# ALISCORE

## A new method for masking of random sequence similarity (RSS)

Patrick Kück

Forschungsmuseum Koenig, Bonn

# Arthropoda relationships

## Supermatrix tree

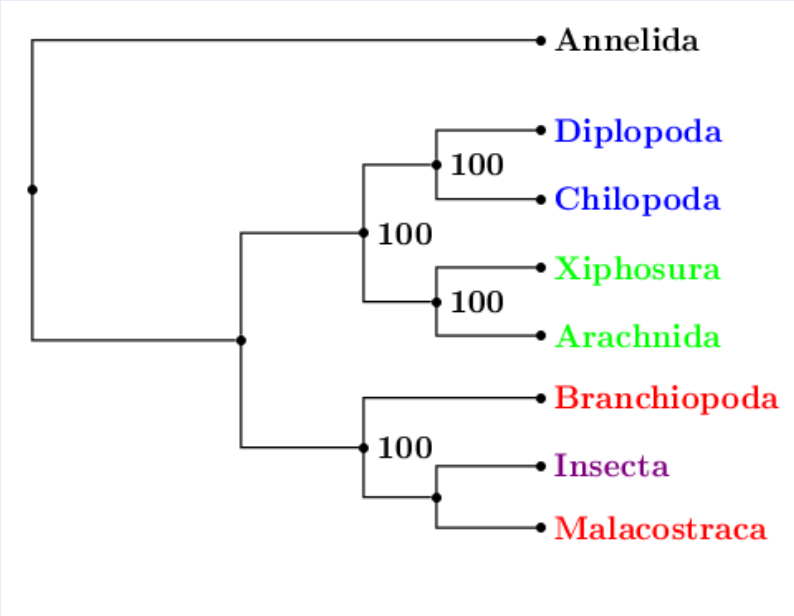



Figure: Bayesian inference
(*Pisani et al. 2004*)

## Used data

- 24 nuclear/mito genes
- 16 taxa, length $>$ 20.000 aa
- 300.000 generations

## Taxon coding

- Outgroup •
- Chelicerata •
- Crustacea •
- Myriapoda •
- Insecta •

# Arthropoda relationships

## Bayesian Inference

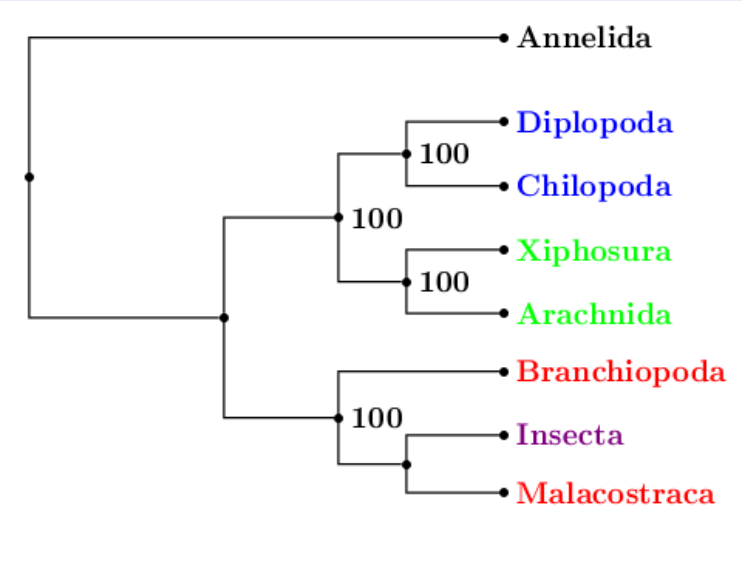



Figure: *Pisani et al. 2004*

## Neighbornet

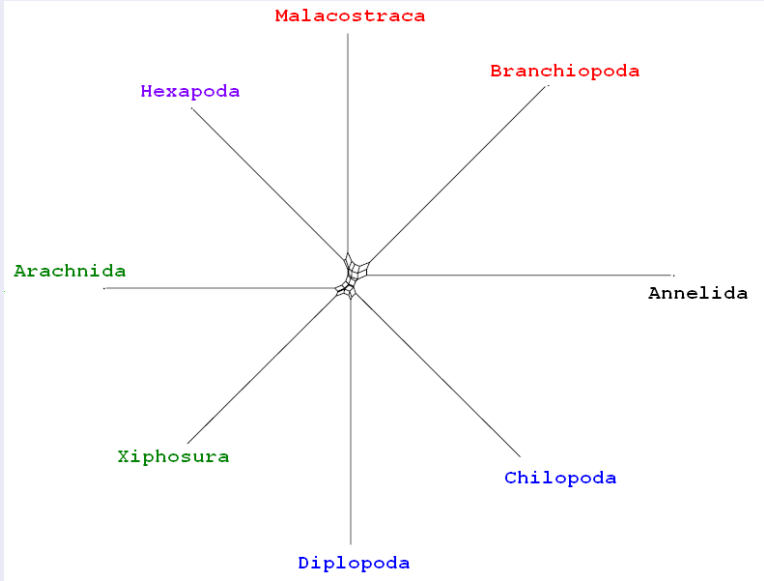



Figure: *Wägele & Mayr 2007*

# Arthropoda relationships

## SAMS

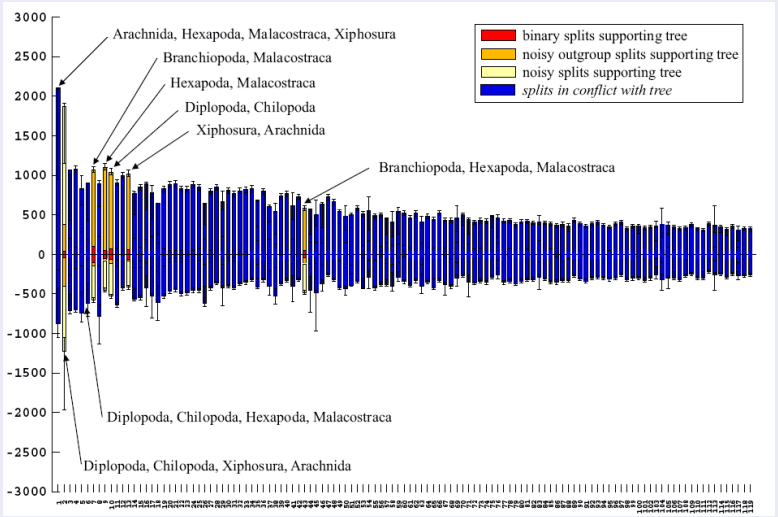



Figure: *Wägele & Mayr 2007*

## Neighbornet

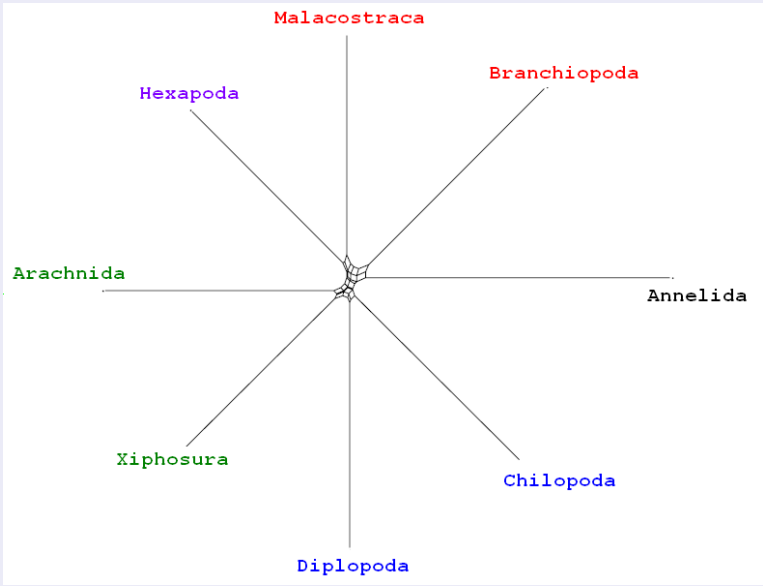


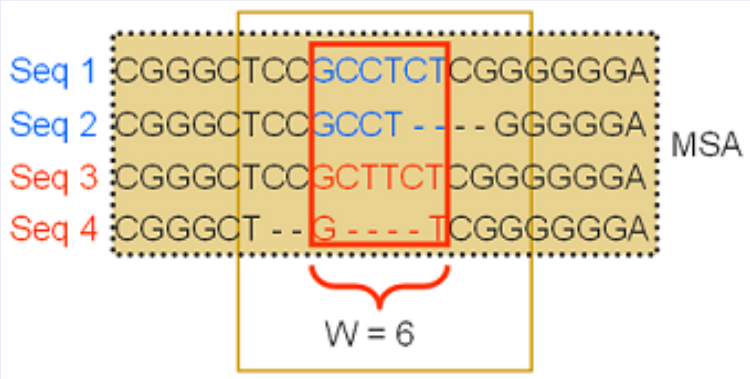

Figure: *Wägele & Mayr 2007*

# ALISCORE *Misof & Misof (in press.)*

## Features

- Sliding window and MC resampling approach

## Pairwise comparisons within single windows

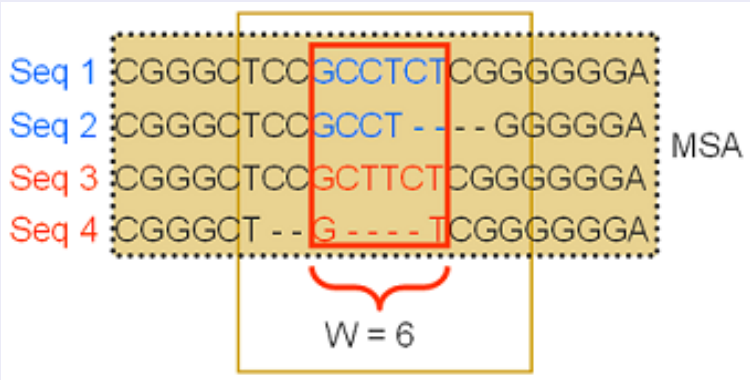


# ALISCORE *Misof & Misof (in press.)*

## Features

- Sliding window and MC resampling approach
- Tree independent

## Pairwise comparisons within single windows

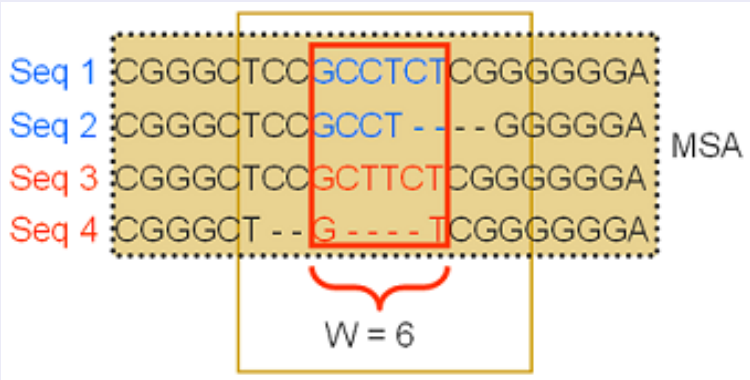


# ALISCORE *Misof & Misof (in press.)*

## Features

- Sliding window and MC resampling approach
- Tree independent
- Without *a priori* rating of parameter space

## Pairwise comparisons within single windows

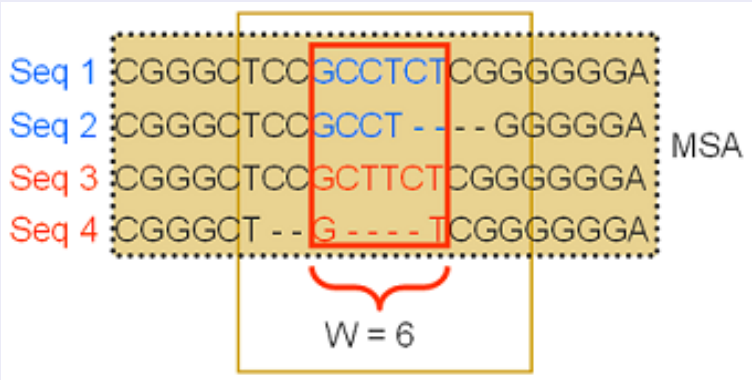


# ALISCORE *Misof & Misof (in press.)*

## Features

- Sliding window and MC resampling approach
- Tree independent
- Without *a priori* rating of parameter space
- Can read nucleotide and protein alignments

## Pairwise comparisons within single windows

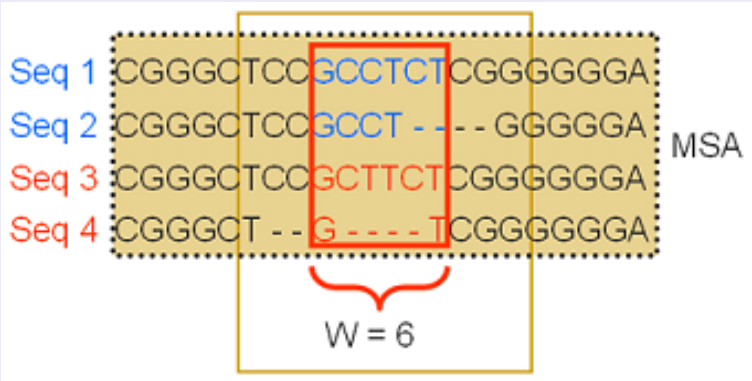


# ALISCORE *Misof & Misof (in press.)*

## Features

- Sliding window and MC resampling approach
- Tree independent
- Without *a priori* rating of parameter space
- Can read nucleotide and protein alignments

## Pairwise comparisons within single windows



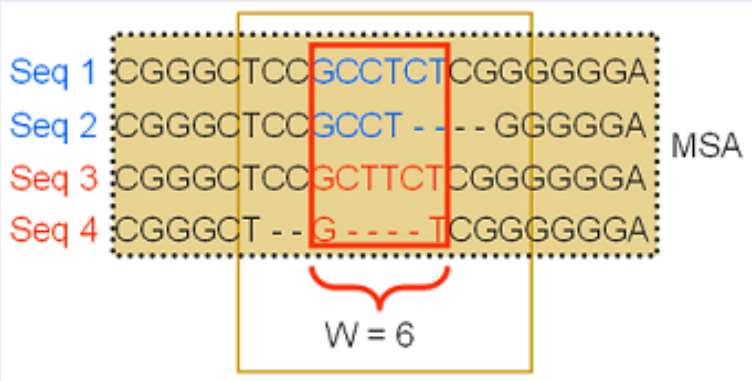

- All pairwise comparisons

# ALISCORE *Misof & Misof (in press.)*

## Features

- Sliding window and MC resampling approach
- Tree independent
- Without *a priori* rating of parameter space
- Can read nucleotide and protein alignments

## Pairwise comparisons within single windows



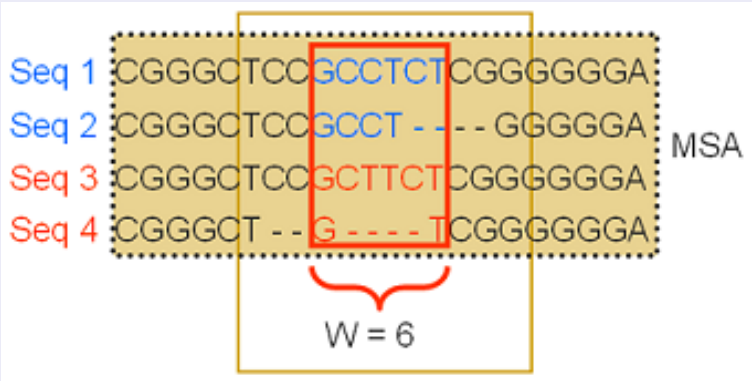

- All pairwise comparisons
- Time expensive
  $P_{xy} = N * (N - 1)/2$

# ALISCORE *Misof & Misof (in press.)*

## Features

- Sliding window and MC resampling approach
- Tree independent
- Without *a priori* rating of parameter space
- Can read nucleotide and protein alignments

## Pairwise comparisons within single windows




- All pairwise comparisons
- Time expensive
  $P_{xy} = N * (N - 1)/2$
- NJ tree

# Pairwise similarity score $S_{obs}$

## infer single scores $S_{si}$ between sequences

- 3 substitution matrices for protein data (Aa)

## Scored $S_{si}$ of AA ($Q_{ij}$)

- BLOSUM62 matrix (PAM250/PAM500)
- Summing scores of single site comparisons (k)
  $(i(k), j(k)), \forall\ k \in (1, 2, \ldots, L)$
- Objective function:
  $S(k) = \sum_{p=0}^{w-1} Q_{ij}(k+p)$

# Pairwise similarity score $S_{obs}$

## infer single scores $S_{si}$ between sequences

- Special scoring function for nucleotides (Nc)

## $S_{obs} = \sum S_{si}/w$ (Nc)

$$S_{obs} = \sum_{k=1}^{i+(w-1)} \begin{cases} 1 & \text{if } seq1_i \bigcap seq2_i \neq 0 \text{ \& non-degenerate} \\ \frac{1}{2} & \text{if } seq1_i \bigcap seq2_i \neq 0 \text{ \& 2-degenerate} \\ \frac{1}{3} & \text{if } seq1_i \bigcap seq2_i \neq 0 \text{ \& 3-degenerate} \\ \frac{1}{4} & \text{if } seq1_i \bigcap seq2_i \neq 0 \text{ \& } \geq \text{4-degenerate} \\ -1 & \text{if } seq1_i \bigcap seq2_i = 0 \end{cases}$$

# Single comparisons

## In relation to the extent of RSS

- $(w = 4, P_{xy} = 1225)$

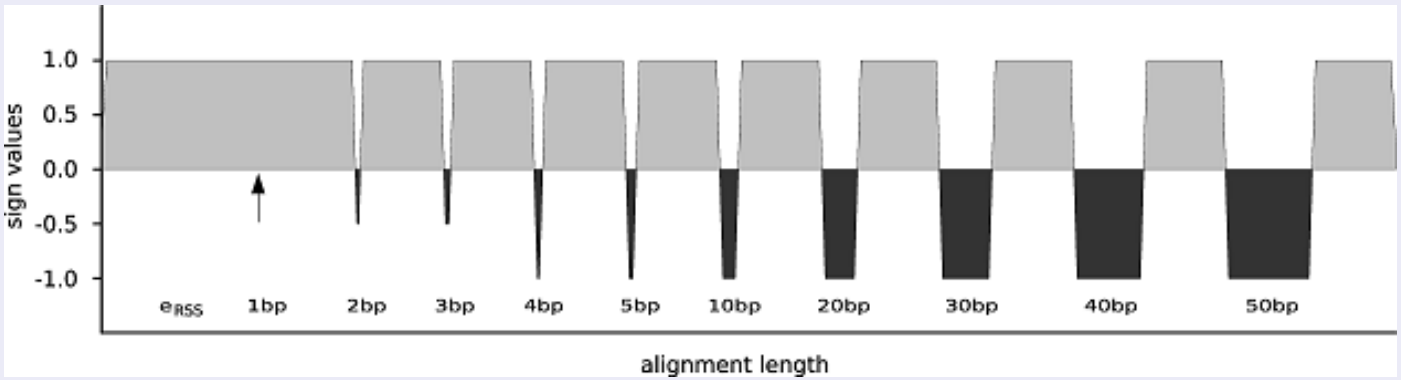



Figure: *Misof & Misof (in press.)*

# Single comparisons

## In relation to the extent of RSS

- $(w = 4, P_{xy} = 1225)$
- Consensus profile identifies RSS greater 1 bp

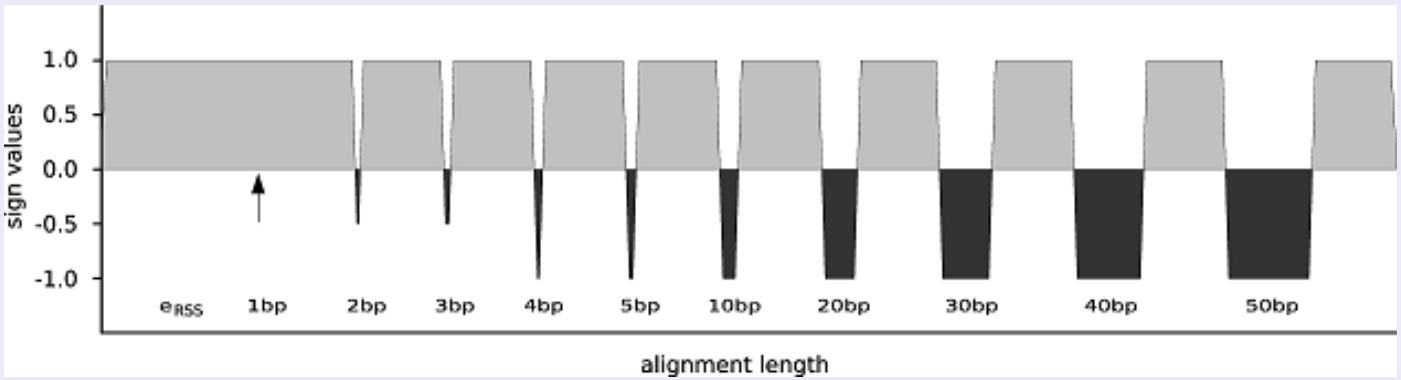



Figure: *Misof & Misof (in press.)*

# Consensus Profiles

## Sensitivity to heterogeneous data

- $w = 4, P_{xy} = 1225, N = 50$

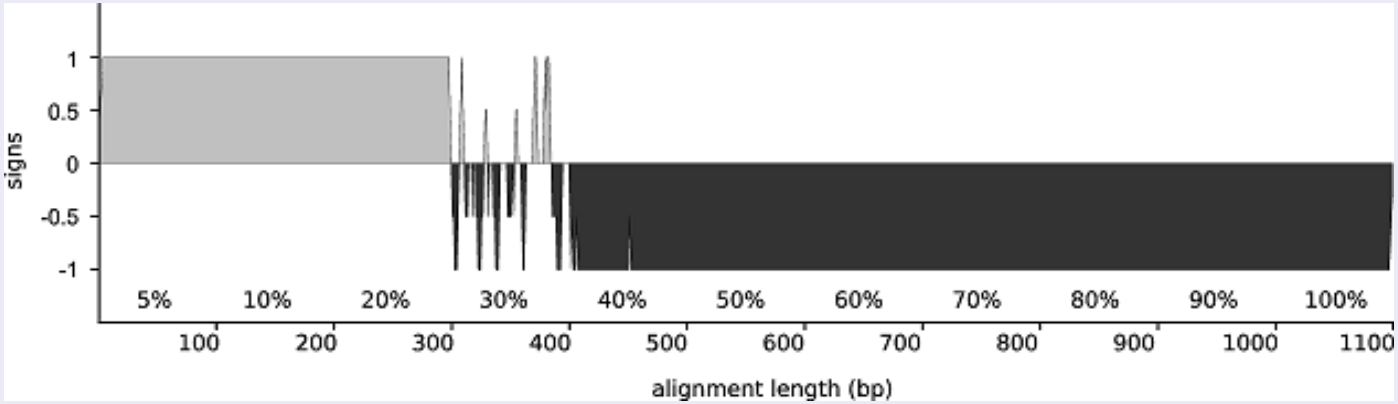



Figure: *Misof & Misof (in press.)*

# Consensus Profiles

## Sensitivity to heterogeneous data

- $w = 4, P_{xy} = 1225, N = 50$
- If random number does not exceed 20%, sites are scored on average $= 0.99$

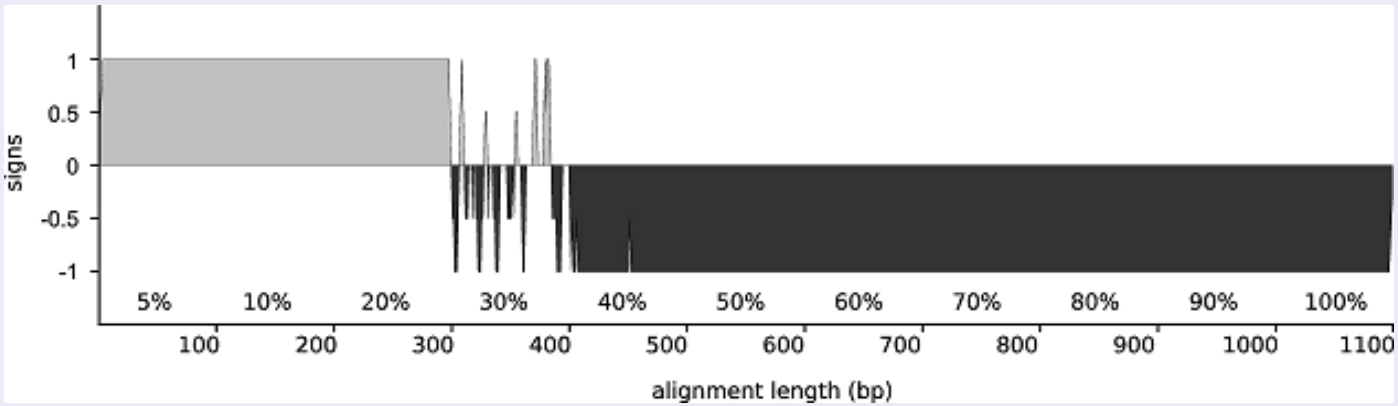



Figure: *Misof & Misof (in press.)*

# Consensus Profiles

## Sensitivity of nodal support

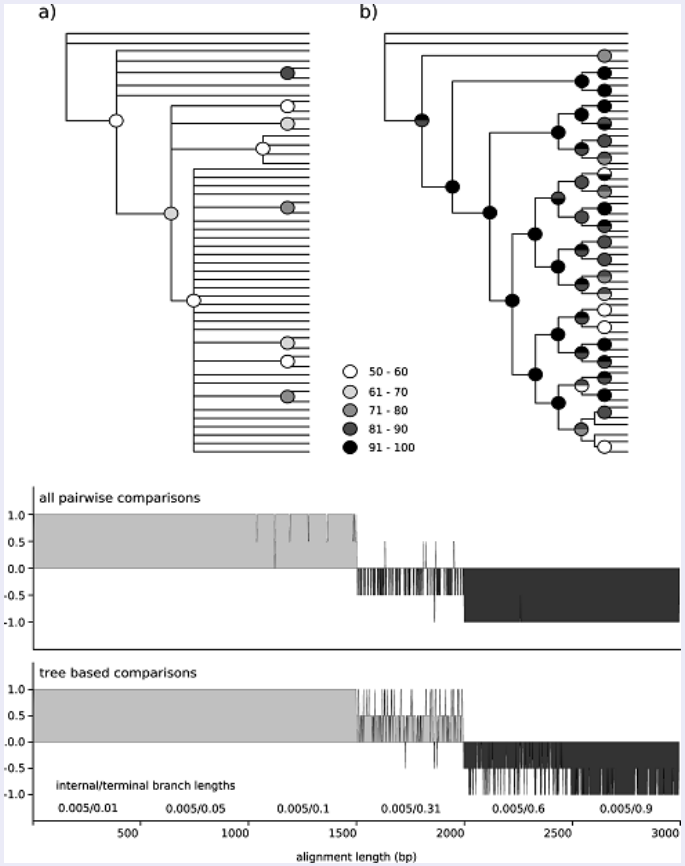



Figure: *Misof & Misof (in press.)*

## Used data

- $N = 50, L = 500$bp

# Consensus Profiles

## Sensitivity of nodal support

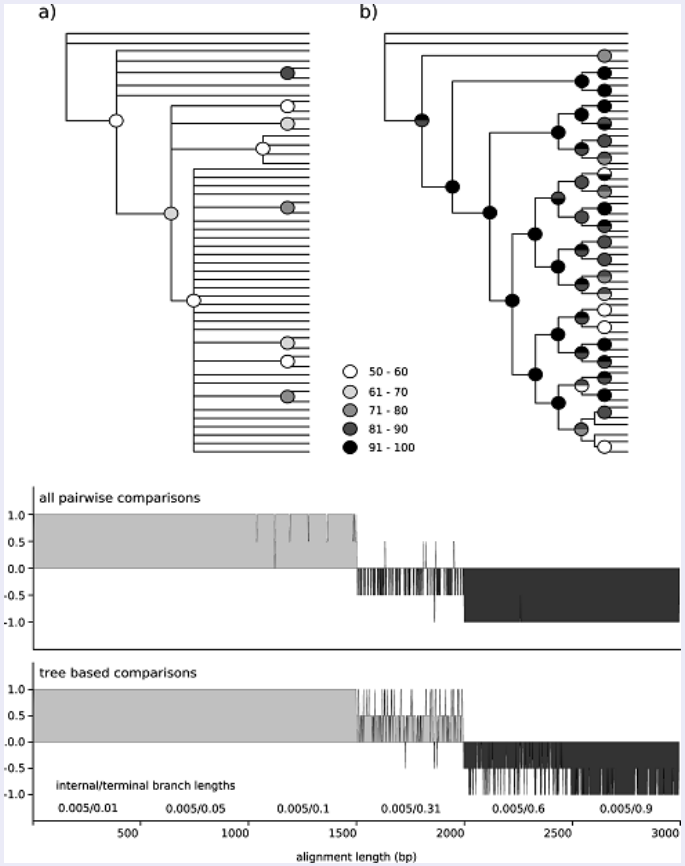



Figure: *Misof & Misof (in press.)*

## Used data

- $N = 50, L = 500$bp
- Variance in internal and terminal branch lengths

# Consensus Profiles

## Sensitivity of nodal support

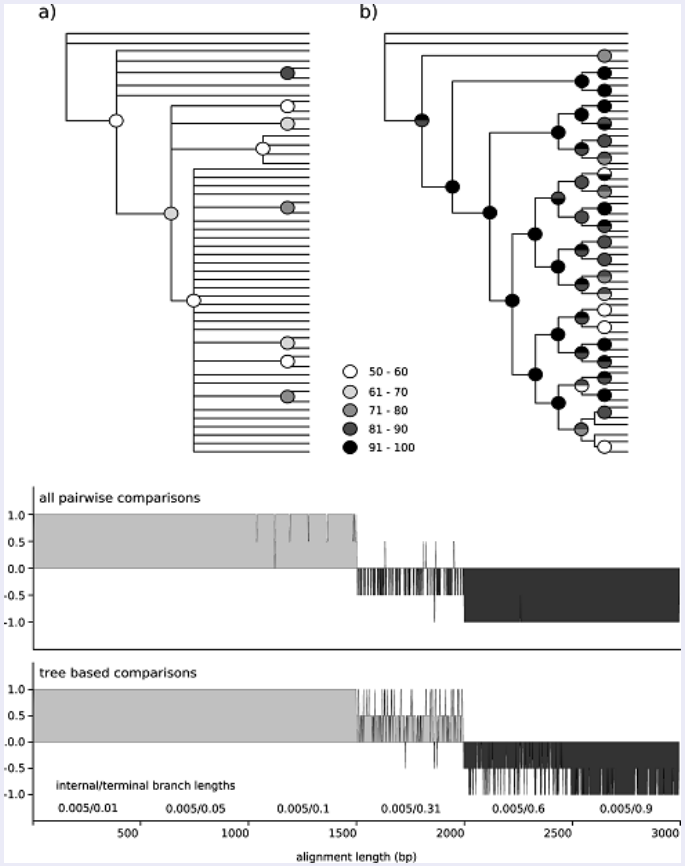



Figure: *Misof & Misof (in press.)*

## Used data

- $N = 50, L = 500$bp
- Variance in internal and terminal branch lengths
- Concatination of simulated data

# Consensus Profiles

## Sensitivity of nodal support

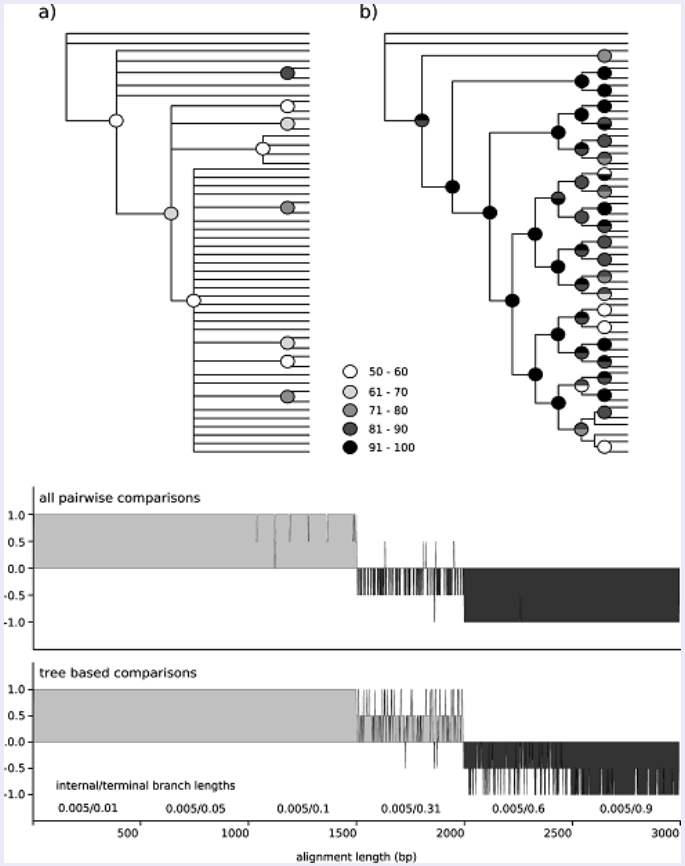



Figure: *Misof & Misof (in press.)*

## Used data

- $N = 50, L = 500$bp
- Variance in internal and terminal branch lengths
- Concatination of simulated data
- a) Parsimony tree

# Consensus Profiles

## Sensitivity of nodal support

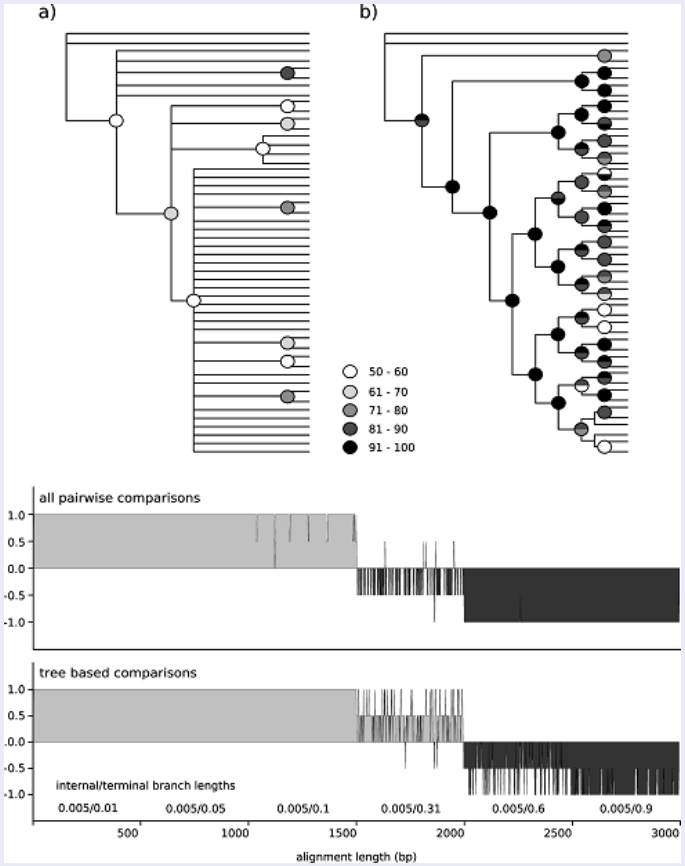



Figure: *Misof & Misof (in press.)*

## Used data

- $N = 50, L = 500$bp
- Variance in internal and terminal branch lengths
- Concatination of simulated data
- a) Parsimony tree
- b) ALISCORE masked

# Material & Methods

## Used data

| | Data | Type | N Genes | Group | N Taxa |
|---|---|---|---|---|---|
| nu I | EST | Aa | 51 | Arthropoda | 26 |
| nu II | rRNA | Nc | 2 | Asellota | 108 |
| mt I | Protein | Aa | 11 | Eukaryota | 17 |
| mt II | Protein | Aa | 5 | Eukaryota | 24 |
| mt III | rRNA | Nc | 2 | Arthropoda | 64 |

## 4 alignment methods (default)

| | |
|---|---|
| CLUSTALX | (*Thompson et al. 1979*) |
| MAFFT | (*Katoh et al. 2005*) |
| MUSCLE | (*Edgar 2004*) |
| T-COFFEE | (*Notredame et al. 2000*) |
| PCMA | (*Pei 2003*) |

## EST data

MSA in total

- $51 * 4 = 204$ MSA

# Material & Methods

## Used data

| | Data | Type | N Genes | Group | N Taxa |
|---|---|---|---|---|---|
| nu I | EST | Aa | 51 | Arthropoda | 26 |
| nu II | rRNA | Nc | 2 | Asellota | 108 |
| mt I | Protein | Aa | 11 | Eukaryota | 17 |
| mt II | Protein | Aa | 5 | Eukaryota | 24 |
| mt III | rRNA | Nc | 2 | Arthropoda | 64 |

## 5 masking settings

| | |
|---|---|
| ALISCORE | w = 6, -N |
| GBLOCKS(none) | stringent |
| GBLOCKS(half) | $>$ 50% gaps per site |
| GBLOCKS(all) | relaxed |
| UNMASKED | original alignment |

## EST data

MSA in total

- $51 * 4 = 204$ MSA

# Material & Methods

## Used data

| | Data | Type | N Genes | Group | N Taxa |
|---|---|---|---|---|---|
| nu I | EST | Aa | 51 | Arthropoda | 26 |
| nu II | rRNA | Nc | 2 | Asellota | 108 |
| mt I | Protein | Aa | 11 | Eukaryota | 17 |
| mt II | Protein | Aa | 5 | Eukaryota | 24 |
| mt III | rRNA | Nc | 2 | Arthropoda | 64 |

## 5 masking settings

| | |
|---|---|
| ALISCORE | w = 6, -N |
| GBLOCKS(none) | stringent |
| GBLOCKS(half) | $>$ 50% gaps per site |
| GBLOCKS(all) | relaxed |
| UNMASKED | original alignment |

## EST data

MSA in total

- $51 * 4 = 204$ MSA
- $204 * 4 = 816$ MSA

# Material & Methods

## Used data

| | Data | Type | N Genes | Group | N Taxa |
|---|---|---|---|---|---|
| nu I | EST | Aa | 51 | Arthropoda | 26 |
| nu II | rRNA | Nc | 2 | Asellota | 108 |
| mt I | Protein | Aa | 11 | Eukaryota | 17 |
| mt II | Protein | Aa | 5 | Eukaryota | 24 |
| mt III | rRNA | Nc | 2 | Arthropoda | 64 |

## 5 masking settings

| | |
|---|---|
| ALISCORE | w = 6, -N |
| GBLOCKS(none) | stringent |
| GBLOCKS(half) | $>$ 50% gaps per site |
| GBLOCKS(all) | relaxed |
| UNMASKED | original alignment |

## EST data

MSA in total

- $51 * 4 = 204$ MSA
- $204 * 4 = 816$ MSA
- $816 + 51 = 867$ MSA

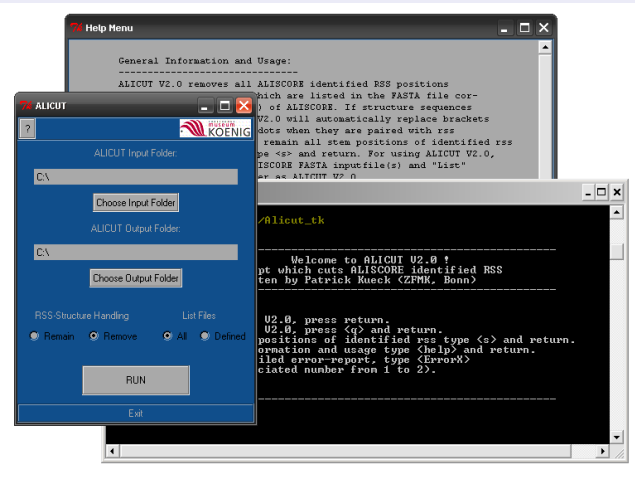
# SPIPES *(Kück)*

## Small pipe written in Perl

$$MSA \rightarrow \begin{cases} \rightarrow ALISCORE/ALICUT & \rightarrow FASconCAT \\ \rightarrow GBLOCKS(none) & \rightarrow FASconCAT \\ \rightarrow GBLOCKS(half) & \rightarrow FASconCAT \\ \rightarrow GBLOCKS(all) & \rightarrow FASconCAT \\ \rightarrow UNMASKED & \rightarrow FASconCAT \end{cases} \rightarrow RAxML$$

## ALICUT (*Kück*)

- Cut of identified RSS
- Considers structure info
- IN: ALISCORE list.txt/.fas
- OUT: *.fas format
- Terminal/GUI- Version



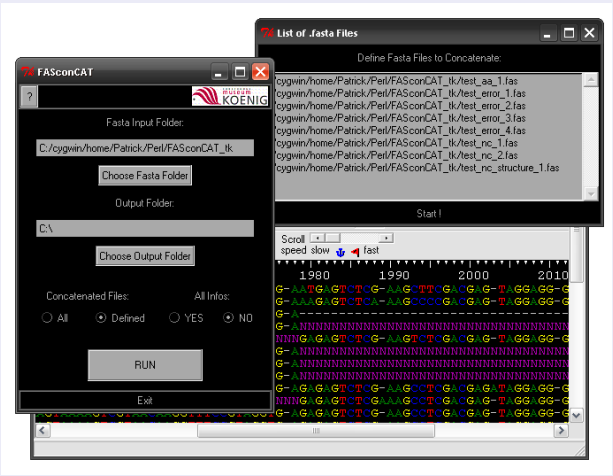
# SPIPES *(Kück)*

## Small pipe written in Perl

$$MSA \rightarrow \begin{cases} \rightarrow ALISCORE/ALICUT & \rightarrow FASconCAT \\ \rightarrow GBLOCKS(none) & \rightarrow FASconCAT \\ \rightarrow GBLOCKS(half) & \rightarrow FASconCAT \\ \rightarrow GBLOCKS(all) & \rightarrow FASconCAT \\ \rightarrow UNMASKED & \rightarrow FASconCAT \end{cases} \rightarrow RAxML$$

## FASconCAT (*Kück*)

- Concatenation of Nc/Aa
- Considers structure info
- IN: *.fas/*.aln format
- OUT: *.fas/*.nex format
- Terminal/GUI- Version



# Identified non-RSS ?

## Remained bp after alignment masking

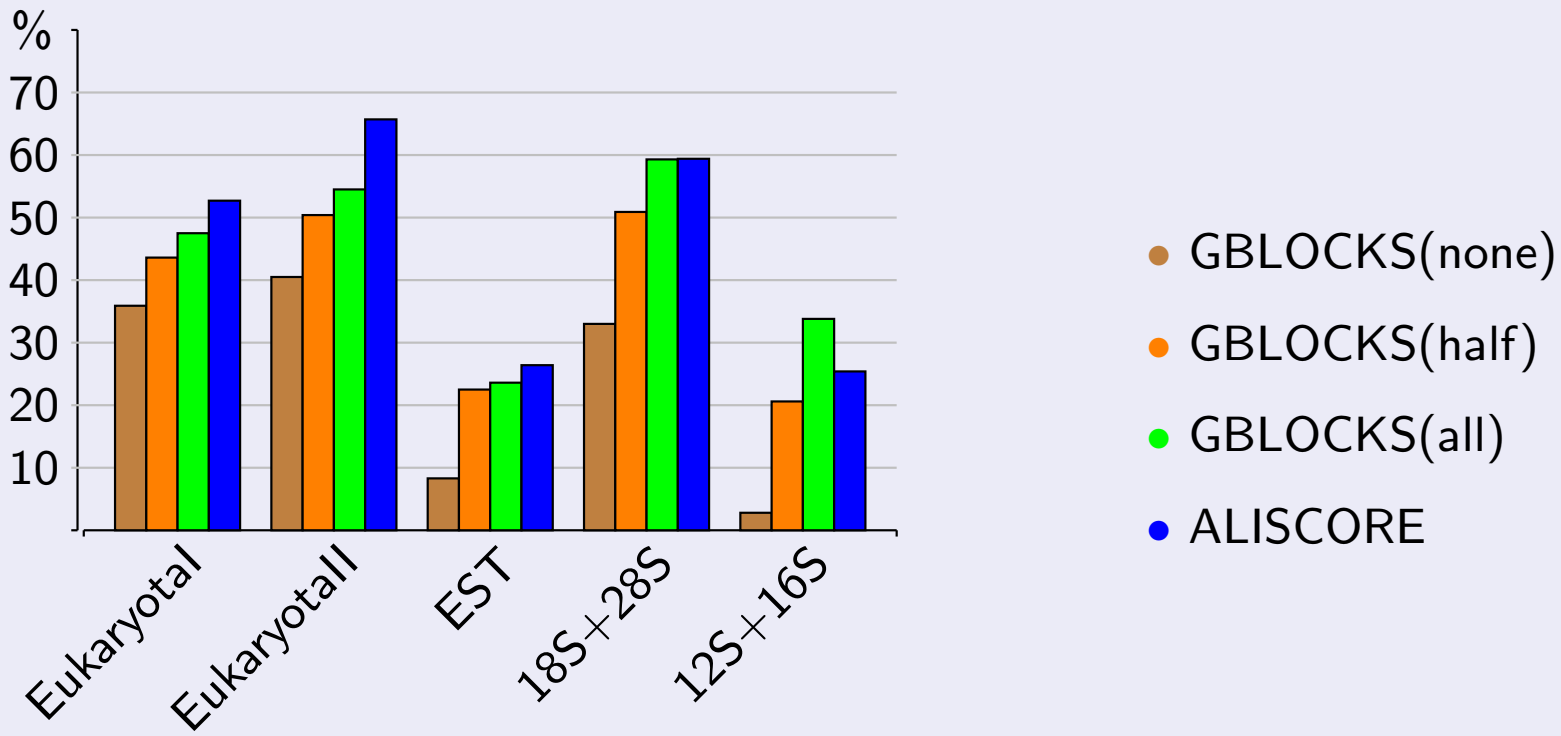
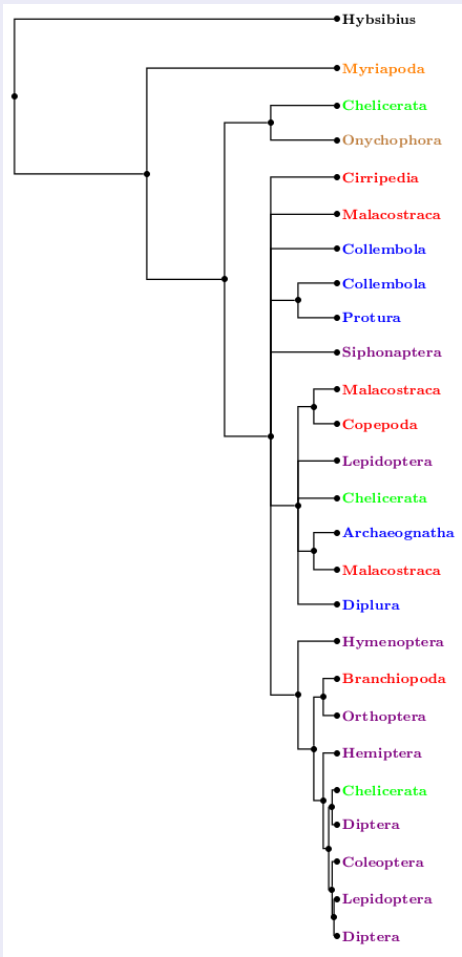



Figure: *Kück et al. (in prep.)*

# Arthropoda relationships (EST's)

## RAxML (UNMASKED)




## Data (T-COFFEE aligned)

- 51 rRNA coding genes
- 26 taxa, length $>$ 35.000 aa
- 100 generations

## Taxon coding

- Outgroup •
- Myriapoda •
- Chelicerata •
- Onychophora •
- Crustacea •
- Apterygota •
- Pterygota •

# Arthropoda relationships (EST's)

## RAxML (ALISCORE)

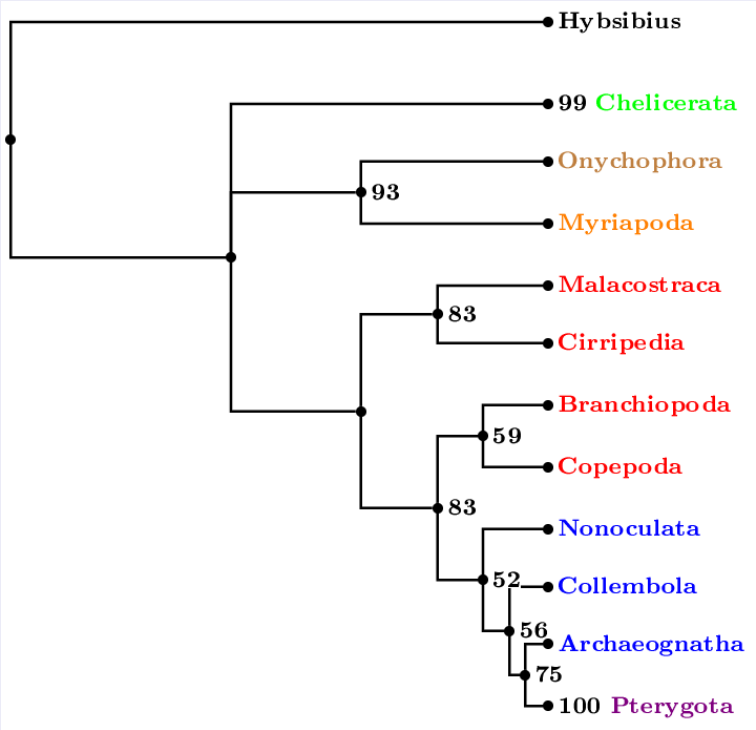



Figure: *Kück et al. (in prep.)*

## Data (T-COFFEE aligned)

- 51 rRNA coding genes
- 26 taxa, length $>$ 35.000 aa
- 100 generations

## Taxon coding

- Outgroup ●
- Myriapoda ●
- Chelicerata ●
- Onychophora ●
- Crustacea ●
- Apterygota ●
- Pterygota ●

# Noise reduction in EST data structure

## ø percentages of resolved nodes after masking

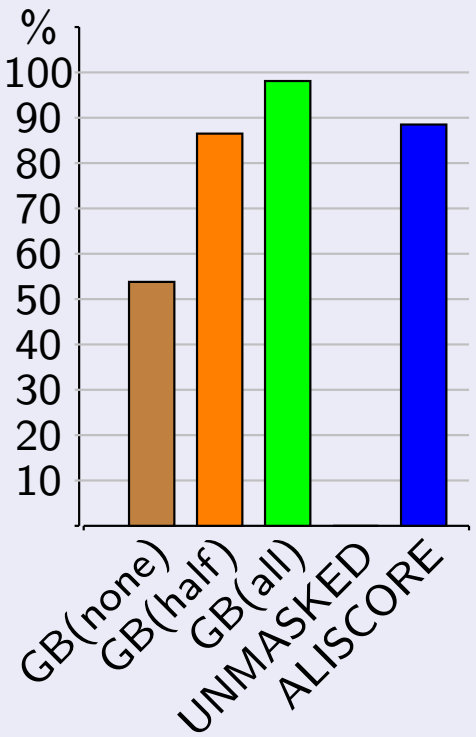



Figure: *Kück et al. (in prep.)*

## ø percentages of bootstrap supports

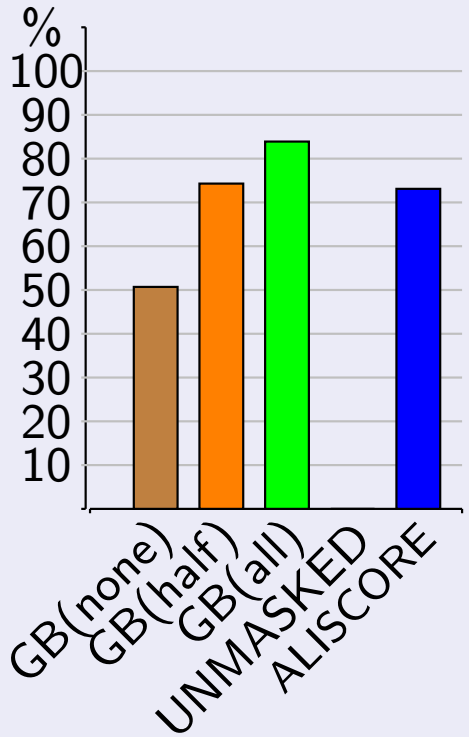



Figure: *Kück et al. (in prep.)*

# Noise reduction in EST data structure

## Neighbornet (UNMASKED)

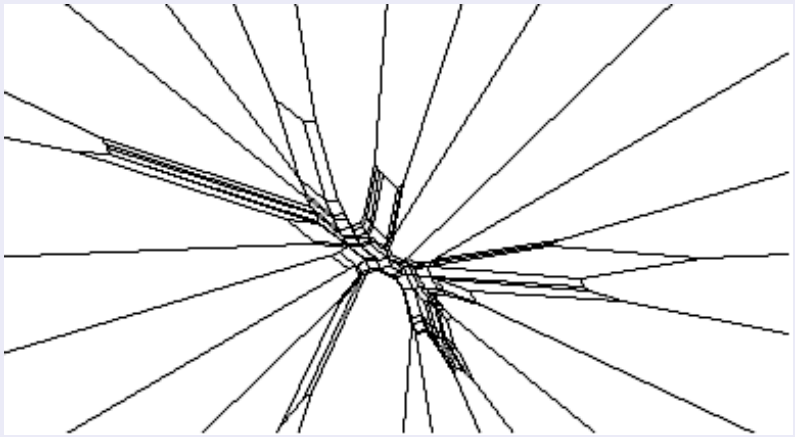


Figure: *Kück et al. (in prep.)*

## Neighbornet (ALISCORE)

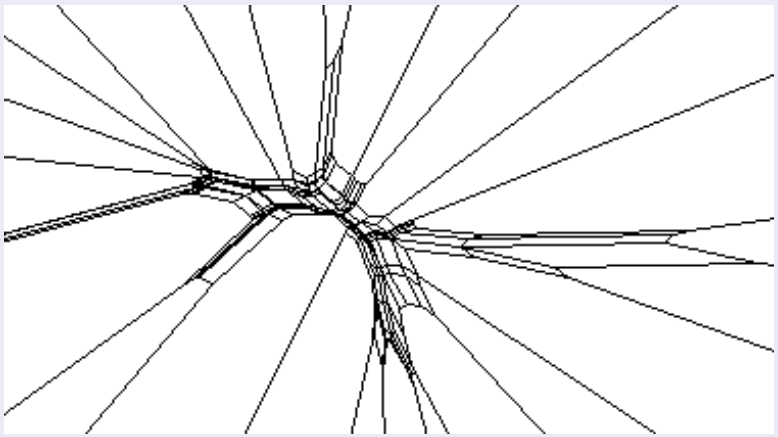


Figure: *Kück et al. (in prep.)*

# Noise reduction in EST data structure

## Topology-distances between best ML trees (TREEDISTANCE)

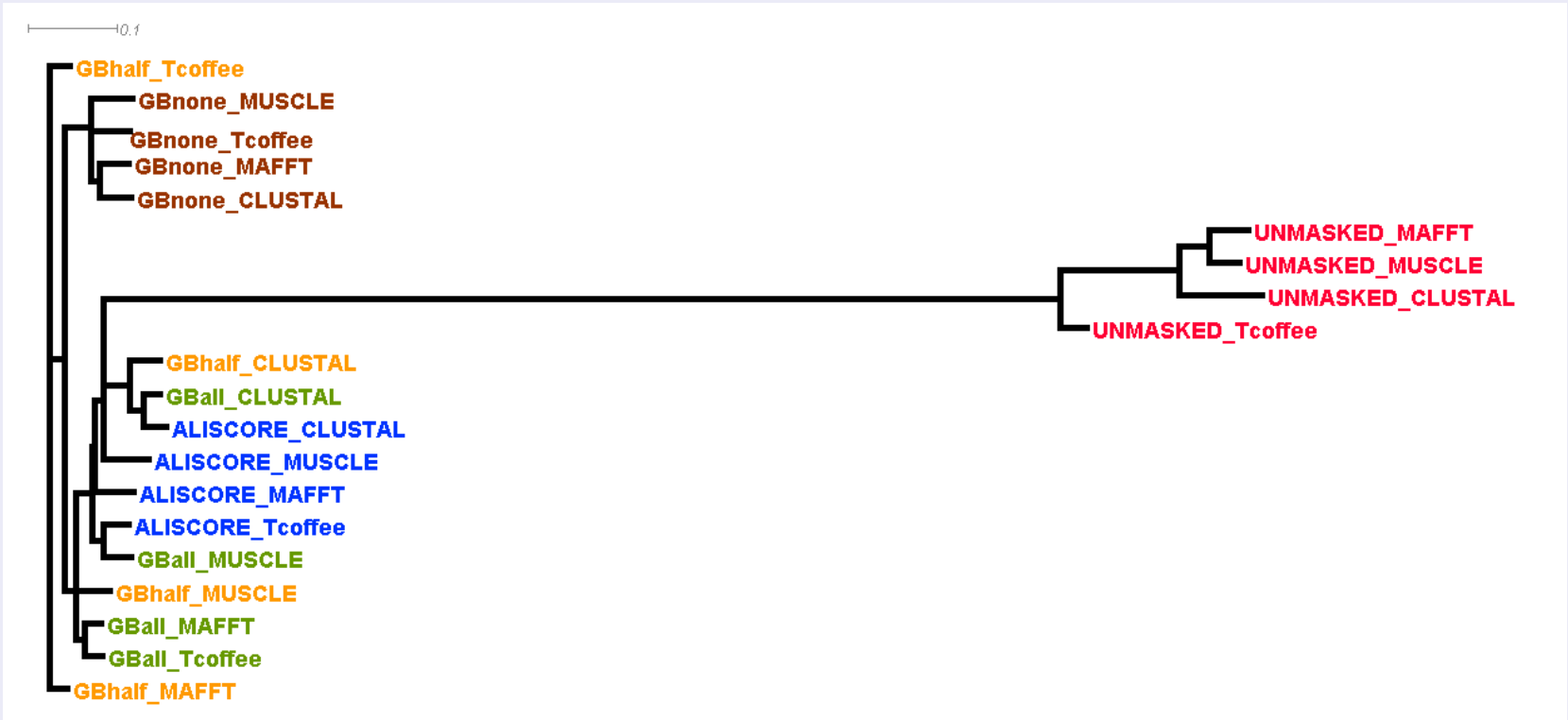



Figure: *Kück et al. (in prep.)*

# Conclusions

## *Kück et al. (in prep.)*

**Masking of randomness in sequence alignments can be improved and leads to better resolved trees**

Patrick Kück *, Karen Meusemann *, Michael Raupach *, Björn v. Reumont *, Wolfgang Wägele * and Bernhard Misof †

*Zoologisches Forschungsmuseum A. Koenig, Bonn, Germany, and †Biozentrum Grindel und Zoologisches Museum, University of Hamburg, Germany

## ALISCORE...

- ...succesfully improves data structure

# Conclusions

## *Kück et al. (in prep.)*

**Masking of randomness in sequence alignments can be improved and leads to better resolved trees**

Patrick Kück *, Karen Meusemann *, Michael Raupach *, Björn v. Reumont *, Wolfgang Wägele * and Bernhard Misof †

*Zoologisches Forschungsmuseum A. Koenig, Bonn, Germany, and †Biozentrum Grindel und Zoologisches Museum, University of Hamburg, Germany

## ALISCORE...

- ...succesfully improves data structure
- ...adapts to heterogeneous base composition

# Conclusions

## *Kück et al. (in prep.)*

**Masking of randomness in sequence alignments can be improved and leads to better resolved trees**

Patrick Kück *, Karen Meusemann *, Michael Raupach *, Björn v. Reumont *, Wolfgang Wägele * and Bernhard Misof †

*Zoologisches Forschungsmuseum A. Koenig, Bonn, Germany, and †Biozentrum Grindel und Zoologisches Museum, University of Hamburg, Germany

## ALISCORE...

- ...succesfully improves data structure
- ...adapts to heterogeneous base composition
- ...adapts to substitution patterns

# Conclusions

## *Kück et al. (in prep.)*

**Masking of randomness in sequence alignments can be improved and leads to better resolved trees**

Patrick Kück *, Karen Meusemann *, Michael Raupach *, Björn v. Reumont *, Wolfgang Wägele * and Bernhard Misof †

*Zoologisches Forschungsmuseum A. Koenig, Bonn, Germany, and †Biozentrum Grindel und Zoologisches Museum, University of Hamburg, Germany

## ALISCORE...

- ...succesfully improves data structure
- ...adapts to heterogeneous base composition
- ...adapts to substitution patterns
- ...identifies consistent RSS already at moderate data size

# Conclusions

## *Kück et al. (in prep.)*

**Masking of randomness in sequence alignments can be improved and leads to better resolved trees**

Patrick Kück *, Karen Meusemann *, Michael Raupach *, Björn v. Reumont *, Wolfgang Wägele * and Bernhard Misof †

*Zoologisches Forschungsmuseum A. Koenig, Bonn, Germany, and †Biozentrum Grindel und Zoologisches Museum, University of Hamburg, Germany

## ALISCORE...

- ...succesfully improves data structure
- ...adapts to heterogeneous base composition
- ...adapts to substitution patterns
- ...identifies consistent RSS already at moderate data size
- ...overlooks random similarity if not present in $\approx 20\%$ of sequences

# Available Programs

## Downloadable under http://www.zfmk.de

- ALISCORE (script & manual)
- ALICUT (terminal & GUI version)
- FASconCAT (terminal & GUI version)
- SPIPES (small pipeline, available soon)

## Aliscore support

- Ali_score@web.de ("help-desk")

# Main topics

## ALISCORE...

- Identification of long branch attracted taxa

# Main topics

## ALISCORE...

- Identification of long branch attracted taxa
- Scoring scheme improvement

# Main topics

## ALISCORE...

- Identification of long branch attracted taxa
- Scoring scheme improvement
- Source code translation into C++

# Main topics

## ALISCORE...

- Identification of long branch attracted taxa
- Scoring scheme improvement
- Source code translation into C++
- Development of the Bonn Package

# Identification of long branch attraction (LBA)

## LBA test tree

# Identification of long branch taxa (LBA)

## Scoring splits between pairwise comparisons

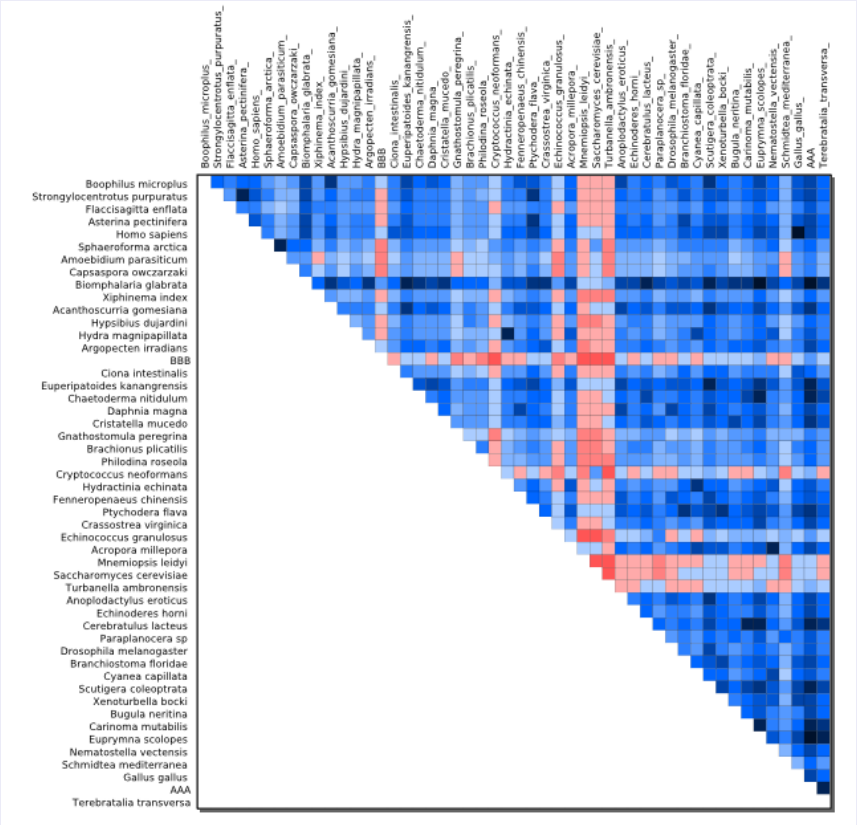



Figure: ALISCORE masked LBA test data

# Identification of long branch taxa (LBA)

## RAxML tree (Eukaryota II)

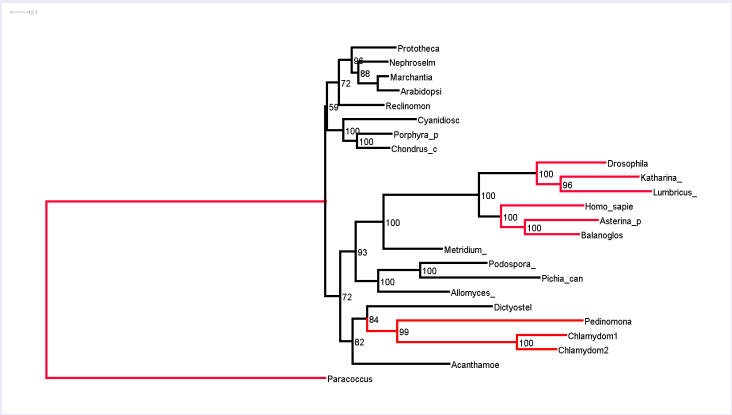



Figure: UNMASKED

## Scoring splits

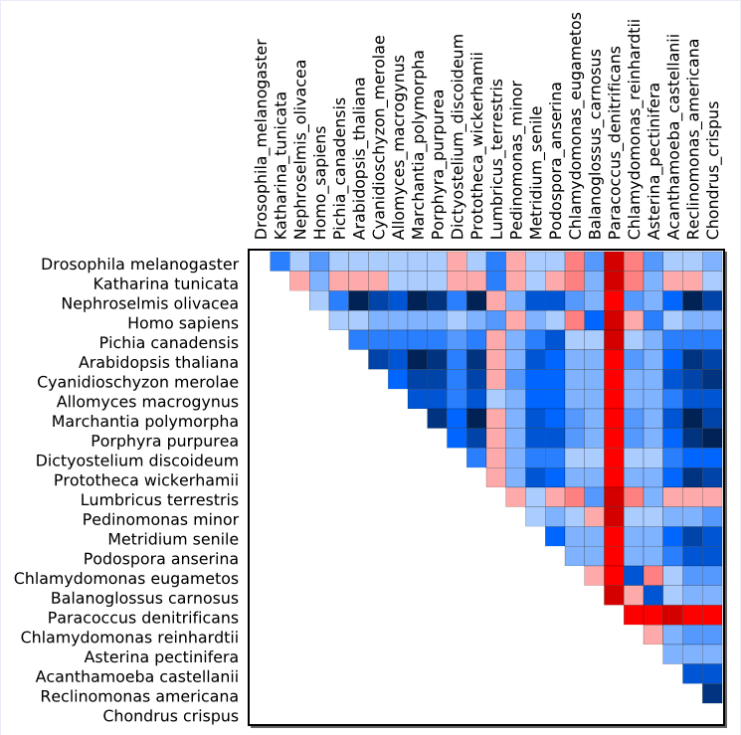



Figure: ALISCORE analysed

# Identification of long branch taxa (LBA)

## RAxML tree (Eukaryota II)

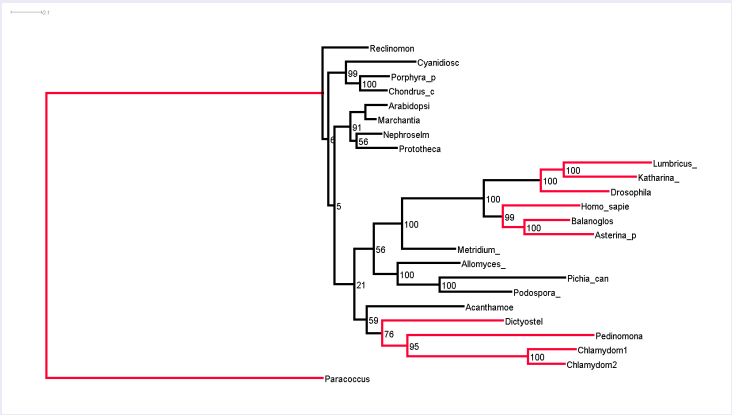



Figure: UNMASKED

## Scoring splits

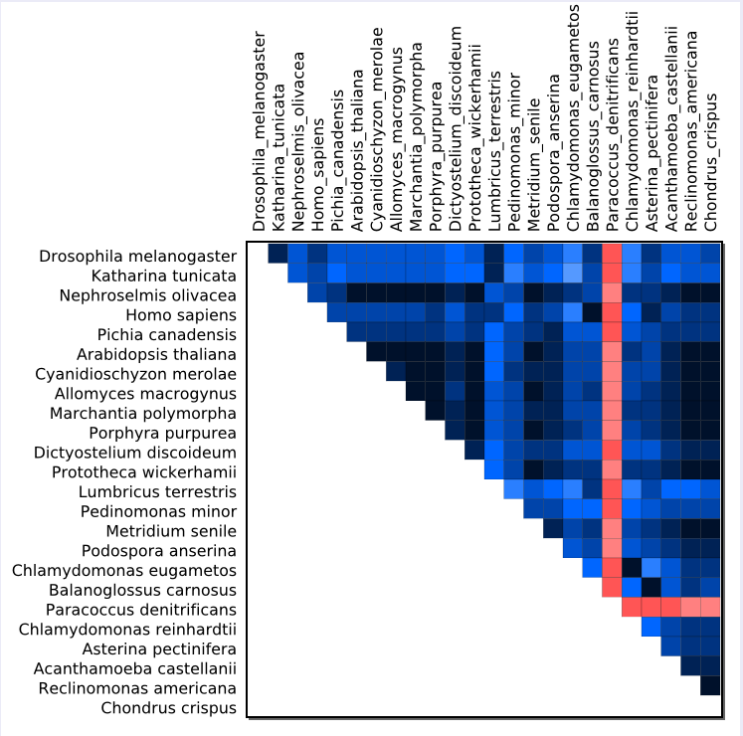



Figure: ALISCORE analysed

# Acknowledgements

## Best thanks to the...

- ...Forschungsmuseum Koenig, Bonn
- ...Deutsche Forschungsgemeinschaft DFG
- ...Leibniz-Gemeinschaft
- ...Karen Meusemann
- ...Prof. Dr. B. Misof
- ...Prof. Dr. J. W. Wägele
- and especially to you !

Deutsche Forschungsgemeinschaft DFG

Forschungsmuseum KOENIG

Leibniz Gemeinschaft

# Consensus Profiles

## Relation to sequence numbers

- Reliability depends on $N_{seq}$ and $P_{xy}$

| $N$ | $e_{\mathrm{RSS}} = 4$ | $e_{\mathrm{RSS}} = 10$ |
|---|---|---|
| 2 | 0.32/0.87* | 0.27/1.00 |
| 5 | 0.58/1.00 | 0.55/1.00 |
| 10 | 0.58/1.00 | 0.65/1.00 |
| 20 | 0.89/1.00 | 0.88/1.00 |
| 30 | 0.94/1.00 | 0.96/1.00 |
| 40 | 0.99/1.00 | 1.00/1.00 |
| 50 | 1.00/1.00 | 1.00/1.00 |

# Consensus Profiles

## Relation to sequence numbers

- Reliability depends on $N_{seq}$ and $P_{xy}$
- $N > 30$, profile reliable for size and position of RSS

| $N$ | $e_{\mathrm{RSS}} = 4$ | $e_{\mathrm{RSS}} = 10$ |
|---|---|---|
| 2 | 0.32/0.87* | 0.27/1.00 |
| 5 | 0.58/1.00 | 0.55/1.00 |
| 10 | 0.58/1.00 | 0.65/1.00 |
| 20 | 0.89/1.00 | 0.88/1.00 |
| 30 | 0.94/1.00 | 0.96/1.00 |
| 40 | 0.99/1.00 | 1.00/1.00 |
| 50 | 1.00/1.00 | 1.00/1.00 |

# Consensus Profiles

## Relation to sequence numbers

- Reliability depends on $N_{seq}$ and $P_{xy}$
- $N > 30$, profile reliable for size and position of RSS
- Frequencies of correct scoring proportional to $N_{seq}$ and size of RSS

| $N$ | $e_{\mathrm{RSS}} = 4$ | $e_{\mathrm{RSS}} = 10$ |
|---|---|---|
| 2 | 0.32/0.87* | 0.27/1.00 |
| 5 | 0.58/1.00 | 0.55/1.00 |
| 10 | 0.58/1.00 | 0.65/1.00 |
| 20 | 0.89/1.00 | 0.88/1.00 |
| 30 | 0.94/1.00 | 0.96/1.00 |
| 40 | 0.99/1.00 | 1.00/1.00 |
| 50 | 1.00/1.00 | 1.00/1.00 |

# MC resampling approach for Nc

## Increased sampling of sequence pairs

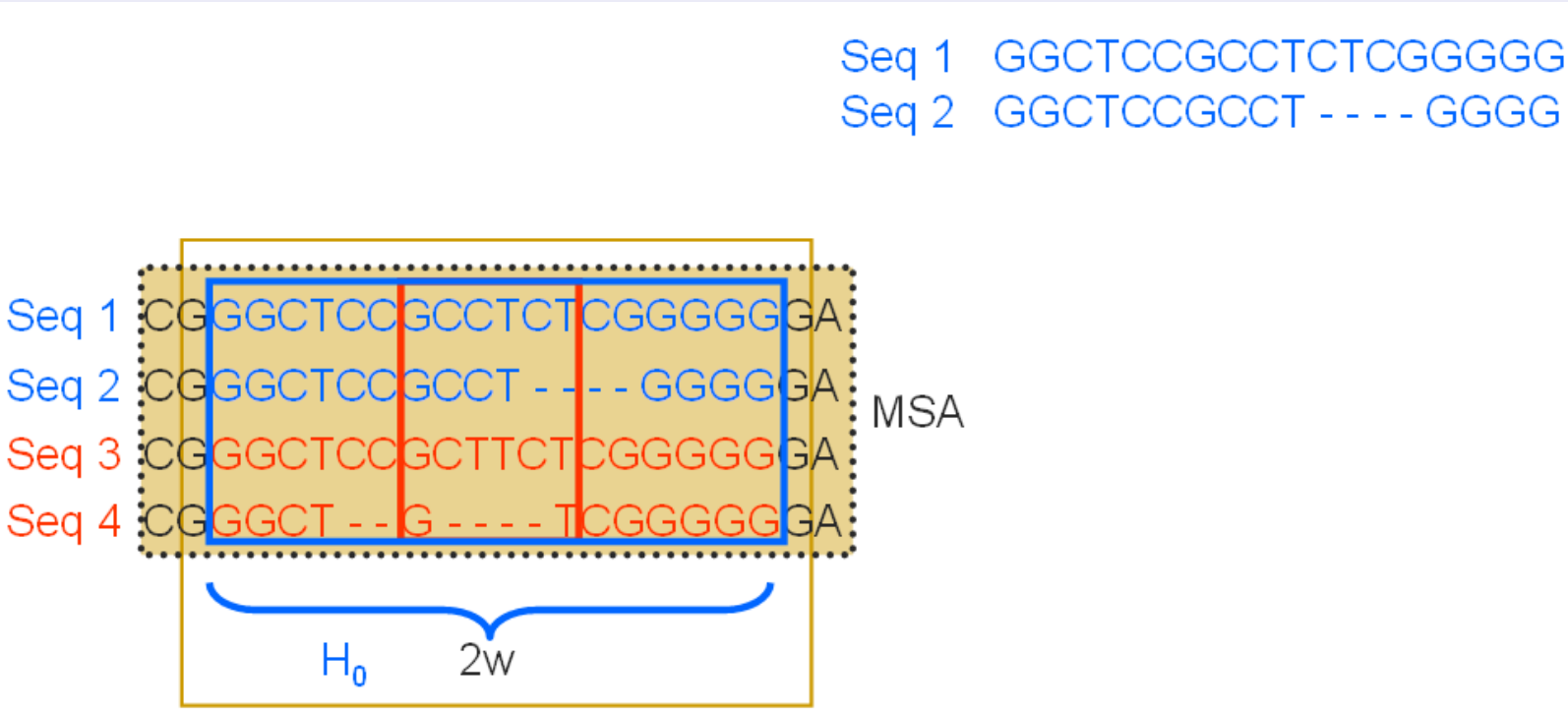



Figure: doublesized window

# MC resampling approach of Nc

## 100 randomly generated sequence pairs

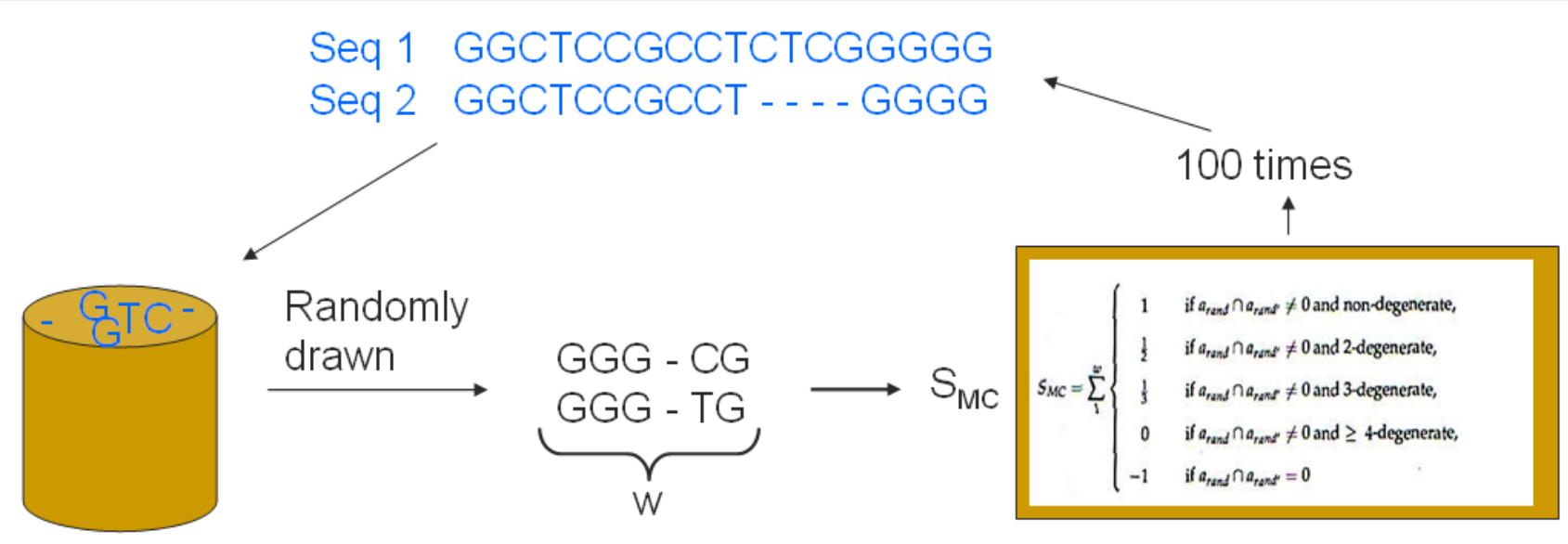



Figure: MC resampling process

# Random resampling of Aa

## Poisson model

- Generating of 100 bootstrap resamples of data matching frequencies

# Random resampling of Aa

## Poisson model

- Generating of 100 bootstrap resamples of data matching frequencies
- Sampling of 100 delete-half bootstrap resamples of each bootstrap replicate

# Null Hypotheses

## $S_{obs} > 95\% \hookrightarrow$ dropped null hypotheses

- Every window bp receives a positive sign

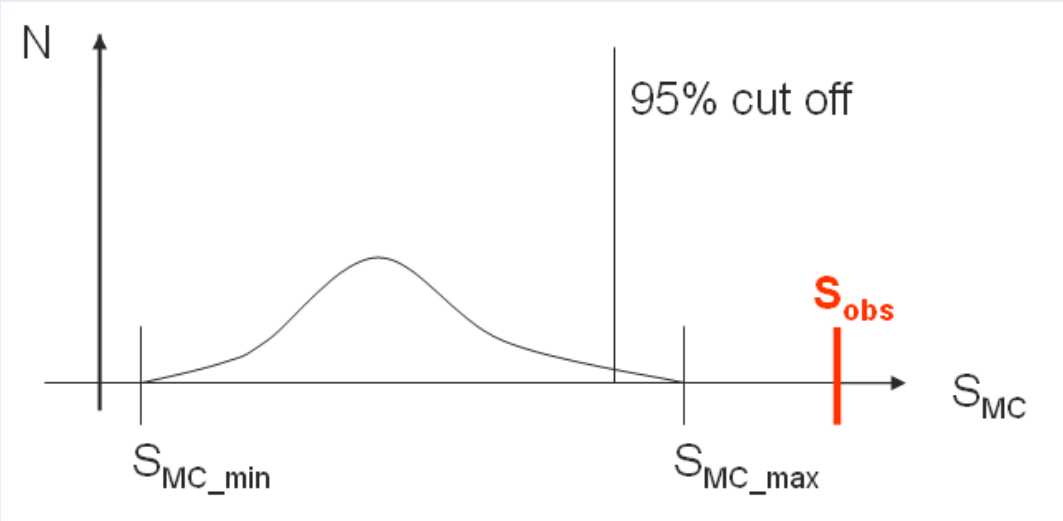



Figure: dropped null hypotheses

# Null Hypotheses

## $S_{obs} > 95\% \hookrightarrow$ dropped null hypotheses

- Every window bp receives a positive sign
- If not dropped, a negative sign

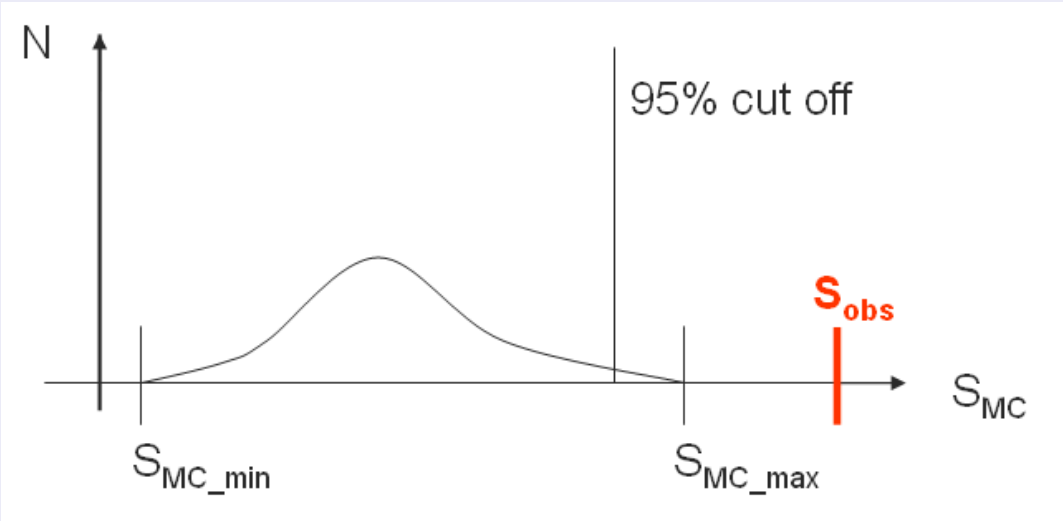



Figure: dropped null hypotheses

# Consensus Profiles

## After collecting all profiles

- Consensus profile inferred by the median of single collected profiles $(-1, 0.66, 1, 1, 1) \hookrightarrow 1$

# Consensus Profiles

## After collecting all profiles

- Consensus profile inferred by the median of single collected profiles $(-1, 0.66, 1, 1, 1) \hookrightarrow 1$
- All pairwise comparisons performed before consensus profile calculated

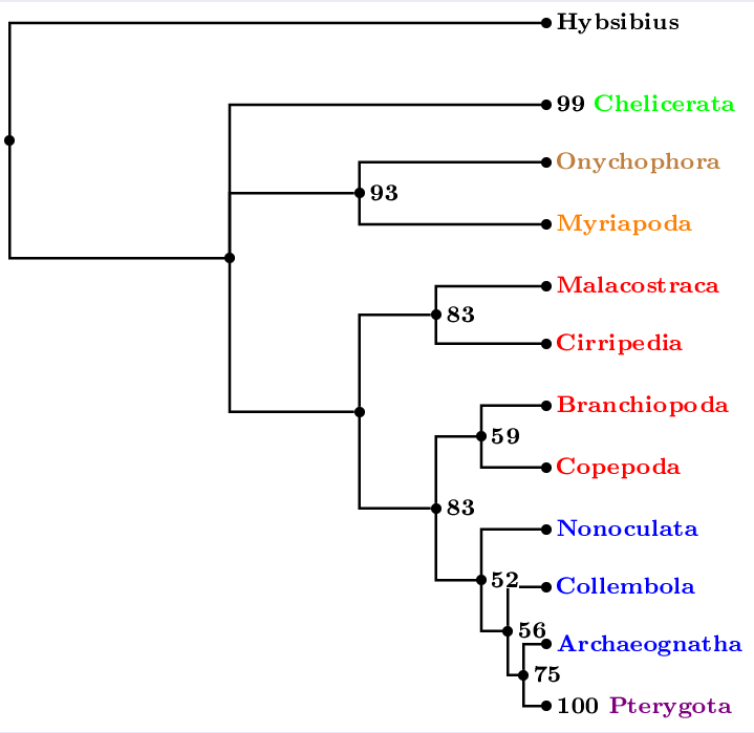
# Consensus Profiles

## After collecting all profiles

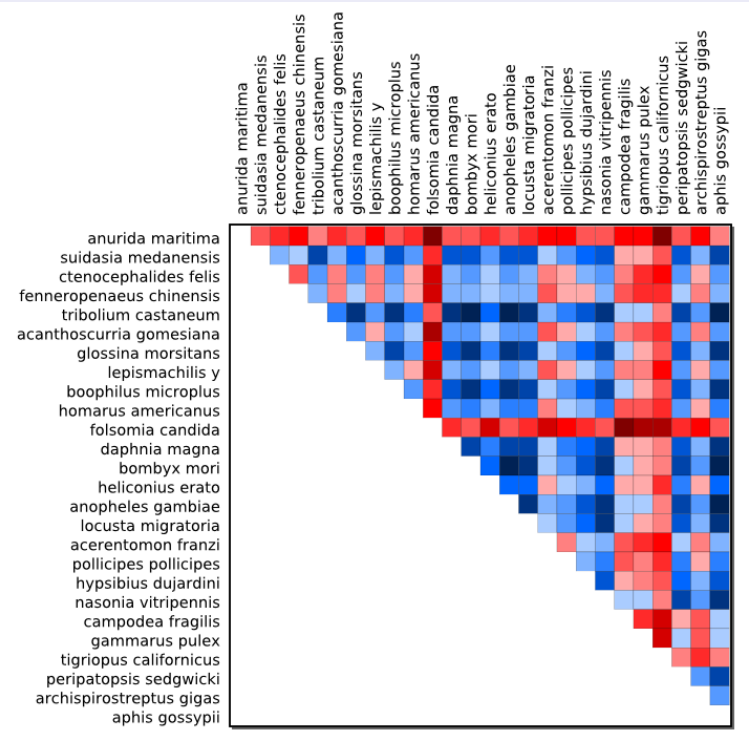
- Consensus profile inferred by the median of single collected profiles $(-1, 0.66, 1, 1, 1) \hookrightarrow 1$
- All pairwise comparisons performed before consensus profile calculated
- Sections of negative signs RSS dominated
  $(-1, -1, -1, -1, 1) \hookrightarrow -1$

# Identification of long branch taxa (LBA)

## RAxML (ALISCORE)

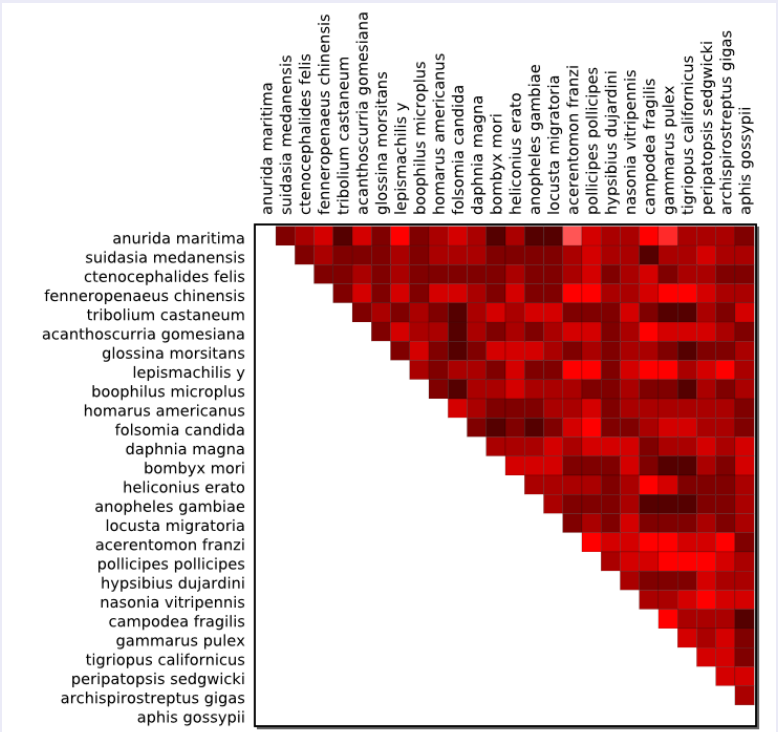



## Pairwise comparisons



# Identification of long branch taxa (LBA)

## UNMASKED




## ALISCORE-masked