Vol. 93, No. 1, 1980 BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
March 13, 1980 Pages 82-86

A SEQUENCE CORRELATION BETWEEN OPPOSITELY CHARGED RESIDUES IN SECRETED PROTEINS

Gunnar von Heijne

Research Group for Theoretical Biophysics
Department of Theoretical Physics
Royal Institute of Technology, S-10044 Stockholm, Sweden

Received November 26, 1979

SUMMARY: Amino acid sequences of 14 secreted proteins have been secreened for correlations between the relative positions of oppositely charged residues. The results show that arginine residues belonging to a particular class of chain segments have glutamic or aspartic acid residues situated four positions away in the sequence significantly more often than expected by chance. This finding is discussed in the light of various models for the trans-membrane translocation process.

During the past few years, considerable progress has been made in the elucidation of the process of trans-membrane translocation of secreted and membrane-spanning proteins. Specifically, the "signal hypothesis" (1) has proved to be very fruitful. However, it is not yet clear whether the nascent chain traverses the membrane through a protein pore (as originally assumed), or if it is simply extruded directly through the lipid bilayer. Experimental (2) as well as theoretical (3,4) results in favour of the latter idea have recently been obtained.

A possible, as of yet unexploited, source of information in this context could be provided by the relatively large number of known amino acid sequences for secreted proteins. In fact, we have already pointed out that, for energetic reasons, charge neutralization between neighbouring, oppositely charged residues should be important if the chain does indeed pass directly through the membrane (4). In this communication, we present data showing that arginine residues belonging to particular, critical chain segments (see below) have glutamic or aspartic acid residues situated four positions away in the sequence significantly more often than expected by chance.

METHODS

From an initial sample of 30 secreted proteins of known sequence, 14 were chosen for further study as a result of a first analysis by the method described in (4); i.e. the physico-chemically based analysis of the translocation process of

0006-291X/80/050082-05$01.00/0

secreted and membrane-spanning proteins developed therein was used to check which proteins were predicted to become extruded without needing charge neutralization between neighbouring residues. Only those proteins predicted to become stuck in the membrane (assuming no neutralization) were kept, and their residues were divided into two groups: (i) segments predicted to halt translocation in the absence of charge neutralization, i.e. the 21 residues spanning the membrane at the moment of halt, and (ii) segments predicted to be extruded even in the absence of charge neutralization, Table I.

The two groups were then subjected to the following test: for each type of positively charged residue (His, Lys or Arg) the total number of occurences N in each group was noted. Moreover, the number of times n that the given type of residue had Glu and/or Asp at one or both of the two positions $\pm j$ residues away was counted for $j=1,\ldots,5$. The observed distributions were compared with hypothetical distributions calculated under an assumption of no correlation between the relative positions of the residues.

Thus, if the frequencies of Glu and Asp residues in group (i) (say) are f_{Glu} and f_{Asp}, the total probability that

TABLE I

Protein	Group (i)-segments	Ref.
E. Coli β-lactamase	53-73, 130-150	6
B. Licheniformis β-lactamase	55-75, 127-147, 232-252	7
L-arabinose binding protein	112-132	8
S. Aureus enterotoxin B	61-81	9
B. Thermoproteolyticus thermolysin	198-218	10
ACTH-β-LPH precursor	94-114	11
Human chorionic gonadotropin, β-subunit	86-106	12
Rat proinsulin	17-37	13
Proparathyroid hormone	42-62	14
Ovalbumin	271-291	15
Bovine pancreatic trypsin inhibitor	32-52	16
Pig phospholipase A2	42-62	17
Bovine prothrombin	43-63, 87-107, 248-268, 378-398, 489-509	18
Human serum albumin	89-109	19

Group (i)-segments for the 14 proteins used in the study. The numbering starts from the N-terminus of the mature protein. A group (i)-segment is defined as the 21 residues spanning the membrane predicted by the method described in (4) to halt translocation when the free energy contribution from charge neutralization between oppositely charged residues in the membrane is not taken into account, see text. All segments not in group (i) are assigned to group (ii).

a positively charged residue at position k has Glu and/or Asp at positions k-j and/or k+j is given by:

$$p = 2(f_{Glu} + f_{Asp})(1 - f_{Glu} - f_{Asp}) + f_{Glu}^2 + f_{Asp}^2 + 2f_{Glu}\, f_{Asp} =$$

$$= 2(f_{Glu} + f_{Asp}) - (f_{Glu} + f_{Asp})^2$$

i.e. as the sum of the probabilities that either k-j or k+j is Glu or Asp and that both k-j and k+j are Glu and/or Asp.

The observed distributions defined by the numbers (n, N-n) were tested against the hypothetical (calculated) distributions (pN, (1-p)N), and the significance of the differences assessed by χ-square analysis.

RESULTS AND DISCUSSION

The reason for distinguishing between groups (i) and (ii) above is that since segments in (i) are not predicted to pass through the membrane (given the assumptions in (4)) unless charge neutralization is taken into account, one is led to expect that evolution should favour charge neutralization in such parts of the proteins. In group (ii) no similar evolutionary tendency should be expected on these grounds.

The results are presented in Table II. Only in two cases does the observed distribution differ significantly from the calculated one: in group (i) for Arg, j=4, and in group (ii) for Lys, j=3. Moreover, the first case displays a clear tendency among the individual proteins: in 12 out of the 14 proteins is the number of observed Arg residues with Glu and/or Asp at j=4 larger than the calculated one.

TABLE II

		Lysine, j=3			Arginine, j=4		
		I	II	SL	I	II	SL
Group (i)	observed	15	18	-	22	21	0.5%
	calculated	10.8	22.2		13.4	29.6	
Group (ii)	observed	76	162	1%	37	115	-
	calculated	58.0	180.0		37.1	114.9	

Observed and calculated distributions of group (i)- and (ii)-residues for Lys, j=3, and Arg, j=4. I: n:o of residues with Glu and/or Asp at positions k±j, II: n:o of residues with no Glu or Asp at positions k±j. SL is the significance level as judged by χ-square analysis. No other distribution with SL≤1% was found for either His, Lys or Arg with j=1,...,5.

This is not so in the second case, where only 8 out of the 14 proteins have more Lys residues with Glu and/or Asp at j=3 than calculated. In fact, the total bias results from only one protein, the L-arabinose binding protein, in which 16 Lys residues out of 27 in group (ii) have Glu and/or Asp at j=3, the calculated number being only 6.6. Presumably, then, some functional or structural constraint peculiar to this one protein, rather than a general property of the whole group, lies behind this observation.

As was stated already in (3), the α-helix should be the most likely conformation for a membrane-spanning polypeptide segment. Thus, the finding that Arg preferentially has Glu and/or Asp at positions k±4 fits well with the observation that residues of opposite charge in helical segments of native proteins tend to have this same distance (5), i.e. with their side-chains close together in the helical structure, thus facilitating electrostatic interactions.

The fact that the only significant correlation found involves Arg rather than His or Lys also makes sense in the light of the energetic analysis presented earlier (3,4), since the estimated gain in free energy obtained when an Arg within the membrane bilayer is neutralized is almos twice as large as for His and Lys.

In conclusion, the observed bias towards Arg - (Glu, Asp) pairs with j=4 in group (i)-segments may be taken as evidence in favour of the idea that secreted and membrane-spanning proteins are extruded directly through the lipophilic interior of the membrane.

REFERENCES

1. Blobel, G., and Dobberstein, B. (1975) J. Cell Biol. 67, 835-851.
2. Wickner, W. (1979) Ann. Rev. Biochem. 48, 23-45.
3. von Heijne, G., and Blomberg, C. (1979) Eur. J. Biochem. 97, 175-181.
4. von Heijne, G. (1979) Eur. J. Biochem., in press.
5. Maxfield, F. R., and Scheraga, H. A. (1975) Macromolecules 8, 491-493.
6. Sutcliffe, J. G. (1978) Proc. Natl. Acad. Sci. USA 75, 3737-3741.
7. Meadway, R. J. (1969) Bioch. J. 115, 12P-13P.
8. Hogg, R. W., and Hermodson, M. A. (1977) J. Biol. Chem. 252, 5135-5141.

9. Dayhoff, M. O. (1972) Atlas of Protein Sequence and Structure, vol. 5, D-227.
10. Dayhoff, M. O. (1973) Atlas of Protein Sequence and Structure, vol. 5 suppl. I, S-31.
11. Nakanishi, S., Inoue, A., Kita, T., Nakamura, M., Chang, A. C. Y., Cohen, S.N., and Numa, S. (1979) Nature (Lond.) 278, 423-427.
12. Morgan, F. J., Birken, S., and Canfield, R. E. (1975) J. Biol. Chem. 250, 5247-5258.
13. Villa-Komaroff, L., Efstratiadis, A., Broome, S., Lomedico, P., Tizard, R., Naber, S. P., Chick, W. L., and Gilbert, W. (1978) Proc. Natl. Acad. Sci. USA 75, 3727-3731.
14. Rosenblatt, M., Habener, J. F., Tyler, G. A., Shepard, G. L., and Potts, J. T. (1979) J. Biol. Chem. 254, 1414-1421.
15. McReynolds, L., O'Malley, B. W., Nisbet, A. D., Fothergill, J. E., Givol, D., Fields, S., Robertson, M., and Brownlee, G. G. (1978) Nature (Lond.) 273, 723-728.
16. Dayhoff, M. O. (1972) ibid, D-169.
17. Dayhoff, M. O. (1972) ibid, D-156.
18. Dayhoff, M. O. (1976) Atlas of Protein Sequence and Structure, vol. 5 suppl. 2, 95.
19. Dayhoff, M. O. (1976) ibid, 266.