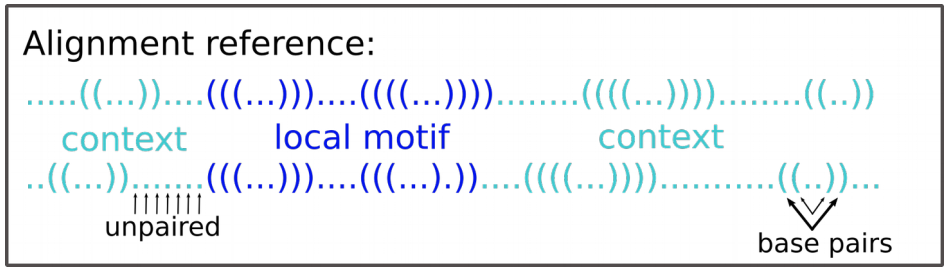
Structure-based RNA alignment: the trouble with locality

Teresa Müller

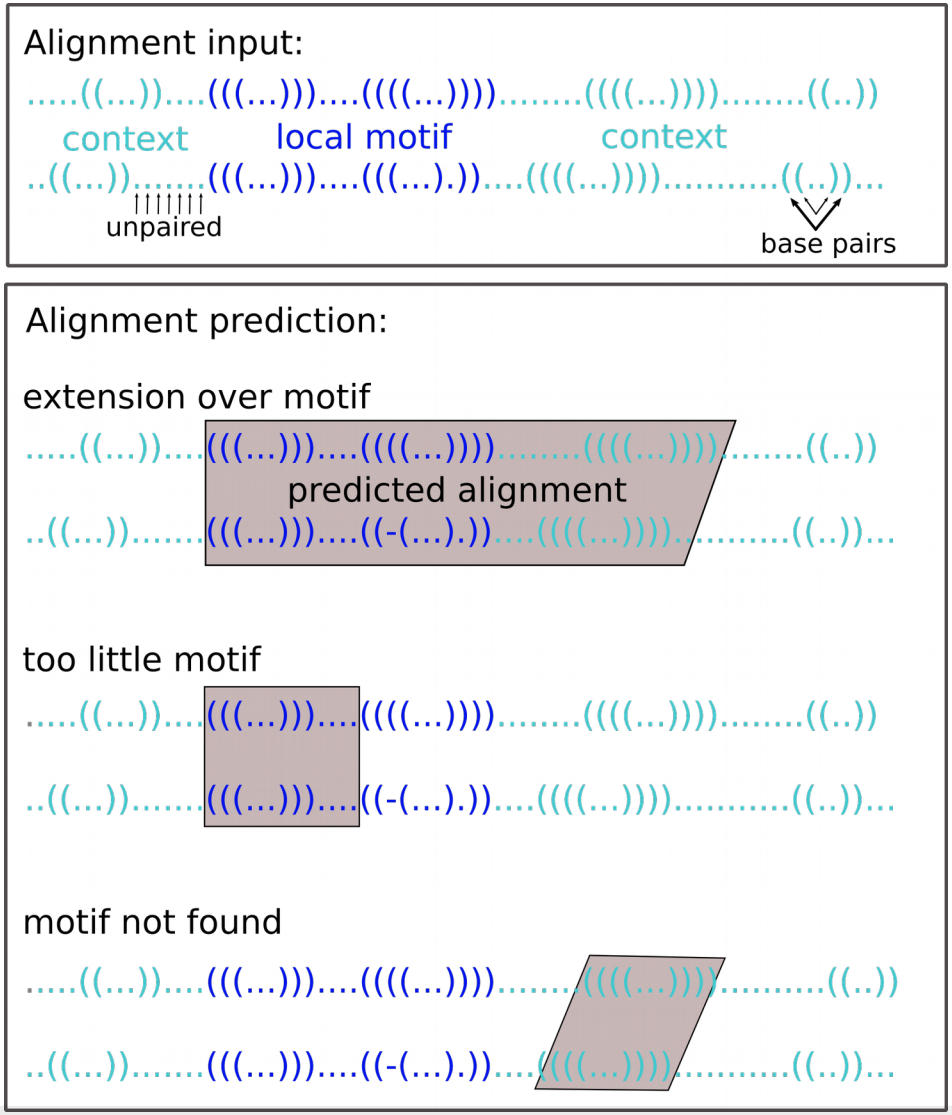
Albert-Ludwigs-Universität Freiburg

UNI FREIBURG

Local alignment: What could go wrong?



Local alignment: What could go wrong?



Local alignment

- Sequence alignment
  - Assumption of negative expected score for alignments of ungapped random sequences holds
  - No length dependency

Local alignment

- Sequence alignment
  - Assumption of negative expected score for alignments of ungapped random sequences holds
  - No length dependency
- Sankoff-like sequence-structure alignment
  - Sequence scoring scheme can easily be transformed
  - Structure scoring uses free energy (Zuker) or base pair probability (McCaskill)

Local alignment

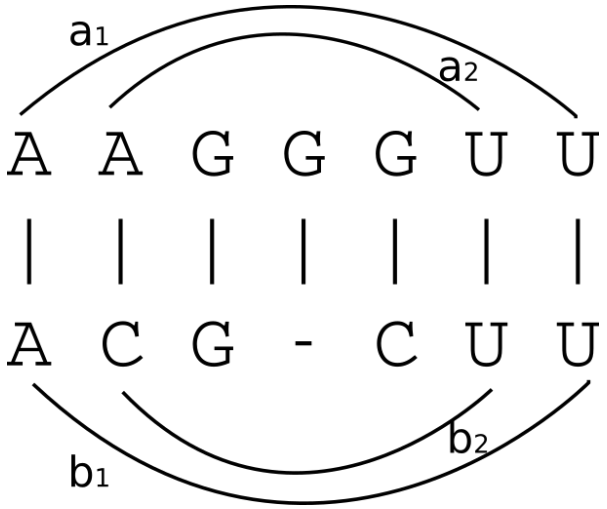
- Sequence alignment
  - Assumption of negative expected score for alignments of ungapped random sequences holds
  - No length dependency
- Sankoff-like sequence-structure alignment
  - Sequence scoring scheme can easily be transformed
  - Structure scoring uses free energy (Zuker) or base pair probability (McCaskill)
- Trouble: structure creates a positive scoring bias in the objective function

Sequence-structure Sankoff-like alignment: LocARNA

- Objective function:

$$\sum_{(ij;kl)\in S} \left(\omega(\Psi^a_{ij} + \Psi^b_{kl}) + \tau\sigma'(a_i, a_j, b_k, b_l)\right) + \sum_{(i,k)\in A_s} \sigma(a_i, b_k) - N_{gap}\gamma - N^o_{gap}\beta$$

- Structure weight ω
- Tau factor τ
- Base pair score ψ
- Ribosome scoring σ
- Gap extension γ
- Gap opening β

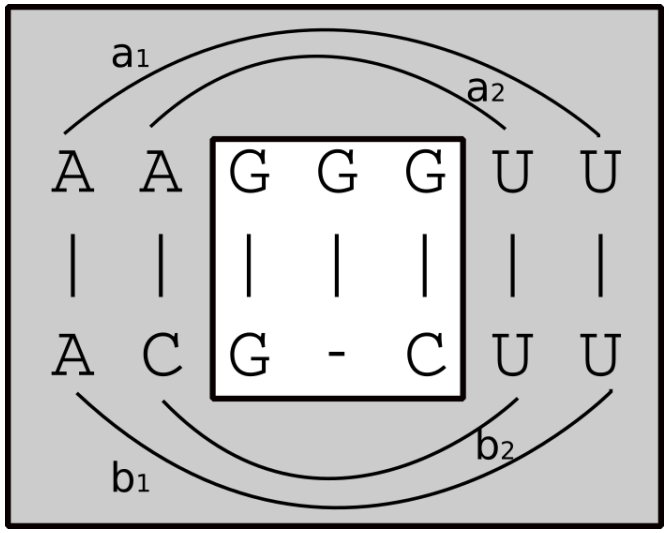


Sequence-structure Sankoff-like alignment: LocARNA

- Objective function:

$$\underbrace{\sum_{(ij;kl)\in S} \left(\omega(\Psi^a_{ij} + \Psi^b_{kl}) + \tau\sigma'(a_i, a_j, b_k, b_l)\right)}_{\text{Structure contribution}} + \sum_{(i,k)\in A_s} \sigma(a_i, b_k) - N_{gap}\gamma - N^o_{gap}\beta$$

- Structure weight ω
- Tau factor τ
- Base pair score ψ
- Ribosome scoring σ
- Gap extension γ
- Gap opening β

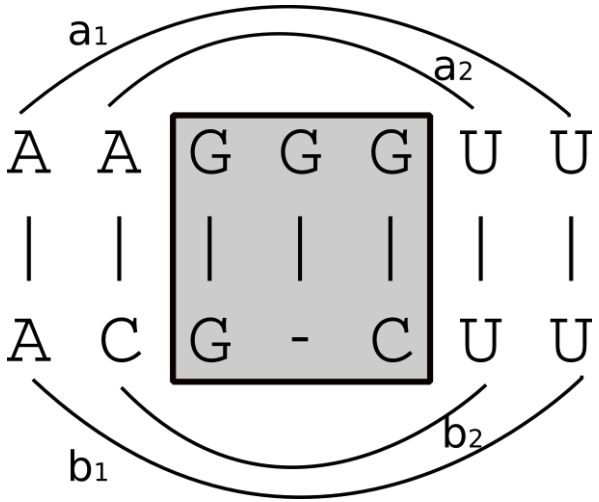


Sequence-structure Sankoff-like alignment: LocARNA

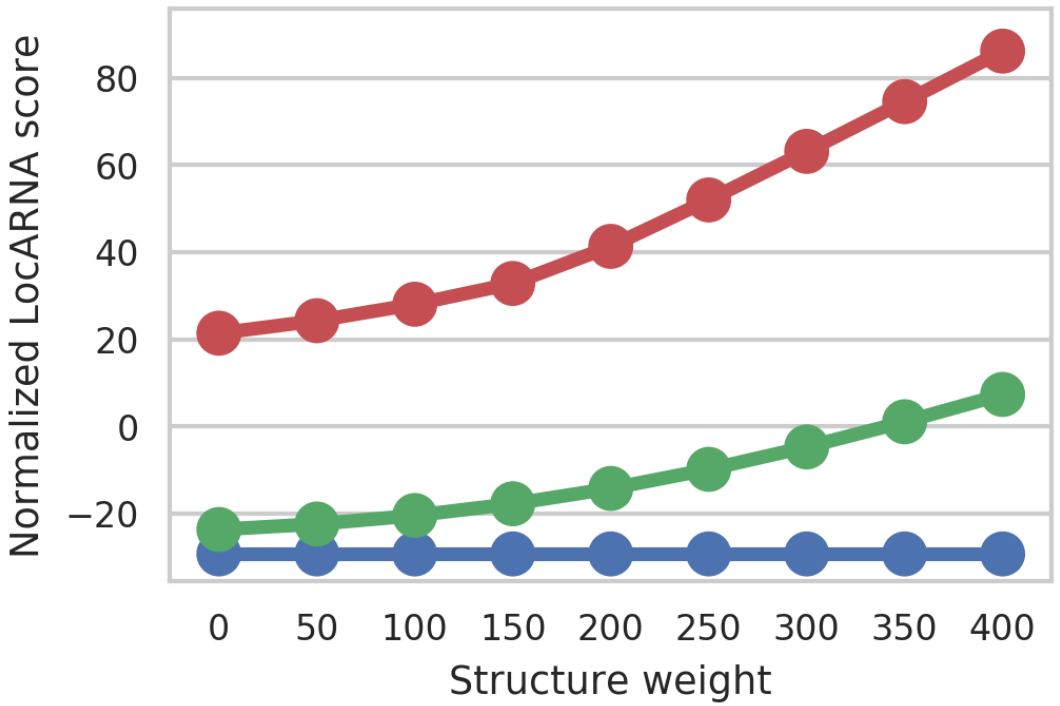
- Objective function:

$$\sum_{(ij;kl)\in S} \left(\omega(\Psi^a_{ij} + \Psi^b_{kl}) + \tau\sigma'(a_i, a_j, b_k, b_l)\right) + \underbrace{\sum_{(i,k)\in A_s} \sigma(a_i, b_k) - N_{gap}\gamma - N^o_{gap}\beta}_{\text{Sequence contribution}}$$

- Structure weight ω
- Tau factor τ
- Base pair score ψ
- Ribosome scoring σ
- Gap extension γ
- Gap opening β

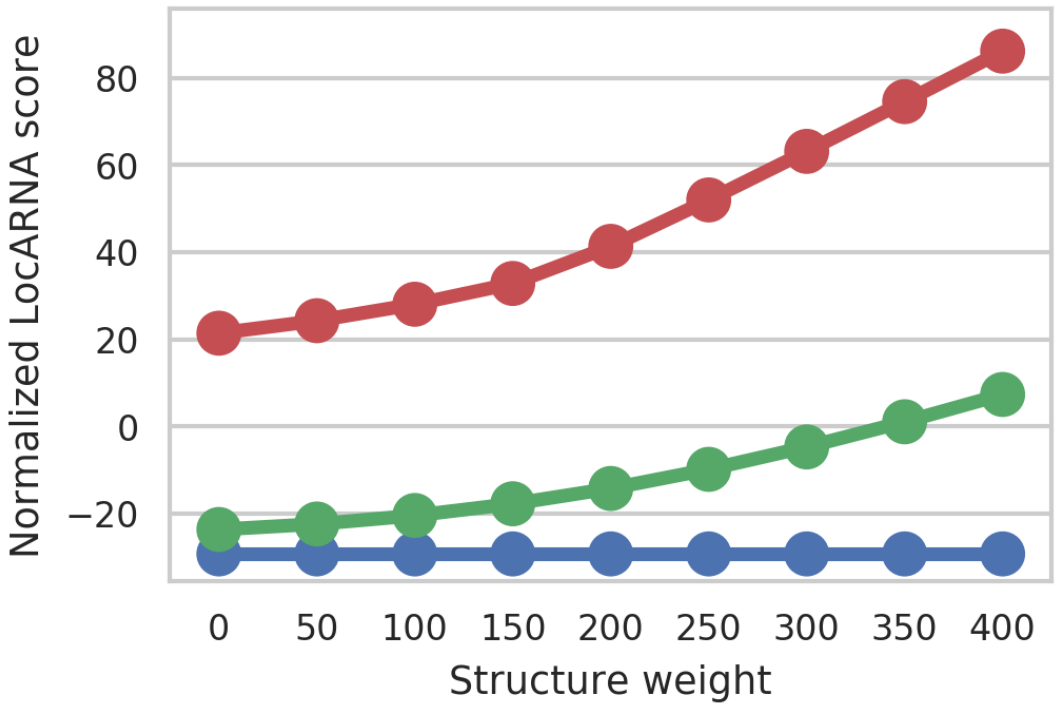


Positive structure scoring bias



- Normalization: score / sum of sequence length

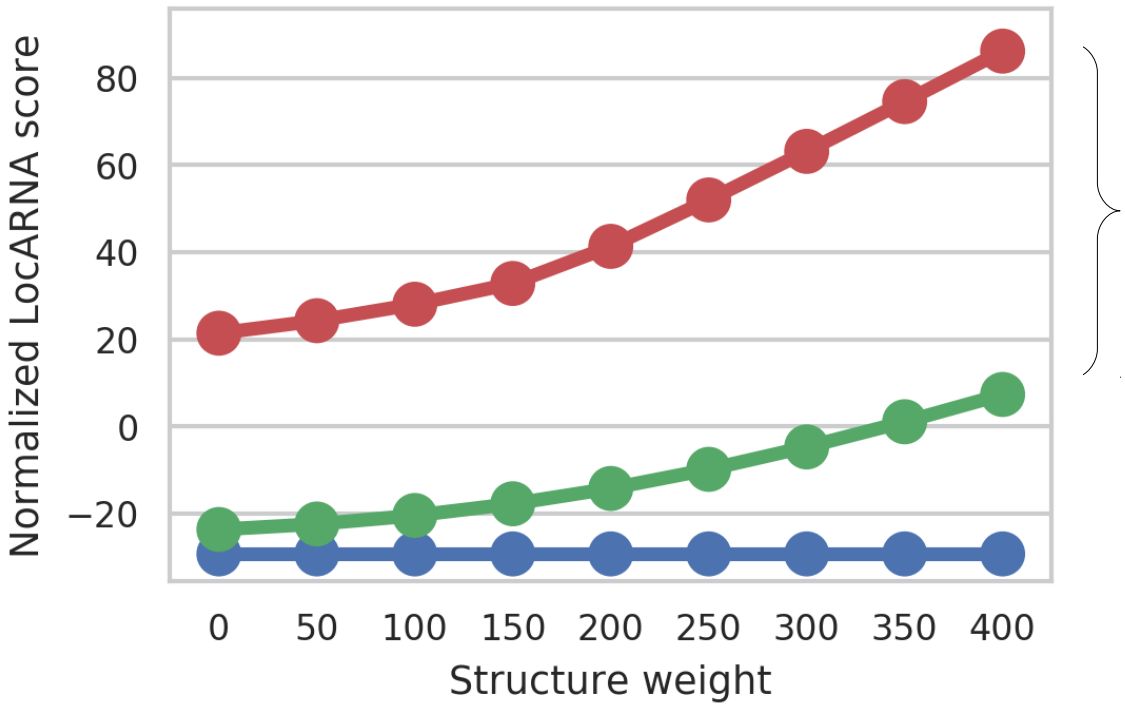
Positive structure scoring bias



	Data	Scoring
●	Bralibase	Sequence structure
●	Shuffled Bralibase	Sequence structure
●	Shuffled Bralibase	Sequence-only

- Normalization: score / sum of sequence length

Positive structure scoring bias

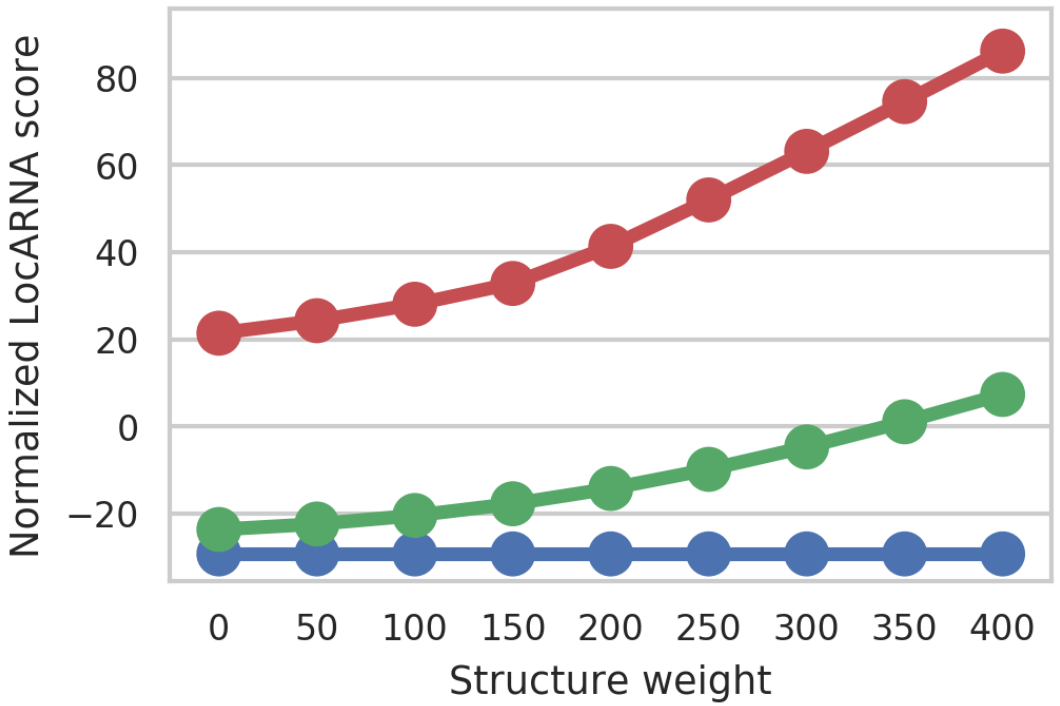


- Clear distinction between ncRNAs and context

	Data	Scoring
●	Bralibase	Sequence structure
●	Shuffled Bralibase	Sequence structure
●	Shuffled Bralibase	Sequence-only

- Normalization: score / sum of sequence length

Positive structure scoring bias



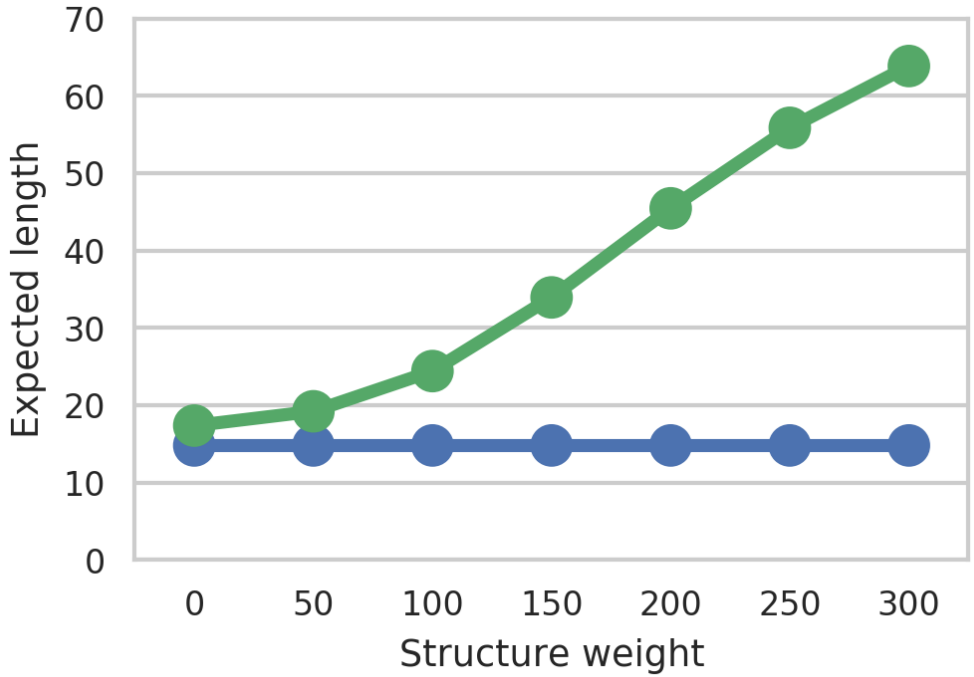
- Clear distinction between ncRNAs and context
- Increased scores already for structure weight 0
- High structure weight positive normalized score

	Data	Scoring
●	Bralibase	Sequence structure
●	Shuffled Bralibase	Sequence structure
●	Shuffled Bralibase	Sequence-only

- Normalization: score / sum of sequence length

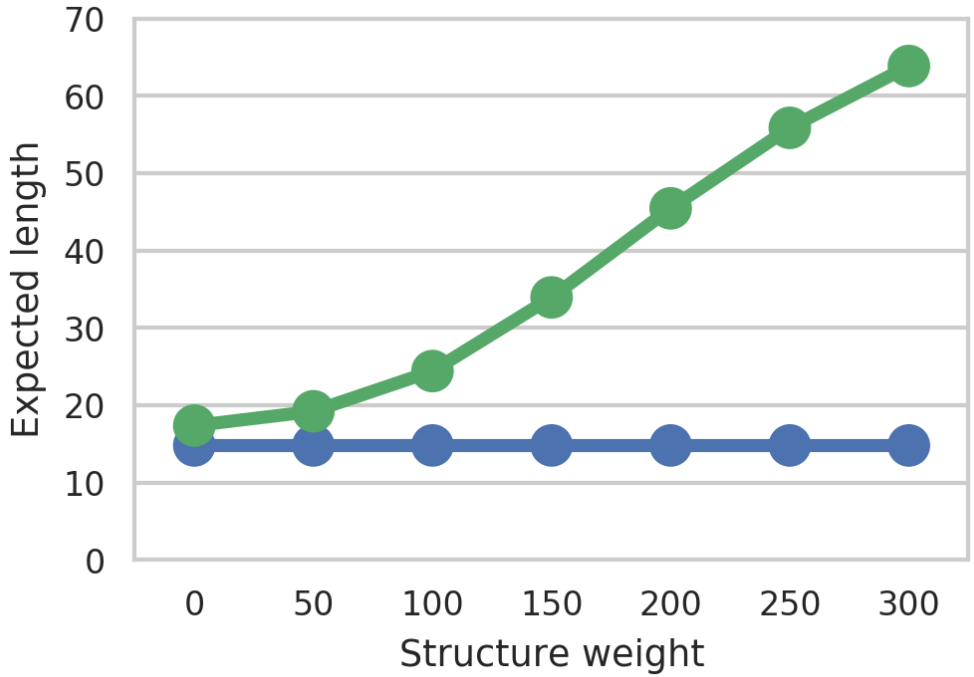
Alignment length growth

Comparing alignment length of random sequences



Alignment length growth

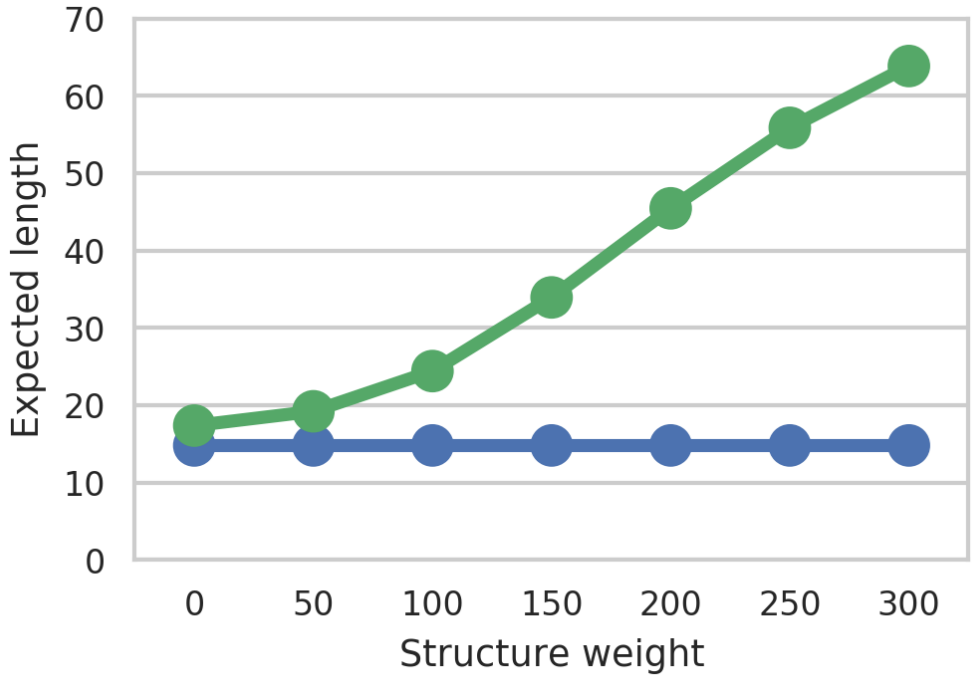
Comparing alignment length of random sequences



	Data	Scoring
●	Random sequences length 100	Sequence structure
●	Random sequences length 100	Sequence-only

Alignment length growth

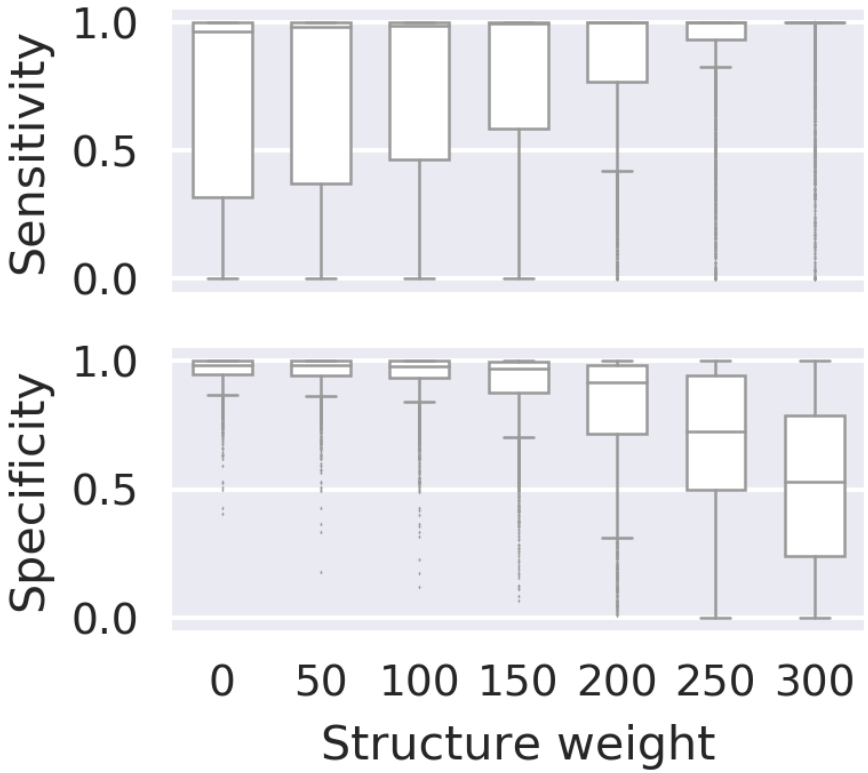
Comparing alignment length of random sequences



- For structure weight 250 > 50 % random sequences aligned
- Using sequence-only < 20 % aligned

	Data	Scoring
●	Random sequences length 100	Sequence structure
●	Random sequences length 100	Sequence-only

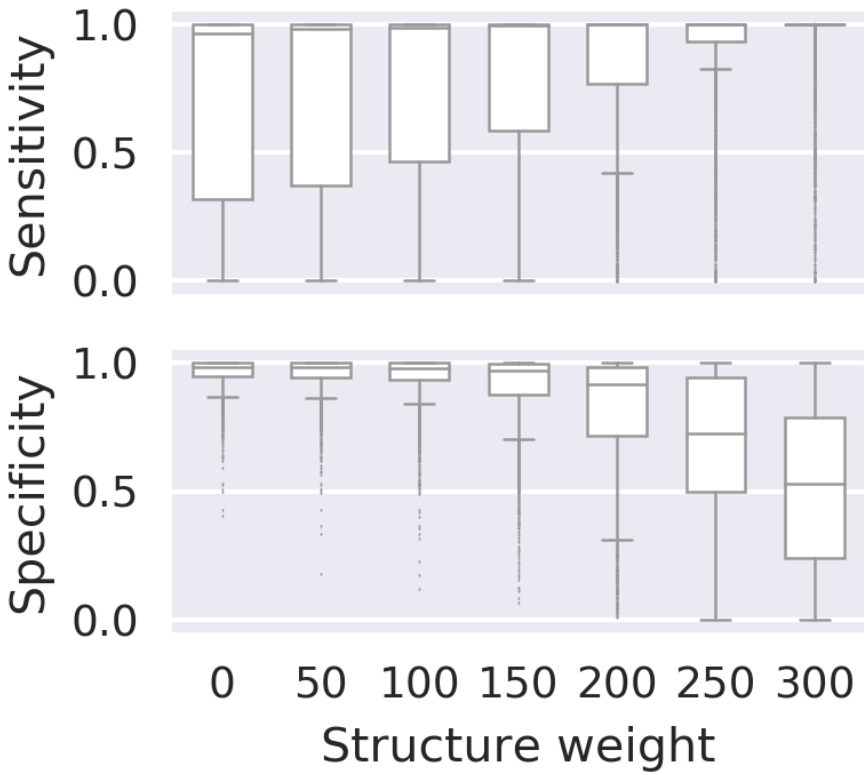
Boundary detection trade-off: structure weight



- Dataset: localBralibase (pairwise alignments)

Boundary detection trade-off: structure weight

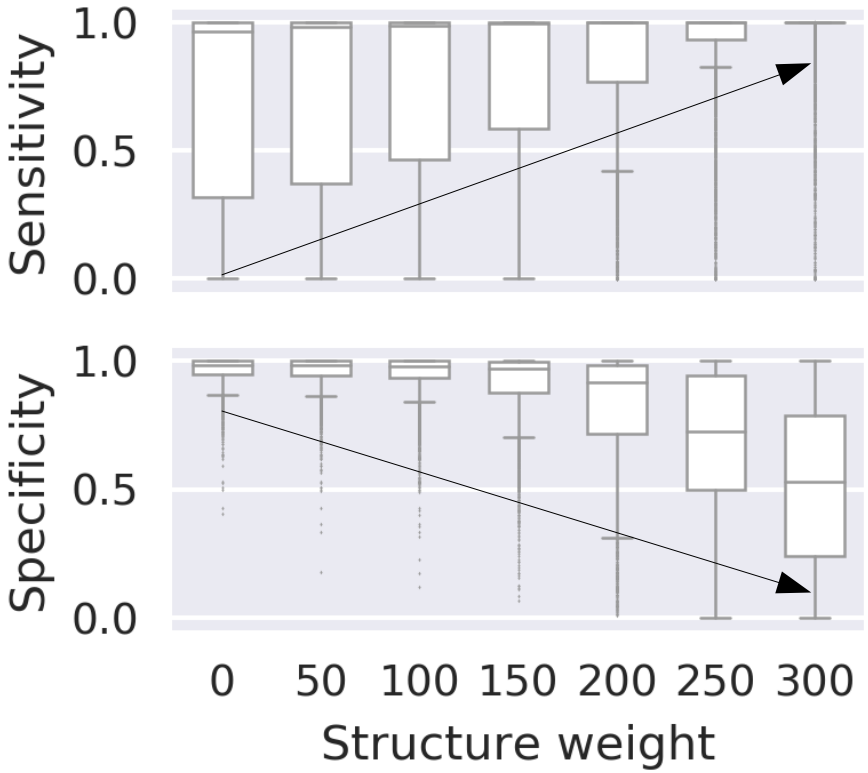
- Sensitivity: how much ncRNA aligned
- Specificity: how much context not aligned



- Dataset: localBralibase (pairwise alignments)

Boundary detection trade-off: structure weight

- Sensitivity: how much ncRNA aligned
- Specificity: how much context not aligned



- Higher structure weight helps covering the complete ncRNA
- Lower structure weight helps to not extend into context

- Dataset: localBralibase (pairwise alignments)

One scoring system (global + local)

- Blackbox parameter optimization:

parameter	Gap γ	Gap opening β	Structure weight ω	Tau factor τ
default	350	500	200	0
Global optimized	68	807	210	72
Local optimized	136	975	115	38
Local optimized (λ 15)	82	883	176	71

One scoring system (global + local)

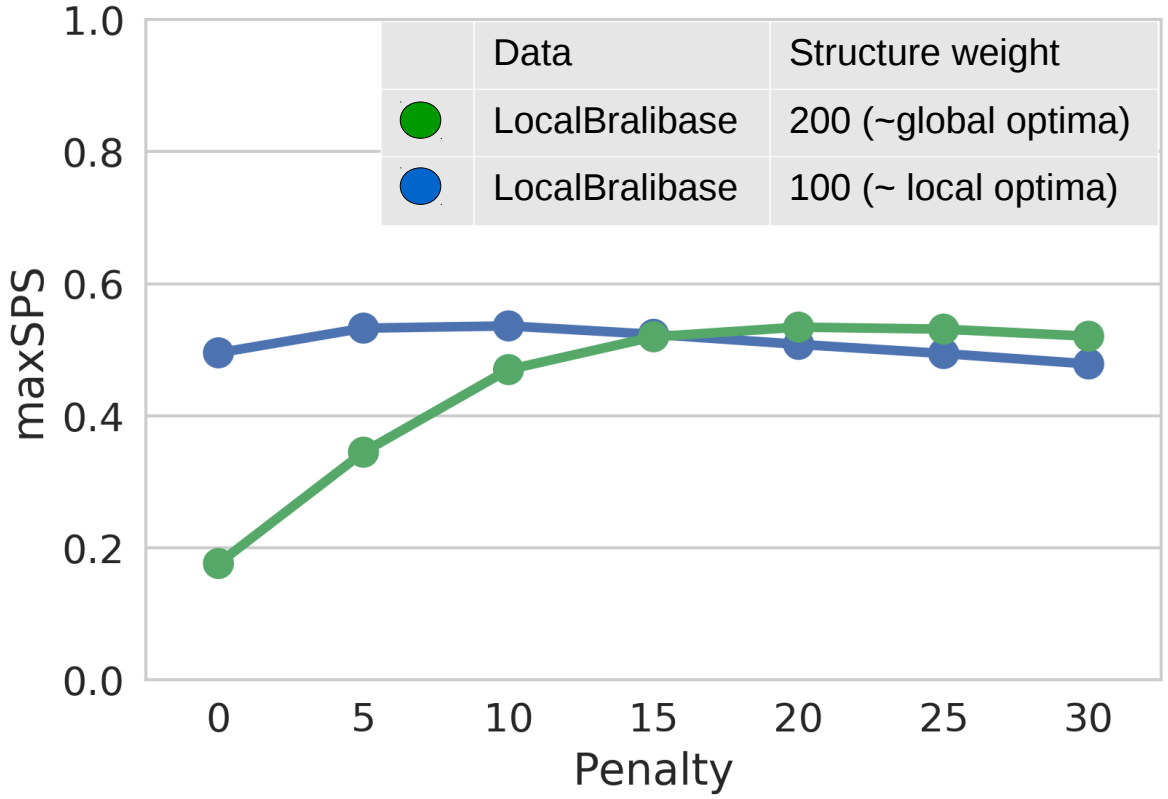
- Blackbox parameter optimization:

parameter	Gap γ	Gap opening β	Structure weight ω	Tau factor τ
default	350	500	200	0
Global optimized	68	807	210	72
Local optimized	136	975	115	38
Local optimized (λ 15)	82	883	176	71

- Solution: position penalty λ
  - Each position of the local alignment is penalized by λ

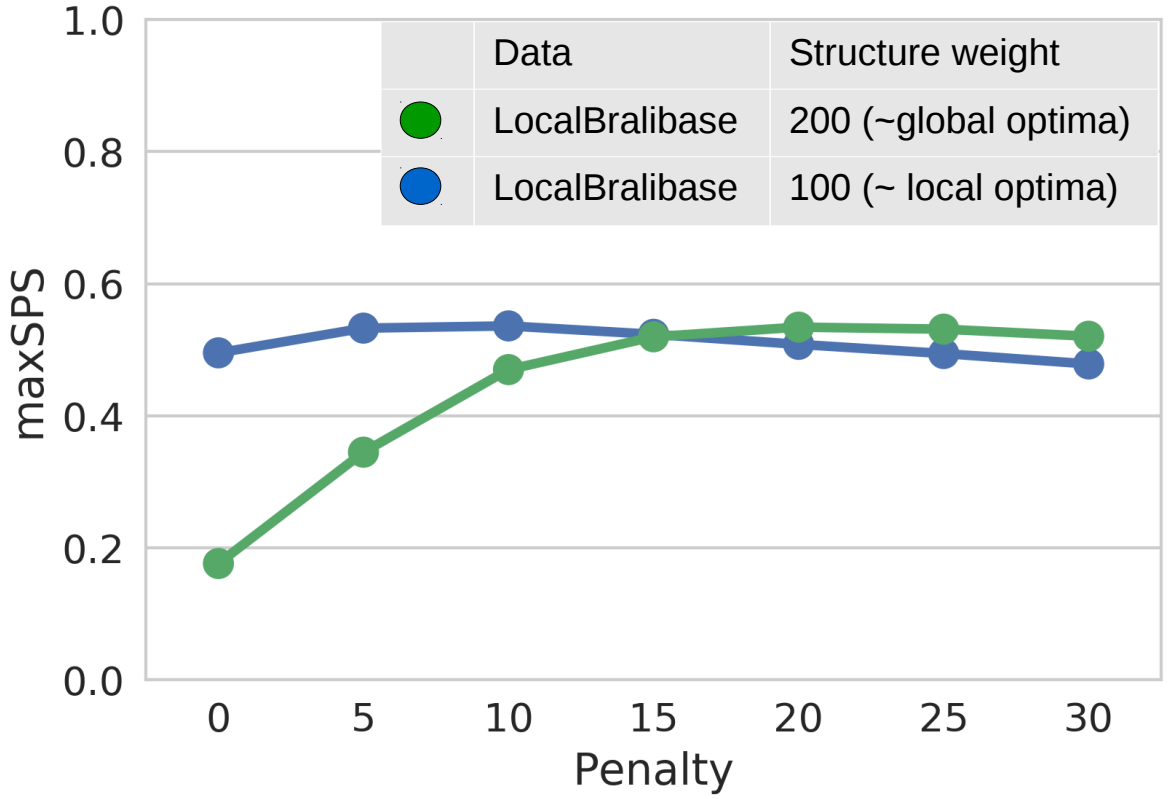
$$\sum_{(ij;kl)\in S} \left(\omega(\Psi^a_{ij}+\Psi^b_{kl})+\tau\sigma'(a_i,a_j,b_k,b_l)-4\lambda\right) + \sum_{(i,k)\in A_s} (\sigma(a_i,b_k)-2\lambda) - N_{gap}(\gamma-\lambda) - N^o_{gap}\beta$$

Structure weight penalty comparison



- maxSPS: alignment quality accounting for context

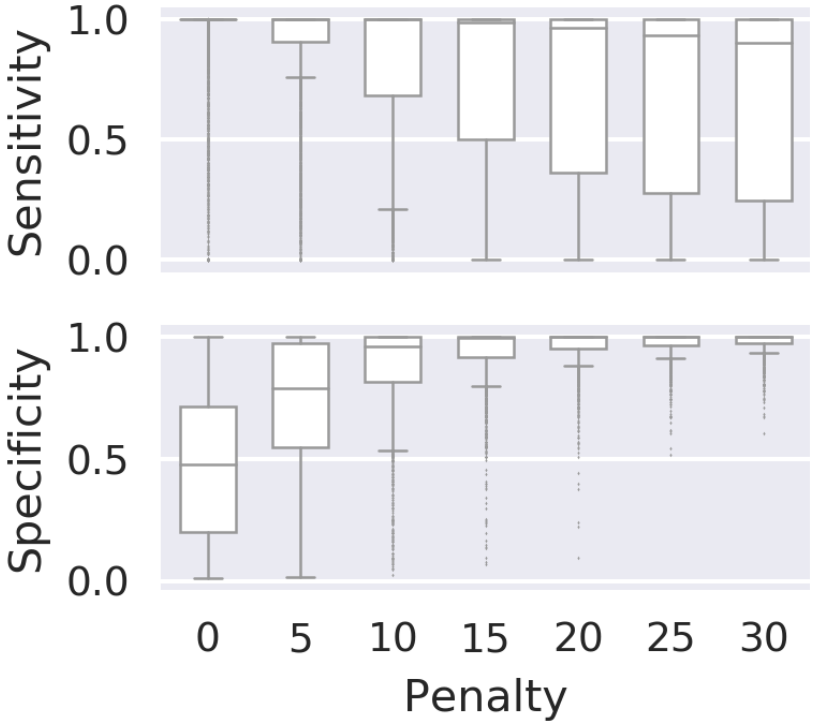
Structure weight penalty comparison



- Low structure weight needs lower penalty
- Higher structure weight needs a higher penalty

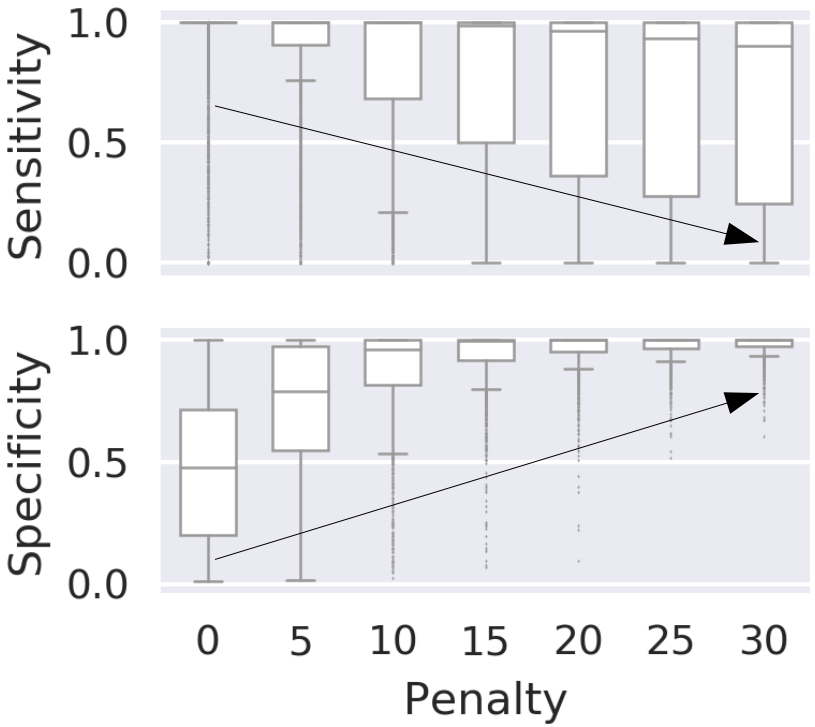
- maxSPS: alignment quality accounting for context

Boundary detection trade-off: position penalty



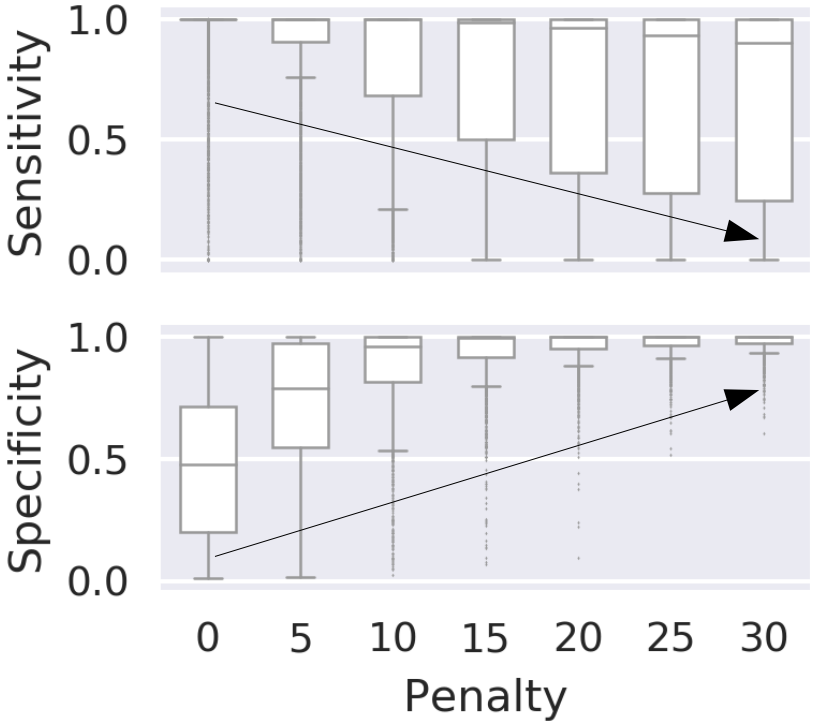
- Structure weight 200

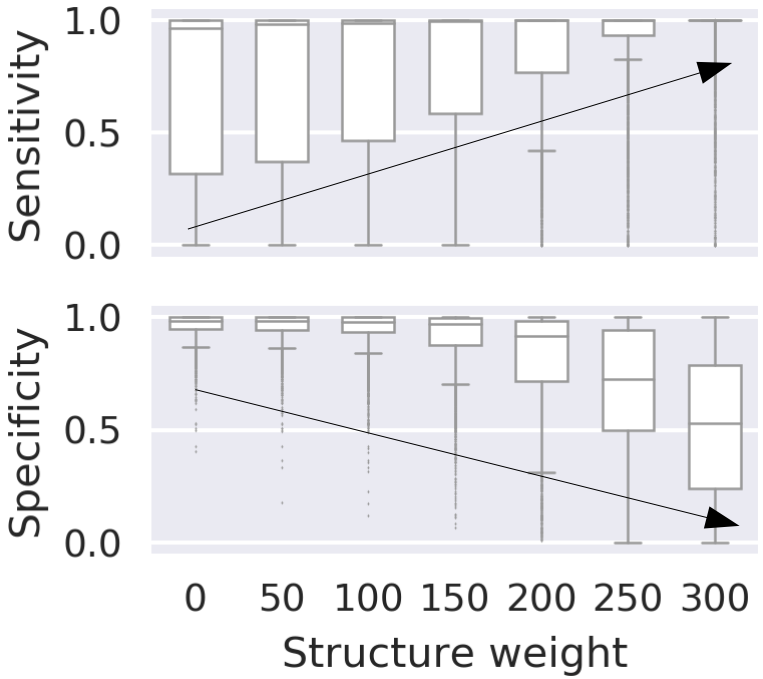
Boundary detection trade-off: position penalty



- Structure weight 200
- Low penalty: ncRNA well covered
- High penalty: context not covered

Boundary detection trade-off: position penalty

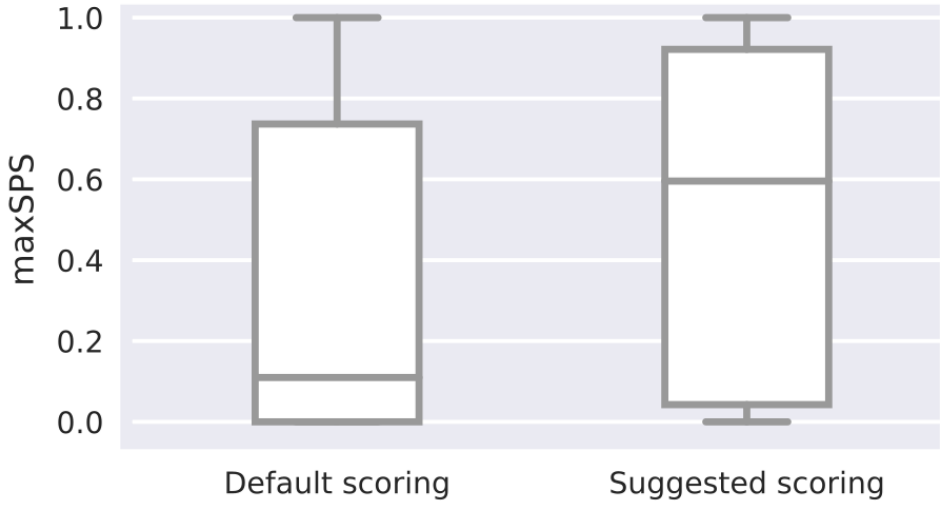




- Structure weight 200
- Low penalty: ncRNA well covered
- High penalty: context not covered

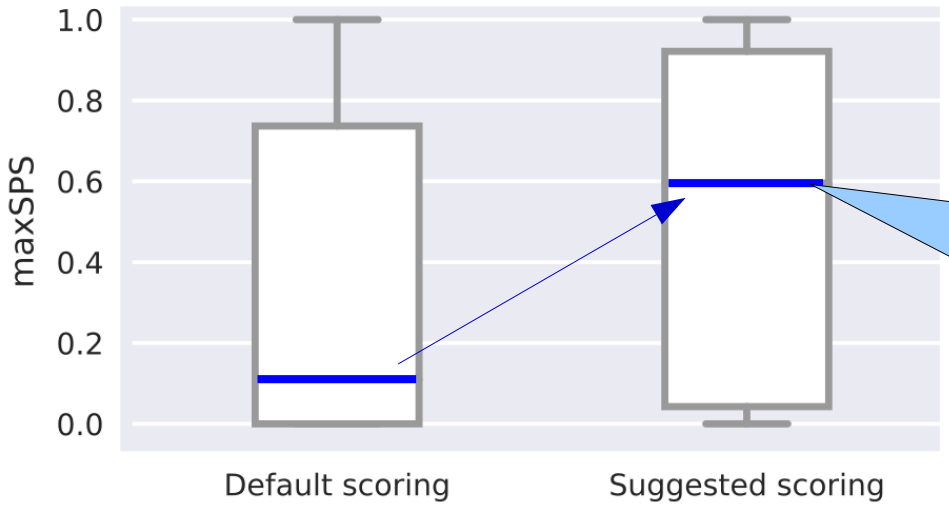
- Remember: Structure weight
- Penalty compensates for context extension

Improved local alignment prediction



Parameter	Gap γ	Gap opening β	Structure weight ω	Tau factor τ	Penalty λ
Default scoring	350	500	200	0	no
Suggested scoring	68	807	210	72	15

Improved local alignment prediction

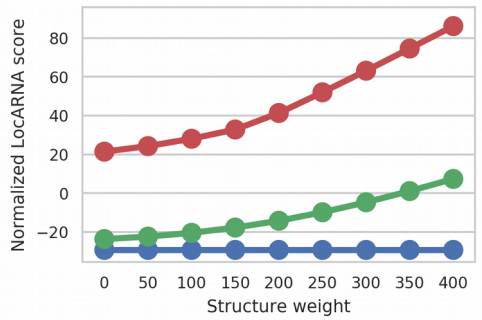


Combination of penalty and optimized parameters improves local alignment prediction!

Parameter	Gap γ	Gap opening β	Structure weight ω	Tau factor τ	Penalty λ
Default scoring	350	500	200	0	no
Suggested scoring	68	807	210	72	15

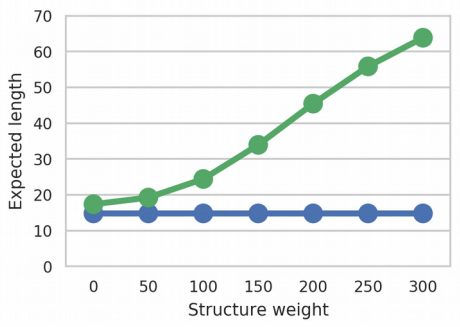
Summary

- Positive scoring bias due to structure



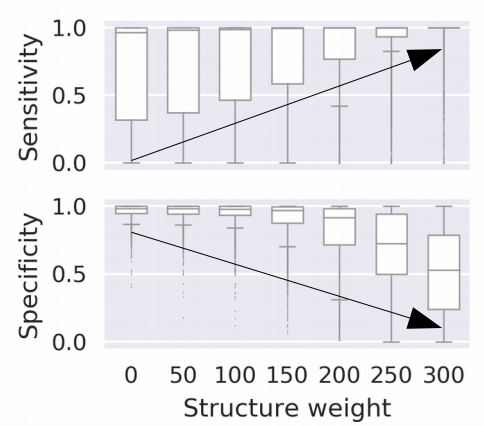
Summary

- Positive scoring bias due to structure
- Bias leads to alignment length extension for random sequences



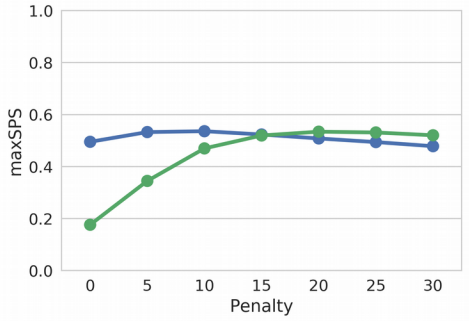
Summary

- Positive scoring bias due to structure
- Bias leads to alignment length extension for random sequences
- Trade-off: increased structure weight leads to higher ncRNA coverage but also extension into context



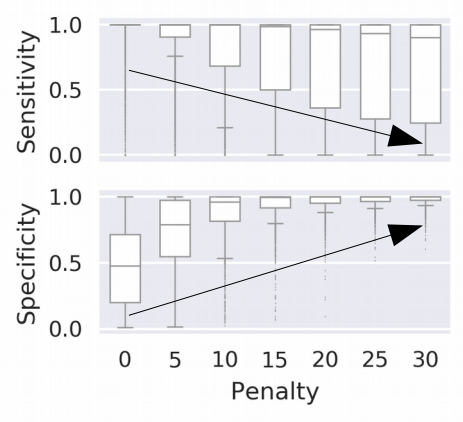
Summary

- Positive scoring bias due to structure
- Bias leads to alignment length extension for random sequences
- Trade-off: increased structure weight leads to higher ncRNA coverage but also extension into context
- Position-specific penalty improves alignment detection



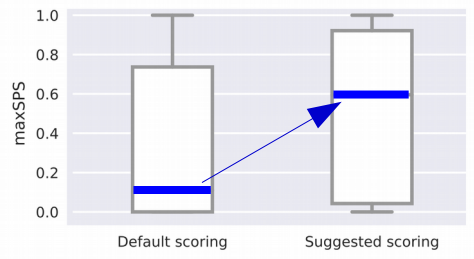
Summary

- Positive scoring bias due to structure
- Bias leads to alignment length extension for random sequences
- Trade-off: increased structure weight leads to higher ncRNA coverage but also extension into context
- Position-specific penalty improves alignment detection
- Compensate context extension with penalty



Summary

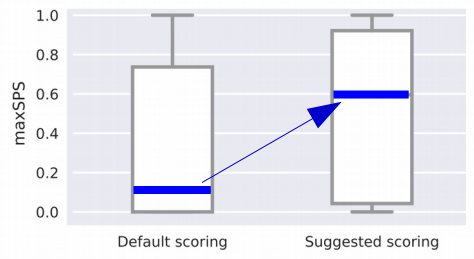
- Positive scoring bias due to structure
- Bias leads to alignment length extension for random sequences
- Trade-off: increased structure weight leads to higher ncRNA coverage but also extension into context
- Position-specific penalty improves alignment detection
- Compensate context extension with penalty
- Suggested scoring improves alignment quality

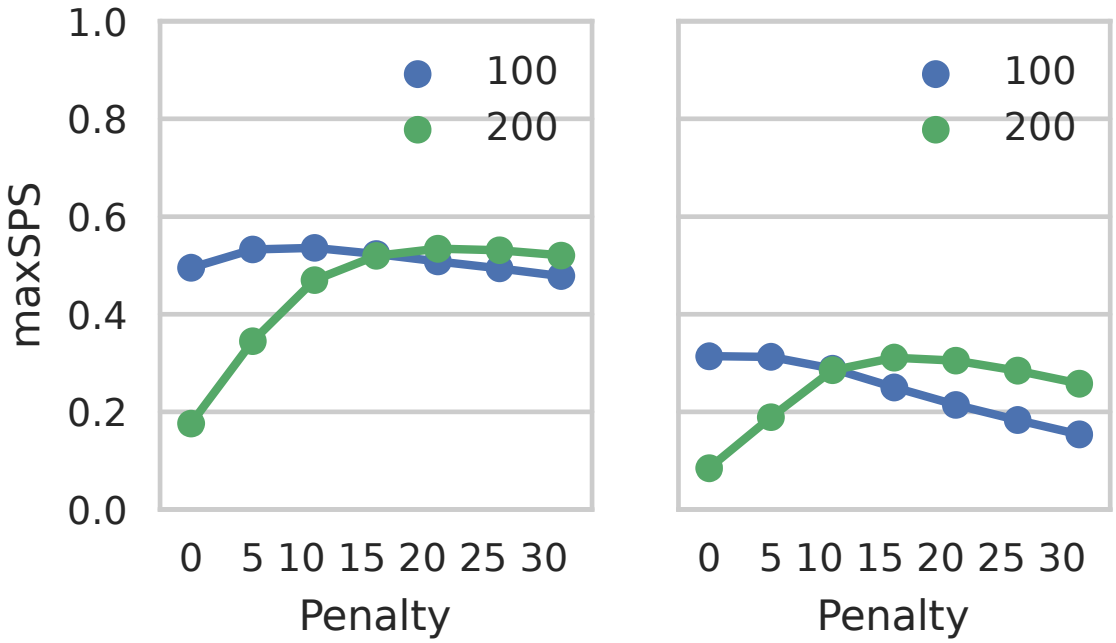


Thanks to:

Summary

- Positive scoring bias due to structure
- Bias leads to alignment length extension for random sequences
- Trade-off: increased structure weight leads to higher ncRNA coverage but also extension into context
- Position-specific penalty improves alignment detection
- Compensate context extension with penalty
- Suggested scoring improves alignment quality

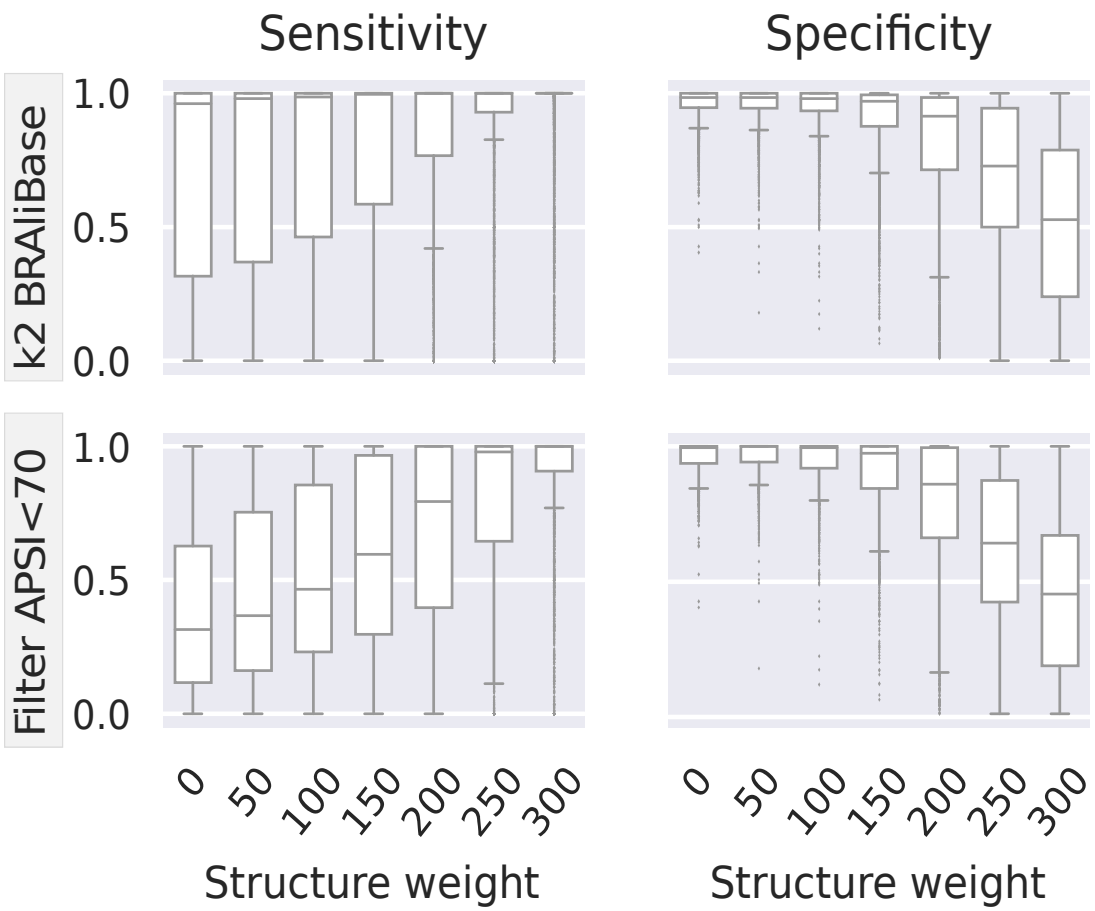




LocalBralibase

APSI < 70

Boundary detection

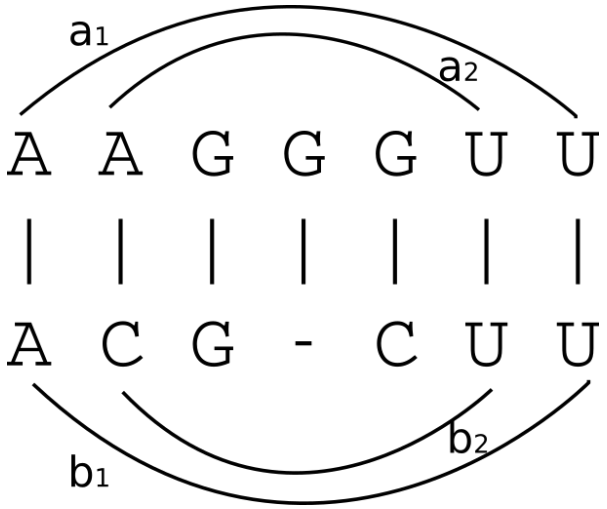


Sequence-structure Sankoff-like alignment: LocARNA

- Objective function:

$$\underbrace{\sum_{(ij;kl)\in S} \left(\omega(\Psi^a_{ij} + \Psi^b_{kl}) + \tau\sigma'(a_i, a_j, b_k, b_l)\right)}_{\text{Structure contribution}} + \underbrace{\sum_{(i,k)\in A_s} \sigma(a_i, b_k) - N_{gap}\gamma - N^o_{gap}\beta}_{\text{Sequence contribution}}$$

- Structure weight ω
- Tau factor τ
- Base pair score ψ
- Ribosome scoring σ
- Gap extension γ
- Gap opening β



MaxSPS example

$$maxSPS = \frac{correct\ predicted\ edges}{maxLength(reference, predicted)}$$

reference alignment

```
UGGCACGCUGC
--|||||||---
CAGGAACCAAG
```

reference length = 6

predicted alignment 1

```
UGGCACGCUGC
--|||||-----
CAGGA-ACCAAG
```

$maxSPS = \frac{3}{6}$

predicted length = 5

predicted alignment 2

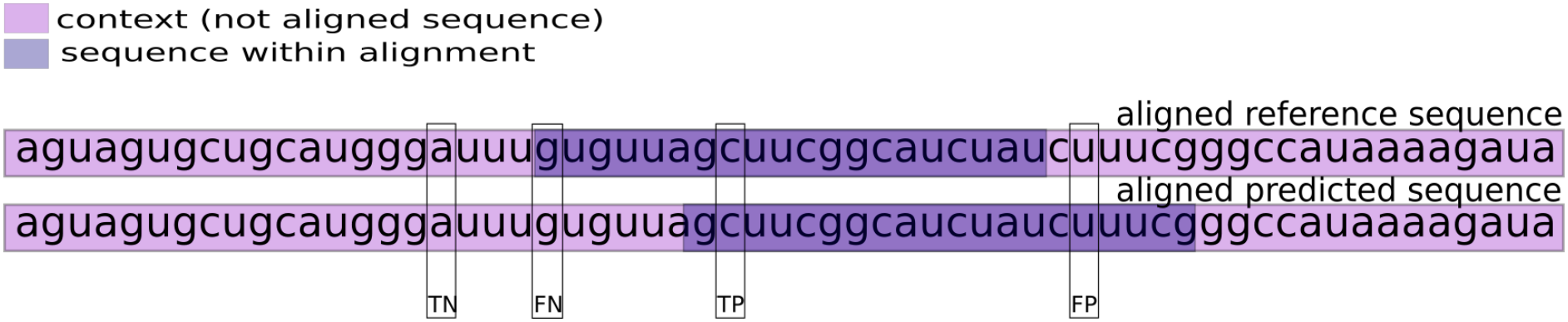
```
UGGCA-CGCUGC
--|||||||||--
CAGGAA-CCAAG
```

$maxSPS = \frac{5}{8}$

predicted length = 8

Sensitivity and Specificity

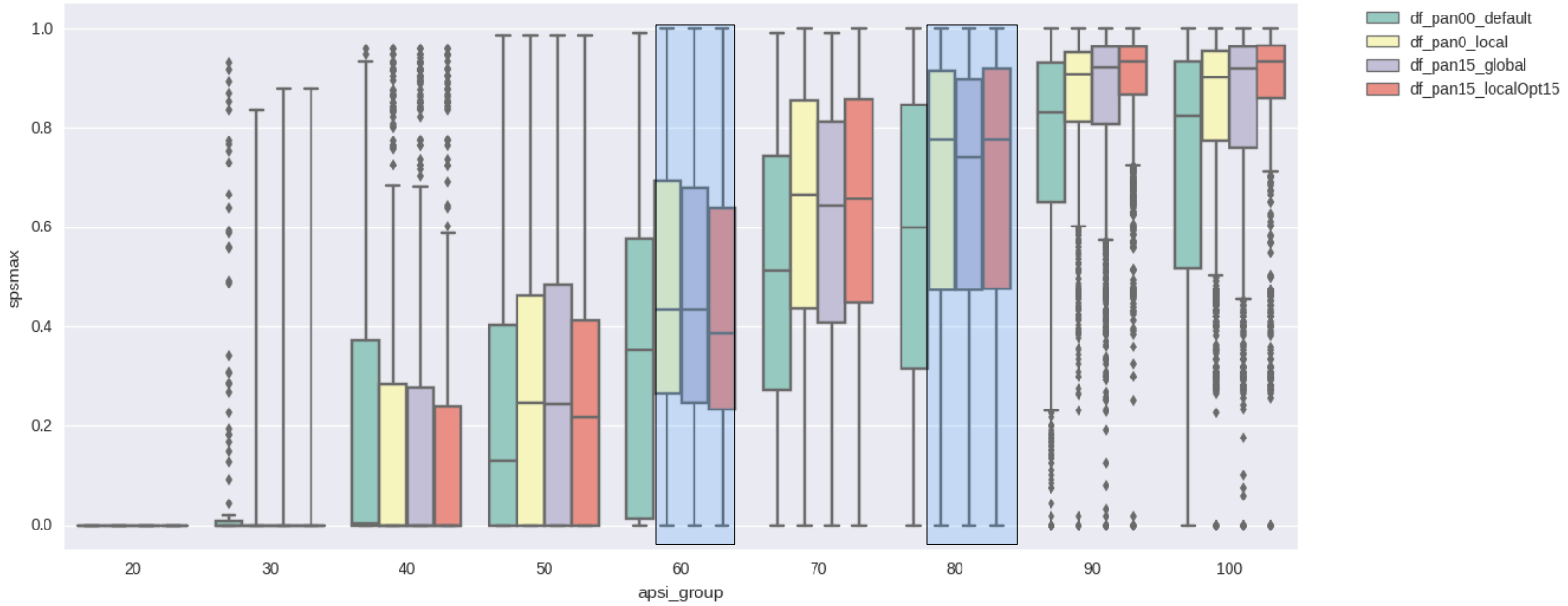
How to count TP, TN, FP and FN:



Sensitivity (RNA) = TP/(TP+FN)

Specificity (Context) = TN/(TN +FP)

Optimization using penalty 15



parameter	Gap	Gap opening	Structure weight	Tau factor
default	350	500	200	0
Global optimized	68	807	210	72
Local optimized	136	975	115	38
Local optimized (pen15)	82	883	176	71