Minimum Factorization Agreement of Spliced ESTs

Paola Bonizzoni

Gianluca Della Vedova

Riccardo Dondi

Yuri Pirola

Raffaella Rizzi

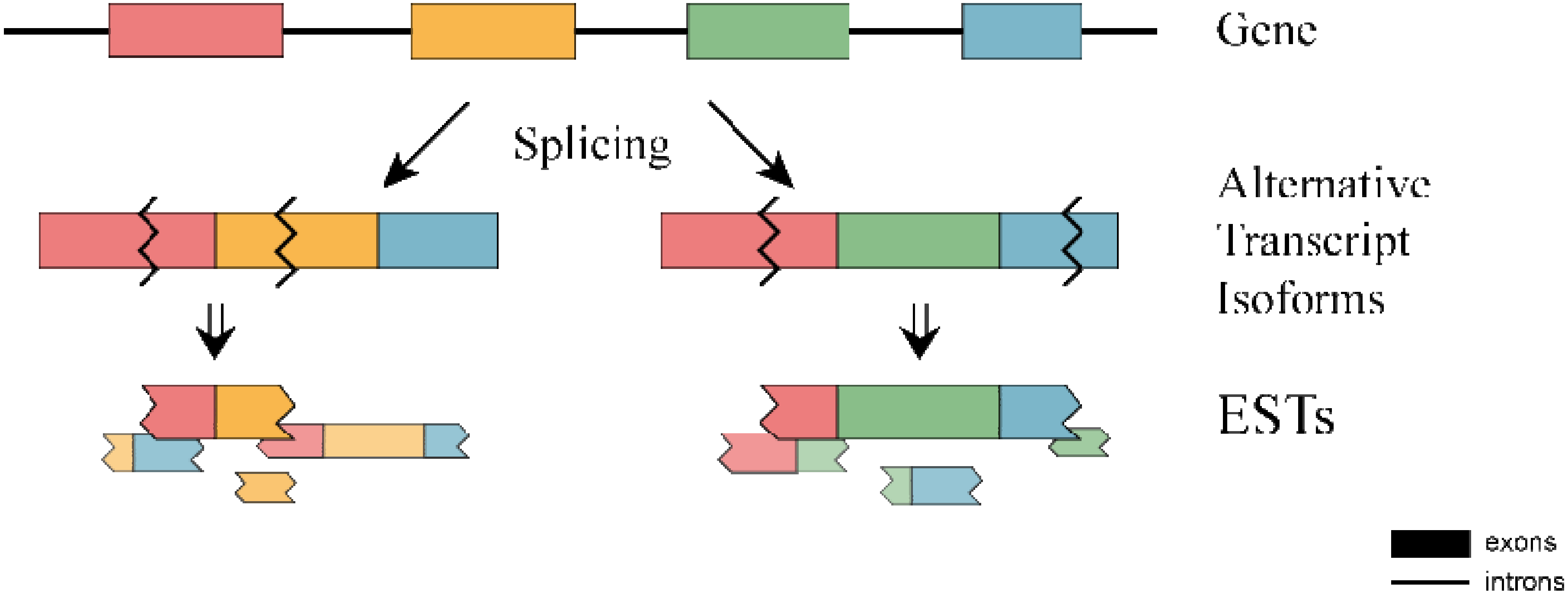
University of Milano-Bicocca

Outline

- Introduction and Motivations
- The Minimum Factorization Agreement Problem
- The Algorithmic Solution
- Experimental Results
- Conclusions

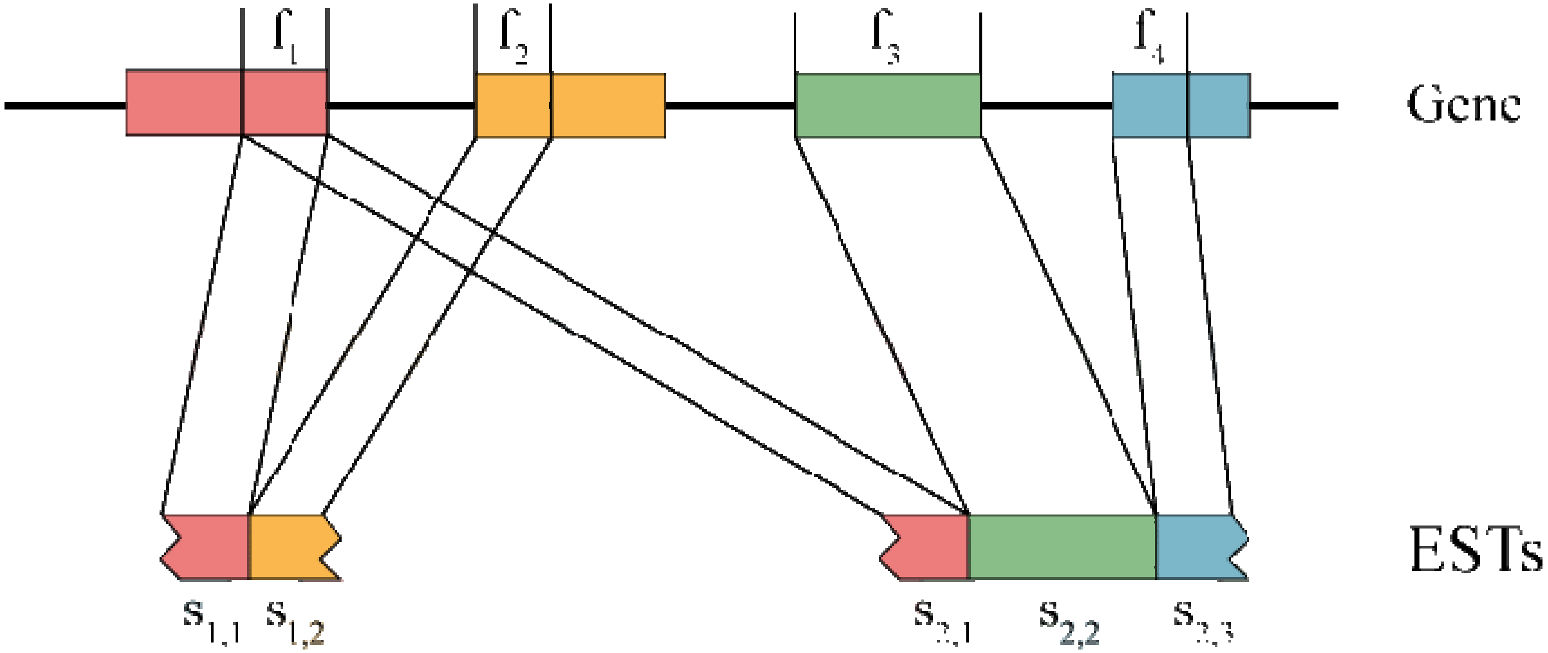
What is an EST?

- **E**xpressed **S**equence **T**ag (**EST**) =
  short fragment of a transcript
- But: Alternative Splicing → 1 gene = *n* transcripts



What is an EST?

f_1, f_2, f_3, and f_4 = **factors**



composition of EST_i = f_1, f_2

spliced EST of EST_i = $\{(s_{1,1}, f_1), (s_{1,2}, f_2)\}$

Why ESTs are used?

- Considerations:
  - ESTs are cheap to obtain
  - ESTs provide some information about transcripts
- Common idea:

  Combining several ESTs to predict:

    - *alternative splicing events*
    - *intron-exon structure*
    - *alternative transcripts*
    - *…*

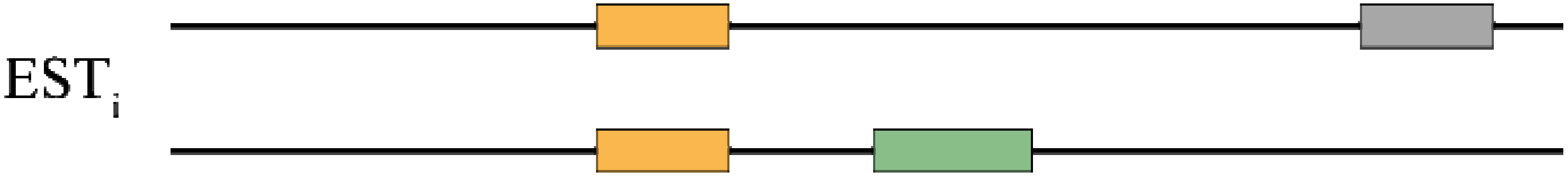
Basic ingredient:
spliced ESTs

Problems

Using ESTs poses several problems due to…

- Sequencing errors
  - especially along the terminal factors
  - near the splice junctions
- Terminal EST factors may be short (10-30bp)
- Genomic sequences may present **repeated substrings**

An Example

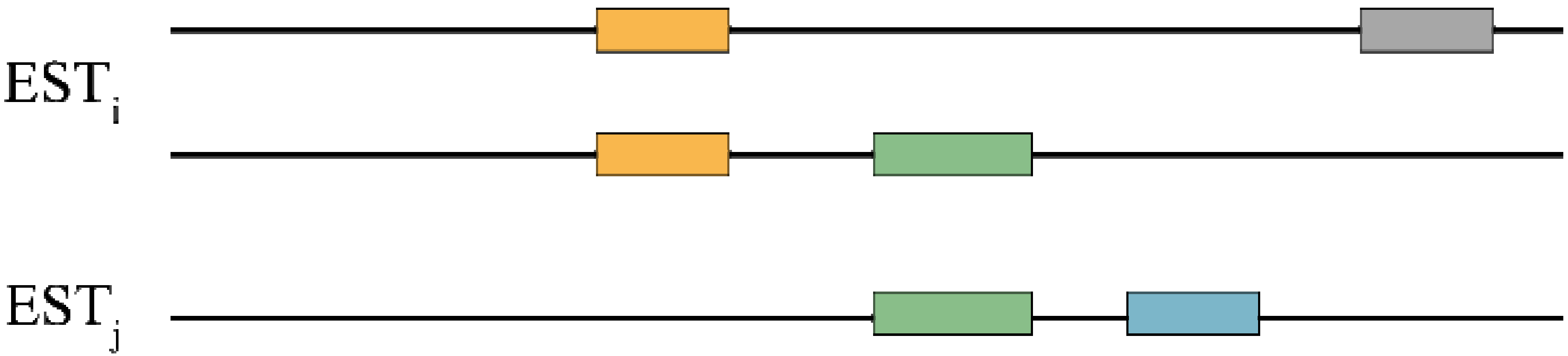


Two possible compositions…

How to choose the “correct” one?

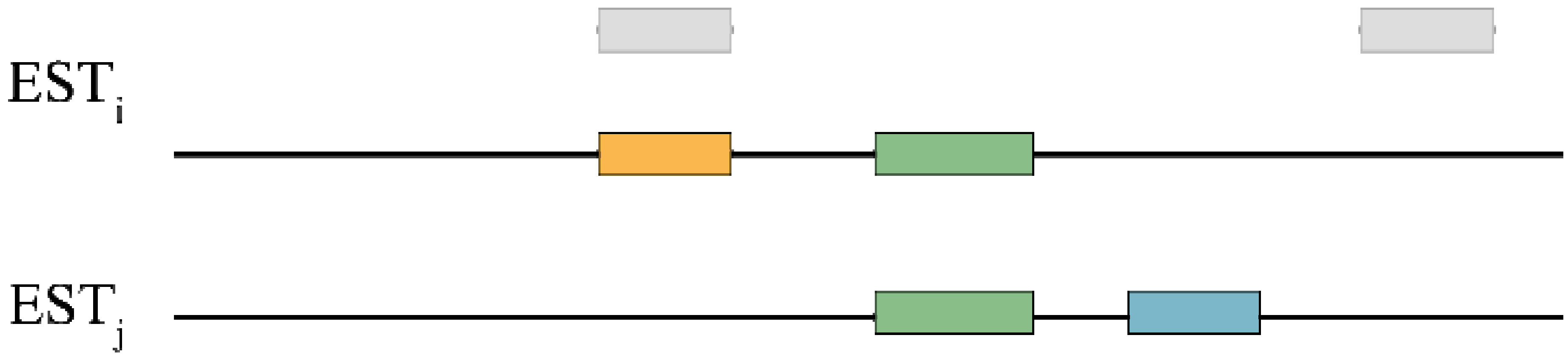
Popular tools (GMap, EST_GENOME, Spidey, …) often report one “best alignment”

An Example



Idea: redundancy can help to choose the “right” one

An Example



Idea: redundancy can help to choose the "right" one

Minimum Agreement Factorization problem

Minimum Agreement Factorization problem (MAF)

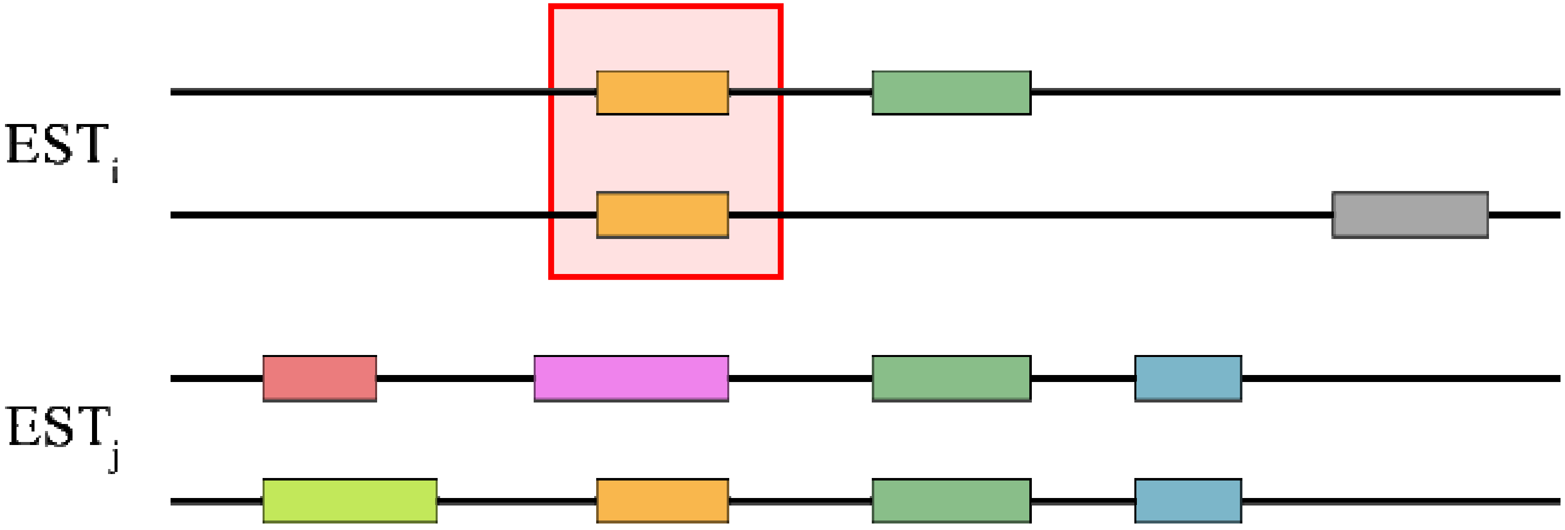
- **Input:**
  The set of compositions C(S) of a family S of EST sequences (over the set of factors F)
- **Output:**
  A minimum-cardinality set F' of factors such that for each EST of S, there exists a composition that uses only factors in F' (F' is a factorization agreement set).

- APX-hard *(by L-reduction from Min Set Cover)*

Real Instances

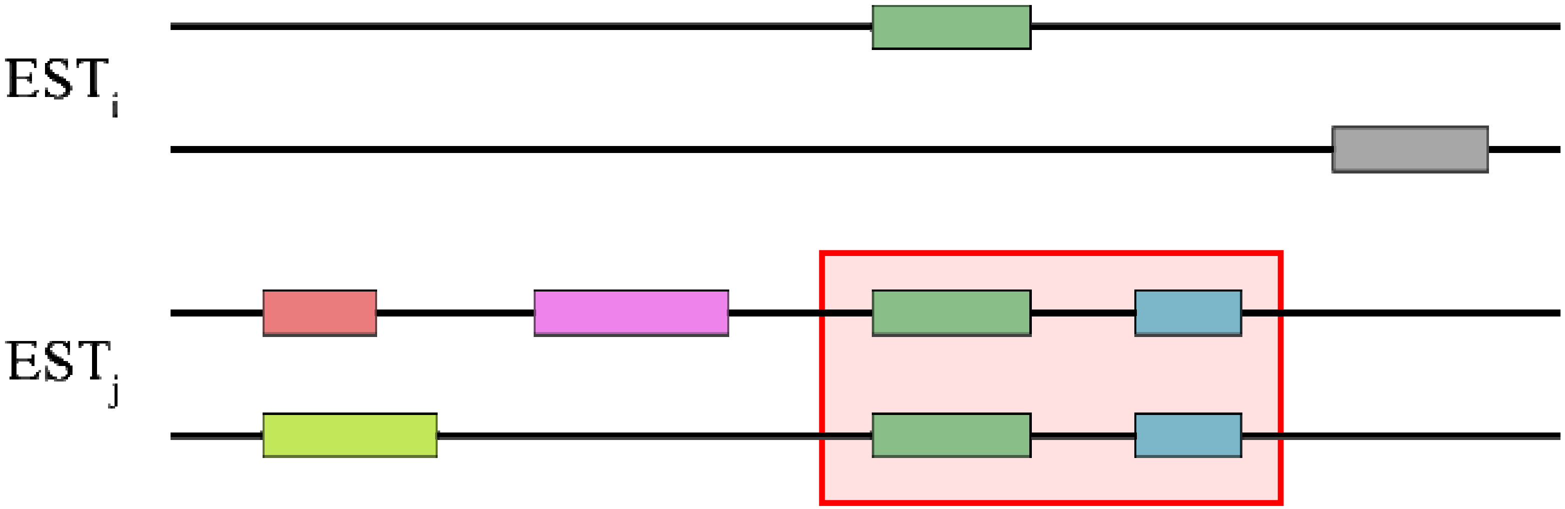
- On real data, several factors that must belong to every optimal solution can be (easily) identified (**necessary factors**)
- *Idea:* identify and remove necessary factors
  - Five rules
  - Efficient (polynomial-time)

Size-reduction *(by example)*

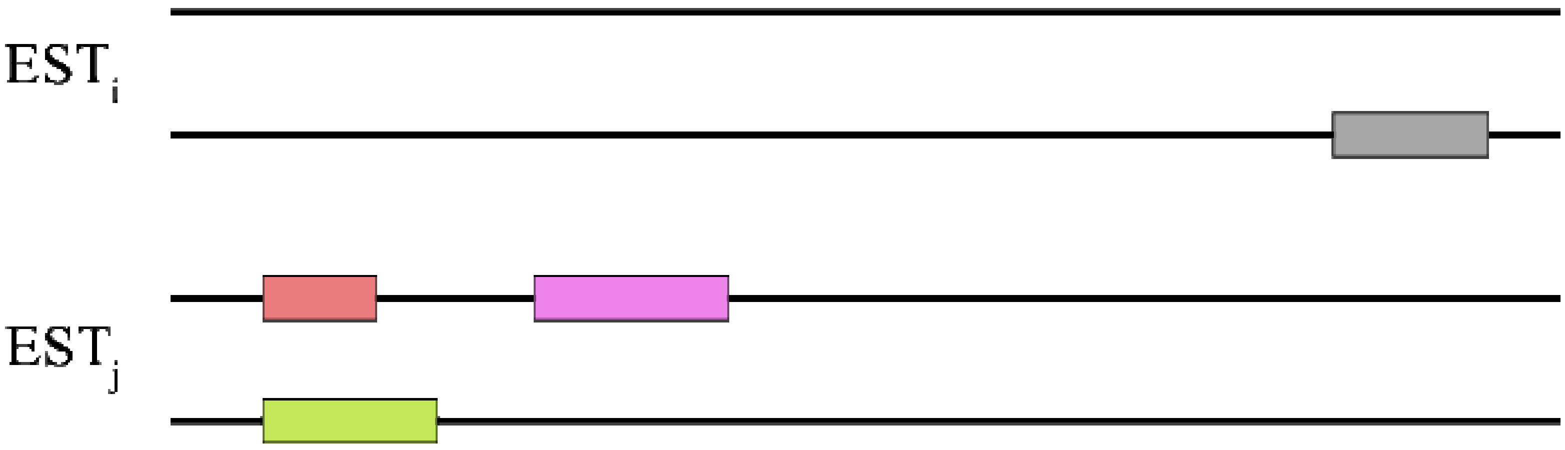


Two ESTs, Four Compositions/spliced ESTs

Size-reduction *(by example)*

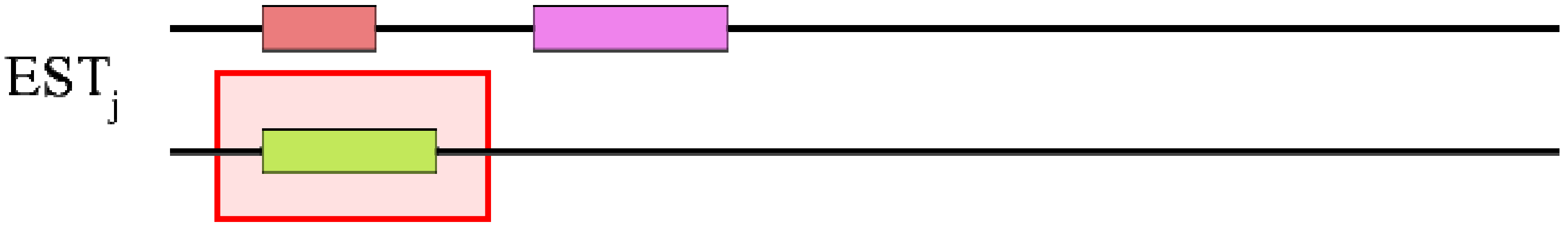


Size-reduction *(by example)*



EST_i can be removed since one of its compositions is empty.

Size-reduction *(by example)*



No more reduction rules can be applied…

…but the problem is easier.

The Algorithmic Solution

- Two-step algorithm:
  1. **Size-reduction**
     (identification of necessary factors)
  2. (Exponential-time) **Exact algorithm**
     (on the remaining factors)

Exact Algorithm

- Exact algorithm: (naïve version)
  - Enumeration of all subsets of factors in non-decreasing order
  - Checking if the subset is a factorization agreement set
  - Exponential-time in $|F|$: $O(2^{|F|} |F| |C(S)|)$

- Usually $|F|$ much smaller than $|C(S)|$

Exact Algorithm

- Naïve version: efficient implementation
  - Bit-parallelism
  - Data locality
- Refined versions may:
  - Discard part of the search space
  - Storing previously computed values
- Refined versions require (often):
  - Extra space
  - Complex implementations

Preliminary Experimentation

- *Data (given a gene):*
  - its genomic sequence
  - its UniGene EST cluster
  - a set of spliced ESTs (compositions) based on the longest common substrings between genomic and EST sequence
- *Results (on 4 genes):*
  - Size-reduction step finds an optimal solution (the exponential algorithm is not needed)
  - The solutions are similar to the ones obtained from another well-known tool (GMap)

Conclusions

- **Conclusions:**
  - A method which exploits redundancy to resolve ambiguity in spliced ESTs
  - Theoretical computational complexity ≠ practical feasibility

- **Future works:**
  - In-depth experimentation *(ongoing)*
  - Associating different meanings to the concept of “factor” (e.g. splice sites, introns, …)

Minimum Factorization Agreement of Spliced ESTs

Paola Bonizzoni

Gianluca Della Vedova

Riccardo Dondi

Yuri Pirola

Raffaella Rizzi

University of Milano-Bicocca