

A GENOME WIDE RNAI SCREEN BY TIME LAPSE MICROSCOPY IN ORDER TO IDENTIFY MITOTIC GENES — COMPUTATIONAL ASPECTS AND CHALLENGES

Thomas Walter^{1*}, Michael Held^{2*}, Beate Neumann¹,
Jean-Karim Hériché³, Christian Conrad¹, Rainer Pepperkok¹ and Jan Ellenberg¹

¹European Molecular Biology Laboratory, Meyerhofstraße 1, 69117 Heidelberg, Germany

²Institute of Biochemistry, ETH Zürich, 8093 Zürich, Switzerland

³Wellcome Trust Sanger Institute, Hinxton CB10 1HH, U.K.

*These authors have equally contributed to the presented work.

ABSTRACT

The MitoCheck project aims at identifying and characterizing the function of genes involved in mitosis in live human cells. The genome wide RNA interference screen developed for this purpose is based on automatic time-lapse microscopy to analyze the phenotypes after knocking down each human gene individually in cultured cells whose chromosomes are fluorescently labeled. Such a screen produces large amounts of digital image data (~ 200.000 video sequences, i.e. over 18 millions of images) which can no longer be handled and interpreted manually.

We have developed an image processing method consisting of segmentation, feature extraction and automatic classification, which assigns to each nucleus in each image one out of several predefined morphological classes. Using the relative cell counts in each of these classes, measured over time for each experiment, we derive a phenotypic fingerprint for each gene that allows clustering of genes by functional similarity.

This paper will give an overview over the computational aspects of this screen. The complete quality controlled data set and phenotypic measurements will be available after publication on <http://www.mitocheck.org/>.

Index Terms— Genome-wide RNAi Screen, Mitosis, Biomedical Image Processing, Pattern Recognition

1. INTRODUCTION

One of the most challenging tasks in current molecular biology is the functional description of genes and proteins on a large scale in order to better understand the molecular regulation of biological processes. An important element in this quest is the use of loss-of-function screens, and in particular RNAi screens. In these screens, genes are characterized by the phenotype resulting from their downregulation. By

The authors would like to acknowledge funding within the MitoCheck consortium by the European Commission FP6

means of fluorescence microscopy, the functional role of the downregulated genes can be studied in their physiological environment with high spatial and temporal resolution. Thanks to recent technical developments in automated microscopy and high-throughput transfection methods, image based assays are now applicable on a genome-wide scale [1].

Most of the microscopy based screens published so far are endpoint assays. Depending on the biological question, this can lead to misinterpreted phenotypes: as in a genome-wide context, phenotypes are normally not predictable in time, it is difficult to state whether an observed phenotype is actually the primary cause or already a consequence of an earlier and not observed phenotype. This limitation can be overcome by using the power of time-resolved live cell imaging [2].

This strategy has been applied to perform a genome-wide RNAi screen in human cells by time-lapse imaging: siRNA transfection mixes have been spotted onto chambered coverglass tissue culture dishes. 18 hours before imaging, HeLa cells stably expressing histone-GFP to report on chromosome segregation and structure, are seeded on top of these arrays and imaged for 48 hours (see [2] and [3] for experimental details). ~ 22.000 protein coding genes have been targeted by at least 2 siRNA each, and for each siRNA, there are at least 3 replicates. The readout of the screen is a data set of ~ 200.000 video sequences over 48 hours, yielding ~ 18 million images ($\sim 30TB$). Obviously, it is neither possible nor desirable to analyze this amount of data manually, but even methods for automatic off-line analysis of this data set must meet strict requirements concerning computational speed: 1s of processing time per image leads to a total processing time of ~ 200 days on a single machine; even with heavy use of parallel computing, this rules out time consuming segmentation or machine learning approaches.

We have developed a method for the automatic analysis of this data set based on automatic recognition of nuclear morphologies. Even though the fundamental biological question we wanted to answer with this screen concerned mitosis,

we can analyze the data set from a different point of view in order to also identify genes involved in other important biological processes, like cell migration. The screen therefore is not only the first genome-wide screen for mitosis by time-lapse imaging in a human cell line, but additionally supplies us with time-resolved phenotypic description for the entire genome. The data will be available after publication on <http://www.mitocheck.org/>.

2. AUTOMATIC IDENTIFICATION OF MITOTIC PHENOTYPES

In the following, we will call each video sequence an *experiment*. For each siRNA (i.e. each downregulated gene) we have at least 3 experiments. The algorithm we have developed in order to identify mitotic phenotypes is based on a classical object recognition pipeline applied to every single image:

1. Images are segmented, i.e. the single nuclei are detected.
2. For each nucleus, features are extracted in order to describe the shape and the texture of the objects.
3. These features are then used in order to classify each nucleus into one out of several predefined morphological classes.
4. Each experiment is characterized by a time series of nuclei counts in all different morphological classes. From this representation, we can derive a score for each experiment.

A detailed description of each of these steps would go over the scope of this paper. We will therefore just give brief descriptions of the developed methods without discussion of the results.

2.1. Segmentation

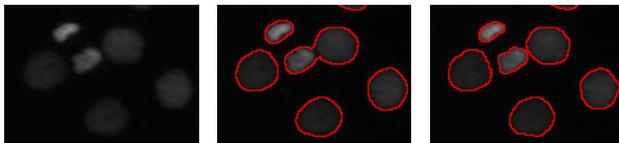


Fig. 1. Segmentation of nuclei: (a) Original image, (b) Application of a local threshold without sharpening, (c) Application of a local threshold with sharpening (toggle-mappings)

The segmentation step aims at identifying the nuclei. As we can see from figure 1(a), the segmentation task is not very difficult: nuclei appear as bright objects on dark background. As the background is not always uniform, some background

approximation must be subtracted from the image before a global threshold is applied. Moreover, close nuclei tend to be segmented as one nucleus due to the blurring of the optical system. As a consequence, we propose to apply a sharpening algorithm (toggle-mappings) in order to separate close nuclei from each other (see figure 1(b) and 1(c)).

As a result of the segmentation step, we obtain a partition of the image plane into disjoint regions, one of which representing the background, and all others the nuclei. We denote the set of foreground pixels as $X = \bigcup_{i=1 \dots N} X_i$, where X_i are the connected components of X , i.e. each X_i corresponds to one nucleus.

2.2. Feature extraction

After the automatic detection of the nuclei, the next step is to characterize them by means of quantitative descriptors. Many features have been proposed in the literature for similar recognition problems [4], [5], but it is often necessary to also use some dedicated features capable of differentiating between visually close classes.

Features can be divided into two groups: shape features describing the geometric properties of X_i and texture features describing the spatial grey level distribution of the image on the regions X_i . From a biological point of view, shape features are supposed to be able to distinguish between normal and abnormal interphases (e.g. multinuclear morphologies), whereas texture features should be able to characterize classes where chromatin is condensed (metaphase, prometaphase, etc.).

As texture features, we have used basic features (like mean intensity, grey level variation, etc.), Haralick features [6], statistical geometric features [7] and features based on mathematical morphology. As shape features, we have used basic descriptors (like size, perimeter, etc.), moment based features [8], features based on mathematical morphology and convex hull features.

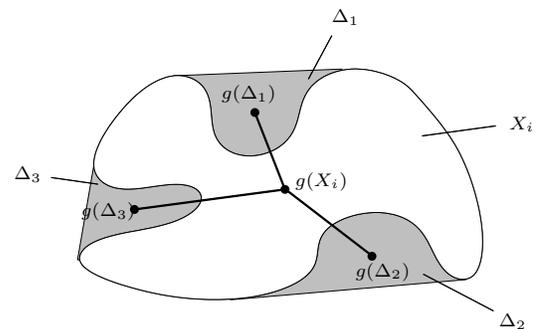


Fig. 2. Features defined for the convex hull of the nuclear region X_i

To illustrate how these features are designed, we will de-

scribe this last group of features. The normal nuclear morphology of an interphase cell has an elliptic 2D projection and appears therefore convex in our images; aberrant morphologies often observed as a consequence of a mitotic defect (e.g. problems in chromosome segregation) mostly show concavities, which can be characterized by features defined on the convex hull $C(X_i)$, i.e. the smallest convex set containing X_i . Let $\Delta = C(X_i) \setminus X_i$ and $\Delta_k, k = 1, \dots, N_c$ its connected components (see figure 2). We define as features:

- the number of connected components N_c .
- the ratio of the surface of X_i to the surface of $C(X_i)$.
- the ratio of the perimeter of X_i to the perimeter of $C(X_i)$.
- the 3 maximal areas of the connected components of Δ . These values give a hint of how symmetric the concavities are.
- average distance of Δ_k to the center of X_i : $\sum_{k=1}^{N_c} \|g(X_i) - g(\Delta_k)\| \frac{\#\Delta_k}{\#\Delta}$, where $g(\cdot)$ denotes the center of gravity (see figure 2) and $\#\cdot$ the number of elements.
- the average area of Δ_k .

This set of features is well suited for the distinction between normal and abnormal morphologies and even describes - to a certain extent - different kinds of abnormalities, relevant for different mitotic or non-mitotic defects.

In total, we have extracted 190 features and have therefore obtained a detailed quantitative description of each nucleus. How these features can then be used in order to assign one out of several predefined morphological classes to each nucleus will be shown in the next section.

2.3. Classification

For the classification step, we have followed a classical supervised learning approach: first, we have defined morphological classes we expect to cover the whole genome-wide data set. Then, we have generated a training set, i.e. a set of labeled nuclei for each of these classes and covering their range of variability. Finally we have used this training set in order to train an SVM classifier.

2.3.1. Class definitions

The most difficult part of this procedure is the definition of meaningful classes covering all possible morphologies in the genome-wide data set, and we are not yet able to present an efficient solution to this problem. It is of course an easy task to define wildtype morphologies (e.g. the morphological classes corresponding to mitotic phases), but unfortunately, this is not appropriate if we are dealing with a genome-wide

data set: we are expecting many more different and possibly unknown morphologies. One way to derive a set of morphological classes is to manually look at a set of experiments and to determine manually the biologically meaningful morphologies. This set of experiments should contain positive and negative controls, genes known from literature and, most importantly, genes for which the proliferation rate was particularly low. Obviously, there is no formal proof that the morphologies we have found in this way are really representative for the whole data set, but it seems a reasonable working hypothesis.

On the other hand, a more straight forward and much more efficient method would be the application of unsupervised learning strategies in order to derive all important morphological classes from the data set. However, our data set consists of ~ 1 billion feature vectors (nuclei), containing 190 features each. Such amounts of data are difficult to handle by classical clustering methods, and over and above that, it is far from being sure to obtain biologically meaningful classes by an unsupervised approach: biological and visual proximity do not always coincide.

We think that the development of methods able to derive morphological classes directly from large data sets will be a major contribution to image based screens: as the morphologies change with the assay (with the marker, cell line, etc.), the definition of a set of meaningful and detectable classes will be an issue also for upcoming screens, and probably be the major computational bottleneck as long as no useful method is available.

2.3.2. Classification with support vector machines

The last step is now to assign to each nucleus represented by the feature vector x one of the predefined classes.

Support Vector Machines (SVM) are one of the most powerful methods in the domain of supervised learning; they have outperformed many other methods in many applications, they can deal with large number of features, are robust, easy to parameterize and computationally effective. SVMs have been used successfully to a number of similar problems, like protein localization ([4], [5]), and the automatic identification of mitotic phases ([2], [9]).

We have used an *RBF* kernel in order to transform the feature vectors into a higher dimensional space before applying a linear classifier. The parameters of the *SVM* have been obtained by the classical grid search strategy. Our training set contained 2957 manually labeled nuclei; we obtained an overall accuracy of 86.1% with 10-fold cross validation.

2.4. Detection of mitotic phenotypes

After classification, each experiment is represented by a set of cell count kinetics. After a smoothing step, these time series can be compared to the corresponding time series of the negative controls: the maximal difference over time between the experimental curve for each morphological class

and the corresponding average curve of the negative controls is a good indicator for the penetrance of the phenotype in the given class.

In order to obtain a score for each siRNA rather than for each single experiment, we take the median of all replicates with the same siRNA. If in a morphological class relevant to the biological process under study, this value is higher than a certain manually fixed threshold we call the siRNA a hit in this class.

In addition to the information if an siRNA is actually giving a hit with respect to the biological question under study, we obtain a phenotypic fingerprint by taking the maximal penetrances in the different classes as descriptive features. Such fingerprints allow one to cluster siRNAs into groups of similar phenotypic characteristics.

3. SCREENING FOR NUCLEAR MOBILITY

In addition to mitosis, this data set contains useful information about other biological processes. By a more thorough exploitation of the time information for instance, it is possible to derive a list of genes involved in cell migration.

For this, we track each individual nucleus, i.e. we identify for each nucleus at time point t its predecessor (i.e. the same nucleus at time point $t - 1$). The euclidean distance between the centers of gravity of the same nucleus at t and $t - 1$ can be seen as a mobility measure for this nucleus. By taking the average mobility measure over all time points and all nuclei, we can easily define a mobility measure for the experiment.

Movement can be caused by many phenomena; dead cells for instance tend to lose adhesion resulting in a high mobility. Therefore we take into account only uncondensed nuclei, i.e. nuclei belonging to morphological classes where the DNA is not condensed (e.g. interphase). After subtracting a correction term in order to compensate for microscope positioning problems, we can easily derive a list of genes resulting in increased nuclear mobility. A more detailed analysis of the total trajectories during the whole experiment leads to a list of genes involved in cell migration.

4. CONCLUSION AND PERSPECTIVES

In this paper, we have presented the computational aspects of a genome-wide RNAi screen by time-lapse microscopy: we have shown how we can derive a phenotypic fingerprint for each of the targeted genes by application of image processing and machine learning methods. The most time-consuming steps of this strategy are the identification of a set of predefined classes and the generation of the corresponding training set. The development of unsupervised or semi-supervised methods able to generate these sets automatically or at least to ease their establishment seem to us the most challenging and important future steps.

Furthermore, we see that the same data set supplies us with more information than just that related to mitosis: we have shown a sketch of an algorithm dedicated to the identification of genes with increased nuclear mobility.

5. REFERENCES

- [1] Rainer Pepperkok and Jan Ellenberg, "High-throughput fluorescence microscopy for systems biology," *Nature Reviews Molecular Cell Biology*, vol. 7, no. 9, pp. 690–696, September 2006.
- [2] Beate Neumann, Michael Held, Urban Liebel, Holger Erfle, Phill Rogers, Rainer Pepperkok, and Jan Ellenberg, "High-throughput rna screening by time-lapse imaging of live human cells," *Nature Methods*, vol. 3, pp. 385–390, 2006.
- [3] Holger Erfle, Beate Neumann, Urban Liebel, Phill Rogers, Michael Held, Thomas Walter, Jan Ellenberg, and Rainer Pepperkok, "Reverse transfection on cell arrays for high content screening microscopy.," *Nat Protoc*, vol. 2, no. 2, pp. 392–399, 2007.
- [4] Robert F Murphy, Meel Velliste, and Gregory Porreca, "Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images," *Journal of VLSI Signal Processing*, vol. 35, pp. 311–321, 2003.
- [5] Christian Conrad, Holger Erfle, Patrick Warnat, Nathalie Daigle, Thomas Lörch, Jan Ellenberg, Rainer Pepperkok, and Roland Eils, "Automatic identification of subcellular phenotypes on human cell arrays," *Genome Research*, vol. 14, pp. 1130–1136, 2004.
- [6] R. M. Haralick, Dinstein, and K. Shanmugam, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, pp. 610–621, November 1973.
- [7] Ross F Walker and Paul Jackway, "Statistical geometric features - extensions for cytological texture analysis," in *ICPR - International Conference on Pattern Recognition*, 1996.
- [8] Richard J. Prokop and Anthony P. Reeves, "A survey of moment-based techniques for unoccluded object representation and recognition," *CVGIP: Graphical Models and Image Processing*, vol. 54, no. 5, pp. 438–460, 1992.
- [9] Meng Wang, Xiaobo Zhou, Fuhai Li, Jeremy Huckins, Randall W King, and Stephen T C Wong, "Novel cell segmentation and online svm for cell cycle phase identification in automated microscopy.," *Bioinformatics*, vol. 24, no. 1, pp. 94–101, Jan 2008.