Fast detection of conserved complementary motifs using gapped-seed associative arrays

Dmitri D. Pervouchine

Roderic Guigó (Center for Genomic Regulation, Spain)
Andrei Mironov & Mikhail Gelfand (Moscow State University, Russia)

- RNA structure prediction appears to be **a completely different business at different scales of RNA sequence length**
- short RNAs, $\leq$ 200nts, thermodynamic model works fine
- long RNAs, kilobases and megabases
  - The requirement of nested RNA structure is the major limitation
  - Only a small corner of the search space is explored
  - $O(n^k)$, $k \geq 3$ is irrelevant as soon as the model is incomplete

Talk outline

- A novel **ultra-fast** method for detecting conserved complementary motifs
  - Dictionary (n-mer $\mapsto$ where it occurs)
  - Complementarity and conservation = intersection of dictionaries
  - **Exhaustive transcriptome-wide** search in linear time
  - Does not require multiple sequence alignment as an input
  - No limit on the distance between complementary motifs
- In application to **RNA (intra-molecular) secondary structure**
  - RNA structures associated with alternative splicing
  - in fruit flies[1]
  - in placental mammals[2]
  - in nematodes
  - (**morning session**)
- In application to **RNA-RNA interaction prediction**
  - non-coding RNAs as possible trans-regulators of pre-mRNA splicing
  - long non-coding RNAs (lncRNAs) and snoRNAs
  - non-coding segments of protein-coding genes
  - (**evening session**)

[1] Raker *et al*, NAR 37(14):4533-44, 2009

[2] Pervouchine *et al*, RNA 18(1):1-15, 2012

Intermolecular RNA (binary) Interaction Search

- Intermolecular RNA Interaction Search = IRIS[1]
  - intra- and inter-molecular structure simultaneously
  - thermodynamic model, dynamic programming, $O(n^3 m^3)$
  - no loop models for RNA structure with pseudoknots
- IRIS + binary search = **IRBIS** (Snow Leopard)
  - No dynamic programming
  - **Conservation** is a powerful and restrictive filter
  - Nearly exact matches, internal loops 2×2
- Workflow
  - Genomic annotation (reference genome)
  - Transcriptome segmentation (by exon boundaries)
  - Boundaries projected to other genomes (blastZ)
  - Orthologous segments
  - **Binary gapped-seed search**
  - Candidate selection
  - Extension, alignment, and visualization
- `http://genome.crg.es/~dmitri/irbis.html`
- soon at `https://github.com/pervouchine/irbis/`

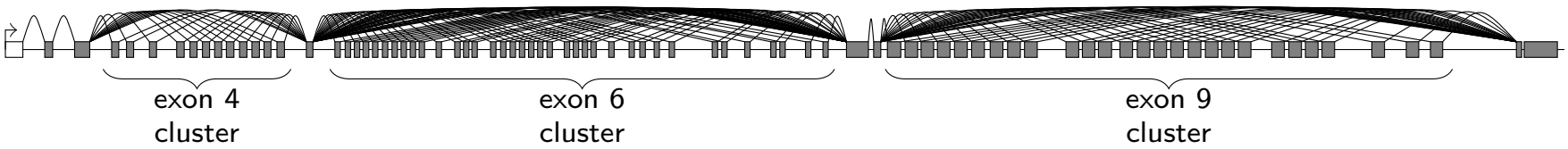
[1] D. Pervouchine, Genome Informatics 15(2), 2004

Part I. Gallery

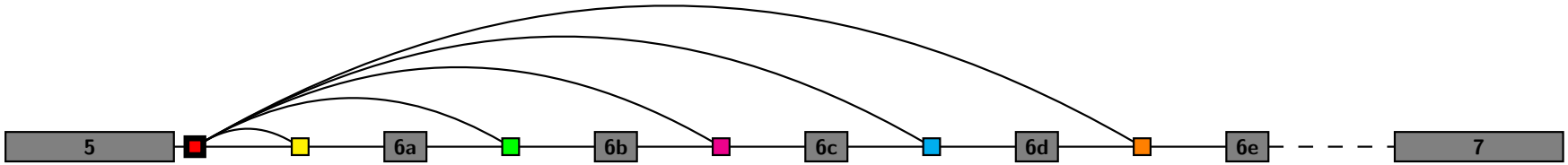
Part II. Algorithm

Part III. Results

Gallery: Mutually exclusive splicing in *Dscam* gene



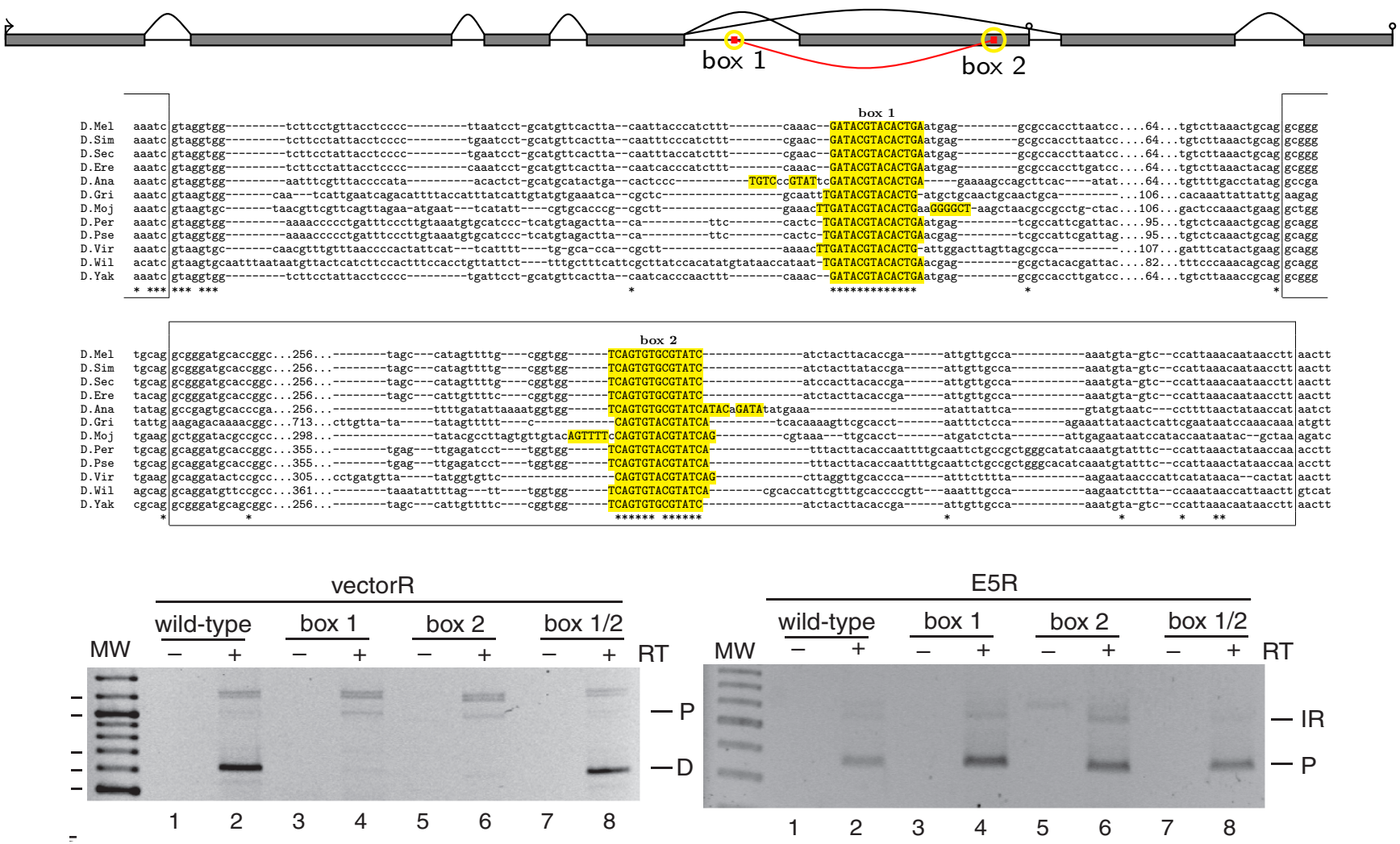
- 12 exons in Exon 4 cluster
- 37 exons in Exon 6 cluster
- 27 exons in Exon 9 cluster
- $12 \times 37 \times 27 \simeq 12,000$ alternative transcripts
- **One and only one** exon from each cluster is included



- Mutually exclusive base-pairing $\Longrightarrow$ mutually exclusive exon choice (May *et al* 2011, Graveley 2005)
- These base-pairings span over 10–15 Kb!

Gallery: splicing and polyadenylation in NMNAT gene

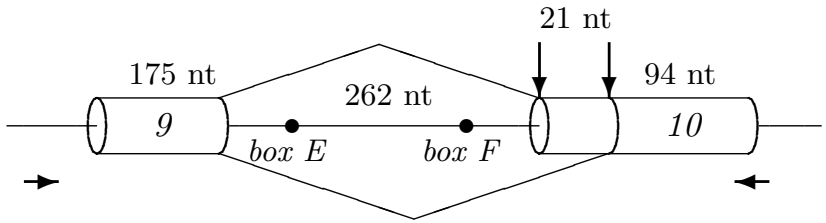
VA Raker, AA Mironov, MS Gelfand, DD Pervouchine, NAR 2009

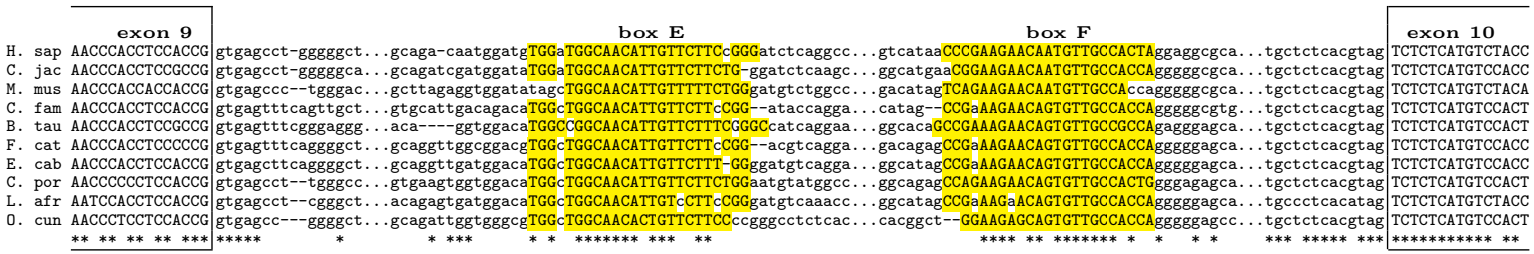


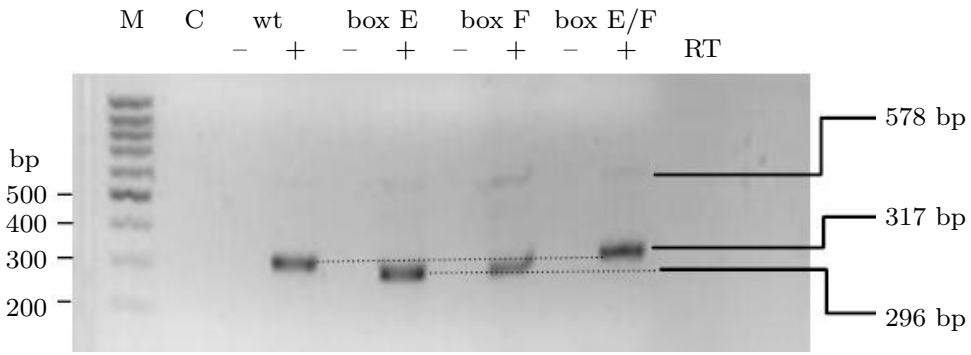
RNA structure affects **both** splicing and polyadenylation

Gallery: Splicing factor 1 (SF1)

Pervouchine *et al*, RNA 2012



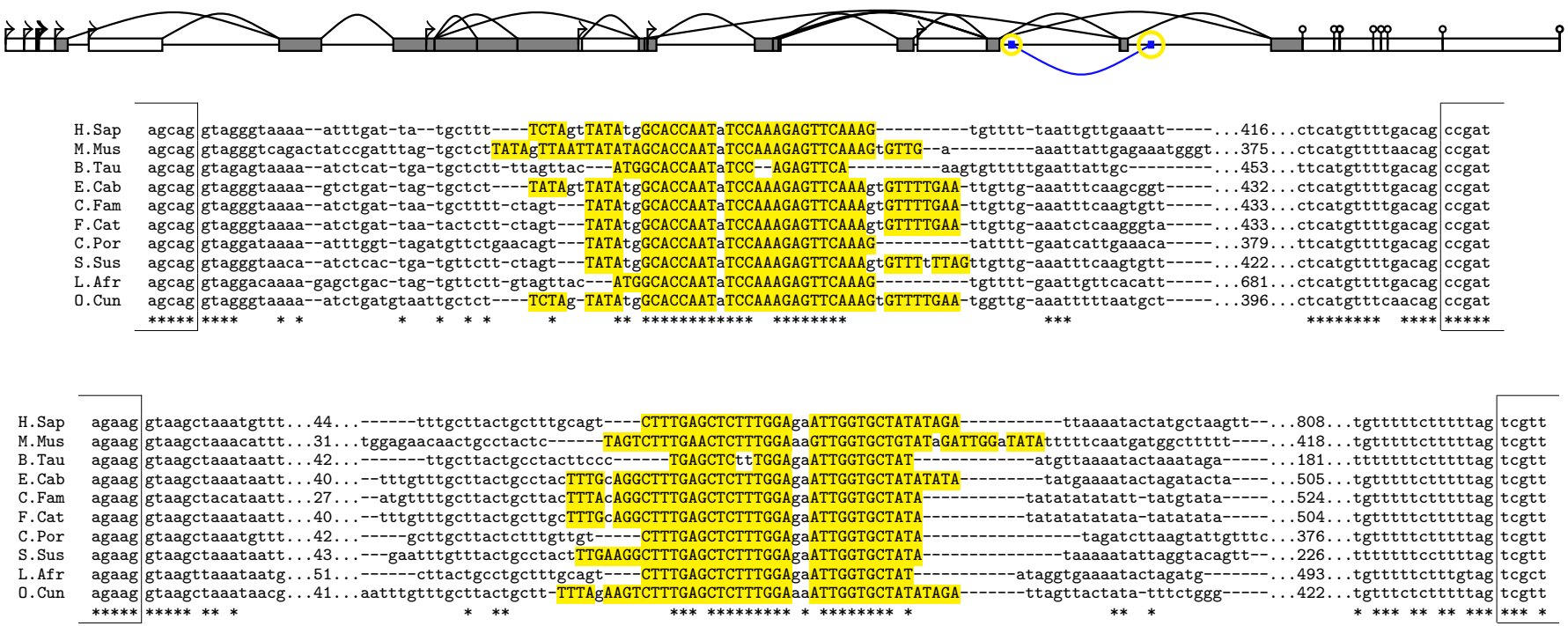




- Intron between exon 9 and 10 contains premature stop codon
- ESTs from breast and uterine adenocarcinoma cell lines support distal acceptor

Gallery: human splicing factor SRSF7

Pervouchine *et al*, RNA 2012



Part I. Gallery

Part II. Algorithm

Part III. Results

Complementarity: no gapsgapped seeds

Sequence A

ATACGAGTCTGATCATTTACGGTCTTATACCGGTCTTATAC

ATACGAGTCT
TACGAGTCTG
ACGAGTCTGA

n-mer	position
ACCGGTCTTA	28
ACGAGTCTGA	2
ACGGTCTTAT	18
AGTCTGATCA	5
ATACCGGTCT	26
ATACGAGTCT	0
ATCATTTACG	11
ATTTACGGTC	14
CATTTACGGT	13
CCGGTCTTAT	29
CGAGTCTGAT	3
CGGTCTTATA	19, 30
CTGATCATTT	8
CTTATACCGG	23
GAGTCTGATC	4
GATCATTTAC	10

ATACGAGTCT

ATACGAGT
ATAC AGTC
ATAC GTCT
TACGAGTCTG
TACGAGTC
TACG GTCT
TACG TCTG
ACGAGTCTGA
ACGAGTCT
ACGA TCTG
ACGA CTGA

n-mer	position:gap

Sequence B

TACCTCGATGCAGAAATCGTCGAGACTCGTATCATTCGAGC

TACCTCGATG >> CATCGAGGTA
ACCTCGATGC >> GCATCGAGGT
CCTCGATGCA >> TGCATCGAGG

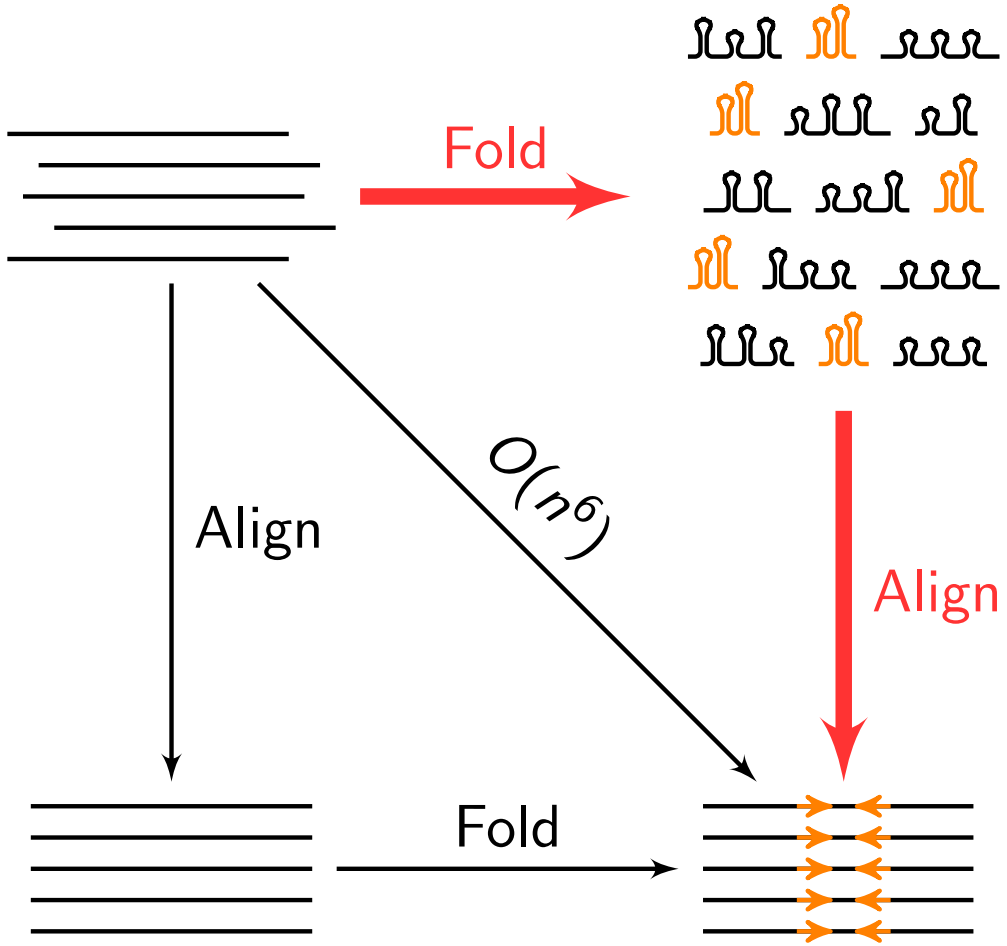
n-mer	position
AATGATACGA	26
ACGAGTCTCG	20
ACGATTTCTG	10
AGTCTCGACG	17
ATACGAGTCT	22
ATGATACGAG	25
ATTTCTGCAT	7
CATCGAGGTA	0
CGAATGATAC	28
CGACGATTTC	12
CGAGTCTCGA	19
CGATTTCTGC	9
CTCGAATGAT	30
CTCGACGATT	14
CTGCATCGAG	3
GAATGATACG	27

TACCTCGATG >> CATCGAGGTA

CATCGAGG
CATC AGGT
CATC GGTA
ACCTCGATGC >> GCATCGAGGT
GCATCGAG
GCAT GAGG
GCAT AGGT
CCTCGATGCA >> TGCATCGAGG
TGCATCGA
TGCA CGAG
TGCA GAGG

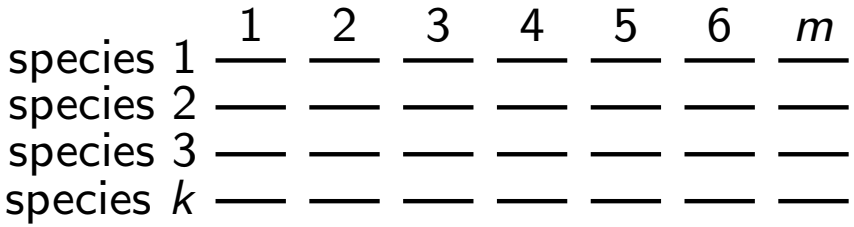
n-mer	position:gap

Align *vs.* Fold: a non-commutative diagram



- $n = 8$, $4^8 = 65535$ words, dictionary size = 8,837,747
- Min number of complementary pairs = 1,191,817,686 (best case)
- For 16 mammals, at least 4 bytes per pair = 342.1 Gb of RAM

Intersection: dropping non-conserved n-mers



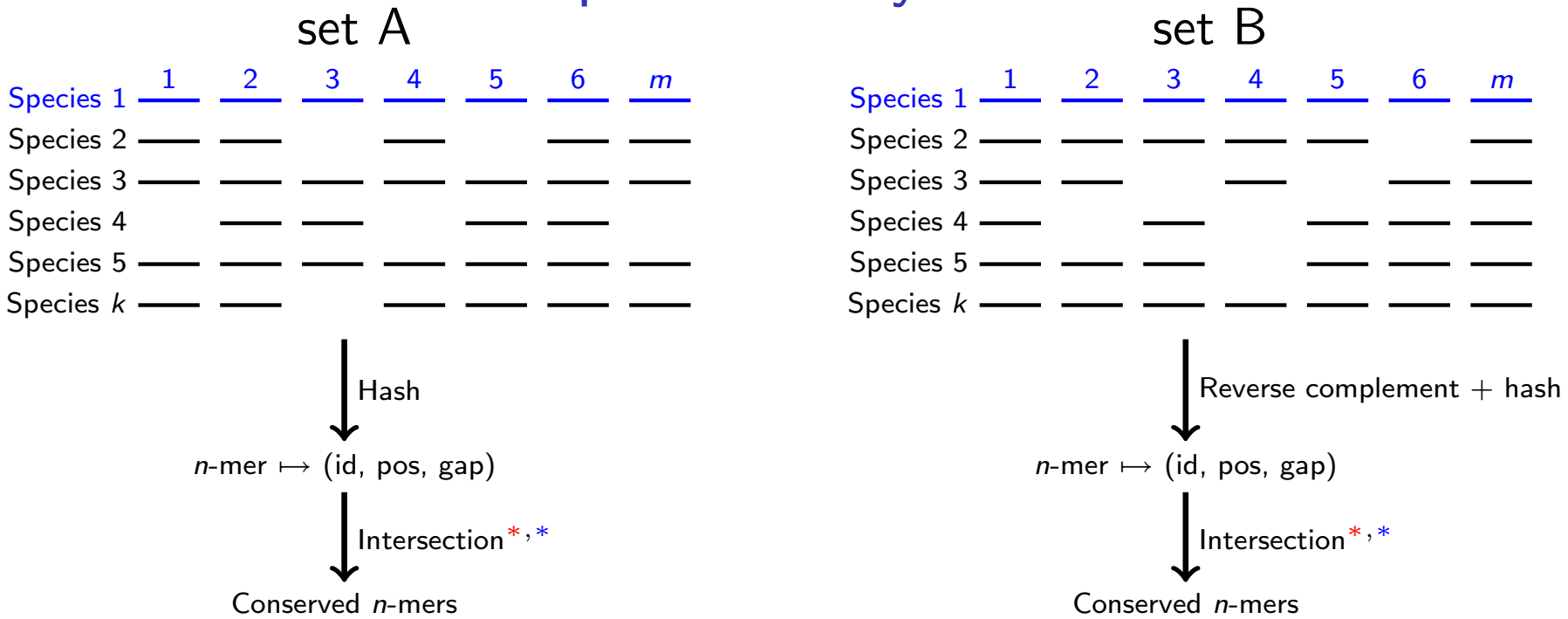
n-mer		segment_id:position:gap
AAAAAAAA	Species 1	1:100:0, 2:100:1, 7:200:1, ...
	Species 2	2:150:1, 4:200:0, 7:100:2...
	...	
	Species k	2:500:0, 7:300:1, 8:400:1...
AAAAAAAC	...	...

blue = pointed at
red = min element
gray = discarded
green = retained

- $i_1{:}p_1{:}g_1 \leq i_2{:}p_2{:}g_2 \iff i_1 < i_2$ or $i_1 = i_2$ & $p_1 < p_2$ or $i_1 = i_2$ & $p_1 = p_2$ & $g_1 \leq g_2$
- $i_1{:}p_1{:}g_1 \simeq i_2{:}p_2{:}g_2 \iff i_1 = i_2$ & $|p_1 - p_2| < M$

for *each n-mer* **do**
  initialize pointers $r_1 = r_2 = \cdots = r_k = 0$;
  while $x = \min\{x_{r_1}, x_{r_2}, \ldots, x_{r_k} | \leq\}$ *is defined* **do**
    compute $c =$ the number of j such that $x \simeq x_{r_j}$;
    keep $x_{r_1}, x_{r_2}, \ldots, x_{r_k}$ if $c >$ threshold;
  end
end

Conservation and complementarity



* seed pattern: 4-2-4; at most 1 GT and at least 2 GC base pairs per seed; sum of weights $\geq 75\%$;
* (id, pos, gap)$\simeq$(id$'$, pos$'$, gap$'$) $\iff$ id=id$'$ & |pos-pos$'$| $< M$
** induced by Cartesian product

One more thing: binary relationship $\mathcal{R} \subseteq A \times B$

Constrain the Cartesian product by a binary relationship $\mathcal{R} \subseteq A \times B$

$A = B =$ segments of protein-coding genes

- $x\mathcal{R}y$ iff $x = y$: local RNA structure
- $x\mathcal{R}y$ iff x and y belong to the same gene: long-range RNA structure within one gene (not necessarily at annotated splicing events)
- **Raker et al, NAR 2009**: $x\mathcal{R}y$ only if the intron $x \longrightarrow y$ is annotated
- **Pervouchine et al, RNA 2012**: $x\mathcal{R}y$ if x and y belong to the same gene

---

- The input to the pipeline: $(A, B, \mathcal{R})$
- $A = B =$ windows around splice sites: RNA structures around splice sites
- $A =$ miRNAs, $B =$ 3'-UTRs, $\mathcal{R} = A \times B$: miRNA targets
- $A =$ snoRNAs, $B =$ windows around splice sites: snoRNA splicing targets
- $A =$ lncRNA segments vs. $B =$ windows around splice sites... **(today at 6pm)**

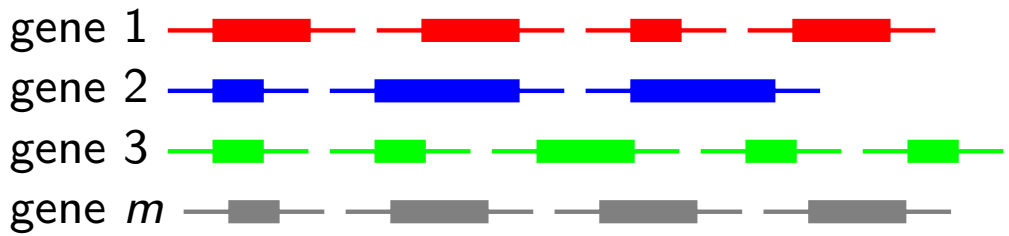
Part I. Gallery

Part II. Algorithm

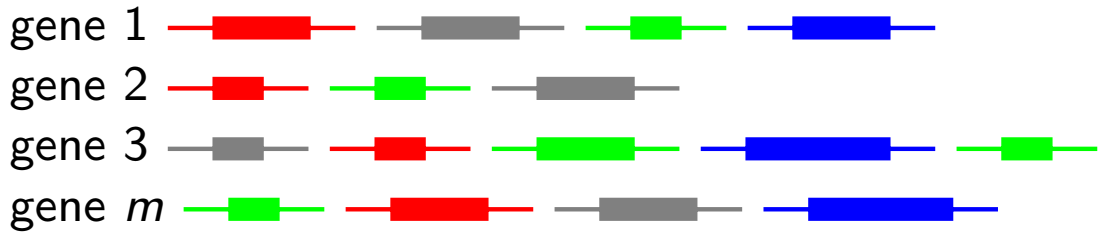
Part III. Results

Statistical control

- Original set of genes

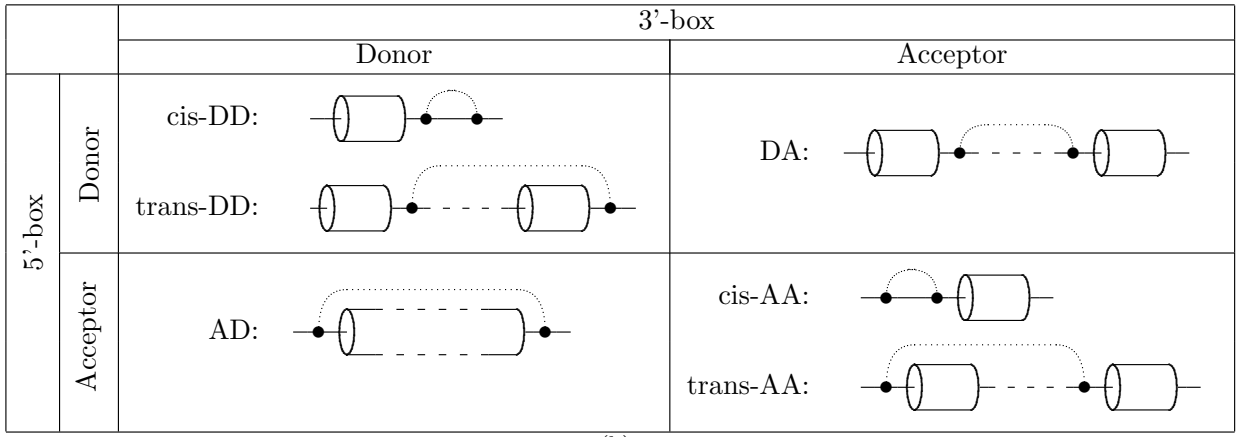


- Control: "re-wired" set



- Look at introns; length reduced to 1000 nts
- Estimate False Positive Rate (FPR)
- Blocking by GC content and/or sequence conservation rate

False positive rate



Repeats	Arrangement	Search	Control	Control GC	Control GC+Cons
Not masked	trans-DD	161	42.5±7.1 (26%±4%)	50.1±7.8 (31%±5%)	72.4±7.5 (45%±5%)
	trans-AA	132	57.0±8.2 (43%±6%)	47.7±7.4 (36%±6%)	60.9±7.1 (46%±5%)
	DA	211	60.1±4.2 (28%±2%)	61.6±4.3 (29%±2%)	76.0±4.1 (36%±2%)
	AD	212	62.6±4.1 (30%±2%)	58.1±4.0 (27%±2%)	80.5±4.7 (38%±2%)
Masked	trans-DD	114	34.2±4.4 (30%±4%)	36.0±4.2 (32%±4%)	27.6±3.5 (24%±3%)
	trans-AA	108	43.1±4.6 (40%±4%)	42.2±4.5 (39%±4%)	43.5±4.1 (40%±4%)
	DA	167	47.4±3.1 (28%±2%)	43.8±3.2 (26%±2%)	50.6±3.0 (30%±2%)
	AD	174	44.7±3.3 (26%±2%)	47.0±3.2 (27%±2%)	42.9±2.9 (25%±2%)

It is not unlikely to find a pair of conserved complementary *n*-mers next to splice sites of mammalian genes

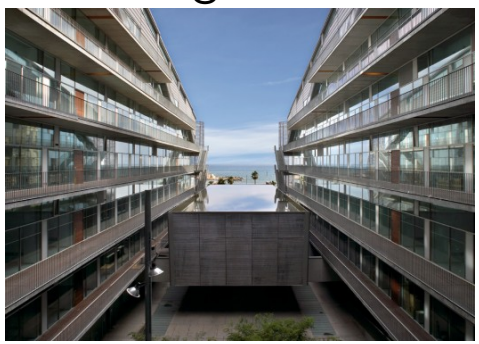
Summary

- IRBIS: a conceptually novel (and computationally realistic) framework for predicting conserved RNA structures and RNA-RNA interactions on genome-wide scale
- Hash table (dictionary) is a natural instrument for simultaneously detecting motif conservation and complementarity
- Implemented as a C++ library

- The set of genes/introns with complementary boxes differs from simple random samples of the same size in many important ways
- Even with FPR as high as 50%, **there is a strong statistical evidence for many stable long-range RNA structures to be conserved and functionally important**

Acknowledgments

Centre de Regulació Genòmica



Roderic Guigó
Alessandra Breschi
Rory Johnson
Angelika Merkel
Andrea Tanzer
Sarah Djebali
Maik Röder
Julien Lagarde
Cedric Notredame
Giovanni Bussotti
Veronica Raker
Juan Valcárcel

Moscow State University

Katya Khrameeva
Marina Pichugina
Ilya Kurochkin
Anya Gerasimova
Petr Rubtsov
Andrei Mironov
Mikhail Gelfand

Oleksii Nikolaienko
Inessa Skripkina
Alla Ryndich

Thank you for your attention

(continued for RNA-RNA interactions)