Coevolution Analysis using Protein Sequences

(CAPS)

Version 1

Mario A. Fares

Suggested citation (Fares and Travers (2006) Genetics)

The author can be reached at:

Mario A. Fares

Evolutionary Genetics and Bioinformatics Laboratory

Department of Genetics

Smurfit Institute of Genetics

University of Dublin, Trinity College

Dublin, Ireland

Table of Contents

1. History 4

2. Introduction 4 – 5

3. General assumptions and requirements of the program 5 – 6

4. Method and Model 6 – 11

  4.1. Theory 7 – 9

  4.2. Removing the phylogenetic coevolution 10

  4.3. Using Atomic Distances as additional information in coevolution analyses 11

5. Running CAPS 11 – 13

6. Control File 13 – 25

  6.1. Input files 14 – 15

  6.2. Output files 15 – 19

  6.3. Coevolution analysis 19

  6.4. Type of data 20

  6.5. 3D test 20

  6.6. Reference sequence 20 – 21

  6.7. Atom interval 21

  6.8. Significance test 21

  6.9. Threshold R value 21 – 22

  6.10. Time correction 22

  6.11. Time estimation 22 – 23

  6.12. Threshold alpha value 23

  6.13. Random sampling 23

  6.14. Gaps 23 – 24

6.15. Minimum R 24

6.16. GrSize 24 – 25

7. Removing the phylogenetic coevolution and testing functional shifts 25 – 27

8. A case study 27 – 32

9. Files and executables in CAPS 32 – 33

10. References 34

1. History

CAPS is the first version of the software aimed at measuring the coevolution between amino acid sites belonging to the same protein (intra-molecular coevolution) or to two functionally or physically interacting proteins (inter-molecular coevolution). This software is written in PERL and runs in various operating systems including Linux/Unix, Windows and Mac Os X. The software implements the method to detect intra-molecular coevolution as published in Genetics (Fares and Travers, 2006) but it also includes several other analyses unpublished to date, such as the preliminary analysis of compensatory mutations or inter-protein coevolution analysis. New modules have been introduced to conduct preliminary analyses of compensatory mutations. These analyses include the detection of correlation in the evolution of the hydrophobic or molecular weight characteristics of two amino acid sites in a multiple sequence alignment. Further, unlike the previous Beta versions developed to detect intra-molecular coevolution by implementing the method of Fares and Travers (2006), this first released version includes other analyses that permit the identification of inter-molecular coevolution and the inference of protein-protein functional interactions.

2. Introduction

CAPS implements a method to analyse amino acid coevolution in a protein multiple sequence alignment and has been successfully applied to detect intra-molecular coevolution in the Gag protein of HIV-I and in the heat-shock proteins 60Kda GroEL and 90Kda Hsp90 (Fares and Travers, 2006). The code is very carefully written and is highly amenable for adding new functions and code modification to perform other analyses. This is possible due to a comprehensive split of the different functions into subroutines and modules, which almost make of CAPS a package rather than a program. The program is highly conservative to detect coevolution and has the ability of disentangling functional/structural/interaction coevolution from stochastic and phylogenetic coevolution. The program is also very straightforward to use and only requires a protein multiple sequence alignment or alternatively a protein-coding multiple sequence alignment. For inter-molecular coevolution, the program requires two alignment files (one per protein), both containing the same number of sequences from the same taxa and in the same order in the input files. Several other parameters are required as shown in the control file of CAPS. The results are generated in an output that includes a comprehensive amount of information about the many different parameters estimated from the alignment.

CAPS is a highly flexible program regarding the number of sequences in the alignment. There is no theoretical limit for the number of sequences and they can be analysed in a reasonable computational time (for example, intra-molecular coevolution analysis of an alignment containing 40 sequences comprising 600 amino acid sites requires just few minutes). The computation time however is exponentially dependent on the number of sites in the alignment.

3. General assumptions and requirements of the program

The program has no limits regarding the number of sequences in the alignment as well as the length of the alignment. However, the program makes several assumptions that might pose limitations in particular data sets:

a) Since no phylogenetic tree is required, the number of sequences in the alignment can be anything between three sequences and few thousand sequences. We have shown however that the minimum number of sequences required to show accurate results minimising the number of false positives is 10 (Fares and Travers, 2006).

b) Pairwise sequence distances can be very broad, however large distance may impose limitations in some of the assumptions made by the method since the time for the divergence of two sequences is assumed to be proportional to the number of synonymous substitutions per synonymous site. For example, if the correction parameter for the time since two sequences have diverged is used in the calculations, the number of substitutions per silent sites should not be greater than 1.25. This is to avoid the underestimation of divergence times due to the saturation of synonymous sites.

c) Blosum values are considered to be proportional to the time since two sequences have diverged (see theory below).

d) No selection shifts are assumed and hence dependence between amino acid sites is assumed to be constant throughout the evolutionary time of a protein. Nonetheless, clade-specific coevolutionary analyses are permitted in CAPS, which allow ameliorating this problem.

4. Method and Model

Coevolution analysis using protein sequences (CAPS) compares the correlated variance of the evolutionary rates at two sites corrected by the time since the divergence of the protein sequences they belong to. Substitutions or conservation at two independent sites cannot be directly compared due to their amino acid composition difference. The method instead compares the transition probabilities between two sequences at these particular sites, using Blocks Substitution Matrix (BLOSUM; Henikoff and Henikoff, 1992). For each protein alignment the correspondent BLOSUM matrix is applied depending on the average sequence identity.

4.1. Theory

Despite the fact that BLOSUM matrices correct for the substitution values due to the estimated divergence between sequence pairs, a given alignment can include sequences whose pair-wise distance is significantly divergent from the mean pair-wise distance. For instance, an alignment including two highly divergent sequence groups (for example gene duplication predating speciation) could show an unrealistic pair-wise average identity level. In this respect, sequences that diverged a long time ago are more likely to fix correlated mutations at two sites by chance (under a Poisson model) compared to recently diverged sequences. BLOSUM values should be hence normalized by the time of divergence between sequences. BLOSUM values (B_{ek}) are thus weighted for the transition between amino acids *e* and *k* using the time (t) since the divergence between sequences *i* and *j*:

$$\left(\theta_{ek}\right)_{ij} = \left(\mathrm{B}_{ek}t^{-1}\right)_{ij} \qquad (1)$$

The assumption made in equation 1 is that the different types of amino acid transitions (slight to radical amino acid changes) in a particular site follow a Poisson distribution along time. The greater the time since the divergence between sequences *i* and *j* the greater the probability of having a radical change. A linear relationship is thus assumed between the BLOSUM values and time. Synonymous substitutions per site (d_{Sij}) are silent mutations, as they do not affect the amino acid composition of the protein. These mutations are therefore neutrally fixed in the gene. Assuming that synonymous sites are not saturated or under constraints, d_S is proportional to the time since the two sequences compared diverged. Time (*t*) therefore is measured as d_S.

Note that convergent radical amino acid changes at two sites in sequences that have diverged recently have larger weights compared to convergent changes in distantly related sequences.

The next step is the estimation of the mean θ parameter for each site $\left(\overline{\theta}_C\right)$ of the alignment, so that:

$$\overline{\theta}_C = \frac{1}{T}\sum_{S=1}^{T}(\theta_{ek})_S \qquad (2)$$

Here S refers to each pairwise comparison, while T stands for the total number of pairwise sequence comparisons, and thus:

$$T = \frac{N(N-1)}{2} \qquad (3)$$

Where N is the total number of sequences in the alignment.

The variability of each pairwise amino acid transition compared to that of the site column is estimated as:

$$\hat{D}_{ek} = \left[(\theta_{ek})_{ij} - \overline{\theta}_C\right]^2 \qquad (4)$$

The mean variability for the corrected BLOSUM transition values is:

$$\overline{D}_C = \frac{1}{T}\sum_{S=1}^{T}\left[(\theta_{ek})_S - \overline{\theta}_C\right]^2 \qquad (5)$$

The coevolution between amino acid sites (A and B) is estimated thereafter by measuring the correlation in the pairwise amino acid variability, relative to the mean pairwise variability per site, between them. This covariation is measured as the correlation between their $\hat{D}_{ek}$ values, such as:

$$\rho_{AB} = \frac{\sum_{S=1}^{T}\left[(\hat{D}_{ek})_S - \overline{D}_A\right]\left[(\hat{D}_{ek})_S - \overline{D}_B\right]}{\sqrt{\sum_{S=1}^{T}\left[(\hat{D}_{ek})_S - \overline{D}_A\right]^2 \sum_{S=1}^{T}\left[(\hat{D}_{ek})_S - \overline{D}_B\right]^2}} \qquad (6)$$

Here *e* and *k* are any two character states at sites A and B. To determine if the correlation coefficient (ρ_{AB}) is significant, either a re-sampling or a simulation analysis can be performed. In the first approach we randomly sample *K* number of pairs of sites and compute equations 1 to 6 for each pair. The mean correlation coefficient and its variance are then estimated as:

$$\overline{\rho} = \frac{1}{K}\sum_{l=1}^{K}\rho_l \quad ; \quad V(\rho) = \frac{1}{K}\sum_{l=1}^{K}\left(\rho_l - \overline{\rho}\right)^2 \tag{7}$$

Correlation coefficients are then tested for significance under a normal distribution:

$$Z = \frac{\rho_{AB} - \overline{\rho}}{\sqrt{V(\rho)}} \tag{8}$$

The second approach consists of the Monte Carlo simulation of *K* sequence alignments. Here the coevolution test is conducted for a number of randomly selected pairs of sites in each simulated alignment computing equations 1 to 6. An average value of the correlation for the simulated alignments and its variance are obtained utilizing equation 7. Finally the real correlation coefficients are tested using equation 8.

The statistical power of the test is optimised by analysing sites showing:

$$\overline{D}_C > \Theta - 2\sigma_\Theta \tag{9}$$

Here, Θ is the parametric value of $\overline{D}_C$ from equation 5 and σ is the standard deviation of Θ. Θ is calculated as:

$$\Theta = \frac{1}{L}\sum_{s=1}^{L}\left(\overline{D}_C\right)_s \tag{10}$$

Where L is the length of the alignment. Pair-wise comparisons including gaps in any or both sites at any sequence are excluded from the analysis.

4.2. Removing the Phylogenetic Coevolution

Coevolution between amino acid sites can be the result of their structural, functional, or physical interaction, their phylogenetic convergence, and their stochastic covariation. The analysis of simulated data to test for significance removes stochastic effects. To disentangle functional, structural, and interaction coevolution from phylogenetic coevolution, the method is applied to the complete alignment and to sub-alignments where specific phylogenetic clades are removed from the tree. Coevolving amino acid sites that are no longer detected following removal of one of the clades will be classified as phylogenetic coevolving sites as they occur in specific branches of the tree. Conversely, coevolving amino acid sites detected irrespective of the tree clades removed will be considered as functional/structural/interaction coevolving sites since they present correlated changes throughout the phylogenetic tree. Notice, that the latter condition means that when one amino acid changes, the covarying amino acid has necessarily to change. In the former condition, a change in one site does not always (in all branches) involve a change in the covarying site. In other words, our method detects phylogenetic-independent coevolution. Clades for coevolution analyses are defined in terms of their biological coherence and/or statistical support (defined as bootstrap values). Consequently, phylogenetic clades are specified prior to conducting the coevolutionary analysis and they include sequences that are either forming a well-defined biological cluster or alternatively the cluster is supported by a high bootstrap value.

4.3. Using Atomic Distances as additional information in coevolution analyses

Spatial proximity between coevolving sites can be used to define their structural or functional interaction. In this method coevolution is not always synonymous with physical interaction but also involves structural and functional coevolution, as has been previously described (Gloor, et al., 2005; Lockless and Ranganathan, 1999; Pritchard and Dufton, 2000; Suel, et al., 2003).

The three-dimensional closeness of two sites is estimated as the vectorial distance between their atomic centers (δ). This distance is obtained comparing the three-dimensional coordinates (X, Y and Z) of atoms A and B for amino acids *i* and *j*:

$$\delta_{A-B} = \vec{A} - \vec{B} = \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2 + (Z_A - Z_B)^2} \tag{11}$$

Since each amino acid consists of several atoms, the mean atomic distance ($\bar{\delta}$) between sites *i* and *j* is taken:

$$\bar{\delta}_{i-j} = \sqrt{\begin{array}{l} \left[\left(\frac{1}{\mu_i}\sum_{m=1}^{\mu_i} X_m\right)_i - \left(\frac{1}{\mu_j}\sum_{m=1}^{\mu_j} X_m\right)_j\right]^2 + \left[\left(\frac{1}{\mu_i}\sum_{m=1}^{\mu_i} Y_m\right)_i - \left(\frac{1}{\mu_j}\sum_{m=1}^{\mu_j} Y_m\right)_j\right]^2 + \\ + \left[\left(\frac{1}{\mu_i}\sum_{m=1}^{\mu_i} Z_m\right)_i - \left(\frac{1}{\mu_j}\sum_{m=1}^{\mu_j} Z_{mj}\right)_j\right]^2 \end{array}} \tag{12}$$

Here, μ refers to the total number of atoms in the amino acid. The significance of the distance is tested by comparing it to a distribution of *K* random amino acid pairs sampled from the three-dimensional structure.

4.4. Solving the inter-dependence problem of the coevolution analyses

Because of the multiple tests conducted for each site of the multiple sequence alignments, some kind of correction for such undesired effect is required to increase the power of the coevolutionary tests conducted. To correct for multiple testing I implemented the step-down permutational correction (Westfall and Young 1993). An application of this correction is also available in a previous report (Templeton, et al., 2005).

5. Running CAPS

CAPS can run in Windows, Unix and Mac OS X operating systems. Windows version runs by simply typing perl CAPS2.pl in the DOS window. Please make sure you have the PERL interpreter installed in your computer. To check whether or not PERL interpreter is installed on your computer please type:

```
perl –v
```

This will also provide information about the version of the perl interpreter available in your computer. If the interpreter is not installed, it can be easily downloaded freely from the web (www.perl.com).

CAPSv1.tgz should be uncompressed and unpacked following the steps below:

- **gunzip CAPSv1.tar.gz**
- **Tar –xvf CAPSv1.tar**

CAPSv1.tgz can be uncompressed using WinZip in Windows.

Once uncompressed CAPS generates four main folders named:

- Documentation: contains the readme file for caps
- Examples: contains an example of the input and output files generated
- Modules: contains all the perl modules required to run CAPS
- Source: contains the source code files (.pl files) and the control files for CAPS.pl and CladesCaps.pl.

All the files should be moved to the same folder to run CAPS without problems.

- Finally the user should make executable the file .pl files in UNIX by typing:

```
chmod +x *.pl
```

The information flow through the program and the computational processes are depicted in figure 1.

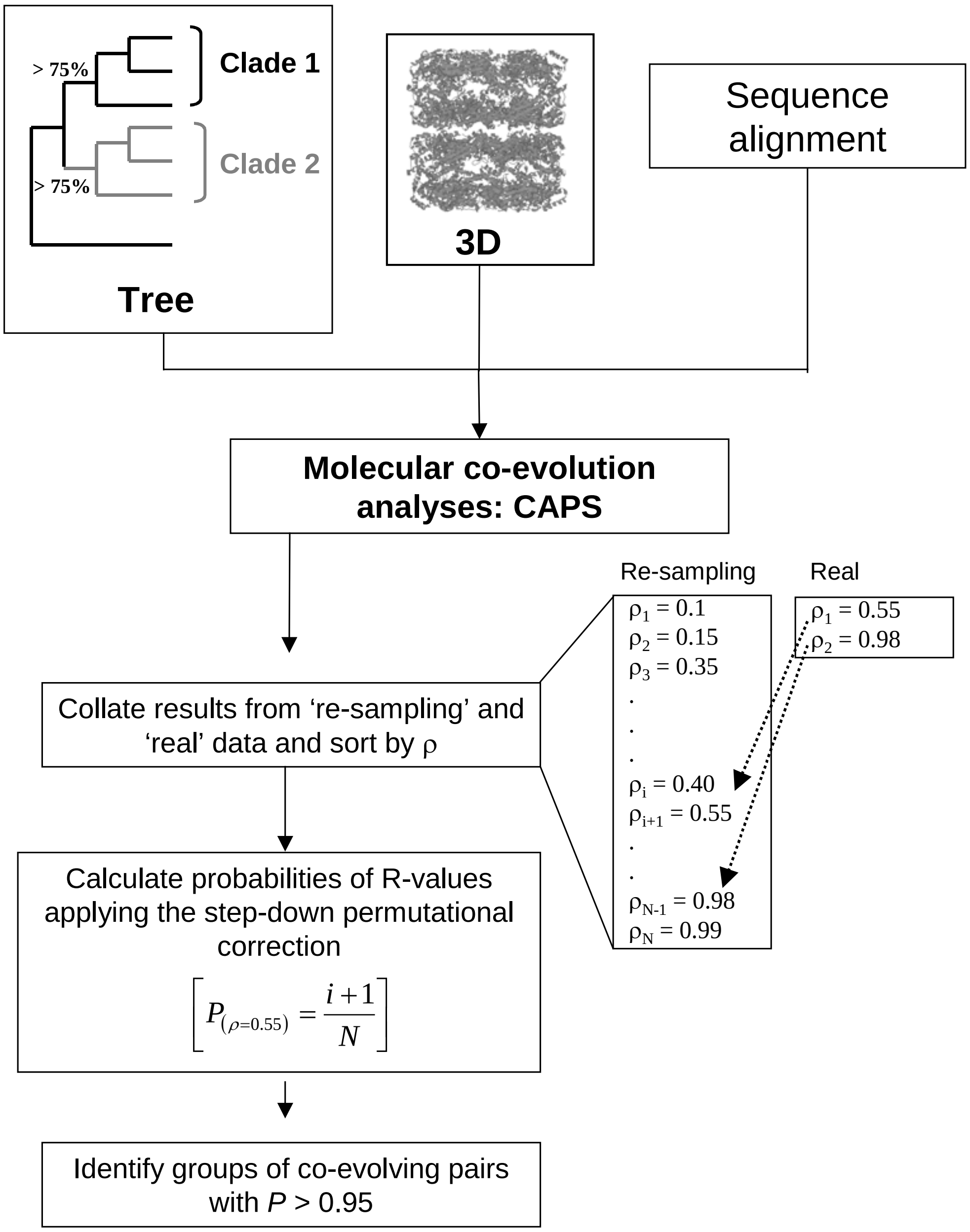

> 75%
Clade 1
Clade 2
> 75%
Tree
3D
Sequence alignment
Molecular co-evolution analyses: CAPS
Re-sampling
Real
$\rho_1 = 0.1$
$\rho_2 = 0.15$
$\rho_3 = 0.35$
$\rho_i = 0.40$
$\rho_{i+1} = 0.55$
$\rho_{N-1} = 0.98$
$\rho_N = 0.99$
$\rho_1 = 0.55$
$\rho_2 = 0.98$
Collate results from 're-sampling' and 'real' data and sort by ρ
Calculate probabilities of R-values applying the step-down permutational correction
$\left[P_{(\rho=0.55)} = \frac{i+1}{N}\right]$
Identify groups of co-evolving pairs with $P > 0.95$

Figure 1. Information flow through the program CAPS

6. Control file

The control file in CAPS is named "CAPS.ctl" and includes several options to run the program under specific conditions. Some of the options in the control file are already computed while some others are future objectives in the subsequent versions of CAPS. The file CAPS.ctl has a very standard format and for each option provides a very brief explanation and guide following the option. The information required by CAPS is presented in figure 2:

```
Input file1: groel.aln              * File containing sequence alignment for the first protein
Input file2: groel.aln              * File containing sequence alignment for the second protein
Out file1: groel.out                * File where the output information should be stored
Co-evolution analysis: 0            * (0) Intra-molecular; (1) Inter-protein
Type of data 1: 1                   * (0) amino acid alignment; (1) codon-based alignment
Type of data 2: 1                   * (0) amino acid alignment; (1) codon-based alignment
3D test: 1                          * Only applicable for intra-protein analysis: (0) perform test; (1) Test is not applicable
Reference sequence 1: 1             * the order in the alignment of the sequences giving the real positions in the protein for 3D analyses
Reference sequence 2: 1             * the order in the alignment of the sequences giving the real positions in the protein for 3D analyses
3D file: groel                      * name of the file containing 3D coordinates
Atom interval: 1-4723               * Amino acid atoms for which 3D coordinates are available
Significance test: 1                * (0) use threshold correlation; (1) random sampling
Threshold R-value: 0.1              * Threshold value for the correlation coeficient
Time correction: 1                  * (0) no time correction; (1) weight correlations by the divergence time between sequences
Time estimation: 1                  * (0) use synonymous distances by Li 1993; (1) use Poisson-corrected amino acid distances
Threshold alpha-value: 0.05         * This option valid only in case of random sampling
Random sampling: 1000               * Use in case significance test option is 1
Gaps: 2                             * Remove gap columns (0); Remove columns with a specified number of gaps (1); Do not remove gap columns (2)
Minimum R: 0.1                      * Minimum value of correlation coeficient to be considered for filtering
GrSize: 3                           * Maximum number of sites in the group permitted (given in percentage of protein length)
```

Figure 2. The control file for the program CAPS.pl

Below I explain each one of the subsections of the control file.

6.1) Input files

Input file asks the user to introduce the name of the file containing the multiple sequence alignment. This sequence alignment can be in any of the most standard formats used in other programs such as Fasta, Phylip or Mega formats **as long as the format is sequential and not interleaved**. For inter-molecular coevolution analyses, the two files containing the sequence alignments must present the same number of sequences and these have to be in the same order. For example, if file 1 contains the sequences in the following order:

File 1:

>*E. coli*

………………….

>*S. typhimurium*

………………….

>*K. pneumoniae*

………………….

File 2 should also have the sequences in that same order:

>*E. coli*

………………….

>*S. typhimurium*

………………….

>*K. pneumoniae*

…………………..

In the next versions of CAPS I will avoid forcing the user to take into account these details.

6.2) Output files

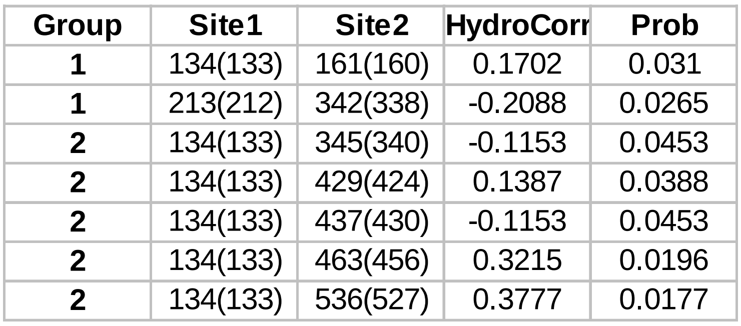
The next option is the output file name, which is up to the user. The output file contains a straightforward interpretable format. The output file includes first a multiple amino acid sequence alignment that is either copied from the input file/s or alternatively generated after translating the protein coding multiple sequence alignment of the input file into a multiple amino acid sequence alignment. Then a hemi-matrix of the pairwise sequence distance is generated. Once the distance are calculated, these are used to correct the pairwise BLOSUM-corrected amino acid differences at a particular amino acid site in the multiple sequence alignment as described in Fares and Travers (2006). As a result of the coevolutionary analyses conducted, a table containing the parameter estimates for the significant coevolving pairs of sites is outputted. This table contains as subheadings, the alignment and real positions for each amino acid site pairs, the mean D values for each site (see Fares and Travers 2006 for a mathematical description of this parameter) and the correlation coefficient for the pairs of sites detected. Please note that real site position depends on the sequence pre-specified as reference sequence by the user. This is a very convenient option especially when the user is thinking on performing analyses of the coevolving amino acid sites in the three-dimensional structure of the protein under analysis. The reference sequence should therefore be that belonging to the organism from which the protein has been crystallised. The exact amino acid position is calculated by subtracting the number of gaps before the position in the sequence alignment from the site position in the multiple sequence alignment. If the crystal structure is being analysed, then the output file also contains a column providing the probability of the distance between the amino acids belonging to that particular coevolution pair. The atomic distance between amino acids is also provided in Angstroms. Finally, CAPS implements an algorithm to join amino acid sites within groups of coevolution as to provide an overview of the networks of coevolving amino acid sites. All the amino acids within each group are coevolving with all the others within the same group.

CAPS also provides in the output information about a preliminary screening for compensatory mutations through the analysis of correlation in the molecular weight, hydrophobicity or both of the amino acid sites detected as coevolving. This is a new feature implemented in CAPS and not included in the manuscript describing the mathematical method (Fares and Travers 2006).

As additional output files CAPS generates comma-separated values files that are readable by Excel or Open Office and enable easy treatment of the results generated by CAPS. These files are summaries of the coevolutionary relationships among the amino acid sites belonging to each one of the coevolution groups. there are three “.csv” files generated:

a) Hydros_coevol.csv

This files contains the pairs of amino acid sites coevolving within each group, their hydrophobicity correlation value and the probability of this correlation estimated after comparing the correlation with a distribution of hydrophobicity correlation coefficients of randomly re-sampled pairs of sites in the multiple sequence alignment (figure 3).

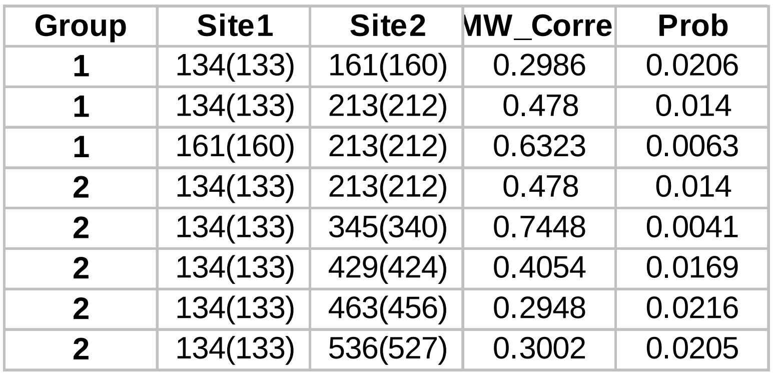


Group	Site1	Site2	HydroCorr	Prob
1	134(133)	161(160)	0.1702	0.031
1	213(212)	342(338)	-0.2088	0.0265
2	134(133)	345(340)	-0.1153	0.0453
2	134(133)	429(424)	0.1387	0.0388
2	134(133)	437(430)	-0.1153	0.0453
2	134(133)	463(456)	0.3215	0.0196
2	134(133)	536(527)	0.3777	0.0177

Figure 3. File containing the information for the correlation analysis in the hydrophobicity values for the pairs of amino acids detected as coevolving in CAPS (file name: Hydros_coevol.csv).

b) MW_coevol.csv

This file contains the pairs of amino acid sites coevolving within each group, the correlation in their molecular weight and the probability of the correlation values as compared to a distribution of molecular weight correlation values for randomly re-sampled pairs of sites in the multiple sequence alignment (Figure 4).

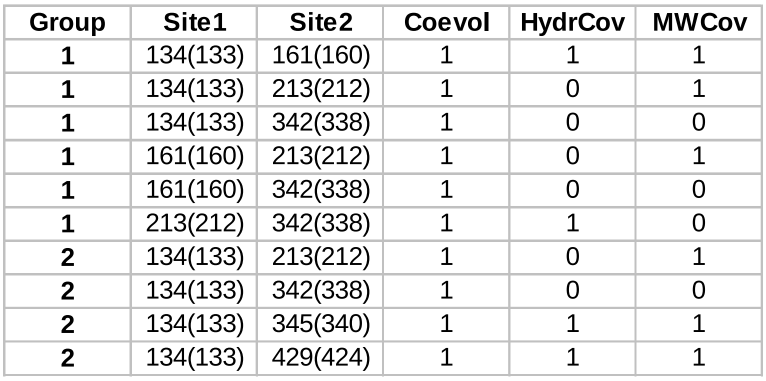


Group	Site1	Site2	MW_Corre	Prob
1	134(133)	161(160)	0.2986	0.0206
1	134(133)	213(212)	0.478	0.014
1	161(160)	213(212)	0.6323	0.0063
2	134(133)	213(212)	0.478	0.014
2	134(133)	345(340)	0.7448	0.0041
2	134(133)	429(424)	0.4054	0.0169
2	134(133)	463(456)	0.2948	0.0216
2	134(133)	536(527)	0.3002	0.0205

Figure 4. File containing the information for the correlation analysis in the molecular weight values for the pairs of amino acids detected as coevolving in CAPS (MW_coevol.csv).

c) CompCoevol.csv

This file contains a summary of all the correlations analysed allowing the user to identify when a pair of amino acid sites that are coevolving and present evidence of positive epistasis. This file contains a column indicating the group the pairs of sites considered belong to, two columns to identify the position of the site in the multiple sequence alignment as well as in the real reference sequence pre-specified by the user and a binary code that indicates whether that particular pair of sites are coevolving (1), present correlation in their hydrophobicities (1) or not (0) and are present correlation in their molecular weights (1) or not (0) (Figure 5).



Group	Site1	Site2	Coevol	HydrCov	MWCov
1	134(133)	161(160)	1	1	1
1	134(133)	213(212)	1	0	1
1	134(133)	342(338)	1	0	0
1	161(160)	213(212)	1	0	1
1	161(160)	342(338)	1	0	0
1	213(212)	342(338)	1	1	0
2	134(133)	213(212)	1	0	1
2	134(133)	342(338)	1	0	0
2	134(133)	345(340)	1	1	1
2	134(133)	429(424)	1	1	1

Figure 5. File containing the summarised information for the correlation analysis in the hydrophobicities, molecular weights and coevolution values for the pairs of amino acids detected as coevolving in CAPS (File name: CompCoevol.csv).

6.3) Coevolution analysis

This option allows the user to conduct two types of coevolutionary analyses. Option 0 is for the intra-molecular coevolution analysis whereas option 1 is for the inter-protein coevolution analysis. Intra-protein coevolution analysis is computed as explained in the article by Fares and Travers (2006). For the inter-protein coevolution analysis, the same approach is held since the average evolutionary rates for each protein do not influence the calculations. Inter-protein coevolution analyses do not allow considering 3D structures, although three-dimensional analysis of the sites from one protein belonging to the same group of inter-protein coevolution is going to be implemented in a future version of CAPS to allow for the detection of protein-protein interfaces based on analysis of coevolution. When providing the multiple sequence alignments for both proteins, the only requirement is to have the same species (taxa) in both files introduced in the same order. Different reference sequences can be used for each protein.

6.4) Type of Data

This option allows the user to specify the type of the multiple sequence alignment being introduced as input file. 0 designates multiple amino acid sequence alignment whereas 1 stands for multiple protein-coding sequence alignment. Type of data has to be very carefully identified since the calculations of the coevolutionary parameters depend on the time-correction parameter, which in turn depends on the type of data used. Only for protein-coding sequences (codon based sequences) could one use synonymous nucleotide substitutions per site as a measure of the divergence time between pairs of sequences.

6.5) 3D test

This option allows considering the three-dimensional structure in the coevolutionary analyses. This information can be very valuable especially when the user is attempting to disentangle functional, structural and interaction coevolution. Option 0 instructs the program to perform the test whereas option 1 means the test is not applicable (no protein crystal structure is available). If the option is set to 0, the program will read the structure from a PDB files provided by the user (and specified in the control file option “3D file”) and will provide information about the structural proximity of the coevolving amino acid site. This information is useful to determine whether two coevolving sites are significantly proximal (indicating structural and probably functional coevolution) or distant (indicating functional coevolution) in the three-dimensional structures.

6.6) Reference sequence

In here the user has to specify the reference sequence order in the input file. This sequence will be used to identify the real site position in the multiple sequence alignment by subtracting the number of gaps until that position in the reference sequence. This option becomes very useful especially if the three-dimensional structure is being analysed, in which case the reference sequence should be that belonging to the organisms from which the protein has been crystallised.

6.7) Atom interval

Many times the structure for the protein being analysed is not completely solved. For example, if only part of the molecule has been successfully crystallised. In such a case, only part of the molecule can be used to conduct three-dimensional analyses of the coevolving amino acids. In this case, user has to provide the atom interval for the crystal structure to be analysed. This atom interval is provided in the PDB files preceded by the word “ATOM”. This option is also useful when the user wishes to analyse only part of the molecule.

6.8) Significance test

In this option, user is allowed to identify important or significant coevolutionary amino acid site pairs by one of two ways, either identify those pairs showing a correlation coefficient above certain user pre-specified threshold (option 0) or alternatively identify significant correlation coefficients compared to a distribution of correlation coefficients for pairs of sites randomly sampled from the multiple sequence alignment (option 1).

6.9) Threshold R value

This option stands for the minimum R-value that a pair of coevolving sites should present to be selected in the final list of groups of coevolution. This option is only taken into consideration if the significance test option is set to 0.

6.10) Time correction

The time correction section permits the user to specify whether a correction for the divergence time between a pair of sequences should be introduced on the calculations as specified in Fares and Travers (2006) (option 1); or whether time should not be introduced in the calculations (option 0). The advice here is to allow for the correction of the coevolutionary parameters by the divergence time between pairs of sequences. Only if the sequences being considered belong to a very recently diverged group of taxa can the time correction be avoided. Time-correction of the transition between amino acids is also possible when sequences have not diverged dramatically (for example, in cases where the nucleotide substitutions per synonymous site do not exceed 1.25, or when Poisson-corrected amino acid changes do not exceed 1).

6.11) Time estimation

This section allows correcting the time by using nucleotide substitutions per synonymous site or alternatively Poisson-corrected amino acid substitutions per amino acid site. Closely related sequences allow for the use of synonymous nucleotide substitutions as a proportional measure of the divergence time between two sequences as long as there is no codon bias or selection on synonymous sites (option 0). If sequences are too divergent as to consider synonymous substitutions, please do use Poisson corrected distances instead (option 1). Assuming that only a small fraction of sites are under strong selective constraints at the amino acid level, on average, Poisson corrected amino acid distances can be proportional to the divergence time between two sequences. This assumption holds specially when we assume the neutral fixation of amino acid replacements.

6.12) Threshold alpha value

In here, the user specifies the threshold type I error acceptable to detect significant covarying amino acid site pairs. Because of the multiple testing problem, the user is advised to specify 0.001. At this level, the amount of possible false positives is negligible.

6.13) Random sampling

This option indicates the number of pseudo-replicates to be sampled from the multiple sequence alignment to generate a distribution of randomly sampled pairs of sites and hence of correlation coefficients. This distribution will be used by CAPS to allocate the correlation coefficient of the coevolving pairs of sites and thus to determine their significance. The user is advised to use a minimum of 10,000 pseudo-replicates and normally it would be more accurate using approximately 10^6 pseudo-replicates.

6.14) Gaps

This option is important to determine the weight that gaps have on the detection of coevolving pairs. Option 0 will remove all columns with gaps; option 1 will allow removing columns presenting a certain proportion of gaps and option 2 will consider columns with gaps as characters of the alignment. When specifying the proportion of gaps allowed in one column (amino acid site), this proportion should not go beyond 25%, although above 30% no amino acid sites are possibly detected as coevolving with other sites in the multiple sequence alignments. Gaps here are only considered at the amino acid multiple sequence alignment.

6.15) Minimum R

The user also has the option of specifying a minimum acceptable correlation coefficient value for the pairs of amino acids considered for the coevolutionary analysis despite the fact that re-sampling had been considered as the method of detecting significant correlation coefficients. This option can be used for example to identify highly correlated pairs of sites among the list of significant covarying amino acid sites.

6.16) GrSize

This option stands for the size (in number of amino acid sites) of the groups of coevolution allowed. This actually performs a first filter to ameliorate the problem of phylogenetic coevolution introducing noise and hampering the detection of functional coevolution. Tillier and Lui (Tillier and Lui, 2003) proposed as a solution to disentangle phylogenetic co-evolution from real co-evolution taking small groups of co-evolution since one important site is most likely to co-evolve with few others. This option is therefore based on the assumption that functional groups of coevolution can only include a small number of amino acid sites, while phylogenetic groups of coevolution include a vast number of sites (see Tillier and Lui 2003, for details). Generally, 3 to 5% of the protein sequence length is enough as to tackle the problem of phylogenetic coevolution. The number in this option indicates the maximum percentage of sites from the protein that a group of coevolution should have to be selected as such. For example, in an amino acid alignment with 200 amino acids, 3 would indicate that the maximum size of the coevolution group should be 6 amino acids (0.03 * 200).

7) Removing the phylogenetic coevolution and testing functional shifts (*CladesCaps.pl*)

CAPS also includes a sub-program able to automatically remove the phylogenetic coevolution in the multiple sequence alignment. This program only requires specifying the sequence names from each clade to be removed and re-do the coevolution calculations. These clades can be specified in the file “Clades” and are defined following the phylogenetic relationships.

In the phylogenetic tree if figure 6 for example we can for use as clades to test and identify functional coevolution and phylogenetic shifts on the coevolutionary pairs the clades:

*R.padi*PS *S.graminum*PS *A.pisum*PS *S.avenae*PS

*T.caerulescens*PS *C.leucomelas*PS

*T.salignus*PS *T.suberi*PS

E.coli S.typhimurium

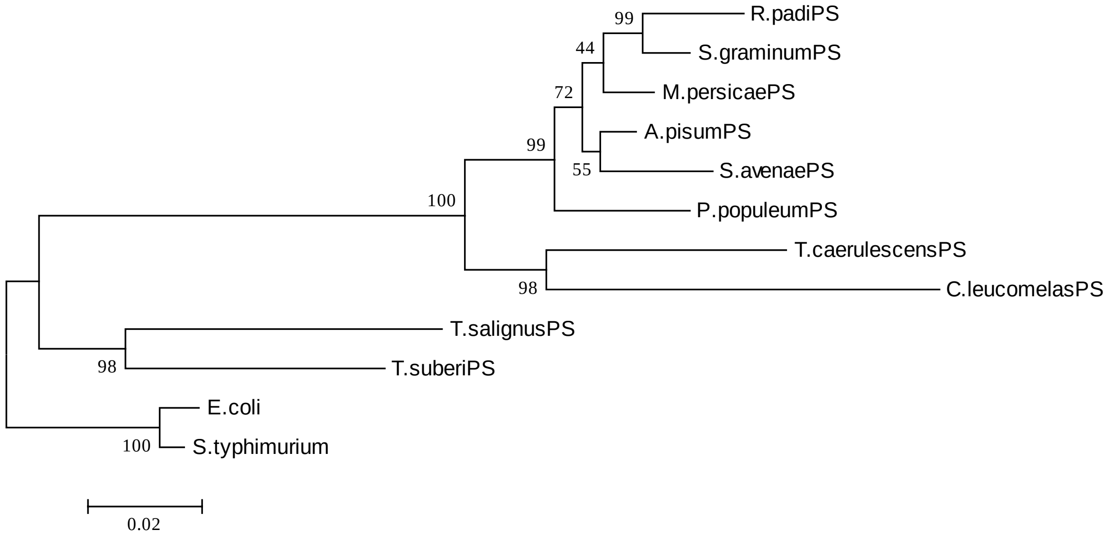


Figure 6. Minimum evolution based tree inferred using Poisson-corrected amino acid distances for the heat-shock protein GroEL from the primary symbiont (PS) of the pahid insects *Ropalosiphum padi, Schizaphis graminum, Myzus persicae, Acertosiphon pisum, Sitobion avenae, Pterocomas populeum, Tetraneura caerulescens*, *Chaitophorus leucomelas*, *Tuberolagnus salignus* and *Telaxis suberi* and the free-living relatives *Escherichia coli* and *Salmonella typhimurium.*

These four clades should be stated in this very same format in the file Clades. The user will then run CladesCaps.pl by typing:

CladesCaps.pl (Unix/Linux)

Perl CladesCaps.pl (Windows)

The program CaldesCaps.pl generates four files each one containing the multiple sequence alignment except the sequences specified in each clade for which the file has been generated. CladesCaps.pl calls the program CAPS.pl to run the coevolution analyses for the entire multiple sequence alignment file and for each one of the sub-alignment files. When analysing each one of the files, CladesCaps reads the options of the control file CladesCaps.ctl, which are equivalent to those in the control file CAPS.ctl. The user has only to set the options for CAPS.ctl. Results of coevolution are then collected from the five files and cross-compared. Those groups of coevolution present no matter what sequence or sequences are removed will be highlighted as groups of functional coevolution and printed to a '.csv' file called "FuncStrucPairs.csv".

User must note that clades are defined in terms of their biological significance or bootstrap support (phylogenetic support). Based on this premise, the user can define as many clades as possible as long as the minimum number of sequences left in the alignment is 10 (as specified in Fares and Travers 2006).

8) A case study: Coevolution in the heat-shock protein GroEL from endosymbiotic bacteria of the aphid insect

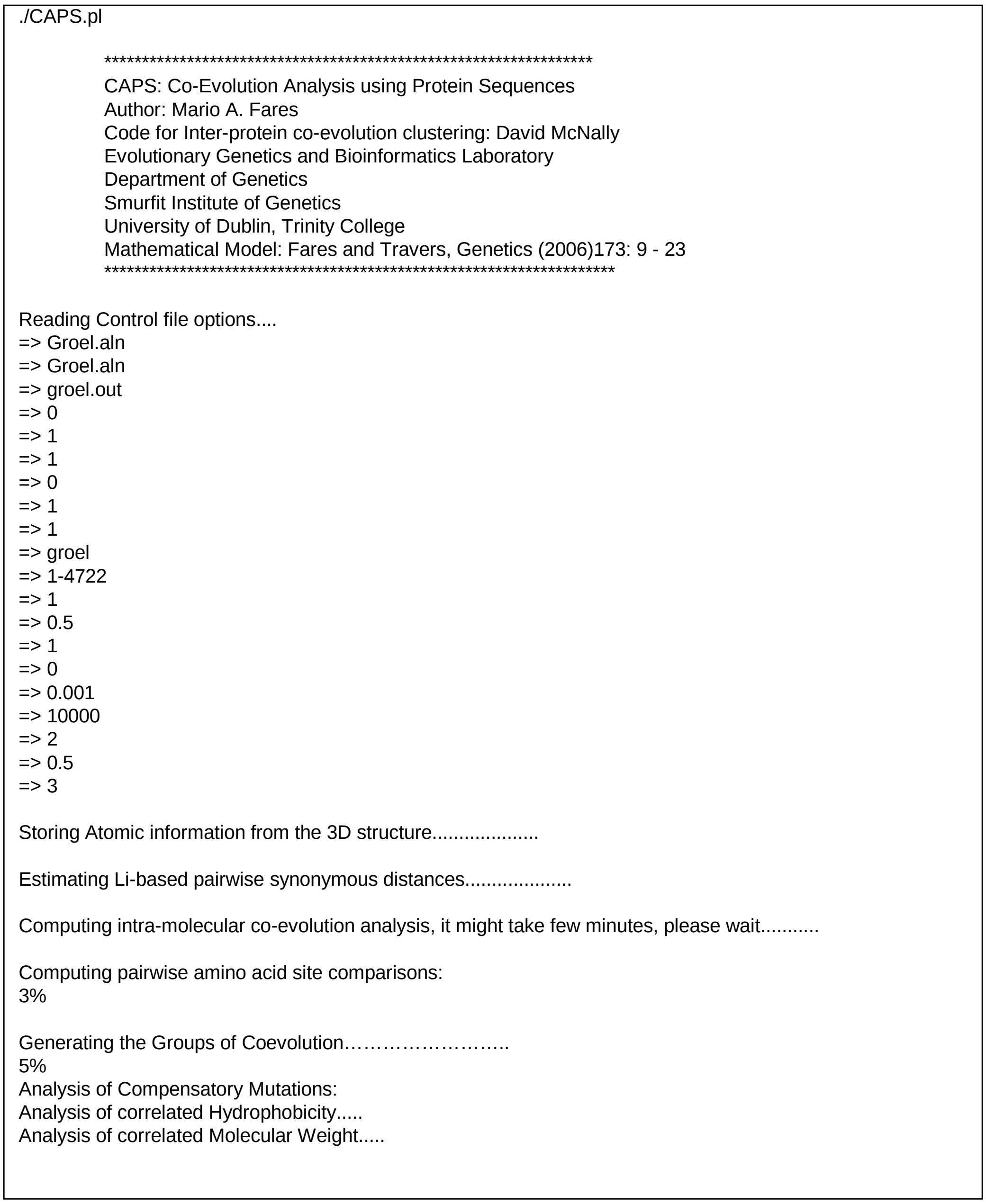
As an example of the output information generated I have run CAPS.pl and CladesCaps.pl on the data set from Figure 6. Here I defined the same clades as those previously indicated and I provided the three-dimensional structure, which is available for GroEL from *E. coli*. For illustrative purposes I am going to show the results generated by CAPS.pl without specifying clades. Note that when clades are considered, the program CladeCaps.pl should be run instead of CAPS.pl. The control file options are as specified in figure 7.

```
Input file1: Groel.aln           * File containing sequence alignment for the first protein
Input file2: Groel.aln           * File containing sequence alignment for the second protein
Out file1: groel.out             * File where the output information should be stored
Co-evolution analysis: 0         * (0) Intra-molecular; (1) Inter-protein
Type of data 1: 1                * (0) amino acid alignment; (1) codon-based alignment
Type of data 2: 1                * (0) amino acid alignment; (1) codon-based alignment
3D test: 0                       * Only applicable for intra-protein analysis: (0) perform test; (1) Test is not applicable
Reference sequence 1: 1          * the order in the alignment of the sequences giving the real positions in the protein for 3D analyses
Reference sequence 2: 1          * the order in the alignment of the sequences giving the real positions in the protein for 3D analyses
3D file: groel                   * name of the file containing 3D coordinates
Atom interval: 1-4722            * Amino acid atoms for which 3D coordinates are available
Significance test: 1             * (0) use threshold correlation; (1) random sampling
Threshold R-value: 0.5           * Threshold value for the correlation coeficient
Time correction: 1               * (0) no time correction; (1) weight correlations by the divergence time between sequences
Time estimation: 0               * (0) use synonymous distances by Li 1993; (1) use Poisson-corrected amino acid distances
Threshold alpha-value: 0.001     * This option valid only in case of random sampling
Random sampling: 10000           * Use in case significance test option is 1
Gaps: 2                          * Remove gap columns (0); Remove columns with a specified number of gaps (1); Do not remove gap columns (2)
Minimum R: 0.5                   * Minimum value of correlation coeficient to be considered for filtering
GrSize: 3                        * Maximum number of sites in the group permitted (given in percentage of protein length)
```

Figure 7. Control file option for the example GroEL.aln

Once the program is running, the user will get information of the process of the coevolutionary analyses. For example, information about the authors of the program, options stated in the control file, the processes being executed, etc. A summary of the information the user gets when running CAPS.pl is depicted in figure 8.

In the file groel.out, an example of the different subsections is depicted as explained in section 6.2 of this manual. It is interesting to see that the correlation coefficients are very high and that different types of coevolution are shown. Some coevolutionary pairs are very close in the three-dimensional structure whereas others are more distant than expected. Plotting these pairs into the three-dimensional structure also gives a flavour of the type of coevolution for each pairs of amino acids (Figure 9). When one of the amino acids involved in the coevolving pair is not included in the crystal structure of the protein, the output show a distance of 9999, as illustrated in the output groel.out.



```
./CAPS.pl

          ******************************************************************
          CAPS: Co-Evolution Analysis using Protein Sequences
          Author: Mario A. Fares
          Code for Inter-protein co-evolution clustering: David McNally
          Evolutionary Genetics and Bioinformatics Laboratory
          Department of Genetics
          Smurfit Institute of Genetics
          University of Dublin, Trinity College
          Mathematical Model: Fares and Travers, Genetics (2006)173: 9 - 23
          *********************************************************************

Reading Control file options....
=> Groel.aln
=> Groel.aln
=> groel.out
=> 0
=> 1
=> 1
=> 0
=> 1
=> 1
=> groel
=> 1-4722
=> 1
=> 0.5
=> 1
=> 0
=> 0.001
=> 10000
=> 2
=> 0.5
=> 3

Storing Atomic information from the 3D structure...................

Estimating Li-based pairwise synonymous distances...................

Computing intra-molecular co-evolution analysis, it might take few minutes, please wait...........

Computing pairwise amino acid site comparisons:
3%

Generating the Groups of Coevolution……………………..
5%
Analysis of Compensatory Mutations:
Analysis of correlated Hydrophobicity.....
Analysis of correlated Molecular Weight.....
```

Figure 8. Running CAPS.pl and the information provided during the analysis of coevolution.

As can be seen from the output file, 12 groups of coevolution are obtained. In these groups it is particularly interesting to see strong coevolution among the amino acid sites Glu338, Ala341 and Gly344, with amino acid sites 338 and 344 being considered as significantly correlated even after performing the analysis of molecular weight and hydrophobicity. These amino acids are also within the 8Å distance that would allow considering them as physically interacting. Note that the amino acid distances taken in CAPS are those referring to the distance from the centre of one amino acid to the geometrical centre of the coevolving amino acid, and is therefore a conservative distance measure.

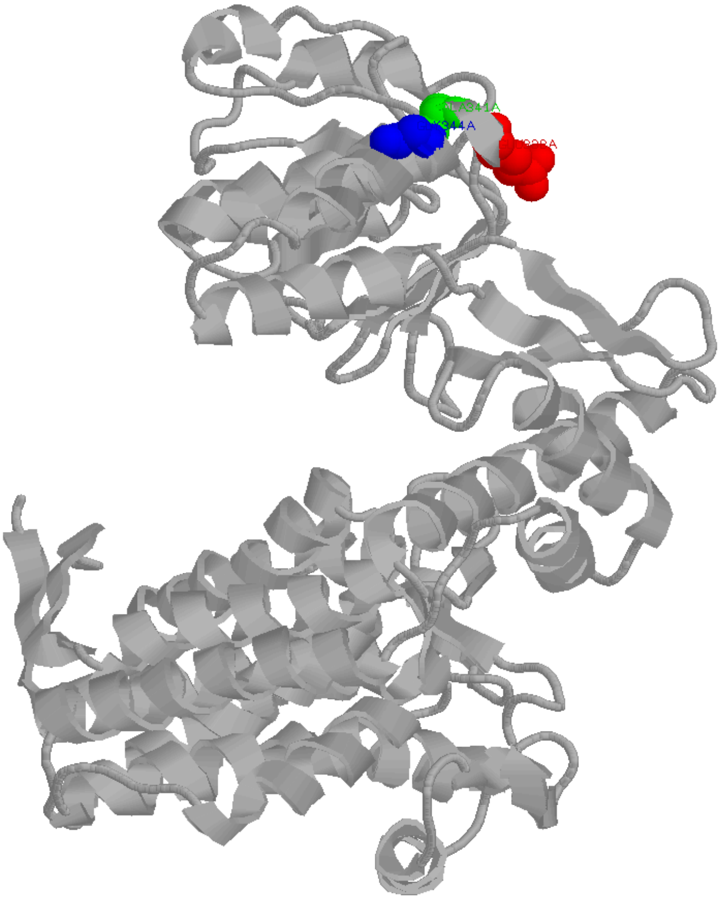


Figure 9. Three-dimensional structure of one of the subunits of GroEL showing coevolving amino acid sites. Red, Green and blue are the amino acid sites belonging to the same group of coevolution and label Glu338, Ala341 and Gly344.

To identify Functional coevolution, I defined two clades just for demonstration purposes. The sequences contained in these clades and defined in the file "Clades" where:

T.salignus T.suberi

R.padi S.graminum

After running CladesCaps.pl we obtain several coevolving pairs, an example of two of the main amino acid groups is depicted in the three-dimensional structure of Figure 10.

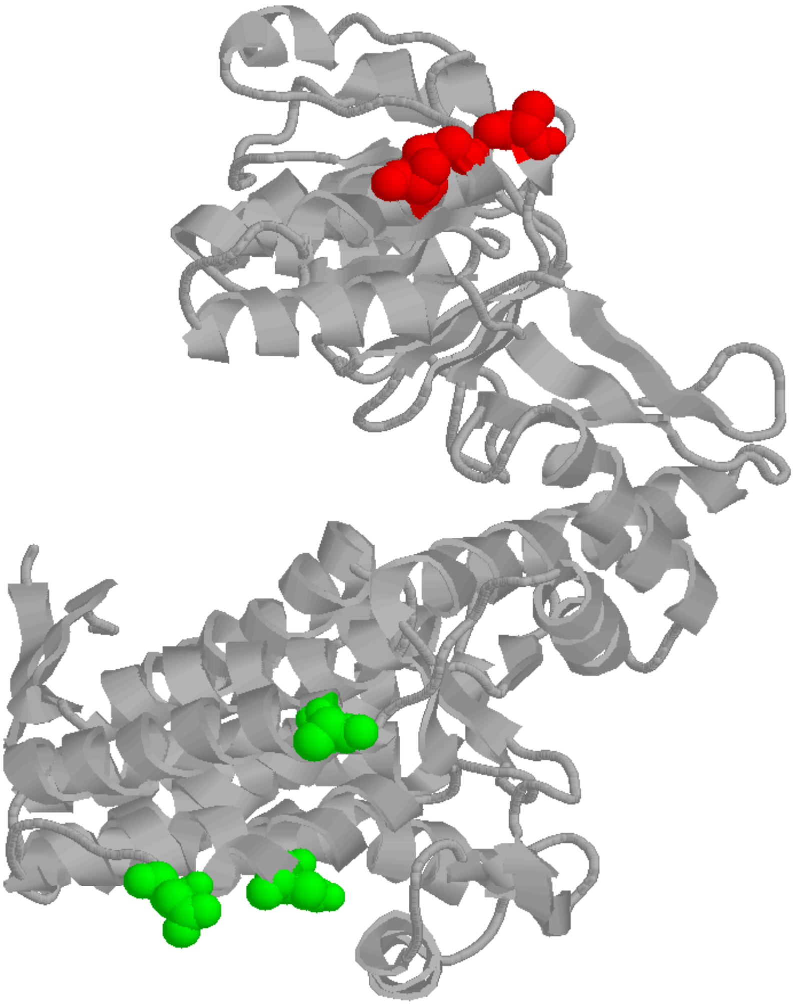


Figure 10. Three-dimensional location of two groups of coevolving amino acid sites. The groups are labelled differently. Green labels amino acid sites 133, 424 and 428, whereas red indicates the group containing the amino acids 344 and 347.

9) Files and executables in CAPS

CAPS is organised into four main types of files:

a) Perl executable files: these files have the extension .pl and are:

a.1) CAPS.pl: this is the main program

a.2) CladesCaps.pl: this program runs CAPS.pl and identifies functional coevolving sites as described above.

b) The control files

b.1) CAPS.ctl: contains the main options to run CAPS.

b.2) CladesCaps.ctl: This file is identical to CAPS.ctl and is used by CladesCaps.pl to modify the control file and run the program CAPS.pl with the different automatically generated subfiles.

c) Perl Modules

This is the most extensive type and contains al the modules with all the subroutines required for running CAPS.pl and CladesCaps.pl automatically. The different module files are:

c.1) caps_module.pm

c.2) CompMut.pm

c.3) File_reader.pm

c.4) Dimen.pl

c.5) Statistics.pm

c.6) Distances.pm

c.7) Automatic_caps.pm

c.8) Seq_manag.pm

c.9) Inter_sort.pm

d) Finally, CAPS contains flat files that provide information about:

d.1) Hydrophobicity: Hydrophobicities for the amino acids as well as their Molecular weights

d.2) BLOSUM: Blosum values for the transition between amino acids

d.3) Clades: Contains the species names that are going to be used for CladesCaps.pl, with each clade specified as a line in the file.

d.4.) Zscores: contains a table with the scores for the normal distribution.

10) **REFERENCES**

Fares, M.A. and Travers, S.A. (2006) A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses, *Genetics*, **173**, 9-23.

Gloor, G.B., Martin, L.C., Wahl, L.M. and Dunn, S.D. (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions, *Biochemistry*, **44**, 7156-7165.

Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks, *Proc Natl Acad Sci U S A*, **89**, 10915-10919.

Lockless, S.W. and Ranganathan, R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families, *Science*, **286**, 295-299.

Pritchard, L. and Dufton, M.J. (2000) Do proteins learn to evolve? The Hopfield network as a basis for the understanding of protein evolution, *J Theor Biol*, **202**, 77-86.

Suel, G.M., Lockless, S.W., Wall, M.A. and Ranganathan, R. (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins, *Nat Struct Biol*, **10**, 59-69.

Templeton, A.R., Maxwell, T., Posada, D., Stengard, J.H., Boerwinkle, E. and Sing, C.F. (2005) Tree scanning: a method for using haplotype trees in phenotype/genotype association studies, *Genetics*, **169**, 441-453.

Tillier, E.R. and Lui, T.W. (2003) Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments, *Bioinformatics*, **19**, 750-755.

Westfall, P.H. and Young, S.S. (1993) Resampling-based multiple testing, John Wiley & Sons, New York.