

Optimizing Multiple Protein Alignments

To optimize the gap penalties I used a simulated tree of the XisC sequence (1494 bp) with an average branch length of 250 accepted mutations. The diversity of those sequences pushes the limits of multiple alignment programs.

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. During the multiple alignment stage study I therefore used Pairwise Alignment penalties of 10 for gap opening and 0.1 for gap extension (default settings) for proteins

For the multiple alignment stage the gap opening penalties were set to 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.

Average Q-scores were recorded.

Results

The optimum was Gap opening of 3 and gap extension of 1.8

The optimum penalties were then compared with the ClustalX default penalties and the penalties that I had previously recommended in *Phylogenetic Trees Made Easy*, both the first edition (2001) and the second edition (2004) for two additional simulated data sets and for four real data sets.

Data Set	Previously Recommended (Pairwise penalties = 35.0 and 0.75, , Multiple penalties = 15 and 0.3)	Default (Pairwise penalties = 10.0 and 0.1, Multiple penalties = 10 and 0.2)	Determined Optimum (Pairwise penalties = 10.0 and 0.1, Multiple penalties = 3.0 and 1.8)
Simulated sequence, root = XisC, 1494 bp, 64 taxa, branch lengths average 250 changes	19.3699	19.4045	20.3474
Simulated sequence, root = TEM-1, 858 bp, 64 taxa, branch lengths average 144 changes	11.3851	12.0	14.1826
Simulated sequence, root = AAC(6')-11, 456 bp, 64 taxa, branch lengths average 76 changes	14.4091	13.6809	15.1872
Class A β -lactamases	18.5833	18.9667	19.0592

75 taxa			
Metallo- β -lactamases Subclass B1+B2 & homologs, 50 taxa	14.8177	14.9589	15.9632
Metallo- β -lactamases Subclass B3 & homologs, 25 taxa	11.8649	12.0534	13.0854
OXA β -lactamases, 34 taxa	22.5936	22.2486	23.0432

The gap penalties in the above table were compared by paired t-tests:

Default vs previously recommended: No significant difference ($p = 0.41$)

Optimum vs previously recommended: Optimum is highly significantly better, mean difference is 1.12, $p = 0.005$

Optimum vs default: Optimum is significantly better, mean difference is 1.08, $p = 0.022$

Note that the "optimum" should always be considered only a starting point for the refinement of your alignment. Poorly aligned regions may be able to be improved by modifying the gap penalties for just those regions as described in *Phylogenetic Trees Made Easy* 2nd Edition by Barry G. Hall, Sinauer Assoc., Publisher, 2004.