Acta Universitatis Latviensis, 2003, Vol. 662, pp. 67–79

A new approach to study the origin of genes and introns

Nora Ievina*, Gunars Chipens

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

Abstract

Eukariotic genomes have two main structural components – different types of repetitive nucleic acids and unique, at first glance nonrepetitive sequences of gene coding parts. A new methodology of sequence analysis based on the structure of the second genetic code can be used to reveal molecular relics in protein and gene structures of tubulins and small G proteins. These are sequences formed of repeat units having identical regularity, as well as ancestral immobile introns distributed in the exon row with the same regularity. A new theory is advanced explaining exon and intron origin from common precursors – highly repetitive simple structure nucleic acids – during the early periods of evolution.

Key words: Multiplication of nucleotides, old ancestral introns, origin of introns, repeat units of genes and proteins, small G proteins, tubulins.

Introduction

During the last decades mainly two main concepts – the exon theory of genes (Gilbert 1987) and the insertional theory of intron origin (Stoltzfus et al. 1994) – have been used to find an answer to the question, whether introns were the media of gene formation by exon shuffling or whether they were inserted later. The problem, however, is not yet solved (Logsdon 1998).

The model of the second genetic amino acid interaction code (or the codon root code; Chipens 1991) has given a posibility to elaborate new methodologies for studies of gene and intron emergence mechanisms, based on comparative amino acid codon root analysis (CAACRA). Codon roots – the second codon letters – are much more conservative and less changed during evolution than amino acids. Twenty natural amino acids determined by the genetic code can be subdivided into four groups of the so called common-root amino acids having identical second codon letters C, G, A and U(T). Natural selection accepts amino acid exchanges in proteins (as a result of point mutations) mainly between the common-root amino acids, indicating that such amino acids are potentially tantamount (Chipens 1991). During evolution, as a result of mutations, amino acids in protein structures may change time and time again, maintaining in many cases the same codon root.

Translating gene exon row nucleotide sequences or protein amino acid sequences

to more conservative codon root sequences in separate cases makes it possible to demonstrate the retained ancestral regularity of genes and proteins. For this purpose several new methods of analysis are useful, such as, autoscanning of protein amino acid and codon root sequences, design of repeat unit piles, calculation and analysis of regularity of intron position co-ordinates and others described below.

The general principle of the advanced model of gene and consequential protein formation is that simple structure nucleotides in early period of evolution were spontaneously saltatory multiplied laterally to generate highly regular polynucleotides consisting of a large number of identical copies termed repeat units (RU). Then, RU diverged in sequences as mutations accumulated in them. At some subsequent time, a group of primary RU from the formed polynucleotide chain could be taken for another (the second step) saltatory multiplication, etc. This model is not principially new – it has been used to study the origin of satellite DNA (Southern 1975). We supplemented this model with the following new theses: (i) exons and introns arose from the same RU-multimer, (ii) formation of exons and introns was induced mainly by emergence and action of the very first splicing machinery, and (iii) splicing sites of introns have been encoded in structures of RU precursors.

Introns evolved much more rapidly than exons, notwithstanding the same rate of mutations in both. Mutations from exons were removed partially by adverse selection as well as by lethal mutations, while introns in the absence of the constraints imposed by the coding function during billions of years accumulated different mutations without any limits. Contrary to these mutations exon structures are under strong control of natural selection, which in accordance with the codon root code (Chipens 1991) accept smainly "symmetric" mutations leading to exchanges between common-root amino acids, which have identical codon roots and are located in the 2D-structure of the genetic translational code symmetrically (Chipens 1991).

Considering that periodic nucleic acids are ancesters of modern genes, the emergence of exons and introns may be easily imagined making only one essential inference. Independently of the biochemical mechanisms of the very first splicing machinery, the splicing sites can be determined only by definite nucleotide structures in the polynucleotide chain – the RU-multimer. Such sites can arise spontaneously by mutations, or alternatively they may be already acidentally encoded in structures of RU precursors. It is well known that nucleotide multiplication reactions form high molecular mass products, e.g. mouse sattelite DNA contains 105 – 108 repeat units (Southern 1975), which not undergoing the splicing after translation can form only giant protein molecules unfit for protein evolution. Evidently the driving force for the evolution of the splicing mashinery is first of all a necessity to diminish the lengh of gene ancestor coding parts. Accidental rare mutations forming splicing site structures, as we suppose, can not have effectively diminished the dimensions of gene ancestors. It more likely seems that splicing sites have been encoded in the RU precursor structures. From this key inference follows a chain of logical conclusions, which create the fundamentals of a new nucleotide multiplication theory of exon and intron origin.

If the splicing site structures are encoded in the nucleotide sequence of a RU precursor then: (i) these sites after the first and the following steps of multiplication reactions are distributed along the nucleotide chain regularly; (ii) intron positions in gene ancestors are regular; (iii) this regularity is the same as the regularity of identical nucleotides or identical

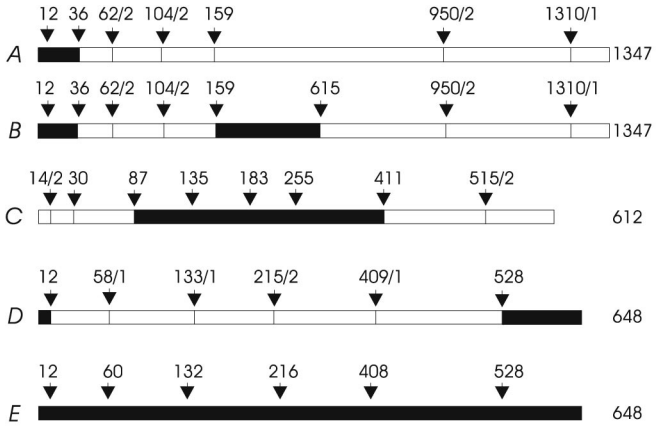

Fig. 1. Regularity of dimensions of intron and exon maps demonstrate the mechanisms of gene emergence by nucleotide multiplication reactions. A and B, intron maps of ß-tubulin genes of the *Aspergillus parasiticus* and *Aspergillus nidulans*. Intron positions are shown as arrows topped with intron co-ordinates/phases (only in cases when the phase differs from zero). Exons containing the whole number of repeat units (nx12nt) are shown as black parts of ribbons. The length of genes coding parts (exon rows, including a stop codon) is shown beside the maps. C, intron map of the green alga *Volvox carteri* gene yptV1. D and E, intron maps of the *Coprinus cinereus* ras gene before and after restoration of intron phases to phase zero (i.e. by changes of intron co-ordinates by ±1 or ±2nt).

amino acids in gene or protein structures; (iv) the birth-positions of introns are strongly determined by the size of RU – introns can be located only between RU (micro-exons) in gene "knot" points. The knot points are situated regularly along the nucleotide chain and denote borderlines of RU; (v) the reactions of nucleotide multiplications determine the formation of long open reading frames of genes with symmetric exons and all introns in a phase-0. Thus, the dominance of symmetric (0,0) exons in natural gene structures (Long et al. 1995) first of all is a signature of gene formation by nucleotide multiplication reactions. And finally (vi), coding parts of gene ancestors have been formed of a whole number of RU. From the essence of the multiplication model, which as a rule postulates formation of only symmetric (0,0) exons follows that during evolution introns can slide and be gradually eliminated, as introns of natural genes in many cases are outside of the gene knot points and have changed phases (Fig. 1).

A stable fundament for the new theory of genes is the possibility to calculate theoretical sizes of exons and an exon row, as well as the potential intron positions in genes, if the size of the repeat unit is known. This allows to compare the calculated and the natural parameters of gene and protein structures and to demonstrate that in many cases contemporary genes have retained some or several introns in the birth positions (ancestral or old immobile introns, OII). The regularity of OII locations is identical with the sequence regularity of exons and exon-coded protein fragments, which can be demonstrated after translation of gene and protein structures into codon-root (second codon letter) symbols and the design of repeat unit piles, or by autoscanning analysis.

The model and nucleotide multiplication theory of exon and intron origin (both further referred to shortly as "the Model") is illustrated by analysis of the tubulin and the small G protein gene families.

Materials and methods

Design and characterization of repeat unit piles

To design a repeat unit pile (RUP), peptide chains and their translations into codon root sequences are cut into fragments corresponding to the repeat unit size. For tabular analysis these RU are laid out horizontally in stacks to form a pile of RU. Immediately after gene formation by multiplication reaction, all the RU in the RUP structure would have identical sequences and the vertical lines (columns) of the RUP would be formed from identical symbols. During the evolution this is disrupted by mutations. The regularity of the repeat unit pile (RUP) structures may be characterized by expressing as a percentage ratio (*f*) of common-root (CR) and identical (I) amino acids, i.e., *f*=CR/I. Similarly, the RUP of gene codon root sequences may be characterized by the number of the preserved nucleotide base structures (also expressed as a percentage). The codon root symbols of gene nucleotide sequences and the corresponding amino acid sequences are identical, therefore identical are also their *f* values, but the amino acid sequence has another *f* value characterising the preserved identity of amino acid symbols. Thus, the gene RUPs are characterized by a single *f* value, but the protein RUPs – by two *f* values as a fraction (the preserved root identity/amino acid identity).

Autoscanning analysis

The protein amino acid and gene exon nucleotide sequence is translated into a sequence of the corresponding codon roots (second codon letters) and moved alongside the sequence duplicate step-by-step (symbol by symbol). The overlapping identical symbols (amino acids and/or codon roots) are counted and registered graphically at each step (a computer-assisted analysis).

For analysis of tubulin structures we used data banks (Dibb, Newman 1989; Liaud et al. 1992) containing information on 109 intron positions and phases in 38 α- and β-tubulin genes, and for analysis of 50 intron positions of SGP – a similar data bank of G-proteins (Dietmaier, Fabry 1994) that included representatives of the following subfamilies: Ras, Rho, Rab/Ypt, Ran/TC4 and Art. The main attention in accordance with the Model was paid mainly to the regularity of intron positions. To analyse intron regularity it is necessary first of all to transform intron position and phase symbols to intron position co-ordinates. By this term we denote the ordinal number of nucleotides of the gene exon row just before the introns (Table 1).

Results and discussion

Regularity of tubulin intron positions

The family of tubulins is composed of highly conserved proteins, which are the principle structural and functional components of eukaryotic microtubules. Previous studies of tubulin family genes have led to different and conflicting conclusions, e.g., Dibb and Newman (1989) consider that intron distribution patterns in tubulin genes could be

Table 1. Revealed (Liaud et al. 1992) and calculated intron co-ordinates of the α- and β-tubulin families. Numerical values of intron co-ordinates marked by asterisks can be expressed as multiples of 12nt

N	Intron, position/phase and co-ordinate (nt)	Nearest gene knot point and intron deviation (nt)	N	Intron, position/phase and co-ordinate (nt)	Nearest gene knot point and intron deviation (nt)
1	2/0, 3	0/+3	21	90/1, 268	264/+4
2	4/1, 10	12/-2	22	95/1, 283	288/-5
3	5/0, 12*	12/0	23	110/1, 328	324/+4
4	9/0, 24*	24/0	24	126/0, 375	372/+3
5	13/0, 36*	36/0	25	134/0, 399	396/+3
6	16/0, 45	48/-3	26	134/1, 400	396/-4
7	17/0, 48*	48/0	27	177/0, 528*	528/0
8	19/1, 55	60/-5	28	208/0, 621	624/-3
9	20/0, 57	60/-3	29	211/1, 631	636/-5
10	21/2, 62	60/+2	30	224/1, 670	672/-2
11	33/0, 96*	96/0	31	257/0, 768*	768/0
12	35/2, 104	108/-4	32	319/2, 956	960/-4
13	38/2, 113	108/+5	33	327/0, 978	972/+6
14	41/0, 120*	120/0	34	346/2, 1040	1044/-4
15	56/0, 165	168/-3	35	351/1, 1051	1056/-5
16	58/1, 172	168/+4	36	353/0, 1056*	1056/0
17	59/1, 175	168/+5	37	407/1, 1219	1224/-5
18	62/0, 183	180/+3	38	412/1, 1234	1236/-2
19	76/1, 225	228/-3	39	437/0, 1308*	1308/0
20	90/0, 267	264/+3	40	448/2, 1343	1344/-1

understood by intron insertion in proto-splice sites after origin, but Liaud and coworkers (Liaud et al. 1992) are convinced that already primordial tubulin genes were rich in introns in agreement with the exon theory of genes (Gilbert 1987), but intron origin is unknown.

We discovered that a sensitive indicator of the birth position changes of introns was the composition of the prime multipliers of intron coordinates. Introns are regular only when the numerical values of co-ordinates had common prime multipliers, because the prime multipliers characterize an internal regularity of the numbers themselves. The Model postulates that immediately after the gene formation all introns in the exon row were situated regularly. Thus, their co-ordinates had to have common prime multipliers. From this followed also that the co-ordinates of regular introns could be used for prognosis of the size of the gene repeat unit. The potential size of the RU could be calculated also from common prime multipliers of exon length or exon length and intron co-ordinates, because in accordance to the Model, internal regularities of these parameters immediately after gene origin were identical.

Analysis of α- and β-tubulin intron position co-ordinates revealed a large group of introns (10 from 40, 25 %; Table 1) having a common set of prime multipliers, 2x2x3, and as a consequence also common regularity of disposition. These intron co-ordinates could be expressed as multiples of 12nt, e.g. for tubulin intron co-ordinates 12, 48 and 1056nt:

12	=		2x2x3		=	1x12;
48	=	2x2x	2x2x3		=	4x12;
1056	=	2x2x2x	2x2x3	x11	=	88x12.

The co-ordinates of regular introns in Table 1 are marked with asterisks.

The small size of the tubulin repeats, and high density of introns in the 5'-terminal parts of tubulin genes (Fig. 1A, B; Table 1) allowed to suppose that the ancestor of tubuline gene most likely originated in a one step multiplication reaction from a simple precursor – a 12-membered nucleotide designated as 4RU or 12nt/4aa (aa, amino acids). The revealed regular introns evidently had not changed their positions since origin of the gene. They crossed the tubulin gene knot points (situated in the exon row after each 12nt) and could be classified as OII.

The second large group was formed by tubulin introns which had slid off their birth positions, but were still in the zone 3nt around the gene knot points (15 introns from 40, 37.5 %). The co-ordinates of these introns were irregular and no common multipliers could be found. Sliding of an intron from the gene knot point even by one nucleotide radically changes the set of prime multipliers, e.g., in the case of a regular tubulin intron with the co-ordinate 48nt:

48	=	2x2x	2x2x3;
49	=	7x7;	
50	=	2x5x5.	

The number of introns crossing the tubulin gene knot points (10 or 25 %) or located near the knot points (15 or 37.5 %) together formed a large group from the analysed intron positions (25 from 40; 62.5 %), and supported the thesis that in the tubulin gene ancestor introns had been situated regularly.

There presently is no plausible molecular mechanism to account for both frequent intron sliding and the actual patterns of intron distribution in genes. One mechanism proposed by Fink (1987) postulates a normal excision of an intron from pre-mRNA, reverse transcription of the modified pre-mRNA, and homologous recombination of the resulting cDNA with the original gene. An additional event that involves imprecise reinsertion of an excised intron back into the pre-mRNA is also suggested by Martinez et al. (1989). The Fink-Martinez model, if correct, should result in unique gene structures – introns should be concentrated near the 5'end of each gene, because reverse transcription begins at the 3'-poly(A) tract of mRNA, but rarely extends completely to the 5'end, and recombination between the gene and cDNA affects the ends less frequently than the middle. As a consequence, intron sliding should be rarely observed at the ends of genes, especially at the 5'end (Martinez et al. 1989). The *Aspergillus* α-tubulin genes (May et al. 1987) supported this model (Fig. 1A, B).

Repeat unit piles of tubulins

In accordance with the Model, regularity of amino acid symbols in protein primary structures, or codon root symbols in the gene exon row, are identical with the regularity of intron positions of the gene ancestor (including OII in modern gene structures).

Exceptions may be only intron positions of mosaic-like genes formed by exon shuffling. Thus, in primary structures of tubulins the identical symbols of amino acids must have appeared with a periodicity of 4, corresponding to the size of a repeat unit 4RU. However, in the vast majority of cases, amino acid sequences of primary RU after billions of years of protein evolution have diverged beyond recognition, even after translation to the codon root sequences. There are also exceptions – in cases when genes contain molecular relics. By this term we denoted the parts of modern gene and protein structures that may be frozen in time, preserving remnants of ancient events that formed the first living molecules (e.g. structures of primary RU, ancestral intron positions, etc.). For example, fibrillar collagens retain many copies of primary RU GPP (GlyProPro) dispersed alongside the protein amino acid sequence (Huang et al. 1991). Similarly the head domains of cytokeratines (e.g. murine CK-15) have preserved several consecutive primary repeats FGGG (Nozaki et al. 1994; Fig. 2A). We revealed similar structures of the RU-relics also in β-tubulin and tubulin prokaryotic homologue FtsZ (named after the filamenting temperature-sensitive mutant Z; Fig. 2C; Erickson 1995). The structures of these and other repeat-relics allowed to conclude that some (if not all) modern proteins evolved from regular simple structure poly-amino acids containing only a few distinctive amino acid residues, and that many genes containing nearly a full set of codons determining 20 natural amino acids evolved from simple almost monotone codon multimers. The driving force for this type of evolution evidently was a prototropic tautomerism of purine and pyrimidine bases and mispairing of nucleotides during nucleic acid replication reactions (Topal, Fresko 1976). As a result of spontaneous mutations during long evolution, non-canonic nucleotide base

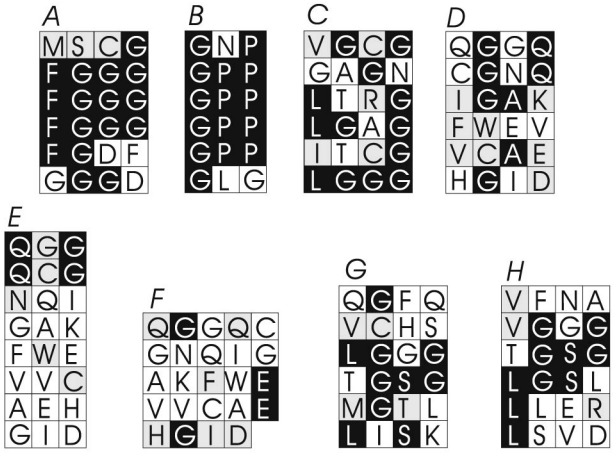

Fig. 2. Repeat unit piles (RUP) of proteins. A, 4RUP of the mouse cytokeratine CK-15 sequence (63-86). Identical symbols in vertical lines are shown against a black backgraund, common-root amino acids are highlighted. B, 3RUP of the human collagen (COL2A1, 158-175). C, three pairs of repeat units 4RU (sequences 43-50, 91-98, and 127-134) from the protein FtsZ (a prokaryotic tubulin homologue). D, 4RUP of the pea ß-1 tubulin, 8-13. E and F, artifically changed sizes of RUP structures (±1 symbol) of the pea ß-1 tubulin diminished the RUP regularity from f = 67/33 (D) to 37/16 (E) and 41/16 (F). G, 4RUP of the pea ß-1 tubulin,131-156. H, 4RUP of the *Arabidopsis thaliana* α-1 tubulin, 137-160.

pairings (such as A-C, G-T, A-A, etc.) could transform polyaminoacids or similar simple structure peptide chains to complex modern protein amino acid sequences.

To study the regularity of protein amino acid sequences we elaborated a new methodology based on CAACRA – design and analysis of repeat unit piles (RUP, see *Materials and methods*). The regularity of RUP could be characterized by the content of dominating isosteric identical and common-root amino acids in vertical lines, using the *f*-factor values. Alongside formation of the original gene sequences (by nucleotide multiplication reactions) during evolution also genes-mosaics were formed by exon shuffling (Gilbert 1987) Therefore, it was necessary to investigate the regularity of all exons and to make definite conclusions of their similarity or distinction. In the case of tubulins this was a difficult task, because the primary RU (micro-exons) of tubulins were small (12nt/4aa), very diverged, and no higher-level regularities (e.g. secondary RU) had been found.

Most of tubulin 4RUPs (formed for analysis of regularity, each containing 6 repeat units and covering the whole amino acid sequence of protein) showed low values of the *f*-factor, which was characteristic for biologically highly specified ("nonregular") sequences. However we noticed that some regions of plant *Pisum sativum* and *Arabidopsis thaliana* tubulins, particularly their N- or C-terminal sequences, had an enlarged content of glycine residues, and that separate tetrapeptide fragments of these regions were formed of hydrophobic U-group amino acid (Chipens 1991) and three glycine residues (e.g., LGGG, IGGG, FGGG, etc.). There were also similar dispersed fragments whose structures had been changed by point mutations and common-root amino acids (e.g., VPGG, VGEG, VGGE, IQGG, etc.), indicating that the potential precursor of tubulin primary RU was related to the RU-relic of the cytokeratine ancestor head domain (FGGG).

In this context the structure of the tubulin prokaryotic homologue was of specific interest – the protein FtsZ had similar primary and 3D-structure with tubulins (Erickson 1995). This protein induces constrictions of the cell wall and cell membranes that leads to the formation of two daughter cells during bacterial cell division. The peptide chain of FtsZ is formed of precisely 91 repeat 4RU (364aa), and in three different regions has retained, as we suppose, ancestor protein repeats LGGG or their mutant forms. The distances between these repeats (calculated by comparison of N-terminal amino acid positions of 4RU, shown in Fig. 2C, in this particular case 91 - 43 = 48 = 4 x 12 and 127 – 91 = 36 = 4 x 9) corresponded precisely to whole number of 4RU multiples. Therefore, the alternative explanation that tetrapeptide sequences had been formed by chance or convergent evolution is not very probable. Most likely, in the early period of evolution, tubulins evolved from RU precursors (containing glycines and U-group amino acids) common to other proteins forming protocell structures. This conclusion was supported by the *f*-factor values of tubulin RUP. The high *f* values were revealed only in cases when repeats forming RUP where rich in glycines (Fig. 2D, G, H). Evidently, only such repeats to some extent reflect the RU precursor structure and as a consequence also the regularity.

Regularity of small G proteins and genes

A family of small GTP-binding or G proteins (SGP) is involved in regulation of very different cellular processes such as signal translation, cytoskeletal organisation, organelle trafic in cells and others (Fabry et al. 1992; Dietmaier, Fabry 1994). Based on amino

Table 2. Revealed (Dietmaier, Farby 1994) and calculated intron co-ordinates of the small G protein genes. Numerical values of intron coordinates marked by asterisk can be expressed as multiples of 12 or 24nt, but marked by two asterisks only as multiples of 12nt

N	Intron, position/phase and co-ordinate (nt)	Nearest gene knot point and intron deviation (nt)	N	Intron, position/phase and co-ordinate (nt)	Nearest gene knot point and intron deviation (nt)
1	6/0, 15	12/+3	26	86/2, 257	252/+5
2	13/1, 37	36/+1	27	88/0, 261	264/-3
3	13/2, 38	36/+2	28	88/1, 262	264/-2
4	25/0, 72*	72/0	29	96/2, 287	288/-1
5	26/0, 75	72/+3	30	97/2, 290	288/+2
6	30/1, 88	84/+4	31	104/0, 309	312/-3
7	34/0, 99	96/+3	32	112/0, 333	336/-3
8	37/1, 109	108/+1	33	114/0, 339	336/+3
9	37/2, 110	108/+2	34	122/0, 363	360//+3
10	38/0, 111	108/+3	35	123/1, 367	372/-5
11	38/1, 112	108/+4	36	125/0, 372**	372/0
12	41/0, 120*	120/0	37	128/1, 382	384/-2
13	46/0, 135	132/+3	38	131/1, 391	396/-5
14	47/1, 139	144/+5	39	132/0, 393	396/-3
15	55/0, 162	156/+6	40	135/1, 403	408/-5
16	56/0, 165	168/-3	41	138/0, 411	408/+3
17	57/0, 168*	168/0	42	138/1, 412	408/+4
18	59/0, 174	172/+2	43	151/0, 450	444/+6
19	63/2, 188	184/+4	44	155/1, 463	468/-5
20	65/2, 194	196/-2	45	162/1, 484	480/+4
21	70/1, 208	204/+4	46	163/0, 486	480/+6
22	77/1, 229	228/+1	47	166/2, 497	492/+5
23	80/0, 237	240/-3	48	171/1, 511	516/-5
24	81/0, 240*	240/0	49	178/0, 531	528/+3
25	86/1, 256	252/+4	50	181/0, 540**	540/0

acid similarities and functions, five different SGP subfamilies are recognized (Ras, Rho, Rab/Ypt, Ran/TC4 and Art). Dietmaier and Fabry (1994) analysed the positions of 125 introns from 28 SGP genes, including representatives of all the tubulin subfamilies, and concluded that most if not all introns in modern SGP genes had arisen by independent insertion events after diversification of the various SGP subfamilies.

We reinvestigated the intron positions of the SGP family genes using the suggested Model and method of analysis, and the same data bank of SGP introns formed by Dietmaier and Farby (1994). About a half of the SGP gene introns were located around the gene knot points in a zone ±3nt (Table 2), confirming the regular intron disposition in a SGP gene ancestor. The green alga *Volvox carteri* gene *yptV1* (Fabry et al. 1992) and

the basidiomycete *Coprinus cinereus* ras gene (Ishibashi, Shishido 1993) can serve as examples of regular organisation of modern SGP. Four exons of the *yptV1* gene have a regular structure and are formed precisely of 4, 4, 6 and 13 repeats (exons 4-7, Fig. 1C):

135 – 87 = 48 (nt); 48 = 4 x 12;
183 –135 = 48 (nt); 48 = 4 x 12;
255 – 183 = 72 (nt); 72 = 6 x 12;
411 – 255 = 156 (nt); 156 = 13 x 12.

The corresponding introns separating these exons all seemed to have "slid off" of the gene knot points (the birth-positions of introns) equally by +3nt, evidently as a result of an indel, because the length of the *yptV1* gene was precisely 51 repeat units (4RU) or 612nt (including the stop codon).

Multiplication reactions of nucleotides, in accordance with the advanced Model, form genes with symmetric exons and all introns in phase-0. Restoration of all introns of the *Coprinus cinereus ras* gene (Fig. 1D) in the birth-positions corresponding to the phase zero (shift by 1-2nt up or down to the nearest gene knot point) revealed a completely regular exon row in which each exon was formed of a whole number of repeats 4RU, but each intron crossed the gene knot points (OII 12, 60, 132, 216, etc.; Fig. 1E). Also in this case the length of gene coding part was formed of a whole number of repeat units (54 x 12nt = 648nt; Fig. 1D, E). Not accidental seemed also the length of the *C. cinereus ras* gene six introns (60, 53, 57, 172, 55 and 61nt), which with one exception all were around 60nt or five repeats (Ishibashi, Shishido 1993).

Visual analysis of long structures of RUP containing many repeats is not handy if large regions of amino acid sequences are investigated. For this purpose we suggested another method termed autoscanning analysis, which was useful for analysis of amino acid as well as codon root sequences. According to this method, a sequence is moved alongside the sequence copy (a duplicate) step-by-step (symbol-by-symbol) and the overlapping identical symbols are counted and registered graphically at each step. When the repeat unit boundaries overlap, the number of overlapping identical symbols is maximal. The scanning graph profiles of regular model sequences resembled a saw and the distances between the teeth of the saw showed the size of the RU.

Natural protein autoscanning graphs, however, in most cases were very complex (as a result of mutations, indels, amino acid sliding, etc.), but there were also exceptions. The autoscanning graph of the green alga *Volvox carteri* Ras protein amino acid sequence encoded by the gene *yptV1* in a small region (steps 38-66) showed maximum positions corresponding to repeat unit sizes 4RU and 8RU (Fig. 3F). The design of Ras protein repeat unit piles (4RUP, 8RUP and 16RUP; Fig. 3) confirmed sequence regularity after 4 as well as 8 symbols. As we supposed, the 8RU was possibly a repeat of the second step multiplication reaction. The *f*-factor values of RUPs were low, but besides this, the density of identical symbols was of special significance for the analysis, e.g., in the 8RUP structure the third column contained four isosteric threonins, but the fourth column – three isosteric isoleucins. There was no possibility of forming such regularities by chance. It is necessary to note that autoscanning analysis of the Ras protein did not show maximum positions which corresponded to the size of 16RU.

The obtained data provided new information about the regular structure of nucleotide and amino acid sequences and is one more step towards understanding intron and gene origin. Does lightening often strike the same place twice? This question was raised during

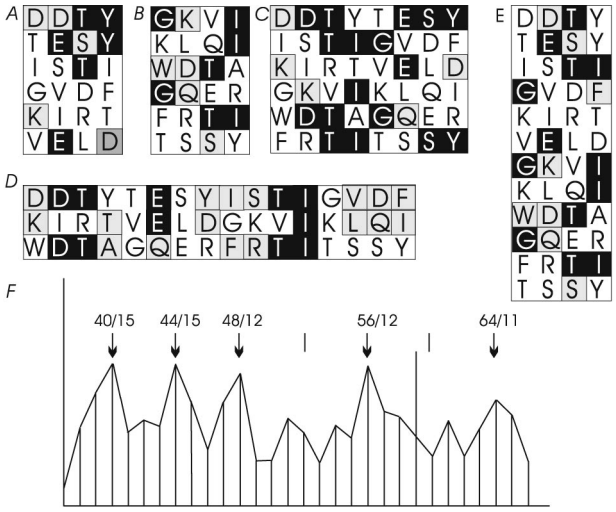

Fig. 3. Analysis of the regularity of the green alga *Volvox carteri* protein SGP encoded by the *yptV1* gene. A-E, repeat unit piles of the protein fragment (sequence 30-77) with different potential sizes of repeats: 4RU, 8RU and 16RU. F, a fragment of an autoscanning graph of the *yptV1* gene encoded amino acid sequence 1-203 (steps 38-66). The bars show numbers of overlapping identical amino acid residues at each step of autoscanning. Maximums are shown by arrows topped with the step number/number of overlapping identical amino acids. Missing maximum positions are shown by vertical dashes.

discussion of origin of the GAPDH (glyceraldehyde-3-phosphate dehydrogenase) gene family introns (Logsdon et al. 1994), and characterizes the absence of strong criteria and parameters of the exon theory of genes and the insertional theory of intron origin that are necessary to study the gene structural organisation and to resolve the intron problem. We suggest such parameters: the co-ordinates of intron location, the sizes of repeat units, exons and gene coding parts as well as the regularity of gene and protein sequences characterized by *f* values of the corresponding repeat unit piles. A stable fundament of a new nucleotide-multiplication theory of exon and intron origin is the possibility to calculate theoretical sizes of exons and an exon row, as well as the potential intron positions in genes if the size of a repeat unit is know. Very important seems also our conclusion that the regularity of ancestral intron (OII) positions is the same as the symbol (amino acids or codon roots) regularity in protein amino acid sequences. Our general conclusion is that introns during evolution have arisen very early, alongside with exons from the same repetitive nucleic acid precursors. Introns are a consequence of gene formation.

Acknowledgements

Grant support No. 01.0171 from the Science Council of Latvia is gratefully acknowledged.

References

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

Dibb N.J., Newman A.J. 1989. Evidence that introns arose at protosplice sites. *EMBO J.* 8: 2015–2021.

Dietmaier W., Fabry S. 1994. Analysis of the introns in genes encoding small G proteins. *Curr. Genet.* 26: 497–505.

Erickson H.P. 1995. FtsZ, a procariote homolog of tubulin. *Cell* 80: 367–370.

Fabry S., Na N., Huber H., Palme K., Jaenicke L., Schmitt R. 1992. The yptV1 gene encodes a small G-protein in the alga *Volvox carteri*: gene structure and properties of the gene product. *Gene* 118: 153–162.

Fink G.R. 1987. Pseudogenes in yeast. *Cell* 49: 5–6.

Gilbert W. 1987. The exon theory of genes. *Cold Spring Harbor Symp. Quant. Biol.* 52: 901–905.

Huang M.-C., Seyer J.M., Thompson J.P., Spinella D.G., Cheah K.S.E., Kang A.H. 1991. Genomic organisation of the human procollagen a1(II) collagen gene. *Eur. J. Biochem.* 195: 593–600.

Ishibashi O., Shishido K. 1993. Nucleotide sequence of a ras gene from the basidomycete *Coprinus cinereus*. *Gene* 125: 233–234.

Liaud M.F., Brinkman H., Cerff R. 1992. The α-tubulin gene family of pea: primary structures, genomic organization and intron-dependent evolution of genes. *Plant Mol. Biol.* 18: 639–651.

Logsdon J.M.,Jr. 1998. The recent origins of splicesomal introns revisited. *Curr. Opin. Genet. Dev.* 8: 637–648.

Logsdon J.M.Jr., Palmer J.D., Stoltzfus A., Cerff R., Martin W., Brinkmann H. 1994. Origin of introns: early or late? *Nature* 369: 526–528.

Long M., Rosenberg C., Gilbert W. 1995. Intron phase correlations and the evolution of the intron/exon structures of genes. *Proc. Natl. Acad. Sci. USA* 92: 12495–12499.

Martinez P., Martin V.F., Cerff R. 1989. Structure, evolution and anaerobic regulation of a nuclear gene encoding cytosolic glyceraldehide-3-phosphate dehydrogenase from maize. *J. Mol. Biol.* 208: 551–565.

May G.S., Tsang M. L.-S., Smith H., Fidel S., Morris N.R. 1987. *Aspergillus nidulans* beta-tubulin genes are unusually divergent. *Gene* 55: 231–243.

Nozaki M., Mori M., Matsushiro A. 1994. The complete sequence of the gene encoded mouse cytokeratin 15. *Gene* 138: 197–200.

Southern E.M. 1975. Long range periodicities in mouse satellite DNA. *J. Mol. Biol.* 94: 51–69.

Stoltzfus A., Spencer D.F., Zuker M., Logsdon J.M.,Jr., Doolitle W.F. 1994. Testing the exon theory of genes: the evidence from protein structure. *Science* 265: 202–207.

Topal M.D., Fresko J.R. 1976. Complementary base pairing and the origin of substitutions mutations. *Nature* 263: 285–289.