Aligning DNA sequences on compressed collections of genomes

Part 1. Reading the DNA: sequencing and assembling

The CODATA-RDA Research Data Science Applied workshop on Bioinformatics

ICTP, Trieste - Italy

July 24-28, 2017

Nicola Prezza

Technical University of Denmark
DTU Compute
DK-2800 Kgs. Lyngby
Denmark

Overview

Overview

The goal of today's lectures is to give an overview of the history, people, techniques, and tools standing behind one of the greatest (still in progress) achievements in human history:
decoding the human genome

Overview

We will explore (in short) the long path that went from the discovery of the molecular structure of DNA to today's most advanced DNA analysis tools.

As we will see, both **biology** and **computer science** played (and still play) a central role in this game

Overview

An overview of the path:

1. Discovering the code: DNA
2. Reading the code: **sequencing and assembling**
3. We have one genome. What about the others? DNA **indexing and alignment**
4. Storing all Human genomes: **data compression**
5. Indexing multiple genomes: **compressed indexes**

Discovering the code: DNA

The discovery of DNA

1869: while trying to isolate and characterize the proteins in white blood cells, swiss chemist Friedrich Miescher discovers a new substance, which he calls *nuclein*:

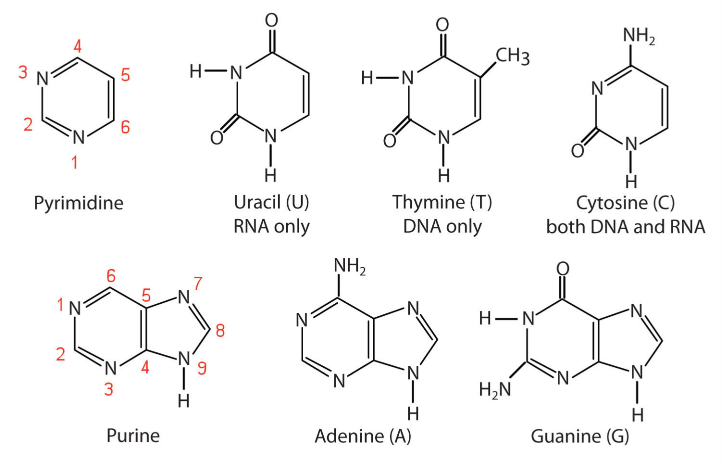
- nuclein is contained in the cell's nuclei
- unlikely proteins, nuclein is not digested by proteolytic enzymes (the guys that digest proteins)
- much higher phosphorous content w.r.t. proteins

The term *nuclein* was later changed to *deoxyribonucleic acid*, or DNA

Scientists later discovered (Walther Flemming, 1878) that DNA is not a single molecule, but is a set of molecules called chromosomes

Relevant for our story is the work of German biochemist Albrecht Kossel (1910 Nobel prize): **DNA is a sequence composed of a series of basic molecules (nucleotides)**

In other words, DNA is a polymer. The monomer units of DNA are Thymine (T), Cytosine (C), Adenine (A), Guanine (G) (RNA, a molecule related to DNA, replaces T with U)



The polymeric structure of DNA and the division in chromosomes are extremely important for us (computer scientists) because they allow us to model DNA as a set of strings on an alphabet of four characters: $\{A, C, T, G\}$. Single characters are called **nucleotides** or **bases**.

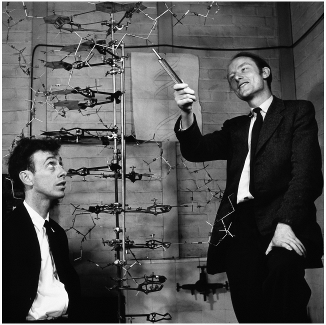
```
Chr1 = ...  CTGGCTCTCAACTTTGTAGATGTAAAAGTTGATTTATCAAT ...
Chr2 = ...  GCTGCGCCCTCCCCGAGCGCGGCTCCAGGACCCCGTCGACC ...
                                ...
ChrY = ...  TTTCCCCGGCGTGTCTGCGGCCATGGTGCGCCCCGCGCCTC ...
```

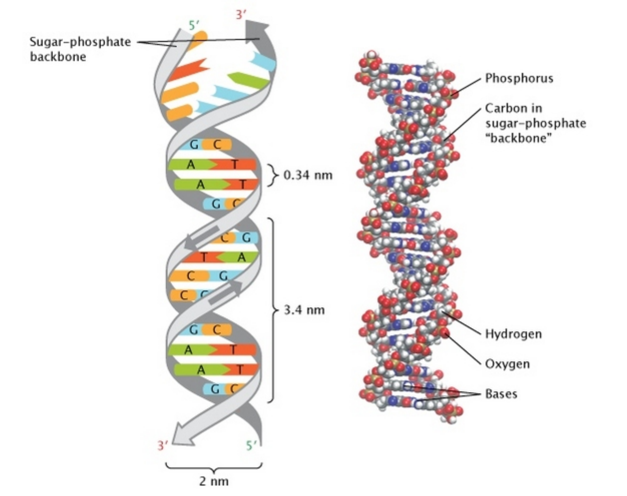
Interestingly enough, until the first half of the 20th century scientists believed that the genetic information was coded in the proteins contained in chromosomes. DNA was regarded as a "too simple molecule" to be the carrier of all life's complexity

In 1944, Oswald Avery, Colin MacLeod and Maclyn McCarty helped demonstrate the role of DNA as the carrier of genetic information

Finally, James Watson and Francis Crick discovered (1953) the three-dimensional structure of DNA (1962 Nobel prize)

Their work built upon results of the American biochemist Linus Pauling (mathematical models of 3D molecular structure using molecular distances and bond angles) and X-ray diffraction results of Rosalind Franklin and her graduate student Raymond Gosling.




Sugar-phosphate backbone
5′
3′
G C
A T
A T
G C
C G
T A
C G
C G
G C
A T
A T
G C
0.34 nm
3.4 nm
3′
5′
2 nm
Phosphorus
Carbon in sugar-phosphate "backbone"
Hydrogen
Oxygen
Bases

The 3D structure of DNA is important for computer scientists for several reasons. The most simple is that DNA bases are paired: C-G and A-T

```
...  CTGGCTCTCAACTTTGTAGATGTAAAAGTTGATTTATCAAT ...
...  GACCGAGAGTTGAAACATCTACATTTTCAACTAAATAGTTA ...
```

The two DNA strands are usually referred to as "Watson" and "Crick".

DNA strands have an orientation (i.e. direction in which they are read by replication enzymes): from 5' to 3'

"Watson" and "Crick" are not only the complement of each other, but also the reverse: they have opposite orientations

```
Watson = 5’ - ... CTGGCTCTCAACTTTGTAGATGTAAAAGTTGA ... - 3’
Crick  = 3’ - ... GACCGAGAGTTGAAACATCTACATTTTCAACT ... - 5’
```

This implies, in particular, that our DNA model is actually a bit more complex: we model DNA as a set of pairs of strings. Every pair is composed of the sequence chromosome and its reverse-complement

For simplicity however, in these lectures we will treat DNA simply as a single string (no chromosomes and their reverse-complements). In the Human genome, the length of this string is approximately equal to $3 \cdot 10^9$ $=$ *3 billion* bases.

DNA sequencing and assembling

How do we read DNA?

Now we know that the human genome is a sequence of 3 billions of letters. How do we obtain this sequence?

First of all, why is this important?

- Discovery of mutations of single individuals w.r.t. the entire population
- Gene annotation: functions of genes
- Genetic testing
- DNA forensics

How do we read DNA?

The genomes of any two individuals are never 100% identical. The "Human genome" therefore must be a **collage** of sequences from different individuals.

Even worse, we do not have the technology to sequence each chromosome in a single run and produce a single sequence.

Standard **DNA sequencers** are able to sequence only $\approx 10^2$ bases at a time. More modern ones increase this to $\approx 10^4$, but are less precise (we will get a closer look in a few slides...)

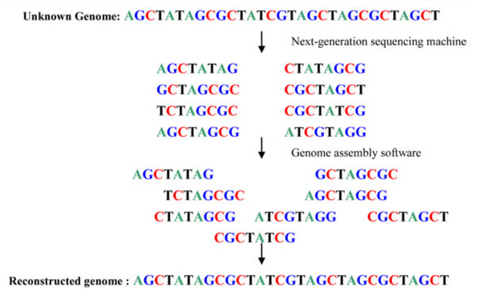
How do we read DNA?

"Reading" a genome

Therefore, the (simplified) general procedure for reconstructing a genome is:

1. **Fragmentation**: break the genome in small pieces (restriction enzymes)
2. **Amplification**: Make a lot of copies of the fragments (Polymerase chain reaction, PCR)
3. **Sequencing**: "Read" each fragment copy with a sequencing machine
4. **Assembly**: put the pieces together with a specialized software (assembler)

Steps 1, 2, 3 are performed in laboratory ("wet lab"). Step 4 is run on a computer ("dry lab")


Unknown Genome: AGCTATAGCGCTATCGTAGCTAGCGCTAGCT
Next-generation sequencing machine
AGCTATAG
CTATAGCG
GCTAGCGC
CGCTAGCT
TCTAGCGC
CGCTATCG
AGCTAGCG
ATCGTAGG
Genome assembly software
AGCTATAG
GCTAGCGC
TCTAGCGC
AGCTAGCG
CTATAGCG
ATCGTAGG
CGCTAGCT
CGCTATCG
Reconstructed genome : AGCTATAGCGCTATCGTAGCTAGCGCTAGCT

How do we read DNA?

Not easy as it seems ...

In practice, the procedure is **much more complicated**:

- DNA is double-stranded (Watson/Crick)
- The sequencer reads both ends ($\approx$ 100 bp) of a fragment ($\approx$ 1000 bp)
- Replication and sequencing errors
- Cutting the genome into pieces using enzymes is not a perfect process (not all regions covered, fragments of different sizes)
- Assembling is a computationally hard problem:
    - Very repetitive genomic regions are hard to reconstruct
    - We do not know exactly the length of the genome
    - There are often multiple ways of assembling the same fragments: how do we choose?

How do we read DNA?

At the end of the 1980s, the technology (PCR+shotgun sequencing+software) was ready. This resulted in a world-wide collaboration:

The Human Genome Project (HGP)

The HGP, started in 1990, used shotgun sequencing and assembling to produce a draft of the Human genome 92% complete and 99.99% accurate. The project was completed in 2003.

Now you can download it on your computer:

`http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz`

A closer look to sequencing: nanopore sequencing

Before passing to the next steps (indexing and compression), let's see closely how DNA sequencing is even possible.

There are several DNA sequencing technologies (more details later). We will focus on one of the most recent and exciting: Nanopore sequencing

Background

DNA sequencing

Problem: determine the sequence of nucleotides within a DNA molecule

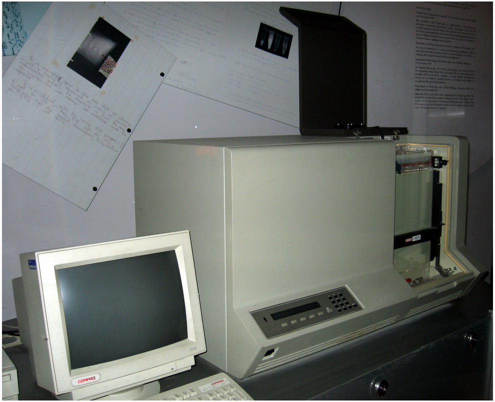
Technology

During the last decades, DNA sequencing technologies underwent dramatic improvements:

- The cost of a 30x sequencing of the Human genome dropped from \$100M (Sanger, 2000) to \$1k (Illumina HiSeq X Ten)
- Length of sequenced fragments increased from 10^2 bp (Sanger, Illumina, SOLiD) to 10^4 bp (PacBio, Oxford Nanopore)
- Throughput increased from 10^3 bp/h (Sanger) to 10^9 bp/h (Illumina)
- Size and cost of sequencing machines *drastically decreased* (next slides) ...

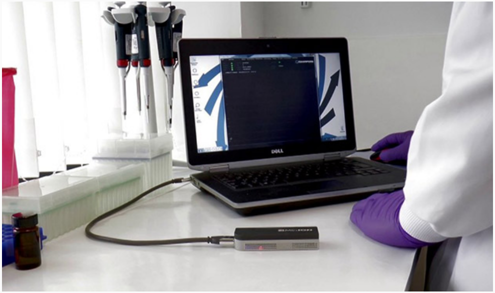
Technology - Sequencer size

From this (Applied Biosystems AB370A, 1987) ...



Technology - Sequencer size

... to this (Oxford Nanopore Technologies MinION, 2014)



In the immediate future:

- Portable clinical genomics (pocket-size? wearable?)
- Personalized medicine
- Routine DNA sequencing

Nanopore Sequencing

Nanopores and DNA sequencing

The idea behind **Nanopore Sequencing** is to "measure" a DNA molecule passing through a nanometer-sized pore

Companies/startups working on NS

- Oxford Nanopore Technologies (ONT)
- Genia Technologies
- Stratos Genomics
- Electronic Biosciences
- ...

We will focus on **ONT nanopore sequencing**

Pros

- Long reads (up to 70 kbp)
- Extremely small and cheap ($1000) sequencing devices
- Little sample preparation needed

The technology still needs to be improved ...

- Average 70%-85% accuracy [1]
- Low throughput (MinION, 10^7 bp/h)[2]

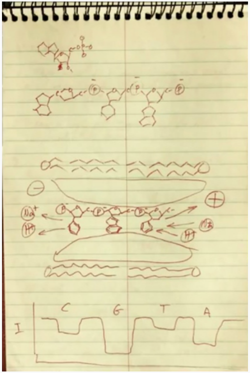
[1] similar to PacBio, and much lower than that—up to 99.8%—of other technologies (e.g. Illumina)

[2] though high-throughput devices are being developed (GridIon, PromethION)

ONT Nanopore Sequencing

Technique

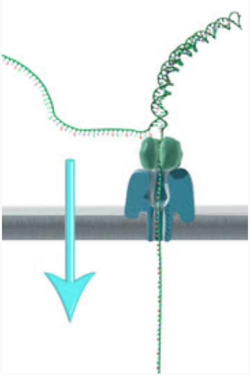
Conceived by the biophysicist David Deamer in 1989



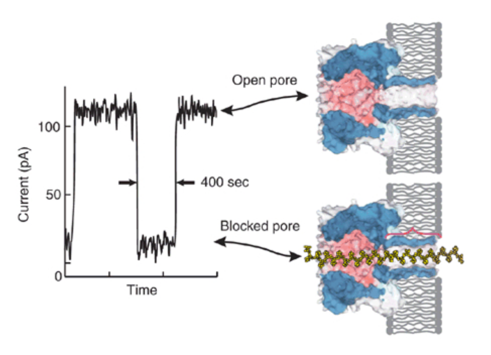
Turned into reality by *Oxford Nanopore Technologies* in 2012

How does it work?

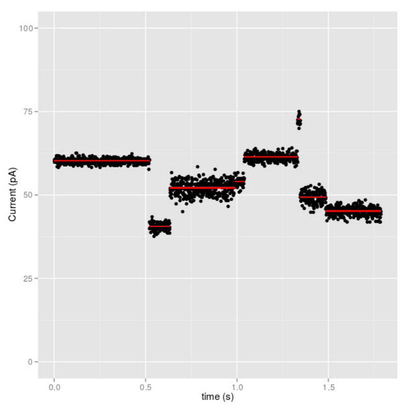
- A *membrane + nanopore* system is immersed in a conducting fluid, and a difference of potential is applied at the sides of the pore
- A DNA molecule slides through the nanopore, pushed by the ions flow



- As DNA slides through the pore, it **modulates the ions flow** passing from one side to the other
- At each moment, $k = 5$ nucleotides occupy the pore. The DNA chain slides at steps of 1 nucleotide at a time
- A **sensor** detects the ions flow at a frequency of 3kHz



- Electric signals are clustered in **real-time** into events by an **event detector**
- Output: a series of events $\beta_1, ..., \beta_m$, each characterized by a mean, standard deviation, and duration.



The *simplicity* of the process, the *absence of bulky and costly signal detectors* (e.g. photosensors), and the *low quantity of reagents* needed makes it possible to build **cheap and small** devices

From DNA *sequencers* to DNA *sensors*

Base calling

Problem (ON base calling)

how to reconstruct the DNA sequence that (most likely) generated

$$\beta_1, ..., \beta_m?$$

What do we know?

- The measured current is expected to be drawn from a **normal distribution**
- e.g. when k-mer TAGCG is in the pore, the measured current is distributed as $N(56.3, 1.2^2)$
- ONT provides a **pore model**, i.e. a set of 4^5 normal distributions (one per k-mer)

Ideal situation

Events $\beta_1, ..., \beta_m$ are in a one-to-one correspondence with contiguous k-mers of the DNA molecule

Ideal situation: example

Sequence: ACTGTTGA ...

$$\begin{array}{ccccccccc}
ACTGT & \rightarrow & CTGTT & \rightarrow & TGTTG & \rightarrow & GTTGA & \rightarrow & ... \\
\Downarrow & & \Downarrow & & \Downarrow & & \Downarrow & & ... \\
\beta_1 & \rightarrow & \beta_2 & \rightarrow & \beta_3 & \rightarrow & \beta_4 & \rightarrow & ...
\end{array}$$

Complications

Reality is far from ideal! we can have skipped k-mers, oversegmentation, noise, backslips, ...

Complications

1. **Skip/undersegmentation**: a k-mer does not emit a signal

$$ACTGA(\beta_i) \rightarrow CTGAT \rightarrow TGATA(\beta_{i+1})$$

2. **Oversegmentation**: k-mer emits multiple signals.

$$ACTGA(\beta_i) \rightarrow ACTGA(\beta_{i+1})$$

3. **backslip**: the DNAP enzyme often slips backwards due to the tension applied by the voltage across the nanopore.

$$ACTGA(\beta_i) \rightarrow GACTG(\beta_{i+1}) \rightarrow ACTGA(\beta_{i+2})$$

4. **Noise**: a signal that does not correspond to any k-mer.

$$ACTGA(\beta_i) \rightarrow (\beta_{i+1}) \rightarrow CTGAT(\beta_{i+2})$$

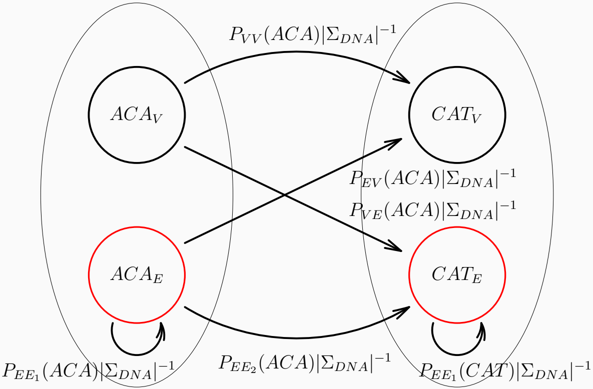
A solution: HMM

This problem can be solved with a Hidden Markov Model:

- **States**: *k*-mers. We need emitting/silent states
- **Transitions**: must model emissions, skipped *k*-mers, oversegmentation, noise, backslips.

Base calling

The base calling problem can be solved by running the **Viterbi algorithm** on the HMM. Output: path (i.e. DNA sequence) that most likely generated the observed events.


$P_{VV}(ACA)|\Sigma_{DNA}|^{-1}$
ACA_V
CAT_V
$P_{EV}(ACA)|\Sigma_{DNA}|^{-1}$
$P_{VE}(ACA)|\Sigma_{DNA}|^{-1}$
ACA_E
CAT_E
$P_{EE_1}(ACA)|\Sigma_{DNA}|^{-1}$
$P_{EE_2}(ACA)|\Sigma_{DNA}|^{-1}$
$P_{EE_1}(CAT)|\Sigma_{DNA}|^{-1}$

Portion of a HMM solving the problem (in this example, $k = 3$). In the complete HMM, two k-mers are adjacent iff they have a $(k-1)$-bases overlap. V states (black) are silent, while E states (red) are emitting.

Conclusions

As understood, assembling a genome is not an easy task. Clearly, we cannot afford doing this for every human being.

Luckily for us, one draft of the genome gives us a toe-hold for efficiently and cheaply reconstructing any other human genome (without assembling) ... how?

... the solution in the next lecture: **indexing and alignment**