A Critical Evaluation of Multiple Sequence Alignment Programs in Aligning Domains of the Bcl-2 Family

INTRODUCTION

Multiple sequence alignments are a valuable tool in the biological sciences. They can help to determine aspects of protein structure, identify important regions for protein function, and classify proteins into families. The advent of the genomic era with the complete sequencing of multiple organisms has increased the importance of correctly aligning similar proteins both within and across species. When only two proteins need to be aligned, it is possible to compare each amino acid of one sequence to that of the other and determine the best path that will maximize the alignment of the two sequences. However, the amount of computational time that is required to perform the same analysis on a larger set of sequences limits the use of this method in generating multiple sequence alignments. Thus, numerous heuristic approaches have been developed to counter this problem. These different methods may enable the programs to perform better under one set of conditions than another. Here, I assess the abilities of five multiple sequence alignment programs – ClustalW, MultAlin, T-Coffee, MAP, and ProAlign – to properly align the Bcl-2 homology domains both within a subfamily and among the subfamilies.

Overview of Multiple Sequence Alignment Programs

Many multiple sequence alignment programs have been developed based solely on one or a combination of two widely used approaches, a progressive or iterative method. In the progressive method, originally introduced by Feng and Doolittle [1], the two most similar sequences are aligned first followed by the incorporation of more divergent sequences into the

alignment. The iterative approach, on the other hand, uses a scoring function to guide the alignment such that a higher score reflects a more biologically correct alignment [2]. Often, this requires repeat iterations of the process until there is no further optimization of the score.

The programs selected for this analysis are based on variations of the progressive method. The majority of these programs first perform pairwise comparisons of the sequences in a given set to determine their relatedness. This information is used to generate a dendrogram or guide tree that reflects the degree of similarity among the sequences. The two most closely related sequences are then aligned first, and the algorithm follows the guide tree to determine the order by which additional sequences will be incorporated. MultAlin is an iterative, progressive alignment that uses the UPGMA method to generate a guide tree [3]. However, this program generates a multiple sequence alignment by first aligning within individual clusters before aligning among the clusters. Once an initial alignment is produced, the program then gives the alignment a score that is the sum of the pairwise alignment scores. It then repeats the hierarchical clustering and continues this process until there is no further change in the dendrogram that is produced. Thus, by taking into account that some subsets of the sequences may be more similar to one another than to the other sequences in the set, MultAlin is expected to work well in data sets containing different families of proteins.

ClustalW is a progressive pairwise sequence alignment that was designed to improve the sensitivity of traditional progressive alignment programs, specifically by addressing the parameter choice problem [4]. The basis for this problem is that traditional progressive alignment algorithms selected a single weight matrix and fixed gap penalties for opening and extending gaps regardless of its position within the sequence. A single weight matrix is problematic when divergent sequences are aligned because there is less sequence identity present

and more mismatches. Determining the proper weight to give to mismatches is important in determining how the sequences are to be aligned. The second issue of having fixed values for gap opening and extension is problematic because gaps do not occur randomly in proteins. Residues within a domain or secondary structure are less likely to possess gaps than linker segments between these structural elements. ClustalW improves upon both parameters by giving different weights to sequences within a set and varying the gap penalties in a position and residue specific manner. Sequences are assigned different weights based upon their evolutionary distance relationships derived from a dendrogram generated using the Neighbor Joining program. Similar sequences get down-weighted while divergent sequences are up-weighted. In addition, gap penalties are varied based upon the likelihood of a gap being present next to each of the 20 amino acids and on the presence of loops as suggested by a string of 5 or more hydrophilic residues. Gaps that occur in loop regions are penalized less than those that occur within a secondary structure. Thus, ClustalW is expected to provide enhanced sensitivity and has become a widely used program in aligning multiple sequences.

The program T-Coffee (Tree-based Consistency Objective Function for alignment Evaluation) was designed to improve upon ClustalW by addressing the problem of "greediness" that is not addressed by ClustalW [5]. The concept of "greediness" refers to the concept that mistakes made early on in an alignment can be propagated to the rest of the alignment since the two most similar sequences are aligned first, and the rest of the alignment follows this initial alignment. To generate a better alignment, T-Coffee generates both local and global pairwise alignments among all the sequences and then builds a library that incorporates both sets of alignments. The program then aligns sequences taking into account how each sequence aligns with its closest neighbor and in relation to all other sequences. By considering information from