ELSEVIER

Expert Systems with Applications 30 (2006) 34–41

Expert Systems with Applications

www.elsevier.com/locate/eswa

A new dot plot-based algorithm for genomes sequences comparison: A preliminary study

Mohamed Tounsi [a,*], Majed Al-Mashari [b]

[a] Department of Computer & Information Sciences, Prince Sultan University, P.O. Box 66833, Riyadh 11586, Saudi Arabia
[b] Department of IS, College of Computer Science, King Saud University, P.O. Box 51178, Riyadh 11543, Saudi Arabia

Abstract

To have efficient data mining systems, we need powerful algorithms to extract and mine the data. In the case of genomes data mining system, the algorithms search for genomes/proteins that share similar properties. Proteins that have a significant biological relationship to one another often share only isolated regions of sequence similarity. When identifying relationships of this nature, the ability to find local regions of optimal similarity is advantageous over global alignments that optimize the overall alignment of two entire sequences. The paper describes a new method for genome sequence comparison. This algorithm can be used in a genomes data mining system. It provides a good theoretical improvement in accuracy with a modest sacrifice in speed as compared to the most commonly used alternatives. The method is based on the popular progressive approach, the dot plot method, but avoids the most serious pitfalls caused by the greedy nature of this technique. The new approach pre-processes a data set of all pair-wise alignments between the sequences. This provides a library of alignment information that can be used to guide the comparison. The algorithm is based on the similar segment method, i.e. having n similar identities in window of size L. The paper presents some results about the termination and correctness of the algorithm and how to include this algorithm into other comparison algorithms. The paper introduces the mechanism to create random sequences. These data will be our main benchmarks for comparing our algorithms.

Keywords: Genome comparison; Optimization algorithm; Dot plot; Sequence; Alignment

1. Introduction

The sequences genome analysis starts with the study of the similarities between the sequences. A sequence is a word build on an alphabet. This alphabet is {A,C,T,G} for nucleic sequences and 20 letters for 20 acid amine for the proteins sequences.

Such a study needs good algorithms and tools to achieve the sequence comparison of protein and nucleic genomes in less time and to propose the best possible alignments.

Consider the following example of an alignment between two sequences 'RDISLVKNAGI' and 'RNILVSDAKNVGI'

```
R D I S L V - - - K N A G I
R N I - L V S D A K N V G I
```

The dashes represent insertions or deletions called indels or gaps. The order in which the residues occur in the alignment is identical to the one in their respective sequences. Two residues that are printed in the same column are called matched.

An alternative alignment for the first three pairs could be .

We align two sequences when we want letter matching between two sequences:

```
ATTAGCCG
|   ||||
AGCTGCCG
```

We can introduce the gap, i.e. the absence of letter from the first or second sequence noted '-', which is biologically an indel or gap.

```
A  TTAGCCG
|  |  ||||
AGCT  GCCG
```

The objective of this study is to develop an algorithm for sequence comparing. This algorithm can be included into the data mining system to have better knowledge discovery, as presented in Hei-Chia et al. (2005), Kim et al. (2005), Kwon et al. (2005), Liao (2005), and King, Wise, and Clare (2004). We show that our algorithm is interesting and may lead to increasing the efficiency. Also, our algorithm has a better selection/sensibility ratio. The algorithm presented in this

* Corresponding author. Tel.: +966 454 8011; fax: +966 454 8317.
E-mail address: mtounsi@cis.psu.edu.sa (M. Tounsi).

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.09.043

paper is similar to the dot plot technique by performing a similar segment analysis in the design. We present a series of inspired algorithms based on the dot plot technique.

2. Sequences comparing

In general, we search the similar parts between two sequences according to three approaches (Fig. 1):

- The exact words (small words, some nucleotides, between two strings are strictly similar)
- Similar segment: we accept a small difference between two words
- Alignment: we accept a difference and the indels.

Some authors distinguish two types of alignments. Kruskal Kruskal (1983) calls them trace and alignment. There are two types of alignments: type I and type II (Fig. 2).

> Type I alignment is one that does not allow for adjacent gaps in opposite sequences.
> Type II alignment is one that does allow for adjacent gaps in opposite sequences.

Type I alignment can therefore be represented as a sequence of successive residue pairs that do not violate the order of residues in the sequences. In at least one of the sequences, no residue may be skipped when going to the next pair. The above example seen as a type I alignment can be expressed by listing the index-pairs for the matched residues: ((1,1),(2,2),(3,3),(5,4),(6,5),(7,9),(8,10),(9,11),(10,12),(11,13))

Example 1. (5,4) means that the 5th residue of the first sequence is matched with the 4th residue of the second sequence. This representation will be chosen to define the type I alignment. To describe a type II alignment, such a representation does not suffice. In the case of , for example the pairs (1,1) and (3,3) do not describe the alignment properly. Instead, it is also necessary to represent the gap-character, say by an index 0. One can use the notation ((1,1),(2,0),(0,2),(3,3)), which exactly describes the given type II alignment. A type II alignment is therefore defined as any sequence of pairs (following the order of residues in the sequences) that may contain either a residue or a gap-character.

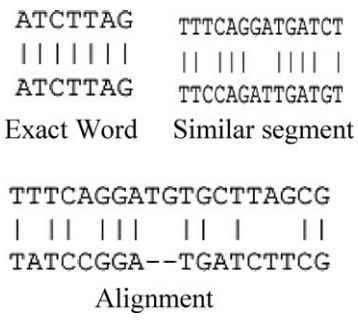

Fig. 1. Sequence comparing type.

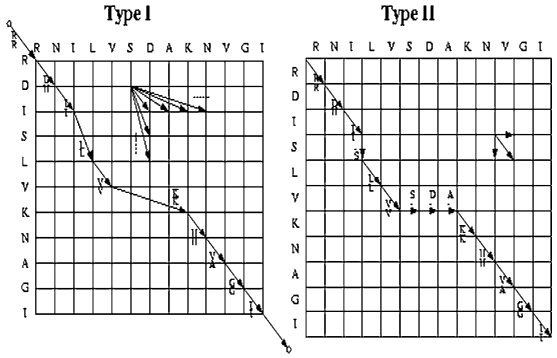

Fig. 2. Type of alignments.

Current researches on the similarities of genomes may help to find the function of such genomes (Needleman & Wunsch, 1970; Sankoff, 1972).

3. Related works

Pairing two gap-characters does not make sense in the context of sequence comparison and therefore is forbidden in the alignment. Type II alignments are richer than type I in that some type II alignments cannot be described in the type I formalism. If one restricts type II alignments to those that do not contain two adjacent columns where one has a gap-character in one sequence and the adjacent one in the other column, then each such alignment corresponds to a type I alignment. The distinction between these two kinds of alignments will be important in explaining some of the differences between algorithms for sequence comparison as they are given in the literature.

In the case of 20 amino acids, identical matched residues are not as frequent as in DNA; this result is also shown in Dayhoff, Barker, and Hunt (1983). Consequently, weighting schemes have been developed, which attribute a value to a pair of matched amino acids. One such scheme was devised by Dayhoff (Dayhoff et al., 1983) and is based on the exchange frequencies between amino acids. This 20×20 matrix attributes different positive values (ranging from $+2$ to $+17$) to exact matches and values between -8 and $+7$ for mismatches. The score of an alignment is then made up of the weights for the matching pairs in the alignment minus a penalty for every gap introduced. The gap penalty will in general be a function g of the length of the gap. The example above scored by the Dayhoff matrix would give:

$$\text{Score}\begin{pmatrix}\text{RDISLV}---\text{KNAGI}\\ \text{RNI}-\text{LVSDAKNVGI}\end{pmatrix}$$

$$\equiv 0+2+5-g(1)+0+4-g(3)+5+2+0+5+5$$

The intention behind such a scoring scheme is that the alignment, which optimizes this score, should best represent the biological similarity between two sequences. Finding such an optimal alignment is the central task in sequence comparison.