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 Next-Generation Sequencing Data Analysis

Presentazione Dottorato

4 Ottobre 2012

Candidato:	Stefano Beretta
Advisors:	Prof.ssa Paola Bonizzoni
	Prof. Gianluca Della Vedova
Tutor:	Prof.ssa Lucia Pomello

Outline

1. Motivations

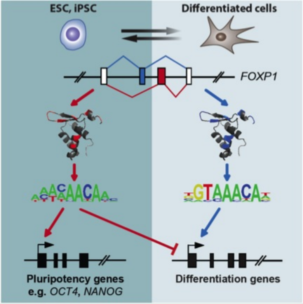
2. Characterization of Alternative Splicing Events

3. Taxonomic Assignment in Metagenomics

Motivations

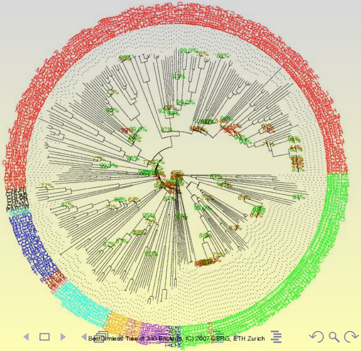
- Context
  - revolution in genome sequencing and analysis: from traditional methods to NGS (Next-Generation Sequencing)
- Computational methods in
  - Transcriptomics
  - Metagenomics
- Objective
  - Transcriptomics: alternative splicing events prediction and transcriptomic analysis from NGS data
  - Metagenomics: classification of organisms of a metagenomic sample from NGS data

Motivations



"In recent years, major strides have been made in the understanding of the and non-coding RNAs in the control of stem cell pluripotency and reprogramming of somatic cells to induced pluripotent stem cells (iPSCs). However, the role of alternative splicing in these processes is largely unknown." [Cell, 2011]

"Assessment of the microbial diversity residing in arthropod vectors of medical importance is crucial for monitoring endemic infections, for surveillance of newly emerging zoonotic pathogens, and for unraveling the associated bacteria within its host." [PLoS One. 2011]



BestDimless Tree of 340 Bacteria, (C) 2007 CBRG, ETH Zurich

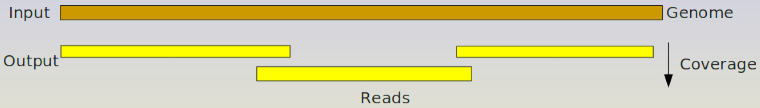
Contributions

- Contributions in Transcriptomics
  - Novel approach to alternative splicing prediction from RNA-Seq data based on splicing graph without a reference sequence
  - Efficient method for building splicing graph (from millions of sequences)
  - Characterization of conditions for reconstructing splicing graphs without a reference sequence
- Contributions in Metagenomics
  - New optimal algorithm for the penalty score calculation based on skeleton tree
  - Improved procedure for the read taxonomic assignments
  - Flexible solution on multiple taxonomies

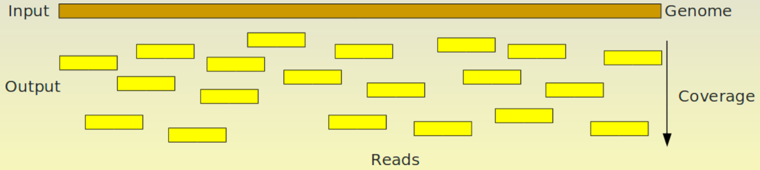
Genome Sequencing

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

- Traditional Methods (Sanger, 1977)

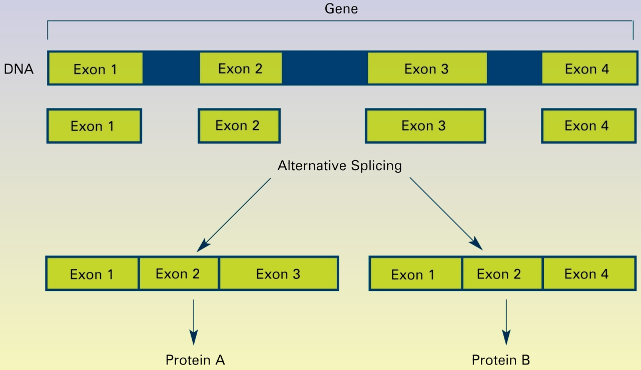


- Next-Generation Sequencing Methods (2005)



Characterization of Alternative Splicing Events

"Alternative splicing has a crucial role in the generation of biological complexity, and its misregulation is often involved in human disease." *Deciphering the splicing code. [Nature, 2010]*



Characterization of Alternative Splicing Events

Core Problem in Transcriptomics

Characterization of Alternative Splicing (AS) variations from RNA-Seq data

- Solution
  - Building a splicing graph that explains all AS events derived from transcripts
    - without explicit transcript assembly
    - without a reference sequence

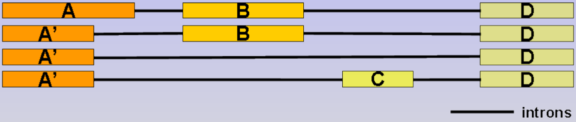
Characterization of Alternative Splicing Events

Detecting Alternative Splicing (AS) variations
from RNA-Seq data without a reference genome

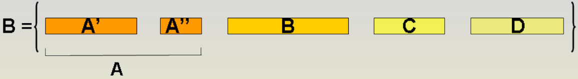
- Limits of Existing Methods:
  - No specific tools for large-scale inference of AS variations among gene transcripts
  - Reference genome is not always available
  - No characterization of differences of transcripts

Characterization of Alternative Splicing Events

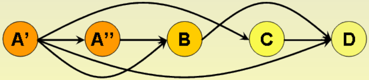
- Gene isoforms



- Set of blocks



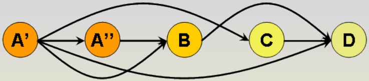
- **Splicing graph**



Characterization of Alternative Splicing Events

Input: set of RNA-Seq reads from unknown gene transcripts
Output: Splicing Graph

$\Longrightarrow$

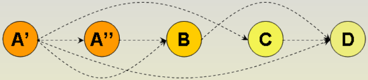


- Question: how to build efficiently the splicing graph, without using a reference genome?

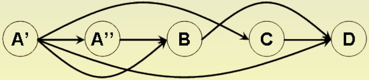
Characterization of Alternative Splicing Events

Solution

- Hashing input reads
  - Input set partitioning → Unspliced/Spliced
  - Constant time access to RNA-Seq reads
- Assembling *unspliced* reads into blocks (graph nodes)

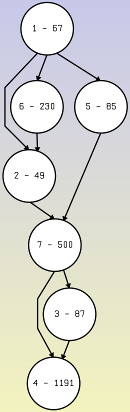


- Linking blocks with *spliced* reads (graph edges)



Characterization of Alternative Splicing Events

- Output: AS Graph



- 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
. . .
```

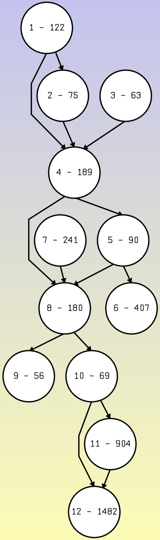
Characterization of Alternative Splicing Events

Experimental validation

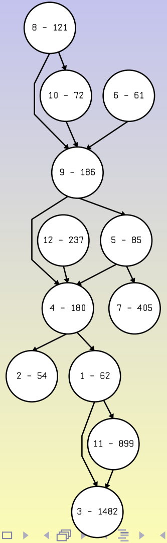
- Data
  - Simulated: 112 genes used as training set in EGASP (separated and mixed reads)
  - Real: RNA-seq data from ENCODE/Caltech
- Analysis
  - Graph mapping (nodes and arcs)
  - Accuracy (S_n and S_p)

Example

- Original Graph

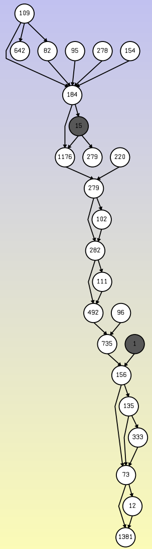


- Predicted Graph

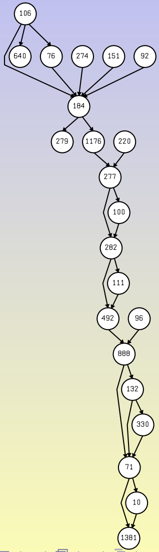


Complex Gene

- Original Graph



- Predicted Graph



Characterization of Alternative Splicing Events

- Results
  - Characterization of necessary and sufficient conditions for reconstructing splicing graph without a reference sequence
  - Efficient algorithm (linear in the number of reads) for splicing graph building
  - Accuracy and stability prediction on simulated data
  - Extremely scalable approach

Conclusions

- Characterization of Alternative Splicing Events
  - Characterization of AS variations from RNA-Seq data without a reference genome
  - Linear time algorithm and engineered a practical implementation for splicing graph construction
  - Experimental validation
- Developments
  - Use of splicing graph for comparing RNA-Seq experiments to detect relevant AS events (graph comparison)
  - Explore the possibility of not using a reference sequence (compare unknown transcripts)
  - Refine transcript predictions

Taxonomic Assignment in Metagenomics

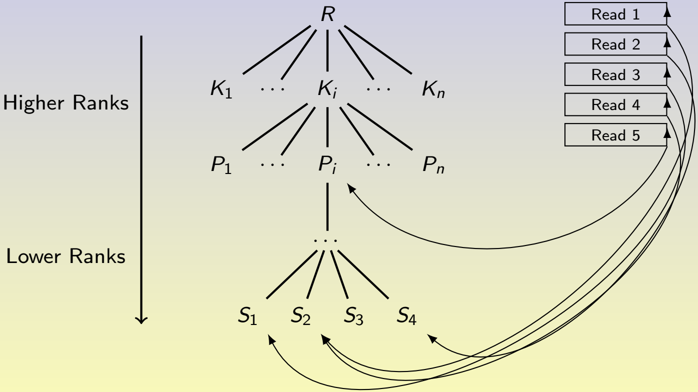
Core Problem in Metagenomics

Determine and quantify the species composition of a sample containing material of different (and possibly unknown) micro-organisms

- Culture-independent genetic studies of complex microbial communities
- Capture the diversity of microbial ecosystems
- Exploration and understanding of multi-organism interaction

Taxonomic Assignment in Metagenomics

- Problem: Taxonomic Assignment
  - **Input:** a set of reads (coming from the analyzed sample) and a reference taxonomic tree
  - **Output:** assignment of the reads to the tree

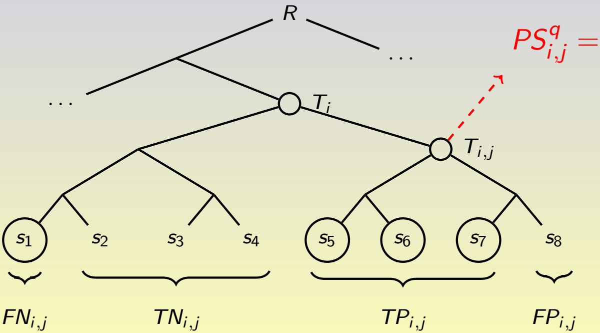


Taxonomic Assignment in Metagenomics

- Taxonomic Tree
  - Leaves are species (with known genomic sequence)
  - Inner node are ancestors (with NO known sequence)
  - All nodes are labeled
- Read Alignment (to the leaves)
  - Unique mapping ($|M| = 1$): unambiguous reads
  - Not unique mapping ($|M| > 1$): **ambiguous reads**
- Assignment (to leaves or internal nodes)
  - Assign reads to the correct node of the taxonomic tree:
    - For unambiguous read: trivial (to the unique leaf)
    - For ambiguous reads: procedure based on Penalty Score (PS) calculation

Taxonomic Assignment in Metagenomics

- $R_i \rightarrow M_i = \{s_1, s_5, s_6, s_7\}$
- $T_i \rightarrow$ subtree rooted at $LCA(M_i)$
- $\forall j \in N(T_i) \rightarrow PS_{i,j}$
- Optimal assignment of $R_i \rightarrow \{j \in N(T_i) : PS^q_{i,j}$ is minimum $\}$



$$PS^q_{i,j} = q \cdot \frac{|FN_{i,j}|}{|TP_{i,j}|} + (1-q) \cdot \frac{|FP_{i,j}|}{|TP_{i,j}|}$$

Taxonomic Assignment in Metagenomics

Problem

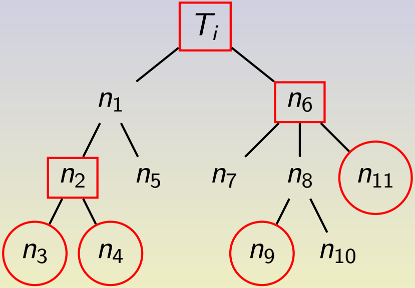
Taxonomic assignment of the reads based on penalty score calculation is slow and strongly depends on the chosen taxonomic tree

- Solution
  1. New optimal algorithm for the minimum penalty score computation
  2. Multiple taxonomic trees (e.g. NCBI, Silva, RDP, Greengene, LTP)

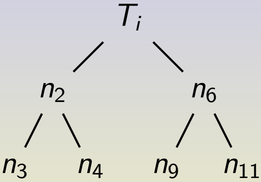
Taxonomic Assignment in Metagenomics

1 - New algorithm for minimum *PS* calculation

- Subtree T_i of T



- Skeleton tree $T_{M_i}^{LCA}$

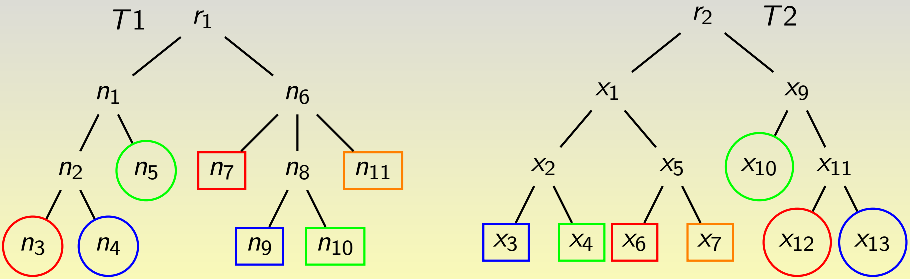


- $\min PS(T_i) = \min PS(T_{M_i}^{LCA})$
- time: $O(|T_i|) \rightarrow O(|M_i|)$

Taxonomic Assignment in Metagenomics

2 - Multiple taxonomic trees

- Taxonomy pre-processing
  - Tree contraction (to valid taxonomic ranks)
  - Leaf mapping (to other taxonomies)
- Input conversion
  - Mapping of valid alignments in another taxonomy



Taxonomic Assignment in Metagenomics

- Results
  - Penalty score can be computed on skeleton tree in an optimal way (proved lemma)
  - Leaf mapping and tree contraction for different taxonomies
  - Read assignments on different taxonomic trees with penalty score calculation
- Work done during the period abroad in collaboration with Prof. Gabriel Valiente (*Technical University of Catalonia - Department of Software, Barcelona, Spain*)

Conclusions

- Taxonomic Assignment in Metagenomics:
  - New optimal algorithm based on skeleton tree for the penalty score calculation
  - Flexible solution to integrate taxonomic information of different taxonomies
  - Fast way to perform the read assignment to new trees
- Developments:
  - Compare assignments on different taxonomies
  - Validate taxonomy structures
  - Refine classification of different organisms