PREDICTED SECONDARY STRUCTURES OF AMINO-TERMINAL EXTENSION SEQUENCES OF SECRETED PROTEINS

B. M. AUSTEN
National Institute for Medical Research, Mill Hill, London NW7 1AA, England

Received 24 April 1979

1. Introduction

Recently it has been shown that the mRNA for many secreted proteins translates in cell-free systems to yield precursors that are larger than mature proteins. The precursors contain amino-terminal extensions known as signal sequences, of 15–30 residues of predominantly hydrophobic amino acid residues [1–3]. Segregation of the secreted protein into vesicles prepared from pancreatic microsomal membranes and proteolytic removal of the extensions occur during translation; the completed pre protein is not processed [4–7]. Small signal peptides have been detected during processing [8] showing that the signal protease is endoproteolytic. Ovalbumin may be an exception in that no proteolysis occurs although nascent ovalbumin and nascent prolactin compete for common receptors at the endoplasmic reticulum (ER) membrane during segregation [9].

According to the signal hypothesis [10] the amino-terminal sequence of the nascent polypeptide chain, as it emerges from the large subunit of the ribosome, directs the ribosomal complex to the ER membrane. Some receptors, which are proteins that are components of the ER, are required for transfer and segregation [11]. The transmembrane glycoprotein from vesicular stomatitis virus is also known to be synthesized initially with a transient 16-residue amino-terminal extension [12], and similar mechanisms operate in prokaryotic cells because several secreted or membrane proteins initially synthesized with transient amino-terminal extensions have been identified in bacteria [13,14]. Empirical methods of secondary structure prediction have been applied to signal sequences and it is shown that there are certain common structural features that could be involved in their transfer across membranes, and that could direct the signal protease to specific residues.

2. Predictions of secondary structure

Sequences of 21 amino-terminal sequences are listed in table 1. The list includes precursors of prohormones, immunoglobulins, milk proteins, egg proteins, exported bacterial proteins and a trans-membrane glycoprotein. The amino-terminal sequence of ovalbumin, which is identical to that of the product translated in vitro except for replacement of the initiator methionine by an acetyl group [9], is also included.

Predictions of secondary structure were made by locating 6 or 5 residue nucleation sites, and 4 residue boundaries [25] with the use of recent conformational parameters [26]. The results are shown in fig.1. The central portions of all the signal sequences were strongly predicted to be involved in repeating structure, either α-helix or β-sheet, of 6–21 residues in length. Where overlap between α-forming and β-forming residues was found, conformational parameters were calculated, and the assignments in fig.1 were made according to the relative values of $\langle P\alpha \rangle$ and $\langle P\beta \rangle$. It is likely, however, that some of the sections shown as α- or β-structure would be capable of folding either way.

Fourteen sequences listed in table 1 contain basic residues within 5 residues of the initiator methionine. The cationic side chains of these residues, and the

Table 1
Signal sequences found in secreted or transmembrane proteins

Precursor	Amino acid sequence	Ref.	Length[a]	HI[b]
	▼			
Pre ovomucoid	M A M A G V F V L F S F V L C G F L P D A A F $\overset{24}{\mathrm{G}}$ A E V D	[3]	19	2.50
Pre lysozyme	M R S L L I L V L C F L P L A A L $\overset{18}{\mathrm{G}}$ K V F X	[3]	16	2.53
Pre conalbumin	M K L I L C T V L S L G I A A V C F $\overset{19}{\mathrm{A}}$ A P P K	[15]	20	2.10
Pre promellitin	M K F L V X V A L V F M V V Y I X Y I Y $\overset{21}{\mathrm{A}}$ A P E P	[16]		
Pre IgG (light chain)				
MOPC41	M D M R A P A Q I F G F L L L L F P G T R $\overset{22}{\mathrm{C}}$ D I Q M	[17]	16	2.41
MOPC321	M E T D T L L L W V L L L W V P G S T $\overset{20}{\mathrm{C}}$ D I V L	[17]	16	2.63
MOPC104E	M A W I S L I L S L L A L S S G A I $\overset{19}{\mathrm{S}}$ Q A V V	[17]	25	1.38
Pre proparathyrin	M M S A K D M V K V M I V M L A I C F L A R S D $\overset{25}{\mathrm{G}}$ K S V K	[18]	12	2.92
Pre proinsulin I	M A L W M R F L P L L A L L V L W E P K P A Q $\overset{24}{\mathrm{A}}$ F V K Q	[19]	11	3.45
Pre growth hormone	M A A D S Q T P W L L T F S L L C L L W P Q E A G $\overset{26}{\mathrm{A}}$ L P A M	[20]	18	2.22
Pre proalbumin	M K W V T F L L L L F I S G S A F $\overset{18}{\mathrm{S}}$ R G V F	[21]	16	2.72
Pre trypsinogen 2	M A K L F L P L A L L L A Y V $\overset{16}{\mathrm{A}}$ F P L D	[1]	16	3.22
Pre α_{s1}-casein	M K L L I L T C L V A V A L $\overset{15}{\mathrm{A}}$ R P K H	[22]	13	2.46
Pre k-casein	M R K S I L L V V T I L A L T L P F L I $\overset{21}{\mathrm{A}}$ Q E Q N	[22]	19	2.68
Pre α-lactalbumin	M M S F V S L L L V G I L F X A T Q $\overset{19}{\mathrm{A}}$ E Q L T	[22]	16	2.34
Pre β-lactoglobulin	M K C L L L A L G L A L A C G V Q $\overset{18}{\mathrm{A}}$ I I V T	[22]	20	2.08
Pre opiocortin	M P R L C S S R S G A L L L A L L L Q A S M E V R $\overset{26}{\mathrm{G}}$ W C L E	[23]	14	1.46
Pre lipoprotein	M K A T K L V L G A V I L G S T L L A $\overset{20}{\mathrm{G}}$ C S S N	[13]	20	1.56
Pre penicillinase	M S I Q H F R V A L I P F F A A F C L P V F $\overset{23}{\mathrm{A}}$ H P E T	[24]	16	2.81
Pre glycoprotein VS Virus	M K C L L Y L A F L F I(H V N) $\overset{16}{\mathrm{C}}$ K F X I	[12]	10 or 12	3.22
Ovalbumin (not cleaved)	M G S I G A A S M E F C F D V F K E L K V H H A N E N I F Y C P I- A I M S A L A M V Y L G A K	[10]	20	2.54

[a] Length of sequences with consecutive uncharged residues
[b] Hydrophobicity index [29] of uncharged segments
(▼) Numbers of residues from amino-termini of sites cleaved by the signal protease

partially charged amino-group at the amino-terminus, could provide possible sites for electrostatic interactions. β-Turns were predicted to occur close to the amino-termini of some of the eukaryotic signal sequences, and ovalbumin fitted into this pattern as its 4 amino-terminal residues were predicted as a β-turn, followed by a long stretch of 21 helical residues.

In 12 of the sequences listed in table 1, proline, which is rarely found in inner helical or sheet structures, occurs within 6 residues of the cleavage sites. In many cases, β-turns are predicted close to the cleavage sites, and tetrapeptide sequences that would be expected to break repeating structures (table 2), with $<P\alpha>$ or $<P\beta>$ of <1.0, occur with high frequency in this region.

The positions of cleavage, which are aligned in table 1, have been ascribed by knowledge of the amino-terminal sequences of the mature or pro-forms of the secreted proteins. Thus, it is assumed that pre trypsinogen is cleaved between residue 16 (Ala) and 17 (Phe) as phenylalanine is the amino-terminal residue of pig trypsinogen [27]. It is apparent that the signal protease cleaves only on the carboxy-terminal side of the uncharged, relatively small amino acids glycine, serine, alanine and cysteine, and amino acid residues with more bulky side chains may not be accommodated in the active side of this protease.

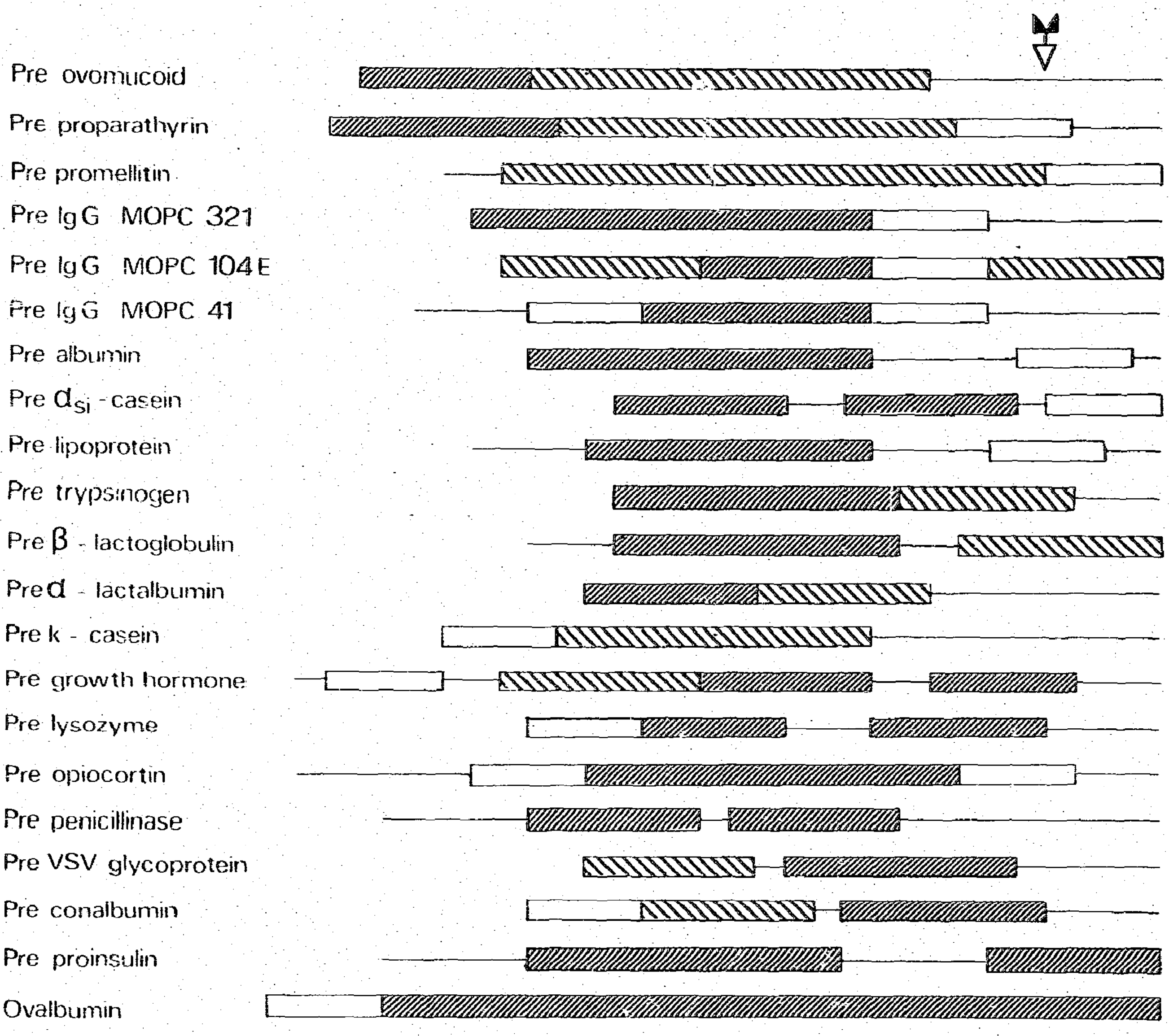

Fig.1. Secondary structures predicted in signal sequences. Close hatching represents α-helix, sparse hatching represents β-structure, and empty rectangles represent β-turns. Sequences are aligned so that the sites of cleavage of the signal protease fall directly under the arrow marker, except for ovalbumin, which is not cleaved.

Although it is not known if proteolytic cleavage of the signal sequence of the initially translated form of the common precursor of corticotrophin and β-endorphin (pre opiocortin) occurs [23], a possible cleavage site exists between glycine (residue 26) and tryptophan (residue 27). This cleavage site is suggested because of similarities in the predicted secondary structures. A β-turn is predicted close to this site (residues 24–27) (see fig.1).

In all transient signal sequences regions of 11–25 residues are found that are free of charged residues. As noted [1,18,21] the amino-terminal extensions are rich in hydrophobic residues. In table 1, hydrophobic residues [28] have been underlined, and it is apparent that in most of the signal sequences these residues are concentrated in central cores of 10–14 residues. The mean hydrophobicity index (HI) [29] for the continuous stretches of uncharged amino acid residues was calculated to be 2.51 ± 0.48 (table 1). This is similar to the HI value of 2.62 for the uncharged

Table 2
Conformational parameters for tetrapeptide sequences occurring in the vicinity of 'signal protease' cleavage sites

Precursor	Residue numbers	$<P\alpha>$	$<P\beta>$
Pre ovomucoid	19–22		0.69
Pre lysozyme	18–21	0.98	
Pre conalbumin	20–23	0.93	
Pre promellitin	22–25		0.56
Pre IgG MOPC 41	18–21	0.74	
MOPC 321	16–19	0.69	
MOPC 104E	14–17	0.88	
Pre proparathyrin	22–25		0.74
Pre proinsulin	18–21	0.93	
Pre growth hormone	22–25	1.15	0.76
Pre trypsinogen	17–20	0.73	
Pre α_{s1}-casein	16–19	0.93	
Pre k-casein	15–18	0.94	
Pre α-lactalbumin	12–15	0.99	
Pre β-lactoglobulin	14–17	0.86	
Pre lipoprotein	20–23	0.70	
Pre penicillinase	23–26	1.00	

segment of 23 residues which comprises the intramembranous stretch of glycophorin from the human erythrocyte [29]. Ovalbumin contains a sequence of 20 uncharged residues commencing 27 residues from the amino-terminus. This segment was found to have an HI value of 2.54, and may be a candidate for performing the same function as similar regions in other signal sequences.

3. Discussion

It could be considered that the regions of uncharged predominantly hydrophobic residues found in signal sequences of secreted or membrane-bound proteins may come into intimate contact with the hydrophobic interior of the ER membrane during transport. Similar uncharged sequences rich in hydrophobic residues, are found in those sections of membrane proteins which are likely to be inserted into membranes [29]. Signal sequences may therefore fold in such a way that hydrophobic side chains form a surface interacting directly with the lipid core of the membrane, or with intramembranous segments of other proteins which are normal constituents of the ER membrane. This interaction may occur after initial binding of the ribosomal complex to receptors exposed to the cytoplasm [9], and these receptors could recognise common structural features of folded signal sequences as they emerge from the ribosomal complex.

A postulated arrangement for a typically folded signal sequence, containing elements of both α-helix and β-strand structures, is depicted in fig.2. Although predictive methods are based on data obtained from water-soluble globular proteins, and their application to segments of proteins in contact with membranes may be less certain, recent studies have shown that predicted secondary structures of intramembranous peptides of cytochrome b_5 are consistent with results of circular dichroism measurements [30]. Folding of signal peptides as predicted in fig.1 would lead to stabilization in the hydrophobic environment of the interior of the membrane as their polypeptide backbones would be involved in hydrogen bond formation. Thus, the more stable structure would be helical, but segments predicted as β-strands might also be stabilised by hydrogen-bonded interactions with suitable intramembranous portions of protein components of the ER membrane (as shown in fig.2), or with other segments of the secreted protein.

Ovalbumin, which is not cleaved after transfer, was predicted to adopt a secondary structure which is sim-

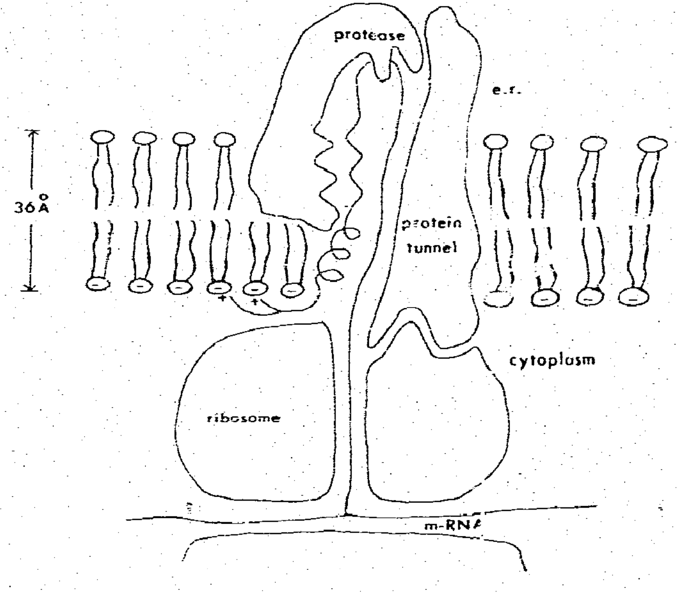

Fig.2. Hypothetical model of how folded signal sequences might interact with lipid and proteinaceous components of the ER membrane.

ilar to structures predicted for other signal sequences. The finding that a mutated form of lipoprotein from *Escherichia coli* with a replacement in its signal sequence is secreted, but not cleaved [31] also suggests that proteolytic cleavage is not essential for transfer. It is not clear why cleavage of ovalbumin at small uncharged residues, such as alanine (residue 24), does not occur, although it is possible that the signal protease recognises additional features of primary sequences, which are lacking in the amino-terminal section of ovalbumin. β-Turns or tetrapeptide sequences which could break α-helix or β-sheet formation are predicted to occur at or near cleavage sites, suggesting that the signal protease is directed to its site of attack by the secondary structures, and it is interesting that no such tetrapeptide sequence is found near alanine (residue 24) in ovalbumin.

It is conceivable that the signal peptide, once inserted into the membrane, could form part of a protein tunnel required to channel the rest of the secreted protein through the membrane (fig.2). It is envisaged that anchoring electrostatic interactions occur between the cationic groups found close to the amino-terminus of the signal peptide, and the anionic head-groups of phospholipids on the cytoplasmic side of the membrane. In the model shown, cleavage sites, in a loop structure, are exposed to the signal protease in the lumen. About 11 residues in β-strand structure, or 23 residues in α-helical structure, would be required to span a membrane of 36 Å dimension, and this might suggest why stretches of 11–25 uncharged residues are found in signal sequences.

By calculating the translation distances of regions of consecutive uncharged residues folded into their predicted structures it was found that these were ⩾36 Å in all but 5 cases. If the additional lengths of the side chains of the charged residues at each end of the uncharged regions were considered, then all these regions could span the membrane as depicted in fig.2. Experiments designed to test the physical and biochemical properties of synthetic signal peptides may provide evidence for this postulated arrangement.

Acknowledgement

I thank Dr D. G. Smyth for his continued support and valuable discussion.

References

[1] Devillers-Thiery, A., Kindt, T., Scheele, G. and Blobel, G. (1975) Proc. Natl. Acad. Sci. USA 72, 5016–5020.

[2] Milstein, C., Brownlee, G. G., Harrison, T. M. and Mathews, M. B. (1972) Nature New Biol. 239, 117–120.

[3] Palmiter, R. D., Gagnon, J., Ericsson, L. H. and Walsh, K. A. (1977) J. Biol. Chem. 252, 6386–6393.

[4] Blobel, G. and Dobberstein, B. (1975) J. Cell. Biol. 67, 852–862.

[5] Shields, D. and Blobel, G. (1978) J. Biol. Chem. 253, 3753–3756.

[6] Kaschnitz, R. and Kreil, G. (1978) Biochem. Biophys. Res. Commun. 83, 901–907.

[7] Maurer, R. A. and McKean, D. J. (1978) J. Biol. Chem. 253, 6315–6318.

[8] Jackson, R. C. and Blobel, G. (1977) Proc. Natl. Acad. Sci. USA 74, 5598–5602.

[9] Palmiter, R. D., Gagnon, J. and Walsh, K. A. (1978) Proc. Natl. Acad. Sci. USA 75, 94–98.

[10] Blobel, G. and Dobberstein, B. (1975) J. Cell. Biol. 67, 835–851.

[11] Warren, G. and Dobberstein, B. (1978) Nature 273, 569–571.

[12] Lingappa, V. R., Katz, F. N., Lodish, H. F. and Blobel, G. (1978) J. Biol. Chem. 253, 8667–8670.

[13] Inouye, S., Wang, S., Sekizawa, J., Halegoua, S. and Inouye, M. (1977) Proc. Natl. Acad. Sci. USA 74, 1004–1008.

[14] Bassford, P. and Beckwith, J. (1979) Nature 277, 538–541.

[15] Thibodeau, S. N., Lee, D. C. and Palmiter, R. D. (1978) J. Biol. Chem. 253, 3771–3774.

[16] Suchanek, G., Kreil, G. and Hermodson, M. A. (1978) Proc. Natl. Acad. Sci. USA 75, 701–704.

[17] Burstein, Y. and Schechter, I. (1978) Biochemistry 17, 2392–2400.

[18] Habener, J. F., Rosenblatt, M., Kemper, B., Kronenberg, H. M., Rich, A. and Potts, J. T. (1978) Proc. Natl. Acad. Sci. USA 75, 2616–2620.

[19] Chan, S. J., Keim, P. and Steiner, D. F. (1976) Proc. Natl. Acad. Sci. USA 73, 1964–1968.

[20] Seeburg, P. H., Shine, J., Martial, J. A., Baxter, J. D. and Goodman, H. M. (1977) Nature 270, 486–494.

[21] Strauss, A. W., Bennett, C. D., Donohue, A. M., Rodkey, J. A. and Alberts, A. W. (1977) J. Biol. Chem. 252, 6846–6855.

[22] Mercer, J. C., Haze, G., Gaye, P. and Hue, D. (1978) Biochem. Biophys. Res. Commun. 82, 1236–1245.

[23] Nakanishi, S., Inoue, A., Kita, T., Nakamura, M., Chang, A. C. Y., Cohen, S. N. and Numa, S. (1979) Nature 278, 423–427.

[24] Ambler, R. P. and Scott, G. K. (1978) Proc. Natl. Acad. Sci. USA 75, 3732–3736.

[25] Chou, P. Y. and Fasman, G. D. (1974) Biochemistry 13, 222–244.

[26] Fasman, G. D., Chou, P. Y. and Adler, A. (1976) Biophys. J. 16, 1201–1238.

[27] Charles, M., Rovery, M., Guidoni, A. and Desnuelle, P. (1963) Biochim. Biophys. Acta 69, 115–129.

[28] Nozaki, Y. and Tanford, C. (1971) J. Biol. Chem. 246, 2211–2217.

[29] Segrest, J. P. and Feldman, R. J. (1974) J. Mol. Biol. 87, 835–838.

[30] Dailey, H. A. and Strittmatter, P. (1978) J. Biol. Chem. 253, 8203–8209.

[31] Lin, J. J. C., Kanazawa, H., Ozols, J. and Wu, H. C. (1978) Proc. Natl. Acad. Sci. USA 75, 4891–4895.