Algorithmics and Bioinformatics

Gregory Kucherov and Philippe Gambette

LIGM/CNRS Université Paris-Est
Marne-la-Vallée, France
Schedule

Course webpage: https://wikimpr.dptinfo.ens-cachan.fr/doku.php?id=cours:c-1-32

Lectures on Monday 8:30 – 12:45 between Sept 16 and Nov 4 (except for Oct 28)
Mode of grading


- written exam (2 hours long on Nov 4): coefficient 2
- one programming exercise (optional, deadline Nov 4): coefficient 1

final grade = max(EX, (2EX + PE)/3)
Plan for today

Part 1: Molecular biology primer for computer scientists

Part 2: Basics of sequence algorithmics

- Dynamic programming and sequence comparison
Bioinformatics

From Wikipedia, the free encyclopedia

For the journal, see Bioinformatics (journal).

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to study and process biological data.

Bioinformatics is both an umbrella term for the body of biological studies that use computer programming as part of their methodology, as well as a reference to specific analysis "pipelines" that are repeatedly used, particularly in the fields of genetics and genomics. Common uses of bioinformatics include
Bioinformatics: different viewpoints

Bioinformatics is …
- [for practicians] software tools and associated analysis methods (often statistical)
- [for biological scientists] a way of thinking of biology (evolution) in terms of information processes
  “Nothing in biology makes sense except in the light of evolution”, Theodosius Dobzhansky
  E.Koonin, The logic of chance, FT Press, 2011
- [for computer scientists] an exciting application area
- [for some theoreticians] source of new problems
  “Ask not what mathematics can do for biology; ask rather what biology can do for mathematics!”, Stanislaw Ulam

Bioinformatics is not the same as DNA computing
- cf e.g. L.Adleman’s experiment of 1994
Part 1: Molecular biology primer
- living organisms consist of **cells**
- cells are complex molecular systems hosting cascades of chemical reactions (metabolic pathways)
- **prokaryotic** and **eukaryotic** are distinguished; eukaryotic cells have **nucleus**
Kingdoms of life

prokaryotes

Eubacteria
- Gram positives
- Spirochetes
- Proteobacteria
- Cyanobacteria
- Planctomyces
- Bacteroides
- Cytophaga
- Thermotoga
- Aquifex

Archaebacteria
- Green filamentous bacteria
- Methanosarcina
- Methanobacterium
- Methanococcus
- T. celer
- Gas vacuoles
- Halophiles
- Thermoproteus
- Pyrodictium

eukaryotes

Animalia
- Fungi
  - Ciliates
  - Slime molds
  - Entamoebae
  - Flagellates
  - Trichomonads
  - Microsporidia
  - Diplomonads
  - Protista

Plantae

Tree of life
Three main types of molecules in living cell

- **main biological molecules are polymers**
  - DNA (Deoxyribonucleic acid) consists of *nucleotides*: A (adenine), C (cytosine), G (guanine), T (thymine)
  - RNA (Ribonucleic acid): A,C,G,U (uracil)
Three main types of molecules in living cell

- DNA
  - holds genetic information
  - has double helix structure
  - contained in chromosomes (linear in eukaryotes, often circular in prokaryotes)
Three main types of molecules in living cell

- **DNA**
  - holds genetic information
  - has double helix structure
  - contained in chromosomes (linear in eukaryotes, often circular in prokaryotes)

- **RNA**
  - participates in protein synthesis (mRNA, tRNA, rRNA)
  - has different regulatory functions
Three main types of molecules in living cell

- **DNA**
  - holds genetic information
  - has double helix structure
  - contained in chromosomes (linear in eukaryotes, often circular in prokaryotes)

- **RNA**
  - participates in protein synthesis (mRNA, tRNA, rRNA)
  - has different regulatory functions

- **proteins**
  - do all essential work in the cell
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
- Human body has $10^{13}$-$10^{14}$ cells, all have the same **genome** (set of DNA)
Central dogma of molecular biology

DNA \rightarrow \text{transcription} \rightarrow \text{mRNA} \rightarrow \text{translation} \rightarrow \text{protein}
Genes

- **Gene** = fragment of DNA that encodes a functional RNA or protein product
- Molecular units of heredity defining *phenotypic traits* (e.g. eye colour)
- Human genome contains ~20,000 protein-coding genes taking ~2% of the genome
- A gene can have different variants (*alleles*)
- Two human individuals differ in ~0.1% of DNA
A few alleles of my chromosome 6

Browsing SNPs on Chromosome 6: positions 301537 through 408833 out of 171115067

Jump to a gene: [__] Go  a SNP: [__] Go

or a chromosome:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT

« Return to your whole genome.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Position</th>
<th>SNP</th>
<th>Versions</th>
<th>Grigori Koutcherov's Genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUSP22</td>
<td>312247</td>
<td>rs17133064</td>
<td>A or G</td>
<td>GG</td>
</tr>
<tr>
<td>DUSP22</td>
<td>327331</td>
<td>rs62390189</td>
<td>A or G</td>
<td>GG</td>
</tr>
<tr>
<td>DUSP22</td>
<td>327560</td>
<td>rs112882910</td>
<td>A or G</td>
<td>GG</td>
</tr>
<tr>
<td>DUSP22</td>
<td>335156</td>
<td>rs201346671</td>
<td>C or T</td>
<td>CC</td>
</tr>
<tr>
<td>DUSP22</td>
<td>348124</td>
<td>i6061548</td>
<td>C or T</td>
<td>CC</td>
</tr>
<tr>
<td>intergenic</td>
<td>359508</td>
<td>rs144826800</td>
<td>C or T</td>
<td>CC</td>
</tr>
<tr>
<td>intergenic</td>
<td>363473</td>
<td>rs116139827</td>
<td>C or T</td>
<td>CC</td>
</tr>
<tr>
<td>intergenic</td>
<td>373578</td>
<td>rs9501943</td>
<td>A or G</td>
<td>no call</td>
</tr>
<tr>
<td>intergenic</td>
<td>376962</td>
<td>rs71550004</td>
<td>C or T</td>
<td>CC</td>
</tr>
<tr>
<td>intergenic</td>
<td>378970</td>
<td>rs9501958</td>
<td>A or G</td>
<td>GG</td>
</tr>
<tr>
<td>intergenic</td>
<td>380341</td>
<td>rs6900267</td>
<td>A or C</td>
<td>AA</td>
</tr>
<tr>
<td>intergenic</td>
<td>382897</td>
<td>rs9378774</td>
<td>A or G</td>
<td>AG</td>
</tr>
<tr>
<td>intergenic</td>
<td>383546</td>
<td>rs1033180</td>
<td>C or T</td>
<td>CT</td>
</tr>
<tr>
<td>intergenic</td>
<td>386933</td>
<td>rs74758148</td>
<td>A or G</td>
<td>GG</td>
</tr>
</tbody>
</table>

« Prev | Next »
Gene expression: promoters
Splicing (eukaryotes)
Alternative splicing
<table>
<thead>
<tr>
<th>3-nucleotide codon</th>
<th>Amino acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>UUU</td>
<td>Phenylalanine (Phe)</td>
</tr>
<tr>
<td>UUC</td>
<td>Leucine (Leu)</td>
</tr>
<tr>
<td>UUA</td>
<td>Leucine (Leu)</td>
</tr>
<tr>
<td>UUG</td>
<td>Leucine (Leu)</td>
</tr>
<tr>
<td>CUU</td>
<td>Leucine (Leu)</td>
</tr>
<tr>
<td>CUC</td>
<td>Leucine (Leu)</td>
</tr>
<tr>
<td>CUA</td>
<td>Leucine (Leu)</td>
</tr>
<tr>
<td>CUG</td>
<td>Leucine (Leu)</td>
</tr>
<tr>
<td>AUU</td>
<td>Isoleucine (Ile)</td>
</tr>
<tr>
<td>AUC</td>
<td>Isoleucine (Ile)</td>
</tr>
<tr>
<td>AUA</td>
<td>Isoleucine (Ile)</td>
</tr>
<tr>
<td>AUG</td>
<td>Methionine (Met)</td>
</tr>
<tr>
<td>GUU</td>
<td>Valine (Val)</td>
</tr>
<tr>
<td>GUC</td>
<td>Valine (Val)</td>
</tr>
<tr>
<td>GUA</td>
<td>Valine (Val)</td>
</tr>
<tr>
<td>GUG</td>
<td>Valine (Val)</td>
</tr>
<tr>
<td>UCU</td>
<td>Serine (Ser)</td>
</tr>
<tr>
<td>UCC</td>
<td>Serine (Ser)</td>
</tr>
<tr>
<td>UCA</td>
<td>Serine (Ser)</td>
</tr>
<tr>
<td>UCG</td>
<td>Serine (Ser)</td>
</tr>
<tr>
<td>CAU</td>
<td>Proline (Pro)</td>
</tr>
<tr>
<td>CCC</td>
<td>Proline (Pro)</td>
</tr>
<tr>
<td>CCA</td>
<td>Proline (Pro)</td>
</tr>
<tr>
<td>CGG</td>
<td>Proline (Pro)</td>
</tr>
<tr>
<td>CAU</td>
<td>Histidine (His)</td>
</tr>
<tr>
<td>CAC</td>
<td>Histidine (His)</td>
</tr>
<tr>
<td>CAA</td>
<td>Histidine (His)</td>
</tr>
<tr>
<td>CAG</td>
<td>Histidine (His)</td>
</tr>
<tr>
<td>UAU</td>
<td>Tyrosine (Tyr)</td>
</tr>
<tr>
<td>UAC</td>
<td>Tyrosine (Tyr)</td>
</tr>
<tr>
<td>UAA</td>
<td>Tyrosine (Tyr)</td>
</tr>
<tr>
<td>UAG</td>
<td>Tyrosine (Tyr)</td>
</tr>
<tr>
<td>UGU</td>
<td>Cysteine (Cys)</td>
</tr>
<tr>
<td>UGC</td>
<td>Cysteine (Cys)</td>
</tr>
<tr>
<td>UGA</td>
<td>Cysteine (Cys)</td>
</tr>
<tr>
<td>UGG</td>
<td>Cysteine (Cys)</td>
</tr>
<tr>
<td>CGU</td>
<td>Arginine (Arg)</td>
</tr>
<tr>
<td>CGC</td>
<td>Arginine (Arg)</td>
</tr>
<tr>
<td>CGA</td>
<td>Arginine (Arg)</td>
</tr>
<tr>
<td>CGG</td>
<td>Arginine (Arg)</td>
</tr>
<tr>
<td>AGU</td>
<td>Asparagine (Asn)</td>
</tr>
<tr>
<td>AGC</td>
<td>Asparagine (Asn)</td>
</tr>
<tr>
<td>AGA</td>
<td>Asparagine (Asn)</td>
</tr>
<tr>
<td>AGG</td>
<td>Asparagine (Asn)</td>
</tr>
<tr>
<td>GAU</td>
<td>Aspartic Acid (Asp)</td>
</tr>
<tr>
<td>GAC</td>
<td>Aspartic Acid (Asp)</td>
</tr>
<tr>
<td>GAA</td>
<td>Aspartic Acid (Asp)</td>
</tr>
<tr>
<td>GAG</td>
<td>Aspartic Acid (Asp)</td>
</tr>
<tr>
<td>GGU</td>
<td>Glycine (Gly)</td>
</tr>
<tr>
<td>GGC</td>
<td>Glycine (Gly)</td>
</tr>
<tr>
<td>GGA</td>
<td>Glycine (Gly)</td>
</tr>
<tr>
<td>GGG</td>
<td>Glycine (Gly)</td>
</tr>
</tbody>
</table>

Note: UGA and UAG are stop codons.
Nucleotide sequence of the *gfp10* cDNA from *Aequorea victoria* and the deduced aa sequence (*Prashera et al, Gene 92*). Green arrows: intron positions (spliced out)
RNA

- folds to form a **secondary structure** that is functionally important

**hydrogen bonds:**
- Watson-Crick: A-U, C-G
- weak: G-U, U-C, G-A
Protein structure

**Primary protein structure**
is sequence of a chain of amino acids

**Secondary protein structure**
occurs when the sequence of amino acids are linked by hydrogen bonds

**Tertiary protein structure**
occurs when certain attractions are present between alpha helices and pleated sheets.

**Quaternary protein structure**
is a protein consisting of more than one amino acid chain.
Some numbers

- bacterial genomes are typically $10^6$-$10^7$bp, human genome is $\sim 3 \cdot 10^9$bp
- gene sizes are very variable, up to a few Mbp ($10^6$) in human genome
- transcript (RNA) sizes for human genome are in range 1 to 20 Kbp ($10^3$)
- proteins are typically a few hundreds of amino acids (100-800aa)
Genome sizes

- human: 3G
- maize (corn): 5G
- salamander: 50G
- trumpet lily: 90G
- toad: 6.9G
- polychaos dubium (amoeba): 670G
Sequencing

- first DNA were sequenced in the 70s (1972-76)
- first proteins were sequenced much earlier (1949)

- How DNA is sequenced?
Genome sequencing
Recovery of shredded newspaper
Sanger sequencing

- **Sanger method** (1977-2000s): sequencing fragments of ~800bp
- whole-genome shotgun sequencing:
  - from multiple copies of genomic DNA obtain *inserts* of certain length (a few Kbp)
  - clone inserts into a *vector* used to infect bacteria
  - sequence DNA collected from bacteria using shotgun approach ⇒ *reads*
  - assemble reads
- 1977, first genome sequenced: bacteriophage phi X 174 (5,386bp)
Milestones: full genome sequencing

- **1995**: *H. influenzae* (~1.8 $\cdot$ 10^6bp, ~1700 genes): first full sequenced genome of a free-living organism
- **1998**: *C. elegans* (~100 $\cdot$ 10^6bp, ~20,000 genes)
- **2000**: *D. melanogaster* (~140 $\cdot$ 10^6bp, ~15,000 genes)
- **2000**: *Arabidopsis thaliana* (~135 $\cdot$ 10^6bp, ~27,000 genes)
- **February 2001**: *H. sapiens* (~3,000 $\cdot$ 10^6bp, ~21,000 genes)
  - published in *Nature* (public consortium) and *Science* (Celera company)
  - $2.7$ billion (public project), several years
  - 30M reads of length <800bp, total 24Gbp or ~8-fold coverage of 3Gbp human genome
~2005-…: Next-Generation Sequencing (NGS) produces millions (1M-1G) of short reads (35-400bp) of DNA in hours

Example: Octopus genome (Nature, August 13, 2015), size ~2.7Gbp sequenced with 60x coverage ⇒ total 162 Gbp
## Main NGS technologies

<table>
<thead>
<tr>
<th>Name</th>
<th>Read Lg</th>
<th>Time</th>
<th>Gb/run</th>
<th>pros / cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>454 GS Flex</td>
<td>700</td>
<td>23 h</td>
<td>0.7</td>
<td>long</td>
</tr>
<tr>
<td>Illumina HiSeq</td>
<td>2*100</td>
<td>48 h</td>
<td>120</td>
<td>short/cost</td>
</tr>
<tr>
<td>SOLID (LifeSc)</td>
<td>85</td>
<td>8 d</td>
<td>150</td>
<td>long time</td>
</tr>
<tr>
<td>Ion Proton</td>
<td>200</td>
<td>2 h</td>
<td>100</td>
<td>new</td>
</tr>
<tr>
<td>PacBio Sciences</td>
<td>3-15000</td>
<td>0.3</td>
<td>3</td>
<td>high error rate</td>
</tr>
</tbody>
</table>

**Illumina HiSeq**

**Oxford Nanopore**
### Main NGS technologies

<table>
<thead>
<tr>
<th>Sequencing Platform</th>
<th>Sequencing Generation</th>
<th>Amplification Method</th>
<th>Sequencing Method</th>
<th>Read Length (bp)</th>
<th>Error Rate (%)</th>
<th>Error Type</th>
<th>Number of Reads Per Run</th>
<th>Time Per Run (Hours)</th>
<th>Cost Per Million Bases (USD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanger ABI 3730xl</td>
<td>1</td>
<td>PCR</td>
<td>Dideoxy chain termination</td>
<td>600–1000</td>
<td>0.001</td>
<td>Indel–Substitution</td>
<td>96</td>
<td>0.5–3</td>
<td>500</td>
</tr>
<tr>
<td>Ion Torrent</td>
<td>2</td>
<td>PCR</td>
<td>Polymerase synthesis</td>
<td>200</td>
<td>1</td>
<td>Indel</td>
<td>$8.2 \times 10^7$</td>
<td>2–4</td>
<td>0.10</td>
</tr>
<tr>
<td>454 Roche GS FLX+</td>
<td>2</td>
<td>PCR</td>
<td>Pyrosequencing</td>
<td>700</td>
<td>1</td>
<td>Indel</td>
<td>$1 \times 10^6$</td>
<td>23</td>
<td>8.57</td>
</tr>
<tr>
<td>Illumina HiSeq 2500; high output</td>
<td>2</td>
<td>PCR</td>
<td>Synthesis</td>
<td>$2 \times 125$</td>
<td>0.1</td>
<td>Substitution</td>
<td>$8 \times 10^9$ (paired)</td>
<td>7–60</td>
<td>0.03</td>
</tr>
<tr>
<td>Illumina HiSeq 2500; rapid run</td>
<td>2</td>
<td>PCR</td>
<td>Synthesis</td>
<td>$2 \times 250$</td>
<td>0.1</td>
<td>Substitution</td>
<td>$1.2 \times 10^9$ (paired)</td>
<td>24–144</td>
<td>0.04</td>
</tr>
<tr>
<td>Illumina MiSeq v3</td>
<td>2</td>
<td>PCR</td>
<td>Synthesis</td>
<td>$2 \times 300$</td>
<td>0.1</td>
<td>Substitution</td>
<td>$3 \times 10^8$</td>
<td>27</td>
<td>0.15</td>
</tr>
<tr>
<td>SOLiD 5500xl</td>
<td>2</td>
<td>PCR</td>
<td>Ligation</td>
<td>$2 \times 60$</td>
<td>5</td>
<td>Substitution</td>
<td>$8 \times 10^8$</td>
<td>144</td>
<td>0.11</td>
</tr>
<tr>
<td>PacBio RS II: P6-C4</td>
<td>3</td>
<td>Real-time single-molecule template</td>
<td>Synthesis</td>
<td>$\sim10,000–15,000$</td>
<td>13</td>
<td>Indel</td>
<td>$3.5–7.5 \times 10^4$</td>
<td>0.5–4</td>
<td>0.40–0.80</td>
</tr>
<tr>
<td>Oxford Nanopore MiniION</td>
<td>3</td>
<td>None</td>
<td>Nanopore</td>
<td>$\sim2000–5000$</td>
<td>38</td>
<td>Indel–Substitution</td>
<td>$1.1–4.7 \times 10^4$</td>
<td>50</td>
<td>6.44–17.90</td>
</tr>
</tbody>
</table>

**Illumina HiSeq**

**Oxford Nanopore**
Data deluge in genomics

- Data flow in DNA sequencing surpassed the flow of advances in hardware and storage
- Moore law (chip performance doubling every two years) vs volume of sequencing data (x1.8 every year)
- Next-generation sequencing technologies generate of order $10^{10}$ B (=10GB) per run
- Recall that the sequencing of human genome (2001) used ~24GB which took several years
- Nowadays, a typical sequencing project may use 150-200 GB of sequence data
Number of complete genomes sequenced as for May 2014.
from http://gregoryzynda.com/ncbi/genome/python/2014/03/31/ncbi-genome.html
Cost of sequencing of human-size genome.
from https://en.wikipedia.org/wiki/Personal_genomics
Nucleotide databases

- International Nucleotide Sequence Database Collaboration (insdc.org)
  - GenBank: NCBI, USA
  - ENA: EMBL, Europe
  - DDBJ: Japan
- Contributors: research labs, sequencing programs, …
- All types of nucleotide sequences - chromosomes, contigs, RNAs, EST, … - but also (huge) NGS read collections
EMBL: nucleotide bank growth

Assembled/annotated sequence growth
17-Aug-2015

Sequences (622.7 millions) — Bases (1,381.5 billions)
ENA: Sequence Read Archive growth

Reads growth
17-Aug-2015

Sequences (15.8 trillions) — Bases (2,072.5 trillions)
Today’s datasets reach teras

- 10^6 megas
- 10^9 gigas
- 10^12 teras
- 10^15 petas
- 10^18

- bacterial genome
- human genome
- human genome sequencing project 2001
- typical eukaryotic genome sequencing project in ~2010
- today’s projects

Sequence Read Archive (e.g. TCGA project is ~2.3Pb)

Current worldwide annual sequencing capacity

(from June 2016)
Clouds are used
Impact of NGS

- NGS made genomic studies cheaper, high-throughput and (importantly) genome-wide

- NGS revolutionized genomic research and made (already) possible a number of unprecedented discoveries:
  - metagenomics
  - ncRNA
  - reconstructing ancient genomes (Neanderthal, mammoth, …)
  - 1000 genomes project
  - low-cost sequencing and personalized medicine (23andme, …)
Protein databases

- UniProt (universal protein resource)
  - Swiss-Prot
  - TrEMBL
  - PIR-PSD

- PDB: Protein Data Bank
  - database of protein structures (as well as structures of other macromolecules)
Co-evolution of technologies/algorithms

- DP-based alignment
- Blast-like alignment
- Specialized alignment
- Alignment-free

- Genetics
- Genomics (Sanger)
- Genomics (NGS)
- Pan-genomics

- 90s
- 2007-2009
- these days