WILEY-BLACKWELL

Cladistics (2015) 1–9

Cladistics

10.1111/cla.12142

Comparison of heuristic approaches to the generalized tree alignment problem

Eric Ford[a,b,*] and Ward C. Wheeler[b]

[a]*Department of Mathematics & Computer Science, Lehman College, CUNY, Bronx, NY 10468, USA;* [b]*Division of Invertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA*

Accepted 10 September 2015

Abstract

Two commonly used heuristic approaches to the generalized tree alignment problem are compared in the context of phylogenetic analysis of DNA sequence data. These approaches, multiple sequence alignment + phylogenetic tree reconstruction (MSA+TR) and direct optimization (DO), are alternative heuristic procedures used to approach the nested NP-Hard optimizations presented by the phylogenetic analysis of unaligned sequences under maximum parsimony. Multiple MSA+TR implementations and DO were compared in terms of optimality score (phylogenetic tree cost) over multiple empirical and simulated datasets with differing levels of heuristic intensity. In all cases examined, DO outperformed MSA+TR with average improvement in parsimony score of 14.78% (5.64–52.59%).

A central goal of biological systematics is mapping the relationships among organisms and groups of organisms—both extant and extinct—based on the reconstruction of phylogenetic trees using comparative character data. The generalized tree alignment problem (GTAP; Sankoff, 1975) is defined as the search for phylogenetic tree(s)—and associated vertex (hypothetical ancestor) sequences—with lowest cost for those data under maximum parsimony.

There has been an ongoing debate in the literature regarding multiple sequence alignment (Katoh et al., 2002; Edgar, 2004; Wheeler, 2007), with several aligners available. In addition, much effort has been expended to improving search on aligned sequences (Goloboff et al., 2003). At the same time, other paradigmata for approaching the GTAP are also available, chief among those being direct optimization (DO) (Wheeler, 1996, 2003; Varón and Wheeler, 2012, 2013). It has been the experience of many investigators that DO gives significantly better results than the two-step process of alignment followed by search for both real and simulated data (e.g. Lindgren and Daly, 2007; Lehtonen, 2008; Liu et al., 2009; Giribet and Edgecombe, 2013). In addition, the high degree of complexity in the settings of the software tools used for alignment and search only confuses the matter, as default settings are often used, and these defaults do not necessarily correspond between aligner and search engine. Here, we compare DO with two-step solutions directly. We also test whether the results of searches where alignment and search setting correspond are better (i.e. more optimal) than those in which they do not. We find that DO results in the discovery of shorter trees, by an average factor of 15%. In addition, using the two-step approach we found significantly (approximately 4%) shorter trees when using settings on alignment that match the settings of subsequent tree search (as opposed to the default settings of multiple sequence alignment (MSA) implementations).

Software tools

We ran comparisons using several pieces of alignment software. What follows is a brief description of each package.

*Corresponding author:
E-mail address: eford@gradcenter.cuny.edu

CLUSTAL OMEGA (Sievers et al., 2011) uses a guide tree to align sequences. It is similar to CLUSTAL W, described below, but speeds the process of creating the guide tree by re-encoding each sequence as an *n*-dimensional vector, which can be viewed as its similarity to *n* reference sequences. This allows for more rapid clustering using hidden Markov models.

CLUSTAL W 2.0.12 (Larkin et al., 2007) pre-dates CLUSTAL OMEGA and uses neighbour-joining (Saitou and Nei, 1987). In addition, CLUSTAL W weights the sequences, giving more-divergent sequences more weight, in an attempt to get improved results in pairwise comparisons.

MAFFT v7.029b (Katoh et al., 2002) performs progressive alignments, also using a guide tree, and uses fast Fourier transform to speed up identification of homologous regions in sequences. Depending on the number of taxa, MAFFT uses either progressive or iterative alignment algorithms. Both MAFFT and MUSCLE, described below, use UPGMA (unweighted pair group method with arithmetic mean, a hierarchical clustering method using a similarity matrix), rather than neighbour-joining, as in CLUSTAL W.

MUSCLE v3.8.31 (Edgar, 2004) also uses a guide tree to direct alignment. To build the initial tree, MUSCLE uses the approximate *k*mer pairwise distances. A *k*mer is a subsequence sample of length *k*, and the *k*mer distance used by MUSCLE is the fraction of *k*mers in common in a compressed alphabet. This initial tree build is followed by progressive alignments using the Kimura distance (Kimura and Ohta, 1972).

In contrast to the above packages, POY (5.0 and antecedents; Wheeler et al., 2015, 2013) uses DO. This approach optimizes median sequences on trees created via heuristic search. The tree search itself is performed using heuristics similar to those implemented in TNT (Goloboff et al., 2003), but because the median optimization and scoring are done as each tree is constructed, a POY search is computationally more complex than a TNT search. In short, TNT is dealing with one NP-Hard optimization (tree search), whereas POY is dealing with two (tree search and tree alignment).

After the alignment stage, we used TNT to search for the shortest trees. In an attempt to compare DO with two-step searches more thoroughly, we also ran TNT on the implied alignment results of POY searches.

Finally, to create synthetic data with length variation in a more controlled manner (i.e. gaps, and thus simulate unaligned sequences), we used DAWG 1.2-release (Cartwright, 2005).

Materials

Biological data

We ran alignment and search on six biological data sets and six synthetic data sets. The biological data sets were chosen for their sizes. Each of the software packages allows for the input of either genetic or protein data. Here, we restricted our comparisons to nucleotide sequence data.

- 62 mantodean small subunit sequences (Svenson and Whiting, 2004);
- 208 metazoan small subunit sequences (Giribet and Wheeler, 1999, 2001);
- 585 archaean small subunit sequence data from the European Ribosomal RNA Database (Wheeler, 2007);
- 1040 metazoan mitochondrial small subunit data from the European Ribosomal RNA Database (Wheeler, 2007);
- 1553 fungal small ribosomal subunit sequences extracted from GenBank (Benson et al., 2013);
- 1766 metazoan small subunit sequences extracted from GenBank (Benson et al., 2013).

Synthetic data

The synthetic data were created, using DAWG, by varying the number of taxa and the distribution of edge (branch) lengths. Each of the synthetic data sets was rooted, and included an outgroup. All had initial sequence lengths of 1800 bp (comparable to metazoan small subunit data). We chose taxon counts to match (approximately) the sizes of the smaller biological data sets, thus creating sets of 51, 201 and 601 taxa, including a root taxon. For each of these taxon sets, we used a Python script to create a random tree, then assigned edge lengths to the tree using two distributions, an exponential distribution with mean of 0.1, and a uniform distribution with values between 0.0 and 0.5. We therefore generated a total of six synthetic data sets using DAWG.

Methods

For each of the 12 data sets, a two-step process was followed. First, the sequences were aligned by each of the aligners (including POY implied alignment). The resulting alignments were then fed into TNT for tree search (specifics below). Thus, for each data set eight total alignments were generated (two each for MAFFT, POY/TNT and CLUSTAL OMEGA, one each for MUSCLE and CLUSTAL W), and for each of the eight alignments, two TNT searches were run. Runs using POY alone (DO

without MSA) were performed on each unaligned data set (labelled "POY" in the tables and figures). This yielded a total of 192 runs through TNT, but 216 searches, including POY solo searches.

On several of the runs, POY found multiple trees. Rather than performing multiple TNT searches on each of the implied alignments, in these cases we used the first tree and implied alignment output. As POY outputs these trees in the order that they were encountered (with a randomized component), this choice was arbitrary. DAWG takes as input a rooted tree with edge lengths and a sequence length, then evolves sequences following the tree. We generated random trees using a Python script, with edge lengths as noted above. Each of the six trees used to create the synthetic data was fed into DAWG. We used a GTR model with defaults settings for the frequency and parameters settings. The lambda parameter was set to 0.05 for both insertions and deletions. The gap model was power law with a gap rate of 1.67 and a maximum gap size of 2000 (these values were based on suggestions of the DAWG wiki). All other settings were default.

DAWG output includes gaps, and is aligned. Thus, after using DAWG to create alignments, the files were run through a Python script to delete gap characters, to generate unaligned sequences, which were then saved in FASTA format.

Due to the complexity of the software packages used, it is often the case that researchers run the applications using default settings. It is worth noting that the default settings of TNT (and probably all other tree-search software) differ from the defaults of the alignment software. That is, TNT uses (in essence) a gap opening cost of 0 and a gap extension cost of 1, whereas each of the aligners has a more sophisticated gap model. In fact, it is impossible to run TNT searches using the same settings as the alignment models, as all of the aligners studied here use affine gap models, and TNT does not (and cannot and maintain character independence). To measure the effect of this difference between the models, each of the data sets was aligned first with default settings in each aligner, and then with a gap cost of 0 and a gap extension cost of 1, which matches TNT's gap-cost scheme. This allowed us to track the effect that this difference in settings has on downstream tree lengths. In addition, when possible, we set all substitutions to a cost of 1, again to match the default of TNT. We hereafter refer to this as 0 : 1 : 1 settings (0 gap cost, 1 gap extension cost, 1 substitution cost). We also note that TNT can treat gaps as missing data (a commonly used option), which is the opposite of an indel—a hypothetical event, and not the same as a missing or mis-read character. To avoid this, make all tree costs comparable and account for all sequence transformation events, gaps were coded as a fifth state.

For the alignment stage, we first ran MAFFT with default settings, then reran it with gap opening cost of 0 and extension cost 1 (`–op 0 –ep 1 –lop 0 –lep 2`). MAFFT uses different alignment settings depending upon the size of the input data. To recreate those settings while changing the defaults for gap costs, three different profiles were created: L-INS-i, with commands `–localpair –maxiterate 1000`, FFT-NS-i (`–retree 2 –maxiterate 2`) and FFT-NS-2 (`–retree 2 –maxiterate 0`). For each data set, MAFFT was first run with default settings. Its terminal output was then read to determine which of the three profiles was used, and that profile was used when running MAFFT again with the custom gap costs. The specifics of MAFFT options are detailed in Katoh et al. (2002). The individual data sets and profiles used are listed in Table 1.

As we were unable to run MUSCLE using a custom gap cost, it was run only with default settings. Likewise, CLUSTAL OMEGA does not provide for the ability to run with custom gap costs, and it was therefore run only with default settings.

As with the other aligners, we ran CLUSTAL W first with default settings on each of the data sets, then with 0 : 1 : 1 settings: `-gapopen = 1 -gapext = 0 -ENDGAPS`, and using a custom base substitution cost matrix with costs of 0 for non-substitution and 1 for substitution.

We in turn ran POY using both less and more aggressive search strategies. The first run used `build()`, which builds ten random addition sequence Wagner trees using DO (referred to as "POY default"). The more aggressive run (`build() swap(transform(static_approx)) swap()`), builds Wagner trees, then creates an implied alignment and

Table 1
MAFFT settings for various data sets

Dataset	MAFFT settings
Mantodea	L-INS-i
Metazoa short	FFT-NS-i
Archaea	FFT-NS-2
Mitochondrial	FFT-NS-2
Fungi	FFT-NS-2
Metazoa long	FFT-NS-2
51-taxon exponential distribution	FFT-NS-i
51-taxon uniform distribution	FFT-NS-i
201-taxon exponential distribution	FFT-NS-i
201-taxon uniform distribution	FFT-NS-i
601-taxon exponential distribution	FFT-NS-2
601-taxon uniform distribution	FFT-NS-2

For the 0 : 1 : 1 runs we needed to match MAFFT's default settings as closely as possible. However, MAFFT's default settings for any given size of data set were unclear from the documentation. These settings were thus chosen by running MAFFT on each data set under default settings, then examining MAFFT's output to the interface to determine which settings were used for each data set. These determinations were then double-checked against the MAFFT manual.

Table 2
Alignment results, biological data

Data set	No. of taxa	Aligner	Settings	Aligned length (bp)	Tree length after standard search	Tree length after aggressive search
Mantodea	62	Clustal Omega	Default	1829	1159	1159
Mantodea	62	Clustal W	Default	1828	1055	1055
Mantodea	62	Clustal W	One-one	1893	1042	1042
Mantodea	62	mafft	Default	1832	1033	1033
Mantodea	62	mafft	One-one-l-ins-i	1872	1006	1006
Mantodea	62	muscle	Default	1853	1044	1044
Mantodea	62	poy		1909	945	939
Mantodea	62	poy/tnt	poy default	1909	942	942
Mantodea	62	poy/tnt	poy aggressive	1912	939	939
Metazoa short	208	Clustal Omega	Default	2863	38 544	38 536
Metazoa short	208	Clustal W	Default	2866	31 105	31 103
Metazoa short	208	Clustal W	One-one	4777	29 317	29 317
Metazoa short	208	mafft	Default	3377	32 842	32 837
Metazoa short	208	mafft	One-one-fft-ns-i	3792	29 024	29 014
Metazoa short	208	muscle	Default	3307	35 237	35 228
Metazoa short	208	poy		6625	26 876	26 817
Metazoa short	208	poy/tnt	poy default	6625	26 861	26 861
Metazoa short	208	poy/tnt	poy aggressive	6651	26 812	26 812
Archaea	585	Clustal Omega	default	1810	39 840	39 814
Archaea	585	Clustal W	Default	1757	39 362	39 340
Archaea	585	Clustal W	One-one	3100	40 560	40 548
Archaea	585	mafft	Default	1837	39 359	39 337
Archaea	585	mafft	One-one-fft-ns-2	2518	41 304	41 292
Archaea	585	muscle	Default	1705	40 760	40 734
Archaea	585	poy		6770	37 914	37 118
Archaea	585	poy/tnt	poy default	6770	37 777	37 775
Archaea	585	poy/tnt	poy aggressive	6391	37 106	37 106
Mitochondrial	1040	Clustal Omega	Default	3908	96 656	96 590
Mitochondrial	1040	Clustal W	Default	3111	91 032	91 014
Mitochondrial	1040	Clustal W	One-one	7198	86 669	86 652
Mitochondrial	1040	mafft	Default	5064	94 205	94 166
Mitochondrial	1040	mafft	One-one-fft-ns-2	5501	87 997	87 938
Mitochondrial	1040	muscle	Default	2435	128 178	128 041
Mitochondrial	1040	poy		14 474	79 191	77 832
Mitochondrial	1040	poy/tnt	poy default	14 474	79 065	79 065
Mitochondrial	1040	poy/tnt	poy aggressive	14 060	77 819	77 819
Fungi	1553	Clustal Omega	Default	2622	83 726	83 654
Fungi	1553	Clustal W	Default	2539	68 973	68 910
Fungi	1553	Clustal W	One-one	4543	67 135	67 059
Fungi	1553	mafft	Default	2844	69 185	69 125
Fungi	1553	mafft	One-one-fft-ns-2	3695	68 501	68 437
Fungi	1553	muscle	Default	2350	76 132	76 014
Fungi	1553	poy		11 610	63 473	62 000
Fungi	1553	poy/tnt	poy default	11 610	63 232	63 202
Fungi	1553	poy/tnt	poy aggressive	11 652	61 944	61 943
Metazoa long	1766	Clustal Omega	Default	4891	243 149	242 941
Metazoa long	1766	Clustal W	Default	4282	198 870	198 780
Metazoa long	1766	Clustal W	One-one	12 535	187 940	187 882
Metazoa long	1766	mafft	Default	7448	205 416	205 263
Metazoa long	1766	mafft	One-one-fft-ns-2	9066	187 180	187 093
Metazoa long	1766	muscle	Default	3554	249 222	249 076
Metazoa long	1766	poy		37 005	169 595	165 563
Metazoa long	1766	poy/tnt	poy default	37 005	169 447	169 443
Metazoa long	1766	poy/tnt	poy aggressive	36 317	165 524	165 522

does a hill-climbing search looking at TBR neighbourhoods, followed by an additional hill-climbing search on TBR neighbourhoods using DO.

In addition, we exported the implied alignment from the resulting poy search, essentially using poy only to do alignment, with search done by tnt.

Table 3
Alignment results, synthetic data

Data set	No. of taxa	Aligner	Settings	Aligned length (bp)	Tree length after standard search	Tree length after aggressive search
Synthetic Exponential	51	Clustal Omega	Default	4099	34 033	34 033
Synthetic Exponential	51	Clustal W	Default	4057	30 210	30 210
Synthetic Exponential	51	Clustal W	One-one	8167	26 551	26 551
Synthetic Exponential	51	mafft	Default	8066	29 345	29 345
Synthetic Exponential	51	mafft	One-one-fft-ns-i	5597	26 537	26 537
Synthetic Exponential	51	muscle	Default	6975	30 477	30 477
Synthetic Exponential	51	poy		9182	22 631	22 261
Synthetic Exponential	51	poy/tnt	poy default	9182	22 631	22 631
Synthetic Exponential	51	poy/tnt	poy aggressive	8696	22 258	22 258
Synthetic Uniform	51	Clustal Omega	Default	4928	48 110	48 110
Synthetic Uniform	51	Clustal W	Default	6233	39 531	39 531
Synthetic Uniform	51	Clustal W	One-one	10 403	33 832	33 832
Synthetic Uniform	51	mafft	Default	12 131	39 781	39 781
Synthetic Uniform	51	mafft	One-one-fft-ns-i	5932	35 978	35 978
Synthetic Uniform	51	muscle	Default	5663	59 214	59 214
Synthetic Uniform	51	poy		12 493	30 698	30 073
Synthetic Uniform	51	poy/tnt	poy default	12 493	30 695	30 695
Synthetic Uniform	51	poy/tnt	poy aggressive	11 626	30 070	30 070
Synthetic Exponential	201	Clustal Omega	Default	7260	205 485	205 469
Synthetic Exponential	201	Clustal W	Default	6705	161 848	161 848
Synthetic Exponential	201	Clustal W	One-one	18 420	139 669	139 662
Synthetic Exponential	201	mafft	Default	20 449	190 213	190 158
Synthetic Exponential	201	mafft	One-one-fft-ns-i	9249	149 846	149 818
Synthetic Exponential	201	muscle	Default	10 748	242 761	242 578
Synthetic Exponential	201	poy		39 379	120 728	119 219
Synthetic Exponential	201	poy/tnt	poy default	39 379	120 714	120 714
Synthetic Exponential	201	poy/tnt	poy aggressive	40 599	119 638	119 638
Synthetic Uniform	201	Clustal Omega	Default	8629	305 445	305 445
Synthetic Uniform	201	Clustal W	Default	8589	243 764	243 724
Synthetic Uniform	201	Clustal W	One-one	24 951	207 577	207 577
Synthetic Uniform	201	mafft	Default	29 779	311 846	311 846
Synthetic Uniform	201	mafft	One-one-fft-ns-i	11 243	239 952	239 890
Synthetic Uniform	201	muscle	Default	9123	351 721	351 536
Synthetic Uniform	201	poy		58 379	179 608	176 956
Synthetic Uniform	201	poy/tnt	poy default	58 379	179 589	179 589
Synthetic Uniform	201	poy/tnt	poy aggressive	60 969	176 937	176 937
Synthetic Exponential	601	Clustal Omega	Default	11 610	697 257	697 196
Synthetic Exponential	601	Clustal W	Default	9922	540 882	540 792
Synthetic Exponential	601	Clustal W	One-one	33 936	470 991	470 945
Synthetic Exponential	601	mafft	Default	27 241	628 005	627 958
Synthetic Exponential	601	mafft	One-one-fft-ns-2	27 680	518 881	518 857
Synthetic Exponential	601	muscle	Default	11 190	837 326	836 892
Synthetic Exponential	601	poy		123 869	401 566	396 841
Synthetic Exponential	601	poy/tnt	poy default	123 869	401 508	401 508
Synthetic Exponential	601	poy/tnt	poy aggressive	125 486	396 802	396 802
Synthetic Uniform	601	Clustal Omega	Default	10 149	751 667	751 587
Synthetic Uniform	601	Clustal W	Default	9775	581 240	581 200
Synthetic Uniform	601	Clustal W	One-one	35 821	509 921	509 885
Synthetic Uniform	601	mafft	Default	30 185	686 561	686 545
Synthetic Uniform	601	mafft	One-one-fft-ns-2	20 684	558 978	558 911
Synthetic Uniform	601	muscle	Default	9620	895 416	894 773
Synthetic Uniform	601	poy		128 617	434 310	427 354
Synthetic Uniform	601	poy/tnt	poy default	128 617	434 271	434 271
Synthetic Uniform	601	poy/tnt	poy aggressive	133 799	427 284	427 284

After every alignment was generated, we used a Python script to convert the resulting FASTA files to Hennig format. In addition, at this point, all gaps were coded as a fifth state, as mentioned above.

To measure the degree to which both search and alignment affect the eventual tree length, we ran TNT with two settings for each alignment. First, a relatively modest run with `mult`, then with much more

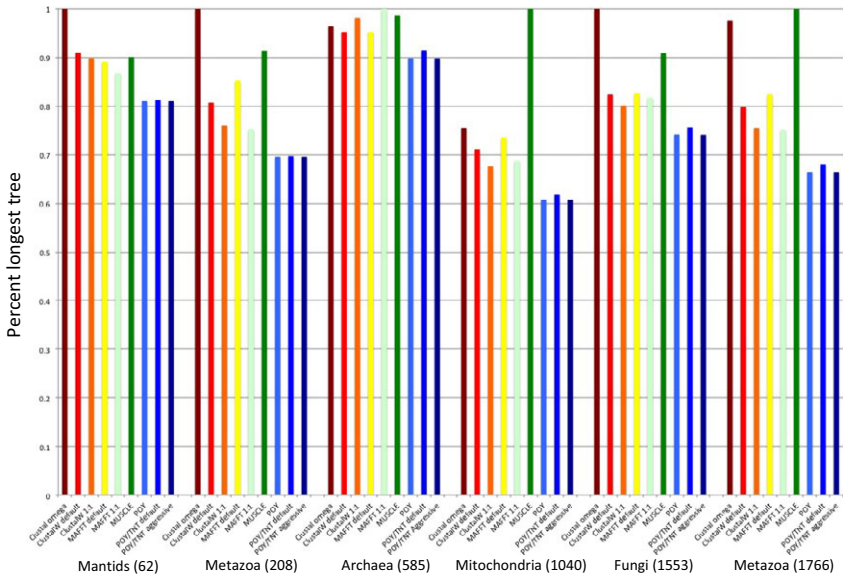

Fig. 1. Graphical depiction of tree lengths based on multiple sequence aligners + TNT and POY for biological sequence data sets. Bars are normalized to the method (for a given data set) that yielded the longest (i.e. least optimal) tree. Data sets, aligners and tree search options are described in the text.

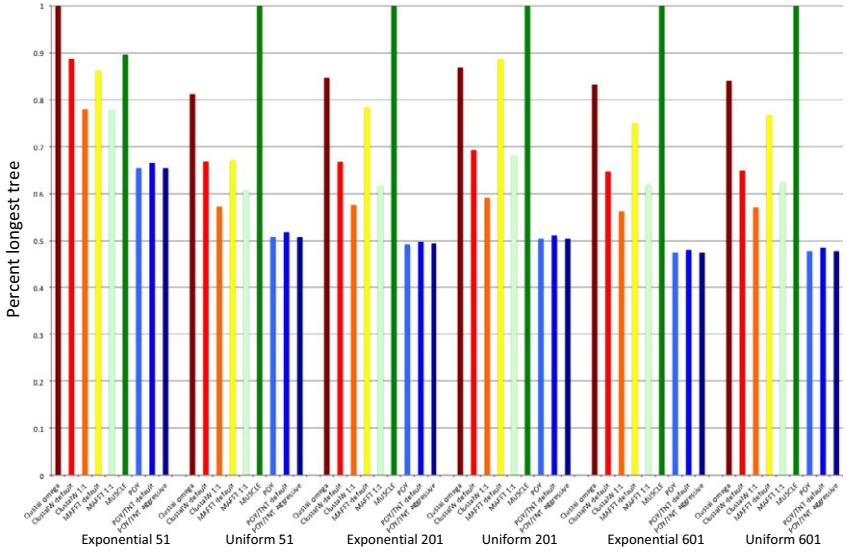

Fig. 2. Graphical depiction of tree lengths based on multiple sequence aligners + TNT and POY for DAWG simulated sequence data sets. Bars are normalized to the method (for a given data set) that yielded the longest (i.e. least optimal) tree. Data sets, aligners and tree search options are described in the text.

aggressive settings: `xmult = replications 10 ratchet 50 drift 20 fuse 5;`.

Results

For each data set, the 0 : 1 : 1 settings gave shorter trees in the subsequent tree search than the default alignment settings. On average, there was a 3.5% improvement in tree length for the 0 : 1 : 1 setting over default MSA values (Tables 2 and 3). In general, running TNT with more aggressive settings gave very little difference in tree length, 0.04% on average. In contrast, there was a more marked improvement, of 1.59% on average, between Wagner build only and more aggressive runs of POY.

TNT was able, in every case, to find a shorter tree than POY, if given POY's implied alignments as input.

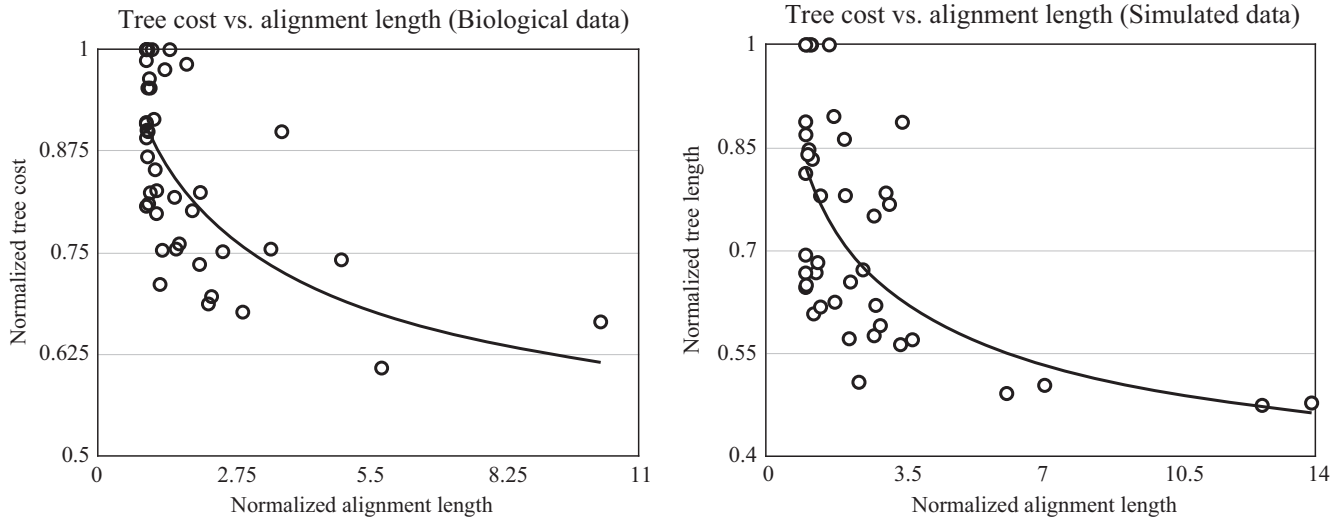

Fig. 3. Relationship between tree cost (normalized by the longest tree) and number of alignment columns for biological (left) and simulated (right) data. Power fit regression lines are shown.

The difference was 0.03% on average. This was due to the more aggressive search options specified for TNT (TBR, simulated annealing and genetic algorithm) as opposed to POY (TBR only). When more aggressive search options were specified in POY (using `transform(static_approx)`) the tree lengths of POY and TNT converged.

The most striking difference was in tree lengths for DO versus two-step alignment. Comparing the aggressive TNT search (based on the MSA aligners) with the aggressive POY search, POY gave 14.78% shorter trees than the two-step alignment. But because the difference in results was so minimal between MSA+ TNT's more- and less-aggressive runs, even the simple Wagner build DO runs were 13.53% better than the more aggressive MSA+ TNT results.

We did not find that one aligner was universally superior to the others. For seven of the data sets, CLUSTAL W (0 : 1 : 1) gave the best results, and for five sets MAFFT (0 : 1 : 1) gave the best results. Likewise for the worst performance: MUSCLE gave the longest tree for seven data sets, CLUSTAL OMEGA for four data sets and MAFFT for one. CLUSTAL OMEGA is specifically designed for use on protein sequences, and gives warnings when presented with DNA data, which might explain its poor showing. These results are summarized in graphical form in Figs 1 and 2.

In addition to tree cost, we also examined variation in alignment length (in terms of aligned "columns") and tree topology [in terms of Robinson and Foulds (1981) (R-F) distance]. For the comparison of alignment length, the number of columns of each alignment was normalized via division by the length of the shortest (i.e. fewest columns) alignment. The linear regression fits for the biological and simulated data were quite similar ($y = -0.0371x + 0.9196$, $R^2 = 0.3174$; and $y = -0.0323x + 0.8146$, $R^2 = 0.2978$, respectively). Power regressions are shown in Fig. 3 (regression fits $y = -0.9058x^{-0.1667}$, $R^2 = 0.4691$; and $y = -0.8222x^{-0.2179}$, $R^2 = 0.4345$). The longest alignments (with lowest costs) are the implied alignments generated by POY, but even apart from this, the highest cost trees were derived from the shortest alignments. Given that indel or "gap" costs were included in the tree costs, this might be regarded as counterintuitive. After all, longer alignments have greater numbers of gaps and could be expected to exhibit higher costs. Clearly, this is not the case, as shorter alignments imply higher numbers of substitutions as well as overall transformations—shorter is not better.

One question that has been raised about tree topology results derived from alternativee methods of analysis is whether it makes a difference. The alignment (and DO) methods might yield alternative homology schemes, but the tree topologies they imply might well be very similar. The differences in tree cost between the various MSA and POY analyses were compared with the (R-F) distances between the MSA analyses and POY (Fig. 4). A normalized R-F distance was used

$$\left(d_{i,j} = \frac{R - F_{ij} + R - F_{ji}}{n\text{Splits}_i + n\text{Splits}_j} \right)$$

to remove the effects of variation in degrees of tree resolution and number of terminal taxa. There is clearly a direct relationship between tree cost difference and R-F distance. If the alignments had shown variation, but the tree topologies not, the normalized R-F values would have been consistently near zero. The linear regression fits of normalized R-F to

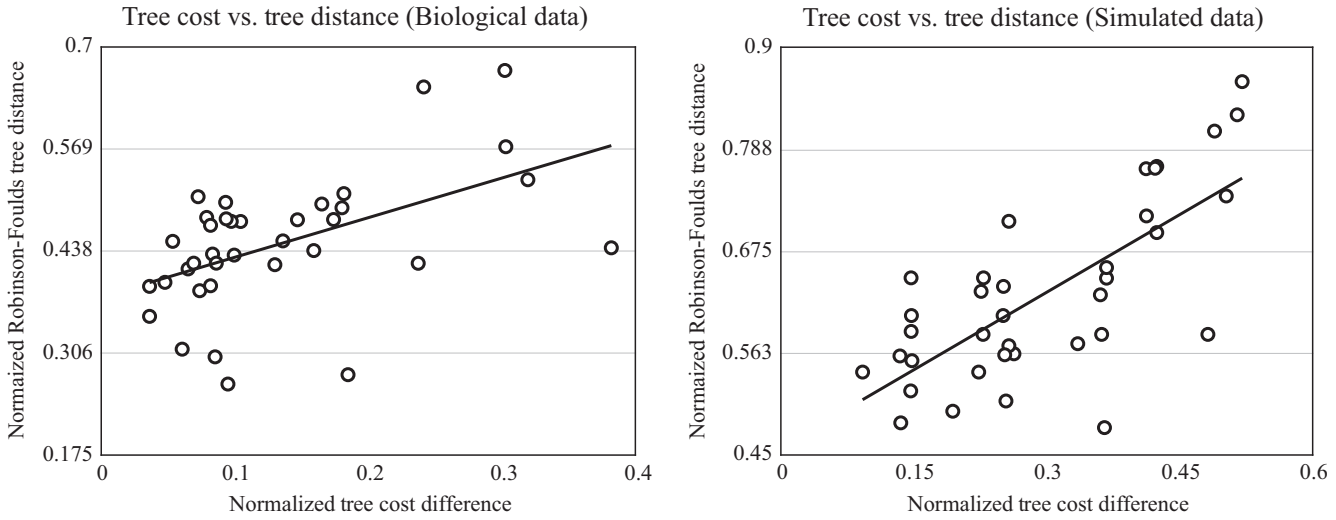

Fig. 4. Relationship of tree cost difference (normalized by the longest tree) between MSA+TNT- and POY-derived trees and R-F tree distance (between MAS+TNT- and POY-derived trees) for biological (left) and simulated (right) data. Linear fit regression lines are shown.

normalized tree cost difference for biological and simulated data were $y = 0.5094x + 0.3786$, $R^2 = 0.2639$; and $y = 0.5727x + 0.4581$, $R^2 = 0.5215$, respectively. As with the alignment length comparisons, the fits for the biological and simulated data were quite similar, with lower levels of scatter for the simulated data.

Discussion

As can be seen readily in the tables and figures, the algorithmic approach taken by DO is consistently and, at times, dramatically more effective at producing optimal trees than that embodied in the separated steps of MSA and tree search. This is true even in the face of "longer" alignments with greater numbers of columns replete with gaps. Perhaps surprisingly, the deficiencies created during the alignment phase cannot be overcome even through the heroics of an implementation as effective as TNT. Furthermore, even using implied alignments (provided by POY) based on ten simple $O(n^2)$ Wagner builds with $O(n^3)$ TBR refinement (of n taxa), subject to all measures of advanced tree search heuristics (tree-fusing, drifting, ratcheting, etc.; Moilanen, 1999; Nixon, 1999; Goloboff, 1999), produces solutions that can be improved only slightly (on the order of 1 part in 10 000). Presumably, applying such higher order heuristics to the GTAP directly (such facilities exist in POY) would generate a similar improvement.

The results found here are consistent with the analyses of Lehtonen (2008) and Wheeler and Giribet (2009) based on the over 5000 simulations of Ogden and Rosenberg (2007). In those comparisons, DO, in every case, yielded superior (in terms of tree length) solutions. Similar improvements have been found for likelihood-based analysis (Whiting et al., 2006).

One component of the relative success of the one-step DO approach over the two step MSA+TR no doubt comes from the number of alignment/sequence homology scenarios examined. In the case of the multiple sequence aligners examined here, a small number of guide trees (usually one) are used to create alignments. Although DO (as implemented in POY) does not use multiple alignments *per se* (implied alignments can only be produced *after* analysis), it does examine larger numbers of sequence homology scenarios. In essence, each tree examined during the search process implies a unique alignment set. A heuristic GTAP search using DO with Wagner builds and TBR refinement will examine $O(n^3)$ homology schemes. For the larger data sets examined here, this easily extends into the billions.

The examination of such a large number of homology scenarios is not without cost, however. The basic time complexity of MSA+TR for n sequences of length m would be on the order of (assuming a single pairwise distance-based guide tree and TBR-based tree search) $O(m^2n^2 + mn^3)$. A one-step DO (at least as implemented in POY) would contain another multiplicative factor of m, $O(m^2n^2 + m^2n^3)$, for Wagner build followed by TBR refinement. The dominant cubic terms would differ by a factor of m, the sequence length, for similar levels of heuristic intensity (as anyone comparing wall-clock times of POY and TNT can attest). However, the optimality efficiency is so much greater, that even $O(n^2)$ DO searches outperform MSA+TR by large factors.

Conclusion

In every case, DO found a shorter tree than the two-step process of alignment and search. In addition,

as TNT's rather more aggressive search strategy was insufficient to find significant improvement in the tree lengths (based on POY implied alignments), it is clear that the majority of the optimality deficit in a two-step search is a result of the alignment stage, which conditions the subsequent search step. These cost improvements are accompanied by changes in tree topology, and hence have phylogenetic significance. Most phylogenetic researchers expend a great deal of computational effort, justifiably, in the tree search stage of analysis. This is often based, however, on MSA procedures that explore little of the tree-alignment space, which is clearly the most important stage in identifying heuristically optimal GTAP solutions. A more efficient use of resources, then, would be to spend more effort on the initial, sequence homology stage or, better, to abandon the approach altogether in favour of more efficient single-step DO approaches.

Acknowledgements

We thank Louise M. Crowley, Prashant Sharma and Katherine St. John for discussion of these ideas and suggesting numerous improvements to the manuscript. We also thank Pablo Goloboff and an anonymous reviewer for their comments and suggestions for additional analyses that enhanced the content and presentation of the work. This material is based upon work supported by, or in part by, the National Science Foundation (BCS-0925978), the US Army Research Laboratory and the US Army Research Office under contract/grant number W911NF- 05-1-0271.

References

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2013. Genbank. Nucleic Acids Res. 41, D36–D42.

Cartwright, R.A., 2005. DNA assembly with gaps (DAWG): simulating sequence evolution. Bioinformatics 21, iii31–iii38.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

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. Phyl. Evol. 10, 1–5.

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

Goloboff, P., 1999. Analyzing large data sets in reasonable times: solutions for composite optima. Cladistics 15, 415–428.

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

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT version 5.25: multiple sequence alignment program. Nucleic Acids Res. 30, 3059–3066.

Kimura, M., Ohta, T., 1972. On the stochastic model for estimation of mutational distance between homologous proteins. J. Mol. Evol. 2, 87–90.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., R, R.L., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948.

Lehtonen, S., 2008. Phylogeny estimation and alignment via POY versus Clustal–PAUP: a response to Ogden and Rosenberg (2007). Syst. Biol. 57, 653–657.

Lindgren, A.R., Daly, M., 2007. The impact of length-variable data and alignment criterion on the phylogeny of Decapodiformes (Mollusca: Cephalopoda). Cladistics 23, 464–476.

Liu, K., Nelesen, S., Raghavan, S., Linder, C.R., Warnow, T., 2009. Barking up the wrong treelength: the impact of gap penalty on alignment and tree accuracy. IEEE Trans. Comput. Biol. Bioinf. 6, 7–20.

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

Nixon, K.C., 1999. The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics 15, 407–414.

Ogden, T.H., Rosenberg, M.S., 2007. Alignment and topological accuracy of the direct optimization approach via POY and traditional phylogenetics via ClustalW + PAUP*. Syst. Biol. 56, 182–193.

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

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Sankoff, D.M., 1975. Minimal mutation trees of sequences. SIAM J. Appl. Math. 28, 35–42.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539.

Svenson, G.J., Whiting, M.F., 2004. Phylogeny of Mantodea based on molecular data: evolution of a charismatic predator. Syst. Entomol. 29, 359–370.

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

Varón, A., Wheeler, W.C., 2013. Local search for the generalized tree alignment problem. BMC Bioinformatics 14, 66.

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

Wheeler, W.C., 2003. Implied alignment. Cladistics 19, 261–268.

Wheeler, W.C. 2007. The analysis of molecular sequences in large data sets: where should we put our effort? In: Hodkinson, T.R., Parnell, J.A.N. (Eds.), Reconstructing the Tree of Life: Taxonomy and Systematics of Species Rich Taxa. Systematics Association, Oxford University Press, Oxford, pp. 113–128.

Wheeler, W.C., Giribet, G. 2009. Phylogenetic hypotheses and the utility of multiple sequence alignment. In: Rosenberg, M.S. (Ed.), Perspectives on Biological Sequence Alignment. University of California Press, Berkeley, CA, pp. 95–104.

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

Wheeler, W.C., Lucaroni, N., Hong, L., Crowley, L.M., Varón, A., 2015. POY version 5: phylogenetic analysis using dynamic homologies under multiple optimality criteria. Cladistics 31, 189–196.

Whiting, A.S. Jr, Sites, J.W., Pellegrino, K.C., Rodrigues, M.T., 2006. Comparing alignment methods for inferring the history of the new world lizard genus *Mabuya* (Squamata: Scincidae). Mol. Phyl. Evol. 38, 719–730.