Journal of Data Mining in Genomics & Proteomics - Open Access

Research Article
OPEN ACCESS Freely available online
doi:10.4172/2153-0602.1000105

Complexity and Entropy Analysis of DNA Methyltransferase

Xiaoli Xie[1,2], Ying Yu[2], George Liu[3], Zhifa Yuan[1] and Jiuzhou Song[2]*

[1]College of Science, Northwest A&F University, Yangling, 712100 P. R. China
[2]Department of Animal & Avian Sciences, University of Maryland, College Park, MD 20742, USA
[3]BFGL-ANRI, ARS, USDA, BARC-East, Beltsville, MD 20705-2350

Abstract

Background: The application of complexity information on DNA sequence and protein in biological processes are well established in this study. Available sequences for *DNMT1* gene were thoroughly explored in the information complexities. *DNMT1* gene is a maintenance methyltransferase responsible for copying DNA methylation patterns to the daughter strands during DNA replication in different species.

Results: We found that the entropy of *DNMT1* gene in different species is DNA base composition dependent, and its complexity in mammals is lower in introns than in coding regions. We also demonstrated the impacts of entropy on domains and non-domain**(s)** of the *DNMT1* gene. The results from DNA and protein sequences indicated that DNA evolution has a tendency toward complexity. The most interest is that the methylation's changes of the gene over aging in a unique chick model showed aging-driven entropy characteristics, which may give an explanation of aging processes.

Conclusion: In summary, the information complexity of *DNMT1* gene is related to its genomic composition, which thereby associates to evolutionary and aging processing, even though the intrinsic mechanism is not to be studied yet.

Keywords: Entropy; Complexity; DNMT1; Epigenetic; Evolution

Introduction

DNA methylation is a main part of epigenetics. It refers to heritable change in gene expression without alteration in DNA sequence, and plays an important role in regulation of gene expression, chromatin structure, normal development, cell differentiation, X chromosome inactivation and genomic imprinting [1]. In higher organisms, the process of DNA methylation is regulated by three DNA methyltransferases, *DNMT1*, *DNMT3a*, and *DNMT3b*. Functionally, *DNMT1* is a maintenance methyltransferase that is responsible for copying DNA methylation patterns to the daughter strands during DNA replication [2], whereas *DNMT3a* and *DNMT3b* are *de nove* methyltransferases [3], and transcriptional repressors with unique localization properties of heterochromatin [4,5]. Most importantly, the methyltransferases are closely related to speciation process, their activities vary among tissues, different cells and development stages as well as aging [6-9]. However, the complexities and structural characteristics of these genes are still unknown among species; the information analysis in entropy view may thus help us elucidate mechanisms and effects of the genes in epigenetic processes.

In present research, we hypothesize that information measure of the gene is related to its genomic complexity, which thereby is associated to relevant biological processing. We try to test the assumption with the information analysis based on available sequences from different species. Among the three methyltransferases, because the available information for *DNMT1* gene is relatively complete and more conserved than *DNMT3a* and *DNMT3b*, we thus focused only on *DNMT1* gene with Shannon entropy. The information measure was introduced by Claude E. Shannon [10]. Theoretically, it reflects an uncertainty associated with a random variable and quantifies the information contained in a message. Complexity of symbolic sequence reflects an ability to represent a compact form based on some structural features of this sequence, indicating regularity or randomness of sequence that encodes complex structure. Right now there are several methods to measure the information complexity analysis, including entropy [11], generalized complexity [12], clustering of cryptically simple sequence stochastic complexity [13], alphabet capacity l-gram [14], generalized lattice graphs [15], promoter QSAR model [16] and grammatical complexity [17]. They are being widely applied in many fields, such as engineer, biology, medical, agriculture and forestry, etc. [18-23]. Among these methods, entropy measure is the simplest method using only the symbol frequencies.

Generally speaking, complex sequences give rise to complex structures. As an important gene controlling DNA methylation, the complexity and structural information stored in the *DNMT1* gene should further be studied whether evolution has a tendency towards complexity. In this paper, we initially explored information complexities in different species, analyzed their DNA sequence, protein sequence, domains and non-domain region coded by the gene as well as checked its methylation status in a unique chick model, which we attempted to elucidate the relationship between complexity information indices of *DNMT1* gene and biological processing.

Methods and Material

Dataset resource

mRNA sequences, genomic DNA sequences and protein sequences of gene *DNMT1* in different species were obtained from the NCBI website (Seen supplementary materials). Intron sequences of *DNMT* were gotten from the website http://genome.ucsc.edu/cgi-

***Corresponding author:** Jiuzhou Song, Department of Animal & Avian Sciences, University of Maryland, College Park, MD 20742, USA, Tel: (301) 405-5943; Fax: (301) 405-7980; Email: songj88@umd.edu

Received November 28, 2010; **Accepted** December 27, 2010; **Published** December 30, 2010

Citation: Xie X, Yu Y, Liu G, Yuan Z, Song J (2010) Complexity and Entropy Analysis of DNA Methyltransferase. J Data Mining in Genom Proteomics 1:105. doi:10.4172/2153-0602.1000105

bin/hgBlat?command=start and organized by perl language. Domain regions of protein sequence of different species were obtained from the website http://pfam.sanger.ac.uk/.

Nucleotide composition

We first calculated the GC content and AT content, which are the measurements of analysis of nucleotide composition [24,25]. Both are defined as:

$$AT(\%)=\frac{\sum_{i=1}^{n}A_i+\sum_{i=1}^{n}T_i}{n} \quad GC(\%)=\frac{\sum_{i=1}^{n}G_i+\sum_{i=1}^{n}C_i}{n} \tag{1}$$

where n is the length of window size, A_i is 1 if there is a nucleotide A in the i^{th} position, and 0 otherwise (the terms T_i, C_i and G_i are defined similarity), so $\sum A_i, \sum T_i, \sum C_i, \sum G_i$ are the numbers of the four nucleotides over the window size. The GC content and AT content depend on the sliding window size.

Kolmogorov-Smirnov test

Kolmogorov-Smirnov test is to compare the distributions of values in the two data vectors X_1 and X_2 with length n_1 and n_2, respectively, representing random samples from some underlying distribution (s). It is non-parametric and distribution free. P value was obtained by statistical toolbox software of Maltlab. Its null hypothesis (H_0) is that X_1 and X_2 are drawn from the same continuous distribution, whereas the alternative hypothesis (H_A) is that they are drawn from different continuous distributions. The H_0 is rejected if the index H is 1. Otherwise, the H_0 is accepted if the index H is 0. Statistically, the test statistic can be written as:

$$\mathrm{Max}(|F_1(x)-F_2(x)|) \tag{2}$$

where $F_1(x)$ is the proportion of X_1 values less than or equal to x and $F_2(x)$ is the proportion of X_2 values less than or equal to x. In this paper, X_1 and X_2 are GC and AT contents of two species, respectively.

Methylation analysis of *DNMT1* gene

Lines 6_3 and 7_2 White Leghorn chickens were initially selected at the Avian Disease and Oncology Laboratory in 1939 for tumor resistance or susceptibility induced by herpesvirus. Tissue samples were collected at 2 weeks and 15 months of age and stored at -20°C until analyses. We then extracted DNA, treated the genomic DNA with bisulfite, amplified with PCR and quantitatively measured the methylation level of the gene with pyrosequencing methods [8].

Information analysis

Shannon entropy is a measure of the uncertainty associated with a random variable. The Shannon entropy of a discrete random variable X taken possible value $\{x_1,\cdots,x_n\}$ is

$$\mathrm{Max}(|F_1(x)-F_2(x)|) \tag{3}$$

where $p(x_i)$ is the probability of $X = x_i$.

Supposed a symbol sequence $\{S_n\}_M^N$, M is the number of symbol in the sequence; N is the length of the sequence. According to *Renyi* definition, the q^{th} order generalized entropy is defined as

$$H_q=-\frac{1}{q-1}\log\sum_{i=1}^{M}p_i^q \tag{4}$$

Where p_i is the occurrence probability of the i^{th} symbol in the sequence? When $p = 0$, $H_0 = H_{max} = \log M$. And if $q \to 1$, $H_1=-\sum_{i=1}^{M}p_i\log p_i$. When $q = 2$, $H_2=-\log\sum_{i=1}^{M}p_i^2$. It is obvious that H_0 is the maximum Shannon entropy, which represents the sequence in a completely random situation. H_1 is Shannon entropy in information theory and refers to a uncertain measure of single character in the sequence. H_2 is an uncertain measure of repeating twice of a single character.

We can prove that generalized entropy is decrease function of q.

Based on the generalized entropy, the following formulas were given: $I = H_0 - H_1$, $GC_R(\mathrm{bit}) = H_1 - H_2$, $R=1-\frac{H_1}{H_0}$. Here, I, GC_R and R are defined as information of ordering, repeatability complexity and redundancy. To compare complexity measurement of sequences with different length in our research, relative repeatability complexity, was defined as follows:

$$GC_R(\%)=(H_1-H_2)/H_0{}^{*}100 \tag{5}$$

In addition, in the paper, we also first computed information entropy of DNA methylation levels of four CpG sites in the exon1 region of *DNMT1* according to methyaltion and unmethylation percentage.

Results

Phylogenetic relationship of *DNMT1* among several species

The mRNA sequences of gene *DNMT1* of different species were downloaded, aligned, and compared. The evolutionary tree, as shown in Figure 1, was built [26]. The distance was calculated with the Jukes-Cantor method. From the Figure 1, it is obvious that the animals are grouped based on the evolution distance of the *DNMT1* mRNA sequences. The smallest evolutionary distance is found between human and chimpanzee, followed by dog and cattle, and then the sequential phylogeny relationships between mouse and zebrafish. The results indicate that the phylogeny relationship of different mRNA sequences of *DNMT1* represents the speciation variations. Thus, we thereby inferred that the epigenetic mechanism controlled by this gene may be vital not only to speciation process, but also to evolution and development for different organisms.

Composition analysis of DNA sequence of *DNMT1* gene

To our knowledge, because of alternative splicing of exons, DNA mutations and RNA editing, there may have been relative big variations in mRNA length and nucleotide compositions for an ortholog gene among different species, which are directly related to

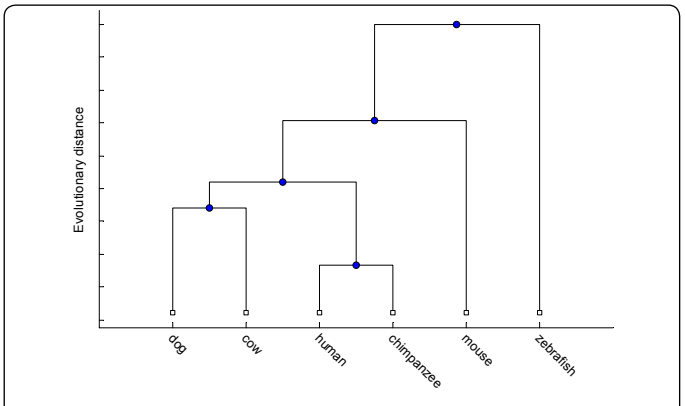

Figure 1: The phylogenetic tree of several species based on mRNA sequences of *DNMT1* gene.

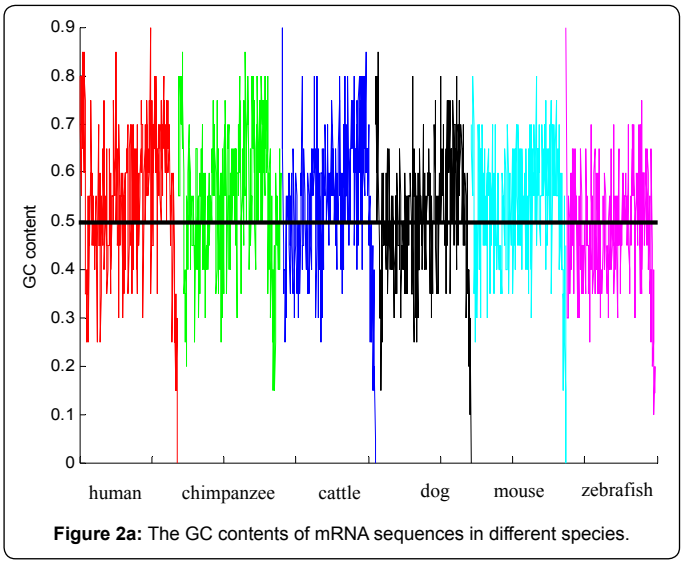

Figure 2a: The GC contents of mRNA sequences in different species.

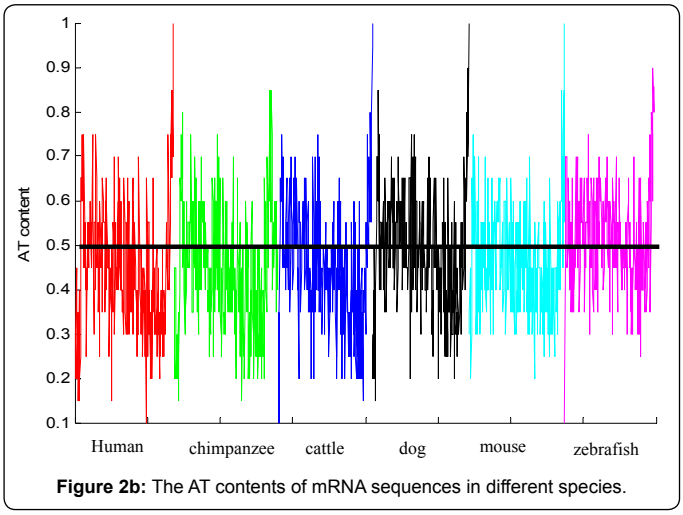

Figure 2b: The AT contents of mRNA sequences in different species.

Species	Human	Chimpanzee	Cattle	Dog	Mouse	Zebrafish
GC%	0.6029	0.6042	0.5385	0.4659	0.5833	0.3496
AT %	0.2904	0.2693	0.3115	0.3864	0.2992	0.4837

Table 1: GC content and AT content ratio in mRNA sequences in different species.

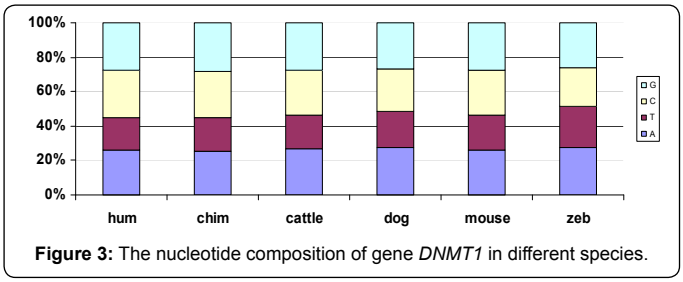

Figure 3: The nucleotide composition of gene *DNMT1* in different species.

species-specific modular formation and gene expression regulation. Therefore, to further explore the nucleotide composition of gene *DNMT1*, we first computed GC and AT contents and then plotted as shown in Figure 2a and Figure 2b. In which, the window size was optimized about 20bp without the overlap between windows. We found that the GC content is larger than 0.5 in most positions, whereas the AT content is less than 0.5 in most positions for mammals. On the other hand, AT content in zebrafish is larger than GC content over most positions (Table1). The compositions of the four DNA nucleotides were calculated in the mRNA sequences. It is obvious that nucleotide distribution of the species is not completely uniform, it varies among species. We also found that, the proportions of four nucleotides as shown in Figure 3, except for Zebrafish and the numbers of nucleotide G and C in mRNA sequence are more than that of A and T, which is consistent with result of the GC and AT contents (Table 2).

To test the distribution of GC and AT contents, a two-sample Kolmogorov-Smirnov (K-S) test was used to compare the distributions among the species. The null hypothesis (H_0) for this test is that the GC or AT contents in species A and species B are drawn from the same continuous distribution. We will accept H_0 when index H is 0. Otherwise we reject H_0 when H is 1. The K-S test results were shown in Table 3 and Table 4. We found that the distribution tests based on the GC and AT contents are the exactly same and may vary with reference. Interestingly, chimpanzee, cattle, mouse and human have the same distribution ($P>0.05$). They are significant different from zebrafish and dog ($P<0.05$, Table 3 and 4).

The entropy analysis of *mRNA* sequence of different species

Because of so different mRNA sequences, we computed the

Species	Human	Chimpanzee	Cattle	Dog	Mouse	Zebrafish
Nucleotide A(%)	0.2609	0.2551	0.2693	0.2762	0.2628	0.2765
Nucleotide T(%)	0.1872	0.1928	0.1930	0.2116	0.2028	0.2415
Nucleotide C(%)	0.2738	0.2724	0.2604	0.2453	0.2592	0.2207
Nucleotide G(%)	0.2781	0.2797	0.2764	0.2669	0.2753	0.2613

Table 2: Nucleotide composition in mRNA sequences in different species.

Species		P value	H
Human	Chimpanzee	0.9999	0
	Cattle	0.6242	0
	Dog	0.0034	1
	Mouse	0.2741	0
	Zebrafish	$3.9*10^{-9}$	1
Chimpanzee	Cattle	0.5827	0
	Dog	0.0092	1
	Mouse	0.2581	0
	Zebrafish	$2.24*10^{-8}$	1
Cattle	Dog	0.2333	0
	Mouse	0.61	0
	Zebrafish	$1.67*10^{-5}$	1
Dog	Mouse	0.0483	1
	Zebrafish	0.0387	1
Mouse	Zebrafish	$1.31*10^{-6}$	1

Table 3: The KS test of AT content.

species		P value	H
Human	Chimpanzee	0.9999	0
	Cattle	0.6242	0
	Dog	0.0034	1
	Mouse	0.2741	0
	Zebrafish	$3.9*10^{-9}$	1
Chimpanzee	Cattle	0.5827	0
	Dog	0.0092	1
	Mouse	0.2581	0
	Zebrafish	$2.24*10^{-8}$	1
Cattle	Dog	0.2333	0
	Mouse	0.61	0
	Zebrafish	$1.67*10^{-5}$	1
Dog	Mouse	0.0483	1
	Zebrafish	0.0387	1
Mouse	Zebrafish	$1.31*10^{-6}$	1

Table 4: The KS test of GC content.

Species	H_1 (bit)	GC_R(bit)	I (bit)	R(%)
Human	1.3748	0.01	0.0115	0.0083
Chimpanzee	1.3765	0.0087	0.0098	0.0071
Cattle	1.3765	0.0082	0.0094	0.0067
Dog	1.3813	0.0048	0.005	0.0036
Mouse	1.3798	0.0059	0.0065	0.0047
Zebrafish	1.3827	0.0035	0.0035	0.0026

Table 5: Entropy information of mRNA sequence of *DNMT1* gene.

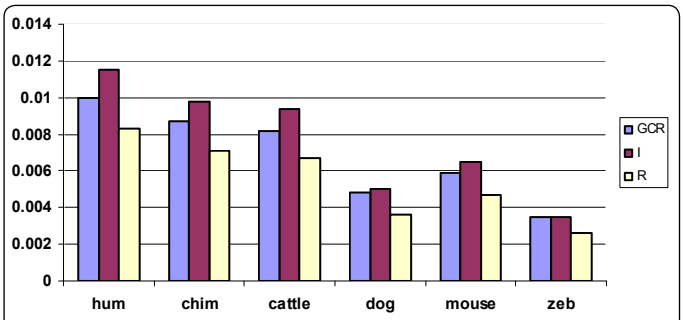

Figure 4a: The information indices of mRNA sequences in different species.

Species	H_1 (bit)	GC_R(%)	I(bit)	R(%)
Human	1v.3820	0.0043	0.0043	0.0031
Dog	1.3813	0.0047	0.0050	0.0036
Mouse	1.3803	0.0058	0.0060	0.0043
Zebrafish	1.3428	0.0423	0.0435	0.0314

Table 6: Entropy information of intron sequence of *DNMT1* gene.

Species	H_1 (bit)	GC_R(%)	I(bit)	R(%)
Human	1.3834	0.0029	0.0029	0.0021
Dog	1.3834	0.0030	0.0029	0.0021
Mouse	1.3822	0.0039	0.0041	0.0030
Zebrafish	1.3639	0.0220	0.0224	0.0162

Table 7: Entropy information of genomic sequence of *DNMT1* gene.

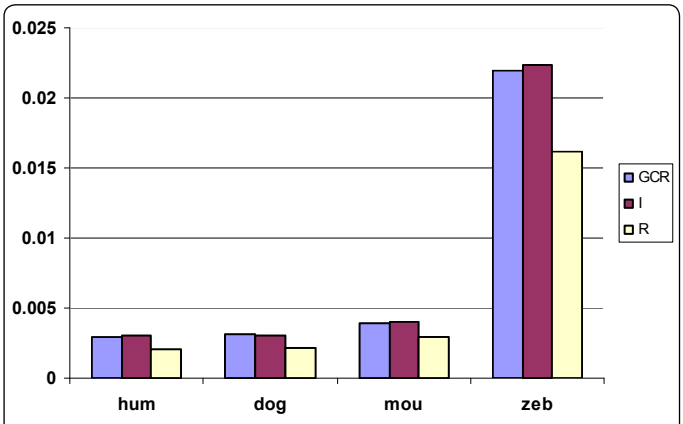

Figure 4b: The information indices of genomic sequences of *DNMT1* gene in different species.

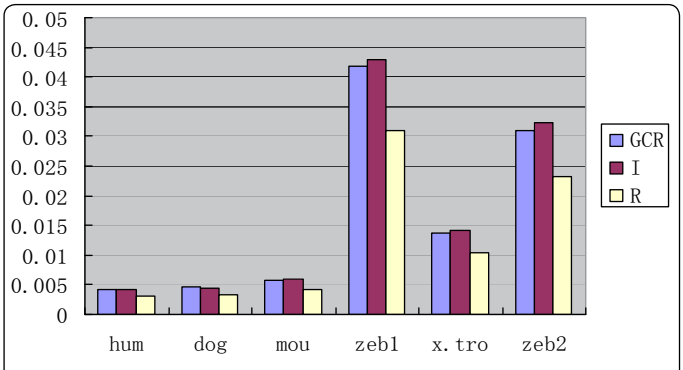

Figure 4c: The information indices of intron sequences of *DNMT1* gene in different species.

several measures related to information entropy of gene *DNMT1* in different species in order to quantitatively compare complexity and information of the mRNA sequences. The information indices include Shannon entropy (H_1), repeatability complexity (GC_R), information of ordering (I) and redundancy (R). The results were shown in the Table 5 and we found that GC_R, I and R of human mRNA sequences are the highest compared to those of other species. In terms of Shannon entropy, the zebrafish's information is the highest in all studied species. As shown in Figure 4a, GC_R, I and R gradually decrease from higher organisms to lower ones. This indicates that the higher organism's mRNA sequence is more complex, regular, and stores more potential information. Therefore, we thought that GC_R, I and R reflect the complicate processing of evolution and development.

Sequentially, the analysis among these species is extended to ascertain the complexity and information in intron region and the whole gene sequences. Because the whole DNA sequences of *DNMT1* gene are not always available in chimpanzee and cattle, we analyzed entropy information only in four species: human, dog, mouse and Zebrafish. The GC_R, I and R on both intron regions and whole genomic DNA are listed in Table 6 and 7. We found that the Shannon entropy information of DNA sequences in intron region and whole genomic sequence of *DNMT1* is alike in three mammals, which indicates the similar DNA nucleotide structure and nucleotide compositions of them. The zebrafish has the lowest entropy information in mRNA sequence and the highest GC_R, I and R in intron region and whole genomic sequence (Figure 4a, 4b and 4c). It was discovered that the tendency of GC_R, I and R of whole DNA sequence and intron is similar in three mammals, which are obviously reverse to that of mRNA. The results, as shown in mRNA sequences, human's information contents are the highest, whereas its value is lowest in intron and whole DNA, suggest that mRNA sequence would play an important role for species evolution. The results confirmed that the entropy information in *DNMT1* gene is DNA nucleotide composition dependent. In term of entropy information of the gene, there are significant correlations and similarities among mammalians followed by relatively bigger divergence between mammalians and veberate animal (Figure 4b and 4c).

Information analysis of protein and domain coded by *DNMT1* gene in different species

The information contents of entropy have demonstrated difference in mRNA sequence of the gene. We postulated that the complexity information measures may result in changes in protein sequences and functions. Because proteins largely constitute the machinery that makes life, they carry out all structural, catalytic, and regulatory functions. Protein sequences also formulate the basis of other structures of protein. Therefore, it is necessary to analyze entropy information of the protein sequences and different domains of the *DNMT1* gene among these species and check the impacts of entropy information on protein domains. To easily compare the entropy information of protein sequences and compositions, we used a relative repeatability complexity (GC_R%), which avoids the different length of protein sequences. The results of different entropy information were shown in Table 8 and it is obvious that the caliber of the relative repeatability complexity is not large, its minimum and maximum are respectively 0.0034, 0.0038. The GC_R% of the *DNMT1* gene in human still is the highest among these species, and is the same for the information analysis of mRNA sequence of the gene. Moreover, information indices R and I of most of mammals, except for mouse, are higher than that of zebrafish (Table 8).

With the aid of computer prediction, we found that there are seven domains, containing four kinds of domain types on the coding region of *DNMT1* gene. The domain types include one DAMP domain, one Cxxc zinc finger domain, and two BAH domains as well as three DNA methylase domains. Information analysis were performed separately

Species	H_1 (bit)	GC_R(%)	I(bit)	R(%)
Human	2.8835	0.0038	0.1123	0.0375
Chimpanzee	2.8809	0.0038	0.1148	0.0383
Cattle	2.8838	0.0037	0.1119	0.0374
Dog	2.8800	0.0038	0.1157	0.0386
Mouse	2.8944	0.0034	0.1013	0.0338
Zebrafish	2.8920	0.0034	0.1037	0.0346

Table 8: Entropy information of protein sequence of *DNMT1* gene.

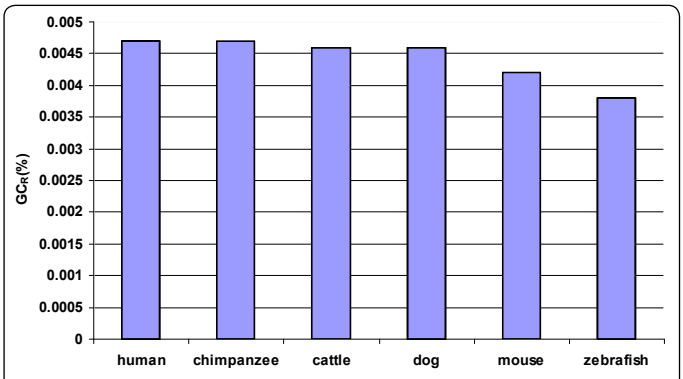

Figure 5: The relative repeatability complexity (GCR%) of nondomain region of *DNMT1* gene in different species.

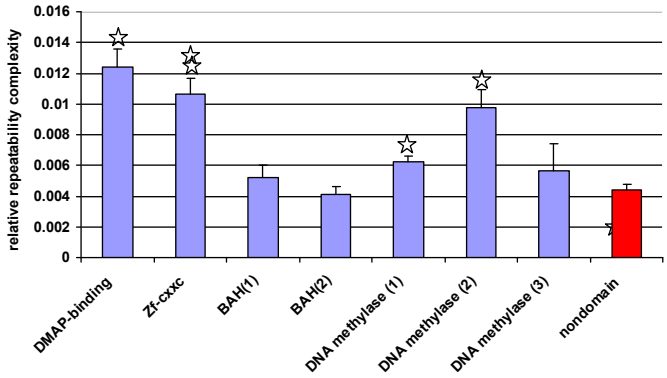

Figure 6a: The relative repeatability complexity (GCR%) of domains of *DNMT1* gene.

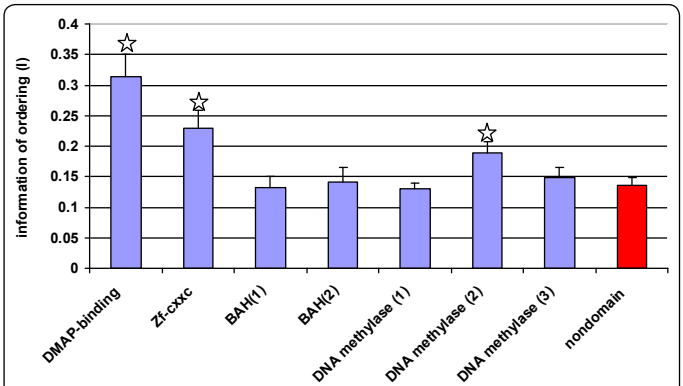

Figure 6b: The information of ordering (I) of domains of *DNMT1* gene.

for the seven domains and non-domain regions, and discovered that the GC_R%, I and R on most of domains and non-domain regions are larger than whole protein sequences. In the species, we found that the GC_R% of non-domain region is related to the evolution distance of the organisms, i.e., the GC_R% gradually decreases from higher organism to lower one (Figure 5). The results suggest that non-domain region may be important regions for biological processes. They imply that there exists some relationships between non-domain regions and domains, and the interactions may lead to a more complicate protein functions. Furthermore, information indices of DMAP binding, CXXC zinc finger domains and the second DNA methylase are not only larger than other domains, but also significantly higher than non-domain region as shown in Figure 6a, 6b and 6c. Interestingly, it was found that same type domains have unequal information (data not shown). Therefore we thought that domain regions stored more information than non-domain regions, which further indicates that protein function is mainly determined by domain regions.

The information analysis of DNA methylation for *DNMT1* gene and aging

In our on-going research, the relationship between methylation

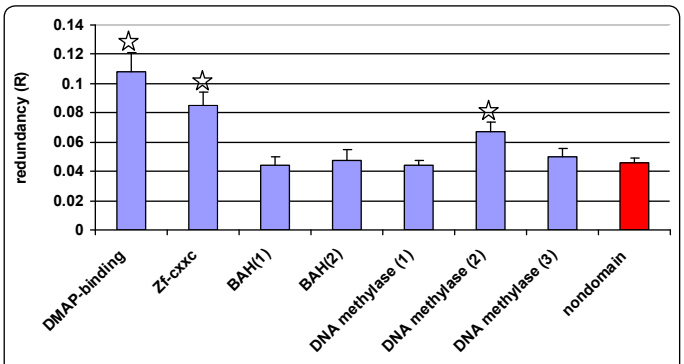

Figure 6c: The redundancy (R) of domains of *DNMT1* gene.

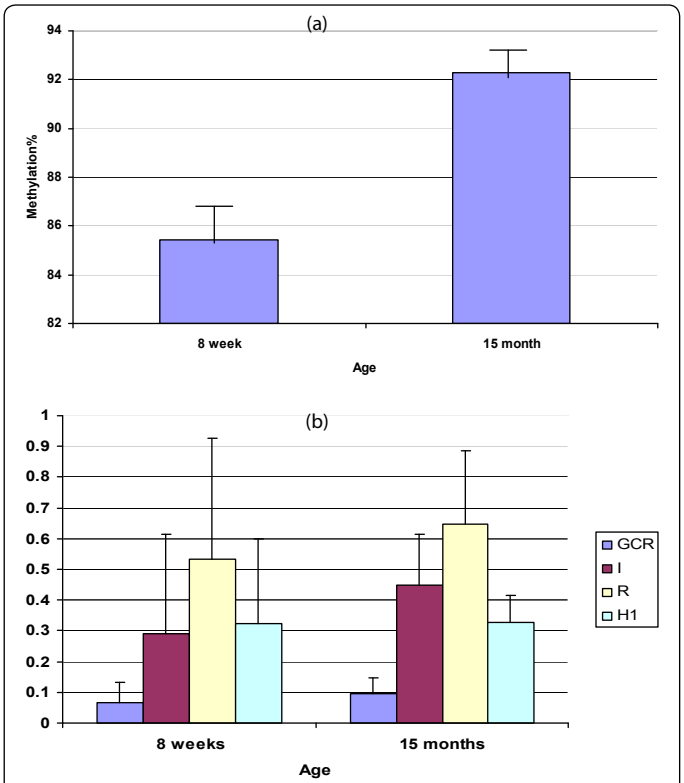

Figure 7: (a) The methylation status of chicken *DNMT1* gene at two ages; **(b)** The complexity analysis of methylation status at two ages.

status of *DNMT1* gene and aging in chicken was discovered in a unique chicken model. This was a quantitative measure of DNA methylation levels of four CpG sites in exon1 region of chicken *DNMT1* gene for 8 weeks of age and 15 months of age (Figure 7a). The methylation level of the CpG sites for 8 weeks of age was lower than that of 15 months of age in liver ($P<0.05$). The results suggested that the methylation status of *DNMT1* is related to aging in chickens. In further information analysis, we treated the methylation percentage as a probability of methyaltion happening at a given CpG site. The probabilities at different CpG sites could then be used to calculate the entropy information and other information contents, thus we set up the relationship between DNA methylation profile and entropy. The results were shown in Figure 7b, except for H_1 Shannon entropy, the GC_R, I and R of DNA methylation profiles in *DNMT1* gene are similar to the methylation levels, i.e., the complexity and entropy of methylation status of the gene are higher at 15 months of age than at 8 weeks of age.

Discussion

The complexity search for DNA regions with different information views is one of the pivotal tasks of structural sequence analysis in post-genomics era. Entropy being an information measure includes several types such as Shannon entropy, spectral entropy and conformational entropy. Shannon entropy was applied to measure global splicing disorders [27], analyzed conserved protein sequences of influenza A viruses and identify vaccine targets [28]. Spectral entropy was used to measure order and correlations in genomic DNA sequences and analyze levels of ordering in coding and noncoding regions of DNA sequences [29,30]. The relationship between spectral entropy and GC content analyses was set up in the β-esterase gene cluster [31,32]. However, the most of researches mentioned above only involved the entropy analysis in DNA sequences. In this study, to overcome the variations of protein sequence length and the types of amino acids, we successfully applied an information measure named as relative repeatability complexity, GC_R (%) = $(H_1 - H_2)/H_0^*100$ in protein sequence. Compared to the information of ordering (I), complexity (GC_R) and redundancy (R), the relative repeatability complexity supplies another relative measure of repeatability complexity. Like Shannon entropy, we believe that the GC_R% could be widely used in complexity analysis of protein sequence and domain identification.

The research revealed the relationship between DNA nucleotide composition of *DNMT1* gene and complexity information. With different approaches, the GC and AT contents as well as different entropy information were explored in various patterns of the *DNMT1* gene from different species. Because of DNA nucleotide composition differences, it was found that the entropy in *DNMT1* gene among species is DNA base composition dependent, which is corresponding to the differences of *DNMT1* gene in AT content and GC content between mammalian animals and zebrafish. We further demonstrated that the complexity of introns of the *DNMT1* gene in mammals is lower than that of coding regions. Most importantly, although the the AT and GC contents among chimpanzee, cattle, mouse and human have the same distribution based on K-S test ($P>0.05$) and they are significant different from zebrafish and dog ($P<0.05$), there are obvious similarities between mammalians in term of the entropy information indices for *DNMT1* gene [31]. All of these results indicate that the complexity information of the *DNMT1* gene may be preconditioned by strong inequality in nucleotide content (based composition) in different species, also by tandem, dispersed repeats or palindrome-hairpin structures, as well as by a combination of all these factors.

We did not only demonstrate the correlations of complexity information between mRNA composition and protein sequence, but we further revealed the impacts of entropy information on domains and non-domain of *DNMT1* gene in different species. The entropy change of protein as a whole is interpreted as a diversification of protein coded by *DNMT1* gene, and is also reflected in complexities of DNA sequence of these organisms. These results indicated that there appears to be a general evolutionary tendency: the similarity of mammalian organisms and diversification with others. The claim is meaningful in biological functions because many other evidences from experiments and protein sequence alignments have confirmed that protein sequence complexity and entropy information have a strong linear correlation to some phenomena such as packing density and hydrophobicity [33], the structural and functional characteristics in G-protein-coupled receptors [34]. Amazingly, DNA binding site, a most important biological interaction between DNA and protein, is a completely entropy-driven process. Some evidence showed that a positive entropy may be due to release of dehydration upon forming the protein/DNA complex [35]. Considering and combining our previous results, to further study the function of *DNMT1* gene on epigenetics, we may redesign its structures via configurational entropy. Therefore, future studies should include identification of functional domains and domain boundaries from sequence alone with entropy information among different organisms, and deeply ascertaining the mechanisms of entropy-driven DNA binding on protein/DNA complex of *DNMT1* gene with aids of bioinformatics methods [36-38].

In our research, the methylation status of *DNMT1* with aging-specific in a unique chick model also implied some aging-driven entropy characteristics. Many evidences have shown that aberrant DNA methylation and histone acetylation have been linked to a number of aging related disorders including cancer, autoimmune disorders and others [39-41]. With rich knowledge supporting aging processes is characterized by entropy changes, we think that an increasing entropy may lead to the loss of molecular fidelity and slow accumulation of overwhelming maintenance system [42,43]. The relationship between methylation status and entropy levels of the *DNMT1* gene, function on maintaining methylation fidelity, shows a huge increase in methylation with aging in the unique chicken model. Nevertheless, the changeable entropy information suggested that aging is a process of energy dispersion. But entropy is to disperse the concentrated energy, resulting in a biologically inactive or malfunctioning. Considering other factors, we propose a hypothesis that aging process is due in part to DNA methylation, DNA damage, mutations and chemical bonds loss, etc., which may cause entropy changes and molecular fidelity loss. To get answer about aging process from DNA methylation changes over ages, the *DNMT1* gene could be a candidate target to pursue in the future research [44].

In summary, information measure of *DNMT1* gene, one of the most important epigenetic genes, is relevant to its genomic complexity, which thereby associates to evolution and aging processing. The intrinsic mechanism is not to be studied yet. In post-genomic era, many unknown genes can be clustered, and analyzed in domain regions and non-domain regions based on complexity of DNA and protein sequences. By applying these entropy information methods to various functional genomic regions, we will have deeper insights on gene functions and genome annotations.

Acknowledgements

The authors are grateful to the foundation for international exchange program of excellent scholar of China and This project was supported by Agriculture and

Food Research Initiative Competitive Grants no. 2010-65205-20588 and 2008-35204-04660 from the USDA National Institute of Food and Agriculture.

References

1. Schermelleh L, Haemmer A, Spada F, Rosing N, Meilinger D, et al. (2007) Dynamics of Dnmt1 interaction with the replication machinery and its role in postreplicative maintenance of DNA methylation. Nucleic Acids Res 35: 4301-4312.
2. Nizami I (2008) Does Norwich's Entropy Theory of Perception avoid the use of mechanisms, as required of an information-theoretic model of auditory primary-afferent firing. J Acoust Soc Am 123: 3857.
3. Rhee I, Bachman KE, Park BH, Jair KW, Yen RW, et al (2002) DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. Nature 416: 552-556.
4. Bachman KE, Rountree MR, Baylin SB (2001) Dnmt3a and Dnmt3b are transcriptional repressors that exhibit unique localization properties to heterochromatin. J Biol Chem 276: 32282-32287.
5. Jones PA, Baylin SB (2007) The epigenomics of cancer. Cell 128: 683-692.
6. Rodin SN, Parkhomchuk DV, Riggs AD (2005) Epigenetic changes and repositioning determine the evolutionary fate of duplicated genes. Biochemistry (Mosc) 70: 559-567.
7. Yokomine T, Hata K, Tsudzuki M, Sasaki H (2006) Evolution of the vertebrate DNMT3 gene family: a possible link between existence of DNMT3L and genomic imprinting. Cytogenet Genome Res 113: 75-80.
8. Yu Y, Zhang H, Tian F, Zhang W, Fang H, et al. (2008) An integrated epigenetic and genetic analysis of DNA methyltransferase genes (DNMTs) in tumor resistant and susceptible chicken lines. PLoS One 3: e2672.
9. Yu Y, Zhang H, Tian F, Bacon L, Zhang Y, et al. (2008) Quantitative evaluation of DNA methylation patterns for ALVE and TVB genes in a neoplastic disease susceptible and resistant chicken model. PLoS One 3: e1731.
10. Shannon C (1948) A mathematical theory of communication. 27: 379-423.
11. Wootton JC, Federhen S (1996) Analysis of compositionally biased regions in sequence databases. Methods Enzymol 266: 554-571.
12. Hang Y, Zhi Q, Yang H, Yingxian S (2007) the comparison between four kinds of vertebrates on the complexities of their DNA sequences. Chinese Journal of Medical physics 24: 110-113.
13. Orlov YL, Filippov V, Potapov V, Kolchanov N (2002) Construction of stochastic context trees for genetic texts. In Silico Biol 2: 233-247.
14. Milosavljevic A, Jurka J (1993) Discovering simple DNA sequences by the algorithmic significance method. Comput Appl Biosci 9: 407-411.
15. Gonzalez-Diaz H, Perez-Montoto LG, Duardo-Sanchez A, Paniagua E, Vazquez-Prieto S, et al. (2009) Generalized lattice graphs for 2D-visualization of biological information. J Theor Biol 26: 136-147.
16. Perez-Bello A, Munteanu CR, Ubeira FM, De Magalhaes AL, Uriarte E, et al. (2009) Alignment-free prediction of mycobacterial DNA promoters based on pseudo-folding lattice network or star-graph topological indices. J Theor Biol 256: 458-466.
17. Jimenez-montano MA, Ebeling W, Pohl T, Rapp PE (2002) Entropy and complexity of finite sequence as fluctuating quantities. Biosystems 64: 23-32.
18. Munteanu CD, J, Pazos Sierra, A. Prado-Prado, F. Pérez-Montoto, LG. Vilar, et al. (2009) Markov Entropy Centrality: Chemical, Biological, Crime and Legislative Networks. Information theory of Complex Networks.
19. Jamasebi R, Redline S, Patel SR, Loparo KA (2008) Entropy-based measures of EEG arousals as biomarkers for sleep dynamics: applications to hypertension. Sleep 31: 935-943.
20. Humeau A, Trzepizur W, Rousseau D, Chapeau-Blondeau F, et al. (2008) Fisher information and Shannon entropy for on-line detection of transient signal high-values in laser Doppler flowmetry signals of healthy subjects. Phys Med Biol 53: 5061-5076.
21. Norris PR, Stein PK, Morris JA Jr (2008) Reduced heart rate multiscale entropy predicts death in critical illness: a study of physiologic complexity in 285 trauma patients. J Crit Care 23: 399-405.
22. Mehl J, Speck T, Seifert U (2008) Large deviation function for entropy production in driven one-dimensional systems. Phys Rev E Stat Nonlin Soft Matter Phys 78: 011123.
23. Adami C, Ofria C, Collier TC (2000) Evolution of biological complexity. Proc Natl Acad Sci U S A 97: 4463-4468.
24. Song J, Ware A, Liu SL, Surette M (2004) Comparative Genomics via Wavelet Analysisfor Closely Related Bacteria. EURASIP J Adv Signal Process 1: 5-12.
25. Song J, Ware A, Liu SL (2003) Wavelet to predict bacterial ori and ter: a tendency towards a physical balance. BMC Genomics 4:17.
26. Drummond A, Rodrigo AG (2000) Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. Mol Biol Evol 17:1807-1815.
27. Ritchie W, Granjeaud S, Puthier D, Gautheret D (2008) Entropy measures quantify global splicing disorders in cancer. PLoS Comput Biol 4:e1000011.
28. Heiny AT, Miotto O, Srinivasan KN, Khan AM, Zhang GL, et al. (2007) Evolutionarily conserved protein sequences of influenza a viruses, avian and human, as vaccine targets. PLoS ONE 2: e1190.
29. Chechetkin VR, Turygin AY (1996) Study of correlations in DNA sequences. J Theor Biol 178:205-217.
30. Chechetkin VR, Lobzin VV (1996) Levels of ordering i coding and noncoding regions of DNA sequence. physics letter 222:354-360.
31. Balakirev ES, Chechetkin VR, Lobzin VV, Ayala FJ (2005) Entropy and GC Content in the beta-esterase gene cluster of the Drosophila melanogaster subgroup. Mol Biol Evol 22:2063-2072.
32. Balakirev ES, Chechetkin VR, Lobzin VV, Ayala FJ (2003) DNA polymorphism in the beta-Esterase gene cluster of Drosophila melanogaster. Genetics 164:533-544.
33. Liao H, Yeh W, Chiang D, Jernigan RL, Lustig B (2005) Protein sequence entropy is closely related to packing density and hydrophobicity. Protein Eng Des Sel 18:59-64.
34. Imai T, Fujita N (2004) Statistical sequence analyses of G-protein-coupled receptors: structural and functional characteristics viewed with periodicities of entropy, hydrophobicity, and volume. Proteins 56: 650-660.
35. Dragan AI, Klass J, Read C, Churchill ME, Crane-Robinson C, et al. (2003) DNA binding of a non-sequence-specific HMG-D protein is entropy driven with a substantial non-electrostatic contribution. J Mol Biol 331: 795-813.
36. Galzitskaya OV, Melnik BS (2003) Prediction of protein domain boundaries from sequence alone. Protein Sci 12: 696-701.
37. Micaelo NM, Victor BL, Soares CM (2008) Protein thermal stabilization by charged compatible solutes: Computational studies in rubredoxin from Desulfovibrio gigas. Proteins 72: 580-588.
38. Bae E, Bannen RM, Phillips GN (2008) Bioinformatic method for protein thermal stabilization by structural entropy optimization. Proc Natl Acad Sci U S A 105: 9594-9597.
39. Wagner EJ, Baines A, Albrecht T, Brazas RM, Garcia-Blanco MA (2004) Imaging alternative splicing in living cells. Methods Mol Biol 257: 29-46.
40. Jang H, Mason JB, Choi SW (2005) Genetic and epigenetic interactions between folate and aging in carcinogenesis. J Nutr 135: 2967S-2971S.
41. Budovsky A, Muradian KK, Fraifeld VE (2006) From disease-oriented to aging/longevity-oriented studies. Rejuvenation Res 9: 207-210.
42. Hayflick L (2007) Entropy explains aging, genetic determinism explains longevity, and undefined terminology explains misunderstanding both. PLoS Genet 3: e220.
43. Hayflick L (2004) "Anti-aging" is an oxymoron. J Gerontol A Biol Sci Med Sci 59: B573-578.
44. Li G, Weyand CM, Goronzy JJ (2008) Epigenetic mechanisms of age-dependent KIR2DL4 expression in T cells. J Leukoc Biol 84: 824-834.