Acta Universitatis Latviensis, 2008, Vol. 745, Biology, pp. 187–198

An alternative model of the origin of genes: quantization of intron dimensions

Gunars Chipens, Nora Ieviņa*, Ivars Kalvinsh

Latvian Institute of Organic Synthesis, Aizkraukles 21, Rīga LV-1006, Latvia
*Corresponding author, E-mail: ievina@osi.lv

Abstract

More than 30 years after the discovery of split genes the problem of intron origin is not yet completely solved. To study the possible germ of life on Earth we developed a quite different point of view – the third way of origin of genes – as an alternative to the Exon theory of genes and the Insertional theory of intron origin. In accordance to the elaborated model the precursors of primeval genes including segments of the future introns were formed by oligonucleotide multiplication and duplication reactions. The gene precursors without introns and primeval genes with the very first introns were regular periodic nucleic acids containing tandemly repeated identical in size and sequence oligonucleotides. Here we demonstrate several contemporary gene families whose members have retained regularity corresponding to oligonucleotide repeats. Regular segments of these gene structures – peculiar molecular relics – contain exons, introns and intron coordinates whose numerical parameters have identical internal regularity and can be quantized – expressed as multiples of identical size oligonucleotide repeat units. The term „gene quantum" in this case shows the number of nucleotides or base pairs in an oligonucleotide named a repeat unit.

Key words: albumin-fetoprotein gene family, globin gene family, regularity of gene structures, superfamily of carbonic anhydrases, tubulin α-1A.

Introduction

The necessity to investigate possible ways of the genetic code and gene origin as a complex problem arose for us in 1964 when at the Institute of Organic Synthesis new directions of Life Sciences were developed, including peptide chemistry and biology, with the aim to synthetize new active analogues of natural peptide bioregulators for use in medicine. Two main problems arose at once – the physico-chemical and functional (biologically tantamount) relatedness of amino acid groups and the genetic (proteomic) code structure of amino acid interaction. However, the problem of origin of the genetic code and the problem of gene and intron origin are intimately and closely connected and thus we studied both of them.

The discovery of split genes in 1977 was completely unexpected (Abate 2001), shocked scientists and promoted very intensive scientific investigation directed to understand intron origin, evolution and functions. Now internet server "Google" offers more than 504 000 files describing different models of intron emergence, thus indicating that the problem is not yet solved – for advanced models there is not yet quite enough convincing and corroborative experimental evidence. Interesting models and discussions about intron

origin can be found in articles Fedorow et al. 2003; Rogozin et al. 2005; Roy, Gilbert 2005; Koonin 2006; Roy, Gilbert 2006, and others.

During the last decade we have been interested in the way of origin of the very first gene precursors – on a day before the creation of the very first primeval genes, when only the primitive nucleic acids existed which, according to our viewpoint, were short simple structure regular and periodic polinucleotides without introns. The fundamental thesis of the supposed third way model is the regularity of gene precursor structures, which were formed of tandemly repeated oligonucleotides named repeat units (RU). Their dimensions correspond to the gene quantum Q of the future genes containing exons and introns. For studies of such a model, taking into account the very long time of evolution and a lot of different mutations, there is only one possibility – to search for regularity of exon and intron dimension numerical parameters, as well as for regularity of a gene itself, as an individual regular entity formed of a whole number of repeat units. We have revealed such a regularity in several gene families.

As an simple example can serve mouse gene tubulin alpha 1A (GenBank accession NC_000081.5; Gene ID:22142; Ensebl release 49 – Mar 2008). The sum of four exons of the coding part of this gene including iniciation and termination codons is 1356 nucleotides (nt) or base pairs (bp) (1356 = 6nt × 226 = 3nt × 454). The sum $\sum_i$ of three intron dimensions which separate the four exon row is 2064nt (2064 = 6nt × 344 =3nt × 688). The sum total of exon (E) and intron (I) dimensions $\sum_{(E+I)}$ is 3420nt (3420 = 6nt × 570 = 3nt × 1140). So, the sum total of exon and intron dimensions separately and totally include a whole number of codons (n × 3nt) where n is a whole number. Below there will be shown similar examples with pentanucleotide and heptanucleotide repeats.

Before the accumulation of a large data base of gene parametres and before elaboration of exact methods of determination of the gene precursor RU sizes we will use the minimal numerical values for the potential gene guantum Q (as a smallest unit of physical parameter). Thus, the mouse tubulin alpha 1A gene quantum Q conditionally is 6nt, but may be also 12 nt; it is necessary to analyse a large amount of the tubulin family genes (Ievina, Chipens 2003).

This gene also confirms our idea that gene exons and introns in the primeval gene were formed of identical in size RU, but a gene must be regarded as a whole regular entity (with the exception of mosaic like genes formed by exon shuffling). During evolution gene component dimensions (particularly introns) were changed, but the internal regularity of a whole gene in many cases was retained unchanged. Completely regular parameters of genes or large regular gene regions must be regarded as gene molecular „fossils" or relics.

Materials and methods

It is impossible to solve very tangled and difficult problems by using the same notions and terms that have led us to these problems. Therefore when investigating intron origin we use principles of information theory and concepts of signature and equivocation (Quastler, 1965; Chipens et al. 1979; Chipens 1980; Chipens et al. 1988; Chipens 1991) and simultaneously study also the origin of the genetic code (Chipens 1996, Chipens, Ievina 2004; Chipens, Ievina 2005), because the problem of origin of genes and introns as well as the problem of origin of genetic translational and amino acid interaction (proteomic) codes are closely connected. To solve the complex problem of gene and intron origin it was

necessary to elaborate many new analytical methods and even new principles of analysis based mainly on the symmetry and antisymmetry of the genetic and proteomic codes and codon root sequences of genes or amino acids shortly enumerated below. These methods (described in our ealier publications: Chipens, Ievina 1994; Ievina, Chipens 2003 and 2004; Chipens, Ievina 2005; Chipens 2006; Chipens et al. 2006; Ievina et al. 2006) were used also in this work.

The discovery of the covert symmetry of the 2D-genetic code, which determines the potential equifunctionality of amino acids having identical second codon letters (the codon roots) was of key significance to reveal gene regularity and intron origin and give a possibility to use widely the principially new method of gene and protein investigation – Comparative Amino Acid Codon Root Analysis (CAACRA) (Chipens 1991; Chipens, Ievina 1994; Chipens 1996). CAACRA allowed to determine and later also to calculate exon and intron internal regularity as well as led to the idea of gene quantum Q and quantization of discrete numerical parametres of genes (Chipens et al. 2005; Ievina et al. 2006).

Functionally tantamount amino acids have identical codon roots (Chipens et al. 1988; Chipens 1991), but the regularity of gene and protein sequences is determined by the translational symmetry of their codon root sequences (Chipens et al 2006). The translational symmetry can be determined only using CAACRA and repeat unit piles (RUP) because mutations change the amino acid sequences, but in many cases retain unchanged their codon roots (Chipens 1991; Chipens, Ievina 1996; Ievina, Chipens 2004).

The physico-chemical basis of CAACRA is the proteomic code, which can be demonstrated by rotational symmetry analysis of the 2D-genetic code structure in a form of a square (Chipens et al. 1988; Chipens 1996) or best of all in a form of windmill vanes (Chipens, Ievina 2004). Methodology and examples of protein and gene codon root sequence analysis using CAACRA and RUP are described in our previous publications (Chipens, Ievina 1999a; Chipens, Ievina 1999b; Ievina, Chipens 2004; Chipens et al. 2005; Ievina et al. 2006).

Results and discussion

The "Third way" model of gene and intron origin

Contrary to the well known „introns early" (Gilbert, 1987) and „introns late" (a review, Logsdon, 1998) theories according to the "third way" model primeval introns and exons are products of internal evolution of one and the same gene polynucleotide chain nucleic acid (most likely of the RNA) world. Gene precursors were highly regular periodic nucleic acids formed of identical in size and sequence oligonucleotides named repeat units (RU). Exons and introns in gene structures originated after the emergence of the very first splicing machinery. Introns in the absence of constraints imposed by the coding function as well as natural selection on the level of proteins during billions of years of evolution accumulated mutations without any limits. If the hypothesis of oligonucleotide multiplication is correct, modern gene structures could have retained some regularity of codon root and possibly also amino acid sequences.

The splicesome is the most complex macromolecular machinery of the contemporary cell (Nilsen, 2003), but independently of the biochemical mechanisms of the very first splicing machinery the splicing sites can be determined only by definite nucleotide

structures in the gene precursor nucleotide chain – the RU multimer. In accordance with our working hypothesis the splicing sites may be accidentally encoded in RU precursors or formed during multistep (the first or higher steps) multiplication reaction by interaction of RU 3'- and 5'-terminal nucleotides. In such cases potential splicing sites were distributed along the nucleotide chain – a multimer of repeats – internally regularly within the intervals n × RU(nt). Consequently, the exon dimensions (expressed by number of nucleotides, nt) as well as intron coordinates (measured by ordinal numbers of 3'-nucleotides of exon row just before introns) were internally regular determined by a whole number of RU (n × RU, nt). Each number can be expand into factors (prime number multipliers). If the numerical values of gene basic parameters have a set of common factors – they possess a comon internal regularity. This regularity is strictly determined by RU size measured by a number of nucleotides (or base pairs, bp) and marked by a special term, the gene quantum Q. According to our hypothesis the basic numerical parameters in regular regions of gene precursors were discrete and can be quantized (n × Q, where n is a whole number).

Thus, multiplication reactions of oligonucleotides form long polynucleotide chains with open reading frames (in cases when RU do not contain or form termination codons).

Oligonucleotide repeat units of the albumin and carbonic anhydrase genes

A scheme of a linear one-dimensional nucleotide chain of gene exon row can be depicted as a stright line with imaginary dots located regularly within intervals corresponding to the RU size. These dots we named the gene knot points (GKP). GKP have a key significance in analysis of regularity of the basic gene numerical parametres (exon length, intron position coordinates and total length of exon row) using as a unit of measure one nucleotide (base

Table 1. Numerical parameters expressed by number of nucleotides (E - exons, I - introns, C - intron coordinates as a sum of the preceeding E length) of the mouse albumin (Alb) in accordance with GenBank data (NC_000075.5, gene ID: 11657). The gene parameters which are multiples of 7nt are framed. *Symbol Σ and the number denoted by asterisk show the sum of exons No 1-14 or introns No 1-13

E/I, No	E (nt)	C (nt)/phase	I (nt)
1	88	88/1	2824
2	49	137/2	206
3	133	270/0	1439
4	212	482/2	847
5	133	615/0	583
6	98	713/2	740
7	130	843/0	5018
8	215	1058/2	13680
9	133	1191/0	2307
10	98	1289/2	817
11	133	1422/0	1288
12	224	1646/2	958
13	133	1779/0	1067
14	48	Σ_E 1827/- *	$\Sigma_{I\ No1\text{-}13}$ 31774nt

pair) and the size of RU. GKP show the borderlines between the neighbour RU. Position of GKP or RU in the exon row polinucleotide chain can be characterized by the corresponding coordinate – the ordinal number of RU 3'-terminal nucleotide, because conditionally each RU ends by a GKP. The reference point usually is nucleotide No. 1 of the first exon.

Theoretically gene exon nucleotide chains must consist of a whole number of RU. Introns crossing GKP split the nucleotide chains, thus forming exon row. If the intron coordinates correspond to the GKP then the exon dimensions are internally regular. It goes without saying that mutations, particularly transversions and indels, as well as intron sliding disrupt this regularity sometimes completely. Therefore it is necessary to search for molecular „fossils" – segments of contemporary genes which have retained the regularity of ancestors.

The potential sizes of RU can be calculated as products of common prime numbers (factors). Such calculation usually shows small numerical values of RU dimensions and a serious problem arises about their real existance. To answer to this important question it was necessary to test and confirm the "third way" model, which was achieved by analysis of the mouse albumine (Alb) gene (Waterston et al. 2002). This gene, if we cut off the 3'-terminal intron and nucleotide sequence of poliadenylic acid (Fig. 1, Table 1) has 14 exon/13 intron structure (GenBank accession number NC_000071.5). More than a half (64.3 %) of Alb exon dimensions can be expressed as a multiple of gene quantum Q = 7nt indicating that the primary repeat unit contains 7 nt and is a heptanucleotide. It is important to underline that some of these exon sizes can be expressed only as products of two prime numbers one of which is always seven (e.g. $49 = 7 \times 7$; $133 = 7 \times 19$) confirming the heptanucleotide size of RU.

It is interesting to note that from the time immemorial the number seven is ascribed as magnificient and lucky number having peculiar strength (Miller 1956). For us it was indeed a lucky number because it firstly introduced the idea that a gene containing exons and introns must be regarded as a whole regular entity and confirmed our idea that separate intron-molecular relic dimensions (4 and 11, Table 1) could be quantized similarly as those of exons.

The sum total (Σ) of the mouse Alb gene thirteen introns (I) 1-13 is 31774 nt (Table 1 and Fig. 1) and differs from the calculated value (if Q = 7nt) only by +1 nt (possibly a nucleotide insertion). Correction of this parameter (i.e."minus" 1 nt) gives the value 31773 nt (31773 : 3 = 10591 codons; 31773 : 7 = 4539 heptanucleotide RU) which together with

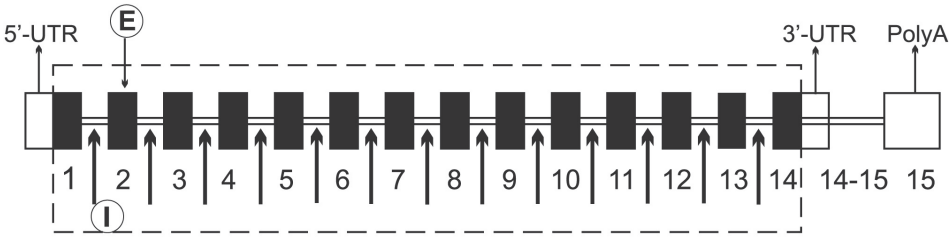

Fig. 1. A scheme of the mouse albumin (Alb) gene structure. The analyzed gene segment is framed by a dashed line. Exons (E) are designated as black boxes. Untranslated 5'-UTR and 3'-UTR of parts of the exons 1 and14 are white. The whole exon 15 – the polyadenilic acid (polyA) is shown as white box. Ordinal numbers of exons and introns (I) are shown below the corresponding gene components. The number (No) of an intron (I, shown as arrow) is the same as the exon that it follows.

Table 2. Numerical parameters of the *Arabidopsis thaliana* beta-carbonic anhydrase-2 gene coding part with introns (CA-2, TAIR: AT5G14740). Designations see Table 1. Parameters which can be expressed as multiples of Q = 5nt are framed, e.g., introns (I) No 3, 4 and 5 dimensions (nt): 135 = 5 × 27; 115 = 5 × 23; 125 = 5 × 5 × 5. The sum total of introns 1-9 is 3130 (5 × 626)

E/I, No	E (nt)	C/phase	I (nt)
1	87	87/0	752
2	118	205/1	1311
3	64	289/2	135
4	150	419/2	115
5	49	468/0	125
6	117	595/1	147
7	54	639/0	257
8	86	725/2	137
9	110	835/1	151
10	158	Σ_E993/-	Σ_I1-9
			3130

the sum of exon (E) dimensions 1-14 ($\Sigma E_{1\text{-}14}$ = 1827, Table 1) reveals the parameter $\Sigma_{(I+E)}$ = 33.600 nt, which corresponds to the whole number of codons (33 600 : 3 = 11.200) as well as the whole number of heptanucleotide repeats (33 600 : 7=4800) supporting our new hypothetical model of gene origin.

Another good idea raised by study of the bovine albumin amino acid sequence came from James Brown publication (1976) indicating that albumin-fetoprotein genes had evolved by number of internal duplications, i.e., the bovine albumine sequence could be divided into three equivalent regions of about 190 amino acids (aa) each. The sequences of two of these were more similar than either was to the third, suggesting that there had been a doubling at one point, from a molecule of about 190 aa to one of 380 aa, and then a second, incomplete, duplication that gave rise to the existing 580 aa-residue structure. There was also evidence that the fundamental macrodomain structure (ca. 190 residues) had itself evolved as the result of internal duplications from a more primitive sequence of about 77 residues (corresponding to the bovine aa sequence ~ 504-581; Brown 1976). This primitive sequence in accordance to the Brown hypothesis arose by duplication and separation of a gene segment encoding C-terminal fragment of very primitive globin family proteins similar to myoglobin and hemoglobin.

If the Brown hypothesis is correct then the primary RU of globin proteins must be 2 and 1/3 aa, i.e. 7nt. Data of our analysis, however, suggest another value – 21nt or 7aa (Chipens et al., 2006; Ievina, Chipens 2004). Therefore we reinvestigated this question once more (see chapter *Globin gene introns retained the primeval gene heptanucleotide repeats, their dimensions can be quantized*).

The superfamily of carbonic anhydrases

Another family of genes confirming the small sizes of oligonucleotide repeats are enzymes – carbonic anhydrases (CA) with a Q value 5nt. As an example can serve beta CA2 of the *Arabidopsis thaliana* (GenBank NC_003076.4, Gene ID: 831326, TAIR: AT5G14740,

NP_001031883, Table 2). The CA-family enzymes are extremely widely distributed among living organisms (Liljas, Laurberg 2000), e.g., the human genome contains 35, and the mouse – 38 different members of the CA superfamily, see for examples the PANTHER (Mi et al. 2007) classification system. Particular attention deserve firstly the revealed observations: (i) dimensions of three neighbour introns (3,4 and 5; Table 2) are multiples of 5nt and have the same internal regularity as exons 4 and 9 and many intron coordinates: 2,6,8,9, and (ii) the sum total Σ of nine introns (1-9; Table 2) can be precisely quantized (1130 : 5 = 226) once more confirming that introns were formed of the same size repeats as exons; (iii) intron drift changed the individual dimensions of introns, but the sum total in separate cases remained the same and the same internal regularity was retained for some introns – the "molecular relics".

Five nucleotides (Q = 5nt) can encode only 1⅔ of amino acids therefore most exons do not contain a whole number of codons (n × 3nt) nor a whole number of pentapeptides (n × 5nt). Not all numbers can be divided with 3 or 5nt without remainder. This is possible only if the exon dimension is a multiple of 15 (3 × 5, e.g., the exon 4; Table 2). Contrary to these arithmetical restrictions 40 % of intron coordinates C (as the sum of preceeding exons length) are multiples of pentanucleotides, i.e. the changes of E dimensions (the result of introns drift or nt insertions and deletions) compensate each other. The sum total Σ_{E+I} 993 + 3130 = 4123nt with deviation „minus" 2nt (as a result of „arithmetical restrictions") correspond to a whole number of codons (n × 3) and a whole number of pentanucleotide (n × 5) repeats : 4123 + 2 = 4125, but 4125 : 15 = 275.

Different classes (alpha, beta, gamma) of carbonic anhidrases do not possess related amino acid sequences. Completely different are also their tertiary and quarternary structures. Only their active sites show essential features of remarkable similarity (Liljas, Laurberg, 2000). According to our viewpoint pentanucleotide multiplication reactions took place before the origin of the genetic code (similarly as the heptanucleotide multiples of albumins) in accordance to the concepts of the fancyful interpretation of early evolution. The underlying assumption of this is that some contemporary processes and molecules had to appear before others and that the evolution of the information processing system involved interaction between separately evolving components (Doolittle, Brown 1994). Possibly a lot of pentanucleotide multiples without introns at first originated forming different gene nucleotide sequences without definite functions. Therefore, only after the emergence of a primitive genetic code (and splicing machinery), very different amino acid sequences were formed. According to this hypothesis introns and exons are indeed very old components of primeval genes whose precursors were formed intramolecularly and simultaneously from a simple structure regular and periodic nucleic acids. Otherwize, according to our viewpoint it is very difficult to explain the evolution of the genetic translational code within about 20 million of years in the preIsuan era about 4000 million years ago (Davis 1999). Thus, the genetic code was established soon after the formation of the Earth.

Globin gene introns retained the primeval gene heptanucleotide repeats, their dimensions can be quantized

Graphical autoscanning of the human globine primary stuctures, e.g., the human hemoglobin subunit HBA2 N-terminal amino acid sequence encoded by the exon 1 (GenBank NC_000016) strongly confirm our previously obtained data (Chipens, Ievina

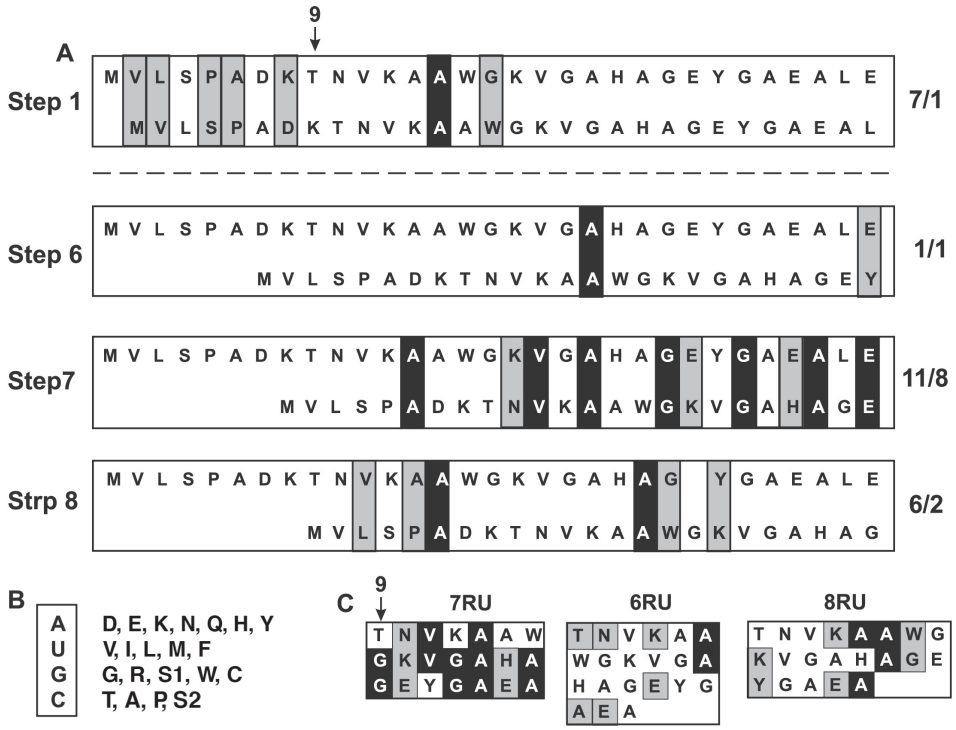

Fig. 2. Methods of analysis of the protein amino acid and the corresponding exon codon root structures using CAACRA methodology. The object of analysis – the human alpha-globin HBA2 exon 1 encoded protein fragment 1-31 (GenBank accession NC_000016). A. The "dynamic" peptide chain (the bottom line) is moved step by step (symbol-by-symbol) relatively to the idenical "static" (the top) chain. The mutual positions of separate steps (1, 6, 7, and 8) are shown. At each step of analysis in front opposite standing symbols (amino acids and corresponding codon root – the second codon letter symbols) are compared, counted and shown at the left side as a fraction of common root amino acids/identical amino acids designated as "factor" or "fraction" (f). In lines identical amino acid pairs are shown against a black background (note that at the same time they have also identical codon roots). Common root but not identical aa are shown against a grey background. B. Glossary of common-root amino acids. C. Analysis of the regularity of amino acid and codon root sequences using the method of repeat unit piles (Ievina, Chipens 2004). Amino acid symbols having an identical codon root (the second codon letter, as well as identical aa number having the same codon root) are enumerated. The highest number of identical symbols in the vertical lines (in this case the "fraction" f is expressed in percent) were when the RUP dimension corresponded to the RU size (7RU). The highest regularity is shown by the HBA2 fragment 9-29. Changes of RU size (6RU or 8RU) significantly decreased the f values (Fig. 3).

2004; Ievina, Chipens 2004) – the RU of the globin peptide chains is 7aa (or on the level of gene – 21nt; Fig. 2). However, these data do not mean that the Brown (1976) hypothesis is not correct, because according to our data of contemporary globin gene intron analysis, RU of the primeval globin genes most likely were 7nt. We find "molecular relics" – the heptanucleotide repeats (only by size, but not by sequences) in the different globin gene intron structures. For analysis we choose globin genes such as the green alga *Chlamydomonas eugametos* hemoglobins (Li637 and Li410) from the protozoan/

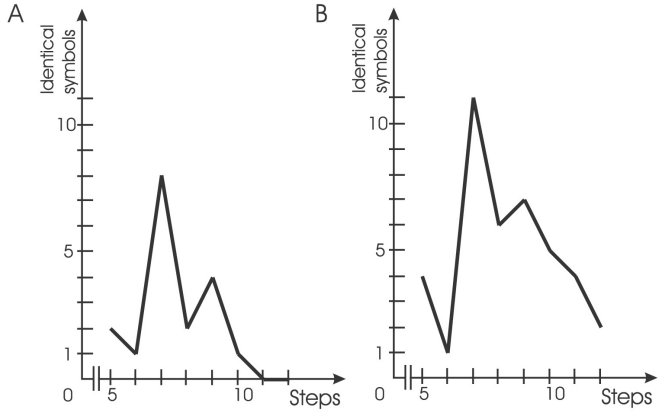

Fig. 3. A fragment of autoscanning curves of the human HBA2 amino acid sequence 1-31, (block A) and the codon root sequence of the corresponding gene exon 1. (block B). Only a fragment of the curve is shown reflecting autoscanning steps 5-12. For the CAACRA and autoscanning principles see also Fig. 2. The maximums of curves correspond to the size of HBA protein RU (7aa, A) or 7 codon roots of the HBA globin gene exons row (B).

Table 3. Intron dimensions of the globin gene family. *Deviations (Δ) are calculated relatively to the nearest GKP, e.g., for HbA-1 $\sum_i$ is 266; 266 : 7 = 38 and Δ is 0, but 266 : 21=12,66; 13 × 21 = 273 and Δ (relatively to $\sum_i$) is 7nt

Globin GenBank accession	**Intron dimensions (nt)**				**Internal regularity of the sum total $\sum_i$ and deviations (Δ, nt)***	
	No 1	**No 2**	**No 3**	**Sum total**	**Number of 7-nt repeats**	**Number of 21-nt repeats**
Homo sapiens myoglobin NC_000022	5925	6056	3440	33327	4761, Δ 0	1587, Δ 0
Homo sapiens HbA-1 NC_000016	117	149	–	266	38, Δ0	273, Δ -7
Homo sapiens HbA-2 NC_000016	117	142	–	259	37, Δ0	252, Δ +7
Homo sapiens HbB NC_000011	130	850	–	980	140, Δ0	987, Δ-7
Chlamidomonas eugametos X65870	181	296	755	1232	176, Δ0	1239, Δ -7
Glycine max U47143	101	114	240	455	65, Δ0	462, Δ-7
Ciona intestinalis AJ548500	254	123	980	1357	194, Δ+1	1365, Δ-8
Ciona intestinalis AJ548501	740	304	–	1044	149, Δ-1	1050, Δ-6

cianobacterial/globin family (Couture et al. 1994), the plant nonsymbiotic hemoglobin gene of the *Glycine max* and the *Ciona intestinalis* (class *Ascidiacea*) globins which are phylogenetically positioned at the base of vertebrates (Ebner et al. 2003). The intron structures in the globin genes family were more conservative than exons – as we supposed they retained and reflect the RU dimensions (7nt or bp) at the first step of multiplication reactions generating globin (and possibly also albumin) gene precursors. Interestingly that not only old globins, but also the sum total of the human hemoglobin gene introns are multiples of 7nt. Independently of number of introns the sum $\sum_i$ of globins can be quantized using Q value 7nt (but not Q = 21, with the exception of myoglobin, Table 3). The human globins A1 and A2 have completely identical amino acid sequences, differing only in their intron No 2 dimension – this difference is precisely 7nt (Table 3) and seven neighbour nucleotides form a cluster before the specific nucleotide sequence of the right splice sight.

Our latest (unpublished) data of analysis of genes encoding several complex enzymes reveal their highly discrete numerical parameters containing segments of alternating neighbour exon-intron-exon-intron sequences formed of identical in size oligonucleotide RU. Biochemical unity of life and the principle biochemical universality (Dayhoff et al. 1972) allow to postulate that the "third way" of gene and intron origin described above most likely is universal for living matter in the early stage of evolution. This, of course, does not exclude other different ways of gene and intron origin and development in later stages of evolution. But according to Susumu Ohno nothing in evolution is created *de novo* – each new gene must have arisen from an already existing gene. This idea is the main motif through the famous monograph "Evolution by gene duplication" (Ohno 1970). He was also the first who postulated that early genes were oligomeric repeats (Ohno 1987).

References

Abate T. 2000. Genome discovery shocks scientists. http://www.euchromatin.org/Abate01.htm

Brown J.R. 1976. Structural origins of mammalian albumin. *Fed. Proc.* 35: 2141–2144.

Chipens G.I., Krikis A., Polevaja L.K. 1979. Physico-chemical principles of information transfer at molecular level. In: Vassileva-Popova J.G., Jensen E.K (eds) *Biophysical and Biochemical Information Transfer in Recognition*. Plenum Press, New York-London, pp. 23–48.

Chipens G. 1980. Using some principles of system analysis for investigation of peptide ligand structures and functions. In: *Structure and Function of Low Molecular Peptides*. Publishing House "Science", Riga, pp. 1–124. (in Russian)

Chipens G., Gnilomedova L.E., Ievina N., Rudzish R., Skliarova S. 1988. Equifunctionality of amino acids, the principle of signature and symmetry of the genetic code. *Proc. Latvian Acad. Sci.* 11/496: 113–116. (in Russian)

Chipens G.I. 1991. The hidden symmetry of the genetic code and rules of amino acid interaction. *Bioorgan. Khim*. 17: 1335–1346. (in Russian)

Chipens G., Ievina N. 1994. Comparative amino acid codon root analysis (CAACRA) of peptide chains. *Proc. Latv. Acad. Sci. Sect. B* 48: 50–54.

Chipens G. 1996. The second half of the genetic code. *Proc. Latv. Acad. Sci. Sect. B* 50: 151–172.

Chipens G., Ievina N. 1999a. Outlines of a nucleotide-multiplication theory of exon and intron origin. *Proc. Latvian Acad. Sci. Sect. B* 53: 65–72.

Chipens G., Ievina N. 1999b. Repeat units are the basic elements of gene and protein structural organisation. *Proc. Latvian Acad. Sci. Sect. B* 53: 54–56.

Chipens G., Ievina N. 2004. Peculiarities of the rotational symmetry of the genetic code two-

dimensional structure. *Latvian J. Chem.* 2004/1: 85–87.
Chipens G., Ievina N. 2005. Connectedness groups of codons/anticodons and origin of the genetic translational code. *Latvian J. Chem.* 2005/3: 282–290.
Chipens G., Ievina N., Kalvinsh I. 2005. A new theory of gene origin and quantisation of aspartate aminotransferase paeameters: mathematical modeling of modern gene structures. *Latvian J. Chem*. 2005/4: 311–324.
Chipens G. 2006. Symmetry and antisymmetry of the genetic code: models of the sense/antisense and the codon root code. *Latvian J. Chem.* 2006/1: 3–18.
Chipens G. Ievina N., Liepina I., Kalvinsh I. 2006. Autoscanning of codon root palindroms. *Latvian J. Chem.* 2006/4: 382–392.
Couture M., Chamberland H., St-Pierre B., Lafontaine J., Guertin M. 1994. Nuclear genes encoding chloroplast hemoglobins in the unicellular green alga *Chlamidomonas eugametos*. *Mol. Gen. Genet*. 243: 185–197.
Davis B.K. 1999. Evolution of the genetic code. *Progr. Biophys. Mol. Biol.* 72: 157–243.
Dayhoff M.O., Eck R.V., Park C.M. 1972. Model of evolutionary change in proteins. In: *Atlas of Protein Sequence and Structure* Vol. 5, Natl. Biomed. Res. Foundation, Washington, pp. 89–99.
Doolitle W.F., Brown J.R. 1994. Tempo, mode, the progenote, and the universal root. *Proc. Natl. Acad. Sci. USA* 91: 6721–6728.
Ebner B., Burmester T., Hankeln T. 2003. Globin genes are present in *Ciona intestinalis*. *Mol. Biol. Evol.* 20: 1521–1525.
Fedorow A., Roy S., Fedorova L., Gilbert W. 2003. Mystery of intron gain. *Genome Res.* 13: 2236–2241.
Gilbert W. 1987. The exon theory of genes. *Cold Spring Harbor Symp. Quant. Biol*. 52: 901–905.
Ievina N., Chipens G. 2003. A new approach to study the origin of genes and introns. *Acta Univ. Latv.* 662: 67–79.
Ievina N., Chipens G. 2004. Origin of globins and mystery of of myoglobin codon root symmetry. *Acta Univ. Latv.* 676: 97–105.
Ievina N., Chipens G., Kalvinsh I. 2006. Internal regularity and quantisation of gene parameters. *Acta Univ. Latv.* 710: 139–153.
Koonin E.V. 2006. The origin of introns and their role in eukaryogenesis: a compromise solution of the introns-early versus introns-late debate. *Biol. Direct.* 14: 22.
Liljas A., Laurberg M. 2000. A wheel invented three times. The molecular structures of the three carbonic anhydrases. *EMBO Reports* 1: 16–17.
Logsdon J.M. Jr. 1998. The recent origins of splicesomal introns revisited. *Curr. Opin. Genet. Dev.* 8: 637–648.
Mi H., Guo N., Kejerival A., Thomas P.D. 2007. PANTHER version 6: protein sequences and functions with expanded representation of biological pathways. *Nucleic Acid. Res.* 35: D247–D252.
Nilsen T.W. 2003. The splicesome: the most complex macromolecular machine in the cell? *Bioessays* 25: 1147–1149.
Ohno S. 1970. *Evolution by Gene Duplication*. Springer-Verlag, Berlin, Heidelberg. 160 p.
Ohno S. 1987. Early genes that were oligomeric repeats generated a number of divergent domains on their own. *Proc. Natl. Acad. Sci. USA* 84: 6486–6490.
Quastler H. 1965. General principles of systems analysis. In: Watterman T.H., Morowitz H.I. (eds) *Theoretical and Mathematical Biology*. Blaisdell Publ. Comp., New York, pp. 313–333.
Rogozin I.B., Sverdlov A.B., Babenko V.N., Koonin E.V. 2005. Analysis of evolution of exon-intron structure of eukariotic genes. *Brief. Bioinform.* 6: 118–134.
Roy S.W., Gilbert W. 2005. Complex early genes. *Proc. Natl. Acad. Sci. USA* 102: 1986–1991.
Roy S.W., Gilbert W. 2006. The evolution of splicesomal introns: patterns, puzzles and progress. *Nature Rev. Genet.* 7: 211–221.
Waterston R.H., et al. 2002. Initial sequencing and comparative analysis of mouse genome. *Nature* 420: 520–562.

Gēnu izveidošanās alternatīvais modelis: intronu dimensiju kvantēšana

Gunārs Čipēns, Nora Ieviņa*, Ivars Kalviņš

Latvijas Organiskās sintēzes institūts, Aizkraukles 21, Rīga LV-1006, Latvija
*Korespondējošais autors, E-pasts: nora.ievina@osi.lv

Kopsavilkums

Intronu problēmas risināšanai ir izstrādāta jauna teorija, principiāli atšķirīga no līdz šim zināmajām. Kvantitatīva gēnu parametru analīze pierāda, ka gēnu, eksonu un intronu koordinātēm piemīt iekšēja identiska regularitāte, kas liecina, ka introni ir veidojušies iekšmolekulāri gēnu priekšteču attīstības gaitā. Mūsdienu gēnu regulāro segmentu skaitliskās vērtības var izteikt ar veselu skaitļu reizinājumu ar jaunizveidotu parametru, kas nosaukts par gēna kvantu un kas izsaka bāzu pāru (nukleotīdu) skaitu gēnu polinukleotīdu virknes atkārtojuma vienībās. Iegūtie dati pierāda, ka gēnu priekšteči kā arī pirmgēni ir bijuši regulāras un periodiskas nukleīnskābes ar vienkāršu struktūru. Pirmgēnu regularitāti ir saglabājuši arī mūsdienu gēnu segmenti („molekulārie relikti") vai pat veselas kodējošo gēnu struktūras. Darbā analizēti peles tubulīns α-1A un un seruma albumīns, *Arabidopis thaliana* β-karbonilanhidrāzes-2, globīnu un hemoglobīnu ģimenes gēni.