Unsupervised modeling of cell morphology dynamics for time-lapse microscopy

Qing Zhong[1,2], Alberto Giovanni Busetto[2,3], Juan P Fededa[1], Joachim M Buhmann[2,3] & Daniel W Gerlich[1,3,4]

Analysis of cellular phenotypes in large imaging data sets conventionally involves supervised statistical methods, which require user-annotated training data. This paper introduces an unsupervised learning method, based on temporally constrained combinatorial clustering, for automatic prediction of cell morphology classes in time-resolved images. We applied the unsupervised method to diverse fluorescent markers and screening data and validated accurate classification of human cell phenotypes, demonstrating fully objective data labeling in image-based systems biology.

Automated live-cell microscopy and large-scale perturbation by RNA interference (RNAi) is a powerful approach to investigate gene function[1–5]. The analysis of huge data sets requires automated procedures, typically comprising computational image processing followed by statistical analysis of cellular morphologies. State-of-the-art methods rely on user-labeled training data to detect cellular phenotypes[5–12], but annotating such data can be time consuming and limited by subjectivity. We introduce an annotation-free method to model cellular morphologies based on unsupervised learning.

As reference data for performance measurements, we used time-lapse microscopy images of human tissue culture cells (HeLa 'Kyoto' cells) expressing a fluorescent chromatin marker (histone H2B–monomeric (m) Cherry; **Fig. 1a**). The task consists of classifying cell morphologies corresponding to interphase and to the five phases of mitosis (prophase, prometaphase, metaphase, anaphase and telophase). Despite the well-defined chromatin morphology of mitotic stages, user annotation can be inconsistent. Dissimilarity analysis of annotations performed by three biologists reveals modest inconsistencies between annotations performed by the same person on different days, whereas annotations performed by different users varied extensively (**Fig. 1b** and **Supplementary Fig. 1**); these problems highlighted the need for objective data-annotation procedures. We obtained a consensus approximation to 'ground truth' (by a majority vote of multiple user annotations) to serve as the gold standard for the labels.

The benchmark data consists of seven image sequences with 326 cell division events, uniformly sampled with a time-lapse interval of 4.6 min[7]. Using CellCognition software[7], we segmented 13,040 individual cell objects. We extracted statistical features describing texture and shape for each cell object and tracked cells over time[7] to assemble a synchronized time series of cell features (**Supplementary Note 1**).We converted these time series using principal component analysis (PCA) to reduce data dimensionality (**Fig. 1c**,**d** and **Supplementary Fig. 2**), and this converted data served as a starting point to model cell morphologies with various methods (**Fig. 1e**). When we applied a basic probabilistic model for multiple Gaussian distributions of cell objects—Gaussian mixture modeling (GMM)—we observed a poor match between clustered individual cell objects and user annotation (**Fig. 1f**,**g**). The low performance of GMM can be explained by the known risk of its convergence to unsatisfactory local maxima and sensitivity to initial conditions. Thus, additional knowledge needs to be incorporated for effective unsupervised learning of cell morphologies.

We developed a temporally constrained combinatorial clustering (TC3) procedure (Online Methods and **Supplementary Fig. 3**) in which temporally linked objects are clustered according to a user-defined number of categories for each cell trajectory. We first detected mitotic subgraphs by binary clustering to generate three subgraphs (**Supplementary Fig. 4a**), which were then individually clustered by the TC3. This strategy makes it practical to conduct an exhaustive search for the best cluster assignment in a large time series (**Supplementary Note 1**). The TC3 labels and user annotations show substantial agreement (**Fig. 1f**,**h**).

To further improve classification accuracy and to build the basis of a dynamic model, TC3 was used to initialize a GMM based on the sample mean and covariances from classes of TC3-labeled cell objects. Indeed, the TC3-initialized GMM generated a label matrix that was more similar to user annotation (**Fig. 1f**,**i**) than to stand-alone GMM (**Fig. 1g** and **Supplementary Note 2**). The process was then extended to hidden Markov models (HMM); in this context, transition probabilities capture the temporal dynamics of cell states, whereas observation densities characterize cell morphologies. This model predicted dynamic cell behavior and generated labels that matched user annotation with further-improved accuracy (**Fig. 1j** and **Supplementary Fig. 4b**).

We compared the performance of our unsupervised method (TC3 + GMM + HMM) with that of supervised methods. State-of-the-art support vector machine (SVM) classification with HMM error correction yields low rates of obvious errors, as visible from post-classification inspection[7] (**Fig. 1k**). Ambiguous morphologies at class transitions, however, make the estimation of absolute error rates challenging. We therefore established the

[1]Institute of Biochemistry, ETH Zurich, Zurich, Switzerland. [2]Department of Computer Science, ETH Zurich, Zurich, Switzerland. [3]Competence Center for Systems Physiology and Metabolic Diseases, ETH Zurich, Zurich, Switzerland. [4]Present address: Institute of Molecular Biotechnology of the Austrian Academy of Sciences (IMBA), Vienna, Austria. Correspondence should be addressed to D.W.G. (daniel.gerlich@imba.oeaw.ac.at).

RECEIVED 22 DECEMBER 2011; ACCEPTED 30 APRIL 2012; PUBLISHED ONLINE 27 MAY 2012; DOI:10.1038/NMETH.2046

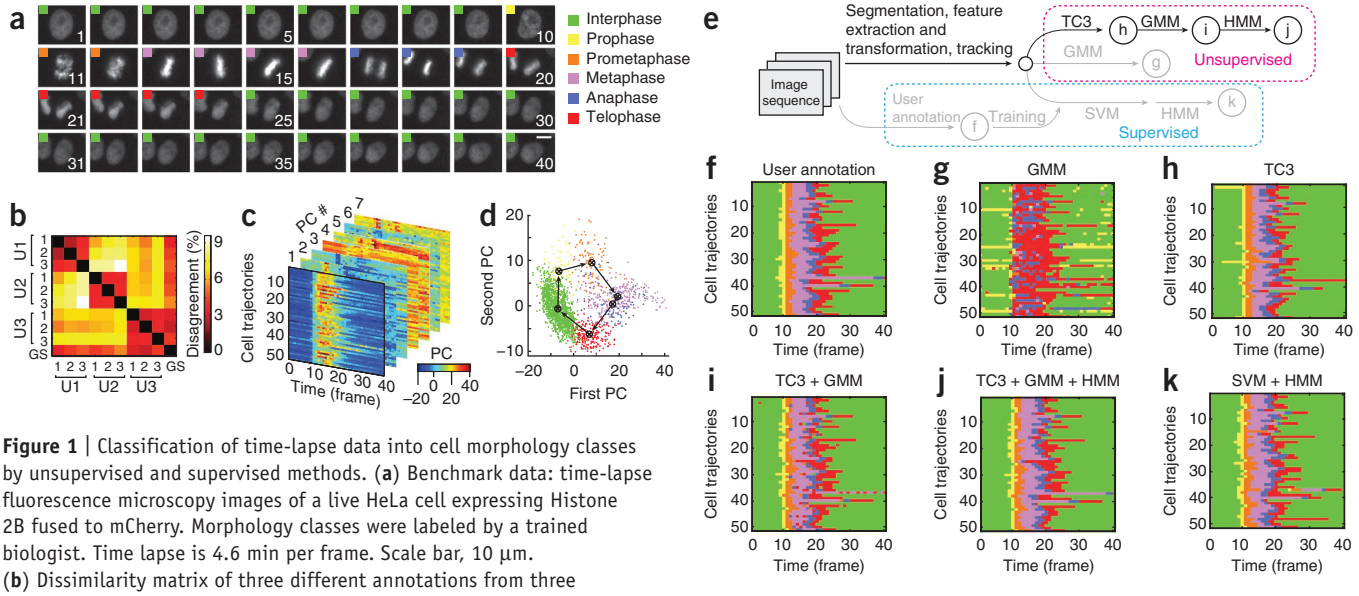

Figure 1 | Classification of time-lapse data into cell morphology classes by unsupervised and supervised methods. (**a**) Benchmark data: time-lapse fluorescence microscopy images of a live HeLa cell expressing Histone 2B fused to mCherry. Morphology classes were labeled by a trained biologist. Time lapse is 4.6 min per frame. Scale bar, 10 μm. (**b**) Dissimilarity matrix of three different annotations from three users (U1, U2, U3). The gold standard (GS) was obtained by majority vote. (**c**) First seven principal-component (PC) features of 51 cell trajectories. (**d**) Cell object distribution in two-dimensional PCA subspace. Dots refer to individual cell objects, color-coded as in **a**. Black crosses indicate sample means of morphology classes derived from human annotations. Arrows indicate temporal progression. (**e**) Workflows of cell morphology classification by different methods. Circled letters correspond to the respective results shown in **f**–**k**. Primary methodology developed in this study is displayed in black and alternative methods in gray. (**f**) Label matrix of user annotation for all cells of the image sequence corresponding to the example shown in **a**, color-coded as in **a**. The matrix consists of 51 cell trajectories (rows) and 40 time frames (columns). (**g**–**k**) Labels generated by GMM (**g**), TC3 (**h**), GMM after TC3-initialization (**i**), HMM extension of TC3 + GMM (**j**) and SVM followed by HMM error correction (**k**).

following validation procedure: the set of cell trajectories derived from each of the seven image sequences was used independently to train an SVM[7]; the cell trajectories from the other six image sequences constituted a test set. SVM classification followed by HMM error correction yielded total accuracies slightly lower than those of the unsupervised method (**Supplementary Fig. 5** and **Supplementary Table 1**; **Supplementary Note 3**). Thus, unsupervised learning can achieve classification accuracy similar to that of supervised methods.

To estimate the extent of user bias on SVM classification, SVM model selection as in ref. 7 was performed for each of nine annotations by three users of a training image sequence. The nine models obtained were evaluated on nine annotations from three users and the gold-standard annotation of a distinct test image sequence. Total accuracies for the labels generated by the nine selected models varied extensively (**Supplementary Fig. 6**), indicating that supervised SVM classification may be inconsistent. Involving multiple users for a gold-standard annotation can reduce this problem, yet this is often difficult in a typical laboratory setting. The independence from user annotation of our unsupervised method hence provides an important advantage over supervised methods. We used biological knowledge about mitosis to define the number of clusters in the benchmark data. When such detailed a priori knowledge is not available, determining the optimal cluster number can be difficult. In such a case, our unsupervised method can be used to automatically visualize labeling schemes with a variable number of clusters (**Supplementary Fig. 7**). Visual inspection of the resulting label matrices then simplifies the determination of an optimal cluster number.

We explored the versatility of the unsupervised method by applying it to additional markers. Live imaging of HeLa cells expressing EGFP-tagged proliferating cell nuclear antigen (PCNA) visualized progression through the G1, S and G2 phases through a changing pattern of replication foci[7] (**Fig. 2a**). Classification by our unsupervised method yielded labels that closely matched user annotation (**Fig. 2b**,**c**; 86.7% total accuracy). Mitotic spindle dynamics of HeLa cells expressing fluorescently tagged α-tubulin (**Supplementary Fig. 8**) were classified with a slightly lower total accuracy (75.5%) because of confusion between midbody and interphase morphologies (**Supplementary Fig. 8d**,**e**). This may have been caused by an under-representation of statistical features responding to single small-scale structures; implementation of appropriate features could solve this problem.

Finally, we analyzed several well-characterized RNAi phenotypes of cell division regulators: a spindle checkpoint protein (Mad2), subunits of the chromosomal passenger complex (INCENP and Aurora B), an activator of the anaphase-promoting complex (Cdc20), a microtubule motor protein (CENP-E), a spindle assembly regulator (Tpx2) and a regulator of sister chromatid cohesion (separase). Time-lapse image sequences of RNAi-treated H2B-mCherry–expressing HeLa cells were acquired on a screening microscope as in refs. 4,7 in three independent experimental replicates (363,120 individual cell objects, **Supplementary Table 2**; examples shown in **Supplementary Fig. 9**). The unsupervised method detected changed timing of mitotic phases (**Fig. 2d** and **Supplementary Fig. 10**) matching the known function of the respective RNAi target proteins. For example, Mad2 RNAi shortened metaphase[13], Cdc20 RNAi prolonged metaphase[14] and INCENP RNAi prolonged prometaphase and metaphase[15]. Some RNAi experiments caused permanent mitotic arrest, demonstrating that the method also detects terminal phenotypes (**Supplementary Fig. 11**).

Visualization of GMM model parameters revealed characteristic morphological deviations for several RNAi phenotypes (**Fig. 2e**

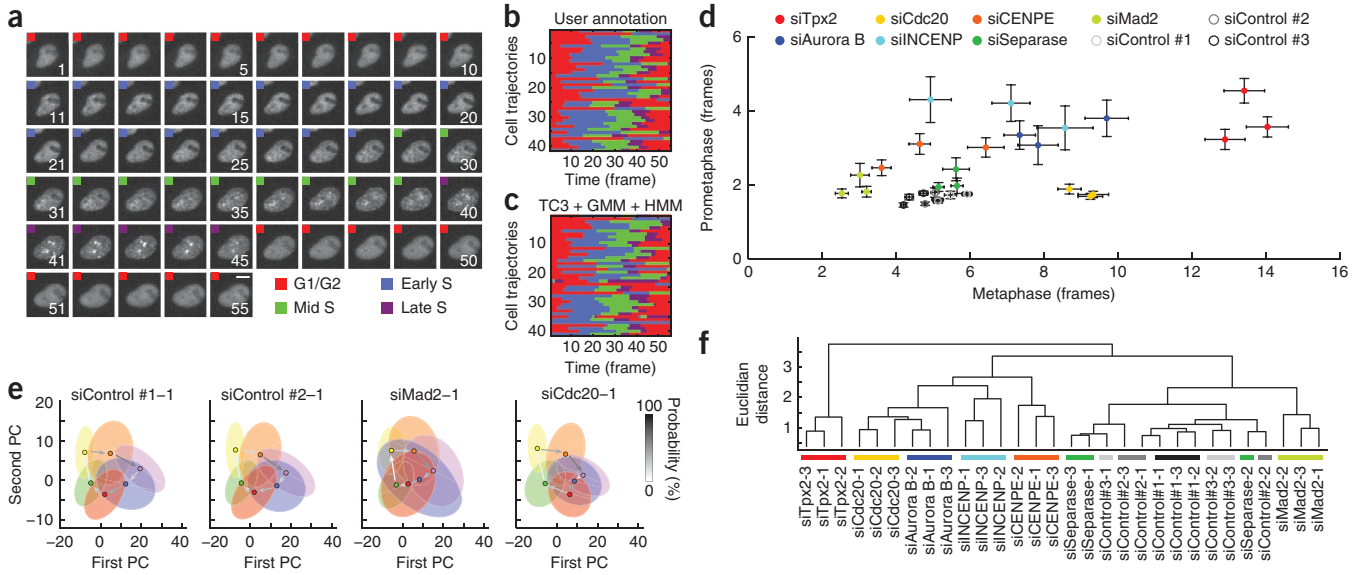

Figure 2 | Unsupervised classification of images with different morphology markers and RNAi phenotypes. (**a**) Time-lapse images of a HeLa cell expressing EGFP-PCNA. Four morphology classes were labeled by a trained biologist. Time lapse is 11.8 min per frame. Scale bar, 10 μm. (**b**) Label matrix of user annotation for EGFP-PCNA image sequence (41 cell trajectories, 55 time frames). (**c**) Labels generated by TC3 + GMM + HMM. (**d**) The plot shows the duration of the mitotic phases in H2B-mCherry–expressing HeLa cells transfected with the indicated siRNAs as determined by the unsupervised TC3 + GMM + HMM method. Time lapse is 4.6 min per frame. Dots indicate mean timing per image sequence; error bars indicate s.e.m. (sample numbers: **Supplementary Table 2**); $n = 3$. Independent experimental replicates on different days. (**e**) Visualization of models derived from data for four RNAi conditions in **d**. Ellipsoids depict GMM component densities at 68% confidence interval, color-coded according to mitotic morphology classes as in **Figure 1a**. Black circles indicate Gaussian component means. Arrows indicate temporal progression. Gray values of arrows indicate transition probabilities derived from HMMs. (**f**) Hierarchical phenotype clustering of RNAi screening data from **d**. Colors correspond to target genes as in **d**.

and **Supplementary Fig. 12**). We therefore included spatial model parameters in the analysis of the RNAi phenotypes. Hierarchical clustering of RNAi experiments based on spatial and temporal features (**Fig. 2f**) showed that siRNAs targeting the same protein generally clustered together (with the exception of one siRNA targeting separase). Our unsupervised method even discriminated between siRNAs targeting two different subunits of the same protein complex (Aurora B and INCENP), indicating that TC3 + GMM + HMM has high sensitivity toward subtle phenotypic differences.

In conclusion, we present a method for consistent and objective annotation of time-resolved cellular imaging data. Such consistency is unattainable by supervised methods because they depend on user labels that can differ among individuals and over extended project durations. Avoiding user annotation in image-based screening facilitates large-scale data integration. Our unsupervised method generates spatio-temporal models of cellular dynamics and provides a powerful means for phenotype profiling.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

The authors thank M. Held for processing image data, D. Scheder and C. Sommer for critical comments on the manuscript and R. Stanyte for user annotations of image data. The Gerlich laboratory has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreements no. 241548 (MitoSys) and no. 258068 (Systems Microscopy), from a European Young Investigator award of the European Science Foundation, from an EMBO Young Investigator Programme fellowship to D.W.G. and from the Swiss National Science Foundation. The Buhmann laboratory has received funding from the SystemsX.ch initiative (LiverX and YeastX projects). J.P.F. was funded by an EMBO long-term fellowship.

AUTHOR CONTRIBUTIONS

Q.Z. contributed to method design and implementation, experiments, data analysis and manuscript writing. A.G.B. contributed to method design, data analysis and manuscript writing. J.P.F. contributed to experiments and data analysis. J.M.B. contributed to method design and data analysis. D.W.G. contributed to method design, data analysis and manuscript writing.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nmeth.2046. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Conrad, C. & Gerlich, D.W. *J. Cell Biol.* **188**, 453–461 (2010).
2. Goshima, G. *et al. Science* **316**, 417–421 (2007).
3. Collinet, C. *et al. Nature* **464**, 243–249 (2010).
4. Schmitz, M.H. *et al. Nat. Cell Biol.* **12**, 886–893 (2010).
5. Neumann, B. *et al. Nature* **464**, 721–727 (2010).
6. Boland, M.V. & Murphy, R.F. *Bioinformatics* **17**, 1213–1223 (2001).
7. Held, M. *et al. Nat. Methods* **7**, 747–754 (2010).
8. Harder, N. *et al. Genome Res.* **19**, 2113–2124 (2009).
9. Wang, M. *et al. Bioinformatics* **24**, 94–101 (2008).
10. Loo, L.H., Wu, L.F. & Altschuler, S.J. *Nat. Methods* **4**, 445–453 (2007).
11. Jones, T.R. *et al. Proc. Natl. Acad. Sci. USA* **106**, 1826–1831 (2009).
12. Conrad, C. *et al. Nat. Methods* **8**, 246–249 (2011).
13. Meraldi, P., Draviam, V.M. & Sorger, P.K. *Dev. Cell* **7**, 45–60 (2004).
14. Wolthuis, R. *et al. Mol. Cell* **30**, 290–302 (2008).
15. Mackay, A.M., Ainsztein, A.M., Eckley, D.M. & Earnshaw, W.C. *J. Cell Biol.* **140**, 991–1002 (1998).

ONLINE METHODS

Benchmark image data and image processing. Published fluorescence live-cell microscopy data[7] of dividing human HeLa cells expressing the chromatin marker H2B-mCherry, the DNA replication marker EGFP-PCNA and EGFP-α-tubulin were used as benchmark data. Object detection, tracking of cells over time and feature extraction were carried out using the image analysis framework CellCognition[7], and the resulting trajectories were converted to synchronized multivariate time series of identical length representing mitotic events. Details of the method are described below and in **Supplementary Note 1**.

The RNAi screening data consisted of 30 image sequences, recorded as three independent experimental replicates on different days for ten different RNAi conditions. Using CellCognition, 9,078 cell division events, or 363,120 individual cell objects, were detected. Raw features of all image sequences were jointly transformed by PCA. TC3 + GMM + HMM models were then identified for each image sequence individually. Timing measurements for mitotic phases were based on the labels obtained from TC3 + GMM + HMM.

Cell lines and RNAi. Experiments (**Fig. 2d–f**) were performed using a monoclonal reporter HeLa Kyoto cell line stably expressing H2B-mCherry[4]. Cells were maintained in Dulbecco's Modified Eagle Medium (DMEM, Gibco/Invitrogen) supplemented with 10% (v/v) FBS (PAA Laboratories), 50 units ml^{-1} of penicillin and streptomycin (Gibco/Invitrogen) and 200 μg ml^{-1} G418 (Gibco/Invitrogen) as a stable reporter's maintenance selection agent.

For live-cell imaging, 3,000 cells per well were reverse-transfected in liquid phase with Hiperfect reagent (Qiagen) using a final siRNA concentration of 10 nM in 96-well plastic-bottom imaging plates (Greiner Bio-One), following the manufacturer's instructions. Cells were grown in DMEM without phenol red or riboflavin (Gibco/Invitrogen), to reduce autofluorescence[4,16], and containing 10% (v/v) FCS and 50 units ml^{-1} of penicillin and streptomycin.

siRNAs (Ambion, Qiagen) were custom-synthesized using sequences identical to previously validated siRNAs as indicated in the cited references (**Supplementary Table 3**): MAD2L1 (ref. 17), CDC20 (ref. 7), separase/ESPL1 (ref. 18), INCENP[19], AURKB[20], CENP-E[21] and TPX2 (si_TPX2_6 validated Qiagen siRNA). The negative controls were as follows: siControl #1, Ambion NC1, cat. no. AM4635; siControl #2, Qiagen AllStars, cat. no. 1027280); and siControl #3, Microsynth NC .

Live-cell microscopy. Live-cell imaging was performed using an ImageXpress Micro screening microscope (Molecular Devices), equipped with a 10×, 0.5 numerical-aperture (NA) S-Fluor Dry Objective (Nikon) and a 5% CO_2, 37 °C humid-chamber stage incubator. A MetaMorph-based custom control macro, *PlateScan*, developed in house was used to control the microscope as described[7]. Light intensity was calibrated and maintained constant between experimental replicas to keep the cell death rate below 5% in siRNA-control transfected cells. Imaging started 16 h post-transfection for a period of 36 h with an average time lapse of 4.6 min.

Feature transformation and reduction. PCA[22] was employed to reduce dimensionality and to transform correlated original feature variables into uncorrelated ones. Normalization was performed by *z*-score standardization[23]. All the features up to 99% accumulative explained variance were kept and sequential feature tensors were obtained.

Temporally constrained combinatorial clustering. Temporally constrained combinatorial clustering (TC3) takes advantage of temporal constraints in cellular morphology dynamics extracted from single cell trajectories. No parameter tuning or model fitting is required to inspect the solution space or to satisfy the incorporated temporal constraints. TC3 can be used to cluster cyclical time series (for example, progression through an entire cell cycle) as well as time series of terminal processes (such as mitotic arrest).

The number of possible temporally constrained combinatorial clustering assignments of a time series with T frames into K non-empty, disjoint and sequential subsets (clusters) is specified by the following recurrence relation:

$$C(T,K)=\begin{cases} C(T-1,K-1)+C(T-1,K) & 1<K<T \\ 1 & K=1 \vee K=T \\ 0 & K>T \end{cases}$$

The full enumeration of all possible cluster assignments is given by the corresponding binomial coefficient

$$C(T,K)=\binom{T-1}{K-1}$$

The limited, sequential and periodic nature of cellular dynamics, as during the cycle, enables the procedure to find globally optimal solutions to initialize cellular morphology clusters. TC3's formal definition and mathematical properties are described in **Supplementary Note 1**.

Mitotic subgraph detection splits each cell trajectory into three disjoint sequential intervals. These are the pre-mitotic subgraph, the mitotic subgraph and the post-mitotic subgraph. Respectively, their lengths are T_1, T_2 and T_3. Summed together, they give the total length of the time series, which is $T = T_1 + T_2 + T_3 = 40$. This strategy breaks down the initial problem into subproblems, and TC3 is applied to each one. Consequently, the task of detecting $K = 6$ clusters (interphase and the five phases of mitosis) in a time series of length T corresponds to the partitioning of T_1 observations into two clusters (interphase and prophase) in the pre-mitotic subgraph, T_2 observations into $K - 2 = 4$ clusters (prometaphase, metaphase, anaphase and telophase) in the mitotic subgraph and T_3 observations into a single cluster (interphase) in the post-mitotic subgraph. Binary clustering is an optional processing step that can be omitted if computational resources are limited.

Rounded mean durations (± s.e.m.) are 10 ± 2 for the pre-mitotic subgraph, 13 ± 4 for the mitotic subgraph and 17 ± 3 for the post-mitotic subgraph. The average number of possible cluster assignments for a time series is $C(10, 2) + C(13, 4) + C(17, 1) = 230$. The search space is thus much smaller than that of assigning $T = 40$ frames into $K = 6$ clusters without detection of mitotic subgraphs. Such a clustering has a search space of cardinality $C(40, 6) = 575{,}757$. Because of the captured temporal structure of the task, it

doi:10.1038/nmeth.2046

is even smaller than the number of cluster assignments of classical combinatorial clustering[24,25], which is given by $S(40, 6) \approx 10^{28}$ (where $S(\cdot,\cdot)$ are Stirling numbers of the second kind). In the mitotic subgraph, domain knowledge can be incorporated by avoiding singleton clusters (for example, setting minimal cluster sizes to 2), thus further reducing the cardinality of the search space from $C(13, 4) = 220$ to $C((13 - 4), 4) = 56$ and enabling a feasible exhaustive search for the best cluster configuration that minimizes the within-cluster point scatter. After TC3, data cluster memberships are converted to class labels.

Modeling of cellular morphologies. Identification of predictive models for cells progressing through mitosis is based on TC3 labels. Continuous HMM (CHMM) as well as discrete HMM (DHMM)[26] are inferred. Each observation density of CHMM is specified by the respective component distribution of GMM. Multivariate features are vector quantized by GMM into discrete symbols for DHMM. TC3 is employed for the initialization of sample means and covariance matrices of the respective GMM components. The HMM of the method TC3 + GMM + HMM in the main text refers to DHMM.

Number of clusters. For the benchmark data, the number of clusters is given as the input of the algorithm. When a priori knowledge is unattainable, the optimal cluster number can be determined either by visual inspection of labeling schemes with a variable number of clusters (**Supplementary Fig. 7**) or by approximation set coding, a general information-theoretic principle for model validation[27].

Hierarchical clustering of RNAi phenotypes. A total of 20 morphology and timing features are extracted from TC3 + GMM + HMM analysis of RNAi screening data. The eight temporal features are mean timing of mitotic subgraph phases and the associated s.e.m. values. The 12 spatial features are Gaussian component means of respective subgraph phases of the first three dimensions. Hierarchical clustering is based on Euclidean distance and complete linkage.

Implementation. The introduced method was coded in MATLAB (R2010a) and based on MATLAB built-in functions such as *K*-means, GMM and HMM. Kevin Murphy's toolbox was employed for CHMMs (http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html). Benchmark data was generated by the CellCognition software[7]. The open source software LIBSVM[28] was used to train and test SVM classifiers on the same benchmark data. The source code and data for the method presented in this manuscript are provided in the **Supplementary Software** and published online at http://www.cellcognition.org/software/tc3/.

16. Schmitz, M.H. & Gerlich, D.W. *Methods Mol. Biol.* **545**, 113–134 (2009).
17. Martin-Lluesma, S., Stucke, V.M. & Nigg, E.A. *Science* **297**, 2267–2270 (2002).
18. Waizenegger, I., Gimenez-Abian, J.F., Wernic, D. & Peters, J.M. *Curr. Biol.* **12**, 1368–1378 (2002).
19. Neumann, B. *et al. Nat. Methods* **3**, 385–390 (2006).
20. Qi, W., Tang, Z. & Yu, H. *Mol. Biol. Cell* **17**, 3705–3716 (2006).
21. Thoma, C.R. *et al. Nat. Cell Biol.* **11**, 994–1001 (2009).
22. Pearson, K. *Philos. Mag.* **2**, 559–572 (1901).
23. Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference* (Springer, 2003).
24. Jain, A.K. & Dubes, R.C. *Algorithms for Clustering Data* (Prentice Hall, 1988).
25. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, 2001).
26. Rabiner, L. *Proc. IEEE* **77**, 257–286 (1989).
27. Chehreghani, M.H., Busetto, A.G. & Buhmann, J.M. *J. Mach. Learn. Res.* **22**, 495–503 (2012).
28. Chang, C.C. & Lin, C.J. *ACM Trans. Intelligent Syst. Technol.* **2** (2011).