maxAlike: Sequence Reconstruction by Maximum Likelihood

Peter Menzel

Bioinformatics Group, Department of Basic Animal and Veterinary Sciences,
Faculty of Life Sciences, University of Copenhagen
Professur für Bioinformatik, Institut für Informatik, Universität Leipzig

February 15, 2010

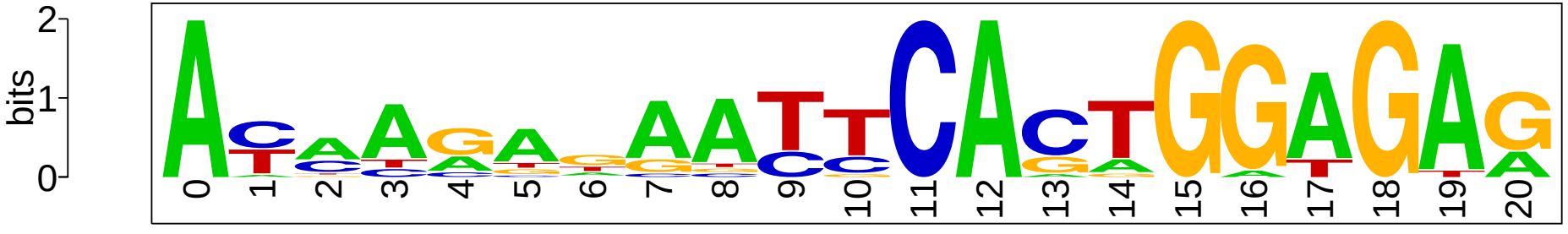
Overview

- Reconstruct / predict sequences in a target species/genome ...
- using an MSA from known homologs to the target sequence and ...
- a phylogenetic tree containing the position of the target species ...
- by applying the maximum likelihood principle ...
- Use predicted sequences for homology search (PSSMs), Primer design for yet un-sequenced genes (gaps in genomes)

PSSMs in homology search

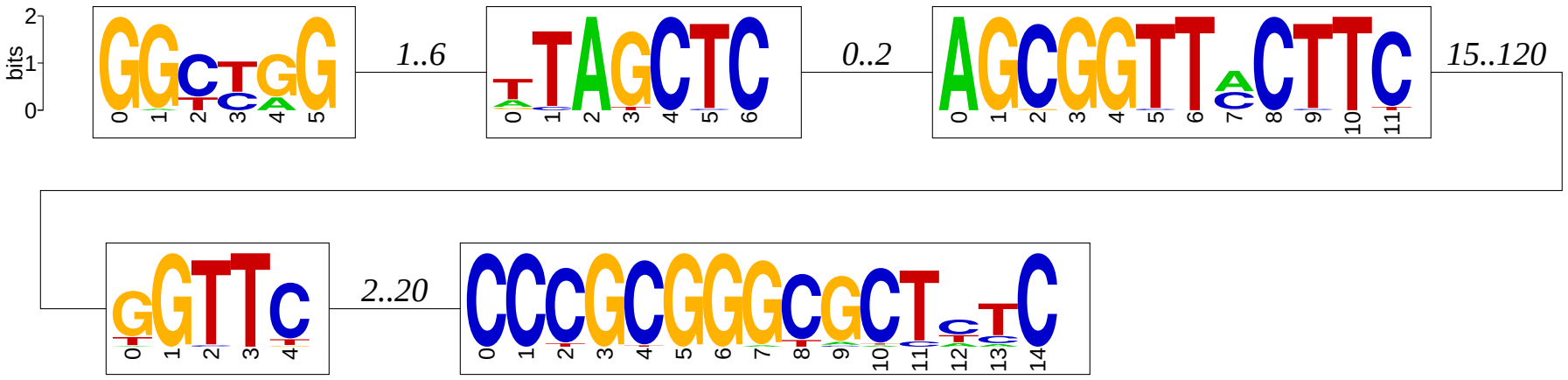
```
          *           **   *   *
   hg18   ATAAGATAACTCAGTGGAGAG
panTro1   ATTAGGTAACTCAGTGGAGAG
rheMac2   ACCAGAGGTTGCACGGGAGAG
   rn4    ACAAGAAAATTCACTGGAGAA
oryCun1   ATCAGCGAATTCACAGGAGAG
   mm8    ACAAAGTAGTCCACTGGAGAG
bosTau2   ATGAAAGAACTCACTGGAGAG
canFam2   ATCACAGAATTCACTGAAGAG
dasNov1   AAAAGTTGCTTCAGTGGAGAA
echTel1   ATATAAGAACTCAGTGGAGAG
monDom4   ATCAGAGAACTCACTGGAGAA
galGal2   ACCAGTGCATCCAAAGGAGAG
xenTro1   ACAAAAAAATTCACTGGAGAG
tetNig1   ACACCACAATTCACAGGTGAA
   fr1    ACACAAGAATCCACTGGTGAG
danRer3   ACATGAAGATCCACTGGTGTA
```

A	16	1	9	12	5	11	3	12	13	0	0	0	16	1	3	0	1	13	0	15	5
C	0	8	5	2	2	1	1	1	1	5	4	16	0	11	0	0	0	0	0	0	0
G	0	0	1	0	9	2	8	3	1	0	1	0	0	4	1	16	15	0	16	0	11
T	0	7	1	2	0	2	4	0	1	11	11	0	0	0	12	0	0	3	0	1	0



`fragrep2`: Search patterns based on PSSMs

- Find conserved blocks
- Use distance constraints between these blocks depending on gaps
- Build search pattern



Problems in PSSM construction

- Bias in phylogenetic distribution of the sequences
- Set of closely related model organisms
- Clustering / weighting of sequences, lower weight to overrepresented species
- Phylogenetic relationships between the sequences are usually not taken into account

maxAlike

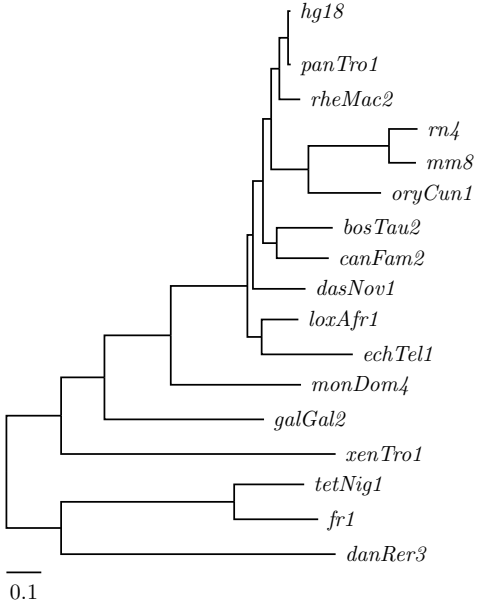
- Overcome overrepresentation bias
- Consider all phylogenetic relationships between species
- Take phylogenetic position of target species into account
- borrow ideas from ancestral reconstruction methods

maxAlike approach

- Instead of reconstructing inner nodes of the tree, we estimate the sequence on a leaf node
- Assume different evolutionary rates for each site
- Given a phylogenetic tree and a multiple alignment, estimate a mutation rate for each site so that the likelihood of the tree is maximized
- Calculate nucleotide probabilities by rooting the tree to the target species

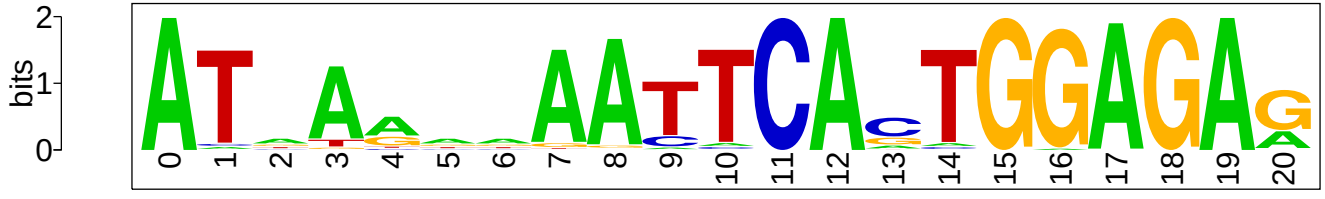
Input and Output

Input:




```
          *              **   *   *
hg18      ATAAGATAACTCAGTGGAGAG
panTro1   ATTAGGTAACTCAGTGGAGAG
rheMac2   ACCAGAGGTTGCACGGGAGAG
rn4       ACAAGAAAATTCACTGGAGAA
oryCun1   ATCAGCGAATTCACAGGAGAG
mm8       ACAAAGTAGTCCACTGGAGAG
bosTau2   ATGAAAGAACTCACTGGAGAG
canFam2   ATCACAGAATTCACTGAAGAG
dasNov1   AAAAGTTGCTTCAGTGGAGAA
echTel1   ATATAAGAACTCAGTGGAGAG
monDom4   ATCAGAGAACTCACTGGAGAA
galGal2   ACCAGTGCATCCAAAGGAGAG
xenTro1   ACAAAAAAATTCACTGGAGAG
tetNig1   ACACCACAATTCACAGGTGAA
fr1       ACACAAGAATCCACTGGTGAG
danRer3   ACATGAAGATCCACTGGTGTA
```

Output for target species *loxAfr1*:



Algorithm 1

1. Delete leaf X (target species) from tree T
2. Determine a mutation rate $\hat{\mu}$ for each alignment column by maximizing the likelihood of the tree (root 0)

$$\hat{\mu} = \underset{\mu}{\operatorname{argmax}}\ L(\mu)$$

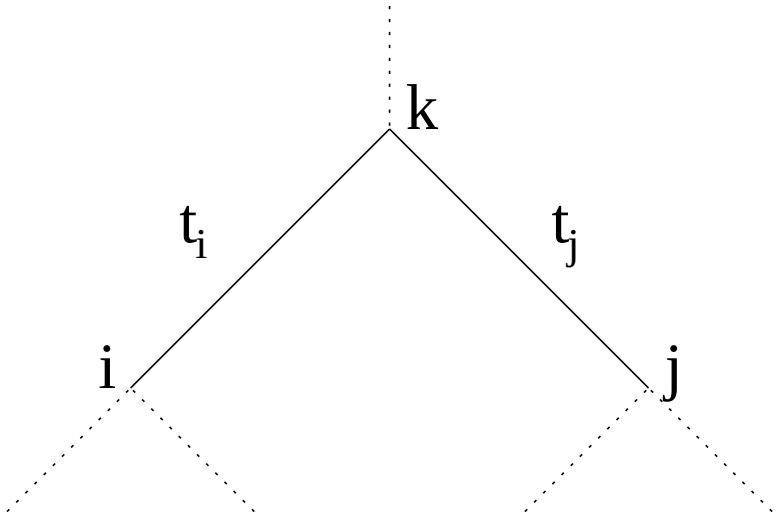
$$L(\mu) = \sum_{s_0} \pi_{s_0} L_{s_0}(\mu)$$

3. Add X to T and root T at X
4. Using the estimated $\hat{\mu}$, recalculate likelihoods for all states s_X of the root node X

Algorithm 2

- Tree likelihood is calculated by post-order traversal of the tree (leafs are evaluated first)
- For interior nodes k

$$L_{s_k}(\mu) = \left(\sum_{s_i} P_{s_k s_i}(t_i, \mu) L_{s_i}(\mu)\right) \cdot \left(\sum_{s_j} P_{s_k s_j}(t_j, \mu) L_{s_j}(\mu)\right)$$

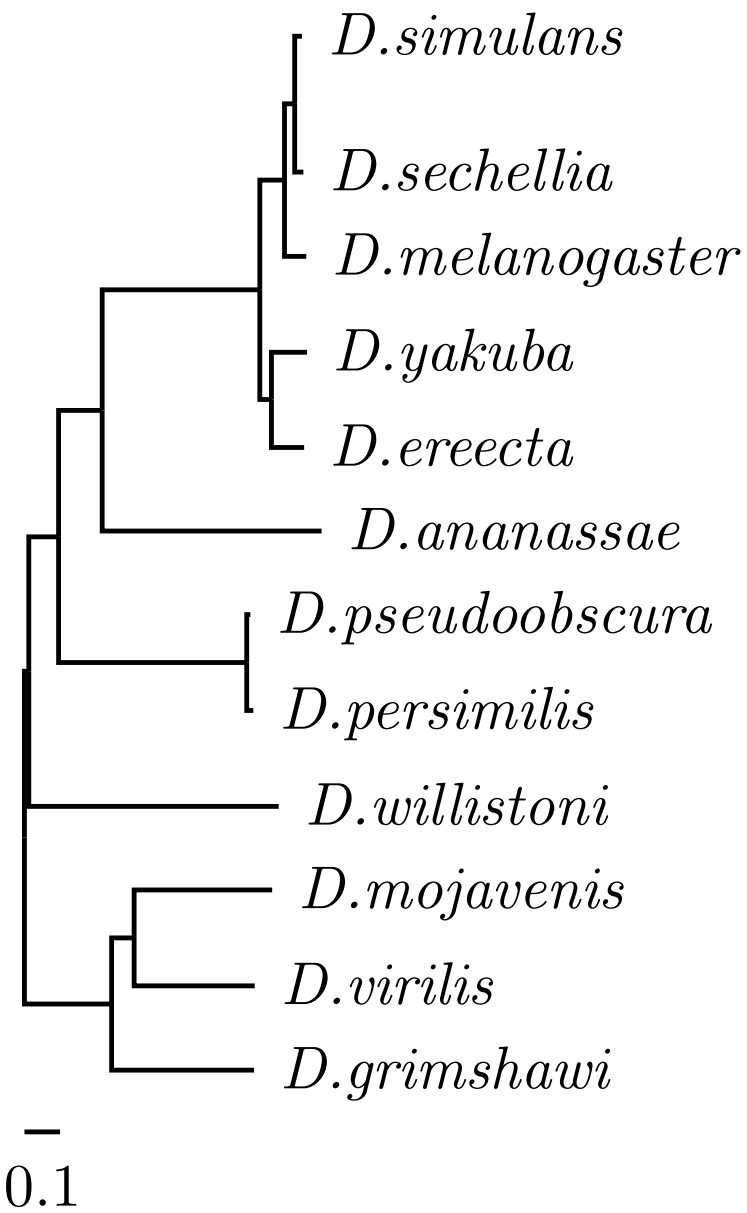


Algorithm 3

- Leaf nodes: $L_{s_k}(\mu) = 1$ if state s_k is in alignment, otherwise $L_{s_k}(\mu) = 0$
- Transition matrix $P_{s_k s_i}(t_i, \mu)$ gives probability for changing s_i into s_k in time t_i given mutation rate μ
- Derived from a rate matrix Q containing instantaneous substitution rates: $P_{xy}(t, \mu) = [e^{t\mu Q}]_{xy}$
- Q is a substitution rate model,use HKY85 atm

Test data set

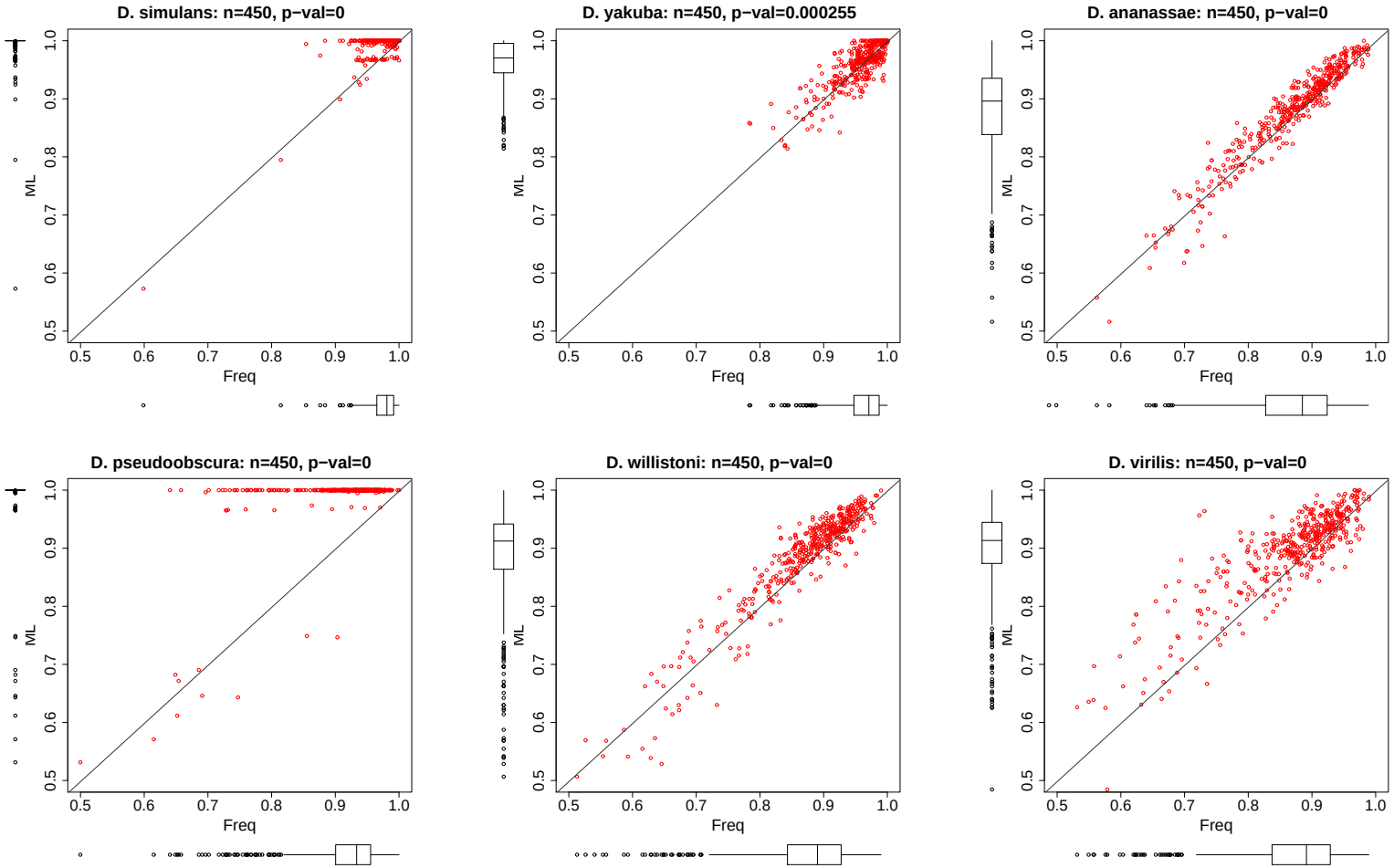
- Full genome multiz alignments of 12 Drosophila species
- Only take alignments which contain all 12 species
- Two sets of alignments:
  - *Set1* (n=45) with high alignment scores with 76.1% avg. sequence id
  - *Set2* (n=56) lower multiz scores, 67.1% id
- Columns with gaps were excluded



Evaluation

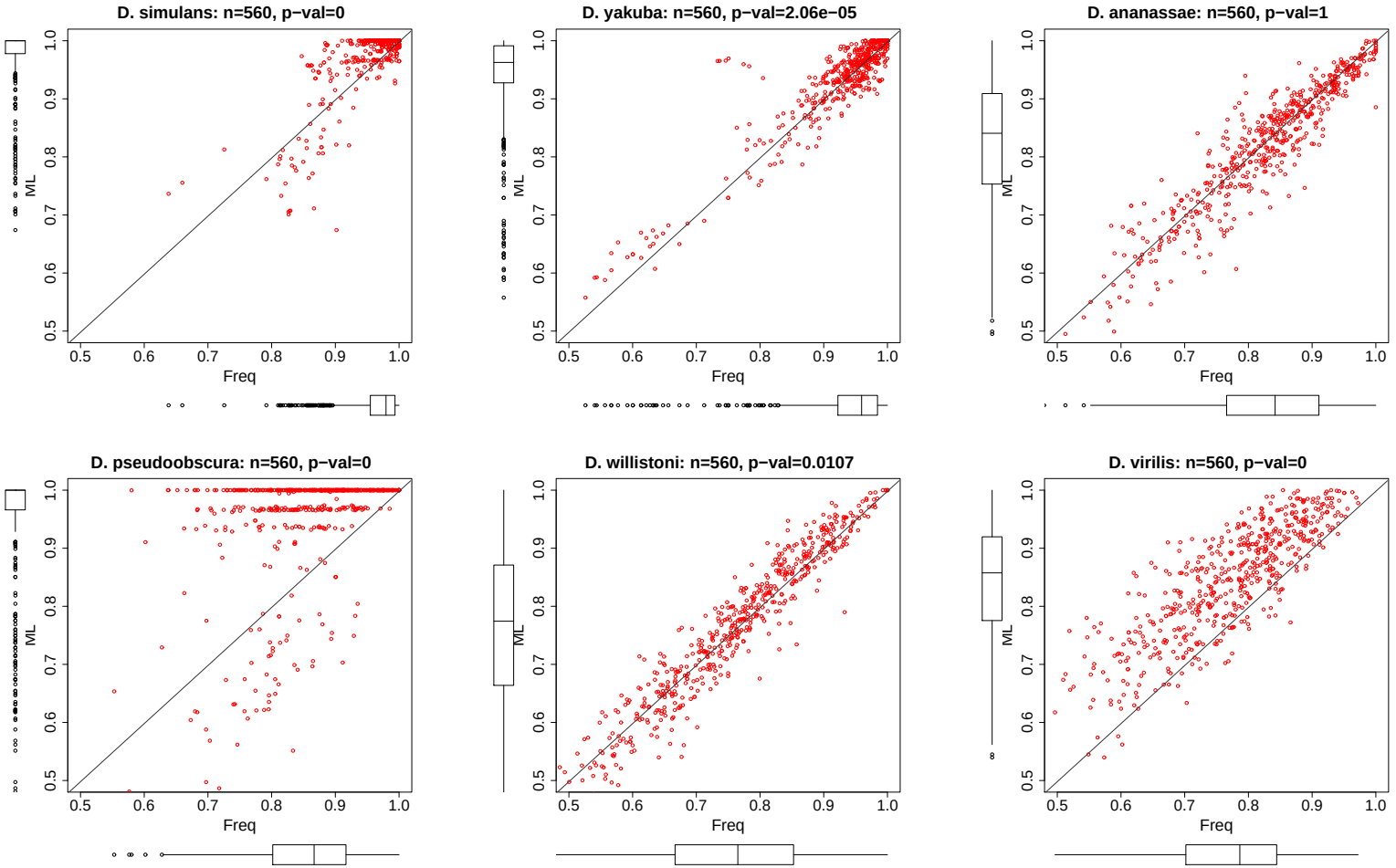
- Remove one species at a time from each alignment
- Calculate residue probabilities for this species with ML algorithm from the remaining 11 sequences
- HKY85 model, transition parameter κ estimated with `paml`
- Convert probabilities to PSSM
- Create another PSSM derived from counting nucleotide frequencies from the 11 species alignment
- `MATCH` similarity score gives a qualitative measure how well a PSSM fits to a nucleotide sequence, score is between 0 and 1
- Randomly select 10 windows of length 30nt from each alignments and match both PSSMs to the previously removed sequence

Drosophila data set: Match scores ML vs Freq



Set 1: `MATCH` scores for ML and Freq ($n = 450$) for randomly drawn windows of length 30nt.

Drosophila data set: Match scores ML vs Freq



Set 2: `MATCH` scores for ML and Freq ($n = 560$) for randomly drawn windows of length 30nt.

Comparison of median match scores

Species	Data set 1			Data set 2		
	ML	Freq	Δ	ML	Freq	Δ
D. sim.	1.000	0.981	0.019	1.000	0.980	0.020
D. sec.	1.000	0.981	0.019	1.000	0.975	0.025
D. mel.	0.986	0.979	0.007	0.970	0.972	-0.002
D. yak.	0.970	0.971	-0.001	0.963	0.959	0.003
D. ere.	0.971	0.972	-0.001	0.959	0.959	0.000
D. ana.	0.896	0.885	0.011	0.841	0.842	-0.001
D. pse.	1.000	0.933	0.067	1.000	0.867	0.133
D. per.	1.000	0.928	0.072	1.000	0.865	0.135
D. wil.	0.912	0.890	0.022	0.774	0.765	0.009
D. moj.	0.912	0.882	0.030	0.838	0.772	0.066
D. vir.	0.913	0.891	0.022	0.858	0.787	0.071
D. gri.	0.877	0.864	0.013	0.824	0.759	0.065

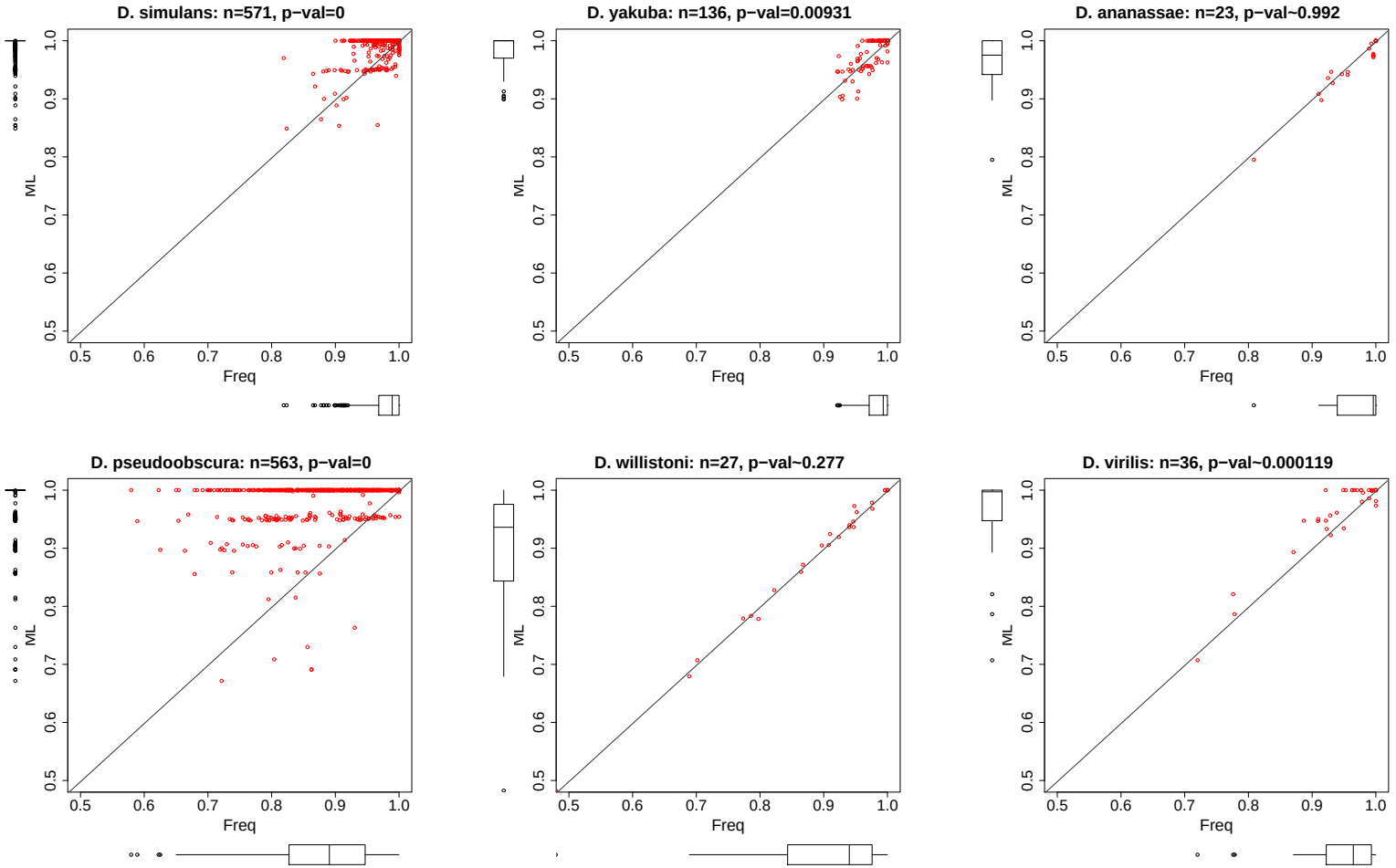
Extracting significant patterns

- Find informative regions that can be used as search pattern
- Shannon information content for each column i (for uniform background distribution):

$$I(i) = 2 - \sum_{s} f_i(s) \cdot \log_2 f_i(s)$$

- Extract windows with certain minimum length and average information content above a user-defined threshold

Drosophila data set: Match scores ML vs Freq



Set2: `MATCH` scores for all non-overlapping windows of length 20 with average information content $I \geq 1.8$.

Discussion

- In most species, the maximum likelihood PSSMs yield higher `MATCH` scores than the frequency-based PSSMs
- For phylogenetically distant species, like *D. willistoni* or *D. ananassae*, no significant difference between both PSSMs for randomly drawn samples
- Average improvement of `MATCH` scores tends to be higher in *Set2* with lower pairwise sequence identity
- When extracting windows with high average IC, we gain significant improvement in match scores in most species,but for distant species the ML approach only gives rise to highly conserved alignment blocks, no difference to Freq PSSMs

Consensus sequences

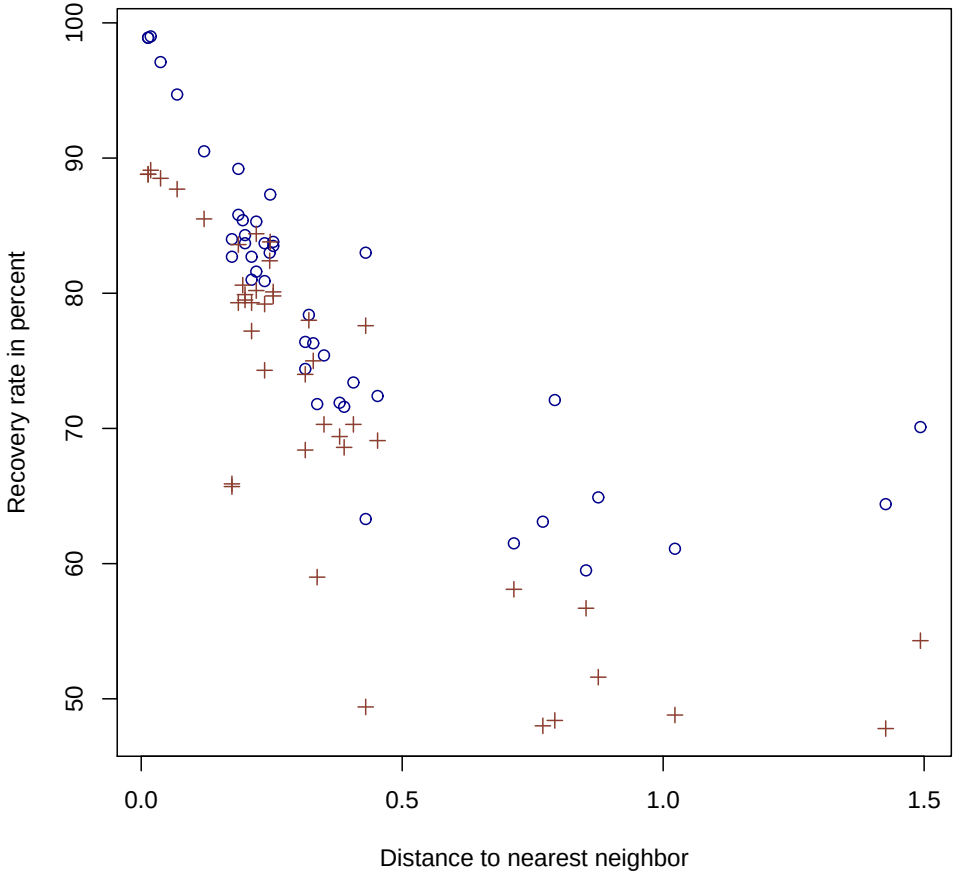
- derive consensus sequences from nucleotide probabilities
- use probability cutoffs, e.g. 0.5

Performance Test

- compare reconstruction ability from maxAlike consensus sequences and nucleotide frequency consensus
- Multiz44 whole genome alignments

Species	Dist.	Threshold 0.5		No threshold		
		Freq	*ML*	*Freq*	*ML*	Nt.pred.
hg18	0.013	88.8	98.9	87.3	98.9	92089
gorGor1	0.018	89.1	99.0	87.5	99.0	61291
ponAbe2	0.037	88.5	97.1	87.0	97.1	88214
calJac1	0.120	85.5	90.5	84.2	90.4	83033
rn4	0.174	65.7	84.0	64.8	83.5	59323
bosTau4	0.186	79.3	85.8	78.1	84.0	76322
felCat3	0.199	79.9	84.3	78.8	83.1	53158
micMur1	0.220	84.4	85.3	83.2	84.5	63285
proCap1	0.236	74.3	80.9	73.3	78.1	47590
equCab2	0.247	83.8	87.3	82.6	86.7	82307
galGal3	0.337	59.0	71.8	56.4	58.9	2000
eriEur1	0.407	70.3	73.4	69.4	71.4	20764
fr2	0.430	77.6	83.0	76.3	80.6	56502
tetNig1	0.430	49.4	63.3	47.1	55.6	3203
gasAcu1	0.770	48.0	63.1	46.6	52.7	3218
oryLat2	0.792	48.4	72.1	47.0	57.9	3234
anoCar1	0.876	51.6	64.9	49.4	53.0	3021
petMar1	1.023	48.8	61.1	46.7	42.4	2833
danRer5	1.426	47.8	64.4	46.0	50.8	4378

Performance difference



maxAlike - web server

Submit Results Help Example

Results

The Job with the ID 205088-2333947107 is **finished**.

Input

- alignment file
- phylogenetic tree
- Name of target species: **canFam2**

(a)

Output

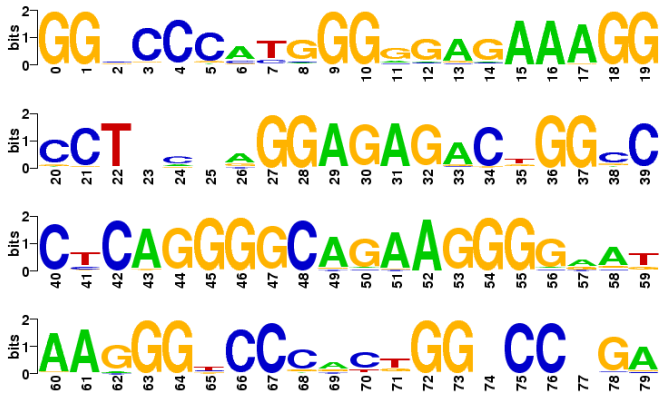
Residue probabilities in target species for each alignment column: (Download this file)

```
0       pair=-1 max_mu=0        IC=2    A=0     G=1     C=0     T=0
1       pair=-1 max_mu=0        IC=2    A=0     G=1     C=0     T=0
2       pair=-1 max_mu=4.40561  IC=0.191463     A=0.157277      G=0.321235      C=0.419056      T=0.
3       pair=-1 max_mu=0.725954 IC=1.36502      A=0.0327656     G=0.050808      C=0.896299      T=0.
4       pair=-1 max_mu=0.405962 IC=1.58541      A=0.0188324     G=0.0292024     C=0.940397      T=0.
5       pair=-1 max_mu=1.09914  IC=1.15784      A=0.0480687     G=0.074537      C=0.847753      T=0.
6       pair=-1 max_mu=1.04938  IC=0.693955     A=0.681427      G=0.0839606     C=0.201352      T=0.
7       pair=-1 max_mu=0.510734 IC=0.862124     A=0.0280441     G=0.0434865     C=0.19959       T=0.
8       pair=-1 max_mu=1.8815   IC=0.926729     A=0.0781329     G=0.788269      C=0.0836965     T=0.
9       pair=-1 max_mu=0        IC=2    A=0     G=1     C=0     T=0
10      pair=-1 max_mu=0        IC=2    A=0     G=1     C=0     T=0
11      pair=-1 max_mu=1.89993  IC=0.656739     A=0.177376      G=0.684152      C=0.0879663     T=0.
```

(b)

Sequence logo

Nucleotide probabilities as sequence logo:



(c)

Consensus sequences

Predicted consensus sequence, with a probability cutoff 0.5 at each site:

```
>canFam2 predicted consensus sequence with probability cutoff 0.5
GGNCCCATGGGGGAGAAAGGCCTNCNAGGAGAGACTGGCCCTCAGGGGCAGAAGGGGAATAAGGGTCCCACTGGNCCNGA
```

Submit to Primer3

Predicted consensus sequence, without cutoff:

(d)

```
>canFam2 predicted consensus sequence
GGCCCCATGGGGGAGAAAGGCCTCCGAGGAGAGACTGGCCCTCAGGGGCAGAAGGGGAATAAGGGTCCCACTGGTCCGGA
```

Submit to Primer3

Conserved Elements

Subsequences having a window length of 10nt with average information content above 1.5:

Summary

- Method for prediction of nucleotide probabilities in a target species
- based on known homologs and phylogenetic relationships between the species
- PSSMs can be used for homology search (by extracting highly informative patterns)
- Consensus sequences for primer design
- Outlook: incorporate base pair substitution models

Acknowledgements

- Peter Stadler, Jan Gorodkin
- People at Bioinformatics groups in Copenhagen and Leipzig

Thanks

Thank you for your attention.