Effects of Gap Open and Gap Extension Penalties

Hyrum Carroll, Perry Ridge[1], Mark Clement, Quinn Snell
Computer Science Department, Brigham Young University
Provo, Utah 84602, USA
{hdc,clement,snell}@cs.byu.edu, perry.ridge@gmail.com

***Abstract*— Fundamental to multiple sequence alignment algorithms is modeling insertions and deletions (gaps). The most prevalent model is to use gap open and gap extension penalties. While gap open and gap extension penalties are well understood conceptually, their effects on multiple sequence alignment, and consequently on phylogeny scores are not as well understood. We use exhaustive phylogeny searching to explore the effects of varying the gap open and gap extension penalties for three nuclear ribosomal data sets. Particular attention is given to optimal phylogeny scores for 200 alignments of a range of gap open and gap extension penalties and their respective distribution of phylogeny scores.**

Keywords: Alignment, gap penalties

I. Introduction

The explosion in DNA sequence data has revolutionized the way scientists perform biological and genetic analysis. By analyzing sequence data for different species, researchers can determine which species are most closely related and make conservation decisions based on these results [1]. Multiple sequence alignment (MSA) is frequently the first step in determining where active regions in proteins are located and plays a critical role in understanding the function of genes and how they govern life. Alignment also plays a central role in sequence analysis as the first step in comparing corresponding regions in the genomes of different organisms (comparative genomics). Since a refined multiple sequence alignment is crucial to so many different types of life-saving research, it is surprising that multiple sequence alignment does not receive more attention from the research community.

Multiple sequence alignment can be performed with DNA nucleotide or amino acid sequences. MSA algorithms insert gaps in order to align the sequences to maximize similarity according to the evolutionary model summarized in the substitution matrix [2]. Gaps correspond to an insertion or deletion of a substring (sometimes a single residue). Gaps can occur because of single mutations, unequal crossover in meiosis, DNA slippage in the replication process or translocation of DNA between chromosomes.

One of the most popular algorithms for MSA is the progressive sequence alignment algorithm [2], [3]. In a progressive sequence alignment algorithm, the substitution matrix is used to determine the likelihood of an observed mismatch (the mismatch may be the result of a mutation or sequencing error). The algorithm then decides to either insert a gap or allow the mismatch to remain in the alignment. In progressive sequence alignment algorithms, inserted gaps are never removed.

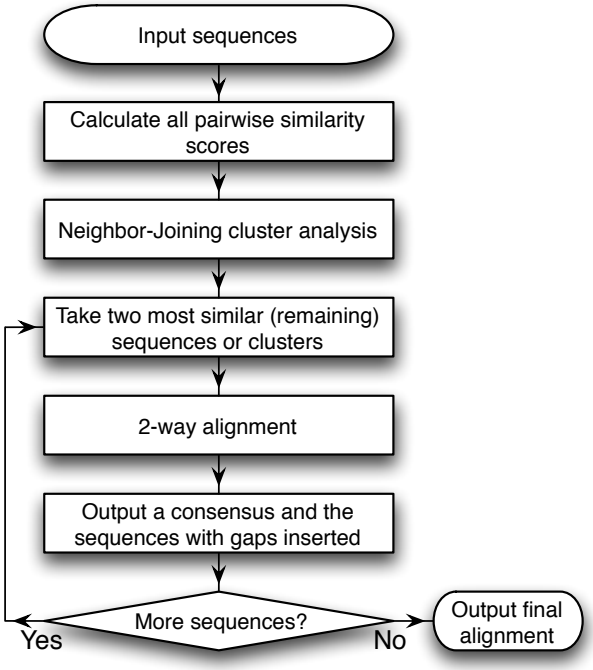

Fig. 1. ClustalW flowchart

The popular alignment program, ClustalW [2], is used in this research. ClustalW utilizes the progressive sequence alignment algorithm (see Figure 1). There are two main phases to progressive alignment. First, a distance matrix is calculated from similarity scores for every possible pair of sequences. ClustalW uses the Wilbur and Lipman algorithm [4] to calculate the distances. These similarity scores are only very general approximations, but work as a starting point [4]. The similarity scores are clustered together with a modified version of the Needleman-Wunsch algorithm [5], producing a guide tree. The second phase consists of following the topology of the guide tree, and at every node aligning the sequences in each of the subtrees until all sequences have been included in the alignment. The first phase generally requires the vast majority of the time and can be skipped by supplying a guide tree.

Since there are several accepted methods for computing a multiple sequence alignment, it is difficult to evaluate the accuracy of an alignment. The alignment score is dependent on the substitution matrix and gap penalties. ClustalW provides

[1]Present Address: University of Nebraska - Lincoln, Lincoln, Nebraska 68505, USA

Data Set	Gene	Number of Sequences	Average Length	Max Length
1	18S	12	1856.25	1998
2	28S	12	332.83	340
3	28S	12	657.083	710

TABLE I

CHARACTERISTICS OF THE THREE RIBOSOMAL DATA SETS USED. AVERAGE LENGTH AND MAX LENGTH ARE MEASURED IN BASE PAIRS.

an alignment score for each multiple sequence alignment performed. However, since this score is dependent on the substitution matrix and gap penalties it cannot be used to compare different alignments of the same data set. The minimum cost for a phylogeny inferred from a given MSA has been suggested as an unbiased measure of the quality of the alignment [6]–[9]. Because no better, unbiased, metric has been commonly used, this research uses the minimum cost phylogeny to determine alignment quality.

Although most phylogenetic search applications use a given multiple sequence alignment as a starting point [10]–[13], multiple sequence alignment has received much less attention than phylogenetic search algorithms [9]. The importance of a quality alignment for the phylogeny search must not be minimized [14], [15]. Morrison *et al.* [14] has even suggested that the resulting phylogeny is affected more by the method used for performing the multiple sequence alignment than the method used to perform the phylogeny search itself.

A. Related Work

The effects of varying parameters for MSA applications was first covered by Fitch and Smith [16] and Williams and Fitch [17]. Several researchers have looked at the effects of parameters for both MSA and phylogenetic search algorithms. These sensitivity analyses have shown that differences in input parameters for MSA have had a greater impact on the phylogeny score then varying the phylogeny search application [14]. Other studies have focused on nodal support and nodal stability [18].

II. DATA SETS

We used three data sets to study the effects of gap open and gap extension penalties (see Table I). The three data sets cover two nuclear ribosomal genes for a wide diversity of hexapod species and are provided by Michael Whiting, a researcher in the Biology Department at Brigham Young University. Data set 1 has twelve 18S gene sequences, while data sets 2 and 3 each of have twelve 28S gene sequences. The sequences in each data set were randomly chosen from larger data sets. We limited the number of sequences included in this study due to the intrinsic computational time limitations of exhaustive searching.

III. RESULTS

Varying the gap open and gap extension costs not only produces very different alignments [14] but produces different distributions of phylogeny scores. Figures 2-4 plot optimal phylogeny scores for gap open penalties (GOP) ranging from 1.0 to 20.0 and gap extension penalties (GEP) evenly distributed between 0 and one half of the respective GOP. For each of the 200 data points in each graph, we used ClustalW to produce the alignment, and then PAUP* [13] to exhaustively generate the phylogenies. While heuristic searches are commonly used, an exhaustive search is necessary to ensure the optimality of the phylogeny score for an alignment. In each of these graphs, the default parameters for ClustalW (GOP 15.0, GEP 6.66) are labeled. Although it is expected that ClustalW's defaults do not produce the optimal alignment with the lowest phylogeny score, it is noteworthy that these scores are 11.0% worse (data set 3) and 106 steps (data set 1) than the best optimal phylogeny score found. Also, the plot for data set 2 clearly reveals that local minima of optimal phylogeny scores exist. For that data set, the local minima has a GOP of 3.0 and a GEP of 1.2. The phylogeny score at that point is 380. Attempts to search over the GOP-GEP space need to incorporate some sort of hill-climbing feature to overcome such local minima.

In addition to gap open and gap extension penalties affecting optimal phylogeny scores, they also greatly affect the distribution of phylogeny scores. Figure 5 illustrate histograms of the phylogeny scores for data sets 1-3. A clear example of the difference in phylogenies scores is exhibited by data set 3. 98% of the possible phylogenies with a GOP of 20.0 and a GEP of 10.0 have of parsimony score worse than any of the phylogenies with a GOP of 2.0 and a GEP of 0.8. In general, varying the GEP parameters *shifts* the histograms of parsimony scores. The shifted distribution retains its general shape and features. For example, the histograms presented for data set 2 each have a second hump aside from another much larger one. Adjusting the gap parameters causes substantial change to both the optimal phylogeny score and its distribution.

IV. CONCLUSION

Gap open and gap extension penalties have long been used to model insertions and deletions. We explored the effects of independently varying the gap open and gap extension penalties for three nuclear ribosomal gene data sets. We employed exhaustive phylogeny searching to guarantee optimal maximum parsimony phylogenies. Varying these parameters not only yields very different optimal phylogenies, but also greatly effects the distribution of possible phylogeny scores. Furthermore, algorithms for traversing the GOP-GEP space need to employ hill-climbing techniques to avoid local minima.

REFERENCES

[1] W.-H. Li, *Molecular Evolution*. Sunderland, Massachusetts: Sinauer Associates, 1997.

[2] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, pp. 4673–4680, 1994.

[3] D. Feng and R. F. Doolittle, "Progressive sequence alignment as a prerequisite to correct phylogenetic trees," *J. Mol. Evol.*, vol. 60, pp. 351–360, 1987.

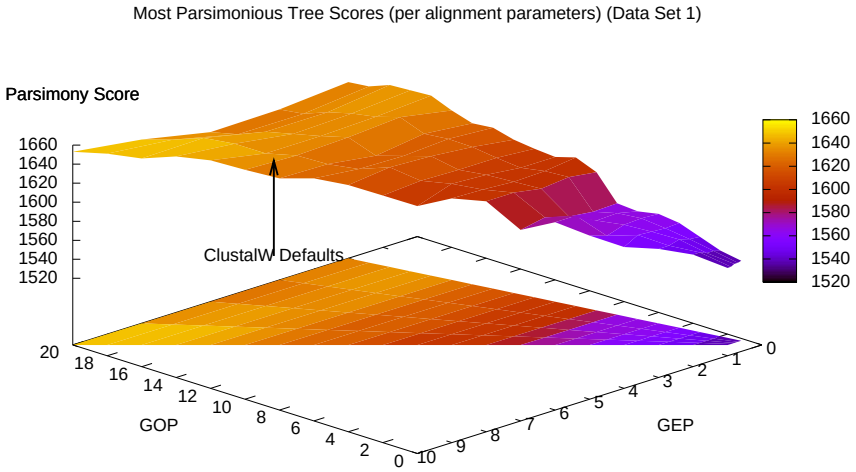

Fig. 2. Optimal parsimony scores for 200 alignments of data set 1. The minimum optimal parsimony score of 1531 has a gap open penalty of 1.0 and two gap extension penalties of 0.4 and 0.5. ClustalW default parameters yield an optimal phylogeny score of 1634.

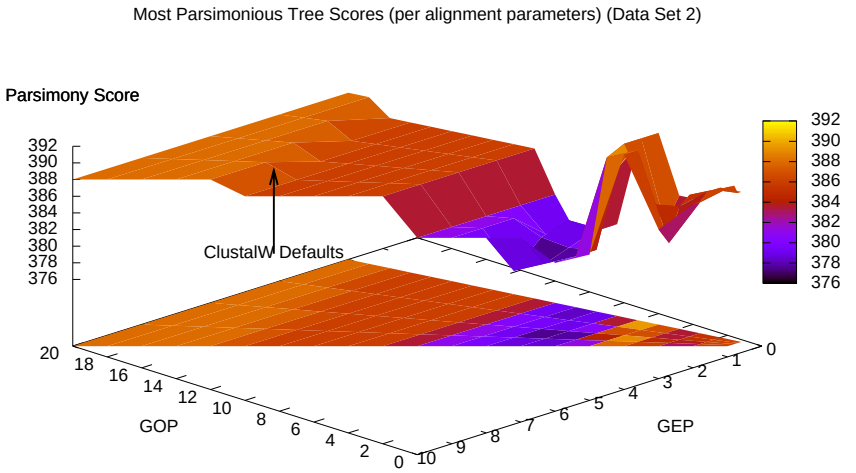

Fig. 3. Optimal parsimony scores for 200 alignments of data set 2. The minimum optimal parsimony score of 376 has a gap open penalty of 8.0 and a gap extension penalty of 3.2. ClustalW default parameters yield an optimal phylogeny score of 388.

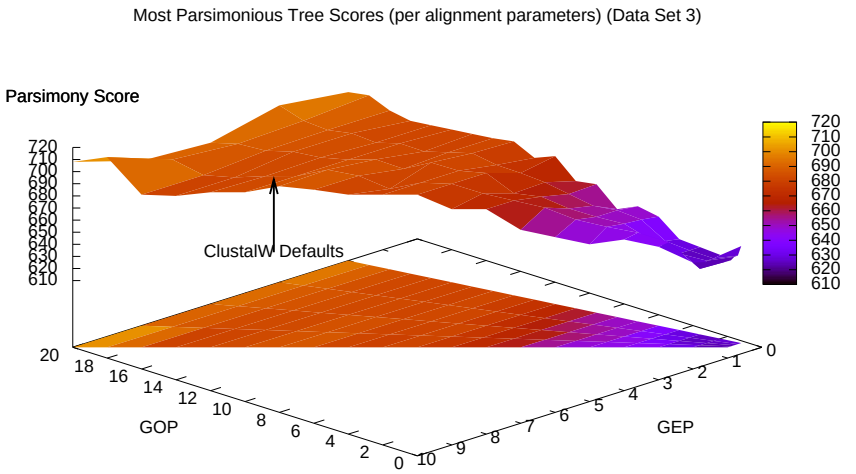

Fig. 4. Optimal parsimony scores for 200 alignments of data set 3. The minimum optimal parsimony score of 618 has a gap open penalty of 2.0 and a gap extension penalty of 0.8. ClustalW default parameters yield an optimal phylogeny score of 686.

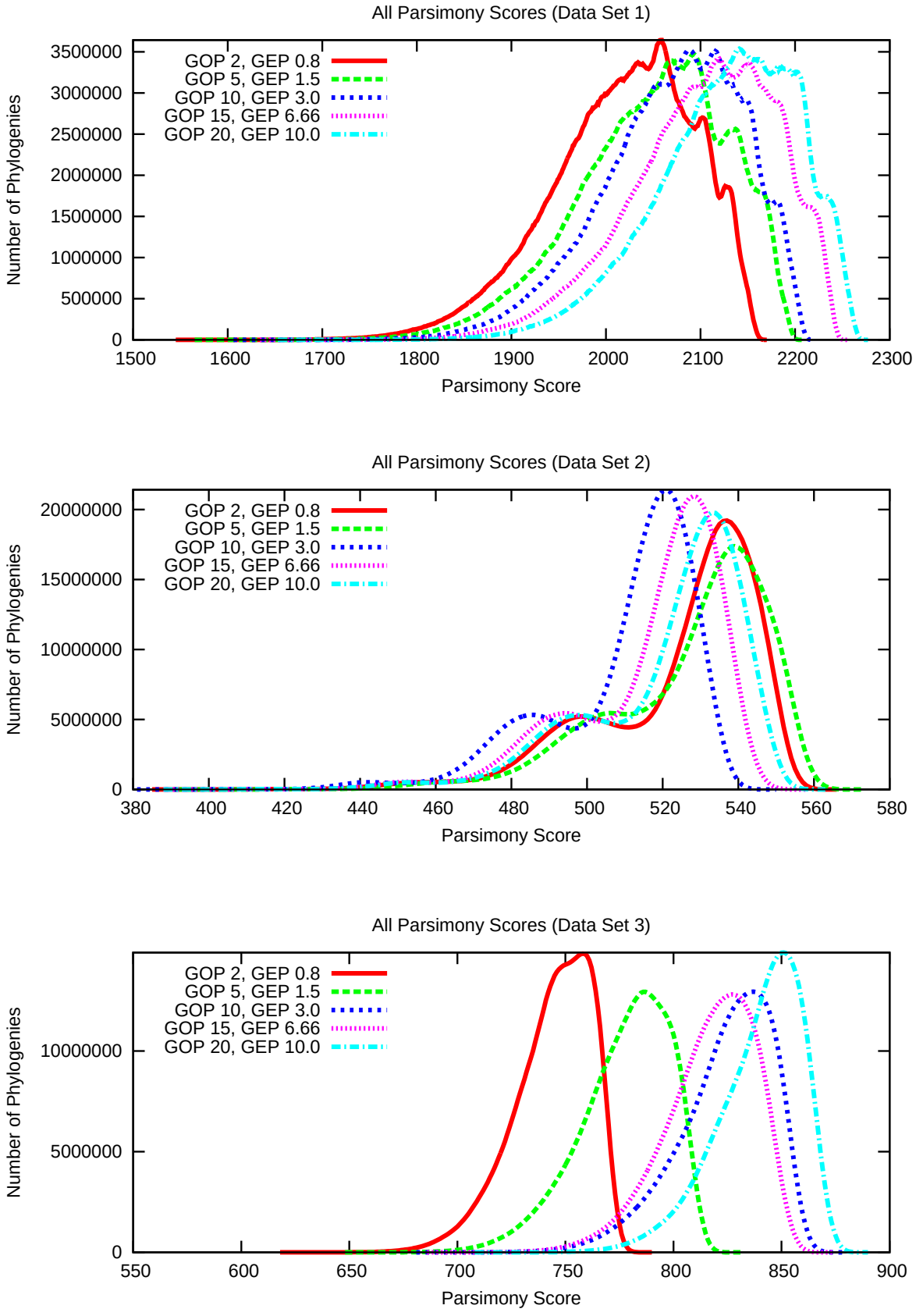

Fig. 5. Representative histograms of all parsimony scores for various alignments parameters for data sets 1-3.

[4] W. Wilbur and D. Lipman, "The context dependent comparison of biological sequences," *SIAM J. Appl. Math*, vol. 44, pp. 557–567, 1984.
[5] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, vol. 48, pp. 443–453, 1970.
[6] A. G. Kluge, "A Concern for Evidence and a Phylogenetic Hypothesis of Relationships Among Epicrates (Boidae, Serpentes)," *Systematic Zoology*, vol. 38, no. 1, pp. 7–25, 1989.
[7] W. C. Wheeler and D. S. Gladstein, "MALIGN: A multiple sequence alignment program," *J. Hered*, vol. 85, pp. 417–418, 1994.
[8] W. C. Wheeler, "Optimization alignment: the end of multiple sequence alignment in phylogenetics?" *Cladistics*, vol. 12, pp. 1–9, 1996.
[9] A. Phillips, D. Janies, and W. C. Wheeler, "Multiple sequence alignment in phylogenetic analysis," *Molecular Phylogenetics and Evolution*, vol. 16, no. 3, pp. 317–330, September 2000.
[10] J. S. Farris, *HENNIG86, version 1.5*, Program and Documentation: Port Jefferson Station, New York, 1988.
[11] J. Felsenstein, "PHYLIP – phylogeny inference package (version 3.2)," *Cladistics*, vol. 5, pp. 164–166, 1989.
[12] P. Goloboff, "Analyzing large datasets in reasonable times: Solutions for composite optima," *Cladistics*, vol. 15, pp. 415–428, 1999.
[13] D. L. Swofford, *PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4*. Sunderland, Massachusetts: Sinauer Associates, 2003.
[14] D. A. Morrison and J. T. Ellis, "Effects of nucleotide sequence alignment on phylogeny estimation: A case study of 18s rDNAs of apicomplexa," *Molecular Biology and Evolution*, vol. 14, no. 4, pp. 428–441, 1997.
[15] N. Mugridge, D. Morrison, T. Jakel, A. Heckeroth, A. Tenter, and A. Johnson, "Effects of sequence alignment and structural domains of ribosomal dna on phylogeny reconstruction for the protozoan family sarcocystidae," *Molecular Biology and Evolution*, vol. 17, pp. 1842–1853, 2000.
[16] W. M. Fitch and T. F. Smith, "Optimal sequence alignments," *Proc. Natl. Acad. Sci. USA*, vol. 80, pp. 1382–1386, March 1983.
[17] P. L. Williams and W. M. Fitch, *The Hierarchy of Life*. Elsevier, Amsterdam, 1989, ch. Finding the minimal change for a given tree.
[18] G. Giribet, "Stability in phylogenetic formulations and its relationship to nodal support," *Systematic Biology*, vol. 52, no. 4, pp. 554–564, 2003.