# DIVERGE MANUAL version 3.0

## (DetectIng Variability in Evolutionary Rates among GEnes)

**Xun Gu**

**Iowa State University**

**Fudan University**

# Table of Contents

**Table of Contents** 1

**1 Preface** 3

1.1 copyright 3
1.2 Introduction 4
1.3 Acknowledgement 5
1.4 Suggested Citations 5

**2 PartⅠ: Getting Started** 6

2.1 System Requirement 6
2.2 Installation 6
2.2.1 Installation Procedure 6
2.2.2 Installation Step and Some Notification for the source code 7

**3 PartⅡ:Input Data Types and File Format** 8

3.1 The Multiple Alignment (Required) 8
3.1.1 FASTA Format 8
3.1.2 CLUSTAL Format 9
3.2 The Tree File Format (optional) 9
3.3 Protein Structure File Format (optional) 10

**4 Part Ⅲ: Procedure** 11

4.1 Sequences Tab 12
4.2 Clustering Tab 13
4.3 Gu99 Tab 14
4.3.1 Exported Data File Format 16
4.3.2 Structure Viewer 17
4.4 Gu 2001 Tab 18
4.5 Rate Variation among Sites(RVS) 19
4.6 Ancestral Sequence Inference 22
4.7 Functional Distance Analysis 23
4.8 Type II Divergence 25
4.9 Type I Analysis 28
4.10 Effective Number of Sites 30
4.11 Asymmetric Test for Type I 31
4.12 False Discovery Rate 32

5 Appendix ....................................33

5.1 Case study: COX (cyclooxygenase) gene family ....................................33

5.1.1 Type I and Type-II functional divergence analysis ....................................33

5.1.2 FDR for Predicted Type I Functional Divergence Related Sites ....................................36

5.1.3 Effective Number of Functional Divergence-Related Sites ....................................36

5.1.4 Site-specific Rate Change between COX1 and COX2 ....................................37

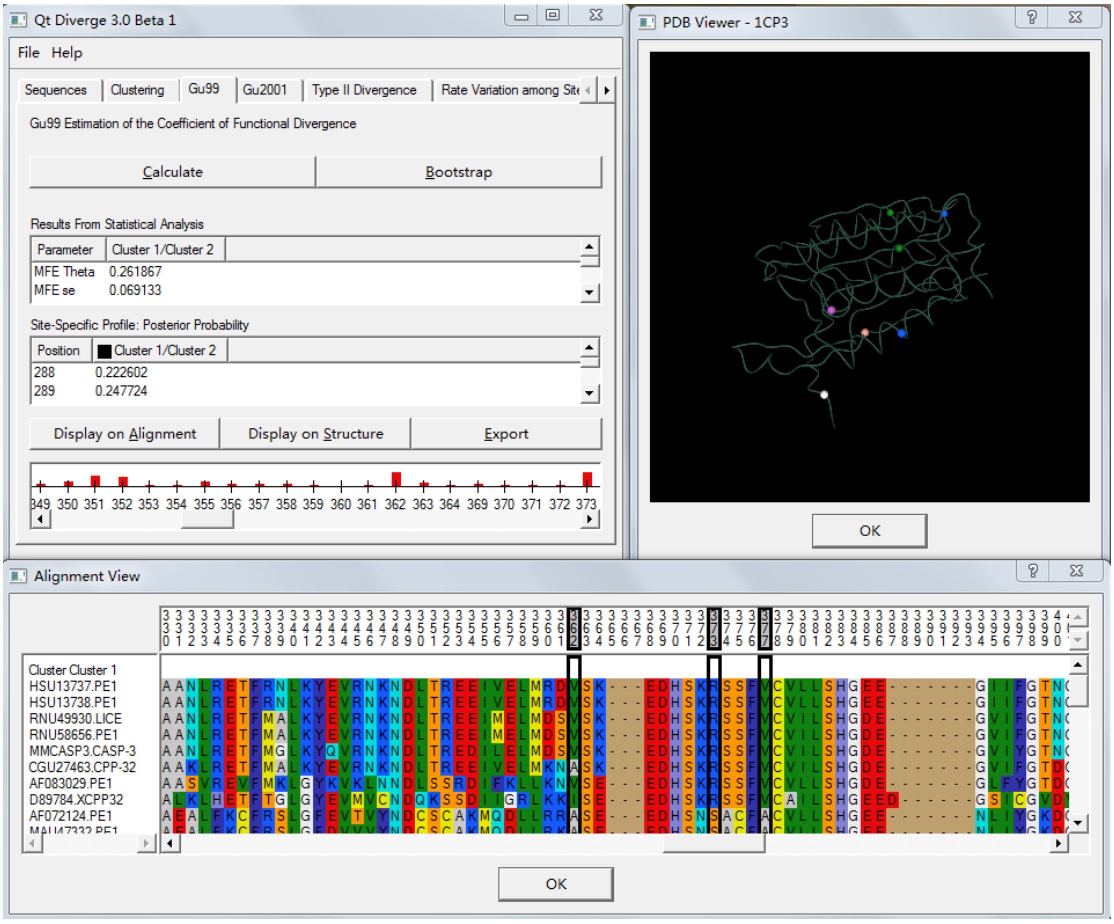
5.2 DIVERGE: Frequently Asked Questions ....................................37

5.3 Source Code Notes ....................................47

6. Reference ....................................49

# 1 Preface

## 1.1 copyright



Gu99 Estimation of the Coefficient of Functional Divergence

Results From Statistical Analysis

| Parameter | Cluster 1/Cluster 2 |
|---|---|
| MFE Theta | 0.261867 |
| MFE se | 0.069133 |

Site-Specific Profile: Posterior Probability

| Position | ■ Cluster 1/Cluster 2 |
|---|---|
| 288 | 0.222602 |
| 289 | 0.247724 |

DIVERGE is distributed free of charge by Xun Gu (Department of Genetics, Development and Cell Biology, 536 Science II, Iowa State University, Ames, IA 50011, Telephone: 515-294-8075 Fax: 515-294-8457, Email: xgu@iastate.edu. It can be downloaded from http://xungulab.com/software.html . )

# 1.2 Introduction

Many organisms have undergone genome-wide or local chromosome duplication events during their evolution[1, 2]. As a result, many genes are represented as several paralogs in the genome with related but distinct functions (gene families)[3]. Since gene duplication is thought to have provided the raw materials for functional innovations[4], it is desirable to identify amino acid sites that are responsible for functional divergence from the sequence analysis of a gene family[2, 5-7]. The software packages DIVERGE developed by our group have been widely used for this purpose[8-11].

The first version DIVERGE 1.0 was developed in 1999[10], which was designed to detect functional divergence between member genes of a protein family, based on (site-specific) shifted evolutionary rates after gene duplication or speciation. Posterior analysis results in a site-specific profile for predicting important amino acid residues that are responsible for functional divergence[12]. Moreover, when the 3D protein structure is available, these predicted sites can be mapped to the 3D structure viewer to explore its structure basis. Then , we updated our software to DIVERGE 2.0[13], which takes the interface structure of DIVERGE and implemented type-II functional divergence. In addition, some new tools that may be useful for data analyses are also implemented. Nevertheless, the main function features in DIVERGE 2.0 is the type-I and type-II functional divergence analyses. Because DIVERGE user body has being increased over years constantly, there is a demanding to have a concise, updated software review for the general readers, dealing with issues related to the principles and the technical issues without much mathematical formulas. In response to popular demand, we released an updated version DIVERGE3.0 with the following improvements[14]: 1) a feasible approach to examining functional divergence in nearly complete sequences by including deletions and insertions (indels); 2) the calculation of the false discovery rate of functionally diverging sites; 3) estimation of the effective number of functional divergence-related sites that is reliable and insensitive to cutoffs; 4) a statistical test for asymmetric functional divergence; and 5) a new method to infer functional divergence specific to a given duplicate cluster.

# 1.3 Acknowledgement

Many friends and colleagues have provided encouragement and assistance in the development of DIVERGE. We are grateful to all the participants for their constructive comments and suggestions.

This work was partly supported by grants from Fudan University and Iowa State University, and a grant from the Ministry of Science and Technology China.

# 1.4 Suggested Citations

[1] Gu, X. (1999) **Statistical methods for testing functional divergence after gene duplication**. Molecular Biology and Evolution 16:1664-1674.

[2] Wang Y, Gu X (2001) **Functional Divergence in the Caspase Gene Family and Altered Functional Constraints: Statistical Analysis and Prediction.** Genetics. 158:1311-1320.

[3] Gu, X (2001) **Maximum likelihood approach for gene family evolution under functional divergence**. Molecular Biology and Evolution 18:453-464.

[4] Gu J, Wang Y, Gu X (2002) **Evolutionary Analysis for Functional Divergence of Jak Protein Kinase Domains and Tissue-Specific Genes.** Journal of Molecular Evolution. 54:725-733.

[5] Gu, X, Vander Velden K (2002) **DIVERGE: Phylogeny-based Analysis for Functional-Structural Divergence of a Protein Family**. Bioinformatics 18:500-501.

[6] Gu X, Zou Y., Su Z., Huang W., Zhou Z., Arendsee A. and Zeng Y. (2013) **An update of DIVERGE software for functional divergence analysis of protein family.** Mol. Biol. Evol. 2013 30: 1713-1719.

# 2 Part Ⅰ: Getting Started

## 2.1 System Requirement

DIVERGE 3.0 can be used on Microsoft Windows operating systems: Windows 95/98, NT, 2000，XP，Windows7, or later. We recommend a computer with at least 32 MB of RAM, 10 MB hard disk space, and an entry-level Pentium processor or equivalent. Our tests show that DIVERGE 3.0 runs well even on computers with an 80486 CPU. However, for analysis of large datasets, you should have a faster processor and larger amount of physical memory for efficient computation.

In addition to Microsoft Windows, We also developed Linux version for DIVERGE 3.0. Later, we will release it on our website. To further facilitate users, we also provide the source code of DIVERGE 3.0 to the public. However, due to the proprietary issues, we can only release a simplified version. Users can easily compile the source code using Microsoft Visual C++ under Windows or GNU g++ under Linux, together with Qt GUI libraries from TrollTech(http://www.trolltech.com).

## 2.2 Installation

### 2.2.1 Installation Procedure

(1) After you fill out the contact information form, you will be able to download different versions of Diverge. Here, I assume you are using Windows for your research project. So, just click on Diverge3.0B1.zip and save it to your local disk.

(2) Now, unzip the downloaded file (using winzip or other tools) and you will find a new directory, DIVERGE3.0B1.

(3) Go into the newly created directory, and double-click on SETUP (an executable file). Please note that, the installation requires system administrative privilege. If you encounter problems related to this issue, please contact your local system administrator.

(4) Now, the installation will continue if you just keep clicking on Next button. When installation finishes, a Setup Complete window will popup and you have Diverge installed successfully.

(5) You can find the Diverge program in the windows start menu. Our software provides some demo data you can play with.

## 2.2.2 Installation Step and Some Notification for the source code

(1) First install the Microsoft Visual C++6.0 Professional Edition.

(2) Then install the Qt 3.3.4 Non-Commercial edition for Windows. Please do remember not to install QT to any directory with space, such as “program files”. Instead, install it to C:\qt.

Look at this link: http://lists.trolltech.com/qt-interest/2002-10/msg00310.html

(3) If the compilation fails because the moc.exe program cannot be found or the Qt header files are missing, check if the PATH contains the %QTDIR%\bin directory and the environment variable QTDIR is set to the directory in which you installed QT. After installing Qt, you may need to reboot your computer to make sure these settings are applied.

To set Qt path, you can go to desktop, right click “my computer”, go to “properties”, then choose “advanced” tab, click on “Environment” variable. On “user variables”, click “New” button. Variable Name is QTDIR. Variable Value is: C:\Qt. You may need to reboot to make it effective.

(4) If the compilation fails because nmake, or the compiler itself cannot be found, you must make sure that environment variables(PATH etc.) are set up for command-line use of the compiler. Run the vcvars32.bat file to setup the environment. This file can be found in the vc98\bin directory of the Visual Studio installation.

(5) If the compilation fails because the compiler does not find the system header files (e.g. stdio.h or windows.h), you must set the INCLUDE and LIB environment variables to contain the paths of the compiler’s header files and library files. The vcvars32.bat file takes care of this.

(6) If tmake fails make sure that the TMAKEPATH environment variable is set to %QTDIR%\tmake\lib\win32-msvc.

(7) When use Microsoft Visual C++ to open the diverge.dsw in dr_gu directory, please do remember first right click on dr_gu directory, and choose “properties”, then unselect the “read-only” and apply to all subdirectory. Then build and execute diverge.exe.

(8) When use Microsoft Visual C++ to execute the diverge.dsw, if you need to rebuilt and execute it several times, please do remember to close the executed diverge first, then rebuilt it. Otherwise, it will generate 1 compilation error.

(9) If QT toolbar in Microsoft Visual C++, you have to copy the “qmsdev.dll” to the Visual C++ Installation path:

C:\Programme\Microsoft Visual Studio\Common\MSDev98\AddIns

Details please see link: http://lists.trolltech.com/qt-interest/2002-03/msg00283.html

# 3 Part Ⅱ:Input Data Types and File Format

The input files of DIVERGE software are: 1) a multiple alignment of amino acid sequences (required), 2) a tree file with the evolutionary relationships of the sequences from the alignment file (optional), and 3) a structure PDB file (optional).

## 3.1 The Multiple Alignment (Required)

The alignment file must be in either FASTA or CLUSTAL form. Either file may contain as many as sequences as required. Only amino acid alignment is allowed in the current version. Gaps (-) in the alignment are allowed.

### 3.1.1 FASTA Format

The FASTA format may contain sequences that are split over numbers of lines or sequences that are on one long line or a mixture. If the sequences are of varying lengths (unaligned), the file will not be loaded and an error message is displayed. FASTA files typically have the extension .fasta. An example of the FASTA follows:

```
>AF025670.MCH2
MTETDGFYRSREVLDPAEQYKMDHKRRGTALIFNHERFFWHLALPERRGTNA
DRDNPTRRFSELGFEVKCFNDLRA
EELLLKIHEVSTSSHVDADCFLCVFLSHGEGNHIYAYDAKIEIQTLTGLFKGDK
CQSLVGKPKIFIIQAC
>AF072124.PE1
MTDDQDCAAELEMADSSTEDGVDAKPDRSTIISSLLWKKKKNASMCPVSTTR
DRVPTYLYRMDFEKMGKCIIINNK
NFDKATGMDVRNGTDKDAEALFKCFRSLGFEVTVYNDCSCAKMQDLLRRAS
EEDHSNSACFACVLLSHGE
>AF078533.PE1
MAEDKHNKNPLKMLESLGKELISGLLDDFVEKNVLKLEEEEKKKIYDAKLQD
KARVLVDSIRQKNQEAGQVFVQTF
LNIDKNSTSIKAPEETVAGPDESVGSAATLKLCPHEEFLKLCKERAGEIYPIKER
KDRTRLALIICNTEF
```

### 3.1.2 CLUSTAL Format

The CLUSTAL format is exactly the output file from the alignment software CLUSTAL. These files normally have the extension .aln. An example of a CLUSTAL aligned file follows. Notice the first line in the example. This line is read by the software to help determine the format of the alignment file. If the word "CLUSTAL" is in the first line the software assumes the file is in CLUSTAL format. If the alignment file is coming from another source and is in this format, you can get the software to read the alignment file by adding CLUSTAL to the top line of the file.

```
CLUSTAL W (1.7) multiple sequence alignment
HSU60521.MCH6          ------------------------------------------------------
CELCED3A.CED-3
-------------------MMRQDRRSLLERNIMMFSSHLKVDEILEVLIAKQVLNSD

HSU60521.MCH6
----MDEADRRLLRRCRLRLVEELQVDQLWDALLSSELFRPHMIEDIQRAGSGS
RRDQAR
CELCED3A.CED-3
NGDMINSCGTVREKRREIVKAVQRRGDVAFDAFYDALRSTGHEGLAEVLEPL
ARSVDSNA
```

## 3.2 The Tree File Format (optional)

Since Gu's (1999) method requires a phylogenetic tree of the gene family, DIVERGE provides an option to generate a neighbor-joining (NJ) tree[15] from the input alignment (see later). If the user decides to use a favored tree rather than the default NJ tree, it can be loaded from a file in the PHYLIP format as demonstrated below. The string representing the tree may be all contained on a single line or broke over a number of lines. Branch lengths (either floating point or integer values) are allowed and read if available but will not be used.

Example without branch lengths:

```
(((AF111345,HSU60519.MCH4),HSU86214.PE1),(HSCASP8S8.CASP8,MMCASP8
S7.PE1)));
```

Example with branch lengths:

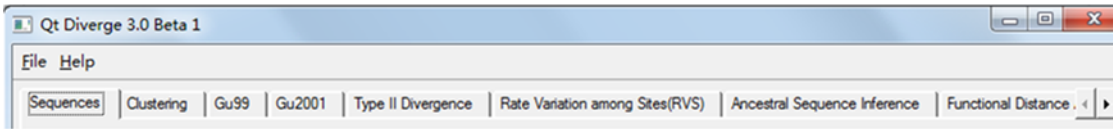
```
(((AF111345:.012,HSU60519.MCH4:.453):.345,HSU86214.PE1:.543):.546,(HSCAS
```

P8S8.CASP8:.954,MMCASP8S7.PE1:.42):.65);

## 3.3 Protein Structure File Format (optional)

A significant feature of DIVERGE is providing a 3D structure image to explore the structural basis of functional divergence of a protein family. The format of a structure file is the typical PDB file and normally has the extension .ent or .pdb. If a filename contains either of these in its extension, but suffixed with .Z or .gz then the file is compressed and will need to be decompressed before it may be used by DIVERGE.

# 4 Part Ⅲ: Procedure



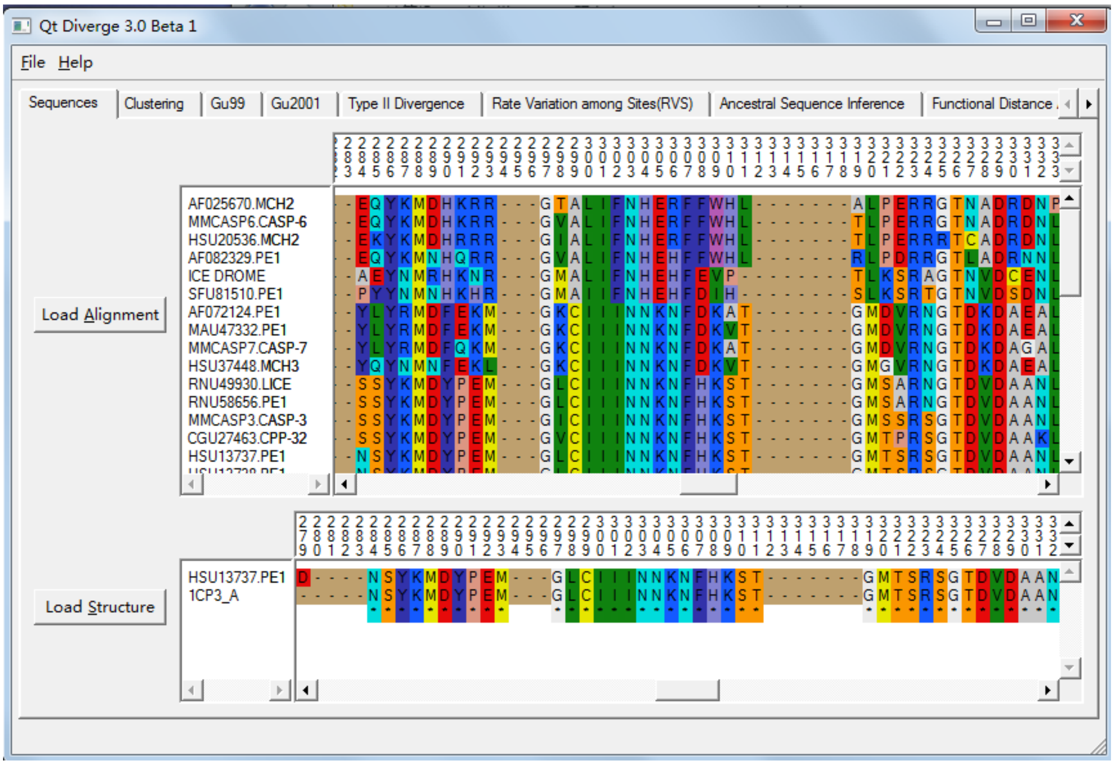
Across the top of the DIVERGE 3.0, there are twelve tabs named: 1) Sequences, 2) Clustering, 3) Gu99, 4)Gu2001, 5)Type II Divergence, 6)Rate Variation among Sites(RVS), 7)Ancestral Sequence Inference, 8)Functional Distance Analysis, 9)Type Ⅰ Analysis, 10)Effective Number of Sites, 11)Asymmetric Test fir Type I,12)False Discovery Rate. The first three steps are the main procedures that must be performed in order to use the DIVERGE software. Start with the leftmost tab (the Sequences tab), complete it, and then move on the next tab. Slowly working your way across. We have listed a brief description for analysis options Implemented in the Updated Version, DIVERGE3.0 (Table 1).

**Table 1. A Brief Description for Analysis Options Implemented in the Updated Version, DIVERGE3.0.**

| Function | Description |
|---|---|
| Gu99 | Detect type-I functional divergence by Gu (1999) method. |
| Gu2011 | Detect type-I functional divergence by Gu (2001) method. |
| Type-II Divergence | Detect type-II functional divergence of gene family. |
| Rate variation among sites (RVS) | Estimate the among-site rate variations for given cluster as described in Gu and Zhang (1997). |
| Ancestral sequence Inference | Infer the ancestral sequence for each internal node |
| Functional Distance Analysis | Estimate the type I functional distance for each pair of clusters and show the type I functional branch length of each cluster when at least three homologous gene clusters are available. |
| Gap involvement | Site-specific posterior profile for sites containing gaps |
| FDR for predictions | Provides more statistical evaluations for predicted sites |
| Asymmetric test for type-I functional divergence | Statistically testing whether the degree of type I functional divergence differs between two duplicate genes |
| Effective number of sites related to functional divergence(type I or type II) | Estimate effectively the number of sites related to type I and type II functional divergences , which is insensitive to the cutoff |

| | |
|---|---|
| Gene-specific type I analysis | Site-specific posterior profiles for predicting gene-specific type I functional divergence-related sites |

## 4.1 Sequences Tab

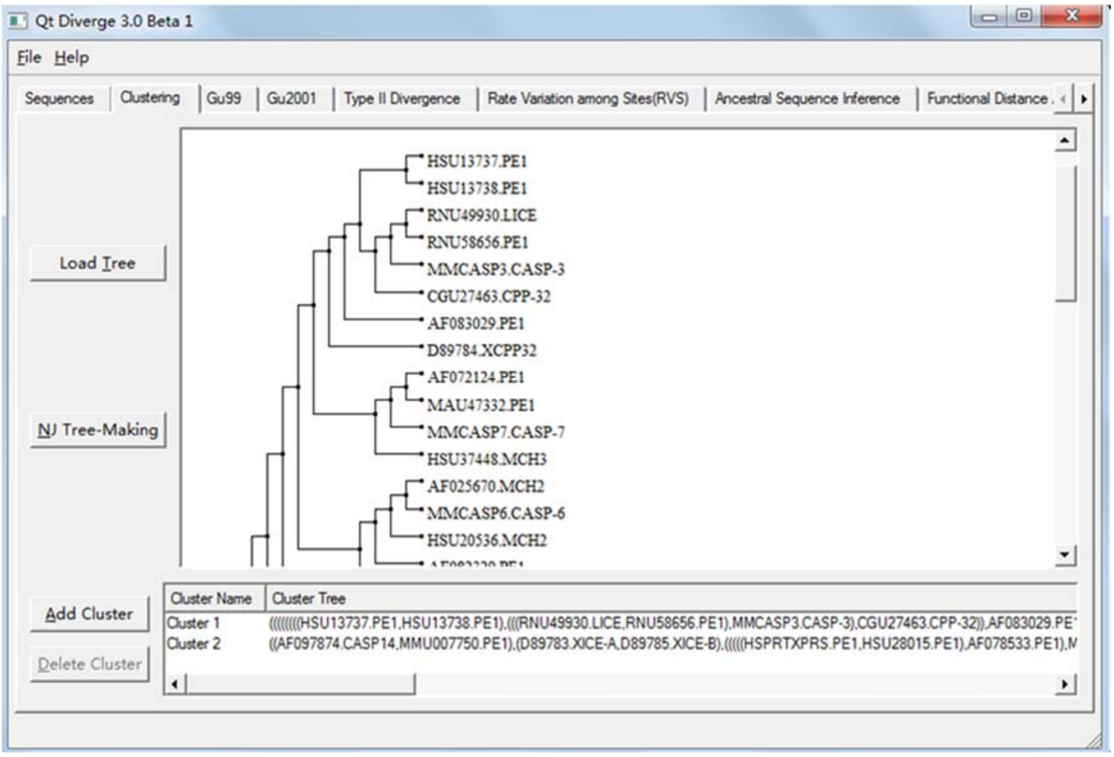


Under the Sequence Tab, the user needs to load an alignment file (FASTA or CLUSTAL), which can be done by selecting the "Load Alignment" button and then selecting the file. After a successful input of the alignment, the area to the right of the button will display the alignment. This area is fully navigable by use of the associate scrollbars. Scrollbars for the taxa names and site positions will become active when needed.

Also the Sequence Tab has the ability to load a structure file by selecting the "Load Structure" button. This structure file is assumed to be in the PDB format and to be related to the sequences in the alignment. The association between the multiple alignments and the 3D structure is done by performing a pairwise alignment between each sequence in the alignment and that of every chain in the structure file. The alignment that has the best score is used to create a map from positions on the alignment to residues in the structure.

Care needs to be taken when a structure is being loaded. Since the software can not judge the relatedness of the sequences in the alignment and that in the structure, the user will need to use their best judgment to decide if the choice of structure is correct.

## 4.2 Clustering Tab



| Cluster Name | Cluster Tree |
|---|---|
| Cluster 1 | (((((((HSU13737.PE1,HSU13738.PE1),(((RNU49930.LICE,RNU58656.PE1),MMCASP3.CASP-3),CGU27463.CPP-32)),AF083029.PE |
| Cluster 2 | ((AF097874.CASP14,MMU007750.PE1),(D89783.XICE-A,D89785.XICE-B),(((((HSPRTXPRS.PE1,HSU28015.PE1),AF078533.PE1),M |

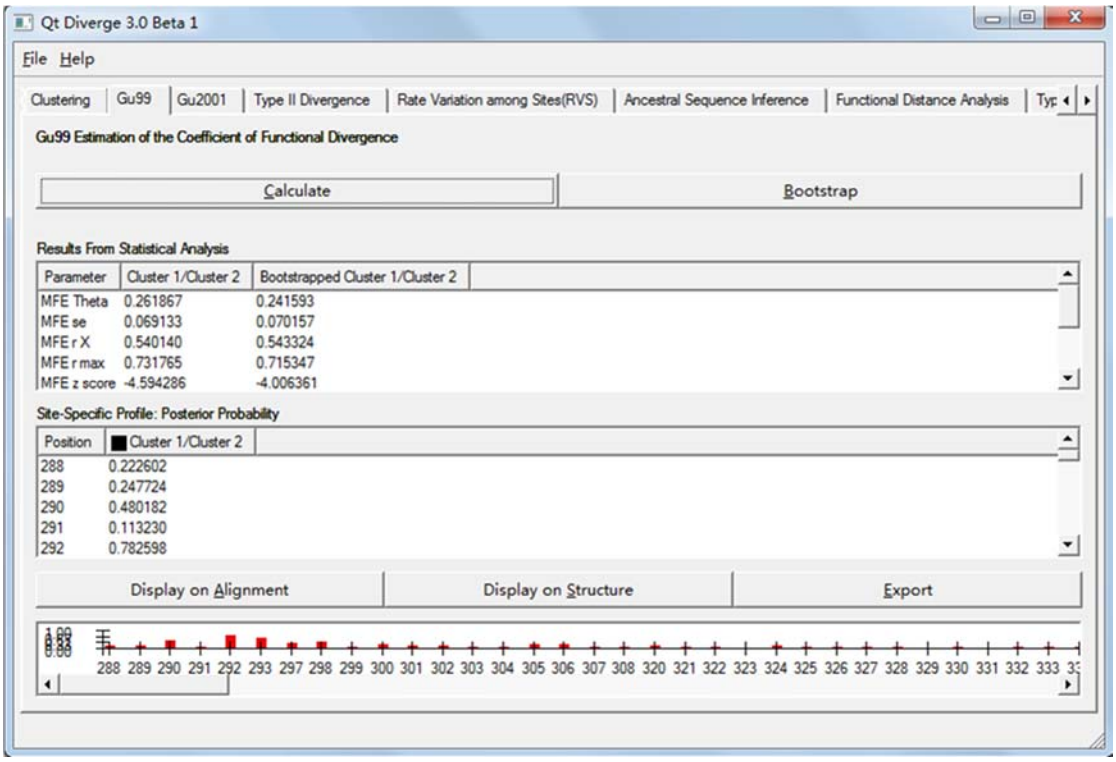
In the clustering tab, the user needs to perform two steps. First, a tree must be either loaded from a file or generated by using the available Neighbor-Joining (NJ) option. Secondly, the user must select subtrees from the main gene tree to represent the independent gene clusters by clicking the ancestral nodes.

Loading a tree from a file requires that the file be in the standard parenthesis PHYLIP form. This is a common form and is generated most phylogenetic tree software systems (e.g., PHYLIP, PAUP*, or CLUSTAL). Branch lengths are optional and if they are available will be used in drawing the tree. Obviously, the taxa names used in the tree must be the same as those used in the alignment. To load a tree file simply select the "Load Tree" button and select the tree file to be loaded.

For many users who are not familiar the with tree presentation, using the Neighbor-Joining option is suggested. By selecting the "NJ Tree-Making" button the software will use the Neighbor-Joining algorithm to quickly generate a tree based on the distance measure (i.e. p-Distance, Poisson, and Kimura) as selected.

After a tree is available, the clusters that are monophyletic (technically, no overlapping along the tree) must be selected. At least two clusters are required. If multiple clusters are selected, pairwise comparisons will be performed. To add a cluster for analysis, first select the node on the tree which forms the root of that cluster. This should highlight that portion of the tree in red. Then select the "Add Cluster" button. After typing a name for the selected cluster, it will be added to the list of clusters below the gene family tree. Since each cluster will be referred to by name later, unique names must be assigned. Clusters in the Cluster List can be viewed in the Tree Viewer by double-clicking them with the mouse and removed by first selecting them with the mouse and then selecting the "Delete Cluster" button.

## 4.3 Gu99 Tab



Gu99 Estimation of the Coefficient of Functional Divergence

Results From Statistical Analysis

| Parameter | Cluster 1/Cluster 2 | Bootstrapped Cluster 1/Cluster 2 |
|---|---|---|
| MFE Theta | 0.261867 | 0.241593 |
| MFE se | 0.069133 | 0.070157 |
| MFE r X | 0.540140 | 0.543324 |
| MFE r max | 0.731765 | 0.715347 |
| MFE z score | -4.594286 | -4.006361 |

Site-Specific Profile: Posterior Probability

| Position | ■ Cluster 1/Cluster 2 |
|---|---|
| 288 | 0.222602 |
| 289 | 0.247724 |
| 290 | 0.480182 |
| 291 | 0.113230 |
| 292 | 0.782598 |

Once all the steps above (i.e. the multiple alignment, the tree, and cluster selection) have been completed, the user can start the statistical analysis by selecting the "Calculate" button. If any steps have been skipped, the software will warn the user of this and cancel the calculations. If everything is fine a progress bar will appear and give the user an estimate of the remaining time required for completion. The time required is mainly a function of the number of selected clusters since all pairwise clusters will be analyzed. The calculation can be canceled by pressing the "Cancel" button.

Once the calculations are complete, the user will be presented with the Statistical Results in the upper most regions, listed for each comparison[10]. Here we try to explain the results presented in the software from the view of users (Table 2); one may refer to the original papers for the technical details.

**Table 2. Interpretations of parameters presented in the output of DIVERGE3.0 for the analysis of type-I functional divergence.**

| Parameters | Interpretations |
|---|---|
| MFE Theta | Estimate of $\theta_I$ by the model-free method |
| MFE se | Standard error of the $\theta_I$ estimated by MFE |
| MFE r_X | The observed coefficient of correlation between two gene clusters |
| MFE r_max | The expected maximum coefficient of correlation between two gene clusters |
| MFE z-score | The z-score for the model-free estimate of $\theta_I$ after Fisher's transformation |
| ThetaML | Maximum likelihood estimate of $\theta_I$ |
| AlphaML | Maximum likelihood estimate of α (the gamma parameter for the among-site rate variation |
| SE Theta | Standard error of the maximum likelihood estimate of $\theta_I$ |
| LRT Theta | The (log) score for the likelihood ratio test against the null $\theta_I$ =0. |

The site-specific profile(s) will be presented in the middle region, which is the posterior probability of a site to be functional divergence-related. Moreover, when the cut-off value is given, they can be viewed in relation to the alignment and/or the protein structure. Finally, these results can be exported to a file for processing by other applications. These actions are controlled by use of the buttons above the graph region.

The alignment and structure viewer options require the user select which pairwise comparisons to examine. Bar graphs for site-specific profile will only appear after selection, which is done by clicking the small square in the header of the site-specific profile. As shown in the example "Cluster 1/Cluster 2" was selected.

Statistically Evaluate the estimate statistical $\theta_I/\theta_{II}$ either from MFE or ML methods:

(1) One may use the standard error (MFE se) for the MFE estimate, or SE Theta for the ML estimate to calculate the p-value. For instance, if one obtains $\theta_I=0.3\pm0.05$, calculate the score=0.3/0.05=6 and then obtain the p-value<0.001 from the Z-score test (normal distribution test).

(2) In the case of MFE estimate, the MFE z-score in the output may be statistically more accurate because it is based on the Fisher's transformation. Note that MFE z-score is usually negative, so the user has to use its absolute value to obtain the p-value from the Z-core test.

(3) In the case of ML estimate, the likelihood ratio test can be applied. The value of LRT Theta in the output is the log-score so that it approximately follows a chi-square distribution with one degree of freedom.

The "Display on Alignment" and "Display on Structure" buttons require that at least one column of site-specific profiles be selected and the user should supply a cutoff value. Those sites that have a posterior probability larger than the cutoff value will be highlighted on the alignment or structure respectfully. If no column has been selected a warning message will appear. Also, as with the graph viewer, multiple columns may be selected. The Structure Viewer is explained further in a later section.

## 4.3.1 Exported Data File Format

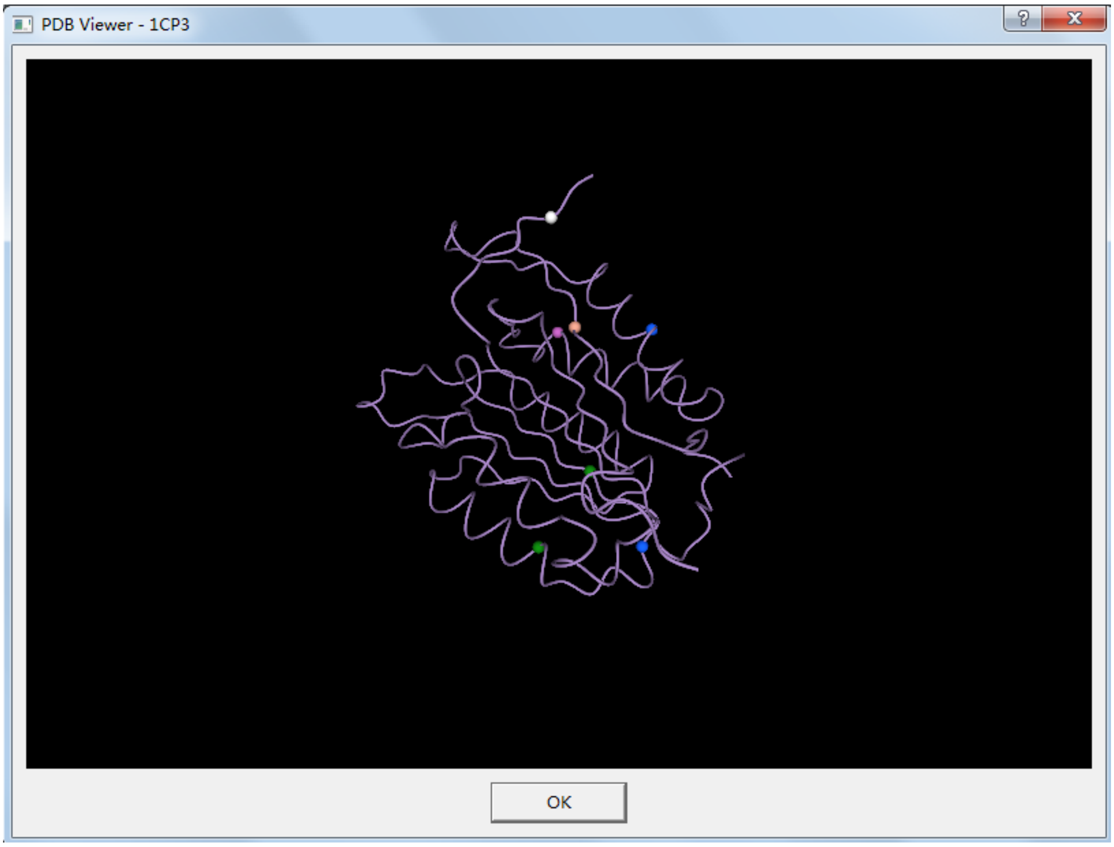
When the "Export" button is selected, all the results are exported to a specified file in tab delimited format for your own personal records and additional processing by other software. Included in the output is the original tree, the selected gene clusters, the Statistical Results, and site-specific profiles for each pair of clusters. An example of this format is shown below.

```
Main Tree:
((((((((HSU13737.PE1,HSU13738.PE1),(((RNU49930.LICE,RNU58656.PE1),MMCASP3.CASP-3),
CGU27463.CPP-32)),AF083029.PE1),D89784.XCPP32),
(((AF072124.PE1,MAU47332.PE1),MMCASP7.CASP-7),HSU37448.MCH3)),((((AF025670.MCH2,
MMCASP6.CASP-6),HSU20536.MCH2),AF082329.PE1),(ICE_DROME,SFU81510.PE1))),
(((AF111345,HSU60519.MCH4),HSU86214.PE1),(HSCASP8S8.CASP8,MMCASP8S7.PE1))),CELC
ED3A.CED-3,((((RRU77933.PE1,MMCASP2.CASP-2),HSU13021.ICH-1),GGU64963.ICH-1),
HSU60521.MCH6)),((AF097874.CASP14,MMU007750.PE1),(D89783.XICE-A,D89785.XICE-B),(((
((HSPRTXPRS.PE1,HSU28015.PE1),AF078533.PE1),MMCASP11.CASP-11),
MMCASP12.CASP-12),((MUSIL1B.PE1,RNU14647),(AF090119,HSIL1BRNA.PE1)))))
Cluster a:
((HSU13737.PE1,HSU13738.PE1),(((RNU49930.LICE,RNU58656.PE1),MMCASP3.CASP-3),CGU2
7463.CPP-32),AF083029.PE1)
Cluster b:      ((AF072124.PE1,MAU47332.PE1),MMCASP7.CASP-7,HSU37448.MCH3)
Cluster c:
(((AF025670.MCH2,MMCASP6.CASP-6),HSU20536.MCH2),AF082329.PE1,(ICE_DROME,SFU81
510.PE1))

            a/b             a/c            b/c
ThetaML 0.560800         0.412800       0.001000
AlphaML 0.319958          1.111486       1.790179
SE Theta    0.168399        0.131543       0.022361
LRT Theta 11.090167        9.847951       0.000000
288        0.538375        0.346385       0.000927
```

| | | | |
|---|---|---|---|
| 289 | 0.800534 | 0.478965 | 0.000784 |
| 290 | 0.538375 | 0.346385 | 0.000927 |
| … | | | |
| 571 | 0.538375 | 0.346385 | 0.000927 |
| 572 | 0.538375 | 0.346385 | 0.000927 |

## 4.3.2 Structure Viewer



While identifying the regions responsible for functional divergence on the aligned sequence can be informative, plotting the regions on the structure can be more rewarding[16]. Identified residues that do not seem to have any relationship to one another on the alignment may form clusters when viewed on the structure[17]. The density of these clusters may be such they would be unlikely to occur at random and suggest a driving force in this particular region of the protein warranting further investigation.

The structure viewer will appear after initially loading a support structure file in the PDB file format. The structure viewer can be dismissed at any time by selecting OK at the bottom of the viewer or pressing Esc while the viewer has active focus. The viewer will appear again after the "Display on Structure" option has been selected or an item or range in the site-specific profile is selected.

Navigation in the Structure Viewer is accomplished mainly with the mouse, with a few keyboard commands of lesser importance.

To use the mouse controls, press the corresponding mouse button and then move the mouse for the appropriate action.

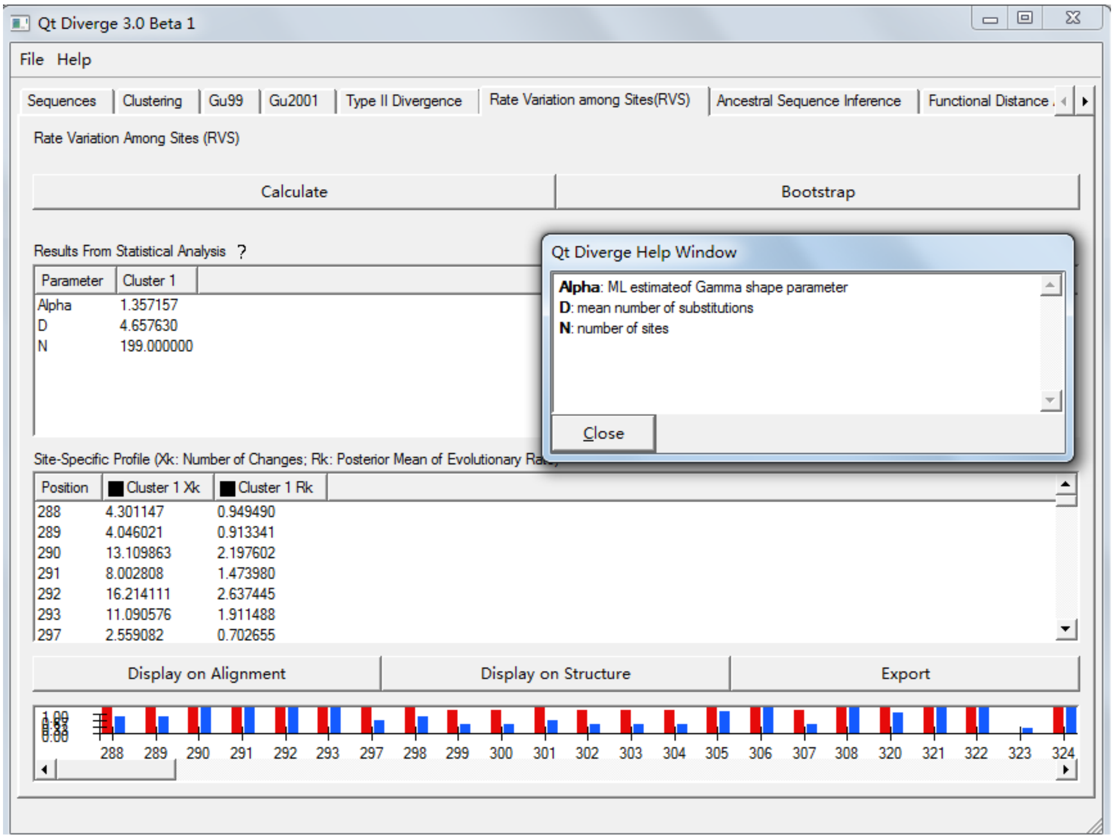
| **Mouse Button** | **Action** |
|---|---|
| Button 1 | Rotate the camera |
| Button 2 | Pan the camera |
| Button 3 | Zoom the camera |

| **Keyboard Keys** | **Action** |
|---|---|
| w | Wireframe mode |
| s | Solid mode |
| r | Reset camera view |
| p | Pick |

# 4.4 Gu 2001 Tab

Using Gu2001 is similar with Gu99. You need to choose at least two clusters first, and then calculate.

# 4.5 Rate Variation among Sites(RVS)



| Parameter | Cluster 1 |
|---|---|
| Alpha | 1.357157 |
| D | 4.657630 |
| N | 199.000000 |

**Alpha**: ML estimateof Gamma shape parameter
**D**: mean number of substitutions
**N**: number of sites

| Position | ■ Cluster 1 Xk | ■ Cluster 1 Rk |
|---|---|---|
| 288 | 4.301147 | 0.949490 |
| 289 | 4.046021 | 0.913341 |
| 290 | 13.109863 | 2.197602 |
| 291 | 8.002808 | 1.473980 |
| 292 | 16.214111 | 2.637445 |
| 293 | 11.090576 | 1.911488 |
| 297 | 2.559082 | 0.702655 |

When the steps like the multiple alignments, the tree, and the cluster selection have been completed, the user can start the statistical analysis by selecting the “Calculation” button. Different from Gu99, which needs to create at least two clusters first, RVS method needs to create only one cluster. If any step above has been skipped, the software will warn the user of this and cancel the calculation. In a normal computation, a progress bar will appear and give the user an estimate of the remaining time required for completion. The time required is mainly a function of the number of selected clusters since all pairwise clusters will be analyzed. The calculation can be canceled by pressing the "Cancel" button.

Once the calculations are completed, the user will be presented with the Statistical Results in the uppermost region, listed for each comparison. Hit the “?” sign and a pop up window will show the meaning of these results.

ML Estimates:

Alpha: Gamma Shape Parameter

D: Mean number of Substitutions

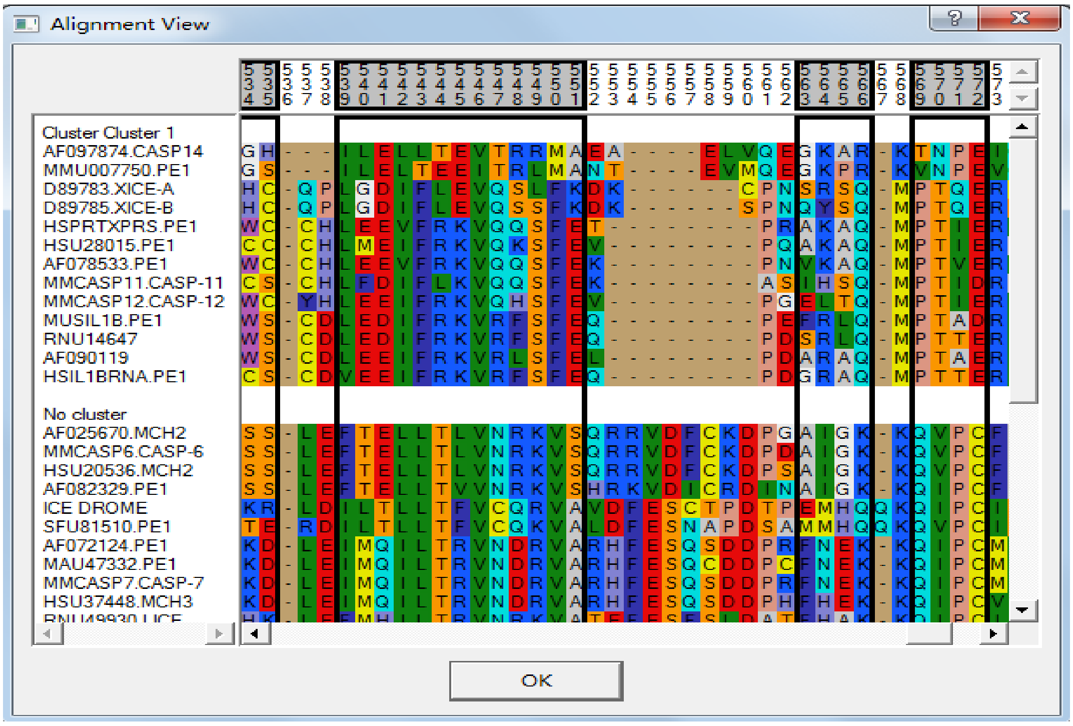
N: Number of Sites

The site-specific profile(s) will be presented in the middle region, which is the posterior probability of a site to be functional divergence-related.

XK: Number of Changes

RK: Posterior Mean of Evolutionary Rate

Moreover, when the cut-off value is given, they can be viewed in relation to the alignment and/or the protein structure. Finally, these results can be exported to a file for processing by other applications. These actions are controlled by using the buttons above the graph regions.



| | 534 | 535 | 536 | 537 | 538 | 539 | 540 | 541 | 542 | 543 | 544 | 545 | 546 | 547 | 548 | 549 | 550 | 551 | 552 | 553 | 554 | 555 | 556 | 557 | 558 | 559 | 560 | 561 | 562 | 563 | 564 | 565 | 566 | 567 | 568 | 569 | 570 | 571 | 572 | 573 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cluster Cluster 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| AF097874.CASP14 | G | H | - | - | - | I | L | E | L | L | T | E | V | T | R | R | M | A | E | A | - | - | - | - | E | L | V | Q | E | G | K | A | R | - | K | T | N | P | E | I |
| MMU007750.PE1 | G | S | - | - | - | I | L | E | L | T | E | E | I | T | R | L | M | A | N | T | - | - | - | - | E | V | M | Q | E | G | K | P | R | - | K | V | N | P | E | V |
| D89783.XICE-A | H | C | - | Q | P | L | G | D | I | F | L | E | V | Q | S | L | F | K | D | K | - | - | - | - | - | - | C | P | N | S | R | S | Q | - | M | P | T | Q | E | R |
| D89785.XICE-B | H | C | - | Q | P | L | G | D | I | F | L | E | V | Q | S | S | F | K | D | K | - | - | - | - | - | - | S | P | N | Q | Y | S | Q | - | M | P | T | Q | E | R |
| HSPRTXPRS.PE1 | W | C | - | C | H | L | E | E | V | F | R | K | V | Q | Q | S | F | E | T | - | - | - | - | - | - | - | - | P | R | A | K | A | Q | - | M | P | T | I | E | R |
| HSU28015.PE1 | C | C | - | C | H | L | M | E | I | F | R | K | V | Q | K | S | F | E | V | - | - | - | - | - | - | - | - | P | Q | A | K | A | Q | - | M | P | T | I | E | R |
| AF078533.PE1 | W | C | - | C | H | L | E | E | V | F | R | K | V | Q | Q | S | F | E | K | - | - | - | - | - | - | - | - | P | N | V | K | A | Q | - | M | P | T | V | E | R |
| MMCASP11.CASP-11 | C | S | - | C | H | L | F | D | I | F | L | K | V | Q | Q | S | F | E | K | - | - | - | - | - | - | - | - | A | S | I | H | S | Q | - | M | P | T | I | D | R |
| MMCASP12.CASP-12 | W | C | - | Y | H | L | E | E | I | F | R | K | V | Q | H | S | F | E | V | - | - | - | - | - | - | - | - | P | G | E | L | T | Q | - | M | P | T | I | E | R |
| MUSIL1B.PE1 | W | S | - | C | D | L | E | D | I | F | R | K | V | R | F | S | F | E | Q | - | - | - | - | - | - | - | - | P | E | F | R | L | Q | - | M | P | T | A | D | R |
| RNU14647 | W | S | - | C | D | L | E | D | I | F | R | K | V | R | F | S | F | E | Q | - | - | - | - | - | - | - | - | P | D | S | R | L | Q | - | M | P | T | T | E | R |
| AF090119 | W | S | - | C | D | L | E | E | I | F | R | K | V | R | L | S | F | E | L | - | - | - | - | - | - | - | - | P | D | A | R | A | Q | - | M | P | T | A | E | R |
| HSIL1BRNA.PE1 | C | S | - | C | D | V | E | E | I | F | R | K | V | R | F | S | F | E | Q | - | - | - | - | - | - | - | - | P | D | G | R | A | Q | - | M | P | T | T | E | R |
| No cluster | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| AF025670.MCH2 | S | S | - | L | E | F | T | E | L | L | T | L | V | N | R | K | V | S | Q | R | R | V | D | F | C | K | D | P | G | A | I | G | K | - | K | Q | V | P | C | F |
| MMCASP6.CASP-6 | S | S | - | L | E | F | T | E | L | L | T | L | V | N | R | K | V | S | Q | R | R | V | D | F | C | K | D | P | D | A | I | G | K | - | K | Q | V | P | C | F |
| HSU20536.MCH2 | S | S | - | L | E | F | T | E | L | L | T | L | V | N | R | K | V | S | Q | R | R | V | D | F | C | K | D | P | S | A | I | G | K | - | K | Q | V | P | C | F |
| AF082329.PE1 | S | S | - | L | E | F | T | E | L | L | T | V | V | N | R | K | V | S | H | R | K | V | D | I | C | R | D | I | N | A | I | G | K | - | K | Q | I | P | C | F |
| ICE DROME | K | R | - | L | D | I | L | T | L | L | T | F | V | C | Q | R | V | A | V | D | F | E | S | C | T | P | D | T | P | E | M | H | Q | Q | K | Q | I | P | C | I |
| SFU81510.PE1 | T | E | - | R | D | I | L | T | L | L | T | F | V | C | Q | K | V | A | L | D | F | E | S | N | A | P | D | S | A | M | M | H | Q | Q | K | Q | V | P | C | I |
| AF072124.PE1 | K | D | - | L | E | I | M | Q | I | L | T | R | V | N | D | R | V | A | R | H | F | E | S | Q | S | D | D | P | R | F | N | E | K | - | K | Q | I | P | C | M |
| MAU47332.PE1 | K | D | - | L | E | I | M | Q | I | L | T | R | V | N | D | R | V | A | R | H | F | E | S | Q | C | D | D | P | C | F | N | E | K | - | K | Q | I | P | C | M |
| MMCASP7.CASP-7 | K | D | - | L | E | I | M | Q | I | L | T | R | V | N | D | R | V | A | R | H | F | E | S | Q | S | D | D | P | R | F | N | E | K | - | K | Q | I | P | C | M |
| HSU37448.MCH3 | K | D | - | L | E | I | M | Q | I | L | T | R | V | N | D | R | V | A | R | H | F | E | S | Q | S | D | D | P | H | F | H | E | K | - | K | Q | I | P | C | V |
| RNU49930.ICE | H | K | - | L | E | F | M | H | I | L | T | R | V | N | R | K | V | A | T | E | F | E | S | F | S | L | D | A | T | F | H | A | K | - | K | Q | I | P | C | I |

**Figure 1. The alignment windows with the sites whose posterior probability is larger than cutoff value**

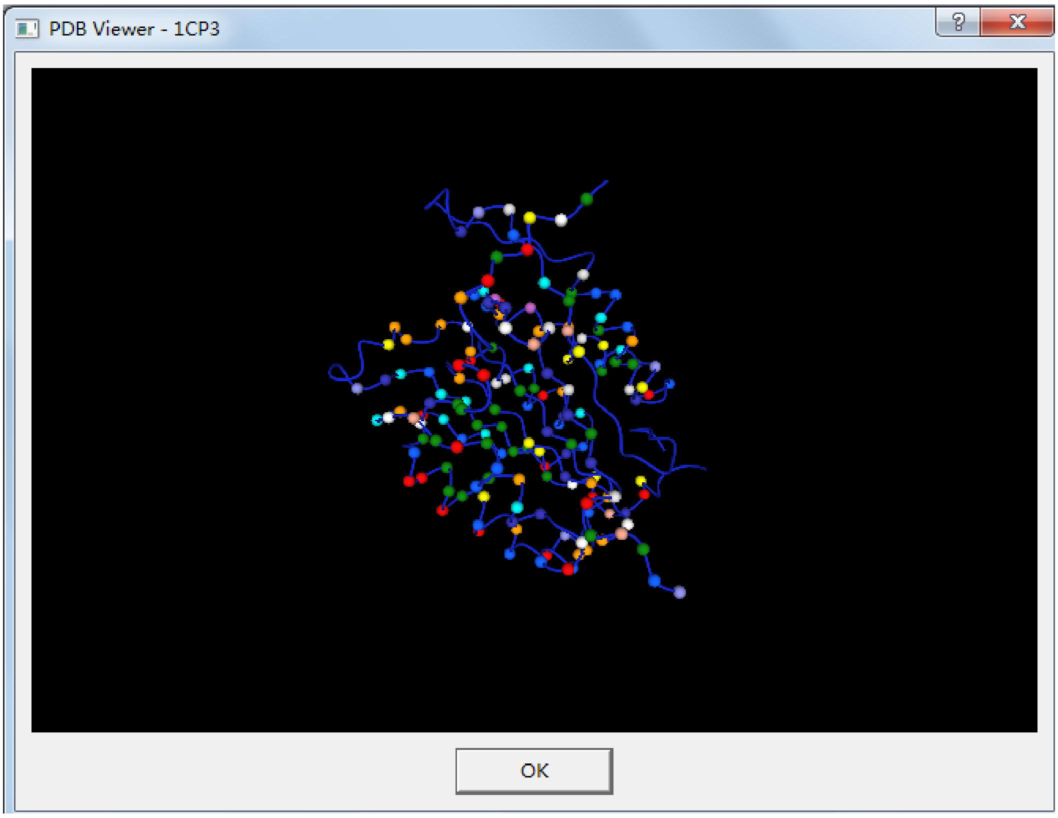
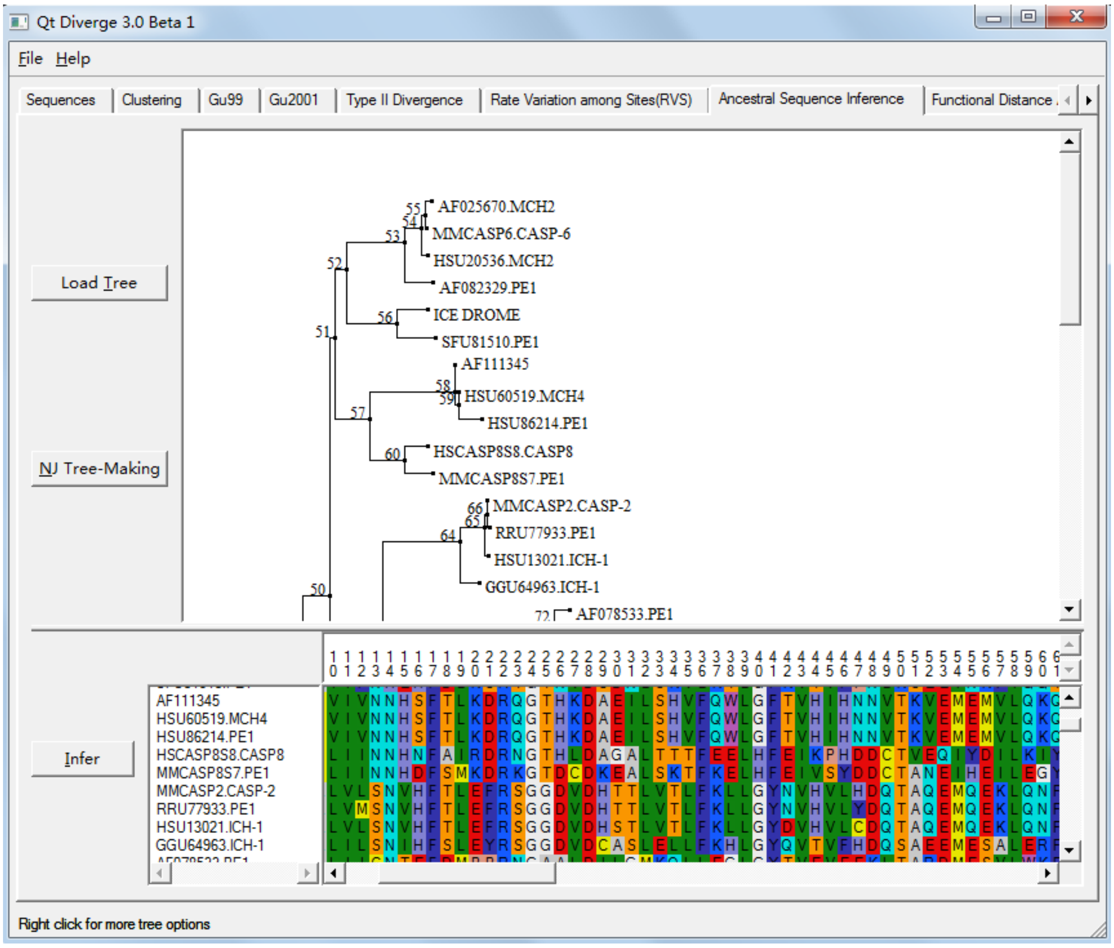


**Figure 2. The protein structure windows with the sites whose posterior probability is larger than cutoff value**

The "Display on Alignment" and "Display on Structure" buttons require that at least one column of site-specific profiles is selected and the user should supply a cutoff value. Those sites that have a posterior probability larger than the cutoff value will be highlighted on the alignment or protein structure, respectively. If no column has been selected, a warning message will appear. The Structure Viewer has already explained in the manual of the first version. For comparison’s purpose, we add a “Bootstrap” function so the user can repeat above step for a given-number times after the first calculation is finished.

# 4.6 Ancestral Sequence Inference



```
                  1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 6 6
                  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
AF111345          V I V N N H S F T L K D R Q G T H K D A E I L S H V F Q W L G F T V H I H N N V T K V E M E M V L Q K Q
HSU60519.MCH4     V I V N N H S F T L K D R Q G T H K D A E I L S H V F Q W L G F T V H I H N N V T K V E M E M V L Q K Q
HSU86214.PE1      V I V N N H S F T L K D R Q G T H K D A E I L S H V F Q W L G F T V H I H N N V T K V E M E M V L Q K Q
HSCASP8S8.CASP8   L I I N N H N F A I R D R N G T H L D A G A L T T T F E E L H F E I K P H D D C T V E Q I Y D I L K I Y
MMCASP8S7.PE1     L I I N N H D F S M K D R K G T D C D K E A L S K T F K E L H F E I V S Y D D C T A N E I H E I L E G Y
MMCASP2.CASP-2    L V L S N V H F T L E F R S G G D V D H T T L V T L F K L L G Y N V H V L H D Q T A Q E M Q E K L Q N F
RRU77933.PE1      L V M S N V H F T L E F R S G G D V D H T T L V T L F K L L G Y N V H V L Y D Q T A Q E M Q E K L Q N F
HSU13021.ICH-1    L V L S N V H F T L E F R S G G D V D H S T L V T L F K L L G Y D V H V L C D Q T A Q E M Q E K L Q N F
GGU64963.ICH-1    L I L S N I H F S L E Y R S G G D V D C A S L E L L F K H L G Y Q V T V F H D Q S A E E M E S A L E R F
```

Right click for more tree options

After constructing the tree, the user can press “Infer” button to infer the ancestral sequence. A progress bar will appear and give the user an estimate of the remaining proportion required for completion. This can be canceled by pressing the "Cancel" button.

After ancestral sequence inference is finished, the results will be shown on the bottom region. The inference results for each internal node (labeled as Internal Node 123) will be displayed in the same row. As indicated in Gu99 paper, the inference procedure is based on the algorithm proposed by Jianzhi Zhang et al[18].

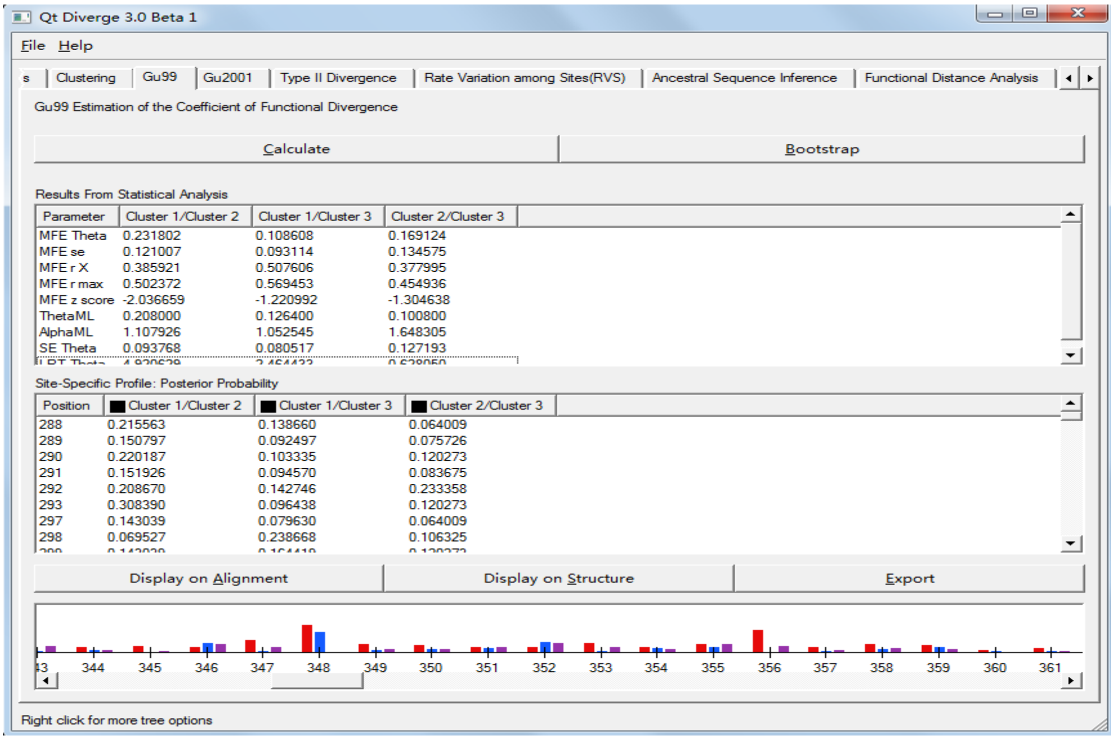
# 4.7 Functional Distance Analysis

The two-cluster analysis in Gu99's cannot tell in which gene cluster the altered functional constraint took place after gene duplication. This problem can be solved by a simple method when at least three homologous gene clusters are available. The functional distance analysis in DIVERGE2.0 is designed for this purpose. The analysis result, functional distance matrix, is based on the coefficient of type I functional divergence ($\theta_{ij}$) of each pair of clusters as follows.

$$d_F(i,j) = -\ln(1-\theta_{ij})$$

For the detailed information about functional distance analysis, please refer to Wang Y, Gu X (2001) Predicting functional divergence of caspase gene family. Genetics. 158:1311-1320.

Using functional distance analysis is very straightforward. First, please use Gu99 method with at least three clusters (Figure 3), Then Functional Distance Analysis results will be calculated and displayed automatically (Figure 4).



Gu99 Estimation of the Coefficient of Functional Divergence

Results From Statistical Analysis

| Parameter | Cluster 1/Cluster 2 | Cluster 1/Cluster 3 | Cluster 2/Cluster 3 |
|---|---|---|---|
| MFE Theta | 0.231802 | 0.108608 | 0.169124 |
| MFE se | 0.121007 | 0.093114 | 0.134575 |
| MFE r X | 0.385921 | 0.507606 | 0.377995 |
| MFE r max | 0.502372 | 0.569453 | 0.454936 |
| MFE z score | -2.036659 | -1.220992 | -1.304638 |
| ThetaML | 0.208000 | 0.126400 | 0.100800 |
| AlphaML | 1.107926 | 1.052545 | 1.648305 |
| SE Theta | 0.093768 | 0.080517 | 0.127193 |

Site-Specific Profile: Posterior Probability

| Position | ■ Cluster 1/Cluster 2 | ■ Cluster 1/Cluster 3 | ■ Cluster 2/Cluster 3 |
|---|---|---|---|
| 288 | 0.215563 | 0.138660 | 0.064009 |
| 289 | 0.150797 | 0.092497 | 0.075726 |
| 290 | 0.220187 | 0.103335 | 0.120273 |
| 291 | 0.151926 | 0.094570 | 0.083675 |
| 292 | 0.208670 | 0.142746 | 0.233358 |
| 293 | 0.308390 | 0.096438 | 0.120273 |
| 297 | 0.143039 | 0.079630 | 0.064009 |
| 298 | 0.069527 | 0.238668 | 0.106325 |

**Figure 3. The functional divergence with 3 clusters**

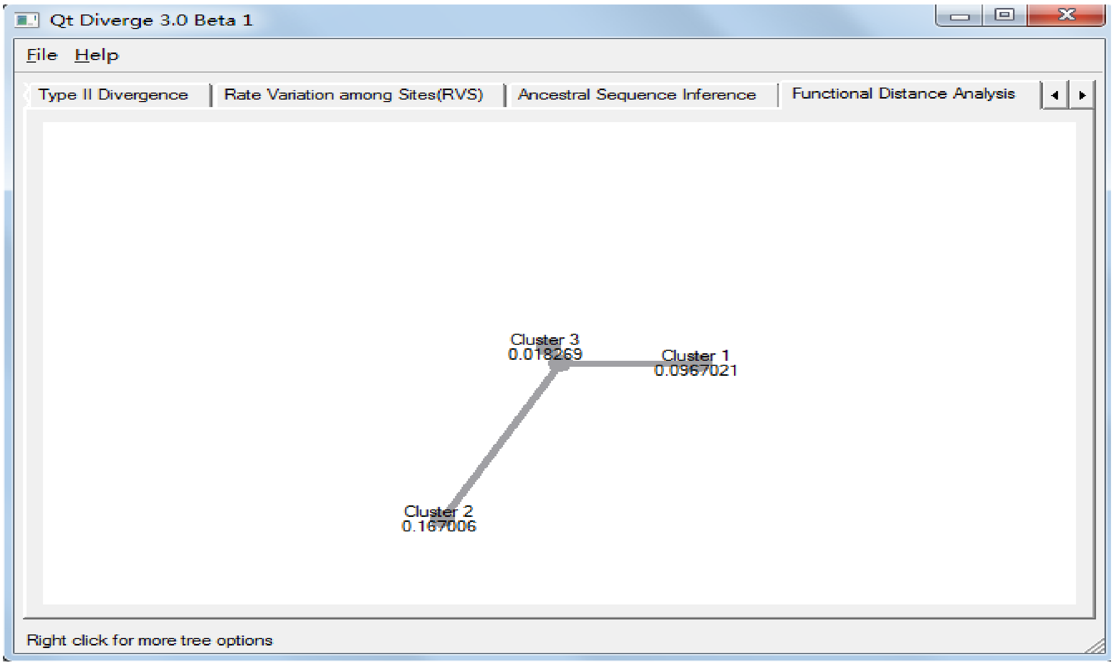
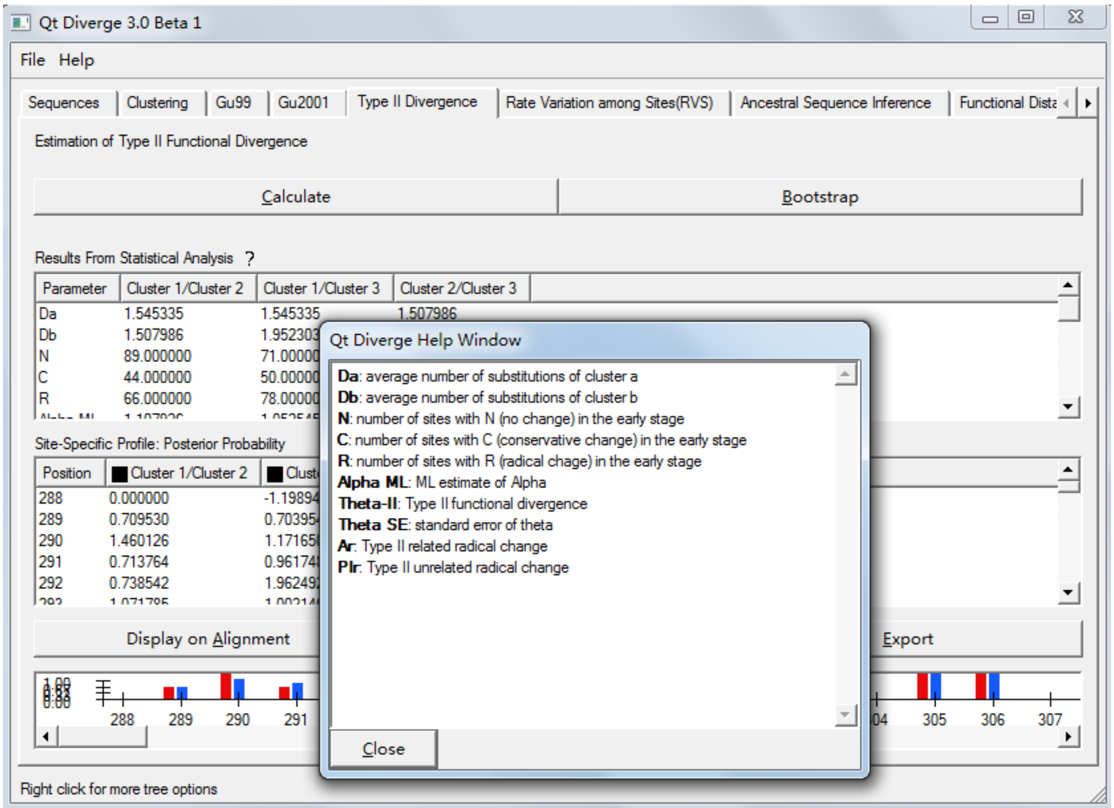

Qt Diverge 3.0 Beta 1
File Help
Type II Divergence
Rate Variation among Sites(RVS)
Ancestral Sequence Inference
Functional Distance Analysis
Cluster 3
0.018269
Cluster 1
0.0967021
Cluster 2
0.167006
Right click for more tree options

**Figure 4. The functional distance between 3 clusters.**

# 4.8 Type II Divergence


Qt Diverge 3.0 Beta 1
File Help
Sequences
Clustering
Gu99
Gu2001
Type II Divergence
Rate Variation among Sites(RVS)
Ancestral Sequence Inference
Functional Dista
Estimation of Type II Functional Divergence
Calculate
Bootstrap
Results From Statistical Analysis ?
Site-Specific Profile: Posterior Probability
Display on Alignment
Export
1.00
0.00
288
289
290
291
04
305
306
307
Qt Diverge Help Window
Close
Right click for more tree options

| Parameter | Cluster 1/Cluster 2 | Cluster 1/Cluster 3 | Cluster 2/Cluster 3 |
|---|---|---|---|
| Da | 1.545335 | 1.545335 | 1.507986 |
| Db | 1.507986 | 1.952303 | |
| N | 89.000000 | 71.00000 | |
| C | 44.000000 | 50.00000 | |
| R | 66.000000 | 78.00000 | |

| Position | ■ Cluster 1/Cluster 2 | ■ Clust |
|---|---|---|
| 288 | 0.000000 | -1.19894 |
| 289 | 0.709530 | 0.70395 |
| 290 | 1.460126 | 1.17165 |
| 291 | 0.713764 | 0.96174 |
| 292 | 0.738542 | 1.96249 |

**Da**: average number of substitutions of cluster a
**Db**: average number of substitutions of cluster b
**N**: number of sites with N (no change) in the early stage
**C**: number of sites with C (conservative change) in the early stage
**R**: number of sites with R (radical chage) in the early stage
**Alpha ML**: ML estimate of Alpha
**Theta-II**: Type II functional divergence
**Theta SE**: standard error of theta
**Ar**: Type II related radical change
**PIr**: Type II unrelated radical change

From the view of molecular evolution, an amino acid residue is said to be functionally or structurally important if it is evolutionarily conserved[19, 20]. Therefore, change of the evolutionary conservation at a particular residue may indicate the involvement of functional divergence[20-22]. Gu (2001) has recognized two types of functional divergence. Type I functional divergence after gene duplication results in altered functional constraints (i.e., different evolutionary rates) between duplicate genes, regardless of the underlying evolutionary mechanisms.Whereas type II results in no altered functional constraints but radical change in amino acid property between them (e.g., charge, hydrophobic, etc.).

We have implemented a statistical method in the software DIVERGE3.0 for detecting type-II functional divergence of gene family. Similarly, the method deals with two duplicate gene clusters, each of which has several (at least four) orthologous sequences. Remember that type-II functional divergence is to statistically test those residues with dramatic amino acid property differences between duplicate genes, but highly conserved within the cluster. The current version of the software tentatively classified twenty amino acids into four groups: charge positive (K, R, H), charge negative (D, E), hydrophilic (S, T, N, Q, C, G, P), and hydrophobic (A, I, L, M, F, W, V, Y). An amino acid substitution is called radical if it changes from one group to another; otherwise it is called conserved. The level of type-II functional divergence is measure by the parameter $\theta_{II}$, called the coefficient of type-II functional divergence; $\theta_{II}=0$ for no type-II functional divergence while $\theta_{II}=1$ for very strong one. Thus, given a sequence alignment, we shall first test whether $\theta_{II} > 0$ significantly. If it is the case, we use the site-specific score based on the posterior ratio to screen amino acid residues related to this type of functional divergence under a cut-off value. Obviously, at the same level of evolutionary conservation within clusters, amino acid residues with radical changes between duplicate clusters will receive a higher score than those with conserved changes in amino acid properties.

Because of the model complexity, at the current version the statistical testing whether $\theta_{II}$ is significantly larger than 0 is based on the estimate (Theta in Table 2) and its standard error (Theta SE). For instance, if one obtains $\theta_{II}=0.2\pm0.05$, calculate the score=0.2/0.05=4 and then obtain the p-value <0.01 from the Z-score test (normal distribution test). It should be noticed that, to our experience, the statistical power of type-II method is relatively lower than that of type-I method, probably due to the fact that less number of amino acid sits are involved in type-II functional divergence between two gene clusters.

Using TypeII is similar with Gu99, too. You need to choose at least two clusters first, and then start the statistical analysis by selecting the "Calculate" button. A progress bar will appear and give the user an estimate of the remaining time required for completion. The calculation can be canceled by pressing the "Cancel" button.

Once the calculations are complete, the user will be presented with the Statistical Results in the upper most regions, listed for each comparison. Hit the “?” sign and a pop up window will show the meaning of these results. Here, Table 3 summarizes the parameters presented in the output of the software for the analysis of type-II functional analysis, as well as their interpretations

**Table 3. Interpretations of parameters presented in the output of the software for the analysis of type-II functional analysis.**

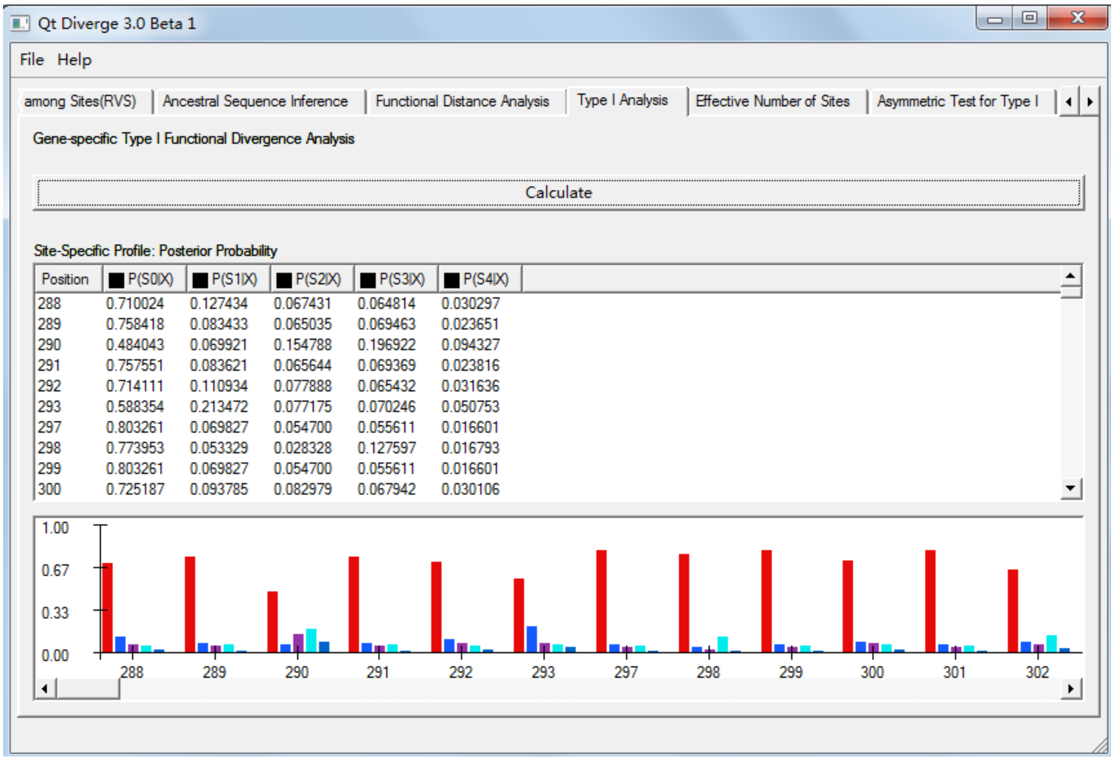
| Parameters | Interpretations |
|---|---|
| Da | Total branch length for gene cluster a |
| Db | Total branch length for gene cluster b |
| N | Number of sites with no change between two clusters |
| C | Number of sites with conserved change between two clusters |
| R | Number of sites with radical change between two clusters |
| p | Proportion of different sites between ancestral nodes of two gene clusters |
| d | Evolutionary distance between ancestral nodes of two gene clusters |
| W | Parameter ad hoc defined in the calculation |
| Z | Parameter ad hoc defined in the calculation |
| Alpha ML | Maximum likelihood estimate of α (the gamma parameter for the among-site rate variation |
| Theta | Estimate of θII by the simplified maximum likelihood method |
| Theta SE | Standard error of the θII estimated by the simplified maximum likelihood method |
| Gc | Proportion of conserved changes |
| Gr | Proportion of radical changes |
| h | Parameter ad hoc defined in the calculation |
| Q | Parameter ad hoc defined in the calculation |
| Ar | Proportion of radical (aR) changes under F2-state (type-II functional divergence) |
| PIr | Proportion of radical (πR) changes under F0-state (no functional divergence) |
| F00,N | Proportion of sites with no change within and between gene clusters |
| F00,R | Proportion of sites with no change within gene clusters, but conserved change between clusters |
| F00,C | Proportion of sites with no change within gene clusters, but radical change between clusters |

The site-specific profile(s) will be presented in the middle region, which is the posterior probability of a site to be functional divergence-related. Moreover, when the cut-off value is given, they can be viewed in relation to the alignment and/or the protein structure. Finally, these results can be exported to a file for processing by other applications. These actions are controlled by use of the buttons above the graph region.

The alignment and structure viewer options require the user select which pairwise comparisons to examine. Bar graphs for site-specific profile will only appear after selection, which is done by clicking the small square in the header of the site-specific profile. As shown in the example "Cluster1/Cluster2" is selected.

The "Display on Alignment" and "Display on Structure" buttons require that at least one column of site-specific profiles be selected and the user should supply a cutoff value. Those sites that have a posterior probability larger than the cutoff value will be highlighted on the alignment or structure respectfully. If no column has been selected a warning message will appear. Also, as with the graph viewer, multiple columns may be selected. When the "Export" button is selected, all the results are exported to a specified file. You could use NotePad or WordPad to open it.

# 4.9 Type I Analysis



Gene-specific Type I Functional Divergence Analysis

Site-Specific Profile: Posterior Probability

| Position | ■ P(S0\|X) | ■ P(S1\|X) | ■ P(S2\|X) | ■ P(S3\|X) | ■ P(S4\|X) |
|---|---|---|---|---|---|
| 288 | 0.710024 | 0.127434 | 0.067431 | 0.064814 | 0.030297 |
| 289 | 0.758418 | 0.083433 | 0.065035 | 0.069463 | 0.023651 |
| 290 | 0.484043 | 0.069921 | 0.154788 | 0.196922 | 0.094327 |
| 291 | 0.757551 | 0.083621 | 0.065644 | 0.069369 | 0.023816 |
| 292 | 0.714111 | 0.110934 | 0.077888 | 0.065432 | 0.031636 |
| 293 | 0.588354 | 0.213472 | 0.077175 | 0.070246 | 0.050753 |
| 297 | 0.803261 | 0.069827 | 0.054700 | 0.055611 | 0.016601 |
| 298 | 0.773953 | 0.053329 | 0.028328 | 0.127597 | 0.016793 |
| 299 | 0.803261 | 0.069827 | 0.054700 | 0.055611 | 0.016601 |
| 300 | 0.725187 | 0.093785 | 0.082979 | 0.067942 | 0.030106 |

DIVERGE provided a site-specific profile based on posterior scores to predict the sites responsible for type I functional divergence.

The updated DIVERGE 3.0 implements a new method that helps the user infer type I functional divergence specific to a given duplicate cluster.

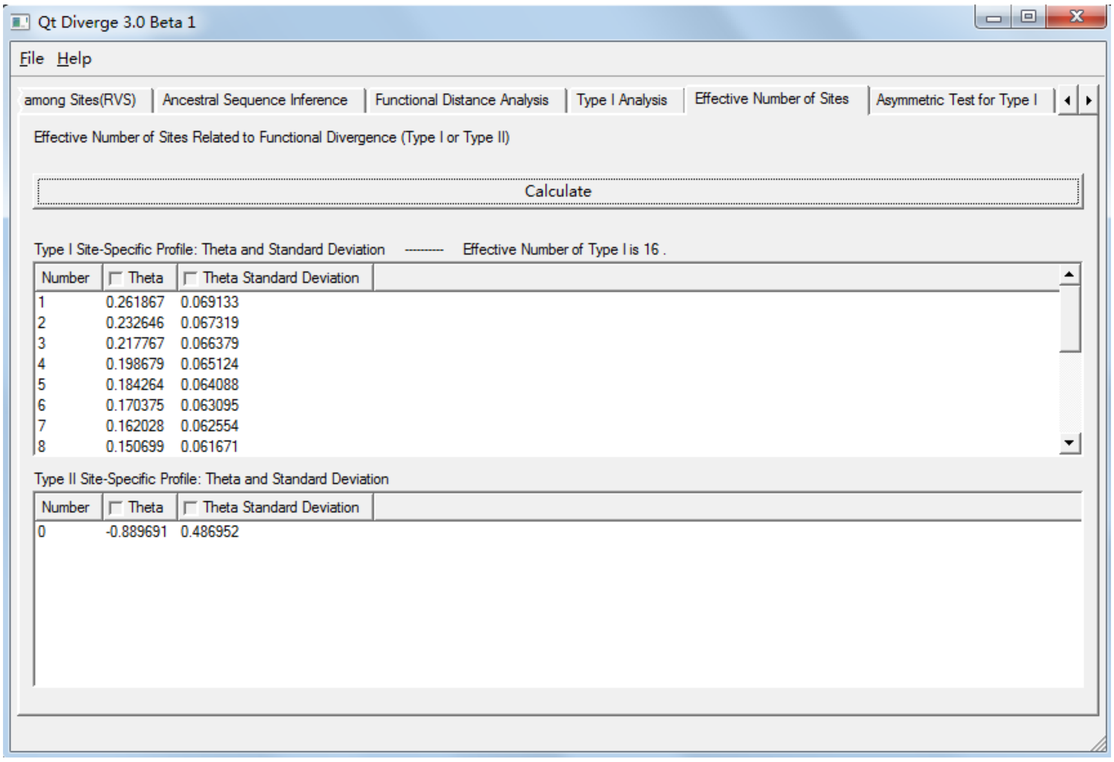
After complete the steps like the multiple alignments, and the tree, selecting three duplicate clusters simultaneously is needed, and then the user can start the statistical analysis by selecting the “Calculation” button. A progress bar will appear and give the user an estimate of the remaining time required for completion. The calculation can be canceled by pressing the "Cancel" button. Once the calculations are complete, the user will be presented with the statistical results, which is the posterior probability of the sites to be functional divergence-related for five possible patterns.

The results contains two parts:

The first part is the site-specific profile with posterior probability. There are six columns in this part. The first column is the site position, the following five columns represent five nondegenerate patterns. Under the two-state model(functional divergence unrelated F0 or related F1), there are eight possible combined states for three duplicate clusters, which can be reduced to five nondegenerate patterns. S0=(F0, F0, F0) means no type I divergence occurred in any clusters. S1=(F1, F0, F0) means type I functional divergence occurred only in cluster 1, and similarly S2 =(F0, F1,F0) and S3 =(F0, F0, F1). The final pattern S4 is for the rest of four states, each of which has two or three clusters that have experienced type I functional divergence.

Choosing the pattern you want to know, it can be viewed in relation to the alignment automatically.

# 4.10 Effective Number of Sites


Qt Diverge 3.0 Beta 1
File Help
among Sites(RVS)
Ancestral Sequence Inference
Functional Distance Analysis
Type I Analysis
Effective Number of Sites
Asymmetric Test for Type I
Calculate

Effective Number of Sites Related to Functional Divergence (Type I or Type II)

Type I Site-Specific Profile: Theta and Standard Deviation --------- Effective Number of Type I is 16 .

| Number | Theta | Theta Standard Deviation |
|---|---|---|
| 1 | 0.261867 | 0.069133 |
| 2 | 0.232646 | 0.067319 |
| 3 | 0.217767 | 0.066379 |
| 4 | 0.198679 | 0.065124 |
| 5 | 0.184264 | 0.064088 |
| 6 | 0.170375 | 0.063095 |
| 7 | 0.162028 | 0.062554 |
| 8 | 0.150699 | 0.061671 |

Type II Site-Specific Profile: Theta and Standard Deviation

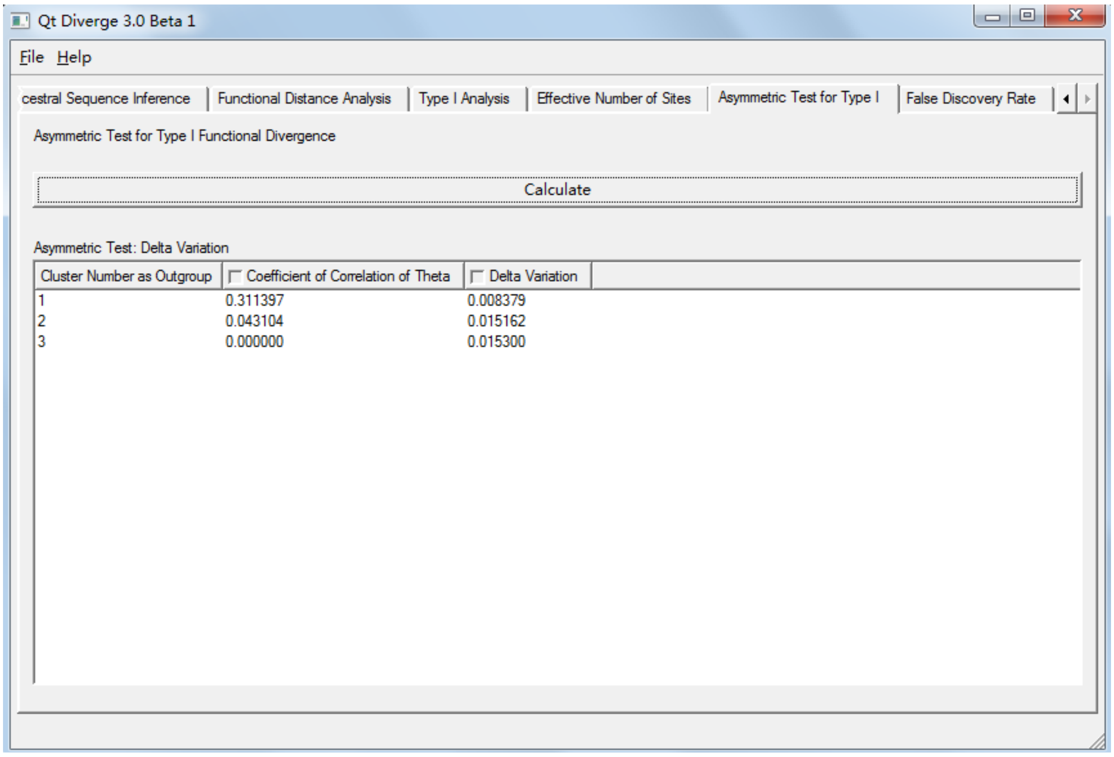
| Number | Theta | Theta Standard Deviation |
|---|---|---|
| 0 | -0.889691 | 0.486952 |

Even though most studies pointed to a small number of sites that can be predicted as type I or type II functional divergence-related, calculation of the average percentage of amino acid sites involved is problematic. From our preliminary analysis, we notice that, after removing those predicted sites with the strongest signals, the functional divergence between duplicate genes for the rest of amino acid sites usually becomes trivial. On the basis of this observation, we have designed a rapid nonparametric procedure to count the effective number of functional divergence-related sites.

This new function estimates effectively the number of sites related to type I and type II functional divergences , which is insensitive to the cutoff. The effective number (ne) of functional divergence-related sites(F sites) is defined as the minimum number of sites, such that, when they are removed, the coefficient of functional divergence for the rest of sites approaches to zero. For the detailed information about the algorithmis, please refer to Gu X.*, Zou Y., Su Z., Huang W., Zhou Z., Arendsee A. and Zeng Y. (2013) An Update of DIVERGE Software for Functional Divergence Analysis of Protein Family.

Using Effective Number of Sites is very straightforward. When the steps like the multiple alignments, the tree, and the cluster selection have been completed, the user can start this statistical analysis by selecting the "Calculation" button. Then effective number of sites with type I and type II results will be calculated and displayed respectively, and the results also show the theta and standard deviation of the sites.

## 4.11 Asymmetric Test for Type I

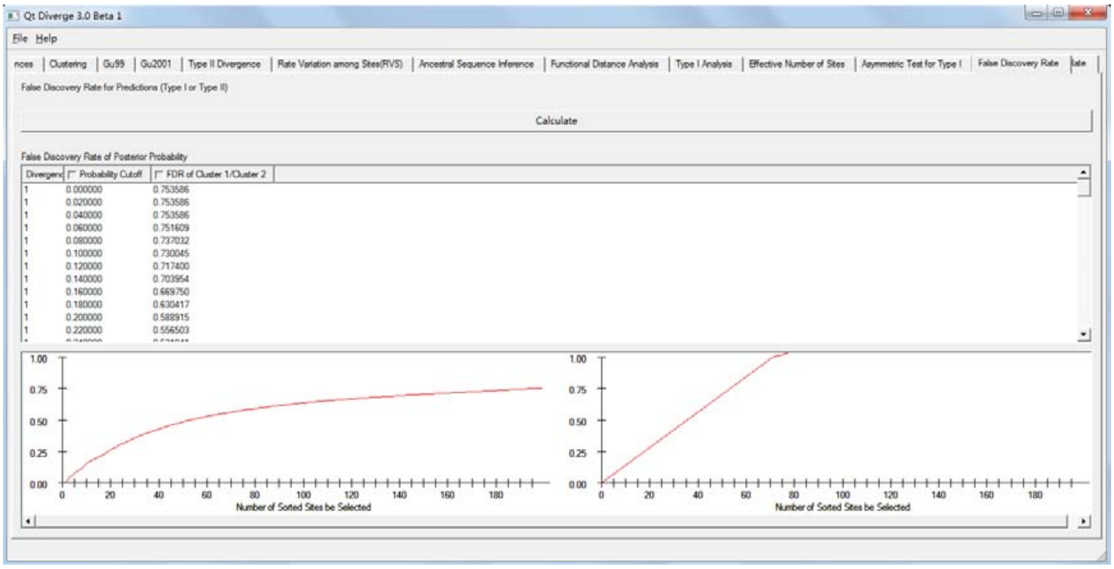

Qt Diverge 3.0 Beta 1
File Help
cestral Sequence Inference
Functional Distance Analysis
Type I Analysis
Effective Number of Sites
Asymmetric Test for Type I
False Discovery Rate
Asymmetric Test for Type I Functional Divergence
Calculate
Asymmetric Test: Delta Variation

| Cluster Number as Outgroup | □ Coefficient of Correlation of Theta | □ Delta Variation |
|---|---|---|
| 1 | 0.311397 | 0.008379 |
| 2 | 0.043104 | 0.015162 |
| 3 | 0.000000 | 0.015300 |

DIVERGE2 implemented functional distance analysis to demonstrate the asymmetry of type I functional divergence but lacked a rigorous statistical basis. Here, the newest version solve this problem by implementing a simple method as follows. Suppose we test whether type I functional divergence is asymmetric between duplicate clusters 1 and 2, given a more ancient duplicate cluster 3 as outgroup. Let $\theta_{12}$, $\theta_{13}$, and$\theta_{23}$ be the coefficients of type I functional divergence between pair-wise duplicate clusters. Under the hypothesis of symmetry between duplicate clusters 1 and 2, we have the null $\theta_{12}=\theta_{13}$ and develop an approximate method to calculate the sampling variance of $\delta=\theta_{13}-\theta_{23}$, for testing whether the null hypothesis $\delta=0$ can be statistically rejected.

Using this function, the user need to create three clusters first, then selecting the "Calculate" button. After the progress finished, the results will show you the coefficency of correlation of $\theta$ and the $\delta$ variation based on different outgroup cluster.

# 4.12 False Discovery Rate



False Discovery Rate for Predictions (Type I or Type II)

False Discovery Rate of Posterior Probability

| Divergenc | Probability Cutoff | FDR of Cluster 1/Cluster 2 |
|---|---|---|
| 1 | 0.000000 | 0.753586 |
| 1 | 0.020000 | 0.753586 |
| 1 | 0.040000 | 0.753586 |
| 1 | 0.060000 | 0.751609 |
| 1 | 0.080000 | 0.737032 |
| 1 | 0.100000 | 0.730045 |
| 1 | 0.120000 | 0.717400 |
| 1 | 0.140000 | 0.703954 |
| 1 | 0.160000 | 0.669750 |
| 1 | 0.180000 | 0.630417 |
| 1 | 0.200000 | 0.588915 |
| 1 | 0.220000 | 0.556503 |

FDR for predictions provides more statistical evaluations for predicted sites. Knowing the false discovery rate (FDR) of the predicted sites is critical to assessing the reliability of the results. In general, FDR is the proportion of predicted sites that are actually unrelated to functional divergence. DIVERGE and DIVERGE2.0 mainly use a site-specific posterior profile, denoted by Qk for site k, as a scoring system to identify functional divergence-related amino acids[23-25].

In the updated version of DIVERGE3.0, FDR was calculated in the following procedure. Let Lc be the number of sites predicted under the posterior cutoff c. Then, we have shown (Gu 2011) that FDR(c) can be approximately calculated by

$$\text{FDR(c)=1-} \sum_{k\ in\ A} \text{Qk}/\text{Lc}\ \ ,$$

where set A is for all sites k that satisfy Qk>c. This value may help to evaluate the cost of experiments caused by false positive predictions.

The use of this new function is also simple and straightforward. After finish the steps like the multiple alignments, and the tree, pressing the button “Calculate” is OK. The lower part of the panel will automatically show you the plot in relation to the number of sorted site with the probibity of the cutoff. With the increase of the number of the sorted sites, the probibility is tend to be paralelle. In this condition, the user can choose a better cutoff.

# 5 Appendix

## 5.1 Case study: COX (cyclooxygenase) gene family

### 5.1.1 Type I and Type-II functional divergence analysis

The cyclooxygenase (COX) enzymes catalyze a key step in the conversion of arachidonate to PGH2, the immediate substrate for a series of cell prostaglandin and thromboxane synthases. Prostaglandins play critical roles in numerous biological processes, including the regulation of immune function, kidney development, reproductive biology, and gastrointestinal integrity. There are two tissue-specific isoforms in mammals: COX-1 and COX-2. Fig.5 shows the phylogenetic tree of COX gene family, inferred by the neighbor-joining method. It is clear that these two isoforms were generated in the early stage of vertebrates.

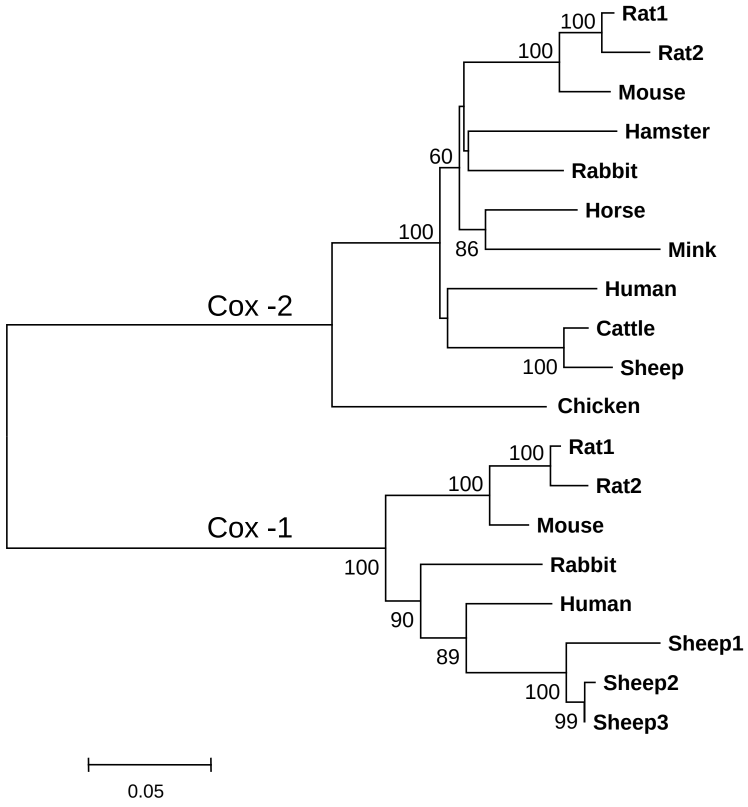


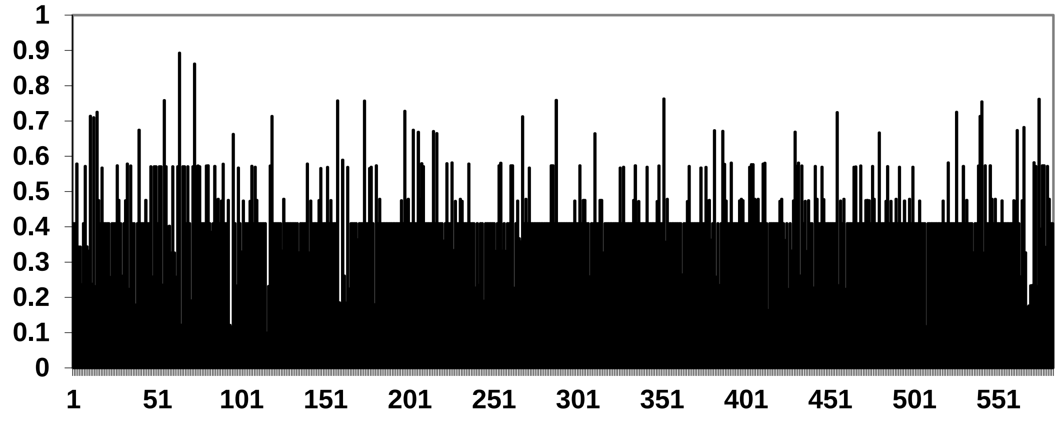
**Fig 5. phylogenetic tree of COX gene family**

Table 4 summarizes the estimates of the coefficients of type-I and type-II functional divergences by various methods implemented in DIVERGE3.0. It appears that the estimate of $\theta_I$ is virtually the same. Indeed, after analyzing numerous cases we have found that this claim is largely correct, as long as the number of sequences in each gene cluster is sufficiently large. That is, as the statistical and computational properties of Gu-1999 ML method is, overall, superior to the other two methods, which is usually recommend.

**Table 4. The estimates of the coefficients of type-I and type-II functional divergences by the methods implemented in DIVERGE3.0**

| Types of functional divergence (method) | Estimate |
|---|---|
| Type-I (Gu-1999, MFE) | $\theta_I = 0.49\pm0.13$ |
| Type-I (Gu-1999, ML) | $\theta_I = 0.46\pm0.09$ |
| Type-I (Gu-2001, ML) | $\theta_I = 0.44\pm0.09$ |
| Type-II (Gu-2006) | $\theta_{II} =0.16\pm0.04$ |

Fig.6 shows the site-specific profiles for type-I and type-II functional divergences, respectively. Not surprisingly, the correlation of site-specific profiles for type-I functional divergence between Gu (1999) and Gu (2001) methods are almost the same ($R^2$=0.96). Given a cut-off, one may select a set of predicted sites for type-I and type-II functional divergences, respectively for the further analysis or experimental verification.



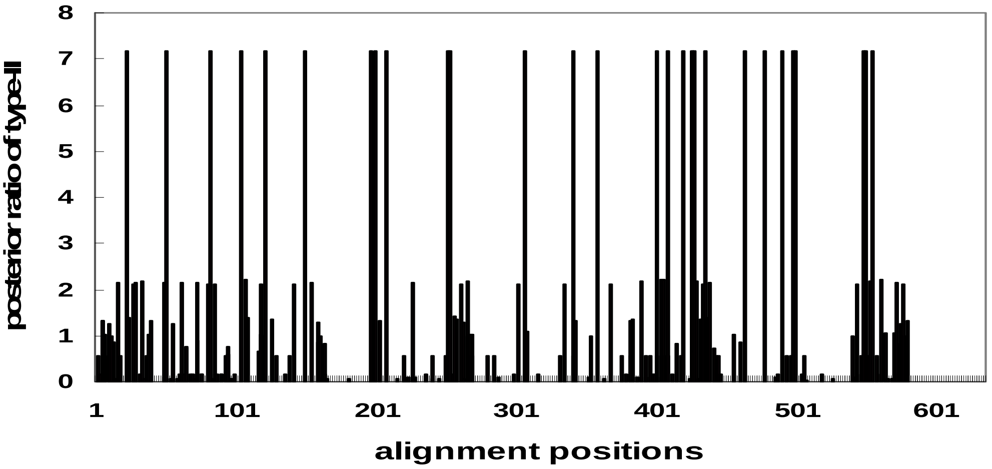


**Fig.6 Site-specific profile for type-I and type-II functional divergence between COX-1 and COX-2 respectively, measured by the posterior ratio.**

Our final comment on the analysis of type-II functional divergence is about the effect of radical amino acid substitutions in the early stage of duplication. For the COX gene family, we found radical substitutions for type II functional divergence in the early stage is about 2.7-fold increasing (aR/πR=2.7). Consequently, an amino acid residue with a radical change between COX1 and COX2 may have a higher score than a conserved change for being type-II functional divergence-related. As shown by many authors, the sites most likely exhibiting type-II behavior are the radical cluster-specific sites, while the conserved cluster-specific sites are less likely, as indicated by a low posterior probability. This case-study clearly shows the important role of statistical analysis, otherwise one cannot objectively justify whether one-less radical cluster-specific sites (i.e., there is one amino acid substitution in the late stage) is more likely to be functional divergence-related than conserved cluster-specific sites.

## 5.1.2 FDR for Predicted Type I Functional Divergence Related Sites

Given the gene phylogeny and 584 aligned amino acid sites, we obtained θI=0.56± 0.11 between COX-1 and COX-2 (Gu 2001a). For the top five predicted sites under the posterior cutoff 0.80, we calculated the corresponding FDR as 11.0%. If we lower the posterior cutoff to 0.7, the FDR value for the 19 predicted sites is 20.1%. This example shows that in practice, one can also use FDR as a primary criterion to make functional predictions.

## 5.1.3 Effective Number of Functional Divergence-Related Sites

As expected, the θ* value decreases rapidly when more sites are removed and is nearly zero when the top 30 sites are removed (roughly at the posterior probability of 0.63). Figure 7 shows the θ*-RemovedSite profile of type I functional divergence between COX1 and COX2. Because this profile has a long tail around zero, we recommend the use of one standard error of θ* to control the long tail problem; in this case, we obtain ne = 27. We thus conclude that about 4.6% of amino acid sites may have been involved in Site-Specific Rate Change between COX1 and COX2.

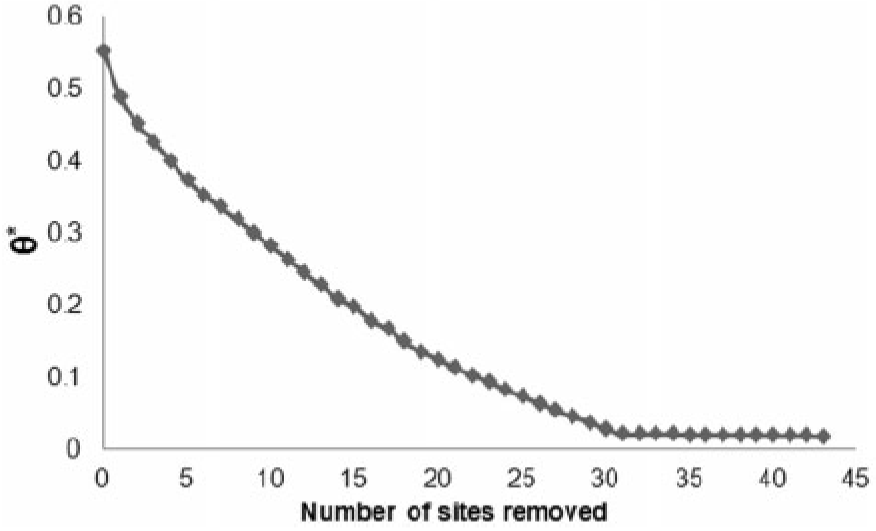


**Fig.7 The θ*-RemovedSite profile of type I functional divergence between COX1 and COX2**

### 5.1.4 Site-specific Rate Change between COX1 and COX2

We calculated the site-specific ω= dN/dS ratio in COX1 and COX2 duplicate clusters (the mean ω= dN/dS for COX1 is 0.076/0.707 = 0.108 and that for COX2 is 0.076/0.551 = 0.137). The change in site-specific rate ranges approximately from 4- to 8-fold at predicted type I functional divergence- related sites, as shown in Fig.8.

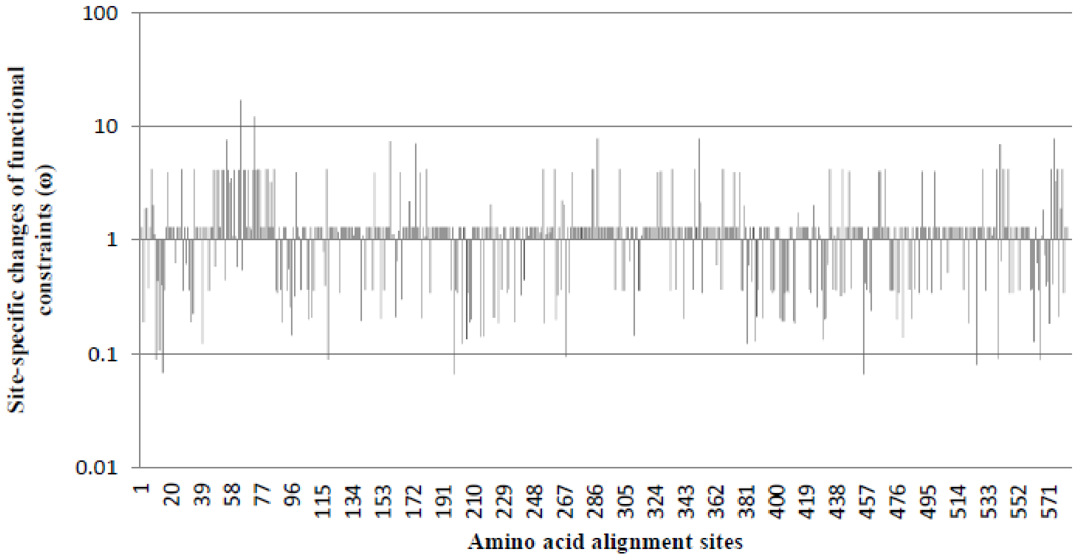


**Fig.8 Site-specific changes of functional constraints between COX-1 and COX-2**

## 5.2 DIVERGE: Frequently Asked Questions

**1.**
**Input data:** Hox11.fasta or test.fasta

**Error Message:**
Using diverge (for windows) to test for functional divergence in a gene family. When hit the Gu99 tab, however, get the following error message:
*The instruction at "0x00413521" referenced memory at "0x00cbace8". The memory could not be read* Or *in Gu99 tab, I always failed calculation with a comment of application error.*
**Reason:**
The sequence data for the cluster you choose are exactly the same or too similar. So it can be almost treated as 1 gene and the variation rate nearly equals to zero. It will

cause no significant statistics meaning. To test this suspicion, you can using GZ97 method to test your clusters separately, and found that which one will cause error since GZ97 is the preliminary condition for GU99 method.

**Suggestions:**

Avoid using sequence data which is too similar with each other

**2.**

**Input data:** BOLA-ALL-DOMAIN.fas bolA-all-7-26-2.fas

**Error Message:**

When using DIVERGE to analysis the functional divergence of 7 group from a family, I have got the export file(in the annex),but there are some conflict in the data such as the ThetaML value >1 or the SE Theta value ThetaML value. How to explain these phenomena?

One file is for full length sequence, the other is for sequence containing only domain. By the way, the family consists of 142 sequences, which come from 2 paralogues in E.coli. How to distinguish their offspring?

**Reason:**

Theoretically, the ThetaML value is between 0 and 1.But it's possible to get the ThetaML value>1 using real data. After studying your export data, I found that the value is almost 1 and 0.9 which means a great chance for divergence and the sequence length is not long.

**Suggestions:**

It is possible that the SE Theta value > ThetaML value, the large variance just says that the estimation is not so accurate, it is possible that the standard error is large than the mean value.

The large variance here can be the result of small sample, if disregarding the method. In this case, the sequence length (in your result only 24 sites counted )and the number of sequences in one cluster is small, which can introduce large variances in the estimation.

For theata value large than 1:

The meaning of theata value is the proportion of sites expected to be functional-divergence-related (F1 state). Of course theata value is between 0 and 1, however, we use the statistical method to estimate the theata, the estimate is not the "true value".

In the model free method, if estimated Theata > 1 , the estimated covariance between X1, X2 ( number of changes in a site ), sigma12 may be negative.

Large theata means that the proportion of sites expected to be functional-divergence-related is large.

In your result, the posterior probability of F1 state of each site is almost 1 in those cases whose theata value is very large. No conflict.

And I went through your alignment, for example, to see this pair of clusters :CEI1/P43781:

I only draw this 8 sequences (domain)from the total data set, and tested in DIVERGE. Actually, I got this result:

```
==============================================================
==========================================
Main
Tree:
((((Q9RHA0_E.coli:0.000000,P43781_E.coli:0.000000):0.011905,Q8XFL8_S.typhi:
0.011905):0.163690,NP_719479_S.oneidensis:0.133929):0.241071,(CAF99621_T.ni
groviridis:0.101190,((Q8CEI1_M.musculus:0.001488,XP_371502_H.sapiens:0.02232
1):0.016369,XP_216181_R.norvegicus:0.031250):0.075892):0.241071):0.000000
Cluster Cluster
2:
((Q9RHA0_E.coli,P43781_E.coli),Q8XFL8_S.typhi,NP_719479_S.oneidensis)
Cluster Cluster
1:
((Q8CEI1_M.musculus,XP_371502_H.sapiens),CAF99621_T.nigroviridis,XP_21618
1_R.norvegicus)

          Cluster 2/Cluster 1
MFE Theta          1.367318
MFE se          0.324612
MFE r X           -0.175607
MFE r max          0.478080
MFE z score           -4.358623
ThetaML           0.999200
AlphaML           0.306165
SE Theta          0.188849
LRT Theta          27.994573
54        0.999040
55        0.999615
56        0.999615
57        0.999782
58        0.999782
59        0.999040
60        0.999040
61        0.999040
62        0.999782
63        0.999714
64        0.999040
65        0.999040
66        0.999964
67        0.999040
68        0.999040
69        0.999601
```

```
70          0.999782
71          0.999040
72          0.999615
73          0.999040
74          0.999040
75          0.999782
76          0.999615
77          0.999040
78          0.999615
79          0.999040
80          0.999040
81          0.999040
82          0.999615
83          0.999040
84          0.999040
85          0.999782
87          0.998101
92          0.999615
93          0.999040
94          0.999040
95          0.999615
96          0.999615
97          0.999040
98          0.999040
100          0.997416
101          0.999615
=============================================================
=============
```

Just as your result, but more sites are counted here.

The method use complete deletion of the gaps, so when the number of your alignment sequences is large, the total number of non-gap sites may be small. If the total number of non-gap sites is small, the variance would be large and the estimation is not accurate.

And the method to estimate the X1, X2(number of changes in a site in one cluster) uses the tree branch length information. The branch length information of the total tree and the tree of only these sequences is not exactly the same.
SO our result is not the same but should be similar.

The posterior probability of F1 state of each site is almost 1. And the ThetaML is 0.999200, it seems to be a quite good result. Note that the

```
======================
MFE Theta 1.367318
MFE se          0.324612
MFE r X           -0.175607
MFE r max             0.478080
MFE z score              -4.358623
==========================
```

are based on the simple "model-free" Method in Gu99 MBEpaper( Gu , Mol. Biol.
Evol.
16(12):1664-1674. 1999).
And the MFE (model free estimation)theata is just similar as your result.
(Generally, the MFE theata is larger than ML ( Maximum likelihood method )

And I am afraid that your alignment is not so good.
The same example:

```
-TEGELKVTQVLKEKFP--RATAIQVTDIS---------GGCG---------AMYEIKIESEEF
KEKRTVQQHQMVNQALKEEIK-G----MHGLRIFTSVPKC-------
------RVTQILKEKFP--RATAIKVTDIS---------GGCG---------AMYEIKIESEEFKEK
RTVQQHQMVNQALKEEIK-E----MHGLRIFTSVPKR-------
-TEGELKVTQVLKEKFP--RATAIQVTDIS---------GGCG---------AMYEIKIESEEF
KAKRMVQQHQMVNQALKEEIK-G----MHGLRIFTSVPKC-------
----------------------------------------------------MYEVHIESMEFKGKRTIQQHQLVNQ
ALKEEIQ-G----MHGLRIFTNV----------

--MENNEIQSVLMNAL---SLQEVHVSGD-----------G-----------SHFQVIAVGELFD
GMSRVKKQQTVYGPLMEYIADN---RIHAVSIK-AYTPAEW-----
--MENHEIQSVLMNAL---SLQEVHVSGD-----------G-----------SHFQVIAVGEMFD
GMSRVKKQQTVYGPLMEYIADN---RIHAVSIK-AYTPAEW-----
--MENNEIQSVLMNAL---SLQEVHVSGD-----------G-----------SHFQVIAVGELFD
GMSRVKKQQTVYGPLMEYIADN---RIHAVSIK-AYTPAEWARDRK
--MECSLIEQILRDAL---ALDEVHASSD-----------G-----------SHYKVIAVGECFDG
MSRVKQQQTIYAPLMSYIASG---ELHALTIK-TFTPTQW-----
```

The alignment in each cluster is good .
But between the two clusters, it seems there is no apparent common
sites( domain?,motif?) and it seem the two cluster are not closely related. So it is not surprise that the theata value is large, says,
in my ML estimation, 0.999200.

And I also checked your last pair of cluster which has a very large theata estimation value.( only use sequence in two clusters )

```
=============================
          Cluster 6/Cluster 7
```

```
MFE Theta            1.263656
MFE se             1.039368
MFE r X             -0.033859
MFE r max             0.128422
MFE z score             -1.219831
ThetaML             0.999200
AlphaML              1.272727
SE Theta            0.358330
LRT Theta             7.775700
11            0.999273
12            0.999171
13            0.998625
14            0.998474
15            0.999433
16            0.999433
21            0.999342
22            0.999302
23            0.998198
24            0.999342
25            0.999424
26            0.999171
27            0.999171
28            0.999302
29            0.999290
53            0.999638
54            0.999069
55            0.999171
56            0.999290
57            0.999342
58            0.999433
59            0.999171
60            0.999231
61            0.999069
62            0.999400
63            0.999236
64            0.999069
65            0.999433
66            0.999069
67            0.999069
68            0.999171
69            0.999151
70            0.999226
71            0.999236
72            0.999623
```

```
73          0.999069
74          0.999723
75          0.999069
76          0.999148
77          0.999171
78          0.999079
79          0.999069
80          0.999723
81          0.999217
82          0.999433
83          0.999280
84          0.999623
85          0.999342
87          0.999171
92          0.999171
93          0.999069
94          0.999069
95          0.999069
96          0.999069
97          0.999069
100           0.999069
101           0.999433
102           0.999171
103           0.999069
==============================
```

The bad thing is I can not actually redo your work since I don' t know how to reroot the tree to get the 7 clusters selected.

**3.**

**Questions:**

1. Do you have a linux version of Diverge?
2. The linux version does not run returning the following:
/diverge: relocation error: ./diverge: undefined symbol: __ti9QGLWidget
Is that a QT related error?
By the way, We are using intel-Pentium4-based machines running Debian Linux.
Specifically: Linux version 2.4.23 (root@ella) (gcc version 3.3.2 (Debian))

**Suggestions:**

1. We do have a Linux version of diverge, which is available from http://xgu.zool.iastate.edu (This copy requires QT library.) Another suggestion is to use Windows version, which is much easier to use.

2. Yes, this is the problem related to QT (GLWidget is a class defined in QT). One suggestion is to download and install QT library for you Debian Linux (check

http://packages.debian.org/stable/devel/libqt-dev). Since we only have RedHat Linux available, it is a little bit difficult for us to figure out and fix your problem.

## 4.
## Questions & Suggestions:

I am using your Diverge v1.04 for functional-structural divergence analysis. However, several trials so far have encountered various problems.
The Caspase example runs smoothly on two different computers (all NT2000 platform).

I can load the alignment (in fasta format) and the structure file (pdb format). Various problems have been observed afterward.
a.. unable to generate tree using the NJ-Tree making function. It seemed it is related to the number of sequence. For an alignment with 240 sequences (600 amino acids in length; identity ranges from 70 % to 97 %), the program just hung up and had to be terminated from "task manager".

To tell the truth, we have not tested our program with such a large data input. With so many sequences, the NJ-Tree making algorithm might have some problem. If you don't mind, could you give us your data set for testing? That will help us to spot the bug right away and solve your problem. We will also try to use some larger data sets by ourselves.

a.. when the number of sequences reduced to half, Tree can be generated within 1-2 seconds. After the assignment of clusters, calculation proceeded to 100 % (then hold for up to several minutes) then either with the calculated results shown or the program crashed. (with error messages such as: the memory could not be read; "The instruction at "0x77fcca14" referenced memory at "0x3fb7421d". The memory could not be "written")

To me, this bug seems to be data dependent. I don't know whether it is related to your input data size. Again, if you can provide some data for testing, that will be great.

Obviously this program is calculation-intensive. My computer is Pentium 3, 800MHz with 512 MB RAM.

I thought it is good enough (as the Caspase example can be done within 2 minutes).

Here are some specific questions:

1.. how do I know when to kill the program and start all over? (something like "hang up for 2 minutes for 50 sequences (500 aa in length) and 3 clusters")

Whenever you need to kill the program manully, that means it is a bug in our program. We would like to hear about such problems. The solution to your problem is to enhance our diverge by showing the progress of tree constrution. That will help you to know the on-going progress. We will incoporate this enhancement into next release of diverge.

2.. if multiple trials were not successful, what is the best way to proceed: reduce the number of sequences? truncate the protein size? reduce the number of clusters?

Generally, you can figure out what is causing the problem based on the stages. For example, if you can not create the NJ-Tree correctly, that means the number of sequences might be too large and you can try to reduce the size. If tree generation is working fine and you can not do calculation using Gu99 method correctly, the cluster number might be the cause of problem. Therefore, you should try to reduce its size. No matter what happens, you can contact with us to discuss about your problem and we would appreciate your inputs.

3.. the structure viewer is very useful. However, I have difficulties in using the following functions "zoom", there is minimum differences between wireframe and solid functions. And I don't know how to use "Pick" function (suppose to identify the residue number, right?)

I am not very sure about your "zoom" question. You can just use the middle button of your mouse to enlarge the structure view. If you only have two buttons on mouse, you should press both of them at the same time. "Pick" function is easy to use. First, use your mouse to click on the spot of PDB structure. After that, press "P". That is it. If not working, try to press P for a few times to see what happens.

It seems Diverge is a nice program for statistical analysis of sequence-function relationship. I would really like to make it work in my hand. Your suggestions are very important and greatly appreciated.

## 5.
## Questions & Suggestions:

I am currently working on analysis of one of the large bacterial families for identifying possible sub-families by looking at both amino acid residue properties as well as selection constraints which could be an influencing factor for functional specificities. I recently downloaded the DIVERGE software for studying the functional divergence due to selection constraints. I had a few questions pertaining to the tool.
a) As a part of my input to the graphical user interface, what would be the format for those "pre-defined clusters" which is a pre-requisite for the program. This brings me to the next question on how and where to add

the clusters in the GUI. The "Add cluster" option is never an active
button, so I am unable to click on it in the version I have on my
computer.

There is not "pre-defined clusters" in DIVERGE. You select your own clusters in the generated tree. What you need to do is to (using your mouse) click on the root of clusters and DIVERGE will highlight it for you. At the same time, the "Add cluster" button will be enabled.

b)Do I necessarily mention any known structure as a part of my input, or
is this optional?

Structure is optional. The purpose is to help you to identify the residues on the PDB structure view.

c) In the latest version (which I downloaded), the system freezes when I
opt for NJ tree making. I was wondering if this is a known bug or it
could be just specific to our computers here. I tested it both on Windows (XP) and Linux, but the problem persists.

Could you please give us more details about the problems? We would like to reproduce your problem. If you can provide us with your test data sets, that will be better.

d) Is the multiple alignment input which is need, an alignment of the
whole family including the pre-defined sub-families (clusters) or is it
an alignment of a cluster?
It is the whole family including the sub-families (clusters).

**6.**

**<u>Questions:</u>**

My name is Javier Forment, and I'm working at the "Institute for Plant

Molecular and Cell Biology (Polytechnic University of Valencia, Spain)". We are

very interested in using your Diverge_1.04 software in a Linux machine with SuSE

The problem is that you provide the binaries for Linux and they are compiled against

the qt2 library, but our machine has the qt3 library and the program gives an error.

I've also tried to install the qt2 library, but other errors arise.Could it be possible to get the sources to compile in our machine?

**Reason:**

Diverge on our website was not compiled with latest Qt library. The easiest solution is to give you a Diverge compiled with Qt 3.

**Suggestions:**

I have created a statically linked version of diverge. It should be free from any Qt libraries related issues. You can download it from

http://xgu.zool.iastate.edu/software/diverge/download/diverge-1.2

## 5.3 Source Code Notes

### Flow:

Main.cpp – MainWindow.cpp – GuWidget.cpp – sequences_tab.cpp – clustering_tab.cpp – gu99_method_tab.cpp – rvs_tab.cpp – ancestral_seq_tab.cpp – func_dist_tab.cpp – gu2001_method_tab.cpp-TypeII_method_tab.cpp

### Details for Each File:

(1) Main.cpp: Main program is the first program to see. It sets the main widget and a popup window.

(2) MainWindow.cpp: Menu Bar popup window items.

(3) GuWidget.cpp: include sequences_tab, clustering_tab, gu99_tab, rvs_tab, ancestral_seq_tab, func_dist_tab, gu2001_tab, typeII_tab and their relative signals.

(4) Sequences_tab.cpp: Design the sequences_tab interface and define two functions:

load_alignment, load_pdb.

(5) Clustering_tab.cpp: Design the clustering_tab interface and define functions like load_tree, nj_cluster, add_cluster and delete_cluster.

(6) Gu99_method_tab.cpp: Realize two methods, which are the "calculate" and "bootstrap". The difference between them is "calculate" is tree_fixed, "bootstrap" is tree_changed(effected by tree).

(7) RVS_tab.cpp: Realize two methods, which are the "calculate" and "bootstrap".

(8) Ancestral_seq_tab.cpp: Load_tree and NJ Tree_making

(9) Func_dist_tab.cpp: Compute Functional Distance Between Clusters.

(10) Gu2001_method_tab.cpp: Similar with gu99_method_tab.cpp, but the difference is Gu2001 method is used.

# 6. Reference

1. Ohno, S., *Evolution by gene duplication*. 1970, Berlin: Springer-Verlag.
2. Wolfe, K.H. and D.C. Shields, *Molecular evidence for an ancient duplication of the entire yeast genome.* Nature, 1997. **387**(6634): p. 708-13.
3. Wang, Y. and X. Gu, *Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction.* Genetics, 2001. **158**(3): p. 1311-20.
4. Huang, Y.F. and G.B. Golding, *Inferring sequence regions under functional divergence in duplicate genes.* Bioinformatics, 2011. **28**(2): p. 176-83.
5. Su, Z. and X. Gu, *Revisit on the evolutionary relationship between alternative splicing and gene duplication.* Gene, 2012. **504**(1): p. 102-6.
6. Tatusov, R.L., E.V. Koonin, and D.J. Lipman, *A genomic perspective on protein families.* Science, 1997. **278**(5338): p. 631-7.
7. Zheng, Y., D. Xu, and X. Gu, *Functional divergence after gene duplication and sequence-structure relationship: a case study of G-protein alpha subunits.* J Exp Zool B Mol Dev Evol., 2007. **308**(1): p. 85-96.
8. Abhiman, S. and E.L. Sonnhammer, *Large-scale prediction of function shift in protein families with a focus on enzymatic function.* Proteins, 2005. **60**(4): p. 758-68.
9. Benitez-Paez, A., S. Cardenas-Brito, and A.J. Gutierrez, *A practical guide for the computational selection of residues to be experimentally characterized in protein families.* Brief Bioinform., 2012. **13**(3): p. 329-36.
10. Gu, X., *Statistical methods for testing functional divergence after gene duplication.* Mol Biol Evol., 1999. **16**(12): p. 1664-74.
11. Gu, X., *Maximum-likelihood approach for gene family evolution under functional divergence.* Mol Biol Evol., 2001. **18**(4): p. 453-64.
12. Capra, J.A. and M. Singh, *Characterization and prediction of residues determining protein functional specificity.* Bioinformatics, 2008. **24**(13): p. 1473-80.
13. Gu, X. and K. Vander Velden, *DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family.* Bioinformatics, 2002. **18**(3): p. 500-1.
14. Gu, X., et al., *An Update of DIVERGE Software for Functional Divergence Analysis of Protein Family.* Molecular Biology and Evolution, 2013.
15. Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* 1987(0737-4038 (Print)).
16. Lopez, G., A. Valencia, and M. Tress, *FireDB--a database of functionally important residues from proteins of known structure.* Nucleic Acids Res., 2007. **35**(Database issue): p. D219-23.
17. Pazos, F. and M.J. Sternberg, *Automated prediction of protein function and detection of functional sites from structure.* Proc Natl Acad Sci U S A., 2004. **101**(41): p. 14754-9.
18. Zhang, J. and M. Nei, *Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods.* J Mol Evol, 1997. **44 Suppl 1**: p. S139-46.
19. Bharatham, K., Z.H. Zhang, and I. Mihalek, *Determinants, discriminants, conserved residues--a heuristic approach to detection of functional divergence in protein families.* PLoS One, 2011. **6**(9): p. e24382.

20. Chakrabarti, S., S.H. Bryant, and A.R. Panchenko, *Functional specificity lies within the properties and evolutionary changes of amino acids.* J Mol Biol., 2007. **373**(3): p. 801-10.

21. Casari, G., C. Sander, and A. Valencia, *A method to predict functional residues in proteins.* Nat Struct Biol., 1995. **2**(2): p. 171-8.

22. Donald, J.E. and E.I. Shakhnovich, *SDR: a database of predicted specificity-determining residues in proteins.* Nucleic Acids Res., 2009. **37**(Database issue): p. D191-4.

23. Gu, X., *A site-specific measure for rate difference after gene duplication or speciation.* Mol Biol Evol., 2001. **18**(12): p. 2327-30.

24. Gu, X., *A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences.* Mol Biol Evol., 2006. **23**(10): p. 1937-45.

25. Gu, X., *Statistical theory and methods for evolutionary genomics*. 2011, Oxford: Oxford University Press.