Linguistic Features in Eukaryotic Genomes

Generating "words" from the structural elements of DNA is the heart of this issue. The words should be representing the structural as well the functional aspects of the genome.

PANAGIOTIS A. TSONIS AND ANASTASIOS A. TSONIS

Panagiotis A. Tsonis is from the Department of Biology, University of Dayton, Dayton, OH 45469-2320 (e-mail: Panagiotis.Tsonis@notes.udayton.edu) and Anastasios A. Tsonis is from the Department of Mathematics, University of Wisconsin-Milwaukee, Milwaukee, WI 53201-0413.

A remarkable feature of natural languages is that they follow a particular law called Zipf's law [1]. According to this law the rank (r) of each word and its frequency (f) are related via a power law, $f \propto r^{-1}$. On a log-log plot this relationship is linear with a slope of -1. The meaningfulness of this law has been debated by studies involving random or shuffled texts, but as it is argued in Reference 2, there are many aspects of the structure and evolution of natural languages that cannot be accounted for by random text. As such Zipf's law is considered by many as one of the most important laws in linguistics and signifies the structure found in languages and efficient information transfer.

During the past few years and prompted by considering that DNA is like an instructive text that provides the information to build organisms, attempts have been made to search whether or not DNA obeys a law similar to Zipf's law for languages. The key issue in such attempts is what could possibly constitute a "word" in DNA sequences. In an initial report [3], n-tuplets (made of the building blocks A,C,G,T) were used arbitrarily to generate words of a fixed length n. This analysis suggested that a form of Zipf's law could be applied for noncoding regions only. However, these results did not settle the issue [4–8], the main reason being the unrealistic choice of words. In addition, the fact that coding sequences, which are the ones that encode the instruction, did not follow Zipf's law was counterintuitive and remains problematic.

Generating "words" from the structural elements of DNA is the heart of this issue. The words should be representing the structural as well the functional aspects of the genome. Recent advances in whole genome sequencing and structural biology have provided important insights on both structural and functional aspects. There is now compelling evidence that genomes of more complex animals have evolved by gene duplication [9,10]. Duplication of DNA sequences that encode for domains in proteins with a particular function have been especially favored. Therefore, after eu-

TABLE 1

A Sample of Domain Rank and Frequency in Different Eukaryotic Genomes

	Domain/Frequency			
Rank	Yeast	*C. elegans*	*Drosophila*	Human
1	WD40 domain/121	Collagen triple helix repeat/384	Zinc finger C2H2 type/771	Zinc finger C2H2 type/4500
2	RNA recognition motif/73	F-box domain/324	Ig domain/291	Ig domain/930
3	Zinc finger C2H2 type/56	Ig domain/323	Ank repeat/269	Cadherin domain/550
4	DEAD/DEAH helicase/52	Ank repeat/223	WD40 domain/226	Fibronectin type III domain/545
10	Ank repeat/20	RNA recognition motif/145	LDL receptor domain class A/152	KRAB box/243
Lower limit of domain frequency	5	10	10	10

TABLE 2

Coefficient of Determination (R^2) and Root Mean Square (RMS) Residual of the Linear Regression between f and r *(Linear Model), between log*f and r (Exponential Model), and between log$f$ and log$r$ (Power Law Model)

	Yeast		*C. elegans*		*Drosophila*		Human	
	R^2	rms	R^2	rms	R^2	rms	R^2	rms
Linear	0.498	16.1	0.662	38.33	0.396	81.49	0.109	370.98
Exponential	0.929	0.111	0.894	0.136	0.929	0.117	0.870	0.178
Power law	0.974	0.057	0.979	0.061	0.958	0.090	0.981	0.068

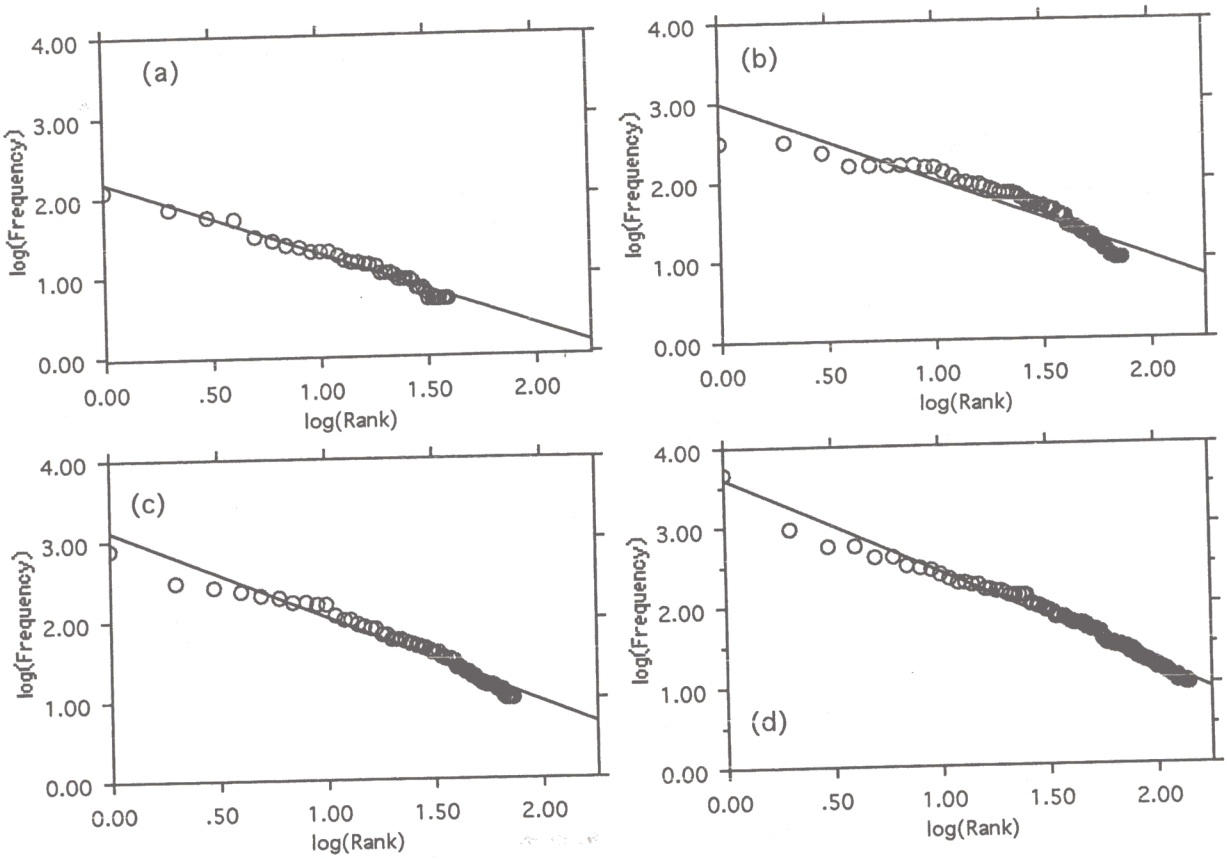
The power law model outperforms the other models in all genomes.

karyotic genomes were sequenced it was no surprise to find that certain protein domains with specific functions are present in many copies in different proteins. A structural domain is involved in a particular function, say DNA binding and, therefore, can be used by different proteins involved in these duties. Also, in agreement with evolutionary considerations, duplication of these domains has occurred more times as we climb the evolutionary ladder.

Based on the above, we focused our attention on genomes (rather than individual genes) and considered that a given genome is a language whose "words" are the different *domains*, which are found in proteins. This is a much more realistic approach because (1) not all domains have duplicated equally (some, like words in texts, are used more often than others), (2) domains constitute the heart of functional genomics (because they are translated into function), and (3) domains show an evolutionary significance. The data used here were taken from a recent study on the comparison of eukaryotic genomes [10]. We considered four genomes (yeast, *Caenorhabditis elegans*, *Drosophila*, and human), and for each genome we found the frequency and rank of each word (domain). For some examples of the ranking and frequency of protein domains in the different genomes see Table 1. Having these data, we investigated which model provides the best fit for the data. The models considered included a linear, an exponential and a power law model. We found (see Table 2) that the power law provided in all cases the best overall fit. Subsequently, Figure 1 shows, for the four different genomes, the corresponding log-log plots. In each of the log-log graphs the linear regression line (best fit) is also plotted. These results indicate that all four genomes obey the law $f \propto r^{-a}$ with a remarkably close to one, which is identical to Zipf's law for natural languages [1,2].

We conclude that Zipf's law can be recovered in genomes if the appropriate definition of a "word" is used. This result suggests that two very different means of information transfer may have shared similar evolutionary mechanisms. This view may lead to important clues about the evolution of languages, DNA, and information processing.

FIGURE 1



Log-log plot of the frequency of a particular domain against its rank. The most frequently appearing domain has rank one, the second most frequently appearing domain has rank two and so on. The straight line is the least squares fit. The slope of this line is (a) 0.92 for yeast, (b) 1.00 for *C. elegans*, (c) 1.10 for *Drosophila*, and (d) 1.15 for human.

REFERENCES

1. Zipf, G. Human Behavior and the Principle of Least Effort; Addison-Wesley: Cambridge, MA, 1949.
2. Tsonis, A.A.; Schultz, C.; Tsonis, P.A. Zipf's law and the structure and evolution of languages. Complexity 1997, 2, 12–13.
3. Mantenga, R.N.; Buldyrev, S.V.; Goldberger, A.L.; Halvin, S.; Peng, C.-K.; Simons, M.; Stanley, H.E. Phys Rev Lett 1994, 73, 3169–3172.
4. Israeloff, N.E.; Kagalenko, M.; Chan, K. Can Zipf law distinguish language from noise in noncoding DNA? Phys Rev Lett 1996, 76, 1976.
5. Bonhoeffer, S.; Hertz, A.V.M.; Boerlijst, M.C.; Nee, S.; Novack, M.A.; May, R.M. No signs of hidden language in noncoding DNA. Phys Rev Lett 1996, 76, 1977.
6. Voss, R.F. Comment on linguistic features of noncoding DNA sequences. Phys Rev Lett 1996, 76, 1978.
7. Chatzidimitriou-Dreismann, C.A.; Streffer, R.M.F.; Larhammar, D. Lack of biological significance in the linguistic features of noncoding DNA—a quantitative analysis. Nucleic Acids Res 1996, 24, 1676–1681.
8. Tsonis, A.A.; Elsner, J.; Tsonis, P.A. Is DNA a language? J Theor Biol 1997, 184, 25–29.
9. Ohno, S. Evolution by Gene Duplication. Heidelberg: Springer-Verlag, 1970.
10. Venter, J.C., et al. The sequence of the human genome. Science 2001, 291, 1304–1351.

Reply to Li and Grosse's "Comments on 'Linguistic Features in Eukaryotic Genomes' "

After receiving this comment by Li and Grosse, two things became very clear to us. One is that Li is very interested in our work. The other is that Li reacts negatively to any article that has the word Zipf in it. But this time again, like the last time (see Tsonis and Tsonis [1]), Li misses the point.

Li may have his opinion about linguistic features of genes. We respect that. He, however, does not respect that in science there may be more than one opinion. There is a large body of scientists that have produced a large amount of evidence that the language metaphor is much more than a simple metaphor [2].

Li fails to understand that the importance of our results [3] is not necessarily that molecules learn languages, but that there may be common principles behind the evolution of languages and genes. Li argues that duplication and random mutations are sufficient to reproduce a power law. We have no problem with that. But how does he know that language did not evolve in a similar way? Primitive languages were very repetitive. In fact, successful decipherment of ancient languages was based on finding repetitive units [4]. In addition, theories on the evolutionary aspects of languages agree that randomness is injected into languages because the transmission of a language from one generation to another is not perfect [5]. It is thus very possible that duplication and mutations underlie the evolution of both languages and genes. The same arguments can be extended to music, which also obeys power laws.

Thus, rather than arguing how can earthquakes learn a language, why don't we look at the bigger picture?

Panagiotis A. Tsonis
Department of Biology
University of Dayton
Dayton, OH 45469
E-mail:
panagiotis.tsonis@notes.udayton.edu

Anastasios A. Tsonis
Department of Mathematical Sciences
University of Wisconsin-Milwaukee
Milwaukee, WI 53201-0413
E-mail: aatsonis@uwm.edu

REFERENCES

1. Tsonis, A.A; Tsonis, P.A. On words and genes. Complexity 2003, 8(5), 12–13.
2. Searls, D.B. The language of genes. Nature 2002, 420, 211–217.
3. Tsonis, A.A; Tsonis, P.A. Linguistic features in eukaryotic genomes, Complexity 2002, 7(4), 13–15.
4. Robinson, A. Lost Languages; McGraw Hill, New York, 2002.
5. Nowak, M.A.; Komarova, N.L.; Niyogi, P. Computational and evolutionary aspects of language. Nature 2002, 417, 611–617.

