Sequence similarity, motif detection and alignments with N-local decoded anchor points

Florian Pitschi

19.12.2009

Abstract

In the post-genomic era, more and more species and with them, an ever increasing number of their genes and other DNA-encoded hereditary material in their genome have become sequenced and made available to the research community in a large number of general, as well as species specific or even more specialized databases. Although being investigated for a long time, detecting and analyzing similarities in a given set of DNA, RNA or amino acid sequences or starting with a single query sequence and looking for similar sequences in a database remains an essential and challenging computer science problem and important for the daily work of many scientists. The topics where sequence similarity comes into play are numerous and diverse, including motif finding, homology search, sequence comparison as well as transcription factor binding site detection and phylogeny and last but not least sequence alignment, because aligned sequences are the basis for many other tasks again. The continuous importance and popularity of similarity search is also proved by the fact that the original BLAST paper, published in 1990, became the most referred scientific paper of the last decade and got cited more than 10.000 times.

Many new ideas have been developed since then and growing data sets and bigger projects (whole genome alignments) ask for efficient algorithms to detect similarities between sequences. In my PhD thesis, I investigated and extended one of them. Applied to a set of sequences, the method's algorithm will output a set of anchor points. These anchor points indicate which positions in the sequences are considered as similar. They can be simply used to visualize which parts of the sequences are conserved and shared between different sequences by coloring positions that carry the same anchor class with the same color in each sequence. This is helpful, because the user can quickly and easily distinguish conserved blocks, which makes motif detection and - if desired - alignment by hand much easier for him. For the later, the user just has to move equally colored letters (nucleotides or amino acids) in the sequences under one another in the same column.

To show the usefulness of the developed anchor computation method, I present how the computed anchors can be further processed (by using newly develeoped graph algorithms to create a consistent subset of the initial anchor equivalence classes) and used in a variety of different applications. First, I showed their ability to create better automatic alignments on the Balibase 2

benchmark database. Secondly, it has be shown that if the anchors are computed for the upstream region of the JUN-gene of different species, they correspond to positions of transcription factor binding sites. On the other hand, I applied them in a very different context, namely to construct correct phylogenetic networks by defining split systems based on the anchors for related input sequences (nuclear ITS region from New Zealand alpine buttercup sequences).