CHAPTER 3
COMBINATION OF HCA, *K*-MEANS CLUSTERING AND LDA FOR EVALUATION OF SIGNALS FROM OFT-ANALYZER TO THE PREDICTION OF PATIENT GROUP OF THALASSEMIA SCREENING

3.1 Introduction

Identification of group of samples in dataset can be performed by 2 methodologies depended on priori of data [1]; unsupervised and supervised pattern recognition.

Unsupervised pattern recognition is clustering method which performed without priori knowledge about class membership, data points are grouped into same group because of the similarity of their feathers. Among such methods are cluster analysis, principle component analysis, correspondence analysis, projection pursuit, Kohonen networks, adaptive resonance theory (ART) and the eigenvector method [2, 3] .

Popular clustering methods such as *k*-means clustering and hierarchical clustering analysis were used in many fields. *K*-means clustering is technique to group data [4]. *K*-means clustering is suitable for sample size larger than 100 data [5]. Awareness of *k*-means clustering are the proper k value which needed to assign from user have to relate with natural group of the considered data and started point of cluster centers had affect to quality of *k*-means clustering [6].

Hierarchical cluster analysis (HCA) dendrogram provide the detailed of relationship between the cases in data [7-10]. HCA expressed distribution of

samples in data in picture of dendrogram. In various position of the combined distance in each level of dengrogram can expressed structure of samples when considered numbers of groups were varied. This is a big good point of HCA. Furthermore, outliers can cluster to the separated group from all samples and they do not affect to the rest groups so it can said that HCA do not affect from outliers. Drawback of HCA is its algorithm takes much time to perform dendrogram so HCA is not proper for large dataset [4, 11].

Principal component analysis (PCA) is tool to extract condensed information of multivariate pattern in data. PCA involves forming a new format of dataset called the principal components (PCs), which are the eigenvectors of the covariance or correlation matrix of the raw data. The first principal component (PC1) engaged extracting, group of data from raw data, the second principal component (PC2), orthogonal to the first one, has been the extracted information from the remain data after PC1 is extracted, and so on. Eigenvalue is amount of relative variance which correlating the considered principal component. Scatter plot present distribution of data in forms of principal componentss [12, 13].

Supervised pattern recognition needs training set of known class membership to set up classification model for use to predict group of data of unknown samples [5, 14-16]. Evaluation of classification model is performed by the use of a validation set comparing predictions with true group [17].

Linear discriminant analysis is one of supervised pattern recognition which widely used in many fields such as food industry [17-35], pharmacy [36-40], forensic science [41], Petroleum [42] and bio-species [43,

44].Furthermore, LDA method was found to use in many clinical studies. LDA were used to screening of many kinds of disorders such as sleep apnoea [45], diabetes [46, 47], asthma [47, 48], cancer [49-52] and thalassemia [53, 54].

A stopped flow system with hydrodynamic injection for red blood cells was developed to generate automatic osmotic fragility test (OFT) analyzer [55]. In [55], slope of OFT was used as variable to discriminate 73 samples in the dataset .

In this study, the chemometrics methods were employed for thalassemia screening. LDA models were performed for this aim. Various types of signals were used to cluster thalassemia patients.

3.2 Chemometrics methodology

3.2.1 Extraction and distribution of data: Principal component analysis (PCA)

Principal component analysis (PCA) is used to extract OFT signals. Approximation of OFT data is performed by PCA and digested into 3 matrices; score, T and loading, P and E, residue error, respectively.

$$X=TP+E$$

Score metric T is principle component which express characteristic of each samples and loading metric P is new variables which extracted by PCA[56]. In this work, the OFT data was extracted by using correlation metric and varimax rotation method, the component score metric was extracted by regression method, only principle components which have eigenvalues

greater than 1 were selected. Score metric can be used in 2 aims that was used as variable for the clustering and was plotted to expressed the distribution of signal of samples.

3.2.2 Clustering using Hierarchical Cluster Analysis (HCA) and *k*-means clustering

3.2.2.1 Hierarchical cluster analysis (HCA)

Agglomerative hierarchical cluster analysis is unsupervised pattern recognition which applied to cluster OFT signals of blood samples. Dissimilarity of any 2 cases was explained by using distance. The distance measure of samples x and y at variable 1-p was calculated based on city block and shown in equation 3.1 [57]:

$$\text{dCity Block}=\sum_{i-1}^{p}|x_i - y_i| \quad \text{.....3.1}$$

The least distance of cases or subclusters was sequential combined to produce dendrogram. In the first step all cases are considered as separate clusters, afterthat 2 cases which have least distance are linked and combined bigger cluster. Distance of rest cases and new subcluster are recalculated. The new distances are considered, the least distance is merge with the formerly subcluster. Linkage cases/subcluster is done sequentially until all cases combined into single cluster. This present work, complete linkage is used as combined method. Complete linkage also called furthest neighbors, the distance between one cluster and another cluster is considered as the furthest distance from any member of one cluster to any member of the another cluster [58]. When HCA was performed, number of natural groups of signals in dataset was considered from distance cluster combine.

3.2.2.2 *K*-means clustering

K-means is a partitioning clustering method which clustering samples into k groups [59, 60]. Number cluster, k, is necessary to assign. In this present work, k was defined by using natural group as same as HCA. Each sample, xi, is inputted and assigned to the cluster whose centroid is closest, using a Euclidean distance metric. Cluster centroids are then updated. Iterative process is done by running means method. The process repeats until no more reassignments of the observational units occur [61].

3.2.3 Identification of group of signals of sample: Linear discriminate analysis (LDA)

Linear Discriminant Analysis (LDA) [62-65] is performed for the classified model on conditions that dataset has normal distribution with the same dispersion (variance–covariance matrix) for each group. Distances of each signal of sample are calculated by Euclidean distance. LDA model gain the boundary to classified groups of signal of samples by considering of eigenvalue. The linear discriminant functions are obtained by linear combinations of initial variables. The number of orthogonal linear discriminant functions or called canonical functions is the number of signal of sample groups minus 1.

3.3 Experimental

3.3.1 Osmotic Fragility Test (OFT) of blood samples [1]

Data used here were taken from reference [1]. All 73 blood samples (21 samples are positive thalassemia patients (RT) and 52 negative thalassemia patients (RN)) were collected by the Thalassemia Research Laboratories, Maharajnakorn Chiang Mai Hospital, Chiang Mai University, Thailand. For thalassemia screening, stopped flow-hydrodynamic injection (SF-HI) system was performed. Solution stream (10 mM phosphate buffer pH 7.5 without NaCl and with0.55% NaCl which used as washing/equilibrating buffer and hypotonic saline solution, respectively) direction was controlled by three-way solenoid valves. Microcontroller was used for controlling of time on the valves. After blood was injected by hydrodynamic force, the mixture solution was stopped in flow cell for 30 s, sample injection volume was 40 ul approximately. The change in transmittance at wavelength 620 nm was observed by Spectronic21. The signal was recorded as voltage with the lab-built computer software and BASIC Stamp data recording system. From the reference66, the slope of osmotic fragility test (OFT) were used as criteria for decision of thalassemia screening. Condition of calculation of slope was optimized until all 73 cases were correctly classified.

In this present work, 4 types of osmotic fragility test signals were used. OFT data matrix obtianed size 73x26 which Row of OFT data matrix expressed case of blood samples and column expressed analysis time of signal.

3.3.2 Apparatus

Chemometrics methods which used in this work were performed on SPSS17, all of software are connected to an AMD Athlon™ 64X2 Dual Core Processor 3800+ 2.20 GHz., 2.00 GB of RAM Physical Address Extension, which were used for the statistical treatment of the data and for the application of PCA, *k*-means clustering, HCA, and LDA methods.

3.3.3 Chemometric methods and data analysis

Clustering analysis methods were used as unbiased tool for screening thalassemia patients. Groups of blood samples were clarified by 2 unsupervised pattern recognition; *k*-means clustering and hierarchical cluster analysis by using 4 types of signal (1 dimension slope, 13 dimension slopes, OFT and PCs of OFT) to be 'PT' (positive patient for thalassemia), 'PN' (negative patient for thalassemia) and 'PU' (unidentified sample). The samples which grouping by 4 types of signal correlated with the database as PT and PN were chosen to use as training set. To identify the signals that discriminate between those two groups, the linear discriminant analysis (LDA) were applied. The statistical analysis of OFT data may be broken down into the following steps in the diagram which contains 4 main parts; data pretreatment, classification of the positive and negative patient using *k*-means and Hierarchical clustering analysis (HCA), selections of training set in LDA and building the LDA model for screening. Visualization is done using principle component analysis.

3.3.4 Data pretreatment

73 thalassemia samples with 21 samples for positive and 52 samples for negative patients of OFT-FIs were recorded by different time frequency and shown in Figure 3.1. From requirement of chemometrics method, each point of signals have to record at the same time period so data pretreatment is needed by using 3 steps.

a) calculating the signal mean per second. The total analysis time is equal for each case.

b) selecting data in range from the working time at 25-50 seconds for chemometrics analysis. The slope of the invested area is showned in Figure 3.2 .The time is rescaled to 1-26 seconds range.

c) smoothing data using the adjacent average with 5 interval point.

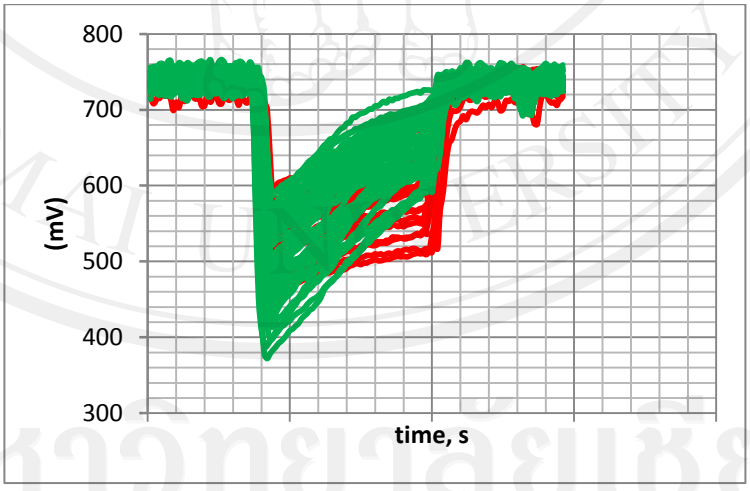


Figure 3.1 Raw OFT signals of 73 cases obtained 21 positive test of thalassemia (red line) and 52 negative test of thalassemia [1]

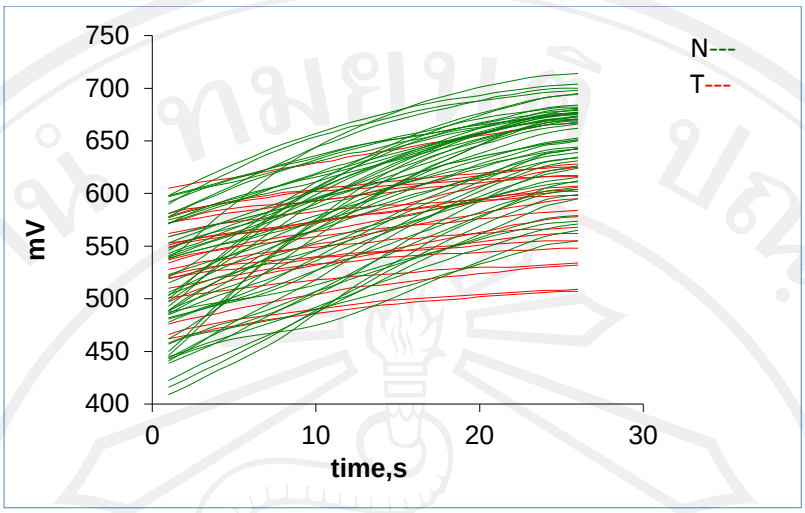


Figure 3.2 The selected range of OFT signals of blood samples

From figure 3.1, in this study the comparison of various types of selected range of signals were used to find the natural group of blood samples by considering of groups of signals of samples in different views obtained. 4 types of signals which used in this study obtained OFT, 1 dimension slope, 13 dimension slopes, and PCs of OFT.

OFT is signal that shows turbidity of solution because of blood broken in range of 1-26 seconds. 1 dimension slope is slope that calculated OFT during 1-26 second range. 13 dimension slopes is gradient slopes within 2 second interval from 1 -26 second range to perform 13 interval slopes. The 2 types of slope expressed change of the signal with overall slope and digested gradient slope. The principle components during 1-26 second range were extracted by using correlation metric and varimax rotation method , the component scores, which was condensed information from OFT, were extracted by regression method, only principle components which have eigenvalues greater than 1 were selected.

3.3.5 Classification methods: *k*-means clustering and hierarchical clustering analysis (HCA)

Hierarchical clustering analysis (HCA), *k*-means clustering, and principal component analysis were performed on SPSS version 17. Clustering results from 8 models of 4 signal types by 2 clustering methods reflected natural group of samples. Signals of samples can be clustered with various probabilities. Samples which were clustered to be negative test of thalassemia with more than 4 of 8 clustering models have probability to be normal person (named as PN). Samples which were clustered to be positive test of thalassemia with more than 4 of 8 clustering models have probability to be thalassemia patient (named as PT). Samples which were clustered to be negative test of thalassemia with 4 of 8 clustering models have probability to be unidentified sample (named as PU). Clearly identified samples to be negative test and positive test of thalassemia were clustered by 8 clustering models to be PN and PT respectively. Finally, the samples which cannot identify clearly will be excluded in order to build the suitable model for positive and negative patients using LDA.

3.3.6 LDA model for screening and visualization using principle component analysis

Linear discriminant analysis was also performed using SPSS version 17. Characteristic of the LDA model was described by eginvalue and canonical correlation. Those terms are defined as:

-Eigenvalue is explained how well the model can separate sample signals in term of ratio of Between-Group Sum Square by Within-Group Sum Square.

-Canonical correlation is correlation between blood samples groups and signal type.

Distributions of various types of samples were plotted and were labeled by group from LDA classification. Furthermore, OFT distribution of blood samples and 13 dimension slopes cannot be easy to notice. Visualization of OFT and 13 dimension slopes distributions can be done by using PCA. Information of distribution of OFT and 13 dimension slopes can be extracted by PCA and shown as PCA score plots.

3.3.7 LDA training set selection and construction of LDA models.

The ratio of each function of LDA models was used to classify samples**.** The sample with more than 3/4 of LDA models was predicted to be negative or positive test of thalassemia was assigned as MN or MT respectively. The sample with 2/4 of LDA models was predicted to be MT and MN was assigned as MU.

3.4 Results and discussions

3.4.1 Grouping of blood samples by hierarchical cluster analysis (HCA) and *k*-means clustering

The comparison of clustering 2 methods (hierarchical clustering and *k*-means clustering) by those 4 types of signal is alternative way to convince group of samples. The training set was chosen by selecting of the samples which clustering groups of 4 types of signals were correlated with the database. For guarantee positive and negative test of thalassemia samples,

the unidentified group was defined as the samples which cluster in PU group or differently cluster among the clustering of 4 types of signals and database.

Dendrogram of each type of signal were shown in figure 3.3. The groups of samples were defined from dissimilarity using city block distance and complete linkage. The groups of the clustering obtain positive with low slope and negative test of thalassemia with high slope and unidentified sample with moderate slope (the group was called PT, PN and PU, respectively).

K-means clustering can be also clustered samples in 3 groups. Summary of *k*-means and HCA group clustering as PT, PN or PU with different signal clustering were shown in table 3.1.

The various patterns of 8 clustering models from using HCA and *k*-means clustering and 4 kind of signals of samples can be found.

In this work, 11 samples which were clustering predicted from all types of signal be the same group and correlated to database were chosen to set the training set. The training set obtained from all correlated data are sample number 1, 2, 8, 9, 12, 13, and 14 in T group and sample number 31, 36, 40 and 71 in the PN group. For uncorrelated data, the predict group from clustering method identify by PN or PT when ratio of clustering was more than 4 of 8 of clustering models that sample was clustered to PN or PT group, respectively, other samples was identified to PU .

Pattern of the clustering to be PT, PN and PU group can be noticed from table 3.1.T group obtained 20 samples which clustered by the ratios of PT:PU:PN are (8PT:0PU:0PN), (6PT:2PU:0PN),(5PT:3PU:0PN) are

7, 2, and 11 samples, respectively. PN group obtained 11 samples which clustered by the ratios of PT:PU:PN are (0PT:0PU:8PN), (0PT:1PU:7PN), (0PT:2PU:6PN), (0PT:3PU:5PN) are 4,2,4, and 1, respectively. Other samples were identified to be PU 42 samples.

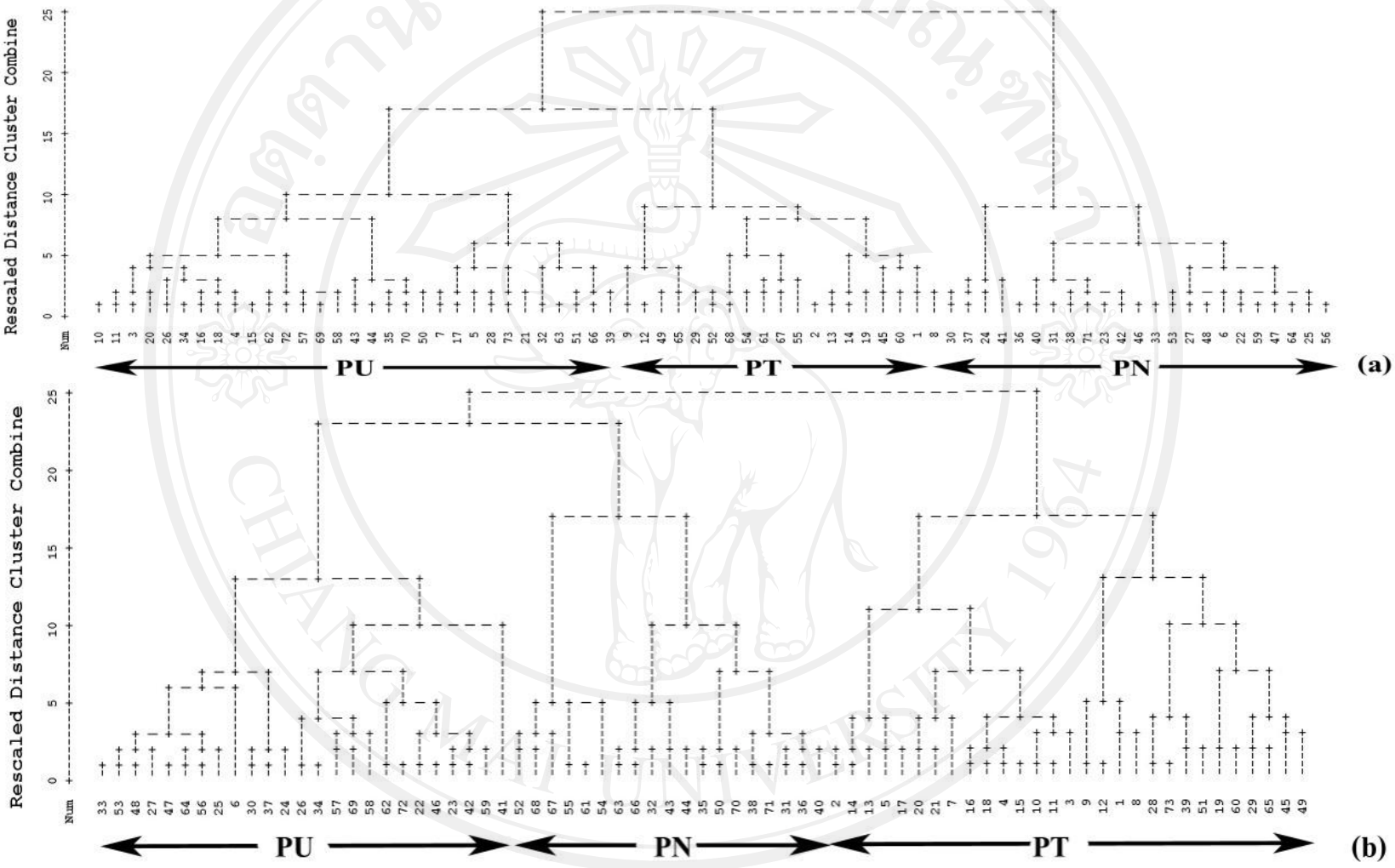


Figure 3.3 Dendrograms of blood samples; OFT (a), and PCs of OFT (b), ;1 dimension slope (c), 13 dimension slopes (d)

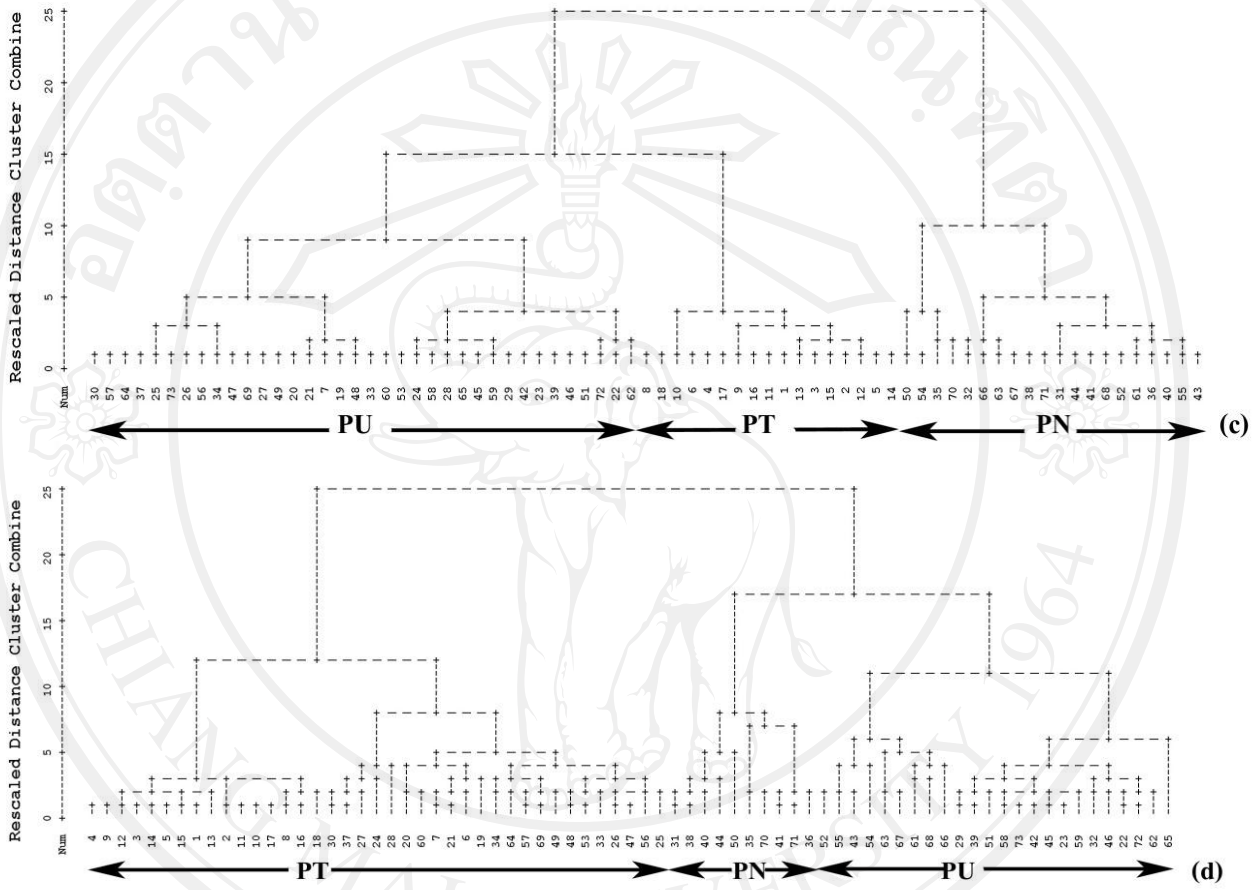


Figure 3.3 Dendrograms of blood samples; OFT (a), and PCs of OFT (b), ;1 dimension slope (c), 13 dimension slopes (d)
-continue-

ble 3.1 Clustering analysis of blood samples (label selected cases which will used to set LDA)

Case	Slope	OFT		PC of OFT		1 dimension slope		13 dimension slopes	
		G_k means	G_HCA	G_k means	G_HCA	G_k means	G_HCA	G_k means	G_HCA
1	**1.32**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**
2	**1.68**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**
3	1.44	PU	PU	PU	PT	PT	PT	PT	PT
4	1.88	PU	PU	PU	PT	PT	PT	PT	PT
5	1.6	PT	PU	PU	PT	PT	PT	PT	PT
6	2.48	PN	PN	PU	PU	PT	PT	PT	PT
7	3.04	PU	PU	PT	PT	PT	PU	PT	PT
8	**2.24**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**
9	**1.8**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**
10	2.2	PU	PU	PU	PT	PT	PT	PT	PT
11	1.96	PU	PU	PU	PT	PT	PT	PT	PT
12	**1.72**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**
13	**1.32**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**
14	**1.52**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**	**PT**
15	1.4	PU	PU	PU	PT	PT	PT	PT	PT
16	1.8	PU	PU	PU	PT	PT	PT	PT	PT
17	1.88	PU	PU	PU	PT	PT	PT	PT	PT
18	2.24	PU	PU	PU	PT	PT	PT	PT	PT
19	3.24	PT	PT	PT	PT	PU	PU	PT	PT
20	2.96	PU	PU	PU	PT	PT	PU	PT	PT
21	2.92	PU	PU	PU	PT	PT	PU	PT	PT
22	5.32	PN	PN	PN	PU	PU	PU	PU	PU
23	5.04	PN	PN	PN	PU	PU	PU	PU	PU
24	4.76	PN	PN	PN	PU	PU	PU	PU	PT
25	4.16	PN	PN	PN	PU	PU	PU	PU	PT
26	3.68	PU	PU	PU	PU	PU	PU	PU	PT
27	3.88	PN	PN	PU	PU	PU	PU	PU	PT
28	4.88	PU	PU	PT	PT	PU	PU	PU	PT
29	5.04	PT	PT	PT	PT	PU	PU	PU	PU
30	4.24	PN	PN	PN	PU	PU	PU	PU	PT
31	**6.96**	**PN**	**PN**	**PN**	**PN**	**PN**	**PN**	**PN**	**PN**
32	6.16	PU	PU	PN	PN	PN	PN	PU	PU
33	3.48	PN	PN	PU	PU	PU	PU	PU	PT
34	3.84	PU	PU	PU	PU	PU	PU	PU	PT
35	8.84	PU	PU	PN	PN	PN	PN	PN	PN
36	**7.36**	**PN**	**PN**	**PN**	**PN**	**PN**	**PN**	**PN**	**PN**
37	4.32	PN	PN	PN	PU	PU	PU	PU	PT
38	7.12	PU	PN	PN	PN	PN	PN	PN	PN
39	4.96	PT	PU	PT	PT	PU	PU	PU	PU
40	**7.64**	**PN**	**PN**	**PN**	**PN**	**PN**	**PN**	**PN**	**PN**

Table 3.1 Clustering analysis of blood samples (label selected cases which will used to set LDA) –continue-

Case	Slope	OFT		PC of OFT		1 dimension slope		13 dimension slopes	
		G_k means	G_HCA	G_k means	G_HCA	G_k means	G_HCA	G_k means	G_HCA
41	7	PN	PN	PN	PU	PN	PN	PN	PN
42	5.04	PU	PN	PN	PU	PU	PU	PU	PU
43	7.48	PU	PU	PN	PN	PN	PN	PN	PU
44	6.96	PU	PU	PN	PN	PN	PN	PN	PN
45	4.84	PT	PT	PT	PT	PU	PU	PU	PU
46	5.56	PN	PN	PN	PU	PU	PU	PU	PU
47	3.84	PN	PN	PU	PU	PU	PU	PU	PT
48	3.2	PN	PN	PU	PU	PU	PU	PT	PT
49	3.92	PT	PT	PT	PT	PU	PU	PU	PT
50	8.32	PU	PU	PN	PN	PN	PN	PN	PN
51	5.56	PU	PU	PT	PT	PU	PU	PU	PU
52	7.32	PT	PT	PT	PN	PN	PN	PN	PU
53	3.36	PN	PN	PU	PU	PU	PU	PU	PT
54	8.12	PT	PT	PT	PN	PN	PN	PN	PU
55	7.6	PT	PT	PN	PN	PN	PN	PN	PU
56	3.72	PN	PN	PU	PU	PU	PU	PU	PT
57	4.24	PU	PU	PN	PU	PU	PU	PU	PT
58	4.64	PU	PU	PN	PU	PU	PU	PU	PU
59	4.83	PN	PN	PN	PU	PU	PU	PU	PU
60	3.46	PT	PT	PT	PT	PU	PU	PU	PT
61	7.29	PT	PT	PT	PN	PN	PN	PN	PU
62	5.75	PU	PU	PN	PU	PU	PU	PU	PU
63	6.49	PU	PU	PN	PN	PN	PN	PU	PU
64	4.25	PN	PN	PU	PU	PU	PU	PU	PT
65	4.89	PT	PT	PT	PT	PU	PU	PU	PU
66	6.27	PU	PU	PN	PN	PN	PN	PU	PU
67	6.61	PT	PT	PT	PN	PN	PN	PN	PU
68	6.9	PT	PT	PT	PN	PN	PN	PN	PU
69	3.8	PU	PU	PU	PU	PU	PU	PU	PT
70	9.2	PU	PU	PN	PN	PN	PN	PN	PN
71	**7.11**	**PN**	**PN**	**PN**	**PN**	**PN**	**PN**	**PN**	**PN**
72	5.5	PU	PU	PN	PU	PU	PU	PU	PU
73	4.45	PU	PU	PT	PT	PU	PU	PU	PU

Table 3.2 Member in each group of clustering

Signal type	HCA			*K*-means clustering		
	PT	PN	PU	PT	PN	PU
OFT	19 (26%)	23(32%)	31(42%)	21 (29%)	21 (29%)	31(42%)
PCs of OFT	29(40%)	19(26%)	25(34%)	23 (32%)	28 (38%)	22 (30%)
1 dimension slope	17 (23%)	20 (27%)	36 (49%)	20 (27%)	20 (27%)	33(45%)
13 dimension slopes	39(53%)	10 (14%)	24 (33%)	22 (30%)	17 (23%)	34 (47%)

Amount of samples in each group of clustering was show in table 3.2. When HCA was used to cluster samples by 4 types of signal, amount of samples in 3 groups were different. From hospital records, 29% should be PT and 71% should be PN. 13 dimension slopes with HCA gave different result from other type of data. Whereas *k*-means clustering gave similar result.

3.4.2 LDA training set selection and construction of LDA models

Identification of samples was considered from 4 LDA models which use 4 different kinds of signal with same training set. Sample were identified as MT or MN when more than 3 of 4 LDA models predicted in the same group, on the other hand, samples were identified as U when 2 of 4 LDA models are shown. The LDA prediction result was shown in Table 3.3.

From the table samples were predicted to MT group 39 samples; 4 of 4 of LDA models (34 samples) and 3 of 4of LDA models (5 samples). Samples were predicted to MN group 25 samples; 4 of 4 of LDA models (12 samples) and 3 of 4 of LDA models (13 samples). 9 samples were unidentified samples.

Table 3.3 Identification of blood samples by 4 LDA models

Case	Record	LDA classified result				Prediction
		OFT	PC of OFT	Overall slope	Gradient slope	
1	RT	MT	MT	MT	MT	MT
2	RT	MT	MT	MT	MT	MT
3	RT	MN	MT	MT	MT	MT
4	RT	MN	MT	MT	MT	MT
5	RT	MT	MT	MT	MT	MT
6	RT	MN	MT	MT	MT	MT
7	RT	MT	MT	MT	MT	MT
8	RT	MT	MT	MT	MT	MT
9	RT	MT	MT	MT	MT	MT
10	RT	MT	MT	MT	MT	MT
11	RT	MT	MT	MT	MT	MT
12	RT	MT	MT	MT	MT	MT
13	RT	MT	MT	MT	MT	MT
14	RT	MT	MT	MT	MT	MT
15	RT	MN	MT	MT	MT	MT
16	RT	MN	MT	MT	MT	MT
17	RT	MT	MT	MT	MT	MT
18	RT	MN	MT	MT	MT	MT
19	RT	MT	MT	MT	MT	MT
20	RT	MN	MT	MT	MT	MT
21	RT	MT	MT	MT	MT	MT
22	RN	MN	MN	MN	MN	MU
23	RN	MN	MN	MN	MT	MU
24	RN	MN	MN	MN	MN	MU
25	RN	MN	MN	MT	MT	MU
26	RN	MN	MT	MT	MT	MT
27	RN	MN	MN	MT	MT	MU
28	RN	MT	MT	MN	MT	MU
29	RN	MT	MT	MN	MT	MU
30	RN	MN	MN	MT	MT	MU
31	RN	MN	MN	MN	MN	MN
32	RN	MT	MN	MN	MT	MU
33	RN	MN	MN	MT	MT	MT
34	RN	MN	MT	MT	MT	MT
35	RN	MN	MN	MN	MN	MN
36	RN	MN	MN	MN	MN	MN
37	RN	MN	MN	MT	MN	MU
38	RN	MN	MN	MN	MN	MN
39	RN	MT	MT	MN	MT	MU
40	RN	MN	MN	MN	MN	MN
41	RN	MN	MN	MN	MN	MN
42	RN	MN	MN	MN	MT	MU
43	RN	MN	MN	MN	MN	MN
44	RN	MN	MN	MN	MN	MN
45	RN	MT	MT	MN	MT	MU
46	RN	MN	MN	MN	MN	MU
47	RN	MN	MN	MT	MT	MT
48	RN	MN	MN	MT	MT	MT
49	RN	MT	MT	MT	MT	MT
50	RN	MN	MN	MN	MN	MN
51	RN	MT	MT	MN	MT	MU

Table 3.3 Identification of blood samples by 4 LDA models -continue-

Case	Record	LDA classified result				Prediction
		OFT	PC of OFT	Overall slope	Gradient slope	
52	RN	MT	MN	MN	MN	MU
53	RN	MN	MN	MT	MT	MT
54	RN	MT	MN	MN	MN	MU
55	RN	MT	MN	MN	MT	MN
56	RN	MN	MN	MT	MT	MT
57	RN	MN	MT	MT	MT	MT
58	RN	MN	MT	MN	MT	MU
59	RN	MN	MN	MN	MT	MU
60	RN	MT	MT	MT	MT	MT
61	RN	MT	MN	MN	MT	MU
62	RN	MN	MN	MN	MT	MU
63	RN	MT	MN	MN	MN	MU
64	RN	MN	MN	MT	MT	MT
65	RN	MT	MT	MN	MT	MU
66	RN	MT	MN	MN	MN	MU
67	RN	MT	MN	MN	MN	MU
68	RN	MT	MT	MN	MN	MU
69	RN	MN	MT	MT	MT	MT
70	RN	MN	MN	MN	MN	MN
71	RN	MN	MN	MN	MN	MN
72	RN	MN	MN	MN	MT	MU
73	RN	MT	MT	MT	MT	MT

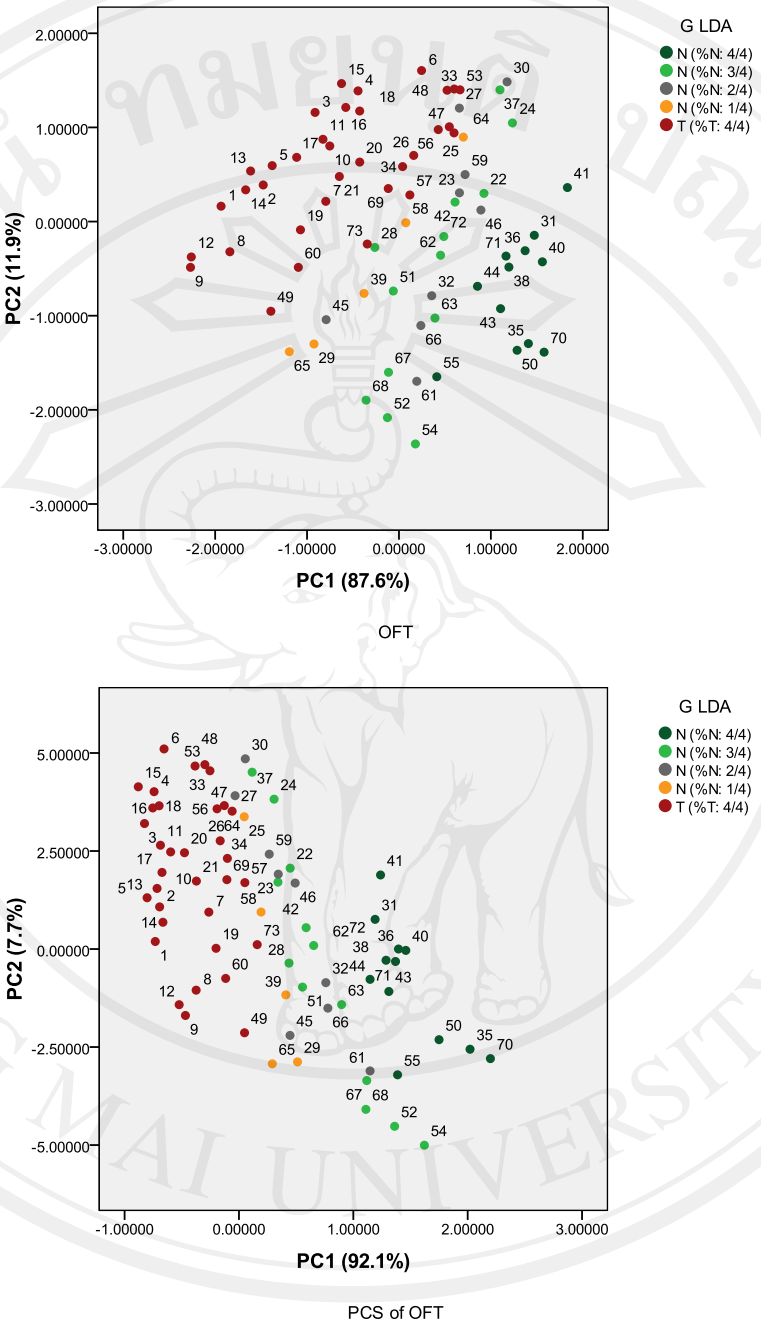


Figure 3.4 Distribution of blood samples defined by groups selected cases for LDA; OFT (top), and PCs of OFT (bottom), 1 dimension slope (top), 13 dimension slopes (bottom) (red dot is sample which predict by 4 of 4 of LDA models to MT, yellow dot is sample which predict by 1 of 4 of LDA models to MN, grey dot is sample which predict by 2 of 4 of LDA models to MN, light green dot is sample which predict by 3 of 4 of LDA models to MN , and dark green dot is sample which predict by 4 of 4 of LDA models to MN)

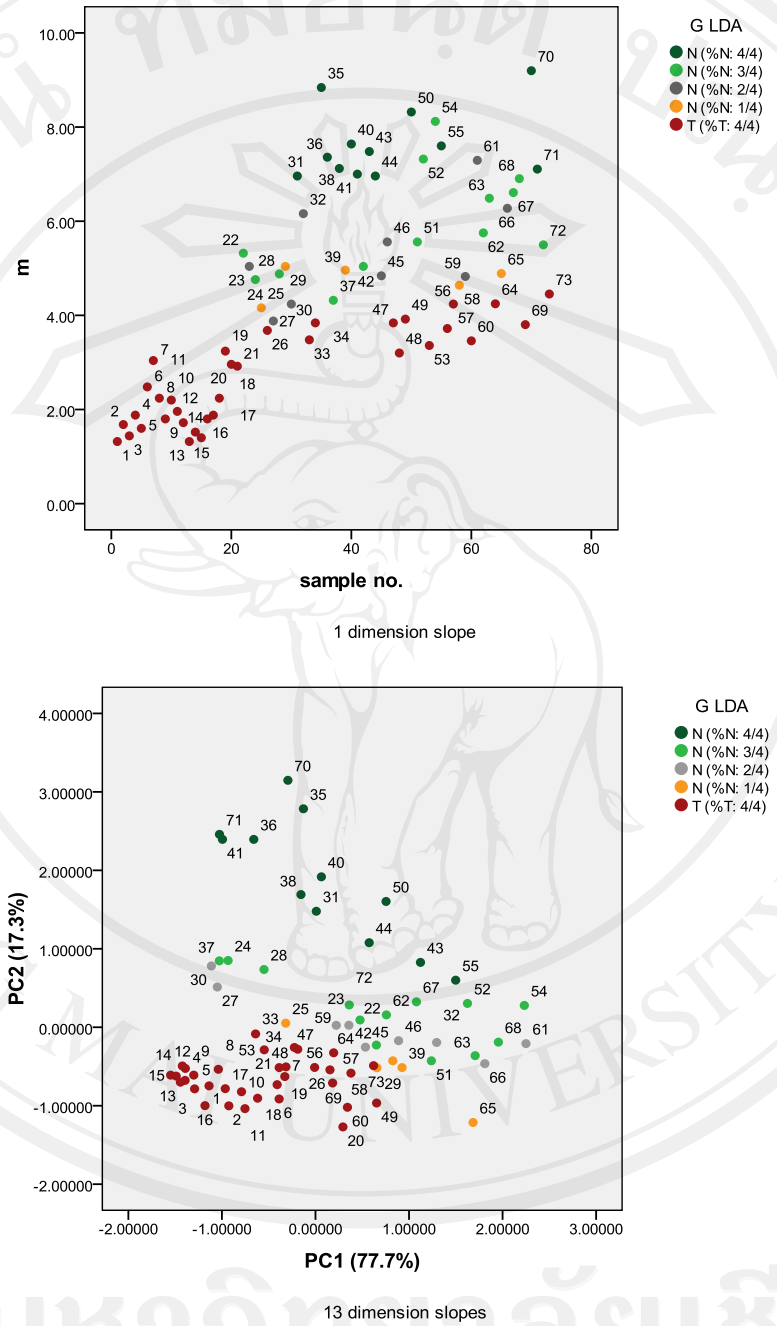


Figure 3.4 Distribution of blood samples defined by groups selected cases for LDA; OFT (top), and PCs of OFT (bottom), 1 dimension slope (top), 13 dimension slopes (bottom) (red dot is sample which predict by 4 of 4 of LDA models to MT, yellow dot is sample which predict by 1 of 4 of LDA models to MN, grey dot is sample which predict by 2 of 4 of LDA models to MN, light green dot is sample which predict by 3 of 4 of LDA models to MN , and dark green dot is sample which predict by 4 of 4 of LDA models to MN) -continue-

In Figure 3.4, the distributions of samples in each case of LDA models predicted by 4 types of signal with different prediction ratio were shown. The samples which were predicted as MN or MT in ratio of 4 of 4 of LDA models were clearly separation. Samples in moderate zone which were in cycles are unreliable signal different from database and line in boundary between MT and MN group. Such signal are samples number 25, 26, 29, 33, 34, 39, 47, 48, 49, 53, 56, 57, 58, 60, 64, 65, 69, and 73 with the moderate slope which indicated in figure 3.5(the red lines).

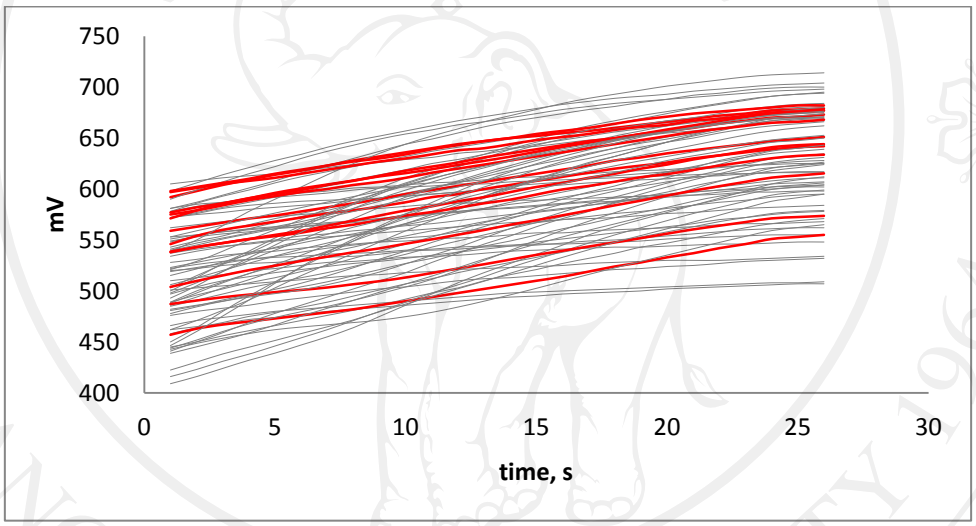
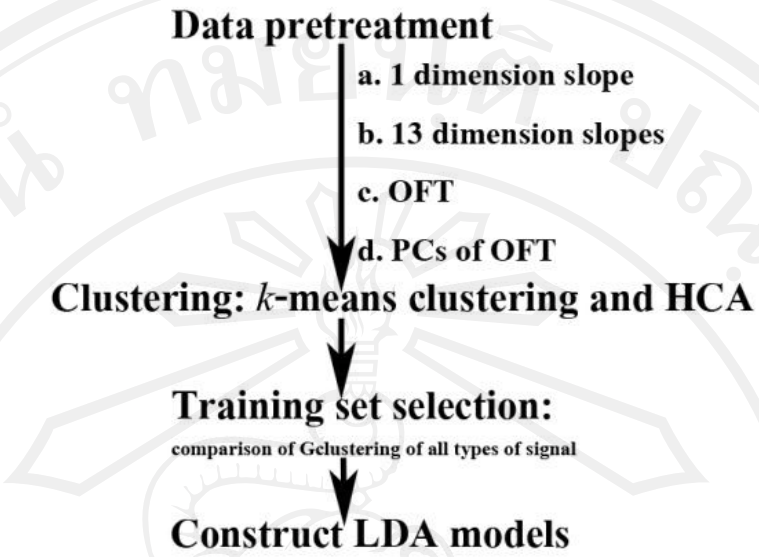


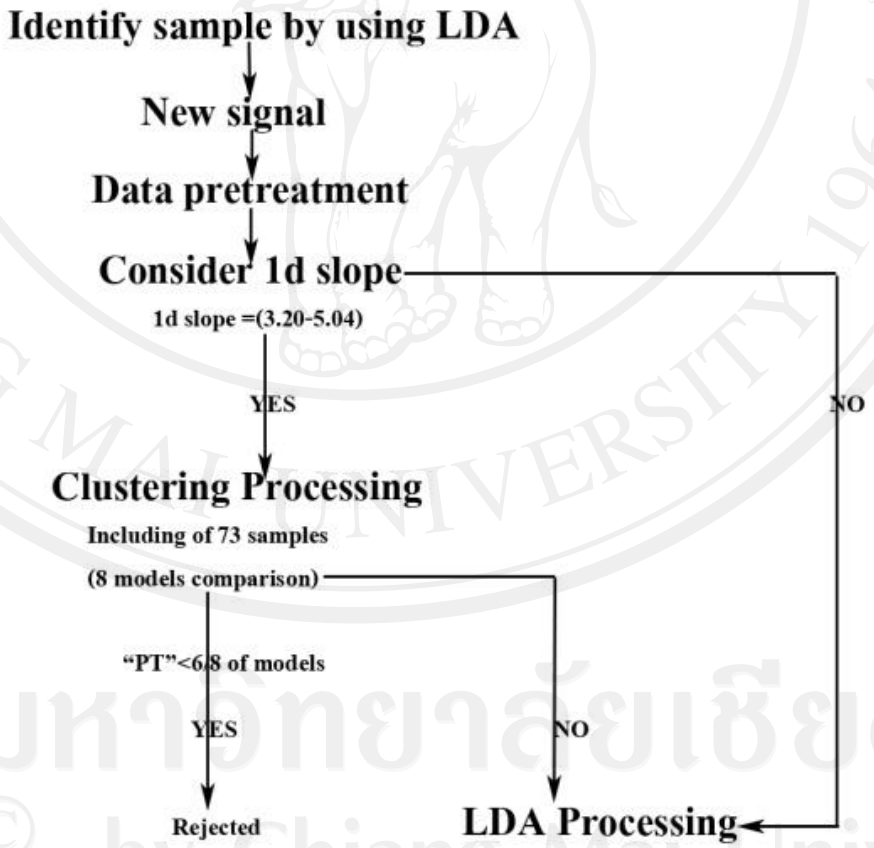
Figure 3.5 Unreliable signals different from database

In application of our model can be clearly predicted signal with high and low slope using LDA model. The LDA models results were compared with the referent method [55] and hospital record. It was found that 21 samples which were predicted to be “MT” were correlated with referent method and hospital record. The comparison between “MN” samples which predicted by LDA and hospital record shown 2 different types of samples. The first one, LDA results was correlated with record as MN and MU. samples which were predicted as MN 25 samples are sample

number 22, 24, 28, 31, 35, 36, 37, 38, 40, 41, 42, 43, 44, 50, 51, 52, 54, 55, 62, 63, 67, 68, 70, 71 and 72. Samples which were predicted as MU 9 samples are sample number 23, 27, 30, 32, 45, 46, 59, 61 and 66. Second one, the predictions (as MT) were uncorrelated with hospital record (as MN) 18 samples are sample number 25, 26, 29, 33, 34, 39, 47, 48, 49, 53, 56, 57, 58, 60, 64, 65, 69, and 73. 1 dimension slope value and ratio of MT of clustering process were used as boundary of LDA prediction. It was found that the moderate signal 3.20 to 5.04 unable to identify the group clearly. The samples which obtained 1 dimension slope within 3.20-5.04 need to check from clustering result before predict with LDA. Only samples which clustering as T more than 6/8 of clustering models can be allowed to analyze with LDA models. The process of the screening of samples is shown in figure 3.6.



(a)



(b)

Figure 3.6 Overview of thalassemia screening by aid of clustering methods; (a) training model and (b) predict sample

The samples which were predicted as other ratios of prediction were still gave overlapping distribution. Although some samples had overlapping distribution but the LDA models results were comparison before identify samples group so the prediction gave high confident result. Characteristic of the LDA models are show in table 3.4.

Table 3.4 LDA canonical function and characteristic of the LDA models

Signal type	Canonical Discriminant Function	Eigenvalue	Canonical Correlation
OFT	D=0.427*t1 - 0.609*t2 - 0.529 *t4 + 0.749*t6 - 27.972	70.226[a]	.993
PCs of OFT	D=13.133*PC1 + 1.532*PC2 - 1.309	192.921[a]	.997
1 dimension slope	D=3.201*m - 11.833	91.177[a]	.995
13 dimension slopes	D=0.496*m1 - 0.288*m2 + 0.244*m3 + 2.250*m4 + 3.369*m5 -4.093*m6 + 0.238*m7 - 1.652*m8 + 2.814*m9 - 17.062	214.463[a]	.998

a. First 1 canonical discriminant functions were used in the analysis.

From table 3.4, Canonical function was used as classification model. From the function, it was found that some variables of 13 dimension slopes and OFT were selected out. The canonical function of 13 dimension slopes contained only m1-m9 and the canonical function of OFT contain only 4 variables; t1, t2, t4 and t6 from the initial of variables of signal. All canonical functions can be used to classify samples with the high eigenvalues. Furthermore the different between groups compared by within group of LDA model of 13 dimension slopes and PCs of OFT are more than 2 times. The canonical coefficient of all LDA model can used for predict group of sample with in 0.99.

When the LDA predictions were compared with hospital records, it was found that predictions of 46 samples were correlated with the records but

predictions of 27 samples were uncorrelated with the records but these samples should be ejected when consider from process which shown in figure 3.6 so the screening still saved.

3.5 References

1 R. G. Brereton, *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*. Editor, John Wiley, New York, **2003**.

2 L. Luo, *X-Ray Spectrometry* **2006**, *35*, 215-225 10.1002/xrs.894.

3 H. Philip K, *Analytica Chimica Acta* **2003**, *500*, 365-377.

4 P. W. T. Krooshof, G. J. Postma, W. J. Melssen, L. M. C. Buydens, T. N. Tran, *TrAC Trends in Analytical Chemistry* **2006**, *25*, 1067-1080.

5 J. Y. Huang, Y. B. Qiu, X. P. Guo, *Electrochimica Acta* **2009**, *54*, 2218-2223.

6 S. J. Redmond, C. Heneghan, *Pattern Recognition Letters* **2007**, *28*, 965-973.

7 A. L. Hsu, S.-L. Tang, S. K. Halgamuge, *Bioinformatics* **2003**, *19*, 2131-2140.

8 J. F. Lu, J. B. Tang, Z. M. Tang, J. Y. Yang, *Pattern Recognition Letters* **2008**, *29*, 787-795.

9 A. Smoliński, B. Walczak, J. W. Einax, *Chemometrics and Intelligent Laboratory Systems* **2002**, *64*, 45-54.

10 M. F. M. Engels, A. C. Gibbs, E. P. Jaeger, D. Verbinnen, V. S. Lobanov, D. K. Agrafiotis, *journal of Chemical information and modeling* **2006**, *46*, 2651-2660.

11 R. G. Brereton, *Multivariate Pattern Recognition in Chemometrics Illustrated by Case Studies*. Editor, Elsevier Amsterdam, The Netherlands, **1992**.

12 M. Penza, G. Cassano, *Analytica Chimica Acta* **2004**, *509*, 159-177.

13 M. Penza, G. Cassano, F. Tortorella, *Measurement Science and Technology* **2002**, *13*, 846.

14 E. Llobet, E. L. Hines, J. W. Gardner, P. N. Bartlett, T. T. Mottram, *Sensors and Actuators B: Chemical* **1999**, *61*, 183-190.

15 L. Kaufman, P. R. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Editor, John Wiley & Sons, New York, **1990**.

16 R. O. Duda, D. G. S. P.E. Hart, *Pattern Classification*. Editor, John Wiley & Sons, New York, **2001**.

17 Y. Roggo, L. Duponchel, C. Ruckebusch, J. P. Huvenne, *Journal of Molecular Structure* **2003**, *654*, 253-262.

18 S. Rebolo, R. M. Peña, M. J. Latorre, S. García, A. M. Botana, C. Herrero, *Analytica Chimica Acta* **2000**, *417*, 211-220.

19 R. M. Alonso-Salces, C. Herrero, A. Barranco, D. M. López-Márquez, L. A. Berrueta, B. Gallo, F. Vicente, *Food Chemistry* **2006**, *97*, 438-446.

20 E. Marengo, M. Aceto, V. Maurino, *Journal of Chromatography A* **2002**, *943*, 123-137 10.1016/s0021-9673(01)01421-2.

21 J. M. Jurado, A. Alcázar, F. Pablos, M. J. Martín, A. G. González, *Talanta* **2005**, *66*, 1350-1354.

22 R. Fernández-Torres, J. L. Pérez-Bernal, M. Á. Bello-López, M. Callejón-Mochón, J. C. Jiménez-Sánchez, A. Guiraúm-Pérez, *Talanta* **2005**, *65*, 686-691.

23 A. Moreda-Piñeiro, A. Fisher, S. J. Hill, *Journal of Food Composition and Analysis* **2003**, *16*, 195-211.

24 M. J. Martín, F. Pablos, A. G. González, *Analytica Chimica Acta* **1996**, *320*, 191-197.

25 L.-J. Ni, L.-G. Zhang, J. Xie, J.-Q. Luo, *Analytica Chimica Acta* **2009**, *633*, 43-50.

26 Y. González Martín, M. C. Cerrato Oliveros, J. L. Pérez Pavón, C. García Pinto, B. Moreno Cordero, *Analytica Chimica Acta* **2001**, *449*, 69-80 .

27 R. M. Alonso-Salces, S. Guyot, C. Herrero, L. A. Berrueta, J.-F. Drilleau, B. Gallo, F. Vicente, *Food Chemistry* **2005**, *91*, 91-98.

28 Q. Zhang, C. Xie, S. Zhang, A. Wang, B. Zhu, L. Wang, Z. Yang, *Sensors and Actuators B: Chemical* **2005**, *110*, 370-376.

29 C. Sola-Larrañaga, I. Navarro-Blasco, *Analytica Chimica Acta* **2006**, *555*, 354-363.

30 F. Marini, A. L. Magrì, F. Balestrieri, F. Fabretti, D. Marini, *Analytica Chimica Acta* **2004**, *515*, 117-125.

31 L. A. Berrueta, R. M. Alonso-Salces, K. Héberger, *Journal of Chromatography A* **2007**, *1158*, 196-214.

32 D. González-Arjona, G. López-Pérez, V. González-Gallero, A. G. González, *Journal of Agricultural and Food Chemistry* **2006**, *54*, 1982-1989.

33 W. Cynkar, D. Cozzolino, B. Dambergs, L. Janik, M. Gishen, *Sensors and Actuators B: Chemical* **2007**, *124*, 167-171.

34 R. M. Alonso-Salces, C. Herrero, A. Barranco, L. A. Berrueta, B. Gallo, F. Vicente, *Food Chemistry* **2005**, *93*, 113-123.

35 F. Marini, F. Balestrieri, R. Bucci, A. D. Magrì, A. L. Magrì, D. Marini, *Chemometrics and Intelligent Laboratory Systems* **2004**, *73*, 85-93.

36 T. Yukio, *Chemometrics and Intelligent Laboratory Systems* **1999**, *49*, 105-115.

37 Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, N. Jent, *Journal of Pharmaceutical and Biomedical Analysis* **2007**, *44*, 683-700.

38 H.-J. Kim, H. Choo, Y. S. Cho, H. Y. Koh, K. T. No, A. N. Pae, *Bioorganic & Medicinal Chemistry* **2006**, *14*, 2763-2770.

39 O. Beckonert, M. E. Bollard, T. M. D. Ebbels, H. C. Keun, H. Antti, E. Holmes, J. C. Lindon, J. K. Nicholson, *Analytica Chimica Acta* **2003**, *490*, 3-15.

40 S. L. Dixon, H. O. Villar, *Journal of Computer-Aided Molecular Design* **1999**, *13*, 533-545.

41 P. Bermejo-Barrera, A. Moreda-Piñeiro, A. Bermejo-Barrera, A. M. a. Bermejo-Barrera, *Analytica Chimica Acta* **2002**, *455*, 253-265.

42 K. M. Pierce, J. L. Hope, K. J. Johnson, B. W. Wright, R. E. Synovec, *Journal of Chromatography A* **2005**, *1096*, 101-110.

43 T. Zhang, N. Y. Edwards, M. Bonizzoni, E. V. Anslyn, *Journal of the American Chemical Society* **2009**, *131*, 11976-11984.

44 S. Hamilton, M. J. Hepher, J. Sommerville, *Sensors and Actuators B: Chemical* **2006**, *113*, 989-997.

45 J. V. Marcos, R. Hornero, D. Álvarez, F. del Campo, C. Zamarrón, *Medical Engineering & Physics* **2009**, *31*, 971-978.

46 J. Yang, G. Xu, Q. Hong, H. M. Liebich, K. Lutz, R. M. Schmülling, H. G. Wahl, *Journal of Chromatography B* **2004**, *813*, 53-58 .

47 R. Madsen, T. Lundstedt, J. Trygg, *Analytica Chimica Acta* **2010**, *659*, 23-33 .

48 S. Carraro, S. Rezzi, F. Reniero, K. Héberger, G. Giordano, S. Zanconato, C. Guillou, E. Baraldi, *American Journal of Respiratory and Critical Care Medicine* **2007**, *175*, 986-990.

49 M. Zhang, P. Tong, W. Wang, J. Geng, Y. Du, *Chemometrics and Intelligent Laboratory Systems* **2011**, *105*, 207-214.

50 V. Mrázová, J. Mocák, E. Varmusová, D. Kavková, A. Bednárová, *Journal of Pharmaceutical and Biomedical Analysis* **2009**, *50*, 210-215.

51 T. Imre, T. Kremmer, K. Héberger, É. Molnár-Szöllősi, K. Ludányi, G. Pócsfalvi, A. Malorni, L. Drahos, K. Vékey, *Journal of Proteomics* **2008**, *71*, 186-197.

52 A. Devos, L. Lukas, J. A. K. Suykens, L. Vanhamme, A. R. Tate, F. A. Howe, C. Majós, A. Moreno-Torres, M. van der Graaf, C. Arús, S. Van Huffel, *Journal of Magnetic Resonance* **2004**, *170*, 164-175.

53 K.-Z. Liu, K. S. Tsang, C. K. Li, R. A. Shaw, H. H. Mantsch1, *Clinical Chemistry* **2003**, *49*, 1125-1132.

54 M. Arjmand, M. Kompany-Zareh, M. Vasighi, N. Parvizzadeh, Z. Zamani, F. Nazgooei, *Talanta* **2010**, *81*, 1229-1236.

55 S. Khonyoung, S. K. Hartwell, J. Jakmunee, S. Lapanantnoppakhun, T. Sanguansermsri, K. Grudpan, *analytical sciences* **2009**, *25*, 819-824.

56 C. Y. Airiau, H. Shen, R. G. Brereton, *Analytica Chimica Acta* **2001**, *447*, 199-210.

57 A. Akbay, A. Elhan, C. ɜ–zcan, S. Demirtas, *Medical Hypotheses* **2000**, *55*, 147-154.

58 J. Mao, J. Xu, *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy* **2006**, *65*, 497-500.

59 A. Ahmad, L. Dey, *Data and Knowledge Engineering* **2007**, *63*, 503-527.

60 J. C. Lindon, E. Holmes, J. K. Nicholson, *Progress in Nuclear Magnetic Resonance Spectroscopy* **2001**, *39*, 1-40.

61 C. Smyth, D. Coomans, Y. Everingham, T. Hancock, *Chemometrics and Intelligent Laboratory Systems* **2006**, *80*, 120-129.

62 F. Marini, F. Balestrieri, R. Bucci, A. D. Magr, A. L. Magrณ, D. Marini, *Chemometrics and Intelligent Laboratory Systems* **2004**, *73*, 85-93.

63 A. Moreda-Piรฑeiro, A. Fisher, S. J. Hill, *Journal of Food Composition and Analysis* **2003**, *16*, 195-211.

64 Y. Gonzรกlez Martญn, M. C. Cerrato Oliveros, J. L. Pรฉrez Pavรณn, C. Garcญa Pinto, B. Moreno Cordero, *Analytica Chimica Acta* **2001**, *449*, 69-80.

65 L. A. Berrueta, R. M. Alonso-Salces, K. Hberger, *Journal of Chromatography A* **2007**, *1158*, 196-214.

66 S. Khonyoung, S. Kradtap Hartwell, J. Jakmunee, S. Lapanantnoppakhun, T. Sanguansermsri, K. Grudpan, *Analytical Sciences* **2009**, *25*, 819-824.