Internet Electronic Journal of Molecular Design

October 2002, Volume 1, Number 10, Pages 545–558

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday
Part 2

Guest Editor: Jun–ichi Aihara

Investigations on Evolutionary Changes in Base Distributions in Gene Sequences

Ashesh Nandy

Environmental Science Programme, Jadavpur University Jadavpur, Calcutta 700032, India

Received: July 21, 2002; Revised: August 11, 2002; Accepted: August 27, 2002; Published: October 31, 2002

Citation of the article:
A. Nandy, Investigations on Evolutionary Changes in Base Distributions in Gene Sequences, *Internet Electron. J. Mol. Des.* **2002**, *1*, 545–558, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Investigations on Evolutionary Changes in Base Distributions in Gene Sequences[#]

Ashesh Nandy*

Environmental Science Programme, Jadavpur University Jadavpur, Calcutta 700032, India

Received: July 21, 2002; Revised: August 11, 2002; Accepted: August 27, 2002; Published: October 31, 2002

Internet Electron. J. Mol. Des. **2002**, *1* (10), 545–558

Abstract

Motivation. We had observed that 2D graphical plots of DNA sequences show apparently systematic variations among members of families of gene sequences and reported these to be due to evolutionary changes. Investigations of phylogenetic relationships between species through studies of gene sequences has been one of the prime areas of interest to molecular biologists. While questions such as rate of mutational changes in DNA sequences have been investigated extensively, attention is being given recently to changes in base distribution within such sequences. Roman–Roldan et al has shown that complexity of base organization in intron segments increases with evolution.

Method. We use the 2D graphical representation method to map the DNA sequences and use four numerical techniques to quantitatively measure the differences observed in the different gene sequences.

Results. We report here our investigations using a graphical technique for DNA sequence representation and related analyses for conserved gene sequences such as kinetoplasts, heat shock proteins, globins and others.

Conclusions. We find that there are strong indications that base composition and distributions in protein coding regions progresses with evolution towards an equi–proportional composition and more complex mixing of the four nucleotides which we term as asymptotic complexity. Thus it is possible that increase in base mixing complexity is a general feature of the evolution of gene sequences.

Keywords. Molecular phylogeny; comparative genomics; evolutionary changes in DNA sequences; 2D graphical representation of DNA sequences; evolutionary changes in coding regions; DNA descriptor index; cluster density.

1 INTRODUCTION

DNA sequences are long molecular double–helical chains linked by four nucleotides or bases, adenine, cytosine, guanine and thymine. DNAs contain, in sections called genes, the molecular codes for generating different proteins. Sequences of genes from the DNAs of different organisms which code for the same protein are known to differ in base distribution and composition due to changes brought about in the course of evolution. These changes arise from various factors like mutations and from processes like gene duplication, amplification and truncation. In the case of

[#] Dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday.

* Correspondence author; E–mail: anandy43@yahoo.com.

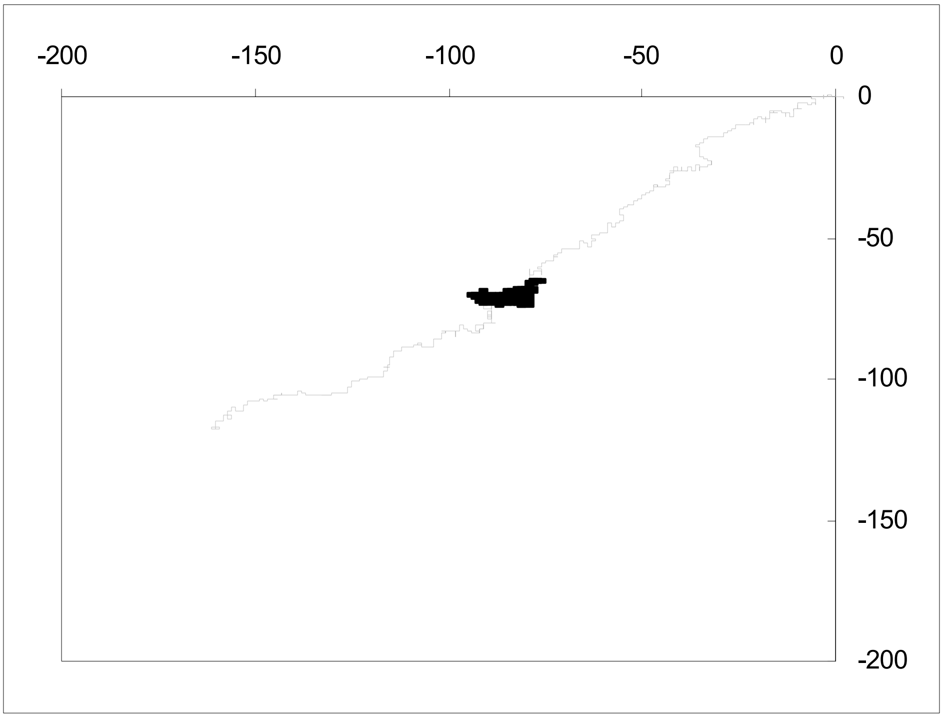
some genes like histones and globins whose protein products are essentially similar across organisms the genetic sequences retain a close similarity and are known as conserved genes; in other cases like the myosin heavy chain genes where a large variety is observable in the end products, the DNA sequences also reflect large variations in base composition and distribution [1,2].

DNAs of eukaryotic organisms generally have a mosaic structure in their genetic sequences consisting of one or more non–coding segments, referred to as introns, that break up the coding regions into smaller fragments known as exons. Evolutionary changes are brought about by point mutations in the bases and also by various other processes, some hypothesized as due to exon shuffling where the exchanges of exons between various genes can lead to the development of new genes [3]. It is evident that in the case of any particular gene taken from the DNAs of various organisms, the necessity of coding for the same or similar functional protein implies evolutionary pressures to retain the same functional structure and, therefore, arrangement of the bases that code for the protein. Yet we know that such genes are homologous to one another but not identical, *i.e.* there are additions, deletions and substitutions in the bases constituting the genetic sequence. It is of interest therefore to investigate the type and extent of such changes that take place during the course of evolution and to enquire whether there are any discernible patterns in evolutionary changes; in particular, how much of such changes take place in the exonic regions where evolutionary pressures to retain the same basic protein structures are prominent.

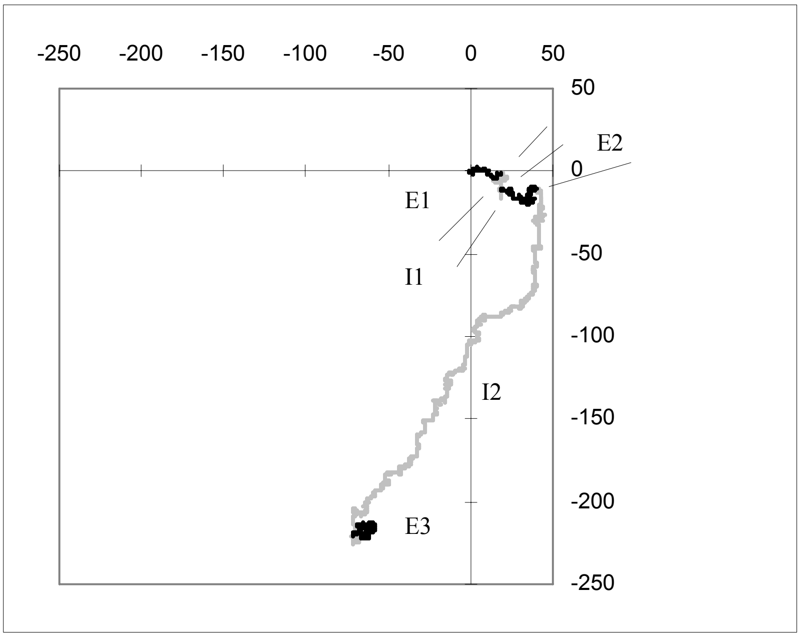
Our investigations on characteristics of conserved gene sequences have shown strong indications that molecular evolution at the gene level progresses towards an equi–proportional composition and mixing of the four nucleotides which we term as asymptotic complexity. We have observed this trend in such genes as the globin family, the tubulins, histones, heat shock proteins, actins and several others. This observation is also supported by the findings of Roman–Roldan, Bernaola–Galvan and Oliver [4] using a modified information content theory that intron complexity increases with evolution. We have used cluster density [5] and similarity measures [6] in a graphical model [7] to study the evolution of base distributions in exons and arrive at essentially the same conclusion, viz., that increase in base mixing complexity may be a general feature of the evolution of gene sequences.

2 METHODS

The basic template for investigating these aspects is provided by a graphical representation of DNA sequences (see review [8]). This is a type of DNA walk where a DNA sequence is represented by a succession of points, one for each base in the sequence, in a two–dimensional Cartesian co–ordinate system in which, to take one possible scheme, we move one step to the left in case the base is adenine, one step up for cytosine, one step to the right for guanine, and one step down for thiamine; this basically provides a running plot with the instantaneous difference between the guanine and adenine residues along the *x*–axis and that between cytosine and thiamine along the *y*–axis.



1a



1b

Figure 1. Two DNA plots generated in the 2D representation scheme with axes ACGT clockwise starting from negative *x*–axis. Figure 1a is a part of the chicken myosin heavy chain gene from base number 13500 to 14800 encompassing exon 22, Figure 1b is a plot of the complete human beta globin gene. The exon and intron segments are clearly marked in dark and light shades, respectively, and shows the clustering of points in the exon regions. The two intron segments shown in Figure 1a are 513 and 562 bases long while the exon consists of 256 bases. In the case of the beta globin gene, the three exons and two introns are marked with the symbols E and I, respectively. Base counts are: E1 = 92, E2 = 223, E3 = 129, I1 = 130, I2 = 850.

The cumulative effect is a graph of the sequence that is characteristic of the local and global base distribution in the sequence. These graphs reveal that in the case of introns with rich repetitive sequences the plot turns out to have a thin almost unidirectional run on the map whereas an exon sequence with a greater admixture of all the four bases tend to concentrate the points in smaller regions often forming dense clusters [9] (see, *e.g.*, Figure 1). We use the following four different methods to investigate this phenomenon.

1. The Cluster Density Method. We define a cluster density of a sequence segment as the number of points, or bases, in the segment divided by the enclosing rectangular box defining the maximum extent of the plot of the segment. This method determines how compactly the points are represented in the graph, but will not differentiate between the distribution of points within the enclosing rectangle. An analysis of the cluster densities calculated in this manner for intron and exon sequences show that while intron cluster densities are usually very low and fall off exponentially rapidly with density, exon densities tend to grow from a small value and then decrease gradually with density [5].

2. The base proportion measure. The second method consists of measuring directly the intra–purine, intra–pyrimidine differences. We define a ratio

$$\varepsilon = (C + G - T - A)/(C + G + A + T)$$

where the A,C,G,T represent the total number of each of these bases in the complete sequence. The parameter ε will tend to zero as the constituents approach equi–proportional representation in the selected sequence segment. This measure is a reflection of the gross deviation from equi–proportion of four bases but does not differentiate between the distribution of the bases within the sequence segment [2].

3. Similarity or Sequence Descriptor Index. The third technique is a quantitative measure of the visual clues of the graphical plots of the gene sequences. We define [6]

$$\mu = \left[(\Sigma x_i / N)^2 + (\Sigma y_i / N)^2\right]^{1/2}$$

where the x_i and the y_i represent the co–ordinate points of each of the bases in the plot and N is the total length of the sequence used for normalizing. This object μ is found to be sensitive to the base distribution in a sequence and forms the basis for use as a descriptor of sequence changes [10].

4. Information content method. Our fourth technique is to measure the information content of coding and non–coding regions of gene sequences. The information content of S may be obtained using Shannon's formula as:

$$I_c(S) = -\sum_i p_i \log_2 p_i$$

where

$$p_i = n_i / n \geq 0, \Sigma p_i = 1, n_i \geq 0$$

and I_C is expressed in bits. In particular, if n is partitioned into 4 disjoint classes, then maximum value of I_C is 2 bits. In the present context the nucleotides of a gene sequence is composed of four bases A, C, G, T forming four classes and the number of the nucleotides in the sequence, namely *n*(A), *n*(C), *n*(G), *n*(T) are the cardinalities of the respective partitioned classes.

In the case of equal representation of the four bases in a DNA sequence segment, i.e. in the case where the bases A, C, G, T are present in equal numbers in the segment, then $p_i = ¼$, $\log_2 p_i = –2$ and I_C will have the maximum value, i.e. $I_C = – [4 \times ¼ \times –2] = 2$. If there is a repeated run of only one base, *i.e.* only one base is represented in the sequence segment, I_C for the segment will reduce to zero; in all other cases I_C will lie in the range $0 < I_C < 2$. Again, this measure also is invariant to changes in base distribution within the sequence segment.

Thus, these four methods provide us information on the following aspects of a DNA sequence segment: Two parameters are dependent upon the total numbers of the four bases in a segment but do not provide any indication of the mixing of the bases within the segment, and two others provide a measure of the mixing of the four bases within the segment.

To summarize: (*a*) the information content parameter provides an estimate of the proportion of the four bases in the segment; (*b*) the mixing parameter determines mismatch in the C–G component and the A–T component as a ratio of the total numbers of the four bases; (*c*) the cluster density sets the upper limit of the mixing of the four bases but cannot determine any finer mixing; and (*d*) the sequence descriptor index provides a measure of the base distribution and composition of a sequence segment and is sensitive to even a single change in the distribution and composition.

3 RESULTS AND DISCUSSION

We have studied gene sequences from the globin family, the histones, kinetoplasts, tubulins, cytochrome C, actins and the myosin heavy chains. We found that in the case of the conserved gene families the 2D plots of all members of the same family have graphs that are shape similar, but the later organisms are represented reduced in extent along one or both the *x*– and the *y*–axes (see, *e.g*., Figure 2); interestingly, these are true for conserved intronless genes such as histone 4, as well as those with introns.

Since these are basically plots of the instantaneous intra–purine (C–T) and intra–pyrimidine (G–A) differences along the two axes, such shrinkages indicate that the intra–purine and intra–pyrimidine differences are less for these later organisms, indicating more equi–proportional representation of the purines and the pyrimidines in the sequences.

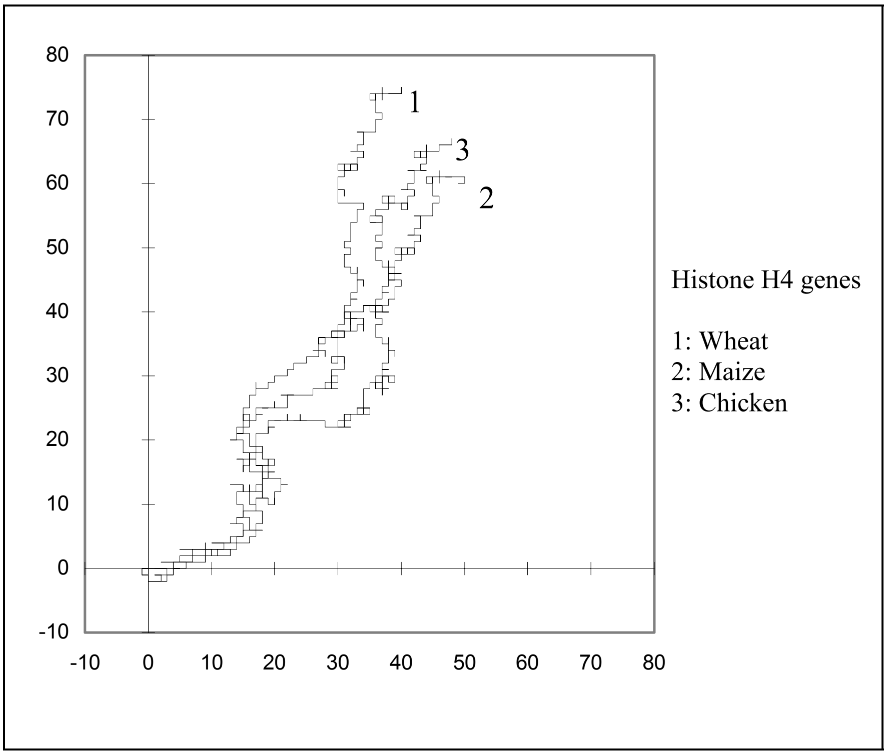


Figure 2. Histone H4 genes of four species plotted in a 2–D scheme (with same axes system as in Figure 1) to show variations with evolution.

Table 1. Genewise table of information content and base proportion parameters

Gene	A	C	G	T	TOTAL	I_C	CG–TA/ Tot%
5S rRNA							
Yeast	31	31	33	37	132	1.9961	–3.03
Mouse	19	20	36	20	95	1.9414	17.89
Rat	22	33	39	27	121	1.9677	19.01
Human (Art)	25	36	42	33	136	1.9760	14.71
Histone H4							
Wheat	62	111	102	37	312	1.8858	36.54
Maize	63	96	111	42	312	1.9092	32.69
Chicken	62	104	108	38	312	1.8913	35.90
Mouse	65	96	100	51	312	1.9479	25.64
Rat	68	93	100	51	312	1.9528	23.72
Human	73	79	100	60	312	1.9756	14.74
Histone H2A							
Xenopus	91	124	118	60	393	1.9489	23.16
Mouse	77	137	133	46	393	1.8820	37.40
Human	94	118	116	65	393	1.9638	19.08
RNA Polymerase II							
Rat	266	514	79	233	1092	1.7576	8.61
Macaque	249	539	91	213	1092	1.7478	15.38
Human	251	537	91	213	1092	1.7498	15.02

Table 1. (Continued)

Gene	A	C	G	T	TOTAL	I_C	CG–TA/ Tot%
Cytochrome–c							
Yeast	121	54	85	82	342	1.9440	–18.71
Mouse	113	57	79	69	318	1.9524	–14.47
Rat	113	57	82	66	318	1.9500	–12.58
Human	109	59	82	68	318	1.9604	–11.32
Alpha Globin							
Horse	16.2	40.4	29.3	14.1	100	1.8711	39.40
Orangutan	15.8	38.4	29.3	16.5	100	1.8987	35.40
Rhesus Mon	15.2	38	30.3	16.5	100	1.8944	36.60
Goat	15.9	37.6	30.1	16.4	100	1.9016	35.40
Mouse	19.8	29.9	28	22.3	100	1.9804	15.80
Human	16.2	37.5	29.7	16.6	100	1.9064	34.40
Beta Globin							
Opossum	21.4	23.6	27.9	27.0	100	1.9920	3.15
Mouse	21.3	27.2	28.8	22.7	100	1.9889	12.02
Rat	22.0	25.2	28.3	24.5	100	1.9941	7.03
Goat	20.4	23.9	31.4	24.3	100	1.9820	10.55
Lemur	19.1	24.3	31.3	25.2	100	1.9785	11.26
Rabbit	21.1	23.6	30.6	24.7	100	1.9862	8.39
Human	19.8	25.7	30.6	23.9	100	1.9826	12.61
Beta Globin x3							
Opossum	27	31	36	35	129	1.9910	3.88
Mouse	23	40	37	26	126	1.9624	22.22
Rat	25	36	36	29	126	1.9835	14.29
Goat	22	36	40	31	129	1.9672	17.83
Lemur	21	33	43	32	129	1.9567	17.83
Rabbit	25	33	34	34	126	1.9891	6.35
Human	27	37	35	30	129	1.9890	11.63
Beta Globin x3 – 2nd half (last 63 bases)							
Opossum	11	19	19	14	63	1.9649	20.63
Mouse	12	20	20	11	63	1.9463	26.98
Rat	13	19	19	12	63	1.9686	20.63
Goat	12	16	20	15	63	1.9763	14.29
Lemur	11	16	21	15	63	1.9631	17.46
Rabbit	13	17	20	13	63	1.9751	17.46
Human	13	16	20	14	63	1.9797	14.29
Myosin Heavy Chain							
Slime Mould	37.8	19.2	18.5	24.5	100	1.9352	–24.60
Nematode	30.8	23.5	23.7	22	100	1.9871	–5.60
Chicken	31	22.6	28.7	17.7	100	1.9677	2.60
Rat	28.8	24.4	29.9	16.9	100	1.9680	8.60
16S rRNA							
CYAN16	353	332	458	268	1411	1.9733	11.98
AG16	357	328	458	293	1436	1.9795	9.47
PURPL16	346	336	478	276	1436	1.9706	13.37
SULF16	321	418	519	232	1490	1.9393	25.77
METHANE16	328	347	460	272	1407	1.9735	14.71

Table 1. (Continued)

Gene	A	C	G	T	TOTAL	I_C	CG–TA/ Tot%
Archaebacteria							
hmsoda	128	200	180	104	612	1.9532	24.18
hssodb1	112	212	195	84	603	1.9041	34.99
hssodc	111	220	193	79	603	1.8904	36.98
hsgyrb	407	658	588	267	1920	1.9225	29.79
hvsodb	112	206	190	92	600	1.9217	32.00
hvsoda	113	207	190	93	603	1.9232	31.67
mbargg	336	293	327	235	1191	1.9868	4.11
mvctsb	453	162	267	312	1194	1.9106	–28.14

Table 2. Genewise comparative information content figures

Table 2a. Information content of coding regions

Gene Type	No of Samples	Coding Region (cDNA)				
		Avg. $I_C \pm$ STD			Max I_C	Min I_C
α–globin	10	1.9554	±	0.0251	1.9954	1.9292
β–globin	10	1.9857	±	0.0077	1.9985	1.9737
a–tubulin	9	1.9773	±	0.0144	1.9943	1.9514
b–tubulin	15	1.9727	±	0.0197	1.9969	1.9349
actin	15	1.9779	±	0.0221	1.9999	1.9167
histone	17	1.9352	±	0.0364	1.9756	1.8820
hsp70	13	1.9563	±	0.0302	1.9974	1.9127
		1.9658	±	0.0222		

Table 2b. Information content of non–coding regions

Gene Type	No of Samples	Non–Coding Regions (Intron)				
		Avg. $I_C \pm$ STD			Max I_C	Min I_C
α–globin	10	1.8244	±	0.1032	1.9828	1.6493
β–globin	10	1.9147	±	0.0508	1.9756	1.7902
a–tubulin	9	1.7762	±	0.2527	1.9695	1.1653
b–tubulin	15	1.8974	±	0.1160	1.9926	1.5733
actin	15	1.8771	±	0.1115	1.9778	1.6552
		1.8580	±	0.1268		

Table 2c. Comparative values for some mammalian I_C

Genes	Information Content Values		
	Mouse	Rat	Human
5S rRNA	1.9414	1.9677	1.9760
Histone H4	1.9479	1.9528	1.9620
Histone H2A	1.8820		1.9638
Cytochrome C	1.9524	1.9500	1.9604
α–globin	1.9775		1.9299
β–globin	1.9889	1.9941	1.9826

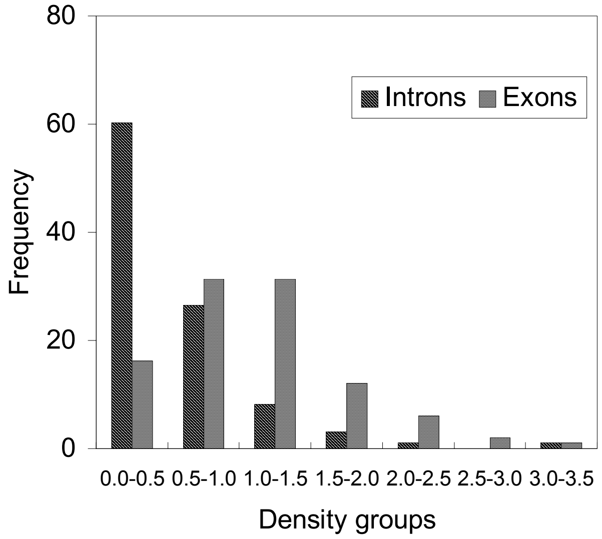


Figure 3a

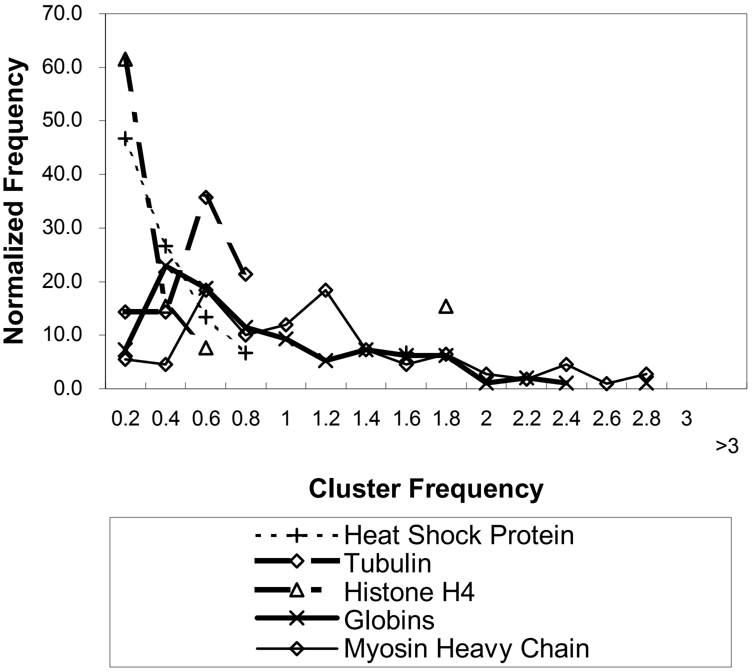


Figure 3b

Figure **3.** Figure 3a shows intron and exon cluster density frequencies of all genes considered, Figure 3b shows the cluster frequencies for exons of various genes. All frequencies are normalised. The earlier genes like kinetoplast and HSPs show low cluster frequencies, the later genes like globins and myosin heavy chain genes show increasing reach of the frequencies.

A study of the base proportion parameter, ε, shows that this is indeed happening across a wide variety of genes and species (Table 1). While the ε for the horse alpha globin gene, inclusive of exons and introns, is 39.4%, that for the human gene is reduced to 28.67%. Similarly, for an intronless gene like the histone H4 too, ε reduces from a large 36.54% for the wheat histone to 25.64% for the mouse histone. Comparing between genes, the alpha globin is known to predate the beta globins by about 300 million years [11]; the base proportion parameter for the alpha and beta globin genes of mouse is seen to be 15.8% and 12.02%.

The information content parameter, I_C, treats all four constituent bases on an equal footing. Its value for all coding sequences turns out to be close to 2, the maximum possible value for four bases, and the differences in the information content between different species of each gene are very small; but here also the trend towards higher information content for the later species is apparent (Table 1). Table 2 shows a summary of the information content for various genes and species. Table 2a of the information content for coding regions shows that I_C is very close to 2 and that the spread between the maximum and minimum values of I_C is very small. In fact, for the later genes like beta globin the spread calculated over 10 sample species is much less than that for the alpha globin genes. Similarly, histones and heat shock proteins which are arguably older than the globin genes have a lower I_C value and higher spreads. Likewise also, comparative figures for different genes across species show that the I_c for humans are generally higher than for mouse and rat except for the alpha–globin gene. Similar trends are noticeable in the case of the intron I_C also: the later genes have higher I_C values and slightly smaller spreads.

Table 3. Average cluster densities of members of the globin gene family

	Cluster Densities			
Gene	Exon 1	Exon 2	Exon 3	Average
Alpha	0.506	0.211	0.474	0.397
Zeta	0.738	0.223	0.377	0.446
Myoglobin	–	–	–	0.491
Beta	0.886	1.371	0.654	0.970
Delta	0.787	1.132	0.765	0.885
Epsilon	0.676	2.059	0.956	1.231

While the above two measures demonstrate that the total number of the four bases in the different genes appear to trend towards equiproportion, the values of the graph radius and the cluster density parameters show that this trend is also consistent with greater mixing of the four bases within the sequences. Detailed presentations of the results of cluster density analysis and of the graph radius parameters have been given elsewhere [5,6,9]. Figure 3 represents cluster frequency analyses for the genes considered in this paper. It is observed that in general, as mentioned hereinbefore, the cluster densities of introns are generally very low, and higher density introns get exponentially rapidly scarcer; the frequency of exons, however, grow from small values to around 0.6–0.8 and die away gradually. Genewise, the earlier genes such as kinetoplasts and heat shock proteins tend to have the base distributions in their exonic segments generally at the low density levels whereas the base distributions in later genes such as globins and myosin heavy chains have increasingly larger presence in the higher cluster density regions, reflecting tighter clustering in the graphical representations. Even amongst the members of a single gene family such as the globins, earlier genes such as alpha globins have low density base distributions, whereas later genes such as the beta and eta globins have higher densities (Table 3) implying increasing complexity in base mixing. Again, the graph radius calculations given in Ref. [9] show that the later species such as goats and monkeys as compared to horses have tighter clustering thus reflecting higher complexity in mixing of bases in the related DNA sequences.

In the event that base composition would have been equi–proportional and distribution completely randomized, clustering would reduce to a minimal four points with maximum density, the graph radius would have decreased to almost zero, the base mixing parameter ε would be zero and the information parameter, I_C, would have stabilized at 2. Such a situation would leave very little room for further changes and mutations in the resultant proteins. In all the genes we have studied so far we have not noticed any sequence with these asymptotic attributes and believe that perhaps this is indicative of some natural process to optimization in base composition and distribution.

Our results thus show that not only the base composition of the sequences is tending towards equi–proportion, but also the base distribution within the sequences is tending towards asymptotic complexity in the mixing of the four bases. This is in tune with the observation of Roman–Roldan *et al*. [4] that intron sequences tend towards greater mixing of the four bases with evolution. It is interesting to note that an analysis using fractal techniques on our 2D–graphical representation technique also reveal the same feature [12].

Table 4. Histone H4 – detailed analysis and comparative studies

1. Base Composition						
		A	C	G	T	ε (%)
Wheat	TAH4091	62	111	102	37	36.5
Maize	ZMH4C1	63	96	111	42	32.7
Chicken		62	104	108	38	35.9
Mouse	MMHIS4	65	96	100	51	25.6
Human	HSHISAD	73	79	104	60	16.0
2. Base substitutions required to revert to wheat H4 sequence						
	Position:	1st	2nd	3rd	Total	
Maize			1	19	20	
Chicken		5		24	29	
Mouse		5		31	36	
Human		6	1	46	53	
3. Base substitutions in Human H4 compared to Wheat H4 Histones (A–C implies A of wheat has been replaced by C in Human H4)						
A–C	3	C–A	9	G–A	5	T–A –
A–G	2	C–G	8	G–C	4	T–C –
A–T	1	C–T	19	G–T	2	T–G –

To investigate a possible explanation for these observations, we note that DNA sequences are known to undergo random mutations in the bases, and also that there is a wide degree of allowable degeneracy in the third codon position which does not affect the final protein. We considered the histone H4 gene in detail to see whether such phenomena could explain the observation of asymptotic complexity. The details are shown in Table 4. We find that while a large part of the changes can be explained by these effects, there still remains a part that requires a separate explanation. The histone H4 genes considered here have exactly 312 bases for each species. The maximum changes in the 104 amino acids between wheat and the rest are 2; the rest of the changes do not affect the amino acids that are coded. There are a total of 52 base changes between the earliest of the species, wheat, and the latest, human; there are 26 changes between chicken and wheat. The greatest deviation from equiproportion of the four bases is in wheat, the least in human. In comparison with the wheat histone, the majority of the base changes required to arrive at the human histone is in reduction in cytosine base content of the wheat histone; other species also follow a similar trend. A null hypothesis would require that codon degeneracy lead to sufficiently random changes such that no trend should be discernible between later and earlier species.

Note that while several amino acids have full degeneracy of the four bases in the third position, others show partial degeneracy as indicated below (for our purposes T in the DNA sequence and U in the amino acid case are equivalent):

Group 1: Complete Degeneracy: Thr, Pro, Ala, Ser, Arg (with C as the first base in the codon), Gly, Leu, Val

Group 2: Partial Degeneracy:

A,G symmetry		C,T symmetry	
Arg	AGA, AGG	Asn	AAC, AAU
Gln	CAA, CAG	Asp	GAC, GAU
Glu	GAA, GAG	His	CAC, CAU
Lys	AAA, AAG	Ile	AUA, AUC, AUU
		Phe	UUC, UUU
		Tyr	UAC, UAU

This implies that random mutations in 3rd base position should be equal for the A and G, and for the C and T for the second group of amino acids, whereas all four bases are equally likely to undergo random mutations in group 1. However, in the case of histone H4 (Table 4.3), there is not a single mutation that changes T to anything else, *i.e.*, the C, T third base codon symmetry is broken.

Analyses of the other genes also show similar trends. Clearly, something more than codon degeneracy is required to explain these phenomena. Miramontes *et al.* [13] had observed that there are possibly hitherto unknown pressures that restrict evolution of base complexities in genomic sequences. We have shown here that evolution tends to move base composition and distribution in a direction of asymptotic complexity in the exons of later genes and later species within the same gene family, while introns seem to have very low cluster densities in our representation and share only in a small way the same evolutionary features. These results for introns are intuitively understandable from the observation that intron segments generally tend to have a bias towards A/T or G/C richness and therefore generate long runs in our graphical plots as can be seen in our Figure 1, whereas exon sequences with more equiproportionate composition and distribution of the bases form dense clusters of points. The apparent lack of evolutionary pressures on intron segments can lead to rapid growth and changes in these segments through gene duplication, addition and truncation, which are not generally available to exonic regions, and therefore will trend with evolution to a lesser extent. The surprising element of the cluster density analysis is that not only the exons have the four bases in almost equal proportion but that they are mixed within the segments in a way that tend asymptotically to optimum complexity as evidenced by the base proportion and other parameters.

Thus, this observation of asymptotic complexity cannot be explained by codon degeneracy alone and we have to look elsewhere for an explanation. We hypothesize this feature could arise from a general entropy consideration that entropy would tend to break down strong asymmetries into greater complexities within the overall constraint of evolutionary pressures. It is interesting to consider thermodynamic free energies in DNAs in this context. The works of Breslauer *et al*. [14] have shown that the stacking energies between consecutive nucleotides that contribute to a chain's stability are greater than the hydrogen bonding energies that keep the two chains of a DNA bound together. A completely random distribution of nucleotides with all four bases present in equal proportion would yield an average stacking energy of –1.99 kcal/nt. Coding regions of DNAs exhibit stacking energies close to this number; *e.g*., chicken myosin heavy chain gene protein coding region has a stacking energy per nucleotide of –1.98, human beta globin has –2.02. Intronic regions, however, show a wide variance: –1.78 kcal per nucleotide for the chicken myosin heavy chain introns, which are mainly AT–rich, to as large numbers as –2.71 for some of the tubulin gene introns which are GC–rich. Thus, introns, especially in higher eukaryotes, that are AT rich have much lower binding energy per nucleotide; GC rich introns have higher energies. We have noticed that coding regions of some of the older genes that have low values for the cluster densities implying comparatively lower base mixing complexities have a greater difference from the median value of –1.99: thus, kinetoplast exons have stacking energies of –1.6 per nucleotide, heat shock proteins (*D melanogaster* hsp 82) have energies of around –1.8 per nucleotide, tubulins –2.3 per nucleotide, etc. Comparing the older and comparatively later genes, these numbers also appear to indicate that over evolutionary time scales the mixing of bases within protein coding regions are tending towards maximizing complexity, a result in keeping with all the other methods of measuring this aspect that we have outlined hereinbefore. For species variations for individual genes, we have to wait for a more precise determination of the stacking energies to pick out the differences.

4 CONCLUSIONS

Thus we have shown that there exists evidence for mixing of bases in protein coding regions to tend towards asymptotic complexity during evolution of genes. The reason for such complexity is however not immediately obvious. Greater mixing of the four bases may imply increasing entropy and greater opportunity for nature to experiment with newer combinations of amino acids. Or it could lead to greater adaptability of the organism with a varied resource. Human species being the last of the evolved organisms have the greatest mixing of the four bases in the gene sequences we have studied; in fact, we find that the ε–factor is almost 15% for most of these genes, leaving comparatively little room for further changes in the proportion of the four bases in base composition and distribution in human gene sequences

5 REFERENCES

[1] R. Nussinov, Compositional variations in gene sequences, *Comput. Appl. Biosci*. **1991**, *7*, 287–293.

[2] A. Nandy and P. Nandy, Graphical analysis of DNA sequence structure: II. Relative abundances of nucleotides in DNAs, gene evolution and duplication, *Current Sci*. **1995**, *68*, 75–85.

[3] W. Gilbert, Genes–in–Pieces Revisited, *Science* **1985**, *228*, 823–824.

[4] R. Roman–Roldan, P. Bernaola–Galvan and J. L. Oliver, Sequence compositional complexity of DNA through an entropic segmentation method, *Phys. Rev. Lett*. **1998**, *80*, 1344–1347.

[5] A. Nandy, Graphical analysis of DNA sequence structure: III. Indications of evolutionary distinctions and characteristics of introns and exons, *Current Sci*. **1996**, *70*, 661–668.

[6] C. Raychaudhury and A. Nandy, Indexing Scheme and Similarity Measures for Macromolecular Sequences, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 243–247.

[7] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. Methodology and Application to Globin Genes, *Current Sci*. **1994**, *66*, 309–314.

[8] A. Ray, C. Raychaudhury, and A. Nandy, Novel Techniques of Graphical Representation and Analysis of DNA Sequences – A Review, J. Biosci. **1998**, *23*, 55–71.

[9] A. Nandy, Two dimensional graphical representation of DNA sequences and intron–exon discrimination in intron–rich sequences, *Comput. Appl. Biosci*. **1996**, *12*, 55–62.

[10] A. Nandy and S. C. Basak, A simple numerical descriptor for quantifying effect of toxic substances on DNA sequences, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 915–919.

[11] M. W. Strickberger, *Evolution*, Jones and Bartlett Publishers Inc, USA **1990**, Ch 12.

[12] S. Tarafdar, P. Nandy, S. Sahoo, A. Som, J. Chakrabarti, and A. Nandy, Self–similarity and scaling exponent for DNA walk model in two and four dimensions, *Indian J. Phys*. **1999**, *73B*, 337–343.

[13] P. Miramontes et al, Structural and thermodynamic properties of DNA uncover different evolutionary histories, J Mol Evol **1995**, *40*, 698–704.

[14] K. J. Breslauer, R. Frank, H. Blocker, and L. A. Marky: Predicting DNA duplex stability from the base sequence, *Proc. Natl. Acad. Sci. USA* **1986**, *83*, 3746–3750.