Adam Abeshouse

CS182 Capstone Abstract

For my capstone, I examined a new method of progressive multiple alignment based on eigenvectors for order.

There's no obviously natural objective function to be maximized for multiple alignment. One strategy that's been tried is progressive alignment, whereby sequences are aligned, and then the next sequence is aligned to the alignment, and so on until they are all aligned in a block. Clearly the order in which the sequences are aligned is crucial.

Some strategies have focused on measures of similarity of the sequences to all the others, choosing to align in decreasing order of 'overall similarity.' There is some evidence[1] that such strategies can achieve a consistent approximation of optimal for the sum of pairs multiple alignment objective function, which is the sum of the pairwise alignment scores of the sequences in the block.

Viewing sequences as vertices in an undirected graph, with weighted edges weighted corresponding to their pairwise alignment score, I explored two measures of similarity, one based on the first eigenvector of the adjacency matrix, and one based on the first eigenvector of the Laplacian matrix. Similar methods have been used to rank sports teams. I compared the new method to a popular progressive alignment method known as Clustal and found that in many cases, with respect to a

[1] Gusfield, Dan. "Efficient methods for multiple sequence alignment with guaranteed error bounds." *Bulletin of mathematical biology* 55.1 (1993): 141-154.

certain alignment scoring scheme, it did at least as well if not better, and in significantly less time. I concluded that the methods are worth more research.

Below are multiple alignments performed by the program, with regions that are conserved in all the sequences boxed in red: