Data mining tools for the *Saccharomyces cerevisiae* morphological database

Taro L. Saito[1,4], Jun Sese[2], Yoichiro Nakatani[2,4], Fumi Sano[3,4], Masashi Yukawa[3,4], Yoshikazu Ohya[3,4] and Shinichi Morishita[2,4,*]

[1]Department of Computer Science, Graduate School of Information Science and Technology, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan and [2]Department of Computational Biology and [3]Department of Integrated Biosciences, Graduate School of Frontier Sciences, University of Tokyo, Building FSB-101, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan and [4]Japan and Institute for Bioinformatics and Research and Development, Japan Science and Technology Corporation, Science Plaza, 5-3, Yonbancho, Chiyoda-ku, Tokyo 102-8666, Japan

ABSTRACT

For comprehensive understanding of precise morphological changes resulting from loss-of-function mutagenesis, a large collection of 1 899 247 cell images was assembled from 91 271 micrographs of 4782 budding yeast disruptants of non-lethal genes. All the cell images were processed computationally to measure ∼500 morphological parameters in individual mutants. We have recently made this morphological quantitative data available to the public through the *Saccharomyces cerevisiae* Morphological Database (SCMD). Inspecting the significance of morphological discrepancies between the wild type and the mutants is expected to provide clues to uncover genes that are relevant to the biological processes producing a particular morphology. To facilitate such intensive data mining, a suite of new software tools for visualizing parameter value distributions was developed to present mutants with significant changes in easily understandable forms. In addition, for a given group of mutants associated with a particular function, the system automatically identifies a combination of multiple morphological parameters that discriminates a mutant group from others significantly, thereby characterizing the function effectively. These data mining functions are available through the World Wide Web at http://scmd.gi.k.u-tokyo.ac.jp/.

MORPHOLOGICAL DATABASE

To study the global regulation of cell morphological characteristics, a number of groups have recently reported genome-wide screening data for yeast mutants with abnormal morphology (1–5). Despite the relatively simple ellipsoidal shape of yeast cells, in the past, cell morphology researchers processed information on cells manually. These time consuming, entirely subjective tasks motivated us to develop image-processing software called CalMorph (6), which automatically extracts yeast cells from micrographs and processes them to measure morphological characteristics such as cell size, roundness, bud neck position angle, nuclear position and actin localization. Using our software, we have retrieved 1 899 247 cells from 91 271 micrographs of 4782 mutants, which cover almost all of the yeast non-essential mutants cultured from the deleted strains available from EURO-SCARF. All cell images, micrographs and quantitative values of morphological parameters are freely available from the SCMD database (7), which presents information that is complementary to the existing sequence and gene-expression databases (8–12).

CELL IMAGE PROCESSING

Our software processes micrographs of cells stained with fluorescein isothiocyanate–Concanavalin A (FITC-ConA) for cell wall identification, with DAPI to localize nuclei and with Rh-ph to visualize the actin distribution. The photos in Figure 1A show three images stained with the respective dyes. Figure 1B presents the result of combining three photos
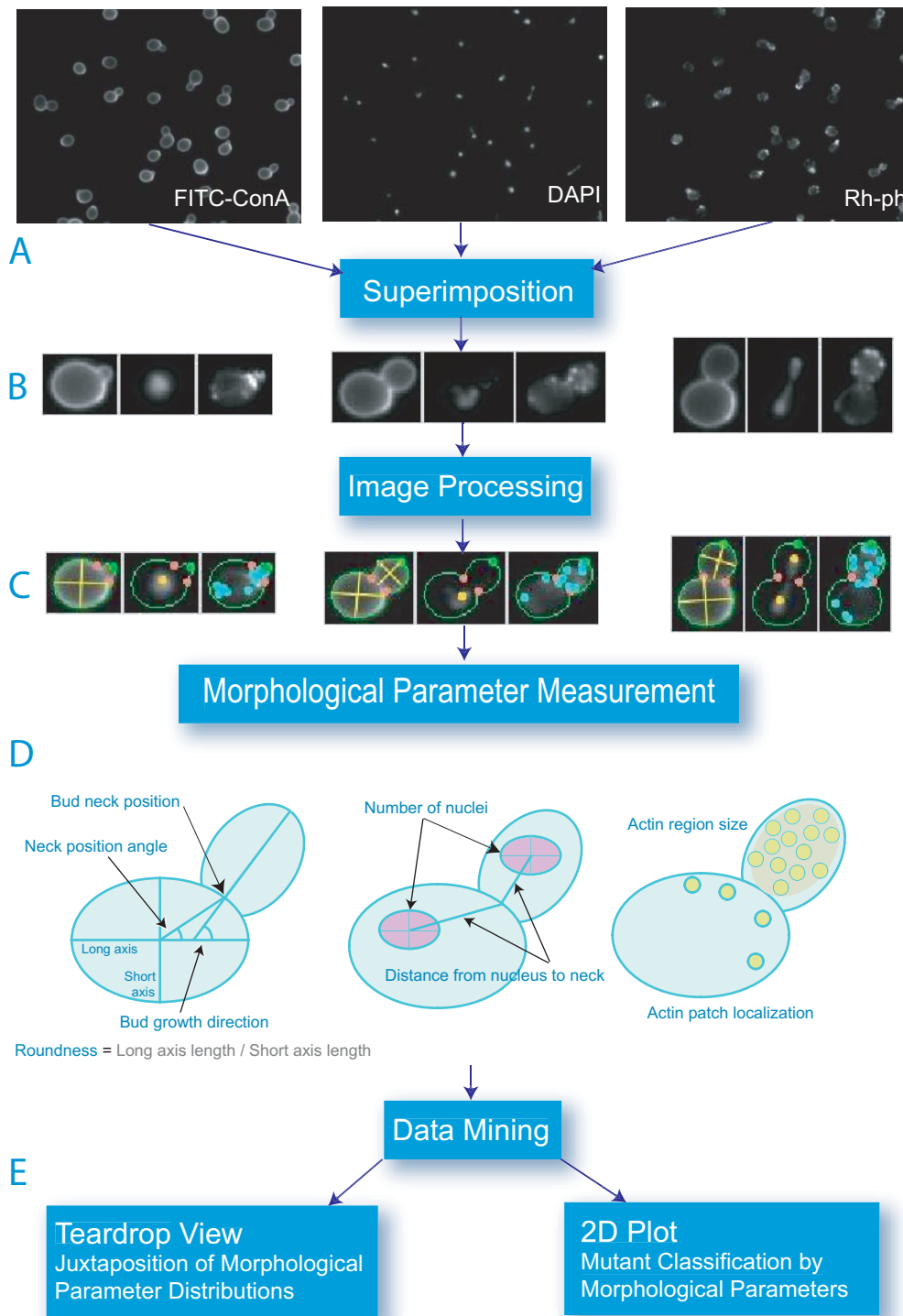
Figure 1. Workflow of image processing and data mining. (**A**) Input photos of cells strained with FITC–ConA, DAPI and Rh-ph to visualize the cell wall, nuclei and actin distribution, respectively. (**B**) Superimposition of three micrographs for individual cells. (**C**) Image-processing results. (**D**) Several examples of ∼500 morphological parameters. (**E**) Data mining processes.

by superimposing images of the cell wall, nuclei and actin for individual cells.

Figure 1C displays image-processing results. Our image-processing software first identifies the cell wall, attempts to fit an ellipse to each mother cell or bud and colors the cell wall green. The yellow lines show the long and short axes of the fitted ellipses. Bud necks that separate mother cells and buds are illustrated by using two red bullets. Identifying the cell

wall makes it easier to determine information on the localization of nuclei and actin patches relative to the cell wall. In Figure 1C, nuclei and actin patches are represented using yellow and light blue bullets, respectively.

Figure 1D shows the primary morphological parameters of cells. The quantitative values of these parameters may change slightly from cell to cell. To perform rigorous statistical analysis of the significance of morphological changes, we need to know the distribution of morphological parameter values for individual cells; this requires that we collect an ample number of image-processed cells and their parameter values. More than 200 image-processed cells were collected for each mutant using a sufficient number of micrographs. Then, ~500 morphological parameters were calculated for the mutants.

DATA MINING

Since there are so many parameters and mutants, some tools for assisting with data mining tasks will help users.

Search

Morphological data should be useful for identifying the morphological changes in particular mutants. Users can query a yeast mutant of interest using its open reading frame name or its gene name. They can also browse average shapes of the mutant, average morphological parameter values, raw and processed micrographs and lists of individual cells associated with morphological parameter values. Users can also

provide a typical morphological shape or a particular mutant as a query and ask the system to search for mutants that are similar in shape to the query. This function is called 'morphology search' (7).

Teardrop view—juxtaposition of morphological parameter distributions

In order to understand which morphological parameters of a particular mutant are abnormal, the system displays the distribution of all mutants for each parameter and highlights the focal mutant value in pink (see Figure 2). The system juxtaposes the distributions of all parameters in parallel, making it easy for users to comprehend the overview of distributions and abnormal parameters at a glance. Parameters are colored blue or pink if their changes are statistically significant in terms of their distributions.

Mutant classification in terms of morphological parameters

Another promising application of morphological parameters is to use them to predict gene functions. For instance, suppose that one is interested in finding a group of genes involved in a particular biological process such as DNA repairs and cell wall construction. You can ask the system to look for a combination of multiple morphological parameters that discriminate disruptants of genes that are known to be relevant to the biological process of interest (see Figure 3). These morphological parameters allow us to define distances between disruptants. If we identify disruptants that are not
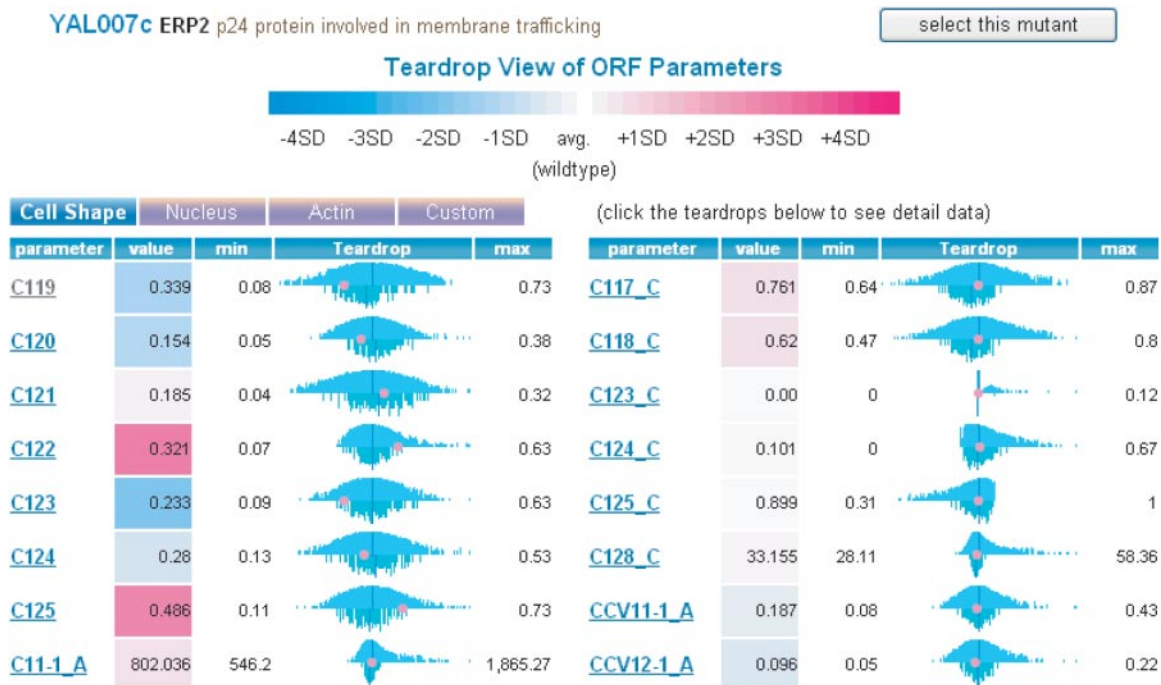

Figure 2. Teardrop view juxtaposes the morphological parameter distributions of all parameters for all mutants and the wild-type *HIS3* (*YOR202w*). For each morphological parameter, the distribution of all mutants and that of the wild type are displayed back-to-back in the upper and lower halves, respectively. The thin central line in each distribution represents the average. The pink dots in the distributions show the data for the focal mutant. Since some wild-type distributions are abnormal and are difficult to fit to any established statistical distribution, the statistical significance of a particular parameter value for a mutant is not assessed in terms of the *P*-value but is estimated using the SD-score (or *Z*-score), the difference between the parameter value and the average of the wild type divided by the standard deviation of the wild-type distribution. The degree of each SD-score is represented by its color.
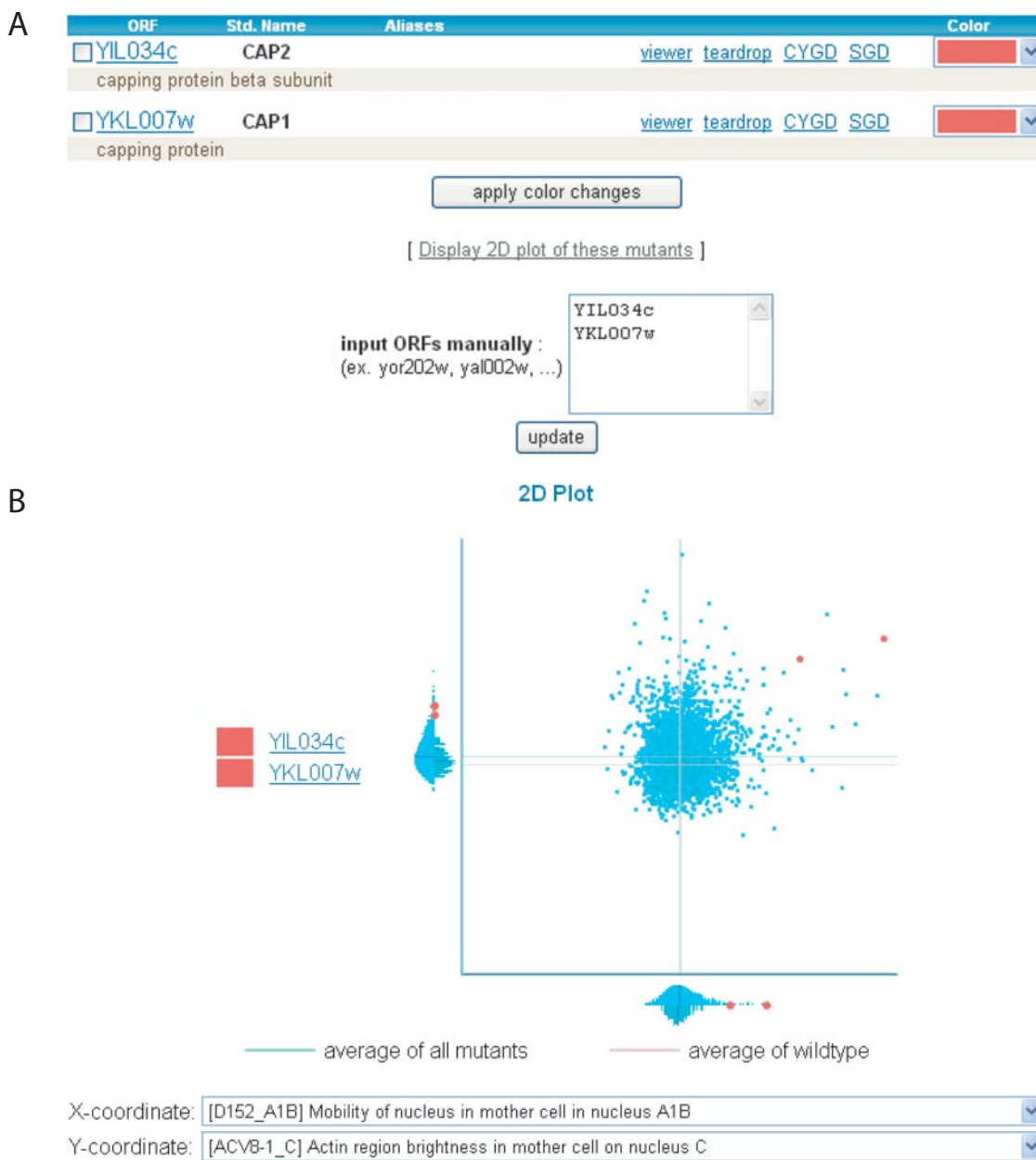
Figure 3. Mutant classification in terms of morphological parameters. (**A**) Select a group of mutants such that the disrupted genes are involved in a biological process of interest. In the example, *CAP1* (*YKL007w*) and *CAP2* (*YIL034c*), capping protein and its beta subunit, are selected. (**B**) The system returns two morphological parameters that best discriminate *CAP1* and *CAP2*, which are represented by two pink dots in the 2D plot. Light blue spots represent mutants, while blue spots are instances of the wild type. Each parameter dimension is associated with the Teardrop view of the morphological parameter distributions.

known to be related to any particular biological process but are closer to disruptants that are relevant to the focal biological process, these disrupted genes are potentially involved in the biological process.

CUSTOMIZATION AND DATA AVAILABILITY

To facilitate customization according to users' interests for the ease of browsing, a dialog-based interface for the parameter selection page helps users choose parameters displayed in datasheets and are memorized in the system. The system also allows users to download the list of selected parameter

values for selected mutants in the XML format or in tabular form. Users can also select particular mutants of interest so that they are always shown in Teardrop View and 2D plot.

UPDATES AND FUTURE DIRECTIONS

The web server currently presents morphological parameter values of disruptants of non-essential genes, but mutants of lethal genes will be processed and available in the future.

ACKNOWLEDGEMENT

Conflict of interest statement. None declared.

REFERENCES

1. Winzeler,E.A., Shoemaker,D.D., Astromoff,A., Liang,H., Anderson,K., Andre,B., Bangham,R., Benito,R., Boeke,J.D., Bussey,H. *et al.* (1999) Functional characterization of the *S.cerevisiae* genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.
2. Jorgensen,P., Nishikawa,J.L., Breitkreutz,B.J. and Tyers,M. (2002) Systematic identification of pathways that couple cell growth and division in yeast. *Science*, **297**, 395–400.
3. Zhang,J., Schneider,C., Ottmers,L., Rodriguez,R., Day,A., Markwardt,J. and Schneider,B.L. (2002) Genomic scale mutant hunt identifies cell size homeostasis genes in *S.cerevisiae*. *Curr. Biol.*, **12**, 1992–2001.
4. Ni,L. and Snyder,M. (2001) A genomic study of the bipolar bud site selection pattern in *Saccharomyces cerevisiae*. *Mol. Biol. Cell*, **12**, 2147–2170.
5. Giaever,G., Chu,A.M., Ni,L., Connelly,C., Riles,L., Veronneau,S., Dow,S., Lucau-Danila,A., Anderson,K., Andre,B. *et al.* (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, **418**, 387–391.
6. Ohtani,M., Saka,A., Sano,F., Ohya,Y. and Morishita,S. (2004) Development of image processing program for yeast cell morphology. *J. Bioinform. Comput. Biol.*, **1**, 695–709.
7. Saito,L.T., Ohtani,M., Sawai,H., Sano,F., Saka,A., Watanabe,D., Yukawa,M., Ohya,Y. and Morishita,S. (2004) SCMD: *Saccharomyces Cerevisiae* Morphological Database. *Nucleic Acids Res.*, **32**, 319–322.
8. Balakrishnan,R., Christie,K., Costanzo,M., Dolinski,K., Dwight,S., Engel,S., Fisk,D., Hirschman,J., Hong,E., Nash,R. *et al.* (2005) Fungal BLAST and Model Organism BLASTP Best Hits: new comparison resources at the *Saccharomyces* Genome Database (SGD). *Nucleic Acids Res.*, **33**, D374–D377.
9. Güldener,U., Münsterkötter,M., Kastenmüller,G., Strack,N., van Helden,J., Lemer,C., Richelles,J., Wodak,S., García-Martínez,J., Pérez-Ortín,J. *et al.* (2005) CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res.*, **33**, D364–D368.
10. Mewes,W., Amid,C., Arnold,R., Frishman,D., Güldener,U., Mannhaupt,G., Münsterkötter,M., Pagel,P., Strack,N., Stümpflen,V. *et al.* (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, **32**, D41–D44.
11. Riffle,M., Malmström,L. and Davis,T. (2005) The Yeast Resource Center Public Data Repository. *Nucleic Acids Res.*, **33**, D378–D382.
12. Lelandais,G., Crom,S., Devaux,F., Vialette,S., Church,G., Jacq,C. and Marc,P. (2004) yMGV: a cross-species expression data mining tool. *Nucleic Acids Res.*, **32**, D323–D325.