User's guide for *ESL* Ver. 0.1

Tae-Kun Seo,
Department of Biological Sciences,
Korea Polar Research Institute,
26 Songdomirae-ro, Yeonsu-gu,
Incheon 406-840,
Republic of Korea,
E-mail) seo.taekun@gmail.com

Jul 28, 2020

This is a first draft of user's guide for *ESL* (pronounced as /esl/; italicized font) program. This is provided mainly for the peer-review of manuscript titled "Measuring Phylogenetic Information of Incomplete Sequence Data" by Tae-Kun Seo, Jeffrey Thorne and Olivier Gascuel (In the following, this manuscript will be referred to as STG manuscript). More detailed manual will be available soon.

ESL program estimates Effective Sequence Length (ESL; pronounced as / i - es - el/; for the definition of ESL, see pages 11-13 of STG manuscript) for given sequence data, tree topology and substitution model.

FORMATS AND OPTIONS

Sequence file : Sequence data format is phylip (sequential) format as shown in Figure 1.

Option file : All options should be represented in the new line within "option.txt" file (Figure 2). In each line of option, only the first item after '>' character is recognized as an input. Do not make any change to the string between '<' and '>' characters.

(1) < seq file >

Sequence file is set here.

(2) < subs model >

Model for nucleotide substitution or amino acid replacement is set here. For nucleotide models, JC (Jukes

```
 12 1000
T1    TTCCGAGGAGCCCTATCACATGGTCGTTTCTAAGATGCCTCCGA...
T2    TTCCGAGGGACTCTCTCACATGATTGTTTCTAAAATGCCTCCGA...
T3    CTCTGAGGGGCCCTGACACATGATTGTTTGTAAAATGCACTCAA...
T4    CTCCGGGAGGCCCTTACATATGACCGTTTGTAAAATTCCTTCAA...
T5    TCCCGAGGGGTACCAGGACATGATAATTTCTAAAATGCCCCTGA...
T6    TCCCGGAGGGTATTGTGACATGATTATTTCTAGAGCGCCTCTGA...
T7    TCCTGGGGGGCCCTACTACATGATAACTCCTAAATTGCCCCTGA...
T8    TCCTGGGGGACCCTATTACATGATAACTCCTAAACTGCCTCTGA...
T9    TTCAGGGGGGCCCCAAGGTATAATTGCTTCCAAAACGCCCTTAG...
T10   TCCAGAAAGGCCCTAAGGCATAATCGCTTCCAAAACGCCCCTAG...
T11   TCCAGGGGAGCCCTGGTTTATAATTGCTTCTGAGGTGCCCCAAG...
T12   CCCAAGGGGGCTCAAGTATATAGTCGCGTCTAAGGTGCCCCCAG...
```

Figure 1: testdata.txt

and Cantor 1969), TN93 (Tamura and Nei 1993), GTR (Tavaré 1986; Yang 1994) can be selected. For amino acid models, WAG (Whelan and Goldman 2001), mtREV24 (Adachi and Hasegawa 1996) , LG (Le and Gascuel 2008), Jones (Jones et al. 1992), Dayhoff (Dayoff et al. 1978) can be selected.

(3) < freq option >

Option for frequencies of nucleotides or amino acids. When nucleotide model is selected in (2), only 'DataFreq' option is valid and empirical nucleotide frequencies are used for rate parameters. When amino acid model is selected in (2), amino acid frequencies provided by the model can be used by choosing 'ModelFreq'

(4) < rand seed >

Random seed number is required for bootstrap procedure.

(5) < tree topo >

Tree topology should be provided with newick format.

(6) < job >

0 should be here to estimate G-Factors and M-Factors. Other job options will be available soon.

(7) < simul length >

Number of times of the length of simulated data with respect to the original data. This corresponds to m of Eq (3) and (4) of STG manuscript.

(8) < boot iter >

```
< seq file >  testdata.txt
< subs model >  GTR // JC, TN93, GTR, WAG, mtREV24, LG, Jones, Dayhoff
< freq option >  DataFreq // DataFreq, ModelFreq
< rand seed >  1
< tree topo >  (((T1,T2),(T3,T4)),((T5,T6),(T7,T8)),((T9,T10),(T11,T12)));
< job >  0
< simul length >  20
< boot iter >  100
< bootstrap type >  0 // 0:RELL, 1:Full
```

Figure 2: esl_options.txt

```
C:\ESL_v01\bin> java ESL_v01.TestMain > scr.txt   +   Enter
```

Figure 3: How to run

Number of bootstrap iteration. This corresponds to B of Eq (15) of STG manuscript.

(9) < bootstrap type >

RELL-like procedure or full bootstrap procedure. See page 9 of STG manuscript.

How to run *ESL*

You must install JAVA runtime environment (JRE; https://java.com/en/download/) to run *ESL*. *ESL* can be run on any OS as long as JRE is pre-installed.

Once you installed JRE, copy ESL_v01 folder to your favourite location. Put sequence and option files in bin folder. Go to bin folder and type the following commands ('TestMain' is case-sensitive) and hit ENTER. Because the results are displayed in the console, you need to capture them by redirecting messages from screen to the appropriate file. In the example of Figure 3, results are redirected from console to scr.txt file.

INTERPRETATION OF THE RESULTS

In the beginning of scr.txt, information for tree topology is shown as in Figure 4. "Node W: X->, <-Y, <-Z" means that "Node X is the ancestor of node W, and nodes Y and Z are the descendants of node W". Also, "brX : Y <–>Y" means that "branch X connects node Y and node Z".

```
......
Tree topology information
Node 0:  T1
Node 1:  T2
Node 2:6->, <-0, <-1
Node 3:  T3
Node 4:  T4
Node 5:6->, <-3, <-4
......
Tree branch information
br0 :  2 <--> 0
br1 :  2 <--> 1
br2 :  6 <--> 2
......
```

Figure 4: scr.txt

In the later part of scr.txt, parameterwise M-Factors, G-Factors and global M-Factor and G-Factor are shown as in Figure 5.

The integer within parenthesis after each standard deviation estimate (0's in this example) represent weird M-Factor and G-Factor estimates during bootstrap (see page 9 of STG manuscript). If this integer is not negligible, full version of bootstrap procedure is required in < bootstrap type >option in Figure 2.

After running *ESL* program, the file named "sequence_file" + "_s-ESL.txt" is generated in the same folder. In our analysis here, "testdata.txt_s-ESL.txt" is generated. Within this file, sitewise and parameterwise ESL (see pages 11-13 of STG manuscript) and R commands for drawing figures are saved (Figure 6). R package "tidyverse" should be pre-installed.

After running *ESL* program, the file named "sequence_file" + "_esl_tree_nex.txt" is generated in the same folder. In our analysis here, "testdata.txt_esl_tree_nex.txt" is generated. Within this file, R commands for drawing Figure 3 of STG manuscript are saved. Move this file to your favourite folder (say, "c:/temp/") and copy only R commands (not tree information of nexus format) to R prompt. Then you will get phylogeny similar to Figure 3 of STG manuscript (Figure 7). If you move the file to different folder, you should change the path within "testdata.txt_esl_tree_nex.txt". Colors, thickness of branches, and font sizes can be adjusted by modifying R commands within "testdata.txt_esl_tree_nex.txt".

```
......
#### Individual beta (M-Factor) estimate +- std
beta[0] = 0.941930 +- 0.053350 (0)
beta[1] = 0.999968 +- 0.032398 (0)
beta[2] = 1.105290 +- 0.049699 (0)
......
#### Individual Rho (G-Factor) estimate +- std
rho[0] = 0.493388 +- 0.020669 (0)
rho[1] = 0.700998 +- 0.014124 (0)
rho[2] = 0.719371 +- 0.017488 (0)
......
#### globaBeta (M-Factor) = 0.998920 +- 0.023108 ; (TestStat = -0.046738;
pValue from bootstrap = 0.510000) (# of extreme bootstrapped samples = 0,
out of 100)
#### globaRho (G-Factor) = 0.602523 +- 0.015369 ; (# of extreme
bootstrapped samples = 0, out of 100)
......
```

Figure 5: scr.txt

```
......
s_ESL <- c(1.427519, 0.902281, -0.047730, ...
......
sp_ESL <- matrix(c(-0.012685,-0.012966,-0.002481, ....
....
x<-1:dim(sp_ESL)[1]
y<-1:1000
library(ggplot2)
library(reshape2)
z <- sp_ESL
p3 <- persp(x,y,z, col="white", theta=120, xlab="br", ylab="site",
zlab="sp-ESL")
p3
......
```

Figure 6: testdata.txt_s-ESL.txt

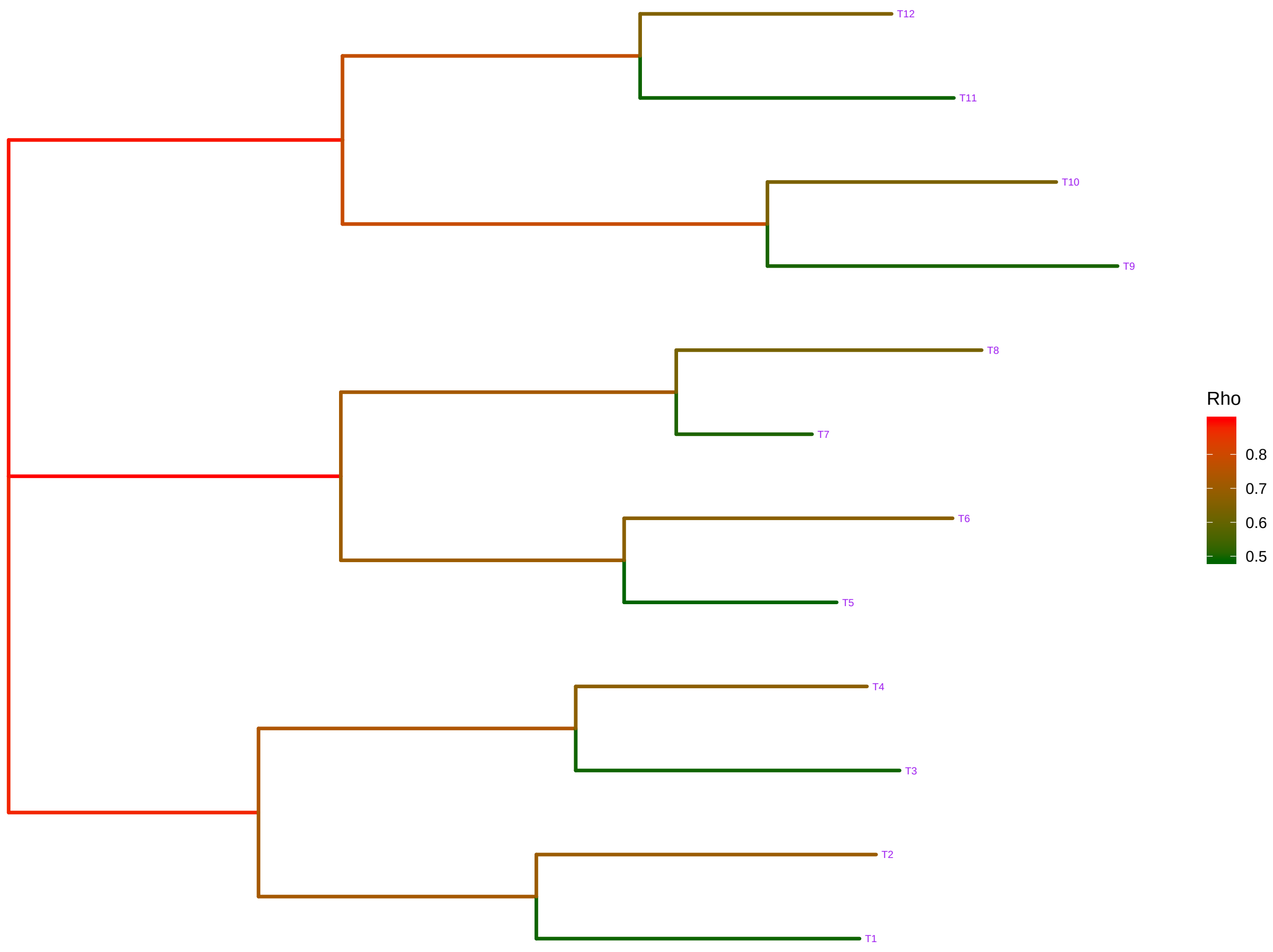

Figure 7: G-Factor (ρ_i) estimates of testdata.txt; For R commands see "testdata.txt_esl_tree_nex.txt"

LITERATURE CITED

Adachi J. and Hasegawa M. 1996. MOLPHY version 2.3:programs for molecular phylogenetics based on maximum likelihood. Comput. Sci. Monogr. 28:1-150.

Dayhoff M.O., Schwartz R. M., Orcutt B.C. 1978. A model of evolutionary change in proteins. Atlas of protein sequence and structure, Vol 5, Suppl. 3, pp. 345–352. National Biomedical Research Foundation, Washington DC.

Jones D.T., Taylor W.R., Thornton J.M. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS 8:275–282.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In Mammalian protein metabolism (ed. H. N. Munro), pp. 21–123. Academic Press, New York.

Le S.Q., Gascuel O. 2008. An Improved General Amino Acid Replacement Matrix. Mol. Biol. Evol. 25(7):1307–1320.

Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math. Life Sci. 17:57–86.

Tamura K., Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10: 512–526.

Whelan S., Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. Mol. Biol. Evol. 18:691–699.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.