Journal of
Molecular Evolution

Sequence Variability in Three Wheat Germ Agglutinin Isolectins: Products of Multiple Genes in Polyploid Wheat

Christine S. Wright[1] and Natasha Raikhel[2]

[1]Department of Biochemistry and Molecular Biophysics, Medical College of Virginia/
Virginia Commonwealth University, Box 614 MCV Station, Richmond, Virginia 23298-0001, USA
[2]Michigan State University–Department of Energy Plant Research Laboratory, Michigan State University,
East Lansing, Michigan, 48824-1312, USA

Summary. Three highly homologous wheat germ isolectins (95–97%) are distinct gene products in hexaploid wheat. The amino acid sequences of two of these [wheat germ agglutinin 1 (WGA1) and 2 (WGA2)] are compared with sequence data derived from a complementary DNA (cDNA) clone for the third isolectin (WGA3). This comparison includes three corrections to earlier amino acid sequence data of both WGA1 and WGA2 at positions 109 (from Ser to Phe), 134 (from Gly to Lys), and 150 (from Gly to Trp). These reassignments are based on new results from crystal structure refinement and amino acid sequence data of WGA1, as well as the recently determined nucleotide sequence of WGA3. In addition, the C-terminal residue of WGA1 has been revised to Gly171 and now differs from WGA2 (Ala171). Four other positions, Asn9, Ala53, Gly119, and Ser123, at which WGA1 and WGA2 are identical but differ from the DNA sequence of WGA3, were also reinvestigated by amino acid sequencing techniques and confirmed.

Variability among the three isolectins is observed at a total of 10 sequence positions: 9, 53, 56, 59, 66, 93, 109, 119, 123, and 171. Pairwise comparisons indicate that WGA3 deviates to a much larger extent from WGA1 (at eight positions) and from WGA2 (at seven positions) than the latter from one another (at five positions). Eight of the 10 mutations are equally distributed between domains B and C, the two interior and more highly conserved of the four WGA domains (A, B, C, D). Correlation of the variable residues with the three-dimensional structure indicates that all except the two previously described B-domain residues, 56 and 59 (Wright and Olafsdottir 1986), are easily accommodated at the dimer surface.

WGA3 displays a higher degree of inter-domain similarity than found in WGA1 and WGA2. Of the seven variable positions that are located in the domain core (residues 3–31), five are in perfect agreement with our earlier predicted domain ancestor sequence. This suggests that of the three isolectins WGA3 is most closely related to the common ancestral molecule.

Key words: Wheat germ agglutinin — Isolectin sequence identity — Domain structure — Gene duplication

Introduction

The presence of multiple forms of an *N*-acetyl-D-glucosamine/*N*-acetyl-D-neuraminic acid (GlcNAc/NeuNAc) binding lectin in wheat has been widely documented (Allen et al. 1973; Ewart 1975; Rice and Etzler 1975; Rice 1976; Etzler 1985). Three variants of wheat germ agglutinin (WGA) were identified as distinct gene products in hexaploid wheat (*Triticum aestivum*): isolectin 1 (WGA1) derives from the A-genome, isolectin 2 (WGA2) from the D-genome, and isolectin 3 (WGA3) from the B-genome (Rice 1976; Peumans et al. 1982; Stinissen et al. 1983). Their biochemical and molecular properties are similar to the extent that under suitable conditions (acid pH) isolectin mixtures undergo

Offprint requests to: C.S. Wright

subunit interchange producing biologically active heterodimer molecules (Rice and Etzler 1975). Although they are immunologically indistinguishable (Raikhel and Pratt 1987), differences exist in their amino acid compositions, electrophoretic mobility, and affinity for specific saccharides (Allen et al. 1973; Ewart 1975; Rice and Etzler 1975; Kronis and Carver 1982).

Although the functional importance of these lectins for wheat is unclear, their sugar specificity has been exploited in a large number of in vitro studies in eukaryotic cell systems (Goldstein and Hayes 1978; Goldstein and Poretz 1986; Lis and Sharon 1986). The dimeric molecule binds saccharide at two pairs of unique sites located in the dimer interface (Wright 1984). Amino acid sequence studies and x-ray crystal structure analysis carried out on the two most abundant isolectins, WGA1 and WGA2, have indicated a subunit size of 171 amino acid residues and four differences between the two sequences (Wright et al. 1984; Wright and Olafsdottir 1986; Wright 1987). The subunit structure consists of four similarly folded 43-residue domains, each stabilized by four identically positioned disulfide bridges (Wright 1980a, 1987). Extensive regions of sequence identity in the polypeptide chain have suggested a fourfold pattern of gene duplication (Wright et al. 1985).

Recently, the complementary DNA (cDNA) nucleotide sequence for WGA3, coding for a 186-residue polypeptide chain, has been determined using a cDNA library from developing grain of tetraploid wheat (Raikhel and Wilkins 1987). This sequence starts with Gln1, as observed in the mature protein, and is extended by a 15-residue hydrophobic tail beyond the C-terminus of the mature protein. A potential glycosylation site at Asn180 was suggested. Subsequent experiments have shown that the initial translation product (pre-pro-WGA) undergoes several processing steps in the endoplasmic reticulum (Mansfield et al. 1988). Removal of the signal peptide and glycosylation are cotranslational events, producing a 23,000 molecular weight precursor (pro-WGA) with sugar-binding ability. This glycosylated precursor is posttranslationally processed by removal of the carboxyl-terminal glycopeptide to yield the mature protein present in the vacuoles.

To carry out a more meaningful evaluation of the evolutionary relationships among the three isolectins than previously reported (Wright et al. 1985), it is first necessary to establish whether all the differences observed among the three isolectin sequences are real. Because it has not been possible to isolate specific clones for WGA1 and WGA2, we have to rely on amino acid sequence information. Because the more reliable cDNA-derived amino acid sequence of WGA3 was found to differ from the WGA1 and WGA2 sequences to a much greater extent (Raikhel and Wilkins 1987) than the latter two from one another (Wright and Olafsdottir 1986), we have reexamined several questionable regions of the WGA1 sequence by improved amino acid sequencing techniques. This isolectin had been sequenced only partially previously. Furthermore, interpretation of amino acid sequence data was complicated by artifacts arising from the multiple internal domain sequence identities. Two of the three Trp in WGA2 could for that reason not be located in the original sequence study (Wright et al. 1984). The position of Trp41 was determined later through crystal structure refinement (Wright and Olafsdottir 1986; Wright 1987).

In this communication we present chemical evidence for the position of the third Trp, and for two additional reassignments in the WGA1 and WGA2 sequences, explaining earlier misleading x-ray and amino acid sequencing data. In addition, we discuss the implications of variability among the three isolectins in terms of structure, stability, and evolution.

Experimental Procedures

Sequence Determination. A nucleotide sequence for WGA3 was deduced by Raikhel and Wilkins (1987) from the cDNA clone pNVR1. The amino acid sequences for WGA1 and WGA2 were established by protein sequencing techniques and knowledge of the three-dimensional structure (Wright et al. 1984; Wright and Olafsdottir 1986). The carboxymethylated protein was digested with thermolysin and trypsin for manual peptide sequencing (Tarr 1977). The method of high performance liquid chromatography (HPLC) was employed for peptide purification and phenylthiohydantoin (PTH)-derivative identification. A complete set of thermolysin peptides could be sequenced because of their small sizes (2–18 residues in length), whereas most tryptic peptides were only analyzed for their amino acid compositions and N-terminal sequences to establish overlap. For the latter, the less definitive methods of dansylation (Gray 1972) and regeneration of the amino acid by back hydrolysis of the PTH derivative using hydriodic acid (Mendez and Lai 1975; Wright and Olafsdottir 1986) were used. Reference to the x-ray structure, which clearly indicated all disulfide bridges and a number of prominent aromatic residues, aided peptide alignment.

More recently, several peptides of WGA1 (T-9, T-13, Th-2b) were resequenced by improved techniques using an automated gas phase Edman degradation sequenator (Model 470A, Applied Biosystems Inc.) interfaced with a PTH analyzer (Model 120A).

C-terminal analysis was repeated on all three isolectins using fresh anhydrous hydrazine (Sigma) (Schroeder 1972). Protein samples of 1–5 nmol were heated in the presence of 50 μl of anhydrous hydrazine at 90°C for 20 h, and subsequently analyzed on an amino acid analyzer after evaporation.

Crystal Structure Refinement. The restrained least squares method of Hendrickson and Konnert (1980) was used to refine the structures of WGA2 at 1.8 Å resolution to an R-factor of 17.9% (F_o > 3 sigma) (Wright 1987) and WGA1 at 2.0 Å resolution to an R-factor of 16.5% (unpublished results).

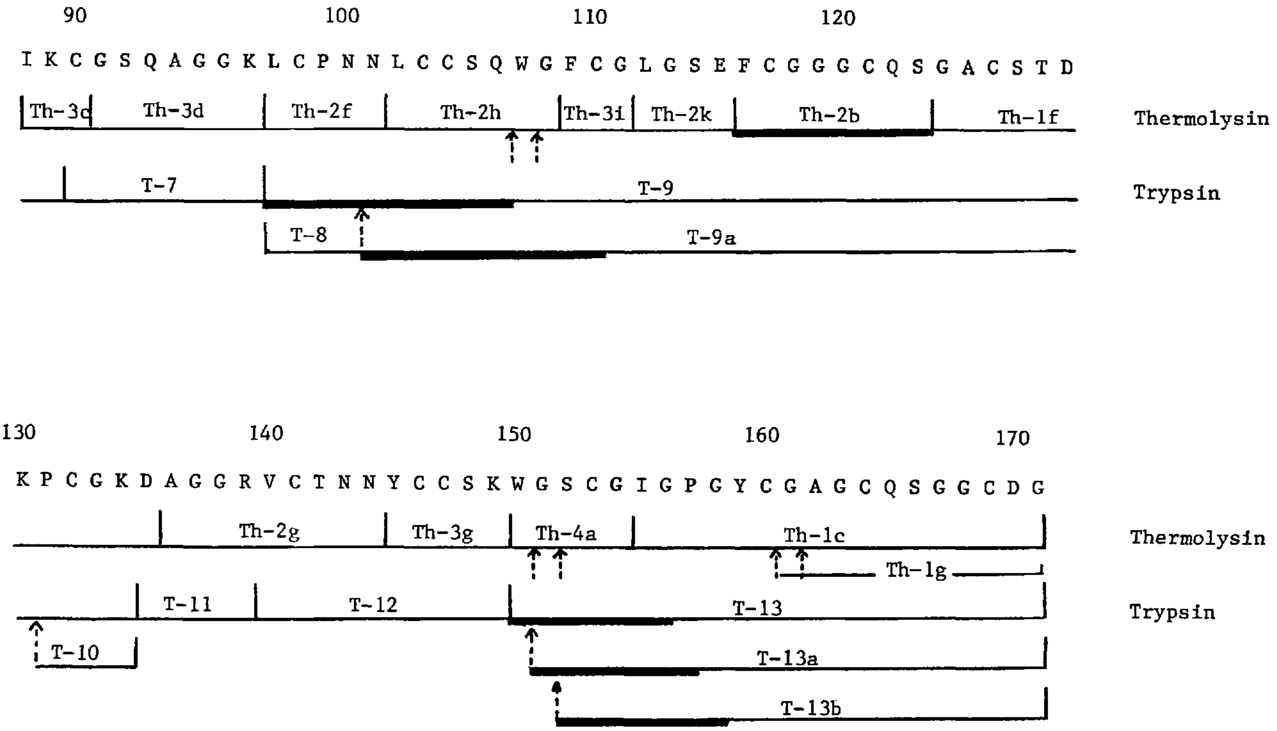
90
100
110
120
I K C G S Q A G G K L C P N N L C C S Q W G F C G L G S E F C G G G C Q S G A C S T D
Th-3c
Th-3d
Th-2f
Th-2h
Th-3i
Th-2k
Th-2b
Th-1f
Thermolysin
T-7
T-9
Trypsin
T-8
T-9a
130
140
150
160
170
K P C G K D A G G R V C T N N Y C C S K W G S C G I G P G Y C G A G C Q S G G C D G
Th-2g
Th-3g
Th-4a
Th-1c
Th-1g
Thermolysin
T-11
T-12
T-13
Trypsin
T-10
T-13a
T-13b

Fig. 1. Pattern of cleavage for trypsin (T) and thermolysin (Th) of the C- and D-domain regions of WGA1. Peptides are numbered as described in Wright et al. (1984). Dashed vertical arrows indicate limited cleavage. The solid bars represent portions of the WGA1 sequence redetermined on the gas phase sequenator.

Results

Figure 1 depicts the digestion pattern for the C- and D-domains of WGA1. We have obtained new sequence data for those regions that are marked by thick solid bars.

Clarification of the T-9 Tryptic Peptide Mixture (Region 97–134)

In the original sequence study (Wright et al. 1984) tryptic peptide T-9 had been assigned to the C-domain region 102–130 (see Fig. 1). Region 97–101 was sequenced as a separate peptide (T-8). Although a tryptic peptide for region 131–134 (T-10) could not be located, identification of these residues was made from thermolytic peptide Th-1f (124–135, see Fig. 1). Assignment of a Gly to 134 was in apparent agreement with the crystal structure, lacking electron density for a sidechain. Prompted by several discrepancies with the nucleotide sequence of WGA3 (Raikhel and Wilkins 1987), recent reexamination of T-9 of WGA1 suggested that this HPLC-purified preparation consists of a mixture of two peptides with different N-terminal sequences (T-9, T-9a, see Fig. 1) and that it contains two Lys (130 and 134) and two Phe (109 and 116).

N-Terminal Region

The T-9/T-9a mixture was sequenced 10 steps to resolve several inconsistencies in the region 107–110: (1) The DNA sequence of WGA3 was found to code for Tyr at 109 and not Ser, as proposed earlier for both WGA1 and WGA2. (2) Amino acid composition data for T-9, lacking Tyr, indicated a high Phe content (1.5) and a Ser content far below the expected value of 5.0 (see Table 1). (3) Thermolysin digestion of the WGA1 T-9 peptide mixture did not yield the Trp-containing fragment Th-4a (TrpGlySerCysGly), a peptide formerly isolated and characterized from the thermolysin digest of the complete WGA2 molecule and assigned to the C-domain positions 107–111 (Wright et al. 1984). (4) Evidence from crystal structure refinement of both WGA1 and WGA2 was inconclusive at residue 109. Its location at the surface adjacent to an extended cluster of water molecules did not suggest the presence of an aromatic sidechain.

The two N-termini of the T-9/T-9a mixture, which we identified by gas phase sequencing, are Leu97 and Asn101 (see Fig. 1). The sequence up to Cys110 was in agreement with the earlier results except that residue 109 sequenced as a Phe, not as Ser or Tyr. This necessitated reassignment of two thermolytic peptides: (1) Th-4a (TrpGlySerCysGly)

Table 1. Peptide compositions for WGA isolectins 1 and 2

	T-9/T-9a Mixture				Th-1f		Th-1g	
Peptide	WGA1	WGA2	Seq	Th-2b	WGA1	WGA2	WGA1	WGA2
CM-Cys	4.7	5.9	(7–8)	1.9 (2)	1.7 (2)	1.6 (2)	1.7 (2)	2.0 (2)
Asp	2.0	2.2	(2–3)		1.3 (1)	1.7 (2)	1.5 (1)	1.1 (1)
Thr	0.8	0.8	(1)		0.7 (1)	0.8 (1)		
Ser	2.3	2.7	(4)	0.9 (1)	0.8 (1)	1.2 (1)	1.0 (1)	0.9 (1)
Gln	2.8	2.3	(3)	1.4 (1)	0.6 (0)	0.5 (0)	1.3 (1)	1.1 (1)
Gly	8.0	5.8	(8)	3.6 (3)	1.6 (2)	2.2 (2)	4.7 (5)	3.8 (4)
Ala	1.0	1.0	(1)		1.0 (1)	1.2 (1)	1.3 (1)	1.8 (2)
Val								
Met								
Ile								
Leu	2.1	1.8	(2–3)					
Tyr								
Phe	1.5	1.3	(2)	0.8 (1)				
His								
Lys	1.7	1.4	(2)		1.8 (2)	1.4 (2)		
Arg								
Trp			(1)					
Region	97–134 101–134	97–134 101–134	97–134 101–134	116–123	124–134	124–135	161–171	161–171

Samples were submitted to 20–24-h hydrolysis in 6 N HCl at 110°C, and values shown in parentheses are derived from the sequence. CM-Cys refers to carboxymethyl-cysteine. Values shown for CM-Cys, Thr, Ser, Tyr, Met, and His are uncorrected for destruction during hydrolysis and thus tend to be low in some cases

that had been placed into region 107–111 was reassigned to the identical D-domain region 150–154 (see Figs. 1 and 2), consistent with the recent identification of a Trp at 150 (see below). (2) Th-3i (PheCysGly) was found to match region 109–111, although earlier believed to belong at positions 116–118 (Wright et al. 1984). Although it is possible that this tripeptide derived from both these closely spaced locations, the fact that thermolysin has a higher specificity for peptide bonds of hydrophobic amino acids (Gly108–Phe109, Gly111–Leu112) than for Gly–Gly bonds (Gly118–Gly119) would suggest that Th-3i originated primarily from position 109–111.

C-Terminal Region

In reexamining the earlier sequence data for Th-1f (124–135) of WGA2, two observations provided evidence for the presence of a Lys at residue 134 in WGA1 and WGA2: (1) amino acid composition data indicated the presence of more than one Lys (1.4–1.8) in peptides T-9 and Th-1f (see Table 1); and (2) in sequencing peptide Th-1f, a small peak for PTH-Lys was observed at position 134 (step 11) (see Fig. 1) but was disregarded in favor of the large PTH-Gly peak. Incomplete cleavage of the Gly133–Lys134 peptide bond coupled with poor recovery of PTH-Lys are possible explanations for the low yield of PTH-Lys at this late step.

Assuming that the Lys130–Pro131 bond is resistant to trypsin cleavage, these observations suggest that peptide T-9 must extend beyond K130 to K134 (see Fig. 1). This would account for the region 131–134 earlier believed to be a separate peptide (T-10). A complete lack of electron density for this lysine in both the WGA1 and WGA2 crystal structures is not uncommon and indicates disorder due to flexibility of the charged sidechain.

Peptide Sequence of Region 116–123 (Th-2b)

Peptide Th-2b was isolated from a thermolysin digest of the WGA1 T-9 peptide (Wright and Olafsdottir 1986) (see Fig. 1). Although its amino acid composition agreed with the earlier sequence data of WGA2 (Wright et al. 1984), we resequenced this peptide because of the discrepancy with WGA3 at residues 119 and 123 (see Fig. 2). High yield recovery of PTH-amino acids at all eight steps confirmed the earlier derived sequence of Th-2b and the identity of Gly119 and Ser123 in both WGA1 and WGA2.

Tryptophan Identification

Observations made recently during refinement of the crystal structure of WGA1 (unpublished results) suggested that residue 150, previously identified as Gly, may possess a disordered sidechain. This prompted reinvestigation of the N-terminal region of tryptic peptide T-13 (150–171) (see Fig. 1). The results from sequencing seven steps were consistent with the presence of a mixture of three similar peptides, T-13, T-13a, and T-13b, differing only by

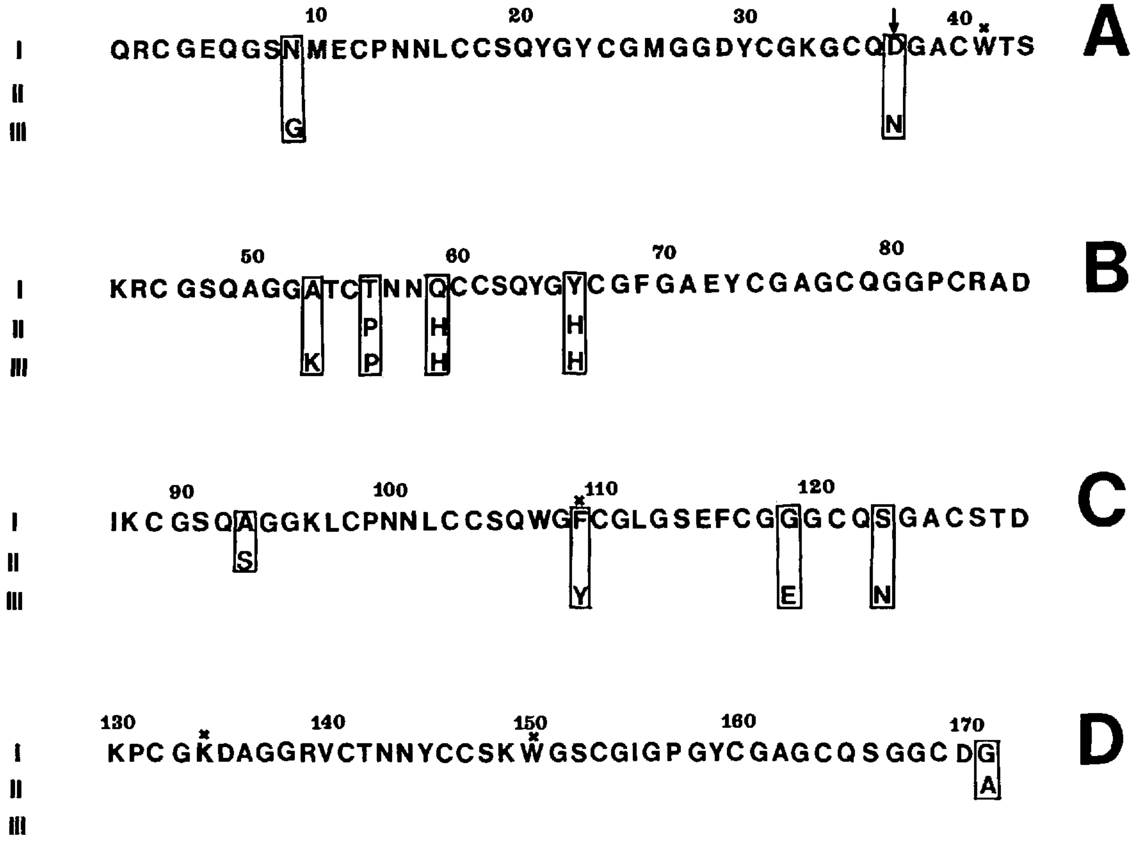
10 20 30 40
I QRCGEQGSNMECPNNLCCSQYGYCGMGGDYCGKGCQDGACWTS
II
III G N
A
50 60 70 80
I KRCGSQAGGATCTNNQCCSQYGYCGFGAEYCGAGCQGGPCRAD
II P H H
III K P H H
B
90 100 110 120
I IKCGSQAGGKLCPNNLCCSQWGFCGLGSEFCGGGCQSGACSTD
II S
III Y E N
C
130 140 150 160 170
I KPCGKDAGGRVCTNNYCCSKWGSCGIGPGYCGAGCQSGGCDG
II A
III
D

Fig. 2. One-letter code representation of the WGA amino acid sequences. I, II, and III refer to the three isolectins WGA1, WGA2, and WGA3 and A, B, C, and D to the four structural domains. The WGA1 sequence is presented in its entirety. Only amino acids at positions that differ from the WGA1 sequence are shown for WGA2 and WGA3. Sequence positions at which reassignments were made are labeled with a cross (×). The arrow (↓) at residue 37 indicates that no true genetic difference exists between the three isolectins (see text).

removal of Trp or TrpGly from their N-termini (see Fig. 1). A low amount of peptide T-2 (Cys3–Lys33) was also found to be present. Identification of residue 150 as Trp can explain the ambiguous results obtained in our earlier study, in which Gly and Ala were observed for the N-terminal residue by PTH back hydrolysis. Although Ala was interpreted as the back hydrolysis product for PTH-Cys of T-2, Gly was assigned to 150, in apparent agreement with the lack of sidechain density observed in the electron density map of WGA2. Both Gly and Ala are, however, also back hydrolysis products for PTH-Trp. With identification of Trp150, the sequence of the region 150–154 (TrpGlySerCysGly) was recognized to be identical to thermolytic peptide Th-4a, earlier incorrectly assigned to residues 107–111 as discussed above.

Determination of the C-Terminus

The identity of the C-terminus in WGA1 and WGA2 has remained uncertain due to inconclusive data from sequencing and C-terminal analysis. Moreover, this terminal residue is disordered in the crystal structure and, therefore, uninterpretable due to insufficient density. Our recent results using the anhydrous hydrazine method suggest a Gly at the C-terminus of WGA1 and WGA3, consistent with the nucleotide sequence for WGA3 and amino acid composition data for peptide Th-1g of WGA1 (see Table 1). The determination for WGA2 suggests both Ala and Gly (ratio 3:2). Although Ala is consistent with earlier sequencing results and amino acid composition data of peptide Th-1g (see Table 1) (Wright et al. 1984), the presence of Gly may be a contaminant from WGA3. A clean chromatographic separation of WGA2 and WGA3 using a salt gradient (Wright 1981) is not always possible.

Discussion

Sequence Corrections

As a result of further experimental evidence three amino acid reassignments were made to the WGA1 and WGA2 sequences at residues 109, 134, and 150 (see Figs. 1 and 2). A third Trp was located at position 150 as a result of crystal structure refinement

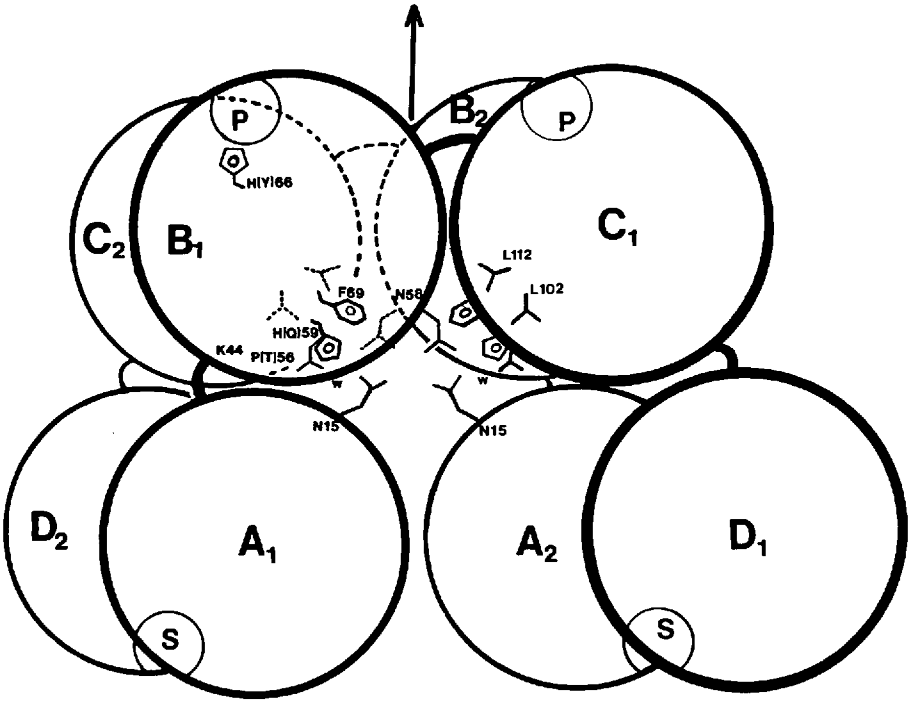

Fig. 3. Schematic representation of the WGA dimer illustrating the relative location of residues that are affected by the B-domain mutations His(Gln)59, Pro(Thr)56, and His(Tyr)66 in the A_1/B_1-, A_1/B_2-, and B_1/C_2-domain contact regions. Domains A, B, C, and D are represented by large circles and their connectivity is suggested by hinges. Subscripts 1 and 2 refer to subunits 1 and 2. The dimer axis is shown as an arrow. Sugar binding sites are labeled P at the B_1/C_2 and B_2/C_1 interface and S at the A_1/D_2 and A_2/D_1 interface.

(Wright 1987) and new amino acid sequence data on tryptic peptide T-13 of WGA1, confirming the presence of three Trp in the molecule as predicted from solution studies (Allen et al. 1973; Rice and Etzler 1975; Monsigny et al. 1978). In agreement with the nucleotide sequence of WGA3 the presence of a Lys at 134 in both WGA1 and WGA2 was confirmed. This allowed the C-terminus of T-9, earlier presumed to be K130, to be reassigned to K134. Identification of F109 was based on new sequence data for the heterogeneous peptide T-9 and constitutes a true difference in the sequences of isolectins 1 and 2 from that of isolectin 3 with Tyr109.

In retrospect, the sequence errors were caused by misleading corroboration between uncertain sequence data and structural information based on the electron density map. Although it is generally possible to interpret electron density features at well-ordered segments of the molecule with a high degree of confidence, interpretations are less reliable at surface regions where exposed sidechains are not restrained by interactions and possess poorly defined electron density shapes.

Limitations in the sequencing techniques employed were encountered for identification of the three Trp in the molecule, Trp41, Trp107, and Trp150. Specificity of thermolysin for peptide bonds on both sides of Trp resulted in reduced yields of each of the fragments. Moreover, poor recovery of PTH-Trp was observed in peptides with Trp at the C-terminus (26–41, 102–108, 145–150), presumably due to destruction during sequencing (Wright et al. 1984). Trp107 was the only Trp recognized in the original electron density map. Its presence was confirmed by fluorescence studies on peptide Th-2h (Leu102–Gly108). Th-4a (TrpGlySerCysGly), the only Trp-containing peptide isolated in pure form in the original study, was therefore believed to belong to region 107–111 in domain C (Wright et al. 1984). Trp41 was misidentified in the original sequence study due to peptide heterogeneity and ambiguities in the sequence data from a peptide mixture of three similar peptides (Wright et al. 1984). However, its identity was established clearly from crystal structure refinement (Wright 1987) and fluorescence studies of tryptic peptide T-3 (Gly34–Ser43) (Wright and Olafsdottir 1986). In contrast, a complete lack of sidechain density in the electron density map at Trp150, even after structure refinement of WGA2, raised no suspicion that a Trp was present, and resulted in misinterpretation of ambiguous sequencing results of tryptic peptide T-13 and omission of techniques that could have identified PTH-Trp directly.

Cross-contamination and heterogeneity due to poor separation on HPLC obscured sequence data for the large tryptic peptides (T-2, T-5, T-9, T-13) (Wright et al. 1984). More precise analysis using a gas phase sequenator has now made it possible to establish the extent of this heterogeneity and resolved formerly observed ambiguities in the amino acid compositions and terminal sequences of peptides T-9 and T-13. Both peptide preparations consisted of mixtures of similar peptides (T-9, T-9a and T-13, T-13a, T-13b), differing only with respect to their N-termini (see Fig. 1).

Variable Sequence Positions

In this study sequence variability among the three distinct WGA isolectins has been confirmed at a total of 10 residues (see Fig. 2). A difference was also observed at residue 37, where the nucleotide sequence codes for Asn (Raikhel and Wilkins 1987) and amino acid sequence data suggests Asp. Although we believe that all three isolectin genes code for Asn, the discrepancy with the amino acid sequence may be a result of deamidation either in the intact protein (Kossiakoff 1988) or in the proteolytically generated peptides used for sequencing. Deamidation is frequently observed in proteins when Asn is succeeded in sequence by Gly (Bornstein and Balian 1977), as is the case here, particularly under acid conditions. Such conditions were routinely used in WGA isolation procedures and in the purification of peptides.

Amino acid composition data for all three isolectins (Rice and Etzler 1975) had suggested earlier that WGA1 is unique (lacking histidines), and that WGA2 and WGA3 are more closely related. However, our inspection of the three amino acid sequences now clearly indicates that the unique isolectin is WGA3. Although WGA1 and WGA2 differ in only five positions, including residue 171 not identified in our earlier comparison (Wright and Olafsdottir 1986), larger differences are observed when comparing the WGA3 sequence with those of WGA1 and WGA2. WGA3 differs from WGA1 at eight positions, 9, 53, 56, 59, 66, 109, 119, and 123 (see Fig. 2). Five of these residues are common differences with WGA2. WGA2 differs from WGA3 in seven sequence positions, 9, 53, 93, 109, 119, 123, and 171. Whereas three of the five differences between WGA1 and WGA2 are concentrated in domain B (residues 56, 59, and 66), and the largest number of differences between WGA2 and WGA3 are located in domain C, the differences between WGA1 and WGA3 are approximately equally distributed between these domains with the exception of residue 9 (see Fig. 2). Considering the importance of these two interior domains for integrity of the molecular structure (see Fig. 3 and Wright 1987) and the fact that the sequences of domains B and C are more highly conserved than those of domains A and D (Wright et al. 1985), this is an unexpected result. However, it is interesting to note that except for residue 53 (Ala versus Lys), all the substitutions could result from only one nucleotide base change and some of these are conserved substitutions (His66 versus Tyr66; Tyr109 versus Phe109; Ala93 versus Ser93).

Whereas new amino acid sequence data has confirmed residues 109, 119, 123, and 171 in WGA1 and WGA2, the identity of residues 9 and 53 was clearly established in the original sequence study (Wright et al. 1984). Although a Gly, as observed in WGA3, is preferable at position 9 on grounds of sequence identity (see Fig. 2), well-defined sidechain density in the refined crystal structures of WGA1 and WGA2 is in accord with amino acid sequence data suggesting an Asn. The presence of a Lys at residue 53 in WGA3 in contrast to an Ala correlates more strongly with Lys96 and Arg139 in domains C and D. Although evidence from the electron density maps for the sidechain identity of this residue was not definitive due to its location in a flexible hair-pin turn, tryptic peptide mixtures of both isolectins did not contain fragments suggesting residue 53 as a cleavage site.

Effect on Structure Stability and Function

We have examined the potential effect of amino acid replacement on the stability of the domain, monomer, and dimer structures. High resolution structure refinement of WGA2 (Wright 1987) and WGA1 (unpublished results) allowed detailed analysis of the interactions that stabilize these structural units. The domain fold is stabilized by four S–S bridges and numerous hydrogen bonds, including 16 between backbone CO and NH groups and 7–8 from sidechains (Arg, Lys, Gln, Ser). Contact between adjacent domains along the polypeptide chain was found to be restricted to a narrow region of identical sequence -Pro(Thr)AsnAsn-X- (domain positions 11–16, see Table 10 and Fig. 8 in Wright 1987) that assumes the conformation of a distorted reverse turn. Although the loose, quasihelical domain arrangement allows for a large surface area in the monomer (see Fig. 4), nonpolar sidechains are strategically positioned to provide an extensive area for association between monomers in the dimer interface. In the sandwich-like dimer structure (see Fig. 3) almost one entire face of each monomer, as viewed in Fig. 4, becomes buried, and individual domains of each monomer associate in opposite polarity. It is reasonable to assume that these stabilizing interactions remain intact in the three isolectin structures.

Although 8 out of the 10 sequence substitutions observed among the three isolectins are easily accommodated at the dimer surface, the two differences at 56 and 59 are involved in monomer and dimer stabilization (see Fig. 3). Structural differences observed between WGA1 and WGA2, as analyzed previously from difference Fourier maps (Wright and Olafsdottir 1986), indicated that replacement of Pro by Thr at position 56 and His by Gln at position 59, did not cause a major disruption in the structure. This is consistent with the fact that WGA1 subunits are able to form hybrid dimers with the other two isolectins. Rearrangement of five water molecules in the immediate vicinity of residues 56

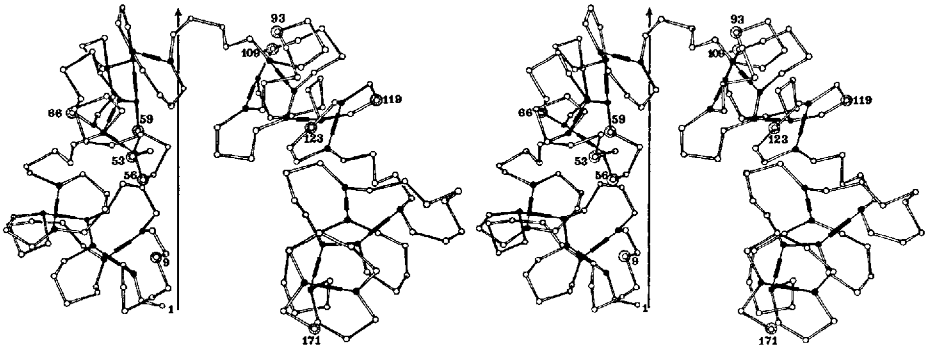

Fig. 4. Stereoscopic view of the α-carbon backbone of WGA. The amino acid residues that were found to vary among the three isolectins are numbered and emphasized by large circles. The approximate position of the twofold dimer axis is inferred by a long arrow.

and 59 (cluster C4, Wright 1987), and sidechain movements of residues Phe69, Lys44, Asn15, Leu102, and Leu112, however, alter the interactions that stabilize contact between domains A and B within the monomer and those in the dimer interface involving domains B and C. Based on its sequence, the structure of WGA3 should be identical with that of WGA2 in this region. Although the crystal structure of WGA3 has not been determined, crystals obtained for WGA3 are morphologically identical and isomorphous with those of WGA1 and WGA2.

A third substitution in domain B, Tyr in WGA1 for His in WGA2 and WGA3 at residue 66, has functional implications as discussed in earlier studies (Kronis and Carver 1982; Wright 1984; Wright and Olafsdottir 1986). Located at the surface in the B/C-domain contact region that forms the primary sugar binding site (Wright 1980b, 1984), it interacts with bound sugar ligands. The presence of a Tyr rather than a His sidechain at this position appears to be responsible for the higher affinity of specific saccharides observed for WGA1. NMR studies carried out on neuraminyl-lactose (NeuNAc) complexes of WGA1 and WGA2 indicated decreased K_d values by factors of 2–5, and a free energy (ΔG) gain of -0.4 to -0.9 kcal/mole for WGA1 complexes (Kronis and Carver 1982). Comparison of the high resolution crystal structures (2.2 Å) of the two isolectin–NeuNAc complexes may now explain those results in structural terms. The presence of a tyrosyl sidechain at residue 66 as opposed to an imidazole group is believed to provide more extensive van derWaals contact with several atoms in the NeuNAc pyranoside ring bound in subsite 1, and an additional hydrogen-bond interaction between the tyrosyl OH-group and the C2 hydroxyl of Gal bound in subsite 2 (details to be published elsewhere).

Table 2. Identity of variable residues in WGA isolectins

Residue	Ancestor	WGA1	WGA2	WGA3
9	Gly (GGC)	Asn (AAC)	Asn (AAC)	Gly (GGC)
53	Lys (AAG)	Ala (GCG)	Ala (GCG)	Lsy (AAG)
56	Pro (CCC)	Thr (ACC)	Pro (CCC)	Pro (CCC)
59	Leu (CTC)	Gln (CAG)	His (CAC)	His (CAC)
66	Tyr (TAC)	Tyr (TAC)	His (CAA)	His (CAC)
93	Ala (GCC)	Ala (GGC)	Ser (TCC)	Ala (GCC)
109	Tyr (TAC)	Phe (TTC)	Phe (TTC)	Tyr (TAC)
119	Ala (GCG)	Gly (GGG)	Gly (GGG)	Glu (GAG)
123	Asn (AAC) Ser (AGC)	Ser (AGC)	Ser (AGC)	Asn (AAC)

As illustrated in Fig. 4, the seven amino acid residues that differ between isolectins two and three (residues 9, 53, 93, 109, 119, 123, 171) are located at the surface of the molecule. Although residues 9, 93, and 109 are involved in intradomain backbone H-bonding and several other residues are solvated by water, the nature of the sidechain substitutions has no effect on the structure. Residues 9, 53, and 93 are located in the highly flexible hook-shaped hairpin turn that constitutes the longest stretch between consecutive Cys residues at the N-terminal end of domains A, B, and C, respectively. Residues 109 and 119 reside in reverse turns stabilized by hydrogen bonds, whereas residue 123 is located in a highly constrained four-residue segment held together by a disulfide bridge (Cys121–Cys126) at the C-terminal end of domain C. The C-terminal residue 171, on the other hand, is completely exposed and its exact position in the x-ray structure is still uncertain due to insufficient electron density and high temperature factors.

Evolutionary Relationships

It was suggested that the four highly similar domains of the WGA molecule evolved in a two-step process by gene duplication and fusion events (Wright et al. 1985). A possible ancestral domain sequence was derived by the present-day ancestor method of Klotz and Blanken (1981), taking into account the sequence of the one-domain rubber tree protein hevein (Walujuno et al. 1976) extremely similar to WGA. Two important facts emerged from this study:

(1) Divergence of the four-domain sequences could be correlated with optimization of quaternary structure requirements. (2) Residues involved in sugar binding were found to be conserved.

The availability of the nucleotide sequence for one of the WGA isolectins may now allow prediction of the nucleotide sequences for the other two. In addition, a sequence for an ancestor gene can be predicted with a much higher degree of precision and earlier ambiguities that were caused by the degeneracy of the amino acid codons eliminated. Although a complete account of such a study will be reported elsewhere, it may suffice to state here that the earlier proposed ancestor sequence of the core region (residues 3–31) has with one exception been confirmed. Residue 23, predicted to be a Ser, has in view of the new sequence data for WGA1 and WGA3 (Tyr at 66 and 109), been reassigned to Tyr. The advantage of an aromatic sidechain at this position in domain B for functional reasons is discussed above.

Judging from the high degree of codon preservation among identically positioned residues in the four domains of WGA3 (Raikhel and Wilkins 1987), it is likely that codon preservation also exists in the regions of perfect homology among the three isolectins. Differences between the three nucleotide sequences would then be restricted to the 10 observed variable sequence positions. In Table 2, codon assignments, requiring the minimum number of base changes from those observed for WGA3, have been made to 9 of the 10 positions in WGA1, WGA2, and to the ancestor. Surveying this table, three interesting observations can be made: (1) Seven of the nine codon replacements require only one base change. (2) The mutation rate is highest in domains B and C; an unexpected result considering that the sequences of these interior domains are more conserved than those of the terminal domains A and D, and that they play an important role in quaternary structure stabilization. (3) WGA3 agrees most closely with the ancestor, particularly in the domain core region where five of the seven variable positions are identical. In contrast, only one of these seven positions is in agreement with the predicted ancestor sequence in WGA1 and WGA2.

As WGA3 is the least abundant of the three isolectins in typical isolation experiments, we suggest that it represents a more ancient and possibly less active form of the lectin that may have diverged at an earlier point in evolutionary time from a common four-domain ancestral molecular intermediate.

Acknowledgments. This research was supported by the National Institutes of Health Grant AI-17992 (C.S.W.), the US National Science Foundation Grant DMB 87-03513 (N.V.R.), and the US Department of Energy Grant under contract DOE-AC02-76ERO-1338 (N.V.R.).

References

Allen AK, Neuberger A, Sharon N (1973) The purification, composition and specificity of wheat germ agglutinin. Biochem J 131:155–162

Bornstein P, Balian G (1977) Cleavage at Asn–Gly bonds with hydroxylamine. Methods Enzymol 47:132–145

Etzler ME (1985) Plant lectins: molecular and biological aspects. Annu Rev Plant Physiol 36:209–234

Ewart JAD (1975) Partial characterization of a wheat germ agglutinin. J Sci Food Agric 26:5–22

Goldstein IJ, Hayes CE (1978) The lectins: carbohydrate-binding proteins of plants and animals. Adv Carbohydr Chem Biochem 35:127–340

Goldstein IJ, Poretz RD (1986) Isolation, physical–chemical characterization and carbohydrate binding specificity of lectins. In: Liener IE, Sharon N, Goldstein IJ (eds) Lectins: properties, function and application in biology and medicine. Academic Press, New York, pp 43–247

Gray WR (1972) Dansyl chloride procedure. Methods Enzymol 25:121–138

Hendrickson WA, Konnert JH (1980) Stereochemically restrained crystallographic least-squares refinement of macromolecular structures. In: Srinivasan R (ed), Biomolecular structure, conformation, function and evolution, vol 1. Pergamon Press, Oxford, pp 43–57

Klotz LC, Blanken RL (1981) A practical method for calculating evolutionary trees from sequence data. J Theor Biol 91:261–272

Kossiakoff AA (1988) Tertiary structure is a principal determinant to protein deamidation. Science 240:191–194

Kronis AK, Carver JP (1982) Specificity of isolectins of wheat germ agglutinin for sialyloligosaccharides: a 360-MHz proton nuclear magnetic resonance binding study. Biochemstry 21: 3050–3057

Lis H, Sharon N (1986) Lectins as molecules and as tools. Annu Rev Biochem 55:35–67

Mansfield MA, Peumans WJ, Raikhel NV (1988) Wheat-germ agglutinin is synthesized as a glycosylated precursor. Planta 173:482–489

Mendez F, Lai CY (1975) Regeneration of amino acids from thiazolinones formed in the Edman degradation. Anal Biochem 68:47–53

Monsigny M, Delmotte F, Heléne C (1978) Ligands containing heavy atoms: perturbation of phosphorescence of a tryptophan residue in the binding site of wheat germ agglutinin. Proc Natl Acad Sci USA 75:1324–1328

Peumans WJ, Stinissen HM, Carlier AR (1982) A genetic basis for the origin of six different isolectins in hexaploid wheat. Planta 154:562–567

Raikhel NV, Pratt L (1987) Wheat germ agglutinin accumulation in coleoptiles of different genotypes of wheat. Localization of monoclonal antibodies. Plant Cell Reports 6:146–149

Raikhel NV, Wilkins TA (1987) Isolation and characterization of a cDNA clone encoding wheat germ agglutinin. Proc Natl Acad Sci USA 84:6745–6749

Rice RH (1976) Wheat germ agglutinin: evidence for a genetic basis of multiple forms. Biochim Biophys Acta 444:175–180

Rice RH, Etzler ME (1975) Chemical modification and hybridization of wheat germ agglutinins. Biochemistry 14:4093–4099

Schroeder WA (1972) Hydrazinolysis. Methods Enzymol 25: 138–143

Stinissen HM, Peumans WJ, Law CN, Payne PI (1983) Control of lectins in *Triticum aestivum* and *Aegilops umbellulata* by homoeologous group 1 chromosomes. Theor Appl Genet 67: 53–58

Tarr GE (1977) Improved manual sequencing methods. Methods Enzymol 47:335–357

Walujuno K, Scholma RA, Mariono A, Hahn AM (1976) Amino acid sequence of hevein. In: Proceedings of the International Rubber Conference. Kuala Lumpur, pp 518–531

Wright CS (1980a) Multi-domain structure of the dimeric lectin wheat germ agglutinin. In: Srinivasan R (ed) Biomolecular structure, conformation, function and evolution, vol 1. Pergamon Press, Oxford, pp 9–17

Wright CS (1980b) Crystallographic elucidation of the saccharide binding mode in wheat germ agglutinin and its biological significance. J Mol Biol 141:267–291

Wright CS (1981) Histidine determination in wheat germ agglutinin isolectin by x-ray diffraction analysis. J Mol Biol 145: 453–461

Wright CS (1984) Structural comparison of the two distinct sugar binding sites in wheat germ agglutinin isolectin 2. J Mol Biol 178:91–104

Wright CS (1987) Refinement of the crystal structure of wheat germ agglutinin isolectin 2 at 1.8 Å resolution. J Mol Biol 194:501–529

Wright CS, Olafsdottir S (1986) Structural differences in the two major wheat germ agglutinin isolectins. J Biol Chem 261: 7191–7195

Wright CS, Gavilanes F, Peterson DL (1984) Primary structure of wheat germ agglutinin isolectin 2. Peptide order deduced from x-ray structure. Biochemistry 23:280–287

Wright HT, Brooks DM, Wright CS (1985) Evolution of the multi-domain protein wheat germ agglutinin. J Mol Evol 21: 133–138

Received April 28, 1988/Revised July 30, 1988