MASS-PRF Manual

Model Averaged Clustering Poisson Random Field

Version 1.1

Zi-Ming Zhao, Ning Li, Zhang Zhang, and Jeffrey P. Townsend

Department of Biostatistics, Yale University

(http://www.yale.edu/townsend/software.html)

December 14, 2016

COPYRIGHT & LICENSE:

MASS-PRF is distributed as open-source software and licensed under the GNU General Public License (Version 3; http://www.gnu.org/licenses/gpl.txt), in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

REFERENCE:

Zi-Ming Zhao, Ning Li, Zhang Zhang, and Jeffrey P. Townsend. (2016) Regions within coding gene sequences experience diverse intensities of natural selection inferred from polymorphism and divergence.

SEE ALSO:

For more information, please see <http://www.yale.edu/townsend/software.html>.

CONTACT:

Please send suggestions or bugs to Zi-Ming Zhao ([ziming.gt@gmail.com)](mailto:ziming.gt@gmail.com)) and Jeffrey Townsend (Jeffrey.Townsend@Yale.edu).

Table of Contents

Introduction 4

Run MASS-PRF 4

How to compile? 4

Required inputs and input formats 4

How to run MASS-PRF? 5

Command for help 5

Run MASS-PRF for an example gene 8

Input files 8

Run MASS-PRF 8

Output files 9

The speed of MASS-PRF 9

Other notes 11

Major updates from last version (v1.0) 12

Gamma calculation using four Lookup Tables 12

Criteria for gamma values 13

Model averaged gamma 13

Exhaustive vs stochastic algorithms for gamma 13

New features 13

Non-coding MASS-PRF 14

Existing problems 14

References 16

## Introduction

MASS-PRF is a software package for estimating selection intensity across sites within coding sequences by using polymorphism and divergence data. MASS-PRF stands for Model Averaged Clustering – Poisson Random Field. It is written by C++, and has been compiled in Linux, Mac, and Windows environment. To address whether the intensity of natural selection is intragenically variable, we apply model-averaged clustering of polymorphism and divergence ([1](#_ENREF_1)), implemented within Sawyer and Hartl’s PRF model ([2](#_ENREF_2)).

## Run MASS-PRF

The MASS-PRF algorithm is run on a command-line called MASS-PRF. If you work with Linux, Unix or Mac, you will work at a shell. If you work with Windows, you will need a command window (Cygwin works like Linux on Windows. gcc compiler package needs to be installed. See more details: <http://netbeans.org/community/releases/60/cpp-setup-instructions.html> ).

Source code is open and free. Complied executed programs are in /bin/ that are for **Linux, Mac and Windows.** Please uncompact the file, and the four Lookup Tables are required in the folder where the program is running.

### How to compile?

1. Within the /MASSPRF main folder, type “make” in Linux or Mac terminal, and the compiled command should be found in /bin/ folder. [make sure there is a ‘bin’ folder in the same folder ‘Makefile’ is in before compiling.]

or

1. Within the folder with all source codes, you can use the compilation command on OS X 10.9.

g++ -O3 -o massprf MASSprf.cpp PRFCluster.cpp base.cpp –w

or

1. The compilation command on Unix or OS X 10.8 or before system is as blow.

c++ -O3 -o massprf MASSprf.cpp PRFCluster.cpp base.cpp –w

### Required inputs and input formats

Two formatted input files in the fasta format are required. One contains polymorphic sequences from one species. The other one contains divergent sequence from another, closely-related species. The sequences must have already been aligned so that sites are ordered as homologous, and gaps have to be removed. Because MASS-PRF is intended for coding sequences, we recommend codon-based approaches for alignment. The sequences must be nucleotide sequences. The two files can be in the fasta format or consensus fasta input format can be used. See the folder ./examples/, in which fasta format files include Attacin-C\_DmDs\_div.fas and Attacin-C\_DmDs\_pol.fas; consensus format input files include consensus\_divergence.txt and consensus\_polymorphism.txt.

### How to run MASS-PRF?

Run polymorphism and divergence sequences, print the output to the screen or direct the output into an output file.

./massprf -p ‘input polymorphism file’ -d ‘input divergence file’ [Other Options] >OutputFile.txt &

### Command for help

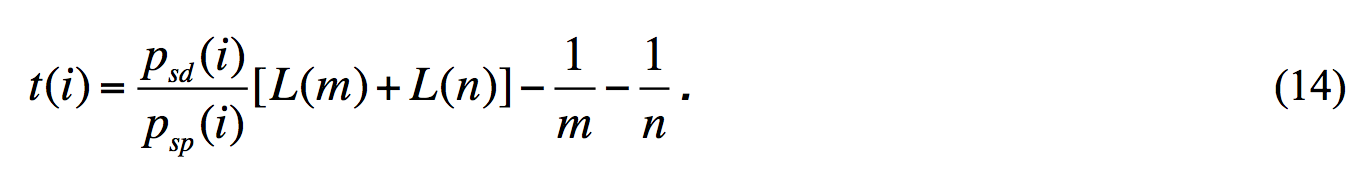
This command is also listed in **Table 1**.

./massprf -h

**Table 1: Input parameters for MASS-PRF**

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter | Values | Default Values | Required ? |
| -p | Polymorphism file | - | Yes |
| -d | Divergence file | - | Yes |
| -o | Choice of output format [integer, optional], {0: amino acid level output || 1: nucleotide level output} | 0 |  |
| -ic | Choice of input format for divergence and polymorphism sequences [integer, optional], {0: DNA sequences || 1: consensus sequences} | 0 |  |
| -sn | Input the number of polymorphism sequences [integer] | - | Required when  -ic=1 |
| -c | Criterion used for clustering  0:BIC; 1:AIC; 2:AICc; 3:LRT | 0 | No |
| -g | Standard genetic codes  (see: <http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi> ) | 1 | No |
| -m | 0: use both model selection and model averaging  1: use only model selection | 0 | No |
| -ci\_m | 95% confidence intervals for model averaging:  0: NOT calculate 95% confidence intervals  1: calculate 95% confidence intervals | 0 | No |
| -s | 0: without clustering results of synonymous sites  1: with clustering results of synonymous sites | 0 | Required when -ssd |
| -r | Estimate gamma  0: NOT estimate scaled selection coefficient  1: estimate scaled selection coefficient | 1 | No |
| -ci\_r | 95% confidence intervals for gamma  0: NOT calculate 95% confidence intervals  1: calculate 95% confidence intervals | 1 | No |
| -exact | Algorithm for 95% confidence intervals for gamma  0: use stochastic algorithm  1: use exact algorithm | 0 | No |
| -mn | Input the number of models to be used in stochastic algorithm [integer] | 10,000 | Optional when -exact=0 |
| -t | Species divergent time in the unit of MY (million years)  [number] | If neither –t or –ssd is used, the gene specific divergence\_time will be calculated based on the total count of SP and SD, as in the formula 10\* below. | No |
| -ssd | - | Site specific divergence\_time is calculated using synonymous clustering, by Formula 14\*\* below.​ | No, but recommended, especially for recombination. |
| -n | Nucleotide is replaced or seen as gap when it is not A, T, G or C  0: see it as gap  1: replace this nucleotide with the most frequently used nucleotide in other sequences | 1 | No |
| -NI | 0: NOT estimate Neutrality Index  1: estimate Neutrality Index | 0 | No |
| -v | Verbose output or not  0: not verbose, concise output  1: verbose output | 1 | No |
| -rMAp | Output gamma calculated using model averaged pr and dr  0: do not use model averaged pr and dr  1: use model averaged pr and dr | 0 | No |
| -h | - | - | - |

Formula 10\*:  

Formula 14\*\*: 

## Run MASS-PRF for an example gene

### Input files

One example of input sequences in ./example is named Attacin-C\_DmDs\_div.fas, Attacin-C\_DmDs\_pol.fas. The *Attacin-C* gene encodes a peptide, transcribed in response to bacterial or fungal infection ([3](#_ENREF_3), [4](#_ENREF_4)). The polymorphism data is from *Drosophila melanogaster*. The divergence data is from *Drosophila simulans*. The two species diverged approximately 2.5 million years ago ([5](#_ENREF_5)). Standard fasta format should be used, '>' should be put before the taxa name and line break should be used for each sequence including the last sequence.

### Run MASS-PRF

Example commands:

Run on the nucleotide fasta format (-t 2.5 means a user input divergence time 2.5 MYA)

./massprf -p ../example/Attacin-C\_DmDs\_pol.fas -d ../example/Attacin-C\_DmDs\_div.fas -o 1 -r 1 -ci\_r 1 -ci\_m 1 -s 1 -exact 0 -mn 30000 -t 2.5 >Attacin-C\_MASS-PRF\_BIC.txt

Run on the consensus input format (-ic 1 is used for consensus file format; and -sn is required when ic=1. For more details, type ./MASS-PRF –h)

./massprf -p ../example/consensus\_polymorphism.txt -d ../example/consensus\_divergence.txt -ic 1 -sn 11 -o 1 -r 1 -ci\_r 1 -ci\_m 1 -s 1 -exact 0 -mn 30000 -t 2.5 >output\_consensus\_MASS-PRF\_BIC.txt

Notes: Output format option: default amino acid sites, alternative nucleotide sites when -o=1.

Run with -ssd option (Site specific divergence time calculation ’-ssd’ should be always together with the option ’-s 1’)

./massprf -p Attacin-C\_DmDs\_pol.fas -d Attacin-C\_DmDs\_div.fas -o 1 -r 1 -ci\_r 1 -ci\_m 1 -s 1 -exact 0 -mn 30000 -ssd >output\_Attacin-C\_MASS-PRF\_SSD.txt &

### Output files

The output file in ./example/ named Attacin-C\_MASS-PRF\_BIC.txt is the analyzed result from analysis of the fasta format sequences Attacin-C\_DmDs\_div.fas and Attacin-C\_DmDs\_pol.fas. The file output\_consensus\_MASS-PRF\_BIC.txt is the analyzed result from analysis of the consensus format sequences consensus\_polymorphism.txt and consensus\_divergence.txt.

### The speed of MASS-PRF

MASS-PRF is computationally intensive; because it performs all possible sub-clusterings in an iterative manner in order to exhaustively sample potential spatial clustering of selection. Its speed scales sub-factorially with gene length (fig. 1). This figure was generated from an independent MASS-PRF analysis we are conducting using an alignment of 1825 genes including *Saccharomyces cerevisiae* polymorphism and *Saccharomyces paradoxus* orthologs. Runtimes were recorded for this data on nodes with eight 2.27GHz cores and 48GB memory in total. The fit line is a least squares linear fit between log(runtime) and gene length (L) with a slope of 0.004 and a y-intercept of 7.3, which is log(runtime) = 7.3 + 0.004L (fig. 1, panel B).

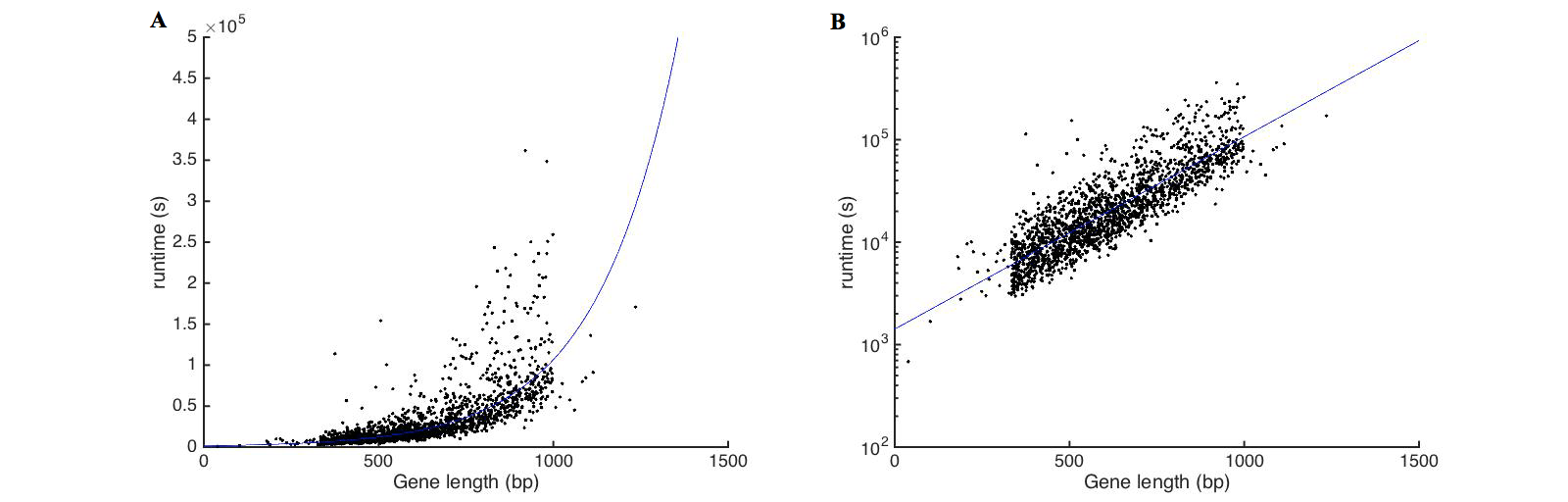


Figure 1: The speed of MASS-PRF on empirical data. The analyses include 1825 genes from alignments of *Saccharomyces cerevisiae* polymorphism *and Saccharomyces paradoxus* divergence sequences. X-axis is the gene length in base pair (bp), and y-axis is the runtime in seconds (s). The fit line is a least square linear fit against and independent axis of gene length (L), depicted for runtimes on an A) linear and on a B) log scale. Approximate log-linearity with gene length can be seen, with some contribution to longer run times arising from higher levels of polymorphism and / or divergence.

With a typical computer cluster, MASS-PRF is suitable for analyses of whole exomes, particularly if long genes (>900 base pairs) are scaled down to more practical lengths. For instance, if a gene is greater than 900 bp in length, it can be scaled by a factor. We recommend using the smallest multiple of 3 (size of each codon) that would produce a scaled consensus sequence less than 900 bp in length. By scaling, we mean dividing all positions by three and retaining all variant sites, then rescaling the result so that it is in the correct units and selective intensity. We are on track to eventually complete MASS-PRF analyses in this fashion on whole exome polymorphism and divergence data on the yeasts *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* and the fruit flies *Drosophila melanogaster* *and Drosophila simulans*, and are also conducting such analyses on other exome data conveying genomic polymorphism and divergence. With a non-trivial amount of patience and sufficient access to high-throughput computing, MASS-PRF is thus suitable for whole genome (or exome) analyses.

We have also performed analyses of 231 simulated genes with the same gene length 900 bp evolved under a neutral selection model. In particular, we used the Hudson’s ms (Hudson 2002) simulation method to generate coalescent trees, and used Fletcher and Yang’s Indelible (Fletcher and Yang 2009) to generate sequences evolved under a neutral codon substitution model (the two ratio model M1 with ω0= 0, ω1= 1 and p0=0.5) as demonstrated by Yang and Nielsen (Yang and Nielsen 2002). More details regarding the method and results of our simulation were included in our manuscript. Depending on the number and position of variant sites, the runtimes on these 231 genes vary by a highly right-skewed five- to ten-fold, even though all are exactly the same gene length (fig. 2). The averaged runtime is approximately 2.7 hours when run on a node of eight 2.27GHz cores with 48GB memory in total). Given that the human genome contains approximately 19,000 genes, we may estimate an upper limit of around 50,000 hours of total runtime for the entire human genome (an upper limit because shorter genes run much faster). If 50 genes could be run simultaneously, this analysis would take just under a month of walltime. Thus, this genome-wide analysis of coding sequences is possible with access to a high performance computer cluster.

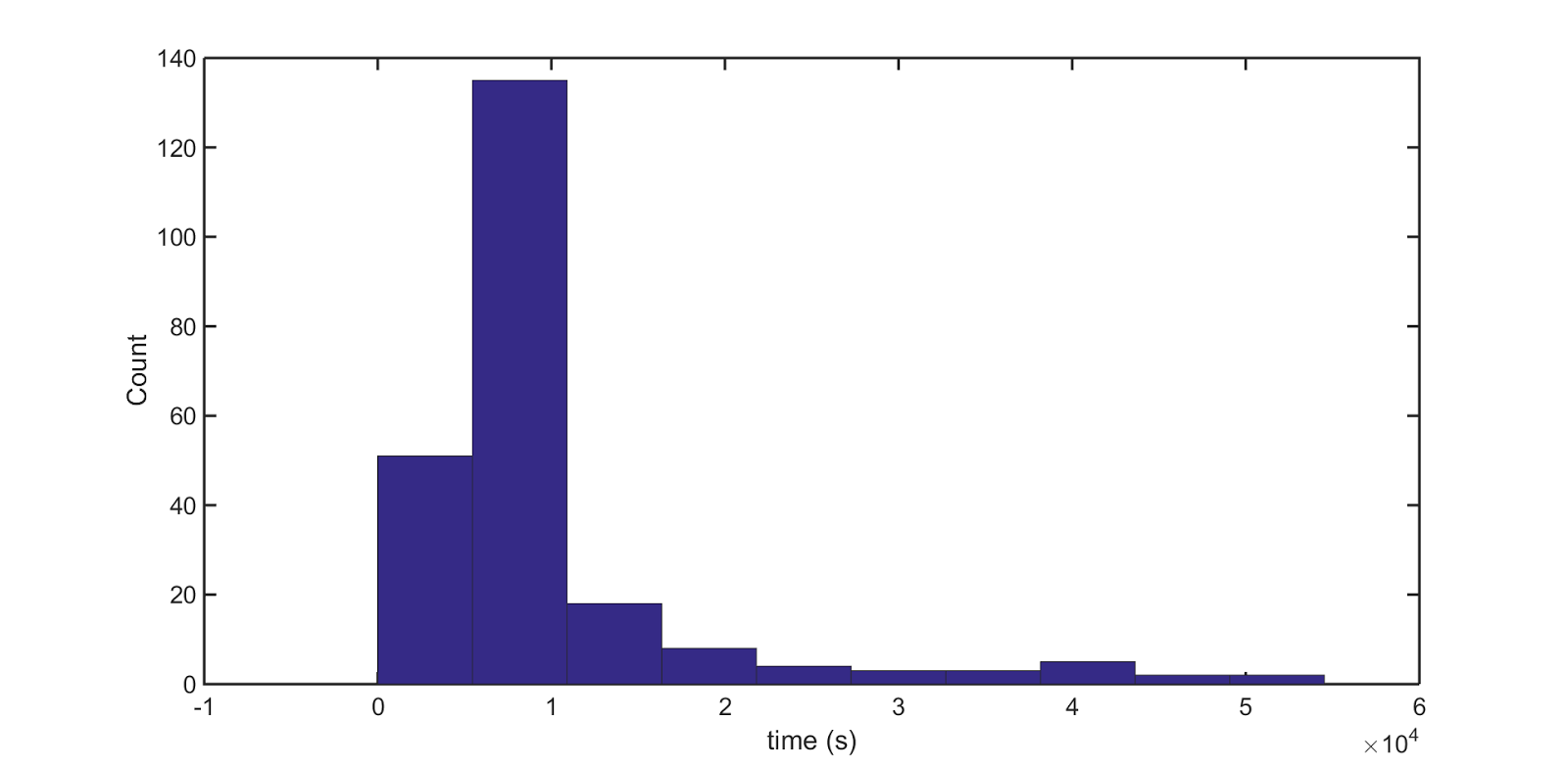


Figure 2: Variation of MASS-PRF runtime on 231 900 bp simulated genes with different numbers and dispositions of polymorphic and divergent sites. The histogram distribution of MASS-PRF runtime for a total of 231 genes with 900 bp length simulated under the neutral codon substitution model (the two ratio model M1 with ω0= 0, ω1= 1 and p0=0.5) using Fletcher and Yang’s Indelible (Fletcher and Yang 2009).

## Other notes

1. Normally, MASS-PRF accepts nucleotide sequences with letter codes of A, T, G, and C. If N/R/Y/M/K/S/W/H/B/V/D is shown in the coding sequences, MASS-PRF accepts them by two different strategies – seen as gap, or as substituted by the most frequently used nucleotide in other sequences accordingly (Table 2). One parameter for running MASS-PRF is ‘-n’, for specifying which strategy will be used for analyzing the data. The default value is 1, indicating to substitute by the most frequently used nucleotide in other sequences. Alternatively it can be set to 0, indicating to treat such codes as gaps.

Table 2 Nucleotide and the symbols.

|  |  |
| --- | --- |
| Symbol | Meaning |
| R | A or G |
| Y | C or T |
| M | A or C |
| K | G or T |
| S | C or G |
| W | A or T |
| H | A, G or T |
| B | C, G or T |
| V | A, C, or G |
| D | A, C or T |
| N | A, T, C, or G |

2. Incomplete lineage sorting (ILS) is associated with considerable difference between species divergence times for different loci. To eliminate the impact of incomplete lineage sorting (ILS) on the performance of MASS-PRF, especially for recently diverged species with large ancestral population sizes that are particularly susceptible to ILS, we do not recommend running MASS-PRF using the option to input species divergence time. Instead, we recommend running MASS-PRF in its default mode, in which silent site divergence is clustered and used to estimate genic and even intragenic site-specific divergence times. This way the silent data is used to estimate the divergence time, capturing any deeper divergence times that might be associated with ILS and incorporating the consequent higher expectation of neutral divergence into the calculation of the intensity of selection.

## Major updates from last version (v1.0)

### Gamma calculation using four Lookup Tables

We implemented four Lookup Tables generated by Mathematica 9.0 to more precisely calculate gamma. Mathematica was used to directly to calculate integrations in the lookup table from 0 to 1. Lookup tables contain a matrix of gamma and the number of sequences *n* in the polymorphism data. The tables were generated, using Mathematica, for gamma from -50 to 50 with the interval of 0.5, for *n* from 1 to 100 with the interval of 1, and for *n* from 100 to 1500 by calculation at each interval of 50 and linear interpolation between. Four Lookup tables are in the ./src/ folder, including 'LookupTable\_gx1\_n\_gamma\_integration\_50\_v9.dat', 'LookupTable\_gx1\_derivative\_n\_gamma\_50\_v9.dat', 'LookupTable\_gx2\_n\_gamma\_integration\_50\_v9.dat', and 'LookupTable\_gx2\_derivative\_n\_gamma\_50\_v9.dat'. The detailed Mathematica codes 'Integration\_gx1\_n\_gamma\_lookup\_table\_50.nb' and 'Integration\_gx2\_n\_gamma\_lookup\_table\_50.nb' to generate the four Lookup tables are in the ./src/ folder. The program won’t run successfully without the LookupTables in the running folder. For gamma=0, gx1, gx1d, gx2, and gx2d are calculated separately (formula shown below) for the series of *n*, by mathematical limitation and integration. The following formulae are generated by Mathematica codes 'Integration-gx1-gamma=0-limit.nb' and 'Integration-gx2-gamma=0-limit.nb' in the ./src/ folder. The formulae for gamma calculation are from equations for the functions of gamma and the derivative of gamma. Details can be found in our manuscript.

### Criteria for gamma values

For the calculation of gamma, if both Prd=0 and Prp=0, gamma will be reported only as NULL. If Prp=0, gamma will be reported as N-INF. If Prd=0, gamma will be reported as INF. If no gamma value was calculated from -50 to 50 meeting the criteria dx<0.4, gamma is reported as NA.

### Model averaged gamma

Model averaged gamma is calculated based on all gamma values calculated from all models (exhaustively, or stochastically sampled). Gamma values from enumerated or sampled models are averaged using their weights, rather than using the model averaged Prd and Prp to calculate a single gamma value.

### Exhaustive vs stochastic algorithms for gamma

Exhaustive vs stochastic algorithms can be optionally used (-exact 1 or 0) for calculating model averaged gamma and 95% confidence intervals of gamma, and different numbers of stochastic models can be used as an user input option. For most sequences of significant length, it is computationally inefficient and impractical to use all possible models. For the exhaustive algorithm, all models were used within bigO(*n*\**m*): *n* is the number of models in dr, and *m* is the number of models in pr. For the stochastic algorithm for gamma, a supernumerary AIC weighting for the gamma model averaging was removed, since pr and dr are sampled based on AIC weights already.

### New features

Multiple changes have been made for Input vs Output options. It is now possible to output gamma across nucleotides or amino acid sequence; to use verbose or concise outputs; to scale consensus input.

Computational efficiency in terms of computational time and memory was improved by using an efficient sort instead of a slow bubble sort, and by improving random model choice of pr and dr in the stochastic sampling calculation of gamma.

The stochastic algorithm for gamma allows the user to assign the number of models of pr and dr to be sampled for the gamma calculation (-mn ModelNumber). The smoothness of the profiles describing selection intensity may be improved by selecting a larger number of samples, at a cost of memory usage and computational time.

MASS-PRF allows three ways of parameterizing divergence time. First, silent sites can be analyzed for clustering which can arise due to mutation rate variation across sites (-ssd), which is recommended, especially with the concern of recombination; Alternatively, intragenic mutation rate can be specified as homogeneous by calculating divergence time with gene-wide counts of polymorphic and divergent silent sites, as in the original PRF framework. Lastly, users of MASS-PRF can specify species divergence time as a parameter (-t NumberOfMYA). When specifying divergence time, selection estimates are driven only by the clustering of replacement divergences relative to the clustering of replacement polymorphisms, avoiding bias caused by the effects of weak selection on silent sites.

### Non-coding MASS-PRF

MASS-PRF can be used for non-coding regions such as cis-regulatory regions for estimating selection intensity, by simply setting -o=1 (nucleotide output format). However, the consensus file format has to be used and all sites should be labeled as R (Replacement). If adjacent coding sequence is available, it can be included; if so, synonymous sites should not be labeled as R and can be used by MASS-PRF for estimating divergence time.

### Existing problems

Due to current memory limitations, the exact algorithm (-exact 1) won’t be feasible for any but the shortest sequences, and the stochastic algorithm should be used.

For genes longer than 1500 base pairs, it can require a lot of memory for models. We have chosen in these circumstances to scale down the nonvariant and variant sequences by a divisor of 3 in the consensus format, so as to achieve a result in reasonable time.

Your job might be killed due to bus errors. This is likely due to the memory issue and you can re-submit the job, and you may have it run successfully if the new job is assigned to a node with larger memory. Due to the requirement of large amount of memory, you always want to reserve all the processors on the node your job is running to maximize your memory access.

Your job might be killed and you might get the error message “blew fast model num …”, we are still debugging this and will keep you updated. Currently, one solution is that you can re-submit your job, and you will likely have it run successfully by another trial.

## References

1. Zhang Z, Townsend JP (2009) Maximum-likelihood model averaging to profile clustering of site types across discrete linear sequences. *PLoS Comput Biol* 5(6).

2. Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. *Genetics* 132(4):1161-1176.

3. Hetru C, Hoffmann D, Bulet P (1998) Antimicrobial peptides from insects. *Molecular mechanisms of immune responses in insects*:40-66.

4. Hedengren M, Borge K, Hultmark D (2000) Expression and evolution of the Drosophila Attacin/Diptericin gene family. *Biochem Biophys Res Commun* 279(2):574-581.

5. Li YJ, Satta Y, Takahata N (1999) Paleo-demography of the Drosophila melanogaster subgroup: application of the maximum likelihood method. *Genes Genet Syst* 74(4):117-127.