ARTICLE

Frequency and Correlation of Nearest Neighboring Nucleotides in Human Genome

Neng-zhi Jin, Zi-xian Liu, Wen-yuan Qiu*

Department of Chemistry, State Key Laboratory of Applied Organic Chemistry, Lanzhou University, Lanzhou 730000, China

(Dated: Received on May 28, 2008; Accepted on December 16, 2008)

Zipf's approach in linguistics is utilized to analyze the statistical features of frequency and correlation of 16 nearest neighboring nucleotides (AA, AC, AG, $\cdots$, TT) in 12 human chromosomes (Y, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, and 12). It is found that these statistical features of nearest neighboring nucleotides in human genome: (i) the frequency distribution is a linear function, and (ii) the correlation distribution is an inverse function. The coefficients of the linear function and inverse function depend on the GC content. It proposes the correlation distribution of nearest neighboring nucleotides for the first time and extends the descriptor about nearest neighboring nucleotides.

Key words: Zipf's law, Nearest neighboring nucleotide, Frequency distribution, Correlation distribution

I. INTRODUCTION

The sequence of human chromosome 1 finished in 2006 indicates the 46 human chromosomes have been entirely sequenced [1]. The human genome contains nearly 2.91Gbp [2]. The crucial question of modern genomics is what kinds of information can be extracted from these data. The statistical analysis of a DNA sequence is of importance for understanding the structure and function of genomes, and several statistic methods have been proposed to study DNA sequences [3-29].

A DNA sequence is composed of four basic chemical letters (A, C, G, and T). It can be considered as a genetic language which includes abundant information; hence, DNA has its words [7] and grammar construction [5], similar to natural language. A remarkable feature of language is Zipf's law which states that the rank (r) of each word and its frequency (P_r) are related via a power law, $P_r=C/r^a$, $a\approx1$ [30]. In 1994, Zipf's approach [30] was first extended to analyze DNA sequence by Mantegna, who defined a word as an n-tuple (n is a free parameter) and found that n-tuple frequency followed Zipf's law [7]. This inaugurated the beginning of study of DNA sequences with Zipf's approach. Subsequent studies proposed that n-tuple frequency follows different distributions [8-16], such as Zipf's law, exponential function, Yule distribution, and nonlinear function. In 1998, Luo *et al.* defined the grammar construction as the correlation between nucleotides, and found that CG and TA models are higher by one or several times than others for most sequences [5]. The research about frequency distribution has mainly paid attention to n-tuple frequency distribution where $n>2$, and gained different results for different selected sequences [7-16]. On the other hand, research about grammar construction [5] only reported the nonuniformity of the correlation distribution of 16 nearest neighboring nucleotides (2-tuple), but did not report the form of correlation distribution clearly.

The study of nearest neighboring nucleotides is important for understanding a DNA sequence. The set of dinucleotide (2-tuple) biases is a remarkable property of the DNA of an organism [17-22]. In addition, neighboring nucleotides may influence the pattern of nucleotide substitution [23-27]. The frequency of dinucleotide also can be used to predict the frequency of oligonucleotide [28,29]. Intriguingly, the universal existence of strong nucleotide correlation in nearest neighboring sites is similar to the grammar construction of natural language [5], which can determine the global statistical property of natural language [31]. Furthermore, the nucleotide correlation in adjacent sites which contain the genetic information is correlated with evolution [5].

Therefore, we focus our attention on the statistical features of frequency and correlation of nearest neighboring nucleotides (2-tuple) in the human genome and establish reasonable mathematical models. Moreover, we investigate the relationship between the distributions and nucleotide composition of the human chromosomes in detail. We find the frequency and correlation distribution of nearest neighboring nucleotides depend on GC content (GC%) in the human genome.

*Author to whom correspondence should be addressed. E-mail: wyqiu@lzu.edu.cn

II. DATA AND METHODS

A. Human chromosomes

Taking into account computing bias and capability, we downloaded 12 human chromosomes (chromosome Y, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, and 12) from ftp://ftp.ncbi.nih.gov/genomes.

B. Frequency of nearest neighboring nucleotides

DNA is composed of 4 nucleotides (A, C, G, and T), where A refers to adenine, C refers to cytosine, G refers to guanine, and T refers to thymine. Hence, there are 4^2=16 2-tuples. To obtain the word frequency for each 2-tuple, we start from the first base pair of the DNA sequence that is under study and progressively shift by 1 base with a window of length 2. For a DNA sequence containing L base pairs, the total number of words is L-1.

C. Correlation of nearest neighboring nucleotides

The correlation of nucleotides is the basis for grammatical construction of genetic language [5]. Probability theory can be used to analyze the correlation of nearest neighboring nucleotides which can be defined as follows [5]:

$$F_{ij} = (P_{ij} - P_i P_j)^2 \quad i, j = \mathrm{A, C, G, T} \tag{1}$$

where P_i is the frequency of nucleotide i occurring in the sequence, and P_{ij} is the frequency of nearest neighboring nucleotides i and j in the sequence. The correlation F_{ij} can be calculated according to Eq.(1).

For the frequency and correlation of nearest neighboring nucleotides, in order to perform Zipf's analysis, we ranked them from high to low. The highest is ranked as 1, the next one as 2, and so forth.

D. Curve fitting and regression analysis

Curve fitting is finding a curve which has the best fit to a series of data points and possibly other constraints. Regression analysis allows for an approximate fit by minimizing the difference between the data points and the curve. Regression analysis is a technique used for the modeling and analysis of numerical data consisting of values of a dependent variable (response variable) and of one or more independent variables (explanatory variables). We utilize regression analysis to analyze the data [32].

Once a regression model has been constructed, it is important to confirm the goodness of fit of the model and the statistical significance of the estimated parameters. Commonly used checks of goodness of fit include R-squared which tell us what percentage of the variability in the dependent variable can be explained by the independent variables, analyses of the pattern of residuals, and hypothesis testing. Statistical significance is checked by an F-test of overall fit, followed by t-tests of individual parameters.

III. STATISTICAL FEATURES OF NEAREST NEIGHBORING NUCLEOTIDES

A. Frequency distribution of nearest neighboring nucleotides

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 determine whether or not DNA obeys a law similar to Zipf' s law for languages.

Here we study the functional form of 2-tuple frequency distribution in analogy to Zipf's analysis of natural language. To implement Zipf's analysis, we define the frequency of nearest neighboring nucleotides which rank is r as P_r. The ranked nearest neighboring nucleotides in 12 human chromosomes are listed in Table I. The log-rank distributions can be found in Ref.[7], and only a portion of the range can be approximated by a linear function when the data are plotted on log-log coordinates. It is necessary to fit the whole range of data. Therefore, we plot in normal scale rather than double logarithmic scale. The plots of P_r against r are shown in Fig.1. The trends of the plots of P_r against rank r for human chromosomes are almost the same. They reveal hierarchies in the frequencies of different 2-tuples. We can conclude that the frequency of a 2-tuple on a given strand approaches its reverse complementary 2-tuple on the same strand (e.g. AA≈TT, TG≈CA). For

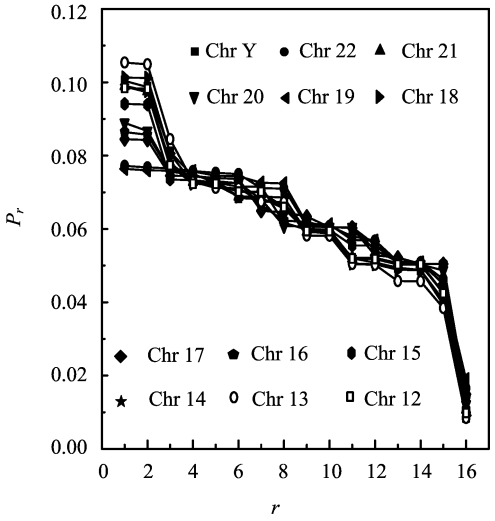

FIG. 1 Frequency P_r of nearest neighboring nucleotides *vs.* its rank r.

DOI:10.1088/1674-0068/22/01/27-33

TABLE I The ranked nearest neighboring nucleotides in human genome in term of their frequencies, the most frequent is rank 1, the next is rank 2, and so forth.

Chr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Y	TT	AA	AT	TG	CA	AG	CT	TA	GA	TC	GT	AC	GG	CC	GC	CG
22	CA	TG	AA	TT	AG	CT	CC	GG	GA	TC	AT	GC	AC	GT	TA	CG
21	AA	TT	AT	CA	TG	AG	CT	TA	GA	TC	GG	CC	AC	GT	GC	CG
20	TT	AA	TG	CA	CT	AG	AT	TC	GG	CC	GA	TA	GT	AC	GC	CG
19	TT	TG	AA	CA	AG	CT	GG	CC	GA	TC	GC	AT	GT	AC	TA	CG
18	TT	AA	AT	TG	CA	CT	AG	TA	GA	TC	GT	AC	GG	CC	GC	CG
17	TT	AA	TG	CA	CT	AG	CC	GG	AT	TC	GA	TA	GC	GT	AC	CG
16	TT	AA	TG	CA	CT	AG	AT	GG	CC	TC	GA	TA	GT	AC	GC	CG
15	AA	TT	CA	TG	AT	CT	AG	TA	TC	GA	CC	GG	AC	GT	GC	CG
14	TT	AA	AT	TG	CA	CT	AG	TA	TC	GA	GG	CC	GT	AC	GC	CG
13	TT	AA	AT	TA	TG	CA	CT	AG	TC	GA	GT	AC	CC	GG	GC	CG
12	TT	AA	AT	CA	TG	CT	AG	TA	TC	GA	CC	GG	AC	GT	GC	CG

most chromosomes, the most frequent 2-tuple is TT, followed by AA. CG is the lowest frequent 2-tuple for all chromosomes. In the human genome, cytosine in 5′-CG-3′ is often methylated, and deamination of 5′-methylcytosine produces thymine [33,34]. Hence, the scarcity of CG and excess of TG in the human genome is explained from the fact that methylation-deamination-mutation would convert CG to TG.

Having this data, we investigate which model provides the best fit for the data (regression analysis). The models considered include a power (y=ax^b), an exponential (y=ae^{bx}), a linear (y=ax+b) and a logarithmic (y=$a\ln x$+b) function. The coefficients of determination R^2 between P_r and r indicate that the linear function leads to more exact results. The linear function is significant after F-test and t-test (significance level a=0.05). Accordingly, the frequency distribution of nearest neighboring nucleotides in human genome is a linear function, which is different from the case for n-tuples where n>2 [7-16].

Of course, it is necessary to determine the factor that influences the fit coefficients of the linear function (y=ax+b). We expect the fit coefficients depend on the GC% of the sequence. Figure 2(a) shows the parameter a *vs.* GC% and Fig.2(b) shows the parameter b *vs.* GC%. We find the values of a and b are best fitted by a polynomial function:

$$a = -0.151\mathrm{GC\%}^2 + 0.1499\mathrm{GC\%} - 0.0402 \quad (2)$$
$$R^2 = 0.9981,\ F = 2403.55,\ P < 10^{-6}$$
$$b = 1.3019\mathrm{GC\%}^2 - 1.2911\mathrm{GC\%} + 0.4078 \quad (3)$$
$$R^2 = 0.9992,\ F = 5345.62,\ P < 10^{-8}$$

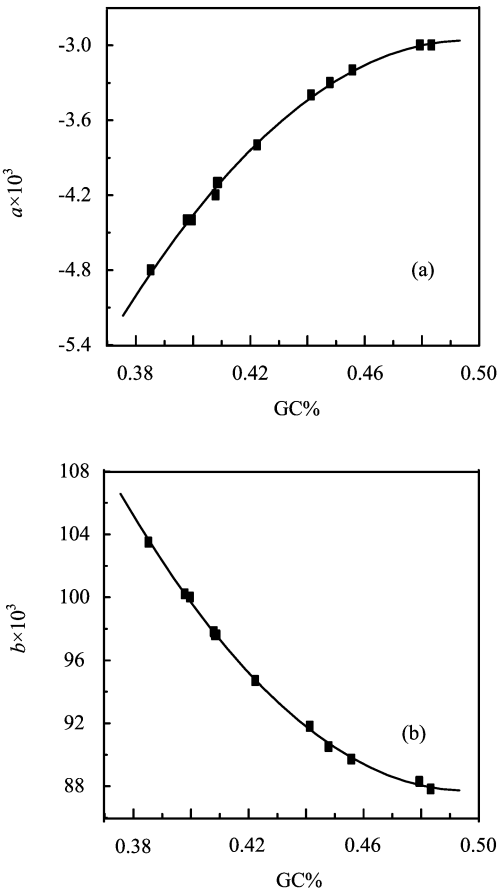

FIG. 2 The coefficients of linear function (y=ax+b) against GC%. (a) a *vs.* GC%; (b) b *vs.* GC%.

B. Correlation distribution of nearest neighboring nucleotides

The correlation of nucleotides where the genetic information is mainly stored is the basis for grammatical

DOI:10.1088/1674-0068/22/01/27-33

TABLE II The ranked nearest neighboring nucleotides in human chromosomes in term of their correlations, the most one is rank 1, the next is rank 2, and so forth.

Chr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Y	CG	TA	TG	CA	TT	AA	CC	AT	GG	AC	GT	AG	CT	GC	TC	GA
22	CG	TA	CA	TG	GG	CC	AG	CT	AC	GT	AT	AA	TT	GA	TC	GC
21	CG	TA	TG	CA	AA	TT	GG	CC	AC	GT	AT	CT	AG	GA	TC	GC
20	CG	TA	CA	TG	CC	GG	AG	CT	GT	AC	AT	TT	AA	TC	GA	GC
19	CG	TA	CC	GG	CA	TG	AG	CT	GT	AC	AT	TT	AA	GC	TC	GA
18	CG	TA	TG	CA	AA	TT	AT	AC	GT	CC	GG	CT	AG	GC	GA	TC
17	CG	TA	CC	GG	TG	CA	CT	AG	AC	GT	AT	AA	TT	TC	GA	GC
16	CG	TA	TG	CA	CC	GG	AC	GT	CT	AG	AA	TT	AT	TC	GA	GC
15	CG	TA	CA	TG	GG	CC	AT	TT	AA	CT	AG	GT	AC	GA	TC	GC
14	CG	TA	CA	TG	TT	AA	CC	AT	GG	GT	AC	AG	CT	TC	GA	GC
13	CG	TA	TG	CA	TT	AA	AT	AC	GT	CC	GG	AG	CT	GC	TC	GA
12	CG	TA	CA	TG	TT	AA	AT	CC	GG	GT	AC	AG	CT	TC	GA	GC

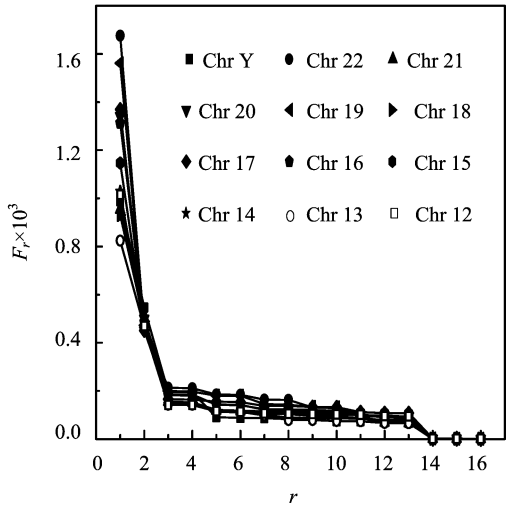

FIG. 3 Correlation F_r of nearest neighboring nucleotides *vs.* its rank r.

construction of genetic language [5]. We also use Zipf's approach to analyze the feature of the correlation of nearest neighboring nucleotides. F_r denotes the correlation of nearest neighboring nucleotides whose rank is r. The ranked correlation of nearest neighboring nucleotides are listed in Table II. The correlation modes CG and TA are strongest for most chromosomes as reported in Ref.[5]. Figure 3 shows plots of F_r against r. Although there are differences in some regions, the plots of F_r against r of human chromosomes have the same trend. To find out which model provides the best fit, a power (y=ax^b), an exponential (y=ae^{bx}), an inverse (y=a+b/x), and a logarithmic (y=$a\ln x$+b) function are used to fit the data (regression analysis). According to the coefficients of determination (R^2) between F_r and r, the inverse function provides the best overall fit in all cases and is significant after F-test and t-test (significance level a=0.05). Therefore, we propose that the correlation distribution of nearest neighboring nucleotides in human genome follows an inverse relationship.

Furthermore, in order to explore the relationship between the distribution of nearest neighboring nucleotide and base composition, we fit the coefficients a and b of inverse function (y=a+b/x) *vs.* GC% of the human chromosomes. The Fig.4(a) shows the parameter a *vs.* GC% and Fig.4(b) shows the parameter b *vs.* GC%. We find the coefficients of a and b are best fitted by linear function:

$$a = -0.00062\mathrm{GC\%} + 0.00021 \quad (4)$$
$$R^2 = 0.8972,\ F = 87.24,\ p < 10^{-5}$$
$$b = 0.007\mathrm{GC\%} - 0.00183 \quad (5)$$
$$R^2 = 0.9580,\ F = 228.31,\ p < 10^{-5}$$

IV. CONCLUSION AND DISCUSSION

As a genetic language, a DNA sequence has its words and grammar construction. In order to establish a mathematical model for genetic language one should first study the statistical characteristics of the language. Some scientists have studied frequency distribution in the coding sequences, and others studied have also considered noncoding sequences. In the case of a coding sequence, the words are the 64 3-tuple which code for the amino acids. However, for a noncoding sequence, the words are not known. In order to unify the value of n, researchers study the features of n-tuple (n>2), however, there have been no reports about the frequency distribution of 2-tuples. Therefore, in this work, in order to explore the common features of the human chromosome, we extended Zipf's approach to analyze the statistical features of nearest neighboring nucleotides in 12 human chromosomes. We calculated the frequencies and correlations of nearest neighboring nucleotides

DOI:10.1088/1674-0068/22/01/27-33

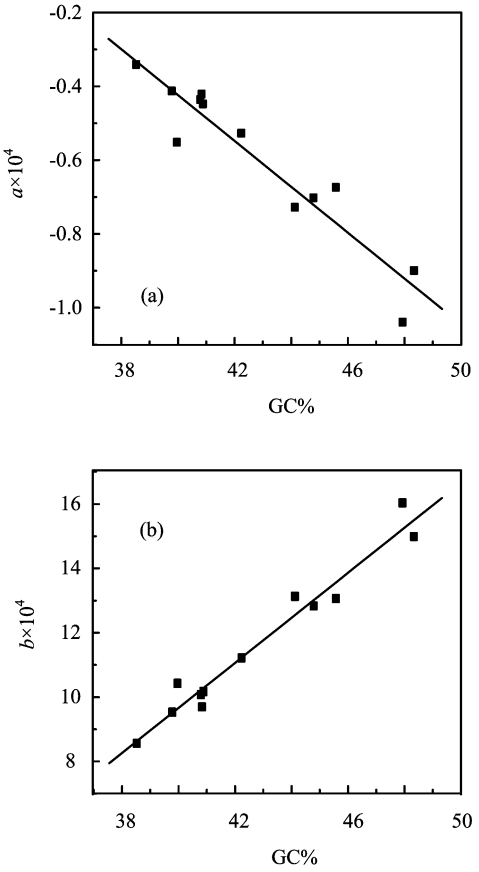

FIG. 4 The coefficients of inverse function (y=a+b/x) against GC%. (a) a *vs.* GC%; (b) b *vs.* GC%.

and ranked them from high to low. Different functions, such as a power (y=ax^b), an exponential (y=ae^{bx}), a linear (y=ax+b), an inverse (y=a+b/x), and a logarithmic (y=$a\ln x$+b) function, were applied in the fitting. The fitting results indicate that the frequency distribution of nearest neighboring nucleotides follows a linear relationship, which is different from n-tuple (n>2) frequency distribution [7-16], and the correlation distribution of nearest neighboring nucleotides follows an inverse relationship, which has not been reported until now. It should be pointed out that the plot of frequency of nearest neighboring nucleotides against its rank is not a straight line because of the suppression of CG, but the fitting is significant at significance level a=0.05.

Furthermore, the frequency and correlation distributions of nearest neighboring nucleotides depend on GC content in human chromosomes. We fit the coefficients of the linear function which represents the frequency distribution of 2-tuples and the inverse function which represents the correlation distribution of 2-tuples versus GC content. The results show that the coefficients a and b of the linear function are best fitted by a polynomial function, and the coefficients a and b of the inverse function are best fitted by a linear function. Therefore, we can conclude that the composition of DNA sequence (GC%) determines the frequency and correlation distribution of nearest neighboring nucleotides in human genome. In other words, GC content mainly decides the statistical features of nearest neighboring nucleotides.

The universal existence of strong nucleotide correlation in adjacent sites of DNA sequence is an important characteristic of genetic language, and the correlation of nearest neighboring nucleotides increases with evolution [5]. The evolution of nucleotide sequences is dominated by two major factors, random mutation which decreases nucleotides correlations and the natural selection which increases nucleotides correlations [5,35,36]. For an uncorrelated random sequence, the frequencies of mononucleotide and nearest neighboring nucleotides are polarized, namely, P_{ij}=P_iP_j=1/16, P_i=1/4, hence, the frequency distribution of nearest neighboring nucleotides is a line whose slope is zero. Furthermore, related to Eq.(1), the correlation of nearest neighboring nucleotides of random sequence F=0. However, the frequency distribution of nearest neighboring nucleotides follows a linear relationship, whereas, frequency distribution of n-tuple (n>2) follows different distributions [8-16], such as Zipf's law, exponential function, Yule distribution, and nonlinear function. This shows that with the increase of n, the frequency distribution of n-tuples presents different features, and also with different studied sequences, the frequency distribution of n-tuple has different features. From the point of evolution, the sequence of human genome has strong noise background due to the pressure of neutral random mutation [5], whereas, the frequency and correlation distributions of nearest neighboring nucleotides of the human genome have hierarchical features which are distinctly different from uncorrelated random sequences because of the pressure of natural selection. Therefore, the distribution of frequency and correlation of nearest neighboring nucleotides can be explained by the long-term evolution of the human genome that is dominated by random mutation and the natural selection.

In conclusion, the Zipf's analysis of the frequency and correlation of nearest neighboring nucleotides of 12 human chromosomes tells us that all chromosomes have common features which are analogous to natural language, and this is the result of long-term evolution of the human genome. This work proposes the correlation distribution of nearest neighboring nucleotides for the first time and extends the descriptor about nearest neighboring nucleotides. Therefore, this work is useful to describe some statistical properties of human chromosomes.

V. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No.20173023 and No.90203012) and the Specialized Research Fund for the Doctoral Program of Higher Education of China

DOI:10.1088/1674-0068/22/01/27-33

(No.20020730006).

[1] S. G. Gregory, K. F. Barlow, K. E. McLay, R. Kaul, D. Swarbreck, A. Dunham, C. E. Scott, K. L. Howe, K. Woodfine, C. C. A. Spencer, M. C. Jones, C. Gillson, S. Searle, Y. Zhou, F. Kokocinski, L. McDonald, R. Evans, K. Phillips, A. Atkinson, R. Cooper, C. Jones, R. E. Hall, T. D. Andrews, C. Lloyd, R. Ainscough, J. P. Almeida, K. D. Ambrose, F. Anderson, R. W. Andrew, R. I. S. Ashwell, K. Aubin, A. K. Babbage, C. L. Bagguley, J. Bailey, H. Beasley, G. Bethel, C. P. Bird, S. Bray-Allen, J. Y. Brown, A. J. Brown, D. Buckley, J. Burton, J. Bye, C. Carder, J. C. Chapman, S. Y. Clark, G. Clarke, C. Clee, V. Cobley, R. E. Collier, N. Corby, G. J. Coville, J. Davies, R. Deadman, M. Dunn, M. Earthrowl, A. G. Ellington, H. Errington, A. Frankish, J. Frankland, L. French, P. Garner, J. Garnett, L. Gay, M. R. J. Ghori, R. Gibson, L. M. Gilby, W. Gillett, R. J. Glithero, D. V. Grafham, C. Griffiths, S. Griffiths-Jones, R. Grocock, S. Hammond, E. S. I. Harrison, E. Hart, E. Haugen, P. D. Heath, S. Holmes, K. Holt, P. J. Howden, A. R. Hunt, S. E. Hunt, G. Hunter, J. Isherwood, R. James, C. Johnson, D. Johnson, A. Joy, M. Kay, J. K. Kershaw, M. Kibukawa, A. M. Kimberley, A. King, A. J. Knights, H. Lad, G. Laird, S. Lawlor, D. A. Leongamornlert, D. M. Lloyd, J. Loveland, J. Lovell, M. J. Lush, R. Lyne, S. Martin, M. Mashreghi-Mohammadi, L. Matthews, N. S. W. Matthews, S. McLaren, S. Milne, S. Mistry, M. J. F. M oore, T. Nickerson, C. N. O'Dell, K. Oliver, A. Palmeiri, S. A. Palmer, A. Parker, D. Patel, A. V. Pearce, A. I. Peck, S. Pelan, K. Phelps, B. J. Phillimore, R. Plumb, J. Rajan, C. Raymond, G. Rouse, C. Saenphimmachak, H. K. Sehra, E. Sheridan, R. Shownkeen, S. Sims, C. D. Skuce, M. Smith, C. Steward, S. Subramanian, N. Sycamore, A. Tracey, A. Tromans, Z. Van Helmond, M. Wall, J. M. Wallis, S. White, S. L. Whitehead, J. E. Wilkinson, D. L. Willey, H. Williams, L. Wilming, P. W. Wray, Z. Wu, A. Coulson, M. Vaudin, J. E. Sulston, R. Durbin, T. Hubbard, R. Wooster, I. Dunham, N. P. Carter, G. McVean, M. T. Ross, J. Harrow, M. V. Olson, S. Beck, J. Rogers, and D. R. Bentley, Nature **441**, 315 (2006).

[2] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. G. Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Qi Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S Hladun, D Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu, Science **291**, 1304 (2001).

[3] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[4] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature **356**, 168 (1992).

[5] L. Luo, W. Lee, L. Jia, F. Ji, and L. Tsai, Phys. Rev. E **58**, 861 (1998).

[6] D. Holste, I. Grosse, S. Beirer, P. Schieq, and H. Herzel, Phys. Rev. E **67**, 061913 (2003).

[7] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. Lett. **73**, 3169 (1994).

[8] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **52**, 2939 (1995).

[9] C. Martindale and A. K. Konopka, Compu. Chem. **20**, 35 (1996).

[10] A. A. Tsonis, J. B. Elsner, and P. A. Tsonis, J. Theor. Biol. **184**, 25 (1997).

[11] D. B. Searls, Nature **420**, 211 (2002).

[12] P. A. Tsonis and A. A. Tsonis, Complexity **7**, 13 (2002).

DOI:10.1088/1674-0068/22/01/27-33

[13] S. P. Li, K. L. Ng, and M. C. Chung, Physica A **321**, 189 (2003).
[14] J. K. Kim, S. I. Yang, Y. H. Kwon, and E. I. Lee, Chaos. Soliton. Fract. **23**, 1795 (2005).
[15] L. Zhang and T. Sun, Chaos, Solitons & Fractals **23**, 1077 (2005).
[16] L. Frappat and A. Sciarrino, Physica A **369**, 699 (2006).
[17] G. J. Russell and J. H. Subak-Sharpe, Nature **266**, 533 (1977).
[18] R. Nussinov, J. Biol. Chem. **256**, 8458 (1981).
[19] R. Nussinov, Nucleic Acids Res. **12**, 1749 (1984).
[20] S. Katlin and C. Burge, Trends Genet. **11**, 283 (1995).
[21] H. Nakashima, K. Nishikawa, and T. Ooi, DNA Res. **4**, 185 (1997).
[22] A. J. Gentles and S. Karlin, Genome Res. **11**, 540 (2001).
[23] R. D. Blake, S. T. Hess, and S. Nicholson-Tuell, J. Mol. Evol. **34**, 189 (1992).
[24] B. R. Morton and M. T. Clegg, J. Mol. Evol. **41**, 597 (1995).
[25] B. R. Morton, Mol. Biol. Evol. **14**, 189 (1997).
[26] P. F. Arndt, C. B. Burge, and T. Hwa, J. Comput. Biol. **10**, 313 (2003).
[27] G. Lunter and J. Hein, Bioinformatics **20**, 216 (2004).
[28] J. Hong, Nucleic Acids Res. **18**, 1625 (1990).
[29] B. V. L. S. Prasad and M. C. Vemuri, J. Biosci. **23**, 255 (1998).
[30] G. K. Zipf, *Human Behavior and the Principal of Least Effort*, Cambridge: Addison-Wesley, MA, (1949).
[31] I. Kanter and D. A. Kessler, Phys. Rev. Lett. **74**, 4559 (1995).
[32] M. A. Golberg and H. A. Cho, *Introduction to Regression Analysis*, Southampton: WIT Press, (2004).
[33] A. P. Bird, Nucleic Acids Res. **8**, 1499 (1980).
[34] K. J. Fryxell and W. J. Moon, Mol. Biol. Evol. **22**, 650 (2005).
[35] M. Kimura, Nature **217**, 624 (1968).
[36] J. L. King and T. H. Jukes, Science **164**, 788 (1969).

DOI:10.1088/1674-0068/22/01/27-33