This software provides code to accompany:

Covariance in protein multiple sequence alignments using groups of columns as the fundamental unit for computation - (submitted)

You will need

(1) A recent version of Java (at least 1.6)
(http://www.oracle.com/technetwork/java/index.html)

A recent version of R for the sampling correction in the last step. This can be accomplished with a number of other packages, but we wrote a custom script in R to do that work for you.
(http://www.r-project.org/)

(2) The McBasc algorithm and COBS algorithm has a requirement for the file Maxhom_McLachlan.metric. This file should be in the subdirectory "data" where you unzipped the distribution zip file.

(3) There is a file in the src directory of the distribution called Energetics.properties. This file needs to be in a directory in your classpath and needs to be edited so that it points to the correct place on your hard drive

(4) You'll need to edit the Energetics.properties file so that the following lines are defined:

```
# this can be downloaded from
#ftp://ftp.sanger.ac.uk/pub/databases/Pfam/current_release/
# alternatively, run scipts.WriteTrunctatedPfamAlignment that will produce a
#file that has only the alignments that COBS will be run over
# (That have >200 sequences, less than 2000 and a matching PDB file).
#FULL_PFAM_PATH=D:\\Pfam\\Pfam-A.full
FULL_PFAM_PATH=c:\\temp\\pfamTruncated.txt.gz

# where the COBS code is installed
COBS_HOME_DIRECTORY=d:\\KyleCleanroom

# A directory to hold PDB files
# PDB_DIR=d:\\pdbDir

# this file is included in the distribution
PDB_PFAM_CHAIN=D:\\git\\cobs\\cobs\\pdbToPfamForCobsViaBlast.txt.gz

# these are where the result files will be held
COBS_CLEANROOM=D:\\COBS_OUT
```

The pfam families that are mapped in our paper are in a file called pdbToPfamForCobsViaBlast.txt.gz included in the distribution. You should set PDB_PFAM_CHAIN in the properties file above to point to this file.

(5) Once this has been done if you run

java scripts.DownloadPDBFiles

from the cobs directory in where you downloaded the code, the PDB files that are needed will be downloaded from the PDB database.

(6) Once the PDB files that have been installed, the next step is to generate pairwise (one-D) scores from McBASC. This can be done with

java cobsScripts.WriteOneDScores

Within the "COBS_CLEANROOM" directory there will be created a "results" sub-directory and then a "oneD" directory. The results for pairwise McBASC will be put in this directory. (By uncommenting the appropriate lines in WriteOneDScores.main(), oneD scores for ConservatonSum, MICovariance and a random control can be generated as needed).

(7) After the oneD scores have been generated, run

java cobsScripts.WriteScores

For each protein family the following files will be placed in the results directory within COBS_CLEANROOM

`AverageAbsMcBASC.txt.gz` – the average from the absolute value of McBASC

`AverageAbsMcBASC_PNormalInitial.txt.gz` – the average from the absolute value of McBASCp

`AverageConservationSum.txt.gz` – The average from conservation sum

`AverageConservationSum_PNormalInitial.txt.gz` – The average from conservation sum pnormalized

`AverageMI.txt.gz` – The average from MI

`2OG-FeII_Oxy_5_AverageMI_PNormalInitial.txt.gz` – The average from MI pnormalized

`2OG-FeII_Oxy_5_AverageMcBASC.txt.gz` – The average McBASC score

`2OG-FeII_Oxy_5_AverageMcBASC_PNormalInitial.txt.gz` – The average McBASC p normalized

`2OG-FeII_Oxy_5_Averagerandom.txt.gz` – A control with scores sampled from the uniform distribution

`2OG-FeII_Oxy_5_COBS_UNCORRECTED.txt.gz` – The cobs score.

Note that the normalization in the above files is applied at the pair of column level not the GOC level. So for example, MI scores are generated and then MIp scores are formed based on phylogenetic correction for each pair of columns and then the average is taken of the MIp scores. In principle, phylogenetic correction could also be applied at the GOC level. This is done in the next step.

(8) To generate ROC curves, start by combining all of the scores across each protein family into single files for each algorithm. This is accomplished by:

java cobsScripts.AbsoluteScoreVsAverageDistance

In the "bigSummaries" folder in COBS_OUT there will be 4 different files for each algorithm.

For example for McBASC:

`bigAverageMI.txt` – Results from average MI with no phylogenetic correction

`bigAverageMI_PNormalInitial.txt` – Results from averaging MIp scores applied at the pair of column level.

`bigAverageMInormedLate.txt` – Results from averaging uncorrected MI and then applying phylogenetic correction at the level of GOC.

`bigAverageMI_PNormalInitialnormedLate.txt` – Results from averging MIp and then again applying phylogenetic correction at the level GOC.

Note that since COBS is only applied at the average level, there are only 2 files for COBS:

`bigCOBS_UNCORRECTED.txt`

`bigCOBS_UNCORRECTEDnormedLate.txt`

since COBS doesn't take an average of pairs of columns, the phylogenetic correction cannot be applied prior to taking the average.

(9) To generate ROC curves run:

java cobsScripts.RocOnBigTable

(10) Finally, to sample the ROC curves, so that combining of the data does not have a bias for number of data points per family, we use an R script that needs the first variable changed to that of the output from step #9. The script is called "Flatten_ROC_Curves.R" and can be found in the same directory as this documentation

and the Energetics.properties file.

This script samples each ROC curve, and produces an identical X-axis (True Positive) for all ROC curves produced in step #9. Namely, the variable:

universal_Path <- '/Users/kkreth/Documents/_results/roc/'

Needs to be changed to match your environment.