Anal. Chem. **2001,** *73,* 64−71

Short-Wave Near-Infrared Spectroscopy of Biological Fluids. 1. Quantitative Analysis of Fat, Protein, and Lactose in Raw Milk by Partial Least-Squares Regression and Band Assignment

Slobodan Šašić and Yukihiro Ozaki*

Department of Chemistry, School of Science, Kwansei-Gakuin University, Uegahara, Nishinomiya 662-8501, Japan

The present study has aimed at providing new insight into short-wave near-infrared (NIR) spectroscopy of biological fluids. To do that, we analyzed NIR spectra in the 800−1100-nm region of 100 raw milk samples. The contents of fat, proteins, and lactose were predicted by partial least-squares (PLS) regression and band assignment in that region was investigated based upon PLS loading plots and regression coefficients. For the fat prediction, the whole set of samples was divided into two groups and the fat concentration was predicted for the samples that were not included in the calibration procedures. The correlation coefficient and root-mean-square error of prediction (RMSEP) in the better prediction run were found to be 0.996 and 0.087 wt %, respectively. Assignment of the bands due to fat was proposed based upon the regression coefficients and PLS loading weights, and the importance of a pretreatment in the prediction was discussed. Milk proteins also yielded sufficient correlation coefficients and RMSEP although the contributions of protein bands to the milk spectra were much smaller than those of the fat bands. The sizes of the calibration models for protein prediction were considered. This is the first time that good correlation coefficients and RMSEP of proteins have ever been obtained for the short-wave NIR spectra of milk. For lactose, noisy regression coefficients with limited prediction ability were obtained. Band assignment was investigated also for bands due to proteins and lactose. We propose the detailed band assignment for the short-wave NIR region useful for various biological fluids. The results presented here demonstrate that the short-wave NIR region is promising for the fast and reliable determination of major components in biological and biomedical fluids.

Recently, the short-wave near-infrared (NIR) region, 700−1100 nm, has received interest because this region is suitable for nondestructive or noninvasive analyses of biological and biomedical materials.[1−5] Because of the high transmittance of light in the 700−1100-nm region and the availability of excellent detectors, it may be possible to construct on-line sensors for the nondestructive determination of the various components in biological and biomedical materials and instruments for noninvasive diagnoses based upon short-wave NIR spectroscopy. Moreover, in the short-wave NIR region, single-beam data can be reliably exploited, which can reduce measurement time,[6] and the problems that arise from the intense water bands in NIR region can be diminished in particular cases.[7] The purpose of the present study is to explore the potential of the short-wave NIR region for quantitative analysis of biological fluids. NIR spectra in the 700−1100-nm region of raw milk have been investigated as an important example. Despite the importance of the short-wave NIR, thus far, very few investigations have been reported for band assignment in this region. To obtain insight into the chemical base for quantitative analysis, the regression coefficients and loading weights for all the main species of raw milk have been investigated in detail in the present study.

The content of major milk components is the most important factor that determines the quality of milk.[8,9] Fast and reliable methods for determining the concentration of the main milk components such as fat, proteins, and lactose, are highly desirable in the dairy industry. NIR spectroscopy seems to be a method that can meet all the analytical demands of everyday milk production.[8−15] The potential of NIR spectroscopy has been proved in many papers that reported the prediction of the milk components.[10−17] Almost all the NIR experiments of milk were

(1) Brazy, J. E. In *Near Infrared Spectroscopy*; *Clinics in Perinatalogy*; Brans, Y. W., Ed.; Saunders: Philadelphia, 1991; p 519.
(2) Ozaki, Y.; Matsunaga, T.; Miura, T. *Appl. Spectrosc.* **1992**, *46*, 180.
(3) Hoshi, Y.; Tamura, M. *Neurosci. Protocols* **1994**, *7*, 1.
(4) Oda, M.; Yamashita, Y.; Nishimura, G.; Tamura, M. *Adv. Exp. Med. Biol.* **1994**, *345*, 861.
(5) Sato, H.; Wada, S.; Ling, M.; Tashiro, H. *Appl. Spectrosc.* **2000**, *54*, 1163.
(6) Bittner, A.; Marbach, R.; Heise, H. M. *J. Mol. Struct.* **1995**, *349*, 341.
(7) Reeves, III J. B. *J. Near Infrared Spectrosc.* **1994**, *2*, 199.
(8) Association of Official Analytical Chemists. *Official Methods of Analysis*, 15th ed.; AOAC: Arlington, VA, 1990.
(9) Lensen, R. G., Ed. *Handbook of Milk Composition*; Academic Press: San Diego, 1995.
(10) Ben-Gera, I.; Norris, K. H. *Isr. J. Agric. Res.* **1968**, *18*, 117.
(11) De Wilder, J.; Bassuyt, R. *Milchwissenschaft* **1983,** *38*, 65.
(12) Baer R. J.; Frank J. F.; Loewenstein, M.; Birth G. S. *J. Assoc. Offic. Anal. Chem.* **1983**, *66*, 858.
(13) Robert, P.; Bertrand, D.; Devaux, M. F.; Grappin, R. *Anal. Chem.* **1987**, *59*, 2191.
(14) Sato, T.; Yoshino, M.; Furukawa, S.; Someya, Y.; Yano, N.; Uozumi J.; Iwamoto, M. *Jpn. J. Zootechnol. Sci.* **1987**, *58*, 698.
(15) Kamishikiryo-Yamashita, H.; Oritani, Y.; Takamura, H.; Matoba, T. *J. Food Sci.* **1994**, *59*, 313.
(16) Tsenkova, R.; Grigorov, T. *Farm Machinery* **1990**, *27*, 64.
(17) Diaz Carillo, E.; Munnoz-Serrano, A.; Alonso-Moraga, A.; Seradilla-Manrique, J. M. *J. Near Infrared Spectrosc.* **1993**, *1*, 141.

10.1021/ac000469c CCC: $20.00
Published on Web 11/23/2000

made for the 1100−2400-nm region. However, the short-wave NIR region, 700−1100 nm, should also be considered for the quantitative and qualitative analysis of milk.

To the best of our knowledge, there are only two papers published recently that give an insight into potential of short-wave NIR region for quantitative analysis of milk composition.[18,19] In the paper by Tsenkova et al.,[18] the calibrations of fat, proteins, and lactose in the 700−1100-nm region were reported and the results were compared with the predictions obtained in the standard NIR region. The regression coefficients were discussed and the wavelengths of particular importance for the calibration of fat, proteins, and lactose were tabulated. Chen et al.[19] investigated the problem of a sample cell in the 700−1100-nm region and were strictly oriented toward fat determination. They reported the first five loading weights from the fat calibration in the 700−1100-nm region.[19]

In the present study, particular attention has also been paid to the standard outputs of the regression analyses, correlation coefficients, root-mean-square error of calibration (RMSEC) and cross validation (RMSECV), and root-mean-square error of prediction (RMSEP), since we have investigated the scope of the short-wave NIR region for practical use. All the regressions have been made by the most popular regression technique, partial least-squares (PLS) regression.[20] The band assignment in the 800−1100-nm region has been proposed based upon the regression coefficients and loading weights for the three major components.

EXPERIMENTAL SECTION

Sample Preparation. Milk samples were supplied directly from cows that were under routine feeding management at the National Institute of Animal Industry, Tsukuba, Japan. Fat and protein contents in the milk samples were determined by Milkoscan 134 A/B (N Foss Electric). A total of 100 milk samples were used as the whole data set. The milk samples were taken twice a day for 2 days, once in the morning and once in the evening.

NIR Measurements. NIR transmission spectra in the 800−1100-nm region were recorded with a step size of 2 nm by an NIR Systems 6500 spectrometer. The milk samples were homogenized and incubated at 40 °C in a water bath prior to the NIR measurements. A liquid sample cell (1.0-mm path length) was used. The NIR measurements were also performed at 40 °C.

Chemometric Analysis. All chemometric analyses were carried out by Unscrambler ver. 6.1 (CAMO AS, Trondheim, Norway) software program. The optimum number of components to be used in the regression was automatically determined by the software.

RESULTS AND DISCUSSION

Fifty NIR spectra of raw milk in the 800−1100-nm region are shown in Figure 1A. A broad feature at 970 nm due to $2\nu_1 + \nu_3$ (ν_1; symmetric stretching, ν_3; antisymmetric streching) vibration of water[21] dominates the spectra and the baseline changes markedly from one spectrum to another. The baseline changes are unavoidable in pure milk spectra, and different types of pretreatment have been applied to the spectra to eliminate them.[18,19,22] One of the pretreatment methods often used is multiplicative scatter correction (MSC).[23] This method is suitable for situations where some common amplification or offset appears. In the milk spectra, the baseline changes are induced mainly by the light scattering due to fat globules.[10] On one hand, this reflection effect may be useful for the determination of the fat content, because it is related to the fat content, but on the other hand, it diminishes the importance of chemically dependent variability especially when other milk components are analyzed. Hence, we have applied MSC to the spectra, and the obtained spectra are shown in Figure 1B. A weak feature near 840 nm arises

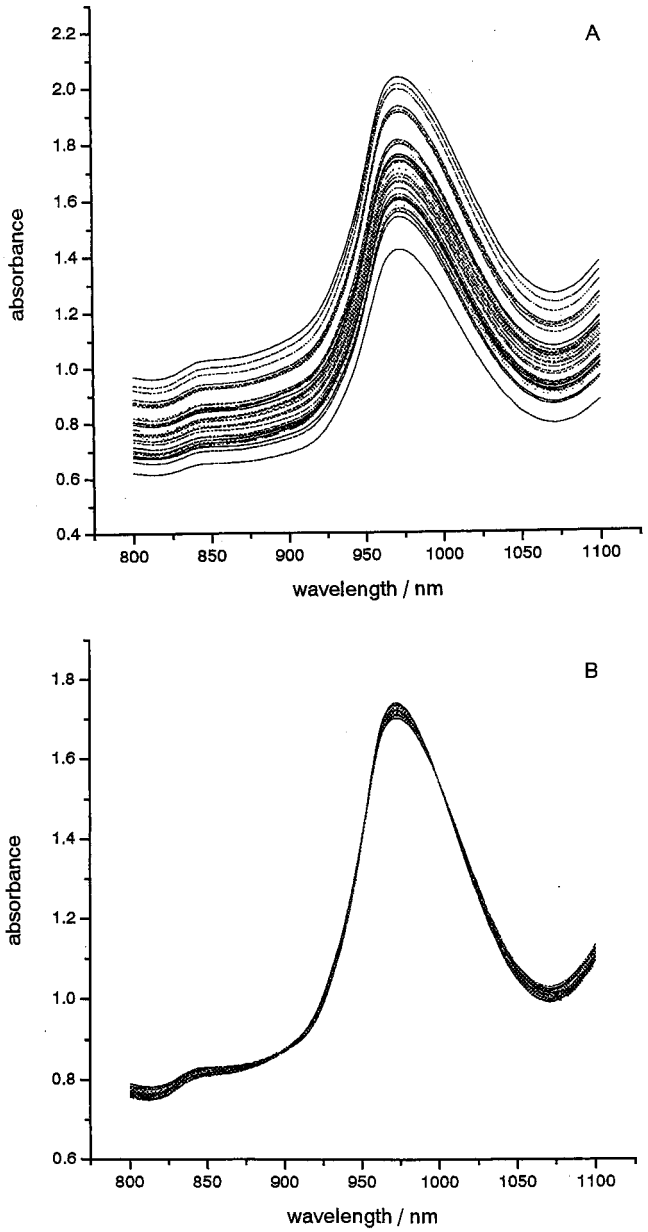

Figure 1. NIR spectra of 50 milk samples in the 800−1100-nm region before (A) and after (B) multiplicative scatter correction.

(18) Tsenkova, R.; Atanassova, S.; Toyoda, K.; Ozaki, Y.; Itoh K.; Fearn, T. *J. Dairy Sci.* **1999**, *82*, 2344.
(19) Chen, Y. J.; Iyo, C.; Kawano S.; Terada, F. *J. Near Infrared Spectrosc.* **1999**, *7*, 265.
(20) Martens, H.; Naes, T. *Multivariate Calibration;* John Wiley and Sons: Chichester, U.K., 1996.
(21) Osborne, B. G.; Fearn, T. *Near Infrared Spectroscopy in Food Analysis;* Longman Scientific and Technical; Essex, U.K., 1986.
(22) Šašić S.; Ozaki, Y. *Appl. Spectrosc.* **2000**, *54*, 1327.
(23) Geladi, P.; MacDougall D.; Martens, H. *Appl. Spectrosc.* **1985**, *39*, 491.

Table 1. Cross Validation Results for Fat Content Determination in Raw Milk from the Short-Wave NIR Spectra[a]

calib set	no. of samples	factors	slope	offset	correlation	RMSEC	RMSECV
1	40	4	1.000	0.000	0.995	0.102	0.119
2	40	6	0.999	0.007	0.994	0.070	0.083

[a] The fat statistics in calibration model 1: min = 1.83 wt %; max = 6.51 wt %; mean = 4.41 wt %; standard deviation (SD) = 1.20 wt %. The fat statistics in calibration model 2: min = 3.20 wt %; max = 5.87 wt %; mean = 4.53 wt %; SD = 0.77 wt %.

from a combination of C−H vibrations and can be ascribed to any of the main milk components.

Fat Calibration. It is well known that fat in milk can be satisfactorily predicted from NIR spectra in the 1100−2500-nm region by use of regression methods.[10−17] Thus, we used the 40 spectra of milk samples for the calibration and validated the model by leave-one-out cross validation. By use of the model developed, the fat concentration was predicted for the remaining 60 milk spectra. Two models were developed using different sample sets consisting of 40 spectra each. The spectra in each set were ordered according to the increase in the fat content. The results of the calibration for these two separate models are shown in Table 1, and the corresponding loading weights and regression coefficients are shown in Figure 2A and B, respectively. Table 1 reveals that only small differences exist between the two models that differ slightly in the ranges of fat content variation. The model with the narrower fat concentration range yields slightly better results in the cross validation. It is interesting to note that the number of components selected by the software was six for the better model while it was four for the poorer one. If the number of factors employed is increased to six in the poorer model, RMSEC and RMSECV become 0.077 and 0.094, respectively, which are quite close to the values obtained for the better model. Thus, very good correlation coefficients and quite small errors in validations can be achieved by small correction of the factors used in the regression coefficients calculation.

The loading weights shown in Figure 2A are obtained from the model that gives the better result in calibration, but the shapes of loading weights are almost identical for both models. The percentages of spectral and concentration variances of the poorer model are 89 and 83%, respectively, for the first loading, 10 and 4% for the second, 0.5 and 11% for the third, and 0.2 and 0.7% for the fourth. By the first two loadings, 99% of the spectral variances is accounted for. However, the remaining, minor, spectral variances contain a significant amount (20%) of the concentration variances. The first loading weights show distinct maximums at 928 and 968 nm. The shorter wavelength probably corresponds to the third overtone of a C−H stretching vibration of fat.[21,24] The same band was observed by Tsenkova et al.,[18] but it was ascribed to a C−O vibration of oil. The peak at 968 nm undoubtedly arises from the $2v_1 + v_3$ stretching vibration of water. The second loading weights develop features similar to those the first one. The main variations are still located at 928 and 972 nm, but two additional peaks appear; the first is observed at 840 nm, and the second is seen at 1018 nm. The later may be assigned to a combination mode, 2 C−H stretching and 3 C−H deformation, of the CH_3 groups originating from fat.[21,24] Besides the peaks at 922 and 972 nm, the third loading weights show significant features at 880, 948, 1018, and 1042 nm. The maximum at 1042 nm is assigned to a combination of 2 C−H stretching and 2 C−H deformation vibrations of oil.[21] The peak at 948 nm probably arises from interaction between fat and water, although it might be caused by protein−water interaction, as will be shown later. As for the band at 880 nm, there are very weak and unclear indications of fat bands near 880 nm in the second-derivative spectra of fat reported by Chen et al.[19] Osborne and Fearn[21] did not tabulate any overtone or combination bands of fat below 900 nm. We infer that the band at 880 nm comes from the third overtone of a C−H stretching vibration of fat. Assignment of this band to fat is supported by regression coefficients (Figure 2B). The fourth loading weights in Figure 2A have peaks mostly at the same positions as the third ones. The wavelengths noted in these four loading weights are in good accordance with those found as important wavelengths for the fat determination in Atlantic salmon and Atlantic halibut fishes by NIR spectroscopy.[25,26]

It must be stressed here that the assignment of the bands based only on PLS loading weights and regression coefficients might be doubtful. PLS loading weights offer information important for calibration of a given species, but these features are not necessarily only from that species. Hence, to avoid mistakes in the tentative assignment proposed here, the relevant data available (spectra of the species of interest and partial assignments published) were extensively used.

The regression coefficients shown in Figure 2B commonly have a strong positive maximum at 928 nm and a negative one at 948 nm. Although the regression coefficient obtained after four factors for calibration set 1 shows better band-structured features and contains all the bands that can be related to fat, its prediction ability is slightly inferior to the ability of the one that is obtained for calibration set 2 after six factors. Figure 3A and B depicts predictions of the fat concentration from the 60 milk samples by the six- and four-factor regression coefficients, respectively. None of the samples from the prediction sets was included into calibration models by which its fat concentration was predicted. A significantly smaller RMSEP obtained by the model with six factors reveals that very important wavelengths for the fat calibration are below 950 nm. Besides the band at 928 nm due to the third overtone of the C−H stretching mode of fat, the bands at 890 and 840 nm as well as the band at 948 nm that is supposed to come from fat−water interaction are demonstrated as useful for the fat calibration. The intensity variations in the water band at 970 nm and fat features at 1018 and 1042 nm are not sufficiently representative for the fat content, although the loading weights (Figure 2) suggest so. On the other hand, the bands at 890 and

(24) Workman, Jr. J. J. *Appl. Spectrosc. Rev.* **1996**, *31*, 251.
(25) Wold J. P.; Jakobsen, T.; Krane, L. *J. Food Sci.* **1996**, *16*, 74.
(26) Nortvedt, R.; Torrisen, O. J.; Tuene, S. *Chemom. Intell. Lab. Syst.* **1998**, *42*, 199.

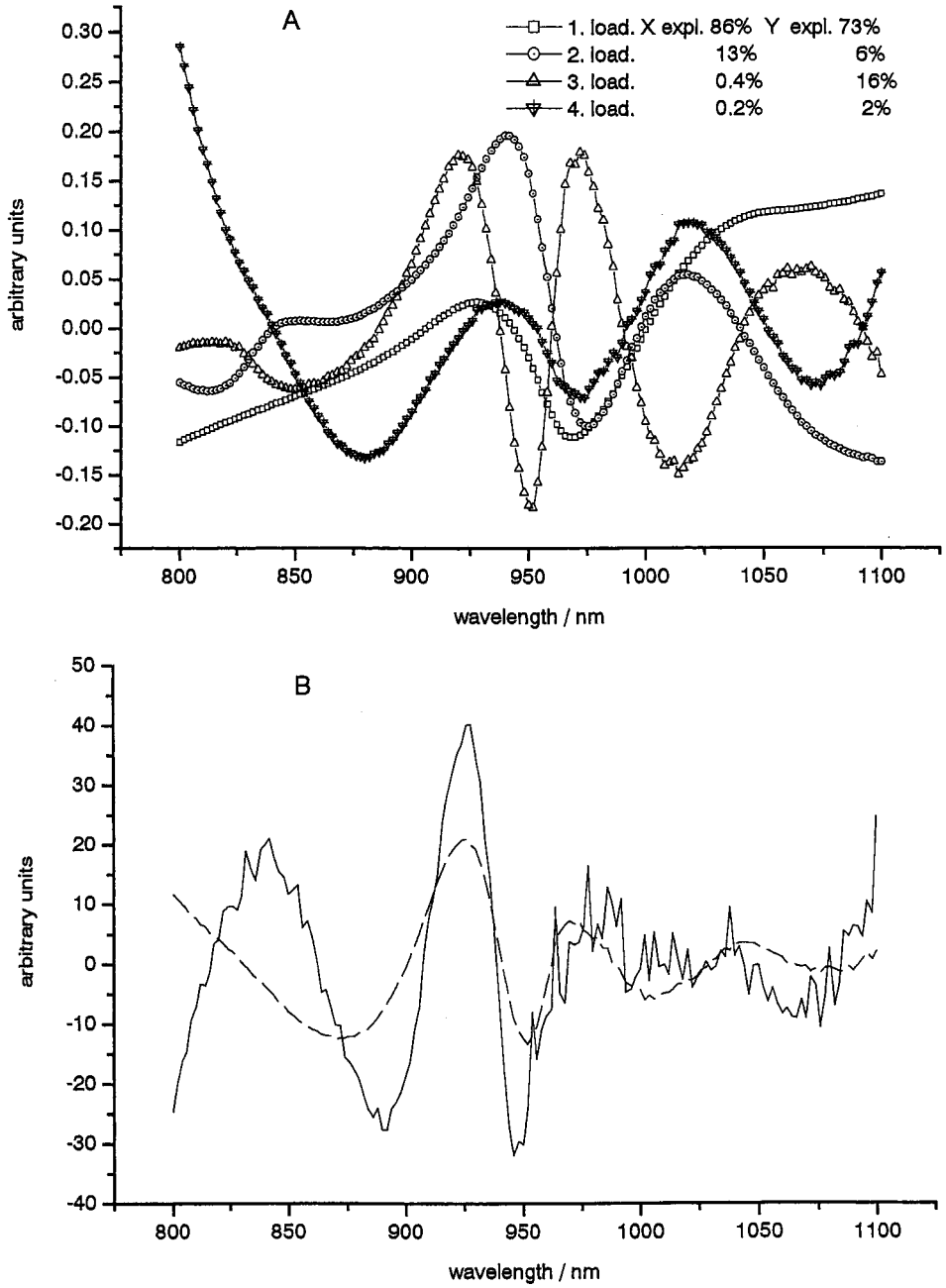

Figure 2. First four loading weights for fat calibration (A) and regression coefficients obtained after four (- - -) and six (—) factors (B).

840 nm are almost negligible for the spectral variances (Figure 2A) and are concerned with not more than 20% of the concentration variances. The regression coefficients obtained in the present study are comparable to those presented by Tsenkova et al.[18] Besides the bands reported in our study, they observed peaks at 968, 990, and 1026 nm. Their cross validation gave the standard error of calibration (SEC) of 0.167 wt % at the best, which is rather higher than the RMSEP presented here. The experimental setup in Chen et al.[19] was closer to the one used here, and comparison with their results is more suitable. The best standard error of prediction (SEP) of a completely independent set of spectra that they reported was 0.11 wt %, which is very close to the RMSEP obtained here. The only difference between our experimental setup and their setup is that we used the 800−1100-nm region while they employed the 700−1100-nm region.

In an NIR study of Atlantic salmon, Wold et al.[25] found that eight wavelengths are important for fat calibration; 890, 910, 930, 950, 970, 1010, 1020, and 1048 nm, although some of these wavelengths are ascribed to proteins. Nortvedt et al.[26] also used these wavelengths for fat prediction in the Atlantic halibut fillets. It was found that reduction from the whole spectra to the set of the above wavelengths did not improve the results.[26] Although the NIR investigations of fish components seem to be far from the studies of milk components, there is significant similarity in the spectra investigated and in the results of spectral analyses. For example, Nortvedt et al.[26] found that fat was modeled with greater success than protein and that the precision of prediction is better for cross validation than for independent test sets. The same findings are true for the milk study. The regression coefficients obtained in the present study confirm not only the

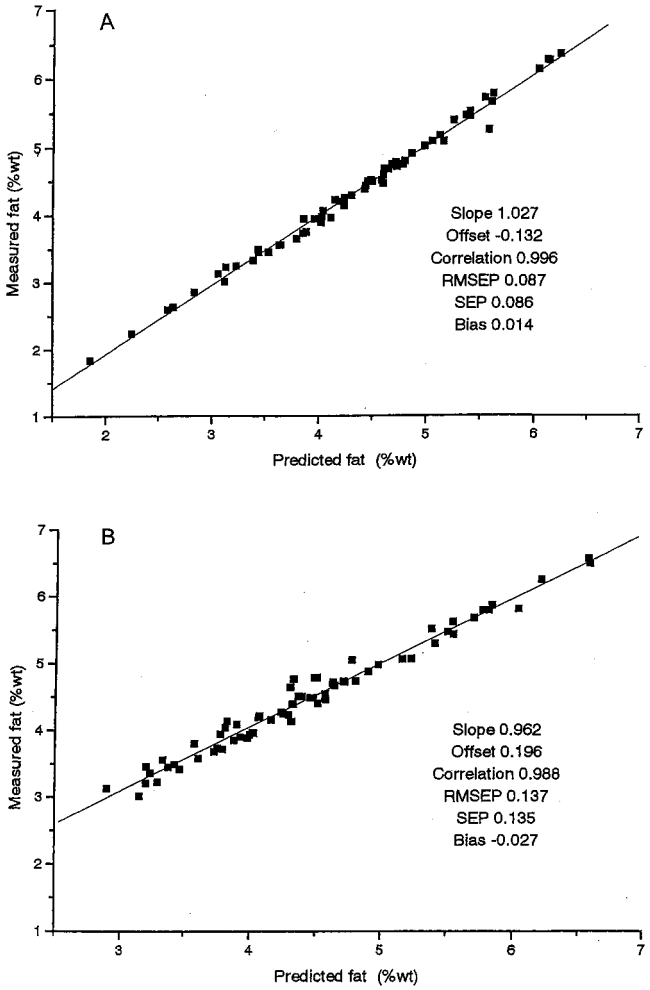

Figure 3. Correlation between measured and NIR predicted values for milk fat content for calibration models obtained after six (A) and four (B) factors.

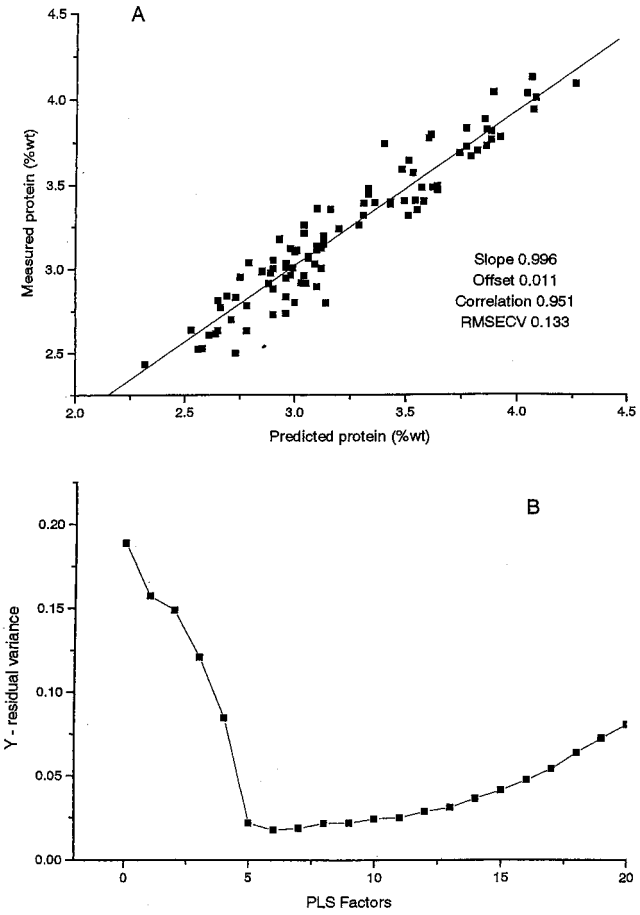

Figure 4. Cross validation plot for protein in milk (A) and protein concentration residual variance versus PLS factors (B).

wavelengths at 890, 930, and 950 nm but also that at 840 nm of particular importance for the fat determination. Our results also show that the wavelengths above 950 nm have small influence on the fat prediction.

For the sake of practical use of the short-wave NIR spectra for the determination of fat in milk, we carefully examined the changes in the baseline with the fat content. As described above, the four groups of milk samples were analyzed. In each of them, 20−25% of samples do not give a higher baseline with increasing fat content. When all the four groups are merged together to form the set investigated, the percentage becomes slightly higher, varying between 30 and 35%. Thus, the baseline level depends almost quantitatively on the fat content. We cannot reveal any reason such as the high concentration of proteins or lactose that could explain the deviation from the proportionality between the baseline level and fat content. Since the baseline itself contains some information about fat, MSC is highly recommended as the pretreatment in the present case. MSC still leaves some distances among the baselines of different samples (Figure 1B), but it allows the fat bands to have more important roles in the calibration. This is clear from comparison of the results of the calibration we obtained from the NIR spectra after MSC with those obtained from the spectra without MSC or the spectra pretreated by baseline correction. All the calibration and prediction parameters are better for the spectra after MSC than the spectra without MSC or the spectra pretreated by baseline removal. RMSEC and RMSECV are found to be 0.175 and 0.208 wt %, respectively, for the spectra without MSC and they are 0.122 and 0.207 wt %, respectively, for those with the baseline correction.

Protein Calibration. The prediction of protein concentration in milk is, in general, inferior in comparison with that of fat concentration. In the short-wave NIR region, only one correlation coefficient of 0.714 was reported for protein calibration in the raw milk[18] while the correlation coefficient of 0.94 was obtained for the NIR spectra of commercial milk in the 1100−2400-nm region.[15] Therefore, we attempted the protein calibration by cross validation of the whole set of 100 samples with 2 samples left out in each validation step. The correlation coefficient of 0.935 was obtained for the 800−1100-nm region, but we found that the result can be improved by narrowing the wavelength region into the 900−1100-nm region. In Figure 4A is shown the result of cross validation from the 900−1100-nm region by use of all the 100 samples. Note that the correlation coefficient is markedly improved compared with previous results based upon the 700−1100-nm region of NIR spectra of 258 milk samples.[18] RMSEC and RMSECV are also small although larger than those presented by Tsenkova et al.[18]

Figure 4B illustrates *Y* residual validation variance versus the number of PLS factors used. The plot in Figure 4B reveals that most of the *Y* variation is explained by lower principal components. The third and fourth principal components seem to carry the main

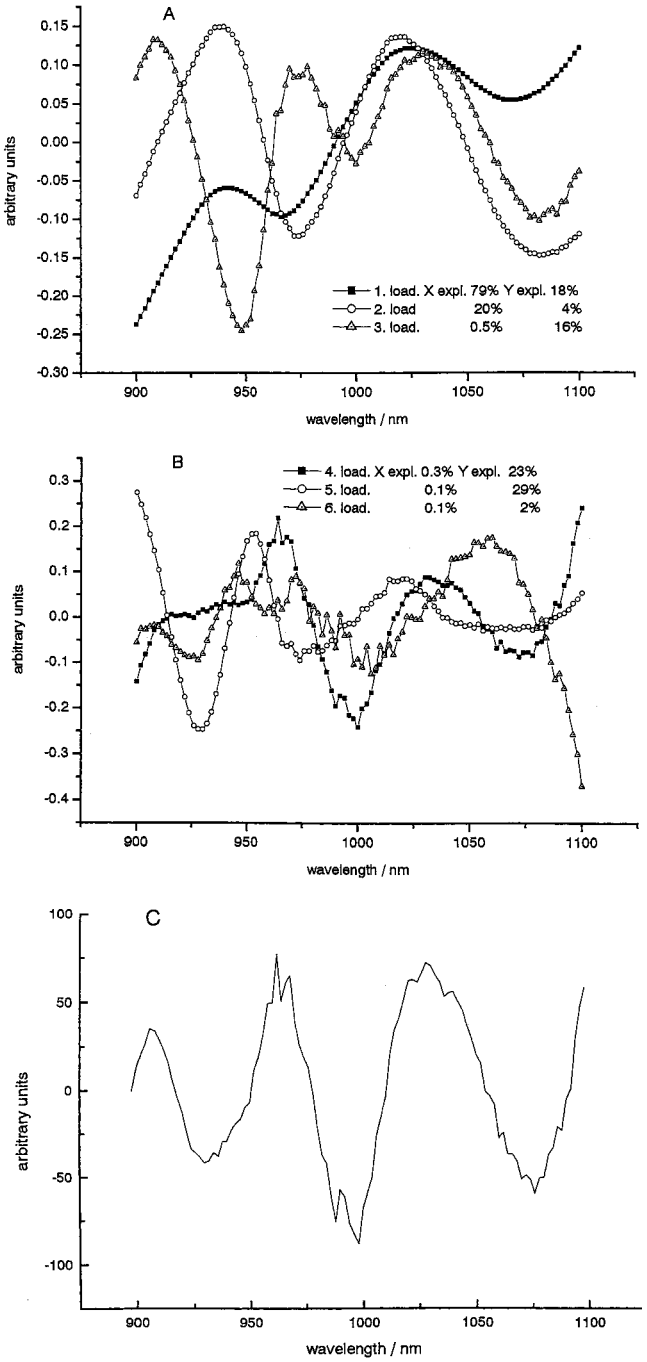

Figure 5. First three loading weights (A), 4−6 loading weights (B), and regression coefficient (C) for protein calibration. The protein statistics: min = 2.32 wt %; max = 4.26 wt %; mean = 3,.22 wt %; SD = 0,43.

information while the optimum number of components is six. The first two loading weights shown in Figure 5A reveal that almost all spectral variance is connected with only 22% of the protein content. The protein bands are strongly obscured and regression of the vector of protein concentration upon the milk spectra cannot emphasize them. The negligible part of spectral variation covers 70% of protein concentration data, mainly in the third, fourth, and fifth loadings (Figure 5A and B). Thus, intensive bands at 936 and 1018 nm in the first and second loadings are weakly related with proteins while the maximums at 906, 946, 972, and 1025 nm in the third, those at 946, 996, and 1032 nm in the fourth, and those at 926 and 950 nm in the fifth reflect the amount of proteins in raw milk. The regression coefficient shown in Figure 5C is similar to the fourth loading weights with a negative band at 1076 nm that appears in the first two loadings.

The band at 906 nm arises from a third overtone of C−H stretching modes of proteins, while that at 1030 nm is due to a second overtone of their N−H stretching vibrations.[21,24] The band at 966 nm probably reflects the interaction between proteins and water while for the band at 990 nm two possibilities exist; the first possibility is that it results from an O−H stretching vibration of water interacting with proteins while the second one is that some O−H bonds in proteins give rise to this band. The band at 930 nm was already assigned to fat. It is likely that the species whose concentration change leads the overall spectral variation affect the calibration of other species with minor spectral contributions. The attempt to assign protein bands can be valuable because there is little information about protein bands in the short-wave NIR spectra of biological materials.

The rather high correlation coefficients found for the whole data set suggested that the prediction of protein content in the samples included in the calibration model could yield satisfactorily results. Thus, we split the whole set into two groups of smaller sets and tried to predict the protein content for the independent sets**.** In the first group, the sets consisting of 40 and 60 samples while in the second group we prepared the sets with 30 and 70 samples. Then, cross validation and cross prediction for all the four sets were performed. Cross prediction means that protein contents in all four sets are predicted by different models. The results are summarized in Table 2. The main conclusion that can be extracted from Table 2 is that the prediction of the samples that were not included into the calibration gives promising results. RMSEP ranges from 0.121 to 0.180 wt %. All the correlation coefficients are higher than 0.920, and the best one is found to be 0.957. Thorough analysis of the data in Table 2 shows that generally the prediction of smaller independent sets by models developed from the larger sets is better than that of larger sets by models developed by smaller sets. It is noted that the prediction ability of the model based upon the whole data set decreases as we increase the number of samples whose protein content should be predicted. RMSEP is higher and the correlation parameters for the prediction of the protein content are poorer for the set that contains 70 samples by the model built from the whole data set than for the set containing 30 samples. If we predict the protein content in the whole data set by the models developed from the 30 and 70 spectra, respectively, the latter model gives considerably smaller errors. All these results put toward the fact that due to the weak spectral variation caused by proteins, the model with high prediction ability must account for a wide range of spectral variances that arise from other milk components, e.g., fat. Thus, only models that are built from a large number of samples can give regression coefficients that satisfactory predict the protein content in smaller, independent sets.

Lactose Calibration. The results of lactose calibration are worse than those obtained for fat and proteins. In Figure 6A, the leave-one-out cross validation plot for the whole set of 95 samples is shown—the validation with the five samples that contained less than 4 wt % of lactose was insufficiently bad, so that these samples

Table 2. Prediction Results Obtained from Several Calibration Models for Protein Content Determination in Raw Milk

predict. no.	no. of samples in calib set	no. of samples in predict. set	correlation in cross validation	factors	slope	offset	correlation in prediction	RMSEP	bias
1	40	60	0.960	6	0.859	0.432	0.920	0.177	0.04
2	60	40	0.928	5	1.080	−0.288	0.957	0.121	0.04
3	30	70	0.924	6	0.888	0.368	0.932	0.165	0.00
4	70	30	0.944	6	1.002	−0.087	0.931	0.147	0.08
5	60	30	0.928	5	1.056	−0.231	0.941	0.131	0.06
6	30	60	0.924	6	0.877	0.400	0.928	0.161	0.01

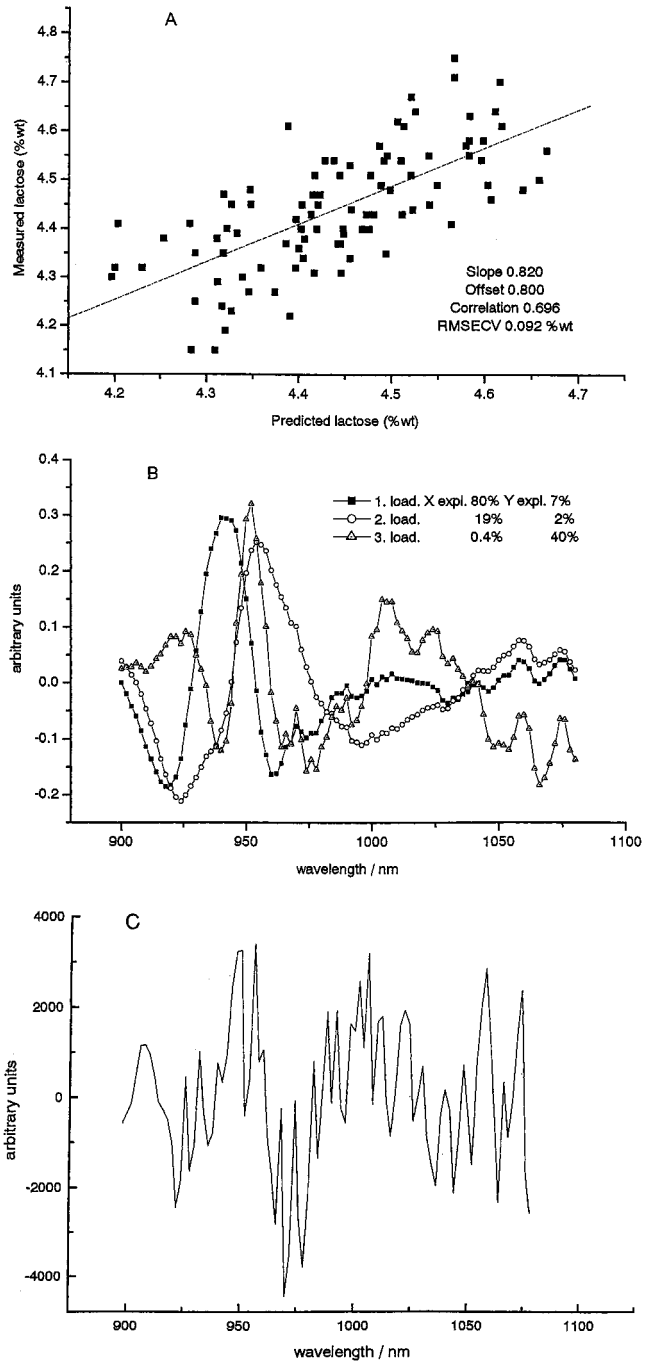

Figure 6. Cross validation plot for lactose in milk (A), the first three loading weights (B), and regression coefficient (C) for lactose calibration. The lactose statistics: min = 4.15 wt %; max = 4.75 wt %; mean = 4.44 wt %; SD = 0.13 wt %

were excluded from the analysis. RMSEC and RMSECV are found to be 0.064 and 0.092 wt %, respectively. The RMSEC is slightly better than that reported by Tsenkova et al.[18] while RMSECV is essentially the same as that in their study.

The loading weights shown in Figure 6B reveal that all the spectral variation in raw milk is weakly dependent on the lactose content while the third loading weights are found to carry most of the information about the lactose content. It can be seen from the third loading weights that wavelengths of 930, 948, 960, and 1025 nm might be of particular importance for the lactose calibration. However, the very noisy regression coefficient shown in Figure 6C calculated from seven factors gives significant weights only in the 950−1000-nm region.

Subsequently, we divided the whole data set of 95 samples into two almost equal sets, made two models from each of these sets, and tried to predict lactose content in each of them by another model. No satisfactorily results are obtained because RMSEP are found to be 0.106 and 0.112 wt %, respectively. The lactose content in the milk samples analyzed was almost constant, and a successful calibration model should recognize small changes in the lactose concentration**.** RMSEP obtained are rather high, and thus, one can conclude that the calibration models give just a rough estimation of the lactose content. The strong water band near 970 nm as well as the fat band at 930 nm located at almost the same wavelength as that for the lactose band[19] strongly obscures any spectral indication from lactose in raw milk, making the calibration of lactose very difficult. The low content of lactose in milk and the absence of characteristic C−H or O−H bands are the reasons for unsuccessful calibration as well.

Band Assignment in the Short-Wave NIR Region. The assignment of all the bands investigated is summarized in Table 3. Reliable band assignment in the short-wave NIR region has been proposed so far only for a few major fat bands. The assignment of bands due to protein has not been discussed on a solid base. This may be the first time that the assignment of the NIR bands of milk in the 800−1100-nm region has been investigated in some detail. Because of the increase in interest for using the short-wave NIR region for milk analysis, the assignment in Table 3 may be valuable.

CONCLUSION

The present study has provided insight into the short-wave NIR region, applied newly for the emerging field of biological and biomedical analysis. PLS calibration of fat, proteins, and lactose in milk using the short-wave NIR region (the 800−1100-nm region for fat and the 900−1100 nm region for proteins and lactose) gave reliable results for fat and proteins but not for lactose. It was found

Table 3. Proposed Assignment of Bands Observed in Loading Weights and Regression Coefficients of PLS Regression Analyses of Short-Wave NIR Spectra of Milk[a]

band position (nm)	species	assignment
840	f	combination mode of C−H stretching and C−H bending vibrations of fat?
880−890	f	third overtone of C−H stretching of fat?
906	p	third overtone of C−H stretching of proteins[21,7]
928	f	third overtone of C−H stretching of fat[21,24]
950−960	w	second overtone of O−H stretching of water interacting with protein (and fat)
968	w	second overtone of O−H stretching of water[21]
996	w	second overtone of O−H stretching of water interacting with protein or second overtone of O−H stretching vibration of proteins
1018	f	2 C−H stretching + 3 C−H deformation of fat
1020	p	2 N−H stretching + 2 amide I[21]
1030	p	second overtone of N−H stretching[21]
1042	f	2 C−H stretching + 2 C−H deformation of fat[21,24]

[a] f, fat; p, proteins; w, water.

that MSC is a useful pretreatment for raw milk spectra. PLS loading weights obtained by the fat calibration show that overall spectral variances in short-wave NIR spectra of milk are proportional to the fat content. The regression coefficients reveal that the band at 930 nm is the most important for fat calibration. The RMSEP of the fat content in the samples that are not included into calibration procedure was less than 0.100 wt %.

The correlation coefficients and RMSEP obtained in the present study for proteins are much better than those reported previously. The RMSEP of proteins in the independent sets varies from 0.121 to 0.180 wt %. The prediction of the protein content may be regarded as very well because only 20% of the protein concentration variation is accounted for by 99% of milk spectral variation. Due to the small contribution of protein to the milk spectra, only regression coefficients calculated from large calibration data sets can predict precisely the protein concentration in smaller, independent data sets. The wavelengths of 906, 926, 950, 964, and 1032 nm are found to play key roles in the protein calibration. The peaks at 906 and 1030 nm arise from the third overtone of a C−H stretching mode and the second overtone of an N−H stretching mode of proteins, respectively. The bands at 966 and 990 nm are supposed to result from water−protein interaction. The regression coefficients for the lactose calibration are noisy and do not provide satisfying prediction results. RMSEC and RMSECV for the lactose concentration in raw milk show that only a rough estimation of the lactose content can be achieved by PLS regression in the short-wave NIR region. On the basis of the PLS regression coefficients and loading weights, we have proposed band assignment in the short-wave NIR region which is very useful for biological, environmental, and material research and application.

ACKNOWLEDGMENT

The authors thank Dr. S. Tanabe, Dr. T. Hayashi, Dr. F. Terada, and Dr. M. Amari of the National Institute of Animal Industry, Japan, Dr. S. Kawano of the National Food Research Institute, Japan, Dr. R. Tsenkova of Kobe University, Japan, and Mr. K. Kiso, Kwansei-Gakuin University, Japan, for the preparation of milk samples, the determination of protein and fat contents in them, and the NIR measurements. This work was supported by the Program for Promotion of Basic Research Activities for Innovative Biosciences (PROBRAIN).

Received for review April 26, 2000. Accepted October 12, 2000.

AC000469C