Special Issue Papers

Serafim Batzoglou
received his PhD from MIT in 2000. He is an Assistant Professor of Computer Science at Stanford, and his research focus is computational biology.

The many faces of sequence alignment

Serafim Batzoglou

Date received (in revised form): 3rd December 2004

Abstract

Starting with the sequencing of the mouse genome in 2002, we have entered a period where the main focus of genomics will be to compare multiple genomes in order to learn about human biology and evolution at the DNA level. Alignment methods are the main computational component of this endeavour. This short review aims to summarise the current status of research in alignments, emphasising large-scale genomic comparisons and suggesting possible directions that will be explored in the near future.

Keywords: *sequence alignment, local alignment, multiple alignment, synteny detection, rearrangements, hidden Markov model*

INTRODUCTION

Sequence alignment is the poster child of bioinformatics. Alignment is the most basic component of biological sequence manipulation, and has diverse applications in sequence assembly, sequence annotation, structural and functional predictions for genes and proteins, phylogeny and evolutionary analysis. The computational problem of sequence alignment was born perhaps in 1966 with the definition of the *edit distance* between two strings as the minimum number of edit operations – insertions, deletions and letter substitutions – needed to transform one string into another.[1] The subsequent literature on alignment has been enormous, and includes seminal papers such as the original Needleman–Wunsch dynamic programming solution,[2] the Smith–Waterman algorithm for local alignment,[3] the introduction of affine gaps,[4] the progressive approach to multiple alignment[5] and the BLAST tool[6] that enabled genome-scale similarity search. Recently, the literature on basic methodology and tools development has been growing rather than shrinking, indicating that the alignment problem is still not solved. How can that be, after nearly 40 years of research and literally hundreds of available tools?

There are two main reasons that alignments should remain an open problem. First, and most important, alignment is not a single problem but rather a collection of many quite diverse questions that all have in common the search for sequence similarity. Starting from the definition of alignment, there are two biologically meaningful formulations – one based on the desire to find evolutionary relationships and one based on the desire to find putative functional relationships. Given some biological sequences, the first formulation may suggest asking for a mapping (an edge) between every pair of letters that are derived from the same ancestral letter through the replication machinery of cells. This definition, however, is both too restrictive by disallowing point mutations or convergent evolution, and too permissive by not seeking to find the truly orthologous segments between the sequences of two species. The second formulation may seek to find all sequence similarities that are more significant than a threshold above random similarity, implying some common function.

Serafim Batzoglou,
Department of Computer Science,
Stanford University,
James H. Clark Center,
318 Campus Drive, RM S-266,
Stanford, CA 94305–5428, USA

E-mail: serafim@cs.stanford.edu