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,

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

and

$$\sum_{i=2}^{K'} n'_i = \sum_{i=1}^{K} n_i - \delta = L - \delta, \tag{5}$$

where

$$\delta = \begin{cases} 0, & s_1 \neq s_N, \\ 1, & s_1 = s_N, \end{cases}$$

and K' is the largest IN distance present in the data sequence. The last equation can be described by the following sentence: the number of IN distances greater than one is equal to the total number of ND distances. As an illustration, the computed values for the previous DNA fragment example are $N=31$, $n'_1 = 14$, $L=18$, $\delta = 1$.

2.2.2. Comparison with an independent random process

In order to calculate some statistical properties of various genomes, we will study the characteristics of the ND distance distribution.

Consider p^A, p^C, p^G and p^T the occurrence probabilities of nucleotides A, C, G and T, respectively. If the nucleotide sequences were generated by an independent and identically distributed (i.i.d.) random process, then each of the distances, W^x, would follow a geometric distribution of parameter $1-p^x$. In fact, the probability distribution of the distance to the nearest dissimilar for symbol x is

$$f^x(k) = P(W^x = k) = P(W = k|x) = (1-p^x)(p^x)^{k-1}, \quad k = 1,2,\ldots \tag{6}$$

Fig. 1 shows the measured and the reference distributions of the ND distance sequences for the *Homo sapiens* complete genome. Although the ND distance distribution from DNA shows an exponential behavior, it differs from the reference distribution. This is expected, since the reference distribution was established under the assumption of an i.i.d. random process (with constant nucleotide relative frequencies estimated from the DNA sequence).

We performed the chi-square goodness of fit test with the result that the distributions are significantly different (all p values $< 10^{-4}$).

In order to measure the differences between the observed and the reference distributions we used the Kullback–Leibler divergence.

Assuming that the DNA sequence was generated by an independent random process with constant parameters, the corresponding global ND distance sequence distribution is given by

$$f(k) = P(W = k) = \sum_{x \in \mathcal{A}} P(W = k|x)P(x) = \sum_{x \in \mathcal{A}} (1-p^x)(p^x)^k. \tag{7}$$

2.3. Numerical procedure

The histograms of the ND distance sequences were computed for each nucleotide and also for the global sequence. For large genomes and for convenience, the sequences were divided into blocks of 500 000 symbols. This procedure does not influence the total nucleotide counts. For eukaryote genomes, the chromosomes were processed separately and the resulting distance histograms were stored separately. All the symbols in the sequence that did not correspond to one of the four nucleotides were removed from the sequences before further processing.

We setup to investigate how similar (or different) are the observed and the reference distributions of:

- the four nucleotides of *Homo sapiens*;
- the chromosomes of *Homo sapiens*;
- various species.

In order to facilitate the visual comparison of the various distance distributions with the theoretical one and to subtract the random background, we used the relative error, as given by

$$r(k) = \frac{f_o(k) - f(k)}{f_o(k)}, \tag{8}$$

where $f_o(k)$ is the observed relative frequency of the distance k and $f(k)$ is the relative frequency of the reference distribution.

For the prokaryote species the values of the relative error of the ND distance were truncated to values between -1 and 1. We also used the IN distance relative error used in Afreixo et al. (2009) and

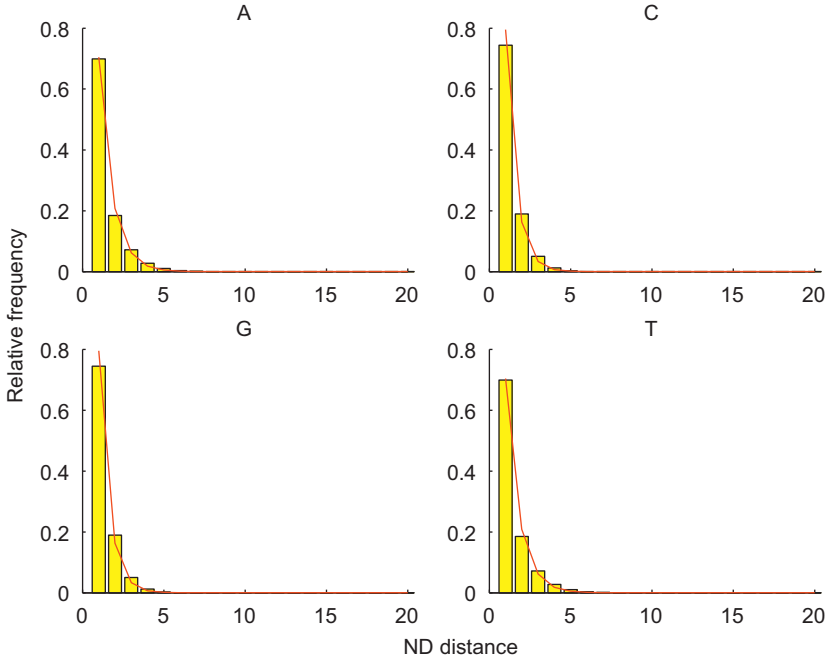

Fig. 1. Distribution of the four nucleotide ND distance sequences for the *Homo sapiens* genome. The histogram is from the observed distances and the solid line shows the reference distribution with parameters estimated from the data.

we concatenated the first 100 IN distances and the first 20 ND distance relative errors. Only the first 20 ND distances were used, because for larger distances the distributions become sparse. We have found that the limitation to the first 20 distances, which was carried out in all numerical experiments described in this paper, provides an adequate compromise between the information content and the vector length. Note that 792 is the largest ND distance for all the studied species and that the maximum proportion of the ND distances above 20 is 0.13%.

3. Results

3.1. Distances analysis

We compared the observed ND distance distributions of all *Homo sapiens* chromosomes and we obtained very small Kullback–Leibler divergence (<0.0008). We have also compared, using the Kullback–Leibler divergence, the ND distance distribution of the four nucleotides of the *Homo sapiens* complete genome. The results of the comparative tests are shown in Table 2, showing that the ND distance distribution of nucleotide A is closer to that of nucleotide T than to the other two nucleotides, and the ND distance distribution of C is closer to that of nucleotide G than to the other two nucleotides. Notice that the DNA complementary sequence was not used in the computation of the distributions. These identical distance distributions for the nucleotides A/T and C/G are present in all the human chromosomes and also in the genome of all the other species used in this work. This may be explained by the Chargaff's second parity rule and the extension of this rule (see for example Qi and Cuticchia, 2001; Albrecht-Buehler, 2006, 2007).

We used the relative error, as defined in (8), to compare the ND distance distributions, both for nucleotides and global sequences, with the reference distributions. This corresponds to removing the contribution of the random background to the ND distance distribution (Qi et al., 2004).

Fig. 2 shows the relative error for the ND distance for each nucleotide of the *Homo sapiens* genome and Fig. 3 the relative error for the global ND distance sequence. A relative error close to zero means that the observed frequency is similar to that of the reference distribution. Values close to one imply observed distances much more frequent than in the reference distribution.

For the global ND distance (Fig. 3) of the human genome, the first two distances have a lower frequency than the one

Table 2
Kullback–Leibler divergence values between the ND distance distribution of the four nucleotides in the *Homo sapiens* genome.

	A	C	G	T
A	0.00	0.04	0.04	0.00
C	0.03	0.00	0.00	0.03
G	0.03	0.00	0.00	0.03
T	0.00	0.04	0.04	0.00

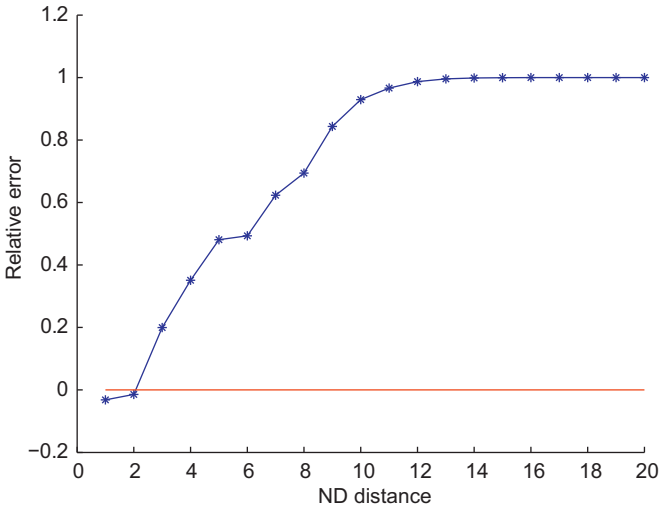

Fig. 3. Relative error for the global ND distance distribution in the complete genome of *Homo sapiens* (first 20 distances).

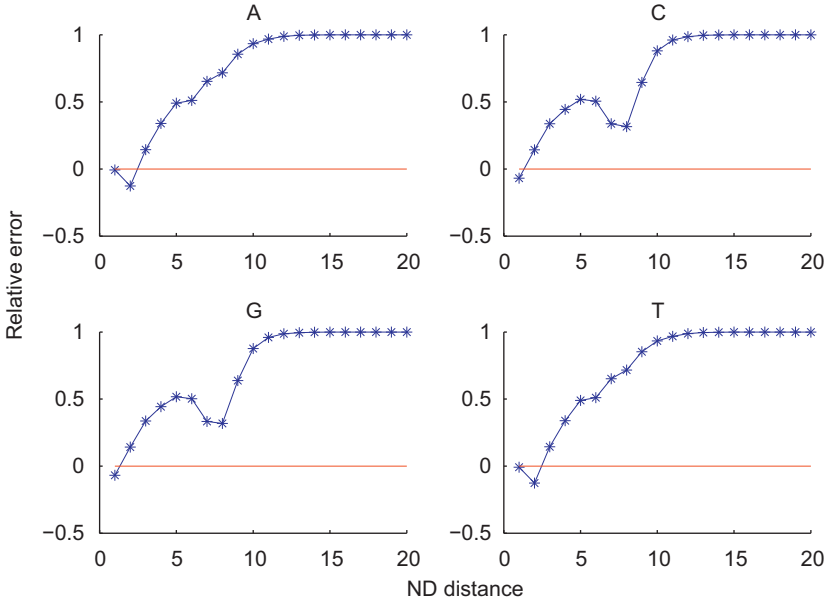

Fig. 2. Relative error of the four ND distances in the complete genome of *Homo sapiens* (first 20 distances).

corresponding to random sequences, whereas the other distances have higher frequencies. The relative frequencies of distances above two are higher than in the reference distribution. The repetition of three or more nucleotides is more frequent in the human genome than in a random sequence. This pervasive existence of repeats is a well known characteristic of the human genome.

3.2. Analysis of multiple organisms

The ND distance relative error vectors of each complete genome may be used as a genomic signature that identifies each species, thus allowing the comparison of species (e.g. by building phylogenetic trees). We used the UPGMA programs in the PHYLIP package (http://evolution.genetics.washington.edu/phylip.html) to build the phylogenetic trees (the similarity matrix was computed using the Euclidean distance).

The comparison of the various phylogenetic trees was carried out with TOPD/FMTS software (Puigbo et al., 2007) and using the split distance (a low split distance value is synonymous of a high number of common branches between the two trees).

Hierarchical clustering was applied to a matrix composed by the first 20 ND distances for all the species used in this study. Fig. 4 shows the phylogenetic tree for all the species in this study. The phylogenetic tree (obtained with complete linkage) displays a first branching between prokaryotes and eukaryotes, showing most of the mammals together. Moreover, all vertebrates are in the same cluster, the only two non-vertebrates in that cluster are *A. mellifera* and *V. vinifera*. Fungi are also grouped together. The bacteria are displayed close together and the archaeota further apart.

With the ND distance relative errors we obtained an interesting tree that shows some of the evolutive features. However, we could point several unexpected clusters which may be due to the use of small vectors (20 elements) to characterize the species. In parallel, we have developed methodologies based on the IN distances that can be complemented with information from the ND distances. These two methodologies are complementary: both describe distances between nucleotides, but in two different ways (see Section 2.2.1). Fig. 5 shows the phylogenetic tree constructed with a species vector that corresponds to the concatenation of the IN (removing the first IN distance) and ND relative errors. We compare the results obtained by the IN–ND concatenation method with the IN, ND and random methods in Table 3. The concatenation method obtains a tree topology similar to the IN method, but the changes introduced produced a better phylogenetic tree, including a higher fungi and archaeota differentiation. The differences between Figs. 4 and 5 may arise from the use of larger vectors in Fig. 5 (119 elements), thus conveying more

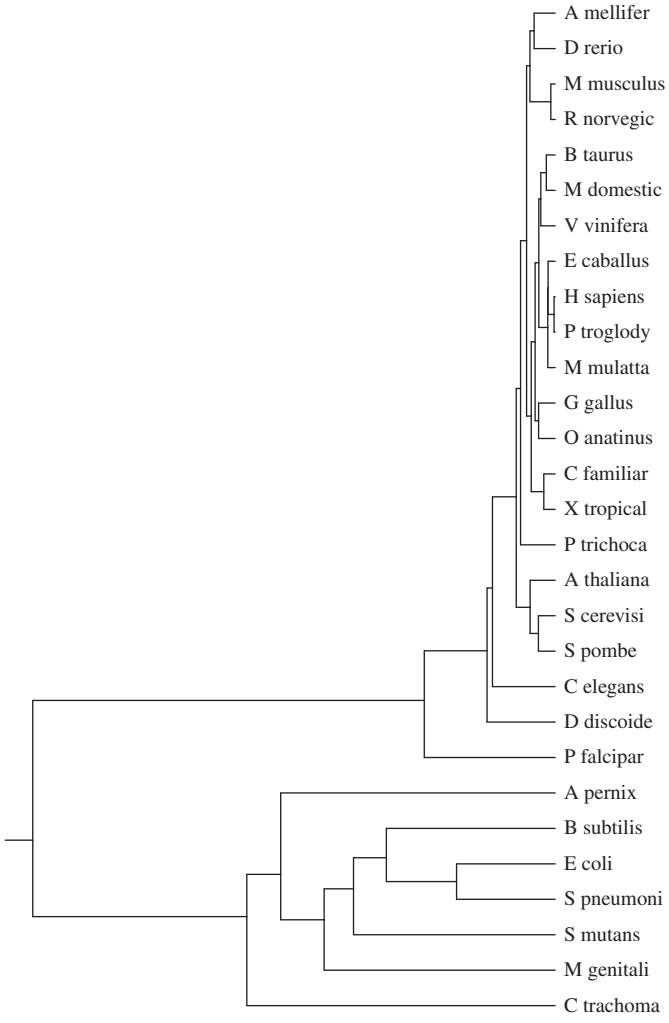

Fig. 4. Phylogenetic tree with the species used in this study, using the 20 ND distances.

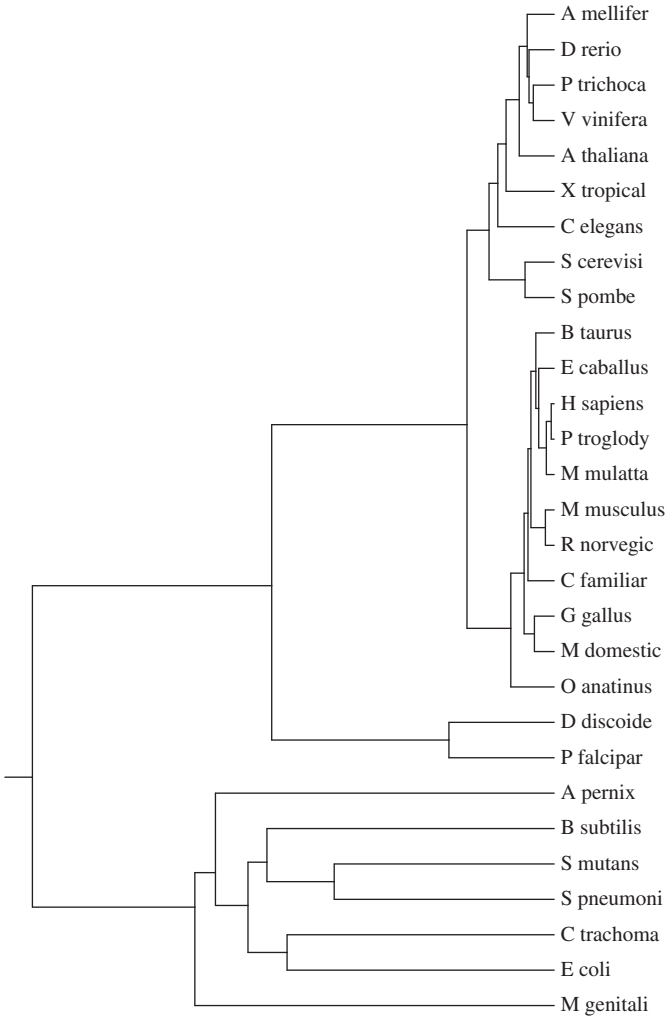

Fig. 5. Phylogenetic tree with the species used in this study, using the IN–ND concatenation method.

Table 3
Normalized split distances of the 29 species.

	ND	IN	Random
IN–ND	0.7692	0.1154	0.9888

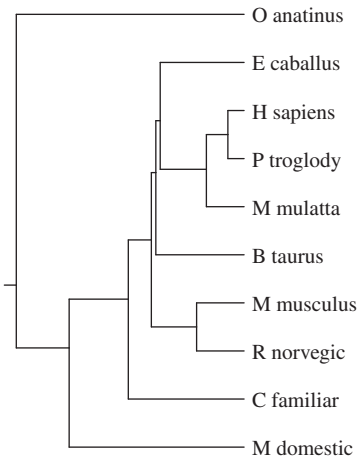

Fig. 6. Phylogenetic tree with the 10 mammals used in this study, using the IN–ND concatenation method.

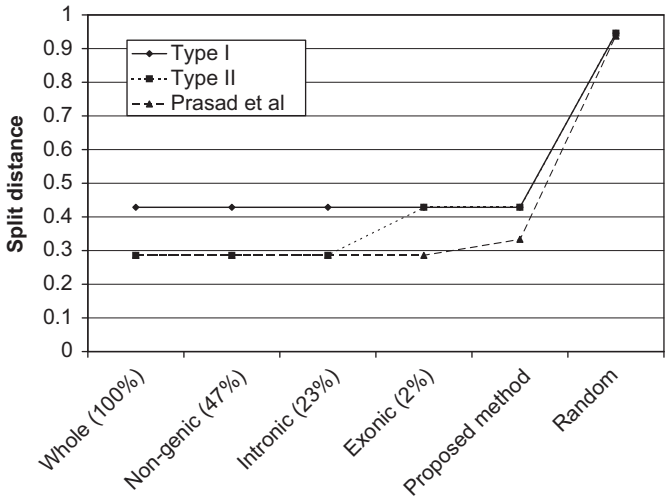

Fig. 7. Normalized split distances between phylogenetic trees constructed with multiple sequence alignment and free-alignment methods.

information than the ND distance vectors (Fig. 4). Moreover, as already mentioned, the ND distance complements the information accumulated in the first IN distance.

In order to compare alignment and alignment-free algorithms to construct phylogenetic trees (Sims et al., 2009) used 10 mammals. Fig. 6 shows the phylogenetic tree constructed for the same set of mammals as used by Sims et al. (2009), using the IN–ND concatenation method. The topology of the phylogenetic tree obtained using the IN–ND concatenation method is similar to the phylogenetic trees presented by Sims et al. (2009) for alignment based algorithms (Type I, Type II, Prasad Arjun et al., 2008).

Fig. 7 shows the normalized split distance between alignment algorithms and alignment-free algorithms (Sims's methods and our methods). The normalized split distance is the number of splits in one tree, but not present in the other, divided by the total number of splits in the tree. The IN–ND concatenation method shows a small split distance value (similar to those obtained by Sims et al., 2009) which reveals that the IN–ND concatenation contains relevant information about the phylogenetic evolution.

4. Conclusion

The ND distance mapping contains information about nucleotide repetition structure in the DNA sequence and characterizes the lengths of the single nucleotide repeats. An interesting feature of the ND distance approach is the strong similarity found for the A/T and C/G nucleotides. This may be explained by the existence of inverted repeats and by the second Chargaff parity rule and its extensions (Qi and Cuticchia, 2001; Albrecht-Buehler, 2006, 2007).

Since the ND distance vectors are small (around 20 elements) and the IN distance characterizes the distance between groups of single nucleotide repeats, we concatenated the IN and ND distances in order to obtain a vector that better differentiates species.

The results obtained in this work suggest that, for the addressed species, there is a genetic signature, a vector with the concatenation of the relative error of the IN and ND distances, that is a distinguishing characteristic of each species.

Acknowledgments

This work was supported in part by the FCT (Fundação para a Ciência e Tecnologia). S.P. Garcia acknowledges funding from the European Social Fund and the Portuguese Ministry of Science, Technology and Higher Education.

References

Afreixo, V., Bastos, C.A.C., Pinho, A.J., Garcia, S.P., Ferreira, P.J.S.G., 2009. Genome analysis with inter-nucleotide distances. Bioinformatics 25, 3064–3070.
Akhtar, M., Epps, J., Ambikairajah, E., 2007. On DNA numerical representation for period-3 based exon prediction. In: Fifth International Workshop on Genomic Signal Processing and Statistics.
Albrecht-Buehler, G., 2006. Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. Proceedings of the National Academy of Sciences of the United States of America 103, 17828–17833.
Albrecht-Buehler, G., 2007. Inversions and inverted transpositions as the basis for an almost universal "format" of genome sequences. Genomics 90, 297–305.
Anastassiou, D., 2001. Genomic signal processing. IEEE Signal Processing Magazine 18, 8–20.
Brodzik, A.K., Peters, O., 2005. Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences. In: Proceedings of IEEE ICASSP, pp. 373–376.
Buldyrev, S., Goldberger, A., Havlin, S., Mantegna, R., Matsa, M., 1995. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. Physical Review E 51, 5084–5091.
Cristea, P.D., 2003. Large scale features in DNA genomic signals. Signal Processing 83, 871–888.
Ding, S., Dai, Q., Liu, H., Wang, T., 2010. A simple feature representation vector for phylogenetic analysis of DNA sequences. Journal of Theoretical Biology 265, 618–623.
Hodge, T., Cope, M.J.T.V., 2000. A myosin family tree. Journal of Cell Science 113, 3353–3354.
Jeffrey, H.J., 1990. Chaos game representation of gene structure. Nucleic Acids Research 18, 2163–2170.
Liao, B., Tan, M., Ding, K., 2005. Application of 2-d graphical representation of DNA sequence. Chemical Physics Letters 401, 196–199.
Nair, A.S.S., Mahalakshmi, T., 2005. Visualization of genomic data using inter-nucleotide distance signals. In: Proceedings of IEEE Genomic Signal Processing.
Ning, J., Moore, N., Nelson, J.C., 2003. Preliminary wavelet analysis of genomic sequences. In: Proceedings of IEEE Bioinformatics Conference, pp. 509–510.
Prasad Arjun, B., Allard Marc, W., Green Eric, D., 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. Journal of Molecular Evolution 25, 1795–1808.
Puigbo, P., Garcia-Vallve, S., McInerney, J.O., 2007. TOPD/FMTS: a new software to compare phylogenetic trees. Bioinformatics 23, 1556–1558.
Qi, D., Cuticchia, A.J., 2001. Compositional symmetries in complete genomes. Bioinformatics 17, 557–559.

Qi, J., Wang, B., Hao, B.I., 2004. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. Journal of Molecular Evolution 58, 1–11.

Randic, M., 2008. Another look at the chaos-game representation of DNA. Chemical Physics Letters 456, 84–88.

Silverman, B.D., Linsker, R., 1986. A measure of DNA periodicity. Journal of Theoretical Biology 118, 295–300.

Sims, G.E., Jun, S.R., Wu, G.A., Kim, S.H., 2009. Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. Proceedings of the National Academy of Sciences of the United States of America 106, 17077–17082.

Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison—a review. Bioinformatics 19, 513–523.

Voss, R.F., 1992. Evolution of long-rang fractal correlations and 1/*f* noise in DNA base sequences. Physical Review Letters 68, 3805–3808.

Zhang, R., Zhang, C.T., 1994. Z curves, an intuitive tool for visualising and analysing the DNA sequences. Journal of Biomolecular Structure and Dynamics 11, 767–782.