Improving the local alignment of LocARNA through automated parameter optimization

Bled 18.02.2016

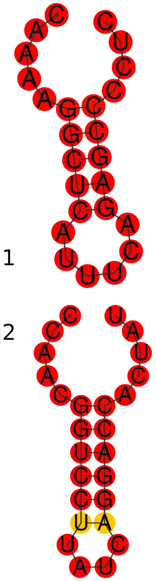
Teresa Müller

Albert-Ludwigs-Universität Freiburg

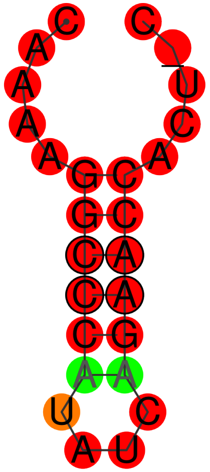
UNI FREIBURG

Introduction

- Non-coding RNA



```
seq1    CAAAAGGCUCAUUUCAGAGCCCCUC
seq2    CCAACGGUCCUUAUCAGGACCACUAU
```



```
seq1    CAAAAGGCUCAUUUCAGAGCCCCU-C
seq2    CCAACGGUCCUUAUCAGGACCACUAU
```

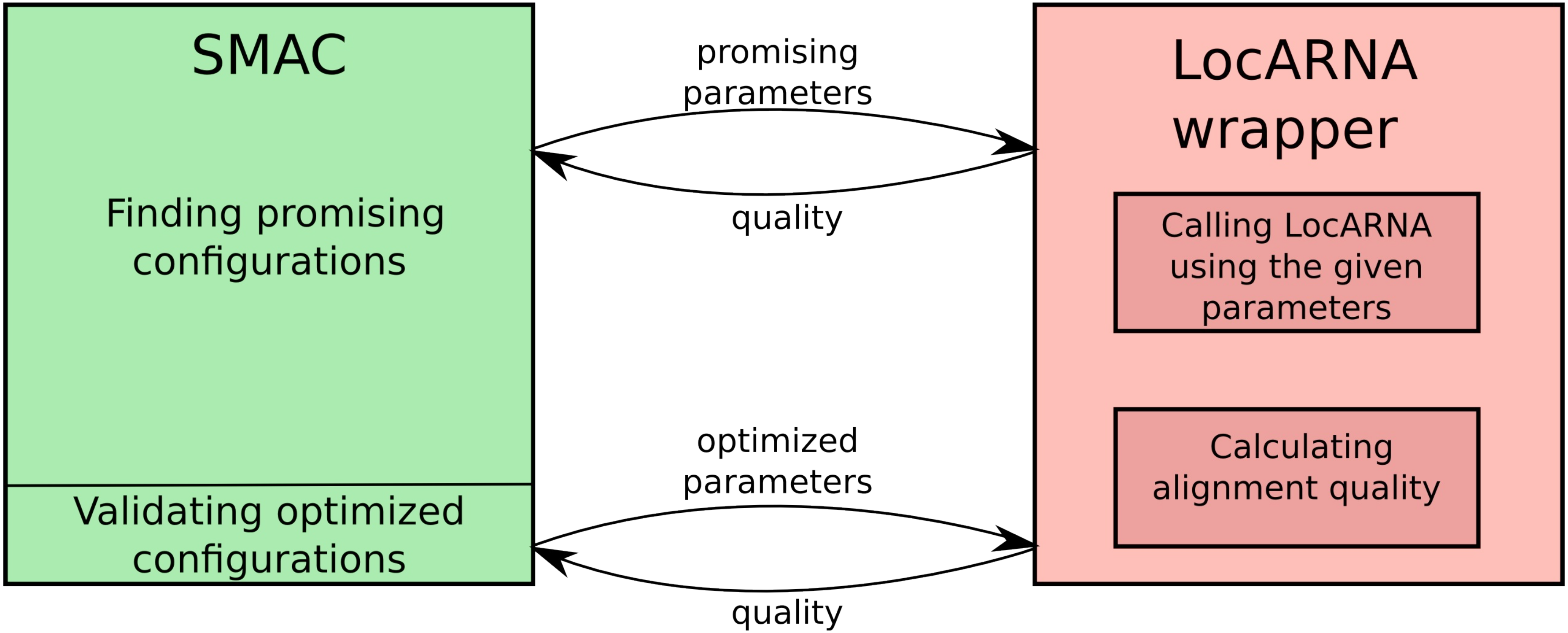
- High performing RNA alignment tool → correct classification
- LocARNA: global and local alignment program
  - Heuristic of Sankoff algorithm [Sankoff, 1985]

SMAC

[Hutter, Hoos, Leyton-Brown, 2011]

- Sequential Model-Based Algorithm Configuration
- Black box tool
- Task: find high-performance parameter settings
- Uses Random Forest model
  - New parameter setting cleverly chosen
- Can optimize categorical parameters

Set-up



SMAC

Finding promising configurations

Validating optimized configurations

promising parameters

quality

optimized parameters

quality

LocARNA wrapper

Calling LocARNA using the given parameters

Calculating alignment quality

Local alignment

- Global alignment

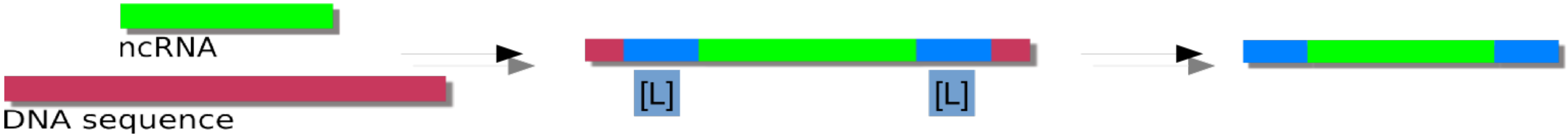
```
--T--CC-C-AGT--TATGT-CAGGGGACACG--A-GCATGCAGA-GAC
  |  || |  ||  | | | |||    || |  | |  | ||||   |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG--T-CAGAT--C
```

- Local alignment

```
                  tccCAGTTATGTCAGgggacacgagcatgcagagac
                     ||||||||||||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

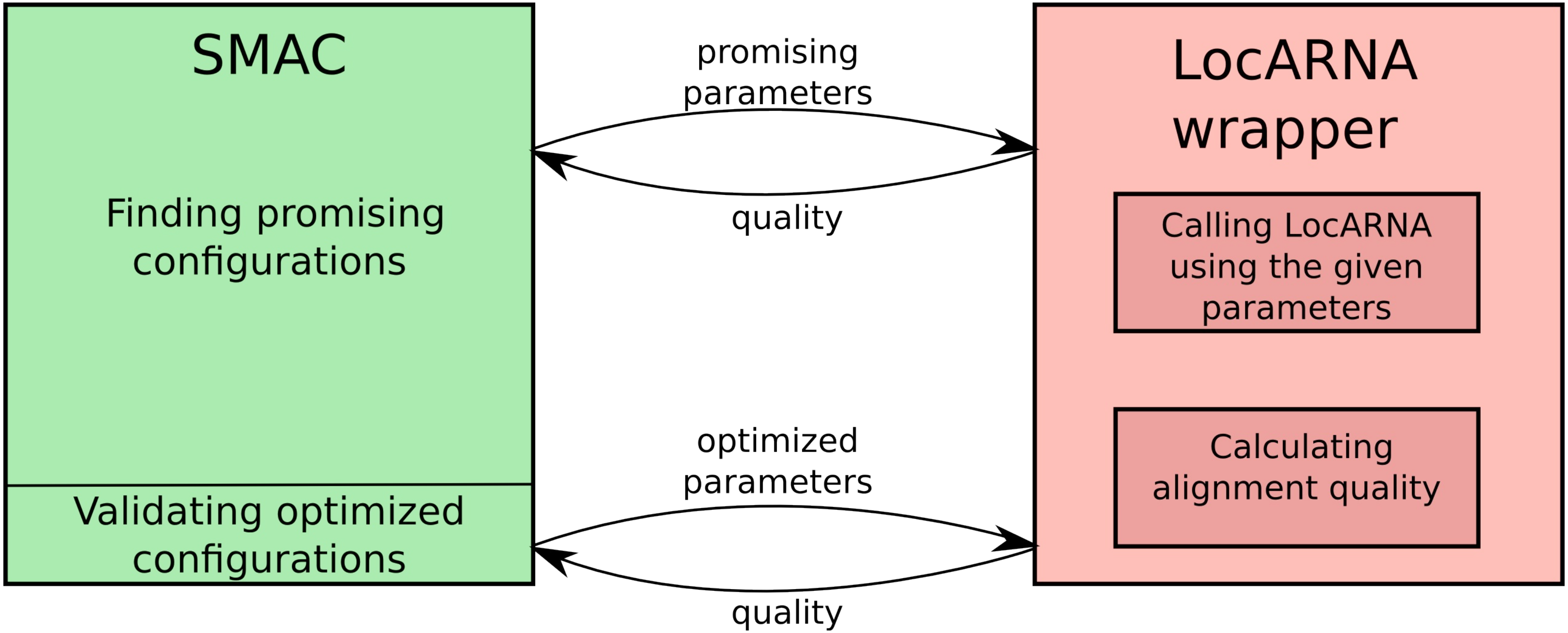
- Lack of accurate local sequence-structure alignment tools
- Challenges of sequence-structure local alignment:
  - Find correct boundaries
  - Find correct alignment edges

Construct local benchmark set from BRAliBase



- BRAliBase ncRNA (green)
- Extract genomic context (red) from European Nucleotide Archive
- Specify a context size [L]
- Extract context parts (blue)
- Shuffled context areas

Set-up


SMAC

Finding promising configurations

Validating optimized configurations

promising parameters

quality

optimized parameters

quality

LocARNA wrapper

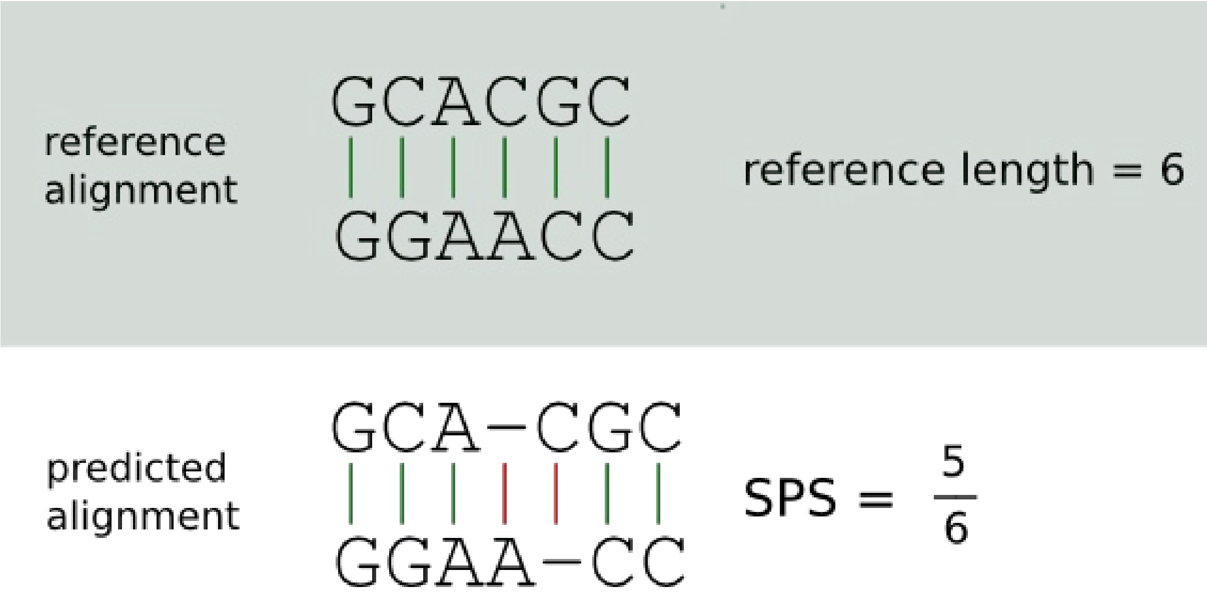
Calling LocARNA using the given parameters

Calculating alignment quality

Sum of Pairs Score

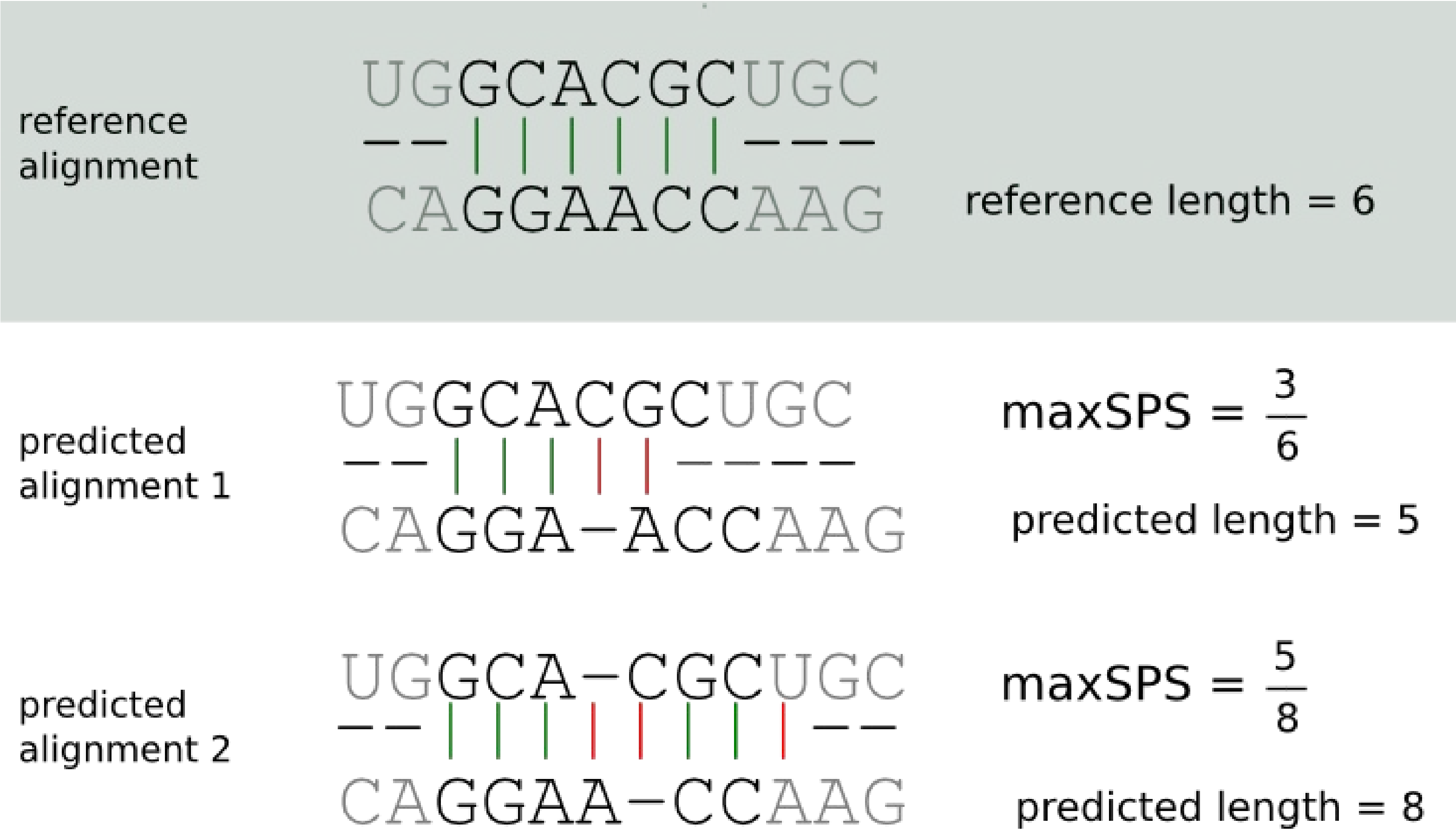
[Thompson et al., 1999]

$$SPS = \frac{correct\ predicted\ edges}{number\ of\ reference\ edges}$$



maxSPS example

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



reference alignment

reference length = 6

predicted alignment 1

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

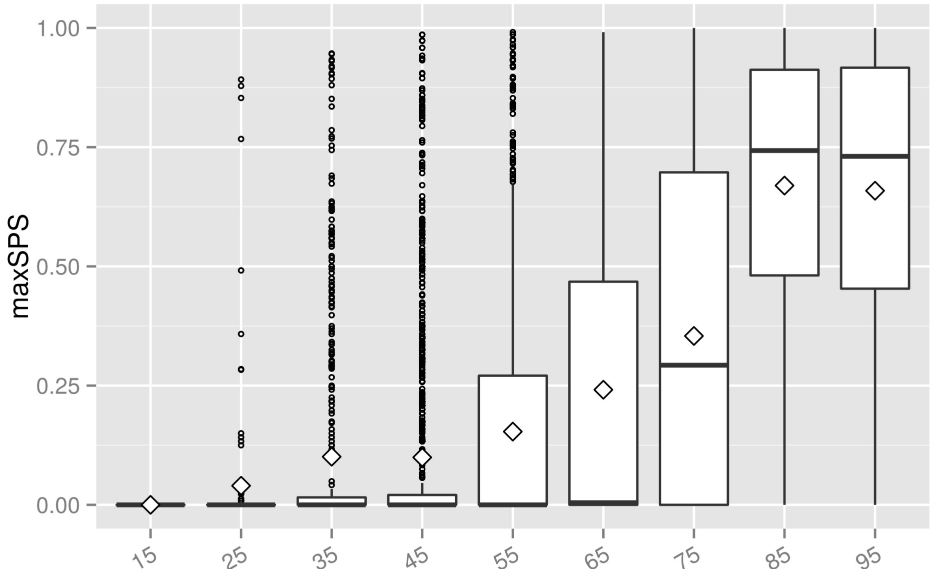
predicted length = 5

predicted alignment 2

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

predicted length = 8

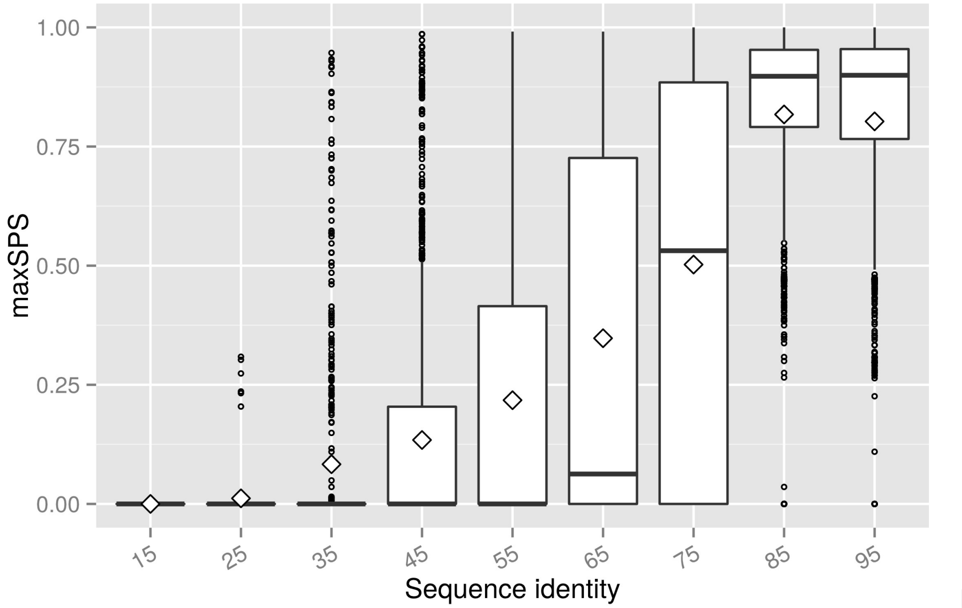
Default vs. Optimized maxSPS



Default parameter setting

Low SI: low maxSPS quality

High SI: more easy to find alignment edges



Optimized parameter setting

Improvement for SI > 40

For low SI 40 no change (less data points)

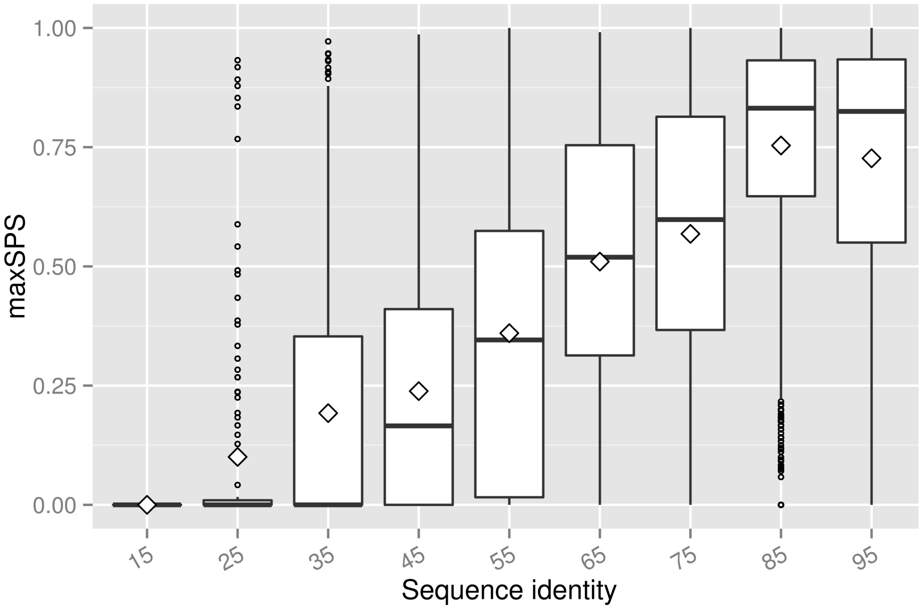
Validation of best run

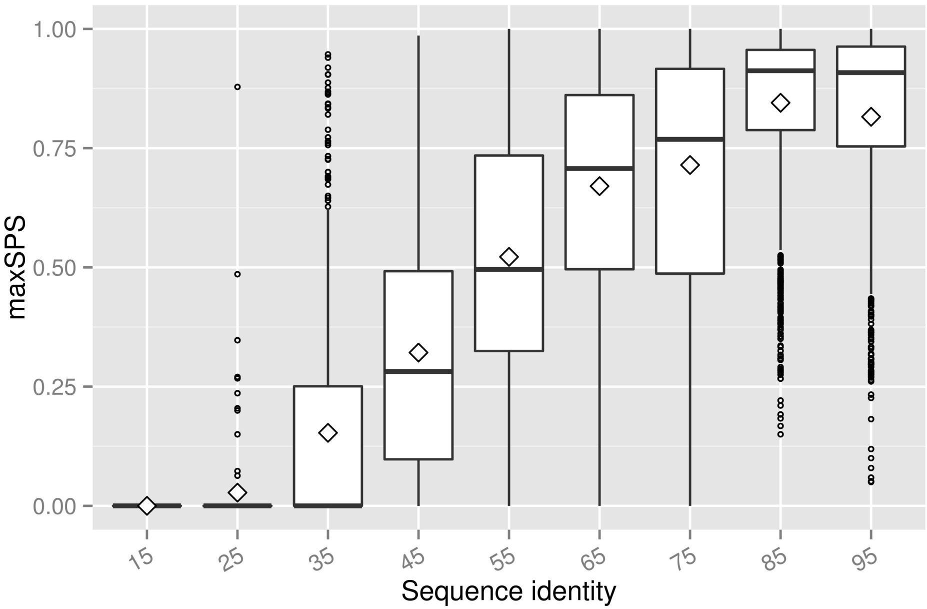
Validation set	Dataset context 20	Dataset context 200
Default parameters	0.393	0.512
Optimized parameters	0.362	0.376
Improvement	8%	27%

Position penalty

- Observation: Background bonus
  - Conserved structures can be found in context
- Solution: position penalty λ
  - Each position of the local alignment is penalized by λ

Position penalty 5 optimization





- Default parameter setting

  using position penalty 5
- Improvement even without optimization

Optimized parameter setting

Parameter optimization based

on dataset SI 40 - 70

Summary

- Novel local benchmark set
- New local quality measure maxSPS
- Learning improves maxSPS (27 %)
- Position penalty solely improves maxSPS
- Additional improvement by learning
- Outlook: more parameters, position penalty validation, additional benchmark set

parameter	Gap	Gap opening	Structure weight	Tau factor
default	350	500	200	0
first optimized	136	975	115	38
Penalty 5 optimized	29	848	127	81

Acknowledgement:

Prof. Dr. Rolf Backofen

Dr. Frank Hutter

Dr. Sebastian Will

Milad Miladi

Christina Otto

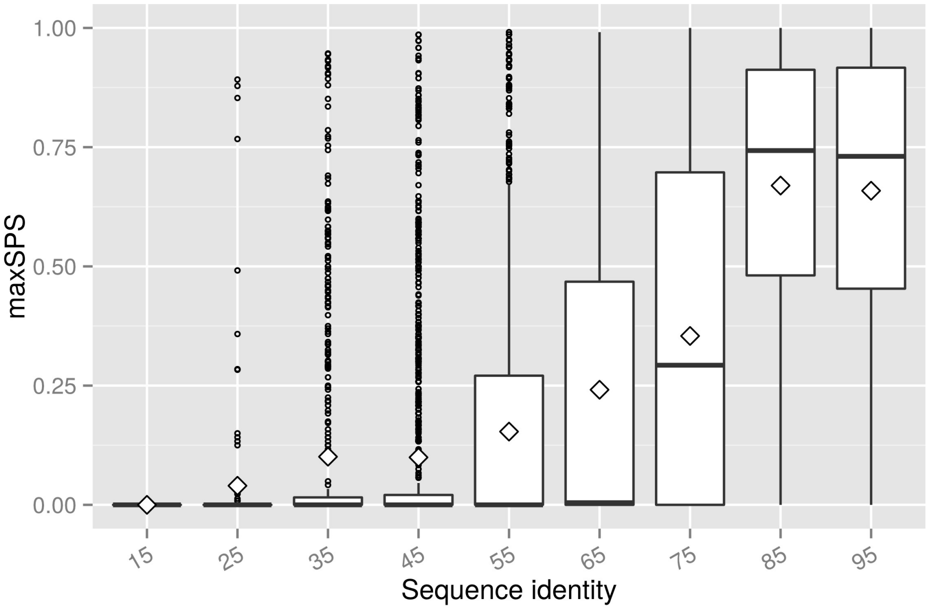
Thanks for your attention

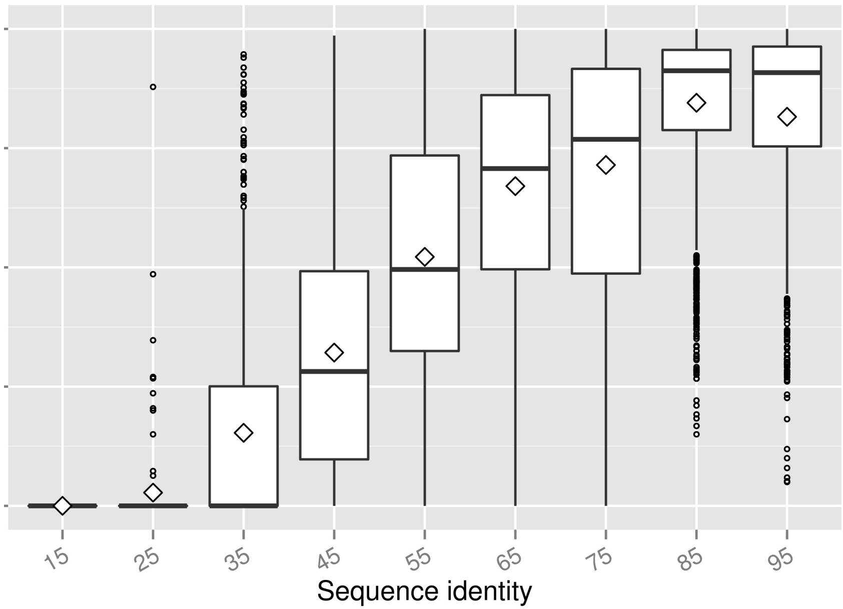
Albert-Ludwigs-Universität Freiburg

Outlook

- Optimization on the exhaustive set of parameter optimization
- Validate the position penalty
- Use different validation set
- Research on failed alignments

Position penalty 5





- Default parameter setting
- Default with position penalty 5

LocARNA scoring function

[Will et.al., 2007]

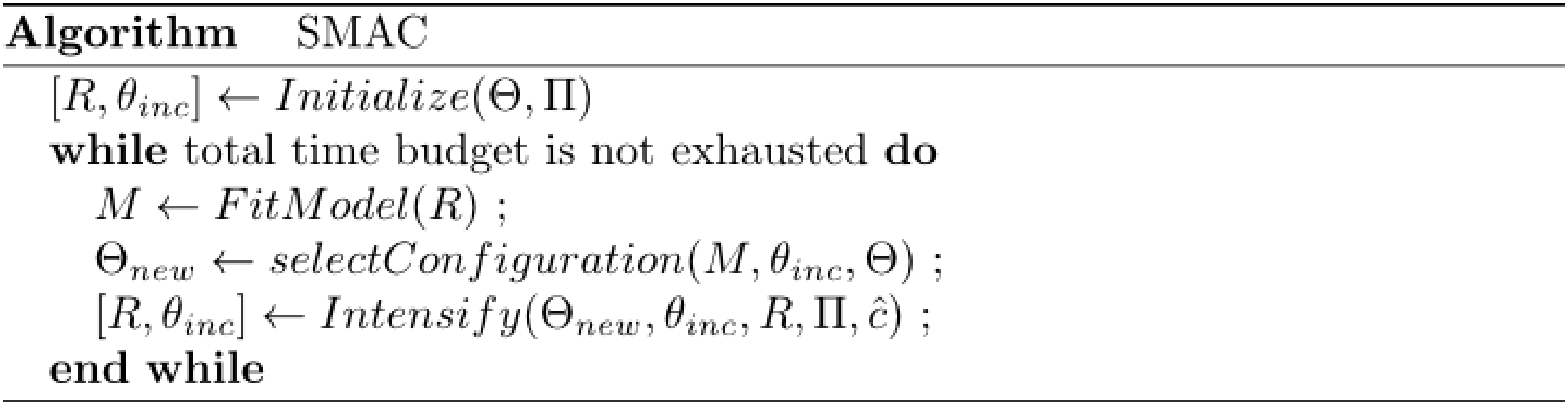
$$sw.\sum_{(ij;\,kl)\in S}\left(\Psi^{A}_{ij}+\Psi^{B}_{kl}\right)+tf.\sum_{(ij\,;kl)\in S}\left(\sigma\left(A_i,B_k\right)+\sigma\left(A_j,B_l\right)\right)+\sum_{(i,k)\in A_S}\sigma\left(A_i,B_k\right)-N_{gap}\gamma-N^{o}_{gap}\beta$$

Ψ_{ij}	Base pair weight
$\sigma\left(A_i,B_k\right)$	(mis-)match score
γ	Gap penalty
N_{gap}	No. of gaps
β	Gap opening penalty
N^{o}_{gap}	No. of gap openings
$sw.$	Structure weight
$tf.$	Tau factor

- Parameter optimization → algorithm configuration

SMAC algorithm

[Hutter, Hoos, Leyton-Brown, 2012]



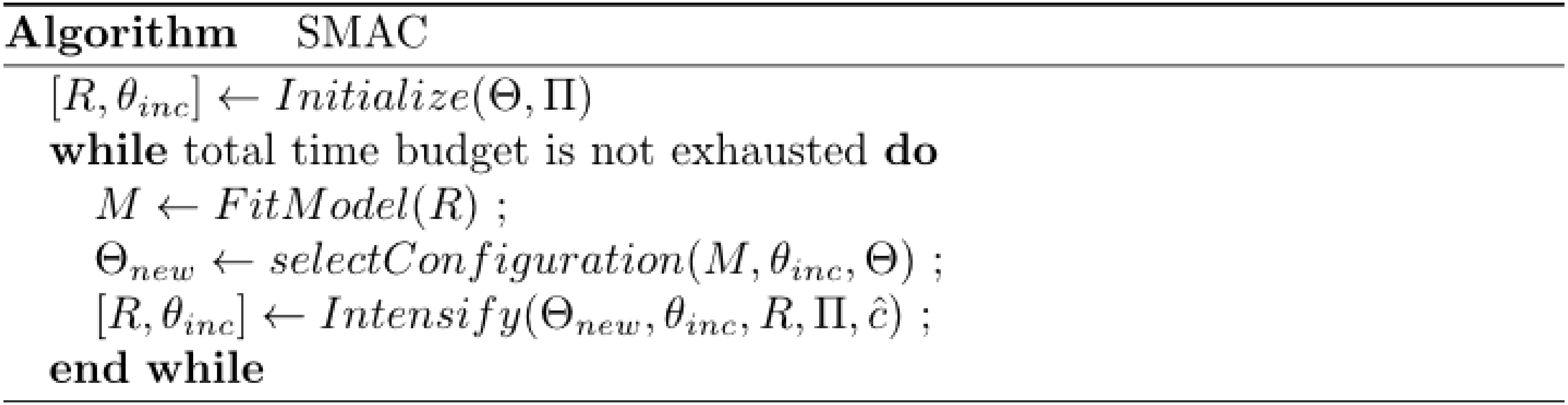
Algorithm SMAC

```
$[R, \theta_{inc}] \leftarrow Initialize(\Theta, \Pi)$
while total time budget is not exhausted do
    $M \leftarrow FitModel(R)$ ;
    $\Theta_{new} \leftarrow selectConfiguration(M, \theta_{inc}, \Theta)$ ;
    $[R, \theta_{inc}] \leftarrow Intensify(\Theta_{new}, \theta_{inc}, R, \Pi, \hat{c})$ ;
end while
```

- Specify parameter configuration space Θ
- Π instance space
- θ_{inc}: best parameter setting seen so far
- R tracks parameter settings and observed performance
- Initialization: set the first incumbent θ_{inc}, and R

Loop iterations

[Hutter, Hoos, Leyton-Brown, 2012]



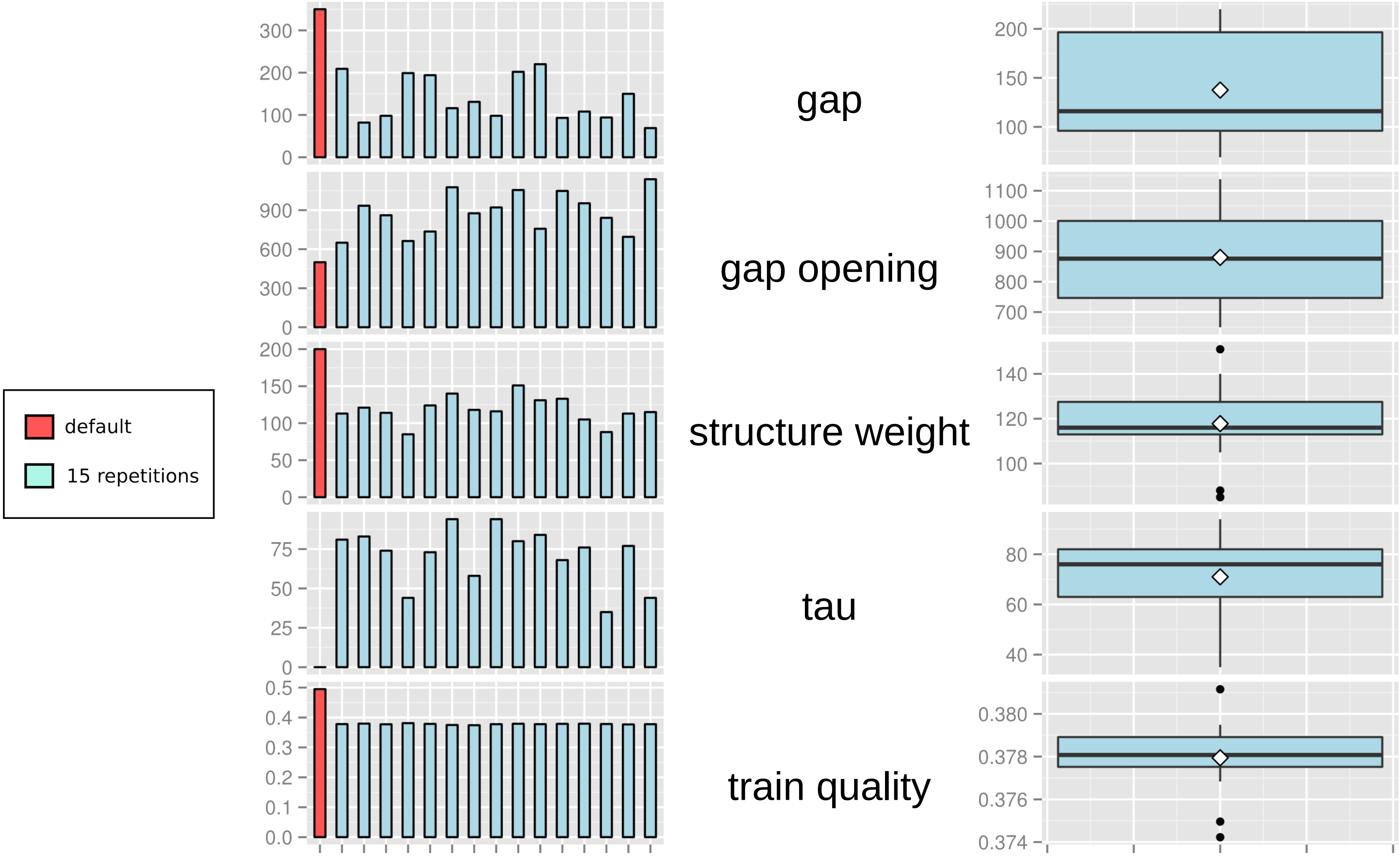
```
Algorithm SMAC
  $[R, \theta_{inc}] \leftarrow Initialize(\Theta, \Pi)$
  while total time budget is not exhausted do
      $M \leftarrow FitModel(R)$ ;
      $\Theta_{new} \leftarrow selectConfiguration(M, \theta_{inc}, \Theta)$ ;
      $[R, \theta_{inc}] \leftarrow Intensify(\Theta_{new}, \theta_{inc}, R, \Pi, \hat{c})$ ;
  end while
```

1. FitModel
   - Built using R
2. SelectConfiguration
   - Model finds promising configurations
3. Intensify
   - Compare promising configurations against incumbent

References

Reference Figure silde 13: http://rosalind.info/media/problems/swat/global_vs_local.png

Local Optimization results (con 100)



ncRNA sensitivity (RS) and context specificity (CS)

- Measuring the aligned areas for each sequence
- Calculate the mean of both values

	Alignment edge in reference alignment	No alignment edge in reference alignment
Alignment edge in predicted alignment	True positive (TP)	False Positive (FP)
No alignment edge in predicted alignment	False Negative (FN)	True Negative (TN)

ncRNA sensitivity (RS)

$$RS_A = \frac{TP_A}{TP_A + FN_A}$$

Context specificity (CS)

$$CS_A = \frac{TN_A}{TN_A + FP_A}$$

Optimization based on uniform k2-BRAliBase

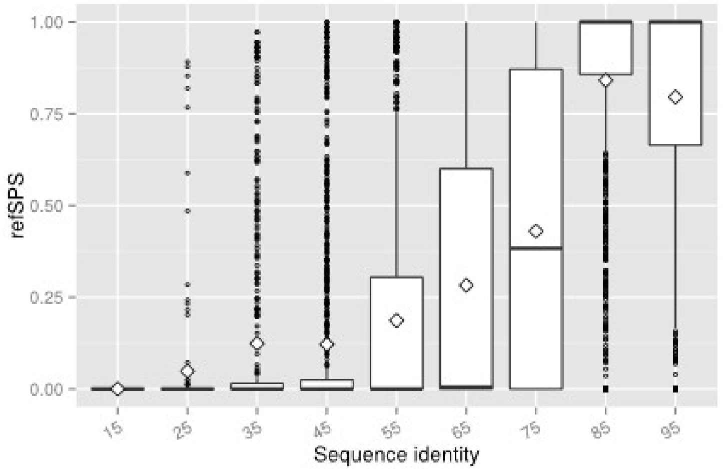
quality	train	default	difference	improvement
SPS	0.119	0.144	0.025	17 %
SP S ∗ MCC	0.226	0.280	0.054	19 %

Default parameter setting: -gap ’350’ -gap-opening ’500’ -struct-weight ’200’ -tau ’0’

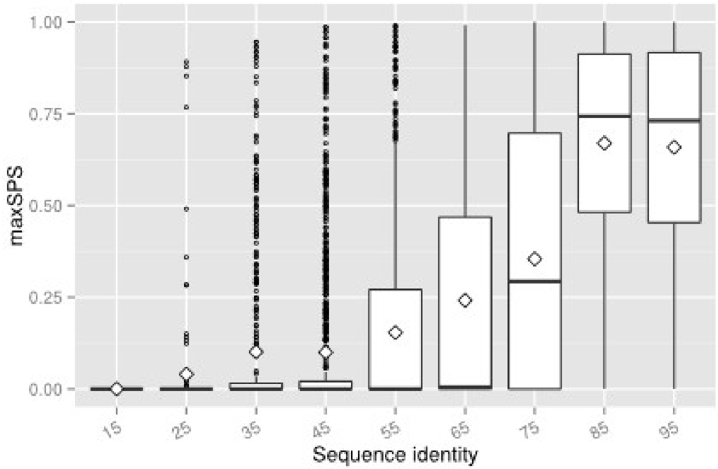
Final parameter setting: -gap ’68’ -gap-opening ’807’ -struct-weight ’210’ -tau ’72’

	Begin	End	Default
gap	0	1000	350
Gap-opening	0	1500	500
struct-weight	0	1000	200
tau	0	100	0

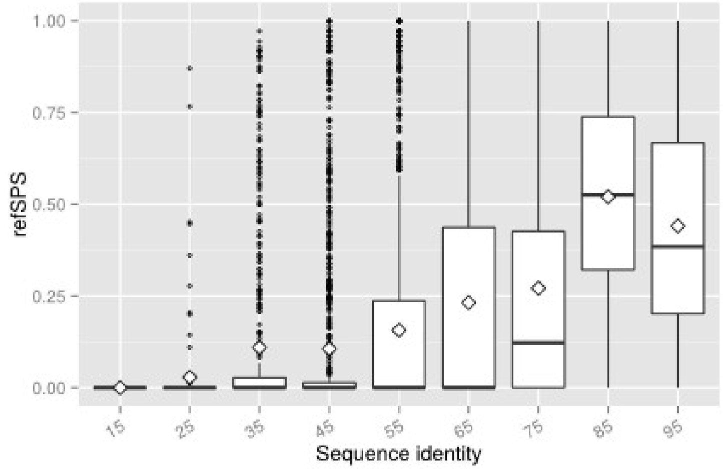
Distribution of refSPS and maxSPS (con 100)



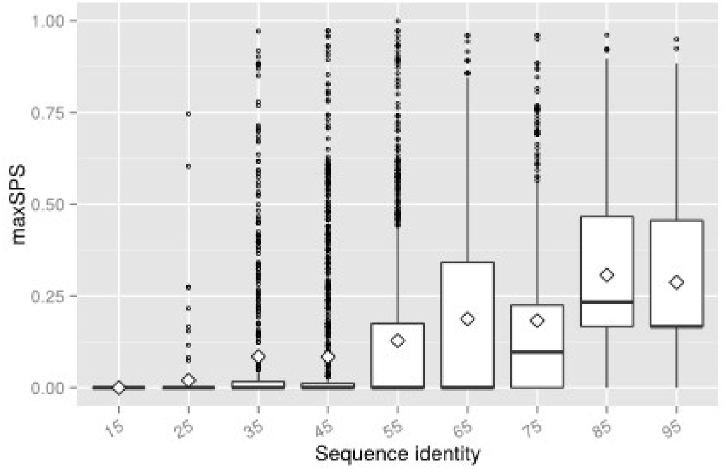
(a) shuffled



(b) shuffled

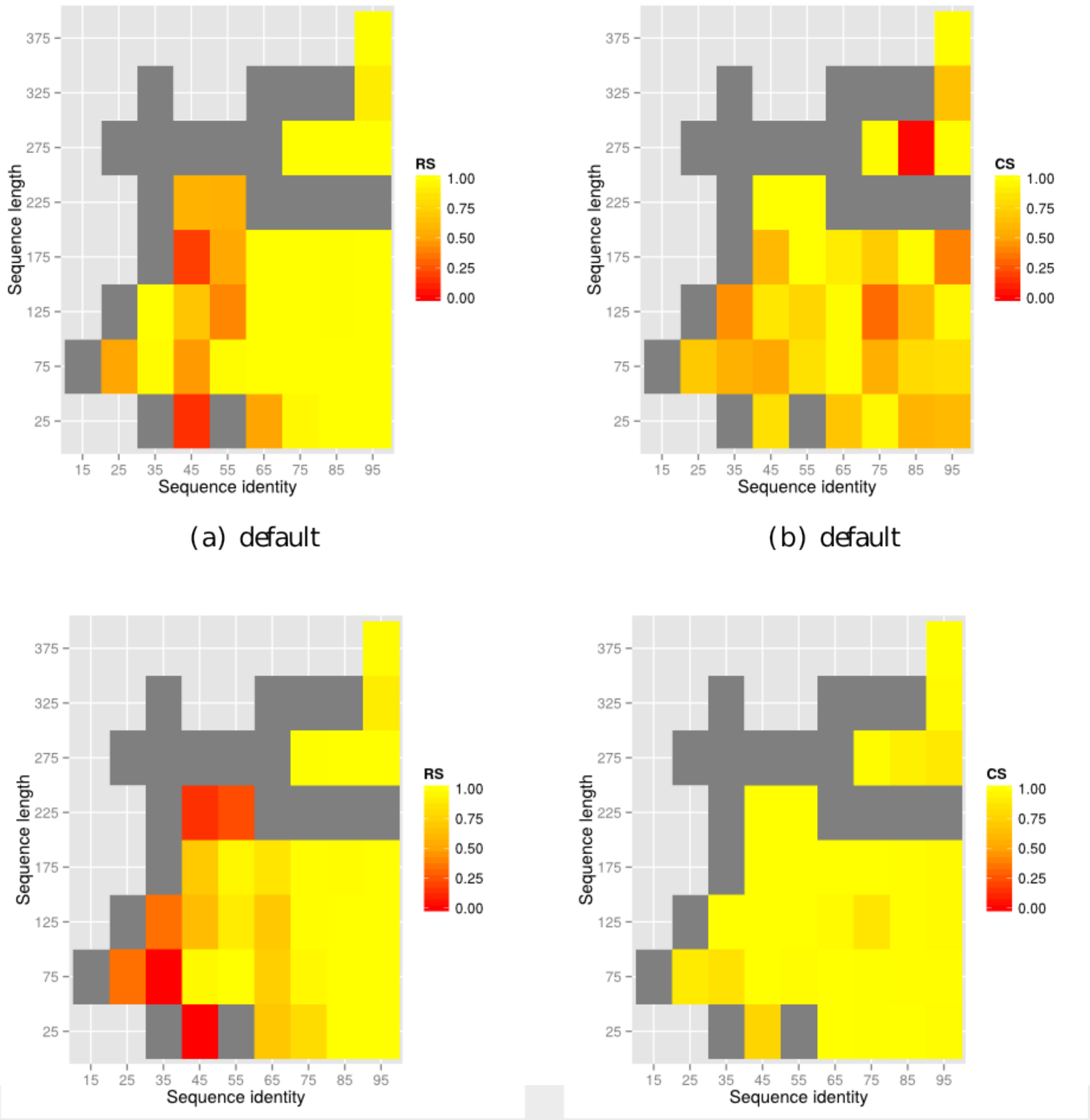


(c) not shuffled



(d) not shuffled

Heatmap default compared to optimized for sens/spec



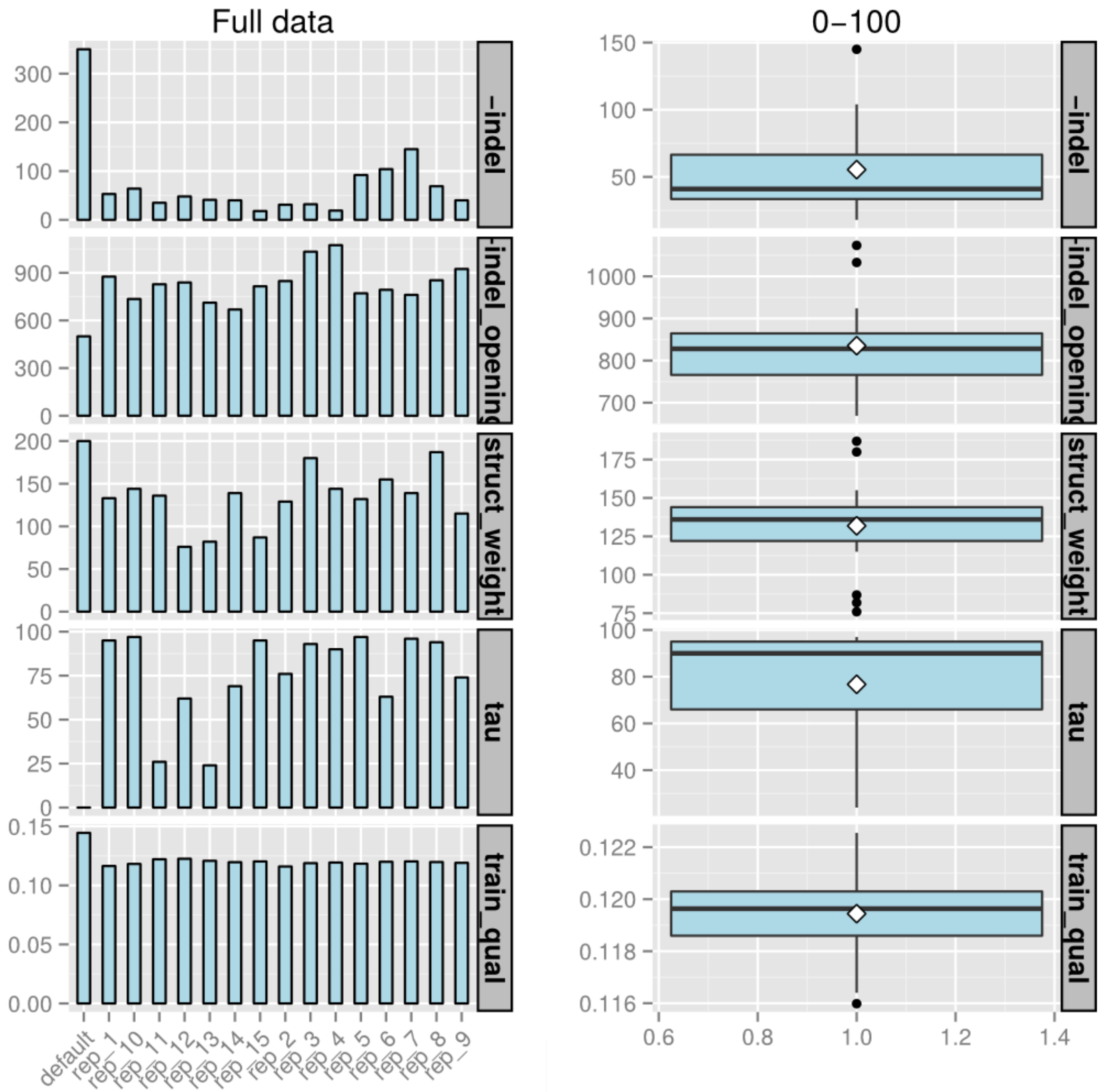
(a) default

(b) default

(c) best-optimized

(d) best-optimized

Parameter distribution of uniform k2-BRAliBase(SPS) global



gap

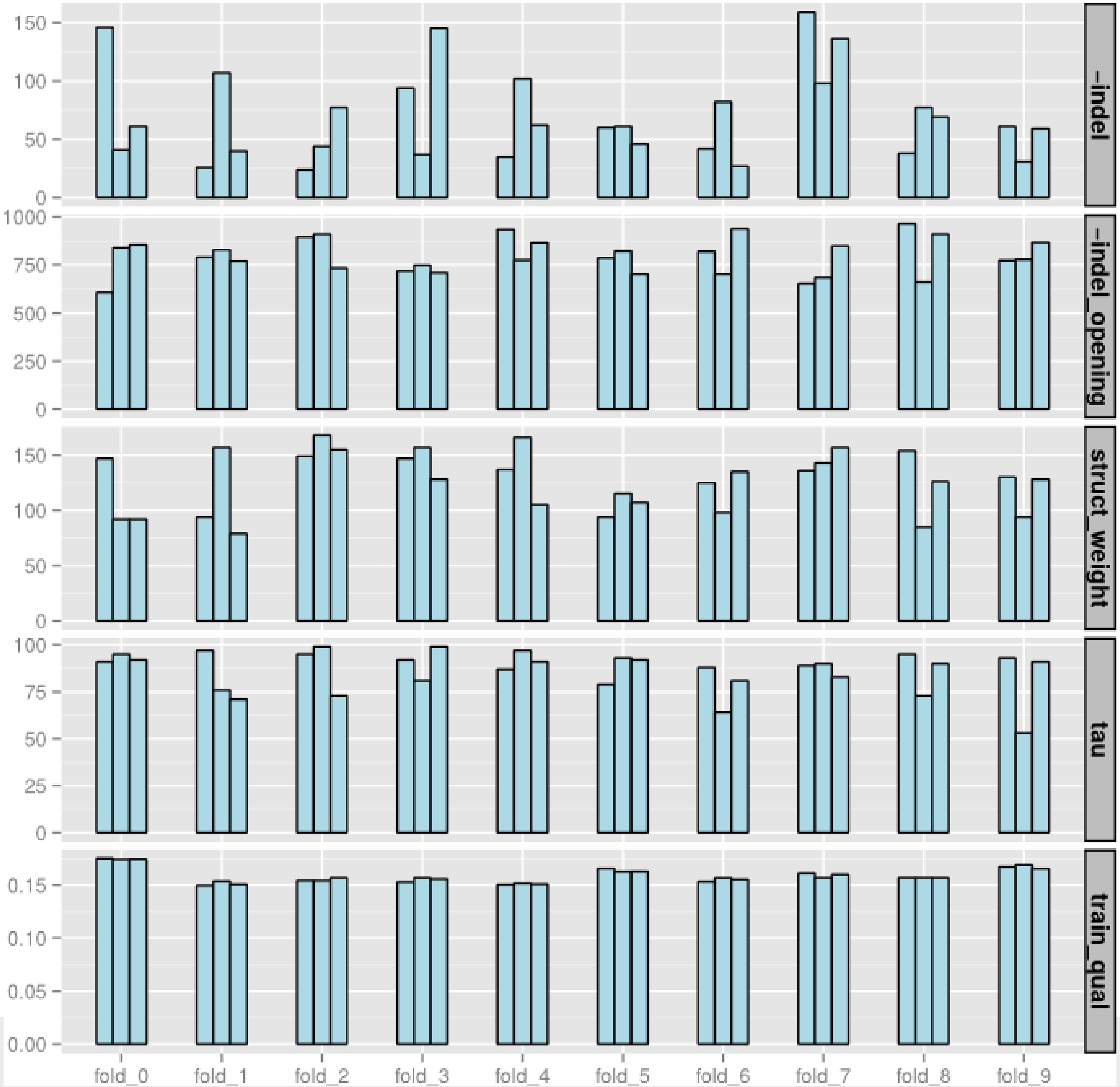
gap opening

structure weight

tau

train quality

K-fold Parameter (SI 50 - 70)



gap

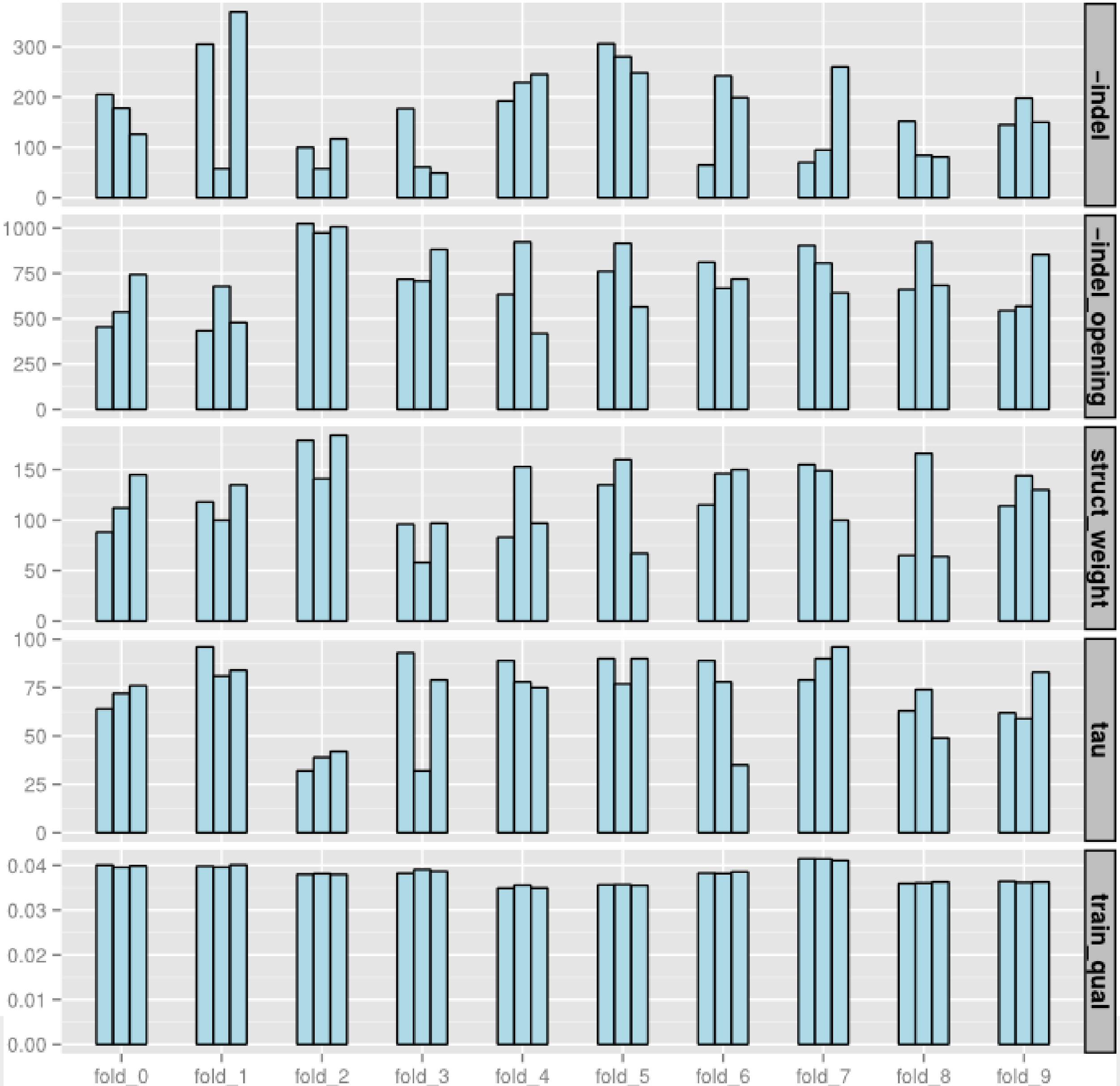
gap opening

structure weight

tau

train quality

K-fold Parameter (SI 71 - 90)



gap

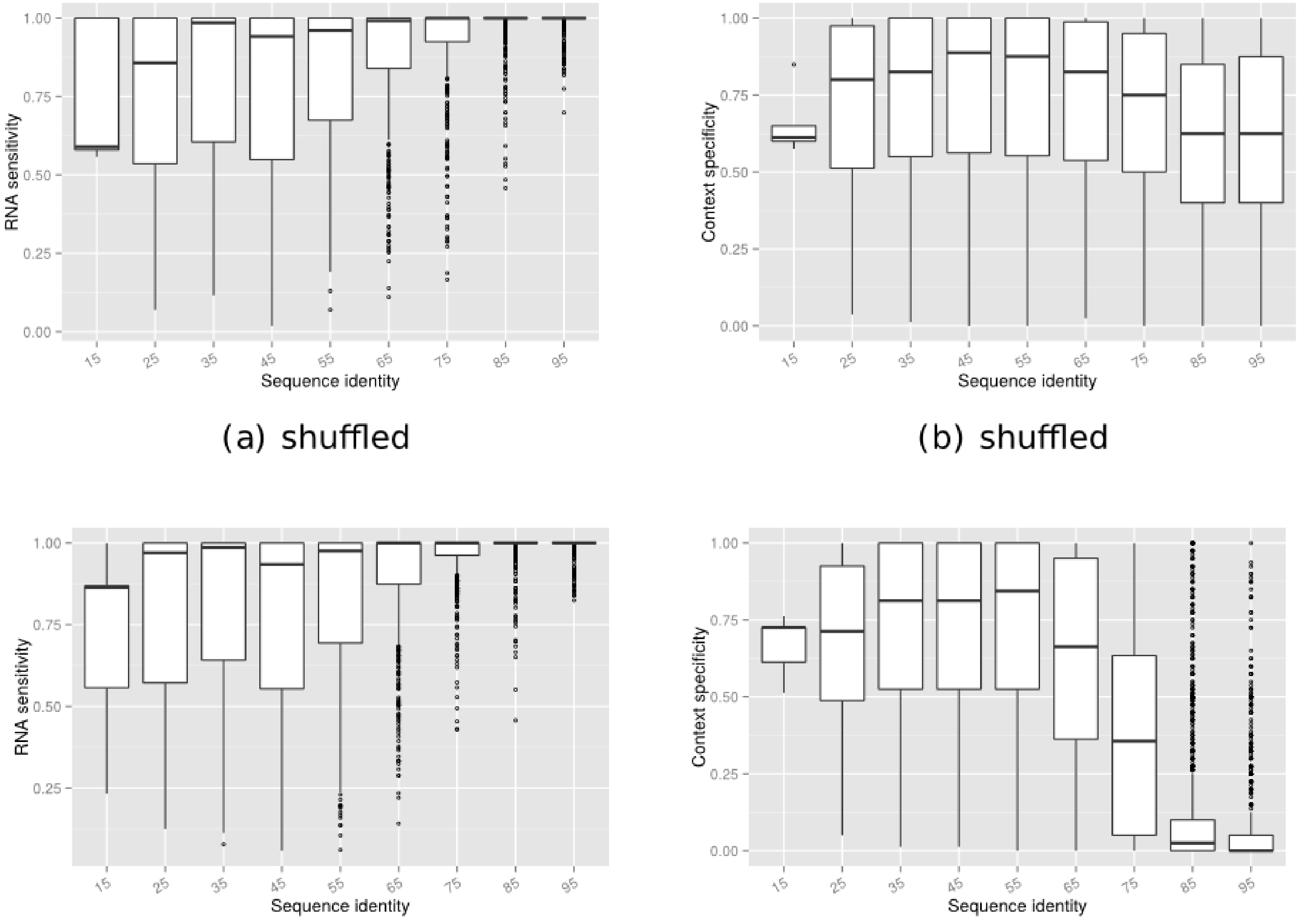
gap opening

structure weight

tau

train quality

Distribution of ncRNA sensitivity and context specificity (con 20)



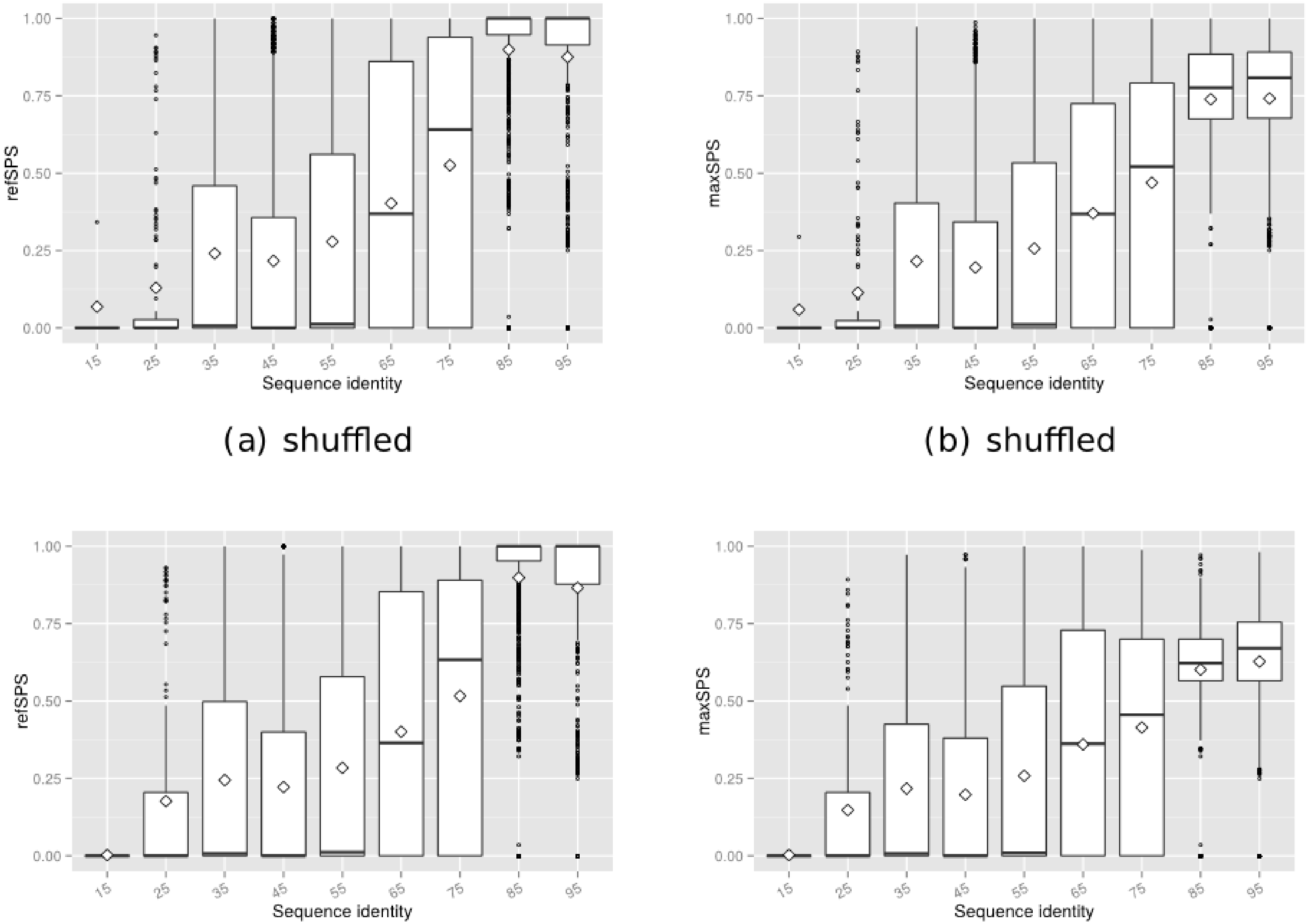
(a) shuffled

(b) shuffled

(c) not shuffled

(d) not shuffled

Distibution of refSPS and maxSPS (con 20)



(a) shuffled

(b) shuffled

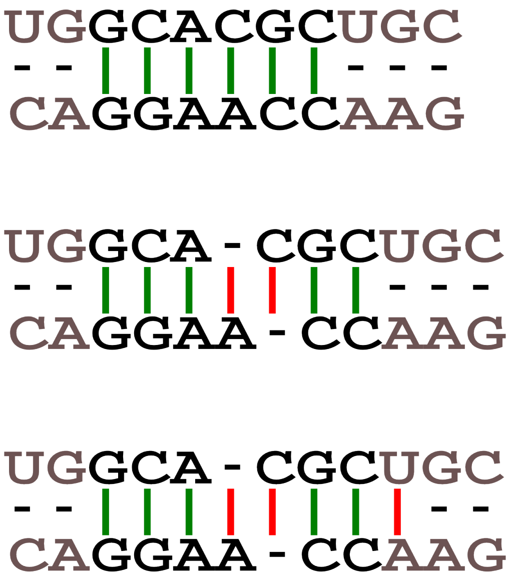
(c) not shuffled

(d) not shuffled

refSPS and maxSPS

$$refSPS = \frac{no.correctEdges}{referenceLength}$$

$$maxSPS = \frac{no.correctEdges}{max(referenceLength, predictedLength)}$$



maxSPS $= \frac{5}{7}$

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

Dataset size

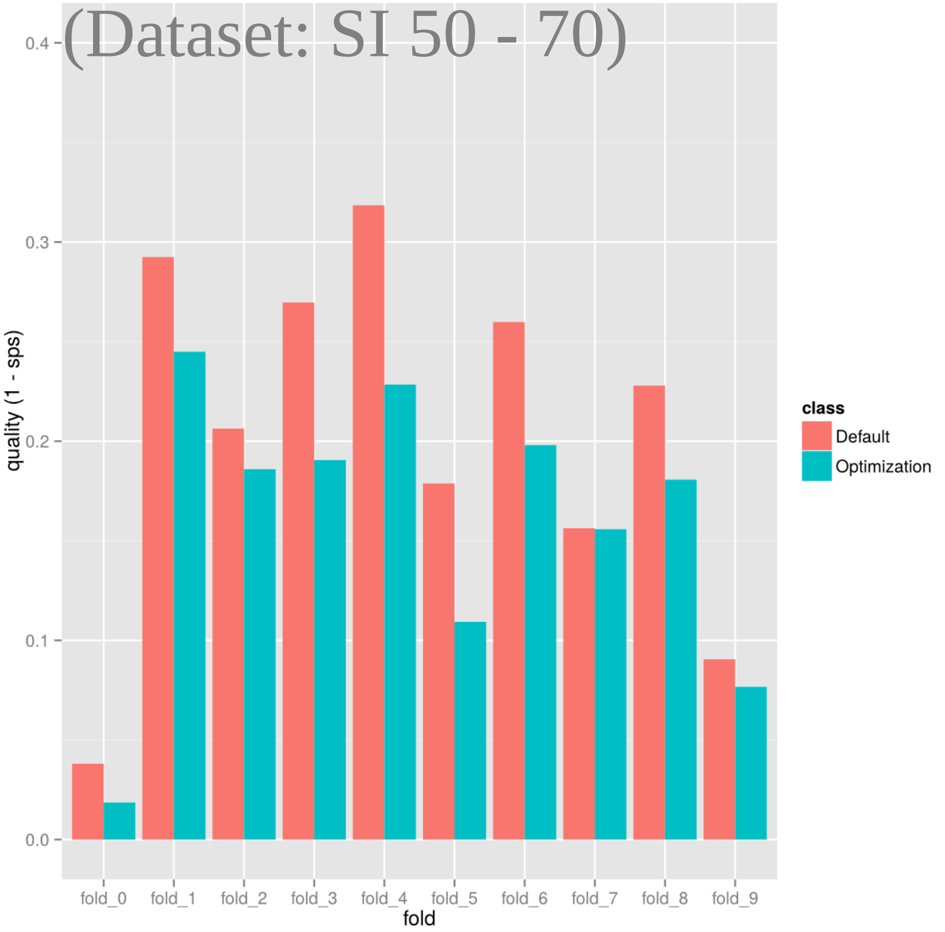
- Full dataset

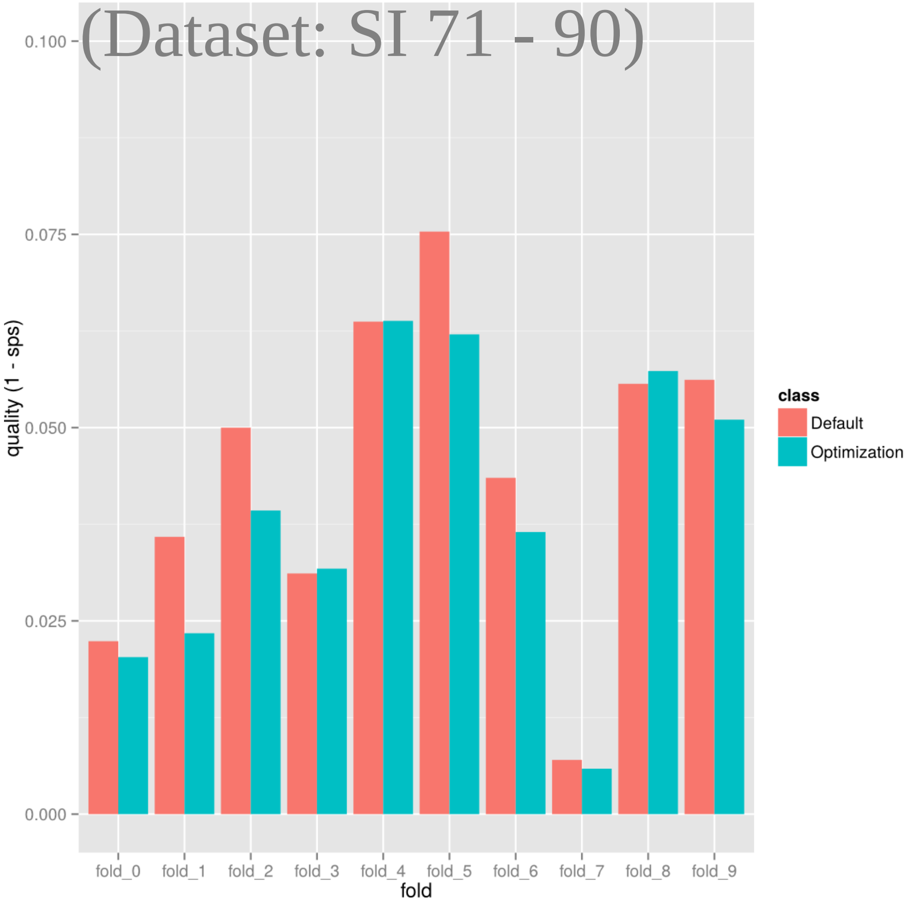
Dataset	Size
Full_Global_Dataset	2090
Full_Local_Dataset	1370

- SI dataset

Dataset	Training size	Validation Size
IS_50-70	513	57
IS_71-90	873	97

K-fold validation and default quality


(Dataset: SI 50 - 70)


(Dataset: SI 71 - 90)

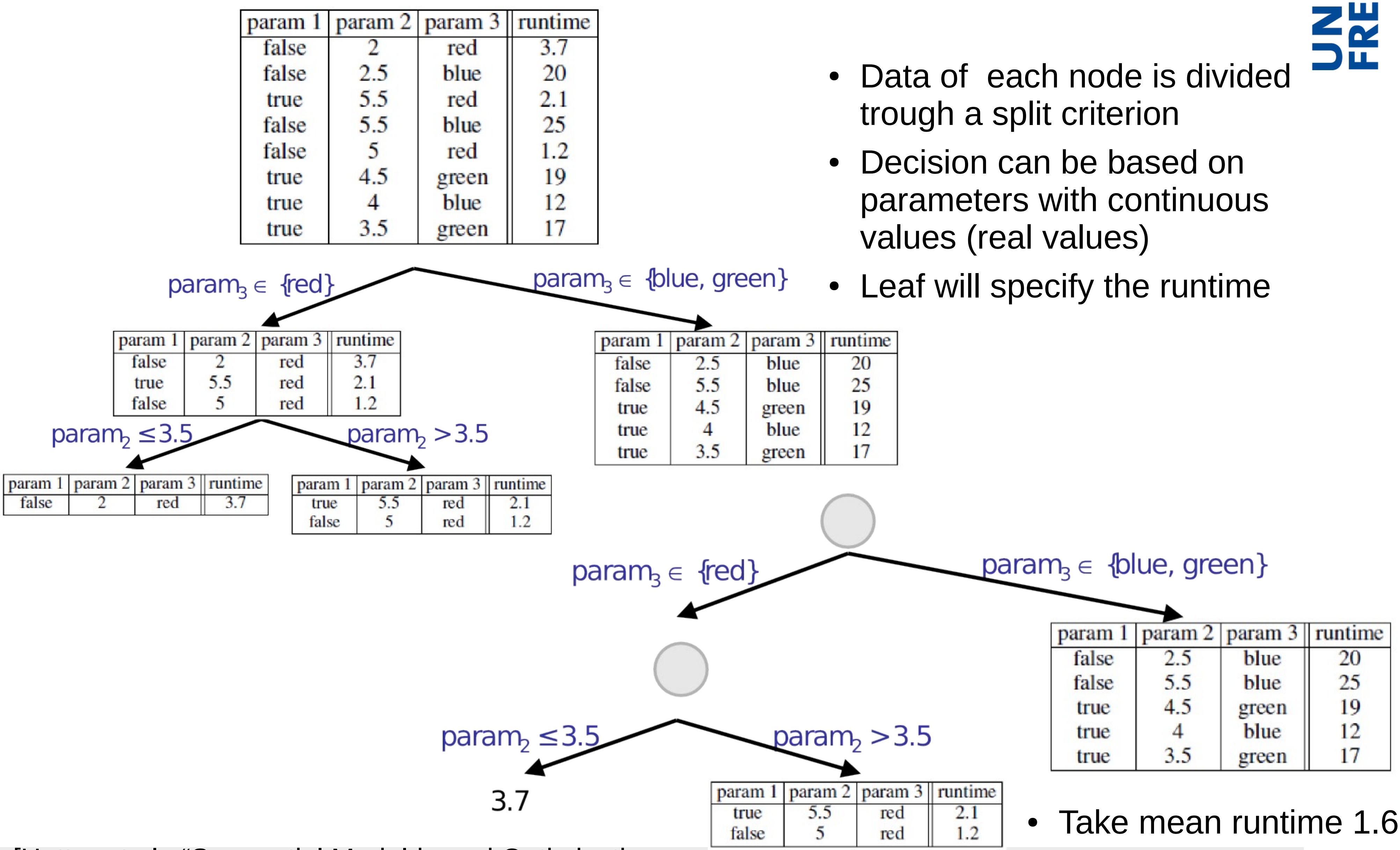
Dataset	Mean difference	Standard deviation
50-70	0.044	0.028
71-90	0.004	0.005

Average difference

dataset	Mean difference	standard deviation
50-70	0.044	0.028
71-90	0.004	0.005
50-70 swaped	0.030	0.013
71-90 swaped	0.003	0.001
50 – 70 (mcc)	0.036	0.026
71 – 90 (mcc)	0.096	0.025

Random Forest

- Data of each node is divided trough a split criterion
- Decision can be based on parameters with continuous values (real values)
- Leaf will specify the runtime



param 1	param 2	param 3	runtime
false	2	red	3.7
false	2.5	blue	20
true	5.5	red	2.1
false	5.5	blue	25
false	5	red	1.2
true	4.5	green	19
true	4	blue	12
true	3.5	green	17

param 1	param 2	param 3	runtime
false	2	red	3.7
true	5.5	red	2.1
false	5	red	1.2

param 1	param 2	param 3	runtime
false	2.5	blue	20
false	5.5	blue	25
true	4.5	green	19
true	4	blue	12
true	3.5	green	17

param 1	param 2	param 3	runtime
false	2	red	3.7

param 1	param 2	param 3	runtime
true	5.5	red	2.1
false	5	red	1.2

param 1	param 2	param 3	runtime
false	2.5	blue	20
false	5.5	blue	25
true	4.5	green	19
true	4	blue	12
true	3.5	green	17

param 1	param 2	param 3	runtime
true	5.5	red	2.1
false	5	red	1.2

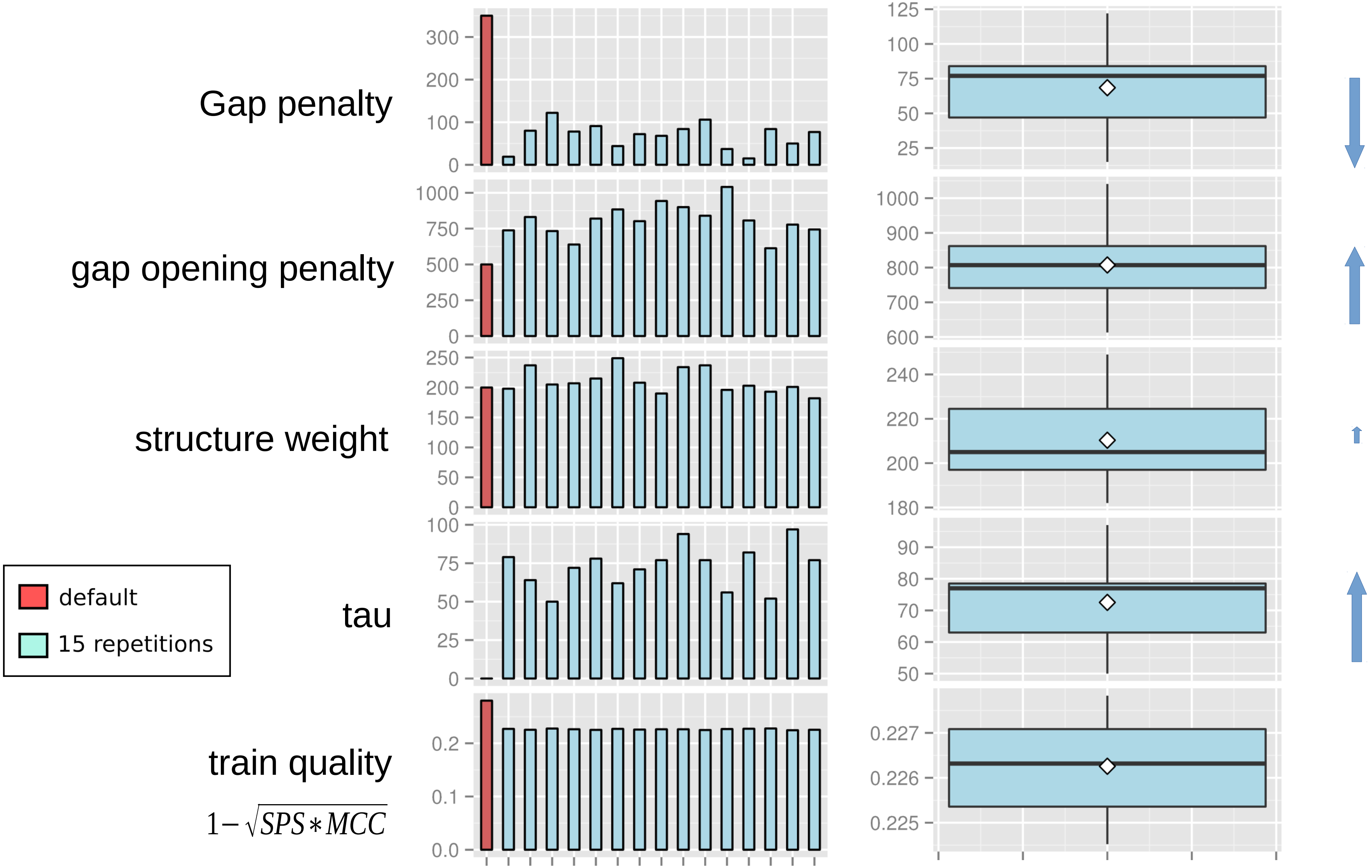
- Take mean runtime 1.65

[Hutter et.al., “Sequential Model-based Optimization for General Algorithm Configuration”, LION 5. Rome. January 18, 2011]

Global alignment dataset

- Dataset: BRAliBase [Wilm et al., 2006]
  - BRAliBase: Benchmark RNA Alignment dataBASE
- Equal number of instances per family
- K-fold cross validation
  - Showed no overfitting

Optimized parameters and quality



Gap penalty

gap opening penalty

structure weight

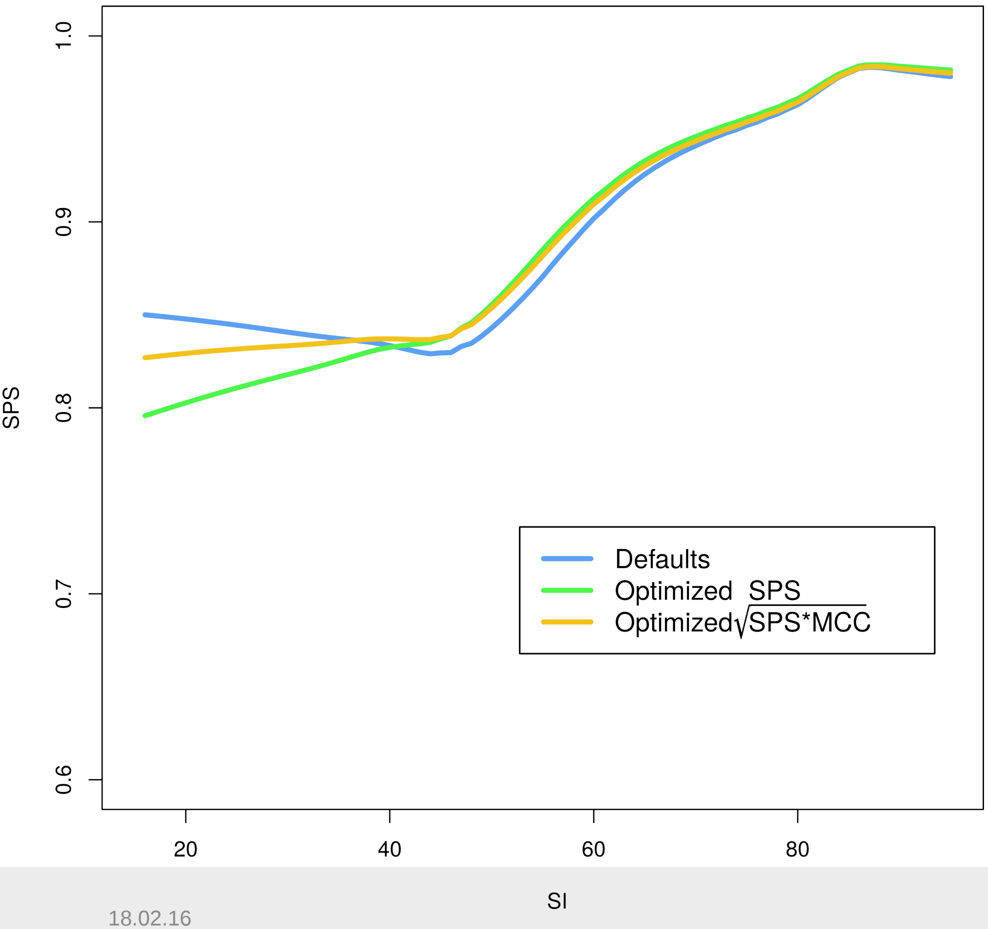
tau

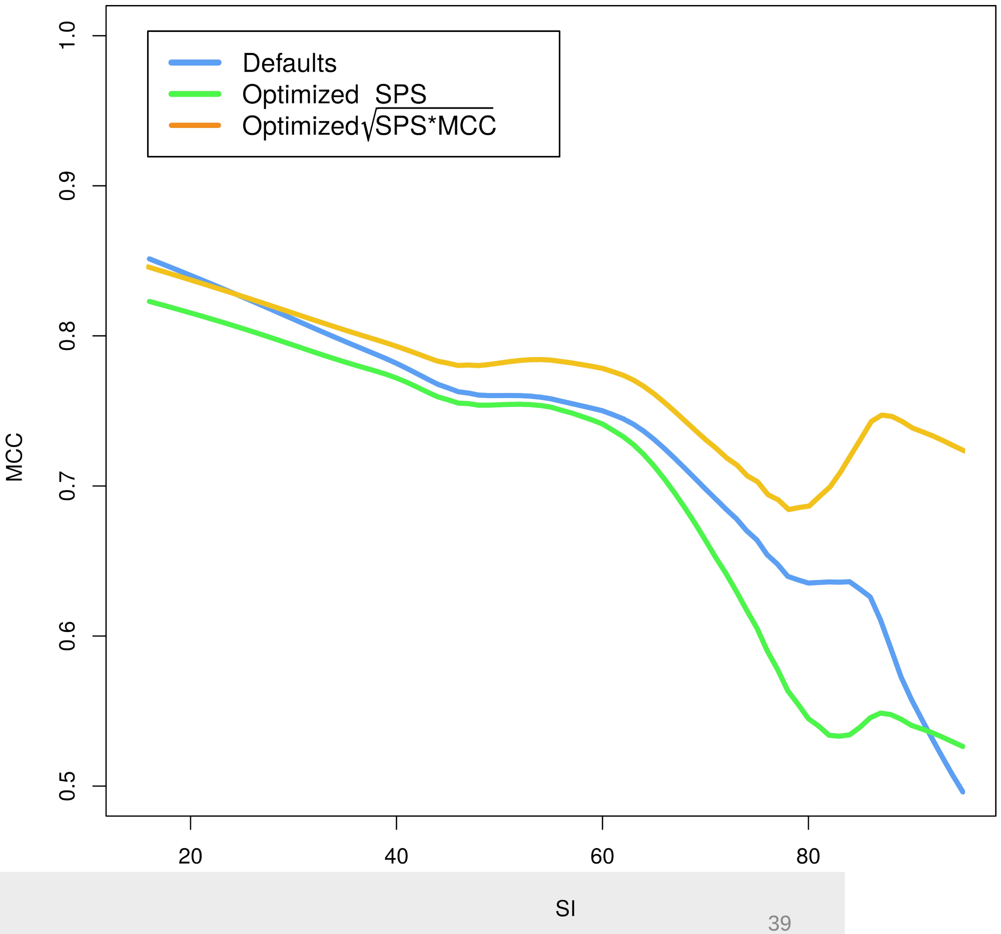
train quality

$1-\sqrt{SPS*MCC}$

Optimized parameter evaluation

quality	train	default	difference	improvement
1 - SPS	0.119	0.144	0.025	17 %
1 - $\sqrt{SPS * MCC}$	0.226	0.280	0.054	19 %


SPS
SI
Defaults
Optimized SPS
Optimized√SPS*MCC


MCC
SI
Defaults
Optimized SPS
Optimized√SPS*MCC