31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

# Cell Cycle Dependence of Protein Subcellular Location Inferred from Static, Asynchronous Images

Taráz E. Buck, Arvind Rao, Luís Pedro Coelho, Margaret H. Fuhrman, Jonathan W. Jarvik, Peter B. Berget, and Robert F. Murphy, *Senior Member, IEEE*

***Abstract*—Protein subcellular location is one of the most important determinants of protein function during cellular processes. Changes in protein behavior during the cell cycle are expected to be involved in cellular reprogramming during disease and development, and there is therefore a critical need to understand cell-cycle dependent variation in protein localization which may be related to aberrant pathway activity. With this goal, it would be useful to have an automated method that can be applied on a proteomic scale to identify candidate proteins showing cell-cycle dependent variation of location. Fluorescence microscopy, and especially automated, high-throughput microscopy, can provide images for tens of thousands of fluorescently-tagged proteins for this purpose. Previous work on analysis of cell cycle variation has traditionally relied on obtaining time-series images over an entire cell cycle; these methods are not applicable to the single time point images that are much easier to obtain on a large scale. Hence a method that can infer cell cycle-dependence of proteins from asynchronous, static cell images would be preferable. In this work, we demonstrate such a method that can associate protein pattern variation in static images with cell cycle progression. We additionally show that a one-dimensional parameterization of cell cycle progression and protein feature pattern is sufficient to infer association between localization and cell cycle.**

## I. INTRODUCTION

The study of subcellular location via imaging is a critical aspect of proteomics that complements studies of sequence, structure, binding interactions, and biochemical activity. Automated determination of protein subcellular localization from microscope images has not only been demonstrated to be feasible for the major organelles [1] but can outperform visual analysis [2]. Protein location varies with numerous factors including cell type, microenvironment, treatment conditions and time. Temporal effects can occur in many places and at many scales, from the millisecond to the day, but one of the most obvious and important temporal processes is the cell cycle. Many proteins interact in orchestrating growth, DNA replication, and cellular division.

The problem of identifying cell-cycle dependent variation in protein localization has been a significant focus of previous work [3-5]. As aberrations in protein localization are invariably related to reprogrammed cell behavior, determining changes in trafficking of proteins through various organelles during the cell cycle can aid understanding of the dynamics of disease and development. An automated method to identify those proteins that might potentially exhibit a cell-cycle dependent localization would be a very useful prospective tool for detailed further investigation of their role in various biological processes.

Previous work examining the cell cycle dependence of protein location usually (1) discretizes the cell cycle into a set of phases (e.g., G0/G1, S, G2, M) or (2) artificially synchronizes the cells under examination; both methods attempt thereby to boost correlative effects observed. Sigal et al. 2006 [3] addressed these limitations by capturing time-lapse images and synchronizing them in silico (i.e., aligning profiles of nuclear intensity of different cells across time). However, time-lapse images can be more difficult to obtain than single images of cells because many microscopes do not maintain a viable environment for the cells they image (e.g., cells die after some time, and even while alive they are not under constant conditions). Furthermore, repeated excitation of dyes for fluorescence imaging causes photobleaching, reducing signal and leading to toxic chemical changes (phototoxicity), further perturbing cells. Lower exposure times reduce these effects but attenuate signal. Time-series images have another limitation: imaging more cells means the microscope takes longer between frames to revisit a particular cell, potentially compromising cell tracking algorithms. A method using unsynchronized cells with single-image capture would have the advantages of avoiding repeated exposure to fluorescence excitation (permitting higher-energy exposure to obtain better signal) and fewer environment viability requirements.

Thus, when imaging proteins in an asynchronous population of cells at a single time point, there is a need to resolve which proteins show a dependence on the cell cycle and which proteins are static across the cell cycle. This paper proposes a method to infer the association between protein location patterns in unsynchronized static cell images and cell cycle progression in an unsupervised manner, i.e., without explicit knowledge of the cell cycle stage for a particular cell.

In this work, we consider images of cells, specifically of

Manuscript received April 23, 2009. This work was funded in part by NIH grant GM075205 [to R.F.M.].

T. E. Buck, L. P. Coelho, and R. F. Murphy are with the Joint Carnegie Mellon University-University of Pittsburgh Ph.D. Program in Computational Biology and the Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213 USA (phone: 412-268-3480; FAX: 412-268-2977; e-mail: murphy@cmu.edu). A. Rao, P.B. Berget and J. W. Jarvik are also with the Lane Center.

M. H. Fuhrman, P. B. Berget, and J. W. Jarvik are with the Department of Biological Sciences, Carnegie Mellon University.

R. F. Murphy is also with the Departments of Biological Sciences, Biomedical Engineering, and Machine Learning at Carnegie Mellon.

their nuclei and of the distribution of a particular tagged protein. Using certain statistics computed on the nuclear image ("nuclear features") as a representation of cell-cycle phase, we infer a one-dimensional statistical manifold (parameterized by $\gamma_1$) for progression in cell cycle. Observing its relationship with features extracted from protein images allows us to identify those protein image features that correlate strongly with cell-cycle progression. The subspace of all such protein features uniquely identifies another statistical manifold along which proteins may show a variation in subcellular localization (which may or may not be associated with the cell cycle). We further demonstrate that variation in the protein distribution due to the cell cycle can be detected and used to rank proteins by how much they vary in this manner. We conclude that this is a feasible task and discuss possible improvements.

## II. Methods

### A. Image Dataset

We used two datasets for our experiments. The first is a single time-series of images of HeLa cells expressing RFP-labeled histone H2B as described previously [6]. Images were taken every half hour with a fixed exposure time, and environmental conditions were kept stable at 37°C and 5% $CO_2$. This dataset was used for validating our proposed method. The second data set consists of single exposures of unsynchronized NIH 3T3 cells expressing fluorescently-tagged proteins, collected as described previously [7]. Our RandTag project generates and images thousands of clones that are CD-tagged to express different GFP fusion proteins under native regulation [8]. We used images for sixteen of these clones in this paper. For each image, DNA was labeled using the viable dye Hoechst 33342. Images were captured using an IC-100 microscope with a 40X objective and a resolution of 0.1613 μm/pixel.

### B. Image Processing

Time-series images were processed as follows. Segmentation and tracking of nuclei were performed as in [6]. Background was removed by subtracting the modal pixel value of all pixels below the mean pixel value for the image. Images were divided by the 95th percentile of pixel intensities from inside nuclear regions, in order to normalize nuclei across images. As fewer than 5% of the nuclei and thus nuclear pixels at any given time had condensed their DNA for mitosis, the 95th percentile should be near the maximum intensity of interphase nuclei. Further computation only included images of nuclei if the rest of each nucleus' cell cycle was also available (mother cell's cytokinesis to next cytokinesis).

Static images were filtered for meaningful signal as follows. Background was removed from both the nuclear and protein channels by the same method as above. An image was removed if its maximum intensity (after background subtraction) was less than 30 in both the nuclear and protein channels (manually selected). Clones for which no images passed this threshold were ignored.

Static images were segmented into individual cell regions as follows. First, the unprocessed nuclear channels were normalized to [0, 1]. A seeded watershed algorithm was used to segment the image into separate nuclei. Regional maxima of the h-maxima transform, which suppresses maxima smaller than some threshold, were used as seeds (using a manually selected threshold of seven times the first quartile of the Gaussian-filtered channel). The watershed surface was the difference of Gaussian-filtered versions of the channel (with standard deviations of the minimum nuclear diameter and half the minimum, set to 5 μm; the former was also morphologically dilated by a disk half the minimum diameter to adjust the edges). A background seed consisting of the border pixels of the image as well as any seeds touching the border was used to ensure compact segmentation of the nuclei. Seeds were then imposed as minima in the watershed surface by morphological reconstruction. Matlab's Image Processing Toolbox was used for most of these operations.

Cellular regions were similarly decided by seeded watershed. Seeds were the nuclei found as above (including the same background seed to prevent inclusion of protein from border cells into the regions of cells of interest). The watershed surface was a Gaussian-filtered version of the unprocessed protein channel (standard deviation of a tenth of the maximum nuclear diameter, 25 μm), also with minima imposed by the seeds.

### C. Feature Extraction

Subcellular Location Features: We have previously described several sets of features for describing protein patterns in fluorescence micrographs and demonstrated that these provide high accuracy for various purposes [1]. We therefore began with the SLF7 set [2], which consists of 84 features including edge, morphological, Haralick texture, and DNA correlation features. To this we added two additional feature sets. The first was a set of 30 wavelet features consisting of the root sum of squares of the detail channels for a 10-level Daubechie-4 wavelet decomposition. The second (to further enhance characterization of textures at different scales), was a set of 13 Haralick texture features for the protein images spatially downsampled by factors of 2, 4 and 8 (giving 39 features). Thus, protein patterns were described by a total of 153 features.

Nuclear features: After binarizing the DNA image to obtain nuclear shapes, we extracted features to represent nuclear appearance. Features include total, minimum, mean, standard deviation of, and maximum intensity, area, perimeter, long, short, and ratio of medial axes, and Haralick texture features. Haralick features were computed on the original nuclei and three lower resolutions obtained by downsampling by factors of two. Haralick features were averaged across horizontal, vertical, and diagonal directions after quantizing the images to eight gray levels. This resulted in a total of 62 features per nucleus.

The intermediate goal is to obtain a scalar field parameterization of this 62-dimensional feature space so that we could study the relationship between cell-cycle stage and its natural parametric progression. As will become clear below, such a parameterization permits the exploration of a possible association between each protein-pattern variation and cell-cycle stage. Isomap manifold embedding is performed for dimension reduction from the feature space (62-D) to a scalar field ($\gamma_1$); this approximately preserves the geometry of the feature space and allows $\gamma_1$ to act as a surrogate for cell cycle phase. A traversal along this scalar field correlates with a corresponding variation in intensity or nuclear area by construction.

### D. Manifold Embedding

The manifold embedding problem is defined as follows: Given data in a high dimensional space (possibly generated from a low dimensional manifold), attempt to recover the underlying low-dimensional structure of data embedded in the high-dimensional space. Isomap [10] is a technique that is used to model the intrinsic geometry of a high-dimensional space using only distances between all pairs of data points. It has three main steps.

First, a nearest-neighbor graph is constructed (we chose to use local determination of dimensionality and tangent space for this construction [11]). Each edge is assigned the weight of the Euclidean distance between its two points. Second, a pairwise geodesic distance matrix is formed from the weight of the shortest path between each pair of vertices. Third, multidimensional scaling applied on the geodesic distance matrix finds the final embedding at a specified dimensionality. Isomap's outputs, the embedding coordinates for the input data points, are returned in order of greatest variance explained, and progressively lower dimensional manifolds omit more of these later coordinates (that is, the target dimensionality of the manifold does not affect the values of the embedding coordinates).

Manifold coordinates for data points not used to compute the manifold are estimated using a modified version of Isomap's coordinate determination method (multidimensional scaling [12].

For time-series data, the manifold was built using half of the training data as input to Isomap, half of which served as landmarks (using a version of Isomap that saves memory and computation time by only preserving distances of all data points to the set of landmarks).

For static images, the 62 nuclear features were given as input to Isomap. The first dimension of the resulting embedding coordinates was taken as a one dimensional manifold and termed the cell cycle parameter.

### E. Regression

The relationship between protein features and $\gamma_1$ was modeled using stepwise polynomial regression. Each protein feature and its powers from two to eight became candidate predictors for $\gamma_1$ to model possible nonlinear relationships. Stepwise regression was used to select a subset of the candidate predictors in order to minimize the number of predictors not contributing improvements to the model. The method of stepwise regression is an iterative heuristic procedure to select the best predictors of the dependent variable that, for each iteration, adds a feature that improves prediction compared with current features, removes one that does not decrease prediction by being eliminated, or exits when neither happens. The criteria of addition or rejection are F-tests below or above specified threshold, respectively.

Stepwise regression was also used to model and check how well the manifold coordinates found on the time-series data correlate with actual time. Time was defined as the number of frames since an individual cell's cytokinesis from its sister cell divided by the total number of frames before the cell divided.

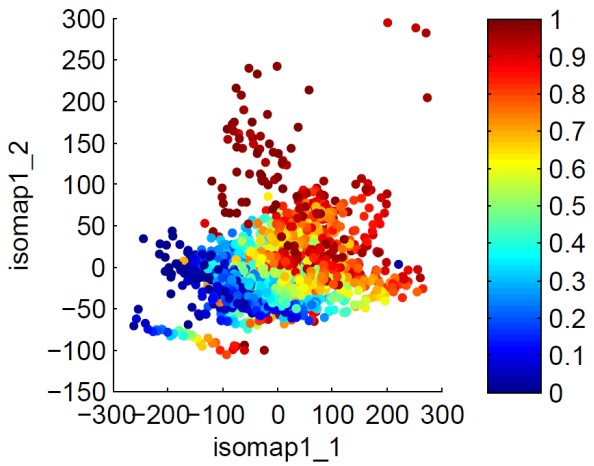


Fig. 1. Relationship between manifold learned on nuclear features of the time-series data and actual cell cycle time. The horizontal axis is first manifold coordinate, and the vertical axis is second. Color indicates fractional time since cytokinesis as shown in the color bar.

## III. RESULTS

### A. Time-Series Evaluation of the Cell Cycle Parameter

We began by determining whether a cell cycle parameter learned from nuclear features could adequately predict the actual time of each frame in a time-series image. Figure 1 shows the correlation between the nuclear manifold learned from time-series data and actual cell cycle time. Cell cycle time clearly progresses in a non-random fashion across the manifold. Using stepwise polynomial regression to regress cell cycle time against the two coordinates, a testing adjusted R-square of 0.70 is achieved (raw nuclear features as predictors produce an R-square of 0.74), indicating that the manifold embedding quite reliably approximates the original geometry of the actual hyperspace, including changes according to time.

### B. Predicting the Cell Cycle Parameter for Static Protein Images

In order to predict the cell cycle parameter for images of randomly-tagged cell clones, we applied the above methods to 16 clones in two combinations: The protein distribution

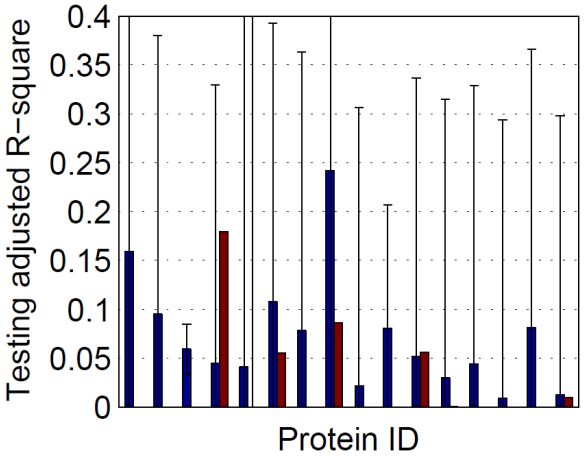


Fig. 2. Cell cycle parameter predictions are grouped by tagged clone (horizontal axis, each pair of blue and red bars). Error bars are standard deviation. Raw protein features (left bar in pairs) predict cell cycle parameter $\gamma_1$ with a greater testing adjusted R-square (vertical axis) than the first 9 dimensions of an Isomap embedding of the same protein features (right bar). However, the Isomap embedding produces reduced-variance estimates across cross-validation folds.

was represented as either the original 153 SLF features or those features reduced by Isomap to a 9-dimensional manifold. As a test of how well variation in protein pattern was correlated with our estimate cell cycle positions, we determined how well the protein features could be used as regression predictors of the cell cycle parameter. Statistics are averages computed by cross-validation. The level of correlation was measured by the testing adjusted R-square.

In Fig. 2, the two tests described above are grouped by protein. The original feature set tended to better predict the cell cycle parameter, while lower variance in estimation of the testing adjusted R-square was observed after Isomap-based dimensionality reduction. Images for various cells sorted by cell cycle parameter for one of these proteins (Trim24) are shown in Fig. 3.

## IV. Conclusion

We have presented a system for inferring correlation of subcellular protein distribution with cell cycle time from unsynchronized images of cells using a one dimensional manifold computed on simple nuclear image features. The cell cycle parameter ($\gamma_1$) can be tested for ability to be predicted on a per-protein basis from protein image features. This relationship provides a way to screen proteins for dependence of their localization on the cell cycle using only static, asynchronous images. Future work will include modifying the cell cycle learning method to incorporate prior knowledge from time-series data, examination of generalizability to other cell lines and nuclear tagging, and comparison of results to curated information regarding cell cycle variation in protein localization.

## Acknowledgment

The authors thank Drs. Xiaobo Zhou and Stephen Wong for providing time-series images, Dr. Elvira Osuna Highley for helpful discussions, Jimmy Xu, Bur Chu, and Charlotte Chou for image acquisition, and Armaghan Naik for critical reading of the manuscript.

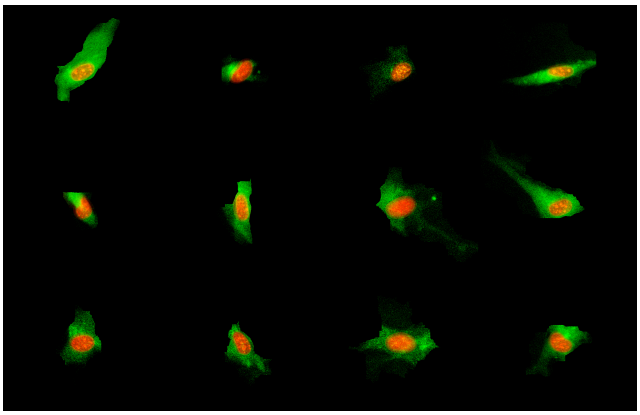

Fig. 3. Images of Trim24 ordered by $\gamma_1$. $\gamma_1$ progresses from left to right, then top to bottom. Trim24 is the second protein from the left in Fig. 2.

## References

[1] M. V. Boland and R. F. Murphy, "A Neural Network Classifier Capable of Recognizing the Patterns of all Major Subcellular Structures in Fluorescence Microscope Images of HeLa Cells," *Bioinformatics*, vol. 17, no. 12, pp. 1213-1223, Dec, 2001.

[2] R. F. Murphy, M. Velliste, and G. Porreca, "Robust Numerical Features for Description and Classification of Subcellular Location Patterns in Fluorescence Microscope Images," *J. VLSI Sig. Proc.,* vol. 35, no. 3, pp. 311-321, November, 2003.

[3] A. Sigal , R. Milo, A. Cohen, N. Geva-Zatorsky, Y. Klein, I. Alaluf, N. Swerdlin, N. Perzov, T. Danon, Y. Liron, T. Raveh, A. E. Carpenter, G. Lahav, and U. Alon, "Dynamic proteomics in individual human cells uncovers widespread cell-cycle dependence of nuclear proteins," *Nature Methods,* vol. 3, no. 7, pp. 525-531, July, 2006.

[4] M. D. Gooden, R. B. Vernon, J.A. Bassuk, and E. H. Sage, "Cell cycle-dependent nuclear location of the matricellular protein SPARC:association with the nuclear matrix," *J. Cell Biochem*., vol. 74, no. 2, pp. 152-67, August, 1999.

[5] M. Miura, H. Watanabe, T. Sasaki, K. Tatsumi, and M. Muto, "Dynamic changes in subnuclear NP95 location during the cell cycle and its spatial relationship with DNA replication foci," *Exp. Cell Res.*, vol. 263, no. 2, pp. 202-8, February, 2001.

[6] X. Zhou, F. Li, J. Yan, S. T. C. Wong, "A Novel Cell Segmentation Method and Cell Phase Identification Using Markov Model," *IEEE Trans. Inf. Tech. Biomed.*, vol. 13, vo. 2, pp. 152-157, March 2009.

[7] E. García Osuna, J. Hua, N.W. Bateman, T. Zhao, P.B. Berget and R.F. Murphy, "Large-Scale Automated Analysis of Location Patterns in Randomly Tagged 3T3 Cells," *Ann. Biomed. Eng.,* vol. 35, no. 6, pp. 1081–1087, June, 2007.

[8] J. W. Jarvik, S. A. Adler, C. A. Telmer, V. Subramaniam, and A. Lopez, "Cd-tagging: A new approach to gene and protein discovery and analysis," *BioTechniques*, vol. 20, no. 5, pp. 896–904, May, 1996.

[9] M. D. Hubar and L. Gerace, "The size-wise nucleus: nuclear volume control in eukaryotes," *J. Cell Biol.,* vol. 179, no. 4, pp. 583–584, November, 2007.

[10] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, vol. 290, no. 5500, 2319–2322, Dec, 2000.

[11] N. Mekuz and J. K. Tsotsos, "Parameterless Isomap with Adaptive Neighborhood Selection," *Lect. Notes Comp. Sci.*, vol. 4174, , pp. 364-373, Sep, 2006.

[12] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet, "Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering," In *Advances in Neural Information Processing Systems*, vol. 16, S. Thrun, L. Saul, and B. Schölkopf, Eds., Cambridge, MA: MIT Press, 2003, pp. 177-184