The EMBO Journal vol.6 no.3 pp.823–831, 1987

# Signal peptide amino acid sequences in *Escherichia coli* contain information related to final protein localization. A multivariate data analysis

Michael Sjöström, Svante Wold, Åke Wieslander[1] and Leif Rilfors[2]

Research Group for Chemometrics, [1]Department of Biochemistry and [2]Department of Physical Chemistry, University of Umeå, S-901 87 Umeå, Sweden

Communicated by C.-I.Brändén

**With few exceptions, the signal peptides from proteins inserted into, or translocated through, the membranes of gram-negative bacteria or the endoplasmic reticulum of eukaryotes have no sequence homologies. Therefore these signal peptides have not been considered to contain information related to the different final localizations of the proteins. In this study, 43 signal peptide amino acid sequences from proteins with different final localizations in *Escherichia coli* have been subjected to a multivariate data analysis. Each amino acid residue was characterized by 20 physico-chemical properties, yielding a multivariate property profile for each peptide. The similarities/dissimilarities in the property profiles for the signal peptides from different classes were compared with each other by generating few-dimensional partial least squares (PLS) discriminant plots. With this approach, signal peptides from proteins localized to the periplasmic space (PS), the outer membrane (OM), and the extracellular surroundings (excreted proteins), were separated into distinct groups. Signal peptides from pili proteins were not separated from the OM signal peptides and only partly from the PS signal peptides, but were clearly different from the signal peptides of the excreted proteins. Signal peptides from inner membrane proteins were similar to those of the PS peptides. The size and the hydrophobicity of different peptide segments were responsible for the separation of the signal peptide classes. For example, the hydrophobicity of the N-terminal segment of the signal peptides increased with an increased distance from the cytoplasm of the final localization for the corresponding proteins. Thus, many signal peptides from proteins with different final localizations in *E. coli* have different discernible physico-chemical profiles.**

*Key words:* signal peptide/amino acid sequence/*Escherichia coli*/targeting information/multivariate data analysis

## Introduction

All eukaryotic and many prokaryotic cells have several membrane systems with compartments between the membranes. Synthesis of proteins usually takes place in the cytoplasmic domain, either on free or membrane-associated ribosomes. Thus some kind of targeting information is needed to direct the proteins to the non-cytoplasmic final localizations. A property common to most proteins which are inserted into a membrane, or translocated through one or several membranes, is an N-terminal peptide with ~15–30 amino acid residues, a so-called signal peptide (Perlman and Halvorson, 1983; von Heijne, 1985). The signal peptide schematically consists of three parts with respect to the amino acid sequence (Inouye and Halegoua, 1980), see Figure 1: (i) an N-terminal part with one or more basic amino acids; (ii) a middle part with mainly hydrophobic amino acids; and (iii) a C-terminal part, the last three amino acids of which define the signal peptidase recognition site (Perlman and Halvorson, 1983; von Heijne, 1983, 1984a). These features seem to be general for signal peptides of prokaryotic proteins and for eukaryotic proteins synthesized in the endoplasmic reticulum, but different from pre-sequences of proteins imported into mitochondria and chloroplasts (Hurt *et al.*, 1986). With few exceptions, sequence homologies between different signal peptides are not found (Magner, 1982; see below).

It is generally accepted that the signal peptide functions as a signal for the initial steps in the protein translocation. However, the signal peptide is not considered to contain information related to the final localization of the protein. Previously the signal pep-

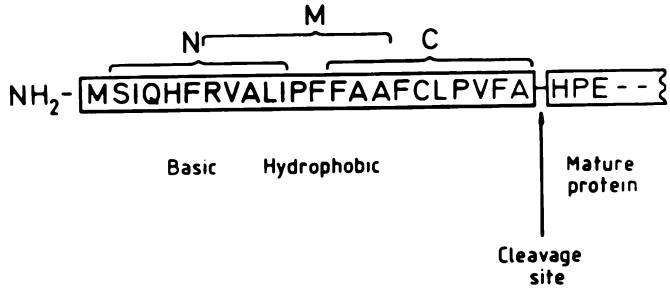


**Fig. 1.** Typical signal peptide from gram-negative bacteria. In the multivariate data analyses the signal peptide amino acid sequences were divided into three overlapping segments (N, M and C), each covering 10 residues. The middle block is shifted one amino acid position to the N-terminal part. This means that for none of the signal peptides the middle block will interfere with the four C-terminal amino acids defining the cleavage site. The signal peptide shown is β-lactamase (1 in Table I).

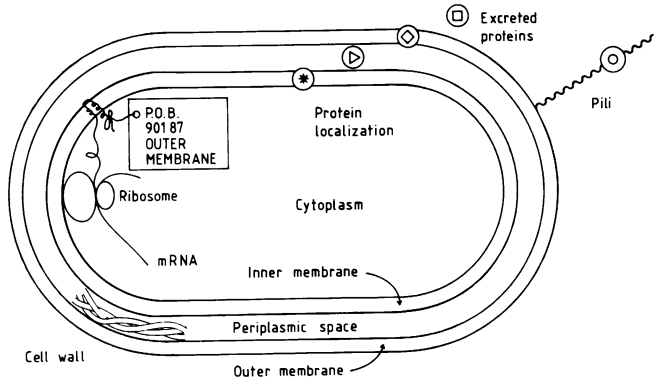


**Fig. 2.** Schematic figure showing the different non-cytoplasmic final localizations of proteins in *E. coli*. Secreted proteins are synthesized on cytoplasmic or membrane-associated ribosomes. Most of the proteins have cleavable signal peptides. The final localizations of the proteins are marked with different symbols. The same symbols are used in the PLS discriminant plots (Figures 3–9).

**Table I.** Amino acid sequences of signal peptides from *E. coli*

| | −20 | −10 | −1+1[a] | |
|---|---|---|---|---|
| **Periplasmic space proteins** | | | | |
| 1 β-lactamase Bla (plasmid) | M S I Q H F R | V A L I P F F A A F | C L P V F A | Ambler and Scott (1978) |
| 2 Alkaline phosphatase PhoA | M K Q S T I A | L A L L P L L F T P | V T K A | Kikuchi *et al.* (1981) |
| 3 Phosphate binding protein PhoS | M K V M R T T | V A T V V A A T L S | M S A F S V F A | Magota *et al.* (1984) |
| 4 Maltose binding protein MalE | M K I K T G A | R I L A L S A L T T | M M F S A S A L A | Bedoulle *et al.* (1980) |
| 5 Leu-spec. binding protein LivK | M K R N A K | T I I A G M I A L A | I S H T A M A | Oxender *et al.* (1980) |
| 6 Leu-Ile-Val binding protein LivJ | M N T K G K | A L L A G L I A L A | F S N M A L A | Landick and Oxender (1985) |
| 7 Galactose binding protein MglB | M N K K V L | T L S A V M A S M L | F G A A A H A | Scripture and Hogg (1983) |
| 8 Arabinose binding protein AraF | M H K F T K | A L A A I G L A A V | M S Q S A M A | Wilson and Hogg (1980) |
| 9 β-lactamase AmpC (chromosomal) | M F K T T L | C A L L I T A S C S | T F A | Jaurin *et al.* (1981) |
| 10 Ribose binding protein RbsB | M N M K K L | A T L V S A V A L S | A T V S A N A M A | Groarke *et al.* (1983) |
| 11 Mercuric resistance MerP | M K K L F | A S L A L A A V V A | P V W A | N.Brown, personal communication |
| 12 Cyclic phosphodiesterase CpdB | M I K F S | A T L L A T L I A A | S V N A | Liu *et al.* (1986) |
| **Outer membrane proteins** | | | | |
| 13 Braun's lipoprotein Lpp | M K A T K | L V L G A V I L G S | T L L A G | Inouye *et al.* (1977) |
| 14 λ-receptor LamB | M M I T L R | K L P L A V A V A A | G V M S A Q A M A | Hedgpeth *et al.* (1980) |
| 15 Outer membrane protein OmpA | M K K T A I | A I A V A L A G F | A T V A Q A | Movva *et al.* (1980) |
| 16 Outer membrane protein OmpC | M K V K V L | S L L V P A L L V | A G A A N A | Mizuno *et al.* (1983) |
| 17 Phosphate limitation protein PhoE | M K K S T L | A L V V M G I V A | S A S V Q A | Overbeeke *et al.* (1983) |
| 18 Outer membrane protein OmpF | M M K R N I | L A V I V P A L L | V A G T A N A | Mizuno *et al.* (1983) |
| 19 Tolerance protein TolC | M K K L L P | I L I G L S L S G | F S S L S Q A | Hackett and Reeves (1983) |
| 20 131 aa protein MolA | M K M N K S | L I V L C L S A G | L L A S A P G I S L A | Clement and Hofnung (1981) |
| 21 Vitamin B12 receptor BtuB | M I K K A | S L L T A C S V T | A F S A W A | Heller and Kadner (1985) |
| 22 Cloacin/aerobactin receptor IutA | M M I S K K | Y T L W A L N P L | L L T M M A P A V A | Krone *et al.* (1985) |
| **Pili proteins** | | | | |
| 23 PapA | M I K S V | I A G A V A M A V | V S F G V N N A | Båga *et al.* (1984) |
| 24 PapE1 | M K K I A | G L C L P V M L G | A V L M S Q H V H A | Lindberg *et al.* (1986) |
| 25 PapE2 | M K K I R | G L C L P V M L G | A V L M S Q H V H A | B.Lund and F.Lindberg, personal communication |
| 26 PapE3 | M K K I A | G L C L P V M L G | A V L M S Q H A H A | B.Lund and F.Lindberg, personal communication |
| 27 PapF1 | M I R L S | L F I S L L L T S | V A V L A | Lindberg *et al.* (1986) |
| 28 PapF2 | M A R L S | L F I S L L L T S | V A V L A | B.Lund and F.Lindberg, personal communication |
| 29 K88ab (gene A) | M K K A F | L L A C V F F L T | G G G V S H A | Mooi *et al.* (1984) |
| 30 K99 | M K K T L | L A I I L G G M A | F A T T N A S A | Roosendal *et al.* (1984) |
| 31 Type 1 FimA | M K I K T | L A I V V L S A L | S L S S T A A L A | Klemm (1984) |
| 32 $F7_2$ | M I K S V | I A G A V A M A V | V S F G A Y A | van Die and Bergmans (1984) |
| 33 Type 1 PilA | M K I K T | L A I V V L S A L | S L S S T T A L A | Orndorf and Falkow (1985) |
| **Excreted proteins (toxins)** | | | | |
| 34 Heat-labile toxin A (human) | M K N I T | F I F F I L L A S | P L Y A | Yamamoto *et al.* (1984) |
| Heat-labile toxin A (porcine) | M K N I T | F I F F I L L A S | P L Y A | Spicer and Noble (1982) |
| 35 Heat-labile toxin B (human) | M N K V K | C Y V L F T A L L | S S L C A Y G | Yamamoto *et al.* (1982) |
| 36 Heat-labile toxin B (porcine) | M N K V K | C Y V L F T A L L | S S L Y A H G | Dallas and Falkow (1980) |
| 37 Heat-stable toxin I | M K K L M | L A I F I S V L S | F P S F S Q S | So and McCarthy (1980) |
| 38 Heat-stable toxin II | −30 / M K K N I | A F L L A S M F V | F S I A T N A Y A | Lee *et al.* (1983) |
| 39 *Bacillus subtilis* amylase | M F A K R F K T S | L L P L F A G F L | L L F H L V L A G P A A A S | Yang *et al.* (1983) |
| **Inner membrane proteins** | | | | |
| 40 M13 major coat protein | M K K S L | V L K A S V A V A | T L V P M L S F A | Sugimoto *et al.* (1977) |
| 41 M13 minor coat protein | M K K L L | F A I P L V V P F | Y S H S | van Wezenbeek *et al.* (1980) |
| 42 Lipoprotein 28 | M K L T T | H H L R T G A A L | L L A G I L L A G | Yu *et al.* (1986) |
| 43 Penicillin binding protein DacA | M N T I F S | A R I M K R L A L | T T A L C T A F I S A A H A | Jackson *et al.* (1985) |

[a]The signal peptide amino acid sequences have been aligned from their cleavage sites. Cleavage *in vivo* takes place between residue −1, the C-terminal residue of the signal peptide, and residue +1, the N-terminal residue of the corresponding mature proteins.

**Table II.** Variables used to characterize the amino acids. The data and references are given in Hellberg *et al.* (1986)

| No. | Property |
|---|---|
| 1. | Mol. wt |
| 2. | $pK_{COOH}$ (COOH on $C_\alpha$) |
| 3. | $pK_{NH_2}$ ($NH_2$ on $C_\alpha$) |
| 4. | pI, pH at isoelectric point |
| 5. | Substituent van der Waals volume |
| 6. | $^1H$ n.m.r. for $C_\alpha-H$ (cation) |
| 7. | $^1H$ n.m.r. for $C_\alpha-H$ (dipolar) |
| 8. | $^1H$ n.m.r. for $C_\alpha-H$ (anion) |
| 9. | $^{13}C$ n.m.r. for C = O |
| 10. | $^{13}C$ n.m.r. for $C_\alpha-H$ |
| 11. | $^{13}C$ n.m.r. for C = O in tetrapeptide |
| 12. | $^{13}C$ n.m.r. for $C_\alpha-H$ in tetrapeptide |
| 13. | $R_f$ for 1-*N*-(4-Nitrobenzofurazono)-amino acids in paper chromatography |
| 14. | Slope of plot of $1/R_f - 1$ versus mol% in $H_2O$ in paper chromatography |
| 15. | d*G* of transfer of amino acid from organic solvent to water phase |
| 16. | Hydration potential of transfer from vapor phase to water |
| 17. | $R_f$, salt chromatography |
| 18. | Log *P*, *P* = partitioning coefficient for amino acids in octanol−water |
| 19. | Log *D*, *D* = partitioning coefficient at pH 7.11 for acetylamide derivatives of amino acids in octanol−water |
| 20. | $dG = RT \ln f$, *f* = fraction of buried to accessible amino acids in 22 proteins |

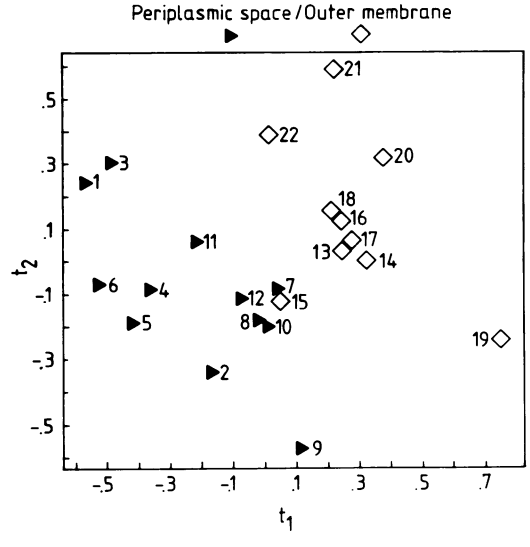


**Fig. 3.** PLS discriminant plot of the periplasmic space class versus the outer membrane class. The properties that contribute to $t_1$ are hydrophobicity of the N-segment (increasing to the right) and hydrophobicity and size of the C-segment (decreasing to the right) and $t_2$ contains information from size and hydrophobicity from the C-segment (increasing upwards). The latent variables $t_1$ and $t_2$ contain predictive information of the class assignments if OmpA (15) is not considered. The signal peptide sequence of OmpA is predicted to belong to the periplasmic space class.

tides have been investigated with respect to frequencies and localizations of hydrophobic, hydrophilic and charged amino acid residues. These studies were performed either with a selection of both eukaryotic and prokaryotic signal peptides (Perlman and Halvorson, 1983), or as a comparison between prokaryotic and eukaryotic sequences (von Heijne, 1984a,b, 1985). So far no

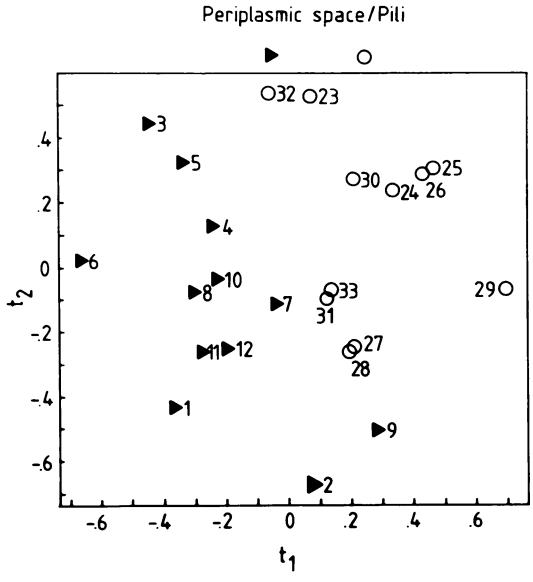


**Fig. 4.** PLS discriminant plot of periplasmic space class versus the pili class. The property that mainly contributes to $t_1$ is hydrophobicity of the N-segment (increasing) and to $t_2$ the size and hydrophobicity of the N- and C-segments (decreasing). The latent variables $t_1$ and $t_2$ contain predictive information of the class alignments.

systematic investigations have been done with signal peptides from a homogeneous system, such as one cell or organelle type. This is necessary in order to find out if signal peptides contain any discernible information that might be related to protein sorting. We have therefore performed an extensive multivariate data analysis of signal peptide sequences from the bacterium *Escherichia coli* to study the similarities/dissimilarities of signal sequences of proteins from different final localizations. The final localizations of the proteins are the inner membrane, the periplasmic space, the outer membrane and the extracellular surroundings (Figure 2). A fifth class is formed by the pili proteins. *E. coli* was chosen because of the large number of signal peptide sequences determined. The signal peptides studied are given in Table I, and the selection of signal peptides is discussed in Materials and methods.

The analysis of the signal peptide sequences includes the following three steps: (i) A multivariate characterization of each signal peptide. This is based on the physico-chemical properties of the amino acids (Table II) and their order in the signal peptide, and gives a multivariate property profile for each signal peptide. (ii) The similarities/dissimilarities in the property profiles for the signal peptides from different classes are calculated. This is done by generating few-dimensional plots, so called partial least squares (PLS) discriminant plots, with the multivariate projection method PLS. (iii) Identification of the class discriminating properties. The steps (i)−(iii) are described in more detail in Materials and methods.

## Results

### *Differences between signal peptide classes*

The cleavable signal peptides from inner membrane proteins are too few to define a class in the multivariate analysis (see Materials and methods). Plots of the two latent variables from pairwise analyses of signal peptide amino acid sequences from the other four classes reveal characteristic features for each of these classes

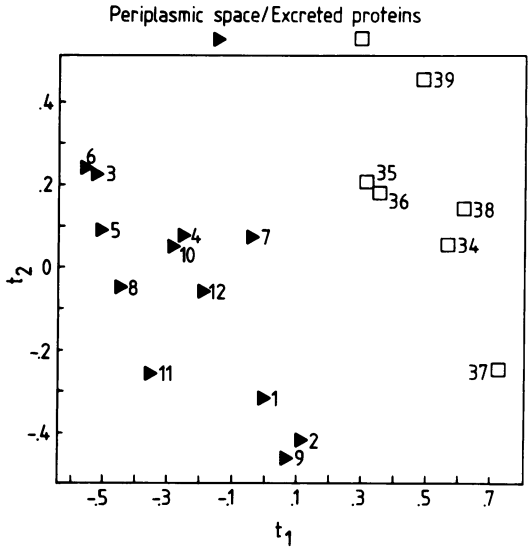


**Fig. 5.** PLS discriminant plot of the periplasmic space class versus the excreted protein class. The properties that contribute to $t_1$ are size and hydrophobicity of the N- and M-segments (increasing) and to $t_2$ the hydrophobicity of the N-segment and the $\beta$-branching of the C-segment (increasing). The latent variables $t_1$ and $t_2$ contain predictive information of the class assignments.

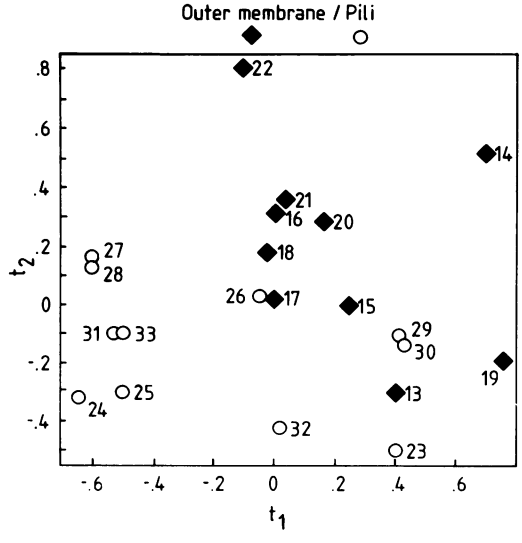


**Fig. 6.** PLS discriminant plot of the outer membrane class versus the pili class. None of the latent variables $t_1$ and $t_2$ contain predictive information about the class assignments.

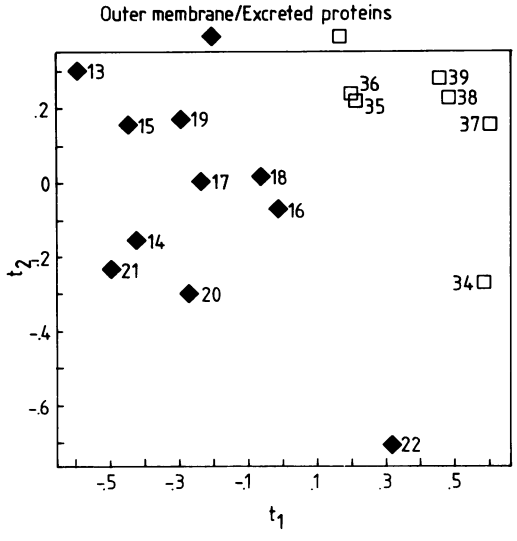


**Fig. 7.** PLS discriminant plot of the outer membrane class versus the excreted protein class. The properties that contribute to $t_1$ are the size and hydrophobicity of the N- and M-segments (increasing) and to $t_2$ the hydrophobicity and size of the N- and M-segments (decreasing). The latent variables $t_1$ and $t_2$ contain predictive information of the class assignments.

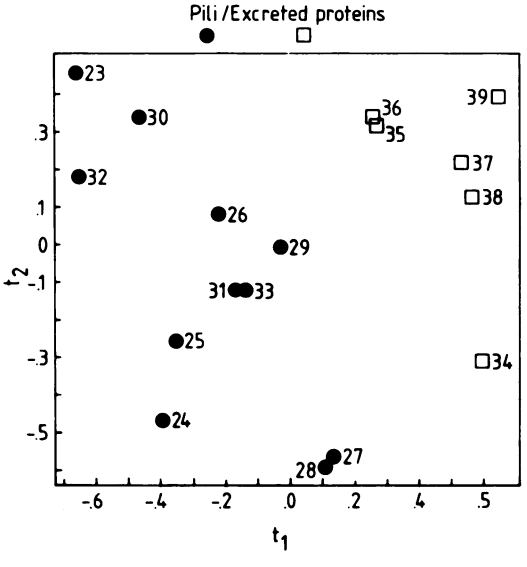


**Fig. 8.** PLS discriminant plot of the pili class versus the excreted protein class. The properties that contribute to $t_1$ are the size and hydrophobicity of the M-segment (increasing) and to $t_2$ the size and hydrophobicity of the C-segment (decreasing). The latent variables $t_1$ and $t_2$ contain predictive information of the class assignments.

(Figures 3–9). The two first latent variables $t_1$ and $t_2$ describe well the main variances in the physico-chemical properties of the amino acid residues in the three peptide segments (Figure 1).

Comparisons of the signal peptides from periplasmic space proteins with signal peptides from outer membrane proteins and excreted proteins, reveal clear groupings of the different classes (Figures 3 and 5). However, the signal peptides of the pili proteins partly overlap with the signal peptides of the periplasmic space proteins (Figure 4). The outer membrane and excreted protein classes are separated from the periplasmic one especially by the properties of their N-segment ($t_1$ in Figures 3 and 5). This segment is usually more hydrophobic in the signal peptides from the outer membrane, pili and excreted proteins. In addition, differences in hydrophobicity and size of the amino acid residues for the C- and M-segments are equally important in separating the outer membrane and excreted proteins from the periplasmic ones (Figures 3 and 5). The hydrophobicity and the size of the amino acid residues in the C-segment, and to a lesser extent the N-segment, give the largest contribution to the variation in the latent variable $t_2$. Hence, the separation of different classes is determined by the physico-chemical properties of more than one part (segment) of the signal peptides.

Signal peptides from outer membrane and pili proteins form overlapping clusters (Figure 6). However, signal peptides from excreted proteins are clearly different from outer membrane protein signal peptides (Figure 7). The hydrophobicity and the size of the amino acid residues in the N- and M-segments are the class-separating properties. Likewise, these properties are the class-separating ones for signal peptides from pili and excreted pro-

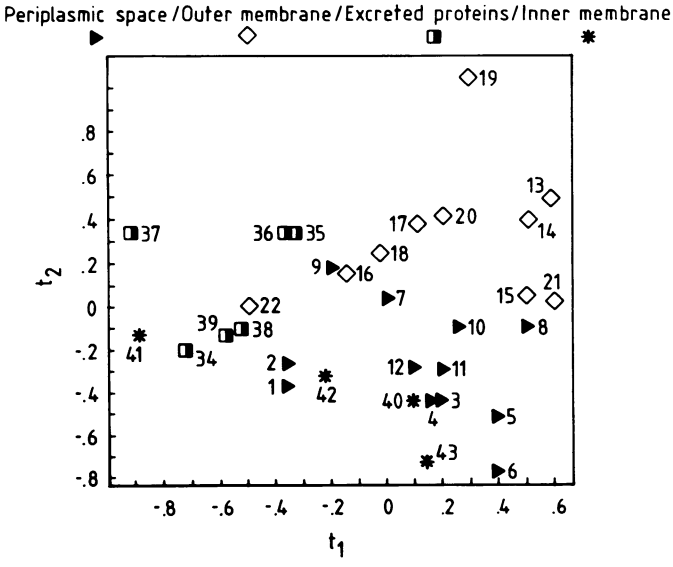


**Fig. 9.** PLS discriminant plot of the periplasmic space class versus outer membrane class versus excreted protein class. The inner membrane class is projected down on the $t_1$ versus $t_2$ plane. The properties that contribute to $t_1$ are the size and hydrophobicity of the N- and M-segments (decreasing) and to $t_2$ the hydrophobicity of the N-segment (increasing) and the size and hydrophobicity of the C-segment (decreasing).

teins (Figure 8), but in this case the properties are confined to the M- and C-segments.

When signal peptides from the inner membrane, periplasmic space, outer membrane and excreted proteins were analysed together, a fair separation of the three latter classes was achieved with only two latent variables (Figure 9). In line with the results from the pairwise analyses, the size and the hydrophobicity of the amino acid residues in the N- and M-segments, and in the N- and C-segments, are the properties that determine the variation in the two latent variables $t_1$ and $t_2$, respectively. Surprisingly, three of the inner membrane signal peptides clearly mapped within the periplasmic class; the fourth peptide was located among the signal peptides from the excreted proteins.

In conclusion, the hydrophobicity of the N-segment of the signal peptides increased with an increased distance from the cytoplasm of the final localization for the corresponding proteins. Moreover, the size of the N-segment and the size and hydrophobicity of the M-segment were larger for the signal peptides from excreted proteins than for signal peptides from proteins confined to the cell envelope.

## Discussion

*Periplasmic versus outer membrane, pili and excreted proteins*

When signal peptides from periplasmic space proteins are compared with signal peptides from outer membrane proteins, pili proteins or excreted proteins, the hydrophobicity of the N-segment is the main differentiating property ($t_1$) in Figures 3–5. The size and the hydrophobicity of the amino acid residues in the C-segment are additional primary separating properties in the plot with periplasmic space and outer membrane signal peptides (Figure 3). This is the only example in all analyses made where the C-segment contributes significantly to the variation in the latent variable $t_1$.

To visualize the differences in physico-chemical properties found by the multivariate analysis to be important for class separation (see Materials and methods), the hydrophobicity and the side chain size of selected signal peptide segments from all plots were estimated (Table III). It is clearly shown that the segments from signal peptides, that are well separated in the plot, have substantially different calculated values of hydrophobicity and size. Large variations in total hydrophobicity were also recently calculated for a large set of eukaryotic signal peptides (von Heijne, 1986).

Several other conclusions can be drawn from these multivariate data analyses. Firstly, the class separation is not correlated to the mol. wt of the mature proteins, since proteins of similar weights are distributed between different classes. Secondly, the signal peptide of Braun's lipoprotein (13 in Table I) clearly belongs to the outer membrane class, although the protein is cleaved by the signal peptidase II, not acting on the other outer membrane proteins. Moreover, lipoprotein-28 from the inner membrane (42 in Table I), although having the same signal peptidase cleavage sequence as Braun's lipoprotein (LLAG-C), does not map within the outer membrane class. Thirdly, the signal peptides of the seven binding proteins from the periplasmic space group (3–8 and 10 in Table I) map as a cluster within the periplasmic space class in Figures 4 and 5. The mature proteins fulfil similar functions, and have similar size and three-dimensional structures (Kubena *et al.*, 1986). Among these proteins the four sugar-binding proteins (4, 7, 8, and 10 in Table I) form a subcluster (Figures 3–5). These observations indicate that a correlation exists between the properties of the signal peptides and the mature proteins, or that these proteins have a common ancestral gene. Fourthly, according to present hypotheses, secreted toxins are considered to make a short halt in the periplasmic space before being excreted from the cell (Hofstra and Witholt, 1984; Hirst *et al.*, 1984; Mackman *et al.*, 1986). However, their signal peptides are different from the signal peptides of periplasmic space proteins (Figure 5). Fifthly, the signal peptide of the outer membrane protein OmpA (15 in Table I) maps in the periplasmic space class (Figure 3). Although being a true outer membrane protein it has several properties that differ from other outer membrane proteins (Mizushima, 1985). Moreover, processed OmpA is temporarily localized in the periplasmic space before insertion into the outer membrane (Freudl *et al.*, 1986).

*Outer membrane versus pili and excreted proteins*

When signal peptides from outer membrane and excreted proteins are compared (Figure 7), the protein OmpA signal peptide clearly maps within the outer membrane class (see above). The receptor for aerobactin and the toxin cloacin (22 in Table I) maps outside the outer membrane class, and has a value of the latent variable $t_1$ similar to those of the excreted proteins (toxins). It is striking that the long signal peptide of amylase from the gram-positive bacterium *Bacillus subtilis* (39 in Table I), cloned and excreted in *E. coli*, maps among the signal peptides of *E. coli* excreted proteins (Figures 5 and 7–9). This *Bacillus* signal peptide can also achieve secretion in *E. coli* of the periplasmic space protein β-lactamase to the extracellular surroundings (Nakazawa *et al.*, 1986).

Certain pili polymer structures are assembled from their monomer subunits in the outer membrane (Gaastra and De Graaf, 1982). It is therefore reasonable that the signal peptides of pili and outer membrane proteins form overlapping clusters (Figure 6), i.e. they have similar physico-chemical properties. A protein subunit of pili K88ab (protein p26), with a mol. wt of 26 000, is considered to make a halt and interact with other proteins in the periplasmic space before being assembled into the pili structure (Mooi and De Graaf, 1985). The signal peptide of this protein maps within the periplasmic space class in our calculations (data not shown). It would be interesting to in-

**Table III.** Hydrophobicity and size of segments from signal peptides selected from Figures 3–8. The main properties of $t_1$ responsible for the separation of the signal peptide classes are shown for well-separated peptides from the pairwise plots

| Plot | Properties[a] | Signal peptide[b] | Segment sequence[c] | | Average hydrophobicity[d] | | Side chain volume (Å$^3$)[e] | |
|---|---|---|---|---|---|---|---|---|
| | | | N | C | N | C | | C |
| Figure 3 | Hydrophobicity of N-segment (+) and hydrophobicity and size of C-segments (−) | 1 | S I Q H F R V A L I | F A A F C L P V F A | +0.96 | +2.27 | | 344 |
| | | 19 | K K L L P I L I G L | L S G F S S L S Q A | +1.44 | +0.51 | | 278 |
| | | | N | | N | | | |
| Figure 4 | Hydrophobicity of N-segment (+) | 6 | N T K G K A L L A G | | −0.16 | | | |
| | | 29 | K K A F L L A C V F | | +1.57 | | | |
| | | | N | M | N | M | N | M |
| Figure 5 | Hydrophobicity and size of N- and M-segments (+) | 6 | N T K G K A L L A G | A L L A G L I A L A | −0.16 | +2.65 | 307 | 280 |
| | | 37 | K K L M L A I F I S | L A I F I S V L S F | +1.45 | +2.66 | 434 | 371 |
| | | | N | M | N | M | N | M |
| Figure 7 | Hydrophobicity and size of N- and M-segments (+) | 13 | K A T K L V L G A V | L V L G A V I L G S | +1.07 | +2.45 | 339 | 285 |
| | | 37 | K K L M L A I F I S | L A I F I S V L S F | +1.45 | +2.66 | 434 | 371 |
| | | | | M | | M | | M |
| Figure 8 | Hydrophobicity and size of M-segment (+) | 32 | | I A G A V A M A V V | | +2.58 | | 250 |
| | | 34 | | I T F I F F I L L A | | +3.06 | | 432 |

[a]Properties of $t_1$ (+, increasing or −, decreasing) responsible for separation of the signal peptide classes according to the legends of Figures 3–8.
[b]Peptides well-separated in the plots (Figures 3–8) are compared. High number objects in the pairs have high values of $t_1$.
[c]Segment sequences (overlapping) of the signal peptides according to Table I. Note the deletion of methionine (M) at all N-termini.
[d]Average hydrophobicity according to Kyte and Doolittle (1982). In this concensus scale more positive figures correspond to an increased hydrophobicity.
[e]Van der Waals volumes of amino acid side chain substituents (Seydel and Schaper, 1979).

vestigate if the signal peptides of the pili proteins, which map close to the periplasmic class (Figure 4), have a similar route.

We note from Figure 8 that even single changes of amino acid residues in the signal peptide of Pap E1 (compare 24–26) are detected by the multivariate data analysis employed. These three closely related proteins are from clinical isolates of uropathogenic *E. coli*, and have certain differences in amino acid composition (B.Lund and F.Lindberg, personal communication). It is tempting to propose that the point changes in the signal peptides are related to the changes in the amino acid composition of the mature proteins.

*Inner membrane proteins*

The four cleavable inner membrane signal peptides mapped together with signal peptides from non-cytoplasmic water-soluble proteins, i.e. periplasmic space and excreted proteins (Figure 9). These inner membrane proteins are anchored in the membrane, but have substantial hydrophilic parts exposed towards the periplasmic space (Wickner, 1975; Davis *et al.*, 1985; Doherty *et al.*, 1986; Pratt *et al.*, 1986). It can be proposed that the signal peptides have the function to facilitate the transfer of certain parts of the proteins through the inner membrane into the periplasmic space. This process is probably halted by membrane-anchoring sequences in the proteins. Stretches of hydrophobic amino acids (stop sequences) have been found in three of these proteins (Sugimoto *et al.*, 1977; van Wezenbeek *et al.*, 1980; Yu *et al.*, 1986). Deletion of the C-terminal stop sequence of the minor coat protein (from strain f1) transforms this membrane protein to a periplasmic one (Davis *et al.*, 1985; Davis and Model, 1985). When the C-terminal membrane anchoring sequence of the fourth protein, Dac A, is removed the truncated protein is released into the periplasmic space (Pratt *et al.*, 1986). These results support the proposition above. The proposal is also similar to the mechanism asserted for mitochondrial inner membrane proteins (Hurt and van Loon, 1986). The matrix-targeting presequence mediates the transfer of some parts of the protein through the mitochondrial inner membrane until the process is halted by a stop sequence.

*Signal peptide exchanges*

To investigate if the final localization of the non-cytoplasmic proteins is completely determined by their signal peptides, a number of these peptides have been tested for their ability to translocate foreign proteins to the same final localization as the natural protein. The signal peptide and the nine amino acid residues from the N-terminal end of Braun's lipoprotein are sufficient for the localization of the periplasmic protein β-lactamase to the outer membrane of *E. coli* (Ghrayeb and Inouye, 1984). This hybrid protein, just like Braun's lipoprotein, is modified with three acyl chains, which most likely is responsible for the anchorage in the outer membrane. The signal peptide of the outer membrane protein OmpA (15) can only achieve secretion of the water-soluble proteins β-lactamase and nuclease A (*Staphylococcus aureus*) to the periplasmic space, but not to the extracellular surroundings (Ghrayeb *et al.*, 1984; Takahara *et al.*, 1985). However, the signal peptide from the outer membrane protein OmpF (18) translocates the small human protein β-endorphin into the culture medium (Nagahari *et al.*, 1985). Of the signal peptides discussed above the one from OmpF maps closest to the class comprising the signal peptides from the excreted proteins (Figure 7) and the one from OmpA maps within the class comprising the signal peptides from the periplasmic space (Figure 3). The protein alkaline phosphatase, which is normally located in the periplasmic space, is secreted to this compartment when fused to a signal peptide from another periplasmic protein (β-lactamase) or to signal peptides from the outer membrane proteins Lam B and OmpF (Hoffman and Wright, 1985). The outer membrane protein OmpF is efficiently translocated into the outer membrane

when fused to the signal peptide of the inner membrane protein Dac A (43) (Jackson *et al.*, 1985), belonging to the periplasmic space class according to our calculations (Figure 9). This can be explained by the fact that the OmpF protein is not water-soluble, and probably inserts spontaneously into the outer membrane after arrival to the periplasmic space. Likewise, the outer membrane protein PhoE (17) fused with a periplasmic space protein signal peptide (Bla) is localized in the outer membrane (Tommassen *et al.*, 1983). These examples indicate that water-soluble proteins (periplasmic space and excreted) cannot be localized to the inner membrane or outer membrane by merely fusing them with signal peptides from membrane proteins; probably a hydrophobic domain is required for the proper membrane anchorage of a protein. Conversely, it does not seem possible to localize membrane proteins to a water-compartment (periplasmic space and extracellular surroundings) by attachment of signal peptides from water-soluble proteins.

*Relevance for protein secretion*

Very little is known about the molecular mechanisms for protein secretion. The significance of the observed differences between *E. coli* signal peptides for the final protein localization can thus only be a matter of speculation. It is possible that the observed differences simply reflect common ancestor genes, for the proteins from each compartment, that have diverged during the evolution. In such instances a fairly high degree of position-specific sequence homologies is usually observed. However, this is not the case for the signal peptides (Magner, 1982; Perlman and Halvorson, 1983; von Heijne, 1985).

The division of signal peptide sequences into three segments was necessary in order to detect differences between the peptide classes. Different segments were responsible for the separation of these classes. Hence, it is possible that information for different steps in the secretion processes is confined in different signal peptide segments. It has been suggested previously that proteins headed for different final localizations may take different routes through the cell envelope (Benson, 1985). This can involve at least two mechanisms. Firstly, receptors in the cell envelope, specific for each protein class, may recognize the appropriate information in the signal peptide segments. In *E. coli* a truncated periplasmic protein is able to impair synthesis and translocation of other periplasmic space proteins, but not proteins headed for the outer membrane (Hengge-Aronis and Boos, 1986). Secondly, the signal peptide segments from different classes have characteristic physico-chemical properties, which in a class-specific way may enable the peptides to reorganize the membrane for passage of their nascent proteins. For example, the differences in hydrophobicity and size for the N-, M- and C-segments shown in Table III are of significant magnitude. The amino acid side chain volumes for the segments shown vary up to 60%. Similar differences in the hydrophobic volume of membrane lipids often have a great influence on the aggregate structure (spheres, rods, lamellae), and thus the phase structure (lamellar, hexagonal, cubic), formed by lipid–water mixtures (Rilfors *et al.*, 1984). Differences in the size of signal peptides may therefore affect their ability to reorganize the membrane lipid

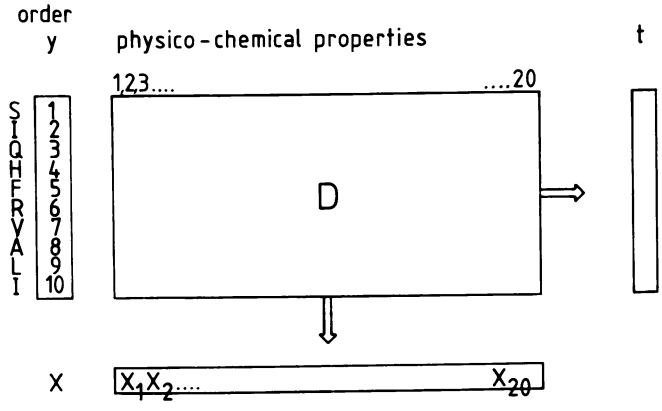


**Fig. 10.** PLS projection of the N-segment data matrix $D$ of $\beta$-lactamase on the order $y$ of the amino acids in the segment.

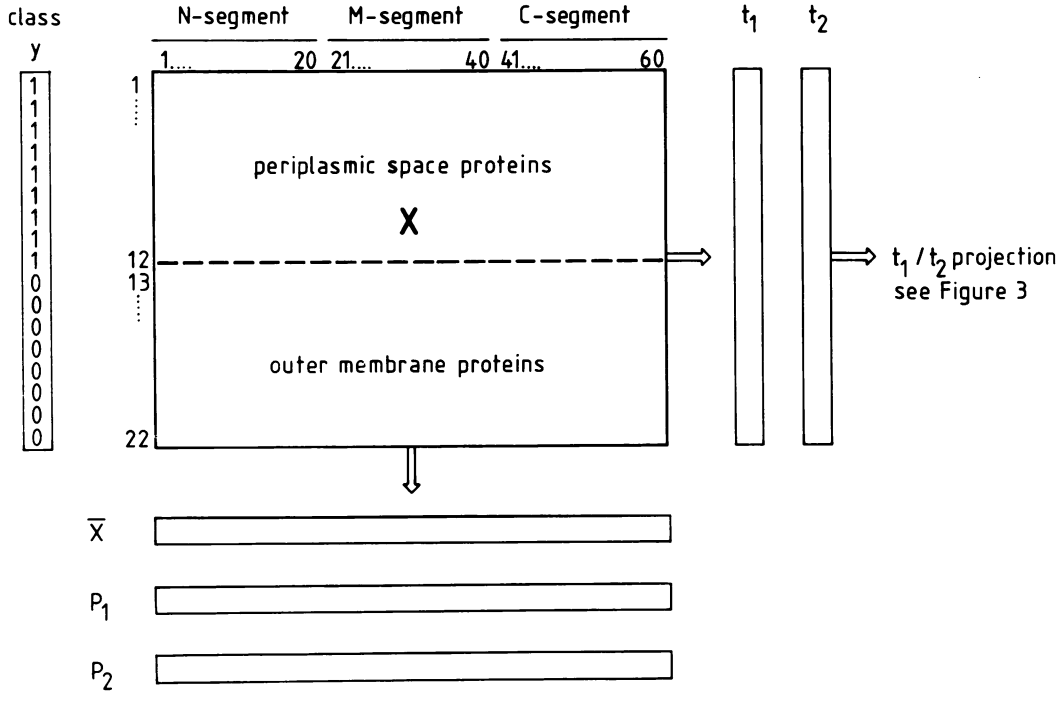


**Fig. 11.** PLS projection of the property profiles of the periplasmic space class versus the outer membrane class.

bilayer. Lipid reorganization was recently suggested to be the cause of the transfer of proteins between the inner and outer membrane (Yatvin *et al.*, 1986). In both the above-mentioned mechanisms, structural features of the mature proteins can be involved as well. For example, long stretches of hydrophobic amino acid residues can act as stop sequences for inner membrane proteins, whereas such sequences rarely occur in outer membrane proteins.

## Materials and methods

*Signal peptides from E. coli*

*E. coli* have four non-periplasmic final localizations for the proteins: the inner membrane, the periplasmic space, the outer membrane and extracellular surroundings (Figure 2). Pili proteins, involved in the attachment of gram-negative bacteria, constitute the largest known group of *E. coli* proteins with a similar function. However, with respect to the final localization, these proteins are appendages to the outer membrane. The signal peptide amino acid sequences used in the calculations are shown in Table I. The majority of these proteins have an established signal peptide cleavage site; a few proteins have predicted cleavage sites according to established rules (Perlman and Halvorson, 1983; von Heijne, 1983, 1984a). Proteins lacking cleavable signal peptides, or consensus cleavable sequences, have been omitted. This reduced the number of signal sequences especially for the inner membrane proteins (Table I). Pili with the protein subunit monomers of similar mol. wt (16–18 kd) were chosen.

*Statistical methods*

As mentioned in the introduction the statistical analysis can be divided into three steps: (i) a multivariate characterization of the signal peptide sequences; (ii) calculation of PLS discriminant plots; (iii) identification of the class discriminating properties.

In the approach below we have utilized the PLS model developed in statistics by H.Wold and co-workers. (Wold,H., 1975; Wold,S. *et al.*, 1984). However, the PLS model is also described in detail in the chemical literature (see e.g. Lindberg *et al.*, 1983).

*Multivariate characterization.* To be able to study a set of signal peptide sequences with multivariate methods, each sequence must be characterized with the same set of variables. Previously we have used a multipositional description of peptide sequences in quantitative structure–activity studies (Hellberg *et al.*, 1986). However, in the present case, where the length of the signal sequences varies considerably (see Table I), we have developed a new approach to generate a multivariate description of the sequences where the variables are the same in different sequences. This approach we have denoted 'fields of amino acid sequences'.

With the field approach each amino acid in a sequence is described multivariately. Here we have used 20 properties earlier collected for the 20 coded amino acids. The properties in the 20 × 20 matrix were first scaled and centered to unit variance and averaged to zero. The properties have been collected with the aim to illuminate different physico-chemical aspects of the amino acids such as size, hydrophobicity, lipophilicity, electronic properties and so on. These properties have earlier been used in a study of the relationship of the amino acid properties and the genetic code (Sjöström and Wold, 1986) and in quantitative structure–activity (QSAR) studies for short peptides (Hellberg *et al.*, 1986). In addition, multivariate descriptions of amino acid properties followed by the use of multivariate projection methods have been used in connection to QSAR studies of peptides (Sneath, 1966) and with the aim to predict tertiary protein structures (Kubota *et al.*, 1982; Kidera *et al.*, 1985).

A peptide sequence is represented by a $n \times 20$ matrix, where $n$ is the number of amino acids in the signal peptide. With the multivariate projection method PLS, the whole data set or part of the data set (as in the present case, see below) is projected by means of a vector describing the sequence order $y$ (see Figure 10). This will result in contraction of the matrix $D$ to two vectors $t$ and $x$. The row vector $x$ contains 20 elements describing the influence of each of the entry variables. A variable value in the $x$ row vector (see e.g. Figure 11) contains mainly information of the average value of an original variable but also some information about the property trends of the amino acids in the sequence. These confounded effects are not directly separated. The column vector $t$ contains elements which are position dependent and is not further used. In the following we have used the vector $x$ as a description of the properties of a sequence. To get a more comprehensive description of a sequence, we have divided each signal peptide sequence into three overlapping segments with 10 amino acids in each (Figure 1). So far no efforts to further improve the description have been made, but this will be dealt with in future investigations. Then an $x$ vector is calculated for each segment. Thus in the present case we have reduced the data matrix as $D$ in Figure 10, which initially describes a signal peptide sequence, to a property profile represented by a row vector with 60 elements for each signal peptide, see Figure 11.

*PLS discriminant plots.* Each signal peptide can now be thought of as a point in a 60-dimensional space, one dimension for each variable. Since the variables are the same for all signal peptide sequences, they can be represented in the same 60-dimensional space. Few-dimensional plots of this 60-dimensional space can be constructed with multivariate projection methods. One type of projection is a principal components score plot or eigenvector projection. A limitation with this projection method is that the information of the class assignments of the signal peptide sequences are not used. Here we have used the PLS projection method instead to extract few-dimensional plots (Wold *et al.*, 1986; Sjöström *et al.*, 1986). With PLS, the assigned class-labels (here inner membrane, periplasmic space, pili, excreted and outer membrane) can be used in the calculation of the projections. With PLS the main variance in the $X$ block (Figure 11) is well described and simultaneously a correlation with the class-labels is sought. The class to which the signal peptides belong is represented by a dummy variable $y$, in which one class of signal peptides are represented by 0 and the other class with 1, as in Figure 11. Thus the construction of the plots is based on the assumption of a correct assignment of the final localization of the proteins. Plots of the two first $t$ vectors from the PLS analysis against each other illustrate the similarities/dissimilarities of the signal peptide sequences within and between classes. Whether a new dimension, expressed by the vectors $t$ and $p$, contains predictive information or not with respect to the class assignments of the signal peptides is investigated by cross-validation. Here some of the signal peptides are kept out of the calculation of a PLS model. Predictions of the class assignments of the withheld signal peptides can then be calculated from the PLS model. In the present investigation all signal peptides have been deleted once for each projection dimension and thus have been used as internal controls. The details of cross-validation in PLS are presented elsewhere (Lindberg *et al.*, 1983; Wold *et al.*, 1984).

*Identification.* The loading vector $p$, which conains 60 elements, gives information of which variables span its corresponding $t$ vector. This information can be used to identify the class-discriminating properties in the signal peptides. Furthermore, the variables 1–20, 21–40 and 41–60 are related to the N-, M- and C-segments respectively. However, the 20 variables for one segment do not directly refer to the variables in Table II, since the data profile vector $x$ is obtained by a contraction of the data matrix $D$. The labels of the $t$-axes (given in the legends) only refer to the average values which seem to dominate. That this is the case is confirmed for the extreme signal peptides in the plots, see Table III. However, it is not sufficient to just calculate, for example, the average values of the discriminating variables to obtain a correct ordering of a set of signal peptides. This can only be achieved with the present multistep projection approach.

## Acknowledgements

Financial support from the Swedish Natural Science Research Council (NFR), the Swedish Council for Planning and Coordination of Research (FRN) and the National Swedish Board for Technical Development (STU) are gratefully acknowledged. We thank Drs B.-E.Uhlin and G.von Heijne for valuable comments, and F.Lindberg and B.Lund for the access to pili protein signal sequences prior to publication.

## References

Ambler,R.P. and Scott,G.K. (1978) *Proc. Natl. Acad. Sci. USA*, **75**, 3732–3736.
Båga,M., Normark,S., Hardy,J., O'Hanley,O., Lark,D., Olsson,O., Schoolnik,G. and Falkow,S. (1984) *J. Bacteriol.*, **157**, 330–333.
Benson,S.A. (1985) *BioEssays*, **2**, 201–205.
Bedouelle,H., Bassford,P.J.,Jr., Fowler,A.V., Zabin,I., Beckwith,J. and Hofnung,M. (1980) *Nature*, **285**, 78–81.
Clement,J.M. and Hofnung,M. (1981) *Cell*, **27**, 507–524.
Dallas,W.S. and Falkow,S. (1980) *Nature*, **288**, 499–501.
Davis,N.G., Boeke,J.D. and Model,P. (1985) *J. Mol. Biol.*, **181**, 111–121.
Davis,N.G. and Model,P. (1985) *Cell*, **41**, 607–614.
Doherty,H., Yamada,H., Caffrey,P. and Owen,P. (1986) *J. Bacteriol.*, **166**, 1072–1082.
Freudl,R., Schwarz,H., Stierhof,Y.-D., Gamon,K., Hindennach,I. and Henning,U. (1986) *J. Biol. Chem.*, **261**, 11355–11361.
Gaastra,W. and De Graaf,F.K. (1982) *Microbiol. Rev.*, **46**, 129–161.
Ghrayeb,J. and Inouye,M. (1984) *J. Biol. Chem.*, **259**, 463–467.
Ghrayeb,J., Kimura,H., Takahara,M., Hsiung,H., Masui,Y. and Inouye,M. (1984) *EMBO J.*, **3**, 2437–2442.
Groarke,J.M., Mahoney,W.C., Hope,J.N., Furlong,C.E., Robb,F.T., Zalkin,H. and Hermodson,M.A. (1983) *J. Biol. Chem.*, **258**, 12952–12956.
Hackett,J. and Reeves,P. (1983) *Nucleic Acids Res.*, **11**, 6487–6495.
Hedgpeth,J., Clement,J.-M., Marchal,C., Perrin,D. and Hofnung,M. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 2621–2625.
Hellberg,S., Sjöström,M. and Wold,S. (1986) *Acta Chem. Scand.*, **B40**, 135–140.
Heller,K. and Kadner,R.J. (1985) *J. Bacteriol.*, **161**, 904–908.

Hengge-Aronis,R. and Boos,W. (1986) *J. Bacteriol.*, **167**, 462–466.
Hirst,T.R., Sanches,J., Kaper,J.B., Hardy,S.J.S. and Holmgren,J. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 7752–7756.
Hoffman,C.S. and Wright,A. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 5107–5111.
Hofstra,H. and Witholt,B. (1984) *J. Biol. Chem.*, **259**, 15182–15187.
Hurt,E.C. and van Loon,A.P.G.M. (1986) *Trends Biochem. Sci.*, **11**, 204–207.
Hurt,E.C., Soltanifar,N., Goldschmidt-Clermont,M., Rochaix,J.-D. and Schatz,G. (1986) *EMBO J.*, **5**, 1343–1350.
Inouye,M. and Halegoua,S. (1980) *CRC Crit. Rev. Biochem.*, **7**, 339–371.
Inouye,S., Wang,S., Sekizawa,J., Halegoua,S. and Inouye,M. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 1004–1008.
Jackson,M.E., Pratt,J.M., Stoker,N.G. and Holland,I.B. (1985) *EMBO J.*, **4**, 2377–2383.
Jaurin,B., Grundström,T., Edlund,T. and Normark,S. (1981) *Nature*, **290**, 221–225.
Kidera,A., Konishi,Y., Oka,M., Ooi,T. and Scheraga,H.A. (1985) *J. Protein Chem.*, **4**, 23–55.
Kikuchi,Y., Yoda,K., Yamasaki,M. and Tamura,G. (1981) *Nucleic Acids Res.*, **9**, 5671–5678.
Klemm,P. (1984) *Eur. J. Biochem.*, **143**, 395–399.
Krone,W.J.A., Stegehuis,F., Koningstein,G., van Doorn,C., Roosendaal,B., De Graaf,F.K. and Oudega,B. (1985) *FEMS Microbiol. Lett.*, **26**, 153–161.
Kubena,B.D., Luecka,H., Rosenberg,H. and Quiocho,F.A. (1986) *J. Biol. Chem.*, **261**, 7995–7996.
Kubota,Y., Nishikawa,K., Takahashi,S. and Ooi,T. (1982) *Biochim. Biophys. Acta*, **701**, 242–252.
Kyte,J. and Doolittle,R.F. (1982) *J. Mol. Biol.*, **157**, 105–132.
Landick,R. and Oxender,D.L. (1985) *J. Biol. Chem.*, **260**, 8257–8261.
Lee,C.H., Moseley,S.L., Moon,H.W., Whipp,S.C., Gyles,C.L. and So,M. (1983) *Infect. Immun.*, **42**, 264–268.
Lindberg,W., Persson,J.Å. and Wold,S. (1983) *Anal. Chem.*, **55**, 643–647.
Lindberg,F., Lund,B. and Normark,S. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 1891–1895.
Liu,J., Burns,D.M. and Beacham,I.R. (1986) *J. Bacteriol.*, **165**, 1002–1010.
Mackman,N., Nicand,J.-M., Ray,L. and Holland,I.B. (1986) *Current Top. Microbiol. Immunol.*, **125**, 159–181.
Magner,J.A. (1982) *J. Theor. Biol.*, **99**, 831–833.
Magota,K., Otsuji,N., Miki,T., Horiuchhi,T., Tsunisawa,S., Kondo,J., Sakiyama,F., Amemura,M., Morita,T., Shinagawa,H. and Nakata,A. (1984) *J. Bacteriol.*, **157**, 909–917.
Mizuno,T., Chou,M.-Y. and Inouye,M. (1983) *FEBS Lett.*, **151**, 159–164.
Mizushima,S. (1985) In Nanninga,N. (ed.), *Molecular Cytology of Escherichia coli*. Academic Press, London, pp. 39–75.
Mooi,F.R. and De Graaf,F.K. (1985) *Current Top. Microbiol. Immunol.*, **118**, 119–138.
Mooi,F.R., van Buuren,M., Koopman,G., Roosendaal,B. and De Graaf,F.K. (1984) *J. Bacteriol.*, **159**, 482–487.
Movva,F.R., Nakamura,K. and Inouye,M. (1980) *J. Biol. Chem.*, **255**, 27–29.
Nagahari,K., Kanaya,S., Munakata,K., Aoyagi,Y. and Mizushima,S. (1985) *EMBO J.*, **4**, 3589–3592.
Nakazawa,K., Takano,T., Sohma,A. and Yamane,K. (1986) *Biochem. Biophys. Res. Commun.*, **134**, 624–631.
Orndorf,P.E. and Falkow,S. (1985) *J. Bacteriol.*, **162**, 454–457.
Overbeeke,N., Bergmans,H., van Mansfeld,F. and Lutgenberg,B. (1983) *J. Mol. Biol.*, **163**, 513–532.
Oxender,D.L., Anderson,J.J., Daniels,C.J., Landick,R., Gunsalus,R.P., Zurawski,G. and Yanotsky,C. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 2005–2009.
Perlman,D. and Halvorson,H.O. (1983) *J. Mol. Biol.*, **167**, 391–409.
Pratt,J.M., Jackson,M.E. and Holland,I.B. (1986) *EMBO J.*, **5**, 2399–2405.
Rilfors,L., Lindblom,G., Wieslander,Å. and Christiansson,A. (1984) In Kates,M. and Manson,L.A. (eds), *Biomembranes, Vol. 12, Membrane Fluidity*. Plenum Press, New York, pp. 205–245.
Roosendal,B., Gaastra,W. and De Graaf,F.K. (1984) *FEMS Microbiol. Lett.*, **22**, 253–258.
Scripture,J.B. and Hogg,R.W. (1983) *J. Biol. Chem.*, **258**, 10853–10855.
Seydel,J.K. and Schaper,K.-J. (1979) *Chemische Struktur und Biologische Aktivität von Wirkstoffen*. Verlag Chemie, Weinheim.
Sjöström,M. and Wold,S. (1986) *J. Mol. Evol.*, **22**, 272–277.
Sjöström,M., Söderström,B. and Wold,S. (1986) In Gelsema,E.L. and Kanal,N.J. (eds), *Pattern Recognition in Practice*. North-Holland, Amsterdam, Vol. II, pp. 461–470.
So,M. and McCarthy,B.J. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 4011–4015.
Sneath,P.H.A. (1966) *J. Theor. Biol.*, **12**, 157–195.
Spicer,E.K. and Noble,J.A. (1982) *J. Biol. Chem.*, **257**, 5716–5721.
Sugimoto,K., Sugisaki,H., Okamoto,T. and Takanami,M. (1977) *J. Mol. Biol.*, **110**, 487–507.
Takahara,M., Hibler,D.W., Barr,P.J., Gerlt,J.A. and Inouye,M. (1985) *J. Biol. Chem.*, **260**, 2670–2674.
Tommassen,J., van Tol,H. and Lugtenberg,B. (1983) *EMBO J.*, **2**, 1275–1279.
van Die,I. and Bergmans,H. (1984) *Gene*, **32**, 83–90.
van Wezenbeek,P.M.G.F., Hulsebos,T.J.M. and Schoenmakers,J.G.G. (1980) *Gene*, **11**, 129–148.
von Heijne,G. (1983) *Eur. J. Biochem.*, **133**, 17–21.
von Heijne,G. (1984a) *J. Mol. Biol.*, **173**, 243–251.
von Heijne,G. (1984b) *EMBO J.*, **3**, 2315–2318.
von Heijne,G. (1985) *J. Mol. Biol.*, **184**, 99–105.
von Heijne,G. (1986) *J. Mol. Biol.*, **189**, 239–242.
Wickner,W.T. (1975) *Proc. Natl. Acad. Sci. USA*, **72**, 4749–4753.
Wilson,V.G. and Hogg,R.W. (1980) *J. Biol. Chem.*, **255**, 6745–6750.
Wold,H. (1975) In Gani,J. (ed.), *Perspectives in Probability and Statistics. Papers in Honour of M.S.Bartlett*. Academic Press, London.
Wold,S., Ruhe,A., Wold,H. and Dunn,W.J.,III (1984) *SIAM J. Sci. Stat. Comput.*, **5**, 735–743.
Wold,S., Albano,C., Dunn,W.J.,III, Edlund,U., Esbensen,K., Geladi,P., Hellberg,S., Johansson,S., Lindberg,W. and Sjöström,M. (1986) In Kowalski,B.R. (ed.), *Chemometrics-Mathematics and Statistics in Chemistry, NATO ASI Series. Series C, Mathematical and Physical Sciences no. 138.* Reidel, Dordrecht, pp. 17–96.
Yamamoto,T., Tamura,T.-A., Ryoji,M., Kaji,A., Yokata,T. and Takano,T. (1982) *J. Bacteriol.*, **152**, 506–509.
Yamamoto,T., Tamura,T.-A. and Takano,T. (1984) *J. Biol. Chem.*, **259**, 5037–5044.
Yang,M., Galizzi,A. and Henner,D. (1983) *Nucleic Acids Res.*, **11**, 237–249.
Yatvin,M.B., Smith,K.M. and Siegel,F.L. (1986) *J. Biol. Chem.*, **261**, 8070–8075.
Yu,F., Inouye,S. and Inouye,M. (1986) *J. Biol. Chem.*, **261**, 2284–2288.

*Received on November 17, 1986; revised on December 29, 1986*