The EMBO Journal vol.4 no.5 pp.1111-1117, 1985

Structural similarity between legumin and vicilin storage proteins from legumes

Patrick Argos[1,3], S.V.L. Narayana[1] and Niels C. Nielsen[2,4]

The United States Department of Agriculture[2] and the Departments of Biological Sciences[1] and Agronomy[2], Purdue University, West Lafayette, IN 47907, USA

[3]On sabbatical leave at the European Molecular Biology Laboratory, Biological Structures Division, Postfach 10.2209, Meyerhofstrasse 1, 6900 Heidelberg, FRG. [4]On temporary assignment at the Plant Breeding Institute, Maris Lane, Trumpington, Cambridge CB2 2LQ, UK until September 1985

Communicated by D. von Wettstein

The primary structures for several members of both the vicilin and legumin families of storage proteins were examined using a computer routine based on amino acid physical characteristics. The comparison algorithm revealed that sequences from the two families could be aligned and share a number of predicted secondary structural features. The COOH-terminal half of the subunits in both families displayed a highly conserved core region that was largely hydrophobic and in which a high proportion of the residues were predicted to be in β-sheet conformations. The central region of the molecules which contained mixed areas of predicted helical and sheet conformations showed more variability in residue selection than the COOH-terminal regions. The NH_2-terminal segments of subunits from the two different families could not be aligned though they characteristically had a high proportion of residues predicted to be in helical conformations. The feature which most clearly distinguished subunits between the two families was an inserted span in the legumin group with a high proportion of acidic amino acids located between the central and COOH-terminal domains. Residues in this insertion were predicted to exist mainly in helical conformation. Since considerable size variation occurs in this area amongst the legumin subunits, alterations in this region may have a minimal detrimental effect on the structure of the proteins.

Key words: legumes/storage proteins/vicilin/legumin/structure

Introduction

Many legumes accumulate large amounts of protein in their seeds. Often these proteins are devoid of catalytic activity and located in specialized subcellular compartments referred to as protein bodies. Because the proteins are mobilized after germination to support the early growth and development of seedlings, they are called storage proteins. For the agronomically important cereals and grain legumes, the molecules make a major contribution to the nutritional quality of the seeds as well as to the 'functional' properties of foods derived from them.

The grain legumes generally contain two types of storage proteins that are distinguishable in size and sugar content (for review, see Darbyshire *et al.*, 1976). Proteins in the first group which have sedimentation coefficients between 10.5S and 13.0S are hexamers devoid of carbohydrate. They are referred to as 'legumin-like' or 11S proteins (Osborne, 1924), although trivial names based on the botanical name of the plants from which they were purified are also used (e.g., glycinin from *Glycine max*). The second group of proteins have smaller sedimentation coefficients (7.0 – 9.0S) and are generally isolated from seed extracts as trimers of glycosylated subunits. This group of proteins is referred to as 'vicilin-like' or 7S proteins (Osborne, 1924), although trivial names are once again used (e.g., conglycinin from *Glycine max*; phaseolin from *Phaseolus vulgaris*). Many grain legumes contain both the legumin-like and vicilin-like proteins although some species contain exclusively either the 11S or 7S protein complexes (Dudman and Millerd, 1975).

While a number of recent studies have provided information about the primary structures of the legumins and vicilins (Nielsen, 1983; Marco *et al.*, 1984; Scallon *et al.*, 1985; Slightom *et al.*, 1983; Lycette *et al.*, 1983,1984; Schuler *et al.*, 1982a,1982b), little is known about their primary, secondary and tertiary structural relationships. During the last several years, computer programs have been developed that allow comparison of sequences and prediction of protein secondary conformation from primary structures. The data presented in this communication were obtained from a computer-assisted comparison and prediction of legume proteins and revealed interesting similarities between the structures of 11S and 7S protein subunits. The studies also revealed the presence of a large hypervariable region in the legumin-like subunits that appears to be a good site for modification of the proteins to improve seed quality.

Results

Nomenclature

The abbreviations adopted for the names of the proteins compared in this study, as well as primary references to their amino acid sequences, are given in Table I. This abbreviated nomenclature will be used throughout the remainder of the manuscript.

Primary structure alignments

Primary structures of the vicilin family included those of VCLN from pea, PHAS from common bean and the Cgy2 subunit of β-conglycinin from soybean. Examples of the legumin family included the Gy2 and Gy4 subunits from soybean glycinin and LEG from pea. The alignment procedure was first used to compare all pairs of proteins within each of the two groups. In each case the alignments were obvious. Figure 1 exemplifies the results for one such intragroup comparison, that of VCLN *versus* PHAS. The search matrix elements used to match the sequences had

Table I. Proteins whose primary structures were aligned in this report

Protein	Species	Abbreviations	References
Vicilin	*Pisum sativum*	VCLN	Lycette *et al.* (1984)
Phaseolin	*Phaseolus vulgaris*	PHAS	Slightom *et al.* (1983)
β-Conglycinin	*Glycine max.*	Cgy2	Schuler *et al.* (1982a,b)
Glycinin 2	*Glycine max.*	Gy2	Nielsen (1983) Marco *et al.* (1984)
Glycinin 4	*Glycine max.*	Gy4	Scallon *et al.* (1985)
Legumin	*Pisum sativum*	LEG	Lycette *et al.* (1984)

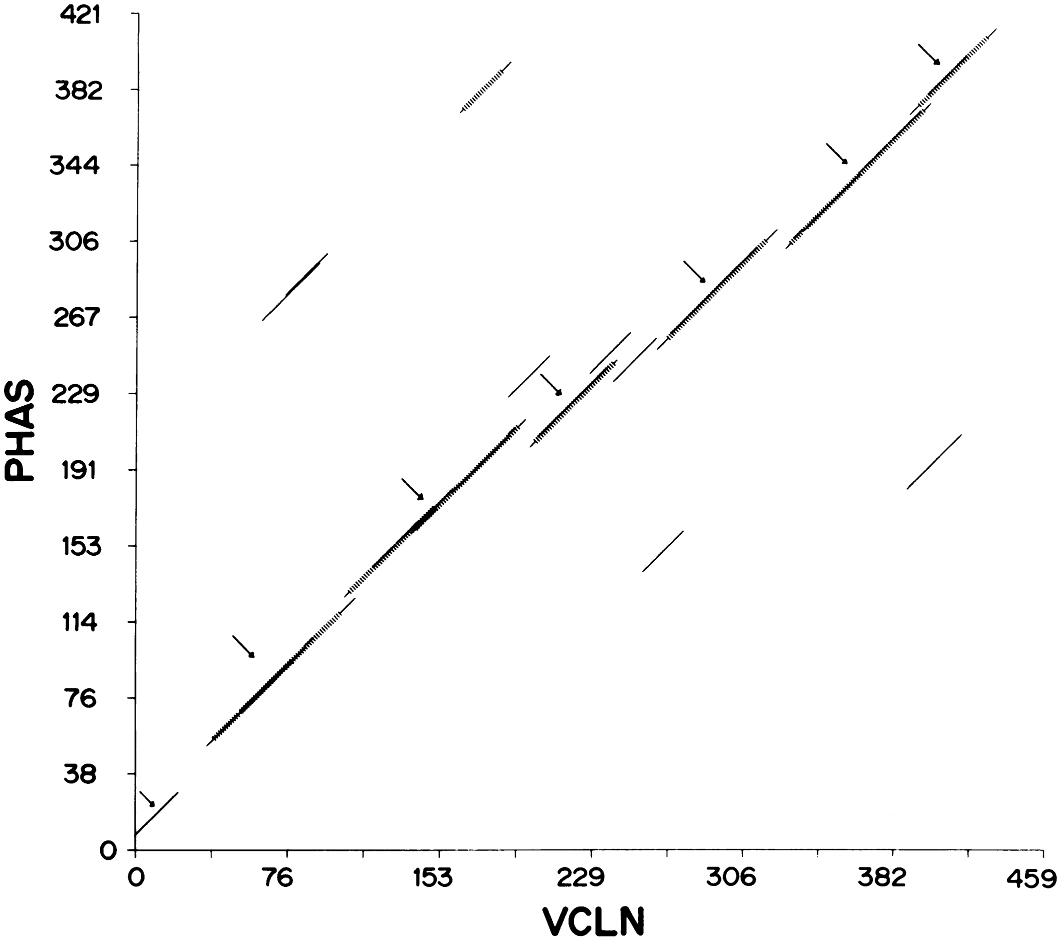

Fig. 1. The structural homology search matrix for PHAS and VCLN. The search window was 30 residues in length. Symbols chosen to indicate the standard deviation (σ) fractional ranges of the search value (S) are: $3.9\sigma \leq S < 4.7\sigma$ (symbol −); $4.7\sigma \leq S < 5.5\sigma$ (symbol 1); $5.5\sigma \leq S < 6.3\sigma$ (symbol ↿); $6.3\sigma \leq S < 7.1\sigma$ (symbol +); and $7.1\sigma \leq S < 7.9\sigma$ (symbol Y). The symbols are placed over the 30 residue segment with appropriate search score. The symbol corresponding to the higher standard deviation fraction was chosen where overlapping positions associated with different matrix values occurred. Arrows indicate the search peaks used to align the sequences.

scores at the 5.0σ level or greater. The theoretical probability of such a matrix value occurring randomly is $\sim 10^{-6}$ (McLachlan, 1971).

The alignment search was then performed amongst all proteins in the vicilin group with all those in the legumin cluster. The best intergroup result was obtained for the PHAS *versus* Gy2 comparison, and the plot of this search matrix is shown in Figure 2. The alignments were made using search values of 3.5σ or greater. While the primary structures for Cgy2 and Gy4 are incomplete, those parts which were tested aligned well with the corresponding regions of the other proteins. The results indicated that there was a clear relationship between the two groups of proteins. The alignments were most obvious in the COOH-terminal half of the polypeptides (see arrows, Figure 2), but substantial homology was also present in the central parts of the molecules.

To obtain maximum alignment, adjustments were made in short transition spans which related high score regions of different stagger relationships (e.g., regions where the difference between matched sequence numbers changed). The adjustments were made visually, while maintaining as much as possible the hydrophobic and polar character of the aligned amino acids. The results are shown in Figure 3. Alignment position numbers were assigned as shown in the figure (every match column is counted regardless of deletions or insertions in some of the proteins), and this nomenclature has been used for the purposes of discussion throughout the remainder of the manuscript.

The minimum base change per codon (MBC/C) was calculated for all possible aligned five-residue spans, and for all possible pairwise comparisons. A match that involved a deleted region was assigned an MBC/C value of 3. The mean MBC/C values were averaged over the 15 possible pairs for all five-residue spans. Two regions (underlined in Figure 3) had mean MBC/C values of 0.7 or less, where random is considered near 1.5 (Keim *et al.*, 1981). In each of these spans, the smallest five-residue mean MBC/C value was near 0.56. The low MBC/C areas generally consisted of hydrophobic residues in the COOH-terminal regions of the respective proteins where the best agreement in matched residues was observed for the intragroup comparisons. Since these storage proteins have no known catalytic function, the highly conserved areas probably exist for structural reasons such as maintenance of interior hydrophobic clusters within subunits or perhaps for oligomeric associations between subunits in the complexes.

Particularly noteworthy in the alignments of Figure 3 are the long insertions (match positions 248–343) in Gy2, and especially Gy4 and LEG, compared with the vicilin group. The comparative magnitude of the insertions account for the previously reported size differences which divide the glycinin subunits into two groups (Nielsen, 1983). The Group-1 glycinins (Gy1, Gy2, Gy3) are

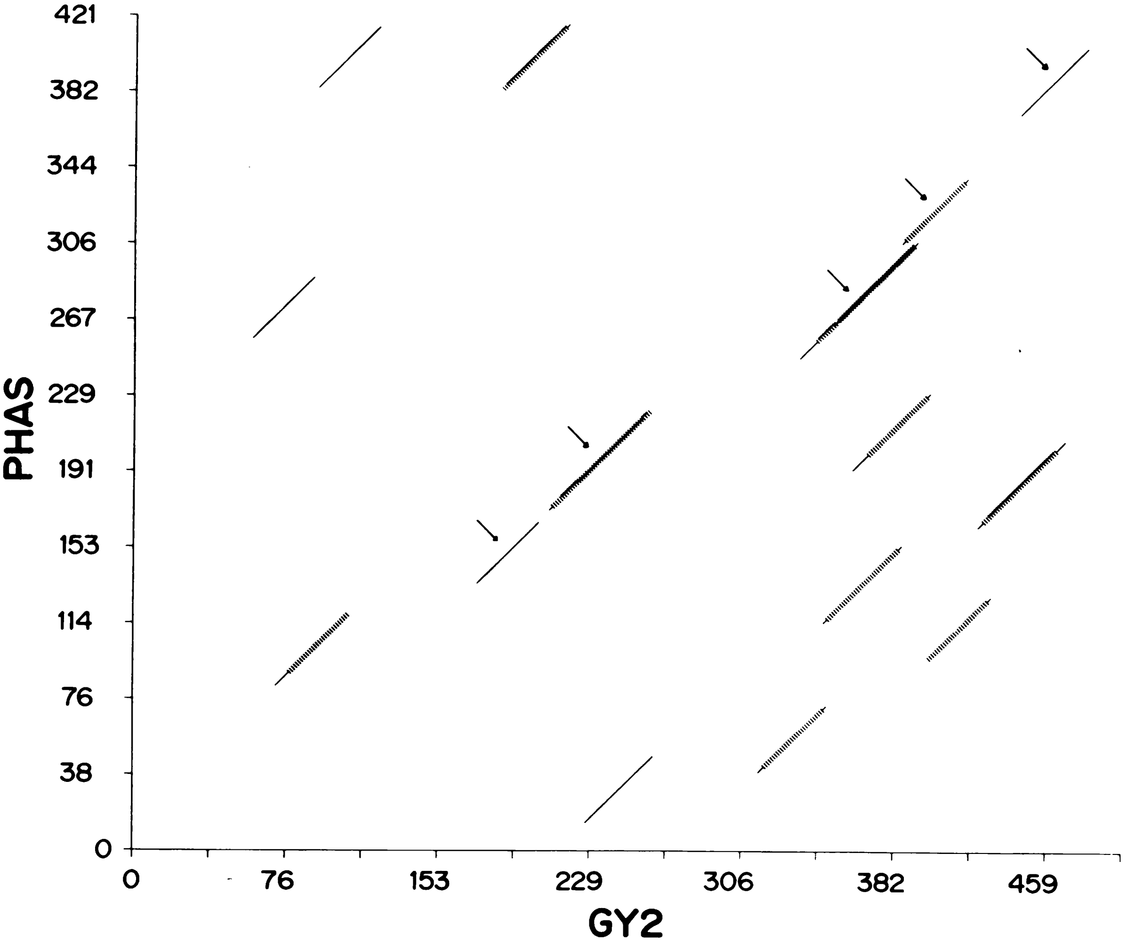

Fig. 2. The structural homology search matrix for PHAS and Gy2. The search window was 30 residues in length. Symbols chosen to indicate the standard deviation (σ) fractional ranges of the search value (S) are: $3.7\sigma \leq S < 4.0\sigma$ (symbol −); $4.0\sigma \leq S < 4.3\sigma$ (symbol 1); $4.3\sigma \leq S < 4.6\sigma$ (symbol ↾); $4.6\sigma \leq S < 4.9\sigma$ (symbol +); and $4.9\sigma \leq S < 5.2\sigma$ (symbol Y). The symbols are placed over the 30 residue segment with appropriate search score. The symbol corresponding to the higher standard deviation fraction was chosen where overlapping positions associated with different matrix values occurred. Arrows annotate the search peaks used to align the sequences.

uniform in size (e.g., 58 000), while the Group-2 (Gy4, Gy5) are larger and more variable in size (e.g., 62 000 − 68 000). Furthermore, the two groups are distinguishable by their degree of sequence homology. Within-group homologies exceed 90% while the between-group value is only 50 − 60% (Nielsen, 1983). The LEG subunits from pea contained a large insert as the Group-2 glycinins in soybean.

While between-group comparisons discussed in this report permitted alignment of the COOH-terminal and central parts of the molecules, the corresponding NH_2-terminal regions could not be aligned. However, it is still possible that the two groups share similar NH_2-terminal structure. To emphasize this difference, the NH_2-terminal regions of the 7S group are shown in Figure 3, while the corresponding regions of the 11S group are given in Figure 4.

Mean correlation coefficients for aligned amino acids were calculated using the six physical characteristics listed in Materials and methods. The values obtained for each of the 15 pairwise combinations are given in Table II. All of them were better than random (0.0) and compared favourably with those proteins whose three-dimensional structures are known to be similar, but whose primary structures are sufficiently different that their relationships are not easily discernible. For example, the NAD-binding domains in the lactate, glyceraldehyde 6-phosphate and alcohol dehydrogenases have mean correlation coefficients near 0.20 for each of the three pairwise comparisons (Otto *et al.*, 1980), while those listed in Table II are 0.37 or greater. These data reinforce the conclusion that the legumins and vicilins exhibit common conformations. As anticipated on the basis of visual examinations, however, within-group comparisons had higher correlation coefficients than in the case of the intergroup comparisons.

Secondary structure predictions

The consensus structure prediction for the six polypeptides compared in this study was calculated as described in Materials and methods. The mean prediction at each aligned position for all six proteins is given in Figure 5. Because the NH_2-terminal regions of the two groups of proteins could not be aligned, the data shown in Figure 5 for the first 93 residues reflect only the vicilins. The corresponding regions for the legumins are given in Figure 6. From alignment positions 93 onwards, however, the data shown in Figure 5 represent the consensus secondary structure prediction for all six polypeptides. To facilitate comparison, the three secondary structural states as well as the coil configuration have also been summarized in Figure 3 for each alignment position.

Examination of these data revealed that the aligned sequences could be separated into four spans (Table III). The NH_2-terminal span (e.g., ~ 100 residues for the 7S group, and 130 for the 11S group) were predicted to be largely in helical and turn confor-

```
       1                10                  20                  30                  40                  50                  60
VCLN                     M L L A I A F L A S V C V S - - S R - S D G E - - - - N P F I F K S N R - F Q T L Y E N E N G H I R L
PHAS   M M R A R V P L L - L L G I L F L A S L S A S F A T S L R E E E E S Q D N P F Y F N S D N S W N T L F K N Q Y G H I R V

       α α α α α α c c c α α α α α α α α α α α α α α α β β β α α α α α α α t t t t t t t t t t t t t β β β c c t t t t β β β β

                        70                  80                  90                 100                 110                 120
VCLN   L Q K F D K R S K I F E N L Q N Y R L L E Y K S K P H T L F L P Q Y T D A D F I L V V L S G K A T L T V L K S N D R N S
PHAS   L Q R F D Q Q S K R L Q N L E D Y R L V E F R S K P E T L L L P Q Q A D A E L L L V V R S G S A I L V L V K P D D R R E
Gy2                                                                     H R E R E G D L I A V P T G - V A W W M Y N N E D T P V
LEG                                                                     N R F R E G D I I A V P T G - I V F W M Y N D Q D T P V
                                                                                     . . . +   . +   .   .   . .     . +

       α α α α α t t t α α α α α α α α α α α α α α t t t t c c α α α α α c α α α c c c β β β β t t t β β β β β β t t t t t t β

                       130                 140                 150                 160                 170                 180
VCLN   F N L E R G D - - - - - - A I K L F A G S I A Y F A N R D D N E E P R V L D L A I P V N K P G Q L Q S F L L S G T Q N Q
Cgy2                                                                 D E N L R M I I L A I P V N K P G R F E/S F F L S S T Q A Q
PHAS   Y F F L T S D N P I F S D H Q K I P A G T I F Y L V N P [illegible] P K E D L R I I Q L A M P V N N P Q I H/E - F F L S S T E A Q
Gy2    V A V S I I D T N S L - E N Q L D Q M P R/R F Y L A G N Q E Q E F L K Y Q Q Q Q Q G G S Q S Q K G K - - Q Q E E E N E G
LEG    I A V S L T D I R S S - N D Q L D Q M P R R F Y L A G N H E Q E F L Q Y Q H Q Q - - G Q E - Q K - - - - - - - - E N E G
       .   .       +                       . .     . + . .         . +   . . .           .   .                           .

       β β β β β t t t t t t t α α α α α t t t t β β β β t t t t t t c c c β β β β β β β t t t t t t t α α α α α α α α t t t t

                       190                 200                 210                 220                 230                 240
VCLN   K S S L S G F S K N I L E A A F N T N Y E E I E K V L L E Q Q E Q E P Q H R R S L K D R R Q E I N E E N V I V K V S R -
Cgy2   Q S Y L Q G F S K N I L E A S Y D/T K F E E I N K V L F G R E E G Q Q Q - - - - - - - - G E E R L Q E S V I V E I S K -
PHAS   Q S Y L Q E F S K H I L E A S F N/S K F E E I N R V L F E E E G Q - - - - - - - - - - - - - - - - Q E G V I V N I D S E
Gy2    S N I L S G F A P E F L K E A F G V N M Q I V R N L Q G E N E E E - - - - - - - - - - - - - - - - D S G A I V T V K G G
Gy4                                                                                                                  I V T V K S -
LEG    N N I F S G F K R D Y L E D A F N V N R H I V D R L Q G R N E D E - - - - - - - - - - - - - - - - E K R A I V K I K G G
           . .   + +   . . . + +     + .     . .   . . . .       . + . .                                 .     . * *   .

       t t β β β t t t α α α α α α α t t t α α α α α α α α α c c c c c t t t α α α α α α α α α c c c c c c c β β β β β β t t t

                       250                 260                 270                 280                 290                 300
VCLN   - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Cgy2   - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
PHAS   Q I K E L S K - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Gy2    L R V T A P A M R K P Q - - - Q E E D D D E E E Q - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Gy4    L S V I S P K W Q E Q Q D E D E D E D E D D E D E Q I P S H P P R R P S H G K R E Q D E D E D E D E D K P R P T R P S Q
LEG    L A I I S P P E K Q A R H Q R G S P Q E E D E D E E - K - - Q P R - H Q R G S R Q E E E E D E D E E R Q P R - - - - H Q

       β β β β α α α α α α t t α α α α α α α α α α α α α α t t t t t t t t t t t t α α α α α α α α α α α α α t t t t t t t t t

                       310                 320                 330                 340                 350                 360
VCLN   - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - D Q I E E L S K N A K S S S K K S
Cgy2   - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - K Q I R E L S K H A K S S S R K T
PHAS   - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - E Q I K E L S K H A K S S S R K S
Gy2    - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - P - Q C V E T D K G C Q R Q S K R S R N G I D E T
Gy4    G K R E Q D Q D Q D E D E D E D E D Q P R K S R E W R S K K T Q P R R P R Q E E P R E R G C E T - - - - - R N G V E E N
LEG    R R R G - - - - - - - - E E E E E D - - - - - - - - - - K K - E - R R G S Q K G K S R R Q G D - - - - - - - N G L E E T
                                                                                             . + . .     + +     .           .

       t t t c c c t t t α α α α α α α α α t t t t t α α α α α α α t t t t t t t t c c c t t α α α α α α α α α α α α τ τ τ β β

                       370                 380                 390                 400                 410                 420
VCLN   V S S E S G P F N L R S R - - N P I Y S N K F G K F F - F I T P E K N Q Q L Q D L D I F V N S V D I K V G S L L L P N Y
Cgy2   I S S E D K P F N L G S R - - D P I Y S N K L G K L F - E I T Q - R D P Q L R D L D V F L S V V D M N E/G A L F L P H E
PHAS   L S K Q D - - - - - - - - - - - - - - - N T I G N E F G N L T E - R T - D N S - L N V L I S S I E M E E/G A L F V P H Y
Gy2    I C T M R L R Q N I G Q N S S P D I Y N P Q A G S I T - T A T S L D F P A L W L L K L S A Q Y G S L R K/N A M F V P H Y
Gy4    I C T L K L H E N I A R P S R A D F Y N P K A G R I S - T L N S L T L P A L R Q F Q L S A Q Y V V L Y K/N G I Y S P H W
LEG    V C T A K L R L N I G P S S S P D I Y N P E A G R I K - T V T S L D L P V L R W L K L S A E H G S L N K N T H F V P H Y
       .               + .               + *       . * . .       . +         +   +     + . .   . . . .   .         . + . * + +

       β β β c c c c t t t t t t t t t t t t t t t t t t β β β β β β β α α α α α α α α α α α β β β β β β c c t t t β β β β t t

                       430                 440                 450                 460                 470                 480
VCLN   N S R A I V I V T V T E G K G D F E L V G Q R N E N Q Q G L G K E N D K E E E Q E E E T S K Q V Q L Y R A K L S P G D V
Cgy2   N S K A I V V L V I N E G E A N I E L V G I K E Q Q Q R Q Q - - - - - - - - - - Q E E Q P L E V R K Y R A E L S E Q D I
PHAS   Y S K A I V I L V V N E G E A H V E L V G P K G N - - - - - - - - - - - - - - - - K E - T L E Y E S Y R A E L S K D D V
Gy4    T L N A N S I I Y A L N G R A L V Q V V N C N G E - - - - - - - - - - - - - - - - - - - - - - - R V F D G E I Q E G G V
Gy2    N L N A N S V I Y V T R G Q G K V R V V N C Q G N - - - - - - - - - - - - - - - - - - - - - - - A V F D G E I R R E Q L
LEG    N L N A N S I I Y A L K G R A R L Q V V N C N G N - - - - - - - - - - - - - - - - - - - - - - - T V F A G K L E A G R A
       +     *     + . . .   . *       . . . *   . . + .                                                 . .     + +       . .

       t t t β β β β β β β β t t t t β β β β β β β t t t t t t t t t t t t t t α α α α α α α α c c c c β β β β c c c c t t t β

                       490                 500                 510                 520                 530                 540
VCLN   F V I P A G H P V A I N A S S D - L N L I G L G I N A E N N E R N - F L A G E E D N V I S Q V E R - - - - - P V K E L A
Cgy2   F V I L A G Y P V V V N A T S D - L N F F A F G I N A E N N Q R N - F L A/G S K D N V I S Q I P S - - - - - Q V Q E L A
PHAS   F V I P A A Y P V A I K A T S N - V N F T G F G I N A N N N N R N - L L A/G K T D N V I S S I G R A L D G K D V L G L T
Gy4    L I V P Q N F A V A A K S Q S D N F E Y V S F K T N D P P S - I G - N L A G - - A N - - - S L L N A L P - E E V I Q H T
Gy2    L V V P Q N F V I A E Q A G D Q G F E Y I V F K T H H N A - - V T S Y L - - - K D - V F R A I P S - - - - - E V L A H S
LEG    L T V P Q N Y A V A A K S L N D R F S Y V A F K T N D R A - G I A R - L A G - T S S V I N D L P L - - - - - D V V A A T
       . + . +     .   + + .   +   + +   . . . .   +     + . .             . * + +     + + + +                     . * .   . .

       β β β β t t t β β β β β t t t t t t β β β β β β t t t t t t t t t c c c c t t t t t β β β β β β c c c c t t t β β β β β

                       550                 560                 570                 580           587
VCLN   F P G S S H E V D R L L E N Q K Q S Y F A N A Q P Q Q R E T R S Q E I K E H L Y S I L G A F
Cgy2   F P R S A K D I E N L I K S Q S E S Y F V D A Q P Q Q K E E G N K G R K G S L S S I L R A F Y
PHAS   F S G S G D E V M K L I N K Q S G S Y F V D A H H H Q - Q E Q Q K G R K G A F - - - - - - V Y
Gy4    F N L K S Q Q A R Q G K N N N P F S F L V - - P P - - - Q E S Q - - - - - - - - - - - - - - -
Gy2    Y N L R Q S Q V F E L K Y E G I G V L W L T L S L - - - - - - N K A H - P V L K S H K - - - -
LEG    F K L Q R D E A R Q L K S N N P F K F L V - - P A - - R Q S E N R A S
       +           . .   . +     . .     + . . +   .           .

       t t t t t c c c c c c c c t t t t β β β β α α α α α α α α α t t t t t t c c β β β β c c β β β
```

Fig. 3. The amino acid alignment of six proteins. The two regions with the lowest mean MBC are shown as underlined. The first residues shown for each sequence are the NH_2-terminal residue (VCLN and PHAS) or the NH_2-terminal most known residue (Cgy2 and Gy4) while the first residue shown for Gy2 and LEG are at their respective sequence positions, 131 and 134. Homology could not be discerned between the NH_2-terminal residues of Gy2 and LEG and the vicilin-like group. An (*) indicates a same-residue conservation in all six proteins while a (+) shows that at least four of six residues at a given position are the same. A (.) indicates conserved residues in at least four of the six proteins according to the following groups: (K,R); (S,T); (P,G); (Q,N,E,D); and (H,Y,W,F,I,L,V,M,C,A), where the latter large group corresponds to the hydrophobic amino acids. The regions predicted to be in a given state (α, helix; β, strand; t, turn; c, coil) are appropriately annotated. The numbering scheme given corresponds to all positions in the alignment of the six proteins regardless of insertions or deletions in some of the proteins. A (/) indicates the position of an intron in the genes which encode the proteins where known (PHAS, Cgy2, Gy2, Gy4).

mations (>80%), while the second span (match positions 101 – 244), which was nearly equal in size to the first span, exhibited mixed helical and β-sheet secondary structure. The third span, which contained the highly variable regions distinctive of the 11S subunits, was predicted to be almost exclusively helix and turn (94%). The fourth span, which corresponded to the COOH-terminal half of all the subunits, was by far the most highly conserved and the most hydrophobic. It was predicted to have mostly β-strand and turn conformations.

```
LEG   M A K L L A L S L S F C F L L L G G C F A L R E Q P Q Q N E C Q L E R L
Gy2   M A K L V - L S L - - C F L L F S G C F A L R E Q A Q Q N E C Q I Q K L
      α α α α α α α α α α α α α α α α t t t α α α α α α α t t t α α α α α α α


LEG   N A L Q P D N R I E S E G G F I E T W N P N N N E F R E C G L D L L R A
Gy2   N A L K P D N R I E S E G G F I E T W N P N N K P F Q S A G V A L S R C
      α α α α t t t t t t t t t t t β β β t t t t t t t t t t t t α α α α α α


LEG   T L Q R N A L R R P Y Y S N A P Q E I F I Q Q G N G Y F G M V F P G C P
Gy2   T L N R N A L R R P S Y T N G P Q E I Y I Q Q G N G I F G M I F P G C P
      α α α α α α α α t t t t t t t t t β β β β β t t t t t t β β β β t t t t


LEG   E T F E E P Q E S E Q G E G R R - Y R D R H Q K V
Gy2   S T Y Q E P Q E S Q Q C G R S Q R P Q D R H Q K V
      t t t α α α α α α α t t t t t t t t t t α α α α α
```

Fig. 4. The alignment of the NH_2-terminal residues for Gy2 and LEG. The remaining residues in the two proteins are given in Figure 3. Abbreviations and symbols are as given in Figure 3.

Table II. Mean correlation coefficients over six physical parameters for aligned residues (lower left of matrix).

	VCLN	Cgy2	PHAS	Gy2	Gy4	LEG
VCLN	X	303	317	307	195	298
Cgy2	0.73	X	266	255	194	246
PHAS	0.70	0.75	X	308	182	294
Gy2	0.38	0.41	0.40	X	212	459
Gy4	0.37	0.40	0.40	0.64	X	262
LEG	0.40	0.42	0.36	0.78	0.64	X

The upper right portion shows the number of matched residues in each pairwise comparison; however, it must be emphasized that some of the sequences are incomplete (Cgy2, Gy4) and that size variation exists amongst the proteins.

Exon structure

The positions of the introns in the genes which encode the 11S and 7S subunits are known for four of the proteins examined (see Table I for references). Their locations are shown by slashes in the primary structures reproduced in Figure 3. Although the intron-exon junctions appear well conserved within groups, this relationship does not extend to the between-group comparisons except near alignment position 412, which divides the conserved COOH-terminal region from the NH_2 segments and is adjacent to the linker region of 11S subunit precursors where post-translational cleavage occurs (Marco *et al.*, 1984; Lycette *et al.*, 1984). The 11S subunit genes maintain the COOH-terminal hydrophobic span (span 4) as one exon, while it was divided into two exons in the case of genes which encode the 7S subunits. The largely helical and turn NH_2-terminal regions (span 1) are contained within the first exon of genes which encode both types of subunits. Interestingly, the highly variable regions (span 3) of the 11S subunits are also contained in a single exon. The intron-exon boundaries are near glutamates or glycines in the 7S genes, while in the 11S subunits they are located near basic residues.

Discussion

Very little is known about the three-dimensional structure of the legumin-like and vicilin-like storage proteins found in legumes. Circular dichroism studies performed with proteins purified from a number of sources have in general indicated a sizeable content of β-structure. The computer predictions described here are consistent with the interpretations given for the spectral data, and indicate that roughly about two-thirds of the amino acid residues in each of the subunits are involved in β-sheet and turn conformations.

The data predict that the subunits for both types of the legume storage proteins have common structural features. For the purposes of discussion, it seems convenient to consider the molecules as being composed of three domains: an NH_2-terminal one (domain I, span 1), a central one (domain II, span 2) and a COOH-terminal one (domain III, span 4) (Figure 7). The COOH-terminal domain III includes the terminal half of all the subunits, is quite hydrophobic and contains a central highly conserved region. The physical characteristics of this domain suggest that it is probably

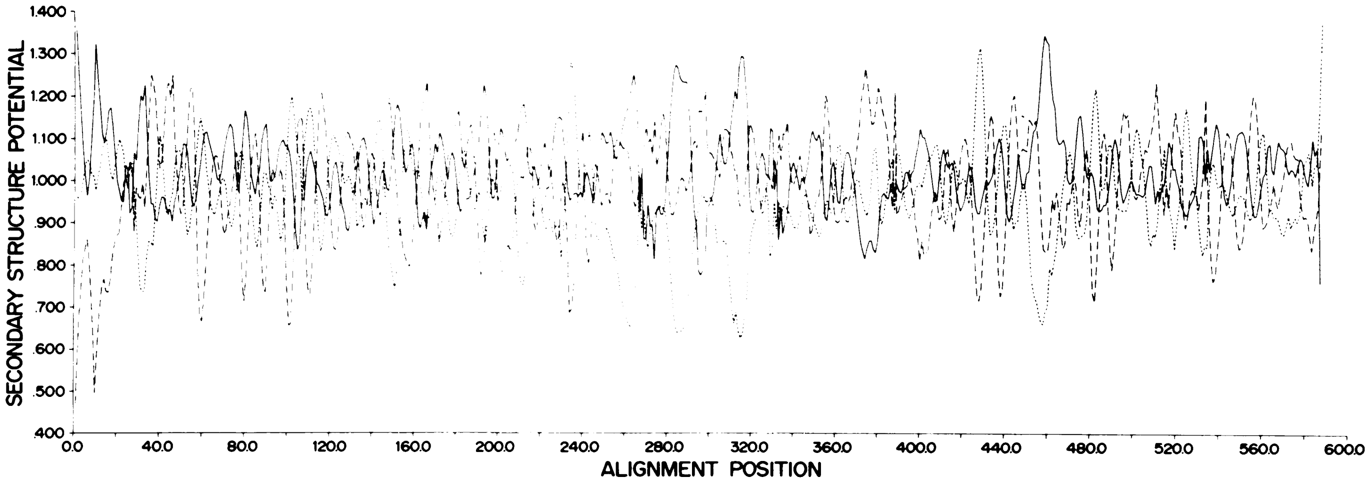

Fig. 5. The mean secondary structural predictions corresponding to the aligned residues of Figure 3. The solid line refers to the helical prediction while the short and long dashed lines indicate respectively the β-strand and turn predictions. A secondary structure potential of 1.0 would indicate a neutral preference for an amino acid to be in a given structural conformation, while values greater (less) than 1.0 refer to a preference (avoidance) of the structure (Chou and Fasman, 1974a,1974b).

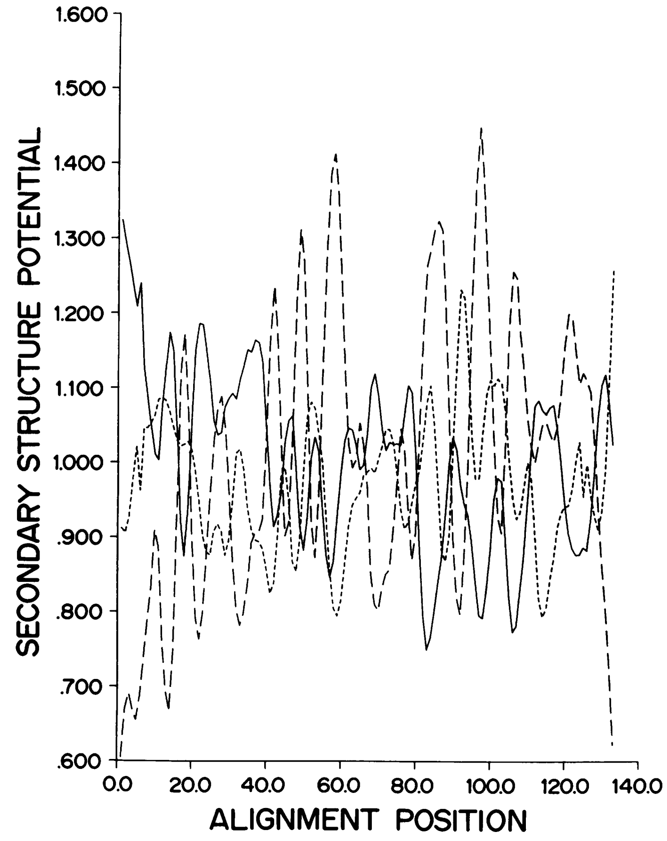

Fig. 6. The mean secondary structural prediction corresponding to the aligned residues in Figure 4. The nomenclature used is the same as for Figure 5.

Table III. Predicted secondary structural composition given as a percentages for four regions in the aligned sequences of Figure 3

Alignment positions	Helix	β-strand	Turn	Coil
1 – 100	56	10	24	10
101 – 244	25	28	36	11
245 – 355	51	0	43	6
356 – 587	13	34	38	15

buried within the subunit and plays an important role in maintenance of structure. The central domain II contains a mixture of helical and β-sheet structures. Where substitutions have occurred, the physical characteristics of the residues are for the most part maintained. The NH_2-terminal I domains consist of the first 100 – 150 amino acids depending on the subunits. These regions are predicted to contain substantially more α-helical structure than the other two domains, which may reflect their location at or near the exterior of the molecules.

The most conspicuous difference between the 11S and 7S subunits occurs at the junction between the central and COOH-terminal domains. Here the 11S molecules apparently are able to tolerate large insertions of variable size, and these insertions were predicted to have largely helical and turn conformations. Such structures are generally observed to be exposed in soluble proteins whose three-dimensional structures are known. It is of interest to point out that post-translational cleavage of the 11S subunits occurs near the COOH-terminal end of the variable regions (Nielsen, 1983; Marco *et al.*, 1984; Lycette *et al.*, 1984).

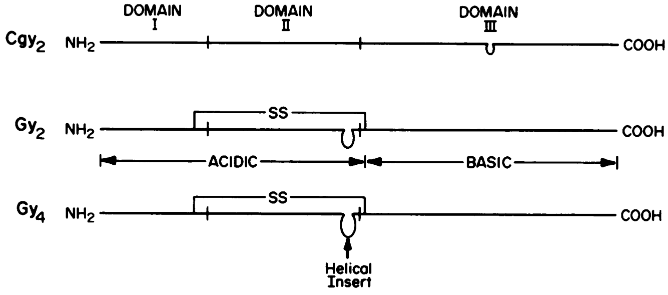

Fig. 7. Illustration of the predicted domain relationships and differences between the legumin and vicilin-like subunits from legumes. Domain I is referred to as span 1 or the NH_2-terminal region in the text. Similarly, domains II and III correspond to span 2 (central region) and 4 (COOH-terminal region), respectively. Residue inserts (span 3) are shown as loops.

The exposed nature postulated for these regions would permit easy access by the enzymes involved in post-translational modification. The two Cys residues (position 12 in Figure 4 and 362 in Figure 3) which participate in the single disulfide bond that links the subunits together after post-translational cleavage (Staswick *et al.*, 1984a, 1984b) are located just outside either end of the variable region. If areas on either side of the variable region are important for maintenance of subunit conformation, the disulfide bond would help maintain their proper relative position. Hence, cleavage near these sites would be expected to disturb minimally the subunit architecture.

The insertions are characterized by a high proportion of aspartate and glutamate. It is difficult to imagine a structurally integral helix composed largely of negatively charged residues which are likely to repel each other at neutral pH. The environment of this region is likely to be protein, and unlikely to be water; apparently the acidic charge is neutralized in some fashion, or it remains flexible.

The 7S subunits also contain small insertions, but they interrupt the highly conserved core regions of domain III. Like the insertions in the 11S subunits, they also have high glutamate and aspartate content. The 7S insertions are, however, considerably shorter than those in the 11S subunits; it is possible that they result in the extension of a short exposed loop. The high content of acidic residues in this region allows speculation that the small insertion in the vicilins occupies the same region in the subunit as the longer one associated with the 11S subunits, despite their different location in the two subunit types.

The amino acid sequence of cruciferin, a storage protein found in rape-seed, has recently been deduced by Crouch and her colleagues (personal communication). Comparison of its primary structure with that for the 11S and 7S legume subunits revealed considerable homology with LEG and Gy2. Cruciferin, however, did not contain a large insertion between its domains II and III. Rather, it contained one of ~40 residues that began near the start of homology between the 7S and the Gy2 and LEG subunits (see position 93, Figure 3). The cruciferin insertion occurs at a position considered here to be at the junction between the NH_2-terminal and central domains. The correlation of the insertion position and domain interfaces would not be expected *a priori*. The existence of such a correlation lends support to the three domain model and to the hypothesis that the extended residue additions are structurally most easily accommodated between domains which are themselves structurally stable entities.

Interest has been expressed in modifying storage protein genes to improve nutritional properties and other seed characteristics. Success in such endeavours will depend upon identification of regions within the genes which can be modified without a detrimental effect on the three-dimensional structure of the protein. The extensive natural variation in the size of the insertions in the 11S subunit suggest that this region is one where a change might be tolerated. The peculiar acidic nature of the legumin insertions, however, indicate that other structural features may severely limit the type of changes which can be made.

It is instructive to consider the general features of the 11S and 7S protein complexes in relation to the structural predictions. Legumins are usually isolated from seed extracts as a hexamer with an aggregate mol. wt. of ~350 000. The vicilins, on the other hand, frequently are purified as trimers with a mol. wt. of ~180 000. It may be significant that several members of the vicilin family undergo reversible ionic strength-dependent aggregation of trimers into hexamers (e.g., *Glycine max*, Thanh and Sibasake, 1974; *Arachis hypogaea*, Johnson and Naismith, 1953; *Vigna unguiculata* and *Lupinus albus*, Jobert, 1957). It is conceivable that the 9S forms of these complexes are the structural equivalent of the vicilin hexamers and occur *in vivo*. If this is correct, the observations by Craig *et al.* (1980a,1980b) that both pea vicilin and legumin are sequestered in the same protein bodies, could indicate that the mass within the protein bodies is a random aggregate of the two protein complexes. It will be important to establish the extent to which the tertiary and quaternary structures of the various 11S and 7S subunits differ, and to what degree they may be substituted for each other. This information could well be useful in efforts to improve seed quality.

Materials and methods

The protein sequences were correlated by comparing every possible span of length L residues in one protein with all such spans in the second protein. Two scoring procedures were used. The first was based on the Dayhoff relatedness odds matrix (Dayhoff, 1969; McLachlan, 1971; Staden, 1982) whose elements express relative weights with which amino acids substitute for one another in aligned sequences of 71 protein families. The second scoring method involved calculation of the mean correlation coefficient for each oligopeptide comparison over six residue physical characteristics thought to be the primary forces that direct protein folding (Creighton, 1978). These were helix, sheet and turn secondary structural conformational preferences, residue polarity, and two amino acid hydrophobicity measures (hydration potential and surrounding hydrophobicity). The physical characteristics have previously been listed and the methodology discussed (Argos *et al.*, 1983; Argos and Siezen, 1983; Kubota *et al.*, 1981,1982). The final search matrix used in this study involved a summation of the values determined by each of the two techniques. It was obtained by averaging the scores from the two techniques after subtracting the mean value of each matrix from all elements and then scaling the two scores such that the sum of the differences between the respective mean matrix values and those of elements greater than the mean were made equal. The standard deviation of the resulting matrix elements was subsequently calculated. The matrix that resulted for the proteins compared here had a lower noise level and indicated longer stretches for alignment than matrices calculated from either method alone.

Search matrix plots were made by attaching symbols to the matrix element values that fell within assigned fractional standard deviation ranges. No matrix value less than 3.5σ was considered. In the $3.5-5.5\sigma$ range, the theoretical probability of such a matrix value occurring randomly is between 10^{-4} and 10^{-6} (McLachlan, 1971). When the matrix data was plotted, related regions in the two primary structures being compared appeared as a series of diagonally colinear, broken lines composed of appropriate symbols. A search length of 30 residues resulted in minimal noise and was used for the data reported. Different symbols were assigned for each of the different standard deviation ranges and were plotted over the entire search length. If there was an overlap of symbols that corresponded to different standard deviation ranges, the symbol for the higher range was allowed to predominate. A search length of 30 residues resulted in minimal noise and was used for the data reported. Once the sequences had been matched, an assessment of the overall structural relatedness of the two proteins was calculated using mean correlation coefficients for all the aligned residues over the six physical characteristics.

Plots of the sequence number *versus* the conformational preference parameter [helix, β-strand, turn (Argos *et al.*, 1983)] for a given amino acid were determined for each protein sequence using a least squares smoothing procedure. Every successive group of three points (i to i+2) were fit by a least squares line and the value at (i+1) was replaced by the one calculated from the line. The smoothing process was repeated for three cycles over each of the parametric plots. The smoothed curves for each potential were averaged over the aligned sequences, a procedure which should yield a better prediction than that from any one sequence (Argos *et al.*, 1976). The structural type assigned at each aligned residue position corresponded to the largest of the three mean potentials that was greater than 1.0, the neutral preference value (Chou and Fasman, 1974a,1974b). Five successive values greater than 1.0 were required for helix initiation and three were used for strands or turns. If a proline or glycine residue occurred at the fifth or smaller position in a region predicted to be helical, the span was assigned to the coil conformation due to the rare appearance of such residues in the central or COOH-terminal parts of helices (Argos and Palau, 1982). For all other conditions (e.g., all mean potential less than 1.0), the coil structure was also predicted.

Acknowledgements

Cooperative research of USDA-ARS and Purdue University. This is Journal number 10256 of the Purdue Agricultural Experiment Station. Mention of a trademark or proprietary product does not constitute a guarantee or warranty of the product by the USDA-ARS or Purdue University. Financial support from the USDA Competitive Grants Program and the American Soybean Association Research Foundation to N.C.N. and the National Institutes of Health (GM 27100) to P.A. are gratefully acknowledged.

References

Argos,P. and Palan,J. (1982), *Int. J. Protein Peptide Res.*, **19**, 380-393.

Argos,P. and Siezen,R.J. (1983) *Eur. J. Biochem.*, **131**, 143-148.

Argos,P., Schwartz,J., Schwartz,J. (1976) *Biochim. Biophys. Acta*, **439**, 261-273.

Argos,P., Hanei,M., Wilson,J.M. and Kelley,W.M. (1983) *J. Biol. Chem.*, **258**, 6450-6457.

Chou,P.Y. and Fasman,G.D. (1974a) *Biochemistry (Wash.)*, **13**, 211-221.

Chou,P.Y. and Fasman,G.D. (1974b) *Biochemistry (Wash.)*, **13**, 222-245.

Craig,S., Goodchild,D.J. and Miller,C. (1980b) *Aust. J. Plant Physiol.*, **7**, 329-338.

Craig,S., Millerd,A. and Goodchild,D.J. (1980b) *Aust. J. Plant Physiol.*, **7**, 339-346.

Creighton,T.E. (1978) *Biophys. Mol. Biol.*, **33**, 231-298.

Darbyshire,E., Wright,D.J. and Boulter,D. (1976) *Phytochemistry*, **15**, 3-26.

Dayhoff,M.O. (1969) *Atlas of Protein Sequences and Structure*, published by National Biomedical Research Foundation, Silver Springs, MD.

Dudman,W.F. and Millerd,A. (1975) *Biochem. Systematics Ecol.*, **3**, 25-40.

Jobert,F.J. (1957) *J. S. Afr. Chem. Inst.*, **10**, 21-31.

Johnson,P. and Naismith,W.E.F. (1953) *Disc. Faraday Soc.*, **13**, 98-105.

Keim,P., Heinrikson,R.L. and Fitch,W.M. (1981) *J. Mol. Biol.*, **151**, 179-197.

Kubota,Y., Takahashi,S., Nishikawa,K. and Ooi,T. (1981) *J. Theor. Biol.*, **91**, 347-361.

Kubota,Y., Nishikawa,K., Takahashi,S. and Ooi,T. (1982) *Biochim. Biophys. Acta*, **701**, 242-252.

Lycette,G.W., Delauney,A.J., Gatehouse,J.A., Gebioy,J., Croy,R.R.D. and Boulter,D. (1983) *Nucleic Acids Res.*, **11**, 2367-2380.

Lycette,G.W., Croy,R.R.D., Shirsat,A.H. and Boulter,D. (1984) *Nucleic Acids Res.*, **12**, 4493-4506.

McLachlan,A.D. (1971) *J. Mol. Biol.*, **61**, 409-424.

Marco,Y.A., Thanh,V.H., Tumer,N.E., Scallon,B.J. and Nielsen,N.C. (1984) *J. Biol. Chem.*, **259**, 13436-13441.

Nielsen,N.C. (1983) *Philos. Trans. R. Soc. Lond., Ser. B.*, **304**, 287-296.

Osborne,T.B. (1924) *The Vegetable Proteins*, 2nd edn., published by Longmans, Green, NY.

Otto,J., Argos,P. and Rossmann,M.G. (1980) *Eur. J. Biochem.*, **109**, 325-330.

Scallon,B.J., Thanh,V.H., Floener,L.A. and Nielsen,N.C. (1985) *Theor. Appl. Gen.*, in press.

Schuler,M.A., Ladi,B.F., Pollaco,J.C., Freyer,G. and Beachy,R.N. (1982a) *Nucleic Acids Res.*, **10**, 8245-8261.

Schuler,M.A., Schmidt,E.S. and Beachy,R.N. (1982b) *Nucleic Acids Res.*, **10**, 2951-2961.

Slightom,J.L., Sun,S.M. and Hall,T.C. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 1897-1901.

Staden,R. (1982) *Nucleic Acids Res.*, **10**, 2951-2961.

Staswick,P.E., Hermodson,M.A. and Nielsen,N.C. (1984a) *J. Biol. Chem.*, **259**, 13424-13430.

Staswick,P.E., Hermodson,M.A. and Nielsen,N.C. (1984b) *J. Biol. Chem.*, **259**, 13431-13435.

Thanh,V.H. and Shibasaki,K. (1979) *J. Agric. Food Chem.*, **27**, 805-811.

Received on 28 January 1985