DOI: 10.12982/CMUJNS.2013.0013

K-means Clustering and Hierarchical Cluster Analysis Coupled with Linear Discriminant Analysis to Classify Signals in Osmotic Fragility Test for Thalassemia Screening

Karuna Jainontee[1,4], Vannajan Sanghiran Lee[2*], Sukon Prasitwattanaseree[3,4*] and Kate Grudpan[1,4]

[1]Department of Chemistry, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand
[2]Department of Chemistry, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia
[3]Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand
[4]Center of Excellence in Innovation for Analytical Science and Technology, Chiang Mai University, Chiang Mai 50200, Thailand

**Corresponding authors: E-mail: vannajan@gmail.com,*
sprasitwattanaseree@gmail.com

ABSTRACT

Investigation has been made in applying chemometric treatment of unbiased approaches to classify signals in osmotic fragility test for thalassemia screening. K-means clustering and hierarchical clustering analysis coupled with linear discriminant analysis were the chemometric techniques employed in this work. A knowledge determined from conventional approaches in osmotic fragility test together with multivariate analysis provides a complementary tool and novel approach for disease diagnosis.

Keywords: *K*-means clustering, Hierarchical clustering analysis, Linear discriminant analysis, Thalassemia screening, Osmotic fragility test, Flow injection

INTRODUCTION

The osmotic fragility test (OFT) is a popular method for thalassemia screening because of its high efficiency and low cost (Hartwell et al., 2005). OFT is a test that measures the ability of red blood cells to retain their integrity in hypotonic saline solution. The surface areas of red blood cells of a thalassemia patient (positive tests) are larger than in a normal person (negative tests), so the blood cells are more likely to breakdown faster than in negative tests. OFT, coupled with stopped flow injection, was developed as an automated thalassemia screening tool that reduce the risk of contamination in the system (Khonyoung et al., 2009). The signals of stopped flow OFT are used to discriminate patients. The method can classify patients by using the optimized slope from the optimization based on hospital records (Khonyoung et al., 2009). However, no evidence exists guaranteeing its accuracy in predicting unknown samples.

Attempts to formulate a screening tool for unknown samples have employed various chemometric algorithms. Approaches to unsupervised learning include *k*-means clustering and hierarchical cluster analysis (HCA). Regression of discriminate signal and validated group from linear discriminant analysis (LDA) can be used to predict groups of unknown signals. In this work, k-means clustering and HCA coupled with LDA of various types of stopped flow OFT signals were used to construct a thalassemia screening tool. The ability to predict thalassemia cases in unknown groups by the proposed methods was evaluated.

MATERIALS AND METHODS

Data and data management

OFT signals used in this paper were obtained from prior research (Khonyoung et al., 2009). From the dataset, 73 blood samples were defined, according to the hospital records, as 21 and 52 samples with positive and negative test defined as RT and RN in Table 1, respectively. From the experiments in Figure 1 (Khonyoung et al., 2009), the signals were recorded at the same time period. In this study, the signals were normalized by considering: (1) the range of OFT signals, (2) rescaling of the time scale and (3) smoothing to reduce noise signals.

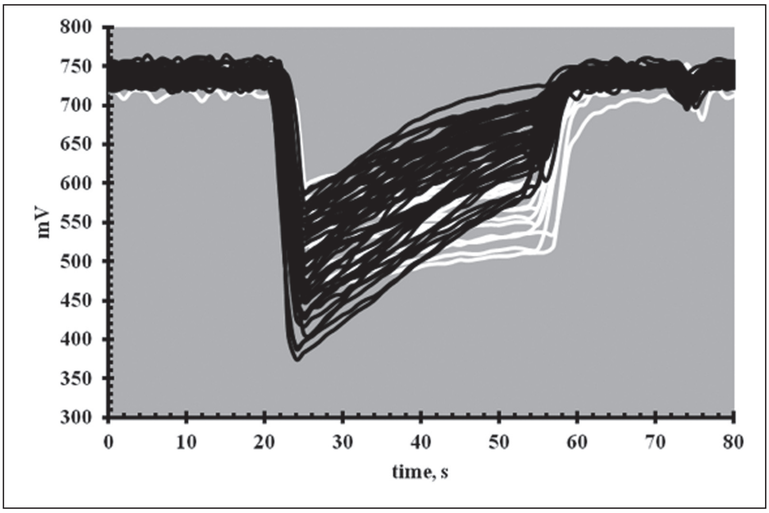

Figure 1. Raw OFT signals of 73 cases (21 positive tests (T) and 52 negative tests (N)).

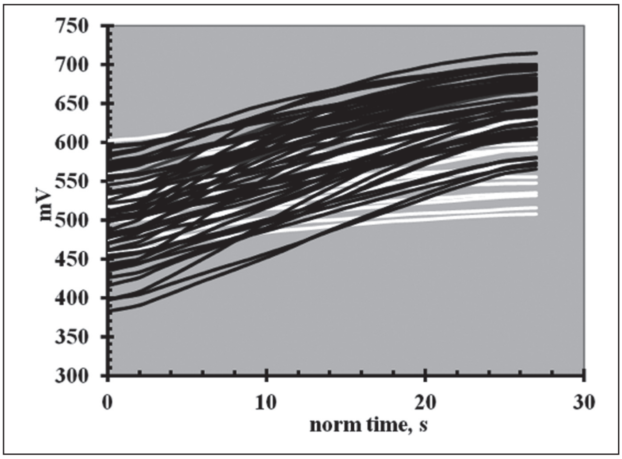

Figure 2. OFT signals from the 23-50 second interval in Figure 1 were rescaled to a 0-27 second range in this study.

Four types of treatment were then employed: (1) normalized signals of OFT in the range of the stopped time period; (2) principal component analysis (PCA) score of normalized OFT signals (PCOFT); (3) slope of the whole range, calculated from the first and the last points (OS) (Figure 2) and (4) interval slope (IS). Each of the four types was then treated by *k*-means clustering and hierarchical clustering analysis (HCA) techniques. The results were considered for the next steps by selecting only the same prediction results for use as training sets. Linear discriminant analysis (LDA) was then applied to predict the group of signals.

Construction of LDA

The selection of training samples for LDA was based on the variation of *k*-means distance. The coverage of *k*-means distance from minimum to maximum of all members in positive test (PT) and negative test (PN) was used as criteria to select training samples. Furthermore, numbers of training samples were also studied by varying from 10-20 samples.

The LDA package in the SPSS program was used in this study. LDA was performed on within-group covariance matrix and prior probabilities of discrimination were set as all groups equal. Canonical function was calculated to maximize the eigenvalue. Groups of OFT signals of training samples were predicted to be positive test (AT) and negative test (AN), respectively. Square mahalanobis distance of sample to centroid (SMDC) value of OFT signals of training samples expressed distribution of the OFT signals on a discriminate scale. The upper bound of SMDC of AT and AN in the training set was used as the boundary of LDA prediction. The prediction was guaranteed when the SMDC was less than the LDA boundary. All LDA models were validated using the same validation set. The sample of which groups of all kinds of signals were concerned as positive (PT) and negative test (PN) was used as the validation sample. In this study, nine samples were used as validation sets, three positive and six negative tests. The percentage correctly classified in the validation set was used to express the LDA prediction ability. The OFT signal predicted within the upper bound of SMDC of the training set as positive (AT) and negative test (AN), respectively; otherwise the signal was predicted as an unidentified case (AU).

OFT signals of unknown samples were normalized using the same process as the training set. Groups of unknown samples were predicted using LDA. The LDA-predicted results were guaranteed within the upper bound of training set SMDC. Other kinds of signals were treated the same as OFT signals for thalassemia screening.

Principal component analysis (PCA)

PCA in the SPSS program was used to calculate linear combination variables, which presented the variance of data (Chatfield and Collins, 1980; Jurado et al., 2005). Correlation matrix was used as variance data. PCA reduces data size while maintaining the important parts of the data by projecting variance data into new eigenvector variables called component (Milde et al., 2007), (De Carvalho et al., 2006). Component contained score that presents the variation on the new dimension of new variables named loading. Eigenvalues express the weight of the eigenvector of the considered component. PCA is a data successive extraction method. After the first component was extracted from the data, the second component was extracted from the remaining data and so on with later components. The linear combination of each component is the orthogonal dimension of the other component. Varimax normalized rotation was performed to maximize (or minimize) the values of the loading matrix in relation to each rotated component (Viana et al., 2006).

The number of principal components was selected using Kaiser criterion in which components were kept only when their eigenvalue were more than 1 (Thanasoulias et al., 2003). In this study, PCA was used for two aims: i) principal component scores were used as discriminate variables and ii) a scatter plot of principal component scores was used to express distribution of multivariate signals (Chatfield and Collins, 1980; Jurado et al., 2005).

Hierarchical clustering analysis (HCA). HCA clusters groups of signals by using distance of signals (Huang et al., 2009; Pham 2001; Smolinski et al., 2002; Teppola et al., 1999). Cityblock distance is one way to express cluster of signals. Cityblock distances calculate the absolute difference of signal *x* that obtained *n* variables of sample *ai* and *aj* (Yiakopoulos et al., 2011).

$$D_{cityblock,(ai,aj)} = \frac{1}{n}\sqrt{\sum_{k=1}^{n} |X_{ik} - X_{jk}|}$$

HCA shows an overview of all signal distances by the dendrogram. All pairs of signals are calculated in terms of distance. The pair of signals that obtained the shortest distance is joined to become the same group by using furthest neighbor linkage. Furthest neighbor uses the maximum distance between members in each group as the point of joining (Mao and Xu, 2006). After the first group was joined, the distances of the remaining signals were calculated again, and the shortest distance is joined to make the bigger group. These processes are continued until all signals are joined to form the same group. The length of the line that combines the samples together expresses the attitude of discrimination of members in the considered level. The longer the length of the combined line means the better clear clusters (Forina et al., 2007). Natural groups of the dataset are expressed when the dendrogram is cut at the longest combined line.

The benefit of HCA is that the user can cluster signals of samples into various groups, depending on various combined line levels (Brereton, 1992; Engels et al., 2006; Massart and Kauffman, 1992). Therefore, the number of clusters does not need to be fixed before HCA treatment. Appearance of distribution and clusters of signals can be seen clearly from the dendrogram. Outliers do not affect the dendrogram, because outliers can be separated into isolated groups. But HCA is suitable only for small numbers of signals, because the HCA algorithm needs time to generate the dendrogram and the dendrogram cannot express natural groups of datasets that obtained unclear clusters.

K-means clustering. *K*-means clustering is another method of cluster analysis, that aims to partition n observations into *k* clusters. Users can fix "*k*" number of clusters freely, when *k*-means clustering is applied. Running means an iterative process is used to find the optimized k centroids. The k centroids are iterated until the sum of squares of the within-cluster distances reach the minimum value (Krooshof et al., 2006). Distance of signals that obtained n variables of a_i to centroid a_j is obtained by using Euclidean distance (Yiakopoulos et al., 2011):

$$D_{\mathbf{Euclidean},(ai,aj)} = \frac{1}{n}\sqrt{\sum\nolimits_{k=1}^{n}(X_{ik}-X_{jk})^2}$$

Euclidean distance of the signals to all centroids is calculated and used as the parameter to identify groups of signals. The signal belongs to the group that obtains the shortest Euclidean distance of the signals to the centroid of the group. The benefit of *k*-means clustering is that the process is flexible. Group of signals can be changed until the maximum difference of all *k* groups is reached (Wu and Chow, 2004). The drawback of *k*-means clustering is that the accuracy of *k*-means clustering is dependent on the size of the data set and number of *k*. The result of *k*-means clustering is reliable only with very large data sets (more than 100 samples) and suitable k number are used (Hartigan, 1975) to express the distribution of all samples in the data as much as possible. This method is sensitive to outliers, so outliers should be removed from the dataset prior to using.

Linear discriminant analysis (LDA)

LDA is a technique that allows the classification of an individual into one of two or more distinctive populations on the basis of a set of measurements. It is a linear combination using the Fisher Transform function (Fisher, 1936;

Melody, 2003; Raghuraj and Lakshminarayanan, 2007; Timm, 2002). LDA assumes a multivariate normal distribution and that all groups have the same covariance matrix. Eigenvector and eigenvalues are iteratively calculated until the maximum ratio of "between groups' and within groups' variances" (Huang et al., 2009) (eigenvalue) is reached. Eigenvalue is one term to show the ability of the LDA model to discriminate data. Canonical function from Fisher's process is used to express weigh principal variables in the discrimination. Stepwise LDA was used to screen only the high impact variables from multivariate data of the discrimination (Huang et al., 2009). Wilks' lambda (λ) (Ren et al., 2006), ratio of variance of discriminant scores, is used as the criteria of a stepwise process to screen variables in LDA. The variable adds into the canonical function when λ is increased, otherwise the variable is rejected. Grouping of signals of samples was predicted by considering the square mahalanobis distance to centroid (Li et al., 1999; Siripatrawan and Harte, 2007; Tomasko et al., 1999).

$$D^2{}_{\mathrm{mahalanobis},c} = (X - \mu_c)S(X - \mu_c)^T$$

when *x* is signal of unknown, *μc* is centroid of group *c*, *T* is transform matrix and *S* is pooled variance-covariance matrix of training set.

RESULTS AND DISCUSSION

Four kinds of signals were employed: (a) OFT, (b) PCs of OFT, (c) overall slope and (d) interval slope. For the PCs of OFT, two principal components were extracted from OFT with the cumulative variance of overall data being 99.5% with percentages of variance of component 1 and 2 being 55 and 44.5, respectively.

Confirmation of group of signal of blood samples from hierarchical clustering analysis (HCA)

Group distributions of four kinds of signals from HCA were expressed by dendrogram (Figure 3). From the dendrogram, the signals were divided at the considered rescaled distance cluster combine values, and the distribution of signals in each cluster is highly dispersed. Two cluster signals for negative and positive tests were subjected for this investigation. Therefore, the dendrograms were cut at a level that signals can be clustered into two groups. The rescaled distance cluster combine of PCA scores of OFT was 24; whereas, rescaled distance cluster combines of other kinds of signals were 20. Numbers of members in the positive (HT) and negative test (HN) groups of OFT, PCA of OFT, overall slope, and the interval slope were (27,46), (38,35), (53,20) and (54,19), respectively. Clearly, the number of groups depends on the type of signal.

Confirmation of group of signal of blood samples from *k*-means clustering analysis

The initial values *k* = 2 were chosen for the *k*-means clustering algorithm.

K-means distance was a direct relation of the distribution of signals in the dataset. The distribution of signals in the negative test group (KN) was less than the positive test group (KT) for all kinds of OFT signals. The *k*-means clustering distance range of positive test and negative test of OFT, PCA score of OFT, overall slope and interval slope were (40.51-251.28, 21.98-272.21), (0.21-1.60, 0.18-2.28), (0.02-1.14,0.24-3.28) and (1.58-9.26, 3.79-16.25), respectively. The comparisons of *k*-means clustering and HCA of four kinds of OFT signals are presented in Tables 1-4. The numbers of members in the positive test group (KT) were usually less than that of the negative test group (KN), except in the case of the overall slope signal. From Table 5, the numbers of members in positive (KT) and negative test (KN) groups of OFT, PCA of OFT, overall slope and interval slope were (30,43), (29,44), (53,20) and (34,39), respectively. The "clustered" column indicated the correlation in identification of positive or negative group, according to the clustering method. The correlated result is labeled CT or CN; uncorrelated result is labeled CU. Similarly, the "concerned" column was the overall comparison of both clustering analysis techniques and the hospital records (report column).

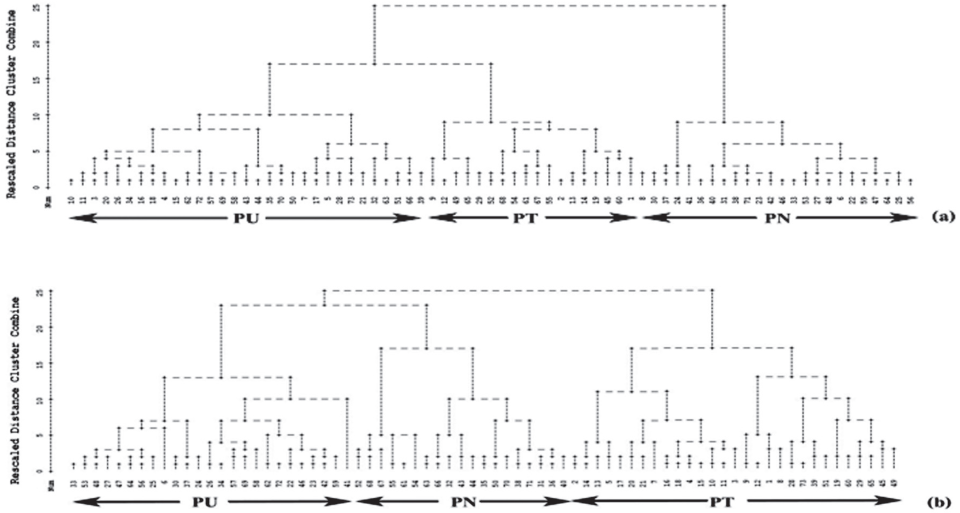

Figure 3. Dendrograms of blood samples: (a) OFT, (b) PCs of OFT, (c) overall slope, (d) interval slope.

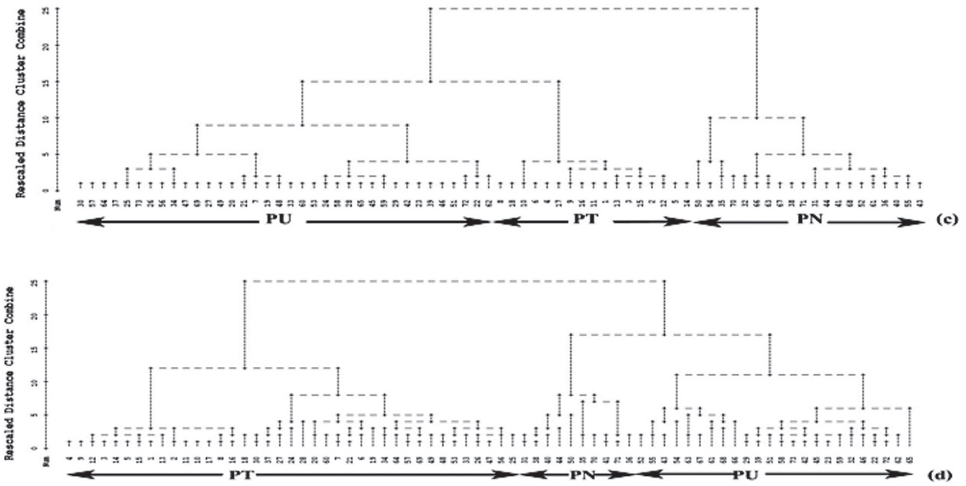

Figure 4. Dendrograms of blood samples: (a) OFT, (b) PCs of OFT, (c) overall slope, (d) interval slope.

Table 1. The clustering results of OFT signals by *k*-means clustering, HCA and hospital records.

Number	*K*-means	HCA	Report	Clustered	Concerned	Number	*K*-means	HCA	Report	Clustered	Concerned
01	KT	HT	RT	CT	PT	38	KN	HN	RN	CN	PN
02	KT	HT	RT	CT	PT	39	KT	HT	RN	*CT*	PU
03	KN	HN	RT	*CN*	PU	40	KN	HN	RN	CN	PN
04	KN	HN	RT	*CN*	PU	41	KN	HN	RN	CN	PN
05	KT	HT	RT	CT	PT	42	KN	HN	RN	CN	PN
06	KN	HN	RT	*CN*	PU	43	KN	HN	RN	CN	PN
07	KT	HN	RT	CU	PU	44	KN	HN	RN	CN	PN
08	KT	HT	RT	CT	PT	45	KT	HT	RN	*CT*	PU
09	KT	HT	RT	CT	PT	46	KN	HN	RN	CN	PN
10	KN	HN	RT	*CN*	PU	47	KN	HN	RN	CN	PN
11	KN	HN	RT	*CN*	PU	48	KN	HN	RN	CN	PN
12	KT	HT	RT	CT	PT	49	KT	HT	RN	*CT*	PU
13	KT	HT	RT	CT	PT	50	KN	HN	RN	CN	PN
14	KT	HT	RT	CT	PT	51	KT	HT	RN	*CT*	PU
15	KN	HN	RT	*CN*	PU	52	KT	HT	RN	*CT*	PU
16	KN	HN	RT	*CN*	PU	53	KN	HN	RN	CN	PN
17	KT	HN	RT	CU	PU	54	KT	HT	RN	*CT*	PU
18	KN	HN	RT	*CN*	PU	55	KT	HT	RN	*CT*	PU
19	KT	HT	RT	CT	PT	56	KN	HN	RN	CN	PN
20	KN	HN	RT	*CN*	PU	57	KN	HN	RN	CN	PN
21	KT	HN	RT	CU	PU	58	KN	HN	RN	CN	PN
22	KN	HN	RN	*CN*	PN	59	KN	HN	RN	CN	PN
23	KN	HN	RN	*CN*	PN	60	KT	HT	RN	*CT*	PU
24	KN	HN	RN	*CN*	PN	61	KT	HT	RN	*CT*	PU
25	KN	HN	RN	*CN*	PN	62	KN	HN	RN	CN	PN
26	KN	HN	RN	*CN*	PN	63	KT	HT	RN	*CT*	PU
27	KN	HN	RN	*CN*	PN	64	KN	HN	RN	CN	PN
28	KT	HT	RN	CT	PU	65	KT	HT	RN	*CT*	PU

Table 1. The clustering results of OFT signals by *k*-means clustering, HCA and hospital records (cont.).

Number	*K*-means	HCA	Report	Clustered	Concerned	Number	*K*-means	HCA	Report	Clustered	Concerned
29	KT	HT	RN	CT	PU	66	KT	HT	RN	*CT*	PU
30	KN	HN	RN	*CN*	PN	67	KT	HT	RN	*CT*	PU
31	KN	HN	RN	*CN*	PN	68	KT	HT	RN	*CT*	PU
32	KT	HT	RN	CT	PU	69	KN	HN	RN	CN	PN
33	KN	HN	RN	*CN*	PN	70	KN	HN	RN	CN	PN
34	KN	HN	RN	*CN*	PN	71	KN	HN	RN	CN	PN
35	KN	HN	RN	*CN*	PN	72	KN	HN	RN	CN	PN
36	KN	HN	RN	*CN*	PN	73	KT	HT	RN	*CT*	PU
37	KN	HN	RN	*CN*	PN						

Note: Italics indicates a result confirmed by *k*-means clustering and HCA, but that did not correlate with the hospital report. Underlining indicates a result confirmed by *k*-means clustering, HCA and hospital report.

Table 2. The clustering results of PCA scores of OFT signals by *k*-means clustering, HCA and hospital records.

Number	*K*-means	HCA	Report	Clustered	Concerned	Number	*K*-means	HCA	Report	Clustered	Concerned
01	KT	HT	RT	CT	PT	38	KN	HN	RN	CN	PN
02	KT	HT	RT	CT	PT	39	KT	HN	RN	CU	PU
03	KT	HT	RT	CT	PT	40	KN	HN	RN	CN	PN
04	KT	HT	RT	CT	PT	41	KN	HN	RN	CN	PN
05	KT	HT	RT	CT	PT	42	KN	HN	RN	CN	PN
06	KN	HT	RT	CU	PU	43	KN	HN	RN	CN	PN
07	KT	HT	RT	CT	PT	44	KN	HN	RN	CN	PN
08	KT	HT	RT	CT	PT	45	KT	HN	RN	CU	PU
09	KT	HT	RT	CT	PT	46	KN	HN	RN	CN	PN
10	KT	HT	RT	CT	PT	47	KN	HT	RN	CU	PU
11	KT	HT	RT	CT	PT	48	KN	HT	RN	CU	PU
12	KT	HT	RT	CT	PT	49	KT	HN	RN	CU	PU
13	KT	HT	RT	CT	PT	50	KN	HN	RN	CN	PN
14	KT	HT	RT	CT	PT	51	KN	HN	RN	CN	PN
15	KT	HT	RT	CT	PT	52	KN	HN	RN	CN	PN
16	KT	HT	RT	CT	PT	53	KN	HT	RN	CU	PU
17	KT	HT	RT	CT	PT	54	KN	HN	RN	CN	PN
18	KT	HT	RT	CT	PT	55	KN	HN	RN	CN	PN
19	KT	HT	RT	CT	PT	56	KN	HT	RN	CU	PU
20	KT	HT	RT	CT	PT	57	KN	HT	RN	CU	PU
21	KT	HT	RT	CT	PT	58	KN	HT	RN	CU	PU
22	KN	HN	RN	CN	PN	59	KN	HN	RN	CN	PN
23	KN	HN	RN	CN	PN	60	KT	HT	RN	*CT*	PU
24	KN	HT	RN	CU	PU	61	KN	HN	RN	CN	PN
25	KN	HT	RN	CU	PU	62	KN	HN	RN	CN	PN
26	KN	HT	RN	CU	PU	63	KN	HN	RN	CN	PN
27	KN	HT	RN	CU	PU	64	KN	HT	RN	CU	PU
28	KT	HN	RN	CU	PU	65	KT	HN	RN	CU	PU
29	KT	HN	RN	CU	PU	66	KN	HN	RN	CN	PN
30	KN	HT	RN	CU	PU	67	KN	HN	RN	CN	PN
31	KN	HN	RN	CN	PN	68	KN	HN	RN	CN	PN

Table 2. The clustering results of PCA scores of OFT signals by *k*-means clustering, HCA and hospital records (cont.).

Number	*K*-means	HCA	Report	Clustered	Concerned	Number	*K*-means	HCA	Report	Clustered	Concerned
32	KN	HN	RN	CN	PN	69	KT	HT	RN	*CT*	PU
33	KN	HT	RN	CU	PU	70	KN	HN	RN	CN	PN
34	KN	HT	RN	CU	PU	71	KN	HN	RN	CN	PN
35	KN	HN	RN	CN	PN	72	KN	HN	RN	CN	PN
36	KN	HN	RN	CN	PN	73	KT	HN	RN	CN	PN
37	KN	HT	RN	CU	PU						

Note: Italics indicates a result confirmed by *k*-means clustering and HCA, but that did not correlate with the hospital report. Underlining indicates a result confirmed by *k*-means clustering, HCA and hospital report.

Table 3. The clustering results of overall slope by *k*-means clustering, HCA and hospital records.

Number	*K*-means	HCA	Report	Clustered	Concerned	Number	*K*-means	HCA	Report	Clustered	Concerned
01	KT	HT	RT	CT	PT	38	KN	HN	RN	CN	PN
02	KT	HT	RT	CT	PT	39	KN	HT	RN	CU	PU
03	KT	HT	RT	CT	PT	40	KN	HN	RN	CN	PN
04	KT	HT	RT	CT	PT	41	KN	HN	RN	CN	PN
05	KT	HT	RT	CT	PT	42	KN	HT	RN	CU	PU
06	KT	HT	RT	CT	PT	43	KN	HN	RN	CN	PN
07	KT	HT	RT	CT	PT	44	KN	HN	RN	CN	PN
08	KT	HT	RT	CT	PT	45	KN	HT	RN	CU	PU
09	KT	HT	RT	CT	PT	46	KN	HT	RN	CU	PU
10	KT	HT	RT	CT	PT	47	KT	HT	RN	*CT*	PU
11	KT	HT	RT	CT	PT	48	KT	HT	RN	*CT*	PU
12	KT	HT	RT	CT	PT	49	KT	HT	RN	*CT*	PU
13	KT	HT	RT	CT	PT	50	KN	HN	RN	CN	PN
14	KT	HT	RT	CT	PT	51	KN	HT	RN	CU	PU
15	KT	HT	RT	CT	PT	52	KN	HN	RN	CN	PN
16	KT	HT	RT	CT	PT	53	KT	HT	RN	*CT*	PU
17	KT	HT	RT	CT	PT	54	KN	HN	RN	CN	PN
18	KT	HT	RT	CT	PT	55	KN	HN	RN	CN	PN
19	KT	HT	RT	CT	PT	56	KT	HT	RN	*CT*	PU
20	KT	HT	RT	CT	PT	57	KN	HT	RN	CU	PU
21	KT	HT	RT	CT	PT	58	KN	HT	RN	CU	PU
22	KN	HT	RN	CU	PU	59	KN	HT	RN	CU	PU
23	KN	HT	RN	CU	PU	60	KT	HT	RN	*CT*	PU
24	KN	HT	RN	CU	PU	61	KN	HN	RN	CN	PN
25	KT	HT	RN	*CT*	PU	62	KN	HT	RN	CU	PU
26	KT	HT	RN	*CT*	PU	63	KN	HN	RN	CN	PN
27	KT	HT	RN	*CT*	PU	64	KN	HT	RN	CU	PU
28	KN	HT	RN	CU	PU	65	KN	HT	RN	CU	PU
29	KN	HT	RN	CU	PU	66	KN	HN	RN	CN	PN
30	KN	HT	RN	CU	PU	67	KN	HN	RN	CN	PN
31	KN	HN	RN	CN	PN	68	KN	HN	RN	CN	PN
32	KN	HN	RN	CN	PN	69	KT	HT	RN	*CT*	PU
33	KT	HT	RN	*CT*	PU	70	KN	HN	RN	CN	PN
34	KT	HT	RN	*CT*	PU	71	KN	HN	RN	CN	PN
Number	*K*-means	HCA	Report	Clustered	Concerned	Number	*K*-means	HCA	Report	Clustered	Concerned

Table 3. The clustering results of overall slope by *k*-means clustering, HCA and hospital records (cont.).

35	KN	HN	RN	CN	PN	72	KN	HT	RN	CU	PU
36	KN	HN	RN	CN	PN	73	KN	HT	RN	CU	PU
37	KN	HT	RN	CU	PU	41	KN	HN	RN	CN	PN

Note: Italics indicates a result confirmed by *k*-means clustering and HCA, but that did not correlate with the hospital report. Underlining indicates a result confirmed by *k*-means clustering, HCA and hospital report.

Table 4. The clustering results of interval slope by *k*-means clustering, HCA and hospital records.

Number	*K*-means	HCA	Report	Clustered	Concerned	Number	*K*-means	HCA	Report	Clustered	Concerned
01	KT	HT	RT	CT	PT	38	KN	HN	RN	CN	PN
02	KT	HT	RT	CT	PT	39	KN	HT	RN	CU	PU
03	KT	HT	RT	CT	PT	40	KN	HN	RN	CN	PN
04	KT	HT	RT	CT	PT	41	KN	HN	RN	CN	PN
05	KT	HT	RT	CT	PT	42	KN	HT	RN	CU	PU
06	KT	HT	RT	CT	PT	43	KN	HN	RN	CN	PN
07	KT	HT	RT	CT	PT	44	KN	HN	RN	CN	PN
08	KT	HT	RT	CT	PT	45	KN	HT	RN	CU	PU
09	KT	HT	RT	CT	PT	46	KN	HT	RN	CU	PU
10	KT	HT	RT	CT	PT	47	KT	HT	RN	*CT*	PU
11	KT	HT	RT	CT	PT	48	KT	HT	RN	*CT*	PU
12	KT	HT	RT	CT	PT	49	KT	HT	RN	*CT*	PU
13	KT	HT	RT	CT	PT	50	KN	HN	RN	CN	PN
14	KT	HT	RT	CT	PT	51	KN	HT	RN	CU	PU
15	KT	HT	RT	CT	PT	52	KN	HN	RN	CN	PN
16	KT	HT	RT	CT	PT	53	KT	HT	RN	*CT*	PU
17	KT	HT	RT	CT	PT	54	KN	HN	RN	CN	PN
18	KT	HT	RT	CT	PT	55	KN	HN	RN	CN	PN
19	KT	HT	RT	CT	PT	56	KT	HT	RN	*CT*	PU
20	KT	HT	RT	CT	PT	57	KT	HT	RN	*CT*	PU
21	KT	HT	RT	CT	PT	58	KN	HT	RN	CU	PU
22	KN	HT	RN	CU	PU	59	KN	HT	RN	CU	PU
23	KN	HT	RN	CU	PU	60	KT	HT	RN	*CT*	PU
24	KN	HT	RN	CU	PU	61	KN	HN	RN	CN	PN
25	KN	HT	RN	CU	PU	62	KN	HT	RN	CU	PU
26	KT	HT	RN	*CT*	PU	63	KN	HN	RN	CN	PN
27	KT	HT	RN	*CT*	PU	64	KT	HT	RN	*CT*	PU
28	KN	HT	RN	CU	PU	65	KN	HT	RN	CU	PU
29	KN	HT	RN	CU	PU	66	KN	HN	RN	CN	PN
30	KN	HT	RN	CU	PU	67	KN	HN	RN	CN	PN
31	KN	HN	RN	CN	PN	68	KN	HN	RN	CN	PN
32	KN	HT	RN	CU	PU	69	KT	HT	RN	*CT*	PU
33	KT	HT	RN	*CT*	PU	70	KN	HN	RN	CN	PN
34	KT	HT	RN	*CT*	PU	71	KN	HN	RN	CN	PN
35	KN	HN	RN	CN	PN	72	KN	HT	RN	CU	PU
36	KN	HN	RN	CN	PN	73	KN	HT	RN	CU	PU
37	KN	HT	RN	CU	PU						

Note: Italics indicates a result confirmed by *k*-means clustering and HCA, but that did not correlate with the hospital report. Underlining indicates a result confirmed by *k*-means clustering, HCA and hospital report.

Confirmation of group of signals using clustering analysis techniques and hospital records

From Table 5, the comparison of clustering and hospital records obtained six predicted patterns. Case 1.1 and 1.5 were correctly identified for the normal (PN) and thalassemia (PT) patients. In other cases, we cannot classify for positive or negative test (PU), because results of *k*-means clustering and HCA were either uncorrelated with each other or with the hospital record. The results in case 1.2 and 1.4 were confirmed by two clustering methods, but classify wrongly according to the hospital records. Signals pattern in these two cases did not have enough selectivity to use in the diagnostic process. Only the correct results, as in cases 1.1 and 1.5, from two clustering analysis techniques and hospital record were used for LDA construction.

Table 5. Confirmation of concerned group by the comparison of clustered results and hospital records.

Considering pattern	Clustering group	Record	Concerned group	OFT	PCA score of OFT	Overall slope	Interval slope
1.1	CN	RN	PN	34(65%)	29(56%)	20(38%)	19(37%)
1.2	CT	RN	PU	18	2	12	33
1.3	CU	RN	PU	0	21	20	0
1.4	CN	RT	PU	9	0	0	0
1.5	CT	RT	PT	9 (43%)	20(95%)	21(100%)	21(100%)
1.6	CU	RT	PU	3	1	0	0

Note: According to the hospital records, there were 21 and 52 samples with positive and negative tests, respectively. The percentage of the correlated classification according to the hospital record are in parenthesis.

The comparison of clustering and hospital records of OFT showed that the correlated results of the positive (PT) and negative tests (PN) were (9/21) 43% and (34/52) 65%, respectively. This shows that OFT is not suitable for thalassemia screening. With the PCA score of OFT, the correlated classification percentage is higher –95% (PT) and 56% (PN), respectively. The OFT signals that can be used for discriminating positive (PT) and negative tests (PN) were 20 and 29 signals, respectively. From the overall and interval slopes, the perfect method to identify thalassemia patients was identified. The correlated results of positive and negative test were 100% and 38% with overall slope and 100% and 37% with the interval slope, respectively. Distributions of signals, which can be visualized easily with PCs score, are shown in Figure 4 as positive test (PT), negative test (PN) and unidentified case (PU). The distribution of PN and PT were clearly separated and suitable for LDA construction. Groups of positive tests of two types of slope were clustered, but for the negative tests a wide distribution was observed.

Table 6. *K*-means distance of signals in PT and PN of various kinds of OFT.

Signal type	K-means distance	
	PT	PN
OFT	67.379-251.277	21.978-272.209
PCA score of OFT	0.211-1.439	0.177-2.278
Overall slope	0.017-1.143	0.241-3.279
Interval slope	1.575-4.852	3.792-16.247

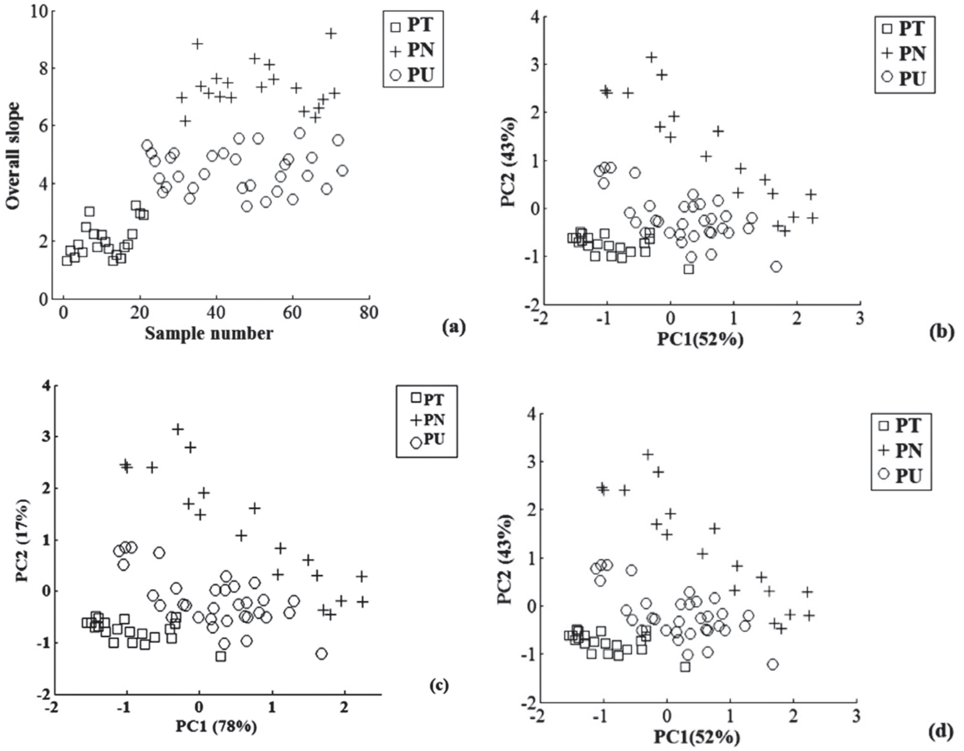

Figure 5. Distribution of various kinds of OFT of 73 samples labeled by concerned group: (a) OFT, (b) PCs score of OFT, (c) PCs of overall slope and (d) PCs of interval slope.

Thalassemia screening by LDA

The selection of positive and negative tests for use as training sets in LDA was considered from k-distance of their concerned groups. *K*-distances of signals in PT and PN were arranged in descending order. Ranges of k-distance of PT and PN of each kind of OFT are shown in Table 6. The training set of LDA contained samples that maximize distribution of *k*-distance. Number of training samples of LDA for each kind of OFT signals model was varied as 10, 16 and 20 to improve the prediction ability. LDA models were named using name of type of signal, numbers of positive and negative test as "type of signal_number of positive test_number of negative test". Symbols used in LDA model to express type of OFT signals were OFT, PCOFT, OS and IS for normalized OFT, PCA score of normalized OFT, overall slope of normalized OFT and interval slope of normalized OFT, respectively. Characteristics of LDA models are shown in Table 7.

Eigenvalue expressed ability of discrimination. The LDA of OFT, PCA score of OFT and overall slope have no significant difference (Table 7). LDA of interval slope models are better predictors than other LDA models. Stepwise LDA was performed for screened variables that contained a high impact of discrimination when OFT and interval slope were used. High impact variable in canonical functions of OFT was OFT at the time of 26th second (t26) as indicated as a variable in the function of OFT_5_5, OFT_5_8, OFT_5_10 and the canonical functions of interval slope were in the function of IS_11_9, IS_7_9, IS_9_7, IS_10_6, IS_11_5, IS_5_5, IS_8_8 and IS3_10_10 with variables of IS4, IS5, IS6, IS7, IS8 and IS11.

Canonical coefficients of all LDA models were more than 0.9. The LDA models could be used to predict groups of unknown signals very well. Eigenvalues of canonical discriminant functions of various numbers of training samples of OFT, PCA score of OFT and overall slope have no significant difference.

The validation samples consisted of three positive and six negative tests. The validation results are presented in Tables 8-9. The correctly classified cases of validation set predictions were increased when the numbers of training samples of LDA of PCA score of OFT and interval slope were increased. The correctly classified validations of LDA of overall slope and OFT when using 10, 16 and 20 training samples were 100%. The ratio of positive and negative tests in the training sets affected the reliability of LDA models.

Table 7. Canonical discriminant function.

Function	Canonical equation	Canonical Correlation	Eigenvalue
OS_5_5	OS_5_5 Function = 1.107*slope - 5.268	0.963	12.692
OS_8_8	OS_8_8 Function = 1.159*slope - 5.531	0.956	10.749
OS_10_10	OS_10_10 Function = 1.183*slope -5.65	0.956	10.742
OFT_5_5	OFT_5_5 Function = 0.038*t26-23.103	0.945	8.373
OFT_5_8	OFT_5_8Function = 0.044*t26-27.66	0.956	10.505
OFT_5_10	OFT_5_10 Function = 0.042*t26-26.388	0.943	8.044
IS_11_9	IS_11_9 Function =0.967*IS6-5.176	0.961	12.037
IS_7_9	IS_7_9 Function =0.375*IS5+0.936*IS7-7.562	0.977	20.707
IS_9_7	IS_9_7 Function =1.025*IS6-5.289	0.966	14.153
IS_10_6	IS_10_6 Function =1.024*IS6-5.120	0.967	14.348
IS_11_5	IS_11_5 Function =2.766*IS8-1.768*IS11-5.249	0.975	19.225
IS_5_5	IS_5_5 Function = 0.312*IS4+1.275*IS8-7.209	0.981	26.127
IS_8_8	IS_8_8 Function = 1.083*IS4-1.244*IS5+2.007*IS7-8.802	0.987	38.426
IS_10_10	IS_10_10 Function = 0.36*IS4+1.033*IS7-7.238	0.980	24.083
PCOFT_5_5	PCOFT_5_5 Function=2.188*PC1-1.872*PC2	0.945	8.326

Table 7. Canonical discriminant function (cont.).

Function	Canonical equation	Canonical Correlation	Eigenvalue
PCOFT_8_8	PCOFT_8_8 Function=2.191*PC1-2.131*PC2	0.949	9.010
PCOFT_10_10	PCOFT_10_10 Function=2.284*PC1-1.858*PC2	0.944	8.150
PCOFT_3_7	PCOFT_3_7 Function=1.771*PC1-1.041*PC2	0.888	3.749
PCOFT_4_6	PCOFT_4_6 Function=2.540*PC1-1.675*PC2	0.951	9.413
PCOFT_6_4	PCOFT_6_4 Function=2.325*PC1-1.866*PC2	0.949	8.996
PCOFT_7_3	PCOFT_7_3 Function=2.321*PC1-1.473*PC2	0.939	7.499
PCOFT_3_13	PCOFT_3_13 Function=1.567*PC1-0.826*PC2	0.838	2.362
PCOFT_4_12	PCOFT_4_12 Function=1.735*PC1-1.296*PC2	0.895	4.024
PCOFT_5_11	PCOFT_5_11 Function=1.828*PC1-1.257*PC2	0.900	4.274
PCOFT_6_10	PCOFT_6_10 Function=2.095*PC1-1.365*PC2	0.922	5.696
PCOFT_7_9	PCOFT_7_9 Function=-1.830*PC1+1.913*PC2	0.931	6.513
PCOFT_9_7	PCOFT_9_7 Function=2.332*PC1-1.964*PC2	0.948	8.955
PCOFT_10_6	PCOFT_10_6 Function=1.937*PC1-1.897*PC2	0.934	6.878
PCOFT_11_5	PCOFT_11_5 Function=-1.760*PC1+1.920*PC2	0.929	6.270
PCOFT_12_4	PCOFT_12_4 Function=-2.059*PC1+2.077*PC2	0.944	8.169

Note: The code “x_y1_y2” for referring x: type of signal, y1: number of positive cases, y2: number of negative cases.

All LDA of PCA scores of OFT models can predict groups of negative tests in validation sets as 100%. The ability of positive test prediction of LDA of PCA score of OFT model was increased when the number of training samples of the model was increased.

The validation results of LDA of interval slope models had different trends. When training samples of LDA models were increased, the model reliability decreased. The boundary of LDA of interval slope models cannot cover the prediction of the validation set because of the limitation of the training set number.

The abilities of validation prediction of LDA of PCA score of OFT models were different when the number and ratio of positive and negative tests of training sets were verified. The LDA of PCA score of OFT models can predict groups of validation as 100% with SMDC of positive and negative less than 1.33 and 2.36, respectively. In this study, the model PCOFT_7_3, PCOFT_6_10, PCOFT_7_9, PCOFT_12_4,PCOFT_10_6 and PCOFT_9_7 can predict the validation set perfectly with the upper bound of SMDC of positive and negative tests as (1.44,2.36), (1.33,3.69), (1.49,3.09), (1.58,3.45), (1.74,3.58) and (2.60,2.80), respectively. Sample number 12 and 40 were predicted as unidentified cases by PCOFT_5_5, PCOFT_8_8 and PCOFT_10_10.

The increase of SMDC range for the predicted group of positive and negative tests on OS_5_5, OS_8_8 and OS_10_10 were (0.44,2.99), (1.67,4.27) and (1.67,4.49), respectively. This was due to the increase of training samples of LDA of overall slope models (10, 16 and 20 samples).

Table 8. Validation results of the LDA models.

Number / Function	2 (RT)		5 (RT)		12 (RT)		31 (RN)		38 (RN)		40 (RN)		43 (RN)		44 (RN)		71 (RN)	
	I	II	I	II	I	II	I	II	I	II	I	II	I	II	I	II	I	II
OFT_5_5	AT	0.71	AT	1.15	AT	1.37	AN	0.52	AN	0.00	AN	0.47	AN	0.01	AN	0.28	AN	0.04
OFT_5_8	AT	0.97	AT	1.56	AT	1.86	AN	0.55	AN	0.01	AN	0.49	AN	0.05	AN	0.52	AN	0.11
OFT_5_10	AT	0.87	AT	1.40	AT	1.67	AN	0.83	AN	0.01	AN	0.76	AN	0.00	AN	0.22	AN	0.01
PCOFT_5_5	AT	1.08	AT	1.40	AU	1.87	AN	0.36	AN	0.35	AN	1.52	AN	1.32	AN	0.06	AN	0.11
PCOFT_8_8	AT	1.20	AT	1.51	AU	3.12	AN	0.95	AN	0.73	AN	2.58	AN	2.19	AN	0.20	AN	0.25
PCOFT_10_10	AT	0.62	AT	0.80	AT	2.11	AN	1.34	AN	1.06	AU	3.07	AN	2.47	AN	0.37	AN	0.51
PCOFT_3_7	AT	0.52	AT	0.61	AU	1.04	AN	0.12	AN	0.10	AN	0.62	AN	0.34	AN	0.00	AN	0.04
PCOFT_4_6	AT	0.59	AT	0.75	AU	1.82	AN	0.94	AN	0.75	AU	2.52	AN	1.83	AN	0.18	AN	0.37
PCOFT_6_4	AT	0.59	AT	0.87	AU	1.78	AN	0.08	AN	0.04	AN	0.66	AN	0.83	AN	0.00	AN	0.05
PCOFT_7_3	AT	0.51	AT	0.69	AT	0.92	AN	0.11	AN	0.12	AN	0.91	AN	0.62	AN	0.00	AN	0.03
PCOFT_3_13	AT	0.52	AT	0.56	AU	1.46	AN	0.43	AN	0.30	AN	1.12	AN	0.58	AN	0.03	AN	0.18
PCOFT_4_12	AT	0.32	AT	0.41	AU	0.94	AN	0.28	AN	0.21	AN	0.98	AN	0.68	AN	0.02	AN	0.07
PCOFT_5_11	AT	0.41	AT	0.49	AU	1.14	AN	0.29	AN	0.21	AN	0.96	AN	0.58	AN	0.01	AN	0.08
PCOFT_6_10	AT	0.34	AT	0.42	AT	1.24	AN	0.65	AN	0.49	AN	1.68	AN	1.17	AN	0.11	AN	0.23
PCOFT_7_9	AT	0.48	AT	0.69	AT	1.15	AN	0.14	AN	0.11	AN	0.79	AN	0.82	AN	0.01	AN	0.00
PCOFT_9_7	AT	0.93	AT	1.23	AT	2.06	AN	0.49	AN	0.43	AN	1.83	AN	1.61	AN	0.08	AN	0.11
PCOFT_10_6	AT	0.61	AT	0.84	AT	1.33	AN	0.15	AN	0.13	AN	0.89	AN	0.82	AN	0.00	AN	0.00
PCOFT_11_5	AT	0.79	AT	1.01	AT	2.48	AN	0.92	AN	0.67	AU	2.23	AN	1.94	AN	0.21	AN	0.22
PCOFT_12_4	AT	0.44	AT	0.69	AT	1.09	AN	0.01	AN	0.01	AN	0.53	AN	0.52	AN	0.06	AN	0.06
OS_5_5	AT	0.05	AT	0.10	AT	0.03	AN	0.56	AN	0.33	AN	0.00	AN	0.03	AN	0.56	AN	0.34
OS_8_8	AT	0.27	AT	0.37	AT	0.22	AN	0.28	AN	0.12	AN	0.07	AN	0.01	AN	0.28	AN	0.13
OS_10_10	AT	0.31	AT	0.42	AT	0.26	AN	0.28	AN	0.11	AN	0.08	AN	0.01	AN	0.28	AN	0.12

Table 8. Validation results of the LDA models. (cont.)

Number / Function	2 (RT)		5 (RT)		12 (RT)		31 (RN)		38 (RN)		40 (RN)		43 (RN)		44 (RN)		71 (RN)	
	I	II	I	II	I	II	I	II	I	II	I	II	I	II	I	II	I	II
IS_11_9	AT	0.71	AT	0.71	AT	0.42	AN	2.14	AN	1.16	AN	0.47	AN	0.09	AN	2.14	AN	2.14
IS_7_9	AT	1.62	AT	0.62	AT	0.01	AN	0.36	AN	1.35	AN	0.17	AN	0.36	AN	1.26	AU	3.23
IS_9_7	AT	0.56	AT	0.56	AT	0.30	AN	2.23	AN	1.17	AN	0.45	AN	0.07	AN	2.23	AN	2.23
IS_10_6	AT	3.65	AT	0.89	AT	0.89	AU	0.54	AN	2.26	AN	1.19	AN	0.47	AU	3.65	AU	3.65
IS_11_5	AT	0.07	AU	10.89	AT	0.34	AN	0.00	AN	0.02	AN	0.74	AN	0.03	AN	0.01	AU	4.23
IS_5_5	AT	0.74	AT	1.41	AT	0.23	AN	0.04	AN	0.00	AN	0.00	AN	1.99	AN	0.00	AN	1.97
IS_8_8	AU	10.68	AT	1.21	AT	0.48	AN	0.04	AN	0.62	AN	4.72	AU	11.03	AN	0.18	AN	0.54
IS_10_10	AU	2.13	AT	0.02	AT	0.04	AN	0.05	AN	0.59	AN	0.78	AN	1.15	AN	0.53	AN	0.69

Note: Number and type of samples used in validation are indicated in the column header. I: Gpred is predicted group of sample from LDA prediction. II: Square Mahalanobis distance of sample to centroid. The code "x_y1_y2" for referring x: type of signal, y1: number of positive cases, y2: number of negative cases. AT and AN are predicted by LDA as positive and negative cases, respectively.

Table 9. SMDC boundary of LDA models and percentage of correctly classified from validation results of nine test cases.

Function	Training SMDC		Correctly classified%	
LDA output	Positive	Negative	Positive	Negative
OFT_5_5	2.11	1.98	100	100
OFT_5_8	2.86	3.02	100	100
OFT_5_10	2.57	2.76	100	100
PCOFT_5_5	1.45	2.70	67	100
PCOFT_8_8	2.58	2.75	67	100
PCOFT_10_10	2.28	2.75	100	83
PCOFT_3_7	0.85	3.47	67	100
PCOFT_4_6	1.44	2.20	67	83
PCOFT_6_4	2.26	1.57	67	100
PCOFT_7_3	1.44	2.36	100	100
PCOFT_3_13	1.25	4.00	67	100
PCOFT_4_12	0.73	4.47	67	100
PCOFT_5_11	0.93	4.24	67	100
PCOFT_6_10	1.33	3.69	100	100
PCOFT_7_9	1.49	3.09	100	100
PCOFT_9_7	2.60	2.80	67	100
PCOFT_10_6	1.74	3.58	100	100
PCOFT_11_5	2.56	2.16	100	83
PCOFT_12_4	1.58	3.45	100	100
OS_5_5	0.44	2.99	100	100
OS_8_8	1.67	4.27	100	100
OS_10_10	1.67	4.49	100	100
IS_11_9	1.08	3.34	100	100
IS_7_9	1.72	2.14	100	83
IS_9_7	1.19	3.96	100	100
IS_10_6	1.83	2.47	100	50
IS_11_5	2.84	2.32	67	83
IS_5_5	3.04	2.11	100	100
IS_8_8	1.35	4.93	67	83
IS_10_10	2.09	3.55	67	100

The stepwise process of LDA of interval slope models affect the containing of impact variables in canonical function. It can predict the validation set as 100% for IS_11_9, IS_9_7 and IS_5_5 with upper bound of positive and negative tests of (1.08,3.34), (1.19,3.96) and (3.04,2.11), respectively. Furthermore, when the different numbers of positive and negative tests in the training set is high, the quality of LDA of the interval slope model was worse.

The number of training samples in the LDA model and upper bound of SMDC value were used as criteria to select the model of each model. The LDA model from the smallest number of training samples with the highest upper bound of SMDC value was selected to predict the group of unknown signal. The selected LDA of OFT, PCA of OFT, overall slope and interval slope were OFT_5_8, PCOFT_9_7, OS_8_8 and IS_5_5 with the upper bound of positive and negative tests as (2.86, 3.02), (2.60, 2.80), (1.67, 4.27) and (3.04, 2.11), respectively.

Table 10. Unknown prediction results of the selected models.

Number	Record	OFT_5_8	PCOFT_9_7	OS_8_8	IS_5_5	Number	Record	OFT_5_8	PCOFT_9_7	OS_8_8	IS_5_5
1	RT	~~AT~~	~~AT~~	~~AT~~	~~AT~~	37	RN	AN	AU	AU	AU
3	RT	AU	AT	AT	~~AT~~	39	RN	AU	AU	AU	AU
4	RT	AU	~~AT~~	AT	~~AT~~	41	RN	AN	~~AN~~	AN	~~AN~~
6	RT	*AN*	AT	~~AT~~	~~AT~~	42	RN	AN	AU	AU	AU
7	RT	AU	~~AT~~	AT	AU	45	RN	**AU**	AU	AU	AU
8	RT	~~AT~~	AT	AT	AT	46	RN	~~AN~~	AN	AU	AN
9	RT	~~AT~~	~~AT~~	AT	AT	47	RN	AN	AU	AU	AU
10	RT	AU	~~AT~~	~~AT~~	AT	48	RN	AN	AU	*AT*	AU
11	RT	AU	AT	AT	AT	49	RN	*AT*	AU	AU	AU
13	RT	AT	AT	AT	AT	50	RN	AN	AU	~~AN~~	AU
14	RT	~~AT~~	AT	~~AT~~	AT	51	RN	**AU**	AN	AU	AU
15	RT	AU	AT	~~AT~~	AT	52	RN	**AU**	AN	AN	AU
16	RT	AU	AT	AT	AT	53	RN	~~AN~~	AU	AU	AU
17	RT	AU	AT	~~AT~~	~~AT~~	54	RN	AU	~~AN~~	AN	AU
18	RT	AU	~~AT~~	AT	AT	55	RN	AU	AN	~~AN~~	AN
19	RT	~~AT~~	~~AT~~	~~AT~~	AU	56	RN	~~AN~~	AU	AU	AU
20	RT	*AU*	~~AT~~	~~AT~~	AU	57	RN	AN	AU	AU	AU
21	RT	AU	~~AT~~	AT	AU	58	RN	~~AN~~	AU	AU	AU
22	RN	AN	~~AN~~	AU	AU	59	RN	AN	AU	AU	AU
23	RN	~~AN~~	AU	AU	AU	60	RN	*AT*	AU	AU	AU
24	RN	~~AN~~	AU	AU	AU	61	RN	AU	~~AN~~	~~AN~~	AU
25	RN	AN	AU	AU	AU	62	RN	AN	AN	AN	AU
26	RN	AN	AU	AU	AU	63	RN	AN	~~AN~~	~~AN~~	~~AN~~
27	RN	AN	AU	AU	AU	64	RN	AN	AU	AU	AU
28	RN	AU	AU	AU	AU	65	RN	*AT*	AU	AU	AU
29	RN	AT	AU	AU	AU	66	RN	AU	AN	AN	~~AN~~
30	RN	~~AN~~	AU	AU	AU	67	RN	AU	AN	AN	~~AN~~
32	RN	AN	AN	~~AN~~	AN	68	RN	AU	~~AN~~	~~AN~~	AN
33	RN	AN	AU	AU	AU	69	RN	AU	AU	AU	AU
34	RN	AN	AU	AU	AU	70	RN	AN	AU	~~AN~~	~~AN~~
35	RN	AN	AU	AN	AU	72	RN	AN	~~AN~~	AU	AU
36	RN	~~AN~~	AN	~~AN~~	AN	73	RN	AU	AU	AU	AU

Note: Crossed-out data were not used in the unknown set for each selected model. Italics indicate a result predicted by LDA, but that did not correlate with the hospital report.

Unknown prediction using selected LDA models

Unknown predictions using selected LDA models are shown in Table 10. Unknown predictions fall within three possible groups: i) within the upper bound of SMDC of training of LDA and correlated with the hospital record; ii) outside the upper bound of SMDC of training of LDA and defined as AU and

iii) within the upper bound of SMDC of training of LDA, but uncorrelated with the hospital record. In this study, all four selected models correctly predict most unknown samples. The uncorrelated predicted results were found when OFT_5_8 and OS_8_8 were used. Unknown samples number 6, 29, 49, 60 and 65 were predicted uncorrelated with hospital records by using OFT_5_8 and unknown sample number 48 was predicted uncorrelated with hospital record by using OS_8_8, so those unknown samples should be rechecked by other methods.

The LDA of PCA score of OFT and interval slope can predict all unknowns correlated with hospital records. In this study, model IS_5_5 was the best model for thalassemia screening due to the high upper bound of SMDC of the training set, since using a small training set of data would yield similar results to a larger number of training set data.

CONCLUSION

Unsupervised learning approaches, *k*-means clustering and HCA coupled with LDA, show the possibility of chemometric techniques to differentiate between thalassemia patients and normal persons by inspecting the signal from osmotic fragility test coupled with stopped flow injection. Various types of signal (OFT, PCs of OFT, overall slope and interval slope) were investigated and only the correlated results from both techniques and the hospital records were used to train LDA models. As a result, LDA models of principal component score of OFT and interval slope can predict correctly for all unknowns correlating with the hospital records. The final model using the PCA score of individual slope (IS_5_5) was the best model for thalassemia screening due to the high upper bound of SMDC of the training set.

ACKNOWLEDGEMENTS

We acknowledge the Development and Promotion of Science and Technology Talents Project (DPST), the Thailand Research Fund (TRF) and the Center of Excellence in Innovation for Analytical Science and Technology (I-ANALY-S-T) for their support.

REFERENCES

Brereton, R.G. 1992. Multivariate Pattern Recognition in Chemometrics Illustrated by Case Studies. Elsevier, Netherlands.

Chatfield, C., and A.J. Collins. 1980. Introduction to Multivariate Analysis. Chapman & Hall, London.

De Carvalho, A.R., M.d.N. Sánchez, J. Wattoom, and R.G. Brereton. 2006. Comparison of PLS and kinetic models for a second-order reaction as monitored using ultraviolet visible and mid-infrared spectroscopy. Talanta 68:1190-1200.

Engels, M.F.M., A.C. Gibbs, E.P. Jaeger, D. Verbinnen, V.S. Lobanov, and D.K. Agrafiotis. 2006. A Cluster-Based Strategy for Assessing the Overlap between Large Chemical Libraries and Its Application to a Recent Acquisition. J. Chem. Inf. Model. 46:2651-2660.

Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems. A Eug. 7:179-188.

Forina, M., S. Lanteri, M. Casale, and M.C.C. Oliveros. 2007. A new algorithm for seriation and its use in similarity dendrograms. Chemom. Intell. Lab. Syst. 87:262-274.

Hartigan, J. 1975. Clustering Algorithms. John Wiley & Sons, New York.

Hartwell, S.K., B. Srisawang, P. Kongtawelert, D. Christian, and K. Grudpan. 2005. Review on screening and analysis techniques for hemoglobin variants and thalassemia. Talanta 65:1149-1161.

Huang, J.Y., Y.B. Qiu, and X.P. Guo. 2009. Cluster and discriminant analysis of electrochemical noise statistical parameters. Electrochim. Acta 54:2218-2223.

Jurado, J.M., A. Alcázar, F. Pablos, M.J. Martín, and A.G. González. 2005. Classification of aniseed drinks by means of cluster, linear discriminant analysis and soft independent modelling of class analogy based on their Zn, B, Fe, Mg, Ca, Na and Si content. Talanta 66:1350-1354.

Khonyoung, S., S. Kradtap Hartwell, J. Jakmunee, S. Lapanantnoppakhun, T. Sanguansermsri, and K. Grudpan. 2009. A stopped flow system with hydrodynamic injection for red blood cells osmotic fragility test: possibility for automatic screening of beta-thalassemia trait. Anal. Sci. 25:819-824.

Krooshof, P.W.T., G.J. Postma, W.J. Melssen, L.M.C. Buydens, and T.N. Tran. 2006. Effects of including spatial information in clustering multivariate image data. Trac-trend. Anal. Chem. 25:1067-1080.

Li, Y., J-H. Jiang, Z-P. Chen, C-J. Xu, and R-Q. Yu. 1999. Robust linear discriminant analysis for chemical pattern recognition. J. Chemometr. 13:3-13.

Mao, J., and J. Xu. 2006. Discrimination of herbal medicines by molecular spectroscopy and chemical pattern recognition. Spectrochim. Acta Part A 65:497-500.

Massart, D.L., and L. Kauffman. 1992. Interpretation of Analytical Data by Use of Cluster Analysis. John Wiley & Sons, New York.

Melody, Y. K. 2003. A comparative assessment of classification methods. Decis. Support Syst. 35:441-454.

Milde, D., J. Machácek, and V. Stuzka. 2007. Evaluation of colon cancer elements contents in serum using statistical methods. Chem. Pap. 61:348-352.

Pham, D.L. 2001. Spatial Models for Fuzzy Clustering. Comput. Vis. Image Understand. 84:285-297.

Raghuraj Rao, K., and S. Lakshminarayanan. 2007. Partial correlation based variable selection approach for multivariate data classification methods. Chemom. Intell. Lab. Syst. 86:68-81.

Ren, Y., H. Liu, C. Xue, X. Yao, M. Liu, and B. Fan. 2006. Classification study of skin sensitizers based on support vector machine and linear discriminant analysis. Anal. Chim. Acta 572:272-282.

Siripatrawan, U., and B.R. Harte. 2007. Solid phase microextraction/gas chromatography/mass spectrometry integrated with chemometrics for detection of Salmonella typhimurium contamination in a packaged fresh vegeTable. Anal. Chim. Acta 581:63-70.

Smolinski, A., B. Walczak, and J.W. Einax. 2002. Hierarchical clustering extended with visual complements of environmental data set. Chemom. Intell. Lab. Syst. 64:45-54.

Teppola, P., S-P. Mujunen, and P. Minkkinen. 1999. Adaptive Fuzzy C-Means clustering in process monitoring. Chemom. Intell. Lab. Syst. 45:23-38.

Thanasoulias, N.C., N.A. Parisis, and N.P. Evmiridis. 2003. Multivariate chemometrics for the forensic discrimination of blue ball-point pen inks based on their Vis spectra. Forensic Sci. Int. 138:75-84.

Timm, N.H. 2002. Applied Multivariate Analysis. Springer, New York.

Tomasko, L., R.W. Helms, and S.M. Snapinn. 1999. A discriminant analysis extension to mixed models. Statist. Med. 18:1249–1260.

Viana, M., X. Querol, A. Alastuey, J.I. Gil, and M. Menéndez. 2006. Identification of PM sources by principal component analysis (PCA) coupled with wind direction data. Chemosphere 65:2411-2418.

Wu, S., and T.W.S. Chow. 2004. Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. Pattern. Recog. 37:175-188.

Yiakopoulos, C.T., K.C. Gryllias, and I.A. Antoniadis. 2011. Rolling element bearing fault detection in industrial environments based on a K-means clustering approach. Expert Syst. Appl. 38:2888-2911.