period

- Computational genomics (4-7 credits, TKK)
- Seminar: Neuroinformatics (3 credits, Kumpula)
- Seminar: Machine Learning in Bioinformatics (3 credits, Kumpula)
- Signal processing in neuroinformatics (5 credits, TKK)

Bioinformatics courses in Helsinki region: 2nd period

- Bayesian paradigm in genetic bioinformatics (6 credits, Kumpula)
- Biological Sequence Analysis (6 credits, Kumpula)
- Modeling of biological networks (5-7 credits, TKK)
- Statistical methods in genetics (6-8 credits, Kumpula)

A good biology course for computer scientists and mathematicians?

- Biology for methodological scientists (8 credits, Meilahti)
  - Course organized by the Faculties of Bioscience and Medicine for the MBI programme
  - Introduction to basic concepts of microarrays, medical genetics and developmental biology
  - Study group + book exam in 1 period (2 cr)
  - Three lectured modules, 2 cr each
  - Each module has an individual registration so you can participate even if you missed the first module
  - www.cs.helsinki.fi/mbi/courses/08-09/bfms/

Bioinformatics courses in Helsinki region: 3rd period

- Evolution and the theory of games (5 credits, Kumpula)
- Genome-wide association mapping (6-8 credits, Kumpula)
- High-Throughput Bioinformatics (5-7 credits, TKK)
- Image Analysis in Neuroinformatics (5 credits, TKK)
- Practical Course in Biodatabases (4-5 credits, Kumpula)
- Seminar: Computational systems biology (3 credits, Kumpula)
- Spatial models in ecology and evolution (8 credits, Kumpula)
- Special course in bioinformatics I (3-7 credits, TKK)

Bioinformatics courses in Helsinki region: 4th period

- Metabolic Modeling (4 credits, Kumpula)
- Phylogenetic data analyses (6-8 credits, Kumpula)

1. What is bioinformatics?
What is bioinformatics?

- Bioinformatics, n. The science of information and information flow in biological systems, esp. of the use of computational methods in genetics and genomics. (Oxford English Dictionary)

- "The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information." -- Fredj Tekaia

What is not bioinformatics?

- Biologically-inspired computation, e.g., genetic algorithms and neural networks
- However, application of neural networks to solve some biological problem, could be called bioinformatics
- What about DNA computing?

Computational biology

- Application of computing to biology (broad definition)
- Often used interchangeably with bioinformatics
- Or: Biology that is done with computational means

Biometry & biophysics

- Biometry: the statistical analysis of biological data
  - Sometimes also the field of identification of individuals using biological traits (a more recent definition)
- Biophysics: "an interdisciplinary field which applies techniques from the physical sciences to understanding biological structure and function" -- British Biophysical Society

Mathematical biology

- Mathematical biology "tackles biological problems, but the methods it uses to tackle them need not be numerical and need not be implemented in software or hardware." -- Damian Counsell

Alan Turing

5
Turing on biological complexity

“It must be admitted that the biological examples which it has been possible to give in the present paper are very limited.

This can be ascribed quite simply to the fact that biological phenomena are usually very complicated. Taking this in combination with the relatively elementary mathematics used in this paper one could hardly expect to find that many observed biological phenomena would be covered.

It is thought, however, that the imaginary biological systems which have been treated, and the principles which have been discussed, should be of some help in interpreting real biological forms.”

– Alan Turing, The Chemical Basis of Morphogenesis, 1952

Related concepts

- **Systems biology**
  - Biology of networks
  - Integrating different levels of information to understand how biological systems work
- **Computational systems biology**

Why is bioinformatics important?

- New measurement techniques produce huge quantities of biological data
  - Advanced data analysis methods are needed to make sense of the data
  - Typical data sources produce noisy data with a lot of missing values
- Paradigm shift in biology to utilise bioinformatics in research

Bioinformatician’s skill set

- **Statistics, data analysis methods**
  - Lots of data
  - High noise levels, missing values
  - #attributes >> #data points
- **Programming languages**
  - Scripting languages: Python, Perl, Ruby, …
  - Extensive use of text file formats: need parsers
  - Integration of both data and tools
- **Data structures, databases**

Bioinformatician’s skill set

- **Modelling**
  - Discrete vs continuous domains
  - Systems biology
- **Scientific computation packages**
  - R, Matlab/Octave, …
- Communication skills!

Communication skills: case 1

Bioligist presents a problem to computer scientists / mathematicians

"I am interested in finding what affects the regulation gene x during condition y and how that relates to the organism's phenotype."

"Define input and output of the problem."
Communication skills: case 2

Bioinformatician is a part of a group that consists mostly of biologists.

Communication skills: case 2

...biologist/bioinformatician ratio is important!

Communication skills: case 3

A group of bioinformaticians offers their services to more than one group.

Bioinformatician’s skill set

How much biology you should know?

Bioinformatician’s skill set

Where would you be in this triangle?

Prof. Juho Rousu, 2006

A problem involving bioinformatics?

- "I found a fruit fly that is immune to all diseases!"
- "It was one of these"

**Molecular biology primer**

*Part 1: What is life made of?*
*Part 2: Where does the variation in genomes come from?*

---

**Life begins with Cell**

A cell is the smallest structural unit of an organism that is capable of independent functioning.

All cells have some common features:

---

**Cells**

- Fundamental working units of every living system.
- Every organism is composed of one of two radically different types of cells:
  - Prokaryotic cells or eukaryotic cells.
- Prokaryotes and Eukaryotes are descended from the same primitive cell.
- All prokaryotic and eukaryotic cells are the result of a total of 3.5 billion years of evolution.

---

**Two types of cells: Prokaryotes and Eukaryotes**

- According to the most recent evidence, there are three main branches to the tree of life.
- Prokaryotes include Archaea ("ancient ones") and bacteria.
- Eukaryotes are kingdom Eukarya and includes plants, animals, fungi and certain algae.

---

**Prokaryotes and Eukaryotes**

Lecture: Phylogenetic trees
All Cells have common Cycles
- Born, eat, replicate, and die

Common features of organisms
- Chemical energy is stored in ATP
- Genetic information is encoded by DNA
- Information is transcribed into RNA
- There is a common triplet genetic code
- Translation into proteins involves ribosomes
- Shared metabolic pathways
- Similar proteins among diverse groups of organisms

All Life depends on 3 critical molecules
- DNAs (Deoxyribonucleic acid)
  - Hold information on how cell works
- RNAs (Ribonucleic acid)
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein
- Proteins
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body’s major components (e.g. hair, skin, etc.)
  - “Workhorses” of the cell

DNA: The Code of Life
- The structure and the four genomic letters code for all living organisms
  - Adenine, Guanine, Thymine, and Cytosine which pair A-T and C-G on complimentary strands.
- Lecture: Genome sequencing and assembly

Discovery of the structure of DNA
- 1952-1953 James D. Watson and Francis H. C. Crick deduced the double helical structure of DNA from X-ray diffraction images by Rosalind Franklin and data on amounts of nucleotides in DNA
- "Photo 51"

DNA, continued
- DNA has a double helix structure which is composed of
  - sugar molecule
  - phosphate group
  - and a base (A,C,G,T)
- By convention, we read DNA strings in direction of transcription: from 5’ end to 3’ end
  - 5’ATTTAGGCC 3’
  - 3’TAATTCGGG 5’
DNA is contained in chromosomes

- In eukaryotes, DNA is packed into chromatids
  - In metaphase, the "X" structure consists of two identical chromatids
- In prokaryotes, DNA is usually contained in a single, circular chromosome

In eukaryotes, DNA is packed into chromatids. In metaphase, the "X" structure consists of two identical chromatids. In prokaryotes, DNA is usually contained in a single, circular chromosome.

Human chromosomes

- Somatic cells in humans have 2 pairs of 22 chromosomes + XX (female) or XY (male) = total of 46 chromosomes
- Germline cells have 22 chromosomes + either X or Y = total of 23 chromosomes

DNA, RNA, and the Flow of Information

- DNA is contained in chromosomes
- RNA is similar to DNA chemically. It is usually only a single strand. T(hyamine) is replaced by U(racil)
- Several types of RNA exist for different functions in the cell.

Length of DNA and number of chromosomes

<table>
<thead>
<tr>
<th>Organism</th>
<th># base pairs</th>
<th>#chromosomes (germline)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prokaryotic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Escherichia coli (bacterium)</td>
<td>4x10^6</td>
<td>1</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae (yeast)</td>
<td>1.35x10^7</td>
<td>1</td>
</tr>
<tr>
<td>Drosophila melanogaster (insect)</td>
<td>1.65x10^8</td>
<td>4</td>
</tr>
<tr>
<td>Homo sapiens (human)</td>
<td>2.9x10^9</td>
<td>23</td>
</tr>
<tr>
<td>Zea mays (corn / maize)</td>
<td>5.0x10^9</td>
<td>10</td>
</tr>
</tbody>
</table>

Hepatitis delta virus, complete genome

RNA

RNA is similar to DNA chemically. It is usually only a single strand. T(hyamine) is replaced by U(racil)

Several types of RNA exist for different functions in the cell.

"The central dogma"

DNA can replicate

Information coded in the sequence of base pairs in DNA is passed to molecules of RNA.

Information in RNA is passed to protein. Information passes from proteins to nucleotides.
Proteins

- Proteins are polypeptides (strings of amino acid residues)
- Represented using strings of letters from an alphabet of 20: AEGLV...WKKLAG
- Typical length 50...1000 residues

Amino acids

How DNA/RNA codes for protein?

- DNA alphabet contains four letters but must specify protein, or polypeptide sequence of 20 letters.
- Dinucleotides are not enough: \(4^2 = 16\) possible dinucleotides
- Trinucleotides (triplets) allow \(4^3 = 64\) possible trinucleotides
- Triplets are also called codons

Proteins: Workhorses of the Cell

- 20 different amino acids
  - different chemical properties cause the protein chains to fold up into specific three-dimensional structures that define their particular functions in the cell.
  - Proteins do all essential work for the cell
  - build cellular structures
  - digest nutrients
  - execute metabolic functions
  - mediate information flow within a cell and among cellular communities.
  - Proteins work together with other proteins or nucleic acids as "molecular machines"
  - structures that fit together and function in highly specific, lock-and-key ways.

Genes

- "A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products" -- Gerstein et al.
- A DNA segment whose information is expressed either as an RNA molecule or protein

Lecture 8: Proteomics
Genes & alleles

- A gene can have different variants
- The variants of the same gene are called alleles

Exons and introns & splicing

- Introns are removed from RNA after transcription
- Exons are joined: This process is called splicing

Alternative splicing

- Different splice variants may be generated

Where does the variation in genomes come from?

- Prokaryotes are typically haploid: they have a single (circular) chromosome
- DNA is usually inherited vertically (parent to daughter)
- Inheritance is clonal
- Descendants are faithful copies of an ancestral DNA
- Variation is introduced via mutations, transposable elements, and horizontal transfer of DNA

Chromosome map of S. dysenteriae, the nine rings describe different properties of the genome
http://www.mgc.ac.cn/ShiBASE/circular_Sd197.htm
Causes of variation

- Mistakes in DNA replication
- Environmental agents (radiation, chemical agents)
- Transposable elements (transposons)
  - A part of DNA is moved or copied to another location in genome
- Horizontal transfer of DNA
  - Organism obtains genetic material from another organism that is not its parent
  - Utilized in genetic engineering

Biological string manipulation

- Point mutation: substitution of a base
  - ...ACGGCT... => ...ACGGCT...
- Deletion: removal of one or more contiguous bases (substring)
  - ...TTGATCA... => ...TTTCA...
- Insertion: insertion of a substring
  - ...GGCTAG... => ...GTTCAACTAG...

Meiosis

- Sexual organisms are usually diploid
  - Germline cells (gametes) contain N chromosomes
  - Somatic (body) cells have 2N chromosomes
- Meiosis: reduction of chromosome number from 2N to N during reproductive cycle
  - One chromosome doubling is followed by two cell divisions

Recombination and variation

- Recap: Allele is a viable DNA coding occupying a given locus (position in the genome)
- In recombination, alleles from parents become shuffled in offspring individuals via chromosomal crossover
- Allele combinations in offspring are usually different from combinations found in parents
- Recombination errors lead into additional variations

Mitosis

- Mitosis: growth and development of the organism
  - One chromosome doubling is followed by one cell division

Recombination frequency and linked genes

- Genetic marker: some DNA sequence of interest (e.g., gene or a part of a gene)
- Recombination is more likely to separate two distant markers than two close ones
- Linked markers: “tend” to be inherited together
- Marker distances measured in centimorgans: 1 centimorgan corresponds to 1% chance that two markers are separated in recombination
Biological databases

- Exponential growth of biological data
- New measurement techniques
- Before we are able to use the data, we need to store it efficiently -> biological databases
- Published data is submitted to databases
- General vs specialised databases
- This topic is discussed extensively in Practical course in biodatabases (III period)

10 most important biodatabases... according to “Bioinformatics for dummies”

- Ensembl  www.ensembl.org  Human/mouse genome
- UniProt  www.expasy.org  Protein sequences
- InterPro  www.ebi.ac.uk  Protein domains
- Enzymes  www.expasy.org  Enzymes
- PDB  www.rcsb.org/pdb/  Protein structures
- KEGG  www.genome.ad.jp  Metabolic pathways

FASTA format

- A simple format for DNA and protein sequence data is FASTA

```
> Hepatitis delta virus, complete genome
atgagccaagttcgaacaagatttcgcggggaggatagatcagcgcccgagaggggtgaga
gtggtaaagaatgagtcagttgtccagttggtttttacaagaaaaggtgaggtgaggtgaggtgaggtgaga
gtcggtaaagagcattggaacgtcggagatacaactcccaagaaggaaaaaagagaaagc
aagaagcggatgaatttccccataacgccagtgaaactctaggaaggggaaagagggaag
gtggaagagaaggaggcgggcctcccgatccgaggggcccggcggccaagtttggaggac
actccggcccgaagggttgagagtaccccagagggaggaagccacacggagtagaacaga
gaaatcacctccagaggaccccttcagcgaacagagagcgcatcgcgagagggagtagaac
acatagcgataggaggggatgctaggagttgggggagaccgaagcgaggaggaaagcaaag
agagcagcggggctagcaggtgggtgttccgccccccgagaggggacgagtgaggctatat
cccggggaactcgacttatcgtccccacatacagactccttcgacccctttcagatga
```