Forest alignment with affine gaps and anchors

Stefanie Schirmer and Robert Giegerich

Practical Computer Science
Bielefeld University

CPM 2011

RNA structure levels

a)

GCGGAUUUAGCUCAGDDGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA

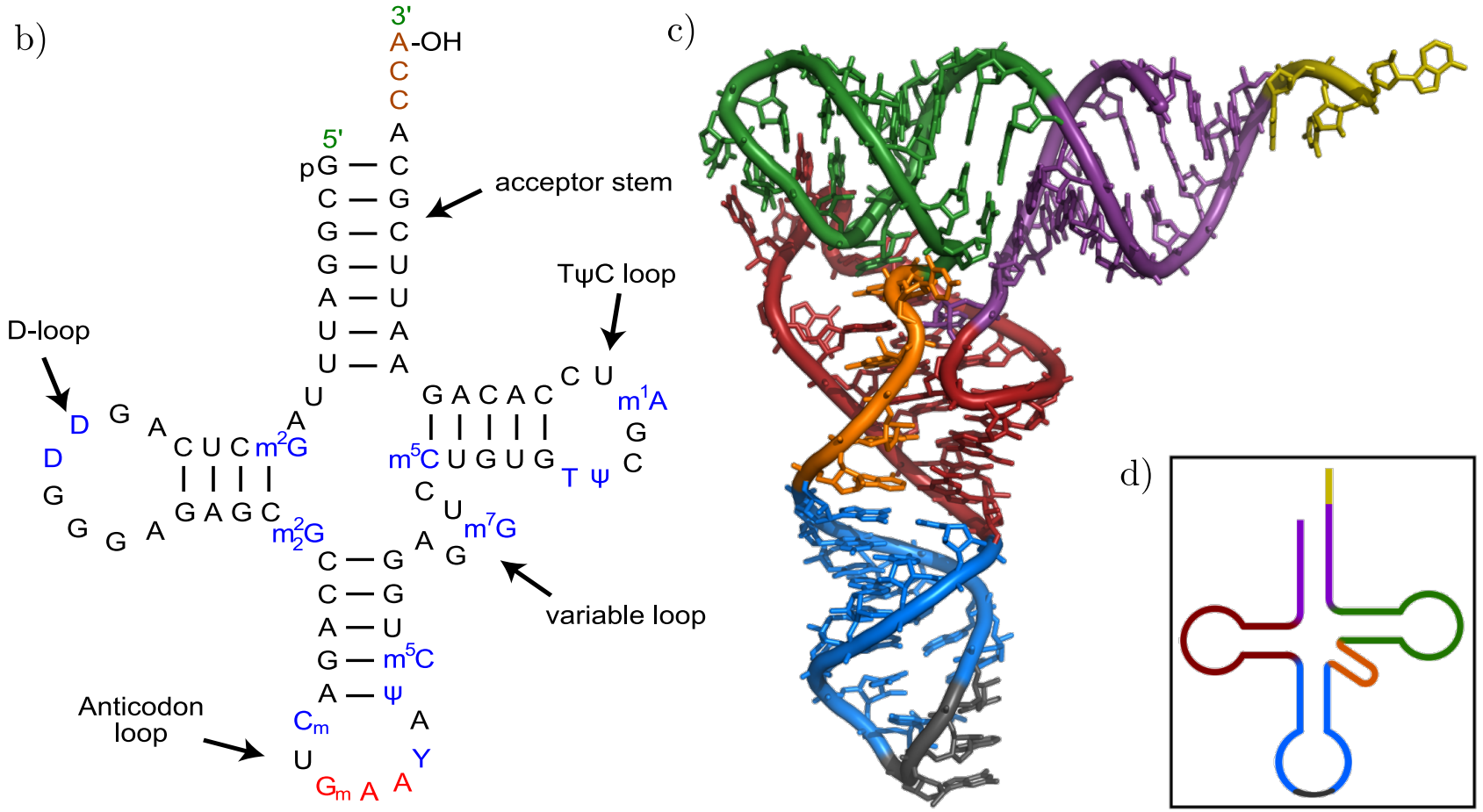


Figure: The three structure levels of an RNA molecule, here: t-RNA for phenylalanine in yeast [9].

Previous works: Forest alignment

Most closely related:

- Tao Jiang, Lusheng Wang, and Kaizhong Zhang. "Alignment of Trees – An Alternative to Tree Edit." In: *Theoretical Computer Science* 143.1 (1995), pp. 137–148
- M. Hoechsmann et al. "Local similarity in RNA secondary structures." In: *Proceedings of the IEEE Computational Systems Bioinformatics Conference (CSB 2003)* 2 (2003), pp. 159–168
- A. Lozano et al. "Seeded tree alignment". In: *IEEE Transactions on computational biology and bioinformatics* (2008), pp. 503–513

Plus 100 articles on RNA structure comparison with other methods

RNA forest representation and forest alignment

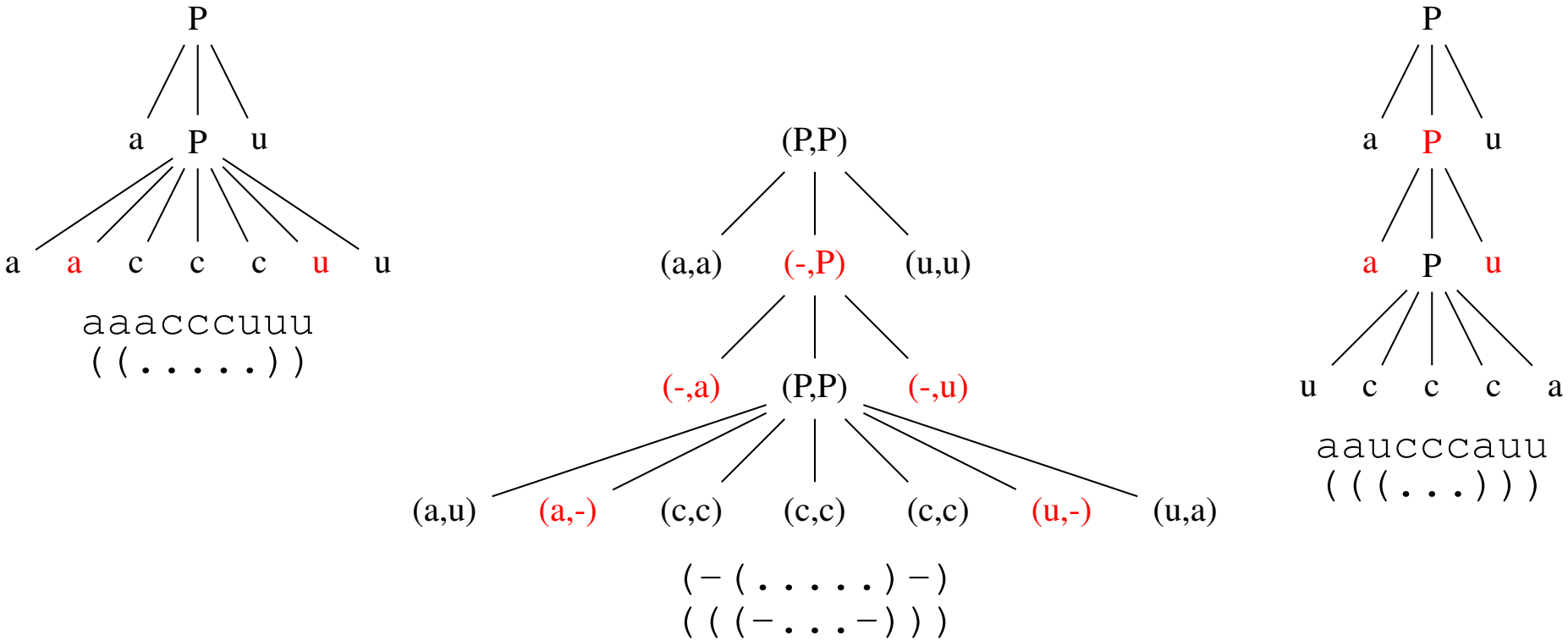


Figure: Example for RNA forest alignment model. Top: two hairpin structures and their tree representations. Bottom: one possible alignment tree.

Tree and Forest alignment

Definition (Alignment tree)

Alignment tree: Tree labeled with pairs from $\{\mathcal{A} \cup \{-\} \times \mathcal{A} \cup \{-\}\}/\{(-,-)\}$.

Definition (Tree alignment)

Tree alignment: Alignment tree $A \in \mathit{Alignments}(F, G)$, which can be transformed

- to F by projecting pair node labels to left component, contracting resulting tree to remove all gaps
- to G in same way after projecting pair node labels to right component

Classic gap model: singleton gap

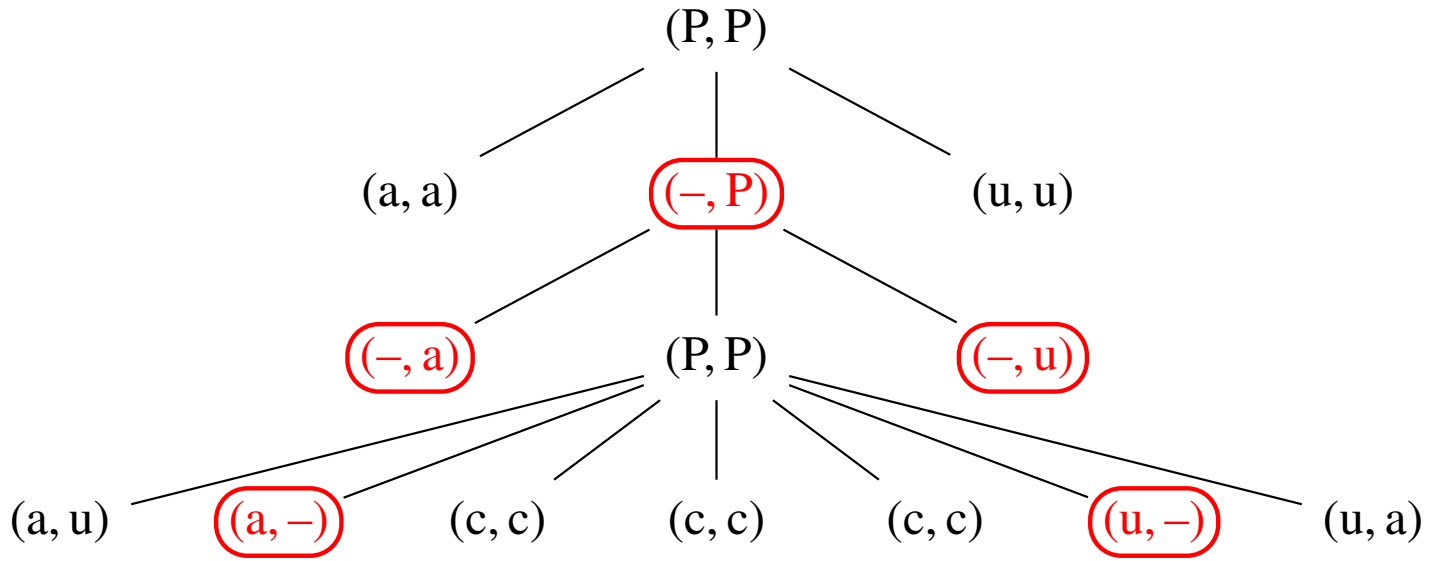


Figure: Singleton Gaps.

Definition (Singleton gap)

Singleton gap: single node in a forest, labeled with gap symbol "–".

New gap model: General gap

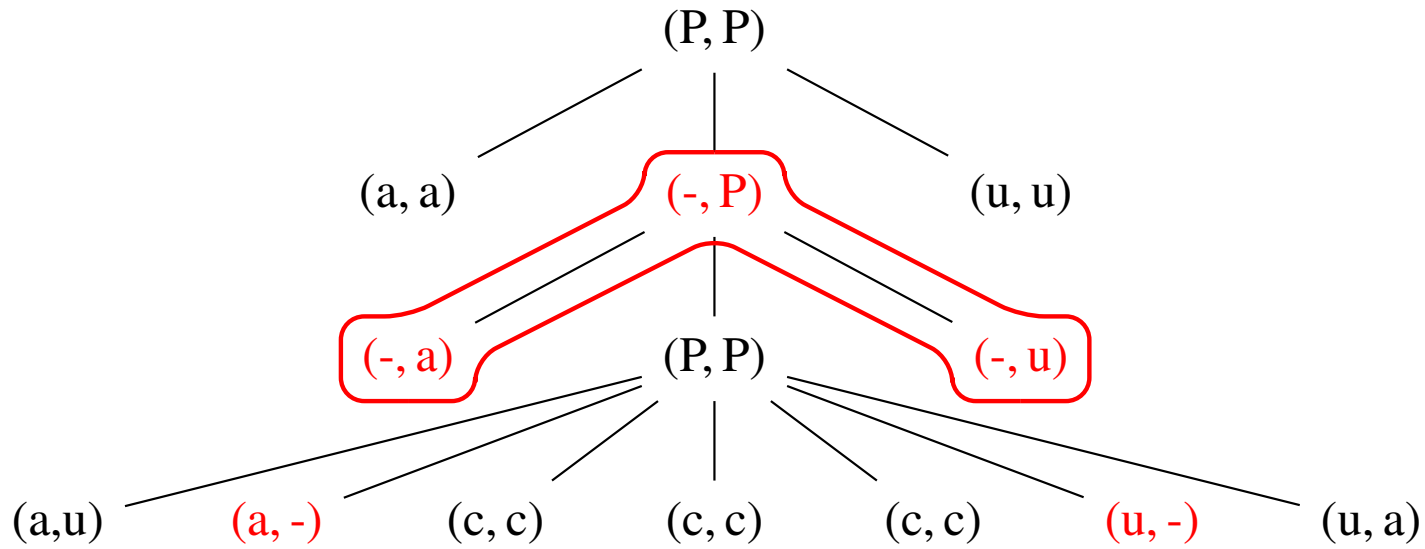


Figure: General gap.

Definition (General gap)

(General) gap: set of singleton gaps in forest, maximal and connected under union of parent-child and direct-sibling relation.
Gap in alignment A of F and G is gap in either left or right projection of A.

New gap model: Oscillating gap

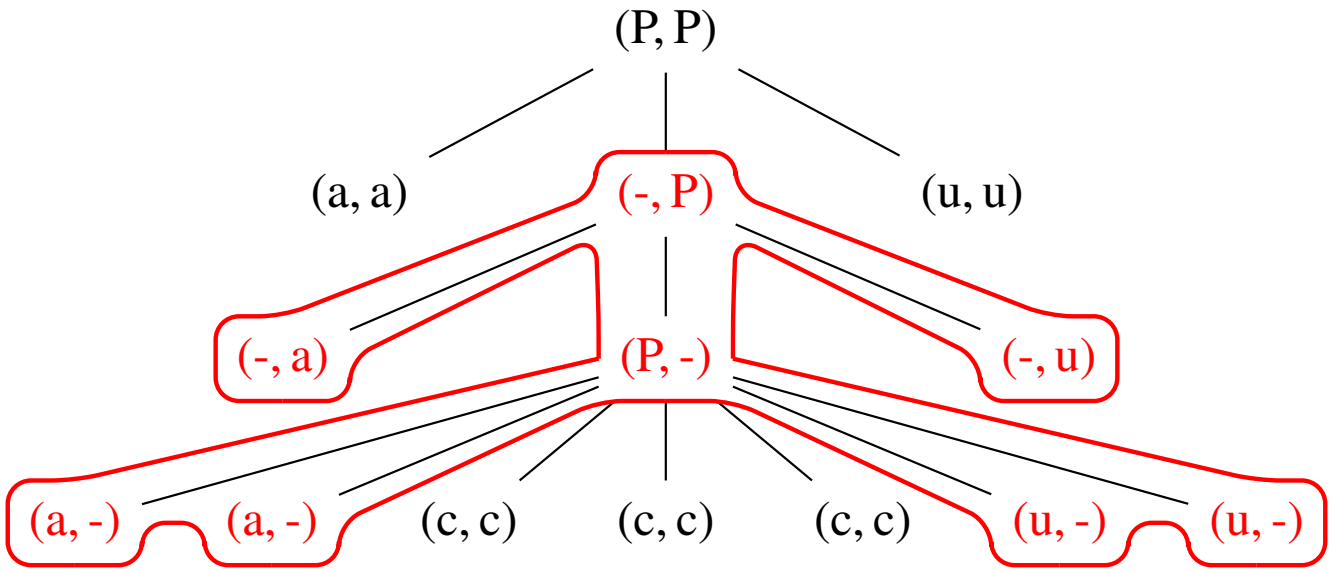


Figure: Oscillating gap.

Definition (Oscillating gap)

Oscillating gap: a gap that may cross from the first to the second component of alignment A and vice versa.

Gap scoring

Classical linear gap model

- each singleton gap adds contribution to score
- score function linear in number of singleton gaps
- procedure can lead to many small gaps / scattered alignment $\rightarrow$ not reasonable (interpreted as evolutionary events)

New affine gap model

- consider adjacent gaps as one unit

$\rightarrow$ *affine gap cost* model: High gap opening costs, low extension costs

- cost function $w(l) = w_{open} + (l-1) \cdot w_{extend}$, l number of singletons
- score of $(a, -)$ depends on context
- algorithm must keep track of opened gaps $\rightarrow$ *gap mode*

Traditional forest alignment

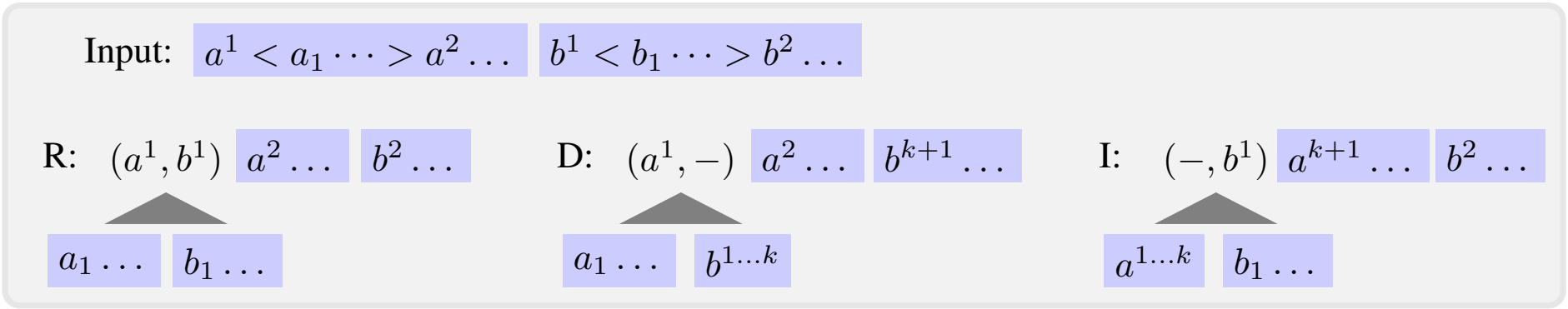


Figure: Recurrences for traditional forest alignment.

Forest alignment with affine gap costs

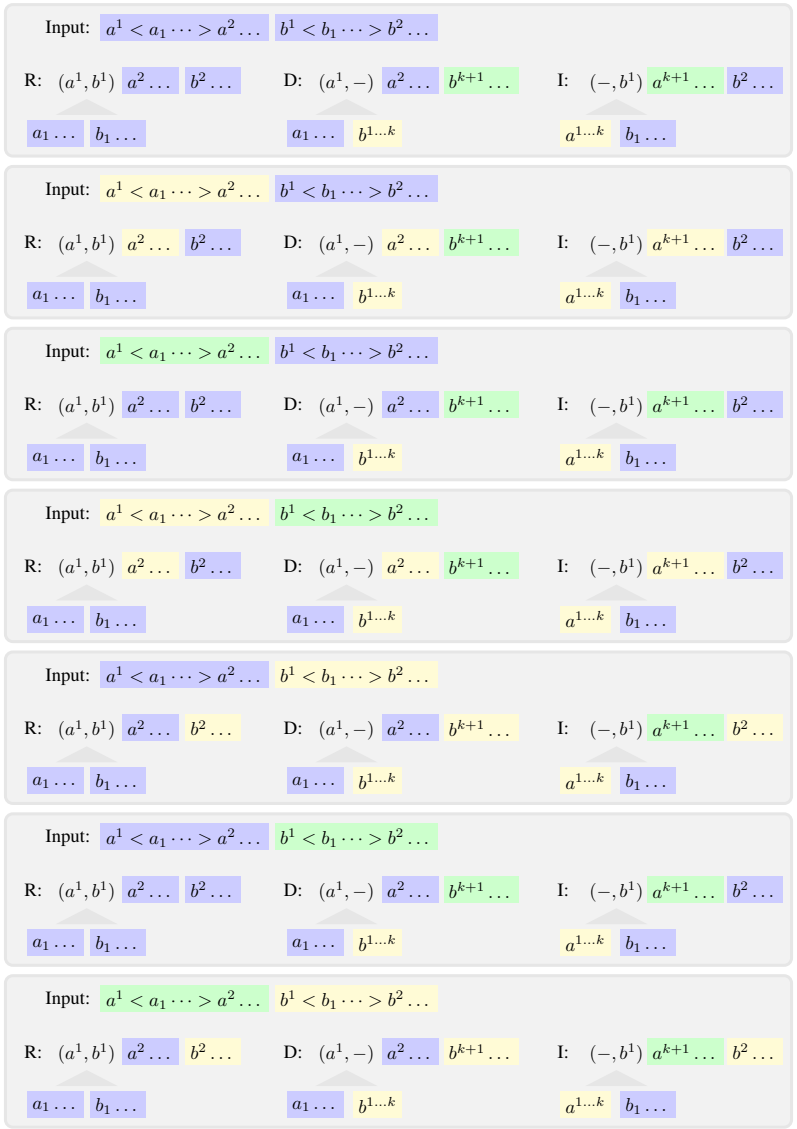


Figure: Affine gap costs: no gap mode, parent gap mode, sibling gap mode.

Forest alignment with affine gap costs: Case 1

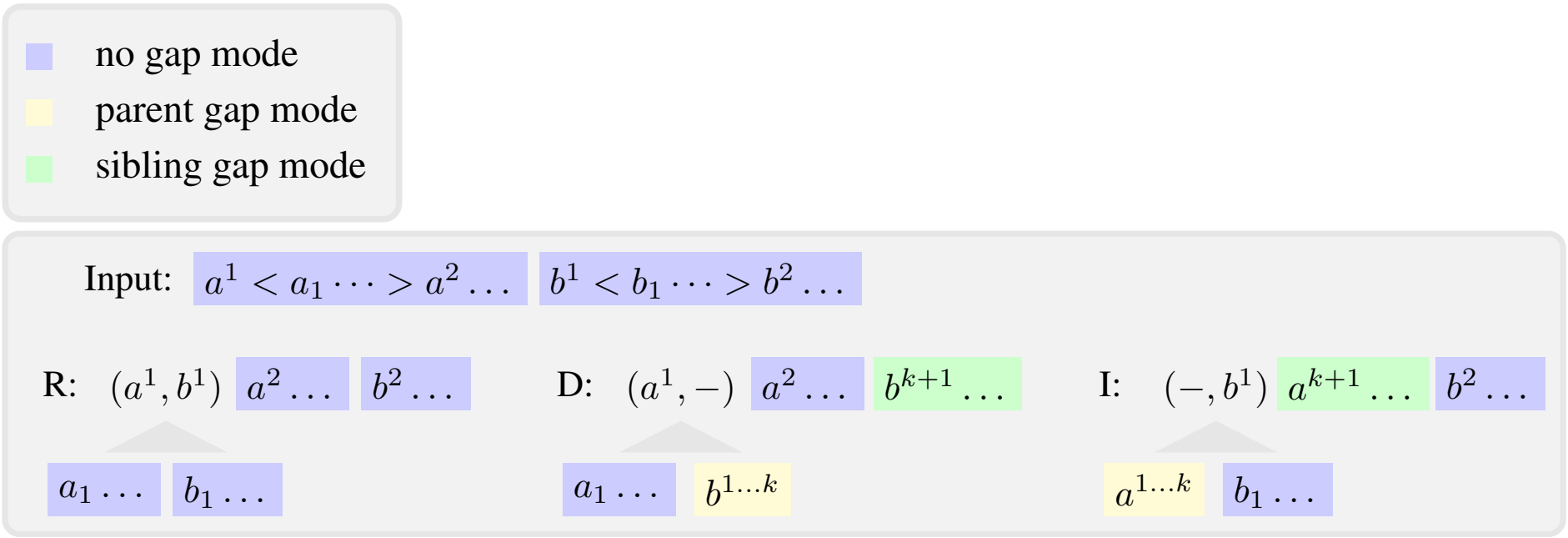


Figure: Case 1 of the forest alignment with affine gap costs.

Shape abstraction

```
CGUCUUAAACUCAUCACCGUGUGGAGCUGCGACCCUUCCCUAGAUUCGAAGACGAG
((((((...(((..(((...)))))))...(((..((.....))..)))))))..
```

```
Shape Level 5:
[[][]]
Shape Level 4:
[[][[]]]
Shape Level 3:
[[[]][[]]]
Shape Level 2:
[[_[]][_[]_]]
Shape Level 1:
[_[_[]]_[_[]_]]_
```

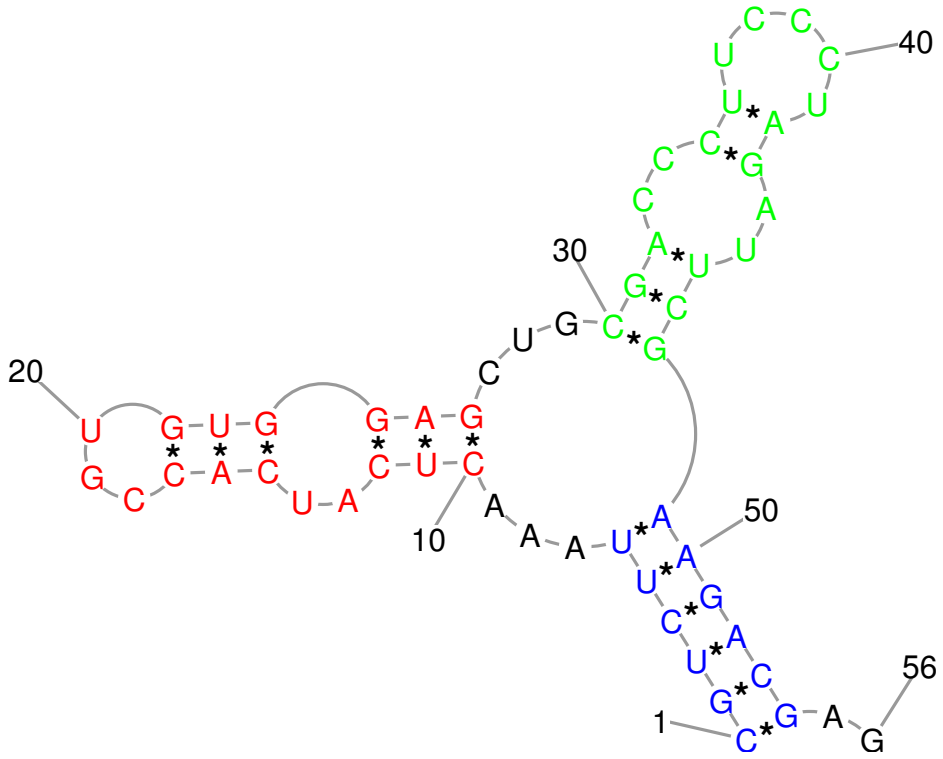


Figure: An example secondary structure and its five shape representations, implemented in the tool *RNAshapes* [7]. Figure from [4]

Anchored alignment

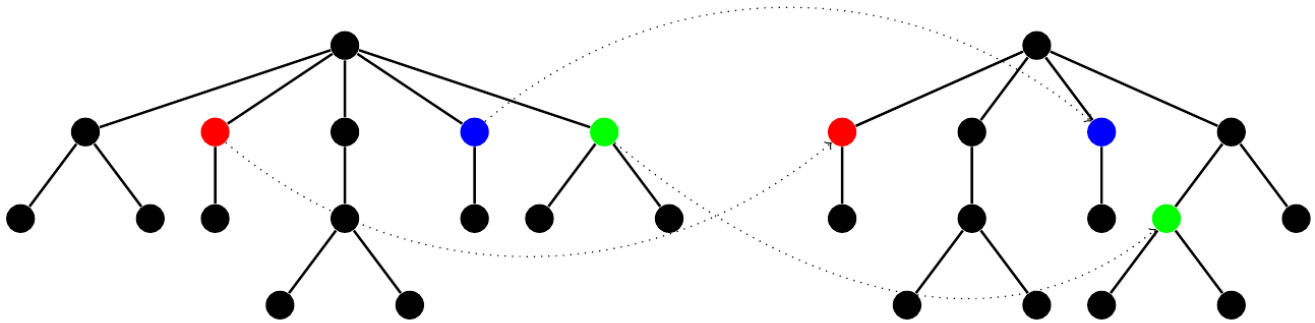


Figure: Anchored alignment input trees with constraints.

- anchoring is a given tree mapping
- tree alignment conforming to an anchoring need not exist

Time complexity: traditional alignment

- let $n = |F| = |G|$
- let $d = deg(F) = deg(G)$
- time complexity for alignment of F and G is $\mathcal{O}(n^2 d^2)$
- RNA: average degree is effectively constant, $d \approx 30$

Time complexity: affine alignment

- remains asymptotically the same with affine gap scoring
- constant factor ≈ 7 is expected
- measured an average runtime factor
  * 8.02 for alignments of folded sequences of ≈ 100 nucleotides
  * 7.79 for those of ≈ 200 nucleotides

Time complexity: anchored alignment

- $k-1$ evenly placed anchors $\rightarrow$ both F and G are split into k parts of about equal size
- n is divided by k in the complexity expression

$\rightarrow$ complexity is reduced from n^2 to $k(\frac{n}{k})^2 = \frac{n^2}{k}$

$\rightarrow$ efficiency with $k-1$ anchors: $\mathcal{O}(\frac{n^2}{k}d^2)$.

Results: speedup by anchoring

Table: Speed-up factor gained by shape anchoring, for ten members of each of the chosen Rfam families, folded by *RNAcast* [5].

RNA Family	Common Shape	Anchors	Avg. Speed-up by Anchoring
5S rRNA (RF00001)	[[] []]	3	1.27
Spot 42 (RF00021)	[] [] []	3	3.30
Cobalamin (RF00174)	[[] [[] [[] []]]]	7	2.96
T-box (RF00230)	[[[] []] [[] []]]	7	3.23

Results: better alignment to true family

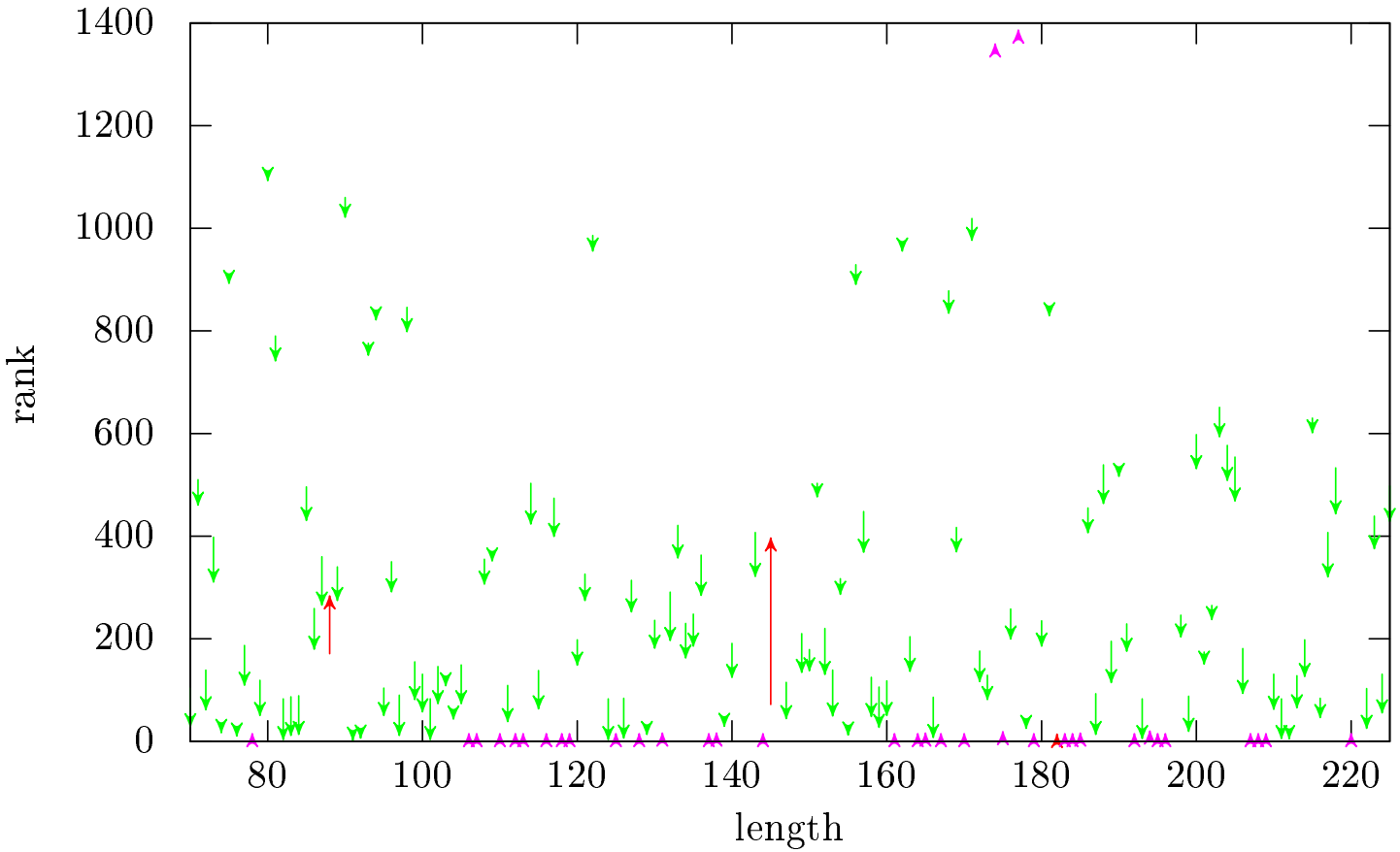
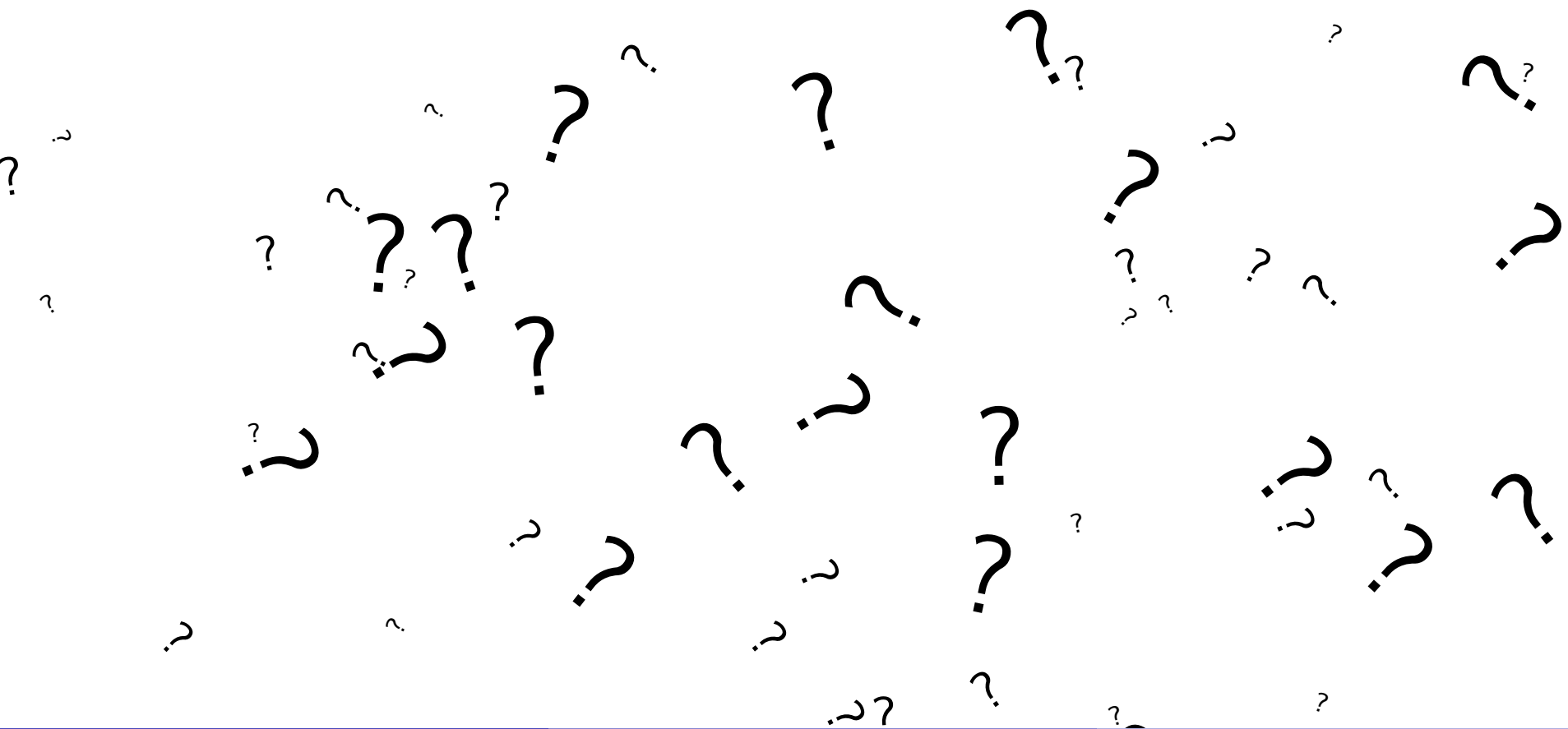


Figure: Improved recognition of RNA family with affine gap costs (green arrow).

Next steps

- submit PhD thesis
- deploy tool on Bielefeld University Bioinformatics Server `http://bibiserv.techfak.uni-bielefeld.de`
- compose shape analysis and anchored alignment for comparative structure prediction

Thanks for your attention.
Questions and remarks are welcome.



Further reading

- Tao Jiang, Lusheng Wang, and Kaizhong Zhang. "Alignment of Trees – An Alternative to Tree Edit." In: *Theoretical Computer Science* 143.1 (1995), pp. 137–148
- M. Hoechsmann et al. "Local similarity in RNA secondary structures." In: *Proceedings of the IEEE Computational Systems Bioinformatics Conference (CSB 2003)* 2 (2003), pp. 159–168
- A. Lozano et al. "Seeded tree alignment". In: *IEEE Transactions on computational biology and bioinformatics* (2008), pp. 503–513
- Francesc Rosselló and Gabriel Valiente. "An algebraic view of the relation between largest common subtrees and smallest common supertrees". In: *Theoretical Computer Science* 362.1 (2006), pp. 33–53
- J. Reeder and R. Giegerich. "Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction." In: *Bioinformatics* 21.17 (2005), pp. 3516–3523
- Hélène Touzet. "Tree edit distance with gaps". In: *Information Processing Letters* 85.3 (2003), pp. 123–129