3s Version 4 Manual

Ziheng Yang, Bo Xu, and Tianqi Zhu

July 11, 2023

Disclaimer. The software package is provided "as is" without warranty of any kind. In no event shall the author or his employer be held responsible for any damage resulting from the use of this software, including but not limited to the frustration that you may experience in using the package. The program is distributed under the GNU GPL v3.

Contents

1 Introduction

The program 3s implements the maximum likelihood method of parameter estimation under the multispecies coalescent (MSC) model either with and without gene flow for three species applied to multilocus genomic sequence data. Four major models are implemented in the program: M0: MSC with no gene flow (complete isolation) (Yang, 2002; Takahata *et al.*, 1995); M1: MSC-beta with variation in species divergence time among loci to approximate gene flow around the time of speciation (Yang, 2010); M2: MSC-M (MSC with migration) also known as isolation-with-migration (IM) model (Zhu and Yang, 2012; Dalquen *et al.*, 2017); and M3: MSci (MSC with introgression) (Flouri *et al.*, 2020). Parameters in the models include the effective population sizes for extant and extinct species (θ), species divergence times (τ), and the rates of gene flow (either migration rates M_{ij} in the MSC-M model or introgression probabilities φ_{ij} in the MSci model).

The simple MSC model without gene flow (M0) is illustrated in figure 1. Given the species tree, $((S_1, S_2), S_3)$, the program can estimate the parameters in the model (θ and τ) using sequence alignments from the extant species. It can also be used to estimate the species tree in the case of three species.

The program can handle loci of arbitrary configurations of two or three sequences. For example, configuration 123 means one sequence from each of the three species, 11 means two sequences from species S_1, 112 means two sequences from species S_1 and one sequence from species S_2, and so on.

The program is written in C. The algorithm averages over gene trees at each locus analytically and calculates the integral over coalescent times in the gene tree for each locus numerically (using Gaussian quadrature). It is limited to three species, with two or three sequences sampled per locus, but can handle over 10,000 loci.

Previously Takahata *et al.* (1995) developed an ML method of parameter estimation under the MSC model with two or three species assuming the infinite-sites mutation model (Kimura, 1969). Yang (2002) extended the method to use the JC model (Jukes and Cantor, 1969). The MSC model has since been implemented in the Bayesian program MCMCcoal (Rannala and Yang, 2003; Burgess and Yang, 2008) and its updated version bpp (`https://github.com/bpp`) (Yang and Rannala, 2010; Yang, 2015; Flouri *et al.*, 2018). Note that bpp can accommodate an arbitrary number of species with an arbitrary number of sequences per locus.

Citations. If you use the simple MSC model with no gene flow, you may cite Yang (2002). If you use the gamma model of variable species divergence times, please citre Yang (2010). If you use the MSC-migration (MSC-M) model, please cite Zhu and Yang (2012); Dalquen *et al.* (2017) and Xu et al. (in prep.). If you use the MSC-introgression (MSci) model, please cite Xu et al. (in prep.).

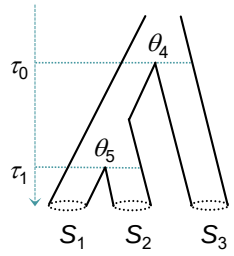


Figure 1: A species tree for three species (S_1, S_2, S_3), illustrating the parameters in the multispecies coalescent (MSC) model.

Table 1: Parameters in the models implemented in 3s

Model	p	parameters
M0 (MSC):	7	$\theta_4, \theta_5, \tau_0, \tau_1, \theta_1, \theta_2, \theta_3$
M1 (MSC-beta):	8	$\theta_4, \theta_5, \tau_0, \bar{\tau}_1, \theta_1, \theta_2, \theta_3, q$
M2 (MSC-M):	15	$\theta_4, \theta_5, \tau_0, \tau_1, \theta_1, \theta_2, \theta_3, M_{12}, M_{21}, M_{13}, M_{31}, M_{23}, M_{32}, M_{53}, M_{35}$
M3 (MSci):	10	$\theta_4, \theta_5, \tau_0, \tau_1, \theta_1, \theta_2, \theta_3, \tau_H, \varphi_{ij}, \varphi_{ji}$ for a pair of extant species (with i, j to be two of species 1, 2, and 3)

Note.— p is the number of parameters in the saturated model with the maximum allowed number of parameters for gene flow (introgression probabilities or migration rates). The fitted model may be simpler with fewer parameters. For example, population sizes for extant species (θ_1, θ_2, or θ_3) are not estimable if at most one sequence is available from the species at any locus, and some parameters may be fixed (e.g., $M_{53} = M_{35} = 0$ in the MSC-M model). For M3 (MSci), introgression may be unidirectional or bidirectional between two of species 1, 2, and 3; introgression involving the ancestral species 4 is not allowed.

2 Models and Likelihood Ratio Tests

2.1 Model M0 (MSC with no gene flow)

The MSC model with no gene flow is the simplest model implemented in 3s (Yang, 2002; Takahata *et al.*, 1995). This involves up to seven parameters: $\theta_1 = 4N_1\mu, \theta_2 = 4N_2\mu, \theta_3 = 4N_3\mu$ for the three extant species (S_1, S_2, S_3); $\theta_4 = 4N_4\mu$ for the common ancestor of S_1, S_2, S_3; $\theta_5 = 4N_5\mu$ for the common ancestor of S_1 and S_2; $\tau_0 = T_0\mu$ and $\tau_1 = T_1\mu$ for the species divergence times, where μ is the mutation rate per site per generation, and N_i are the extant and ancestral (effective) population sizes, T_0 and T_1 are the species divergence times in generations (fig. 1). Both parameters θs and τs are measured by the expected number of mutations per site. Note that $\theta_1, \theta_2, \theta_3$ are identifiable only if the dataset includes loci with at least two sequences from the same species (e.g., configurations 113 for θ_1, 223 for θ_2, and 233 for θ_3).

2.2 Model M1 (discrete beta)

The MSC-beta (M1) model in 3s allows τ_1 (fig. 1) to vary according to a beta distribution (Yang and Rannala, 2010). This is an extension of the model of Osada and Wu (2005), which uses two values

for the divergence time between two species to approximate the effects of gene flow around the time of their divergence. The M1 model has up to eight parameters: $\theta_4, \theta_5, \tau_0, \bar{\tau}_1, \theta_1, \theta_2, \theta_3, q$ (table 1), where $\bar{\tau}_1$ is the mean of the beta and q is another parameter of the beta. Here the beta distribution beta(p, q), with mean $\frac{p}{p+q}$ and variance $\frac{pq}{(p+q)^2(p+q+1)}$, is reparametrized as $\bar{\tau}_1 \equiv p/(p+q)$ and q. Thus a small q means greater variation in τ_1 while a large q means little variation (Yang and Rannala, 2010).

As under M0, $\theta_1, \theta_2, \theta_3$ are estimable only if the dataset includes loci with two or more sequences from the same species. Parameter $\bar{\tau}_1$ in M1 should be comparable with τ_1 under M0. In version 2.0, the continuous beta model is implemented, and the likelihood is calculated using the 3-D numerical integration, so that the computation is proportional to K^3 with K points used in the quadrature in each dimension. It was noted that in large datasets with > 10000 loci (say), $K = 16$ was not large enough, and $K = 32$ or even higher was necessary. Since version 2.1, the model has been modified so that a discrete beta model is used (Liu *et al.*, 2015). The beta distribution for τ_1 is broken into $B = 5$ bins, and the median in each bin is used to represent all τ_1 values in that bin. (This is the same idea as the use of a "discrete gamma" to account for variable rates among sites by Yang 1994.) The computation is then proportional to BK^2.

2.3 Model M2 (MSC-with-migration or MSC-M model)

Model M2 (MSC-M) is also known as the isolation-with-migration (IM) model. In the case of three species, it involves up to 15 parameters: $\theta_4, \theta_5, \tau_0, \tau_1, \theta_1, \theta_2, \theta_3, M_{12}, M_{21}, M_{13}, M_{31}, M_{23}, M_{32}, M_{53}, M_{35}$ (table 1). The migration rate $M_{ij} = N_j m_{ij}$ is defined as the expected number of migrants into population j from population i per generation, where m_{ij} is the proportion of immigrants in population j from population i. We define the migration rate using the real-world view with time running forward. Note that m_{ij} is the proportion of migrants in the recipient population rather than the proportion in the donor population (see Appendix A).

Zhu and Yang (2012) implemented a symmetrical version of the migration model, assuming migration between the two ingroup species (S_1, S_2) at the same rate, with $M_{12} = M_{21}$ and $\theta_1 = \theta_2$. Dalquen *et al.* (2017) extended the model to the asymmetrical case so that M_{12} and M_{21} may be different (as may θ_1 and θ_2), and to loci of arbitrary configurations (with 2 or 3 sequences per locus). Xu et al. (in prep) extended the model further to allow for migration involving species S_3 and also migration between species S_3 and S_5 (the ancestor of species 1 and 2). In the saturated model for three species, there are eight migration rates.

The Drosophila dataset D2 (noncoding) analyzed in Dalquen *et al.* (2017) is included in the package.

The migration model is specified using the `migration` keyword, using the same syntax as in BPP:

```
migration = 3
            1 2
            2 1
            5 3
```

The first line specifies the number of migration events or migration rate parameters, followed by as many lines, each of which specifies the source and target populations for each migration event. In the example here, there are three migration rates in the model: M_{12}, M_{21}, and M_{53}.

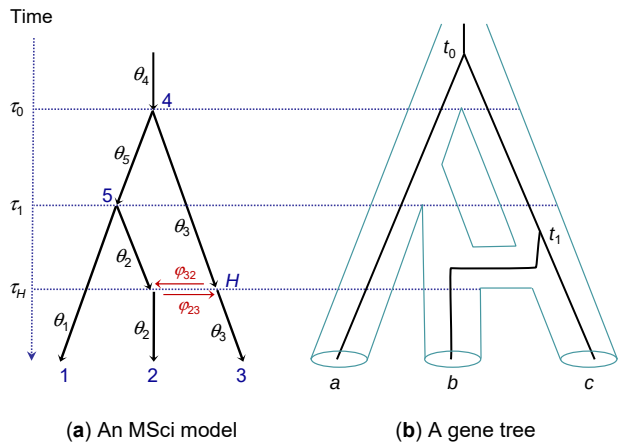


Figure 2: (**a**) An introgression (MSci) model for three species and (**b**) a gene tree for a locus of configuration 123 (with one sequence sampled from each of the three species 1, 2, 3).

One way may be to fit the 6-rates model, allowing migration for every pair of the three extant species. If this model has a very similar log likelihood value as M0 (no gene flow), there is no point of fitting simpler models with fewer migration rates. Otherwise if the six-rates model has a much larger log likelihood than model M0, you can remove migration rates that are close to 0 and fit the smaller model with fewer migration rate parameters again.

2.4 Model M3 (MSC-with-introgression or MSci model)

The introgression (MSci) model (M3) for three species implemented in 3s involves up to 10 parameters: θ_4, θ_5, τ_0, τ_1, $\theta_1, \theta_2, \theta_3$, the introgression time τ_H, and a pair of introgression probabilities φ_{ij} and φ_{ji}, where i and j can be any two of species 1, 2, and 3 (table 1). Introgression may be unidirectional or bidirectional and occurs at time $\tau_H < \tau_1$. Introgression probability φ_{ij} is defined as the proportion of introgressed individuals in population j from population i (fig. 2).

The introgression model is specified using the keyword `introgression`. For example

```
introgression = B 1 2
```

specifies the bidirectional introgression model with introgression probabilities φ_{12} and φ_{21}, while

```
introgression = U 1 3
```

specifies the unidirectional introgression model with introgression probability φ_{13}.

Our current implementation does not allow multiple introgressions involving more than one pair of extant species (such as $\varphi_{12}, \varphi_{13}$), or introgressions involving ancestral species 5 (such as $\varphi_{35}, \varphi_{53}$).

2.5 The likelihood ratio test

The likelihood ratio test (LRT) is commonly used to compare nested hypotheses. Suppose H_0 is the null hypothesis with p_0 free parameters and H_1 is alternative hypothesis with p_1 free parameters. The two hypotheses are nested, so that $p_0 < p_1$. Let ℓ_0 and ℓ_1 be the log likelihood values under the two models, calculated at the maximum likelihood estimates (MLEs). Then the LRT statistic, $2\Delta\ell = 2(\ell_1 - \ell_0)$ can be compared with $\chi^2_{p_1-p_0}$, the χ^2 distribution with $p_1 - p_0$ degrees of freedom

to decide whether the null hypothesis H_0 is rejected in favor of the alternative hypothesis H_1. For example, if the two models differ by one parameter, $p_1 - p_0 = 1$, the critical values are $\chi^2_{1,5\%} = 3.84$ at the 5% significance level and $\chi^2_{1,1\%} = 6.63$ at the 1% level.

Note that the log-likelihood values calculated by 3s are averages over the gene tree topologies and coalescent times at each locus.

The LRT can be used to compare M0 (MSC with no gene flow) and M1 (MSC-beta), as M1 reduces to M0 if $q = \infty$ under M1 and the two models are nested. However $q = \infty$ is at the boundary of the parameter space in M1 so that the regularity conditions for the LRT are not satisfied. In this case, the LRT statistic or twice the log likelihood difference, $2\Delta\ell = 2(\ell_1 - \ell_0)$, should be compared with the 50:50 mixture of 0 and χ^2_1 (Self and Liang, 1987). The critical values are 2.71 at 5% and 5.41 at 1% (as opposed to 3.84 for 5% and 6.63 for 1% for χ^2_1). To calculate the p-value, find the p-value from χ^2_1 and halve it.

Similarly the migration model (M2: MSC-M) reduces to M0 (MSC with no gene flow) if all migration rates are 0, so that the LRT can be used to compare the models to test for the presence of gene flow. Suppose there is only one migration rate (M) in the MSC-M model. The null hypothesis is then M0: MSC with no gene flow ($M = 0$), while the alternative hypothesis is $M > 0$. Since $M = 0$ is at the boundary of the parameter space, the null distribution is again the 50:50 mixture of 0 and χ^2_1, and the critical values are 2.71 at 5% and 5.41 at 1% levels.

The introgression model (M3: MSci) reduces to the model of no gene flow (M0) if the introgression probability is zero ($\varphi_{ij} = 0$, where i, j are any two of species 1, 2, and 3) or if the time of hybridization coincides with the time of species divergence ($\tau_H = \tau_1$). If the species tree is fixed, the two nested models can thus be compared using an LRT.

In Version 4, 3s will calculate the 5% significance level of the test if M2 or M3 is used, and indicate whether the rate of gene flow is significant at the 5% level.

The MSC-M and MSci models are not nested and the χ^2 distribution cannot be used. Nevertheless, one can use the log likelihood values or AIC to compare them.

Note that when the models are nested, the log likelihood value for the more-complex alternative model H_1 should not be lower than that for the null model H_0. However the opposite may occur because of the quadrature approximation or numerical problems with the optimization algorithm. You should always run the analysis under each model multiple times and choose the results corresponding to the highest log likelihood. If ℓ_1 is only slightly less than ℓ_0 (by less than 1 unit, say), you can set $\ell_1 = \ell_0$ (so that the LRT statistic is 0). If the difference is large (by 10 units, say), something may be wrong. Perhaps the iteration algorithm did not converge or the quadrature approximation did a poor job. Increase `npoints` and/or run the program again.

2.6 Species tree estimation

Parameter estimation and inference of gene flow is typically done with a fixed species tree (with the default option `speciestree = 0`). In the case of no gene flow, you can specify `speciestree = 1` so that the program will compare the three possible species trees automatically.

The option `speciestree = 1` does not work with either the introgression model or the migration model.

3 Compiling and running the program

A Windows executable is included. To compile for Mac OSX or UNIX/linux, try something like the following:

```
cc -o 3s -O3 3s.c tools.c lfun3s.c -lm
```

If you have the GNU Scientific Library (GSL) installed on your system, you can use it to speed up computation in 3s. To link to GSL, use a command similar to this:

```
cc -o 3s -O3 -DUSE_GSL 3s.c tools.c lfun3s.c -lm -lgsl -lgslcblas
```

You might have to point your compiler to the directory where GSL is installed:

```
cc -o 3s -O3 -DUSE_GSL -I/usr/local/include -L/usr/local/lib \
  3s.c tools.c lfun3s.c -lm -lgsl -lgslcblas
```

If your compiler supports OpenMP for parallel processing on multi-core processors, you can further speed up computation by enabling OpenMP support:

```
cc -o 3s -O3 -DUSE_GSL -fopenmp -I/usr/local/include -L/usr/local/lib \
 3s.c tools.c lfun3s.c -lm -lgsl -lgslcblas
```

To run the program, type the following at the command line

```
3s
3s <controlfilename>
```

The program reads in the data and runs an iteration algorithm to maximize the log-likelihood function. For M0, the program uses random numbers as starting points so you can run the same analysis at least twice to confirm that the iteration finishes at the same MLEs. For other models, you can choose either random numbers or around the MLEs under M0 by the option initialvalues. The SEs of parameters and the variance-covariance matrix are calculated by approximating the curvature of the log likelihood surface by the difference method.

4 Data format and example files

4.1 Sequence data file

The sequence data file is in PHYLIP/PAML format. Alignments for all loci are in one file, one after another. The number of loci is specified using the variable `nloci` in the control file. The alignment at each locus consists of two or three sequences. Look at the example file `ChenLi3s.txt`, which includes the 53 loci from Chen and Li (2001).

We use two approaches to assign sequences to species and specify the species tree. The first is the one used in BPP. Each sequence in the sequence alignment file is tagged with the name of the specimen (individual), and an `Imapfile` maps individuals to species. Note that in the control file the third species on the list of species must be the outgroup:

```
  species&tree = 3  A B C    * the 3rd species is the outgroup
```

In the second approach, the tags in the sequence file mean species, so that an Imap file is not needed.

Sites with alignment gaps or ambiguity characters in any of the three sequences are removed, so the variable `cleandata` in the control file has no effect.

4.2 Control file format

```
              seed = -1
            outfile = out
            seqfile = seq.txt     * sequence alignment file

           Imapfile = Imap.txt    * map of sequences to species
*          ratefile = Rate.txt    * for variable rates among loci

             nloci = 53

            usedata = 1  * 1: sequence  2: tree
            verbose = 1  * 1: more output in outfile (site patterns at loci)
      initialvalues = 1  * 0: random  1: around MLEs from M0
           nthreads = 1  * # of threads (-1: all available)
            npoints = 16 5 1   * use 8, 16 or 32
              getSE = 1
         Small_Diff = 0.5e-9

        speciestree = 0     * 0: species tree fixed  1: estimate species tree
       species&tree = 3  A B C  * the 3rd species is to be the outgroup

             models = 0 2 3     * models to use, 0, 1, 2 or 3
                                 * 0: MSC
                                 * 1: DiscreteBeta
                                 * 2: Isolation-with-Migration (MSC-M)
                                 * 3: Introgression (MSci)

          simmodel = 0   * 1: symmetric migration model of Zhu&Yang (2012)
       GIM_2species = 0   * 1: generalised IM model for 2 species (IIM, SC)

          migration = 3   * 3 migration rates, listed below
```

```
                    1 2
                    2 1
                    5 3
   introgression = B 1 3  * B: bidirectional introgression
```

`seed` (an integer) is the random number seed used to generate initial values for parameters. A negative value means random initial values.

The next few lines specify the input file names (`seqfile`, `treefile`, `Imapfile`, `ratefile`) and output file name (`out.txt`).

`ratefile` is used to specify relative mutation rates for loci. For example, the example sequence file `ChenLi3s.txt` (with control file `3s.ChenLi.ctl` or `3s.ctl`) can be used to duplicate the results of Yang (2002, table 2). To duplicate the results for "Variable rates among loci" in the table, uncomment the line

```
   ratefile = ChenLi3s.rates.txt    * for variable rates among loci
```

`npoints` is the number of points (K) in the Gaussian quadrature. Use 8, 16, or 32 (the default is 16). The computation is proportional to K^2 under M0 (MSC) and M2 (MSC-M) and to BK^2 under M1 (MSci), where $B = 5$ is the number of discrete categories in the discrete-beta model (Yang, 2010).

`models` specifies the models to be fitted to the data. For example. `models = 0 2 3` will only estimate parameters for M0, M2 and M3. M0 is compulsory, if option models is applied.

To specify the symmetric migration model of Zhu and Yang (2012), with $\theta_1 = \theta_2$ and $M_{12} = M_{21}$, use the option `simmodel = 1`. The default value is 0.

`speciestree = 1` is used to infer the species tree or model, doing the same analysis using all possible species trees automatically. The default is `speciestree = 0`, and the order of the species implies the assumed species tree. Thus

```
  speciestree = 0
  species&tree = 3  A B C
```

specifies 3 species (A, B, C). The third species is the outgroup, which means that the assumed species tree is $((A, B), C)$.

The program can also be used to analyze data from only two species $(S1, S2)$. This option is specified using

```
  species&tree = 2  S1 S2
```

In this case, θ_4 is for the ancestral species and τ_0 is the species divergence time.

`usedata = 1` (default) means that sequence alignments will be used and analyzed. `usedata = 2` means using gene trees with coalescent times (branch lengths) as data, instead of sequence alignment.

`initialvalues` controls how initial values of parameters for model M1, M2 and M3 are generated before the ML optimization. It can take two possible values: 0 and 1. `initialvalues = 0` means setting initial values to random numbers between the lower and the upper bound of the parameters. The default value is 1, which means setting initial values to random numbers around the MLE of model M0. See notes in the section "Issues with the optimization algorithm".

To specify initial values for parameters, you can include a file named `in.3s` in the working directory. This takes precedence over the option variable `initialvalues`. If a file named `in.3s` exists in the working directory, initial values will be read from the file, and the variable

`initialvalues` will have no effect. The initial-value file `in.3s` has a rigid format, and the easiest approach is to edit a template file included in the release, which includes comments (lines beginning with *, which are treated as comments and ignored). There are four lines in the file, one each for the four models: M0, M1, M2, and M3. Even if a model is not specified in the control file (on the `models` line), the corresponding line for the model exists in `in.3s`; this is read and ignored.

For each line for the specified model in `in.3s`, the parameters are in a fixed position and order, and some parameters may not exist in the specified model. You should edit the values rather than deleting some values. The value 0 for a parameter that exists in the model means that a random initial value will be used for that parameter.

`nthread` specifies the number of threads to be used with the OpenMP version of the program. The default value is 1. `nthreads = -1` means using all cores detected on the computer.

`GIM_2species = 0` specifies the generalised IM model for two species, including the isolation-with-initial migration (IIM) and the secondary contact (SC) models (Costa and Wilkinson-Herbots, 2021).

4.3 Example datasets

The file `ChenLi3s.rates.txt` includes 53 relative rates from table 1 of Yang (2002). Run the program with the command

```
3s
```

The example file `ApeC1.txt` (control file `3s.ApeC1.txt`) is for duplicating the results in Yang and Rannala (2010, table 3, chromosome 1). Run the program with

```
3s 3s.ApeC1.txt
```

5 Issues with the optimization algorithm

The 3s program uses the BFGS algorithm from PAML (Yang, 1997) to minimize the negative log likelihood function, with parameters having bounds (lower and upper limits). The likelihood function is calculated by summing over the gene tree topologies and integrating over the coalescent times on each topology numerically.

The BFGS algorithm is one of the so-called conjugate gradient algorithms, which requires first derivatives. For the MSC models implemented in 3s, the first derivatives are not available analytically, so they are calculated using the difference approximation. The first derivative is also called the gradient or slope. The slope of $y = f(x)$ at x, written as $\frac{\mathrm{d}y}{\mathrm{d}x}$, can be approximated by changing x by a small amount h and see how much the function value changes:

$$\frac{\mathrm{d}y}{\mathrm{d}x} \approx \frac{f(x+h) - f(x)}{h}. \tag{1}$$

This is called the *forward difference* method for approximating the derivative.

Similarly

$$\frac{\mathrm{d}y}{\mathrm{d}x} \approx \frac{f(x+\frac{h}{2}) - f(x-\frac{h}{2})}{h} \tag{2}$$

is called the *central difference* method.

If the step length h is too large, the approximation may be too crude, whereas if h is too small, the approximation may be because of cancellation errors. When the daaset is large the likelihood function is highly concentrated, it may be hard to choose the step length sensibly. In 3s, the step length h is affected by the variable `SmallDiff` in the control file, so you can change its value (in the range from 10^{-9} to 10^{-6}) to see whether it helps.

Note that when the function $f(x)$ reaches its maximum or minimum, the slope is 0. This should be the case when the optimum occurs inside the parameter space, but not at the boundary of the parameter space.

Figure 3 shows the screen output for a healthy run, indicated by `p` (which is a kind of slope disregarding parameters at the boundary of the search space) goes to 0 and `lnL` stabilizes.

The optimisation algorithm may fail, in particular, in large datasets with many loci. The most common reason for failures appears to be the approximate calculation of the derivatives, as the step length (h in eqs. 1 or 2) may be poorly chosen. The algorithm seems to work better under the simple model M0 (no gene flow) and fails more often under models with gene flow.

Figure 4 shows an example of algorithmic failure, indicated by `p` being far away from 0 even though the log likelihood has not stabilized.

Our practical suggestion to deal with the problem is to run the program multiple times, by using different initial values (and different `SmallDiff`. Using initial values that are close to the MLEs in the `in.3s` file appears to be the most effective approach. See also descriptions of the keyword `initialvalues`.

The runs will be exactly identical if the same `in.3s` file is used. As the file name is hard-coded, you should conduct duplicate runs in different folders. For example, you can write some scripts to generate slightly different `in.3s` files in different folders and launch multiple runs in different folders. Also you can change SmallDiff (in the range from 1e-6 to 1e-9). Then use the results corresponding to the highest log likelihood.

```
   *** Model 0 (M0) ***

Initials & bounds
theta4    theta5      tau0 tau1/tau0    theta1    theta2
0.000300  0.000300  0.000100  0.500000  0.000200  0.000200
0.00001   0.00001   0.00001   0.00001   0.00001   0.00001
0.59900   0.59900   0.59900   0.99900   0.59900   0.59900

lnL0 = -7850.245956

Iterating by ming2
Initial: fx=  7850.245956
x=  0.00030  0.00030  0.00010  0.50000  0.00020  0.00020

1 h-m-p  0.0000 0.0000 159379.6972 YCCC   7848.257984  3 0.0000    21 | 0/6
2 h-m-p  0.0000 0.0000 224593.6260 CCCCC  7843.393497  4 0.0000    44 | 0/6
3 h-m-p  0.0000 0.0000 108580.8419 YCYYYC  7832.936904  5 0.0000    66 | 0/6
4 h-m-p  0.0000 0.0000 10339.3085 CYC    7832.763348  2 0.0000    84 | 0/6
5 h-m-p  0.0000 0.0000 1886.2813 YC     7832.754271  1 0.0000   100 | 0/6
6 h-m-p  0.0002 0.0760   6.5752 ++YCC  7832.670855  2 0.0039   120 | 0/6
7 h-m-p  1.3243 7.3072   0.0193 CCC    7831.396921  2 2.0481   139 | 0/6
8 h-m-p  0.6357 4.8832   0.0621 CYCCC  7830.959311  4 1.1882   161 | 0/6
9 h-m-p  0.8073 4.5021   0.0914 YCCCC  7830.104274  4 0.8592   183 | 0/6
10 h-m-p  0.2779 1.3897   0.1965 ++     7828.645676  m 1.3897   198 | 0/6
...
28 h-m-p  0.0160 8.0000   0.0030 +CC    7827.586409  1 0.1017   518 | 0/6
29 h-m-p  0.7915 8.0000   0.0004 ----------------C  7827.586409  0 0.0000   548
    | 0/6
30 h-m-p  0.0160 8.0000   0.0001 +C     7827.585454  0 0.0598   564 | 0/6
31 h-m-p  1.6000 8.0000   0.0000 --------------C  7827.585454  0 0.0000   593

lnL  = -7827.585454  (  594 lfun calls)
MLEs
theta4    theta5      tau0      tau1    theta1    theta2
0.000416  0.000012  0.000067  0.000062  0.002510  0.003456
SEs:
0.000027  0.000002  0.000011  0.000010 -1.000000  0.007673

Time used:  2:23
```

Figure 3: A healthy run under the M0 model (no gene flow) with six parameters. When the optimization algorithm succeeds, p goes to 0 and lnL stabilizes.

```
*** Model 0 (M0) ***

Initials & bounds
    theta4    theta5      tau0 tau1/tau0    theta1    theta2
  0.003321  0.006953  0.003370  0.739209  0.005144  0.004216
  0.000010  0.000010  0.000010  0.000010  0.000010  0.000010
  0.499000  0.499000  0.499000  0.999000  0.499000  0.499000

lnL0 = -18797.993044

Iterating by ming2
Initial: fx= 18797.993044
x=  0.00332  0.00695  0.00337  0.73921  0.00514  0.00422

  1 h-m-p  0.0000 0.0000 2815391.7451 ++     9393.714339  m 0.0000    10 | 1/6
  2 h-m-p  0.0000 0.0000 594147.8378 +YYYCC  7896.204914  8 0.0000    33 | 1/6
  3 h-m-p  0.0000 0.0000 230.9390 +YYYYCCYY  7883.412668 10 0.0000    56 | 0/6
  4 h-m-p  0.0000 0.0000 1332649.9668 -..  | 0/6
  5 h-m-p  0.0000 0.0000 282271.1898  | 0/6
  6 h-m-p  0.0000 0.0000 290941.7757

lnL  = -7883.412668  (   85 lfun calls)
MLEs
    theta4    theta5      tau0      tau1    theta1    theta2
  0.000445  0.006316  0.000010  0.000007  0.003718  0.000033
SEs:
  0.000026 -1.000000  0.000010  0.000011 -1.000000  0.000049

Time used:  0:42
```

Figure 4: A failed run under the same model as in figure 3, with initial values generated at random. Here p is far from 0 and lnL has not stabilized, but the iteration has aborted.

6 History

- 3s v4, May 2023: Bo Xu extended M2 (MSC-M) to allow migration involving species S_3 and the ancestor S_5 (the common ancestor of S_1 and S_2). Also Xu Bo added the introgression model M3 (MSci), and the function of species tree estimation under model M0. A few control variables are also added including `treefile`, `speciestree`, `species&tree`, `verbose`, `initialvalues`, and `nthreads`.

- 3s v3, February 2015: Daniel Dalquen extended the program to accommodate the asymmetrical model (with θ_1, θ_2 and M_{12}, M_{21} as independent parameters) and to loci of arbitrary configurations (such as 111, 112, 133, etc.).

- 3s v2.1, July 2012: Modified the program to read sequence names that start with 123 or ABC or abc. Added a page about definitions of the migration rate in this document. Replace the model M1 (beta) of Yang (2010) to use a discretized beta distribution for τ_1, with $K_b = 5$ categories used, while for each τ_1, the 2-D integrals in the probability of data at each locus is calculated using quadrature methods, in which K (npoints) can be adjusted by the user. The old implementation of model M1 assuming the continuous beta distribution for τ_1 and using quadrature to calculate the 3-D integrals can be found in version v2.0 or v2.0a. Also I added the SIM3s model of Zhu and Yang (2012).

- 3s v2.0a, June 2011: Added the option variable `Small_Diff`. It may help to adjust this parameter if the SEs are printed out as –1.

- 3s version 2, September 2009: Changed name to 3s. Used quadrature to replace Mathematica for numerical integration. Added a model of variable τ_1 among loci.

- NE3sML version 1.1, 11 February 2005: added back variable rates among loci (by including the file `Ne3slnL.Rates.m`).

- NE3sML version 1.0: 28 June 2003.

References

Burgess, R. and Yang, Z. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol. Biol. Evol.*, 25(9): 1979–1994.

Chen, F.-C. and Li, W.-H. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.*, 68: 444–456.

Costa, R. J. and Wilkinson-Herbots, H. M. 2021. Inference of gene flow in the process of speciation: Efficient maximum-likelihood implementation of a generalised isolation-with-migration model. *Theor. Popul. Biol.*, 140: 1–15.

Dalquen, D., Zhu, T., and Yang, Z. 2017. Maximum likelihood implementation of an isolation-with-migration model for three species. *Syst. Biol.*, 66: 379–398.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. 2018. Species tree inference with bpp using genomic sequences and the multispecies coalescent. *Mol. Biol. Evol.*, 35(10): 2585–2593.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. 2020. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Mol. Biol. Evol.*, 37(4): 1211–1223.

Hobolth, A., Andersen, L., and Mailund, T. 2011. On computing the coalescence time density in an isolation-with-migration model with few samples. *Genetics*, 187: 1241–1243.

Jukes, T. and Cantor, C. 1969. Evolution of protein molecules. In H. Munro, editor, *Mammalian Protein Metabolism*, pages 21–123. Academic Press, New York.

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to the steady flux of mutations. *Genetics*, 61: 893–903.

Liu, J., Zhang, D.-X., and Yang, Z. 2015. A discrete-beta model for testing gene flow after speciation. *Methods Ecol. Evol.*, 6: 715–724.

Osada, N. and Wu, C. I. 2005. Inferring the mode of speciation from genomic data: a study of the great apes. *Genetics*, 169: 259–264.

Rannala, B. and Yang, Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4): 1645–1656.

Self, S. and Liang, K.-Y. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.*, 82: 605–610.

Takahata, N., Satta, Y., and Klein, J. 1995. Divergence time and population size in the lineage leading to modern humans. *Theor. Popul. Biol.*, 48: 198–221.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39: 306–314.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, 13: 555–556.

Yang, Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, 162(4): 1811–1823.

Yang, Z. 2010. A likelihood ratio test of speciation with gene flow using genomic sequence data. *Genom. Biol. Evol.*, 2: 200–211.

Yang, Z. 2015. The BPP program for species tree estimation and species delimitation. *Curr. Zool.*, 61: 854–865.

Yang, Z. and Rannala, B. 2010. Bayesian species delimitation using multilocus sequence data. *Proc. Natl. Acad. Sci. USA*, 107: 9264–9269.

Zhu, T. and Yang, Z. 2012. Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. *Mol. Biol. Evol.*, 29: 3131–3142.

Appendix A. Definitions of the migration rate in different programs

Different definitions of the migration rate have been used in the literature or in different computer programs, as summarized in table 2. Both IMa and GENETREE use the backward migration rate, with migration events occuring backward in time because time runs backwards in the coalescent processing of tracing the genealogical history of the sample. In those programs, an $i \rightarrow j$ migration actually means a migration from populations j to i in the real world. MIGRATE and 3s/BPP use the forward or natural migration rates, so that an $i \rightarrow j$ migration means a migration from populations i to j in the real world.

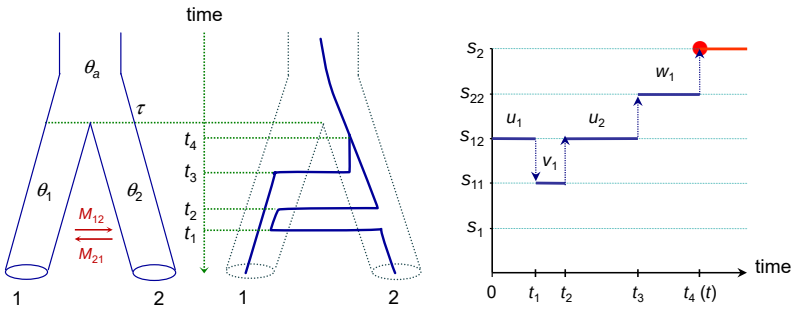
The term "backward migration" has another interpretation, that is, immigration as opposed to emigration; in this case time still runs forward. In many population genetics models, the proportion of individuals that emigrate out of the current population (that is, the forward migration rate) is typically hard to deal with, as emigrants may get drowned trying to cross the ocean or they may move to unsampled locations, etc. It is often easier to deal with the proportion of individuals in the current population that are immigrants, or the backward migration rate. We do not use this interpretation and will refer to the proportion of individuals in population j that are immigrants from population i simply as the migration rate m_{ij}, and reserve the term "backward migration" for the coalescent-world view interpretation in which time runs backwards.

Table 2: Definitions of migration rate used in several programs

Biological parameter	3s/BPP	MIGRATE3.2	IMA2	GENETREE
The proportion of individuals in population j that are immigrants from population i.	$m_{i\rightarrow j}$ or m_{ij}	$M_{j\rightarrow i}$		
The expected number of immigrant individuals in population j (from population i) per generation.	$M_{i\rightarrow j} = N_j m_{ij}$	$M_{ij}\Theta_j/4$ (Θ in MIGRATE is θ in 3s/BPP, both in expected number of mutations per site).	$\theta_j m_{j\rightarrow i}/4$	**-m** Allows the specification of a backward migration rate matrix. If there are s subpopulations this matrix has the dimensions s by s.
Parameters used in the program, expressed using parameters in 3s/BPP.		M_{ij} in MIGRATE is $4M_{ij}/\theta_j$ in 3s/BPP.	$m_{j\rightarrow i}$ in IMA2 is $4M_{ij}/\theta_j$ in 3s/BPP.	

Note.— We assume a diploid autosomal locus with $\theta = 4N\mu$, where N is the (effective) population size and μ is the mutation rate per site per generation, so that θ is the expected proportion of differences between two DNA sequences sampled at random from the population.

Appendix B. Parametrization of the MSC-M model in the case of two populations



(**a**) Species tree and a gene tree (**b**) Migration history

Figure 5: (**a**) The migration (MSC-M) model for two species and (**b**) Markov chain description of the backwards-in-time process of coalescent and migration. The starting state is s_{12} with one sequence sampled from the two species each.

This note explains the change of parameters when one changes the time scale from one generation to the expected time to accumulate one mutation per site. Suppose each locus has two sequences, one from each of populations 1 and 2. With time measured in generations, the Q matrix for the Markov chain describing the backwards-in-time process of coalescent and migration is as follows.

	s_{11}	s_{12}	s_{22}	$s_{1\|2}$
s_{11}	$-\left(2m_{21}+\frac{1}{2N_1}\right)$	$2m_{21}$	0	$\frac{1}{2N_1}$
s_{12}	m_{12}	$-(m_{12}+m_{21})$	m_{21}	0
s_{22}	0	$2m_{12}$	$-\left(2m_{12}+\frac{1}{2N_2}\right)$	$\frac{1}{2N_2}$
$s_{1\|2}$	0	0	0	0

(3)

Here q_{ij} is the rate of transition from state i to state j, with time measured in generations and running backward into the past. In population i with the sample size n_i, coalescent occurs at rate $n_i(n_i-1)/2\times\frac{1}{2N_i}$, while migration from j (to i) occurs at rate $n_i m_{ji}$.

Let the matrix of transition probabilities over time t be $P(t)=\{p_{uv}(t)\}=\mathrm{e}^{Qt}$. The coalescent time t in the number of generations has the following density

$$f(t|\Theta)=p_{s_{12},s_{11}}\times\frac{1}{2N_1}+p_{s_{12},s_{22}}\times\frac{1}{2N_2} \tag{4}$$

(Hobolth *et al.*, 2011; Zhu and Yang, 2012).

Note that under the JC model (Jukes and Cantor, 1969), the probability of observing x_i differences out of n_i sites at locus i given that the divergence time is t is

$$\mathbb{P}(x_i|t)=\left(\tfrac{3}{4}-\tfrac{3}{4}\,\mathrm{e}^{-4\mu t/3}\right)^{x_i}\left(\tfrac{1}{4}+\tfrac{3}{4}\,\mathrm{e}^{-4\mu t/3}\right)^{n_i-x_i}. \tag{5}$$

where μ is the mutation rate per site per generation. Averaging over the distribution of the coalescent time t, we have the marginal probability

$$\mathbb{P}(x_i|\Theta)=\int_0^\infty \mathbb{P}(x_i|t)f(t|\Theta)\,\mathrm{d}t. \tag{6}$$

Now define one time unit to be the expected time to accumulate one mutation per site, which means rescale Q matrix to have Q/μ, and change $1/2N_i$ into $2/\theta_i$ and m_{ij} into $4M_{ij}/\theta_j$. The new Q matrix becomes

Q matrix when one time unit is one mutation per site

	s_{11}	s_{12}	s_{22}	$s_{1\|2}$
s_{11}	$-\left(\frac{8M_{21}}{\theta_1}+\frac{2}{\theta_1}\right)$	$\frac{8M_{21}}{\theta_1}$	0	$\frac{2}{\theta_1}$
s_{12}	$\frac{4M_{12}}{\theta_2}$	$-\left(\frac{4M_{12}}{\theta_2}+\frac{4M_{21}}{\theta_1}\right)$	$\frac{4M_{21}}{\theta_1}$	0
s_{22}	0	$\frac{8M_{12}}{\theta_2}$	$-\left(\frac{8M_{12}}{\theta_2}+\frac{2}{\theta_2}\right)$	$\frac{2}{\theta_2}$
$s_{1\|2}$	0	0	0	0

(7)

Here the time unit is one mutation per site, $w_{21} = \frac{4M_{21}}{\theta_1} = \frac{m_{21}}{\mu}$ is the *mutation-scaled migration rate* into species 1 and $w_2 = \frac{4M_{12}}{\theta_2} = \frac{m_{12}}{\mu}$ is the rate into 2. Note that the Markov chain runs backwards in time while the migration rates (e.g., M_{12} and m_{12}) are defined under the real-world forward-in-time view. For example, in the first row, the transition from s_{11} to s_{12} represents migration from 2 to 1 in the real world, and either sequence in 1 can be the migrant, so that the rate is $2m_{21}$ per generation or $2m_{21}/\mu = 2w_{21}$ per mutational time unit. The transition from s_{11} to $s_{1|2}$ means that the two sequences coalesce in 1, with rate $\frac{2}{\theta_1}$. State s_{22} is not reachable from s_{11} instantaneously.

This is equivalent to applying a change of variable in the above integral $x = \mu t$, so that the density of the divergence time t of eq. 4 becomes

$$f(t|\Theta) = \begin{cases} p_{s_0 s_{11}}(t)\frac{2}{\theta_1} + p_{s_0 s_{22}}(t)\frac{2}{\theta_2}, & \text{if } t < \tau, \\ \left[1 - p_{s_0 s_{1|2}}(\tau)\right]\frac{2}{\theta_5}\,\mathrm{e}^{-\frac{2}{\theta_5}(t-\tau)}, & \text{if } t \geq \tau. \end{cases} \tag{8}$$