

Software

PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change

Mathieu Fourment and Mark J Gibbs*

Address: School of Botany and Zoology, The Australian National University, Canberra ACT 0200, Australia

Email: Mathieu Fourment - m.fourment@gmail.com; Mark J Gibbs* - gibbslab@hotmail.com

* Corresponding author

Published: 03 January 2006

Received: 18 August 2005

BMC Evolutionary Biology 2006, **6**:1 doi:10.1186/1471-2148-6-1

Accepted: 03 January 2006

This article is available from: <http://www.biomedcentral.com/1471-2148/6/1>

© 2006 Fourment and Gibbs; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Phylogenies are commonly used to analyse the differences between genes, genomes and species. Patristic distances calculated from tree branch lengths describe the amount of genetic change represented by a tree and are commonly compared with other measures of mutation to investigate the substitutional processes or the goodness of fit of a tree to the raw data. Up until now no universal tool has been available for calculating patristic distances and correlating them with other genetic distance measures.

Results: PATRISTICv1.0 is a java program that calculates patristic distances from large trees in a range of file formats and allows graphical and statistical interpretation of distance matrices calculated by other programs.

Conclusion: The software overcomes some logistic barriers to analysing signals in sequences. In addition to calculating patristic distances, it provides plots for any combination of matrices, calculates commonly used statistics, allows data such as isolation dates to be entered and reorders matrices with matching species or gene labels. It will be used to analyse rates of mutation and substitutional saturation and the evolution of viruses. It is available at <http://biojanus.anu.edu.au/programs/> and requires the Java runtime environment.

Background

Phylogenetic trees are standardly used to analyse gene families as well as species ancestry [1]. A patristic distance is the sum of the lengths of the branches that link two nodes in a tree, where those nodes are typically terminal nodes that represent extant gene sequences or species. A matrix of patristic distances calculated from a tree for all pairs of genes or species summarizes the genetic change, or phylogenetic change, represented in the tree and the data. The distances may be used to analyse the rate of change and may be compared with other measures of genetic difference, such as the total change (evolutionary distance) or components of change such as those indi-

cated by transitions, transversions, gene rearrangements or recombination [2-4]. In one significant example, the origin of the HIV pandemic was investigated by comparing the patristic distances of dated HIV gene sequences [5]. There are many methods and programs for finding, building, or testing trees but only one is known to us that calculates patristic distances and it does so only from maximum parsimony trees, which by definition only represent a fraction of genetic change [6].

Implementation

PATRISTICv1.0 is a Java program that can be used as an applet on our website or downloaded. It calculates patris-

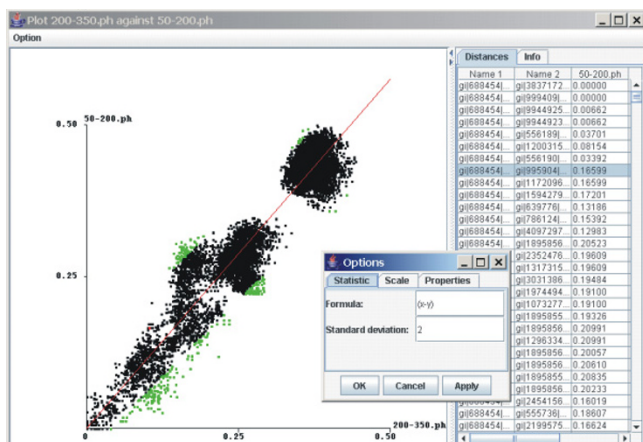


Figure 1
A screenshot of PATRISTICv1.0 in operation with a plot of patristic distances calculated from two different regions of bunyavirus segment S sequences. The green points represent paired distances that are more than two standard deviations from the mean difference. A regression is shown in red, the distances matrices from which the plot was calculated are shown as columns on the right and an window for entering formulae and altering the scale is shown in the foreground.

tic distances from trees, generates scatter plots from ordered pairs of distances and calculates correlation coefficients and other statistics from distance matrices. It reads trees in variants of the Newick format, including the NEXUS variant used by the package PAUP [6] and the variants used by the programs MEGA [6], PHYLIP [8], CLUSTALX [9] and TREEPUZZLE [10]. An algorithm that traverses the various textual representations of trees [11] was used to calculate the patristic distances, along with code that permitted different tree-text formats to be read, permitted the easy selection of matrices for plotting from a large number of stored matrices and permitted matrices and plots to be displayed. PATRISTICv1.0 runs on Windows, Mac and Linux systems with the Java Runtime Environment. A patristic distance matrix from a tree of 187 gene sequences was calculated in 12 seconds in a PC with an AMD CPU at 2.2 GHz and 256 RAM using the JRE 1.5. PATRISTICv1.0 was tested by calculating patristic distances by hand across several small trees and in every case the results of the program were found to be accurate.

The program also recognises distance matrices calculated by other programs from other components of sequence data, such as evolutionary distances calculated from pair-wise sequence comparisons. It reads distance matrices generated by the programs MEGA, PAUP and PHYLIP. For the current version, these externally generated matrices must be presented as upper-right or lower-left hemimatrices or a column. Other measures of that can be converted into distances, such as the isolation dates of virus

samples, may be entered for each species or gene as real numbers. If sample times are entered directly PATRISTICv1.0 will generate a matrix of time differences between the species or sequences.

The order of sequences or species represented in a tree almost always differs from the order in the original data file from which the tree was found. Hence to plot patristic distances against distances calculated by other methods or patristic distances from two trees, the program automatically reorders matrices with matching sequence or species labels. Matrices may also be edited and reordered within an editing window.

A regression is calculated from the ordered pairs of distances when two matrices are plotted against each other (Figure 1) and simple statistics such as the sums of the distances are displayed. Correlation coefficients, differences and quotients between the ordered pairs of distances may also be calculated using PATRISTICv1.0, as may the mean and standard deviation of the differences or quotients, and the program has a facility to enter other formulae so that other statistics can be calculated.

Points on a plot that lie outside a chosen multiple of the standard deviation are identified by colour. Points are also automatically identified on a plot when the mouse cursor is moved over them. A zoom feature allows the user to focus on a specific part of a plot by choosing the minimum and maximum distances for the two axes which correspond to the two distance matrices. Plots of distance matrices may be inverted relative to the axes by a single mouse click. The user can also determine the scale used on the axes as well as the dot size.

Plots may be saved as postscript or jpeg files allowing editing in graphics software in a vectorial or bitmap format. The program also allows distance matrices to be saved in a coma separated value format (CSV) as a full matrix or as columns so that they may be entered into a spreadsheet program. Matrices can also be saved in the DIP format used by the software DIPLOMO [12].

Results and Discussion

Maximum likelihood (ML) methods provide the best estimates of evolutionary change and genetic difference by modelling substitutions. Models are used to calculate branch-lengths (genetic change) that take into account the superimposition of substitutions and the similarities between one sequence or taxon and all others represented in the dataset [13]. Hence, some patristic distances from ML trees are greater than the equivalent estimates of genetic distance, when the genetic distances are based on pair-wise comparisons between the raw sequences. The converse also occurs if patristic distances are calculated

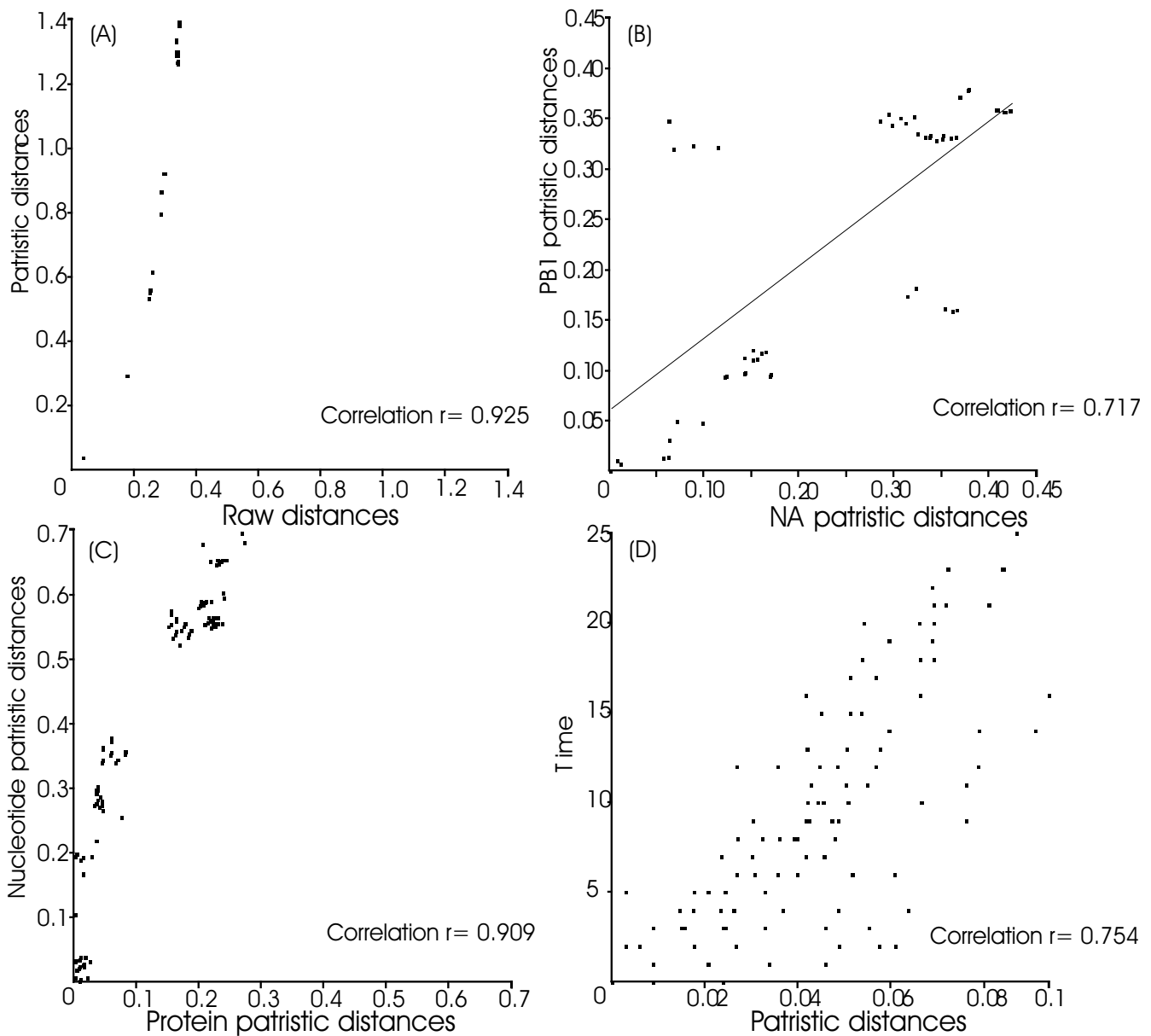


Figure 2

Plots generated using PATRISTICv1.0 of: (A) patristic distances from an ML tree and matching uncorrected evolutionary distances calculated from bunyavirus RNA polymerase gene sequences, (B) patristic distances calculated from the neuraminidase and RNA cap-binding protein genes of a set of influenza A virus N1 subtype isolates with a regression line drawn for all points, (C) patristic distances calculated from the nucleocapsid protein amino acid sequences and the equivalent nucleotide sequences of bunyaviruses, (D) patristic distances from an ML tree of the neuraminidase genes of influenza virus isolates and the differences in years between times of isolation of the isolates.

from a maximum parsimony tree, since characters that do not conform to the cladistic definition of phylogenetic information (autapomorphies and homoplasies) are discounted. To demonstrate one of those effects we used PATRISTICv1.0 to plot patristic distances against uncor-

rected evolutionary distances, where the patristic distances were calculated from an ML tree of bunyavirus RNA polymerase gene sequences and the evolutionary distances were obtained from pair-wise comparisons of the same sequences without reference to a tree (Figure 2A). It

can be seen that whereas the evolutionary (raw) distances increased linearly the patristic distances were greater and increased exponentially.

Another use of patristic distances is to investigate recombinational processes. A low correlation between distances obtained from different genes may indicate a recombinational process such as horizontal gene transfer [14]. Influenza A viruses undergo a recombinational process, known as reassortment, that produces incongruities much like those produced by horizontal gene transfer, but on a smaller scale [15], and the incongruities may be identified as poorly correlated patristic distances. We detected incongruities probably due to reassortment by plotting patristic distances calculated from the neuraminidase (NA) and RNA cap-binding protein (PB2) genes of a set of influenza A virus N1 subtype isolates (Figure 2B); we found that the majority of the points lie close to the regression line and appear to increase linearly, whereas some points fall well off the line. Even when there has been no recombination, different measures of genetic difference from the same set of organisms may not correlate linearly. The rates of change in nucleotide sequences and the amino acid sequences they encode do not vary linearly because of the redundancy in the genetic code and because some sites reach a saturation point of superimposed substitutions. We demonstrated the effect by plotting patristic distances calculated from the nucleocapsid protein amino acid sequences and the equivalent nucleotide sequences of bunyaviruses (Figure 2C); the patristic distances from the nucleotide sequence tree increased at a greater rate than the equivalent patristic distances from the amino acid tree.

Patristic distances have broader uses as it is sometimes of value to compare them with other data, such as data related to time or geographic distribution. For example, comparisons of the times of isolation and the distances between the sequences of influenza A virus strains describe an important process whereby the older strains are being eliminated from global circulation and being replaced by new strains. Figure 2D shows that the patristic distances from an ML tree of the neuraminidase genes of a set of influenza virus isolates correlated with the differences in years between the times of isolation of the isolates.

Conclusion

A phylogenetic tree usually represents signals drawn from many sites in the sequences from which it was inferred, and it represents a hierarchy of those signals that may be nested. Hence, the relationships between phylogenetic trees and sequences of interest are often complex and obscure. Comparisons of patristic distances and other measures of genetic change allow the relationships to be

analysed. PATRISTICv1.0 was developed to permit expedient comparisons and analyses. The program reads trees and distance matrices produced by the most commonly used software without editing. Re-formatting is a time consuming element of multiple sequence and phylogenetic analysis and for that reason PATRISTICv1.0 creates no new intermediate formats. The need to re-order distance data from gene sequences or species was a major barrier that previously hindered analyses and that problem has been solved. The program is likely to be used to analyse recombination, rates of mutation and substitutional saturation and the evolution of rapidly evolving entities such as viruses.

Availability and requirements

Project name: Patristic

Project home page: <http://biojanus.anu.edu.au/projects/>

Operating system(s): Platform independent

Programming language: Java

Other requirements: Java 1.41 or higher <http://java.sun.com>

Any restrictions to use by non-academics: None

References

1. Grus WE, Shi P, Zhang YP, Zhang J: **Dramatic variation of the vomeronasal pheromone receptor gene repertoire among five orders of placental and marsupial mammals.** *Proc Natl Acad Sci USA* 2005, **102**:5767-72.
2. Hayward DC, Bastiani MJ, Trueman JWH, Truman JW, Riddiford LM, Ball EE: **The sequence of *Locusta* RXR homologous to *Drosophila* Ultraspiracle, and its evolutionary implications.** *Dev Genes Evol* 1999, **209**:564-71.
3. Broughton RE, Stanley SCE, Durrett RT: **Quantification of Homoplasy for Nucleotide Transitions and Transversions and a reexamination of Assumptions in Weighted Phylogenetic Analysis.** *Syst Biol* 2000, **49**:617-627.
4. Herniou EA, Luque T, Chen V, Vlcek JM, Winstanley D, Cory JS, O'Reilly DR: **Use of Whole Genome Sequence Data To Infer Baculovirus Phylogeny.** *J Virol* 2001, **75**:8117-8126.
5. Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Bhattacharya T: **Timing the Ancestor of the HIV-1 Pandemic Strains.** *Science* 2000, **288**:1789-1796.
6. Swofford DL: **PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4.** Sinauer Associates, Sunderland, Massachusetts; 2003.
7. Felsenstein J: **PHYLIP – Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5**:164-166.
8. Kumar S, Tamura K, Nei M: **MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment.** *Briefings in Bioinformatics* 2004, **5**:150-163.
9. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Research* 1997, **25**:4876-4882.
10. Schmidt HA, Strimmer K, Vingron M, Von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18**:502-504.
11. Newick [\[http://evolution.genetics.washington.edu/phylip/newick_doc.html\]](http://evolution.genetics.washington.edu/phylip/newick_doc.html)

12. Weiller GF, Gibbs A: **DIPLOMO: the tool for a new type of evolutionary analysis.** *Comput Appl Biosci* 1995, **5**:535-40.
13. Hillis DM, Moritz C, Mable BK: *Molecular Systematics* 2nd edition. Sinauer Associates; 1996.
14. Farahi K, Pusch GD, Overbeek R, Whitman WB: **Detection of lateral gene transfer events in the prokaryotic tRNA synthetases by the ratios of evolutionary distances method.** *J Mol Evol* 2004, **58**:615-31.
15. Holmes EC, Ghedin E, Miller N, Taylor J, Bao Y, St George K, Grenfell BT, Salzberg SL, Fraser CM, Lipman DJ, Taubenberger JK: **Whole-Genome Analysis of Human Influenza A Virus Reveals Multiple Persistent Lineages and Reassortment among Recent H3N2 Viruses.** *PLoS Biology* 2005, **3**(9):e300.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

