Analytica Chimica Acta, 277 (1993) 239–253
Elsevier Science Publishers B.V., Amsterdam

DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures

S. Wold, J. Jonsson, M. Sjöström, M. Sandberg and S. Rännar

Research Group for Chemometrics, Umeå University S-901 87 Umeå (Sweden)

(Received 3rd September 1992)

Abstract

Biopolymer sequences (e.g., DNA, RNA, proteins and polysaccharides) and chemical processes (e.g., a batch or continuous polymer synthesis run in a chemical plant) have close similarities from the modelling point of view. When a set of sequences or processes is characterized by multivariate data, a three-way data matrix is obtained. With sequences the position and with processes the time is one direction in this matrix. The multivariate modelling of this matrix by principal component analysis (PCA) or partial least-squares (PLS) methods for the following purposes is discussed: classification of sequences; quantitative relationships between sequence and biological activity or chemical properties; optimizing a sequence with respect to selected properties; process diagnostics; and quantitative relationships between process variables and product quality variables. To obtain good models, a number of problems have to be adequately dealt with: appropriate characterization of the sequence or process; experimental design (selecting sequences or process settings); transforming the three-way into a two-way matrix; and appropriate modelling and validation (modelling interactions, periodicities, "time series" structures and "neighbour effects"). A multivariate approach to sequence and process modelling using PCA and PLS projections to latent structures is discussed and illustrated with several sets of peptide and DNA promoter data.

Keywords: Process analysis/on-line analysis; Pattern recognition; Biopolymer sequences; DNA sequences; Multivariate modelling; Partial least squares; Peptide sequences; Principal component analysis

With the increasing analytical abilities of biochemists, analytical chemists, microbiologists and molecular biologists, the sequences of peptides, proteins, DNA and other biopolymers are being determined at an increasing rate. To bring some order to these data and to use them for purposes such as the understanding of the evolutionary relationships between organisms and relationships between sequence and biological activity and chemical properties, various kinds of models are needed. With the complexity of sequences, i.e., many positions times several properties, these models will necessarily be multivariate [1,2].

Correspondence to: S. Wold, Research group for Chemometrics, Umeå University, S-901 87 Umeå (Sweden).

The monitoring and control of chemical processes is another area that is currently undergoing a data explosion. Sensors and on-line instrumentation provide multitudes of data characterizing the state of the process at any given point in time [3]. A common objective is to relate these time sequences of data to qualities of the resulting products of the process, such as yield, purity, crystal size, polymer strength and viscosity. Again, multivariate modelling provides the tools for handling these data analysis problems [4–6].

Interestingly, these two application areas of multivariate modelling show close similarities. A biopolymer sequence, such as a peptide or a DNA, has a univariate direction from beginning to end, analogous to a time sequence of process

data. The consecutive observations of a time sequence of process data are usually not independent. This has necessitated the use of special types of models, time series models [7], for the analysis of process data. It may be of interest to apply these models also to polymer sequences to investigate whether nature introduces "time series patterns" in the biological sequences such as neighbour effects, periodicities and auto-correlations.

The modelling of a set of polymer sequences characterized by a multivariate description of each position leads to the analysis of a three-way matrix (sequence × position × descriptor). The same situation appears in the modelling of a set of, say, batch process time sequences, with multivariate observations at each time point (process × time × variable). Before the analysis with standard methods, these three-way matrices must be transformed into two-way matrices, which can be done in different ways. These ways are not equivalent, but correspond to different assumptions about the underlying action mechanisms of the modelled sequences or processes.

In this paper some aspects of sequence and process modelling are discussed, in particular their common features. Examples of peptide and nucleic acid sequence models are used as illustrations.

SCOPE OF SEQUENCE AND PROCESS MODELLING

The first problem in any investigation is to collect quantitative and relevant data characteriz-

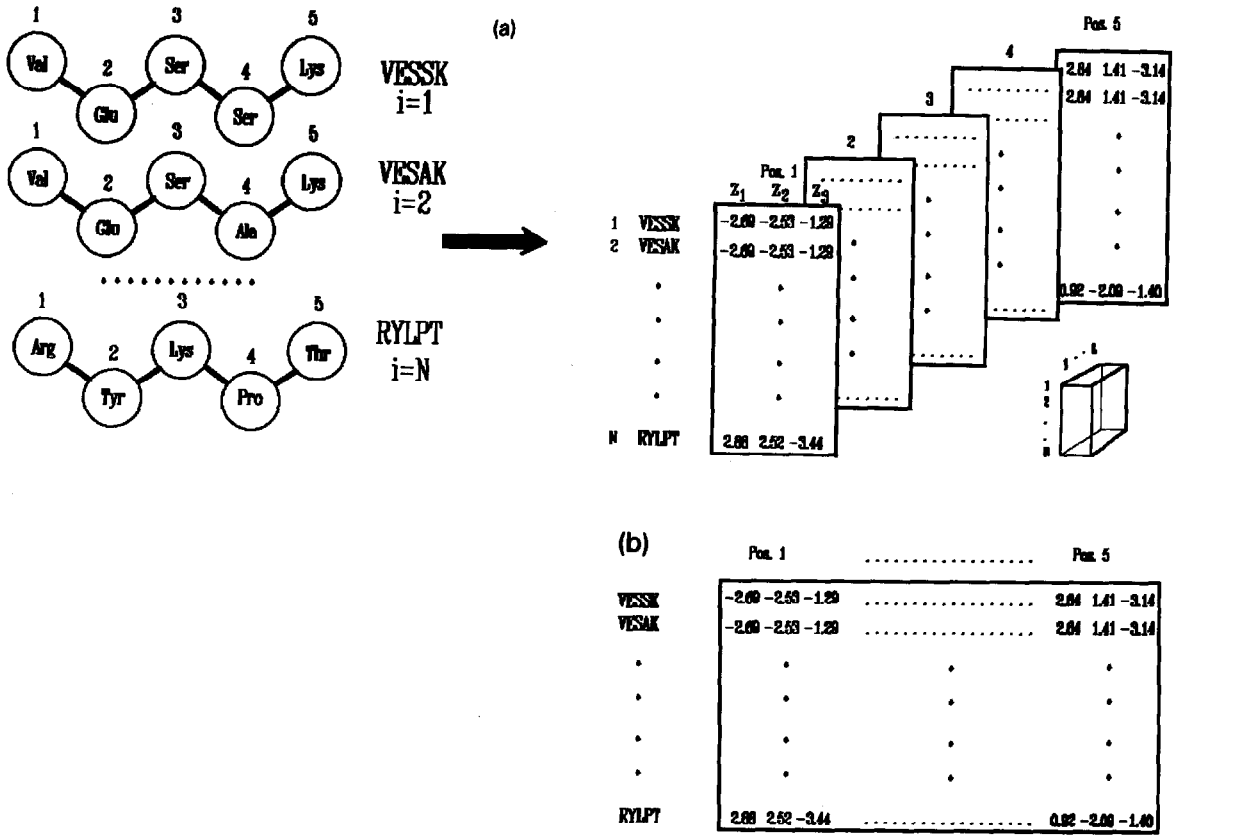

Fig. 1. (a) A set of biopolymer sequences can be quantified by using descriptor scales such as z_1, z_2 and z_3 for each varying position. The resulting descriptor data matrix (**X**) is three-way (sequence × position × descriptor). (b) In the case of sequences of the same length, the three-way table can be directly "unfolded" to give a two-way table with several columns for each sequence position.

ing the investigated objects (here sequences). This gives a data matrix in which each row corresponds to an object (sequence) and each column corresponds to a variable, calculated or measured on the objects. With sets of polymer or time sequences, this is often a three-way data matrix (Figs. 1 and 2), as will be discussed below.

Once quantitative and relevant data have been obtained, one needs to analyse the data, i.e., relate them to various more or less clearly formulated objectives. These objectives can be grouped into three broad categories: overview, classification and quantitative modelling.

Overview; finding groups, patterns

When little is known about the problem or system under investigation, the scope of the analysis is usually to obtain an overview of the objects in the data (here biopolymer or process time sequences). The result is a "map" of the similarities and dissimilarities of the objects, a model of the system. Sometimes groups of objects (classes) can be distinguished in this map, but often the objects are distributed over the map without any clear gaps. Other "patterns" may still be seen in the map, such as trends in time or other coordinates.

The analysis also informs on which variables (object characteristics) are related to directions in the map and other discernible patterns.

The data analysis tools useful for this type of analysis are scatter plots when the total number of variables, K, is smaller than ca. 5, and principal components analysis (PCA) and factor analysis (FA) when K is larger [8,9].

Diagnostics

The model obtained in PCA or FA can be used in a second stage of the data analysis to scrutinize new objects and their relationships to

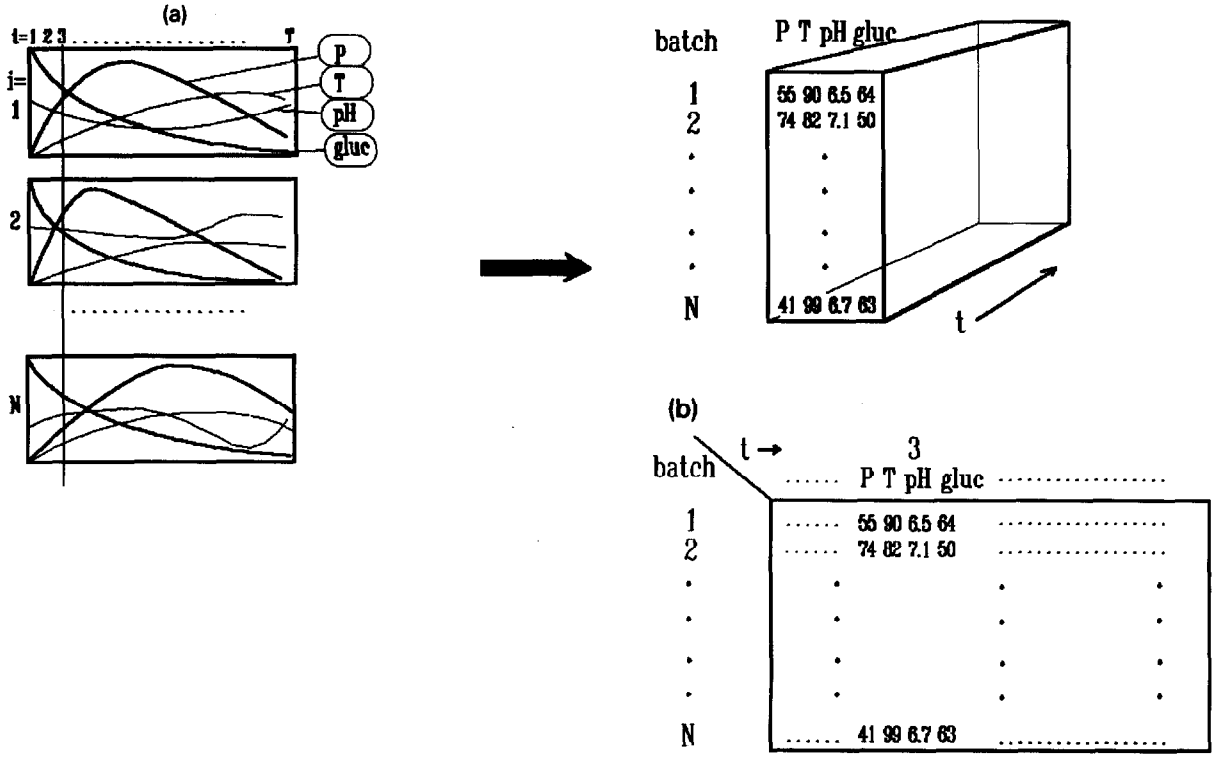

Fig. 2. (a) A set of process time sequences is quantified by process variables such as pressure, temperature, pH and glucose concentration. The resulting process data matrix (X) is three-way (process sequence × time × process variable). (b) In the case of process time sequences of the same length, the three-way table can be directly "unfolded" to give a two-way table with several columns for each process time point.

those previously analysed and to the model. This gives a "diagnosis" of each new object, whether it is similar to the set of previously modelled objects and, if so, in which region of the map the object falls.

For analytical chemistry, correlations between the position in the map and chromatographic retention, solubility, etc., are of interest.

With biopolymer sequences it may be of interest to correlate the inter-point distances in the map to evolutionary or genetic resemblance. Relationships between the place in the map and various diseases are relevant for medical diagnostic objectives.

Analogously, multivariate process diagnosis has recently started to develop, where patterns in multivariate time sequences are used to diagnose the state of the process, whether it is running normally or whether problems are occurring [4–6,10].

Often the diagnostic objective is formulated in terms of a classification, as discussed further below.

Classification into known classes (groups)

When more is known about the investigated system, this knowledge is often formulated as a class structure; peptides are classified according to their type of biological activity as enkephalins, bradykinins, signal peptides, etc. A class may be subdivided into sub-classes, e.g., signal peptides are divided into periplasmic or outer membrane and other types.

A process may be divided in classes such as "normal and under control", "raw material problems", "temperature problems in reactor 1" and "stack-up problems before reactor 2".

The scope of the analysis now is usually to see how the data collected for the objects relate to this class structure, and thereafter to classify new objects accordingly. The result of the analysis is still a "map" of the similarities and dissimilarities of the objects, a model of the system, often with sub-maps for the different classes. Again, the analysis will inform on which variables are important for the class separation and which are not. Also, the model informs about the degree of separation of the classes, which are well resolved and not, etc.

The model can then be used in a second stage of the data analysis to classify new objects. The result may be one of three: that an object is uniquely assigned to one of the classes on a certain probability level, or that the object is similar to several classes or that it is similar to none of the classes, an outlier, a new type.

The tools for classification, or pattern recognition, or discriminant analysis as this analysis is often called, are many, and are covered in reviews and textbooks [11,12]. When the number of characterizing variables (K) is larger than the number of objects (N), which is common in the analysis of sequences, projection methods such as SIMCA [13] and PLS discriminant analysis [14,15] are advantageous.

Quantitative relationships between sequence and biological activity or process variables and product properties

On the highest ambition level of data analysis, the scope is to find relationships between the characteristics of the objects (sequences) and other quantitative variables (response variables), e.g., the measured levels of biological activity variables, enzyme binding constants or other physico-chemical variables, such as chromatographic retention times and solubility. With processes, the responses are "properties" of the product and process such as yield, purity, stability, strength and viscosity, including "negative" results such as costs and pollution levels.

Often the data analysis is a combination of a classification and a quantitative modelling: the first objective is to find the class or type of an object, and thereafter to obtain a prediction of values of quantitative property variables that are defined for this class.

In the analysis, a quantitative model relating the characteristics of the objects to the response variables is derived with the help of a training set of objects with known values of the response variables. If several classes are at hand, one usually develops one such model for each class.

The developed model(s) are then used to predict the response values for new objects, after a classification if so needed.

Projection methods such as PLS (partial least-squares projection to latent structures) are suitable tools for this type of analysis [1,2,4,11,14–17]. In simple cases when few variables are needed to characterize the objects, and when statistical experimental design [18,19] has been applied for the selection of the training set, and when the number of responses is small (say a maximum of 5) and independent, multiple regression can be used for the model development and data analysis [20].

Optimization

A common objective is to achieve optimum properties of a biopolymer sequence or optimum quality and yield of a process. The models can be used for this purpose in several ways. With simple models, one can search mathematically for their maxima and minima, but with more complicated models simulations are necessary.

SELECTING A SET OF SEQUENCES (DESIGN OF A TRAINING SET)

To develop a model of a family of sequences, it is necessary to have a basis in the form of a representative sub-sample of family members. The selection of this sub-sample is far from trivial, and only recently has a rational approach become available with the methodology of statistical experimental design [2,18,19,21]. The intuitive way of generating a sub-set by modifying one position at a time in a "lead sequence" does not give a sub-set with sufficient information for good model development [2,21]. Unfortunately, this intuitive approach is still used in most applications, which is a major reason for the poor quality of structure–activity models in general.

EXAMPLES

To illustrate the principles of biopolymer sequence modelling, some sets of short peptides and a set of fairly short DNA sequences that were previously modelled by position based quantitative structure–activity relationships (QSARs) will be used. These sets of biopolymers will be modelled in two ways, with position-based description (where possible), and auto- and cross-covariance (ACC, see below)-based description which is alignment independent. In addition, some simulations have been made to demonstrate the non-equivalence between these two ways of modelling.

TABLE 1

Results of the PLS modelling of (1) the position-based (pos.) and (2) the auto-correlation (ACC)-transformed biopolymer sequence data [a]

Parameter	Ex. 1, enkephalins	Ex. 2, bradykinins	Ex. 3, signal peptides	Ex. 4, DNA promoters	Ex. 5, simulations
Length (L)	5	9	18 –28	68	5, 10
No. of sequences (N)	31	43	22	25	16, 32
No. of X-variables, K (pos.)	16	34	60	195	15, 30
R^2, mult. corr. (pos.)	0.59	0.52	0.65	0.85	0.88, 0.85
Q^2, R.S.D.–mult. corr. (pos.)	0.46	0.25	0.12	0.40	0.70, 0.30
No. of X-variables, K (ACC)	48	48	54	27	27, 27
R^2, mult. corr. (ACC)	0.58	0.48	0.91	0.79	0.30, 0.31
Q^2, R.S.D.–mult. corr. (ACC)	0.49	0.29	0.63	0.41	0.0, 0.0

[a] R^2 and Q^2 refer to the ordinary and cross-validated multiple correlation coefficients. $R^2 = 1 - [SS\ Y\text{-resid}/SS\ Y]$, where SS denotes sum of squares, Y-resid the residuals and Y the observed values of the response variable. The number of X-variables in the models is denoted by K and the number of cases by N.

(1) MVD enkephalins (length L = 5)

These $N = 31$ pentapeptides are of interest because they contain some D-amino acids in the sequences of mainly L-amino acids. They were previously modelled by position-based QSARs [22], where each of the four varying positions (position 1 is constant) were coded by the three z-scales of Hellberg [23] and Jonsson et al. [24]. An additional indicator variable was included to code for the D- or L-form of the amino acid. Hence the position-based description takes sixteen X-variables: four positions × (three z-scales and one indicator). A PLS model with two significant components explains $R^2 = 0.59$ of the Y-variation (enkephalin activity), with a cross-validated multiple correlation coefficient (Q^2) of 0.46 (see Table 1).

(2) Bradykinins (L = 9)

These $N = 43$ nonapeptides (length $L = 9$) were previously modelled [22] in the same way as the enkephalins in example 1. These bradykinins also had both D- and L-amino acids, except in positions 5 and 8, which had only L-forms. Hence, a position-based characterization with the same four variables as in the example 1 takes 34 variables $[(9 \times 3) + (7 \times 1)]$. A two-dimensional PLS model gives $R^2 = 0.52$ and $Q^2 = 0.25$.

(3) Signal peptides (18 ⩽ L ⩽ 28)

Signal peptides are N-terminal peptides sitting on proteins that, after their synthesis in the cytoplasmic domain of the cell, are translocated to a membrane or through one or several membranes. These signal peptides usually consist of 15–30 amino acid residues. Previously a set of *Escherichia coli* signal peptides were analysed and it was shown that the sequences do indeed contain patterns related to the site to which they direct the protein [25]. This analysis was made with a modified position-based description of the sequences based of the three z-scales [23,24].

The multi-positional description of peptide sequences used in QSAR studies [2,22,23] is not applicable here because the numbers of amino acids in the sequences are different. Hence the

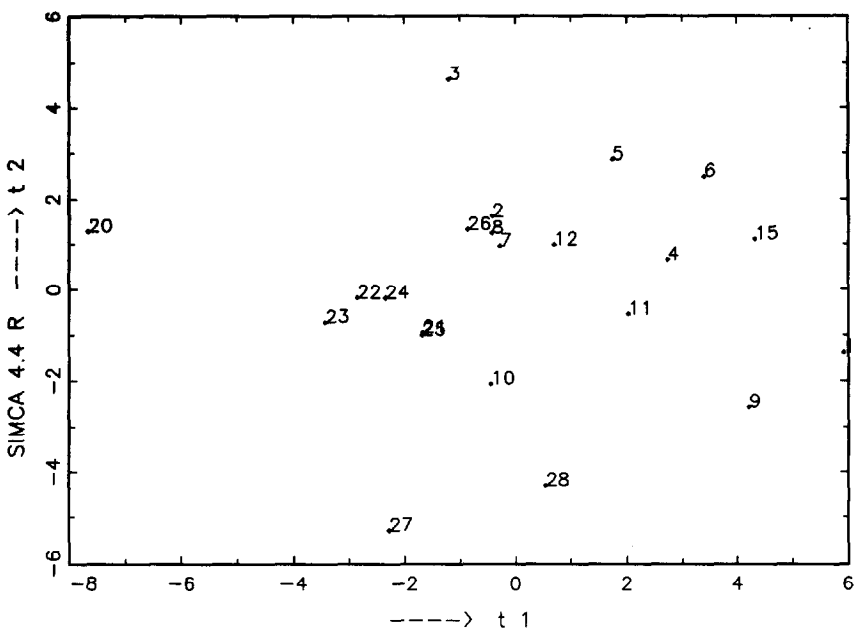

Fig. 3. Score plot of the PLS discriminant model of the signal peptides quantified by the "field" approach [25]. Nos. 1–15 are periplasmic space signal peptides and Nos. 20–28 are outer membrane signal peptides. The periplasmic peptide 10 and outer membrane peptide 26 are seen to fall in wrong regions.

sequences would be represented by different numbers of variables and it is not obvious which variables correspond to each other. In an effort to circumvent these problems, the so-called field approach was previously developed, where the two-dimensional matrix of one sequence is projected down on a vector with specified length. Hence the result is a specific number of variables for each sequence, independent of sequence length. The field approach is described in detail in a previous paper [25].

The field approach was used to generate a data matrix for 43 signal peptide sequences with a length between 18 and 32 amino acids. The final non-cytoplasmic locations of the corresponding proteins in *E. coli* were used as class labels. The number of classes was five: periplasmic space, outer membrane, pili, toxins and inner membrane proteins. The difference between pair-wise classes was investigated with PLS discriminant analysis. Significant differences according to cross-validation were obtained between most pairs of classes. However, the cross-validated explained variance of the discriminant variable was fairly low, which made the classification of sequences with unknown class membership uncertain.

Here the results from a PLS discriminant analysis based on the field approach are compared with those based on the auto- and cross-covariances of the *z*-scales coding the amino acid positions. To limit this preliminary study, only two classes of signal peptide sequences are considered, namely those of periplasmic space proteins and those of outer membrane proteins. The sequences are Nos. 1–22 in Table I in [25]. The location of the protein corresponding to protein No. 20 (MolA) has been reclassified from an outer membrane to a periplasmic protein since the previous analysis.

A PLS discriminant analysis [14,15] was done on a data matrix with 60 variables generated with the field approach. Here the first PLS dimension was barely significant (cross-validated 0.88) and the second was not. In the score plot (Fig. 3) the resolution of the classes is seen to be moderate. We note that two of the sequences fall in the wrong domains in the plot, and hence are wrongly classified.

(4) DNA promoters ($L = 68$)

Promoters are specific DNA sequences that govern the binding of the sigma unit of the RNA polymerase holoenzyme (RNAP), thereby punctuating the onset of transcription. Jonsson et al. [2,26] recently investigated the relationship between the nucleic acid sequence of a set of $L = 68$ long promoters in *E. coli* and their in vivo promoter efficiency. Each position was coded by three descriptors, giving a position-based description of 195 variables (three positions were constant in the set). A PLS model with three significant components modelled $R^2 = 0.85$ of the variation of the *Y*-variable (promoter efficiency), with a cross-validated $Q^2 = 0.40$. This model was then used to predict how to construct sequences with even higher activity than the most active in the training set. The predicted "optimum" sequences were synthesized and indeed confirmed to be stronger promoters than the previously existing ones [2,27].

(5) Simulations of penta- and decapeptides ($L = 5$ and 10)

In the present examples the peptide and DNA sequence models based on either a position-based description or an auto- and cross-covariance (ACC) description give similar results, except for signal peptides where the ACC model is superior. This might be interpreted as that these two ways of translating three-way to two-way data are mathematically equivalent. To investigate this possibility, a limited set of simulations generating position-based data for artificial sets of (a) $N = 20$ penta- and (b) $N = 32$ decapeptides ($L = 5$ and 10) were made. The peptide sets were generated according to Plackett Burman or fractional factorial designs [18,19,21], respectively. An activity variable, *Y*, was generated as a linear model in the three *z*-scales times the L positions ($3 \times L$ variables) with uniformly $(-1, 1)$ randomly distributed coefficients. Normally distributed residuals (noise) with a standard deviation (S.D.) equal to 10% of the S.D. of *Y* was added to the *Y*-variable.

The generated data were then analysed analogously to examples 1, 2 and 4 with a position-based model and an ACC-based model. The position-

based model, as expected, adequately recovered the generated model (see Table 1). The cross-validation is seen to underestimate grossly the predictive power of the model for the data generated according to the fractional factorial designs (decapeptides), however. This phenomenon is well known and understood; the roundness and orthogonality of this design in X-space make regression and PLS models change direction substantially in X-space for each deletion of an observation in the cross-validation scheme, with the consequence that the noise in the data is overestimated.

For these simulations, the ACC model failed to give significant models according to the cross-validation. This result shows that the two ways of translating three-way into two-way data are not equivalent, and that the results of examples 1–4 indicate that periodicities and neighbour effects are important in these sets of biopolymer sequences.

TRANSLATING SEQUENCES AND PROCESSES TO DATA MATRICES

Once a "training set" of sequences is available, and their biological or/and physico-chemical properties have been measured, one needs to translate the actual sequences of the training set into numbers to be able to develop a relationship between sequence and activity or other properties.

Description of each position giving a three-way matrix

The translation of a sequence to quantitative variables is a difficult problem because it touches the very essence of chemistry and molecular biology, namely what structural features are important in a sequence and how to quantify them. There is the theoretical approach where a large number of quantum mechanical "indices" such as charges, partial molecular volumes and energy levels are calculated and used as descriptor variables.

Another approach is based on the so-called analogy principle, where one tries to use measurements on model systems to derive scales that are then used to describe to actual sequences. Along this line, three scales have been developed for the individual amino acid positions, the so-called z-scales [23,24]. These quantitative scales are developed from multivariate measurements in model systems and roughly correspond to hydrophilicity (z_1), steric bulk (z_2), and polarity/charge (z_3). Using these scales, one can hence translate a peptide sequence of length L into $3 \times L$ numbers. These can be arranged in a three-way matrix as in Fig. 1.

With a process time sequence, one usually measures values of the process variables such as pH, temperature or pressure, at various locations in the process set-up (feeders, reactors, distillation towers, etc.). These multivariate process measurements makes the data for each batch process be a matrix, and the data of a set of batch processes be a three-way matrix, here denoted **X**.

Output variables (responses, Y) measuring quality, yield, and cost are often measured less frequently, and in batch processes only at the end of the time sequence. Hence, the response matrix, **Y**, is often smaller than the process data matrix, **X**, and is often just a two-way matrix (batch × response).

When the sequences (polymer or time) have the same length and can be assumed to be alignable, these three-way matrices have data elements everywhere. For a set of sequences of different lengths, however, the three-way matrix will have parts without data, which is discussed further below.

Transforming the three-way matrix into a two-way data matrix

Depending on the degree of diversity between sequences or processes and assumptions about the "auto-correlation structure" in the sequences, the translation into the two-way matrix should be done in different ways.

In the modelling of time sequences of process data, it has long been recognized that the values of process variables at a time point t are not independent of the corresponding values at the time points $t-1$ and $t+1$. This has lead to the

use of time series models, where the correlations over time are taken into account [7]. One may argue that similar types of dependences might occur in biological sequences, and that time series tools here may have an interesting new area of application.

The simplest, and not always sufficient, way to account for dependences between consecutive observations is to use Fourier transforms, auto-correlation or auto-covariance transforms, or other suitable transforms of the data. Indeed, Van Heel [28] has recently shown that a qualitative 1–2 auto-correlation transform of protein sequences contains sufficient information to classify the proteins into known classes of great detail [28]. Here quantitative 1–2, 1–3 and 1–4 auto-correlation transforms of sequence descriptors will be used and compared, where possible, with the traditional position-based description.

Position / time aligned unfolding for similar sequences of almost the same length without neighbouring effects. When the biopolymer or process time sequences in a three-way matrix all are of the same length (or nearly so, see below), and a sequence position or process time point indeed corresponds to the same position or time point in the other sequences, one says that the sequences are alignable. One can then unfold the three-way table to give a two-way table as shown in Figs. 1b and 2b. This two-way matrix, **X**, can then be modelled and analysed according to one of the objectives stated above; overview, classification or quantitative relationship to Y.

When the sequences have almost the same length in chain or time, one can reach an approximate alignment by means of insertions of gaps at suitable positions in the shorter sequences. These gaps can then be parameterized by special values of the characterizing variables, or be left as missing data. The resulting three-way table, **X**, is then analysed in the same way.

This unfolding of **X**, followed by its analysis by a linear PC, classification, regression or PLS model, corresponds to the assumption that the positions of the sequences are truly independent, i.e., there are no neighbour–neighbour interactions. In process data this is a fairly unrealistic assumption, and one can use other ways of translating the three-way **X** into a two-way matrix as discussed below.

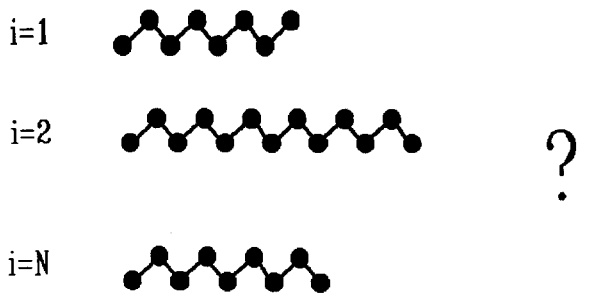

Fig. 4. Biopolymer sequences of different lengths cannot be directly quantified as in Fig. 1, because the sequences would give a two-way table with different numbers of columns, and the three-way table cannot be unfolded.

Transformations independent of alignment based on auto- and cross-covariance (ACC) structures. Whenever sequences in the same analysed set differ much in length (Figs. 4 and 5), one must use principles other than position-based to go from the three-way **X** to the two-way **X** table. However, also in the modelling of sequences of the same length one may wish to use a translation that takes neighbour effects, i.e., lack of independence between subsequent positions (time points), into account. This can be done by using ACC, and other transforms of each sequence, followed

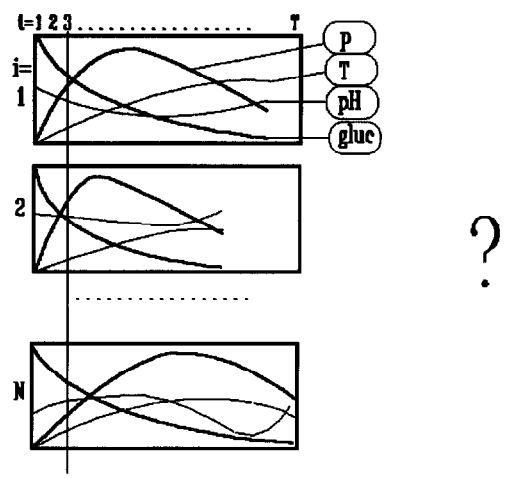

Fig. 5. Process time sequences of different lengths cannot be directly quantified as in Fig. 1, because the time sequences would give two-way table with different numbers of columns, and the three-way table cannot be unfolded.

TABLE 2

Example of the generation of auto- and cross-covariance functions (ACC) of a pentapeptide (RYLPT) described in each amino acid position by the three z-scales [23,24] [a]

Pos. (i)	AA (Amino acid)	z_1	z_2	z_3	$A_{11}(1)$ $z_1(i)$ $*z_1(i+1)$	$C_{12}(1)$ $z_1(i)$ $*z_2(i+1)$	$C_{13}(1)$ $z_1(i)$ $*z_3(i+1)$	$C_{21}(1)$ $z_2(i)$ $*z_1(i+1)$	$A_{22}(1)$ $z_2(i)$ $*z_2(i+1)$	$C_{23}(1)$ $z_2(i)$ $*z_3(i+1)$	$C_{31}(1)$ $z_3(i)$ $*z_1(i+1)$	$C_{32}(1)$ $z_3(i)$ $*z_2(i+1)$	$A_{33}(1)$ $z_3(i)$ $*z_3(i+1)$
1	Arg, R	1.13	−2.36	1.26	−4.0454	2.3391	−0.0452	8.4488	−4.8852	0.0944	−4.5108	2.6082	−0.0504
2	Tyr, Y	−3.58	2.07	−0.04	−13.4608	−4.1528	8.3056	7.7832	2.4012	−4.8024	−0.1504	−0.0464	0.0928
3	Lys, L	3.76	1.16	−2.32	0.0752	1.7672	11.5056	0.0232	0.5452	3.5496	−0.0464	−1.0904	−7.0992
4	Pro, P	0.02	0.47	3.06	0.011	−0.0446	−0.0298	0.2585	−1.0481	−0.7003	1.683	−6.8238	−4.5594
5	Thr, T	0.55	−2.23	−1.49									
Sum					−17.42	−0.0911	19.7362	16.5137	−2.9869	−1.8587	−3.0246	−5.3524	−11.6162
Sum/4		ACC(1)			−4.355	−0.023	4.934	4.128	−0.747	−0.465	−0.756	−1.338	−2.904
Pos. (i)	AA (Amino acid)	z_1	z_2	z_3	$A_{11}(2)$ $z_1(i)$ $*z_1(i+2)$	$C_{12}(2)$ $z_1(i)$ $*z_2(i+2)$	$C_{13}(2)$ $z_1(i)$ $*z_3(i+2)$	$C_{21}(2)$ $z_2(i)$ $*z_1(i+2)$	$A_{22}(2)$ $z_2(i)$ $*z_2(i+2)$	$C_{23}(2)$ $z_2(i)$ $*z_3(i+2)$	$C_{31}(2)$ $z_3(i)$ $*z_1(i+2)$	$C_{32}(2)$ $z_3(i)$ $*z_2(i+2)$	$A_{33}(2)$ $z_3(i)$ $*z_3(i+2)$
1	Arg, R	1.13	−2.36	1.26	4.2488	1.3108	−2.6216	−8.8736	−2.7376	5.4752	4.7376	1.4616	−2.9232
2	Tyr, Y	−3.58	2.07	−0.04	−0.0716	−1.6826	−10.9548	0.0414	0.9729	6.3342	−0.0008	−0.0188	−0.1224
3	Lys, L	3.76	1.16	−2.32	2.068	−8.3848	−5.6024	0.638	−2.5868	−1.7284	−1.276	5.1736	3.4568
4	Pro, P	0.02	0.47	3.06									
5	Thr, T	0.55	−2.23	−1.49									
Sum					6.2452	−8.7566	−19.1788	−8.1942	−4.3515	10.081	3.4608	6.6164	0.4112
Sum/3		ACC(2)			2.082	−2.919	−6.393	−2.731	−1.451	3.360	1.154	2.205	0.137

[a] The calculations of the auto and cross covariance functions of lag d, $A(d)$ and $C(d)$ respectively, are described in the text. For lag $d = 1$, the contributions to an ACC, for example $C_{12}(1)$, are formed by multiplying the z_1-value of position i by the z_2-value of position $i+1$ for $i = 1$, 2, 3, and 4. Thereafter these four terms are summed and divided by 4 to give the average = $C_{12}(1)$.

by the digitization and unfolding of suitable parts of these transforms.

In pioneering work, Van Heel [28] transformed a large number of protein sequences into nearest neighbour amino acid frequency histograms. In each sequence, he counted how many times the pair Ala–Ala appears, the pair Ala–Asn, the pair Ala–Asp, etc., until the pair Tyr–Tyr. This gave 400 relative frequencies, one for each of the 20×20 amino acid pairs. A characterization based on these 400 values can be used as variables to characterize any peptide sequence, independent of its length or alignment. Van Heel then showed that these variables contain sufficient information about similarities and dissimilarities of the proteins to classify them into known classes with very fine detail.

The advantages of Van Heel's sequence characterization are that it is general and independent of alignment, and that it accounts for nearest neighbour interactions. It has some drawbacks, however, in that it is qualitative (all pairs of amino acids are equivalent) and difficult to interpret.

Here a simplification and quantification of Van Heel's scheme is proposed by using first the three z-scales (and a fourth for D-/L- if warranted), and then calculating the ACC functions of this description along the sequence. With three z-scales this gives nine nearest neighbour (lag 1) ACCs, nine next nearest neighbour (lag 2), etc., because one computes also the ACCs between z_1 and z_2, between z_1 and z_3, etc. To simplify further the description and calculations, the auto- and cross-covariance functions are used instead of auto- and cross-correlations (see below). The abbreviation ACC will be used, somewhat loosely, for both auto- and cross-correlations and covariances, which is acceptable as they are almost identical after scaling.

Table 2 gives an example of the calculation of the lag 1 and lag 2 ACCs of a pentapeptide as an illustration.

As pointed out by Van Heel [28], it may be warranted to compute the ACCs separately for, say, the beginning of a sequence, the middle of a sequence and the end of a sequence, particularly when the sequences are long. However, in all examples including the signal peptides a single ACC function is used over the whole sequence. Van Heel's proposal will be investigated in the future in connection with a more systematic study of sequence modelling. To complement the ACC description, averages of the z-values for different parts of the sequences, say the beginning, middle and end, may provide additional variables of possible modelling value.

Auto- and cross-correlation and covariance (ACC)-based transformations. The lag d auto-correlation function of a descriptor x (say z_1) over a sequence of length L is (the subscript i is the sequence position index running from 1 to L) [7] is given by

$$A(d) = \sum_i (x_i - \bar{x})(x_{i+d} - \bar{x}) / \sum_i (x_i - \bar{x})^2$$

The corresponding auto-covariance function is [7]

$$A(d) = \sum_i (x_i - \bar{x})(x_{i+d} - \bar{x}) / L$$

A lagged cross-covariance function between two descriptors x and u, e.g., $x = z_1$ and $u = z_2$, is analogously defined as [7]

$$C_{xu}(d) = \sum_i (x_i - \bar{x})(u_{i+d} - \bar{u}) / L$$

The summations are done from 1 to $L - d$ because of terms such as x_{i+d}. Also, it is important to note that ACC_{xu} functions differ from ACC_{ux}. Table 2 gives an example of the calculation of the lag 1 and lag 2 ACC functions of a pentapeptide.

In the calculation of ACC functions of the biopolymer sequences centred terms were not used as above, i.e., subtracting $\bar{x}$, etc., but rather used uncentered terms corresponding to using $\bar{x} = 0$, because the z-scales are already centred over all amino acids in their derivation.

MULTIVARIATE MODELLING AND ANALYSIS

Once a set of sequences have been translated into a two-way data matrix, standard methods of multivariate analysis such as PCA, and PLS are used to develop the models relating sequence to properties (or class) [1,2,8,9,14–17]. With both polymer and process time sequences, evolving

factor analysis (EFA) [29] and evolving PLS [30] may be interesting alternatives aimed at an early diagnosis of sequence "type" or class, and possibly capable of recognizing "clean" parts of a sequence corresponding to a single action mechanism.

In this study PLS modelling and PLS discriminant analysis [8,14–17] were used in all examples. The two-way data matrices were all auto-scaled to unit column variance before the analysis, and cross-validation was used to ensure predictive significance of the models. SIMCA-4R software [31] was used for the computations.

RESULTS

Table 1 summarizes the results of the simulations and the four examples. Only example 3 will be discussed in more detail because it is the most general with sequences of different lengths, and because the ACC model gives such good results.

Simulations

The main result is that the position-based models and ACC-based models are not equivalent. Data generated according to a position-based model are, in general, very poorly modelled by an ACC model unless periodic structures are explicitly generated.

Examples 1, 2 and 4

The interesting result of these examples is that the ACC models give the same fit (R^2) and slightly better cross-validated predictive power than the position-based models, even for the very short peptides (see Table 1). This indicates that the positions in these sequences are not independent, which may have consequences for the design and modifications of biopolymer sequences as discussed further below.

Example 3, signal peptides

Each signal sequence was first multi-positionally described by the three z-scales. Then the ACC model was used to generate 54 variables. Thus all possible cross-terms were generated between amino acid position i and position $i+d$, and where d is between 1 and 6. ($3 \times 6 \times 3 =$ scales $\times$ lagged positions $\times$ scales $= 54$). These variables are labeled a to f for lags 1 to 6, followed by the two subscripts of the cross-covari-

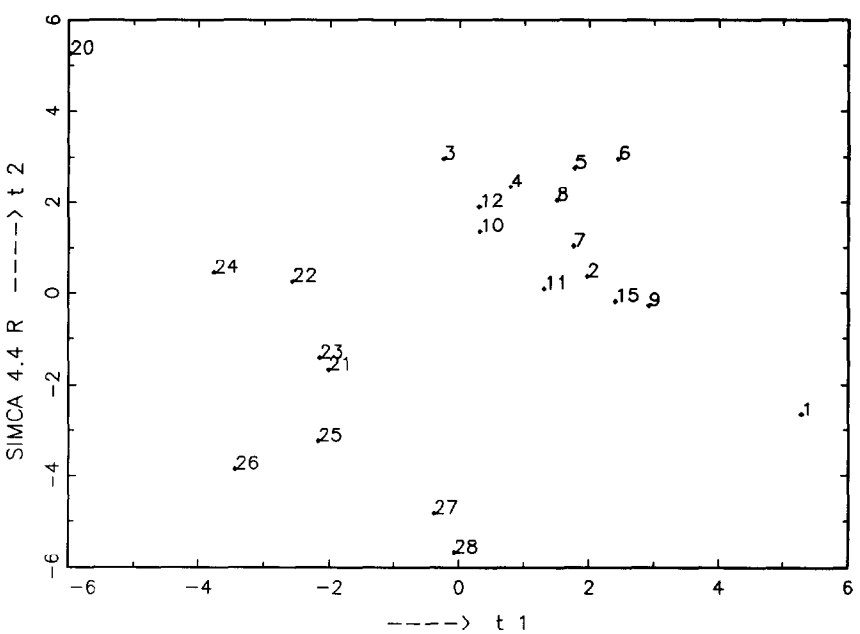

Fig. 6. Score plot of the PLS discriminant model of the signal peptides quantified by the ACC functions of the z-scaled. Nos. 1–15 are periplasmic space signal peptides and Nos. 20–28 are outer membrane signal peptides.

ated z-scales. Hence d_{31}, for example, denotes the lag 4 cross-covariance between z_3 and z_1.

A dummy variable with the value $y = 1$ was used for the periplasmic space proteins and $y = -1$ for the outer membrane proteins. The PLS discriminant analysis, based on auto-scaled data, showed that two components were significant according to cross-validation, with $Q^2 = 0.63$. The two PLS dimensions model 64% and 26%, respectively of the variation in y.

A score plot of the two significant components (t_2/t_1) is given in Fig. 6 and the corresponding loading plot in Fig. 7. The score plot reveals a very good separation between the classes, with all training sequences correctly classified. The PLS weight plot (Fig. 7) shows that many of the ACC functions contribute with information to the class separation, dominated by the terms of gap 4, z_3, z_1, and z_1, z_1 (d_{31} and d_{11}) and the gaps 3, and 6, z_2, z_1 terms (c_{21} and f_{21}).

The negative signs of the d_{31} terms show that in signal peptides of outer membrane proteins d_{31} is high (because $y = -1$ for these). This means that in these sequences there is a strong tendency to have lipophilic amino acids such as Val, Leu and Phe (negative z_1 values), situated four positions after amino acids with positive z_3 values (Pro, Cys, Asp), or hydrophilic amino acids with positive z_1 values (e.g., Arg, Asn, Asp, Glu) four positions after amino acids with negative z_3 values (e.g., Lys, Arg).

Signal peptides of periplasmic proteins, in contrast, are characterized by positive d_{11}, c_{21}, and f_{21} values, i.e., amino acids separated by a gap of 4 have a tendency of being either both hydrophilic or both lipophilic (d_{11}), and there are positive correlations between large size and hydrophilicity, or small size and lipophilicity, respectively, 3 and 6 positions later (c_{21} and f_{21}).

These coefficients can possibly be interpreted in terms of helix-making or helix-breaking patterns, and these possibilities will be pursued later in a more detailed investigation of a larger set of signal peptides. It is worth noting that the strongest patterns in both classes involve lagged cross-covariances between different properties which easily escape traditional sequence analyses of one separate property at a time.

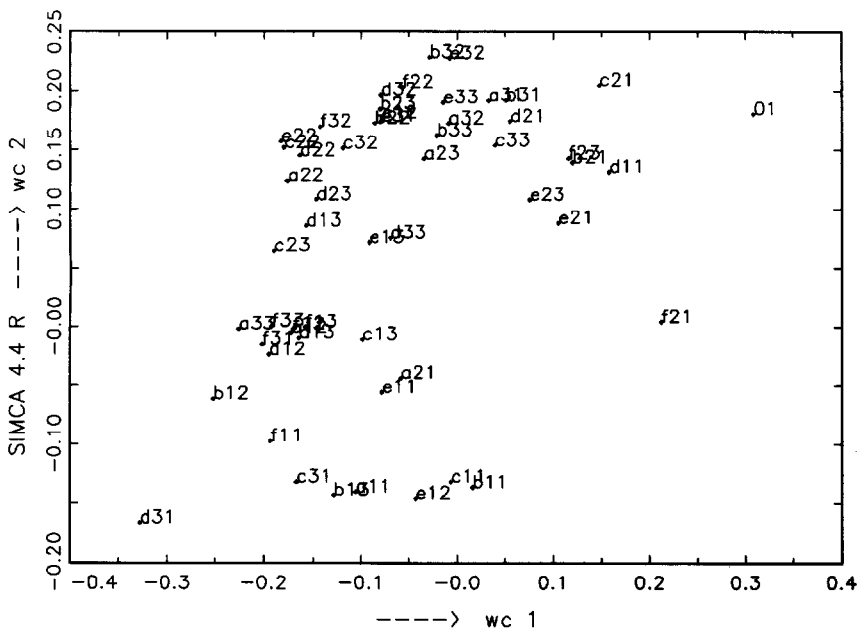

Fig. 7. Plot of the PLS coefficients w and c of the two dimensions of the ACC PLS discriminant model of the signal peptides. The point 01 indicates the position of the dummy y-variable. The model is seen to be dominated by the terms d_{31} (lag 4, z_3, z_1 cross correlation), c_{21}, d_{11}, and f_{21} terms.

ADVANTAGES AND DISADVANTAGES WITH ACC-BASED MODELS

ACC modelling has some clear advantages in that the description and modelling of a set of polymer or time sequences become independent of alignment. Whenever a set of sequences have widely different lengths, as in the signal peptide example, position-based models requiring alignment are difficult or impossible. Moreover, consistent dependences between neighbouring sequence positions can be modelled.

However, a clear difficulty with ACC models, which they share with any multivariate models where interactions between variables are important, is the interpretation and understanding of the model. This is very much an educational problem; we are taught to see the world in terms of independent variables and individual causal factors, even when it is obvious that many natural systems display strong interactions, multiple causes and lack of independence. Auto- and cross-covariance patterns correspond to repeated interactions of the same kind over the sequence, and are hence difficult to comprehend.

This difficulty of interpretation corresponds to difficulties in predicting sequences with higher or lower values of the response variables and constructing a typical sequence belonging to a certain class, say a periplasmic signal peptide. Only systematic simulations, starting with a given sequence and making multiple simulated changes, and inserting each modified sequence into the model, will result in sequences predicted to be "optimal" for a given purpose. These multiple simulated changes of the starting sequence must, of course, be made according to an appropriate statistical design [18,19,21], in order for the results to be reliable.

DISCUSSION AND CONCLUSION

There are many ways to quantify biopolymer and process data time sequences. The use of theoretical calculations by molecular mechanics and quantum chemistry (molecular modelling) is always possible, but has the drawback of being complicated and time consuming, especially for longer sequences. The molecular modelling of sets of sequences of different length introduces problems of equivalence; it is not certain that theoretically derived parameters are directly comparable for sequences of different length.

Process modelling has similar problems: knowing what to measure, where and how often requires great insight and experience.

The examples in this paper demonstrate that the simple and straightforward quantification of biopolymer sequences by z-scales that are derived from physico-chemical properties of free amino acids (or nucleotides) contains much information about the properties of the sequences.

Position-based modelling of the resulting data works as long as sequences are of the same length and similar so that indeed a position in different sequences has the same chemical and biological significance, and as long as sequence periodicities lack biological significance; and analogously for process time sequences. Interactions between different parts of the sequence could be modelled by interaction terms between variables at different positions. With a sequence of some length, the number of possible interactions becomes very large, however; in a decapeptide there are $10 \times 9/2$ possible position–position interactions, which, when multiplied by 9 for all possible z–z cross-interactions, becomes 405.

For sets of biopolymer sequences of different length, the further preprocessing of the sequence characterizing data by means of ACC (auto- and cross-correlation or -covariance) transformations gives alignment-independent and general descriptions of the sequences, which also seem to preserve substantial parts of the information in the data. Averages of the z-values for different parts of the sequences may provide complementary information regarding the "average" structure of sequences, particularly in classification problems.

Similarly, multivariate ACC modelling (including cross-correlations) of sets of process data time sequences is an interesting and simple alternative to process time series modelling, e.g., of batch processes. Batch processes often differ in length because the time of "completion" is influenced by changes in process variables such as pH, tem-

perature and pressure during the course of the process. Fermentation processes in biotechnology provide typical examples.

The fact that ACC modelling of biopolymers works so well indicates that nature indeed assigns significance to "periodic" patterns. The biochemical interpretation of this finding is far from clear and much more experimental validation and experience of the ACC models is necessary. Some periodicities related to helices of peptide sequences are known and possible to recognize, but the extent of other periodic patterns in biopolymer sequences is little understood. The possibility that the recognition of sequences by biological "receptors" may be based on auto-correlation patterns is most interesting, however, and may have profound consequences for biochemical theory. The results of van Heel [28] and, to some extent the present investigation raise this possibility.

This line of investigation is continuing, looking at a wider range of data and at such questions as the use of a single or several ACC structures, possible relationships between ACC structures and protein folding and other secondary and tertiary structures, i.e., helices, sheets, bends, and motifs involving their combinations. The use of ACC structures with QSARs of ordinary organic molecules that are not sequences is also an intriguing possibility. In addition, the use of these models in analytical chemistry to predict chromatographic separation and other properties of peptides and nucleotides is of great practical interest.

Support from the Swedish Natural Science Research Council (NFR), the Swedish Board for Technical Development (STU–NUTEC) and the Centre for Environmental Research (CMF) is gratefully acknowledged.

REFERENCES

1 W.J. Dunn and S. Wold, in C. Hansch and C. Ramsden (Eds.), Comprehensive Medicinal Chemistry, Vol. 4, Pergamon Press, Oxford, 1990, Chap. 22.3.
2 J. Jonsson, Thesis, Research Group for Chemometrics, Umeå University, Umeå, 1992.
3 A. Lorber and B.R. Kowalski, J. Chemometr., 2 (1988) 67.
4 B. Skagerberg, J.F. MacGregor and C. Kiparissides, Chemometr. Intell. Lab. Syst., 14 (1992) 341.
5 J.V. Kresta, J.F. MacGregor and T.E. Marlin, Can. J. Chem. Eng., 69 (1991) 35.
6 P. Nomikos and J.F. MacGregor, in NATO ASI Symposium on Batch Processes, Antalia, Turkey, May 1992, Springer, New York, 1993.
7 G.E.P. Box and G.M. Jenkins, Time Series Analysis, Holden-Day, Oakland, CA, 2nd edn., 1976.
8 J.E. Jackson, A User's Guide to Principal Components, Wiley, New York, 1991.
9 S. Wold, K. Esbensen and P. Geladi, Chemometr. Intell. Lab. Syst., 2 (1987) 37.
10 Integrated Process Intelligence (IPI), Manual of SIMCA-P, Version 1.0, Umetri, Umeå, and ABB–AUT, Luleå, 1992.
11 C. Albano, W.J. Dunn, U. Edlund, E. Johansson, B. Norden, M. Sjöström and S. Wold, Anal. Chim. Acta, 103 (1978) 429.
12 K.V. Mardia, J.T. Kent and J.M. Bibby, Multivariate Analysis, Academic Press, London, 1979.
13 S. Wold, Pattern Recognition, 8 (1976) 127.
14 M. Sjöström, S. Wold and B. Söderström, in E.S. Gelsema and L.N. Kanal (Eds.), Pattern Recognition in Practice, II, Elsevier, Amsterdam, 1986, p. 486.
15 L. Ståhle and S. Wold, J. Chemometr., 1 (1987) 185.
16 H. Wold, in K.-G. Jöreskog and H. Wold (Eds.), Systems Under Indirect Observation, Vol. II, North-Holland, Amsterdam, 1982, Chap. 1.
17 A. Höskuldsson, J. Chemometr., 2 (1988) 211.
18 G.E.P. Box, W.G. Hunter, J.S. Hunter, Statistics for Experimenters, Wiley, New York, 1978.
19 E. Morgan, Chemometrics: Experimental Design, ACOL, London, and Wiley, New York, 1991.
20 N.R. Draper and H. Smith, Applied Regression Analysis, Wiley, New York, 2nd edn., 1978.
21 S. Hellberg, M. Sjöström, B. Skagerberg, C. Wikström and S. Wold, Acta Pharm. Jugosl., 37 (1987) 53.
22 L. Eriksson, J. Jonsson, S. Hellberg, F. Lindgren, B. Skagerberg, M. Sjöström and S. Wold, Acta Chem. Scand., 44 (1990) 50.
23 S. Hellberg, Thesis, Research Group for Chemometrics, Umeå University, Umeå, 1986.
24 J. Jonsson, L. Eriksson, S. Hellberg, M. Sjöström and S. Wold, Quant. Struct. Activ. Relat., 8 (1989) 204.
25 M. Sjöström, S. Wold, Å. Wieslander and L. Rilfors, EMBO J., 6 (1987) 823.
26 J. Jonsson, L. Eriksson, S. Hellberg, F. Lindgren, M. Sjöström and S. Wold, Acta Chem. Scand., 45 (1991) 186.
27 J. Jonsson, T. Norberg, L. Carlsson, C. Gustafsson and S. Wold, Nucl. Acid Res., (1993) in press.
28 M. van Heel, J. Mol. Biol., 220 (1991) 877.
29 H. Gampp, M. Maeder, C.J. Meyer and A.Z. Zuberbühler, Talanta, 32 (1985) 1133.
30 K. Helland, H.E. Berntsen, O.S. Borgen and H. Martens, Chemometr. Intell. Lab. Syst., 14 (1992) 129.
31 Simca-R 4.4, Manual, Umetri, Umeå, 1992.