J Mol Evol (1985) 22:272–277

Journal of
Molecular Evolution

A Multivariate Study of the Relationship Between the Genetic Code and the Physical-Chemical Properties of Amino Acids

Michael Sjöström and Svante Wold

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

Summary. The 20 naturally occurring amino acids are characterized by 20 variables: pK_{NH_2}, pK_{COOH}, pI, molecular weight, substituent van der Waals volume, seven ^1H and ^{13}C nuclear magnetic resonance shift variables, and eight hydrophobicity–hydrophilicity scales. The 20-dimensional data set is reduced to a few new dimensions by principal components analysis. The three first principal components reveal relationships between the properties of the amino acids and the genetic code. Thus the amino acids coded for by adenosine (A), uracil (U), or cytosine (C) in their second codon position (corresponding to U, A, or G in the second anticodon position) are grouped in these components. No grouping was detected for the amino acids coded for by guanine (G) in the second codon position (corresponding to C in the second anticodon position). The results show that a relationship exists between the physical–chemical properties of the amino acids and which of the A (U), U (A), or C (G) nucleotide is used in the second codon (anticodon) position. The amino acids coded for by G (C) in the second codon (anticodon) position do not participate in this relationship.

Key words: Principal components analysis — Pattern recognition — Multivariate analysis of physical properties of amino acids — Genetic code

Introduction

The origin of the genetic code has been a matter of much discussion during the past 30 years (see, e.g., Woese et al. 1966; Crick 1968; Jukes 1978; Shimizu 1982). Theories have been formulated that can be divided schematically into four groups: (a) The genetic code is a frozen accident and there has never been a direct interaction between the properties of amino acids and their encoding nucleotides. It should be possible to construct other genetic codes in which the amino acids are coded for by arbitrary combinations of the four nucleotides in the three codon positions. (b) At an early stage in evolution some kind of interaction between the amino acids and the nucleotides took place. A pattern that relates amino acids with similar properties to similar codons has been conserved as a relic. (c) A mechanism exists that relates the chemical and physical properties of the amino acids and the nucleotides to each other. (d) Evolution has favored a genetic code in which similar amino acids are coded for by similar combinations of nucleotides, which should minimize the consequences of mutations in mRNA [for a review of theory (d) and related theories, see Labouygues and Figureau (1982)].

At present, theories (b) and (c) seem to be the most frequently discussed. In the case of theory (c) Shimizu (1982) has proposed a lock and key mechanism acting at the tRNA. He suggests that a specific site in the tRNA forms a lock into which the amino acid fits. The site consists of the three anticodon bases plus a discriminating base at the fourth position from the 3′ end. The evolution proposed by theory (d) must have taken place early, since basically the same code is used in taxa ranging from viruses to vertebrates. For (b), models of how the code evolved have been proposed by Woese, Crick, and others. Crick (1968) discussed at length a theory in which the code at an early stage could discern only similar groups of amino acids. At a later stage the code became more amino acid specific and then

Offprint requests to: M. Sjöström

Table 1. Variables used to characterize the amino acids

No.	Property
1	Molecular weight
2	pK_{COOH} (COOH on C_α)[a]
3	pK_{NH_2} (NH_2 on C_α)[a]
4	pI, pH at isoelectric point[b]
5	Substituent van der Waals volume[c]
6	^1H NMR for C_α–H (cation)[d]
7	^1H NMR for C_α–H (dipolar)[d]
8	^1H NMR for C_α–H (anion)[d]
9	^{13}C NMR for C=O[e,f]
10	^{13}C NMR for C_α–H[e,f]
11	^{13}C NMR for C=O in tetrapeptide[e,f]
12	^{13}C NMR for C_α–H in tetrapeptide[e,f]
13	R_f for 1-N-(4-nitorbenzofurazono)-amino acids in ethyl acetate–pyridine–water[g]
14	Slope of plot of $1/R_f - 1$ vs mol% in H_2O in paper chromatography[h]
15	dG of transfer of amino acid from organic solvent to water[i]
16	Hydration potential or free energy of transfer from vapor phase to water[j]
17	R_f, salt chromatography[k]
18	Log P, P = partition coefficient for amino acids in octanol–water[l]
19	Log D, D = partition coefficient at pH 7.1 for acetylamide derivatives of amino acids in octanol–water[m]
20	dG = RT ln f; f = fraction of buried to accessible amino acids in 22 proteins[n]

NMR, nuclear magnetic resonance; R_f, rate of flow; dG, difference in free energy

References: [a] Merck Index (1977); [b] CRC Handbook of Biochemistry (1968); [c] Seydel and Schaper (1979); [d] Roberts and Jardetzky (1970); [e] Horsley et al. (1970); [f] Rosenthal and Fendler (1976); [g] Aboderin (1971); [h] Woese et al. (1966); [i] Nozaki and Tanford (1971); [j] Wolfenden et al. (1981); [k] Weber and Lacey (1978); [l] Pliška et al. (1981); [m] Fauchère and Pliška (1983); [n] Janin (1979)

froze. Woese et al. (1966) have been major proponents of theory (b), and more recently this theory has been discussed by Nagyvary and Fendler (1974). They proposed that the code evolved by interactions between amino acids, nucleotides, and small micelles and that a selective compartmentalization of amino acids and nucleotides according to their polarities has taken place. Recently, Reuben and Polk (1980) have shown that the dissociation constants of the complexes of AMP with methyl esters of amino acids in aqueous solution exhibit correlations with features in the genetic code. The main arguments for theory (b) are the findings that U in the second codon position codes for amino acids with hydrophobic side chains and that amino acids coded for by C in the second codon position seem to have similar polar requirements (Woese et al. 1966). Some other recognized regularities have been reported (Woese et al. 1966; Crick 1968; Jukes 1978). Weber and Lacey (1978), Lacey and Mullins (1983), and Jungck (1978) have reported correlations between the hydrophobicities of amino acids and the properties of the anticodon nucleotides. Thus it seems that similar amino acids are coded for by the same nucleotides in some cases. We note that the relationship between the genetic code and the physical–chemical properties of the amino acids plays an important role in theories (b)–(d).

By inspection of the chemical and physical properties of the amino acids, it is easy to recognize that a single variable (for example, a single hydrophobicity scale) is not sufficient to characterize the similarity or dissimilarity among the amino acids. A multivariate description of the amino acids is better suited for characterization. This paper is the first in a series addressing the relationship between the physical–chemical properties of the amino acids and the genetic code based on a multivariate description of the amino acids and using multivariate statistical methods.

Choice of Data

The data set is compiled from the literature. Basically the variables are arbitrarily chosen, since for the given problem it is reasonable to try to characterize the amino acids with many types of variables. In the present case, the amino acids are characterized both free and as derivatives, at different pHs, and in different solvent systems. Size and the electronic properties of the side chains also are represented. We can expect that the different hydrophobicity scales contain information on the hydrogen bond donor and acceptor properties of the side chains. The variables used are listed in Table 1. Only variables for which measurements exist for most of the amino acids have been considered, and in the data set only two measurements are missing (the measurements of variable 13 for Cys and of variable 16 for Pro).

Principal Components Analysis (PCA)

A single-variable description of the amino acids (for example, a hydrophobicity scale) can be visualized by plotting the measurements along a tally line. Amino acids that are similar with respect to the variable lie closer to each other on this line than do amino acids that have different properties. In the same way, if two or three variables (for example, two or three different hydrophobicity scales) are studied, the properties can be plotted against each other in two- or three-dimensional plots. By analogy to the one-dimensional plot, amino acids that lie close to each other in two- or three-dimensional plots can be considered more similar to each other than those far away from each other. This interpretation also holds in a multivariate space with more than three dimensions. However, if one has a multi-

variate description of the amino acids (as in the present case, where we have 20 variables), no direct illustration of the data set is possible. Different strategies for resolving this problem are possible. For example, pairwise plots of the variables can be studied, but these will not give a comprehensive picture of the relationship among the amino acids and the plots will be numerous. It can be informative to illustrate a multivariate data set by plotting a few latent variables, which are linear combinations of the original variables, against each other. Thus the original data are projected onto a few two-dimensional planes. The latent variables are the coordinates in the planes. Latent variables can be determined by PCA (see, e.g., Wold 1976, 1978; Wold and Sjöström 1977; Mardia et al. 1979; Malinowski and Howery 1980; Wold et al. 1984). Karhunen–Loeve expansion, factor analysis, and singular-value decomposition are names for similar or closely related techniques.

The first principal component (PC) explains the main variance in the data, the second PC explains the next largest variance in the data, and so on. This means that the new dimensions will describe much more of the variance in the data than the same number of original variables would. The PCA also acts as a filter in that random errors of measurement influence the first PCs very little. Furthermore, all or nearly all of the original variables will contribute to a few PCs. The PCA also shows which variables contribute to the new dimensions. By plotting the PC scores for each amino acid for, say, the first and second PCs and for the second and third PCs, a good illustration of the relationships among the amino acids can be obtained. In such plots similar samples can be found as clusters or show covariance with each other. We can regard such two-dimensional plots as two-dimensional windows onto a multivariate space.

The PCA model has the form

$$x_{ki} = \bar{x}_i + \sum_{a=1}^{A} b_{ai} t_{ka} + e_{ki} \qquad (1)$$

Here x_{ki} are the measurements in the multivariate characterization of the amino acids. The index k is used for the amino acids and the index i for the variables. From the measurements the parameters $\bar{x}_i$ (the mean value for each variable), the loadings b_{ai}, and the PC scores t_{ka} are determined in the present analysis by minimizing the squared sum of the residual ($\Sigma_i \Sigma_k e^2$). The number of estimated PCs is given by A. The absolute value of b tells how much a variable contributes to the PC and the sign tells whether the original variable is negatively or positively correlated with the PC. In the PCA the variables are weighted according to their variances. To give all variables the same importance, the data are usually scaled so that all variables will have the same variance. This type of scaling was used in the present analysis.

Methods for describing classes in multivariate data and finding class differences usually are called pattern recognition (PaRC) methods (Wold 1976; Sjöström and Kowalski 1979; Varmuza 1980; Wold et al. 1984). Numerous techniques are available for solving PaRC problems. In the present case, in which the number of amino acids in each case is few compared with the number of variables, traditional PaRC methods like linear discriminant analysis and linear learning machines are not applicable (Sjöström and Kowalski 1979). The PCA does not have this drawback. We have used PCA to create two-dimensional projections (often called eigenvector projections) of the multivariate data set, with the aim of illuminating the relation between the genetic code and the properties of the amino acids.

The PCA was done with the multivariate data analytical package SIMCA (Wold and Sjöström 1977; Wold et al. 1984) implemented on a microcomputer. This package utilizes an iterative PCA algorithm (NIPALS) in which the PCs are determined consecutively.

Sneath (1966) has previously presented a PCA of amino acids in connection with a quantitative structure–activity study of polypeptides. He described the amino acids with 134 qualitative variables. Examples of these variables are the presence or absence of β-CH_2, γ-CH_2, δ-CH_2, ω-COOH, ω-NH_2, and phenyl rings. A similarity matrix was calculated to which PCA was applied. However, continuous variables and variables based on chemical and physical measurements contain more information than do qualitative measures.

Other multivariate techniques such as cluster analysis and multidimensional scaling have recently been used in molecular genetics (see Rowe et al. 1984; Swanson 1984). Cluster analysis has no advantage over PCA, since PC plots can be inspected directly for groupings without previous assumptions about the number of classes and class structures, which must be made in cluster analysis. Multidimensional scaling is closely related to PCA in that it determines a number of so-called principal coordinates (similar to PCs) from multivariate data. However, multidimensional scaling is applicable only to the special case of symmetrical data matrices of the type (objects × objects) in which the elements are distances between pairs of objects.

Results and Discussion

A number of new dimensions were determined from the data set by PCA. Figure 1 presents plots of the

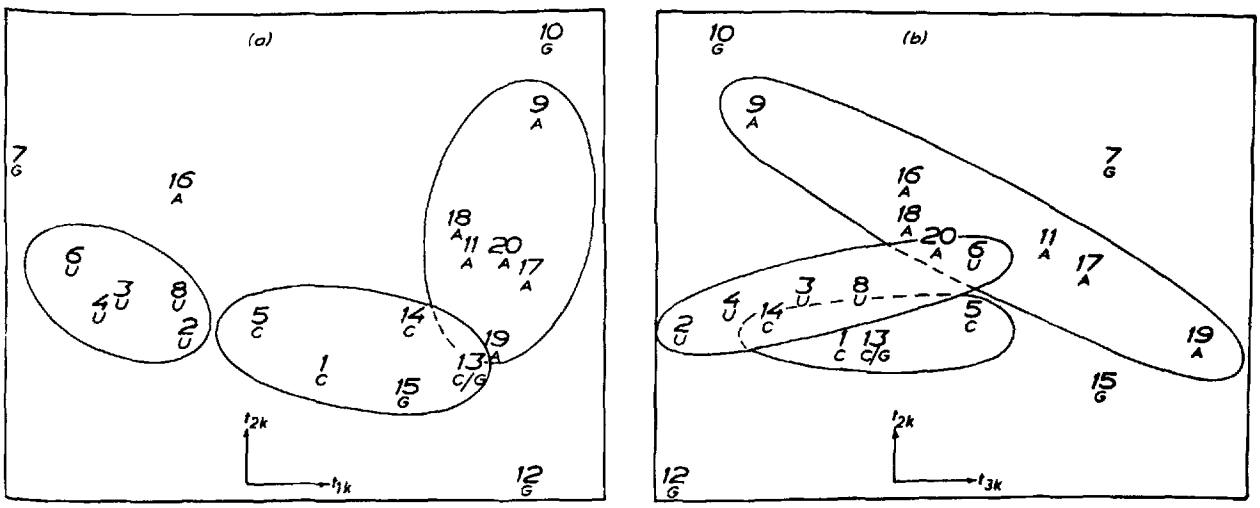

Fig. 1a,b. Plots of the first and second (a) and third and second (b) PCs against each other. The nucleotides in the second codon position are given. Amino acids coded for by U, A, and C in this position form clusters in the plots. The U, A, and C classes are encircled so that the class structures may be found easily.

Table 2. Scores (t_{ka}) for the three first PCs

No.	Amino acid	t_{k1}	t_{k2}	t_{k3}
1	Ala	−0.24	−1.74	−0.39
2	Val	−2.03	−1.03	−2.33
3	Leu	−2.90	−0.22	−1.01
4	Ile	−3.22	−0.59	−1.79
5	Pro	−1.07	−0.82	0.83
6	Phe	−3.66	0.34	0.83
7	Trp	−4.41	2.18	2.32
8	Met	−2.20	−0.41	−0.38
9	Lys	2.76	3.25	−1.57
10	Arg	2.85	4.59	−2.00
11	His	2.00	0.61	1.63
12	Gly	2.41	−4.08	−2.47
13	Ser	1.78	−1.80	−0.45
14	Thr	0.99	−0.80	−1.38
15	Cys	0.90	−2.17	2.30
16	Tyr	−2.25	1.87	0.12
17	Asn	2.57	0.19	2.00
18	Gln	1.59	1.02	0.08
19	Asp	2.10	−1.19	3.27
20	Glu	2.10	0.60	0.39

Table 3. Contribution of each variable (b_{ia}) to the three first PCs

Variable	b_{1i}	b_{2i}	b_{3i}	mpow[a]
1	−0.13	0.42	0.19	0.64
2	−0.23	0.05	−0.31	0.30
3	−0.13	−0.13	−0.12	0.03
4	0.07	0.30	−0.29	0.32
5	−0.13	0.47	0.03	0.75
6	0.06	−0.06	0.28	0.06
7	−0.11	−0.16	0.48	0.56
8	−0.01	0.01	0.54	0.54
9	−0.09	0.19	0.08	0.03
10	−0.16	0.13	0.01	0.04
11	−0.15	0.30	0.03	0.22
12	−0.16	0.11	−0.03	0.03
13	−0.38	−0.04	−0.14	0.81
14	0.32	0.01	0.24	0.51
15	−0.35	0.13	0.05	0.51
16	−0.23	−0.35	−0.15	0.67
17	0.31	−0.16	−0.20	0.51
18	−0.27	−0.17	0.17	0.35
19	−0.36	−0.10	0.02	0.57
20	−0.26	−0.33	0.03	0.55

[a] The modeling power (mpow) tells how much of the standard deviation of a variable is explained by the three-component PC model [A = 3 in model (1)]. Mpow = 1 means that the model explains all of the variation and mpow = 0, none of the variation for a variable

first and second and second and third PCs against each other. The three first PCs describe 27, 18, and 13%, respectively, of the variance in the data, for a total of 58%. The three PCs are listed in Table 2 and the contributions of each variable [the b-values in model (1)] to the PCs are listed in Table 3. The variables that contribute most to the first PC are 13, 14, 15, 17, and 19, which are the hydrophobicity measures. The size variables; additional information from the hydrophobicity variables 1, 5, 14, and 16; and variable 4 (pI) are the main contributing variables to the second PC. The third PC contains information from the electronic descriptors (variables 2, 4, 7, and 8). Except for variables 3, 6, 9, 10, and 12, all the variables make considerable contributions to the three first PCs, as seen from the modeling power measure (mpow) given in Table 2.

In Fig. 1 the nucleotides in the second codon position are given for the amino acids. From these plots we can see that the amino acids with A, U, and C in the second codon position seem to form partly overlapping clusters. Closer inspection reveals that this overlap is an artifact. In Fig. 1a, for example, Val (point 2) and Pro (point 5) are close to each other, as are Ser (point 13) and Asp (point

19). However, this is not the case in Fig. 1b. If one merged Fig. 1a and b to form a three-dimensional plot, the three classes would be well separated from each other. We also want to stress that the groupings found hold also for the second anticodon position because of the codon–anticodon base pairings A–U and G–C. However, in the following we refer only to the codon bases.

Note that in Fig. 1a Tyr (point 16) does not fall into the A class, despite having A in the second position. This is also confirmed by a SIMCA classification (Wold 1976, 1978; Wold and Sjöström 1977) in the following way: A PCA model like (Eq. 1) was determined from Lys (point 9), His (point 11), Asn (point 17), Gln (point 18), Asp (point 19), and Glu (point 20). The PCA gives the typical profile of the behavior of these amino acids. Then whether Tyr fell into this profile was checked. This was not the case, which confirmed the atypical properties of Tyr compared with those of the other amino acids coded for by A in the second codon position.

In contrast to the U, A, and C classes, the amino acids coded for by G in the second codon position seem to behave randomly in the plots. The quite different amino acids Cys (point 15), Trp (point 7), Arg (point 10), Ser (point 13), and Gly (point 12) are coded for by G. In the plots Trp and Arg are distant from Cys, Ser, and Gly. Furthermore, Ser has two alternative nucleotides (G or C) in the second codon position and falls into the C class.

Thus we have a partially asymmetric classification problem (Dunn and Wold 1980) with three well-defined classes with inherent similarity and a fourth "non-class." From an information-theoretical point of view this is sufficient for a unique class assignment of all 20 amino acids. Biologically this partial class asymmetry is interesting because it allows or is an indication of the inclusion of arbitrary new amino acids into the non-class defined by G in the second codon position.

The difference between the U, A, and C classes has also been confirmed by so-called partial least-squares (PLS) discriminant analysis (Wold et al. 1984). With the PLS method it is possible to find the optimal class-separating latent variables (Sjöström et al. 1985). However, we have chosen not to present such plots here to save the reader from a lengthy presentation of methodology. PLS discriminant plots have been shown to be informative also in the study of the relationship between the physical–chemical properties of amino acids and the other codon positions. The relationship between the suggested fourth discriminating nucleotide and the properties of the amino acids will also be investigated with this method.

In conclusion, the present investigation supports theories based on similarities between the amino acids coded for by the same nucleotide in the second codon or anticodon position. Such theories must also account for the found inconsistent behavior of the amino acids coded for by G (C) in the second codon (anticodon) position. Basically the present analysis cannot distinguish theories (b), (c), and (d) from each other. The atypical behavior of the G class is not in contradiction with theory (c), since relationships between the nucleotides and amino acid properties might also be partially asymmetric. The class-separating information provided by the hydrophobicity and hydrophilicity variables is probably relevant. These properties are important for theory (b) but probably also for theory (d). These two theories can in principle be distinguished if the properties of the nucleotides also are studied in relation to the properties of the amino acids. Some support for such relations has been given by Weber and Lacey (1978) and Jungck (1978), as mentioned in the Introduction. A multivariate method like the PLS approach (Lindberg et al. 1983) is well suited to determining whether such relations are valid; we plan to conduct such investigations.

The PCA of the amino acids supplies a quantitative metric for their study (Ninio 1983). It would be of interest to investigate if this metric is useful for describing relationships between the amino acid sequence and nucleation sites, secondary structures, and tertiary structures of proteins. The design of biologically active polypeptides (Kaiser and Kézdy 1984) is also an area where a multivariate metric could be useful, since a multitude of variables probably are important for determining the properties of polypeptides. Thus in a recent study (Hellberg et al., in press) we have found that this metric contains information predictive of the biological activity of bradykinins and some other peptide families.

Acknowledgment. We are grateful for financial support from the Swedish Natural Science Research Council (NFR).

References

Aboderin AA (1971) An empirical hydrophobicity scale for α-aminoacids and some of its applications. Int J Biochem 2:537–544

CRC Handbook of Biochemistry (1968) CRC Press, Boca Raton, Florida

Crick FHC (1968) The origin of the genetic code. J Mol Biol 38:367–379

Dunn WJ III, Wold S (1980) Structure–activity analyzed by pattern recognition: the asymmetric case. J Med Chem 23: 595–599

Fauchère J-L, Pliška V (1983) Hydrophobic parameters π of amino acid side chains from partitioning of N-acetyl amino acid amides. Eur J Med Chem 18:369–375

Hellberg S, Sjöström M, Wold S (in press) The prediction of bradykinin potentiating potency of pentapeptides. An ex-

ample of peptide quantitative-structure–activity relationship. Acta Chem Scand B

Horsley W, Sternlicht H, Cohen JS (1970) Carbon-13 magnetic resonance studies of amino acids and peptides. II. J Am Chem Soc 92:680–686

Janin J (1979) Surface and inside volumes in globular proteins. Nature 277:491–492

Jukes TH (1978) The amino acid code. Adv Enzymol 47:375–432

Jungck JR (1978) The genetic code as a periodic table. J Mol Evol 11:211–224

Kaiser ET, Kézdy FJ (1984) Amphiphilic secondary structure: design of peptide hormones. Science 223:249–255

Labouygues JM, Figureau A (1982) L'Origine et l'évolution du code génétique. Rev Can Biol Exp 41:209–216

Lacey JC Jr, Mullins DW Jr (1983) Experimental studies related to the origin of the genetic code and the process of protein synthesis: a review. Orig Life 13:3–42

Lindberg W, Persson J-Å, Wold S (1983) Partial least squares method for spectrofluorimetric analysis of mixtures of humic acids and ligninsulfonates. Anal Chem 55:643–648

Malinowski ER, Howery DG (1980) Factor analysis in chemistry. John Wiley & Sons, New York

Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis. Academic Press, London

Merck Index, 9th ed (1977) Merck & Co, Rahway, New Jersey

Nagyvary J, Fendler JH (1974) Origin of the genetic code: a physical chemical model of primitive codon assignments. Orig Life 5:357–362

Ninio J (1983) L'Explosion des séquences: les années folles 1980–1990. Biochem Syst Ecol 11:305–313

Nozaki Y, Tanford C (1971) The solubility of amino acids and the glycine peptides in aqueous ethanol and dioxane solution. J Biol Chem 246:2211–2217

Pliška V, Schmidt M, Fauchère J-L (1981) Partition coefficients of amino acids and hydrophobic parameters π of side chains as measured by thinlayer chromatography. J Chromatogr 216: 79–92

Reuben J, Polk FE (1980) Nucelotide–amino acid interactions and their relation to the genetic code. J Mol Evol 15:103–112

Roberts, GCK, Jardetzky O (1970) Nuclear magnetic resonance spectroscopy of amino acids, peptides, and proteins. Adv Protein Chem 24:447–545

Rosenthal SN, Fendler JH (1976) ^{13}C nmr spectroscopy in macromolecular systems of biochemical interest. Adv Protein Chem 13:279–424

Rowe, GW, Szabo VL, Trainor LEH (1984) Cluster analysis of genes in codon space. J Mol Evol 20:167–174

Seydel JK, Schaper KJ (1979) Chemische Structur und biologische Aktivität von Wirkstoffen. Verlag Chemie, Weinheim

Shimizu M (1982) Molecuar basis for the genetic code. J Mol Evol 18:297–303

Sjöström M, Kowalski BR (1979) A comparison of five pattern recognition methods based on the classification results from six real data bases. Anal Chim Acta 112:11–30

Sjöström M, Söderström B, Wold S (1985) PLS discriminant plots. In: Gelsema ED, Kanal LN (eds) Pattern recognition in practice II. North-Holland Publishing Company (in press)

Sneath PHA (1966) Relations between the structure and biological activity in peptides. J Theor Biol 12:157–195

Swanson R (1984) A unifying concept for the genetic code. Bull Math Biol 46:187–203

Varmuza K (1980) Pattern recognition in chemistry. In: Berthier G, et al. (eds) Lecture notes in chemistry 21. Springer-Verlag, Berlin

Weber AL, Lacey JC Jr (1978) Genetic code correlations: amino acids and their anticodon nucleotides. J Mol Evol 11:199–210

Woese CR, Dugre DH, Saxinger WC, Dugre SA (1966) The molecular basis for the genetic code. Proc Natl Acad Sci USA 55:966–974

Wold S (1976) Pattern recognition by means of disjoint principal components models. Pattern Recognition 8:127–139

Wold S (1978) Cross validatory estimation of the number of components in factor and principal components models. Technometrics 20:397–405

Wold S, Sjöström M (1977) SIMCA: a method for analyzing chemical data in terms of similarity and analogy. In: Kowalski BR (ed) Chemometrics: theory and application. American Chemical Society, Washington DC, pp 243–282 (ACS symposium series no 52)

Wold S, Albano C, Dunn WJ III, Edlund U, Esbendsen K, Geladi P, Hellberg S, Johansson E, Lindberg W, Sjöström M (1984) Multivariate data analysis in chemistry. In: Kowalski BR (ed) Chemometric—mathematics and statistics in chemistry. Reidel, Dordrecht, pp 17–96 (NATO ASI series C: mathematical and physical sciences, no 138)

Wolfenden R, Andersson L, Cullis PM, Southgate CCB (1981) Affinities of amino acid side chains for solvent water. Biochemistry 20:849–855

Received May 29, 1984/Revised May 20, 1985