THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB Article

Similarity Metrics for Subcellular Analysis of FRET Microscopy Videos

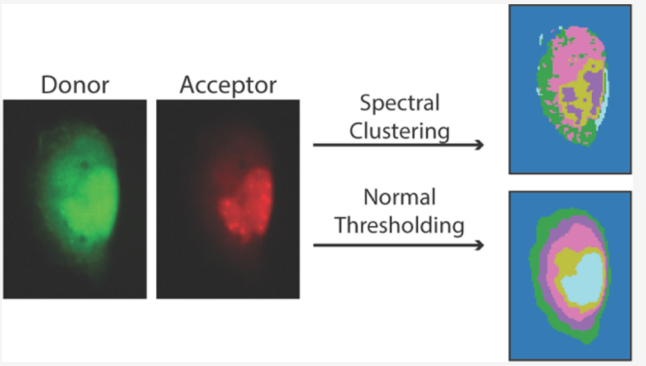
Published as part of The Journal of Physical Chemistry B special issue "Advances in Cellular Biophysics".

Michael J. Burke, Victor S. Batista,* and Caitlin M. Davis*

Cite This: *J. Phys. Chem. B* 2024, 128, 8344–8354 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Understanding the heterogeneity of molecular environments within cells is an outstanding challenge of great fundamental and technological interest. Cells are organized into specialized compartments, each with distinct functions. These compartments exhibit dynamic heterogeneity under high-resolution microscopy, which reflects fluctuations in molecular populations, concentrations, and spatial distributions. To enhance our comprehension of the spatial relationships among molecules within cells, it is crucial to analyze images of high-resolution microscopy by clustering individual pixels according to their visible spatial properties and their temporal evolution. Here, we evaluate the effectiveness of similarity metrics based on their ability to facilitate fast and accurate data analysis in time and space. We discuss the capability of these metrics to differentiate subcellular localization, kinetics, and structures of protein-RNA interactions in Forster resonance energy transfer (FRET) microscopy videos, illustrated by a practical example from recent literature. Our results suggest that using the correlation similarity metric to cluster pixels of high-resolution microscopy data should improve the analysis of high-dimensional microscopy data in a wide range of applications.



INTRODUCTION

Recent advances in instrumentation, alongside enhanced computational power and more affordable data storage solutions, have transformed the field of microscopy.[1] These developments have facilitated the acquisition of extensive and intricate multidimensional data sets, particularly in the context of time series data.[2] To gain a deeper understanding of complex chemical and biological interactions, new imaging techniques are being developed to investigate phenomena with both temporal and spatial resolution.[3−6] Despite the exciting prospects of collecting times-resolved images, a significant challenge accompanies them: the effective analysis and extraction of meaningful insights from the abundant information available.[7,8]

Many microscopy video analysis approaches focus on iterative image processing that treats each frame independently, whether for object tracking, interaction analysis, or monitoring morphology changes.[9−11] Fewer analysis approaches consider the relationship between frames in the time-series data, although there has been progress in similar multidimensional imaging fields.[12−14] Time-series analysis is ordered and, in the case of microscopy data, regularly sampled data over a continuous event.[15] To fully resolve an event, it is often necessary to consider the entire time series rather than analyzing specific frames individually.

Clustering methods are particularly valuable, as they allow for unsupervised classification of large data sets. While deep learning and Kernel-based methods have shown great promise in microscopy image analysis, they require the use of labels or significant tuning of hyperparameters as compared to unsupervised methods which can be applied without a labeled data set.[16−18] When applied to time-resolved microscopy, clustering pixels with high similarity distinguishes regions where the pixels exhibit more similarity among themselves than with other pixels in the image.[19−23] By effectively handling the large and complex data sets generated by microscopy, time-series clustering approaches advance our ability to analyze molecular composition and behavior at the microscopic level, paving the way for advancements in fields ranging from biomedical research to materials science. While many of the methods presented have been available for a while, our goal is to provide general guidelines for selecting and implementing

Received: April 30, 2024
Revised: July 16, 2024
Accepted: August 14, 2024
Published: August 26, 2024

https://doi.org/10.1021/acs.jpcb.4c02859
J. Phys. Chem. B 2024, 128, 8344−8354

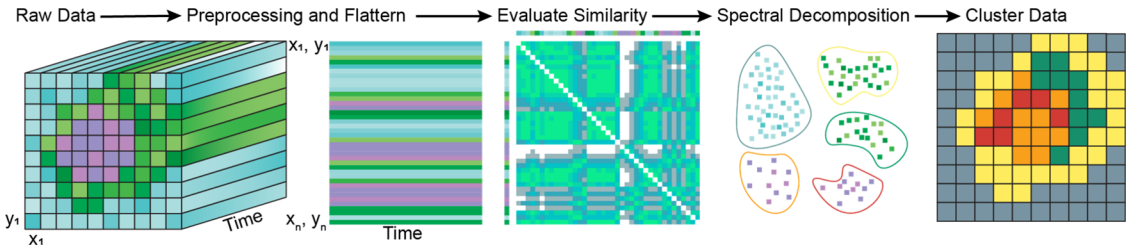


Figure 1. Data workflow for spectral clustering of FRET microscopy videos. Each pixel in a microscopy video is in a 3-dimensional matrix, with two positional coordinates and a time coordinate. To compare signal across time, the data is flattened into a 2-dimensional matrix where preprocessing methods such as normalization and smoothing can be applied. For spectral clustering, an adjacency matrix is generated by evaluating the similarity of each pixel or sample. This similarity metric is chosen based on the desired comparison. Since the matrix is symmetric, it can be efficiently spectrally decomposed to obtain the eigenvalues of the adjacency matrix. This acts as a dimensionality reduction so that clustering can be performed more efficiently in fewer dimensions.

clustering approaches and to demonstrate their application in the context of time-resolved Forster resonance energy transfer (FRET) microscopy videos.

Challenges of Clustering Multidimensional Data. Multidimensional data, such as time-series videos, can pose challenges for clustering methods when organized into manageable groups. The three main problems are noise in the sample, high dimensionality, and the ordered nature of the data.

How to Deal With Noise? Noise in data poses a significant challenge for time-series clustering, because it can distort and obscure underlying patterns within temporal sequences. Clustering algorithms aim to group similar time series, but noise disrupts this process by introducing spurious similarities and causing dissimilar sequences to be erroneously grouped. This not only diminishes the accuracy and reliability of clustering results but also hinders the identification of meaningful patterns and trends within the data. Robust and effective time-series clustering requires techniques that can mitigate the impact of noise, such as noise filtering, outlier detection, and noise-tolerant distance measures, to ensure accurate and meaningful grouping of temporal sequences.

Noise filtering in time series analysis involves removing or reducing unwanted variations, disturbances, or inaccuracies present in the data, making it easier to identify and analyze meaningful patterns and trends. This process allows for a more accurate and reproducible interpretation of the data.[24] Despite these benefits, implementing noise filtering in analysis workflows is not always practical. When dealing with data of high dimensionality, it is challenging, if not impossible, to ensure that the filtering performs as intended across every dimension. Aggressive noise filtering can lead to the loss of important information or introduce artifacts, both of which affect the resultant clustering.[25] This issue is exacerbated by the nature of time-series data, as noise filtering may not work well for nonstationary data where noise characteristics change over time, making it challenging to find a single filtering approach that suits the entire data set.[26]

Dimension reduction techniques play a crucial role in reducing noise within data by capturing the most salient and informative features, while discarding or minimizing the impact of noise-related dimensions. These techniques work by transforming the original high-dimensional data into a lower-dimensional representation. In doing so, they inherently filter out noise and emphasize the underlying patterns and structures within the data. By focusing on dimensions that explain the most variance or exhibit the strongest relationships, dimension reduction effectively highlights the meaningful variations in the data, while attenuating the influence of noisy and less informative dimensions. This reduction also simplifies the complexity of the data, making them more easily manageable and interpretable.

Curse of Dimensionality. The primary aim of clustering methods is to analyze numerous data points and identify discernible communities or subsets. Analyzing thousands of time series for patterns poses a significant challenge for researchers. Similarity metrics require comparing each pixel against all others, causing the number of comparisons to increase quadratically with the number of pixels. Consequently, strategies that reduce the number of pixels while preserving the essential processes can substantially reduce the computation time required.

Many data dimensionality approaches reduce the amount of information in the time dimension while maintaining a constant number of compared samples or pixels.[27,28] However, it is also important to consider reducing the number of pixels before further analysis. The simplest example of this is thresholding, which focuses the analysis on key spatial regions. Nevertheless, even when the analysis is focused on key areas, one may still encounter a large number of data points. Thresholds can also be used to categorize the pixels into a predetermined number of regions. Although these binning strategies are effective and efficient, they reduce the spatial resolution of the imaging technique without taking their behavior in the time domain into account.

Agglomerative hierarchical clustering methods reduce the number of pixels by considering their similarity in time and space.[29] However, they are much slower than the thresholding and binning strategies described above. Indeed, because they require n^2 calculations for n samples, which must be recalculated as the number of samples decreases, these approaches can consume more computational time than spectral clustering. Using local hierarchical agglomerative clustering that only compares similarity with neighboring pixels for binning (Figure S1), and therefore scales linearly with number of pixels, we have found that each round of this informed binning reduces the number of data points by a factor of 3−5.

Relating Ordered Data. Time series data are particularly challenging to analyze for similarity because it is an ordered data set. A data point at time t is related to time points $t-1$ and $t+1$. Depending on the shape and nature of the data, various methods of quantifying this similarity may prove to be useful.

Spectral clustering is a graph-based technique used for clustering data points by leveraging the relationships between them. (Figure 1) This method can accommodate various

https://doi.org/10.1021/acs.jpcb.4c02859
J. Phys. Chem. B 2024, 128, 8344−8354

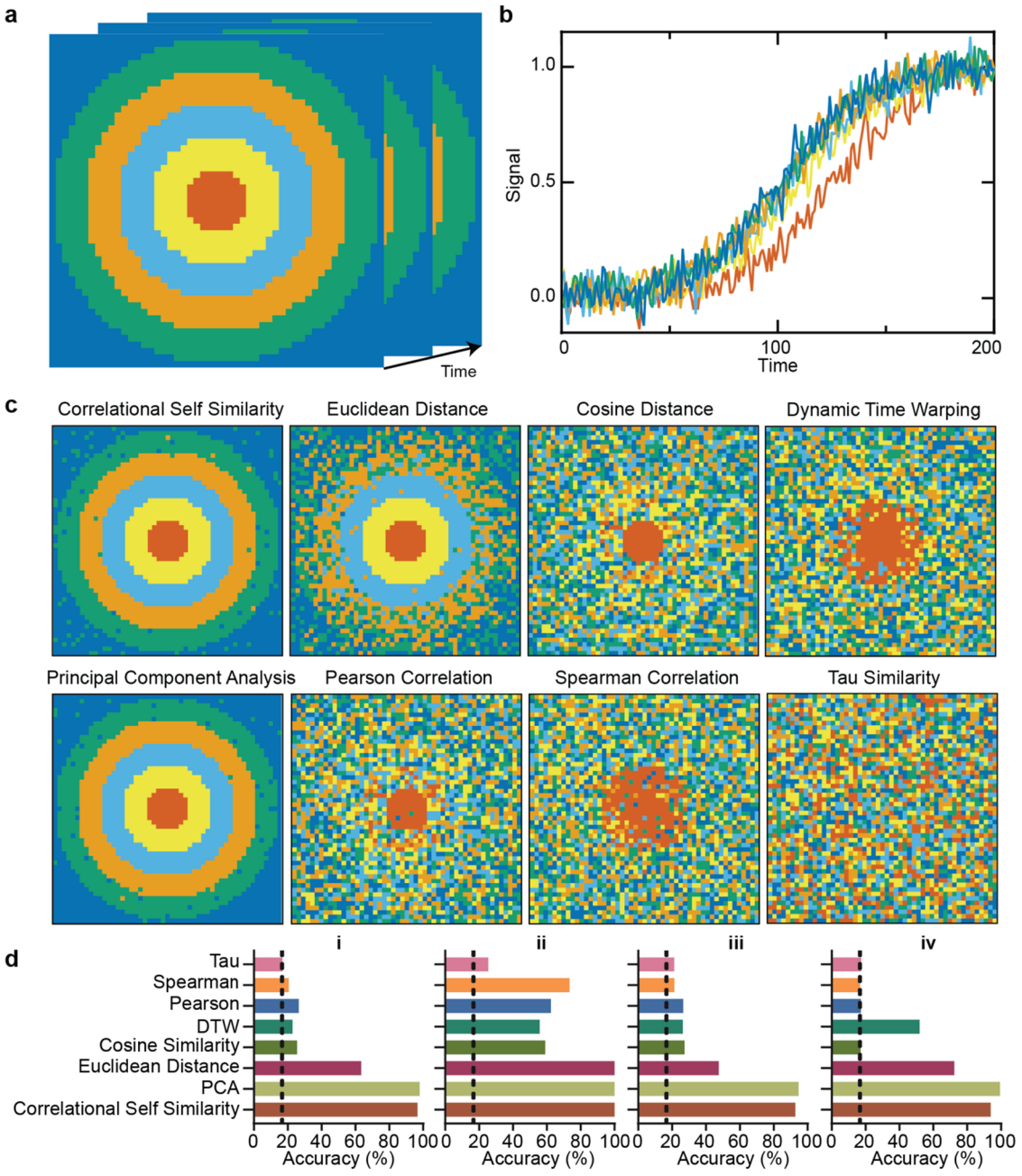


Figure 2. Ability of similarity metrics to distinguish function parameters. (a) A 51 × 51 pixel grid was subdivided into six regions, each simulating a specific function whose parameters are varied across groups, as detailed in Table S1. These parameter variations model changes within the function's characteristics. (b) Representative time series from each region in (a) are depicted, with color coding consistent with the respective group. This helps illustrate how sensitive the clustering is to the different time series. (c) Clustering outcomes of each similarity metric, tested against simulated sigmoidal curves with varying midpoints. Each metric's ability to cluster pixels according to the underlying parameter variations is displayed, highlighting the differences in performance and suitability for this analysis context. (d) Quantitative evaluation of the clustering accuracy for each similarity metric, applied to two different functions, sigmoid and exponential decay, with varying parameters. The accuracy is shown across four scenarios: (i) sigmoid functions with a shifted midpoint, (ii) sigmoid functions with an altered slope, (iii) exponential decay functions with modified decay rates, and (iv) exponential decay functions with changing amplitude. These scenarios are designed to test each metric's sensitivity to specific types of parameter alterations within the function.

similarity metrics, adapt to different data distributions, and efficiently solve problems using standard linear algebra methods.[30,31] The relationships between them are defined by the construction of an affinity matrix that quantifies the similarity or dissimilarity between each pair of data points based on the selected similarity metric. This matrix is then transformed into a graph representation, where each data point becomes a node and edges represent pairwise similarities. Clusters are formed by partitioning the graph's Laplacian matrix, aiming to minimize the normalized cut or other graph-based criteria.[23,24] This process involves finding the eigenvectors associated with the smallest eigenvalues of the Laplacian matrix, which capture the optimal embedding of data points in a lower-dimensional space that enhances cluster separability.

Principal component analysis (PCA) is a technique related to spectral clustering, though it focuses less on clustering and more

https://doi.org/10.1021/acs.jpcb.4c02859
J. Phys. Chem. B 2024, 128, 8344−8354

on dimensionality reduction and feature transformation.[32−34] Similar to spectral clustering, PCA constructs the covariance matrix, a type of affinity matrix where the similarity is measured by the covariance between each point. However, instead of generating a graph Laplacian and using graph-based partitioning methods to cluster data, PCA-based clustering operates directly on the covariance matrix. Like spectral clustering, this process involves finding the eigenvectors and eigenvalues of the covariance matrix, but PCA-based methods typically use a k-means-based clustering scheme on a lower-dimensional representation of the data, thereby reducing the complexity of the data.

Due to the similarities between spectral clustering and PCA, and their ability to reduce the data to a lower-dimensional space, we explore both approaches in analyzing microscopy videos. While other methods of dimension reduction and similarity metrics exist,[7,28,30,35,36] we focus on spectral clustering and more commonly used similarity metrics to make this approach accessible to microscopists.

METHODS

Implementation of Spectral Clustering. Simulations were carried out in Python using a commercial laptop (Dell XPS 13, 7390 with an Intel(R) Core i7-10710U CPU). Custom code was used to calculate the adjacency matrices for correlational self-similarity, Euclidean distance, and cosine distance. SciPy[37] was used to calculate the adjacency matrices for the Pearson Correlation, Spearman Correlation, Kendall's Tau. These adjacency matrices were then used for the spectral clustering algorithm in Scikit-learn.[38] PCA clustering was performed with Scikit-Learns implementation of PCA with their K-means clustering algorithm.

Similarity Metrics. Similarity can be assessed according to data shapes, extracted features, and fitted models. Shape-based methods operate on raw data, while methods for extracted features and model parameters typically require a more supervised approach.[15] Similarity scores are defined using generalized distances that quantify the closeness of two data points within a data set. These distances vary depending on the chosen similarity metric. There is no one-size-fits-all solution for selecting a similarity metric. The effectiveness of a metric depends on the nature of the data and the specific information on interest. To illustrate this, we generate test data sets to help researchers choose similarity metrics most useful for their microscopy applications (Figure 2). For the similarity metrics below, we assess distance-based, correlational, and convolutional methods for time series X and Y, each consisting of n measurements.

Distance-Based Metrics. We compared three types of distance-based metrics: Euclidean distance, cosine similarity, and dynamic time warping (DTW). In these metrics, a greater distance indicates that the two vectors are less similar.

The Euclidean distance is the straight-line distance between two points in the n-dimensional space.[39] It is defined as follows:

$$\text{Euclidean distance}\ (\boldsymbol{X}, \boldsymbol{Y}) = \sqrt{\sum (\boldsymbol{x_i} - \boldsymbol{y_i})^2} \tag{1}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are the ith measurements of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. The Euclidean distance is useful for low-dimensional problems but becomes less practical in higher-dimensional spaces or when the data involve vectors with different magnitudes, different orientations, or significantly different amounts of noise. Additionally, the Euclidean distance does not take into account the order of the time-series data.

The cosine similarity metric measures the cosine of the angle between two vectors, providing a measure of orientation irrespective of magnitude.[39] It is especially useful for high-dimensional space analysis. The cosine similarity is given by

$$\text{cosine distance}\ (\boldsymbol{X}, \boldsymbol{Y}) = \frac{\boldsymbol{X}\cdot\boldsymbol{Y}}{\|\boldsymbol{X}\|\,\|\boldsymbol{Y}\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{2}$$

where $\boldsymbol{X}\cdot\boldsymbol{Y}$ is the dot product between vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, and $\|\boldsymbol{X}\|$ and $\|\boldsymbol{Y}\|$ are the norms (or magnitude) of vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. The coordinates used for assessing the Euclidean distance and cosine similarity are extremely restrictive and must be in sync. However, in practice, microscopy data may not be evenly sampled or may be out-of-phase, which effectively accelerates or decelerates the features.

DTW measures the similarity between two-time series, which may vary in time or speed. For instance, similarities in walking patterns could be detected, even if one walker was walking faster than the other. DTW is defined through an optimal alignment of the sequences. It is obtained by nonlinearly warping the time series to minimize their differences, as follows:[40]

$$\text{DTW}\ (\boldsymbol{X}, \boldsymbol{Y}) = \min_{\pi \in A(\boldsymbol{X},\boldsymbol{Y})} \sum_{(i,j)\in\pi} \|\boldsymbol{x}_i - \boldsymbol{y}_j\| \tag{3}$$

where π is an alignment path consisting of index pairs and $A(\mathbf{X},\mathbf{Y})$ represents the set of all admissible paths. These paths satisfy constraints that ensure that the sequences start and end are aligned, and the indices monotonically increase, with each index appearing at least once. Therefore, DTW allows the identification of similar shapes of data sets, even when the coordinates or magnitude of the data are quite different. For each of these concepts of generalized distances, a larger distance implies greater dissimilarity.

Correlation-Based Metrics. We evaluated the capabilities of three correlation metrics based on generalized distances: the Pearson correlation coefficient, Spearman correlation, and Kendall's Tau Similarity. To address issues of magnitude, these correlation metrics measure the degree to which two sets of data are linearly related by comparing trend or shape similarity.

The Pearson correlation metric measures the strength of the linear relationship between two variables.[41] It assesses the extent to which one variable increases or decreases with another one. The Pearson correlation coefficient is calculated as follows:

$$\text{Pearson correlation}\ (\boldsymbol{X}, \boldsymbol{Y}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{4}$$

Here, $\bar{x}$ and $\bar{y}$ are the means of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. This coefficient ranges from +1, indicating a perfect positive linear correlation, to −1, indicating a perfect anticorrelation, with 0 indicating no correlation.

Spearman Rank Correlation Coefficient metrics first convert the data into ranks and then calculate the Pearson correlation coefficient of these ranks. Spearman's rank correlation can be expressed, as follows:[42]

https://doi.org/10.1021/acs.jpcb.4c02859
J. Phys. Chem. B 2024, 128, 8344−8354

$$\text{Spearman correlation } (\boldsymbol{X}, \boldsymbol{Y}) = \frac{\mathbb{E}[(R_X - E[R_X])(R_Y - E[R_Y])]}{\sigma_{R_x}\sigma_{R_y}} \quad (5)$$

where R_X and R_Y are the rank transformations of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, $E[R_X]$ and $E[R_Y]$ are the expected values of these ranks, and σ_{Rx} and σ_{Ry} are the standard deviations of the ranks. The coefficient also ranges from +1 to −1. Unlike the Pearson correlation, Spearman correlation is particularly effective at identifying monotonic relationships, regardless of whether the correlation is linear or not. However, it is highly sensitive to errors and outliers.[43]

The Kendall's Tau coefficient metric also uses a ranking system and assesses the degree of correspondence between the orders of the data points. It compares pairs of observations to see if their ranks are correlated, as follows:[44]

$$\text{Tau similarity } (\boldsymbol{X}, \boldsymbol{Y}) = \frac{2}{n(n-1)} \sum_{i<j} \text{sgn}(x_i - x_j)\text{sgn}(y_i - y_j) \quad (6)$$

where sgn is the sign function, indicating the direction of the difference. Kendall's Tau measures the difference between the number of concordant and discordant pairs, normalized by the number of possible pairs. It ranges from +1 (perfect agreement) to −1 (perfect disagreement), with 0 indicating no correlation.

Convolutional-Based Metrics. Autocorrelation analysis enables the assessment of self-similarity between data and its lagged version. This analysis is particularly useful for detecting repeatable patterns within a time-ordered data set, utilizing scoring metrics such as those described above.[45] This approach is widely used to analyze time- and frequency-dependent data, such as fluorescence correlation microscopy and infrared spectroscopy data.[5,14] Autocorrelation analysis often requires additional analysis or parameter fitting before it can be compared with other data points.

To generalize and facilitate the comparison of the time series data, we used a correlational self-similarity metric that involves both normalization and compression. It measures the cross-correlation between two data sets, normalized by the maximum autocorrelation of each data set. For two pixels represented by time series $\boldsymbol{X}$ and $\boldsymbol{Y}$ over a window size w, the correlation self-similarity (CSS) is calculated as

$$\text{CSS } (\boldsymbol{X}, \boldsymbol{Y}) = \frac{\sum_{n=0}^{w} \text{corr}(\boldsymbol{X} \times \boldsymbol{Y})[[n]]}{\max(\sum_{n=0}^{w} \text{corr}(\boldsymbol{X} \times \boldsymbol{X})[[n]], \sum_{n=0}^{w} \text{corr}(\boldsymbol{Y} \times \boldsymbol{Y})[[n]])} \quad (7)$$

The cross-correlation $\text{corr}(X \times Y)[n]$ for a lag n is

$$\text{corr } (X \times Y)[[n]] = \sum_{m=0}^{N-1} X[m]Y[m+n] \quad (8)$$

This measure is normalized by the maximum of the sums of the autocorrelations of X and Y over the window w, allowing a relative comparison of similarity that accounts for the strongest internal correlations of each data set. Window sizes were 0 (Figures 2 and 5) and 3 (Figures 3 and 4). These convolution-based metrics, through the use of correlation calculations over lagged intervals, provide a framework for understanding the dynamics and similarities within and between time-series data sets.

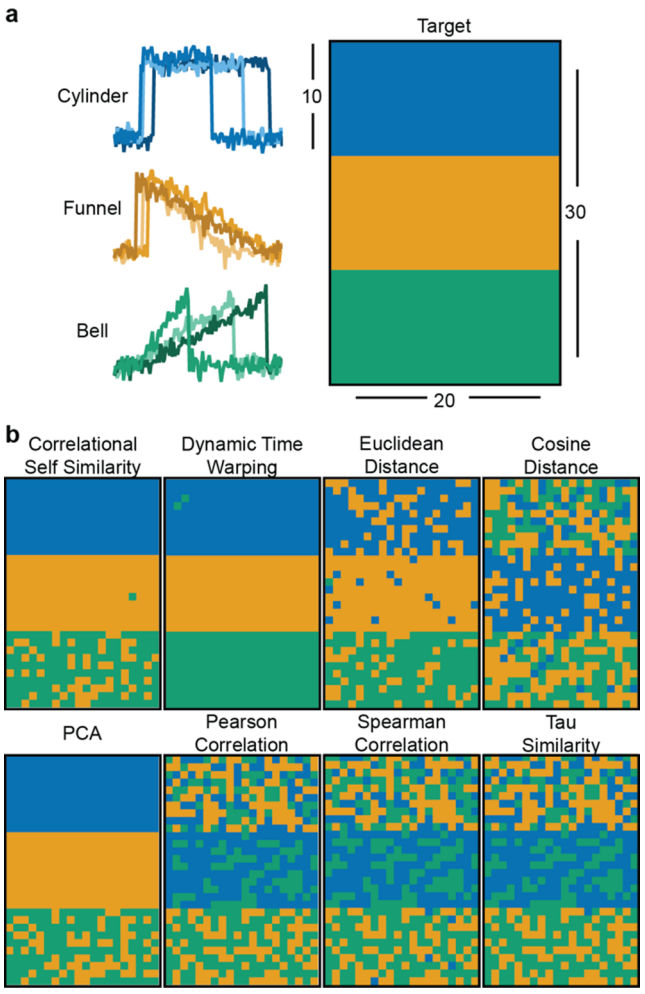


Figure 3. Comparative clustering of similarity metrics on cylinder-bell-funnel (CBF) data set. (a) Example time series for the three illustrative of the three function types: cylinder (blue), funnel (orange), and bell (green). A target segmentation map delineates the ideal clustering arrangement with each function type assigned to discrete segments represented by different colors in the vertical layout: cylinder (top), funnel (middle), and bell (bottom), signifying the benchmark for subsequent clustering comparison. (b) Results depicted are a typical clustering performance of each similarity metric when applied to the CBF data set with color consistency indicating greater accuracy. The variation in patterns shows the sensitivity of each similarity metric to the different shapes of the time series. Accuracies are quantified in Figure S4a.

RESULTS AND DISCUSSION

Evaluation of Similarity Metrics. In the upcoming discussion, we evaluate how similarity metrics assess the resemblance between various functions or time-series data. We divide our discussion into three sections, each examining how different similarity metrics measure and quantify similarity based on specific criteria. These tests will rigorously evaluate the performance of similarity metrics: measuring their sensitivity to changes in parameters, assessing their ability to capture different behaviors, and gauging their effectiveness in detecting changes in events occurring over time. Simulated time-series data provide a known target to test the ability and sensitivity of different similarity metrics. By controlling the cluster size, function parameters, and noise, we can quantify the merits and limitations of each approach. We aim to provide a

https://doi.org/10.1021/acs.jpcb.4c02859
J. Phys. Chem. B 2024, 128, 8344−8354

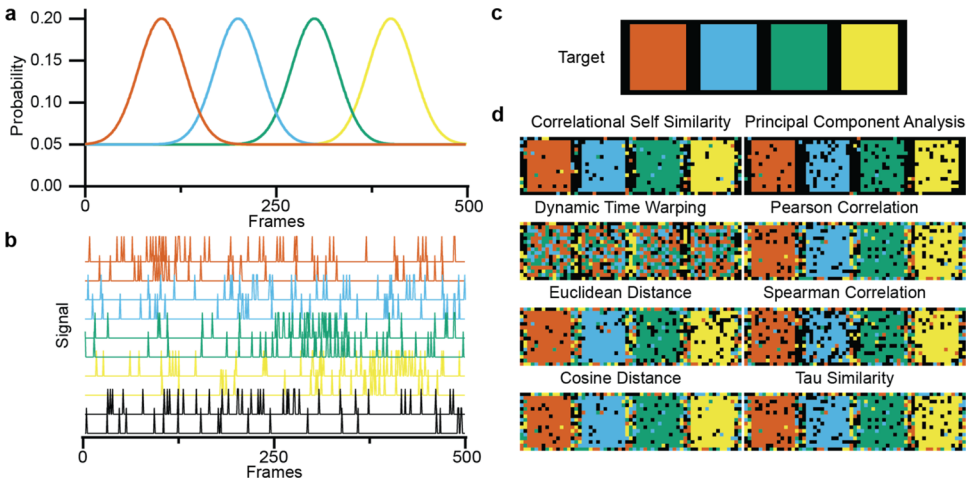


Figure 4. Similarity metric sensitivity to event frequencies. (a) The underlying probability distributions for spike occurrence across different regions of a video (red, blue, green, yellow, and black), consisting of overlapping Gaussian functions peaking at 20%. (b) Representative time-series data of spikes for two pixels taken from each region. It is unlikely that two pixels in the same region behave identically; however, they behave similarly based on their underlying probability distribution. The probability distributions in (a) are aligned with representative time-series data of individual pixels in (b). Pixels from the same region have a high probability of spiking when the probability distribution in that region is high. These distributions are combined with a baseline spike probability set at 5% (black) to model the stochastic nature of the spike generation. (c) A target data segmentation map is created based on the probability parameters, displaying the idealized classification of data into four categories, each color-coded to represent a different spike probability region. This map serves as a reference standard for clustering accuracy. (d) Results depicted are a typical clustering performance of each similarity metric when applied to the data set with color consistency indicating greater accuracy. The variation in patterns visualizes the sensitivities of each similarity metric to resolving the different regions in comparison to the background. Accuracies are quantified in Figure S4b.

comprehensive understanding of their utility and applicability in various microscopy scenarios.

Changes in Parameter. To examine changes in parameters within the same function, a sigmoid curve was chosen because sigmoids present a challenging clustering task due to the regions of high similarity at the beginning and end of the curves and smaller regions of change. The simulated data shown in Figure 2a consisted of a 51 × 51 array, with the distance from the center used to generate six different regions in a "target". Importantly, these clustering approaches do not consider spatial proximity and can just as easily cluster separated regions together. Based on their region (color), the pixels are populated with sigmoid functions of varying parameters with Gaussian noise added to each time point (Figure 2b). The center of the image represents the region with the highest dissimilarity. As the rings extend outward, the difference in parameters decreases to test the sensitivity of each metric. The accuracy was evaluated by comparing the defined simulated groups to the clustered results. A visual indication of the different abilities of similarity metrics to cluster sigmoids with varying midpoints is represented in Figure 2c. The midpoints used to generate each curve are 5, 2.5, 1, 0.5, 0.2, and 0.0. With a temporal resolution of 0.1, these represent changes in the midpoint of 50, 25, 10, 5, and 2 frames from the outer ring. Gaussian noise was added to each point, with a standard deviation of 0.05 applied to each time point. Results with varying levels of Gaussian noise are presented in Figure S2.

PCA and correlational self-similarity performed the best among the metrics, clustering the image with overall accuracies of 96.7 and 92.9%, respectively (Figure 2c,d). Both PCA and correlational self-similarity showed no errors in the three inner regions. For PCA, the accuracies of the outer three regions were 99.3, 96.2, and 91.7%, from the middle to outer regions, respectively.

Euclidean distance was able to cluster the image with an accuracy of 63.4%, perfectly resolving the first two groups and with few errors in the third group (95.7% accuracy). However, it poorly resolved the three outer groups, where the midpoint was the most similar. They showed accuracies around 50% (56.6, 38.8, and 58.9% from inside out). This demonstrates that Euclidean distance struggles to resolve neighboring groups that are more similar due to the noise present in the data. When tasked with clustering the data into 6 clusters, the rest of the metrics perform poorly.

It is worth noting that most metrics perform reasonably well when asked to cluster into only 2 regions (Figure S3). Each similarity metric, besides Kendall's Tau similarity, was able to divide the data set in Figure 1b into two communities. However, only Correlational Self Similarity separated the most dissimilar group from the others, even though it represented the smallest community. The other metrics separated the data set into two clusters but with a more equal distribution.

Changes in Function. The previous simulation compared the same function with different parameters (such as the midpoint, decay rate, and slope), but it is also important to test the ability of an approach to detect differences in shape. For this, we apply a well-known artificial 1-dimensional test data set for shape determination that uses three classes of time series.[46] The cylinder, bell, and funnel are generated based on 3 random variables, a, b, and n (Figure 3a). The variables adjust the functions to change the shape within each set where n adjusts the amplitude, a adjusts the starting point of the function, and b adjusts the ending point:

$$\text{Cylinder}\ (i) = (6 + n)\cdot\chi_{[a,b]}(i) + \in (i) \tag{9}$$

$$\text{Bell}\ (i) = (6 + n)\cdot\chi_{[a,b]}\cdot{(i - a)}/{b - a} + \in (i) \tag{10}$$

$$\text{Funnel}\ (i) = (6 + n)\cdot\chi_{[a,b]}\cdot{(b - i)}/{b - a} + \in (i) \tag{11}$$

$X_{[a,b]} = 1$ if $a \le t \le b$ and 0 if $a \ge t$ or $t \ge b$. This creates flat regions of Gaussian noise, $\epsilon(i)$, with regions of difference either

https://doi.org/10.1021/acs.jpcb.4c02859
J. Phys. Chem. B 2024, 128, 8344−8354

flat, decreasingly linear, or increasing linearly on the interval $[a,b]$. To evaluate the similarity metrics, the noise was gradually increased until all of the methods started to make errors. As anticipated, DTW outperforms the other metrics in its ability to cluster different shapes, correctly identifying 99.7% of the points. However, correlational self-similarity performs fairly well, correctly classifying 91.0% of the curves (Figure 3b). It confuses some cylinders and bells for funnels. Principal component analysis makes the same mistakes but with slightly less accuracy (88.8%).

DTW outperforms other metrics in separating data sets from different types of distribution functions. Intriguingly, DTW performs the worst of all metrics when tasked with separating the same shape of either bells, funnels, or cylinders, when the a and b are varied (Figure S5). DTW warps the two signals to minimize the differences between them, which allows it to group similar function shapes even if the parameters defining that function differ. This benefit is a detriment when tasked with separating similar data from different underlying variables. In this task, the warping likely masks the small differences and prevents it from properly clustering the data. For this reason, both Euclidean, correlational self-similarity metrics, and principal component analysis outperform dynamic time warping when similar curves are separated with different parameters. This demonstrates that DTW is best at perceiving differences in function, while correlational self-similarity and PCA are best at separating similar shapes with different parameters.

Changes in Time. In the realm of time-series clustering, understanding the frequency of events over time is a crucial aspect that can significantly impact the performance and accuracy of clustering algorithms. The frequency of events, such as peaks, troughs, or other distinctive patterns, can carry valuable information about the underlying dynamics and behavior of the time-series data. Incorporating frequency-based testing in time-series clustering serves several essential purposes that enhance the reliability and interpretability of the clustering results.

In this experiment, we aim to assess the efficacy of various similarity metrics in evaluating different time-series data generated through microscopy, particularly focusing on the ability to discern changes over time. The generated data consists of a sequence of 500 frames to represent a video (Figure 4a). Within these frames, five distinct groups are embedded, each characterized by unique temporal behaviors. One group serves as the background, exhibiting a consistent spike rate at a frequency of 0.05. The remaining four groups are centered at frames 100, 200, 300, and 400, respectively, and follow a Gaussian probability distribution. These groups also exhibit background spike frequencies of 0.05; however, around their center points, the probability peak is 0.20 (Figure 3b).

By subjecting this synthesized data set to various similarity metrics, we aim to explore their effectiveness in capturing the temporal changes introduced by the distinctive spike patterns within the data. This experiment offers valuable insights into the suitability of different similarity metrics for detecting and quantifying temporal variations in microscopy-generated time-series data, facilitating an informed choice of metrics for specific analytical contexts.

Through comparison to the target regions, we see that PCA clustering and correlation self-similarity most accurately produce the original groups with accuracies of 88.3 and 87.5%, respectively (Figure 3c,d). Pearson, Spearman, Kendall's Tau, Euclidean, and cosine scored accuracies of 68.9, 68.8, 68.6, 66.6, and 70.2%, respectively. Interestingly, while these metrics can find the signal groups, they struggle to separate the background from the signal. The correlation metrics seem to get it wrong the opposite way, assigning more background points to the signal areas. DTW also performs poorly on this task with a 31.4% accuracy, because each time series has a similar shape to each other.

Choosing a Similarity Metric. When choosing a similarity metric, it is important to consider how the time for a method to run scales off the size of the initial conditions. The number of samples (e.g., pixels in microscopy images), frames, or associated measurements in the time-series data set determines the overall computational cost of calculations based on the various similarity metrics (Table S2). In FRET microscopy videos the upper bound of samples is determined by the number of pixels on the camera, while the upper bound for the number of frames is determined by the data storage capacity of the camera.

The computational cost of spectral clustering depends on the number of samples, whereas the computational cost of PCA depends on the number of frames. PCA and spectral clustering are performed in two similar but different steps: first, the creation of a similarity matrix (spectral clustering) or a covariance matrix (PCA) and then the eigendecomposition of these matrices into a lower-dimensional space for clustering. The computational complexity, how the method's computational cost scales with the number of samples and frames, is dependent on how long it takes to create these matrices and the resultant size of the similarity or covariance matrix. The sizes of the matrices are different for spectral clustering and PCA. For spectral clustering, the similarity matrix size is determined by the number of samples. For PCA, the size of the covariance matrix is determined by the number of frames. The computational cost of the eigendecomposition of a matrix is dependent on the size of the matrix and scales with the third power of the matrix size.

The time required for each method with different numbers of frames and samples is evaluated in Figure S6. In general, similarity metrics scale linearly with respect to the number of frames and quadratically with respect to the number of points. The primary decider of similarity metric speed is whether it can be efficiently calculated with matrix manipulation or if it requires a direct comparison between every data point. Thus, Euclidean, cosine, Pearson, Spearman, correlational self-similarity, and PCA, which can all be calculated with transformations on an array, are faster than DTW and Kendall's Tau. DTW and Kendall's Tau both require a direct comparison of each point and n^2 calculations for a time series of n frames.

Considering only the faster metrics, Euclidean, cosine, and correlational self-similarity are the fastest with respect to increasing frames, whereas Spearman and PCA scale poorly (Figure S6b). Spearman scales poorly because it must first rank all of the data, a task that becomes more complicated with increasing numbers of frames. PCA scales quadratically with respect to the number of frames. However, PCA is the fastest with respect to increasing the number of samples to be compared, with correlational self-similarity performing the worst (after DTW and Kendall's Tau) (Figure S6d). PCA scales linearly with respect to an increasing number of samples, while the spectral clustering approaches will scale quadratically with respect to the number of sample comparisons. Correlational self-similarity scales poorly because of the normalization term, which must be applied in a second step.

In terms of computational time, the speed and flexibility of the Euclidean distance make it a reasonable starting point for many

https://doi.org/10.1021/acs.jpcb.4c02859
J. Phys. Chem. B 2024, 128, 8344−8354

researchers. Euclidean scales well with respect to the number of points and performs adequately in many types of applications. Alternatively, if the behavior of the system is known, an informed decision can be made based on computational time and the analyses performed in previous sections.

If Euclidean distance does not perform well and no preknowledge about the behavior of the system is known, correlational self-similarity, the best performer across all three tests, is a good option. Where the Euclidean distance, as well as Pearson and Spearman correlation metrics, are more susceptible to noise,[47,48] the correlational self-similarity metric is less susceptible to noise since it compares the data sets by cross-correlation, effectively "sliding" the time-series data sets with each other. The computational cost of correlational self-similarity is determined by the window width (i.e., the number m of elements in the window). For each slide, the correlational self-similarity method requires a series of $n \times m$ comparisons, scaling linearly with n and m. The choice of window size depends on the desired comparison; larger window sizes are appropriate when there is a preknowledge that grouped behaviors may be time-shifted. Nevertheless in a spectral clustering workflow, a well-tuned window parameter is less critical because two time-shifted time series can be clustered together by a third time-series that is similar to both. A window size of 0 will better separate difference in time (e.g., Figures 2 and 5), whereas a larger window will pregroup time-shifted data together depending on the window set (e.g., Figures 3 and 4 had a window size of 3). The resulting regularization introduced by cross-correlation reduces the impact of the noise and allows for the differentiation of similar time series with differences in their underlying parameters.

Although we did not explore combinations of similarity metrics here, in practice the computational cost of a combination of low-complexity techniques or a low- and high-complexity technique may be capable of classification at a lower computational cost. For example, DTW, the costliest metric tested, could be used to cluster data sets from different families of distribution functions after another clustering metric is used that is more effective at distinguishing data sets from the same type of distribution but with different parameters.

Limitations of Spectral Clustering. A limitation of this study is that many of the methods are presented according to their ability to separate relatively simple functions. This was done to help the reader understand how these methods are parsing out differences in information. However, if researchers are analyzing data with well-defined functions, they may wish to base their clusters based on derived or fitted parameters.

While useful for microscopy analysis, these methods may not be fast enough for big data applications, approaching millions of comparisons. If data sets contain more than hundreds of thousands of samples or frames, the best course of action would be to limit the amount of data using strategies presented in the introductory section Curse of Dimensionality. The similarity metrics investigated here scale linearly with frames and quadratically in samples; therefore, reducing the number of samples is more effective in lowering computation time. We suggest an informed binning scheme for data reduction, local hierarchical agglomerative clustering (Figure S1), that considers the relationship between data in time and space.

Application to Real Microscopy Videos. To demonstrate the application of the correlational self-similarity metric to real microscopy data, we cluster videos of the spliceosome protein U1A with its RNA binding partner stem-loop 2 (SL2) collected in living cells.[49] The binding process is monitored by probing the FRET efficiency between a donor fluorophore on U1A (Alexa 488) and an acceptor fluorophore on SL2 (Alexa 594) (Figure 5a,b). A two-color Fast Relaxation Imaging (FReI) setup enabled the investigation of thermodynamics and kinetics of binding between U1A-SL2.[50]

In FReI, the binding equilibrium is perturbed by temperature jumps induced by a 2 μm infrared laser, which rapidly heats the sample. The equilibrium effect of the temperature increase

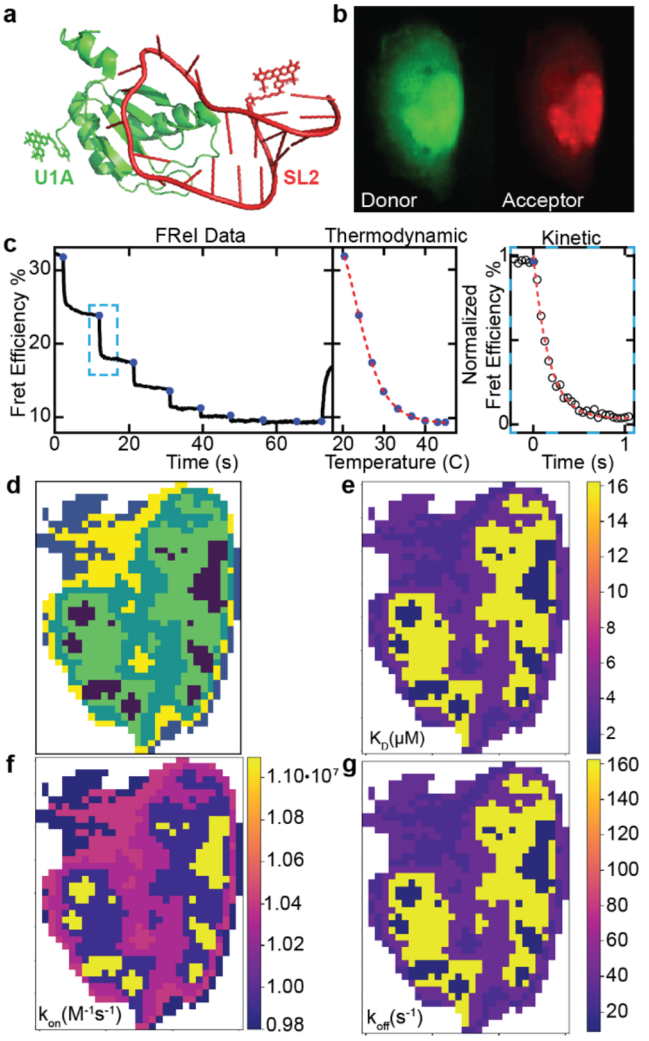


Figure 5. Analysis and clustering of U1A-SL2 binding dynamics using spectral clustering. (a) Schematic representation of the U1A-SL2 FRET construct showing the U1A protein (green) and SL2 RNA (red) bound together. (b) Fluorescence images from the initial frame of the FReI experiment displaying the donor (U1A labeled with Alexa 488, green) and acceptor (SL2 labeled with Alexa 594, red). (c) Example data of FReI experiment which gives us three outputs. (Left) FRET efficiency data over time, showing the change in energy transfer efficiency as a function of the temperature jumps and changes in interactions. (Middle) Thermodynamic profile of the binding interaction, with FRET efficiency plotted against temperature to assess the stability and binding changes under varying thermal conditions. (Right) Kinetic analysis data, depicting normalized FRET efficiency decay over time, indicative of the rates of binding and unbinding. (d) Clustering results from spectral clustering using correlational self-similarity. (e–g) Calculated K_d, k_{on}, and k_{off} values based on the average signal for each cluster. Results show that the changes in K_D (e) are from changes in the dissociation rates (g) and not the association rates (f).

https://doi.org/10.1021/acs.jpcb.4c02859
J. Phys. Chem. B 2024, 128, 8344−8354

provides thermodynamic information, while the response to the almost instantaneous temperature jump provides kinetic information, as the relaxation of the new bound and unbound populations are monitored in time (Figure 5c).[51] We have applied the correlation self-similarity metric to this data set to test whether U1A-SL2 binding is regulated by its surrounding local cellular environment. Our analysis was focused on the nucleus by using Otsu thresholding on the U1A fluorescence to separate the nucleus from the cytoplasm (Figure 5d). The data was clustered using correlational self-similarity, and the average signal of each resultant cluster was examined in the context of the bimolecular binding reaction. The results from the other similarity metrics are presented in Figure S7. Of the methods, only correlational self-similarity and PCA successfully segment the video, but correlational self-similarity is more computationally efficient due to the large number of frames (4560) compared to the number of pixels (933). Additionally, the data from each cluster was fit to the integrated rate laws for the bimolecular reaction to determine the binding affinity as well as the association and dissociation rates (Figure 5e,f).

The previous study found that the binding affinities in the nucleus ($K_D = 4.4 \pm 0.6 \times 10^{-6}$ μM) were slightly higher than those in the cytoplasm ($K_D = 3.0 \pm 0.8 \times 10^{-6}$ μM). Here, we further resolve two regions of binding in the nucleus, a lower binding affinity region exhibiting 16.3×10^{-6} μM affinity and a region of higher affinity of 0.9×10^{-6} μM. Consistent with earlier observations, we find that the differences in affinities result from differences in dissociation rates (Figure 5g) while the association rates are not significantly different (Figure 5f).

These findings are significant since the nucleus is rather heterogeneous and organized into compartments with distinct functional roles. The high-affinity regions are likely Cajal bodies, a site central for nuclear splicing and creation of the snRNP. The U1A protein and its RNA binding partner SL2 RNA are known to colocalize in Cajal bodies in the nucleus, which determines distinct binding properties since Cajal bodies are small and highly conserved subnuclear structures (i.e., 0.2−2 μm) in size depending on the organism.

Several different properties of Cajal bodies could be the source of this difference in affinity. One potential explanation is that the increased macromolecular crowding in the Cajal bodies could stabilize the U1A-SL2 complex compared to the rest of the nucleus. It was determined for *Xenopus* oocytes that the Cajal bodies had a macromolecule density of 0.136 mg/mL compared to the 0.106 g/mL value for the surrounding nucleoplasm.[52] However, while U1A-SL2 data fit with excluded volume theory and experiments that predict stabilization they do not agree with predictions of an increase in association rates.[53−56] While it is possible that an increase in association rates may be offset by lower diffusion from more crowded environments, our observations also support recent work that suggests that macromolecular crowding does not have large effects on association rates.[49,57,58] This is supported by the *in vitro* study of the U1A-SL2 complex with lysis buffer and cell lysate which suggested that weak nonspecific interactions destabilized the complex.[21]

CONCLUSIONS

Here we evaluate various similarity metrics and their effectiveness in distinguishing different signal types in microscopy videos. Lessons learned from clustering simulations were then applied to actual microscopy videos obtained by FReI microscopy. Although we provide a single practical example, we believe that the observations made about clustering methods generally apply to the clustering of ordered data found in microscopy videos and are broadly relevant to other areas of time-resolved fluorescence microscopy, such as fluorescence correlation spectroscopy (FCS), single-molecule localization microscopy (SMLM), and fluorescence lifetime imaging microscopy (FLIM).

Cluster analysis has proven useful for IR and Raman imaging for classification,[12,59] and segmentation of cell or material types[60,61] for samples or tissues.[62] We have employed the workflows described in this paper to better understand the heterogeneity of *de novo* lipogenesis and to identify previously overlooked regions with higher concentrations of free fatty acids in Huh-7 cells using OPTIR microscopy.[61] Further applications of quantitative clustering and iterative image analysis lie in the *in vivo* study of the thermodynamic and kinetic properties of biomolecules in different organelles, cellular regions, and biocondensates. They have also shown promise in the investigation of organism behavior.[63] Microscopy videos contain a wealth of information that often goes unused. We believe that this work can facilitate the application of these methods to other time series-based methods, furthering our understanding of complex biological processes.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.4c02859.

> Workflow of L-HAC method; clustering as a function of noise; two group clustering from Figure 1A; accuracy analysis; single function performance from cylinder, funnel, bell data set; time scaling with samples and frames; clustering of FReI data with different similarity metrics; table of parameters for simulations used in Figure 1; and formulation and time complexity for similarity metrics (PDF)

AUTHOR INFORMATION

Corresponding Authors

Victor S. Batista − *Department of Chemistry, Yale University, New Haven, Connecticut 06520, United States;* orcid.org/0000-0002-3262-1237; Email: Victor.Batista@yale.edu

Caitlin M. Davis − *Department of Chemistry, Yale University, New Haven, Connecticut 06520, United States;* orcid.org/0000-0003-4340-4577; Email: c.davis@yale.edu

Author

Michael J. Burke − *Department of Chemistry, Yale University, New Haven, Connecticut 06520, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.4c02859

Author Contributions

M.J.B. conceived the work, wrote and tested the code, analyzed the data, and wrote the manuscript. V.S.B. conceived the work and wrote the manuscript. C.M.D. conceived the work, acquired the FReI data, and wrote the manuscript.

Notes

The authors declare no competing financial interest.

https://doi.org/10.1021/acs.jpcb.4c02859
J. Phys. Chem. B 2024, 128, 8344−8354

■ ACKNOWLEDGMENTS

M.J.B. was partially supported by the Peter B. Moore Yale University Fellowship. V.S.B. acknowledges support from the NIH grant R01 GM136815 and a generous allocation of high-performance computing time from NERSC. C.M.D. acknowledges support from the NIH grant R35 GM151146.

■ REFERENCES

(1) Shou, J.; Oda, R.; Hu, F.; Karasawa, K.; Nuriya, M.; Yasui, M.; Shiramizu, B.; Min, W.; Ozeki, Y. Super-Multiplex Imaging of Cellular Dynamics and Heterogeneity by Integrated Stimulated Raman and Fluorescence Microscopy. *iScience* **2021**, *24* (8), No. 102832.

(2) Huang, L.; Wong, C.; Grumstrup, E. Time-Resolved Microscopy: A New Frontier in Physical Chemistry. *J. Phys. Chem. A* **2020**, *124* (29), 5997−5998.

(3) Schueder, F.; Rivera-Molina, F.; Su, M.; Marin, Z.; Kidd, P.; Rothman, J. E.; Toomre, D.; Bewersdorf, J. Unraveling Cellular Complexity with Transient Adapters in Highly Multiplexed Super-Resolution Imaging. *Cell* **2024**, *187* (7), 1769−1784.e18.

(4) Gross, N.; Kuhs, C. T.; Ostovar, B.; Chiang, W.-Y.; Wilson, K. S.; Volek, T. S.; Faitz, Z. M.; Carlin, C. C.; Dionne, J. A.; Zanni, M. T.; et al. Progress and Prospects in Optical Ultrafast Microscopy in the Visible Spectral Region: Transient Absorption and Two-Dimensional Microscopy. *J. Phys. Chem. C* **2023**, *127* (30), 14557−14586.

(5) Yu, L.; Lei, Y.; Ma, Y.; Liu, M.; Zheng, J.; Dan, D.; Gao, P. A Comprehensive Review of Fluorescence Correlation Spectroscopy. *Front. Phys.* **2021**, *9*, No. 644450.

(6) Prater, C.; Bai, Y.; Konings, S. C.; Martinsson, I.; Swaminathan, V. S.; Nordenfelt, P.; Gouras, G.; Borondics, F.; Klementieva, O. Fluorescently Guided Optical Photothermal Infrared Microspectroscopy for Protein-Specific Bioimaging at Subcellular Level. *J. Med. Chem.* **2023**, *66* (4), 2542−2549.

(7) Nunez-Iglesias, J.; Kennedy, R.; Parag, T.; Shi, J.; Chklovskii, D. B. Machine Learning of Hierarchical Clustering to Segment 2D and 3D Images. *PLoS One* **2013**, *8* (8), No. e71715.

(8) Belevich, I.; Joensuu, M.; Kumar, D.; Vihinen, H.; Jokitalo, E. Microscopy Image Browser: A Platform for Segmentation and Analysis of Multidimensional Datasets. *PLoS Biol.* **2016**, *14* (1), No. e1002340.

(9) Huth, J.; Buchholz, M.; Kraus, J. M.; Schmucker, M.; von Wichert, G.; Krndija, D.; Seufferlein, T.; Gress, T. M.; Kestler, H. A. Significantly Improved Precision of Cell Migration Analysis in Time-Lapse Video Microscopy through Use of a Fully Automated Tracking System. *BMC Cell Biol.* **2010**, *11* (1), 24.

(10) Popescu, G.; Park, Y.; Lue, N.; Best-Popescu, C.; Deflores, L.; Dasari, R. R.; Feld, M. S.; Badizadegan, K. Optical Imaging of Cell Mass and Growth Dynamics. *Am. J. Physiol.: Cell Physiol.* **2008**, *295* (2), C538−C544.

(11) Tinevez, J.-Y.; Perry, N.; Schindelin, J.; Hoopes, G. M.; Reynolds, G. D.; Laplantine, E.; Bednarek, S. Y.; Shorte, S. L.; Eliceiri, K. W. TrackMate: An Open and Extensible Platform for Single-Particle Tracking. *Methods* **2017**, *115*, 80−90.

(12) Byrne, H. J.; Knief, P.; Keating, M. E.; Bonnier, F. Spectral Pre and Post Processing for Infrared and Raman Spectroscopy of Biological Tissues and Cells. *Chem. Soc. Rev.* **2016**, *45* (7), 1865−1878.

(13) Varley, T. F.; Sporns, O. Network Analysis of Time Series: Novel Approaches to Network Neuroscience. *Front. Neurosci.* **2022**, *15*, No. 787068.

(14) Crase, S.; Hall, B.; N Thennadil, S. Cluster Analysis for IR and NIR Spectroscopy: Current Practices to Future Perspectives. *Comput. Mater. Contin.* **2021**, *69* (2), 1945−1965.

(15) Aghabozorgi, S.; Seyed Shirkhorshidi, A.; Ying Wah, T. Time-Series Clustering—A Decade Review. *Inf. Syst.* **2015**, *53*, 16−38.

(16) Camps-Valls, G.; Bruzzone, L. Kernel-Based Methods for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43* (6), 1351−1362.

(17) von Chamier, L.; Laine, R. F.; Jukkala, J.; Spahn, C.; Krentzel, D.; Nehme, E.; Lerche, M.; Hernández-Pérez, S.; Mattila, P. K.; Karinou, E.; et al. Democratising Deep Learning for Microscopy with ZeroCostDL4Mic. *Nat. Commun.* **2021**, *12* (1), No. 2276.

(18) Xing, F.; Xie, Y.; Su, H.; Liu, F.; Yang, L. Deep Learning in Microscopy Image Analysis: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29* (10), 4550−4568.

(19) Andronov, L.; Michalon, J.; Ouararhni, K.; Orlov, I.; Hamiche, A.; Vonesch, J.-L.; Klaholz, B. P. 3DClusterViSu: 3D Clustering Analysis of Super-Resolution Microscopy Data by 3D Voronoi Tessellations. *Bioinformatics* **2018**, *34* (17), 3004−3012.

(20) Chang, Y.-H.; Yokota, H.; Abe, K.; Tang, C.-T.; Tasi, M.-D. Automated Detection and Tracking of Cell Clusters in Time-Lapse Fluorescence Microscopy Images. *J. Med. Biol. Eng.* **2017**, *37* (1), 18−25.

(21) Cribben, I.; Yu, Y. Estimating Whole-Brain Dynamics by Using Spectral Clustering. *J. R. Stat. Soc., C: Appl. Stat.* **2017**, *66* (3), 607−627.

(22) Kobrina, Y.; Rieppo, L.; Saarakkala, S.; Jurvelin, J. S.; Isaksson, H. Clustering of Infrared Spectra Reveals Histological Zones in Intact Articular Cartilage. *Osteoarthritis Cartilage* **2012**, *20* (5), 460−468.

(23) Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J. Chem. Theory Comput.* **2007**, *3* (6), 2312−2334.

(24) Kostelich, E. J.; Schreiber, T. Noise Reduction in Chaotic Time-Series Data: A Survey of Common Methods. *Phys. Rev. E* **1993**, *48* (3), 1752−1763.

(25) Vanrullen, R. Four Common Conceptual Fallacies in Mapping the Time Course of Recognition. *Front. Psychol.* **2011**, *2*, 365.

(26) Gao, J.; Sultan, H.; Hu, J.; Tung, W.-W. Denoising Nonlinear Time Series by Adaptive Filtering and Wavelet Shrinkage: A Comparison. *IEEE Signal Process. Lett.* **2010**, *17* (3), 237−240.

(27) Keogh, E.; Chakrabarti, K.; Pazzani, M.; Mehrotra, S. Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. *Knowl. Inf. Syst.* **2001**, *3* (3), 263−286.

(28) Belkin, M.; Niyogi, P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Comput.* **2003**, *15* (6), 1373−1396.

(29) Johnson, S. C. Hierarchical Clustering Schemes. *Psychometrika* **1967**, *32* (3), 241−254.

(30) Baek, M.; Kim, C. A Review on Spectral Clustering and Stochastic Block Models. *J. Korean Stat. Soc.* **2021**, *50* (3), 818−831.

(31) Nascimento, M. C. V.; de Carvalho, A. C. P. L. F. Spectral Methods for Graph Clustering—A Survey. *Eur. J. Oper. Res.* **2011**, *211* (2), 221−231.

(32) Jolliffe, I. T. *Principal Component Analysis*; Springer Series in Statistics; Springer: New York, NY, 1986.

(33) Jesse, S.; Kalinin, S. V. Principal Component and Spatial Correlation Analysis of Spectroscopic-Imaging Data in Scanning Probe Microscopy. *Nanotechnology* **2009**, *20* (8), No. 085714.

(34) de Carvalho, C.; Pons, M.; Manuela R da Fonseca, M. Principal Components Analysis as a Tool to Summarise Biotransformation Data: Influence on Cells of Solvent Type and Phase Ratio. *Biocatal. Biotransform.* **2003**, *21* (6), 305−314.

(35) Kutz, J. N.; Fu, X.; Brunton, S. L. Multiresolution Dynamic Mode Decomposition. *SIAM J. Appl. Dyn. Syst.* **2016**, *15* (2), 713−735.

(36) Newman, M. E. J. Modularity and Community Structure in Networks. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (23), 8577−8582.

(37) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17* (3), 261−272.

(38) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (85), 2825−2830.

(39) Black, P. Dictionary of Algorithms and Data Structures, 2024. https://www.nist.gov/dads/.

(40) Sakoe, H.; Chiba, S. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Trans. Acoust., Speech, Signal Process.* **1978**, *26* (1), 43−49.

(41) Lee Rodgers, J.; Nicewander, W. A. Thirteen Ways to Look at the Correlation Coefficient. *Am. Stat.* **1988**, *42* (1), 59−66.
(42) Spearman, C. Demonstration of Formulæ for True Measurement of Correlation. *Am. J. Psychol.* **1907**, *18* (2), 161−169.
(43) Gideon, R. A.; Hollister, R. A. A Rank Correlation Coefficient Resistant to Outliers. *J. Am. Stat. Assoc.* **1987**, *82* (398), 656−666.
(44) Kendall, M. G. A New Measure of Rank Correlation. *Biometrika* **1938**, *30* (1.2), 81−93.
(45) Maharaj, E. A.; Pierpaolo, D. U.; Caiado, J. *Time Series Clustering and Classification*; CRC Press, 2019.
(46) Saito, N. Local Feature Extraction and Its Applications Using a Library of Bases, Yale University, 1994. https://www.math.ucdavis.edu/~saito/publications/saito_phd.pdf.
(47) Keogh, E.; Ratanamahatana, C. A. Exact Indexing of Dynamic Time Warping. In *Knowledge and Information Systems*; Springer, 2005; Vol. 7, pp 358−386.
(48) Berndt, D. J.; Clifford, J. In *Using Dynamic Time Warping to Find Patterns in Time Series*, Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, AAAI Technical Report, 1994; p 12.
(49) Davis, C. M.; Gruebele, M. Cellular Sticking Can Strongly Reduce Complex Binding by Speeding Dissociation. *J. Phys. Chem. B* **2021**, *125* (15), 3815−3823.
(50) Guo, M.; Xu, Y.; Gruebele, M. Temperature Dependence of Protein Folding Kinetics in Living Cells. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109* (44), 17863−17867.
(51) Kubelka, J. Time-Resolved Methods in Biophysics. 9. Laser Temperature-Jump Methods for Investigating Biomolecular Dynamics. *Photochem. Photobiol. Sci.* **2009**, *8* (4), 499−512.
(52) Handwerger, K. E.; Cordero, J. A.; Gall, J. G. Cajal Bodies, Nucleoli, and Speckles in the Xenopus Oocyte Nucleus Have a Low-Density, Sponge-like Structure. *Mol. Biol. Cell* **2005**, *16* (1), 202−211.
(53) Minton, A. P. Influence of Macromolecular Crowding upon the Stability and State of Association of Proteins: Predictions and Observations. *J. Pharm. Sci.* **2005**, *94* (8), 1668−1675.
(54) Cayley, S.; Lewis, B. A.; Guttman, H. J.; Record, M. T. Characterization of the Cytoplasm of *Escherichia Coli* K-12 as a Function of External Osmolarity: Implications for Protein-DNA Interactions in Vivo. *J. Mol. Biol.* **1991**, *222* (2), 281−300.
(55) Akabayov, B.; Akabayov, S. R.; Lee, S.-J.; Wagner, G.; Richardson, C. C. Impact of Macromolecular Crowding on DNA Replication. *Nat. Commun.* **2013**, *4* (1), No. 1615.
(56) Minton, A. P. How Can Biochemical Reactions within Cells Differ from Those in Test Tubes? *J. Cell Sci.* **2006**, *119* (14), 2863−2869.
(57) Phillip, Y.; Kiss, V.; Schreiber, G. Protein-Binding Dynamics Imaged in a Living Cell. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109* (5), 1461−1466.
(58) Wieczorek, G.; Zielenkiewicz, P. Influence of Macromolecular Crowding on Protein-Protein Association Rates—a Brownian Dynamics Study. *Biophys. J.* **2008**, *95* (11), 5030−5036.
(59) Ali, S. M.; Bonnier, F.; Lambkin, H.; Flynn, K.; McDonagh, V.; Healy, C.; Lee, T. C.; Lyng, F. M.; Byrne, H. J. A Comparison of Raman, FTIR and ATR-FTIR Micro Spectroscopy for Imaging Human Skin Tissue Sections. *Anal. Methods* **2013**, *5* (9), 2281−2291.
(60) Oust, A.; Møretrø, T.; Kirschner, C.; Narvhus, J. A.; Kohler, A. FT-IR Spectroscopy for Identification of Closely Related Lactobacilli. *J. Microbiol. Methods* **2004**, *59* (2), 149−162.
(61) Shuster, S. O.; Burke, M. J.; Davis, C. M. Spatiotemporal Heterogeneity of De Novo Lipogenesis in Fixed and Living Single Cells. *J. Phys. Chem. B* **2023**, *127* (13), 2918−2926.
(62) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559* (7715), 547−555.
(63) Girdhar, K.; Gruebele, M.; Chemla, Y. R. The Behavioral Space of Zebrafish Locomotion and Its Neural Network Analog. *PLoS One* **2015**, *10*, No. e0128668.

https://doi.org/10.1021/acs.jpcb.4c02859
J. Phys. Chem. B 2024, 128, 8344−8354