RNAalignClust: Sequence-structure-based clustering of multiple alignments

Alexander Junge[1]
Milad Miladi[2]

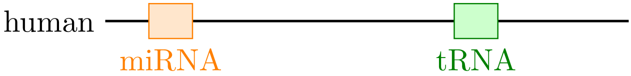
[1] Center for non-coding RNA in Technology and Health (RTH), University of Copenhagen
[2] Bioinformatics Group, University of Freiburg

Computational Analysis of RNA Structure and Function
Benasque, Spain
July 2015

Clustering ncRNA sequences

Annotated human ncRNAs:

- Rfam 11.0 seed alignments: 1.850
- GENCODE v20: 23.989

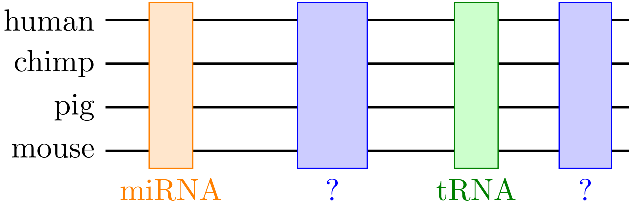


Clustering single RNA sequences identifies ncRNA classes, e.g.,

- GraphClust [Heyne et al., Bioinformatics, 2012]: clustering based on local sequence and structure

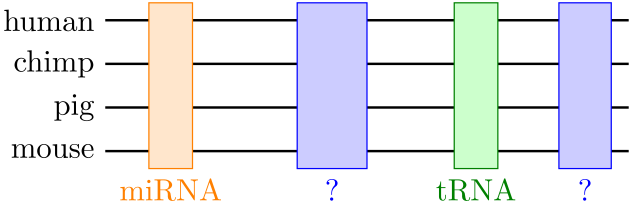
Clustering ncRNA alignments

Screens in genome-wide alignments (using e.g., RNAz, CMfinder) for structured RNAs yield multiple structural alignments of conserved ncRNAs



Clustering ncRNA alignments

Screens in genome-wide alignments (using e.g., RNAz, CMfinder) for structured RNAs yield multiple structural alignments of conserved ncRNAs



We are developing a program for clustering multiple alignments:

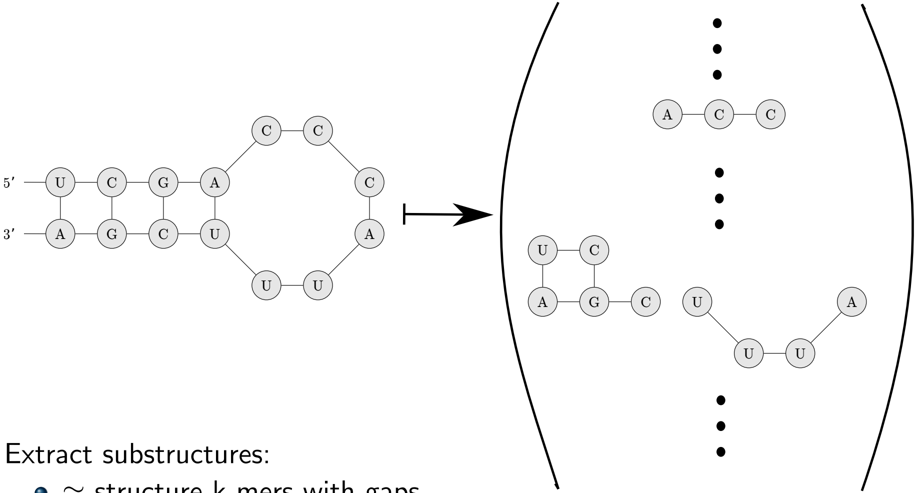
- Improve clustering quality compared to single sequence approaches
- Derive evolutionary conserved sequence and secondary structure

Goals for RNAalignClust

1. Identify **sequence-structure similarities** between ncRNAs
2. Leverage **evolutionary information (covariation)** contained in multiple sequence alignments in clustering
3. Perform **clustering** to:
    - Find new members of existing ncRNA families
    - Unravel new ncRNA families/classes

Identifying similarities of secondary structures

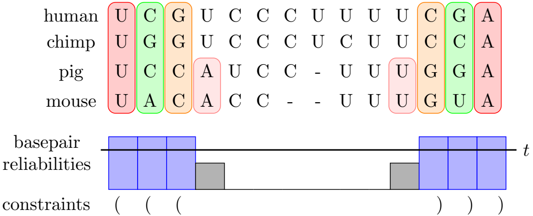
Neighborhood Subgraph Pairwise Distance (NSPD) Kernel used in GraphClust [Heyne et al., Bioinformatics, 2012]



Extract substructures:

- ≈ structure k-mers with gaps
- ncRNAs highly similar if many shared substructures

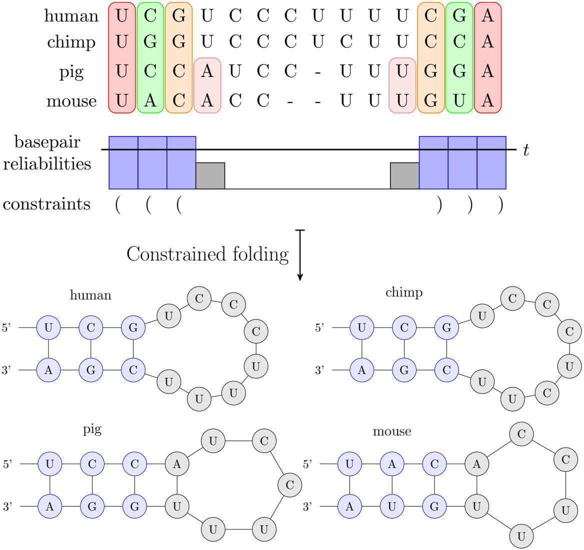
Measuring structure similarity of multiple alignments



```
human   U C G U C C C U U U U C G A
chimp   U G G U C C C U C U U C C A
pig     U C C A U C C - U U U G G A
mouse   U A C A C C - - U U U G U A
```

Evolutionary folding

Measuring structure similarity of multiple alignments



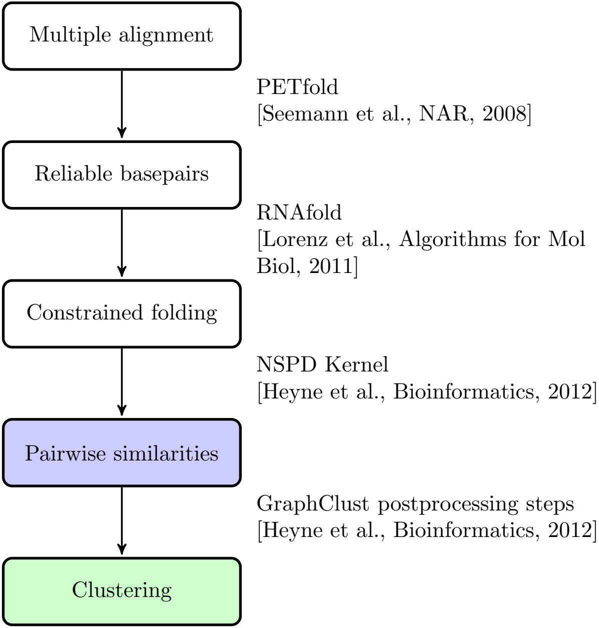
human	U	C	G	U	C	C	C	U	U	U	U	C	G	A
chimp	U	G	G	U	C	C	C	U	C	U	U	C	C	A
pig	U	C	C	A	U	C	C	-	U	U	U	G	G	A
mouse	U	A	C	A	C	C	-	-	U	U	U	G	U	A
constraints	(	(	(									)	)	)

Evolutionary folding

Set of secondary structures

→ use NSPD graph kernel to compare alignments

RNAalignClust - From input alignments to clustering



Multiple alignment

PETfold
[Seemann et al., NAR, 2008]

Reliable basepairs

RNAfold
[Lorenz et al., Algorithms for Mol Biol, 2011]

Constrained folding

NSPD Kernel
[Heyne et al., Bioinformatics, 2012]

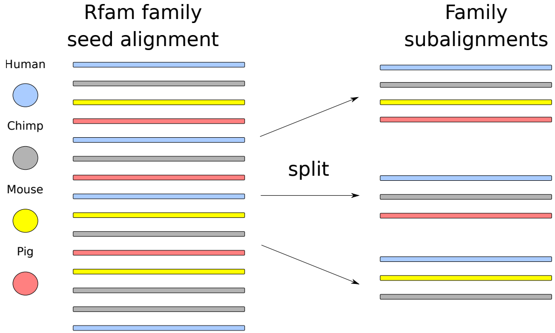
Pairwise similarities

GraphClust postprocessing steps
[Heyne et al., Bioinformatics, 2012]

Clustering

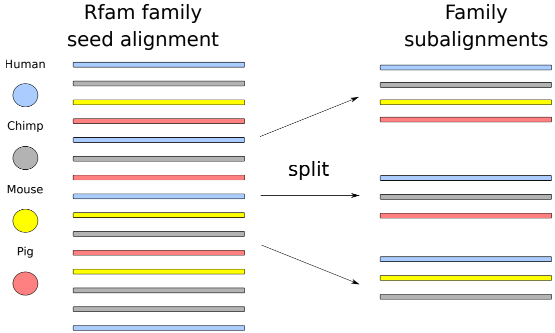
Constructing a benchmark data set

Split Rfam 12 family seed alignments into subalignments.
Similar sequences from *different* species form a subalignment.



Constructing a benchmark data set

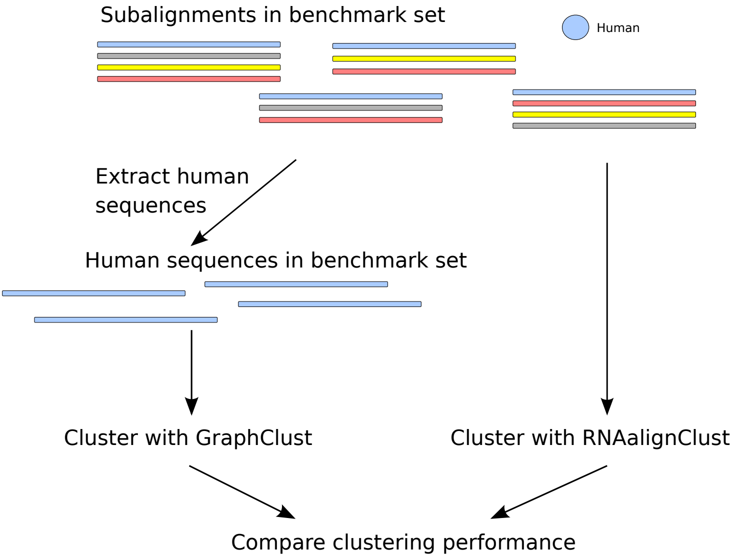
Split Rfam 12 family seed alignments into subalignments.
Similar sequences from *different* species form a subalignment.



1. Benchmark contains subalignments from all Rfam families
2. Each subalignment contains one human sequence

A good clustering puts all subalignments from same Rfam family in one cluster and does not mix families.

Comparing sequence to alignment clustering



Subalignments in benchmark set

● Human

Extract human sequences

Human sequences in benchmark set

Cluster with GraphClust

Cluster with RNAalignClust

Compare clustering performance

Using alignments improves clustering performance

- V-measure is harmonic mean of *homogeneity* and *completeness*
- homogeneity: each cluster contains only members of a single family
- completeness: all members of a given family are in same cluster

	GraphClust	RNAalignClust
V-measure	0.871	0.909

Using alignments improves clustering performance

- $a =$ number. of object pairs from same family correctly assigned to same cluster
- $b =$ number of object pairs from different families correctly assigned to different clusters

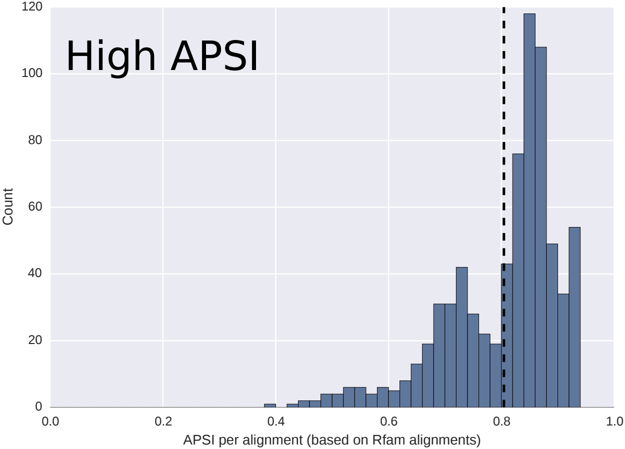
$$\text{Rand Index} = \frac{a+b}{\binom{n}{2}}$$

- $n =$ number of alignments
- Adjusted Rand Index is R adjusted for chance

	GraphClust	RNAalignClust
Adjuste Rand Index	0.672	0.887

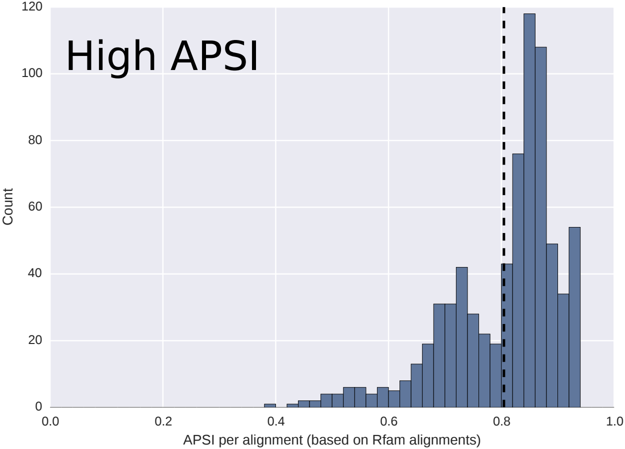
Low covariation in the benchmark data set

The benchmark data set has high average pairwise sequence identity (APSI) in the alignments



Low covariation in the benchmark data set

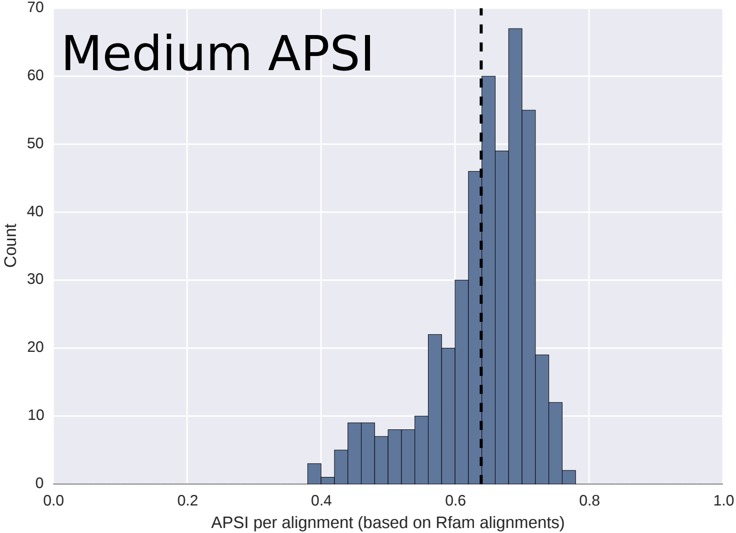
The benchmark data set has high average pairwise sequence identity (APSI) in the alignments



→ limit APSI to study effect of covariation on clustering performance

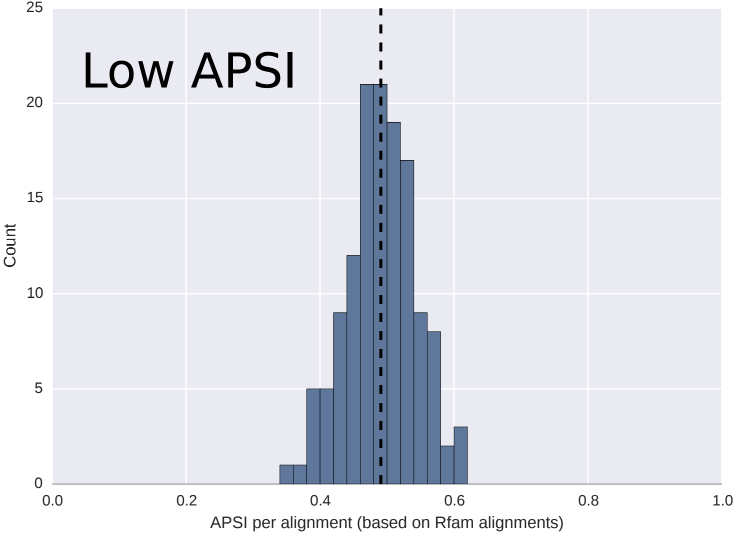
More benchmark sets with different degrees of covariation

Create 2 additional benchmark data set with bounded APSI in alignments



More benchmark sets with different degrees of covariation

Create 2 additional benchmark data set with bounded APSI in alignments



More covariation improves alignment clustering

V-measure

	Mean APSI	Covariation	GraphClust	RNAalignClust
High APSI	0.81	Low	0.87	0.91
Medium APSI	0.62	Medium	0.86	0.91
Low APSI	0.49	High	0.85	0.94

More covariation improves alignment clustering

V-measure

	Mean APSI	Covariation	GraphClust	RNAalignClust
High APSI	0.81	Low	0.87	0.91
Medium APSI	0.62	Medium	0.86	0.91
Low APSI	0.49	High	0.85	0.94

Adjusted Rand Index

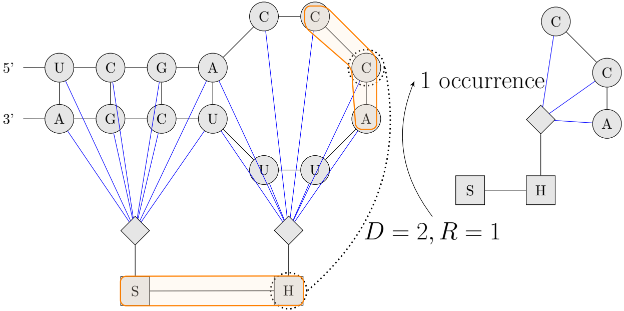
	Mean APSI	Covariation	GraphClust	RNAalignClust
High APSI	0.81	Low	0.67	0.89
Medium APSI	0.62	Medium	0.70	0.95
Low APSI	0.49	High	0.72	0.99

Ongoing Work

- **Additional** benchmark data sets
- **Fine tune parameters**
  - Compare different clustering algorithms/postprocessing steps
- **Genome-scale clustering** of potential ncRNAs

Ongoing work - Extended graph representation

as e.g. GraphProt [Maticzka et al., Genome Biology, 2014]



Add additional abstract nodes:

- **S**tem
- Loop (**H**airpin, **M**ulti, **I**nternal, **B**ulge)
- **E**xternal regions

Extract subgraphs at

- Radius **R**
- Distance **D**

→ ncRNAs highly similar if many shared substructures

Conclusion

- Similarity function derived from **NSPD Graph Kernel**
- Leverage **evolutionary information** contained in multiple alignments:
  1. Conserved basepairs as **folding constraints**
  2. **Set of secondary structures** represents each alignment
- RNAalignClust has potential to cluster **large** (>100.000) data sets (locality sensitive hashing)

Acknowledgements

UNI FREIBURG

RTH
CENTER FOR NON-CODING RNA
IN TECHNOLOGY AND HEALTH

Bioinformatics Group,
University of Freiburg:

- Milad Miladi
- Fabrizio Costa
- Rolf Backofen

RTH, University of Copenhagen:

- Stefan Seemann
- Jakob Hull Havgaard
- Jan Gorodkin

Funding:
Danish Center for Scientific Computing, Innovation Fund Denmark, Danish Cancer Society

Acknowledgements

UNI FREIBURG

Bioinformatics Group,
University of Freiburg:

- Milad Miladi
- Fabrizio Costa
- Rolf Backofen

RTH
CENTER FOR NON-CODING RNA
IN TECHNOLOGY AND HEALTH

RTH, University of Copenhagen:

- Stefan Seemann
- Jakob Hull Havgaard
- Jan Gorodkin

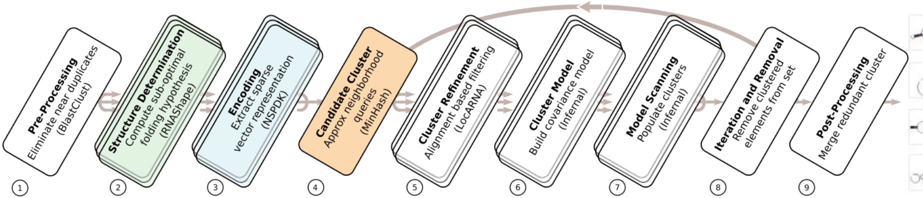
Funding:
Danish Center for Scientific Computing, Innovation Fund Denmark, Danish Cancer Society

Thank you for your attention!

Data set size

	Families with >3 subalignments (Number of alignments)
High APSI	48 (234)
Medium APSI	26 (166)
Low APSI	10 (92)

GraphClust full pipeline

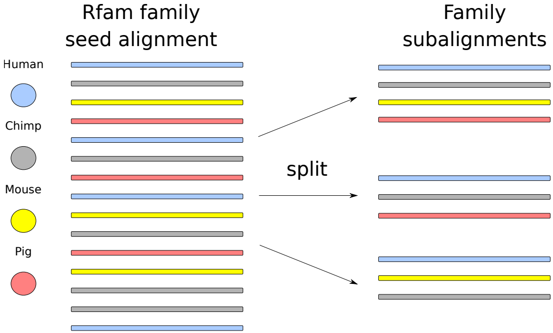


V-measure

- homogeneity: each cluster contains only members of a single class
- completeness: all members of a given class are assigned to the same cluster
- V-measure is harmonic mean of homogeneity and completeness
- 0.0 is as bad as it can be, 1.0 is a perfect score
- not normalized wrt. random labeling

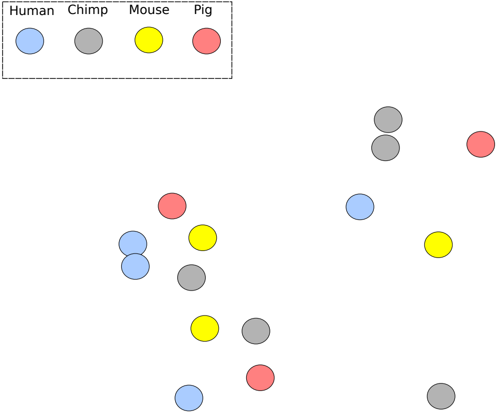
Constructing a benchmark data set

Split each Rfam 12 family seed alignment into subalignments. *Similar* sequences from *different* species form a subalignment.



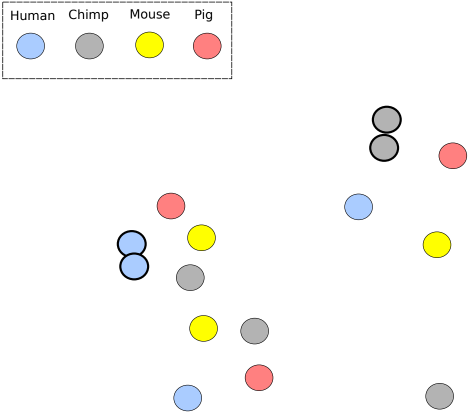
Constructing a benchmark data set

1) Each sequence in the alignment is represented as a node in a graph.



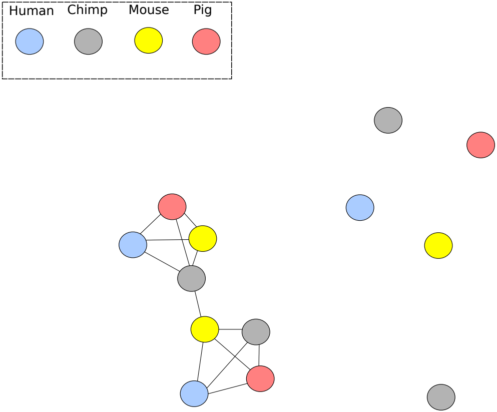
Constructing a benchmark data set

2) Remove sequences with pairwise sequence identify (PSI) $>$ 0.95.



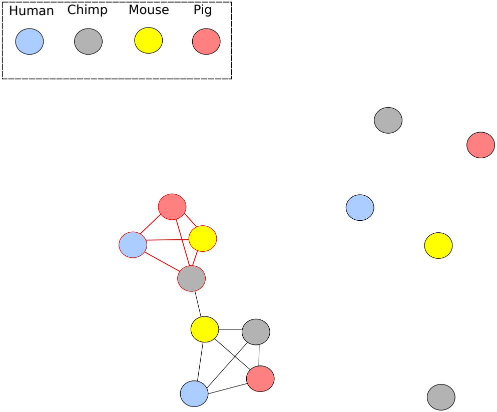
Constructing a benchmark data set

3) Add edge between sequences from diff. species with PSI $\in (0.9, 0.95]$.



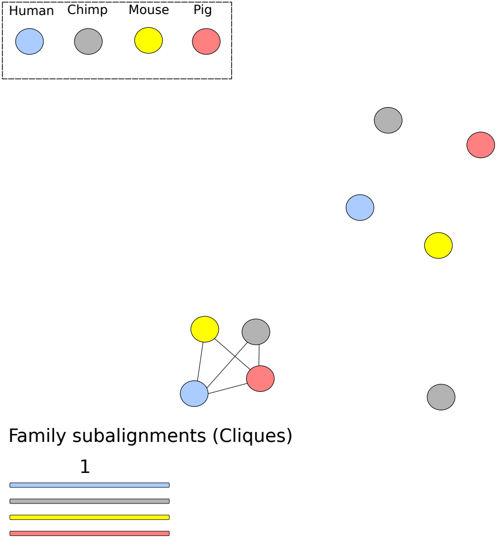
Constructing a benchmark data set

4) Search for cliques in graph.



Constructing a benchmark data set

5) Add clique as subalignment to benchmark data set.

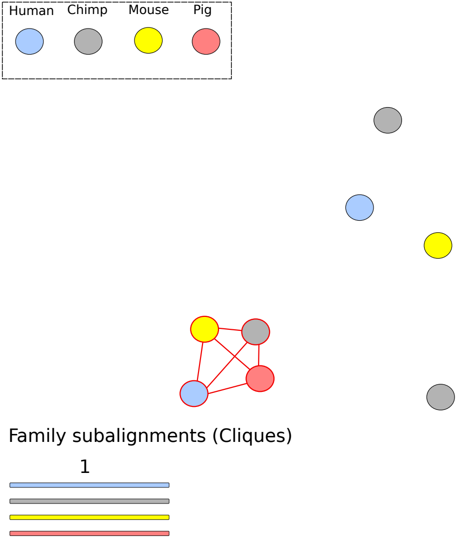


Family subalignments (Cliques)

1

Constructing a benchmark data set

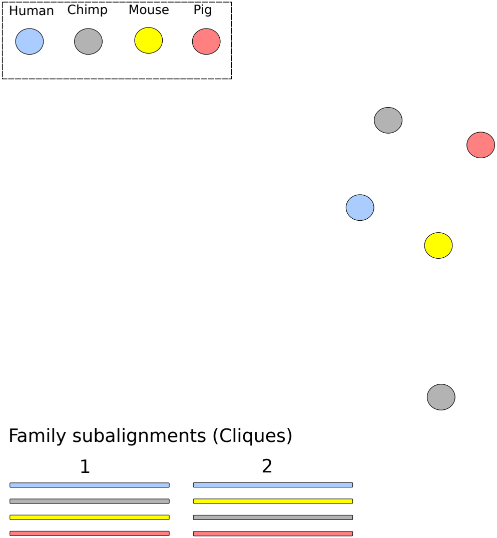
6) Add edge between sequences from diff. species with PSI $\in (0.8, 0.9]$.



Family subalignments (Cliques)

Constructing a benchmark data set

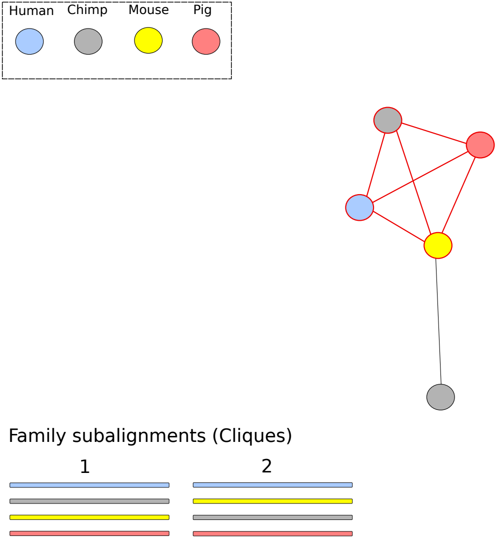
7) Add clique as subalignment to benchmark data set.



Family subalignments (Cliques)

Constructing a benchmark data set

8) Add edge between sequences from diff. species with PSI $\in (0.7, 0.8]$.



Human Chimp Mouse Pig

Family subalignments (Cliques)

1 2