WILEY-BLACKWELL

Cladistics (2014) 1–8

Cladistics

10.1111/cla.12083

POY version 5: phylogenetic analysis using dynamic homologies under multiple optimality criteria

Ward C. Wheeler*, Nicholas Lucaroni, Lin Hong, Louise M. Crowley and Andrés Varón

Division of Invertebrate Zoology, American Museum of Natural History, Central Park West @ 79th Street, New York, NY, 10024-5192, USA

Accepted 28 April 2014

Abstract

We present POY version 5, an open source program for the phylogenetic analysis of diverse data types including qualitative, aligned sequences, unaligned sequences, genomic data, and user-defined sequences. In addition to the maximum-parsimony optimality criterion supported by POY4, POY5 supports several types of maximum likelihood as well as posterior probability. To make these analyses feasible, new heuristic search algorithms and parallelization options have been implemented for all criteria.

Background

POY is an open source phylogenetic analysis program for use on a diversity of data types including morphology, aligned and unaligned sequence data, and genomic sequences. Previous versions were released in 2004 (POY 3.0.11, Wheeler et al., 1996–2005),[1] 2009 (POY 4.1.3, Varón et al., 2008, 2010), and now version 5 (POY 5.1.1, source code Wheeler et al., 2013). All features of POY4 have been maintained and many improved. The presentation here concentrates on areas of new functionality and efficiency. The main areas of enhancement in POY5 are the addition of likelihood and Bayesian optimality criteria and the implementation of within-trajectory parallelization.

*Corresponding author:
E-mail address: wheeler@amnh.org

[1]Although there were earlier versions back to 1997 (Gladstein and Wheeler, 1997), POY4 was completely new code.

New/improved character types

Continuous characters

'Continuous' characters are, as is customary, optimized as additive discrete characters (Farris, 1972) with a large number of states. POY5 improves the implementation over POY4. Following the format of TNT (Goloboff et al., 2003), this character type is specified by an additional, prepended specification in the Hennig86/NONA input file (Fig. 1). Character states are specified as integer values that are then decimalized by character weighting. Hence, a character with states from 0.0 to 1.000 are coded as 0 to 1000 and set to a weight of 0.001. Characters are entered as ranges (`[low high]` or `[value]` if a single value—this diverges from TNT). The maximum character value is $2^{62}-1$.

Prealigned for custom and amino acid sequences

In previous versions of POY, only nucleotide sequence data could be treated as prealigned data (unless recoded as qualitative data), respecting the placement of gap ("-") states. Without this specification, input gaps are stripped out and data are treated as unaligned. POY5 allows the input of amino acid and

```
nstates cont;
xread
'My continuous character matrix'
2 5
Alpha       [0 100] [2000 2100]
Beta        [0] [700 800]
Gamma       [3000 3100][3100 3200]
Delta       [900 1000] [3350 3450]
Epsilon     [1220 1320] [1900 2000]
;
cc + 0.1;
proc /;
;
```

Fig. 1. An example file specifying continuous characters. All characters are scored as ranges "[*x y*]" and must be in their own (no other data type) input file. Reweighting is required to decimalize the states.

custom (user-specified alphabets such as gene synteny or developmental) sequence data (e.g. `read (prealigned:(amino acids:("*.fas"), tcm:"matrix1"))`. In addition to allowing the user additional flexibility in analytical assumptions, this new functionality also permits the use of static approximation heuristics for searching on these characters either manually (`transform(static_approx)`) or automatically via the `search()` command.

Levels for large alphabet sizes (amino acid and custom)

The direct optimization algorithm (DO; Wheeler, 1996; Varón and Wheeler, 2012) in both its $O(n^2)$ and $O(n^3)$ forms, employs a precalculated cost and state matrix to reduce sequence optimization time. This matrix has dimensions defined by the number of potential element combinations at each sequence position ($\left[\Sigma_{i=1}^{i=a}\binom{a}{i}-1\right]^2$ where a=|*elements*|). This is quite manageable for sequences with small alphabets (i.e. DNA with 225 values) but rapidly becomes unworkable for alphabets presented by protein sequences (4,398,042,316,801 values). Larger alphabets presented by linguistic (Wheeler and Whiteley, 2014), behavioural, and developmental sequences (Schulmeister and Wheeler, 2004) present an even greater challenge. The strength and speed of the DO heuristic depends on this matrix.

To make larger alphabets feasible in unaligned sequences, we implement a "level" parameter. The level specifies the maximum number of state combinations for which medians and costs are stored in the precalculated matrix. For DNA sequences, standard DO has level 5. Smaller levels reduce the number of combinations and storage that are required, but also weakens the heuristic (Fig. 2).

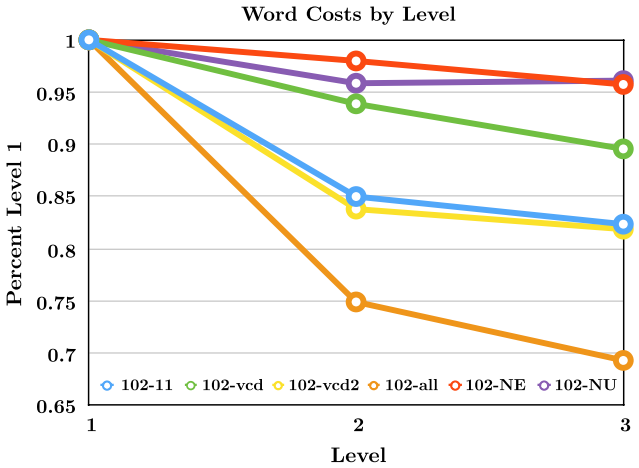

Fig. 2. Effect of increased heuristic level on optimality score for linguistic data under a variety of models (102-11 etc.). Data from Wheeler and Whiteley (2014).

Chromosome and genome

Although POY versions 3 and 4 have facilities for optimizing chromosomal data (sequences with potential rearrangement), these required prior annotation of the data (Vinh et al., 2006, 2007; Wheeler, 2007). POY5 has implemented the MAUVE pairwise genomic aligner algorithm (Darling et al., 2004), and in concert with Fixed States optimization (Wheeler, 1999), removes this limitation, allowing the direct analysis of unannotated chromosomal data.

This operates via the `transform()` command after reading in chromosomal data (Fig. 3). The transform to Fixed States optimization also allows for a stem string to be specified ("my_chrom_aligns" in Fig. 3) that generates a series of files (one for each pair of input chromosomes) that can be processed by MAUVE to create graphical images of the sequence similarity and rearrangements occurring between the two chromosomes. This is shown for a set of complete mitochondrial genomes of Heteroptera in Fig. 4 where the transformations between each optimized node on the tree are displayed. At present, only Fixed States optimization is possible with the MAUVE aligner.

New optimality criteria

In addition to the parsimony criterion of POY4, POY5 implements two forms of maximum likelihood (ML; Wheeler, 2006) and a form of Bayesian posterior probability (Wheeler, 2014). In each case, facilities exist for analysis of morphological[2] and sequence data in both static and dynamic homology frameworks. In all

[2]Characters with large numbers of states (e.g. 'continuous' characters) can be optimized under these criteria, but the user would have to create appropriate models and weight regimes.

```
read(chromosome:("my_chromosome.fasta"))
transform(tcm:(1,1), gap_opening:1)
transform(chromosome:(locus_inversion:100,locus_indel:(10,0.9)))
transform(chromosome:(annotate:(mauve,25.0,0.30,0.005,0.25)))
transform(fixed_states:("my_chrom_aligns", ignore_polymorphism))
```

Fig. 3. An example script showing the reading of an unnanotated chromosome sequence in FASTA format, setting costs for within-locus substitution and affine indels, setting values for the cost of various chromosomal events (e.g. insertion-deletion of a locus, rearrangement cost), and pairwise annotation and alignment via the MAUVE algorithm.

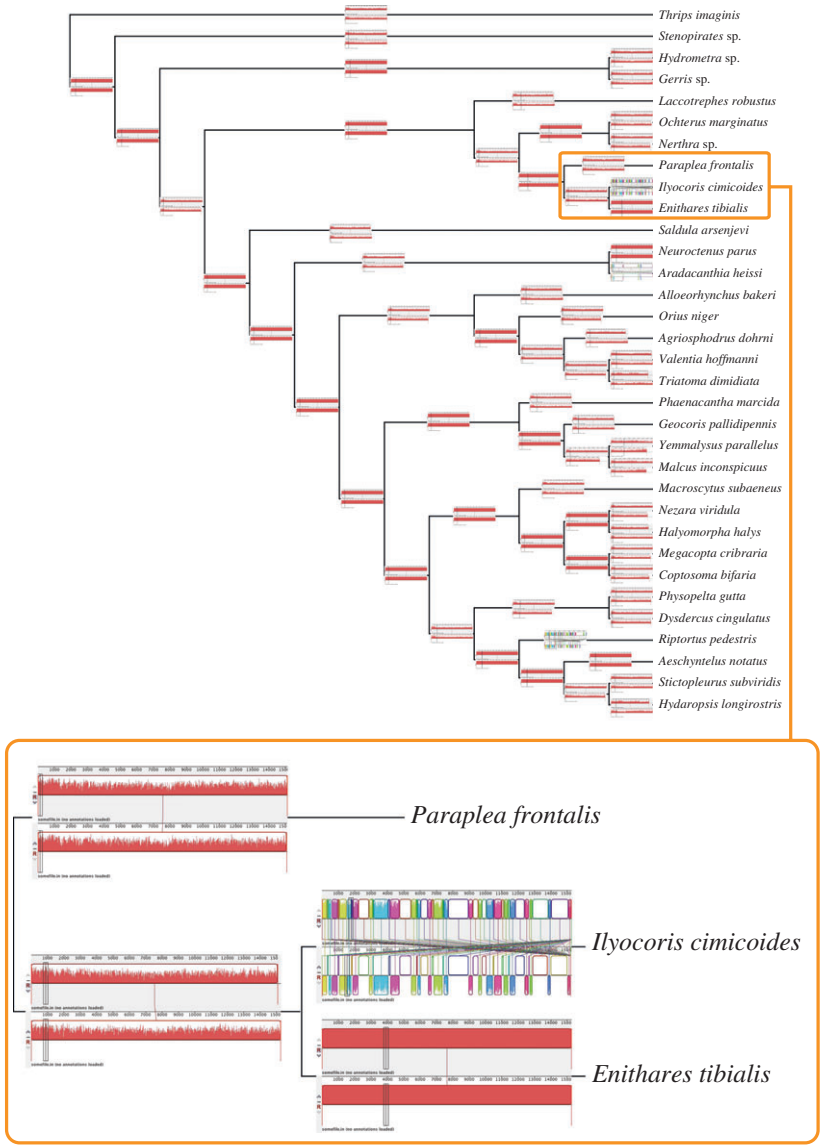

Fig. 4. MAUVE (Darling et al., 2004) annotations of edge transformations between vertex mitochondrial genomes of Heteroptera (and *Thrips* outgroup).

cases, a variety of models are available from the completely symmetrical (Jukes and Cantor, 1969; Neyman, 1971) to parameter-rich (Tavaré, 1986) with a variety of indel model options.

The optimality criterion is specified using the `transform()` command. The criterion can be switched back and forth between parsimony and likelihood at any time, allowing comparisons between

methodologies and optimality criteria (Giribet and Edgecombe, 2013).

Maximum average likelihood

Following the ML taxonomy of Steel and Penny (2000) (reviewed by Wheeler, 2012), maximum average likelihood (MAL) is the form of maximum relative likelihood where tree likelihoods are integrated (or averaged) over all possible internal state assignments. This is the most popular form of likelihood and is implemented for prealigned sequences in packages such as RAxML (Stamatakis et al., 2005). POY5 implements MAL for prealigned sequences and allows gap characters to be treated as missing data (`gap:missing`), with a single parameter (`gap:coupled`), or multiple with individual parameters for indels to and from A, C, G, T or other sequence alphabet elements (`gap:character`). Comparisons among MAL implementations of aligned sequences with gaps treated as missing are reported in Denton and Wheeler (2012).

When sequence data are unaligned (dynamic data), MAL in POY requires indel parameterization (i.e. not missing) and is restricted to the Fixed States heuristic with summing over input sequences as potential internal states (DO is implemented via most parsimonious likelihood, MPL, below).

Qualitative data (as well as prealigned sequence data) can be analysed by common (CM; Neyman, 1971) and no-common mechanism (NCM; Tuffley and Steel, 1997) models. In both cases, the Neyman (1971) transition model is employed with either a single estimated rate for all characters (CM) or potentially unique rates for each character (NCM). The alphabet size for qualitative data is not restricted. Because alternative models can be specified for different data partitions (e.g. morphological and molecular data), combined or total evidence (Kluge, 1989) analyses can be performed under ML.

Most parsimonious likelihood

MPL (Barry and Hartigan, 1987) is a form of ML in which ancestral states are uniquely identified as those with highest likelihood. Single state assignments are made and alternative assignment likelihoods are not summed at internal nodes. The "dominant" alignment in likelihood alignment (Thorne et al., 1991) is an expression of this idea. POY5 implements a likelihood DO procedure (Wheeler, 2006) to create median (ancestral) sequences with heuristically maximal MPL.

Like MAL above, this form of likelihood can be applied to prealigned and unaligned sequences, as well as qualitative data. Unlike MAL, unaligned sequence optimization is not restricted to Fixed States.

Under both MAL and MPL, when unaligned sequence data are analysed, rate class parameters are ignored because classes of characters have no meaning in a dynamic homology context.

Model specification, estimation, and choice

As with other implementations, POY5 allows for the specification of a broad variety of models from the effectively zero parameter JC69/Neyman (Jukes and Cantor, 1969; Neyman, 1971), through F81 (Felsenstein, 1981), K2P/K80 (Kimura, 1980), F84 (Felsenstein, 1984), HKY85 (Hasegawa et al., 1985), TN93 (Tamura and Nei, 1993), to GTR (Tavaré, 1986). Single and multiple rate distributions are also permitted (for static and prealigned data) with the discrete Γ and invariant sites models. The NCM (Tuffley and Steel, 1997) model is also available. Neyman, NCM, and GTR models allow any alphabet size (i.e. not limited to DNA nucleotides), and hence can be employed for qualitative data and larger alphabet sequences such as proteins. When multiple data partitions are present, they may be assigned different models.

The estimation of model parameters can be accomplished in a variety of ways and with user-specified intensity and frequency from `optimize(model:never, branch:never)` (i.e. only during the initial transformation to likelihood, set branch lengths to the proportion of changes on edge) to `optimize(model:always, branch:all_branches)` (i.e. optimize model every time a new tree is encountered, optimize all branch lengths every time a new tree is encountered). Initial parameter values may be estimated to optimize the likelihood of an input tree, or based on the parsimony character optimization of an input tree. The model and branch optimization intensity are set within search options (e.g. `swap(spr,all,optimize:(model:never,branch:join_region))`) and can vary throughout the search, for instance, to begin with relatively coarse options and progressively refine results with increasingly demanding options.

Model parameters may also be set directly by the user and not optimized further. For example, reading the model (`model:`"file_name"):

0.0	2.0	1.0	2.0	4.0
2.0	0.0	0.0	1.0	4.0
1.0	2.0	0.0	2.0	4.0
2.0	1.0	2.0	0.0	4.0
4.0	4.0	4.0	4.0	0.0

will be normalized so that the mean rate is 1, and the rows sum to 0.0. This becomes (with equal equilibrium frequencies):

−0.865	0.192	0.096	0.192	0.384
0.192	−0.865	0.192	0.096	0.384
0.096	0.192	−0.865	0.192	0.384
0.192	0.096	0.192	−0.865	0.384
0.384	0.384	0.384	0.384	−1.538

Q-matrix class constraints can also be specified (and later optimized) with the (`custom:"file_name"`) command. Diagonal elements are ignored, but necessary, and any character can be used as a placeholder (below we use a dash, "–"). Any ASCII character can be used to define an associated rate, but the matrix must be symmetric. For example, the following matrix will create a model in which three parameters are ultimately optimized (given that the base parameter is not optimized).

-	*a*	*c*	*d*	*e*
a	-	*c*	*d*	*e*
c	*c*	-	*d*	*e*
d	*d*	*d*	-	*e*
e	*e*	*e*	*e*	-

In addition to specifying the model explicitly, POY5 allows the user to specify one of a number of information-theoretic criteria to select a model based on an input tree. The Akaike Information Criterion (AIC; Akaike, 1974), corrected AIC (AICc; Sugiura, 1978), and Bayesian Information Criterion (BIC; Schwarz, 1978) are available. POY5 will optimize the model to the tree of all available models, and select the best (based on the information criterion selected) to be kept in memory. A report is also printed to show the scores and analysis of model selection (see Table 1).

Bayesian maximum a posteriori assignment

A form of Bayesian posterior probability analysis, maximum a posteriori assignment (MAP-A; Wheeler, 2014), is implemented in POY5 via a non-standard cost matrix (`transform(tcm:"matrix_name")`) and dynamic programming under the parsimony criterion. In short, a weight function is created either analytically or numerically (via the accessory program MAPA.ml, distributed with POY5 source) that allows the identification of that tree which maximizes a posteriori vertex state assignments. This weight matrix is derived from the character change model (e.g. GTR) and branch-length distributions (e.g. exponential). As opposed to the MAP approach (Rannala and Yang, 1996), which sums all possible state assignment contributions to the posterior probability, MAP-A uses only those with maximum probability. In this way, MAP-A is to MAP what MPL is to MAL. The benefit of this approach lies largely in its computational efficiency. POY5 does not offer MC^3 search. The MAPA.ml program allows the specification of a variety of models and prior distributions on model and branch-length parameters.

MAP-A weights may be employed for either static or dynamic characters of any alphabet size and, as with ML models, may vary among partitions and even individual characters. As with the parsimony and ML implementations, MAP-A allows for multiple partition, combined, or "total evidence" analysis.

Performance enhancements

In addition to new features, POY5 has several performance enhancements over POY4. For standard,

Table 1

Model	$-\ell$	K	n	AICc	Δ	ω	Cum(ω)
JC69	538.819	31	49	1256.345	0.000	0.863	0.863
JC69 + G	532.219	32	49	1260.438	4.093	0.111	0.975
K81	533.777	32	49	1263.554	7.209	0.023	0.998
K81 + G	526.763	33	49	1269.127	12.782	0.001	1.000
F81	539.512	34	49	1317.024	60.678	$5.754\ 10^{-14}$	1.000
HKY	532.270	35	49	1328.387	72.041	$1.961\ 10^{-16}$	1.000
F84	533.315	35	49	1330.477	74.131	$6.899\ 10^{-17}$	1.000
F81 + G	533.682	35	49	1331.210	74.865	$4.780\ 10^{-17}$	1.000
HKY+G	526.033	36	49	1346.067	89.721	$2.840\ 10^{-20}$	1.000
F84 + G	526.518	36	49	1347.037	90.691	$1.749\ 10^{-20}$	1.000
TN93	531.899	36	49	1357.799	101.453	$8.052\ 10^{-23}$	1.000
TN93 + G	525.841	37	49	1381.318	124.973	$6.290\ 10^{-28}$	1.000
GTR	530.724	39	49	1486.116	229.770	$1.102\ 10^{-50}$	1.000
GTR+G	523.365	40	49	1536.731	280.386	$1.125\ 10^{-61}$	1.000

Example of POY5 output showing scores and analysis of model selection using aicc. Model type, negative log likelihood values ($-\ell$), penalty parameters, which include the number of branches, and model parameters (K), number of characters (n), AICc values, AICc differences (Δ), Akaike weights (ω), and cumulative Akaike weights (Cum (ω)) are reported. In this example, JC69 garners the best information-theoretic score. "+G" indicates the addition of the Γ parameter.

direct optimization of unaligned nucleic acid sequence data, POY5 is 2–4 times as fast as POY4 (using the data of Arango and Wheeler, 2007). Beyond simple reduction in run times, this speed up significantly enhances the efficacy of timed `search()` trajectories.

Parallelization

POY4 implemented parallelization strategies based on independent trajectories [e.g. individual Random Addition Sequence (RAS) + TBR replicates assigned to their own processes]. Parallelization of multiple random addition sequence Wagner builds, SPR and TBR branch swapping, as well as Tree-Fusing (Goloboff, 1999; Moilanen, 1999) and the `search()` function was accomplished by dividing the tasks and assigning them to processes with little to no communication until these operations were complete. This resulted in near linear speed up when large numbers of RAS replicates or independent `search()` commands were required. The option of parallelizing a single trajectory was unavailable. POY5 meets this need with the swap option "parallel" (e.g. `swap(tbr, all, parallel)`).

Parallel swap tree refinement breaks up the tree search neighbourhood among parallel processes increasing communication load, and hence is likely to be most efficient for large trees and those with time-consuming optimization options (e.g. likelihood, iterative pass optimization). Figure 5 shows the relationship between execution time (a DO TBR swap on a dataset of 208 18S rRNA genes) and number of processes. The linear regression of time on log number of processes is −0.427 in comparison with the perfect −0.693. This suggests that for each doubling in processor number, the execution time decreases by a factor of roughly 1.5. Interestingly, at least for this data set, this factor continued up to 64 processes for a reduction in wall-clock execution time by a factor of nearly 20.

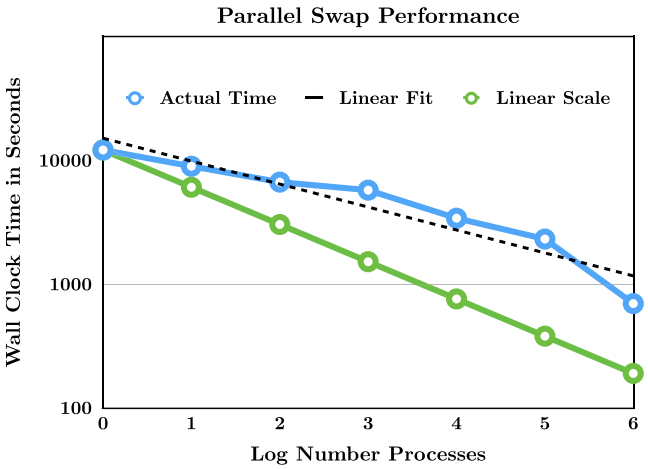

Fig. 5. Linearity of parallel swap. Data set from Giribet and Wheeler (1999, 2001).

Parallelization of large fixed state optimizations (such as unannotated chromosomes using the MAUVE algorithm above) can be achieved using the "–enable–parmap" compiler option and `set (parmap:n)` command (where *n* is the number of parallel processes). Parmap is not an Message Passing Interface-based paralellization and requires alternative source code compilation (Danelluto and DiCosimo, 2012; see POY5 documentation).

Reporting

In addition to the ASCII, Newick, and PDF tree, consensus, and diagnosis outputs of POY4, POY5 has added several new reporting options.

Branch lengths

POY5 will report trees with branch lengths. When the optimality criterion is specified as likelihood, these are the optimized branch lengths and can be read back into POY5 or other likelihood implementations. Under parsimony, the branch lengths are based on optimized character changes on edges, which can be ambiguous. To deal with this, the default option "single" as well as alternatives "min" and "max" are supported.

Likelihood models and other information

The option "lkmodel" (`report(lkmodel)`) reports the likelihood model, costs, and tree length in a style similar to that of PHYLIP.

Graphic diagnosis

Hyperlinked graphical character diagnosis is available via `report("file_name", graphdiagnosis)`. This is especially useful for large data sets.

XML interface with Supramap

To promote an interface with Supramap (Janies et al., 2007; `https://supramap.renci.org/supramap/home`), the report command can output diagnosis information in XML. This is specified by ending the report output file name with ".xml".

Tree distances

Robinson–Foulds (Robinson and Foulds, 1981) tree distances can be output with `report`

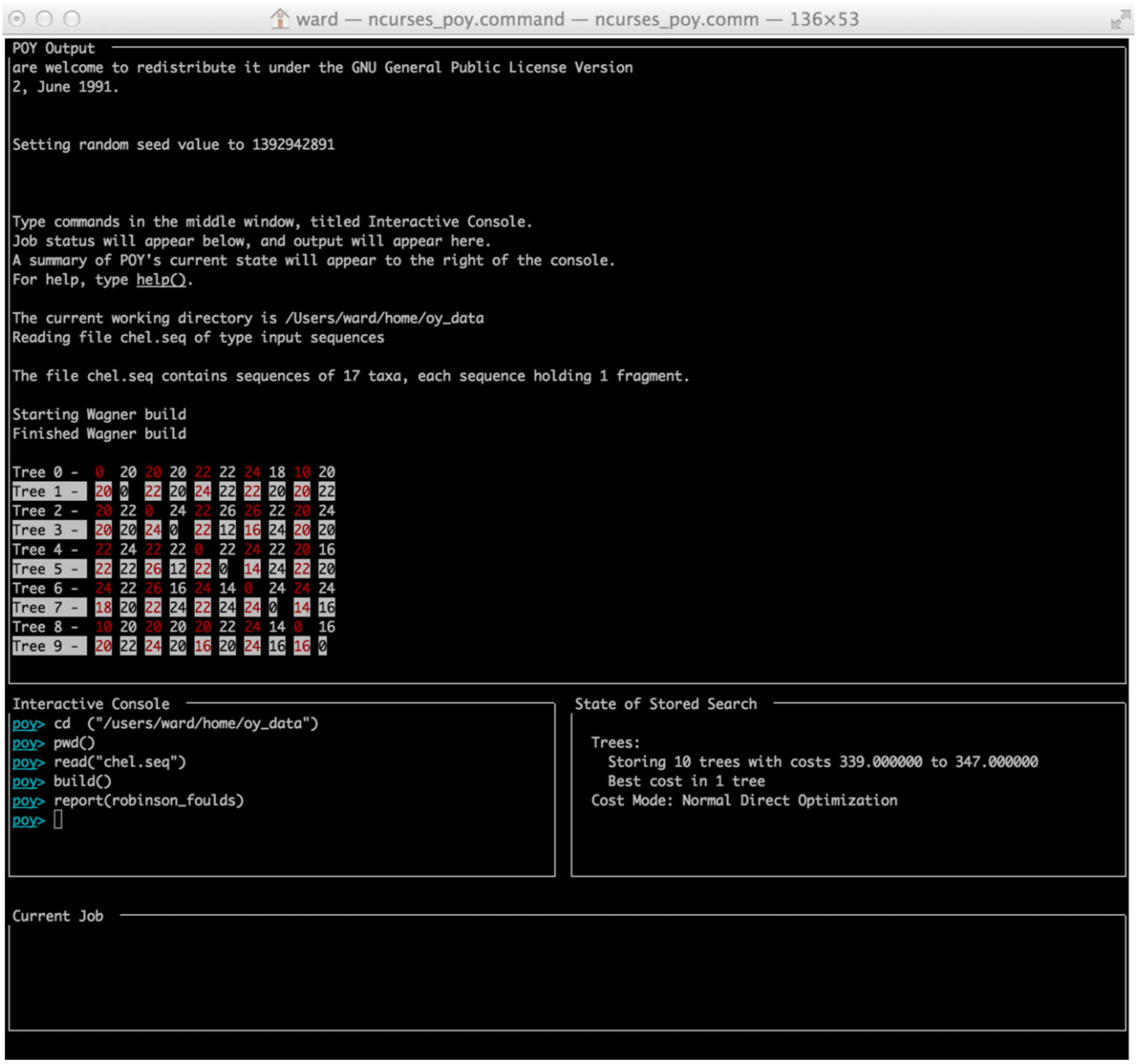

Fig. 6. Robinson–Foulds (Robinson and Foulds, 1981) tree distances for a set of ten trees using interactive console.

`(robinson_foulds)`. As with other report options, this may be redirected to a file (Fig. 6).

Distribution

As before, POY5 is available as binaries as well as source code and documentation at the AMNH (http://www.amnh.org/our-research/computational-sciences/research/projects/systematic-biology/poy/download) and GoogleCode (https://code.google.com/p/poy). Distributions are now also available through GitHub (https://github.com/AMNH/POY).

Acknowledgements

Many people have improved POY5 by suggesting functionality, identifying bugs, reviewing previous versions of the manuscript, and testing pre-release code, especially Megan Cevasco, John Denton, Eric Ford, Cyrille D'Haese, Lauren Heller, Gonzalo Giribet, Taran Grant, Denis Machado, and Fernando Marques. This material is based upon work supported by, or in part by, the US Army Research Laboratory and the US Army Research Office under grant number W911NF- 05-1-0271.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19, 716–723.

Arango, C.A., Wheeler, W.C., 2007. Phylogeny of the sea spiders (Arthropoda, Pycnogonida) based on direct optimization of six loci and morphology. Cladistics 23, 255–293.

Barry, D., Hartigan, J., 1987. Statistical analysis of hominid molecular evolution. Stat. Sci. 2, 191–210.

Danelluto, M.R., DiCosimo, A., 2012. A "minimal disruption" skeleton experiment: seamless map & reduce embedding in ocaml. Procedia Comput. Sci. 9, 1837–1846.

Darling, A.C., Mau, B., Blattner, F.R., Perna, N.T., 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14, 1394–1403.

Denton, J., Wheeler, W.C., 2012. Trivial alignments in maximum likelihood analysis of nucleotide data. Cladistics 28, 514–528.

Farris, J.S., 1972. Estimating phylogenetic trees from distance matrices. Am. Nat. 106, 645–668.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.

Felsenstein, J., 1984. Distance methods for inferring phylogenies: a justification. Evolution 38, 16–24.

Giribet, G., Edgecombe, G.D., 2013. Stable phylogenetic patterns in scutigeromorph centipedes (myriapoda: Chilopoda: Scutigeromorpha): dating the diversification of an ancient lineage of terrestrial arthropods. Invertebr. Syst. 27, 485–501.

Giribet, G., Wheeler, W.C., 1999. The position of arthropods in the animal kingdom: Ecdysozoa, islands, trees and the 'parsimony ratchet'. Mol. Phylogenet. Evol. 10, 1–5.

Giribet, G., Wheeler, W.C., 2001. Some unusual small-subunit ribosomal DNA sequences of metazoans. Am. Mus. Novit. 3337, 1–14.

Gladstein, D.S., Wheeler, W.C. 1997. POY version 2.0. program and documentation available at http://research.amnh.org/scicomp/projects/poy.php. American Museum of Natural History, New York.

Goloboff, P.A. 1999. NONA (No Name) ver. 2. Published by the author. Tucumán, Argentina.

Goloboff, P., Farris, S., Nixon, K., 2003. TNT (Tree analysis using New Technology) version 1.0 ver. beta test v. 0.2. program and documentation available at http://www.lillo.org.arg/phylogeny/tnt. Published by the authors. Tucumán, Argentina.

Hasegawa, M., Kashino, H., Yano, T., 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22, 160–174.

Janies, D., Hill, A.W., Guralnick, R., Habib, F., Waltari, E., Wheeler, W.C., 2007. Genomic analysis and geographic visualization of the spread of avian influenza (H5N1). Syst. Biol. 56, 321–329.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules, pp. 21–132. In Munro, N.H. (ed.), Mammalian Protein Metabolism. Academic Press, New York.

Kimura, M., 1980. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111–120.

Kluge, A.G., 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst. Zool. 38, 7–25.

Moilanen, A., 1999. Searching for most parsimonious trees with simulated evolutionary optimization. Cladistics 15, 39–50.

Neyman, J. 1971. Molecular studies in evolution: a source of novel statistical problems. In Gupta, S.S., Yackel, J. (eds.), Statistical Decision Theory and Related Topics, pp. 1–27. Academic Press, New York.

Rannala, B., Yang, Z., 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. 43, 304–311.

Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147.

Schulmeister, S., Wheeler, W.C., 2004. Comparative and phylogenetic analysis of developmental sequences. Evol. Dev. 6, 50–57.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 19, 205–221.

Stamatakis, A., Ludwig, T., Meier, H., 2005. Raxml-iii: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21, 456–463.

Steel, M., Penny, D., 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17, 839–850.

Sugiura, N., 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. Commun. Stat. Theory Methods 7, 13–27.

Tamura, H., Nei, M., 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10, 512–526.

Tavaré, S., 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math. Life Sci. 17, 57–86.

Thorne, J.L., Kishino, H., Felsenstein, J., 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 33, 114–124.

Tuffley, C., Steel, M., 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59, 581–607.

Varón, A., Wheeler, W.C., 2012. The tree-alignment problem. BMC Bioinformatics 13, 293.

Varón, A., Vinh, L. S., Bomash, I., Wheeler, W.C. 2008. Poy 4.0. American Museum of Natural History. http://research.amnh.org/scicomp/projects/poy.php.

Varón, A., Vinh, L.S., Wheeler, W.C., 2010. POY version 4: phylogenetic analysis using dynamic homologies. Cladistics 26, 72–85.

Vinh, L., Varón, A., Wheeler, W.C., 2006. Pairwise alignment with rearrangement. Genome Inform. 17, 141–151.

Vinh, L., Janies, A.V.D., Wheeler, W.C., 2007. Towards phylogenomic reconstruction. In Proceedings of the 2007 International Conference on Bioinformatics and Computational Biology, pp. 98–104, Las Vegas, NV, USA.

Wheeler, W.C., 1996. Optimization alignment: the end of multiple sequence alignment in phylogenetics? Cladistics 12, 1–9.

Wheeler, W.C., 1999. Fixed character states and the optimization of molecular sequence data. Cladistics 15, 379–385.

Wheeler, W.C., 2006. Dynamic homology and the likelihood criterion. Cladistics 22, 157–170.

Wheeler, W.C., 2007. Chromosomal character optimization. Mol. Phylogenet. Evol. 44, 1130–1140.

Wheeler, W.C., 2012. Systematics: A Course of Lectures. Wiley-Blackwell, Chichester.

Wheeler, W.C., 2014. Maximum a posteriori probability assignment (MAP-A): an optimality criterion for phylogenetic trees via weighting and dynamic programming. Cladistics, 30, 282–290.

Wheeler, W.C., Whiteley, P.M., 2014. Historical linguistics as a sequence optimization problem: the evolution and biogeography of Uto-Aztecan languages. Cladistics, 10.1111/cla.12078.

Wheeler, W.C., Gladstein, D.S., De Laet, J. 1996-2005. POY version 3.0. program and documentation available at http://research.amnh.org/scicomp/projects/poy.php (version 3.0.11). documentation by D. Janies and W. C. Wheeler. commandline documentation by J. De Laet and W. C. Wheeler. American Museum of Natural History, New York.

Wheeler, W.C., Lucaroni, N., Hong, L., Crowley, L., Varón, A. 2013. Poy 5.0. American Museum of Natural History. http://research.amnh.org/scicomp/projects/poy.php.