Partial Least-Squares Regression of Fourth-Derivative Ultraviolet Absorbance Spectra Predicts Composition of Protein Mixtures: Application to Bovine Caseins

Guillermo E. Arteaga,* Yasumi Horimoto, Eunice Li-Chan, and Shuryo Nakai

Department of Food Science, The University of British Columbia,
Vancouver, British Columbia, Canada V6T 1Z4

A method is described for the quantitative determination of the composition of protein mixtures using fourth-derivative ultraviolet spectroscopy. The methodology was demonstrated by quantifying mixtures of the three main bovine caseins (α_{s1}-, β-, and κ-casein). The 244–296-nm spectra region of protein mixtures of known composition, dissolved in 4 M guanidine hydrochloride, was used for the development of prediction models using the multivariate method of partial least-squares regression analysis (PLSR), implemented in a commercial software. The standard errors of prediction for 16 test samples were 13.4, 5.5, and 11.9% for α_{s1}-, β-, and κ-casein, respectively, and the correlations between measured and predicted composition were 0.91, 0.99,.and 0.94 for the three proteins, respectively. This method was found to be faster and simpler than alternate procedures (e.g., gel electrophoresis), suggesting the possibility of using this methodology as a quality control tool in industrial protein fractionation environments.

Keywords: *Caseins; protein determination; PLS; multivariate analysis*

INTRODUCTION

Protein purification is one of the most common procedures in protein research and is also becoming more important as an industrial process. Two groups of techniques are normally involved in a protein purification problem. The first group includes the methodologies for the purification/extraction of the desirable protein(s) (e.g., precipitation and chromatographic techniques). The second group encompasses techniques for the quantitative evaluation of the purification scheme. Protein recovery, biological activity, and purity need to be measured to assess the efficiency of the different purification stages (Ersson et al., 1989). Electrophoresis is frequently used for judging the reproducibility and outcome of each separation step. Depending on the type of protein, other chemical, enzymatic, and immunological methods can also be applied.

Gel electrophoresis gives to the researcher a clear picture of the complexity of the sample and, in particular, what other proteins (contaminants) are present. The amount of each component, its molecular weight, its isoelectric point, and even some immunological properties, if the proper antiserum is available, can be obtained in two or three runs (Ersson et al., 1989). However, electrophoresis has some limitations. Electrophoresis is usually only able to give a qualitative or semiquantitative estimation of the relative purity of the protein components. Different proteins vary in the extent of binding of stains such as Coomassie blue or silver stain, so a separate standard curve for each protein in needed for quantitative analysis. Furthermore, stain uptake is dependent on the electrophoresis gel matrix and staining conditions (e.g., temperature and time), which are not always easy to control precisely. The incorporation of a known quantity of an internal standard protein with each gel may be used to correct for these variations.

UV spectroscopy is widely used for the quantification of concentration of homogeneous protein solutions. This requires knowledge of the extinction coefficient of the protein to be measured. The UV absorption spectra of a protein depend on the properties of the constituent aromatic amino acids. Although the spectra of these amino acids are different, their combined presence in many proteins produces the broad UV absorption spectra typical of most proteins. Due to the similarity of the UV spectra of most proteins, the application of UV absorption for the estimation of individual proteins in a protein mixture has been assumed to be rather limited.

In this paper we present a new approach for estimating the concentration of the individual components in a protein mixture. This quantitative analysis of protein mixtures involves the analysis of the fourth-order derivative of the UV absorption spectra with the multivariate calibration technique of partial least squares (PLS). A similar approach was reported by Mach et al. (1989), who used second-derivative UV spectra analysis and a proprietary multicomponent analysis to simultaneously determine the concentration of the three main classes of proteins (the α-, β-, and κ-crystalline) in mammalian ocular lenses. However, fourth-derivative spectroscopy provides better resolution than the first and second derivatives, and locations of maxima occur at the same wavelength as for the original absorption spectrum (Padrós et al., 1984). This technique has been used to study the state of aromatic residues in proteins (Padrós et al., 1982, 1984; Mozo-Villarías et al., 1991), and recently it has also been used for the rapid quantification of tyrosine in proteinaceous materials (Botsoglou et al., 1993). Since fourth-derivative spectroscopy is essentially a resolution enhancement technique, differences between the UV spectra of various proteins will tend to be magnified (Padrós et al., 1984),

* Address correspondence to this author at the Centro de Investigacion en Alimentacion y Desarrollo, A.C., Carretera a la Victoria Km. 0.6, Apartado Postal 1735, Hermosillo, Sonora, Mexico CP 83000 [fax (62) 80-0421].

0021-8561/94/1442-1938$04.50/0

Table 1. Ten-Point Experimental Design for Protein Mixtures

	proportion		
mixture	α_{s1}-casein	β-casein	κ-casein
M1	1.00	0.00	0.00
M2	0.00	1.00	0.00
M3	0.00	0.00	1.00
M4	0.33	0.33	0.33
M5	0.50	0.50	0.00
M6	0.00	0.50	0.50
M7	0.50	0.00	0.50
M8	0.67	0.17	0.17
M9	0.17	0.67	0.17
M10	0.17	0.17	0.67

making possible the quantitative analysis of the composition of protein mixtures.

To demonstrate this methodology, the simultaneous quantification of the three main bovine milk caseins (α_{s1}, β, and κ) in model mixtures is presented. Casein and caseinate are widely used as functional food proteins. Fractionation of caseins may yield products with improved functionality, and in recent years there has been increasing interest in new separation processes (e.g., ultrafiltration, microfiltration, selective precipitation, and chromatographic processes) for milk proteins, in recognition of the potential for application of discrete protein fractions as functional, nutritional, and pharmaceutical agents (Donnelly, 1991). The enrichment of the β-casein component or the depletion of the α_{s1}-casein component from crude bovine casein, to make it more similar to human milk, has been the subject of several investigations (Murphy and Fox, 1991; Li-Chan and Nakai, 1988). The potential establishment of commercial casein separation processes will create a need for simple methods as alternatives to gel electrophoresis for quantitative evaluation of casein fractionation.

MATERIALS AND METHODS

Materials. α_{s1}- and κ-casein were prepared from skim milk as described by Zittle and Custer (1963). β-Casein, *N*-acetyl-L-tyrosine ethyl ester, *N*-acetyl-L-tryptophan amide, and *N*-acetyl-L-phenylalanine ethyl ester were obtained from Sigma Chemical Co. (St. Louis, MO). Sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE) using the Phast system (Pharmacia) with Coomassie staining indicated that the purity of each casein was greater than 85%.

Preparation of the Mixtures. An essential step in any multivariate calibration is the collection of the training set of experiments which is used to develop the prediction model. An adequate experimental design is needed to minimize the number of experiments, while maximizing the amount of useful information (Martens and Næs, 1989). The 10-point augmented simplex-centroid design of Cornell (1986) (Table 1) was used to formulate three single-protein systems, three two-protein mixtures, and four three-protein mixtures. Each of the pure caseins was dissolved in 4 M guanidine hydrochloride (Gdn-HCl) (Sigma) to a final protein concentration of 5 mg/mL. Gdn-HCl was used to avoid any possible interactions between the proteins. The different mixtures were made by mixing the pure protein solutions in the volume proportions given by the experimental design. Each of the 10 points was prepared to give a final total protein concentration of 0.75 mg/mL in 4 M Gdn-HCl. The solutions of the pure proteins and their mixtures were prepared and analyzed in duplicate.

Spectrophotometric Measurements. UV absorbance (300–240 nm) was measured with a Shimadzu UV-160 double-beam spectrophotometer with a scan speed setting of "slow" (50 nm/min). All spectra were recorded against a reference cell containing 4 M Gdn-HCl. After the UV absorbance spectra were measured, the fourth-derivative spectra were calculated using the built-in program of the spectrophotometer. As recommended by Padrós et al. (1982), a derivative wavelength difference ($\Delta\lambda$) value of 1.8 nm was used. This value depends on the measuring wavelength range and the setting of the parameter *N*. For a range $\lambda \leq 100$ nm, a setting of $N = 3$ was used. The resulting data at 1-nm intervals were then transferred to a microcomputer for analysis.

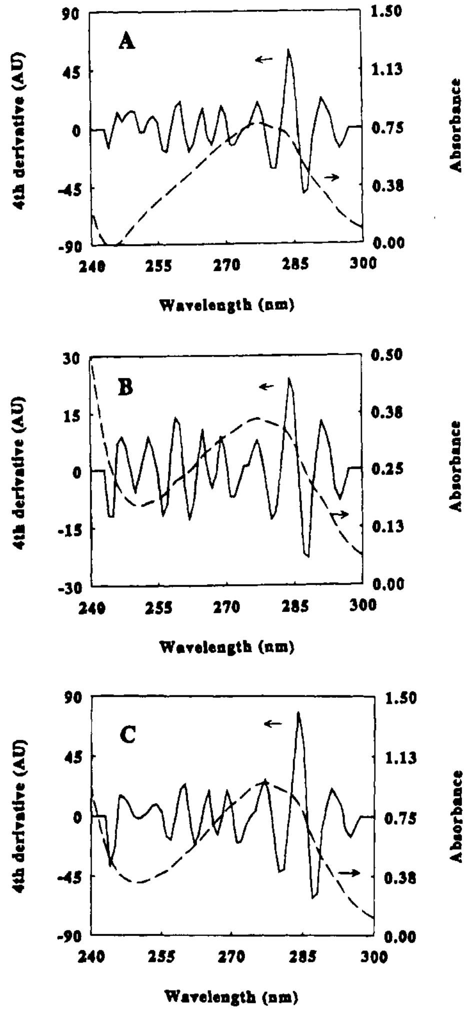

Figure 1. Ultraviolet absorption spectra (- - -, right *Y*-axis) and fourth derivative (—, left *Y*-axis) of the caseins (0.75 mg/mL) dissolved in 4 M Gdn-HCl: (A) α_{s1}-casein; (B) β-casein; (C) κ-casein.

Multivariate Calibration. Multivariate calibration was performed using the commercial software PLSplus, version 2.1, an add-on software to the spectroscopic/chromatographic software system LabCalc (Galactic Industries Corp., Salem, NH). The software was run on an IBM-PC compatible 486/25MHz personal computer. The optimum number of PLS factors was determined using cross-validation procedures as described in Martens and Næs (1989). The best region for calibration was selected by computing the correlation of the fourth derivative at every wavelength in the training spectra to the concentration of the proteins. The 244–296-nm region showed overall higher correlations and was selected for the calibration.

RESULTS AND DISCUSSION

Absorption and Derivative Spectra of Pure Caseins and Model Amino Acids. Ultraviolet absorption spectra and the corresponding fourth-derivative spectra of the three pure caseins in 4 M Gdn-HCl are shown in Figure 1. The aromatic amino acid contents per mole of casein have been reported (Eigel et al., 1984)

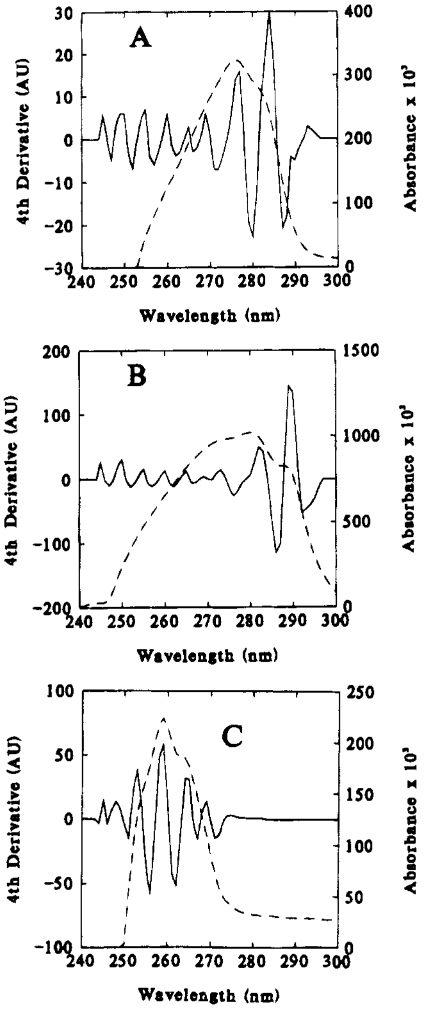

Figure 2. Ultraviolet absorption spectra (- - -, right Y-axis) and fourth derivative (—, left Y-axis) of the aromatic amino acid models dissolved in 4 M Gdn-HCl: (A) 0.7 mM N-acetyl-L-tyrosine ethyl ester; (B) 0.18 mM N-acetyl-L-tryptophan amide; (C) 1 mM N-acetyl-L-phenylalanine ethyl ester.

to be as follows: 10 Tyr, 8 Phe, and 2 Trp for α_{s1} (B-8P variant); 4 Tyr, 9 Phe, and 1 Trp for β-casein (A^2-5P variant); and 9 Tyr, 4 Phe, and 1 Trp for κ-casein (B-1P variant). The molar extinction coefficients follow the order Trp > Tyr $\gg$ Phe. Since β-casein has a lower content of Tyr and Trp than α_{s1}- and κ-casein, its UV absorbance values are lower. The maximum and minimum peak positions of the fourth-derivative spectra of proteins are somewhat related to protein conformation (Padrós et al., 1982, 1984; Mozo-Villarías et al., 1991). The main peaks observed in the range 290–275 nm are due to Trp and Tyr, while the fine structure observed at the lower wavelengths (270–240 nm) is due to Phe (Figure 2).

Multivariate Calibration of Protein Mixtures. Accurate estimation of the composition of protein mixtures requires analysis of the full range of the fourth-derivative UV spectra from 240 to 300 nm. For this type of "full spectrum" problem, multivariate calibration techniques are recommended (Martens and Næs, 1989). According to Martens and Næs (1989), the term multivariate calibration refers to the process of determining how to simultaneously use many measured variables (X_1, X_2, ..., X_k) for quantifying some target variable(s). In our case, the X variables are the value of the fourth derivative ($d^4A/d\lambda^4$) at wavelengths from 240 to 300 nm and the target variables are the concentrations of the individual proteins (α_{s1}-, β-, and κ-casein) forming the mixture. Of the several multivariate calibration techniques available, the technique known as partial least-squares regression (PLSR) has been shown to be the most useful. A detailed description of the PLSR can be found in Martins and Næs (1989).

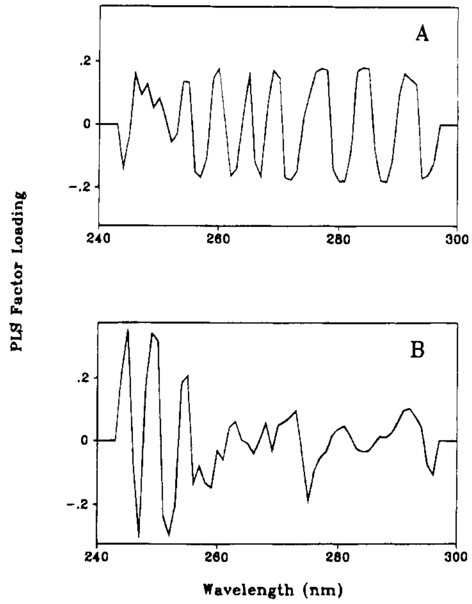

Figure 3. Factor loadings for the two PLSR factors (A and B). Large loading values indicate regions that are important for prediction.

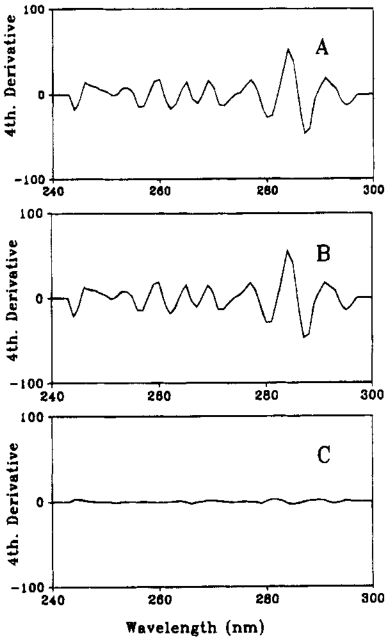

Figure 4. Original (A), PLSR reconstructed (B), and residual spectra (C) for sample M4. This sample had the following fractional composition: α_{s1}-casein, 0.33; β-casein, 0.33; and κ-casein, 0.33. The residual spectrum is the difference between the experimental and reconstructed spectra.

PLSR was applied to the fourth-derivative spectra from 244 to 296 nm of the 10 protein solutions in the training set (Table 1). To estimate the real prediction ability of the PLSR model, a full cross-validation was used. In full cross-validation one repeats the calibration n times, where n is the number of calibration samples, each time taking out one sample from the whole calibration set. At the end of each calibration, the composition (i.e., α_{s1}-, β-, κ-casein proportions in the

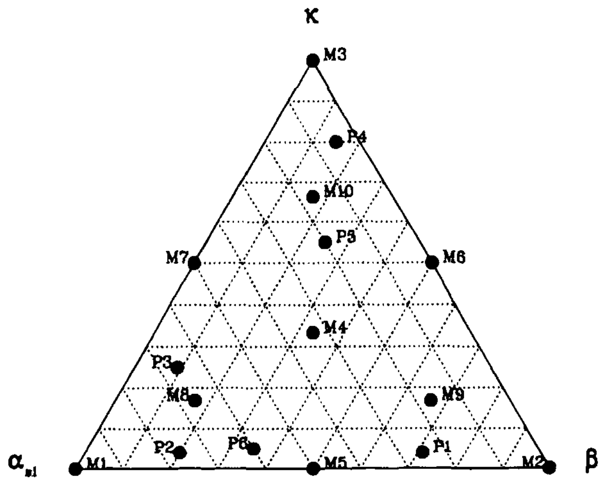

Figure 5. Triangle plot showing the composition of the 16 mixtures used as testing set (M1–M10 and P1–P6). Ten samples with composition M1–M10 were used to develop the calibration models and formed the 10-point-augmented simplex-centroid design of Cornell (1986).

Table 2. Sixteen-Point Testing Set for Protein Mixture

	proportion		
mixture	α_{s1}-casein	β-casein	κ-casein
P1	0.25	0.71	0.04
P2	0.76	0.20	0.04
P3	0.66	0.09	0.25
P4	0.05	0.15	0.80
P5	0.20	0.25	0.55
P6	0.60	0.35	0.05
M1	1.00	0.00	0.00
M2	0.00	1.00	0.00
M3	0.00	0.00	1.00
M4	0.33	0.33	0.33
M5	0.50	0.50	0.00
M6	0.00	0.50	0.50
M7	0.50	0.00	0.50
M8	0.67	0.17	0.17
M9	0.17	0.67	0.17
M10	0.17	0.17	0.67

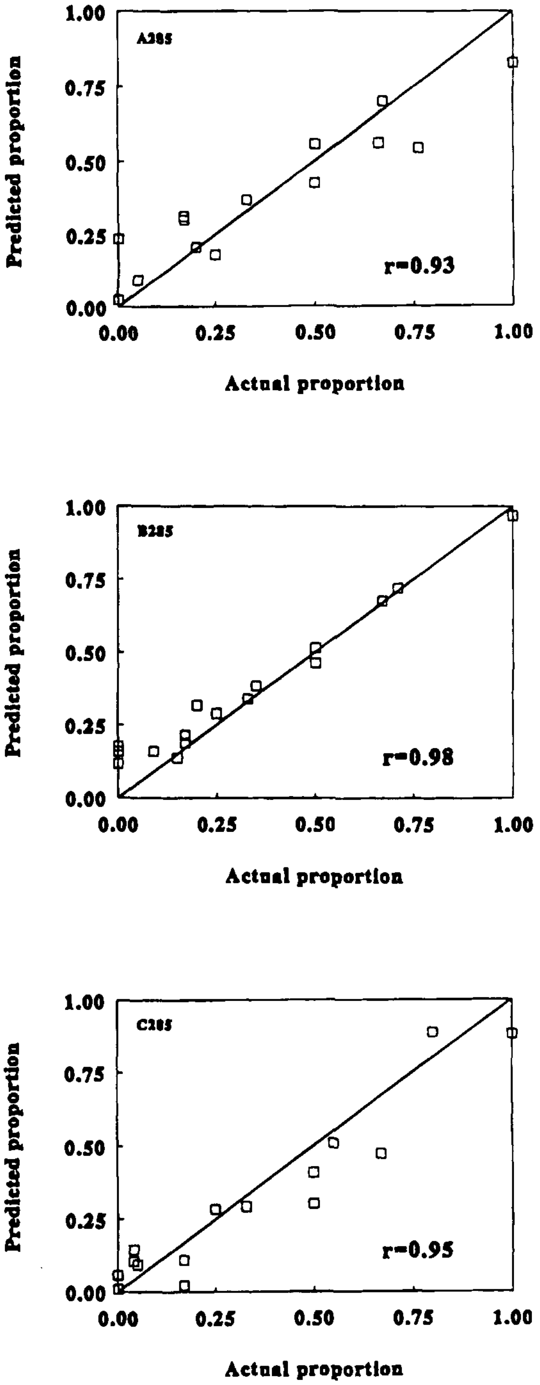

Figure 6. Prediction ability of the PLSR models on the testing set (cross-validated results); predicted proportion for (A) α_{s1}-casein, (B) β-casein, and (C) κ-casein. The solid lines represent perfect matches between actual and predicted proportions.

mixture) of the excluded sample is calculated and compared to the known composition values. In the end, all of the calibration samples have been treated as prediction objects and a standard error of prediction (SEP) is calculated (Martens and Næs, 1989). The standard errors of prediction for the cross-validated PLS results were 10.5, 5.0, and 8.3% for α_{s1}-, β-, and κ-casein, respectively. The correlation coefficients between actual and PLS-predicted mixture composition were 0.892, 0.975, and 0.931 for α_{s1}-, β-, and κ-casein, respectively.

The optimum model required two PLSR factors. These factors, also called loading, loading vectors, or dimension, are a mathematical representation of all the independent variations in the data. The loadings are shown in Figure 3. Large loading values indicate spectral regions that are more important for the prediction. Linear composition of the factors can be used to reconstruct the measured spectra to within the noise and model error. For all samples very good reconstruction of the fourth-derivative absorbance spectra was obtained. A typical result is presented in Figure 4.

To further confirm the adequacy of the PLSR model, the compositions of 16 test samples were predicted using the developed PLSR model. The compositions of these 16 samples are shown in Figure 5 and Table 2. Although the compositions of 10 of these samples (M1–M10) coincided with those of the training set, it is important to point out that these samples were prepared separately and not the ones used in the development of the calibration model.

The prediction results for these samples are shown in Figure 6. The standard errors of prediction for these sample were 13.4, 5.5, and 11.9% for α_{s1}-, β-, and κ-casein, respectively. These results confirm the cross-validation results that the prediction ability for α_{s1}-casein and κ-casein was not as good as for β-casein, which may be due to the similarity of the absorption spectra between α_{s1}-casein and κ-casein (Figure 1).

Conclusions. It has been demonstrated that fourth-derivative UV absorption spectroscopy coupled with the multivariate calibration method of PLSR can provide useful estimates of the composition of a protein mixture. The small prediction errors (5–10.5%) suggest that this technique could be very useful as a quality control tool in industrial protein fractionation environments. Another potential application of this methodology is in the quantification of other chemical components (e.g., nucleic acids) in protein mixtures.

Notwithstanding some necessary mathematical manipulation of the data, this methodology is significantly faster and experimentally simpler than gel electrophoresis. Because of its advantages, we believe methodologies such as the one described have great potential as a complementary tool to electrophoresis in quantitative estimation of the composition of protein mixtures.

LITERATURE CITED

Botsoglou, N. A.; Fletouris, D. J.; Papageorgiou, G. E.; Mantis, A. J. Derivative Spectrophotometric Method for the Analysis of Tyrosine in Unhydrolyzed Protein, Food, and Feedstuff Samples. *J. Agric. Food Chem.* **1993**, *41*, 1635–1639.

Cornell, J. A. A Comparison Between Two Ten-Point Designs for Studying Three Component Mixture Systems. *J. Qual. Technol.* **1986**, *18*, 1–15.

Donnelly, W. J. Applications of Biotechnology and Separation Technology in Dairy Processing. *J. Soc. Dairy Technol.* **1991**, *44*, 67–73.

Eigel, W. N.; Butler, J. E.; Ernstrom, C. A.; Farrell, H. M.; Harwalkar, V. R.; Jenness, R.; Whitney, R. McL. Nomenclature of Proteins of Cow's Milk: Fifth Revision. *J. Dairy Sci.* **1984**, *67*, 1599–1631.

Ersson, B.; Rydén, L.; Janson, J. C. Introduction to Protein Purification. In *Protein Purification: Principles, High Resolution Methods, and Design*; Janson, J. C., Rydén, L., Eds.; VCH: New York, 1989; Chapter 1.

Li-Chan, E.; Nakai, S. Rennin modification of bovine casein to simulate human casein composition: effect on acid clotting and hydrolysis by pepsin. *Can. Inst. Food Sci. Technol. J.* **1988**, *21*, 200–208.

Mach, H.; Thomson, J. A.; Middaugh, C. R. Quantitative Analysis of Protein Mixtures by Second Derivative Absorbtion Spectroscopy. *Anal. Biochem.* **1989**, *181*, 79–85.

Martens, H.; Næs, T. *Multivariate calibration*; Wiley: Chichester, U.K., 1989.

Mozo-Villarías, A.; Morros, A.; Andreu, J. M. Thermal Transitions in the Structure of Tubulin: Environments of Aromatic Amino Acids. *Eur. Biophys. J.* **1991**, *10*, 295–300.

Murphy, J. M.; Fox, P. F. Fractionation of Sodium Caseinate by Ultrafiltration. *Food Chem.* **1991**, *39*, 27–38.

Padrós, E.; Morros, A.; Mañosa, J.; Duñach, M. The State of Tyrosine and Phenylalanine Residues in Proteins Analyzed by Fourth-Derivative Spectrophotometry: Histone H1 and Ribonuclease A. *Eur. J. Biochem.* **1982**, *127*, 117–122.

Padrós, E.; Duñach, M.; Morros, A.; Sabés M.; Mañosa, J. Fourth-Derivative Spectrophotometry of Proteins. *Trends Biochem. Sci.* **1984**, *9*, 508–510.

Zittle, C. A.; Custer, J. H. Purification and Some of the Properties of α_{s1}-Casein and κ-Casein. *J. Dairy Sci.* **1963**, *46*, 1183–1190.

Received for review December 16, 1993. Revised manuscript received June 8, 1994. Accepted June 17, 1994.[⊗]

[⊗] Abstract published in *Advance ACS Abstracts,* July 15, 1994.