THE OCCURRENCE IN PROTEINS OF THE TRIPEPTIDES ASN-X-SER AND ASN-X-THR AND OF BOUND CARBOHYDRATE*

Lois T. Hunt and Margaret O. Dayhoff
National Biomedical Research Foundation
Silver Spring, Maryland 20901

Received April 14, 1970

SUMMARY

The 101 occurrences of the tripeptides Asn-X-Ser and Asn-X-Thr in the available protein sequence data are tabulated; carbohydrate is found, attached to the asparagine, in not more than 20 of the 101 tripeptides. A statistical analysis of the data from all completely sequenced proteins shows that the observed frequency of occurrence of the two kinds of tripeptides is only about 65% of the expected. This lowered frequency is evidence for a newly postulated kind of limitation — which we call a "restricted sequence" — imposed by natural selection on the primary structure of proteins.

In a number of glycoproteins the carbohydrate (CHO) prosthetic group appears to be bound N-glycosidically to an asparagine (Asn) residue in the polypeptide chain; the tripeptide sequence is generally reported to be Asn-X-Ser (serine) or Asn-X-Thr (threonine) (1,2,3), where X is any amino acid. We have recently received several requests for information on the occurrence of these two tripeptide sequences in those proteins which have been sequenced.

We have therefore searched the sequences appearing in the *Atlas of Protein Sequence and Structure 1969* (4) for occurrences of the sequences Asn-X-Ser and Asn-X-Thr. There are about 18,000 tripeptide amino acid links in the complete or almost complete sequences, and another 10,000 such links in the fragmentary data, reported in this latest edition of the *Atlas*. We have found a total of 88 occurrences of these two tripeptides in these data (see Tables I and II). The proteins containing 79 of the tripeptide regions lack bound carbohydrate, while those containing 7 of the tripeptide sequences probably or certainly

*Supported by NIH Grant GM-08710 and NASA Contract 21-003-002.

have carbohydrate attached to asparagine. Two other proteins have both an Asn-X-Thr sequence and bound carbohydrate; however, the carbohydrate is not bound to the asparagine, but to a threonine which is not part of the tripeptide. In addition, 13 tripeptides, from 11 protein fragments not reported in the 1969 *Atlas*, are recorded in Tables I and II; these increase the total tripeptide occurrences in these Tables to 101. The asparagine in these tripeptides is known to have carbohydrate attached (5-16). Several Asx-X-Ser/Thr tripeptides are included (Asx indicates that the presence or absence of an amide group has not been determined), in which it is probable that the first residue is asparagine.

Carbohydrate in glycoproteins is not exclusively bound to asparagine. It is also known to be bound O-glycosidically to serine and threonine, as well as to the modified amino acids hydroxylysine and hydroxyproline (2,3). Glycoproteins containing more than one carbohydrate group may have more than one type of carbohydrate-protein linkage.

We performed a statistical analysis of the occurrences of the tripeptides in the complete (or nearly complete) sequences tabulated in the *Atlas*. A total of 18,251 tripeptide sequences from 264 proteins were grouped by computer into the 400 possible tripeptide combinations of the 20 amino acids in which the second position was ignored. The sequences Asn-X-Ser and Asn-X-Thr occur 36 and 25 times, respectively. The detailed statistics for the expected and observed frequencies of occurrence of these tripeptides are given in Table III.

The number of Asn-X-Ser/Thr sequences observed is only 60% of that expected. The total number of these sequences actually counted by computer was 61, whereas the total number expected, on the basis of a random distribution of the amino acids within each sequence, is 101.9. The standard deviation of the expected frequency is theoretically approximated by $\sqrt{101.9}$ or 10.1. The observed frequency is then 4 standard deviations lower than the expected, a highly unusual count to have resulted from chance variation ($P < 0.0001$ for a normal distribution).

It may be misleading to accumulate data from each protein sequence sepa-

All Asn-X-Ser and Asn-X-Thr tripeptide regions found in the protein sequence data from the *Atlas of Protein Sequence and Structure 1969* (4) are listed. Data from eleven sequences not reported in the *Atlas* are also included; reference numbers are in parentheses. A few of the glycoproteins listed above contain more than one CHO unit; some of these other units are not necessarily attached to Asn-X-Ser/Thr tripeptides. Data used are from complete sequences and from fragments. In the case of a fragment whose real position within a protein is not known, the position numbers are enclosed in parentheses and are taken from the data section of the 1969 *Atlas*. The presence or absence of bound CHO is indicated in the 4th column by the symbols + or -; a question mark indicates that we have no information concerning its presence. In the 5th column, a + or - indicates whether or not the bound CHO is attached to the tripeptide asparagine, and a question mark, that we have no information on the CHO binding site. In the punctuation of sequences, a hyphen is placed between residues whose position is determined, and a period between residues sequenced by homology only. The abbreviation (*fr*) follows a fragmentary sequence.

TABLE I

Occurrence of Asn-X-Ser and Asn-X-Thr Tripeptides

	TRIPEPTIDE		CARBOHYDRATE	
			PRE-	BOUND
PROTEIN and ORGANISM	SEQUENCE	POSITION	SENT	TO ASN
Cytochrome c - Tuna Fish	Asn-Lys-Ser Asn.Asp)Thr	52-54 61-63	- -	
Cytochrome c - Puget Sound Dogfish	Asn-Leu-Ser Asn-Lys-Ser	31-33 52-54	- -	
Cytochrome c - Lamprey	Asn-Lys-Ser	52-54	-	
Cytochrome c - *Neurospora crassa*	Asn-Leu-Thr	27-29	-	
Cytochrome c - Baker's Yeast (*S. oviformis*)	Asn-Met-Ser	68-70	-	
Cytochrome c - Baker's Yeast (*S. cerevisiae*)	Asn-Met-Ser	68-70	-	
Cytochrome c_3 - *Desulfovibrio vulgaris*	Asn-His-Ser	21-23	-	
Cytochrome c_{551} - *Pseudomonas fluorescens*	Asn-Gly-Ser	50-52	-	
Cytochrome b_5 - Bovine	Asn-Asn-Ser	16-18	-	
Azurin - *Pseudomonas fluorescens*	Asn-Leu-Ser	32-34	-	
Azurin - *Alcaligenes faecalis*	Asn-Asp-Ser	9-11	-	
Ferredoxin - *Micrococcus aerogenes*	Asn-Asp-Ser	5-7	-	
Ferredoxin - *Clostridium butyricum*	Asn-Asp-Ser	5-7	-	
Ferredoxin - Alfalfa	Asn-Gln-Ser	57-59	-	
High Potential Iron Protein - *Chromatium* D	Asn-Ala-Thr	11-13	-	
Hemoglobin α Chain - Rabbit	Asn-Val-Ser	131-133	-	
Hemoglobin α Chain - Pig	Asx-Val-Ser	131-133	-	
Hemoglobin α Chain - Bovine	Asn-Val-Ser	131-133	-	
Hemoglobin γ Chain - Human	Asn-Leu-Ser	47-49	-	
Fibrinogen γ (C) Chain - Human (*fr*)	Asn-Lys-Thr	52-54	+	+
*Immunoglobulin G1 γ Chain - Human EU (5)	Asx-Ser-Thr	297-299	+	+

Immunoglobulin G γ Chain - Rabbit (*fr*)(4,6)	Asx-Ser-Thr	(194-196)	+	+
*Immunoglobulin G γ Chain - Bovine (*fr*)(7)	Asx-Ser-Thr	?	+	+
Immunoglobulin G γ Chain - Mouse ADJPC5 (*fr*)	Asx.Ser.Thr	?	+	+
Immunoglobulin G γ Chain - Mouse MOPC21 (*fr*)	Asx.Ser.Thr	?	+	+
Immunoglobulin M μ Chain - Human OU (*fr*)	Asn-Asp-Ser	74-76	+	?
Bence Jones λ Chain - Human HA	Asn-Gly-Thr	28-30	-	
Bence Jones λ Chain - Human BO	Asn-Asp-Thr	70-72	-	
Immunoglobulin G κ Chain - Guinea Pig (C-t *fr*)	Asn-Arg-Ser	(3-5)	?	
*Bence Jones κ Chain - Mouse MOPC46 (*fr*)(8)	Asx-Ile-Ser	28-30	+	+
Trypsinogen - Bovine	Asn-Ser-Ser	151-153	-	
Elastase - Pig	Asn-Gly-Thr	66-68	-	
	Asn-Asn-Ser	123-125	-	
	Asn-Val-Thr	215-217	-	
Subtilisin - *Bacillus amyloliquifaciens*	Asn-Asn-Ser	76-78	-	
	Asn-Met-Ser	123-125	-	
	Asn-Gly-Thr	218-220	-	
	Asn-Trp-Thr	240-242	-	
	Asn-Thr-Thr	252-254	-	
Subtilisin - *Bacillus subtilis* Carlsberg	Asn-Asn-Thr	76-78	-	
	Asn-Ser-Ser	96-98	-	
	Asn-Met-Ser	122-124	-	
	Asn-Gly-Thr	217-219	-	
	Asn-Leu-Ser	239-241	-	
Pepsinogen - Pig (*fr*)	Asn-Asn-Ser	(253-255)	-	
Carboxypeptidase A - Bovine (*fr*)	Asn-Pro-Ser	93-95	-	
*Bromelain - Pineapple stem (*fr*)(9)	Asn-Glu-Ser	?	+	+
*Deoxyribonuclease - Bovine (*fr*)(10)	Asn-Ala-Thr	?	+	+
Nuclease - *Staphylococcus aureus* V8	Asn-Asn-Thr	118-120	?	
Nuclease - *S. aureus* Foggi	Asn-Asn-Thr	118-120	?	
Ribonuclease (B,C,D) - Bovine	Asn-Leu-Thr	34-36	+	+
Ribonuclease - Rat	Asn-Cys-Thr	97-99	-	
	Asn-Thr-Thr	101-103	-	
*Ribonuclease - Pig (11)	Asn-Ser-Ser	21-23	+	+
	Asn-Met-Thr	34-36	+	+
	Asn-Ser-Thr	76-78	+	+
Lactalbumin - Bovine	Asn-Ile-Ser	74-76	+	?
Lysozyme - Duck II (*fr*)	Asn-Gly-Ser	48-50	-	
Lysozyme - Duck III (*fr*)	Asn-Gly-Ser	48-50	-	
Lysozyme - Bacteriophage T2	Asn-Gln-Thr	140-142	-	
Lysozyme - Bacteriophage T4	Asn-Gln-Thr	140-142	-	
Tryptophan Synthetase α Chain - *E. coli*	Asn-Ala-Thr	65-67	-	
Glyceraldehyde 3-PO_4 Dehydrogenase - Pig	Asn-Ala-Ser	146-148	-	
	Asn-Val-Ser	236-238	-	
	Asn-Asp-Ser	284-286	-	
Glyceraldehyde 3-PO_4 Dehydrogenase - Lobster	Asn-Ala-Ser	145-147	-	
	Asn-Arg-Ser	286-288	-	

Catalase - Bovine (*fr*)	Asn-Leu-Ser	(258-260)	-	
	Asn-Val-Thr	(285-287)	-	
	Asn-Phe-Ser	(339-341)	-	
Penicillinase - *Staphylococcus aureus*	Asn-Lys-Ser	181-183	-	
	Asn-Lys-Ser	236-238	-	
Growth Hormone - Horse (*fr*)	Asn-Cys-Ser	(9-11)	-	
*Thyroglobulin - Human (*fr*)(12)	Asx-Ala-Thr	?	+	+
Thyrocalcitonin - Pig	Asn-Leu-Ser	3-5	-	
Coat Protein - Tobacco Mosaic Virus *vulgare*	Asn-Pro-Thr	101-103	-	
	Asn-Arg-Ser	140-142	-	
Coat Protein - Tobacco Mosaic Virus OM	Asn-Pro-Thr	101-103	-	
	Asn-Arg-Ser	140-142	-	
Coat Protein - Tobacco Mosaic Virus U2	Asn-Ser-Thr	73-75	-	
Coat Protein - Tobacco Mosaic Virus HR	Asn-Ile-Thr	3-5	-	
	Asn-Ala-Thr	109-111	-	
Coat Protein - Bacteriophage F2	Asn-Phe-Thr	3-5	-	
	Asn-Val-Thr	17-19	-	
Coat Protein - Bacteriophage MS2	Asn-Phe-Thr	3-5	-	
	Asn-Val-Thr	17-19	-	
Coat Protein - Bacteriophage R17	Asn-Phe-Thr	3-5	-	
	Asn-Val-Thr	17-19	-	
Myosin - Rabbit (*fr*)	Asn-Phe-Thr	(49-51)	-	
	Asn-Glu-Thr	(110-112)	-	
Haptoglobin α 1 - Human F and S	Asn-Asp-Ser	2-4	-	
Haptoglobin α 2 - Human F/S	Asn-Asp-Ser	2-4	-	
*α_1-Acid Glycoprotein - Human (*fr*)(13)	Asn-Gly-Thr	?	+	+
*Ba-α_2-Glycoprotein - Human (*fr*)(14)	Asx-Asx-Thr	?	+	+
κ Casein - Bovine A (*fr*)	Asn-Val-Thr	(28-30)	+	-
κ Casein - Bovine B (*fr*)	Asn-Val-Thr	(28-30)	+	-
Acyl Carrier Protein - *E. coli* E-26	Asn-Ala-Ser	25-27	-	
*Avidin - Chicken egg-white (*fr*)(15)	Asn-Met-Thr	17-19	+	+
*Ovalbumin - Chicken egg-white (*fr*)(16)	Asn-Leu-Thr	?	+	+

*Sequence data not in 1969 *Atlas*, but published later.

rately, because many related sequences repeat certain tripeptides preferentially. Therefore, we have also considered the alternative extreme hypothesis, namely, that average values for each family, instead of values for the individual sequences, should be used. For this calculation, the data were grouped into 69 families of related sequences. Each family contributed to the total tripeptide count the number which would be derived from one protein in the family.

TABLE II

Numerical Tabulation of Asn-X-Ser and Asn-X-Thr Tripeptide Regions Listed in Table I

	Number of Tripeptide Sequences
1a) Protein Lacks Carbohydrate (CHO)	76
1b) Protein Probably Lacks CHO	3
2) Protein Has Bound CHO and Tripeptide	
a) CHO Bound to Tripeptide Asn	18
b) CHO May be Bound to Tripeptide Asn	2
c) CHO Not Bound to Tripeptide Asn	2
TOTAL	101

TABLE III

Statistics for the Frequencies of Occurrence of Two Tripeptide Sequences

Sequences given equal weight:

Sequence	Observed	Expected	Obs.-Exp.	# St. Dev. of Obs. from Exp.	Obs./Exp.
Asn-X-Ser	36	54.9 ± 7.4	-18.9	-2.5	0.66
Asn-X-Thr	25	47.0 ± 6.9	-22.0	-3.2	0.53
Asn-X-S/T	61	101.9 ± 10.1	-40.9	-4.0	0.60

69 families of related sequences averaged:

Sequence	Observed	Expected	Obs.-Exp.	# St. Dev. of Obs. from Exp.	Obs./Exp.
Asn-X-Ser	18.9	25.1 ± 5.0	-6.2	-1.2	0.75
Asn-X-Thr	13.4	19.6 ± 4.4	-6.2	-1.4	0.69
Asn-X-S/T	32.3	44.7 ± 6.7	-12.4	-1.9	0.72

Observed and expected values were accumulated for each sequence in a family and were scaled to obtain the family contribution to the total. The total number for all of the sixty-nine families is 6,884 tripeptides.

The observed number of weighted occurrences of Asn-X-Ser/Thr is 32.3, while the expected number is 44.7, with a theoretical standard deviation of 6.7 (see Table III). The observed number is 1.9 standard deviations less than the expected number. By chance, one would obtain such a low count less than 3% of the time. It is much more likely that there is a systematic reason for the low count.

The two methods of averaging are seen to give similar results for the ratio of observed to expected occurrences (60% and 72%). Any other reasonable scheme for weighting related sequences would give an answer intermediate between those given by the two methods we have used.

Because the low frequency of the Asn-X-Ser/Thr tripeptides might be due to a general chemical or steric factor related to the tripeptide sequence itself, we investigated the occurrences of other chemically similar tripeptides: Asp-X-Ser/Thr, Glu-X-Ser/Thr, Gln-X-Ser/Thr, Ser/Thr-X-Asn, Ser/Thr-X-Asp, Ser/Thr-X-Glu and Ser/Thr-X-Gln. Our statistical analysis shows that, in contrast to the low frequency of occurrence of the two asparagine tripeptides, the occurrences of the chemically similar tripeptides are more frequent than expected in some cases and are never less frequent than one standard deviation below random expectation. It thus appears likely that the restricted occurrence of the Asn-X-Ser/Thr sequences is due to enzymatic recognition rather than to an effect of the protein structure itself.

It is reasonable to suppose that, whether or not a protein is normally a glycoprotein, carbohydrate may be attached if the two tripeptides are generally recognizable by specific enzymes (glycosyltransferases), and if they are located on the outside of a protein (where such hydrophilic residues as asparagine, serine and threonine usually occur) (17) and thus are accessible to glycosyltransferases. In those proteins whose tertiary configuration has been determined, the asparagine tripeptides are located on the outside, whether or not carbohydrate is bound to the asparagine (18). Some reasons that carbohydrate may not be bound to the asparagine tripeptides are: first, there may be compartmentalization of the protein and the glycosyltransferase in the cell; second, there may not be synthesis of glycosyltransferases within a particular cell type; third, there may be steric hindrance (not related to the tripeptide itself).

We suggest that the frequency of occurrence of the Asn-X-Ser/Thr tripeptides in the available protein sequences, which is considerably lower than expected, reflects a restriction by natural selection on the occurrence of the two tripeptides in proteins. Selection would reject a protein which acquired the tripeptide(s) by mutation, if carbohydrate, bound to the tripeptide by the enzyme, subsequently interfered with a normal interaction or function of the protein.

At the present time, relatively few glycoprotein sequences are available for studies such as this one. Of the more than 500 protein sequences that we examined, only 20 are glycoproteins and only 4 of these are complete sequences. The determination of many more glycoprotein sequences is required to provide sufficient data for a statistical investigation of several interesting questions. For example, all of the asparagine tripeptides in the 4 completely sequenced glycoproteins have carbohydrate attached; an examination of the frequency of occurrence, in glycoproteins only, of these two tripeptides would be expected to reveal a low frequency of those lacking carbohydrate. Also, a comparison between glycoproteins and other proteins, with regard to the frequency of occurrence of Asn-X-Ser/Thr tripeptides, might demonstrate significantly different frequencies. The determination of complete sequences of more types of bacterial and viral glycoproteins will permit comparisons between these and the proteins of higher organisms concerning the frequencies of occurrence of the two tripeptides with and without bound carbohydrate.

In conclusion, we find that only about 65% of the expected number of Asn-X-Ser/Thr tripeptides occur in those proteins whose sequences are known. We suggest that this low frequency has resulted from the rejection by selection of a certain number of mutations to these carbohydrate-binding tripeptides in proteins. It is evidence for the existence of a newly observed type of constraint, a sequence restriction, imposed at the molecular level by natural selection and effected through the presence of intracellular glycosyltransferases. The Asn-X-Ser/Thr "restricted sequence" may be a "word" for carbohydrate binding.

Acknowledgement. We thank Drs. Leroy Hood and Harold Sox, of the National Cancer Institute, NIH, for their helpful discussions of the manuscript.

REFERENCES

1. Eylar, E.H., *J. Theor. Biol.*, 10: 89 (1965)
2. Gottschalk, A., *Nature*, 222: 452 (1969)
3. Spiro, R.G., *New England Journal of Medicine*, 281: 991, 1043 (1969)
4. *Atlas of Protein Sequence and Structure 1969*, ed. Dayhoff, M.O., 4: 361 pg., Natl. Biomedical Research Found., Silver Spring, Md. (1969)
5. Edelman, G.M., Cunningham, B.A., Gall, W.E., Gottlieb, P.D., Rutishauser, U. and Waxdal, M.J., *Proc. Nat. Acad. Sci. U.S.*, 63: 78 (1969)
6. Nolan, C. and Smith, E.L., *J. Biol. Chem.*, 237: 446 (1962)
7. Howell, J.W., Sanders, B.G. and Hood, L., *J. Mol. Biol.*, 30: 555 (1967)
8. Melchers, F., *Biochemistry*, 8: 938 (1969)
9. Takahashi, N., Yasuda, Y., Kuzuya, M. and Murachi, T., *J. Biochem. (Tokyo)*, 66: 659 (1969)
10. Catley, B.J., Moore, S. and Stein, W.H., *J. Biol. Chem.*, 244: 933 (1969)
11. Jackson, R.L. and Hirs, C.H.W., *J. Biol. Chem.*, 245: 624, 637 (1970)
12. Rawitch, A.B., Liao, T.-H. and Pierce, J.G., *Biochim. Biophys. Acta*, 160: 360 (1968)
13. Satake, M., Okuyama, T., Ishihara, K. and Schmid, K., *Biochem. J.*, 95: 749 (1965)
14. Ishihara, K. and Schmid, K., *Biochim. Biophys. Acta*, 133: 56 (1967)
15. DeLange, R.J., *Fed. Proc.*, 28: 343 (1969)
16. Cunningham, L., Ford, J.D. and Rainey, J.M., *Biochim. Biophys. Acta*, 101: 233 (1965)
17. Shotton, D.M. and Hartley, B.S., *Nature*, 225: 802 (1970)
18. Dickerson, R.E. and Geis, I., *The Structure and Action of Proteins*, Harper and Row, New York (1969), 120pg.