



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

30 Settembre 2010

Candidato: Stefano Beretta

Tutor: Prof.ssa Lucia Pomello

Outline

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

Outline

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

Motivations

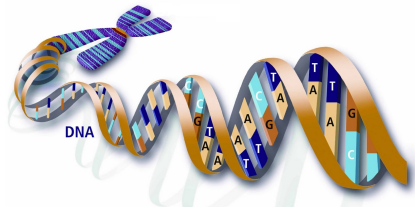
- Revolution in genome sequencing and analysis: from traditional methods to NGS (Next-Generation Sequencing)¹
- Need to develop novel computational frameworks to analyze NGS data
- **Goal:** design algorithms to analyze NGS data for detecting sequence **variations**

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

Genome Sequencing

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

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

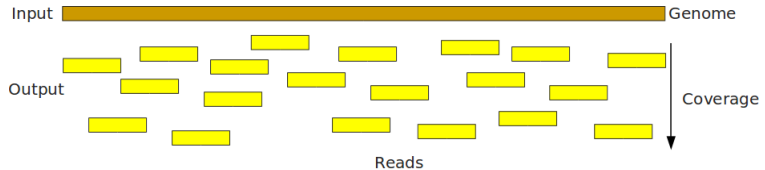


Sanger Vs. Next-Generation Sequencing

- Sanger (1977)



- NGS (2005)



Sanger Vs. Next-Generation Sequencing

Sanger (1977)

- 1 Long Reads (~ 1000 bp)
- 2 Low Throughput
($\sim 10^6$ bp/day)
- 3 Low Coverage ($\sim 1x$)
- 4 Expensive (10^3 bp/\$)

NGS (2005)

- 1 Short Reads (25-300 bp)
- 2 High Throughput
($\sim 10^9$ bp/day)
- 3 High Coverage ($> 10x$)
- 4 Low Costs ($> 10^5$ bp/\$)

Sanger Vs. Next-Generation Sequencing

Sanger (1977)

- ① Long Reads (~ 1000 bp)
- ② Low Throughput ($\sim 10^6$ bp/day)
- ③ Low Coverage ($\sim 1x$)
- ④ Expensive (10^3 bp/\$)

NGS (2005)

- ① Short Reads (25-300 bp)
- ② High Throughput ($\sim 10^9$ bp/day)
- ③ High Coverage ($> 10x$)
- ④ Low Costs ($> 10^5$ bp/\$)



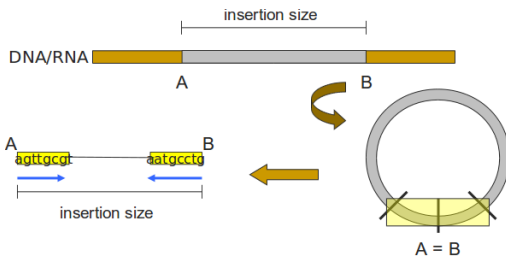
~~Traditional Algorithms, Models and Tools~~

NGS Data

- Short Read



- Paired-Ends



Biological Problems

- 1 Compare a known genome reference with an unknown genome (but sequenced by NGS)
- 2 Infer transcripts data using short reads sampled by NGS



Detect Variations

Algorithmic Challenges

- More than 10^9 short sequences \Rightarrow Linear time algorithms
- Need for data compression / succinct data structures
- New computational model and data structure for pattern matching
(es. hashing, Burrows-Wheeler transf., suffix array)^{2 3 4}

²Langmead B, et al., Genome Biology (2009), *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*

³Dalca V.A. and Brudno M., Briefings in Bioinformatics (2010), *Genome variation discovery with high-throughput sequencing data*

⁴Li H. and Homer N., Briefings in Bioinformatics (2010), *A survey of sequence alignment algorithms for next-generation sequencing*


Outline

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

Computational Problems

- ① Identification of differences (Structural Variations) between a known genome (*reference*) and an unknown genome sequenced by NGS (*donor*).
- Biological Motivations:
 - SV are **common** in human individuals and are related to **diversity** and **disease susceptibility** ^{5 6}
 - Detecting SV is crucial in medical and biological studies of several diseases

⁵ Korbelt, et al., Science (2007), *Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome*

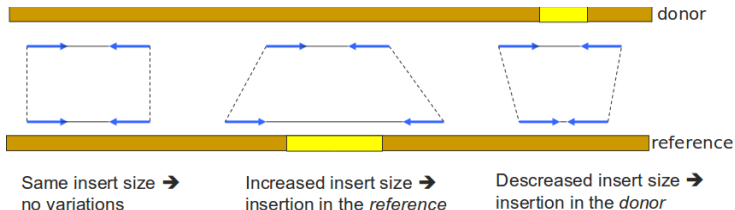
⁶ Tuzun, et al., Nature Genetics (2005), *Fine-scale structural variation of the human genome* 

Computational Problems

- ② Characterization of variations (i.e alternative splicing events) among different transcripts sequences (sequenced by NGS) of the same gene.
- Biological Motivations:
 - Human genes undergo AS (alternative splicing)
 - AS is the key process in determining **transcriptomes diversity**

Identification of Structural Variations (SVs)

- Structural Variations (SVs)
 - Insertions
 - Deletions
 - Inversions (>5 Kb)



Identification of Structural Variations (SVs)

- Structural variation discovery using maximum parsimony is NP-hard ⁷
- Actual tools consider only one alignment (for each short read) and discard all the others ⁸
- Probabilistic frameworks have been designed for the identification of specific SVs

⁷ Hormozdiari F., Alkan C., Eichler E., Sahinalp C., Genome research (2009), *Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes*

⁸ Medvedev P., Stanciu M., Brudno M., Nature methods (2009) *Computational methods for discovering structural variation with next-generation sequencing*

Identification of Structural Variations (SVs)

- **Problem:** predicting structural variations
 - **Input:** a set S of paired-ends (PEs) from a donor genome D and a reference genome R .
 - **Goal:** compute the set of structural variations that explains how D differs from R
- **Previous approaches:**
 - consider each SV separately
 - adopt probability based formulation
- **Our Approach:**
 - design a specific tool for PEs
 - develop an integrated approach for all SVs

Identification of Structural Variations (SVs)

• Algorithmic Solution

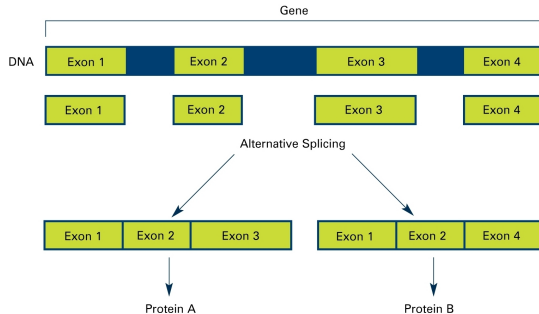
- PEs are aligned to the reference genome R (> 1 locations and > 1 orientations) and clustered into
 - Concordant mapped PEs \Rightarrow Donor = Reference
 - Discordant mapped PEs \Rightarrow Structural Variations
- Discordant PEs are analyzed to detect
 - Insertion / Deletion, Inversion, Other complex cases

• Issues

- Large SVs are hard to detect
- Detecting combinations of different SVs

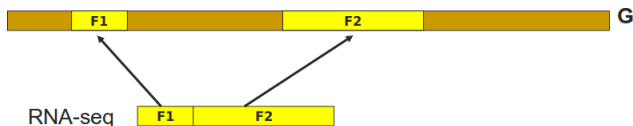
Characterization of Alternative Splicing Events

- Alternative Splicing



Characterization of Alternative Splicing Events

- 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) *Supersplated RNA-seq alignment*

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

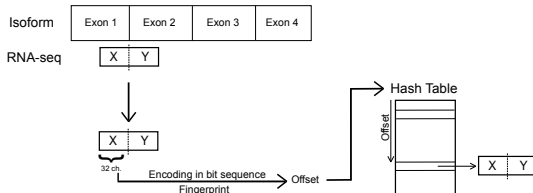
Characterization of Alternative Splicing Events

- **Problem:** 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 few) and discard similarities (too many)
 - no alignment to the reference genome

Characterization of Alternative Splicing Events

Algorithmic Solution

- We index short reads with a hash table in order to:
 - assembly chains of short reads to compose Exons
 - identify junction points of Exons



Issues

- Not unique identification of splicing junction
- Some AS events are hard to characterize
- Needs for a topological sort of Exons

Outline

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

Conferences and Journal Plan

● Conferences

- Recomb (Research in Computational and Molecular Biology)
- WABI (Workshop on Algorithms in Bioinformatics)
- ISMB (Intelligent Systems in Molecular Biology)
- ECCB (European Conferences on Computational Biology)

● Journals

- Journal of Computational Biology
- BMC Bioinformatics
- Bioinformatics
- Journal of Bioinformatics and Computational Biology
- Genome Research
- Theoretical Computer Science
- ACM / IEEE Transaction on Computational Biology and Bioinformatics

Courses

- Soft Computing Techniques for Software Engineering*
(Prof. F. Arcelli)
- Algorithmic Game Theory
(Prof. N. Gatti) - Politecnico di Milano
- Modelli di Calcolo e Complessità Computazionale
(Prof. G. Mauri) - Autunno 2010

*Esame sostenuto

Courses

- English
(Prof. J. Weekes)
- Laboratorio di Comunicazione
(Prof. C. Torelli)
- Scuola di Dottorato di Scienze M.M. F.F. N.N.
 - Gestione dei Progetti di Ricerca
(Prof. G. Barozzi)
 - Elementi di Organizzazione Aziendale
(Prof. Fedeli - De Vita)

Conclusions

