UNIVERSITÀ DI PISA

DNA polymorphisms and RNA-Seq alternative splicing blow bubbles in de Bruijn Graphs

Nadia Pisanti

University of Pisa & Leiden University

Universiteit Leiden

Outline

New Generation Sequencing (NGS), and the importance of detecting DNA polymorphisms and RNA alternative splicing events.

de Bruijn graphs: definition, examples, and why SNPs and AS events correspond to *bubbles* in the de Bruijn graph.

An optimal algorithm to detect bubbles (in any kind of graph).

Experiments on RNA-Seq data and comparison with other tools specifically designed for this task.

Why sequencing?

The knowledge of DNA and RNA sequences has become a crucial tool for basic research in biology, pharmacology and medicine.

With NGS (New Generation Sequencing), experiments are much larger and cheaper, opening the way to many new applications.

More and more is understood in terms of correlations between DNA sequence, congenital diseases and predisposition to diseases or to drug response... and more and more in needed to investigate!

What is NGS great for

- re-sequencing: no assembly, just mapping on a known reference genome.
- Metagenomics
- Transcriptome Sequencing: RNA-Seq
- Chromatin immunoprecipitation combined with DNA sequencing: ChIP-Seq

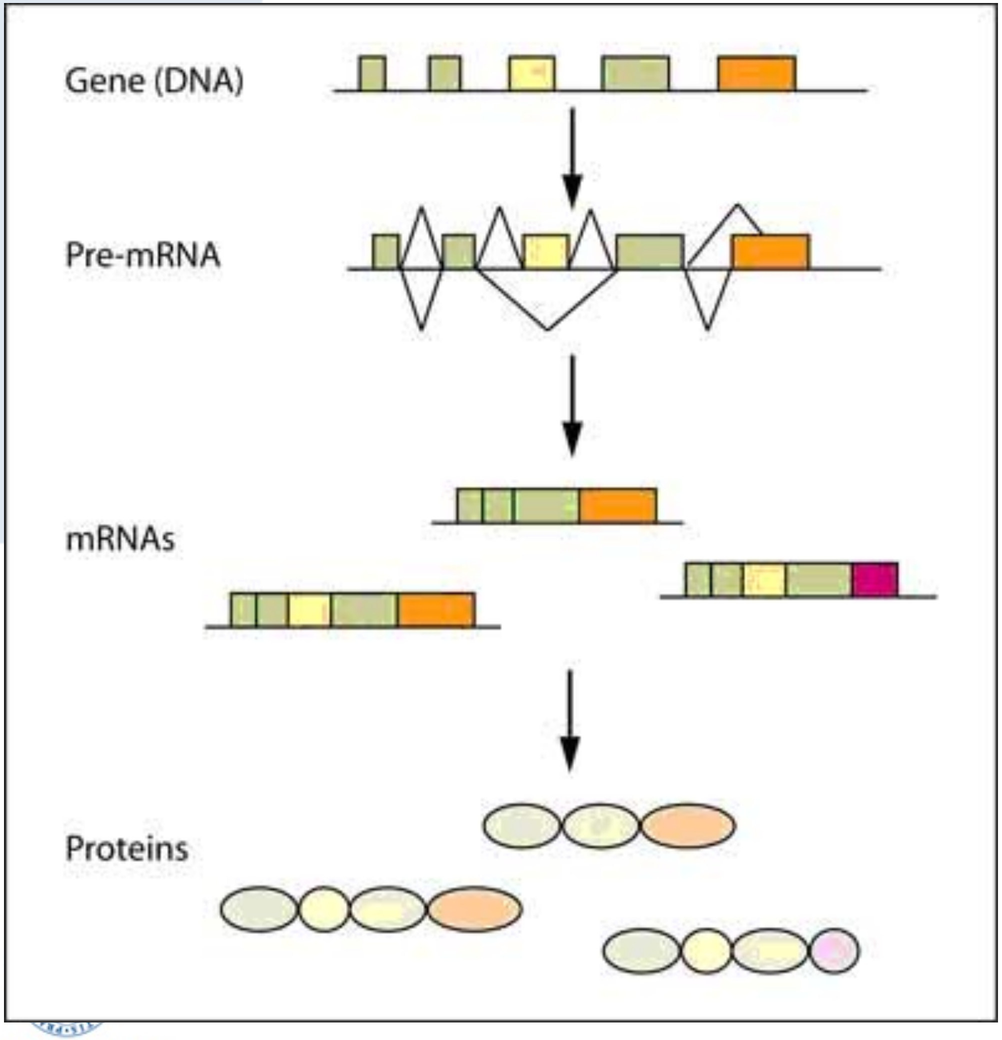
re-sequencing

- Sequencing a new individual of a species for which the reference genome is know (and (well) annotated).

- What for?
  - Genotyping (testing for known mutations).
  - Variations analysis: finding polymorphisms such as SNPs (Single Nucleotide Polymorphisms), CNVs (Copy Number Variations), and SVs (Structural Variants).

Universiteit Leiden

Alternative Splicing



- AS is when several mRNAs can be produced from a unique pre-mRNA

- E.g. in humans there are approximately 30,000 genes and it is estimated that 70% of human protein-coding genes undergo alternative splicing to generate up to 150,000-200,000 mRNAs and proteins through alternative splice site usage.

- In 2008, an experiment revealed that 34% of human transcripts were not from known genes [Science 321]

RNA-Seq

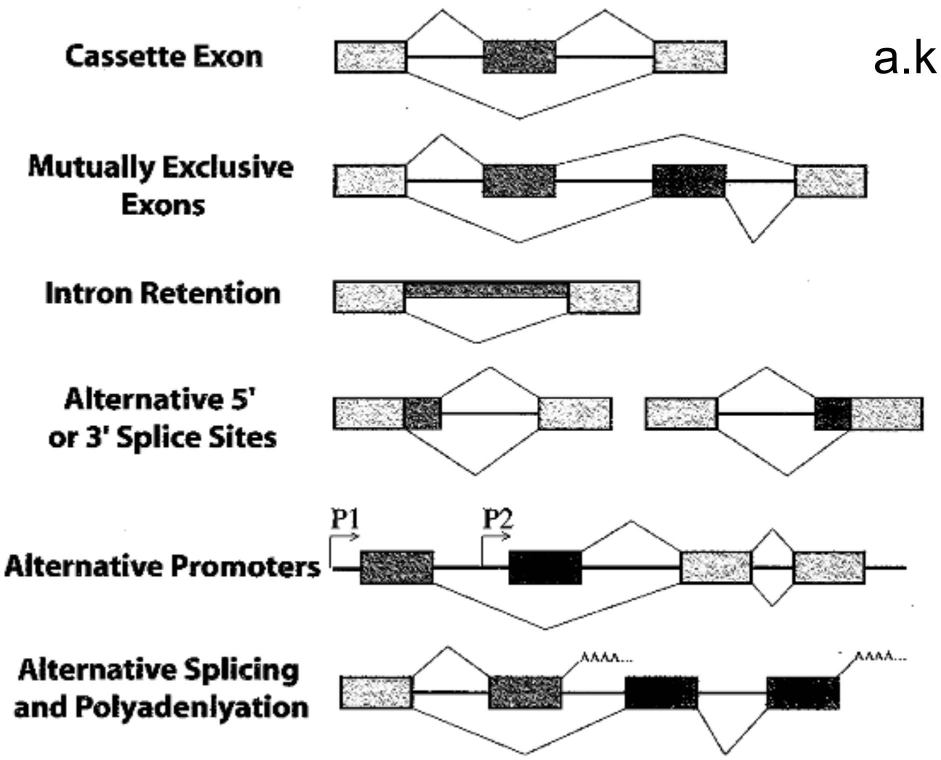
Among the goals of old good Human Genome Project there was the mapping and genotyping to associate DNA sequences to diseases (predispositions). This task kind of failed...

It is now very clear (and it was not at the time) that reading the genome is not enough: the very same DNA, in different situations (e.g. environmental conditions such as food, pollution and life style, or different tissues) expresses in different ways.

In particular, the same gene can express more or less, or with different (frequency of) alternative splicing events.

RNA-Seq is a NGS experiment that “takes a picture” of what genes are transcribed in that cell at that moment, and how, and how much.

Alternative Splicing events



a.k.a. *exon skipping*

a.k.a. *alternative donor/acceptor*

Universiteit Leiden

The bad side of NGS

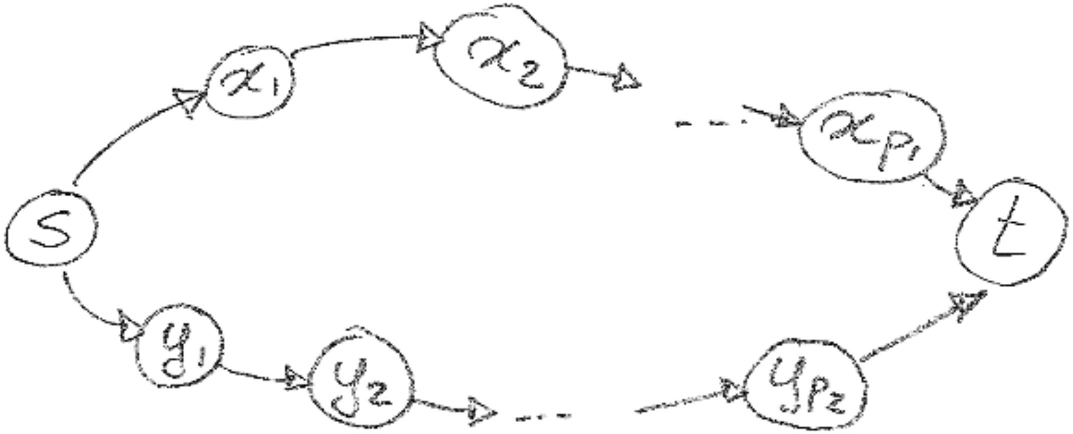
- Even shorter fragments (now called *reads*): from 1000 of Sanger technology to 25, then 50, then 75, now ~100 bases.
- Even more errors (when new size is released).

Fragment assembly is even harder !!

Bubbles in graphs

A bubble in a directed graph G=(V,E) is, for two fixed nodes s,t ϵ V (source and target), a pair of node disjoint paths from s to t.

$x_1 x_2 \cdots x_{p_1}$ is a path from s to t

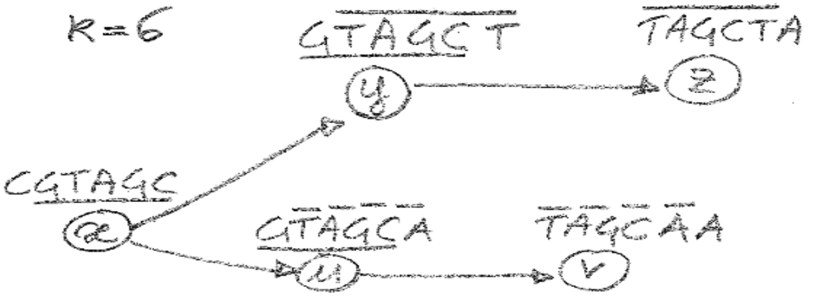


$y_1 y_2 \cdots y_{p_2}$ is a path from s to t

$\forall\, 1 \le i \le p_1 \ \forall\, 1 \le j \le p_2 \quad x_i, y_j \in V$ and $x_i \neq y_j$

The de Bruijn Graph

- The outcome of a NGS experiment is a huge set of *reads*.
- For space and time computational complexity reasons, the most used data structure to represent the outcome of a NGS experiment is the so (unproperly) called “de Bruijn Graph”.
- Fixed an integer k, there is a node per each k-mer (sequence of length k) that occurs in at least one read, and an edge (x,y) if the sequences x and y overlap in k-1 positions.
- Example:



reads: CGTAGCTA x-y-z

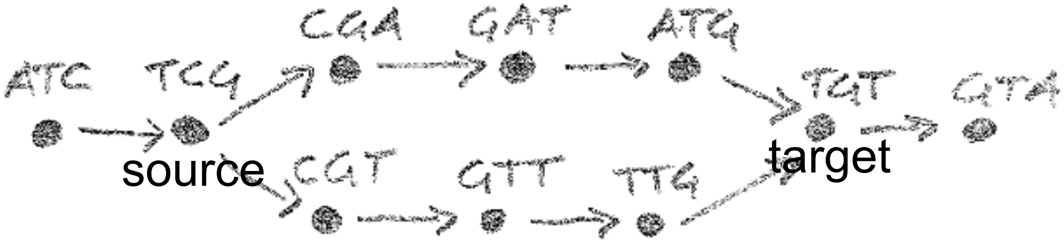
CGTAGCAA x-u-v

SNPs in de Bruijn Graph

- If a de Bruijn graph contains reads from two individuals (or alleles) that have a SNP:

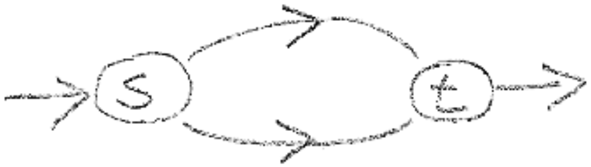
two sequences ATCGATGTA (a SNP)
ATCGTTGTA

k=3



source

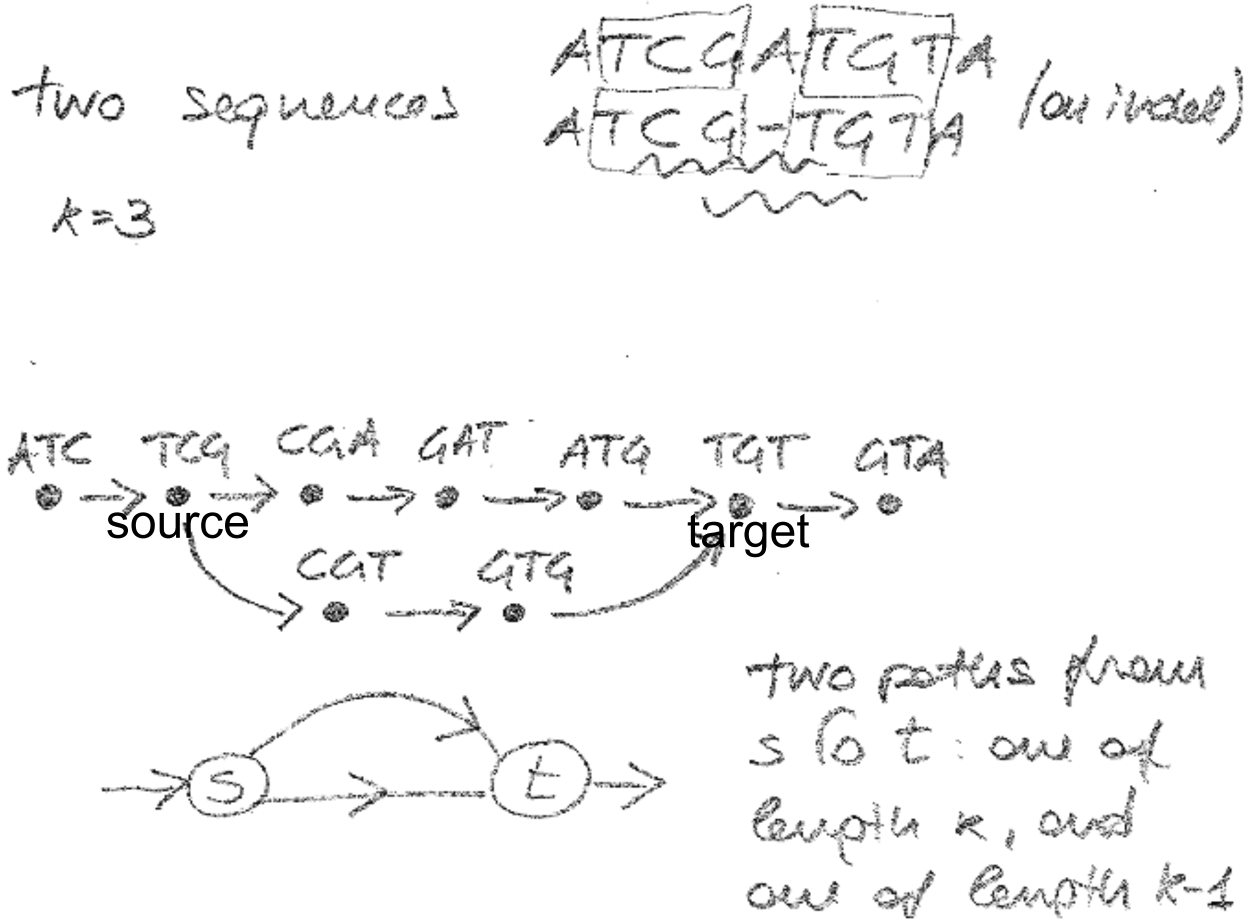
target



two paths of length k from s to t

Indels in de Bruijn Graph

- If a de Bruijn graph contains reads covering two fragments that differ for an <u>indel</u>:



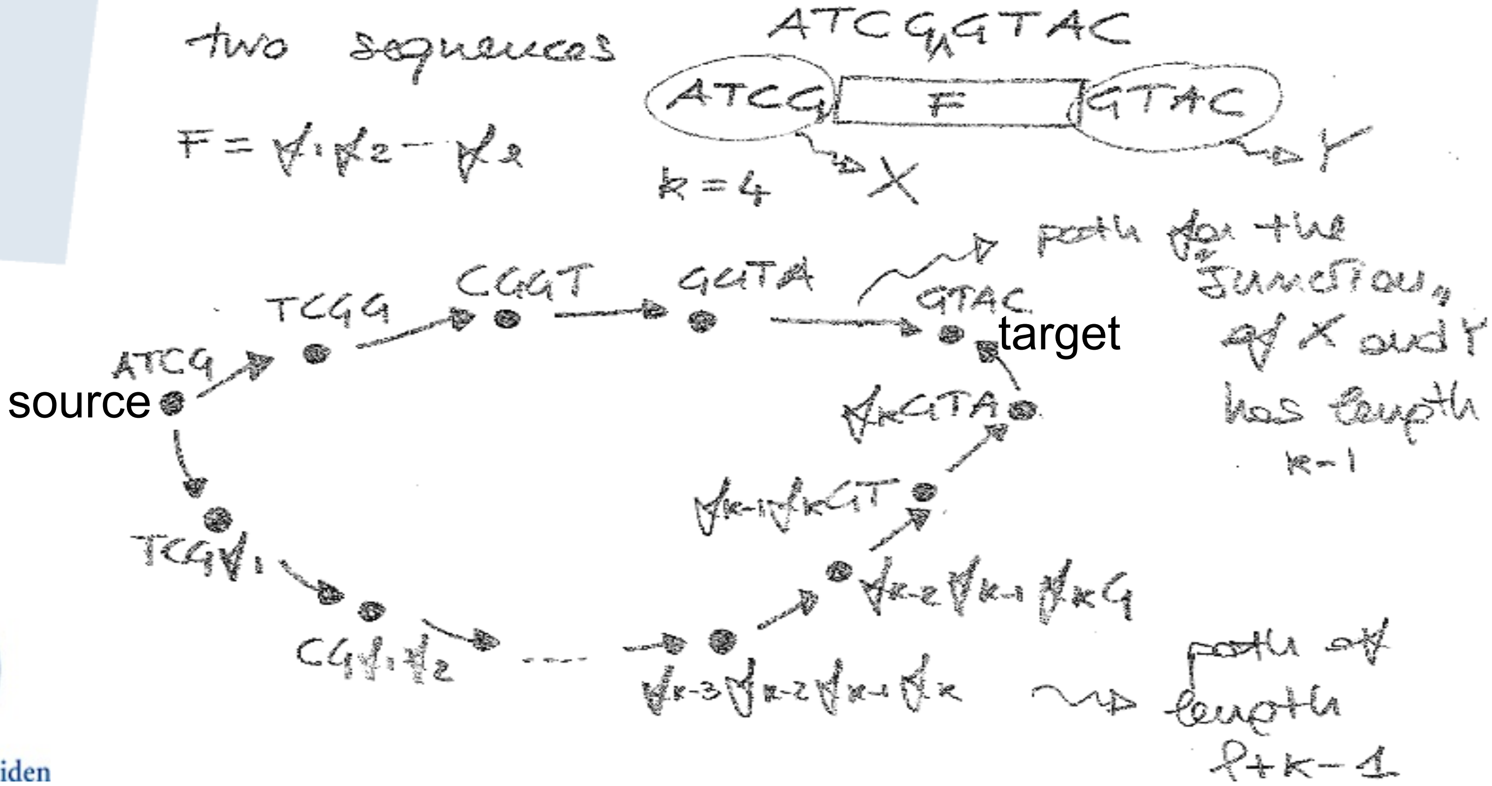
source

target

Indels of fragments in de Bruijn Graph

If a de Bruijn graph contains reads covering two fragments that differ for an insertion or a deletion of a whole segment:

Remark: this is the case of CNVs in DNA, and AS such as exon skipping and intron retention in RNA.

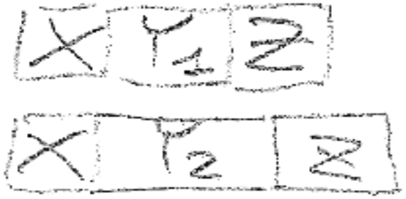

two sequences
ATCG$_F$GTAC
ATCG
F
GTAC
X
Y
$F = \alpha_1 \alpha_2 \dots \alpha_\ell$
k=4
path for the junction of X and Y has length k−1
source
ATCG
TCGG
CGGT
GGTA
GTAC
target
α_kGTA
$\alpha_{k-1}\alpha_k$GT
$\alpha_{k-2}\alpha_{k-1}\alpha_k$G
TCGα_1
CG$\alpha_1\alpha_2$
$\alpha_{k-3}\alpha_{k-2}\alpha_{k-1}\alpha_k$
path of α → length ℓ+k−1

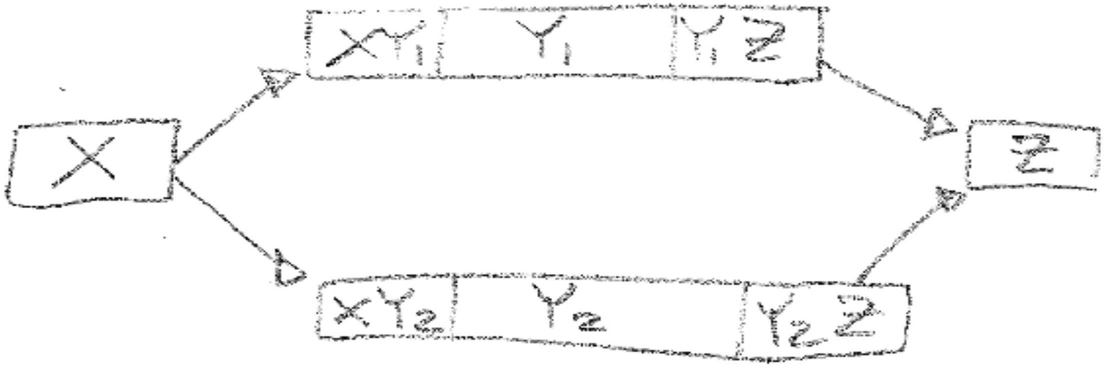
Substitution of fragments in de Bruijn Graph

If a de Bruijn graph contains reads covering two fragments that begin and end with the same fragment but contain two different segments inbetween.

Remark

two RNA isoforms





two different paths of any ($\geq k-1$) size from s to t

Detecting bubbles in de Bruijn Graph

Fragment Assembly is particularly hard for NGS data.

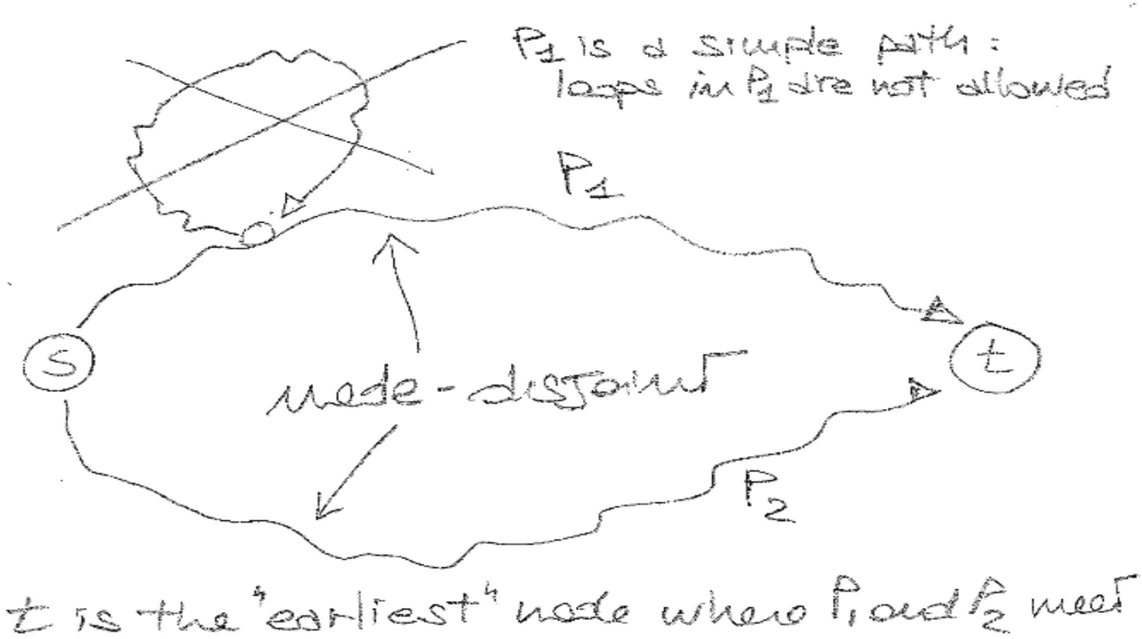
Mapping the reads to a reference genome (when available) is also a computationally heavy task.

Finding bubbles in the de Bruijn graph can help the detection of biologically interesting features without the need of assemblying nor mapping.

An algorithm to detect bubbles in **any** directed graph

The computational problem: Given a directed graph G=(V,E), we seek for all *s* ϵ V, all the bubbles from *s* to any *t* ϵ V.

Recall: for s,t ϵ V, a bubble is a pair of node-disjoint paths P_1 and P_2 from s to t.



P_1 is a simple path: loops in P_1 are not allowed

t is the "earliest" node where P_1 and P_2 meet

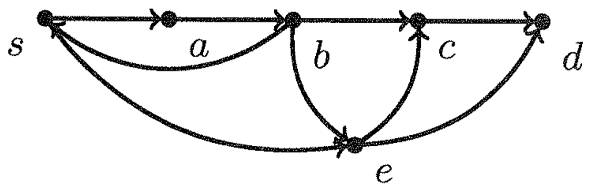
Observation: if you swap the direction of all the edges of P_1 (or P_2), then the bubble turns into a cycle.

The idea

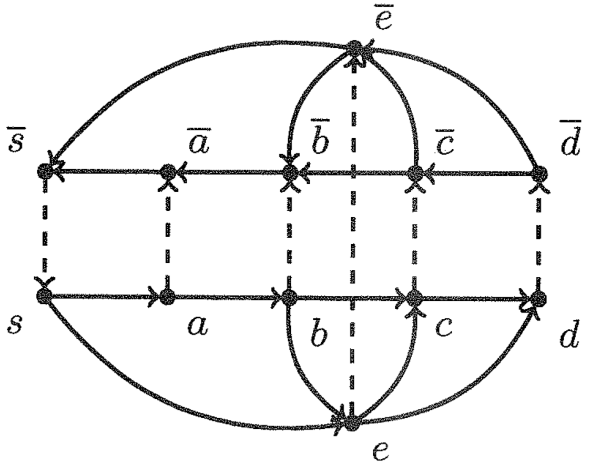
Given G=(V,E) and s ϵ V, we build a graph G'_s as follows:

- G'_s is made of the original G plus a copy G'_s=(V,E) of it.
- In G'_s all edges directions are swapped.
- Any edge entering s is deleted.
- For each node x ϵ V different from s, we add an edge (x,x) where x is the copy of x in G'_s.
- We add the edge (s',s): *must edge*

certain types of cycles in G'_s correspond to bubbles of G



(a) Graph G

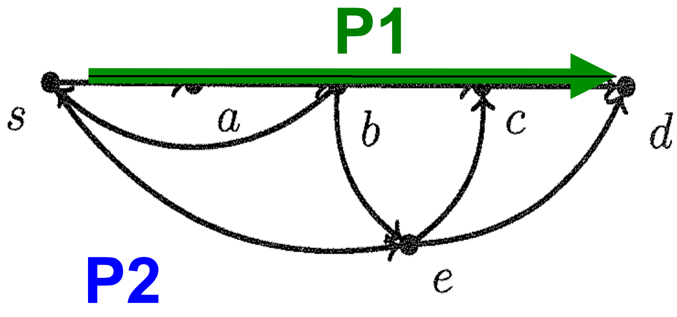


(b) Graph G'_s

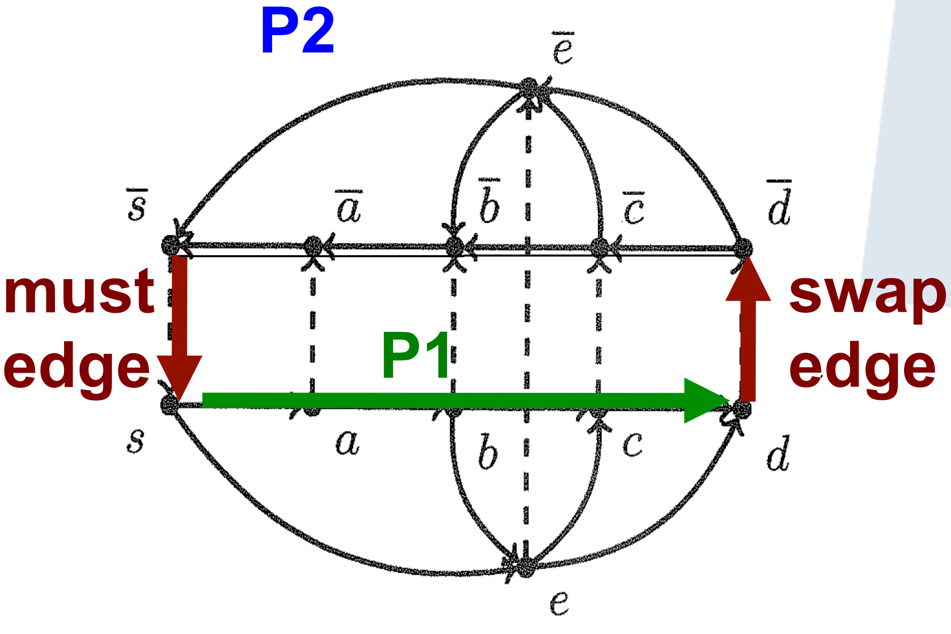
Algorithm BubbleCycle

For each *s* ϵ V builds G'_s and there seeks cycles from *s* such that:

- no twin edges are used, except for s and another one: the *swap edge*
- the swap edge is *t*, the target of the bubble.
- a certain nodes numbering order is preserved.



(a) Graph G



(b) Graph G'_s

Testing BubbleCycle (BC)

No specific algorithm existed for bubbles detection. What to compare with?

KISSPLICE* is a tool for AS events detection whose pipeline includes a step that detects a special type of bubbles (bubbles whose shortest path can't be longer than k-1: it finds intron retention, exon skipping, and alternative splice site events only).

We adapted the BubbleCycle to find this special type of bubbles and replaced with this the bubble finding step of KISSPLICE: name KBC the new pipeline.

We compared the performances of KISSPLICE (KS) with those of the new tool KBC

Although the comparison in unfair for BubbleCycle, KBC outperformed KS !!!

* Sacomoto et al. *KISSPLICE: de-novo calling alternative splicing events from RNA-seq data*. BMC Bioinformatics 13(6):55, 2012.

Experiments on human tissues

Dataset: Human Body Map 2.0 data (ERP00546). Two tissues: liver and brain. Size 32M and 29M. Read length 75 bp.

Choosing k=25, KBC and KS found the same results, but KBC was faster: 35'10” rather than 54'01”.

KISSPLICE could not manage the search of bubble of unlimited size (that is, releasing the constraint of shortest path of size k-1), and thus could not detect mutually exclusive exons.

We run KBC without the constraint and the computation ended successful: the analysis of the results is work in progress...

References and Acknowledgements

This talk includes the description of a some joint work with:

E.Birmele (Univ. of Evry), P.Crescenzi (Univ. of Florence), R.Ferreira (Microsoft Research), R.Grossi (Univ. of Pisa), V.Lacroix (Univ. of Lyon), A.Marino (Univ. of Florence), G.Sacomoto (INRIA Rhone Alpes), M.-F.Sagot (INRIA Rhone Alpes).

It also describes results obtained by my student Paolo Cois in his master thesis to be discusses in October 2012 in Pisa.

REFERENCES

E.Birmele, P.Crescenzi, R.Ferreira, R.Grossi, V.Lacroix, A.Marino, N.Pisanti, G.Sacomoto and M.-F.Sagot *Efficient bubble enumeration in directed graphs*, proceedings of String Processing and Information REtrieval (SPIRE) 2012, Springer LNCS 7608, 2012.