A new sequence distance measure based on the Burrows-Wheeler Transform

Sciortino M.,Mantaci S., Restivo A., Rosone G.

Dipartimento di Matematica ed Applicazioni, Via Archirafi 34, 90123 Palermo

Motivation

The recent developments in sequencing genomes have given a new direction to bioinformatic research. Actually, the possibility of sequencing the whole genome has raised the question of discovering common features between biological sequences corresponding to different species, reflecting on common evolutionary and functional mechanisms. This reason has led researchers to look for a definition of a distance measure on sequences able to capture these common mechanisms. Most of the traditional methods for comparing biological sequences were based on the technique of sequence alignment. Nevertheless, sequence alignment considers only local mutations of the genome, therefore it is not suitable to measure events like segment rearrangements, that involve longer genomic sequences. For this reason some alignment free distance measures have been recently introduced (see [VinAl] for a survey) and most of them are based on the concept of information theory and data compression (cf. [OS, LV, CV, EMS, BCL]). Such measures are more suitable to deal with the problem of whole genome phylogeny. The intuitive idea is that the more similar two sequences are, the more effective their joint compression is than their independent compression. We introduce a new alignment free method for comparing sequences that, differently from other ones, is combinatorial by nature and does not make use of any compressor nor any information theoretic notion.

Methods

Our method is based on an extension of the Burrows-Wheeler Transform recently given in [DCC05]. The Burrows-Wheeler Transform (denoted by BWT) is a well founded mathematical transformation on sequences introduced in 1994 (cf.[bwt94]), widely used in the context of Data Compression (cf. [Man99], [Fen96]) and recently studied also from a combinatorial point of view (cf. [MRS00]). It has been remarked (cf. [CDP]) that there exists a close relation between BWT and a technique, used by Gessel and Reutenauer in [GeRe] for stating a correspondence between finite words and a set of permutations with a given cyclic structure and a given descent set. Loosely speaking, BWT is a transformation that produces a permutation BWT(w) of an input sequence w, such that we can easily retrieve w from BWT(w), i.e. the transformation is reversible, and, at the same time, BWT(w) is much easier to compress than w. The new transformation, introduced in [DCC05] and denoted by E, works analogously to BWT, but takes as input a multiset S of sequences. Such a transformation has been also inspired by the above mentioned technique of Gessel and Reutenauer. A fundamental step in the computation of E(S) consists in sorting all the symbols occurring in the sequences in S, using as a sort key for each symbol its context, i.e. the segment following it in the sequence. Such a step is realized by sorting the conjugates of all sequences in S according to an order relation different from the lexicographical one. We use the transformation E in order to define a new method for comparing sequences. Such a method is based on the following idea: when E is applied to S={u,v}, if the same segment s occurs both in u and v, then the conjugates of u and v starting by s are likely to be close in the sorted list of conjugates. This implies that the greater is the number of segments shared by u and v, the greater is the mixing of the conjugates of u and v in the sorted list. The

comparison method based on transformation E will measure how similar u and v are, by taking into account how their conjugates are mixed. This intuition has different possible formalizations. We propose a distance measure that computes the number of alternations in the above list between the conjugates of u and those of v.

Results

The computation of our distance is simple and efficient, and it is particularly advantageous in the case of a multiple sequences comparison. We show also the validity of the method by applying the distance introduced to a data set for the whole mitochondrial genome phylogeny problem. The results we have obtained are very close to the ones derived, with other approaches, in most of the papers in which the considered species are the same. Remark that, however, our goal is not to confirm or refute previous phylogenetic studies but rather to introduce new methods and tools to the comparative genomics research community.

Contact email: mari@math.unipa.it

Supplementary Information: [BCL] D. Benedetto, E. Caglioti, and V. Loreto. Zipping out relevant information. Computing in Science and Engineering, pages 80-85, 2003. [bwt94] M. Burrows and D.J. Wheeler. A block sorting data compression algorithm. Technical report, DIGITAL System Research Center, 1994. [CV] R. Cilibrasi and P. Vitanyi. Clustering by compression. IEEE Trans. Information Theory (sub- mitted), 2005. [CDP] M. Crochemore, J. Désarménien, and D. Perrin. A note on the Burrows-Wheeler transformation. Theoret. Comput. Sci., 332:567-572, 2005. [EMS] F. Ergun, S. Muthukrishnan, and C. Sahinalp. Comparing sequences with segment rearrangements. Lecture Notes in Comput. Sci, pages 183-194, 2003. Proc. of the FSTTCS'03, Bombay, India. [Fen96] P. Fenwick. The Burrows-Wheeler transform for block sorting text compression: principles and improvements. The Computer Journal, 39(9):731-740, 1996. [GeRe] I. M. Gessel and C. Reutenauer. Counting permutations with given cycle structure and descent set. J. Combin. Theory Ser. A, 64(2):189-215, 1993. [LV] M. Li, X. Chen, X. Li, B. Ma, and P. Vitanyi. The similarity metric. IEEE Trans. Inform. Th., 12(5):3250-3264, 2004. [MRS00] S. Mantaci, A. Restivo, and M. Sciortino. Burrows-Wheeler transform and Sturmian words. Informat. Proc. Lett., 86:241-246, 2003. [DCC05] S. Mantaci, A. Restivo, and M. Sciortino. An extension of the Burrows-Wheeler Transform to k words. Technical Report 267, University of Palermo, Dipartimento di Matematica ed Appl., December 2004. Extended abstract in Data Compression Conference 2005. [Man99] G. Manzini. The Burrows-Wheeler transform: Theory and practice. In Proc. of the 24th International Symposium on Mathematical Foundations of Computer Science (MFCS '99), pages 34-47. Springer-Verlag LNCS n. 1672, 1999. [OS] H.H. Otu and K. Sayood. A new sequence distance measure for phylogenetic tree construction. Bioinformatics, 19(16):2122-2130, 2003. [VinAl] S. Vinga and J. Almeida. Alignment-free sequence comparison - a review. Bioinformatics, 19(4):513-523, 2003.