

# Phylogenetic hypotheses and the utility of multiple sequence alignment

Ward C. Wheeler<sup>1</sup> and Gonzalo Giribet<sup>2</sup>

<sup>1</sup> Division of Invertebrate Zoology, American Museum of Natural History

Central Park West at 79th Street, New York, NY 10024-5192

wheeler@amnh.org

<sup>2</sup> Museum of Comparative Zoology

Department of Organismic & Evolutionary Biology, Harvard University

16 Divinity Avenue, Cambridge, MA 02138, USA

ggiribet@oeb.harvard.edu

November 13, 2007

## Abstract

The role of Multiple Sequence Alignment in phylogenetic analysis is discussed in the context of data and hypothesis. Alignments cannot be observed in nature, hence are neither data nor “real” in the scientific sense. Observers gather sequence data as strings of nucleotides and phylogenetic hypotheses (= topologies) are tested with them on the basis of quantitative optimality criteria. This optimization problem, the Tree-Alignment problem, is known to be NP-Hard, hence extremely unlikely to have an exact solution in polynomial time. Multiple sequence alignment can play a role, however, as a tool in identifying heuristic solutions to this problem. As such, it must be evaluated against other such tools. Real data sets and recent simulations offer tests of the performance of multiple sequence alignment as a heuristic approach to the tree-alignment problem.

# 1 Introduction

Multiple sequence alignment (MSA) is not a necessary, but is a potentially useful, technique in phylogenetic analysis. By this, we mean that we can construct and evaluate phylogenetic hypotheses without MSA, but it may be productive in terms of time or optimality, to do so. In order to evaluate this statement, we must first define phylogenetic hypothesis, define the problem, define the criteria we will use to assay the relative merits of hypotheses, define what we mean by the utility of a technique, and then finally compete alternate techniques.

In the following sections, each of these terms and operations are defined. The final section will compete a “one-step” optimization heuristic (Direct Optimization; Wheeler, 1996) embodied in POY4 (Varón et al., 2007) with the Multiple Sequence Alignment + Search approach (“two-step” phylogenetics *sensu* Giribet, 2005) embodied by CLUSTAL (Higgins and Sharp, 1988) using a large number of small data set simulations run under a variety of conditions (Ogden and Rosenberg, 2007) and a few larger (hundreds to over 1000 taxa) real data sets.

## 2 Phylogenetic Hypotheses

A phylogenetic hypothesis is a topology ( $T$ ), a tree linking terminal taxa (leaves or OTUs) through internal vertices (or HTUs) without cycles. More formally,  $T = (V, E)$  where  $V$  are the vertices both terminal leaves and internal, and  $E$  the edges or branches that link  $V$ . Furthermore, there must be an assignment  $\chi$  of observed data  $D$  to  $V$ , and a cost function  $\sigma$  that specifies the transformation costs between sequence elements (for this discussion A, C, G, T, and GAP or “-”). The phylogenetic hypothesis ( $H$ ) then can be expressed as  $H = (T, \chi_D, \sigma)$ . For simplicity, the discussion and examples here (following Ogden and Rosenberg, 2007) will use the homogenous  $\sigma = 1$ . In addition,  $\chi_D$  will always be a function of  $D$ , hence we can rewrite as  $H = (T, D)$  or just  $T(D)$ .

This topology may represent historical relationships or be simple summaries of hierarchical variation, but their form is the same. All hypotheses explain all potential data, just not to the same extent. Hence, those hypotheses that explain the data “best” are favored over others (summarized by Giribet and Wheeler, 2007).

## Observations and Data

In order to test hypotheses, we require data. Data are the observations an investigator makes in nature. DNA sequence data are gathered as contiguous strings of nucleotides from individual taxa. These data are observed without reference to the sequences of other creatures. Entire genomes can be sequenced without knowledge of any other entity. Nucleotides are observed only in reference to those that are collinear in the same taxon. MSAs are highly structured, inferential objects constructed by scientists either automatically or manually. They do not exist in nature, they can not be observed, they are not data.

## 3 The Tree-Alignment Problem

The problem of assigning vertex sequences such that the overall tree cost is minimized when sequences may vary in length is known as the Tree-Alignment-Problem (TAP; Sankoff, 1975; Sankoff et al., 1976). A phylogenetic search seeks to minimize the TAP cost over the universe of possible trees. Such an approach embodies the notion of “dynamic homology” (Wheeler, 2001) as opposed to “static homology” where predetermined correspondences and putative homologies are established prior to analysis and applied uniformly throughout tree search.

Unfortunately, this problem is known to be NP-Hard (Wang and Jiang, 1994), meaning that no polynomial time solution exists (unless  $P=NP$ ). In other words, the search for vertex median sequences is as hard as the phylogeny search problem over

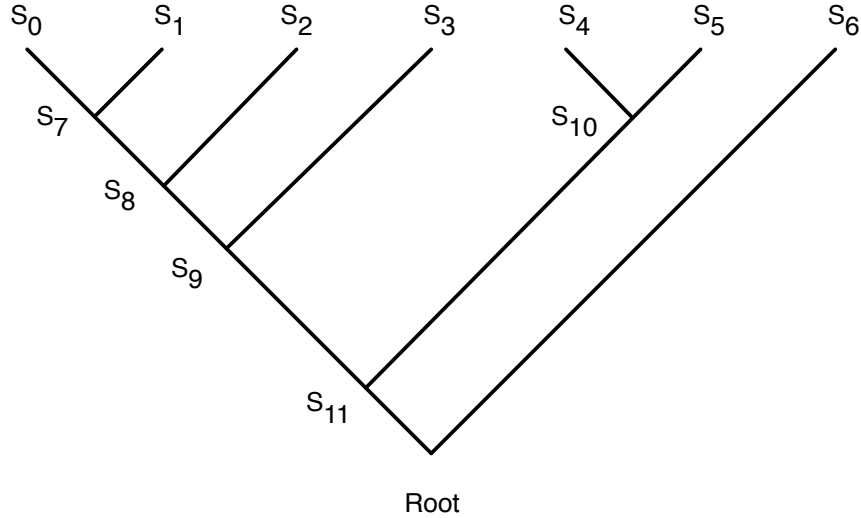


Figure 1: The tree alignment minimization assigns medians  $\{S_7 \dots S_{11}\}$  given observed leaf sequences  $\{S_0 \dots S_6\}$  such that the overall tree cost summed over all edges  $E$  is minimized. The root sequence can be any sequence between  $S_{11}$  and  $S_6$ .

tree space for static homology characters. As with tree searches, other than explicit or implicit exhaustive searches for trivial cases, we will always be limited to heuristic solutions (see [Slowinski, 1998] for numbers of homology scenarios).

## 4 Criteria to Evaluate Hypotheses

In order to compare hypotheses, there must be an explicit objective criterion. At its most fundamental level, a distance function  $d$  is specified to determine the pairwise cost of transforming each ancestor vertex into its descendent along an edge. This distance may be minimization-based (such as parsimony) or statistical in nature (such as likelihood), the TAP itself is agnostic. The distance must, at minimum, accommodate substitutions, insertions, and deletions (although substitutions are not strictly necessary, we will proceed as if they were).

Implementations of these criteria under dynamic homology have been proposed for parsimony (Wheeler, 1996), likelihood (Steel and Hein, 2001; Wheeler, 2006), and

posterior probability (Redelings and Suchard, 2005). Each of these criteria will allow comparison of topological hypotheses (and their associated vertex sequence assignments). The central idea is that there is such an explicit value that can be calculated and compared. The remainder of the discussion here will use equally weighted parsimony as the criterion of choice. Hence, results and conclusions are specific to this flavor of this criterion.

## 5 Heuristic Techniques

In principle, we could solve the TAP by examining all possible vertex median sequences. This has been proposed by Sankoff and Cedergren (1983) through  $n$ -dimensional alignment using  $T$  to determine cell costs. Another method using dynamic programming over all possible sequences (Wheeler, 2003c) would yield the same result. Only trivial data sets are amenable to such analysis.

Discussing heuristic solutions, Wheeler (2005) categorized these heuristic approaches into two groups: those that attempt to estimate vertex median sequences directly and those that examine candidate medians from a predefined set. Estimation methods calculate medians based on the sequences of the vertices adjacent to them. Direct Optimization (DO; Wheeler, 1996) uses the two descendent vertices for an (length  $n$  sequences)  $O(n^2)$ . Iterative Pass (Sankoff et al., 1976; Wheeler, 2003b) uses all three connected vertices and revisits vertices for improved medians, but at a time complexity of  $O(n^3)$ . Search methods such as “lifted” alignments (Gusfield, 1997), Fixed-States (Wheeler, 1999), and Search-Based (Wheeler, 2003c) employ predefined candidate sequence medians in increasing number (for  $m$  taxa, lifted uses  $m/2$ , fixed-states  $m$ , and search-based  $> m$ ). Time complexity of lifted alignments is linear with  $m$  sequences and quadratic for both fixed-states and search-based (after an  $O(m^2n^2)$  edit cost matrix setup). A polynomial time approximation (PTAS) exists for TAP (Wang and Gusfield, 1997) but the time complexity is too great to be of any practical use.

## Multiple Sequence Alignment and The Tree Alignment Problem

As mentioned above, the determination of the cost (for any  $d$ ) of each individual candidate tree is NP-Hard. This is then compounded by the complexity of the tree search. The simultaneous optimization of both these problems can be extremely time consuming. The motivation behind MSA is to separate these problems, performing the homology step (MSA) only once. The determination of tree cost for these now static characters is linear with the length of the aligned sequences and tree search can proceed with alacrity. This is a reasonable heuristic procedure whose behavior will depend on the appropriateness of using that single MSA for all tree evaluations. Obviously, for trivial cases this will be as effective as more exhaustive approaches. The method can be further refined by linking MSA more closely with the tree search by generating new MSAs based on a TAP “Implied Alignment” (Wheeler, 2003a) and iteratively alternating between static and dynamic searches until a local minimum is found (“Static Approximation;” Wheeler, 2003a).

Each of these techniques can be evaluated on two bases, the quality of the solution in terms of optimality value, and execution time. Here, we concern ourselves with the optimality of the solution, although it is clear that a good solution (such as PTAS or exact) may be “better” by optimality, but of little use due to their time complexity.

## 6 Evaluation of Heuristic Techniques

In order to examine the relative effectiveness of MSA, we will use equally weighted parsimony as our optimality criterion. Equal weighting is not used because of some innate superiority of this form of analysis (see Grant and Kluge, 2005; Giribet and Wheeler, 2007 for some acrimony), but because it offers a clear and simple test (similar reasoning motivated Ogden and Rosenberg, 2007). Other indices could be used, and

other results found, hence our conclusions are restricted.

There has been some discussion in the literature about “true” alignments and their place in evaluating phylogenetic methods (Kjer et al., 2007; Ogden and Rosenberg, 2007), more specifically MSA and DO. These authors, and others, have set up the test of these methods as the recovery of “true” alignment (known by simulation or other inferred qualities). It is the position taken here that this is incorrect. Alignments are not an attribute of nature. They cannot be observed, only created by automated or manual means. Whether or not a method can create a MSA directly or as a tree adjuvant that matches a notion based in simulation or imagination is irrelevant to its quality as a solution to the TAP.

Ogden and Rosenberg (2007) performed an admirably thorough set of simulations (15,400) on small set of taxa (16) for realistically sized sequences (2000 nucleotides) under a variety of tree topology-types and evolutionary conditions/models. To summarize, 100 replicate simulations were performed on seven tree topologies (balanced, pectinate, and five “random” topologies) under ultrametric, clock-like, and non-clocklike evolution, with two rates of change for 154 combinations. Ogden and Rosenberg then stripped out evolved (= true) gaps and reanalyzed the sequences in two ways. The first was the traditional “two-step” phylogenetics of alignment and subsequent analysis of static data. This was accomplished with CLUSTAL (Higgins and Sharp, 1988) under default conditions and PAUP\* (Swofford, 2002). The second was “one-step” analysis using POY3 (Wheeler et al., 2005). Ogden and Rosenberg compared the implied alignments (Wheeler, 2003a) generated by POY and offered by CLUSTAL with the simulated “true” alignments. The POY implied alignments were found to be more dissimilar to the simulated alignments than were those of CLUSTAL.

Although the specifics of Ogden and Rosenberg’s use of POY could be a subject of discussion, the objective here is not to take issue with the details of their analysis, but their general approach. Although Ogden and Rosenberg did look at topologies in a secondary comparison, the authors never examined the optimality effectiveness of their



competing approaches. No tree costs were reported.

We reanalyzed all 15,400 simulations performed and generously provided us by Ogden and Rosenberg. Three analyses were performed in each case yielding 46,200 runs. In the first, we used the CLUSTAL alignments of Ogden and Rosenberg, running them as static non-additive characters (all transformations equal), in the second the “true” alignments were used again as static non-additive characters, and in the third the unaligned data were analyzed under equal transformation costs (indels=1) using DO. All analyses were performed using POY4 beta 2398 (Varón et al., 2007) with 10 random addition sequences and TBR branch swapping on several Mac Intel machines. POY4 replaced PAUP\* in the static analyses for consistency of heuristic approach. For such small data sets, large differences are unlikely to appear. Given the settings and problem here, POY4 differs from POY3 mainly in the efficiency of implementation, the core algorithms for DO are the same. The POY4 runs on unaligned data took approximately 10-20x those of the pre-aligned data.

In every one of the 15,400 comparison cases (Table 1), POY4 yielded a lower cost than CLUSTAL+POY4. The average tree cost differences for the 154 experimental combinations (over the 100 replicate simulations) were as low as 2% and as high as 20%. The higher rates of evolution (maximum distance = 2) had greater deficits compared to POY4 than the lower. Interestingly, analysis of the simulated and CLUSTAL analyses showed tree costs that were similar, with neither obviously producing lower cost trees.

As far as these simulations are concerned, the one-step heuristic approach of POY is overwhelmingly superior to that of the two step alignment + tree search approach advocated by Ogden and Rosenberg (2007).

## **Real Data and Heuristics**

As a reality check, we performed the same pairs of analyses on four larger, real data sets (Table 2) used in Wheeler (2007). These data sets were all ribosomal DNA and varied in

Model	Balanced			Pectinate			Random A			Random B		
	C/P	T/P	T/C	C/P	T/P	T/C	C/P	T/P	T/C	C/P	T/P	T/C
a	1.0299	1.0266	1.00	1.0726	1.0604	0.989	1.0433	1.0335	0.9906	1.0527	1.0423	0.9902
b	1.0993	1.1172	1.0164	1.1795	1.2068	1.02	1.1276	1.1267	0.9992	1.1327	1.1437	1.01
RBL-1a	1.0525	1.0372	0.9855	1.0749	1.0562	0.9826	1.0522	1.0526	1.0004	1.0548	1.0416	0.988
RBL-1b	1.1680	1.1690	1.0009	1.188	1.2068	1.0161	1.12	1.1513	1.0268	1.1643	1.1680	1.0032
RBL-2a	1.0502	1.034	0.9843	1.0763	1.0592	0.9843	1.0565	1.0386	0.9830	1.0600	1.0430	0.9840
RBL-2b	1.1811	1.1664	0.9876	1.1728	1.1961	1.0198	1.21	1.1884	0.9821	1.1700	1.1713	1.0011
RBL-3a	1.0447	1.0330	0.99	1.0765	1.0634	0.9879	1.0440	1.0340	0.9904	1.0612	1.052	0.9918
RBL-3b	1.1453	1.1420	0.9971	1.1836	1.2143	1.0260	1.1271	1.1260	0.9990	1.1449	1.1652	1.0178
RBL-4a	1.0476	1.0356	0.9886	1.0732	1.0578	0.9858	1.0696	1.0507	0.9825	1.0654	1.0476	0.9833
RBL-4b	1.1464	1.1435	0.9975	1.1664	1.1877	1.0183	1.2042	1.2162	1.0100	1.2017	1.2013	0.9997
RBL-5a	1.0423	1.0303	0.9885	1.0892	1.0645	0.9774	1.0511	1.0441	0.9934	1.0602	1.0432	0.9839
RBL-5b	1.1356	1.1391	1.0031	1.2104	1.239	1.0238	1.1296	1.148	1.0160	1.1990	1.1888	0.9916
RBLNoC-1a	1.0368	1.0186	0.9825	1.0517	1.0480	0.9966	1.0351	1.0262	0.9915	1.0745	1.0536	0.9808
RBLNoC-1b	1.0827	1.0716	0.9898	1.0942	1.1116	1.0159	1.0984	1.0898	0.9922	1.1155	1.1467	1.0279
RBLNoC-2a	1.0340	1.0188	0.9853	1.0563	1.0366	0.9814	1.0470	1.0325	0.9862	1.0616	1.0450	0.9845
RBLNoC-2b	1.0815	1.0676	0.9872	1.1406	1.1310	0.9916	1.1090	1.1110	1.0018	1.1074	1.1315	1.0217
RBLNoC-3a	1.0436	1.0209	0.978	1.0805	1.0530	0.9747	1.0396	1.0285	0.9893	1.0480	1.034	0.9868
RBLNoC-3b	1.0934	1.0807	0.9884	1.1468	1.1771	1.0265	1.1013	1.1013	1.0001	1.1066	1.1065	0.9999
RBLNoC-4a	1.0390	1.0195	0.9812	1.056	1.0449	0.9898	1.038	1.0274	0.9898	1.0513	1.0405	0.9898
RBLNoC-4b	1.0860	1.0757	0.991	1.1274	1.1349	1.0066	1.0860	1.0801	0.9946	1.0987	1.1105	1.0107
RBLNoC-5a	1.0389	1.0236	0.9854	1.0656	1.0518	0.9871	1.0345	1.0260	0.9918	1.0353	1.0190	0.9843
RBLNoC-5b	1.0967	1.093	0.9966	1.1258	1.1520	1.0232	1.0897	1.0872	0.998	1.095	1.0741	0.9805
Model	Random C			Random D			Random E					
	C/P	T/P	T/C	C/P	T/P	T/C	C/P	T/P	T/C			
a	1.0621	1.054	0.9924	1.048	1.0351	0.9880	1.0478	1.038	0.9904			
b	1.1440	1.1736	1.0259	1.1397	1.1412	1.0014	1.1296	1.1344	1.004			
RBL-1a	1.057	1.0494	0.9929	1.0613	1.0445	0.9842	1.0479	1.0358	0.9885			
RBL-1b	1.1478	1.165	1.015	1.1934	1.1939	1.0004	1.140	1.1327	0.9939			
RBL-2a	1.0630	1.0576	0.9949	1.0618	1.0573	0.9958	1.0485	1.0430	0.99			
RBL-2b	1.1624	1.1870	1.0212	1.1540	1.1860	1.0278	1.1193	1.133	1.0123			
RBL-3a	1.06	1.0439	0.9872	1.0507	1.0378	0.9877	1.0638	1.0471	0.9843			
RBL-3b	1.1729	1.1710	0.9984	1.1451	1.1497	1.0040	1.2059	1.2019	0.9967			
RBL-4a	1.0621	1.0474	0.9861	1.0594	1.0425	0.9841	1.0585	1.0418	0.9843			
RBL-4b	1.1716	1.1837	1.0104	1.1909	1.1876	0.9973	1.1936	1.1848	0.9926			
RBL-5a	1.0574	1.0503	0.9933	1.0476	1.0373	0.9902	1.0622	1.0558	0.9940			
RBL-5b	1.1442	1.1596	1.0135	1.1442	1.1444	1.0002	1.168	1.1958	1.0242			
RBLNoC-1a	1.0600	1.0396	0.9808	1.0449	1.0280	0.9839	1.0466	1.0326	0.9867			
RBLNoC-1b	1.1330	1.1425	1.0084	1.1103	1.1108	1.0005	1.1028	1.105	1.0022			
RBLNoC-2a	1.0500	1.0403	0.991	1.0430	1.0307	0.9883	1.0457	1.0301	0.99			
RBLNoC-2b	1.1097	1.1131	1.003	1.097	1.0981	1.0008	1.1164	1.112	0.9958			
RBLNoC-3a	1.0519	1.0411	0.9897	1.0358	1.0243	0.9889	1.0351	1.023	0.9882			
RBLNoC-3b	1.1256	1.1300	1.0039	1.097	1.0783	0.9826	1.0966	1.0872	0.9915			
RBLNoC-4a	1.0682	1.0545	0.9872	1.0420	1.0246	0.9834	1.0323	1.0196	0.9876			
RBLNoC-4b	1.1197	1.1490	1.0262	1.1120	1.0830	0.9740	1.0937	1.074	0.9819			
RBLNoC-5a	1.0615	1.0484	0.9877	1.0431	1.0237	0.9814	1.0369	1.0265	0.9900			
RBLNoC-5b	1.1248	1.1413	1.0147	1.1050	1.0909	0.9873	1.1005	1.0943	0.9946			

Table 1: Tree cost comparisons of simulated data of Ogden and Rosenberg (2007). The model column specifies the evolutionary model and rate of the simulation, Balanced, Pectinate, and Random A-E the tree topologies. C/P denoted the average (over 100) trials of the cost ratio of the CLUSTAL+ Search trees and POY trees, T/P the average ratios of “true” alignment to POY costs, and T/C the ratios of “true” alignment to CLUSTAL+Search trees.

sized from 62 taxa to 1040. They were analyzed with CLUSTAL under the same default conditions as Ogden and Rosenberg (2007) (gap opening=15, extension=6.66, delay divergent=30% transition weight=0.50, DNA weight matrix=IUB) and performed the same 10 replicates + TBR for aligned and unaligned data. In each of these four, cases the single-step POY4 tree costs were from 6% to 17% lower.

## 7 Conclusions

Given that alignments are not “real” in a any natural sense, their role can only be as a heuristic tool in the solution of phylogenetic problems. The core hypothesis of phylogenetic analysis is that topology which optimizes some measure of merit. The simulations of Ogden and Rosenberg (2007) and the analyses presented here clearly show that for this type of analysis, MSA is an inferior heuristic as far as generating low cost solutions to the TAP. MSA may be a useful tool in accelerating searches (such as in Static-Approximation) heuristics, but on its own it falls short. In fact, the only case where MSA could be self-consistent would be if a complete set of *optimal* implied alignments (there are likely many for any given tree) were to be generated for the *optimal* set of trees (again there may be many). In this case, static (=two-step) analysis of this complete set would return the same set of optimal source trees. Only then would there be a solid relationship between optimal alignments and optimal trees. Given that both of these sets of objects are unlikely to be found and recognized for

Data Set	Taxa	POY	CLUSTAL+POY	Cost Ratio
Mantid 18S	62	956	1052	1.1004
Metazoa 18S	208	26697	30983	1.1605
Archaea SSU	585	37003	39193	1.0592
Mitochondrial SSU	1040	77753	90685	1.1663

Table 2: One and two step analyses of four larger real data sets. All transformations were set to unity (indels=1). The mantid data are from Svenson and Whiting (2004) the other data collected in Wheeler (2007).

any non-trivial data set, this situation exists only in theory.

Multiple sequence alignments are neither real, nor particularly useful. So what keeps them around other than tradition, inertia, and luddism?

## **8 Acknowledgments**

The National Science Foundation for financial support. Heath Odgen and Michael Rosenberg for supplying there simulation files. Andrés Varón and Louise Crowley for discussion of this manuscript.

## References

- Giribet, G. 2005. Tree analysis using new technology. *Syst. Biol.* pp. 176 – 178.
- Giribet, G. and Wheeler, W. C. 2007. The case for sensitivity: a response to grant and kluge. *Cladistics* 23:1–3.
- Grant, T. and Kluge, A. G. 2005. Stability, sensitivity, science, and heurism. *Cladistics* 21:597–604.
- Gusfield, D. 1997. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press.
- Higgins, D. G. and Sharp, P. M. 1988. Clustal: A package for performing multiple sequence alignment on a microcomputer. *Gene* 73:237–244.
- Kjer, K. M., Gillespie, J. J., and Ober, K. A. 2007. Opinions on multiple sequence alignment, and an empirical comparison of repeatability and accuracy between poy and structural alignment. *Syst. Biol.* 56:133–146.
- Ogden, T. H. and Rosenberg, M. S. 2007. Alignment and topological accuracy of the direct optimization approach via poy and traditional phylogenetics via clustalw + paup\*. *Sys. Biol.* 56:182–193.
- Redelings, B. D. and Suchard, M. A. 2005. Joint bayesian estimation of alignment and phylogeny. *Syst. Biol.* 54:401–418.
- Sankoff, D., Cedergren, R. J., and Lapalme, G. 1976. Frequency of insertion-deletion, transversion, and transition in the evolution of 5s ribosomal rna. *J. Mol. Evol.* 7:133–149.
- Sankoff, D. M. 1975. Minimal mutation trees of sequences. *SIAM J. Appl. Math.* 28:35–42.

- Sankoff, D. M. and Cedergren, R. J. 1983. Simultaneous comparison of three or more sequences related by a tree, pp. 253–263. *In* D. M. Sankoff and J. B. Kruskal (eds.), Time warps, string edits, and macromolecules: the theory and practice of sequence comparison, chapter 9. Addison Wesley, Reading, Massachusetts.
- Slowinski, J. B. 1998. The number of multiple alignments. *Mol. Phylogen. Evol* 10:264–266.
- Steel, M. and Hein, J. 2001. Applying the thorne-kishino-felsenstein model to sequence evolution on a star-shaped tree. *Appl. Math. Let.* 14:679–684.
- Svenson, G. J. and Whiting, M. F. 2004. Phylogeny of Mantodea based on molecular data: evolution of a charismatic predator. *Syst. Ent.* 29:359–370.
- Swofford, D. L. 2002. Paup\*: Phylogenetic analysis using parsimony (\* and other methods), version 4.0b 10. Sinauer Associates, Sunderland, Massachusetts.
- Varón, A., Vinh, L. S., Bomash, I., and Wheeler, W. C. 2007. Poy 4.0 beta 2013. American Museum of Natural History. <http://research.amnh.org/scicomp/projects/poy.php>.
- Wang, L. and Gusfield, D. 1997. Improved approximation algorithms for tree alignment. *J. of Alg.* 25:255–273.
- Wang, L. and Jiang, T. 1994. On the complexity of multiple sequence alignment. *Journal of Computational Biology* 1:337–348.
- 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. 2001. Homology and the optimization of DNA sequence data. *Cladistics* 17:S3–S11.
- Wheeler, W. C. 2003a. Implied alignment. *Cladistics* 19:261–268.
- Wheeler, W. C. 2003b. Iterative pass optimization. *Cladistics* 19:254–260.
- Wheeler, W. C. 2003c. Search-based character optimization. *Cladistics* 19:348–355.
- Wheeler, W. C. 2005. Alignment, dynamic homology, and optimization, pp. 73–80. *In* V. Albert (ed.), *Parsimony, Phylogeny, and Genomics*. Oxford University Press.
- Wheeler, W. C. 2006. Dynamic homology and the likelihood criterion. *Cladistics* 22:157–170.
- Wheeler, W. C. 2007. The analysis of molecular sequences in large data sets: where should we put our effort? *In* T. R. Hodkinson and J. A. N. Parnell (eds.), *Reconstructing the Tree of Life: Taxonomy and Systematics of Species Rich Taxa*, pp. 113–128. Systematics Association, Oxford University Press.
- Wheeler, W. C., Gladstein, D. S., and De Laet, J. 1996–2005. POY version 3.0. program and documentation available at <http://research.amnh.org/scicomp/projects/poy.php> (current 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.