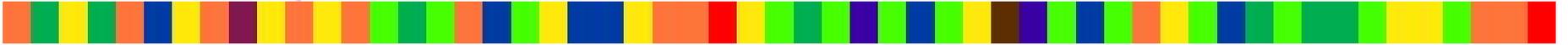




PLANAR: RNA Sequence Alignment using Non-Affine Gap Penalty and Secondary Structure



Ofer Hirsch Gill*,
Naren Ramakrishnan** &
Bhubaneswar Mishra*

(*)Courant Institute, NYU & (**)Virginia Tech



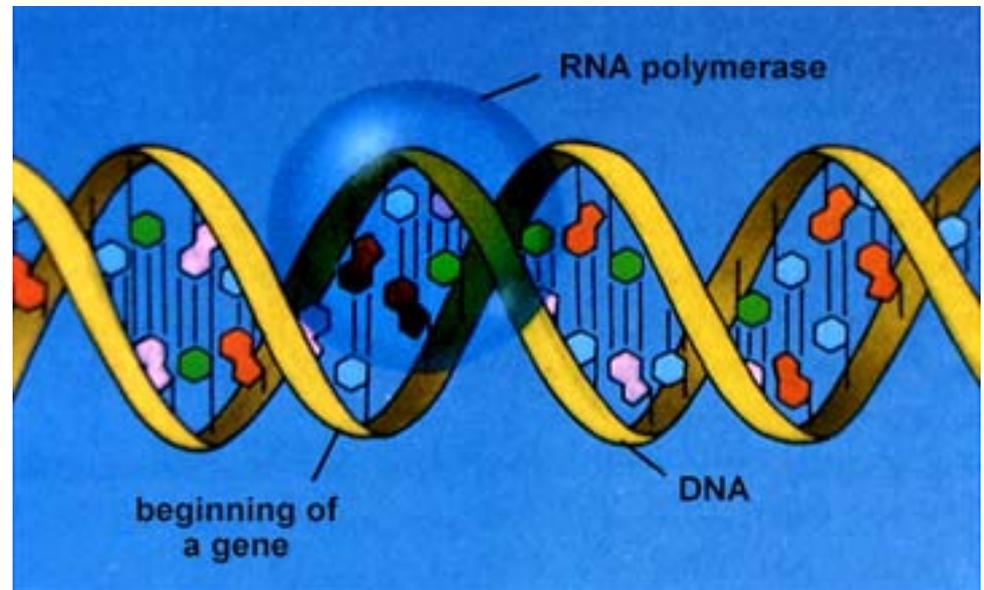
Outline

- **Introduction**
- PLAINS (for DNA Alignment)
- PLANAR (for RNA Alignment)
- SEPA (for Alignment Evaluation)
- Results
- Conclusions and Future Work



Motivation

- Why Align (or Match)?
 - Find similarities between sequences
 - Identify genes and their cellular functions
 - Learn not just what the Genome sequence is, but what it does!



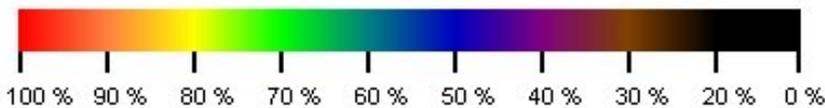
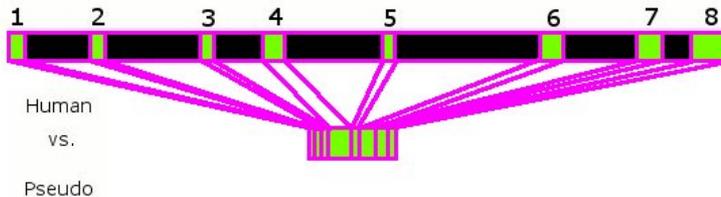
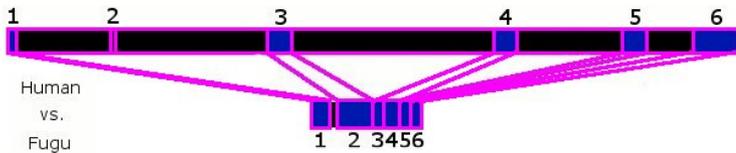


Comparing Fugu vs. Human Genome



- Traditional SWAT (Smith-Waterman) algorithm does not work well, because

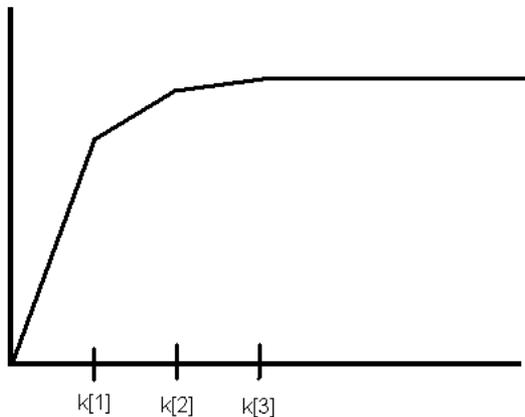
- Gaps do not follow an exponential distribution
- Log likelihood penalty is not “Affine”
- Exons have been conserved, but yet, the homology level is low
- The region to be compared is rather long.
- A more “Global” Alignment is sought.





Piecewise-Linear Approximation of Gap Functions

- Can approximate any Gap Function
- Lets us align faster than most Gap Functions
 - Almost as fast as aligning with Linear Gap Functions



- A non-affine gap-penalty function that models the evolutionary process better
- It approximates a logarithmic function quite well

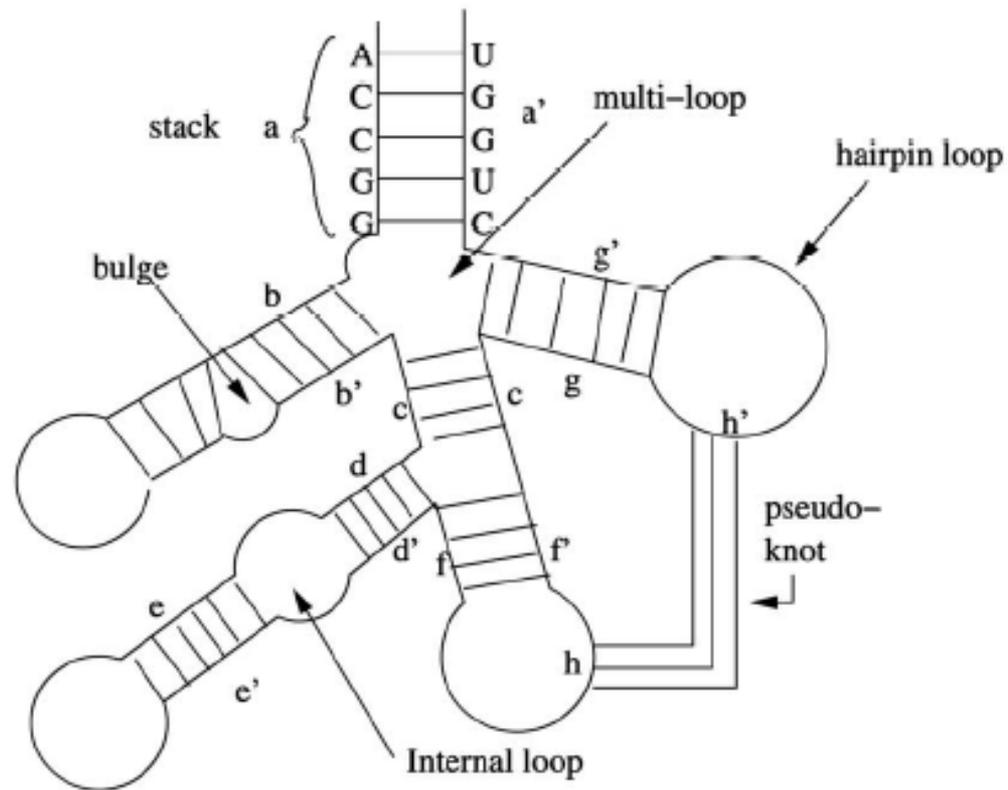


DNA / RNA Alignment

- Normally, sequence similarities in DNA or proteins are used to identify functional correlations
- But for RNA, this is not enough.
 - RNA functionality is also tied to secondary structure



Secondary Structure Example





Motivation

- Given an alignment, how do we measure its accuracy?
 - Which alignments are chance occurrences and which are biologically meaningful?
 - Can we measure “reliability”?

```
58  tAG-AGGTgttcTtTAgcAtA---GaTTTA-ACCTT-CAtTATTTaC--TT 99
    ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
47  cAGtAGGTtaaaTaTAttAaaTgcGtTTTAgACTTaCAaTATTTtCaaTT 96

100  TAGccTGtGaTTCtggATgAgccagctGAAtTcAAcCccagGgAgGcaGA 149
    ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
97  TAGgaTGAg-TTC---ATtA-----aGAcATaAAaCaattGtAaGttGA 136

150  GcTtGcTgTCagTttC-TCtAgtTgGctgGGtCTCTgccTttGctcT-GA 197
    ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
137  G-gaG-TaTCtaTaaCaTCaAtcTtGaacGGaCTCTtttTaaGaaaTaGA 184

198  gTTTggatGC-GGctgC--TGGCTgA-gCtTcaccTtCCAGCAagactAc 243
    ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
185  aTTTaaaggGCcGGacaCagTGGCTcAcaCcTgtaaTcCCAGCActttgAg 234

244  tGGCtggccactGCA-ATCaggcGGaG-CtGctGgcTGgGA--ActgTGA 289
    ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
235  aGGCcaagacggGCagATCAcctGGgGtCaGgaGttTGaGaccAaccTGA 284

290  tt-GCtTctAatAtCtggCagCgtcACaAAgtGTAttTcccTgttgGGGa 338
    ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
285  ccaGCaTggAgaAaCcccCAtCtctACTAAaaGTAcAaatTagcaGGGc 334
```



p-Value

- Computing p-values for “important” segments of an alignment
 - These are segments with higher similarities and scores
- p-value denotes the probability a segment is coincidental
 - If segment has score s , the p-Value is denoted as $\Pr(x \geq s)$
 - x is the score of an arbitrary segment
 - p-Value is contrasted to Null Hypothesis
 - If segment comes from the Null Hypothesis, its p-Value should be > 0.5 (most certainly coincidental)



Outline

- Introduction
- **PLAINS (for DNA Alignment)**
- PLANAR (for RNA Alignment)
- SEPA (for Alignment Evaluation)
- Colorgrids (for Alignment Visualization)
- Results
- Conclusions and Future Work



PLAINS

- **Piecewise Linear Alignment with Important Nucleotide Seeker**
 - Pure DP-based algorithm over DNA
 - Miller-Meyers reduction (+)
 - Linear-space worst-case(*) and memory efficient
 - Species customization

(+) Miller-Myers, 1988.



Outline

- Introduction
- PLAINS (for DNA Alignment)
- **PLANAR (for RNA Alignment)**
- SEPA (for Alignment Evaluation)
- Results
- Conclusions and Future Work



PLANAR

- **Piecewise Linear Alignment for Nucleotides Arranged as RNA**
 - Pure DP-based Algorithm over RNA
 - Efficient like Single Secondary Structure Algorithms
 - Adjusts Alignments to Account for Both Secondary Structures (*)
 - CMSAA reduction (+)
 - Similar to Miller-Meyers, except for RNA
 - Species customization

(+) Eddy 2002.



PLANAR

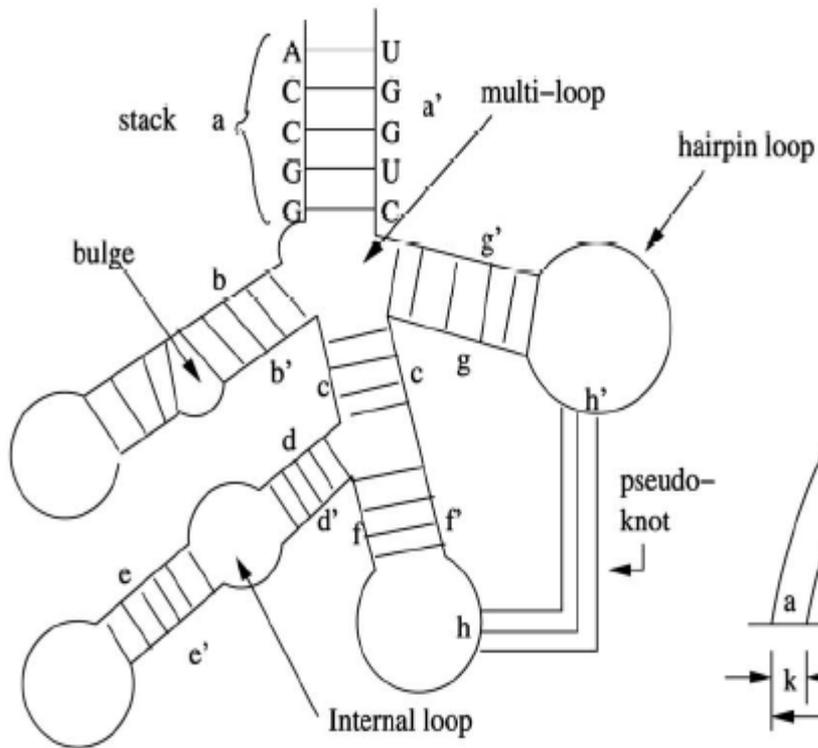
■ Strengths

- Biological consistency
- Secondary structure consistency
- Identifies key correlations

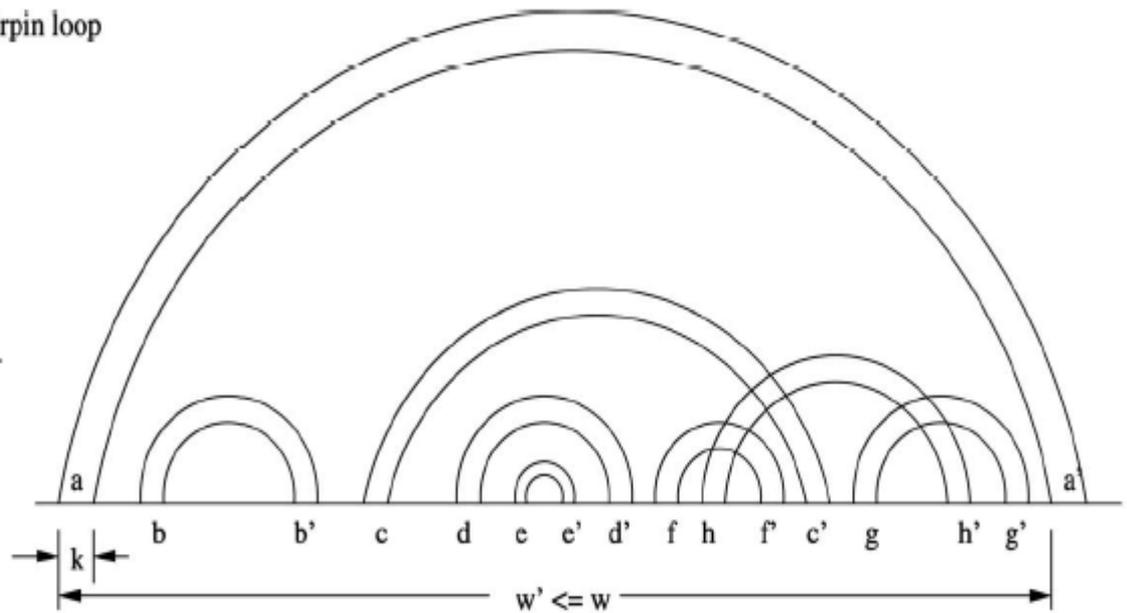
■ Weaknesses

- Speed
- Calibration techniques need a theoretical justification

Secondary Structure Unfolding



(a)

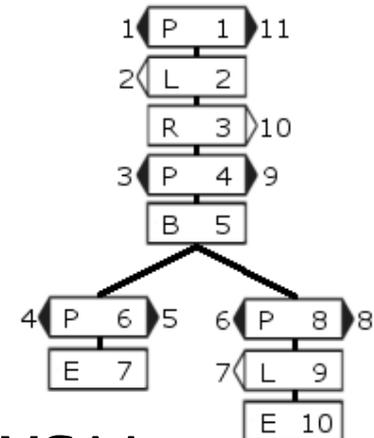
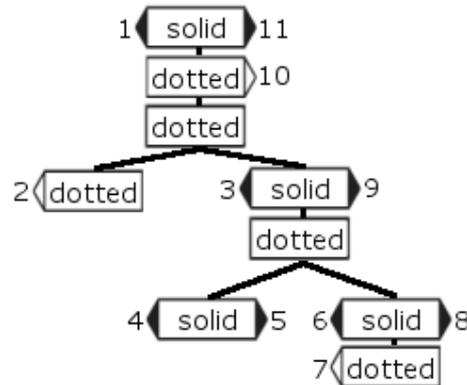
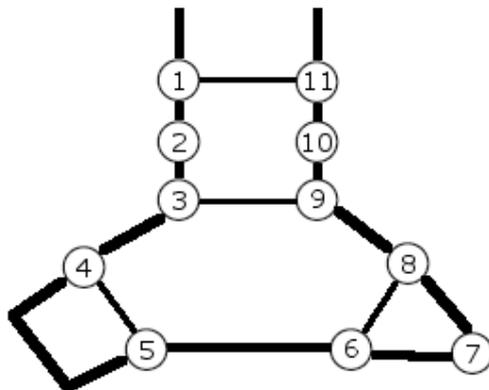


(b)



Binarization

- Convert a given secondary structure into a tree.
- Different Binarization algorithms give different trees for the same structure.



(+) Zhang-Haas-Eskin-Bafna, 2005.

FastR(+)

CMSAA



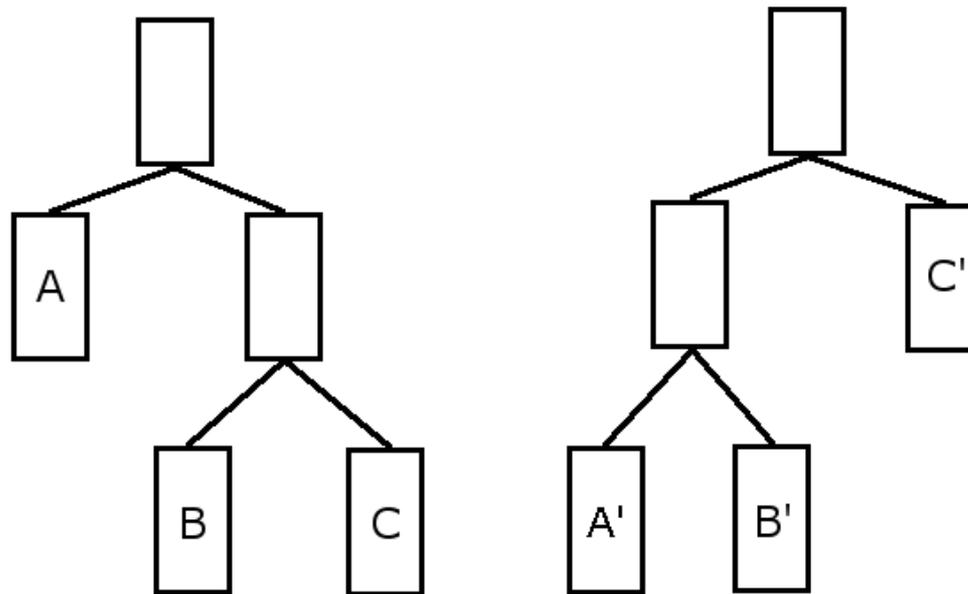
Binarization

- We ignore pseudoknots in unwinding RNA
 - Pseudoknots slowdown runtime, but do not affect the final results drastically
- “Bulking” adjacent nucleotides of a hairpin into the same linear chain is helpful because:
 - Intuitive conceptualization
 - Fewer bifurcations Faster runtime
 - Allows simpler implementation of length-dependent gap functions
 - Allows for “reduced” gap penalties at bound positions



Secondary Structures

- Drawback to considering two secondary structures at a time:





Node Labeling for $u \in T_x$

- 'L' for Left-Character Only
- 'R' for Right-Character Only
- 'P' for Paired Position
 - Bound Position with both Left and Right Characters
- 'B' for Bifurcation
- 'E' for Endpoint (Leaf Node)
 - Serves as Base-Case in Alignment



PLANAR Alignment Formulation (*)

- If u 's label is 'E':
 - $V(u, i, j) = w(j - i + 1)$
- If $i > j$:
 - $V(u, i, j) = w(|u|)$
- If u 's label is 'B':
 - $V(u, i, j) = \max_{i-1 \cdot k \cdot j} [V(u.\text{left}, i, k) - w(u.\text{right}, k+1, j)]$
- If u 's label is not 'B':
 - $V(u, i, j) = \max\{D(u, i, j), E(u, i, j), F(u, i, j), G(u, i, j)\}$
 - $D(u, i, j) = \max_{i+1 \cdot k \cdot j+1} [V(u, k, j) - w(k-i)]$
 - $E(u, i, j) = \max_{i-1 \cdot k \cdot j-1} [V(u, i, k) - w(j-k)]$
 - $F(u, i, j) = \max_{t \text{ s.t. } \text{LCB}(t,u)} [V(t, i, j) - w(|u| - |t|)]$



PLANAR Alignment Formulation

- If u 's label is 'L':
 - $G(u, i, j) = V(u.child, i+1, j) + s(X[l_u], Y[i])$
- If u 's label is 'R':
 - $G(u, i, j) = V(u.child, i, j-1) + s(X[r_u], Y[j])$
- If u 's label is 'P' and $i < j$:
 - $G(u, i, j) = V(u.child, i+1, j-1) + b(X[l_u], X[r_u], Y[i], Y[j])$
- Otherwise:
 - $G(u, i, j) = -1$
- Space Reduction in this table using CMSAA's Generic Splitter
 - Identical to Hirschberg, except we "split" at halfpoints of linear chains and bifurcations in T_x .
 - Cubic runtime and quadratic space.



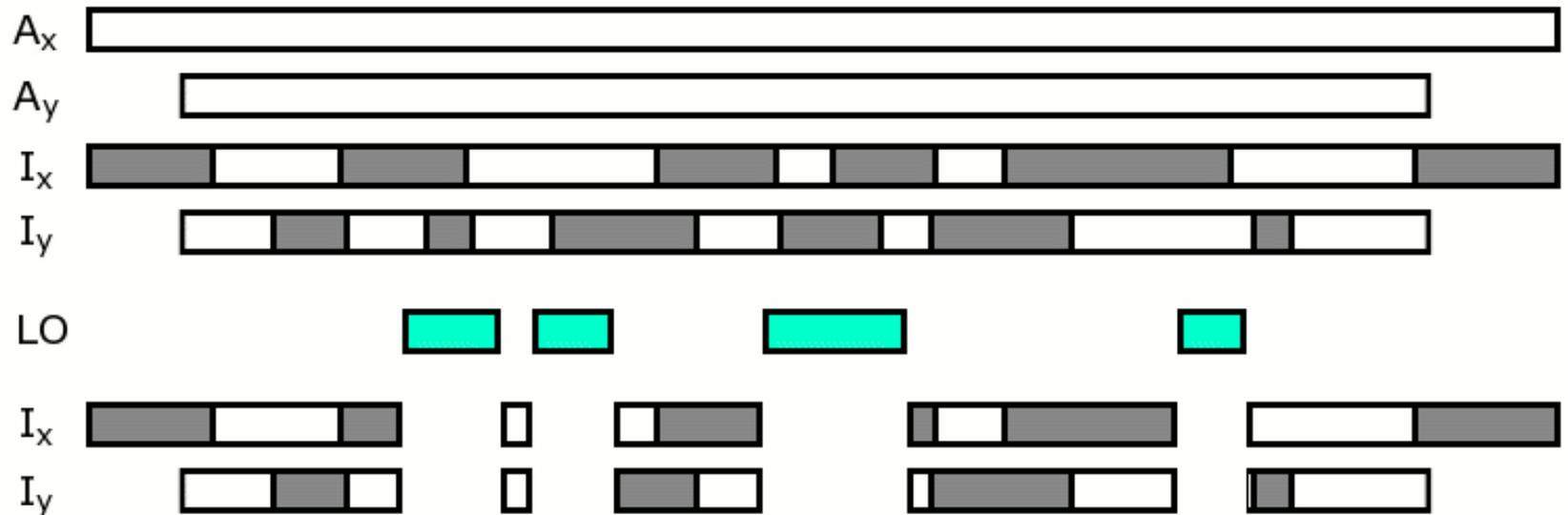
Double Secondary Structure Correction (*)

- We align T_x to Y to get an alignment A_x
- We align T_y to X to get an alignment A_y
- Given A_x and A_y , our goal is to get the final result A .
- We want in A :
 - Segments that A_x and A_y have in common
 - Non-overlapping segments of A_x and A_y with exceptionally high similarities.



Double Secondary Structure Correction

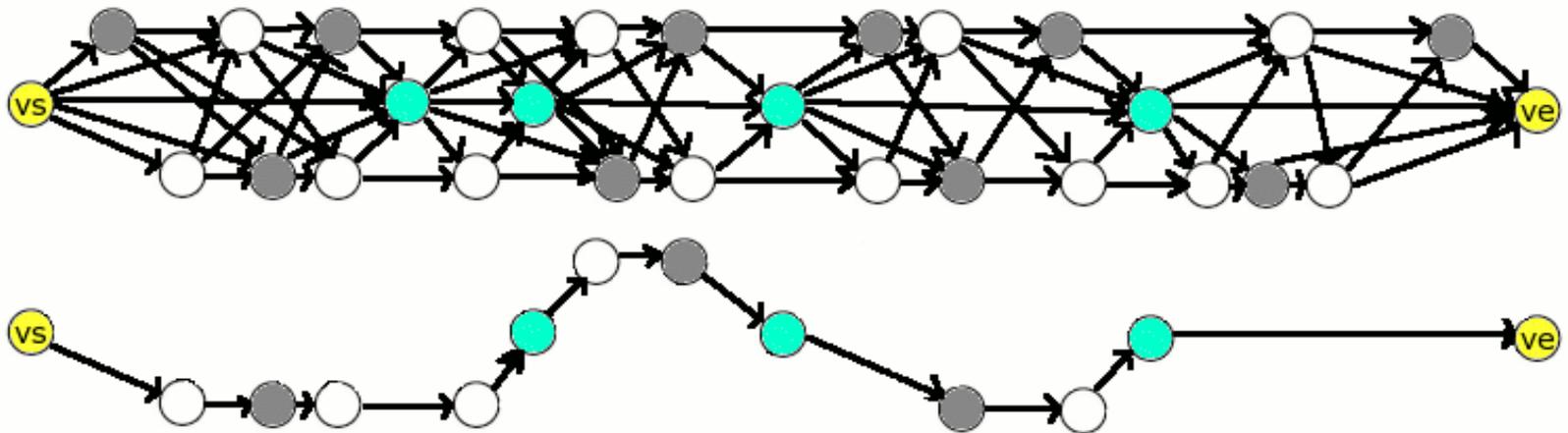
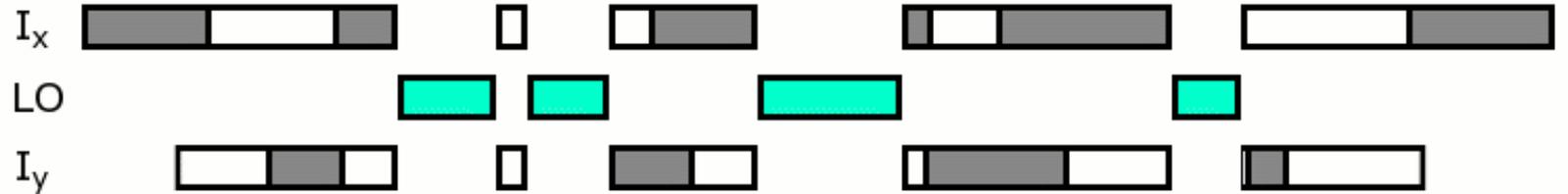
- Merging A_x and A_y to make A . (Part 1)





Double Secondary Structure Correction

■ Merging A_x and A_y to make A . (Part 2)



A





Learning Penalty Parameters

- The match/mismatch/gap parameters are dictated by five variables (α , β , d , m_s , m_b)
 - Parameters are identical to PLAINS, except for the introduction of m_b (the “extra reward” for bound position match)
 - Parameter-Optimization is identical to that of PLAINS, except taking slightly longer due to longer time for each alignment. (Cubic vs. Quadratic, and SS Corrections)
 - Empirical evidence shows species customizations from parameters work here too.



Outline

- Introduction
- PLAINS (for DNA Alignment)
- PLANAR (for RNA Alignment)
- **SEPA (for Alignment Evaluation)**
- Results
- Conclusions and Future Work



SEPA

- **Segment Evaluator for Pairwise Alignments**
 - Can evaluate any alignment, not just PLAINS or PLANAR.
 - Identifies important segments from any alignment, regardless of homology levels
 - Assigns p-Values (that is $P(x, s)$) to each segment
 - Assigns ζ value for coincidental probability of all important segments identified. This acts as a single “alignment measure”
 - Compares against a Null Hypothesis, based on Unrelated Sequences Calibration
 - Identifies Non-obvious Correlations in Sequences



SEPA

■ Strengths

- Estimations based on thorough segment behavioral analysis for Null Hypothesis
- Regardless of similarities, we catch:
 - Important segments, exon regions, and unknown correlations
- Estimation successfully identifies segments from random DNA alignments as “coincidental”

■ Weaknesses

- ζ value is overly sensitive to the number of segments identified
- Estimation has little theoretical justification
- Estimation does not yet account for secondary structures in evaluating RNA alignments



Methodology(*)

- We score each possible segment of length W .
- We compute average μ and deviation σ for the scores.
- Any segment scoring above $\mu + \omega\sigma$ is marked as important
- We trim segments to start/end with a match
- We merge overlapping segments and score them, and do our p-Value estimation
- If necessary, we remove segments with p-Value higher than ρ



Analyzing Segments(*)

- For each thousand-length from 1000 to 8000, we generated 25 random sequences.
 - We also generated 25 random sequences of length 500
- For all combinations of length pairs, we used PLAINS to generate 625 possible alignments, analyzing with SEPA length-dependent behavior
 - No ρ filtering was used here



Outline

- Introduction
- PLAINS (for DNA Alignment)
- PLANAR (for RNA Alignment)
- SEPA (for Alignment Evaluation)
- **Results**
- Conclusions and Future Work



RNA Alignment Tools Compared

- RSMATCH(+)
 - Assumes input is generic
 - Uses pure DP algorithm based on SS loops
 - Aligns using SS of both sequences
 - Uses linear gap penalty
 - Fastest pure-DP algorithm for RNA

(+) Liu-Wang-Hu-Tian, 2005.



PLANAR vs. RSMATCH

Test Name	PLANAR			RSMATCH		
	t	r	ζ'	t	r	ζ'
rnase.1_2	204.67	3	3.37	118.58	1	5.99
rnase.1_3	134.35	2	4.53	85.55	1	5.99
rnase.3_5	120.76	2	4.51	42.88	1	5.98
rnase.4_5	104.17	1	5.98	93.08	2	4.49
telomerase.1_2	29.58	2	4.47	14.89	2	4.47
telomerase.1_3	79.23	2	4.48	24.67	2	4.48
telomerase.2_3	17.06	1	6.10	2.60	1	6.10



Discussion

- PLANAR does not always have the highest ζ'
 - The nature of piecewise-linear gap functions is to incorporate as many regions as possible
 - Esp. when sequences have high expected gap and low homology regions
 - This process raises the r , hence penalizing ζ'
 - However, if r is fixed, their t (and hence ζ') is stronger.
 - This is because the PLAINS and PLANAR results have higher homologies in most of the important segments identified by SEPA.



Outline

- Introduction
- PLAINS (for DNA Alignment)
- PLANAR (for RNA Alignment)
- SEPA (for Alignment Evaluation)
- Results
- **Conclusions and Future Work**



Conclusion

- PLAINS and PLANAR show promise because:
 - They can run on a single regular PC
 - Although they identify “too many” important segments, for fixed r , their segments are stronger
 - They show promise of identifying unknown correlations
 - Parameters are user-adjustable, and optimization techniques require no user-knowledge.
- SEPA shows promise because:
 - Its estimation method distinguishes important regions from unimportant ones
 - It models p -Values for DNA accurately



Future Work

- Possible improvements to PLANAR include:
 - Speeding up the DP methods
 - Learn expected alignments to various species, instead of just approximating parameters
 - Refine the results of locally identified interval regions for global alignments
 - Use scoring matrix for scoring certain letters (instead of pure match/mismatch model)



Bibliography

- Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N., “Combinatorial microRNA target predictions.” *Nature Genetics*, 37(5): 495--500, 2005.
- Miller, W., and Myers E.W., “Sequence Comparison with Concave Weighting Functions,” *Bulletin of Mathematical Biology*, 50:97--120, 1988.
- Miller, W., and Myers E.W., “Optimal Alignments in Linear Space,” *CABIOS*, 4:11--17, 1988.
- Hromkovic J, “Heuristics.” *Algorithms for Hard Problems, Second Edition*, 6:439-467, 2003.
- Eddy S.R., “A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure.” *BMC Bioinformatics*, 3:18, 2002.
- Zhang S, Haas B, Eskin E, Bafna V, “Searching Genomes for Noncoding RNA Using FastR.” *IEEE/ACM Trans. on Comp. Bio. and Bioinf.*, 2(4): 366--379, 2005.



Bibliography (contd.)

- Karlin S, Altschul S.F., “Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes,” Proc. Natl. Acad. Sci. USA, 87:2264--2268, March 1990.
- Karlin S, Altschul S.F., “Applications and statistics for multiple high-scoring segments in molecular sequences,” Proc. Natl. Acad. Sci. USA, 90:5873--5877, June 1993.
- Siegmund, D., Yakir, B.: “Approximate p-Values for Local Sequence Alignments,” The Annals of Statistics, 28 (3) (2000) 657--680
- Rice P, Longden I, Bleasby A., “EMBOSS: the European Molecular Biology Open Software Suite,” Trends Genetics, Jun 16(6):276-7, 2000.
- Michael Brudno, Chuong Do, Gregory Cooper, Michael F. Kim, Eugene Davydov, Eric D. Green, Arend Sidow, Serafim Batzoglou, “LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA,” Genome Research, 13(4):721-31, 2003 Apr.
- Liu J, Wang JTL, Hu J, Tian B, “A method for aligning RNA secondary structures and its application to RNA motif detection,” BMC Bioinformatics, 6:89, 2005.