mkprf

Carlos D. Busmante
cdb28@cornell.edu
Dept. Biological Statistics and Computational Biology
422 Warren Hall, Cornell University
Ithaca, NY 14853

April 17, 2006

1 Preliminaries and Acknowledgements:

This is the README file for the program `mkprf`. It provides a description of the input file format, output file format, and instructions on how to run the program using the Cornell Computational Biology Serive Unit web server `http://ser-loopp.tc.cornell.edu/cbsu/mkprf.htm`. The initial version of `mkprf` was written by Carlos Bustamante in ANSI C. Adi Fledel-Alon greatly helped speed up the program, ported it to MS Windows environment, and with Jarek Pillardy, wrote the MPI code for it to run on the CBSU web server. Stanley Sawyer (Wash. U) and Rasmus Nielsen (Cornell) were original beta testers for the command line version of `mkprf`. Your running of `mkprf` is made possible through the support of the Cornell CBSU.

Citing `mkprf`

When publishing results based on the program, please cite Bustamante, C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan, and D. L. Hartl. 2002. "The cost of inbreeding in *Arabidopsis*." Nature 416: 531534. We ask that you also please acknowledge the Cornell Computational Biology Service Unit (CBSU).

2 Disclaimer and Copyright

3 Introduction

`mkprf` implements a Markov Chain Monte Carlo algorithm for the Bayesian analysis of polymorphism and divergence for neutral and selected silent and replacement coding DNA sites across a set of genes for a pair of closely related species (see Bustamante et. al, 2002). For each locus, the program will expect four observations and several locus statistics (e.g., sample size, total number of silent sites, etc.)

In `mkprf` we implement a gaussian hirerchical model where the selection coefficient ($\gamma = 2N_e s$) on non-lethal amino acid replacement mutations for each locus is drawn from a distribution with mean μ and variance σ^2. The user specifies the genes to group together through the use of *classes*. For example, if one wishes to assume a single genomic distribution determines the selection coefficients for all the loci, all genes would belong to the same class. Alternatively, one may wish to classify the loci *a priori* based on molecular function, developmental timing, expression level, etc. Likewise, the user may wish to not use a hierarchical model for the data by specifying the flag `FIXED_VARIANCE = 1`, in which case the σ^2 parameter is not update and the initial user setting `sigmasqr0` is used throughout.

The other relevant population genetic parameters are $\theta_R = 4N_e\mu_r$ (locus non-synonymous mutation rate), $\theta_S = 4N_e\mu_s$ (locu synonymous mutation Rate), and τ (# of generations since species divergence / $2N_e$). The prior distribution on θ_S and θ_R are independent Gamma distributions with parameters α_R, β_R, α_S, and β_S. The prior distribution on τ is uniform on the interval $(0, 100)$. Each locus has its own θ_R and θ_S, while τ is a shared parameter across the genome.

`mkprf` assumes that data entered in silent sites column have $\gamma = 0$. Conditional on the number of silent SNPs and silent fixed difference *across all genes*, we sample from the posterior distribution of τ; given τ we then sample from the posterior distribution of γ for each locus given replacement data and the hierarchical model structure. If one has other "neutral" data (e.g., intron, UTR, non-coding, etc.), this can be substituted for silent site data and estimates of γ for silent sites obtained.

4 Infile Format

The infile is a tab-delimitted plain ASCII file consisting of the data for individual genes, with each gene starting on a separate line. **Please note the list below fixes a typo in previous documentation**. There are 11 columns for each line:

1. class name
2. gene name
3. FS : # fixed silent silents
4. SS : # segregating silent sites
5. FR : # fixed replacement sites
6. SR : # segregating replacement sites
7. n_1 : number of sequences sampled in species 1.
8. n_2 : number of sequences sampled in species 2.
9. TS : total number of silent sites in the alignment.
10. TR : total number of replacement sites in the alignment.
11. *Ratio* : haploid ratio (1.0 for autosomal, 0.75 for X-linked, 0.25 for Y-linked).

If a line has less than 11 columns, the program will terminate with an error message. For example, Barrier et. al (2003) implemented `mkprf` to analyze two types of genes, those from an earlier publication (Nature, with 12 genes) and those from an EST screen (EST, with 7 genes). The input file for the data set is:

```
Nature  Adh1    36      13      15      7       17      1       270.75  812.25  1
Nature  AP1     11      8       8       10      15      1       188.25  564.75  1
Nature  AP3     14      9       6       19      19      1       174     522     1
Nature  CAL     14      5       14      15      17      1       141     423     1
Nature  CHI     36      2       26      2       20      1       183     549     1
Nature  ChiA    16      25      16      19      17      1       226.5   679.5   1
Nature  LFY     37      4       18      2       15      1       317.25  951.75  1
Nature  PgiC    55      16      8       15      21      1       418.5   1255.5  1
Nature  PI      18      4       9       12      16      1       156     468     1
Nature  TFL     13      1       3       3       14      1       132.75  398.25  1
Nature  FAH1    36      17      5       5       20      1       258     774     1
Nature  F3H     34      10      7       3       20      1       339.75  1019.25 1
EST EST.1       24      10      19      7       13      5       150     450     1
EST EST.2       22      6       27      6       10      5       255.75  767.25  1
EST EST.3       5       2       3       0       10      5       87.75   263.25  1
EST EST.5       18      9       23      28      11      4       239.25  717.75  1
EST EST.6       8       10      12      7       12      2       106.5   319.5   1
EST EST.7       2       0       5       5       13      5       63      189     1
```

4.1 Pooled or Split Gene Data

The theory behind the method also applies to pooled data across loci. That is, if instead of considering individual loci one was concerned with either subsets of the loci (e.g., individual protein domains or exons) or classes of mutations (e.g., radical amino mutations across globular enzymes), the "loci" entries would correspond to what we call mutational classes. As explained above, the silent mutations across loci are treated as "exchangeable", so one could enter data on mutation classes by having a line for replacement changes with 0 in the column entry for silent sites.

For example, considering a recent analysis of amino acid polymorphism in the human genome we can pool data across 301 autosomal genes sequenced in $n_1 = 90 \times 2$ human chromosomes and $n_2 = 1 \times 2$ chimp chromosomes (NIEHS data from Williamson et al., 2005). We classified the mutations based on 75 types of amino acid changes (each possible single step change), 18 classes of silent changes (one for each amino acid that is encoded by more than one codon), and classes for intron, UTR, and flanking DNA variation. Furthermore, we classified the amino acid mutation types into "moderate", "radical", and "conservative" classes based on various physicochemical measures. Below are the first three lines of the input file. The "gene" names are the mutation type names (e.g., ALA $\leftrightarrow$ ALA = AA) and the class names correspond the mutation class. Note that if we have a neutral standard (say from non-coding DNA), then we can estimate selection on the silent substitutions relative to the neutral standard.

Conservative	AA	148	76	0	0	180	1	10914	0	1
Radical	AD	0	0	1	4	180	1	0	2879	1
Radical	AE	0	0	4	2	180	1	0	3255	1
Moderate	AG	0	0	15	5	180	1	0	2906	1
Conservative	AP	0	0	17	7	180	1	0	2753	1
Moderate	AS	0	0	18	10	180	1	0	3955	1
Moderate	AT	0	0	59	42	180	1	0	12552	1
Moderate	AV	0	0	37	29	180	1	0	13398	1

Parameter	interpretation
alphas	Scale parameter for Gamma prior on θ_S
betas	Shape parameter for Gamma prior on θ_S
alphar	Scale parameter for Gamma prior on θ_R
betar	Shape parameter for Gamma prior on θ_R
mu0	mean of prior distribution on μ for each class
sigmasqr0	variance of prior distribution on μ for each class
kappa0	along with nu0 determines weight to give prior for μ
nu0	
sigma0	if `FIXED_VARIANCE = 1`, all loci will have a prior with mean mu0 and standard deviation, sigma0

Table 1: Description of parameters in the prior distribution.

Parameter	interpretation
chains	number of separate starting points for the MCMC algorithm ($\geq$2)
howmany	number of draws per chain ($\geq 10^3$)
burnin	number of initial steps of the MCMC chain to throw out before retaining draws ($\geq 10^3$)
step	number of steps between retained draws in the MCMC chain (≥ 10)
delta_t	step size in Metropolis sampling step for τ ($0.5 \leq \delta_\tau \leq 2$)
rho	ratio of N_e for species 2 to N_e for species 1
rhosd	quantifies user uncertainty in rho

Table 2: Description of MCMC parameters.

Parameter	interpretation
CUTOFF_R	minimum number of variable replacement sites required to include a gene
CUTOFF_S	minimum number of variable silent sites required to include a gene
CHECK_POLY	if 1, omit genes that have 0 marginals for SNPs counts
FIXED_VARIANCE	if 1, do not update μ and σ vectors. Equivalent to having a prior mu0 and sigma0 for each gene.
OUTPUT_MCMC	if 1, return a file with the posterior draws for $\mu_1, \mu_2, \ldots, \gamma_1, \gamma_2, \ldots, \theta_{R,1}, \theta_{R,2}, \ldots,$. Each complete iteration of the algorithm will be on a separate line.
OUTPUT_THETA	if 1, return summary of posterior distribution on θ_R and θ_S
USE_FAST_T	if 1, uses the average sample size across loci so that the program can be run in parallel
USE_GENOMIC_DS	report the dn/ds ratio by the genomic average rather than the locus average.
Force parallel run	click to run the program in parallel (WARNING: sample size must be the same across loci).

Table 3: Optional flags for large data sets and control output.

5 Parallel Processing

If the samplesize is the same across all genes, you will want to take advantage of running the program on more than one processor. To do so, check flags `USE_FAST_T`, USE_GENOMIC_POLYDIV, `USE_GENOMIC_DS`.

6 Output

There are two output files available from `mkprf`. The first is a summary of the marginal posterior distributions of the parameters. The columns in the file are as follows

1. Parameter name
   - All of the gamma parameters begin with "gamm."
   - If the flag `OUTPUT_THETA` is checked, the posterior distribution of $\log \frac{\theta_R}{\theta_S}$ is also summarized (scaled by the total number of silent and replacement sites).
2. Average rejection rate (number of proposals rejected / total number of proposals; will be 1.0 for mu and sigma since these are updated via Gibbs Sampling)
3. Posterior mean
4. Posterior s.d.
5. 2.5% posterior quantile
6. 97.5% posterior quantile
7. Gelman Rubin statistic (if the chains have converged, value should be close to 1.0)
8. Tail probability that posterior distribution of parameter is below 0.

The second output file (produced if flag `OUTPUT_MCMC` is checked) are the raw outputs from the MCMC runs. Each parameter is a column and each draw is a separate row. The first line of the file gives column (parameter) headings.