Original article

Genetic multivariate calibration for near infrared spectroscopic determination of protein, moisture, dry mass, hardness and other residues of wheat

Durmuş Özdemir*

Department of Chemistry, Faculty of Science, Izmir Institute of Technology, Gülbahce, 35430 URLA/IZMIR, Turkey

(Received 1 November 2005; Accepted in revised form 28 February 2006)

Summary Determination of wheat flour quality parameters, such as protein, moisture, dry mass by wet chemistry analyses takes long time. Near infrared spectroscopy (NIR) coupled with multivariate calibration offers a fast and nondestructive alternative to obtain reliable results. However, due to the complexity of the spectra obtained from NIR, some wavelength selection is generally required to improve the predictive ability of multivariate calibration methods. In this study, two different wheat data sets are investigated with the aim of establishing successful calibration models using NIR spectra of wheat samples. The first data set (material 1) was obtained from the ftp address (ftp://ftp.clarkson.edu/pub/hopkepk/Chemdata/) and contained 100 NIR spectra of wheat of which wet chemical analysis of protein and moisture content were done with reference methods. The second data set (material 2) contained 176 spectra and was downloaded from http://www.spectroscopynow.com/Spy/basehtml/SpyH/1,1181,2-1-2-0-0-newsdetail-0-74,00.html. This wheat data set was given with the quality parameters, such as protein content, moisture content, other residues, dry mass, protein content in dry mass and hardness that were determined previously. Multivariate calibration models generated with genetic inverse least squares method demonstrated very good prediction results for the parameter mentioned here. Overall, the average per cent recoveries (APR) ranged between 99.23% and 100.34% with a standard deviation (SD) ranging from 0.34 to 3.15 for all the parameters investigated, except hardness. The APR value of hardness was 103.32 with the SD of 14.97.

Keywords Genetic algorithms, moisture, multivariate calibration, near infrared spectroscopy, protein, wheat.

Introduction

Total protein content of wheat flour is one of the most important factor in determining the quality and market value (Wesley *et al.*, 2001). However, there are other parameters such as moisture and hardness that also play a primary role in the end use of wheat. The conventional determinations of these parameters often requires large amount of samples and are generally time-consuming and expensive. The reference methods most commonly used in grain analysis are the Kjeldahl method and the combustion nitrogen analysis (CNA) method. A simple thermogravimetric analysis method is accepted as the official method of determining the moisture content of wheat. There are several United States Department of Agriculture approved protocols for moisture and protein content determination of wheat (http://www.usda.gov). Determining the hardness of wheat with a number of different methods was described in literature (Famera *et al.*, 2004).

Near infrared (NIR) spectroscopy (McClure, 1994) has become a popular method for simultaneous chemical analysis and is studied extensively in a number of different fields, such as process monitoring (DeThomas *et al.*, 1994), biotechnology (Arnold *et al.*, 2000) and the pharmaceutical industry (Tran *et al.*, 2004) because of the potential for on-line, nondestructive and noninvasive instrumentation. Traditionally, NIR spectroscopy has found its widest application area in agriculture and food industry (Puchwein & Eibelhuber, 1989; Sorvaniemi *et al.*, 1993; Hareland, 1994; Kalivas, 1997; Delwiche, 1998; McCaig, 2002; Miralbés, 2004; Ferrioa *et al.*, 2005). The NIR portion of the electromagnetic spectrum covers the range from 780 to 2500 nm and most of the absorption bands observed in this region are because of

*Correspondent: Fax: +90 232 750 7509;
e-mail: durmusozdemir@iyte.edu.tr

doi:10.1111/j.1365-2621.2006.01243.x

overtones and combinations of the fundamental mid-IR molecular vibrational bands. Although all the fundamental vibrational modes can have overtones, the most commonly observed bands arise from the C–H, O–H and N–H bonds in the molecules.

Advances in computers and automation technology have made today's instruments incredibly fast, so they can produce hundreds of spectra in a few minutes for a given sample that contains multiple components. Unfortunately, univariate calibration methods are not suitable for this type of data, as they require an interference-free system. Thanks to the chemometrics, multivariate calibration methods make it possible to relate instrument responses that consist of several predictor variables to a chemical or physical property of a sample. Several classical multivariate calibration methods have been developed (Lindberg *et al.*, 1983; Geladi & Kowalski, 1986; Haaland & Thomas, 1988; Wentzell *et al.*, 1997) in the last couple of decades for the analysis of complex chemical mixtures. The choice of the most suitable calibration method is very important in order to generate calibration models with high predictive ability for future samples. In some cases, conventional methods may not offer a satisfactory solution to a given problem due to the complexity of the data and it may be necessary to apply some sort of variable selection. There have been many mathematical methods of variable selection (Lindgren *et al.*, 1994; Centner *et al.*, 1996; Forina *et al.*, 1999), and genetic algorithm is one of them that offers a fast and effective solution for large-scale problems (Leardi *et al.*, 1992; Lucasius & Kateman, 1993; Hörchner & Kalivas, 1995).

Inverse least squares (ILS) is based on the inverse of Beer's Law where concentrations of an analyte are modelled as a function of absorbance measurements. Genetic inverse least squares (GILS) is the modified version of original ILS methods in which a small set of wavelengths are selected from a full spectral data matrix and evolved to an optimum solution using a genetic algorithm (GA) and has been applied to a number of wavelength selection problems (Özdemir & Dinç, 2004; Özdemir & Öztürk, 2004; Özdemir, 2005). GAs are nonlocal search and optimisation methods that are based upon the principles of natural selection (Hibbert, 1993; Paradkar & Williams, 1997; Pizarro *et al.*, 1998; Mosley & Williams, 1998; Özdemir & Williams, 1999).

In this work, GILS method, a multivariate calibration method based on a GA, was tested with the aim of establishing calibration models that have a high predictive ability for the NIR spectroscopic determination of several chemical and physical parameters of two wheat data sets. The first data set (material 1) was obtained from the ftp address ftp://ftp.clarkson.edu/pub/hopkepk/Chemdata/ and the second data set (material 2) was downloaded from http://www.spectroscopynow.com/Spy/basehtml/SpyH/1,1181,2-1-2-0-0-news_detail-074,00.html.

Genetic inverse least squares

The major drawback of the classical least squares (CLS) method is that all of the interfering species must be known and their concentrations included in the model. This need can be eliminated by using the ILS method, which uses the inverse of Beer's Law. In the ILS method, concentration of a component is modelled as a function of absorbance measurements. Because modern spectroscopic instruments are very stable and provide excellent signal-to-noise (S/N) ratios, it is believed that the majority of errors lie in the reference values of the calibration sample, and not in the measurement of their spectra. In fact, in many cases the reference data of the calibration set is generated from another analytical technique that already has its inherent errors, which might be higher than those of the spectrometer (e.g. Kjeldahl protein analysis used to calibrate NIR spectra).

The ILS model for m calibration samples with n wavelengths for each spectrum is described by:

$$\mathbf{C} = \mathbf{AP} + \mathbf{E_C} \tag{1}$$

where $\mathbf{C}$ is the $m \times l$ matrix of the component concentrations, $\mathbf{A}$ is the $m \times n$ matrix of the calibration spectra, $\mathbf{P}$ is the $n \times l$ matrix of the unknown calibration coefficients relating l component concentrations to the spectral intensities and $\mathbf{E_C}$ is the $m \times l$ matrix of errors in the concentrations not fit by the model. In the calibration step, ILS minimises the squared sum of the residuals in the concentrations. The biggest advantage of ILS is that eqn (1) can be reduced for the analysis of a single component at a time as analysis is based on an ILS model, which is invariant with respect to the number of chemical components included in the analysis. The reduced model is given as:

$$\mathbf{c} = \mathbf{Ap} + \mathbf{e_c} \tag{2}$$

where $\mathbf{c}$ is the $m \times 1$ vector of concentrations for the component that is being analysed, $\mathbf{p}$ is the $n \times 1$ vector of calibration coefficients and $\mathbf{e_c}$ is the $m \times 1$ vector of concentration residuals not fit by the model. During the calibration step, the least-squares estimate of $\mathbf{p}$ is:

$$\hat{\mathbf{p}} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}' \cdot \mathbf{c} \tag{3}$$

where $\hat{\mathbf{p}}$ is the estimated calibration coefficient. Once $\hat{\mathbf{p}}$ is calculated, the concentration of the analyte of interest can be predicted with the equation that follows:

$$\hat{c} = \mathbf{a}' \cdot \hat{\mathbf{p}} \tag{4}$$

where $\hat{c}$ is the scalar estimated concentration and $\mathbf{a}$ is the spectrum of the unknown sample. The ability to predict one component at a time without knowing the concentrations of interfering species has made ILS one of the most frequently used calibration methods.

The major disadvantage of ILS is that the number of wavelengths in the calibration spectra should not be more than the number of calibration samples. This is a big restriction as the number of wavelengths in a spectrum will generally be much more than the number of calibration samples and the selection of wavelengths that provide the best fit for the model is not a trivial process. Several wavelength selection strategies, such as stepwise wavelength selection and all possible combination searches, are available to build an ILS model, which fits the data best.

GA are global search and optimisation methods based upon the principles of natural evolution and selection as developed by Darwin. Computationally, the implementation of a typical GA is quite simple and consists of five basic steps including initialisation of a gene population, evaluation of the population, selection of the parent genes for breeding and mating, crossover and mutation and replacing parents with their offspring. These steps have taken their names from the biological foundation of the algorithm.

GILS is an implementation of a GA for selecting wavelengths to build multivariate calibration models with reduced data set. GILS follows the same basic initialise/breed/mutate/evaluate algorithm as other GAs to select a subset of wavelengths, but is unique in the way it encodes genes. A gene is a potential solution to a given problem and the exact form may vary from application to application. Here, in GILS method, the term 'gene' is used to describe the collection of instrumental responses, such as the absorbance values at certain wavelengths determined randomly at the wavelength range used to collect the spectrum (refer eqn 1). In other words, a gene is basically a subspectrum sample at a few wavelengths of the full spectrum. The term 'population' is used to describe the collection of individual genes in the current generation. The number of genes in a population must be an even number as it will be explained later.

In the initialisation step, the first generation of genes is created randomly with a fixed population size. Although random initialisation helps to minimise bias and maximise the number of possible recombinations, GILS is designed to select initial genes in a somewhat biased random fashion in order to start with genes better suited to the problem than those that would be randomly selected. Biasing is done with a correlation coefficient by plotting the predicted results of initial population against the actual component concentrations. The size of the gene pool is a user-defined even number in order to allow breeding of each gene in the population. It is important to note that the larger the population size, the longer the computation time. The number of instrumental responses in a gene is determined randomly between a fixed low limit and high limit. The lower limit was set to 2 in order to allow single-point crossover whereas the higher limit was set to eliminate overfitting problems and reduce the computation time. Once the initial gene population is created, the next step is to evaluate and rank the genes using a fitness function, which is the inverse of the standard error of calibration (SEC).

The third step is where the basic principle of natural evolution is put to work for GILS. This step involves the selection of the parent genes from the current population for breeding using a roulette wheel selection method according to their fitness values. The goal is to give a higher chance to those genes with high fitness so that only the best performing members of the population will survive in the long run, and will be able to pass their information to the next generations. Because of the random nature of the roulette wheel selection method, however, genes with low fitness values will also have some chance to be selected. Also, there will be genes that are selected multiple times and some genes will not be selected at all and will be thrown out of the gene pool. After the selection procedure is completed, the selected genes are allowed to mate top–down in pairs whereby the first gene mates with the second gene and the third one with the fourth one and so on as illustrated in the following example:

Parents

$$\mathrm{S}_1 = (A_{347}, A_{251}, \#A_{379}, A_{218}) \quad (5)$$

$$\mathrm{S}_2 = (A_{225}, A_{478}, \#A_{343}, A_{250}, A_{451}, A_{358}, A_{231}, A_{458}) \quad (6)$$

The points where the genes are cut for mating are indicated by #.

Offspring

$$\mathrm{S}_3 = (A_{347}, A_{251}, A_{343}, A_{250}, A_{451}, A_{358}, A_{231}, A_{458}) \quad (7)$$

$$\mathrm{S}_4 = (A_{379}, A_{218}, A_{225}, A_{478}) \quad (8)$$

where A_{347} represents the instrument response at the wavelength given in subscript, S_1 and S_2 represent the first and second parent genes and S_3 and S_4 are the corresponding genes for the offspring. Here, the first part of S_1 is combined with the second part of the S_2 to give the S_3; likewise, the second part of the S_1 is combined with the first part of the S_2 to give S_4. This process is called the single-point crossover and is common in GILS. Single-point crossover will not provide different offsprings if both parent genes are identical, which may happen in roulette wheel selection, when both genes are broken at the same point. Also, note that mating can increase or decrease the number of instrument responses in the offspring genes. After crossover, the parent genes are replaced by their offsprings and the offsprings are evaluated. The ranking process is based on their fitness values following the evaluation step. Then, the selection for breeding/mating starts all over again. This is repeated until a predefined number of iterations are reached.

Mutation which introduces random deviations into the population was also introduced into the GILS during the mating step at a rate of 1% as is typical in GA. This is usually done by replacing one of the responses in an existing gene with a randomly selected new one. Mutation allows the GR to explore the search space and incorporate new material into the genetic population. It helps keep the search moving and can eject GILS from a local minimum on the response surface. However, it is important not to set the mutation rate too high as it may keep the GA from being able to exploit the existing population. Also, the GILS method is an iterative algorithm and therefore, there is a high possibility that the method may easily overfit the calibration data so that the predictions for independent sets might be poor. To eliminate possible overfitting problems, cross validation is used in which one spectrum is left out of the calibration set and the model is constructed with $m-1$ sample. Then, this model is used to predict the concentration of the left out sample. This process is continued until all samples are left out at least once in each iteration. As long as the number of spectra in the calibration set is not too large, cross validation is an effective method of eliminating overfitting. If the number of calibration spectra is very large, then the GILS method has the option of half validation approach in which the half of the spectra in the calibration set is used to validate the model in each iteration.

In the end, the gene with the lowest SEC (highest fitness) is selected for the model building and this model is used to predict the concentrations of component analysed in the prediction (test) sets. The success of the model in the prediction of the test sets is evaluated using standard error of prediction (SEP). Because random processes are heavily involved in GILS as in the entire GA, the program has been set to run several times for each component in this study. The best run (i.e. the one generating the lowest SEC for the calibration set and at the same time producing SEP for prediction sets that are in the same range with the SEC) is subsequently selected for evaluation and further analysis. The termination of the algorithm can be done in many ways. The easiest way is to set a predefined iteration number for the number of breeding/mating cycles.

GILS has some major advantages over classical univariate and multivariate calibration methods. First of all, it is quite simple in terms of the mathematics involved in the model building and prediction steps, but at the same time it has the advantages of the multivariate calibration methods with a reduced data set as it

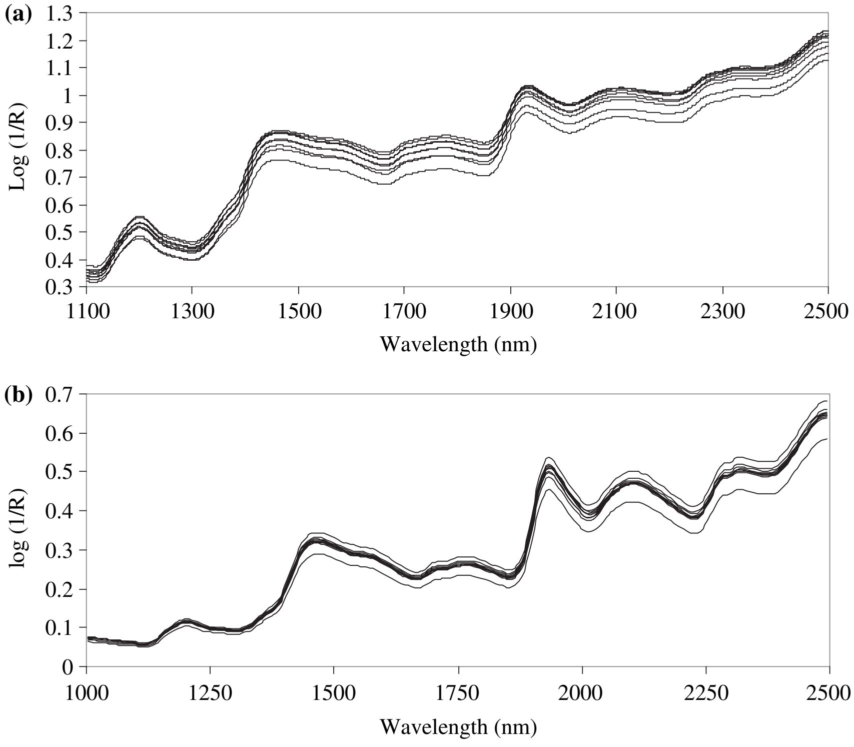

Figure 1 Near infrared (NIR) diffuse reflectance spectra of ten wheat samples: (a) from material 1 and (b) from material 2.

uses the full spectrum to extract genes. By selecting a subset of instrument responses, it is able to eliminate nonlinearities that might be present in the full spectral region.

Materials and methods

Two different NIR spectra of wheat data sets are investigated. Material 1 was obtained from the ftp address (ftp://ftp.clarkson.edu/pub/hopkepk/Chemdata/) and contained 100 NIR spectra of wheat of which wet chemical analysis of protein and moisture content were done with reference methods. The NIR spectra were recorded in diffuse reflectance mode as log (1/*R*) from 1100 to 2500 nm at 2-nm intervals and eighty-seven of them were used as described in Kalivas (1997). The composition of calibration and prediction sets were also kept the same as given in Kalivas (1997) with fifty

Table 1 Reference values of protein and moisture determined previously in prediction sets 1 and 2 of material 1

	Prediction set 1		Prediction set 2	
Sample number	Protein as is (w/w %)	Moisture as is (w/w %)	Protein as is (w/w %)	Moisture as is (w/w %)
1	11.36	12.70	13.01	12.96
2	10.40	13.13	12.58	13.14
3	12.23	13.11	11.93	12.80
4	11.65	12.79	10.78	12.99
5	11.81	13.16	11.06	13.16
6	12.48	13.46	10.36	13.57
7	13.72	13.34	11.78	13.14
8	13.08	12.94	14.02	13.14
9	11.72	16.68	11.09	16.94
10	11.23	16.50	11.75	16.33
11	11.08	16.19	11.13	16.12
12	10.26	15.99	10.86	16.31
13	11.42	15.88	11.24	16.31
14	11.27	15.51	11.48	15.80
15	10.90	15.59	11.85	15.36
16	10.98	15.18	11.07	15.34
17	11.98	15.11	11.05	15.69
18	13.15	15.62	10.74	15.65
19	11.77	15.78	10.90	15.56
20	11.87	15.68	11.64	15.10

Table 2 Predicted protein and moisture content in the prediction sets 1 and 2 of material 1 along with standard error of calibration (SEC), standard error of prediction (SEP), average percent recoveries (APR) and standard deviations (SD)

	Prediction set 1		Prediction set 2	
Sample number	Protein as is (w/w %)	Moisture as is (w/w %)	Protein as is (w/w %)	Moisture as is (w/w %)
1	11.77	12.83	12.24	13.30
2	9.55	13.16	12.21	13.36
3	12.35	12.72	12.53	12.69
4	11.46	13.04	11.30	12.72
5	11.88	13.13	10.54	12.98
6	12.83	13.31	10.43	13.38
7	13.88	13.24	11.40	12.96
8	13.05	12.98	13.76	13.20
9	11.86	16.84	11.33	16.95
10	11.72	16.47	11.89	16.22
11	11.25	16.05	11.32	15.48
12	10.58	15.63	11.13	16.21
13	11.23	16.23	11.12	16.16
14	11.08	15.80	11.06	16.13
15	11.25	15.69	11.25	15.47
16	10.84	15.31	10.96	15.75
17	12.22	15.14	11.15	15.72
18	12.41	15.24	10.74	15.62
19	11.84	15.96	10.76	15.95
20	12.12	15.58	11.23	15.15
SEC	0.08	0.12	0.08	0.12
SEP	0.34	0.21	0.37	0.25
APR	100.34	99.99	99.23	99.99
SD	3.09	1.45	3.15	1.72

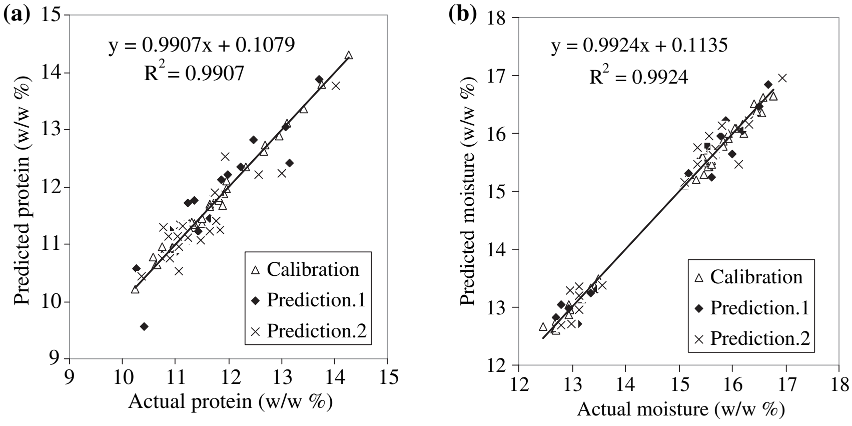

Figure 2 Actual versus genetic inverse least squares (GILS)-predicted: (a) moisture and (b) protein content of wheat samples in material 1.

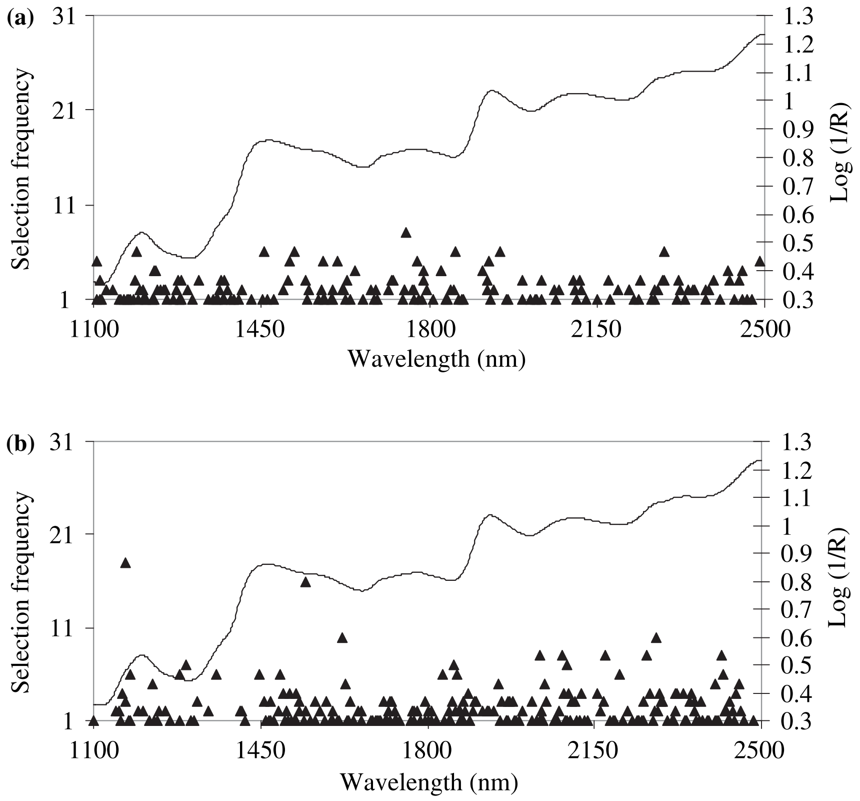

Figure 3 Frequency distribution of genetic inverse least squares (GILS)-selected wavelengths on the near infrared (NIR) diffuse reflectance spectra of material 1: (a) protein and (b) moisture.

Table 3 Reference values of protein, moisture and others (after subtracting the amount of protein and moisture from the raw sample), dry mass of samples (after removing the moisture from sample), protein in dry mass and hardness of wheat determined previously in prediction set of material 2

Sample number	Protein as is (w/w %)	Moisture as is (w/w %)	Others as is (w/w %)	Dry mass (w/w %)	Protein in dry mass (w/w %)	Hardness
1	9.83	11.10	77.58	88.90	8.74	56.6
2	10.94	12.76	74.93	87.24	9.55	77.1
3	11.07	13.67	73.75	86.33	9.56	59.7
4	11.27	13.17	74.17	86.83	9.79	54.6
5	11.60	14.61	72.30	85.39	9.91	51.1
6	11.89	10.77	75.85	89.23	10.61	45.5
7	12.19	11.53	74.61	88.47	10.78	59.5
8	12.55	14.26	71.72	85.74	10.76	50.1
9	12.95	13.45	72.25	86.55	11.21	59.4
10	13.42	10.95	73.96	89.05	11.95	83.8
11	13.66	11.09	73.36	88.91	12.14	67.3
12	13.78	13.52	71.41	86.48	11.92	65.1
13	13.95	13.58	70.84	86.42	12.06	79.7
14	14.02	10.64	73.73	89.36	12.53	30.1
15	14.36	11.15	72.68	88.85	12.76	72.7
16	14.45	10.73	73.15	89.27	12.90	95.3
17	15.81	10.59	73.09	89.41	13.95	24.7
18	14.68	13.56	70.14	86.44	12.69	80.2
19	14.84	11.28	72.11	88.72	13.16	39.5
20	14.99	11.25	72.04	88.75	13.30	22.0
21	15.16	12.87	70.39	87.13	13.21	30.9
22	15.22	12.79	70.41	87.21	13.28	23.2
23	15.40	10.14	72.59	89.86	13.84	28.9
24	15.58	13.62	69.20	86.38	13.46	24.9
25	15.87	10.84	71.55	89.16	14.15	48.2
26	15.95	10.82	71.66	89.18	14.22	88.2

samples in calibration set and twenty samples in each of the two prediction sets in order to be able to compare the results in literature.

Material 2 was downloaded from http://www.spectroscopynow.com/Spy/basehtml/SpyH/1,1181,2-1-2-0-0-newsdetail-0-74,00.html and contained 176 NIR diffuse reflectance spectra in the wavelength rage from 1000 to 2500 nm at 10-nm intervals. This wheat data set was given with the quality parameters like protein content, moisture content, other residues, dry mass, protein content in dry mass and hardness that were determined previously. Here, the data set was organised in a way, so that the first 150 samples assigned into calibration set and the remaining twenty-six samples were reserved for prediction set as the data given at the previous website. These samples were assigned into calibration and prediction sets with the constraints that the values of all parameters for prediction set embedded in the calibration range.

The GILS method was written in MATLAB programming language using Matlab 5.3 (MathWorks Inc, Natick, MA).

Results and discussion

The wheat data sets used in this study were selected to demonstrate the applicability of NIR spectroscopy coupled with genetic multivariate calibration for the determination of several physical and chemical quality parameters using NIR diffuse reflectance spectra of wheat samples. Figs 1a and b show the ten diffuse reflectance spectra as log $(1/R)$ between 1000 and 2500 nm wavelength range for the materials 1 and 2, respectively. Because of structural similarities, the spectral features of these wheat samples are very much alike and only minute differences exist in some parts of the whole spectrum. Throughout the multivariate calibration process, it is expected that these differences will reveal the information necessary to build successful calibration models otherwise almost impossible with univariate calibration methods.

The GILS method was first applied to the material 1 and calibration models for moisture and protein content were prepared. The calibration models were prepared with fifty spectra and then these models were tested with

Table 4 Predicted properties along with standard error of calibration (SEC), standard error of prediction (SEP), average percent recoveries (APR) and standard deviations (SD) in the prediction set for the six parameters investigated in material 2

Sample number	Predicted protein as is (w/w %)	Predicted moisture as is (w/w %)	Predicted others as is (w/w %)	Predicted dry mass (w/w %)	Predicted protein in dry mass (w/w %)	Predicted hardness
1	9.78	11.22	77.68	88.39	8.78	53.84
2	10.32	12.66	75.37	87.56	9.22	79.08
3	11.24	13.82	73.59	86.36	9.88	64.24
4	10.96	13.00	74.56	86.78	9.67	45.35
5	11.70	14.33	72.18	85.56	10.04	52.16
6	11.78	10.86	75.90	88.93	10.27	56.08
7	11.91	11.64	74.91	88.54	10.59	52.21
8	12.74	14.61	71.44	85.73	10.92	57.52
9	12.98	13.36	72.18	86.39	11.17	68.24
10	13.56	11.75	72.39	88.34	12.14	78.40
11	13.77	10.91	73.27	89.02	12.22	71.68
12	13.45	13.75	71.11	86.34	11.79	56.37
13	14.40	13.73	70.75	86.51	12.28	70.94
14	14.13	11.27	73.03	88.90	12.66	21.00
15	14.18	11.19	72.88	88.94	12.71	75.20
16	14.47	10.87	73.65	89.01	12.41	89.95
17	15.87	11.27	71.68	88.83	13.98	24.93
18	14.86	12.82	70.76	86.66	13.11	79.99
19	14.79	11.33	72.70	89.05	12.81	27.80
20	14.81	11.33	72.36	88.93	13.11	20.13
21	15.25	12.70	70.66	86.73	13.07	31.06
22	15.27	12.72	70.04	87.26	13.31	26.61
23	15.84	10.13	72.20	89.99	14.16	31.47
24	15.48	13.07	70.11	86.86	13.35	25.69
25	15.63	10.95	71.86	89.04	13.88	47.41
26	15.68	10.48	72.24	89.28	14.12	87.44
SEC	0.27	0.17	0.43	0.25	0.29	4.24
SEP	0.34	0.34	0.57	0.30	0.29	5.87
APR	100.25	99.63	100.00	100.06	100.27	103.32
SD	1.92	2.82	0.80	0.34	1.96	14.97

twenty independent prediction spectra with two separate prediction sets as shown in Table 1 (prediction set 1 and prediction set 2), which are not used in the calibration step. Because of the random nature of the GILS method, the program was set to run thirty times with twenty genes and fifty iterations. As the GILS program is iterated fifty times in each run, full cross validation is applied during the model building step to avoid possible overfitting problems. The model with the lowest SEC and at the same time, generating an SEP value in agreement with the SEC is chosen as the best model. Table 2 shows the predicted moisture and protein content of prediction sets together withSEC and SEP results for calibration and prediction sets, respectively. The average percent recoveries (APR) along with the standard deviations (SD) of APR were also given for both prediction sets. As can be seen from the Table 2, the best SEC and SEP values were ranged between 0.08% and 0.37% by mass for both moisture and protein content. These values are very similar with the values reported in the literature using the same data set obtained by partial least squares (PLS) method (Kalivas, 1997). APR and associated SD values for the prediction sets are also given in the last two rows of Table 2. As can be seen, both moisture and protein determinations were resulted with SD values ranging from 1.45% to 3.15%, indicating that the GILS method was able to generate successful calibration models. The plot of actual versus GILS predicted concentrations for both moisture and protein are illustrated in Fig. 2. While the R^2 values were ranged between 0.991 and 0.992 for the calibration models of both properties, predictions were slightly spread out. However, good correlations between predicted and actual values were observed.

Because GILS is a wavelength selection-based method, it is interesting to observe the distribution of selected wavelengths in multiple runs over the entire full spectral region. Figure 3 illustrates the frequency distribution of selected wavelengths in thirty runs for both moisture and protein content. Although there is not a very strong dominance of any particular wavelength range over the entire full spectral region, there are some distinct regions indicating a higher selection frequency as seen in the figure and in each run, GILS method was able to generate successful calibration models.

Material 2 contained 176 diffuse reflectance spectra and were split into calibration and prediction sets with the first 150 of them as calibration set and the remaining 26 of them as the prediction set. This second data set was given with the quality parameters like protein

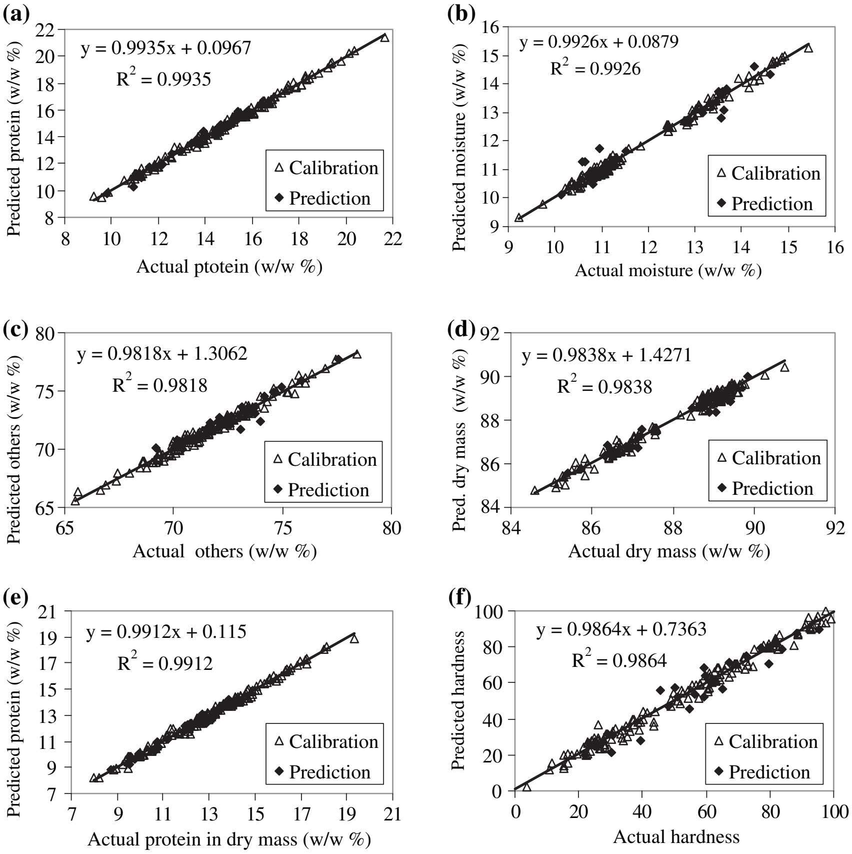

Figure 4 Actual versus genetic inverse least squares (GILS)-predicted values of parameters investigated by the GILS method in material 2: (a) protein, (b) moisture, (c) other residues in wheat, (d) dry mass, (e) protein in dry mass and (f) hardness.

content, moisture content, other residues, dry mass, protein content in dry mass and hardness that were determined previously. Table 3 shows the reference values of these six properties in the prediction set with twenty-six samples. For example, the protein content in raw sample ranges between 9.83% and 15.95% by mass and the moisture values ranged from 11.10% to 14.61% by mass. Several calibration models were generated for each property using GILS method in the same way outlined here with one difference. The number of calibration samples was 150 and this number was too large to apply full cross validation method to avoid possible overfitting of the models. To eliminate the problem, half validation approach was used in the GILS method in which all the odd numbered samples in the original calibration set were selected for model building step and the even numbered samples were reserved for model validation in each iteration. This approach not only eliminates the overfitting, but also significantly reduces the iteration time.

Table 4 shows the GILS-predicted values of the six properties studied in the second data set along with the SEC, SEP, APR and SD of APR. Very similar SEC and SEP values for all the properties except hardness were obtained ranging from 0.17% to 0.57% by mass indicating good fit for the models generated. The SEC and SEP values for hardness were 4.25% and 5.87% (arbitrary), respectively. The SD of APR were ranged between 0.34 and 2.82 for all the properties except hardness. The SD obtained for hardness was 14.97 indicating somewhat poor prediction. The plot of actual versus GILS-predicted concentrations for all properties are illustrated in Fig. 4. The R^2 values were ranged between 0.982 and 0.993 for the calibration models of all properties. As can be seen from Fig. 4, very good predictions were observed for the prediction set with the exception of a few samples.

The frequency distributions of selected wavelengths in thirty runs for all the properties are illustrated in Fig. 5. Once again, the distribution of selected wavelengths does not seem to indicate a strong localization of the algorithm, but regions around 1300 and 1900 nm show higher selections than the rest of the spectrum. This could be considered as an indication that the genetic algorithm incorporated into the GILS method focus the regions, where the most concentration-related information is contained. As a result, it

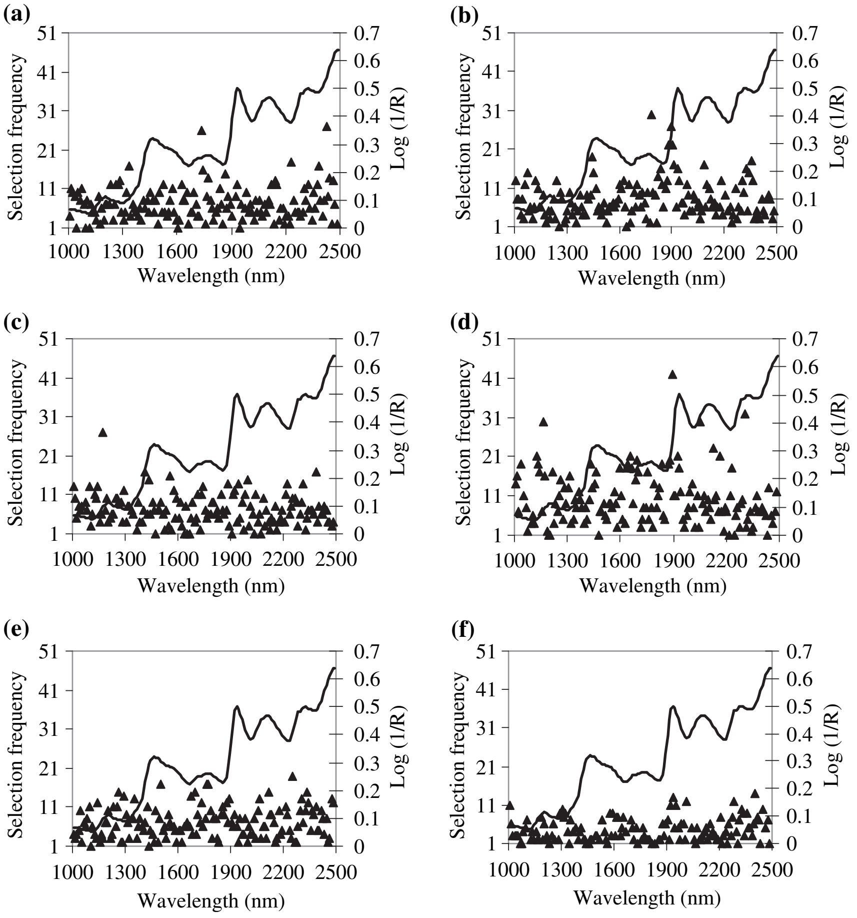

Figure 5 Frequency distribution of genetic inverse least squares (GILS)-selected wavelengths on the near infrared (NIR) diffuse reflectance spectra of material 2 for the parameters: (a) protein, (b) moisture, (c) other residues in wheat, (d) dry mass, (e) protein in dry mass and (f) hardness.

can be said that the GILS method can be used for fast and simultaneous determination of several chemical and physical properties of wheat samples using their NIR spectra.

Conclusions

This study has demonstrated the application of NIR spectroscopy with genetic multivariate calibration to simultaneous determination of several properties of wheat samples. The fact that the SEP values are below 0.50% by mass for moisture, protein content, dry mass and other residues, show that NIR spectroscopy can be used for simultaneous determination of chemical and physical properties of wheat. On the other hand, the GA used in the GILS method is able to select and extract the most relevant information to build successful calibration models that have high predictive ability for the independent test samples.

Acknowledgment

I would like to thank and acknowledge Prof. Dr Philip K. Hopke and Prof. Dr Harald Martens for the wheat data sets used in this study.

References

Arnold, S.A., Crowley, J., Vaidyanathan, S. et al. (2000). At-line monitoring of a submerged filamentous bacterial cultivation using near infrared spectroscopy. *Enzyme and Microbial Technology*, **27**, 691–697.

Centner, V., Massart, D.L., De Noord, O.E., De Jong, S., Vandeginste, B.M. & Sterna, C. (1996). Elimination of uninformative variables for multivariate calibration. *Analytical Chemistry*, **68**, 3851–3858.

Delwiche, S.R. (1998). Protein content of single kernels of wheat by near-infrared reflectance spectroscopy. *Journal of Cereal Science*, **27**, 241–254.

DeThomas, F.A., Hall, J.W. & Monfre, S.L. (1994). Real-time monitoring of polyurethane production using near infrared spectroscopy. *Talanta*, **41**, 425–431.

Famera, O., Hruskova, M. & Novotna, D. (2004). Evaluation of methods for wheat grain hardness determination. *Plant Soil and Environment*, **50**, 489–493.

Ferrioa, J.P., Villegasb, D., Zarcob, J., Apariciob, N., Arausc, J.L. & Royob, C. (2005). Assessment of durum wheat yield using visible and near-infrared reflectance spectra of canopies. *Field Crops Research*, **94**, 126–148.

Forina, M., Casolino, C. & Pizarro, M.C. (1999). Iterative predictor weighting (IPW) PLS: a technique for the elimination of useless predictors in regression problems. *Journal of Chemometrics*, **13**, 165–184.

Geladi, P. & Kowalski, B.R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, **185**, 1–17.

Haaland, D.M. & Thomas, E.V. (1988). Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry*, **60**, 1193–1202.

Hareland, G.A. (1994). Evaluation of flour particle size distribution by laser diffraction, sieve analysis and near-infrared reflectance spectroscopy. *Journal of Cereal Science*, **20**, 183–190.

Hibbert, D.B. (1993). Genetic algorithms in chemistry. *Chemistry of Intelligent Laboratory Systems*, **19**, 277–293.

Hörchner, U. & Kalivas, J.H. (1995). Further investigation on a comparative study of simulated annealing and genetic algorithm for wavelength selection. *Analytica Chimica Acta*, **311**, 1–13.

Kalivas, J.H. (1997). Two data sets of near infrared spectra. *Chemistry of Intelligent Laboratory Systems*, **37**, 255–259.

Leardi, R., Boggia, R. & Terrile, M. (1992). Genetic algorithms as a strategy for feature selection. *Journal of Chemometrics*, **6**, 267–281.

Lindberg, W., Persson, J.A. & Wold, S. (1983). Partial least-squares method for spectrofluorimetric analysis of mixtures of humic acid and lignin sulfonate. *Analytical Chemistry*, **55**, 643–648.

Lindgren, F., Geladi, P., Rännar, S. & Wold, S. (1994). Interactive variable selection (IVS) for PLS. Part 1: Theory and algorithms. *Journal of Chemometrics*, **8**, 349–363.

Lucasius, C.B. & Kateman, G. (1993). Understanding and using genetic algorithms. Part 1. Concepts, properties and context. *Chemistry of Intelligent Laboratory Systems*, **19**, 1–33.

McCaig, T.N. (2002). Extending the use of visible/near-infrared reflectance spectrophotometers to measure colour of food and agricultural products. *Food Research International*, **35**, 731–736.

McClure, W.F. (1994). Near infrared spectroscopy-the giant is running. *Analytical Chemistry*, **66**, 43A–53A.

Miralbés, C. (2004). Quality control in the milling industry using near infrared transmittance spectroscopy. *Food Chemistry*, **88**, 62–68.

Mosley, R.M. & Williams, R.R. (1998). Determination of the accuracy and efficiency of genetic regression. *Applied Spectroscopy*, **52**, 1197–1202.

Özdemir, 2005.Özdemir, D. (2005). Determination of octane number of gasoline using near infrared spectroscopy and genetic multivariate calibration methods. *Petroleum Science and Technology*, **23**, 1139–1152.

Özdemir and Dinç, 2004.Özdemir, D. & Dinç, E. (2004). Determination of thiamine HCl and pyridoxine HCl in pharmaceutical preparations using uv–visible spectrophotometry and genetic algorithm based multivariate calibration methods. *Chemical and Pharmaceutical Bulletin*, **52**, 810–817.

Özdemir and Öztürk, 2004.Özdemir, D. & Öztürk, B. (2004). Genetic multivariate calibration methods for near Infrared (NIR) spectroscopic determination of complex mixtures. *Turkish Journal of Chemistry*, **28**, 497–514.

Özdemir and Williams, 1999.Özdemir, D. & Williams, R.R. (1999). Multi-instrument calibration in uv-visible spectroscopy using genetic regression. *Applied Spectroscopy*, **53**, 210–217.

Paradkar, R.P. & Williams, R.R. (1997). Genetic regression as a calibration technique for solid phase extraction of dithizone-metal chelates. *Applied Spectroscopy*, **51**, 92–100.

Pizarro, M.C., Forina, M., Casolino, M.C. & Leardi, R. (1998). Extraction of representative subsets by potential functions methods and genetic algorithms. *Chemistry of Intelligent Laboratory Systems*, **40**, 33–51.

Puchwein, G. & Eibelhuber, A. (1989). Outlier detection in routine analysis of agricultural grain products by near-infrared spectrometry. *Analytica Chimica Acta*, **223**, 95–103.

Sorvaniemi, J., Kinnunen, A., Tsados, A. & Mälkki, Y. (1993). Using partial least squares regression and multiplicative scatter correction for FT-NIR data evaluation of wheat flours. *Food Science and Technology*, **26**, 251–258.

Tran, C.D., Oliveira, D., Grishko, V.I. (2004). Determination of enantiomeric compositions of pharmaceutical products by near-infrared spectrometry. *Analytical Biochemistry*, **325**, 206–214.

Wentzell, P.D., Andrews, D.T. & Kowalski, B.R. (1997). Maximum likelihood multivariate calibration. *Analytical Chemistry*, **69**, 2299–2311.

Wesley, I.J., Larroque, O., Osborne, B.G., Azudin, N., Allen, H. & Skerritt, J.H. (2001). Measurement of gliadin and glutenin content of flour by nir spectroscopy. *Journal of Cereal Science*, **34**, 125–133.