EMOGEE and EMOGEE Tools

Michael Rosskopf (rosskopf@cs.uni-duesseldorf.de)

Dept. of Computer Science, University of Düsseldorf, Germany

June 19, 2007

1 Introduction

Different studies describe that the level of gene expression between species is positively correlated with the time that has passed since the species split from a common ancestor [Ranz and Machado, 2006]. Moreover, Khaitovich et al. [2004] found a linear relationship between divergence time and expression differences. This linearity can be explained by the neutral theory [Kimura, 1983]. Consequently, a neutral model for gene expression evolution was suggested [Khaitovich et al., 2005]. The model describes mutations in the regulatory region of a gene by a compound Poisson process. The strength of changes in the expression level is described by a continuous distribution which is called here mutation effect distribution (MED). That is, whenever a mutation occurs, the gene expression level changes according to the mutation effect distribution.

The EMOGEE (=(E)stimator for (MO)dels of (G)ene (E)xpression (E)volution) package implements the model by Khaitovich et al. [2005] and several extensions of that model described in the PhD thesis "Development and Applications of Neutral Models for Evolution of Gene Expression". In a first extension a gamma distribution is used to describe mutation effects which is more flexible than the distributions used in the original model (M-gamma model). In a second extension, non-mutational effects are taken into account (M&E model). These effects (e.g., metabolism and environmental effects) overlay mutational changes of gene expression. To describe them a new parameter is introduced which provides a better fit to real data. This makes it possible to estimate influences of mutational and non-mutational changes of the gene expression level.

According to a variant of M&E model using normal distributed mutation effects (M&E-normal model), two applications were implemented. They are located in the EMOGEE Tools package. The first application is a Bayesian method to detect genes with mutations in their regulatory regions. The second one is a non-neutrality test which can be applied to gene expression data sampled from individuals of a population. Based on this test one can detect those genes that show a significant deviation from expression levels under neutrality. The test is an adaptation of the widely used Tajima's D test [Tajima, 1989]. Before using the Bayesian method or the Tajima-type test, it is necessary to estimate the model parameters of the corresponding data with EMOGEE. The respective results have to be fed into the configuration file of EMOGEE Tools.

2 Installation

EMOGEE and EMOGEE Tools were both written in C++ on a Linux system with GNU C++ compiler. The sources are available on the homepage

`http://www.cibiv.at/software/emogee`

It should be possible to compile the sources on every computer with a C++ compiler. However, if you use a Linux system, it is very easy to install the packages. At first you have to unzip the downloaded files with `unzip emogee.zip` and `unzip emogeetools.zip`, respectively. If you like to install EMOGEE (for EMOGEE Tools analogous), go to the `emogee` directory and start the configuration script with `./configure`. After that start the makefile with `make`. Then switch to the directory `src`. You can start EMOGEE now with `./emogee`. Additionally, if you want to use EMOGEE Tools, it is necessary to copy the program `ms` by Hudson [1991] into the directory `src`. This program generates coalescent trees used by the Tajima-type test. You can find `ms` on the lab homepage of Richard Hudson:

`http://home.uchicago.edu/~rhudson1/source.html`

Table 1: The table is an example for a data set with four genes and two samples with the labels "1" and "2", respectively, which contain three individuals each

ProbeSetID	1	1	1	2	2	2
Gene1	5.2	5.4	4.8	6.1	5.9	6.2
Gene2	8.6	7.5	7.9	8.2	7.7	8.3
Gene3	7.4	6.9	7.2	5.4	5.5	6.7
Gene4	9.6	7.4	8.4	7.2	9.3	8.6

3 How to use the packages

EMOGEE and EMOGEE Tools, respectively, are batch programs[1] which can be controlled by configuration files. These files are written in plaintext and can be edited via a text editor. Thus, the program options can be set up and data sets can be assigned for analysis. After that the packages can be started from the shell. The basic program output is printed to the screen and can be written to a file with '>' on UNIX systems. More detailed output is printed into specific output files.

Data which should be applied to EMOGEE or EMOGEE Tools as well have to be arranged in a special table format. In this format each line of the table represents one gene and each column represents one individual. Thereby, each individual belongs to a sample taken from a specific group, for example, a species. Samples are characterised by a label which is a positive integer number. The first row is used for these labels which have to be sorted in an increasing order starting at "1" followed by "2" and so forth. On the other hand, the first column of a table is used for the probe set IDs of the corresponding genes. The remaining entries are used for the gene expression values which have to be on a logarithmic scale. It is necessary to separate the columns by tabulators. Additionally, the last row have to be empty. Table 1 is an example for such a data set. Standard spreadsheet programs are advisable to set up those data sets.

[1] This is not quite correct, since EMOGEE needs user input in the mode "Data generation".

Table 2: Models and estimation methods supported by EMOGEE. Please note that the models referred here as M-normal and M-extreme model were described by Khaitovich et al. [2005]. Thereby, an analytical solution is returned automatically, and the choice of the optimisation method is not taken into account.

MED	M-models (no non-mutational effects)	M&E models (with mutational effects)
Normal	M-normal model (Analytical solution)	M&E-normal model (Maximum-likelihood/χ^2-fit)
Extreme value	M-extreme model (Analytical solution)	M&E-extreme model (χ^2-fit)
Gamma	M-gamma model (Bracketing)	-

4 EMOGEE

The software package EMOGEE implements the different gene expression evolution models. The task of EMOGEE can be configured by editing the file `config.txt`. The program can be started with an argument which specifies the output file:

```
./emogee <outputfile.txt>
```

If EMOGEE is started without this argument, the output is written to the file `_output_emogee.txt`. The use of different names for the output file is reasonable when starting EMOGEE multiple times on computer clusters (c.f. section "Parallel use" below). If EMOGEE is used for parameter estimation (see below), the output is formated as follows, starting with the parameter estimates (parameters which are not part of the selected model are set to "0"):

```
<run-number> <alpha-estimate> <sigma_m/beta-estimate> <sigma_e-estimate>
<d_1-estimate> <d_2-estimate> <d_3-estimate> <d_4-estimate>
<d_3-estimate + d_4-estimate>
```

After a *-symbol, the quality of the result is described:

```
<chi^2-value/maximum-likelihood-value> <Valid/Not valid>
```

In case of `Not valid` the limits of the search space have been reached. The moments of the applied data set follow after a *-symbol:

```
<var within sample 1> <var within sample 2> <var within sample 3>
<var between 1 and 2> <var between 1 and 3> <var between 2 and 3>
<skew between 1 and 2> <skew between 1 and 3> <skew between 2 and 3>
<kurt between 1 and 2> <kurt between 1 and 3> <kurt between 2 and 3>
<2nd coefficient of Pearsons skewness of sample-1-intermediate genes>
<2nd coefficient of Pearsons skewness of sample-2-intermediate genes>
```

In the following sections all program parameters are described which can be determined by the file `config.txt`. Please note that some combinations of parameter selections are not possible, for example, a gamma distributed MED can not be combined with non-mutational effects. Table 2 gives an overview about possible models and estimation methods. EMOGEE returns an error message in cases which are not supported.

4.1 General options

Setting the program mode

```
// Mode: 1 -> Data generation
//       2 -> Estimation on real data
//       3 -> Estimation on bootstrap data
//       4 -> Estimation on generated data
mode = <number>;
```

In the mode "Data generation" a synthetic data set is generated according to the chosen model and the chosen model parameter values. When starting EMOGEE the user is asked for the number of individuals in the samples. The generated data set is written to the file `_output_emogee.txt` if not stated otherwise by an argument. In the mode "Estimation on real data" a data set is loaded from an assigned file. Accordingly, it is analysed with the chosen model. When using "Estimation on bootstrap data", a bootstrap data set is generated by sampling with replacement from the loaded data set [Efron, 1979]. After

that it is analysed like in the previous mode. The mode "Estimation on generated data" can be used to explore the stochastic process described by the selected model and its parameters. In this mode a data set is first generated like in the mode "Data generation". The size of the samples is set to one in this case. Subsequently, the chosen parameter estimation method is applied. This mode as well as "Estimation on bootstrap data" are naturally repeated a lot of times to get a stable distribution of estimates.

Assigning a parameter file

```
modelParameterFilename = <path>;
```

The parameter file specifies the model parameters used for "Data generation" and "Estimation on generated data". The parameter file must be composed as follows:

```
alpha = <value>;
sigma_m/beta = <value>;
sigma_e = <value>;
d_1 = <value>;
d_2 = <value>;
d_3 = <value>;
d_4 = <value>;
```

Depending on the model only a part of the parameters are used. The M-gamma model needs all parameters except `sigma_e`, since it does not contain non-mutational effects. The M&E models do not consider outgroup data. Thus, `d_3` and `d_4` are not required. Further, alpha is not needed, since the mutation effects are completely specified by the parameter `sigma_m/beta` for a normal distributed MED (M&E-normal model) and an extreme value distributed MED (M&E-extreme model), respectively. For the M&E-normal model the parameter `d_1` describes the expected number of mutations between both samples. Thus, also `d_2` is not required. Parameters which are not required can be set to an arbitrary value, but it is advised to set those parameters to zero for clarity. Examples for parameter files can be found in the folder `modelparameters`. These files represent the test cases used in the thesis.

Assigning a data file

```
dataTableFilename = <path>;
```

The data table file contains the data set which should be analysed with the modes "Estimation on real data" or "Estimation on bootstrap data".

Selection of samples for the analysis

```
sample1 = <number>;
sample2 = <number>;
sample3 = <number>;
```

This option assigns the two samples and the outgroup sample (if used) to the labels in the data file. Please set `sample3` to "0" if an outgroup is not applied (always except when using the M-gamma model).

Toggle between normal analysis / all pair analysis

```
// Analysis type: 1 -> Mean of all pairwise comparisons
//                       (between groups)
//                2 -> Estimation for all pairs of individuals
//                       (within and between groups)
analysisType = <number>;
```

In the first mode, the moments are estimated for all pairs of individuals between two samples. Then the mean values of the moments are calculated and used further on for estimation. If the distribution of gene expression differences is required (in case of the χ^2-method), it is calculated from the pairwise comparisons between all individuals between the samples for all genes. In the second mode, exactly one parameter estimation is performed for all pairs of individuals within and between the first and the second sample. Thereby, only the mode "Estimation on real data" is supported. Furthermore, this type of analysis is not supported for the M-gamma model which needs a third sample as an outgroup.

Determining the number of genes for data generation

```
numberOfGenes = <number>;
```

With this option the number of genes is specified which is generated in the data sets if one of the modes "Data generation" or "Estimation on generated data" is used.

Determining the number of simulations

```
numberOfSimulations = <number>;
```

If an optimisation method is used to estimate parameters, synthetic data is also generated internally within each step. With this parameter the number of gene expression differences which are simulated in one step is specified. The preset value for `numberOfSimulations` is $10,000,000$.

Determining the number of runs

```
numberOfRuns = <number>;
```

In the modes "Estimation on bootstrap data" and "Estimation on generated data" it is useful to set this parameter to a large value, for example, $1,000$ to get a stable distribution of results. Please note that a large number of runs can take a long time to compute on a single computer. However, it is possible to run EMOGEE on computer clusters (c.f. section "Parallel use" below).

Selection of the mutation effect distribution

```
// Mutation effect distribution (MED): 1 -> Normal
//                                     2 -> Extreme value
//                                     3 -> Gamma
mED = <number>;
```

With this option the used mutation effect distribution can be determined. This distribution is used for data generation as well as for estimation. In the latter case, it specifies the assumed model. Thereby, there are some restrictions, since it is not possible to combine the used MED with all other option. When using a gamma distributed MED (M-gamma

model), it is not possible to use non-mutational effects, but it is necessary to apply an outgroup sample. Against, if non-mutational effects are used (M&E model), outgroups are not supported.

Determining the stop criterion of the optimisations

```
quality = <value>;
```

The program parameter `quality` defines the stop criterion of the optimisation method. It is a number between 0 and 1. The sense of the parameter depends on the used method which can be a root find method (used for the M-gamma model), a Brents Method (χ^2-fit for the M&E-normal and the M&E-extreme model) or a Downhill Simplex method (maximum-likelihood method for the M&E-normal model). Overall, a larger quality increases the computation time. Preset values for `quality` are 0.9999 for the Downhill Simplex method and 0.999 for the remaining methods.

Choosing the precision of the results

```
outputPrecision = <number>;
```

The program parameter `outputPrecision` defines the number of digits in the output file.

4.2 Special options related to parameter optimisation of the M-gamma model

Adjusting the search space

```
standardLowerBoundAlpha = <value>;
minimalAlpha = <value>;
standardUpperBoundAlpha = <value>;
maximalAlpha = <value>;
```

If the M-gamma model is used, the parameter estimates are found by a root find method on the range of model parameter α (c.f. chapter 3). The program parameters `standardLowerBoundAlpha` and `standardUpperBoundAlpha` initialise the search space. If the search space contains no root, it is expanded exponentially up to the limits given

by `minimalAlpha` and `maximalAlpha`. Please note, that EMOGEE return an "invalid" result if the root is still not found in the expanded search space. Preset values are `standardLowerBoundAlpha = 0.25`, `standardUpperBoundAlpha = 10`, `minimalAlpha = 0.01` and `maximalAlpha = 4096`.

4.3 Special options related to the M&E model

Turn on/off non-mutational effects

```
// Modelling non-mutational effects: 1 -> No
//                                   2 -> Yes
nonMutationalEffects = <number>;
```

With this option non-mutational effects can be set on or off. Please note that it is not possible to use non-mutational effects when using a gamma distributed MED. If a normal distributed or extreme value distributed MED but no non-mutational effects are used, the parameter estimation is performed analytically (c.f. basic model by [Khaitovich et al., 2005]).

Selection of the optimisation method

```
// Method: 1 -> Chi^2-fit (M&E-normal and M&E-extreme model)
//         2 -> Maximum Likelihood (only for the M&E-normal model)
method = <number>;
```

Two optimisation methods can be selected if non-mutational effects are enabled. Please note that the maximum-likelihood method is only supported for the M&E-normal model.

Selection of the number of bins for the χ^2-fit method

```
numberOfBins = <number>;
```

The χ^2-fit method fits to a discrete version of the distribution of gene expression differences. This parameter specifies the number of bins used for the discretisation. The preset value for `numberOfBins` is 100.

Search space restriction for the χ^2-fit method

```
maximalD = <number>;
```

MaximalD determines the maximal value for the sum of `d_1` and `d_2` during the parameter estimation. Since large values for `d_1` and `d_2` decrease the computation speed rapidly, this option is available. Please note that EMOGEE returns an "invalid" warning if the optimum was not found in the search space. The preset value for `maximalD`is 512.

Number of repeats of the maximum-likelihood method

```
repeats = <number>;
```

The likelihood function is optimised by a Downhill Simplex Search (c.f. Nelder and Mead [1965]). This method might stop in local optima. To amplify the probability to reach the global optimum, the method is repeated several times. After that, the result with the highest likelihood is returned. The number of restarts of the method can be determined with the program parameter `repeats`. The preset value for `repeats` is 10.

Selection of the initial search space for the maximum-likelihood method

```
sigmaMLo = <value>;
sigmaMHi = <value>;
sigmaELo = <value>;
sigmaEHi = <value>;
dLo = <value>;
dHi = <value>;
```

The Downhill Simplex Search is initialised with random values for the parameters taken from uniform distributions. The limits of these distributions are defined for each of the three model parameters. The preset values are 0 for the lower bounds and 2 for the upper bounds.

4.4 Parallel use

The modes "Estimation on bootstrap data" and "Estimation on generated data" requires a lot of computational power, since a large number of repetitions, e.g. $1,000$, are typically

performed. Thus, it is advisable to run EMOGEE on a computer cluster and start it parallel on a large number of nodes. In this case, the program should be started by a shell script, to assign different output files as an argument. Normally, output is written to the file `_output_emogee.txt`. To avoid problems with the random number generator, a second argument can be used. This argument must be an integer number called `seedmodifier` which is added to the system time. Then the program is started as follows:

```
./emogee <output.txt> <seedmodifier>
```

If EMOGEE is started multiple times, an ascending value should be used for the seedmodifier. This technique avoids that two programs get the same seed, if they are started in the same second by the schedular of the cluster which would lead to equal results. If all runs have been accomplished, the output files can be concatenated to one file which can be used for further analysis.

5 EMOGEE Tools

The software package EMOGEE Tools implements the mutation detection method and the Tajima-type test. Before using it with real data, it is necessary to apply that data to EMOGEE to estimate the model parameters. Subsequently, the configuration file of EMOGEE Tools must be adapted according to these model parameters. Please note that EMOGEE Tools always assumes the M&E-normal model. Other models are not supported at the moment.

Selection of the tool

```
// Tool: 1 -> Mutation detection
//       2 -> Tajima-type test
tool = <number>;
```

The tool is selected by this program parameter. Please read the thesis for more information about the tools. The output of the mutation detection method is written to `_output_mutation-detection_results.txt`. Extended output which contains details about each gene is written to `_output_mutation-detection_gene-list.txt`. If synthetic gene expression data is generated, it is written to `_datatable_synthetic_microarray.txt`.

The output of the Tajima-type test is written into five different files:
The main output is written to `_output_tajima-type-test_results.txt`
The distribution of Δ-values of the analysed data is written to
`_output_tajima-type-test_Delta-dist_real-data.txt`
The distribution of Δ-values of the simulated data (neutral case) is written to
`_output_tajima-type-test_Delta-dist_neutral-case.txt`
Lists with the directional selected genes and balancing selected genes are written to `_output_tajima-type-test_directional_selected_genes.txt` and `_output_tajima-type-test_balancing_selected_genes.txt`, respectively. Generated coalescent trees used by the test are stored in `_treefile.txt`. Results of simulations on these trees are stored in `_datatable_synthetic_microarray.txt`.

Toggle between test mode and data analysis

```
// Mode: 1 -> Test mode
//       2 -> Perform a full analysis on real data
mode = <number>;
```

In the "Test mode" no data analysis is performed. Instead, simulations are accomplished in order to generate data by the model which corresponds to the neutral case. The output consists of statistics of the simulations. It depends on the used tool which is explained further.

Assigning a data file

```
dataTableFilename = <path>;
```

The data table file contains the data set which should be analysed in mode 2.

Selection of samples for the analysis

```
sample1 = <number>;
sample2 = <number>;
```

This option assigns the two samples to the labels in the data file. If the Tajima-type test is used, only the first group has a meaning.

Choosing the model parameters

```
sigma_m = <value>;
sigma_e = <value>;
d/theta = <value>;
```

If a data analysis is performed, these estimates define the neutral case which corresponds to the majority of genes. The parameters had to be estimated with EMOGEE before. In the test mode the parameters can be chosen unrestricted. Please note, that `d` is equal to θ in the Tajima-type test (the expected number of mutation events between two individuals).

Determining the number of simulations

```
numberOfSimulations = <number>;
```

When using the tool "Mutation detection", this value is used for the number of genes in the test mode (one simulation for each gene). In the Tajima-type test the value describes the number of genealogies which are generated. This parameter has a large influence on the computational time. The preset value for `numberOfSimulations` is $100,000$.

Choosing the sample size

```
sample1Size = <number>;
sample2Size = <number>;
```

With these values the number of individuals in the samples can be determined in "Test mode". For the Tajima-type test merely the first group has a meaning. When performing a full data analysis this number is pretended by the loaded data.

Determining the maximal number of mutation events (Mutation detection)

```
maxMut = <number>;
```

This parameter describes the maximal number of mutation events which is regarded in "Mutation detection". If there are genes which mutated more times than `maxMut`, their number of mutations is is estimated with `maxMut`.

Determining the number of mutations in test mode (Mutation detection)

`fixedMut = <number>;`

This program parameter is meaningful only when using the "Test mode" of "Mutation detection". In this mode no Poisson process is used to determine the number of mutations within a gene. Instead, the number of mutation events taking place is defined by this parameter.

Determining the confidence interval (Tajima-type test)

`convidenceInterval = <value>;`

This program parameter is meaningful only when using the "Tajima-type test". It is used to select the confidence interval. The value must be between 0 and 1. For example, for `convidenceInterval = 0.95` (preset value) a 95 % confidence range is selected.

References

B. Efron. Bootstrap method: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979.

R.-R. Hudson. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, 7:1–49, 1991.

P. Khaitovich, S. Pääbo, and G. Weiss. Towards a neutral evolutionary model of gene expression. *Genetics*, 170:929–939, 2005.

P. Khaitovich, G. Weiss, M. Lachmann, I. Hellmann, W. Enard, B. Muetzel, U. Wirkner, W. Ansorge, and S. Pääbo. A neutral model of transcriptome evolution. *PLoS Biology*, 2(5):682–689, 2004.

M. Kimura. *The neutral theory*. Cambridge University Press, Cambridge, UK, 1983.

J.A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7(4):308–313, 1965.

J.-M. Ranz and C.-A. Machado. Uncovering evlutionary patterns of gene expression using microarrays. *Trends in Ecology and Evolution*, 21(1):29–37, 2006.

F. Tajima. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123:585–595, 1989.