Journal of Theoretical Biology 275 (2011) 52–58

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Genome analysis with distance to the nearest dissimilar nucleotide

Vera Afreixo [a,*], Carlos A.C. Bastos [b,c], Armando J. Pinho [b,c], Sara P. Garcia [b], Paulo J.S.G. Ferreira [b,c]

[a] Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal
[b] Signal Processing Lab, IEETA, University of Aveiro, 3810-193 Aveiro, Portugal
[c] Department of Electronics, Telecommunications and Informatics, University of Aveiro, 3810-193 Aveiro, Portugal

ARTICLE INFO

Article history:
Received 6 September 2010
Received in revised form
24 January 2011
Accepted 24 January 2011
Available online 2 February 2011

Keywords:
Alignment-free genome comparison
Inter-nucleotide distances
Nearest dissimilar distances
DNA sequences

ABSTRACT

DNA may be represented by sequences of four symbols, but it is often useful to convert those symbols into real or complex numbers for further analysis. Several mapping schemes have been used in the past, but most of them seem to be unrelated to any intrinsic characteristic of DNA. The objective of this work was to study a mapping scheme that is directly related to DNA characteristics, and that could be useful in discriminating between different species.

Recently, we have proposed a methodology based on the inter-nucleotide distance, which proved to contribute to the discrimination among species. In this paper, we introduce a new distance, the distance to the nearest dissimilar nucleotide, which is the distance of a nucleotide to first occurrence of a different nucleotide. This distance is related to the repetition structure of single nucleotides. Using the information resulting from the concatenation of the distance to the nearest dissimilar and the inter-nucleotide distance, we found that this new distance brings additional discriminative capabilities. This suggests that the distance to the nearest dissimilar nucleotide might contribute with useful information about the evolution of the species.

1. Introduction

DNA sequences have been converted to numerical signals using different mappings. A commonly used mapping is to consider binary sequences that describe the position of each symbol (Voss, 1992). The binary representation is certainly one of the earliest and one of the most popular mappings of DNA. However, several other different mappings have been proposed (see for example Silverman and Linsker, 1986; Jeffrey, 1990; Zhang and Zhang, 1994; Buldyrev et al., 1995; Anastassiou, 2001; Cristea, 2003; Ning et al., 2003; Brodzik and Peters, 2005; Liao et al., 2005; Akhtar et al., 2007; Randic, 2008; Nair and Mahalakshmi, 2005; Afreixo et al., 2009).

Some of the mappings used in DNA processing do not have a simple numerical interpretation and others do not have biological motivation. Also, some of the representations are not reversible and do not take into account the sequence structure. Currently, there is no ideal mapping to analyze every type of correlation in DNA sequences.

In a previous work, we explored the inter-nucleotide (IN) distance, the distance to the first occurrence of the same symbol, to perform a comparative analysis between species (Afreixo et al., 2009). In this work, we present a new DNA numerical profile and a new mapping to explore the correlation structure of DNA: the distance to the nearest dissimilar (ND) nucleotide. This representation converts any DNA sequence into a unique numerical sequence with lower length, where each number represents the distance of a symbol to the next occurrence of a different symbol. We introduced also four sequences, one for each nucleotide, to represent the ND distances. This allows to perform comparative analysis between the behavior of the four nucleotides distance sequences and the global sequence.

From the perspective of molecular evolution, DNA sequences may reflect both the results of random mutation and selective evolution. One should subtract the random background from the simple counting result in order to highlight the contribution of selective evolution (Qi et al., 2004; Ding et al., 2010). Therefore, we present an analysis of the relative error to highlight the contribution of selective evolution of the DNA of each species. This residual analysis may be used, for example, to perform multiple organism comparisons.

Phylogenetic trees reproduce the evolutionary tree that represents the historical relationships between the species. Recent phylogenetic tree algorithms use nucleotide sequences. Typically, these trees are constructed with multiple sequence alignment (Hodge and Cope, 2000), which is a computationally demanding task. Recently, alignment-free methods have been proposed and present some advantages over multiple sequences alignment

* Corresponding author.
E-mail address: vera@ua.pt (V. Afreixo).

0022-5193/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.01.038

methods (see for example Sims et al., 2009; Vinga and Almeida, 2003). The distance that we address in this paper seems to possess discriminating properties that might be helpful in inferring phylogenies. We do believe that this claim is supported by the examples of trees that are provided. However, we also believe that, by itself, this distance measure does not convey all necessary information for building phylogenies. Instead, it should be regarded as potentially useful for working in cooperation and complementing other measures.

2. Materials and methods

2.1. DNA sequences

In this study, we used the complete DNA sequences of 29 species: 27 were obtained from the National Center for Biotechnology Information (NCBI) (ftp://ftp.ncbi.nih.gov/genomes/); *Populus trichocarpa* (California poplar) obtained from the Joint Genome Institute (http://genome.jgi-psf.org/) and *Xenopus tropicalis* (Western clawed frog) from Xenbase (http://www.xenbase.org/). The species used in this work are listed in Table 1.

2.2. Distance to the nearest dissimilar

Consider the alphabet $\mathcal{A}=\{A,C,G,T\}$ and let $s=(s_k)_{k\in\{1,\ldots,N\}}$ be a symbolic sequence defined in $\mathcal{A}$. Consider a numerical sequence, w^x, that represents the distance to the nearest dissimilar of symbol $x\in\mathcal{A}$. As an example, the four ND distance sequences for the short DNA fragment CAAACCGTTTAAGTAACAGGGATATTGGCCC are

$$w^A=(3,2,2,1,1,1),\quad w^C=(1,2,1,3),$$

$$w^G=(1,1,3,2),\quad w^T=(3,1,1,2).$$

Table 1
List of DNA builds used for each species.

Species	Reference
Homo sapiens (human)	Build 36.3
Pan troglodytes (chimpanzee)	Build 2.1
Macaca mulatta (Rhesus macaque)	Build 1.1
Mus musculus (mouse)	Build 37.1
Rattus norvegicus (brown rat)	Build 4.1
Equus caballus (horse)	Build 2.1
Cannis familiaris (dog)	Build 2.1
Bos taurus (cow)	Build 4.1
Ornithorhynchus anatinus (platypus)	Build 1.1
Monodelphis domestica (opossum)	Build 2
Gallus gallus (chicken)	Build 2.1
Xenopus tropicalis (Western clawed frog)	Build 4.1
Danio rerio (zebrafish)	Build 3.1
Apis mellifera (honey bee)	Build 4.1
Caenorhabditis elegans (nematode)	NC003279
Vitis vinifera (grape vine)	Build 1.1
Populus trichocarpa (California poplar)	Build 1.0
Arabidopsis thaliana (thale cress)	AGI 7.2
Saccharomyces cerevisiae str.S228C (budding yeast)	SGD 1
Schizosaccharomyces pombe (fission yeast)	Build 1.1
Dictyostelium discoideum str.AX4 (amoeba)	Build 2.1
*Plasmodium falciparum*3D7 (protozoon)	Build 2.1
Escherichia coli str.K12 substr.MG1655 (bacterium)	NC000913
Bacillus subtilis str.168 (bacterium)	NC000964
Chlamydia trachomatis str.D/UW-3/CX (bacterium)	NC000117
Mycoplasma genitalium str.G37 (bacterium)	NC000908
Streptococcus mutans str.UA159 (bacterium)	NC004350
Streptococcus pneumoniae str.ATCC 700669 (bacterium)	NC011900
Aeropyrum pernix str.K1 (archaeota)	NC000854

The global sequence of ND distances for this example is

$$w=(1,3,2,1,3,2,1,1,2,1,1,3,1,1,1,2,2,3).$$

Note that the ND distance of each nucleotide corresponds to the repeat length of that nucleotide.

Algorithm 1. Computation of w for sequence s.

```
p := 1
p' := 1
while p' ≤ N do
  i := 0
  while s_{p'} = s_{p'+i} do
    i := i+1
  end while
  w_p := i
  p := p+1
  p' := p'+i
end while
```

Algorithm 2. Computation of w^x with $x\in\mathcal{A}$.

```
p := 1
p' := 1
while p' ≤ N do
  i := 0
  while s_{p'+i} = 'x' do
    i := i+1
  end while
  if i=0 then
    p' := p'+1
  else
    w_p := i
    p := p+1
    p' := p'+i
  end if
end while
```

Note that

$$\sum_{i=1}^{L} w_i = N, \tag{1}$$

where L is the length of w. Let n_i be the number of occurrences of ND distance i, then

$$\sum_{i=1}^{K} n_i i = N, \tag{2}$$

where K is the largest ND distance present in the data sequence. The mean distance is

$$\frac{\sum_{i=1}^{L} w_i}{L} = \frac{N}{L}. \tag{3}$$

2.2.1. Relationship between ND and IN distances

In general, the ND distance distribution complements the information that is accumulated in the first IN distance. We recall that the IN distance is the distance to the first occurrence of the same nucleotide. Let n'_i be the absolute frequency of the ith IN distance,

$$n'_1 = \sum_{i=2}^{K} n_i(i-1)+\delta = N-L+\delta \tag{4}$$