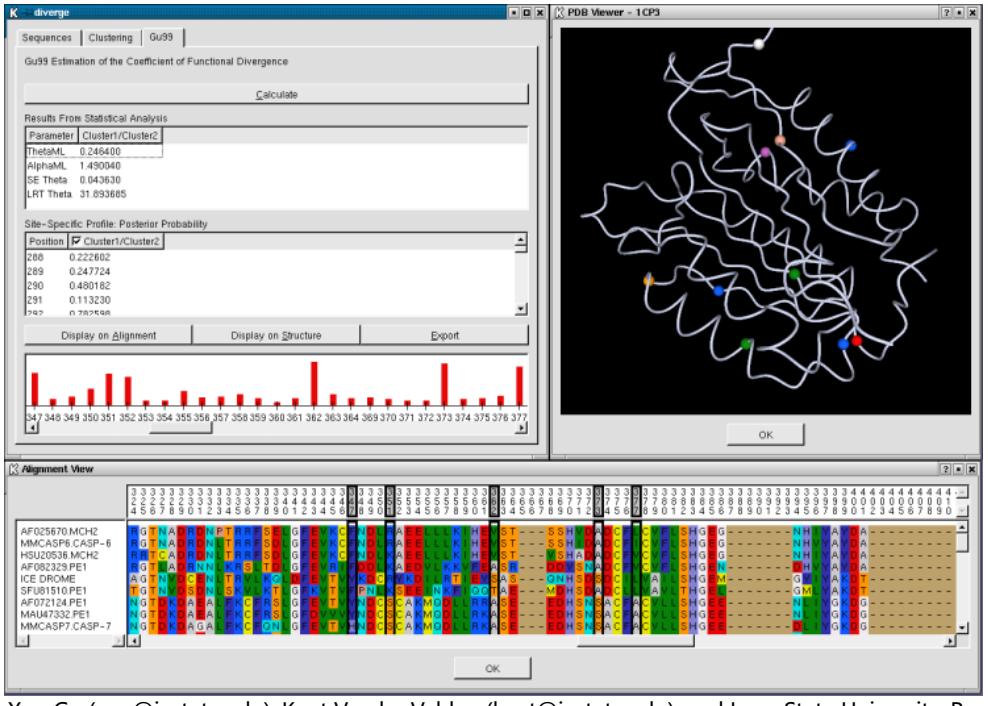
DIVERGE v1.04



Xun Gu
Department of Zoology and Genetics
LHB Center for Bioinformatics and Biological Statistics
332 Science II
Iowa State University
Ames, IA 50011

Telephone: 515-294-8075
Fax: 515-294-8457
Email: xgu@iastate.edu

and

Kent Vander Velden
Bioinformatics and Computational Biology
301 Science II
Iowa State University
Ames, IA 50011

Telephone: 515-294-4567
Fax: 515-294-8457
Email: kent@iastate.edu

and can be downloaded from http://xgu1.zool.iastate.edu.

Suggested citation:

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

1 Introduction

DIVERGE is 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. Moreover, when the 3D protein structure is available, these predicted sites can be mapped to the 3D structure viewer to explore its structure basis. The program can be used on Linux and Microsoft Windows operating systems.

2 Installation

DIVERGE can be downloaded from http://xgu1.zool.iastate.edu. Two versions are available one for use on Linux and another for use on Microsoft Windows. Each package include the exe file, installation procedure, README document, and a list of example files.

The Microsoft Windows version uses an InstallShield script to automate the installation process of DIVERGE. After retrieving the software installer one simply needs to run the executable to begin the installation. The installer walks a person through the entire process. After the installation is complete a new DIVERGE entry will be present in the Start menu. If for some reason the InstallShield script does not run on a particular machine, a zip file is available containing the installation files.

The Linux version does not come with an automated installer, but should not be of much difficulty for those that have installed software on Linux before. The DIVERGE package is distributed as a compressed tar file. To extract the files simply run `tar -zxvf diverge.tar.gz`. This will create the main directory DIVERGE under which will be the executable and associated support files. The executable was compiled on a Mandrake 8.0 system.

The included example is from:

```
Wang and Gu (2001)
Functional divergence in the caspasegene family and altered functional
constraints: Statistical analysis and prediction.
Genetics 158:1311-1320.
```

It include casp.aln (amino acid sequence alignment of caspase gene family), casp.pdb (caspase-1 protein 3D file), and casp.tree (the tree file of caspase gene family).

3 Input File Formats

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
MTETDGFYRSREVLDPAEQYKMDHKRRGTALIFNHERFFWHLALPERRGTNADRDNPTRRFSELGFEVKCFNDLRA
EELLLKIHEVSTSSHVDADCFLCVFLSHGEGNHIYAYDAKIEIQTLTGLFKGDKCQSLVGKPKIFIIQAC
>AF072124.PE1
MTDDQDCAAELEMADSSTEDGVDAKPDRSTIISSLLWKKKKNASMCPVSTTRDRVPTYLYRMDFEKMGKCIIINNK
NFDKATGMDVRNGTDKDAEALFKCFRSLGFEVTVYNDCSCAKMQDLLRRASEEDHSNSACFACVLLSHGE
>AF078533.PE1
MAEDKHNKNPLKMLESLGKELISGLLDDFVEKNVLKLEEEEKKKIYDAKLQDKARVLVDSIRQKNQEAGQVFVQTF
LNIDKNSTSIKAPEETVAGPDESVGSAATLKLCPHEEFLKLCKERAGEIYPIKERKDRTRLALIICNTEF
```

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           ----MDEADRRLLRRCRLRLVEELQVDQLWDALLSSELFRPHMIEDIQRAGSGSRRDQAR
CELCED3A.CED-3          NGDMINSCGTVREKRREIVKAVQRRGDVAFDAFYDALRSTGHEGLAEVLEPLARSVDSNA
```

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 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,MMCASP8S7.PE1)));
```

Example with branch lengths:

```
(((AF111345:.012,HSU60519.MCH4:.453):.345,HSU86214.PE1:.543):.546,
  (HSCASP8S8.CASP8:.954,MMCASP8S7.PE1:.42):.65);
```

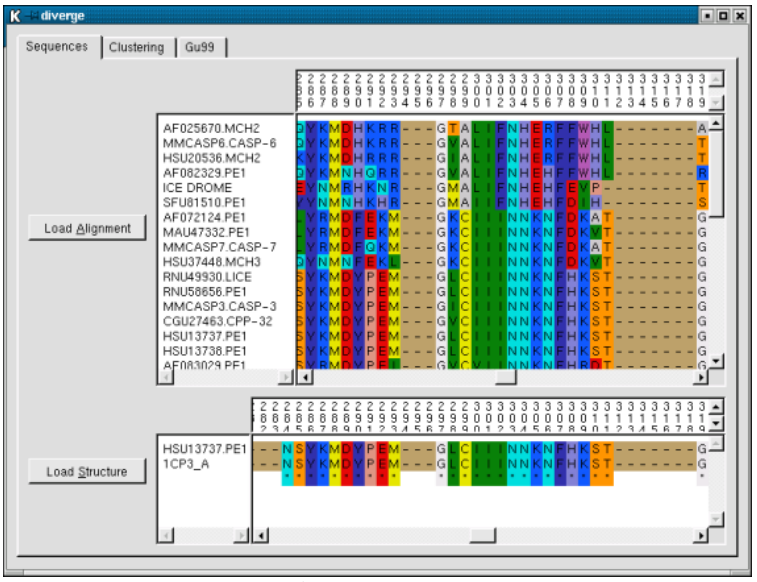
3.3 Protein Structure File Format (optional)

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 have 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 Procedure

Across the top of the software are three tabs named: 1) Sequences, 2) Clustering, and 3) Gu99. Think of these as the three main steps 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.

4.1 Sequences Tab

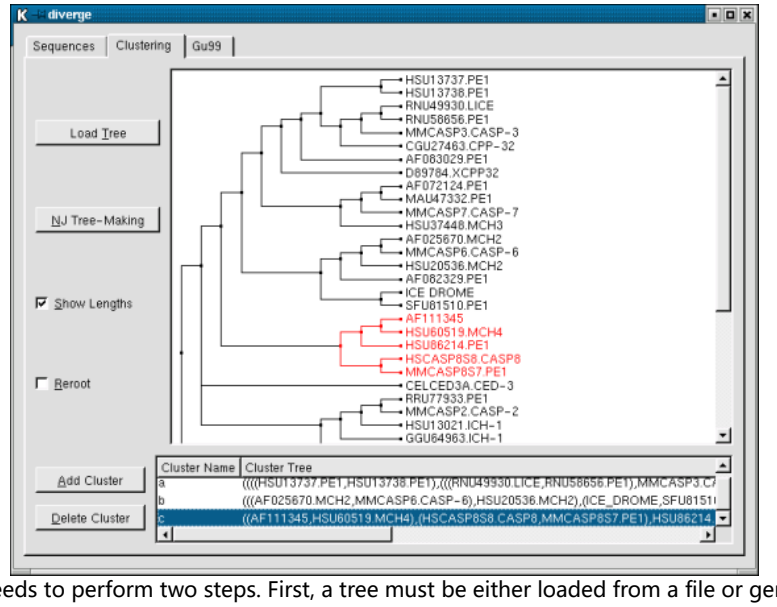


Under the Sequence Tab, the user needs to load an alignment file (FASTA or CLUSTAL), which can be done by selecting the "Load Alignment" button and then selecting the file. After a successful input of the alignment, the area to the right of the button will display the alignment. This area is fully navigable 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 alignment 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



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 selects subtrees from the main gene tree to represent the independent gene clusters by clicking the ancestral nodes.

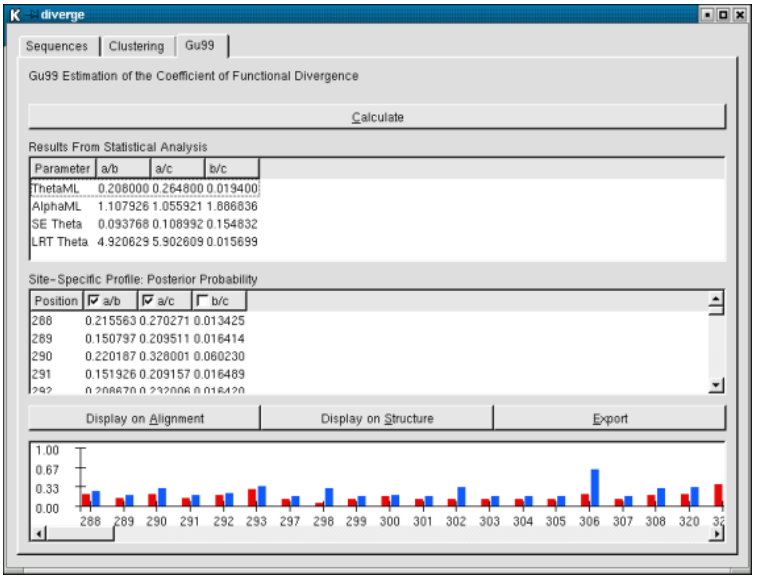
Loading a tree from a file requires that the file be in the standard parenthesis PHYLIP form. This is a common form and is generated most phylogenetic tree software systems (e.g., PHYLIP, PAUP*, or CLUSTAL). Branch lengths are optional and if they are available will be used in drawing the tree. Obviously, 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.

Rerooting of this tree provides a convenient way to select gene clusters. It can be done by selecting the "Reroot" toggle and then selecting the node, represented by a small square on the tree, to form the out-group. The "Reroot" toggle will turn-off after a node has been selected but can also be deactivated explicitly by selecting the toggle again. A toggle was used to allow the user to cancel the reroot option if needed.

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



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 region, listed for each comparison.

- Theta ML: maximum likelihood estimate for theta, the coefficient of functional divergence
- Alpha ML: maximum likelihood estimate for alpha, the gamma shape parameter for rate variation among sites
- SE Theta: standard error of the estimate theta
- LRT Theta: 2 log-likelihood-ratio against the null hypothesis of theta=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 result 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 "a/b" and "a/c" are both 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. The Structure Viewer is explained further in a later section.

4.3.1 Exported Data File Format

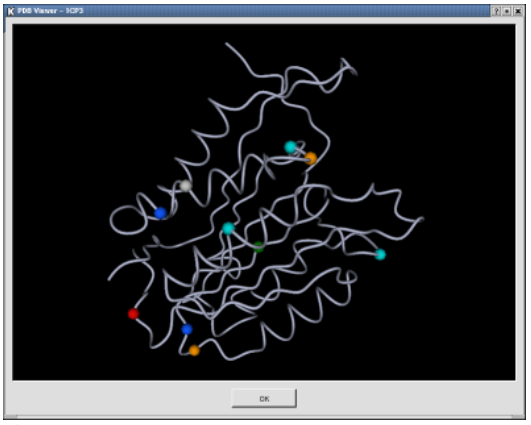
When the "Export" button is selected, all the results are exported to a specified file in tab delimited format for your own personal records and additional processing by other software. Included in the output is the original tree, the selected gene clusters, the Statistical Results, and site-specific 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))),CELCED3A.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),CGU27463.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,SFU81510.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. Identified residues that do not seem to have any relationship to one another on the alignment may form clusters when viewed on the structure. 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 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

Last modified: Tue Aug 14 21:49:24 CDT 2001