User's guide for *ESL* Ver. 0.2

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

June 30, 2021

This is a first draft of the user's guide for the *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, Olivier Gascuel and Jeffrey Thorne. In the following, this manuscript will be referred to as the SGT manuscript. A more detailed manual will be available soon.

The *ESL* program estimates Effective Sequence Length (ESL; pronounced as / i - es - el/; for the definition of ESL, see pages 13–14 of the SGT manuscript) for given sequence data, tree topology and substitution model.

FORMATS AND OPTIONS

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

Option file : All options should be represented as a separate line within the "esl_option.txt" file (Figure 2). In each line that specifies an option, only the first item after the '>' character is recognized as an input. Do not change the string between the '<' and '>' characters.

(1) < seq file >

The sequence file is set here.

(2) < subs model >

```
 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

The model for nucleotide substitution or amino acid replacement is set here. For nucleotide models, JC (Jukes and Cantor 1969), TN93 (Tamura and Nei 1993), and 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), and Dayhoff (Dayoff et al. 1978) can be selected.

(3) < freq option >

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

(4) < rand seed >

A random seed value is required for the bootstrap procedure. Any integer between -2^{63} and $2^{63} - 1$ can be selected for random seed number.

(5) < tree topo >

The tree topology should be provided in the newick format. Tree topology should be unrooted. Branch lengths can be included in newick format.

(6) < job >

0 should be here to estimate G-Factors and M-Factors. Other job options are not allowed in current ver-

```
< 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
< short br constraint >  1 // 0:R0, 1:R1, 2:R2
< CI for boundary check >  0 // 0:95%, 1:99%
```

Figure 2: esl_options.txt

sion.

(7) < simul length >

Number of times by which the length of the simulated data is longer than the original data. This corresponds to m of Eq (3) and (4) of the SGT manuscript.

(8) < boot iter >

Number of bootstrap iterations. This corresponds to B of Eq (15) of the SGT manuscript.

(9) < bootstrap type >

RELL-like procedure or full bootstrap procedure. See page 9–11 of the SGT manuscript.

(10) < short br constraint >

Three constraints (R_0, R_1 and R_2) for selecting moderately long branches. See supplementary information [C]

(11) < CI for boundary check >

Confidence level for constraints R_1 and R_2. See Eqs (S13) and (S14) of supplementary information [C]

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 JRE is installed, copy the ESL_v02 folder to your favorite location. Put the sequence and option files in the bin folder. Go to the bin folder and type the commands of Figure 3 ('TestMain' is case-sensitive) and hit ENTER. Because the results are displayed in the console, you need to capture them by redirecting

```
C:\ESL_v02\bin>  java ESL_v02.TestMain   >  scr.txt   ↵
```

Figure 3: How to run

```
......
### Tree topology node information ###
Node 0:  T1
Node 1:  T2
Node 2:  T3
......
Node 18:17->, <-0, <-1
......
### Tree topology branch information
br0 :  18 <--> 0 ; terminal br to T1
br1 :  18 <--> 1 ; terminal br to T2
br2 :  21 <--> 2 ; terminal br to T3
......
br18 :  17 <--> 18 ; MRCA of T1 and T2
......
```

Figure 4: scr.txt

messages from screen to the appropriate file. In the example of Figure 3, results are redirected from the console to the scr.txt file.

INTERPRETATION OF THE RESULTS

In the beginning of scr.txt, information about 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 <–>Z" means that "branch X connects node Y and node Z".

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 pages 9–11 of SGT manuscript) that have values that are outside the range [0,2]. If this integer is substantially greater than zero, the full version of the bootstrap procedure is recommended in < bootstrap type >option of Figure 2.

```
......
#### Individual beta (M-Factor) estimate +- std
beta[0] = 0.919046 +- 0.051048 (0)
beta[1] = 1.012447 +- 0.036901 (0)
beta[2] = 0.971255 +- 0.057786 (0)
......
#### Individual Rho (G-Factor) estimate +- std
rho[0] = 0.511539 +- 0.017178 (0)
rho[1] = 0.704503 +- 0.022012 (0)
rho[2] = 0.512716 +- 0.015787 (0)
......
#### globaBeta (M-Factor) = 1.010271 +- 0.023745 ; (TestStat = 0.432551;
pValue from bootstrap = 0.695000) (# of extreme bootstrapped samples = 0,
out of 200)
#### 0.590381 +- 0.018680 ; (# of extreme bootstrapped samples = 0, out of
200)
......
```

Figure 5: scr.txt

After running the *ESL* program, various result files are generated in the same folder. The file named "sequence_file" + "_sp-ESL_score_only.txt" ("testdata.txt_sp-ESL_score_only.txt" in this example) contains sp-ESL, s-ESL and p-ESL values. These values can be read-in by R by running "source()" command in R prompt.

The file named "sequence_file" + "_sp-ESL_info.txt" ("testdata.txt_sp-ESL_info.txt" in this example) contains aligned columns and corresponding s-ESL as Figure 6. Columns sorted with descending s-ESL follows. R command to draw figure with sp-ESL values are included in the end.

The file named "sequence_file" + "_loglike_surface.txt" ("testdata.txt_loglike_surface.txt" in this example) contains information for log-likelihood curvature. It contains R codes to draw figure as shown in Figure S2(B–D) of supplementary information.

The file named "sequence_file" + "_esl_tree_nex.txt" ("testdata.txt_esl_tree_nex.txt" in this example) contains information for individual G-Factors and M-Factors of lineages. Within this file, R commands for drawing Figure 3 of SGT manuscript are included. Move this file to your favourite folder (say, "c:/temp/") and copy only R commands (not tree information of nexus format) to the R prompt. Then you will get phylogeny similar to Figure 3 of SGT manuscript (Figure 7 of this manual). If you move the file to a different folder, you should change the path within "testdata.txt_esl_tree_nex.txt". Colors, thickness of

```
......
### Sitewise ESL (unsorted) ###
# site num | column pattern | sitewise ESL
#0th site:  TTCCTTTTTTTC | 1.291302
#1th site:  TTTTCCCCTCCC | 0.879693
#2th site:  CCCCCCCCCCCC | -0.047019
......
### Sitewise ESL (sorted, start) ###
site num | column pattern | sitewise ESL | brID(max sp-ESL), brID(min sp-ESL)
#i=0 | 191th site:  CCTTCTGTGCTT | 9.877353 | br6(34.262372), br2(-0.518841)
#i=1 |  74th site:  TTTCCCACCCCC | 9.664993 | br6(40.639917), br16(-0.749125)
#i=2 | 311th site:  CCCCCCTCCCCC | 7.963127 | br6(35.725106), br16(-0.264207)
......
### R commands for figures
x<-1:dim(sp_ESL)[1]
y<-1:1000
library(ggplot2)
library(reshape2)
z <- sp_ESL
p3 <- persp(x,y,z, col=‘‘white", theta=45, phi=30, xlab=‘‘br", ylab=‘‘site", zlab=‘‘sp-ESL")
p3
```

Figure 6: testdata.txt_sp-ESL_info.txt

branches, and font sizes can be adjusted by modifying R commands within “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.

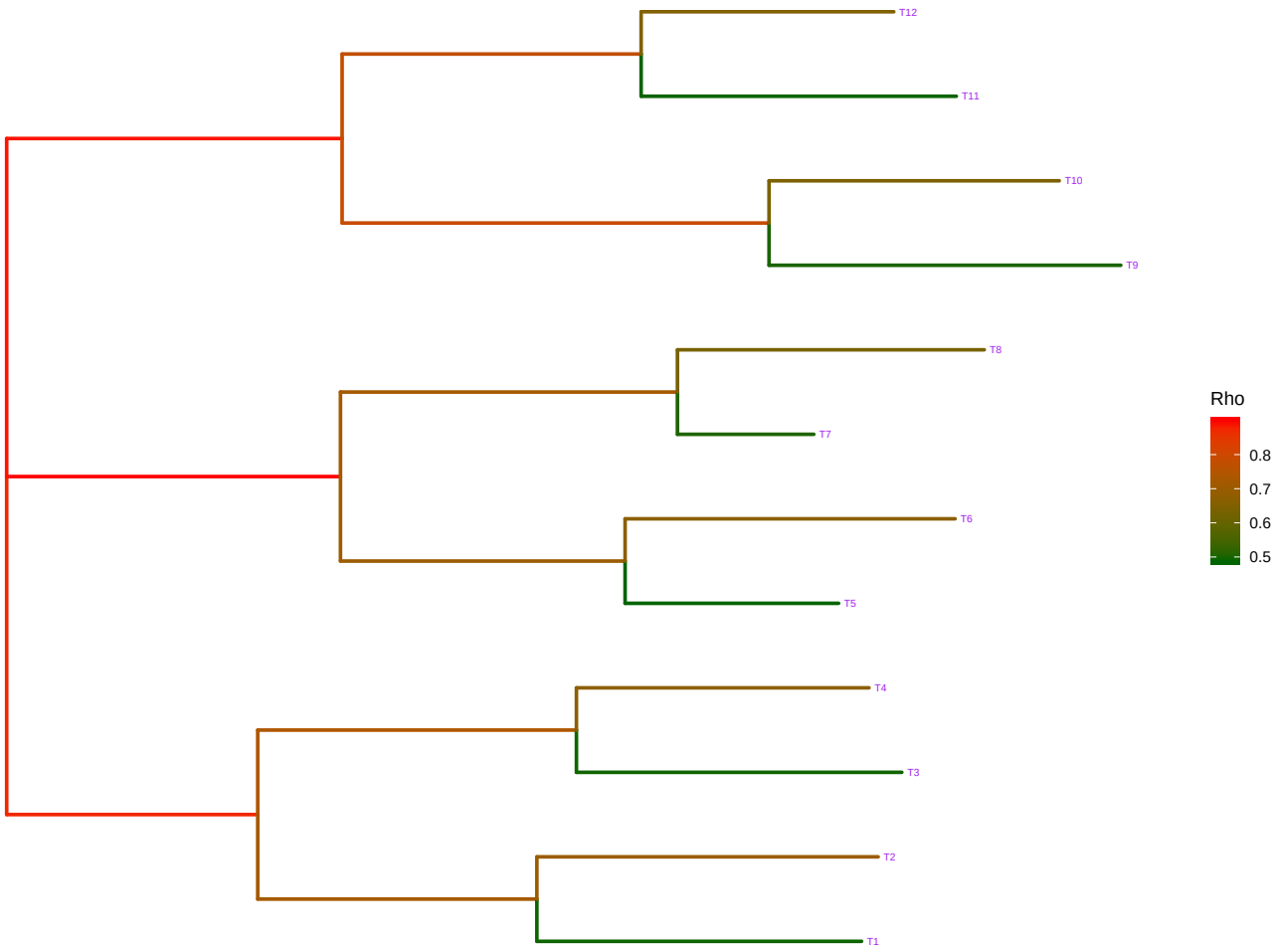

Figure 7: G-Factor estimates of testdata.txt; For R commands see "testdata.txt_esl_tree_nex.txt"

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.