J|A|C|S
ARTICLES
Published on Web 07/27/2002

Principles of Mucin Architecture: Structural Studies on Synthetic Glycopeptides Bearing Clustered Mono-, Di-, Tri-, and Hexasaccharide Glycodomains

Don M. Coltart,[†] Ajay K. Royyuru,[‡] Lawrence J. Williams,[†,⊥] Peter W. Glunz,[†,⊥] Dalibor Sames,[†,⊥] Scott D. Kuduk,[†,⊥] Jacob B. Schwarz,[†,⊥] Xiao-Tao Chen,[†,⊥] Samuel J. Danishefsky,[*,†,§] and David H. Live[||]

Contribution from the Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota Medical School, Minneapolis, Minnesota 55455, the Laboratory for Bioorganic Chemistry, Sloan-Kettering Institute for Cancer Research, 1275 York Avenue, New York, New York 10021, the Department of Chemistry, Columbia University, New York, New York 10027, and the Computational Biology Center, IBM Thomas J. Watson Research Center, Yorktown Heights, New York, New York 10598

Received February 12, 2002

Abstract: The structural characteristics of a mucin glycopeptide motif derived from the N-terminal fragment STTAV of the cell surface glycoprotein CD43 have been investigated by NMR. In this study, a series of molecules prepared by total synthesis were examined, consisting of the peptide itself, three glycopeptides having clustered sites of α-O-glycosylation on the serine and threonine side chains with the Tn, TF, and STF carbohydrate antigens, respectively, and one with the *β*-O-linked TF antigen. Additionally, a glycopeptide having the sequence SSSAVAV, triglycosylated with the Le^y epitope, was investigated. NMR data for the tri-STF-STTAV glycopeptide were used to solve the structure of this construct through restrained molecular dynamics calculations. The calculations revealed a defined conformation for the glycopeptide core rooted in the interaction of the peptide and the first *N*-acetylgalactosamine residue. The similarity of the NMR data for each of the α-O-linked glycopeptides demonstrates that this structure persists for each construct and that the mode of attachment of the first sugar and the peptide is paramount in establishing the organization of the core. The core provides a common framework on which a variety of glycans may be displayed. Remarkably, while there is a profound organizational effect on the peptide backbone with the α-linked glycans, attachment via a *β*-linkage has little apparent consequence.

Introduction

Large, architecturally sophisticated proteins arrayed with elaborate carbohydrate domains dominate the cell surface landscape. Mucin glycoproteins comprise one of the most significant classes of these cell surface molecules.[1,2] While their roles are not fully understood, it has become increasingly clear that mucins are structurally intricate systems, serving a wide range of functions. For example, it has long been appreciated that mucins form a highly stable, viscous gel that provides a protective barrier over internal epithelial surfaces and protects against chemical, physical, and microbial agents.[2] These properties are attributed to the extensive glycosylation characteristic of mucins. Similarly, glycosylation appears to enhance protein stability by reducing vulnerability to proteolytic degradation.[1,2]

Recent advances in glycobiology point to several other essential functional roles for mucins, including mediation of cellular interactions and signal transduction events by the mucin-based ectodomains of many transmembrane proteins.[3−7] The mucin glycoproteins CD43 and CD45, which are estimated to occupy ~30% of the surface of T-cells,[8] are two such O-linked glycoproteins that are known to act in just such a manner. CD45 serves as an important marker in various cell lines, resulting from cell-specific variation in its three N-terminal mucin glycodomains.[7,9−11] These domains play a role in the localization

[*] Author to whom correspondence should be addressed. E-mail: s-danishefsky@ski.mskcc.org.

[†] Sloan-Kettering Institute for Cancer Research.

[‡] IBM Thomas J. Watson Research Center.

[§] Columbia University.

[||] University of Minnesota Medical School.

[⊥] Current addresses. L.J.W.: Department of Chemistry, Rutgers University, Piscataway, NJ 08854. P.W.G.: Bristol Myers Squibb Co., Wilmington, DE, 19880. D.S.: Department of Chemistry, Columbia University, New York, NY 10027. S.D.K.: Merck & Co., West Point, PA 19486. J.B.S.: Pfizer Inc., Ann Arbor, MI 48105. X.-T.C.: Bristol Myers Squibb Co., Wilmington, DE 19880.

[1] Van den Steen, P.; Rudd, P. M.; Dwek, R. A.; Opdenakker, G. *Crit. Rev. Biochem. Mol. Biol.* **1998**, *33*, 151−208.

[2] Brockhausen, I. In *Glycoproteins*; Montreuil, J., Vliegenthart, J. F. G., Schachter, H., Eds.; Elsevier Science: New York, 1995; pp 201−259.

[3] Lowe, J. B. *Cell* **2001**, *104*, 809−812.

[4] Santanna, M. A.; Pedraza-Alva, G.; Olivares-Zavaleta, N.,; Madrid-Marin, V.; Horejsi, V.; Burakoff, S. J.; Rosenstein, Y. *J. Biol. Chem.* **2000**, *275*, 31460−31468.

[5] Ashwell, J. D.; D'Oro, U. *Immunol. Today* **1999**, *20*, 412−416.

[6] (a) Thomas, M. L.; Brown, E. J. *Immunol. Today* **1999**, *20*, 406−411. (b) Pace, K. E.; Lee, C.; Stewart, P. L.; Baum, L. G. *J. Immunol.* **1999**, *163*, 3801−3811.

[7] Johnson, K. G.; Bromley, S. K.; Dustin, M. L.; Thomas, M. L. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10138−10143.

[8] Shaw, A. S.; Dustin, M. L. *Immunity* **1997**, *6*, 361−369.

10.1021/ja020208f CCC: $22.00 © 2002 American Chemical Society

of CD45 within cell surface microdomains, thereby affecting the activity of its intracellular phosphatase domain in signal transduction processes.[5,6] The extracellular domain of CD43 functions in cell−cell recognition, and CD43 also serves as an accessory protein in the signal transduction process via conserved phosphorylation sites contained within its intracellular region.[12,13]

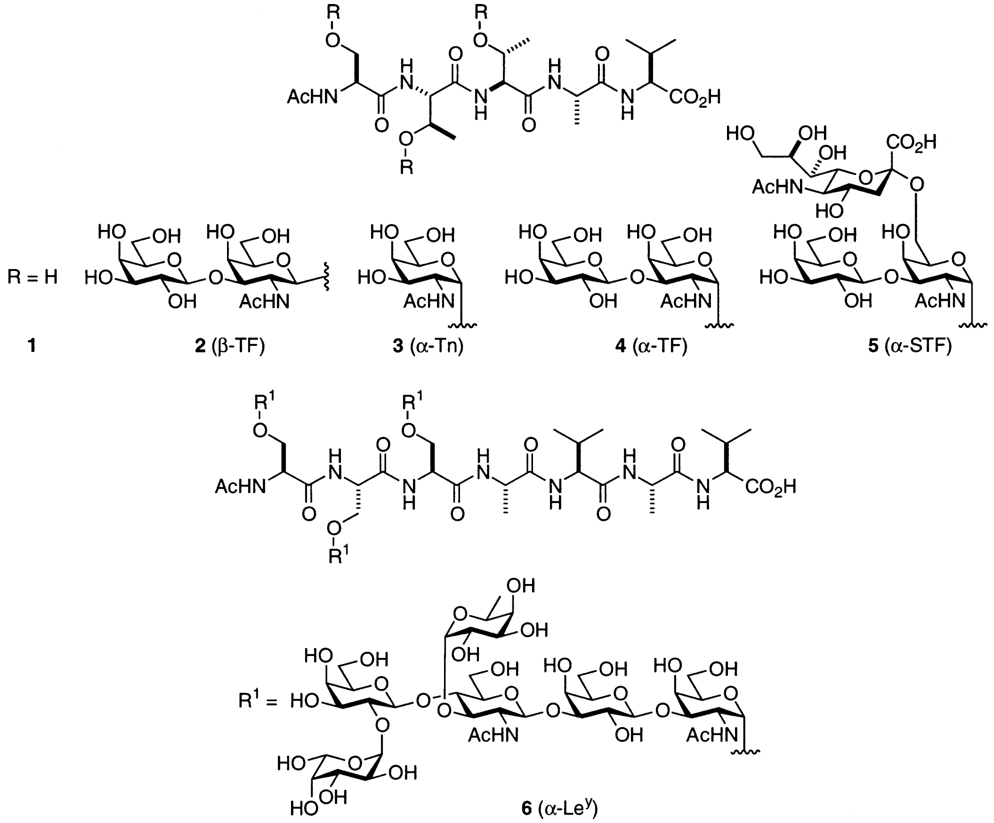
The diversity of glycans within the broad mucin family, including CD43 and CD45, appears to vary with cell type, development, and physiological state.[14−16] Indeed, in certain cases, mucin expression levels and architecture can serve as markers for the onset of disease. For instance, upon carcinogenic transformation, the expression level and structure of cell surface carbohydrates are often significantly altered, thus differentiating cancerous cells from normal cells. Cell surface carbohydrates, therefore, may be especially suited for targeting by the immune system, as they provide potential for a directed therapeutic approach.[16−19] In an effort to exploit this variation in glycan composition between normal and transformed cells, we have been engaged in a program to develop synthetic carbohydrate-based cancer vaccines to combat epithelial cancer recurrence.[17] Our strategy has been to prepare, by total synthesis, sophisticated tumor-associated carbohydrate antigens known to be overexpressed on certain cancer cells. Presentation of such constructs in an appropriate immunological context could well elicit an effective anticancer immune response. One set of vaccines that we have recently advanced to human clinical trials consists of a glycopeptide containing carbohydrate antigens displayed as a mucin motif.[17]

Mucins are characterized as having sequentially arrayed serine and threonine residues in the protein primary structure, most of which are glycosylated with polysaccharides via the side-chain hydroxyl substituent.[1,2] Importantly, the first carbohydrate residue is a conserved GalNAc, α-O-linked to the protein domain. Beyond this regularity, relatively little detailed structural information has been acquired for mucins. Obviously, a detailed description of the molecular architecture of these systems could be valuable in rationalizing their biological roles. However, to date, structural definition of mucins at high resolution has been elusive, which is largely due to the molecular size and microheterogeneity in glycosylation of samples isolated from natural sources. Structural information pertaining to mucins that has been gathered thus far has relied primarily on NMR analysis and, in some cases, has involved computer-generated modeling based on the limited NMR data.[20−41] The results of such studies have been expressed in largely qualitative terms, and the information that has been gleaned generally suggests that native mucin glycoforms have a stiffened, extended structure and do not prefer conventional globular forms.[20−23] Under such circumstances, intramolecular interactions of sequentially remote segments in the native glycoprotein are unlikely. Consequently, we wondered whether more tractable glycopeptide segments of suitable complexity levels could serve as realistic models for the full mucin structure. Such systems could provide a realistic basis on which to develop a high-resolution model of larger mucin domains. The fact that antibodies generated in response to immunization by smaller glycopeptides based on a given tumor-associated carbohydrate antigen recognize corresponding epitope patterns on tumor cells further supports the potential relevance of these truncated structures to the native molecules as key components of the cell surface landscape.[17,42−45]

As a direct result of the synthetic efforts related to our cancer vaccine program,[17] we have overcome the difficulties associated with obtaining homogeneously pure glycopeptides that strictly conform to the complex features of the clustered mucin motif. Using methodologies developed in our laboratory, we have synthesized a collection of mucin glycopeptide segments in which the core carbohydrate and the pendant glycans are varied (see Chart 1). These compounds are based on the pentapeptide

[9] Luqma, M.; Johnson, P.; Trowbridge, I.; Bottomly, K. *Eur. J. Immunol.* **1991**, *21*, 17−22.
[10] Novak, T. J.; Farber, D.; Leitenberg, D.; Hong, S.-C.; Johnson, P.; Bottomly, K. *Immunity* **1994**, *1*, 109−119.
[11] Johnson, P. A.; Maiti, A.; Ng, D. H. W. In *Wier's Handbook of Experimental Immunology*, V. 2; Wier, D. M., Herzenberg, L. A., Blackwell, C., Eds.; Blackwell Science, Cambridge, MA, 1997; pp 62.1−62.16.
[12] Ostberg, J. R.; Barth, R. K.; Frelinger, J. G. *Immunol. Today* **1998**, *19*, 546−550.
[13] Rosenstein, Y.; Santana, A.; Pedraza-Alva, G. *Immunol. Res.* **1999**, *20*, 89−99.
[14] Fukuda, M. *Glycobiology* **1991**, *1*, 347−356.
[15] Fukuda, M.; Carlsson, S. R.; Klock, J. C.; Dell, A *J. Biol. Chem.* **1986**, *261*, 12796−12806.
[16] (a) Hakomori, S.-I. *Adv. Exp. Med. Biol.*, **2001**, *491*, 369−402. (b) Keissling, L. L.; Gestwicki, J. E.; Strong, L. E. *Curr. Opin. Chem. Biol.* **2000**, *4*, 696−703.
[17] Danishefsky, S. J.; Allen, J. R. *Angew. Chem., Int. Ed.* **2000**, *39*, 836−863.
[18] Sikut, R.; Nilsson, O.; Baeckstrom, D.; Hansson, G. C. *Biochem. Biophys. Res. Commun.* **1997**, *238*, 612−616.
[19] Zhang, S.; Walberg, L. A.; Ogata, S.; Itzkowitz, S. H.; Koganty, R. R.; Reddish, M.; Gandhi, S. S.; Longenecker, B. M.; Lloyd, K. O.; Livingston, P. O. *Cancer Res.* **1995**, *55*, 3364−3368.
[20] Cyster, G. C.; Shotten, D. M.; Williams, A. F. *EMBO J.* **1991**, *10*, 893−902.
[21] Shorgren, R.; Gerken, T. A.; Jentoft, N. *Biochemistry* **1989**, *28*, 5525−5536.
[22] Li, F.; Erickson, H. P.; James, J. A.; Moore, K. L.; Cummings, R. D.; McEver, R. P. *J. Biol. Chem.* **1996**, *271*, 6342−6348.
[23] Gerken, T. A.; Butenhof, K. J.; Shogren, R. *Biochemistry* **1989**, *28*, 5536−5543.
[24] Liang, R.; Andreotti, A. H.; Kahne, D. *J. Am. Chem. Soc.* **1995**, *117*, 10395−10396.
[25] Kirnarsky, L.; Prakash, O.; Vogen, S. M.; Nomoto, M.; Hollingsworth, M. A.; Sherman, S. *Biochemistry* **2000**, *39*, 12076−12082.
[26] Simanek, E. F.; Huang, D. H.; Pasternack, L.; Machajewski, T. D.; Seitz, O.; Millar, D. S.; Dyson, H. J.; Wong, C. H.; *J. Am. Chem. Soc.* **1998**, *120*, 11567−11575.
[27] Wu, W. G.; Pasternack, L.; Huang, D. H.; Koeller, K. M.; Lin, C. C.; Seitz, O.; Wong, C. H. *J. Am. Chem. Soc.* **1999**, *121*, 2409−2417.
[28] Lane, A. N.; Hays, L. M.; Feeney, R. E.; Crowe, L. M.; Crowe, J. H. *Protein Sci.* **1998**, *7*, 1555−1563.
[29] Bush, C. A.; Feeney, R. E. *Int. J. Pept. Protein Res.* **1986**, *28*, 386-397.
[30] Pavia, A. A.; Ferrari, B. *Int. J. Pept. Protein Res.* **1983**, *22*, 539−548.
[31] Pepe, G.; Siri, D.; Oddon, Y.; Pavia, A. A.; Reboul, J.-P. *Carbohydr. Res.* **1991**, *209*, 67−81.
[32] Liu, X.; Sejbal, J.; Kotovych, G.; Koganty, R. R.; Reddish, M. A.; Jackson, L.; Ganghi, S. S.; Mendonca, A. J.; Longenecker, B. M. *Glycoconjugate J.* **1995**, *12*, 607−617.
[33] Pieper, J.; Ott, K.-H.; Meyer, B. *Nat. Struct. Biol.* **1996**, *3*, 228−232.
[34] Klich, G.; Paulsen, H.; Meyer, B.; Meldal, M.; Bock, K. *Carbohydr. Res.* **1997**, *299*, 33−48.
[35] Braun, P.; Davies, G. M.; Price, M. R.; Williams, P. M.; Tendler, S. J. B.; Kunz, H. *Bioorg. Med. Chem.* **1998**, *6*, 1531−1545.
[36] Schuman, J.; Qui, D.; Koganty, R. R.; Longenecker, B. M.; Campbell, A. P. *Glycoconjugate J.* **2000**, *17*, 835−848.
[37] Schuster, O.; Klich, G.; Sinnwell, V.; Kranz, H.; Paulsen, H.; Meyer, B. *J. Biomol. NMR* **1999**, *14*, 33−45.
[38] Naganagowda, G. A.; Gururaja, T. L.; Satyanarayana, J.; Levine, M. J. *J. Pept. Res.* **1999**, *54*, 290−310.
[39] Live, D. H.; Williams, L. J.; Kuduk, S. D.; Schwarz, J. B.; Glunz, P. W.; Chen, X. T.; Sames, D.; Kumar, R. A.; Danishefsky, S. J. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 3489−3493.
[40] Andreotti, A. H.; Kahne, D. *J. Am. Chem. Soc.* **1993**, *115*, 3352−3353.
[41] Huang, X.; Barchi, J. J.; Lung, F. D. T.; Roller, P. P.; Nara, P. L.; Muschik, J.;, Garrity, R. R. *Biochemistry* **1997**, *36*, 10846−10856.
[42] Glunz, P. W.; Hintermann, S.; Schwarz, J. B.; Kuduk, S. D.; Chen, X. T.; Williams, L. J.; Sames, D.; Danishefsky, S. J.; Kudryashov, V.; Lloyd, K. O *J. Am. Chem. Soc.* **1999**, *121*, 10636−10637.
[43] Kudryashov, V.; Glunz, P. W.; Williams, L. J.; Hintermann, S.; Danishefsky, S. J.; Lloyd, K. O. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *98*, 3264−3269.
[44] Vichier-Guerre, S.; Lo-Man, R.; Bay, S.; Deriaud, E.; Nakada, H.; Leclerc, C.; Cantacuzene, D. *J. Pept. Res.* **2000**, *55*, 173−180.
[45] Kuduk, S. D.; Schwarz, J. B.; Chen, X. T.; Glunz, P. W.; Sames, D.; Ragupathi, G.; Livingston, P. O.; Danishefsky, S, J. *J. Am. Chem. Soc.* **1998**, *120*, 12474−12485.

Chart 1



1 **2** (β-TF) **3** (α-Tn) **4** (α-TF) **5** (α-STF)

6 (α-Le^y)

STTAV, which is found at the N-terminus of mature CD43. In each of the compounds that we have synthesized, the three side-chain hydroxyl substituents of the pentapeptide bear a particular carbohydrate motif. In a preliminary communication, we reported on the structure of a mucin peptide that displayed the STF antigen, a complex carbohydrate associated with certain carcinomas.[46] The number of NOE interactions and the large couplings between amide and α-protons observed indicated a stable and extended structure, which is unusual for linear peptides of this length.[39] Remarkably, investigations conducted on related constructs possessing α-linked TF and Tn antigens and β-linked TF antigen served to indicate, at least qualitatively, a clear structural distinction between the glycopeptides having the natural mucin α-linked carbohydrate and their unnatural β-linked anomeric forms.[39] This result spoke to possible through-space communication between the saccharide and peptidal domains.

The collection of molecules that we have synthesized provided a unique opportunity to conduct a detailed investigation into the information encoded by the mucin constitution and its value in establishing three-dimensional presentation. The data gathered clearly indicated that a common structure is exhibited by the α-O-linked series of glycopeptides (i.e., the mucin systems). In fact, for the STTAV peptide sequence, the α-linked clusters exhibit a distinct NOE fingerprint. This intriguing result suggested the possibility that the structural motif observed might conceivably be conserved in typical mucins. To test this possibility we thoroughly examined compounds **1**−**5**. Furthermore, we prepared and studied compound **6**, a mucin-based glycopeptide having the sequence SSSAVAV, in which each of the three side-chain hydroxyls bears the full Le^y blood group determinant. As discussed below, the NOESY spectrum includes an NOE fingerprint that reveals a clear and striking similarity to the α-linked STTAV systems. This finding indicated that the overall organization is similar to the STTAV constructs, despite size and variation of the glycans and sequence differences of the peptide core. We now report a refined high-resolution molecular model of the mucin motif which clearly demonstrates that nature has organized mucins in a conserved manner as rigid entities capable of displaying a variety of glycans with relatively little intrinsic change to their overall structure.

Results

Resonance Assignments. The proton signals within individual residues were assigned by COSY in D_2O and TOCSY and NOESY experiments in H_2O and D_2O.[47,48] The NOESY results from the amide regions of the spectra provided sequential assignments of the peptide backbone. $^1H-^{13}C$ HMQC[49] and HSQC[50,51] experiments aided these assignments through exploiting the known resonance positions of specific carbon sites and the greater dispersion associated with the ^{13}C nucleus to resolve ambiguities. The ^1H and ^{13}C correlations provided the basis for sequential assignments for the carbon signals. $^1H-^{13}C$ HMBC[52] data augmented the homonuclear through-bond coupling experi-

[46] Sames, D.; Chen, X. T.; Danishefsky, S. J. *Nature* **1997**, *389*, 587−591.

[47] Wuthrich, K. *NMR of Proteins and Nucleic Acids*; John Wiley and Sons: New York, 1986.

[48] Rance, M. *J. Magn. Reson.* **1987**, *74*, 557−564.

[49] Bax, A.; Griffey, R. H.; Hawkins, B. L. *J. Magn. Reson.* **1983**, *55*, 301−315.

[50] Bodenhausen, G.; Ruben, D. J. *Chem. Phys. Lett.* **1980**, *69*, 185−189.

[51] Palmer. A. G., III; Cavanagh, J.; Wright, P. E.; Rance, M. *J. Magn. Reson.* **1991**, *93*, 151−170.

[52] Bax, A.; Summers, M. F. *J. Am. Chem. Soc.* **1986**, *108*, 2093−2094.

Table 1. ^{1}H and ^{13}C Chemical Shifts (ppm) Relative to Random Coil Values for Peptide Backbone

	S1	T2	T3	A4	V5
			Compound 1		
Cα	0.01	−1.84	−0.07	0.07	0.81
Hα	0.03	0.11	0.02	0.05	−0.06
HN	0.06	0.19	−0.01	0.10	−0.27
			Compound 2		
Cα	−1.56	−1.41	−1.74	−0.05	1.52
Hα	0.09	0.14	0.12	0.07	−0.10
HN	0.08	0.07	0.03	−0.06	−0.25
			Compound 3		
Cα	−1.73	−2.02	−2.34	−0.98	0.96
Hα	0.25	0.41	0.18	0.03	−0.17
HN	0.19	0.55	0.61	0.14	−0.08
			Compound 4		
Cα	−1.48	−2.00	−2.26	−0.98	1.29
Hα	0.30	0.47	0.21	0.07	−0.15
HN	0.25	0.70	0.73	0.15	−0.13
			Compound 5		
Cα	−2.02	−2.02	−2.26	−1.12	1.57
Hα	0.35	0.47.	0.21	0.06	−0.21
HN	0.19	0.67	0.74	0.16	−0.25

ments and were particularly valuable in establishing the connectivity across the glycosidic linkage from the amino acid to the GalNAc residues. The sequence-specific assignment of the methyl groups of the *N*-acetyl moieties relied on the NOEs to the assigned respective amide protons. With the amide proton signals assigned, the origin of the ^{15}N signals could be readily determined from the $^{1}H-^{15}N$ HSQC correlations. All studies were carried out at natural abundance.

Chemical Shifts. Amide proton chemical shifts are sensitive to peptide organization and were used as a basis for initial qualitative comparisons of molecules **1**−**5** as shown in our initial report.[39] Table 1 presents quantitative data for the shifts of the amide and α-protons and α-carbons for molecules **1**−**5**. Shifts of these nuclei within amino acid residues depend on the conformational context, whether helical or extended, and the relative value of these compared to the respective sites of the same amino acid in a random coil environment has been calibrated.[53] The average shift contributions associated with the transition to a *β*-sheet extended structure from a random coil are 0.38 ppm for Hα, −1.4 ppm for Cα, and 0.29 ppm for amide proton. Those associated with an α-helix are of similar magnitude but opposite sign. As the shift data are sensitive to conformation, the close similarity of the values for the α-linked constructs **3**−**5** are indicative of the consistency of the peptide backbone structures among the constructs. The values presented in Table 1 for the α-linked series are in accord with those associated with extended *β*-sheet conformations,[53] although a contribution to the α-carbon shifts of the threonine and serine residues may arise from known substituent effects of glycosylation on the isolated amino acids.[54] Overall the values are distinct from those for the *β*-linked glycopeptide **2** and the peptide **1**. Apart from the relative shifts of the Cα sites on the glycosylated amino acids, the values for **2** are reminiscent of a peptide backbone in a random coil state. $^{1}H-^{15}N$ HSQC spectra of **1**, **2**, and **4** in water were recorded. ^{15}N shifts are a sensitive measure of peptide backbone interactions, and a sharp contrast exists between the ^{15}N shifts of the *β*-TF glycopeptide **2** and the α-TF glycopeptide **4**, with significant upfield shifts exhibited for the amide nitrogens and particularly for A4 of **4** (Figure 1B). There was very little difference between the ^{15}N shifts of peptide **1** and *β*-TF glycopeptide **2** (Figure 1A). These results tend to reinforce our earlier suggestion of a conformational distinction between the α-linked series on one hand and the *β*-linked example on the other.

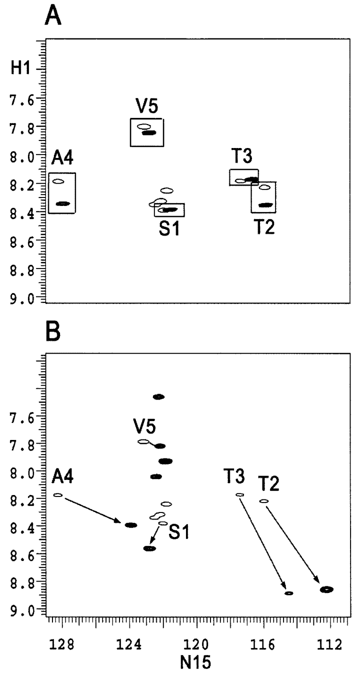


Figure 1. Overlay of $^{1}H-^{15}N$ HSQC spectra (A) of STTAV peptide (filled cross-peaks) and *β*-TF STTAV glycopeptide (open cross-peaks) and (B) of α-TF STTAV glycopeptide (filled cross-peaks) and *β*-TF STTAV glycopeptide (open cross-peaks). In panel A, boxes surround cross-peaks from the same residues of the respective molecules, and in (B) arrows show the change in shifts between respective sites on **2** and **4**. Spectra were acquired at 800 MHz (^{1}H) in 90% H_2O/10% D_2O. Cross-peaks whose assignments are not indicated arise from GalNAc amides.

CD Analysis. The similarity between the parent peptide **1** and the *β*-TF glycopeptide **2** and its contrast with the α-TF glycopeptide **4** is supported by the CD analysis of these molecules (Figure 2). Interestingly, the shapes of the curves for the first two systems are essentially identical, while that of the latter deviates substantially.

Coupling Constants. The peptide backbone $^{3}J_{HN-H\alpha}$ couplings were measured from resolved splittings in 1D spectra in H_2O (Table 2).[47] The consistent values for the respective individual residues in the α-linked group of glycopeptides **3**−**5** suggest a preferred extended structure. The relatively high values observed for the threonine residues make substantial torsional averaging unlikely. By contrast, in the *β*-linked analogue **2**, the backbone couplings are smaller. This difference could reflect a less extended structure or greater conformational flexibility. The results for glycopeptide **2** are similar to those for peptide **1** with the exception that the values at residues T2 and T3 are somewhat higher. This difference presumably reflects a local conformational effect. The $^{3}J_{HN-H2}$ couplings in the GalNAc residues are all ∼10 Hz, thus suggesting the torsion angle between H2 and NH protons to be ∼180°.[47] Accordingly, the orientation of the *N*-acetyl group relative to the sugar framework seems essentially fixed. The $^{3}J_{H\alpha-H\beta}$ were determined in order to provide insight into the χ_1 angles for side chains of threonine

[53] Wishart, D. S.; Sykes, B. D. *Methods Enzymol.* **1994**, *239*, 363−392.
[54] Dill, K.; Allerhand, A. *FEBS Lett.* **1979**, *107*, 26−29.

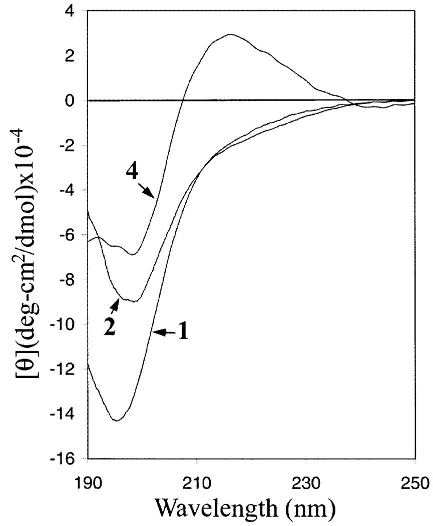


Figure 2. CD spectra of STTAV peptide **1**, α-TF STTAV glycopeptide **4**, and *β*-TF STTAV glycopeptide **2** in H_2O at room temperature.

Table 2. $^3J_{\mathrm{HN-H\alpha}}$ Coupling Constants (Hz)

cmpd	S1	T2	T3	A4	V5
1	6.83	8.33	8.23	6.51	8.97
2	6.35	7.57	7.33	6.35	8.55
3	6.77	9.00	8.97	6.96	7.70
4	6.71	9.16	9.28	7.21	8.06
5	7.03	8.79	9.23	7.42	8.34

and serine. The couplings of both threonine residues for **3**−**5** have a remarkably small variation, falling between 1.8 and 2.3 Hz. These low values preclude any significant torsional averaging and tend to limit torsion angles between the protons on Cα and C*β* to a narrow region near either 90° or 270° (corresponding to angles of −150° and 30° in the conventional definition of χ_1 angles).[55]

The couplings associated with the χ_1 angle of S1 have been determined for **5**, giving values of 3.3 and 5.2 Hz for the coupling between the α- and the two *β*-protons. The observation of a comparatively large chemical shift difference of ~0.2 ppm between the two serine *β*-protons, as well as the ^{13}C relaxation data,[56] argues against substantial conformational averaging about the central bond. In conjunction with the essentially equal intensity for the two Hα to H*β* NOEs in this residue, this finding suggests a fixed conformation with either an approximately 60° or 120° angle between the protons. Dihedral angle constraints based on these results for the serine and threonine residues were used in the final structure refinement calculations (vide infra).

For **2**, which contains the *β*-linked sugars, the side-chain couplings are significantly different from those observed for the α-linked series above. The threonine residues have values of 4.6 and 4.9 Hz, notably larger values than for **3**−**5**, and S1 has values of 5.5 and 8 Hz. For **2**, the difference in chemical shift between the two serine *β*-protons is reduced to ~0.07 ppm, compared to 0.2 ppm for the α-series. The results for **2** bear some similarity to what is found for the peptide **1**, in which the threonine couplings are 4.2 and 5.0 Hz. The multiplet patterns for the S1 α- and *β*-protons in the peptide can be fit to coupling values of 5.8 and 5.9 Hz and the shift difference between the two *β*-protons is only 0.035 ppm.

Table 3. NOE Interaction between GalNAc and Amino Acid Residues for STF Glycopeptide (**5**)[a]

sugar residue/ proton	peptide proton	intensity
Ser1-GalNAc		
Me	T3 Hα	w
	T3 H*β*	w
	T3 *γ*Me	m
NH	S1 NH	w
	S1 Hα	w
	S1 H*β*1	w
H1	S1 NH	w
	S1 *β*1	m
	S1 *β*2	m
H3	S1NH	m
T2 GalNAc		
H1	T3 NH	m
	A4 NH	w
	T2 *γ*Me	m
	T2 H*β*	m
NH	T2 NH	m
	T3 Hα	w
	T2 *γ*Me	w
H3	T2 *γ*Me	w
H4	T2 *γ*Me	w
	T2 NH	w
Me	A4 NH	m
	A4 H*β*	m
T3 GalNAc		
NH	T3 NH	m
	T3 H*β*	w
	T2 *γ*Me	w
	T3 *γ*Me	w
	T2 H*β*	w
	T3 Hα	w
H27	T3 NH	w
7Me	T3 NH	w
	A4 Hα	w
7H4	T3 NH	w
	T3 *γ*Me	w
H71	T3 *γ*Me	w
	T3 H*β*	w
	A4 NH	w
H73	T3 *γ*Me	w

[a] Four additional ambiguous NOEs between carbohydrate and peptide were used in the calculations, T3 NH to H3 of either T2 GalNAc or T3 GalNAc, T2 NH to H3 of either T2 GalNAc of T3 GalNAc, T2 NH to H1 of either S1 GalNAc or T2 GalNAc, and T2 GalNAc NH to T2 Hα or T3 GalNAc H1. The distance ranges for NOE constraints were 1.8−3.0, 1.8−4.0, and 1.8−5.0 Å for strong, medium, and weak, respectively.

NOEs. NOE interactions of the protons of the amino acids and the resonances of the proximal GalNAc residues, as well as several to the H1 protons of the galactose residues, were identified by conducting NOESY experiments at 600 and 800 MHz and employing 3D TOCSY−NOESY[57] homonuclear measurements at 600 MHz as well. The regions of the NOESY spectra involving interactions between amidic protons and those in the aliphatic region of the α-STF glycopeptide **5**, presented in Figure 3, illustrate the interaction of sugar and peptide components. The same pattern of NOEs is also seen for the α-Tn glycopeptide **3** and the α-TF glycopeptide **4** (data not shown). The NOE interactions observed in **2**, **4**, and **6** between the *N*-acetyl methyl groups and the peptide backbone NH groups are shown in Figure 4. The NOE data spanning several residues from the T2 GalNAc methyl to the A4 NH in **5** is of particular structural interest. The NOEs between the GalNAc moieties and the peptide residues of **5** are given in Table 3. Enumeration and analysis of the NOEs in the amide region of *β*-linked TF glycopeptide **2** was complicated by a comparative lack of

[55] Bystrov, V. F. *Prog. Nucl. Magn. Reson. Spectrosc.* **1976**, *10*, 41−81.
[56] Live, D. H., unpublished results.

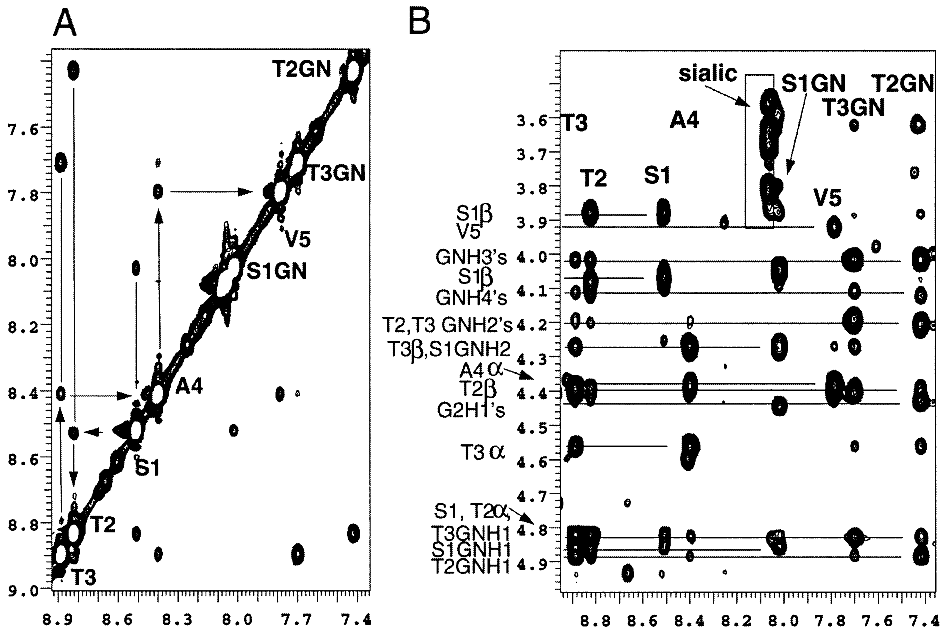


Figure 3. Sections of the 350-ms NOESY spectra (600 MHz) in 90% H_2O/10% D_2O, 18 °C, of α-STF STTAV **5** showing (A) amide−amide and (B) amide aliphatic cross-peaks. Arrows in (A) trace the sequential connectivities, with extensions showing NOEs to the NH protons on the GalNAc residues. In (B), the labels within the box refer to the NH assignments and those outside to the sites to which they have NOEs. GN refers to sites on the GalNAc residues attached to the amino acid residue indicated, G2 to the galactose residues.

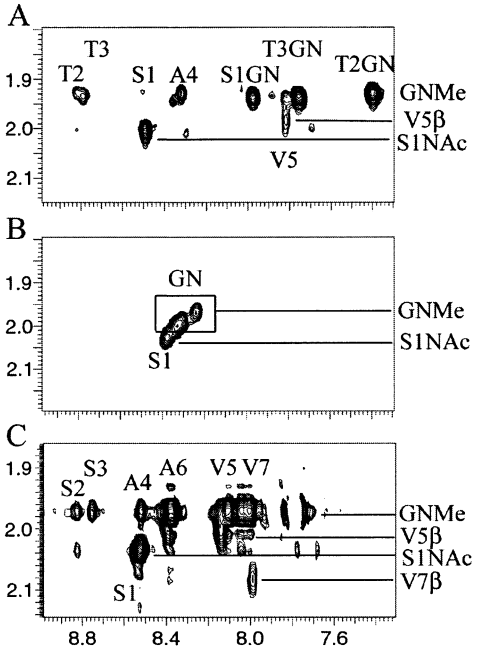


Figure 4. Sections of the NOESY spectra in 90% H_2O/10% D_2O showing interactions between amide protons and *N*-acetyl methyl protons for α-TF STTAV (**4**) (A), *β*-TF STTAV (**2**) (B), and α-Le^y-SSSAVAV (**6**) (C).

Table 4. Amide Proton Exchange Lifetimes (s)

cmpd	S1	T2	T3	A4	V5	G1	G2	G3	sial
1	38	44[a]	26	44[a]	932				
2	278	450	399[a]	399[a]	1348	550	994	1043	
3	235	357	405	833	1477[a]	1477[a]	12310	10540	
4	233	369	437	770	1855	1522	22240	14410	
5	406	931	1505	1812	1718	3230	44000	32350	189

[a] The individual intensities could not be accurately determined because of overlap. The components appeared qualitatively to decay at similar rates and gave reasonable fits to a single exponential.

dispersion. However, one dramatic difference between **2** and its α-linked analogue can be seen in the interactions involving the *N*-acetyl methyl groups of the GalNAc residue. In contrast to the α-linked case, (Figure 4A) no NOEs to the peptide backbone amides for **2** were observed (Figure 4B). This result is indicative of a much less compact structure then is the case with the α-linked peptides. Similar dispersion of amide protons and pattern of NOEs for **5** and the more complex **6** indicate a common structural motif.

Exchange and Temperature Coefficients of Amidic Protons. The amide H/D exchange rates[58] and the amide proton shift temperature coefficients were measured.[59] As seen in Table 4, introduction of an O-linked carbohydrate linkage causes significant extension in the amide proton lifetime regardless of the stereochemistry of the linkage. Comparison of the three α-linked glycopeptides clearly establishes that the amide proton lifetime increases with increasing size of the carbohydrate epitope. Two particularly noteworthy features are evident in comparing the α- to the *β*-linked glycopeptides. First, the lifetimes for the A4 and V5 amides in the α-linked group are significantly longer than in the *β*-linked construct, and second, lifetimes for the amide protons of the GalNAc residues on T2 and T3 in the α-linked cases are dramatically extended beyond those of the *β*-linked construct.

The temperature dependence of the amide proton chemical shifts are presented in Table 5. The results for the backbone of the peptide **1** and the *β*-TF glycopeptide **2** are very similar, whereas the α-linked glycopeptides are distinct from **1** and **2** in several ways. The amides of A4 and V5 and the T2 GalNAc

(57) Oschkinat, H.; Cieslar, C.; Holak, H. A.; Clore, G. M.; Gronenborn, A. M. *J. Magn. Reson.* **1989**, *83*, 450−472, with modifications incorporating gradients and water supression, Live, D. *J. Magn. Reson.*, submitted.

(58) Roeder, H. *Methods Enzmol.* **1989**, *176*, 446−473.

(59) Dyson, H. J.; Rance, M.; Houghton, R. A.; Lerner, R. A.; Wright, P. *J. Mol. Biol.* **1988**, *201*, 161−200.

Table 5. Amide Proton Shift Temperature Coefficients (ppb/°C)

cmpd	S1	T2	T3	A4	V5	G1	G2	G3	sial
1	−7.0	−7.1	−5.3	−7.1	−8.4				
2	−6.9	−7.4	−5.7	−6.2	−8.1	−7.4	−7.5	−7.4	
3	−6.0	−10.8	−8.9	−5.2	−5.6	−6.3	−4.0	−4.4	
4	−6.8	−11.6	−8.6	−5.9	−6.1	−7.7	−4.4	−4.8	
5	−6.0	−11.7	−8.9	−6.1	−6.4	−7.2	−5.2	−3.4	−5.8

Table 6. Statistical Analysis of NMR Restraints and Computed Structure for **5**

distance restraints	
total	107
intraresidue	
peptide	21
glycan	15
sequential ($\|i-j\| = 1$)	
peptide	20
glycan	9
peptide to glycan	
within the same glycosylated residue	29
between glycosylated residues	12
glycan to glycan	
between glycans	1
3-bond J-coupling restraints	
peptide	5
glycans	6
dihedral restraints	
peptide	3

structural statistics	
NOE violations	
number >0.15 Å	0
3 bond J-coupling violations	
number >0.5 Hz	0
dihedral violations	
number >5^o	0
deviations from ideal geometry (peptide and S1GalNAc, T2GalNAc, and T3GalNAc)	
bond length (Å)	0.0113 ± 0.0003
bond angle (deg)	1.226 ± 0.035
impropers (deg)	0.578 ± 0.024
average pairwise rmsd among 59 final structures (Å)	
peptide backbone+ (S1GalNAc, T2GalNAc, T3GalNAc) heavy atoms	0.81 ± 0.25
peptide heavy atoms and (S1GalNAc, T2GalNAc, T3GalNAc) heavy atoms	1.26 ± 0.31

and T3 GalNAc amides in the α-series show reduced sensitivity to temperature, consistent with shielding of these sites from the solvent or involvement in intramolecular interactions. The two glycosylated threonine residues (T2 and T3), however, show an enhanced temperature sensitivity. This observation is consistent with a recent report involving a peptide having two sequential threonine residues, in which introduction of an α-O-linked carbohydrate on one of these gave rise to an enhancement in the temperature dependence of the backbone amide proton shift.[38] Furthermore, this dependence was even greater when both successive residues were glycosylated.[38] In our case, we did not observe a differential effect related to glycosylation of serine on the temperature dependence of its amide proton resonances.

Structure Calculations and Refinement. Using the NMR data described here, structural calculations were carried out on the α-STF glycopeptide **5** using the X-PLOR program.[60] A total of 107 NOE constraints were employed, originating mostly from protons on the amino acid residues and the proximal GalNAc residues, with a few identified involving the anomeric proton on the Gal residues.[61] Values of the eight amide proton couplings were also incorporated as constraints. As mentioned above, the values of $^3J_{\mathrm{HN-H\alpha}}$ for T2, and T3 are large and substantially limit the associated ϕ angle to a region of ∼−120°.[47] The value for A4 is lower and has possible solutions at angles of about −160°, −90°, and +60°. The angles of the amides in the GalNAc residues are uniquely established based on the value of the couplings.

An initial set of 100 structures was calculated using these constraints. Of these, the 19 best results were identified using, as criteria, a lack of distance constraint violations greater than 0.5 Å and a lack of coupling constraint violations greater than 1 Hz. Significantly, in the absence the χ_1 restraints, the angles for S1, T2, and T3 were all highly clustered in the vicinity of 60°. These results are consistent with one of the two possible solutions derived from the experimental coupling measurements and, remarkably, indicate that the structure is virtually static. The consistency of χ_1 results with experimental data, even though not included at this level in the calculations, also provides a measure of validation for the NOE and backbone amide coupling constraints. A round of calculations was subsequently carried out from the same starting point as above, but with the information from the first round used to discriminate which of the possible χ_1 dihedral constraints to use. The S1 χ_1 angle was then constrained to a range of 60° ± 25°, and T2 and T3 to 35° ± 25°. An ensemble of 200 structures was calculated. Fifty-nine of these structures had no distance violations greater than 0.15 Å, no coupling violations greater than 0.5 Hz, and no dihedral violations greater than 5°. As an additional check on the selection of angles, calculations were carried out with all possible permutations of the experimentally allowed χ_1 values for T2 and T3. The structures based on the other choices resulted in substantially higher levels of constraint violations. The structure statistics are given in Table 6. The list of NOE and dihedral constraints, and coordinates of the 59 best structures, have been deposited in the Protein Data Bank and are available from their web site under the identifier 1KYJ. A superposition of the heavy atoms of the peptide backbone and first sugar for the 59 best structures is shown in Figure 5. The full structure of **5** closest to the average of the 59 is shown in Figure 6. To provide a better sense of the three-dimensional organization, a

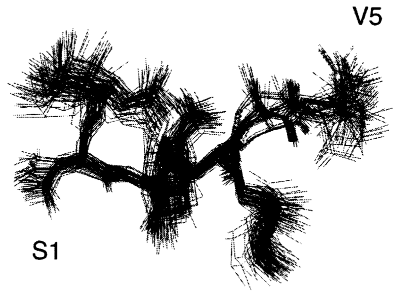


Figure 5. Superposition of the peptide backbone and first sugar residue of the 59 best calculated structures of **5** from the restrained dynamics calculations.

[60] Brunger, A. T. *X-PLOR Version 3.1, A System for X-ray Crystallography and NMR*; Yale University Press: New Haven. CT. 1992.

[61] In the earlier report,[39] there were only 98 unique NOEs. The indication of 116 arose from an inadvertent bookkeeping error where several of the symmetric NOE cross-peaks were counted on both sides of the diagonal.

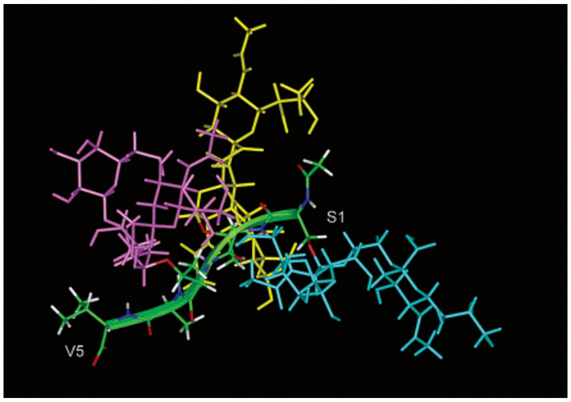


Figure 6. Structure closest to the average of the 59 best calculated models for tri-α-STF−STTAV glycopeptide (**5**). The carbohydrate epitopes are color coded magenta for the one on T3, yellow for the one on T2, and turquoise for the one on S1.

stereopair of the common inner glycopeptide core of **3−5,** formed by the peptide and the α-GalNAc residues, is illustrated in Figure 7.

Several features of this model are noteworthy. Briefly, the distribution of the ϕ and ψ angles of the internal T2, T3, and A4 residues for the 59 best structures shows strong clustering of these angles for T2 around −100/−155 and for T3 around −100/+150. About two-thirds of the A4 residues are in two roughly equally populated groups with ϕ/ψ coordinates of −85/−145 and −155/+115, respectively. The ϕ/ψ loci of A4 in the remaining third of the structural models are in two groups of equal size with ϕ/ψ coordinates of −85/+100 and −155/+150. The position of A4 as the penultimate C-terminal residue and the limited number of constraints involving the terminal valine may contribute to reduced definition of the backbone in that region. The superposition of structures illustrates that this variation is significant only outside the cluster (i.e., the alanine residue and beyond) and has only a modest effect on the position of the A4 β-methyl group. The backbone ϕ and ψ angles for the internal amino acid residues fall in ranges associated with β-sheet structures, with those for T3 being most favored and those for T2 and A4 in allowed but less favored regions.[62] The β-character of the peptide is consistent with the chemical shift trends mentioned above. Importantly, the shift indices, which point to β-strand type organization, were not employed in the refinement, providing further independent validation of the results. The family of structures have an average pairwise rmsd for the heavy atoms of the peptide backbone and the first sugar residue of only 0.81 Å with the values ranging from 0.25 to 1.6 Å, supporting a surprisingly highly defined structure for the molecule.

Discussion

The external cell surface milieu is dominated by a variety of carbohydrate-containing molecules. Notable among these are mucin glycopeptides, which display clustered polyvalent carbohydrates connected to serine and threonine residues via a conserved α-O-linked GalNAc moiety. Mucins have been implicated in a number of important biological processes, including protection of mucosal surfaces, cellular recognition and adhesion, signal transduction, and oncogenic transformation.[1,2] Although the molecular architecture of mucins is highly complex in nature, we have surmounted the synthetic challenges associated with their construction,[17] thus enabling the preparation of a number of glycopeptides possessing the authentic mucin framework. Consequently, our totally synthetic mucin segments have been obtained in homogeneously pure form and in sufficient quantity to allow for both detailed immunological and high-resolution structural analyses. Immunological investigations have culminated in advancement of some of our constructs to human clinical trials as conjugate cancer vaccines.[17]

In the present investigation, we provide a comprehensive and detailed structural analysis showing that the clustering of GalNAc-based glycans via an α (axial) linkage induces a remarkably stable and extended structure within mucin-based glycopeptides. Attachment of the first α-O-linked GalNAc residue has a profound effect on the overall conformation of its

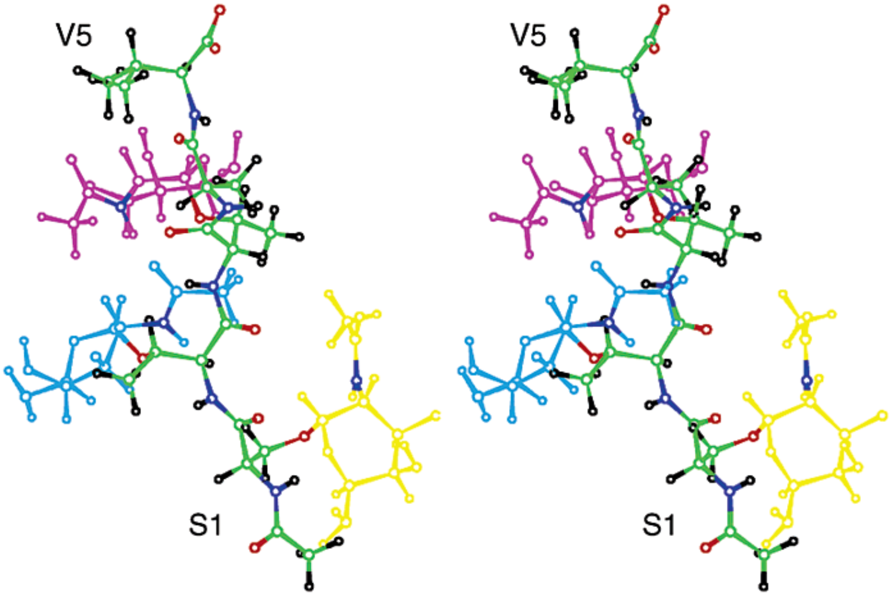


Figure 7. Stereopair of the glycopeptide core composed of the peptide and the attached GalNAc residue for the structure closest to the average of the tri-α-STF−STTAV (**5**) models. S1GalNAc is in yellow, T2GalNAc is in turquoise, and T3GalNAc is in magenta, and the nitrogen atoms of the *N*-acetyl groups are in blue.

peptide backbone, in which specific nonbonding interactions occurring between the peptide and the carbohydrate moiety play a key role. Several glycopeptides possessing the native mucin-like connectivity (**3**−**6**) were analyzed in our study. These systems were compared to one another, as well as to a similar glycopeptide (**2**) having a non-native β-linkage, and to the nonglycosylated peptide (**1**). A particularly intriguing feature was revealed on examination of the NOE data; *across the spectrum of the* α*-O-linked glycopeptides, including the very large construct* ***6****, a distinct NOE fingerprint was evident, clearly indicating that a high level of molecular organization exists within the mucin-like structures examined. Remarkably, this effect appears to be independent of the extent and nature of glycosylation beyond the initial* α*-O*−*GalNAc residue.* A large number of constraints were revealed from our analysis of these systems, and these were used as the basis for a computed three-dimensional model. Interestingly, the number of constraints and level of resolution achieved for the model is reminiscent of a stable folded globular protein rather than what would be expected for a relatively short linear glycopeptide.

It is apparent that the mode of carbohydrate attachment (via α- or β-glycosidic bonds) within a clustered locus has a profound effect on the conformation of the peptide. The α-linked systems exhibit an extraordinary level of structural organization for molecules of the size examined, particularly in light of their linear nature. In sharp contrast, the β-linked system **2**, which is trigylcosylated with the TF disaccharide, exhibited spectral features and physical properties comparable to the largely unstructured, nonglycosylated peptide **1**. This similarity was most clearly revealed by the substantial uniformity of the ^{1}H and ^{15}N shifts of the two systems (Figure 1A) but was also evident from comparison of the backbone and side-chain couplings, as well as the amide proton H/D exchange rates and the temperature dependence of the amide proton chemical shifts. Additionally, whereas a distinct NOE fingerprint was apparent in each of the α-linked systems (**3**−**6**) examined, no such distinguishing pattern was apparent for either the β-linked system (**2**) or the parent peptide (**1**).

Several structural studies on β-linked glycopeptides with single sites of glycosylation have been reported.[26,27] The pattern of NOEs determined in those simplified cases are consistent with what we have observed in the more elaborate clustered β-linked glycopeptide **2**. In general, β-linked systems appear to resemble the parent peptides and exhibit considerable flexibility in the peptide backbone, unlike their α-linked analogues. The β-linked derivative **2** is dynamic and relatively unrestrained, whereas α-linked clusters are confined to few conformations.

Ordinarily, in an extended molecule, it is difficult to acquire the level insight into its three-dimensional structure as has been obtained for the α-linked series of compounds examined here. This results primarily from the fact that the range of NOE interactions is limited to nearest neighbors. For this series of molecules, however, several other factors, such as the definition of the glycosylated side-chain orientations from the *J*-coupling values, as well as numerous GalNAc to peptide NOEs (including those to *next* nearest neighbors), provided additional constraints leading to the calculation of a well-defined structure. Locally, the influence of the exo-anomeric effect is likely to limit the conformational options at the glycosidic linkage of the α-GalNAc to the amino acid, and the bulk of the attached groups further limits the accessible conformational space.[63] However, these factors alone would not necessarily give rise to such a specific and stable backbone structure. With such a thoroughly resolved structure now in hand, it is well to consider in detail the specific interactions contributing to the high level of organization that exists within the glycopeptide core.

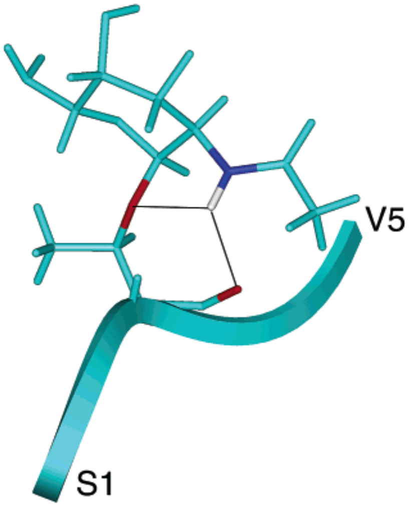


Figure 8. Illustration of the hydrogen-bonding-type interactions of *N*-acetyl amide proton of the T3 GalNAc, shown by black lines. These are typical for all the glycosylated residues of α-STF STTAV (**5**). The relevant oxygens are highlighted in red and the nitrogen in blue. The backbone trace of the rest of the peptide portion is shown as a ribbon.

The first of these factors is hydrogen bonding. Examination of our model reveals that, while the carbonyl groups of the amides of the GalNAc are oriented away from the peptide backbone, the NH bonds approach and participate in hydrogen-bonding interactions with the peptide. The average distances and angles measured over the collection of 59 best structures between the nitrogens of these amides and possible carbonyl hydrogen-bonding partners of the peptide backbone are 3.5 Å for the S1 GalNAc NH to T2 carbonyl, with an N−H−O angle of ∼96°, 4.1 Å for the S1 GalNAc NH to S1 carbonyl, with an angle of ∼152°, 3.7 Å for the T2 GalNAc NH toT2 carbonyl, with an angle of ∼122°, and 3.2 Å for the T3 GalNAc NH to T3 carbonyl, with an angle of ∼114°. These interactions are indicative of weakly stabilizing hydrogen bonds.[64] The proximity of the GalNAc amide proton to the glycosidic oxygen and the angle made by that atom, the proton, and the relevant carbonyl acceptor is ∼90°.[64] This arrangement suggests that the GalNAc amide protons are participants in bifurcated hydrogen bonds (Figure 8), although the N−H−O angle is more acute than is usually found for such interactions. While it is difficult to assert with absolute certainty whether these interactions are more appropriately viewed as electrostatic, or should be seen as conventional hydrogen bonds, the quantitative analysis of the structural geometry presented above indicates stabilizing interactions. The presence of these interactions is supported independently by the substantial increase in the amide proton

[62] Assessed with PROCHECK: Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. *J. Appl. Crystallogr.* **1993**, *26*, 283−291.

[63] Magnusson, G.; Chernyak, A. Y.; Kihlberg, J.; Kononov, L. O. In *Neoglycoconjugates*; Lee, Y. C., Lee, T. R., Eds.; Academic Press: New York, 1994; pp 53−144.

[64] Baker, E. N.; Hubbard, R. E. *Prog. Biophys. Mol. Biol.* **1984**, *44*, 97−179.

lifetimes (Table 4) and the reduction in shift temperature dependencies (Table 5) noted for these GalNAc amide protons.

Prior to the availability of high-resolution experimental data pertaining to a mucin motif, a computational approach was applied to assess the structure and possible interactions that stabilize the extended character of mucin glycoproteins.[65] A sequence containing a diad of two glycosylated threonines surrounded by alanine residues was used as a model. Hydrogen bonding between the GalNAc residues and the peptide backbone was identified as a major factor from the computations, but the favored arrangement identified by this method was between the GalNAc amide proton and the oxygen atom of the carbonyl group of the adjacent amino acid residue. This is in contrast to our experimentally determined structure, which shows hydrogen bonding between the GalNAc amide proton and the backbone carbonyl of its own residue. Computations conducted during the previous work also suggest that the latter arrangement gives rise to a less stable hydrogen bond. It should be noted, however, that the χ_1 angles for the threonines derived from the experimentally measured $^3J_{H\alpha H\beta}$ couplings in our study indicate a smaller torsion angle than what emerged from the purely computational model. These couplings placed an experimental constraint on the disposition of the GalNAc residue. This factor, along with the sequence variation between the glycopeptide in the earlier model, and the construct we investigated, may well contribute to the different conclusions arrived at regarding the nature of the stabilizing hydrogen-bonding interaction between the glycan and peptide.

The structure appears to be further stabilized by the close association of several methyl groups (Figure 9A). Over the 59 structures surveyed, the T2 GalNAc methyl carbon is ~3.6 Å removed from the A4 β, and the T3 γ carbon comes to within less than 4.1 Å of the S1 GalNAc methyl carbon. This pattern of medium-range interactions, where the GalNAc residues loop over in such a way that their *N*-acetyl methyl groups interact with methyl groups on the amino acid residues two residues down the chain, could well provide adequate shielding of hydrophobic surfaces. This type of organization also provides a mechanism by which the effects of glycosylation could potentially propagate beyond the immediate carbohydrate residue. The covalent linkage anchors the GalNAc residue to the peptide backbone, and the noncovalent interactions buttress and extend the peptide so as to present the distal glycans in an ordered way. Viewed from this perspective, the overall structural role of the α-O-linked GalNAc residue is functionally reminiscent of a partner strand in a β-sheet.

The inner glycopeptide core of the α-linked clusters appears remarkably well ordered and stable in dynamic terms for a peptide-related construct of its size. As pointed out above, some of the coupling constant values are inconsistent with extensive conformational averaging, and the large induced shifts argue against this as well. NMR relaxation measurements of T_1, T_2, $T_{1\rho}$, and NOE values for the Cα, Cβ, and anomeric carbons for gylcopeptide **3** (excluding the Cβ of the alanine since it is a methyl group), provide additional experimental support for concluding that conformational rigidity persists through residues 1−4.[56] Such conformational stability implies that the structure is energetically favorable. It also suggests that the observed structure might be maintained in longer mucin sequences and, perhaps, in the presence of flanking peptide sequences as well. Support for such an assertion is found in the initial analysis of the more complex construct **6**, with α-O-linked Le^y glycosylated SSSAVAV. In particular, this is revealed by comparison of panels A and C of Figure 6 that illustrate the retention of the sequence-specific amide shifts and NOE patterns between molecules **4** and **6**, including A4 amide to *N*-acetyl methyl and A4 methyl to *N*-acetyl methyl NOEs. Thus, despite the variation in peptide sequence and antennary glycan, the organization of the peptide backbone in the glycosylated region of **6** is consistent with the common features of the glycopeptide core **3**−**5**.

Beyond the first GalNAc residues in the α-linked glycopeptides **4**−**6**, the more distal carbohydrate residues cannot be distinguished spectroscopically from the proximal one, since the chemical shifts of the respective nuclei of the former seem to be degenerate. Identification of interresidue NOEs for these residues is similarly complicated, although this may be due to peak overlap and limited dispersion of signals within this region. Based on these observations, one can conclude that the antennary

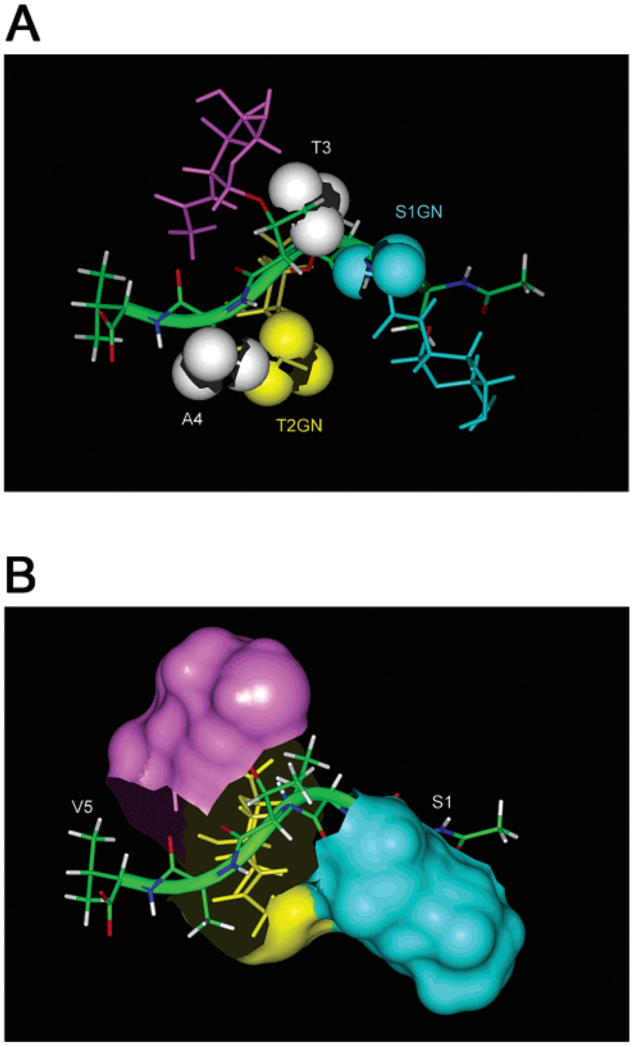


Figure 9. Glycopeptide core composed of the peptide and the attached GalNAc residue for the structure closest to the average of the tri-α-STF−STTAV (**5**) models, showing in (A), the van der Waals surfaces for the methyl protons of S1 GalNAc (turquoise), T2 GalNAc (yellow), and T3γ methyl and Ala β methyl (white) and, in (B), the Connolly surfaces of the three GalNAc residues, S1GalNAc (turquoise), T2 GalNAc (yellow), and T3 GalNAc (magenta).

[65] Butenhof, K. J.; Gerken, T. A. *Biochemistry* **1993**, *32*, 2650−2663.

glycans are in essentially identical environments, possibly static, but likely with greater conformational averaging than the core.

As indicated above, the stereochemical constraints imparted by the α-linkage place the first carbohydrate residue in close proximity to the peptide backbone. This assertion is supported by the number of NOEs between this sugar and the peptide (Table 3). Moreover, the dispersion of amide protons in the α-linked series is comparable to that observed for organized protein secondary structure. Thus, in a mucin domain, a compact structural core emerges that is sustained even when the distal glycans are removed. This fact is convincingly demonstrated by the spectroscopic data for the α-linked series of molecules **3**−**5**. Even when the second and third distal residues of the STF cluster **5** are deleted, the remaining Tn (GalNAc) antigen is seen to substantially shroud the peptide in the immediate vicinity of the glycosylation site (Figure 9B). Direct comparison of α- and β-linked clusters strongly suggests that the conformational properties are not simply due to a nonspecific phenomenon but rather appear to be a *specific* consequence of the stereochemistry at the glycosidic linkage joining the carbohydrate domain with the peptide domain. Thus, the unusual stability of the core structure is a consequence of the specific steric, hydrogen-bonding and nonbonded interactions associated with the attachment of the first α-linked GalNAc residue.

The results obtained from our studies on mucin glycopeptides have provided a high-resolution model that indicates the formation of extended structures and the intramolecular interactions from which these arise. Comparison can be made to the limited number of NMR-based structures of clustered glycopeptides that have been reported, but only at a somewhat qualitative level, since no atomic coordinates for these structures are available. One example is a glycopeptide segment from glycophorin A,[37] HT*S*T*S*S*S*VTL, where the asterisks indicate the six sequential sites of α-O-linked glycosylation to GalNAc residues. The structure deduced was based on 52 constraints, somewhat fewer than the number employed in our calculations. Backbone J couplings are high and reminiscent of those for **3**−**5**. The structure presented is an extended arrangement reminiscent of what we have found. The disposition of the carbohydrate moieties is somewhat different, however, and they are shown as projecting from alternate sides sequentially down the peptide backbone. The $^3J_{\mathrm{H}\alpha\mathrm{H}\beta}$ for the threonines are higher than for **3**−**5**, indicating different χ_1 angles, which may be related to this particular carbohydrate arrangement. Structures of glycopeptides from MUC7 have also been described.[38] These sequences are relatively rich in proline, suggesting that the backbone might adopt a polyproline type II structure, which is indeed supported by CD data. Computer models for the MUC7 peptides containing glycosylated threonine residues, built on the basis of this motif, were minimized by including NMR-derived constraints. In this analysis, the peptide backbone was found to maintain the polyproline type II motif, although the experimental backbone couplings near the sites of glycosylation suggest more negative ϕ angles than are associated with the canonical polyproline motif derived from crystallographic data of collagen-related peptides.[66] These compounds also showed evidence of NOE interactions between the GalNAc methyl and the β-methyl of an $i + 2$ alanine residue, which is similar to what is reported in our study. This suggests that the stabilizing methyl interactions we describe may have been important in this case as well. The comparatively limited number of NOEs observed in the MUC7 study may be a reflection of the lower field at which the data were collected and the size of the molecules in question. We have observed that NOEs are more numerous for such glycopeptides in spectra at 600 and 800 MHz than at 500 MHz. NMR data have been presented for a second MUC7 peptide with three consecutive serine residues, although no structure was presented.[38] The disposition of the amide proton shifts is quite similar to what we observe for the three glycosylated serine residues in **6**, arguing in favor of the persistence of the kind of structural motif we report. Both the glycophorin and MUC7 peptides are comparatively rigid as well. The characteristics shared by these glycopeptides and the constructs we have studied point to the generality of the organizational aspects of the peptide backbone we have identified. Greater variation between these systems likely results from the way in which the carbohydrate is presented by the peptide scaffold, which is expected to depend on the density of glycosylation. Further investigations would be required to characterize the scope of structural diversity accessible to mucins, en route to a conformational definition of a full mucin domain.

One of the means that nature has devised for cell−cell signaling and for the display of cell surface markers is the presentation of carbohydrates in a mucin motif. Mucin domains are composed of clustered glycosylated sites with the carbohydrate appended through an α-O-linked GalNAc residue, which may be further elaborated, depending on the physiological state of the cell. The information that the cell presents in this way is controlled by the selective expression of members of its glycotransferase repertoire.[16] Thus, cell surface signaling events can be effected by glycoproteins with the same core structure. Through the study of glycopeptides based on the N-terminal region of the cell surface glycoprotein CD43, insights into the structural basis of this phenomenon have been obtained and a quantitative model of mucin organization has been advanced. Importantly, it has been possible to illuminate how nature, through choice of glycan and stereochemistry of glycosylation, has devised an economical and effective strategy for meeting two key conditions for cell signaling. The first is the maximization of carbohydrate exposure through enforcement of an extended structure, thus preventing the carbohydrate surface from being obscured by protein collapse. The second is the presentation of a locally high concentration of carbohydrate molecules, which is often a decisive factor in cell signaling and cell−cell communication. Both of these conditions are admirably met by the admittedly abbreviated mucin motifs demonstrated herein.

Experimental Section

Synthesis of the glycopeptides used in this study has been reported.[17] The STTAV peptide was prepared in the Microchemical Facility of Sloan-Kettering Institute. NMR spectra were obtained on Varian INOVA 600 and 800-MHz instruments. The samples were dissolved in H_2O or D_2O with 10 mM phosphate buffer at a pH of ∼4.5 and had concentrations in the range of 5−20 mM. Studies were done at 18 °C except for the determination of the temperature dependence of the amide shifts. ^{1}H 1D and 2D TOCSY and NOESY experiments in H_2O were

[66] Kramer, R. J.; Bells, J.; Brodsky B.; Berman, H. M. *J. Mol. Biol.* **2001**, *311*, 131−147.

recorded using a WATERGATE[67] solvent suppression scheme. 2D $^1H-^{15}N$ HSQC data were collected using coherence selected field gradient-based experiments with sensitivity enhancement available from the Varian ProteinPack distribution. 1D, double quantum filtered COSY, TOCSY, NOESY, and $^1H-^{13}C$ HMQC and HMBC spectra in D_2O were obtained using pulse sequences in the Varian standard software library. The 3D TOCSY−NOESY experiment used WATERGATE solvent suppression[57] and was carried out in H_2O, with a pulse sequence written by D.H.L. The TOCSY mixing time was 45 ms, and the NOESY mixing times were between 350 and 400 ms. 1D and 2D data were processed using the Varian software, and the 3D data were processed with nmrPipe software.[68] ^1H and ^{13}C chemical shifts are given relative to DSS, and ^{15}N are given relative to NH_3.

The temperature dependence of the amide chemical shifts was measured from peak positions in 1D spectra over a range of 3−25 °C, with the data being fit to a line by the linear regression utility in the Vnmr software. The amide proton lifetimes were determined in several experiments covering different time regimes to allow for the substantial variation in lifetimes. Signal intensity data were fit to a single decaying exponential using the Vnmr regression analysis package.

CD spectra were obtained at room temperature on a Jasco J 710 instrument on solutions diluted from NMR samples. Concentrations were determined from quantitative amino acid analysis.

Cross-peak intensities from NOESY spectra were classified as strong, medium, or weak, with distance ranges of 2.4 ± 0.6, 2.9 ± 1.1, and 3.4 ± 1.6 Å, respectively. PARALLHDG.PRO (Version 4.02, M. Nilges) was used in assigning parameters for the pentapeptide. For the glycans, ideal geometry values were taken from CHARMM19 (MSI) with the bond, angle, and improper force constants set to match those in PARALLHDG.PRO. The electrostatic term was turned off, and van der Waals terms were used for the nonbonded interactions. The initial extended structure generated by INSIGHTII/BIOPOLYMER was initially minimized to idealize the covalent geometry. Peptide backbone and sugar linkage torsions were randomized for generating starting conformations. Structure refinement was carried out with distance, three-bond *J* coupling, and dihedral angle restrained torsion angle dynamics in X-PLOR Version 98.0[69] with an optimized version of the TAD protocol.[70] The conformations of the sugar rings were held fixed in the chair conformation. The simulated annealing consisted of 30 ps of dynamics computed with a time step of 15 fs at 50 000 K followed by cooling to 0 K in 30 ps and 2000 steps of energy minimization.

Acknowledgment. This paper is dedicated to Professor Bruce Merrifield on the occasion of his 80th birthday. The NMR data were collected at the Structural Biology NMR Resource of the University of Minnesota, supported by grants from the Minnesota Medical Foundation and NSF BIR-961477. Computation and visualization was carried out on facilities of the Minnesota Supercomputing Institute. This work was supported by the National Institute of Health (N.I.H.) (Grants: AI-16943, CA-28824 (S.J.D.)). Postdoctoral fellowship support is gratefully acknowledged by D.M.C. (Natural Science and Engineering Research Council of Canada, PDF-230654-2000; and Alberta Heritage Foundation for Medical Research, 199901330), L.J.W. (N.I.H., F32CA79120), P.W.G. (American Cancer Society, PF 98026), D.S. (The Irving Institute and M. R. Bloomberg), S.D.K. (U.S. Army Breast Cancer Grant, DMAD 17-98-1-8154), and J.B.S. (N.I.H., F3218804). We thank Dr. George Sukenick of the Sloan-Kettering Institute Core Facility for mass spectral analyses. We thank Judit Puche of the Chemistry Department at the University of Minnesota for providing amino acid analyses. The receipt of a Pfizer Award to S.J.D. for Creative Work in Organic Synthesis is gratefully acknowledged.

Supporting Information Available: Four regions of the $^1H-^{13}C$ HMQC spectrum of **5**. This material is available free of charge via the Internet at http://pubs.acs.org. See any current masthead page for ordering information and Web access instructions.

JA020208F

[67] Piotto, M.; Saudek, V.; Sklenar, V. *J. Biomol. NMR* **1992**, *2*, 661−665.
[68] Delaglio, F.; Grzesiek, S.; Vuister, G.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277−293.
[69] Stein, E. G.; Rice, L. M.; Brunger, A. T. *J. Magn. Reson.* **1995**, *124*, 154−164.
[70] Badger, J.; Kumar, R. A.; Yip, P.; Szalma, S. *Proteins* **1999**, *35*, 25−33.