CPN

(Cladograms ⏵ Phylograms ⏵ Network)

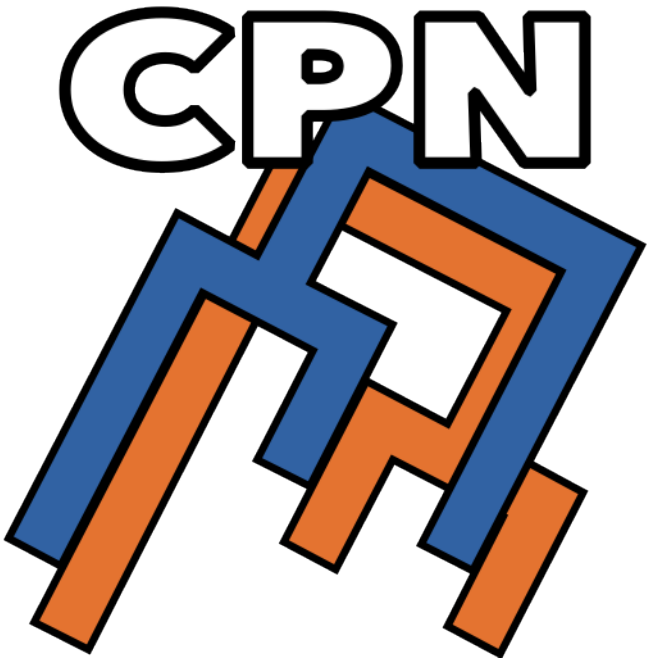
User manual

Branders Vincent & Mardulyn Patrick

March 2015

Program available at : http://ebe.ulb.ac.be/ebe/CPN.html

Contact information : vbranders@gmail.com; pmarduly@ulb.ac.be



Introduction

CPN provides a tool for the inference of allele (or haplotype) networks under a global maximum parsimony approach. It should be used in combination with a phylogenetic inference program capable of inferring all most parsimonious (MP) trees from a DNA sequence alignment. CPN reads this set of MP trees and the corresponding DNA sequence alignment, and generates a network graph.

CPN can be separated in two major components:

1. **Cladograms to Phylograms (CP).**

Many phylogeny inference programs are capable of inferring the MP trees (cladograms) from a DNA sequence alignment. While most of these programs are also able to infer one or a few MP phylograms for each inferred cladogram, to the best of our knowledge, no one has a function to produce all MP phylograms associated with the set of inferred MP cladograms. While such function would indeed be unnecessary for inferring species trees (the original purpose of these programs), it becomes crucial for inferring a MP network graph, because each different phylogram has the potential to add new MP paths in the network. This component of the program finds all MP phylograms corresponding to the set of MP cladograms and DNA sequence alignment provided by the user.

2. **Phylograms to Network (PN).**

This component combines a set of MP phylograms (normally, the set generated by the CP component) into a network graph, using the algorithm described in Cassens *et al.* (2005)[1], modified to allow reducing the length of cycles, by merging some of its edges and nodes, where possible (see end of the manual for a detailed description of this modification).

[1] Cassens I, Mardulyn P, Milinkovitch MC. 2005. Evaluating intraspecific "Network" construction methods using simulated sequence data: do existing algorithms outperform the global maximum parsimony approach? *Systematic Biology*, 54:363-372

Graphical user interface

The Graphical User Interface (GUI) of CPN is divided into three major parts: the menu bar (top) which controls the processing of data and program outputs, the *Progress* Panel (left), displaying the state of a run, and the *Input Data* Panel (right), summarizing the content of the input files and of current results. *Figure 1* shows the different parts of the GUI.

The *menu bar* is used to control the processing of data and program outputs, and allows to modify the parameters of a run. The remaining of the GUI shows information over the running job and the results produced.

Figure 1 shows a snapshot of the GUI: production of phylograms can be monitored on section 2 and information on these are given in section 6; the progress of trees combination is presented on section 4 and results are listed on section 7; section 3 shows general information over the run as well as errors or missing data. Section 5 mimics a *console* and displays all outputs from the software.

Illustration

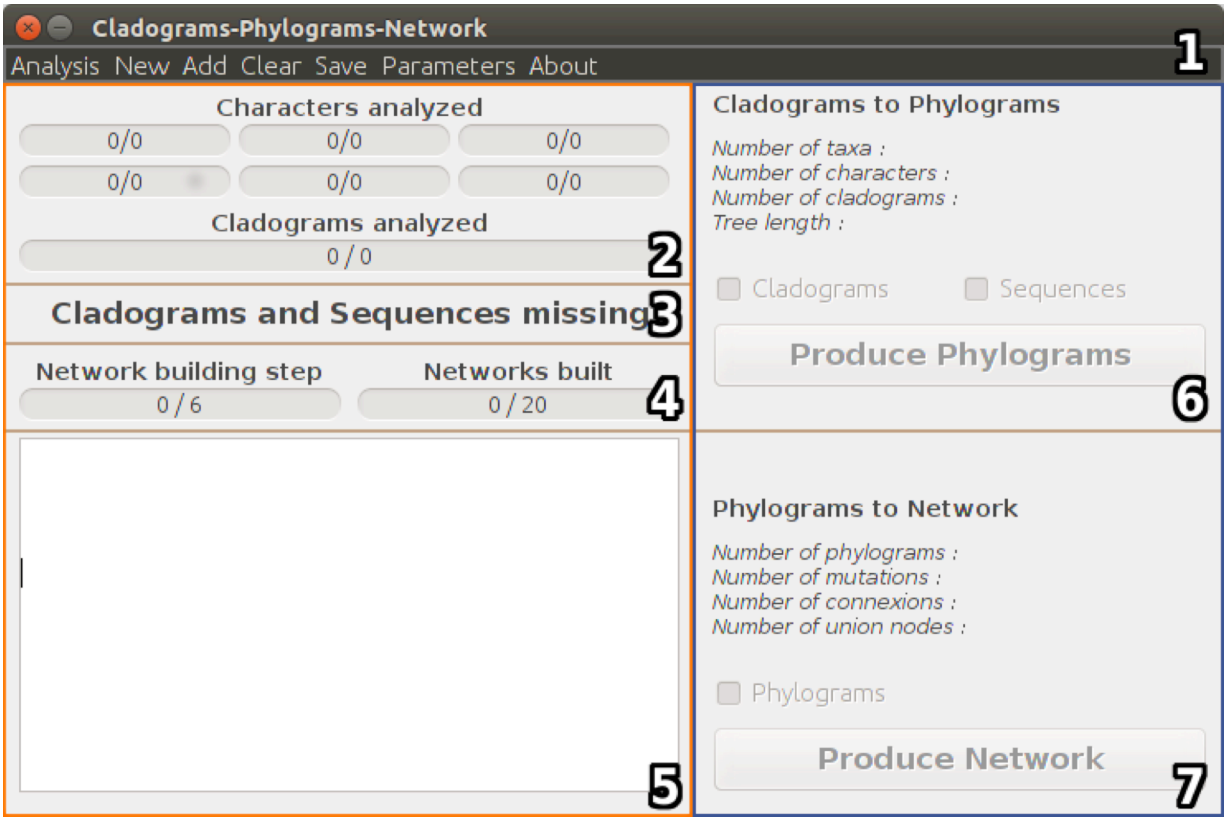
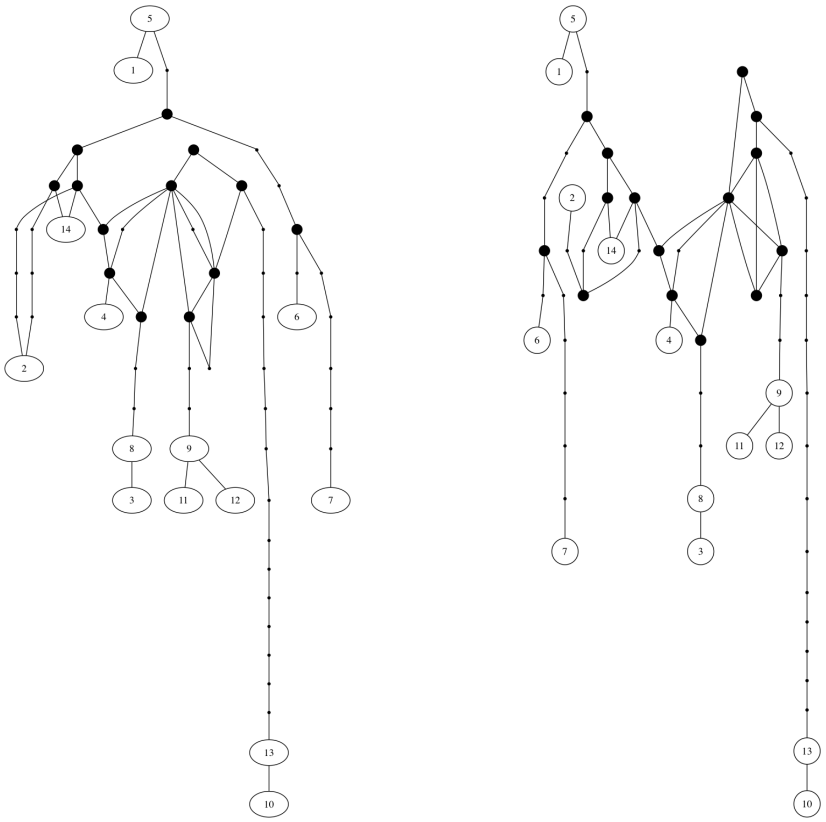


Figure 1: Partitioning of the GUI
Menu bar *(gray section n°1).*
Progress Panel *(orange sections n°2+3+4+5).*
Input Data Panel *(blue sections n°6+7).*
lighter sections are related to the production of phylograms and darker sections are related to the combination of trees into a network graph.

Running an analysis (tutorial)

- Open the program by double-clicking the CPN icon
- Choose *New* > *File*, and select the Nexus or Phylip file containing the sequence alignment (only one sequence per allele/haplotype allowed; see Data input on page 13)
- Choose *Add* > *Cladograms*, and select the Nexus or Newick file with the MP cladograms (inferred using a phylogeny inference program; see Data input on page 13)
- Choose *Parameters* > *Reduce cycles* (activates optional extra step implemented to reduce the length of cycles, after the network is generated, by merging nodes and edges, where possible)
- Check that the button "Produce Phylograms" is active, and click on it.
- When all phylograms are found, click on button "Produce Network", which should have become active by then.
- At the end of the analysis, the program has created a Nexus file (containing the phylograms) and two DOT files (containing the inferred networks, the second file containing the network after the attempt to reduce cycles length) in the folder where your input files are located.

Below, the two networks obtained when running CPN on the "example 1" data file (provided with the program); best network before (left) and after (right) cycles reduction (DOT files open in Graphviz; http://www.graphviz.org).



Detailed description of available commands

The menu bar gives access to all options (commands) available.

Overview of available menus:

- **Analysis:**
  Infer phylograms, combine trees, or quit the software.
- **New:**
  Load data from files while erasing data previously in memory. You can define the type of data to import (sequences, cladograms or phylograms). If you select a specific type, CPN will only erase the data of the same type previously in memory. You can also delete all data currently used by CPN.
- **Add:**
  Load data from files; similar to *New* menu, except nothing is removed from memory.
- **Clear:**
  Remove some or all type of data present in memory.
- **Save:**
  Save your current results, save the software state (this allows you to use current parameters as default for future uses), and choose to automatically store the results produced in a pre-defined folder, in the folder containing the input data, or not at all.
- **Parameters:**
  Define how many independent networks to infer from a given set of phylograms, while randomly changing the order of added phylograms and connections between runs (and identifying the best one), how many CPU cores of your computer to use, and the end of line type of the ouput files. Define if phylograms are read as phylograms or cladograms. Determine if the results should contain sequence data and if every phylogram has to be saved in a *.dot* format, which allows to visualize each of them as a network without cycle.
- **About:**
  Review of the software version, how to get it, and contact information.

Detailed description of commands:

Analysis

Produce Phylograms:

Initiates the process of inferring all most parsimonious phylograms from the set of all most parsimonious cladograms and the corresponding sequence alignment. This command becomes active when the following conditions are met:

1. There is at least one uploaded cladogram;
2. A sequence alignment is uploaded;
3. Sequences in the alignment are of the same length;
4. The number of sequences in the alignment matches the number of tip branches in the cladograms.

Produce Network:

Initiates the process of combining all most parsimonious phylograms into a single network graph. This command becomes active when there is at least one phylogram in memory. Because the network inferred may depend on the order in which phylograms, and their specific connections, were added, it is possible to launch several runs, each time randomizing this order, the program finally identifying then the best network found (identified as the networks associated with the minimum number of connections). This option can be set *via Parameters>Number of networks to generate*.

The number of mutations, connections and union nodes is computed for each network. Union nodes are inferred nodes (i.e., absent from the alignment) with at least three connections (i.e., unsampled nodes of degree > 2).

Quit:

Closes the software.

New

- ***File****:*
  Removes all sequences, cladograms and phylograms from memory. It then opens up a new window to select an input file of one of the following formats: Nexus, Newick, or Phylip. Information over the new data are outputted in the console, and key parameters are shown in the input data panel.
- ***Sequences****:*
  Removes all sequences data from memory. It then opens up a new window to select an input

file. Only sequences data are extracted from the file that can be of the following formats: Nexus or Phylip. Information over the new data is outputted to the console, and key parameters are shown in the input data panel.

- ***Cladograms****:*
  Removes all cladograms from memory. It then opens up a new window to select an input file. Only cladograms data are extracted from the file that can be of the following formats: Nexus or Newick. Information over the new data is outputted to the console, and key parameters are shown in the input data panel.
  Warning: Newick trees with information over branch length are phylograms and can't be loaded through this command.

- ***Phylograms****:*
  Removes all phylograms from memory. It then opens up a new window to select an input file. Only phylograms data are extracted from the file. Input files can be of the following formats: Nexus or Newick. Information over the new data is outputted to the console, and key parameters are shown in the input data panel.
  Warning: if the file contains cladograms (i.e., trees without branch length information), they won't be loaded.

Add

- ***File:***
  Opens a new window to select a file. The content of the file is added to the existing data, and can be of the following formats: Nexus, Newick, and Phylip. Information over the new data is outputted to the console, and key parameters are shown in the input data panel.

- ***Sequences:***
  Opens a new window to select a file. Sequences from the file are added to the existing data. Only sequences data are extracted from the file that can be of the following formats: Nexus or Phylip. Information over the new data is outputted to the console, and key parameters are shown in the input data panel.

- ***Cladograms:***
  Opens a new window to select a file. Cladograms from the file are added to the existing data. Only cladograms data are extracted from the file that can be of the following formats: Nexus or Newick. Information over the new data is outputted to the console, and key parameters are shown in the input data panel.
  Warning: Newick trees with information over branch length are phylograms and can't be loaded through this command.

- ***Phylograms :***
  Opens a new window to select a file. Phylograms from the file are added to the existing data. Only phylograms data are extracted from the file. Input files can be of the following formats: Nexus or Newick. Information over the new data is outputted to the console, and

key parameters are shown in the input data panel.
Warning: if the file contains cladograms (i.e., trees without branch length information), they won't be loaded.

Clear

- ***Everything***:
  Removes all sequences, cladograms and phylograms data from memory.
- ***Sequences***:
  Removes all sequences from memory.
- ***Cladograms***:
  Removes all cladograms from memory.
- ***Phylograms***:
  Removes all phylograms from memory.

Save

- ***Current state***:
  Many parameters (such as the number of independent networks to infer, the number of CPU cores to use, etc.) can be defined in CPN. Saving the current state of the program allows to save all parameter settings, and to keep those as default for next time CPN is used.

- ***Phylograms as***:
  Allows you to select a folder, in which all phylograms can be saved. Used in combination with *Parameters>Output phylograms as .dot* allows to produce a network graph from each of the phylogram stored in memory.
- ***Network as***:
  Allows selection of a folder, in which the best networks are saved.
  Best networks are selected as those associated with the lowest number of connections.
- ***Automatically in***:
  This option is used by CPN after every job completion (either inferring phylograms or producing the network).
  - **None**: when a process is completed, results are stored in memory without being saved in a file.
  - **Selected folder**: when a process is completed, the results are automatically saved in the selected folder.
  - **Input folder**: when a process is completed, the corresponding results are automatically saved to the folder of the last imported file. This allows storage of

results according to the organization of your input files. By default, this is the selected option.

Parameters

- ***Number of networks to generate:***
  Allows defining the number of networks to generate from the same set of phylograms, but each time randomly modifying the order in which the phylograms, and their inner connections, are added to build it. This option is implemented because it was found that the order in which phylograms and connections are added may influence the final network graph (Cassens *et al.* 2005). Default value is set to 20.

- ***Read phylograms as:***
  This only affects the *File* menu.

  - **Phylograms**: when using the *File* menu (both *New* and *Add* commands), cladograms will be read as cladograms and phylograms as phylograms. This means phylograms won't be converted. This is the default option.

  - **Cladograms**: when using the *File* menu (both from *New* and *Add* commands), both cladograms and phylograms will be read as cladograms. Thus, every type of phylogenetic tree will be read as cladograms and used for the production of the set of most parsimonious phylograms.

- ***Availables cores:***
  Defines the number of CPU cores you want to use to run the program. More CPU cores means less computation time (in most cases). However, using too much cores will slow down others applications. If you are not sure what to do, we suggest you keep the default value for this parameter, or decrease this number if your computer becomes too slow. This parameter can only be changed before starting a run. By default, the program chooses the total number of CPU cores on the computer, minus one.
  Warning: this command only affects the "produce phylogram" component of CPN. When producing networks, only one CPU core is used, because the job cannot be shared among multiple cores.

- ***New line type:***
  Depending on the program used to open the output, you may need specific line breaks, which are different among operating systems. In most cases, your best bet will be to use the line break type of the OS you use (by default, the program automatically selects the line break type of the OS it runs on).

- ***Output with sequences in file:***
  Phylograms are outputted in Nexus format. If this option is selected, the output file will also contain the sequence alignment used for inferring phylograms. By default, this option is not activated.

- ***Output phylograms as .dot:***
  In addition to saving the inferred network, it may be useful to save all phylograms under the dot format, which allows their comparison with the resulting networks. If this option is selected, every phylogram will be saved in a separate file. By default, this option is not activated.

- ***Reduce cycles:***
  If selected, the program attempts to reduce the length of cycles, by merging some of its edges and nodes (see end of manual for a detailed description of this optional step). By default, this option is not activated. When selected, a new network will be produced and stored in a separate file. *Note that CPN takes sometimes more time to achieve this additional step than the initial network building process*, *which may thus considerably increase the overall runtime of the analysis.*

- ***Reset CPN launch:***
  This command will restore the initial parameter options of CPN (those that were in effect the first time you launched the program). This is the opposite of the *Save>Current State* menu.

About

- ***CPN :***
  This command triggers the display of general information about the software, including version number and contact info.

Progress Panel

The P**rogress Panel** shows the progress of the running job.

- ***Characters analyzed:***
  The number of progress bars mirrors the number of CPU cores available for the job. Each of them indicates the proportion of nucleotides for which most parsimonious ancestral states have been inferred, for the purpose of defining all MP phylograms corresponding to the given set of cladograms. Each cladogram is treated independently from the others, so that each cladogram is handled by a specific CPU core.

- ***Cladograms analyzed:***
  This progress bar shows the proportion of cladograms for which the analysis is over. If requested, results will be saved automatically upon completion of the progress bar.

- ***Network building step:***
  Proportion of steps accomplished for building one network from the set of all most parsimonious phylograms. Because the total number of steps may increase during the run, this is not a good indication of the overall progress, but it is also the best possible representation for a job, as run length cannot be predicted *a priori*. When the progress bar is full, one network has been built and will be stored in memory, or not (depending on the previously saved networks).

- ***Networks built:***
  The current number of inferred networks compared to the total number of networks required by the user. If requested, results will be saved automatically upon completion of the progress bar.

- ***Console:***
  This part of the GUI gradually adds information over the job in progress.

Input data Panel

This panel gives an overview of the data loaded in the program. The data for the two major functions of CPN are shown separately.

Produce Phylograms

This button becomes active when the boxes *Cladograms* and *Sequences* are checked (i.e., when appropriate data are available to perform the analysis).

- ***Number of taxa:***
  Number of taxa found in the loaded input files.
- ***Number of characters:***
  Number of nucleotides contained in the loaded DNA sequence alignment.
- ***Number of cladograms:***
  Number of cladograms for which the program is ready to infer phylograms.
- ***Tree length:***
  Tree length of cladograms.

Produce Network

This button becomes active when the box *Phylograms* is checked, which is the case when at least one phylogram (loaded or inferred) is in memory.

- ***Number of phylograms:***
  Number of phylograms that will be combined into a network graph.
- ***Number of mutations:***
  Number of mutations in the network currently in memory (i.e., the network identified as the one associated with the minimum number of connections).
- ***Number of connections:***
  Number of connections in the network currently in memory (i.e., the network currently identified as the one associated with the minimum number of connections).
- ***Number of union nodes:***
  Number of union nodes in the network currently in memory (i.e., the network currently identified as the one associated with the minimum number of connections). Union nodes are unsampled nodes of degree > 2.

Data input and output

Input files can be in one of the following format: Nexus (sequential), Newick, Phylip (sequential or interleaved).

Nexus files can contains the following blocks:

- Data, Characters (or Character).
- Taxa.
- Trees (or Tree).
- CPN.

However, the Nexus input file containing the trees cannot include a “Translate” block (assigning a number to each sequence/taxon name). For example, if using PAUP to infer the MP trees, use the format=altnex option with the command “savetrees” (i.e., the name of each sequence should be included directly in the Newick representation of the trees).

Phylip files can be interleaved or sequential, and tabulated or not. Sequences can contain A, C, T, G, or any other IUPAC character, plus “-” and "?" which, by default, identifies gaps and missing characters, respectively. Other characters can also be assigned to gaps or missing characters, if specified in the data block of a Nexus file. Any other character will not be recognized by CPN, and will generate an error message indicating that the number of sequences loaded is inconsistent with the expected number of sequences, as defined in the beginning of the Nexus file.

Note that gaps will be treated as missing characters by the program.

Important for preparing the input file:

1. include only one sequence per allele (haplotype). If the data files contain more than one sequence per allele (i.e., two or more sequences are identical), the program will stop when attempting to generate the network and ask the user to remove identical sequences from both the sequence alignment and trees, while indicating which sequences are redundant. Note that when determining whether two sequences are identical, two IUPAC characters with at least one nucleotide in common (e.g., Y and C, or R and M) will be considered separated by a distance of 0 (no substitution).

2. avoid the use of '#' (unless when initiating a Nexus file).

3. sequence names should contain only the following type of characters: letters (A-Z), numbers (0-9), underscore (_), minus (-) or dot (.). Sequence names cannot start with the letter “U”, and cannot include blank spaces, tab characters, commas (,), colons (:), or semi colons (;).

4. make sure sequence names in the sequence alignment and in the cladograms/phylograms are identical, otherwise, CPN will not be able to match a sequence in the trees with that same sequence in the alignment (an error message will be reported and the analysis will stop).

For convenience of job automation, we defined a new type of Nexus block. It must be delimited by

"Begin CPN;" and ended with "End;". Each line in-between should correspond to parameters and corresponding values separated by an "=" character and lines should end with semi-colon. Table 1 shows parameters and authorized values. Parameters and values are not case sensitive. *Italic* words are word associated with particular function and should be correctly written.

Output files containing the inferred phylograms are in Nexus format, and can be visualized by different programs, e.g., FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Output files with the final network are written in the DOT language, and can be visualized by a standard network drawing program such as GraphViz (http://www.graphviz.org).

Table 1 : CPN block (Nexus format)

Cores	Unsigned integer	Number of CPU cores used for phylograms production
Output	*None*	Results will not be stored
	Input	Results will be stored in the folder of the last input file imported
	Path to a folder	Results will be stored in the folder
FileAsPhylograms	*True* or *1*	Phylograms are read as phylograms
	Other	Phylograms are converted to cladograms
KeepSeq	*True* or *1*	Sequences are stored with resulting phylograms
	Other	Sequences are not stored
KeepDot	*True* or *1*	Every phylogram is stored in a separate .dot file
	Other	Only the networks are stored in .dot file
NetworksToGenerate	Unsigned integer	Number of independent networks to generate
ReduceCycles	*True* or *1*	The program attempts to reduce the length of cycles in the best networks, and save the resulting networks in separate files
	Other	No reduced network is produced
LineBreak	*Windows* or *W*	Line break used: \r\n
	Mac or *M*	Line break used: \n
	Linux or *L*	Line break used: \r
UseAsDefault	*True* or *1*	Use parameter specifications in this file as default when launching CPN. Parameters in the files becomes initial default parameters
	Other	Default parameters are unchanged.
Files	File paths separated by coma (",")	It should be set at then end of the block. Those files are loaded in CPN.
Reset	*True* or *1*	Removes previously loaded data
	Other	-

Automation Pipelines / CommandLine

CPN can be used as a command line program in the following way:

- Java –jar CPN.jar <option1>=<value1> <option2>=<value2>… <filename1>=< file path> <filename2>=< file path>

You can include as many options as you wish (see option description below). If an option is unrecognized, CPN interpret it as the name of a new file you want to import.

Exemple of run:

- Java –jar CPN.jar file1=/path/sequenceAlignment.nex file2=/path/mpCladograms.tre

This line will simply load the two files *sequenceAlignment.nex* and *mpCladograms.tre* and run the analysis. If options are defined, they should precede the input files on the command line, because as soon as sequences and cladograms are available to the program, the analysis is launched.

Options can be one of the following:

- Cores
  Associated value must be any unsigned integer. This defines the number of CPU cores to use for phylogram production. By default, the program uses all available CPU cores, minus one.
- Save
  Associated value must be a path to a file. When this option is read, it immediately saves results into the selected file. Otherwise, results are stored in folder of the last imported file.
- Out
  Associated value must be a path to a folder. This configures the path to the folder where results are stored. If no output folder is defined, results are saved in the folder of the last imported file. If this option is absent from the command line, results will not be stored.
- StoreUniquePhylograms
  Associated value must be "True" or "1", anything else meaning "False". In some cases, different Newick representations can correspond to an identical phylogram. When loading phylograms, CPN only imports those that are not already in memory. If prune is set to "True" or "1", the program will keep non redundant phylograms and store them in the folder of the imported file. This allows a user to determine the number of phylograms from the input file that are really different. Default value is "False".
- FileAsPhylograms
  Associated value must be "True" or "1", anything else meaning "False". If this is set to "True" or "1", phylograms are read as phylograms. See *Parameters* menu for details. Default value is "True".
- MakePhylograms
  Associated value must be "True" or "1", anything else meaning "False". If this is not set to "True" or "1", CPN will not try to infer new phylograms (and will produce the network only from available phylograms). Default value is “True”.
- MakeNetworks
  Associated value must be "True" or "1", anything else meaning "False". If this is not set to "True" or "1", CPN will not build a network. Default value is "True".
- NetworksToGenerate
  Associated value must be any unsigned integer. This defines the number of independent networks to generate. By default, this number is set to 20.

- KeepSeq
  Associated value must be "True" or "1", anything else meaning "False". If this is set to "True" or "1", the sequence alignment will be stored in the Nexus file created to store the produced phylograms. By default, it is set to "False".
- KeepDot
  Associated value must be "True" or "1", anything else meaning "False". If this is set to "True" or "1", every phylogram built will be stored in a separate .dot file. By default, this is set to "False".
- ReduceCycles
  Associated value must be "True" or "1", anything else meaning "False". If this is set to "True" or "1", the program will attempt to reduce the length of cycles in the best networks, and save the resulting networks in separate files. Default value is set to "False".
- LineBreak
  Associated value must be "Mac", "Linux" or "Windows". This defines the line break type of the output files. By default, the program recognizes the operating system it runs in, and uses the corresponding line break.

Description of additional step to the algorithm combining multiple trees in a network

Additional procedure attempting to reduce cycles length

x	Labeled node of the network graph (sampled alleles or inferred alleles)
C(**x**;**y**)	Direct connection between nodes **x** and **y**.
LEN(**x**;**y**)	Length of the path between nodes **x** and **y**.

Store de final network graph built according to algorithms described in Cassens *et al.* (2005).

List all labeled nodes **n** with at least two connections of length > 1.

For each node **n** :

- For each labeled node pair **m1** and **m2** connected to node **n** with length > 1:
    - List all paths of the network linking nodes **m1** and **m2** without going through **n**.
    - If there are no such path, move to the next node pair.
    - Otherwise, take the shortest path of size LEN(**m1**;**m2**).
        - Identify LEN(reduced) = MIN(LEN(**n**;**m1**);LEN(**n**;**m2**)) – LEN(**m1**;**m2**).
        - If LEN(reduced) > 0 ; cycle length can be reduced :
            - Change C(**n**;**m1**) to C(new**Node**;**m1**) with length LEN(new**Node**;**m1**)=LEN(**n**;**m1**)-LEN(reduced)
            - Change C(**n**;**m2**) to C(**newNode**;**m2**) with length LEN(new**Node**;**m2**)=LEN(**n**;**m2**)-LEN(reduced)
            - Create new connection C(**n**;new**Node**) of length LEN(**n**;new**Node**)=LEN(reduced)
            - This reduces the number of mutations in the final network by LEN(reduced), and adds one new connection and one new node of degree ≥ 3 (union node).
            - check whether the modified network still contains every MP phylograms that were considered to build it; if not, undo this specific modification