Processing Local Alignments

Steffi Gebauer-Jung and Thomas Wiehe

Max Planck Institute for Chemical Ecology, Department of Molecular Genetics and Evolution,
Tatzendpromenade 1a, 07745 Jena, Germany,
Phone: +49-3641-64 3631, Fax: +49-3641-64 3668, e-mail: gebauer-jung@ice.mpg.de

Local alignment algorithms belong to the most important tools in comparative genomics. However, when large genomic sequences are aligned, several problems arise: depending on parameter choices and the scoring method local alignments may be highly fragmented; nucleotides in either sequence may not be uniquely aligned; or significant matches are interspersed with insignificant ones. These problems make it difficult to distinguish – on inspection or computationally – conserved from non-conserved regions and to assess the level of conservation in different stretches of the genomic sequences. To deal with these problems Z. Zhang et al., (1999), recently suggested a method to post-process local alignments which decomposes a given local alignment into a hierarchy of subalignments. Here, we present methods for post-processing which rest upon filtering, score-optimization or calculation of a summary match-statistic. Given a local alignment, which had been obtained for example with the Smith-Waterman method, one of our programs, `sim2lean`, generates a so-called *lean alignment*. This is a disjoint (any pair of retained aligned fragments is non-overlapping), maximal (the combined score of the retained fragments is maximal) and – optionally – monotonous (the retained aligned fragments can be ordered as a monotonous sequence with respect to their DNA-sequence positions) subalignment. The algorithm is based upon a dynamic programming approach. Another program, `sim2compile`, compiles all aligned fragments found within a sliding window and renormalizes the compiled match percentages with the average of the lengths of the genomic sequences. The result is a single, quasi-continuous function, CMP, the compiled match-percentage. Alternatively, when gene structures are known, CMP can be computed for all gene elements such as introns, exons, UTR's, individually. This makes it possible to directly compare the conservation level within and between elements – a typical problem for gene annotators.

We will give a detailed description of the programs and show examples of their use in alignments of large genomic sequences.