Proc. Nat. Acad. Sci. USA
Vol. 72, No. 8, pp. 2964–2968, August 1975
Biochemistry

Amino-acid sequence and oligosaccharide attachment sites of human erythrocyte glycophorin

(sialoglycoprotein/membrane)

MOTOWO TOMITA AND VINCENT T. MARCHESI

The Department of Pathology, Yale University School of Medicine, New Haven, Connecticut 06510

Communicated by George E. Palade, May 14, 1975

ABSTRACT **Glycophorin, the major sialoglycoprotein of the human erythrocyte membrane, is composed of 131 amino acids and an average of 16 oligosaccharide chains. Fifteen oligosaccharides are linked to threonine/serine residues via *O*-glycosidic bonds, and one more complex unit is attached to asparagine. The location of each of these oligosaccharides and the complete amino-acid sequence of this molecule have been determined by Edman degradation techniques. Glycophorin appears to be organized into three distinct "domains" on the basis of the locations of glycosylated amino acids and the clustering of residues of similar type. These include (*i*) a glycosylated segment composed of approximately 64 residues from the NH_2-terminus, (*ii*) a "hydrophobic" segment of approximately 32 nonpolar residues, and (*iii*) a COOH-terminal segment, composed of approximately 35 residues, which has an unusual concentration of hydrophilic amino acids.**

This unique structure is consistent with the earlier suggestions that glycophorin is one of the major "intrinsic" membrane proteins which has a transmembrane orientation.

Glycophorin is the major sialoglycoprotein of the human erythrocyte membrane (1, 2). It can be isolated from the membrane by a number of different techniques (3–6), and the purified product is a water-soluble complex which carries a number of different biological activities, including MN blood group- and influenza virus-binding activity. Glycophorin also contains covalently-bound oligosaccharides which bind kidney bean phytohemagglutinin and wheat germ agglutinin.

The properties of glycophorin are of great interest to investigators concerned with the molecular organization of cell membranes since this molecule is one of the principal "intrinsic" membrane proteins of erythrocytes. Glycophorin is tightly associated with the lipids of the membrane and also seems to be in a trans-membrane orientation (7, 8).

We decided to determine the complete amino-acid sequence of this molecule with the expectation that an understanding of its primary structure would give us some insight into the arrangement of its external receptor units, the basis for its association with membrane lipids, and the structural features of its "internalized" segment that may be relevant for further studies on the interactions between glycophorin and other membrane components.

MATERIALS AND METHODS

Glycophorin was isolated as described (3). The two principal sets of peptides, obtained by tryptic and chymotryptic cleavages, were isolated by gel filtration on Sephadex G-150 superfine (2.5 × 150 cm), and subsequently by ion exchange chromatography on DEAE-cellulose, using procedures that were slight modifications of methods described earlier (9). Peptides T-6 and CH-6 were isolated as insoluble precipitates produced as a result of proteolytic digestion. Some of the glycopeptides were subjected to secondary cleavages with trypsin, chymotrypsin, thermolysin, or cyanogen bromide after the desialation of the glycopeptides with a purified neuraminidase from *Clostridium perfringens* (10). The secondary peptides were isolated by gel filtration on Sephadex G-50 superfine (2.5 × 100 cm). For amino-acid analysis, dry samples in small test tubes were hydrolyzed at 110° for 24 hr with a vapor of 6 M HCl containing 2% phenol in a glass chamber equipped with a Viton O-ring. Six to 12 samples were usually hydrolyzed at the same time, and the analyses were performed on a Durrum D-500 analyzer. Carbohydrate compositions were determined as trimethyl silyl derivatives of methyl glycosides by gas-liquid chromatography (Hewlett-Packard 5710B gas chromatograph with an integrator 3373B) as described elsewhere (11). Inositol was used as an internal standard for the carbohydrate analysis. Amino acids released from the COOH-terminus of peptides by the action of carboxypeptidase A and B (12) or by hydrazinolysis (13) were identified on the Durrum analyzer. In some cases the COOH-terminal residue was identified by amino-acid analysis of the degraded peptide after Edman degradation (14). The amino-acid sequences of the peptides were determined by the manual Edman technique (15), which was slightly modified for glycopeptides. Phenylthiohydantoin (PTH) amino acids were identified by gas-liquid chromatography using a 10% DC560 column (14), thin-layer chromatography on polyamide layer (16), and amino-acid analysis after hydrolysis of the samples with 55% HI (17).

Since PTH derivatives of the glycosylated amino acids were not extractable in organic solvents, the Edman-dansyl procedure (18) and the subtractive Edman procedure (19) were used in these cases. Alkaline borohydride treatment of glycopeptides was also used to identify *O*-glycosylated serines; these were identified as PTH-alanine (20). Methanolysis with 0.5 M HCl in methanol was also used to remove carbohydrate from some of the glycopeptides.

RESULTS AND DISCUSSION

The complete amino-acid sequence of human erythrocyte glycophorin is shown in Fig. 1. This molecule is composed of a single polypeptide chain of 131 amino acids to which are attached 16 oligosaccharide units. The sequence presented here was derived from studies of seven unique tryptic peptides, six unique chymotryptic peptides, and one cyanogen bromide fragment prepared from purified glycoprotein

Abbreviation: PTH, phenylthiohydantoin.

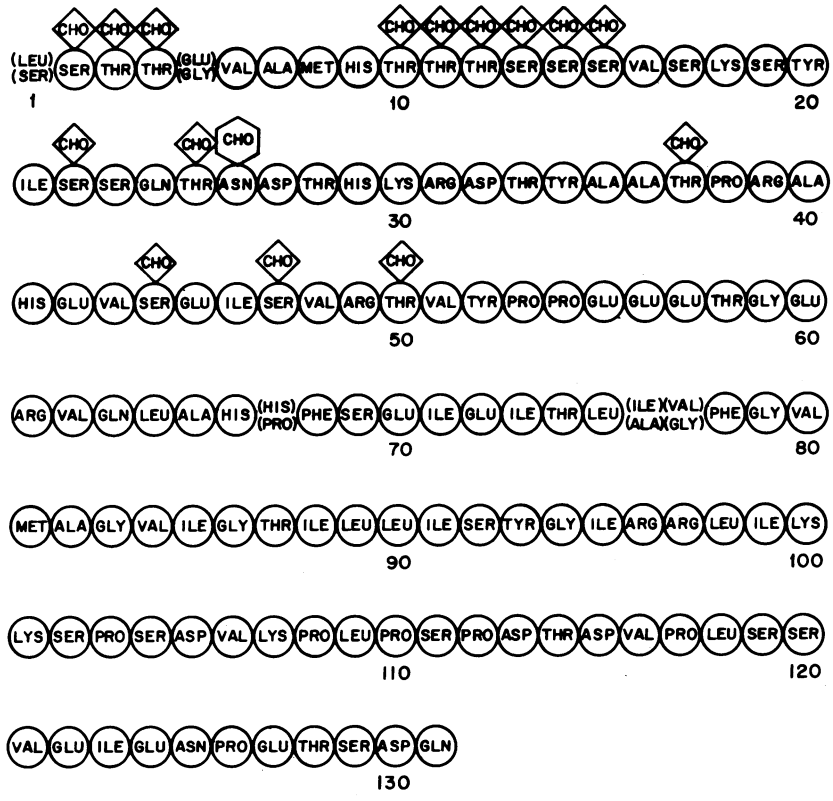

FIG. 1. Amino-acid sequence of human erythrocyte glycophorin. CHO, oligosaccharide: *diamond-shaped*, *O*-glycosidic linkage; *hexagon*, *N*-glycosidic linkage.

preparations isolated from a number of different human donors. Based on composition and sequence analysis, the correct arrangement of these different peptides in the native molecule is presented in Fig. 2.

The arrangement of different types of peptides illustrated in Fig. 2 bears out the earlier speculations concerning the "tripartite" character of the glycophorin molecule. The part of the molecule containing all the sugar residues is referred to as the *glycosylated segment* and is composed of approximately 64 amino acids and 16 oligosaccharide chains. Three peptides, representing a 35 amino-acid segment, are derived from the *COOH-terminal end* of the molecule and are referred to as such. A stretch of polypeptide composed of approximately 32 residues connects these two domains and, for reasons described earlier (21), is referred to as the *hydrophobic domain*.

Amino-acid sequence of the glycosylated domain

The sequence of the glycosylated portion of the molecule was determined by analyzing four tryptic glycopeptides (T-1, T-2, T-3, and T-5) and three chymotryptic glycopeptides (CH-1, CH-2, and CH-3). The amino-acid and carbohydrate compositions of these are shown in Tables 1 and 2. Four of these glycopeptides (T-1, T-2, CH-1, and CH-2) were found to have the same NH_2-terminal sequence as the intact molecule and were subsequently found to represent different lengths of the same polypeptide chain. Each of these peptides was subjected to sequential Edman degradation, but the entire sequence of none could be determined, so that complete overlaps could not be obtained. Our initial efforts to produce smaller peptides from these by additional incubations with thermolysin, subtilisin, and Pronase were unsuccessful. However, smaller secondary peptides were eventually generated by incubating the purified primary peptides with neuraminidase before subjecting them to a second proteolytic degradation. Secondary peptides were produced by tryptic and thermolysin cleavage of desialated tryptic peptides (Fig. 2). Secondary peptides of T-1 were also generated by cyanogen bromide cleavage.

A second problem encountered in determining the sequence of the glycopeptides was the difficulty in identifying the amino acids that were glycosylated. Although degradation occurred through glycosylated residues, the PTH-derivatives of the glycosylated amino acids were insoluble in the organic solvent used to identify the components and, hence,

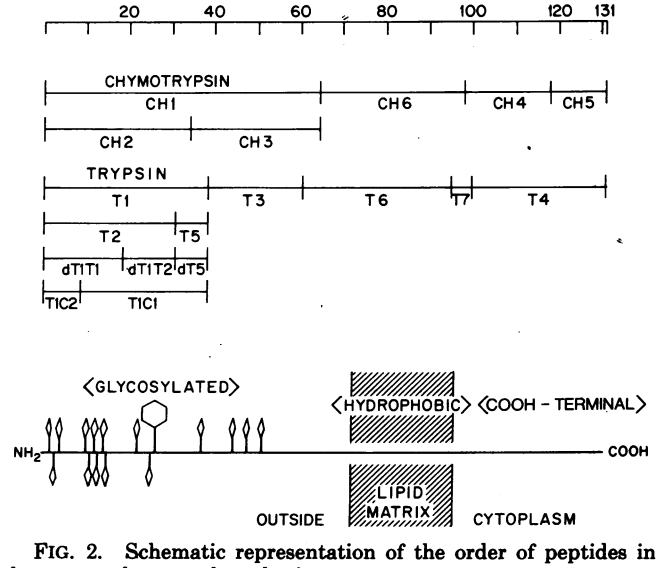

FIG. 2. Schematic representation of the order of peptides in human erythrocyte glycophorin.

Table 1. Chemical composition of the trypsin digest

	T1	T2	T3	T4	T5	T6
Asp	3.0 (3)	1.8 (2)	0.7 (0)	5.1 (5)	1.0 (1)	1.3 (0)
Thr	8.4 (9)	6.7 (7)	2.0 (2)	2.0 (2)	1.5 (2)	2.7 (2)
Ser	7.4 (8.5)	7.3 (8.5)	2.0 (2)	5.2 (6)	0.3	2.9 (2)
Glu	1.4 (1.5)	2.1 (1.5)	5.6 (6)	4.1 (4)	0.3	2.9 (3)
Pro	0.9 (1.0)	0	2.6 (2)	6.2 (6)	1.0 (1)	1.4 (0.5)
Gly	0.7 (0.5)	1.1 (0.5)	1.0 (1)	0.1	0.2	3.7 (4.5)
Ala	2.9 (3)	1.1 (1)	1.3 (1)	0.2	1.7 (2)	2.3 (1.5)
Cys	0	0	0	0	0	0
Val	2.0 (2)	2.3 (2)	3.1 (3)	3.3 (3)	0	2.6 (3.5)
Met	0.7 (1)	0.6 (1)	0	0	0	0.9 (1)
Ile	1.0 (1)	1.1 (1)	1.0 (1)	1.0 (1)	0	5.0 (6.5)
Leu	0.4 (0.5)	0.9 (0.5)	0.2	2.0 (2)	0.2	3.5 (4)
Tyr	1.4 (2)	1.0 (1)	0.7 (1)	0	0.6 (1)	1.1 (1)
Phe	0	0	0	0	0	1.8 (2)
His	1.9 (2)	2.1 (2)	1.1 (1)	0	0	1.4 (1.5)
Lys	1.9 (2)	1.8 (2)	0.2	2.1 (2)	0	0.8 (0)
Arg	1.7 (2)	0.6 (0)	2.0 (2)	0	0.8 (1)	1.3 (1)
Fuc	1.4	0.8	0	0	0	
Man	4.0	2.0	0.1	0	0	
Gal	15.5	12.2	4.2	0	0.8	
Glc	2.2	0.6	0.2	0	0.2	
GalNAc	10.0	10.5	3.0	0	0.7	
GlcNAc	5.8	3.6	0	0	0.1	
NANA	18.6	18.5	7.0	0	1.5	

NANA, acetylneuraminic acid.

could not be identified. The glycosylated amino acids were provisionally identified as dansylated derivatives by a combined Edman-dansyl procedure. The degradation process seemed to proceed through *O*-glycosylated amino acids with the same yield obtained with nonglycosylated residues, but the yield was considerably lower (approximately 30% of expected) in the case of *N*-glycosylated asparagine, as judged by the yield of the PTH derivative of the next amino acid. A complete discussion of the different approaches used to identify the glycosylated residues in different glycopeptides will be presented elsewhere (Tomita and Marchesi, manuscript in preparation).

Another unusual feature of the glycosylated segment of the glycophorin was that both trypsin and chymotrypsin

Table 2. Chemical composition of the chymotrypsin digest

	CH 1	CH 2	CH 3	CH 4	CH 5	CH 6
Asp	3.1 (3)	2.8 (3)	0.2	3.1 (3)	2.2 (2)	1.4
Thr	10.3 (11)	7.7 (8)	3.3 (3)	1.1 (1)	1.0 (1)	2.7
Ser	8.6 (10.5)	6.9 (8.5)	2.3 (2)	2.5 (3)	2.5 (3)	2.9
Glu	7.9 (8.5)	2.2 (1.5)	6.9 (7)	0.2	4.1 (4)	2.0
Pro	3.1 (3)	0	2.4 (3)	5.3 (5)	1.1 (1)	2.7
Gly	1.8 (1.5)	1.2 (1.5)	1.2 (1)	0.1	0.2	2.9
Ala	4.5 (4)	1.3 (1)	3.5 (3)	0.1	0.3	2.5
Cys	0	0	0	0	0	0
Val	6.1 (6)	2.2 (2)	4.6 (4)	2.3 (2)	1.1 (1)	2.3
Met	0.6 (1)	0.6 (1)	0.1	0	0	0.9
Ile	2.0 (2)	1.1 (1)	1.0 (1)	1.0 (1)	1.1 (1)	4.6
Leu	1.3 (1.5)	0.8 (0.5)	1.0 (1)	2.0 (2)	0.2	3.5
Tyr	2.6 (3)	1.4 (2)	0.7 (1)	0	0	1.0
Phe	0	0	0	0	0	1.4
His	3.0 (3)	1.8 (2)	0.9 (1)	0	0	1.2
Lys	2.1 (2)	1.8 (2)	0.2	2.9 (3)	0.2	1.0
Arg	3.0 (4)	0.6 (1)	2.6 (3)	0	0.1	1.4
Fuc	1.3	1.5	0	0	0	
Man	5.2	4.0	0.1	0	0	
Gal	24.9	18.7	5.2	0	0	
Glc	0	0	0.6	0	0	
GalNAc	14.5	12.4	4.0	0	0	
GlcNAc	8.1	7.3	0.5	0	0	
NANA	28.7	19.2	7.3	0	0	

NANA, acetylneuraminic acid.

produced two sets of peptides that were ostensibly derived from the same segment of the polypeptide chain (Fig. 2). The composition and sequence of T-1 are essentially identical to those of T-2 and T-5 together, yet a second digestion of T-1 with trypsin did not result in the production of T-2 and T-5. A similar set of glycopeptides are also derived by chymotryptic digestion (CH-1 compared with CH-2 and CH-3). However, tryptic digestion of the CH-1 glycopeptide produced T-1, T-3, and the tripeptide Val-Gln-Leu, but neither T-2 nor T-5 was generated. We feel that these results are due to a difference in the degree of glycosylation of a single polypeptide chain. As mentioned earlier, removal of sialic acids significantly affects the sensitivity of these glycopeptides to proteolytic cleavage. When desialated T-1 is treated with trypsin a second time, "new" glycopeptides, labeled dT1-T1 and dT1-T2 in Fig. 2, are generated, as is the desialated form of T-5.

Second amino acids were detected at the amino-terminus and in the fifth position, which we attribute to heterogeneity rather than contamination. Both sets of residues appeared in all the samples we analyzed, and in the secondary peptides, T_1-C_2.

The "hydrophobic" domain

In a previous study we isolated and determined the partial sequence of the peptides derived from glycophorin that had a high content of nonpolar amino acids and solubility properties which suggested that they might be derived from a hydrophobic segment of the polypeptide chain (21). The complete sequence of a tryptic peptide (labeled T-6 here) and the partial sequence of a cyanogen bromide fragment were obtained with a Beckman automatic sequenator (21), and similar results have been obtained during the present study. However, in both cases a number of "minor" amino acids were detected at several steps, and it is not clear whether these represent amino-acid substitutions in the polypeptide chain or residues derived from contaminating peptides. We have attempted to purify these peptides by a variety of different ways, but have not yet succeeded in isolating material that gives a single, predominant residue at each step of the sequence. Our present interpretation is that the double amino acids in positions 67, 76, and 77 (Fig. 1) are due to contaminating peptides rather than to amino acid substitutions in different glycophorin molecules. However, this question will not be resolved until this part of the sequence is redone with pure peptides.

The COOH-terminal segment

On the basis of earlier labeling studies from this and other laboratories, it appeared that a segment of the polypeptide chain of the glycophorin molecule was situated internal to the lipid barrier of the membrane. Since the NH_2-terminal end of the molecule was clearly located outside the cell membrane, it seemed simplest to postulate that the COOH-terminal end of the polypeptide chain extended into the cytoplasm of the cell (7, 8) although other, more elaborate, "loop-type" arrangements could not be ruled out.

Tryptic peptide T-4 and chymotryptic peptides CH-4 and CH-5 are derived from this segment, and their sequences are included in Fig. 1. Part of the sequence of this segment of the molecule had been obtained earlier by determination of the sequence of 30 residues of a cyanogen bromide fragment (21), and the sequences of both T-4 and CH-4 overlapped and confirmed these earlier data. Although a number of threonine and serine residues and one asparagine residue are present in this segment of the molecule, none seem to be glycosylated. This is consistent with previous suggestions that all the sugar residues of this molecule are located on the external side of the lipid bilayer.

Structure of the carbohydrate units

Each polypeptide chain of glycophorin has an average of 15 *O*-glycosidically linked oligosaccharides and a single *N*-glycosidic unit. Two different structures have been proposed for the complex oligosaccharide that is linked *N*-glycosidically to glycophorin (22, 23), but neither one is consistent with our findings. Our data indicate that the asparagine-linked oligosaccharide should contain approximately four residues of mannose and seven of *N*-acetyl glucosamine rather than one-half the number of each, which have been reported. Since the glycosylated asparagine is linked to aspartic acid, it is conceivable that the previous studies were carried out on glycopeptides containing two aspartate residues, and carbohydrate values normalized to a single aspartic acid residue would be misleading.

Our values for the carbohydrate composition of the *O*-glycosidically linked oligosaccharides are consistent with the tetrasaccharide structure proposed previously (24). These units seem to be involved in specifying MN blood group specificity (1), but we have as yet no way of determining which of the 15 units is involved. We also do not know whether some threonine/serine residues are only partially glycosylated.

Molecular weight of the monomeric unit of glycophorin

A variety of different values have been reported for the molecular weight of glycophorin, ranging from 15,000 to 400,000 (4–6, 25). It is generally acknowledged that heavily glycosylated* proteins cannot be analyzed by gel filtration or by sodium dodecyl sulfate-gel electrophoresis with great confidence, but it has only become evident more recently that amphiphatic proteins (like glycophorin) are difficult to dissociate into their monomeric units by the usual denaturing agents (26). It has been suggested recently that glycophorin might exist as a stable dimeric unit even in the presence of high concentrations of sodium dodecyl sulfate (27). The molecular weight of the glycoprotein unit described here is approximately 31,000, as calculated from the amino-acid sequence and carbohydrate content. This value is remarkably similar to that recently obtained by analyzing glycophorin by sedimentation equilibrium in the presence of sodium dodecyl sulfate (28).

The present results clarify our earlier preliminary findings which were based on an analysis of peptides of glycophorin produced by cyanogen bromide cleavage (8). We now know that the glycoprotein preparations used in these earlier studies contained small amounts of a different glycopeptide that copurified with the glycophorin molecule. The presence of these contaminants accounts for the higher molecular weight and the higher content of methionine residues and additional cyanogen bromide fragments reported earlier (8). An improved method for the isolation of glycophorin has been developed (Furthmayr, Tomita, and Marchesi, manuscript in preparation), and peptides produced by cyanogen bromide cleavage of this material are those expected from a molecule containing two methionine residues.

* It is not clear what percentage of sugar is required to put a protein in this category; glycophorin is 60% carbohydrate by weight.

Molecular anatomy of glycophorin

The amino-acid sequence and the location of oligosaccharides are completely consistent with the trans-membrane orientation of glycophorin postulated previously (8). The sugar attachment sites are confined to the NH_2-terminal 50 amino acids and are thus entirely exposed to the outside of the cell. The sequence of the COOH-terminal region of the molecule is particularly interesting in that it contains a large number of acidic amino acids, along with threonine, serine, and proline. Thus, this segment of glycophorin is well suited to bind cations such as Ca^{2+} or interact electrostatically with basic peptides or with the amino groups of phospholipids in the internal half of the lipid bilayer.

We wish to thank Manyee Chung for excellent technical assistance. This work was supported by USPHS Research Grant GM-19929 and Membrane Center Grant no. P17-GM 21714.

1. Winzler, R. J. (1969) in *Red Cell Membrane, Structure and Function*, eds. Jamieson, G. A. & Greenwalt, T. J. (J. B. Lippincott Co., Philadelphia, Pa.), pp. 157–171.
2. Marchesi, V. T., Tillack, T. W., Jackson, R. L., Segrest, J. P. & Scott, R. E. (1972) *Proc. Nat. Acad. Sci. USA* **69**, 1445–1449.
3. Marchesi, V. T. & Andrews, E. P. (1971) *Science* **174**, 1247–1248.
4. Springer, G. F., Nagai, Y. & Tegtmeyer, H. (1969) *Biochemistry* **5**, 3254–3272.
5. Blumenfeld, O. O. & Zvilichovsky, B. (1972) in *Methods in Enzymology*, ed. Ginsburg, V. (Academic Press, New York), Vol. 28, pp. 245–252.
6. Fukuda, M. & Osawa, T. (1973) *J. Biol. Chem.* **248**, 5100–5105.
7. Bretscher, M. S. (1971) *Nature New Biol.* **231**, 229–232.
8. Segrest, J. P., Kahane, I., Jackson, R. L. & Marchesi, V. T. (1973) *Arch. Biochem. Biophys.* **155**, 167–183.
9. Jackson, R. L., Segrest, J. P., Kahane, I. & Marchesi, V. T. (1973) *Biochemistry* **12**, 3131–3138.
10. Cuatrecasas, P. & Illiano G. (1971) *Biochem. Biophys. Res. Commun.* **44**, 178–184.
11. Reinhold, V. N. (1972) in *Methods in Enzymology*, eds. Hirs, C. H. W. & Timasheff, S. N. (Academic Press, New York), Vol. 25, 244–249.
12. Ambler, R. P. (1972) in *Methods in Enzymology*, eds. Hirs, C. H. W. & Timasheff, S. N. (Academic Press, New York), Vol. 25, 143–154.
13. Schroeder, W. A. (1972) in *Methods in Enzymology*, eds. Hirs, C. H. W. & Timasheff, S. N. (Academic Press, New York), Vol. 25, 138–143.
14. Sauer, R. T., Niall, H. D., Hogan, M. L., Keutmann, H. T., O'Riodan, J. L. H. & Pott, J. T., Jr. (1974) *Biochemistry* **13**, 1994–1999.
15. Peterson, J. D., Nehrlich, S., Oyer, P. E. & Steiner, D. F. (1972) *J. Biol. Chem.* **247**, 4866–4871.
16. Summers, M. R., Smythers, G. W. & Oroszlan, S. (1973) *Anal. Biochem.* **53**, 624–628.
17. Smithies, O., Gibson, D., Fanning, E. M., Goodfliesh, R. M., Gilman, J. G. & Ballantyne, D. L. (1971) *Biochemistry* **10**, 4912–4921.
18. Gray, W. R. (1972) in *Methods in Enzymology*, eds. Hirs, C. H. W. & Timasheff, S. N. (Academic Press, New York), Vol. 25, 121–138.
19. Konigsberg, W. (1972) in *Methods in Enzymology*, eds. Hirs, C. H. W. & Timasheff, S. N. (Academic Press, New York), Vol. 25, 326–332.
20. Spiro, R. G. & Bhoyroo, V. D. (1974) *J. Biol. Chem.* **249**, 5704–5717.
21. Segrest, J. P., Jackson, R. L., Marchesi, V. T., Guyer, R. B. & Terry, W. (1972) *Biochem. Biophys. Res. Commun.* **49**, 964–969.
22. Kornfeld, R. & Kornfeld, S. (1970) *J. Biol. Chem.* **245**, 2536–2546.
23. Thomas, D. B. & Winzler, R. J. (1971), *Biochem. J.* **124**, 55–59.
24. Thomas, D. B. & Winzler, R. J. (1969) *J. Biol. Chem.* **244**, 5943–5946.
25. Janado, M., Azuma, J. & Onodera, K. (1973) *J. Biochem.* **74**, 881–887.
26. Maddy, A. H. & Kelly, P. G. (1971) *Biochim. Biophys. Acta* **241**, 114–116.
27. Marton, L. S. G. & Garvin, J. E. (1973) *Biochem. Biophys. Res. Commun.* **49**, 964–969.
28. Grefrath, S. P. & Reynolds, J. A. (1974) *Proc. Nat. Acad. Sci. USA* **71**, 3913.