Waveband selection of reagent-free determination for thalassemia screening indicators using Fourier transform infrared spectroscopy with attenuated total reflection

Xiaoli Long
Guisong Liu
Tao Pan
Jiemei Chen
Waveband selection of reagent-free determination for thalassemia screening indicators using Fourier transform infrared spectroscopy with attenuated total reflection

Xiaoli Long, Guisong Liu, Tao Pan and Jiemei Chen
Jinan University, Key Laboratory of Optoelectronic Information and Sensing Technologies of Guangdong Higher Educational Institutes, Huangpu Road West 601, Tianhe District, Guangzhou 510632, China

Abstract. A reagent-free determination method for the thalassemia screening indicators hemoglobin (Hb), mean corpuscular Hb (MCH), and mean corpuscular volume (MCV) was developed based on Fourier transform infrared spectrometers equipped with an attenuated total reflection accessory. A random and stability-dependent rigorous process of calibration, prediction, and validation was conducted. Appropriate wavebands were selected using the improved moving window partial least squares method with stability and equivalence. The obtained optimal wavebands were 1722 to 1504 cm\(^{-1}\) for Hb, 1653 to 901 cm\(^{-1}\) for MCH, and 1562 to 964 cm\(^{-1}\) for MCV. A model set equivalent to the optimal model was proposed for each indicator; the public waveband of Hb equivalent wavebands was 1717 to 1510 cm\(^{-1}\), and the public equivalent waveband for MCH and MCV was 1562 to 901 cm\(^{-1}\). All selected wavebands were within the MIR fingerprint region and achieved high validation effects. The sensitivity and specificity were 100.0% and 96.9% for the optimal wavebands and 100.0% and 95.3% for the equivalent wavebands, respectively. Thus, the spectral prediction was highly accurate for determining negative and positive for thalassemia screening. This technique is rapid and simple in comparison with conventional methods and is a promising tool for thalassemia screening in large populations. © 2014 Society of Photo-Optical Instrumentation Engineers (SPIE)

Keywords: thalassemia screening; Fourier transform infrared; attenuated total reflection; waveband optimization; stability; equivalence.

Paper 140342R received Jun. 2, 2014; revised manuscript received Jul. 24, 2014; accepted for publication Jul. 25, 2014; published online Aug. 19, 2014.

1 Introduction

Thalassemia comprises a group of genetic disorders of hemoglobin (Hb) synthesis, and it affects individuals from many parts of the world, including South China, where it has a high prevalence and incidence and has caused serious health damage. Conservative estimates show that at least 345 million people around the world carry the genes responsible for this disease. In China, the rates of gene carriers are as high as 24.50% and 11.07% in the population of Guangxi and Guangdong provinces, respectively. This serious disorder is caused by partial or total mutations that reduce or abolish the synthesis of \(\alpha\)- or \(\beta\)-globin chains of the Hb molecule, which will result in hemolytic anemia. At present, the disease cannot be cured, except through hematopoietic stem cell transplantation. The most fundamental prevention methods include premarital and prenatal thalassemia screening in a large population. Heterozygote screening and genetic counseling are essential for preventing and controlling severe thalassemia. The ultimate diagnosis for the different types of thalassemia depends on DNA analysis using the polymerase chain reaction. However, a direct DNA approach without a precise biochemical hematological indication is highly time consuming, is expensive, and often subjects to false-negative results or misinterpretations. The complex relationship between the genotype and phenotype makes diagnosis difficult.

Thus, a combination of different tests is required for accurate diagnosis.

The fundamental test for thalassemia is hematological phenotype screening. First, two main hematological parameters, the mean corpuscular Hb (MCH) and mean corpuscular volume (MCV), are screened. Following this, Hb analysis is further performed on the basis of the parameters of total Hb, HbF, and Hb\(\alpha\)2 to classify the condition into \(\alpha\)- or \(\beta\)-thalassemia. DNA analysis is an accurate and comprehensive method of thalassemia screening, but this approach is relatively complicated and is not suitable for screening large populations.

MCH and MCV are preliminary screening parameters for thalassemia that can effectively evaluate microcytic hypochromic anemia in a large population. The discrimination threshold (cut-off value) of the measured MCH or MCV values is determined on the basis of the highest sum of sensitivity and specificity for discriminating phenotype-positive subjects from phenotype-negative subjects. In the conventional method, the cut-off values corresponding to MCH and MCV are 27.0 pg and 80.0 fL, respectively. Phenotype-positive subjects for thalassemia are those with MCH \(\leq 27.0\) pg or MCV \(\leq 80.0\) fL. Individuals with low MCH or MCV (MCH \(\leq 27.0\) pg or MCV \(\leq 80.0\) fL) are usually further assessed using Hb and DNA analyses for identifying the type of defect.
At present, Hb, MCH, MCV, and other red blood cell indicators are mainly measured by using a blood cell analyzer, which uses an electrical impedance method to perform blood cell count and volume measurement. The blood cell analyzer needs ancillary reagents, such as electrolyte solution and hemolytic agent. Therefore, this procedure is relatively complicated because it requires chemical reagents. Other label-free optical methods like quantitative digital holographic phase microscopy also discussed the measurement of thalassemia screening indicators MCH and MCV; however, this method needs sample pretreatment. A rapid, simple, and reagent-free method may help in pre-marital and prenatal thalassemia screening in population prevention and control programs.

The application of spectroscopy for thalassemia screening, particularly in the mid-infrared (MIR) region, was desirable as a high potential implementation. Actually, the MIR spectrum has rich information on the molecular structure and contents of the material, which could be used to determine the molecular structure of a protein. However, the interference caused by the strong absorption of water molecules in the MIR region is the main difficulty in the use of MIR for direct measurement of biological samples (e.g., blood). Therefore, complex preparation and handling of samples are always required; this process is neither rapid nor simple.

The described difficulty associated with the handling of biological samples has been circumvented by developing the attenuated total reflection (ATR) technique. Incidentally, parallel to the development of Fourier transform infrared (FTIR) spectrometers and the ATR technique, the spectroscopic method using FTIR equipped with an ATR accessory (FTIR/ATR) provides substantial potential as a quantitative tool based on the molecular structure and on the interactions between the molecule and its environments. In this case, the MIR absorption of water molecules is greatly reduced, and the MIR spectra generated with the ATR method could be used for direct measurement of samples that contain water molecules. The FTIR/ATR technique has been extensively applied in the areas of life sciences, clinical medicine, and others.

Literature first reported a method for infrared spectroscopic identification of β-thalassemia based on hemoglobin samples. Literature proposed a FTIR/ATR quantitative analysis method of thalassemia screening indicators (MCV, MCH) based on hemolysate samples. Their results show that the MIR spectrum can reflect molecule absorbance information of thalassemia gene molysate samples. Their results show a significant correlation between MCV and MCH (refer to the results presented behind), so MCV could indirectly correspond to the absorption of the FTIR/ATR spectrum. Therefore, the three indicators have the mechanism of quantitative analysis by using the FTIR/ATR method. In the present study, we aimed to confirm the feasibility of quantitative analysis of MCH and MCV with FTIR/ATR spectroscopy.

Human blood is a complex system with multiple components; therefore, spectroscopic analysis of some components in human blood must mitigate noise disturbance from its other components. An appropriate spectral wavenumber selection method is important for rapid and chemical-free measurement of a complex system using FTIR/ATR spectroscopy; however, this is a difficult aspect. Therefore, it is essential to improve the effectiveness of spectral prediction, reduce the complexity of the method, and design a specialized spectrometer with a high signal-to-noise ratio (SNR). Furthermore, appropriate chemometric methods must be utilized for optimizing the wavenumber.

Partial least squares (PLS) regression is a popular multivariate calibration method that has been widely applied in multicomponent spectral analysis, particularly in vibrational spectroscopy such as MIR. It could be used to comprehensively screen spectroscopic data, extract information variables, and overcome spectral co-linearity. The moving window PLS (MW-PLS) method is an effective method of spectral analysis and has a high prediction capability. In the present study, the MW-PLS method was improved in terms of stability and equivalence. The appropriate MIR wavebands for the reagent-free simultaneous measurement of Hb, MCH, and MCV in human peripheral blood samples were selected, which correspond to the preliminary thalassemia screening.

The stability of the spectrum analysis model is very important in actual practice. Numerous experiments have shown that differences in the partitioning of calibration and prediction sets can result in fluctuations in predictions and parameters, thereby generating unstable results. In the present study, a rigorous process of calibration, prediction, and validation based on randomness and stability was performed to achieve the goal of spectroscopic analysis. Initially, some samples were randomly selected as a validation set and were not subjected to the modeling optimization process. The remaining samples were used as modeling samples and were further divided several times into calibration and prediction sets. Thereafter, to obtain a stable result, optimal models of the MW-PLS method were selected on the basis of different divisions of the calibration and prediction sets in the modeling array. To solve the restrictions of position and size of wavebands caused by costs and material properties in the instrument design, the set that includes various wavebands that are equivalent to the optimal MW-PLS waveband was then proposed for each indicator. Finally, the selected models were revalidated against the validation set.

2 Materials and Methods

2.1 Experimental Materials, Instruments, and Measurement Methods

In total, 380 human peripheral blood samples were collected and placed in 0.2% ethylenediaminetetraacetic acid-containing tubes. Hb, MCH, and MCV values of these samples were measured via a routine clinical method using a BC-3000Plus blood cell analyzer (Shenzhen Mairui Ke Technology Co., Ltd., China). The data measured by the apparatus were used in the calibration, prediction, and validation sets as reference values for spectroscopic analysis. Statistical analysis of the measured values of the 380 samples for the three indicators Hb, MCH, and MCV.
Table 1 Statistical analysis of measured Hb, mean corpuscular Hb (MCH), and mean corpuscular volume (MCV) values of 380 human peripheral blood samples.

<table>
<thead>
<tr>
<th>Sample types</th>
<th>Number of samples</th>
<th>Hb (gL⁻¹)</th>
<th>MCH (pg)</th>
<th>MCV (fL)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Min</td>
<td>Max</td>
<td>Mean</td>
</tr>
<tr>
<td>All samples</td>
<td>380</td>
<td>61</td>
<td>173</td>
<td>119.4</td>
</tr>
<tr>
<td>Negative</td>
<td>200</td>
<td>61</td>
<td>173</td>
<td>126.2</td>
</tr>
<tr>
<td>Positive</td>
<td>180</td>
<td>74</td>
<td>168</td>
<td>111.9</td>
</tr>
</tbody>
</table>

Note: SD is the abbreviation of standard deviation.

MCV is shown in Table 1. Based on the cut-off values of MCH and MCV, 200 samples were negative and 180 were positive.

Spectra were collected using a VERTEX 70 FTIR Spectrometer (BRUKER Co., Germany) equipped with a KBr beam splitter and a deuterated triglycine sulfate KBr detector. The MIR spectra were obtained from 4000 to 600 cm⁻¹ using a horizontal ATR sampling accessory with a diamond internal reflection element on a zinc selenide crystal (SPECAC Co., UK). Thirty-two scans of symmetrical interferograms at 4 cm⁻¹ resolution were added to each spectrum. The instrument was allowed to purge for several minutes prior to the acquisition of spectra for minimizing the spectral contribution from atmospheric water vapor. Each peripheral blood sample (0.075 mL) was measured thrice, and the mean value of the measurements was used for modeling and validation. The spectra were measured at 25°C ± 1°C and 46% RH; the time of acquisition of an FTIR/ATR spectrum was about 1 min.

2.2 Attenuated Total Reflection

The design of an ATR accessory is based on the principle of internal reflection of light. Infrared light that is emitted by a light source through a crystal with a large refractive index can be projected onto the sample surface using a small refractive index; total reflection occurs when the angle of incidence is greater than the critical angle. Actually, an attenuated evanescent wave is formed on the contact surface; therefore, not all the infrared light is reflected back; however, it partially penetrates to a certain depth beneath the surface of the specimen and then returns to the surface. During this process, the sample selectively absorbs the resulting incident light frequency region as the intensity of the reflected light is decreased, which then generates a spectrogram that is similar to a transmission absorption spectrogram. The penetration depth of the evanescent wave is significantly less than the optical path of ordinary transmission accessories.

As shown in Fig. 1, when the refractive index \( n_1 \) of medium 1 (reflecting element) is greater than the refractive index \( n_2 \) of medium 2 (sample) and the incident angle \( \theta \) is greater than the critical angle \( \theta_c \) (\( \sin \theta_c = n_2 / n_1 \)), the incident light was totally reflected. In fact, infrared light is reflected after penetrated to a certain depth beneath the sample surface. According to Maxwell’s theory, the penetration depth \( d_p \) is defined as follows:

\[
    d_p = \frac{\lambda}{2\pi n_1 \sqrt{\sin^2 \theta - (n_2/n_1)^2}},
\]

where \( \lambda \) is the wavelength of infrared light in the reflected medium. The \( d_p \) value depends on the infrared wavelength, the refractive indexes of reflecting element and sample, and the incident angle. From Eq. (1), we can estimate that \( d_p \) and \( \lambda \) have approximately equal magnitude, whereas the commonly used MIR wavelength \( \lambda \) is ranged from 2.5 to 25 μm.

2.3 Sensitivity and Specificity for Thalassemia

Sensitivity and specificity are two evaluation indicators in clinical diagnostic tests that are used for quantitative recognition of patients and nonpatients. In the present study, a new approach for quantitative analysis of thalassemia based on the simultaneous determination of Hb, MCH, and MCV using FTIR/ATR spectra was established. Therefore, the sensitivity and specificity values of the new approach must be evaluated using the existing standard clinical biochemistry method.

According to the cut-off values of MCH and MCV for thalassemia, the number of true-positive, false-negative, false-positive, and true-negative samples was \( a, b, c, \) and \( d \), respectively. Therefore, the sensitivity and specificity values of the spectroscopic analysis method used in this study were calculated as follows:

\[
    \text{Sensitivity} = \frac{a}{a + b} \times 100\%;
\]
\[
    \text{Specificity} = \frac{d}{c + d} \times 100\%.
\]  

2.4 Sample Set Division and the Calibration, Prediction, and Validation Process

A new framework was developed for the calibration, prediction, and validation process and sample division on the basis of randomness and stability. Some samples were randomly selected.
from all samples as validation samples and were not subjected to
the modeling optimization process. The remaining samples were
used as modeling samples and were divided multiple times into
calibration and prediction sets. For obtaining objective and sta-
table results, calibration and prediction models were established
for all divisions of the sample sets, and model parameters were
optimized depending on the mean prediction effects of all
divisions.

All the calibration, prediction, and validation sets must con-
tain negative and positive samples to ensure modeling repre-
sentativeness and integrity. Therefore, the negative and pos-
tive samples must be divided into calibration, prediction, and
validation sets. The following specific procedure was fol-
lowed. First, 64 of the 200 negative samples were randomly
selected as the validation set. The remaining 136 samples
were used as modeling samples and were further randomly
divided 200 times into calibration (68 samples) and prediction (68
samples) sets. Second, 86 of the 180 positive samples were ran-
domly selected as the validation set. The remaining 94 samples
were used as modeling samples and were also further randomly
divided 200 times into calibration (47 samples) and prediction (47
samples) sets. Finally, the positive and negative samples
used for validation were merged into a single validation set
(150 samples). Similarly, the positive and negative samples
used for calibration and prediction were merged into whole cal-
boration (115 samples) and prediction (115 samples) sets for
each division, respectively. Figure 2 shows the type and number
of samples in the calibration, prediction, and validation sets.

Calibration and prediction were performed for each division
i, i = 1, 2, · · · , 200. The root mean square error and the corre-
lation coefficients for prediction in the modeling set are denoted
as MSEP, and MRp,i, respectively. Calculation formulas are as
follows:

\[
M_{\text{SEP}}_i = \frac{\sqrt{\sum_{k=1}^{n} \left[ C_k^{(i)} - \bar{C}_k^{(i)} \right]^2}}{n - 1},
\]

\[
M_{\text{R}_p,i} = \frac{\sqrt{\sum_{k=1}^{n} \left[ C_k^{(i)} - \bar{C}_k^{(i)} \bar{C}_k^{(i)} - \bar{C}_k^{(i)} \bar{C}_k^{(i)} \right]^2}}{\sqrt{\sum_{k=1}^{n} \left[ C_k^{(i)} - \bar{C}_k^{(i)} \right]^2 \sum_{k=1}^{n} \left[ C_k^{(i)} - \bar{C}_k^{(i)} \right]^2}},
\]

where \( n \) was the number of prediction samples; \( C_k^{(i)} \) and \( \bar{C}_k^{(i)} \)
were the measured and predicted values for \( i \)’th division and
\( k \)’th prediction sample, respectively; \( \bar{C}_k^{(i)} \) and \( \bar{C}_k^{(i)} \)
were the mean measured value and mean predicted value of all prediction
samples for \( i \)’th division, respectively.

The mean value and the standard deviation (SD) of the
MSEP, and MRp,i of all the divisions were denoted as
MSEP,Ave, MRp,Ave, MSEP,SD, and MRp,SD, respectively. Calculation formulas are as follows:

\[
M_{\text{SEP}}_\text{Ave} = \frac{\sum_{i=1}^{200} M_{\text{SEP}}_i}{200},
\]

\[
M_{\text{R}_p,\text{Ave}} = \frac{\sum_{i=1}^{200} M_{\text{R}_p,i}}{200},
\]

\[
M_{\text{SEP}}_\text{SD} = \frac{\sqrt{\sum_{i=1}^{200} \left[ (M_{\text{SEP}}_i) - (M_{\text{SEP}}_\text{Ave}) \right]^2}}{200 - 1},
\]

\[
M_{\text{R}_p,\text{SD}} = \frac{\sqrt{\sum_{i=1}^{200} \left[ (M_{\text{R}_p,i}) - (M_{\text{R}_p,\text{Ave}}) \right]^2}}{200 - 1}.
\]

These values were used for evaluation of modeling prediction
accuracy and stability. The equation

\[
M_{\text{SEP}}^+ = M_{\text{SEP}}_\text{Ave} + M_{\text{SEP}}_\text{SD},
\]

was used as a comprehensive indicator of modeling prediction
accuracy and stability of a model. A smaller value of \( M_{\text{SEP}}^+ \)
indicated higher accuracy and stability of the model. The model
parameters were selected for achieving a minimum \( M_{\text{SEP}}^+ \).
The selected model was then revalidated against the validation
set. The root mean square error and correlation coefficients of
prediction in the validation set were then calculated and denoted
as VSEP and VRp, respectively. The calculation formulas are as
follows:

\[
V_{\text{SEP}} = \sqrt{\frac{\sum_{k=1}^{m} \left( \bar{C}_k - \bar{C}_k \right)^2}{m - 1}},
\]

\[
V_{\text{R}_p} = \frac{\sqrt{\sum_{k=1}^{m} \left( \bar{C}_k - \bar{C}_k \right)^2 \sum_{k=1}^{m} \left( \bar{C}_k - \bar{C}_k \right)^2}}{\sqrt{\sum_{k=1}^{m} \left( \bar{C}_k - \bar{C}_k \right)^2}},
\]

where \( m \) was the number of validation samples; \( \bar{C}_k \) and \( \bar{C}_k \)
were the measured and predicted values of \( k \)’th validation samples;
\( \bar{C}_k \) and \( \bar{C}_k \) were the mean measured values and mean pre-
dicted values of all validation samples.

Quantitative analyses of Hb, MCH, and MCV were inde-
pendently performed on the basis of the same modeling process
mentioned earlier. The selections of wavebands for three indica-
tors were obtained independently.

2.5 Optimization Frame of the Improved MW-PLS
Method with Stability

For the MW-PLS method, consecutive spectral data on \( N \)
adjacent wavenumbers were designated as a window. For all
the windows in a predetermined search region of the spectrum,
PLS models were established, and the optimal analytical wave-
bands were selected by moving and varying the window size
(see also Fig. 2). By considering the position and length of
the wavebands as well as the PLS factor, the search parameters
were set as follows: (1) initial wavenumber \( I \), (2) number of
wavenumbers \( N \), and (3) number of PLS factors \( F \). The
search range of the parameters \( I, N, \) and \( F \) can be selected.
according to the actual chemical, physical, and statistical significance. PLS models can be established for any combination of \((I, N, \text{ and } F)\) depending on different divisions of calibration and prediction sets. The corresponding \(M_{\text{SEP}_{\text{AVE}}}, M_{\text{RP}_{\text{AVE}}}, M_{\text{SEP}_{\text{SD}}}, M_{\text{RP}_{\text{SD}}}, \text{ and } M_{\text{SEP}^+}\) values were then calculated. For achieving stable results, the optimal waveband with minimum \(M_{\text{SEP}^+}\) was selected.

### 2.6 Selection of the Number of PLS Factors for Stability

PLS regression can comprehensively screen spectroscopic data, extract information variables, and overcome spectral colinearity. The number of PLS factors \((F)\) is an important parameter that corresponds to the number of integrated spectral variables corresponding to sample information. The selection of a reasonable \(F\) is necessary as well as difficult. In the present study, \(F\) was selected by considering the number of divisions of the calibration and prediction sets. Thus, the optimum number of PLS factors exhibited stability and practicality. Each waveband corresponded to a unique combination of \((I, N) = (I_0, N_0)\); the optimal PLS model of the waveband was selected according to the following expression:

\[
M_{\text{SEP}^+}(I_0, N_0) = \min_{I, N} M_{\text{SEP}^+}(I_0, N_0, F),
\]

and the corresponding \(M_{\text{RP}_{\text{AVE}}}, M_{\text{SEP}_{\text{SD}}}, M_{\text{RP}_{\text{SD}}}, \text{ and } M_{\text{SEP}^+}\) values were determined.

### 2.7 Global Optimal Model

The global optimal model was selected according to the following equation:

\[
M_{\text{SEP}^+} = \min_{I, N, F} M_{\text{SEP}^+}(I, N, F),
\]

and the corresponding \(M_{\text{RP}_{\text{AVE}}}, M_{\text{SEP}_{\text{SD}}}, M_{\text{RP}_{\text{SD}}}, \text{ and } M_{\text{SEP}^+}\) values were determined.

### 2.8 Local Optimal Model Corresponded to a Single Parameter

The instrument design typically involves some limitations of position and number of wavenumbers (such as costs and material properties). At some instances, the demand of actual conditions is not met by the global optimal waveband. Therefore, local optimal wavebands that correspond to different positions and the number of wavelengths are significant. For any fixed \(I = I_0\), the local optimal model was selected according to the following equation:

\[
M_{\text{SEP}^+}(I_0) = \min_{N, F} M_{\text{SEP}^+}(I_0, N, F),
\]

and the corresponding \(M_{\text{RP}_{\text{AVE}}}, M_{\text{SEP}_{\text{SD}}}, M_{\text{RP}_{\text{SD}}}, \text{ and } M_{\text{SEP}^+}\) values were determined. Meanwhile, for any fixed \(N = N_0\), the local optimal model was selected according to the following expression:

\[
M_{\text{SEP}^+}(N_0) = \min_{I, F} M_{\text{SEP}^+}(I, N_0, F),
\]

and the corresponding \(M_{\text{RP}_{\text{AVE}}}, M_{\text{SEP}_{\text{SD}}}, M_{\text{RP}_{\text{SD}}}, \text{ and } M_{\text{SEP}^+}\) values were determined.

In the present study, the search range for the MW-PLS method spanned the entire scanning region of 4000 to 600 cm\(^{-1}\) with 1764 wavenumbers. Furthermore, \(I\) and \(F\) were set as follows: \(I \in \{4000, 3998, \ldots, 600\}\) and \(F \in \{1, 2, \ldots, 30\}\) for Hb, MCH, and MCV. To reduce the workload and maintain representativeness, \(N\) was set as follows: \(N_{\text{Hb}} \in \{1, 2, \ldots, 150\} \cup \{170, 190, \ldots, 1730\} \cup \{1764\}, N_{\text{MCH}} \in \{1, 2, \ldots, 150\} \cup \{170, 190, \ldots, 270\} \cup \{271, 272, \ldots, 430\} \cup \{450, 470, \ldots, 1730\} \cup \{1764\}, N_{\text{MCV}} \in \{1, 2, \ldots, 150\} \cup \{170, 190, \ldots, 270\} \cup \{271, 272, \ldots, 430\} \cup \{450, 470, \ldots, 1730\} \cup \{1764\}\) for Hb, MCH, and MCV, respectively. With \((I, N) = (1717, 108)\) as an example, Fig. 3 shows a schematic diagram for the waveband and screening mode for moving windows.

The computer algorithms for the abovementioned method were designed using the MATLAB 7.6 version software.

### 3 Results and Discussion

#### 3.1 Global Optimal Models Using the MW-PLS Method

The FTIR/ATR spectra of the 380 human peripheral blood samples on the overall scanning region of 4000 to 600 cm\(^{-1}\) are shown in Fig. 4. For comparison, the spectrum of distilled water is represented by a dotted line in Fig. 4. Figure 4 had a red dotted line and 380 black solid lines, where the red dotted line was the spectrum of distilled water, and the black solid lines were the spectra of 380 samples of human peripheral blood. Figure 4 shows that the water molecules had an absorption within the range of 3290 and 1637 cm\(^{-1}\), the absorption of hemoglobin composition and other blood components mainly

![Fig. 3 Schematic diagram for the waveband and screening mode for moving windows.](image1)

![Fig. 4 FTIR/ATR spectra of 380 human peripheral blood samples in the entire scanning range (4000 to 600 cm\(^{-1}\)).](image2)
appeared within the MIR fingerprint region 1800 to 800 cm⁻¹, and different samples had significantly different ranges of absorption in the fingerprint region.

Using the previously mentioned method, PLS models for Hb, MCH, and MCV were first established on the basis of the entire scanning region. The prediction accuracy and stability results ($M_{\text{SEP Ave}}, M_{\text{RP Ave}}, M_{\text{SEP SD}}, M_{\text{RP SD}},$ and $M_{\text{SEP}^+}$) are summarized in Table 2. The results show that the predicted and clinically measured values have a certain correlation for each indicator. The number of wavenumbers employed was 1764, and the models showed high complexity. Further waveband optimization was performed using the MW-PLS method to improve the prediction accuracy and to reduce complexity. Depending on the minimum $M_{\text{SEP}^+}$ value, the optimal MW-PLS models were selected for Hb, MCH, and MCV. The corresponding parameters $I$, $N$, and $F$ and the prediction effects are summarized in Table 2. The results show that the optimal values of $I$ and $N$ are 1722 cm⁻¹ and 114 for Hb, 1653 cm⁻¹ and 391 for MCH, and 1562 cm⁻¹ and 311 for MCV, respectively. The corresponding waveband intervals were 1722 to 1504 cm⁻¹ for Hb, 1653 to 901 cm⁻¹ for MCH, and 1562 to 964 cm⁻¹ for MCV, all of which were within the MIR fingerprint region. The number of wavenumbers for the three wavebands was all less than one-fourth of that for the entire scanning region. Therefore, the model complexity was reduced for each indicator. Table 2 shows that the five values of the prediction effects ($M_{\text{SEP Ave}}, M_{\text{RP Ave}}, M_{\text{SEP SD}}, M_{\text{RP SD}},$ and $M_{\text{SEP}^+}$) of the optimal MW-PLS models were all significantly better than those of the overall scanning region for each indicator. Thus, the prediction accuracy and stability of the optimal MW-PLS models were significantly improved.

### 3.2 Local Optimal Models

$M_{\text{SEP}^+}$ values of the local optimal models, which correspond to each $I$ and $N$, are shown in Figs. 3(a) and 4(a) for the three indicators. Figures 3(a) and 4(a) show the minimum $M_{\text{SEP}^+}$ achieved for Hb when $I = 1722$ (cm⁻¹) and $N = 114$. Figures 3(b) and 4(b) show the minimum $M_{\text{SEP}^+}$ achieved for MCH when $I = 1653$ (cm⁻¹) and $N = 391$. Figures 3(c) and 4(c) show the minimum $M_{\text{SEP}^+}$ achieved for MCV when $I = 1562$ (cm⁻¹) and $N = 311$. The results indicate that these parameters have the best prediction accuracy and stability. These data may serve as a valuable reference for designing the splitting system of spectroscopic instruments. Some local optimal models whose prediction parameters are close to that of the global optimal model remain a good choice. These models address restrictions such as cost and material properties as well as the position and number of wavenumbers in the instrument design.

### 3.3 Model Set with Equivalence

As mentioned, the optimal MW-PLS wavebands for Hb, MCH, and MCV were screened according to the minimum $M_{\text{SEP}^+}$, and the obtained minimum $M_{\text{SEP}^+}$ values were 7.96 g L⁻¹ for Hb, 2.309 pg for MCH, and 5.433 fL for MCV. However, statistically speaking, because modeling samples are random and limited, the models with slightly fluctuating prediction accuracy are considered equivalent. Therefore, the optimal $M_{\text{SEP}^+}$ could float upward at an appropriate range (take 0.6%, for example, in the present report; the upward range could be adjusted accordingly).

For the Hb indicator, the minimum $M_{\text{SEP}^+}$ floated upward from 7.96 to 8.00 g L⁻¹; the model set included 75 wavebands that were equivalent to the optimal MW-PLS waveband, all of which were within the MIR fingerprint region, and the corresponding min($N$) and max($N$) were 108 and 122, respectively. The public range of the 75 equivalent wavebands was from 1717 to 1510 cm⁻¹, with 108 wavenumbers, and just one of the equivalent wavebands had the lowest $N$. Thus, the waveband 1717 to 1510 cm⁻¹ can be employed instead of other

### Table 2. Modeling prediction accuracy and stability of PLS models based on the entire scanning region, the optimal MW-PLS wavebands, and the equivalent wavebands for Hb, MCH, and MCV.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Waveband (cm⁻¹)</th>
<th>$N$</th>
<th>$F$</th>
<th>$M_{\text{SEP Ave}}$</th>
<th>$M_{\text{SEP SD}}$</th>
<th>$M_{\text{RP Ave}}$</th>
<th>$M_{\text{RP SD}}$</th>
<th>$M_{\text{SEP}^+}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entire scanning region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hb</td>
<td>4000 – 600</td>
<td>1764</td>
<td>22</td>
<td>10.3</td>
<td>0.9</td>
<td>0.876</td>
<td>0.023</td>
<td>11.2</td>
</tr>
<tr>
<td>MCH</td>
<td>30</td>
<td>2.75</td>
<td>0.29</td>
<td>0.852</td>
<td>0.032</td>
<td>3.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCV</td>
<td>30</td>
<td>6.74</td>
<td>0.59</td>
<td>0.832</td>
<td>0.030</td>
<td>7.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimal MW-PLS wavebands</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hb</td>
<td>1722 – 1504</td>
<td>114</td>
<td>6</td>
<td>7.5</td>
<td>0.4</td>
<td>0.932</td>
<td>0.008</td>
<td>8.0</td>
</tr>
<tr>
<td>MCH</td>
<td>1653 – 901</td>
<td>391</td>
<td>11</td>
<td>2.17</td>
<td>0.14</td>
<td>0.904</td>
<td>0.013</td>
<td>2.31</td>
</tr>
<tr>
<td>MCV</td>
<td>1562 – 964</td>
<td>311</td>
<td>10</td>
<td>5.09</td>
<td>0.35</td>
<td>0.899</td>
<td>0.014</td>
<td>5.43</td>
</tr>
<tr>
<td>Equivalent wavebands</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hb</td>
<td>1717 – 1510</td>
<td>108</td>
<td>6</td>
<td>7.6</td>
<td>0.4</td>
<td>0.931</td>
<td>0.009</td>
<td>8.0</td>
</tr>
<tr>
<td>MCH</td>
<td>1562 – 901</td>
<td>344</td>
<td>10</td>
<td>2.17</td>
<td>0.15</td>
<td>0.904</td>
<td>0.013</td>
<td>2.32</td>
</tr>
<tr>
<td>MCV</td>
<td>10</td>
<td>5.09</td>
<td>0.35</td>
<td>0.900</td>
<td>0.014</td>
<td>5.44</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Journal of Biomedical Optics 087004-6 August 2014 • Vol. 19(6)
equivalent wavebands, which contained enough information on Hb.

For the MCH indicator, the minimum $M_{\text{SEP}^+}$ floated upward from 2.309 to 2.323 pg; the model set included 79 wavebands that were equivalent to the optimal MW-PLS waveband, all of which were within the MIR fingerprint region. The $\min(N)$ and $\max(N)$ were 343 and 403; the $\min(I)$ and $\max(I)$ were 1562 and 1674 cm$^{-1}$; and the minimum and maximum ending wavenumber were 874 and 903 cm$^{-1}$, respectively.

For the MCV indicator, the minimum $M_{\text{SEP}^+}$ floated upward from 5.433 to 5.466 fl; the model set included 53 wavebands that were equivalent to the optimal MW-PLS waveband, all of which were within the MIR fingerprint region. The $\min(N)$ and $\max(N)$ were 343 and 403; the $\min(I)$ and $\max(I)$ were 300 and 480 cm$^{-1}$; and the minimum and maximum ending wavenumber were 947 and 1320 cm$^{-1}$, respectively.

Fig. 5 Optimal $M_{\text{SEP}^+}$ corresponding to initial wavenumber for (a) Hb, (b) mean corpuscular Hb (MCH), and (c) mean corpuscular volume (MCV).

Fig. 6 Optimal $M_{\text{SEP}^+}$ corresponding to number of wavenumbers for (a) Hb, (b) MCH, and (c) MCV.

Fig. 7 Relationship between the measured values of MCH and MCV for 380 human peripheral blood samples.
min(N) and max(N) were 303 and 347; the min(I) and max(I) were 1562 and 1574 cm\(^{-1}\); and the minimum and maximum ending wavenumbers were 897 and 980 cm\(^{-1}\), respectively.

Interestingly, the equivalent model sets of MCH and MCV showed the same waveband range of 1562 to 901 cm\(^{-1}\); thus, this waveband range can be used for the simultaneous quantification of MCH and MCV. In fact, the present study aimed to confirm the feasibility of quantitatively analyzing Hb, MCH, and MCV using FTIR/ATR spectroscopy. FTIR/ATR spectroscopy cannot be used for direct measurement of MCV, which refers to the average volume of individual erythrocytes. However, clinical examination results show a significant correlation between MCV and MCH; the obtained correlation coefficients (R) of the measured MCH and MCV values for all 380 human peripheral blood samples reached 0.975 (Fig. 8). The correlation between MCV and MCH may be utilized for quantitative analysis of MCV using FTIR/ATR spectroscopy.

Therefore, it is reasonable to select the public equivalent waveband range of 1562 to 901 cm\(^{-1}\) for MCH and MCV.

The public waveband (1717 to 1510 cm\(^{-1}\)) of Hb equivalent wavebands and the public equivalent waveband (1562 to 901 cm\(^{-1}\)) of MCH and MCV were used as the typical examples of equivalent model sets; the corresponding prediction accuracy and stability are shown in Table 2. The positions of all equivalent wavebands for Hb, MCH, and MCV and the corresponding \(M_{SEP}^{+}\) are shown in Fig. 8.

The equivalent model set provided various waveband selections and circumvented the restrictions of position and size of wavebands caused by costs and material properties in the instrument design.
3.4 Model Validation

The randomly selected validation samples, which were excluded in the modeling optimization process, were used for validating the optimal MW-PLS wavebands (1722 to 1504 cm\(^{-1}\) for Hb, 1653 to 901 cm\(^{-1}\) for MCH, and 1562 to 964 cm\(^{-1}\) for MCV) as well as the public waveband (1717 to 1510 cm\(^{-1}\)) of the Hb equivalent wavebands and the public equivalent waveband (1562 to 901 cm\(^{-1}\)) of MCH and MCV. The regression coefficients were calculated using the spectral data and clinically measured values of the entire modeling set depending on the corresponding parameters. The predicted values of the validation samples were then calculated using the obtained regression coefficients and the spectra of the validation samples.

The relationship between the predicted and clinically measured values of the 150 validation samples for Hb, MCH, and MCV is shown in Figs. 9 and 10, respectively. The evaluation values for validation (\(V_{\text{SEP}}\) and \(V_{\text{RP}}\)) are summarized in Table 3. The results indicate that the six cases have high validation prediction accuracy. The prediction values of Hb, MCH, and MCV of the validation samples are close to those of the clinically measured values. Satisfactory validation effects were achieved for the random validation samples because stability was considered in the modeling optimization process.

The classification of the negative and positive samples for thalassemia can be observed in the two-dimensional (2-D) diagram of (Hb, MCH) and (Hb, MCV). Among the 150 validation samples, 64 are negative and 86 are positive based on their clinical measured values and the cut-off line (\(\text{MCH} = 27.0; \text{MCV} = 80.0\); Fig. 11). Figures 12 and 13 show the 2-D diagram of the FTIR/ATR predicted values of the 150 validation samples, corresponding to the optimal MW-PLS wavebands and the equivalent wavebands, respectively. The sensitivity and specificity were 100.0% and 96.9% for the optimal MW-PLS wavebands and 100.0% and 95.3% for the equivalent wavebands, respectively. The sensitivity and specificity values for two cases were high, and the prediction errors were distributed primarily around the cut-off line. The region neighboring the cut-off line is blurry, and a few prediction errors in this region are understandable. The results also confirmed the feasibility of negative and positive screening of thalassemia samples that feature microcytic hypochromic anemia.

### Table 3 Validation effects of PLS models based on the optimal MW-PLS wavebands and the equivalent wavebands for Hb, MCH, and MCV.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Waveband (cm(^{-1}))</th>
<th>(N)</th>
<th>(F)</th>
<th>(V_{\text{SEP}})</th>
<th>(V_{\text{RP}})</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Optimal MW-PLS wavebands</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hb</td>
<td>1722 – 1504</td>
<td>114</td>
<td>6</td>
<td>7.7</td>
<td>0.928</td>
<td>100.0%</td>
<td>96.9%</td>
</tr>
<tr>
<td>MCH</td>
<td>1653 – 901</td>
<td>391</td>
<td>11</td>
<td>2.28</td>
<td>0.900</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCV</td>
<td>1562 – 964</td>
<td>311</td>
<td>10</td>
<td>5.14</td>
<td>0.906</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Equivalent wavebands</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hb</td>
<td>1717 – 1510</td>
<td>108</td>
<td>6</td>
<td>7.6</td>
<td>0.928</td>
<td>100.0%</td>
<td>95.3%</td>
</tr>
<tr>
<td>MCH</td>
<td>1562 – 901</td>
<td>344</td>
<td>10</td>
<td>2.22</td>
<td>0.901</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCV</td>
<td></td>
<td>10</td>
<td>5.25</td>
<td>0.901</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4 Conclusions

The main objective of the present study was to conduct preliminary thalassemia screening. A novel method for simultaneous quantitative analysis of Hb, MCH, and MCV was proposed using FTIR/ATR spectroscopy. In addition, a rigorous process of calibration, prediction, and validation based on randomness and stability was performed to generate objective and stable models.

The key technology used in this study is the high SNR of the waveband selection technique that corresponds to the information on Hb, MCH, and MCV. The selections of appropriate wavebands were accomplished using the improved MW-PLS method in terms of stability and equivalence on the basis of different divisions of calibration and prediction sets. The obtained optimal MW-PLS wavebands were 1722 to 1504 cm$^{-1}$ for Hb, 1653 to 901 cm$^{-1}$ for MCH, and 1562 to 964 cm$^{-1}$ for MCV. A model set equivalent to the optimal MW-PLS model was proposed for each indicator; the obtained public waveband for the Hb equivalent wavebands was 1717 to 1510 cm$^{-1}$, and the public equivalent waveband of MCH and MCV was 1562 to 901 cm$^{-1}$. All the waveband selections were within the MIR fingerprint region, and their models achieved high validation effects. The results confirm the feasibility of simultaneous quantitative analysis of Hb, MCH, and MCV using FTIR/ATR spectroscopy. The sensitivity and specificity were 100.0% and 96.9% for the optimal MW-PLS wavebands and 100.0% and 95.3% for the equivalent wavebands, respectively. Thus, spectral prediction was also highly accurate for determining negative and positive samples for thalassemia screening.

This is a reagent-free and accurate determination technique that is simple and rapid in comparison with the conventional methods; thus, it is a promising tool for screening, diagnosing, preventing, and controlling thalassemia in large populations. The strategies used for the development of this method may be useful in studying other important parameters of thalassemia. The waveband selections also provide valuable references for designing specialized spectrometers.

**Fig. 11** 2-D diagrams for the clinical measured values of validation samples classified as negative and positive: (a) (Hb, MCH) and (b) (Hb, MCV).

**Fig. 12** 2-D diagrams for the FTIR/ATR predicted values of validation samples classified as negative and positive on the basis of PLS models with the optimal MW-PLS wavebands: (a) (Hb, MCH) and (b) (Hb, MCV).

**Fig. 13** 2-D diagrams for the FTIR/ATR predicted values of validation samples classified as negative and positive on the basis of PLS models with the equivalent wavebands: (a) (Hb, MCH) and (b) (Hb, MCV).
Acknowledgments
This work was supported by National Natural Science Foundation of China (No. 61078040), the Science and Technology Project of Guangdong Province of China (No. 2012B031800917), and the Science and Technology Project of Guangzhou of China (No. 2014Y2-00002).

References

Xiaoli Long is a master’s student in optoelectronic engineering from the Department of Optoelectronic Engineering at Jinan University, Guangzhou, China.

Guisong Liu is a master’s student in applied mathematics from the Department of Mathematics at Jinan University, Guangzhou, China.

Tao Pan is a professor in the Department of Optoelectronic Engineering at Jinan University. He received his BS degree in mathematics from Sichuan University, China, and his PhD degree in biologic information engineering from Mie University, Japan. He is director of the Applied Spectroscopy Laboratory at Jinan University, College of Science and Engineering. He is engaged in studies of spectroscopy, biomedical information, chemometrics, pattern recognition, and partial differential equations. He has published more than 80 peer-reviewed papers.

Jiemei Chen is associate professor in the Department of Biological Engineering at Jinan University. She received her BS and MS degrees in microbiology from Sichuan University and Guangxi University, China, and her PhD degree in biology from Mie University, Japan. She is engaged in studies of microbiology, spectroscopy, biomedical information, and so forth. She has published more than 50 peer-reviewed papers.