Potential β-Sheet Surfaces of Soybean Seed Proteins

John A. Rothfus*

Biopolymer Research, NCAUR, USDA, ARS, Peoria, Illinois 61604

ABSTRACT: Volume and amphiphilicity profiles computed for β-sheet conformations of soybean seed proteins (the acidic and basic subunits of glycinin, the α subunit of β-conglycinin, an extensin-like high-proline protein, and a lipid body oleosin) were compared to profiles for selected nonseed proteins and random-sequence polypeptides. The major soy proteins resemble fibrinogen more than silk or collagen but are differentiated from each other by surface polarity. Polarity in acidic glycinin fluctuates somewhat regularly and symmetrically along both sides of the β-sheet, but in basic glycinin, polarity on one side is fourfold that on the other throughout 70% of the protein. Polar residues distribute equally on either side of the β-conglycinin subunit, but half of the acidic and basic residues concentrate in the N-terminal third of the molecule. The remainder of the β-conglycinin contains 90% of the molecule's tyrosine, of which 70% is along one side. In the high-proline protein, 90% of the tyrosine distributes to one side of a β-sheet; 95% of the acidic residues to the other. Only soy oleosin approximates the per-residue volume, polarity, and uniformity of silk or collagen. Its 85-residue central lipophilic domain is much less polar than silk and nearly as uniform.

JAOCS 72, 501–506 (1995).

KEY WORDS: Amphiphilicity, computational chemistry, conglycinin, glycinin, molecular volume, oleosin.

Certain proteins, such as silk and collagen, have relatively uniform dimensions and nonpolar molecular surfaces that allow them to self-associate into tough, pliable, quite valuable materials. Specially engineered proteins promise equally valuable properties that, furthermore, can be regulated by chemical synthesis (1) or by systematic application of recombinant genetics (2). In addition, proteinaceous mixtures that are not as uniform nor as apolar as silk or collagen can be rendered more durable by various chemical crosslinks analogous to those used in plastics. Salt bridges, for example, strengthen cured casein glues (3), and covalent intermolecular condensations stabilize fibrin in clotting blood (4).

Such knowledge of protein versatility strengthens conviction that plant proteins, which are among the world's most abundant and renewable polymers, should be examined systematically for elements that lend themselves, directly or after chemical modification, to uses presently served by petroleum-based polymers. Similar conviction led Henry Ford to incorporate soybean proteins into automobile parts a half century ago (5), only to see them replaced by petrochemical resins for economic and performance reasons.

Unlike Ford's time, energy and environmental concerns now justify biofuels development and promise cheap protein by-products from fuel processes. With market conditions more favorable, it is again appropriate to concentrate on the performance of agricultural commodities in nonfood materials. Fortunately, knowledge of protein structure, the ultimate determinant of performance, is accumulating rapidly, and molecular biology provides the means to adjust and magnify desirable structures once they are identified.

Because useful properties of industrial materials arise as much from weakly bonded interactions of regular molecular surfaces as from covalent structure, it is important to identify the kinds of surfaces allowed by the primary structures of plant proteins. Accordingly, this work describes the distribution of volume and polarity along molecular surfaces generated from known sequences for soybean seed proteins in β-sheet conformations.

EXPERIMENTAL PROCEDURES

Volume and amphiphilic character profiles were computed from amino acid sequence data by a moving window analysis as described by Rose *et al.* (6). Molecular hydrated volumes and amphiphilicities are summations of amino acid residue dimensions and amphipathic characteristics collected by others (Table 1). Unless specified otherwise, sequence residue numbers are inclusive, volumes are in $Å^3$, and amphiphilicities are in arbitrary units (a.u.). Values assigned to alternate residues were summed separately to estimate bulk and polarity on either side of each extended polypeptide chain. Side A thus represents odd-numbered residues; Side B, even-numbered. Molecular values, which remain unmodified for terminal structure and charge, are expressed on a per-residue basis to facilitate comparisons between different-size molecules or segments.

Uniformity of volume and amphiphilicity on each side of a molecule or segment was estimated on a per-residue basis by calculating volume or amphiphilicity dispersions, i.e., the standard deviation of each property for a specific sequence or side. Thus, for example, homopolypeptides have volume and

*Address correspondence at NCAUR, 1815 North University St., Peoria, IL 61604.

TABLE 1
Amino Acid Residue Volumes and Amphiphilicities

Acid	Volume[a] (Å3)	Amphiphilicity[b] (a.u.)
Isoleucine	169	−73
Phenylalanine	203	−61
Valine	142	−54
Leucine	168	−53
Tryptophan	238	−37
Methionine	171	−26
Alanine	92	−25
Glycine	66	−16
Cysteine	106	−4
Tyrosine	204	−2
Proline	129	7
Threonine	122	18
Serine	99	26
Histidine	167	40
Glutamic acid	141	62
Asparagine	135	64
Glutamine	161	69
Aspartic acid	114	72
Lysine	176	110
Arginine	181	176

[a]Chen and Bendedouch (Ref. 7). [b]Eisenberg *et al.* (Ref. 8).

TABLE 2
Proteins

Designation	Protein	Residues	Reference
Soybean			
SG2AGLY	G2 acidic glycinin	278	10
SG2BGLY	G2 basic glycinin	185	10
SABCNGLY	β-conglycinin, α subunit	583	11
SHIPRO	pro-rich protein (SbPRP2)	203	12
SOLSN	24-kDa oleosin	223	13
Miscellaneous			
SILK2	*Nephila clavipes* Spidroin 2	627	14
HCOLA1X	human α1(X)collagen	680	15
HGFBGN	human fibrinogen, γ-chain	411	16
BCASEINB	bovine βA^2-casein	209	17

amphiphilicity dispersions of zero while peptides with uniform distributions of equimolar quantities of each of the twenty common amino acids on each side give volume and amphiphilicity dispersion values of 42 and 62, respectively. Certain alternating sequences, e.g., polyglycyltryptophan or polyisoleucylarginine, have the greatest volume (86 $Å^3$) and amphiphilicity (125 a.u.) dispersions overall while maintaining zero dispersions on either side of the peptide backbone.

Sequences were generally derived from nucleic acid data. Each side of a sequence was scanned from N-terminus to C-terminus at a window width of 5 residues (i.e., index residue ± 2 residues). Larger or smaller window widths diminished or obscured profile details.

A general frame of reference was established by similar analyses of fifty 100-residue random sequences and five 1000-residue random sequences generated from a randomized amino acid population, weighted according to the frequency of occurrence of amino acid residues in known protein sequences (9). For comparison purposes, several nonseed and nonplant proteins were also examined. Proteins discussed in this work are listed in Table 2 along with designations by which they are identified.

RESULTS AND DISCUSSION

With backbone chains of carbon and nitrogen atoms, the simplest polypeptides resemble hydrocarbon polymers. Their monomeric units, which range from 66 to 238 $Å^3$, are, furthermore, about the size of those in common plastics such as polyethylene (55 $Å^3$) and polystyrene (190 $Å^3$). Thus, silk, in which the small amino acids glycine, alanine, and serine predominate, readily assumes a stable β-sheet conformation with many of its backbone atoms arranged in zigzag chains analogous to molecules in hydrocarbon plastics. Depending on composition, sequence, and hydrogen bond structure, other proteins adopt three-dimensional configurations that are far more complex but also more susceptible to change with changes in their molecular environments. Knowledge of parameters that control conformations of seed proteins is thus fundamental to adapting them to nonfood uses. Idealized β-sheet conformations provide convenient means to inspect the proteins for features that can affect their surface properties, even though their primary structures, which often include numerous proline residues, may favor alternative conformations.

Random sequences. Arithmetic averages suggest that completely random polypeptide sequences should have average residue volumes of 149 $Å^3$, which is slightly larger than the average monomeric volume for typical hydrocarbon polymers. Amphiphilicities would average 14.6 a.u. per residue. This is much more polar than −9 a.u. per methylene, which might be expected for polyethylene if an average of the amphiphilicity differences between valine and isoleucine or leucine is an accurate estimate of methylene polarity.

Sequences selected randomly from an amino acid population that reflects natural abundance (9) are somewhat different from arithmetic means. Analyses of 55 such sequences, representing 10,000 residues, led to the per-residue values given in Table 3 for an average random β-structure. In general, these values are consistent with the natural predominance of small and polar amino acids (9), the range of amino acid volumes from 66 to 238 $Å^3$, and the fact that amino acid amphiphilicities span a range from −73 to 176 a.u. As antici-

TABLE 3
Per-Residue Volume and Amphiphilicity of a Random β-Sheet Protein[a]

	Volume ($Å^3$)	Dispersion ($Å^3$)	Amphiphilicity (a.u.)	Dispersion (a.u.)
Complete sequence	142 ± 2	39 ± 1	14.9 ± 4.0	66 ± 3
Side A	141 ± 3	40 ± 2	16.4 ± 5.9	67 ± 3
Side B	142 ± 3	39 ± 2	13.4 ± 4.3	64 ± 4

[a]Average values ± SD from fifty 100-residue and five 1000-residue sequences selected randomly from a randomized population consistent with natural abundance (Ref. 9); a.u., arbitrary units.

pated, values for volume and dispersions in random sequences indicate little or no variation from side to side or from sequence to sequence. Amphiphilicity averages, however, suggest substantial variation. Furthermore, distributions of specific amino acids to either side of random β-sheet sequences averaged ±3.5 residues from half the total for each type; not zero or ±0.5.

The asymmetry implied by these unexpected deviations from means was especially apparent in two of the five 1000-residue random sequences. Volume and amphiphilicity profiles for segments from these two sequences are given in Figure 1. Approximately 20–25% of each sequence is noticeably different from the remainder, due to random aggregation of extreme values. In random natural-frequency polypeptide 1 (RNDNF1), residues 305–485 have the volume expected of a random sequence, but this segment's polarity is reduced well below average on Side B (−1.9 a.u.) and increased above average on Side A (21.3 a.u.), due to a concentration of polar residues near 405. In RNDNF5, a segment, 160–295, that has a relatively small per-residue volume (135 $Å^3$) overlaps another, 110–240, in which volume (136 $Å^3$) and polarity (3.5 a.u.) are both below average.

Examination of random sequences provides a tentative context within which to recognize features that differentiate protein surfaces. Rigorous guidelines await much additional data. Significant differences in properties, however, should accompany per-residue differences of more than 4–6 $Å^3$ and 8–12 a.u. Likewise, dispersion differences of more than 5% should indicate significant differences in surface uniformity, provided clusters of extreme properties are recognized and profiles justify comparisons. Fortunately, nonrandomness is often quite obvious in natural protein sequences.

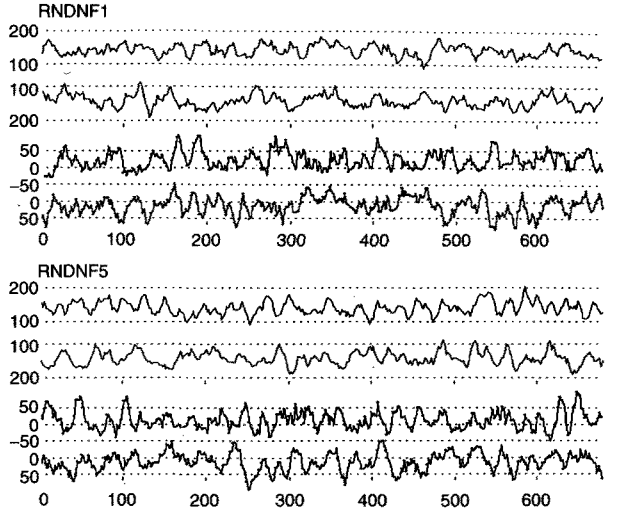

FIG. 1. Volume (upper panel) and amphiphilic (lower panel) profiles of two random polypeptide sequences [random natural-frequency polypeptide (RNDNF1, RNDNF5)] in extended ß-sheet conformation. Upper curve in each pair represents Side A of the peptide chain; lower curve, Side B. Properties are displayed, left to right, from N-terminus to C-terminus. Volumes are in $Å^3$; amphiphilicities, arbitrary units. Tick marks indicate 100 residues.

Nonplant proteins. Compared to the random sequences, spider silk (Fig. 2, SILK2) is remarkably narrow and apolar. Its uniform, nearly hydrocarbon-like architecture appears ideal for self-association and flexibility. Regular notches of reduced volume and polarity along these profiles coincide with repeated short sequences of polyalanine, which are suspected (18) to undergo reversible helix formation and thereby impart elasticity to silk fibers. Though it is not obvious from the SILK2 profiles, it is extremely interesting from the standpoint of chemical reactivity that two-thirds of the protein's tyrosine residues and nearly 60% of its serine residues are aligned along Side A of the β-sheet.

The portion of human collagen (Fig. 2, HCOLA1X) that survives post-translational editing (residues 57–519), likewise, has a smaller diameter and is more uniform in volume than random sequences, but its amphiphilicity profile evidences a surface more polar and varied than that of silk. Such differences appear consistent with differences in the physical properties of these two materials. Analogous to silk, collagen concentrates 60–70% of its tyrosine, histidine, and serine residues on Side A of the β-sheet.

In contrast, the γ-chain of human fibrinogen (Fig. 2, HGFBGN), the principal constituent of fibrin, exhibits none of the uniformity seen in silk or collagen, even though it likewise contains elements of a durable solid. This attests to the

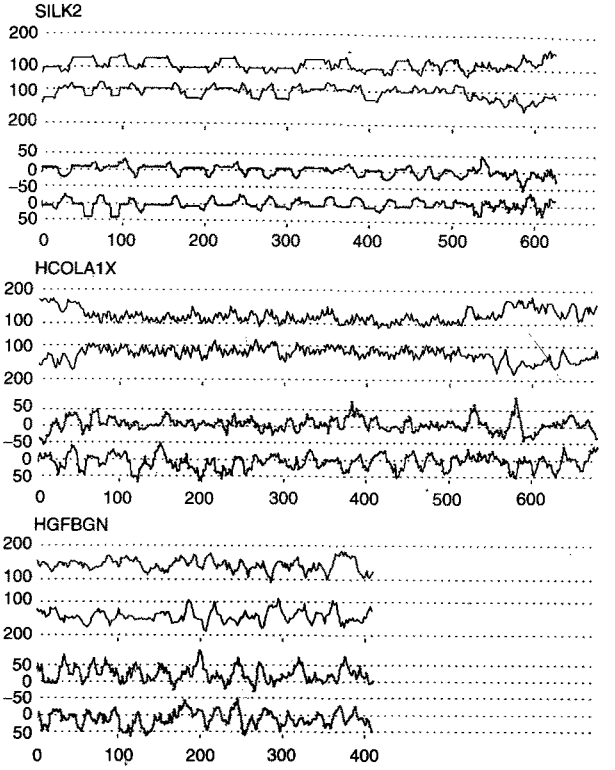

FIG. 2. Volume and amphiphilic profiles of selected nonseed proteins. Properties are displayed as in Figure 1. HCOLA1X profiles include N-terminal and C-terminal noncollagenous sequences of 56 and 161 residues, respectively. SILK2, *Nephilia clavipes* Spidroin 2; HCOLA1X, human α1(X) collagen; HGFBGN, human fibrinogen, γ-chain.

importance of multiple mechanisms of molecular interaction in the formation of useful materials.

Soybean seed proteins. Profiles of major soybean cotyledon proteins in Figure 3 clearly demonstrate that they are neither random nor are they as uniform as collagen or silk. Though different in molecular weight, the soybean proteins appear quite similar overall in the distribution of volume along each sequence. Only SG2BGLY, which has a slightly smaller per-residue volume, is noticeably asymmetric from residues 30 through 160. This segment, which accounts for 70% of the molecule, is about 4% larger on Side A than Side B.

Amphiphilicity profiles also attest to the unusual nature of SG2BGLY, which is much less polar (12.9 a.u.) than SG2AGLY (27.6 a.u.) or SABCNGLY (30.2 a.u.). In SG2AGLY, concentrations of polar residues occur to about the same extent on either side of the molecule at regular intervals. SG2BGLY lacks this regularity. Most intriguing, the 30–160 segment in SG2BGLY exhibits a definite nonpolar side that is not obvious in the other proteins: Side A, 12.8 a.u.; Side B, 3.0 a.u. Peptides that exhibit such alternation of hydrophobicity and hydrophilicity between alternate residues tend to form β-sheets (19). Accordingly, the results lead to conjecture that the bulk of SG2BGLY should readily assume the β-sheet conformation and might, therefore, easily form membranous molecular complexes.

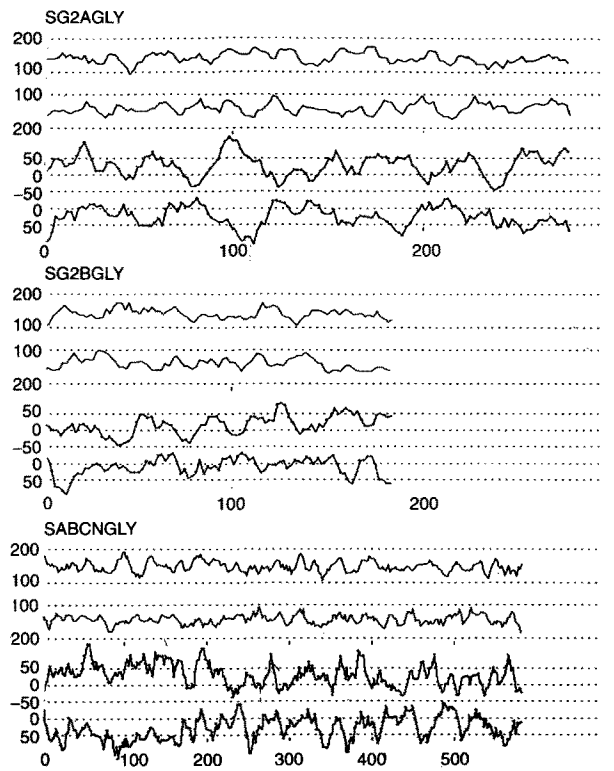

FIG. 3. Volume and amphiphilic profiles of major soybean cotyledon proteins. Properties are displayed as in Figure 1. SG2AGLY, G2 acidic glycinin; SG2BGLY, G2 basic glycinin; SABCNGLY, β-conglycinin, α subunit.

Conversely, it seems likely that SG2AGLY, which exhibits a much different amphiphilicity profile, will interact differently with self and neighboring molecules. The synchronous distribution of polarity along both sides of SG2AGLY is reminiscent of SILK2 profiles (Fig. 2), even though the soy protein is more polar and irregular. Silk fiber is thought to consist of a series of ordered regions, stacked antiparallel β-sheets derived from multiple peptide chains, separated at regular intervals by less polar polyalanine segments of undefined structure (18). Perhaps SG2AGLY contributes texture to soy protein preparations by analogous association of ordered regions that are stabilized through ionic bonding.

Considering that ionizable residues constitute 20 and 28% of the residues in SG2BGLY or SG2AGLY, respectively, it is obvious that pH will affect substantially any interactions in which they participate. The same is true for SABCNGLY, in which 36% of the amino acids are ionizable.

Both sides of SABCNGLY are essentially equal overall, but less than a third of the molecule (residues 1 through 175) concentrates essentially half of the acidic and basic residues and only 11% of the most hydrophobic residues. In β-casein (17), a component of commercial casein glues, half of the molecule's acids and bases are likewise concentrated near the N-terminus, but overall the molecule is less polar than SABCNGLY by nearly 20 a.u. per residue.

If SABCNGLY could be cleaved at residue 175, the per-residue amphiphilicity of the N-terminal fragment would nearly double to 52.6 a.u. while that of the C-terminal fragment would fall to 20.6 a.u., leaving it still more polar than a random sequence (Table 3). Polarity and volume would be about the same on either side of the C-terminal fragment, but 70% of its tyrosines would be on Side A, 60% of its serine and threonine on Side B.

Because charged residues enhance solubility, and bulky aromatic or basic residues commonly provide sites for enzymic proteolysis, it is not surprising that the soy cotyledon proteins, with such residues distributed throughout their sequences, find good use in foods. Unfortunately, these same advantages tend to diminish any ability of the unmodified proteins to substitute directly for hydrocarbon polymers.

Other soybean proteins appear better suited to nonfood use in films, coatings, plastics, etc. Figure 4 gives profiles for two examples of unusually repetitive sequences. One of these, SHIPRO, produces a remarkably uniform profile due to highly repetitive structure in which the sequence Pro-Val-Glu-Pro-Pro-Val-Tyr-Lys-Pro is duplicated some 18 times with only five to six amino acid replacements. Averaging inherent in the moving window analysis can produce essentially invariant profiles for such repetitive sequences. Actually, the volume dispersion for SHIPRO is 25 $Å^3$ and the amphiphilicity dispersion is ±53. In the β-sheet conformation, SHIPRO is as asymmetric as SG2BGLY. Differences between Side A and Side B evidence size and polarity distributions that would be consistent with a natural tendency toward β-sheet structure and possible film formation (19). This asymmetry is due primarily to the curious distribution of essentially all of the

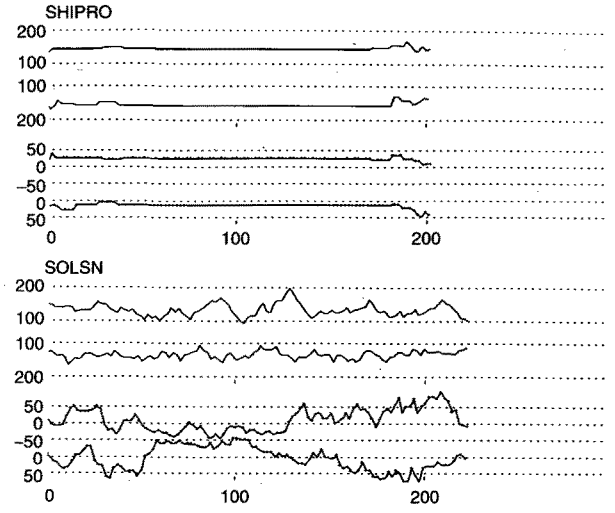

FIG. 4. Volume and amphiphilic profiles of a proline-rich protein from soybean seed coat (SHIPRO) and a soy oil-body oleosin (SOLSN). Properties are displayed as in Figure 1.

molecule's acidic residues along Side A and more than 90% of its tyrosine residues along Side B, which is an even more extreme segregation than that seen in SILK2.

Though treated here as a high-proline sequence, SHIPRO is analogous in both structure and occurrence to a group of hydroxyproline-rich glycoproteins known as extensins (20). The gene for this soybean protein is highly expressed in the root and immature seed coat (12). Quite likely it is hydroxylated and glycosylated like other extensins if it survives in the mature seed. Hydroxylation, of course, would increase the protein's volume and amphiphilicity from levels given above. Although its uniform structure would not change markedly, a hydroxylated SHIPRO probably would find stable conformations that are quite different from those preferred by the non-hydroxylated version.

Another soybean protein, SOLSN (Fig. 4), is especially interesting because it likely occurs unmodified in the mature bean and, therefore, should offer better prospects for availability. Oil bodies isolated from seeds of several oil-bearing plants contain oleosins, which are proteins that share highly conserved and extremely hydrophobic sequences, generally 70–80 residues in length (21). Tzen and co-workers (22) estimate that such proteins may account for as much as 7% of total protein in rapeseed. They further propose that the hydrophobic domains in these oleosins form antiparallel β-structure loops that protrude into the triglyceride-rich matrix of the oil body while polar portions of the proteins combine with phospholipids to constitute the lipid body membrane (21).

Overall, the soybean oleosin identified by Kalinski and co-workers (13), SOLSN, is smaller per-residue (134 Å^3) and much less polar (8.2 a.u.) than a random protein (Table 3). An 85-residue peptide that should be released by enzymic digestion of SOLSN with trypsin would have about the same per-residue size, but it would be much more hydrophobic (−24.2 a.u.) and might, therefore, make an excellent candidate for inclusion in plastics.

Table 4 summarizes pertinent quantitative data from analyses of the proteins discussed above. A comparison of residue volumes and amphiphilicities evidences marked differences among the examined proteins. The major soy cotyledon proteins and fragments from them are more like the fibrin-forming element of fibrinogen than silk or collagen.

Volume dispersion values suggest that, in terms of space filling and packing, soy proteins are as uniform or perhaps more uniform than most proteins (Tables 3 and 4). However, they definitely lack the uniform polarity typical of hydrocar-

TABLE 4
Residue Volumes and Amphiphilicities of β-Sheet Proteins[a]

Protein	Volume (Å^3)	Dispersion[b] (Å^3)	Amphiphilicity (a.u.)	Dispersion[b] (a.u.)
SG2AGLY				
Side A	142	37	27.4	63
Side B	141	38	27.7	66
SG2BGLY				
Side A	140	37	14.9	60
30–160[c]	143	37	12.8	61
Side B	140	38	10.9	66
30–160	137	39	3.0	58
SABCNGLY				
Side A	148	35	28.9	69
1–175	150	33	50.8	67
176–583	146	35	19.4	68
Side B	143	35	31.5	67
1–175	143	29	54.5	55
176–583	144	37	21.8	69
SHIPRO				
Side A	146	20	25.9	55
Side B	153	31	15.3	51
SOLSN				
Side A	134	39	9.9	64
49–133	134	46	−19.4	43
Side B	134	37	6.4	61
49–133	134	38	−29.2	38
SILK2				
Side A	110	42	0.3	34
Side B	106	39	0.4	37
HCOLA1X				
Side A	125	44	5.0	49
80–500	115	41	5.6	46
Side B	124	43	7.3	54
80–500	116	43	10.6	55
HGFBGN				
Side A	141	41	21.9	61
Side B	142	43	12.0	58
BCASEINB				
Side A	147	32	11.0	61
1–48	142	29	39.5	65
49–209	149	33	2.6	57
Side B	145	29	10.4	53
1–48	147	27	29.8	60
49–209	144	30	4.5	49

[a]Average per-residue volume or amphiphilicity. See Table 2 for protein designation.
[b]Standard deviation of residue volume or amphiphilicity values for sequence. a.u., Arbitrary units.
[c]Inclusive residue numbers.

bon polymers. Polarity varies about twice as much in soybean proteins as it does in silk.

Only values for the soy oleosin fragment approach the per-residue volumes and uniform polarity of silk or collagen while exceeding their hydrophobicities. Similarly, a fragment from SG2BGLY approximates the polarity of collagen better than does the whole soy protein. The challenge to make useful materials from soybean proteins may be as much a challenge to identify and secure useful parts from them as it is to enrich single natural components.

REFERENCES

1. Choma, C.T., J.D. Lear, M.J. Nelson, P.L. Dutton, D.E. Robertson and W.F. DeGrado, *J. Am. Chem. Soc. 116*:856 (1994).
2. Creel, H.S., M.J. Fournier, T.L. Mason and D.A. Tirrell, *Macromolecules 24*:1213 (1991).
3. Pocius, A.V., in *Encyclopedia of Chemical Technology*, Vol. 1, 4th edn., edited by J.I. Kroschwitz, Wiley, New York, 1992, p. 458.
4. Chen, R., and R.F. Doolittle, *Proc. Natl. Acad. Sci. USA 66*:472 (1970).
5. Windish, L.G., *The Soybean Pioneers*, L.G. Windish, Galva, 1983, pp. 31–35.
6. Rose, G.D., L.M. Gierasch and J.A. Smith, *Adv. Protein Chem. 37*:1 (1985).
7. Chen, S.H., and D. Bendedouch, *Methods Enzymol. 130*:79 (1986).
8. Eisenberg, D., R.M. Weiss and T.C. Terwilliger, *Proc. Natl. Acad. Sci. USA 81*:140 (1984).
9. McCaldon, P., and P. Argos, *Proteins 4*:99 (1988).
10. Nielsen, N.C., C.D. Dickinson, T-J. Cho, V.H. Thanh, B.J. Scallon, R.L. Fischer, T.L. Sims, G.N. Drews and R.B. Goldberg, *Plant Cell 1*:313 (1989).
11. Sebastiani, F.L., L.B. Farrell, M.A. Schuler and R.N. Beachy, *Plant Mol. Biol. 15*:197 (1990).
12. Hong, J.C., R.T. Nagao and J.L. Key, *J. Biol. Chem. 265*:2470 (1990).
13. Kalinski, A., D.S. Loer, J.M. Weisemann, B.F. Matthews and E.M. Herman, *Plant Mol. Biol. 17*:1095 (1991).
14. Hinman, M.B., and R.V. Lewis, *J. Biol. Chem. 267*:19320 (1992).
15. Thomas, J.T., C.J. Cresswell, B. Rash, H. Nicolai, T. Jones, E. Solomon, M.E. Grant and R.P. Boot-Handford, *Biochem. J. 280*:617 (1991).
16. Watt, K.W.K., T. Takagi and R.F. Doolittle, *Proc. Natl. Acad. Sci. USA 75*:1731 (1978).
17. Ribadeau-Dumas, B., G. Brignon, F. Grosclaude and J-C. Mercier, *Eur. J. Biochem. 25*:505 (1972).
18. Xu, M., and R.V. Lewis, *Proc. Natl. Acad. Sci. USA 87*:7120 (1990).
19. Zhang, S., T. Holmes, C. Lockshin and A. Rich, *Ibid. 90*:3334 (1993).
20. Cassab, G.I., and J.E. Varner, *Annu. Rev. Plant Physiol. 39*:321 (1988).
21. Tzen, J.T.C., G.C. Lie and A.H.C. Huang, *J. Biol. Chem. 267*:15626 (1992).
22. Tzen, J.T.C, Y-z. Cao, P. Laurent, C. Ratnayake and A.H.C. Huang, *Plant Physiol. 101*:267 (1993).

[Received August 25, 1994; accepted January 16, 1995]