Acta Chemica Scandinavica, 1998: 52: 517–523
Printed in UK – all rights reserved

ACTA
CHEMICA SCANDINAVICA
ISSN 0904–213X

Review Article

Chemometrics and its Roots in Physical Organic Chemistry

Svante Wold* and Michael Sjöström

Research Group for Chemometrics, Department of Organic Chemistry, Umeå University, S-901 87 Umeå, Sweden

Dedicated to Professor Lennart Eberson on the occasion of his 65th birthday

Wold, S. and Sjöström, M., 1998. Chemometrics and its Roots in Physical Organic Chemistry. – Acta Chem. Scand. 52: 517–523. © Acta Chemica Scandinavica 1998.

Linear free energy relationships (LFERs) and extra-thermodynamic relationships (ETRs), i.e., similarity and analogy models of physical organic chemistry, are mathematically and statistically equivalent to the models much used in chemometrics and data analysis, i.e., PCA, PLS, and SIMCA. Examples of early LFERs and ETRs include the Brönsted, Hammett, Taft, and Hansch relationships.

Much of the early development of chemometrics derives from this equivalence. Thus, the interpretation and derivation of LFERs and ETRs as the first terms of serial expansions of perturbation theory applied to moderate structural change lead first to the SIMCA method for classification and discriminant analysis (pattern recognition), then to the approach of principal properties for the characterization of structural fragments, compounds, and materials, and finally also strongly influenced the development of PLS and its use in structure–effect relationships such as quantitative structure–activity relationships (QSARs).

The interpretation of chemical data by a combination of physical organic chemistry models and chemometric principles often leads to interesting conclusions as illustrated by some examples.

Chemometrics in its present form was started in the 1960s to cope with the ever increasing size of chemical data sets.[1–3] In analytical chemistry, spectroscopy and gas chromatography started to provide many variables per analytical sample, often several hundreds. Similarly, other branches of chemistry were becoming increasingly flooded by large data sets from spectroscopy, kinetics, electrophoresis, process sensors, etc.

Strangely, no statistical methods were available that could cope with data with very many variables, particularly if the number of observations (objects, samples, cases) was relatively small which was often the case in chemical data sets.

Analytical chemometrics. To begin with, the analytical chemometricians borrowed data-analytical methods from electrical engineering and computer science [4,5] to solve problems related to 'pattern recognition', i.e., classification and discriminant analysis. This tradition has continued with the use of expert systems, neural networks, and genetic algorithms – methods that have never gained much popularity outside analytical chemical applications of chemometrics.

Organic chemometrics. In organic chemometrics, however, methods were transplanted from psychology, i.e., factor and principal components analysis (PCA) and similar approaches, for the analysis of both reactivity and other data.[6,7]

The latter methods model a data matrix $\boldsymbol{Y}$ with the elements y_{ik} ($i=1, 2, \ldots, N$, and $k=1, 2, \ldots, K$) as a constant plus an expansion of A product terms [eqn. (1)]

$$y_{ik} = c_k + r_{1k}t_{i1} + r_{2k}t_{i2} + r_{3k}t_{i3} + \cdots + r_{Ak}t_{iA} + e_{ik} \qquad (1)$$

Here the constant c_k and the parameters r_{ak} are specific for the kth column of $\boldsymbol{Y}$, containing data from, for instance, reaction series **k**, while the parameters t_{ia} are specific for the ith row of $\boldsymbol{Y}$, corresponding, for instance, to substituent **i**. In organic chemistry applications, the

* To whom correspondence should be addressed.

number of 'components' (product terms), A, is usually 1, 2 or 3. The model is not exact, which is due to experimental variability, measurement errors, and model inadequacies. The residuals, e_{ik}, express this non-exactness and are the deviation between the data and the model.

Malinowski *et al.*[6] and ourselves[7–13] pointed out the equivalence of model (1) and relationships used in physical organic chemistry for relating reactivity data to each other. The Brönsted, Hammett, Taft, Yukawa–Tsuno and Marcus relationships are examples.[14,15] In the Brönsted relationship the single t-parameter scale relates to the pK_a of a series of bases or acids, and the data (y) are typically the logarithmic rate constants of reactions catalyzed by the bases or acids. The Hammett relationship defines a substituent parameter scale ($t_i = \sigma_i$) based on the pK_a of substituted benzoic acids, which then models y_i = logarithmic rate or equilibrium constants of reaction series differing in the corresponding substituents (**i**).

From a chemical point of view, model (1) can be interpreted as the quantification of the 'analogy principle' of organic chemistry. As well put by Hammett:[16] 'From its beginning the science of organic chemistry has depended on the empirical and qualitative rule that like substances react similarly, and that similar changes in structure produce similar changes in reactivity ... Linear free energy relationships constitute the quantitative specialization of this fundamental principle.'

Latent variables and physical organic chemistry 'effects'. The physical organic chemistry foundation of models (1) and (2) were initially formulated as the existence of one 'effect' for each term in the model. Thus, the Hammett equation indicated the presence of a single mechanism of interaction between substituent and reaction center – the inductive effect – and the multiple term extensions of Taft, Yukawa–Tsuno, etc. indicate the presence of two or more 'effects'. With the 'similarity model' interpretation,[7–13] however, the resulting 'components' need not be seen as clean 'chemical effects', but rather local directions in a multidimensional space which best summarize the given data, so-called latent variables. This interpretation has caused some controversy, but at the same time made the interpretation of linear free energy relationships (LFERs) and extra-thermodynamic relationships (ETRs) easier in that their foundation is less related to first principles models than initially believed.

Principal components and factor analysis in chemistry

The equivalence between model (1) – the factor analysis and principal components model – and the 'linear free energy relationships' of physical organic chemistry then led us to two things: (a) we derived 'optimal' substituent parameters for the Hammett relationship using principal components analysis of a fairly large data base of organic reactivity data,[10] and (b) we generalized somewhat the derivation and interpretation of model (1) as a second-order 'Taylor expansion of data matrices'.[11,13] Polanyi, Hammett, Leffler and Grunwald, and Palm[17] had already started this derivation, where basically model (1) is shown always to be valid as long as there is only limited variation – i.e., similarity – between the 'objects' ($i = 1, 2, \ldots, N$).

Principal properties. Eqn. (1) can be used to model *any* data measured on *any* series of similar objects – e.g., substituents, solvents, amino acids, detergents, chromatographic columns, catalysts or materials – as long as the measurements are related to this similarity. The score values (t_{ia}) resulting from the analysis of a table of 'properties' measured on a set of similar objects can be used as quantitative scales modelling the inclusion of these objects in other systems, just as the Hammett substituent scale, σ_i, can be used to model the effect of an aromatic substituent (**i**) in any reaction. Indeed the derivation of new types of 'substituent scales' has been continued with the work of Hellberg,[18] Skagerberg,[19] Jonsson[20] and Sandberg[21,22] (amino acid and nucleoside scales), and Carlson *et al.*[23] (solvents, carbonyl compounds, amines, Lewis acid catalysts). The scales are now often called *principal properties.*[24]

The SIMCA method for classification and discriminant analysis. The derivation of model (1) by means of second-order perturbation theory led us to postulate[25,26] that a method of 'pattern recognition' could be based on the separate modeling of each class of similar objects by means of a principal components model, i.e., eqn. (1). This led to the development of the SIMCA method for 'pattern recognition' that has turned out to be one of the more useful methods of chemometrics.[25,26]

The principle of SIMCA is very simple. A 'training set' of multivariate data measured on a class of similar objects can be used to develop a model of type (1). If there are several classes, this results in several models. New incoming objects for which the same multivariable data have been measured are then classified according to how similar their data vectors are to the various class models. This similarity is measured by (a) the score values (t_{ia}) resulting from fitting a data vector to a class model, and (b) the residuals (e_{ik}) after the fitting. To be judged similar to a class, the score values of the new object should all be within the typical ranges of the class model (expressed as a Hotelling's T^2), and the residual standard deviation (often called the distance to the model) should be within a tolerance interval defined by the F-distribution and the residual standard deviation of the class model.

The SIMCA method has been used to classify chemical compounds as toxic or not,[27] beta-adrenergic or not,[28] etc., and also to classify chemical reactions according to their mechanism.[29,30] In an early application it was used in the non-classical carbonium ion controversy with

conclusive results supporting the Winstein interpretation.[30]

PLS (partial least squares) projections to latent structures

With principal components models one could often see a relationship between the position in the 'class' (described by the score values, t_{ia}) and the values of other measurements made on the 'objects' (compounds, samples, etc.) such as biological activity, chronological age, etc. With the PLS models developed by Herman Wold and co-workers between 1975 and 1982,[31–33] these relationships were formally modeled by a generalization of model (1), eqns. (2a) and (2b).

$$x_{ik} = c_k + r_{1k}t_{i1} + r_{2k}t_{i2} + r_{3k}t_{i3} + \cdots + r_{Ak}t_{iA} + e_{ik} \quad (2a)$$

$$y_{im} = c_m + s_{1m}t_{i1} + s_{2m}t_{i2} + s_{3m}t_{i3} + \cdots + s_{Am}t_{iA} + f_{im} \quad (2b)$$

Because the 'scores' t are the same in the models of X (2a) and Y (2b), PLS provides a relationship between X and Y built up as a sequence of linear components ($r_{1k}t_{i1}$ and $s_{Am}t_{iA}$ above), eqn. (2c).

$$y_{im} = c_m + b_{1m}x_{i1} + b_{2m}x_{i2} + b_{3m}x_{i3} + \cdots + b_{Km}x_{iK} + f_{im} = c_m + \Sigma_k b_{km}x_{ik} + f_{im} \quad (2c)$$

PLS is a quantitative similarity/analogy model and has been used for a large array of quantitative relationships in chemistry, as well as biology, psychology, technology, and other areas. In chemistry, the most well known are multivariate calibration (MC) in analytical chemistry, and multivariate QSAR in medicinal and bio-organic chemistry.

Multivariate calibration (MC)

Here a number of 'signals', e.g., spectroscopic absorbances at given wavelengths/frequencies, are related to the amounts of various analytes in a set of 'calibration samples'.[34] Apparently, this is a different type of model from the LFER/ETR of physical organic chemistry, but interestingly it has the same shape and form, and hence may be a related interpretation.

Clearly, a linear relationship between a set of 'analyte' concentrations and a set of absorbencies at different wavelengths/frequencies is valid only for very similar samples. Hence, we can see the 'multivariate standard curve' of multivariate calibration as the idealized form (first order perturbations) of a general relationship between analyte concentration and spectral absorption at different wavelengths/frequencies. According to the LFER/ETR interpretation, this type of relationship should be valid between *any* property (here denoted y) – not only analyte concentrations – and any type of signals measured on the samples, provided that the samples are similar, including a limited variation of the property, y. This explains how we can 'calibrate' on y= [strength of paper] using near infrared reflectance (NIR) spectra as 'signals', as well as on y=[taste of whisky] using gas chromatograms as 'signals'.

Example, multivariate calibration. As an example, we present the first analytical chemical application of multivariate calibration, based on the fluorescence spectroscopic determination of lignin sulfonate (LS), humic acid (HA), and optical whitener (OW), in sea water.[35,36] The main difficulty at the time was that all three analytes – LS, HA, and OW – had very similar excitation–emission spectra, and hence were difficult to determine precisely in the presence of the others.

The UV emission spectra of 16 mixtures of three constituents were recorded between 320 and 540 nm. The emission intensities at 27 equally distributed wavelengths were used. In this way a calibration set consisting of a 16×27 data matrix $\boldsymbol{X}$ was formed, which described the emission at the frequencies. The concentrations of the three constituents for the spectra formed a 16×3 matrix ($\boldsymbol{Y}$).

Interestingly, the appropriate use of the whole excitation–emission spectra in the data analysis gave an attractive solution; the three constituents could be precisely determined ($R2 = 0.99$ and $Q2 = 0.94$) even in the presence of the others, and the (overlapping) spectral regions associated with each analyte could be identified as illustrated in Fig. 1a–c.

The predictive capability of the model was further investigated by preparing nine new mixtures. Their spectra were recorded and digitized as before. These samples were not used to update the model, their spectra ($\boldsymbol{X}$-data) were fitted to the model and the concentration of the constituents were predicted. The composition of the new samples (17–25) were well predicted, except for the sample 25, see Fig. 2a–c. Interesting, this sample did not fit well to the $\boldsymbol{X}$-space of the model, i.e. the spectrum was atypical for the samples in the calibration set shown by its class distance (DModX) which is much larger than of the other samples.

Quantitative structure–activity relationships, and quantitative sequence–activity models, QSAR and QSAM

The similarity/analogy models such as the Hammett relation can be extended to more complicated systems including those of biology. It was found by Hansch that a model of type (1) could model many data sets where the biological activity had been measured on a series of similar chemical compounds.[37] These models are often called QSARs for quantitative structure–activity relationships.[38–39] The Hansch models needed at least two 'scales', the ordinary Hammett scale (σ_i), plus a 'lipophilicity scale' denoted as π_i. Typically the square of the latter is also needed in the model, indicating the need for third-order terms in the perturbation model.

Instead of a separate estimation of the substituent scales such as π_i and σ_i, one may directly include a large

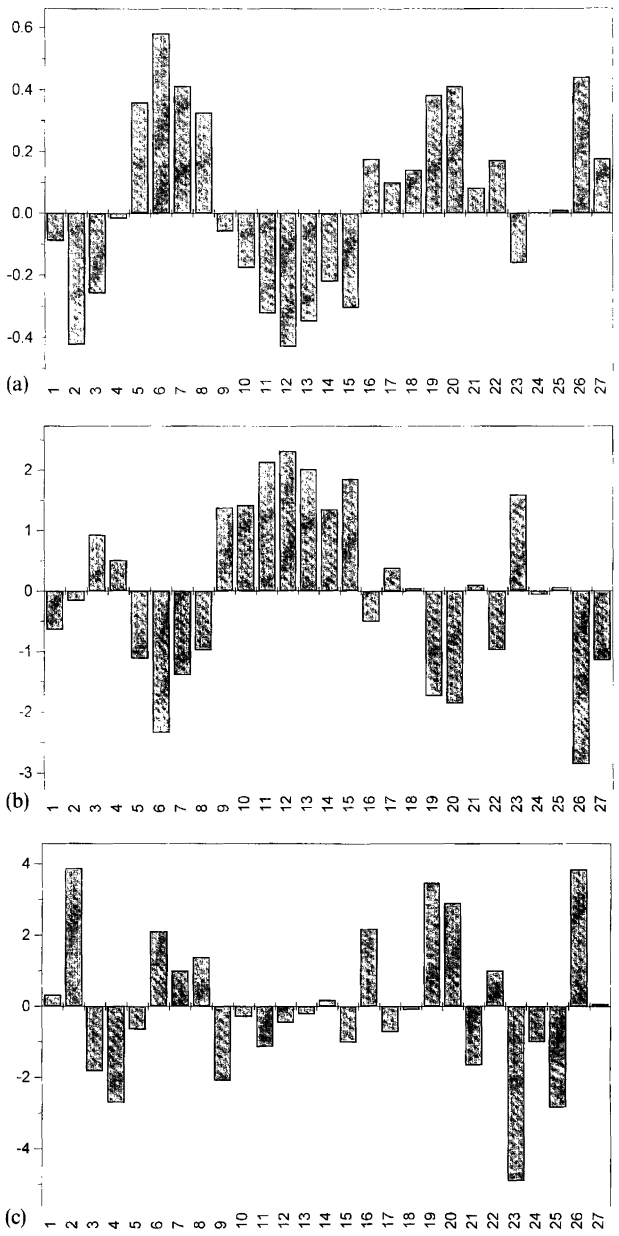

Fig. 1. The PLS regression coefficients plotted against the variable number (ordered with increasing wavelength) for the three different constituents: (a) ligninsulfonate, (b) detergent and (c) humic acid.

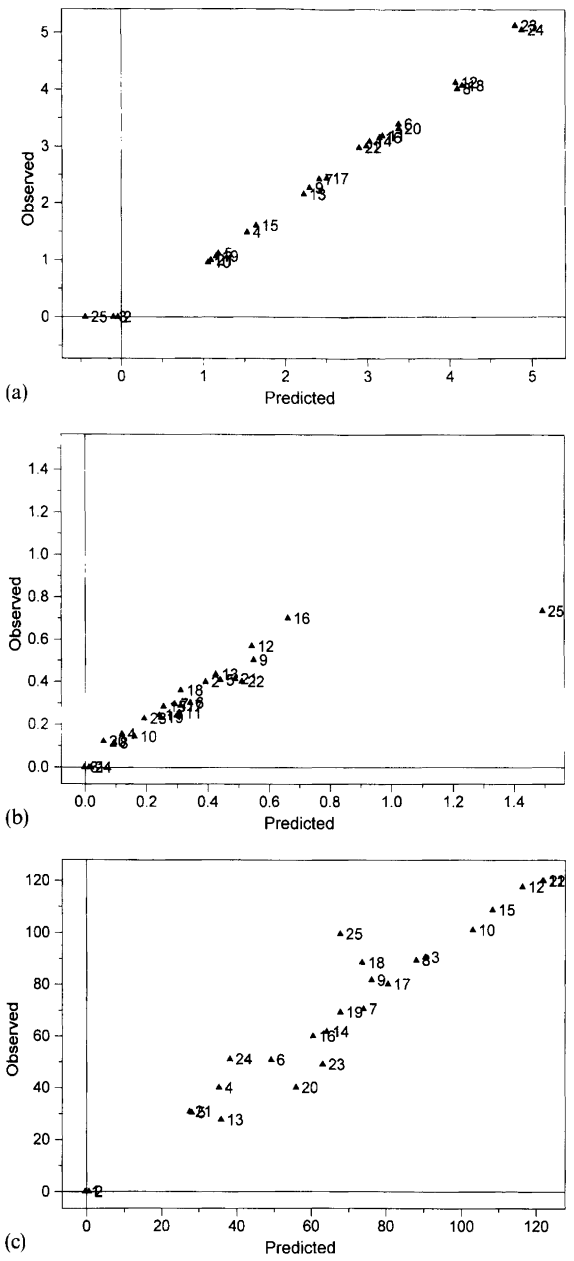

Fig. 2. Predicted values of the three constituents. Samples 1–16 are used as the training-set and 7–25 as the test-set.

number of 'raw properties' in the QSAR model. If, in addition, many sites in the molecule are modified, the resulting number of variables is large, even if only a few parameters are used to describe the change at each site.

An interesting property of the PLS model, as well as the other chemometric models based on eqn. (1), is that any number of collinear (correlated) variables can be incorporated into the analysis without difficulty. As long as a small number of model dimensions, components (A), is used in the modeling, the model parameters as well as its predictions are stable.

The PC and PLS components have the same shape as the 'effects' in the physical organic chemistry models, making it tempting to interpret them in the same way. This is closely related to the concept of latent variables – inherent properties – that we indirectly observe via a number of measured (manifest) variables.

Hence, we arrive at PLS models as generalizations of the few-term similarity/analogy models, and simultaneously, as generalizations of linear regression models with very many and collinear variables. The fact that PLS works well in practice may indicate that these interpretations are related, and that either or both may be used for the interpretation of actual cases.

PLS modeling is widely used in QSAR, from few term models to 3D QSARs with tens of thousands of structure

descriptors in COmFA and GRID models.[40–42] This type of application concerns some of the most complicated systems that chemistry deals with, and it is a credit to the philosophical foundations of these models in physical organic chemistry, i.e., quantification of the analogy principle, that they work so remarkably well.

We illustrate this approach with some examples, a QSAR of dipeptides, structure classification of a set of enzymes and a general overview of all enzymes in an entire genome.

Example, dipeptides (inhibiting angiotensin converting enzyme. A series of 58 dipeptides which inhibit angiotensin-converting enzyme was characterized by the three principal properties z_1, z_2 and z_3 in each of the amino acid positions. Thus each dipeptide was characterized by six values. In addition, to account for weak non-linear behavior between the biological data and the physicochemical characterization, square terms ($S1*S1$, etc.) and cross-terms ($C1*C2$, etc.) of the z-scales were added. The biological activity was expressed as $6+\log(1/I_{50})$, where I_{50} is the concentration (in μM) inhibiting 50% of angiotensin-converting enzyme. The biological data are from a compilation by Cheung *et al.*[43]

A QSAR was calculated based on the complete set of 58 dipeptides.[24,44] PLS analysis resulted in a model with two significant latent variables according to cross-validation ($R^2=0.78$ and $Q^2=0.68$) and thus explained well the variation in biological data (Fig. 3). The joint influence of the two latent variables on the original variables can be expressed as scaled and centered PLS regression coefficients (CoeffCS) as shown in Fig. 4. The plot shows that it is the z_1 and z_2 scales in position 2 are the most influential scales. The response surface in Fig. 5 illustrates the relationship between the settings of the two most influential scales and the biological activity.

The system studied in this example is complex but still it is possible to derive a rather simple model with good predictive capability. One possible explanation is that the LFER principle also is applicable in this and in similar systems.

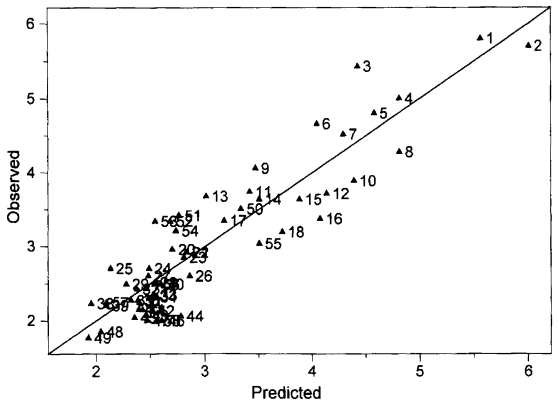

Fig. 3. The observed biological activities plotted against the calculated activities for the dipeptides.

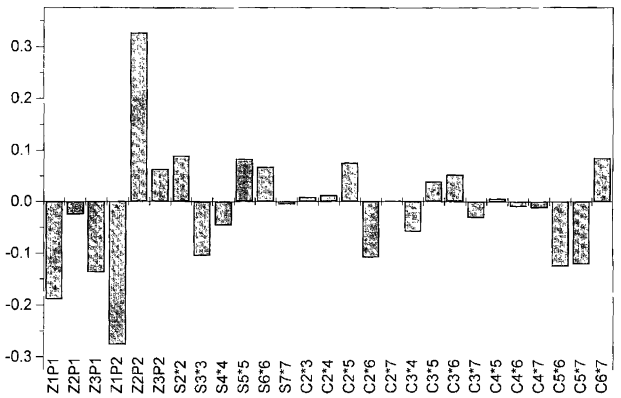

Fig. 4. Variable influence of the biological activity in the dipeptide example expressed as regression coefficients.

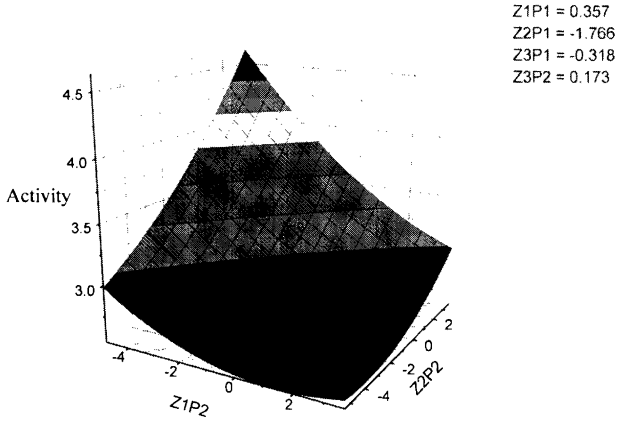

Fig. 5. Response surface plot showing the influence of the z_1 and z_2 scales in position 2 on the biological effect. The other variables (z-scales) are kept constant at their average values.

Examples of quantification of peptide sequences based on an ACC description followed by multivariate analysis. In many peptide QSAR and classification problems the number of amino acids in the investigated sequences differs between the compounds in the set. For example, so called signal peptides usually consist of 15–35 amino acids. Other examples are enzymes that can vary in magnitude by a factor of 10 (commonly between 100 and 1000 amino acids). It is also imperative to use multivariate methods to describe and discriminate parts of proteins, e.g., sequences with similar 3-D folding.

The multipositional description used in the dipeptide example above also has a limitation in that a shift of an important sequence of amino acids by one or a few amino acid positions will result in a remarkable shift in the multivariate description. To cope with this problem, we have proposed auto cross covariances (ACC) of sequences based on principal properties of amino acids as a preprocessing method.[45,46] This preprocessing method gives a set of variables that are independent of the sequence length. The number of created variables is dependent only on the number of lags used in the ACC approach and the number of principal properties used.

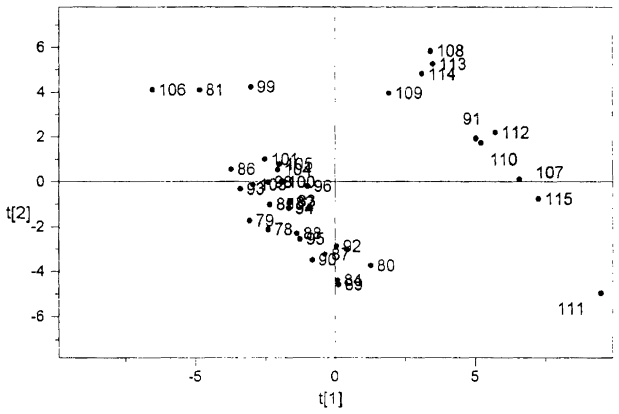

Fig. 6. PLS discriminant score plot based on an ACC description showing a difference between TIM barrels (Nos. 91, 107–115) and the remaining cyfoplasmic proteins.

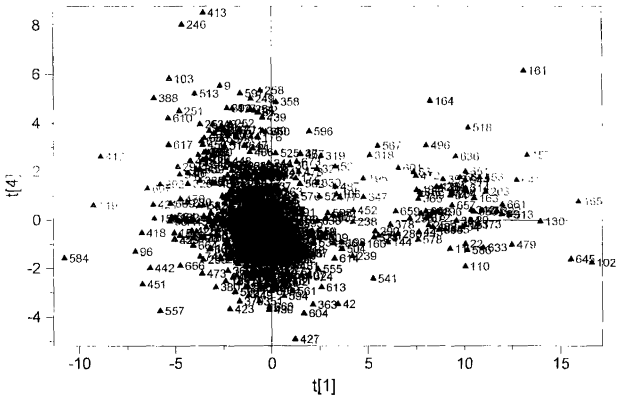

Fig. 7. A PCA score plot based on an ACC description of all the enzymes in *Mycoplasma pneumoniae.*

Some classification applications investigated by principal component or PLS discriminant score plots of ACC variables are shown in Figs. 6 and 7. The first example, shows a score plot of a series of proteins, nine of which have a specific highly symmetrical folding type, the so-called TIM barrels. The TIM barrels are easily discriminated from cyfoplasmatic proteins, despite the low sequence homology among most of the TIM barrels (Fig. 6). The second example shows a principal component score plot based on an ACC representation of all the enzymes in mycoplasma, one of the first determined whole genomes.[47] A score plot (Fig. 7) shows interesting features, however, the interpretation is outside the scope of the present paper.

Discussion and conclusion

Chemistry provides a very interesting field for scientific modeling. On the one hand, chemical systems are simple enough to allow some interpretation of resulting model parameters in terms of first principles, and this also allows good experimental control and hence a reasonable experimental reproducibility. On the other hand, chemical systems are so complicated that any modelling involves large approximations of various kinds – strict first principles or *ab initio* models are not possible for any chemical system of interest if solvent and reactants are included in the model.

The ETR and LFER models of physical organic chemistry, exemplified by the Brönsted, Hammett, Taft and Hansch models, have the same theoretical foundation as thermodynamics. They can be derived as first- and second-order perturbations of continuous multivariable systems, and hence have general, but local applicability. The latter means that the ranges (usually of structure variation) that these models cover are limited.

This theoretical foundation also makes these models 'similarity models', which are always applicable to data measured on a set of 'similar' systems. This, in turn, lends the same interesting interpretation to the isomorphous chemometrical and statistical models, namely the principal components, factor, and PLS models.

This isomorphism between physical organic and chemometric models could provide a needed bridge between the fields, and encourage a cultural and informational exchange to everybody's benefit. This is well illustrated by the profound importance of PLS modeling to solve multivariate calibration problem in analytical chemistry as illustrated by the example above.

In analogy to the Hammett substituent scales we have also shown how to develop scales for other types of substituent and series of molecules, in particular scales for amino acids. The scales can then in turn be used in similarity models with the same foundations as the ETRs.

We also foresee that this type of scale, to characterize series of substituents or molecules as well as QSAR modeling, will strongly increase in the future. The reason is the fast development of new screening methods for biologically active compounds in combination with use of combinatorial chemistry. This means that the accessibility of biological data for series of compounds will increase and consequently the need for QSAR modeling.

Substituent scales will also have a profound importance for the development of combinatorial chemistry. This is because substituent scales can be used as design variables in statistical experimental designs,[44,48–50] which makes it possible to make balanced series of compounds for combinatorial libraries. Such series will be much more suited to QSAR analysis followed by rational optimization of the structure than to randomly synthesized libraries. This is because designed libraries will span the possible analog space more efficiently than a randomly generated library.

The fast development in biochemistry and molecular biology also means that there will be many and interesting problems for chemometrics to deal with, where here we have just shown some examples. The genome projects create enormous amounts of sequence data and here questions about similarities and differences between sets of sequences are well suited to analysis by chemometrical methods. The new field of bioinformatics would thus benefit much from the development of chemometrics.

Thus we foresee increasing interest in chemometrical methods and a bright future for chemometrics. However, of utmost importance for chemometrics to be used and to develop in the right way is that chemometrics stays close to applications in chemistry and related fields.

Acknowledgements. Support from the Swedish Natural Science Research Council is gratefully acknowledged.

References

1. Jurs, P. and Isenhour, T. L. *Chemical Applications of Pattern Recognition,* Wiley, New York 1975.
2. Kowalski, B. R. and Bender, C. F. *J. Am. Chem. Soc. 94* (1972) 5632; *ibid. 95* (1973) 686.
3. Kowalski B. R., Ed., *Chemometrics, Mathematics and Statistics in Chemistry*, Reidel Dordrecht, Holland 1984.
4. Nilsson, N. J. *Learning Machines*, McGraw Hill, New York 1965.
5. Cover, T. M. and Hart P. E. *IEEE Transaction on Information Theory,* Vol. IT-13, No. 1 (1967) 21.
6. Malinowski E. R. and Howery D. G. *Factor Analysis in Chemistry*, Wiley, New York 1980.
7. Wold, S. and Sjöström, M. *Chem. Scr. 2* (1972) 49.
8. Wold, S. and Sjöström, M. *Eur. Fed. Chem. Engin. Congr. Use of Computers in Chemical Engineering*, 1973, Paris, France, Vol. 4, p. 25.
9. Sjöström, M. and Wold, S. *Chem. Scr. 6* (1974) 114.
10. Sjöström, M. and Wold, S. *Chem. Scr. 9* (1976) 200.
11. Wold, S. *Chem. Scr. 5* (1974) 97.
12. Sjöström, M. and Wold, S. *Acta Chem. Scand., Ser. B 35* (1981) 537.
13. Sjöström, M. and Wold. S. In: Chapman, N. B. and Shorter, J., Eds., *Correlation Analysis in Chemistry*, Plenum Press, London 1978, Ch. 1, pp. 1–54.
14. Exner, O. In: Chapman, N. B. and Shorter, J., Eds., *Advances in Linear Free Energy Relationships*, Plenum, London 1972, Ch. 1. pp. 1–69.
15. Eberson, L. *Electron Transfer Reactions in Organic Chemistry*, Springer, New York 1987.
16. Hammett, L. P. In: Chapman, N. B. and Shorter, J.,Eds., *Advances in Linear Free Energy Relationships*, Plenum, London 1972, foreword.
17. Palm, V. A. *Osnovy Kolichestvennoi Teorii organicheskikh Reaktsii (Izdateslstvo Khimiya, Leningrad, 1967) German translation, Grundlagen der Quantiativen Theorie Organischer Reaktionen*, Akademie Verlag, Berlin, DDR 1971.
18. Hellberg, S. Ph.D. Thesis. *A Multivariate Approach to QSAR*, Umeå University, Umeå, Sweden 1986.
19. Skagerberg, B. Ph.D. Thesis. *Principal Properties in Design and Structural Description in QSAR*, Umeå University, Umeå, Sweden 1989.
20. Johnsson, J. Ph.D. Thesis, *Quantitative Sequence–Activity Modeling*, Umeå University, Umeå, Sweden 1992.
21. Sandberg, M. Ph.D. Thesis, *Decoding Information from Sequences. A Multivariate Approach*, Umeå University, Umeå, Sweden 1997.
22. Sandberg, M., Eriksson, L., Jonsson, J. Sjöström M. and Wold, S. *J. Med. Chem. In press.*
23. Carlson. R. *Design and Optimization in Organic Synthesis*, Elsevier, Amsterdam 1992.
24. Hellberg, S., Sjöström, M., Skagerberg, B. and Wold, S. *J. Med. Chem., 30* (1987) 1126.
25. Sjöström, M. and Wold, S. In: Kowalski, B. R., Ed., *Chemometrics: Theory and Application*, American Chemical Society Symposium Series No. 52 (1977) pp. 243–282.
26. Dunn, W. J. III. and Wold, S. In: Van de Waterbeemd, H., Ed., *Methods and Principles in Medicinal Chemistry, QSAR: Chemometric Methods in Molecular Design Vol 2*, Verlag Chemie, Weinheim, Germany 1994, pp. 179–193.
27. Eriksson, L. Ph.D. Thesis, *A Strategy for Ranking Environmentally Occuring Chemicals*, Umeå University, Umeå, Sweden 1991.
28. Dunn, W. J. III., Wold, S. and Martin, Y. C. *J. Med. Chem. 21* (1978) 922.
29. Albano, C. Ph. D. Thesis, *Multivariat Analys av Solvolytiska Reaktivitetsdata*, Umeå University, Umeå, Sweden 1981.
30. Albano, C. and Wold, S. *J. Chem. Soc., Perkin Trans. 2* (1980) 1447.
31. Wold, H. In: Jöreskog, K. G. and Wold, H., Eds., *System Under Indirect Observation*, Vol. II, North Holland, Amsterdam 1982.
32. Wold, S., Ruhe, A., Wold, H. and. Dunn, W. J. III. *SIAM, J. Sci. Statist. Comput. 5* (1984) 735; also in Report UMINF-83.80 (ISSN 0348-0542).
33. Wold, S., Martens, H. and Wold, H. In: Ruhe, A. and Kågström, B., Eds., *Proceedings of the Conference on Matrix Pencils*, Piteå, Sweden, March 1982; *Lecture Notes in Mathematics*, Springer, Heidelberg 1983, pp. 286–293.
34. Martens, H. and Naes, T. *Multivariate Calibration*, Wiley, New York 1989.
35. Lindberg, W., Persson, J.-Å. and Wold, S. *Anal. Chem. 55* (1983) 643.
36. Sjöström, M., Wold, S., Lindberg, W., Persson, J.-Å. and Martens, H. *Anal. Chim. Acta 150* (1983) 61.
37. Hansch, C., Maloney, P. P., Fuijita, T. and Muir, R. M. *Nature 194* (1962) 178.
38. Hansch C. and Leo, A. *Exploring QSAR, Fundamentals and Applications in Chemistry and Biology*, ACS, Washignton, DC 1995; *QSAR: Chemometric Methods in Molecular Design*, Vol. 2, Verlag Chemie, Weinheim, Germany 1994.
39. Kubinyi, H. In: Mannhold, R., Kogsgaard-Larsen, P. and Timmerman, H., Eds., *QSAR: Hansch Analysis and Related Approaches. Methods and Principles in Medicinal Chemistry*, Vol. I, VCH, Weinheim 1993.
40. Cramer, R. D. III, Patterson, D. E. and Bunce J. D. *J. Am. Chem. Soc. 110* (1988) 5959.
41. Goodford, P. J. *J. Med. Chem. 28* (1985) 849.
42. Kubinyi, H., Ed., *3-D QSAR in Drug Design: Theory, Methods and Applications*, ESCOM, Leiden 1993.
43. Cheung, H.-S., Wang, F.-L. Ondetti, M. A., Sabo, E. F. and Cuchman, D. W. *J. Biol. Chem. 255* (1980) 401.
44. Hellberg, S., Eriksson, L., Jonsson, J., Lindgren, F., Sjöström, M., Skagerberg, B., Wold, S. and Andrews, P. *Int. J. Peptide Protein Res. 37* (1991) 414.
45. Wold, S., Jonsson, J., Sjöström, M., Sandberg, M. and Rännar, S. *Anal. Chim. Acta 277* (1993) 239.
46. Sjöström, M., Wold, S., Wieslander, Å. and Rilfors, L. *EMBO 6* (1987) 823.
47. Himmelreich, R., Hibert, H., Plagens, H., Pirki, E., Li, B.-C. and Herman, R. *Nucleic Acids Res. 24* (1996) 4420.
48. Hellberg, S., Sjöström, M., Skagerberg, B., Wikström, C. and Wold, S. *Acta Pharm. Yugoslavica 37* (1987) 53.
49. Sjöström, M. and Eriksson, L. In: van de Waterbeemd, H., Ed., *Methods and Principles in Medicinal Chemistry, QSAR: Chemometric Methods in Molecular Design*, Vol. 2, Verlag Chemie, Weinheim, Germany 1994, pp. 63–90.
50. Austel. V. In van der Waterbeemd, H., Ed., *Methods and Principles in Medicinal Chemistry, QSAR: Chemometric Methods in Molecular Design*, Vol. 2, Verlag Chemie, Weinheim, Germany 1994, pp. 48–62.

Received May 1, 1997.