Genomic Signature Is Preserved in Short DNA Fragments

Patrick Deschavanne deschavanne@imed.jussieu.fr
Alain Giron giron@imed.jussieu.fr
Joseph Vilain
Alexandra Vaury
Bernard Fertil fertil@imed.jussieu.fr

INSERM U 494, CHU Pitié-Salpêtrière, 91 bvd de l'hôpital,
75634 Paris cedex 13 - France

Keywords: genome, genomic signature, sequence length, word length, parametric images

1 Introduction

The recent availability of complete genomes opens a new field of research devoted to the general analysis of their global structure without regard to gene interpretation. The Chaos Game Representation of DNA sequence [2], when modified to allow for quantification, displays the whole set of frequencies of words found in a given genomic sequence under the form of images where the value of each pixel corresponds to the frequency of a specific word in the sequence. The specific image, which is associated with each species, may therefore be considered as a genomic signature [1]. The question arises about the amount of species-specific information (in terms of word usage) remaining in short DNA fragments?

2 Method and Results

Sixteen genomes were selected to evaluate the variability of word frequency distribution along the genomes. Series of (100, 25, 10, 5, 1, 0.5 and 0.1 kb) non-overlapping DNA fragments were randomly chosen for each genome. The corresponding images were generated. In general, patterns are clearly maintained when length of fragments decreases. However, a blurring effect may develop for fragments below 5 kb. In order to test to what extent short DNA fragments share properties of the species they derive from, a unsupervised clustering technique has been used to group fragments as a function of the characteristics of their word distribution. The grouping of fragments obtained can be compared to their origin, using an index which quantifies the proportion of well-classified fragments. Obviously, variability of word frequencies within genome decreases when the size of fragment increases. In the meantime, the length of words used may also be an interesting factor to consider. Longer words may be more species-specific although their frequency along the genome may be more variable. As expected, the proportion of well-classified fragments roughly decreases with the size of the fragment, whatever the length of the words considered in the calculation (Fig. 1). However, a surprising high proportion of 1kb fragments is properly classified. Similarly, the proportion of well-classified fragments increases with the length of the words, whatever the size of the fragment. While a perfect classification is already achieved using 3 letter-words and 100 kb fragments, the analysis of longer words is required to get good results when small fragments are considered. The usage of long words (5-letters) seems to be quite constant along each genome while being also very species-specific.

We have shown that word usage in short fragments of genomic DNA (as short as 1 kb) is similar to that of the whole genome, thus providing a strong support to the concept of genomic signature. As a consequence, small DNA fragments can be used for classification purposes.

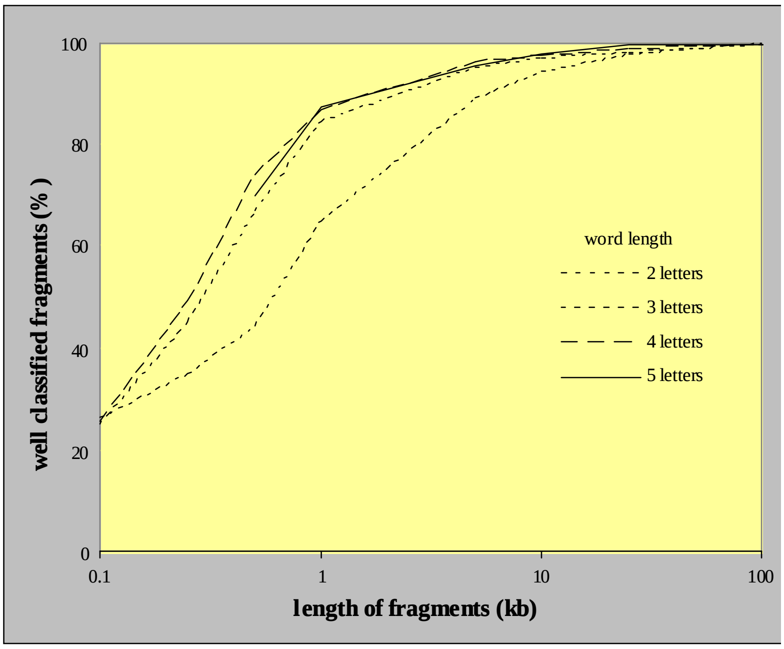

Figure 1: Classification of DNA fragments (100 fragments per species, except for 100 kb fragments) from 16 species. Percentage of well-classified fragments is expressed as a function of length of fragments and size of words.

References

[1] Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G. and Fertil, B., Genomic signature: characterization and classification of species assessed by chaos game representation of sequences, *Mol. Biol. Evol.*, 16(10):1391–1399, 1999.

[2] Jeffrey, H.J., Chaos game representation of gene structure, *Nucleic Acids Res.*, 18(8):2163–2170, 1990.