Mining and Processing Biomedical Data

Dr. rer. nat. Krisztian Buza
adiunkt naukowy
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw, Poland
chrisbuza@yahoo.com

Analysis of NGS data

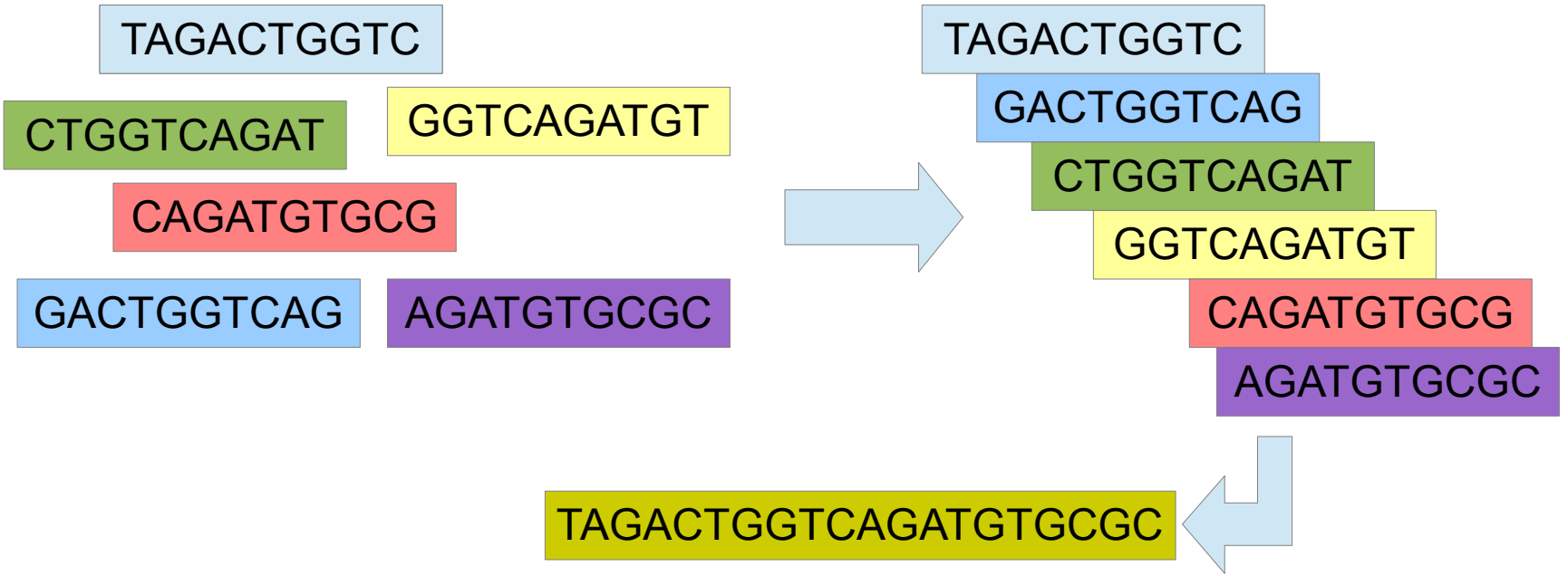
Outline of the Module

- NGS data = Next Generation Sequencing data
- Introduction to
  - Genome Assembly
  - Variant Analysis
  - Analysis of ChipSeq data
  - Analysis of methylation data

ASSISTED ASSEMBLY PROBLEM

Genome Assembly

- Goal: determine the DNA sequence of an organism
- Short *reads* have to be put together like a puzzle (using the overlaps between them)



The output of (assisted) assembly algorithms

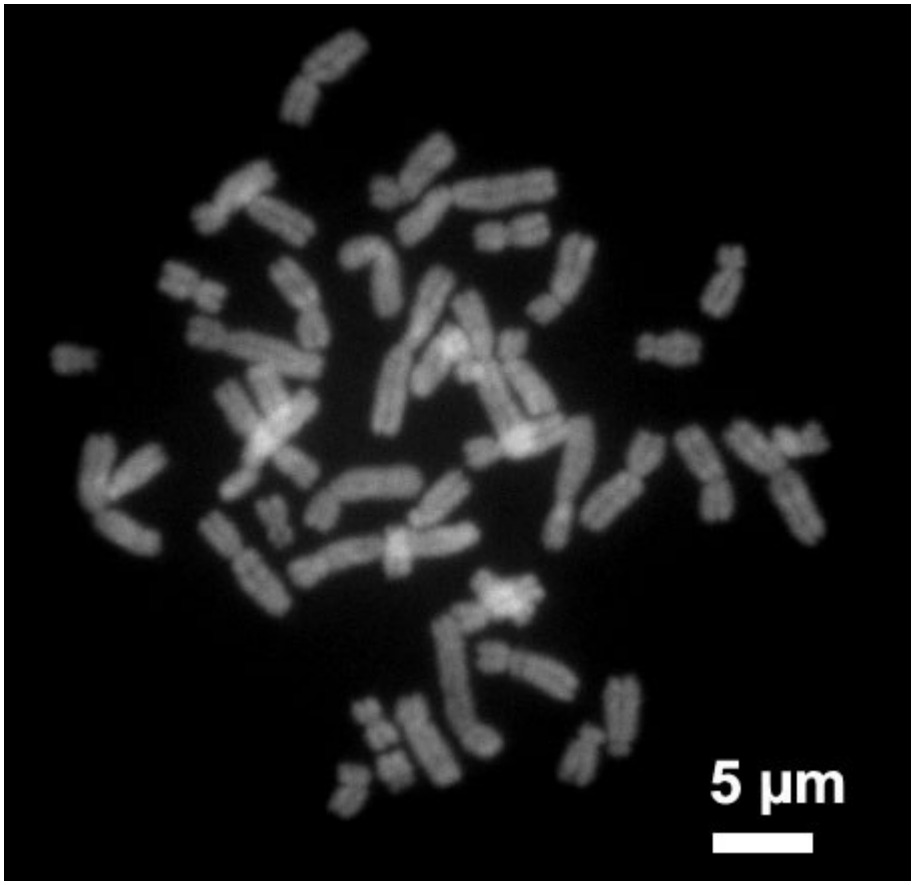


Image: from Wikipedia

- Ideally: a few sequences of A,C,G,T, each one corresponding to a chromosome
- In reality: a set of relatively large sequences (contigs, scaffolds).

Working with NGS data: some of the challenges

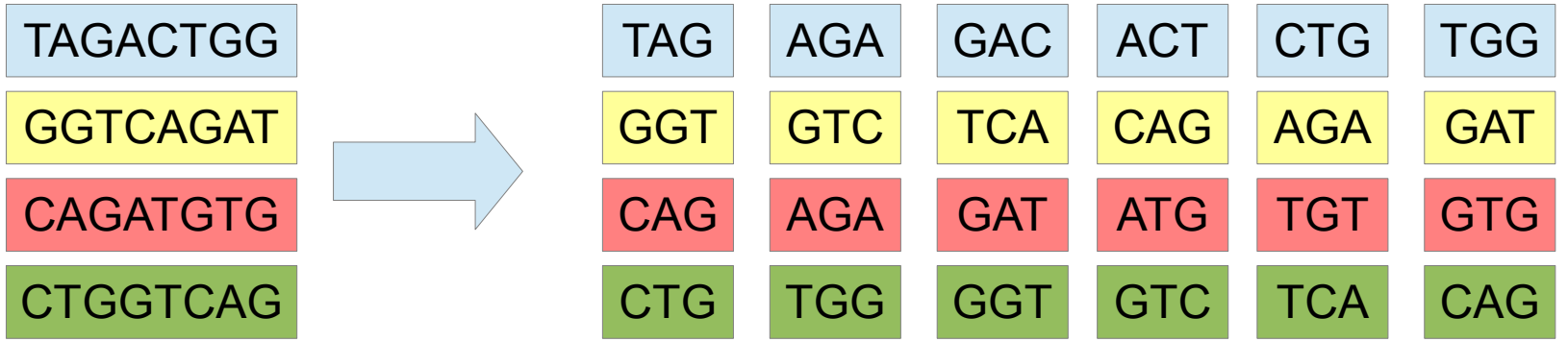
- Read errors – quality scores
- Paired end reads
- Repeats
  - How many times is a pattern repeated?
- Reads are randomly distributed
  - Length of overlaps is often different
  - Some regions might be uncovered

DE BRUIJN GRAPHS FOR GENOME ASSEMBLY

k-Mers

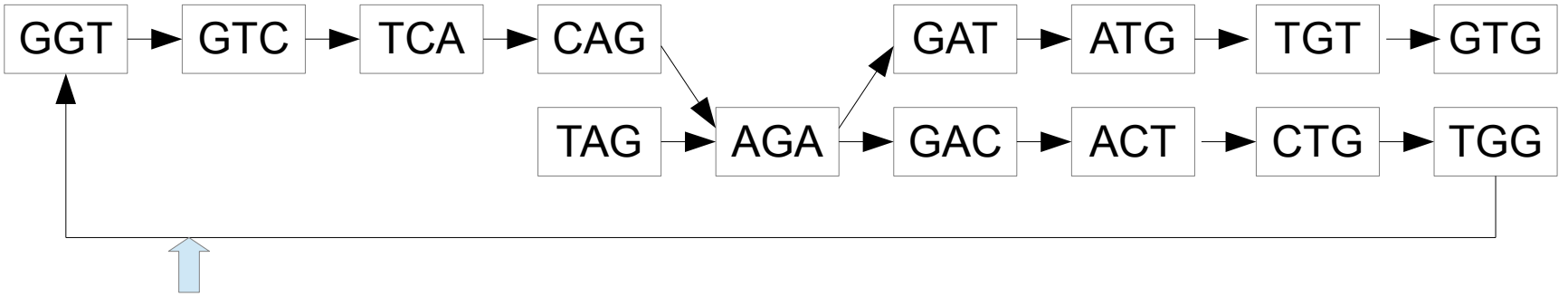
- *k*-Mer: Nucleotid sequence of length *k*
- For example
  ATC, GTT, AGA are 3-Mers
  GCTTC, TCAGG, CGCGA are 5-Mers

De Bruijn Graphs for Genome Assembly



Short reads

K-mers (overlapping subsequences of the short reads)

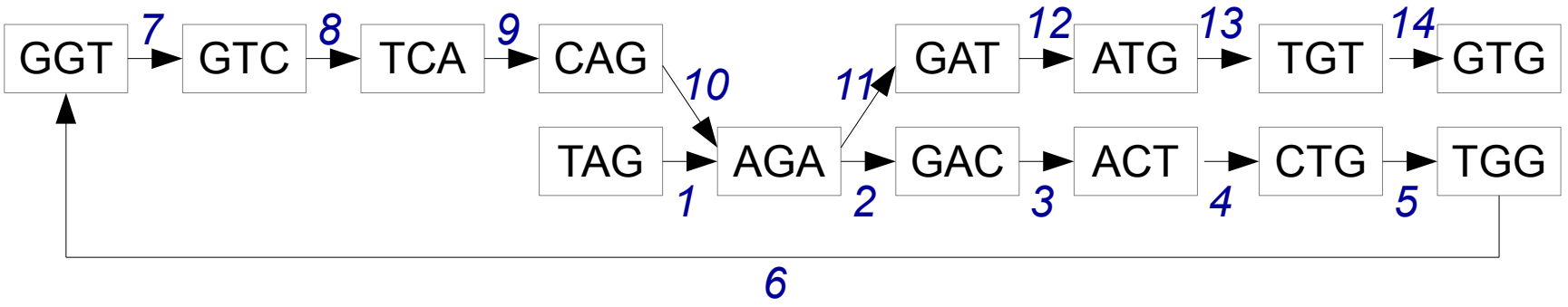


de Bruijn graph

In the last (green) short read, the k-mer "GGT" is subsequent to "TGG". This is represented by this edge.

De Bruijn Graphs for Genome Assembly

de Bruijn graph



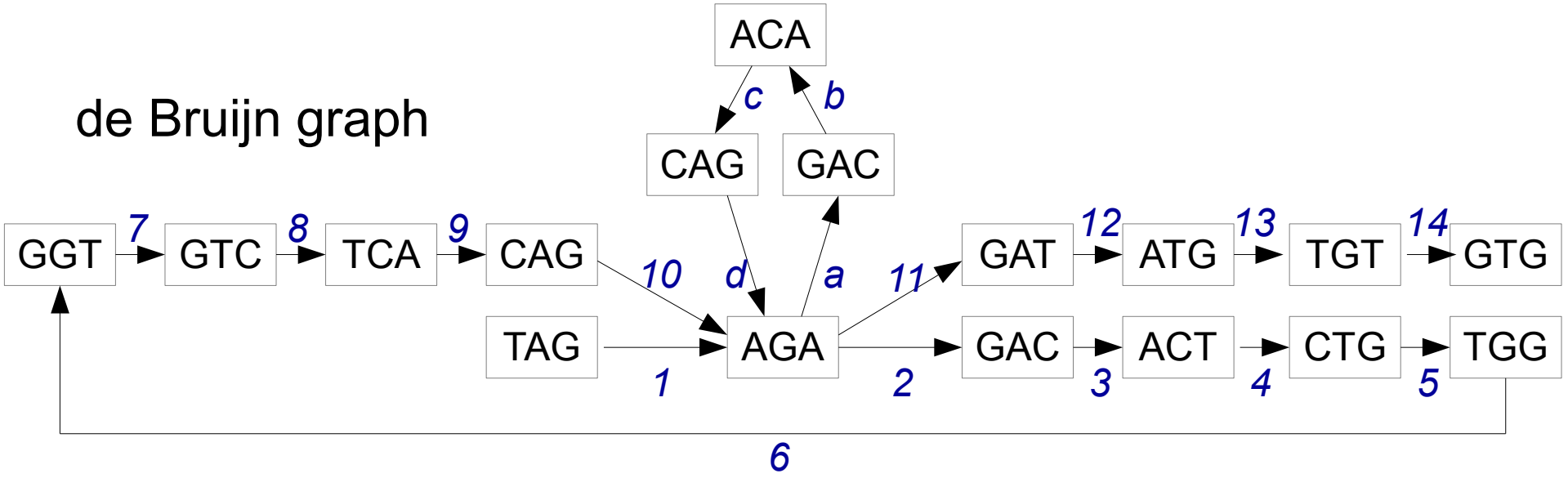
Edges are numbered according to their order in the Euler-path.

Euler-path: traverses through all the edges, each edge is visited exactly once, the same node may be visited several times

The Euler-path corresponds to the result of the assembly:
TAGACTGGTCAGATGTG

De Bruijn Graphs for Genome Assembly

de Bruijn graph



The Euler-path is not always unique:
in this graph edges may be visited in two orders:

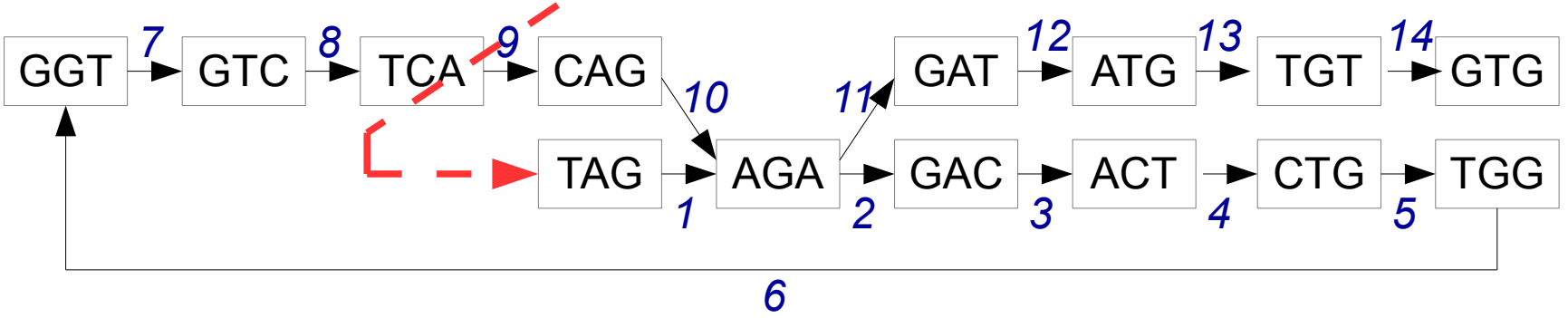
1, **a, b, c, d,** 2, 3, 4, …, 14 and

1, 2, 3, … 9, 10, **a, b, c, d**, 11, 12, 13, 14

The Intuition behind Reference-assisted Assembly

“CGCTAGG”

“TAGACTGGTCAGATGTG”



The genome of a related organism (called “reference”)

...ACATCAG**CTAGGTAGACT**GAAGGTACCAGA...

Finally, the contig “CGCTAGGTAGACTGGTCAGATGTG” is produced.

Assisted Assembly

- For the description of a reference-assisted genome assembly, see:

  Nathaniel Parrish, Benjamin Sudakov and Eleazar Eskin (2013): *Genome reassembly with high-throughput sequencing data*, The Eleventh Asia Pacific Bioinformatics Conference

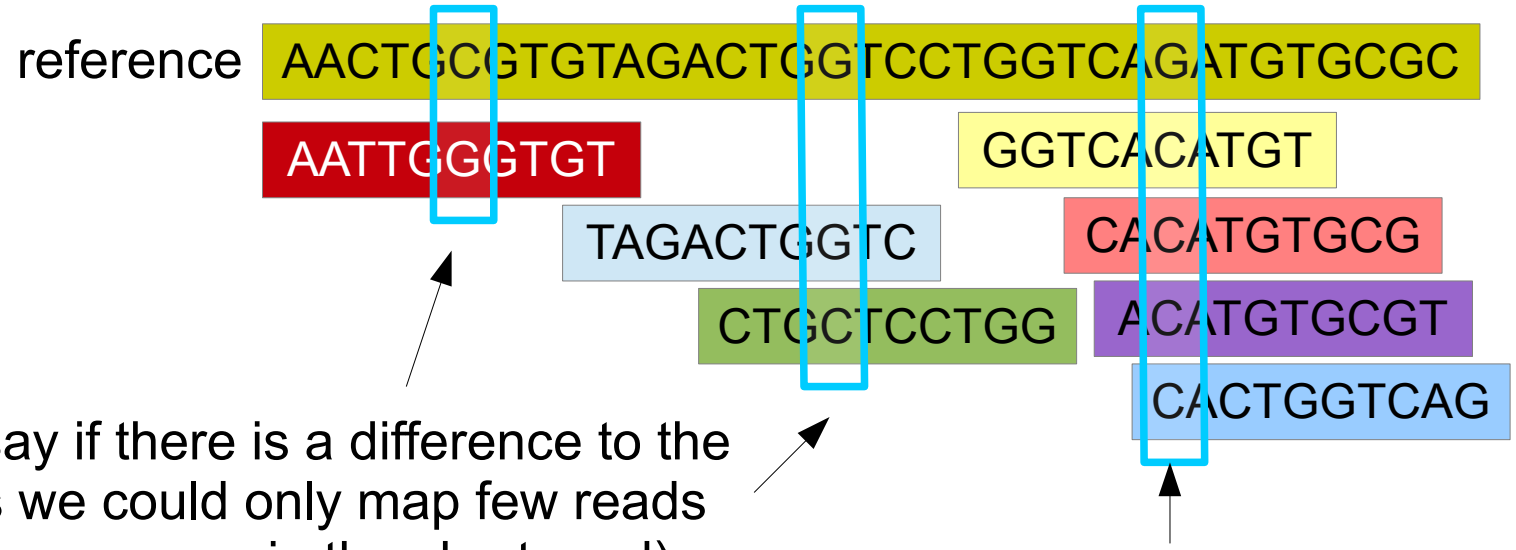
- You may check out the FastQ and FastA file formats (see e.g. Wikipedia), if you are interested in how the short reads and the contigs are usually stored.
- A few software tools for genome assembly:

  Velvet, Amos, …

VARIANT ANALYSIS

Variation Analysis

- DNA of each individual is slightly different from the “reference” DNA of the species
- Various types of differences
- Single Nucleotid Polymorphisms (SNPs) can be detected via mapping short reads to the reference



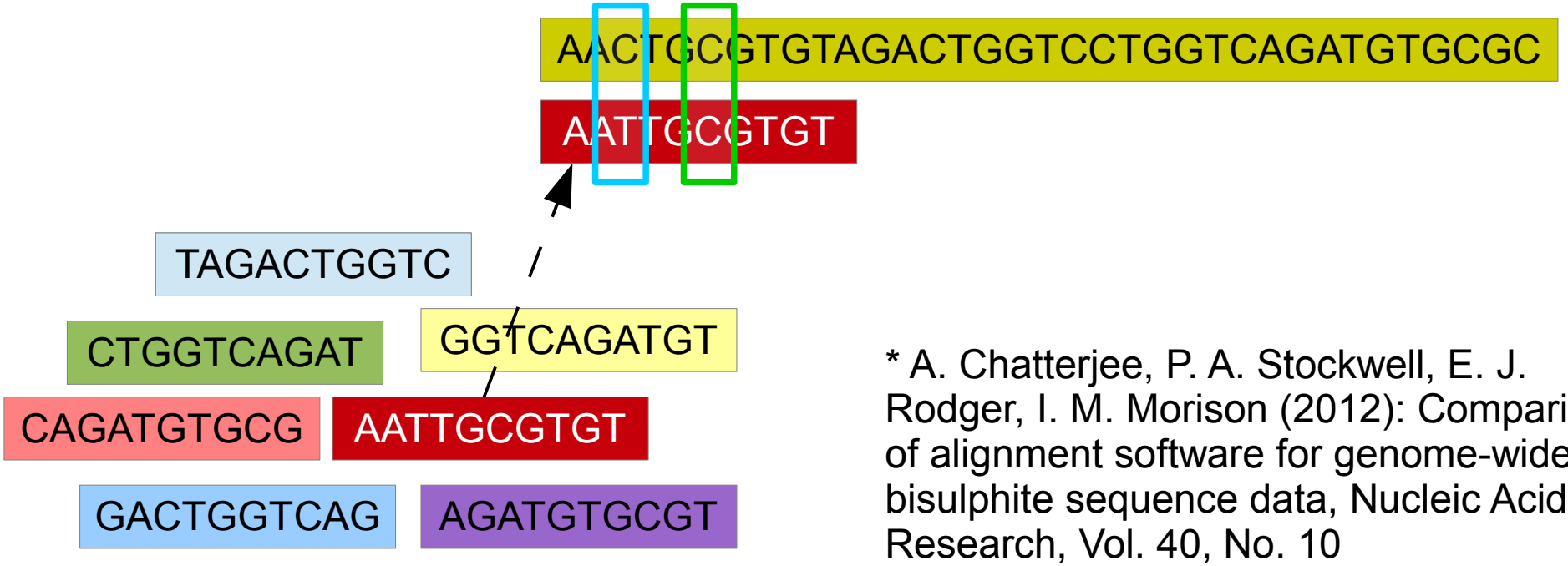
We can not say if there is a difference to the reference, as we could only map few reads (there might be an error in the short read)

At this position, the genome of this individual is probably different from the reference

ANALYSIS OF DNA-METHYLATION

DNA methylation

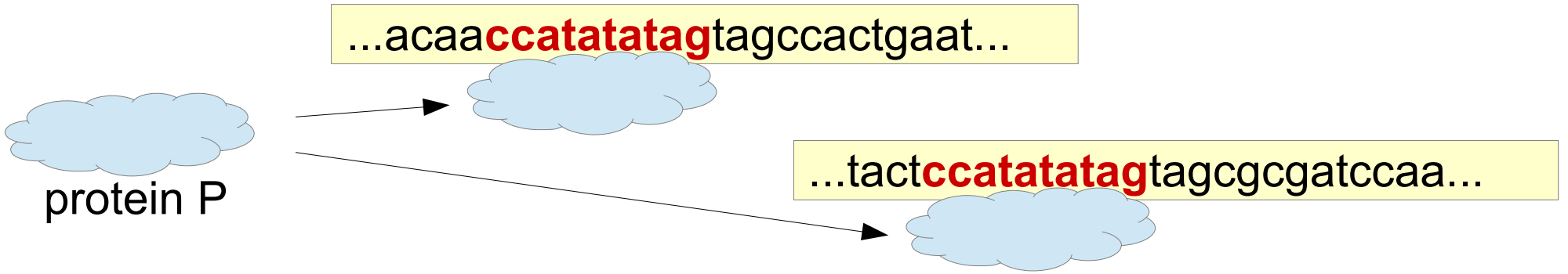
- One of the mechanisms to turn genes “on” / “off”
- Methylated C: a methyl group is added to the C
- Key feature: “Sodium bisulphite treatment of DNA converts unmethylated C into T …, but methylated Cs remain unchanged.”



* A. Chatterjee, P. A. Stockwell, E. J. Rodger, I. M. Morison (2012): Comparison of alignment software for genome-wide bisulphite sequence data, Nucleic Acids Research, Vol. 40, No. 10

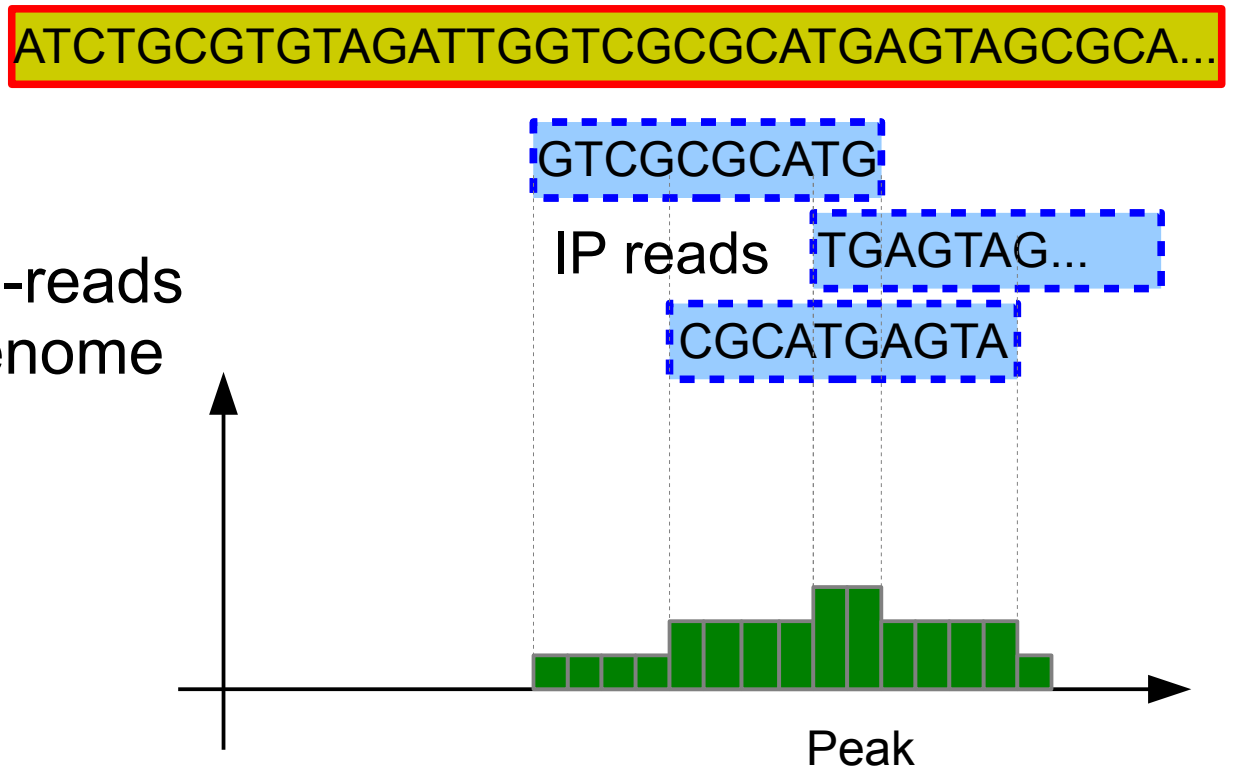
ChIP-Seq

ChIP-Seq Experiments at the first sight
(from the point of view of data processing)



- We want to determine the positions where protein P binds to the DNA
- INPUT reads: short sequences more-less uniformly distributed over the genome
- IP-reads (immunoprecipitated reads): short sequences from the regions where the protein binds

Analysis of ChIP-Seq data



- Method:
  - Map INPUT and IP-reads to the reference genome (Bowtie)
  - Detect peaks (MACS)
- “Challenges”:
  - DNA is double stranded (reads from any of them, “peaks” do not look like peaks)
  - Read errors → approximate matching
  - Coverage not uniform: map both IP reads and INPUT reads
  - Variations between reference genome and target genome

Final Remarks

- This lecture only presents the intuition behind various analysis tasks!
- Some of the software tools
  - Assembly: Velvet, Amos
  - Mapping: Bowtie, Bowtie2, BWA
  - Methylation analysis: Bismark
  - ChIP-Seq (peak calling): MACS