A brief tutorial for using `TransmicBS`

Identification of putative disease transmission links using hierarchical clustering in phylogenetic trees

Kaveh Pouran Yousef
kaveh.yousef@fu-berlin.de

October 5, 2015

1 What is `TransmicBS`?

`TransmicBS` is a Python program for identifying clusters in phylogenetic trees by combining inter-taxon patristic distances with bootstrap-based significance values. The corresponding clades of phylogenetic trees indicate sets of evolutionary close species. This program is customized for the identification of putative disease transmission clusters. In particular, it allows for multiple follow-up viral sequences extracted from the same individual (for a detailed description of the underlying method see [1]).

2 How to use `TransmicBS`?

2.1 Files needed by the program

Besides the program itself, three additional files are needed for running it:

1. The control file of the program called `controlfile.tbsc` .
2. A file containing the phylogenetic tree in Newick format.
3. An .xls-file containing meta information for each genetic sequence

The control file must be placed in the same directory as the program. The other two files may also be located in other directories than the program.

2.2 Format of genetic sequence identifiers

Since the program allows for two different modes (individual- and sequence-based clustering, see subsection 2.3) the format of taxa in the phylogenetic tree needs to be adopted accordingly. In the individual-based mode, the name of the taxa must contain the sequence ID and the individual/patient ID separated by an underscore. For instance, a taxon with the name `12-1023_0321` indicates that the sequence ID is `12-1023` and this sequence belongs to the individual/patient ID is `0321`.

In contrast, in the sequence-based mode the underscore is not needed and the names of the taxa should consist of the sequence ID only.

2.3 Configuring the file `controlfile.tbsc`

The program configuration is conducted through a control file called `controlfile.tbsc`. It must be placed in the same directory as the program. A sample control file is contained in the program package. In order to customize the program for your needs, the parameters in the double quotes have to be modified. Consider the following example of a control file:

```
-----------------------------------------------------------------
#TransmicBS control file

#Please only edit parameter values within double quotes.

#Name of the tree file in Newick format
#(or absolute path if file not in same directory as the programm)
tree_file="RAxML_bipartitions_WithFP.BS_TREE"

#Cut-off value for the mean pairwise patristic distances (mppd)
#i.e. a clade is a transmission cluster if  mppd < cutoff_mppd (in percent)
cutoff_mppd="4.8"

#Cut-off for the bootstrap support value (in percent)
#i.e. a clade is a transmission cluster if  value > cutoff_bootstrap
cutoff_bootstrap="95"

#Patient-wise clustering
#This option only works with sequence identifiers consisting
#of a unique sequence ID followed by a patient ID separated by
#underscore e.g. seqID_patID (you can then set it to true).
#In all other cases (e.g. simple tree-based clustering) set this
#option to false.
patWiseClust="true"

#Name of excel file with patient meta-information
RKIfile="RKIdatabase.xls"

#Enforce for sequence IDS the format
#of the Robert Koch Institute
enforce_RKI_format="true"

#Database ID of the outgroup sequence
outgroupSeqId="SIVcpz_ant"
-----------------------------------------------------------------
```

All the lines beginning with # are ignored by the program as comments and all control parameters are entered within quotation marks "...". In the following we explain the control parameters step by step:

1. `tree_file`: this parameter determines the name of the file containing the phylogentic tree. If this file is located in the same directory then simply its name should be typed. If the file is located in another directory, in Linux you may either type the absolute path to it e.g. `/home/somedirectory/ExampleTree` or the relative path (depending on the location of the program).

2. `cutoff_mppd`: distance threshold (mean pairwise patristic distance, MPPD) for the hierarchical clustering method (units: %).

3. `cutoff_bootstrap`: bootstrap support threshold which determines the candidate clades as potential clusters (units: %).

4. `patWiseClust`: boolean parameter which determins if the clustering should be based on individuals (e.g. virus carriers). Setting this parameter to `"true"` invokes a clustering method were the distance between individuals is computed as the minimal tree distance between any pair of viral sequences belonging to the two individuals (for details see Pouran Yousef *et al.* [1] and the corresponding supplementary information). This method is useful if multiple sequences per individual are allowed. Otherwise, if clustering should be conducted based on genetic sequences only then set this parameter to `"false"`.

5. `write_out_table`: boolean parameter which determins if, in addition to the text ouput, the identified clusters should be written out into an table file with meta information. To this end an additional .xls file is required containing tbe meta information (see below). The clusters are printed out into table, where the first column contains the sequence names (used for naming the tree taxa), and the second column contains the sequence IDs (corresponding to the first column in the `database_file`). Also, all further columns correspond to the further columns (if any) of the `database_file`. The file has an ending .tsv (tab-separated values) and may be impoted into any table manipulation program.

6. `database_file`: Name of the .xls file containing the meta-information about the genetic sequences (this parameter is only considered by the program if `write_out_table` is `true`.). The first column of this program needs to contain the sequence IDs which correspond to the first part of the sequence names in the phylogenetic tree (see subsection 2.2). The following columns can contain further information corresponding to each genetic sequence, such as collection year and subtype of the viral sequence or the characteristics of the individuals carrying the virus etc. Note that currently only the .xls format is supported (not .xlsx or .csv). Therefore, in order to use a table file containing meta-information for the genetic sequences, simply save your table file in the .xls format (this can be easily done using Microsoft Excel or the Linux program Gnumeric).

7. `enforce_RKI_format`: Enforces a special format for the sequence names (`yy-abcde`), where `yy` denotes the collection year of the genetic sequence and `abcde` is a 5-digit number. Setting this parameter to `"true"`, results in converting the sequence IDs from the phylogenetic tree file into this format. E.g. (`01-3650`) is then converted to (`01-03650`) by adding a leading digit 0. Note that this parameter only should be set to `"true"` if the sequence IDs in table with the meta-information are also in this format.

8. `outgroupSeqId`: Identifier of the outgroup sequence used for rooting the phylogenetic tree. Note that if the tree is already rooted then it will be re-rooted using this genetic sequence as an outgroup.

2.4 Running the program

After ensuring that Python is installed on your system, all required files are available in the program folder and the configuration of the control file is finished, the program can be started in Linux by

```
me@mycomputer:~$ python transmicBS.py
```

2.5 Text output

As an example, if three clusters have been identified, the text output file in the directory `text_out` may have the following form:

```
This is an output file of TransmicBS.

Transmission clusters computed according to the distance
threshold of 1.3 % and a bootstrap threshold of 95.0 %.

Number of sequences:  17
Number of identified clusters:  3
Average number of sequences in a cluster:  2.6667
Number of sequences not included in clusters:  9

1.
Mean value of pairwise patristic distances 0.0127
Bootstrap support (%): 68
['F', 'G']

2.
Mean value of pairwise patristic distances 0.0298
Bootstrap support (%): 68
['M', 'N', 'O']

3.
```

```
Mean value of pairwise patristic distances 0.1481
Bootstrap support (%): 52
['B', 'C', 'D']
```

References

[1] K. Pouran Yousef, K. Meixenberger, M. Smith, K. Jansen, B. Gunsenheimer-Bartmeyer, O. Hamouda, M. von Kleist, and C. Kücherer. Inferring HIV-1 transmission dynamics in Germany from genotypes of recently transmitted viruses. *Submitted.*