Antonini, E., & Brunori, M. (1971) in *Hemoglobin and Myoglobin in Their Reactions with Ligands* (Neuberger, A., & Tatum, E. L., Eds.) North-Holland Publications, Amsterdam, Holland.

Bennett, J. C. (1967) *Methods Enzymol. 11*, 330–339.

Bunn, H. F., & McDonough, M. (1974) *Biochemistry 13*, 988–993.

Bunn, H. F., & Forget, B. G. (1986) *Hemoglobin: Molecular, Genetic and Clinical Aspects*, pp 381–451, W. B. Saunders Co., Philadelphia, PA.

Carrell, R. W., & Kay, R. (1972) *Br. J. Haematol. 23*, 615–619.

Clegg, J. B., Naughton, M. A., & Weatherall, J. D. (1966) *J. Mol. Biol. 19*, 91–108.

Edelstein, S. J., Rehmar, M. J., Olsen, J. S., & Gibson, Q. H. (1970) *J. Biol. Chem. 245*, 4372–4381.

Fermi, G., & Perutz, M. F. (1981) *Atlas of Molecular Structures in Biology: Haemoglobin and Myoglobin*, Clarendon Press, Oxford, England.

Gibson, Q. H. (1959) *Prog. Biophys. Chem. 9*, 1–53.

Gray, R. D. (1974) *J. Biol. Chem. 249*, 2879–2885.

Harano, T., Harano, K., Shibata, S., Ueda, S., Imai, K., & Seki, M. (1983) *Hemoglobin 7*, 85–90.

Ip, S. H. C., Johnson, M. L., & Ackers, G. K. (1976) *Biochemistry 15*, 654–660.

Kazim, L., & Atassi, M. Z. (1981) *Biochem. J. 197*, 507–510.

Kellett, G. L., & Gutfreund, H. (1970) *Nature (London) 227*, 921–926.

Kilmartin, J. V., & Hewitt, J. A. (1971) *Cold Spring Harbor Symp. Quant. Biol. 36*, 311–314.

Kilmartin, J. V., Hewitt, J. A., & Wootton, J. (1975) *J. Mol. Biol. 93*, 203–218.

Kleihauer, E. F., Reynolds, C. A., Doxy, A. M., Wilson, J. B., Moores, R. R., Berenson, M. P., Wright, C. S., & Huisman, T. H. J. (1968) *Biochim. Biophys. Acta 154*, 220–222.

Landon, M. (1970) *Methods Enzymol. 47*, 145–149.

McDonald, M. J., Lund, D. W., Bleichman, M., Bunn, H. F., DeYoung, A., Noble, R. W., Foster, B., & Arnone, A. (1980) *J. Mol. Biol. 140*, 357–375.

McDonald, M. J., Turci, S. M., Bleichman, M., & Stinson, R. A. (1985) *J. Mol. Biol. 183*, 105–112.

Moo-Penn, W. F., Jue, D. L., Johnson, M. H., McDonald, M. J., Turci, S. M., Shih, T.-B., Jones, R. T., Therrell, B. L., & Arnone, A. (1984) *J. Mol. Biol. 180*, 1119–1140.

Moo-Penn, W. F., Jue, D. L., Johnson, M. H., Olsen, K. W., Shih, D., Jones, R. T., Lux, S. E., Rogers, P., & Arnone, A. (1988) *Biochemistry 27*, 7614–7619.

Park, C. M. (1973) *Ann. N.Y. Acad. Sci. 209*, 237–257.

Perutz, M. F. (1987) in *The Molecular Basis of Blood Diseases* (Stamatoyannopoulos, G., Nienhuis, A. W., Leder, P., & Majerus, P. W., Eds.) pp 127–178, W. B. Saunders Co., Philadelphia, PA.

Raferty, M. A., & Cole, R. D. (1963) *Biochem. Biophys. Res. Commun. 10*, 467–472.

Reider, R. F. (1970) *J. Clin. Invest. 49*, 2369–2376.

Riggs, A. F., & Wolbach, R. A. (1956) *J. Gen. Physiol. 39*, 585–605.

Schneider, R. G. (1978) *Crit. Rev. Clin. Lab. Sci. 9*, 203–271.

Schneider, R. G. Brimhall, B., Jones, R. T., Bryant, R., Mitchell, C. B., & Goldberg, A. I. (1971) *Biochim. Biophys. Acta 243*, 164–169.

Schroeder, W. A. (1985) *Hemoglobin 9*, 609–612.

Singer, K., Chernoff, A. I., & Singer, L. (1951) *Blood 6*, 413–428.

Smyth, D. G. (1967) *Methods Enzymol. 11*, 214–236.

Turner, B. W., Pettigrew, D. W., & Ackers, G. K. (1981) *Methods Enzymol. 76*, 596–628.

Prohormonal Cleavage Sites Are Associated with Ω Loops[†]

Eugene Bek and Robert Berry*

Department of Cell Biology and Anatomy, School of Medicine, Northwestern University, Chicago, Illinois 60611

Received May 11, 1989; Revised Manuscript Received August 14, 1989

ABSTRACT: Secretory peptides are generated from larger precursor proteins, or prohormones, by proteolytic cleavage at sites consisting of one or more basic amino acids. We have investigated the association of these cleavage sites with the various classes of secondary structure in the prohormones. In particular, we determined the association of cleavage sites with the newly defined category of Ω loops. We developed an algorithm for predicting the occurrence of such loops from the primary structure of the precursor and validated this procedure by comparison to crystallographic data. When this method was applied to prohormones, we found that about one-third of the cleavage sites previously assigned to reverse turns were actually associated with Ω loops. Moreover, sites that delimit secreted peptides are most often associated with loops and are concentrated in the neck regions of the loops. These data are interpreted in terms of a model in which the processing endoprotease interacts with two sites on the prohormone: a recognition site in the middle of a loop and the cleavage site at its neck.

Proteolytic processing plays an essential role in the generation of secretory peptides from their larger precursors. Cleavage is known to occur at lysine and arginine residues, most commonly at a Lys–Arg, Lys–Lys, or Arg–Arg pair but also at single residues or strings of three or four basic amino acids (Gluschankof & Cohen, 1987). Thus, the placement of cleavage sites in a protein is encoded in its primary structure. However, sites containing the same set of basic amino acids can be cleaved differentially, suggesting that some aspect of the structure of the region surrounding the cleavage site determines the kinetics of cleavage and, indeed in some cases,

[†] Supported by NIH Grant GM-35115.

[*] To whom correspondence should be addressed.

0006-2960/90/0429-178$02.50/0

whether the site will be cleaved at all. The precursors to many hormones are polyproteins, in which the structures of several secretory peptides are contained within a single precursor. In such cases, the cleavages that generate the different peptides are not made at random but in a time-ordered sequence, indicating that the sites are not treated equivalently by the processing endoprotease(s) (Berry, 1981; Berry & Yates, 1986). Moreover, the multiple cleavage sites in the polyhormone proopiomelanocortin are cleaved differently by different cell types (Mains & Eipper, 1980). Finally, purified putative processing endoproteases exhibit definite substrate specificity (Creminon et al., 1988; Gluschankof et al., 1988).

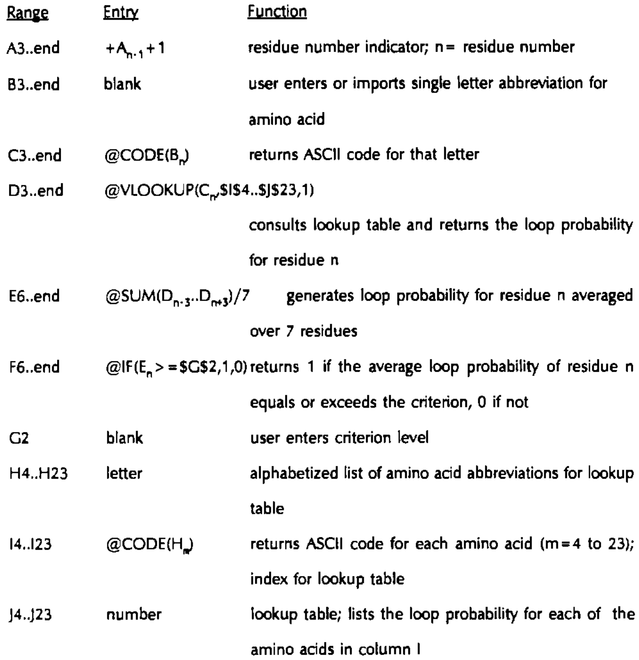
With the exception of single arginine cleavage sites (Schwartz, 1986), there has been no indication that the primary structure of the surrounding region influences cleavage (Gluschankof & Cohen, 1987; Rholam et al., 1986). However, aspects of the secondary or tertiary structure of a cleavage site are likely to be important determinants of the kinetics of cleavage at that site (Gluschankof & Cohen, 1987). The secondary structures immediately surrounding cleavage sites were first investigated by Rholam et al. (1986), who applied predictive methods to 20 precursors of known amino acid sequence. These methods indicated a preponderance of cleavage sites in or near reverse turns separating regions of α helix or β sheet. Such an orientation is appealing on theoretical grounds, since it could expose the site on the end of a fingerlike projection in the aqueous environment. However, it seemed likely to us that at least some of the structural assignments made by these workers were incorrect, since reverse turns are only 3–5 amino acids in length (Rose et al., 1985), whereas some of the proposed turns were as much as 35 residues long. The distinction is not trivial, since turns are strictly defined stereochemical entities (Rose et al., 1985).

The subsequent description of Ω loops by Leszczynski and Rose (1986) suggested a resolution to this problem. On the basis of X-ray crystallographic data, Ω loops were proposed to be structures of 6–16 amino acids in length, containing no ordered secondary structure (α helix or β sheet), and having an end-to-end distance no greater than 1 nm. Significantly, such loops were almost invariably found to occur on the external surface of the protein, which would be a necessary condition for interaction with proteolytic enzymes. Finally, data on the frequency of occurrence of various amino acids in loop regions indicated that residues having a high probability of being found in turns also had a high probability of being found in loops.

Thus, we reasoned that many of the excessively long reverse turns described by Rholam et al. might actually be Ω loops and that proteolytic cleavage sites might be preferentially distributed in loops as well as turns. Accordingly, we created a loop prediction program, validated it against a subset of the proteins studied by Leszczynski and Rose, and applied it to the prohormones studied by Rholam et al. Our results indicate that proteolytic cleavage sites are indeed associated with Ω loops as well as with reverse turns.

Experimental Procedures

At the time we undertook this work, loop prediction programs were not yet available, although one has since appeared (Ralph et al., 1987). Consequently, we designed a simple Lotus 1-2-3 spreadsheet for loop prediction, implementing it on a Compu-Add 286 computer. The basic structure of the spreadsheet is outlined in Figure 1. The program is based on the values given by Leszczynski and Rose (1986) for the normalized frequency of an amino acid's occurrence in loops. It computes for each residue the average of the normalized loop frequencies of that residue and its six nearest neighbors, or *L* value. Our justification for the use of a seven-residue window in the sliding average is that the minimum loop size is six residues and this size window produces acceptable smoothing of the plots. In preliminary studies, we also investigated a five-residue window and found that while the assignment of individual residues in or out of loops was sometimes altered, the number and approximate positions of loops were not. The program next identifies those residues whose *L* value equals or exceeds a user-specified criterion value. The setting of this criterion level is discussed below. Finally, the user eliminates by inspection those loops whose length is less than 6 or greater than 16 residues. It should be noted that the averaging method might artificially elevate the *L* values of residues immediately adjacent to a loop, causing it to be rejected on the basis of exceeding the length restriction, but in practice, only one putative loop had to be rejected because of excessive length (a 19-residue loop in proopiomelanocortin). In some cases, residues near the boundary of a loop fell below the criterion level and then rose above it at the boundary. In these instances the entire span was counted as a loop if the residues in question fell no more than 0.3 unit below the criterion and were flanked by residues that exceeded the criterion level. It should be noted that Ω loops are bounded by structurally defined "necks" (Leszczynski & Rose, 1986). The current version of the program does not take this into account, so that, in general, the actual residues bounding a loop may not be predicted accurately.



Range	Entry	Function
A3..end	$+A_{n-1}+1$	residue number indicator; n= residue number
B3..end	blank	user enters or imports single letter abbreviation for amino acid
C3..end	@CODE(B_n)	returns ASCII code for that letter
D3..end	@VLOOKUP(C_n,I4..$J$23,1)	consults lookup table and returns the loop probability for residue n
E6..end	@SUM(D_{n-3}..D_{n+3})/7	generates loop probability for residue n averaged over 7 residues
F6..end	@IF(E_n>=G2,1,0)	returns 1 if the average loop probability of residue n equals or exceeds the criterion, 0 if not
G2	blank	user enters criterion level
H4..H23	letter	alphabetized list of amino acid abbreviations for lookup table
I4..I23	@CODE(H_m)	returns ASCII code for each amino acid (m=4 to 23); index for lookup table
J4..J23	number	lookup table; lists the loop probability for each of the amino acids in column I

FIGURE 1: Loop prediction spreadsheet used in this study. "Entry" denotes the contents of each cell in the range. Functions are written in Lotus 1-2-3, release 2.01. Column E contains the averaged loop probability of each residue except the first and last three and is useful for plots. Column F contains a string of 1's and 0's, respectively denoting regions of loop probability above and below the criterion level.

To set the criterion level and determine the accuracy of our prediction methods, we analyzed 23 of the proteins studied by Leszczynski and Rose that are known to contain loops on the basis of crystallographic data. These proteins are listed in Table I. From an initial inspection of *L* value plots for several of these proteins, we determined that criterion levels from 1.06 through 1.10 would probably provide the best fit to Leszczynski and Rose's data. We then generated loop predictions for criterion values at increments of 0.01 between these limits and

Table I: Proteins Used To Determine the Accuracy of Loop Predictions

Brookhaven abbrev	protein
155C	cytochrome *c*-550
156B	cytochrome *b*-562
1ABP	arabinose-binding protein
4ADH	alcohol dehydrogenase
1AZU	azurin
2BP2	phospholipase A_2
2C2C	cytochrome c_2
2CAB	carbonic anhydrase B
2CHA	chymotrypsinogen A
3CNA	concanavalin A
3FXN	flavodoxin
1GPD	glyceraldehyde-3-phosphate dehydrogenase
7LYZ	lysozyme
1LZM	lysozyme
2PAB	prealbumin
8PAP	papain
1REI	Bence–Jones immunoglobulin
1RHD	rhodanese
1SBT	subtilisin BPN
2SNS	nuclease
2SOD	superoxide dismutase
2SSI	subtilisin inhibitor
3TLN	thermolysin

compared these to the crystallographic data in several ways. First, for each criterion level, we determined the total number of residues correctly assigned to loops in the 23 proteins (TP = true positives), as well as the total number of residues erroneously predicted to reside in loops (FP = false positives), the number of residues wrongly predicted not to lie in loops (FN = false negatives), and the number of residues correctly predicted to fall outside of loops (TN = true negatives). These values were used to compute the following measures of predictive accuracy at each criterion level: specificity, or the fraction of residues actually residing in loops that are predicted correctly, TP/(TP + FN); sensitivity or the fraction of nonloop residues predicted to reside in nonloop regions, TN/(FP + TN); grade, or the fraction of correct minus incorrect predictions, (TP + TN − FP − FN)/(TP + TN + FP + FN); correlation coefficient

$$\frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{[(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})]^{1/2}}$$

In addition to the above measures that are based on the accuracy of the assignment of individual residues, we also generated measures based on the accuracy of assignment of entire loops. For this calculation, a true positive prediction was scored if the position of the predicted loop overlapped that of an actual loop by 50% or more. We calculated predictive accuracy in two ways: the fraction of actual loops that were correctly predicted (true positives) and the fraction of predicted loops that were actually present in these proteins (correct predictions). In addition, at each criterion level, we calculated the fraction of predicted loops that did not occur (false positives) and the fraction of actual loops that were not predicted (false negatives).

Results

Table II presents the results of tests of the predictive accuracy of the program on a per residue basis. Not unexpectedly, as the criterion level is raised, the fraction of loop residues that are correctly assigned (specificity) drops, with a concomitant increase in the fraction of correct assignments of nonloop residues (sensitivity). As the criterion level is increased, false negatives increase at about the same rate as false positives decrease, leaving the "grade" measure nearly constant. The correlation coefficient shows a nonmonotonic dependence on the criterion level, peaking at 1.07. These values compare favorably with those reported for the loop prediction algorithm of Ralph et al. (1987) (specificity = 0.57, sensitivity = 0.63, grade = 0.24, and correlation coefficient = 0.21). The average predicted loop in the 23 proteins we analyzed was 9.7 residues long, which agrees closely with the mean of 9.8 determined for 67 proteins by Leszczynski and Rose (1986).

Table II: Accuracy of the Loop Prediction Program on a per Residue Basis

	criterion					
	1.06	1.07	1.08	1.09	1.10	1.11
specificity	0.63	0.64	0.56	0.51	0.48	0.44
sensitivity	0.84	0.88	0.90	0.91	0.92	0.93
grade	0.60	0.65	0.65	0.65	0.65	0.65
correlation	0.45	0.49	0.47	0.45	0.45	0.43

Table III: Accuracy of the Loop Prediction Program on a per Loop Basis

	criterion					
	1.06	1.07	1.08	1.09	1.10	1.11
correct predictions	0.47	0.54	0.56	0.53	0.49	0.49
true positives	0.72	0.72	0.69	0.64	0.56	0.53
false positives	0.34	0.26	0.19	0.17	0.14	0.09
false negatives	0.27	0.27	0.30	0.36	0.43	0.45

Table III presents the results of tests of the predictive accuracy of the program on a per loop basis. Regardless of the criterion level, approximately half of the predicted loops match real ones. This measure shows a slight dependency on criterion level, peaking at 1.08, because the increase in the number of correct predictions that occurs as the criterion level is reduced is less than the increase in the total number of loops predicted. The fraction of true positives, or actual loops that are correctly predicted, increases monotonically and more rapidly with decreasing criterion level, such that approximately 70% of loops are correctly predicted at 1.08 and lower. As expected, as the criterion level increases, fewer overestimates of loops are made (false positives), along with more underestimates (false negatives). The two errors are approximately equivalent at a criterion level of 1.07. Because a balance between overestimation and underestimation is a desirable characteristic in a prediction algorithm and because the 1.07 level achieved the highest or second highest score in the other tests of accuracy, this level was chosen for use in subsequent calculations.

We next applied this loop prediction method to the prohormones analyzed by Rholam et al. Since procedures for predicting regions of ordered secondary structure are well developed, we assumed that regions of α helix, β sheet, and reverse turns that did not satisfy the criteria for loops were as predicted by these authors. However, if a predicted loop overlapped a predicted turn, the region was scored as a loop. A predicted region of secondary structure was taken as being associated with a potential or actual processing site if it overlapped it or was separated from it by no more than a single intervening residue. Basic sites were further categorized as to whether or not they flanked peptides that are known to be secreted, regardless of experimental evidence for cleavage.

The results of this analysis are presented in Table IV and summarized in Table V. The 20 prohormones contained 43 predicted loops. Of these, 18 (58%) were associated with basic sites. The mean length of these loops was 8.8 residues. Nineteen of the turns predicted by Rholam et al. were reassigned as loops, leaving 46 turns associated with basic sites. In each case, the length of the loop and that of the turn it

Table IV: Loops, Turns, and Basic Sites in Prohormones[a]

site	flanking?	structure
procalcitonin		
K(83)–R(84)	Y	loop (72–87)
K(118)–K–R(120)	N	turn (115–120)
procalcitonin gene related peptide		
K(81)–K(82)	Y	loop (72–87)
R(121)–R–R(123)	N	turn (110–123)
procorticotropin releasing factor		
		loop (48–54)
		loop (79–86)
		loop (96–105)
R(123)–R(124)	N	turn (121–131)
R(152)–R(153)	Y	loop (154–159)
R(188)–K(189)	N	turn (186–189)
R(196)	Y	helix (190–196)
proenkephalin A		
		loop (27–33)
		loop (85–91)
K(98)–K(99)	Y	loop (99–110)
K(105)–R(106)	Y	loop (99–110)
K(112)–K(113)	Y	loop (99–110)
		loop (121–128)
K(134)–R(135)	Y	turn (134–139)
K(141)–K(142)	N	helix (140–144)
		loop (162–168)
K(186)–R(187)	Y	loop (176–191)
K(197)–R(198)	Y	turn (196–201)
K(210)–R(211)	Y	turn (210–216)
R(217)–R(218)	Y	
K(230)–R(231)	Y	loop (230–235)
K(237)–R(238)	Y	helix (236–242)
		loop (243–255)
K(261)–R(262)	Y	turn (261–266)
proglucagon		
K(51)–R(52)	Y	turn (53–65)
		loop (56–65)
R(69)–K(70)	N	helix (66–78)
K(82)–R(83)	Y	turn (79–84)
K(89)–R(90)	Y	turn (90–101)
R(122)–R(123)	N	helix (112–123)
progonadotropin releasing hormone		
K(−2)–R(−1)	Y	loop (−8 to +1)
K(54)–K(55)	N	turn (49–53)
pro-growth hormone releasing hormone		
		loop (16–24)
R(30)–R(31)	Y	helix (27–38)
R(42)–K(43)	N	turn (39–43)
R(51)–K(52)	N	helix (44–62)
R(77)	Y	turn (72–77)
proinsulin		
R(31)–R(32)	Y	turn (19–31)
		loop (46–54)
K(64)–R(65)	Y	turn (53–66)
		loop (69–76)
proinsulin-like growth factor II		
R(37)–R(38)	N	loop (28–36)
R(68)	Y	turn (60–66)
R(103)–R(104)	N	helix (95–136)
R(113)–R(114)	N	helix (95–136)
K(129)–R(130)	N	helix (95–136)
R(155)–K(156)	N	turn (137–180)
		loop (167–172)
prooxytocin–neurophysin I		
K(11)–R(12)	Y	loop (4–9)
		loop (23–34)
		loop (36–44)
		loop (85–98)
proopiomelanocortin		
		loop (−72 to −64)
R(−57)–K(−56)	Y	turn (−89 to −55)
R(−43)–R(−42)	N	loop (−42 to −34)
proopiomelanocortin		
K(−28)–R(−27)	Y	helix (−31 to −22)
K(−2)–R(−1)	Y	turn (−21 to +3)
K(15)–K–R–R(18)	Y	turn (10–28)
		loop (24–29)
K(40)–R(41)	Y	helix (29–43)
K(82)–K(83)	Y	loop (83–89)
K(102)–R(103)	N	loop (96–104)
K(131)–K(132)	N	helix (117–206)
proparathyroid hormone		
K(−3)–K–R(−1)	Y	loop (−10 to −5)
R(25)–K–K(27)	N	helix (14–41)
R(52)–R–K(54)	N	turn (42–54)
prorelaxin		
K(29)–R(30)	Y	turn (25–31)
K(105)–K(106)	N	turn (94–108)
K(134)–K–R–R(137)	Y	turn (126–137)
		loop (142–147)
K(153)–R(154)	N	turn (151–154)
prorenin		
K(34)–K(35)	N	helix (27–59)
K(62)–R(63)	N	turn (60–64)
K(311)–R(312)	N	helix (298–318)
R(352)–R(353)	Y	turn (332–354)
R(382)–K(383)	N	helix (373–401)
R(401)	Y	
prosomatostatin		
R(−15)	Y	turn (−16 to −2)
R(−2)–K(−1)	Y	
protachykinin		
		loop (26–36)
R(57)	Y	turn (58–62)
K(70)–R(71)	Y	turn (69–77)
K(96)–R(97)	Y	turn (99–102)
K(109)–R(110)	Y	turn (108–111)
R(127)–R–R–K(130)	N	turn (123–127)
prothyrotropin releasing hormone		
R(51)–R(52)	N	turn (48–52)
K(75)–R(76)	Y	helix (68–77)
K(81)–R(82)	Y	turn (78–83)
K(107)–R(108)	Y	helix (104–109)
R(113)–R(114)	Y	turn (110–114)
K(152)–R(153)	Y	helix (151–154)
R(158)–R(159)	Y	turn (155–161)
K(170)–R(171)	Y	helix (162–173)
R(176)–R(177)	Y	turn (173–177)
K(200)–R(201)	Y	turn (198–202)
K(206)–R(207)	Y	turn (205–210)
		loop (212–221)
prourotensin		
K(101)–R(102)	Y	loop (100–109)
R(137)–K(138)	N	turn (136–141)
K(145)	Y	turn (142–145)
provasopressin–neurophysin II		
K(11)–R(12)	Y	loop (4–10)
		loop (23–34)
		loop (36–44)
		loop (85–99)
R(105)–R(106)	N	sheet (101–107)
R(108)	Y	turn (108–123)
provasoactive intestinal peptide		
		loop (69–75)
R(80)	Y	turn (71–91)
K(100)–K(101)	N	helix (92–110)
K(109)–R(110)	Y	helix (92–110)
K(123)–R(124)	Y	turn (111–121)
R(138)–K(139)	N	helix (136–152)
K(144)–K(145)	N	
K(154)–R(155)	Y	loop (154–165)

[a] Amino acid sequences were taken from the references given in Rholam et al. (1986). Abbreviations: K, lysine, R, arginine.

replaced were equal to within two residues. Significantly, 32 of the 46 site-associated turns are greater than five residues in length, yet fail to meet our criteria for loops. Placing these sites in the unassigned category yielded the following distribution of the 93 potential processing sites: 14 in turns, 18 in loops, 24 in helices, 1 in a β sheet, and 36 unassigned.

Differentiating those sites which flank the hormonal products from those which do not, as in Table V, reveals that 70%

Table V: Secondary Structures Associated with Basic Sites in Prohormones

	turns	loops	helices	sheets	unassigned
flanking	32	15	10	0	3
nonflanking	14	3	14	1	1

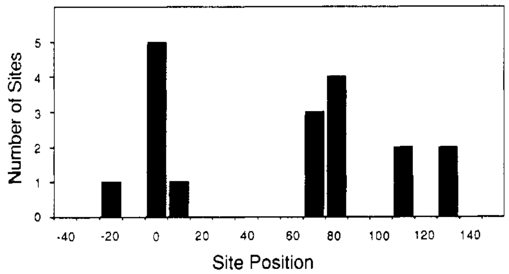


FIGURE 2: Normalized location of basic sites within and near predicted loops. The X axis represents site position with respect to the N- (0) and C-termini (100) of the loop with which that site is associated, measured as a percentage of the length of that loop. The Y axis is the number of loop-associated sites whose C-terminal residue falls within a given 0.1-unit interval on the X axis. The distribution is significantly nonrandom ($p < 0.01$) by the χ^2 test.

of the turn-associated sites are flanking sites, compared with 83% of the loop-associated sites and only 42% of the helix-associated sites. However, flanking sites represent only 50% of the sites associated with turns smaller than six residues in length, that is, with true turns. Thus, if a site is associated with a loop, it is likely to be a flanking site, whereas sites associated with helices and true turns are approximately equally likely to be flanking or not. The distribution of flanking sites among the various categories of secondary structure does not reveal a preponderance in any one category, other than an absence of association with sheets: 23% of the flanking sites are associated with turns less than six residues in length, 25% with loops, 17% with helices, and 0% with sheets; 35% are unassigned. Since we did not attempt to reproduce the entire range of structural assignments made by Rholam et al., we cannot specify the percentage of total residues residing in helices, loops, sheets, and turns. However, the average protein analyzed by Leszczynski and Rose contained 26% helix, 21% loop, 19% sheet, and 15% turn. If this is the approximate composition of the proteins under consideration, then on a per residue basis an assignable flanking site has a 19% probability of being associated with a helix, a 35% likelihood of being associated with a loop, and a 45% probability of residing in a turn.

There is no indication in the data that loops tend to associate with basic sites of a particular composition. All but one of the loop-associated sites are two residues in length, two-thirds of them Lys–Arg, with the remainder divided equally between Arg–Arg and Lys–Lys sites. This approximates the distribution of site types in the proteins studied (59% Lys–Arg, 17% Lys–Lys, 24% Arg–Arg). The sites are not distributed evenly along the lengths of the loops, but tend to cluster near the N- and C-termini (Figure 2).

Discussion

Our algorithm for predicting loops is entirely heuristic, being based merely on the frequency of occurrence of individual amino acids in known loops. Nonetheless, when compared to crystallographic data, it performs reasonably well: Half of its loop predictions are correct, with approximately equal over- and underpredictions, and it correctly matches 70% of those loops that actually occur. Somewhat surprisingly, it significantly exceeds the algorithm of Ralph et al. (1987) in several tests of accuracy. The two methods both employ a sliding average of Leszczynski and Rose's frequency values, but differ in that ours uses a 7-residue window, as opposed to 6–16-residue windows, and in that we apply a criterion test to each residue, whereas Ralph et al. (1987) assigned the maximum value in a given range to each residue in that range. This last feature may be significant, as it would tend to increase the number of false positives. In any event, the accuracy of our method could still stand improvement. One path for future exploration would be to seek common features of amino acids found in the boundaries, or neck regions, of Ω loops. Such an approach to the boundaries of α helices has proven fruitful (Richardson & Richardson, 1988).

Even within the limits of predictive accuracy afforded by the present algorithm, it does appear that a significant fraction (approximately one-third) of the reverse turns identified by Rholam et al. may well be Ω loops. Although turns are allowed to exist within loops (Leszczynski & Rose, 1986), our results cannot be explained by such a relationship, because the predicted loops closely match in length the predicted turns they replace. Approximately two-thirds of the predicted turns not reassigned as loops exceed five residues in length and thus cannot be turns in the strict sense, although they do not meet our criteria for loops. It is likely that these unassigned regions will contain members of both categories, but we feel it is safest not to categorize them at present.

Even disregrading these regions, our results call into question the notion of an exclusive or predominant association of proteolytic processing sites with reverse turns. Of the potential cleavage sites in this data set, 25% occur in true turns, 32% in loops, and 42% in helices. Only 1 of 99 potential sites occurred in a predicted β sheet, suggesting that the constraints imposed by the formation of this rigid structure are incompatible with the formation of a potential cleavage site. If one only considers sites which flank known hormones, that is, sites whose cleavage is well established, the case for cleavage in loops becomes even stronger: approximately half of the flanking sites studied here are associated with loops. Taking into account the relative percentages given over to the various categories of secondary structure in a typical protein, the probability of a flanking site being associated with a loop is about twice its probability of being associated with a helix or true turn.

The fact that cleavage sites are not invariably associated with any single category of predicted secondary structure may reflect more than the inaccuracy of the predictions. In the first place, a given category of secondary structure can have multiple functions. For example, in a given prohormone, some loops may function in ordering tertiary structure, others may provide recognition sites for binding proteins involved in routing, and others may be associated with cleavage sites. Moreover, more than one category of secondary structure may be capable of carrying out a given function. In the context of the present investigation, it may be hypothesized that the role of secondary structure in proteolytic processing is twofold: first, to present both the cleavage site and a separate recognition site to the aqueous environment, where they would be accessible to the processing endoprotease; and second, to maintain the proper spatial relationship between the two. That cleavage sites are not restricted to any one category of secondary structure may simply reflect the fact that there are multiple means by which these functions can be accomplished.

Our finding that cleavage sites associated with Ω loops cluster at their margins is somewhat surprising, since these locations ought to be relatively inaccessible, compared to the midregion. One potential explanation would be that the basic amino acids play a beneficial role in forming the necks of a loop. Under this hypothesis, lysine and arginine should occur more frequently in the necks of loops than elsewhere. However, these amino acids are somewhat underrepresented in these regions: of the proteins in the Leszczynski and Rose data set, lysine and arginine account for 6.9% and 2.4% of the residues at loop borders and 7.1% and 3.1% of the total residues, respectively.

An alternative possibility is suggested by the dual-site model considered above. Under this hypothesis, binding of a processing endoprotease to a recognition site on the exposed portion of the loop would induce a conformational change that would bring the basic site into juxtaposition with the catalytic site on the enzyme. An interesting feature of this mechanism is that the conformational shift induced by enzyme binding could conceivably strain peptide bonds in the neck region, aiding in bond scission.

ACKNOWLEDGMENTS

We thank L. Douglas for advice and encouragement, Dr. A. Tesler for graphics, and Dr. R. Chisholm for database searches.

Registry No. Procalcitonin, 56645-65-9; procalcitonin gene-related peptide, 95077-07-9; procorticotropin-releasing factor, 99533-36-5; proenkephalin A, 88402-54-4; proglucagon, 55963-74-1; progonadotropin-releasing hormone, 123774-88-9; pro-growth-hormone-releasing factor, 96511-22-7; proinsulin, 9035-68-1; proinsulin-like growth factor II, 93928-21-3; proopiomelanocortin, 66796-54-1; proparathyroid hormone, 50815-54-8; prorelaxin, 87004-01-1; prorenin, 39364-01-7; prosomatostatin, 74315-46-1; prothyrotropin-releasing hormone, 98616-54-7; prourotensin, 123774-89-0; provasoactive intestinal peptide, 81032-42-0.

REFERENCES

Berry, R. W. (1981) *Biochemistry 20*, 6200–6205.

Berry, R. W., & Yates, M. E. (1986) *Peptides 7*, 637–643.

Creminon, C., Rholam, M., Bousetta, H., Marakchi, N., & Cohen, P. (1988) *J. Chromatogr. 440*, 439–448.

Gluschankof, P., & Cohen, P. (1987) *Neurochem. Res. 12*, 951–958.

Gluschankof, P., Gomez, S., Lepage, A., Creminon, C., Nyberg, F., Terenius, L., & Cohen, P. (1988) *FEBS Lett. 234*, 149–152.

Leszczynski, J. F., & Rose, G. D. (1986) *Science 234*, 849–855.

Mains, R. E., & Eipper, B. A. (1980) *Ann. N.Y. Acad. Sci. 343*, 94–108.

Ralph, W. W., Webster, T., & Smith, T. F. (1987) *Comput. Appl. Biosci. 3*, 211–216.

Rholam, M., Nicolas, P., & Cohen, P. (1986) *FEBS Lett. 207*, 1–6.

Richardson, J. S., & Richardson, D. C. (1988) *Science 240*, 1648–1652.

Rose, G. D., Gierasch, L. M., & Smith, J. A. (1985) *Adv. Protein Chem. 37*, 1–109.

Schwartz, T. W. (1986) *FEBS Lett. 200*, 1–10.

Sequential Resonance Assignment and Secondary Structure Determination of the *Ascaris* Trypsin Inhibitor, a Member of a Novel Class of Proteinase Inhibitors[†]

Angela M. Gronenborn,*,[‡] Michael Nilges,[‡,∥] Robert J. Peanasky,[§] and G. Marius Clore*,[‡]

Laboratory of Chemical Physics, Building 2, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, and Department of Biochemistry, School of Medicine, The University of South Dakota, 414 East Clark, Vermillion, South Dakota 57069-2390

Received May 2, 1989; Revised Manuscript Received July 20, 1989

ABSTRACT: The solution conformation of the *Ascaris* trypsin inhibitor, a member of a novel class of proteinase inhibitors, has been investigated by nuclear magnetic resonance spectroscopy. Complete sequence-specific assignments of the ^{1}H NMR spectrum have been obtained by using a number of two-dimensional techniques for identifying through-bond and through-space (<5-Å) connectivities. Elements of regular secondary structure have been identified on the basis of a qualitative interpretation of the nuclear Overhauser enhancement, coupling constant, and amide exchange data. These are two β-sheet regions. One double-stranded antiparallel β-sheet comprises residues 11–14 (strand 1) and 37–39 (strand 2). The other triple-stranded sheet is formed by two antiparallel strands comprising residues 45–49 (strand 4) and 53–57 (strand 5) connected by a turn (residues 50–52), and a small strand consisting of residues 20–22 (strand 3) that is parallel to strand 4.

Over the last two decades a large number of primary sequences of protein proteinase inhibitors from a variety of species have been determined [see Laskowski and Kato (1980) for a review]. On the basis of sequence comparisons and functional studies, the serine proteinases have been classified into at least 10 families among which the best known are (i) the pancreatic trypsin inhibitors (Kunitz type), (ii) the pancreatic secretory trypsin inhibitors (Kazal type), and (iii) the *Streptomyces* subtilisin inhibitor family (Laskowski & Kato, 1980). Classification into families is based on extensive amino acid sequence homology, in particular at and surrounding the reactive site, as well as on the topological relationship between disulfide bridges and the location of the reactive site loop. Most members of these families inhibit proteinases according to a common mechanism, forming a substrate-like enzyme–

[†] This work was supported by the Intramural AIDS Targeted Antiviral Program of the Office of the Director of the National Institutes of Health (A.M.G. and G.M.C.) and by the National Institute of Allergy and Infectious Diseases (R.J.P.).

[‡] National Institutes of Health.

[∥] Present address: The Howard Hughes Medical Institute and Department of Molecular Biochemistry and Biophysics, Yale University, 260 Whitney Ave., P.O. Box 6666, New Haven, CT 06511.

[§] The University of South Dakota.

0006-2960/90/0429-183$02.50/0