An Application of Information Theoretical Measures for DNA Structure

DNA Yapısı İçin Teorik Bilgi Ölçülerinin Uygulaması

Özlem EGE ORUÇ[a]
Ömer DURSUN[a]

[a]Department of Statistics,
Dokuz Eylul University
Faculty of Arts & Sciences
İzmir

Geliş Tarihi/*Received:* 25.06.2010
Kabul Tarihi/*Accepted:* 07.10.2010

Yazışma Adresi/*Correspondence:*
Özlem EGE ORUÇ
Dokuz Eylul University
Faculty of Arts & Sciences,
Department of Statistics, İzmir,
TÜRKİYE/TURKEY
ozlem.ege@deu.edu.tr

ABSTRACT Objective: This study makes use of the application of entropy in information theory in the DNA structure. We applied the information theoretic notion of entropy and Kullback-Leibler distance to characterize the distribution of exons, introns and amino acid in DNA sequences. **Material and Methods:** Entropy, relative entropy and mutual information values for exons, introns and amino acid are computed based on their probability distribution. An application to perform a comparison exon / intron and amino acid sequences in some human genome is briefly discussed. **Results:** The two major divisions of an Eukaryote's DNA that are transcribed into RNA are protein-coding sections called exons, and nonprotein-coding sections called introns. It makes good intuitive sense that introns and exons probability distributions may be different, since they are subject to different random process. Although, introns are usually more highly diversified than neighboring exons, we obtain that the probability distribution of exons and introns are similar. In analyzing the splice site regions of exons and introns, it is observed that the probability distributions of the bases are very different than the probability distributions of all the bases of exons and introns. **Conclusion:** The entropy of a number of DNA coding and non-coding sequences collected from different genomes was calculated for finite sequences. A distance measure was applied to compare exon and intron information content and it was found that they are structurally quite similar. Although, introns are usually more highly diversified than neighboring exons, we obtain that the probability distribution of exons and introns are similar. This unforeseen result exhibits that introns can carry nearly as much information as exons. In analyzing the splice site regions of exons and introns, it is observed that the probability distributions of the bases are very different than the probability distributions of all the bases of exons and introns. We may claim that one may obtain information on the splice site region of the genes by examining the probability distributions of the last bases of exons before the GT pair and the first bases of introns after the GT pair.

Key Words: Entropy; probability theory; information theory; sequence analysis DNA

ÖZET Amaç: Bu çalışmada, bilgi teorisindeki entropi kavramı kullanılarak DNA dizilim yapısı için uygulama yapılmıştır. Bilgi teorisindeki entropi ve Kullback-Leibler uzaklığı ile DNA dizilimindeki ekzonların, intronların ve amino asitlerin olasılık dağılımları karakterize edilmeye çalışılmıştır. **Gereç ve Yöntemler:** Ekzon, intron ve amino asitlerin bazları temel alınarak oluşturulan olasılık dağılımlarından, entropi, göreli entropi ve ortak bilgi değerleri hesaplanmıştır. Ekzonların, intronların ve amino asitlerin olasılık dağılımlarını karşılaştırmak için insan genomuna ait bazı genler ile uygulama yapılmış ve sonuçlar yorumlanmıştır. **Bulgular:** Ökaryot hücrelerin DNA dizi yapısı iki ana bölümden oluşur. Bunlar, protein kodlayan bölüm (ekzon) ve kodlamayan bölüm (intron) olarak adlandırılır. Bu iki bölüm farklı görevler yaptıkları için sezgisel olarak olasılık dağılımlarının da farklı olacağı düşünülmektedir. İntronlar, komşuları ekzonlardan çok farklı olmalarına rağmen yapılan çalışma sonucunda olasılık dağılımlarının benzer olduğu gösterilmiştir. İntron ve ekzonların ayrılma bölgesinde bazların olasılık dağılımları incelendiğinde ekzonların baz dizilimlerinin ve intronların baz dizilimlerinin olasılık dağılımlarının bu bölgelerde farklı olduğu gösterilmiştir. **Sonuç:** Farklı genlerden toplanan DNA'nın protein kodlayan ve kodlamayan bölümlerinin sonlu dizileri için entropi değerleri hesaplanmıştır. Ekzon ve intronların bilgi içeriklerini karşılaştırmak için uzaklık ölçüleri hesaplanmış ve iki yapının da benzer olduğu görülmüştür. Bu beklenmedik sonuç, intronların da ekzonlar kadar bilgi taşıyabileceklerinin bir göstergesi olabilir. İntron ve ekzonların ayrılma bölgesindeki bazların olasılık dağılımları incelendiğinde ekzonların baz dizilimlerinin ve intronların baz dizilimlerinin olasılık dağılımlarının bu bölgelerde farklı olduğu gösterilmiştir. Buna göre, genlerdeki ekzon ve intronların ayrılma bölgesinin neresi olduğuna dair bilgi sahibi olmak için ekzonların GT baz çiftinden önceki son bazlarının, intronların da GT baz çiftinden sonraki ilk bazlarının dizilimlerinin olasılık dağlımını incelemenin uygun olabileceği ileri sürülmüştür.

Anahtar Kelimeler: Entropi; olasılık teorisi; bilgi teorisi; dizi analizi DNA

Turkiye Klinikleri J Biostat 2011;3(1):1-7

Living systems are not only characterized by metabolism and reproduction, but also by flow of information. On the molecular level, information is carried by the sequences of DNA and proteins.[1] Entropy, the main concept of Shannon's theory of information and communication, is often play transparent role when applied to statistical ensembles of symbolic sequences. There have been a many studies on the amount of information per base in DNA sequences. Entropies of DNA sequences have been discussed by various authors Mantegna et al. ,Schmitt and Herzel, and Herzel et al., and Chun and Wang.[2-4] The two major divisions of an Eukaryote's DNA that are transcribed into RNA are protein-coding sections called exons, and nonprotein-coding sections called introns. Investigating these major divisions has attracted researchers' attention. Applications of information theory for DNA structure in finding ways of information transmitted in living organisms have proved useful. It makes good intuitive sense that introns and exons probability distributions may be different, since they are subject to different random process. In this paper we apply information theoretic concept of entropy and Kullback-Leibler distance to characterize the distribution of exons, introns and amino acid in DNA sequences. Although, introns are usually more highly diversified than neighboring exons, we obtain that the probability distribution of exons and introns are similar. This unforeseen result exhibits that introns can carry nearly as much information as exons. The effects of the similarity of the probability distributions of the amino acids in the genes and the base sequence on the mutual information are also examined.

FUNDAMENTALS OF DNA

DNA is the genetic material in organisms and it stores the instruction needed by the cell to perform daily life function. Central Dogma (Figure 1) is the term that tells us how we get the protein from the DNA. This process is also called gene expression. The expression of gene consists of two steps: *Transcription* and *Translation* see.[5]

The two major divisions of an Eukaryote's DNA that are transcribed into RNA are protein-coding sections called exons, and nonprotein-coding sections called introns. Intron is a segment of gene situated between exons. It is not responsible for the coding of protein. So the introns will be ultimately spliced out of the mRNA. An exon is a nucleotide sequence in DNA that carries the code for the final mRNA molecule and thus defines the amino acid sequence during protein synthesis. The process of removing the introns for the mRNA sequence is called RNA splicing. Intron in eukaryotic genes generally satisfies the GT-AG rule. That is, an intron begins with GT and ends with AG.[6] In living cells, the set of rules that provide the link between the sequence of the DNA with the bases in the RNA molecule as its counterpart and the proteins to be synthesized is called the genetic code. Apart from some exceptions, all living objects use the same genetic code called Standard Genetic Code. The amino acids produced via these codes are presented in Table 1.[5]

FIGURE 1: Central Dogma in a Cell.

TABLE 1: The Genetic Code.

	U	C	A	G	
U	Phe (UUU)	Ser (UCU)	Tyr (UAU)	Cys (UGU)	U
	Phe (UUC)	Ser (UCC)	Tyr (UAC)	Cys (UGC)	C
	Leu (UUA)	Ser (UCA)	**Stop(UAA)**	**Stop(UGA)**	A
	Leu (UUG)	Ser (UCG)	**Stop(UAG)**	Trp (AGG)	G
C	Leu (CUU)	Pro (CCU)	His (CAU)	Arg (CGU)	U
	Leu (CUC)	Pro (CCC)	His (CAC)	Arg (CGC)	C
	Leu (CUA)	Pro (CCA)	Gln (CAA)	Arg (CGA)	A
	Leu (CUG)	Pro (CCG)	Gln (CAG)	Arg (CGG)	G
A	Ile (AUU)	Thr (ACU)	Asn (AAU)	Ser (AGU)	U
	Ile (AUC)	Thr (ACC)	Asn (AAC)	Ser (AGC)	C
	Ile (AUA)	Thr (ACA)	Lys (AAA)	Arg (AGA)	A
	Met (AUG)	Thr (ACG)	Lys (AAG)	Arg (AGG)	G
G	Val (GAU)	Ala (GCU)	Asp (GAU)	Gly (GGU)	U
	Val (GAC)	Ala (GCC)	Asp (GAC)	Gly (GGC)	C
	Val (GAA)	Ala (GCA)	Glu (GAA)	Gly (GGA)	A
	Val (GAG)	Ala (GCG)	Glu (GAG)	Gly (GGG)	G

CONCEPTS OF ENTROPY

Shannon entropy is a quantitative measure of uncertainty in a data set. This section briefly defines Shannon entropy, relative entropy (Kullback-Leibler divergence), joint entropy and mutual information. Let X be a discrete random variable, taking a finite number of possible values $x_1, x_2,...x_n$ with respective probabilities $p_i \geq 0$ for i=1, 2,...,n and $\sum_{i=1}^{n} p_i = 1$. The Shannon entropy H(X) is defined by

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \,. \qquad (1)$$

in the works Cover and Thomas, and Shannon.[7,8] The joint entropy measures how much entropy is contained in a joint system of two random variables. If the random variables are *X* and *Y*, the joint entropy *H(X;Y)* given in Cover and Thomas is [7]

$$H(X,Y) = -\sum_{x=i}\sum_{y=j} p_{ij}(x,y)\log[p_{ij}(x,y)] \,. \qquad (2)$$

The mutual information of two random variables is a quantity that measures the mutual dependence of the two variables. The interpretation is that when mutual information is absent, marginal distributions are independent and their entropies add up to the total entropy. When mutual information is positive, marginal distributions are dependent as some combinations occur relatively more often than other combinations do, and marginal entropies exceed total entropy by an amount equal to the mutual information. Mutual information I is evaluated by the formula[7]

$$I(X,Y) = H(X) + H(Y)\ \ H(X,Y) \,. \qquad (3)$$

The Kullback-Leibler divergence (KL=$D(p\|q)$) is a non-commutative measure of the difference between two probability distributions p and q. KL is also sometimes called the information gain about X achieved if p can be used instead of q. It is also called the relative entropy, for using q instead of p. The relative entropy is an appropriate measure of the similarity of the underlying distribution. It may be calculated from

$$D(p\|q) = \sum_{x\in X} p(x)\log\frac{p(x)}{q(x)} \,. \qquad (4)$$

The properties of the relative entropy equation make it non-negative and it is zero if both distributions are equivalent p=q. The smaller the relative entropy is the more similar the distribution of the two variables and vice versa.[9]

MATERIAL AND METHODS

The data sets pertaining to the human genome genes used in this paper are obtained via the web sites of the NCBI (National Center for Biotechnology Information) and BDGP (Berkeley Drosophila Genome Project). The human genes HUMGALK1A, HUMCD19A and HSALADG are used as data sets. The Human Galactokinase (HUMGALK1A) gene extracted from the 17th chromosome of the human genome comprises of 8 exons and 7 introns. The Human CD19 (HUMCD19A) gene extracted from the 16th chromosome is composed of 14 exons and 13 introns. As for the Homosapiens ALAD (HSALADG) gene, it has 14 exons and 13 introns. This study consists of two separate applications concerning the ways of using information theory in investigating the DNA structure. The purpose of the first application is to show that the probability distributions of the exons and the introns of genes are the same and to emphasize that introns also include information as exons do, and to conduct an analysis for the splice site regions of exons and introns, considering that exons always begin with a GT base pair. The second application aims at providing an example for information theory on the amino acid sequences in the genes. The Kullback-Leibler distance is used in order to measure the similarity among the probability distributions that are obtained using the bases in the exons, introns and the amino acids of the genes. Moreover, various entropy values are computed from the probability distributions. Also, the distance of the probability distributions of the base sequences of the amino acids in the exons and introns to the uniform distribution, where each base has the same chance to be seen and the entropy is at the maximum value, is examined by calculating the positional relative entropy values. Besides, some interpretations on the randomness in the sequences are made with respect to the distances.

TABLE 2: Distance from the Uniform Distribution of Exons and Introns Distribution.

	EXON&INTRON vs UNIFORM	
GENES		**K-L DIVERGENCE**
HUMGALK1A	EXON	0,068914955
	INTRON	0,017947962
HUMCD19A	EXON	0,035600501
	INTRON	0,029262792
HSALADG	EXON	0,025785072
	INTRON	0,004980854

TABLE 3: Distance between the Distribution of Exons and Distribution of Introns Distribution.

	EXON vs INTRON
GENES	**K-L DIVERGENCE**
HUMGALK1A	0,036551825
HUMCD19A	0,037031963
HSALADG	0,012455115

APPLICATIONS

In the first application, the distances between the uniform distribution and the probability distributions obtained from the bases in the exons and introns of the three genes under inspection are examined separately. According to the results presented in Table 2, the probability distributions of the exons and introns in each gene are proximate to the uniform distribution. The similarity to the uniform distribution shows that the bases in the exons and introns occur in equal probability which indicates that the sequences are random.

The similarities between the probability distributions of the exons and introns, grounded on the bases, are examined via the calculation of the Kullback-Leibler distance values. The values obtained are presented in Table 3. These values are close to zero in each of the genes. This closeness shows that the probability distributions of exons and introns are similar to each other.

In the last part of the first application, an analysis is carried out for the splice site regions of exons and introns, considering that introns always begin with a GT base pair. It is known that introns always begin with a GT base pair and end with an AG base pair. We observed this condition in the splice site regions of exons and introns. When the protein sequence is examined, each GT base pair observed is not always an intron beginning. Similarly, each AG base pair observed is not always an intron ending. In various studies on determining how introns begin, it is seen that the last bases of the exon before the GT pair and the first bases of the intron after, is important. As we examine the probability distributions of the sequences of the nine last bases of the exon before the GT pair and the sequences of the first nine bases of the intron after the GT pair in the genes in question, it is observed that exons ending before introns beginning with a GT pair most probably end with a Guanine base. Using Kullback-Leibler distance scale one may deduce that the probability distributions of the splice site region bases of exons and introns are different from the probability distributions of the bases in exons and introns in the whole sequence. Table 4 demonstrates this result. Information on splice site region can be obtained by analyzing the probability distributions of the last nine bases of exons before the GT pair and of the first nine bases of introns after the GT pair.

For the second application, we acquired the probability distributions of the amino acids belonging to the three genes under examination. The results obtained are presented in Table 5. The entropy values calculated using these probability distributions are 4,119339bits for HSALADG, 4,071260bits for HUMCD19A and 3,992027bits for

TABLE 4: Distance from the splice site region bases distributions of exons to all exons distributions.

Genes		**K-L Divergence**
HUMGALK1A	EXON	0,041399401
	INTRON	0,160774519
HUMCD19A	EXON	0,131393025
	INTRON	0,133212327
HSALADG	EXON	0,084580981
	INTRON	0,127826935

TABLE 5: Probability Distributions of amino acid in HSALASG, HUMCD19A, HUMGALK1A genes.

Amino Acid	HSALADG	HUMCD19A	HUMGALK1A
Phenylalanine (F)	0,03371	0,02551	0,02657
Leucine (L)	0,09831	0,10714	0,11353
Isoleucine (I)	0,03371	0,01361	0,02174
Methionine (M)	0,03090	0,02381	0,02174
Valine (V)	0,07022	0,04252	0,07488
Serine (S)	0,05618	0,09184	0,07488
Proline (P)	0,06461	0,09864	0,05556
Threonine (T)	0,06461	0,06803	0,06522
Alanine (A)	0,11236	0,04592	0,11594
Tyrosine (Y)	0,03371	0,02381	0,02657
STOP	0,00281	0,00170	0,00242
Histidine (H)	0,02528	0,01531	0,02174
Glutamine (Q)	0,02809	0,03741	0,04831
Asparagine (N)	0,01404	0,03061	0,01208
Lysine (K)	0,03371	0,02891	0,01691
Aspartic Acid (D)	0,05056	0,05272	0,03140
Glutamic Acid(E)	0,08708	0,09524	0,09420
Cysteine (C)	0,02247	0,01361	0,01932
Tryptophan (W)	0,00843	0,02891	0,00242
Arginine (R)	0,06461	0,05442	0,07488
Glycine (G)	0,06461	0,10034	0,07971

HUMGALK1A. Because the HUMGALK1A gene has the smallest entropy value according to the values found, it can be said that the estimation of the amino acids of this gene is easier.

The entropy value decreases when new information is added. We checked this condition for the genes we analyzed in our application. The results obtained are presented in Table 6. The entropy value that calculated when we do not know the first base of the amino acids can be seen in the Genes row of the table. The values when the first base is known are shown in the other rows. Any additional base information leads to a decrease in the entropy value. When the additional information confirms the realization of the amino acid, the entropy value is found zero as expected. We may implement these calculations for other base sequences. Among the three genes under examination, HUMGALK1A is the easiest one to estimate the amino acids when new information is added.

The similarity between the probability distributions based on the amino acids comprising the three genes in our application is analyzed by computing Kullback-Leibler value. According to the results presented in Table 7, the genes where the amino acid distributions are the furthest are HUMCD19A and HUMGALK1A. It is seen that the amino acid distributions of HUMGALK1A and HSALADG are very similar. Therefore the amino acid sequences of genes with similar distribution may be estimated easily by examining other genes' distributions.

The joint entropy values of all genes and all base position are calculated separately from the joint probability distribution. The joint entropy values are given in Table 8. The same result is also valid for the other variables.

TABLE 6: Entropy Values for Additional Information.

Entropy	HSALADG	HUMCD19A	HUMGALK1A
Genes	4,1193386	4,07126012	3,99202660
Initial T	2,5185482	2,40382757	2,36292184
Initial C	2,1732272	2,01671637	2,16791716
Initial A	2,7419248	2,74608344	2,69412214
Initial G	2,2589254	2,24306345	2,22170036

TABLE 7: Relative Entropy Values for HSALADG, HUMCD19A and HUMGALK1A genes.

	HSALADG	HUMCD19A	HUMGALK1A
HSALADG	0	0,143158427	0,046260314
HUMCD19A	0,148137697	0	0,156046159
HUMGALK1A	0,050428046	0,183402414	0

TABLE 8: Joint Entropy for Amino Acid Variables and Adenine, Cytosine, Guanine, Thymine Base Position Variables for Three Genes.

	Joint Entropy			
Amino Acid	Adenine	Guanine	Thymine	Cytosine
HSALADG	4,371859	4,094576	4,152196	4,283639
HUMCD19A	4,298459	4,168234	4,028408	4,193326
HUMGALK1A	4,206562	4,083357	3,867612	4,267013

The consequence is the joint entropy: $H(X; Y) = 4; 371859$ where X is Adenine is and Y is HSALADG. It shows how much entropy is contained in a joint system of Adenine position and HSALADG amino acid. In probability theory and information theory, the mutual information, or transformation, of two random variables is a quantity that measures the mutual dependence of the two variables. Intuitively, mutual information measures the information that X and Y share: It measures how much we know about one of these variables and hence reduces the uncertainty about the other. For example, if X and Y are independent, then knowing X does not give any information about Y and vice versa, so their mutual information is zero. In this study, the mutual information value calculated for the Urasil base position-HSALADG amino acid can be interpreted as follows. Those two variables seem to have a lot of information in common, 0.899653 bits of information. The mutual information values also found for the Urasil base position-HUMCD19A amino acid variables and the Urasil base position-HUMGALK1A amino acid variables are interpreted in the same way. It was observed that the variable of the Urasil base position, among the mutual information values obtained in this work, was able to restrict the uncertainty on other variables so much. Table 9 exhibits the shared information between pairs of all base and amino acid variables. The pair sharing the most information is Adenine base position - HSALADG, while the least is Cytosine base position - HUMCD19A amino acid variables.

TABLE 9: Mutual Information for Amino Acid Variables and Adenine, Cytosine, Guanine, Thymine Base Position Variables for Three Genes.

	Mutual Information			
Amino Acid	**Adenine**	**Guanine**	**Thymine**	**Cytosine**
HSALADG	1,126493	0,893481	0,899653	0,765229
HUMCD19A	0,985267	0,794084	1,009066	0,717221
HUMGALK1A	1,040720	0,784268	0,908829	0,773480

CONCLUSION

Initially, the probability distributions of the bases in exons and introns of three genes belonging to human genome are examined. As a result, it is observed that the base sequences of both exons and introns are *equally random* and it is found that the probability distributions of exons are very similar to probability distributions of introns. Hence it is shown that introns can also carry information as exons do, in contrast to general agreement. If the study is repeated for the other data sets belonging to the human genome, we may obtain results concerning the similarity of the probability distributions of base sequences of exons and introns. Our work suggests that relative entropy (Kullback-Leibler distance) is useful tool in exploring the distribution of intron and exons. In analyzing the splice site regions of exons and introns, it is observed that the probability distributions of the bases are very different than the probability distributions of all the bases of exons and introns. It may be said that the last base of exons, before the GT base pair in the splice site region of genes in data set, is most probably guanine. And the first base after the GT base pair is most probably Adenine or Thymine. We may claim that one may obtain information on the splice site region of the genes by examining the probability distributions of the last bases of exons before the GT pair and the first bases of introns after the GT pair. Furthermore, when the entropy values calculated using the probability distributions of the amino acid sequences in each three genes, it is observed that HUMGALK1A has the smallest entropy value and this makes the estimation of this gene's amino acids easier. When the similarity of the amino acid distributions of the genes examined it is seen that some of them are quite close. These analyses using this method can be applied to different genes, and the amino acid sequences of genes with similar distribution may be estimated easily by examining other genes' distributions. Finally, the computation of the mutual information value between the amino acids in the genes and the sequence of bases reveals how much information the knowledge on the base sequence value provides to acknowledge the amino acids in the genes.

REFERENCES

1. Öztaş S, Gül D, Tatar A. [Human genome, genes and DNA]. Turkiye Klinikleri J Pediatr Sci 2005;1(2):18-23.
2. Mantegna RN, Buldyrev SV, Goldberger AL, Havlin S, Peng CK, Simons M, et al. Linguistic features of noncoding DNA sequences. Phys Rev Lett 1994;73(23):3169-72.
3. Schmitt AO, Herzel H. Estimating the entropy of DNA sequences. J Theor Biol 1997;188(3): 369-77.
4. Herzel H, Ebeling W, Schmitt AO. Entropies of biosequences: The role of repeats. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 1994;50(6):5061-71.
5. Tompa M. Lecture notes biological sequence analysis. University of Washington Department of Computer Science & Engineering Technical Report. Washington: Department of Computer Science and Engineering University of Washington; 2000. p.1-12, 39-43.
6. Farach M, Noordewier M, Savari S, Shepp L, Wyner A J, Ziv J. On the entropy of DNA: algorithms and measurements based on memory and rapid convergence. In: SIAM Activitiy Group on Discrete Mathematics, eds. Prog. ACM-SIAM Symposium on Discrete Algorithms. Philadelphia, PA: Soc.Industr. Appl. Math; 1995. p.48-57.
7. Cover TM, Thomas JA. Entropy, relative entropy and mutual information. Elements of Information Theory. 2nd ed. New Jersey: John Wiley and Sons; 2006. p.13-22.
8. Shannon CE. A Mathematical theory of communication. The Bell System. Technical Journal 1948;27:379-423, 623-56.
9. Kullback S. The Kullback-Leibler distance.The American Statistician 1987;41(2):340-1.