Protein splicing and its evolution in eukaryotes

P. L. Starokadomskyy

Institute of Molecular Biology and Genetics, NAS of Ukraine
150, Academician Zabolotnogo Str., Kyiv, Ukraine, 03680

p.l.starokadomski@imbg.org.ua

Inteins, or protein introns, are parts of protein sequences that are post-translationally excised, their flanking regions (exteins) being spliced together. This process was called protein splicing. Originally inteins were found in prokaryotic or unicellular eukaryotic organisms. But the general principles of post-translation protein rearrangement are evolving yielding different post-translation modification of proteins in multicellular organisms. For clarity, these non-intein mediated events call either protein rearrangements or protein editing. The most intriguing example of protein editing is proteasome-mediated splicing of antigens in vertebrates that may play important role in antigen presentation. Other examples of protein rearrangements are maturation of Hg-proteins (critical receptors in embryogenesis) as well as maturation of several metabolic enzymes. Despite a lack of experimental data we try to analyze some intriguing examples of protein splicing evolution.

Keywords: protein splicing, protein editing, proteasome, evolution.

Introduction. Splicing is well known to be a processing mechanism that excises an internal region from a precursor molecule with subsequent ligation of the flanking sequences. Until recently splicing was revealed only for nucleic acids. However, in 1990, Stevens et al. have found that one yeast protein – VMA1 – acquire an ability of unusual posttranslational modification via process, similar to DNA and RNA splicing [1, 2]. By analogy they named this process «protein splicing». Below we briefly describe the main concepts of this phenomenon.

Protein splicing is an autocatalytic process and is independent of any cofactor or enzyme, differing thereby from posttranslational processing (Fig. 1). It was proposed to use the term intein (internal protein) for the central protein region, which is subject to excision; terms N- and C-extein (external protein) for the corresponding flanking sequence [2, 3].

Primary structure of inteins. More than 100 inteins are describe for today [4]. Usually a size of inteins vary from 134 to 650 amino acid (aa) residues [5, 6], although there are some exceptions (giant inteins TerDnaE-1 and TerDnaB-1 – 1308 and 1650 residues correspondingly) [6, 7]. Inteins can be divided into four large groups: maxi-inteins (with central endonuclease domain), mini-inteins (lacking the endonuclease domain), trans-splicing inteins (the splicing junction belong to different proteins) and Ala-inteins (with non-canonical Ala as the N-terminal amino acid, see later) (Fig. 2) [8–10].

Analysis of sequences found the most important sites and motifs within inteins (Fig. 2) [4, 11–13]. Position 1 at the N-terminus of an intein is almost always occupied by Cys or Ser or, in extremely rare cases (in Ala-inteins) by Ala. Second most important motif is a short C-terminal sequence of 8 residues, seven of which belong to the intein and the eighth one is the N-terminal residue of C-extein. These two last
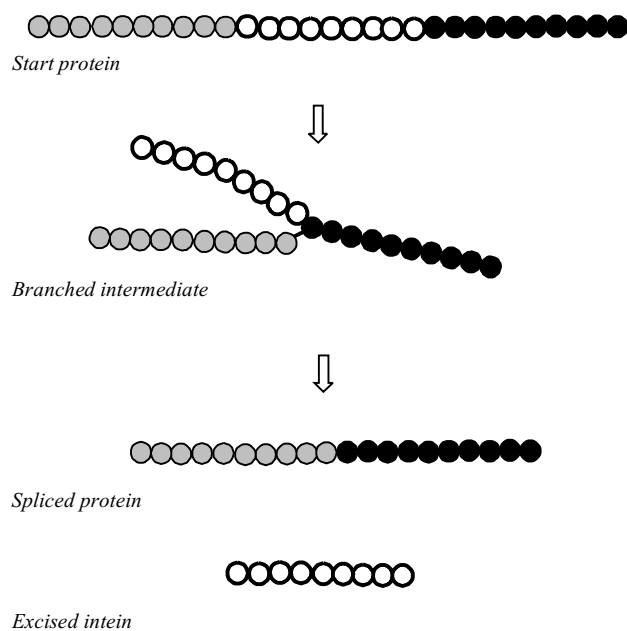
Start protein

Branched intermediate

Spliced protein

Excised intein

Fig.1. General scheme of protein splicing

lation factors [15]. However, their distribution is highly nonuniform. For instance, intein-coding sequences were found in 19 genes only in *Methanococcus jannaschii*, whereas no intein-coding sequences are in the genomes of more than 30 species (prokaryotes and eukaryotes), including *Arabidopsis thaliana, Agrobacterium tumefaciens, Escherichia coli* K12, *Drosophila melanogaster, Homo sapiens* etc. [4].

The majority of intein sequences were found in the genomes of prokaryotes, mostly *archaebacteria* and *eubacteria*. A gene usually harbors one or, in rare cases, two intein-coding sequences (3–4 very rarely). In eukaryotes, inteins were found in genomes of about 50 taxons mostly of primitive eukaryotes (*Saccharomyces cerevisiae, D. hansenii, Rhodothermus marinus* etc.) [4, 9, 16–20]. Also inteins were found in plastids of some unicellular organisms (*Algae porphyra, Euglena gracilis, Chlorella vulgaris,* etc. [4, 15, 21–23]).

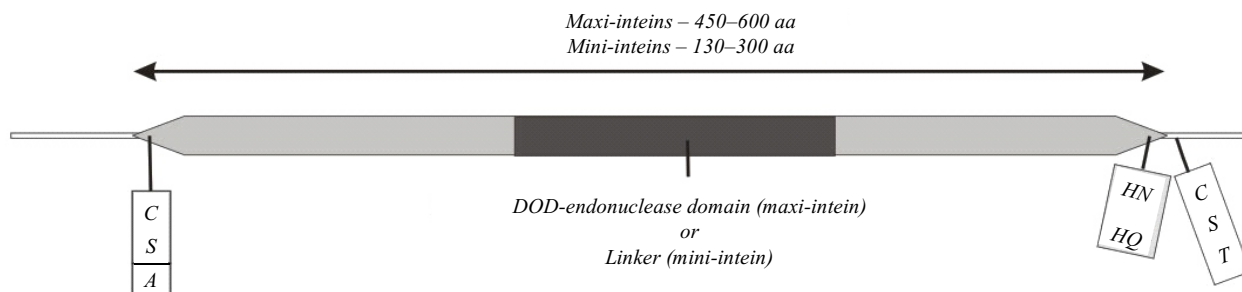Horizontal transfer is an important mechanism

Fig. 2. General structure of inteins

residues play an important role in hydrolysis of a peptide bond between intein and C-extein [4, 13], that is necessary for the final step of splicing.

A more detailed description of the functional motifs and chemical mechanism of the splicing is omitted here and can be found in the previous works (e. g. [8, 14]).

Evolution of inteins in prokaryotes and primitive eukaryotes. Apparently, inteins have appeared early in evolution [5, 15]. This is indirectly supported by the fact that they are spread in all three domains of life. Intein host proteins are very diverse, including DNA and RNA-polymeres, ATPases, proteases, metabolic enzymes, transcription and trans-

allowing the spreading of intein sequences. This is evident from the fact that inteins of homologous proteins from different species often share many features [8]. The possibility of horizontal transfer of inteins between prokaryotes and eukaryotes is evident from some data. For instance, inteins within the *dnaB* of two different organisms (*R. marinus* and *Synechocystis sp. PCC6803*) display far higher homology as compared to the exteins [22, 24]. Also transfer of inteins can occur between non-allelic genes, which is possibly due to endonucleolytic cleavage of a similar, though not identical, site [15]. This hypothesis is supported by the fact that the *M. leprae* and *M. tuberculosis* genomes have three homologous inteins in
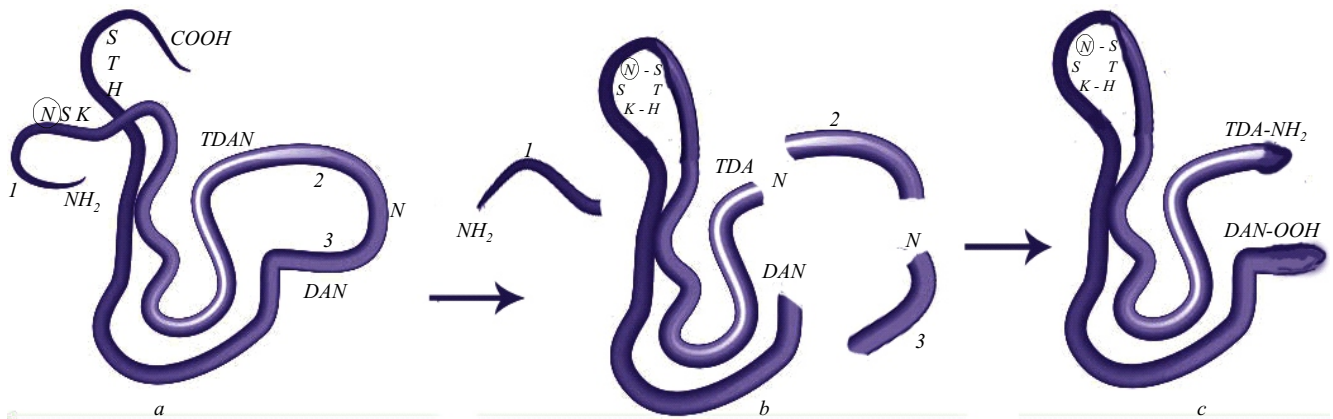
Fig. 3. Post-translation modification of ConA: *a* – excision of N-term peptide; *b* – ligation of N- and C-terminus; *c* – excision of central loop

pps1, recA, and *dnaB*, while the intein insertion sites are non-homologous [4].

Inter-specific horizontal transfer of inteins is mediated first at all by viruses. Intein-coding sequences were found in the genomes of several prokaryotic and eukaryotic viruses [19, 25–27]. The presence of intein sequences in virus genomes (*i. e.* intein APMVPol in *polB* of the *Mimivirus* [19, 27]) provides a brilliant illustration for the mechanism ensuring horizontal transfer of inteins.

As mentioned above, horizontal transfer is an important mechanism of spreading intein sequences resulting in similarity of inteins of homologous proteins from different species [8]. An evident example is provided by eukaryotic chitin synthase intein PanCHS2. This intein was found in *P. anserine*, but its analogs were not detected in the chitin synthase genes of other species. Unexpectedly, PanCHS2 is highly homologous to glutamate synthase intein PanGLT1 [15]. This finding suggests a horizontal transfer of intein-coding sequences between nonallelic genes and different hosts.

Evolution of protein splicing's principle in higher eukaryotes. It's fairly to presume, that a principle of protein splicing should inherits and upgrades in eukaryotic organisms. Different papers have described non intein-mediated protein rearrangements in eukaryotes. The mechanism of these rearrangements is not autocatalytic and modifications are catalyzed either by specific protease [20, 28] or by proteasome [16–19, 29]. For clarity, it was suggested to call these non-intein mediated events either protein rearrangements or protein editing. Here we consider some of interesting cases.

Post-translational rearragement by reverse proteolysis in plants. One of examples of post-translational modification, namely, protein splicing by reverse proteolysis, was discovered in plants. It was observed in case of lectin concanavalin A (ConA) [29] – protein which binds specifically to certain structures found in various sugars, glycoproteins, and glycolipids, mainly internal and nonreducing terminal alpha-mannosyl groups.

Some time ago it was showed, that maturation of this protein occured in non-typical way. Although the exact mechanism of this splicing in plants is still not clear, the fact that all digestion occurs before Asn, suggests a participation of some endopeptidase. Indeed, *in vitro* study showed that specific enzyme as- paraginyl endopeptidase (AEP) could digest ConA and then re-ligate the fragments by its reverse proteolytic activity [20, 30].

The initial precursor of ConA (glyco-pro-ConA) is first activated by deglycosylation to pro-Con A (Fig. 3). Pro-Con A is then cleaved to produce two distinct proteins that are transposed and relegated to become a mature ConA. A biological role of such rearrangement it is not clear.

Proteasome-mediated protein splicing. Recently it has been shown that proteasomes are cell compartments where not only protein degradation but also post-translational protein modifications occur [16, 17].
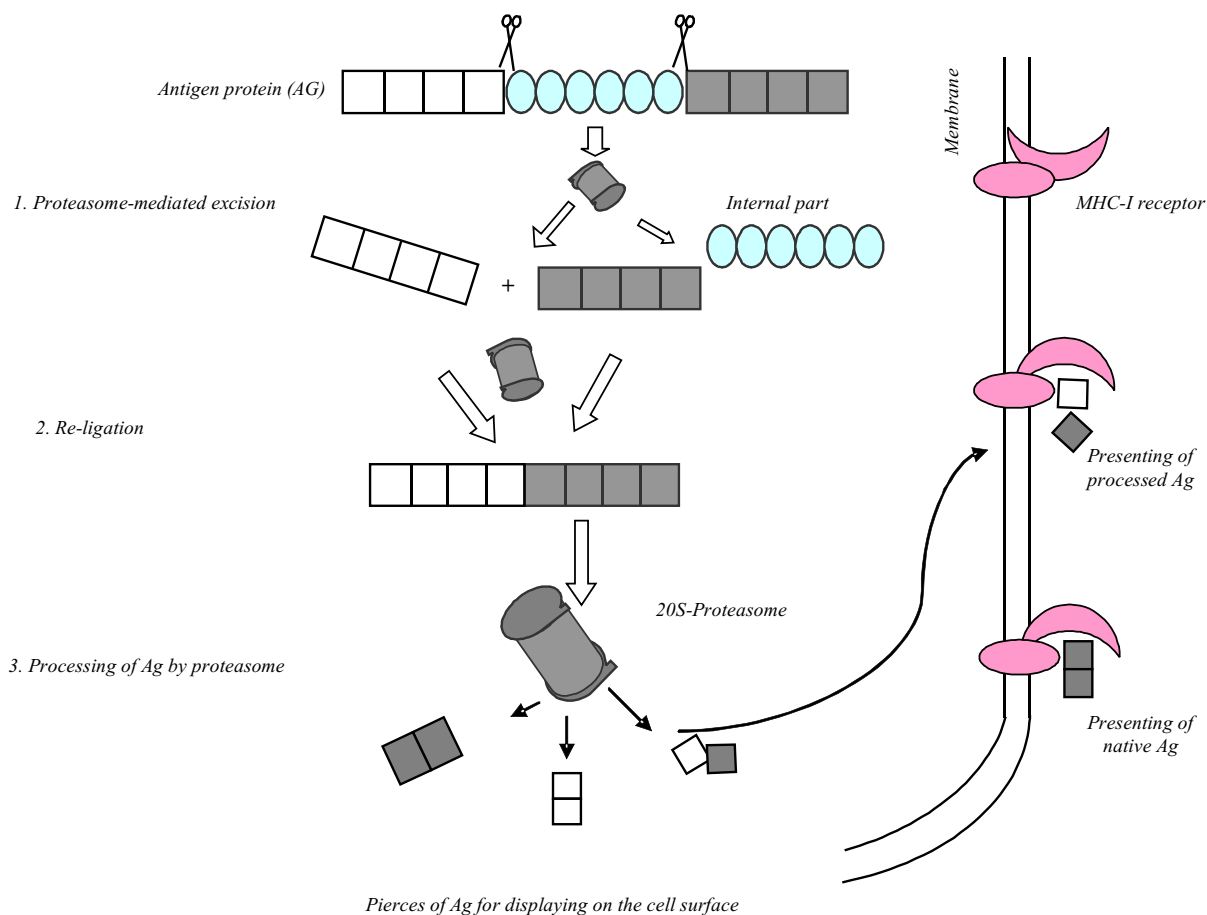
Fig. 4. Scheme of proteasome-associated protein rearrangements

Hanada et al. have discovered that antigen maturation and presentation for cytotoxic T-lymphocytes (CTLs) is performed by proteasome through protein splicing mechanism (peptide excision and re-ligation) (Fig. 4).

They found, that CTLs, killing cancer cells overexpressing fibroblast growth factor-5 (FGF-5), recognize presented by MHC class I molecules a nine-residue FGF-5 peptide generated by protein splicing. Using the proteasome inhibitor, the authors evidenced that FGF-5 protein splicing observed in their system is proteasome-mediated. This splicing was conducted by 20S proteasome *in vitro* and *in vivo*. They showed that in contrast to intein splicing, that usually excised polypeptide of not less than 134 aa in length, the excised polypeptides in their system were only 18–40 aa [17]. Thus, the proteasome-mediated splicing overcomes the limitation, associated with the length of

intervening sequences, and can be considered as a next step of protein splicing evolution. The detailed mechanism of proteasome-mediated splicing you can find in [17]. Shortly, it was shown that proteasome -subunits directly catalyze protein rearrangement in a way very similar to protein *trans*-splicing.

Implications of these process allow the immune system to monitor non-contiguous peptide sequence generated post-translationally. This capability represents an enormous increase in the ability of immune cells to recognize intrinsic and foreign proteins.

Intein-like processing of embryonic signal proteins of the Hedgehog (Hh) family. Signal proteins of the Hh family were found in human, mouse, insects, and other multicellular organisms. The Hh family proteins play a crucial role for embryo development [5, 8, 24, 31–33]. In context of our review the autocatalytic maturation of
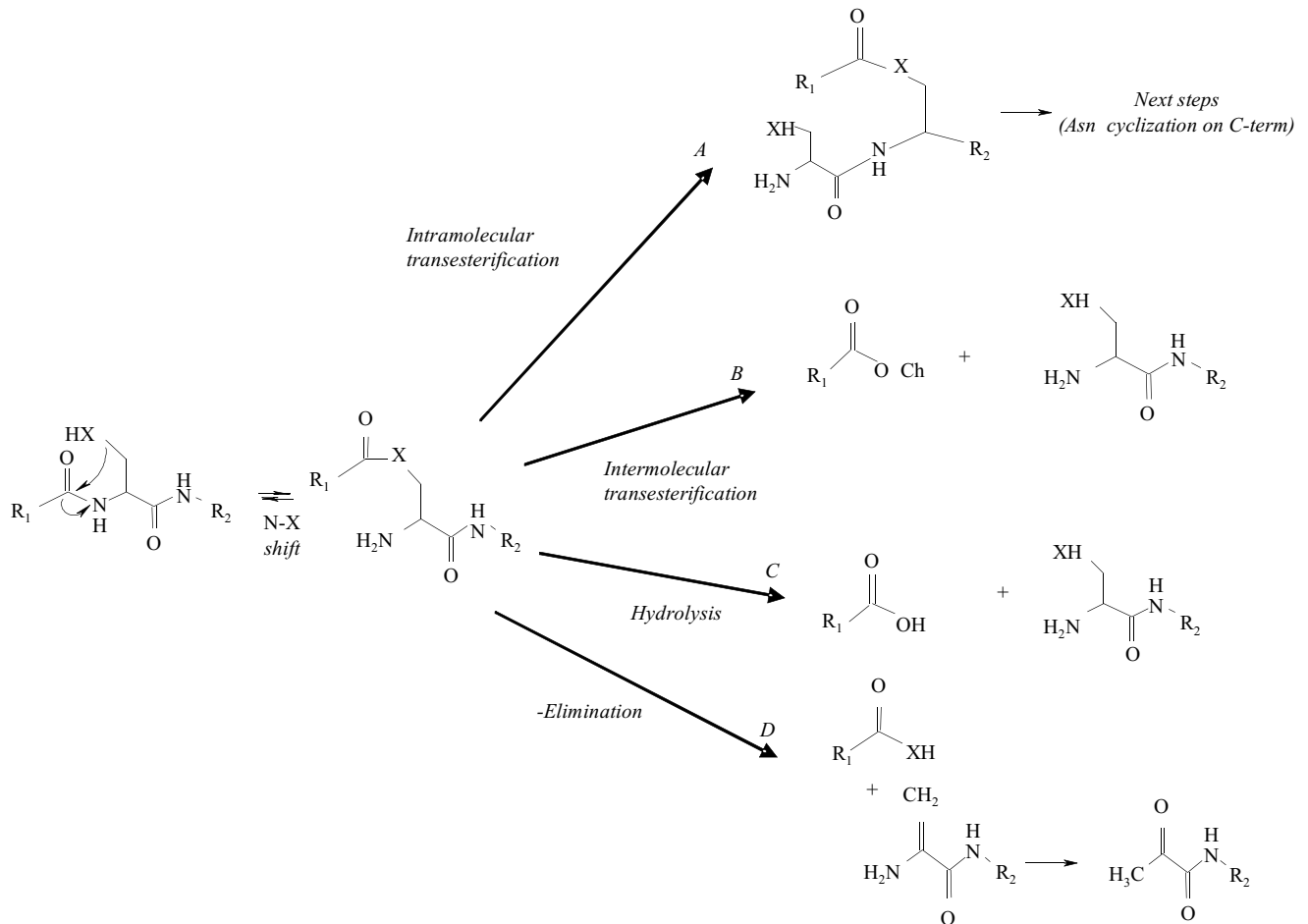
Fig. 5. Role of N–O and N–S acyl shifts in autoprocessing of various proteins [31]: *A* – protein splicing *X* = S or O; *B* – autoprocessing of Hedgehog proteins Ch-OH = cholesterol; *C* – autoprocessing of glycosyl asparaginase *X* = O; *D* – autoprocessing of pyruvoyl enzymes *X* = O

Hh proteins will be considered. The Hh proteins are synthesized as inactive precursors consisting of two domains, N-terminal signal domain (Hh-N) and C-terminal catalytic domain (Hh-C) (Fig. 5, *B*). Hh-C forms an intein-like domain, containing conserved motifs similar to intein. The structural similarity suggests a common origin of the inteins and Hh proteins. Yet their host organisms differ: the inteins occur mostly in unicellular, whereas the Hh proteins were found only in multicellular organisms [8, 32].

In contrast to the inteins, during the Hh proteins during maturation excise themselves onto two parts. The reaction requires an endogenous inductor, cholesterol. As a result, the peptide bond between Hh-N and Hh-C parts is disrupted and cholesterol is covalently linked to Hh-N. Modified with cholesterol, Hh-N migrates toward the plasma membrane as a re-

ceptor or is exported from the cell to perform its function [30, 32].

Autocatalytic modification of enzymes. Similar to protein splicing, the reactions with subsequent hydrolysis of the peptide bond are involved in maturation of several enzymes like N-terminal nucleophilic hydrolases, N-terminal aminotransferases, and pyruvate-dependent enzymes (Fig. 5, *C*, *D*) [22, 31, 33, 34]. Interestingly that their autoprocessing is catalyzed by domains that totally differ in sequence and three-dimensional structure from the inteins, but chemical principles are the same.

Catalytic activity of enzymes from the first two families depends on highly reactive Ser, Thr, or Cys located at the N-end. The mechanism of autoprocessing can be considered on the example of *Flavobacterium meningosepticum* glycosyl asparaginase, since a si-

milar mechanism is involved in maturation of glycosyl glycosyl asparaginase of other organisms, including *H. sapiens*. The mature enzyme is a heterodimer encoded by one gene. An inactive protein precursor is converted into an active form as a result of autocatalytic cleavage of the peptide bond before Thr152, yielding two separate subunits,   and . The reaction is triggered by an N–O shift with generation of the ester bond, which is then hydrolyzed (Fig. 5, *C*) [22, 31]. As a result, Thr152 becomes the N-terminal residue of the -subunit. Substitution of Thr152 with an amino acid residue of a different type (other than Ser or Cys) totally abolishes cleavage.

Biological functions of protein splicing into the cells. The function of inteins was initially explained in terms of «egoistic genes». An intein utilizes the endonuclease domain to spread itself into the genome, and Hint domain to cut-and-paste host protein, preserving its biological activity (*i. e.*, viability of the host cell). Located within a protein sequence, an intein utilizes the cell mechanism of protein expression, including the regulatory sequences of the host gene and its mRNA, that testifies to the selfishness of the intein sequence.

However, this point of view is rather naive. The intein sequences can indeed be classified as mobile elements according to some of their features, first and foremost, processes of homing [8, 15]. However, while retrotransposons occur in many copies in the genome, intein sequences are restricted to certain sites of particular genes, and their genomic distribution directly correlates with the copy number of an intein-coding gene. Also because inteins are integrated in highly conserved regions, the functions of host protein before and after excision of an intein should be different.

It seems more probable that free inteins (after splicing) perform still unknown regulatory and enzymatic functions in the cell [35]. For instance, inteins may regulate the activity of the host enzymes through regulation of their maturation via *cis*- or *trans*-splicing. The most evident example is the arrangement of active enzyme (DnaE in *Synechocystis* sp. PCC6803) from two different proteins through the *trans*-splicing mechanism [14, 15, 36]. The N- and C-terminal halves of DnaE (catalytic subunit of DNA polymerase III) are encoded by two separate genes, *dnaE-N* and *dnaE-C*, respectively.

The *dnaE-N* product consists of a N-extein sequence followed by a 123-aa intein sequence, whereas the *dnaE-C* product consists of a 36-aa intein sequence followed by a C-extein sequence. The two intein sequences together reconstitute a split mini-intein, which mediates reconstitution of the full active DnaE. Its means that here the intein can be an integrated regulator of polymerase activity.

In addition, the cleaved off inteins may catalyze certain reactions in the cell, using its activated N-terminal nucleophilic group and acting similarly to activated glycosyl asparaginase, ConA or pyruvoyl enzymes [22].

The occurrence of protein splicing in vertebrates has important implications for the complexity of the vertebrate proteome and for the immune recognition of own and foreign peptides. A principle of protein splicing is fully used by immune system of mammalians and plants. Its usage allows to combine different parts of antigens providing a logarithmic increase of possible combinations.

In eukaryotes, DNA recombination and RNA splicing were already known to increase the number of different proteins produced by each of mammalian genes. The discovery of protein splicing adds new tools to this kit. In a broader context, the existence of protein splicing in vertebrates greatly increases the cell's options for converting genetic information on the post-translation level. All together splicing (DNA, RNA, protein) is a powerful mechanism allowing to realize the main principle of multi-functionality – transformation of original compact genetic information into wide diversity of protein structures and functions after its realization.

П. Л. Старокадомський

Білковий сплайсинг та його еволюція в еукаріотних організмів

Резюме

Інтеїни, або білкові інтрони, – це білкові послідовності, які автокаталітично вирізаються після трансляції, зшиваючи фланкуючі ділянки (екстеїни). Такий процес названо білковим сплайсингом. Спочатку інтеїни виявлено у прокаріотів і одноклітинних еукаріотів. Але загальний принцип посттрансляційної перестановки білків еволюціонував разом із живою матерією. На сьогодні ціла низка білків у багатоклітинних організмів модифікується за механізмами, принципово схожими з білковим сплайсингом у бактерій. Ці процеси отримали назву білкового редагування. Найцікавішим випадком білкового

редагування є «тасування» епітопів антигенів протеасомою, що відіграє важливу роль у процесі презентації антигену клітинами. Іншими прикладами слугують дозрівання Hg-білків (рецепторів, критичних для коректного протікання ембріогенезу) або дозрівання низки метаболічних ферментів. Незважаючи на брак експериментальних даних, в огляді зроблено спробу описати найяскравіші приклади еволюції у вищих еукаріотів.

Ключові слова: білковий сплайсинг, білкове редагування, протеасома, еволюція.

П. Л. Старокадомский

Белковый сплайсинг и его эволюция у эукариотных организмов

Резюме

Интеины, или белковые интроны, являются частью белковой последовательности, которая самовырезается после трансляции, сшивая фланкирующие участки (экстеины). Такой процесс назван белковым сплайсингом. Изначально интеины обнаружены у прокариотов и одноклеточных эукариотов. Но общий принцип посттрансляционной белковой перестановки эволюционировал вместе с живой материей. На сегодня целый ряд белков у многоклеточных организмов подвергается модификациям, принципиально схожим с белковым сплайсингом у бактерий. Эти процессы получили название белкового редактирования. Наиболее интересным случаем белкового редактирования является «перетасовка» эпитопов антигенов протеасомами, что, по-видимому, может играть определенную роль в процессе презентации антигена клетками. Другими примерами служат дозревание Hg-белков (рецепторов, критических для корректного протекания эмбриогенеза) или дозревание ряда метаболических ферментов. Несмотря на недостаток экспериментальных данных, в обзоре сделана попытка описать наиболее яркие примеры эволюции белкового сплайсинга у высших эукариотов.

Ключевые слова: белковый сплайсинг, белковое редактирование, протеасома, эволюция.

REFERENCES

1. *Kane P. M., Yamashiro C. T., Wolczyk D. F., Neff N., Goebl M., Stevens T. H.* Protein splicing converts the yeast TFP1 gene product to the 69 kDA subunit of the vacuolar H$^+$-adenosine triphosphatase // Science.–1990.–**250**, N 4981.–P. 651–657.

2. *Anraku Y., Mizutani R., Satow Y.* Protein splicing: Its discovery and structural insight into novel chemical mechanisms // IUBMB Life.–2005.–**57**, N 8.–P. 563–574

3. *Perler F., Davis E., Dean G., Gimble F., Jack W., Neff N., Noren C., Thorner J., Belfort M.* Protein splicing elements: Inteins and exteins, a definition of terms, and recommended nomenclature // Nucl. Acids Res.–1994.–**22**, N 7.–P. 1125–1127.

4. *Perler F.* InBase, the intein database // Nucl. Acids Res.–2002.–**30**, N 1.–P. 383–384.

5. *Noren C., Wang J., Perler F.* Dissecting the chemistry of protein splicing and its applications // Angew. Chem. Int. Ed.–2000.–**39**, N 3.–P. 450–466.

6. *Perler F.* Protein splicing mechanisms and applications // IUMBM Life. –2005. –**57**, N 7.–P. 469–476.

7. *Yang J., Meng Q., Liu X.* Intein harbouring large tandem repeats in replicative DNA helicase of *Trichodesmium erythraeum* // Mol. Microbiol.–2004.–**51**, N 4.–P. 1185–1192.

8. *Starokadomskyy P. L.* Protein splicing // Mol. Biol. (Moscow).–2007.–**41**, N 2, P. 278–293.

9. *Starokadomskii P. L.* Protein splicing // Ukr. Biochem. J.–2005.–**77**, N 4.–P. 14–29.

10. *Evans T., Xu M.* Mechanistic and kinetic considerations of protein splicing // Chem. Rev.–2002.–**102**, N 12.–P. 4869–4883.

11. *Pietrokovski S.* Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins // Protein Sci.–1994.–**3**, N 12.–P. 2340–2350.

12. *Perler F., Olsen G., Adam E.* Compilation and analysis of intein sequences // Nucl. Acids Res.–1997.–**25**, N 6– P. 1087–1093.

13. *Pietrokovski S.* Modular organization of inteins and C-terminal autocatalytic domains // Protein Sci.–1998.–**7**, N 1–P. 64–71.

14. *Choi J. J., Nam K. H., Min B., Kim S. J., Soll D., Kwon S. T.* Protein trans-splicing and characterization of a split family B-type DNA polymerase from the hyperthermophilic archaeal parasite *Nanoarchaeum equitans* // J. Mol. Biol.–2006.–**356**, N 5.–P. 1093–1106.

15. *Pietrokovski S.* Intein spread and extinction in evolution // Trends Genet.–2001.–**17**, N 8.–P. 465–472.

16. *Hanada K., Yewdel J. W., Yang J. C.* Immune recognition of a human renal cancer antigen through post-translational protein splicing // Nature.–2004.–**427**, N 6971.–P. 252–256.

17. *Hanada K., Yang J. C.* Novel biochemistry: post-translational protein splicing and other lessons from the school of antigen processing // J. Mol. Med.–2005.–**83**, N 6.–P. 420–428.

18. *Rammensee H. G.* Immunology: protein surgery // Nature.–2004.–**427**, N 6971.–P. 203–204.

19. *Saska I., Craik D.* Protease-catalysed protein splicing: a new post-translational modification? // Trends Biochem. Sci.–2008.–**33**, N 8.–P. 363–368.

20. *Hara-Nishimura I., Inoue K., Nishimura M.* A unique vacuolar processing enzyme responsible for conversion of several proprotein precursors into the mature forms // FEBS Lett.–1991.–**294**, N 1–2.–P. 89–93.

21. *Sun P., Ye S., Ferrandon S., Evans T. C., Xu M. Q., Rao Z.* Crystal structures of an intein from the split *dnaE* gene of *Synechocystis* sp. PCC6803 reveal the catalytic model without the penultimate histidine and the mechanism of zinc ion inhibition of protein splicing // J. Mol. Biol.–2005.–**353**, N 5.–P. 1093–1105.

22. *Bjornsdottir S. H., Blondal T., Hreggvidsson G. O., Eggertsson G., Petursdottir S., Hjorleifsdottir S., Thorbjarnardottir S. H., Kristjansson J. K.* Rhodothermus marinus: Physiology and molecular biology // Extremophiles.–2006.–**10**, N 1.–P. 1–16.

23. *Khan M. S., Khalid A. M., Malik K. A.* Intein-mediated protein trans-splicing and transgene containment in plastids // Trends Biotech.–2005.–**23**, N 1.–P. 217–220.

24. *Gogarten J., Senejani A., Zhaxybayeva O., Olendzenski L., Hilario E.* Inteins: Structure, function, and evolution // Annu. Rev. Microbiol.–2002.–**56**.–P. 263–287.

25. *Hertveldt K., Lavigne R., Pleteneva E., Sernova N., Kurochkina L., Korchevskii R., Robben J., Mesyanzhinov V., Krylov V. N., Volckaert G.* Genome comparison of *Pseudomonas aeruginosa* large phages // J. Mol. Biol.–2005.–**354**, N 3.–P. 536–545.

26. *Nagasaki K., Shirai Y., Tomaru Y., Nishida K., Pietrokovski S.* Algal viruses with distinct intraspecies host specificities include identical intein elements // Appl. Environ. Microbiol.–2005.–**71**, N 17.–P. 3599–3607.

27. *Suhre K., Audic S., Claverie J. M.* Mimivirus gene promoters exhibit an unprecedented conservation among all eukaryotes // Proc. Nat. Acad. Sci. USA. 2005.–**102**, N 41.–P. 14689–14693.

28. *Carrington D. M., Auffret A., Hanke D. E.* Polypeptide ligation occurs during post-translational modification of concanavalin A // Nature.–1985.–**313**, N 5997.–P. 64–67.

29. *Zaretsky J., Wreschner D.* Protein multifunctionality: Principles and mechanisms // Translat. Oncogenom.–2008.–**3**.–P. 99–136.

30. *Min W., Jones D. H. In vitro* splicing of concanavalin A is catalyzed by asparaginyl endopeptidase // Nat. Struct. Biol.–1994.–**1**, N 8.–P. 502–504.

31. *Paulus H.* Protein splicing and related forms of protein autoprocessing // Annu. Rev. Biochem.–2000.–**69**.–P. 447– 496.

32. *Ingham P. W., McMahon A. P.* Hedgehog signaling in animal development: Paradigms and principles // Genes Develop.–2001.–**15**, N 23.–P. 3059–3087.

33. *Amitai G., Belenkiy O., Dassa B., Shainskaya A., Pietrokovski S.* Distribution and function of new bacterial intein-like protein domains // Mol. Microbiol.–2003.–**47**, N 1.–P. 61–73.

34. *Albert A., Dhanaraj V., Genschel U., Khan G., Ramjee M. K., Pulido R., Sibanda B. L., von Delft F., Witty M., Blundell T. L., Smith A. G., Abell C.* Crystal structure of aspartate decarboxylase at 2.2 Å resolution provides evidence for an ester in protein self-processing // Nat. Struct. Biol.–1998.–**5**, N 4.–P. 289–293.

35. *Klabunde T., Sharma S., Telenti A., Jacobs W. R., Sacchettini J. C.* Crystal structure of GyrA intein from *Mycobacterium xenopi* reveals structural basis of protein splicing // Nat. Struct. Biol.–1998.–**5**, N 1.–P. 31–36.

36. *Wu H., Hu Z., Liu X.* Protein trans-splicing by a split intein encoded in a split *DnaE* gene of *Synechocystis* sp. PCC6803 // Proc. Nat. Acad. Sci. USA.–1998.–**95**, N 16.–P. 9226–9231.