Primary structural comparison of the preprohormones cholecystokinin and gastrin

Robert J. Deschenes[+], S.V.L. Narayana[*], Patrick Argos[*] and Jack E. Dixon[+,°]

Departments of [*]Biological Sciences and [+]Biochemistry, Purdue University, W. Lafayette, IN 47907, USA

Received 21 December 1984

The nucleotide and amino acid sequences of rat preprocholecystokinin and porcine preprogastrin have been aligned and their secondary structures predicted. Both precursors are predicted to be largely in turn and helical configurations. The sequence homology and position of an intron which splits the preprohormones suggest their evolution from a common ancestral protein.

Preprohormone structure *Cholecystokinin* *Gastrin*

1. INTRODUCTION

Cholecystokinin (CCK) and gastrin (GSN) are polypeptide hormones that were originally isolated and characterized from the mammalian gastrointestinal tract. GSN regulates gastric acid secretion and growth of cells of the stomach epithelium [1], while CCK causes gall bladder emptying and pancreatic enzyme secretion [2]. CCK and GSN have also been detected in the central nervous system [3,4], and there is some evidence that CCK may have a satiety effect in some species [5].

CCK and GSN contain the same carboxyl terminal sequence, GWMDF-amide, but have sequence variation beyond the common pentapeptide [6]. The amino acid sequence of porcine and human preprogastrins have been deduced from the nucleotide sequence of a cDNA clones [7,8]. Recently, the sequence of rat preprocholecystokinin was reported [9] as well as a description of the corresponding rat gene [10]. Here, the nucleotide and amino acid sequences of the CCK and GSN preprohormones are aligned. Based on the sequence alignment and secondary structural predictions, the precursors to CCK and GSN are shown to be closely related.

° To whom correspondence should be addressed

Abbreviations: CCK, rat cholecystokinin; GSN, porcine gastrin; CCK-N, CCK N-terminal residues 20–60; CCK-C, 73–115; GSN-N and GSN-C, GSN residues 21–62 and 65–104, respectively

2. METHODS

Searches for homologous nucleotide sequences in the two preprohormones involved comparisons of every possible span of length L bases from the first cDNA sequence with all possible stretches of length L in the second cDNA. At each oligobase match the total base difference was determined with only a consideration of the first and second base positions due to the degeneracy of the third base in amino acid coding. The length L was chosen as 30 bases (of which only 20 were compared) to allow for possible gaps yet maintain statistical significance. A plot of the observed base differences vs their frequency, generated from all the oligobase alignments, allowed the calculation of the mean base difference and standard deviation to test for significance.

To examine a possible structural relationship between amino acid sequences where nucleotide homology proved statistically insignificant, various residue physical characteristics were used. They included the experimental hydration poten-

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/85/$3.30

tial; the residue 'surrounding' hydrophobicity; the conformational preference parameters for α-helices, β-strands, and reverse turns; and the residue polarity. The parameters have been listed and discussed previously [11]. The characters were selected as they represent the major forces thought to be required for protein folding [12].

Plots of the amino acid sequence number versus a given physical characteristic value for a particular residue in the primary structure were calculated for various sequence fractions of the two preprohormones to test for possible gene duplication. The plots are 'smoothed' by determining a least squares line for all successive 5-point groups, (i) to ($i + 4$), to calculate the ($i + 2$) point on the smoothed curve. The entire smoothing process was repeated for 3 cycles. Cross-correlation coefficients for a particular physical characteristic can then be determined between the smoothed plots. To ascertain the best phase relationship between the plots, correlations are calculated for various sequence registers (lag values) of the two curves. An additive combination of correlation-vs-lag curves for the 6 physical characteristics can yield a possible structural relationship between sequence segments, especially if the maximum correlation sum is above a threshold value determined to be about 2.0 (i.e., an average correlation of 0.33 for each of the physical characters). This method has been discussed in detail elsewhere [13,14].

3. RESULTS AND DISCUSSION

Searches for nucleotide homology in precursor GSN and CCK sequences were performed using a probe length of 30 bases of which only 20 were compared as the third codon bases were ignored. The integral mean base difference and standard deviation of the frequency distributions were 14 and 2, respectively. It was possible to align the precursor to CCK and GSN base sequences (fig.1) using 65% of the observed oligobase matches with base differences at the 2σ level or greater (i.e., a base difference of 10 or less). (A table listing the statistics for the alignment can be obtained from the authors upon request.) In the alignment of fig.1, 63% of the first and second bases in precursors to CCK and GSN are identical, a percentage well above the random value of 25%. Even if the obviously homologous hormone region at the carboxyl termini is removed from these statistics, the identity in first and second bases is still 59%.

Secondary structure predictions for the precursors to GSN and CCK were effected by plotting the helix, strand, and turn conformational preference parameters against sequence number. The curves were then subjected to 3 cycles of smoothing; the results are shown in fig.2. A smoothed preference greater than the normalized, neutral preference value of 1.0 would suggest a particular structural conformation. The two preprohormones are predicted to be largely in a twisted, turn configuration due to the extensive number of prolines and glycines; no strand predictions were observed. The major cleavage sites in the precursors to CCK and GSN are also shown and generally occur within predicted turn regions likely to be exposed at the prohormone surface. The possible sites include proteolysis on the C-terminal side of CCK residues 45, 64, 70, 95 and 103 and of GSN residues 58 and 75 (fig.1).

Correlation coefficients were calculated for 4 physical characteristics between the two aligned sequences. Cross-correlation values of 1.0, 0.0, and -1.0 would, respectively, refer to perfect, random, and oppositely phased correlations. The correlations showed good structural correspondence in the two preprohormones with respective coefficients for hydration potential, polarity, helical preference, and turn potential being 0.52, 0.56, 0.48, and 0.57 (see [15] for parameter listings). These correlation coefficients compare favorably with values relating proteins whose 3-dimensional structures are known to be similar, but whose primary sequence homologies are not easily discernable. For example, NAD binding domains in 3 different dehydrogenases (lactate, glyceraldehyde-6-phosphate, and alcohol) [15] show mean correlation coefficients near 0.20 for all 3 pair-wise comparisons. The predicted structure of the two preprohormones are generally homologous except between CCK 84–95 and GSN 76–84 where CCK shows a predicted turn region while GSN is strongly helical (fig.2).

In the alignment of the precursor sequences GSN and CCK, a long insertion of 11 residues in the CCK sequence (positions 60–70 between GSN residues 62 and 63) is observed (fig.1) and predicted to be in a helical conformation (fig.2). Characterization of the rat CCK gene has revealed

Fig.1. The nucleotide and amino acid sequence alignments of the rat tumor CCK [12] and porcine gut GSN [10] preprohormones. Amino acid identities are boxed. The putative initiator methionine in each case is designated +1. Deletions are indicated by (---).

an intron which splits the gene within the codon for Ala-72 [10]. This suggests that the evolutionary events resulting in GSN and CCK divergence utilized an intron-exon junction to alter the coding sequence of the gene.

Boel et al. [10] have suggested duplication in the nucleotide sequences of the precursors to human and porcine gastrins (GSN residues 29–54 with 62–87). Given the intron insertion at CCK residue 72 and the alignment of the precursors CCK in GSN in fig.1, it would be expected that, if the gene duplication occurred, CCK residues 20–60 (signal peptide and 10-residue CCK insertion removed) and GSN residues 21–62 would display nucleotide and/or structural homology with the CCK segment 73–115 and the corresponding GSN span 65–104.

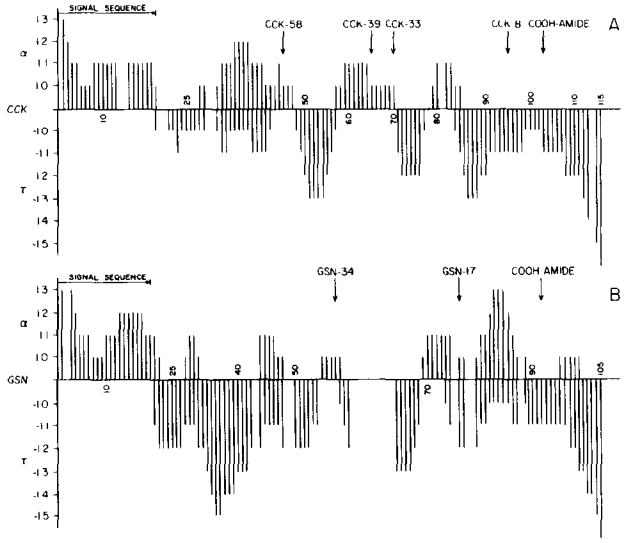

Fig.2. The smoothed conformational preference values are plotted vs residue position for CCK (A) and GSN (B). The letters 'α' and 'τ' refer respectively to the helical and turn smoothed conformational preference values. A value of $+1.0$–1.3 indicates a helical preference, and values -1.0 to -1.5 indicate a turn preference. There were no regions of β sheet predicted in either of the preprohormones.

Boel et al. [8] have suggested that GSN residues 29–54 (CCK residues 26–52) are homologous to GSN 62–87 (CCK 60–97). Nucleotide searches, using a probe length of 21 bases, among the various N- and C-terminal segments suggested by the gene duplication model yielded no statistically significant homology. Plots of 6 physical characteristics were calculated and smoothed by 3 cycles for the CCK and GSN N- and C-terminal segments. Combined characteristic correlation values were determined over a lag range of -15 to $+15$ for all possible comparisons of the CCK and GSN N-terminal spans with the C-terminal segments (CCK-N with CCK-C, CCK-N with GSN-C, GSN-N with CCK-C, and GSN-N with GSN-C). Using the Boel et al. [8] spans, all the correlation maxima were much less than 2.0, the minimum value for statistical significance [15]. Furthermore, the average correlation coefficient over 4 residue characteristics (helical and turn preferences, hydration potential and residue polarity) for the 4 possible GSN and CCK N- and C-terminal segment comparisons was -0.05.

ACKNOWLEDGEMENTS

This work was supported in part by grants from the National Institutes of Health, AM18849, AM20542, and support from predoctoral training grant GM07211 to R.J.D. This is Journal Paper no.9997 from the Purdue University Agricultural Experiment Station.

REFERENCES

[1] Walsh, J.H. and Grossman, M.I. (1975) N. England J. Med. 292, 1377–1384.
[2] Mutt, V. and Jorpes, E. (1971) Biochem. J. 125, 57–58.
[3] Dockray, G.J., Gregory, R.A., Hutchison, J.B., Harris, J.I. and Runswick, M.J. (1978) Nature (London) 274, 711–713.
[4] Larsson, L.-I. and Rehfeld, J.F. (1981) Science 213, 768–770.
[5] Della-Fera, M.A., Baile, C.A., Schneider, B.S. and Grinker, J.A. (1981) Science 212, 687–689.
[6] Mutt, V. (1980) in: Gastrointestinal Hormones (Jerzy, G.B. ed.) pp.169–221, Raven Press, NY.
[7] Yoo, O.J., Powell, C.T. and Agarwal, K.L. (1982) Proc. Natl. Acad. Sci. USA 79, 1049–1053.
[8] Boel, E., Vuust, J., Norris, F., Norris, K., Wind, A., Rehfeld, J.F. and Marcker, K.A. (1984) Proc. Natl. Acad. Sci. USA 80, 2866–2869.
[9] Deschenes, R.J., Lorenz, L.L., Haun, R.S., Roos, B.A., Collier, K.J. and Dixon, J.E. (1984) Proc. Natl. Acad. Sci. USA 81, 726–730.
[10] Deschenes, R.J., Haun, R.S., Funckes, C. and Dixon, J.E. (1985) J. Biol. Chem., in press.
[11] Argos, P., Hanei, M., Wilson, J.M. and Kelly, W.N. (1983a) J. Biol. Chem. 258, 6450–6457.
[12] Creighton, T.E. (1978) Prog. Biophys. Mol. Biol. 33, 231–297.
[13] Argos, P. and Siezen, R.J. (1982) Eur. J. Biochem. 131, 143–148.
[14] Argos, P., Taylor, W.L., Minth, C.D. and Dixon, J.E. (1983b) J. Biol. Chem. 258, 8788–8793.
[15] Otto, J., Argos, P. and Rossmann, M.G. (1980) Eur. J. Biochem. 109, 325–330.