SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
DYNAMICAL SYSTEMS
TIME SERIES

Received
30 September 2014

Accepted
6 February 2015

Published
11 March 2015

Correspondence and requests for materials should be addressed to Z.N. (nikoloski@mpimp-golm.mpg.de)

Segmentation of biological multivariate time-series data

Nooshin Omranian[1,2], Bernd Mueller-Roeber[1] & Zoran Nikoloski[2]

[1]Department of Molecular Biology, University of Potsdam, Karl-Liebknecht-Str. 24-25, Haus 20, 14476 Potsdam, Germany, [2]Systems Biology and Mathematical Modelling Group, Max Planck Institute for Molecular Plant Physiology, Am Muehlenberg 1, 14476 Potsdam, Germany.

Time-series data from multicomponent systems capture the dynamics of the ongoing processes and reflect the interactions between the components. The progression of processes in such systems usually involves check-points and events at which the relationships between the components are altered in response to stimuli. Detecting these events together with the implicated components can help understand the temporal aspects of complex biological systems. Here we propose a regularized regression-based approach for identifying breakpoints and corresponding segments from multivariate time-series data. In combination with techniques from clustering, the approach also allows estimating the significance of the determined breakpoints as well as the key components implicated in the emergence of the breakpoints. Comparative analysis with the existing alternatives demonstrates the power of the approach to identify biologically meaningful breakpoints in diverse time-resolved transcriptomics data sets from the yeast *Saccharomyces cerevisiae* and the diatom *Thalassiosira pseudonana*.

Time-series data gathered from biological and technological systems capture the underlying dynamics of the ongoing processes. For a single component of the system, the corresponding time-series can be partitioned into intervals with predominant trends, e.g., increasing or decreasing[1]. The identified breakpoints are usually associated with major events contributing to the behavior of the components[2]. However, determining consensus intervals over multiple observed components of a given system, referred to as *multivariate time-series segmentation* (MTS-seg), remains a challenging computational problem[3]. MTS-seg has wide applications in computational systems biology[4,5], market analysis[6], and process control[7].

In a multicomponent system, the progression of processes is a result of interactions among the components whose dependencies are reflected in the correlation structure of the respective data read-outs. For instance, in a biological system, the components include genes, proteins, and metabolites, whose changes can be monitored with high-throughput technologies[8]. Moreover, changes in the behavior of system's components may cause shifts in the correlation structure. MTS-seg can therefore be applied on time-resolved biological data to detect major changes as breakpoints in the systems behavior based on the temporal correlation structure.

Various approaches have been developed for the MTS-seg problem[9–11], which can be categorized into four classes based on the computational methodology used: (1) clustering[4,12–14], (2) graphical models[15–20], (3) genetic algorithms[21–23], and (4) regression.

While the first three categories have been well-investigated (see Ref. 5 for further details), the regression-based approaches provide a novel strategy for addressing the MTS problem, especially if regularization techniques (*e.g.*, least absolute shrinkage and selector operator (LASSO)[24]) are considered[25]. The regression-based approaches must account for the realistic scenario with a small number of time points (n) and large number of variables (p), typically arising in biological settings. For instance, Davis *et al.*[26] applied minimum description length to detect the best fitting autoregressive (AR) model for each segment. This approach was recently extended in Ref. 23, where each segment is represented by a piecewise quantile regression model penalized for description length. Another regression-based approach addresses the MTS-seg problem by applying piecewise constant function on the MTS data[27]. The breakpoints are estimated by using total variation penalty while small jumps from the zero-mean are discarded. This method has also been extended to solve the MTS-seg problem by reformulating it as group LASSO regression[28].

A further approach uses a discrete hidden logistic process which allows for smooth or abrupt changes in polynomial regression models. This method was first suggested to solve the problem of univariate time series segmentation/clustering[29,30] which was then extended to multivariate time-series data sets[31]. This method uses an expectation-maximization algorithm to estimate the model parameters in an unsupervised fashion; however, the

presented formulation of this approach relies on a pre-specified number of latent processes. Another recent approach that relies on dynamic programming is based on the simple piecewise polynomial regression mixture[32]; however, this approach may result in discontinuous segmentation, which is not meaningful in the analysis of time-series data from biological systems.

Finally, Preuß *et al.*[33] introduced a nonparametric approach to infer breakpoints in the autoco-variance structure of the multivariate piecewise stationary process which relies on comparison of spectral distribution on different segments.

Despite these recent developments, however, the existing regression-based approaches suffer from several shortcomings related to the applicability on large data sets and the necessity to *a priori* specification of the breakpoints in the system. Moreover, most of these approaches are designed and applicable only for the case where components of the system are independent, which does not hold true in biological applications[34]. Our contribution is threefold: First we propose a formulation of the MTS-seg problem based on a fused LASSO regression, whereby the natural order of time points (features) is imposed in the fusion. We then propose a novel method to estimate the significance of the determined breakpoints which relies on clustering within each of the detected segments. Finally, we identify the key components for the determined segments, based on quantifying the effect that the removal of a set of components has on the established segments. To this end, we employ two criteria: structural and ontology-based homogeneities. We extensively illustrate the biological relevance of the proposed method through a comparative case study with the state-of-the-art alternative methods on transcriptomics MTS data from various experimental scenarios on the yeast *Saccharomyces cerevisiae* and the diatom *Thalassiosira pseudonana*.

Results

Fused LASSO formulation of the MTS-seg problem. We formulate the approach for the MTS-seg problem by using fused LASSO to consider the inherent order of time points. Therefore, each variable, corresponding to a time point, is described by a vector over the considered components. Since the relationship between variables changes at a breakpoint, it is expected that the variables for the preceding time points have negligible explanatory power in the regression model for the breakpoint variable; analogously, a breakpoint variable is expected to have small explanatory contribution for the time points following it.

Therefore, in a regression setting with a given variable (time point) as a response and the variables corresponding to the preceding time points as regressors, the breakpoint variable is expected to have zero regression coefficient, provided that all variables are scaled and centered. This idea can be readily captured by the fused LASSO formulation: Given time-series for m variables over n time points, $T=\{t_l\}_{l=1}^{n}$, represented by a data matrix $M_{m\times n}$, we aim at determining a model for partitioning the time-series into k non-overlapping contiguous segments $P=[t_{i_0},t_{i_1}],[t_{i_1+1},t_{i_2}],\ldots,[t_{i_{k-1}+1},t_{i_k}]$, $1\le i_j < i_{j+1}\le n, 0\le j<k$, that span the whole series, *i.e.*, $i_0=1$ and $i_k=n$. Let $y_{t_{res}}$ be the profile of time point t_{res}, $(3\le res\le n)$, and let the matrix $x_{t_{reg}}$ includes the profiles of all subsequent time points t_{reg}, $(reg \in [1, res))$ which will be used as regressors. The fused LASSO is then given by:

$$\hat{\beta}=\mathrm{argmin}_{\beta}\left\|y_{t_{res}}-\sum_{reg=1}^{res-1}x_{t_{reg}}\beta_{t_{reg}}\right\|_2 + \lambda_1\sum_{reg=1}^{res-1}\left|\beta_{t_{reg}}\right|+\lambda_2\sum_{reg=2}^{res-1}\left|\beta_{t_{reg}}-\beta_{t_{reg-1}}\right|, \quad (1)$$

with β as a vector of regression coefficients. Solving the fused LASSO with every time point as a response, we obtain a lower-triangular matrix $C_{n\times n}$ containing the regression coefficients of the n models.

Breakpoints are then determined by examining the sequence A obtained by averaging the absolute values from each column of the matrix C. This sequence A summarizes the overall behavior and temporal patterns of the components over the examine time domain. The breakpoints correspond to the local minima of A, given by the time point i for which $A_{i-1}>A_i$ and $A_i<A_{i+1}$.

The assumption is that each segment captures a specific trend in the system's behavior. Therefore, the data profiles of the time points in a given segment should be similarly explained by the preceding time points. As a result, the breakpoints are those with weak relation. To this end, the "fusion" helps detect the consecutive time points, preceding the time point corresponding to the response, which can similarly explain the behavior of the response time point. The similarity of the explanatory behavior is captured by the closeness of the respective coefficients, assessed by the respective absolute value of their difference. While the LASSO constraint helps identify the likely true dependencies between regressors and the response (in case of large time-series data), the fusion penalty contributes to minimizing the differences of the coefficients (capturing the dependencies) between successive time points.

Estimating the statistical significance of the segments. Segmentation constraints require that segments are internally homogeneous and externally heterogeneous with respect to the behavior of the components. These constraints can be readily examined by considering the quality of the clusters extracted from the different segments by using cluster quality indices[35]. The significance level of the resulting segmentation can be estimated by permutation testing in the following fashion: First, the segmentation approach is applied on the original data set and the average cluster quality index of choice is estimated over the resulting segments, denoted by S_0:

$$S_0=\frac{1}{k}\sum_{i=1}^{k}Q_i \quad (2)$$

where k is the number of clusters and Q_i is the quality index of the i'th cluster (segment). Then, the time points are randomly permuted and the segmentation approach is applied on the permuted data set. The resulting breakpoints are then used to segment the original data set. The empirical p-value is obtained as:

$$p=\frac{\sum_{j=1}^{B}I(S_j>S_0)+1}{B+1} \quad (3)$$

where S_j is calculated using Eq. 2 with the results obtained from the j'th permutation, I is the indicator function, and B is the number of permutations.

Determining key components. The breakpoints partition the time-series data into the sequence of segments. Each segment can reveal information about the temporal behavior of the system.

Given a set of segments $P=[t_{i_0},t_{i_1}],[t_{i_1+1},t_{i_2}],\ldots,[t_{i_{k-1}+1},t_{i_k}]$ for which $1\le i_j<i_{j+1}\le n$, $0\le j<k$ while $i_0=1$ and $i_k=n$, next we introduce an approach to discover the key components for each segment which may be responsible for the breakpoints. To this end, we rely on the idea that a component is considered key if the removal of its time-series disturbs the estimated breakpoints on the entire data set. Determining the key components can be obtained by considering two criteria:

- *Structural homogeneity*: This criterion can be applied in cases that variables in the time-series data are not well characterized (annotated). In this case, the following steps are repeated for all except the last segment in P (Algorithm 1). First, the data profiles located in the time interval corresponding to the segment $[t_{i_j},t_{i_{j+1}}]$ are clustered by using an algorithm of choice (the partitioning around medoid (pam) clustering is used in this analysis[36]). Then, the profiles of the components in each cluster are iteratively removed,

and the segmentation is determined on the remaining data profiles, followed by inspection of any change in the breakpoints. To determine the maximum number of key components in a feasible manner, the clustering is performed starting with $l = 2$ clusters. The number of clusters, l, is increased by one if none of the l clusters affected the breakpoints. The clustering procedure is repeated until at least one cluster is found to be the key for the segmentation or the number of clusters equals the number of components.

- *Ontology-based homogeneity*: Biological data can be grouped based on conceptual features given in an ontology. For example, in case of transcriptomics data, genes can be grouped based on the pathways or biological processes in which they participate. Therefore, an analogous approach as for the structural homogeneity can be applied. The divisive step can be readily applied due to the hierarchical nature of the existing ontologies[37]. More specifically, to cluster the profiles of genes based on their biological homogeneity, we proceeded as follows: (1) The GO terms were obtained for the genes to be clustered. (2) The *k*means clustering was applied for the $k = 2$: 40 number of clusters. (3) For each k, the biological homogeneity of the clusters was estimated using biological homogeneity index[38]. (4) The value for k which was associated with the maximum biological homogeneity index was stored. (5) The procedure was repeated 1000 times starting from step (2). (6) The histogram of the obtained values for k from the 1000 runs was used to determine the most frequent value of k which resulted in the highest biological homogeneity for the clusters.

Applications of the approach. We applied the proposed approach for MTS-seg problem based on fused LASSO regression to several data sets, including: synthetic, yeast's metabolic and cell cycles, and *Thalassiosira pseudonana*'s diel growth state transition. Moreover, we compared its performance with the other regression-based approach intended for solving the MTS-seg problem.

Synthetic data. To investigate the performance of the algorithm, we created synthetic time-series data for 80 variables over 40 time points (see Figure 1 and Methods). The segmentation points corresponded to time points 5, 13, 25, 28 and 35. Figure 1 illustrates the segments obtained by the proposed approach with the *p–value* of 0.04 for 1000 permutations. The synthetic MTS data are segmented into 8 segments with breakpoints at 6, 13, 18, 20, 25, 29, and 35. All real breakpoints were captured, and only two additional breakpoints were included. This was likely due to the apparent change in the behavior of the time-series between the time points 13 and 25, which was not controlled in the generation of the data (see Methods).

In contrast, the group fused LASSO approach from Bleakley *et al.*[27] on the same data resulted in the five segments with following breakpoints: 6, 12, 16, and 24. However, this approach could not detect the late breakpoints at 28 and 35. The extensive comparison of the contending methods with our approach is given in the Supplementary information S1 (including Fig. S1 and Table S1).

The time-series data were next structurally clustered in each segment to obtain the key components. The colored curves at each segment show the key components which led to a structural change at the specified breakpoint. Since detection of key components is based on clustering, only the segments with more than two time points were inspected for the key components.

Yeast's metabolic and cell cycles. Motivated by the predictions from applying the approach on the synthetic data set, we next investigated the MTS-seg on the transcriptomics data sets from yeast metabolic cycle[39], cell cycle[40], and the experiment capturing the effect of oxidative stress, induced by hydrogen peroxide (HP) treatment, on the yeast's cell cycle[41]. In all data sets, we filtered out the genes which: (1) contain missing values, (2) have no gene ontology (GO) annotation, and (3) their coefficients of variation are smaller than 1. We focused on yeast's metabolic cycle, and the results for the other data sets as well as the comparison with other methods are detailed in the Supplementary Information S1 (including Fig. S2–S4 and Table S2–S4).

The yeast metabolic cycle (YMC) consists of the following three successive phases spanning each ~5 h: (1) a reductive charging (R/C) phase, involving non-respiratory metabolism (glycolysis and fatty acid oxidation) and protein degradation, (2) oxidative metabolism (Ox), in which respiratory processes are used to generate adenosine triposphate (ATP), (3) reductive metabolism (R/B), marked by a decrease in oxygen uptake and dominance of DNA replication, mitochondrial biogenesis, ribosome biogenesis, and cell division[39]. The data set included the time-resolved expression of 6555 genes (with 9335 microarray probes) over 36 time points (separated by ~25-min intervals) over three consecutive cell cycles. Clustering of the obtained transcript profiles was employed in Tu *et al.*[39] to show that YMC controls the timing of key cellular and metabolic processes to allow coordination of anabolic and catabolic processes for efficient energy production and usage. Therefore, this data set can serve as a

Algorithm 1: Key components detection.

Data: T time-series data with n time points

$$P = [t_{i_0}, t_{i_1}], [t_{i_1+1}, t_{i_2}], \ldots, [t_{i_{k-1}+1}, t_{i_k}]$$

Result: *key*, list of key components for each breakpoint

```
begin
    l ← 2
    for each segment p in P do
        while key is empty do
            cls ← cluster the segment p into l clusters
            for each cluster cl in cls do
                Ť ← remove cl from T
                segment Ť (Algorithm 2)
                if the estimated breakpoints disturbed then
                    key ← cl
            if key is empty then
                increase l
```

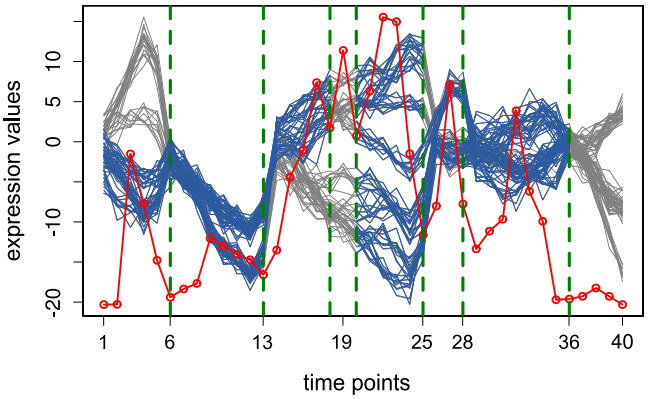

Figure 1 | Segmentation over synthetic data. The green dashed lines show the obtained breakpoints. The blue colored curves at each segment illustrate the respective key components. The gray colored parts of the time-series denote the variables not involved in the local changes at the corresponding breakpoints. The red dots, connected by a red line, represent the sequence A (column-averages of the absolute values of the regression coefficients in the matrix C).

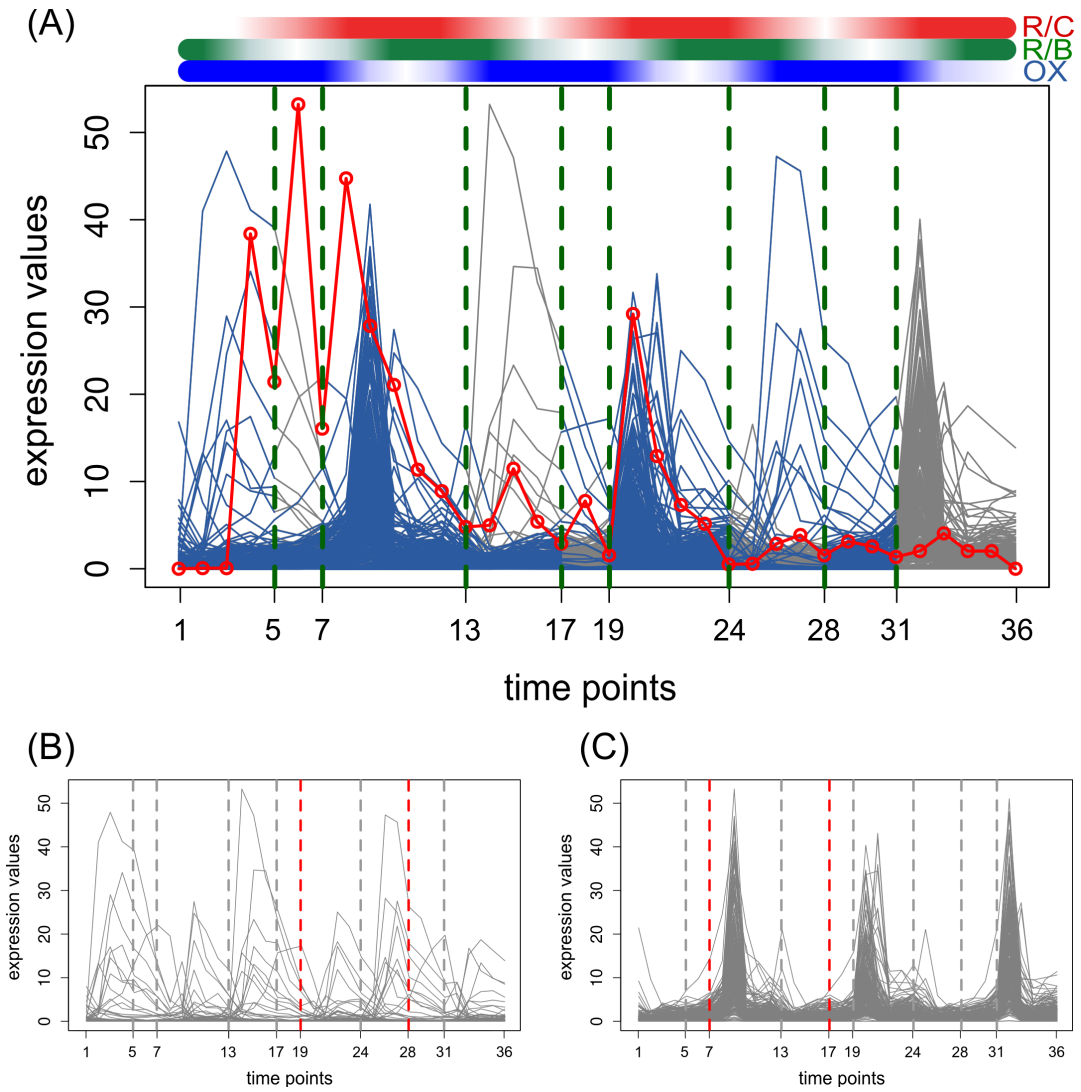

Figure 2 | Segmentation over yeast's metabolic cycle. (A) The expression profiles of 255 genes over 36 time points (separated by ~25-min intervals) over three consecutive cell cycles. The green dashed lines denote the obtained breakpoints. The blue colored curves at each segment illustrate the key components. The gray colored parts of the time-series denote the variables not involved in the local changes at the corresponding breakpoints. The red dots, connected by a red line, represent the sequence *A* (column-averages of the absolute values of the regression coefficients in the matrix *C*). Genes were grouped into two clusters (B) and (C) based on their biological homogeneity. (B) includes 41 genes which are predominantly involved in M, G1, and G1/S phases and (C) contains 214 genes related to the metabolic processes. Each cluster is specifically responsible for the breakpoints represented by red dashed lines.

benchmark for testing our proposed algorithms for the MTS-seg problem.

With the filtering steps mentioned above, the number of genes was reduced from 6555 to 255. The latter were employed to determine the segmentation based on the proposed approach. Due to the presence of recurrent changes on the global level, two segmentation points, at 12–13 and 24–25 should be detected. These breakpoints delineate the three considered cell cycles. In addition, due to the presence of the alternation phases in the metabolic cycle, each of the three cycles should contain at least one more segmentation point.

Applying the proposed approach on YMC time-series data resulted in 9 segments which was significant at the level of 0.088. Figure 2.A illustrated the segmentation and the Supplementary Table S2 compares the segmentation results with the previous studies[4,5,27]. As it is shown in the Figure 2.A, the starting points of the three aforementioned cycles were captured by this approach. The inferred breakpoints are at the 5, 7, 13, 17, 19, 24, 28 and 31 time points, and the breakpoints delineating the three considered cell cycles (time points 13 and 24) could be precisely detected. The approach of Bleakley *et al.*[27] failed to infer all three cycles, as well as the other breakpoints in each cycle (breakpoints at 8, 9, 26, 31, and 32).

We used biological homogeneity index to cluster the genes based on the biological process (BP). Two clusters were generated as a result, including 41 and 214 genes, respectively. Genes in the first cluster were predominantly involved in the M, G1, and G1/S phases of the cell cycle (Figure 2.B), while genes in the second cluster were related to the metabolic processes (Figure 2.C), as determined by gene enrichment analysis (Supplementary Table S5).

In Figures 2.B and 2.C, the red dashed lines visualize the points at which the specific cluster highly contributed to the inferred breakpoints. Considering the third cycle (time interval between 25 h and 36 h), genes that were involved in M, G1, and G1/S phases were identified to contribute to the detection of a breakpoint at time point 28 h, while the genes involved in metabolic processes were involved at the breakpoint at 31 h.

Diel growth state transition of diatom Thalassiosira pseudonana. The proposed approach is also applicable to short time-series data. To

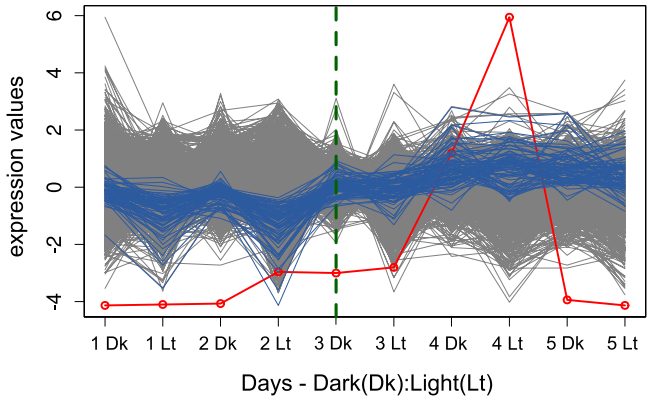

Figure 3 | Segmentation over diel growth state transition of diatom *Thalassiosira pseudonana*. The expression profile of 5417 genes illustrated over cycles of 12 : 12 h dark(Dk) : light(Lt) in 5 days. The green dashed line shows the obtained breakpoint. The blue colored curves illustrate the key components within the first segment which led to the only structural changes at [3 Dk]. The gray colored parts of the time-series denote the variables that were not involve in the local changes at the corresponding breakpoints, as detected by the approach. The red dots, connected by a red line, represent the sequence A (column-averages of the absolute values of the regression coefficients in the matrix C).

illustrate this application, we investigated the MTS-seg on transcriptomics MTS data from *Thalassiosira pseudonana*'s diel growth state transition[42]. Ashworth *et al.*[42] measured the transcript level of *Thalassiosira pseudonana* in five days on a 12 : 12 h dark(Dk) : light(Lt) cycle to find the key regulators responsible for exponential and stationary phase modulation as well as diel phase reversal (Fig. 3). They reported a major shift on the third day between 12 h dark and 12 h light (3 Dk and 3 Lt). Up to the third day 12 h light (3 Lt), genes labeled as "exponential" showed higher expression than genes labeled as "stationary"; however, from the third day light (3 Lt) on, the stationary genes showed higher expression. Consequently, the time between 3 Dk and 3 Lt could be considered as a transition period due to the shift between exponential and stationary phase causing major change in gene expression levels.

The proposed MTS segmentation approach applied on the transcription profiles of 5417 genes (with the coefficient of variation above 0.3) over 10 time points resulted in 2 segments with breakpoint at 3 Dk with the significance level of 0.14 illustrated in Figure 3.

Next, we used three groups of genes annotated with exponential and stationary phase modulation as well as diel phase reversal (Supplementary Table S7 obtained from Supplementary Dataset1 in Ref. 42). However, the removal of all these genes did not affect the result of the clustering which indicates further genes that also affect the shift between 3 Dk and 3 Lt. Therefore, and due to the lack of GO annotation for *Thalassiosira pseudonana*, we used structural clustering in order to detect the key components. We could obtain 82 genes as key components whose behavior led to the break point at 3 Dk (Supplementary Table S6). The profile of these genes are highlighted in blue in Figure 3. Among the key components 9 genes were in the list of genes which were more highly expressed at dawn during the exponential and diel phase, based on the Supplementary Dataset1 from Ref. 42.

Algorithm 2: Regularized segmentation.

Data: T time-series data with n time points
Result: BP, breakpoints
begin
  for i *in* n: 3 **do**
    $y \leftarrow T[,t_i]$
    $x \leftarrow t(T[,1:(t_i-1)])$
    estimate β based on Eq. 1
    $C[i,] \leftarrow c(rep(0, i), \beta)$
  $A \leftarrow$ average over columns of matrix C
  $BP \leftarrow$ time point i for which $A_{i-1} > A_i$ and $A_i < A_{i+1}$

Discussion

Here we introduced a regression formulation of the MTS-seg problem based on fused LASSO. The breakpoints were determined by inspecting the changes in the regression coefficients over a series of regression models. In addition, we proposed a method to determine the statistical significance analysis of the inferred breakpoints by applying a cluster-based approach and a cluster quality index of choice.

We note that all findings on the statistical significance of the found breakpoints were obtained based on the cluster quality measure. Due to the difference in behavior of different cluster quality indices[43], the determined p-values should be carefully interpreted based on the properties of the investigated data sets. Nevertheless, the proposed procedure provides a general method to couple clustering in segments with the determined breakpoints with the aim of establishing the significance of the latter.

Moreover, we devised a clustering-based approach to identify the key components giving rise to the abrupt changes in the system. We could identify the order of processes for a metabolic cell cycle from yeast data set. In addition, application of this approach to diel growth state transition of diatom result in a group of key components which are highly expressed at dawn during exponential and diel phase. This approach elaborate on biological aspect of the dynamic relationships underlying biological processes.

Applying the method to different data sets supported the reliability and significance of the determined breakpoints in the well-documented cases of yeast's cell and metabolic cycles. Unlike other approaches, we did not impose restriction on the number of time points included in each segment, rendering our method applicable to short time-series typical in biological studies. In addition, the comparative analysis demonstrated that the regression-based approach performs better in comparison to the state-of-the-art algorithm.

Improvement to the proposed approach can be obtained by imposing constraints to the fused LASSO regression such as in the recent study from Sue and Tibshirani *et al.*[44]. To further investigate on the accuracy and sparsity of the breakpoint set, various constraints can be imposed to the fused LASSO. For instance, if a time point as a regressor does not have explanatory power for the response time point, the regression coefficients of all preceding time points can be neglected.

Formulation of the segmentation with additional constraints, including different segmentations for groups of entities, would render our approach widely applicable in different fields.

Methods

Segmentation algorithm. The segmentation algorithm, given in Algorithm 2, is implemented in R [http://www.R-project.org] by using the package penalized [http://cran.r-project.org/web/packages/penalized]). The regression coefficients were robustly estimated by K-fold cross validation (K selected based on the available time points) together with the optimal values for λ_1 and λ_2 from the range [1,50]. We assumed that no breakpoints could occur in the first and the last three time points of the time-series, as robust changes could only be detected after at least 4 consecutive time points[5]. The implementation is available at http://mathbiol.mpimp-golm.mpg.de/Segmentation-fLASSO/index.html.

Significance and key components of segmentation. In the current implementation, the significance level of the segmentation was based on the average silhouette width[45]. To determine the structural homogeneity, partitioning around medoids algorithm (pam) function in the R package cluster [http://stat.ethz.ch/R-manual/R-patched/library/cluster/html/pam.html] with Pearson correlation was employed to cluster data profiles. For biological data, the ontology-based homogeneity was inspected based on the biological process (BP) category of the Gene Ontology[46]. The GO enrichment analysis for each cluster was performed using hypergeometric test for which we used phyper function from R package stats [https://stat.ethz.ch/R-manual/

R-patched/library/stats/html/Hypergeometric.html] and annotation package ygs98.db [http://www.bioconductor.org/packages/release/data/annotation/html/ygs98.db.html].

Synthetic data set. To create these segmentation points, a number of data profiles were generated for each segment by simulating a zero-mean autoregressive moving average (ARIMA) model (described in Ref. 5). The number of profiles simulated for the six segments, [1–5], [6–13], [14–25], [26–28], [29–35] and [36–40], was set to 5, 2, 8, 3, 7 and 4, respectively. Each of the 80 variables was obtained by randomly sampling a characteristic data profile in each segment. In addition, a normally distributed error term, $N(0, 1)$, was added to the sampled profile value at each time point. The code for generating synthetic data is available at http://mathbiol.mpimp-golm.mpg.de/Segmentation-fLASSO/index.html.

1. Bellman, R. & Roth, R. Curve fitting by segmented straight lines. *J. Am. Statist. Assoc.* **64**, 1079–1084; doi:10.1080/01621459.1969.10501038 (1969).
2. Keogh, E., Chu, S., Hart, D. & Pazzani, M. Segmenting time series: A survey and novel approach. *Work* **57**, 1–21; doi:10.1142/9789812565402 0001 (2003).
3. Fan, J., Lv, J. & Qi, L. Sparse high dimensional models in economics. *Annu. Rev. Econom.* **3**, 291–317; doi:10.1146/annurev-economics-061109-080451 (2011).
4. Ramakrishnan, N. *et al.* Reverse engineering dynamic temporal models of biological processes and their relationships. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 12511–12516; doi:10.1073/pnas.1006283107 (2010).
5. Omranian, N., Klie, S., Mueller-Roeber, B. & Nikoloski, Z. Network-based segmentation of biological multivariate time series. *PLoS ONE* **8**, e62974; doi:10.1371/journal.pone.0062974 (2013).
6. McCarty, J. A. & Hastak, M. Segmentation approaches in data-mining: A comparison of rfm, chaid, and logistic regression. *J. of Bus. Res.* **60**, 656–662; doi:10.1016/j.jbusres.2006.06.015 (2007).
7. Zou, C., Jiang, W. & Tsung, F. A lasso-based diagnostic framework for multivariate statistical process control. *Technometrics* **53**, 297–309; doi:10.1198/TECH.2011.10034 (2011).
8. Malone, J. H. & Oliver, B. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.* **9**, 34; doi:10.1186/1741-7007-9-34 (2011).
9. Horváth, L. & Rice, G. Rejoinder on: Extensions of some classical methods in change point analysis. *TEST* **23**, 287–290; doi:10.1007/s11749-014-0375-5 (2014).
10. Hušková, M. & Prášková, Z. Comments on: Extensions of some classical methods in change point analysis. *TEST* **23**, 265–269; doi:10.1007/s11749-014-0368-4 (2014).
11. Horváth, L. & Rice, G. Extensions of some classical methods in change point analysis. *TEST* **23**, 219–255; doi:10.1007/s11749-014-0368-4 (2014).
12. Abonyi, J., Feil, B., Nemeth, S. & Arva, P. Modified gath–geva clustering for fuzzy segmentation of multivariate time-series. *Fuzzy Set. Syst.* **149**, 39–56; doi:10.1016/j.fss.2004.07.008 (2005).
13. Duchene, F., Garbay, C. & Rialle, V. Learning recurrent behaviors from heterogeneous multivariate time-series. *Artif. Intell. Med.* **39**, 25–47; doi:10.1016/j.artmed.2006.07.004 (2007).
14. Tadepalli, S., Ramakrishnan, N., Mishra, B., Watson, L. T. & Helm, R. F. Deriving kripke structures from time series segmentation results. In *Discrete Event Systems, 2008. WODES 2008. 9th International Workshop on,* 406–411; doi:10.1109/WODES.2008.4605980 (IEEE, 2008).
15. Bai, J. & Perron, P. Computation and analysis of multiple structural change models. *J. Appl. Econometr.* **18**, 1–22; doi:10.1002/jae.659 (2003).
16. Yin, J., Shen, D., Yang, Q. & Li, Z.-N. Activity recognition through goal-based segmentation. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 1,* 28–33 (AAAI Press, 2005).
17. Xuan, X. & Murphy, K. Modeling changing dependency structure in multivariate time series. In *Proceedings of the 24th international conference on Machine learning,* 1055–1062; doi:10.1145/1273496.1273629 (ACM, New York, NY, USA, 2007).
18. Dobigeon, N., Tourneret, J.-Y. & Scargle, J. D. Joint segmentation of multivariate astronomical time series: Bayesian sampling with a hierarchical model. *IEEE Trans. Signal Process.* **55**, 414–423; doi:10.1109/TSP.2006.885768 (2007).
19. Picard, F., Lebarbier, E., Budinska, E. & Robin, S. Joint segmentation of multivariate gaussian processes using mixed linear models. *Comput. Stat. Data Anal.* **55**, 1160–1170; doi:10.1016/j.csda.2010.09.015 (2011).
20. Angelosante, D. & Giannakis, G. B. Sparse graphical modeling of piecewise-stationary time series. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 1960–1963; doi:10.1109/ICASSP.2011.5946893 (2011).
21. Tucker, A., Liu, X. & Ogden-Swif, A. Evolutionary learning of dynamic probabilistic models with large time lags. *Int. J. Intell. Syst.* **16**, 621–645; doi:10.1002/int.1027 (2001).
22. Graves, D. & Pedrycz, W. Multivariate segmentation of time series with differential evolution. In Carvalho, J. P., Dubois, D., Kaymak, U. & da Costa Sousa, J. M. (eds.) *IFSA/EUSFLAT Conference*, 1108–1113 (2009).
23. Aue, A., Cheung, R. C., Lee, T. C. & Zhong, M. Segmented model selection in quantile regression using the minimum description length principle. *J. Am. Stat. Assoc.* To appear; doi:10.1080/01621459.2014.889022 (2014).
24. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288, doi:10.1111/j.1467-9868.2011.00771.x (1996).
25. Vert, J.-P. & Bleakley, K. Fast detection of multiple change-points shared by many signals using group lars. In *NIPS*, 2343–2351 (2010).
26. Davis, R. A., Lee, T. C. M. & Rodriguez-Yam, G. A. Structural break estimation for nonstationary time series models. *J. Am. Stat. Assoc.* **101**, 223–239; doi:10.1198/016214505000000745 (2006).
27. Bleakley, K. & Vert, J.-P. The group fused lasso for multiple change-point detection. *arXiv preprint arXiv:1106.4199* URL http://arxiv.org/abs/1106.4199 (2011).
28. Yuan, M. & Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Series B Stat. Methodol.* **68**, 49–67; doi:10.1111/j.1467-9868.2005.00532.x (2006).
29. Chamroukhi, F., Samé, A., Govaert, G. & Aknin, P. Time series modeling by a regression approach based on a latent process. *Neural Networks* **22**, 593–602; doi:10.1016/j.neunet.2009.06.040 (2009).
30. Samé, A., Chamroukhi, F., Govaert, G. & Aknin, P. Model-based clustering and segmentation of time series with changes in regime. *Adv. Data Anal. Classif* **5**, 301–321; doi:10.1007/s11634-011-0096-5 (2011).
31. Chamroukhi, F., Mohammed, S., Trabelsi, D., Oukhellou, L. & Amirat, Y. Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing* **120**, 633–644; doi:10.1016/j.neucom.2013.04.003 (2013).
32. Chamroukhi, F. Piecewise regression mixture for simultaneous functional data clustering and optimal segmentation. *arXiv:1312.6974 [stat.ME]* 1312.6974; (2013).
33. Preuß, P., Puchstein, R. & Dette, H. Detection of multiple structural breaks in multivariate time series. *J. Am. Stat. Assoc.* to appear; doi:10.1080/01621459.2014.920613 (2014).
34. Pósfai, M., Liu, Y.-Y., Slotine, J.-J. & Barabási, A.-L. Effect of correlations on network controllability. *Sci. Rep.* **3**; doi:10.1038/srep01067 (2013).
35. Wagner, S. & Wagner, D. Comparing clusterings- an overview Technical Report 2006-04, ITI Wagner, Informatics, Universität Karlsruhe. (2007).
36. Reynolds, A. P., Richards, G., de la Iglesia, B. & Rayward-Smith, V. J. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *JMMA* **5**, 475–504; doi:10.1007/s10852-005-9022-1 (2006).
37. Rhee, S. Y., Wood, V., Dolinski, K. & Draghici, S. Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.* **9**, 509–515; doi:10.1038/nrg2363 (2008).
38. Datta, S. & Datta, S. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics* **7**, 397; doi:10.1186/1471-2105-7-397 (2006).
39. Tu, B. P., Kudlicki, A., Rowicka, M. & McKnight, S. L. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science* **310**, 1152–1158; doi:10.1126/science.1120499 (2005).
40. Spellman, P. T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297; doi:10.1091/mbc.9.12.3273 (1998).
41. Shapira, M., Segal, E. & Botstein, D. Disruption of yeast forkhead-associated cell cycle transcription by oxidative stress. *Mol. Biol. Cell* **15**, 5659–5669; doi:10.1091/mbc.E04-04-0340 (2004).
42. Ashworth, J. *et al.* Genome-wide diel growth state transitions in the diatom thalassiosira pseudonana. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 7518–7523; doi:10.1073/pnas.1300962110 (2013).
43. Delling, D., Gaertler, M., Görke, R., Nikoloski, Z. & Wagner, D. How to evaluate clustering techniques (University of Karlsruhe, Faculty of Informatics, 2006).
44. Suo, X. & Tibshirani, R. An ordered lasso and sparse time-lagged regression. *arXiv preprint arXiv:1405.6447* URL http://arxiv.org/abs/1405.6447 (2014).
45. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65; doi:10.1016/0377-0427(87)90125-7 (1987).
46. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29; doi:10.1038/75556 (2000).

Author contributions

Conceived and designed the study: N.O. and Z.N. Developed the model: N.O. and Z.N. Implemented: N.O. Wrote the paper: N.O., B.M.R. and Z.N. All authors read the manuscript, made comments and approved the final version submitted.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Omranian, N., Mueller-Roeber, B. & Nikoloski, Z. Segmentation of biological multivariate time-series data. *Sci. Rep.* **5**, 8937; DOI:10.1038/srep08937 (2015).