Master Course
Sequence Analysis

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

Sequence Analysis course schedule
Lectures

[wk 44] 30/10/04 Introduction Lecture 1
wk 44 02/11/04 Sequence Alignment 1 Lecture 2
wk 45 06/11/04 Sequence Alignment 2 Lecture 3
wk 45 09/11/04 Sequence Alignment 3 Lecture 4
wk 46 13/11/04 Substitution Matrices Lecture 5
wk 46 16/11/04 Multiple Sequence Alignment 1 Lecture 6
wk 47 20/11/04 Multiple Sequence Alignment 2 Lecture 7
wk 47 23/11/04 Sequence Entropy Lecture 8
wk 48 27/11/04 Sequence Motifs Lecture 9
wk 48 30/11/04 Sequence Database Searching 1 Lecture 10
wk 49 04/12/04 Sequence Database Searching 2 Lecture 11
wk 49 07/12/04 Genome Analysis Lecture 12
wk 50 11/12/04 Phylogenetics Lecture 13
wk 50 14/12/04 Wrapping up Lecture 14

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 Pairwise alignment</td>
<td>1/10 1/8</td>
</tr>
<tr>
<td>3. Assignment Multiple sequence alignment</td>
<td>1/10 1/8</td>
</tr>
<tr>
<td>4. Assignment Sequence Entropy</td>
<td>1/10 1/8</td>
</tr>
<tr>
<td>5. Assignment Database searching</td>
<td>1/10 1/8</td>
</tr>
<tr>
<td>6. Optional assignment Dynamic programming</td>
<td>1/10</td>
</tr>
</tbody>
</table>

Bioinformatics staff for this course

- Anton Feenstra – Postdoc (1/09/05)
- Walter Pirovano – PhD (1/09/05)
- Radek Szklarczyk - PhD (1/03/03)
- Bart van Houte – PhD (1/09/04)
- Jaap Heringa – Grpldr (1/10/02)

Sequence Analysis course schedule
Practical assignments

- There will be four practical assignments you have to carry out. Each assignment will be introduced and placed on the IBIVU website:
  1. Pairwise alignment (DNA and protein) – assignment 1A and 1B
  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.

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
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....
Human DNA

- There are about 3bn (3 x 10^9) nucleotides in the nucleus of almost all of the trillions (5-10 x 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 ~10^23 nucleotides!
- Many DNA regions code for proteins, and are called genes (1 gene codes for 1 protein in principle)
- Human DNA contains ~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).

<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>Mycoplasma genitalium</td>
<td>580,000</td>
</tr>
<tr>
<td>Hemophilus Influenza</td>
<td>1.8 x 10^6</td>
</tr>
<tr>
<td>Yeast (S. Cerevisiae)</td>
<td>12.1 x 10^6</td>
</tr>
<tr>
<td>Human</td>
<td>3.2 x 10^9</td>
</tr>
<tr>
<td>Wheat</td>
<td>16 x 10^9</td>
</tr>
<tr>
<td>Lilium longiflorum</td>
<td>90 x 10^9</td>
</tr>
<tr>
<td>Salamander</td>
<td>100 x 10^9</td>
</tr>
<tr>
<td>Amoeba dubia</td>
<td>670 x 10^9</td>
</tr>
</tbody>
</table>

Humans have spliced genes...

A gene codes for a protein

<table>
<thead>
<tr>
<th>DNA</th>
<th>CCTGAGCCAACTATTGATGAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>transcription</td>
<td></td>
</tr>
<tr>
<td>mRNA</td>
<td>CCGAGCCAAACUAUGAGAA</td>
</tr>
<tr>
<td>translation</td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>PEPTIDE</td>
</tr>
</tbody>
</table>

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
New areas interfacing bioinformatics

- 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

Protein Sequence-Structure-Function

Ab initio prediction and folding

Sequence

Structure

Function

Homology searching (BLAST)

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)
- EMBL databank (http://www.ebi.ac.uk/embl/)
- trEMBL databank (http://www.ebi.ac.uk/trembl/)

Sequence -- Structure/function gap

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

Ancestral sequence

Sequence 1  Sequence 2

1: ACCTGTAATC
2: ACGTGCGATC

D = 3/10 (fraction different sites (nucleotides))

(a) G (b) G
     C     A
     A     C

One substitution - one visible
Two substitutions - one visible

(c) G (d) G
     C     A
     A     C

Two substitutions - none visible
Back mutation - not visible

G
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

Serine proteinase (subtilisin) and chymotrypsin
- Different evolutionary origins, there are
- 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 clan (SA) is ordered HDS, but is ordered DHS in the subtilisin clan (SB) and SDH in the carboxypeptidase clan (SC).

subtilisin and chymotrypsin

Very different tertiary structures…

A protein sequence alignment
MSTGAFLY--TSILIKECHAMPAGNE----- ---GGILLFHRTHELIKEHAMANDEGGSNNS
* * * **** ***

A DNA sequence alignment
attcgttgcaaatgcctctaccgccgctttaa
att---tggcggtatcg-cctctacgggcc----
*** **** **** ** *****

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.)
Functional Genomics

*From gene to function*

- Genome
- Expressome
- Proteome
- Metabolome

---

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