Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Nuclear magnetic resonance-based screening of thalassemia and quantification of some hematological parameters using chemometric methods

Mohammad Arjmand[a,*], Mohsen Kompany-Zareh[b,*], Mahdi Vasighi[b], Nastran Parvizzadeh[a], Zahra Zamani[a], Fereshteh Nazgooei[a]

[a] Department of Biochemistry, Pasteur Institute of Iran, Tehran, Iran
[b] Department of Chemistry, Institute for Advanced Studies in Basic Sciences, Zanjan, Iran

ARTICLE INFO

Article history:
Received 26 August 2009
Received in revised form
13 December 2009
Accepted 7 February 2010
Available online 16 February 2010

Keywords:
Thalassemia
Nuclear magnetic resonance
Metabonomics
Blood serum
Linear discriminant analysis
Radial basis function neural network

ABSTRACT

High-resolution ^{1}H NMR spectroscopy of biofluids is a good representation of metabolic pattern and offers a high potential noninvasive technique for pathological diagnosis. Diagnosis of thalassemia and quantification of some blood parameters can be performed by using ^{1}H NMR spectra of human blood serum in parallel with chemometric techniques. Spectra of 28 samples were collected from 15 adult male and female thalassemia patients as experimental set and 13 healthy volunteers as control set. Principal component analysis (PCA) as a dimension reduction tool was used for transforming spectra to abstract factors. The abstract factors were introduced to linear discriminant analysis (LDA), which is a common technique for classification, in order to establish adequate model for discrimination of healthy and unhealthy samples. In addition, these abstract factors were used for calibration of some blood parameters using radial basis function neural network (RBFNN) as an artificial intelligence modeling method. Different test sets (left out samples in training algorithm) were used for evaluating the quality and robustness of the built models. PCA abstract factors were employed as input for LDA model and successfully classified all the members of the test sets except one member of third test set. RBFNN also has a good capability for modeling the most of blood parameters according to proposed network parameters optimization procedure. We conclude that ^{1}H NMR spectroscopy, LDA and RBFNN assisted by PCA provide a powerful method for thalassemia diagnosis and prediction of some blood variants.

1. Introduction

Thalassemia is an inherited autosomal recessive blood disease that is commonly found in many parts of the world [1,2]. In thalassemia, genetic defect results in reduced rate of synthesis of one of the globin protein chains that makes up hemoglobin and this causes the anemia and several related diseases like bone related disorders such as deformities, scoliosis and osteoporosis which are the characteristic of presenting symptom and finally lead to death [3,4].

Many techniques have been used for screening and diagnosis of hemoglobin variants and thalassemia [5]. Determination of the genetic make up of the person in question and characterization of human blood using complete blood cell count (CBC) are the most reliable methods for diagnosis of thalassemia. However, there is still a limitation in the analysis of data due to a large number of possible candidate characteristics and various types of thalassemia and thalassemia trait [6]. However, using such methods, there would be no information about alterations in the patterns of metabolites present in the biological materials that can give valuable diagnostic information and mechanistic insight into the biochemistry of disease processes and related abnormalities. Due to the complexity of metabolites data in the biological samples, chemometrics methods are necessary to extract the information content of data. The obtained information from metabolite patterns can be employed for classification and diagnosis purposes, or quantification of different factors and metabolite concentration in the complex biological samples such as blood serum. Early attempts to formulate an automated classification of blood related diseases were performed using image analysis [7]. In addition, statistics based diagnosis of hematological abnormality [8], and clustering of anemia, based on ferrokinetic parameters [9] are reported in literatures. Recently, an implementation of a neural network, a k-nearest neighbor technique and a support vector machine [10] as a thalassemia diagnostic tool based on blood related parameters was reported. Chaiyaratana and co-workers [11] presented a method using neural network and a decision tree which evolved by

* Corresponding authors.
E-mail addresses: Arjmand@pasteur.ac.ir, arjmand1@yahoo.com (M. Arjmand), kompanym@iasbs.ac.ir (M. Kompany-Zareh).

0039-9140/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2010.02.014

genetic programming, in thalassemia classification by inspecting characteristics of red blood cells, reticulocytes and platelets.

High-resolution ^{1}H NMR spectroscopy of biofluids is a good representation of metabolic patterns and provides information about both structure and composition of low molecular mass metabolites in biological fluids. It is also a powerful, noninvasive technique for investigating disease states in clinical studies [12,13]. Attempts to in vivo and in vitro study of body iron overload by monitoring iron level in the fraction of tissue were performed by NMR spectroscopy and reported in several papers [14–16].

Application of chemometrics techniques has made a large improvement in the performance of NMR-based methods for investigation of biological fluids. Pattern recognition (PR) and data reduction methods have been applied on complex urine NMR spectra to enhance the interpretation of spectral data [17]. Dimensionality of the ^{1}H NMR data can be reduced by using principal component analysis (PCA). NMR in parallel with pattern recognition techniques has been used to classify several inborn errors of metabolism using PCA of urine spectra [18] and blood spots [19]. Recently, the use of metabonomics and chemometrics to identify patients suffering from coronary artery occlusion based on ^{1}H NMR spectra of blood serum has been highlighted [20]. In many cases, PCA models can be applied for data pretreatment [21], solving classification problems in NMR spectroscopy [22] and NMR-based metabonomics [23]. However, in PCA, new basis sets are oriented according to the largest variance and not the largest class separation ability. Linear discriminant analysis (LDA) is frequently used as supervised pattern recognition technique for analysis of complex NMR data sets [22,24]. As far as we know, there is no report on the application of ^{1}H NMR spectroscopy for diagnosis of thalassemia using pattern recognition techniques, and first part of this report is on this subject.

Second part of this article deals with quantification of some hematological parameters using available ^{1}H NMR spectra of healthy and unhealthy volunteers employing different multivariate calibration techniques. ^{1}H NMR spectroscopy of human blood serum used for the calibration of blood metabolites which characterize thalassemia and related diseases, such as various types of bone disorder, will be the focus of our future work.

Artificial neural networks (ANNs) are non-linear calibration techniques and have a great power to handle non-linear problems as well as linear ones. Application of neural networks in quantification of lipid content of human blood plasma and metabolites using NMR spectroscopy was reported in early works [25–27]. Among ANNs, radial basis functions neural networks (RBFNNs) offer some advantages over other ANNs. RBFNNs allow modeling of non-linear data using a linear approach and parameters can be adjusted by fast linear methods with small training times and is ensured to reach the global minimum of error surface during training [28]. Recently, RBFNNs have been successfully applied in many multivariate calibrations and model development studies [29–31]. In order to bring up high predictive power of network we defined a criterion to find best network parameters and guarantee minimum possible error for quantification of blood related parameters. PCA scores are used as input for RBF network and thereupon number of scores as dimensionality reduction parameter also should be considered and optimized simultaneously with network parameters according to defined criterion. The data set used in this investigation is highly informative and required minimum pretreatment and could be used for both classification and calibration purposes with defined models in order to have discriminatory and predictability power as high as possible. This work is the first application of RBFNN in a NMR-based quantification in a biological sample. In comparison with PLSR, significantly better performance was obtained using RBFNN, as an intelligent flexible modeling technique.

2. Materials and methods

2.1. Sample collection and preparation

Twenty-eight blood samples were collected from 15 adult male and female β-thalassemia patients undergoing iron chelation and blood transfusion as experimental set and 13 healthy individual as control set from Imam Khomeini hospital at Tehran. Serum and plasma samples were drawn in Vacutainers (BD Company, USA) before patients being transfused and serum kept in 4 °C before set of experiments. Complete blood cell count (CBC) was done within 1 h.

2.2. Instrumentation

All spectra were recorded at 25 °C on a Bruker 400 MHz NMR spectrometer operating at 399.69 MHz for ^{1}H (Fig. 1). For each serum sample (150 μL 90%/10% mixtures of serum/D_2O), the free induction decay (FID) was weighted by an exponential function with a 0.3 Hz line-broadening factor prior to Fourier transformation (FT). Water pre-saturation pulse sequence (D-90-t1-90-tm-90-acquired FID) with relaxation delays of 5 s and flip angle 90 were used. The water signals and broad protein resonances were suppressed by a combination of pre-saturation and the Carr–Purcell–Meiboom–Gill (CPMG) (90-(t-180-tn-acquisition) $t' = 200$, $n = 100$) pulse sequence [32].

Calcium and phosphate level of serum samples were analyzed by Roch-Hitachi-912 Autoanalyser and complete blood cell count (CBC) performed using Automated Sysmex KX21-N hematology counter.

2.3. Data analysis

All calculations were carried out using an Intel Celeron 3.20 GHz computer running Windows XP operating system. PCA and LDA were coded manually in MATLAB (Mathwork Inc.) for pattern recognition analysis. Normalization and binning as preprocessing methods were coded manually in MATLAB software too. RBFNN was performed on dataset using MATLAB ANN toolbox.

2.3.1. PCA

PCA, as a linear projection method is based on variance, transforms the original measurement variables into new uncorrelated variables called principal components [33,34]. In many cases, only a few axes can define most of variation in data. In this way, PCA can be used as a dimension reduction method [35] to simplify the working with huge data sets like NMR data sets and features extraction. In many cases, PCA models can be applied for solving classification

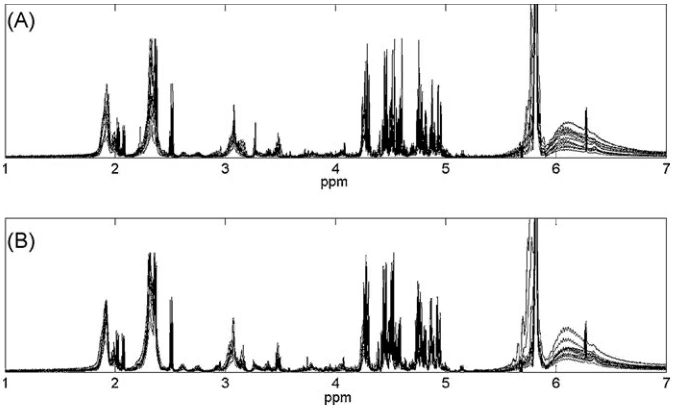

Fig. 1. ^{1}H NMR spectra of (a) 13 healthy and (b) 15 unhealthy samples.

problems, if the class memberships are known in advance. However, in PCA, new basis sets are oriented according to the largest variance, not according to the largest class separation ability.

2.3.2. *LDA*

Linear discriminant analysis (LDA) is a frequently used technique for dimension reduction and feature extraction. It projects points to a smaller dimension hyperplane using a linear function of the variables, which maximizes between-class variance and minimizes the within-class variance, which fulfill maximum separation among the given classes. Description of the LDA algorithm can be found in Ref. [36] in detail.

2.3.3. *PLS*

PLS-regression (PLSR) provides an approach to the quantitative modeling of the complicated relationships between predictors, X, and responses, Y, by a linear multivariate model, but goes beyond traditional regression by using structures of X and Y in modeling procedure instead of using the only structure of X.

2.3.4. *RBFNN*

RBFNN [37] is a three-layer network in which the activation function in the nodes of the hidden layer is a radial type function. Euclidean distance between the input object and the center of the radial basis function is the input of network. Adjusting the weights connecting the output layer and the hidden layer implemented to minimize the mean square error of the net output. Gaussian function is used here as activation function of RBFNN which is characterized by two parameters, i.e. center (basis function) and width (spread). Number of basis functions and spread value are the critical issues in constructing RBF networks. Some efforts are made in solving this problem in intelligent way [38]. In our work, the best structure of RBFNN is designed by defining an error related criterion which is minimized in optimal network parameters.

3. Results and discussion

3.1. *Classification of ^1H NMR spectra*

3.1.1. *Principal component analysis of the ^1H NMR spectral data*

The data matrix which rows are ^1H NMR spectra of samples (Fig. 1) was subjected to PCA analysis. The large ^1H NMR multivariate dataset transformed by PCA into a low-dimensional space that is more conducive to handling and visualization. The majority of the structure in data can be represented in a small number of PCs space instead of high dimension variable space by the scores associated with these PCs for each observation. Fig. 2a shows the 3D and 2D score plots from PCA analysis on raw spectral data. Each ^1H NMR spectrum denotes by a point in principal component space. 3D score plot revealed that healthy and patient samples are not clearly

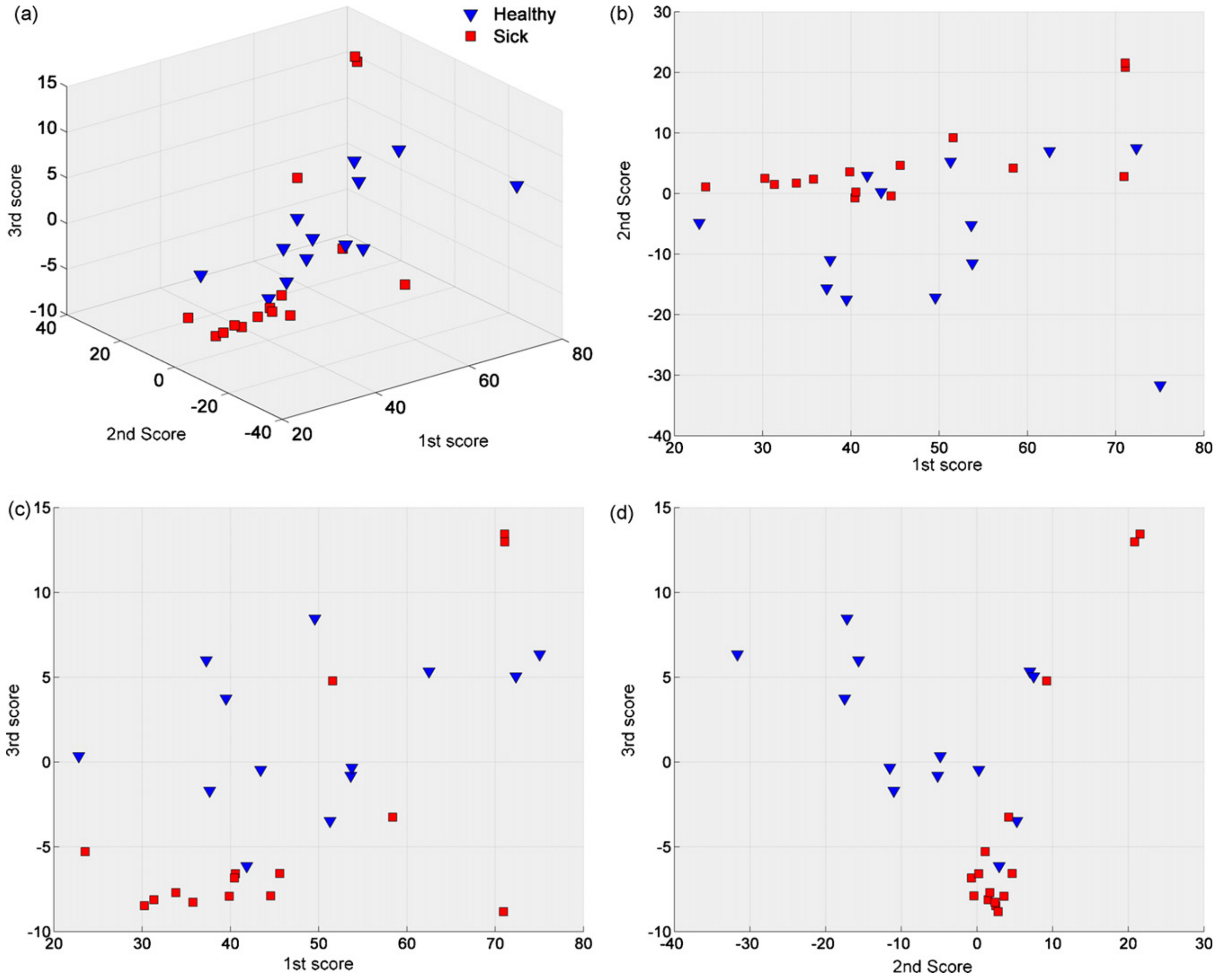

Fig. 2. (a) 3D score plot from PCA analysis on raw spectral data; 2D score plot of (b) 1st vs. 2nd PC, (c) 1st vs. 3rd and (d) 2nd vs. 3rd.

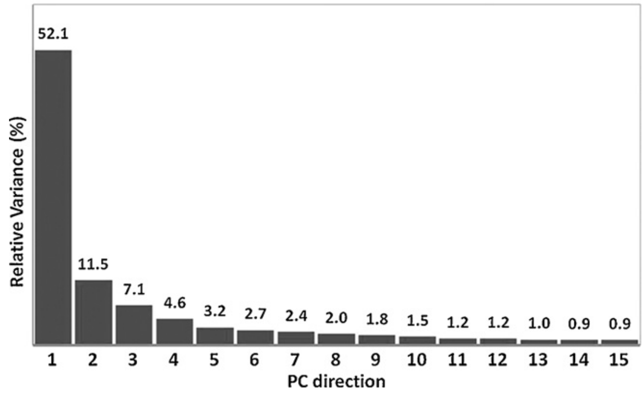

Fig. 3. Percentage of variance in first 15 PCs direction relative to total variance in raw data matrix.

discriminated in the space spanned by first three PCs and we cannot separate two groups by a straight line, although they approximately formed different clusters. In other words, information content in the first three scores is not enough for discriminatory purpose and visual investigation of more than three scores is not possible. Fig. 3 shows percentage of variance in first 15 PCs direction relative to total variance in raw data matrix. Although about 70% of variance is included in first three PCs (as shown in Figs. 2 and 3), the results show that three PCs are not enough for perfect discrimination in the considered data set. Presence of 52.1% of variance in the first PC indicates the extent of similarity of the spectral data from different samples (Fig. 3). In this way, determination of optimum number of scores is a crucial point. When class memberships are available, coordination of the points, which are PCA scores, were introduced into a supervised pattern recognition technique like linear discriminant analysis (LDA).

3.1.2. Linear discriminant analysis (LDA)

The number of variables in the data matrix (27,500) was too large to apply LDA directly on the data. In this way PCA as a simple tool for dimension reduction with preserving information was used, and then, LDA was applied on scores from PCA because the traditional LDA algorithm cannot be used directly due to singularity in the within-class scatter matrix. In addition, the high-dimensional vectors lead to computational difficulty. Application of PCA before LDA is reported in the literature [39]. Accordingly, the number of scores as input for LDA was a crucial parameter that should be considered. In order to find optimal number of PCs for building an optimal model for classification, total populations were split into seven sub-groups. One sample from each of seven sub-groups was used to build test set and other three members in each sub-group were chosen to form the training set. In this way, test set included seven members and training set included 21 members. In second step, leaved-one out cross-validation (LOOCV) was applied on training set. To build a proper classification model using LOOCV, different number of PCA scores was introduced into LDA and the left out sample was classified, as internal validation sample. All samples in training set left out once and classified as mentioned.

As each sub-group included four members, four different arrangements for training and test sets, consisting different members were built in a similar way and were investigated concurrently. Number of misclassification for training and test sets is related to number of input PCA scores introduced into LDA. Different preprocessing techniques like normalization and binning were performed on raw data matrix before PCA-LDA based classification. Table 1 shows the number of misclassifications for different preprocessing approaches applied on data at different number of scores. Results reveal that intensity of peaks in NMR spectra (norm of vectors) is an important parameter in classification and normalization can decrease the performance of discrimination according to increasing number of misclassifications compared to the raw data set. Raw ^{1}H NMR spectra was also compressed using binning technique (50 point integration width). Binning preprocessing has shown superior results with no misclassification using first 11 PCs as shown in Table 1.

Table 1
Number of misclassifications for calibration set (first arrangement) using different preprocessed data at different number of PCs.

Number of scores	Raw data	Normalized spectra	Binned spectra
2	9	6	10
3	4	4	9
4	5	6	5
5	2	2	2
6	1	4	3
7	1	2	3
8	1	4	2
9	3	4	3
10	2	4	1
11	1	6	0
12	1	3	0
13	2	4	0
14	2	3	0
15	3	4	0
16	1	2	0
17	2	3	1
18	3	1	3
19	3	7	1
20	2	4	1

According to Table 1, binning of spectra (with a 50 points integrating width) and using first 11 scores (SC = 11) from PCA resulted in no misclassification. In this way, binning was employed as pretreatment before building the classification model. This model was checked using training and test sets from four different arrangements and all the samples in test set were classified correctly except just one sample in the third arrangement that showed magnificent predictability for built models. Number of scores (SC) for building the model was chosen according to Table 1. Similar results were obtained for raw (unpreprocessed) spectra too. The only difference between these binned and unpreprocessed data results is the number of misclassifications in LOOCV, which shows that the results for binned data are slightly better. Accordingly binning of spectra, as a compression technique, can be used as preprocessing technique for best discrimination between healthy and thalassemia patients.

3.2. Quantification of hematological parameters

After performing CBC analysis and determination of calcium and phosphate level (mg dL^{-1}), it is apparent that parameters related to thalassemia disorder are significantly different for healthy (control) and unhealthy (experimental) blood samples in hematological and biochemical aspects according to box-plots in Fig. 4. Hematocrit (HCT) and mean cell volume (MCV) values are scaled (divided by 10) for better representation in the figure. These differences between healthy and unhealthy populations could lead us to conclude that the difference between NMR spectra of healthy and unhealthy blood samples is related to biochemical and hematological difference of them and are informative enough to build calibration models for prediction of these parameters.

3.2.1. Partial least squares

To develop a linear calibration model for quantification of each blood variants, a PLS model was built. Number of PLS factors for building the calibration model was the influencing factor on quality of model. Accordingly, a criterion for determining optimal number

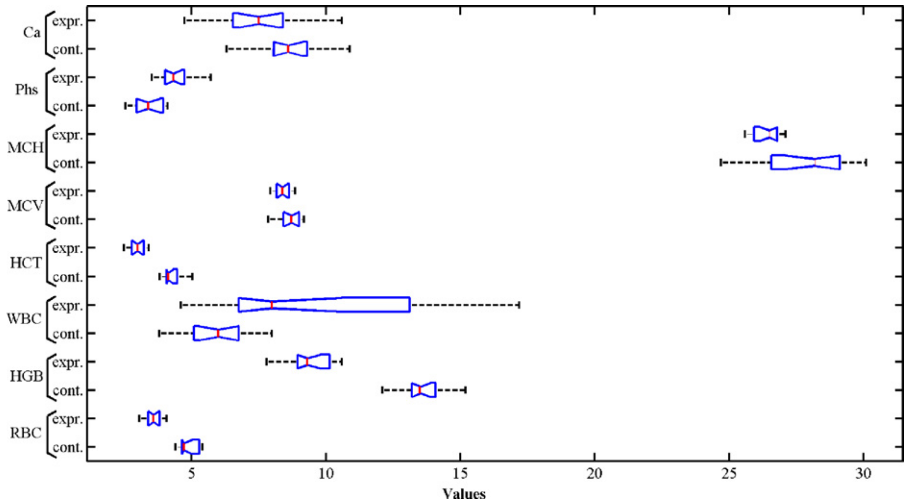

Fig. 4. Box plot representation of some biochemical (calcium (Ca) and phosphate (Phs)) and hematological parameters (red blood cell count (RBC), hemoglobin (HGB), white blood cell count (WBC), hematocrit (HCT), mean corpuscular volume (MCV) and mean corpuscular hemoglobin (MCH)) for healthy (cont.) and unhealthy (expr.) samples. For better representation, HCT and MCV values are scaled (divided by 10).

of PLS factors was defined. The criterion is representative for values of error from both calibration and validation sets.

$$\text{Criterion} = \text{RMSEP\%} \times |\text{RMSEP\%} - \text{RMSECV\%}| \qquad (1)$$

This criterion fulfills minimal difference between cross validation error of calibration set (RMSECV%) and prediction error of validation set (RMSEP%), and also minimum value for RMSEP% as possible. At different number of PLS factors, RMSECV% was obtained using leaved-one out cross validation (LOOCV) on calibration set and RMSEP% was calculated using built PLS model. In this way, for each number of PLS factors we have a criterion value. Best predictive model was obtained using optimum number of PLS factors at minimum value of criterion. Four different arrangements for calibration and validation sets were utilized, according to classification section. Table 2 shows RMSECV% and RMSEP% for four different arrangements of calibration and validation sets. Appropriate number of PLS factors was obtained using first calibration and validation set as mentioned before and used for analysis of other sets. According to results, only for platelet count (PLT), phosphate and lymphocyte percent (LYM%), RMSECV% and RMSEP% are below 10% just for one of arrangements and for other parameters these errors are high. This mean that considered biochemical system is too complex to be modeled by a linear regression method like PLS. This led us for applying artificial neural networks to improve modeling performance.

3.2.2. *Radial basis function neural network (RBFNN)*

Same arrangements (training and validation sets) as in classification section were used for calibration of different hematological parameters using radial basis function neural networks (RBFNN). Score values from PCA as input and values of hematological parameters as target were introduced into RBFNN for building the model. Number of neurons in the network (MN), spread parameter in radial basis function (SP) and number of scores as input (SC) were considered as the influencing factors on the estimated calibration and validation error values. RMSECV% and RMSEP% values were estimated for all possible combinations of different values for MN, SP and SC. To fulfill minimal difference between RMSECV% and RMSEP% and also the minimal value for RMSEP%, similar criterion as used in PLS calibration (Eq. (1)) was utilized.

At different levels of SC, MN and SP, RBFNN was trained using first training set and validated using first validation set (arrangement 1). In each level of SC, MN and SP, defined criterion was calculated using RMSECV% and RMSEP% and finally a cube of esti-

Table 2
RMSECV% and RMSEP% for different arrangements of calibration and validation sets at optimum number of factors using PLSR.

Hematological parameters	Number of factors[a]	Arrangement 1		Arrangement 2		Arrangement 3		Arrangement 4	
		RMSECV%	RMSEP%	RMSECV%	RMSEP%	RMSECV%	RMSEP%	RMSECV%	RMSEP%
HGB	1	35.31	13.84	35.56	15.23	34.54	5.15	42.25	45.76
HCT	1	36.66	12.61	35.69	10.77	35.09	8.17	43.24	44.32
MCH	1	35.42	8.62	34.89	16.72	35.73	30.20	43.14	41.40
RBC	2	20.58	31.04	23.56	20.63	22.61	6.86	28.36	28.48
WBC	7	37.60	37.06	18.34	44.37	0.61	84.04	1.50	16.56
MCV	1	37.84	5.37	36.30	13.89	37.37	32.72	45.34	38.89
PLT	2	27.19	28.16	21.41	2.56	14.34	7.98	5.93	1.99
MCHC	2	20.09	24.89	5.53	11.36	10.68	12.56	20.33	13.71
Phosphate	3	8.30	4.33	12.58	3.67	16.27	0.51	10.95	29.56
Calcium	16	21.55	0.26	9.00	72.32	6.87	29.12	7.91	29.36
LYM%	2	17.69	5.90	11.28	24.89	3.60	5.88	12.39	39.84
MXD%	16	33.06	34.40	13.18	1.84	7.77	23.78	43.78	10.80
NEUT%	1	28.45	13.82	30.42	34.47	26.41	28.49	36.62	7.65
LYM#	5	54.26	0.09	34.89	11.38	1.07	88.92	4.89	67.90

HGB: hemoglobin; HCT: hematocrit; MCH: mean corpuscular hemoglobin; RBC: red blood cell count; WBC: white blood cell count; MCV: mean corpuscular volume; PLT: platelet count; MCHC: mean corpuscular hemoglobin concentration; LYM%: lymphocyte count percent; MXD%: mixed cell count; NEUT%: neutrophil count; LYM#: lymphocyte count.

[a] Optimum number of factors was obtained using arrangement 1 for calibration and validation.

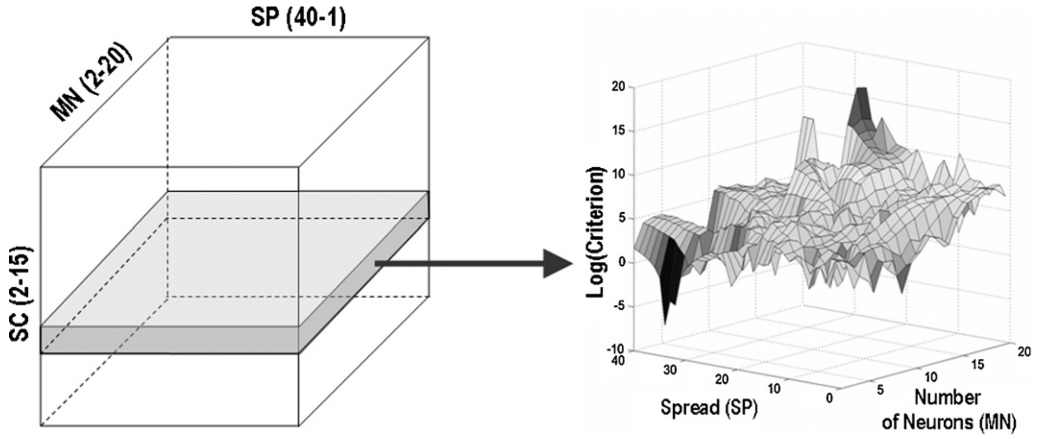

Fig. 5. Criterion cube calculated from Eq. (1) at different MN, SP and SC. The surface plot shows values of log(criterion) at constant SC and different combinations of MN and SP.

mated criteria was obtained (Fig. 5). We can plot a surface (MN vs. SP) for criterion at constant SC as shown in Fig. 5 but should find minimum value of criterion in the whole cube and representing the surface is just for better explanation. Minimum value in this cube shows the best MN, SP and SC to build the most predictive model. For each hematological parameter, the cube was calculated separately. In this way, for each hematological parameter an individual model was built using different values of MN (from 2 to 20), SP

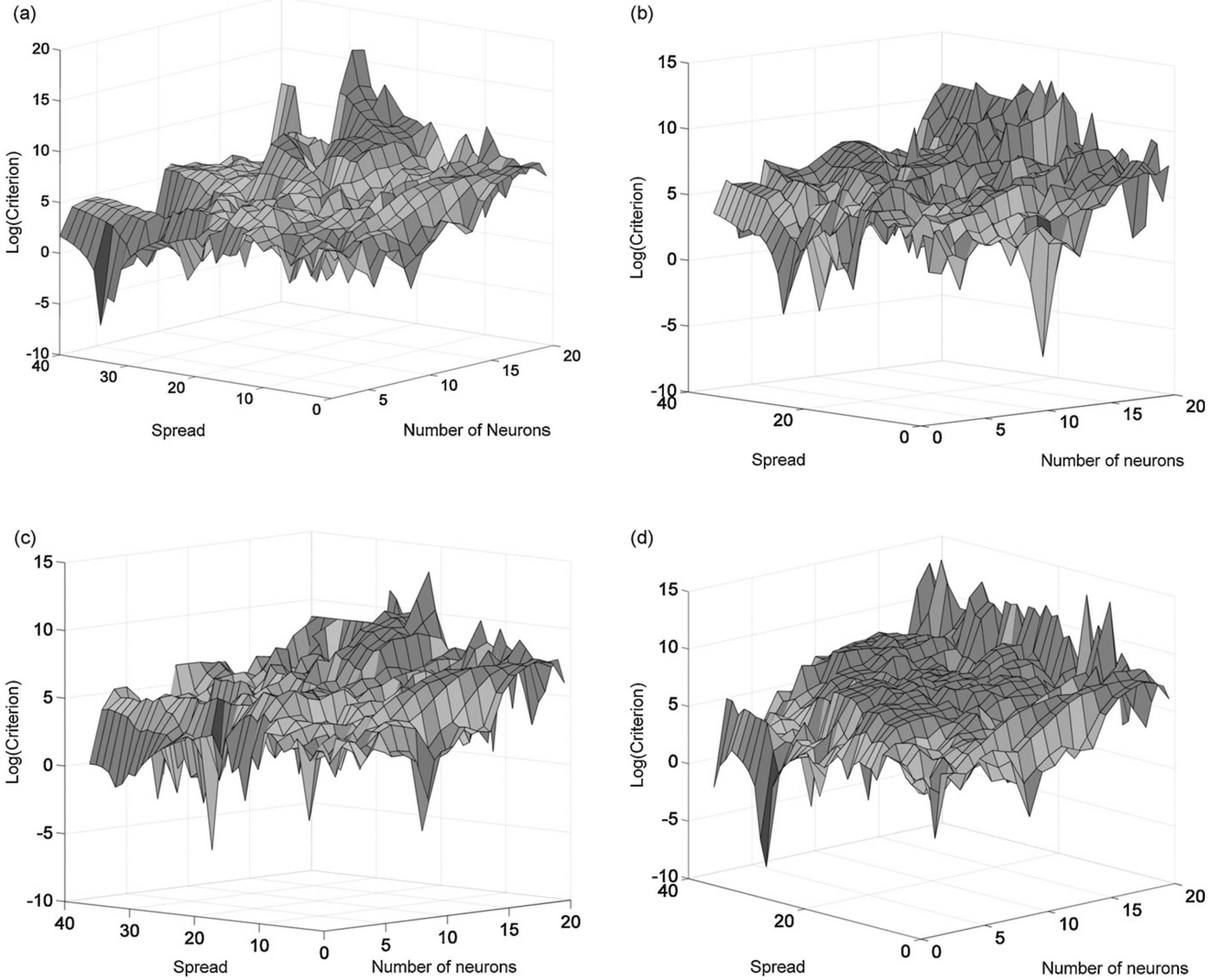

Fig. 6. Surface plot of criterion at a constant SC (a) (SC = 13) in logarithmic scale for HGB, (b) (SC = 10) in logarithmic scale for PHS, (c) (SC = 9) in logarithmic scale for MCH, and (d) (SC = 6) in logarithmic scale for HCT.

Table 3
RMSECV% and RMSEP% for different arrangements of calibration and validation sets at optimum number of factors using RBFNN.

Hematological parameters	Network parameters[a]			Arrangement 1		Arrangement 2		Arrangement 3		Arrangement 4	
	SC	MN	SP	RMSECV%	RMSEP%	RMSECV%	RMSEP%	RMSECV%	RMSEP%	RMSECV%	RMSEP%
HGB	13	2	34	0.37	0.37	5.22	3.96	6.01	19.46	3.22	0.24
HCT	6	2	31	0.19	0.19	7.78	1.21	5.34	17.73	2.88	4.32
MCH	9	3	29	0.24	0.25	2.94	1.17	0.72	8.34	0.42	5.15
RBC	2	3	38	0.62	0.008	6.95	1.98	0.66	15.24	12.94	5.90
WBC	5	3	15	0.34	0.30	9.78	10.74	11.23	47.55	11.09	23.98
MCV	10	9	28	0.28	0.00	7.22	3.66	11.08	9.64	2.87	3.03
PLT	2	8	1	0.68	0.67	2.84	1.07	8.40	23.11	1.52	1.97
MCHC	5	12	22	0.03	0.03	8.23	7.70	2.18	4.73	1.16	2.87
Phosphate	10	11	3	2.47	2.47	2.70	4.22	11.93	9.46	1.22	16.75
Calcium	13	4	18	0.20	0.16	6.92	36.77	4.70	7.93	2.75	5.92
LYM%	5	11	22	0.11	0.26	16.27	12.56	30.14	40.45	4.67	38.07
MXD%	4	3	23	5.10	5.10	0.78	43.50	17.41	32.33	22.24	23.53
NEUT%	7	15	6	2.85	2.86	70.86	4.20	143.13	41.71	189.71	60.29
LYM#	5	2	8	18.85	18.85	8.89	4.02	1.16	76.94	15.77	32.65

HGB: hemoglobin; HCT: hematocrit; MCH: mean corpuscular hemoglobin; RBC: red blood cell count; WBC: white blood cell count; MCV: mean corpuscular volume; PLT: platelet count; MCHC: mean corpuscular hemoglobin concentration; LYM%: lymphocyte count percent; MXD%: mixed cell count; NEUT%: neutrophil count; LYM#: lymphocyte count.

[a] Optimum number of scores and values for network parameters were obtained using arrangement 1 for calibration and validation.

(from 1 to 40) and SC (from 2 to 15) with increment equal to 1 and cube dimension 19 × 40 × 14 in this study. To determine the performance of the formed network, optimum values of MN, SC, and SP from the first arrangement was applied to build the calibration models for second, third and fourth arrangements. Fig. 6a shows the surface plot of criterion in logarithmic scale at SC equal to 13 (~98.1% variance) which contains global minimum in defined criterion cube which was calculated using first arrangement of training and validation set. The logarithmic scale was just used for better presentation of the surfaces. In ordinary scale, the plots were not clear and understandable. Fig. 6b–d also shows similar surface plots for optimization procedure of phosphate (PHS) (SC = 10), mean corpuscular hemoglobin (MCH) (SC = 9) and hematocrit (HCT) (SC = 6) respectively. As shown in these figures, we utilized grid search (trying all possible combinations of the parameters) to find the minimum criterion value (the best condition) and to avoid trapping in the local minima. This may be a bit time-consuming procedure but we expected the noisy surface and presence of many local minima. Globally, increasing the number of neurons in the network architecture results in higher criterion value in all cases that is due to overtraining of the RBFNN on training set and increasing RMSEP%.

At SC = 13, MN = 2 and SP = 34 for RBFNN resulted in minimum criterion value for HGB, which means the minimum and comparable RMSECV% and RMSEP% values for HGB (Fig. 6a). Similar methodology was used for other hematological parameters and the results are shown in Table 3. As shown in Table 3, different hematological parameters were modeled using different SC, MN and SP and good results were obtained for most of hematological parameters.

For HGB, HCT, MCH, MCV, PLT and MCHC, RMSECV% and RMSEP% are low for the first, second and fourth arrangements and error is high for the third arrangement. This could be due to members of calibration and validation set in the third arrangement. Members in validation set are acting an important role in building a predictive model and when we left them out, as validation set remaining ones cannot build a true predictive model. Hence, error values for the third arrangements are below 20% for these parameters except for PLT. In addition, error values for some hematological parameters such as MCH and MCHC are low for all arrangements that mean that they can easily be modeled by this procedure. According to RMSECV% and RMSEP% for LYM%, LYM#, MXD% and NEUT%, these parameters was not models as well as others and model failed in these cases.

4. Conclusions

In this paper, the ^{1}H NMR spectra of human blood serum were used for almost exact classification of normal persons and those suffering from thalassemia. Quantification of some blood cell parameters and variants was performed, as well. In the first part, obtained ^{1}H NMR spectra were binned and subjected to PCA. Four arrangements were considered for calibration and validation sets and accordingly the corresponding LDA models were built using scores from PCA. Small number of misclassified samples for calibration and validation sets showed the power of LDA based model for discrimination of the two classes.

In the second part, quantification of some blood cell parameters and variants were performed using multivariate calibration methods. PLSR, as a common linear regression method, was applied. The obtained prediction error values were considerable and showed that the complex relation between spectra and blood parameters cannot be modeled in a linear manner. RBFNN as a non-linear calibration tool performed to build a model using scores from PCA. Error of prediction for most of blood parameters was low compared to PLSR method.

Totally, the high information content of ^{1}H NMR in parallel with specific modeling ability of RBFNN can be used as a tool for quantification of metabolite content of biofluids which lead to a direct insight in metabolism of many disorders and diseases which is among our future outlooks.

Acknowledgement

The authors acknowledge Dr. Mina Izadyar, Immam Khomieni hospital, Tehran for her timely help.

References

[1] D.J. Weatherall, J.B. Clegg, The Thalassemia Syndromes, fourth ed., Blackwell Science, Malden, MA, 2001.
[2] F.H. Bunn, B.G. Forget, H.M. Ranney, Hemoglobinopathies, WB Saunders, Philadelphia, PA, 1997.
[3] A. Filosa, S. Maio, S.Di. Vocca, A. Saviano, G. Esposito, L. Pagano, Acta Paediatr. 86 (1997) 342.
[4] P.E. Vichinsky, Ann. N. Y. Acad. Sci. 850 (1998) 344.
[5] S.K. Hartwell, B. Srisawanga, P. Kongtawelert, D. Christianc, K. Grudpana, Talanta 65 (2005) 1149.
[6] N.I. Birndorf, J.O. Pentecost, J.R. Coakley, K.A. Spackman, Comput. Biomed. Res. 29 (1996) 16.
[7] P.R. Lund, R.D. Barnes, The Lancet 300 (7775) (1972) 463.

[8] R.L. Engle, B.J. Flehinger, S. Allen, R. Friedman, M. Lipkin, B.J. Davis, L.L. Leveridge, Bull. N. Y. Acad. Med. 52 (1976) 584.
[9] G. Barosi, M. Cazzola, C. Berzuini, S. Quaglini, M. Stefanelli, Br. J. Haematol. 61 (1985) 357.
[10] S.R. Amendolia, G. Cossu, M.L. Ganadu, B. Golosio, G.L. Masala, G.M. Mura, Chemometr. Intell. Lab. Syst. 69 (2003) 13.
[11] W. Wongseree, N. Chaiyaratana, K. Vichittumaros, P. Winichagoon, S. Fucharoen, Sciences 177 (2007) 771.
[12] R.H. Barton, J.K. Nicholson, P. Elliott, E. Holmes, Int. J. Epidemiol. 37 (2008) i31.
[13] O. Beckonert, M.E. Bollard, T.M.D. Ebbels, H.C. Keun, H. Antti, E. Holmes, J.C. Lindon, J.K. Nicholson, Anal. Chim. Acta 490 (2003) 3.
[14] P.D. Jensen, F.T. Jensen, T. Christensen, J. Ellegaard, Br. J. Haematol. 87 (1994) 171.
[15] R.M. Dixon, P. Styles, F.N. al-Refaie, G.J. Kemp, S.M. Donohue, B. Wonk, A.V. Hoffbrand, G.K. Radda, B. Rajagopalan, Hepatology 19 (1994) 904.
[16] P. Liu, M. Henkelman, J. Joshi, P. Hardy, J. Butany, M. Iwanochko, M. Clauberg, M. Dhar, D. Mai, S. Waien, N. Olivieri, Can. J. Cardiol. 12 (1996) 155.
[17] E. Holmes, P.J.D. Foxall, J.K. Nicholson, G.H. Neild, S.M. Brown, C.R. Beddell, B.C. Sweatman, E. Rahr, J.C. Lindon, M. Spraul, P. Neidig, Anal. Biochem. 220 (1994) 284.
[18] J.C. Lindon, J.K. Nicholson, J.R. Everett, Ann. Rep. NMR Spectrosc. 38 (1999) 1.
[19] M.A. Constantinou, E. Papakonstantinou, D. Benaki, M. Spraul, K. Shulpis, M.A. Koupparis, E. Mikros, Anal. Chim. Acta 511 (2004) 303.
[20] J.T. Brindle, H. Antti, E. Holmes, G. Tranter, J.K. Nicholson, H.W.L. Bethell, S. Clarke, P.M. Schofield, E. Mckilligan, D.E. Mosedale, D.J. Grainger, Nat. Med. 8 (2002) 1439.
[21] A.A. Christy, S. Kasemsumran, Y.P. Du, Y. Ozaki, Anal. Sci. 20 (2004) 935.
[22] S. Rezzi, D.E. Axelson, K. Heberger, F. Reniero, C. Mariani, C. Guillou, Anal. Chim. Acta 552 (2005) 13.
[23] M.A. Constantinou, E. Papakonstantinou, M. Spraul, S. Sevastiadou, C. Costalos, M.A. Koupparis, K. Shulpis, A. Tsantili-Kakoulidou, E. Mikros, Anal. Chim. Acta 542 (2005) 169.
[24] S. Rezzi, I. Giani, K. Heberger, D.E. Axelson, V.M. Moretti, F. Reniero, C. Guillou, J. Agric. Food Chem. 55 (2007) 9963.
[25] Y. Hiltunen, E. Heiniemi, M. Alakorpela, J. Magn. Reson. 106 (1995) 191.
[26] T.F. Bathen, J. Krane, T. Engan, K.S. Bjerve, D. Axelson, NMR Biomed. 13 (2000) 271.
[27] Y. Hiltunena, J. Kaartinenb, J. Pulkkinenc, A. Häkkinend, N. Lundbome, R.A. Kauppinenc, J. Magn. Reson. 154 (2002) 1.
[28] B. Walkzak, D.L. Massart, Chemom. Intell. Lab. Syst. 50 (2000) 179.
[29] Q.F. Li, X.J. Yao, X.G. Chen, M.C. Liu, R.S. Zhang, X.Y. Zhang, Z.D. Hu, Analyst 125 (2000) 2049.
[30] Y. Akhlaghi, M. Kompany-Zareh, J. Chemom. 20 (2006) 1.
[31] Y. Akhlaghi, M. Kompany-Zareh, J. Chemom. 21 (2007) 239.
[32] S. Lasic, J. Stepisnik, A. Mohoric, J. Magn. Reson. 182 (2006) 208.
[33] S. Wold, K. Esbensen, P. Geladi, Chemom. Intell. Lab. Syst. 2 (1987) 37.
[34] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S.D.E. Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part B, Elsevier, Amsterdam, The Netherlands, 1998, p. 88.
[35] G. Ivosev, L. Burton, R. Bonner, Anal. Chem. 80 (2008) 4933.
[36] M. Otto., Chemometrics Statistics and Computer Application in Analytical Chemistry, Wiley-VCH, Weinheim, Germany, 1999.
[37] B. Walczak, D.L. Massart, Anal. Chim. Acta 331 (1996) 177.
[38] H. Sarimveis, A. Alexandridis, G. Tsekouras, G. Bafas, Ind. Eng. Chem. Res. 41 (2002) 751.
[39] J. Yang, J. Yu Yang, Pattern Recogn. 36 (2003) 563.