BMC Bioinformatics

RESEARCH

Open Access

Temporal classification of short time series data

Benedikt Venn[1], Thomas Leifeld[2], Ping Zhang[2] and Timo Mühlhaus[1*]

*Correspondence:
timo.muehlhaus@rptu.de

[1] Computational Systems Biology, RPTU Kaiserslautern, 67663 Kaiserslautern, Germany
[2] Institute of Automatic Control, RPTU Kaiserslautern, 67663 Kaiserslautern, Germany

Abstract

Motivation: Within the frame of their genetic capacity, organisms are able to modify their molecular state to cope with changing environmental conditions or induced genetic disposition. As high throughput methods are becoming increasingly affordable, time series analysis techniques are applied frequently to study the complex dynamic interplay between genes, proteins, and metabolites at the physiological and molecular level. Common analysis approaches fail to simultaneously include (i) information about the replicate variance and (ii) the limited number of responses/shapes that a biological system is typically able to take.

Results: We present a novel approach to model and classify short time series signals, conceptually based on a classical time series analysis, where the dependency of the consecutive time points is exploited. Constrained spline regression with automated model selection separates between noise and signal under the assumption that highly frequent changes are less likely to occur, simultaneously preserving information about the detected variance. This enables a more precise representation of the measured information and improves temporal classification in order to identify biologically interpretable correlations among the data.

Availability and implementation: An open source F# implementation of the presented method and documentation of its usage is freely available in the *TempClass* repository, https://github.com/CSBiology/TempClass [58].

Keywords: Time series analysis, Smoothing spline, Profile classification, Omics analyis

Introduction

Biological systems are constantly regulating their genes, proteins, and metabolites to maintain an optimal internal state. Optimal, however, is context-dependent and contingent upon prevailing environmental factors. Disturbances such as alterations in light, temperature, moisture, or mineral concentrations necessitate metabolic adjustments to mitigate potential stress and restore optimal conditions according to the external influence. These acclimation responses are meticulously orchestrated and follow a defined sequence. Their primary objectives are to mitigate the adverse effects of unfavourable environmental conditions or optimally exploit positive alterations.

With the declining costs associated with high throughput technologies like RNA-Seq and MS proteomics, the utilization of time series analyses has gained popularity as a valuable tool to study the kinetics/temporal dynamics of biological molecules. Nonetheless, challenges complicate comprehensive analyses of such data. Time series datasets often comprise a limited number of measurement points, and due to the substantial investment required in growing biological material and the still relatively high costs of these analyses, experiments are typically designed with a modest number of measurement points (typically 4 to 8) and a few replicates (2–4).

When characterizing the cellular response characteristics for individual molecules, it is imperative to assign lower significance to measurement points characterized by elevated uncertainty [1]. While this assignment is often intuitive when performed manually, an automated evaluation method necessitates the explicit incorporation of this consideration.

Simultaneously, the biological response capacity is constrained. High amplitude fluctuations are improbable from a regulatory standpoint, and they would entail substantial synthesis and degradation costs for the biological system in question. In the absence of new stimuli, one can reasonably anticipate smooth kinetics in biological molecules, especially for more complex molecules like proteins. This assumption provides valuable additional information that enhances the precision and utility of biological models. However, it mandates the development of novel analytical techniques to effectively incorporate such information.

Our proposed approach addresses the dual challenges of variable measurement uncertainties and the expectation of low signal fluctuation. We have employed smoothing splines, which impose continuity in function, slope, and curvature, while also permitting the weighting of individual measurement points and the imposition of shape constraints. Unlike existing methods that use predefined profiles [2], or require a preselected number of clusters [3], our approach classifies the data where a single classification is uncoupled from the remaining data.

Methods

Time series data can be assumed as functions of time with superimposed heteroscedastic biological variance and technical noise [4].

$$y_i = f(t_i) + \varepsilon_i, i = 1, \ldots, n$$

where f(·) resembles the function of the true abundance time course at the ith of n time points. The error term ε_i combines biological variation and noise introduced by sample processing and measurement devices, leading to the blurring of the true relationship f(·) to the final reading y at time point i. The interval widths between measuring time points are defined by:

$$h_i = y_{i+1} - y_i, i = 1, \ldots, n-1$$

While in many analysis strategies, e.g. common clustering procedures or statistical testing frameworks, interval widths are not taken into account, they may possess valuable information regarding the dynamic of the underlying kinetic [5, 6]. High amplitude changes within a short time period can be considered unlikely and thus penalized by the

model. However, for biological regulatory responses, the model time point spacing may vary from the actual experimental time point spacing as discussed below.

Time point spacing

As indicated by its name, the independent variable is time (e.g. hours since experiment start). For experiments with no perturbation, these time intervals can be directly used for curve fitting. Cell cycle regulation may be measured with fixed time intervals because the expected rate of change is evenly distributed between the time points. If the biological system, however, faces sudden condition perturbation, the spacing according to time intervals is insufficient for modelling. A perturbation causes the biological system to react immediately. This regulation of molecular processes has to occur quickly with regulation in later time points being less fluctuating. To account for this asymmetric regulation, samples are taken according to the expected rate of system response. The measurements of the presented experiment were taken by doubling the time interval at each measurement. This is according to the estimated change apparent within two measurements. Hence, samples can be spaced uniformly in time. The presented approach is not restricted to uniformly spaced time points and works with any univariate time series.

Smoothing spline

To investigate the underlying kinetic, the application of smoothing splines offers a valuable approach to model the data, striking a balance between data fidelity and the smoothness of the fit. While various other fitting techniques exist, i.e. interpolation strategies or (non-) linear regression, most of them rely on a predefined template function or ignore point uncertainty as they are forced to interpolate the sample means. Further explanations for choosing smoothing splines over other fitting techniques are given in the discussion. In this method, piecewise cubic polynomials are employed to model each subinterval of the data while smooth transitions are ensured by enforcing the equality of function values, slopes, and curvatures at each designated knot. Considering that knots are positioned at each time point, there are $n-1$ intervals to analyze.

Splines within the interval $[t_i, t_{i+1}]$ are defined as

$$f_i(t) = a_i\phi_{0i}t + a_{i+1}\phi_{1i}t + c_i\gamma_{0i}t + c_{i+1}\gamma_{1i}t \tag{1}$$

with $a_i = f_i(t_i)$, $c_i = f''_i(t_i)$, and basis functions ϕ_0, ϕ_1, γ_0, and γ_1 defined in [7] and listed in the supplement. The vector a is the vector that only contains the function values at the measured time points, c contains the function curvature (second derivative) at the measured time points and the basis functions describe how the adjacent knots influence the curve shape between them.

The estimation of spline segments is subject to the minimization of the following cost function [7, 8]:

$$\int_{t_1}^{t_n} \left[f''(t)\right]^2 dt + \frac{\lambda}{n}\|W(a-y)\|^2, W_{i,i} = w_i \tag{2}$$

W is a diagonal matrix of observation weights introduced in Eq. 10, a is a rowvector of the splines function values at the knots, y is a rowvector of the observation values, and

$||\cdot||$ denotes the Euclidean norm. While the first term serves as a roughness penalty, the second term ensures the required fidelity to the data [9]. A smoothing factor λ mediates between these opposing error terms. When $\lambda = 0$, the resulting spline results in a straight least squares regression line, while $\lambda \to \infty$ leads to an interpolating cubic spline.

In-depth spline theory is given in [8–12]. The minimization of Eq. 2 can be rewritten as a quadratic optimization problem [7].

$$\min_a \frac{1}{2} a^T G_\lambda a + c_\lambda^T a, \tag{3}$$

with $G_\lambda = 2\left(H^T D^{-1} H + \frac{\lambda}{n} W^T W\right)$ and $c_\lambda = -2\frac{\lambda}{n} y^T W^T W$. Band matrices D and H are defined as:

$$D_{n-2\times n-2} = \begin{bmatrix} d_1^a & d_2^b & 0 & 0 & \cdots & 0 & 0 & 0 \\ d_2^b & d_2^a & d_3^b & 0 & \cdots & 0 & 0 & 0 \\ 0 & d_3^b & d_3^a & d_4^b & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & d_{n-3}^b & d_{n-3}^a & d_{n-2}^b \\ 0 & 0 & 0 & 0 & \cdots & 0 & d_{n-2}^b & d_{n-2}^a \end{bmatrix} \tag{4}$$

$$d_i^a = (h_i + h_{i+1})/3 \tag{5}$$

$$d_i^b = h_i/6 \tag{6}$$

$$H_{n-2\times n} = \begin{bmatrix} e_1^b & e_1^a & e_2^b & 0 & \ldots & 0 & 0 & 0 \\ 0 & e_2^b & e_2^a & e_3^b & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & e_{n-2}^b & e_{n-2}^a & e_{n-1}^b \end{bmatrix} \tag{7}$$

$$e_i^a = -(h_i + h_{i+1})/(h_i \cdot h_{i+1}) \tag{8}$$

$$e_i^b = 1/h_i \tag{9}$$

Measurement weighting

A crucial part of spline smoothing is the determination of the weighting matrix W. For each signal, the time point weighting w_i relies on the signal's standard deviation that is divided by the average standard deviation of all time points. As smoothing splines—under the given smoothness constraints—aim to minimize the distance of the original data points to the resulting prediction (sum of squares), outlier values would negatively impact the prediction function. As time points with outliers often are affected by high uncertainty, this variance can be exploited to reduce the outlier impact. When calculating the sum of squares, high variances are encoded as low weights that reduce the impact of points, which should have reduced influence on the fit (Eq. 2).

$$W = \begin{bmatrix} w_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & w_i & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & w_n \end{bmatrix}, w_i = \frac{w_{init}(i)}{\overline{w_{init}}} \tag{10}$$

$$w_{init}(i) = \sigma_{y_i} \tag{11}$$

Shape constraints

To control the spline's shape, monotonicity can be enforced for each interval separately. To obtain a single maximum within interval h_i, combinations of monotonicity constraints are applied, so that intervals within $[t_0, t_{i-1}]$ are monotonically increasing, and intervals in $[t_{i+1}, t_n]$ are monotonically decreasing. Monotonicity constraints are well studied and can be enforced by the following conditions [13–16].

For monotonically increasing polynomials in $[t_i, t_{i+1}]$ ensure that

$$a_{i+1} - a_i \geq 0 \tag{12}$$

$$f'_i \geq 0 \tag{13}$$

$$f'_{i+1} \geq 0 \tag{14}$$

$$3\frac{a_{i+1} - a_i}{h_i} - f'_i \geq 0 \tag{15}$$

$$3\frac{a_{i+1} - a_i}{h_i} - f'_i \geq 0 \tag{16}$$

For monotonically decreasing polynomials signs are switched from $\geq$ to $\leq$. The derivative f'_i is calculated from $B_i a$ where B is a $n \times n$ Matrix defined by $B = P^{-1}U$. P and U are described in [7] and can be seen in the supplement. The conditions for every interval are summarized in a constraint matrix $Ca \geq [0]$ representing linear inequality constraints. When solving (12–16) with respect to $Ca \geq [0]$ some constraints will be satisfied as equality constraints. This constraint set where $C_A a = [0]$ is termed 'Active set'. Matrix Z which columns form a basis for the null space of C_A is used for determining the smoothing factor λ [7, 17]. Using smoothing splines in combination with the described monotonicity constraints allows the construction of smooth curves with oscillations allowed in specified regions that are not constrained to be monotone.

For every signal to fit, several monotonicity constraints can be applied, resulting in a range of potential parent shapes. These parent shapes span a spectrum from monotonically in- or decreasing curves to those featuring 1, 2, 3, or 4 extrema, with each type starting either with a maximum or a minimum. This results in a total of 10 distinct parent shapes, along with a single unconstrained scenario, that can be applied to a single time series signal.

Under the specified slope constraints, the case of monotonically increasing or decreasing curves offers only one shape, while curves containing at least one extremum necessitate the consideration of multiple shape possibilities (as summarized in Additional file 1: Table S1). Consequently, it is necessary to fit curves corresponding to all conceivable shapes to each signal. Subsequently, the most appropriate shape is selected as the descriptor for elucidating the underlying molecule kinetics.

Model selection

Besides the data points with an associated weighting matrix, smoothing splines rely on a smoothness parameter λ that controls how curved the resulting curve is going to be. The smoothing strength must be determined individually for each signal. The ideal λ value is estimated by minimizing the modified generalized cross validation (mGCV), which is an estimate for the model prediction error and—in contrast to other cross validation techniques—only requires a single passage. A minimal mGCV hints at the optimal compromise between over- and underfitting respectively [12, 18, 19].

$$\min_{\lambda} n \frac{\|W(a-y)\|^2}{(Tr(I-\rho A_\lambda))^2} \tag{17}$$

Matrix I is a $n \times n$ identity matrix and $A_\lambda = 2\frac{\lambda}{n} Z(Z^T G_\lambda Z)^{-1} Z^T W^T W$ an influence matrix, so that $a = Ay$ and $\rho = 1.3$ as compensatory factor for small sample sizes [7, 8, 20]. Because the minimization of Eq. 17 with respect to Eq. 3 leads to a non-convex and non-continuous optimization problem, a grid search approach is used to choose from a wide range of λ.

For every parent shape, a single candidate is reported and selected by minimizing mGCV. When a maximum of four extrema is allowed, this results in 11 final fits to choose from ((i) 5 fits with 0–4 extrema starting with a positive slope, (ii) 5 fits with 0–4 extrema starting with a negative slope, and (iii) an unconstrained spline that has no assumptions regarding its monotonicity). While mGCV proves suitable for making reasonable determinations of the smoothing strength within the parent shape class, it encounters challenges when confronted with the task of selecting the correct shape from the remaining 11 shape options. As mGCV does not include information about the number of allowed extrema it tends to favour fits with increased flexibility over conservative ones.

An adapted version of the Akaike information criterion (AIC) is used to select the final shape [21]. This modified approach incorporates a correction factor tailored for cases with limited sample sizes (AICc) as originally proposed in [22] and a term that incorporates the number of extrema present in the curve.

$$AIC_c = n \cdot \ln\left(\frac{\|W(a-y)\|^2}{n}\right) + 2k + \frac{2k^2+2k}{n-k-1} \tag{18}$$

Here k equals the enforced number of extrema.

Among all shape assumptions, the model that minimizes AIC_c (Eq. 18) is assumed to represent the underlying function at best and is used for temporal classification.

Extrema extraction and classification

It is trivial to identify the splines extrema, since all polynomial coefficients can easily be obtained for every interval from function values a and their second derivatives c at adjacent knots. Polynomial template:

$$s(t) = k_1 t^3 + k_2 t^2 + k_3 t + k_4, \tag{19}$$

with $s(t_i) = a_i$, $s(t_{i+1}) = a_{i+1}$, $s''(t_i) = c_i$, and $s''(t_{i+1}) = c_{i+1}$
Basic calculus leads to the polynomial coefficients:

$$k_1 = -\frac{1}{6}(c_i + c_{i+1}) \tag{20}$$

$$k_2 = \frac{1}{2}c_i \tag{21}$$

$$k_3 = -\frac{1}{3}c_i - \frac{1}{6}c_{i+1} - a_i + a_{i+1} \tag{22}$$

$$k_4 = a_i \tag{23}$$

Extreme points are determined by setting $s'(t) = 0$ and $s''(t) < 0$ for maxima and $s''(x) > 0$ for minima within $[t_i, t_{i+1}]$. Additional file 1: Fig. S6 gives a visual impression of the extrema extraction process. The location of extreme points is used to group similar shaped time series. Although the location of extreme values is determined by the model, it is necessary to determine the exact position after a model is fitted to the data. This downstream identification of extrema enables previous filtering of quasi-constant signals and ensures that extrema can be assigned to their nearest knots. The classifier may result in `Min3,Max4` indicating the spline having a minimum at the third time point and a maximum at the fourth time point. If there is no extremum, the classifier results in either `I` or `D`, depending on whether the spline is monotonically in- or decreasing. If necessary, these two monotone classes may be further refined using the second derivative to identify prominent changes in curvature (plateau regions).

Iterative clustering

As a common time-series clustering procedure, k-means clustering is used and compared to the presented classification method. The k-means algorithm iteratively recalculates the position of k initial centroids. All points are assigned to the nearest centroid, while the squared Euclidean distance is used as distance measurement. Convergence is reached when no reallocation of points to different centroids occurs [23, 24].

The optimal cluster number k is determined by the gap statistics method [25].

If relative changes of two signal slopes behave the same but their absolute values differ, most distance measures would report high distances even if their parallel change would suggest a difference of 0 (or similarity of 1). To prevent dominance of variables with differing ranges a normalization step often is needed to equalize all amplitudes. A popular

normalization is the z-score, which transforms a single time series to have zero mean and unit variance by

$$y'_k = \frac{y_k - \mu}{\sigma} \tag{24}$$

when μ denotes the arithmetic mean and σ is the standard deviation of the given data y_k [26]. This analysis was performed using FSharp.Stats v0.5.0 [27].

Comparison of leave one out cross validation

In order to assess the robustness of the used constrained smoothing spline, protein time series were fitted using four different curve fitting methods: (i) constrained smoothing spline as presented in this work, (ii) polynomial interpolation of sample averages, (iii) linear spline interpolation of sample averages, (iv) cubic spline interpolation of sample averages. All but the constrained smoothing spline procedures are available at FSharp.Stats v0.5.0. For leave one out cross correlation all three replicates of an internal sample were deleted from the time series.

After deletion all four fitting procedures were applied to the modified time series. The distance of the prediction at the time point of missing data to the original prediction is determined. For each protein signal, this leads to 6 distances. The same procedure was applied using the distance from the prediction at the time point of missing data to the sample mean (Additional file 1: Figure S1).

Visualization

All visualizations presented in this manuscript were prepared using Plotly.NET v4.0.0 [28].

Enrichment analysis

Gene set enrichment analysis is performed using extended MapMan annotations (for example "PS.lightreaction.LHC" becomes "PS.lightreaction.LHC", "PS.lightreaction" and "PS") [29–31]. All 1292 proteins served as background if overrepresentation of functional annotations is studied within classes. Enrichment was performed using multiple hypergeometric tests while p values were corrected for multiple testing using Storey's q value method [32, 33].

Results

Our approach assumes biological molecule kinetics have the intrinsic constraint to avoid curvature and therefore be smooth. To underpin this assumption with data, a time series dataset with 12 time points was analysed, which were analysed with 7 replicates each. Naturally, the more replicates measured, the more accurate the abundance estimator is supposed to be. To validate the smoothness assumption 2, 3, 5, and 7 replicates of each protein were randomly selected and analysed. The signals were interpolated with both linear splines and cubic splines, and the slope (linear spline) and curvature (cubic spline) at each knot were extracted. Both slopes and curvatures of the signals decrease with increasing number of replicates (Fig. 1). The higher the measurement accuracy (increased number of replicates), the higher

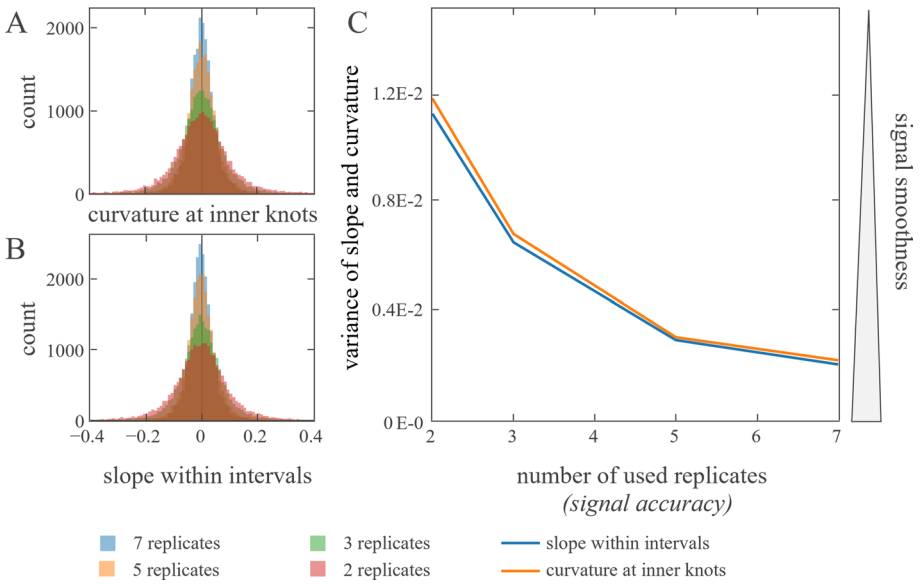

Fig. 1 In a circadian time series proteomics experiment, protein intensities were measured every four hours for two days (PXD019431). To investigate the smoothness of the signal, replicates at each time point were shuffled randomly and reduced to a replicate number given in the legend. In each data reduction iteration curvature and slopes were determined. **A**: Protein intensity means at 12 time points were interpolated with cubic splines using natural boundary conditions [27]. The second derivative of the spline at the inner knots was determined as smoothness measure. **B**: Protein intensity means at 12 time points were interpolated with linear splines. The first derivative of the lines within the knot intervals was determined as smoothness measure. **C**: Variances of curvatures **A** and slopes **B** of the signal interpolation are calculated and plotted against the number of used replicates

the signals smoothness (reduced variance in slope and curvature) supporting the assumption of biological molecule kinetics showing the tendency of being smooth.

Based on the constrained smoothing splines, we classified protein abundances from a heat acclimation experiment conducted in the green algae *Chlamydomonas reinhardtii* [34]. The utilized data subset consists of measurements taken at 8 time points during 40 °C heat treatment (0 h, 0.5 h, 1 h, 2 h, 4 h, 8 h, 16 h, and 24 h respectively). 1,292 proteins are measured in three biological replicates. An exemplary comparison of four fitting methods is given in Fig. 2. Constrained smoothing splines ensure smooth curves while being flexible enough to recognize relevant changes.

The signals were classified by the location of their extrema as described in *Extrema extraction and classification*. For 8 measured time points, 327 possible curve configuration possibilities (shapes) exist (Additional file 1: Table S1).

The analysis of the smoothed protein signals resulted in 77 classes to be filled with at least one protein signal. As expected, the classes become smaller with increasing specificity (Fig. 3). Classes with a high number of extreme points must have strong evidence of such behaviour in the form of low variances and high amplitude differences. The smoothness constraint therefore leads to less class occupancies the more complex the class gets, despite the fact, that the shape possibilities increase with more extrema (Fig. 3B).

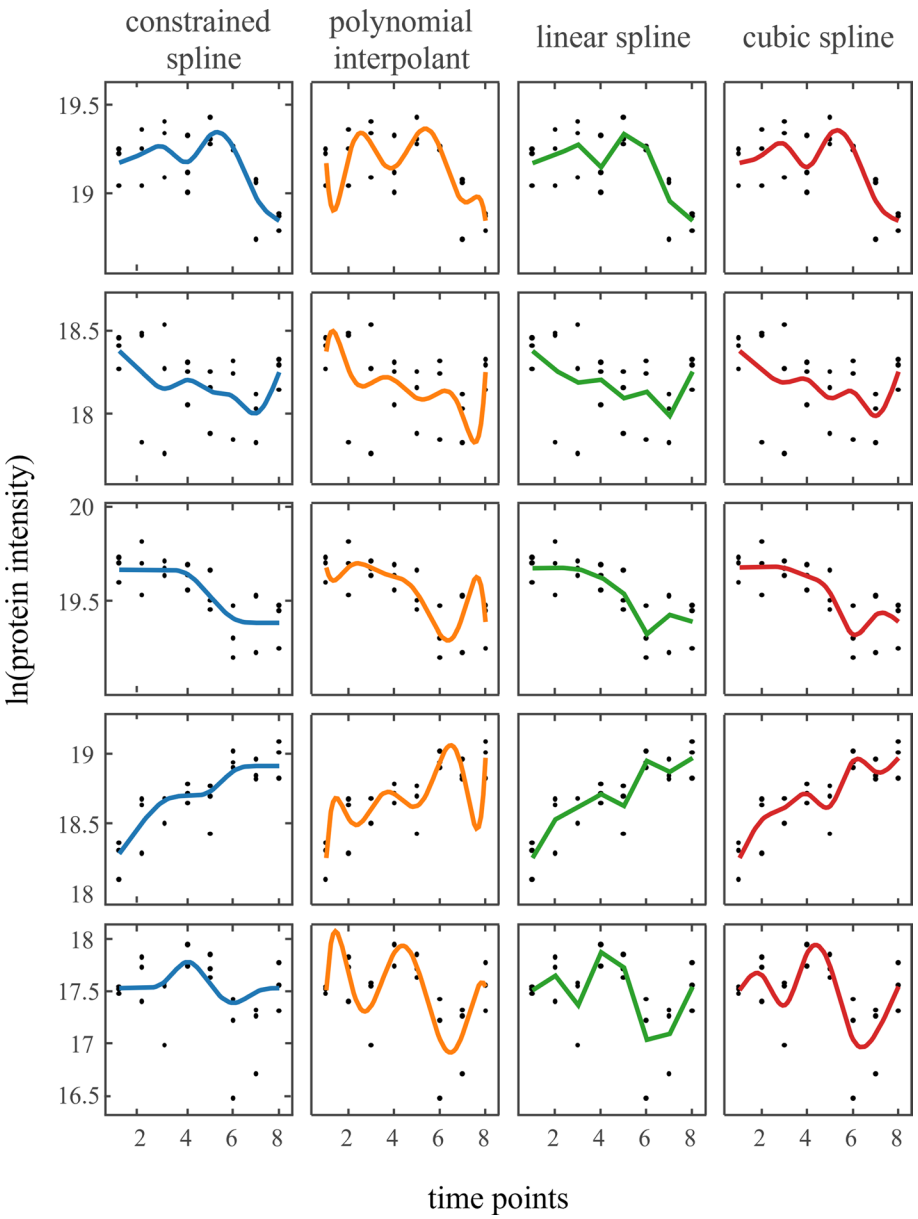

Fig. 2 Four curve fitting approaches on five proteins. Five exemplary protein abundance signals (top to bottom) modelled using four approaches: (i) constrained smoothing spline, (ii) polynomial interpolation of arithmetic means, (iii) linear spline interpolation of arithmetic means, and (iv) cubic interpolating splines with natural boundary conditions

Determination of fit robustness

To compare the use of constrained smoothing splines with other fitting methods, the data were cross-validated as explained in *Comparison of leave one out cross validation*. All replicates of an inner time point were removed, and the remaining time points fitted with the presented constrained smoothing spline, interpolating polynomial, linear spline, and cubic spline. The distance of the predicted value at the time of the missing data to the original prediction serves as a robustness measure. If a data point is deleted from the time series, the model curve may change in shape.

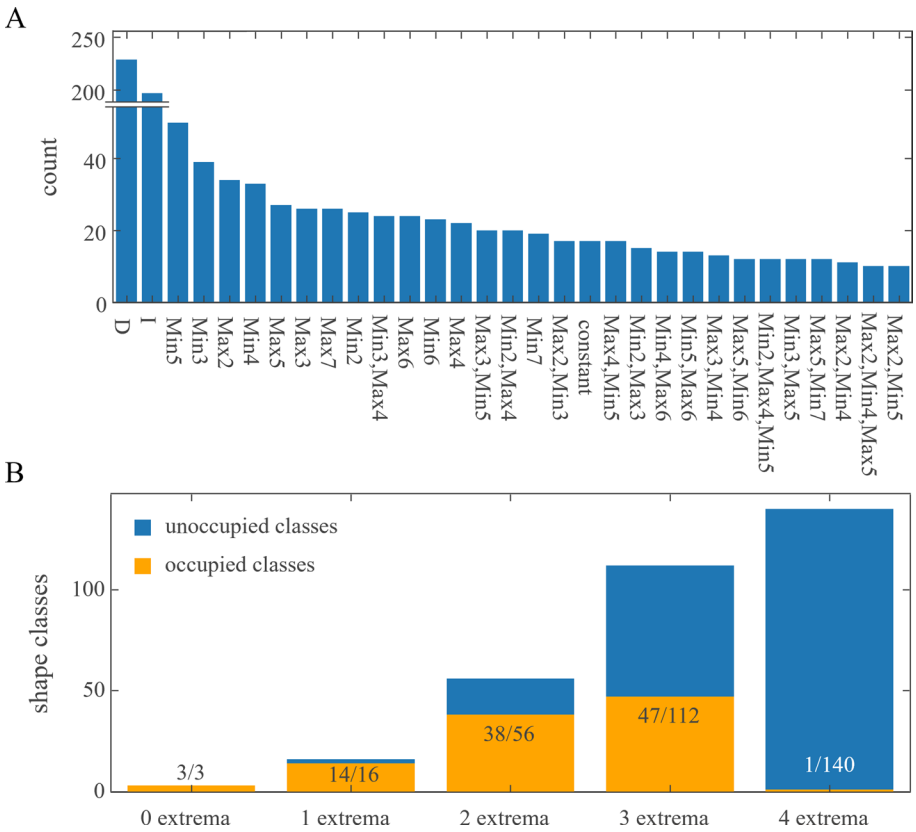

Fig. 3 Class occupancy **A** All classes that had more than 10 proteins assigned are depicted. Min3 indicates a single minimum at the third time point. Min3, Max4 indicates a minimum at the third time point followed by a maximum at the fourth time point. **B** Number of possible classes (blue) and number of occupied classes in the presented data set. No extremum is present in constant, monotonically in- and decreasing signals

The higher this change, the more prone to overfitting a model is. Simultaneously high distances indicate a high influence of the particular point for the model. The protein signals range from 16 to 24 with a median standard deviation of 0.156. As expected, polynomial interpolation leads to massive overfitting (compare curve shapes in Fig. 2). With increasing variance at the point of interest, the overfitting tendencies of linear and cubic splines tend to increase (Fig. 4B, C).

Especially when variance is high in the missing time point replicates, the constrained spline assigned lower weightings to this point and shows a reduced distance to the original curve (Fig. 4). To examine whether this robustness is solely due to a conservative fitting, the same procedure was performed using the distance of the prediction of the sparse data at the time point of missing replicates to the original sample mean instead of the original prediction. Low distances in this measure would hint at underfitting, indicating that the model isn't at all influenced by the signal manipulation. Because the polynomial as well as the linear and cubic spline interpolates the mean, the distances stay the same. The constrained spline shows similar distances as the other fitting techniques, indicating a comparable fidelity to the data and not giving suspicion for underfitting tendencies (Additional file 1: Fig. S2).

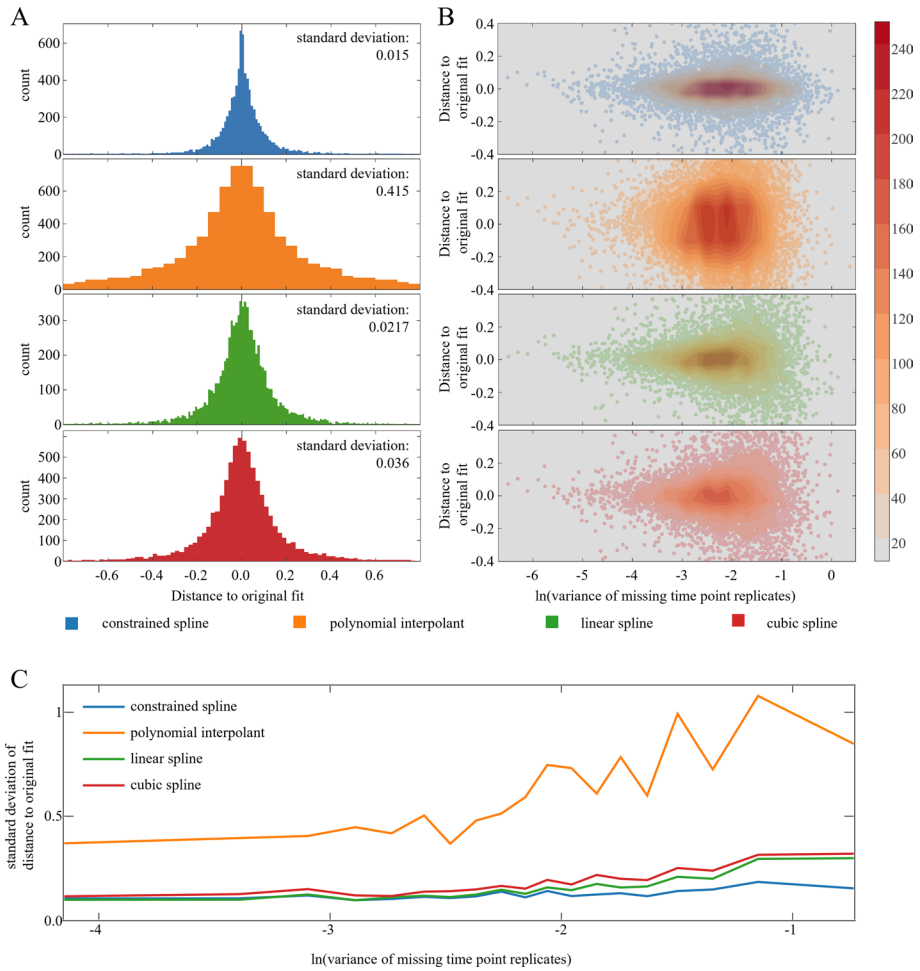

Fig. 4 Robustness analysis. **A**: Four fitting techniques were applied to each protein signal. After deleting every inner time point once, the distance from the original prediction to the prediction using the sparse signal is measured. For each protein, 6 distances are reported (number of inner time points) and summarized in a histogram. The histogram's standard deviation is given in the top right. **B**: The same data was used as in **A** but additionally separated by the variance of the time point replicates that were deleted. **C**: The distance data is separated in 20 equally large bins depending on the time point variance of the missing time point replicates (**B** x axis). The standard deviation of the distances within each bin is calculated and plotted against the average time point variance

Comparison to clustering approaches

Besides statistical methods, clustering approaches are the most common analyses of biological time series and thus are a genuine reference for our approach.

For this purpose, the signals should be transformed in advance so that they have zero mean and unit variance. However, signals whose abundance does not change would be strongly distorted by this transformation. In order to increase the clustering quality, such signals can be filtered out. A simple to use quality filtering approach often seen on biological time series data is the application of a one-way ANOVA. Its main purpose in clustering approaches is not the detection of significances, but to filter signals whose average did not change during the time course. As a common threshold a p value of 0.05 was chosen. Note that the presented method of temporal

classification does not require this step, since signal classification is not affected by the presence of other signals.

Clustering of the remaining 720 protein signals was performed using the k means algorithm with Euclidean distance [27]. The optimal number of clusters was determined to be 5 (Additional file 1: Figure S3). It is obvious that five clusters do not sufficiently represent all regulatory responses within a biological system. It is a suitable methodology for obtaining a global impression of the data set and a summary of the major protein kinetic groups. But for a detailed description of the response of fine-tuned biological processes, this approach is too rough. Interesting regulatory details are blurred by the sheer amount of data within a single cluster (Fig. 5, Additional file 1: Fig. S4). Regression analysis reveals that for every cluster 30% of the protein signals are not within the 95% prediction interval of the cluster mean. If the regulatory response of proteins is studied, in at least 30% of the cases a classification based on the cluster would be inaccurate at best.

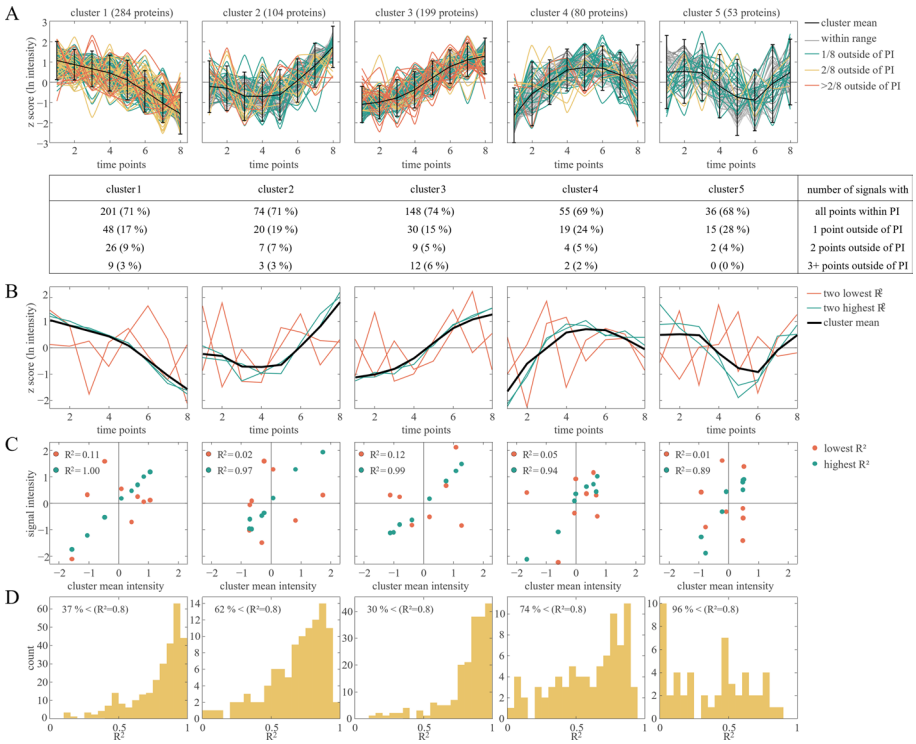

Fig. 5 Iterative clustering result. 720 protein signals were individually transformed to z scores and subsequently clustered by k means clustering ($k = 5$). **A**: Five clusters are depicted with cluster mean and 95% prediction interval (PI). Signal colours indicate whether 0 (grey), 1 (green), 2 (yellow), or more (orange) points of a signal lie outside of the PI. The table below shows the percentage of each group. **B**: The cluster mean intensity is visualized together with signals that showed the highest (green) and lowest (red) coefficient of determination (R^2) to the cluster mean. **C**: The cluster mean intensity is plotted against the intensities of the signal of highest (green) and lowest (red) R^2 within the cluster. R^2 of both signals is given in each panel. **D**: Histogram of all R^2 values between signals and cluster mean. The percentage in each panel depicts the percentage of signals whose R^2 is lower than 0.8

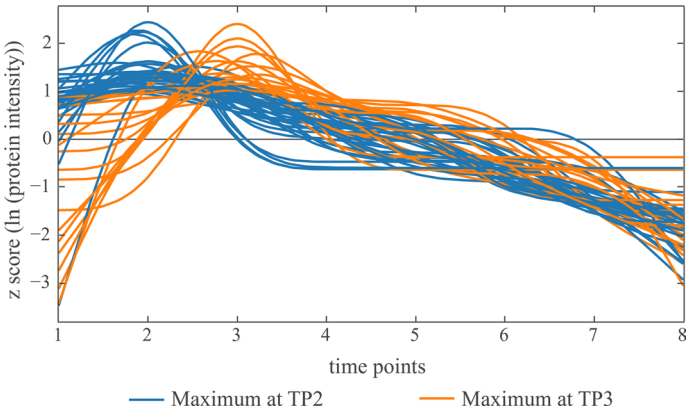

Fig. 6 Visualization of two classes that show a single maximum at time points 2 or 3 respectively. The intensity signals of 50 proteins are transformed to have zero mean and unit variance

Table 1 Enrichment result of early responder class

Functional term	q value	Trivial names (if annotated)
Protein.synthesis.ribosomal protein	0.0164	rps9; PRPL1; MRPL1; RPL23A; UBQ2; RPL40; UBQ1; RPS7; RPL18; RPL12; RPS27E1; RPL7; PRPL19;
Transport	0.0462	ATPVH;ATPVH;MPC1;AAA1;
Protein.synthesis.ribosomal protein. eukaryotic.60S subunit	0.0472	RPL23A; UBQ2; RPL40; UBQ1; RPL18; RPL12; RPL7
Protein.degradation.ubiquitin	0.0161	PKL1; UBC2; UBQ2; RPL40; UBQ1; UBQ2; RPL40; UBQ1; EIF3F; RPT4
RNA.RNA binding	0.0472	UBC2;REF1;HNR1
Transport.metabolite transporters at the mitochondrial membrane	0.0472	MPC1;AAA1
Polyamine metabolism.synthesis	0.0161	SPS1;SPD1

Functional annotations based on the MapMan ontology are listed together with the associated q value and trivial names if proteins had such

Biological interpretation

Besides using the smoothed signals for comparative analysis (e.g. co-expression networks), the smoothed and classified protein signals can be used for exploratory data analysis. To elucidate early, but short-term responders of the heat treatment, signals of the classes “Maximum at 2 or 3" can be isolated and used for categorizing the acclimation response.

Furthermore, global analysis strategies can be applied to classified signals. Gene set ontology enrichments of molecular functions can be applied to identify function overrepresentation.

Obviously, due to the sensitivity of the classification, the number of classes is by far greater than the number of clusters using common clustering strategies. This results in sparse occupancy of shape classes, which impedes enrichment strategies. However, the possibility of subsequent class aggregation presents valuable opportunities for analysing different combinations of response types. A gene set enrichment analysis was conducted on the early responder classes (Fig. 6) using MapMan functional annotations for *Chlamydomonas reinhardtii* genome version 5.5 [29, 30, 32]. Functional annotations that were overrepresented within the early responders can be seen in Table 1.

All functional annotations that are overrepresented in regulation shortly after heat onset were previously described to be involved in early heat acclimation regulation. (i) ribosomal proteins are required for the fast production of proteins; (ii) the transport group contains proteins predominantly involved meeting the increased demand of energy [35]; (iii) ubiquitin related proteins are necessary to both, degrade proteins that interfere with a heat acclimation, and remove proteins that aggregated due to the increased heat [36]; (iv) RNA binding proteins are involved in processing, stabilizing and exporting newly transcribed mRNA [37, 38]; (v) proteins of the polyamine synthesis group have been described to increase thermos-tolerance in algae [39].

The biological dissemination reveals that the classification approach is capable of elucidating the time-resolved orchestration of cellular responses and differentiating between different forms of regulation within a functional set of biological molecules.

Discussion

The era of high throughput technologies enabled researchers to analyse the abundance of thousands of molecules in a time-resolved manner. Scientists once needed days to take samples and measure the signals of a few proteins individually. Nowadays, it is possible to quantify the entire transcriptome or proteome in one fell swoop. Although it is possible to measure the kinetics of hundreds of proteins at a time, the number of robust strategies for an analysis of temporally resolved cell responses remains small [40–42].

Clustering methods have always been a popular tool to analyse time series experiments, as these are great options for unsupervised methods that work well with only a few assumptions to be made. For example, k-means clustering of time point averages represents the most commonly used algorithm for unsupervised analysis as its computation is efficient and the resulting clusters can easily be interpreted [43–45].

Although great findings have been made by this approach [46–48], it poses two problems when used for temporal characterization of regulation responses: (i) Most commonly used distance measures consider each coordinate separately. The time series vectors can be shuffled in pairs and still obtain the same distance. This behaviour is contrary to biological intuition because transcript or protein quantities are strongly dependent on the previous time points. This dependence is not taken into account in the model and inevitably leads to a decrease in the quality of the signal-to-noise separation [49, 50]. (ii) Although there are also distance measures that act in an environment-dependent manner (Dynamic Time Warping), in clustering methods it is necessary to specify the number of clusters in advance. This reduces accuracy, since small definable groups may be sorted into large groups, and their identification is thus only possible manually. Numerous ways of determining the optimal number of clusters have been developed (Elbow criterion, Silhouette index, Gap statistics) [25, 51], but in most cases these underestimate the number of biological response forms present. Furthermore, clustering approaches often are used to subsequently classify the data based on features that are visible when looking at whole clusters, but not necessarily are valid for individual cluster elements [52, 53]. As shown, this is prone to result in a huge number of misclassifications (Fig. 5). These and other similarity-based techniques are not able to dissect delicate regulation responses, but instead, these signals might be blurred by averaging effects.

Our approach models the kinetic response by constrained smoothing splines with the incorporation of measurement variance. Several other fitting techniques can be applied, each of which addresses different assumptions. Linear splines are the least complex fitting model for time series as they just connect the point estimates (median or average) at each time point. Point weighting is not possible and there is no separation of signal and noise. For the same reasons, other interpolating methods such as interpolating polynomials or cubic splines are not suited.

An exception is a polynomial-based function approximation with Chebyshev knots. A major problem with interpolating polynomials is Runge's Phenomenon. This is manifested by high frequency oscillation of the function in the outer knot intervals [54]. By clever rearrangement of the knots away from the curve centre and towards the peripheral areas such an oscillation can be prevented [55]. At the same time, the function no longer passes through the original data points. However, disadvantages here are both the non-obvious selection of knots and their weighting, as well as the lack of methods to comply with monotonicity constraints. Linear or nonlinear regression techniques seem inappropriate as they require either (i) the selection of a polynomial degree that does not represent any meaningful biological interpretation, or (ii) an already predefined function that is fitted to the signal.

The modelling of the time series by constrained smoothing splines is based on smoothness assumptions and the consideration of measurement point variances. Additionally, this regression approach preserves the existing dependence between neighbouring time points and therefore enforces monotonicity where excessive oscillation is unlikely. When compared to other interpolation methods (polynomials, linear splines, cubic splines) our approach showed high robustness while being flexible enough to capture characteristic events during the time course (Figs. 2, 4). Due to the overfitting tendency and presence of Runge's spike oscillation, polynomial interpolation is unsuited for classification analysis that handles extrema position as its primary characterization criterion. Linear splines show a high sensitivity for false declaration of extrema and perform poorly when it comes to predicting function values between the measured time points. Despite that cubic interpolating splines inherently aim to reduce heavy oscillations and perform great when it comes to predicting within intervals, their interpolating nature and inability to be weighted lead to little, but noticeable oscillations that interfere with an extrema-based classification strategy. While the number of shape classes can go to the hundreds, we could show that shapes with high flexibility and oscillations are found rarely (Fig. 3). This corresponds to the biological intuition of smooth protein regulation which was confirmed in Fig. 1.

Hybrid approaches are available that extend clustering approaches with prior smoothing regression [56], or by selecting meaningful expression profiles [1]. These approaches still rely on unsupervised approaches to group the data without predefining group labels. If faced with short time series data not exceeding 10 measurement time points, we propose that the number of group labels is manageable, hence all possible response shapes could be examined. However, a comparison of classification and clustering approaches remains difficult since the ground truth is unknown and both approaches address different questions. Most clustering approaches measure distances between signals, while classifications are concerned with the dissection

of signal features. The incorporation of the information that biological signals tend to be smooth and not oscillating leads to a feature extraction that corresponds to the intuition regarding the regulation of biological molecules.

With our temporal classification approach for studying time resolved regulation, it is possible to not only find an optimal fit of the data, but also assign shape classes to large time series data sets. This makes it possible to analyse the temporal orchestration of acclimation response and actively search for patterns of interest. We were able to show that our method provides a robust estimator when faced with sparse data. Furthermore, a well-studied process of heat acclimation of *Chlamydomonas reinhardtii* was presented as an example of the method enabling a detailed and supervised analysis of specific acclimation responses.

The smoothing and classification algorithms can be accessed as F# implementation at https://github.com/CSBiology/TempClass (Additional file 1: Figure S5) [58].

Limitations

This classification strategy is based on the selection of the optimal combination of monotone regions and extreme points. This approach requires an inner optimization to obtain the best fit of any enforced combination of monotonicity constraints, and an outer optimization to select the most ideal of the best shapes. With an increasing number of measured time points, there is a combinatorial explosion of the number of potential curve configurations. This not only increases the calculation time exponentially, but also the large number of resulting classes becomes unmanageable. Therefore, we have limited the number of allowed extreme points to 4 and recommend temporal classification for time series with 4 to 12 measurement time points.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05636-6.

Additional file 1: Spline basis functions, Table S1, and Figures S1 - S6.

Author contributions

BV and TM wrote the manuscript text. BV, TL, and TM implemented the method. All figures were prepared by BV. TM and PZ. supervised the work. All authors reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work was supported by Deutsche Forschungsgemeinschaft (TR175, project D02).

Availability of data and materials

The data set the method was applied on is available in the repository: ***Systems-wide investigation of responses to moderate and acute high temperatures in the green alga Chlamydomonas reinhardtii***, accessible at https://doi.org/10.60534/9e5jx-75d83 [57].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 26 October 2023 Accepted: 3 January 2024
Published online: 17 January 2024

References

1. Ernst J, Nau GJ, Bar-Joseph Z. Clustering short time series gene expression data. Bioinformatics. 2005;21(Suppl 1):i159–68. https://doi.org/10.1093/bioinformatics/bti1022.
2. Ernst J, Bar-Joseph Z. STEM: a tool for the analysis of short time series gene expression data. BMC Bioinformatics. 2006;7:191. https://doi.org/10.1186/1471-2105-7-191.
3. Lloyd S. Least squares quantization in PCM. IEEE Trans Inform Theory. 1982;28:129–37. https://doi.org/10.1109/TIT.1982.1056489.
4. Bathia N, Yao Q, Ziegelmann F. Identifying the finite dimensionality of curve time series. Ann Statist. 2010. https://doi.org/10.1214/10-AOS819.
5. Huang X, Ye Y, Xiong L, Lau RY, Jiang N, Wang S. Time series k-means: a new k-means type smooth subspace clustering for time series data. Inf Sci. 2016;367–368:1–13. https://doi.org/10.1016/j.ins.2016.05.040.
6. Warren LT. Clustering of time series data—a survey. Pattern Recogn. 2005;38:1857–74. https://doi.org/10.1016/j.patcog.2005.01.025.
7. Wood SN. Monotonic smoothing splines fitted by cross validation. SIAM J Sci Comput. 1994;15(5):1126–33. https://doi.org/10.1137/0915069.
8. Leifeld T, Venn B, Cui S, Zhang Z, Mühlhaus T, Zhang P. Curve form based quantization of short time series data. In: pp. 3710–3715. doi:https://doi.org/10.23919/ECC.2019.8795870.
9. de Boor C. A practical guide to splines. New York, N.Y.: Springer; 2001.
10. Lancaster P. Curve and surface fitting: an introduction. London: Academic Press; 1986.
11. Eubank RL. Nonparametric regression and spline smoothing. 2nd ed. Boca Raton: Chapman and Hall/CRC; 1999.
12. Fahrmeir L, Kneib T, Lang S. Regression: modelle, methoden und anwendungen. Berlin: Springer; 2007.
13. Fn F. Monotone piecewise cubic interpolation. SIAM J Numer Anal. 1980;17:238–46.
14. Ramsay JO. Monotone regression splines in action. Stat Sci. 1988;1:425–41.
15. Meyer MC. Constrained penalized splines. Can J Stat. 2012;40:190–206. https://doi.org/10.1002/cjs.10137.
16. Turlach BA. Constrained smoothing splines revisited. Statistics Research Report SRR 008-97. Center for Mathematics and Its Applications. Australian National University Canberra. 1997.
17. Anderson E, Bai Z, Bischof C, Blackford S, Demmel J, Dongarra J, et al. LAPACK users' guide. 3rd ed. Philadelphia: Society for Industrial and Applied Mathematics; 1999.
18. Craven P, Wahba G. Smoothing noisy data with spline functions. Numer Math. 1978;31:377–403. https://doi.org/10.1007/BF01404567.
19. Hutchinson MF, Gessler PE. Splines—more than just a smooth interpolator. Geoderma. 1994;62:45–67. https://doi.org/10.1016/0016-7061(94)90027-2.
20. Lukas MA. Robust generalized cross-validation for choosing the regularization parameter. Inverse Prob. 2006;22:1883–902. https://doi.org/10.1088/0266-5611/22/5/021.
21. Akaike H (1998) Information theory and an extension of the maximum likelihood principle. In: Parzen E, Tanabe K, Kitagawa G, editors. Selected papers of Hirotugu Akaike. Springer Series in Statistics. New York: Springer. https://doi.org/10.1007/978-1-4612-1694-0_15
22. Hurvich CM, Tsai C-L. Regression and time series model selection in small samples. Biometrika. 1989;76:297–307. https://doi.org/10.1093/biomet/76.2.297.
23. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1: Statistics: The Regents of the University of California; 1967.
24. Hartigan JA, Wong MA. A K-means clustering algorithm. Appl Stat. 1979;28:100. https://doi.org/10.2307/2346830.
25. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B Stat Methodol. 2001;63:411–23. https://doi.org/10.1111/1467-9868.00293.
26. Jain A, Nandakumar K, Ross A. Score normalization in multimodal biometric systems. Pattern Recogn. 2005;38:2270–85. https://doi.org/10.1016/j.patcog.2005.01.012.
27. Venn B, Mühlhaus T, Schneider K, Weil L, Zimmer D. fslaborg/FSharp.Stats: release 0.5.0: Zenodo; 2023.
28. Schneider K, Venn B, Mühlhaus T. Plotly. NET: a fully featured charting library for NET programming languages. F1000Res. 2022; 11: 1094. https://doi.org/10.12688/f1000research.123971.1.
29. Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, et al. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. 2004;37:914–39. https://doi.org/10.1111/j.1365-313x.2004.02016.x.
30. Usadel B, Poree F, Nagel A, Lohse M, Czedik-Eysenberg A, Stitt M. A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species. Maize Plant Cell Environ. 2009;32:1211–29. https://doi.org/10.1111/j.1365-3040.2009.01978.x.
31. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, et al. The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science. 2007;318:245–50. https://doi.org/10.1126/science.1143609.
32. Venn B. CSBiology/OntologyEnrichment: release 0.0.1: Zenodo; 2022.
33. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA. 2003;100:9440–5. https://doi.org/10.1073/pnas.1530509100.
34. Zhang N, Mattoon EM, McHargue W, Venn B, Zimmer D, Pecani K, et al. Systems-wide analysis revealed shared and unique responses to moderate and acute high temperatures in the green alga Chlamydomonas reinhardtii. Commun Biol. 2022;5:460. https://doi.org/10.1038/s42003-022-03359-z.

35. Vale RD. AAA proteins. Lords of the ring. J Cell Biol. 2000;150:F13–9. https://doi.org/10.1083/jcb.150.1.f13.
36. Galves M, Rathi R, Prag G, Ashkenazi A. Ubiquitin signaling and degradation of aggregate-prone proteins. Trends Biochem Sci. 2019;44:872–84. https://doi.org/10.1016/j.tibs.2019.04.007.
37. Pokora W, Tułodziecki S, Dettlaff-Pokora A, Aksmann A. Cross talk between hydrogen peroxide and nitric oxide in the unicellular green algae cell cycle: how does it work? Cells. 2022. https://doi.org/10.3390/cells11152425.
38. Pandey M, Stormo GD, Dutcher SK. Alternative splicing during the chlamydomonasreinhardtii cell cycle. G3 Bethesda. 2020;10:3797–810. https://doi.org/10.1534/g3.120.401622.
39. Liu S, Zhang J, Sun X, Xu N. Characterization of spermidine synthase (SPDS) gene and RNA—Seq based identification of spermidine (SPD) and spermine (SPM) involvement in improving high temperature stress tolerance in gracilariopsis lemaneiformis (Rhodophyta). Front Mar Sci. 2022. https://doi.org/10.3389/fmars.2022.939888.
40. Tripto NI, Kabir M, Bayzid MS, Rahman A. Evaluation of classification and forecasting methods on time series gene expression data. PLoS ONE. 2020;15: e0241686. https://doi.org/10.1371/journal.pone.0241686.
41. Androulakis IP, Yang E, Almon RR. Analysis of time-series gene expression data: methods, challenges, and opportunities. Annu Rev Biomed Eng. 2007;9:205–28. https://doi.org/10.1146/annurev.bioeng.9.060906.151904.
42. Wang X, Wu M, Li Z, Chan C. Short time-series microarray analysis: methods and challenges. BMC Syst Biol. 2008;2:58. https://doi.org/10.1186/1752-0509-2-58.
43. Jain AK, Dubes RC. Algorithms for clustering data; 1988.
44. Maigné É, Noirot C, Henry J, Adu Kesewaah Y, Badin L, Déjean S, et al. Asterics: a simple tool for the ExploRation and Integration of omiCS data. BMC Bioinformatics. 2023;24:391. https://doi.org/10.1186/s12859-023-05504-9.
45. Datta S, Datta S. Evaluation of clustering algorithms for gene expression data. BMC Bioinformatics. 2006;7(Suppl 4):S17. https://doi.org/10.1186/1471-2105-7-S4-S17.
46. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999;22:281–5. https://doi.org/10.1038/10343.
47. Saadeh H, Fayez RQA, Elshqeirat B. Application of K-means clustering to identify similar gene expression patterns during erythroid development. IJMLC. 2020;10:452–7. https://doi.org/10.18178/ijmlc.2020.10.3.956.
48. Nies H, Zakaria Z, Mohamad M, Chan W, Zaki N, Sinnott R, et al. A review of computational methods for clustering genes with similar biological functions. Processes. 2019;7:550. https://doi.org/10.3390/pr7090550.
49. Abanda A, Mori U, Lozano JA. A review on distance based time series classification. Data Min Knowl Disc. 2019;33:378–412. https://doi.org/10.1007/s10618-018-0596-4.
50. Bagnall A, Lines J, Bostrom A, Large J, Keogh E. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Min Knowl Discov. 2017;31:606–60. https://doi.org/10.1007/s10618-016-0483-9.
51. Kodinariya TM, Makwana PR. Review on determining number of cluster in K-means clustering. Int J. 2013;1(6):90–5.
52. Babichev S, Škvor J. Technique of gene expression profiles extraction based on the complex use of clustering and classification methods. Diagnostics. 2020. https://doi.org/10.3390/diagnostics10080584.
53. Datta S, Datta S. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. BMC Bioinformatics. 2006;7:397. https://doi.org/10.1186/1471-2105-7-397.
54. Boyd JP. Defeating the Runge phenomenon for equispaced polynomial interpolation via Tikhonov regularization. Appl Math Lett. 1992;5:57–9. https://doi.org/10.1016/0893-9659(92)90014-Z.
55. Trefethen LN. Approximation theory and approximation practice. Philadelphia: Society for Industrial and Applied Mathematics; 2013.
56. Déjean S, Martin PG, Baccini A, Besse P. Clustering time-series gene expression data using smoothing spline derivatives. EURASIP J Bioinform Syst Biol. 2007;2007:1. https://doi.org/10.1155/2007/70561.
57. Zhang N, Mattoon E, McHargue W, Venn B, Zimmer D, Pecani K, Jeong J, Anderson C, Chen C, Berry J, Xia M, Tzeng SC, Becker E, Pazouki L, Evans B, Cross F, Cheng J, Czymmek K, Schroda M, Mühlhaus T, Zhang R. Systems-wide investigation of responses to moderate and acute high temperatures in the green alga Chlamydomonas reinhardtii [Data set]. DataPLANT. 2023. https://doi.org/10.60534/9e5jx-75d83
58. Venn B, Mühlhaus T. CSBiology/TempClass: release 0.0.1: Zenodo; 2023.

Publisher's Note