Anacê: Phylogenetic Trees Drawing Web Service

Hélio Augusto Sabóia Moura
MPComp - Integrated Master on Applied Computation
State University of Ceará
Fortaleza, CE, Brazil
helio.moura@uece.br

Gerardo Valdísio Rodrigues Viana
Dept. of Computer Science
State University of Ceará
Fortaleza, CE, Brazil
valdisio@uece.br

***Abstract*—In this paper, we describe a tool, called Anacê[1], to draw phylogenetic trees in diverse topologies. Available via Web service, and developed in Scala, this tool can be used in any computational platform in the interactive form or as subprograms in any application or programming language. The generated trees is exported in SVG (Scalable Vector Graphics) image formats that are independent of computational platforms. The tool easily draws trees from the distinct forms of tree's representation, generated by other software. It is meant to be used by researchers and as a learning tool.**

***Keywords*-computational biology; phylogenetic trees; Web service.**

I. Introduction

A phylogenetic tree can be depicted in several topologies. For example, a given phylogenetic tree can be represented as a rectangular cladogram, an inclined cladogram, a phylogram, a radial tree, a free tree with or without root or still in the textual form using the Phylip standard [1]. In this paper, we propose a tool, called Anacê, to draw trees in any[2] of these formats. It is available via Web service and is implemented in Scala [2]. We created a site with a Anacê's tutorial [3].

Since a user developing a Web service does not need to know on which programming language a client will implement his/her applications, Anacê uses RESTful [4] Web services, and so, the only resource needed is a library to use HTTP protocol, wich is available in the most of the programming languages.

The main objective with this tool is to have a unique point of execution, preventing each different computational environment to have a copy of the library installed in the client application. Another advantage is that the improvements in this service and new methodologies developed become automatically available on the Anacê web site, thus assuring that the latest version of the tool is used.

Several tools have been developed to draw phylogenetic trees. For example, DrawTree [1], TreeView [5], PhyloDraw [6] and Spectrum [7]. In order to be used all these tools need to be installed in the user machine, and so they request specific computational resources. There exists also the Web server called Phyml [8] that uses the maximum-likelihood method to infer phylogenetic trees.

In this work our intention is not to infer [8] neither to reconstruct [9] phylogenies. Actually, we aim to draw trees from its distinct forms of representation generated from other software.

In Section II, we review concepts about Phylogenetic Systematics with emphasis on phylogeny, cladograms and phylogenetic trees. In Section III, we describe methods for phylogeny construction from matrices of distances and sequences of genes and proteins [8]. In Section IV, we describe the functionalities of our tool. Finally, in Section V, we present the conclusions of the work.

II. Phylogenetic Systematics

The fundamental concept of the evolution is that for any two species there was at least one common ancestral species; for three species, the hypothesis is that two of them have an ancestral that is not common to the third one [10]. Following this reasoning for all species, we get a sequence of fragmented divisions from the first ancestral species. The diagram that represents this evolutive history of the species is called, generically, of phylogeny or phylogenetic tree [11].

In modern biology there exists a research area called Phylogeny Systematics that has as objective to understand the relationships between all living beings and then to infer the history of their lives and origins [12]. The term phylogeny is used to designate any diagram that presents the phylogenetic relations between the species in study. A cladogram corresponds to the relationships of a group of species (taxons) with common ancestor, whereas a phylogenetic tree, moreover, expresses the relations of the type ancestral-descendants. There exists still the term phylogram that is a special rectangular cladogram, in which the size of its branches is proportional to the evolutive distances between taxons.

To illustrate these differences, we present in Figure 1.a an inclined cladogram with four recent species A, B, D, E and a fossil species $C^{\star}$. One of the possible phylogenetic trees for this cladogram is shown in Figure 1.b, where F is

[1] The name Anacê comes from Tupi, a brazilian native idiom, and means parenthood.

[2] In time of this article not all formats are available yet.

ISBN: 978-1-61208-137-3

a common ancestral to A and B and G is a species common to all other species.

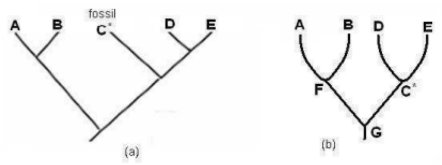


Figure 1. (a) Inclined cladogram with four recent species and one fossil. (b) A possible phylogenetic tree [10].

Phylogenetic trees can be rooted or not. In a rooted tree it is possible to introduce the notion of ancestral traces (plesiomorphics) and derivatives (apomorphics). It is observed in Figure 2.a that the evolutive sequence of some tetrapodies (vertebrate terrestrial that possess four members) is clear and the control group *fishes* is identified, what does not occur in the non-rooted tree shown in Figure 2.b.

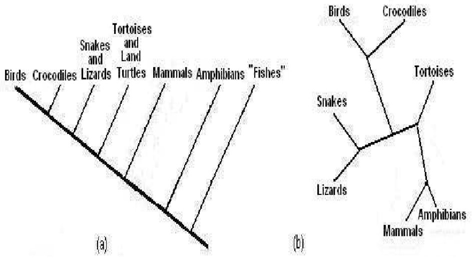


Figure 2. (a) A rooted tree. (b) A non-rooted free tree [13].

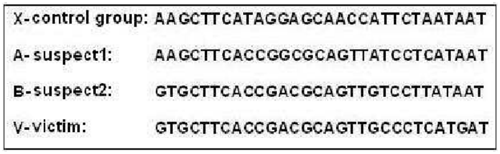
In general, in a phylogeny it is not necessary to identify ancestors nor the dates in the ramifications. However, the sequential order of the evolution must always be shown. Therefore, if the external group is biased to the left, recent will be biased to the right, or vice versa, in such a way that it is possible to distinguish the ancestral characteristics from the derivatives [14].

A phylogenetic tree can also express the evolutive distance between the species, in a such way that the length of its branches or edges is proportional to that distance. Thus, two species that have higher similarity will be near, otherwise they will be distant from each other.

III. Methods for Construction of Phylogenetic Trees

Anacê has as input, in the basic format, a text file that contains the distance matrix or a representation enclosed in parentheses in the Phylip standard format [1]. Both generated from programs that analyze and make alignments of sequences of nucleotides or amino acids [15].

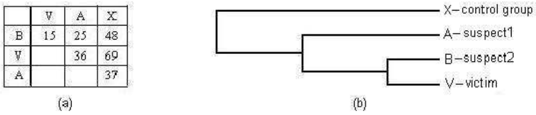
To illustrate some of these forms, we use an example of DNA test to identify which of the two suspects A or B had transmitted the HIV/Aids virus to a victim V of rape. In Figure 3 are shown sequences with 30 nucleotides of the involved ones in the test, where sequence X belongs to a person carrying the virus, however, not related with the crime. In this case, he/she corresponds to the called control group, or external group.



X-control group:	AAGCTTCATAGGAGCAACCATTCTAATAAT
A-suspect1:	AAGCTTCACCGGCGCAGTTATCCTCATAAT
B-suspect2:	GTGCTTCACCGACGCAGTTGTCCTTATAAT
V-victim:	GTGCTTCACCGACGCAGTTGCCCTCATGAT

Figure 3. DNA Sequences [13].

Comparing every pair of sequences in Figure 3 we have the following percentages of distinct characters: $XA = 8/30$, $XB = 12/30$, $XV = 13/30$, $AB = 5/30$, $AV = 6/30$ and $BV = 3/30$. From these values we obtain the results in Figure 4.a, represented by the matrix of distances in relative values. Figure 4.b shows the corresponding rectangular cladogram. Similar matrix, with proportional values, can be obtained by running the programs Clustal and ProtDist in Phylip package [1].



	V	A	X
B	15	25	48
V		36	69
A			37

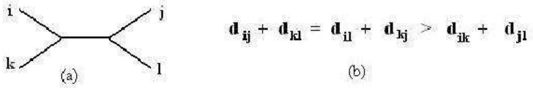
Figure 4. (a) Matrix of distances of the elements in the test. (b) Corresponding phylogeny represented by a rectangular cladogram [13].

From Figure 4.b we conclude that suspect2 (B) is the culprit for the crime, since that one has more common characteristics with the victim (V).

The method used to generate the phylogenetic tree shown in Figure 4.b was the Neighbor-Joining [12]. This method uses the concept of distance in a metric space [11] given by the function $d : E \times E \to R$, such that are valid the following properties for any distinct elements x, y and z of E:

- $d(x,x) = 0$ and $d(x,y) > 0$ (i.e., d is a non-negative function)
- $d(x,y) = d(y,x)$ (i.e., d is symmetric)
- $d(x,y) \leq d(x,z) + d(y,z)$ (i.e., d satisfies the triangle inequality)

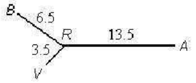
We say that a metric space is additive if, and only if, given four elements i, j, k and l, represented, for example, in Figure 5.a, the relations shown in Figure 5.b are valid.



$$\mathbf{d}_{ij} + \mathbf{d}_{kl} = \mathbf{d}_{il} + \mathbf{d}_{kj} > \mathbf{d}_{ik} + \mathbf{d}_{jl}$$

Figure 5. Valid relations in an additive metric space.

In the matrix in Figure 4.a, B and V are the nearest neighbors, and $AV > AB$, and so in an additive metric space we would have the situation presented in Figure 6.

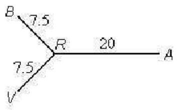


$$\begin{cases} AR + RV = AV = 17 \\ AR + RB = AB = 20 \\ RB + RV = BV = 10 \end{cases} \quad \begin{cases} AR = 13.5 \\ RV = 3.5 \\ RB = 6.5 \end{cases}$$

Figure 6. (a) Determining the ramification point (R) of the line segment BV in an additive metric space.

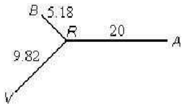
We observe that the system in Figure 6 is always feasible, with unique solution given by $AR = (AV + AB - BV)/2$, $RV = (AV+BV-AB)/2$ and $RB = (AB+BV-AV)/2$. In this case, the ramification RB starting at BV always exists if the solution of the system is positive.

In case this does not occur, the space is not metric and point R is not defined. To fix this problem, in order to show the relationship between the species, we use the methods Unweighted Pair-Group Method using Arithmetic average (UPGMA) when the evolution taxes are approximately constant between different species and Weighted Pair-group Method using Arithmetic average (WPGMA) corresponding to a weighed mean, in order to better reflect the neighborhoods. For example, for $AB = 19$, $AV = 36$ and $BV = 15$, the solution of the system would be $AR = 20$, $RV = 16$ and $RB = -1$. Using the methods mentioned above, we obtain the solutions shown in Figure 7 (method UPGMA) and in Figure 8 (method WPGMA).



$$\begin{cases} BR = \frac{BV}{2} = \frac{15}{2} = 7.5 \\ RV = BV - BR = 15 - 7.5 = 7.5 \\ AR = \frac{(AB - BR) + (AV - RV)}{2} = 20 \end{cases}$$

Figure 7. Solution obtained by the method UPGMA.



$$\begin{cases} BR = \frac{AB}{AB + AV} * BV = \frac{19}{19+36} * 15 = 5.18 \\ RV = BV - BR = 15 - 5.18 = 9.82 \\ AR = \frac{(AB - BR) + (AV - RV)}{2} = 20 \end{cases}$$

Figure 8. Solution obtained by the method WPGMA.

We observe that UPGMA halves a segment, while WPGMA makes it proportionally, better reflecting the similarities. For this reason, we chose this method in our implementation.

The algorithm finds, in the distance matrix, the nearest neighbors, and then the ramification point is computed, initially in accordance with the criterion presented in Figure 6 (Neighbor-Joining), otherwise, with that described in Figure 8 (WPGMA). We observe that segment BV was divided proportionally and accurately, while segment RA has reduced size to make the considered points pertaining to a *new* metric space. These procedures are successively repeated until all elements in the matrix have been considered.

Applying this method for the matrix in Figure 4.a, we obtain the phylogenetic tree indicated in two distant forms in Figure 9, whose representation in the text format using the Phylip standard is given by $(X : 25, A : 2, (B : 3.5, V : 6.5) : 11.5)$.

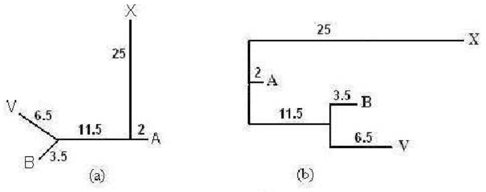


Figure 9. (a) Free and non-rooted phylogenetic tree. (b) Corresponding phylogram.

IV. Functionalities of the Anacê

Anacê is a set of functions written in the Scala [2] programming language available as a set of REST [16] services. Using the RESTful [4] pattern, any program in any programming language can access this services. The only resource needed is a library to use HTTP protocol, wich is available in the most of the programming languages.

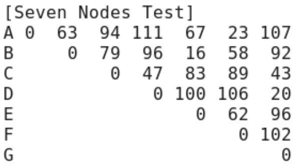
The interactive use of this service via Web is by selecting options in forms. The input file is a symmetric matrix of distances in text format that must be pasted in a specific area of work with the structure indicated in Figure 10. Note that de distance's matrix starts with the heading between bracktes, following the heading are the lines of de superior-triangular matrix of distances, stating from zero, that corresponds to the element in the main diagonal in each line.

An alternative input form for Anacê is the Phylip standard format that represents a tree via a text. Figure 12 shows the phylogenetic tree generated by Anacê for the following data that use the Phylip standard format: ((((Y arrowia : 0.57, (Orthopsilosis : 0.03, Candida : 0.01) : 0.43) : 0.20, M archantia : 0.52) : 0.35, Caenorhabditis : 1.71) : 0.23, (Drosophila : 0.30, M elipona : 0.61) : 0.54, ((Rattus : 0.13, (P an : 0.03, Homo : 0.03) : 0.10) : 0.15, (Cobitis : 0.46, Oreochromis : 0.27) : 0.33) : 0.30).

A. *Using* curl *command to access Anacê*

The command *curl* [17] is a tool to transfer data from or to a server, using one of the supported protocols (DICT, FILE, FTP, FTPS, GOPHER, HTTP, HTTPS, IMAP, IMAPS, LDAP, LDAPS, POP3, POP3S, RTMP, RTSP, SCP, SFTP, SMTP, SMTPS, TELNET and TFTP). The command is designed to work without user interaction.

In the next examples the word *prefix* must be replaced by *http://anace.uece.br:9080/anace/rs/*.



```
[Seven Nodes Test]
A 0  63  94 111  67  23 107
B     0  79  96  16  58  92
C         0  47  83  89  43
D             0 100 106  20
E                 0  62  96
F                     0 102
G                         0
```

Figure 10. Input example in the basic format.

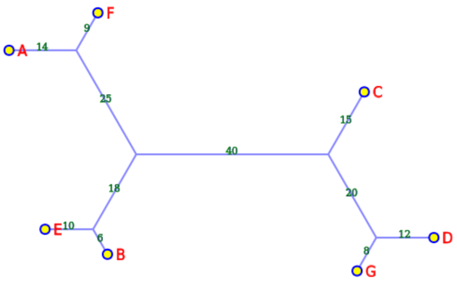


Figure 11. Phylogenetic tree for the input in Figure 10.

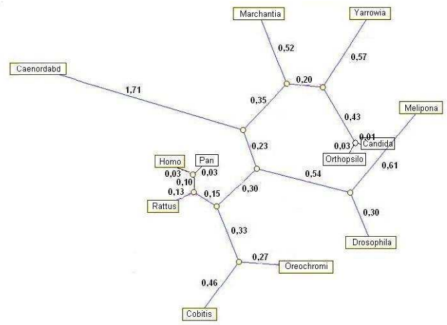


Figure 12. Phylogenetic tree for 12 species generated by Anacê.

The example that follow informs a distance's matrix and get a phylogenetic tree in Phylip's format.

```
curl -X POST prefix/toTree \
-d matrix="[7 Nodes] A 0 63 94 111 67 \
23 107 B 0 79 96 16 58 92 C 0 47 83 89 \
43 D 0 100 106 20 E 0 62 96 \
F 0 102 G 0"
```

Note that all the commands might be written in a single line, here we use de \ character to continue the command in the next line only for visualization purpose. The result is:

```
[&&HM:title=7 Nodes]((F:9.00000,A:14.0000)
:25.0000,(E:10.0000,B:6.00000):18.0000,((
G:8.00000,D:12.0000):20.0000,C:15.0000)
:40.0000)
```

The example that follow informs a phylogenetic tree in Phylip's format and get a distance's matrix.

```
curl -X POST prefix/toMatrix \
 -d tree="((F:9,A:14):25,(E:10,B:6):18, \
          ((G:8,D:12):20,C:15):40)"
```

The result is:

```
[no title]
E 0   62   67   96   16   83  100
F 0   23  102   58   89  106
A 0  107   63   94  111
G 0   92   43   20
B 0   79   96
C 0   47
D 0
```

Note that this distance's matrix can be rewroted like:

```
[no title]
E 0 62   67   96   16  83  100
F    0   23  102   58  89  106
A         0  107   63  94  111
G              0   92  43   20
B                   0  79   96
C                       0   47
D                            0
```

The example that follow informs a phylogenetic tree in Phylip's format and get a SVG image for the tree like a cladogram.

```
curl -X POST prefix/toCladogram \
  /400/400/1.0/0.2 \
  -d tree="((F:9,A:14):25,(E:10, \
     B:6):18,((G:8,D:12):20, \
     C:15):40)"
```

The result is:

```
<svg:svg width="400" height="400" ...
 ... here comes all the SVG's commands
     to trace the phylogenetic tree ...
</svg:svg>
```

The Anacê's site, with its tutorial, may be visited at:

```
http://anace.uece.br:9080/anace/home
```

V. Conclusion

In this paper, we have described a tool, called Anacê, that makes it simple the task of drawing phylogenetic trees. This tool is especially useful to researchers in computational biology that work with phylogeny. The Anacê is also useful for people working in graph theory, since this tool makes more enjoyable the process of checking visually if a given graph satisfies a specific property.

References

[1] J. P. Felsenstein, "Phylogeny inference package computer programs for inferring phylogenies," URL:http://evolution.genetics.washington.edu/phylip.html, Seattle - WA, EUA, 1993, last time accessed: January 2011.

[2] Odersky, Martin, Spoon, Lex, and Venners, Bill, *Programming in Scala*. EUA: Artima, 2008.

[3] G. V. R. Viana and H. A. S. Moura, "Anacê," URL:http://anace.uece.br/anace/home, Fortaleza, CE, Brazil, 2011, last time accessed: January 2011.

[4] Richardson, Leonard and Ruby, Sam, *RESTful Web Services*. USA: OReilly, 2007.

[5] R. D. M. Page, "Treeview for win32," URL:http://taxonomy.zoology.gla.ac.uk/rod/rod.html, last time accessed: January 2011.

[6] Choi, J., Jung, H., KIM, H., and Cho, H., "Phylodraw: A phylogenetic tree drawing system," *Bioinformatics*, vol. 16, pp. 1056–1058, 2000.

[7] M. A. Charleston, "Spectrum: Spectral analysis of phylogenetic data," *Bioinformatics*, vol. 11, p. 9899, 1998.

[8] O. Gascuel, "A web server for fast maximum likelihood-based phylogenetic inference," 2004.

[9] Swofiord, D. L. and Olsen, G. L., "Phylogeny reconstruction molecular systematics," Massachusetts - EUA, 1990.

[10] D. S. Amorim, *Elementos Básicos da Sistemática Filogenética*. Holos, 1997.

[11] J. C. Setubal and J. Meidanis, *Introduction to Computational Molecular Biology*. Brooks/Cole Publishing Co., 1997.

[12] N. Saitou and N. Nei, "The neighbor-joining method: A new method for reconstructiong phylogenetic trees," *Molecular Biology and Evolution*, vol. 4, p. 4.

[13] S. C. Stearns, *Evolution: an Introduction*. Oxford University Press, 2000.

[14] Purves, W.K., Sadawa, D., Orians, G. H., and Heller, C., *Life: The Science of Biology*. W.H. Freeman Co., 2003.

[15] Kumar, S., Tamura, K., and Nei, M., "Mega3: Integrated software for molecular evolutionary. genetic analysis and sequence alignment." *Briefing in Bioinformatics*, vol. 5(2), pp. 150–163, 2004.

[16] T. Fielding, "Architectural styles and the design of network-based software architectures," Doctor of Philosophy, University of California, Irvine - CA, USA, 2000.

[17] "curl," URL:http://curl.haxx.se, last time accessed: January 2011.