

Improving ClustalX Alignments with TuneClustalX

Barry G. Hall
drbh@mail.Rochester.edu

Abstract

TuneClustalX facilitates adjusting the gap penalties of ClustalX multiple sequence alignments in order to improve the quality of the alignments. As an example the application is used to determine initial gap penalties that are superior to the default gap penalties for protein alignments. The choice of gap penalties strongly affects the quality of the alignment. TuneClustalX, compiled for Macintosh (Classic and OSX operating systems) and for Windows, and source code for compilation on Unix platforms, together with documentation, is available at <http://www.bellinghamresearchinstitute.com/> by clicking the Software button.

Introduction

Multiple sequence alignment (MSA) is one of the most fundamental tasks in both bioinformatics and phylogenetics. Despite the recent plethora of new MSA methods and associated programs (Edgar 2004; Lee et al. 2002; Morgenstern 1999; Morgenstern 2004; Notredame et al. 2000; Van Walle et al. 2004), the progressive alignment approach (Feng and Doolittle 1987) remains the most widely used method. The progressive alignment approach involves two stages. The first stage involves all possible pairwise alignments of the sequences. Scores of the pairwise alignments are used to create a guide tree. Finally, in the multiple alignment stage, the two most similar sequences are aligned, then additional sequences are added to the alignment in an order that is determined by the guide tree. ClustalW (Thompson et al. 1994) and ClustalX (Thompson et al. 1997) remain two of the most popular implementations of that method. ClustalW is a command line program, and ClustalX is a graphical interface version of ClustalW that includes pull-down menus and a scrollable window that displays the alignment. Although the alignments produced by the two programs are the same, the user can better evaluate the alignments in the ClustalX graphical interface (Chenna et al. 2003)

ClustalX provides a variety of tools to assess the quality of the alignment. One of those tools is a histogram that displays the "quality score" for each column in the alignment immediately beneath that column. The peaks and valleys of that histogram provide a visual representation of the quality of the alignment.

The user can adjust gap penalties for both the pairwise and the multiple alignment stages, and the choice of those gap penalties largely determines the quality of the alignment. Typically the user aligns a set of sequences using the default gap penalties, then attempts to improve the alignment by adjusting those penalties and observing the effects of those adjustments on the histogram of column quality scores. The process of refining the gap penalties can be very frustrating because it is typically impossible to see all of the histogram at one time, thus it is difficult to assess the overall consequences of gap penalty adjustments. Comparisons of gap penalties are also frustrating because it is impossible to

hold a visual image of a previous histogram in mind in order to compare it with a new histogram following a modification of gap penalties. It is not uncommon for there to be manual adjustments of gaps after the user has arrived at an apparently optimal alignment. This implies that the alignment is not, after all, optimal and there is a strong possibility of subjectivity in choice of residues and placement of gaps during such manual adjustments. That subjectivity can lead to strong biases that affect downstream analyses such as phylogenetic analyses.

To facilitate comparisons of alignment qualities ClustalX provides a "Save Column Scores to File" option under the Quality menu. Selection of that option generates a text file with the extension ".qscore" in which the residues in each column are written to a single line that is followed by the quality score for that column. Again, however, examination of that file does not provide a good sense of the overall quality of the alignment, nor does it permit an easy comparison with a previous alignment.

Results

TuneClustalX is a simple program that reads the Q-score file and calculates the average Q-score. To improve the alignment quality the qscores are saved to a file, the average Q-score is computed by TuneClustalX, the gap penalties are modified and the sequences realigned, and the Q-scores of the new alignment are again saved to a file and the average Q-score is calculated by TuneClustalX. Comparison of the two average Q-scores permits the user to judge whether the modified gap penalties improved, had no effect, or made worse the quality of the alignment. Because the pairwise gap penalties are only used to construct the guide tree, typically the user need only modify the multiple alignment gap penalties in order to maximize the quality of the alignment.

ClustalX also provides, under the Quality menu, tools that allow the user to identify misaligned or low-scoring regions. Once the best overall alignment, based on the highest average Q-score, is obtained it is often possible to further improve the alignment by adjusting the gap penalties again and realigning only selected low-scoring regions by using the "Realign selected residue range" option under the Alignment menu.

TuneClustalX provides the option of calculating the average Q-score for a specified residue range, permitting the user to quickly determine the effects of such local gap penalty modifications.

Application of TuneClustalX to improving default gap penalties for protein alignments

ClustalX provides default gap penalties for both the pairwise and the multiple alignment stages. For protein sequences the default penalties for the pairwise alignment stage are 10 for gap opening and 0.1 for gap extension, and for the multiple alignment stage are 10.0 for gap opening and 0.2 for gap extension.

To evaluate the default gap penalties for the pairwise alignment stage the program EvolveAGene 2.2 (Hall 2005) was used to evolve a set of 64 sequences descending via a bifurcating tree from a 1494 bp root sequence with an average of 250 changes per branch. That program creates a true protein sequence tree in which the correct topology is known. The program CompareTrees (Hall 2005) was used to compare the topology of the true tree with the topology of the guide trees produced by ClustalX when the set of protein sequences were aligned using pairwise alignment gap opening penalties ranging from 1 to 40 and gap extension penalties ranging from 0.1 to 2.0. There is a broad optimum for the gap opening and gap extension penalties of the pairwise alignment stage for protein sequences: Gap opening penalties of 10 & 15 with gap extension penalties from 0.1 to 2.0 all

result in identical topology scores for the guide tree. The default gap penalties for the pairwise stage therefore appear to be very satisfactory.

To evaluate the gap penalties for the multiple alignment stage the same set of protein sequences was aligned using the default gap penalties for the pairwise alignment stage and for the multiple alignment stage using gap opening penalties of 5, 10, 15, and 20. For each gap opening penalty the gap extension penalties were set to 0.1, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, and 2.0. For each alignment a Q-scores file was saved and TuneClustalX was used to calculate the average Q-score. A sharp optimum was found at a gap opening penalty of 3.0 and a gap extension penalty of 1.8.

Table 1 compares the average Q-scores for alignments of three different sets of simulated sequences and for four different sets of real sequences using the default gap penalties and the "optimum" penalties of 3.0 and 1.8. The Q-scores in Table 1 were compared by a paired t-test. The optimum gap penalties were significantly better ($p=0.022$) with a mean difference of 1.08.

A similar study of gap penalties for DNA alignments showed that the default penalties for both the pairwise and multiple alignment stages are very satisfactory (data not shown).

Discussion

TuneClustalX is a useful tool for improving nucleic acid and protein sequence alignments by adjusting gap penalties both globally and locally. Sometimes there will be a broad range of gap penalties, or several configurations of gap penalties, that are equally suitable to the data set in question. It should be understood, however, that the "optimum" alignment, i.e. the alignment that maximizes the average Q-score, is not the "true" alignment in which each gap represents an actual indel in the history of the sequences. The optimal alignment is simply the best alignment that can be obtained under the algorithm and assumptions employed by ClustalX. That "optimum" alignment is, however, devoid of the subjective placement of gaps and the inherent bias that results from human intervention in alignments.

Acknowledgement

I am grateful to Harmit Malik for helpful suggestions during the preparation of this report.

REFERENCES

- Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **31**, 3497-500.
- Edgar RC. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-7.
- Feng DF, Doolittle RF. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol* **25**, 351-60.
- Hall BG. (2005) Comparison of the Accuracies of Several Phylogenetic Methods Using Protein and DNA Sequences. *Mol. Biol. Evol.* **Submitted**.
- Lee C, Grasso C, Sharlow MF. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics* **18**, 452-64.
- Morgenstern B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15**, 211-8.

- Morgenstern B. (2004) DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. *Nucleic Acids Res* **32**, W33-6.
- Notredame C, Higgins DG, Heringa J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205-17.
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nuc. Acid. Res.* **25**, 4876-82.
- Thompson JD, Higgins DG, Gibson TJ. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680.
- Van Walle I, Lasters I, Wyns L. (2004) Align-m--a new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics* **20**, 1428-35.

Table 1 Average Q-scores

| Data Set | Default (Pairwise alignment penalties: gap opening =10.0 and gap extension = 0.1, Multiple alignment penalties: gap opening =10 and gap extension = 0.2) | Optimum (Pairwise alignment penalties: gap opening =10.0 and gap extension = 0.1, Multiple alignment penalties: gap opening =3.0 and gap extension = 1.8) |
|---|---|--|
| Simulated sequences, root = XisC, 1494 bp, 64 taxa, branch lengths average 250 changes | 19.40 | 20.35 |
| Simulated sequences, root = TEM-1, 858 bp, 64 taxa, branch lengths average 144 changes | 12.0 | 14.18 |
| Simulated sequence, root = AAC(6')-11, 456 bp, 64 taxa, branch lengths average 76 changes | 13.68 | 15.19 |
| Class A β -lactamases 75 taxa | 18.97 | 19.06 |
| Metallo- β -lactamases Subclass B1+B2 & homologs, 50 taxa | 14.96 | 15.96 |
| Metallo- β -lactamases Subclass B3 & homologs, 25 taxa | 12.05 | 13.09 |
| OXA β -lactamases, 34 taxa | 22.25 | 23.04 |