Genes, Genomes and Genomics ©2009 Global Science Books

Phylip and Phylogenetics

Ahmed Mansour*

Genetics Department, Faculty of Agriculture, Zagazig University, Zagazig, Egypt

Correspondence: * **alzohairy@yahoo.com**; **amansour@zu.edu.eg**

ABSTRACT

Phylogenetics studies are mainly concerned with evolutionary relatedness among various groups of organisms. Recently, phylogenetic analyses have been performed on a genomic scale to address issues ranging from the prediction of gene and protein function to organismal relationships. Computing the relatedness of organisms either by phylogenetic (gene by gene analyses) or phylogenomic (the whole genome comparison) methods reveals high-quality results for demonstrating phylogenies. In this regard, Phylip (Phylogeny Inference Package) software is a free package of programs for inferring phylogenies of living species and organisms. It is now one of the most widely used packages for computing accurate phylogenetic trees and carrying out certain related tasks. This paper provides an overview on Phylip package and its applications and contribution to phylogenetic analyses.

Keywords: bioinformatics, evolutionary relatedness, genetic diversity

INTRODUCTION

The word phylogenetics is derived from the Greek words, *phylon*, which means tribe or race, and *genetikos*, which means birth. Phylogenetic analysis could be applied in classification of living species and organisms, genetic testing and forensics and inferring functions of new genes. All these applications allow the researcher to collect data and to conduct multiple sequence alignments (MSA) and subsequently build phylogenetic tree followed by evaluation and visualization of the produced tree. In this regards, building trees are usually based on three major types of methods, Distance matrix, Maximum parsimony, and Maximum likelihood. However Phylip package, does not include any program based on Bayesian inference method, it can build trees for many kinds of data with those three different methods mentioned above (Felsenstein 1981). In addition, one of the nice things about Phylip package is that it can check the reliability of the tree by bootstrapping (Felsenstein 1992). Data types that can be handled by Phylip modules include DNA molecular sequences (e.g. gene frequencies, restriction sites), protein sequences (Felsenstein 1996), quantitative data (Felsenstein 2005), distance matrices or even 0/1 (binary) discrete characters (Felsenstein 2008). This short review is aiming to briefly introduce the molecular biologist to advantages and disadvantages of this package and its application in biology.

Phylip: History

Since October of 1980, Joseph Felsenstein, University of Washington, has created Phylip as a free package of programs for inferring phylogenies. Phylip was first released in that month, and has been substantially improved in subsequent releases. Subsequent versions have been enhanced dramatically by adding more programs and methods for trees drawing and accepting different data types (Felsenstein 1996). Phylip is currently one of the most widely distributed phylogenetic analysis software packages since 1980 (Guo et *al*. 2008; He *et al*. 2008; Belyaeva *et al*. 2009; Völgyi *et al*. 2009) and integrated into other biological knowledge base such as BioBIKE (Elhai *et al*. 2009).

Phylip: Different useful programs

The PHYLIP programs could be classified into five categories (**Table 1**):

1- Programs for molecular sequence data (such as DNAPARS, PROTPARS, etc.);
2- Programs for distance matrix data (such as FITCH, KITSCH and NEIGHBOR);
3- Programs for gene frequencies and continuous characters (such as CONTML, GENDIST and CONTRAST);
4- Programs for 0-1 discrete state data (such as MIX, MOVE, CLIQUE, etc.);
5- Programs for plotting trees and consensus trees (such as DRAWTREE, CONSENSE, etc.).

Phylip: Advantages and disadvantages

Advantages:

- Phylip is very easy to use and free of charge.
- Phylip is stand-alone software that can run on any computer.
- The recent version contains many programs for different kinds of data.
- Some sites make running of Phylip programs available as a web server which can also run on (almost) any computer that has access to the Internet and can return the results online or by e-mail.
- It works on as many computer systems such as Windows, Macintosh MacOS X, Macintosh Mac OS and Linux.
- Phylip contains different programs to analyze DNA and amino acid sequence data.

Disadvantages:

- The format requirements for Phylip are rather stringent, and any deviation will result in an error message "Unable to allocate memory" and then the program terminates.
- Phylip is used in a sequential way; the output from the first program is used as an input in the next program.
- The input or output files in the program folder have

Received: 7 January, 2009. Accepted: 15 September, 2009.

Research Note

Table 1 List of the programs and the documentation files classified based on the used method (also see **Fig. 1**).

Method	**Name**	**Function**
A- Molecular sequence methods	Protpars	Protein parsimony documentation file
	dnapars	DNA sequence parsimony documentation file
	dnapenny	DNA parsimony branch and bound documentation file
	dnamove	interactive DNA parsimony documentation file
	dnacomp	DNA compatibility documentation file
	dnaml	DNA maximum likelihood documentation file
	dnamlk	DNA maximum likelihood with clock documentation file
	proml	Protein sequence maximum likelihood documentation file
	promlk	Protein sequence maximum likelihood with clock documentation file
	dnainvar	DNA invariants documentation file
	dnadist	DNA distance documentation file
	protdist	Protein sequence distance documentation file
	restdist	Restriction sites and fragments distances documentation file
	restml	Restriction sites maximum likelihood documentation file
	seqboot	Bootstrapping/Jackknifing documentation file
B- Distance matrix methods	fitch	Fitch-Margoliash distance matrix method documentation file
	kitsch	Fitch-Margoliash distance matrix with clock documentation file
	neighbor	Neighbor-Joining and UPGMA method documentation file
C- Gene frequencies and continuous characters	contml	Maximum likelihood continuous characters and gene frequencies documentation file
	contrast	Contrast method documentation file
	gendist	Genetic distance documentation file
D- Discrete characters methods	pars	Unordered multistate parsimony documentation file
	mix	Mixed method parsimony documentation file
	penny	Branch and bound mixed method parsimony documentation file
	move	Interactive mixed method parsimony documentation file
	dollop	Dollo and polymorphism parsimony documentation file
	dolpenny	Dollo and polymorphism branch and bound parsimony documentation file
	dolmove	Dollo and polymorphism interactive parsimony documentation file
	clique	0/1 characters compatibility method documentation file
	factor	Character recoding program documentation file
E- Tree drawing, consensus, tree editing, tree distances	drawgram	Rooted tree drawing program documentation file
	drawtree	Unrooted tree drawing program documentation file
	consense	Consensus tree program documentation file
	treedist	Tree distance program documentation file
	retree	interactive tree rearrangement program documentation file

always to be named as "infile" ,"outfile" or "outtree" respectively which may replace exiting files. Thus the user needs to be sure to rename the files he/she wants to save Phylip package does not have any program based on Bayesian inference method.

Phylip: Software design

Phylip contains more than 35 programs. The source code is written in C and precompiled executables are available for Windows (95/98/NT/2000/me/XP/Windows Vista), Mac OS 8 and 9, Mac OS X, and Linux systems The programs are controlled through a menu, which contains many options which can be changed, and then allows the user to start. The input data should be in flat ASCII or Text Only format. Some sequence alignment programs, like ClustalX and T-Coffee, can write data files in the Phylip format. Currently Phylip has reached version 3.68. Executables were made available also by others for Red Hat Linux RPMs, Debian linux executables and executables for FeeBSD. Trees written onto outtree, the output file of tree drawing software, are in the Newick format which can be used by many other programs.

Limitations:

- Commands that the user should type are written with 12 pt Courier font. In addition, file names are written with 12 pt Courier New font and Output from the programs is should represented with 10 pt Courier font (Tuimala 2006).
- The Input files it should be in flat ASCII or Text Only format.
- The font files need to be in the same folder as the Drawtree or Drawgram program(s).

Phylip availability online

http://evolution.gs.washington.edu/Phylip©.html (main Phylip web page)
http://portal.litbio.org/Registered/Help/phylip/phylip.html
http://bioweb2.pasteur.fr/phylogeny/intro-en.html.

Phylip: Methods and modules

Phylip programs can be divided into these categories of methods (**Fig. 1**):

- *Distance methods:* These programs make and use matrices of distances and are intended to be used sequentially, the output of one program is the input for the other, such as Dnadist, Protdist, Fitch, Kitsch, and Neighbor.
- *Character based methods*: These programs read in a sequence alignment, and produce either one or multiple trees in the output files, outfile and outtree, such as Dnapars, Dnapenny, Dnaml, Dnamlk, Protpars, Proml.
- *Resampling tool:* This tool reads in a sequence alignment, and generates a specified number of random samples into a file outfile, such as (Seqboot).
- *Tree drawing:* These programs draw a tree from the specifications in the Newick-format such as (Drawgram, Drawtree, Retree).
- *Consensus trees:* This program constructs a consensus tree from multiple trees such as (Consense)
- *Tree distances:* This program computes, *e.g.*, a topology-based distance between two or more trees such as (Treedist).

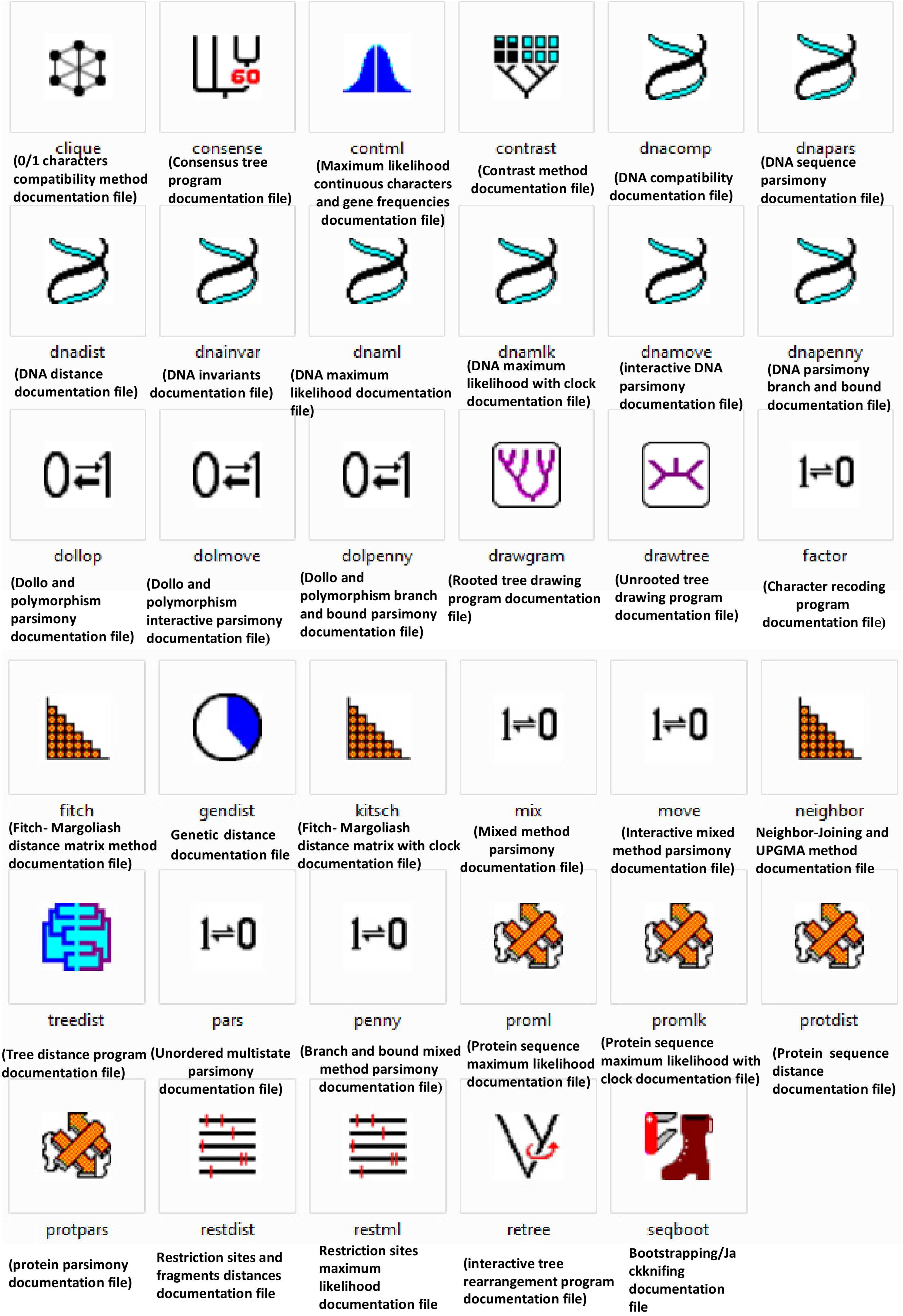
clique
(0/1 characters compatibility method documentation file)
consense
(Consensus tree program documentation file)
contml
(Maximum likelihood continuous characters and gene frequencies documentation file)
contrast
(Contrast method documentation file)
dnacomp
(DNA compatibility documentation file)
dnapars
(DNA sequence parsimony documentation file)
dnadist
(DNA distance documentation file)
dnainvar
(DNA invariants documentation file)
dnaml
(DNA maximum likelihood documentation file)
dnamlk
(DNA maximum likelihood with clock documentation file)
dnamove
(interactive DNA parsimony documentation file)
dnapenny
(DNA parsimony branch and bound documentation file)
dollop
(Dollo and polymorphism parsimony documentation file)
dolmove
(Dollo and polymorphism interactive parsimony documentation file)
dolpenny
(Dollo and polymorphism branch and bound parsimony documentation file)
drawgram
(Rooted tree drawing program documentation file)
drawtree
(Unrooted tree drawing program documentation file)
factor
(Character recoding program documentation file)
fitch
(Fitch- Margoliash distance matrix method documentation file)
gendist
Genetic distance documentation file
kitsch
(Fitch- Margoliash distance matrix with clock documentation file)
mix
(Mixed method parsimony documentation file)
move
(Interactive mixed method parsimony documentation file)
neighbor
Neighbor-Joining and UPGMA method documentation file
treedist
(Tree distance program documentation file)
pars
(Unordered multistate parsimony documentation file)
penny
(Branch and bound mixed method parsimony documentation file)
proml
(Protein sequence maximum likelihood documentation file)
promlk
(Protein sequence maximum likelihood with clock documentation file)
protdist
(Protein sequence distance documentation file)
protpars
(protein parsimony documentation file)
restdist
Restriction sites and fragments distances documentation file
restml
Restriction sites maximum likelihood documentation file
retree
(interactive tree rearrangement program documentation file)
seqboot
Bootstrapping/Jackknifing documentation file

Fig. 1 Some icons represent the modules included in the PHYLIP© package which carry out different tasks.

CASE STUDIES

Recently, utilization of Phylip modules provided the foundation for population genetics in different phyla. it was used recently in humans population research to study 215 independent Hungarian male samples and the genetic distances to 23 other populations using 49 Y-chromosomal single nucleotide polymorphisms (SNPs) and 11 Y-chromosomal STR loci (Völgyi *et al.* 2009). It was used to study the phylogeny of short-chain dehydrogenases/reductases (SDRs) in both *Drosophila melanogaster* and human retinol dehydrogenase 12 (RDH12) indicating a common ancestry early in bilaterian evolution (Belyaeva *et al.* 2009). PHYLIP© version 3.2 software were used to investigate human migration and evolution analysis and construction of phylogenetic tree based on the analysis of human platelet alloantigens (HPA) polymorphisms, in five systems which served as the genetic marker (Feng *et al.* 2008), to construct a Familial Mediterranean Fever (FMF) cumulative database and to propose a MEFV based phylogenetic tree (Papadopoulos *et al.* 2008) and to conduct phylogenetic analysis of transmembrane regions of GPCRs (Murakami *et al.* 2008).

In plants, the MaturaseK (*Mat*K) chloroplast gene, which is highly conserved in plant, was used for defining the inter- and intra-generic relationships among family of Zingiberaceae family which comprises 47 genera with medicinal values using PHYLIP© modules (Selvaraj *et al.* 2008). In the same manner PHYLIP© modules were used in animals to analysis of genetic diversity, for instance, in wild common carp (*Cyprinus carpio* L.) populations using microsatellite DNA marker (Li *et al.* 2007).

In prokaryotes, Phylip DNA modules were used to reveal genetic sequence evolution using 1500 bp fragments of the 16S rDNA gene from *Klebsiella pneumoniae* strains isolated from diarrhea specimens (Guo *et al.* 2008). In addition, it was used in the phenotypic classification and phylogeny estimation of the mycobacterial strains (Mignard and Flandrois 2008) and to analyze the genetic differences of *Orientia tsutsugamushi* (Ot) Sta56 gene between isolates from Shandong, China and other strains deposited in GenBank (Liu *et al.* 2007).

In virology, Phylip was used for studying phylogeny and genotype analysis among viruses such as enterovirus type 71, detected from hand-foot-mouth disease patients (He *el al.* 2008) and metapneumoviruses isolated in Chongqing, China (Mao *et al.* 2008).

CONCLUSION

Many researchers in the field of Molecular biology have used Phylip modules during their original research in Phylogenetics studies. Recently, many genomic evolutionary relatedness were discovered in different organisms using Phylip new features using published sequences in NCBI database. There are many reasons why Phylip become an instant hit these days. With more than 40,000 citations, Phylip is one of the most widely cited scientific programs in the history of biology. The freeware license and its efficient modules beside its quick ability to produce results make it the third popular programs for Phylogenetics studies nowadays after PAUP* and MrBayes.

ACKNOWLEDGEMENT

The author would like to thank Professor Joseph Felsenstein, Department of Genome Sciences and the Department of Biology, University of Washington, the original author of PHYLIP© program for his kind help in editing and correcting some information in this manuscript (Felsenstein 2008).

REFERENCES

Belyaeva OV, Lee SA, Kolupaev OV, Kedishvili NY (2009) Identification and characterization of retinoid-active short-chain dehydrogenases/reductases in *Drosophila melanogaster*. *Biochimca et Biophysca Acta* **1790**, 1266-1273

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17 (6)**, 368-376

Felsenstein J (1989) PHYLIP© - Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164-166

Felsenstein J (1992) Estimating effective population size from samples of sequences: a bootstrap *Monte Carlo* integration method. *Genetics Research* **59 (2)**, 139-147

Felsenstein J (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle

Felsenstein J (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods in Enzymology* **266**, 418-427

Felsenstein J (2005) Using the quantitative genetic threshold model for inferences between and within species. *Philosophical Transactions of the Royal Society of London Series* **360**, 1427-1434

Felsenstein J (2008) Comparative methods with sampling error and within-species variation: contrasts revisited and revised. *American Naturalist* **171 (6)**, 713-725

Feng ML, Huang H, Shen T, Zhang X, Yin B, Yang JH, Liu DZ (2008) Estimating genetic distance and phylogenetic tree of HPA-1-3, 5, and 15 in different populations. *Genetics* **30**, 838-42

Guo XL, Wang DC, Zhang YM, Wang XM, Zhang Y, Zuo Y, Zhang DM, Kan B, Wei L, Gao Y (2008) Isolation, identification and 16S rDNA phylogenetic analysis of *Klebsiella pneumonia* from diarrhea specimens. *Chinese Epidemiology Journal* **29**, 1225-1229

He YQ, Yang H, Li LL, Tan J, Zhou L, Mao LS, Yang F, Liu JJ, Lu X (2008) Genotype analysis of enterovirus type 71 detected from patients with hand-foot-mouth disease in Shenzhen. *Chinese Epidemiology Journal* **29**, 790-793

Li D, Kang D, Yin Q, Sun X, Liang L (2007) Microsatellite DNA marker analysis of genetic diversity in wild common carp (*Cyprinus carpio* L.) populations. *Journal of Genetics and Genomics* **34**, 984-93

Liu YX, Zhang Q, Zhao ZT, Yang ZQ, Yang LP, Zhang PH, Yang H, Yuan YE, Wei H, Suo JJ, Xing YB, Jia N, Gao Y, Cao WC (2007) Amplification and restriction fragment length polymorphism analysis on the complete sequence of Sta56 gene of *Orientia tsutsugamushi* isolated from Shandong area. *Chinese Epidemiology Journal* **28**, 886-890

Mao HW, Yang XQ, Zhao XD (2008) Characterization of human metapneumoviruses isolated in Chongqing, China. *Chinese Medical Journal* **121**, 2254-2257

Mignard S, Flandrois JP (2008) A seven-gene, multilocus, genus-wide approach to the phylogeny of mycobacteria using supertrees. *International Journal of Systematic and Evolutionary Microbiology* **58**, 1432-1441

Murakami M, Shiraishi A, Tabata K, Fujita N (2008) Identification of the orphan GPCR, P2Y(10) receptor as the sphingosine-1-phosphate and lysophosphatidic acid receptor. *Biochemical and Biophysical Research Communication* **371**, 707-712

Papadopoulos VP, Giaglis S, Mitroulis I, Ritis K (2008) The population genetics of familial Mediterranean fever: a meta-analysis study. *Annals of Human Genetics* **72**, 752-761

Selvaraj D, Sarma RK, Sathishkumar R (2008) Phylogenetic analysis of chloroplast *mat*K gene from Zingiberaceae for plant DNA barcoding. *Bioinformation* **3 (1)**, 24-27

Völgyi A, Zalán A, Szvetnik E, Pamjav H (2009) Hungarian population data for 11 Y-STR and 49 Y-SNP markers. *Forensic Science International: Genetics* **3 (2)**, e27-e28