ComPhy User Manual

Digital Biology Laboratory
Computer Science Department
271C Christopher S. Bond Life Sciences Center
1201 East Rollins Road
University of Missouri-Columbia
Columbia, MO 65211-2060
http://digbio.missouri.edu/ComPhy

I. System Requirements

1. Software:
JRE 5.0 or later version
Reference URL: http://java.sun.com/javase/downloads/index.jsp
2. Operating Systems:
Windows or Linux/Unix platforms
3. Hardware:
At least 1GB memory

II. Installation guide

1. To begin installation save the package to any folder, for example: C:\ComPhy_win.zip and uncompress it.
*Please do not use any space in folder name, because it will cause some potential errors for java programs.

- Windows: Using Winzip or any other upzip tool to uncompress the command line viersion of the tool ComPhy_win.zip.
Before running ComPhy on Windows platforms, please use the following command to set the class path: **set classpath=.;%classpath%**

- Linux/Unix: tar -xvzf ComPhy_linux-x32.tgz

2. ComPhy directory contains following:
ComPhy.jar: jar file that runs ComPhy tool
Input: will contains all input files

- a file contains list of interested genome (Provided by user)
- SeqDir: a directory contains protein sequence files for the interested genomes (Could be empty or provided by the user)
- GeneLocDir: a directory contains gene physical location files for interested genomes (CAN NOT BE EMPTY)
- OrthologDir: a directory contains ortholog list files for interested genomes (Could be empty or provided by the user)

Output – will contains all output files, including genome distance matrix file, tree file in Newick format or phylogram format and performance evaluation result file.

Othertools – some third party tools, such BLAST, Phylip.

III Input files formats

All the input files need to be put in "Input" directory in "ComPhy".

1. One File of interested genomes: all genome names are listed as in one column
For example:

```
A+A2.1.1.1.1-NC_000916
A+A2.4.1.1-NC_008212
B+B2.1.1.1.1-NC_000853
B+B4.1.1.1.1-NC_008025
B+B10.1.1.1.11-NC_005071
...
```

2. All genome protein sequence files need to be in "SeqDir" directory in "Input" folder.
The sequence file needs to be Fasta format, for example:

>gi|15678032|ref|NP_275146.1| hypothetical protein MTH1 [Methanothermobacter thermautotrophicus str. Delta H]
MNRVDLSIFIPDSLTAETGDLKIKTYKVVLIARAASIFGVKRIVIYHDDADGEARFI
RDILTYMDTPQYL
RRKVFPIMRELKHVGILPPLRTPHHPTGKPVTGEYRQGLTVKRVKKGTLVDIGAD
KLALCREKLTVNRIM
SFRVVRLGKEILIEPDEPEDRYWGYEVLDTRRNLAESLKTVGADVVVATSRNASP
ITSILDEVKTRMRGA
REAAILFGGPYKGLPEIDADIWVNTLPGQCTETVRTEEAVLATLSVFNMLTQIDE
KDE
>gi|15678033|ref|NP_275147.1| 50S ribosomal protein L3 [Methanothermobacter thermautotrophicus str. Delta H]
MARHHQPRKGSVAFSPRKRAARETPRVKSWPQVDEPGLLALAGYKAGMTHVM
MVDNQKNSPTEGMEVSTP
VTILEVPPLTVMAVRTYEKTSRGLKTLGEVLATETKDDLRRKLTPPADDYDQEA
AIEKIRSNMEYVADVR

Note:
1. Each sequence has an annotation line, starting with ">", either starts just with gene ID, then space or tab, or the line starts with format "gi|15678032|ref|NP_275146.1|", where gene ID follows "gi", then space or tab (Which is the NCBI sequence format).
Note: All gene IDs in sequence files must be consistent with gene IDs in gene location files.
All sequence files need to be stored in directory "SeqDir" in "Input" folder.

2. All gene physical location files need to be in "GeneLocDir" in "Input" folder.
The location file is in the format of following:
Line 1: "genome_size #" shows genome sequence length, can not leave empty
Line 2: "gene_number #" shows total number of genes in the genome, can not be emoty

Line 3: “PID StartPos EndPos”, this is the title tabs of each gene information, which are gene ID, physical starting position of the gene and physical ending position of the gene.

All data in the files have to be tab-delimited.

For example:

genome_size 1751377
gene_number 1873
PID StartPos EndPos
15678032 171 977
15678033 997 2010
…

3. If user chooses to use their own list of orthologs, instead uses ComPhy to define ortholog lists, then all the ortholog list files should be in folder “orthologDir” in “Input” directory.

File name would be “1st SpeciesName_VS_2nd SpeciesName.txt” and the format of ortholog should be as follow:

!1st species name (separated by tab) 2nd species name
Ortholog1 from species1 (separated by tab) ortholog1from species2
Ortholog2 from species1 (separated by tab) ortholog2from species2
Ortholog3 from species1 (separated by tab) ortholog3from species2
…

For example:

!A+A2.1.1.1.1-NC_000916 A+A2.4.1.1-NC_008212
GI|15679411|REF|NP_276528.1| GI|110666977|REF|YP_656788.1|
GI|15679411|REF|NP_276528.1| GI|110666977|REF|YP_656788.1|
GI|15679404|REF|NP_276521.1| GI|110666979|REF|YP_656790.1|

IV. How to run the tool

ComPhy takes 4 parameters:

1. Absolute ComPhy directory path
2. name of the genome list file
3. Y for yes and N for no for selection if use provide own orthologs or not
4. Protein sequence file extension, “faa” or “fasta” or “txt”.

For example:

If the ComPhy directory is saved under C: diver, then the command line for running the ComPhy is:

java -jar C:\ComPhy\ComPhy.jar C:\ComPhy genomeList.txt N faa

I. Bug report

ComPhy is developed by the Digital Biology Laboratory, University of Missouri-Columbia. Your feedback on using ComPhy will be greatly appreciated. If you find any bug or error in the program or have any suggestion, please feel free to contact us:

Guan Ning Lin: gnln66@mizzou.edu
Or Dr. Dong Xu: xudong@missouri.edu