Analytical Letters, 39: 2073–2083, 2006

ISSN 0003-2719 print/1532-236X online
DOI: 10.1080/00032710600724088

Taylor & Francis
Taylor & Francis Group

CHEMOMETRICS

A Partial Least Squares-Based Consensus Regression Method for the Analysis of Near-Infrared Complex Spectral Data of Plant Samples

Zhenqiang Su
Department of Chemistry, University of Science and Technology of China, Hefei, Anhui, P. R. China and Center for Toxico-informatics, National Center for Toxicological Research (NCTR), US Food and Drug Administration (FDA), Jefferson, AZ, USA

Weida Tong and Leming Shi
Center for Toxico-informatics, National Center for Toxicological Research (NCTR), US Food and Drug Administration (FDA), Jefferson, AZ, USA

Xueguang Shao and Wensheng Cai
Department of Chemistry, University of Science and Technology of China, Hefei, Anhui, P. R. China and Department of Chemistry, Nankai University, Tianjin, P. R. China

Abstract: A consensus regression approach based on partial least square (PLS) regression, named as cPLS, for calibrating the NIR data was investigated. In this approach, multiple independent PLS models were developed and integrated into a single consensus model. The utility and merits of the cPLS method were demonstrated by comparing its results with those from a regular PLS method in predicting moisture, oil, protein, and starch contents of corn samples using the NIR spectral data. It was

Received 21 October 2005; accepted 15 February 2006
The authors gratefully acknowledge the National Center for Toxicological Research (NCTR) of the U.S. Food and Drug Administration (FDA) and SAS Inc. for postdoctoral support through the Oak Ridge Institute for Science and Education.
Address correspondence to Wensheng Cai, Department of Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China. E-mail: wscai@ustc.edu.cn

found that cPLS was superior to regular PLS with respect to prediction accuracy and robustness.

Keywords: Near-infrared spectroscopy, partial least squares, consensus modeling, multivariate calibration

INTRODUCTION

Near-infrared spectroscopy (NIR) combined with chemometric methods; e.g., multivariate calibration techniques, has been widely used for the detection and identification of the composition of plant samples such as tobacco, tea, traditional Chinese medicine, and corn, etc. (Corti et al. 1993; Blanco and Romero 2001; Shao et al. 2004; Schulz et al. 2005). Such approaches are both faster and less expensive than conventional wet chemical methods in estimating the plant sample constituents.

An NIR spectrum consists of fundamental vibration bands of molecules that are usually derived from anharmonic *X*-H (mainly C-H, N-H, and O-H) stretching modes. Thus, NIR provides rich information about the structural and physical properties of the samples (Liu et al. 1994; Millar et al. 1996). However, most NIR spectra contain overlapping bands, which poses a challenge for extracting sample-specific peaks and further for developing robust models to predict unknown samples based on the spectra. Accordingly, many chemometric methods have been investigated for the analysis of NIR data, including signal preprocessing techniques such as multiplicative scatter correction (MSC) (Helland et al. 1995), orthogonal signal correction (OSC) (Sjoblom et al. 1998), wavelet transform (WT) (Chen et al. 2003; Chen et al. 2004), and data modeling methods such as partial least squares (PLS) (Inon et al. 2005) and soft independent modeling of class analogies (SIMCA) (Candolfi et al. 1999).

Chemometric methods have been effectively used to develop predictive models that relate spectral information with sample characteristics (Thomas and Haaland 1990). In these approaches, a model is first developed to correlate the peaks in the spectra with sample characteristics across known samples, and then the model can be used to predict unknown samples. Most of such regression techniques used for NIR spectra are based on a single model. Though these are good approaches, the single model tends to fit the calibration data to a single spectral pattern, which could result in the loss of some information in the richly complex spectra that may contain multiple superimposed patterns. Consensus modeling approaches that are able to extract separate spectral patterns could improve the fidelity of the correlation between spectral features and sample characteristics that would be otherwise lost by a single model's fit to a single pattern.

In this paper, a consensus regression approach named cPLS was presented. In cPLS, rather than selecting one PLS model on the basis of best fit, several

PLS models satisfying a predefined criterion were selected and combined into one. The effectiveness of cPLS was demonstrated by comparing the prediction results to those from the regular PLS in an application for calibration of the NIR spectra of corn samples. The results suggested that combining multiple individual PLS models by cPLS could improve not only the accuracy of prediction, but also the robustness of the model.

METHODOLOGY

Consensus Modeling for Regression–Theory

Consensus modeling combines the results of multiple individual models (called member models hereafter) to obtain a single prediction. The use of consensus modeling in many fields has increased significantly in the last few years (Wrabl and Shortle 1996, Hilser 2001, Prasad et al. 2003; Gramatica et al. 2004; Baurin et al. 2004; Svetnik et al. 2005), especially in the studies of quantitative structure-activity relationships (QSARs). Other examples are in simple averaging of individually trained neural networks (Perrone and Cooper 1993) and in the combination of hundreds of decision trees by boosting weak classifiers in random forest (Breiman 2000). The underlying assumption in consensus modeling is that multiple models will effectively identify and encode more aspects of the relationship between independent and dependent variables than will a single model.

Among two types of model development, regression and classification, research has shown that consensus classification models achieved better performance than single classifier due to better fidelity in extracting discriminating features within data as well as lower sensitivity to noise (Tong et al. 2003). For regression problems, the consensus model error $e(\bar{x})$can be represented by (Krogh and Vedelsby 1995):

$$e(\bar{x}) = \bar{e}(\bar{x}) - \bar{a}(\bar{x}) \tag{1}$$

where $\bar{e}(\bar{x})$ is the average error across all member models, while $\bar{a}(\bar{x})$ is the variance of the member models with respect to the results of the consensus model, and $\bar{a}(\bar{x})$ measures the disagreement among member models on input vector $\bar{x}$. These two terms are defined as:

$$\bar{e}(\bar{x}) = \frac{1}{N_m}\sum_{i=1}^{N_m}(y - f_i(\bar{x}))^2 \tag{2}$$

$$\bar{a}(\bar{x}) = \frac{1}{N_m}\sum_{i=1}^{N_m}(f_i(\bar{x}) - \bar{f}(\bar{x}))^2 \tag{3}$$

where N_m is the number of member models, $\bar{x}$ is the vector of the independent variables (i.e., a set of peaks in the NIR spectra in this study), y is the

dependent variable (i.e., the moisture, oil, protein, or starch content modeled in this study), $f_i(\bar{x})$ is the prediction result of the ith member model, while $\bar{f}(\bar{x})$ is the prediction of the consensus model, which can either be a linear combination of the prediction results of multiple member models: $\bar{f}(\bar{x}) = 1/N_m \sum_{i=1}^{N_m} f_i(\bar{x})$ (used in this study) or a weighted average:

$$\bar{f}(\bar{x}) = \sum_{i=1}^{N_m} w_i f_i(\bar{x}) \text{ with } \sum_{i=1}^{N_m} w_i = 1.$$

Clearly, the consensus model error $e(\bar{x})$ can be minimized in two ways, decreasing $\bar{e}(\bar{x})$ by enhancing the predictive quality of individual member models or increasing $\bar{a}(\bar{x})$ by diversifying the member models. Therefore, a robust consensus model should comprise multiple, high quality (low $\bar{e}(\bar{x})$), but mutually uncorrelated (high $\bar{a}(\bar{x})$) models.

cPLS Algorithm

Suppose that there are two matrices, $X(n, p)$ constituting p spectral signals of n samples and $Y(n, 1)$ presenting dependent variables for the n samples. The cPLS operates on these two matrices as depicted in Figure 1, which includes four steps:

1. Determine the number of the samples used to assess the quality of member models (N_t): N_t is a key parameter of cPLS. It controls the diversity of member models derived from $(n - N_t)$ samples. Although a large N_t will reduce the correlation between cPLS member models, a large N_t will also reduce the quality of the individual member model. A proper N_t must be determined prior to development of the consensus model. To determine N_t, n samples are randomly divided into a training set and an assessing set arbitrarily. In this study, the number of member models, N_m, is set to 100 (this is somewhat arbitrary). Then N_t is increased from 1 to 30 with a step size of one. The optimal N_t will be determined if the root mean squared error of prediction (*RMSEP*) begins to increase, as shown in Figure 2A. For each N_t, a cPLS model is developed (as in step 3) from the training set, and the model is then used to predict the assessing set and the *RMSEP* is calculated by:

$$RMSEP = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i^c)^2} \quad (4)$$

where m is the number of the samples in the assessing set, y_i is the experimental value of the ith sample in the assessing set, and $\hat{y}_i^c$ is the prediction of the cPLS model for the ith sample.

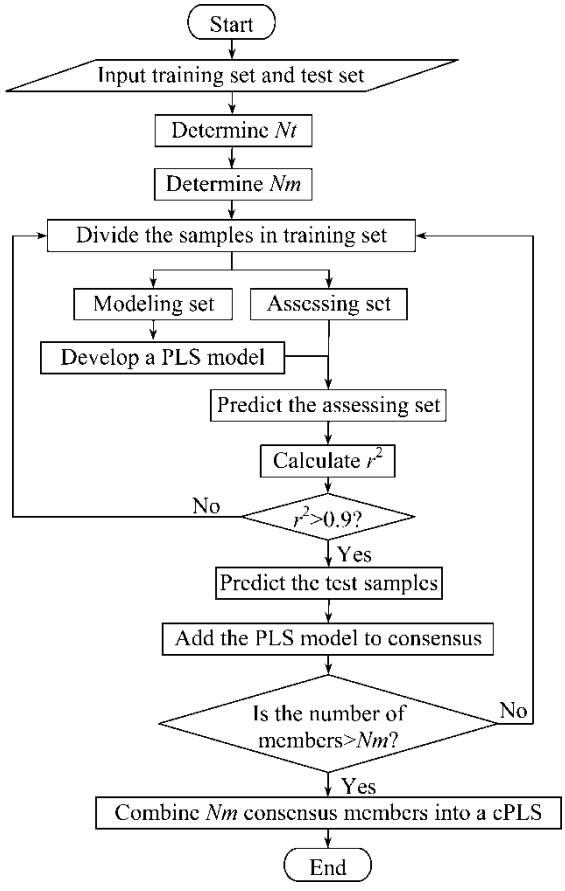

Figure 1. Flowchart of the cPLS algorithm.

2. Determine the number of member models (N_m): N_m affects the performance and stability of cPLS. For a given N_t, N_m can be determined as follows: the data set is randomly divided into a training set and an assessing set arbitrarily, then N_m increases from 1 to 500 with a step size of five. For each N_m, a cPLS model is developed from the training set, and the model is used to predict the assessing set with the *RMSEP* computed. N_m is determined if the *RMSEP* does not decrease, as shown in Figure 2B.
3. Develop N_m member models: n samples are randomly divided into two sets, with $(n - N_t)$ samples for model development and the remaining N_t samples for model assessment. A PLS model is then constructed on the training set, where the number of principal components used for the model is determined through an external validation. The number of

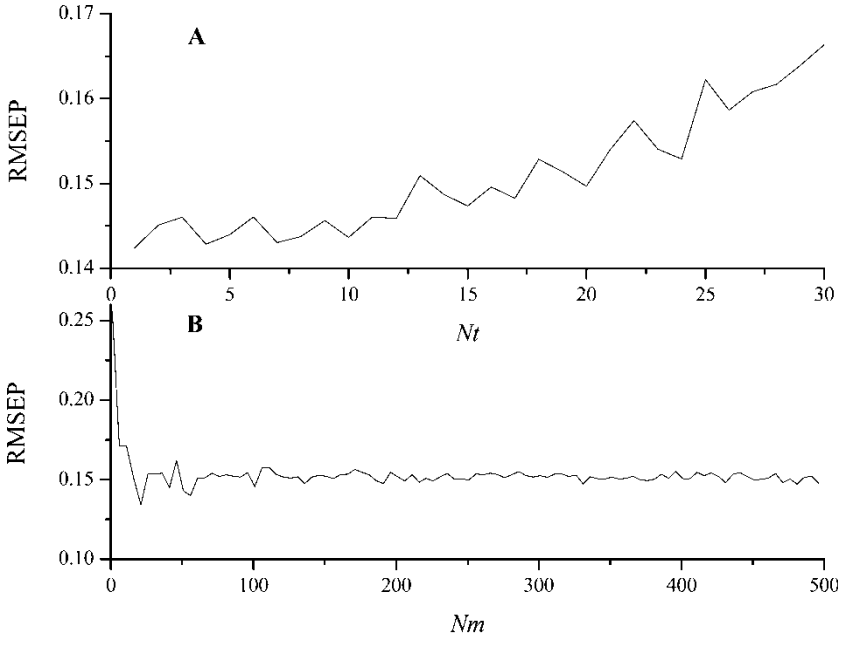

Figure 2. (**A**) *RMSEP* versus the number of the samples used in the test set (N_t) during the cPLS model development, and (**B**) *RMSEP* versus the number of member models (N_m) combined to form a cPLS model.

principal components yielding the lowest *RMSEP* is selected as optimal (as shown in Figure 3). The model is then used to predict the assessing set. The model will become a member of the cPLS if its performance in predicting the assessing set meets an acceptable criterion of $r^2 > 0.9$, otherwise it will be discarded. The r^2 describes how much variance is explained by the regression, which can be obtained by:

$$r^2 = 1 - \frac{\sum_{i=1}^{N_t}(y_i - \hat{y}_i^c)^2}{\sum_{i=1}^{N_t}(y_i - \bar{y})^2} \tag{5}$$

where y_i is the experimental value of the ith sample in the assessing set, $\hat{y}_i^c$ is the prediction for the ith sample, and $\bar{y}$ is the mean of the experimental data in the assessing set. The process is repeated until the number of the member models equals to N_m.

4. Combine N_m member models into one: the final cPLS model averages the prediction results of N_m member models to obtain a single prediction:

$$\hat{y}^c = \frac{1}{N_m}\sum_{i=1}^{N_m}\hat{y}_i \tag{6}$$

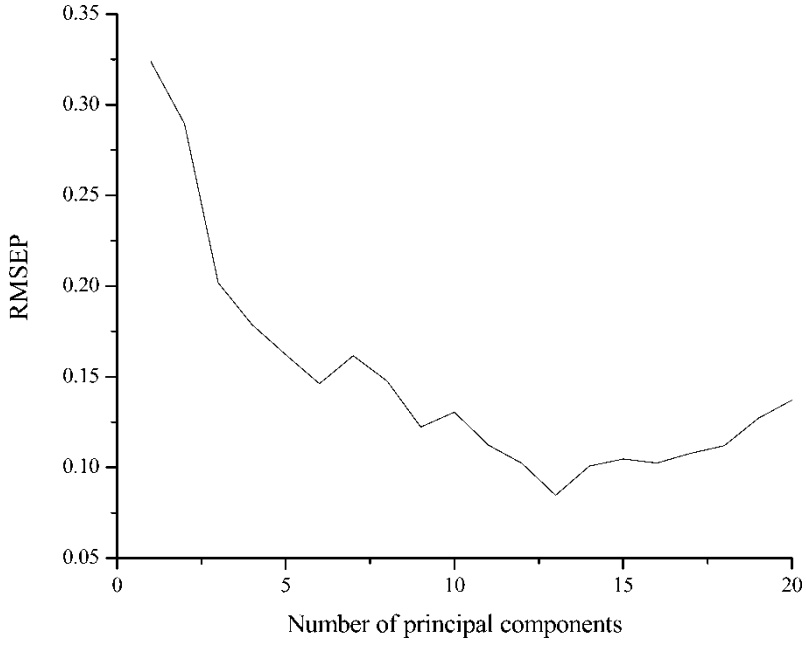

Figure 3. Relationship between *RMSEP* and the number of the principal components used to develop the member models of cPLS.

where $\hat{y}^c$ is the prediction of the cPLS, and $\hat{y}_i$ denotes the prediction of the ith member model.

Software

The cPLS algorithm was developed using the programming language MATLAB® 7.0, running on a personal computer equipped with two Intel Pentium 4 1.6 GHz processors and 2 GB of RAM. The program is available upon request.

RESULTS AND DISCUSSION

The cPLS algorithm was tested by comparing with the regular PLS regression method on a corn NIR data set. The data set, available from http://software.eigenvector.com/Data/Corn/corn.mat, has NIR spectra of 80 corn samples measured at Cargill Inc. (Minneapolis, MN, USA) using a spectrometer MP6. The values associated with the moisture, oil, protein, and starch content are the dependent variables. The independent variables are the wavelengths ranging from 1100–2498 nm at 2 nm intervals (700 variables). The

objective of the analysis is to predict the moisture, oil, protein, and starch content of the samples based on their NIR spectra.

To compare performance of regular PLS and cPLS for modeling the moisture, oil, protein, and starch content of the corn samples based on the NIR spectra, 80 samples were arbitrarily divided into two sets, with 60 samples used for model development and the remaining 20 used as an external validation set to challenge the model. A regular PLS model and a cPLS model were developed from the 60-samples training set, respectively. For the regular PLS, the number of principal components included in the PLS model was determined through the leave-one-out cross-validation based on the 60-samples training set. For the cPLS, the moisture content was used to determine the parameters N_t and N_m (Figures 2A and 2B). The optimal values for N_t and N_m are 10 and 100, respectively. Then 50 of 60 samples were randomly selected for member model development and the model was then assessed through predicting the remaining 10 samples (N_t is equal to 10). Thus, the number of samples used for development of member

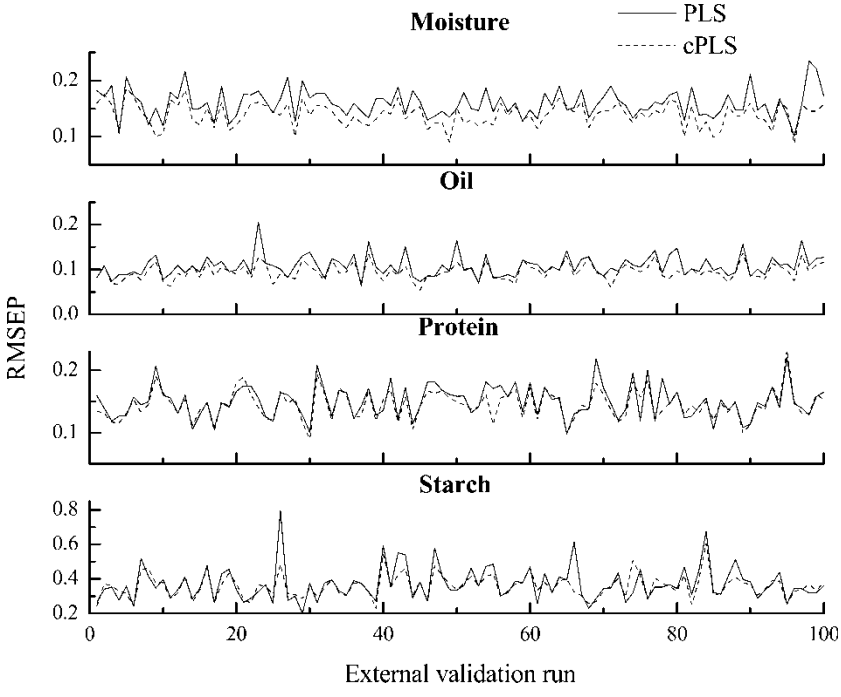

Figure 4. Comparison in prediction accuracy (*RMSEP*) between PLS and cPLS for the contents of moisture, oil, protein and starch. In this comparison, 80 corn samples were randomly divided into two sets, with 60 samples used for model development and 20 samples used as an external test set. The process was repeated 100 times and the prediction accuracy of the 100 runs was plotted for four endpoints.

Table 1. Comparison between cPLS and PLS in terms of *RMSEP* based on averaging of 100 runs of predicting the external test set

	Method	Moisture	Oil	Protein	Starch
Mean	PLS	0.159	0.107	0.150	0.370
	cPLS	0.139	0.0948	0.145	0.358
Standard deviation	PLS	0.025	0.024	0.026	0.098
	cPLS	0.021	0.018	0.024	0.068

models of cPLS (i.e., 50 samples) is actually less than that used for development of regular PLS (i.e., 60 samples). Both PLS and cPLS models were then evaluated by predicting the 20 samples not used for model calibration. The process was repeated 100 times and the prediction results are summarized in Figure 4, Table 1, and Table 2.

Figure 4 compares *RMSEP* between cPLS and regular PLS in predicting the external validation sets. For moisture and oil, it is clear that the majority of *RMSEP*s for cPLS is smaller than those for regular PLS, but for protein and starch, the *RMSEP*s for both models are almost identical. Table 1 lists the averages and standard deviations of *RMSEP*s for both cPLS and PLS over 100-times external validations on moisture, oil, protein, and starch. Both averages and standard deviations of *RMSEP*s for the cPLS are consistently smaller than those for regular PLS, indicating that the cPLS was more accurate and robust than regular PLS. To validate whether the calculated difference between cPLS and regular PLS is significant or not, a student *t*-test on the *RMSEP* results based on the 100-times external validations was conducted. Table 2 shows that the p values are extremely minute for moisture, oil, and protein, and that the largest p value for starch is also less than 0.05, the most commonly used level of statistical significance.

In this paper, the utility of the cPLS was demonstrated by comparing cPLS predictions of corn samples content based on NIR spectra with corresponding predictions from a regular, single PLS model. The results

Table 2. Statistical significance in a Student t-test by comparing the *RMSEP* values derived from cPLS with those from regular PLS in 100 runs of prediction

	p value
Moisture	3.6×10^{-21}
Oil	1.4×10^{-14}
Protein	0.00017
Starch	0.045

suggested that the modeling power of PLS was further enhanced by adopting the consensus approach described in this paper. Moreover, consensus modeling offered a generically effective means to obtain more accurate and robust regression models based on complex spectral data. This approach should be extensible to other fields when chemometric methods are used for predictive models.

REFERENCES

Baurin, N., Mozziconacci, J.C., Arnoult, E., Chavatte, P., Marot, C., and Morin-Allory, L. 2004. 2D QSAR consensus prediction for high-throughput virtual screening. An application to COX-2 inhibition modeling and screening of the NCI database. *J. Chem. Inf. Comput. Sci.*, 44 (1): 276–285.

Blanco, M. and Romero, M.A. 2001. Near-infrared libraries in the pharmaceutical industry: a solution for identity confirmation. *Analyst*, 126 (12): 2212–2217.

Breiman and Leo 2000. Randomizing outputs to increase prediction accuracy. Machine learning. *Machine Learning*, 40: 229–242.

Candolfi, A., De Maesschalck, R., Massart, D.L., Hailey, P.A., and Harrington, A.C. 1999. Identification of pharmaceutical excipients using NIR spectroscopy and SIMCA. *J. Pharm. Biomed. Anal.*, 19 (6): 923–935.

Chen, D., Hu, B., Shao, X., and Su, Q. 2004. Removal of major interference sources in aqueous near-infrared spectroscopy techniques. *Anal. Bioanal. Chem.*, 379 (1): 143–148.

Chen, D., Wang, F., Shao, X., and Su, Q. 2003. Elimination of interference information by a new hybrid algorithm for quantitative calibration of near infrared spectra. *Analyst*, 128 (9): 1200–1203.

Corti, P., Dreassi, E., and Leonardi, S. 1993. Near infrared reflectance analysis: features and applications in pharmaceutical and biomedical analysis. *Farmaco*, 48 (1): 3–20.

Gramatica, P., Pilutti, P., and Papa, E. 2004. Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling. *J. Chem. Inf. Comput. Sci.*, 44 (5): 1794–1802.

Helland, I.S., Naes, T., and Isaksson, T. 1995. Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. *Chemom. Intell. Lab. Syst.*, 29 (2): 233–241.

Hilser, V.J. 2001. Modeling the native state ensemble. *Methods Mol. Biol.*, 168: 93–116.

Inon, F.A., Llario, R., Garrigues, S., and de la Guardia, M. 2005. Development of a PLS based method for determination of the quality of beers by use of NIR: spectral ranges and sample-introduction considerations. *Anal. Bioanal. Chem.*, 382 (7): 1549–1561.

Krogh, A. and Vedelsby, J. 1995. *Neural Network Ensembles, Cross Validation, and Active Learning*. In *Advances in Neural Information Processing Systems 7*, Touretzky, D.S. and Leen, T.K. (eds.), MIT Press: Cambridge.

Liu, Y., Cho, R., Sakurai, K., Miura, T., and Ozaki, Y. 1994. Studies on spectra/structure correlations in near-infrared spectra of proteins and polypeptides. Part I. A marker for hydrogen bonds. *Appl. Spectrosc.*, 48: 1249–1254.

Millar, S., Robert, P., Devaux, M.F., Guy, R.C.E., and Maris, P. 1996. Near-infrared spectroscopic measurements of structural changes in starch-containing extruded products. *Appl. Spectrosc.*, 50: 1134–1139.

Perrone, M.P. and Cooper, L.N. 1993. *When Networks Disagree: Ensemble Methods for Hybrid Neural Networks*. In *Neural Networks for Speech and Image Processing*. and Mammone, R.J. (eds.), Chapman and Hall.

Prasad, J.C., Comeau, S.R., Vajda, S., and Camacho, C.J. 2003. Consensus alignment for reliable framework prediction in homology modeling. *Bioinformatics*, 19 (13): 1682–1691.

Schulz, H., Baranska, M., Quilitzsch, R., Schutze, W., and Losing, G. 2005. Characterization of peppercorn, pepper oil, and pepper oleoresin by vibrational spectroscopy methods. *J. Agric. Food Chem.*, 53 (9): 3358–3363.

Shao, X., Wang, F., Chen, D., and Su, Q. 2004. A method for near-infrared spectral calibration of complex plant samples with wavelet transform and elimination of uninformative variables. *Anal. Bioanal. Chem.*, 378 (5): 1382–1387.

Sjoblom, J., Svensson, O., Josefson, M., Kullberg, H., and Wold, S. 1998. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. *Chemom. Intell. Lab. Syst.*, 44 (1): 229–244.

Svetnik, V., Wang, T., Tong, C., Liaw, A., Sheridan, R.P., and Song, Q. 2005. Boosting: an ensemble learning tool for compound classification and QSAR modeling. *J. Chem. Inf. Model.*, 45 (3): 786–799.

Thomas, E.V. and Haaland, D.M. 1990. Comparison of multivariate calibration methods for quantitative spectral analysis. *Anal. Chem.*, 62: 1091–1099.

Tong, W., Hong, H., Fang, H., Xie, Q., and Perkins, R. 2003. Decision forest: combining the predictions of multiple independent decision tree models. *J. Chem. Inf. Comput. Sci.*, 43 (2): 525–31.

Wrabl, J.O. and Shortle, D. 1996. Perturbations of the denatured state ensemble: modeling their effects on protein stability and folding kinetics. *Protein Sci.*, 5 (11): 2343–2352.