Automated image analysis of protein localization in budding yeast

Shann-Ching Chen¹, Ting Zhao¹, Geoffrey J. Gordon² and Robert F. Murphy¹,²,³,*

¹Center for Bioimage Informatics, Department of Biomedical Engineering, ²Department of Machine Learning and ³Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA

ABSTRACT

Motivation: The yeast Saccharomyces cerevisiae is the first eukaryotic organism to have its genome completely sequenced. Since then, several large-scale analyses of the yeast genome have provided extensive functional annotations of individual genes and proteins. One fundamental property of a protein is its subcellular localization, which provides critical information about how this protein works in a cell. An important project therefore was the creation of the yeast GFP fusion localization database by the University of California, San Francisco, USA (UCSF). This database provides localization data for 75% of the proteins believed to be encoded by the yeast genome. These proteins were classified into 22 distinct subcellular location categories by visual examination. Based on our past success at building automated systems to classify subcellular location patterns in mammalian cells, we sought to create a similar system for yeast.

Results: We developed computational methods to automatically analyze the images created by the UCSF yeast GFP fusion localization project. The system was trained to recognize the same location categories that were used in that study. We applied the system to 2640 images, and the system gave the same label as the previous assignments to 2139 images (81%). When only the highest confidence assignments were considered, 94.7% agreement was observed. Visual examination of the proteins for which the two approaches disagree suggests that at least some of the automated assignments may be more accurate. The automated method provides an objective, quantitative and repeatable assignment of protein locations that can be applied to new collections of yeast images (e.g. for different strains or the same strain under different conditions). It is also important to note that this performance could be achieved without requiring colocalization with any marker proteins.

Availability: The original images analyzed in this article are available at http://yeastgfp.ucsf.edu, and source code and results are available at http://murphylab.web.cmu.edu/software

Contact: murphy@cmu.edu

1 INTRODUCTION

Yeast is the first eukaryote whose genome was sequenced completely. Large-scale analysis has been performed by different research groups and extensive annotations and experimentally determined properties have been organized into various databases, such as the Saccharomyces Genome Database (SGD) (Cherry et al., 1998) and the Comprehensive Yeast Genome Database (CYGD) (Guldener et al., 2005).

One critical protein property is its subcellular localization, which provides critical information about how a protein works inside a cell. This property is frequently determined by visual interpretation of fluorescence microscope images. A major advance therefore came from the creation of the yeast GFP fusion localization database (Huh et al., 2003) at the University of California, San Francisco, USA (UCSF). In order to maximally preserve the wild-type levels of protein expression, this GFP-tagged library was constructed by inserting the coding sequence of Aequorea Victoria GFP (S65T) (Tsien, 1998) into the yeast genome in-frame immediately preceding the stop codon of each open reading frame (ORF). Annotation of the images was done manually: two human scorers initially independently classified these proteins into one or more of 12 subcellular localization categories and then refined these categories by performing a series of co-localization experiments with monomeric red fluorescent proteins (mRFP) (Campbell et al., 2002) whose location had been characterized previously. They defined 22 unique location categories (one of which was ‘ambiguous’) and assigned each protein to either one or mixtures of these categories. This database provides localization data for 75% of the proteins encoded by the yeast genome.

Over the past 10 years, we have developed automated systems for comprehensive and quantitative analysis of protein subcellular location patterns in mammalian cells. By utilizing informative numerical descriptors for high resolution images and statistical machine-learning methods for pattern recognition, these systems can recognize the major subcellular patterns (Boland et al., 1998, 2001; Murphy et al., 2000) and learn new patterns directly from fluorescence microscope images (Chen and Murphy, 2005; Chen et al., 2003). Previous work showed that our system is able to distinguish similar patterns better than visual examination (Murphy et al., 2003).

In this work, we present computational methods to classify the protein images in the yeast GFP fusion localization database. The results demonstrate that highly accurate and objective assignments can be made by a fully automated system.

2 METHODS

An overview of our method is presented in Figure 1 and described briefly below. Our initial desire was to utilize machine-learning algorithms to classify proteins into one of 22 clearly predefined...
sets, while the category 'ER_to_Golgi' has only 6 image sets. The four major categories, 'cytoplasm', 'nucleus', 'mitochondrion' and 'ER', contain 2059 image sets in each category is not uniform. The number of images in each category in the yeast GFP fusion localization database

| Category               | Size | Category    | Size |
|------------------------|------|-------------|------|
| cytoplasm              | 819  | endosome    | 34   |
| nucleus                | 490  | late_Golgi  | 33   |
| mitochondrion          | 484  | actin       | 27   |
| ER                     | 266  | peroxisome  | 21   |
| vacuole                | 121  | lipid_particle | 19 |
| punctate_composite     | 72   | Golgi       | 15   |
| nucleolus              | 68   | bud_neck    | 14   |
| cell_periphery         | 57   | early_Golgi | 11   |
| vacuolar_membrane      | 54   | microtubule | 10   |
| nuclear_periphery      | 53   | ER_to_Golgi | 6    |
| spindle_pole           | 39   |             |      |

2.2 Feature extraction and image segmentation

2.2.1 Field-level features To analyze the location patterns in an image, we can extract features that summarize the pattern at the level of the entire field (assuming that a field only contains a homogenous population of cells expressing a single labeled protein). The advantage of using field-level features is that it makes it unnecessary to segment each image into subregions containing individual cells, and we have previously shown that such features can be used to obtain good classification accuracy for the major subcellular patterns in mammalian cells (Huang and Murphy, 2004a). The feature set used previously consisted of 26 features, 13 morphological and 13 Haralick texture features using 256 gray levels for images resampled to 1.15 micron/pixel. We extended this feature set by adding 60 Gabor texture features and calculated the Haralick texture features using 128 gray levels at the original image resolution.

2.2.2 Cell-level features If images are segmented into single cell regions, additional features that are not appropriate for whole fields can be calculated. These include Zernike moment features, additional morphological features and wavelet features. The details for different feature sets (which we term SLF) have been reviewed previously (Huang and Murphy, 2004b). The advantage of the cell-level features is that more information in one image can be extracted and utilized. The disadvantage is the requirement for a cell segmentation step, which is crucial because its accuracy determines the accuracy of the downstream analysis. We first developed a novel segmentation method based on a graphical model approach (Chen et al., 2006) which is particularly suitable for this yeast dataset. We then extracted 245 cell-level features for each cell subregion. These features were calculated after background subtraction and normalization of the total intensity of the image to one. Thus differences in protein expression and background (reflected in brightness and contrast of the image) are removed. This feature set contains 49 Zernike moment features, 16 morphological features, 30 wavelet features, 6 DNA features, 5 edge features, 1 non-object feature and 78 Haralick texture features using 128 gray levels at the original image resolution and various resolutions with downsampling factors ranging from 2 to 6.

2.2.3 Image segmentation with graphical model approach Yeast cells are often clustered together, so it is hard to define good initial seeding for conventional techniques, such as the seeded watershed algorithm (Bengtsson et al., 2004; Velliste and Murphy, 2002) and Voronoi segmentation (Rodenacker and Bischoff, 1990). We developed a graphical-model-based segmentation approach.

---

Fig. 1. Overview of the approach for cell-level classification.

Localization categories. Since some of the proteins in the UCSF database were assigned to an ‘ambiguous’ category or to a mixture of localization categories, we initially excluded them to obtain a set of proteins showing only one location pattern. We also excluded the ‘punctate_composite’ class given the inherent ambiguity of this designation. We used a novel graphical model approach to segment each image into single cell regions and extracted numerical features to describe the pattern in each cell. These were used to train and test support vector machine classifiers. Plurality voting was used to combine results from different cells and classifiers to obtain a unique localization category for each image.
2.3 Support vector machine classification

Support vector machines were originally designed for binary classification by finding a maximum margin hyperplane between two classes (Cortes and Vapnik, 1995). They can be extended to solve multi-class classification problems by combining several binary classifiers. For the work described here, we used the LIBSVM package (http://www.csie.ntu.edu.tw/~cjlin/libsvm) with the one-against-one method (Hsu and Lin, 2002) for multi-class classification. To improve accuracy for small classes, we used class-dependent weights. The weight for class C was the size of the smallest class divided by the size of class C. For SVMs, the proper choice of parameters will greatly affect the classification performance. There are two parameters to be adjusted, C (the penalty parameter of the error term) and γ (the width of Gaussian function in the kernel function), for a SVM with a radial basis function kernel:

\[ K(x_i,y_j) = e^{-\gamma \|x_i-x_j\|^2}, \gamma > 0 \]

For parameter tuning with cell-level features, we randomly split all the data into 6 folds, 5 of them for training and one of them for testing, ran stepwise discriminant analysis (SDA) to select the most discriminative features on the training set, and conducted a grid-search on C and γ using 6-fold cross-validation. The best set of parameters for the cell level classification was \((C, \gamma) = (2^{15}, 2^{-6})\). We have previously evaluated eight different feature reduction methods in the context of subcellular pattern analysis and obtained the best performance using SDA (Huang et al., 2003).

2.4 Plurality voting

Given a yeast image which was segmented into several cell regions, and results for the classification of each cell, we can utilize information from more than one cell to improve the classification accuracy. Previous work (Boland and Murphy, 2001) has shown individual classification accuracy of sets of HeLa cells can be improved from 83 to 98% by allowing cells to vote for a single classification for the entire slide (and choosing the class that receives a plurality of the vote). We adopt the same strategy here. We used 25 seeds to randomly split the data into 6 folds, 5 of them for training and one of them for testing, ran stepwise discriminant analysis (SDA) to select the most discriminative features on the training set, and conducted a grid-search on C and γ using 6-fold cross-validation. The best set of parameters for the cell level classification was \((C, \gamma) = (2^{15}, 2^{-6})\). We have previously evaluated eight different feature reduction methods in the context of subcellular pattern analysis and obtained the best performance using SDA (Huang et al., 2003).

For parameter tuning with cell-level features, we randomly split all the data into 6 folds, 5 of them for training and one of them for testing, ran stepwise discriminant analysis (SDA) to select the most discriminative features on the training set, and conducted a grid-search on C and γ using 6-fold cross-validation. The best set of parameters for the cell level classification was \((C, \gamma) = (2^{15}, 2^{-6})\). We have previously evaluated eight different feature reduction methods in the context of subcellular pattern analysis and obtained the best performance using SDA (Huang et al., 2003).

We first evaluated the classification of entire, unsegmented fields of cells. Using 6-fold cross-validation on the four largest classes, we obtained a 92.7% overall accuracy (Table 2). Here the 6-fold cross-validation just evenly split the data into 6 folds for each class. However, the classification accuracy decreased significantly when this approach was extended to all 21 classes (data not shown). The overall accuracy (number of correctly classified images out of all images) was 76.2% but the adjusted accuracy (the average over the diagonal of the confusion matrix) was 33.9%. This suggests that although most images from large classes were classified correctly, most images from small classes were classified incorrectly. It also illustrates the difficulty of classifying uneven data described in Section 2.4. This is at least partially due to the small number of training images available for the small classes (since only one set of features is calculated for each field).

3.2 Cell-level classification

We next evaluated cell-level classification as described above. The results are shown in Table 3. A total of 2139 out of 2640 images (81%) were classified correctly. Most images in the large classes were classified correctly, but many images in the small classes were not (probably due to the small number of training images available for these classes). However, the accuracy of the classifications can be significantly improved by considering only those that have high confidence. This can be done by setting a threshold on the fraction of classifiers that must agree in order for a classification to be assigned. When this number is varied from 28% to 100%, the precision of the classification results (the number of assignments that agree with the UCSF assignments divided by the number of assignments made) increases from 81.02% to 94.70%. The recall (fraction of all images whose classification agrees with the UCSF assignments) is over 80% at this high accuracy (Fig. 2). The accuracies in Table 3 assume that the UCSF labels were all correct. Since the vast majority is likely to be correct, the results strongly confirm the feasibility of automated classification of yeast images, even without using co-localization. However, it is possible that at least some of the visually assigned labels are incorrect. We therefore next examined the difference between our

| | Cytoplasm | Nucleus | Mitochondrion | ER |
|---|---|---|---|---|
| Cytoplasm | **95.9** | 0.5 | 1.2 | 2.4 |
| Nucleus | 2.6 | **94.9** | 2.4 | 0 |
| Mitochondrion | 6.2 | 1 | **92** | 0.8 |
| ER | 15.4 | 1.1 | 3.4 | **80.1** |

In the confusion matrix representation, the rows and columns indicate true and predicted class, respectively. The diagonal entries represent correct classifications, while the off-diagonal entries represent incorrect ones. The overall accuracy (percentage of correct images) is 92.7% and the adjusted accuracy (average over the diagonal of the confusion matrix) is 90.7% for the field-level classification.
automatically-derived labels and the labels from the UCSF database.

3.3 Identification of potentially incorrect labels

Of the 2640 images classified in Table 3, there are 501 images whose computer-assigned labels did not match the previously assigned labels. Of these, 96 were classified with 100% confidence (25 out of 25 classifiers agreed). A list of these proteins can be found at http://murphylab.web.cmu.edu/data. (The complete list is available at http://murphylab.web.cmu.edu/data.). We plan to further explore these proteins and evaluate evidence from other databases to get further insight into which may be correct. As a preliminary analysis, we show images for two of the proteins in Figures 3 and 4. According to the CYGD database, the ORF YAL009W in Figure 3 is a meiotic protein which should have strong expression in the M (mitosis) phase. Our method classified it as vacuole protein with 52% confidence. (13 of the 25 classifiers classified it as vacuole, while another 11 classifiers classified it as cytoplasm and one classifier classified it as mitochondrion). We hypothesize that these cells are not in the M phase so that the proteins are not expressed in the nucleus. Another example is the ORF YGR130C in Figure 4. According to the CYGD database, YGR130C is a protein of unknown function localized to cytoplasm, and it is classified as a punctate_composite protein by the UCSF database. We can tell from the image that the proteins are localized at cell periphery, which is consistent with the automated classification. (91 of the 150 classifiers classified it as cell_periphery, 45 classifiers classified it as cytoplasm and 14 classifiers classified it as ER). These two examples suggest that automated classification can provide a better measurement of the protein location patterns, at least in some cases. The validations of other proteins with potentially incorrect labels are still under investigations.

Table 3. The confusion matrix of cell-level classification on 20 classes of pure single location patterns. The overall accuracy (percentage of images whose classification agrees with the UCSF assignments) is 81%

|              | Cytoplasm | nucleus | mitochondrion | ER | vacuole | nucleolus | cell_periphery | vacuolar_membrane | nuclear_periphery | spindle_pole | late_Golgi | actin | peroxisome | lipid_particle | Golgi | bud_neck | early_Golgi | microtubule | ER_to_Golgi |
|--------------|-----------|---------|---------------|----|---------|-----------|---------------|-------------------|-------------------|--------------|-----------|------|-----------|----------------|-------|----------|-------------|------------|------------|
| cytoplasm    | 94.4      | 0.6     | 1.5           | 2.1| 0       | 0.7        | 0             | 0                 | 0                 | 0.2          | 0         | 0.2  | 0         | 0.2            | 0     | 0        | 0           | 0.2        | 0         |
| nucleus      | 0.6       | 88.6    | 1.8           | 0  | 5.5     | 2.4        | 0             | 0                 | 0                 | 0.2          | 0         | 0.2  | 0         | 0.2            | 0     | 0        | 0           | 0.2        | 0         |
| mitochondrion| 3.9       | 0.4     | 90.9          | 1.2| 2.5     | 0.2        | 0             | 0                 | 0                 | 0.2          | 0         | 0.2  | 0         | 0.2            | 0     | 0        | 0           | 0.2        | 0         |
| ER           | 20.0      | 0       | 1.5           | 70.2| 3.8     | 0           | 1.9           | 0                 | 0                 | 0.8          | 0         | 0.4  | 0.4       | 0.8            | 0     | 0        | 0           | 0.4        | 0         |
| vacuole      | 7.4       | 9.1     | 0.8           | 7.4| 73.6    | 0           | 0             | 1.7               | 0                 | 0            | 0         | 0    | 0         | 0              | 0     | 0        | 0           | 0          | 0         |
| nucleolus    | 0         | 20.6    | 0             | 0  | 2.9     | 73.5       | 0             | 0                 | 1.5               | 1.5          | 0         | 0    | 0         | 0              | 0     | 0        | 0           | 0          | 0         |
| cell_periphery| 14.0     | 0       | 1.8           | 7.0| 1.8     | 0           | 75.4          | 0                 | 0                 | 0            | 0         | 0    | 0         | 0              | 0     | 0        | 0           | 0          | 0         |
| vacuolar_membrane| 11.1 | 5.6    | 1.9           | 7.4| 16.7    | 0           | 0             | 51.9              | 1.9               | 1.9          | 0         | 0    | 0         | 0              | 0     | 0        | 0           | 0          | 0         |
| nuclear_periphery| 0   | 3.8     | 0             | 3.8| 11.3    | 0           | 0             | 3.8               | 77.4              | 0            | 0         | 0    | 0         | 0              | 0     | 0        | 0           | 0          | 0         |
| spindle_pole| 2.6       | 7.7     | 12.8          | 0  | 5.1     | 2.6         | 0             | 2.6               | 64.1              | 2.6          | 0         | 0    | 0         | 0              | 0     | 0        | 0           | 0          | 0         |
| endosome     | 11.8      | 2.9     | 11.8          | 0  | 8.8     | 2.9         | 0             | 8.8               | 5.9               | 47.1         | 0         | 0    | 0         | 0              | 0     | 0        | 0           | 0          | 0         |
| late_Golgi  | 18.2      | 0       | 45.5          | 18.2| 0       | 0           | 0             | 0                 | 0                 | 6.1          | 12.1      | 0    | 0         | 0              | 0     | 0        | 0           | 0          | 0         |
| actin        | 14.8      | 0       | 37.0          | 14.8| 0       | 0           | 0             | 3.7               | 0                 | 0            | 0         | 3.7  | 22.2      | 3.7             | 0     | 0        | 0           | 0          | 0         |
| peroxisome  | 4.8       | 0       | 66.7          | 4.8| 0       | 0           | 0             | 0                 | 0                 | 0            | 4.8       | 14.3 | 0         | 0              | 0     | 0        | 0           | 0          | 0         |
| lipid_particle| 0       | 0       | 63.2          | 21.1| 15.8    | 0           | 0             | 0                 | 0                 | 0            | 0         | 0    | 0         | 0              | 0     | 0        | 0           | 0          | 0         |
| Golgi        | 26.7      | 0       | 13.3          | 20.0| 20.0    | 0           | 0             | 0                 | 13.3              | 0            | 6.7       | 0    | 0         | 0              | 0     | 0        | 0           | 0          | 0         |
| bud_neck    | 50.0      | 0       | 28.6          | 0  | 14.3    | 0           | 0             | 0                 | 0                 | 0            | 0         | 0    | 0         | 0              | 0     | 0        | 0           | 0          | 0         |
| early_Golgi | 0         | 0       | 36.4          | 27.3| 0       | 9.1         | 0             | 0                 | 27.3              | 0            | 0         | 0    | 0         | 0              | 0     | 0        | 0           | 0          | 0         |
| microtubule | 30.0      | 20.0    | 0             | 40.0| 0       | 0           | 0             | 0                 | 10.0              | 0            | 0         | 0    | 0         | 0              | 0     | 0        | 0           | 0          | 0         |
| ER_to_Golgi | 16.7      | 0       | 66.7          | 0  | 16.7    | 0           | 0             | 0                 | 0                 | 0            | 0         | 0    | 0         | 0              | 0     | 0        | 0           | 0          | 0         |
3.4 Classification of proteins with ambiguous class

Besides the proteins identified as being in only one location, we applied the classifier on the 156 proteins within the ‘ambiguous’ category. We sought location labels of these proteins in the CYGD database, but the labels in CYGD are all mixtures of ‘ambiguous’ and other categories. The results of automated classification of these proteins are available at http://murphylab.web.cmu.edu/data. An example is shown in Figure 5. ORF YFL034W was classified as ambiguous class in both UCSF and CYGD databases, while our system classifies it as a ER protein with 46.67% confidence. (70 of the 150 classifiers classified it as ER, while 48 classifiers classified it as cytoplasm and 32 classifiers classified it as cell_periphery.) The results for 37 proteins for which all automated classifiers agree are listed in Table 4.

Table 4. Localizations assigned by automated classification for proteins classified as ambiguous in UCSF database. Only proteins for which all 25 classifiers agreed are listed

| ORF          | Automated localization |
|--------------|------------------------|
| YBR007C      | cytoplasm              |
| YBR193C      | nucleus                |
| YBR225W      | cytoplasm              |
| YBR259W      | cytoplasm              |
| YDL025C      | cytoplasm              |
| YDL203C      | cytoplasm              |
| YDR276C      | cytoplasm              |
| YDR319C      | cytoplasm              |
| YDR502C      | cytoplasm              |
| YEL062W      | cytoplasm              |
| YER070W      | nucleolus              |
| YFR048W      | cytoplasm              |
| YGL140C      | cytoplasm              |
| YGL249W      | cytoplasm              |
| YGR149W      | cytoplasm              |
| YHR025W      | mitochondrion          |
| YHR115C      | cytoplasm              |
| YHR138C      | cytoplasm              |
| YIL004C      | cytoplasm              |
| YIL055C      | cytoplasm              |
| YIL160C      | cytoplasm              |
| YJL005W      | cytoplasm              |
| YJR083C      | cytoplasm              |
| YKL176C      | cytoplasm              |
| YLR036C      | cytoplasm              |
| YLR260W      | cytoplasm              |
| YLR262C      | cytoplasm              |
| YLR405W      | cytoplasm              |
| YLR433C      | cytoplasm              |
| YML020W      | cytoplasm              |
| YMR159C      | cytoplasm              |
| YOL031C      | cytoplasm              |
| YOR084W      | cytoplasm              |
| YOR231W      | cytoplasm              |
| YPL202C      | cytoplasm              |
| YPL241C      | cytoplasm              |
| YPL250C      | cytoplasm              |

4 CONCLUSIONS AND DISCUSSIONS

We have presented computational methods to annotate localization categories of yeast proteins in the yeast GFP fusion localization database. This measurement of protein
localization is objective, quantitative and repeatable. Our results have a high accuracy of 81.02% for all proteins and 94.70% precision at 80.22% recall, compared to human labels for those proteins with pure location pattern. The high accuracy confirms the feasibility of automated classification of yeast images.

The most important future work is to seek evidence which explains the differences between the human and computer generated labels. This can be done by looking at the literature for these proteins, checking information from other yeast databases. In addition, there are 1294 images with mixed location patterns. We plan to extend the existing system to handle more than one location pattern in the same cell.

This work can be placed in the context of growing efforts to automate analysis of yeast phenotypes. For example, automated analysis of yeast morphology changes in response to mating pheromone has been described (Narayanaswamy et al., 2006).

The system we have described should be valuable for automating image analysis of yeast images. This work can be placed in the context of growing efforts to automate analysis of yeast phenotypes. For example, automated analysis of yeast morphology changes in response to mating pheromone has been described (Narayanaswamy et al., 2006).

The system we have described should be valuable for automating image analysis of yeast images. This work can be placed in the context of growing efforts to automate analysis of yeast phenotypes. For example, automated analysis of yeast morphology changes in response to mating pheromone has been described (Narayanaswamy et al., 2006).

ACKNOWLEDGEMENTS

We thank Adam Carroll and Erin O’Shea for making the yeast images available. This work was supported in part by NSF grant EF-0331657, by NIH National Technology Center for Networks and Pathways grant U54 RR022241, and by NIH National Center for Biomedical Computing grant National U54 DA021519.

Conflict of Interest: none declared.

REFERENCES

Bengtsson,E. et al. (2004) Robust cell image segmentation methods. Pattern Recogn. Image Anal., 14, 157–167.

Boland,M.V. and Murphy,R.F. (2001) A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. Bioinformatics, 17, 1213–1223.

Boland,M.V. et al. (1998) Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. Cytometry, 33, 366–375.

Campbell,R.E. et al. (2002) A monomeric red fluorescent protein. Proc. Natl Acad. Sci. USA, 99, 7877–7882.

Chen,X.-C. et al. (2006) A novel graphical model approach to segmenting cell images. Proc. IEEE Symp. Comput. Intell. Bioinform. Comput. Biol., (CIBCB ’06), 1079, 1–8.

Chen,X. and Murphy,R.F. (2005) Objective clustering of proteins based on subcellular location patterns. J. Biomed. Biotechnol., 2005, 87–95.

Chen,X. et al. (2003) Location proteomics building subcellular location trees from high-resolution 3D fluorescence microscope images of randomly-tagged proteins. Proc. SPIE, 4962, 298–306.

Cherry,J.M. et al. (1998) SGD: Saccharomyces Genome Database. Nucleic Acids Res., 26, 73–79.

Cortes,C. and Vapnik,V. (1995) Support vector networks. Mach. Learn., 20, 1–25.

Gu¨ldener,U. et al. (2005) CYGD: the comprehensive yeast genome database. Nucleic Acids Res., 33, D364–D368.

Hsi,C.-W. and Lin,C.-I. (2002) A comparison of methods for multi-class support vector machines. IEEE Trans. Neural Net., 13, 415–425.

Huang,K. and Murphy,R.F. (2004a) Automated classification of subcellular patterns in multicell images without segmentation into single cells. Proc. 2004 IEEE Int. Symp. Biomed. Imaging, 1139–1142.

Huang,K. and Murphy,R.F. (2004b) From quantitative microscopy to automated image understanding. J. Biomed. Opt, 9, 893–912.

Huang,K. et al. (2003) Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images. Proc. SPIE, 4962, 307–318.

Huh,W.-K. et al. (2003) Global analysis of protein localization in budding yeast. Nature, 425, 686–691.

Murphy,R.F. et al. (2000) Towards a systematics for protein subcellular location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. Proc. Int. Conf. Intell. Syst. Mol. Biol., 8, 251–259.

Murphy,R.F. et al. (2003) Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. J. VLSI Signal. Proc., 35, 311–321.

Narayanaswamy,R. et al. (2006) Systematic profiling of cellular phenotypes with spotted cell microarrays reveals mating-pheromone response genes. Genome Biol., 7, R6.

Rodenacker,K. and Bischoff,P. (1990) Quantification of tissue sections: graph theory and topology as modelling tools. Pattern Recognit. Lett., 11, 275–284.

Tsien,R.Y. (1998) The green fluorescent protein. Ann. Rev. Biochem., 67, 509–544.

Velliste,M. and Murphy,R.F. (2002) Automated determination of protein subcellular locations from 3D fluorescence microscope images. Proc. IEEE Int. Symp. Biomed. Imaging, (ISBI-2002), 867–870.