

META-ALIGNMENT OF BIOLOGICAL SEQUENCES

Enrique Blanco García

The sequences are very versatile data structures. In a straightforward manner, a sequence of symbols can store any type of information. Systematic analysis of sequences is a very rich area of algorithmics, with lots of successful applications. The comparison by sequence alignment is a very powerful analysis tool. Dynamic programming is one of the most popular and efficient approaches to align two sequences. However, despite their utility, alignments are not always the best option for characterizing the function of two sequences. Sequences often encode information in different levels of organization (meta-information). In these cases, direct sequence comparison is not able to unveil those higher-order structures that can actually explain the relationship between the sequences.

We have contributed with the work presented here to improve the way in which two sequences can be compared, developing a new family of algorithms that align high level information encoded in biological sequences (meta-alignment). Initially, we have redesigned an existent algorithm, based in dynamic programming, to align two sequences of meta-information, introducing later several improvements for a better performance. Next, we have developed a multiple meta-alignment algorithm, by combining the general algorithm with the progressive schema. In addition, we have studied the properties of the resulting meta-alignments, modifying the algorithm to identify non-collinear or permuted configurations.

Molecular life is a great example of the sequence versatility. Comparative genomics provide the identification of numerous biologically functional elements. The nucleotide sequence of many genes, for example, is relatively well conserved between different species. In contrast, the sequences that regulate the gene expression are shorter and weaker. Thus, the simultaneous activation of a set of genes only can be explained in terms of conservation between configurations of higher-order regulatory elements, that can not be detected at the sequence level. We, therefore, have trained our meta-alignment programs in several datasets of regulatory regions collected from the literature. Then, we have tested the accuracy of our approximation to successfully characterize the promoter regions of human genes and their orthologs in other species.

GBL Dissertation Series

Universitat Politècnica de Catalunya

Meta-Alignment of Biological Sequences — Enrique Blanco García

META-ALIGNMENT OF BIOLOGICAL SEQUENCES

Enrique Blanco García

PhD Thesis

Barcelona, November 2006