⁂The Structure and Complexity of the 11S Polypeptides in Soybeans[1]

NIELS C. NIELSEN, USDA/ARS Agronomy Department, Purdue University, West Lafayette, IN 47907

ABSTRACT

The 11S soybean proteins called glycinin are isolated as a 350,000 dalton complex that consists of six nonidentical subunits. Each subunit consists of an acidic polypeptide component linked to a basic component by a single disulfide bond. Initial translation products of glycinin subunits are single polypeptides of ca. 60,000 daltons that undergo both co- and posttranslational modification. The precursors have a short signal sequence, followed by the acidic component, a short linker polypeptide, the basic component and a short trailer peptide. Five major subunit types have been purified and characterized by amino acid sequence analysis. While all of them are clearly synthesized by a family of homologous genes, they can be separated into two groups based on sequence homologies. Group I subunits ($A_{1a}B_2$, $A_{1b}B_{1b}$, A_2B_{1a}) are uniform in size (M_r = 58,000), relatively rich in methionine, and exhibit ca. 90% sequence homology among members in the group. The group II subunits (A_3B_4, $A_5A_4B_3$) exhibit a similar level of homology among themselves, although they contain less methionine and are larger ($M_r \cong$ 62,000-69,000) than group I subunits. Sequence homology between a member of one group and a member of the other is only 60-70%. Since the sulfur amino acid content of subunits is variable and genetic polymorphism in subunit composition has been documented, alteration of the functional and nutritional properties of these seed proteins by genetic manipulation may be possible.

INTRODUCTION

Soybeans have been grown on a commercial basis in the United States since the early part of the 20th century. While recognized primarily as an oil crop, they also are a rich source of protein. About 20% of the seed dry mass is oil, and about 40% is protein. In this country, the protein traditionally has been used as animal feed, although texturized soy protein recently has seen increased use as an additive in foods for human consumption. While problems associated with the use of soy proteins, such as factors associated with off-flavor and flatulence, remain serious impediments to more widespread acceptance of soy-based food products, considerable effort is being expended to understand features of the proteins that influence nutritional and functional properties of food products derived from them. Clearly, identification of the predominant polypeptides and elucidation of their structure and mechanisms of assembly are important prerequisites to effective manipulation of soy-based food products.

Seed proteins traditionally have been studied with a solubility fractionation scheme similar to that developed by Osborne (1) around the turn of this century. The scheme calls for sequential extraction of seed meal by a solvent series. Extraction with water yielded a fraction defined as albumins, dilute salt gave globulins, ethanol the prolamins, and acid/alkali producted glutelins. Since its inception, seed proteins from many plant species, varieties and breeding lines have been fractionated by various versions of this solubility scheme and compared. The studies have shown that proteins in legumes such as soybean are extracted predominantly into the globulin fraction and contain suboptimal levels of the sulfur amino acids, methionine and cystine. By contrast, the proteins of many grain cereals are extracted mainly into the prolamin fraction and have low lysine and tryptophan contents. Exceptions to this generalization exist, however, such as the high globulin content of oats and high glutelin content of rice.

While the solubility extraction methods provided a convenient means to compare the proteins from various seeds, the techniques suffered from serious drawbacks. The techniques were empirical, and there was no assurance that the same polypeptides would not be extracted into more than one solubility class due to their association with other proteins. A more precise means of identifying the predominant polypeptides in various protein fractions was clearly desirable. Early work on soybeans, reviewed by Wolf (2), soon established that the globulin fraction could be separated into two fractions that accounted for 70% or more of the total seed protein. One fraction sedimented in the ultracentrifuge between 7 and 8S. The major component in that fraction has come to be known by the trivial name, β-conglycinin. The other fraction sedimented between 11 and 12S and was called glycinin. Both of these proteins have been shown to be sequestered within specialized subcellular compartments and are termed storage proteins because they have no known catalytic function. Glycinin and β-conglycinin are thought to function as a mobilizable source of carbon and nitrogen used to support seedling growth and development immediately after germination.

Structural Features

The structure of soybean seed proteins has been studied extensively, and this work was reviewed recently (3). Both glycinin and β-conglycinin are complexes that consist of nonidentical subunits (Table I). β-Conglycinin generally is purified from dilute salt extracts of soybean meal and is found to be a trimer with a molecular weight of about 180,000. When ionic strength is low, however, β-conglycinin complexes associate to form 9S hexamers. Three prevalent types of subunits are associated with β-conglycinin and are referred to as α′, α and β. Another member of the β-subunit family, termed β′, is present in some soybean cultivars (4) and also has been described. Each of the subunits bears one or two N-linked glycosyl groups that consist of $Asn(NAcGlc)_2(Man)_{7-9}$.

By contrast, glycinin is devoid of sugar and does not undergo the ionic strength dependent association-dissociation phenomena characteristic of β-conglycinin. When purified from dilute salt extracts of soybeans, glycinin is about a 12S hexamer and generally is reported to have a molecular weight of about 360,000. Each glycinin subunit consists of two polypeptide components, one with an acidic (A_n) and the other with a basic (B_n) isoelectric point. The

TABLE I

Properties of the Two Major Soybean Proteins

Protein	Size	Mol wt[a]	Subunits	Sugar	% Sulfur
Glycinin	12.2S	309-393 kD	6	None	1.8
β-Conglycinin	7.9S	105-193 kD	3	$Asn(NAcGlc)_2$-$(Man)_{7-9}$	0.6

[a]Range of reported values. Molecular weights reported generally are around 180,000 for β-conglycinin and 350,000 for glycinin.

[1] Based on a paper presented at the AOCS Meeting, April 1984, Dallas, TX.

two polypeptide components (e.g., A_n-SS-B_n) are linked by a single disulfide bond (5).

Methods employed by many workers have relied on urea to dissociate the glycinin complex and disulfide reductants to increase resolution. The acidic and basic polypeptide components separate under these conditions (Fig. 1), and in the older literature have been referred to erroneously as subunits. As data collected during the past several years show, the fundamental units for assembly of glycinin complexes consist of A-SS-B single gene products. The disulfide linkage between the acidic and basic polypeptide components forms after subunit synthesis and may help stabilize the subunit after posttranslational modification. The term intermediate complex (6) has been used in the older literature and refers to complexes of an acidic and basic polypeptide. The term was coined when it was thought that acidic and basic polypeptides were synthesized from separate genes and associated randomly during subunit assembly. Since recent data have shown that this assumption is incorrect, use of the term intermediate complex should be discontinued in favor of the term subunit.

The initial objective of our studies was to purify each of the major glycinin polypeptides from a genetically defined soybean cultivar (CX635-1-1-1) and to establish the structural relationships between them. The early studies, summarized in Table II, revealed that the acidic and basic polypeptides each exhibited considerable NH_2-terminal sequence homology (7). Subsequent studies showed that the high degree of homology extended into the interior portions of the acidic polypeptides (8), and it was concluded that the proteins were synthesized from a family of homologous genes which had evolved from a common ancestral gene. Clones prepared from glycinin mRNAs by Goldberg and his collaborators (9) and used as probes to locate glycinin genes in the soybean genome yielded results consistent with this view. In addition, this group's data indicated that the glycinin gene family was small, because only three to five genes hybridized with the probe they used.

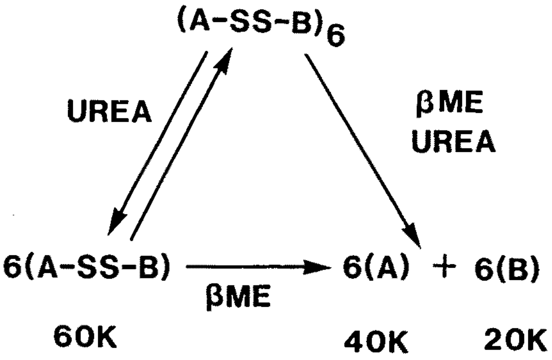

FIG. 1. Mechanisms for diassembly of glycinin precursors.

With the exception of B_{1b} and B_2, differences in the first 10-15 residues of the NH_2-terminal sequence permitted unambiguous identification of each acidic and basic polypeptide component (Table II). Attempts were therefore made to separate the A_n-SS-B_n subunits and ascertain the identity of their polypeptide components by sequence analysis (10). The principal conclusion drawn from this study was that association of acidic and basic polypeptides in subunits was nonrandom. Sequence analysis firmly established the polypeptide pairings A_2B_{1a} and A_3B_4. Polypeptides A_5B_3 were disulfide linked, but A_4 was found not to be associated with a basic component. Subsequent data have established that $A_5A_4B_3$ together form a subunit, but that the acidic component contains a proteolytic cleavage site about 100 amino acids from the NH_2-terminal end of the precursor polypeptide. Consequently, the A_4-component separates from disulfide linked A_5B_3 component upon denaturation (5,11). Small differences in isoelectric focusing patterns between B_{1b} and B_2 were used in addition to NH_2-terminal sequences to make the polypeptide assignments $A_{1a}B_2$ and $A_{1b}B_{1b}$.

Comparison of the physical properties of the glycinin subunits revealed they can be separated into two distinct groups (Table II). Subunits in group I have more uniform apparent molecular weights and contain more methionine than members of group II. This feature may be important in efforts to manipulate seed nutritional quality because legumes in general contain suboptimal levels of this essential amino acid. The extent of sequence homology between subunits provides an additional criteria with which to distinguish the two subunit groups. NH_2-terminal and internal sequence homology among members of the same group exceeds 80% while that between members of different groups is reduced to 40-50% (Table II) (8) (Nielsen, unpublished results).

The nomenclature and subunit assignments given in Table II were developed to identify the polypeptide components in glycinin subunits from cultivar CX635-1-1-1. A consideration which remains to be fully evaluated is how applicable these assignments are to subunits from other soybean varieties. Allelic variation undoubtedly will be present in the soybean population and potentially could lead either to changes in primary structure or to elimination of certain subunits from the seed. The extent to which such variation exists can be evaluated by examining large numbers of accessions from soybean germplasm collections. So far about 1200 accessions of *Glycine max* and *G. soja* from the Northern USDA collection, and a somewhat smaller number of cultivated varieties in the Brazilian collection, which are adapted to short day lengths, have been screened. For these studies single dimension SDS-electrophoretic

TABLE II

Comparison of the NH_2-Terminal Sequences of the Acidic and Basic Polypeptides in Glycinin Subunits from Breeding Line CX635-1-1-1

Group	Subunit	Mol wt[a]	No. met[b]	Acidic	Basic
I	$A_{1a}B_2$	58,000	5-6	FSSREQPQQNECQIQKLNALKPD...	GIDETICTMRLRQNIGQTSSPDIY
I	$A_{1b}B_{1b}$	58,000	5-6	FSFREQPQQNECQIQKLNALKPD...	GIDETICTMRLRHNIGQTSSPDIY
I	A_2B_{1a}	58,000	7-8	LREQAQQNECQIQKLNALKPD...	GIDETICTMRCRHNIGQTSSPDIF
II	A_3B_4	62,000	3	ITSSKF NECQLNNLNALQPD...	GVEENICTMKLHENIARPSWARFY
II	$A_5A_4B_3$	69,000	3	ISSSKL NECQLNNLNALEPD...	GVEENICTLKLHENIARPSWARFY

[a] Apparent molecular weight estimated by SDS-electrophoresis. Molecular mass for A_2B_{1a} determined from primary structure is 31,600 ± 100 (A_2) + 19,900 ± 100 (B_{1a}) = 51,500 ± 200 (Staswick et al., 1984).

[b] Variability in methionine content reflects heterogeneity among subunits (Staswick et al., 1984b).

and/or isoelectric focusing methodology were used, as well as two dimensional separations involving both techniques (Nielsen and Moreira, unpublished experiments). Few major charge or size variants were detected that affected glycinin subunits and also were genetically inherited. The results suggested that most major subunits observed in CX635-1-1-1 also were present in other soybean accessions. The single major exception to this generalization is the absence of subunit $A_5A_4B_3$ in a small proportion (estimate <5%) of the soybean population. Its lack in the seed has been shown to be due to a recessive allele of the structural gene for this subunit (11). The paucity of size and charge variants is perhaps not unexpected considering the narrow genetic base of cultivated soybeans. Moreover, the glycinin polypeptides in seeds of perennial species closely related to *G. max* are quite similar to those in CX635-1-1-1 (12).

Substantial heterogeneity in primary structure may, however, be present. Each purified polypeptide component from cultivar CX635-1-1-1 has exhibited charge microheterogeneity when subjected to analysis in analytical isoelectric focusing gels (8). To determine the basis for this heterogeneity and to provide a means for identification of gene coding regions, the complete amino acid sequence of the A_2B_{1a} subunit was determined (13). Structural heterogeneity was encountered at a number of locations in both the acidic and basic polypeptides, and accounts, at least in part, for the observed charge microheterogeneity (Fig. 2). The fact that the heterogeneity is not confined to the ends of the polypeptides, and that amino acids other than those which undergo deamidation are involved, implies that several different coding sequences contribute to the A_2B_{1a} polypeptides analyzed. It is likely that the charge microheterogeneity observed for the other acidic and basic polypeptides can also be attributed to structural heterogeneity.

While different coding sequences appear to account for the heterogeneity, the genetic relationships between them are obscure. That there was residual genetic heterogeneity among the population of inbreds from which the seed polypeptides were purified, or for that matter any commercially available variety, cannot be excluded. Alternatively, there may be more structural genes for glycinin subunits than predicted on the basis of experiments of Goldberg et al. (9) and Scallon et al. (14). These considerations remain to be evaluated more carefully.

Determination of the A_2B_{1a} primary structure permitted direct comparison of true molecular mass with the value estimated from SDS-electrophoresis. The acidic polypeptide contains 278 amino acids with an aggregate molecular mass of 31,600 ± 100. The uncertainty reflects structural heterogeneity. The true molecular mass of A_2 is substantially less than the values of about 37,000-40,000 estimated by SDS-electrophoresis and reported by most workers. In contrast, the B_{1a}-polypeptide contains 180 amino acids with an aggregate size of 19,900 ± 100 daltons, and is similar to the value of 20,000 daltons estimated electrophoretically. The basis for the non-ideal behavior of the acidic polypeptides in SDS-gels is unknown.

The non-ideal behavior of the group II subunits is accentuated when 6 $\underset{\sim}{M}$ is included in SDS-polyacrylamide gels (Fig. 3). Under these conditions, both A_3 and A_4 exhibit higher apparent molecular weights (15). This has proved to be a valuable feature because it permits the separation of A_4 from the group I acidic polypeptides and facilitates identification of cultivars with null-alleles for $A_5A_4B_3$ subunits (11,16). Interestingly, A_5 electrophoretic mobility remains unchanged in 6 $\underset{\sim}{M}$ urea relative to the basic polypeptides. The structural features which respond to urea, therefore, are located in the COOH-terminal two-thirds of the molecule.

Gene Structure and Expression

In vitro translation of purified glycinin mRNA results in 60,000 dalton precursors (17,18). The precursors consist of a signal sequence, followed by the acidic polypeptide, a short linker and then the basic polypeptide (18). Co- and

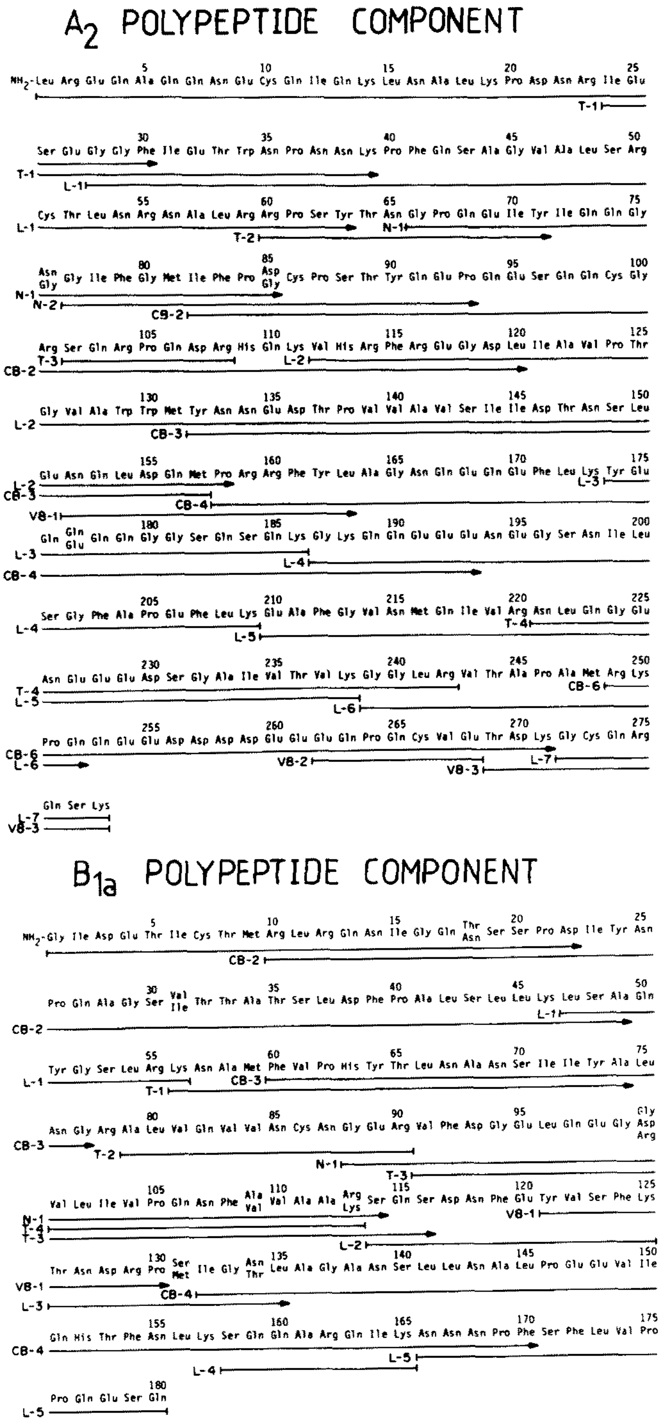
A2 POLYPEPTIDE COMPONENT
5 10 15 20 25
NH2-Leu Arg Glu Gln Ala Gln Gln Asn Glu Cys Gln Ile Gln Lys Leu Asn Ala Leu Lys Pro Asp Asn Arg Ile Glu
T-1
30 35 40 45 50
Ser Glu Gly Gly Phe Ile Glu Thr Trp Asn Pro Asn Asn Lys Pro Phe Gln Ser Ala Gly Val Ala Leu Ser Arg
T-1
L-1
55 60 65 70 75
Cys Thr Leu Asn Arg Asn Ala Leu Arg Arg Pro Ser Tyr Thr Asn Gly Pro Gln Glu Ile Tyr Ile Gln Gln Gly
L-1
T-2
N-1
80 85 90 95 100
Asn Gly Gly Ile Phe Gly Met Ile Phe Pro Asp Gly Cys Pro Ser Thr Tyr Gln Glu Pro Gln Glu Ser Gln Gln Cys Gly
N-1
N-2
CB-2
105 110 115 120 125
Arg Ser Gln Arg Pro Gln Asp Arg His Gln Lys Val His Arg Phe Arg Glu Gly Asp Leu Ile Ala Val Pro Thr
T-3
L-2
CB-2
130 135 140 145 150
Gly Val Ala Trp Trp Met Tyr Asn Asn Glu Asp Thr Pro Val Val Ala Val Ser Ile Ile Asp Thr Asn Ser Leu
L-2
CB-3
155 160 165 170 175
Glu Asn Gln Leu Asp Gln Met Pro Arg Arg Phe Tyr Leu Ala Gly Asn Gln Glu Gln Glu Phe Leu Lys Tyr Glu
L-2
CB-3
L-3
CB-4
V8-1
180 185 190 195 200
Gln Gln Glu Gln Gln Gly Gly Ser Gln Ser Gln Lys Gly Lys Gln Gln Glu Glu Glu Asn Glu Gly Ser Asn Ile Leu
L-3
L-4
CB-4
205 210 215 220 225
Ser Gly Phe Ala Pro Glu Phe Leu Lys Glu Ala Phe Gly Val Asn Met Gln Ile Val Arg Asn Leu Gln Gly Glu
L-4
T-4
L-5
230 235 240 245 250
Asn Glu Glu Glu Asp Ser Gly Ala Ile Val Thr Val Lys Gly Gly Leu Arg Val Thr Ala Pro Ala Met Arg Lys
T-4
L-5
CB-6
L-6
255 260 265 270 275
Pro Gln Gln Glu Glu Asp Asp Asp Asp Glu Glu Glu Gln Pro Gln Cys Val Glu Thr Asp Lys Gly Cys Gln Arg
CB-6
L-6
V8-2
L-7
V8-3
Gln Ser Lys
L-7
V8-3
B1a POLYPEPTIDE COMPONENT
5 10 15 20 25
NH2-Gly Ile Asp Glu Thr Ile Cys Thr Met Arg Leu Arg Gln Asn Ile Gly Gln Thr Asn Ser Ser Pro Asp Ile Tyr Asn
CB-2
30 35 40 45 50
Pro Gln Ala Gly Ser Val Ile Thr Thr Ala Thr Ser Leu Asp Phe Pro Ala Leu Ser Leu Leu Lys Leu Ser Ala Gln
L-1
CB-2
55 60 65 70 75
Tyr Gly Ser Leu Arg Lys Asn Ala Met Phe Val Pro His Tyr Thr Leu Asn Ala Asn Ser Ile Ile Tyr Ala Leu
L-1
CB-3
T-1
80 85 90 95
Asn Gly Arg Ala Leu Val Gln Val Val Asn Cys Asn Gly Glu Arg Val Phe Asp Gly Glu Leu Gln Glu Gly Gly Asp Arg
CB-3
T-2
N-1
T-3
105 110 115 120 125
Val Leu Ile Val Pro Gln Asn Phe Ala Val Val Ala Ala Arg Lys Ser Gln Ser Asp Asn Phe Glu Tyr Val Ser Phe Lys
V8-1
N-1
T-4
T-3
L-2
130 135 140 145 150
Thr Asn Asp Arg Pro Ser Met Ile Gly Asn Thr Leu Ala Gly Ala Asn Ser Leu Leu Asn Ala Leu Pro Glu Glu Val Ile
V8-1
CB-4
L-3
155 160 165 170 175
Gln His Thr Phe Asn Leu Lys Ser Gln Gln Ala Arg Gln Ile Lys Asn Asn Asn Pro Phe Ser Phe Leu Val Pro
CB-4
L-5
L-4
180
Pro Gln Glu Ser Gln
L-5

FIG. 2. Primary structure of polypeptides from glycinin subunit A_2B_{1a}. Standard three-letter abbreviations are used to denote amino acids. Horizontal lines indicate regions sequenced for each peptide. The beginning and end of peptides are denoted with vertical bars. Arrowheads indicate the peptide continues but was not sequenced beyond that point. Peptides were generated by cleavage with CNBr (e.g., CB-1, CB-2, etc.), NH_2OH (e.g., N-1, N-2, etc.), trypsin on citraconylated peptides (e.g., T-1, T-2, etc.), *Staphylococcus aureus* V-P protease (e.g., V8-1, V8-2, etc.) and endoproteinase Lys-C (e.g., L-1, L-2, etc). Data from Staswick et al. (13).

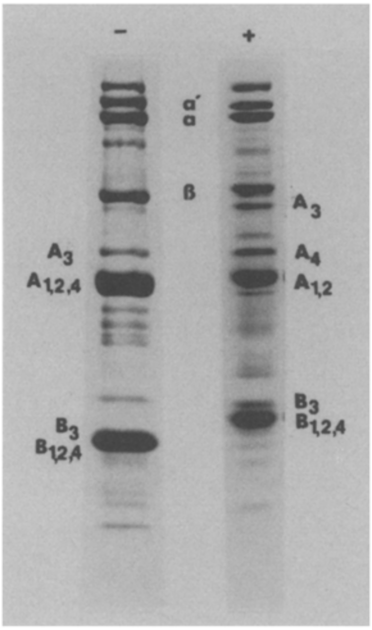

FIG. 3. Electrophoretic separation of seed extracts in SDS-gels with or without urea. Urea elicits lower mobility of A_3 and A_4 (right lane). Data from Fontes et al. (15).

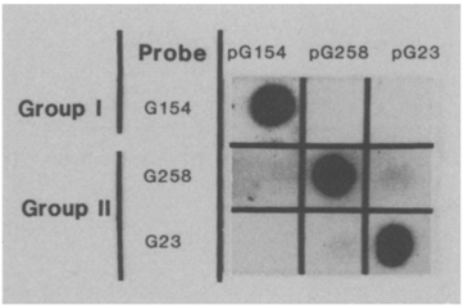

FIG. 4. Cross hybridization among glycinin probes. About 5 ng of DNA was bound to nitrocellulose. It was denatured in 0.1 N NaOH at 50 C for 5 min and equilibrated in 10× SSC. Southern hybridization was at 40% formamide, 5× SSC at 64 C for 16 hr and the blot was washed at 64 C in 0.2× SSC and 0.1% SDS.

posttranslational events appear to result in sequential removal of the signal and linker polypeptides, and account for the nonrandom association of the various acidic and basic polypeptide components of glycinin. Similar maturation processes have been observed for pea legumin (20,21), as well as 11S globulins from oats (22) and rice (23).

Recent experiments have provided a more complete description of the structural genes involved in glycinin synthesis and the posttranscriptional and posttranslational events which occur. Messenger RNAs isolated from mid-maturation stage seeds have been used to prepare three cDNA clones that are denoted pG154, pG23 and pG258. They encode A_2B_{1a}, A_3B_4 and $A_5A_4B_3$ subunits, respectively (14). Under stringent conditions of hybridization (64 C, 5XSSC, 40% formamide), pG154, which contains a group I glycinin subunit insert, does not interact with the two clones which contain group II inserts and vice versa (Fig. 4). The insert from pG23 hybridizes strongly with itself, but weakly with pG258. These results again emphasize the structural differences between the two groups of glycinin subunits and in addition provide a means to distinguish between sequences for group II subunits.

TABLE III

Comparison of Glycinin Genes in the Variety 'Dare'

Subunit group	Subunit ident[a]	Eco RI fragment size (kb)[b]	Hybridization pG154	pG23	pG258	Glycinin gene
I	$A_{1a}B_{1b}$	5.3	+	–	–	Gy_1
I	A_2B_{1a}	4.1	+	–	–	Gy_2
I	$A_{1b}B_2$	3.2	+	–	–	Gy_3
II	$A_5A_4A_3$	13	–	±	+	Gy_4
II	A_3B_4	9	–	+	±	Gy_5

[a] NH_2-terminal sequence of A_{1a}, A_{1b}, B_{1b}, B_2 are similar but not identical to those of CX635-1-1-1.

[b] Probes include about the 3′-terminal half of coding sequence for A_2B_{1a} (pG154), A_3B_4 (pG23) and $A_5A_4B_3$ (pG258). In the case of the Group I genes, other Eco RI fragments which originate from the 5′-end of the gene may hybridize to probes that contain additional coding sequence.

Leaf DNA has been purified from seedlings, cut with EcoRI and the resulting fragments separated in 0.5% agarose gels. When southern blots from these gels were probed with pG154, fragments of about 5.3, 4.1 and 3.2 kb hydridized (Table III). These three fragments correspond to G1, G2 and G3 identified by Fischer and Goldberg (24), and are located at the 3′-end of three group I subunit genes we refer to as glycinin-1, glycinin-2 and glycinin-3, respectively. Clones specific for group II sequences do not hybridize to the G1, G2 as G3 Eco RI fragments at moderate or high stringency, but do hybridize to fragments of about 13 kb and 9 kb in the same DNA digests. The glycinin insert in pG258 hybridizes strongly to the 13 kb fragment, but more weakly to the 9 kb fragment, while the converse result is obtained when the insert from pG23 is used as probe. A more complete analysis of a genomic clone which includes most of the 13 kb fragment revealed that the entire $A_5A_4B_3$ coding sequence (e.g., glycinin-4) is contained within the fragment (14). The 9 kb fragment therefore likely encodes the A_3B_4 subunits (glycinin-5). Genomic reconstructions using the probes suggest that there is about one copy of each glycinin gene per genome (14,24).

The complete nucleotide sequence of the glycinin-2 gene has been determined on a clone purified from a genomic library prepared by Goldberg's group using DNA from the variety "Dare" (Nielsen et al., in preparation). The gene product it encodes corresponds closely with A_2B_{1a} from soybean breeding line CX635-1-1-1. Only five differences in amino acid sequence were detected. They were scattered throughout the subunit and were due to single base changes. Tom Sims, in Bob Goldberg's laboratory, has since determined the structure of glycinin-1, while we have completed the sequence for glycinin-3 and glycinin-4 from the Dare library. The glycinin genes sequenced have a number of structural features in common. Each gene encoded a precursor with a signal peptide followed by the acidic component, a linker and then the basic component. The coding sequences in both group I and group II genes are interrupted three times by introns, twice in the acidic component region and once in the basic component region. The introns occur in each gene at the same position, although their length is variable. Not surprisingly, the intron-exon junction sequences correspond with the consensus GT/AG splicing rule for 5′ and 3′ borders in other eukaryotic genes (25). The introns exhibit regions with long runs of A or T, a characteristic also typical of DNA flanking the genes. When comparing corresponding introns among genes, sequence homology is found toward the intron borders but breaks down at the central part of the intervening sequence. Of

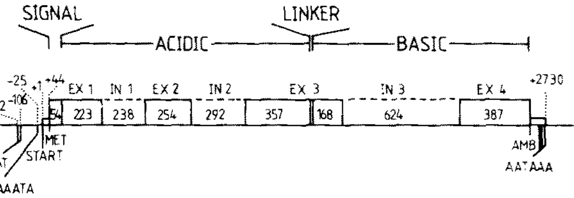

FIG. 5. Schematic representation of the A_2B_{1a} gene from the genomic clone λDA28-30. The genomic clone was isolated from an Alu-Hae III linker library prepared by Fischer and Goldberg. Hind III fragments from λDA28-30 that contained the gene were subcloned into pBR322 and mapped by standard technique. The nucleotide sequence of the gene was obtained by the method of Maxam and Gilbert. Data from Nielsen et al. (in preparation).

more potential interest are regions in the glycinin introns that exhibit 70-80% homology with the "TACTAAC-box" in introns of nuclear genes from yeast. This structure has been implicated as a splicing signal (26,27), although its significance in eukaryotic genes other than yeast has been questioned (28).

S_1-nuclease mapping located the site for initiation of transcription in the glycinin-2 gene (Fig. 5). In these experiments, polysomal mRNA enriched for glycinin mRNA (17) was used to protect the cloned gene. A region rather than a unique site was identified around 43 bp upstream from the putative AUG initiation codon. Failure to detect a single nucleotide in this region perhaps reflects the heterogeneous nature of the glycinin mRNA populations used to protect the gene in the mapping experiments.

Consensus regulatory sequences were identified in the 5′-flanking region of the glycinin-2 gene. A TATAAATA promoter site was located 25 bp upstream from the putative "cap" site and double CCAT sequences were found at positions -106 and -112. While the two CCAT sequences are located further from the promoter than in many other eukaryotic genes, they are similarly placed in genes for maize storage proteins (29). The significance, if any, of two CCAT sequences in glycinin-2 is unknown, but they also have been observed in 7S phaseolin (30).

Three consensus polyadenylation sequences were located in the 3′-flanking region of the glycinin-2 (Fig. 5). Comparison of this region with that for the trailer sequence of clone pG154 revealed that the polyadenylated tail began 10-12 bp after the third signal. Multiple polyadenylation signal sequences are a common feature of many eukaryotic genes (31 to 34), and they are used with different efficiencies. The factors involved in selection of a particular polyadenylation site for use are poorly understood.

The sequence analysis of full length cDNA clones and genomic clones has shown that glycinin genes each encode a short signal sequence (Table IV). As anticipated from analysis of glycinin precursors made in vitro (19), a high degree of homology among them was evident. In each case, cleavage of the signal occurred after a PheAla sequence. While variability in the identity of the next residue was tolerated, the NH_2-terminal amino acid of the mature protein was always a hydrophobic residue. As for secreted precursors of animal cells (35), the signal peptide contained a hydrophobic core and is followed by a region whose structure is more open where posttranslational cleavage takes place.

TABLE IV

Comparison of Signal Sequences of Glycinin Precursors

Gene	Signal sequence ↑ acidic polypeptide[a]
Gy_1	MAK. . . LVF. SLC. FLLFSQCCFA↑FSSREQ. . . . . . .
Gy_2	MAK. . . LVL. SLC. FLLFSGC. FA↑LREQAQ. . . . . . .
Gy_3	MAK. . . LVL. SLC. FLLFSGCCFA↑FSFREQ. . . . . . .
Gy_4	MGKPFTLSLSSLCLLLLSSA. CFA↑ISSSKL. . . . . . .

[a]Underlined amino acids hydrophobic. Dots in the signal sequence indicate the presence of residues at equivalent positions in other signals that are absent in that one.

TABLE V

Comparison of the Linker Regions of Glycinin Precursors

Gene	Acidic polypeptide ↑ linker region ↑ basic polypeptide[a]
Gy_1	GKDKHCQRPRGSQSK S↑RRN↑GID
Gy_2	PQCVETDKGCQRQSK ↑ R SRN↑GID
Gy_3	EECPDCDEKDKHCQS Q SRN↑GID
Gy_4	QP↑RRPRQEEPRERGC E TRN↑GVE
Gy_5	SRPEQQEPRGRGC Q TRN↑GVE

[a]In the case of Gy_2, the actual linker is known from comparison of the protein sequence of the mature polypeptides and the sequence deduced from clones. Cleavage at COOH-terminal of linker region for other subunits is predicted on basis of NH_2-terminal sequence determined for basic polypeptides (Moreira et al., 1979). Where indicated, cleavage at NH_2-terminals of linker region is predicted on the basis of paired basic amino acids being present.

The acidic and basic polypeptide domains in the precursor are joined by a linker of variable size (19). The structure in this region for each of the five major subunits of glycinin has been deduced (Table V). In each case, the ArgAsn↑Gly tripeptide at the junction between the linker and the basic polypeptide has been conserved. Since all basic polypeptides have NH_2-terminal glycine (6,7), cleavage occurs just before this residue. Proteolytic enzymes which cleave asparagine-glycine bonds are to our knowledge unknown, although an enzyme with similar specificity would be required to cleave glutamine-glycine bonds in coat polyprotein precursors of poliovirsus (36).

The site of cleavage at the junction between the acidic polypeptide and the linker region is not as well conserved (Table V). In the case of A_2B_{1a}, comparison of the sequence deduced from Edman degradation of the acidic polypeptide indicated cleavage occurred between lysine and asparagine. Paired basic amino acids have been implicated as recognition signals for proteolytic enzymes that modify prohormones posttranslationally (37), and are suggested to be important in processing of legumin precursors (38,39). Consistent with this view, paired basic amino acids have been identified in most of the linker regions in the glycinin precursors.

Cleavage at other sites in the glycinin subunits also implicates dibasic amino acids as proteolytic recognition signals. Unlike other glycinin subunits, the acidic polypeptide of $A_5A_4B_3$ undergoes a cleavage about 100 residues from its NH_2-terminal. A_4, the COOH terminal product of this cleavage, has paired arginine residues at its NH_2-terminal (7,16). Likewise, the basic polypeptide component in the A_2B_{1a} precursor appears to have a pentapeptide associated with it that is not found in the mature protein. The pentapeptide also has paired NH_2-terminal arginine residues (Nielsen et al., in preparation). By analogy with pulse-chase experiments done in pea cotyledons, it is likely that fragmented seed protein polypeptides such as glycinin are processed proteolytically in protein bodies as suggested by Chrispeels (40). The function, if any, of this proteolytic processing is poorly understood.

Relationship of Glycinin to Other Storage Proteins

The primary structures for a number of 11S and 7S subunits have been deduced from the nucleotide sequences of the DNA which encodes them. These include 7S proteins from common garden bean (phaseolin) (30), pea (vicilin) (41) and soybean (β-conglycinin) (4,42,43), as well as the 11S proteins from pea (legumin) (44) and soybean (glycinin-2 and glycinin-4) (14,37,45). These amino acid sequences have been compared using computer generated algorithms to detect alignments in primary structure and to predict secondary structure relationships (46). In addition to showing that there was considerable sequence homology and predicted secondary structural identity among members within each major group of storage proteins, the results also indicated there could be a high degree of structural relatedness between members of different groups. Thus, the 11S and 7S proteins apparently share common structural features and could be related to a common ancestral gene.

The algorithms were interpreted as showing each legume storage protein subunit consisted of three primary domains (Fig. 6). The COOH-terminal one, which includes nearly half of each molecule, was the most highly conserved region among subunits compared, contained extensive areas predicted to be in β-sheet or turn conformation, and was quite hydrophobic. It is considered likely that the COOH-terminal domain is buried within the molecule and has an important function in maintaining subunit structure. The other two domains split the remaining NH_2-terminal half of the subunit into two parts. The central domain was predicted to contain mixed helical, β-sheet and turn regions, while the NH_2-terminal domain was predicted to exist largely in helical and turn conformations. Unlike the central domain where regions of common predicted secondary structure lined up across all subunits compared, the positions of predicted helical and turn conformations in the NH_2-terminal domain varied between 11S and 7S subunits. Thus, the NH_2-terminal domains may tolerate modification better than either the central or COOH-terminal ones.

Upon comparing the 7S and 11S legume subunits, it was clear that the principal difference between them was due to the insertion of a hypervariable region between the central and COOH-terminal domains (Fig. 6). Size variation within this region also accounted fully not only for the size differences between the group I and group II glycinin subunits, but also between the two group II glycinin subunits (Nielsen, unpublished experiments). The insertions had an acidic nature due to high aspartate and glutamate content and were predicted to exist mainly in helical conformation. The natural variation that occurred within this region indicates that it does not fulfill a critical structural function and can tolerate modification.

Increased knowledge about the structure of storage proteins similar to that described for soybean glycinin has been gained for many of the grain legumes and cereals during the past half decade (47). As the structures of these proteins are compared, similarities and differences are bound to become apparent that will signal where changes in structure can be tolerated without adversely affecting maturation processes required for deposition of these protein in seeds. The information can be used with new biotechnology to understand these processes more fully. Both genes and complete coding sequences of legume storage proteins have been inserted into and are being expressed by a variety of bacteria, yeast and whole plants (Nielsen, unpublished experiments) (Beachy, personal communication) (48). Modification of the inserted genes should permit identification and study of parts of precursor molecules important in the maturation process the legume storage proteins undergo between synthesis in the endoplasmic reticulum and deposition in protein bodies. Similar approaches could also generate important information about regions of the molecules required for "functional" properties of the proteins in food systems.

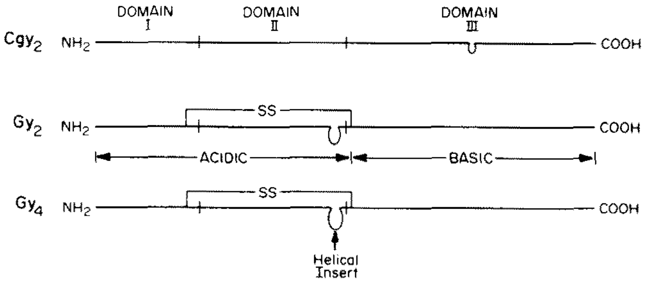

FIG. 6. Illustration of the predicted domain relationships and differences between the legumin and vicilin-like subunits from legumes. Residue inserts are shown as loops (data from Ref. 46).

ACKNOWLEDGMENTS

The work summarized was done by M.A. Moreira, P.E. Staswick, J.S. Medeiros, N.E. Tumer, Y. Marco, B. Scallon, L. Floener, T.J. Cho, C. Dickinson, C.S. Davies, K. Kitamura and T. Vu while they were associated with the author's laboratory. The work described has been cooperative research between USDA/ARS and the Purdue Agricultural Experiment Station. It was supported in part with grants from the American Soybean Association Research Foundation, the USDA Competitive Grants Program and the US-Brazil Bilateral Scientific Exchange Program of the National Science Foundation.

REFERENCES

1. Osborne, T.B., The Vegetable Proteins, 2nd Ed., Longmans and Green, New York, 1924.
2. Wolf, W.J., and J.C. Cowan, Soybeans as a Food Source, 2nd Ed., CRC Press, Cleveland, 1975.
3. Nielsen, N.C., in "New Protein Foods," Vol. 5, edited by A.M. Altschul and H.L. Wilcke, Academic Press, New York, pp. 27-58, 1985.
4. Coates, J.B., J.S. Medeiros, V.H. Thanh and N.C. Nielsen, Arch. Biochem. Biophys. (in press).
5. Staswick, P.E., M.A. Hermodson and N.C. Nielsen, J. Biol. Chem. 259:13431 (1984a).
6. Kitamura, K., T. Takagi and K. Shibasaki, Agric. Biol. Chem. 40:1837 (1976).
7. Moreira, M.A., M.A. Hermodson, B.A. Larkins and N.C. Nielsen, J. Biol. Chem. 254:9921 (1979).
8. Moreira, M.A., M.A. Hermodson, B.A. Larkins and N.C. Nielsen, Arch. Biochem. Biophys. 210:633 (1981).
9. Goldberg, R.B., G. Hoscheck, G.S. Ditta and R.W. Breidenbach, Developmental Biol. 83:218 (1981).
10. Staswick, P.E., M.A. Hermodson and N.C. Nielsen, J. Biol. Chem. 256:8752 (1981).
11. Kitamura, K., C.S. Davies and N.C. Nielsen, Theor. Appl. Genet. 68:253 (1984).
12. Staswick, P.E., P. Broué and N.C. Nielsen, Plant Physiol. 72: 114 (1983).
13. Staswick, P.E., M.A. Hermodson and N.C. Nielsen, J. Biol. Chem. 259:13424 (1984b).
14. Scallon, B.J., V.H. Thanh, L.A. Floener and N.C. Nielsen, Theor. Appl. Genet. 70:510 (1985).
15. Fontes, E.P.B., M.A. Moreira, C.S. Davies and N.C. Nielsen, Plant Physiol. 76:840 (1984).
16. Staswick, P.E., and N.C. Nielsen, Arch. Biochem. Biophys. 223:1 (1983).
17. Tumer, N.E., V.H. Thanh and N.C. Nielsen, J. Biol. Chem. 256:8756 (1981).
18. Barton, K., J. Thompson, J. Madison, R. Rosenthal, N. Jarvis and R. Beachy, Ibid. 257:6089 (1982).
19. Tumer, N.E., J.D. Richter and N.C. Nielsen, Ibid. 257:4016 (1982).
20. Spencer, D., and T.J.V. Higgins, Biochem. Int. 1:502 (1980).
21. Croy, R.D., J.A. Gatehouse, I.M. Evans and D. Boulter, Planta 148:49 (1980).
22. Brinegar, A.C., and D.M. Peterson, Plant Physiol. 70:1767 (1982).
23. Yamagata, H., T. Sugimoto, K. Tanaka and Z. Kasai, Ibid. 70: 1094 (1982).
24. Fischer, R., and R.B. Goldberg, Cell 29:651 (1982).

25. Breathnach, R., and P. Chambon, Annu. Rev. Biochem. 50:349 (1981).
26. Langford, C.J., and D. Gallwitz, Cell 33:519 (1983).
27. Langford, C.J., F.-J. Klinz, C. Donath and D. Gallwitz, Ibid. 36:645 (1984).
28. Mount, S.M., Nature 304:309 (1983).
29. Pedersen, K., J. Devereux, D.R. Wilson, E. Sheldon and B.A. Larkins, Cell 29:1015 (1982).
30. Slightom, J.L., S.M. Sun and T.C. Hall, Proc. Natl. Acad. Sci. U.S.A. 80:1897 (1980).
31. Alt, F.W., A. Bothwell, M. Knapp, E. Siden, E. Mather, M. Koshland and D. Baltimore, Cell 20:293 (1980).
32. Rosenfeld, M.G., J.J. Mermod, S.A. Amara, L.W. Swanson, P.E. Sawchenko, J. River, W.W. Vale and R.M. Evans, Nature 304:129 (1983).
33. Henikoff, S., J.S. Sloan and J.D. Kelly, Cell 34:405 (1983).
34. Setzer, D.R., M. McGrogan, J.H. Nunberg and R.T. Schimke, Ibid. 22:361 (1980).
35. Harwood, R., in "The Enzymology of Post-translational Modification of Proteins," edited by R.B. Freedman and H.C. Hawkins, Academic Press, New York, 1980, p. 3.
36. Kitamura, N., B. Semler, P. Rothberg, G. Larsen, C. Adler, A. Dorner, E. Emini, R. Hanecak, J. Lee, S. van der Werf, C. Anderson and E. Wimmer, Nature 291:547 (1981).
37. Docherty, K., R.J. Carroll and D.F. Steiner, Proc. Natl. Acad. Sci. U.S.A. 79:4613 (1982).
38. Croy, R.R.D., G.W. Lycett, J.A. Gatehouse, J.N. Yarwood and D. Boulter, Nature 295:76 (1982).
39. Nielsen, N.C., Phil. Trans. R. Soc. Lond. B304:287 (1983).
40. Chrispeels, M.J., Ibid. B304:309 (1983).
41. Lycette, G.W., A.J. DeLauney, J.A. Gatehouse, J. Gilroy, R.R.D. Croy and D. Boulter, Nucleic Acids Res. 11:2367 (1983).
42. Schuler, M.A., B.F. Ladin, J.C. Pollaco, G. Freyer and R.N. Beachy, Ibid. 10:8245 (1982).
43. Schuler, M.A., E.S. Schmidt and R.N. Beachy, Ibid. 10:8225 (1982).
44. Lycette, G.W., R.R.D. Croy, A.H. Shirsat and D. Boulter, Ibid. 12:4493 (1984).
45. Marco, Y.A., V.H. Thanh, N.E. Tumer, B.J. Scallon and N.C. Neilsen, J. Biol. Chem. 259:13436 (1984).
46. Argos, P., S.V.L. Narayana and N.C. Nielsen, EMBO Journal 4:1111 (1985).
47. Larkins, B.A., in "The Biochemistry of Plants, A Comprehensive Treatise," Vol. 6, edited by P.K. Stumpf and E.E. Conn, Academic Press, New York, 1981, p. 450.
48. Cramer, J.H., K. Lea and J.L. Slightom, Proc. Natl. Acad. Sci. U.S.A. 82:334 (1985).

[Received October 5, 1984]

Modification of Surface Charges of Soy Protein by Phospholipids

W. S. CHEN and **W. G. SOUCIE***, Kraft Inc. R&D,
801 Waukegan Rd., Glenview, IL 60025

ABSTRACT

Lecithin is used to prevent soy protein isolates from clumping in food processing. A PenKem Inc. System 3000 Electrokinetic Analyzer was used to investigate how the phospholipid modified the surface charge of soy protein. Electrophoretic mobility-pH curves showed that a commercial soy lecithin lowered the isoelectric point (pI) and increased the electrical mobility of soy protein more than did a pure phosphatidylcholine. The modification of the surface charge of the protein was a function of the phospholipid added. Lecithinated soy isolate was more negatively charged and thus more dispersible in water than the nonlecithinated soy control.

INTRODUCTION

Protein phospholipid interactions related to food problems have been studied and reviewed extensively (1-7). However, few studies have been made on how phospholipids change colloidal properties of proteins. Because proteins form colloids in water and the surface charge plays an important role in dispersion (8), it was of interest to investigate how phospholipids modify the surface charges of soy protein and how these modifications affect colloidal properties.

In dilute aqueous suspensions the electrophoretic mobility is proportional to the zeta potential which, in turn, is proportional to the surface charge (8-11). Therefore, we have measured electrophoretic mobility in order to characterize the surface charge of the particles of soy isolate, phospholipid and phospholipid/soy protein complex.

MATERIALS AND METHODS

L-alpha-phosphatidylcholine was from Sigma Chemical Co. (St. Louis, Missouri), and Centrolene A (a food grade, hydroxylated soybean lecithin) was from Central Soya Co. (Ft. Wayne, Indiana). Soy protein isolate obtained from Kraft, Inc. (Glenview, Illinois) had the following composition: nitrogen, 13.3%; fat, 2.7%; moisture, 6%; and ash, 2.8%. Lecithinated soy isolate was prepared by spraying 0.2% (w/w) liquified lecithin onto the soy protein while mixing the powder in a ribbon blender.

L-alpha-phosphatidylcholine and Centrolene A were added into distilled water in small portions with stirring to form 1 mg/ml dispersions. These phospholipid dispersions were then mixed with various concentrations of soy protein solution to obtain dispersions with different phospholipid/protein ratios. The dispersions of phospholipids, soy protein and phospholipid/soy protein were adjusted to various pH values with either 0.1 M HC1 or NaOH. In order to obtain the mean mobility values for the electrophoretic mobility vs. pH plots (Figs. 1, 5, and 6) all protein/phospholipid mixtures were stirred for 30 min or until a single peak was obtained. To capture the presence of multiple peaks (Fig. 2), mobility measurements were taken approximately 1 min after mixing.

Because electrophoretic mobility was unaffected by the concentration of the colloid dispersion used in this study (0.4-1.0 mg/ml), the overall concentration of protein and phospholipids was not adjusted to the same final value.

Electrophoretic mobilities of phospholipids, soy protein and phospholipid-soy protein dispersions were measured at various pH values with a PenKem System 3000 automated electrokinetic analyzer at 25 C. The procedure described in the instruction manual was followed. One mobility unit is equal to 1.0×10^{-8} meters/sec/volt/meter.

Particle size analysis of 1% (w/v) dispersions of protein in distilled water, pH 6.3 and 25 C, were measured in an Electrozone/Celloscope manufactured by Particle Data, Inc., Elmhurst, Illinois.

RESULTS AND DISCUSSION

The electrophoretic mobility-pH curves of food grade lecithin and soy isolate in Figure 1 show that lecithin has a higher negatively charged surface than soy protein. As a result, two groups of peaks appear immediately after the phospholipid

*To whom correspondence should be addressed.