Master Course
Sequence Analysis

Anton Feenstra, Bart van Houte, Walter Pirovano, Jaap Heringa

Centre for Integrative Bioinformatics VU
vrije Universiteit amsterdam

heringa@cs.vu.nl, http://ibi.vu.nl, Tel. 020-5987649, Rm P.128
Bioinformatics staff for this course

- Anton Feenstra – Postdoc (1/09/05)
- Walter Pirovano – PhD (1/09/05)
- Bart van Houte – PhD (1/09/04)
- Jaap Heringa – Grpldr (1/10/02)
Sequence Analysis course schedule

Lectures

[wk 49] 03/12/07 Introduction
[wk 49] 05/12/07 Sequence Alignment 1
[wk 49] 06/12/07 Sequence Alignment 2
[wk 50] 10/12/07 Sequence Alignment 3
[wk 50] 12/12/07 Substitution Matrices
[wk 02] 07/01/08 Multiple Sequence Alignment 1
[wk 02] 09/01/08 Multiple Sequence Alignment 2
[wk 03] 14/01/08 Sequence Entropy
[wk 03] 16/01/08 Sequence Motifs
[wk 04] 21/01/08 Sequence Database Searching 1
[wk 04] 23/01/08 Sequence Database Searching 2
[wk 05] 28/01/08 Genome Analysis
[wk 05] 30/01/08 Phylogenetics
There will be four practical assignments you will have to carry out. Each assignment will be introduced and placed on the IBIVU website:

1. Pairwise alignment (DNA and protein) – assignment 1A, 1B, 1C
2. Multiple sequence alignment (Insulin family)
3. Sequence entropy
4. Database searching
5. Programming your own sequence analysis method (assignment ‘Dynamic programming’ supervised by Bart). If you have no programming experience whatsoever, you can opt out for this assignment. But it’s a ‘must’ for bioinformatics master students.
## Sequence Analysis course final mark

<table>
<thead>
<tr>
<th>Task</th>
<th>Fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Oral exam</td>
<td>1/2</td>
</tr>
<tr>
<td>2. Assignment <em>Pairwise alignment</em></td>
<td>1/10 1/8</td>
</tr>
<tr>
<td>3. Assignment <em>Multiple sequence alignment</em></td>
<td>1/10 1/8</td>
</tr>
<tr>
<td>4. Assignment <em>Sequence Entropy</em></td>
<td>1/10 1/8</td>
</tr>
<tr>
<td>5. Assignment <em>Database searching</em></td>
<td>1/10 1/8</td>
</tr>
<tr>
<td>6. Optional assignment</td>
<td>1/10</td>
</tr>
<tr>
<td><em>Dynamic programming</em></td>
<td></td>
</tr>
</tbody>
</table>

Bioinformaticians and others with programming experience
Gathering knowledge

- Anatomy, architecture
- Dynamics, mechanics
- Informatics (Cybernetics – Wiener, 1948)
  (Cybernetics has been defined as the science of control in machines and animals, and hence it applies to technological, animal and environmental systems)
- Genomics, bioinformatics, Systems Biology
  “The Science of the 21st century”
Bioinformatics

Mathematics
Statistics

Chemistry

Biology
Molecular biology

Computer
Science
Informatics

Physics

Medicine
Bioinformatics

“Studying informational processes in biological systems”
(Hogeweg, early 1970s)

• No computers necessary
• Back of envelope OK

“Information technology applied to the management and analysis of biological data”
(Attwood and Parry-Smith)

Applying algorithms with mathematical formalisms in biology (genomics) -- USA
Bioinformatics in the olden days

• Close to Molecular Biology:
  – (Statistical) analysis of protein and nucleotide structure
  – Protein folding problem
  – Protein-protein and protein-nucleotide interaction

• Many essential methods were created early on (BG era)
  – Protein sequence analysis (pairwise and multiple alignment)
  – Protein structure prediction (secondary, tertiary structure)
Bioinformatics in the olden days (Cont.)

• Evolution was studied and methods created
  – Phylogenetic reconstruction (clustering – NJ method)
But then the big bang....
The Human Genome -- 26 June 2000

Dr. Craig Venter
Celera Genomics
-- Shotgun method

Dr. Francis Collins / Sir John Sulston
Human Genome Project
Saving the HGP

- The ISCB has awarded the Overton Prize for 2003 to W. James Kent, an assistant research scientist at the University of California, Santa Cruz. The award, which recognizes outstanding achievement in the field of computational biology, was presented at ISMB2003, where Kent delivered the annual Overton Lecture on July 1, 2003.
- Kent is best known as the researcher who "saved" the human genome project, a feat chronicled in the New York Times. With little more than a month before the company Celera was to present a complete draft of the human genome to the White House in 2000, Kent wrote GigAssembler, a program that produced the first full working draft assembly of the human genome, which kept the data freely available in the public domain.

http://www.iscb.org/overton.shtml
Human DNA

- There are about $3 \times 10^9$ nucleotides in the nucleus of almost all of the trillions ($5-10 \times 10^{12}$) of cells of a human body (an exception is, for example, red blood cells which have no nucleus and therefore no DNA) – a total of $\sim 10^{23}$ nucleotides!
- Many DNA regions code for proteins, and are called genes (1 gene codes for 1 protein in principle)
- Human DNA contains $\sim 30,000$ expressed genes
- Deoxyribonucleic acid (DNA) comprises 4 different types of nucleotides: adenine (A), thiamine (T), cytosine (C) and guanine (G). These nucleotides are sometimes also called bases
Human DNA (Cont.)

- All people are different, but the DNA of different people only varies for 0.2% or less. So, only 1 letter in ~1400 is expected to be different. Over the whole genome, this means that about 3 million letters would differ between individuals.
- The structure of DNA is the so-called double helix, discovered by Watson and Crick in 1953, where the two helices are cross-linked by A-T and C-G base-pairs (nucleotide pairs – so-called Watson-Crick base pairing).
- The Human Genome has recently been announced as complete (in 2004).
# Genome size

<table>
<thead>
<tr>
<th>Organism</th>
<th>Number of base pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>φX-174 virus</td>
<td>5,386</td>
</tr>
<tr>
<td>Epstein Bar Virus</td>
<td>172,282</td>
</tr>
<tr>
<td><em>Mycoplasma genitalium</em></td>
<td>580,000</td>
</tr>
<tr>
<td><em>Hemophilus Influenza</em></td>
<td>1.8 \times 10^6</td>
</tr>
<tr>
<td>Yeast (<em>S. Cerevisiae</em>)</td>
<td>12.1 \times 10^6</td>
</tr>
<tr>
<td><strong>Human</strong></td>
<td><strong>3.2 \times 10^9</strong></td>
</tr>
<tr>
<td>Wheat</td>
<td>16 \times 10^9</td>
</tr>
<tr>
<td><em>Lilium longiflorum</em></td>
<td>90 \times 10^9</td>
</tr>
<tr>
<td>Salamander</td>
<td>100 \times 10^9</td>
</tr>
<tr>
<td><em>Amoeba dubia</em></td>
<td>670 \times 10^9</td>
</tr>
</tbody>
</table>
Humans have spliced genes...
A gene codes for a protein

DNA

transcription

mRNA

translation

Protein

CCTGAGCCAACTATTGATGAA

CCUGAGCCACUAUUUGAUGAA

PEPTIDE
Orthologous genes are homologous (corresponding) genes in different species (genomes) relating to the speciation event.

Paralogous genes are homologous genes (repeats) within the same species (genome).
Orthology/paralogy

• >50% of the human genome consists of repeats (microsatellites, minisatellites, LINE, SINE, MIR…)
• Many proteins consist of many repeats
  • Sometimes to gain function
  • Sometimes leading to disease (e.g. single-residue repeats)
Fibronectin repeat example

Fig. 3. (A) Schematic outline of rat fibronectin sequence repeats. Three repeat types are involved. (B) Multiple alignment of 15 type I fragments (Patel et al., 1987) which are designated their start and end sequence positions. Amino acid substitution weights used were those from the PAM-250 log odds matrix (Dayhoff et al., 1983) and penalties for gap initiation and extension were respectively 10 and 1. Residues conserved in six or more fragments are shown in black boxes. Similar amino acids are boxed in various grey shades.
Genome revolution has changed bioinformatics

- More high-throughput (HTP) applications (cluster computing, GRID, etc.)
- More automatic pipeline applications
- More user-friendly interfaces
- Greater emphasis on biostatistics
- Greater influence of computer science (machine learning, software engineering, etc.)
- More integration of disciplines, databases and techniques
Protein Sequence-Structure-Function

- Ab initio prediction and folding
- Function prediction from structure
- Homology searching (BLAST)
- Threading
- Structure
- Sequence
Luckily for bioinformatics…

- There are many annotated databases (i.e. DBs with experimentally verified information)
- Based on evolution, we can relate biological macromolecules and then “steal” annotation of “neighbouring” proteins or DNA in the DB.
- This works for sequence as well as structural information
- Problem we discuss in this course: how do we score the evolutionary relationships; i.e. we need to develop a measure to decide which molecules are (probably) neighbours and which are not
- *Sequence – Structure/function gap*: there are far more sequences than solved tertiary structures and functional annotations. This gap is growing so there is a need to predict structure and function.
Some sequence databases

- UniProt (formerly called SwissProt) (http://www.expasy.uniprot.org/)
- PIR (http://pir.georgetown.edu/home.shtml)
- NCBI NR-dataset (all non-redundant GenBank CDS translations+RefSeq Proteins+PDB+SwissProt+PIR+PRF)
- EMBL databank (http://www.ebi.ac.uk/embl/)
- trEMBL databank (http://www.ebi.ac.uk/trembl/)
Boston Globe:

“Using a strategy called 454 sequencing, Rothberg's group reported online July 31 in Nature that they had decoded the genome -- mapped a complete DNA sequence -- for a bacterium in four hours, a rate that is 100 times faster than other devices currently on the market. A second group of researchers based at Harvard Medical School, published a report in last week's Science describing how ordinary laboratory equipment can be converted into a machine that will make DNA sequencing nine times less expensive.

Mapping the first human genome took 13 years and cost $2.7 billion. Current estimates put the cost of a single genome at $10 million to $25 million.”
A bit on divergent evolution

1: ACCTGTAATC
2: ACGTGCAGATC
D = 3/10 (fraction different sites (nucleotides))
A protein sequence alignment
MSTGAVLIY--TSILIKECHAMPAGNE-----
---GGILLLFHRTHELIERKESHAMANDEGGGSNNS
        *       *       *       ****     ***

A DNA sequence alignment
attcgttggcaaatcgccccctatccggcccttaa
att---tggcggatcg-cctctacgggccc----
***       ****       ****       **       *******
A word of caution on divergent evolution

Homology is a term used in molecular evolution that refers to common ancestry. Two homologous sequences are defined to have a common ancestor.

This is a Boolean term: two sequences are homologous or not (i.e. 0 or 1). Relative scales (“Sequence A and B are more homologous than A and C”) are nonsensical.

You can talk about sequence similarity, or the probability of homology. These are scalars.
Convergent evolution

- Often with shorter motifs (e.g. active sites)
- Motif (function) has evolved more than once independently, e.g. starting with two very different sequences adopting different folds
- Sequences and associated structures remain different, but (functional) motif can become identical
- Classical example: serine proteinase and chymotrypsin
- Convergent evolution is now often referred to as non-orthologous displacement
Serine proteinase (subtilisin) and chymotrypsin

- Different evolutionary origins
- Similarities in the reaction mechanisms. Chymotrypsin, subtilisin and carboxypeptidase C have a catalytic triad of serine, aspartate and histidine in common: serine acts as a nucleophile, aspartate as an electrophile, and histidine as a base.
- The geometric orientations of the catalytic residues are similar between families, despite different protein folds.
- The linear arrangements of the catalytic residues reflect different family relationships. For example the catalytic triad in the chymotrypsin subfamily is ordered HDS (histidine, aspartatic acid, serine), but is ordered DHS in subtilisins and SDH in the carboxypeptidase clan.

<table>
<thead>
<tr>
<th>H</th>
<th>D</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>chymotrypsin</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D</th>
<th>H</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>subtilisin</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S</th>
<th>D</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>carboxypeptidase</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
subtilisin and chymotrypsin

Very different tertiary structures…
Functional Genomics

From gene to function

Genome
Expressome
Proteome
Metabolome
Modern bioinformatics is closely associated with genomics

• The aim is to solve the genomics information problem

• Ultimately, this should lead to biological understanding how all the parts fit (DNA, RNA, proteins, metabolites) and how they interact (gene regulation, gene expression, protein interaction, metabolic pathways, protein signalling, etc.)

• Genomics will result in the “parts list” of the genome
New areas interfacing bioinformatics

• Translational Medicine
• Systems Biology
  – Cellular networks
  – Quantitative studies
    • Time processes
    • Cellular compartmentation
    • Multi-scale modelling
  – Link with experiment
• Neurobiology
  – From genome information to behaviour
  – Brain modelling
  – Link with experiment
Translational Medicine

- “From bench to bed side”
- Genomics data to patient data
- Integration
Systems Biology

is the study of the interactions between the components of a biological system, and how these interactions give rise to the function and behaviour of that system (for example, the enzymes and metabolites in a metabolic pathway). The aim is to quantitatively understand the system and to be able to predict the system’s time processes

- the interactions are nonlinear
- the interactions give rise to emergent properties, i.e. properties that cannot be explained by the components in the system
Systems Biology

understanding is often achieved through modeling and simulation of the system’s components and interactions.

Many times, the ‘four Ms’ cycle is adopted:

**Measuring**
**Mining**
**Modeling**
**Manipulating**
METABOLIC PATHWAYS

- Metabolism of Complex Carbohydrates
- Biodegradation of Xenobiotics
- Metabolism of Complex Lipids
- Nucleotide Metabolism
- Metabolism of Other Amino Acids
- Carbohydrate Metabolism
- Amino Acid Metabolism
- Lipid Metabolism
- Metabolism of Cofactors and Vitamins
- Energy Metabolism
- Biosynthesis of Secondary Metabolites
Endomesoderm Specification to 30 Hours

Ubiq = ubiquitous; Mat = maternal; activ = activator; rep = repressor; unkn = unknown; Nucl = nuclearization; γ = γ-catenin source; nβ-TCF = nuclearized b-TCF; ES = early signal; ECNS = early cytoplasmic nuclearization system; Zyg N = zygotic Notch

Copyright © 2001–2006 Hamid Bolouri and Eric Davidson
A system response

Apoptosis: programmed cell death

Necrosis: accidental cell death
Neuroinformatics

• Understanding the human nervous system is one of the greatest challenges of 21st century science.

• Its abilities dwarf any man-made system - perception, decision-making, cognition and reasoning.

• Neuroinformatics spans many scientific disciplines - from molecular biology to anthropology.
Neuroinformatics

- **Main research question:** How does the brain and nervous system work?
- **Main research activity:** gathering neuroscience data, knowledge and developing computational models and analytical tools for the integration and analysis of experimental data, leading to improvements in existing theories about the nervous system and brain.
- **Results for the clinic:** Neuroinformatics provides tools, databases, network technologies and models for clinical and research purposes in the neuroscience community and related fields.
A word on the Bioinformatics Master

- Concerning study points (ECTS), mandatory courses are on half time basis
- You need to combine those with either an optional course, or with an internship (project)
- Talk to your mentor about how to structure your master
Please remember

• DNA makes RNA makes Protein
• Sequence encodes structure encodes function
• “Mind the Gap” - sequence versus Structure and Function