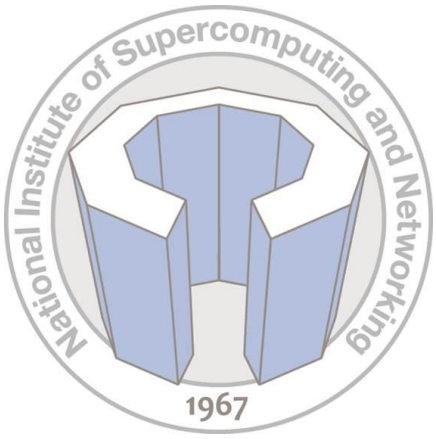
SimFlu [ver 1.0]

A Simulation Tool for

Predicting the Evolutionary Patterns of Influenza A virus

User Manual



Insung Ahn

National Institute of Supercomputing and Networking @ KISTI

Table of Contents

Table of Figures

Introduction

What is SimFlu?

SimFlu stands for "a **Sim**ulation tool for in**flu**enza virus", and it perfoms the sequence simulations using the codon variation patterns of influenza A viruses over time. This program can be installed on *Linux* and *Macintosh OS X* as well as *Windows XP* or *7*, and it automatically searches the various types of input files from user input folder of working directory.

Installation

SimFlu is freely available from the 'PRODUCT' in SimFlu homepage (http://lcbb.snu.ac.kr/simflu/products.html). In this page, you may download the compressed SimFlu files for your operating systems. Detailed installation method for each operating system is as follows.

Windows machine

If your operating system is *Windows XP* or *7*, you can download a zip-file named 'SimFlu.zip', and then, extract it to a preferred folder in your computer. Because SimFlu is developed using C^{++} language, the full file path of your folder must be in English. The SimFlu zip-file contains five different subfolders, including [exe], [lib], [examples], [manual] and [user_input]. The executable file for SimFlu, 'SimFlu.exe', is located in [exe], and the library files for simulation can be found in [lib]. In [lib] folder, you can find another subfolder named [v_1]. The integer after the prefix 'v_' is the version number for SimFlu library, so, in this case, the version number is '1'. During the simulations, SimFlu automatically finds the latest version of the library by comparing the version number of this folder, and select the largest number to use the latest version of library. If you want to move the SimFlu.exe from its original folder to the other place, you must create a 'shortcut' of SimFlu.exe file first, and then, move or copy it to your preferred location. SimFlu also provided a designed icon of createad shortcut for Windows named 'SimFlu_icon.ico' in [manual] folder.

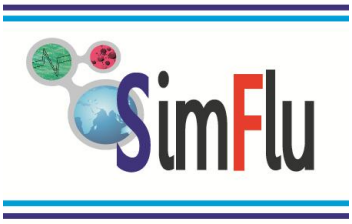


Figure 1. Shortcut icon for SimFlu (SimFlu_icon.ico)

The PDF file of this manual document is also located in [manual], and [user_input] is a default folder where SimFlu searchs your input files. Several sample files are located in that folder for the beginner. If you assign other directory as a working directory in <Working Directory Setting> step, SimFlu will try to find your input files in 'user_input' folder within your working directory.

In [examples] folder, you can see the three different subfolders, such as [Sample_command_file], [Sample_seed_sequence] and [Sample_user_parameter], which contain some example files for the beginners. In [Sample_command_file], you may find a sample command file named 'test_command.cmd', which contains sample command options with their values. The 'test_seed.seq' file in [Sample_seed_sequence] folder is a sample seed sequence of hemaggutinin gene of influenza A virus, and 10 different text files in [Sample_user_parameter] is the sample user parameter files.

Linux or Mac OS X machine

If the operating system of your PC is *Linux* (*Red Hat*-based system) or *Mac OS X* (version 10 or upper), download a tgz-file named 'SimFlu.tgz', and extract and expand it to the preferred location. As for the *Linux* system, you can use the 'tar' command with specific options as described below.

```
[user@userPC]$ tar xvfz SimFlu.tgz
[user@userPC]$ cd SimFlu
[user@userPC SimFlu]$ ls
 examples  exe  lib  manual  user_input
[user@userPC SimFlu]$ cd exe
```

```
[user@userPC SimFlu exe]$ ls
 SimFlu.exe
[user@userPC SimFlu]$ chmod 755 ./SimFlu.exe
```

SimFlu contains five different subfolders, such as [exe], [lib], [examples], [manual] and [user_input], and an executable file, "SimFlu.exe (for *Linux*)", or "SimFlu (for *Mac*)", can be found in [exe]. Due to the simulation process is dependent on the library files in [lib], you must execute the SimFlu program within the [exe] folder. So, if you want to move the SimFlu.exe file to the other location in *Mac* machine, you must create a 'shortcut' and then move or copy it to the preferred location. We also provided a shortcut icon for *Mac* named 'SimFlu_Mac_icon.bmp' in [manual]. Sometimes, SimFlu may not work on *Linux* system because of the absence of the gcc compiler. In that case, you can install the latest version of gcc by using yum command as follow.

```
[user@userPC]$ yum install gcc-c++
```

In *Fedora Linux*, you may see an error message for the absence of C^{++} library in your PC. In this case, go to the "add/remove softwares" in *Linux* setting options, and search the "libstd" in the searching window to find "Compatibility standard C^{++} libraries" package, and then, just install it.

Getting started

Calculation of the SimFlu library

SimFlu program provides pre-calculated variation parameters of influenza A virus genes between two different year-of-isolation groups as library. The source nucleotide sequences of these library files were collected from the Influenza Virus Resource (http://www.ncbi.nlm.nih.gov /genomes/FLU/) of US National Center for Biotechnology Information (NCBI). In the current version of SimFlu library (ver. 1.0), we collected the messanger RNAs (mRNAs) of 10 kinds of genes (HA, NA, NP, PA, PB1, PB2, M1, M2, NS1 and NS2) for 3 major influenza A virus subtypes (H1N1, H3N2 and H5N1). The target year-of-isolations were from 2000 to 2011, target host species were human and swine for H1N1

and H3N2, and human and avian for H5N1.

The first step of calculating the SimFlu library files is to perform multiple sequence alignments (MSAs) among all the possible pairs of year-of-isolations between 2000 and 2011. When you execute the SimFlu program, you have to choose the interval type of target years as well as the initial and final target year-of-isolations in <Library Settings> step. The interval type can be devided into 2 categories, such as type A, and type B. If you choose the 'type A', the time intervals between the initial (*time T*) and final target (*time T'*) years will be one-year, whereas the initial year is fixed and only the final year is increased by 1-year in 'type B'. Next figure (Fig. 2) is the calculation process of SimFlu libraries when you choose the 'type A' interval time with a range between 2000 and 2011.

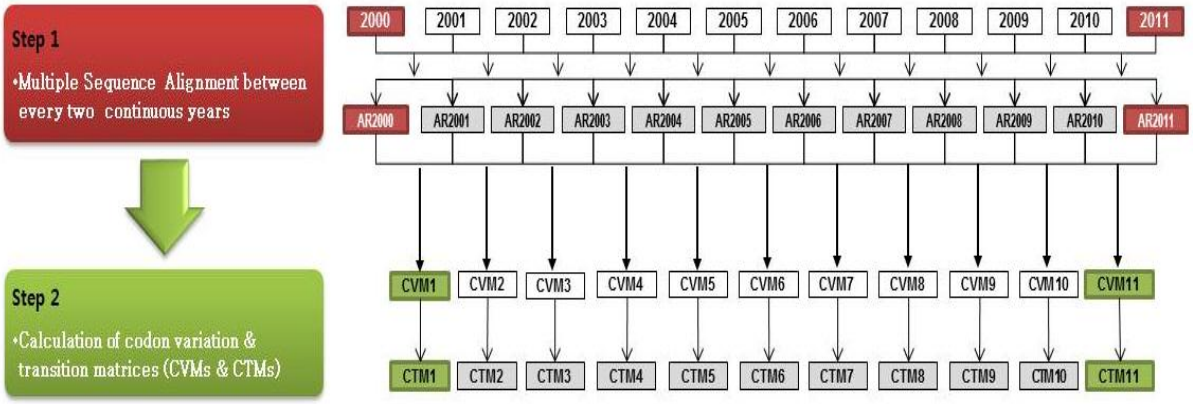


Figure 2. Calculation process of SimFlu Libraries (Interval type A)

Each pair of target years must be aligned togather using MSA program such as ClustalW, and then their aligned output sequences are devided into 2 different files according to their year of isolations for the further process (AR2000, AR2001, …, AR2011). Just be sure that you need to save the aligned sequences in FASTA-format.

In the second step, we counted all the possible codon variations between 2 MSA result files that contain the 'gap' information. Detailed comparing and counting process is described as follow.

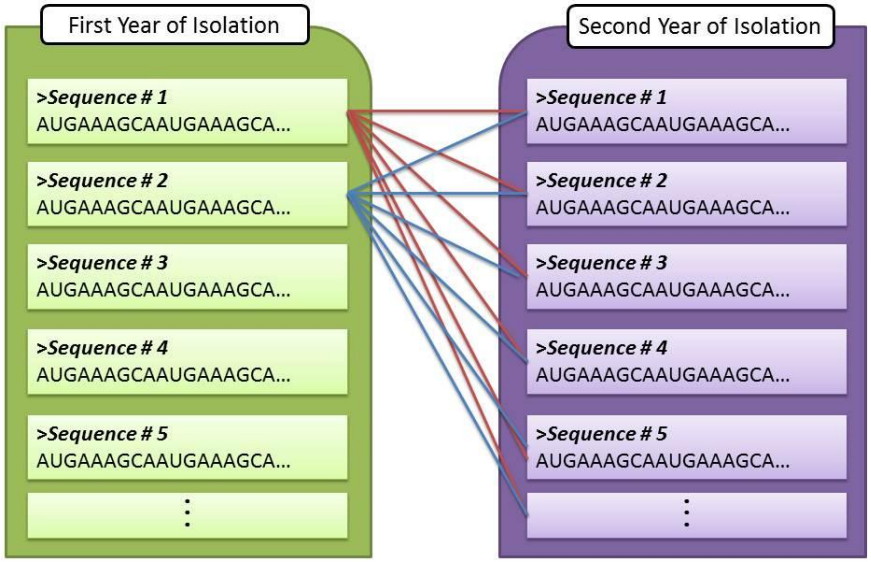


Figure 3. Comparison of aligned sequences between the first and second year of isolations

First of all, each sequence (Sequence #1, #2, …, #*n*, *n* = total number of sequences in 'First year-of-isolation') in the MSA result file named 'First year of isolaiton' is compared with the sequences (Sequence #1, #2, …, #*p*, *p* = total number of sequences in 'Second year of isolation') in the other MSA result file named 'Second year of isolation'. As a result, total *n* x *p* times of comparisons will be conducted. Each codon in each sequence region along the aligned result of the first year of isolation is compared with that in the second year of isolation result, and counted variation is saved in the 61 x 61 matrix named codon variation matrix (*CVM*) in Figure 2. In the final step, all the CVMs are converted into codon transition matrix (*CTM*) using the Markov model. The names of calculated *CTM*s are encoded, and then, saved in [lib] folder as packaged with their version number.

Interactive mode

If you are a beginner of SimFlu program, I recommend you to choose the ‘Interactive mode’ in <Mode Selection> step.

Figure 4. Mode Selection: Interactive mode

In this mode, SimFlu automatically search the parent folder of SimFlu.exe file located in the subfolder [exe], and will ask you whether to use it as a working directory or not (Fig. 5). If you want to use other directory as your working directory, type the ‘yes’ in the first question and write the exact file path to the preferred working directory folder. SimFlu will check the existence of your working directory, and move to the next step if it is normally exist in your PC.

```
<<<<<<<<<<<<<<<<<<<< Working Directory Setting >>>>>>>>>>>>>>>>>>>>

 !! Note: All the Directories and File names in file path MUST be in "English" !!

 * SimFlu will automatically set the Working Directory as "/Users/isahn_note/Desktop/working/".
   Do you want to change the Working Directory? [Y or N]
   Your Answer: y

 * Type the Full Path of the Working Directory you want to use (ex. /home/test).
   Your Answer: ../test

 * You changed the Working Directory as "/Users/isahn_note/Desktop/working/test/".
   Is that correct? [Y or N]
   Your Answer: y
```

Figure 5. Working Directory Setting

In the next step, SimFlu ask you the name of seed sequence file, and all you have to do is just to type the exact name of your seed file, not the whole file path (Fig. 6). Once you write the seed file name, SimFlu will search it in the [user_input] folder in your working directory, and if it is normally exist, SimFlu will show up “Okay” sign in the screen and move to the next step. Note that the seed sequence file in the [user_input] folder must be in FASTA-format and its extension must be ‘*.seq’.

```
<<<<<<<<<<<<<<<<<<<<     Input File List     >>>>>>>>>>>>>>>>>>>>

 !! Note: You Input File Path is "/Users/isahn_note/Desktop/working/test/user_input/". !!

 * Type the Full name of Seed Sequence File in "/Users/isahn_note/Desktop/working/test/user_input/" .
   (Note: Only FASTA-format file (*.seq) is acceptable)
   Your Answer: test_seed.seq

     "Okay, I checked your Seed Sequence File."
```

Figure 6. Seed sequence file setup

Using the SimFlu Library

```
<<<<<<<<<<<<<<<<<<<<   Parameter Type Setting  >>>>>>>>>>>>>>>>>>>>

 * Do you want to use the Parameter Files in SimFlu's Library? [Y or N]
   Your Answer: y

     "The Most Recent Version of Library, "version (v_1)" will be used in this Simulation."
```

Figure 7. Using SimFlu Library

SimFlu provides pre-calculated codon variation parameters as libraries in SimFlu package, and you can use them by selecting the library option (Fig. 7). In library ver. 1.0, SimFlu provide parameter files for 3 different subtypes of influenza A virus such as H1N1, H3N2 and H5N1, and accessible year-of-isolation range is between 2000 and 2011. If you choose to use the Library, SimFlu will automatically search for the latest version of library from the subfolder [lib], and inform you the version number. More detailed setup method for SimFlu library is as follows.

```
<<<<<<<<<<<<<<<<<<<<   Library File Setting    >>>>>>>>>>>>>>>>>>>>

 * Type the Target "Hemagglutinin" (HA) type of the Libraries you want to uese. (Ex.: 1 / 3 / 5)
  (Note: Currently H1N1, H3N2 and H5N1 subtypes are available..)
   Your Answer: 1

 * Type the Target "Neuraminidase" (NA) type of the Libraries you want to uese.
  (Note: Currently H1N1, H3N2 and H5N1 subtypes are available..)
   Your Answer: 1

     "You choose H1N1 subtype~!"
```

Figure 8. Library selection: Subtype setup

The first work to do in the <Library File Setting> part is to choose the target subtype of influenza A virus to import from the library (Fig. 8). The version 1.0 of SimFlu library provides 3 subtypes such as H1N1, H3N2 and H5N1. Once you type the target hemagglutinin and neuraminidase types, SimFlu checks the presence of subtype in the library and show up an error message if it’s not exist.

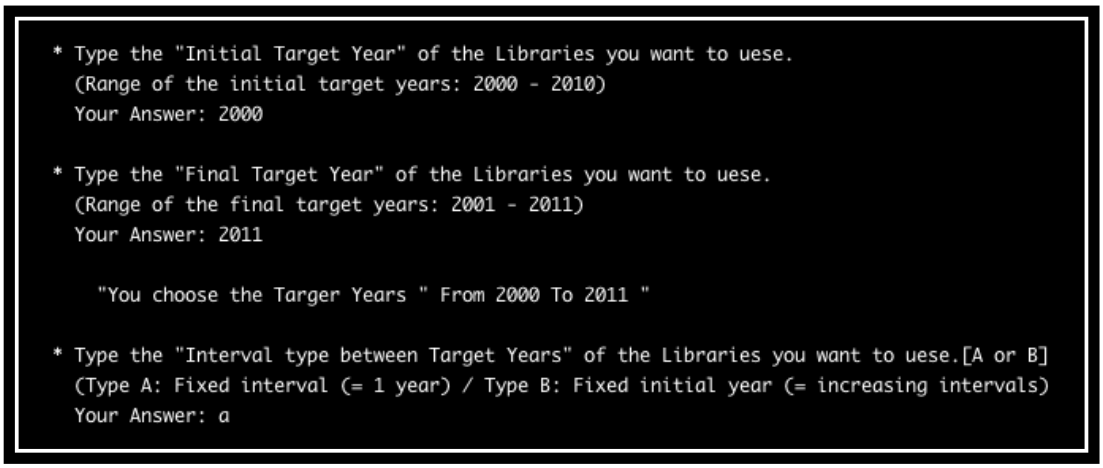


Figure 9. Library Selection: target year and interval type setup

Next step of library setting is to choose the target range of year-of-isolations. The SimFlu Library (ver. 1.0) provides simulation parameters of influenza A viruses isolated from 2000 to 2011. After selecing the target range of years, you also have to choose the interval type between target years. If you choose the 'type A', the time intervals between the initial and final target years will be one-year, whereas the initial year is fixed and only the final year is increased by one-year in 'type B'. More detailed explanations are shown in the 'Calculation of the libraries' part. Once you type all the informations, SimFlu checks the presence of initial and target years in the library and if there is no error, you can move to the next step.

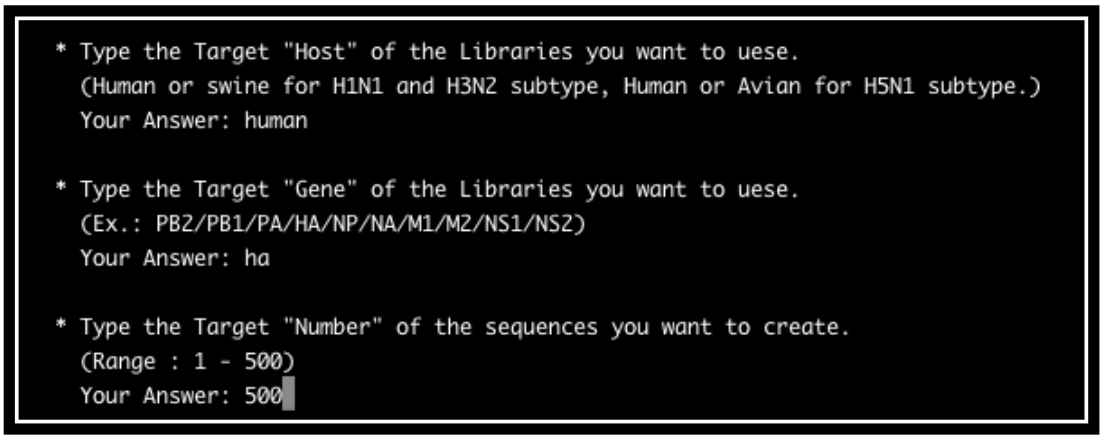


Figure 10. Library Selection: target host, gene and simulation number setup

In this final step, you need to choose other library conditions such as the target host, target gene, and total number of simulations you want to perform (Fig. 10). The SimFlu Library (ver.

1.0) provides 2 kinds of host species per each subtype, such as human and swine for H1N1 and H3N2, and human and avian for H5N1 subtype. According to the gene of your seed sequence, you can choose the target gene name for library. SimFlu now provides the library files for 10 genes, including PB2, PB1, PA, HA, NP, NA, M1, M2, NS1 and NS2. Finally, SimFlu ask you the last question about the total number of simulation you want to perform. In order to prevent the overloading of your PC, SimFlu restricts the maximum number of simulations as 500.

Using the User Paremeter

If you have your own codon transition matices (CTMs) of your influenza sequences, you can choose the 'User parameter' option in <Parameter Type Setting>.

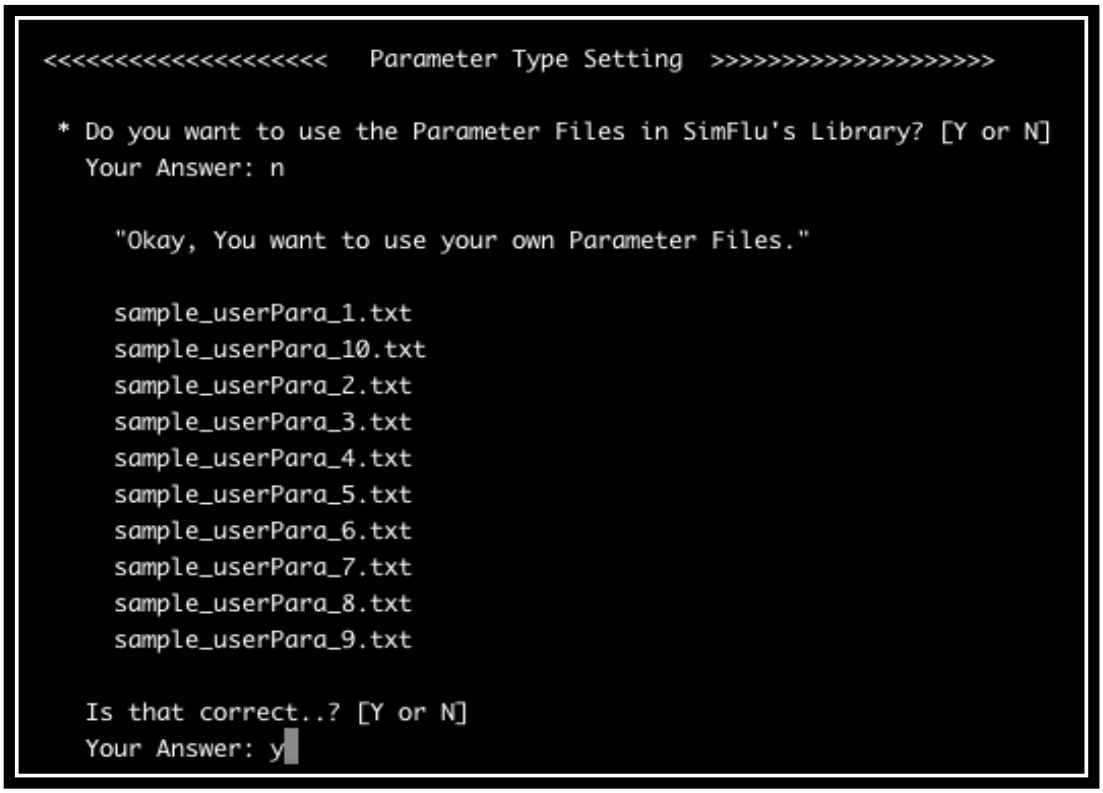


Figure 11. User Parameter Search

In order to use your own parameter files, you have to answer 'No' in the SimFlu's question above. Once you type 'n', SimFlu will start to find your parameter files (*.txt) in the subfolder [user_input] of your working directory, and if the parameter files are normally exist, it will show you the whole list of parameters.

Command file mode

Because the ‘Interactive mode’ is designed for the beginner of SimFlu, user has to follow the step-by-step process per each job. It is quite good for the new users, but it may not so good for the advanced users because of the long setup time. For the advanced users who have large data size and short analysis time, SimFlu recommands the ‘Command file mode’.

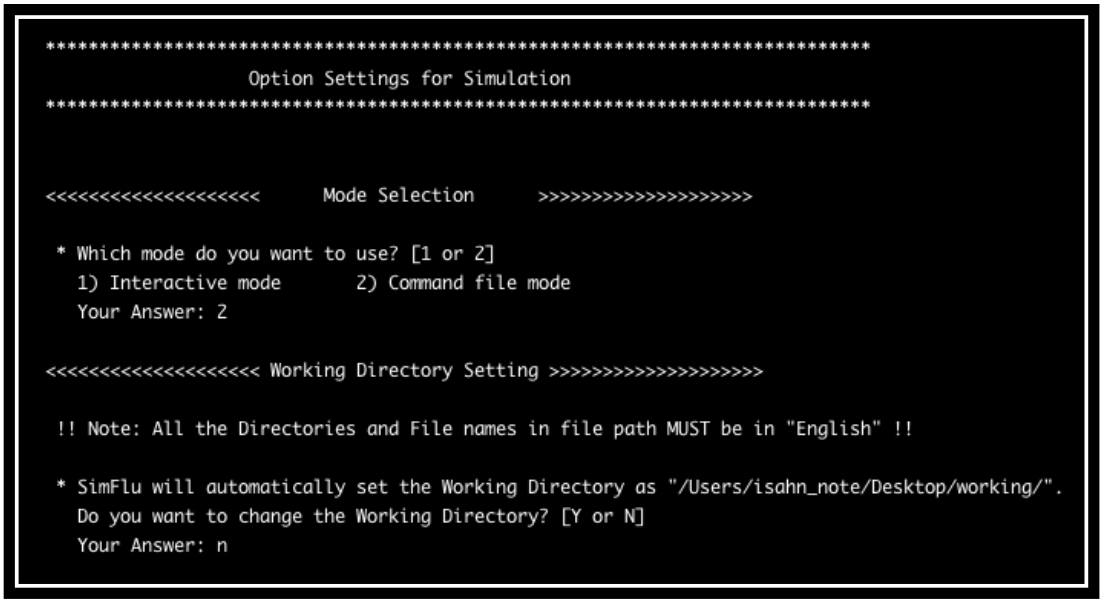


Figure 12. Mode Selection: Command file mode

If you want to use the ‘Command file mode’, choose the number 2 in the <Mode Selection> part (Fig. 12). After choosing the ‘Command file mode’ in the first step, SimFlu automatically search the parent folder of SimFlu.exe file located in the subfolder [exe], and will ask you whether to use it as a working directory or not. If you want to use other directory as your working directory, type the ‘yes’ in the first question and write the exact file path to the preferred working directory folder. SimFlu will check the existence of your working directory, and move to the next step if it is normally exist in your PC.

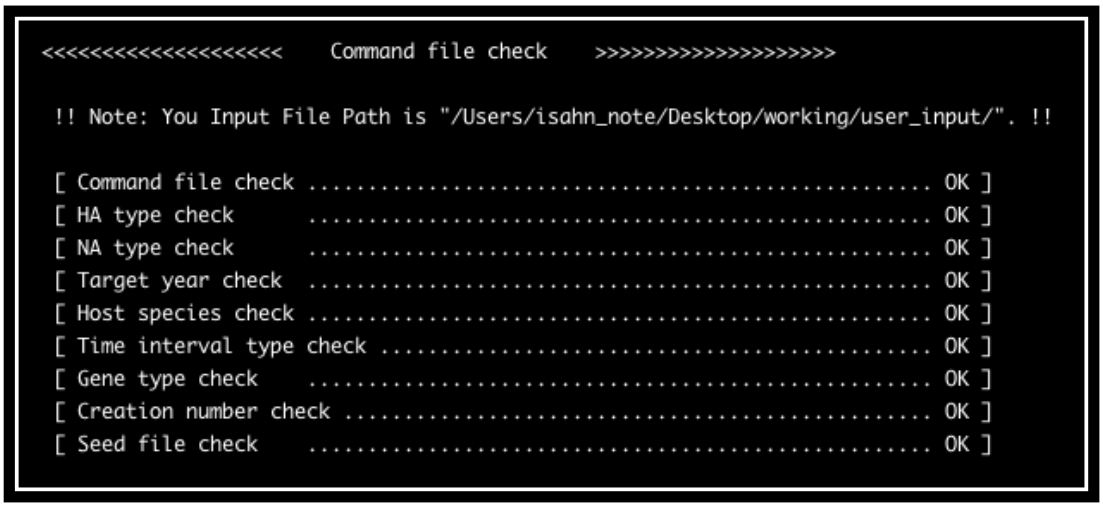


Figure 13. Checking list of command file mode

Once you setup your working directory, SimFlu searches the command file (*.cmd) in the subfolder [user_input] in your working directory. If your command file is normally exist, SimFlu will start to simulate automatically using the working conditions described in your command file. It evaluates the usabilities of defined conditions in command file, and finally starts to simulate if all the target parameters are valid. Note that there must be only one command file (*.cmd) in your input folder. Detailed setting indicators in the command file are as follows.

```
#$ -s test_seed.seq
#$ -l yes
#$ -ha 1
#$ -na 1
#$ -iy 2000
#$ -fy 2003
#$ -ho human
#$ -it b
#$ -g m1
#$ -num 500
```

Figure 14. Example of setting indicators

In the command file, all the lines which contain the setting indicators must be started with ‘#$’, and an indicator and its value is followed. Setting indicators and their values are

described in the next Table.

Table 1. Setting indicators and their preferred valuese in SimFlu command file

Indicator	Description	Values
-s	Seed sequence file name	*.seq
-l	Whether to use SimFlu Library or not	Y or N
-ha	Hemagglutinin type	H1N1 / H3N2 / H5N1
-na	Neuraminidase type	H1N1 / H3N2 / H5N1
-iy	Initial target year	2000 – 2010 (fy > iy)
-fy	Final tager year	2001 – 2011 (fy > iy)
-it	Interval type between years	A or B
-ho	Target host species	Human or swine for H1N1, H3N2 / human or avian for H5N1
-g	Target gene name	PB2, PB1, PA, HA, NP, NA, M1, M2, NS1 or NS2
-num	Total simulation number	1 - 500

Option mode

'Option mode' in SimFlu is very similar to the 'Command file mode', but the main purpose of these two modes is quite different. The 'Option mode' is designed to facilitate the highly advanced users who will execute the SimFlu program in various conditions using 'queuing system', while the 'Command file mode' is for only one command file condition. You can execute the SimFlu program with all the working conditions at one go. Detailed option parameters are as follows.

```
[user@userPC SimFlu exe]$ ./SimFlu.exe –wd /home/test -s test_seed.seq -l yes -ha 3 -na 2 –iy
      2000 –fy 2011 –it b –ho swine –g ha –num 500
```

All the option indicators are same as those of the command file mode except for the '-wd' option, which defines the file path of working directory. If you want to simulate many times in one single action, you can create a job script file that contains all the simulation conditions, and then, submit that file to your queuing system.

Simulation algorithm

SimFlu is a simulation tool than create the hypothetical future nucleotides from the real influenza A virus sequence (= seed sequence) using the codon variation parameters, such as SimFlu's library or user parameter files. Detailed working process is described as follow.

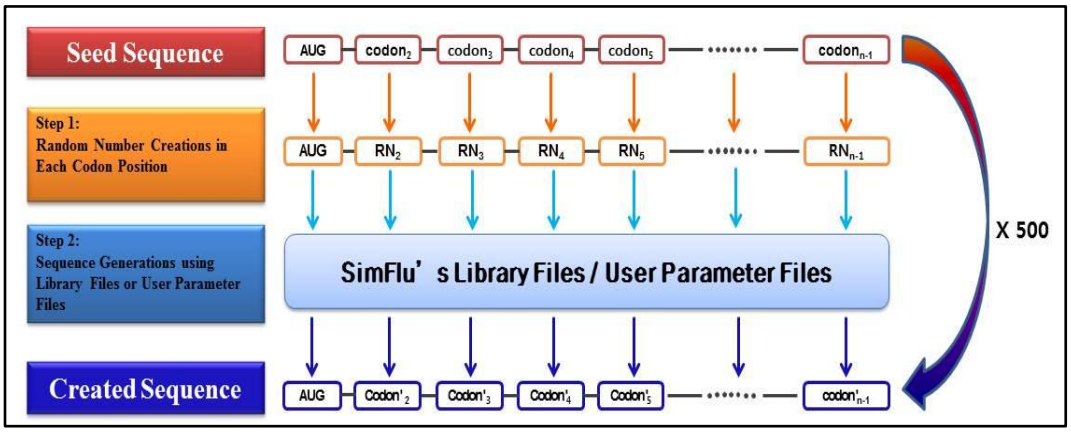


Figure 15. Simulation process of SimFlu

The simulation process of SimFlu begins with importing the seed sequence in units of codons. Once all the codons of seed sequence are read, SimFlu generates a random number between 0 and 1, respectively, in each codon position except for the start and termination codon, and when this task is completed, SimFlu converts each codon of the seed sequence to a new coodn which is changed by the probability of random number based on the SimFlu library or user parameter. It repeats the same process as many times as you ordered. During the working process, SimFlu creates a temporal folder named [_tmp] in your working directory to conduct many intensive works, and this folder will be removed when SimFlu finish the simulation.

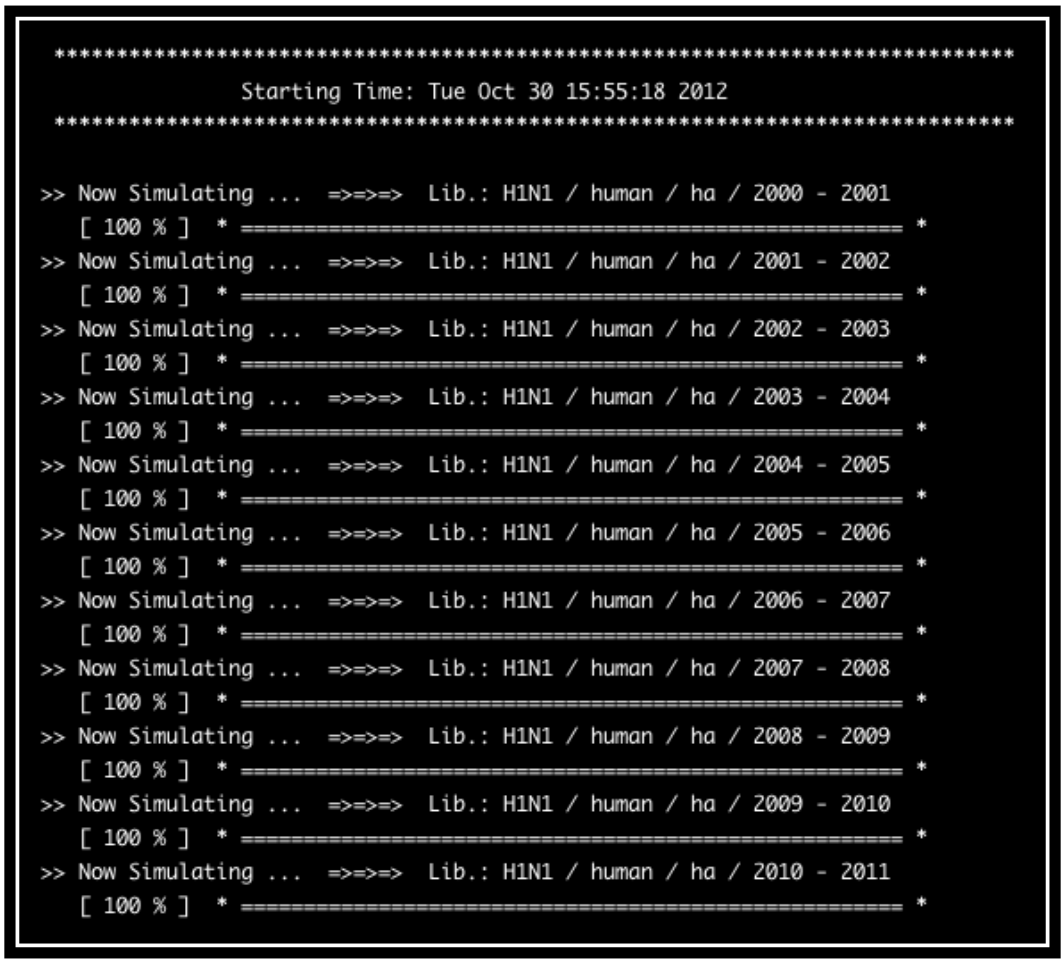


Figure 16. Job processing of SimFlu

This figure is a screenshot of job processing of SimFlu. In the first line, SimFlu presents the starting time of simulation, and each simulation process is represented as a bar graph as shown above. In this case, the user uses the library of HA (hemagglutinin) gene of H1N1 subtype isolated between 2000 and 2011. Because the initial and final target years are increased equally, such as 2000-2001, 2001-2002, …, the interval type of target years is thpe 'A'. If you select to use your own user parameter files instead of SimFlu's library, you can see your parameter file names after 'User Para.:'. After finishing all the simulation processes, SimFlu also informs the 'Ending Time' and 'Total Processing Time'.

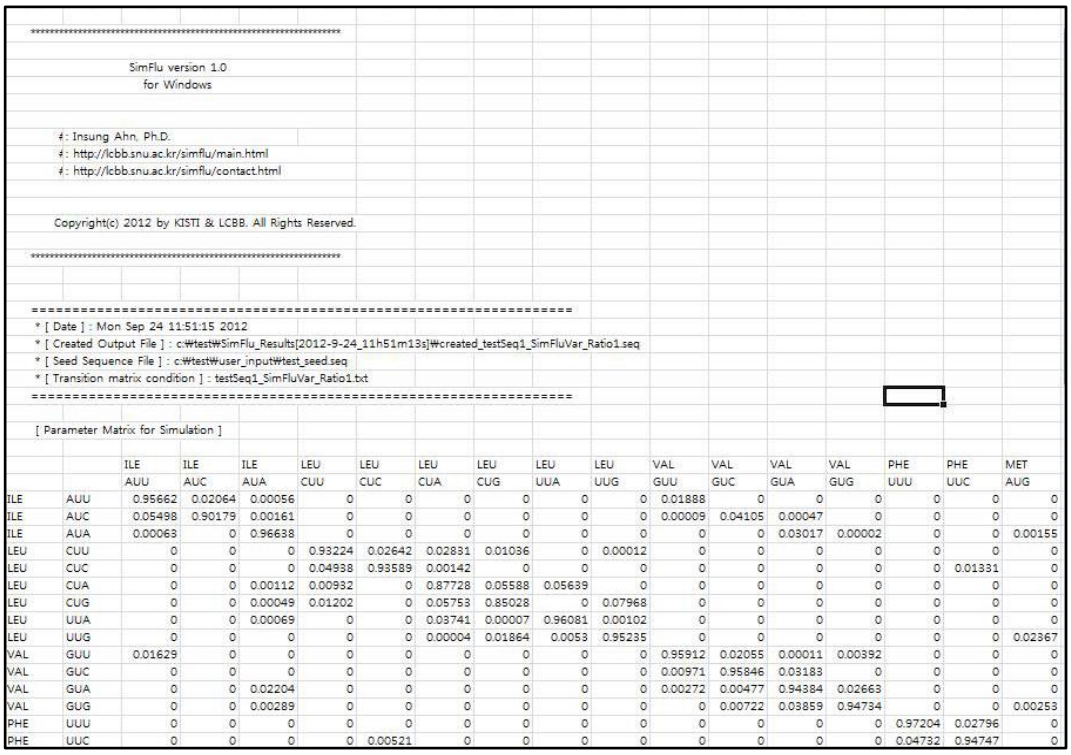
Outputs

Once all the simulation jobs are finished, SimFlu creates an output folder named ‘SimFlu_Results[year_month_day_hour_minute-sec]’ automatically in your working directory, and all the calculated output files are located in that folder. The output files of SimFlu can be divided into two categories, such as created sequence and information files. In order to minimize the user’s work, SimFlu automatically assigns the sequence and information file names. SimFlu attaches the subtype, host, gene, and the initial and final target years after the prefix, ‘created_’ with the extension, *.seq, for the simulated sequence file, while the extension for the information file is ‘_info.txt’.

```
>CreatedSeq_1
AUGAAAGUAAAACUACUGGUCCUGUUAU ...
>CreatedSeq_2
AUGAAAGUAAAACUACUGGUCCUGUUAU …
.
.
.
>CreatedSeq_500
AUGAAAGUAAAACUACUGGUCCUGUUAU …
```

Figure 17. SimFlu Output: Created sequence file

<Figure 17> is the contests of the created sequence file. The output format is FASTA and each sequence title is automatically assigned with the created number of sequence with the prefix, ‘CreatedSeq_’.



SimFlu version 1.0
for Windows

#: Insung Ahn, Ph.D.
#: http://lcbb.snu.ac.kr/simflu/main.html
#: http://lcbb.snu.ac.kr/simflu/contact.html

Copyright(c) 2012 by KISTI & LCBB. All Rights Reserved.

* [Date] : Mon Sep 24 11:51:15 2012
* [Created Output File] : c:₩test₩SimFlu_Results[2012-9-24_11h51m13s]₩created_testSeq1_SimFluVar_Ratio1.seq
* [Seed Sequence File] : c:₩test₩user_input₩test_seed.seq
* [Transition matrix condition] : testSeq1_SimFluVar_Ratio1.txt

[Parameter Matrix for Simulation]

		ILE	ILE	ILE	LEU	LEU	LEU	LEU	LEU	LEU	VAL	VAL	VAL	VAL	PHE	PHE	MET
		AUU	AUC	AUA	CUU	CUC	CUA	CUG	UUA	UUG	GUU	GUC	GUA	GUG	UUU	UUC	AUG
ILE	AUU	0.95662	0.02064	0.00056	0	0	0	0	0	0	0.01888	0	0	0	0	0	0
ILE	AUC	0.05498	0.90179	0.00161	0	0	0	0	0	0	0.00009	0.04105	0.00047	0	0	0	0
ILE	AUA	0.00063	0	0.96638	0	0	0	0	0	0	0	0	0.03017	0.00002	0	0	0.00155
LEU	CUU	0	0	0	0.93224	0.02642	0.02831	0.01036	0	0.00012	0	0	0	0	0	0	0
LEU	CUC	0	0	0	0.04938	0.93589	0.00142	0	0	0	0	0	0	0	0	0.01331	0
LEU	CUA	0	0	0.00112	0.00932	0	0.87728	0.05588	0.05639	0	0	0	0	0	0	0	0
LEU	CUG	0	0	0.00049	0.01202	0	0.05753	0.85028	0	0.07968	0	0	0	0	0	0	0
LEU	UUA	0	0	0.00069	0	0	0.03741	0.00007	0.96081	0.00102	0	0	0	0	0	0	0
LEU	UUG	0	0	0	0	0	0.00004	0.01864	0.0053	0.95235	0	0	0	0	0	0	0.02367
VAL	GUU	0.01629	0	0	0	0	0	0	0	0	0.95912	0.02055	0.00011	0.00392	0	0	0
VAL	GUC	0	0	0	0	0	0	0	0	0	0.00971	0.95846	0.03183	0	0	0	0
VAL	GUA	0	0	0.02204	0	0	0	0	0	0	0.00272	0.00477	0.94384	0.02663	0	0	0
VAL	GUG	0	0	0.00289	0	0	0	0	0	0	0	0.00722	0.03859	0.94734	0	0	0.00253
PHE	UUU	0	0	0	0	0	0	0	0	0	0	0	0	0	0.97204	0.02796	0
PHE	UUC	0	0	0	0	0.00521	0	0	0	0	0	0	0	0	0.04732	0.94747	0

Figure 18. SimFlu Output: Information file

<Figure 18> is the contents of the information file of SimFlu. First of all, the version number of SimFlu program is shown on the top, and more detailed informations such as simulation date, created sequence file path, the seed sequence file path, and the user parameter or library category are followed. If you use the SimFlu's library in your simulation, you can see the detailed library information, such as subtype, host species, gene name, the initial and final target years in '[Transition matrix condition]'. Because the SimFlu provides the detailed information for the inputs and outputs of SimFlu per each simulation job, user don't need to remember those informations if he keep this information files with other output files. After this information part, you can see a 61 x 61 transition matrix of codon variations from user parameter or SimFlu's library files used in this simulation work.