maxAlike: Sequence Reconstruction by Maximum Likelihood Estimation

Peter Menzel [1,2], Peter F. Stadler [2–6] and Jan Gorodkin [1,*]

[1]Center for non-coding RNA in Technology and Health, Division of Genetics and Bioinformatics, IBHV, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg, Denmark
[2]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
[3]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
[4]Fraunhofer Institut für Zelltherapie und Immunologie, Perlickstraße 1, D-04103 Leipzig, Germany
[5]Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Vienna, Austria
[6]The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico

Received

ABSTRACT

The task of reconstructing a piece of sequence in a particular species is gaining more and more importance in the light of the rapid development of high-throughput sequencing technologies and their limitations. Applications include not only compensation for missing data in unsequenced genomic regions but also the preparation of customized queries for homology-based searches. Here, we introduce the maxAlike web server. It takes a multiple sequence alignment and a phylogenetic tree that also contains a target species as input. For the target species, it computes nucleotide probabilities as well as the most likely sequence, which can be used for primer design or homology search. Furthermore, position specific scoring matrices (PSSMs) of regions of high confidence are available for download. We show that as much as 99% of a sequence can be reconstructed correctly using the maxAlike algorithm, when the sequence of a closely related species is available, compared to only 89% reconstructed positions using only the consensus sequence from the input alignment. For more distant species, the reconstruction rate of maxAlike drops to a plateau value of about 60–70% for the maxAlike approach, compared to 50–60% for the consensus sequence. The web server is freely accessible at: http://rth.ku.dk/resources/maxAlike.

INTRODUCTION

With increased opportunities for high-throughput sequencing, many more organisms will be sequenced in near future. Due to inherent limitations in these technologies, it is also likely that routinely produced genomic sequences will be incomplete at various levels, but it is still essential to infer as much knowledge as possible about these missing regions in relation to specific analyses. On the other hand, the relative position of the target species in the evolutionary tree of species is typically known or can be inferred from the sequence information that is already available. With the rapidly growing collection of sequence information from diverse organism it also becomes increasingly important and useful to utilize not only consensus sequences but also the information contained in the patterns of variation. We present a web server – maxAlike – that aims at reconstructing sequences in a particular target species, using a phylogenetic tree and sequences from other species (1).

The maxAlike algorithm uses a multiple sequence alignment and a corresponding phylogenetic tree annotated with phylogenetic distances to estimate substitution rates for each column. The same tree augmented by the target species is then employed to infer the nucleotide probabilities for the homologous sequence in the target species. The technical details of applying this approach to homology search are discussed in (1). Clearly, the homology search approach can be extended to annotate partial sequences. Sequence reconstruction may be of particular interest for the targeted (re-)sequencing of a particular region that is not or not completely represented e.g. in a genome assembly. The web server therefore links the reconstructed sequence directly to the `Primer3` web server (2) for designing appropriate primers for the target species.

Related work includes in particular the `primers4clades` server (3), which derives a phylogenetic tree from a multiple sequence alignment as input. From this, the user can, through manual intervention though, restrict the input sequences to a phylogenetic group which is considered for the primer design. We have decided against estimating the tree from the input alignment because the usually relatively short regions of interest in the alignments often yield poor phylogenetic estimates. The `uniprime2` server (4) employs a pipeline of publicly available homology search, multiple alignment and primer design software to derive primers from conserved parts of a given gene. However, no phylogenetic information is taken into account.

*To whom correspondence should be addressed, Tel: +45 353 33578; Fax: +45 353 333042; Email: gorodkin@genome.ku.dk

SEQUENCE RECONSTRUCTION

The maxAlike algorithm (1) aims at reconstructing a sequence homolog in a given target species by employing a maximum likelihood computation over a phylogenetic tree, which relates the target species to species with already known homologs. This approach is in spirit similar to the reconstruction of ancestral sequences from their extant offsprings, see e.g. (5) for a review. The input for the computation is a multiple sequence alignment M and a phylogenetic tree T, which represents the phylogenetic relationships and distances among the species. Additionally, one of the species in the tree is chosen as the target species for the reconstruction. In a first step, T is restricted to the sequences contained in the alignment, and a relative substitution rate $\hat{\mu}_i$ is estimated for each alignment column i, by performing a maximum likelihood computation over T, which follows Felsenstein's pruning algorithm (6). The HKY85 (7) substitution model is used for computing the nucleotide transition probabilities. In a second step, we re-root T to the target species and use the estimated $\hat{\mu}_i$ to compute the likelihood of the tree again. From the likelihoods of each state at the root of T, we directly obtain the nucleotide probabilities in each alignment column i.

The calculated probabilities depend explicitly on the relative position of the target species to the other species in T. If the target is in close proximity to one or several other species with known homologs, then high probabilities will be assigned to the nucleotides present in these neighboring species. With increasing distance to its closest neighbor, the residue probabilities in the target species will converge to an equilibrium distribution, which depends on the parameters of the substitution model. This equilibrium is reached faster, the higher the substitution rate $\hat{\mu}_i$ is. The algorithm thus tells us which alignment columns or regions can be expected to be informative for a particular target sequence. To this end, we also compute the Shannon information content at each site, from which in turn, we can derive subsequences exceeding a user-defined minimum length that have an average information content above a certain threshold.

WEB SERVER

Input

Two files are required as input to the web server. First, the user must supply a multiple sequence alignment of homologous DNA or RNA sequences. The second input file is a phylogenetic tree in Newick format, where the species names must match the sequence names in the alignment file. Branch lengths are required. Since only relative values are required, these can be taken from a broad variety of sources. The tree may contain fewer or more species than the sequence alignment. Only taxa contained in the tree will be considered in the computation. Example trees for different groups of organisms are provided on the help page of the web server. At last, the user has to specify one of the species names from the phylogenetic tree – the target species for sequence reconstruction. Additionally users can enter their e-mail address to receive a notification including a link to the results, once the job has finished running.

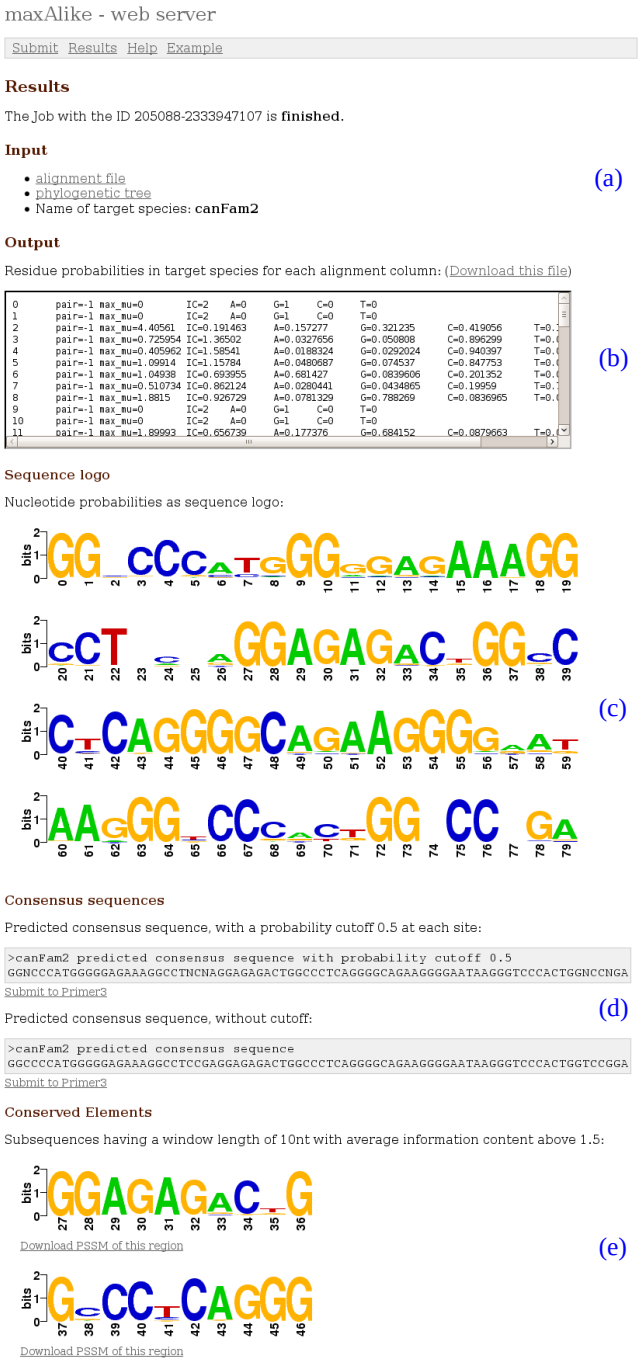

Figure 1. Screen-shot of the maxAlike web server's result page: **(a)** Input data **(b)** Text output of nucleotide probabilities for each site **(c)** Probabilities converted to sequence logo for entire sequence **(d)** Consensus sequences, with and without probability cutoff **(e)** Sequence windows with average information content and length above a user-defined threshold

Output

The results of each submission are displayed on a HTML page (see figure 1), which is accessible up to 30 days after completion of the computation. The main output is a text file which contains one line for each column of the

alignment. An entry contains the probabilities of occurrence for each nucleotide in the target species. Additionally the Shannon information content and the estimated mutation rate $\hat{\mu}$ are included. Columns in the input alignments that contain gaps will not be predicted and get zero probability for each nucleotide. The nucleotide probabilities are converted to a sequence logo, using the information content of each alignment column. This logo gives a visual representation of the expected sequence in the target species, and highlights sites with high predicted nucleotide probabilities. Additionally, two "consensus" sequences are constructed from the nucleotide probabilities: one sequence taking the most probable base at each site into account, and another sequence which only considers nucleotides having probabilities above a user defined threshold. Possible gaps and sites with probabilities below the threshold are shown as "N".

The consensus sequences can then be used for submission to primer design software, or directly submitted to the Primer3 web server (2). The last part of the output contains a list of non-overlapping subsequences of a minimum length yielding an average information content greater than a user-defined threshold. Position specific scoring matrices for these subsequences are also available for download. These PSSMs can directly be used with homology search programs, e.g. fragrep2 (8).

Implementation

The web server runs Apache (www.apache.org) and uses `PHP` programs (www.php.net) to submit jobs to a queueing system that distributes jobs on a compute cluster. The maxAlike algorithm is implemented in C++, making use of the `Bio++` libraries (9). Sequence logos are created with the `weblogo` package (10).

PERFORMANCE EVALUATION

We tested the prediction performance on several vertebrate full-genome multiple alignments. The first data set (*MZ44*) is derived from the human genome (hg18) alignments to 43 vertebrate species (multiz44way)[1]. The second data set (*ENC*) is derived from the December 2007 ENCODE (11) Multi-Species Sequence Analysis sequence freeze, which contains sequences orthologous to the human ENCODE regions from 36 vertebrate species[2].

Both data sets were divided into two parts each: the first part contains alignments with higher sequence conservation (*MZ44-1*: multiz score from 1M to 4M and at least 20 species per alignment with minimum length 200nt, number of alignments $n=254$, *ENC-1*: multiz score from 1M to 3M and at least 15 species per alignment with minimum length 300nt, $n=114$), whereas the second part contains alignments with lower sequence conservation only from chromosome 1. (*MZ44-2*: multiz score from 10k to 1M and at least 20 species per alignment, $n=886$, *ENC-2*: multiz score from 10k to 1M and at least 10 species per alignment, $n=1276$). Columns containing gaps were not included in the computation and thus removed from the alignments.

From each alignment, we removed one species at a time and used the remaining sequences to reconstruct the homolog in the removed species (target species) with the maxAlike algorithm using the phylogenetic trees from the multiz and ENCODE phastCons models respectively (see supplementary materials). The transition bias parameter κ for the HKY85 substitution model is estimated using `PAML` (12). The predicted nucleotide probabilities were converted into position specific scoring matrices (PSSM) for each predicted sequence. In addition, consensus sequences were created by considering the most probable nucleotide at each site with a probability greater than a threshold of 0.5 and by considering the most probable nucleotide without using a threshold. We also created PSSMs by counting the nucleotide frequencies in each input alignment and derived consensus sequences, again using either a relative frequency threshold of 0.5 or no threshold, respectively. In order to evaluate the effect of including phylogenetic information into the PSSM and consensus sequence construction, we compared the predictions from the maxAlike probabilities (*ML*) and the nucleotide frequencies (*Freq*).

For the comparison of the *ML* and *Freq* PSSMs, we computed the MATCH scores (13) for both the *ML* and *Freq* PSSMs using the previously removed sequence as reference. The MATCH algorithm is designed for matching short matrix profiles (originally for transcription factor binding sites) to a primary sequence, thus we restrict the tested PSSMs to windows of size 30nt and randomly draw 10 windows from each alignment to achieve a minimum coverage of the predicted sequences. The MATCH score takes values between 0.0 and 1.0, with 1.0 denoting a perfect match of the matrix to the sequence.

Table 1 compares the median match scores for all species of the *MZ44-2* data set. For almost all species, we see a significant improvement of the score when using the *ML* PSSMs compared to the *Freq* PSSMs. In particular, predictions for target species with a close neighbor in the tree gain most from the inclusion of phylogenetic information to reconstruction algorithm. All bony fishes (teleostei) show improved scores, since the sequences from the tetrapoda have much less impact on prediction of the homologs. In frequency matrices the large fraction of mammals among the sequences genomes causes a substantial bias. Conversely, the impact of bony fish sequences on mammalian targets is highly overestimated in *Freq* PSSMs compared to the *ML* PSSMs.

The full-length consensus sequences from each alignment were evaluated in terms of the percentage of correctly predicted nucleotides at each position compared to the previously removed homolog in the target species. We excluded absolutely conserved sites from the evaluation, since both methods perform identically on this subset. Only species with at least 10 predicted sequences were evaluated to retain a sufficient sample size. Table 2 shows the recovery rates in percent of both the *ML* and the *Freq* consensus sequences for threshold 0.5 and no threshold, resp., in the *MZ44-2* data set. We observe similar effects as for the PSSM comparison. In general the amount of correctly predicted positions is higher in the *ML* consensus sequences compared to the *Freq* consensus sequences. When requiring a probability of at least

[1] http://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz44way/
[2] http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encode/MSA/DEC-2007/

Table 1. Data set *MZ44-2*: Median MATCH scores of the maxAlike PSSMs (*ML*) and the frequency PSSMs (*Freq*) for 10 randomly selected 30nt windows from each alignment. The last column Δ shows the difference between both medians. Only species with more than 100 predicted PSSMs were included. The Dist. column shows the distance to the phylogenetically nearest neighbor in the tree.

Species	Dist.	*ML*	*Freq*	Δ
hg18	0.013	1.000	0.957	0.043
panTro2	0.013	1.000	0.957	0.043
gorGor1	0.018	1.000	0.958	0.042
ponAbe2	0.037	0.997	0.954	0.043
rheMac2	0.069	0.967	0.946	0.021
calJac1	0.120	0.934	0.925	0.010
mm9	0.174	0.895	0.760	0.135
rn4	0.174	0.895	0.762	0.134
bosTau4	0.186	0.901	0.870	0.030
turTru1	0.186	0.935	0.908	0.026
vicPac1	0.195	0.898	0.880	0.018
canFam2	0.199	0.890	0.869	0.021
felCat3	0.199	0.903	0.876	0.027
choHof1	0.211	0.890	0.867	0.023
dasNov2	0.211	0.872	0.847	0.025
micMur1	0.220	0.901	0.899	0.002
otoGar1	0.220	0.871	0.864	0.006
loxAfr2	0.236	0.900	0.867	0.034
proCap1	0.236	0.867	0.822	0.045
tarSyr1	0.246	0.875	0.884	-0.008
equCab2	0.247	0.908	0.905	0.003
myoLuc1	0.253	0.881	0.866	0.015
pteVam1	0.253	0.885	0.877	0.008
ochPri2	0.314	0.811	0.769	0.042
oryCun1	0.314	0.850	0.822	0.028
tupBel1	0.321	0.840	0.847	-0.008
speTri1	0.330	0.833	0.830	0.002
galGal3	0.337	0.840	0.758	0.083
echTel1	0.350	0.829	0.794	0.035
cavPor3	0.380	0.795	0.781	0.014
dipOrd1	0.389	0.794	0.777	0.017
eriEur1	0.407	0.799	0.787	0.012
fr2	0.430	0.780	0.688	0.091
tetNig1	0.430	0.745	0.696	0.048
sorAra1	0.453	0.795	0.774	0.021
monDom4	0.714	0.710	0.680	0.030
gasAcu1	0.770	0.761	0.674	0.087
oryLat2	0.792	0.835	0.699	0.136
ornAna1	0.852	0.727	0.686	0.041
anoCar1	0.876	0.791	0.686	0.105
petMar1	1.023	0.732	0.649	0.082
danRer5	1.426	0.767	0.648	0.119
xenTro2	1.493	0.817	0.732	0.085

0.5 for a valid prediction, the prediction quality becomes better on average. Results for the *MZ44-1* and the two *ENC* data sets are similar, see supplementary materials for additional tables. In the *MZ44-2* data set the difference in the prediction performance is smaller compared to *MZ44-1*. This is explained by the higher overall sequence conservation in the alignments, which decreases the impact of phylogenetic tree information in the predictions.

Figure 2 shows the correlation between the overall recovery rate in data set *MZ44-2* in each target species and its distance to its phylogenetically closest neighbor. Species in the upper left corner are primates, both points in the bottom right corner are frog and zebrafish.

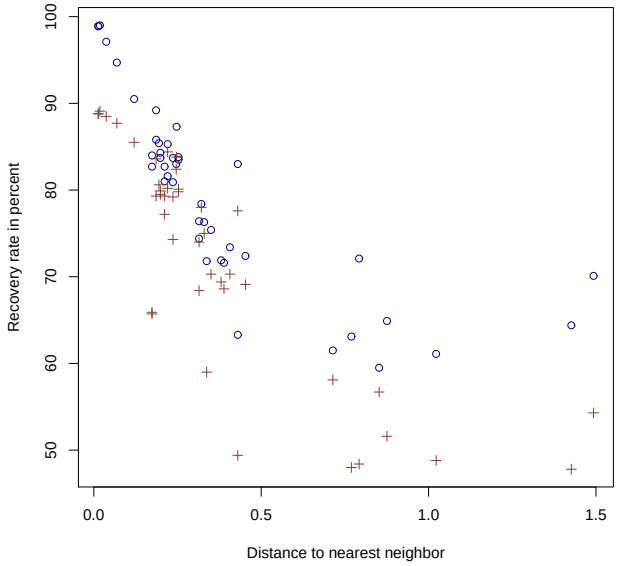

Figure 2. Recovery rate in percent for each species compared to its distance to the phylogenetically closest neighbor. Points denoted by a circle are *ML* consensus sequences, points denoted by a plus are *Freq* consensus sequences. The recovery rate only takes sequence positions with a nucleotide probability ≥ 0.5 into account.

DISCUSSION

The maxAlike web server estimates the nucleotide probabilities at each sequence position and estimates a reconstructed consensus sequence using a combination of homology information from a multiple sequence alignment and a phylogenetic tree. We have demonstrated that phylogenetic information significantly improves the prediction performance by about 10% in all target species compared to standard models based on consensus sequences and position frequency matrices. The prediction results improve when the target species is surrounded by an increasing number of closely related species with known homologs. The reconstruction rate can reach as much as 99% accuracy.

In the typical scenario, a phylogenetic tree which contains the target species needs to be available first. Ideally this tree is constructed by all the already available sequence data from the target species, combined with the homologs of these sequences in other closely related model organisms. The more sequence data is included in the tree construction, the more accurate the relative position of the target species within its phylogenetic neighborhood can be determined. An automation

of this process might be an further enhancement of web server. On the other hand, more and more phylogenetic trees become available in public databases, e.g. (14). Other extensions might include a calculation of the sequence logos with respect to a non-uniform background distribution, as outlined e.g. on the `MatrixPlot` web server (15).

SUPPLEMENTARY MATERIALS

The supplementary materials contain both phylogenetic trees and the results of the performance evaluation of all data sets.

FUNDING

PM is supported by the Danish research council for Technology and Production through and the Danish research school in biotechnology. This work was supported by the Danish Center for Scientific Computation.

Conflict of interest statement. None declared.

REFERENCES

1. Menzel, P., Gorodkin, J., and Stadler, P. F. (2009) Maximum Likelihood Estimation of Weight Matrices for Targeted Homology Search. In Grosse, I., Neumann, S., Posch, S., Schreiber, F., and Stadler, P., (eds.), *Lecture Notes in Informatics: Proceedings of the German Conference on Bioinformatics 2009*, pp. 212–220.
2. Rozen, S. and Skaletsky, H. (2000) Primer3 on the www for general users and for biologist programmers.. *Methods Mol Biol*, **132**, 365–386.
3. Contreras-Moreira, B., Sachman-Ruiz, B., Figueroa-Palacios, I., and Vinuesa, P. (Jul, 2009) primers4clades: a web server that uses phylogenetic trees to design lineage-specific pcr primers for metagenomic and diversity studies.. *Nucleic Acids Res*, **37**(Web Server issue), W95–W100.
4. Boutros, R., Stokes, N., Bekaert, M., and Teeling, E. C. (Jul, 2009) Uniprime2: a web service providing easier universal primer design.. *Nucleic Acids Res*, **37**(Web Server issue), W209–W213.
5. Thornton, J. W. (May, 2004) Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat Rev Genet*, **5**(5), 366–375.
6. Felsenstein, J. (November, 1981) Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, **17**(6), 368–376.
7. Hasegawa, M., Kishino, H., and Yano, T. (1985) Dating the human-ape split by a molecular clock of mitochondrial dna. *Journal of Molecular Evolution*, **22**, 160–174.
8. Mosig, A., Sameith, K., and Stadler, P. F. (2006) `fragrep`: Efficient Search for Fragmented Patterns in Genomic Sequences. *Geno. Prot. Bioinfo.*, **4**, 56–60.
9. Dutheil, J., Gaillard, S., Bazin, E., Glmin, S., Ranwez, V., Galtier, N., and Belkhir, K. (2006) Bio++: a set of c++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics.. *BMC Bioinformatics*, **7**, 188.
10. Crooks, G., Hon, G., Chandonia, J., and Brenner, S. (2004) Weblogo: A sequence logo generator. *Genome Research*, **14**, 1188–1190.
11. ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1the encode pilot project.. *Nature*, **447**(7146), 799–816.
12. Yang, Z. (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol*, **24**(8), 1586–1591.
13. Kel, A., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O., and Wingender, E. (2003) Match: a tool for searching transcription factor binding sites in dna sequences. *Nucl. Acids Res.*, **31**(13), 3576–3579.
14. Morell, V. (1996) TreeBASE: The Roots of Phylogeny. *Science*, **273**(5275), 569–0.
15. Gorodkin, J., Staerfeldt, H. H., Lund, O., and Brunak, S. (Sep, 1999) MatrixPlot: visualizing sequence constraints. *Bioinformatics*, **15**, 769–770.

Table 2. Data set *MZ44-2*: Recovery rate in percent for consensus sequences derived from maxAlike probabilities (*ML*) and nucleotide frequencies (*Freq*), with a nucleotide probability/relative frequency above a 0.5 threshold at each site and without a threshold, i.e. consensus sequences contain nucleotides with highest probability/frequency at each site. A value of e.g. 70 means that 70% of the nucleotides in the consensus sequence were predicted correctly. The Dist. column shows the distance to the phylogenetically nearest neighbor in the tree. Only species with at least 10 reconstructed sequences were considered. The last column shows the total number of reconstructed nucleotides for each species, excluding fully conserved columns. See supplementary material, for the percentage of nucleotides above the 0.5 threshold for both methods.

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
panTro2	0.013	88.8	98.9	87.3	98.9	90693
gorGor1	0.018	89.1	99.0	87.5	99.0	61291
ponAbe2	0.037	88.5	97.1	87.0	97.1	88214
rheMac2	0.069	87.7	94.7	86.2	94.7	87992
calJac1	0.120	85.5	90.5	84.2	90.4	83033
mm9	0.174	65.9	82.7	65.0	81.7	65472
rn4	0.174	65.7	84.0	64.8	83.5	59323
bosTau4	0.186	79.3	85.8	78.1	84.0	76322
turTru1	0.186	83.6	89.2	82.4	88.4	75751
vicPac1	0.195	80.6	85.4	79.4	84.4	55289
canFam2	0.199	79.5	83.7	78.3	82.7	80552
felCat3	0.199	79.9	84.3	78.8	83.1	53158
choHof1	0.211	79.3	82.7	78.2	81.1	49130
dasNov2	0.211	77.2	81.0	76.2	79.0	48334
micMur1	0.220	84.4	85.3	83.2	84.5	63285
otoGar1	0.220	80.2	81.6	79.1	80.6	57445
loxAfr2	0.236	79.2	83.7	78.1	81.9	53365
proCap1	0.236	74.3	80.9	73.3	78.1	47590
tarSyr1	0.246	82.4	83.0	81.4	82.3	60541
equCab2	0.247	83.8	87.3	82.6	86.7	82307
myoLuc1	0.253	79.8	83.5	78.4	82.7	31309
pteVam1	0.253	80.1	83.8	78.9	83.1	59906
ochPri2	0.314	68.4	74.4	67.6	69.6	40361
oryCun1	0.314	74.0	76.4	73.1	74.0	49759
tupBel1	0.321	78.0	78.4	76.9	77.4	53481
speTri1	0.330	75.0	76.3	74.0	74.6	50565
galGal3	0.337	59.0	71.8	56.4	58.9	2000
echTel1	0.350	70.3	75.4	69.3	72.7	31358
cavPor3	0.380	69.4	71.9	68.5	69.7	70481
dipOrd1	0.389	68.6	71.6	67.7	69.2	41512
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
sorAra1	0.453	69.1	72.4	68.1	70.3	22024
monDom4	0.714	58.1	61.5	56.8	56.6	11424
gasAcu1	0.770	48.0	63.1	46.6	52.7	3218
oryLat2	0.792	48.4	72.1	47.0	57.9	3234
ornAna1	0.852	56.7	59.5	55.9	54.3	3610
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
xenTro2	1.493	54.3	70.1	52.0	51.4	3660