Chemometric determination of the length distribution of single walled carbon nanotubes through optical spectroscopy

Rongmei Si^{a,1}, Ke Wang^{a,1}, Tao Chen^{a,b,*}, Yuan Chen^{a,*}

^a School of Chemical and Biomedical Engineering, Nanyang Technological University, 62 Nanyang Drive, Singapore 637459, Singapore

^b Division of Civil, Chemical and Environmental Engineering, University of Surrey, Guildford GU2 7XH, UK

Abstract

Current synthesis methods for producing single walled carbon nanotubes (SWCNTs) do not ensure uniformity of the structure and properties, in particular the length, which is an important quality indicator of SWCNTs. As a result, sorting SWCNTs by length is an important post-synthesis processing step. For this purpose, convenient analysis methods are needed to characterize the length distribution rapidly and accurately. In this study, density gradient ultracentrifugation was applied to prepare length-sorted SWCNT suspensions containing individualized surfactant-wrapped SWCNTs. The length of sorted SWCNTs was first determined by atomic force microscope (AFM), and their absorbance was measured in ultraviolet-visible near-infrared (UV-vis-NIR) spectroscopy. Chemometric methods are used to calibrate the spectra against the AFM-measured length distribution. The calibration model enables convenient analysis of the length distribution of SWCNTs through UV-vis-NIR spectroscopy. Various chemometric techniques are investigated, including pre-processing methods and non-linear calibration models. Extended inverted signal correction, extended multiplicative signal correction and Gaussian process regression are found to provide good prediction of the length distribution of SWCNTs with satisfactory agreement with the AFM measurements. In summary, spectroscopy in conjunction with advanced chemometric techniques is a powerful analytical tool for carbon nanotube research.

Keywords: Chemometrics; Density gradient ultracentrifugation; Jensen-Shannon divergence; Length distribution; Multivariate calibration; Single walled carbon nanotube.

1. Introduction

Single walled carbon nanotubes (SWCNTs) are hollow cylinders with one atomic layer of graphene wall, with diameter around 1 nm and length ranging from tens of nanometers to centimeters. SWCNTs have versatile applications including transistors, conductive films, sensors, probes, mechanical reinforcements, hydrogen storage and catalytic supports because of their excellent electrical, mechanical, thermal, optical and biological properties [1]. The macroscopic properties of SWCNTs depend on the underlying atomic characteristics, such as diameter, chirality and length [2]. Current synthesis methods do not ensure the uniformity of the structure and properties of SWCNTs, a fact that has hindered various potential applications. To overcome this barrier, significant research efforts have been dedicated to sorting SWCNTs according to the diameter, metallicity and chirality [2-4]. Separation by length is especially useful because the electronic, optical and chemical properties of SWCNTs are significantly affected by the length [5-7]. Length sorting may

¹ These authors contributed equally to the presented work.

* Corresponding authors. Tel.: +44 1483686593; Fax: +44 1483686581; Email: t.chen@surrey.ac.uk (T. Chen). Tel: +65 63168939; Fax: +65 67947553; Email: chenyan@ntu.edu.sg (Y. Chen).

also be necessary as a preliminary step to obtain high purity monodisperse samples [8, 9]. Furthermore, toxicity of SWCNTs is often associated with their length. Length-dependent SWCNT uptake by cells has been observed [10].

In recent years, several techniques have been developed to sort SWCNTs by length, including size exclusion chromatography [6, 11], gel electrophoresis, capillary electrophoresis, shear-aligned photoluminescence anisotropy [8], field flow fractionation [12], and density gradient ultracentrifuge (DGU) [13-15]. Separation of SWCNTs by length using centrifugation in a density gradient medium was developed by Fagan *et al.* [14]. The key advantages of DGU originate from its tenability in operating parameters and its scalability, because both the density of SWCNTs and the density gradient of medium can be easily adjusted to yield different separation selectivity [16-20]. A common need for SWCNT length sorting is to determine the length distribution accurately, because to date, no separation method has been able to yield samples with exactly one length. Transmission electron microscope (TEM) and atomic force microscope (AFM) are the most widely used methods to quantify the length distribution of SWCNTs. In microscopic studies, well-dispersed SWCNT suspensions are deposited on substrates, and hundreds of nanotubes are counted individually to gain reliable statistic information. The sample preparation for TEM and AFM is often time-consuming and tedious. Visual inspection and image processing also introduce errors. Particularly, only a very small fraction of samples can be inspected in microscopic studies. Dynamic light scattering (DLS) has been used to quantify the length of bulk SWCNT dispersions. However, lengths were usually calculated by assuming nanotubes to be thin rigid rods in DLS [6]. Because of their large aspect ratios (100 to 10000) and mechanical flexibility, SWCNTs may curl up in liquid suspensions, resulting in inaccurate length prediction.

In addition, Fagan *et al.* observed the length-dependent optical effects in SWCNTs [6]. In general, when a light beam encounters particles, such as nanotubes, both scattering and absorption can consume energy, a phenomenon known as light extinction [21]. The law of conservation of energy requires that $C_{\text{ext}} = C_{\text{sca}} + C_{\text{abs}}$, where C_{ext} , C_{sca} , and C_{abs} are the cross sections of the particle from extinction, scattering, and absorption, respectively [21]. Having the dimension of area, the cross sections reflect the probability for an event to occur. When a particle's radius (r) is comparable to the wavelength of light (λ), the Mie theory [21] indicates that C_{sca} varies in the order of r^6 , while C_{abs} varies in the order of r^3 only. Therefore, for nano-scale particles, absorption is predominant over scattering [22]. This has been reported in both experimental and theoretical studies [22]. For instance, the $C_{\text{sca}}/C_{\text{abs}}$ ratio of gold nanospheres with radius of 10 nm under 521 nm light is only 0.01 [23]. A recent study showed that scattering from larger diameter multiple walled carbon nanotubes (MWCNTs) is minor (<8%) in light extinction over the wavelength range of 300-1300 nm [24]. Thus, it is reasonable to hypothesize that absorption spectroscopy may be an effective tool to measure the length of SWCNTs. Indeed, Fagan *et al.* showed that the optical absorption of SWCNTs has a nearly linear dependence on the mean (average) length [6]. The absorption peak intensity at optical resonance (absorbance at 984 nm) verses absorbance baseline (at 775 nm) was found to scale almost linearly with the mean length of nanotubes up to 1 μm (note that the diameter of SWCNTs is around 1 nm only). A simple physical argument was also proposed to relate the length dependence in SWCNT absorption to the localization of a bound exciton along the backbone of the nanotube, suggesting a quantitative link between nanotube length and oscillator strength for quasi-one-dimensional ordered nanostructures [6]. In the Rayleigh regime, the bound exciton is localized on a length scale much smaller than the length of the nanotube, implying that the imaginary part of the SWCNT dielectric response function is proportional to nanotubes' length. Later, an empirical formula, given below, was applied to predict the mean length of SWCNTs obtained from length sorting based on their optical absorbance [14]:

$$l(\text{nm}) = \left(\frac{\text{Absorbance (984 nm)}}{\text{Absorbance (775 nm)}} - 0.842 \right) \times 160.4 \text{ nm} \quad (1)$$

However, because of overlapping of optical absorbance from SWCNTs at different lengths, predicting length distribution was not explored in the literature.

Multivariate chemometric calibration methods are to utilize measured spectra to predict the corresponding response variables (the SWCNT length distribution in this work). For this purpose, partial least squares (PLS) regression is a widely used linear calibration method. However, non-linear variations can be induced to the spectra because of external disturbances, mainly baseline shift/noise in this study. This is usually the result of the difference in sample homogeneity, the temperature variation during the analysis, and so on, giving rise to non-linear spectral data. In the literature, baseline and other noises can be addressed by using pre-processing methods and/or non-linear calibration models [25]. Typical pre-processing methods to remove non-linearity in the spectral data include first (D1) and second derivatives (D2) [26, 27], standard normal variate (SNV) [28], extended multiplicative signal correction (EMSC) [29], and extended inverted signal correction (EISC) [30]. Alternatively, non-linear calibration methods, such as artificial neural network (ANN) [31-33], and Gaussian process regression (GPR) [34], are capable of directly modelling the non-linearity in the spectra.

In this contribution, we used DGU to prepare length-sorted SWCNT suspensions containing individualized surfactant wrapped SWCNTs. The length distribution of the sorted SWCNTs was first determined by AFM and it is used as the reference value for calibration. Their absorbance was then measured in UV-vis-NIR absorption spectrometer. An existing method [14] to determine the mean length of SWCNTs from absorption spectroscopy is extended in this work to predict the entire length distribution. In addition, advanced multivariate chemometric methods are adopted and compared for the calibration purpose, in contrast to using an empirical formula (eq. (1)) in the previously reported work [14]. The results showed that satisfactory prediction accuracy can be achieved by using chemometric techniques. In summary, absorption spectroscopy in conjunction with proper calibration technology is an effective analytical tool to complement the time-consuming, laborious and expensive AFM (or TEM) analysis for SWCNT research.

2. Experimental

SWCNTs (8, 10, and 12 mg, respectively) synthesized by CO decomposition on cobalt-molybdenum catalysts (Sigma-Aldrich, batch number 94496KJ) were dispersed in 10 mL of 2 wt% aqueous sodium deoxycholate (DOC) solution. The suspensions were sonicated in an ultrasonicator (SONICS, VCX-130) for 1 hr at 30 W in an ice bath, prior to being centrifuged at 53,000 g (Hitachi-Koki, CS150GXL) for 1 hr to remove large nanotube bundles and other impurities. After centrifugation, supernatants of SWCNT suspensions containing individualized nanotubes was further sorted by length using the DGU method reported previously in [14]. Briefly, the density gradients were formed by mixing iodixanol (60% (w/v), Sigma) with appropriate amount of 2 wt% DOC solutions or SWCNT supernatants. As sketched in Figure 1, four different layers were introduced in a centrifugation tube (12 mL), from the bottom to the top: 0.5 mL of 40 % iodixanol, 0.5 mL of 30 % iodixanol, 0.8 mL of 20 % iodixanol containing 0.267 mL dispersed SWCNTs, and 8.2 mL of 18 % iodixanol. Length sorting was performed in an ultracentrifuge (Hitachi CP80WX) with a rotor (P40ST) at 24,600 g for 94 hr at 5 °C. Afterwards, the dispersion in the centrifugation tube was extracted in ten individual fractions (1 mL each) by a density gradient fractionator (Hitachi, DGF-U), and labelled from “1f” (the topmost layer) to “10f” (the bottom layer), respectively.

(Figure 1 about here)

AFM was employed to characterize the length distribution of the sorted SWCNT fractions. Sliced silica wafers (1×1 cm) (Bonda) were first bath-sonicated in piranha solution (H₂SO₄ and H₂O₂) at 100 °C for 30 min, and then were rinsed with deionized water to obtain clean surfaces. Subsequently, SWCNTs were deposited on wafer surfaces by spin-coating for 10 s at 300 rpm, followed by 3,000 rpm for 600 s.

Deionized water was further applied to wash away surfactant residues on silica surfaces. AFM images were obtained by an MFP3D microscope (Asylum Research, Santa Barbara, CA). The scan rate was set to 1 Hz at various scan sizes, while all data resolutions were set to 512 pixels $\times$ 512 lines. To eliminate the Z offset between scan lines, we applied the first-order flattening to AFM amplitude profiles. Then, we analyzed the AFM amplitude profiles using the Adobe Photoshop CS4 software to determine the length of each SWCNT on 5 $\mu\text{m} \times 5 \mu\text{m}$ areas.

UV-vis-NIR absorption spectra were collected on a Varian Cary 5000 spectrophotometer from 400 to 1,350 nm. Measurements were made on the extracted SWCNT fractions in a 1 mm light path length quartz cuvette in transmission mode. Aqueous DOC solvents (2 wt %) were used as references. Data was recorded at 1 nm increments with an integration time of 0.1 s per increment.

3. Chemometric methods

Chemometric methods are applied to predict the length distribution of sorted SWCNTs from the corresponding absorption spectroscopy. From the measured nanotube length by AFM, it was found that the length approximately follows the lognormal distribution. Indeed, lognormal distribution is widely accepted to describe the distribution of size/length of particles [35-37]. A lognormal distribution is solely determined by two parameters: the mean (μ) and standard deviation (σ) of the natural logarithm of the random variable (i.e. the length). Therefore, μ and σ are treated as the response variables to be predicted by the chemometric calibration models. Relative root mean square error (RRMSE) and the coefficient of determination (R^2) are used to evaluate the performance of the calibration methods; RRMSE is defined as

$$\sqrt{\frac{1}{I} \sum_{i=1}^I \left(\frac{y_i - \hat{y}_i}{y_i} \right)^2} \quad (2)$$

where y_i is the actual response (either μ or σ) for the i^{th} sample determined by AFM, $\hat{y}_i$ is the chemometric prediction from the spectral data, and I is the number of samples. RRMSE and R^2 on the parameters of the log-normal distribution are indirect measures for the prediction accuracy, since it is the distribution itself that is of ultimate interest. Another measure, the Jensen-Shannon (J-S) divergence originally developed in the community of information theory and statistics [38], is used to directly quantify the difference between the predicted and reference distributions. For continuous distributions with density functions $p(x)$ and $q(x)$, e.g. $p(x)$ being the reference (true) distribution and $q(x)$ the predicted distribution, the J-S divergence is defined as

$$D(p, q) = \frac{1}{2} \int \left[p(x) \log \frac{2p(x)}{p(x) + q(x)} + q(x) \log \frac{2q(x)}{p(x) + q(x)} \right] dx \quad (3)$$

J-S divergence is between zero and unity; zero means that the two distributions are exactly the same whilst one means they are extremely different.

In the rest of this section, the chemometric techniques for pre-processing (D1, D2, SNV, EMSC, EISC) and non-linear calibration methods (ANN, GPR) are briefly introduced. As for notation, x_{ij} denotes the measured spectra of the i^{th} ($i = 1, \dots, I$) sample at the j^{th} ($j = 1, \dots, J$) wavelength (variable). $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{ij}]$ and y_i represents the corresponding response variable for the i^{th} sample.

3.1. Pre-processing methods

The pre-processing methods are used to remove the non-linearity originating from baseline noise and other disturbances, prior to the application of linear calibration methods (e.g. PLS).

First and second derivatives (D1 and D2)

The first derivative (D1) can be obtained by subtracting every variable of a sample from its immediate neighboring variables. By repeating this procedure on D1, the second derivative (D2) can be calculated [26]. To reduce the amplified high frequency measurement noise, Savitzky-Golay algorithm is used to attain smoother derivatives for calibration purposes [27].

Standard normal variate

The model of standard normal variate (SNV) is written as $\mathbf{x}_i = a_i + b_i \mathbf{x}_{i,chem}$, where $\mathbf{x}_{i,chem}$ denotes the theoretical (corrected) spectra. The coefficients, a_i and b_i , are the mean and standard deviation of the i^{th} sample with respect to wavelengths. The corrected spectra are obtained by $\mathbf{x}_{i,corr} = (\mathbf{x}_i - a_i)/b_i$.

Extended multiplicative signal correction

The model of extended multiplicative signal correction (EMSC) is given by $\mathbf{x}_i = a_i + b_i \mathbf{x}_{i,chem} + d_i \lambda + e_i \lambda^2$, where λ is the wavelength; d_i and e_i represent the wavelength-dependent spectral variations [29]. The model parameters can be estimated by the least square regression of measured spectra $\mathbf{x}_i$ and theoretical spectra $\mathbf{x}_{i,chem}$, which can be calculated from pure spectra $\mathbf{s}_k$ and the concentration y_{ik} of the k^{th} constituent through $\mathbf{x}_{i,chem} = \sum_{k=1}^K y_{ik} \mathbf{s}_k$. Afterwards, corrected spectra are calculated as $\mathbf{x}_{i,corr} = (\mathbf{x}_i - a_i - d_i \lambda - e_i \lambda^2)/b_i$. When pure spectra are not available, mean spectra of the entire data set are typically used in place of $\mathbf{x}_{i,chem}$ for estimating parameters.

Extended inverted signal correction

Extended inverted signal correction (EISC) is based on the “reverse” expression of EMSC as $\mathbf{x}_{i,chem} = a_i + b_i \mathbf{x}_i + c_i \mathbf{x}_i^2 + d_i \lambda + e_i \lambda^2$ [30]. The parameters (a_i, b_i, c_i, d_i, e_i) can be estimated by the least square regression, and corrected spectra are given by $\mathbf{x}_{i,corr} = a_i + b_i \mathbf{x}_i + c_i \mathbf{x}_i^2 + d_i \lambda + e_i \lambda^2$.

3.2. Non-linear calibration methods

The rationale of non-linear calibration is to directly model the non-linearity in the spectra with no regard to the source of the non-linearity. This is in contrast to the pre-processing methods that are targeted at a presumed cause of non-linearity (e.g. wavelength-dependent multiplicative effect due to disturbances). In the literature, non-linear calibration has shown comparable prediction performance with pre-processing techniques [25].

Artificial neural networks

A typical feed-forward artificial neural network (ANN) consists of three layers (input, hidden and output layer). Each layer comprises of multiple neurons. For calibration modeling, the response y_i from the output layer can be expressed as $y_i = f[\sum_{n=1}^N \omega_n g(\sum_{j=1}^J (w_j x_{ij} + \phi_j)) + \varepsilon_n]$ [32], where N is the number of hidden-layer neurons, w_j represents the weights connecting the input- and hidden-layer neurons, ω_n denotes the weights connecting the hidden- and output-layer neurons, ϕ_j and ε_n are the biases in the hidden and output layers. The “transfer functions”, $f(\cdot)$ and $g(\cdot)$, are typically taken as linear and sigmoid functions. In this study, the parameters are estimated by a Bayesian regularized “back-propagation” algorithm based on a Gaussian approximation to the posterior distribution [39]. The Bayesian approach automatically determines the “effective” number of parameters, and thus is less sensitive to the pre-chosen number of neurons and less likely to over-fit the data when compared with traditional ANN.

Gaussian process regression

In the Gaussian process regression (GPR) model [34], the response variable y_i is modelled by a joint Gaussian distribution with a zero mean as $\mathbf{y} = (y_1, \dots, y_I) \sim G(\mathbf{0}, \mathbf{C})$. The covariance matrix $\mathbf{C}$ is defined as $C_{ik} = C(\mathbf{x}_i, \mathbf{x}_k) = a_0 + a_1 \sum_{j=1}^J x_{ij}x_{kj} + v_0 \exp\left(-\sum_{j=1}^J w_j(x_{ij} - x_{kj})^2\right) + \sigma^2 \delta_{ik}$, where the first two terms represent the constant bias and linear correlation, respectively. The third term is similar to the form of the radial basis function, and the fourth term corresponds to the random error. The model parameters $\boldsymbol{\theta} = (a_0, a_1, v_0, w_1, \dots, w_J, \sigma^2)$ can be estimated by maximizing the log-likelihood of the data. For a new data point $\mathbf{x}^*$, the response y^* can be predicted as $(\hat{y}^*) = \mathbf{k}^T(\mathbf{x}^*)\mathbf{C}^{-1}\mathbf{y}$, where $\mathbf{k}(\mathbf{x}^*) = [C(\mathbf{x}^*, \mathbf{x}_1), \dots, C(\mathbf{x}^*, \mathbf{x}_I)]^T$.

4. Results and discussion

The photographs of centrifugation tubes and the schematic illustration of nanotube length sorting are shown in Figure 1. The layer containing SWCNTs is clearly visible before separation, which was placed near the bottom of the centrifugation tube to ensure the separation efficiency. After 94 hrs of centrifugation, the narrow SWCNT layer was spread out in the centrifugation tube. Longer nanotubes travel to the top of the centrifugation tube, while shorter nanotubes remain near the bottom. Absorption spectra of SWCNT fractions extracted from the centrifugation tube were shown in Figure 2. Sorted SWCNTs show clear length dependent absorption spectra, similar to previous reports [6, 14]. Longer SWCNTs display stronger optical features. In order to assess the influence of nanotube length on optical absorption, we focused on the (7, 6) SWCNTs, which are individually dispersed in 2 wt% DOC solution having the first interband transition near 1120 nm. All absorption data were scaled to the same concentration at 775 nm, following the method proposed by Fagan et al. [6]. The 775 nm was selected as the reference for scaling because there is little contribution from any SWCNTs or graphitic carbon features to the absorption around this wavelength. Such scaling helps alleviate the impact of the concentration of SWCNTs on the spectra, and thus, scaled absorption spectra are mainly a function of the length of SWCNTs. The decrease of absorption intensity above 1200 nm comes from the density medium. Afterwards, different chemometric calibration techniques were applied to predict the mean and standard deviation (SD) of SWCNTs' log-length based on their scaled absorption spectra from 400 to 1350 nm.

(Figure 2 about here)

In AFM investigation, only individualized SWCNTs were counted to ensure the accuracy of the reference length distribution. Approximately, 100 SWCNTs were measured for each length-sorted nanotube fraction. Three sample AFM images are illustrated in Figure 3. The length of sorted SWCNTs generally ranges from 100 to 800 nm. Figure 3 indicates that most nanotubes from 3f are longer than those from 5f and 9f, which is consistent with the observation of more intense absorption peaks from 3f than those from 5f and 9f. In total, we collected the length distributions of 21 length-sorted SWCNT samples, which were from three different initial SWCNT concentrations (referring to the concentration prior to DGU sorting): eight samples (3f-10f) from 0.8 mg mL⁻¹, seven samples (4f-10f) from 1 mg mL⁻¹, and six samples (5f-10f) from 1.2 mg mL⁻¹. The three concentrations correspond to three batches of centrifugation, thus referred to as “batches” hereafter, whilst the extracted layers of SWCNTs are termed “samples”. Some topmost layers from sorted samples obtained using higher initial SWCNT concentrations were discarded, because those sorted layers have a poor separation resolution. The histogram and fitted length distributions of nine selected nanotube samples obtained by AFM studies are presented in Figure 4. AFM results confirmed that the length of these sorted SWCNTs follows a lognormal distribution.

(Figure 3 and 4 about here)

The method of leave-one-out cross-validation (LOO-CV) is used to evaluate the chemometric techniques. In LOO-CV, one sample is chosen for prediction, while the remaining 20 samples are used as reference data

for developing a calibration model. This step is repeated until each sample has been used once for prediction. However, the left-out sample is related to the other samples from the same centrifugation batch and thus not independent. To resolve this issue, the method of leave-one-batch-out (LOBO) is also investigated, where the samples from one batch are treated as testing data while the samples from the other two batches are used for developing the calibration model. This is repeated for three times so that the samples from each centrifugation batch have been used once for testing.

Five pre-processing techniques in conjunction with PLS (D1, D2, SNV, EMSC, and EISC) and two non-linear calibration models (ANN and GPR) are evaluated to search for the optimal method. The results of PLS without pre-processing are reported as benchmark. The number of latent variables (LVs) in PLS is selected by five-fold random-splitting cross-validation. In addition, a simple “average prediction” method is used as the bottom-line technique to test the usefulness of the sophisticated calibration models, where the average of the response variables of the calibration data set is used as prediction. Prior to developing the non-linear calibration models (ANN and GPR), PLS is first applied to reduce the dimension of the data, which is a common method used to reduce the computational cost of parameter estimation in non-linear models [34]. Conceptually, non-linear models are capable of direct modelling of the non-linearity; we also empirically observed that pre-processing methods, which aim to remove the non-linearity of the data, do not improve the non-linear calibration models [25]. As a result, these scattering correction techniques are not used for non-linear models. The appropriate number of neurons in the hidden layer of ANN was preliminarily screened and fixed to six. Note that since ANN was trained using the Bayesian regularization method, the choice of the number of neurons did not appear to have significant impact on the results.

Table 1 shows the number of latent variables (LVs), RRMSE, R^2 and J-S divergence of different chemometric calibration techniques by using the method of LOO-CV. The values of J-S divergence in the table are the average of all the testing samples. All methods are better than the baseline average prediction (labelled “Ave. Pred.” in the table). The measures of RRMSE and R^2 indicate that EMSC, EISC and GPR attain outstanding accuracy in predicting the log-mean length (GPR being the best according to RRMSE). In addition, EMSC and EISC achieve superior performance for the log standard deviation (EISC being the slightly better). A general trend in Table 1 is that the prediction of mean is more accurate than the standard deviation. This phenomenon suggests that the prediction of the first-order information carried by the mean is easier than that of the second-order information carried by the standard deviation, which is consistent with statistical intuitions. In addition, lower-order information usually has higher influence on the entire probability distribution. The overall prediction accuracy can be assessed by the J-S divergence, which suggests that EMSC, EISC and GPR are effective calibration methods for this data set. The effectiveness of EISC as a representative pre-processing method is further illustrated in Figure 5, which clearly shows that the baseline shift/noise has been largely removed, giving rise to improved calibration models

(Table 1 and Figure 5 about here)

Using the predicted log-mean by GPR and log standard deviation by EISC, the lognormal distribution of the length of SWCNTs is recovered and displayed in Figure 6, overlaid with the histogram and fitted reference distribution from the AFM measurement. The J-S divergence values are also given for each sample. Visual inspection reveals that most distributions are well predicted. In addition, all the J-S divergence measures, except for sample 20, are less than 0.1 and the majority are fairly close to zero, indicating excellent calibration results.

(Figure 6 about here)

Next, the results of LOBO-CV are reported. This is a more challenging scenario since the amount of data for calibration modelling is very limited. The results are summarized in Table 2(a), (b) and (c), corresponding to the testing samples from the three batches with SWCNT concentration of 0.8, 1.0 and 1.2

mg mL⁻¹, respectively. Clearly, the prediction of the left-out batches is not as accurate as that of the left-out samples, and the prediction of the log standard deviation seems especially disappointing. In particular, some R^2 values become negative, suggesting that the sophisticated models are even worse than simply using the average response variables of the testing data for prediction, which would give $R^2=0$. However, the average of testing data is not available *a priori* when the calibration model is used in practice. Instead, only the average of the *calibration* data can be used for “prediction”, and the results are given in the last column of Table 2. The R^2 values of this averaging method are also negative, especially for the log standard deviation of the length distribution. More importantly, compared with this bottom-line method, it appears that GPR is still effective in predicting the log-mean of the length distribution, whilst EMSC and EISC provide satisfactory results on the log standard deviation. The measure for the accuracy of the overall distribution, J-S divergence, also indicates the improved results of the calibration modelling. Nevertheless, whether the improvement over the baseline averaging method is practically useful should be carefully assessed. Figure 7 illustrates the predicted length distributions with LOBO-CV, where the log-mean is predicted by GPR and the log standard deviation by EISC. There exists significant mismatch between the predicted and reference distribution for a few samples, the most pronounced being sample 1. However, for most samples, the overall location and shape characteristics of the length distribution have been properly captured, and the J-S divergence values are largely within 0.1, indicating that the calibration models are satisfactory in practice. In addition, the calibration accuracy is expected to be significantly improved if more batches of data can be collected.

(Table 2 and Figure 7 about here)

From the results of LOO-CV and LOBO, it appears that the prediction for mid-layer samples (5f-8f) at moderate “post-DGU” concentration (the actual concentration after DGU treatment and it can be inferred from the absorbance at 775 nm) are superior to samples from the top or bottom layers at either very low (absorbance at 775 nm < 0.01) or very high concentration (absorbance at 775 nm > 0.32). The relatively poor results at the low- and high-end post-DGU concentration may be the result of model extrapolation. In general, comfortable prediction performance has been achieved.

5. Conclusions

In this work, DGU method was applied to sort CoMoCAT SWCNTs by length. The length distribution was measured by AFM as a reference method, and it follows a lognormal distribution. Sorted SWCNT suspensions have length-dependent UV-vis-NIR absorption spectra, which provided the opportunity to calibrate the spectroscopy against AFM-measured length distribution for rapid analysis. Using absorption data, five linear calibration methods (D1, D2, SNV, EMSC, EISC) and two non-linear calibration methods (ANN and GPR) are evaluated in terms of prediction accuracy. Two chemometric calibration methods (GPR and EISC) are able to predict the length distribution with satisfactory agreement with AFM measurement. Two evaluating methods of LOO-CV and LOBO are investigated. The prediction performance of LOBO method is not as good as LOO-CV. This is largely due to the limited amount of data (only 13-15 samples) for calibration. Overall, the predicted distributions of most samples capture the major characteristics of the measured distributions which is the foremost goal of this work.

It should be noted that there are several limitations of the current technique. Spectroscopy, once properly calibrated, is to complement but not completely replace the time-consuming and costly AFM measurements. The calibration procedure itself requires centrifugation, sample preparation, and AFM analysis to obtain the reference data. Proper execution of the calibration stage is crucial to the success of the prediction model. In addition, SWCNT dispersions with very high or low concentration as discussed above may induce errors in their length distribution prediction by absorption spectra, which is a well-known extrapolation problem encountered in all data-based chemometric techniques. Extrapolation could also be an issue for SWCNTs

taken from different sources, in which situation a new calibration model may be required if the samples have significantly different diameters or lengths.

Acknowledgements

This work was partially supported by National Research Foundation, Singapore (Grants: NRF-CRP2-2007-02 and NRF2010-POC001-021).

References

- [1] A. Jorio, G. Dresselhaus, M.S. Dresselhaus, Carbon Nanotubes, Advanced Topics in the Synthesis, Structure, Properties and Applications, Springer, Berlin, 2008, p. 1.
- [2] M.C. Hersam, Nature Nanotechnology, 3 (2008) 387.
- [3] S. Banerjee, T. Hemraj-Benny, S.S. Wong, J. Nanosci. Nanotechnol., 5 (2005) 841.
- [4] J. Liu, M.C. Hersam, MRS Bull., 35 (2010) 315.
- [5] Y. Matsuo, K. Tahara, E. Nakamura, Org. Lett., 5 (2003) 3181.
- [6] J.A. Fagan, J.R. Simpson, B.J. Bauer, S.H.D. Lacerda, M.L. Becker, J. Chun, K.B. Migler, A.R.H. Walker, E.K. Hobbie, J. Am. Chem. Soc., 129 (2007) 10607.
- [7] A. Rajan, M.S. Strano, D.A. Heller, T. Hertel, K. Schulten, Journal of Physical Chemistry B, 114 (2010) 4938.
- [8] J.P. Casey, S.M. Bachilo, C.H. Moran, R.B. Weisman, Acs Nano, 2 (2008) 1738.
- [9] M. Zheng, E.D. Semke, J. Am. Chem. Soc., 129 (2007) 6084.
- [10] M.L. Becker, J.A. Fagan, N.D. Gallant, B.J. Bauer, V. Bajpai, E.K. Hobbie, S.H. Lacerda, K.B. Migler, J.P. Jakupciak, Advanced Materials, 19 (2007) 939.
- [11] X.Y. Huang, R.S. McLean, M. Zheng, Analytical Chemistry, 77 (2005) 6225.
- [12] J. Chun, J.A. Fagan, E.K. Hobbie, B.J. Bauer, Analytical Chemistry, 80 (2008) 2514.
- [13] J.A. Fagan, M.L. Becker, J. Chun, E.K. Hobbie, Advanced Materials, 20 (2008) 1609.
- [14] J.A. Fagan, M.L. Becker, J.H. Chun, P.T. Nie, B.J. Bauer, J.R. Simpson, A. Hight-Walker, E.K. Hobbie, Langmuir, 24 (2008) 13880.
- [15] E.K. Hobbie, J.A. Fagan, J. Obrzut, S.D. Hudson, Acs Applied Materials & Interfaces, 1 (2009) 1561.
- [16] C.W. Lee, X.D. Han, F.M. Chen, J. Wei, Y. Chen, M.B. Chan-Park, L.J. Li, Advanced Materials, 22 (2010) 1278.
- [17] M.S. Arnold, A.A. Green, J.F. Hulvat, S.I. Stupp, M.C. Hersam, Nature Nanotechnology, 1 (2006) 60.
- [18] E.J.F. Carvalho, M.C. dos Santos, Acs Nano, 4 (2010) 765.
- [19] N. Stiirzl, F. Hennrich, S. Lebedkin, M.M. Kappes, J. Phys. Chem. C, 113 (2009) 14628.
- [20] K. Yanagi, T. Iitsuka, S. Fujii, H. Kataura, Journal of Physical Chemistry C, 112 (2008) 18889.
- [21] H.C.v.d. Hulst (Ed.), Light Scattering by Small Particles, Dover Publications, 1981.
- [22] M.A. van Dijk, A.L. Tchegotareva, M. Orrit, M. Lippitz, S. Berciaud, D. Lasne, L. Cognet, B. Lounis, Physical Chemistry Chemical Physics, 8 (2006) 3486.
- [23] P.K. Jain, K.S. Lee, I.H. El-Sayed, M.A. El-Sayed, Journal of Physical Chemistry B, 110 (2006) 7238.
- [24] S. Sarkar, A.A. Gurjarpadhye, C.G. Rylander, M.N. Rylander, Journal of Biomedical Optics, 16 (2011).
- [25] K. Wang, G.Y. Chi, R. Lau, T. Chen, Analytical Letters, 44 (2011) 824
- [26] C.D. Brown, L. Vega-Montoto, P.D. Wentzell, Applied Spectroscopy, 54 (2000) 1055.
- [27] A. Savitzky, M.J.E. Golay, Analytical Chemistry, 36 (1964) 1627.
- [28] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Applied Spectroscopy, 43 (1989) 772.
- [29] H. Martens, J.P. Nielsen, S.B. Engelsen, Analytical Chemistry, 75 (2003) 394.
- [30] D.K. Pedersen, H. Martens, J.P. Nielsen, S.B. Engelsen, Applied Spectroscopy, 56 (2002) 1206.

- [31] J.R. Long, V.G. Gregoriou, P.J. Gemperline, *Analytical Chemistry*, 62 (1990) 1791.
- [32] Z.Y. Wang, J.N. Hwang, B.R. Kowalski, *Analytical Chemistry*, 67 (1995) 1497.
- [33] H.H. Thodberg, *IEEE Trans. Neural Netw.*, 7 (1996) 56.
- [34] T. Chen, J. Morris, E. Martin, *Chemometrics and Intelligent Laboratory Systems*, 87 (2007) 59.
- [35] A. Matsuura, L.S. Fan, *Aiche J.*, 30 (1984) 894.
- [36] M.I. Hossain, T. Chen, Y.H. Yang, R. Lau, *Ind. Eng. Chem. Res.*, 48 (2009) 10136.
- [37] J.P.M. Syvitski, Cambridge University Press: New York (2007).
- [38] J. Lin, *IEEE Trans. Inform. Theor.* 37 (1991) 145.
- [39] D.J.C. Mackay, *Neural Comput.*, 4 (1992) 448.

Figure 1. Photographs of centrifugation tubes before and after length sorting, and a schematic illustration of sorted SWCNTs.

Figure 2. Absorbance spectra of length-sorted SWCNT fractions after scaling at 775 nm. Note that “1f” and “2f”, corresponding to the sample from the top of the centrifugation tube, contains few nanotubes, and thus their spectra are removed from analysis. The absorption spectrum of the starting material before separation (dashed line and denoted “AP”) is shown as a reference.

Figure 3. AFM images of nanotube fractions 3, 5, and 9 extracted from length-sorted SWCNT suspensions (initial concentration at 0.8 mg mL^{-1}). The size of all the images is $5 \mu\text{m} \times 5 \mu\text{m}$ and the scale bar is at $1 \mu\text{m}$.

Figure 4. The length distribution of sorted SWCNT samples obtained from AFM analysis. The red solid line represents the fitted log-normal distribution. The bars represent the histogram from the AFM measure.

Figure 5. The spectra before and after pre-processing by EISC.

Figure 6. LOO-CV calibration results where the log-mean is predicted by GPR and log standard deviation by EISC. The red solid line represents the reference fitted distribution, while the blue dotted line indicates the predicted distribution. The bars represent the histogram from the AFM measure.

Figure 7. LOBO-CV calibration results where the log-mean is predicted by GPR and log standard deviation by EISC. The red solid line represents the reference fitted distribution, while the blue dotted line indicates the predicted distribution. The bars represent the histogram from the AFM measure.

Table 1. The LOO-CV calibration results.

		PLS	D1	D2	SNV	EMSC	EISC	ANN	GPR	Ave. Pred.
Mean	LVs	5.7	5.2	5.8	5.7	10.1	10.9	5.7	5.7	-
	RRMSE	20.5%	18.3%	36.0%	21.6%	7.3%	7.2%	22.3%	6.7%	48.8%
	R^2	0.78	0.78	0.05	0.75	0.95	0.95	0.77	0.93	-0.33
Standard Deviation	LVs	3.9	2.3	3.3	3.9	5.4	4.4	3.9	3.9	-
	RRMSE	23.9%	22.2%	21.1%	24.7%	13.3%	12.0%	35.0%	22.5%	57.8%
	R^2	0.41	0.37	0.34	0.40	0.83	0.85	0.01	0.42	-20.71
J-S Divergence		0.073	0.075	0.221	0.065	0.021	0.020	0.074	0.032	0.318

Table 2. The LOBO-CV calibration results.(a) Testing samples from the concentration of 0.8 mg mL⁻¹.

		PLS	D1	D2	SNV	EMSC	EISC	ANN	GPR	Ave. Pred.
Mean	LVs	7	2	4	7	5	6	7	7	-
	RRMSE	58.6%	31.3%	54.0%	60.3%	36.1%	29.7%	59.8%	26.9%	48.8%
	R²	-0.55	0.40	-0.20	-0.64	0.41	0.59	-0.62	0.68	-0.33
Standard Deviation	LVs	3	2	3	3	4	4	3	3	-
	RRMSE	28.5%	28.1%	43.1%	48.1%	17.8%	19.8%	53.4%	54.0%	57.8%
	R²	-4.55	-4.34	-13.36	-16.91	-1.39	-1.85	-17.45	-18.94	-20.71
J-S Divergence		0.301	0.208	0.204	0.266	0.163	0.129	0.289	0.118	0.318

(b) Testing data: samples from the concentration of 1.0 mg mL⁻¹.

		PLS	D1	D2	SNV	EMSC	EISC	ANN	GPR	Ave. Pred.
Mean	LVs	4	5	3	4	4	4	4	4	-
	RRMSE	9.9%	41.2%	57.3%	9.3%	16.2%	21.8%	26.4%	8.3%	26.1%
	R²	0.86	-0.12	-2.16	0.82	0.58	0.33	-0.13	0.89	-0.21
Standard Deviation	LVs	5	4	4	5	4	4	5	5	-
	RRMSE	28.7%	30.8%	31.0%	29.6%	29.9%	31.1%	43.8%	43.2%	43.2%
	R²	-11.76	-13.71	-14.46	-12.34	-13.10	-14.13	-30.00	-11.76	-28.69
J-S Divergence		0.055	0.150	0.380	0.063	0.122	0.172	0.258	0.114	0.262

(c) Testing data: samples from the concentration of 1.2 mg mL⁻¹.

		PLS	D1	D2	SNV	EMSC	EISC	ANN	GPR	Ave. Pred.
Mean	LVs	5	8	7	5	7	7	5	5	-
	RRMSE	6.3%	11.2%	15.7%	7.4%	8.6%	10.4%	5.1%	6.1%	20.0%
	R²	0.90	0.74	0.57	0.87	0.89	0.82	0.94	0.91	-0.23
Standard Deviation	LVs	4	2	3	4	4	5	4	4	-
	RRMSE	26.6%	18.0%	18.6%	27.4%	19.7%	16.2%	31.9%	29.5%	32.0%
	R²	-2.38	-0.33	-0.47	-2.31	-0.71	-0.43	-3.25	-2.38	-3.25
J-S Divergence		0.040	0.057	0.091	0.043	0.041	0.051	0.035	0.035	0.165