Research Article

BIOMEDICAL Journal of Scientific & Technical Research

ISSN: 2574 -1241

DOI: 10.26717/BJSTR.2019.18.003164

Application of General Information Model to DNA Identification

Li Ying Wang[1], Meng Suo[1] and Yong Chang Huang[1,2]*

[1]*Institute of Theoretical Physics, China*

[2]*CCAST (World Lab), China*

***Corresponding author:** Yong Chang Huang, Institute of Theoretical Physics, China

ARTICLE INFO

Received: May 24, 2019

Published: May 31, 2019

Citation: Li Ying Wang, Meng Suo, Yong Chang Huang. Application of General Information Model to DNA Identification. Biomed J Sci & Tech Res 18(3)-2019. BJSTR. MS.ID.003164.

Keywords: General Information Model; Information Quantity; k-Tuple; Frequency PACS number(s): 87.10. +e, 02.50.-r, 05.40. +j, 87.15.-v

ABSTRACT

We deduce the Fisher information equation in terms of DNA research from its formal definition. By finding the similar rules corresponding to different methods to measure information, we apply the general information model to four kinds of information quantities and attain respective equations in DNA research domain. Then we find that the essence of information quantities is the measurement of similar rules of systems, e.g., we discover that Shannon's entropy is a kind of description on the combination of some different similar rules with respective weights. Because of the difference of measurement, the similar rules of Shannon's entropy differ from those of Fisher information. We discuss the relations between each of the four information quantities and the k-tuple word length *k* of DNA sequences of 16 typical genomes, respectively. Then we find that the difference between Shannon's entropy versus *k* and Fisher information versus *k* may be in accordance with the discrepancy between their similar rules. Since DNA sequences are generally close to random sequences, the similar rules of Shannon's entropy of different living organisms in word domain are similar and such similarity provides a convincing explanation to the universal linear relation existing among the studied species.

Introduction

Research on DNA sequences is the focus of modern life science. Analysis of statistical attributes of DNA sequences is significant for evolutional biology and for technologies to identify living organisms. Several attempts have been made to identify relatively small size (microbial) genomes by using the distribution of the appearance of short consecutive nucleotide strings of length *k* called k-tuple [1-7]. To describe the distribution, many scholars used information quantities including Shannon entropy and Fisher information [8-16]. Some scholars also utilized some nonlinear methods or models to analyze the DNA sequences [17,18]. In this paper, we generalize the geographical remote sensing information model [19] to a general information model which is image joining equation calculated. Such model is a grey non-linear equation that is from formal logic inferring to dialectical logic calculation and from abstract thinking to both of abstract and visual thinking [19]. We apply the general information model to deduce four kinds of information quantities used in DNA sequence analysis and then analyze the statistical results cited from Ref. [15] on the scope of similar rules.

General Information Model

The remote sensing information model is established according to geographical regularity and complexity. This model, a grey nonlinear equation, combines both certainty and uncertainty in an equation from the point of view of dialectical logic. In geographic research, scholars usually transform non-remote sensing data into images and then generalize the remote sensing information model to the geographical image information model. Here, we take any data that can be portrayed as images into account and apply the model further to any field that is characteristic of complexity and non-linearity as geography. Then we obtain a general information model. Based on the analysis in terms of the formal logic principle, (Table 1) gives the analysis of the remote sensing information model and the general information model on intension and extension of certainty and uncertainty [19]. In logic, intension is the set of attributes constituting the meaning of a term [20].

Table 1: Intension and Extension of certainty and uncertainty.b

Formal Logic	Mathematics System	Mathematics Method
Extension of certainty & intension of certainty	White system	Mathematics- Physics equation
Extension of uncertainty & intension of certainty	Fuzzy system	Fuzzy equation
Extension of certainty & intension of uncertainty	Grey system	Grey equation
Extension of uncertainty & intension of uncertainty	Black system	Stochastic equation

It is often contrasted with extension, which is the class of objects to which the term applies. For a particular research subject, intension indicates all parameters that affect the results, and extension indicates the set of objects that the subject refers to. As Table 1 shows, when both intension and extension are certain, the studied phenomenon can be considered as a white system and it can be solved by using Mathematics- Physics equations. When both intension and extension are uncertain, the studied phenomenon can be considered as a black system and it can be solved by using stochastic equations. When intension is certain and extension is uncertain, the studied phenomenon is the transition of two certain phenomena, it can be considered as a fuzzy system and can be solved by using fuzzy equations. When intension is uncertain and extension is certain, the studied phenomenon is known incompletely, it can be considered as a grey system and can be solved by using grey equations. According to complexity of studied phenomena, a multifactor function y influencing phenomena can be expressed as [19].

$$y = f(x_1, x_{2,}..., x_n) \qquad (1)$$

where $x_1, x_{2,}..., x_n$ are n influencing factors. By doing dimensional analysis on y, $x_1, x_{2,}..., x_n$ and applying the generalized π theorem which breaks the limitation of centimeter-gram-second system and gives the similar descriptions of all relative kinds of phenomena, one can attain the similar rules $\pi_y, \pi_{x_1}, \pi_{x_2,}..., \pi_{x_n}$ [19]. Writing the similar rules as a general non-linear expression, one can generally obtain [19]

$$\pi_y - a_0 \pi_{x_1}^{a_1} \pi_{x_2}^{a_2} ... \pi_{x_n}^{a_n} = 0 \qquad (2)$$

where π_y is a non-dimensional factor group of the seeking phenomena, in which factors may have definite physical, ecological, environmental, genetic etc significances, respectively, for different studied phenomena. In Eq. (2), we call a_0 as information coefficient and $a_1, a_2, \cdots, a_n$ as information exponents. To find out the information coefficient and information exponents, one can rewrite Eq. (2) as a logarithmic equation [19]

$$\lg \pi_y = \lg a_0 + a_1 \lg \pi_{x_1} + a_2 \lg \pi_{x_2} + \cdots + a_n \lg \pi_{x_n} \qquad (3)$$

$$Y = X_0 + a_1 X_1 + a_2 X_2 + \cdots + a_n X_n \qquad (4)$$

It is easy to see that Eq. (4) is a general multiple regression equation, one can solve the equation to obtain the value of X_0, $a_1, a_2, \cdots, a_n$ and solve $X_0 = \lg a_0$ to obtain the value of a_0. When both the two sides of Eq. (3) equate to 1, it is easy to see that $a_1, a_2, \cdots, a_n$ means the weight, the belonging degree in fuzzy mathematics [19].

In Eq. (2), such non-dimensional factor groups (π_{x_i}) are some similar rules in white system resulting from causal analysis, which satisfies quantitative causal relation [21-23], e.g. Ref.[24] uses the no-loss-no-gain homeomorphic map transformation satisfying the quantitative causal relation to gain exact strain tensor formulas in Weitzenböck manifold. In fact, some changes (causes) of some quantities in Eq. (2) must result in the relative some changes (results) of the other quantities in Eq. (2), so that Eq. (2)'s right side keeps no-loss-no-gain, i.e., zero, namely, Eq. (2) also satisfies the quantitative causal relation. The ellipses in Eqs. (2-4) mean the factors which have not been included in the equations for the lack of actual measured data or having not been recognized [19]. Thus, Eq. (2) leaves room for improving description on studied phenomena. With further research, some new similar rules might be found and then they can be added into the general information model [19]. Before finding the new similar rules, they are included in the information coefficient phenomenologically. The analysis of the information exponents suggests that:

(1) when $a_i (i = 1,2,\cdots,n) = 0$ there is no relation between $\pi_{x_i}^{a_i}$ and π_y; (2) when $a_i (i = 1,2,\cdots,n) = 1$, there is a linear relation between $\pi_{x_i}^{a_i}$ and π_y; (3) when $a_i (i = 1,2,\cdots,n) = \frac{N}{M}$ (fraction), there is a fractal self-similar relation between $\pi_{x_i}^{a_i}$ and π_y; (4) when $a_i (i = 1,2,\cdots,n) = \frac{N(x,y,z,t)}{M(x,y,z,t)}$, there is a general space-time relation between $\pi_{x_i}^{a_i}$ and π_y [19]. From the above review, it is easy to see that both the completely determinative physical equations and the completely indeterminate stochastic equations are special cases of Eq. (2). Both the certainty and the uncertainty are combined in this grey non-linear equation, in which each similar rule, the information coefficient and each information exponent can be depicted as images. So, the general information model is a dialectical logic calculation combining abstract thinking with visual thinking as Figure 1 shows [19].

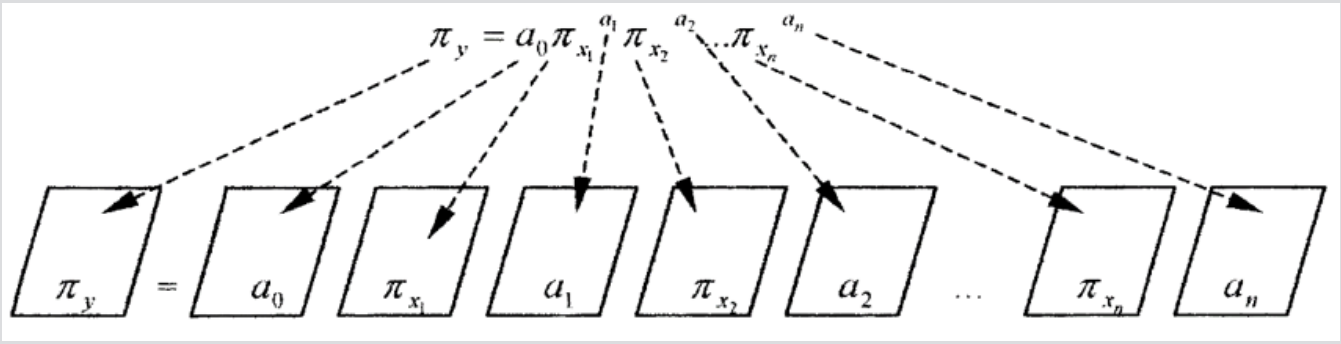

Figure 1: Correspondence between images and data.

Four Kinds of Information Calculation Methods of DNA in terms of the General Information Model

As long as we find the similar rules of different information quantities according to their significances of informatics, we can achieve their particular equations by using the general information model. In this section, we deduce four kinds of information quantity equations usually applied in the DNA sequence research.

The nucleic acid sequence can be regarded as a linear text composed by four vocabularies—A (adenine), C (cytosine), G (guanine), T (thymine). A segment of length k (3~9) of nucleic acid sequence is called as a k-tuple [15]. There are 4^k kinds of k-tuples and we can count the frequency (of occurrence) of each k-tuple by moving the window of length k with step length 1 along a DNA sequence [15]. It may be conjectured that the distribution of k-tuples in a genome would be the "equivalent representation" of the genome when k is big enough, namely, the genome sequence can be determined by the distribution of k-tuples uniquely [7].

Shannon's Entropy in Word Domain

There are 4^k kinds of k-tuples for a certain k. For example, when k = 2, there are 16 kinds of 2-tuples (AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT). We use f_i to denote the frequency of the i th kind of k-tuple and N to denote the total frequency of k-tuples, determined by the equation: N= the length of the genome sequence – k + 1. Considering the similar rules as the probability of the frequency f_i of each k-tuple, we attain $\pi_{x_i} = \frac{f_i}{N}$, $a_0 = 1$, ($i = 1,2,\cdots 4^k$) where the minus is due to $\log_2 \frac{f_i}{N} < 0$. Then Eq. (3) can be embodied as following:

$$H_k = -\sum_{i=1}^{4^k} \frac{f_i}{N} \log_2 \frac{f_i}{N} \qquad (5)$$

This is just Shannon's entropy equation and is the same as the result in Ref. [15]. Therefore, we first discover that Shannon's entropy is a kind of description on the combination of some different similar rules with respective weights.

Logarithm of Rate of Standard Deviation and Mean value in word domain (log*DTM*)

Considering the quantity of A+T in a k-tuple, we can divide the k-tuple into k+1 subsets, each of which called the m th subset [15]. For example, when k = 2, there are 3 subsets: (1) m = 0, there are four 2-tuples including neither A nor T—CC, CG, GC, GG; (2) m =1, there are eight 2-tuples including one A or T; (3) m =2, there are four 2-tuples only including A or T. We use L_{km} to denote the length of the m th subset for a certain k and f_{km} to denote the average frequency of all k-tuples for a certain k in the m th subset [15]. The corresponding standard deviation is d_{km}^2 [15]. Regarding the similar rules as the contribution of the standard deviation d_{km}^2 to the average frequency f_{km} of each m th subset, we obtain that $\hbar \pi_{x_i} = \frac{L_{km}}{N} \frac{d_{km}^2}{f_{km}^2}$ namely, $\pi_{x_i} = e^{\frac{L_{km}}{N} \frac{d_{km}^2}{f_{km}^2}}$, $a_0 = 1$, $a_i = 1$ (i = 0,1,..., k+1). Then Eq. (3) can be embodied as

$$\log DTM = \hbar \sum_m \frac{L_{km}}{N} \frac{d_{km}^2}{f_{km}^2} \quad (\sum_m L_{km} = N, \quad k=3\sim 9) \qquad (6)$$

This is the log *DTM* equation, and is the same as the result in Ref. [15]. It could be proved that log *DTM* is related to Shannon's entropy.

Shannon's Entropy in Frequency Domain

We use m_i to denote the number of k-tuples occurring in a certain interval and T to denote the number of intervals: T = the total sequence length / the interval length. Considering the similar rules as the probability of the number of k-tuples occurring in different intervals m_i, we gain that $\pi_{x_i} = \frac{m_i}{4^k}$, $a_0 = 1$, $a_i = -\frac{m_i}{4^k}$ ($i = 1,2,\cdots,T$) . Then Eq. (3) can be incorporated as

$$H_k = -\sum_{i=1}^{T} \frac{m_i}{4^k} \log_2 \frac{m_i}{4^k} \qquad (7)$$

This is Shannon's entropy equation in frequency domain, and is the same as the result in Ref. [15].

Fisher Information Quantity in Frequency Domain

So far, Fisher information quantity equation used in research on DNA sequence was defined directly according to its formal definition in statistics. We deduce the result firstly. As Fisher information quantity is defined

$$F = \int p(x)\left(\frac{\partial \ln p(x)}{\partial x}\right)^2 dx = \int \frac{1}{p(x)}\left(\frac{\partial p(x)}{\partial x}\right)^2 dx \qquad (8)$$

we discretize Eq. (8) to

$$F = \sum_i \frac{1}{p(x_i)} \frac{(p(x_{i+1}) - p(x_i))^2}{\Delta x_i} \qquad (9)$$

We may take $p_i = \frac{\Delta x_i}{4^k} \cdot \frac{m_i}{L}$, which represents the probability of the number of k-tuple occurrence m_i when the total interval length L changed by Δx_i. Then Eq. (9) can be rewritten as

$$F = \sum_i 4^{-k}(m_{i+1} - m_i)^2 / m_i \qquad (10)$$

This is Fisher information quantity equation in frequency domain and is the same as the result in Ref. [15]. Thus, we complete the deduction of the Fisher information equation in terms of DNA research from its formal definition. Considering the similar rules as the distribution of k-tuples in different intervals measured by Fisher information quantity in frequency domain, we get that $\ln \pi_{x_i} = 4^{-k}(m_{i+1} - m_i)^2 / m_i$, namely, $\pi_{x_i} = e^{4^{-k}(m_{i+1}-m_i)^2/m_i}$, $a_0 = 1$, $a_i = 1$ ($i = 0,1,\cdots 4^k$). Then Eq. (3) can be embodied as Eq. (10).

Based on above discussion, it is easy to see that the general information model is the uniform expression of different information measurements. Such uniform expression is not merely the combination of four equations but a model that reveals the essence of information quantity-the measurement of similar rules of systems. Put it clearly, Shannon's entropy is the measurement of frequency occurrence of different k-tuples in word domain; log *DTM* is the measurement of the deviation of k-tuples frequency to mean value in different *m* th subset; Shannon's entropy in frequency domain is the measurement of the number of k-tuples occurring in different intervals in global measurement; Fisher information in frequency domain is the measurement of the distribution of k-tuples occurring in different intervals in local measurement. On the ground of the correspondence of different similar rules to their information measurements, we can achieve the respective equations.

Application of the General Information Model to DNA Identification

In research of living organisms, humans understand the macrocosm more concretely and amply than they know the microcosm. The general information model (which integrates deduction, induction and analogy into a whole and is characterized with image calculating function) can satisfy the demand of the status of modern research well and is helpful to the identification and analysis of DNA sequences. Ref. [15] shows the statistical results of the four information quantities calculated by Eqs. (5), (6), (7) and (10). The four information quantities are denoted by $H^{(1)}$, $H^{(2)}$, $H^{(3)}$ and $H^{(4)}$, respectively. The relations between the four kinds of information quantities and the k-tuple word length *k* of DNA sequences of 16 typical genomes are displayed in Figure 2 [15].

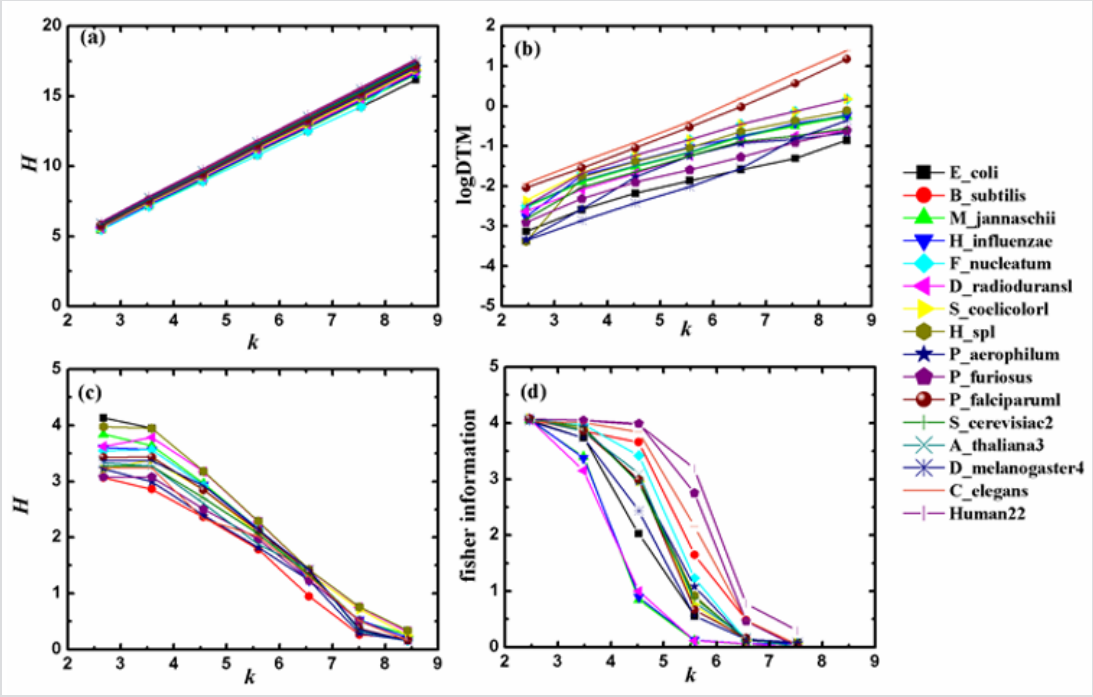

Figure 2: The relations between four kinds of information quantities and the word length *k*.
(a) $H^{(1)}$ versus *k* (b) $H^{(2)}$ versus *k* (c) $H^{(3)}$ versus *k* (d) $H^{(4)}$ versus *k*

According to the results shown in Figure 2 (a-c), it is easy to find that the three Shannon's information quantities have good linear relation with *k* (For the relation between log *DTM* and Shannon's entropy, log *DTM* can be seen as a kind of Shannon's entropies). Moreover, such linear relation is universal for different species on the whole. Shannon's entropy is the measurement of uncertainty about a stochastic variable. It only depends on the certain probability distribution of the stochastic variable rather than its value. In the perspective of the general information model, similar rules of different information quantities are relative to the probability distribution of their respective stochastic variables, and the corresponding information exponents describe the weight each similar rule contributes. The above graphs show that the linear relation between $H^{(1)}$ and k is the most strong among the all. For $H^{(1)} = k + b$, the regression coefficient a is close to 2 and the interception b is close to 0. For a random sequence with infinite length, $H^{(1)} = 2k$ [7]. This suggests that $H^{(1)}$ of different species is close to that of random sequences; namely, their corresponding similar rules, information coefficients and information exponents are close to those of the random sequences. Actually, the neutral mutation causes all DNA sequences of different species close to random ones [16]. Thus, the similar rules, information coefficients and information exponents of them are alike, so the relation between $H^{(1)}$ and *k* shows the universality which is nearly independent of species.

As Figure 2d shows, the relation between Fisher information $H^{(4)}$ and *k* varies with different species, the level of linearity of which is heterogeneous with Shannon's entropy. The reason is that the two measure information by using different methods in different measurements. The heterogeneity between their corresponding similar rules and information exponents results in the difference of relation between information quantities and *k*. For other two information quantities, the forms of their equations are different with Shannon's entropy though, the three are defined by the same measurement method, i.e., the measurement of uncertainty about a stochastic variable associated with a certain probability distribution. Therefore, their similar rules and information exponents have commonness to some extent which may explain the similar linear relation.

Summary and Conclusion

This paper generalizes the concept of geographical remote sensing information model and extends it to a general information model. This grey non-linear equation has many advantages to the application in molecular biological research which contains both certainty and uncertainty associated with a myriad of data that may be described in images. We deduce the Fisher information equation in DNA domain from its formal definition. Then we obtain the equations of four kinds of information quantities in terms of DNA research by using the general information model and find that all kinds of information quantities we discussed are measurements of the similar rules of systems. The contribution of each similar rule to the whole system can be described in the form as the a_i th power of π_{x_i} and a_i can be considered as the weight.

The statistical data and graphs of 16 typical genomes cited from Ref. [15] show that three kinds of Shannon's information quantities (Shannon's entropy in word domain and in frequency domain and log *DTM*) have good linear relation between them and *k* but Fisher information has not. We explain the results from a new perspective the different statistical results are the manifestation of the heterogeneity between similar rules of Shannon's entropy and those of Fisher information, which results from the difference between measurements of the two when people describe the smoothness of their probability functions $p^{(X)}$. On the other hand, the universality of the linear relation between Shannon's entropy and k suggests that the more the similar rules, information coefficients and information exponents of different systems are similar, the more their characters and qualities are similar. Such universality also demonstrates that the general information model is an effective and advantageous way to analyze these problems regarding DNA research.

Acknowledgement

The authors are grateful for Prof. A. N. Ma for useful discussion and comment. The work is supported by National Natural Science Foundation of China (Grant No. 11875081).

References

1. Y Fofanov, Y Luo, C Katili, Wang J, Belosludtsev Y, et al. (2004) How independent are the appearances of n-mers in different genomes? Bioinformatics 20(15): 2421-2428.
2. R Nussinov (1984) Doublet frequencies in evolutionary distinct groups. Nucleic Acids Res 12(3): 1749-1763.
3. S Karlin, I Ladunga (1994) Comparisons of eukaryotic genomic sequences. Proc Natl Acad Sci USA 91: 12832-12836.
4. R Sandberg, G Winberg, CI Branden, Kaske A, Ernberg I, et al. (2001) Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. Genome Res 11(8): 1404-1409.
5. S Karlin, J Mrazek, AM Campbell (1997) Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179(12): 3899-3913.
6. H Nakashima, K Nishikawa, T Ooi (1994) Differences in dinucleotide frequencies of human, yeast and Escherichia coli genes. DNA Res 4(3): 185-192.
7. HM Xie, BL Hao (2003) Visualization of k-tuple distribution in procaryote complete genomes and their randomized counterparts. IEEE Proc comp sys Bioinf 24(7): 31-42.
8. I Grosse, P Bernaola Galván, P Carpena, Román Roldán R, Oliver J, et al. (2002) Analysis of symbolic sequences using the Jensen-Shannon divergence. Phys Rev E 65: 041905.
9. G Pesole, M Attimonelli, C Sacconne (1994) Linguistic approaches to the analysis of sequence information. Trends Biotech 12(10): 401-408.
10. P Bernaola Galván, I Grosse, P Carpena, Oliver JL, Román Roldán R, et al. (2000) Finding borders between coding and noncoding DNA regions by an entropic segmentation method. Phys Rev Lett 85(6): 1342-1345.

11. CH Chang, LC Hsieh, TY Chen, Hong Da Chen, Liaofu Luo, et al. (2005) Shannon information in complete genomes. Journal of Bioinformatics and Computational Biology 3(3): 587-608.

12. HD Chen, CH Chang, LC Hsieh, HC Lee (2005) Divergence and Shannon information in genomes. Phys Rev Lett 94(17): 178103.

13. AO Schmitt, H Herzel (1997) Estimating the entropy of DNA sequences. Journal of Theoretical Biology 188(3): 369-377.

14. BR Frieden (2004) Science from Fisher information. A Unification: Cambridge University Press.

15. LF Luo (2005) Discussion on the universality of information content in DNA sequences. Journal of Hefei University (Natural Sciences) 15(1): 1-6.

16. LF Luo, WJ Lee, LJ Jia, Fengmin Ji, Lu Tsai (1998) Statistical correlation of nucleotides in a DNA sequence. Phys Rev E 58(1): 861-871.

17. YH Chen, SL Nyeo, CY Yeh (2005) Model for the distributions of k-mers in DNA sequences. Phys Rev E 72(1): 011908.

18. J Barral P, A Hasmy, J Jiménez, A Marcano (2000) Nonlinear modeling technique for the analysis of DNA chains. Phys Rev E 61(2): 1812-1815.

19. AN Ma (2001) Remote sensing information model and geographic mathematics. Acta Scientiarum Naturalium Universitatis Pekinensis 37(4): 557-562.

20. (2004) The American Heritage Dictionary of the English Language, Fourth Edition. Houghton Mifflin Company.

21. YC Huang, XG Lee, MX Shao (2006) Mod Phys Lett A21: 1107; L Liao, YC Huang (2007) Phys Rev D75 025025.

22. YC Huang, FC Ma, N Zhang (2004) Mod Phys Lett B18: 1367.

23. YC Huang, CX Yu (2007) Phys Rev D75 (2007) 044011.

24. YC Huang, BL Lin (2002) Phys Lett A299: 644.

ISSN: 2574-1241

DOI: 10.26717/BJSTR.2019.18.003164

Yong Chang Huang. Biomed J Sci & Tech Res

This work is licensed under Creative Commons Attribution 4.0 License

Submission Link: https://biomedres.us/submit-manuscript.php

Assets of Publishing with us

- Global archiving of articles
- Immediate, unrestricted online access
- Rigorous Peer Review Process
- Authors Retain Copyrights
- Unique DOI for all articles

https://biomedres.us/