Journal of Theoretical Biology 374 (2015) 179–182

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Letter to Editor

Necessary relations for nucleotide frequencies

ARTICLE INFO

Keywords:
Genome composition
Combinatorics
Dinucleotide frequency
k-mer Analysis

ABSTRACT

Genome composition analysis of di-, tri- and tetra-nucleotide frequencies is known to be evolutionarily informative, and useful in metagenomic studies, where binning of raw sequence data is often an important first step. Patterns appearing in genome composition analysis may be due to evolutionary processes or purely mathematical relations. For example, the total number of dinucleotides in a sequence is equal to the sum of the individual totals of the sixteen types of dinucleotide, and this is entirely independent of any assumptions made regarding mutation or selection, or indeed any physical or chemical process. Before any statistical analysis can be attempted, a knowledge of all necessary mathematical relations is required. I show that 25% of di-, tri- and tetra-nucleotide frequencies can be written as simple sums and differences of the remainder. The vast majority of organisms have circular genomes, for which these relations are exact and necessary. In the case of linear molecules, the absolute error is very nearly zero, and does not grow with contiguous sequence length. As a result of the new, necessary relations presented here, the foundations of the statistical analysis of di-, tri- and tetra-nucleotide frequencies, and k-mer analysis in general, need to be revisited.

Theoretical work aimed at deciphering features of molecular evolution and the processes bearing on these features can only be effective if the most fundamental properties of sequences are clearly understood. In particular, patterns or regularities of a purely mathematical nature need to be separated from evolutionary signal. In the following, I describe such a type of mathematical pattern or regularity. I present it in its simplest form, since that is most likely to contribute to actual understanding.

The analysis of nucleic acid sequences is a fundamental component of modern genomic and evolutionary analysis. The base composition of DNA has been studied since the earliest days of molecular biology (Freese, 1962), beginning with the empirical observations of Chargaff and the rules he derived from them (Chargaff and Davidson, 1955). As a step towards understanding the genetic code, combinatorics was successfully used (Brenner, 1957) to exclude the possibility of overlapping triplet codes (Gamow, 1954), and the results presented here continue that line of thought which associates pure mathematical proof with molecular reality. Dinucleotide frequencies are the frequencies of neighbouring pairs of nucleotides in the order they appear in a given sequence. For standard nucleic acid sequences, there are sixteen such pairs, and their relative abundances have long been known to be biologically and evolutionarily informative (Freese, 1962; Karlin and Burge, 1995). Here we show that four (25%) of these can be expressed as simple sums and differences of the remaining twelve, for purely combinatorial reasons. This is the first time that this level of dependency has been recognised. The new relations are mathematically exact for any circular molecule, including plasmids and the genomes of many viruses and the vast majority of cellular organisms, irrespective of assumed mutation rates or models, with or without selection. For linear molecules, the worst case error does not increase with contiguous sequence length. The same type of dependency exists between words of more than two nucleic acids, and the relations apply to metagenomic analysis. Since statistical analysis of any kind relies upon an understanding of what is independent and what is not (Walker, 1940), the assumptions made in statistical analyses of nucleic acid composition need to be revisited.

Early studies of base composition tended to focus on the frequencies of base pairs (Freese, 1962; Karlin and Burge, 1995), but it has also become common, particularly in the context of metagenomics, to count triplets, quadruplets (tetranucleotides) (Pride et al., 2003; Teeling et al., 2004b) or longer words (k-mers) (Alneberg et al., 2014; Chor et al., 2009; Ragan and Chan, 2013), and these are applied in a wide variety of contexts (Hohl and Ragan, 2007; Karlin and Ladunga, 1994; Mrazek, 2009; Sims et al., 2009). Dinucleotide frequencies have continued to receive attention (Baran and Ko, 2008; Liu and Li, 2008; Palmeira et al., 2006; Simmonds et al., 2013), as exemplified by studies of CpG islands and CpG methylation (Arndt and Hwa, 2005; Bernardi, 2012; Zemach et al., 2010). Of particular relevance here is a careful study of CpG and TpA deficiencies in human isochores that revealed covariation due to mathematical consequences of dinucleotide overlap (Duret and Galtier, 2000). Starting from that observation, that overlap implies mathematical constraints, one can consider these constraints from an entirely abstract point of view. Here, an exact and complete account of all the mathematical consequences of dinucleotide, trinucleotide and tetranucleotide overlap is provided for the first time. This will impact statistical analysis of the composition of any type of nucleic acid, since what is revealed are necessary relations which until now had not been apparent, and such relations directly determine, among other things, the number of degrees of freedom (Walker, 1940). It is likely

http://dx.doi.org/10.1016/j.jtbi.2015.03.025
0022-5193/

that patterns seen in the analysis of raw sequence and gene expression data which are not specific to biological source (Zheng et al., 2011) are in part reflections of the dependencies exposed here. The new exact relations can also be used to design more logically consistent mutation models, since any consistent model must respect these dependencies. For a detailed case study of what this might entail, the reader is referred to the study by Duret and Galtier (2000), which improved upon an earlier mutation model (Sved and Bird, 1990) by taking aspects of dinucleotide overlap into account. The results presented here allow one to go to a more fundamental level, and design models which include the complete set of correct dependencies in their basic formulation. The positive impact of improved statistical analysis can be expected to increase the signal to noise ratio in metagenomics, and could be built in to the algorithms associated with large databases (Kryukov et al., 2012; Teeling et al., 2004a). To illustrate what is meant on a practical level, a new relation which can be directly applied in metagenomic composition analysis is provided below, and derived in the Supplementary material.

The main result for dinucleotides is best illustrated by concrete examples. The number of times the pair "CG" appears in the DNA sequence of a circular genome can be computed using nothing more than the numbers of times the pairs "GC", "AC", "CA", "TC" and "CT" appear, via the following exact formula:

$$\#CG = \#GC + \#AC - \#CA + \#TC - \#CT. \quad (1)$$

The circular mitochondrial genome of the fish *Sardinops melanostictus* (Inoue et al., 2000) is a non-trivial example. It has 956 "GC" pairs, 1138 "AC" pairs, 1181 "CA" pairs, 1131 "TC" pairs and 1389 "CT" pairs. Inserting these into the equation, one finds that there should be $956+1138-1181+1131-1389$ or 655 "CG" pairs, and this is indeed the case. Note that it was not necessary to know the length of the genome, but the fact that it is circular does play a role. Before explaining our general approach, let us describe one direct way of deriving Eq. (1), using "N" as a wildcard which will match any nucleotide: One expects that #CN equals #NC, since both are effectively estimates of #C. Writing $\#CN=\#NC$, removing #CC from both sides and solving for #CG gives a rough derivation of Eq. (1). The role of circularity of the molecule becomes clear when one realises that #CN is *exactly equal to* #NC on a circular genome, but they may differ by at most one for a linear genome, with any difference due to a "C" at only one end. For example, the linear sequence "CACGT" can be broken down into the four dinucleotides "CA", "AC", "CG" and "GT". Two begin with a "C", so $\#CN=2$. Only one ends in a "C", so $\#NC=1$. One can see that $\#CG=1$, but $\#GC+\#AC-\#CA+\#TC-\#CT=0+1-1+0-0=0$, so Eq. (1) has an error of one (the left hand side equals 1, but the right hand side equals 0). If it were a circular molecule, there would be an extra dinucleotide "TC", completing the circle by linking the final nucleotide back to the intitial nucleotide of "CACGT". Eq. (1) would then be satisfied exactly.

The general approach is elementary, and does not require full formal treatment. It will be useful to imagine a bracelet made of a piece of string threaded through a number of variously coloured beads and then tied in a loop (Fig. 1). Whatever the colours are, it will be possible to divide them up into two types. If the knot is positioned such that it is visible between two specific beads, then, once a direction has been chosen, the type of each bead can be read out in order, beginning with the first bead after the knot, and ending with the last bead reached before returning to the knot. Every bead has a successor, from which it may or may not differ in type. If, in passing from one bead to the next, the type changes from that of the first bead to the other type, this can be imagined as a step away from the starting type. If the next bead is of the same type as the first bead, and the current bead is not, then that can be imagined as a step back to the starting type. Now, since the bracelet is a loop, we must finally return to the first bead, so the numbers of steps away and back must always be equal. This is a mathematical fact, not related to the materials the bracelet is made of, nor to the procedure by which the colours were divided, nor the position of the knot (Fig. 1).

If one strand of a circular genome or plasmid takes the place of the bracelet, the same logic can be applied. Given the standard four nucleotides, "G", "A", "T" (for DNA) or "U" (for RNA), and "C", then there are a number of choices which can be made concerning their type assignments. One could consider purines (i.e. "A" and "G") as one type, and pyrimidines ("T" or "U", and "C") as another. In that case, the logic of the argument leads to the conclusion that the number of nucleotide pairs classified as purine – pyrimidine must equal the number of pairs classified as pyrimidine – purine. This equality translates into the equation $\#\text{purine}-\text{pyrimidine}=\#\text{pyrimidine}-\text{purine}$, or, if the actual nucleotide combinations are written out in full, assuming DNA, $\#AT+\#AC+\#GT+\#GC=\#TA+\#CA+\#TG+\#CG$.

A linear chromosome corresponds to an open string — a cut bracelet. Let the string be cut between the final bead and the knot. The last bead no longer has a successor, and so, if the first and last beads differ in type, one change in type may be missed. The equation may then only be an approximation, but since the worst case is that only one step has been missed, the greatest error is plus or minus one.

Experimentally determined genome sequences can have many gaps. These correspond to many cuts in the string. One can treat each uncut, or contiguous, piece as a linear molecule. For each one, the worst case error is plus or minus one, so the worst case error for a gapped sequence is plus or minus the total number of contiguous subsequences. Human chromosome 1 (Gregory et al., 2006) (NCBI RefSeq (Pruitt et al., 2014) Accession NC_000001.11) has 10145272 "GC", 11598278 "AC", 16768284 "CA", 13844699 "TC" and 16444797 "CT" dinucleotides. Using Eq. (1), one arrives at the approximation

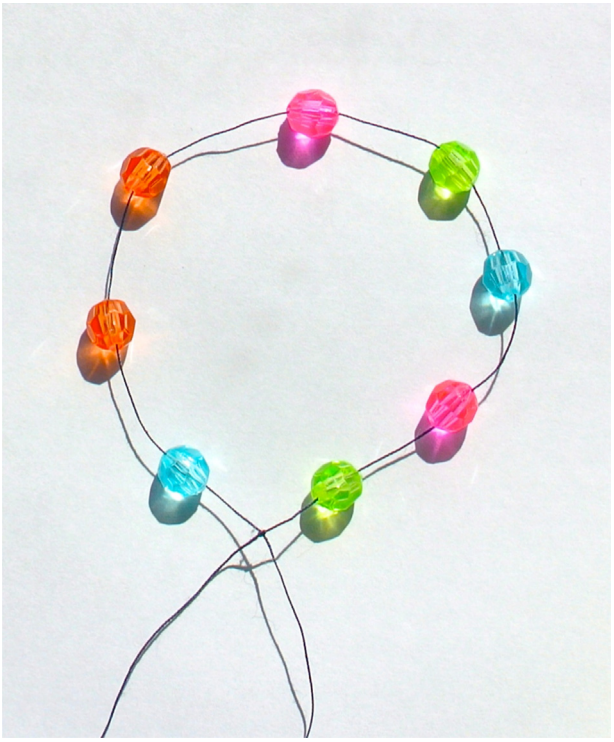

Fig. 1. A bracelet made from eight coloured beads threaded on a black string which has been knotted. One could decide to group the reddish ("R") and the greenish ("G") colours together. Starting at the bead to the immediate right of the knot, the bracelet reads GRGGRRRGG, including the first bead once more at the end. The number of times "GR" appears (two) equals the number of times "RG" appears. Such combinatorial equalities are the basis of our approach.