C.A.P.S. *v*2.0 MANUAL

Contents

1. Suggested Citation

Awaiting publication

2. History

CAPS version 2.0is the second version of 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 C++ and runs in various operating systems including Linux/Unix, Windows(needs Cygwin) and Mac Os X. The software implements an improved method to detect **intra**-molecular coevolution as published in Genetics (Fares and Travers, 2006a), improved versions of **inter**-molecular coevolution and analysis of compensatory mutations documented in Bioinformatics(Fares and McNally, 2006) and additional methods to predict whether pairs of proteins are interacting and create user friendly output to be mapped to crystal structures of a protein.

3. Introduction

CAPS version 1.0 and its model have been used in many studies (Fares and Travers, 2006b; Travers et al, 2007;Huang et al 2011), cited many(80-90) times and our caps server receives roughly 800 queries a year. The updated version aims to improve the analysis by

(1) increase the speed of coevolutionary analyses through optimised implementation of the program in C++ making it applicable to proteome-wide analyses in a reasonable time interval
(2) implements an explicit test of coevolution between two proteins
(3) phylogeny between sequences is accounted for and lineage specific contribution of coevolution is measured

4. Method

4.1. **Overview.** Below(Figure 1) is an overview of the workings of the program.

4.2. **Phylogenetic Relationships.** The choice of whether to provide a phylogenetic relationship or not is up to the user.

Calculated trees:

If the user decides not to provide a tree for the analysis then the software will calculate a phylogenetic relationship. Trees are built under the JTT model of protein sequence evolution(Jones et al, 1992) and using the BIONJ(Gascuel, 1997) tree building method.

User defined trees:

The user may provide a phylogenetic relationship in Newick format which has been calculated from any tree building software, e.g., RAxML, PhyML.

4.3. **Ancestral Reconstruction.** For each alignment tested the ancestral sequences are reconstructed. The amino acids on the inner nodes are predicted using a maximum likelihood approach under the JTT model of protein sequence evolution which is implemented with the use of the Bio++ libraries(Dutheil et al, 2006). Figure 1 and 2 show how this will lead to improved scores for amino acid sites. It is shown that when we use this approach there is only one radical transition which gives a Pearson's correlation coefficient of $r = 0.85$. In Figure 2 we show the ancestral reconstruction can lead to a much different scenario, here we see that there are many radical substitutions. Using all of these transitions correctly yields a higher Pearson's correlation coefficient of $r = 0.96$. The comparison of these two examples outlines the necessity to test according to this procedure, even though the majority of the ancestrally reconstructed sites are identical, the number of radical changes is very different.

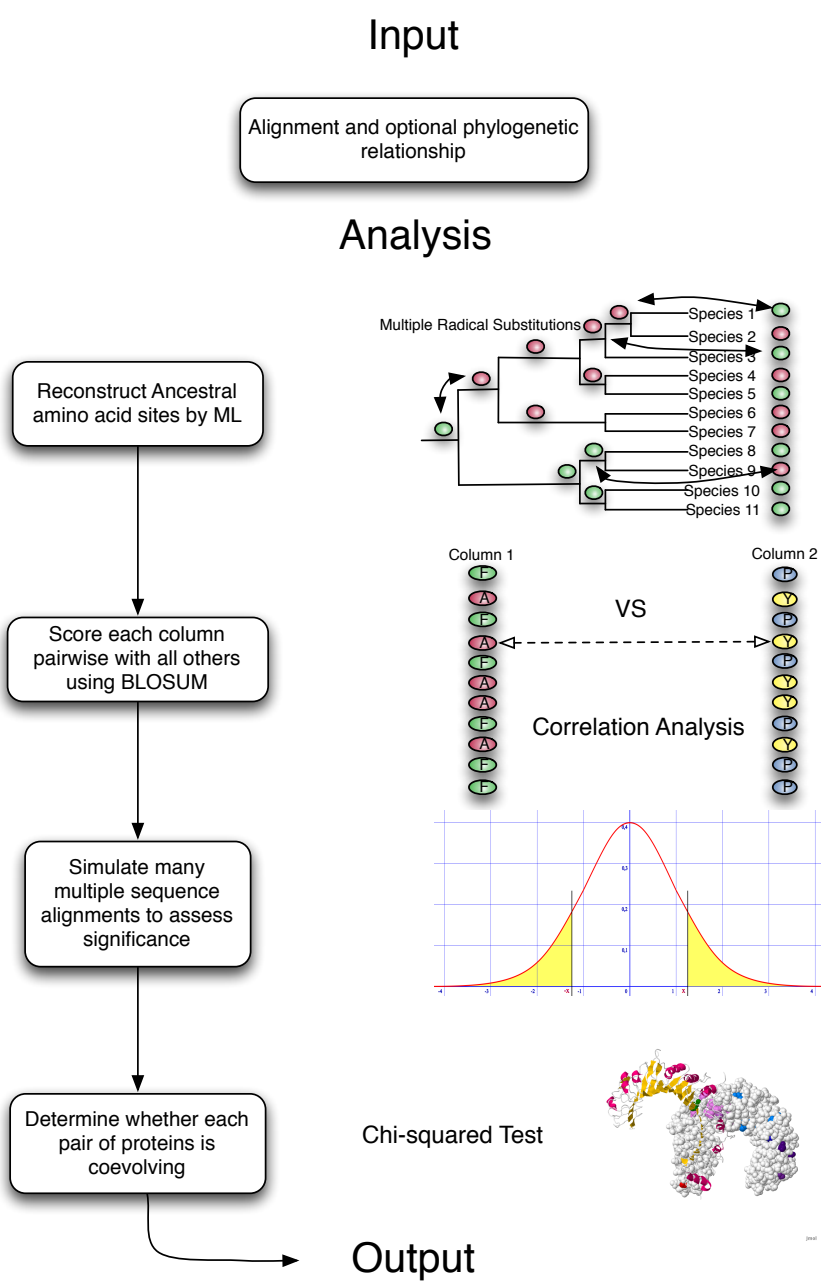

FIGURE 1. Firstly ancestral sites are reconstructed and BLOSUM corrected values are applied. Following this a correlation coefficient is performed. Simulations are carried out to assess significance. Finally each pair of proteins are tested to see if the levels of coevolution are significant to predict interaction.

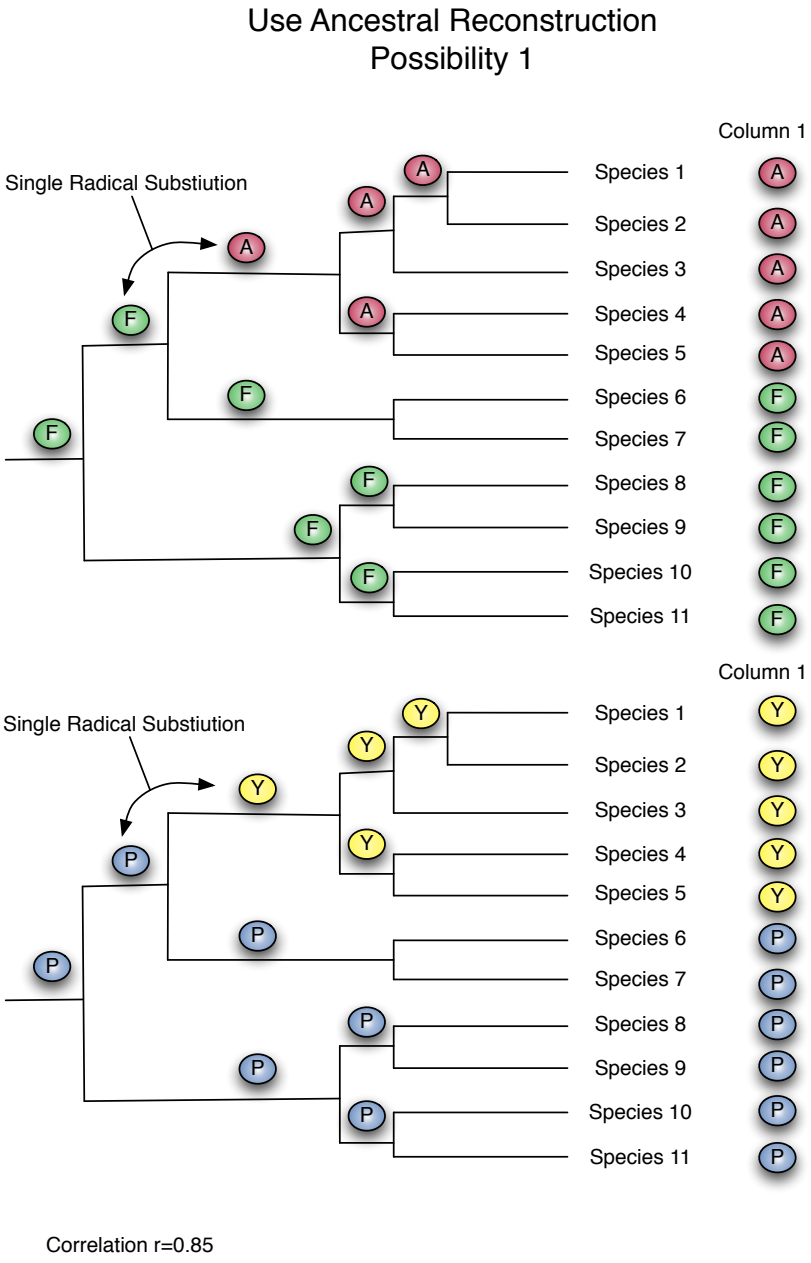

FIGURE 2. Ancestral reconstruction shows only one radical substitution and a raw Pearson's correlation coefficient of 0.85

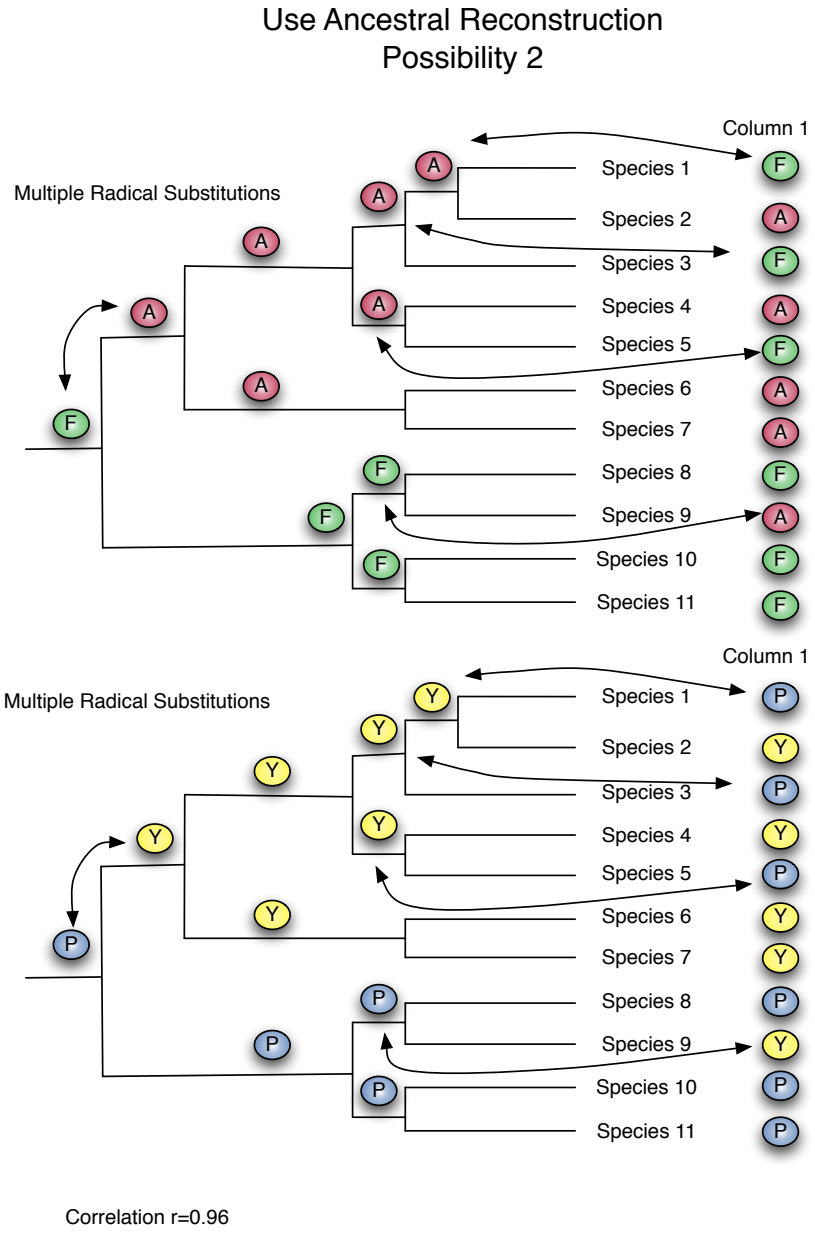

FIGURE 3. Ancestral reconstruction shows only one radical substitution and a raw Pearson's correlation coefficient of 0.96. This value is considered much more significant than the example shown in Figure 3.2.

4.4. **Scoring Coevolution.** Following the parsing/calculation a phylogenetic tree and ancestral reconstruction of sequences at all inner nodes, we first calculate the mean transition parameter θ for each column, given by:

$$\bar{\theta} = \sum_{S=1}^{n} (\theta_{ij})_S \tag{1}$$

Where θ_{ij} is the transition score given by the weight matrix for an amino acid substitution from state i to state j and n is the number of transitions on the phylogenetics tree in each column. The next step is to calculate the variability of each amino acid transition compared to the mean transition parameter. This is given by:

$$D_{ij} = [(\theta_{ij} - \bar{\theta}]^2 \tag{2}$$

Correlation coefficients are then calculated according to Pearson's correlation coefficient ρ_{AB} :

$$\rho_{AB} = \frac{\sum_{S=1}^{n} [(D_A)_S - \bar{D_A}][(D_B)_S - \bar{D_B}]}{\sqrt{\sum_{S=1}^{n} [(D_A)_S - \bar{D_A}]^2} \sqrt{\sum_{S=1}^{n} [(D_B)_S - \bar{D_B}]^2}} \tag{3}$$

The values of ρ_{AB} are kept for significance testing either by using a threshold cut-off supplied by the user or a cut-off evaluated by simulations performed by the program.

4.5. **Simulations and Significance test.** For each alignment tested, simulated sequence alignments are created using the JTT model(Jones et al, 1992) according to the gene-specific phylogenetic tree calculated above or specified at input. That is, the distance matrix for the simulated data is equal to that of the real data. The simulated alignments

are tested according to the equations above, resulting in a null distribution of the test score against which P-values for the real data are calculated. Following this, the values of ρ_{AB} which are below the cut-off value are disgarded.

4.6. **Prediction of Interactions.**

Chi-squared tests Three chi squared tests are carried out on the results of multiple alignment runs of **inter**-molecular coevolution. Each of which use the following equation.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{4}$$

Where O_i is the observed frequency of pairs of sites predicted as coevolving, E_i is the expected frequency and n is the number of possible outcomes of each event. We use these tests to identify whether each pair of alignments is interacting and therefore possibly involved in protein-protein interactions. The Expected frequencies are the only value to change in each of the three tests.

(1) We simulate 100 pairs of alignments which have the same relative distances on the phylogenetic trees as the input sequences and use the mean number of coevolving pairs as the expectated value(E) in equation 4.
(2) We calculated the mean number of pairs of coevolving sites for all pairwise **inter**-molecular analysis and use this number as the expectated value(E) in equation 4.
(3) We calculated the mean number of pairs of coevolvings sites for all pairwise **inter**-molecular analyses performed for each protein. i.e. mean number of sites for Protein A versus all other Proteins. Again this is use as the expectated value(E) in equation 4.

5. Obtaining and installing

The source code and a series of binaries is available on the lab website at: http://bioinf.gen.tcd.ie/ faresm/softwar

Binaries need not be installed just run using the options documented in the Running section below.

5.1. **Prerequisites for Compiling from source.**

(1) gcc(comes packaged in xcode tools on mac)
(2) Bio++
(3) gnu scientific library

With these prerequisites installed on a terminal change directory to where you downloaded the package using *cd folderpath*(perhaps Downloads).
Next unzip the package using *unzip caps2.zip*
Next change directory to the extracted folder using *cd caps2*
Finally compile using *make*
Continue to the next section for instructions on running the program and viewing results.

6. Running

6.1. **Input.** The minimum one needs to run **intra**-molecular coevolution analysis is a multiple sequence alignment in Fasta Format. The minimum one needs to run **inter**-molecular coevolution analysis is two multiple sequence alignments in Fasta Format with at least **some** common species names. Files to be analyzed should be placed in a folder together.

Optionally a set of corresponding pdb files can be supplied. Read section on preparing

pdb files.

NOTE: species names should be **EXACTLY** the same in alignments to be tested against each other.

6.2. **Running options.** There are multiple ways to run the analysis, examples are given below:

- ./caps

  This will print a splash of the options available.

- -F foldername/

  This option should be used to run analysis on a folder of files.

  e.g. ./caps -F sample_folder/

  Results in **intra**-molecular coevolution analysis on all files in the folder "sample_folder".

- - -inter (double minus)

  This option toggles intra(default) analysis to **inter**-molecular coevolution analysis.

  e.g. ./caps -F sample_folder/ - -inter

  Results in **inter** molecular coevolution analysis on all files in the folder "sample_folder".

- - -intra (double minus)

  This option toggles analysis to **intra**-molecular coevolution analysis. This is somewhat redundant since the program defaults to intra anyway but is here for clarity.
  e.g. ./caps -F sample_folder/ - -intra
  Results in **intra**-molecular coevolution analysis on all files in the folder "sample_folder".

- -S structure_folder/

  This option allows the user supply a folder with pdb files corresponding to the alignments used with the -F option.
  e.g. ./caps -F sample_folder/ –inter -S structure_folder/
  Results in **inter**-molecular coevolution analysis with a set of corresponding pdb structures. **NOTE:** Refer to section of preparing pdb files.

- -r

  This option allows the user to specify the number of random cycles that should be carried out. The default number is one hundred simulated alignments of the same size and phylogenetic distribution as the input alignments.
  e.g. ./caps -F sample_folder/ –inter -r 1000
  Results in **inter**-molecular coevolution analysis with one thousand simulations carried out. **NOTE:** Altering the default number can **drastically** alter runtime, the default value should be ok in most circumstances.

- -a

  This option allows the user to specify an alpha value for threshold cut-off. The default value is 0.001. e.g. ./caps -F sample_folder/ -a 0.0001

  Results in **intra**-molecular coevolution analysis the results of which will have a P-value less than 0.0001.

- -T

  This option allows the user to specify a folder of Newick formatted tree files, either with or without distances. If there are distances supplied then simulations alignments will have pairwise distances equal to those on the tree. If there are no distances on the tree then the program will attempt to put distances on the tree keeping the topology fixed.

  e.g. ./caps -F sample_folder/ -T mytreefolder

  Results in **intra**-molecular coevolution analysis calculated according to the phylogenetic relationship contained in the file newick.nh.

6.3. **Output files and analysis.** For each file or pair of files the following output files will be provided.

6.3.1. *Intra.* Inputfile.fa.out

Inputfile.fa.csv

Inputfile-overlap.csv

In the .out file you will find a list of all of the coevolving pairs of amino acid sites, groupings of all of the pairs, i.e. with residue pair 1 & 2 and pair 1 & 4 and pair 2 & 4 we get a group of 1 & 2 & 4. If a pdb file was provided, the distances between pairs is given. There is also a list of overlapping groups, i.e. with residue pair 8 & 10 and pair 10 & 12 we get

an overlapping group of 8 & 10 &12. In this case whilst residue 8 & 12 are not directly coevolving there is some indirect pressure to coevolve.
The .csv files contains residue annotations of groups of coevolving residues which can be mapped to jmol via pfaat. The first contains residue annotations for the non-overlapping groups. The second for the overlapping groups.

6.3.2. *Inter.*
Inputfile1.fa_inputfile2.out
Inputfile1.fa_inputfile2.csv
Inputfile2.fa_inputfile1.csv

In the .out file you will find a list of all of the coevolving pairs of amino acid sites, groupings of all of the pairs, i.e. with pair 1 & 2 and pair 1 & 4 and pair 2 & 4 we get a group of 1 & 2 & 4. If a pdb file was provided, the distances between pairs is given. There are also overlapping groups listed, i.e. with residue pair 8 & 10 and pair 10 & 12 we get an overlapping group of 8 & 10 &12. In this case whilst residue 8 & 12 are not directly coevolving there is some indirect pressure to coevolve.
The Inputfile1.fa_inputfile2.csv file contains residue annotations of groups of coevolving residues for Inputfile1.fa which can be mapped to jmol(structures) via pfaat, or simply viewed in pfaat.
The Inputfile2.fa_inputfile1.csv file contains residue annotations of groups of coevolving residues for Inputfile1.fa which can be mapped to jmol(structures) via pfaat, or simply viewed in pfaat.
The Inputfile1.fa_inputfile2-overlap.csv and Inputfile2.fa_inputfile1-overlap.csv contain residue annotations of overlapping groups of coevolving residues for Inputfile1.fa and Inputfile2.fa

respectively.

6.4. **Visualisation of Groups via pfaat.** pfaat can be launched directly from your browser by going to this link:
http://pfaat.sourceforge.net/webstart/pfaat.jnlp

To view the alignment and structure with residue annotations mapped follow the instructions below, figures are at bottom of pdf.

7. Preparing pdb files.

This should be considered very important if one wants to visualize/view the correct information on pdb structures. With this in mind we suggest the following preparation for pdb files.

(1) Firstly make a copy of the pdb file, we don't want to lose the original and have to go download it again
(2) Now open the pdb in your favourite editor, vim, emacs, textedit etc.
(3) Find all lines which are not of use to the user. By this we mean if there are multiple chains in a dimer, say, with chains A & B, find the lines which begin with ATOM and identify the chain corresponding to the alignment.
(4) Now delete all lines which don't pertain to the alignment, i.e. the other lines. Its ok to do this because we saved the original and these are the only lines needed for distance calculations
(5) Now save the file as myfile.pdb, where myfile.fa is the fasta aligned file which corresponds to this pdb file.

(6) When viewing the structure use the ORIGINAL file with all the lines.

8. Example Run

Whichever version you downloaded, either source or binary it should have come packages with a folder called sample_folder/ and structure_folder/. These can be used as example files to test the running of the program, see how it works and as examples of the input needed for the program.

For example you could run **intra**-molecular coevolution analysis of the three alignments in sample_folder by running:

./caps -F sample_folder/