# An Image Informatics Method for Automated Quantitative Analysis of Phenotype Visual Similarities

Lior Shamir*, D. Mark Eckley, John Delaney,
Nikita Orlov, Ilya G. Goldberg
Laboratory of Genetics
National Institute on Aging, NIH
251 Bayview Boulevard, Baltimore, MD 21224
shamirl@mail.nih.gov

***Abstract*— The post genomic era introduced the need to define single gene functions within biological pathways. A systems biology approach can be realized by automating image acquisition and phenotype classification. While machinery for automated data acquisition have been developing rapidly in the past years, the main bottleneck remains the effectiveness of the computer vision algorithms. Here we describe a fully automated process for finding phenotype similarities within a dataset acquired from an RNAi screen. The source code for the algorithms is available for free download.**

## I. INTRODUCTION

In the past few years, pipelines providing high-throughput biological imaging have been becoming increasingly popular, and are supported by the increasing availability of automated microscopy and high-performance computing. The availability of such instruments enables quantitative measurements of phenotype similarities in very large datasets. Applications include profiling drug responses [9], screening for small molecules [16], and classification of sub-cellular localization [2].

The availability of full genome sequences introduces the opportunity to perform large-scale analysis of gene functionality and reveal relationships between gene function and phenotype traits [11], [10]. However, due to the complex nature of the data, a successful implementation is subjected to the limitations of the computer vision algorithms, manifesting a large barrier to overcome.

Image analysis based on a specific morphology of the cell such as the size or shape [5], [6] may not provide clear relationships between gene knockout and function due to the variety of the terminal phenotypes expected. Instead, the algorithms should be able to handle phenotype similarities in a more general sense, covering a wide range of phenotype instances.

Here we describe an automated method that can be used for the purpose of automatic mining for phenotype similarities. This is a new approach to finding genes with similar functionality, and is different from finding gene similarity by comparing sequences (e.g., BLAST).

*This research was supported by the Intramural Research Program of the NIH, National Institute on Aging.

## II. METHODS

For the purpose of measuring phenotype similarities, we used WND-CHARM image classification algorithm. WND-CHARM [12], [14] makes use of a large set of 1025 image features extracted from each image. Each image feature (or a set of image features) can be found useful in finding similarities (or differences) between several different types of images.

Image features extracted by WND-CHARM can be divided into four groups: High contrast features, which include edge and object information; Polynomial decomposition, which is statistics based on the polynomial representation of the image; Statistics, which include multi-scale histograms and moments; and textures, which include Tamura [15] and Haralick [7] textures. This set of image features is computed by WND-CHARM on the raw pixels, but also on the Fourier, Chebyshev and Wavelet (Symlet 5) transforms of the image, and also on several compound transforms (Fourier-Wavelet and Fourier-Chebyshev). This variety of image features makes WND-CHARM effective for finding similarities between the different phenotypes, not known prior to the experiment. A more detailed description of WND-CHARM can be found at [12], [14], and the MATLAB and C implementations of the algorithm are available for free download at http://www.openmicroscopy.org.

Before image features are computed, each image is broken into 16 equal-sized tiles, and image features are computed for each tile separately. Then, the images are split into a training set and test set with two thirds of the images in the training set, and the remaining images used to evaluate the trained classifier.

Since very many feature values are being computed, some features are assumed to represent noise. To increase signal and remove noise, the features ranked by their discriminative power using simple Fisher scores [1]. The lowest 35% of the features are rejected. Once Fisher scores are computed, the weighted Euclidean distance dt,s between a tile t from the training set and a tile s from the test set is computed by $d_{t,s} = \sum_{f \in F} w_f (t_f - s_f)^2$, such that $F$ is the set of 1025 image features, $t_f$ and $s_f$ are the values of image feature $f$ in tile $t$ and $s$, respectively, and $w_f$ is the Fisher score of feature $f$.

Phenotype similarities are determined based on how images in the test set are classified using the images in the training set. When an image from the test set is classified, each of its 16 tiles is assigned with a similarity value to each of the genes in the training set. This is performed using a Weighted Nearest Neighbor rule [3], such that the similarity value $t_g$ of tile $t$ to gene $g$ is $t_g = (1/d_g)/\sum_{i<G} 1/d_i$, where $d_g$ is the distance from tile $t$ to its closest tile of gene $g$ in the training set, and $G$ is the set of all genes participating in the experiment. The similarity values of an image to any of the genes is computed simply by summing the 16 similarity vectors of the 16 tiles. The sum improves the signal-to-noise ratio of image similarity, compared to the similarity vector computed from one tile.

The resulting phenotype similarity values are computed by averaging the similarity values of all images for each of the genes participating in the experiment. That is, the similarity of the phenotype produced by knocking down gene $g_1$ to the phenotype produced by knocking down gene $g_2$ is the average similarity values of all images of gene $g_1$ to images of gene $g_2$. This results in a matrix of similarity values between all pairs of genes. This similarity matrix can be used for finding similar phenotypes that were produced by knocking down different genes, those findings may be used to reconstruct biological pathways.

Manually observing the similarity matrix and searching for high similarity values can become an exhausting task, especially when very many genes are involved. In order to make this task more convenient, we visualize the phenotype similarities using phylogenies (evolutionary trees) inferred automatically by Phylip package [4]. The phylogenies provide a tree of phenotypes with the lengths of the edges correlated with how similar the phenotypes are reflecting the values taken from the similarity matrix.

Deducing image similarities is considered a complicated task for computer programs due to the complex nature of the data, and therefore the presence of noise in the similarity values is unavoidable. Due to the noise generated by the image classifier, some of the phenotype similarities might not be symmetric. That is, the similarity of $g_1$ to $g_2$ may be 0.9, while the similarity of $g_2$ to $g_1$ is 0.86. Since the distances in the phylogeny are undirected, we simply average the two values to obtain one distance value between the two genes.

## III. RESULTS

To assess the efficacy of the image analysis we utilized a small dsRNA library (Open Biosystems) to cause single gene knockdown in cultured Drosophila cells. The library included 14 genes as listed in Table I, and can be divided into five expected phenotypic classes, which are Apoptotic (dIAP1), G1 arrest (pavarotti, CyclinE, MCM2, Rad17), G1 delay (MAPk-AK2), DNA damage (p38-MPK2, FANC-M, Cul-4) and one unknown Phenotype (CHD3).

Each gene had 50 experiments done on the same slide. After fixation, cells were stained with DAPI, washed and mounted, then deconvloved 1024×1024 images (one per experiment) were acquired using a Deltavision (Applied Precision, Inc., Issaqua, WA) microscope setup. Sample images are shown by Figure 1.

TABLE I

GENES INCLUDED IN THE TESTED DSRNA LIBRARY.

| ID | Gene name | CG # | Predicted phenotype |
|---|---|---|---|
| 1 | Pavarotti | 1258 | Binucleate |
| 2 | MAPk-AK2 | 3086 | G2 delay |
| 3 | CHD1 | 3733 | ?(chromodomain, helicase domain) |
| 4 | CyclinE | 3938 | G1 arrest |
| 5 | p38-MPK2 | 5475 | G1 delay |
| 6 | MCM2 | 7538 | G1 arrest |
| 7 | Rad17 | 7825 | DNA repair, 911 complex loading |
| 8 | FANC-M | 7922 | DNA repair |
| 9 | Pebble | 8114 | Binucleate |
| 10 | Cul-4 | 8711 | G1 arrest |
| 11 | Dmp53 | 10873 | Proliferation deficient |
| 12 | Loki(Rad53) | 10895 | Chk2 DNA damage siganlling |
| 13 | Diap1 | 12284 | Cell death |
| 14 | none/untreated | - | none |

After the images were acquired, WND-CHARM image classifier provided the following phenotype similarity matrix, shown by Table II. The values for each gene are normalized such that the similarity of each gene to itself is 1.

Figure 2 shows the corresponding phylogeny that visualizes the similarities values of Table II. As can be seen in Figure 2, the proposed method detected very similar phenotypes for gene 11 and 12. This observation is backed up by a clear link to previously reported studies, indicating that gene 11 (Dmp53) is a substrate for gene 12 (Loki).

Untreated cells (14) were not found similar to any of the other phenotypes, and so were genes 2 and 9, which dont have a similar phenotypes in the tested group of genes. Gene 13 (dIAP1) causes cell death, and was also found by the proposed method not to share similarities with any of the other tested genes. Gene 3 (CHD1) does not have a well-defined phenotype reported in the literature, and does not appear to be associated with any of the tested genes.

## IV. CONCLUSIONS

Mining for gene similarities has been attracting a considerable attention in the field of bioinformatics. Due to difficulties in processing and comparing large sets of different phenotypes, most attempts of finding genes with similar functionality are based on sequence analysis methods (e.g., BLAST). These methods heavily rely on the contention that genes with similar functionality should also have similar sequences, generating similar proteins. However, in many cases genes with different sequences can be part of the same biological mechanisms [8], [13].

Here we described an automated process that can be used for the purpose of automatic mining of phenotype similarities. This is a new approach of finding genes with similar functionality, and is different from finding gene similarity by comparing sequences.

Clearly, the proposed method can only sense phenotypic features that are visible using a microscope, and due to

TABLE II

PHENOTYPES SIMILARITY VALUES COMPUTED BY WND-CHARM.

| ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.00 | 0.839 | 0.913 | 0.906 | 0.730 | 0.928 | 0.955 | 0.730 | 0.597 | 0.782 | 0.853 | 0.822 | 0.653 | 0.571 |
| 2 | 0.571 | 1.000 | 0.637 | 0.592 | 0.510 | 0.513 | 0.505 | 0.506 | 0.595 | 0.509 | 0.470 | 0.442 | 0.651 | 0.533 |
| 3 | 0.446 | 0.353 | 1.000 | 0.364 | 0.279 | 0.454 | 0.304 | 0.293 | 0.197 | 0.360 | 0.499 | 0.400 | 0.169 | 0.218 |
| 4 | 0.868 | 0.812 | 0.766 | 1.000 | 0.737 | 0.836 | 0.936 | 0.726 | 0.671 | 0.745 | 0.753 | 0.735 | 1.124 | 0.639 |
| 5 | 0.743 | 0.749 | 0.598 | 0.746 | 1.000 | 0.753 | 0.754 | 0.956 | 0.980 | 0.908 | 0.833 | 0.888 | 0.626 | 1.027 |
| 6 | 0.905 | 0.733 | 0.966 | 0.830 | 0.695 | 1.000 | 0.840 | 0.705 | 0.562 | 0.782 | 0.870 | 0.823 | 0.569 | 0.534 |
| 7 | 0.792 | 0.668 | 0.659 | 0.792 | 0.681 | 0.791 | 1.000 | 0.683 | 0.636 | 0.694 | 0.714 | 0.713 | 0.755 | 0.560 |
| 8 | 0.788 | 0.755 | 0.659 | 0.759 | 1.001 | 0.796 | 0.774 | 1.000 | 0.929 | 0.922 | 0.890 | 0.958 | 0.586 | 0.960 |
| 9 | 0.427 | 0.568 | 0.299 | 0.459 | 0.603 | 0.450 | 0.497 | 0.606 | 1.000 | 0.537 | 0.417 | 0.459 | 0.569 | 0.819 |
| 10 | 0.841 | 0.882 | 0.828 | 0.825 | 0.968 | 0.839 | 0.814 | 0.990 | 0.978 | 1.000 | 0.922 | 0.946 | 0.722 | 0.936 |
| 11 | 0.832 | 0.683 | 0.964 | 0.769 | 0.783 | 0.861 | 0.764 | 0.802 | 0.579 | 0.833 | 1.000 | 0.973 | 0.506 | 0.585 |
| 12 | 0.830 | 0.644 | 0.835 | 0.788 | 0.865 | 0.851 | 0.751 | 0.889 | 0.646 | 0.876 | 0.980 | 1.000 | 0.495 | 0.662 |
| 13 | 0.248 | 0.333 | 0.162 | 0.353 | 0.265 | 0.232 | 0.328 | 0.262 | 0.355 | 0.251 | 0.167 | 0.184 | 1.000 | 0.330 |
| 14 | 0.387 | 0.394 | 0.301 | 0.376 | 0.564 | 0.400 | 0.382 | 0.542 | 0.668 | 0.502 | 0.443 | 0.485 | 0.360 | 1.000 |

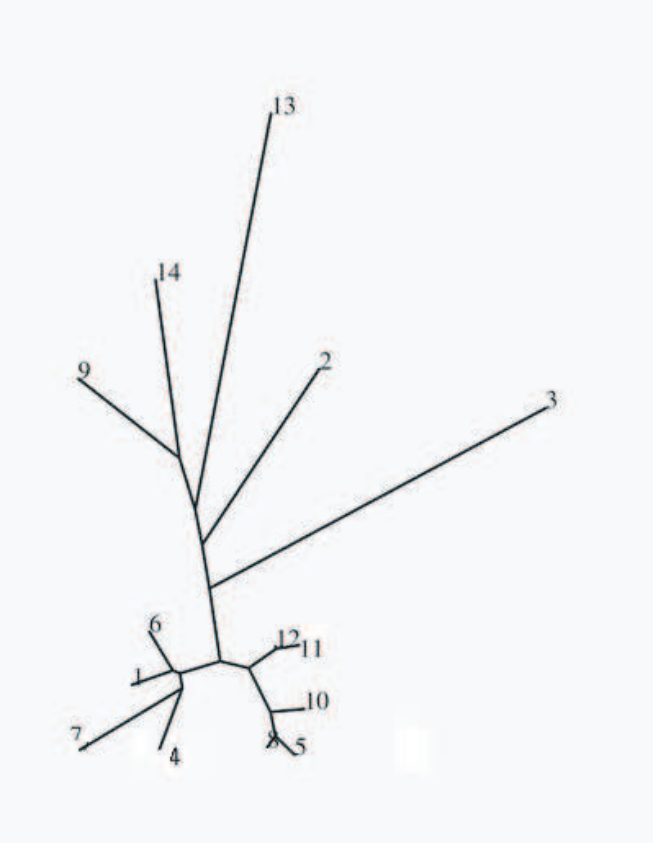


Fig. 2. The phylogeny of phenotype similarities generated from the similarity values of Table II.

the complex nature of the quantification of the phenotype morphology it is not expected to detect all genes with similar functionality. However, given that very many of the genes in any organism are not mapped to any known function, applying this method on large sets of phenotypes with single gene knockdown can potentially reconstruct biological pathways.

## ACKNOWLEDGMENT

This research was supported entirely by the Intramural Research Program of the NIH, National Institute on Aging.

## REFERENCES

[1] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[2] M. V. Boland, R. F. Murphy, neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells, *Bioinformatics*, vol. 17, 2001, pp 1213-1223.

[3] T. K. J. Brown, The weighted nearest neighbor rule for class dependent sample sizes, *IEEE Trans. on Information Theory*, vol. 25, 1979, pp 617-619.

[4] J. Felsenstein, PHYLIP Phylogeny Inference Package, Version 36, 2004.

[5] A. G. Fraser, et al., Functional genomic analysis of C. elegans chromosome I by systematic RNA interference, *Nature*, vol. 408, 2000, pp 325-330.

[6] G. Giaever, et al., Functional profiling of the Saccharomyces cerevisiae genome, *Nature*, vol. 418, 2002, pp 387-391.

[7] R. M. Haralick, K. Shanmugam, I. Dinstein, Textural Features for Image Classification, *IEEE Tran. on Systems, Man, and Cybernetics*, vol. 6, 1973, pp 269-285.

[8] G. Lettre, et al., Identification of ten loci associated with height highlights new biological pathways in human growth, *Nat. Genet.*, vol. 40, 2008, pp 584-591.

[9] L. Loo, L. F. Wu, S. J. Altschuler, Image-based multivariate profiling of drug responses from single cells, *Nature Methods*, vol. 4, 2007, pp 445-453.

[10] R. Pepperkok, J. Ellenberg, High-throughput fluorescence microscopy for systems biology, *Nature Reviews Molecular Cell Biology*, vol. 7, 2006, pp 690-696.

[11] Y. Ohya, et al. High-dimensional and large-scale phenotyping of yeast mutants, *Proc. Natl. Acad. Sci.*, vol. 102, 2005, pp 19015-19020.

[12] N. Orlov, L. Shamir, T. Macura, J. Johnston, D. M. Eckely, I. Goldberg, WND-CHARM: Multi-purpose image classification using compound image transforms, *Pattern Recognition Letters*, vol. 29, 2008, pp. 1684-1693.

[13] S. Sanna, et al., Common variants in the GDF5-UQCC region are associated with variation in human height, *Nat Genet*, vol. 40, 2008, pp 198-203.

[14] L. Shamir, N. Orlov, T. Macura, D. M. Eckley, J. Johnston, I. G. Goldberg, Wndchrm - An Open Source Utility for Biological Image Analysis, *BMC Source Code for Biology and Medicine*, vol. 3, 2008, pp 13.

[15] H. Tamura, S. Mori, T. Yamavaki, Textural features corresponding to visual perception, *IEEE Trans. on Syst., Man and Cyber.*, vol. 8, 1978, pp 460-472.

[16] M. Tanaka, et al., An unbiased cell morphology-based screen for new, biologically active small molecules, *PLoS Bio.*, vol. 3, 2005, p e128.

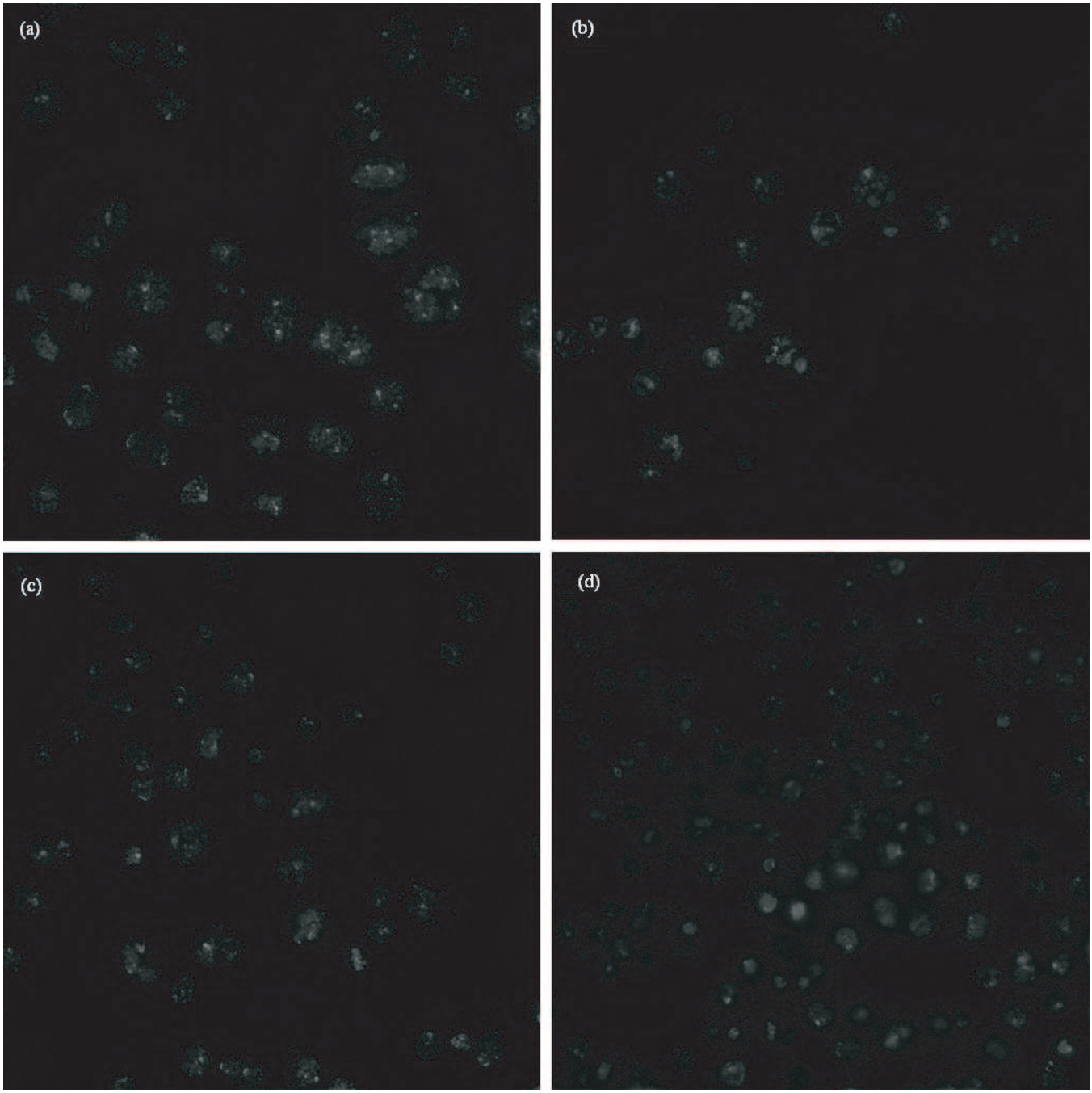


Fig. 1. Sample images of genes Pavarotti (a), CyclineE (b), p38-MPK2 (c) and untreated cells (d)