Сборник трудов Зоологического музея МГУ им. М.В. Ломоносова
Archives of Zoological Museum of Lomonosov Moscow State University
Том / Vol. 54 Стр. / Pp. 99–115

COMPUTATIONAL ASPECTS OF THE PHYLOGENETIC ANALYSIS OF COMPARATIVE SEQUENCE DATA

Ward C. Wheeler

Division of Invertebrate Zoology, American Museum of Natural History; *wheeler@amnh.org*

The construction of phylogenetic trees based on sequence data presents interesting computational challenges usually not encountered in the analysis of anatomical or other qualitative data. In addition to the more familiar problems inherent in tree searching, comparative sequence data require the additional step of alignment. Each of these procedures is computationally “hard”. When these two operations are coupled, the joint activity is referred to as the General Tree Alignment Problem (GTAP), perhaps the most appropriate form of analysis for comparative sequence data. Since this optimization is also NP-hard, heuristic techniques have been brought to bear to allow researchers to identify useful solutions to empirical problems. Various exact and heuristic approaches are discussed here with reference to their computational burden. Advances in sequencing technology are greatly increasing the quantity of sequence data requiring analysis. Increasing time complexity in this light of new data streams is discussed.

ВЫЧИСЛИТЕЛЬНЫЕ АСПЕКТЫ ФИЛОГЕНЕТИЧЕСКОГО АНАЛИЗА СРАВНИТЕЛЬНЫХ ДАННЫХ ПО СИКВЕНСАМ

Уорд К. Уилер

Построения филогенетических деревьев, основанные на сиквенсах, представляют интересные вычислительные проблемы, которые, как правило, не встречаются при анализе анатомических или других качественных данных. Сравнительные данные по сиквенсам, в дополнение к стандартным проблемам поиска деревьев, требуют специфического этапа выравнивания. Каждая из этих процедур является вычислительно «трудной». Объединение этих двух операций известно как General Tree Alignment Problem (GTAP): возможно, это наиболее подходящая форма анализа сравнительных данных по сиквенсам. Так как эта оптимизация также является NP-трудной, в неё задействованы эвристические методы, позволяющие выявлять практические

решения эмпирических проблем. В статье обсуждаются различные точные и эвристические подходы с указанием их вычислительных нагрузок. Достижения в области технологии секвенирования резко увеличивают количество сиквенсов, включаемых в анализ. В этом контексте обсуждается увеличение времени вычислений при включении новых потоков данных.

1. Introduction

As evolutionary biologists, we seek to accommodate and explain the broadest possible sample of comparative information including both phenotypic and genomic information. The construction of phylogenetic trees based on sequence data presents interesting computational challenges usually not encountered in the analysis of anatomical or other qualitative data. Although both types of information rely on tree-searching procedures, sequence data do not present themselves with pre-ordained correspondences. This adds to the process an additional computational challenge usually referred to as alignment.

Unfortunately, as is well-known, both of these operations (tree-searching and alignment) are NP-hard optimizations (Foulds, Graham, 1982; Day, 1987; Wang, Jiang, 1994; Roch, 2006), hence exact solutions for non-trivial data sets will not only be unfindable (at least with guarantee), but potentially exponential in number. This creates challenges in the definition of time-efficient and effective heuristic procedures. An additional factor lies within the specification of types of analysis based on different types of alignment, tree-searching and their interaction.

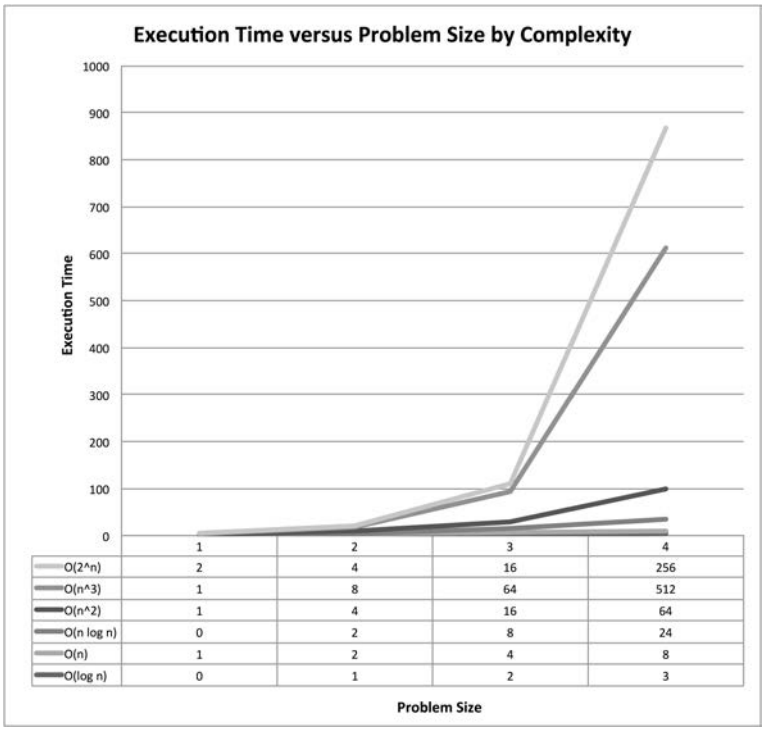
Here, I will discuss alternate methods of multiple sequence alignment (MSA) and their interaction with tree searching procedures. Several existing tools and oncoming empirical challenges will also be presented. The overall objective being to identify effective (in terms of optimality score) and efficient (in terms of time effort) analytical procedures as we encounter mounting availability of sequence data.

2. Types of alignment

The core operation of all multiple alignment procedures (MSA) is pairwise alignment. Whether a pairwise alignment is to minimize dissimilarity, maximize similarity, and whatever method used to score the quality of the alignment (e.g. edit costs, probabilistic models), the time-complexity of the operation is proportional to the product of the lengths of the two input sequences ($O(n_1 n_2)$) or more simply the square of the length of the longest sequence in an analysis ($O(n^2)$). This string-match procedure is often referred to as the Needleman–Wusch (Needleman, Wunsch, 1970) algorithm and is based on dynamic programming. The quadratic complexity means that if the sequences are doubled in length, the operation will take four times as long and so forth (reviewed in Wheeler, 2012).

This level of time complexity is actually more than just a worst-case scenario. Given that biological sequences are highly related and non-random (at least in the phylogenetic case), the algorithm of Ukkonen (1985) (or variants included in implementations such as Wheeler et al., 2013, 2015) can be used and time-complexity reduced to an average case of $O(n\log n)$ (Fig. 1). Storage (memory requirements) can also follow the time complexity.

The Needleman–Wunsch and Ukkonen procedures (as originally published) only align pairs of sequences. For phylogeneti-



	1	2	3	4
O(2^n)	2	4	16	256
O(n^3)	1	8	64	512
O(n^2)	1	4	16	64
O(n log n)	0	2	8	24
O(n)	1	2	4	8
O(log n)	0	1	2	3

Fig. 1. The increase in execution time (by problem size) for several complexities for logarithmic $O(\log n)$, $O(n\log n)$, linear $O(n)$, quadratic $O(n^2)$, cubic $O(n^3)$, and exponential $O(2^n)$.

cally meaningful data sets, alignments of large numbers of sequences are required. Pairwise algorithms can be naively extended beyond two, but rapidly become unmanageable. For three sequences a cube with n^3 elements would be required, and for each element, 7 evaluations would have to take place as opposed to the 3 for pairwise. In the general case for m sequences of length n, n^m elements would need to be stored and $(2^m - 1)$ operations required at each element, for a time complexity of $O((2^m - 1)n^m)$. Even the smallest data sets would be beyond our analytical capabilities. This issue has lead to the development of lower time and space complexity "progressive" alignment procedures (Feng, Doolittle, 1987). Progressive alignment, in essence, breaks down the exponential (in the number of sequences) to a series of pairwise alignments performed in an order determined in a variety of ways, most usually by a "guide-tree" (Fig. 2). A guide tree is not a phylogenetic tree, but simply a way to order pairwise alignments and their amalgamation into a single MSA. Progressive alignment reduce the exponential time complexity down to one linear in the number of sequences and quadratic in their length, $O((m - 1)n^2)$.

This reduction in complexity is welcome and makes many large data sets tractable. It does not, however, solve the entire problem. A key question remains of what determines a "good" alignment versus a "bad" one, or more precisely, how do we attach an optimality value (or score) to an MSA.

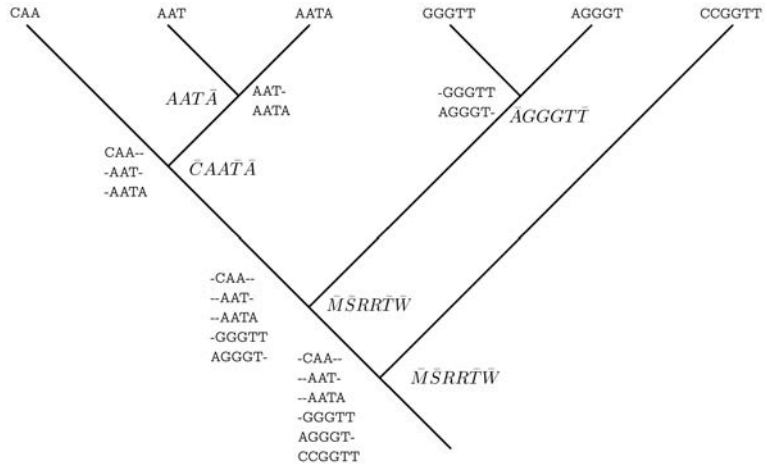


Fig. 2. Progressive alignment of Feng, Doolittle (1987). Pairwise alignments are performed in post-order tree traversal to create partial MSAs following once a gap, always a gap (left of vertices) or profile sequences (right of vertices) of IUPAC symbols for each column with a bar if a gap is also present in that aligned column. Redrawn after Wheeler (2012).

For pairwise alignment, scoring is straight forward, the distance between to two sequences. We might do this in a variety of ways (see optimality criteria below), but once we have a distance between base pairs, or amino acids, we can apply this without fuss to a pair of sequences. However, once we move to three sequences, the process becomes more complicated. Two common approaches would be the sum of the (three) pairwise distances among the three aligned sequences. A second would be to sum the distance between each aligned sequence and a common or "centroidal" sequence that represents the center, or average of the three. The first approach is referred to as "Sum-of-Pairs", or SP alignment (Carrillo, Lipman, 1988) and the second "Consensus" alignment (Gusfield, 1997). Furthermore, with four sequence a third approach is possible, that of "Tree" alignment (Sankoff, 1975), where alignments are created such that an overall evolutionary tree cost (in terms of summed edge costs) is minimized. Each of these methods can result in different MSAs even though basic cost parameters (such as substitution and indel costs) are the same. An important question is which method is most appropriate, or best in a phylogenetic context.

3. Alignment in the context of phylogeny

Once we have dispensed with the historical atrocity of "by-eye" alignment (e. g. Kjer, 2004), we are left with three common approaches to automated MSA generation when we analyze comparative sequence data — SP, Consensus, and Tree. There are two general approaches to discussing this choice — those based on first principles of historical homology and those based on efficacy and time complexity, often of implementations.

On the theoretical side, the argument has been made recently (Padial et al., 2014) and not so recently (e. g. Giribet et al., 2002; Faivovich et al., 2004; Prendini et al., 2005)

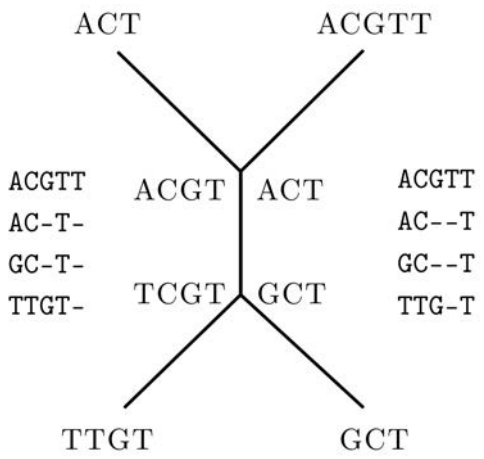


Fig. 3. "Tree" alignment showing median sequences (at internal vertices) and MSAs (multiple) on left and right. Note that there are two equally costly vertex assignments and derived alignments for this simple case. Redrawn after Wheeler (2012).

that since it is phylogenetic trees we are interested in, an alignment method based on trees is the only appropriate approach. Padial et al. (2014) in their forceful analysis also adduce empirical data as to the efficacy (in terms of tree optimality) of the tree alignment approach at least as embodied in the program POY (Wheeler et al., 2013, 2015). They explore and criticize what they have named "similarity alignment" (i. e. SP) on the grounds that similarity as a means of reconstructing phylogeny (as with UPGMA) has been long rejected in favor of homology (and synapomophy) based schemes. The homology implications of a potentially unique scheme for each possible phylogenetic scenario, termed "dynamic" homology (Wheeler, 2001), as opposed to the universal "static" homology statements produced by SP and consensus similarity alignments are argued to be vastly superior in explanatory power and more firmly rooted in the historical notions of derived homology (Fig. 3).

In addition to the theoretical and epistemological rationales for favoring tree alignment, heuristic efficacy is also an important factor. Since the original description of Direct Optimization (DO) (Wheeler, 1996; Varón, Wheeler, 2012) as an approach to the analysis of sequence data, many empirical analyses have shown substantial improvement in tree optimality (discussed more below) over other MSA methods. (e. g. Whiting et al., 2006; Lindgren, Daly, 2007; Giribet, Edgecombe, 2013). Ford and Wheeler (2016) in analyzing data sets from 62–1766 rDNA sequences (Giribet, Wheeler, 1999, 2001; Wheeler, 2007; Benson et al., 2013) as well as simulated data showed optimality improvements of up to 50% over SP and Consensus alignment implementations (Katoh et al., 2002a; Edgar, 2004; Larkin et al., 2007; Sievers et al., 2011). These analyses compared MSAs (implied alignments [Wheeler, 2003a] in the case of POY5) when subsequently analyzed by TNT (Goloboff et al., 2003) the preeminent parsimony tree search program. These results confirmed those of (Wheeler, Giribet, 2009) where in the case of over 5000 simulated data sets of Ogden and Rosenberg (2007) (also reanalyzed by Lehtonen, 2008), similarity MSAs *always* underperformed with respect to DO-based analysis.

To my knowledge, no analysis has ever been published where tree-alignment (at least as far as implemented by POY) has resulted in phylogenetic trees of inferior optimality value (i. e. parsimony score) to those based on any other MSA method.

4. Complexity of alignment

Commonly used MSA implementations — e. g. CLUSTALW (Higgins and Sharp, 1988), MAFFT (Katoh et al., 2002b, MUSCLE (Edgar, 2004) — are based around progressive alignment after pairwise align-

ments of input sequences, hence have a time complexity that is quadratic in both sequence length (n) and number (m), hence $O(m^2n^2)$. This may seem rather efficient, but the general MSA optimization problem itself, no matter whether SP, Consensus, or Tree is NP-hard (Wang, Jiang, 1994). As mentioned above, this implies both that optimal solutions (nor non-trivial data sets) will not be found, and that there are a potentially exponential numbers of such solutions. Basically, there is likely no identifiable, single "optimal" MSA (Wheeler, 1994).

Polynomial time approximations schemes (PTAS) have been developed for SP alignment with guaranteed bounds (Wang et al., 2000) employing a mix of exhaustive and lifted alignment (Wang, Gusfield, 1997; Wheeler, 1999). Mainly of analytical interest, a 1.47 bound can be ensured with a time complexity of $O(m^2n^9)$. This is clearly well beyond empirical utility (and for not such a great bound), but underscores the difficulty in determining high quality solutions.

The conditions of guarantee are rather loose, including all possible scenarios, and the reality of historical biological sequences is that they are much more "well-behaved" than the most general case would allow. There are no known guaranteed bounds for the DO algorithm (Wheeler, 1996), but performance in comparison with exact solutions in the three-sequence case has been examined (Varón, Wheeler, 2012). Depending on the rates of sequence (evolutionary) change, the DO algorithm has been shown to yield solutions within 3% to 10% of the optimal solution in biological realistic conditions. An $O(n^3)$ version of DO (Wheeler, 2003b) can perform better (exactly, unsurprisingly, in the three sequence case), but with an additional cost factor of the length of the sequences, n. When employed in empirical cases (with comparable time complexity tree searches i. e. m^2 to m^3 for overall complexity of $O(m^{2-3}n^2)$, this still rather loose bound has been shown to be effective and more than competitive with other heuristics (Ford, Wheeler, 2016).

5. Complexity of tree searching

For most empirical systematists, the search for optimal (or at least optimal enough) trees occupies the greatest portion of their computational effort. For this reason, algorithmic efficiency and quality of implementation have been extremely important to practicing systematists since the first phylogenetic (i. e. non cluster-based) tree reconstruction software was produced in the 1980's (Farris, 1978, 1988; Felsenstein, 1980; Mickevich, Farris, 1980; Swofford, 1990). Tree searching, as an operation, is not unique to sequence data, but its interaction with the dynamic homology concept yields a tight connection between the search for optimal homology schemes and optimal trees (General Tree Alignment Problem, GTAP, below).

The search for the optimal or "best" phylogenetic tree is well-known to be NP-hard. As for MSA, this implies that exact solutions for most datasets will be unavailable (time complexity $O(2^n)$) and potentially exponential in number. This is the case for all optimality criteria that have been examined including distances (Day, 1987), parsimony (Foulds, Graham, 1982), and likelihood (Roch, 2006). For smaller data sets (< 25 or so taxa), exact solutions can be found via Branch-and-Bound approaches (Land, Doig, 1960; Hendy, Penny, 1982). These methods rely on the examination of intermediate (partial) solutions and pruning large segments of the solution space. For this to perform well, the data need to be relatively clean and still may not yield much of an improved execution time — and may even be slower in

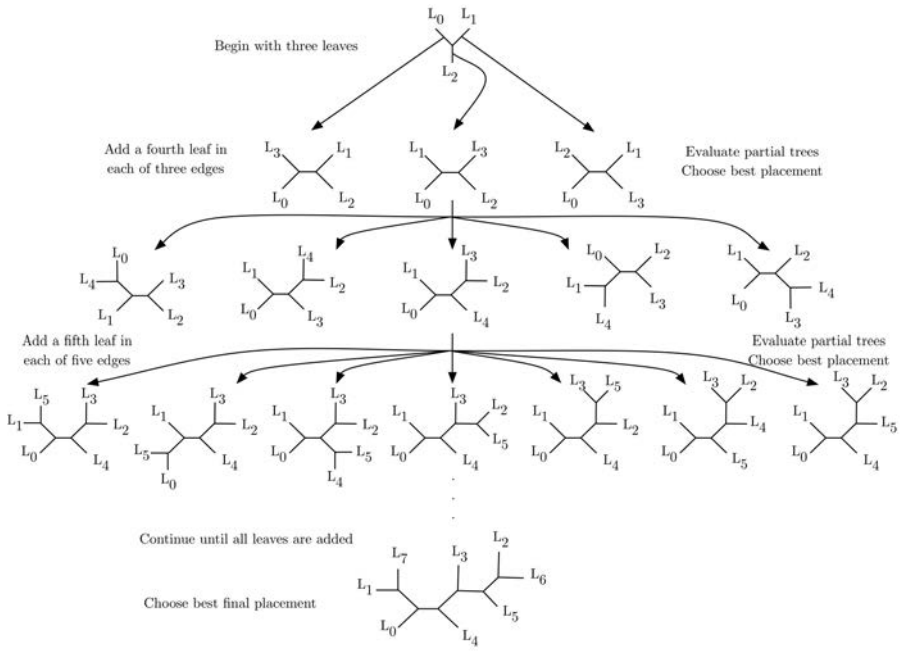


Fig. 4. Initial tree construction trajectory via the Wagner algorithm (Farris, 1970).

pathological cases. Modern empirical data sets are almost exclusively of the size that such an approach is impractical.

Again as with MSA, heuristic approaches are the main tool used to identify optimal trees. These can be organized into two sorts of approaches that are most frequently used in combination: trajectory and perturbation. Trajectory-based searches identify local neighborhoods of solutions, choose the best of them and advance to another neighborhood until no better solutions are found. Perturbation techniques, on the other hand, take a given solution and attempt to improve it by modifying the tree in ways which may not be immediately better, but may result in superior solutions at a later stage.

The nearly universally employed initial trajectory algorithm is the "Wagner" algorithm, named after a botanist (Wagner, 1961) but conceived by a computational systematist (Farris, 1970) constructs trees via sequential addition of taxa to a growing tree. Initially, three taxa are chosen (either randomly or by distance), and each taxon added to the tree in each possible place. The optimality value is calculated for each of these candidate solutions and the tree with the best value chosen. This is continued until all taxa have been added to the tree (Fig. 4).

The time complexity of this operation is quadratic ($O(m^2)$), however, common practice is to perform a number of these, randomizing the addition order (this number can be thought to grow with the problem size adding an additional factor of m, see below).

These randomized Wagner build solutions are not, on their own, usually felt to be satisfactory and much improvement can be garnered by what is commonly referred to as branch-swapping, a form of tree refinement. There are two fundamental operations in branch-swapping: tree division and reattachment. In the first, the tree is divided

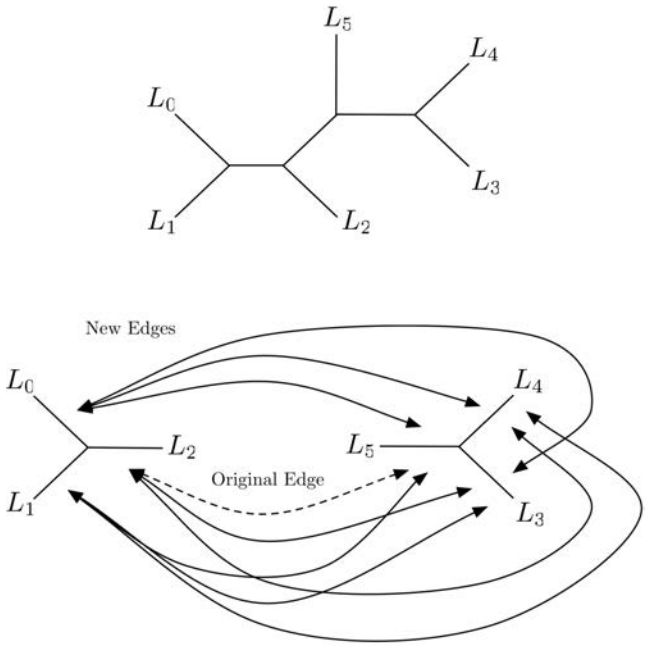


Fig. 5. Tree-Bisection and Regrafting (TBR) rearrangement neighborhood.

by removing an edge and in the second, reattached in a new location via the creation of a new edge. The named forms of swapping differ in the set of new edges that can be created. In each case, a neighborhood of new tree solutions is identified and evaluated. The size of this neighborhood drives the time complexity of the operation.

Nearest-Neighbor-Interchange (NNI; Camin, Sokal, 1965; Robinson, 1971) generates the smallest neighborhood with the lowest time complexity ($O(m)$). In NNI, as with all swapping procedures, each internal edge is examined in turn. All edges are defined by two vertices, each of which is connected to two further edges. NNI deletes one of these connecting edges and creates two new trees by creating edges between the now unconnected vertex and the two remaining edges. Given that there are $(m - 3)$ edges in a tree with m taxa, the total number of tree rearrangements examined is $2(m - 3)$, hence the linear time complexity (assuming tree optimality can be determined in constant time — a fallacy we will hold to for now).

Larger neighborhoods can be generated through the commonly referred to Subtree-Pruning-and-Regrafting (SPR) and Tree-Bisection-and-Regrafting (TBR) algorithms (initially undocumented and known under various names; Mickevich, Farris, 1980; Swofford, 1990; see Wheeler, 2012). SPR refinement involves breaking an edge and reattaching the unconnected vertex via new edges to each remaining edge (as opposed to only the proximate two of NNI). There are then $(2m - 7)$ reattachments for a total of $2(m - 3)(2m - 7)$ (Allen, Steel, 2001). Hence, SPR has quadratic time complexity in the number of taxa. TBR yields an additional factor of m by allowing edge connection not only to the root vertex of the disconnected tree, but to its internal edges as well. The exact neighborhood size depends

on the tree shape, but is cubic in the number of taxa (Fig. 5).

A typical trajectory search would involve a series of randomized Wagner builds followed by TBR branch-swapping refinement. This is often referred to as RAS+TBR (Goloboff, 1999) and given that the number of random additions usually scales with the number of terminals the overall strategy can be thought of as quartic, $O(m^4)$.

This rather daunting polynomial factor led Goloboff (1999) to search for methods with reduced time complexity based on the notion of "composite optima". The RAS+TBR strategy just scales too poorly for large data sets over the entire tree. Yet, it may well arrange smaller components of thee tree properly, at least for some subset of RAS+TBR runs. Goloboff reasoned that segments of the data set (perhaps 50 taxa or so; later named "sectors") might well be in optimal or near optimal configuration, but the odds of getting a large number of sectors simultaneously well configured would be very small, hence requiring prohibitive RAS+TBR iterations. By combining this notion of sectors with breadth-first searching (Cormen et al., 2001), Goloboff (1999) proposed Sectorial-Searching (SS). In this branch-swapping refinement operation, the edge set available for initial deletion and reattachment is limited to those not in sectors (commonly, but not invariantly subtrees). The time complexity of this operation is still cubic, but with a reduction in constant factor of the cube of the number of sectors, k. If these sectors are roughly equal in size, time complexity can be reduced dramatically to $O((n/k)^3)$. Disc-Covering methods (Huson et al., 1999; Roshan et al., 2004) have many similarities with sectorial methods, but have several important differences in the way they define and resolve intermediate subproblems. This has been shown to result generally inferior performance (Goloboff, Pol, 2007).

In apposition to trajectory heuristics, perturbation approaches accept immediately poorer solutions in hope of finding better ones further on down the search path. The motivation behind accepting sub-optimal solutions is that the optimality landscape has multiple "peaks" and that solutions between them are suboptimal. Perturbation methods are specifically designed to go from local to hopefully global solutions by traversing solution areas with reduced quality. The first use of this idea as an improvement on trajectory search came from Nixon (1999) in the form of the "parsimony ratchet." The ratchet is a technique that begins with a local (i. e. trajectory) solution and strives to break through intermediate optimality barriers by reweighting subsets of characters and searching (via TBR) on the new (but related) data set. The reweighting scheme shifts the optimality landscape such that new optima may be reached via standard swapping, A second search, again typically via TBR, is then performed beginning with the reweighted result, but with weights returned to their original values. This is performed multiple times, each instance offering an opportunity to find a path from a local to a (more) global solution. The ratchet had an immediately salutary affect on analyses, first on the "Zilla" RBCL dataset (Chase et al., 1993; Nixon, 1999) and later on others (Giribet, Wheeler, 1999). The ratchet approach has had an enormous effect on tree searching not only by directly improving search results, but also by opening up thinking that lead to a number of other innovative approaches.

Following on the heals of Nixon's ratchet, other perturbation techniques were developed, more directly adapted from simulated annealing. First developed for atomic bomb calculation in the 1950's (Metropolis et al.,

1953; Kirkpatrick et al., 1983), simulated annealing mimics the process of the annealing of metals via stepped reductions in temperature. The key concepts are the analogues to temperature and energy state. When temperatures are relatively high, the probability of accepting a lower quality solution (higher energy, in this case inferior optimality score) increases, allowing the search to find global solutions through optimality valleys. Better (lower energy, superior optimality) solutions are always accepted. If the temperature is high enough, the search becomes a random walk with no influence of optimality score at all. As the temperature is gradually reduced, the probability of transitions to lower optimality is increasingly diminished, until they are forbidden entirely and the search becomes a standard trajectory. The trick to using the technique effectively is identifying the proper connections between the physical model of annealing and the optimization at hand, and the appropriate heating schedule. Goloboff's "Tree-Drifting" (Goloboff, 1999) uses elements of such a simulated annealing approach. In his method, the difference in tree scores between candidate trees is the analogue of energy difference with random factor moderating acceptance of sub-optimal tree solutions.

A method employing elements of both trajectory and perturbation actions, and employing populations of trees is referred to as Genetic Algorithm, or GA. The idea behind this optimization technique is to simulate the evolutionary process with mutation, recombination, and selection (Holland, 1975). As applied to tree searching (Moilanen, 1999, 2001), GA starts with a set of initial solutions (such as those form RAS+TBR) and mutates these via some type of perturbation. The pool of trees then undergoes a selection step, where the optimality value for the trees determines their continued presence in the population. Those that survive selection then undergo recombination-corresponding components of trees (subtrees with the same leaf set) are exchanged. The order of these operations may vary, but these core operations are always components of GA optimization. Goloboff (1999) emphasized the recombination step in his "Tree-Fusing" method, which also adds in trajectory searches to improve solutions. At least in phylogenetic applications (including MSA; Notredame, Higgins, 1996), GA has not been shown to be very effective in generating solutions *de novo*, but has been extremely useful in improving existing solutions.

Each of the methods discussed above has found utility in empirical tree searching, and are nearly always used in combination. Implementations such as TNT (Goloboff et al., 2003; Goloboff, Pol, 2007) and POY (Wheeler et al., 2013, 2015) are explicit in their efforts to make these techniques available. Large analyses of up to thousands of unaligned (Ford, Wheeler, 2016) and tens of thousands of aligned (Goloboff et al., 2009) sequences demonstrate their combined effectiveness.

6. The GTAP and heuristic efficiency

The previous sections have described alignment and tree search as two separate operations. That is, in fact, how most phylogenetic analyses are performed, and how most systematists understand the problem of deriving phylogenetic trees from sequence data. This is, however, a limited and largely incorrect notion of the basic challenge. This does not mean that separate (i. e. 2-step) analyses do not have heuristic utility, merely that as a conceptualization and problem definition (with concomitant solution space and complexity) it is an erroneous path.

As discussed above, tree-alignment is the proper form of MSA in a phylogenetic

context, hence, the alignment be constructed to minimize the optimality score (originally parsimony) for a given tree. Known as the TAP (for Tree Alignment Problem), it is the "small" parsimony problem for a tree of unaligned sequence data and is NP-hard. If this tree is unknown, as is usually the case, then a tree search must also be performed with each tree evaluated in turn based on its tree alignment cost. This is the "large" parsimony problem for unaligned sequences and is referred to as the General Tree Alignment Problem or GTAP. Wrapping one NP-hard optimization inside of another makes this an exceptionally difficult, but empirically important, challenge.

Distinct alignment and tree searching operations can be thought of as one sort of GTAP heuristic. From the alignment perspective alone, there were only two efforts (of which I am aware) to construct alignments specifically with the objective of yielding optimal trees MALIGN (Wheeler, Gladstein, 1994, 1998) and TreeAlign (Hein, 1989a,b, 1990). These implementations would produce MSA results (one or more MSAs) that would then be fed into tree search programs to complete phylogenetic analysis. Although working towards tree-alignment, as with other MSA efforts the production of a single (or small number) of alignments used as a basis for the evaluations of a large number of trees (easily $> 10^9$) allows for a certain efficiency, but at a significant penalty in result quality.

The fundamental idea of the GTAP is that each tree, in essence, needs to be evaluated on the basis of its own alignment. That is, the alignment optimal (at least heuristically) for that tree. As tree space is searched, a potentially unique MSA would need to be generated for each candidate tree. Currently the most efficient heuristic procedure for this operation is Direct Optimization (DO; Wheeler, 1996; Varón, Wheeler, 2012) implemented in successive versions of POY (Gladstein, Wheeler, 1997; Wheeler et al., 2005; Varón et al., 2008; Wheeler et al., 2013). DO creates a series of vertex sequences via tree traversal in acceptable time with worst case complexity $O(mn^2)$, but average case complexity on the order of $O(mn\log n)$. This has allowed analyses of sequences of length 2000 or more for more than 1000 sequences with unmatched optimality scores (Varón, Wheeler, 2013; Ford, Wheeler, 2016). Higher time complexity flavors of DO have been defined: e. g. "Iterative-Pass" optimization (Wheeler, 2003b) with the much greater time complexity of $O(n^3)$.

While DO constructs internal vertex sequences, other GTAP approaches have been based on using observed sequences as candidate vertex labels. These include "lifted" alignments (Wang et al., 1996; Wang, Gusfield, 1997), and the stronger "fixed-states" (Wheeler, 1999) and "Search-Based" optimization (Wheeler, 2003c). Each of these has complexities that are cubic in the number of taxa $O(m^3)$ after a quadratic setup phase $O(m^2n^2)$. In general, these methods do not yield results competitive with DO, but do have guaranteed bounds, hence are extremely useful for analytical purposes.

Direct GTAP solution attempts will likely always have a higher time complexity than tree searching on pre-aligned sequences, currently by a factor of n (all over factors being equal, which they are not). But two factors have to be noted on this front, first is the time expended in alignment part of the two-step approach, and second the quality of the resultant solution.

7. Optimality criteria and complexity

This discussion has so far been rooted in minimizing overall evolutionary score — parsimony. Each of the techniques discussed above (alignment, tree search, direct GTAP

optimization) can and often is accomplished within the framework of other optimality criteria, namely maximum likelihood and Bayesian posterior probability.

Sequence alignment in a maximum likelihood context was first proposed by Bishop and Thompson (1986) and later in successive efforts by Thorne and coworkers (Thorne et al., 1991, 1992; Thorne, Kishino, 1992). A tree alignment approach was proposed by Wheeler (2006) (reviewed in: Denton, Wheeler, 2012) implemented in POY3 and POY5.

Tree alignment in the general context of Bayesian analysis has produced a number of implementations including Profile-HMM (Krogh et al., 1994), HANDEL (Holmes, Bruno, 2001), SATCHMO (Edgar, Sjölander, 2003), ProAlign (Löytynoja, Milinkovitch, 2003), ALIFRITZ (Fleissner et al., 2005), BEAST (Lunter et al., 2005), and BaliPhy (Redelings, Suchard, 2005; Suchard, Redelings, 2006). Wheeler (2014) proposed a version of MAP based (as opposed to MC^3) Bayesian tree alignment implemented in POY5.

The basic time complexity factors of statistical alignment, tree search and GTAP are still present as they are for parsimony, but usually have large additional constant factors resulting in vastly increased execution time. This is due to their need to evaluate statistical models of sequence evolution including a variety of parameters, most importantly branch/edge lengths. The most commonly used ML tree search tool in use today, RAxML (Stamatakis et al., 2005) as well as that for Bayesian, MrBayes (Huelsenbeck, Ronquist, 2003), consume many thousands of times more CPU cycles than that for parsimony, at least in the guise of TNT (Goloboff et al., 2003). Again, this is not sue to any shortcoming in implementation, but the reality of parameter-rich model-based tree searching.

The same factors, yet increased again, affect Bayesian GTAP methods as well. In addition to even more enhanced parameters and prior distributions involved in Bayesian calculations, the reliance on MC^3 methods adds enormous and problematic (in terms of stationarity requirements and their recognition) optimization factors (Mossel, Vigoda, 2005, 2006). In short, MSA, tree search, and GTAP can be approached in a variety of ways, and with a variety of tools, but all are burdened with the weight of NP-hard optimizations.

8. The coming deluge

In the past 10 years sequencing technology has rapidly advanced. This is largely due to the Human Genome Project and these technological fruits are now being reaped by systematic biology. Two important sources of comparative data are transcriptomic and whole genome sequencing. The first metazoan whole-genome phylogeny was only published in 2007 and then only for twelve species of very closely related *Drosophila* species (Clark et al., 2007). Even though whole genome costs are rapidly decreasing, faster even than Moore's Law, the large data sets produced now are based on the expressed portion of the genome, the transcriptome. Although the transcriptome represents less than 1% of the genome of most organisms, analyses based on these data have increase our available genomic information by three orders of magnitude in a short time. As an example, the metazoan analysis of Dunn et al. (2008) relied on 150 genes, followed the next year by Hejnöl et al. (2009) with nearly 1500. Riesgo et al. (2012) and Sharma et al. (2014) have produced data sets with greater than 20,000 genes. Concomitant with this, is an increase in the number of taxa from which these samples are derived. At present, transcriptomic analysis requires relatively freshly obtained specimens, and

this has limited the taxonomic expansion in some cases.

DNA-based techniques are now in development that will likely allow the determination of whole genomes from preserved museum specimens. This development will no doubt result in an explosion of taxonomic data that will threaten to overwhelm our current computational abilities. As we have seen, the general complexity of analysis of a genetic sequence is at worse quadratic in length, and linear in number of genes. This is a complexity we can handle at present. The taxon number factor is more daunting with a potentially quartic time complexity. Whole genomes will also provide a great deal more information than transcriptomes do presently and this will take effort to extract.

The challenges coming our way in the next few years are clear. We will have to identify computational methods that will produce, at least heuristically, optimal results. Furthermore, these methods will have to scale into the hundreds of thousands of terminals and potentially billions of base pairs. This will no doubt be assisted by the commodity parallelism now broadly available. Yet that will only generate a linear factor of improvement. To get to where we will need to be, only improved algorithms and potentially novel analytical approches offer a path.

Acknowledgements

Thanks to Igor Pavlinov for the opportunity to review this topic.

References

Allen B., Steel M. 2001. Subtree transfer operations and their induced metrics on evolutionary trees. — Annals of Combinatorics, 5 (1): 1–13.

Benson D.A., Cavanaugh M., Clark K. et al. 2013. Genbank. — Nucleic Acids Research, 41 (D1): D36–D42.

Bishop M.J., Thompson E. A. 1986. Maximum likelihood alignment of DNA sequences. — Journal of Molecular Bioliology, 190 (2): 159–165.

Camin J.H., Sokal, R.R. 1965. A method for deducing branching sequences in phylogeny. — Evolution, 19 (3): 311–326.

Carrillo H., Lipman D. 1988. The multiple sequence alignment problem in biology. — SIAM Journal on Applied Mathematics, 48 (5): 1073–1082.

Chase M.W., Soltis D.E., Olmstead R.G., Morgan et al. 1993. Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene RBCL. — Annals of the Missouri Botanical Garden, 80 (3): 528–580.

Clark A.G., Eisen M.B., Smith D.R. et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. — Nature, 450: 203–218.

Cormen T.H., Leiserson C.E., Rivest R.L., Stein C. 2001. Introduction to algorithms. Cambridge (MA): The MIT Press, 2nd edition. 1180 p.

Day W.H.E. 1987. Computational complexity of inferring phylogenies from dissimilarity matrices. — Bulletin of Mathematical Biology, 49 (4): 461–467.

Denton J., Wheeler W.C. 2012. Trivial alignments in maximum likelihood analysis of nucleotide data. — Cladistics, 28 (5): 514–528.

Dunn C.W., Hejnöl A., Matus D.Q. et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. — Nature, 452: 745–749.

Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. — Nucleic Acids Research, 32 (5): 1792–1797.

Edgar R.C., Sjölander K. 2003. SATCHMO: sequence alignment and tree construction using hidden Markov models. — Bioinformatics, 19 (11): 1404–1411.

Faivovich J., Garca P.C., Ananias F. et al. 2004. A molecular perspective on the phylogeny of the *Hyla pulchella* species group (Anura, Hylidae). — Molecular phylogenetics and evolution, 32 (3): 938–950.

Farris J.S. 1970. A method for computing Wagner trees. — Systematic Zoology, 19:83–92.

Farris J.S. 1978. Wag78. Software and documentation. Publ. by author.

Farris J.S. 1988. Hennig86. Version 1.5. Publ. by author.

Felsenstein J. 1980. PHYLIP Version 1.0: Phylogeny Inference Package. http://www0.nih.go.jp/~jun/research/phylip/main.html?ref=herseybedava.info

Feng D.-F., Doolittle R.F. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. — Journal of Molecular Evolution, 25 (4): 351–360.

Fleissner R., Metzler D., von Haeseler R. 2005. Simultaneous statistical multiple alignment and phylogeny reconstruction. — Systematic Biology, 54 (4): 548–561.

Ford E., Wheeler W. 2016. Comparison of heuristic approaches to the general-tree-alignment problem. — Cladistics, 32. in press. http://onlinelibrary.wiley.com/doi/10.1111/cla.12142/abstract?userIsAuthenticated=false&denied.

Foulds L.R., Graham R.L. 1982. The Steiner problem in phylogeny is NP-complete. — Advances in Applied Mathematics, 3 (1): 43–49.

Giribet G., Edgecombe G.D. 2013. Stable phylogenetic patterns in scutigeromorph centipedes (Myriapoda : Chilopoda : Scutigeromorpha): dating the diversification of an anci ent lineage of terrestrial arthropods. — Invertebrate Systematics, 27 (5): 485–501.

Giribet G., Wheeler W.C. 1999. The position of arthropods in the animal kingdom: Ecdysozoa, islands, trees and the “parsimony ratchet”. — Molecular Phylogenetics and Evolution, 13 (3): 619–623.

Giribet G., Wheeler W.C. 2001. Some unusual small-subunit ribosomal DNA sequences of metazoans. — American Museum Novitates, 3337: 1–14.

Giribet G., Wheeler W.C., Muona J. 2002. DNA multiple sequence alignments. — DeSalle R., Giribet G., Wheeler W.C. (eds). Molecular systematics and evolution: Theory and practice. Basel: Birkhauser Verlag. P. 107–114.

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

Goloboff P. 1999. Analyzing large data sets in reasonable times: solutions for composite optima. — Cladistics, 15 (4): 415–428.

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

Goloboff P.A., Catalano S.A., Mirande J.M. et al. 2009. Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups. — Cladisitcs, 25 (3): 211–230.

Goloboff P.A., Pol D. 2007. On divide-and-conquer strategies for parsimony analysis of large data sets: Rec-I-DCM3 versus TNT. — Systematic Biology, 56 (3): 485–495.

Gusfield D. 1997. Algorithms on strings, trees, and sequences: Computer science and computational biology. Cambridge: Cambridge Univ. Press. 534 p.

Hein J. 1989a. A new method that simultaneously aligns and reconstruct ancestral sequences for any number of homologous sequences, when the phylogeny is given. — Molecular Biology and Evolution, 6 (6): 649–668.

Hein J. 1989b. A tree reconstruction method that is economical in the number of pairwise comparisons used. — Molecular Biology and Evolution, 6 (6): 669–684.

Hein J. 1990. Unified approach to alignment and phylogenies. — Methods in Enzymology, 183: 626–645.

Hejnöl A., Obst M., Stamatakis A. et al. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. — Proceedings of the Royal Society, ser. B, Biological Sci., 276: 4261–4270.

Hendy M.D., Penny D. 1982. Branch and bound algorithms to determine minimal evolutionary trees. — Mathematical Biosciences, 60: 133–142.

Higgins D.G., Sharp P.M. 1988. Clustal: A package for performing multiple sequence alignment on a microcomputer. — Gene, 73 (1): 237–244.

Holland J.H. (ed.). 1975. Adaptation in natural and artificial systems. Ann Arbor (MI): University of Michigan Press. 211 p.

Holmes I., Bruno W.J. 2001. Evolutionary HMMs: a Bayesian approach to multiple alignment. — Bioinfomratics, 17 (9): 803–820.

Huelsenbeck J.P., Ronquist F. 2003. MrBayes: Bayesian inference of phylogeny, 3.1.2 edition. Program and documentation. at http://morphbank.uuse/mrbayes/.

Huson D., Nettles S., Warnow T. 1999. Disk-covering, a fast converging method for phylogenetic tree reconstruction. — Journal of Computational Biology, 6 (3): 368–386.

Katoh K., Misawa K., Kuma K., Miyata T. 2002a. MAFFT: a novel method for rapid mutliple sequence alignment based on fast fourier transform. — Nucleic Acids Research, 30 (14): 3059–3066.

Katoh K., Misawa K., Kuma K., Miyata T. 2002b. MAFFT version 5.25: multiple sequence alignment program. — Nucleic Acids Research, 30 (14): 3059–3066.

Kirkpatrick S., Gelatt C.D., Vecchi M.P. 1983. Optimization by simulated annealing. — Science, 220: 671–680.

Kjer K. M. 2004. Aligned 18S and insect phylogeny. — Systematic Biology, 53 (3): 506–514.

Krogh A., Brown M., Mian I.S. et al. 1994. Hidden Markov models in computational biology: applications to protein modeling. — Journal of Molecular Bioliology, 235 (5): 1501–1531.

Land A.H. and Doig, A.G. 1960. An automatic method of solving discrete programming problems. Econometrica, 28 (3): 497–520.

Larkin M.A., Blackshields G. et al. 2007. Clustal W and Clustal X version 2.0. — Bioinformatics, 23 (21): 2947–2948.

Lehtonen S. 2008. Phylogeny estimation and alignment via POY versus Clustal–PAUP: A response to Ogden and Rosenberg (2007). — Systematic Biology, 57 (4): 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 (4): 464–476.

Löytynoja A., Milinkovitch M. C. 2003. ProAlign, a probabilistic multiple alignment program. — Bioinformatics, 19 (11): 1505–1513.

Lunter G., Drummond A.J., Miklós I., Hein J. 2005. Statistical alignment: Recent progress, new applications, and challenges.— Nielsen R. (ed.). Statistical methods in molecular evolution. Springer. P. 375–406.

Metropolis N.A., Rosenbluth A., Rosenbluth M. et al. 1953. Equation of state calculations by fast computing machine. — J. Chem. Phys, 21 (6): 1087–1092.

Mickevich M.F., Farris J.S. 1980. PHYSYS: Phylogenetic analysis system. Publ. by authors.

Moilanen A. 1999. Searching for most parsimonious trees with simulated evolutionary optimization. — Cladistics, 15 (1): 39–50.

Moilanen A. 2001. Simulated evolutionary optimization and local search: Introduction and application to tree search. — Cladistics, 17 (1): S12–S25.

Mossel E., Vigoda E. 2005. Phylogenetic mcmc algorithms are misleading on mixtures of trees. — Science, 309: 2207–2209.

Mossel E., Vigoda E. 2006. Limitations of markov chain monte carlo algorithms for bayesian inference of phylogeny. — The Annals of Applied Probability, 16 (4): 2215–2234.

Needleman S.B., Wunsch C.D. 1970. A general method applicable to the search for similarities in the amino acid sequences of two proteins. — Journal of Molecular Bioliology, 48 (3): 443–453.

Nixon K.C. 1999. The parsimony ratchet, a new method for rapid parsimony analysis. — Cladistics, 15 (4): 407–414.

Notredame C., Higgins D.G. 1996. SAGA: sequence alignment by genetic algorithm. — Nucleic Acids Research, 24 (8): 1515–1524.

Ogden T.H., 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.

Padial J., Grant T., Frost D.R. 2014. Molecular systematics of terraranas (Anura: Brachycephaloidea) with an assessment of the effects

of alignment and optimality criteria. — Zootaxa, 3825: 1–132.

Prendini L., Weygoldt P., Wheeler W.C. 2005. Systematics of the *Damon variegatus* group of African whip spiders (Chelicerata: Amblypygi): Evidence from behaviour, morphology and DNA. — Organisms Diversity and Evolution, 5 (3): 203–236.

Redelings B.D., Suchard M.A. 2005. Joint Bayesian estimation of alignment and phylogeny. — Systematic Biology, 54 (3): 401–418.

Riesgo A., Andrade S.C., Sharma P.P. et al. 2012. Comparative description of ten transcriptomes of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa. — Frontiers in Zoology, 9 (1): 1–24.

Robinson D.F. 1971. Comparison of labelled trees with valency three. — Journal of Combinatorial Theory, 11 (2): 105–119.

Roch S. 2006. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. — IEEE/ACM Transactions on Computational Biology and Bioinformatics, 3 (1): 92–94.

Roshan U., Moret B., Williams T., Warnow T. 2004. Rec-I-DCM3: A fast algorithmic technique for reconstructing large phylogenetic tree. — Proc. IEEE Computer Society Bioinformatics Conference CSB 2004, Stanford (CA). P. 98–109.

Sankoff D.M. 1975. Minimal mutation trees of sequences. — SIAM Journal on Applied Mathematics, 28 (1): 35–42.

Sharma P.P., Kaluziak S.T., Pérez-Porro A.R. et al. 2014. Phylogenomic interrogation of arachnida reveals systemic conflicts in phylogenetic signal. — Molecular Biology and Evolution, 31 (11): 2963–2984.

Sievers F., Dineen D., Gibson T.J. et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. — Molecular Systems Biology, 7 (1): 539.

Stamatakis A., Ludwig T., Meier H. 2005. Raxml-iii: A fast program for maximum likelihood-based inference of large phylogenetic trees. — Bioinformatics, 21 (4): 456–463.

Suchard M.A., Redelings B.D. 2006. Bali-Phy: Simultaneous Bayesian inference of alignment and phylogeny. — Bioinformatics, 22 (16): 2047–2048.

Swofford D.L. 1990. PAUP: Phylogenetic Analysis Using Parsimony. Version 2.4. Distributed by the Illinois Natural History Survey: Champaign, Illinois.

Thorne J.L., Kishino H. 1992. Divergence time and evolutionary rate estimation with multilocus datafreeing phylogenies from the artifacts of alignment. — Molecular Biology and Evolution, 9 (6): 1148–1162.

Thorne J.L., Kishino H., Felsenstein J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. — Journal of Molecular Evolution, 33 (2): 114–124.

Thorne J.L., Kishino H., Felsenstein J. 1992. Inching toward reality: an improved likelihood model of sequence evolution. — Journal of Molecular Evolution, 34 (1): 3–16.

Ukkonen E. 1985. Finding approximate patterns in strings. — Journal of Algorithms, 6 (1): 132–137.

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

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.

Wagner W.H. 1961. Problems in the classification of ferns. — Recent Advances in Botany. Toronto: University of Toronto Press. P. 841–844.

Wang L., Gusfield D. 1997. Improved approximation algorithms for tree alignment. — Journal of Algorithms, 25 (2): 255–273.

Wang L., Jiang T. 1994. On the complexity of multiple sequence alignment. — Journal of Computational Biology, 1 (4): 337–348.

Wang L., Jiang T., Gusfield D. 2000. A more efficient approximation scheme for tree alignment. — SIAM J. Comput., 30 (1): 283–299.

Wang L., Jiang T., Lawler E.L. 1996. Approximation algorithms for tree alignment with

a given phylogeny. — Algorithmica, 16 (3): 302–315.

Wheeler W.C. 1994. Sources of ambiguity in nucleic acid sequence alignment. — Schierwater B., Streit B., DeSalle R. (eds). Molecular ecology and evolution: Approaches and applications. Basel: Birkhäuser Verlag. P. 323–352.

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

Wheeler W.C. 1999. Fixed character states and the optimization of molecular sequence data. — Cladistics, 15 (4): 379–385.

Wheeler W.C. 2001. Homology and the optimization of DNA sequence data. — Cladistics, 17 (1): S3–S11.

Wheeler W.C. 2003a. Implied alignment. — Cladistics, 19 (3): 261–268.

Wheeler W.C. 2003b. Iterative pass optimization. — Cladistics, 19 (3): 254–260.

Wheeler W.C. 2003c. Search-based character optimization. — Cladistics, 19 (4): 348–355.

Wheeler W.C. 2006. Dynamic homology and the likelihood criterion. — Cladistics, 22 (2): 157–170.

Wheeler W.C. 2007. The analysis of molecular sequences in large data sets: where should we put our effort? — Hodkinson T.R., Parnell J.A.N. (eds). Reconstructing the Tree of Life: Taxonomy and systematics of species rich taxa. Oxford: Oxford Univ. Press. P. 113–128.

Wheeler W.C. 2012. Systematics: A course of lectures. Wiley-Blackwell. 460 p.

Wheeler W.C. 2014. Maximum a posteriori probability assignment (MAPA): An optimality criterion for phylogenetic trees via weighting and dynamic programming. — Cladistics, 30 (3): 282–290.

Wheeler W.C., Giribet G. 2009. Phylogenetic hypotheses and the utility of multiple sequence alignment. — Rosenberg M.S. (ed.). Perspectives on biological sequence alignment. Berkeley (CA): University of California Press. P. 95–104.

Wheeler W.C., Gladstein D.S. 1991–1998. MALIGN. Program and documentation. Documentation by Janies D., Wheeler W.C. New York. http://research.amnh.org/scicomp/projects /malign.php.

Wheeler W.C., Gladstein D.S. 1994. MALIGN: A multiple sequence alignment program. — Journal of Heredity, 85 (5): 417–418.

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

Wheeler W.C., Lucaroni N., Hong L. et al. 2013. POY version 5.0. New York: American Museum of Natural History. http://research.amnh.org/scicomp/projects/poy.php.

Wheeler W.C., Lucaroni N., Hong, L. et al. 2015. POY version 5: Phylogenetic analysis using dynamic homologies under multiple optimality criteria. — Cladistics, 31 (2): 189–196.

Whiting A.S., Pellegrino K.C., Rodrigues M.T. 2006. Comparing alignment methods for inferring the history of the new world lizard genus *Mabuya* (Squamata: Scincidae). — Molecular Phylogenetics and Evolution, 38 (3): 719–730.