Phaseolin: Structure and Evolution

Chandrakanth Emani and Timothy C. Hall*

Institute of Developmental and Molecular Biology; Department of Biology, Texas A&M University, College Station, TX 77843-3155, USA

Abstract: Phaseolin is the salt-soluble glycoprotein or the group of polypeptides of the French bean (*Phaseolus vulgaris* L.) that account for some 50% of the total protein in mature bean seeds. It was one of the first plant proteins to be translated *in vitro* from mRNA and one of the first plant genes isolated. It was also the first developmentally regulated plant gene to be expressed in a heterologous plant species through *Agrobacterium*-mediated transformation. Studies on phaseolin have provided insight to many aspects of plant protein synthesis, from fundamental molecular mechanisms to practical goals such as the improvement of the French bean's nutritional quality. The present review is a comprehensive account of the structural and functional features of phaseolin that have implications regarding its evolution. Additionally, future directions in phaseolin evolutionary studies and suggestions regarding effective and safe biotechnological approaches for the nutritional improvement of French bean seed are outlined.

Keywords: Phaseolin, seed storage proteins, seed protein evolution.

INTRODUCTION – HISTORICAL BACKGROUND

Understanding of the fundamental concepts leading to human intervention in plant culture that resulted in specific crops and farming practices appears to have initiated about 20,000 years ago, laying a vital foundation for mankind's civilization. Neolithic man (or, perhaps, more correctly, woman) developed the processes of seed collection, storage and planting that underlie systematic agricultural practices [1]. Seed proteins are intrinsic to these processes as their regulated degradation and assimilation enables proper seedling germination. They represent major food resources that can be stored and used over winter both for man and his livestock, further establishing civilization by enabling the development of stable communities. Archaeological evidence of domesticated grain seeds such as wheat and barley was found in 7000 year old dwellings and graves in Egypt [2].

The first accurate botanical descriptions of the common bean or French bean (*Phaseolus vulgaris* L.) are in European herbals dating back to 1542. Archaeological remains of bean domestication of over 7000 years antiquity were found in the Tehuacan Valley of Mexico [3]. The discovery of the wild common bean in Argentina [4] and Guatemala [5] along with the findings of archaeological remains in the Americas [6] finally established the origins of bean to the Americas, rather than India, as was ascribed by Linnaeus. French bean is an important legume crop with protein contents (>20%) higher than those of cereals like rice or wheat (<15%) [7]. The earliest recorded investigations related to seed protein composition in bean were attributed to Ritthausen [8]. Modern studies on seed proteins leading to immunological and crystallographic research were made possible by the ground-breaking contributions of Thomas Osborne, ably assisted by George F. Campbell [reviewed in 9]. Osborne's classification of seed proteins as globulins, prolamins, glutelins and gliadins was based on differential solubility in aqueous and non-aqueous solutions of various pH levels [10]. While our current knowledge related to seed proteins has extended to the realms of amino acid sequence homologies and differences, the pioneering foundations laid down by Osborne [9] still serve as valuable operational definitions. Osborne [11] was the first to realize that while French bean seeds contained globulin proteins, the major storage protein differed significantly from legumin and vicilin of pea and soybean in terms of solubility, heat stability and chemical composition. Osborne named this major seed storage protein of French bean as "Phaseolin."

PHASEOLIN – THE EARLY YEARS

McLeester *et al.* [12] first recognized that phaseolin can be readily precipitated by dilution from a crude acidic saline solution, thus providing a simple technique for isolating an essentially pure form of the protein. The ability to obtain large quantities of purified phaseolin facilitated electrophoretic studies using one dimensional SDS gels [12] and characterization by equilibrium sedimentation [13]. These initial forays paved the way for biophysical characterization [14-17] and the subsequent crystallographic structural analysis [18-21].

A meticulous study employing electrophoretic analysis to observe protein accumulation during seed development [14] showed the discrete onset of phaseolin synthesis some 12 days after anthesis followed by a dramatic rapid accumulation over the following week. These results suggested the use of developing seed as an ideal tissue for the isolation of mRNA, leading to phaseolin being one of the first plant proteins to be translated *in vitro* [22]. Important insights into the role of glycosylation were obtained from this study. The wheat germ extracts used in mRNA translation *in vitro* were devoid of membranes, resulting in non-glycosylated products

*Address correspondence to this author at the Institute of Developmental and Molecular Biology - Mail Stop 3155, Texas A&M University, Biological Sciences Building West Suite 403, College Station, TX 77843-3155, USA; Tel: 979-845-7750; Fax: 979-862-4098; E-mail: tim@idmb.tamu.edu

1874-4044/08

that migrated faster than the native glycosylated polypeptides during electrophoresis. The study also provided evidence that the non-glycosylated phaseolin polypeptides are more susceptible to proteolytic degradation than the native forms.

The ability to isolate substantial amounts of mRNA led to the cloning of phaseolin as a cDNA [23]; subsequently, nick-translated cDNA clones were used as probes to identify *phas* genomic clones generated in lambdaphage libraries, resulting in the isolation of the β-phaseolin gene as one of the first plant genes to be cloned. Although most cDNA clones were not full-length, sequence comparison with that of the β-phaseolin gene resulted in the first demonstration of introns in a plant gene [24]. Subsequently, full-length cDNA and genomic clones were isolated and characterized [24, 25]. The availability of the complete nucleotide sequence of phaseolin was important in revealing its amino acid sequence, an essential step towards the twin goals of improving digestibility by eliminating the asparagine residues at which the *N*-glycans were linked and improving the nutritional value of the protein by the insertion of additional methionine codons into the β-phaseolin gene [26]. Attainment of these goals required the construction of clones bearing the desired sequence alterations and the exciting but, at that time, very challenging, transfer to a suitable host plant. In a keen competition, a partial β-phaseolin clone was initially transferred into sunflower callus *via* the tumor-inducing plasmid vector *Agrobacterium tumefaciens* [27]. Subsequently, the full-length genomic sequence was transferred to fertile tobacco plants and substantial quantities of phaseolin accumulated in the seeds but none was detected in leaves, mimicking the spatial and temporal regulation seen in *P. vulgaris*. This was the first unequivocal demonstration of the transfer of a developmentally regulated plant gene from one plant species to another and its spatially-correct expression [28, 29].

PHASEOLIN PROTEIN STRUCTURE

The biophysical characterization of phaseolin, including crystallography, and its post-translational processing and intracellular trafficking, can be found in our earlier extensive review [30]. Therefore, the following focuses on important protein structural and functional features that are relevant to phaseolin evolution.

Phaseolin was initially characterized as a globulin on the basis of its solubility [11]. Using equilibrium sedimentation, Sun *et al.* [13]showed that phaseolin reversibly associates from a monomeric form (3S; M_r ~50,000) at alkaline pH through a trimeric configuration (7.1 S, M_r 150,000) at neutral pH to a dodecameric form (18.2S; M_r 596,000) at pH 4.5. The dodecameric form is readily soluble in 0.5 NaCl, pH 4.5, but decreasing the salinity to 0.1 M results in a flocculent white precipitate. Even though earlier researchers knew that there were significant differences between vicilin and phaseolin, Derbyshire *et al.* [31] concluded on the basis of its sedimentation coefficient of 7S at neutral pH that phaseolin belongs to the 7S or 'vicilin-like' family of legume seed storage proteins. Using optical rotatory dispersion and circular dichroism, Blagrove *et al.* [19] confirmed the observations of Sun *et al.* [13] and predicted the presence in phaseolin of a low portion of α-helices, coupled with a high level of β-sheets in its secondary structure, which was subsequently confirmed by Plietz *et al.* [18]. The definitive crystal structure of phaseolin was solved by Lawrence *et al.* [20, 21] who elucidated the structural features as follows: "the trimeric protein has its monomers arranged around a 3-fold symmetric axis and each monomer has two α and β modules related by a pseudo-dyad perpendicular to the trimeric 3-fold axis. Each module further contains a b-barrel with an elaborated jelly-roll folding motif followed by a α-helical domain comprising of three helices, two of which exhibit a helix-turn-helix motif. The *N*-terminal α-helical domain is linked to the *C*-terminal β-barrel through a segment containing the fourth helix followed by a putatively extended portion." These three-dimensional structural features led Lawrence *et al.* [21] to postulate a canonical structure for 7S globulins, and it was also observed that structural features of bean phaseolin and jack bean canavalin show considerable similarity. Further, Lawrence *et al.* [21] also laid down the basis for interpreting the structure of the 11S family seed storage proteins for which X-ray diffraction data are not available.

Nascent phaseolin contains a signal peptide of 24 amino acids, spanning from Met^{1} to Ala^{24} with positively charged arginyl residues at positions 2 and 4, followed by a long stretch of hydrophobic amino acids that is characteristic of signal peptides that aid in the sequestration of soluble proteins in endoplasmic reticulum [30]. Two asparagine residues Asn^{252} and Asn^{341} [24] are present in canonical Asn-X-Ser/Thr *N*-glycosylation sites [32]. Phaseolin polypeptides with two glycans have $Man_7(GlcNAc)_2$ attached to Asn^{252} and $Man_9(GlcNAc)_2$ attached to Asn^{341}, whereas polypeptides with only one glycan have a complex oligosaccharide Xylose-$Man_3(GlcNAc)_2$ at Asn^{252} [32]. The differential glycosylation at Asn^{341} seems to control the presence of the oligosaccharide residues that in turn is responsible for the heterogeneity among the phaseolin polypeptides seen as α (M_r 51,000-53,000), β (M_r 47,000-48,000) and γ (M_r 43,000-46,000) [33-35].

PHASEOLIN GENE STRUCTURE

Phaseolin electrophoretic profiles revealed existence of three molecular weight variants of dissociated peptides termed α,β and γ [12]. The evaluation of cultivars showed three distinct patterns [15] that were named T, C and S after the cultivars Tendergreen (edible bean pods), Sanilac (dry beans) and Contender (another dry bean cultivar). The number of copies of the phaseolin gene present in the genome remains somewhat ambiguous. Whereas reconstruction Southern blots indicated that 10-15 copies are present in Tendergreen, Sanilac and Contender cultivars [36], only 6-8 copies were predicted for Tendergreen genomic DNA from solution hybridization kinetics and genomic DNA blot hybridization [37]. The banding patterns obtained by single and two-dimensional gel electrophoretic analyses, indicate that Contender is a composite of Tendergreen and Sanilac types, showing a good agreement between the number of phaseolin protomers and structural genes. Genetic analysis revealed a tight linkage for the entire phaseolin gene family [17]. Heteroduplex analysis of the phaseolin genomic clones [36] revealed that whereas the phaseolin DNA coding sequences were similar, considerable divergence occurred close to their 5' and 3' termini [37].

The first report of phaseolin gene structure involved comparison of the partial sequence of a cDNA and the corresponding genomic DNA [24]. This revealed for the first

time, the presence of introns in plant DNA, namely IVS-A (88bp), IVS-B (124bp) and IVS-C (129bp). The intervening sequences all begin with nucleotides GT and end with AG [24] similar to those of animal and virus genes [38]. The sequence conservation in regions adjacent to the 5' and 3' intron boundary sequences suggested an involvement in recognition by enzymes responsible for intron excision and exon splicing [39]. However, 6 of the 11 nucleotides at the 3'end of IVS-C differ from the generalized sequence [24].

The complete nucleotide sequence of phaseolin [25] revealed a gene structure that includes a total of 1990 bp distributed as: 80 bp of 5' untranslated region, 1263 bp of protein-encoding DNA interrupted by five introns (a total of 515 bp: IVS1, 72bp; IVS2, 88bp; IVS3, 124 bp; IVS4 128 bp and IVS5, 103 bp) and 135 bp of 3' untranslated DNA. Thus, the original mRNA transcript of 1990 bp must be processed by five or more RNA splicing events to result in a 1475 bp mRNA molecule. Three TATA box sequences are located upstream from the mRNA cap at positions -28, -37 and -39. Slightom *et al.* [25] postulated that the sequence at -28 was likely to be the most important in driving high expression from the *phas* promoter; this was later confirmed by Grace *et al.* [40] who used an *in vitro* transcription system that revealed that the sequence and spacing of the TATA box elements are critical for accurate initiation from the β-phaseolin promoter. The A+T content of the introns is similar to that in soybean proteins, but is considerably higher than those of non-plant species. Two CCAAT box-like sequences are located at positions -67 (CCAT) and -74 (CCAAAT), similar to those of soybean leghemoglobin genes [41]. Zein, the major storage prolamin storage protein of maize [42] is encoded by a family of highly similar genes bearing a TATA motif at -32, similar to that of most eukaryotes. In contrast, the zein CCAT sequence is located at -112 bp [25]. The hexanucleotide poly (A) addition signal sequence AATAAA is located 16 bp 5' to the first nucleotide of poly (A).

Complex *in vivo* footprinting profiles of the *phas* promoter in transgenic tobacco seeds led to the hypothesis that individual *cis* elements possess autarkical functions in disparate modules of the embryo [43]. Transcriptional activation of *phas* genes [43-45] revealed that specific promoter regions confer expression in discrete modules such as the radicle, hypocotyl, or cotyledons of the embryo. Site-directed substitution mutations of 10 locations within the *-295phas* promoter, made to explore these module-specific factor-DNA interactions [46], revealed that only 2.6% of the promoter activity remained after the mutation of the G-box. In contrast, high levels of expression in embryo tissues were retained after mutation of specific CCAAAT box, E-box and RY elements. The proximal (-70 to -64) RY motif was found to bestow expression in the hypocotyl while all the RY elements contributed to expression in cotyledons but not to the vascular tissue during embryogenesis. RY elements at positions -277 to -271, -260 to -254 and -237 to -231 orchestrated radicle-specific expression. The study also established that the G-box (-248 to -243) is a functional abscisic acid responsive element and the E-site (-163 to -158) is probably a coupling element. The similar patterns of expression from the *phas* promoter in transgenic tobacco and *Arabidopsis*, two distantly related plants, provide evidence for a generality of function for the observed factor-element interactions.

PHASEOLIN POLYPEPTIDES AS EVOLUTIONARY MARKERS – FRENCH BEAN DOMESTICATION AND DISPERSAL

Electrophoretic analysis of seed proteins has proven to be a valuable tool in tracing the evolution of crop plants, especially for identification of the wild progenitors of the respective crops and gathering additional information on the evolutionary and domestication patterns [47, 48]. The presumed ancestral forms and evolutionary patterns of chickpea [49], maize [50], wheat [51] and soybean [52, 53] have been identified and established by studying the electrophoretic variability of their respective seed proteins. Gepts [54] used phaseolin as an evolutionary marker in his insightful review of the domestication pattern and world-wide dispersal of the French bean.

The characteristic structural and functional features of phaseolin render it as a useful evolutionary marker. As the major seed storage protein of French bean, it accounts for 50% of the total protein stored in the cotyledons [36], and 35-46% of the total seed nitrogen [12, 55]. It is now rigorously established that spatial regulation of phaseolin expression is mandated through a combination of epigenetic [56, 57, 58] and genetic [46] events (chromatin structure and transcription factors). Phaseolin levels were found to be positively correlated with total available methionine levels [59]. These characteristics, along with properties such as the genotype, influence the amount and rate of storage protein accumulation [15] such that the concentration of phaseolin shows a positive correlation with that of the total cotyledon protein. That both traits respond similarly to selection render phaseolin as a major determinant of the quantitative and qualitative protein composition of bean seeds [54].

The cluster of closely related genes coding for phaseolin [37] may have arisen by successive duplications of an ancestral gene followed by divergence [54]. The divergence process included the insertions, duplications and deletions as demonstrated by the presence and absence of direct repeats [36, 60-62], and point deletions represented by nucleotide substitutions [33, 36]. In addition to the divergence observed at the DNA level, co- and post-translational modifications, including signal peptide cleavage upon the polypeptide transit into the lumen of endoplasmic reticulum [35, 36], glycosylation of polypeptides leading to variation in the polysaccharide side chains [34, 35] and amino acid substitutions leading to charge differences [36] resulted in the formation of a group of similar, but slightly heterogeneous phaseolin polypeptides. The electrophoretic patterns of these molecular entities in terms of molecular weight and isoelectric point changes reflect genotypic divergence and have been widely used in analyzing evolutionary relationships among bean cultivars [54].

Phaseolin proved to be a valuable evolutionary marker mainly owing to the complexity of the sequence of molecular events that lead to the variable electrophoretic patterns. It is highly improbable that such patterns would arise at different geographical locations or at different times, which suggests that each phaseolin type is unique and would have arisen only once in the evolutionary history of the French bean [54]. This property of uniqueness in pattern rendered phaseolin as a useful tool in tracing the domestication patterns of French bean cultivars [54]. Gepts [54] followed the

domestication and dissemination patterns of French bean and found that domestication occurred repeatedly along the distribution range of its wild relative. The Middle American domestication gave rise to small-seeded, 'S' phaseolin cultivars, while large-seeded, 'T' phaseolin (and possibly 'A', 'C' and 'H') cultivars were seen in southern Andes. The Colombian domestication rendered small-seeded, 'B' phaseolin cultivars. Dispersal of these domesticates then occurred to the rest of the Americas, Europe, Africa and the Caribbean. Based on the molecular complexity principle, Gepts [54] suggested a well-defined region in the west-central Mexico as the actual Middle American domestication center. The phaseolin data on domestication and dispersal of French bean cultivars was consistent with the existing archeological, botanical, historical and linguistic data (Gepts, 1988) [54]. This report, and that of Hall *et al.* [36], initiated the use of molecular complexity to explain the value of a gene cluster and its product as a marker in crop evolution. Gepts [54] suggested that other seed proteins encoded by multi-gene families such as legumin and vicilin in pea, conglycinin and glycinin in soybean, zein in maize, B-hordein and amylase in barley can be explored from evolutionary aspects in a similar manner, as, indeed, can non-seed protein, and also other multi-gene proteins such as leghemoglobin, chlorophyll a/b binding protein and glutamine synthetase. Gepts [54] draws parallels to research exploiting the molecular complexity principle to identify the geographic origin of sickle-cell anemia mutations [63, 64] and the mapping of the human β-globin gene cluster by restriction endonuclease analysis. The identification of the 'S', 'T' and 'C' phaseolin gene sequences was by EcoRI restriction polymorphisms [37].

DOMAIN DUPLICATION IN PHASEOLIN EVOLUTION

Gibbs *et al.* [65] examined the primary sequences of the jack bean protein canavalin along with other vicilin-type proteins: pea vicilin, French bean phaseolin, and soybean conglycinin; legumin-type sequences: pea legumin and four sequences of soybean glycinins. The hypothesis that the pseudodyad seen in the three-dimensional structure of the canavalin, pea vicilin and phaseolin arises from an ancestral gene duplication was tested by comparing the sequences using the computer programs based on the FASTP algorithm [66]. An ancient sequence duplication was found to account for 80% of the amino acid residues in canavalin of jack bean and the orthologous proteins phaseolin and pea vicilin [65]. The observed sequence duplication was also stated to adequately account for the presence of a pseudodyad axis in the crystalline protein. Gibbs *et al.* [65] also searched the National Biomedical Research Foundation (NBRF) protein data base for sequences similar to canavalin and found that best scores were obtained with phaseolin and pea vicilin, and a significant match was found for a partial pea legumin sequence. A library constructed with known legumin and vicilin sequences was then examined for representative sequences of each family, and significant matches and similarities in sequences were observed between the vicilin and legumin families. Gibbs *et al.* [65] concluded from these observations that there appears to have been a common precursor to much of the legumin and vicilin sequences, that included part of the *N*-terminal repeat and the entire *C*-terminal region of the vicilin-type proteins including phaseolin. To account for the sequence similarity findings between vicilin and legumin families, the researchers proposed an evolutionary scheme, in which an ancestral gene encoding one copy of the repeat domain first underwent duplication, by either homologous recombination or, more likely, an unequal crossing over to yield a gene similar in structure to the modern vicilins. The duplication of this gene would enable one copy to evolve as a vicilin. To support their hypothesis, the authors point to the fact that evidence for a gene family for vicilin proteins is seen in *Phaseolus vulgaris* [25] and *Glycine max* [67]. The other copy of the domain was hypothesized to yield the legumin family. The authors further contend that at least one domain of the legumin subunit, as well as the internal redundant domains of vicilins are derived from a common precursor. This putative evolutionary scheme was also used to explain certain physical properties of the proteins, such as the apparent dyad axis in vicilins.

PHASEOLIN NUCLEOTIDE SEQUENCE DIVERSITY – PRESUMED ANCESTRAL SEQUENCES

Hall *et al.* [36] characterized nine phaseolin cDNAs that revealed a high degree of sequence conservation, and the molecular weight differences between α and β forms were found to result from the absence in β-phaseolin of two direct repeat sequences: a 15 bp repeat in exon 4 and a 27 bp repeat in exon 6. The cultivar Sanilac contained the 27 bp repeat, but not the 15 bp repeat [60]. Lines containing at least one repeat are considered α types.

According to Kami and Gepts [61], the presence of repeats and smaller imperfect duplications in all the reported phaseolin sequences implies that repeated sequences were formed prior to the expansion of a phaseolin progenitor into a multigene family, with the repeats arising repeatedly during phaseolin evolution. The study also examined amino acid replacements in diverse phaseolin sequences, and it was suggested that divergence of α and β phaseolin genes predates the divergence of S and T phaseolins. Further, the mutations responsible for amino acid replacements must have occurred after the initial duplication of the original gene and formation of the 27 bp repeat, but before the divergence of the S and T phaseolins and a subsequent introduction of the 15 bp repeat [61]. The study also opined that since it is plausible that repeats were generated from preexisting sequences, the simplest sequences of the β-phaseolins lacking both the 15 and 27 bp repeats may be considered the progenitors of α-phaseolin genes. The presence of this progenitor sequence in both S and T multigene families further indicates that β-phaseolins may have undergone a duplication event followed by the introduction of repeated sequences. Further, since the 27 bp repeat is present in both S and T varieties, its introduction might have predated the divergence of *P. vulgaris* into its two major geographic, Middle American and Andean, gene pools [54, 68]. The introduction of the 15 bp repeat occurred later among the Andean α-phaseolin sequences, and was followed by additional duplications that further expanded the gene family [61]. PCR analysis conducted in our lab (G. Li and T.C. Hall, unpublished) and by the Gepts group revealed that all permutations of repeats can be detected. The specific amplification and sequencing of members of the phaseolin multigene family provided evidence for the accumulation of tandem direct repeats in both introns and exons during its evolution [62]. This study identified *P. vul-*

garis cv. Inca as a possible ancestral line since its I-type phaseolin genes (designated I-type by the Gepts group [62], based on SDS-PAGE protein profile studies) lack both repeats. A third 21 bp repeat in intron 3 was found to be present only in several nearly extinct wild bean populations in Peru and Ecuador.

PHASEOLIN SEQUENCE STUDIES – IMPLICATIONS FOR NUTRITIONAL IMPROVEMENT

The phaseolin direct repeat studies not only have important implications in phaseolin evolution, but in nutritional improvement of the French bean and legumes in general, as the sequence analysis of various phaseolin types make it possible to deduce potential sites for amino acid replacement to improve the overall methionine content of phaseolin [55, 59]. Hall *et al.* [26] also stated that a thorough characterization of the phaseolin sequence can be aimed at the ultimate goals of improving both *in vivo* and *in vitro* enzyme digestibility by eliminating the asparagine residues to which the *N*-glycans are linked and improving nutritional value by inserting additional methionine codons into the β-phaseolin gene.

Gepts and Bliss [59] demonstrated that nutritional availability of methionine in French bean positively correlated with phaseolin content. Since it was shown that S phaseolin can provide a higher nutritional value than T due to increased methionine residues [60, 61], the smaller S α-phaseolin provides a higher molar ratio of methionines than its larger T phaseolin counterpart, and may prove to be valuable material for sequence modifications to enhance phaseolin methionine content [61]. Future sequence comparison and analysis studies involving phaseolin can take advantage of the deposited sequences [25, 33] and the ever increasing deposits in the sequence databases to conduct multiple alignment studies aimed at revealing additional amino acid sites potentially amenable for sequence modifications to improve amino acid balance for nutritional improvement of French bean.

PHASEOLIN'S ANCESTRAL PRECURSOR?

The evolutionary conservation of phaseolin within the genus *Phaseolus* is striking, as seen from the cDNA sequence comparisons [36] and the direct repeat analysis [60, 62]. The seed storage globulins of the legumin and vicilin type exhibit widespread existence both in angiosperms and gymnosperms. Crystallographic studies of phaseolin [21] and sequence comparisons of legumin and vicilin-like seed storage proteins [69] revealed the existence of a characteristic framework of highly conserved amino acid positions as well as the partial conservation of the exon-intron structure between legumin and vicilin subunits. This suggested that

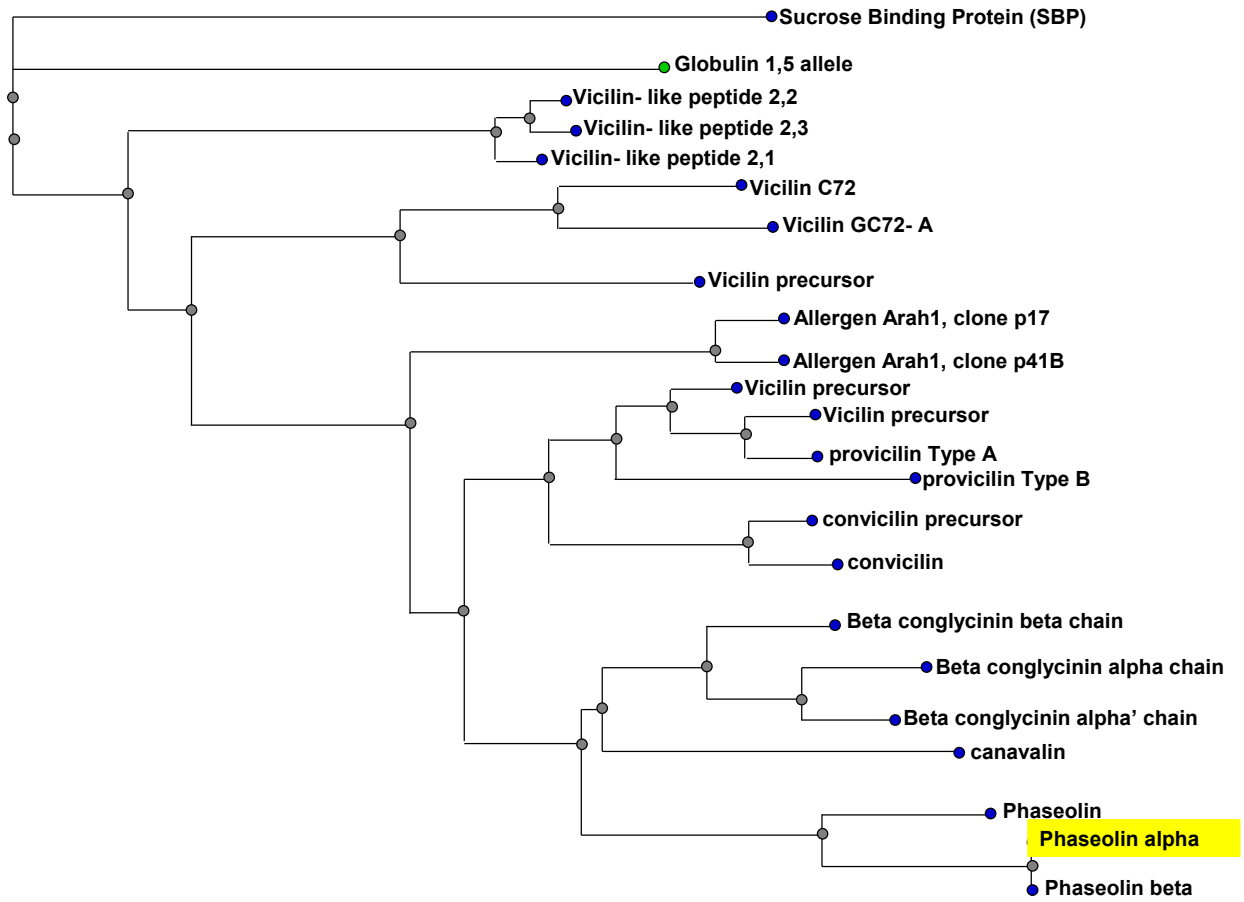

Fig. (1). Distance tree of phaseolin and its homologous polypeptides automatically generated by a blast-p search. The highlighted sequence is the input sequence of phaseolin alpha-type precursor used in the analysis to retrieve the homologous sequences. Events at the node points may give insight to functional changes contributing to the evolutionary history of the polypeptide. The putative ancestral precursor in this coarse distance tree is the soybean sucrose binding protein (SBP).

seed globulin genes are derivatives of a common single-domain ancestor and the evolution from this ancestor was by an early duplication or triplication event [69]. The highly conserved amino acid positions in modern seed storage proteins trace back their ancestors to spherulin-like proteins of myxomycetes probably involved in basic cellular dessication/hydration processes [70]. Braun *et al.* [71] isolated and characterized a vicilin-like gene expressed in the cycad *Zamia furfuraceae*. Sequence comparisons revealed remarkable similarities to a sucrose-binding protein (SBP) of soybean, *Glycine max*. Among the highly conserved amino acids in vicilin-like proteins, 24 out of 28 residues were recognized in the SBP sequence of *Zamia*. The single most important difference was seen in the characteristic β-bulge of the C-strand of vicilin domains where the proline residue of the vicilins is replaced by an unconserved isoleucine in case of the SBP C-terminal, but not the N-terminal domain [71]. Shutov and Baumlein [72] conducted extensive sequence comparisons and analyses of gene structures of legumins, vicilins, germins and spherulins to reconstruct an evolutionary pathway for seed storage globulins. The model proposed that a prokaryotic ancestral molecule involved in basic cellular desiccation/ hydration processes evolved into molecules with the basic features of extant single-domain plant germins and fungal spherulins. This molecule evolved into the two-domain globulins by a single duplication event and the storage globulin ancestor might have been recruited from a limited set of developmentally-regulated proteins specific for tissues which tolerate desiccation and rapid dehydration [72]. An extant fern-specific vicilin-like protein seems to fit in the structural description as a two-domain progenitor common ancestor of the storage globulins vicilin and legumin as well as several related non-storage proteins [72]. Based on these observations and the fact that there are sequence similarities between canavalin and phaseolin, and the common ancestor from a domain-duplication event [65], it is tempting to think that an SBP-like protein might be a prime candidate for a possible ancestral precursor of phaseolin (as shown in a preliminary coarse distance tree generated from a blastp search, Fig. **1**). More recent studies by Khuri *et al.* [73] trace back the evolution of seed proteins to the cupin superfamily [74], a small group of functionally diverse proteins found in all three kingdoms of life, Archaea, Eubacteria and Eukaryota. The conserved domain seen in these proteins is the characteristic six-stranded β-barrel structure, based on which seed storage proteins are termed as bi-cupins (Fig. **2**) because of a two-domain structure [73]. It remains to be seen whether the evolution of phaseolin can be traced back to a prokaryotic ancestor based on the cupin superfamily evolution pathway. Fig. (**3**) illustrates a hypothetical scenario for phaseolin evolution from our present knowledge of seed protein evolution [69-74].

CONCLUSION

Recapitulating the evolutionary pathway of phaseolin should now take advantage of the wealth of sequence, genome and structural data available today. For example, Fig. (**1**) shows a coarse distance tree of phaseolin and homologous sequences mostly within the papilionoidea subfamily of plants where distinct monophyletic groups are identified. Using such a tree with a carefully constructed multiple alignment, the different clades can be examined for the corresponding changes in their protein sequences and structure with particular emphasis paid to the sequence-structure correspondence of the changes. The resulting comprehensive evolutionary pathway can be utilized to investigate several interesting questions. The possibility that phaseolin might have roles other than seed storage can be examined. An important property that needs additional detailed analysis is the role of glycosylation in phaseolin and its relatives. The evolution of introns and promoter structures in phaseolin and its correspondence with the phylogenetic profile might gather insights into various important structural features. Another important area of investigation is to examine the phylogenetic profile of the proteases that cleave phaseolin and its relationships vis-à-vis the phaseolin-like seed storage proteins. The insight gained from deciphering such important structural and functional features might define/predict specific biotechnological approaches for the nutritional improvement of French bean.

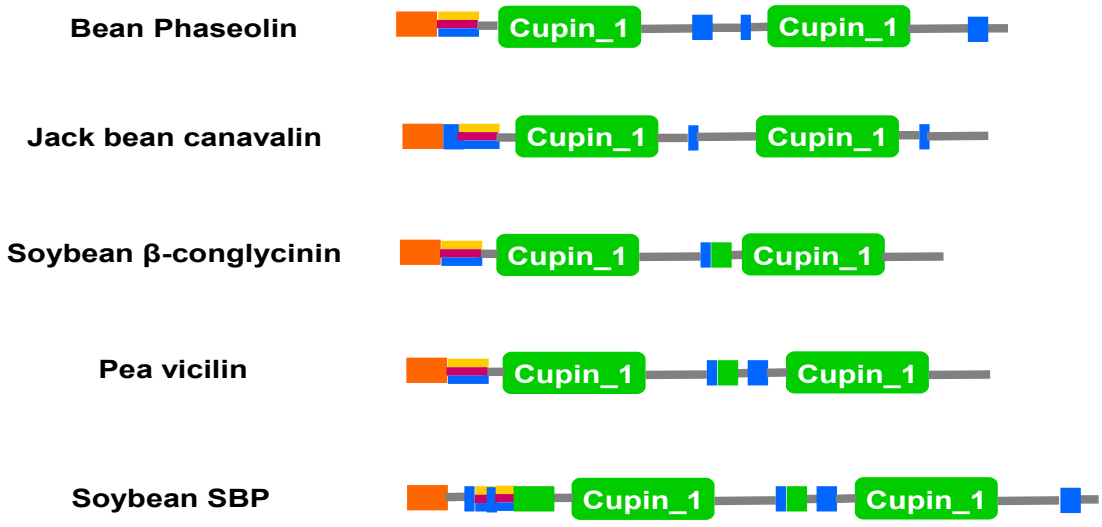

Fig. (2). Domain structure of phaseolin and related seed storage proteins (including the ancestral soybean sucrose-binding protein precursor) exhibiting the bi-cupin architecture (from the Pfam database, http://pfam.sanger.ac.uk/). Key for the colors according to the pfam database: Orange-signal peptides; small rectangle with yellow, purple and cyan stripes-unannotated clusters; cyan-low-complexity region; lime green-coiled coils; grey-base sequence.

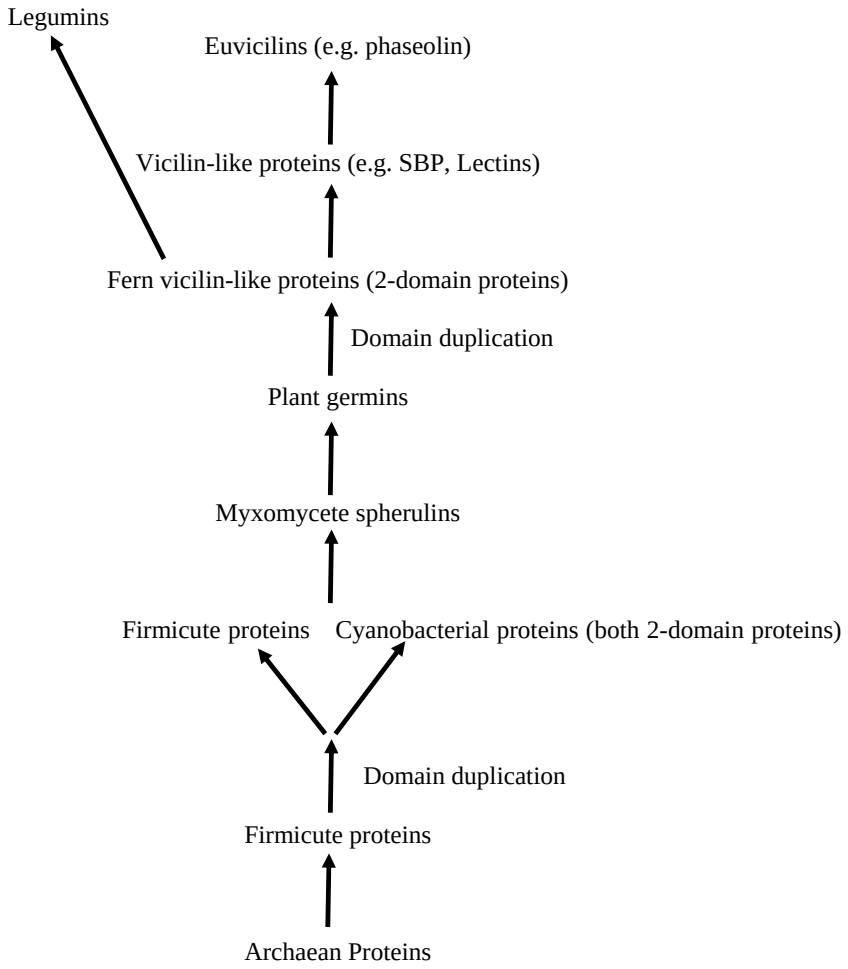

Fig. (3). Schematic representation of a hypothetical scenario for phaseolin evolution [from the findings of 69-74].

CONFLICT OF INTEREST

The authors hereby declare that there is no conflict of interest related to this research. Aspects of the work described here were funded by the National Science Foundation (grant MCB-0346681 and predecessors).

REFERENCES

[1] Borlaug, N. *The Green Revolution, Peace and Humanity*. Nobel Peace Prize Lecture. Nobel Peace Prize. Nobel Museum. Stockholm, **1970**.

[2] Stanley, D.J.; Warne, A.G. Sea level and initiation of Predynastic culture in the Nile delta. *Nature*, **1993**, *363*, 435-438.

[3] Kaplan, L. Archeology and domestication in American Phaseolus (beans). *Econ. Bot.*, **1965**, *19*, 358-368.

[4] Burkart, A.; Brücher, H. Phaseolus aborigineus burkart, die mutmabliche andine Stammform der Kulturbohne. *Theor. Appl. Genet.*, **1953**, *23*, 65-72.

[5] McBryde, F.W. Cultural and Historical Geography of Southwest Guatemala. *Smithsonian Inst. Publ.*, **1947**, 4, 1-184.

[6] Kaplan, L.; Kaplan, L.N. Phaseolus in archaeology. In *Genetic resources of Phaseolus beans; their maintenance, domestication, evolution, and utilization.;* Gepts, P., Ed.; Kluwer Academic Publishers: Dordrecht, Holland, **1988**, pp. 125-142.

[7] Bewley, J.D.; Black, M. *Seeds: Physiology of Development and Germination*. Plenum Press, New York, **1994.**

[8] Ritthausen, H. Die proteinsubstanz der esben, wicken, saubohnen, linsen und bohnen, das pflazen-casein oder legumin. *J. Prakt. Chim.*, **1863**, *53*, 193.

[9] Osborne, T.B. *The Vegetable Proteins.* 2nd ed.; Longmans Green: London, **1924,** pp. 21-28.

[10] Shewry, P.R.; Casey, R. *Seed proteins*. Kluwer Academic Press: Boston, **1999**.

[11] Osborne, T.B. The proteins of kidney bean. *J. Am. Chem. Soc*., **1894**, *16*, 633-764.

[12] McLeester, R.C.; Hall, T.C.; Sun, S.M.; Bliss, F.A. Comparison of globulin proteins from *Phaseolus vulgaris* with those from *Vicia faba*. *Phytochem.*, **1973**, *12*, 85-93.

[13] Sun, S.M.; McLeester, R.C.; Bliss, F.A.; Hall, T.C. Reversible and irreversible dissociation of globulins from *Phaseolus vulgaris* seed. *J. Biol. Chem.*, **1974**, *249*, 2118-2121.

[14] Sun, S.M.; Mutschler, M.A.; Bliss, F.A.; Hall, T.C. Protein synthesis and accumulation in bean cotyledons during growth. *Plant Physiol.*, **1978**, *61*, 918-923.

[15] Mutschler, M.A.; Bliss, F.A.; Hall, T.C. Variation in the accumulation of seed storage protein among genotypes of *Phaseolus vulgaris* (L.). *Plant Physiol.*, **1980**, *65*, 627-630.

[16] Brown, J.W.S.; Bliss, F.A.; Hall, T.C. Microheterogeneity of globulin-1 storage protein from French bean with isoelectrofocusing. *Plant Physiol.*, **1980**, *66*, 838-840.

[17] Brown, J.W.S.; Bliss, F.A.; Hall, T.C. Linkage relationships between genes controlling seed proteins in French bean. *Theor. Appl. Genet.*, **1981**, *60*, 251-259.

[18] Plietz, P.; Damaschun, G.; Zirwer, D.; Gast, K.; Schlesier, B. Structure of 7S seed globulin from *Phaseolus vulgaris* L. in solution. *Int. J. Biol. Macromol.*, **1983**, *5*, 356-360.

[19] Blagrove, R.J.; Lilley, G.G.; Van Donkelaar, A.; Sun, S.M.; Hall, T.C. Structural studies of a French bean storage protein: phaseolin. *Int. J. Biol. Macromol.*, **1984**, *6*, 137-141.

[20] Lawrence, M.C.; Suzuki, E.; Varghese, J.N.; Davis, P.C.; Van Donkelaar, A.; Tulloch, P.A.; Colman, P.M. The three-dimensional structure of the seed storage protein phaseolin at 3 A resolution. *EMBO J.*, **1990**, *9*, 9-15.

[21] Lawrence, M.C.; Izard, T.; Beuchat, M.; Blagrove, R.J.; Colman, P.M. Structure of phaseolin at 2.2 A resolution. Implications for a common vicilin/legumin structure and the genetic engineering of seed storage proteins. *J. Mol. Biol.*, **1994**, *238*, 748-776.

[22] Hall, T.C.; Ma, Y.; Buchbinder, B.U.; Pyne, J.W.; Sun, S.M.; Bliss, F.A. Messenger RNA for G1 protein of French bean seeds: Cell-free translation and product characterization. *Proc. Natl. Acad. Sci. USA*, **1978**, *75*, 3196-3200.

[23] Hall, T.C.; Sun, S.M.; Ma, Y.; McLeester, R.C.; Pyne, J.W.; Bliss, F.A.; Buchbinder, B.U. The major seed storage protein of French bean seeds: characterization *in vivo* and translation *in vitro*. In *The Plant Seed: Development, Preservation and Germination*, Rubenstein, I.; Phillips, R.L.; Green, C.E.; Genegenbach, B.G., Eds. Academic Press, Inc.: New York, **1979**; pp 3-26.

[24] Sun, S.M.; Slightom, J.L.; Hall, T.C. Intervening sequences in a plant gene-comparison of the partial sequence of cDNA and genomic DNA of French bean phaseolin. *Nature*, **1981**, *289*, 37-41.

[25] Slightom, J.L.; Sun, S.M.; Hall, T.C. Complete Nucleotide sequence of a French bean storage protein gene: phaseolin. *Proc. Natl. Acad. Sci. USA*, **1983**, *80*, 1897-1901.

[26] Hall, T.C.; Sun, S.M.; Ma, Y.; Buchbinder, B.U.; Pyne, J.W.; Bliss, F.A.; Kemp, J.D. Bean seed globulin mRNA: translation, characterization, and its use as a probe towards genetic engineering of crop plants. In *Genome Organization and Expression in Plants*, Leaver, C.J., Ed. Plenum: New York, **1980**; pp 259-272.

[27] Murai, N.; Kemp, J.D.; Sutton, D.W.; Murray, M.G.; Slightom, J.L.; Merlo, D.J.; Reichert, N.A.; Sengupta-Gopalan, C.; Stock, C. A.; Barker, R.F. Phaseolin gene from bean is expressed after transfer to sunflower *via* tumor-inducing plasmid vectors. *Science*, **1983**, *222*, 476-482.

[28] Hall, T.C.; Reichert, N.A.; Sengupta-Gopalan, C.; Cramer, J.H.; Lea, K.; Barker, R.F.; Slightom, J.L.; Klassy, R.; Kemp, J.D. Regulation of the b-phaseolin gene expression in yeast and tobacco seed. In *Molecular form and function of the plant genome*, van-Vloten-Doting, L.; Groot, G.S.P.; Hall, T.C., Eds. Plenum press: New York, **1985**; pp. 517-529.

[29] Sengupta-Gopalan, C.; Reichert, N.A.; Barker, R.F.; Hall, T.C.; Kemp, J.D. Developmentally regulated expression of the bean ß-phaseolin gene in tobacco seed. *Proc. Natl. Acad. Sci. USA*, **1985**, *82*, 3320-3324.

[30] Hall, T.C.; Chandrasekharan, M.B.; Li, G. Phaseolin: its past, properties, regulation and future. In *Seed Proteins*, Shewry, P.R.; Casey, R., Eds. Kluwer Academic Publishers: Netherlands, **1999**; pp. 209-240.

[31] Derbyshire, E.; Wright, D.J.; Boulter, D. Legumin and vicilin, storage proteins of legume seeds. *Phytochemistry*, **1976**, *15*, 3-24.

[32] Sturm, A.; Van Kuik, J.A.; Vliegenthart, J.F.; Chrispeels, M.J. Structure, position, and biosynthesis of the high mannose and the complex oligosaccharide side chains of the bean storage protein phaseolin. *J. Biol. Chem.*, **1987,** *262*, 13392-13403.

[33] Slightom, J.L.; Drong, R.F.; Klassy, R.C.; Hoffman, L.M. Nucleotide sequences from phaseolin cDNA clones: the major storage proteins from *Phaseolus vulgaris* are encoded by two unique gene families. *Nucleic Acid Res.*, **1985**, *13*, 6483-6498.

[34] Bollini, R.; Vitale, A.; Chrispeels, M.J. *In vivo* and *in vitro* processing of seed reserve protein in the endoplasmic reticulum: evidence for two glycosylation steps. *J. Cell Biol.*, **1983**, *96*, 999-1007.

[35] Lioi, L.; Bollini, R. Contribution of processing events to the molecular heterogeneity of four banding types of phaseolin, the major storage protein of *Phaseolus vulgaris* L. *Plant Mol. Biol.*, **1984**, *3*, 345-353.

[36] Hall, T.C.; Slightom, J.L.; Ersland, D.R.; Murray, M.G.; Hoffman, L.M.; Adang, M.J.; Brown, J.W.S.; Ma, T.; Mathews, J.A.; Cramer, J.H.; Barker, R.F.; Sutton, D.W.; Kemp, J.D. Phaseolin: Nucleotide sequence explains molecular weight and charge heterogeneity of a small multigene family and also assists vector construction for gene expression in alien tissue. In *Structure and function of Plant Genome*, Cifferi, O.; Dure, L.S., Eds. Plenum: New York, **1983**; pp 123-142.

[37] Talbot, D.R.; Adang, M.J.; Slightom, J.L.; Hall, T.C. Size and organization of a multigene family encoding phaseolin, the major seed storage protein of *Phaseolus vulgaris* L. *Mol. Genet. Gen.*, **1984**, *198*, 42-49.

[38] Breathnach, R.; Benoist, C.; O'Hare, K.; Gannon, F.; Chambon, P. Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. *Proc. Natl. Acad. Sci., USA*, **1978**, *75*, 4853-4857.

[39] Rogers, J.; Wall, R. A mechanism for RNA splicing. *Proc. Natl. Acad. Sci., USA*, **1980**, *77*, 1877.

[40] Grace, M.L.; Chandrasekharan, M.B.; Hall, T.C.; Crowe, A.J. Sequence and spacing of TATA box elements are critical for accurate initiation from beta-phaseolin promoter. *J. Biol. Chem.*, **2004**, *279*, 8102-8110.

[41] Hyldig-Nielsen, J.J.; Jensen, E.O.; Paludan, K.; Wiborg, O.; Garrett, R.; Jorgensen, P.; Marcker, K.A.. The primary structures of two leghemoglobin genes from soybean. *Nucleic Acids Res.*, **1982**, *10*, 689-701.

[42] Pedersen, K.; Devereux, J.; Wilson, D.R.; Sheldon, E.; Larkins, B.A. Cloning and sequence analysis reveal structural variation among related zein genes in maize. *Cell*, **1982**, *29*, 1015-1026.

[43] Li, G.; Hall, T.C. Footprinting *in vivo* reveals changing profiles of multiple factor interactions with the beta-phaseolin promoter during embryogenesis. *Plant J*., **1999**, *18*, 633-641.

[44] Bustos, M.M.; Begum, D.; Kalkan, F.A.; Battraw, M.J.; Hall, T.C. Positive and negative cis-acting DNA domains are required for spatial and temporal regulation of gene expression by a seed storage protein promoter. *EMBO J*., **1991**, *10*, 1469-1479.

[45] Van Der Geest, A.H.M.; Frisch, D.A.; Kemp, J.D.; Hall, T.C. Cell ablation reveals that expression from the phaseolin promoter is confined to embryogenesis and microsporogenesis. *Plant Physiol.*, **1995**, *109*, 1151-1158.

[46] Chandrasekharan, M.B.; Bishop, K.J.; Hall, T.C. Module-specific regulation of the beta-phaseolin promoter during embryogenesis. *Plant J*., **2003**, *33*, 853-866.

[47] Ladizinsky, G.; Hymowitz, T. Seed protein electrophoresis in taxonomic and evolutionary studies. *Theor. Appl. Genet.*, **1979**, *54*, 145-151.

[48] Ladizinsky, G. Study of evolutionary problems by means of seed protein electrophoresis. In *Seed proteins: biochemistry, genetics, nutritive value*; Gottschalk, W.; Muller, H.P., Eds. NiJhoff/Junk: Netherlands, **1983**, pp. 481-498.

[49] Ladizinsky, G.; Adler, A. The origin of chickpea *Cicer. arietinum* L. *Euphytica*., **1976**, *25*, 211-217.

[50] Paulis, J.W.; Wall, J.S. Comparison of the protein compositions of selected corns and their wild relatives, teosinte and Tripsacum. *J. Agric. Food Chem.*, **1977**, *25*, 265-270.

[51] Johnson, B.L. Seed Protein Profiles and the Origin of the Hexaploid Wheats. *Am. J. Bot.*, **1972**, *59*, 952-960.

[52] Hymowitz, T.; Kaizuma, N. Dissemination of soybean (*Glycine max*): seed protein electrophoresis profiles among Japanese cultivars. *Econ. Bot.*, **1979**, *33*, 311-319.

[53] Hymowitz, T.; Kaizuma, N. Soybean seed protein electrophoresis profiles from 15 Asian countries or regions: hypotheses on paths of dissemination of soybeans from China. *Econ. Bot.*, **1981**, *35*, 10-23.

[54] Gepts, P. Phaseolin as an evolutionary marker. In *Genetic resources of Phaseolus beans: their maintenance, domestication, evolution, and utilization*.; Gepts, P. Ed.; Kluwer Academic Publishers: Dordrecht, Holland, **1988**, 215-241.

[55] Ma, Y.; Bliss, F.A. Seed proteins of common bean. *Crop. Sci.*, **1978**, *18*, 431-437.

[56] Li, G.; Chandrasekharan, M.B.; Wolffe, A. P.; Hall, T.C. Chromatin structure and phaseolin gene regulation. *Plant Mol. Biol.*, **2001**, *46*, 121-129.

[57] Ng, D.W.K.; Chandrasekharan, M.B.; Hall, T.C. Ordered histone modifications are associated with transcriptional poising and activation of the phaseolin promoter. *Plant Cell*, **2006**, *18*, 119-132.

[58] Murray, M.G.; Kennard, W.C. Altered chromatin conformation of the higher plant gene phaseolin. *Biochemistry*, **1984**, *23*, 4225-4232.

[59] Gepts, P.; Bliss, F.A. Enhanced available methionine concentration associated with higher phaseolin levels in common bean seeds. *Theor. Appl. Genet.*, **1984**, *69*, 47-53.

[60] Anthony, J.L.; Haar, R.A.V.; Hall, T.C. Nucleotide sequence of an alpha-phaseolin gene from *Phaseolus vulgaris*, *Nucleic Acids Res.*, **1990**, *18*, 3396.

[61] Kami, J.A.; Gepts, P. Phaseolin nucleotide sequence diversity in *Phaseolus*. I. Intraspecific diversity in *Phaseolus vulgaris*, *Genome*, **1994**, *37*, 751-757.

[62] Kami, J.; Velasquez, V.B.; Debouck, D.G.; Gepts, P. Identification of presumed ancestral DNA sequences of phaseolin in *Phaseolus vulgaris*, *Proc. Natl. Acad. Sci., USA*, **1995**, *92*, 1101-1104.

[63] Orkin, S.H.; Kazazian, H.H. The mutation and polymorphism of the human beta-globin gene and its surrounding DNA. *Ann. Rev. Genet.*, **1984**, *18*, 131-171.

[64] Pagnier, J.; Mears, J.G.; Dunda-Belkhodja, O.; Schaefer-Rego, K.E.; Beldjord, C.; Nagel, R.L.; Labie, D. Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. *Proc. Natl. Acad. Sci. USA*, **1984**, *81*, 1771-1773.

[65] Gibbs, P.E.; Strongin, K.B.; McPherson, A. Evolution of legume seed storage proteins--a domain common to legumins and vicilins is duplicated in vicilins. *Mol. Biol. Evol.*, **1989**, *6*, 614-623.

[66] Lipman, D.J.; Pearson, W.R. Rapid and sensitive protein similarity searches. *Science*, **1985**, *227*, 1435-1441.

[67] Harada, J.J.; Barker, S.J.; Goldberg, R.B. Soybean beta-conglycinin genes are clustered in several DNA regions and are regulated by transcriptional and posttranscriptional processes. *Plant Cell*, **1989**, *1*, 415-425.

[68] Gepts, P. The use of molecular and biochemical markers in crop evolution studies. *Evol. Biol.*, **1993**, *27*, 51-94.

[69] Shutov, A.D.; Kakhovskaya, I.A.; Braun, H.; Bäumlein, H.; Müntz, K. Legumin-like and vicilin-like seed storage proteins: evidence for a common single-domain ancestral gene. *J. Mol. Evol.* **1995**, *41*, 1057-1069.

[70] Bäumlein, H.; Braun, H.; Kakhovskaya, I.A.; Shutov, A.D. Seed storage proteins of spermatophytes share a common ancestor with desiccation proteins of fungi. *J. Mol. Evol.*, **1995**, *41*, 1070-1075.

[71] Braun, H.; Czihal, A.; Shutov, A.D.; Bäumlein, H. A vicilin-like seed protein of cycads: similarity to sucrose-binding proteins. *Plant Mol. Biol.*, **1996**, *31*, 35-44.

[72] Shutov, A.D.; Blattner, F.R.; Bäumlein, H. Evolution of a conserved protein module from Archaea to plants. *Trends Genet.*, **1999,** *15*: 348-349.

[73] Khuri, S.; Bakker, F.T.; Dunwell, J.M. Phylogeny, function, and evolution of the cupins, a structurally conserved, functionally diverse superfamily of proteins. *Mol. Biol. Evol.*, **2001**, *18*, 593-605.

[74] Dunwell, J.M. Cupins: a new superfamily of functionally diverse proteins that include germins and plant storage proteins. *Biotechnol. Genet. Eng. Rev.*, **1998**, *15*, 1-32.

Received: September 30, 2008 Revised: October 07, 2008 Accepted: October 15, 2008