Volume 14 Number 11 1986 Nucleic Acids Research

A new method for predicting signal sequence cleavage sites

Gunnar von Heijne

Research Group for Theoretical Biophysics, Department of Theoretical Physics, Royal Institute of Technology, S-100 44 Stockholm, Sweden

Received 5 March 1986; Revised and Accepted 5 May 1986

ABSTRACT

A new method for identifying secretory signal sequences and for predicting the site of cleavage between a signal sequence and the mature exported protein is described. The predictive accuracy is estimated to be around 75-80% for both prokaryotic and eukaryotic proteins.

INTRODUCTION

The transient N-terminal signal sequence found on most secretory proteins serves to initiate export across the inner membrane (in prokaryotes) or the endoplasmic reticulum (in eukaryotes). Three structurally and, possibly, functionally distinct regions have been identified as the basic building-blocks of a secretory signal sequence: a basic N-terminal region (n-region), a central hydrophobic region (h-region), and a more polar C-terminal region (c-region) (1). The structural determinants for cleavage of the signal sequence from the mature protein once export is under way seems to reside in the n- and h-regions, with positions -3 and -1 relative to the cleavage site being the most important ones (2,3). Indeed, this "(-3,-1)-rule" has been used quite successfully to predict the most likely site of cleavage directly from the primary sequence (2).

In view of the great interest in secretory proteins and the fact that most such proteins are known only from their DNA sequence, it is important to assess and, if possible, to improve upon the predictive accuracy of the original method. In this paper, I present a new scheme based on a weight-matrix approach that can be expected to give correct predictions about 75-80% of the time when applied to new sequences (both prokaryotic and eukaryotic). This represents a substantial gain over the old method, which is shown to be around 65% and 45% accurate for eukaryotic and prokaryotic proteins, respectively.

METHODS

161 eukaryotic and 36 prokaryotic non-homologous signal sequences with known cleavage sites were chosen from my collection of signal sequences totalling at the present time some 450 eukaryotic and 80 prokaryotic entries. The prokaryotic sample did not include any sequences known to be cleaved by the lipoprotein signal peptidase (signal peptidase II) (4).

Weight-matrices W(a,i) (see below) were calculated from the observed amino acid counts in each position, N(a,i), (i.e. the number of residues of type a in position i) with all sequences aligned from their known site of cleavage between positions -1 and +1, by first dividing all counts by their respective expected abundance in proteins in general, <N(a)> (Tables 1 & 2, last column), and then taking the natural logarithms of these quotients: $W(a,i) = \ln(N(a,i)/\langle N(a)\rangle)$. To correct for the limited size of the data base, all zero-elements in the amino acid count matrices were put equal to one before the divison. Zero-counts in positions -3 and -1 were treated differently: they were also put equal to one, but then divided by the total number of sequences in the sample, N, rather than the expected number of residues, e.g. $W(a,-1) = \ln(1/N)$ if $N(a,-1) = 0$.

The most probable cleavage site was identified by scanning the sequence in question with the appropriate weight-matrix and summing the weights for each position, i.e. $S(i) = W(a_{i-p},i-p) + W(a_{i-p+1},i-p+1) + \ldots + W(a_{i+q},i+q)$ where the summation window extends from position i-p to i+q. The predicted cleavage site j is the one with the highest S-value, $S(j) = \max[S(i);\ i=1-p,..,L-q]$, where L is the length of the sequence analyzed. As shown below, maximum predictive accuracy was obtained for p=-12 and q=2.

RESULTS

The (-3,-1)-rule

Based on previous statistics (2), acceptable cleavage sites were suggested to conform to the following rules: the residue in position -1 must be small, i.e. either Ala, Ser, Gly, Cys, Thr, or Gln; the residue in position -3 must not be aromatic (Phe, His, Tyr, Trp), charged (Asp, Glu, Lys, Arg), or large and polar (Asn, Gln). Further, it was suggested that Pro must be absent from positions -3 through +1. The new amino acid counts presented in Tables 1 & 2 are based on more than twice as many sequences; nevertheless, the (-3,-1)-rule is seen to hold remarkably well. The only exceptions found to date among eukaryotic proteins are one sequence with Leu in -1, one with Pro in -2, and three with Pro in -1. Thus, barring sequencing errors, we must

Table 1 **Amino acid counts for eukaryotic signal sequences**
The average composition (last column) is from Ref.(10)

	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	Expected
A	16	13	14	15	20	18	18	17	25	15	47	6	80	18	6	14.5
C	3	6	9	7	9	14	6	8	5	6	19	3	9	8	3	4.5
D	0	0	0	0	0	0	0	0	5	3	0	5	0	10	11	8.9
E	0	0	0	1	0	0	0	0	3	7	0	7	0	13	14	10.0
F	13	9	11	11	6	7	18	13	4	5	0	13	0	6	4	5.6
G	4	4	3	6	3	13	3	2	19	34	5	7	39	10	7	12.1
H	0	0	0	0	0	1	1	0	5	0	0	6	0	4	2	3.4
I	15	15	8	6	11	5	4	8	5	1	10	5	0	8	7	7.4
K	0	0	0	1	0	0	1	0	0	4	0	2	0	11	9	11.3
L	71	68	72	79	78	45	64	49	10	23	8	20	1	8	4	12.1
M	0	3	7	4	1	6	2	2	0	0	0	1	0	1	2	2.7
N	0	1	0	1	1	0	0	0	3	3	0	10	0	4	7	7.1
P	2	0	2	0	0	4	1	8	20	14	0	1	3	0	22	7.4
Q	0	0	0	1	0	6	1	0	10	8	0	18	3	19	10	6.3
R	2	0	0	0	0	1	0	0	7	4	0	15	0	12	9	7.6
S	9	3	8	6	13	10	15	16	26	11	23	17	20	15	10	11.4
T	2	10	5	4	5	13	7	7	12	6	17	8	6	3	10	9.7
V	20	25	15	18	13	15	11	27	0	12	32	3	0	8	17	11.1
W	4	3	3	1	1	2	6	3	1	3	0	9	0	2	0	1.8
Y	0	1	4	0	0	1	3	1	1	2	0	5	0	1	7	5.6

admit the possibility that residues other than the classical (-3,-1)-kinds can be used in position -1, but only when no better cleavage site is available in the vicinity (this is true for all five exceptions).

A few other points can also be made. First, the constraints on the prokaryotic sequences in the (-3,-1)-region seem even stronger than for the eukaryotic ones: only Ala, Gly, Ser and Thr have been found in -1, and only Ala, Gly, Leu, Ser, Thr, and Val in -3. Second, Leu is abundant in the prokaryotic sample up to and including position -8, but its incidence drops precipitously in position -7, where it is replaced by the likewise hydrophobic but less strongly helix-inducing residues Val and Phe. Only from position -6 do we find predominantely polar residues. Finally, there is a notable imbalance between the basic residues Arg and Lys in the c-region of the eukaryotic signal sequences, with 26 Arg and only 6 Lys (Arg/Lys = 4.3). This is in sharp contrast to the n-region where Arg/Lys = 66/72 = 0.9 and to proteins in general where the expected ratio is 0.6 (Table 1, last column).

Table 2 **Amino acid counts for prokaryotic signal sequences**
The average composition (last column) is from Ref.(10)

	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	Expected
A	10	8	8	9	6	7	5	6	7	7	24	2	31	18	4	3.2
C	1	0	0	1	1	0	0	1	1	0	0	0	0	0	0	1.0
D	0	0	0	0	0	0	0	0	0	0	0	0	0	2	8	2.0
E	0	0	0	0	0	0	0	0	0	0	0	1	0	4	8	2.2
F	2	4	3	4	1	1	8	0	4	1	0	7	0	1	0	1.3
G	4	2	2	2	3	5	2	4	2	2	0	2	2	1	0	2.7
H	0	0	1	0	0	0	0	1	1	0	0	7	0	1	0	0.8
I	3	1	5	1	5	0	1	3	0	0	0	0	0	0	2	1.7
K	0	0	0	0	0	0	0	0	0	1	0	2	0	3	0	2.5
L	8	11	9	8	9	13	1	0	2	2	1	2	0	0	1	2.7
M	0	2	1	1	3	2	3	0	1	2	0	4	0	0	1	0.6
N	0	0	0	0	0	0	0	1	1	1	0	3	0	1	4	1.6
P	0	1	1	1	1	1	2	3	5	2	0	0	0	0	5	1.7
Q	0	0	0	0	0	0	0	0	2	2	0	3	0	0	1	1.4
R	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1.7
S	1	0	1	4	4	1	5	15	5	8	5	2	2	0	0	2.6
T	2	0	4	2	2	2	2	2	5	1	3	0	1	1	2	2.2
V	5	7	1	3	1	4	7	0	0	4	3	0	0	2	0	2.5
W	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.4
Y	0	0	0	0	0	0	0	0	0	3	0	1	0	0	0	1.3

Construction of weight-matrices

Weight-matrix methods have been used for a number of years to locate signals in nucleic acid sequences (see (5) for a thorough discussion). Their use for pattern recognition in protein sequences requires a larger data base (20 amino acids rather than 4 bases must be scored for in each position), but is no different in principle. Basically, one converts the observed number of each kind of residue in each position in a sample of aligned "signals" into a measure of the probability of finding that particular kind of residue in that particular position - the probability weight-matrix - by a suitable normalization. Then, any new sequence can be scanned by a moving window (looking up the respective probabilities in the weight-matrix and multiplying together for each position of the window) to get a measure of the fit to the sample used in the construction of the weight-matrix. The highest-scoring window-position is then taken as the prediction for the location of the signal, if the score is above some minimum value.

To score for possible signal sequence function, and to locate the most probable cleavage site in a putative signal sequence, weight-matrices for prokaryotic and eukaryotic signal sequences were constructed as follows. The raw amino acid counts for the two samples (Tables 1 & 2) were divided by the expected number <N(a)> of each kind of residue given amino acid frequencies as in soluble proteins in general (last columns). Except for positions -3 and -1 relative to the cleavage site, all matrix elements with zero counts were normalized as $1/\langle N(a)\rangle$. For positions -3 and -1, where there is good reason both from previous statistical and experimental studies to believe that only a subset of all residues are allowed (2,6), the more stringent normalization 1/N was used for the zero-count elements (where N is the total number of sequences in the sample). The final weight-matrix was obtained by taking the natural logarithms of the normalized values, thus reducing the ensuing probability calculations to summations rather than multiplications of the weight-matrix elements.

Assessment of the predictive accuracy

When the two weight-matrices were used to predict the cleavage sites in the samples used in their construction, virtually all sites were correctly identified (87% in the eukaryotic sample, 100% in the prokaryotic sample). However, these sequences are at an advantage relative to sequences not included in the matrix: when correctly aligned with the weight-matrix, all residues in a sequence included in the weight-matrix sample will correspond to a count, and a spuriously high predictive accuracy will be found.

To avoid this problem, the eukaryotic sample was divided into 7 subsamples, each of 23 sequences. For each subsample, the remaining 138 sequences were used to construct a new weight-matrix, and this matrix was then applied to the subsample. Similarly, the prokaryotic sample was divided into 4 subsamples, each of 9 sequences. All subsequent calculations were carried out by summing the results for the subsamples.

Weight-matrices including positions -15 to +5 were first used to determine the effect of ignoring residues at either end in the predictions. It was found that positions -13 to +2 were sufficient to obtain maximal predictive accuracy (for the prokaryotic sample, positions -5 to +2 were sufficient but the full -13 to +2 range was used nevertheless): with this choice, 125 out of 161 eukaryotic and 32 out of 36 prokaryotic cleavage sites (78% and 89%) were correctly identified with a standard deviation of about ±10% in each case. For an additional 19 eukaryotic and 2 prokaryotic sequences, the correct site had the second-highest score. These values should

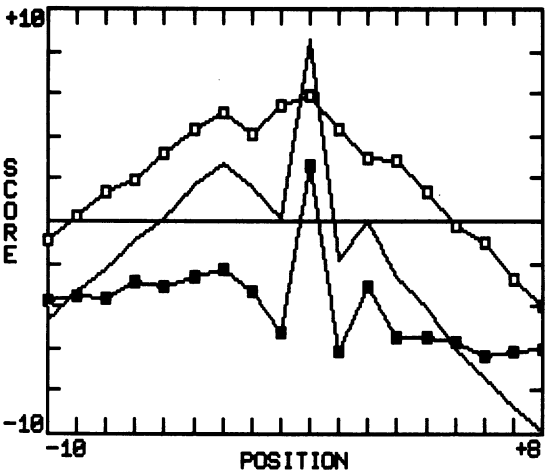

Figure 1 **Average h- and c-region scores as a function of the position of the moving window.** Open squares: h-region; solid squares: c-region; full line: total score.

be compared with the predictive accuracy of the older method (as implemented in a program kindly communicated by Dr. H.S. Ip, Rockefeller University). When this method was applied to the 121 sequences in the eukaryotic sample that were not included in the original statistics (2), 77/121 (64%) of the known cleavage sites were correctly identified, and only 17/36 (47%) of the prokaryotic ones were found.

With -13 to +2 weight-matrices, the contribution to the overall success from individual positions was also investigated. Only positions -3 and -1 had any strong impact; when one or the other was left out in the calculations the percentage of correctly identified eukaryotic sites dropped to 61% and 53%, respectively (81% and 69% for the prokaryotic sample).

As has been shown previously (1,7), residues -13 to -6 correspond to the h-region in the "average" eukaryotic signal sequence, residues -5 to -1 correspond to the c-region, and residues +1 and +2 seem to be selected such that few alternative cleavage sites should exist in the vicinity of the correct one (i.e. residues -5 to +2 can be included in an extended c-region). Thus, it is possible to calculate the scores for the h- and c-regions separately by summing the contributions from positions -13 to -6 and -5 to +2, respectively. As shown in Fig.1, the average h-region score for the eukaryotic sample increases slowly as the window is moved up to position -1 (the known cleavage site), and then decreases. The average c-region score

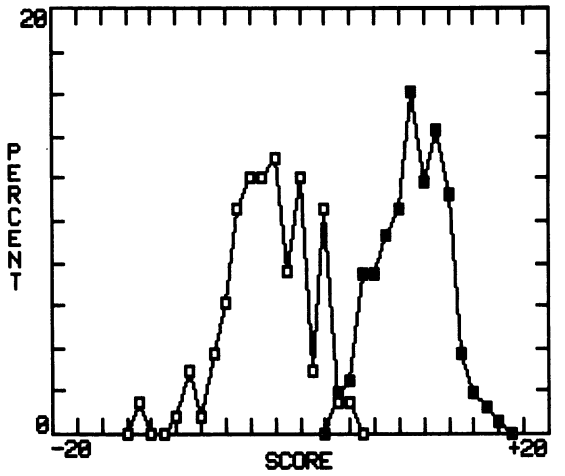

Figure 2 **Distribution of maximum scores for signal sequences and cytosolic proteins.** Open squares: cytosolic proteins; solid squares: signal sequences.

shows a more dramatic behaviour, with a pronounced peak in position -1 and troughs in positions -2 and +1, reflecting the match to the (-3,-1)-pattern and the tendency to have residues in position -2 that do not fit this pattern (see Tables 1 & 2). Similar curves are obtained for the prokaryotic sample (not shown).

Interestingly, 35 out of the 36 erroneous predictions for the eukaryotic sequences fall on the N-terminal side of the correct cleavage site, mostly in the region -6 to -3 (30/36). About half of these result from matches with a higher score in the h-region but a lower one in the c-region than calculated for the correct site, whereas only 6 out of 36 have higher c- and lower h-region scores than the correct site. I have thus tried to improve the predictive accuracy in various ways, e.g. by multiplying the -3 and -1 weigths or the whole c-region score by an extra factor, or by allowing a small variation in the distance between the h- and c-regions, but have not been able to obtain more than marginal improvements on the order of 2-4% in the overall success-rate.

The method described here not only allows prediction of the most likely cleavage site in new signal sequences, it also makes it possible to discriminate quite efficiently between putative signal sequences and the N-terminal regions of cytosolic proteins. The distribution of maximum scores for the eukaryotic signal sequences is shown in Fig.2, together with the

corresponding distribution obtained for a sample of 132 40-residues long N-teminal regions of cytosolic eukaryotic proteins (8). Only 3/161 (2%) of the signal sequences have maximum scores < 3.5; conversely, only 2/132 (2%) of the cytosolic sequences have maximum scores > 3.5. This level of discrimination compares favourably with that obtained with a recently published signal-sequence detecting algorithm (9).

DISCUSSION

Using a standard weight-matrix approach easily implemented even on a micro-computer, it is possible to set up a prediction method that (i) provides a clean discrimination between signal sequences and the N-terminal region in cytosolic proteins, and (ii) can be expected to identify the correct cleavage site 75-80% of the time when applied to new sequences not included in the data base (both prokaryotic and eukaryotic). This represents a significant improvement over previous methods.

Since the first submission of this work, another 36 eukaryotic signal sequences with known cleavage sites have been added to the data base. Using the same weight-matrix as above (Table 1), 75% of these sites were correctly predicted.

ACKNOWLEDGEMENT

This work was supported by a grant from the Swedish Natural Sciences Research Council.

REFERENCES

(1) von Heijne,G. (1985) J.Mol.Biol. 184, 99-105.
(2) von Heijne,G. (1983) Eur.J.Biochem. 133, 17-21.
(3) Perlman,D., and Halvorson,H.O. (1983) J.Mol.Biol. 167, 391-409.
(4) Mollay,C. (1985) in The Enzymology of Post-translational Modification of Proteins, Vol.2, pp. 1-23, Academic Press, London.
(5) Staden,R. (1984) Nuc.Acids Res. 12, 505-519.
(6) Kuhn,A., & Wickner,W. (1985) J.Biol.Chem. 260, 15914-15918.
(7) von Heijne,G. (1984) J.Mol.Biol. 173, 243-251.
(8) Flinta,C., Persson,B., Jörnvall,H., and von Heijne,G. (1986) Eur.J.Biochem. 154, 193-196.
(9) McGeoch,D.J. (1985) Virus Res. 3, 271-286.
(10) Klapper,H.M. (1977) Biochem.Biophys.Res.Commun. 78, 1018-1024.