RNAscClust – clustering RNAs using structure conservation and graph-based motifs

Milad Miladi[1]
Alexander Junge[2]

miladim@informatik.uni-freiburg.de

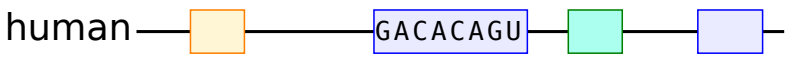
ajunge@rth.dk

[1]Bioinformatics Group, University of Freiburg

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

31st TBI Winterseminar
Bled, Slovenia
February 2016

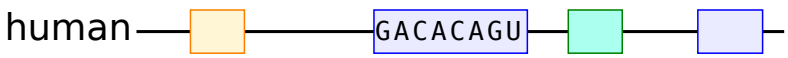
Clustering ncRNA sequences using structure conservation

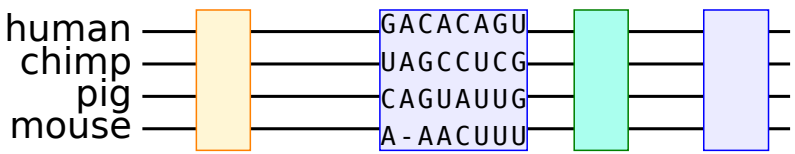


`GraphClust` [Heyne et al., Bioinformatics, 2012]:

- Clusters ncRNA sequences
- Can find paralogs belonging to same ncRNA class
- Features based on local sequence and structure

Clustering ncRNA sequences using structure conservation





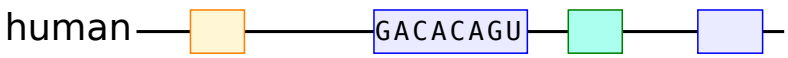
`GraphClust` [Heyne et al., Bioinformatics, 2012]:

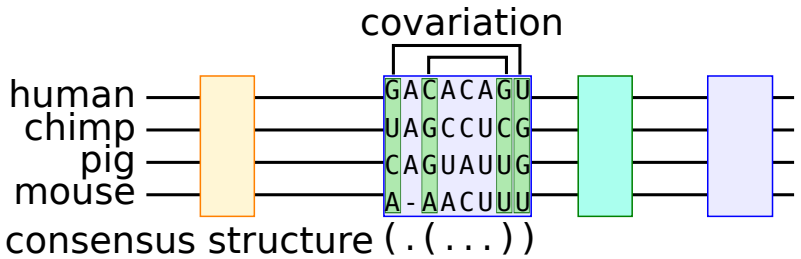
- Clusters ncRNA sequences
- Can find paralogs belonging to same ncRNA class
- Features based on local sequence and structure

`RNAscClust`:

- Clusters paralogous RNA sequences structurally aligned to their orthologs
- Extends `GraphClust` approach:

Clustering ncRNA sequences using structure conservation





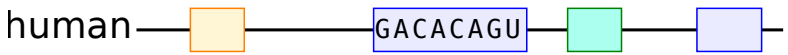
`GraphClust` [Heyne et al., Bioinformatics, 2012]:

- Clusters ncRNA sequences
- Can find paralogs belonging to same ncRNA class
- Features based on local sequence and structure

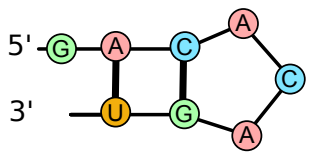
`RNAscClust`:

- Clusters paralogous RNA sequences structurally aligned to their orthologs
- Extends `GraphClust` approach:
  $\Rightarrow$ Derives evolutionary conserved sequence and secondary structure

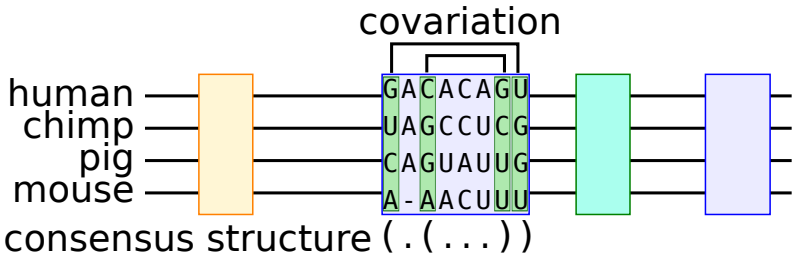
Single sequence vs alignment clustering



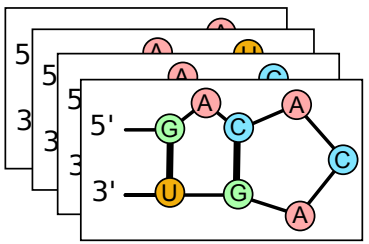
structure prediction
from single sequence



wrong structure

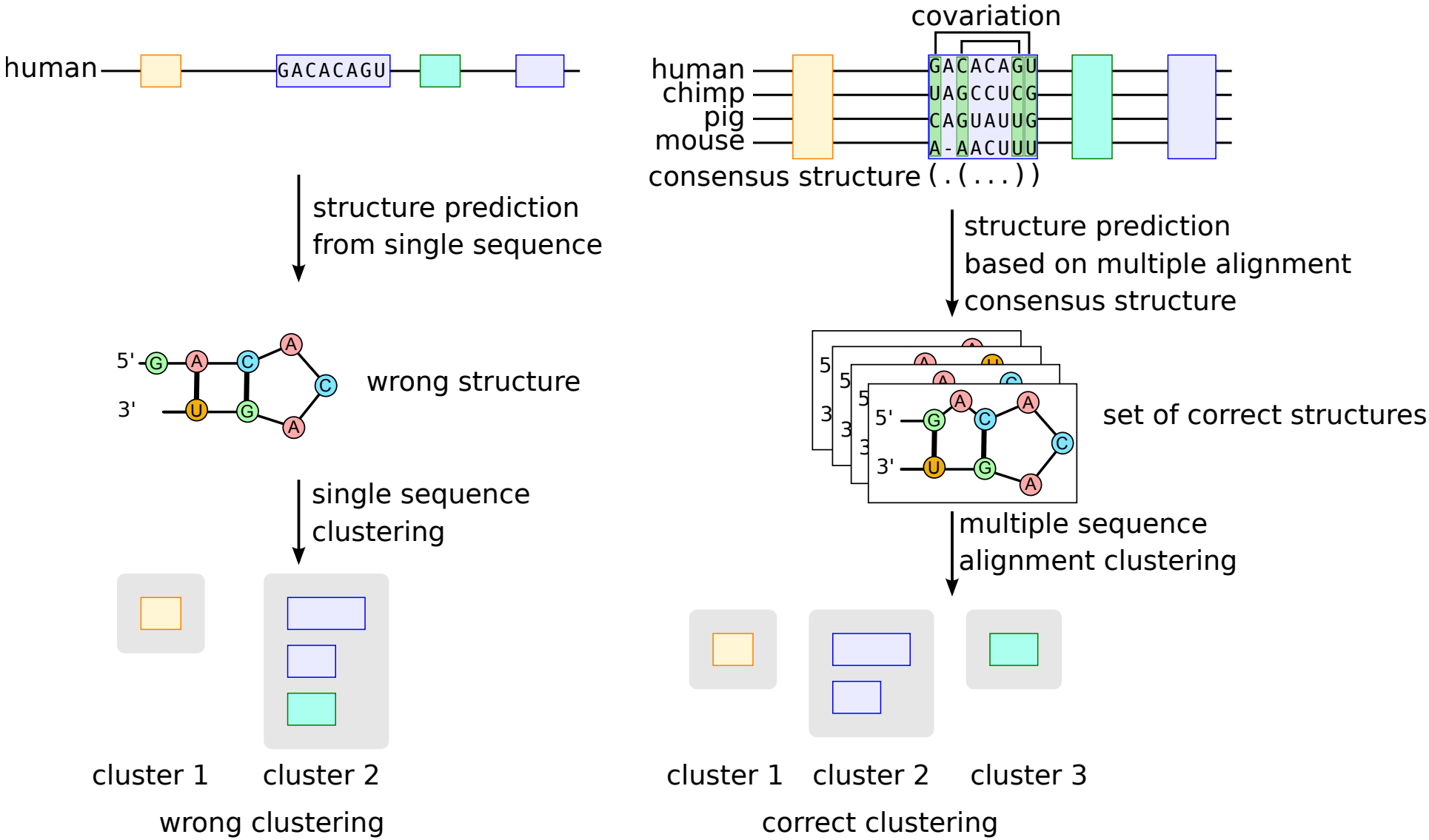


structure prediction
based on multiple alignment
consensus structure



set of correct structures

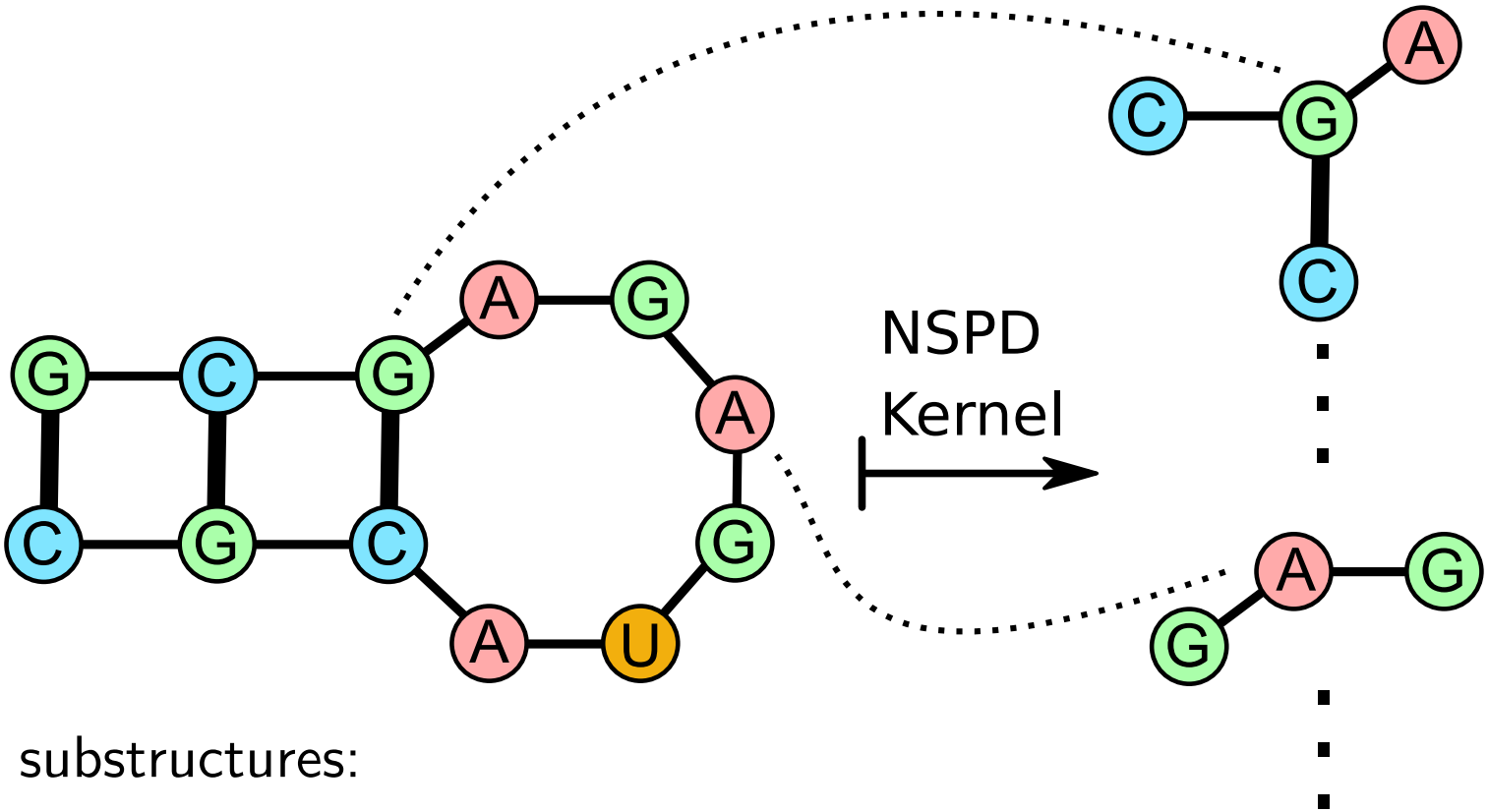
Single sequence vs alignment clustering



Identifying similarities of secondary structures

Neighborhood Subgraph Pairwise Distance (NSPD) Graph Kernel

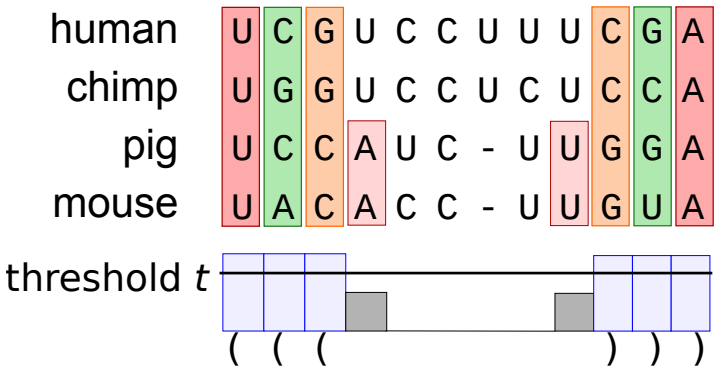
[Costa and De Grave, Proceedings ICML 10, 2010]



Extract substructures:

- intuitively: structure k-mers
- ncRNAs highly similar if many shared substructures

Measuring structure similarity of multiple alignments



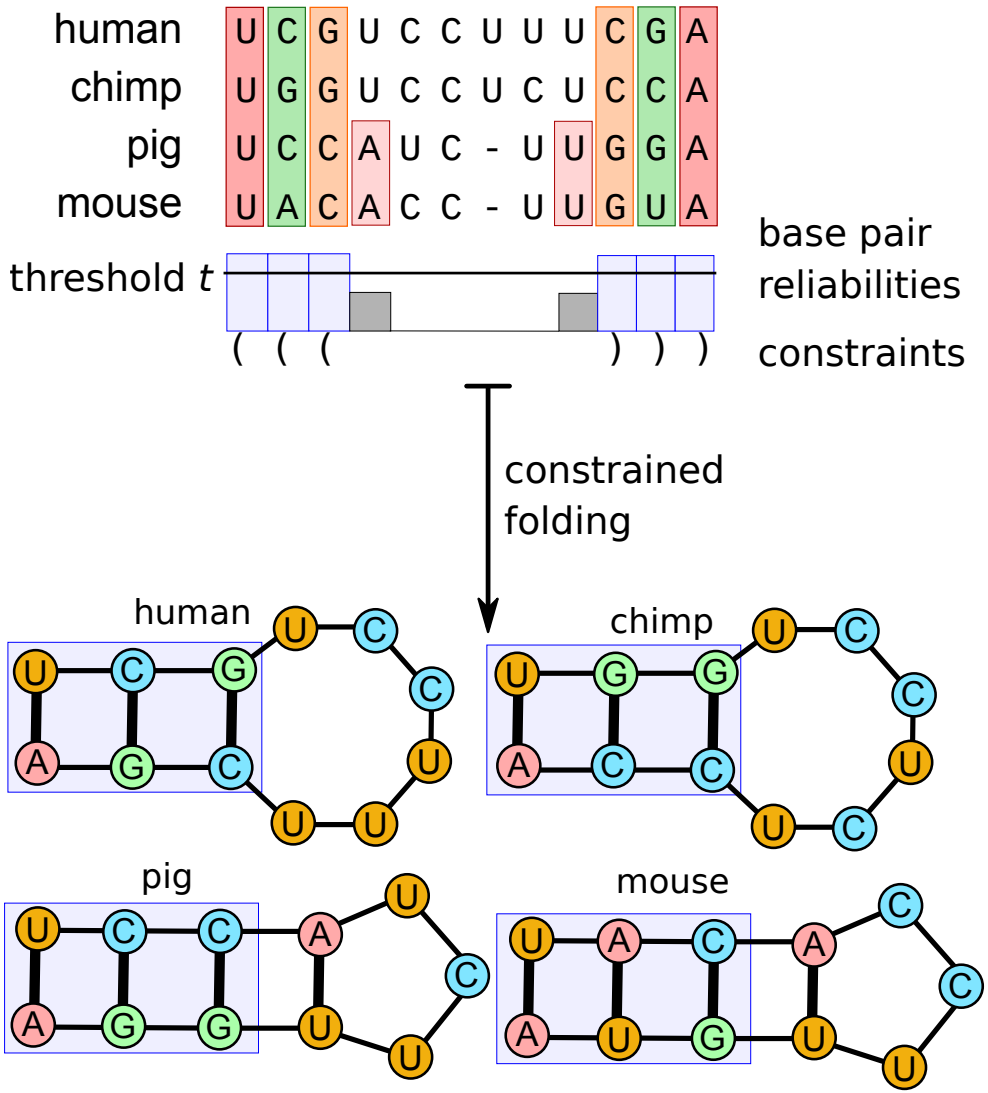
base pair reliabilities

constraints

`PETfold`

[Seemann et al., NAR, 2008]

Measuring structure similarity of multiple alignments

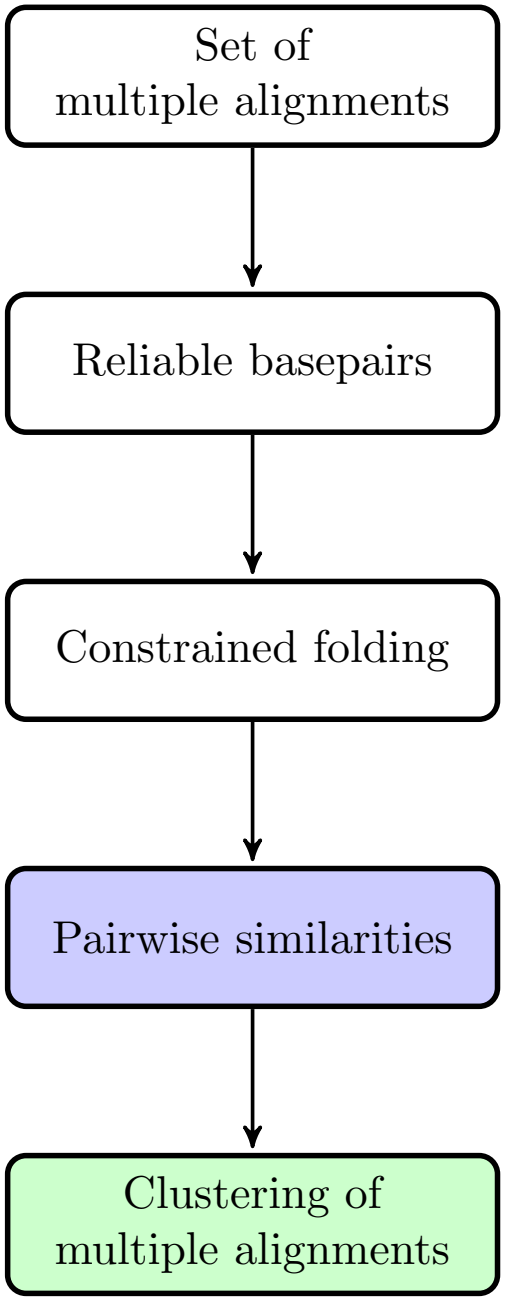


`PETfold`
[Seemann et al., NAR, 2008]

`RNAfold` for each sequence
[Lorenz et al., Alg for Mol Biol, 2011]

$\rightarrow$ use NSPD Graph Kernel to compare alignments

RNAscClust pipeline: From input alignments to clustering

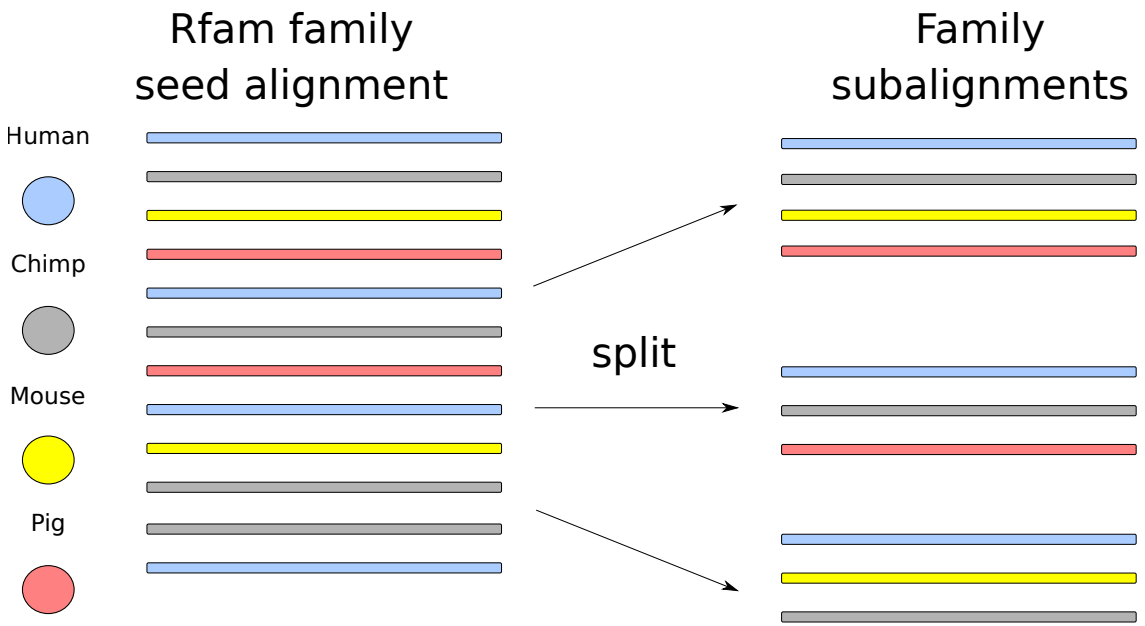


NSPD Graph Kernel

`GraphClust` postprocessing steps

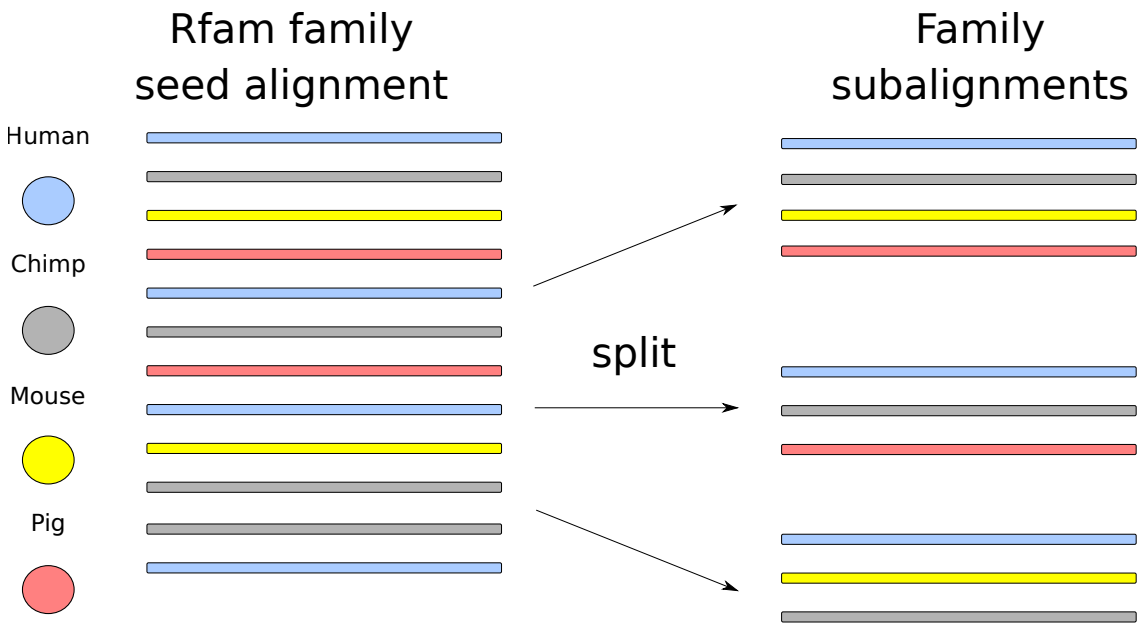
Constructing a benchmark data set

Split Rfam 12 family seed alignments into subalignments:



Constructing a benchmark data set

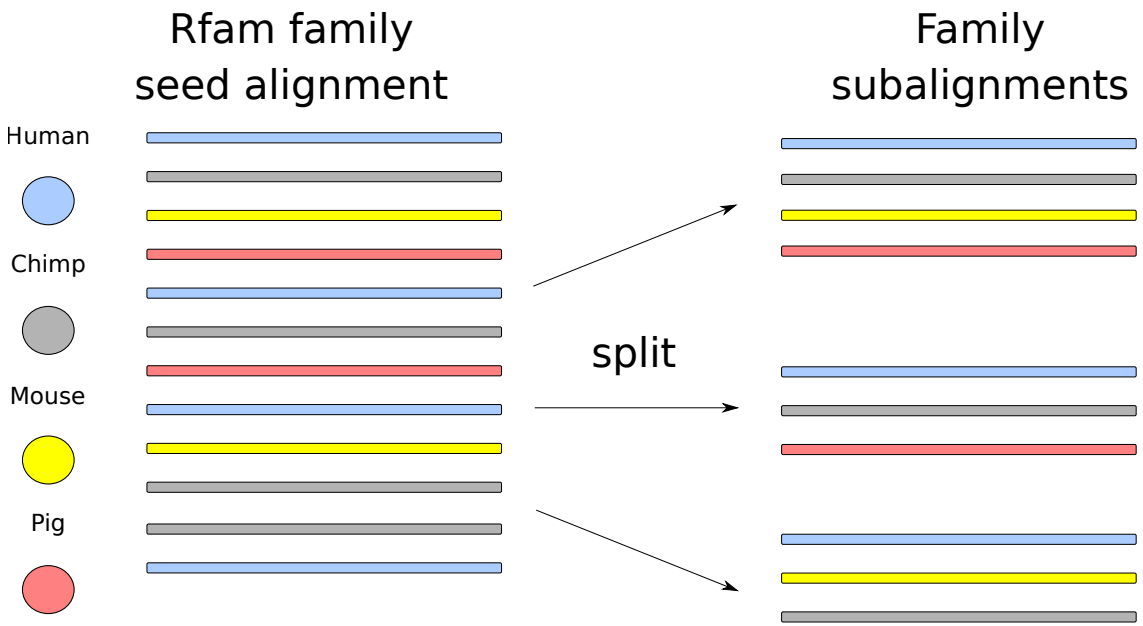
Split Rfam 12 family seed alignments into subalignments:



1. Each subalignment contains one human sequence
2. Similar sequences from different species form a subalignment
   $\Rightarrow$ subalignments have max. sequence identity

Constructing a benchmark data set

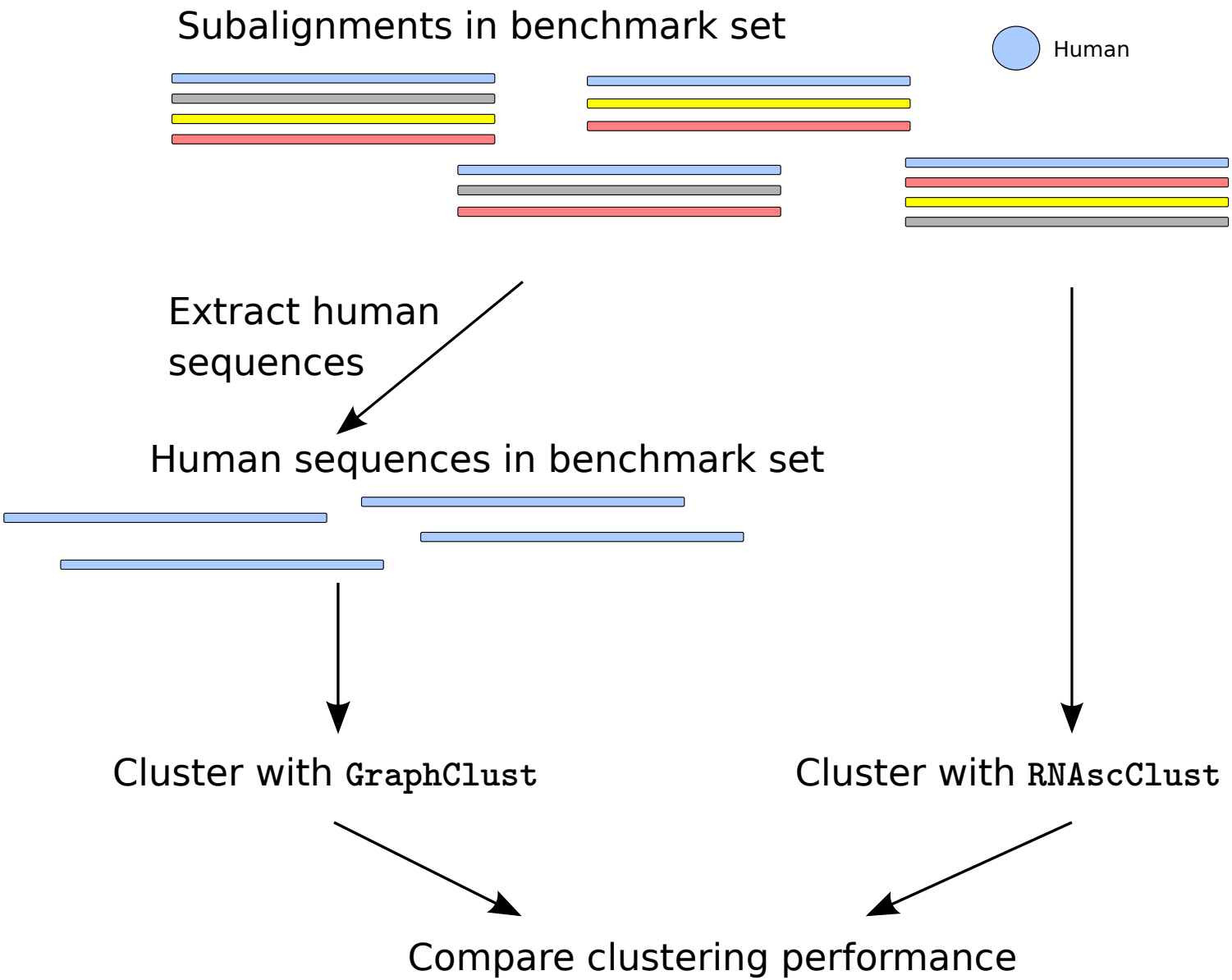
Split Rfam 12 family seed alignments into subalignments:



1. Each subalignment contains one human sequence
2. Similar sequences from different species form a subalignment
   $\Rightarrow$ subalignments have max. sequence identity

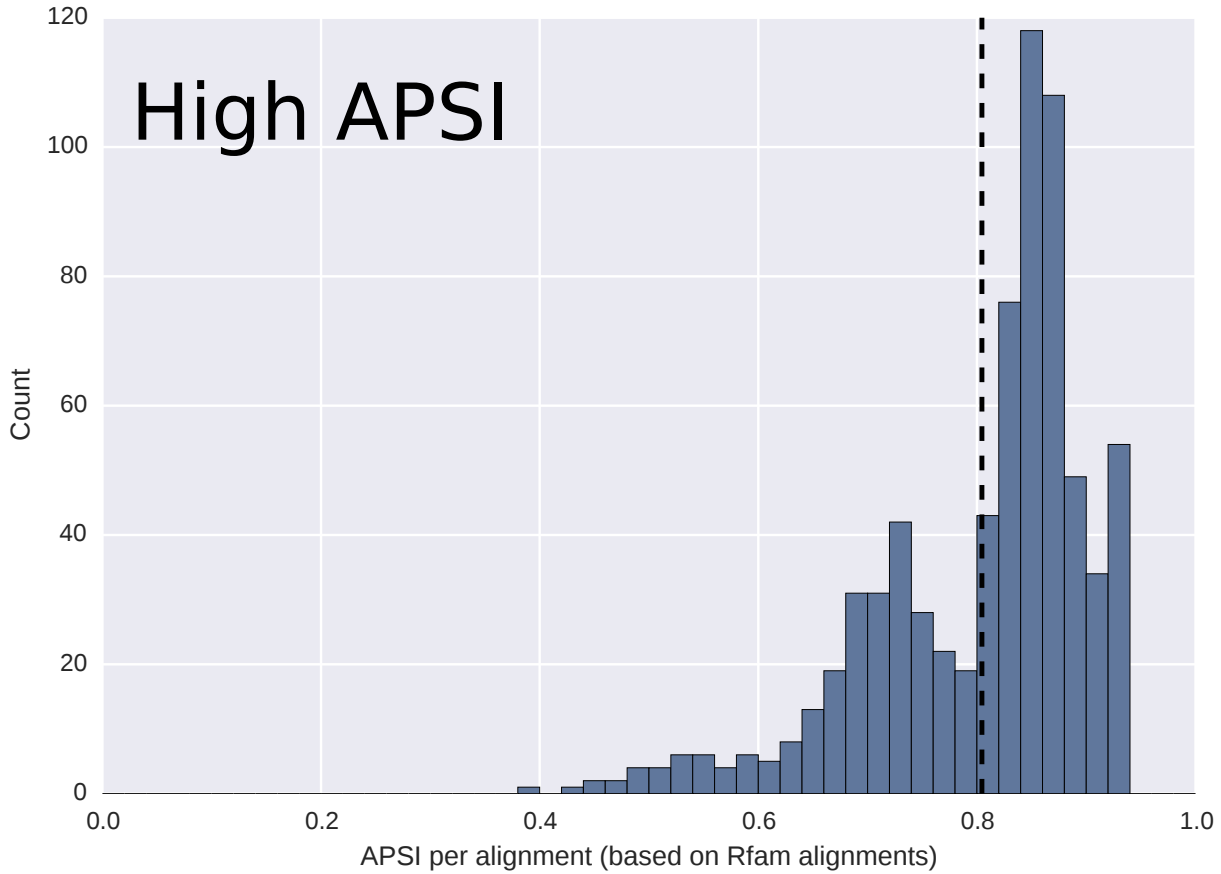
Ideal clustering groups only subalignments from same Rfam family

Comparing sequence to alignment clustering



Low covariation in the benchmark data set

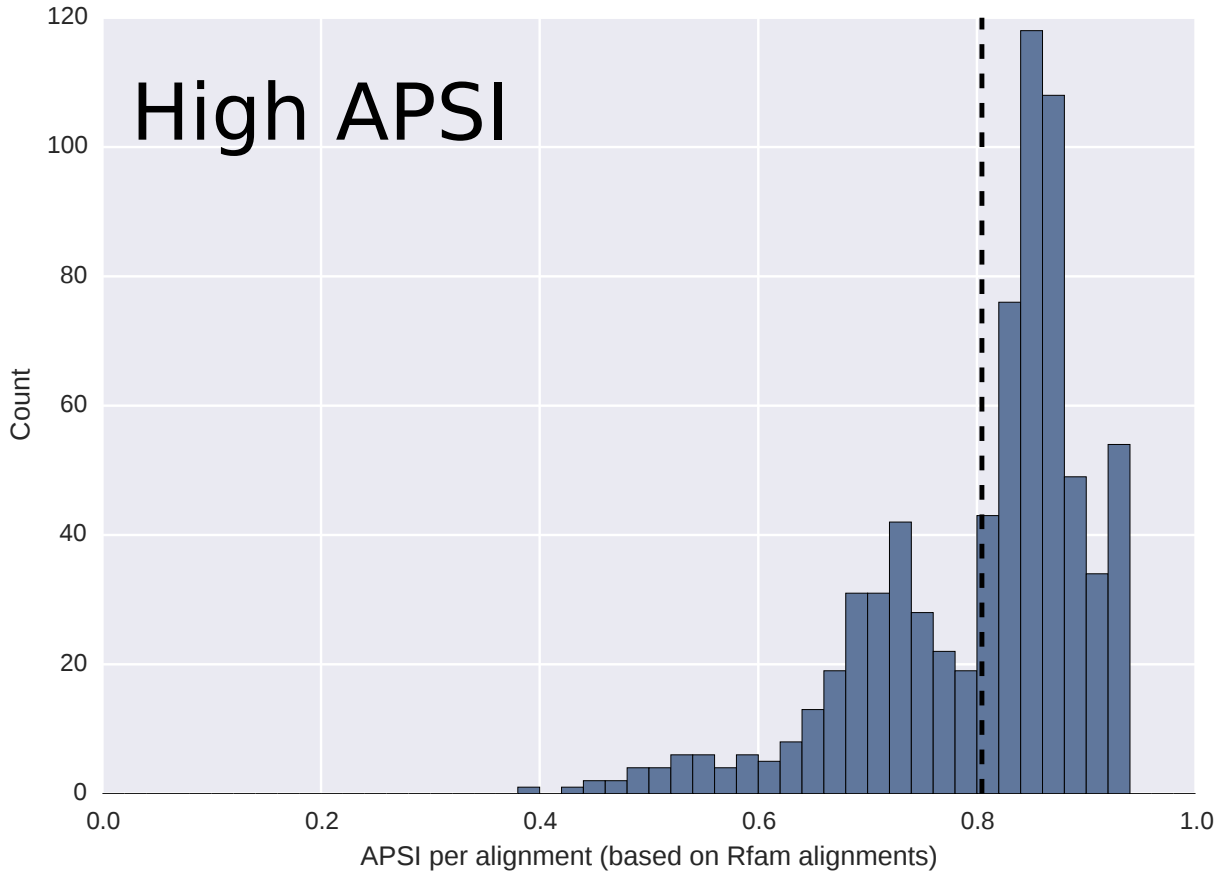
Benchmark set has high **A**verage **P**airwise **S**equence **I**dentity (APSI) in alignments



0.81 mean APSI
48 families
234 alignments

Low covariation in the benchmark data set

Benchmark set has high **A**verage **P**airwise **S**equence **I**dentity (APSI) in alignments

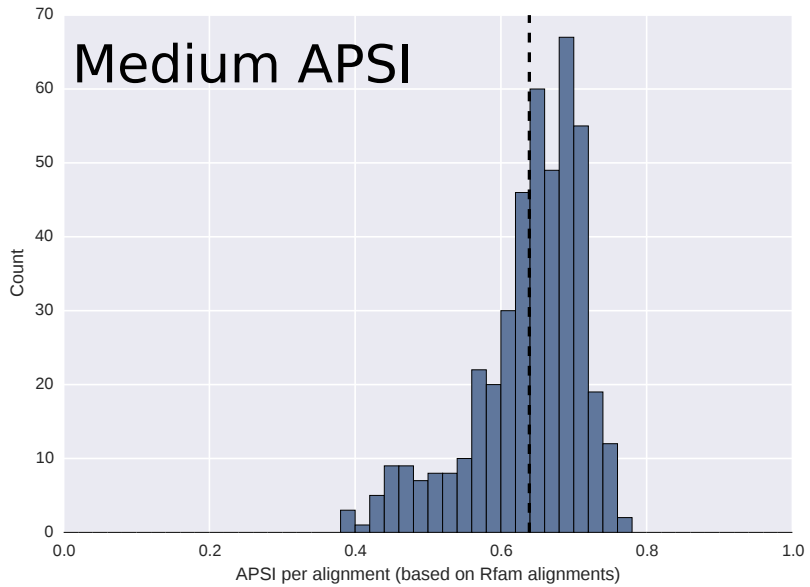


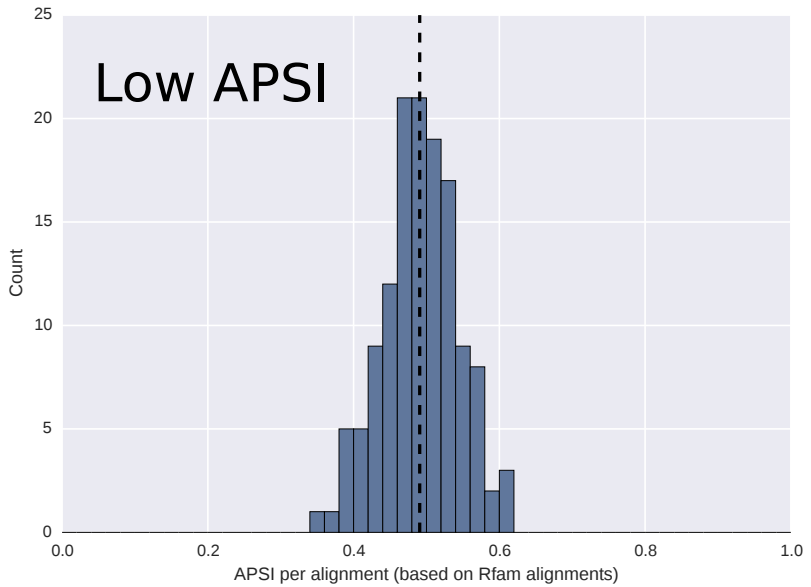
0.81 mean APSI
48 families
234 alignments

$\rightarrow$ limit APSI to study effect of covariation on clustering performance

Benchmark sets with different degrees of covariation

Create 2 additional benchmark data set with controlled APSI in alignments





0.64 mean APSI
26 families
166 alignments

0.49 mean APSI
10 families
92 alignments

Clustering performance metrics - V-measure

- Homogeneity H: each cluster contains only members of a single family
- Completeness C: all members of a given family are in same cluster
- V-measure [Rosenberg and Hirschberg, 2007] is harmonic mean of H and C:

$$V = \frac{2 \cdot H \cdot C}{H + C}$$

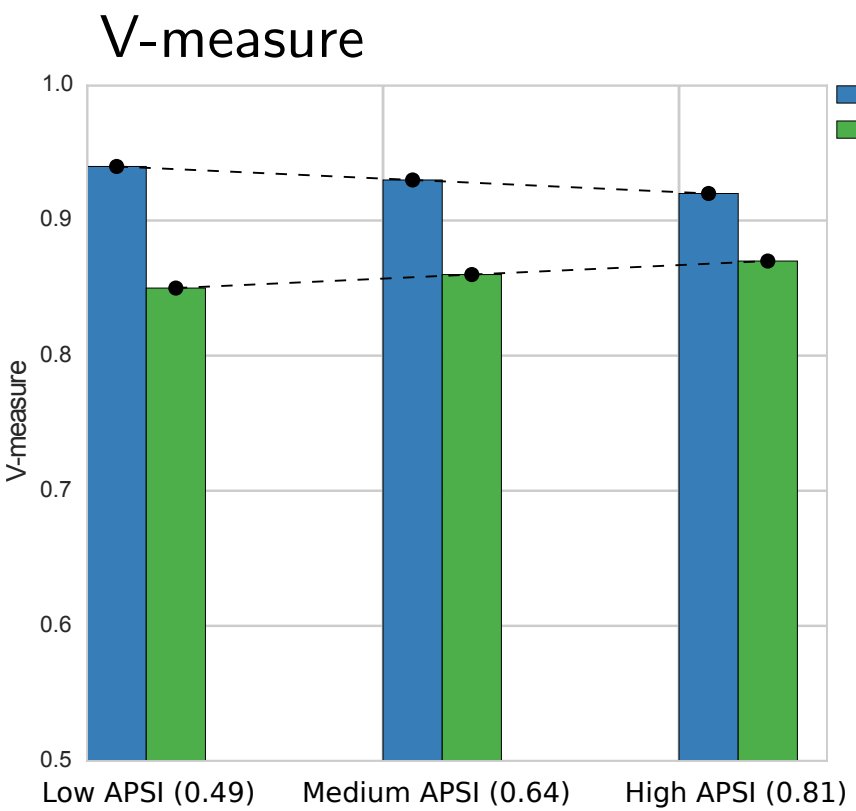
Clustering performance metrics - Adjusted Rand Index

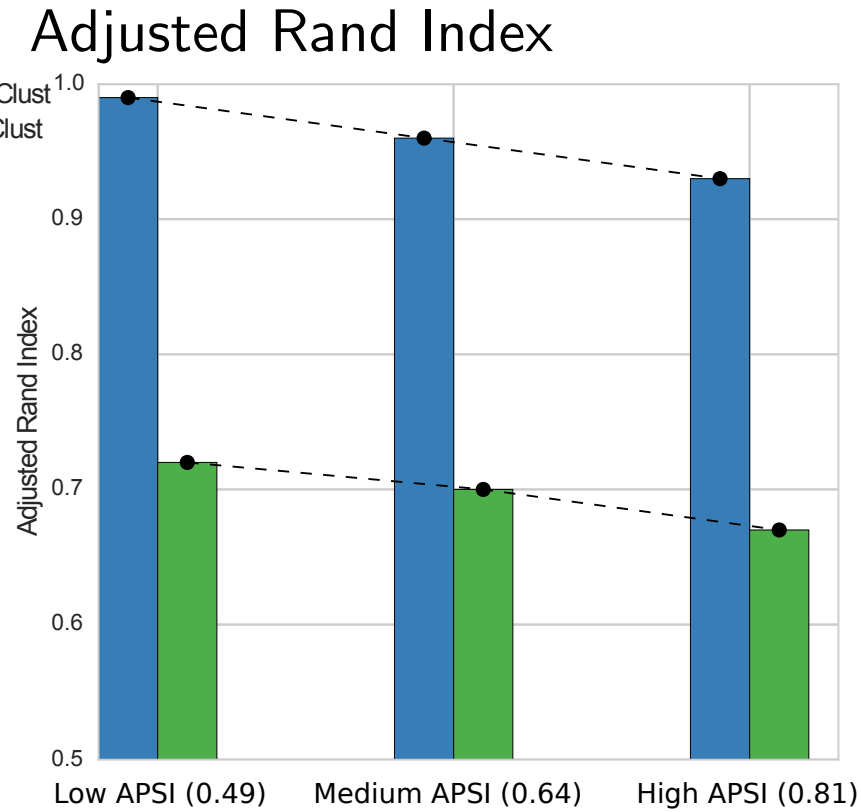
- a = #object pairs from same family assigned to same cluster
- b = #object pairs from different families assigned to different clusters
- n = number of alignments

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

- Adjusted Rand Index [Hubert and Arabie, 1985] is Rand Index [Rand, 1971] adjusted for chance

More covariation improves RNAscClust performance





RNAscClust – Conclusion

- Leverage conserved sec. structure derived from multiple alignments
- NSPD Graph Kernel as similarity measure
- Improved clustering compared to GraphClust

RNAscClust – Conclusion

- Leverage conserved sec. structure derived from multiple alignments
- NSPD Graph Kernel as similarity measure
- Improved clustering compared to GraphClust

Next step:

- Genome-scale clustering of potential ncRNAs

Acknowledgements

Bioinformatics Group,
University of Freiburg:

- Milad Miladi
- Fabrizio Costa
- Rolf Backofen

RTH, University of Copenhagen:

- Stefan Seemann
- Jakob Hull Havgaard
- Jan Gorodkin

Funding:

DFG, Danish Center for Scientific Computing, Innovation Fund Denmark,
Danish Cancer Society

Acknowledgements

Bioinformatics Group,
University of Freiburg:

- Milad Miladi
- Fabrizio Costa
- Rolf Backofen

RTH, University of Copenhagen:

- Stefan Seemann
- Jakob Hull Havgaard
- Jan Gorodkin

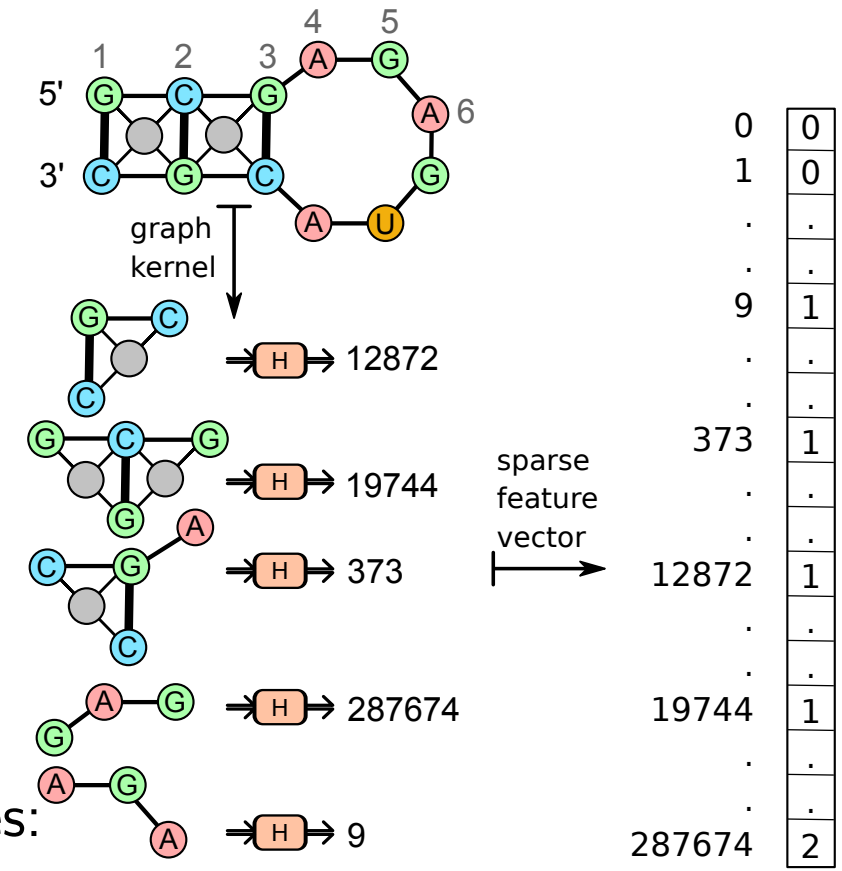
Funding:

DFG, Danish Center for Scientific Computing, Innovation Fund Denmark,
Danish Cancer Society

Thank you for your attention!

Identifying similarities of secondary structures

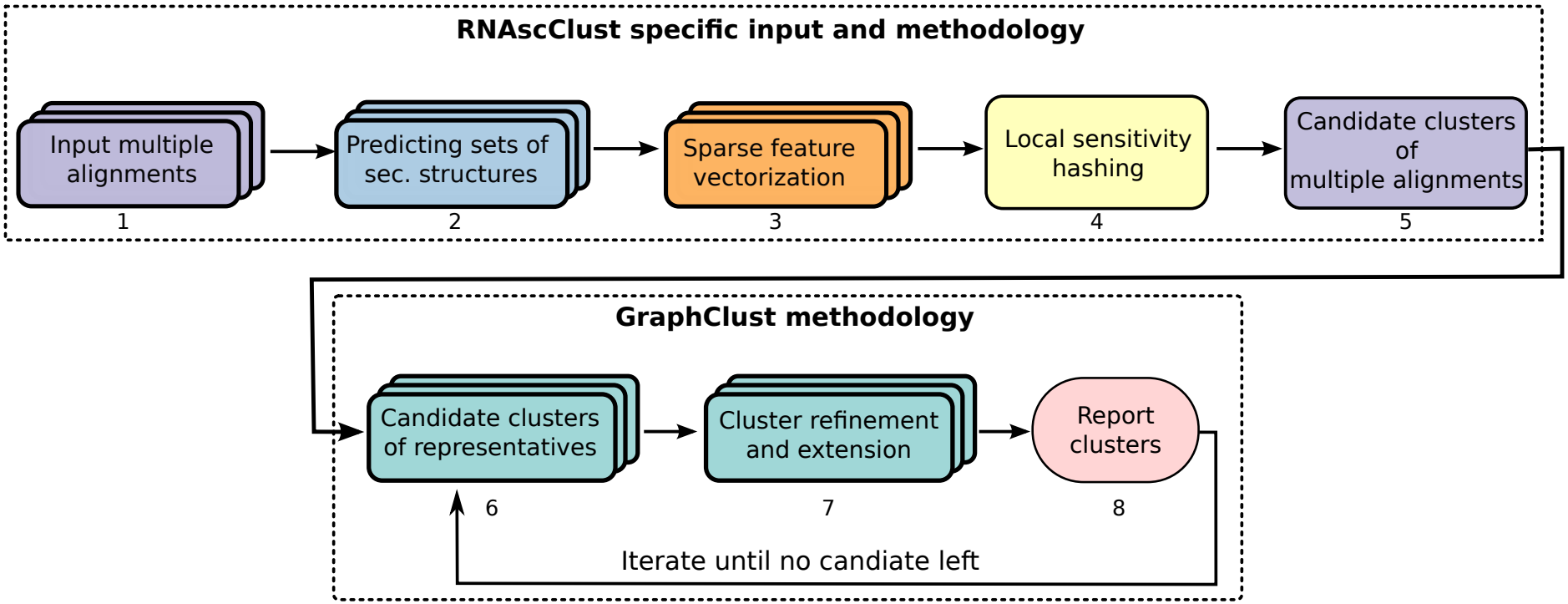
Neighborhood Subgraph Pairwise Distance (NSPD) Kernel used in `GraphClust` [Heyne et al., Bioinformatics, 2012]



Extract substructures:

- $\approx$ structure k-mers
- ncRNAs highly similar if many shared substructures

RNAscClust full pipeline



Steps executed in parallel are shown as stacks

V-measure

Clusters $K = \{K_1, \ldots, K_m\}$; true classes $C = \{C_1, \ldots, C_n\}$. Homogeneity h is defined as:

$$h = \begin{cases} 1 & \text{if } H(C|K) = 0 \\ 1 - \frac{H(C|K)}{H(C)} & \text{else} \end{cases}$$

$H(C|K)$ is the conditional entropy of the classes given the clustering and $H(C)$ is the entropy of the classes, i.e.,

$$H(C|K) = -\sum_{k=1}^{|K|}\sum_{c=1}^{|C|} \frac{a_{ck}}{N} \log \frac{a_{ck}}{\sum_{c=1}^{|C|} a_{ck}}$$

$$H(C) = -\sum_{c=1}^{|C|} \frac{\sum_{k=1}^{|K|} a_{ck}}{n} \log \frac{\sum_{k=1}^{|K|} a_{ck}}{n}$$

V-measure

Clusters $K = \{K_1, \ldots, K_m\}$; true classes $C = \{C_1, \ldots, C_n\}$.
On the other hand, completeness c is defined as:

$$c = \begin{cases} 1 & \text{if } H(K|C) = 0 \\ 1 - \frac{H(K|C)}{H(K)} & \text{else} \end{cases}$$

where $H(K|C)$ is the conditional entropy of the clustering given the classes and $H(K)$ is the entropy of the clustering, i.e.,

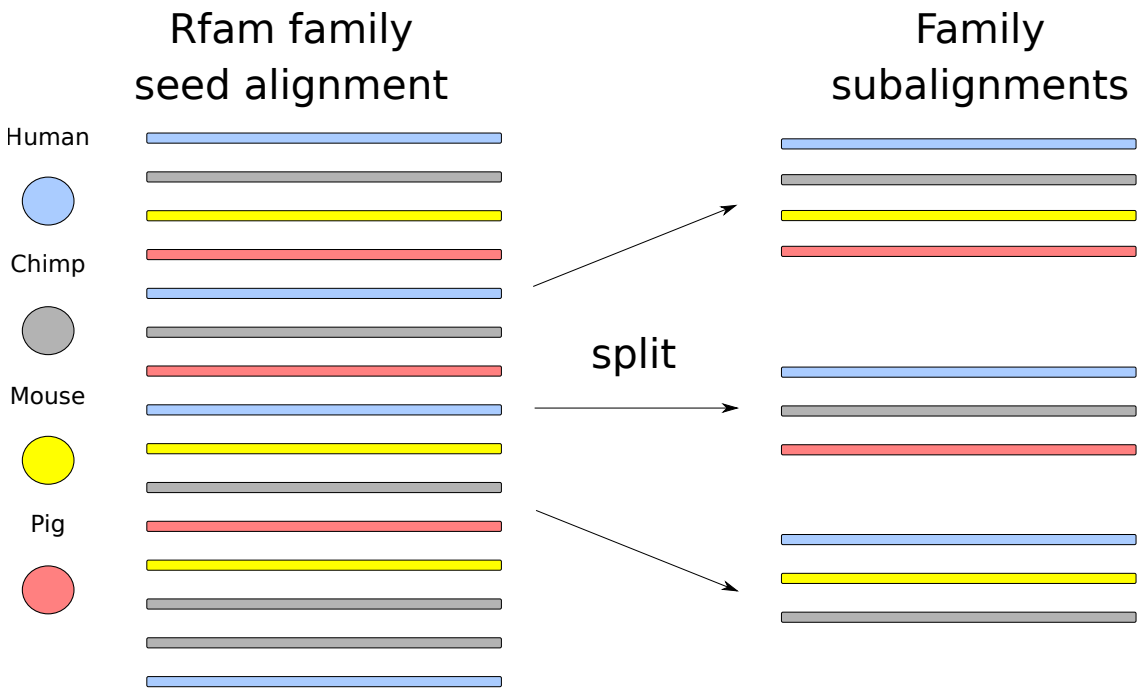
$$H(K|C) = -\sum_{c=1}^{|C|} \sum_{k=1}^{|K|} \frac{a_{ck}}{N} \log \frac{a_{ck}}{\sum_{k=1}^{|K|} a_{ck}}$$

$$H(K) = -\sum_{k=1}^{|K|} \frac{\sum_{c=1}^{|C|} a_{ck}}{n} \log \frac{\sum_{c=1}^{|C|} a_{ck}}{n}$$

V-measure is harmonic mean of homogeneity and completeness and is not normalized wrt. random labeling. 0.0 is as bad as it can be, 1.0 is perfect.

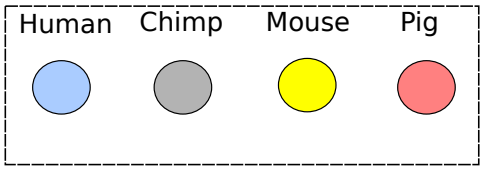
Constructing a benchmark data set

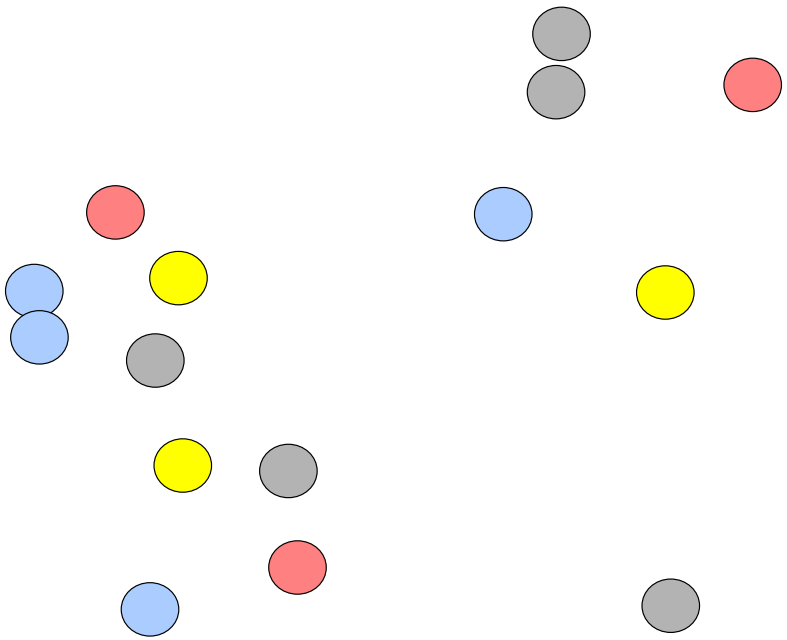
Split each Rfam 12 family seed alignment into subalignments. *Similar* sequences from *different* species form a subalignment.



Constructing a benchmark data set

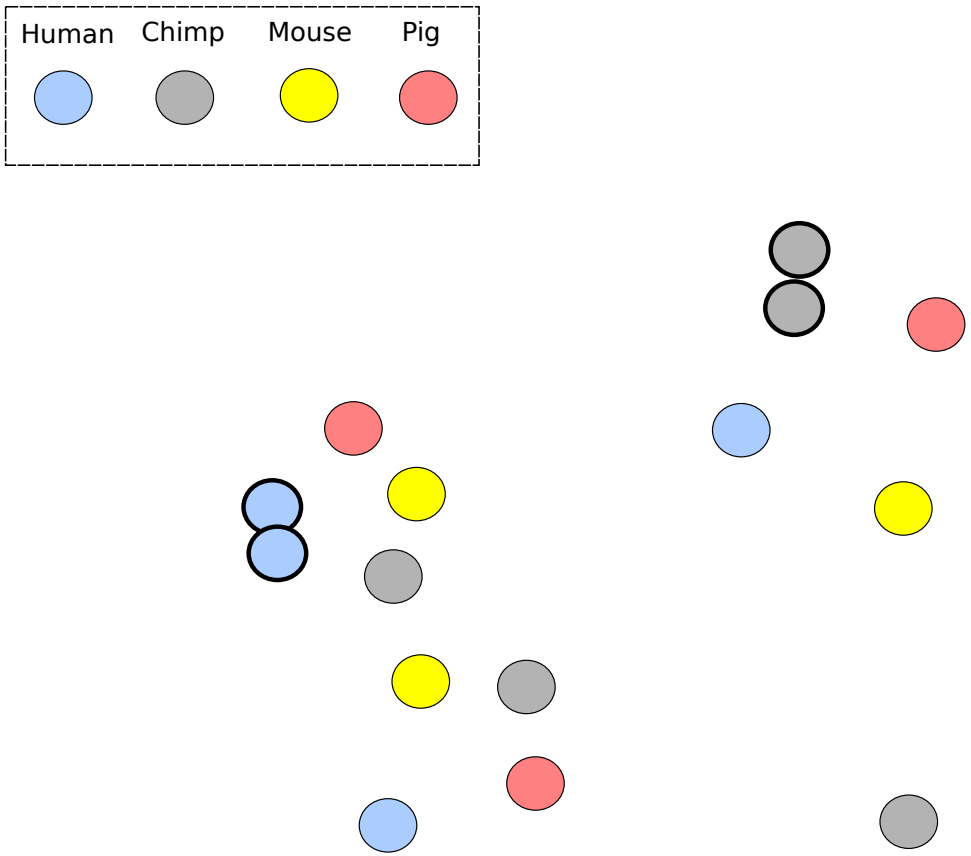
1) Each sequence in the alignment is represented as a node in a graph.





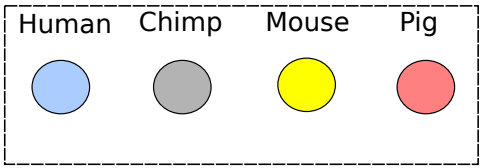
Constructing a benchmark data set

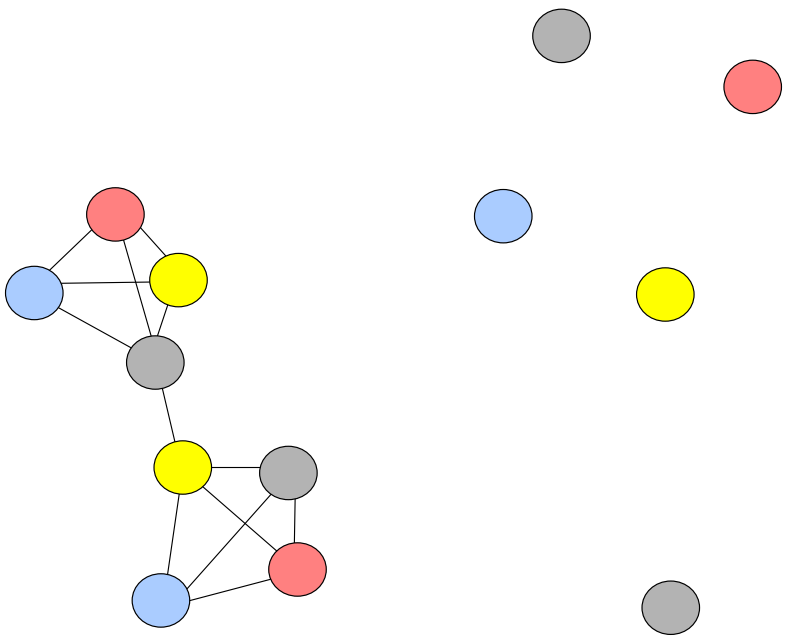
2) Remove sequences with pairwise sequence identify (PSI) > 0.95.



Constructing a benchmark data set

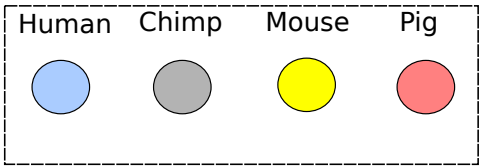
3) Add edge between sequences from diff. species with PSI $\in (0.9, 0.95]$.

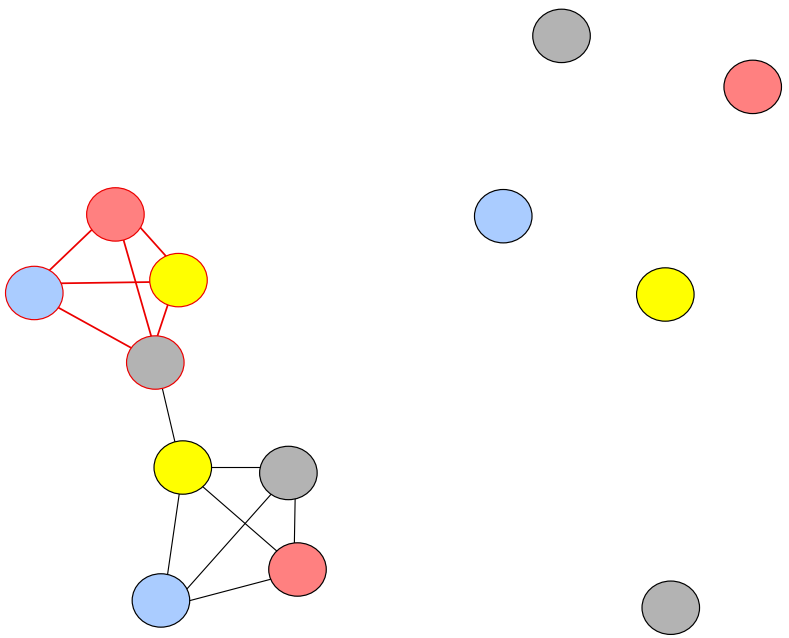




Constructing a benchmark data set

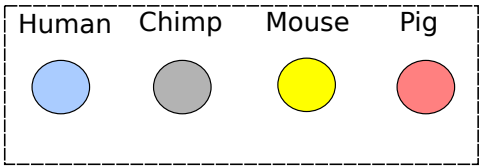
4) Search for cliques in graph.

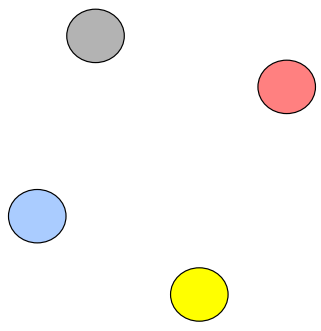


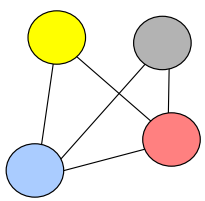


Constructing a benchmark data set

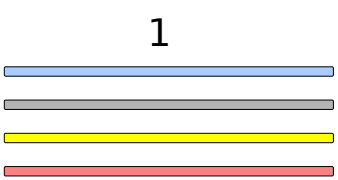
5) Add clique with max. APSI as subalignment to benchmark data set.





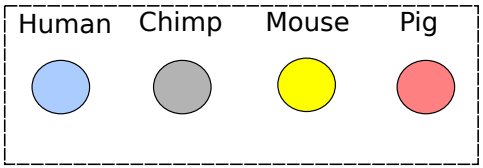


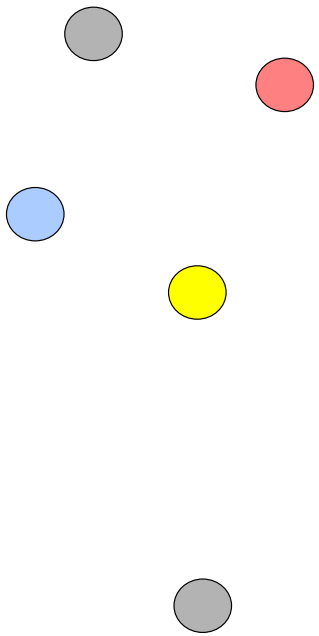
Family subalignments (Cliques)

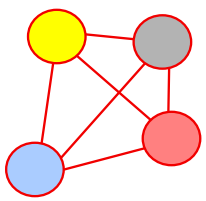


Constructing a benchmark data set

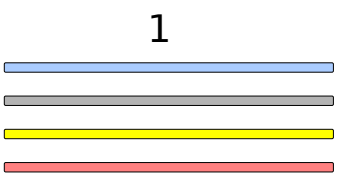
6) Add edge between sequences from diff. species with PSI $\in (0.8, 0.9]$.





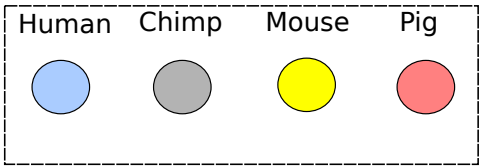


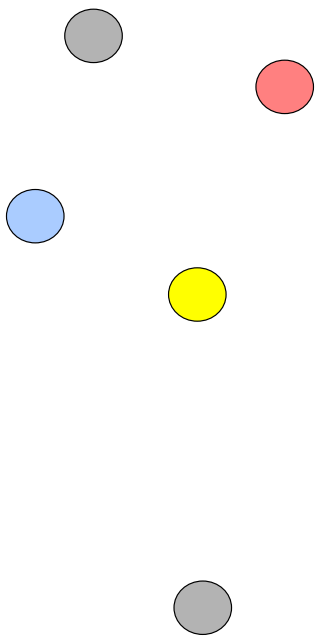
Family subalignments (Cliques)



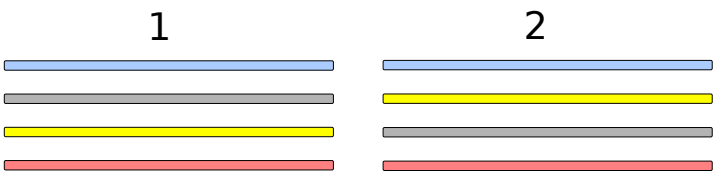
Constructing a benchmark data set

7) Add clique as subalignment to benchmark data set.



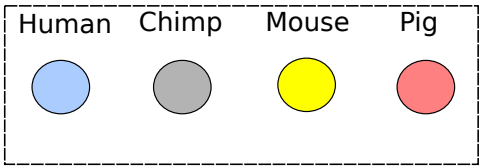


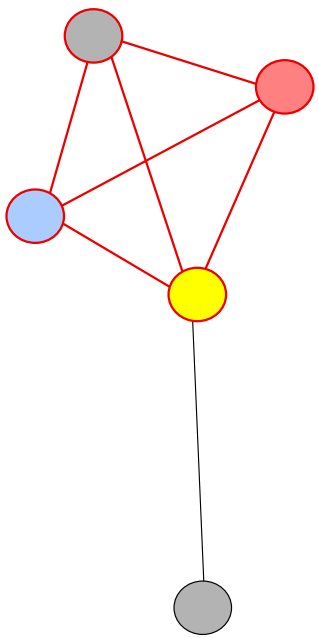
Family subalignments (Cliques)



Constructing a benchmark data set

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





Family subalignments (Cliques)