Noncoding DNA, Zipf's Law, and Language

Faye Flam (Research News, 25 Nov. 1994, p. 1320) reports that Eugene Stanley and his colleagues (*1*) have found that Zipf's law (*2*) applies somewhat better to noncoding than to protein-coding DNA sequences. The article implies that the statistical difference between protein-coding and noncoding DNA sequences is a surprising new discovery and that noncoding DNA resembles some sort of language.

The fact that nucleotide sequences of protein-coding regions have a different statistical structure than those of various kinds of noncoding regions (such as introns or intergenic spacers) has been well known since at least 1981 (*3*). In fact, many routine methods for discriminating between coding and noncoding DNA regions are based on such differences (*4*). It is therefore difficult to appreciate the alleged novelty of the findings of Stanley and his colleagues.

Zipf's distribution is not specific to language. Zipf himself said that it is far more general. Diverse examples of log-rank distributions that fit Zipf's law include relative sizes of cities (*2*, p. 416), income (*2*, p. 484; *5*), number of species per genus (*2*, p. 231), and number of papers per scientist in a given field of research (*2*, p. 514; *6*). There is no reason to conclude that a general population is a language even if a sample drawn from this population is characterized by Zipf's distribution.

The oligonucleotide frequency distribution in noncoding DNA does not appear to fit Zipf's law any better than does the distribution in coding regions. As may be seen clearly in the figure accompanying Flam's article, both log-rank distributions are similar and both display a nonlinear, rather than a linear, trend. In both cases, only a portion of the range can be approximated by a linear function when the data are plotted on log-log coordinates. A reasonable conclusion is that both coding and noncoding regions fit Zipf's law rather poorly, if at all.

Andrzej K. Konopka
Biolingua Research, 1415 Key Parkway, Frederick, MD 21702, USA
Colin Martindale
Department of Psychology, University of Maine, Orono, ME 04469, USA

References

1. R. N. Mantegna *et al.*, *Phys. Rev. Lett.* **73**, 3169 (1994).
2. G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Boston, 1949).
3. M. J. Shulman *et al.*, *J. Theor. Biol.* **88**, 409 (1981); J. W. Fickett, *Nucleic Acids Res.* **10**, 5303 (1982); J.-M. Claverie and L. Bougueleret, *ibid.* **14**, 179 (1986); P. Salamon and A. K. Konopka, *Comput. Chem.* **16**, 117 (1992).
4. E. C. Uberbacher and R. J. Mural, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 11261 (1991); M. Borodovsky and J. McIninch, *Comput. Chem.* **17**, 123 (1993); E. E. Snyder and G. D. Stormo, *Nucleic Acids Res.* **21**, 607 (1993); S. Karlin and L. R. Cardon, *Annu. Rev. Microbiol.* **48**, 619 (1994); A. K. Konopka, *Biocomputing: Informatics and Genome Projects*, D. Smith, Ed. (Academic Press, San Diego, CA, 1994), pp. 119–174.
5. V. Pareto, *The Mind and Society* (Harcourt Brace, New York, 1935).
6. A. J. Lotka, *J. Washington Acad. Sci.* **16**, 317 (1926).

Corrections and Clarifications

In the report "Continent-ocean chemical heterogeneity in the mantle based on seismic tomography" by Alessandro M. Forte *et al.* (21 Apr., p. 386), note 14 (p. 388) should have included the following sentence at the end. "We note, however, that this classical measure of significance does not take into account the red spectrum of the observed nonhydrostatic geoid, whose harmonic coefficients cannot be properly regarded as a random distribution; therefore, the statistical significance of the measured correlation coefficient is possibly less than 99%."