Indian Journal of Biochemistry & Biophysics
Vol. 38, February & April 2001, pp. 124-127

Identification of human proteins using the linguist's tools

S Chattopadhyay [1], J Chakrabarti [1,*], D Bandyopadhyay [2] and A Som [1]

[1]Department of Theoretical Physics, Indian Association for the Cultivation of Science, Calcutta 700 032;

[2] Linguistics Research Unit, Indian Statistical Institute, Calcutta 700 035

Accepted 3 October 2000

The symbolic sequences of the exons that make human proteins are subjected to methods of statistical linguistics. The ideas developed for the natural languages by G. K. Zipf, when applied to these sequences, show significant promise. In particular, we argue, the Zipf's exponent differentiates, and hence, identifies disparate human sequences.

Introduction

Natural languages share some gross scaling properties as noted by G. K. Zipf[1]. Later, through the second half of the 20th century, the other common, perhaps more substantive feature of natural languages, namely the generative grammar[2-4] has been exhaustively investigated. Literary pieces, somewhat like music, appear to have long-range power law correlations of $1/f^{\beta}$ type[5,6].

The DNA sequences are texts in the symbols A, C, G and T. The gross scaling properties, the Zipf's law, could characterize the DNA sequences the same way it does for the literary pieces. The Zipf's exponent could be used to identify and label the sequences.

The DNA sequences may be broadly divided into two parts: the coding (exons) and the non-coding (introns and the intergenic parts or flanks). The coding parts, dominated by the degenerate, triplet codons, typically have the Fourier spectra as shown in Fig. 1. The non-coding regions on the other hand are not dominated by single, unique periodicities, but have Fourier spectra of the type given in Fig. 2.

The first hurdle in applying the ideas of Zipf to DNA texts is the identification of words that carry biological sense. In the case of coding regions, the triplets dominate. The Fourier spectra have the overwhelming 3 period, i.e. a sharp peak at 1/3 frequency. Biological words may be viewed to be the triplet codons.

For introns and the flanks, the clear absence of any unique dominant feature makes identification of words ambiguous. The applicability of the Zipf's law in these cases is somewhat circumscribed[7].

The curious aspect is the long-range power law correlations. The introns and the flanks show a clear evidence of this behaviour. Notice the large fall-off for the ultra low frequencies in Fig. 2. On the other hand, such behaviour, if present in the coding regions, is weak and muted. The long-range order in exon regions is certainly not as universal as for the introns and the flanks. Some exons do show long-range correlations while for the others, these correlations are at best weak or non-existent. Thus the coding regions, where the identification of words is clear, appear to be weakly similar to the languages or music, while the introns and the flanks, though similar to the languages in the sense of long-range correlations, do not have the well-understood word structure.

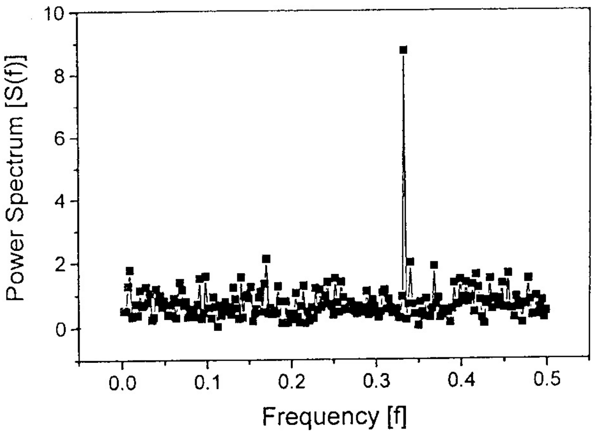

Fig. 1—Frequency [f] is plotted[11] against Power Spectrum [S(f)] for the exons of human alpha globin gene (GenBank V00488). [The exon regions show a sharp peak at f=1/3 suggesting the ubiquitous presence of the 3 *periodicity*. The peak at f=0 is muted. The exon texts, therefore, show a weak resemblance to the characteristic $1/f^{\beta}$ correlations in the languages.]

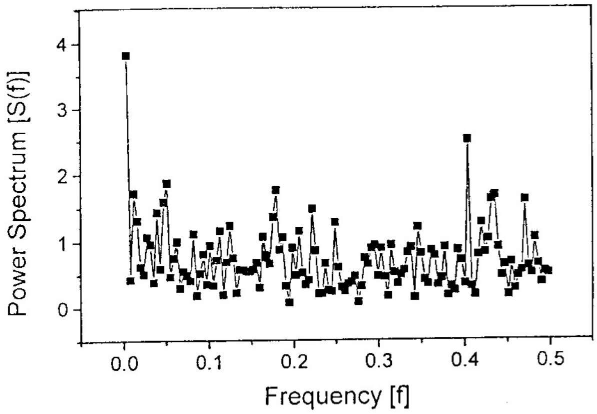

Fig. 2—Frequency [f] is plotted[11] against Power Spectrum [S(f)] for the introns of human alpha globin gene (GenBank V00488). It is difficult to isolate any unique periodicity that dominates the intron regions. Note the sharp fall-off near f=0, characteristic of the long-range $1/f^{\beta}$ correlations.

It is known, however, that the gross Zipf behaviour is applicable not just to the literary works, but to any random sequence of letters and spaces created playfully on a typewriter. Indeed, it has been argued that the idealized Zipf's exponent corresponds to the random sequence. It is the deviation of the exponent from this ideal value that characterizes the order and the organization in the sequences. With this in mind, we have applied the Zipf's law to coding sequences. The other, equally important, purpose of our work is to study if the exponent can be used to classify and label the sequences.

Methodology

Zipf proposed that in a text of certain length, the vocabulary V (i.e., the number of different words) and the text length T (i.e., the total number of words) hold a power-law relationship:

$$V = T^{\alpha} \qquad \text{... (1)}$$

where α is the Zipf's exponent.

The law was originally aimed at the analysis of natural human languages. Here we explore the application of this law to understand the exon regions of a few protein-coding human genes.

Exons are the coding parts of a gene, which consist of triplet codons, each of which gives rise to an amino acid in the ultimate protein. We considered each triplet a word, thereby, for a single gene, the maximum possible vocabulary could be 62 (because 61 for all amino acid coding triplets, and 1 for any of the three stop codons). The increase in vocabulary suggests the increase in the variability of triplets. If we take the log of both sides in equation (1), the relationship becomes:

$$\log V = \alpha \log T$$

$$\alpha = \frac{\log V}{\log T} \qquad \text{... (2)}$$

Therefore, the more the exponent value, more the variation of words in the sequence. Under idealized Zipf's law conditions, the value of the exponent would be 0.873. In other words, the closeness or the deviation of the exponent value from 0.873 gives the most pertinent information about the text concerned.

Results and Discussion

The Zipf's exponents for 29 human proteins were calculated. These proteins can be divided into nine different classes: oxidoreductase, kinase, transferase, high mobility group, histone, globin, globulin, albumin and insulin. The exponent rises steadily starting from albumin, which has the lowest exponent to histone H4 having the highest exponent. In Table 1 the results are summarized. A plot of the α-values against the proteins appear in Fig. 3. In Fig. 3, the five histones are subdivided into four different classes: H1, H2, H3 and H4.

Despite the wide variety in exon-lengths, the exponent value for a specific class of proteins was found to be nearly constant. Within the class the variation is marginal, the exceptions being H1, H2A and 2B, H3 and H4. These histones have widely varying Zipf-exponents. Curiously, the class oxidoreductase and histone H1 have overlapping exponent values (Fig. 3). It is interesting further that H2 (both A and B), H3 and the high mobility group (HMG) have exponent values close to one another.

The protein classes, leaving aside the exceptions noted above, are characterized by unique α values. Since the proteins belonging to a particular class perform nearly similar functions, the character α classifies the protein functions. In that sense, α gives us a measure of the structural similarity of the protein conformations.

The value of α varies from about 0.64 (albumin) to 0.81 (histone H4) and the internal organization of the peptides vary significantly. The difference of α provides a measure of the variation and, therefore, the organization (and information) in the amino acid sequences. We presume, therefore, the Zipf exponent is related to the entropy of the exon/polypeptide sequences.

Table 1—Zipf's exponent values for a few human protein-coding exon sequences

Category	Protein	Exon-length	Zipf's exponent
Oxidoreductase	Glyceraldehyde-3-phosphate dehydrogenase	1008	0.6825
	Lactate dehydrogenase		
	Dihydrodiol dehydrogenase	999	0.6991
	L-glycerol-3-phosphate : NAD oxidoreductase	972	0.7083
		1050	0.6961
Kinase	Glucokinase (liver)	1395	0.6524
	Glucokinase (pancreas)	1398	0.6492
	Pyruvate kinase M2 type	1596	0.6550
	Pyruvate kinase L type	1632	0.6526
Transferase	Glutathione transferase	669	0.7307
	Hypoxanthine phosphoribosyl transferase	657	0.7367
High mobility group	HMG1	648	0.7420
	SRY	615	0.7528
Histone	H1	666	0.6962
	H2A	393	0.7762
	H2B	378	0.7777
	H3	408	0.7837
	H4	312	0.8098
Globin	Alpha globin	429	0.7221
	Beta globin	444	0.7279
Globulin	Thyroxine-binding globulin	1248	0.6761
	Sex hormone-binding globulin	1209	0.6853
	Corticosteroid-binding globulin	1218	0.6789
Albumin	Serum albumin	1830	0.6435
	Alpha albumin	1800	0.6374
	Alpha fetoprotein	1830	0.6384
	Afamin	1800	0.6374
Insulin	Insulin	333	0.7986
	Insulin-like growth factor precursor	462	0.7919
	Insulin-like growth factor 1 (breast tumour cell line)	414	0.7980

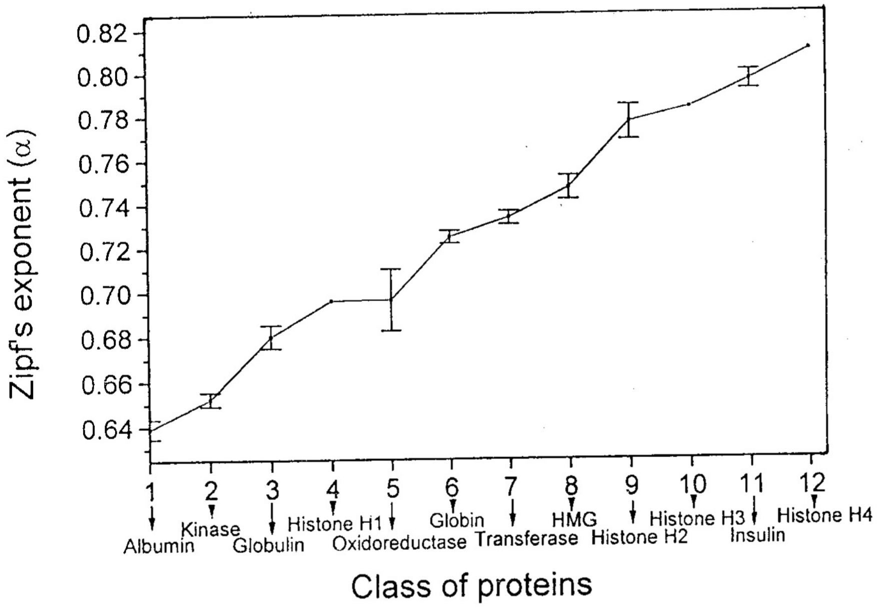

Fig. 3—The Zipf's exponent values for 12 different classes of protein-coding exon sequences. The five histones are subdivided into four different classes: H1, H2, H3 and H4. The error-bars simply indicate the maximum deviations from the average values.

Since α characterizes the classes, it provides a quick measure for the identification of unknown query sequence. In that sense, the study of the exponent is useful.

The natural as well as the computer languages are characterized by gross statistical features studied by Zipf and subsequently, by many others. The DNA texts of the letters A, C, G and T are subjected to this statistical analysis. The results show that the gross statistical quantities are linked to the DNA functions.

The other, and deeper, feature of the languages is the generative grammar[8-10]. Work is in progress in our laboratory on the generative grammar of the DNA sequences.

Acknowledgement

Our interest in the sequences owe a great deal to Prof. Ashesh Nandi. This interest has been further reinforced by our constant interactions with Prof. Anjali Mookerjee and Dr. Chaitali Mukhopadhyay. Prof. Rabi Majumdar, with his active interest, has inspired us to follow through. Satyabrata Sahoo, our companion in the laboratory, has been helping us on many fronts.

References

1 Zipf G K (1949) *Human Behaviour and the Principle of Least Effort* Addison-Wesley, Cambridge, Massachusetts, USA

2 Chomsky N (1957) *Syntactic Structures* MIT Press, Cambridge, Massachusetts, USA

3 Chomsky N (1965) *Aspects of the Theory of Syntax* MIT Press, Cambridge, Massachusetts, USA

4 Chomsky N (1995) *The Minimalist Programme* MIT Press, Cambridge, Massachusetts, USA

5 Ebeling W & Neiman A (1995) *Physica* A 215, 233-241

6 Havlin S (1995) *Physica* A 216, 148-150

7 Mantegna R N, Buldyrev S V, Goldberger A L, Havlin S, Peng C K, Simons M & Stanley H E (1995) *Physical Review* E 52, 2939-2950

8 Searls D B (1992) *American Scientist* 80, 579-591

9 Bairoch A, Bucher P & Hofmann K (1997) *Nucleic Acids Research* 25, 217-221

10 Durbin R, Eddy S R, Krogh A & Mitchison G (1998) *Biological Sequence Analysis*, Cambridge University Press, Cambridge, USA

11 Chattopadhyay S, Som A, Sahoo S & Chakrabarti J (2000) *Indian J Phys* 74B(1), 1-39