Additional Paper for “Multiple Sequence Alignment”


Abstract: The authors presented a generalized concept of a multiple alignment as a “threaded blockset”, which describes a multiple alignment as blocks of locally aligned subsequences (blockset) linked together by a path. By using such a formulation, each individual sequence can be used as the “reference” sequence upon which the other sequences can be projected using the cross-mapping information in a series of blocks (threaded blocks). The authors proposed a method based on a simulated divergence from a known sequence to evaluate the performances of different multiple aligners. In addition, the authors also introduced an implementation of the Threaded Blockset Aligner (TBA) called MULTIZ and made the aligned dataset available on the UCSC Genome Browser.

Discussion: As the genomes of more and more species become available, it is hard to overstate the importance of identifying and comparing orthologous regions from different species. This problem is further complicated by the need to visualize these alignments using base “reference” sequences in an ungapped way that is key to any genome browsers. Therefore the authors proposed a way to constructs blocks of local alignments that contains one or all of the sequences represented in the alignment. To build an ungapped alignment using a reference sequence, all it takes then is to thread through these blocks in one direction according to that reference sequence. The building of the actual alignment relies on the geometric intuition of finding a highest scoring path in a grid and does so by a scoring matrix that is refined through dynamic programming. In addition, there is an algorithm wherein TBA can invoke MULTIZ such that it can build transitive blocksets beyond a pairwise alignment by using a guide tree. To demonstrate the capabilities of the program, the authors demonstrated how TBA can display the HoxA cluster alignment between 4 tetrapods and 4 teleosts by using either tilapia or human as the base sequence and how the corresponding alignment changes. An additional problem the authors tackled was the evaluation of alignment algorithms. Since the use of existing benchmarks are subjected to biases both in the construction of those alignments and unknown errors, the authors resorted to a different method which simulated evolutionary divergence of a number of “species” from a known ancestral sequence. Then the performances of different alignment algorithms are compared against each other. The results from such comparison suggest that TBA and MULTIZ performs consistently better than a number of other existing programs for the dataset used. In addition, an estimate of the sensitivity and specificity according to Brudno, Batzoglou and Pachter also seem to favor TBA. This advantage also seems to carry over to a running time comparison, where TBA and MULTIZ completes the operation faster than other methods, with the exception of MAVID, which seems to have a low sensitivity to specificity ratio, according to the authors. Finally, the authors presented an easy way to access the data through the UCSC Genome Browser and other software.