PROGRAM MLYA

a tool for simulation of DNA and protein data set

Program simulates two populations of two sister species in changing environment and changing possibilities of secondary gene flow between the species.

D.Sherbakov

Limnological Institute SB RAS
Ulanbatorskaya 3
664033 Irkutsk
Russia

---

e-mail: sherb@lin.irk.ru

November 1, 2009

Contents

1 INSTALLATION

1.1 Dependencies

Program depends on GNU Scientific Library sources of which can be downloaded from here.A compiled version of GSL is available as part of Cygwin on Windows, but so far I did not succeed to compile if, thuse mlya is available for Linux and MacOSX only.

Output of the program on a Linux machina may be analyzed with gnuplot. MacOSX users will find the software and installation instructions on this page, but a freeware program plot performs even better for the graphs required.

Simulated data sets currently consist of fixed number of sequences from each species. The sequences are outputted in PHYLIP non-interleaved format and may be processed with any software understanding it.

1.2 Linux

Usual procedure for lightweight programs on UNIX systems. If GSL library if not present or installed weirdly, it will fail silently.

1. make

2. become root or SUDO

   ```
   make install
   ```

   it will copy the program *mlya* into `/usr/local/bin` .

3. alternatively

   ```
   make local
   ```

   will copy it into your local `bin` directory and other users will have no access to it. This is very convenient if you do not want to get into discussions with your system administrator(s).

4. Finally, use make clean to clean up the waste created during compilation.

1.3 MacOSX

Will be added later

2 Simulation

Simulation is individual-oriented. This means that the program creates a set of virtual "organisms" with a number of properties. They interact and evolve according to a set of predefined rules. By the end of a simulation the program produces some data on demographic history, heterozygocity etc. and the objects ("organisms") are sampled for their genes just like during a real sequencing project.

2.1 Organisms

Each object has predefined set of properties depending on the particular aim of the simulation. Most of values of the properties may be chanved via control files described below. This version includes the following:

1. **Age.** When an object is "born", age=0 is assigned to it. The age increases with every successful mating, but cannot exceed "maximal age". I recommend to change this value: it impacts the result of the simulation strongly;

2. **Gender.** All organisms currently are considered to be hermaphrodite. They choose their effective gender before each mating season. Sex ratio is approximately 1:1. The only difference between genders is that the maternal organism becomes the source of "mitochondrial DNA" for it's progeny;

3. **Mitochondrial DNA**. An array of elements $x_i \in (A, C, G, T)$, $i_{max} = N_{max}$. At the moment all four nucleotides are equiprobable. I see no use in implementing a complicated model, but it is possible in the next version(s).

4. **Mutation rate of mitochondrial DNA**. This value defines the probability of mutation an a whole mitochondrial sequence. A sequence may be chosen for mutation once per reproductive cycle. As soon as it is chosen, at a randomly selected position the base will be changed. Again, all changes have the same probability, so that only JC molecular evolution model is implemented[1].

5. **Nuclear gene**. Two allels of independently mutating amino acid arrays, in this simulation they are not more than a neutral markers, so mutations do not affect viability of an organism. During mating one progeny allel is randomly chosen from one paren, the other comes from the other parent. Segregation is fully random. All amino acids are equiprobable.

6. **Mutation rate of nuclear genes**. This value is defined separately from the same thing for a mitochondrial marker. It defines the probability of

[1] Accordingly, while processing DNA further (and amino acid) sequences resulting a simulation, only simple models may be used.

mutation an a one of the two nuclear allels. An organism may be chosen for mutation once per reproductive cycle. As soon as it is chosen, ane allel is targeted at a randomly selected position. The amino acid there will be changed. All changes are equiprobable.

2.2 Generation

All organisms of the same species existing at the same time are considered to belong to the same generation. They interactions are simple: competition for resources and mating. The rules for intraspecific interactions are:

1. Competition for resources is implemented as the age independent probability of survival of an organism to the next generation depending on environmental capacity and population density. In the beginning of each round the survival probability for each organism is calculated.

2. Each organism surviving this mating period decides randomly with $p = 0.5$ which gender it has this time, after which males and females make couples. For each couple number of progeny is decided.

3. The projeny if produced and organisms of zero age **compete among themselves only** for vacant places freed by the mature organisms failed to survive this reproductive cycle. Again, it is important to remember that there is no competition between juveniles and mature organisms.

2.3 Sister species

This simulation deals with two sister species. These species may be sensitive to different environmental fluctuations or/and respond differently to the same challenges. This is required see if a hypthesis concerning evolutionary or demograpic history of a species may be distinguished from a different hypthesis with the given amount of DNA (protein) sequence data.

I did not say "alternative", did I?

In fact, there are three separate taxonomic entities in every simulation, the additional one is the outgroup.

Outgroup formally is one single organism initially exactly at the same mutational distance from the ancestors of the sister species. In cource of a simulation it's sequences keep mutating at the same rate as in the sister species. It serves as an internal standard

At the initial stage of a simulation one sequence is created for the mitochondrial markes and one for the nuclear marker. These sequences are mutated several times to create the orthological ancestors for the simulation (the outgroup sequences are mutated as well). After that the founding organisms are "cloned" in order to produce monomorphic ancestor founding groups. After that they evolve independently unless there is some gene flow between the species.

It is important to remember that in spite of origin, migrants **to** a species compete with the locals for the local resources according to the local rules.

3 USING THE PROGRAM

In the directory from which you launch your simulation 4 text files with extension .ctl must present. These files contain parameters which you specify and are ultimately needed by the program. By the end of simulation there will be several output files.

3.1 Shared parameters

Parameters shared between species are defined in file params.ctl:

```
400 #length of the mitochondrial sequence;
200 #length of "nuclear" amino acid sequence
3 #mean number of kids (Poisson distributed)
5 #max lifespan measured in reproductive cycles
.03 #nucl mutation prob
.02 #aa substitution
1 #numRep
20 #epyN
.18 #Dw width of parwise dist distribution
```

length of mitochondrial sequence: length of sequence consisting of A,G,T,C which will be transfered maternally;

length of nuclear amino acid sequence: length of sequence consisting of twenty amino acids which will be transfered to progeny so that one randomly chosen allel wil come from one of the parents, the other also randomly chosen one will come from the other parent;

mean number of kids: the number of progeny produced by each pair of animals is random. It is Poisson distributed with average number defined by this parameter. It may be a float number such as 2.4. Normally this number must be less then 4 because of computer memory restrictions;

maximal lifespan: life length are measured in reproductive cycles. At each cycle depending on population density each individual has good chance to die. This value put absolute limit to any participant of the simulation. It is non-inclusive, therefore lifespan of 1 is nonsense;

mutation probability of a nucleotide sequence: between two successive reproductive cycles each "mitochondrial" sequence has this chance to mutate. If it enters mutation, the nucleotides are randomly chosen and then changed;

mutation probability of an amino acid sequence: between two successive reproductive cycles each "nuclear", or amino acid sequence has this chance to mutate. If an organisms enters mutation state, first the allel (there are 2 of them) is chosen randomly, then the positionis chosento be changed. At the moment all possible substitutions are equiprobable;

number of replications is the number of simulations. One would need few hundreds of them in order to determine confidence limits of some parameters. But in most of cases 1 is enough to obtain simulated data sets in order to process them further.

Other parameters are not used at the moment and reserved for future.

ATTENTION! THE PARAMETERS MUST APPEAR IN THIS CONTROL FILE EXACTLY IN THE ORDER THEY ARE PRESENTED HERE! ALL OF THEM MUST BE THERE EVEN IF YOU CONSIDER THEM AS NOT NEEDED!

3.2 Fluctuations of the environment

Environmental impact on each of the sister species is described separately in two files, envy.ctl and envy1.ctl as changing carrying capacity of the environment. It determines the observed number of organisms as

$$\frac{dP}{dt} = P\left(1 - \frac{P}{K}\right) \tag{1}$$

where P is the number of organisms and K is environmental capacity. It is expressed in organisms number. This number means the maximal amount of organisms maintained by the niche. This means that before every reproductive cycle of the simulation the survival rate of each individual not reached the age limit (defined in 3.1) will have survival probability

$$s = 1 - \frac{P}{K} \tag{2}$$

This means that at a constant will fluctuate around $2/3$ of the environmental capacity[2].

File envy.ctl looks as follows:

```
5
l 400 400 1000
l 400 5000 1000
l 5000 5000 200
l 5000 500 100
l 500 500 700
```

The first line gives the number of segments into which the whole length of the simulation is split. The columns are separated by spaces or tabs. Each of the other lines of the file describe separate segments so that one line corresponds to one segment. First column must be "l" (for **l**inear). In previous versions there were some other options like "**c**onstant" and "**r**andom" [3] but at the moment

[2] This means that the youngest generation which did not pass through reproductive cycle does not compete for critical resources with the "adults". Instead it fills the vacancies left after the aged ones or the losers of the survival lotter die out. The competition for the vacancies is constrained by the same logistic rule as defined in (1).

[3] Poisson distribution, required 1 parameter

they seem to be useless. Next column is the starting environmental capacity expressed in number of organisms. The third column is the ending number of organisms. In case if the two values are not equal, the value of every reproductive cycle of a segment changes in a linear fasion. The last column containes the length of a segment expressed in the number of reproductive cycles.

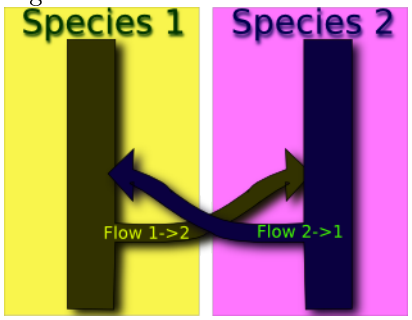
For each of the two species there is a separate file. This means that the same fluctuations of the environment may couse different responses in species: for example, warm-loving species would prosper when cold-adapted ones decline due to too high temperature.

3.3 Gene flow (migration) between species

Gene flow occurs when parents belong to different species. For each species it is determined by the probability that it's representative will seek a partner from the other species. Gene flow may be asymmetrical. Scenario of gene flow is desctibed in control file naturally called swing.ctl. It's listing is given below:

```
3
2100 0 0
200 0.10 0.01
700 0 0
```

First line gives the number of segments (the sum of lengths of all segments must be equal to the length of whole simulatin defined in files envy.ctl and envy1.ctl). Each segment is defined by three values: the first digit is the length of a segment expressed as a number of reproductive cycles, second value is the ratio of organisms of species 1 who form couples with the representatioves of species 2, second number is the same ratio for species 2.

Figure 1: 

The difference between the directions is shown on Figure 1: In case of 1->2 the members of Species 1 mate with the members of species 2 and their progeny become members of the latter. Naturally they compete for resources together with the original members species 2 both at the stage of "generation 0" (juveniles) and later. In case of the opposite flow Species 1 receives the gene supply from Species 2.

Possible assymetry of genders is not yet implemented, but there is not a big problem.

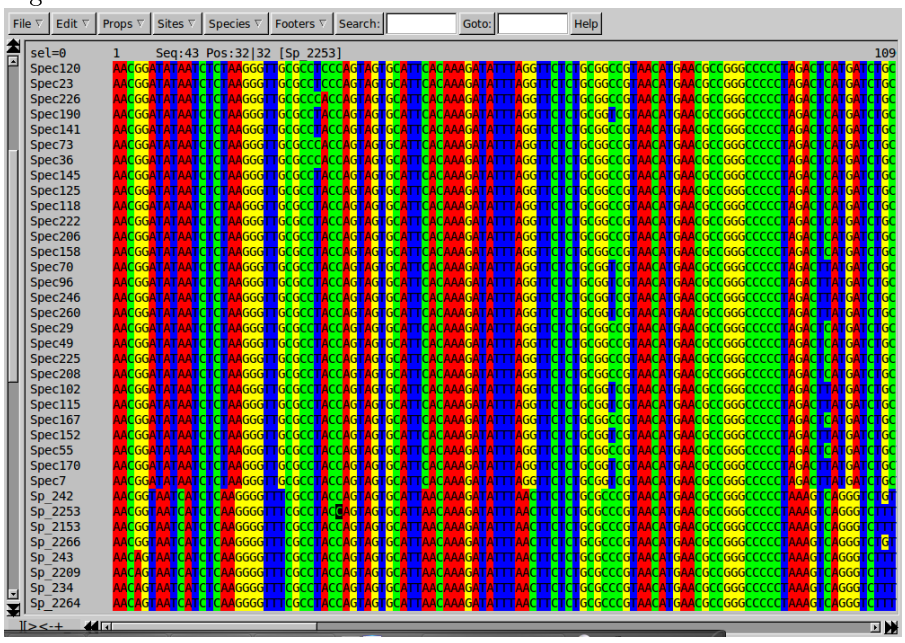
4 Output files

Program outputs several files:

1. This (Fig. 2) is a fragment of file sample.phy containing 81 sequences, which are: outgroup (on the top, not shown on this picture) and 40 sequences randomly sampled from the last generation of each species. Note the difference in OTU names: representatives of one species have names consisting of "*Spec*" concatenated with unique number. OTUs coming from the other species consist of *"Sp_2"* and unique number. This file is in an accordance with PHYLIP non-interleaved standard. Lines ends are UNIX.

2. File sampleAA.phy (Fig.3). Follows the same rules as above but containes amino acid sequences of 40 haplotypes from each species. One may see here how different the sequence of the outgroup may become sometimes.

Figure 2:



3. File demogr.dat is text TAB-delimited file containing demographic information from the run:

```
0 100 400 100 400 0.100 0.010
1 229 400 251 400 0.100 0.010
2 302 400 292 400 0.100 0.010
3 237 400 234 400 0.100 0.010
4 196 400 205 400 0.100 0.010
5 215 400 213 400 0.100 0.010
6 225 400 222 400 0.100 0.010
7 222 400 223 400 0.100 0.010
8 212 400 219 400 0.100 0.010
9 196 400 211 400 0.100 0.010
10 239 400 209 400 0.100 0.010
```

First column is the number of the simulation step. Second is the number of organisms of species 1 followed by the niche capacity at this moment, columnts 4 and 5 contain the same information for species 2. Columns 6 and 7 list migration probabilities, 1►2 and 1►2 respectively.

Figure 3: