# CoEv Software Manual

The goal of the document is to describe CoEv web-platform usage

The CoEv web platform provides two main phylogenetic-based methods:

1. prediction (evaluation) of coevolving positions and their evolutionary profile based on the aligned sequences and a phylogenetic tree
2. simulation of coevolving pairs of sites.

# Evaluating a coevolving score:

The evaluation method is the implementation of the model described in Dib et al. 2014, where we present a new substitution model that describes the coevolving process along the tree and reconstructs the ancestral states of coevolving pairs of positions. We propose the implementation of a probabilistic Markov model that not only identifies coevolving positions but also estimates the associated coevolving profile. The Coev Markov model instantenous matrix Q contains 4 continuous parameters and a discrete parameter. The implementation of the model is written in C.

# How to use the evaluate method using the web-platform?

Given an alignment file, the rooted phylogenetic tree in nexus format and a pair of positions the user will be able to evaluate the score of coevolution and estimate the profile with either the maximum likelihood or bayesian framework for amino acid and nucleotide sequences.

Step1: The user selects the alphabet, i.e amino-acid or nucleotide

Step2: The user uploads the aligned sequences in FASTA format\*

Step3: The user uploads the binary and rooted phylogenetic tree in newick format

Step4: The user specifies the framework to use: either Bayesian or maximum likelihood

Step5: The user specifies the positions that will be tested in position 1 and position 2 fields

Step6: The user provides its e-mail address

The user’s input is controlled immediately to verify if the corresponding field is correctly filled. If this is not the case, an error message is displayed over the form to inform the user about the issue. The form can only be submitted when all the fields are correctly completed. Note that the time to complete the job depends on the complexity of tree as well as some other parameters such as the number of iterations, the sampling frequency, etc.

Once the job is finished the user receives a mail that summarizes the maximum likelihood results or a tracer-readable file for the Bayesian results. An explanation on how to read these results is given below.

\*: we took particular care of the alignment input file where we first checked if its content is written in nucleotide or amino-acid alphabet. Then we filtered all conserved positions (>90% conservation) and gapped positions. Once processed the alignment file can be downloaded by the user to validate the filtering.

# Simulating of a coev olving pair:

The simulation method is based on the same instantaneous matrix (Dib et al., 2014). Given a rooted tree in nexus format and the values of the 4 continuous parameters, the software randomly picks a coevolving profile, a state at the root and lets this state evolve along the branches according to CoEv substitution matrix and a frequency vector that favours the coevoling states defined by the profile. The software therefore simulates nucleic and proteic pairs of positions along a tree by assigning a state composed of two letters from the amino acid or nucleotides alphabet to the leafs.

How to use the simulate method using the web-platform?

Given an alignment file, the rooted phylogenetic tree in nexus format and a pair of positions the user will be able to simulate pairs of coevolving sites.

Step1: \*The user uploads the binary and rooted phylogenetic tree in newick format

Step2: The user selects the alphabet, i.e amino-acid or nucleotide

Step3: \*\*The user specifies the 4 continuous evolutionary rates (s, d, r1 and r2).

Step4: The user set the number of pair to simulate under the same coev olving profile

Step5: The user provides its e-mail address

The user’s input is controlled immediately to verify if the corresponding field is correctly filled. If this is not the case, an error message is displayed over the form to inform the user about the issue. The form can only be submitted when all the fields are correctly completed. Note that the time to complete the job depends on the complexity of tree.. Once the job is finished the user will receive a mail with the attached simulated pairs in a fasta format.

\* if the branch length are too small you might obtain conserved positions

\*\* there is a relation between s/d ratio and coevolution (Dib et al. 2014). The smaller the ratio the more coevolving the pair is.

# Reading output file of the evaluate method:

To read the output, the user can either look at the log file or at the output console. In the log file, each line corresponds to a profile.

## *In the case of a Maximum Likelihood framework evaluation*:

## Each line can be truncated in 9 values:

Value 1 corresponds to the profile set: it is a binary vector of length 16 where 1 corresponds to a combination that belong to a profile and 0 when it doesn't. The binary vector follows the following order AA,AC,AG,AT,CA,CC,CG,CT,GA,GC,GG,GT,TA,TC,TG,TT. For example when the value is 1000010000000000 then it means that the profile is {AA,CC}.

Value 2 corresponds to the estimated value of w1 in the Null model: it is a double, e.g.:6.928542e-02

Value 3 corresponds to the estimated value of w2 in the Null model: it is a double, e.g.:2.014202e-01

Value 4 corresponds to the estimated value of Log Likelihood in the Null model: it is a doube, ex:-7.405995e+01

Value 5 corresponds to the estimated value of s in the model: it is a double, e.g.: 2.735160e-02

Value 6 corresponds to the estimated value of d in the CoEv model: it is a double, e.g.: -4.970079e-01

Value 7 corresponds to the estimated value of r1 in the CoEv model: it is a double, e.g.: 3.103843e-02

Value 8 corresponds to the estimated value of r2 in the CoEv model: it is a double, e.g.: 2.922180e-01

Value 9 corresponds to the estimated value of Log Likelihood in the of the CoEv model: it is a double, e.g.:-6.814893e+01

## *In the case of a Bayesian framework* evaluation:

The log file is conceived to be read by Tracer (<http://tree.bio.ed.ac.uk/software/tracer/>).

The first line is a header.

The other lines contain values where

Value 1 corresponds to the iteration identifier: it is an integer

Value 2 corresponds to the posterior probability: it is a double

Value 3 corresponds to the Log Likelihood value of CoEv: it is a double

Value 4 corresponds to the prior value:  it is a double

Value 5 corresponds to s: it is a double

Value 6 corresponds to d: it is a double

Value 7 corresponds to r1: it is a double

Value 8 corresponds to the r2: it is a double

The rest of the values correspond to the profile vector where 1 corresponds to a combination that belong to a profile and 0 when it doesn't. The binary vector follows the following order AA,AC,AG,AT,CA,CC,CG,CT,GA,GC,GG,GT,TA,TC,TG,TT. For example when the value is 1000010000000000 then it means that the profile is {AA,CC}.

# For local executions:

System requirements: the software CoEv uses several external tools that should be installed. We optimized the implementation using LAPACK (Linear Algebra PACKage) and nlopt (library for nonlinear optimization) downloadable at <http://ab-initio.mit.edu/wiki/index.php/NLopt> and tested it on Linux and Mac operating systems.

# Instructions for running CoEv with a command line:

To execute run:

coev [OPTIONS]

-method s: Use method s (either 'bayes', 'ml' or 'sm').

Default is 'ml: maximum likelihood'.

-data s: Use data type s (either 'nt', 'aa' ). Default is 'nt'.

-tree s: The name of the file containing the input rooted tree in Newick

format. Default is s=treeInput.txt.

-align s: When method is bayes or ml, the name of the file containing the sequence alignment in FASTA format. Default is s=alignment.txt.

-cols v1 v2: The columns in the alignment to use for the analysis.

Default is n1=1 and n2=2.

-s v: The s parameter rate. Default is v=1.

-d v: The s parameter rate. Default is v=1.

-r1 v: The r1 parameter rate. Default value is 1.

-r2 v: The r2 parameter rate. Default value is 1.

-out s: The name of the log file to write the results to. Default s=output.log.

-IT n: When method is bayes, run for n iterations. Default is n=10000.

- ns n : when the method is sm the parameters sets the number of pairs to simulate. Default value is 1.

-sfreq n: When method is bayes, write every n'th iteration to file. Default is n=1000.

-pfreq n: When method is bayes, print every n'th iteration on the screen.

Default is n=1000.

-burn-in n: Number of burn-in iterations. Default is n=0.

-h: Print this help screen and exit.

Example to simulate data:

./coev -data nt -method sm -tree example/tree.txt -out example/simulate.out -s 1 -d 10 -r1 5 -r2 5 –ns 10

Example to evaluate score of co-evcoevolution using maximum likelihood framework

./coev -data nt -method sm ml –align example/align.txt -tree example/tree.txt –out example/outML.log –cols 1 2

Example to evaluate score of co-evcoevolution using Bayesian framework

./coev -data nt -method bayes -align example/align.txt -tree example/tree.txt -out example/outTracer.out -IT 100000 -sfreq 1000 -burnin 1000 –cols 1 2

# Reference:

When using this software, please cite:

L. Dib, D. Silvestro and N. Salamin (2014). Evolutionary footprint of coevolving positions in genes. Bioinformatics. 30(9): 1241-9.

L. Dib et al. CoEv platform, in review.

# Contacts:

For questions, please send an email to www.coev [at] unil.ch