Dottorato di Ricerca in Informatica - Ciclo XXV
Dipartimento di Informatica, Sistemistica e Comunicazione
Facoltà di Scienze Matematiche, Fisiche e Naturali
Università degli Studi di Milano - Bicocca

Algorithms for detecting variations from Next-Generation Sequencing Data

Presentazione Dottorato

11 Ottobre 2011

Candidato:	Stefano Beretta
Supervisor:	Prof.ssa Paola Bonizzoni
Tutor:	Prof.ssa Lucia Pomello

Outline

1. Motivations
2. State of the Art & Ongoing Works
3. Conclusions

Motivations

- Revolution in genome sequencing and analysis: from traditional methods to NGS (Next-Generation Sequencing)[1] [2] [3]
- Need to develop novel computational frameworks to analyze NGS data
- Goal: design algorithms to analyze NGS data for detecting sequence variations

[1] Venter, J.Craig:*Multiple personal genomes await.* Nature, (2010)

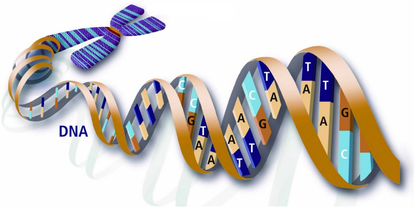
[2] Mardis, E.R.: *The impact of next-generation sequencing technology on genetics.* Trends in genetics, (2008)

[3] Metzker, M.L.: *Sequencing technologies - the next generation.* Nature reviews. Genetics, (2010)

Genome Sequencing

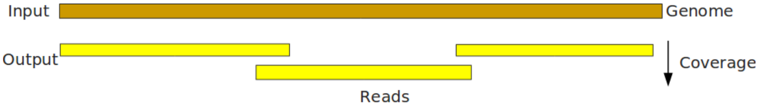
Determination of the primary structure of a molecule
DNA/RNA → sequence of nucleotides

- Traditional Methods (Sanger, 1977)
- Next-Generation Sequencing Methods (2005)

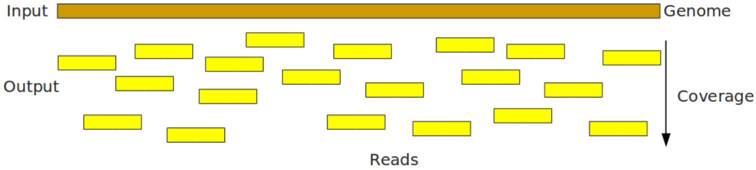


Sanger Vs. Next-Generation Sequencing

- **Sanger (1977)**



- **NGS (2005)**



Challenges

- Algorithmic Challenges:
  - More than 10^9 short sequences ⇒ Linear time algorithms
  - Need for data compression / succint data structures
  - New computational model and data structure for pattern matching and indexing of NGS reads
    (es. hashing, Burrows-Wheeler transf., suffix array)[4] [5] [6]

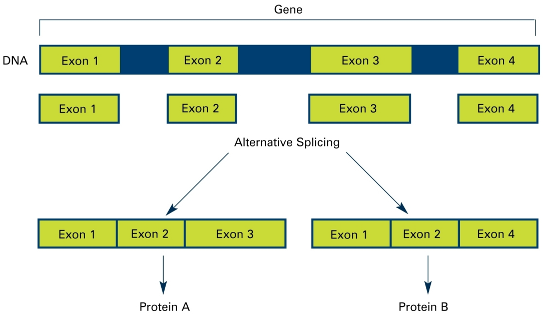
[4] Langmead B., et al.: *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biology (2009)

[5] Dalca, V.A., Brudno, M.:*Genome variation discovery with high-throughput sequencing data.* Briefings in Bioinformatics (2010)

[6] Li H., Homer N.:*A survey of sequence alignment algorithms for next-generation sequencing.* Briefings in Bioinformatics (2010)

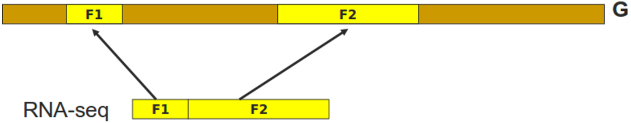
Computational Problem

- Characterization of variations (i.e alternative splicing events) among different transcript sequences (sequenced by NGS) of the same gene.
    - Human genes undergo AS (alternative splicing)



Characterization of Alternative Splicing Events

- Limits of Existing Methods:
  - No techniques based on short reads comparison
  - No characterization of differences of transcripts
  - Developed algorithms map the NGS data into the given reference genome to infer splice junctions[9] [10]



[9] Bryant, et. al., Bioinformatics (2010) *Supersplatspliced RNA-seq alignment*

[10] Trapnell, et. al., Bioinformatics (2009) *TopHat: discovering splice junctions with RNA-Seq*

Characterization of Alternative Splicing Events

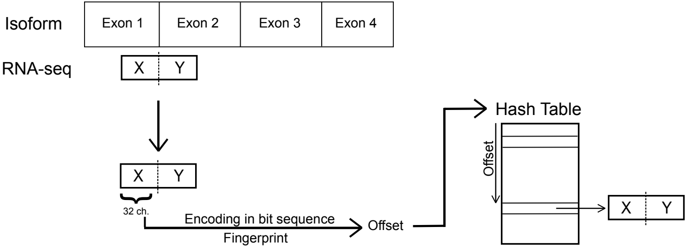
- Problem 1: inference of alternative splicing (AS) events
  - **Input**: a set of short reads from transcripts of a gene
  - **Goal**: graph representation of AS events (genome scale)
- Previous approaches:
  - Detect splice junctions
  - Validate transcripts
- Our Approach:
  - Detect differences (which are a few) and discard similarities (too many)
  - No alignment to the reference genome

Characterization of Alternative Splicing Events

- Problem 2: detecting gene structure and AS events
  - **Input**: a set of short reads from transcripts of a gene and Refseq data
  - **Goal**: complete gene structure
- Novel approach:
  - Compare different NGS experiments
  - Quite powerful in detecting gene structure
  - Efficient annotation of short reads

Characterization of Alternative Splicing Events

- Algorithmic Solution
  - We index short reads with a hash table in order to:
    1. *De Novo Assembly* of short reads to compose Exons
    2. Identify junction points of Exons



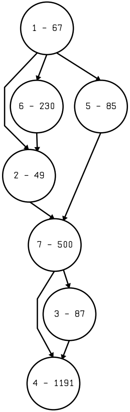
Characterization of Alternative Splicing Events

- Input: Short Reads

```
>ARHGAP4@U52112.4-002 311AFAAXX_HWI-EAS229_90:8:1:1405:899/1
GGAGCAGTCCCGAGGGCCTCCTGGCATCGGAGCTGGTCCACCGGCCAGAGCCATG
>ARHGAP4@U52112.4-022 311AFAAXX_HWI-EAS229_90:8:1:1699:670/1
AACTGATGGACCAGGCCTCTCGAGCCATGATAGAGAACTTCAATGCCAAATATGT
>ARHGAP4@U52112.4-012 311AFAAXX_HWI-EAS229_90:8:3:1740:1221/1
CCAGACCAGCCCCTCCACCGAGTCCCTCAAGTCCACCAGCTCAGACCCAGGCAGC
>ARHGAP4@U52112.4-001 311AFAAXX_HWI-EAS229_90:8:4:1458:1519/1
GCCCCGAAGCCCAAAGGCCCCGCCCAGCAGCCGCCTGGGCAGGAACAAAGGCTTC
>ARHGAP4@U52112.4-005 311AFAAXX_HWI-EAS229_90:8:4:1149:370/1
CCGGAGGCGCGGCCAGCAGCAGGAGACCGAAACCTTCTACCTCACGAAGCTCCAG
...
```

- Results Summary
  - Composed chains (graph nodes): 7
  - Linked chains (graph arcs): 9
  - Sensitivity (nodes): 1
  - Positive Predictive Value (nodes): 1
  - Sensitivity (arcs): 1
  - Positive Predictive Value (arcs): 1

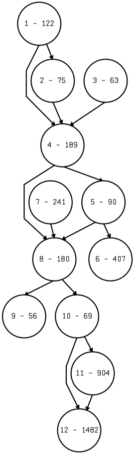
- Output: AS Graph



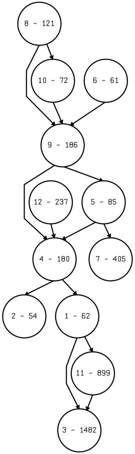
Example (with correct prediction)

- Results Summary
  - Pr. nodes: 12
  - Or. nodes: 12
  - Pr. arcs: 14
  - Or. arcs: 14
  - S_n nodes: 1
  - PPV nodes: 1
  - S_n arcs: 1
  - PPV arcs: 1

- Original Graph



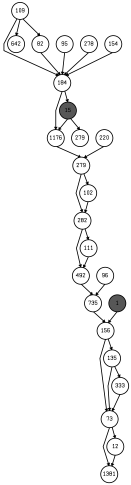
- Predicted Graph



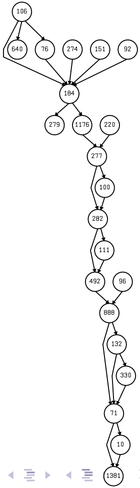
Example (with no correct prediction)

- Results Summary
  - Pr. nodes: 22
  - Or. nodes: 25
  - Pr. arcs: 27
  - Or. arcs: 31
  - S_n nodes: 0.84
  - *PPV* nodes: 0.95
  - S_n arcs: 0.71
  - *PPV* arcs: 0.81

- Original Graph



- Predicted Graph



Characterization of Alternative Splicing Events

- Results on Problem 1[*][†]
  - Efficient algorithm (linear in the number of reads)
  - Accuracy prediction on simulated data
  - Robustness to typical error scenario
  - Extremely scalable approach
- Results on Problem 2
  - Fast annotation of short reads from different NGS experiments
  - Fast detection of the gene structure with hashing and clustering techniques

[*] *Alternative Splicing from RNA-seq Data without the Genome.*, 8th Special Interest Group meeting on Alternative Splicing (AS-SIG), 2011, Vienna

[†] *Identification of Alternative Splicing variants from RNA-seq Data.*, Next Generation Sequencing Workshop, 2011, Bari

Characterization of Alternative Splicing Events

- Thesis Structure and Ongoing Works
  - Manage short reads data from different technologies (length increasing): main algorithmic issues
  - General problem: gene structure prediction via reads with or without Refseq and algorithmic inference of AS events from the produced graphs
  - Experimental work: testing our approach at genome-wide scale on real data (human , mouse,...) *

* *Body Map 2.0 (Illumina HiSeq) http:www.broadinstitute.org igvdataBodyMaphg19IlluminaHiSeq2000_BodySites*

Conclusions

- Linear time algorithm for NGS data analysis
- Efficient data structure for short reads
- Characterization of variations in AS events
- Experimental validation on simulated data