Information Sciences 181 (2011) 1454–1462

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Minimal-dot plot: "Old tale in new skin" about sequence comparison

V. Kirzhner *, S. Frenkel, A. Korol

Institute of Evolution and Department of Evolutionary and Environmental Biology, University of Haifa, Haifa 31905, Israel

ARTICLE INFO

Article history:
Received 22 April 2010
Received in revised form 13 October 2010
Accepted 16 December 2010
Available online 23 December 2010

Keywords:
Human genome
Sequence comparison
Dot-plot
Compositional spectra
Track asymmetry

ABSTRACT

The authors propose a simple version of the dot-plot scheme to be used in the case when the distances between sequence elements may take more than two values. The method is applicable, in particular, to the case of the sequences of large-length windows when the sets of distance values are continuous. The proposed technique is simple to implement and the results can produce readable maps for further analysis. To illustrate its potentialities, the method has been applied to the comparison of genomic sequences. The asymmetry in the number of direct and reverse tracks for the *Homo sapience* genome has been discovered.

1. Introduction

The first method developed to assess the similarity between biological symbol sequences was the dot-plot analysis [7,9]. Such symbol sequences are used for describing DNA or protein molecules, where four or twenty symbols are employed, respectively. In certain contexts below, relatively short symbol sequences are referred to as words.

The dot-plot procedure (in its original form) can be described as follows. Let us consider two symbol sequences, *ATTCGACG* and *TCGACAAGC*, as DNA texts written in the {A,T,C,G} alphabet. To compare these sequences, the following technique is used. The two fragments are located along the sides of an 8×9 rectangle (Fig. 1). Designate each pair of identical symbols along the two sequences by a dot and thus obtain a two-dimensional dot-plot diagram. The main output of the dot-plot method is such dot sequences that the indexes of subsequent dots along the whole sequence either both increase by a unit or one index increases by a unit, while the other one decreases by a unit. In geometrical representation, such dot sequences are parallel to the bisector of some angle of the diagram. Each of the dot sequences, which will be referred to as *tracks*, signifies a symbol-to-symbol coincidence of the fragments of the two sequences under consideration. For example, the track of length 5 marked in Fig. 1 corresponds to the word *TCGAC*, which is present in both sequences. The other marked track of length 3 corresponds to the word *AGC*, which also appears in both sequences, but in opposite orientations. Thus, the tracks located along the main bisector (which we choose to be the bisector of the top left angle) or along the secondary bisector signify the coincidence or the reverse coincidence, respectively, of the words of the two symbol sequences being compared. Correspondingly, we will distinguish between *direct* and *reverse* tracks.

The tracks can also be used as the basis for a more general comparison of two sequences, performed, e.g., in the framework of the dot-plot diagram filtration approach, proposed by Maizel and Lenk [23]. According to this approach, the two tracks that continue each other, but are separated by a short gap, can be joined together. Indeed, the omission of some dots caused by a mismatch of symbols can be attributed, in this case, to substitution mutations. It this way, the main elementary

* Corresponding author.
E-mail address: valery@esti.haifa.ac.il (V. Kirzhner).

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.12.009

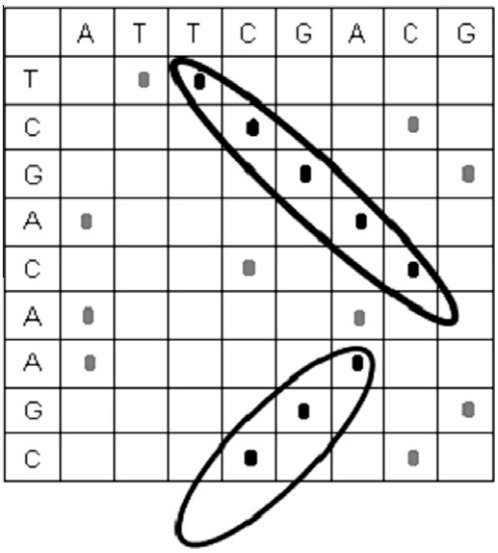

Fig. 1. Example of a dot-plot diagram. The compared sequences are located along the upper and the left sides of the rectangle. The marked tracks correspond to identical fragments of these sequences.

events of molecular evolution – substitutions, insertions, and deletions of symbols as well as transpositions of large fragments – can be detected by studying track peculiarities (see, e.g., [10,20]).

The dot-plot technique described above can be called a "classical" one. In the course of bioinformatics development, there appeared an approach to sequence comparison based on the preliminary division of original sequences into fragments (windows), which are usually of equal length. As a result, an original sequence of symbols is transformed into a sequence of windows, which, in turn, can be considered as certain formal symbols. The novelty of the approach consists in the comparison of the latter sequences through defining pair-wise distances between the formal symbols (windows).

For further consideration, it is important to emphasize that the distance functions employed in the above approach are no longer binary and may have any, even a continuous, set of values. The dot-plot method can be applied in this case, too, but the main pattern will differ from the "classical" one. Namely, for a distance scale with a relatively large set of possible values, a colored axis of dots is used in accordance with the values of the distance function at these dots. For example, in the case of an infinite set of distance values, an arbitrary discrete distance scale can be introduced. The resulting main pattern is a certain dot-plot area which corresponds to fairly close, with respect to the distance values, sets of sequence fragments. As an example of such colored dot-plots, consider Fig. 1A from [25], where two genome sequences are compared. The characteristic rectangular zones in the dot-plot area are accounted for by the pairs of fragments with almost the same distances between them. Actually, in this figure, only one track can be distinguished, namely, the line along the diagonal with the origin in the top-left corner. However, this is a trivial track since the compared sequence are almost identical.

Obviously, such formulation of the problem results in the requirement of local scale regulation (local zoom) for visualization of non-trivial tracks, similar to that used in the dot-plot method implementations described in [11,30]. At present, the dot-plot method is a part of various packages used in the field of bioinformatics [1,6,12,28,30,34], where, as a rule, it is possible to make comparisons for any distance scale.

1.1. The locally-minimal track approach

Obviously, obtaining a track in the case of a multivalent (continuous) distance function is not a visualization problem. It is rather a conceptual problem, which is connected, to a great extent, with the formal definition of the track notion. Let us consider the following simple example (Fig. 2).

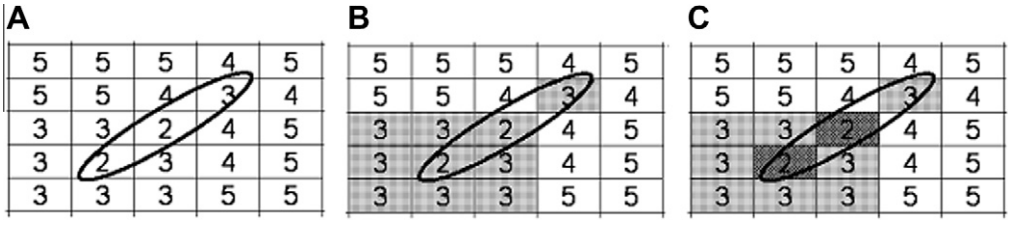

Fig. 2. The scale effect illustrated on a fragment of a dot-plot matrix.