Theor Appl Genet (1989) 78:852–856

An attempt to elucidate the origin of cultivated soybean via comparison of nucleotide sequences encoding glycinin B_4 polypeptide of cultivated soybean, *Glycine max*, and its presumed wild progenitor, *Glycine soja*

E.S. Zakharova*, S.M. Epishin and Yu.P. Vinetski

Institute of Genetics and Selection of Industrial Microorganisms, Moscow, USSR

Received May 13, 1989; Accepted June 27, 1989
Communicated by Yu. Gleba

Summary. Nucleotide sequences of cDNAs encoding soybean glycinin B_4 polypeptide were compared for three soybean cultivars and two introductions of wild soybean, *G. soja*. For three *G. max* cultivars, only two nucleotide substitutions were found, while *G. max* and *G. soja* nucleotide sequences had four substitutions. These data give added proof that *G. max* originated from *G. soja*. On the other hand, the time required for the accumulation of four nucleotide substitutions (calculated from the parameters of 11S globulin molecular evolution) appeared to be longer than the duration of the soybean domestication period.

Key words: Storage protein – Glycinin – Nucleotide sequences – soybean – molecular evolution

Introduction

In modern systems of the *Papilionaceae* family, *Glycine max* and *Glycine soja* belong to the *Soja* subgenus (Hermann 1962) and the latter is considered a wild progenitor of cultivated *G. max* (Hymowitz 1970). According to Vavilov (1926), China is the centre of origin of cultivated soybean *G. max*. Wild soybean may be found in Manchuria, adjoining regions of the USSR, Korea and Taiwan. Both species are cytologically and biochemically similar: they have 40 chromosomes in the diploid set, identical sets of polypeptides of glycinin seed storage protein and restriction-fragment pattern of rRNA genes (Hymowitz 1970; Aleksenko et al. 1985; Doyle and Beachy 1985; Doyle 1988). Nevertheless, all these data seem insufficient to solve the problem of the cultivated soybean progenitor. We made an attempt to approach this problem by studying the nucleotide sequences of glycinin genes of the soybean species. Apart from unravelling the origin of one of the main species of cultivated plants, these data give us a unique possibility to compare the longevity of the plant domestication period with a calculated individual gene divergency time of the soybean species.

Glycinin is a storage protein of soybean seed. It belongs to the 11S globulin class. The molecules of this globulin include six subunits. Each subunit contains acidic (A) and basic (B) polypeptides, which are encoded by one gene and are translated as a high-molecular-weight precursor subjected to post-translational modification. Five genes forming a multigene family and encoding these subunits were identified. This family can be subdivided into two subfamilies by their nucleotide sequence homology. Genes encoding subunits A_2B_{1a}, $A_{1b}B_{1b}$, $A_{1a}B_2$ belong to the first subfamily, and genes A_3B_4 and $A_5A_4B_3$ to the second (Nielsen 1984). cDNAs of these subunits were isolated and completely or partially sequenced (Momma et al. 1985a, b; Scallon et al. 1985; Fukazawa et al. 1985; Epishin et al. 1986; Zakharova et al. 1986).

Storage 11S seed globulins are a convenient model for studying the molecular evolution of plant proteins and their genes. It was shown recently that some 11S soybean globulin polypeptides have antigenic homology, with the 11S globulins of taxonomically distant plants, gymnosperms and cereals among them (Aleksenko et al. 1988). Comparison of amino acid sequences of separate 11S globulin subunits from soybean, peas, rapeseed, cotton and pumpkin revealed homology with rice glutelin and 12S oat globulin (Walburg and Larkins 1986; Borroto and Dure 1987; Hayashi et al. 1987; Higuchi and Fukazawa 1987). These and other data allowed us to

* To whom correspondence should be addressed

calculate the parameters of 11S plant globulin molecular evolution at the amino acid and nucleotide levels (Aleksenko et al. 1989).

In the present study, cDNAs encoding B_4 polypeptide of the seed soybean storage protein glycinin in three *G. max* cultivars and in two different plant introductions of the presumed wild progenitor of cultivated soybean were isolated and sequenced. We discovered that the three compared cultivar sequences differ in two nucleotides. The sequences of two different plant introductions were absolutely similar and differed from *G. max* polypeptides in four nucleotides only. This provides additional molecular evidence for the origin of *G. max* from *G. soja*. Nevertheless, the time required for the revealed number of substitutions, calculated from the parameters of 11S globulin molecular evolution, appeared to be significantly longer than the duration of the soybean domestication period. This discrepancy allows us to suggest several hypotheses concerning the origin of cultivated soybean.

Materials and methods

Characterization of soybean cultivars and wild soybean introductions. Soybean cultivars "Mandarin" and "Mukden" of Chinese provenance were used. They are among the six main maternal ancestors of all the USA north and south cultivars (frequencies of occurrence in the parentage are 51% and 4%, respectively), according to the Committee on Genetic Vulnerability on Major Crops (1972). The seeds of *G. soja* plant introductions L23 and L582 (VIR catalogue N8432 and N8427, respectively) were provided by T.S. Sedova (Laboratory of Field Cultures of the VIR Far East Experimental Station). Phenotypically different plant introductions L23 and L582 were collected by A. Ala and T.S. Sedova in the Amur region.

mRNA isolation. Maturating seeds of cultivated and wild soybean were kindly provided by V.I. Sychkar (Institute of Genetics and Selection, Odessa). The material was harvested 21–23 days after flowering. The harvested seeds were frozen in liquid nitrogen and stored at –70 °C. Total RNA isolation by phenol method and poly(A)-RNA isolation was performed as described by Maniatis et al. (1982).

ds-cDNA synthesis and cloning. ds-cDNA (bouble-stranded complementary DNA) synthesis and cloning were performed according to Maniatis et al. (1982). ds-cDNA was cloned by using dG/dC connectors in vector plasmid pTZ 18 after cleavage with Sma1 restrictase.

Transformation of *E. coli* TG 1: Δ (lac, pro), thi, str A, sup E, end A, sbc B, hsd R^-, [F^- tra D36, pro AB, lac^{Iq}, zΔM15], was made according to Hanahan (1983).

Recombinant clones selection. Recombinant clones were selected via hybridization of colonies on nitrocellulose filters. cDNA of glycinin A_3B_4 subunit cloned in pSPG 204 plasmid was used as a probe (Epishin et al. 1986). Hybridization was made at 65 °C in 5 × Denhardt's solution (1 × is 0.02% bovine serum albumin, 0.02% Ficoll 400, 0.02% polyvinylpyrrolidone), 0.2% SDS (sodium dodecyl sulfate), 5 × SSC (1 × SSC is 0.15 *M* sodium citrate, pH 7.0), 100 mg/ml denatured DNA. After hybridization, the filters were washed twice at 68 °C, each time for 60 min in a solution 2 × SSC, 0.2% SDS and once for 30 min in 0.2 × SSC, 0.2% SDS solution. Thus the clones with A_3B_4 and $A_5A_4B_3$ glycinin subunit sequences that have nucleotide homology exceeding 90% were selected. Restriction analysis with HindIII and TaqI was carried out for clones with cDNA exceeding 700 bp. Clones, comprising A_3B_4 subunit cDNA were selected by using the restriction map for cDNA (complementary DNA) of the glycinine subunits A_3B_4 and $A_5A_4B_3$.

Determination of nucelotide sequences. The method of Sanger et al. (1977) modified by Messing (1983) was employed. Upon subcloning of ds-cDNA fragments in vectors M13 tg130 and tg131, SmaI, BamH1, Alu1, HaeIII and Sau 3a restrictases were used. All these revealed substitutions were tested by sequencing several independent M13-cloned fragments containing both DNA strands.

Results

Comparison of cultivar nucleotide sequences. The length and position of all cloned and sequenced ds-cDNAs are shown at Fig. 1. The four cloned sequences completely overlap the region encoding polypeptide B and only in the case of *G. soja* pSPG L1 sequence, is a small part of B_4 polypeptide C-terminal region dropped from analysis.

Figure 2 gives the nucleotide sequences of A_3B_4 glycinin cDNA for the cultivars "Mandarin" "Mukden" sequenced in this study and cultivar "Rannaya-10", cloned and sequenced by us earlier (Epishin et al. 1986). For comparison, the sequence of the "Mandarin" cultivar was taken as standard. In subunit A_3B_4 cDNA of "Mukden" and "Rannaya-10" cultivars, only two nucleotide substitutions in the region encoding polypeptide B_4 were found. One substitution was in the cDNA of the "Mukden" cultivar, where the second nucleotide of codon TTG was changed to C, leading to the substitution

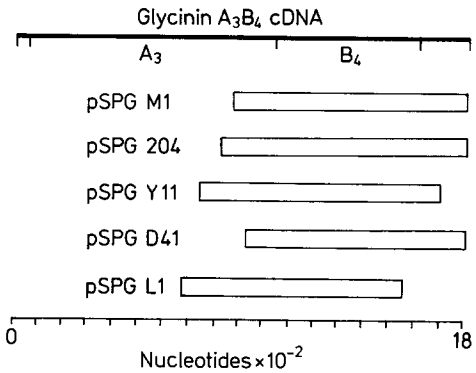

Fig. 1. Cloning and sequencing of ds-cDNA of A_3B_4 glycinin subunit. Plasmids pSPG M1, pSPG 204, pSPG Y11 comprised cloned cDNAs of A_3B_4 glycinin subunit of "Mandarin", "Rannaya-10" and "Mukden" cultivars, respectively. Plasmids pSPG D41 and pSPG L1 comprised cloned cDNA of wild soya *G. soya* introductions L582 and L23, respectively

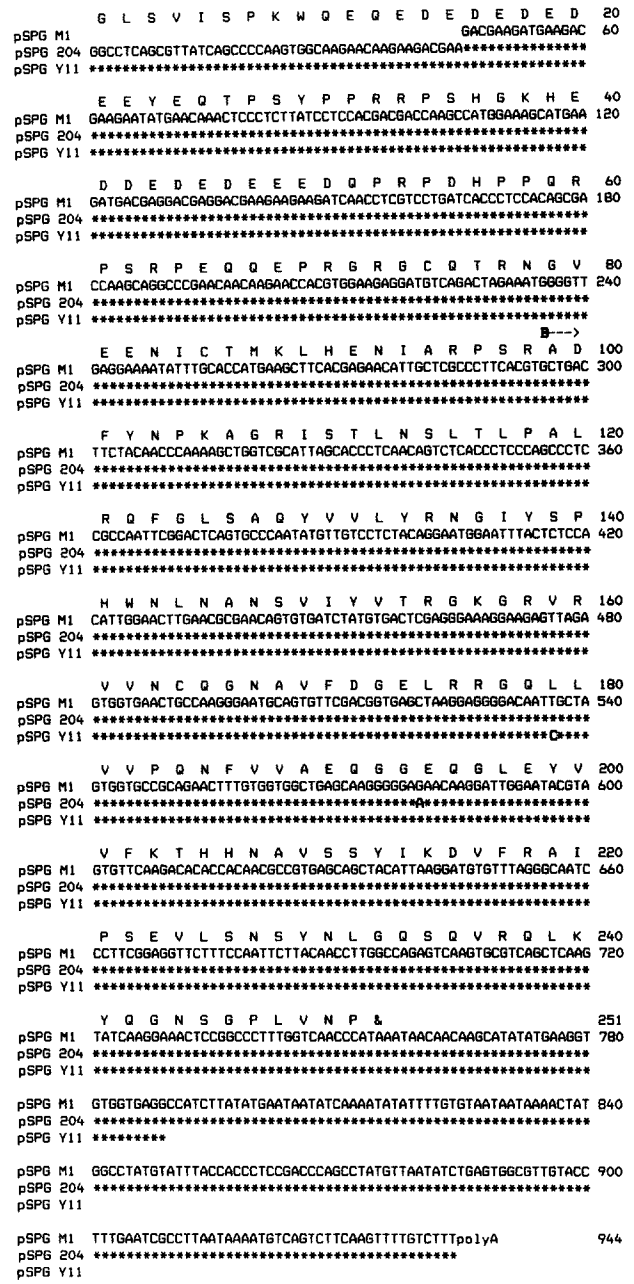

Fig. 2. Comparison of cDNA sequences encoding A_3B_4 glycinin subunit of soybean cultivars "Mandarin", "Rannaya-10" and "Mukden". The limit of B_4 polypeptide is underlined by *dotted arrow*

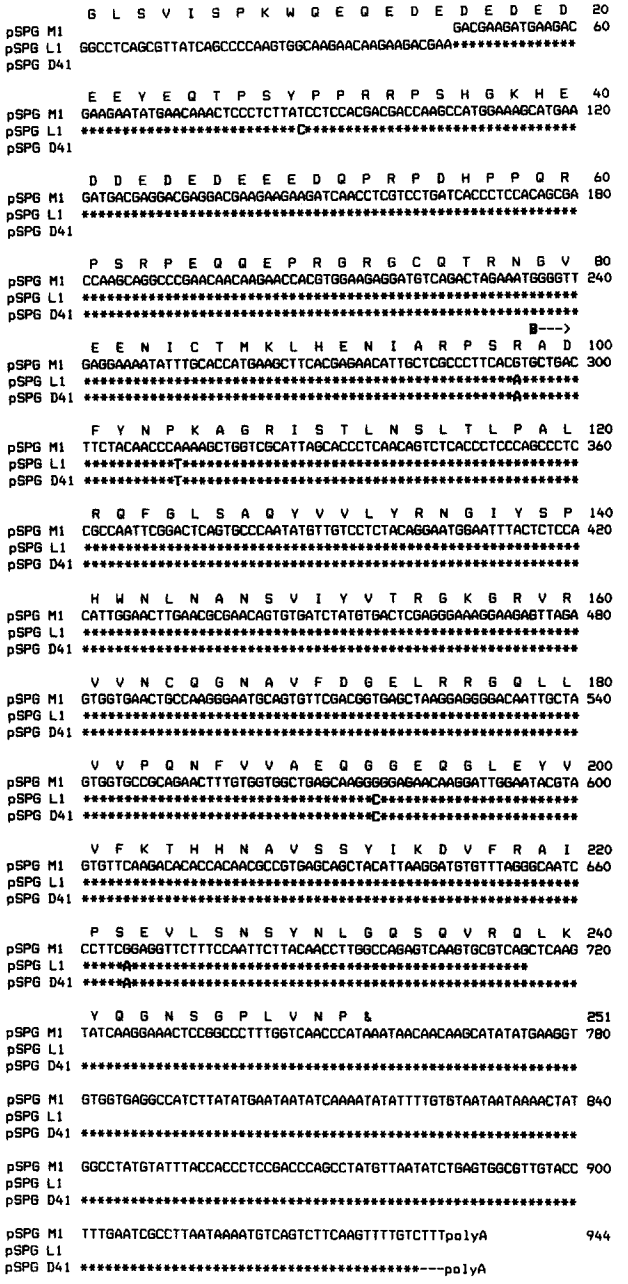

Fig. 3. Alignment of cDNA sequences encoding A_3B_4 glycinin subunit of two wild soybean introductions with that of soybean cultivar "Mandarin". The limit of B_4 polypeptide is underlined by *dotted arrow*

of leucine for serine. One more substitution was revealed in cultivar "Rannaya-10": the first nucleotide of the GAA codon was changed for A, leading to the substitution of glutamic acid for lysine.

Comparison of cDNA nucleotide sequences of wild soybean and cultivar "Mandarin". The nucleotide sequences within the cDNA regions of two plant introductions of wild soybean studied were completely identical (3). At the same time, the comparison with cDNA of cultivar "Mandarin" revealed that the wild soybean introductions differ in four nucleotides within the region encoding polypeptide B_4 and in one nucleotide encoding the C-terminal part of polypeptide A_3. The four substitutions were silent and one substitution in the second position of codon CGT led to the replacement of arginine for histidine.

Discussion

Polymorphism level of glycine subunits. Glycinin subunits from different cultivars of soybean were cloned and sequenced almost simultaneously in several laboratories (Momma et al. 1985a, b; Scallon et al. 1985; Fukazawa et al. 1985; Epishin et al. 1986; Zakharova et al. 1986). In these papers the sequences of cDNAs of four different soybean cultivars were published, including the cultivar "Rannaya-10". The comparison of the nucleotide sequences of the A_2B_{1a} subunit cDNA from two different cultivars revealed several substitutions. Studies on another subunit of the same cultivars also revealed substitutions (nine in the first case and eight in the second). Taking into account these comparisons, Utsumi et al. (1987a, b) came to the conclusion that polymorphism is a hereditary characteristic of the glycinin subunit genes.

In our previous work the sequences of cDNA of glycinin A_3B_4 and $A_5A_4B_3$ subunits and their counterparts from other soybean cultivars were compared (Zakharova et al. 1986). We found that the number of nucleotide substitutions in the A_3B_4 subunit cDNA is several times higher than in subunit $A_5A_4B_3$. This fact has had no reasonable explanation up to now, since the nucelotide sequences of these subunits (within the regions compared encoding polypeptides B_3 and B_4) have 90% homology. We propose that the variations in the sequencing method employed in different laboratories may be the source of variation in the number of registered substitutions. Thus, the conclusion of Utsumi et al. (1987a) concerning the polymorphism of the glycinin subunits' primary structure causes doubts.

Therefore, we studied the A_3B_4 cDNA sequences encoding polypeptide B_4 of the glycinin A_3B_4 subunit in two cultivars of Chinese origin, representative of the USA assortment. The comparison of these cultivars "Rannaya-10", selected in the USSR, showed that three cDNAs have only two nucleotide substitutions. These data do not support the conclusion on the expressed polymorphism of the cultured soybean glycinin subunits, at least not for the main B_4 polypeptide. We cannot but notice that Fuchman (1985) investigated leghemoglobin from 69 *G. max* cultivars and 18 plant introductions of *G. soja* and came to the conclusion that the discrepancies between the amino acid sequences published can be due to sequencing errors rather than to the leghemoglobin structural polymorphism.

Comparison of cDNA encoding polypeptide B_4 in G. max and G. soja. The nucleotide sequences of two independently collected wild soybean plant introductions were absolutely identical, although the material for this study was collected at different times in two regions of the Amur District of the USSR. However, it remains unclear to what extent these plant introductions reflect the populational structure of *G. soja.*

The cDNA sequences of wild and cultured soybeans differed in four nucleotides within the region encoding polypeptide B_4. Altogether, five substitutions were revealed in the investigation region of the A_3B_4 subunit gene. According to the diffeences in the cDNA sequneces, the polymorphism of polypeptide B_4 gene revealed by comparing wild and cultured soybean roughly corresponds to that among the cultivars of cultured soybean. Thus, the results of nucleotide sequencing add further proof to the concept that *G. soja* is the ancestor of cultured soybeans.

The time calculated from the rate of nucleotide divergence of the glycinin B_4 polypeptides significantly exceeds the longevity of soybean domestication. The mean rates of nucleotide substitution fixation in the basic polypeptide genes of plant 11S globulins were estimated to be 3.0×10^{-9} nucleotides/year, which corresponds to amino acid substitution fixation in the polypeptides of approximately $3.4-3.6 \times 10^{-9}$ site/year. Such estimates are close to those obtained for an "average" protein. Our data were obtained from analyses of homology of cDNA sequences and their amino acid derivatives (Aleksenko et al. 1989). On the basis of these molecular evolution parameters, the time period necessary for the accumulation of the estimated number of nucleotide substitutions is $0.9-2.4 \times 10^6$ years. Let us compare these results with the historic-archeological record.

During the early Chinese Yang-Shao culture (starting about 5000 B.C.) soybean was not among the cultivated plants, rice, hemp, mulberry and millet (*Sertania italica*) (Hymowitz 1970; Bellwood 1978); thus, the maximum period that passed after soybean domestication is no more than five to seven thousand years. This period is evidently shorter than the time required for the accumulation of the number of nucleotide substitutions revealed, if the calculation is based on the parameters of molecular evolution of plant 11S globulins.

Thus, the nucleotide substitutions we observe in cultivars could not have arisen during cultivation, but appeared beforehand.

Another explanation may be that the rate of molecular evolution in the process of soybean domestication increased, which is manifested in the increase of the nucleotide substitution rate in B_4 polypeptide gene. If we compare the calculated time of molecular evolution and the longevity of the cultivation period, this increase is 100- to 200-fold, but this hypothesis seems unreasonable. We are inclined to think that there may be other explanations of the results obtained. It seems possible that both cultured and wild soybean had a common ancestor, which gave rise to several lines. Within these lines to genes of B_4 polypeptide accumulated substitutions for one to two million years, which seems reasonable by

molecular evolution criteria. The lines of cultivated soybean were obtained by simultaneous introduction of various varieties of wild soybean lines. The process of selection resulted in the natural exhaustion of the gene pool. Therefore, the majority of present soybean cultivars have common progenitory with respect to the monogenic traits of the B_4 polypeptide structure and differ only slightly from one another. At the same time the wild soybean, apart from the selection process, could maintain a certain population variability of this polypeptide structure.

References

Aleksenko AYu, Sichkar VI, Timokhina EA, Musatova AA, Vinetski YuP (1985) Conservatism of the subunit composition of glycine from various samples of *Glycine max* and *Glycine ussuriensis*. Genetika 21:1185–1191 (in Russian)

Aleksenko AYu, Nikolaev IV, Vinetski YuP (1988) Soybean 11S globulin polypeptides have an antigenic homology with 11S globulins from various plants. Theor Appl Genet 76:143–147

Aleksenko AYu, Epishin SM, Zakharova ES, Vinetski YuP (1989) Phylogeny and molecular evolution rate of plant 11S globulins. J Evol Biochem Physiol (in press)

Bellwood P (1978) Man's conquest of the pacific. The prehistory of Southeast Asia and Oceania. Collins, Auckland p 5/7

Borroto K, Dure III L (1987) The globulin seed storage proteins of flowering plants are derived from 2 ancestral genes. Plant Mol Biol 8:113–131

Committee on Genetic Vulnerability on Major Crops (1972) Genetic vulnerability of major crops. Proc Natl Acad Sci USA

Doyle JJ (1988) 5S ribosomal gene variation in the soybean and its progenitor. Theor Appl Genet 75:621–624

Doyle JJ, Beachy RN (1985) Ribosomal gene variation in soybean and its relatives. Theor Appl Genet 70:369–376

Epishin SM, Kotenko SV, Zakharova ES, Kiselev SL, Vinetski YuP (1986) Cloning and structural analysis of cDNA encoding a portion of the A_3B_4 glycinin subunit. Genetika 22:1067–1072 (in Russian)

Fuchman WH (1985) Discrepancies among published amino acid sequences of soybean leghemoglobins: Experimental evidence against cultivar differences as the sources of the discrepancies. Arch Biochem Biophys 243:454–460

Fukazawa C, Momma T, Hirano H, Harada K, Udaka K (1985) Glycinin A_3B_4 mRNA: cloning and sequencing of double-stranded cDNA complementary to a soybean storage protein. J Biol Chem 260:6234–6239

Hanahan D (1983) Studies on transformation of *E. coli* with plasmids. J Mol Biol 166:557–587

Hayashi M, Mori H, Nishimura M, Anazawa T, Hara-Nishimura I (1987) Nucleotide sequence of cloned cDNA coding for pumpkin 11S globulin subunit. Eur J Biochem 172:627–632

Hermann FJ (1962) A revision of the genus *Glycine* and its immediate allies. US Dep Agric Bull 1268:1–10

Higuchi W, Fukazawa C (1987) A rice alutelin and soybean glycinin have evolved from a common ancestor gene. Gene 55:245–253

Hymowitz T (1970) On domestication of the soybean. Econ Bot 24:408–421

Maniatis T, Fritsch EF, Sambrook J (1982) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor/NY

Messing J (1983) New M13 vectors for cloning. Meth Enzymol 101:20–78

Momma T, Negoro T, Hirano H, Matsumoto A, Udaka K, Fukazawa C (1985a) Glycinin $A_5A_4B_3$ mRNA: cDNA cloning and nucleotide sequencing of a splitting storage protein subunit of soybean. Eur J Biochem 149:491–496

Momma T, Negoro T, Udaka K, Fukazawa C (1985b) A complete cDNA coding for the sequence of glycinin A_2B_{1a} subunit precursor. FEBS Lett 188:117–122

Nielsen NC (1984) The chemistry of legume storage proteins. Philos Trans R Soc London Ser B 304:287–296

Sanger F, Nicklen S, Coulson AR (1987) DNA sequencing with chain terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Scallon B, Thanh VH, Floener LA, Nielsen NC (1985) Identification and characterization of DNA clones encoding group II glycinin subunits. Theor Appl Genet 70:510–519

Utsumi S, Kim C-S, Kohno M, Kito M (1987a) Polymorphism and expression of cDNAs encoding glycinin subunits. Agric Biol Chem 51:3267–3273

Utsumi S, Kohno M, Mori T, Kito M (1987b) An alternate cDNA encoding glycinin $A_{1a}B_x$ subunit. J Agric Food Chem 35:210–214

Vavilov NI (1926) Studies on the origin of cultivated plants. Cit. by: Vivilov NI (1987) The origin and geography of cultivated plants. Nauka, Leningrad, pp. 26–126 (in Russian)

Walburg G, Larkins BA (1986) Isolation and characterization of cDNAs encoding oat 12s globulin mRNAs. Plant Mol Biol 6:161–169

Zakharova ES, Kotenko SV, Lomakin IB, Epishin SM, Vinetski YuP (1986) Nucleotide sequences of the glycinin subunits cDNA (subfamily A_3): different intraspecific variability deduced from comparative study. Genetika 22:2741–2749 (in Russian)