Clusterfunc User Manual

June 2, 2011

Contents

1 Suggested citation

Caffrey BE, Williams TA, Jiang X, Toft C, Hokamp K, Fares MA (2011). Proteome-wide analysis of functional divergence reveals the molecular basis of ecological adaptations in bacteria . *Submitted*..

2 Introduction

2.1 Manual overview

Clusterfunc is a simple and fast method for **Cluster**ing functionally divergent genes by **Func**tional **Category**. Before using the software, we recommend you familiarize yourself with the theoretical basis of the method, and the kinds of biological question you can ask using the software. This **Introduction** and the following **Methods** sections explain the background and aims of the software, how it works, and how the outputs might be interpreted. **Obtaining and installing** explains how to set up the software, including how to get the source code, libraries, and/or pre-compiled binaries. **Running** details the command-line operation of the software, including available options.

2.2 Functional divergence

Changes in protein structure and function are reflected at the level of amino acid sequence. This principle suggests that functional divergence—changes in protein function, for instance following gene duplication or new selective pressures—can be identified by analysis of primary sequence data. However, many amino acid substitutions have a negligible effect on protein function (Kimura, 1983). This means that a simple comparison of the sequence differences between two clusters of homologous proteins will not reveal the subset of changes responsible for functional divergence.

2.3 What is Clusterfunc?

The method implemented in Clusterfunc is one of several sequence-based methods for identifying the "interesting" subset of substitutions that might underpin functional divergence—in particular, see also Gu & Velden (2002). These methods are based on the idea of Kimura (1983) that functionally-important residues are highly conserved, so that evolutionary rates tend to be low at important sites. Functional divergence can then by identified by comparing rates (or levels of conservation) between two clades of proteins at a homologous site. Alternatively, a significant change in amino acid identity (such as a large, positively-charged residue in one group of sequences versus a small, neutral residue in the other) could indicate functional divergence even without a change in rate.

2.4 Functional divergence analyses with Clusterfunc

Clusterfunc employs a simple and fast distance method for identifying amino acid positions under functional divergence in protein multiple sequence alignments. This speed makes single-gene or whole-proteome analyses possible on an average desktop computer, which may not be practical for existing likelihood or Bayesian methods. Two different kinds of analysis can be performed by Clusterfunc, each of which can be used to address a specific biological question:

- *Single gene*: Given a protein multiple sequence alignment and a user-generated phylogenetic tree, test for functional divergence on each node of the tree and report the number of significant FD sites for each branch. Used to identify the branches with the greatest amount of functional divergence (e.g. as used in Williams et al. (2010)), or to examine the specific sites under FD on certain lineages or over the whole tree.

- *Multiple genes/whole proteome*: Given a set of multiple sequence alignments that have been tagged using some ontology system (e.g. COG), perform a FD analysis on each and identify the functions and species enriched for functional divergence across the whole dataset. This method also clusters species according to shared patterns of category enrichment (see Caffrey et al. (2010) for a case study).

3 Method

The following text describes the steps of the Clusterfunc method for a whole-proteome analysis. Analyzing a single gene does not involve the clustering and enrichment testing step, and only one alignment is tested. This text is based on that of Caffrey et al. (2010).

3.1 Building gene trees

When analyzing entire proteomes for functional divergence, the use of a species tree to infer events on each branch is problematic: extensive horizontal gene transfer (HGT), particularly among prokaryotes, means that genomes may not be related in a tree-like way (Dagan & Martin, 2006). We therefore calculate a tree for each gene (set of homologous sequences) in the dataset using BIONJ (Gascuel, 1997), under the JTT model of protein sequence evolution (Jones et al., 1992). Calculations for that gene are then made exclusively using the resulting tree.

3.2 Scoring functional divergence

We walk through the phylogenetic tree and calculate functional divergence scores at each of the inner nodes. Clades on either side of the bifurcation are compared to the closest available outgroup with respect to one of several possible substitution matrices (see "Running" below). Scores for each column are given by:

$$FD_{score} = \frac{\overline{X_1} - \overline{X_2}}{S_{X_1 - X_2}} \tag{1}$$

where $X_{1,2}$ are the mean substitution scores for mutation from clades on either side of the bifurcation in the phylogenetic tree relative to the outgroup and the standard error for unequal sample sizes with unequal variances is given by:

$$S_{X_1 - X_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{2}$$

3.3 Significance testing

For each inner node tested, a simulated sequence alignment is created using the JCprot model according to the gene-specific phylogenetic tree calculated above. That is, the distance matrix for the simulated data is equal to that of the real data, but sequences evolve under a model that represents molecular evolution without functional constraint, as all substitutions have equal probability. The simulated alignment is tested according to equation (1), resulting in a null distribution of the test score against which P-values for the real data can be evaluated, with significance taken at the 5% level. These values are then corrected for multiple testing by the False Discovery Rate method (**?**). Following this procedure, branches on the tree that still possess at least one significant site are considered to be under functional divergence for the purposes of enrichment and clustering.

3.4 Enrichment analysis and clustering

Once all alignments have been analyzed, we perform three different enrichment tests to ask three different biological questions of the data. The enrichment analysis is performed for each species, for each functional category, and for species-by-functional category. We use a chi-squared test:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{3}$$

Where O_i is the observed frequency of genes/alignments under functional divergence, E_i is the expected frequency and n is the number of possible outcomes of each event. The result is three tests that assess:

1. Whether each species is enriched (either under or over enrichment) relative to all other species: that is, whether some species experience more or less functional divergence than others.

2. Whether each functional category is enriched (either under or over enrichment) relative to all other categories: whether some groups of genes undergo more or less functional divergence than others.

3. Whether each functional category in each species is enriched (either under or over enrichment) relative to all other categories in each species:

whether particular species undergo more or less functional divergence in particular categories than others.

The results of the third test can be read into R viewed in the form of a heatmap which clusters species according to functional divergences.

4 Obtaining and installing

The source code and a selection of binaries can be found on http://www.bioinformatics.org/clusterfunc/. We advise that source code be used if possible as it will most likely run faster, however this involves installation of 2 other libraries which means there is more effort involved.

4.1 Binaries:

After download the binaries should work straight off the bat so continue to the "Running" section. Just make sure you have the correct binary for the system you are using(Mac/Linux). Also if you want to be able to use this program from any path on Linux/Mac, copy the binary to /usr/bin/

4.2 Source Code Compilation(Mac/Nixes/Windows)

After downloading you should perform the following instructions:

4.2.1 Mac/Nixes

1. Firstly make sure to have Apple Development tools installed if on Mac. This can be found on the OS install disk under optional installs folder.

2. There are two pre-required libraries which are needed to compile the program. These are found at the following addresses:
   Gnu Scientific Library
   Bio++

3. Unzip the package, this can be done with a host of extractors(stuffit, winrar, tar, etc etc).

4. On the command line change directory to the unzipped package.

   cd clusterfunc

5. To install clusterfunc perform the following:
   make

6. Your program should be ready to run.

4.2.2 Windows with Cygwin

After downloading you should perform the following instructions:

1. First install Cygwin, available at: Cygwin

2. There are two pre-required libraries which are needed to compile the program. These are found at the following addresses:
   Gnu Scientific Library
   Bio++

3. Run Cygwin, then on the command line change directory to the unzipped package.

   cd clusterfunc

4. To install clusterfunc perform the following:
   make

5. Your program should be ready to run.

5 Running

5.1 Single Gene Run

This mode performs a branch-specific analysis of functional divergence on a single gene (multiple sequence alignment), optionally with a user-supplied phylogenetic tree. The output gives you the number of sites under functional divergence on each branch of the tree as well as specific residues under functional divergence.

input files: FASTA alignment(required), Newick tree file(optional).

output files: A set of files which correspond to the name of the input alignment file but are numbered from zero upwards.

The command line option to be used is for a single gene run is -f.

Examples:
./cfc -f alignment.fasta
./cfc -f alignment.fasta -t treefile.tre

5.2 Multiple Gene Run

This mode performs all of the analysis of the single gene run but on multiple sequences alignments. The results of all of the alignments are analysed if funcitonal categories are specified and enrichment analysis is performed. The command line option to be used is -F

input files: A folder of fasta formatted alignments(required) and a tree file(optional).
output files: A files which gives the enrichment of each species and functional category analyzed. A file which gives the enrichment of each species for each functional category and optionally the information for each single gene run.
Examples:
./cfc -F myfolder
./cfc -F myfolder -t mytree.tre

Note: You **MUST** use either -f or -F followed by a file or folder respectively.

5.3 Optional parameters:

amino acid information: To supress specific amino acid information(e.g. for larger runs) use the -a option.

Example: ./cfc -F alingment_folder -a

Force a tree topology: To force the topology of a phylogenetic tree one can pass a pre-calculated tree with the -t option. The tree file should be in the Newick format.

Example: ./cfc -F myfolder -t mytree.tre

Specify an alternative matrix: To use an alternative matrix to the one provided with the program(BLOSUM62) use the -m option.

Example: ./cfc -f alignment.fasta -m BLOSUM80.tab

Note: Any different matrix file used should be in the same format as the matrix provided

Threshold cut-off: To chose a threshold cut-off for convergences use the -c option. This option specifies the value to which the change in standard deviation on subsequent runs of the simulations must have converged within.

Example: ./cfc -F alingment_folder -c 0.001

5.4 Alternative uses

The options given above can be used in multiple ways, there are no limitations, the following are examples of other legitimate run commands:

./cfc -f alignment.fasta -t mytree.tre -m BLOSUM80.tab
/cfc -F myfolder -a

6 Clustering

After analyzing many proteomes with Tags assigned to each gene a file named matrix.txt will be output. To cluster this using R you need gplots, then use the following commands.

1. Open R by typing R on command line then complete the following.
2. cfc=read.table(”matrix.txt”, header=TRUE);
3. mylabels=cfc$X.
4. cfc$X. = NULL
5. x<- as.matrix(cfc)
6. pdf(”heatmap.pdf”)
7. library(gplots)
8. heatmap.2(x,key=FALSE,col=colorpanel(15, ”yellow”, ”blue”) ,trace=”none”,xlab=”COG Tag”,ylab=”Bacteria”, labRow=mylabels)
9. dev.off()

This will output a pdf called ”heatmap.pdf”

7 Other scripts: merging

With the source code is packaged a script for merging the results from multiple runs of the Clusterfunc program. Clusterfunc outputs a file called matrix.txt. This matrix contains all of the functional divergence analysis in raw numeric form. For one to merge the results of multiple runs(on a cluster say) or also even just one run which has included paralogs and orthologs they must do the following:

1. Species which are to be merged should be the same in all alignments, meaning they should have the exact same spelling without trailing spaces etc.

2. When using paralogs in alignments one should name ALL species in the alignments with the same number of characters. For example, in the case study we used the 5 character codes used in the OMA database. All paralogs should be followed by a number. So for example one could use the following species names HUMAN, ECOLI, ECOLU, LEUCK. Then with paralogs use HUMAN2, ECOLI2, etc. If a species name has numbers this is OK but should still stick to the same number of letters, e.g. STRP4 and for paralogs STRP41.

3. Copy the matrix files to the same folder, making sure to rename them from 0..*n*, i.e. matrix0.txt, matrix1.txt,.....matrixn.txt

4. Now to merge the output of the matrix files perform the following: ./*merge* 4 5 , where 4 is the number of matrix files and 5 is the number of letters in the sequence names discounting the numbers attributed for

5. A single file is output called Enrichment.txt which has analyzed the combined matrices.

N.B. Examples of the naming/numbering conventions to be used are given in the sample alignments.

8 Case studies

References

Dagan, T., & Martin, W. (2006). The tree of one percent. *Genome Biology*, *7*(10), 118. PMC1794558.

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, *14*(7), 685–695. PMID: 9254330.

Gu, X., & Velden, K. V. (2002). DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics*, *18*(3), 500–1. 11934757.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences: CABIOS*, *8*(3), 275–282. PMID: 1633570.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge: Cambridge University Press.

Williams, T. A., Codoer, F. M., Toft, C., & Fares, M. A. (2010). Two chaperonin systems in bacterial genomes with distinct ecological roles. *Trends in Genetics: TIG*. PMID: 20036437.