Constructing Phenetic and Phylogenetic Relationship Using Clad'97

Brian Rahardi[1*], Estri Laras Arumningtyas[1], Wayan Firdaus M[2]

[1]Biology Department, Faculty of Sciences, Brawijaya University, Malang-Indonesia
[2]Computer Science Program, Faculty of Sciences, Brawijaya University, Malang-Indonesia

ABSTRACT

Relationship construction has a very important position in classification process for arranging taxonomy of organism. In the world of taxonomy, there are two the most familiar relationship diagram, cladogram and phenogram. In every construction activity, a researcher is always facing character state data from taxa that becomes components of the diagram. Calculation that is used for construction is often incorporate iterative or repetitive process that needs time and precision. The existence of calculating tools that produces both text and graphical output are hopefully decrease time and error during construction. Basic algorithm that is used in calculation is for phylogenetic construction by Kluge and Farris in 1969,for phenetic construction using cluster analysis with slight modification. Basic common algorithm used in the software is by calculating two dimensional arrays of taxa x characters matrix and creating distance or similarity matrix. In more detail the program creates one dimensional array of taxonomical object and each object has some other one dimensional array containing data commonly exist in a taxonomic unit. The relationship between one object and theother are regulated by an object that created by class representing taxonomic tree. Cladogram is constructed by calculating nearest distance between each taxon (OTU) and creating one HTU in every bifurcation. Phenogram is constructed agglomeratively by searching highest similarity between taxon then grouped into new taxon. Program calculates numerical data after we do character scoring. Final result for each user may be different; this may be due to decision by user during construction process. This paper hopefully attracts people from systematic computation to develop further into open source software and multi-platform feature.

Keywords: software, phenetic, phylogenetic.

INTRODUCTION

In the science of taxonomy classification process has a very important position. Indonesia is well known as a country that has a wealth of biodiversity. Biodiversity requires data collection and inventory building. Taxonomy is a branch of biological sciences in addressing the most appropriate data collection and inventory of species.

The purpose of this paper is to introduce the working principle and way of operation of a calculation with an additional tool of the graphical display of relationship diagrams and general information required in the analysis of relationship. To further this program called Clad'97 to facilitate reference to the construction program and the phylogenetic relationship are phenetic this. In the preparation or construction of relationship, and character state, first character is converted to the form of numbers that can be done with or without weighting.

Corresponding address:
Brian Rahardi
Biology Department, Faculty of Sciences,
Brawijaya University, Malang, Indonesia 65145
Email : brian_rhardi@ub.ac.id

MATERIALS AND METHODS

Algorithms that used in the preparation of relationship are the phenetic algorithm by Sneath-Sokal (1973) and for phylogenetic construction using algorithm by Kluge Farris (1969). Classification is done by grouping of taxa by phylogenetic lineage, as illustrated by a cladogram. In cladogram construction of the character x taxon matrix can be done by checking the status of each character and then subsequently breaks it, the lineages, taxa that have the characteristics derived. If, however, there are many taxa, more characters, or a lot of characters that are not compatible (indicated by homoplasy), a more rigorous methods may be necessary to determine cladogram those who have the least number of evolutionary steps in terms of character state changes.

In contrast to the phylogenetic classification, namely the taxa grouped by inherited traits,

phenetic classification is the grouping of taxa from the overall similarity, regardless of whether these similarities or synapomorphy and symplesiomorphy in the phylogenetic sense. Of the many methodologies contained in phenetic group, (including quantitative phylogenetic analysis, multivariate statistical analysis, and non-hierarchical classification), cluster analysis is the most commonly used in phenetic determine a classification scheme [1].

This software was written using C++ programming language. Most importantly, C++ adds object-oriented programming style in C [2]. A taxonomic unit can have several properties symbolized by an object in programming:

```
struct CTaxonomicUnit
{
BOOL    m_bComponentUnit[MAXTAXA]; ending marker.
Int     m_nMyNearest; nearest object marker used in phenetic.
Int     m_nCharacter[MAX]; character it contains.
float   m_nDistance[MAX]; distance to other unitsin phylogenetic.
CString m_strName; it stores name of this unit.
BOOL    m_bIsSelected; this is marker for unit still operable or not.
STATE   m_state;
};
```

Basic algorithms used phylogenetic construction

Clad'97 program uses an algorithm based on the construction of phylogenetic relationship Kluge and Farris in Radford (1986):

1. Establishment of an array of objects that OTU, HTU and Ancestor from the class CTaxonomicUnit with each object have the status of the character and name of each :

```
CTaxonomicUnit unit;
unit.status = OTU; or
unit.status = HTU; or
unit.status = ANC;
```

Creation of an array to store these values can be denoted by the following variables:

```
cTextBuffer[] = contents of text file;
nNumberOfTaxa = part from cTextBuffer[] contains number of taxa;
nNumberOfCharacterStates = part from cTextBuffer[] containing
number of character state;
nCharacterStatesData[nNumberOfTaxa X
nNumberOfCharacterStates] = part from cTextBuffer[] containing
data of character state;
```

2. From the character x taxon matrix, compute the distance between each pair of taxa (including ANC) and Tabulate the data in the distance matrix. The distance is defined as the sum total of the difference in character state between the two taxa character. Calculations were performed using following equation :

$$d(X,Y) = \sum_{i=1}^{n} |Vx - Vy| \quad \text{..... (1)}$$

3. From the status of the characters can be searched distance from the each using the distance() function and the results will be stored in an array from the structure member variable m_nDistance from CTaxonomicUnit:

```
unit.distance[to_taxa] = distance();
```

4. Each object has a value of taxa that have a distance to each other taxa object can determine the object closest taxa to find the smallest distance value using *SeekNearestTaxa()* member function from the class *CPhylogeneticTree* and the results will be stored in member variables *MyNearest:*

```
unit.MyNearest = SeekNearestTaxa();
```

5. Next is to determine the location of HTU by the shortest distance from an OTU object remaining with two taxa that have been placed:

```
unit.Distance = distanceHTU();
```

6. HTU which has been formed is then determined the status of the character based on the status of characters from the the three taxa that form the branching object by calling the member function *SetHTUCharacter():*

```
tree.SetHTUCharacter();
```

7. Steps 5 and 6 is repeated until the entire OTU and HTU are placed in the diagram.

Basic algorithms used in phenetic construction

Clad'97 program uses an algorithm based on the construction of relationship phenetic Sneath and Sokal in Radford (1986):

1. The input status of the characters is handled by member functions *SetCharacter():*

```
taxonUnit[].Character[] = value;
```

and input the name of taxa are handled by calling the member function *setName():*

```
taxonUnit[].strName = name;
```

2. Calculate the similarity between each pair of taxa with overall similarity coefficient (Sjk). Coefficient of the overall similarities is defined by various formulas, one of which is generally used are:

$$Sjk = 1 - \frac{\sum_{i=1}^{n} \frac{|Xij - Xik|}{Ri}}{n} \quad \text{..... (2)}$$

3. For phenetic construction, which is the coefficient from the similarity was calculated using the overall *similarity()* function. The calculation result is stored in member variable *distance:*

```
unit.distance[to_taxa] = similarity();
```

4. Furthermore, taxa pair searched with the greatest similarity value. *SeekNearestTaxa()* member function allows each object knows the object taxa other taxa that have the greatest similarity:

```
unit.myNearest = SeekNearest();
```

and the greatest similarity value can be found by comparing the *distance* of each object taxa.

5. Therefore, it can set up a replacement taxa object, from the two taxa objects to be represented by objects taxa HTU, for further calculations are performed using the *similarityNew()* members function.

6. Step 2 to 5 repeated as long as there is still the object of taxa that have variable *IsSelected* contains FALSE value.

RESULTS AND DISSCUSION

The construction process of relationship

When you first run the program, will be shown the welcome screen (splash screen) with the inscription Clad'97 like figure 1 below:

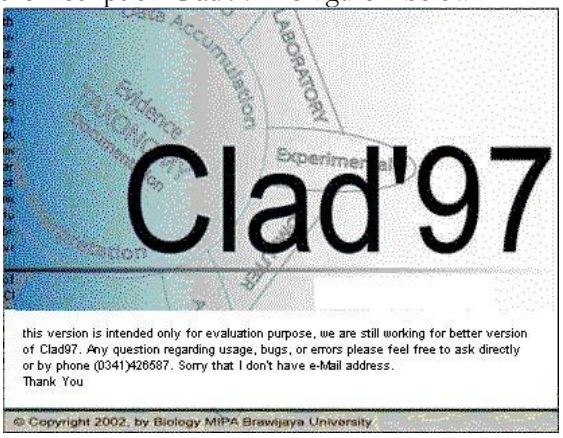


Figure 1. Clad'97 program start-up screen

After that the main window will appear as in Figure 2 which will display the results of calculations, diagrams and information graphics taxa formed. The main window is divided into two parts, the upper display screen is for displaying graphics, and panel for text output at the bottom. Bottom panel has three output windows those are a text box "Output Summary" which displays the number of taxa information, the number of characters the status of each taxa, the status of the overall character of the taxa included and the results of calculations during the construction process. Display on the right is a list box with the name "Taxa created so far" that will display the taxa that have been created during construction, if one taxon name which is displayed is selected, it will display information about the taxon is in the "Short Information" below .

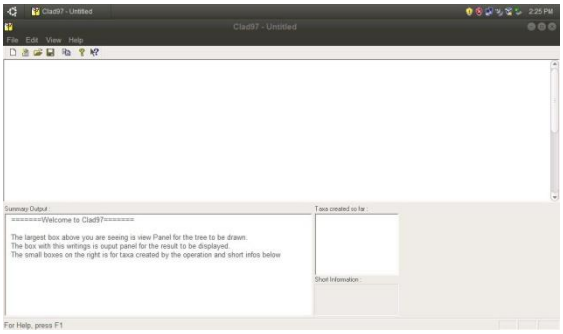


Figure 2. The main window screen

Clad'97 has four menus in its menu bar as shown in Figure 2, the menu is:
Files, which have sub menus:
- New, sub menu will start the construction work of relationship through a wizard that will guide the steps of inputting data. If the window you have open already contains data then this menu will open a new window.
- New Text, a sub menu will call the notepad text editing program which is a program that is integrated in every installation of Microsoft Windows. We enter data by creating a text file that contains the number of taxa, the character state, the label for character state of the taxa, for then we open the text file with a program Clad'97.
- Open, sub menu allows you to open a text file containing the data to be calculated which we created earlier.
- Exit, sub menu will end the session and exit calculations Clad'97

On the File menu also featured four file names that was last opened. Edit, the menu has only one sub menus Copy will copy the view from the main window into the clipboard, so it can be included in any image processing program for text processing and then can be saved.

Clad'97 program taking input in two ways. The first way is to use wizard that guides determination to enter data and analysis at each step. The second way is to use a text file created using a text editor. At the panel manufacturing output is still using a dialog box with the size of the patent so it can not adjust the screen size larger or smaller.

Data Input Through the Data Input Wizard

By selecting the File menu and sub menu New users will be given the choice of construction type of relationship that will be done through the choice of radio buttons, the buttons that only allow the user selects one of two options offered, as shown in Figure 3 below:

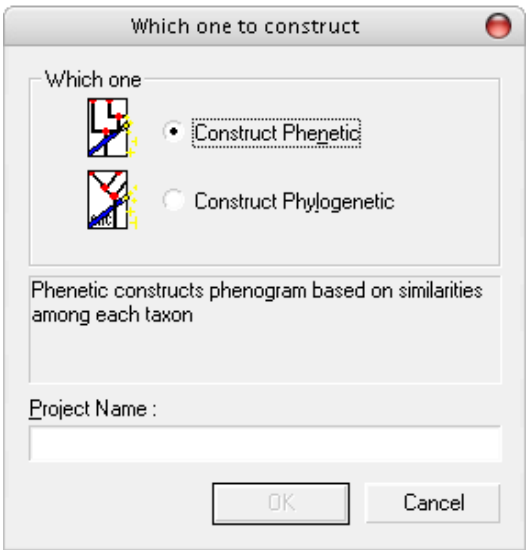


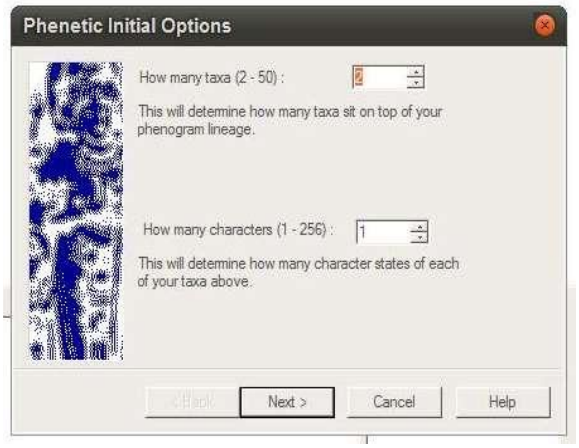
Figure 3. Construction options dialog box

Users can choose one of two construction types offered relationship, that is, phenetic when the construction is based on overall similarity of taxa and the phylogenetic distance is construction is based on the evolution of each taxon. To be able to continue the user must fill in project name. The user will be guided in the steps of inputting data by the system wizard dialog (Figure 4.).

Input text files

In figure 4A, the first wizard dialog box provides a text box for the number of taxa will be operated and the number of character state are owned by their respective taxa. In this version there is a limitation of the maximum number of taxa, 50 taxa and the number of character state that can be operated the maximum characters is 200 character states. This limitation is caused by the use of arrays for data storage in memory and requires improvements such as more flexible memory usage according to the size of the data.

A.



B.

C.

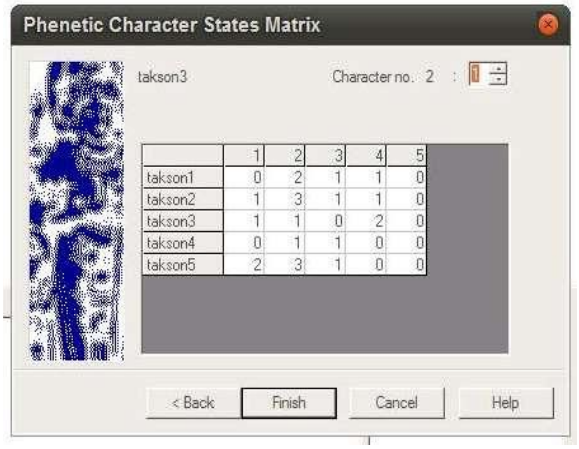


Figure 4: Wizard dialog boxes input data: A. Entering the number of taxa and characters, B. Enter the name of taxa, C. Inserting the value of character

In figure 4B, the second wizard dialog box provides a charging box taxa name and a box containing another list of names of taxa that have been entered. Users can remove the name of the taxon from the list by selecting a name that will

be deleted and pressing the Delete key. If the number of names have not reached the number of taxa are loaded to be operated then the wizard will not allow the user to advance to the next dialog box. In figure 4C, the third wizard dialog box provides a form filling state of the character values, each cell can be filled with numbers 0 to 9 which is a numerical value after the conversion, quantification or weighting of characters. In everyday, the use of wizard dialog is felt less practical and more users are using direct input from a text file. But help users who do not know the format of the input text file data to still be able to use the program Clad'97

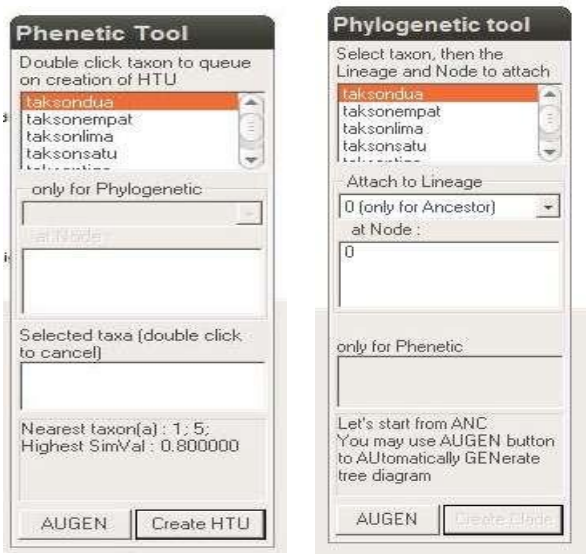
.File menu and sub menu New Text will call the notepad program used to create a text file to be opened using the program Clad'97. The text file has a format with an example:

```
04
005
*
"A"=12210
"B"=12100
"C"=01010
"D"=00001
```

Two digits at the top show the number of taxa, maximum allowable taxa is 50 taxa. The three digits on the next line are the number of character state of each taxon, the maximum allowed is 200 character state of each taxon. Any construction used relationship least 60 characters and it is considered to have been sufficient. An asterisk marks the beginning of the reading of character state. Taxon name is written between two double quote character in front of the data with granting the sign "=" between the character and state data label. The use of the input text file is more practical for those who are familiar or frequently make changes in the input data because it does not have to go through the stages of the wizard dialog.

Construction of relationship through Phenetic / Phylogenetic toolbox

Tool in figure 5 contains information regarding the ongoing process and is used to make decisions during the construction process takes place. This toolbox will close automatically when the construction process has been completed. Taking into account one OTU or group at a time, then the result is a dichotomous phenogram [3]. In principle, cluster analysis is a generic term for a mathematical method that shows which objects of concern in a set, along with many others [4].



A **B**

Figure 5: Tool for: A. Construction Phenetic, B.

Construction of Phylogenetic

Presentation of the results of calculations

The results of calculations are presented in the form of text and graphical diagrams. In the results in text form are divided into three parts, the first is the number of taxa and number of the character state that have been operated, the second is the raw data of the character state that has been entered, the third part is the overall similarity coefficient table produced by the construction phenetic or distance table taxa generated by phylogenetic construction. Output (output) in the form of graphs has not been shown in a common format but tried to be the results of calculations can be represented visually. For output (output) phenetic construction, the new taxon is formed from HTU named taxa merging with its formation and serial number along with the overall distance similarity coefficient of taxa that formed it. Output graphs to illustrate the construction of phylogenetic lineages between HTU and OTU nearby. Under each HTU and OTU are the character state of the HTU or OTU.

The design software is still primitive for calculating the basic construction process automation in phenetic and phylogenetic relationship. The Use of this program can shorten the construction time which at first took place in a matter of hours or days become a matter of minutes. The application of this program may have little problem for binary stated characters, that means character consist of only zero and one for example in microbiological research which generally use the results of positive and negative readings [5]. The program is also relatively

commonly used in molecular research base level using data from the fragmentation of DNA strands as in the RAPD [6] and RFLP analysis [7]. In general, DNA fragmentation data rated the use of positive integers. Quantification was based on the thickness of the fragments fit with the theory that determines the weighting of the thickness of the fragment. The program is still at an early stage because it is compiled only for Windows while recent software developments require a program that runs multi-platform. With a program that is able to run multi-platform open the possibility of making the program as an open source program (Free Open Source Software). At the time of the source code available to the public and announced by the creator or initiator to a variety of media and open source platform to attract people interested in the project [8].

CONCLUSIONS

The Use of this program can shorten the construction time which at first took place in a matter of hours or days become a matter of minutes. The application of this program may have little problem for binary stated characters, that means character consist of only zero and one for example in microbiological research

REFERENCES

[1] Radford, Albert E., 1986, Fundamentals of Plant Systematics, Harper & Row Publishers, Inc., New York

[2] Booch, Grady, 1994, Object-Oriented Analysis and Design with Application, The Benjamin/Cummings Publishing company, Inc., California

[3] Panchen, Alec L., 1994, Classification, Evolution, and The Nature of Biology, Cambridge University Press, USA

[4] Romesburg, Charles H., 1984, Cluster Analysis for Researchers, Lifetime Learning Publications, California

[5] Suharjono et al., 2007. Sistematik Numerik Strain-Strain Anggota Genus Pseudomonas Pendegradasi Alkilbenzen Sulfonat Liniar Berdasarkan Sifat Fenotip dan Protein Fingerprinting. Biota, 12(1), pp.47-54

[6] Arumingtyas, E.L. et al., 2010. Polymorphism Analysis of Kenaf (Hibiscus CannabinusL.) Mutants Based on Random Amplified PolymorphicDNAs (RAPDs). Journal of Materials Science and Engineering, 4(2), pp.56-62.

[7] Pahlevi, M.R., 2007. A Study of Genetic Variation And Relationships Among Bali Cattle from P3bali Pulukan and Tabanan by RFLP. In International Conference on Molecular Biology of Life Sciences. International Conference on Molecular Biology of Life Sciences. Brawijaya University.

[8] Sturmer, M. 2005. Open Source Community Building. Faculty of Economics and Social Science of the University of Bern, Switzerland