RNA-seq: analysis and de novo assembly

Yakovlev Pavel

St-Petersburg Academic University
Supervisor: Segrey Nurk

April 6, 2013

What?

We want to know everything about RNA high throughput sequencing.

- How RNA is sequenced?
- What are the features and differences from DNA sequencing data?
- How to assembly this data?
- How to validate assembled data?

Why?

Assemble and analyze RNA is very cool, because of:

- we get only what we interested in genes (mRNA) and ncRNA,
- we can get useful information about ogranism without full-genome sequencing,
- **Everybody can do it, but we cannot yet!**
- everybody do it awfully.

Theory

In RNA-seq:

- cDNA is sequenced, so we can get paired reads
- eukaryotic mRNA has cap and poly-A tail, which help to identify start and end
- prokaryotic mRNA do not have alternative splicing, so it is easy to assemble them (**FALSE**)
- all the headache starts with alternative splicing...
  - different isoforms have different expression
  - different isoforms have same exons
  - and more, than you can imagine

What have we done?

✔ Read dozens of papers.

✔ Understood the process of RNA sequencing.

✔ Created the table of (dis)advantages of all popular assemblers.

✔ Understood most of RNA De Bruijn graph reducing algorithms.

✖ SPAdes 2.4.0 fails on prokaryotic data ☹.

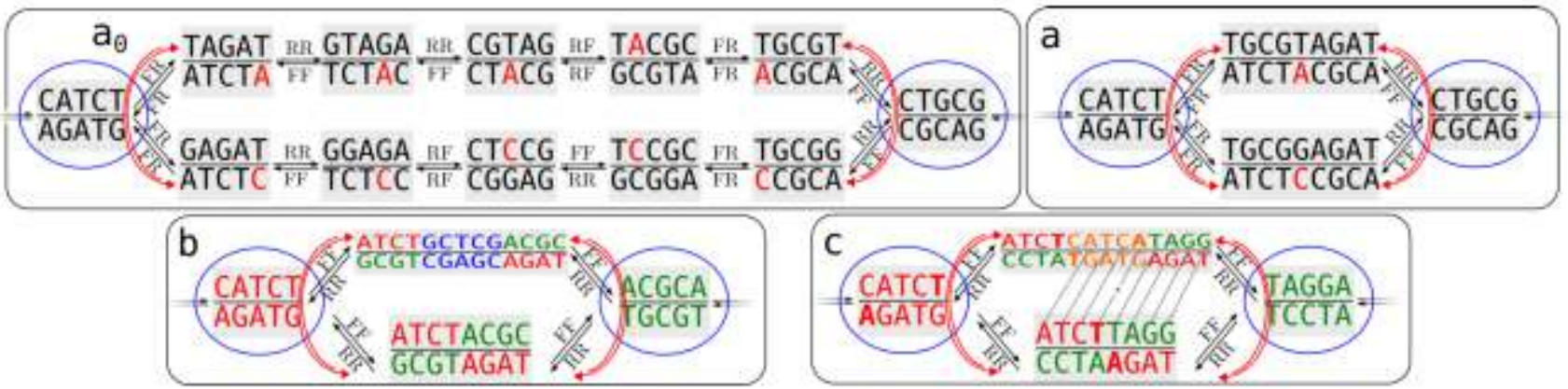
⊶ Alternative splicing reducing algorithms are in development.

What's the problem with SPAdes?

To assemble 25 Gb paired reads of E. Coli's mRNA SPAdes took:

- 26 hours of work on ACE server
- 144 Gb RAM peak load
- Average contig length is $(k + 1)$

After 26 hours of work SPAdes failed with std::bad_alloc exception!



Let‘s do it!