Available online at www.sciencedirect.com
ScienceDirect
Computational Biology and Chemistry 32 (2008) 147
Computational Biology and Chemistry
www.elsevier.com/locate/compbiolchem

Technical Comment

'Genome order index' should not be used for defining compositional constraints in nucleotide sequences

Eran Elhaik [a,*], Dan Graur [a], Krešimir Josić [b]

[a] *Department of Biology & Biochemistry, University of Houston, Houston, TX 77204-5001, USA*
[b] *Department of Mathematics, University of Houston, Houston, TX 77204-3008, USA*

Received 18 August 2007; accepted 29 November 2007

Abstract

A "genome order index," defined as $S = a^2 + c^2 + t^2 + g^2$, where a, c, t, and g are the nucleotide frequencies of A, C, T, and G, respectively, was used to suggest that there exist genome-specific constraints on nucleotide composition. We show that the "evidence" for constraint, $S < 1/3$, is in fact a mathematical property that is always true regardless of data. Moreover, we show that S is strictly equivalent to and derivable from the Shannon H-function and has no advantage over it.
Published by Elsevier Ltd

Keywords: Nucleotide composition; Genomic G + C content; Shannon H-function; Genome order index; Isochores

Zhang and Zhang (2004) discussed a "genome order index," defined as $S = a^2 + c^2 + t^2 + g^2$, where a, c, t, and g are the nucleotide frequencies of A, C, T, and G, respectively. The fact that the numerical value of S is smaller than 1/3 for almost all DNA sequences of 809 genomes have been erroneously interpreted as supporting evidence for the existence of genome-specific constraints on nucleotide composition of naturally occurring DNA, i.e., isochores. Zhang and Zhang (2004) also provided an apparently incorrect geometric explanation mistakenly suggesting that $S < 1/3$ is due to the actual nucleotide composition of DNA sequences, while in reality $S < 1/3$ is a mathematical property of S that should be always valid regardless of specific data. It should also be mentioned that contrary to the apparently inaccurate claim of Zhang and Zhang (2004), there is no obvious advantage of using S instead of Shannon (1948) H-function in sequence analysis. In fact, S and H functions are strictly equivalent to and derivable from each other, because they both are numerical measures of deviation from discrete uniform distribution of alphabetic symbol frequencies over sufficiently long strings (sequences) composed of these symbols.

Taking the foregoing comments into account we believe that the "genome order index" is a misconceived mathematical tool that should not be used in a meritorious sequence analyses.

Acknowledgments

KJ was supported in part by NSF Grant DMS-0071735. DG and EE were supported in part by NSF Grant DBI-0543342. We wish to thank the Editors of CBAC for carefully reading the manuscript and suggesting its current format.

References

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 24, 379–432.
Zhang, C.T., Zhang, R., 2004. A nucleotide composition constraint of genome sequences. Comp. Biol. Chem. 28, 149–153.

* Corresponding author. Tel.: +1 713 743 2312; fax: +1 713 743 2636.
E-mail address: eelhaik@gmail.com (E. Elhaik).

1476-9271/$ – see front matter. Published by Elsevier Ltd
doi:10.1016/j.compbiolchem.2007.11.003