How to select best substitution model by Kakusan4/Aminosan?

Protocols for comparing molecular evolution models to get better phylogenetic tree by Treefinder and MrBayes

November 12, 2010

Akifumi S. Tanabe

Table of Contents

Chapter 1

Installing Kakusan4/Aminosan

Kakusan4 is a computer program to automate nucleotide substitution model selection. Aminosan is a software for comparing amino-acid substitution models. Both Kakusan4 and Aminosan is licensed under GNU General Public License Version 2. You can download Kakusan4 and Aminosan freely from the following URLs.
http://www.fifthdimension.jp/products/kakusan/
http://www.fifthdimension.jp/products/aminosan/

As all requirements are integrated to the distributed files, there is no other requirement on Windows and MacOS X. If you want to install source distribution, Kakusan4 and Aminosan require Perl execution environment, 2 Perl modules (Statistics::Distri-butions and Statistics::ChisqIndep), ReadSeq, PHYLIP and Treefinder. Perl is provided at the following URL.
http://www.perl.org/
You can install the Perl modules from CPAN. You can get ReadSeq from the following URL.
ftp://ftp.bio.indiana.edu/molbio/readseq/
You can download PHYLIP and Treefinder from the following URLs, respectively.
http://evolution.genetics.washington.edu/phylip.html
http://www.treefinder.de/
Kakusan4 and Aminosan also requires modified baseml and codeml, respectively. The source codes of modified baseml and codeml are contained in the source distribution of Kakusan4 and Aminosan. Note that Treefinder, PHYLIP, ReadSeq, modified baseml, and codeml are not licensed under GNU GPL2.

1.1 Windows

Executable installers for Windows are available at public web sites of Kakusan4 and Aminosan. When you run the installer, you will see the install wizard and can easily install the programs to your PCs. After the installation is finished, you can execute Kakusan4 and Aminosan from start menu or

"Send to" of context menu of files.

1.2 MacOS X

You can get zipped files for MacOS X from public web sites. After the files extracted, you can find executables. Please move the executables to "Applications" folder. I recommend registering Kakusan4 and Aminosan to Dock.

1.3 Linux

You can download zipped files of source distribution from public web sites. If downloaded zipped files put into "~/temp", you can install Kakusan4 and Aminosan to your Linux PCs as the following commands.

```
cd ~/temp
unzip kakusan4-4.0.yyyy.mm.dd.zip
cd kakusan4-4.0.yyyy.mm.dd
make -f Makefile.UNIX
sudo mkdir -p /usr/local/share/kakusan4
sudo cp kakusan4.pl /usr/local/share/kakusan4/
sudo cp baseml /usr/local/share/kakusan4/
sudo mkdir -p /usr/local/bin
sudo echo '#!/bin/sh
export PATH=/usr/local/share/kakusan4:$PATH
perl /usr/local/share/kakusan4/kakusan4.pl $*
' > /usr/local/bin/kakusan4
sudo chmod 755 /usr/local/bin/kakusan4
cd ..
unzip aminosan-1.0.yyyy.mm.dd.zip
cd aminosan-1.0.yyyy.mm.dd
make -f Makefile.UNIX
sudo mkdir -p /usr/local/share/aminosan
sudo cp aminosan.pl /usr/local/share/aminosan/
sudo cp codeml /usr/local/share/aminosan/
sudo mkdir -p /usr/local/bin
sudo echo '#!/bin/sh
export PATH=/usr/local/share/aminosan:$PATH
perl /usr/local/share/aminosan/aminosan.pl $*
' > /usr/local/bin/aminosan
sudo chmod 755 /usr/local/bin/aminosan
```

Note that you must install Statistics::Distributions and Statistics::ChisqIndep before running Kakusan4 and Aminosan. Because these modules are available at CPAN, you can install the modules as the following commands.

```
cpan -i Statistics::ChisqIndep
```

Chapter 2

Preparing sequence files for model selection and tree inference

Kakusan4 and Aminosan can treat many sequence file formats such as GenBank, FASTA, PHYLIP, and NEXUS because Kakusan4 and Aminosan use ReadSeq to convert file format. However, characters of sequence names (OTU names) and file names are restricted to alphanumeric characters and underscore. Do not use other symbols and multibyte characters in sequence names and file names. Kakusan4 and Aminosan can treat ambiguous characters such as gaps (-), missing data (?) and degenerate characters (`MRWSYKVHDBN` of DNA), but cannot treat period (`.`) as same character as the character of the same site of the first sequence. If you want to analyze multilocus sequence data and to use interactive interface, do not blend protein-coding nucleotide, rRNA/tRNA-coding nucleotide, non-coding nucleotide, and amino-acid sequences to the same file.

2.1 Protein-coding nucleotide sequences

Name of the file which contain protein-coding nucleotide sequences must terminate "`_P`" (underscore and p) to indicate that the data is coding sequence. For example, "`ND5_P.fas`", "`Cyt-b_P.nex`", and "`EF1alpha_P.gbk`". Codon positions (reading frame) of the sequences shouldn't be shifted from beginning to end.

2.2 Non-coding nucleotide, rRNA/tRNA-coding nucleotide, and amino-acid sequences

There is no restriction of name of the file which contain non-coding nucleotide, rRNA/tRNA-coding nucleotide, and amino-acid sequences, except characters. You can restrict the candidate rate matrices of amino-acid sequence data by adding "_nc" (Dayhoff / JTT / BLOSUM62 / VT / WAG / PMB / LG), "_mt" (mtREV / mtMam / mtArt / mtZoa), "_cp" (cpREV), "_rt" (rtREV / HIVb / HIVw), "_modelname" (specific model) to the tail of the input file name except extension. For example, "ND5_mt.fas", "Cyt-b_mtMam.nex", and "EF1alpha_nc.gbk".

2.3 Multilocus sequences

Kakusan4 and Aminosan can treat multilocus sequence data by 2 methods. The easier way to treat multilocus data is giving multiple sequence files to Kakusan4 and Aminosan. If Kakusan4 and Aminosan receive multiple files, the sequences in the different files are treated as different partitions. The multiple input files must have completely same set of OTU names. The another way to treat multilocus data is specifying the command line option in non-interactive mode. For example, "--partition=NonCodingLocus:1-300,ProteinCodingLocus_P:301-600".

Chapter 3

Molecular evolution models and information criteria

3.1 Information criteria

Kakusan4 and Aminosan calculate AIC (Akaike, 1974), AICc (Sugiura, 1978), and BIC (Schwarz, 1978) to compare the candidate models. The model which has less value of these criteria is better for the data. Because AICc and BIC require "sample size" and it is unclear in phylogenetic analysis, Kakusan4 and Aminosan calculate a number of AICc and BIC. Association between AICc/BIC number and sample size is as the following.

AICc1 / BIC1: the minimum number of substitutions across the tree (parsimonious tree length)
AICc2 / BIC2: the sum of the minimum number of substitutions at each site
AICc3 / BIC3: the sum of the minimum number of character states at each site
AICc4 / BIC4: the number of sites (alignment length)
AICc5 / BIC5: the number of variable sites
AICc6 / BIC6: the number of all of the characters (NSites×NOTUs)

The most used sample size is the number of sites (AICc4/BIC4).

3.2 Rate matrices of nucleotide substitution

The DNA substitution rate matrix is described as below.

From To	A	C	G	T
A	-	$r_{AC}\pi_C$	$r_{AG}\pi_G$	$r_{AT}\pi_T$
C	$r_{AC}\pi_A$	-	$r_{CG}\pi_G$	$r_{CT}\pi_T$
G	$r_{AG}\pi_A$	$r_{CG}\pi_C$	-	$r_{GT}\pi_T$
T	$r_{AT}\pi_A$	$r_{CT}\pi_C$	$r_{GT}\pi_G$	-

Kakusan4 can compare the following rate matrices.

	$\pi_A = \pi_C = \pi_G = \pi_T$	$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$
$r_{AC} = r_{AG} = r_{AT} = r_{CG} = r_{CT} = r_{GT}$	JC69 (Jukes and Cantor, 1969)	F81 (Felsenstein, 1981)
$r_{AC} = r_{AT} = r_{CG} = r_{GT} \neq r_{AG} = r_{CT}$	K80/K2P (Kimura, 1980)	HKY85 (Hasegawa *et al.*, 1985)
$r_{AC} = r_{AT} = r_{CG} = r_{GT} \neq r_{AG} \neq r_{CT}$	TN93ef	TN93 (Tamura and Nei, 1993)
$r_{AC} = r_{GT} \neq r_{AT} = r_{CG} \neq r_{AG} = r_{CT}$	K81/K3P (Kimura, 1981)	K81uf/K3Puf
$r_{AC} = r_{CG} \neq r_{AG} \neq r_{AT} = r_{GT} \neq r_{CT}$	J1ef	J1 (Jobb, 2008)
$r_{AC} = r_{AT} \neq r_{AG} \neq r_{CG} = r_{GT} \neq r_{CT}$	J2ef	J2 (Jobb, 2008)
$r_{AC} = r_{GT} \neq r_{AG} \neq r_{AT} = r_{CG} \neq r_{CT}$	TIMef	TIM (Posada, 2003)
$r_{AC} \neq r_{AG} = r_{CT} \neq r_{AT} \neq r_{CG} \neq r_{GT}$	TVMef	TVM (Posada, 2003)
$r_{AC} \neq r_{AG} \neq r_{AT} \neq r_{CG} \neq r_{CT} \neq r_{GT}$	SYM (Zharkikh, 1994)	GTR (Tavaré, 1986)

r_{XY} indicates substitution rate parameter between nucleotide X and Y. π_X indicates frequency parameter of nucleotide X.

Kakusan4 can use modified baseml, Treefinder, and PAUP* to calculate maximum likelihoods of the candidate models. If Kakusan4 use Treefinder to calculate maximum likelihoods of the candidate models, the maximum likelihoods of K81/K3P and K81uf/K3Puf models cannot be calculated because Treefinder cannot apply K81/K3P and K81uf/K3Puf models.

3.3 Rate matrices of amino-acid substitution

Aminosan can consider the following rate matrices.

	π_X are given by model	π_X are given by data
Poisson	Poisson	Poisson+F
Nuclear	Dayhoff-DCMut (Kosiol and Goldman, 2005)	Dayhoff-DCMut+F
	JTT-DCMut (Kosiol and Goldman, 2005)	JTT-DCMut+F
	BLOSUM62 (Henikoff and Henikoff, 1992)	BLOSUM62+F
	VT (Müller and Vingron, 2000)	VT+F
	WAG (Whelan and Goldman, 2001)	WAG+F
	PMB (Veerassamy *et al.*, 2003)	PMB+F
	LG (Le and Gascuel, 2008)	LG+F
Mitochondrial	mtREV24 (Adachi and Hasegawa, 1996)	mtREV24+F
	mtMam (Cao *et al.*, 1998)	mtMam+F
	mtArt (Abascal *et al.*, 2007)	mtArt+F
	mtZoa (Rota-Stabelli *et al.*, 2009)	mtZoa+F
Chloroplast	cpREV (Adachi *et al.*, 2000)	cpREV+F
Retroviral	rtREV (Dimmic *et al.*, 2002)	rtREV+F
	HIVb (Nickle *et al.*, 2007)	HIVb+F
	HIVw (Nickle *et al.*, 2007)	HIVw+F

3.4 Models for among-site rate variation

Kakusan4 can use modified baseml, Treefinder, and PAUP* to calculate maximum likelihood of the candidate models. Aminosan can use codeml and Treefinder to calculate maximum likelihood of

the candidate models. There are differences in applicable models for among-site rate variation among modified baseml, codeml, Treefinder, and PAUP*.

Modified baseml and codeml can apply discrete gamma (`Gamma`) (Yang, 1994), autocorrelated discrete gamma (`AGamma`) (Yang, 1995), codon position specific rate (`CodonPos`), gene (partition) specific rate (`Gene`), gene (partition) and codon position specific rate (`GeneCodonPos`), combinations of codon position specific rate and shared discrete gamma (`CodonPos1Gamma`), codon position specific rate and codon position specific discrete gamma (`CodonPos3Gamma`), codon position specific rate and shared autocorrelated discrete gamma (`CodonPos1AGamma`), codon position specific rate and codon position specific autocorrelated discrete gamma (`CodonPos3AGamma`), gene (partition) specific rate and shared discrete gamma (`Gene1Gamma`), gene (partition) specific rate and gene (partition) specific discrete gamma (`GeneNGamma`), gene (partition) specific rate and shared autocorrelated discrete gamma (`Gene1AGamma`), gene (partition) specific rate and gene (partition) specific autocorrelated discrete gamma (`GeneNAGamma`), gene (partition) and codon position specific rate and shared discrete gamma (`GeneCodonPos1Gamma`), gene (partition) and codon position specific rate and gene (partition) and codon position specific discrete gamma (`GeneCodonPosNGamma`), gene (partition) and codon position specific rate and shared autocorrelated discrete gamma (`GeneCodonPos1AGamma`), and gene (partition) and codon position specific rate and gene (partition) and codon position specific autocorrelated discrete gamma (`GeneCodonPosNAGamma`).

Treefinder can apply discrete gamma (`Gamma`) (Yang, 1994), proportion of invariable sites (`Invariant`) (Hasegawa *et al.*, 1985), and combination of `Gamma` and `Invariant` (`GammaInvariant`).

PAUP* can apply discrete gamma (`Gamma`) (Yang, 1994), proportion of invariable sites (`Invariant`) (Hasegawa *et al.*, 1985), combination of `Gamma` and `Invariant` (`GammaInvariant`), codon position specific rate (`CodonPos`), gene (partition) specific rate (`Gene`), and gene (partition) and codon position specific rate (`GeneCodonPos`).

3.5 Mixed model

We can apply different rate matrices and different models for among-site rate variation to different genes and/or codon positions. Such models are called as "mixed models" or "partitioned models". Mixed models contain 2 types of models. One is proportional model. The other is separate model. Proportional model assumes that branch lengths are proportional among genes and/or codon positions. Separate model assumes that each gene and/or each codon position has an independent set of branch lengths.

Kakusan4 and Aminosan can compare nonpartitioned, proportional, and separate models. Here, I describe procedure of the comparison. First, model selections on all genes and all codon positions concatenated sequences (hereafter, called "whole" partition), each gene and/or each codon position are performed. Next, maximum likelihoods of proportional and separate models are calculated on the whole partition. The best-fit rate matrices and among-site rate variation models which are selected in the first step are applied to the sequences of the appropriate partitions in these optimizations. Finally,

AIC (Akaike, 1974), AICc (Sugiura, 1978), and BIC (Schwarz, 1978) of the models are calculated and compared.

Because Kakusan4 and Aminosan use Treefinder to calculate maximum likelihood of proportional model, Treefinder-incompatible rate matrices and among-site rate variation models cannot be applied in proportional model. Treefinder-incompatible rate matrices and among-site rate variation models cannot be applied in comparison among nonpartitioned, proportional, and separate models because of likelihood compatibility. For this reason, be careful to interpret the result of this comparison.

Chapter 4

Performing model selection analysis by Kakusan4/Aminosan

Here, I describe procedure of the model selection on the whole partition, each gene and/or each codon position.

First, Kakusan4 and Aminosan test homogeneity of nucleotide or amino-acid composition among all OTUs. This is needed because compositional heterogeneity biases phylogenetic estimates (**?**). Kakusan4 and Aminosan test this assumption by a χ^2 test which is almost same as that of PAUP* (Swofford, 2003). The difference between Kakusan4/Aminosan and PAUP* is only handling of ambiguous characters. PAUP* uses ambiguous characters to calculate χ^2 statistics but Kakusan4 and Aminosan do not use them. If homogeneity is rejected, users should consider using data recoding techniques such as RY coding (Woese *et al*., 1991), and Dayhoff coding (Hrdy *et al*., 2004), or applying a nonhomogeneous model (Blanquart and Lartillot, 2006, 2008).

Next, Kakusan4 and Aminosan generate tree topology for maximum likelihood optimizations by neighbor-joining method (Saitou and Nei, 1987) using JC69 (Jukes and Cantor, 1969) (Kakusan4) or K83 (Kimura, 1983) (Aminosan) distances. Then, Kakusan4 and Aminosan calculate maximum likelihoods under the candidate substitution models and the tree topology. In this calculation, Kakusan4 and Aminosan can perform processing in parallel to accelerate huge computing for the maximum likelihoods estimations under the candidate models at multi-core CPU or multi-CPU machines.

Finally, Kakusan4 and Aminosan compare candidate models based on AIC (Akaike, 1974), AICc (Sugiura, 1978) and BIC (Schwarz, 1978).

Kakusan4 and Aminosan have 2 operation mode. One is interactive wizard mode. The other is non-interactive command line mode.

4.1 Interactive mode

Because I demonstrate all queries by Kakusan4/Aminosan in the following, but the quiries depend on circumstances, you don't encount all quiries.

If Kakusan4 and Aminosan is executed and receive no input file or "`--interactive=enable`" option is specified, Kakusan4 and Aminosan enter the interactive mode. Kakusan4 and Aminosan query input file name in the first step of interactive mode.

```
Kakusan4 4.0.2010.10.27
========================================================================

This is a script to select nucleotide substitution model for multi-
partitioned data set. Official web site of this script is
http://www.fifthdimension.jp/products/kakusan/ .
To know script details, see above URL.

Copyright (C) 2006-2010  Akifumi S. Tanabe

This program is free software; you can redistribute it and/or modify
it under the terms of the GNU General Public License as published by
the Free Software Foundation; either version 2 of the License.

This program is distributed in the hope that it will be useful,
but WITHOUT ANY WARRANTY; without even the implied warranty of
MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
GNU General Public License for more details.

You should have received a copy of the GNU General Public License along
with this program; if not, write to the Free Software Foundation, Inc.,
51 Franklin Street, Fifth Floor, Boston, MA 02110-1301 USA.

Parsing command line options...
No input files are specified.
Entering interactive mode.
Specified options are ignored.
Specify an input file name.
Note that you can use wild card.
```

On Windows (other than Vista) or MacOS X, drop a file into the window to enter full path of the file. On Windows Vista, shift+right click on a file and choose "`Copy as Path`", and then, right click on the window title bar and choose "`Paste`" in "`Edit`" menu. On Linux, type a full path of the input file.

```
Specify an input file name.
Note that you can use wild card.
"C:\Users\akifumi\Desktop\SampleData\CYTBnuc_P.fas"
```

Then, just press enter key to give full path to Kakusan4/Aminosan.

```
"C:\Users\akifumi\Desktop\SampleData\CYTBnuc_P.fas"
"C:\Users\akifumi\Desktop\SampleData\CYTBnuc_P.fas" was accepted.
```

```
Specify an input file name or just press enter to leave input file specification.
```

If you want to analyze multilocus data, repeat above operation to give multiple files. If you use wild card in the path, you can give many files to Kakusan4 or Aminosan in the same time. If multiple files are given, Kakusan4 and Aminosan will consider applying mixed models which is partitioned among loci (files). After all files are given, press enter key in blank to exit input file specification.

```
Specify an input file name or just press enter to leave input file specification.

OK. Input file specification have terminated.

Log, result and configuration files will be output to "C:\Users\akifumi\Desktop\
SampleData\CYTBnuc_P.fas.kakusan".
```

As described above, Kakusan4 and Aminosan make a folder in the folder which contain first given file and will save all output files to the folder.

```
OUTPUT OPTIONS

Which is a target analysis software? (MrBayes/Treefinder/PAUP/PHYML/RAxML)
(default: Treefinder)
```

If Treefinder is selected, the model configuration files for Treefinder will be output and applying mixed models which is partitioned among loci and codon positions are forced to consider. If MrBayes is selected, the model configuration files for MrBayes will be output and applying mixed models which is partitioned among loci and codon positions are forced to consider. If RAxML is selected, the model configuration files for RAxML will be output and both partitioning and nonpartitioning of codon positions are forced to consider. If PAUP* or PHYML is selected, the model configuration files for PAUP* or PHYML will be output and applying nonpartitioned models to the whole partition are forced to consider because PAUP* and PHYML cannot apply mixed models.

```
ANALYSIS OPTIONS

You input protein coding sequence.
Do you want to consider partitioning of codon positions? (y/n)
(default: n)
```

If this option is enabled, applying mixed models which is partitioned among codon positions to protein-coding nucleotide sequences will be considered.

```
You enabled partitioning of codon positions.
Do you want to consider nonpartitioning of codon positions? (y/n)
If you say yes, applying nonpartitioned models to all-codon position-concatenate
d sequences will be considered on each locus.
(default: n)
```

If this option is enabled, applying nonpartitioned models to all codon positions concatenated sequences will be considered on each locus.

```
You input multiple files.
Do you want to consider nonpartitioning of loci? (y/n)
If you say yes, applying nonpartitioned models to all-loci-concatenated sequence
s will be considered.
(default: n)
```

If this option is enabled, applying nonpartitioned models to the whole partition will be considered.

```
You input multiple files or protein coding sequence.
Do you want to compare nonpartitioned, proportional and separate models on all-
loci concatenated sequences? (y/n)
Note that this function needs Treefinder.
(default: y)
```

If this option is enabled, nonpartitioned, proportional and separate models will be compared on the whole partition.

```
Which do you want to use the program for likelihood calculation? (baseml/tf/paup)
(default: baseml)
```

This query have different choices in Aminosan. The default choice depends on above settings. The specified program will be used to calculate maximum likelihoods. If you want to compare nonpartitioned, proportional, and separate models, I strongly recommend Treefinder but you can specify other choices.

```
Do you want to optimize the parameters of base composition? (y/n)
(default: n)
```

If this option is enabled, the parameters of nucleotide or amino-acid composition will be optimized by maximum likelihood method. Otherwise, the parameters will be fixed to observed value from the data.

```
How many rate categories of discrete gamma rate heterogeneity do you want to con
sider? (integer)
(default: 8)
```

This value should be larger than 4. The larger value raises more accurate likelihoods but the computations are more time-consuming.

```
Do you want to consider invariant model for among-site rate variation? (y/n)
(default: n)
```

If this option is disabled, the invarinat model for among-site rate heterogeneity will not be considered.

```
Do you want to consider N-GAM model for among-site rate variation? (y/n)
Note that this model is very time-consuming.
(default: n)
```

If this option is enabled, "`NGamma`" (and "`NAGamma`") for among-site rate variations will be considered. Because this option does not influence mixed models, NGamma enabled mixed models still will be considered.

```
Do you want to consider autocorrelated discrete gamma model for among-site rate
variation? (y/n)
Note that this model is very time-consuming.
(default: n)
```

If this option is enabled, "`AGamma`" (and "`NAGamma`") for among-site rate variations will be considered.

```
Do you want to use different tree topology for parameter optimization on each lo
cus? (y/n)
(default: n)
```

If multilocus data is not given, this query does not appear. If this option is enabled, different tree topologies are used to optimize among loci (partitions).

```
If you want to give tree(s) for parameter optimization, specify an input file na
me.
Otherwise, just press enter.
```

If a tree file is given, tree topology which is contained in the file will be used to optimize likelihoods. Newick or NEXUS tree file is accepted.

```
How many processes do you want to run simultaneously? (integer)
(default: 1)
```

If the integer number more than 1 is specified, multiple processes will run simultaneously. I recommend that the number is same size as the logical number of CPU cores of your computer.

```
All configurations have been completed.
Just press enter to run!
```

If type Enter key, the model selection analysis will begin.

4.2 Non-interactive mode

Type the following in command prompt, console, or terminal to show help message.

```
"C:\Program Files\Kakusan4\kakusan4" --help (Windows)
perl /Applications/Kakusan4.app/Contents/MacOS/kakusan4.pl --help (MacOS X)
kakusan4 --help (other operating systems)
"C:\Program Files\Aminosan\aminosan" --help (Windows)
perl /Applications/Aminosan.app/Contents/MacOS/aminosan.pl --help (MacOS X)
aminosan --help (other operating systems)
```

Type the following in command prompt, console, or terminal to run model selection under default settings.

```
"C:\Program Files\Kakusan4\kakusan4" options inputfile (Windows)
perl /Applications/Kakusan4.app/Contents/MacOS/kakusan4.pl options inputfile (MacOS X)
kakusan4 options inputfile (other operating systems)
"C:\Program Files\Aminosan\aminosan" options inputfile (Windows)
perl /Applications/Aminosan.app/Contents/MacOS/aminosan.pl options inputfile (MacOS X)
aminosan options inputfile (other operating systems)
```

You can give multiple input files to Kakusan4 and Aminosan. You can also give partition settings to Kakusan4 and Aminosan. However, both multiple input files and partition settings cannot be given in the same time.

Chapter 5

Browsing analysis results

As described above, the folder that is named "`*.kakusan`" or "`*.aminosan`" will be made in the folder that contain the first given file, and all output files will be saved in this folder. A structure of output folder is shown below.

```
Outpur folder
|- Chisq
| | chisq_partition.txt (Result of chi-square test at each partition)
| +- ...
|- Results
| |- partition_criterion.txt (Result of model selection at each partition)
| |- whole_criterion_comparemix.txt
| | (Comparison result among nonpartitioned, proportional, and separate models on the whole)
| +- ...
|- MrBayes
| |- partition_criterion_xxx.nex
| | (NEXUS file which contain sequence data and model configuration commands for MrBayes)
| +- ...
|- PAUP
| |- partition_criterion.nex
| | (NEXUS file which contain sequence data and model configuration commands for PAUP*)
| +- ...
|- PHYML
| |- partition.phy (Sequence data of each partition)
| |- partition_criterion_singlesearch.bat (Batch file for single tree search)
| |- partition_criterion_shotgunsearch.bat (Batch file for shotgun tree search)
| |- partition_criterion_bootstrap.bat (Batch file for bootstrap analysis)
| |- partition_criterion_shotgunbootstrap.bat (Batch file for shotgun bootstrap analysis)
| +- ...
|- RAxML
| |- partition.phy (Sequence data of each partition)
| |- partition_criterion_xxx.partition (Model configuration for each partition)
| |- partition_criterion_xxx_singlesearch.bat (Batch file for single tree search)
| |- partition_criterion_xxx_shotgunsearch.bat (Batch file for shotgun tree search)
| |- partition_criterion_xxx_bootstrap.bat (Batch file for bootstrap analysis)
| +- ...
|- Treefinder
| |- partition_xxx.tf (Sequence data of each partition)
| |- partition_criterion_xxx.model (Model configuration for each partition)
| |- partition_criterion_xxx.rates (Model configuration of proprtional or separate)
| |- partition_criterion_comparemodels.tl
| | (TL script for comparing nonpartitioned, proportional, and separate models)
| |- partition_criterion_xxx_singlesearch.tl (TL script for single tree search)
| |- partition_criterion_xxx_shotgunsearch.tl (TL script for shotgun tree search)
| |- partition_criterion_xxx_bootstrap.tl (TL script for bootstrap analysis)
| +- ...
|- Scores
| |- partition_model.txt (Maximum log-likelihood under each model at each partition)
| +- ...
+- Logs
   +- ...
```

where "partition" is the partition name, "criterion" is the name of information criterion, "xxx" is the model name of nonpartitioned, proportional, and separate models. Note that the partition of all-loci concatenated sequences is called "whole" and nonpartitioned model which applied to protein-coding sequences is called "codonnonpartitioned". Shell scripts are generated instead of the batch files on MacOS X and Linux

5.1 Viewing results of χ^2 test of homogeneity of nucleotide or amino-acid composition

Open "chisq_partition.txt" in "Chisq" folder by text editor or viewer. Then, you can see a test result on each partition like below.

Content of file 5.1 a result of χ^2 test

```
Type of Nucleotides: 4
Number of Taxa: 46
Degree of Freedom: 135
Total Count: 42771
Chi-square Statistic: 85.8329297978372
p-value: 0.99968

                A           C           G           T    rtotal
OTU1          342         191         156         451      1140
      351.294101 189.613523 155.550256 443.542120

                       snip

ctotal     13180        7114        5836       16641     42771
```

If p-value is equal to or less than 0.05, the homogeneity of nucleotide or amino-acid composition among all OTUs is rejected. However, there are the requirements that faithfull p-value is calculated (Cochran, 1954). If the data does not meet the requirements, you can find the message at the tail of this file.

5.2 Viewing model selection results on each partition

Open "partition_criterion.txt" in "Results" folder by text editor or viewer. Then, you can see a result of model selection on each partition like below.

Content of file 5.2 a result of model selection

```
model                              criterion  weight                -LnL  nparam
SYM_GeneCodonPos1Gamma   5.237279083000e+004 0.98496 2.606139541500e+004    125
J2ef_GeneCodonPos1Gamma  5.238115467800e+004 0.01504 2.606757733900e+004    123
SYM_Gamma                5.288409574800e+004 0.00000 2.631904787400e+004    123
snip
```

The model name consist of the name of substitution rate matrix and the name of among-site rate variation model. The characters in front of underscore is the name of substitution rate matrix. The

characters behind underscore is the name of among-site rate variation model.

Note that the best model in this file and the applied model in the analysis might be different because the list of the file contains all considerable candidate models and the applicable models are different among analysis softwares. Check the contents of "partition_criterion.nex" for PAUP*, "partition_criterion_xxx.nex" for MrBayes, "partition_criterion_xxx.model" for Treefinder, "partition_criterion_*.bat" for PHYML, and both "partition_criterion_xxx_*.bat" and "partition_criterion_xxx.partition" for RAxML.

5.3 Viewing comparison results among nonpartitioned, proportional, and separate models on the whole data

Open "whole_criterion_comparemix.txt" in "Results" folder by text editor or viewer. Then, you can see a result of comparison among nonpartitioned, proportional, and separate models on the whole partition like below.

Content of file 5.3 a result of comparison among nonpartitioned, proportional, and separate models

```
1 model                                criterion                -LnL  nparam
2 Separate_CodonProportional       1.286036307191e+004 6.373181535953e+003     57
3 Proportional_CodonProportional   1.286895735412e+004 6.385478677060e+003     49
4 Separate_CodonSeparate           1.288258125450e+004 6.352290627248e+003     89
5 Proportional_CodonNonpartitioned 1.401815088065e+004 6.983075440327e+003     26
6 Separate_CodonNonpartitioned     1.402149556766e+004 6.976747783830e+003     34
7 Nonpartitioned                   1.413466486467e+004 7.049332432334e+003     18
```

The characters in front of underscore is the model name among loci. The characters behind underscore is the model name among codon positions. Because Kakusan4 and Aminosan use Treefinder to calculate maximum likelihood of proportional model, Treefinder-incompatible rate matrices and among-site rate variation models cannot be applied in proportional model. Treefinder-incompatible rate matrices and among-site rate variation models cannot be applied in comparison among nonpartitioned, proportional, and separate models because of likelihood compatibility. For this reason, be careful to interpret the result of this comparison. Because Treefinder can apply less among-site rate variation models than MrBayes, comparison among "Nonpartitioned", "CodonNonpartitioned", "Proportional_CodonNonpartitioned", "Separate_CodonNonpartitioned" models and the others must be interpreted carefully for Bayesian inference by MrBayes.

Chapter 6

Performing phylogenetic analyses with application of selected models

6.1 Maximum likelihood tree inference by Treefinder

You can use "partition_criterion_xxx_singlesearch.tl" in "Treefinder" folder. This file contains the commands of simple tree search with application of selected model for Treefinder. To use this file, specify this file in the dialogue of "Load TL Script ..." in "Kernel" menu of Treefinder graphical interface. After the processing, you can find ML tree file ("*_optimum.nwk", Newick format), ML substitution model/rates file ("*_optimum.model" and "*_optimum.rates", Treefinder format), and log file ("*_treesearch.log", TL Report format) in "Treefinder" folder. You can visualize the result by opening log file by "Open Image ..." in "File" menu of Treefinder graphical interface. Moreover, you can edit tree by "Redraw ..." in "View" menu and output to PostScript image by "Save" in "File" menu of Treefinder graphical interface.

You can also use "tf" command to execute tree search like below.

```
tf partition_criterion_xxx_singlesearch.tl
```

If you want to explore broader, you can use "partition_criterion_xxx_shotgunsearch.tl" after running simple search or "pgtfratchet" command in Phylogears software package to perform likelihood ratchet (Vos, 2003). You can get Phylogears from the following URL.
http://www.fifthdimension.jp/products/phylogears/

6.2 Bayesian tree inference by MrBayes(5D)

You can use MrBayes to infer the maximum posterior probability tree, but MrBayes cannot apply several amino-acid replacement models. MrBayes5D is the extended version of MrBayes. MrBayes5D can apply more amino-acid substitution models. If you want to analyze amino-acid sequence data, you should use MrBayes5D. MrBayes5D is available at the following URL.
http://www.fifthdimension.jp/products/mrbayes5d/

You can use "`partition_criterion_xxx.nex`" in "`MrBayes`" folder to run MCMC analysis. To input this file, run MrBayes(5D), and "`Execute`" this file like the following.

```
mrbayes5d
MrBayes > Execute partition_criterion_xxx.nex
```

After that, run "`MCMC`" command to start MCMC.

References

Abascal, F., Posada, D., and Zardoya, R., 2007, "MtArt: a new model of amino acid replacement for Arthropoda", *Molecular Biology and Evolution*, **24**, 1–5.

Adachi, J. and Hasegawa, M., 1996, "MOLPHY version 2.3: programs for molecular phylogenetics based in maximum likelihood", *Computer Science Monographs*, **28**, 1–150.

Adachi, J., Waddell, P. J., Martin, W., and Hasegawa, M., 2000, "Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA", *Journal of Molecular Evolution*, **50**, 348–358.

Akaike, H., 1974, "New look at statistical-model identification", *IEEE Transactions on Automatic Control*, **19**, 716–723.

Blanquart, S. and Lartillot, N., 2006, "A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution", *Molecular Biology and Evolution*, **23**, No. 11, 2058–2071, Nov.

——— , 2008, "A site- and time-heterogeneous model of amino acid replacement", *Molecular Biology and Evolution*, **25**, No. 5, 842–858, May.

Cao, Y., Janke, A., Waddell, P. J., Westerman, M., Takenaka, O., Murata, S., Okada, N., Pääbo, S., and Hasegawa, M., 1998, "Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders.", *Journal of Molecular Evolution*, **47**, 307–322.

Cochran, W. G., 1954, "Some methods for strengthening the common χ^2 tests", *Biometrics*, **10**, 417–451.

Dimmic, M. W., Rest, J. S., Mindell, D. P., and Goldstein, R. A., 2002, "rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny", *Journal of Molecular Evolution*, **55**, 65–73.

Felsenstein, J., 1981, "Evolutionary trees from DNA sequencies - a maximum-likelihood approach", *Journal of Molecular Evolution*, **17**, 368–376.

Hasegawa, M., Kishino, H., and Yano, T., 1985, "Dating of the human-ape splitting by a molecular phylogenetics", *Journal of Molecular Evolution*, **22**, 160–174.

Henikoff, S. and Henikoff, J. G., 1992, "Amino acid substitution matrices from protein blocks", *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 10915–10919.

Hrdy, I., Hirt, R. P., Dolezal, P., Bardonová, L., Foster, P. G., Tachezy, J., and Embley, T. M., 2004, "Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I", *Nature*, **432**, No. 7017, 618–622, Dec.

Jobb, G., 2008, "Treefinder version of April 2008", Software distributed by the author at http://www.treefinder.de/.

Jukes, T. H. and Cantor, C. R., 1969, "Evolution of protein molecules", in Munro, H. N. ed. *Mammalian protein metabolism*, New York: Academic Press, 21–132.

Kimura, M., 1980, "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences", *Journal of Molecular Evolution*, **16**, 111–120.

——— , 1981, "Estimation of evolutionary distances between homologous nucleotide sequences", *Proceedings of the National Academy of Sciences of the USA*, **78**, 454–458.

——— , 1983, *The neutral theory of molecular evolution*: Cambridge University Press.

Kosiol, C. and Goldman, N., 2005, "Different versions of the Dayhoff rate matrix", *Molecular Biology and*

Evolution, **22**, 193–199.

Le, S. Q. and Gascuel, O., 2008, "An improved general amino acid replacement matrix", *Molecular Biology and Evolution*, **25**, 1307–1320.

Müller, T. and Vingron, M., 2000, "Modeling amino acid replacement", *Journal of Computational Biology*, **7**, No. 6, 761–776.

Nickle, D. C., Heath, L., Jensen, M. A., Gilbert, P. B., Mullins, J. I., and Pond, S. L. K., 2007, "HIV-specific probabilistic models of protein evolution", *PLoS ONE*, **2**, e503.

Posada, D., 2003, "Using MODELTEST and PAUP* to Select a Model of Nucleotide Substitution", in *Current Protocols in Bioinformatics*, New York: John Wiley & Sons.

Rota-Stabelli, O., Yang, Z., and Telford, M. J., 2009, "MtZoa: a general mitochondrial amino acid substitutions model for animal evolutionary studies", *Molecular Phylogenetics and Evolution*, **52**, No. 1, 268–272, Jul.

Saitou, N. and Nei, M., 1987, "The neighbor-joining method: a new method for reconstructing phylogenetics trees", *Molecular Biology and Evolution*, **4**, 406–425.

Schwarz, G., 1978, "Estimating the dimension of a model", *Annals of Statistics*, **6**, 461–464.

Sugiura, N., 1978, "Further analysis of the data by Akaike's information criterion and the finite corrections", *Communications in Statistics: Theory and Methods*, **A7**, 13–26.

Swofford, D. L., 2003, *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*, Sunderland, Massachusetts: Sinauer Associates.

Tamura, K. and Nei, M., 1993, "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees", *Molecular Biology and Evolution*, **10**, 512–526.

Tavaré, S., 1986, "Some probabilistic and statistical problems in the analysis of DNA sequences", *Lectures on Mathematics in the Life Sciences*, **17**, 57–86.

Veerassamy, S., Smith, A., and Tillier, E. R. M., 2003, "A transition probability model for amino acid substitutions from blocks.", *Journal of Computational Biology*, **10**, 997–1010.

Vos, R. A., 2003, "Accelerated likelihood surface exploration: the likelihood ratchet", *Systematic Biology*, **52**, 368–373.

Whelan, S. and Goldman, N., 2001, "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach", *Molecular Biology and Evolution*, **18**, 691–699.

Woese, C. R., Achenbach, L., Rouviere, P., and Mandelco, L., 1991, "Archaeal phylogeny: reexamination of the phylogenetic position of Archaeoglobus fulgidus in light of certain composition-induced artifacts", *Systematic and Applied Microbiology*, **14**, No. 4, 364–371.

Yang, Z., 1994, "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods", *Journal of Molecular Evolution*, **39**, 306–314.

——— , 1995, "A space-time process model for the evolution of DNA sequences", *Genetics*, **139**, 993–1005.

Zharkikh, A., 1994, "Estimation of evolutionary distances between nucleotide sequences", *Journal of Molecular Evolution*, **39**, 315–329.