STUDIA UNIVERSITATIS
Revistă științifică a Universității de Stat din Moldova, 2007, nr.7

PROTEIN STRUCTURE NOVELTY CREATED VIA INSERTION OF NON-CODING ELEMENTS INTO CODING REGION OF SOME *ARABIDOPSIS* GENES

Andrei SHUTOV, Helmut BÄUMLEIN*, Lothar ALTSCHMIED*

Laboratory of Plant Biochemistry
**Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany*

Prezența unui număr mare de repetări scurte (duplete) este caracteristică genomilor eucariotelor. În genomul *Arabidopsis* am detectat peste 400 gene, în care segmentul regiunii de codificare este repetat în regiunea necodificată. Pentru estimarea abilității potențiale a dupletelor cu crearea noutății în regiunile de codificare a genelor prin încorporarea elementelor din regiunea necodificată am analizat 138 gene de *Arabidopsis* care conțineau duplete, pentru care este cunoscută succesiunea mARN. Analiza evolutivă a acestor gene ne-a permis a urmări istoria formării dupletelor. A fost demonstrat că abilitatea potențială a dupletelor studiate de a crea noutatea în regiunile de codificare a genelor se realizează la membrii a cel puțin două familii de gene. În prezentul articol va fi pusă în discuție contribuția posibilă a dupletelor în evoluția genomilor eucariotelor.

Recently, short paired duplications in mammalian and *Arabidopsis* genomes have been described [1]. The overall frequency of these duplications was shown high enough to suggest that the formation of short sequence copies (doublets) might be a key process in the evolution of protein structure and gene regulation. A potential ability of the doublets to create novelty in protein gene coding regions due to incorporation of non-coding sequence elements was hypothesized [2]. More than 400 genes were detected in the *Arabidopsis* genome, which exemplify presence of precise copies of a sequence element in both coding and non-coding regions. In this paper we got proofs that the duplication of a pre-existed non-coding sequence element into protein coding region created protein structure novelty in members of at least two different *Arabidopsis* gene families. Possible contribution of this kind of duplications in the evolution of eukaryotic genomes is discussed.

Materials and Methods

The *Arabidopsis* genome was chosen because the annotation of protein-coding genes is well advanced (http://mips.gsf.de). Only 138 genes whose models were proved by mRNA sequencing were analyzed among 415 genes in the *Arabidopsis* genome found to contain precise copies (≥20 nucleotides) in their coding/non-coding regions [2].

CLUSTAL program was used for nucleotide and amino acid alignments. Evolutionary trees based on amino acid sequences were constructed using TREECON [3]. Orthologous sequences closely related to the *Arabidopsis* paralogs have been used as outgroups. Most conserved part of sequences aligned with minimum gaps was used for tree construction. In the figures, the numbers along branches refer to bootstrap values (% from 100 replicates). Sequence numbering corresponds to that of unspliced genes in MIPS annotations.

Results and Discussion

The only possibility to describe the history of formation coding/non-coding doublets is to trace back the history of respective genes, each being a member of a family of paralogous genes. At least three members of such a family are expected to be required to show whether a doublet has been formed either A) by insertion of a doublet unit (DU) pre-existed as a coding region motif into non-coding region, or B) by insertion of a DU pre-existed in non-coding region into coding region. These members are ancestral gene A and genes B and C (Fig.1), which reflect gene structure immediately before and after duplication event, respectively. The presence of a DU inside the gene B non-coding region would demonstrate the direction of duplication defined as B in Fig.1. Obviously, the non-coding sequence regions of genes B and C should retain homology as a basis for detection of either absence or presence of the DU in the gene B non-coding region.

The identity of the gene C DUs indicates that the duplication had happened very recently (the non-coding DU was not in time to be modified). Therefore, when the gene B containing a single copy of DU motif is a closest gene C relative, it can be considered to reflect the structure of gene C immediate ancestor. In this case the gene A and evolutionary analysis of a gene family are desirable but not strongly necessary.

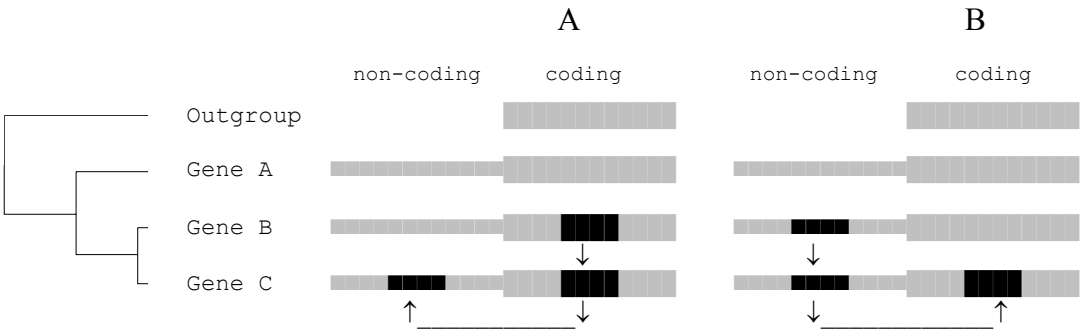


Fig.1. Principle possibility to demonstrate insertion of coding sequence into non-coding region (A), and insertion of a non-coding sequence into coding region (B).

The detection of gene families, which exemplify formation of a novelty in the gene C coding region due to incorporation of DU derived from non-coding region (Fig.1B), meet following difficulties.

1. Only the genes C of known mRNAs can be used for analysis to be sure that their structure is evolutionarily adopted and the gene is actually expressed.

2. The presence of paralogous genes in the *Arabidopsis* genome and the existence of an ortholog as outgroup usually is a prerequisite for analysis. The three members of the *Arabidopsis* family (genes A, B and C, Fig.1) have to be conserved and closely related to their ortholog. This is a prerequisite for reliable alignment and statistically well-supported evolutionary tree.

3. The evolutionary distance between genes B and C should be short. Only in this case the unspliced sequence of the gene B do reflect gene C features before the duplication event. This is the major problem because of usual instability of gene non-coding regions (5'-leader, introns and 3'-trailer).

The entire collection of 138 genes of known mRNAs found containing coding/non-coding doublets have been analysed aiming to detect examples that demonstrate formation of coding region novelty due to insertion of non-coding sequence elements. The obtained results can be divided into three categories.

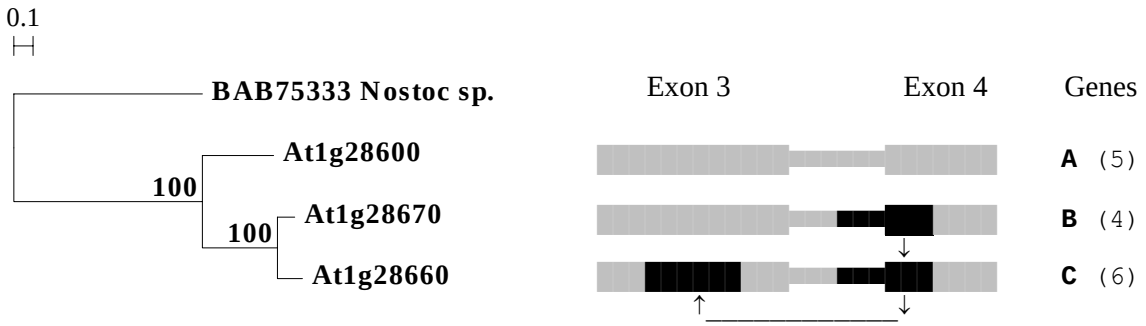
Largest category I: The history of doublet formation baffles description. Sometimes, evolutionary relationships of potential members of a gene family are indefinite, and the history of the gene C cannot be traced back. Although in most cases the relationships between members of the family are well defined, the evolutionary distance between genes B and C (Fig.1) is too long, and their non-coding regions cannot be aligned. Thus, the doublets cannot be characterised.

Large category II. Well-defined history of doublets demonstrates that non-coding sequences do not affect coding regions.

Smallest category III: The DU present in gene C coding region exhibits dissimilarity to relevant region of the genes B and A. Therefore, sequence features of such a coding region specific for only gene C might be resulted from duplication of non-coding region into coding region. In most cases the analysis of the category III cannot be irreproachable because of instability of relevant non-coding regions in genes B and C. Nevertheless, at least two gene families described below (Fig.2,4) convincingly demonstrate formation of certain protein novelty via insertion of a DU pre-existed in non-coding region into a coding region. Naturally, these examples are less simple than that formalized in Fig.1B.

Family of lipases (Fig.2). Conserved 5'-edge of the exon 4 is characteristic of the family of at least 15 *Arabidopsis* genes (including the gene C) coding lipases and related proteins (Fig.3). The exon 3 3'-edge also is conserved in all genes of the family but excluding the gene C in the region where the DU is located. Sequences of intron 3/exon 4 borders of genes C and B are almost identical. In combination, these observations are sufficient to conclude that the intron/exon DU sequence, which pre-existed in gene C before duplication, was copied into exon and thereby created its novelty. Formation of the doublet did not cause extension of the site of insertion. This implies that the originally conserved sequence motif inside exon 4 of the gene C immediate ancestor was substituted with the DU sequence. The structural novelty created by this substitution becomes clearly recognizable when the relevant sequence segment of the gene C is compared with homologous sequences present in *Arabidopsis* and *Oryza* genomes (Fig.3). Most convincing novelty consists in non-conserved substitution of Pro, which is strictly conserved in all other sequences. Possibly, this peculiar substitution became permissible only in combination with other substitutions (and single-gap deletion), although they are less dramatic. If so, the adopted substitution of conserved Pro might be achieved only via duplication event.

A



B

```
                 K  E  V  E  E  L  V  P  F  V  I  A  S  I  S  S  T  I  T                 E  L  I  G  M  G  G
A At1g28600 1501 AAGGAAGTTGAAGAGCTGGTTCCATTTGTAATCGCTTCTATTTCTTCTACAATAACGgtat//tgttgaatccagGAGTTGATTGGTATGGGGGGT 1709
                 N  E  I  K  E  L  V  P  L  I  V  K  A  I  S  S  A  I  V       t  k  l  q  D  L  I  D  L  G  G
B At1g28670 1335 AATGAAATCAAAGAGCTTGTTCCTCTAATCGTCAAAGCTATTTCTTCTGCTATTGTGgtaa//acaaaactacagGATTTGATTGATTTAGGGGGC 1502
                 N  E  T  K  -  L  Q  D  L  I  I  K  A  I  S  S  A  I  V       t  k  l  q  D  L  I  A  L  G  G
C At1g28660 1289 AATGAAACCAAA---CTACAGGATTTGATCATCAAAGCTATTTCTTCTGCAATTGTGgtaa//accaaactacagGATTTGATCGCTTTAGGGGGC 1435
                          inserted DU                                                  pre-existed DU
```

Fig. 2. The gene family of lipases: sequence segment covering intron 3/exon 4 border was copied into exon 3 in the At1g28660 gene. A, evolutionary relationships between genes A, B and C as they specified in Fig. 1 B. In total 292 positions of aligned amino acid sequences coded by the genes A-C and by a gene of bacterial lipase (outgroup) have been used for the tree construction. The tree branches are supplemented by schemes describing the structures of respective genes in the region of exons 3 and 4 (thick rectangles) and intron 3 (thin rectangle). Black rectangles indicate position of the gene C duplication units and homologous (90 % identities) sequence segment in the gene B. The number of mRNAs known for the genes A-C is shown in brackets. B, aligned nucleotide sequences and translated amino acid sequences of exons 3 and 4 (capitals) and intron 3 (low case letters). Underlined characters printed in bold correspond to sequence segments shown schematically as black rectangles in A. Similarity between aligned full-length sequences of introns 3 of the genes B and C (93 positions) exceeds chance level (68.1 % identities).

```
                                                    ⇓
At1g28660 ILMGEIGGNDFFYPSSEGKSINETK-LQDLIIKAISSAIVDLIALGGKTFLVPGGFPAG   Arabidopsis
At1g28670 ILMGEIGGNDYNYPFFEGKSINEIKELVPLIVKAISSAIVDLIDLGGKTFLVPGGFPTG
At1g28650 ILMGEIGGNDYNYPFFEGKSINEIKELVPLIIKAISSAIVDLIDLGGKTFLVPGNFPIG
At1g28640 ILMGEIGVNDYNYPFFEGKSINEIKQLVPLVIKAISSAIVDLIDLGGKTFLVPGNFPLG
At1g28570 ILMGEIGGNDYNYAFFVGKNIEEIKELVPLVIETISSAITELIGMGGKTFLVPGEFPLG
At1g31550 ILMGEIGANDYNFPFFQLRPLDEVKELVPLVISTISSAITELIGMGGRTFLVPGGFPLG
At1g28580 ILMGEIGGNDYNYAFFVDKGIEEIKELMPLVITTISSAITELIGMGGRTFLVPGEFPVG
At1g28600 ILMGEIGGNDYNFPFFNRKPVKEVEELVPFVIASISSTITELIGMGGKTFLVPGEFPIG
At1g27360 ILIGEIGGNDYNFPLFDRKNIEEVKELVPLVITTISSAISELVDMGARTFLVPGNFPLG
At1g28610 IIMGEIGGNDFNFAFFVNKT-SEVKELVPLVITKISSAIVELVDMGGRTFLVPGNFPLG
At1g28590 ILMGEIGGNDYNFALFQRKPVKEVEELVPFVIATISSAITELVCMGGRTFLVPGNFPIG
At5g03980 FMVGEIGGNDYNYGFFQGKPMEEIRSYIPHVVGAITAAAREVIRAGAVNVVVPGNFPVG
At5g45910 FLVGEIGGNDYNYPLLAFRSFKHAMDLVPFVINKIMDVTSALIEEGAMTLIVPGNLPIG
At1g09390 LYMIDIGQNDIADSFSKGLSYSRVVKLIPNVISEIKSAIKILYDEGGRKFWVHNTGPLG
At1g56670 LYMIDIGQNDIARSFARGNSYSQTVKLIPQIITEIKSSIKRLYDEGGRRFWIHNTGPLG
NP_917249 FLVGEIGGNDYNHPLICGVSIRKIRSFTPSVIAEISSTITELIRLGAKTLVVPGNLPIG   Oryza
NP_913345 FFMGEFGGNDYVFLQAAGKTVEQLIPYVPKVVGAISAGIEAVIKEGAVQVVVPGELPNG
NP_913332 FFVGELGWNDYSAVLLAGRGVDEARSLTPRVVGTIRAATQKLIDGGARTVFVSGITPMG
NP_913325 FIVGEFGGNDYNAPLFGGKSMDEVKGYVPQIIAKITSG--TLIGLGAVDIVVPGVMPIG
BAB89190  FLVGEIGGNDYNLPLMSGMSIEKIRNFTPSVIAKISSIITELIGLGAKTLVVPGNIPIG
BAD68794  FLVGEIGGNDYNYPLMSGVPIEKIRSFTPSVIAKISSTITELIGLGAKTLVVPGNLPIG
BAD61220  FLVGEVGGNDYNHLIVRGKSLDELHELVPKVVGTITSAITELINLGAKKLVVPGNFPIG
BAA94228  FLVGEIGGNDYNYAFFKGKSLDDAKSYVPTVAGAVADATERLIKAGAVHLVVPGNLPIG
AAT01388  FVFGEFGGNDYSFAWKAEWSLEKVKTMVPSVVASMAGGIERLLDEGARHVVVPGNLPAG
BAA94220  FFMGEFGGNDYVFLLAAGKTVDEAMSYVPKVVGVISAGVEAVIEEGARYVVVPGQLPTG
BAA94224  FFMGEIGGNDYVFLYAAGKTVDEAMSYVPKVVQAISAGVEAVIKEGARYVVVPGQLPTG
BAD73016  FFMGEFGGNDYVFILAAGKTLEELVPYVPKVVQAISAGIEAVIKEGARYVVVPGELPNG
AAT44175  FVVGEFGGSDYRYLLSGGKSLEQAKSFVPEVVRAICRGVERLVEEGARYVVVTGTPPAG
BAD54230  FVVGEFGGNDYNAPLFSGVAFSEVKTYVPLVAKAIANGVEKLIELGAKDLLVPGVLPIG
AAT44169  FVVGEFGGNDYNAPLFAGRAMTEVRDYVPQVVSKIIRGLETLIRMGAVDVVVPGVLPIG
BAD19158  FVVGEFGGNDYNAPLFAGKGLEEAYKFMPDVIQAISDGIEQLIAEGARELIVPGVMPTG
BAD54227  FVVGEFGGNDYNAPLFAGKDLREAYNLMPHVVQGISDGVEQLIAEGARDLIVPGVMPSG
BAC16480  FMVGEFGGNDYLHPLFQNKTLEWVRPLVPRVVRYIAGAVEELVGLGATTVYVPGLFPLG
```

Fig.3. Amino acid sequences of the gene families of lipases and related proteins in *Arabidopsis* and *Oryza* genomes aligned in the region homologous to exons 3 and 4 of the At1g28660 gene.
The symbol ⇓ indicates position of intron 3 in the At1g28660 gene. Sequences of DU inside exon 3 and a part of DU inside exon 4 are underlined. Pro residues strictly conserved in all sequences excluding At1g28660 are printed in bold.

The family of vesicular-associated proteins (VAMP) (Fig.4). BLAST searches revealed that similarly sized conserved 5'-edge of the exon 1 is characteristic of VAMPs and related proteins from *Arabidopsis*, *Oryza* and other species. The 5'-edge of the exon 1 of the gene C is peculiarly extended due to presence of two copies (identical and almost identical) of non-coding/coding border designated as DU2 in Fig.4. Sequence segment homologous to DU2 (83% identities) forms related non-coding/coding border in the gene B. In combination, these observations are sufficient to conclude that the non-coding/coding DU2 sequence pre-existed in immediate gene C ancestor was copied two times into exon 1 and thereby created its novelty.

The family of VAMP genes represented by the genes A'-C provides sufficient information for a more deep reconstruction of the history of events finalized by formation of the extant gene C structure. The genes A', A and B structures reflect situation formalized in Fig.1A: black segment designated as DU1 in Fig.4, pre-existed in coding regions of the genes A' and A, in the gene B was duplicated into non-coding region generating its extension. According to the summary scheme (Fig.4), the non-coding black segment of the gene B has been served as an intermediate between black exon segment of the same gene and black 5'-leader segment of the gene C. Similarity relationships between these three segments (Fig.4C) support this conclusion. The fourth black segment in the gene C designated as DU1 was inherited from black exon segments of the genes A', A and B and retained their sequence features (Fig.4A).

In summary, coding region novelty in the gene C was created in two-steps: insertion of a segment from coding region into 5'-leader followed by its modification, and triplication of the modified 5'-leader/exon border into coding region. Partial overlapping of sequence segments subjected to successive but independent duplication/triplication events exemplifies suggestive mutual exchange by sequence information between non-coding/coding regions.

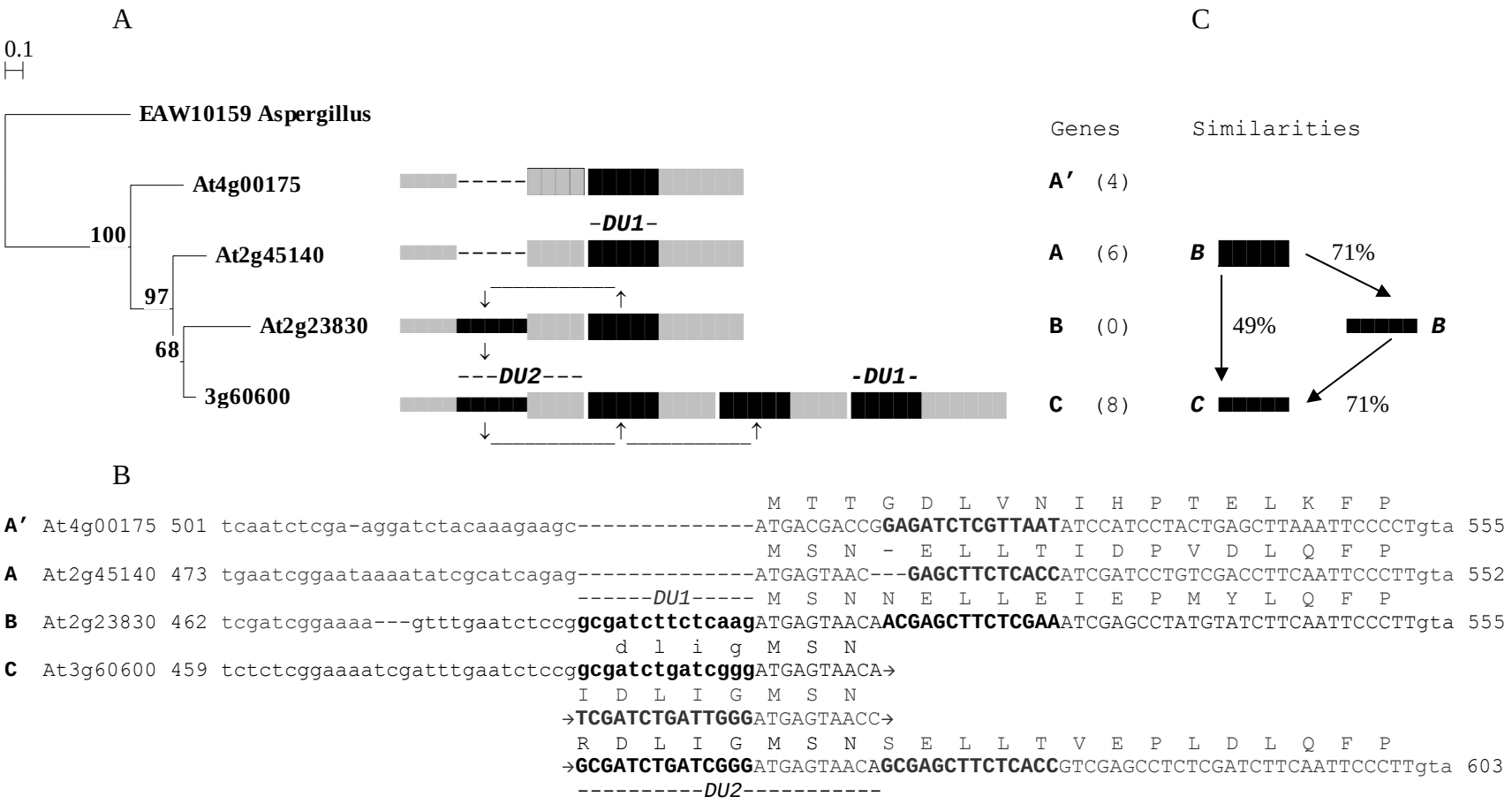


Fig.4. The gene family of vesicular-associated proteins (VAMP) exhibits mutual exchange by sequence information between 5'-leader and flanking exon.

A, evolutionary relationships between genes A', A, B and C as they specified in Fig.1. In total 119 positions of aligned amino acid sequences coded by the genes A'-C and by a gene of fungal VAMP (outgroup) have been used for the tree construction. The tree branches are supplemented by schemes describing the structures of respective genes in the region of 5'-leaders and exons 1 (thin and thick rectangles, respectively). Black rectangles indicate position of putative doublet units DU1 in the gene B and homologous sequence segments in other genes. In the gene C extended doublet unit DU2 is shown two times copied into exon 1. The number of mRNAs known for the genes A'-C is shown in brackets. B, aligned nucleotide sequences and translated amino acid sequences of exon 1 (capitals) and 5'-leader (low case letters). Three copies of the doublet unit DU2 are shown folded in the alignment. Similarity between aligned regions of 5'-leaders upstream from segments involved in duplication/triplication in the genes B and C (28 positions) exceeds chance level (79.6% identities). C, similarity relationships (identities, %) between coding/non-coding sequence segments homologous to DU1 in the genes B and C.

Concluding remarks

The deduced ability of non-coding sequence elements duplicated into coding regions to create their novelty has been convincingly shown realized only in two genes of the *Arabidopsis* genome. However, it should be taken into account that only recent duplications that left detectable traces are available for description. Meanwhile, all adopted coding region alterations derived from non-coding/coding duplications should be accumulated during genome evolution. Therefore, in many cases sequence features of gene coding regions (especially those features that are specific within individual members of a gene family) might have been derived from pre-existed non-coding elements. Moreover, a single-step duplication process can be hypothesized to be an exclusive evolutionary tool for creation of a novelty advantageous inside highly conserved coding region; such a novelty might be unattainable by means of several successive independent point mutations. The described evolutionary process focused to a restricted gene coding region might speed up the fine-tuning and increase the flexibility of the structural and functional adaptation of a protein.

References:

1. Thomas E.E., Srebro N., Sebat J., Navin N., Healy J., Mishra B., Wigler M. Distribution of short paired duplications in mammalian genomes // Proc. Natl. Acad. Sci. U.S.A. - 2004. - Vol.101. - P.10349-10354.
2. Shutov A., Bäumlein H., Altschmied L. Large numbers of tandem duplications in the *Arabidopsis* genome contribute novelty to coding sequences // Third Plant GEMs Meeting, Lyon. - 2004. - Poster 040923.
3. Van de Peer Y., De Wachter R. TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environement // Comp. Appl. Biosci. - 1994. - Vol.10. - P.569-570.

Prezentat la 30.07.2007