Biological Characteristics of Cell Similarity Measure

Qingqing Hu, Hongfei Zhang, Yusi Zheng, Hongwei Qin, Bo Hu,* Qiang Yu,* and Jie Tian*

1. Introduction

Similarity measures play an important role in many data analysis fields. However, the current cell similarity measures are poor used in data characteristics. Herein, a novel similarity measure named segment weighting similarity (SWS) is developed for the analysis of single-cell Raman spectra. SWS segments the spectra by the cell biological characteristics and quantifies the significant factors per region, which can increase the contribution of intrinsic biological features and reduce noise. The similarity heat maps of SWS and three other kinds of traditional similarities, including cosine similarity, Pearson correlation coefficient, and Euclidean distance, show that SWS has high accuracy and low bias in distinguishing cell spectra. K-nearest-neighbor classifiers have the identification accuracy, sensitivity, and specificity of 0.852, 0.853, and 0.965, respectively. The purity of the clustering model could increase by 0.31 in some tasks of the K-means and spectral clustering. The classification and clustering results demonstrate that SWS is more effective than common ones. SWS, based on the basic data and intrinsic biological characteristics, provides a new thought and formula in the similarity measure for most of the Raman and infrared technologies, and has great potential for enhancing the performance of machine learning algorithms.

Similarity measures play an important role in many data analysis fields. However, the current cell similarity measures are poor used in data characteristics. Herein, a novel similarity measure named segment weighting similarity (SWS) is developed for the analysis of single-cell Raman spectra. SWS segments the spectra by the cell biological characteristics and quantifies the significant factors per region, which can increase the contribution of intrinsic biological features and reduce noise. The similarity heat maps of SWS and three other kinds of traditional similarities, including cosine similarity, Pearson correlation coefficient, and Euclidean distance, show that SWS has high accuracy and low bias in distinguishing cell spectra. K-nearest-neighbor classifiers have the identification accuracy, sensitivity, and specificity of 0.852, 0.853, and 0.965, respectively. The purity of the clustering model could increase by 0.31 in some tasks of the K-means and spectral clustering. The classification and clustering results demonstrate that SWS is more effective than common ones. SWS, based on the basic data and intrinsic biological characteristics, provides a new thought and formula in the similarity measure for most of the Raman and infrared technologies, and has great potential for enhancing the performance of machine learning algorithms.

Similarity is an important metric in machine learning, which has been applied widely in many fields, such as text matching,[3] image recognition,[2] and biomedical applications.[1–4] Cosine similarity (CS), Pearson correlation coefficient (PC), and Euclidean distance (ED), as the common similarity measures, have the classical and universal similarity metrics for the analysis of various data. However, these universal methods lose a part of the specific information and hardly utilize the intrinsic characteristics. Each feature of the data has a special meaning and contributes differently to the results, which may be positive or negative.[7,8] Common similarity measures treat all features as the same. A strategy should be considered on how to optimize the feature. Weighting is an efficient approach to quantify the important coefficient of a feature by the data characteristics, which can reduce the influence of weak features and improve the contribution of useful features. Therefore, appropriate weightings can improve the similarity performance to obtain good results in data analysis.[8]

Raman spectroscopy based on molecular vibrational scattering is known as the molecular fingerprint.[9,10] The analysis of Raman data is challenging due to the low signal-to-noise ratio, dispersive signals, and signal overlap.[11,12] More and more researchers have explored the analysis of Raman spectra by machine learning. Carey et al. optimized the mineral spectra matching performance using careful preprocessing and a weighting-neighbors classifier of a vector similarity metric.[13] Fenn has presented a novel data analysis framework named fisher-based feature selection support vector machines (FFSVM) for classification and has got high accuracy in five cancerous and noncancerous breast cell lines.[14] However, similarity is rarely used for the analysis of Raman spectra by machine learning due to the limited performance of the common similarity.

Raman spectroscopy reflects variations of DNA/RNA, proteins, lipid, carbohydrates, and other small-molecule metabolites, which makes it an excellent tool for monitoring biochemical changes on the cellular level.[14–16] However, the data analysis methods ignore the biological characteristics of Raman spectra and treat each feature as equally important. Raman spectra are related to the biological characteristics of molecular vibration,
in which the contribution of each feature to the similarity results varies greatly. Therefore, it is necessary to mine the biological characteristics of Raman spectra and apply the information to machine learning methods. Study based on the Fourier transform infrared (FTIR) spectroscopy of bacteria has shown a new strategy that segmented the spectra according to the major contribution of biomacromolecules in each region.\(^{[17,17\text{--}19]}\)

Although this strategy has matched the biological characteristics of bacteria with the Raman spectra for clearly explaining data characteristics and the biological characteristics can clearly interpret the spectra, it has not been applied in data analysis.

Here, a novel similarity measure based on cell biological characteristics, named segment weighting similarity (SWS), is provided. SWS segments spectra according to biological characteristics and assigns weightings to each feature properly. In the segment part, a new scheme based on cell Raman spectra was designed. In the weighting part, the weighting method is used in each region of the segments, which can reduce the influence of weak features and noise as well as increase the good regions. Then, a comprehensive similarity value is calculated. A similarity heat map is used to study the superiority of SWS over other similarity measures, such as CS, ED, and PC. SWS is also used to establish a K-nearest neighbor (KNN) model,\(^{[20]}\) a classifier based on similarity calculation. Compared with other KNN models, SWS has a better classification performance for five kinds of cell spectra. SWS shows significantly improved results in the clustering algorithm,\(^{[21]}\) such as K-means and spectral clustering.

### 2. Results and Discussion

#### 2.1. SWS

The SWS method is proposed to make full use of the abundant cell information in the Raman spectrum, which has a broad application, such as classification and clustering (Figure 1).

The process of SWS is shown in Figure 1a. The Raman spectra of a single cell can be divided into five regions according to the main contribution of biomolecules, and the segment scheme is shown in Table 1. Then, each region is weighted separately. The five weighting subspectra are used to calculate the data similarity in the database. A comprehensive distance is generated by summing the five similarities. Therefore, SWS considers both data characteristics and cell biology through segment and weighting.

#### 2.2. Segment Scheme of Raman Spectra

An innovative Raman spectra segment scheme (Figure 1a) is proposed taking the single cell Raman characteristics and preliminary experimental results into account (Figure S1 and S2, Supporting Information), which segments the spectra into five regions, including the fingerprint region, protein I region, mixed region, protein II region, and genetic material region. In addition to the segment scheme, Table 1 summarizes the biological meaning of the important wavenumbers in each segment in detail. Each color corresponds to a type of biomolecule. For the fingerprint region in 600–928 cm\(^{-1}\), the vital wavenumbers mainly express genetic material, protein, and saccharide.\(^{[7]}\) Rich and complex biological characteristics of this region named “fingerprints” lead to the special meaning. For the second region in the 929–1010 cm\(^{-1}\), the selected wavenumbers shown in Table 1 are the main peak locations correlating to phenylalanine closely. This subset maybe shows the difference of the cell metabolism. For the mixed region 1011–1430 cm\(^{-1}\), the first half is mainly related to lipids and the second one is mainly related to the genetic material. In the protein II region from 1431 to 1626 cm\(^{-1}\), most are proteins except 1528 cm\(^{-1}\), which is carotenoid. For the last region in 1627–1800 cm\(^{-1}\), the most prominent wavenumbers are correlated to the genetic material, and the main characteristics of this region are 1662 and 1663 cm\(^{-1}\). The segment scheme expresses the correlation of Raman spectra and biological characteristics and provides the strategy and basis for SWS.

The selection of the length of the segment is the main part of the SWS. Most importantly, the intrinsic biological characteristics are the first principle. The detailed summary of the biological meaning for the Raman peaks in each segment is necessary. Figure S3, Supporting Information, shows the mean curve of Raman spectra and the original weighting curves for classification and clustering without sampling, windowing, and iterations, in which Raman peaks with high values are used to determine the important wavenumbers. After the selection process of the length of the segment, every segment region should have the typical biological characteristics with high correlation between important wavenumbers. The calculation of the classification accuracy of each segment shows the similar ranks between the segments (Table S1, Supporting Information). This proves the feasibility of the segment scheme and the necessity of the weighting method. Further, the basic data characteristics are used to improve the selection results. In each segment, the weighting curve of the classification is compared with the mean curve of Raman spectra. Figure S4, Supporting Information, shows the good correlation between the mean curve of the Raman spectra and the weighting curve of the classification in five segment regions, and the important wavenumbers of Raman peaks are well kept in the weighting curve. In contrast, the poor correlation in four segment regions is shown in Figure S5, Supporting Information, and some of the important wavenumbers of Raman peaks are missing in the weighting curve. Further, sampling frequency and iterations are two important parameters in the weighting curve. Figure S6, Supporting Information, shows the effect of the sampling frequency and iterations on the weighting curve in five segment regions. While the sampling frequency and number of iterations are 2 and 2, respectively, the important wavenumbers of the Raman peaks are kept, and unconsidered wavenumbers are filtrated.

#### 2.3. Weighting Process

An effective weighting method is also necessary because the contribution of each segment is different for the data analysis. Each feature corresponds to a matching weighting in the weighting part. The right side of Figure 1a shows the process of calculating weighting. The main steps are spectra sampling, weighting formula, and windowing. Spectra sampling can effectively reduce the calculation number of high-dimensional...
Figure 1. Design of SWS and its typical application. a) The process of SWS. b) The weighting spectrum based on the classification weighting formula, which is the blue curve, and the original spectrum shown as the orange one. c) The process of the classification and clustering based on SWS. These are two applications of SWS. The weighting formulas for the classification and clustering are different, shown in the weighting process part.
spectra by selecting a feature at intervals of \( n \) (sampling frequency) and calculating the weighting. The weighting formula is one of the most important parts of the total process. It measures the contribution of the features to the experimental results after sampling. This formula can be used to assign big weightings for good features and small weightings for poor features. Good features contain more information. More specifically, it means a feature with obvious difference in different labeled data.

Top \( K \) of other labeled spectra are used to train the weighting curve, whereas classification requires difference between samples. The weighting formula is

\[
W(A) = \sum_{C \in \text{Class}(R)} \left[ \frac{p(C)}{1 - p(\text{Class}(R))} \sum_{j=1}^{k} \frac{|R(A) - M_j(C)|}{mk} \right]
\]

The formula calculates the weighting of the feature \( A \). \( R \) is a random sample. \( K \) is the number of neighbor samples. \( M \) is the database containing \((F - 1) \times K\) nearest-neighbor samples with the different labels, such as \( R \). \( F \) is the number of spectra kinds. The \( m \) is the iterations. \( M_j(C) \) is the \( j \)th nearest neighbor sample in the class \( C \).

Table 1. The segment scheme of the single-cell Raman spectra.

| Name               | Range [cm\(^{-1}\)] | Peak [cm\(^{-1}\)] | Biological characteristics                  | References |
|--------------------|----------------------|--------------------|---------------------------------------------|------------|
| **Fingerprint region** | 600–928             | 625                | 600–800: nucleotide conformation           | [16,25]    |
|                    |                      | 725                | A (ring breathing mode of DNA/RNA bases)    | [25]       |
|                    |                      | 752                | V15 (porphyrin breathing mode)             | [26]       |
|                    |                      | 753, 754           | Symmetric breathing of tryptophan (protein assignment) | [27-29]    |
|                    |                      | 831                | Asymmetric O—P—O stretching, tyrosine       | [25]       |
|                    |                      | 856                | Amino acid side chain vibrations of proline hydroxy proline | [30]       |
|                    |                      | 867                | Ribose vibration one of the distinct RNA modes | [25]       |
|                    |                      | 898                | Monosaccharides (B-glucose), (C—O—C) skeletal mode | [31]       |
|                    |                      | 928                | Protein band                               | [32]       |
| **Protein I region** | 929–1010            | 941                | Skeletal modes (polysaccharides, amylose)   | [31]       |
|                    |                      | 1008               | Phenylalanine                              | [33]       |
|                    |                      |                   | \( n (CO), n (CC), d (OCH), \) ring (polysaccharides, pectin) | [31]       |
|                    |                      | 1000–1008          | Phenylalanine                              | [25,33-35] |
| **Mixed region**  | 1011–1430           | 1016               | Carbohydrates peak for solids             | [36]       |
|                    |                      | 1073               | Triglycerides (fatty acids)                | [37]       |
|                    |                      | 1124, 1129         | V(C—C) skeletal of acyl backbone in lipid (transconformation) | [34]       |
|                    |                      | 1131               | Palmitic acid                              | [38]       |
|                    |                      | 1188               | Fatty acid                                 | [38]       |
|                    |                      | 1257               | 1185–300: Antisymmetric phosphate vibrations | [39]       |
|                    |                      | 1315, 1317         | Guanine (B, Z-marker)                      | [40]       |
|                    |                      | 1342               | G (DNA/RNA)                                | [35]       |
|                    |                      | 1450               | CH\(_2\) bending (proteins)                | [41]       |
|                    |                      | 1453               | Structural protein modes of tumors         | [42]       |
|                    |                      | 1528               | Carotenoid (absent in normal tissues)       | [43]       |
|                    |                      | 1560               | Tryptophan                                 | [32]       |
|                    |                      | 1587, 1588         | Phenylalanine, hydroxyproline              | [34]       |
| **Protein II region** | 1431–1626           | 1637               | Amide I band                               | [44]       |
|                    |                      | 1657               | Fatty acids                                | [36,45]    |
|                    |                      | 1662               | Nucleic acid modes                         | [43]       |
|                    |                      | 1663               | DNA                                        | [27]       |
|                    |                      | 1747               | C=O, lipids                                | [41]       |

| Saccharide | Protein | DNA,RNA | Lipids | Others |
|------------|---------|---------|--------|--------|

| Genetic material | 1627–1800 | 1637 | Amide I band | [44] |
|                  |           | 1657 | Fatty acids  | [36,45] |
|                  |           | 1662 | Nucleic acid modes | [43] |
|                  |           | 1663 | DNA          | [27] |
|                  |           | 1747 | C=O, lipids  | [41] |
Formula (1) quantifies the importance of feature A by the difference between spectral features. The differences in spectral features are fundamental to machine learning. Differences in other labeled data are especially important, and differences in the same labeled data are small enough to be ignored.

There is no labeled information in the clustering. The feature difference is the standard deviation of all the data in the feature and is used to study the spectral difference. The weighting formula is:

$$W(A) = \frac{1}{N} \sum_{j=1}^{N} (A_j - \mu_A)^2$$  \hspace{1cm} (2)

$A$ is a feature, and $N$ is the data number. $\mu_A$ is the mean of the feature $A$. $A_j$ is the $j$th value of the feature $A$.

The SWS formula is defined as

$$S = \sum_{k=1}^{s_1} (W_k \times x_{1k} - W_k \times x_{2k})^2 + \sum_{k=s_1+1}^{s_2} (W_k \times x_{1k} - W_k \times x_{2k})^2 \ldots$$

$$+ \sum_{k=s_{4}+1}^{s_5} (W_k \times x_{1k} - W_k \times x_{2k})^2$$

where $s_1$, $s_2$, ..., $s_5$ is the end of each segment, and $x_{1k}$ is the $K$th feature of data $x_1$.

Windowing is the next important step after weighting. The windowing function is the extension of the high-weighting subset. For example, in Figure 1a, the green imaginary line represents the mean value of the weighting curve. The point is filtered when it is smaller than the mean, and the point is retained when it is larger than mean. Windowing increases the weighting around the retained weighting to a high-weighting subset. This process is repeated several times, which is called the iteration. When the iterations reach the target value, the loop finishes. The unimportant weighting is assigned to a small fixed value in the final weighting curve, such as the mean value or 0. The three aforementioned steps can effectively reduce the calculation amount, cut down the influence of noisy data, and enhance robustness.

The weighting spectra based on the classification weighting formula are shown in Figure 1b and the curve for the clustering is shown in Figure S7, Supporting Information. The weighting spectrum has shown that the weighting has a significant enhancement effect on some points in the fingerprint region, protein 1 region, and genetic material region. For example, 754 cm$^{-1}$ in the fingerprint region is related to symmetric breathing of tryptophan, and 938 cm$^{-1}$ in the protein I is related to proline, hydroxyproline, and $\eta$(C–C) skeleton of the collagen backbone. This indicated that some proteins are the key to distinguishing different cells. The 1663 cm$^{-1}$ in the genetic material region is related to DNA or RNA. Therefore, protein and DNA are two special materials for cell identification and show obvious differences between the original spectrum and the weighting curve.

2.4. Effectiveness of SWS

Similarity heat map describes the similarity in a batch of data by color changes. Heat maps can be used to evaluate the performance of similarity measures where data labels are determined. Figure 2 has shown that the heat maps are calculated by four similar measures: SWS, CS, ED, and PC. The spectra of five cell lines, including MCF-10a, MDA-MB-231, BT-474, SK-BR3, and T-47D, have been used, each containing 40 spectra. The plots in the black box are intraclass similarity, which should be higher than the interclass similarity. The intraclass similarity and interclass similarity of each cell line are calculated separately and shown by the histogram.

SWS and ED have shown a larger difference between the intraclass similarity and interclass similarity in Figure 2, which are obviously better than CS and PC. In the table in Figure 2, the performance of SWS and ED was analyzed by the intraclass similarity, interclass similarity, and differences. In particular, difference is an important index of the ability in the similarity measure to distinguish cell spectra, which is calculated by subtracting interclass similarity from intraclass similarity. SWS and ED have the same mean difference, and the deviation of SWS is small. CS and PC have poor performance and the difference is small. Comparing SWS with ED, SWS has a smaller deviation for the same difference. Therefore, SWS has the best performance in the similarity heat map.

2.5. Application of SWS in the Classification

A general classifier of KNN based on the similarity measure has been used to demonstrate the performance of SWS, ED, PC, and CS compare with SWS, in which the accuracy, specificity, sensitivity, confusion matrix, receiver operating characteristic (ROC) curve and area under the curve (AUC) values are studied for the performance of the four similarity measures. Figure 1c shows the classification process. The similarity of the preprocessed spectra is calculated by SWS. Then, spectra are classified by the KNN algorithm. The database contains 1003 spectra from five cell lines, and tenfold cross-validation is used to reduce errors.

$K$ is studied from 5 to 30 as an important parameter. As shown in Figure 3a, the accuracy and sensitivity of SWS can be up to 0.852 and 0.853, respectively. CS is the best in the other three formulas, with the accuracy and sensitivity of 0.816 and 0.816, respectively. In contrast, the results of ED are poor, and the accuracy and sensitivity are 0.651 and 0.663, respectively. All the four similarity formulas are high specificity, up to 0.9. It is interesting to demonstrate the effectiveness of biological characteristics, because SWS adds biological characteristics to ED for the calculation. Therefore, SWS has the best capability of the four similarity formulas in the $K$ values. Figure S8–S11, Supporting Information, show the $K$ values from 1 to 30 without interval. The ROC curve and confusion matrix are two common parameters to evaluate the classification results. $K$ of 20 is chosen to calculate the two indexes of the four similarity formulas based
on KNN. T-47D has poor classification results in Figure 3c, which is different from other data. The AUCs of SWS, CS, and ED are above 0.95. However, the AUC of PC is slightly poor: it equals only 0.89. Therefore, the performance of SWS in the six indexes is better than that of the other three similarity measures. CS and ED are second, and PC is the worst.

2.6. Application of SWS in Clustering

Clustering is chosen as the second application of SWS to verify its performance, and the choice of similarity measure is important\cite{22,23} in clustering. Three indexes are used to evaluate the performance of SWS and other traditional similarities. Figure 1c shows the use of SWS in the clustering. The preprocessed batch of data is clustered after SWS calculation, and then the advantages of SWS are shown through three indicators.

The database includes 1003 spectra, and there are six clustering tasks designed to compare the formula performance roundly. Figure S12, Supporting Information, shows the six tasks in detail, and the information of each clustering task is illustrated by colored dots. Figure 4a shows the clustering results based on K-means, in which each colored bar represents a cell line in the heat map. Every task has two rectangular colored strips: the one at the top is the result based on SWS and the one at the bottom is calculated by ED. Figure 4b shows the purity, precision, and recall of each task for the evaluation of the index results.

In task 1, 1003 spectra are clustered into two classes, normal cells and tumor cells. Compared with the standard color scales, the purity, precision, and recall of SWS are 0.73, 0.932, and 0.711, respectively. These three indexes of ED are 0.644, 0.848, and 0.672, which are worse than SWS. In task 2-1, it is a complex condition that the data are clustered into five clusters.

Figure 2. The similarity heat maps of five kinds of cells were calculated by a) SWS, b) CS, c) ED, and d) PC. The histograms represent the difference of intraclass similarity and interclass similarity. Each of the columns in the histogram is a cell line. The heat map from left to right and table from top to bottom represent MCF-10A, MDA-MB-231, BT-474, SK-BR3, and T-47D, respectively. Tables sum the average of the intraclass, interclass, and the difference between them.
those based on SWS are 0.573, 0.593, and 0.573, respectively. Therefore, SWS is better than ED in three indicators. In task 2-3, the data are from the spectra of MCF-10A and BT-474. The purity of SWS is 0.824, and the result based on ED is 0.513. The purity of SWS can increase by 0.31. In the other three tasks, the performance of SWS is better than that of ED, in which MCF-10A versus MDA-MB-231, SK-BR3, and T-47D is used. The purity of the three tasks of SWS is 0.01–0.04 higher than that based on ED in Table S2, Supporting Information. Therefore, SWS can improve the performance of K-means distinctly and describes the degree of similarity more accurately. To further verify these results, the results of PC and CS are also compared in detail in Table S3, Supporting Information. The purity of CS is less than 0.8, and most of them are less than 0.6. PC performs well in some clustering tasks, such as task 2-2 and task 2-3, and the purity can reach up to 0.99. However, it is unstable, and performance degrades in other cluster tasks. PC and CS perform poorly on most tasks. Spectral clustering, another common clustering method, is also used to study the performance of SWS and other similarity measures in Table S4 and S5, Supporting Information. SWS can improve the purity of the original spectral clustering in five clustering tasks, except that the purity of SWS in task 2-2 is lower than that of ED. Especially in task 2-3, SWS increases the purity by 0.34. Therefore, SWS can significantly improve the performance of original clustering models compared to other similarity measures.

SWS considers both the basic data and the intrinsic biological characteristics in the process of the segment and weighting. It is a novel strategy for the similarity measure to make good use of the basic data and intrinsic characteristics, which can produce better performance than using only the basic data features. It has been demonstrated that the single-cell biological characteristics are important factors affecting the data analysis of the Raman spectra in SWS. Further, SWS can be used to other Raman and infrared technologies such as surface-enhanced Raman spectra, tip-enhanced Raman spectra, coherent anti-Stokes Raman spectra, stimulated Raman spectra, infrared spectra, and Fourier transform infrared spectra. These Raman and infrared spectra also have shown the rich basic data and intrinsic structural or biological characteristics of the analytical molecules, materials, proteins, cells, and tissues. Intelligent microsystem has achieved great progress recently, which can well integrate these Raman and infrared technologies in it. Therefore, SWS would
not only provide new thoughts and solutions in the similarity measure on the complex data characteristics of single-cell Raman spectroscopy, but also inspire more practical applications in other Raman and infrared technologies.

Figure 4. The application of SWS in the clustering. a) Based on the K-means, the clustering results of six tasks that used SWS and ED. Task 1 clustered the five kinds of cells into normal cells and cancer cells. Task 2 includes five tasks, five kinds of spectra clustering, MCF-10A versus MDA-MB-231, MCF-10A versus BT-474, MCF-10A versus SK-BR3, and MCF-10A versus T-47D, respectively. b–g) The statistics results of the six clusterings.
3. Conclusion

In conclusion, SWS, as a novel similarity measure, makes good use of both the basic data and intrinsic biological characteristics by the process of segmenting and weighting. SWS strongly demonstrates that similarity correction considering intrinsic characteristics is feasible, which is different from the traditional similarity measures considering the basic data characteristics. Compared with CS, ED, and PC, SWS is the best in the similarity heat map for the Raman spectra of five cell lines. SWS can be effectively used in classification and clustering, and has the best performance in Raman data analysis and can significantly improve the performance of general machine learning algorithms based on similarity. Taken together, SWS not only opens an avenue for constituting a similarity measure of well-used basic data and intrinsic characteristics, but also highlights the possibility of data analysis in most of the Raman and infrared technologies, showing promise to achieve wide application in machine learning.

4. Experimental Section

Database: The database contains 1003 spectra of five mammmary cell lines, including 207 spectra of MCF-10A, 210 spectra of MDA-MB-231, 220 spectra of BT-474, 168 spectra of SK-BR3, and 220 spectra of T-47D cells. The classification was verified by tenfold cross-validation. The database was divided into ten parts, among which nine parts were training databases and one part was the test database.

All cell lines were acquired from the Cell Resource Center (Peking Union Medical College, China). The cell lines were confirmed to be free of mycoplasma contamination by polymerase chain reaction (PCR) and culture. They were cultivated in Dulbecco's Modified Eagle's Medium (DMEM), which included 1% penicillin–streptomycin solution and 10% fetal bovine serum (FBS). Cells were incubated at 37 °C, 5% CO2, and 95% air. Before use, the cells were washed three times with phosphate buffered saline (PBS) solution. For distributing the cells evenly, cells were spun at 1000 rpm for 3 min after each cleaning.

Single cells were distributed on the silicon wafer to obtain the single-cell Raman spectrum. The Finder One Raman spectrometer (Zolix, China), including a multimode diode laser with an excitation wavelength of 532 nm and a Peltier-cooled back-illuminated deep depletion CCD detector, was used to acquire spectra with an NA = 0.65, 40× long working distance objective lens. The integration time was 5 s.

Preprocessing: The spectral data were preprocessed and analyzed by Python and MATLAB. All spectra were from 600 to 1800 cm⁻¹. When measured, the silicon wafer was the substrate, and the spectra data had a strong silicon peak near 968 cm⁻¹. First, silicon was used as the internal parameter to normalize the cell data and the blank control data, and then all spectra data were subtracted from the mean spectra of the blank control group to remove noise. Following the normalization, an adaptive iterative reweighting penalty least squares (airPLS) algorithm was used to calibrate the baseline of the spectrum.

Biological Analysis of Spectra: Raman spectra contain complex biological characteristics, and the contribution of different information to experimental results varies greatly. A Raman spectrum segment scheme was designed to divide the spectrum into five parts. The typical discriminative features of each segment were investigated based on the Raman spectrum literature of biological specimens (Table 1).

Segment Weighting Similarity: SWS describes the similarity in the spectra by considering the basic data and intrinsic biological characteristics of the spectra in the calculation. It utilizes a segment scheme that treats each segment as a unit. In each section, the weighting method is used to help the “good features” get higher weightings, thereby increasing the contribution of these data to the results. “Bad features” get lower weightings to reduce the noise signal.

Similarity Heat Map: The similarity heat map provides a visual expression of similarity across the five kinds of spectra and reflects the ability of each similarity measure to distinguish the degree of similarity among spectra. To quantify the results of these maps, three indexes named “intraclass,” “interclass,” and “difference” are provided. “Intraclass” is the mean similarity of the same labeled data. “Interclass” is the mean similarity of the different labeled data. The difference between intraclass and interclass is calculated, which is named “difference.” Similarity measures with large differences perform better.

There are the other three similarity measures

\[
D = \sqrt{\sum_{k=1}^{n} (x_{1k} - x_{2k})^2}
\]  

\[
D = ED to calculate the straight distance of vector \(x_1\) and vector \(x_2\).
\]

\[
\cos = \frac{\sum_{k=1}^{n} x_{1k} x_{2k}}{\sqrt{\sum_{k=1}^{n} x_{1k}^2} \sqrt{\sum_{k=1}^{n} x_{2k}^2}}
\]  

Cos is CS to describe the relation of angle between vector \(x_1\) and vector \(x_2\) in space.

\[
r = \frac{\sum_{k=1}^{n} (x_{1k} - \bar{x}_1)(x_{2k} - \bar{x}_2)}{\sqrt{\sum_{k=1}^{n} (x_{1k} - \bar{x}_1)^2} \sqrt{\sum_{k=1}^{n} (x_{2k} - \bar{x}_2)^2}}
\]  

\[r\] is PC, which represents the degree of correlation about two vectors. \(r\) is from -1 to 1. The closer \(r\) gets to 1, the higher the degree of similarity is; otherwise, the lower the degree of similarity is.

To convert distance into similarity, the following formula is used.

\[
Sim = \frac{1}{1 + Dis}
\]

Classification: KNN is a classical distance-based classifier, which is widely used in text classification, image recognition, and spectral recognition. The reference spectra were sorted according to the similarity between the reference spectra and the query spectra, and the unknown spectral information was determined according to the top K reference spectral labels. The performance of the four similarity measures is discussed through classification results.

Clustering: Clustering is an unsupervised machine learning method to group a batch of unlabeled data. The key of a clustering method is the similarity measure that divides the data into specific clusters by similarity. Based on the four similarity measures, six clustering tasks were clustered by k-means and spectral clustering.

Evaluation Methods: Evaluation methods are vital to measure the performance of similarity formulas. Classification results are evaluated by accuracy, sensitivity, specificity, ROC curves, and AUC.

\[
Acc = \frac{TP + TN}{P + N}
\]  

\[
Sen = \frac{TP}{TP + FN}
\]  

\[
Spe = \frac{TN}{TN + FP}
\]  

Clustering results are evaluated by purity, recall, and accuracy from the labels of the given data. The calculation of purity is the same as that of accuracy.

\[
Pre = \frac{TP}{TP + FP}
\]

Adv. Intell. Syst. 2022, 4, 2100093 2100093 (9 of 11) © 2021 The Authors. Advanced Intelligent Systems published by Wiley-VCH GmbH
Rec = \frac{TP}{TP + FN} \quad (12)

For formula (8)-(12), P means the number of positive samples. N means the number of negative samples. TP means true positive, which represents the number of positive examples that are correctly classified as positive samples. TN means true negative, which represents the number of negative samples that are correctly classified as negative samples. FP means false positive, which represents the number of negative samples that are wrongly classified as positive samples.

Data Availability: The training and test database were set up on the cell line samples. Raw data to reproduce Figure 2 can be shared upon request.

Code Availability: All the analyses can be reproduced by Spyder 4.2.1 and MATLAB 2014. The SWS code is available on GitHub (https://github.com/biolightlab/Segment-Weighted-Similarity). Users can install all Python packages automatically. The packages used are: Python packages (scikit-fusion V1).

Supporting Information
Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements
This work was supported by a grant from the National Natural Science Foundation of China (81772017) and the National Key R&D Program of China (2017YFA0205202). The equations (1), (2), and (3) were corrected on March 14, 2022, after initial publication online.

Conflict of Interest
The authors declare no conflict of interest.

Data Availability Statement
Research data are not shared.

Keywords
biological characteristics, machine learning, segments, similarity, weighting

Received: May 17, 2021
Revised: August 10, 2021
Published online: October 17, 2021

[1] E. Petrakis, K. Tzeras, Software Pract. Exper. 2015, 50, 1447.
[2] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, Y. Wu, in 2014 IEEE Conf. on Computer Vision and Pattern Recognition, Vol. 1, IEEE, Piscataway, NJ 2014, pp. 1386.
[3] X. Xu, L. Zhao, Q. Xue, J. Fan, Q. Hu, C. Tang, H. Shi, B. Hu, J. Tian, Anal. Chem. 2019, 91, 7973.
[4] I. W. Schie, R. Kiselev, C. Krafft, J. Popp, Analyst 2016, 141, 6387.
[5] E. Baria, R. Cicchi, F. Malentaclchi, I. Mancini, P. Pinzani, M. Pazzagli, F. Pavone, J. Biophotonics 2020, 14, 202000365.
[6] C. S. Ho, N. Jean, C. A. Hogan, L. Blackmon, S. S. Jeffrey, M. Holodniy, N. Banaei, A. A. E. Saleh, S. Ermon, J. Dionne, Nat. Commun. 2019, 10, 4927.
[7] S. Y. Ji, D. H. Jeong, M. Hassan, I. K. Ilev, IEEE J. Sel. Top. Quant. 2018, 25, 7209008.
[8] S. S. Khan, M. G. Madden, Chemommetr. Intell. Lab. 2012, 114, 99.
[9] L. S. Lawson, J. D. Rodriguez, Anal. Chem. 2016, 88, 4706.
[10] Z. Lin, X. Zhao, J. Huang, W. Liu, Y. Zheng, X. Yang, Y. Zhang, M. de la Chapelle, W. Fu, Analyst 2019, 144, 2803.
[11] F. Cui, Y. Yue, Y. Zhang, Z. Zhang, H. S. Zhou, ACS Sensors 2020, 5, 3346.
[12] L. Yi, L. Zhao, X. Xue, H. Cheng, J. Tian, Lab Chip 2020, 20, 931.
[13] C. Carey, T. Boucher, S. Mahadevan, P. Bartholomew, M. D. Dyer, J. Raman Spectrosc. 2015, 46, 894.
[14] M. B. Fenn, V. Pappu, P. C. Georgeiv, P. M. Pardalos, J. Raman Spectrosc. 2013, 44, 939.
[15] A. Hufner, W. L. Kuan, S. L. Mason, R. A. Barker, S. J. C. S. Mahajan, Chem. Sci. 2019, 11, 525.
[16] Z. M. Movaghasi, S. Rehan, I. U. Rehman, Appl. Spectrosc. Rev. 2007, 42, 493.
[17] J. Schmitt, H. C. Flemming, Int. Biodeter. Biodegr. 1998, 41, 1.
[18] D. Naumann, App. Spectrosc. Rev. 2001, 36, 239.
[19] K. Z. Liu, M. Xu, D. A. Scott, Brit. J. Haematol. 2007, 136, 713.
[20] G. Bhattacharya, K. Ghosh, A. S. Chowdhury, Pattern Recogn. 2017, 66, 425.
[21] E. Arvaniti, M. Claassen, Nat. Commun. 2017, 8, 14825.
[22] K. Seddiki, P. Saudemont, F. Precioso, N. Ogrinc, M. Wizsztorski, M. Salzet, I. Fournier, A. Droit, Nat. Commun. 2020, 11, 5595.
[23] D. M. Camacho, K. M. Collins, R. K. Powers, J. C. Costello, J. J. Collins, Cell 2018, 173, 1581.
[24] W. Liu, Y. Tao, R. Xue, C. Song, Q. Wu, Y. Ren, Electrophoresis 2021, 42, 939.
[25] J. Chan, D. Taylor, T. Zwerdling, S. Lane, K. Ihara, T. Huser, Biophys. J. 2006, 90, 648.
[26] J. Deng, Q. Wei, M. Zhang, Y. Wang, Y. Li, J. Raman Spectrosc. 2005, 36, 257.
[27] J. Binoy, J. P. Abraham, I. H. Joe, V. S. Jayakumar, G. R. Pettit, O. F. Nielsen, J. Raman Spectrosc. 2004, 35, 939.
[28] Z. Huang, A. McWilliams, H. Lui, D. McLean, S. Lam, H. Zeng, Int. J. Cancer 2003, 107, 1047.
[29] N. Stone, C. Kendall, N. Shepherd, P. Crow, H. Barr, J. Raman Spectrosc. 2002, 33, 564.
[30] M. Gniadecka, H. C. Wulf, N. N. Mortensen, O. F. Nielsen, D. H. Christensen, J. Raman Spectrosc. 1997, 28, 125.
[31] G. Shetty, C. Kendall, N. Shepherd, N. Stone, H. Barr, Brit. J. Cancer 2006, 94, 1460.
[32] D. Lau, Z. Huang, H. Lui, C. Man, K. Berean, M. Morrison, H. Zeng, Laser. Surg. Med. 2003, 32, 210.
[33] D. Naumann, in Infrared Spectroscopy: New Tool in Medicine, Proc. of the SPIE, San Jose, Calif, USA, 1998, 3257, pp. 245–257.
[34] W. T. Cheng, M. T. Liu, H. N. Liu, S. Y. Lin, Microsc. Res. Techin. 2005, 68, 75.
[35] I. Notingher, C. Green, C. Dyer, E. Perkins, N. Hopkins, C. Lindsay, L. L. Hench, J. R. Soc. Interface 2004, 1, 79.
[36] R. K. Dukor, in Handbook of Vibrational Spectroscopy, Vol. 5, John Wiley & Sons, Ltd, New York 2006, pp. 3335.
[37] L. Silveira, S. Sathiaiah, R. Zangaro, M. Pacheco, M. Chavez, C. Pasqualucci, Laser. Surg. Med. 2002, 30, 290.
[38] C. Krafft, L. Neudert, T. Simat, R. Salzer, Spectrochim Acta A 2005, 61, 1529.
[39] P. G. L. Andrus, R. D. Strickland, Biospectroscopy 1998, 4, 37.
[40] A. J. Ruiz-Chica, M. A. Medina, F. Sánchez-Jiménez, F. J. Ramírez, *J. Raman Spectrosc.* 2004, 35, 93.
[41] R. J. Lakshmi, V. B. Kartha, C. M. Krishna, J. G. R. Solomon, G. U. U. Devi, *Radiat. Res.* 2002, 157, 175.
[42] U. Utzinger, D. Heintzelman, A. Mahadevan-Jansen, A. Malpica, M. Follen, R. Richards-Kortum, *Appl. Spectrosc.* 2001, 55, 955.
[43] A. Mahadevan-Jansen, R. Richards-Kortum, in *Proc. of the 19th Annual Int. Conf. of the IEEE*, Vol. 6, 1997, pp. 2722.
[44] E. O. Faolain, M. B. Hunter, J. M. Byrne, P. Kelehan, F. M. Lyng, *Vib. Spectrosc.* 2005, 38, 121.
[45] E. B. Hanlon, R. Y. Manoharan, T. W. Koo, K. E. Shafer, M. S. Feld, *Phys. Med. Biol.* 2000, 45, R1.