Algorithms in Bioinformatics - Sequences
Algorithms in Bioinformatics - Sequences

Lecturer
Christian Nørgaard Storm Pedersen
E-mail: cstorm@birc.au.dk
Office: 1110.325

Course homepage
http://www.birc.au.dk/~cstorm/courses/AiBS_e14
Lectures

Mondays, 12.15-15.00, in Room 1111-100

Exercises, lectures, discussions

Info about week $X$ available on Wednesday in week $X-1$ at

http://www.birc.au.dk/~cstorm/courses/AiBS_e14/schedule.html
Week 1: Biological sequences. Pairwise global alignment.
Week 2: Other gapcosts. Local alignment.
Week 3: Pairwise alignment in linear space (Dan).
Week 4: Multiple alignment using SP-score, exact and approx (Dan).
Week 5: Multiple alignment, continued.
Week 6: Speeding up dynamic programming (Four Russian).
Week 7: Student presentations.
Literature

Book chapters, research papers etc. will be available via WWW

Dan Gusfield
Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology

Carlos Setubal and Joao Medanis
Introduction to Computational Molecular Biology

Neil C. Jones and Pavel A. Pevzner
An Introduction to Bioinformatics Algorithms
Mandatory Projects and Exam

Besides weekly exercises, there are mandatory projects.

- Small exercise about pairwise alignment (25/8 – 1/9)
- Implementing pairwise alignment (1/9 – 15/9)
- Implementing multiple alignment (15/9 – 29/9)
- Experimenting with multiple alignments (29/9 – 6/10)

Work in groups of 2-3 students

Implementation, experiments, report or presentation, discussion

The exam is an individual 20 minute oral exam which (surprise) includes a presentation of a topic covered in the class.
What is bioinformatics?

Construction and application of algorithms and programs for collecting, handling, and analysis of biological data ...

Biological questions, models of biology, formulation of computational problems, construction of effective algorithms, development of useful tools ...
Multiple disciplines

- Bioinformatics
- Information Technology
- Computer Science
- Statistics
- Biology
- Chemistry
- Medicine
How similar are strings?

Why is this bioinformatics? Why is this a relevant question?

... we need to know some biology ...
The Tree of Life

Similar biochemistry: proteins and nucleic acids (DNA, RNA)

The tree of life web project: http://tolweb.org/tree
The Tree of Life

The tree of life web project: http://tolweb.org/tree

Similar biochemistry: proteins and nucleic acids (DNA, RNA)

The tree of life web project: http://tolweb.org/tree
Proteins

The building blocks of all living organisms

Structural proteins: tissue binding block
Enzymes: catalysts of chemical reactions
Chain of amino acids

\[ S \in \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}^* \]
<table>
<thead>
<tr>
<th>Name</th>
<th>Three- and one-letter code</th>
<th>R group (side chain)</th>
<th>Polarity</th>
<th>Molecular weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala, A</td>
<td>−CH₃</td>
<td>Nonpolar</td>
<td>89.10</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu, L</td>
<td>−CH₂CH(CH₃)₂</td>
<td>Nonpolar</td>
<td>131.18</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile, I</td>
<td>−CH(CH₂CH₃)</td>
<td>Nonpolar</td>
<td>131.17</td>
</tr>
<tr>
<td>Valine</td>
<td>Val, V</td>
<td>−CH(CH₃)₂</td>
<td>Nonpolar</td>
<td>117.15</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro, P</td>
<td>(complete structure)</td>
<td>Nonpolar</td>
<td>115.13</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe, F</td>
<td>−CH₂ −苯基</td>
<td>Nonpolar</td>
<td>165.19</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Trp, W</td>
<td>−CH₂ −嘌呤基</td>
<td>Nonpolar</td>
<td>204.23</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met, M</td>
<td>−CH₃CH₂SCH₃</td>
<td>Nonpolar</td>
<td>149.21</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly, G</td>
<td>−H</td>
<td>Polar(?)</td>
<td>75.07</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser, S</td>
<td>−CH₂OH</td>
<td>Polar</td>
<td>105.10</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr, T</td>
<td>−CHOH</td>
<td>Polar</td>
<td>119.12</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr, Y</td>
<td>−CH₂ −酪氨酸基</td>
<td>Polar</td>
<td>181.19</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys, C</td>
<td>−CH₂S−H</td>
<td>Polar</td>
<td>121.20</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn, N</td>
<td>−CH₂CONH₂</td>
<td>Polar</td>
<td>132.10</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gin, Q</td>
<td>−CH₂CH₂CONH₂</td>
<td>Polar</td>
<td>146.15</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>Asp, D</td>
<td>−CH₂CO₂H</td>
<td>− Charged</td>
<td>133.10</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>Glu, E</td>
<td>−CH₂CH₂CO₂H</td>
<td>− Charged</td>
<td>147.10</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys, K</td>
<td>−CH₂CH₂CH₂CH₂NH₂</td>
<td>+ Charged</td>
<td>146.19</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg, R</td>
<td>−CH₂CH₂CH₂NH−C−NH₂</td>
<td>+ Charged</td>
<td>174.20</td>
</tr>
<tr>
<td>Histidine</td>
<td>His, H</td>
<td>−CH₂ −组氨酸基</td>
<td>+ Charged(?)</td>
<td>155.16</td>
</tr>
</tbody>
</table>
Protein Structure

... predicting structural elements from the sequence is hard ...
Nucleic Acids - RNA and DNA

The carrier of genetic information - The blueprints of proteins
Nucleotides

Nucleotides are the subunits of the nucleic acids.

**Bases**
- Adenine (A)
- Guanine (G)
- Cytosine (C)
- Thymine (T)
- Uracil (U)

**Sugars**
- 2′-deoxyribose (in DNA)
- Ribose (in RNA)
DNA - deoxyribonucleic acid

A DNA molecule is two complementary strands of nucleotides

**DNA:** $s \in \{A,C,G,T\}^*$  
**RNA:** $s \in \{A,C,G,U\}^*$
A DNA molecule is two complementary strands of nucleotides.

**DNA:** $s \in \{A,C,G,T\}^*$

**RNA:** $s \in \{A,C,G,U\}^*$
Cells and Chromosomes

Any organisme consists of one to billions of cells

Chromosomes: Large DNA molecules
Genome: The collection of chromosomes

Human: 23 pairs, 3.100.000.000 bp
Fruit fly: 4 pairs, 200.000.000 bp
Yeast: 16 pairs, 10.000.000 bp
Genes - Blue prints of proteins

Each protein is encoded in a stretch of DNA. A gene ...

Which is expressed when the protein is needed ...

Important problem
Locating genes on the genome and determining how they get expressed ...
## Genetic Code

### Cracked Full Story

<table>
<thead>
<tr>
<th>1st 2nd</th>
<th>3rd 1st 2nd</th>
<th>3rd 1st 2nd</th>
<th>3rd 1st 2nd</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHE</td>
<td>SER</td>
<td>TYR</td>
<td>CYS</td>
</tr>
<tr>
<td>LEU</td>
<td>SER</td>
<td>Ochre</td>
<td>Opal</td>
</tr>
<tr>
<td>PRO</td>
<td>HIS</td>
<td>ARG</td>
<td></td>
</tr>
<tr>
<td>ILEU</td>
<td>THR</td>
<td>ASPN</td>
<td>LYS</td>
</tr>
<tr>
<td>MET</td>
<td>THR</td>
<td>LYS</td>
<td></td>
</tr>
<tr>
<td>VAL</td>
<td>Ala</td>
<td>ASP</td>
<td>Gly</td>
</tr>
<tr>
<td>VAL</td>
<td>Ala</td>
<td>Ala</td>
<td>Gly</td>
</tr>
<tr>
<td>VAL</td>
<td>Ala</td>
<td>Ala</td>
<td>Gly</td>
</tr>
</tbody>
</table>

### Key
- Leu
- Pro
- Histidine
- Tryptophan
- Tyrosine
- Cysteine
- Leucine
- Proline
- Alanine
- Valine
- Lysine
- Glycine

---

...and it is. The code for each of the twenty amino acids is simple, isn't it? Read the table and...
The Central Dogma

\[ s \in \{A,C,G,T\}^* \]

\[ s \in \{A,C,G,U\}^* \]

\[ s \in \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}^* \]

Everything are strings! A good model of important biology.
Biological sequences

Everything are strings! A good model of important biology

Lots of data


Available data grows faster than computational resources ...

Human genome sequence published in Nature and Science in Feb. 2001 ...

3.100.000.000 bp

... influenzae, mouse, rat, zebrafish, fugu, mosquito, fruitfly, C. elegans ...
FASTA format for sequences

From Wikipedia: In bioinformatics, FASTA format is a text-based format for representing both nucleic and protein sequences, in which base pairs or proteins are represented using a single-letter code. The format also allows for sequence names to precede the sequences and for comments (preceded by a semicolon).

>gb:A14666|acc:A14666 PRLB promoter - Bacteriophage lambda, 281 bp.;
;NCBI gi: 579066;
;SEQIO retrieval from GenBank database entry. 07-Feb-1996
gatcagctgc gacacaacta gtttacctac tcgcttatta aaccagaccc acaatctttt
acacagatac aatatataaa gttgaaacctt cttgacatxtt cggcccatga cctttactct
gttataaaatt acttttatgg gggacgatca cactagcaaa ggagttacct aagccccgaa
tgttcaatgg gaagaccttcc ccaatcatga cccacattac gggaaaaataa gtggcggaga
agaagggcgat gttaactgtgc aaagcaatca cagagatgat c

http://en.wikipedia.org/wiki/Fasta_format
A fundamental problem

**Evolution**: genetic material evolve over time by mutations ...

**Biological theory**: the simplest explanation is a good estimate of evolution, i.e. similar sequences are related ...

**A reasonable question**: How similar are sequences?

GGCCTAAAGG CGCCGGTCTTT TCGTACCCCCA AAATCTGGGC ATTTTAAGAT
AAGTGAGTGT TGCCTTACAC TAGCGATCTA CCGCTCTTA TACTTAAGCG
TATGCCCTAGA TCTGACTAAT CGTGCCCCTGG GATTAGACGG GCTTGATGGG
AAAGAACAGC TCGTCTGTTTT ACGTATAAAC AGAATCGCCT GGGTTCGC

GGGCTAAAGG TTAGGGTCTTT TCACTACAAAA GAGTGAGTGGC TATCGTGGCT
AATGTACCAG TCTCTGTATAC GTGGCTTACG GCCAGACCTA CAAGTACTAG
ACCTGAGAAC TAATCTTGTCC GAGCCTTCCA TTGAGGGTAA TGGGAGAGAA
CATCGAGTCA GAAGTTATTC TTGTTTACGT AGAATCGCCT GGGTCCGC
You have been introduced to **DNA**, **RNA**, and **proteins** and the **central dogma** of molecular biology.

You should know that **DNA**, **RNA**, and **protein sequences** can be represented as **strings over finite alphabets**, and how these strings can be represented in **FASTA format**.