Acta Universitatis Latviensis, 2006, Vol. 710, Biology, pp. 155–158

Determination of the human epidermial growth factor precursor *hEGFP* gene repeat unit size by quantization of exon 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

A new method for determination of eventual regularity and periodicity of contemporary gene exon rows is suggested. The essence of this method lies in the common internal regularity of gene parameters measured by the number of nucleotides (nt) – mainly of gene intron coordinates and exon lengths (or dimensions). For this purpose it is necessary to calculate the prime number multipliers of all these gene parameters, to find a set of common ones, and, finally, to calculate the product of the set of revealed common multipliers. The obtained product shows the potential size (nt) of the gene repeat unit (or gene quantum). Here we demonstrate that the potential size of gene repeats can be determined using only exon dimensions. The method is used for *hEGFP*. The first exson of this gene includes a long and unregular 5'-untranslated region. This does not allow to determine the correct reference point for determination of regularity of intron coordinates. The size of the calculated primary repeat unit of *hEGFP* is 21 nt.

Key words: gene quantum, internal regularity of genes, repeat units.

Introduction

Continuing our studies of the origin and structural organization of genes and proteins (Chipens et al. 2005) we analyzed the precursor of human epidermal growth factor (hEGFP). Epidermal growth factor (EGF) is a 53 amino acid polypeptide that has many different biological properties – it is a potent mitogen for cells *in vitro* and stimulates proliferation and differentiation of cells *in vivo* (Carpenter, Cohen 1990). Human EGFP consists of 1207 amino acids (aa), and the corresponding cDNA coding part – of 3621 nucleotides (nt) (Bell et al. 1986). The precursor is processed to EGF in different tissues. The sequence of hEGFP includes not only EGF, but also eight EGF-like units (egf; Fig. 1) and near the carboxyl terminus a hydrophobic sequence characteristic for an integral membrane protein with its NH_2-terminus external to the cell surface (Doolitle et al. 1984).

The sequence of EGF has been reported to be similar to fragments of several blood coagulation factors, LDL-receptor and tumor growth factor (Doolitle et al. 1984; Bell et al. 1986). It is supposed also that tumor growth factor and EGFP arose as a result of common ancestral gene (Doolitle et al. 1984). According to our viewpoint, to investigate the relatedness of these and other bioregulators and to study the evolution of hEGFP structure itself (Bell et al. 1986), first of all it is necessary to know the potential sizes of

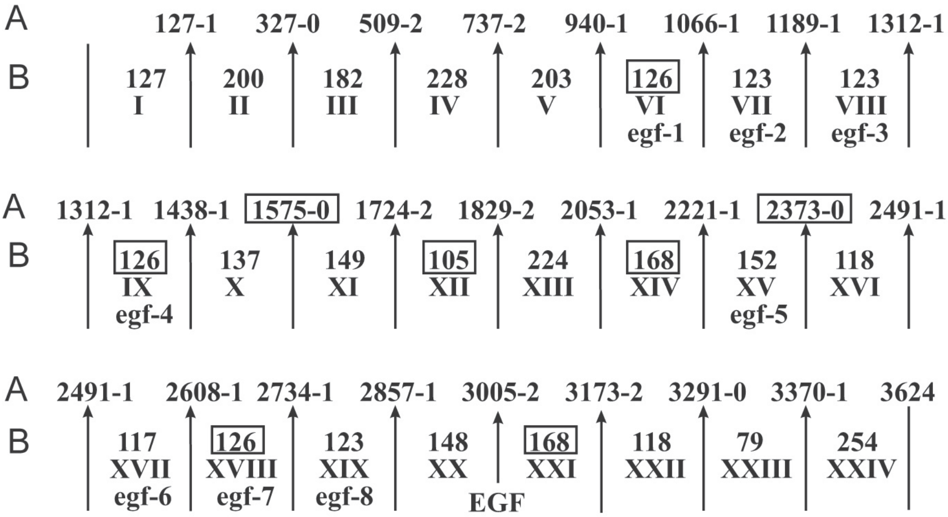

Fig. 1. Scheme of structural organisation and numerical parameters of exon length and intron coordinates of human *EGFP* cDNA, in accordance with the data in the GenBank X04571 and SwissProt P01133. A, arrows topped with intron coordinates (nt) and phases separate exons in the exon row. Intron coordinates are calculated as a sum of preceeding exons length. B, exon length (nt) and ordinal (Roman) numerals of exons. Parameters which can be expressed as multiples of the gene quantum (21 nt) are framed. EGF written in small letters (egf) denote the location of EGF-like amino acid sequences encoded by the given exon.

their repeat units expressed as the number of nt which is termed gene quantum (Ievina, Chipens 2003; Chipens et al. 2005). To quantize means to select a discrete set of values from a continuous range of possibilities.

Methods

The aim of our work was to determine the potential dimensions of the repeat unit (RU) of *hEGFP* gene and protein. For this purpose we use cDNA structure of *EGFP* and a scheme of exon-intron organisation described in the literature (Bell et al 1986). The first exon of the human *EGFP* contains a long unregular 5'-untranslated region about 451 nt (Bell et al. 1986), which due to potential indels in the nucleotide sequence do not allow to determine the exact reference point (the first nucleotide of the whole first exon). As a consequence, this does not allow to calculate correct intron coordinates. Therefore, to determine the potential sizes of *hEGFP* gene quantum and RU we used only the dimensions of exons expressed by number of nucleotides (Fig. 1).

Determination of RU dimension and a gene quantum is based on a model of gene precursor (highly repetitive and periodic nucleic acids) origin by oligonucleotide multiplication reactions (Ievina, Chipens 2003; Chipens et al. 2005). According to this model exons of gene precursors were formed of a whole number of RU, and introns were located between exon and RU boundaries and crossing the gene knot points (the 3'-terminal nucleotides of RU). The gene quantum characterize the size of primary RU of a

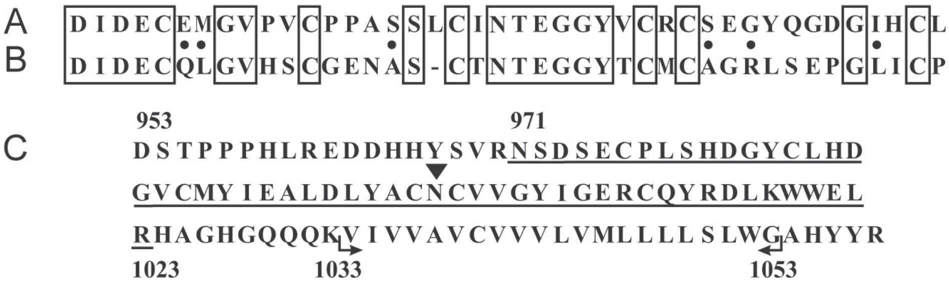
A
D I D E C E M G V P V C P P A S S L C I N T E G G Y V C R C S E G Y Q G D G I H C L
B
D I D E C Q L G V H S C G E N A S - C T N T E G G Y T C M C A G R L S E P G L I C P
953
971
C
D S T P P P H L R E D D H H Y S V R N S D S E C P L S H D G Y C L H D
G V C M Y I E A L D L Y A C N C V V G Y I G E R C Q Y R D L K W W E L
R H A G H G Q Q Q K V I V V A V C V V V L V M L L L L S L W G A H Y Y R
1023
1033
1053

Fig. 2. Two EGP-like amino acid sequences and location of epidermial growth factor (EGF) of the *hEGFP* gene encoded protein. A, EGF-like sequence (amino acids 870-911 encoded by the exon XVIII). B, EGF-like sequence (amino acids 912-952 encoded by the exon XIX). Identical symbols of amino acid residues are framed. Common-root amino acids with identical second codon letters are denoted by dots. C, Amino acid sequence encoded by the human EGFP exons XX and XXI (amino acids 953-1058). EGF structure (971-1023) is underlined. The intron position which crosses the EGF sequence (intron coordinate 2878-1, Fig. 1) is shown by a filled arrowhead. The transmembrane domain (amino acid residues 1033-1053) is marked by broken arrows.

given gene. Both parameters – RU and a gene quantum – can be calculated on the basis of common prime multipliers of exon dimensions and/or intron coordinates. The method of calculations in detail is described in Chipens et al. (2005) and Ievina et al. (2006).

Results and discussion

The human *EGFP* gene consists of 24 exons and 23 introns (Bell et al. 1986). Determination of prime multipliers of exon dimensions allow to select six exons (VI, IX, XII, XIV, XVIII and XXI; Fig. 1) whose dimensions had common prime multipliers 3 × 7. For example 126 = 2 × 3 × 3 × 7 = 6 × 21; 168 = 2 × 2 × 2 × 3 × 7 = 8 × 21; 105 = 3 × 7 × 5 = 5 × 21. Thus, the potential values of the primary RU and the gene quantum are 21 nt or 7 aa. As we suppose, the *EGFP* gene was formed by multiplication of the secondary RU – 126 nt long nucleotide – a hexamer of primary RU. Among the *EGF*-like sequences (egf; Fig. 1), there are also 123-nt-long repeats (exons VII, VIII, XIX), the size of which changed by intron sliding (drift) or one codon deletion.

The supposed secondary RU had homology of amino acid sequences, e.g., exons XVIII (126 nt/42 aa) and XIX (123 nt/41 aa, Fig. 2 A, B). The human epidermial growth factor (EGF, 159 nt/53 aa; Fig. 2 C) was formed during evolution, most likely from two neighbour exons XX and XI (Fig. 1, Fig. 2 C). The *hEGFP* nucleotide sequence contained also several exon dimensions and intron coordinates whose numerical parameters (as a result of intron drift during evolution) differed from those calculated on the basis of *hEGFP* gene quantum multiples by 1-2 nt, e.g., exon I (6 × $\underline{21}$+1), exon XI (7 × $\underline{21}$ + 2), exon XX (7 × $\underline{21}$ + 1), intron 1724-2 (82 × $\underline{21}$ + 2), intron 2857-1 (136 × $\underline{21}$ + 1), intron 3005-2 (143 × $\underline{21}$ + 2), etc., supporting our viewpoint, that introns and exons were formed from one and the same regular precursor having identical size and structure of RU. The exon row length of the human *EGFP* gene, including the nontranslated parts of the 5'-terminal and 3'-terminal exons is 4871 nt (Bell et al. 1986), which differs from the gene quantum multiple (232 × 21 = 4872) only by one nucleotide.

References

Chipens G., Ievina N., Kalvinsh I. 2005. A new theory of gene origin and quantization of aspartate aminotransferase parameters: mathematical modeling of modern gene structures. *Latvian J. Chem.* 4: 311–324.

Carpenter G., Cohen S. 1990. Epidermial growth factor. *J. Biol. Chem.* 265: 7709–7712.

Bell G., Fong N., Stempien M., Wormsted M., Caput D., Ku L., Urdea M., Rall L. Sanchez-Pescador R. 1986. Human epidermial growth factor precursor: cDNA sequence, expression *in vitro* and gene organisation. *Nucleic Acids Res.* 14: 8427–8446.

Doolittle R., Feng D., Johnson H. 1984. Computer-based characterisation of epidermial growth factor precursor. *Nature* 307: 558–560.

Ievina N., Chipens G. 2003. A new approach to study the origin of genes and and introns. *Acta Univ. Latv.* 666: 67–79.

Ievina N., Chipens G., Kalvinsh I. 2006. Internal regularity and quantization of gene parameters. *Acta Univ. Latv.* 710: 139–153.

Cilvēka epidermiāā augšanas faktora priekšteča gēna atkārtojuma vienības izmēru noteikšana kvantējot eksonu dimensijas

Gunārs Čipēns, Nora Ieviņa*, Ivars Kalviņš

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

Kopsavilkums

Mēs esam izstrādājuši jaunu metodi mūsdienu gēnu eksonu rindu iespējamās regularitātes noteikšanai. Metodes būtība ir kopējās iekšējās regularitātes noteikšana gēnu parametriem, galvenokārt intronu koordinātēm un eksonu garumiem (dimensijām), kas izteikti ar nukleotīdu skaitu (nt). Šim nolūkam jāatrod visu šo parametru pirmreizinātāji un no tiem jāatlasa kopējo faktoru kopa. Faktoru reizinājums parāda iespējamo gēna atkārtojuma vienības lielumu (nt). Šeit mēs demonstrējam, ka atkārtojuma vienības lielumu var noteikt, izmantojot vienīgi eksonu dimensijas. Metode ir pielietota cilvēka epidermiālā augšanas faktora priekšteča gēnam. Pirmais šī gēna eksons ietver garu neregulāru 5'-netranslēto rajonu. Tas neļauj izvēlēties pareizu atskaites punktu intronu koordinātu regularitātes noteikšanai. Aprēķinātais gēna pirmējās atkārtojuma vienības lielums ir 21 nt.