Physica A 414 (2014) 216–226

Contents lists available at ScienceDirect

Physica A

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

A genome signature derived from the interplay of word frequencies and symbol correlations

Simon Möller [a], Heike Hameister [b], Marc-Thorsten Hütt [a,*]

[a] Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany
[b] Merck Serono GmbH, Alsfelder Strasse 17, 64289 Darmstadt, Germany

HIGHLIGHTS

- Statistical properties of DNA sequences come from a wide range of biological processes.
- We derive a novel genome signature combining word frequencies and symbol correlations.
- Our genome signature performs better in a metagenomics clustering example.
- It reveals strong differences in eukaryotic microsatellite distribution.

ARTICLE INFO

Article history:
Received 3 October 2013
Received in revised form 11 March 2014
Available online 23 July 2014

Keywords:
DNA sequences
Symbol correlations
Autoregressive processes
Phylogeny
Genome evolution
Markov processes

ABSTRACT

Genome signatures are statistical properties of DNA sequences that provide information on the underlying species. It is not understood, how such species-discriminating statistical properties arise from processes of genome evolution and from functional properties of the DNA. Investigating the interplay of different genome signatures can contribute to this understanding. Here we analyze the statistical dependences of two such genome signatures: word frequencies and symbol correlations at short and intermediate distances.

We formulate a statistical model of word frequencies in DNA sequences based on the observed symbol correlations and show that deviations of word counts from this correlation-based null model serve as a new genome signature. This signature (i) performs better in sorting DNA sequence segments according to their species origin and (ii) reveals unexpected species differences in the composition of microsatellites, an important class of repetitive DNA.

While the first observation is a typical task in metagenomics projects and therefore an important benchmark for a genome signature, the latter suggests strong species differences in the biological mechanisms of genome evolution.

On a more general level, our results highlight that the choice of null model (here: word abundances computed via symbol correlations rather than shorter word counts) substantially affects the interpretation of such statistical signals.

1. Introduction

Null models are of fundamental importance in quantitative biology, as they provide the opportunity to compare statistical properties of biological systems with random expectations (provided by the null model). Deviations from randomness can often be functionally interpreted, as the signature of a biological process or an evolutionary shaping of the biological system.

* Corresponding author. Tel.: +49 421 200 3238.
E-mail address: m.huett@jacobs-university.de (M.-T. Hütt).

http://dx.doi.org/10.1016/j.physa.2014.07.048
0378-4371/

The details of the null model (i.e. the specific formulation of random systems implemented to generate the reference data) then determine the scope and reliability of the functional interpretation.

Even though this is a well known statistical problem, the recent literature in bioinformatics, systems biology and statistical physics contains a wide range of null-model debates—from metabolic networks [1] and network motifs [2–7] to sequence alignment [8] and gene regulation [9,10].

Statistical properties of genomes have always served as particularly fascinating examples of such null model comparisons and the subsequent functional and biological interpretation of the results. Peng et al. [11] demonstrated for example that the mosaic structure of DNA contributes to long-range correlations but is not sufficient to explain it. Based on their previous work [12] (see also Refs. [13,14]) on long-range correlations as a universality class, Messer et al. [8] explored the influence of the choice of null model on sequence alignment scores. They compared a model in which the symbols are drawn independently from an identical distribution to a model that incorporates long-range correlations and found that the score distribution differed strongly between the two. Herzel et al. [15] discussed the influence of protein structure and DNA folding on symbol correlations, relating periodicities with DNA supercoiling for bacterial genomes.

In the past, both correlation-based and word-based genome signatures have been employed to analyze the species information contained in DNA sequences.

Josse et al. [16] discovered by biochemical experiments that different species have a different distribution of dinucleotides (i.e. words with length 2 also called 2-mer). To compare these distributions between species with different nucleotide frequencies Subak-Sharpe et al. [17] introduced a normalization for which the observed dinucleotide frequency was divided by the product of the single nucleotide frequencies, thus normalizing it to the expected dinucleotide frequency in random sequences. This reflects the assumption that there are biological processes acting specifically on the dinucleotide composition of an organism's DNA and that the effect of such processes might be hidden by distortions on the single nucleotide level.

In a range of studies, phylogenies (i.e. evolutionary relationships among species or 'trees of life') have been constructed from these dinucleotide statistics (e.g., Refs. [18,19]). To extend the considerations to longer words, Schbath [20] defined a z-score for the abundancies of d-mer by taking into account the expected frequency and variance based on shorter words as they evidently influence the distribution of d-mer. This score was used by Woyke et al. [21] to analyze data from a metagenomics project. Hao and Qi [22] used abundancies of nucleotide- and amino acid-words to successfully build a phylogeny of prokaryotes.

The origin of the species specificity of word frequencies is still not clear. Karlin et al. [23] suggested that it is caused by DNA replication- and repair-enzymes that behave slightly differently for different organism. Zhao et al. [24] gave support for this hypothesis by finding that the presence of a subunit of the DNA polymerase affects the GC content of the genome. A recent review by Bohlin [25] also lists some further possible explanations like DNA structure and the physical and chemical properties of an organism's typical environments.

Zipf's law has been seen as an evidence for 'structured language' underlying genomic sequences. These studies emphasized very early that from a statistical perspective, the non-coding regions of DNA sequences are non-random (see, e.g., Mantegna et al. [26,27], Stanley et al. [28]). These analyses have complemented previous work on long-range correlations in DNA sequences [29] by relating these observations with 'linguistic' features of the sequences.

Several studies investigated the phylogenetic signal contained in short-range correlations quantified using different methods, either as parameters of a discrete autoregressive process [30–33], via representations related to fractals (e.g., Deschavanne et al. [34,35]) or via compositional entropies [36,37] or a variant of the mutual information function [38].

Here we explore the questions, to what extent correlations and word counts contain independent evolutionary signals. To this end we have computed predictions of word counts in a model of symbol correlations. We then use the deviation of the true word counts from these predictions as a novel statistical observable in DNA sequences.

In traditional null models the random expectations of word counts come from shorter words. As described above, the motivation is that the statistical influence of some biological processes, which take place on a specific scale, is hidden by a distortion on the scale of shorter words. Subtracting this 'statistical noise' allows to investigate and interpret the processes taking place on the larger scale. We consider word counts not in the sense of such abundancies but rather with respect to a new correlation-based null model, where the random expectations of word counts come from the observed correlations in the DNA sequence at hand.

At the core of our study is a specific and well-established model of statistical correlations in DNA sequences, a discrete autoregressive process of the order p, DAR(p) (introduced in Ref. [39] and applied to DNA sequences in Ref. [40]; see also Refs. [30,41,31]). The relative correlation strengths α_i between symbols at distance i (up to p) are model parameters. Expressing the word frequencies as a function of these model parameters allows us to use correlations as a null model for word frequencies.

We apply this new view of correlation-based null models for word counts in DNA sequences to two biological scenarios serving as case studies: The automatized clustering of sequence fragments in metagenomics data and the microsatellite inventory of several eukaryotic genomes.

2. Correlation-based null model

Our correlation-based null model is derived from a discrete autoregressive process of order p, DAR(p). This is a parameter-efficient way of modeling a subset of Markov processes of order p containing short- and medium-range correlations [39,40].