A Deterministic Algorithm for Alpha-Numeric Sequence Comparison with Application to Protein Sequence Detection

A. Kheniche†[a], N. Brahimi†[b] and A. Salhi†[d] *;
†Department of Mathematical Sciences,
The University of Essex, Colchester CO4 3SQ, UK
[c]as@essex.ac.uk, Tel: 00 44 1206 873022, Fax: 00 44 1206 873043

Received: 11/11/2014; Accepted: 11/02/2015

ABSTRACT
This paper is an extension of a deterministic algorithm, [1], that was initially designed to measure the rate of similarity between DNA sequences, and any sequences made up with symbols of alphabets of cardinality 4. Here, a modified and extended version to handle sequences of symbols from alphabets of cardinality > 4 is presented. This extension opens up its application area. As a test ground, we search for peptides within a protein database. Computational results on real data and a comparison with BLAST will be discussed.

Keywords: BLAST, Deterministic Algorithm, Alpha-Numeric Sequence, Numeration System, Database

1. INTRODUCTION

Sequence analysis is at the core of bioinformatics. A lot of algorithms have been developed to that effect, most of which attempt to measure the similarity between sequences in one way or another. They are, generally, in two broad classes: deterministic and stochastic. The deterministic algorithms are often hard to implement and are invariably slow. Stachastic algorithms, on the other hand, are often easy to implement and efficient. However, not guaranteeing to return the correct results, may be a high price to pay for efficiency particularly in situation where time, as opposed to determinism, is not the issue. Situations where determinism is essential are numerous. For instance, in clinical contexts, it is

---

* Contact Author: A.Salhi, Dept. of Mathematical Sciences, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK