Primary Structure of Wheat Germ Agglutinin Isolectin 2. Peptide Order Deduced from X-ray Structure[†]

Christine Schubert Wright,* Francisco Gavilanes,[‡] and Darrell L. Peterson

ABSTRACT: The complete amino acid sequence of wheat germ agglutinin isolectin 2 has been determined by the method of sequential Edman degradation and with the aid of the three-dimensional structure known from X-ray crystallography. Peptides ranging from 2 to 18 residues in length were obtained by thermolysin digestion of the S-carboxymethylated protein and purified by gel filtration and high-performance liquid chromatography. The peptide order was established primarily by matching (carboxymethyl)cysteines with the clearly defined half-cystine positions in the X-ray structure, thereby satisfying the disulfide repeat pattern observed in all four isostructural domains (A, B, C, and D) of wheat germ agglutinin, and by examination of amino acid compositions and terminal sequences of ten tryptic peptides. The unique assignment of peptides to these domains was consistent with all invariant half-cystines and glycines, as well as the single tryptophan, the two closely spaced histidines, and a number of other residues clearly identified in the X-ray structure analysis. Discrepancies between the chemical and X-ray sequences lie exclusively in poorly defined regions of the electron density map, at the N- and C-termini, and at the first intercystine loop of each domain. The latter loop was found to be eight instead of six residues in length, thus extending the size of domains A, B, and C from 41 to 43 residues and that of domain D to 42 residues. Regions of extensive interdomain homology, in addition to that of the half-cystines, are clustered at the central portion of each domain fold and are likely to be important for the integrity of the three-dimensional structure of the dimer molecule.

Wheat germ agglutinin (WGA),[1] a dimer of two identical subunits, is one of the most extensively studied plant lectins and has been shown to possess numerous in vitro biological activities [review by Goldstein & Hayes (1978) and Lis & Sharon (1977)]. Although the crystal structure of WGA had been determined at high resolution some time ago (Wright, 1977a), a chemical amino acid sequence had not become available.

Early physicochemical studies on WGA revealed the unusual character of this lectin as compared to other plant lectins. Amino acid composition data indicated an unusually large half-cystinyl (18%) and glycyl (20–23%) content as well as large amounts of glutamyl, asparaginyl, and seryl residues. The total length of the polypeptide chain was estimated to comprise 165 ± 12 amino acid residues (Allen et al., 1973; Nagata & Burger, 1974; Rice & Etzler, 1975; Wright, 1977a). Recent reports indicate that lectins from other cereal grains possess very similar molecular properties to those of WGA (Tsuda, 1979; Peumans et al., 1982). Thus, it is quite likely that the structure of WGA will recur in many other cereal lectins. The strong internal homology observed in WGA, which is dominated by a disulfide structural fold (Drenth et al., 1980), distinguishes this molecule from most other protein structures. The protomer consists of an assembly of four isostructural domains, each stabilized by four interlocking and homologously placed disulfide bonds. The two unique carbohydrate binding sites, which had been predicted earlier from solution studies (Nagata & Burger, 1974; Privat et al., 1974a), were found to be located in the dimer interface between nonidentical domains (Wright, 1980b). In the crystal, both of these sites are specific for the binding of *N*-acetyl-D-glucosamine, while only one of them binds *N*-acetyl-D-α-neuraminic acid.

Two to four genetically distinct species, called isolectins, which are presumably due to the polyploid nature of wheat, were characterized by Allen et al. (1973) and Rice & Etzler (1975). An amino acid sequence for WGA isolectin 2 based on the electron density map and model building has been proposed, but in numerous poorly defined regions of the map, only tentative assignments could be made (Wright, 1981). Because of the absence of histidines in isolectin 1, the positions for the two histidines in isolectin 2 could be determined by application of the difference Fourier technique, on the basis of high-resolution data of crystals of both isolectins (Wright, 1981). Chemical investigation of the amino acid sequence of an isolectin mixture and later of isolectin 3 was begun by Dr. Y. Nagata (Tokyo University) at the time that the crystal structure was being determined. However, his proposed preliminary sequence (unpublished data) was found to be inconsistent with the X-ray structure.

For these reasons, as well as the necessity to identify side chains in the sugar binding sites, the primary structure of WGA isolectin 2 (WGA2) was reinvestigated. The amino acid sequence proposed here has been derived from a set of 30 cleavage fragments generated by thermolysin digestion and from ten tryptic peptides. The strategy followed to establish peptide order rests in large part on the X-ray structure.

Experimental Procedures

Materials

WGA isolectin 2 was isolated from raw wheat germ (Sigma Chemical Co.) as described previously (Wright, 1981). Thermolysin, carboxypeptidase Y, iodoacetic acid, dithioerythritol (DTE), thiodiglycolic acid, and the standard dansyl-amino acid kit were also purchased from Sigma, and trypsin

[†] From the Department of Biochemistry, Medical College of Virginia, Richmond, Virginia 23298. *Received June 10, 1983.* This research was supported by National Institutes of Health Research Grant AI 17992.

[‡] Present address: Departamento de Bioquimica, Facultad de Ciencias Quimicas, Ciudad Universitaria, Madrid 3, Spain.

[1] Abbreviations: WGA2, wheat germ agglutinin isolectin 2; HPLC, high-performance liquid chromatography; PTH, phenylthiohydantoin; DTE, dithioerythritol; CM-WGA2, carboxymethylated wheat germ agglutinin isolectin 2; SDS, sodium dodecyl sulfate; Tris, tris(hydroxymethyl)aminomethane; TFA, trifluoroacetic acid.

0006-2960/84/0423-0280$01.50/0

was from Worthington Millipore Corp. Sequencing reagents were obtained from Pierce Chemical Corp. and sequenal- or HPLC-grade solvents either from Pierce or Burdick & Jackson Laboratories. Amino acid analyzer reagents and solvents came from Durrum.

Methods

Carboxymethylation. The native protein (10 mg) was carboxymethylated with iodoacetic acid as suggested by Erni et al. (1980) in the absence of denaturant. A 1:50 molar ratio of WGA2 (SH groups) to DTE was used, and the reaction was carried out at pH 8.5 (2.3 M Tris buffer) for 30 min with a 100-fold excess of iodoacetic acid over WGA SH groups. The modified protein was dialyzed against distilled water and frozen after being concentrated to a small volume. The presence of a single band on a 15% SDS–acrylamide gel, positioned above that of native WGA, indicated homogeneity of the carboxymethylated protein (CM-WGA2).

Digestion. Nine milligrams of CM-WGA2 was digested with thermolysin (enzyme to substrate concentration 1:50) at 37 °C for 4.5 h in a buffer consisting of 0.05 M ammonium bicarbonate and 0.0025 M $CaCl_2$ (pH 8.0). Trypsin digestion was carried out on 6 mg of CM-WGA2 in 0.08 M ammonium bicarbonate buffer (pH 8.0) for 4 h at 37 °C (1:50 enzyme to substrate ratio) and on the blocked thermolysin peptide Th-lb with a 1:100 enzyme to substrate ratio under the same conditions. C-Terminal sequencing on the intact protein (0.5 mg) was performed with carboxypeptidase Y at enzyme to substrate concentration ratios of 1:25 and 1:12.5. Norleucine was used as an internal standard. The reaction was allowed to procede for 45 min at 37 °C in 50 mM sodium acetate buffer (pH 5.5). Aliquots were removed at 10-min intervals and analyzed for amino acid composition.

Peptide Separation. The peptide mixture was applied to a 1.7 × 80 cm column of Sephadex G-25 (Pharmacia), fine, which had previously been equilibrated with a solution of 0.015 N NH_4OH (adjusted to pH 9.5 with HOAC) and 10% 1-propanol. The effluent was monitored at 220 and at 280 nm. Individual G-25 peak fractions (or two to four such fractions at the down slopes of the peaks) were further analyzed for the presence of peptide mixtures by high-performance liquid chromatography (HPLC) (Model 5000, Varian Instruments, Inc.). Evaporated samples were dissolved in 100 μL of HPLC buffer [0.1% trifluoroacetic acid (TFA)] and applied to a reversed-phase C_{18} peptide column (MCH-5 Micro-Pak, Varian Instruments, Inc.) A 0.1% TFA buffer was used and an acetonitrile gradient. Optimal conditions for setting up the solvent gradient for separation of each peptide mixture were determined by trial and error. Typically, the duration of one run was 90–100 min with acetonitrile gradients ranging from 0–35% to 0–70%. Fractions of 1–1.4 mL were collected at a flow rate of 1 mL/min, monitored at 215 nm. Peak fractions were evaporated in a Model RH20-12 Speed Vac concentrator (Savant Instruments, Inc.) and appropriate amounts were hydrolyzed in 6 N HCl, 0.5% in thiodiglycolic acid. Amino acid compositions of all major HPLC peaks were determined on a Durrum MBF amino acid analyzer equipped with a fluoropa detection system and a Varian CDS-111 integrator. Peaks possessing identical compositions, but derived from different G-25 fractions, were pooled for sequence analysis.

Sequence Determination. The sequential manual Edman degradation procedure of Tarr (1977) was employed for sequencing of all thermolysin peptides 2–18 residues in length, with 5–15-nmol quantities. In the cleavage step, the peptides were incubated for 90 s at 50 °C with 20 μL of concentrated HCl. Conversion of the anilinothiazolinones to the phenylthiohydantoin (PTH) derivatives of the amino acids was accomplished by incubation at 50 °C for 10 min in 1 N acetyl chloride/MeOH. PTH amino acids were identified by HPLC on a Beckman PTH-C_{18} reversed-phase column, correlating their elution times with those obtained from standard PTH-amino acids (from Pierce). The solvent gradient procedure suggested by Beckman was used: 5.75 mM acetate buffer (pH 5.0–5.5) with or without tetrahydrofuran and an acetonitrile gradient (0–58%).

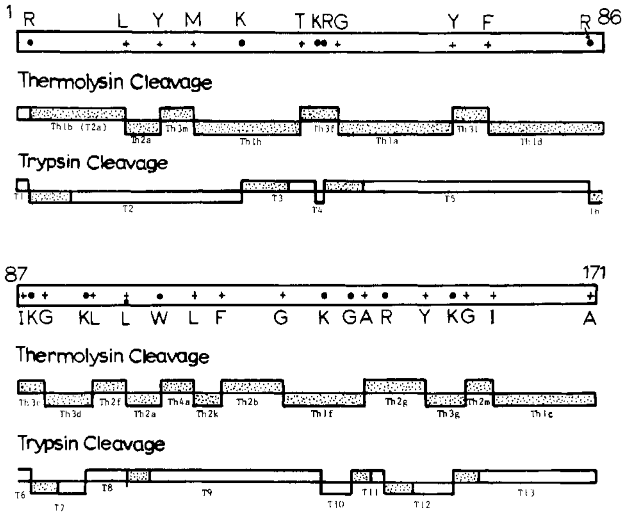


FIGURE 1: Schematic outline of position of all major fragments generated by thermolysin and trypsin digestion. Cleavage points, identified by the one-letter amino acid code, are marked (+) for thermolysin and (●) for trypsin on the straight bars, which represent the intact molecule. Dotted areas indicate the portions sequenced.

Tryptic peptides were sequenced by a combination of the Edman and dansyl chloride (Gray, 1972) procedures. Dansyl-amino acids were identified on 5 × 5 cm polyamide thin-layer sheets (Chen Chin Trading G. Ltd., Taiwan) visually by inspection with UV light and comparison to the spot positions of standard dansyl-amino acids applied to the reverse side of the sheet.

Tryptophan Detection. The presence of tryptophan in peptide Th-2h was determined on a 4800 series scanning fluorometer (SLM Instruments) equipped with a Hewlett-Packard 9815 and 9862A calculator and plotter. Fluorescence emission was scanned from 300–500 nm at an excitation wavelength of 294 nm.

C-Terminal Analysis. Samples of 2 nmol were heated under vacuum at 80 °C for 18 h in the presence of 50 μL of anhydrous hydrazine (Pierce), similar to the method of Schroeder (1972).

Results

(*a*) *Sequencing Strategy.* For the determination of the amino acid sequence of WGA, we have used classical sequencing techniques in combination with knowledge from the three-dimensional structure. The molecule is known to consist of a 4-fold structural domain repeat. Figure 1 summarizes schematically our chemical approach. The positions of 20 primary fragments, resulting from thermolysin digestion, and 10 tryptic overlaps are shown in proper alignment. Figure 2 provides the summary proof of the various peptide sequences constituting the complete primary structure.

Two independent procedures for establishing the peptide sequence were carried out as follows. First, fragmentation of CM-WGA into a collection of some 30 thermolysin peptides produced a complete set of primary fragments and also numerous overlaps resulting from limited hydrolysis. Sequencing

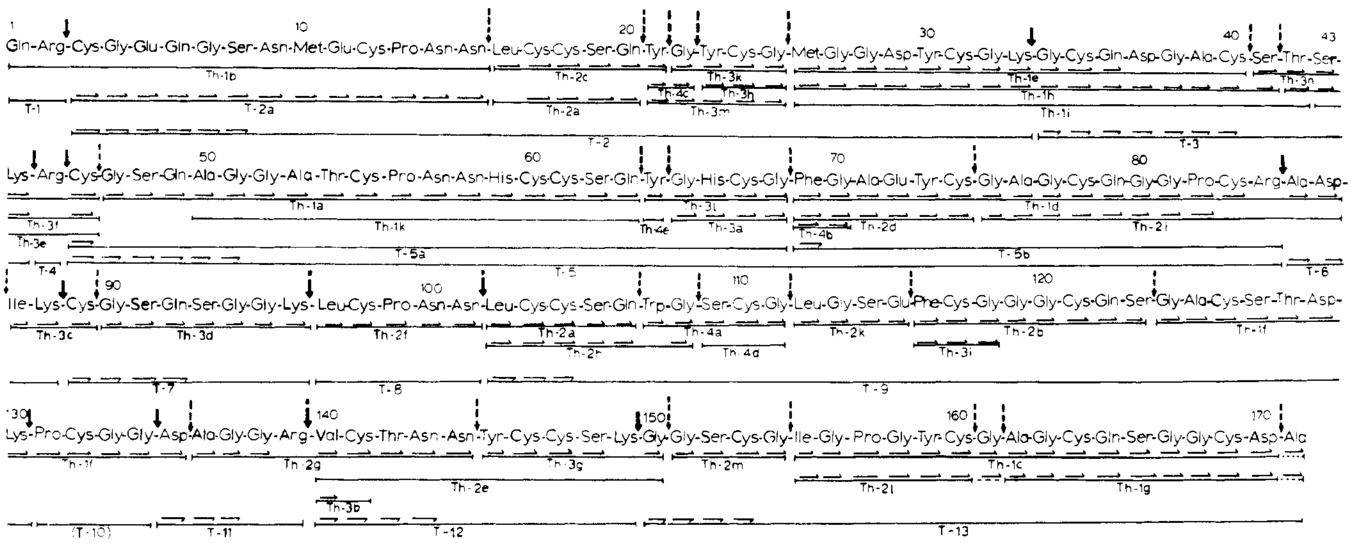

Gln-Arg-Cys-Gly-Glu-Gln-Gly-Ser-Asn-Met-Glu-Cys-Pro-Asn-Asn-Leu-Cys-Cys-Ser-Gln-Tyr-Gly-Tyr-Cys-Gly-Met-Gly-Gly-Asp-Tyr-Cys-Gly-Lys-Gly-Cys-Gln-Asp-Gly-Ala-Cys-Ser-Thr-Ser-
Lys-Arg-Cys-Gly-Ser-Gln-Ala-Gly-Gly-Ala-Thr-Cys-Pro-Asn-Asn-His-Cys-Cys-Ser-Gln-Tyr-Gly-His-Cys-Gly-Phe-Gly-Ala-Glu-Tyr-Cys-Gly-Ala-Gly-Cys-Gln-Gly-Gly-Pro-Cys-Arg-Ala-Asp-
Ile-Lys-Cys-Gly-Ser-Gln-Ser-Gly-Gly-Lys-Leu-Cys-Pro-Asn-Asn-Leu-Cys-Cys-Ser-Gln-Trp-Gly-Ser-Cys-Gly-Leu-Gly-Ser-Glu-Phe-Cys-Gly-Gly-Gly-Cys-Gln-Ser-Gly-Ala-Cys-Ser-Thr-Asp-
Lys-Pro-Cys-Gly-Gly-Asp-Ala-Gly-Gly-Arg-Val-Cys-Thr-Asn-Asn-Tyr-Cys-Cys-Ser-Lys-Gly-Gly-Ser-Cys-Gly-Ile-Gly-Pro-Gly-Tyr-Cys-Gly-Ala-Gly-Cys-Gln-Ser-Gly-Gly-Cys-Asp-Ala
1
10
20
30
40
43
50
60
70
80
90
100
110
120
130
140
150
160
170
Th-1b
Th-2c
Th-3k
Th-4c
Th-3h
Th-1e
Th-1h
Th-3n
T-1
T-2a
Th-2a
Th-3m
Th-1i
T-2
T-3
Th-1a
Th-3l
Th-1d
Th-3f
Th-1k
Th-4e
Th-3a
Th-4b
Th-2d
Th-2i
Th-3e
T-5a
T-5b
T-4
T-5
T-6
Th-3c
Th-3d
Th-2f
Th-2a
Th-4a
Th-2k
Th-2b
Th-1f
Th-2h
Th-4d
Th-3i
T-7
T-8
T-9
Th-1f
Th-2g
Th-3g
Th-2m
Th-1c
Th-2e
Th-2l
Th-1g
Th-3b
T-10
T-11
T-12
T-13

FIGURE 2: Alignment of thermolysin and tryptic peptides of WGA2. Layout of the sequence into four segments is designed to match the approximate boundaries of each structural domain (A, B, C, and D). Peptide fragments are represented by solid lines labeled Th for thermolysin and T for trypsin. Thermolysin peptides are numbered according to the peak in Figure 1 from which they derive. Tryptic peptides are numbered sequentially. (⇀) refers to PTH-amino acid identification by HPLC, and (→) indicates dansyl-amino acid identification on polyamide thin-layer sheets. Cleavage sites for thermolysin and trypsin are marked by dashed and solid arrows, respectively.

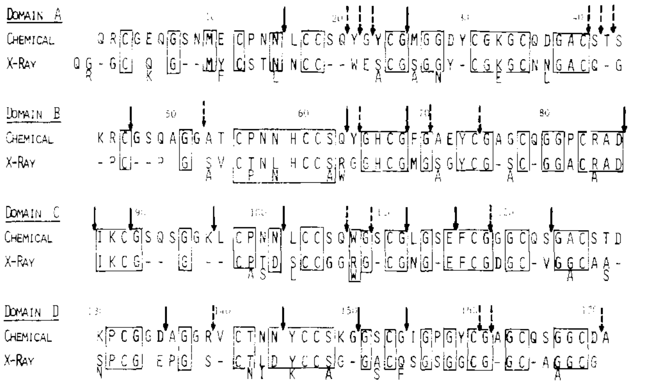


FIGURE 3: Comparison of chemically determined amino acid sequence for WGA2 with the tentative sequence derived from the X-ray structure. Sequences are represented by the one-letter code. Solid arrows indicate complete cleavage by thermolysin, while dashed arrows designate partial cleavage. Regions of perfect agreement of the two sequences are shown boxed in. Gaps in the X-ray sequence are an indication that these residues could not be located due to lack of electron density. Dashes designate that it was not possible to assign a definite side chain to the density.

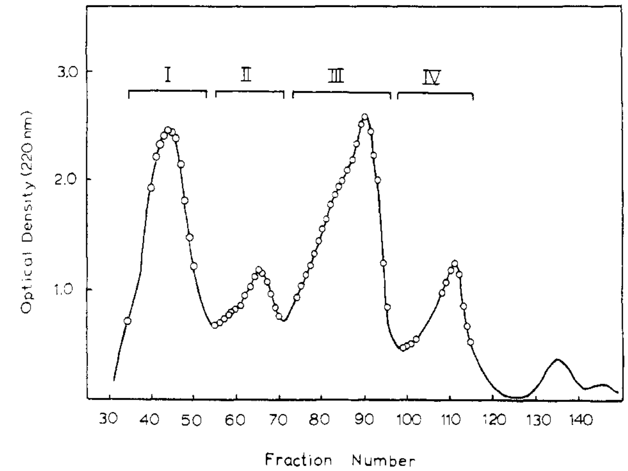


FIGURE 4: Elution profile of thermolysin digest of CM-WGA2 from a Sephadex G-25 column equilibrated with 0.015 N ammonium hydroxide, 10% in 1-propanol. The volume collected in each tube was 1.2 mL. Circles indicate the fractions further analyzed by HPLC. Peaks are numbered for the purpose of peptide identification.

by the manual Edman procedure of Tarr (1977) of these peptides was straightforward, although difficulties due to salt accumulation and incomplete cleavage at Gly–Gly bonds were encountered. PTH-Arg could not be identified by HPLC nor dansylation. Thus, blank steps were assigned to PTH-Arg, if the presence of Arg was known from the amino acid composition. The yield of PTH-CM-Cys varied from peptide to peptide and was often found to be very low. In our approach to place these peptides into proper order, we utilized the preliminary sequence derived from the X-ray structure, which clearly indicated the positions of all half-cystines and those of the two histidine residues (Wright, 1981; see Figure 3). For this process, three main criteria were followed: (i) matching of the Cys positions to locate the peptide within the domain (this was possible, because the four disulfide bonds in each domain are represented by very strong density in the electron density map and thus are the most reliable feature of the structure); (ii) occurrence of reliably interpreted marker residues, such as Trp, Met, His, Tyr, Phe, Leu, or Ile, to ascertain into which domain the peptide belongs; (iii) peptide length and location within the three-dimensional structure (surface or interior) to fill remaining gaps with peptides that did not possess marker residues and could not be fitted by the first two criteria.

Second, in order to verify by chemical means correct alignment of the thermolysin fragments, the CM protein was digested with trypsin. Although only 10 of the 12 peptides could be recovered from the digest, terminal sequences of these peptides sufficed to confirm the positions of all Lys and Arg residues. In addition, amino acid composition data for the 10 peptides (Table IV) confirmed their placement but also indicated an erroneous exchange of two similar six-residue peptides (Th-3m, Th-4a) between domains A and C.

(*b*) *Thermolysin Peptides.* Thermolysin digestion of CM-WGA (500 nmol) was performed at 37 °C, pH 8.0 (4.5 h). Although thermolysin is fully active at higher temperatures, the (carboxymethyl)cysteines appeared to become unstable when digestion was carried out at 55 °C. Initially, a gross separation of this digest into four peaks was achieved by gel filtration on Sephadex G-25 (see Figure 4). Sixty-three fractions distributed among all four peaks as marked in Figure

Table I: Amino Acid Compositions of HPLC-Purified Thermolysin Peptides from Peak I (Sephadex G-25 Gel Filtration)[a]

amino acid	Th-1a	Th-1b	Th-1c	Th-1d	Th-1e	Th-1f	Th-1g	Th-1i	T-1	T-2a	WGA2[b]
CM-Cys	2.2 (3)	1.5 (2)	2.5 (3)	2.4 (3)	2.2 (3)	1.6 (2)	2.0 (2)	2.8 (3)		1.7 (2)	32.0 (32)
Asp	2.2 (2)	3.5 (3)	1.5 (1)	1.1 (1)	2.4 (2)	1.7 (2)	1.1 (1)	2.4 (1)		2.8 (3)	15.3 (15)
Thr	1.2 (1)					0.8 (1)		0.5 (1)			3.7 (4)
Ser	1.7 (2)	0.9 (1)	1.0 (1)			1.2 (1)	0.9 (1)	0.8 (1)		1.1 (1)	9.7 (16)
Glu	2.2 (2)	4.3 (4)	1.4 (1)	2.1 (2)	1.2 (1)	0.5 (0)	1.1 (1)	1.5 (1)	1.1 (1)	3.4 (3)	15.7 (15)
Pro	(1)	(1)	(1)	(1)		(1)				(1)	5.5 (6)
Gly	2.4 (3)	1.7 (2)	6.2 (6)	5.3 (5)	3.8 (5)	2.2 (3)	3.8 (4)	3.2 (5)		2.0 (2)	40.4 (42)
Ala	1.5 (2)		1.3 (2)	2.9 (3)	0.9 (1)	1.2 (1)	1.8 (2)	1.5 (1)			9.2 (10)
Val											0.8 (1)
Met		0.9 (1)			0.9 (1)			0.4 (1)		1.0 (1)	1.2 (2)
Ile			0.9 (1)								2.3 (2)
Leu											4.1 (4)
Tyr			0.7 (1)	0.9 (1)	1.0 (1)			0.9 (1)			6.5 (7)
Phe				1.0 (1)							2.3 (2)
His	0.8 (1)										2.0 (2)
Lys					0.8 (1)	1.4 (1)		0.6 (1)			7.9 (6)
Arg		1.0 (1)		1.0 (1)					0.9 (1)		3.4 (4)
total	17	15	17	18	15	12	11	17	2	13	171
position	47–63	1–15	155–171	69–86	26–40	124–135	161–171	26–42	1–2	2–15	1–171

[a] All values presented are a result of 20–24-h hydrolysis and represent mol/mol of peptide. Values in parentheses were taken from the sequence in Figure 6. Thermolysin peptides are designated Th and tryptic peptides T. CM-Cys refers to (carboxymethyl)cysteine. Hydrolysis values for CM-Cys, Thr, Ser, Tyr, His, and Met are uncorrected for destruction during hydrolysis and thus tend to be low in some cases. [b] Values are quoted from Wright (1981) for crystals of WGA2.

Table II: Amino Acid Compositions of HPLC-Purified Thermolysin Peptides from Peak II (Sephadex G-25 Gel Filtration)[a]

amino acid	Th-2a	Th-2b	Th-2c	Th-2d	Th-2f	Th-2g	Th-2h	Th-2i	Th-2k	Th-2l	Th-2m
CM-Cys	1.8 (2)	2.1 (2)	1.9 (2)	1.4 (1)	0.5 (1)	0.9 (1)	1.7 (2)	1.6 (2)		1.2 (1)	0.9 (1)
Asp					2.2 (2)	2.3 (2)		1.9 (1)			
Thr						1.1 (1)					
Ser	0.9 (1)	1.0 (1)	0.8 (1)	0.5 (0)			1.3 (1)	0.6 (0)	0.9 (1)		1.0 (1)
Glu	0.9 (1)	1.2 (1)	1.3 (1)	1.4 (1)			1.2 (1)	1.3 (0)	1.2 (1)		
Pro					(1)			(1)		(1)	
Gly		2.7 (3)		1.4 (1)		2.3 (2)	1.3 (1)	4.0 (4)	0.9 (1)	1.9 (2)	2.1 (2)
Ala				1.1 (1)		1.1 (1)		1.9 (2)			
Val						1.0 (1)					
Met											
Ile										1.0 (1)	
Leu	1.0 (1)		1.1 (1)		1.1 (1)		1.0 (1)		1.2 (1)		
Tyr			1.0 (1)	0.9 (1)						0.8 (1)	
Phe		1.0 (1)		1.0 (1)							
His											
Lys											
Arg						1.2 (1)		0.8 (1)			
Trp							(1)[b]				
total	5	8	6	6	5	9	7	12	4	6	4
position	16–20, 102–106	116–123	16–21	69–74	97–101	136–144	102–108	76–86	112–115	155–160	151–154

[a] Conditions are as specified in Table I. [b] Determined spectrophotometrically.

4 were further analyzed by reversed-phase HPLC to yield pure, salt-free peptides for sequencing. Representative HPLC profiles, indicating the position of individual peptides and a more detailed discussion of this separation step, are given in the supplementary material (see paragraph at end of paper regarding supplementary material).

Peptides are designated 1, 2, 3, and 4 according to the peak in Figure 4 from which they were derived. Their amino acid compositions as well as that for the intact molecule are listed in Tables I–III. The compositions derive from G-25 fractions, whose HPLC analysis indicated clean separation of the particular peptide. The percent yield for each peptide could not be calculated, as not all G-25 fractions were submitted to HPLC analysis and peptides often overlapped several of the fractions. In addition, recovery of each peptide from the HPLC column was often found to be only 50%. A number of special difficulties were encountered with the following peptides.

N-Terminal Peptide (*Residues 1–15*). Peptide Th-lb was the only peptide found to be blocked. As Nagata & Burger (1974) had reported a blocked N-terminus, this peptide was assumed to derive from the N-terminus. The amino acid composition indicated the presence of one Met (Table I), in agreement with X-ray evidence for a Met at position 10 (Wright, 1981). Moreover, the presence of an Arg made it possible to obtain trypsin subfragments and thus derive the sequence, at least in part. Two well-separated cleavage fragments were isolated by HPLC (C_{18} column), and amino acid composition data (Table I) revealed that a dipeptide (T-1) and a 13-residue peptide (T-2a) had been generated. Dansylation of the dipeptide (Glx-Arg) produced a negative result, implying that it must represent the blocked N-terminal peptide. Blockage was assumed to be due to pyrrolidonecarboxylic acid formation, confirming the earlier prediction by Nagata (1978; Y. Nagata, unpublished data). In sequencing peptide T-2a, we encountered difficulties due to a tendency of this peptide

Table III: Amino Acid Compositions of HPLC-Purified Thermolysin Peptides from Peaks III and IV (Sephadex G-25 Gel Filtration)[a]

amino acid	Th-3a	Th-3c	Th-3d	Th-3f	Th-3g	Th-3i	Th-3k	Th-3l	Th-3m	Th-4a
CM-Cys	1.1 (1)	1.2 (1)		0.9 (1)	1.8 (2)	0.9 (1)	0.9 (1)	1.0 (1)	0.9 (1)	0.7 (1)
Asp										
Thr				1.0 (1)						
Ser			1.8 (2)	0.9 (1)	0.9 (1)					1.0 (1)
Glu			1.1 (1)							
Pro										
Gly	1.9 (2)		2.9 (3)		1.1 (1)	0.8 (1)	2.8 (2)	2.0 (2)	2.4 (2)	2.6 (2)
Ala										
Val										
Met										
Ile		1.0 (1)								
Leu										
Tyr					1.0 (1)		1.0 (1)	1.0 (1)	1.6 (2)	
Phe						1.0 (1)				
His	1.0 (1)							1.0 (1)		
Lys		0.9 (1)	1.0 (1)	1.0 (1)	1.2 (1)					
Arg				1.1 (1)						
Trp										(1)
total	4	3	7	5	6	3	4	5	5	5
position	65–68	87–89	90–96	42–46	145–150	116–118	22–25	64–68	21–25	107–111

[a] Conditions are as specified in Table I.

to emulsify heavily during the heptane/ethyl acetate extractions in the manual Edman procedure, resulting in low PTH-amino acid yields.

Methionine Peptide (*Residues 26–42*). Cleavage at Met-26 generated a mixture of three closely related peptides that were separable as individual peaks by HPLC and varied only in their Ser and Thr contents. Peptides Th-1e and -1i could be sequenced with confidence only up to step 11 (Gln-36) at which point the remaining portion became blocked. This blockage was thought to be a result of diketopiperazine or cyclic imide formation between the aspartate carboxyl group and the adjacent glycine –NH– group (Bornstein, 1969; Ondetti et al., 1968). In peptide Th-1h, which was most abundant but could only be isolated as an equimolar mixture associated with peptide Th-1a, this difficulty seemed to be absent. As the sequence of Th-1a had already been established, side by side sequencing of these two peptides was feasible. Beyond step 11, only tentative PTH assignments could be made, due to the similarity in these two peptide sequences. However, the C-terminal sequence (Asp-37-Ser-41) was later confirmed by tryptic peptide T-3. It is also worthwhile noting that a large portion of this region is in good agreement with the X-ray sequence (Figure 3).

Tryptophan Peptide (*Residues 102–112*). The single Trp residue present in the molecule was found to be located in two overlapping peptides. This was due to incomplete digestion at a Gln–Trp and a Gly–Ser bond. The PTH-Trp at the N-terminus of peptide Th-4a was observed in high yield by HPLC. However, in peptide Th-2h, no PTH-amino acid could be detected after four sequencing attempts at step 6, although the following step (7) yielded PTH-Gly in high amount. Thus, the presence of a Trp at step 6 in this peptide was suspected and confirmed by fluorescence emission spectroscopy (see Methods).

C-Terminal Sequence (*Residues 155–171*). A long and a short peptide (Th-1c and Th-1g, respectively) were assigned to the C-terminal region of the molecule. The short one, Th-1g, heterogeneous at both its N- and C-terminus and separable as four distinct peptides by HPLC, is generated by cleavage at either Cys-160–Gly-161 or Gly-161–Ala-162 as a subfragment of Th-1c (see Figure 2). Th-1c was also isolated as two components differing in Ala content. While sequencing of Th-1c did not yield the identity of the last three residues with certainty, due to limitations of the manual sequencing technique, the complete sequence was obtained from the largest component of Th-1g with Ala at position 162 and at the C-terminus. The C-terminus was further confirmed by carboxypeptidase Y digestion. Ala was found to be released in largest yield followed by Gly. Hydrazinolysis performed on 3 nmol of WGA2 also produced Ala predominantly. The presence of Gly could not be explained and was presumed to be an impurity.

(*c*) *Tryptic Peptides*. Preliminary fractionation of the tryptic digest by gel filtration (Sephadex G-25) was necessary, as direct analysis by HPLC did not resolve this peptide mixture. We purified ten major tryptic peptides by subsequent HPLC analysis of 13 of the G-25 fractions. Illustrations for these fractionations and further discussion of the purification and sequence determination of these tryptic peptides are given in the supplementary material. Their amino acid compositions are shown in Table IV. Six of these peptides terminate in Lys (T-2, -3, -6, -7, -9, and -12) and two in Arg (T-5 and T-11), while the two remaining ones (T-8 and T-13) contained neither Arg nor Lys. Amino acid analysis indicated that T-13 is the C-terminal peptide and T-8 is identical with thermolysin peptide Th-2f (Table IV). The latter peptide resulted from secondary specificity of trypsin, requiring a Pro three residues removed (at P_3) from the scissile bond (Wright, 1977b).

The N-terminal dipeptide Gln-Arg (T-1) and a tetrapeptide, residues 131–134 (T-10), could not be recovered from the hydrolysate. The latter peptide, T-10 (Pro-Cys-Gly-Gly), is also due to trypsin's secondary specificity. T-1 had already been characterized as the N-terminal sequence of thermolysin peptide Th-1b (Table I), and the sequence of T-10 was also firmly established in thermolysin peptide Th-1f, in good agreement with the X-ray sequence. Evidence for the presence of the single Arg was found by direct amino acid analysis of the tryptic digest.

These peptides were sequenced in 5–8-nmol quantitites a sufficient number of steps from the N-terminus by the Edman/dansyl method to establish satisfactory overlap with the thermolysin sequences (see Figure 2). As three of the large peptides (T-2, -5, and -13) could only be partially resolved by HPLC, we used peak fractions from the edges of the HPLC

Table IV: Amino Acid Compositions of HPLC-Purified Tryptic Peptides[a]

amino acid	T-2	T-3	T-5	T-5a	T-5b	T-6	T-7	T-8	T-9	T-11	T-12	T-13
CM-Cys	5.2 (0)	1.9 (2)	7.2 (8)	4.6 (5)	2.6 (3)		1.0 (1)	1.3 (1)	5.9 (6)		2.3 (3)	3.9 (4)
Asp	3.6 (4)	1.4 (1)	2.7 (2)	2.2 (2)	0.4 (0)	1.0 (1)		1.6 (2)	2.2 (1)	1.1 (1)	2.0 (2)	1.4 (1)
Thr		1.0 (1)	1.0 (1)	1.0 (1)					0.8 (1)		1.0 (1)	
Ser	2.1 (2)	1.3 (2)	2.4 (2)	2.0 (2)	0.2 (0)		1.8 (2)		2.7 (5)		1.5 (1)	1.5 (2)
Glu	3.9 (4)	1.3 (1)	3.9 (4)	2.0 (2)	2.0 (2)		1.1 (1)		2.3 (3)		0.7 (0)	1.5 (1)
Pro	(1)		(2)	(1)	(1)			(1)				(1)
Gly	8.0 (7)	2.8 (2)	9.5 (10)	5.2 (5)	4.8 (5)	0.5 (0)	4.6 (3)		5.8 (7)	2.2 (2)		8.5 (9)
Ala	0.6 (0)	1.1 (1)	3.5 (4)	2.0 (2)	1.8 (2)	1.1 (1)			1.0 (1)	1.0 (1)		2.0 (2)
Val											1.0 (1)	
Met	1.7 (2)											
Ile						1.0 (1)						0.8 (1)
Leu	1.0 (1)							1.0 (1)	1.8 (2)			
Tyr	2.6 (3)		1.8 (2)	0.4 (1)	0.2 (1)						0.9 (1)	1.1 (1)
Phe			1.0 (1)		0.9 (1)				1.3 (1)			
His			1.7 (2)	2.0 (2)								
Lys	1.2 (1)	1.2 (1)				1.0 (1)	1.0 (1)		1.4 (1)		1.2 (1)	
Arg			1.1 (1)		1.0 (1)					1.0 (1)		
Trp									(1)[b]			
total	31	11	39	23	16	4	8	5	29	5	10	22
position	3–33	34–44	46–84	46–68	69–84	85–88	89–96	97–101	102–130	135–139	140–149	150–171

[a] Conditions are as specified in Table I. [b] Determined spectrophotometrically.

peaks for amino acid analysis and sequencing.

Discussion

Sequence determination of the 171-residue polypeptide chain of WGA has been based on 32 fully sequenced thermolysin peptides, 10 partially sequenced tryptic peptides, and a tentatively proposed amino acid sequence derived from the X-ray structure.

(*A*) *Enzymatic Digestion.* In the native state, the heavily disulfide cross-linked WGA molecule is completely resistant to most proteases, although protease-sensitive amino acids are located at readily accessible surface regions of the molecule. However, thermolysin, a protease active at elevated temperatures, is able to digest native WGA completely at 55 °C. To eliminate potential complications caused by the S–S bonds, we chose to work with the S-carboxymethylated protein. Carboxymethylation of the reduced protein was accomplished in complete absence of denaturant, suggesting that the S–S bonds play an important role in maintaining the secondary and tertiary structure of the molecule.

Due to the broad specificity of thermolysin, numerous overlapping peptides could be generated, and in that respect, the choice of this protease proved to be most advantageous. Of the 35 susceptible peptide bonds, complete cleavage occurred only at leucine (4), isoleucine (2), phenylalanine (2), and at one of the two methionines (no. 26). The second Met (no. 10) was completely resistant to digestion. We observed incomplete hydrolysis at almost all of the tyrosine residues. Cleavage occurred at peptide bonds in which Tyr participated at both the N- and C-terminal sides. Incomplete cleavage to a higher degree also occurred at several Gly, Ala, Ser, and Thr and at the single Trp and Val, in which their amide groups contributed to the susceptible peptide bond. Complete trypsin cleavage occurred at all Lys (6) and Arg (4) residues and at two additional bonds as a result of secondary trypsin specificity.

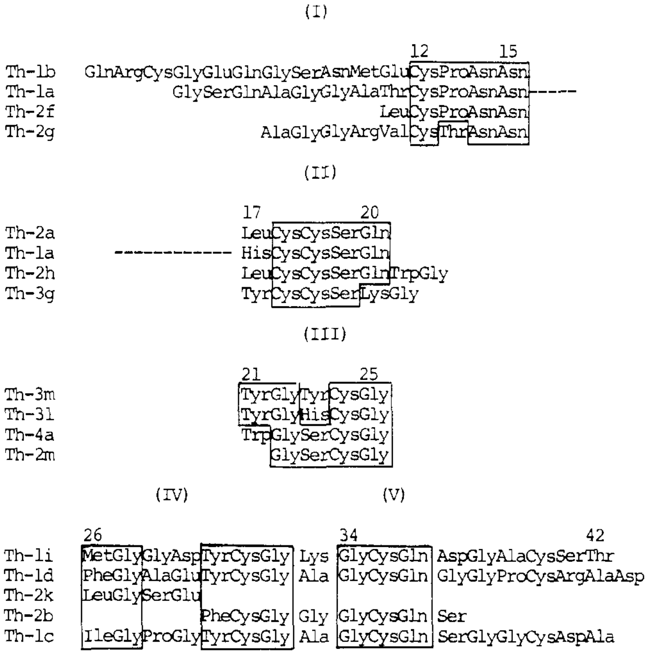
(*B*) *Peptide Order.* Reference to the preliminary X-ray sequence (see Figure 3) served as the primary tool in ordering the thermolysin peptides into a linear sequence. In addition, thermolysin subfragments generated by hydrolysis at less specific bonds provided in many cases a means to establish overlaps. Furthermore, ten tryptic peptides confirmed the peptide arrangement illustrated in Figure 2. Several peptides, Th-1a, Th-1b, Th-1c, Th-2b, Th-2g, Th-3c, and Th-3g, could immediately be placed into the X-ray sequence with a high degree of confidence. The sequences of Th-1a and Th-1b established the presence of an eight-residue instead of a six-residue intercystine loop at the N-terminal end of domains A and B, most likely also to be expected in domains C and D. Peptide Th-2g was positioned in front of Th-3g, as it overlaps with Th-2e. A dilemma in placing the large number of highly homologous short peptides into their proper domain was anticipated initially. However, except for peptides Th-4a and Th-3m, all these peptides (Th-2c, -2h, -2m, and -3l) could be fitted unequivocally to the X-ray sequence on the basis of distinguishable marker residues. The rest of the peptides could easily be inserted into the remaining gaps by virtue of their size as well as by matching their Cys positions. A detailed account of our rationale for the determination of peptide order in each domain is presented in the supplementary material. Suffice it to discuss here briefly some of the more ambiguous peptide assignments. (1) We were uncertain in placing the Trp-containing peptides (Th-2h and Th-4a) and two closely homologous peptides (Th-2c and Th-3k), in which Trp and Ser are replaced by Tyr. The X-ray data suggested two possibilities: positions 16–25 (domain A) and positions 102–111 (domain C). Initially, we had assigned Th-4a and Th-2h to the A domain, since the electron density at position 21 is more characteristic of a tryptophan than a Tyr. However, the amino acid composition of tryptic peptide T-2 (Table IV) indicated the presence of more than one tyrosine, while that of T-9 completely lacked tyrosine. A fluorescence emission scan performed on T-9 provided further evidence that the Trp peptide, Th-4a, must reside in domain C and, thus, Th-3 m in domain A in homologous location. (2) Placement of the second methionine residue in domain A at position 26 had not been clear-cut, as the X-ray data suggested an excellent fit at position 69 (see Figure 3). However, the three Met peptides, heterogeneous at their C-termini, exhibited an overall better fit to the X-ray sequence at the C-terminal region of domain A, complimenting the contiguous peptides Th-3e, -3f, and -3n heterogeneous at their N-termini. Reexamination of the electron density at residue 26 led to the conclusion that a methionine and not a serine (see Figure 3) is acceptable at this position, because the major $PtCl_4^{2-}$ binding site, a heavy atom compound used in the original structure determination, had been located here sandwiched between a disulfide and the side chain of residue 26. $PtCl_4^{2-}$ is known to bind specifically to methionines in protein structures (Blundell & Johnson, 1975; Wright et al., 1969). (3) It is noteworthy that while the two His-containing peptides (Th-1a and Th-31) clearly match the X-ray sequence best in domain B (residues 47–68), confirming the His positions earlier deduced from an electron density difference map for the two WGA isolectins (1 and 2) (Wright, 1981), alternative assignments might have been entertained without this structural knowledge. (4) In domain C, the least satisfactory overlap due to the large number of short peptides is observed. For instance, no overlap exists for peptide Th-2f, identical with tryptic peptide T-8. It was placed at positions 97–101 purely on homology grounds and by process of elimination. (5) Peptide Th-3g was found to be heterogeneous as to its Gly content, the Gly content differing from zero to two residues per peptide. In addition, Lys (149) could not be identified as PTH-Lys by HPLC. Satisfactory overlap between Th-3g and Th-2m was provided by the terminal sequence of tryptic peptide T-13, which is consistent with cleavage at Lys-149. The amino acid composition of T-13 containing Ile and Tyr also confirms placement of Th-1c at the C-terminus of the molecule.

(*C*) *Agreement with X-ray Sequence*. In agreement with the X-ray sequence, the positions of all invariant Cys and Gly residues, as well as the single Trp, one Ile, Met, Phe and Tyr, the two histidines, and many other residues could be confirmed, producing a 50–60% agreement factor between the two sequences. However, two of the four leucines had been interpreted as isomorphic asparagine in the X-ray structure. All the lysines, many glutamines, most of the tyrosines, and two of the arginines, located at flexible surface regions, were not recognized in the electron density map due to uncharacteristic shape or complete lack of density beyond the α-carbon. A number of additional discrepancies were observed, clearly a result of poorly defined regions of electron density. These lie in three areas. (i) At the N-terminus, the first two residues of the X-ray sequence had to be truncated to fit the chemical sequence. These two residues are represented by obscure density, part of which, in retrospect, seems to belong to the Arg-2 side chain. Gln-1 (formerly thought to be residue number 3) associates with its pyrrolidone ring across the dimer axis with the corresponding Gln of the other protomer. (ii) Most of the D domain is found to be represented by poorly defined electron density as compared with the other three domains (Wright, 1977a). Thus, it was not too surprising to find a discrepancy at the C-terminus, where the density could accommodate only a Gly instead of the Asp-Ala found chemically. This domain consists of only 42 amino acid residues, as compared to 43 in the other domains. (iii) The external loops, spanning the distance between the first two half-cystines (Cys-3 and Cys-12) and poorly defined in all four domains, are two residues longer according to the chemical sequence. This change is easily accommodated, but the domain size is now extended from 41 to 43 residues.

In all remaining regions, a satisfactory fit of the sequence presented here to the 2.2-Å electron density map could be obtained by model building. Crystallographic refinement of this model is currently in progress.

The complete polypeptide chain comprises a total of 171 residues. The calculated molecular weight based on the sequence is 20 600. The amino acid composition of the protein derived from the sequence (Table I) compares reasonably well with earlier published values (Rice & Etzler, 1975; Nagata & Burger, 1974; Privat et al., 1974b; Allen et al., 1973). The Gly and Ser contents were, however, mostly underestimated, while three to four Trp (Privat et al., 1974b) and a higher Cys content had been predicted.

(*D*) *Homology*. Peptides created by thermolysin cleavage, with hydrophobic *N*-termini, display easily recognizable homology with one another. This was not unexpected, as a high degree of structural homology observed between the four WGA domains (Wright, 1977a, 1981) and some sequence homology beyond that of the half-cystines had already been observed. The degree of sequence homology found here, however, is surprising. Regions of identical sequence repeats are clustered in five groups in interior regions of each domain:


(I)
12
15
Th-1b GlnArgCysGlyGluGlnGlySerAsnMetGluCysProAsnAsn
Th-1a GlySerGlnAlaGlyGlyAlaThrCysProAsnAsn-----
Th-2f LeuCysProAsnAsn
Th-2g AlaGlyGlyArgValCysThrAsnAsn
(II)
17
20
Th-2a LeuCysCysSerGln
Th-1a ------------- HisCysCysSerGln
Th-2h LeuCysCysSerGlnTrpGly
Th-3g TyrCysCysSerLysGly
(III)
21
25
Th-3m TyrGlyTyrCysGly
Th-31 TyrGlyHisCysGly
Th-4a TrpGlySerCysGly
Th-2m GlySerCysGly
(IV)
(V)
26
34
42
Th-1i MetGlyGlyAspTyrCysGly Lys GlyCysGln AspGlyAlaCysSerThr
Th-1d PheGlyAlaGluTyrCysGly Ala GlyCysGln GlyGlyProCysArgAlaAsp
Th-2k LeuGlySerGlu
Th-2b PheCysGly Gly GlyCysGln Ser
Th-1c IleGlyProGlyTyrCysGly Ala GlyCysGln SerGlyGlyCysAspAla

These regions represent portions of the molecule important for the integrity of the three-dimensional structure. For instance, the group I homologies are located at hairpin loops that are contact points between domains A–C around the dimer axis and also around the pseudo 2-fold relating domains AB of protomer 1 to domains CD of protomer 2. Model building shows that a H-bond network between the Asn side chains exists in the interior of the dimer. The group II homologies represent a central stretch of chain within each domain. Careful examination of these regions in terms of secondary structural elements and hydrogen-bonding capability is currently in progress. Residues, which have been shown to interact with specific N-acetylated saccharides in crystal complexes (Wright, 1980b), are homologous when comparing the two unique binding locations. These fall into region II (Ser-62 and Ser-148) and region IV (Asp-29, Glu-115, Tyr-73, and Tyr-159) and are identified in heavy outline in Figure 5.

It had been recognized earlier that the 43-residue protein hevein (from rubber trees) has a Cys distribution very similar to that of the WGA domains. In the light of the amino acid sequence reported here, reexamination of the similarity between these two proteins shows that the homologies seen between the WGA domains in groups I–III are also preserved in hevein. A more detailed account of these sequence comparisons in terms of evolutionary significance will appear elsewhere (C. S. Wright, D. Brooks, and H. T. Wright, unpublished results).

The amino acid sequence presented here (see Figure 5) is the first example of a sequence completed for one of the Cys-rich, chitin-binding lectins from the cereal grain or grass family (Graminaea) (Mishkind et al., 1983). A great deal of homology among these lectins from the Triticeae subfamily

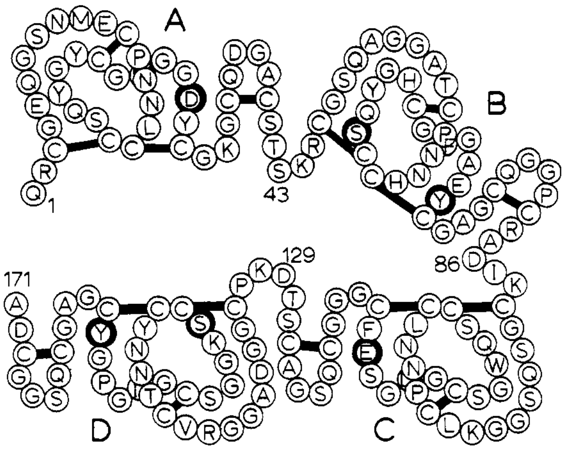


FIGURE 5: Diagrammatic representation of covalent structure of WGA isolectin 2. Amino acid residues are designated by the standard one-letter amino acid code. Disulfide bridges are shown by thick black bars. The residues believed to interact with specific sugar ligands are denoted by circles in heavy outline. Structural domains are labeled A, B, C, and D.

is expected, when one considers the similarity in their amino acid compositions (Peumans et al., 1982).

Acknowledgments

We thank Dr. Yoshiho Nagata for communicating his sequence data to C.S.W. during 1975–1977, when the interpretation of the WGA electron density map was in progress. We are also grateful to Drs. Jose Gonzalez-Ros and J. Michael Shaw for taking the fluorescence spectra.

Supplementary Material Available

One figure illustrating the gel-filtration (Sephadex G-25) profile of the tryptic digest and nine figures showing the HPLC analysis of individual fractions from the gel-filtration runs of the thermolysin and tryptic digests and further details discussing the peptide separation procedure by HPLC, sequencing techniques, and their limitations and rationale for determining peptide order (9 pages). Ordering information is given on any current masthead page.

Registry No. WGA2, 88179-78-6.

References

Allen, A. K., Neuberger, A., & Sharon, Y. (1973) *Biochem. J. 131*, 155–162.

Blundell, T. L., & Johnson, L. N. (1975) in *Protein Crystallography* (Horecker, B., Kaplan, N. O., Marmur, J., & Scheraga, A. A., Eds.) pp 138–239, Academic Press, New York, San Francisco, and London.

Bornstein, P. (1969) *Biochem. Biophys. Res. Commun. 36*, 957–964.

Drenth, J., Low, B. W., Richardson, J. S., & Wright, C. S. (1980) *J. Biol. Chem. 255*, 2652–2655.

Erni, B., DeBoeck, H., Loontiens, F., & Sharon, N. (1980) *FEBS Lett. 120*, 149–154.

Goldstein, I. J., & Hayes, C. E. (1978) *Adv. Carbohydr. Chem. Biochem. 35*, 127–340.

Gray, W. R. (1972) *Methods Enzymol. 25*, 121–138.

Lis, H., & Sharon, N. (1977) *Antigens 4*, 465–513.

Mishkind, M. L., Palevitz, B. A., & Raikhel, N. V. (1983) *Science (Washington, D.C.) 220*, 1290–1292.

Nagata, Y., & Burger, M. M. (1974) *J. Biol. Chem. 249*, 3116–3122.

Ondetti, M. A., Deer, A., Sheehan, J. T., Pluscec, J., & Kocy, O. (1968) *Biochemistry 7*, 4069–4075.

Peumans, W. J., Stinissen, H. M., & Carlier, A. R. (1982) *Biochem. J. 203*, 239–243.

Privat, J. P., Delmotte, F., & Monsigny, M. (1974a) *FEBS Lett. 46*, 229–232.

Privat, J. P., Delmotte, F., Mialonier, G., Bouchard, P., & Monsigny, M. (1974b) *Eur. J. Biochem. 47*, 5–14.

Rice, R. H., & Etzler, M. E. (1975) *Biochemistry 14*, 4093–4099.

Schroeder, W. A. (1972) *Methods Enzymol. 25*, 138–143.

Tarr, G. E. (1977) *Methods Enzymol. 47*, 335–357.

Tsuda, M. (1979) *J. Biochem. (Tokyo) 86*, 1451–1461.

Wright, C. S. (1977a) *J. Mol. Biol. 111*, 439–457.

Wright, H. T. (1977b) *Eur. J. Biochem. 73*, 567–578.

Wright, C. S. (1980a) *J. Mol. Biol. 139*, 53–60.

Wright, C. S. (1980b) *J. Mol. Biol. 141*, 267–291.

Wright, C. S. (1981) *J. Mol. Biol. 145*, 453–461; *152*, 181 (erratum).

Wright, C. S., Alden, R. A., & Kraut, J. (1969) *Nature (London) 221*, 235–242.