# Automatic Classification of M-FISH Human Chromosome Images using Fuzzy Classifier and Statistical Classifier

**B. Baheti[1], G. Ahuja[2], and A. Parode[3]**

[1] Shri Guru Gobind Singji Institute of Engg. and Technology, Nanded, Maharashtra
[2] Amazon Web Services Inc., USA
[3] iNautix Technologies India Pvt Ltd, Pune, Maharashtra

{bahetibhakti@gmail.com, ahuja.gulsheen@gmail.com, parodeashwini@gmail.com}

**Abstract.** Identification of abnormalities in the chromosomes is a tedious job. Conventionally gray-scale imaging was used for chromosome analysis and was based on features like relative length, banding pattern, centromere position. As an alternate, Multiplex fluorescence in-situ hybridization (M-FISH) is a combinatorial labeling technique which does not require these features and has many advantages over the conventional method. Our contribution includes implementation of joint segmentation classification technique with a) high classification accuracy, b) low computational complexity and c) high speed for automated Karyotyping of M-FISH chromosome images. We show effect of image preprocessing on classification accuracy. We propose the use of univariate approach instead of multivariate in Fuzzy and Statistical Classifier for pixel-by-pixel joint segmentation classification which results in significant reduction in computational and time complexity with increased accuracy. The overall classification accuracy with Fuzzy and Statistical classifier is 96.47 % and 97.32 % respectively.

**Keywords:** ***M-FISH, Expectation Maximization, Fuzzy, Bayes***

## 1 Introduction

Chromosomes are the structures in cells that contain genetic information and are basic building blocks of life. These genetic structures contain important information about health of an individual. Chromosome image analysis is important in the study of genetic disorders as well as cancer. Normally, in humans there are 46 chromosomes of 24 distinct classes including 22 homologous pairs and 2 gender determining chromosomes X and Y. Chromosome karyotyping is the process of classification of chromosomes in a cell according to standard nomenclature. Chromosome analysis is of vital importance for detection of abnormalities which include unusual number of chromosomes, deletion of a part of chromosome, duplication of genetic material within a chromosome, translocations where genetic information is exchanged between two chromosomes etc [1]. Detection of these abnormalities is vital as they are reliable indicators of genetic diseases and damage. Their study and analysis can lead to new insights about these disorders.

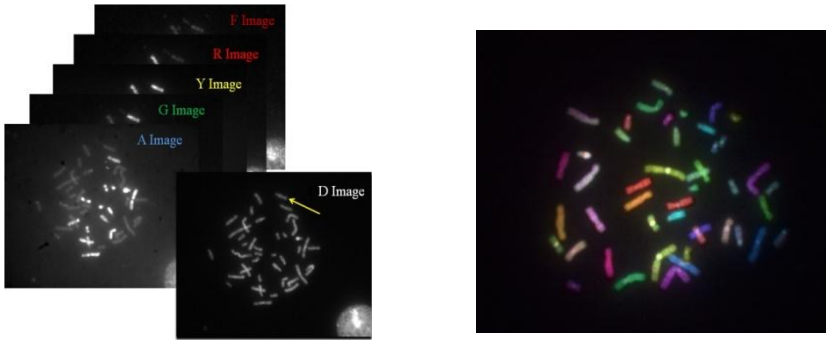
Conventional methods of classifying chromosomes involved manual or semiautomatic image segmentation of grayscale images. Since, the manual karyotyping is a difficult, tedious and time consuming process; it has motivated many medical image processing and computer vision researchers to investigate automatic karyotyping techniques to make it faster and more accurate. Conventional automatic karyotyping methods are based on grayscale images that use features like relative length, banding pattern and centromere position of chromosomes. But these methods involve possibility of error and misclassification [2]. On the other hand, M-FISH image analysis provides an efficient and more accurate method for analysis of chromosome images and identification of abnormalities in them. This enables karyotyping to be completely automated using digital image processing techniques.

The remainder of the paper is organized as follows. Section 2 elaborates concept of M-FISH imaging and the chromosome labeling chart. It also briefs about the previous work in M-FISH image analysis. Technical approach used for automatic classification of chromosomes is explained in Section 3. In section 4, results of various stages are discussed. Section 5 presents the conclusion and section 6 presents acknowledgement.

B. Iyer, S. Nalbalwar and R. Pawade (Eds.)
ICCASP/ICMMD-2016. Advances in Intelligent Systems Research.
Vol. 137, Pp. 557-564.

## 2 What is M-FISH?

Multispectral imaging allows extraction of additional information that human eye fails to capture. Multiplex fluorescence in-situ hybridization (M-FISH) is a combinatorial labeling technique used for chromosome analysis. M-FISH images are acquired by microscopy under CCD camera.




(a) M-FISH Image set (6 Images) (b) Pseudo-colored M-FISH Image

**Fig.1.** M-FISH Images

**Table 1:** Vysis M-FISH Chromosome Labeling Chart

| Chromosome Class | DAPI | Aqua | Green | Yellow | Red | Far Red |
|---|---|---|---|---|---|---|
| 1 | x | | | x | | |
| 2 | x | | | | x | |
| 3 | x | X | | | | |
| 4 | x | | x | | x | X |
| 5 | x | | | x | | X |
| 6 | x | | x | | | |
| 7 | x | | | | | X |
| 8 | x | | | | x | X |
| 9 | x | | | x | x | |
| 10 | x | X | | x | | |
| 11 | x | X | | | x | |
| 12 | x | | x | x | | |
| 13 | x | X | x | | | |
| 14 | x | | x | x | x | |
| 15 | x | X | | x | x | |
| 16 | x | | x | | | X |
| 17 | x | | x | | x | |
| 18 | x | | | x | x | X |
| 19 | x | | x | x | | X |
| 20 | x | X | | | x | X |
| 21 | x | X | x | x | | |
| 22 | x | X | x | | x | |
| X | x | X | | | | X |
| Y | x | X | | x | | X |

This multispectral staining technique uses 5 color dyes (also called fluorophores or color channels) namely, Aqua, Green, Yellow, Red and Far-red, that attach to various chromosomes differently. DAPI is a counter staining

dye that is absorbed by all the 24 classes of chromosomes. Fig. 1 (a) represents an M-FISH image set consisting of 6 images and Fig. 1 (b) is a pseudo colored chromosome image in which each color signifies a different class. Each class of chromosome absorbs a unique combination of dyes and based on which association of a pixel to a class can be determined. Thus it is possible to envision new and improved methods for the location and classification of chromosome images by exploiting the color information in M-FISH images. For example, class 5 chromosomes absorb only Dapi, Yellow and Far-red fluorophores. So they should be visible in only images corresponding to Yellow and Far-red dyes and not in others. The standard VYSIS chromosome labelling chart tells us what combinations of the six fluorophores are absorbed by different classes of chromosomes and is shown in Table 1. It is used for classifying each pixel into one of the 24 classes.

Since past few years, various methods for M-FISH chromosome classification are being explored. In 2002, Mehul P. Sampat et al. [3] proposed multivariate bayes classifier for automatic chromosome classification with pixel-by-pixel approach. The classification was approached as a 25 class (24 chromosome classes + background) 6 feature pattern recognition problem and the overall accuracy reported is 95% considering only non-overlapping cases. M. P. Sampat et al. has also reported and compared classification by nearest neighbor, K-nearest neighbor and Maximum-Likelihood estimation methods. The highest classification accuracy was achieved with the K-nearest neighbor method with k=7 [4]. Alan C. Bovik et al. [5] suggested a pixel classification and a probabilistic model of chromosome features to select from among a set of segmentation possibilities. Since the model was function of both, segmentation and classification can be achieved simultaneously. Average chromosome classification accuracy obtained was 68% with standard deviation of 17.5%. Hyohoon Choi et al. [6] explained an unsupervised classification method based on fuzzy logic classification and a prior adjusted reclassification. It reported an increase in overall classification accuracy. A watershed based method for segmenting the chromosomes was proposed by Petros S. Karvelis et al. [7]. They implemented a region based classifier to classify the segmented chromosomes and the results were evaluated on a publicly available chromosome database. The reported overall accuracy is 82.4 % and the time required for classification is 36.4s (±10.9s) on a Pentium P4 2-GHz PC, with 512 MB RAM. Hyohoon Choi et al. [8] proved that a significant improvement is achieved on pixel classification accuracy after performing a new technique called feature normalization. He reported an increase of 20% in overall pixel classification accuracy. Authors also explained the algorithm for removal of non-flat background from the M-FISH images [9]. The paper discussed needs and effects of background correction and also reports an improvement in classification accuracy after Background correction.

The approaches reported in the literature survey use multivariate analysis techniques for classification of M-FISH chromosome images which increases the computational complexity of the classifiers as simultaneous processing on five images is complex in space and time. This motivated us to formulate a classification technique which exploits advantage of multivariate data and univariate analysis to reduce computation and time complexity of M-FISH image analysis maintaining good classification accuracy and we could achieve it with the following approach.

## 3 Technical Approach

Main steps in M-FISH chromosome image analysis include Image Pre-processing, Feature Normalization and Classification.

### *3.1 Image Pre-processing*

M-FISH images in their original form are not appropriate for processing [10]. We expect that in a color channel, only those chromosome pixels which are sensitive to that particular dye should be bright and background should be perfectly black but practically it is not so. Factors like non-homogeneity of staining, overlap between emission spectra of fluorophores, DC offset of the CCD device, autofluoroscence of the slide and noise added during image capturing process give rise to the non-flat intensity elevation of the background. Hyohoon Choi et al. [9] proposed a signal model as shown in eq. 1 to recover the true signal by removing elevated background from the observed signal.

$$y = E\{Cx + b\} + n \qquad (1)$$

where,

$y$ : Observed signal

$x$ : 6 × 1 vector of true signal

$b$ : DC offset of the device and various factors causing background elevation

$n$ : noise of the imaging device
$E$ : 6 × 6 diagonal matrix of exposure times
$C$ : 6 × 6 color spread matrix

Using the approach in [9], 2-D cubic background surface is estimated and subtracted from the observed signal to obtain background corrected signal.
After background correction, DAPI image is used to segment background and foreground to separate chromosome pixels for further processing. This is achieved by performing edge detection using Laplacian of Gaussian (LoG) operator followed by morphological operations.

## *3.2 Feature Normalization*

Each dye or fluorophore has different sensitivity towards the excitation wavelength and hence different integration times. This uneven hybridization and unequal fluorophore sensitivities cause intensity variations among chromosomes in a channel and also between channels. In M-FISH image analysis, relative intensities across six channels is the only feature used. These features i.e. intensity distribution must be normalized before classification to avoid misclassification.
An M-FISH image is composed of six gray-scale images $\{I_1, I_2 \ldots I_6\}$ and each gray-scale image $I_m$ has intensity values $y$ that belong to either chromosomes $I_c(m)$ or the background$I_b(m)$. The distribution of $I_c(m)$ is assumed to be mixture of two Gaussians:

$$p(y|C_1) \sim N(\mu_1, \sigma_1) \ \ and \ \ p(y|C_2) \sim N(\mu_2, \sigma_2)$$

where,

$C_1$ is intensity due to non-fluorophore and $C_2$ is intensity due to fluorophore and $\mu_i$, $\sigma_i$ are the mean and variance of respective classes.
One might question that what significance of non-fluorophore class is after foreground segmentation. Consider a color channel, say Aqua, there are some particular classes of chromosomes that are sensitive to that fluorophore and are expected to be bright in that channel. Remaining chromosome classes should not absorb the fluorophore at all. But practically the remaining classes also show some sensitivity to that dye and have non-zero intensities. So non-fluorophore class refers to those chromosome pixels which were expected to be dark in that channel but are actually not and hence care should be taken to avoid misclassification.
So, $I_c(m)$ is a set of unlabeled samples independently drawn from mixture density shown in eq. 2.

$$p(y) = p(y|C_1)p(C_1) + \ p(y|C_2)p(C_2) \qquad (2)$$

Parameter vector $\Theta$ contains $(\mu_1, \sigma_1, p(C_1), \mu_2, \sigma_2, p(C_2))$ where $p(C_1)$ and $p(C_2)$ are prior class probabilities.
The normalization process should find the decision boundary between classes $C_1$ and $C_2$ from the marginal density function and its parameters. The unknown parameters of this function are extracted using Expectation-Maximization algorithm as proposed by Hyohoon Choi [8]. To increase the convergence rate of Expectation-Maximization, kmeans clustering is used to initialize parameters of EM.
After the parameters of distribution are estimated, the decision boundary between $C_1$ and $C_2$ is found by eq. 3.

$$T = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \qquad (3)$$

where,

$$A = \sigma_2^2 - \sigma_1^2$$

$$B = 2\sigma_1^2\mu_2 - \ 2\sigma_2^2\mu_1$$

$$C = \sigma_2^2\mu_1^2 - \ \sigma_1^2\mu_2^2 - 2\sigma_2^2 \ln\left(\frac{\sigma_2 p(C_1)}{\sigma_1 p(C_2)}\right)$$

Once we have the parameters and the decision boundary i.e. threshold, the original sample distribution is normalized by linear piece-wise transformation function. The input intensity $r$ is mapped to output intensity $s$ as shown in eq 4.

$$f(r) = \begin{cases} \frac{64}{(\mu_1 - \min(r))}(r - \min(r)), & \min(r) \le r < \mu_1 \\ \frac{64}{(T - \mu_1)}(r - \mu_1) + 64, & \mu_1 \le r < T \\ \frac{64}{(\mu_2 - T)}(r - T) + 128, & T \le r < \mu_2 \\ \frac{63}{(\max(r) - \mu_2)}(r - \mu_2) + 192, & \mu_2 \le r < \max(r) \end{cases} \quad (4)$$

## *3.3 Classification*

Classification is the process of assigning input pixel data into one of the 24 classes of chromosomes based on feature extraction. From literature survey, it has been noticed that MFISH chromosome images are analyzed in multivariate manner. In that, each of the 5 channels is modeled as a mixture of 24 Gaussians corresponding to 24 classes of chromosomes giving rise to high computational complexity. For example, the mean vector becomes of dimension 24 × 5 (24 : Chromosome classes and 5: Spectral channels). On the other hand, in our proposed univariate approach, each channel is modeled as mixture of only 2 Gaussians (Fluorophore and Non-Fluorophore class). And thus there is drastic reduction in computational complexity, e.g. now mean vector becomes only 5 × 2.
We have implemented classifier with supervised parametric technique using two different approaches.

**Fuzzy Classifier**

We have implemented fuzzy classifier using Sugeno modeling in MATLAB in which output membership functions are constant. There are five inputs to FIS system corresponding to five spectral channels and there are twenty four output membership functions. The rule base has 24 rules for 24 classes of chromosomes. Output of FIS is only one, i.e. class of chromosome. Rules of classification are based on standard Vysis chromosome labeling chart.

**Statistical Classifier using Bayes Theorem**

We know, Bayes theorem is given as,

$$p(C_i|x) = \frac{p(x|C_i) * p(C_i)}{p(x)} \quad (5)$$

where,
$i$: Class, flurophore and non-flurophore
$p(C_i|x)$: Posterior probability
$p(x|C_i)$ : Class-conditional probability distribution function
$p(C_i)$: Prior class probability
$p(x)$ : $\Sigma\, p(x|C_i) * p(C_i)$
$p(x|C_i)$ and $p(C_i)$ are estimated in the training phase.

Using eq. 5, two posterior probabilities for each pixel are obtained and each pixel is assigned as either chromosome pixel or non-chromosome pixel by comparing them. Thus each spectral image is converted into binary image having only two values 0 for non-chromosome pixels and 1 for chromosome pixels. For example, initial pattern [68 205 150 138 10] is now converted to [0 1 1 1 0]. This new feature vector for each pixel is compared pixel by pixel with standard pattern as per given in standard labeling chart which is basis of classification.

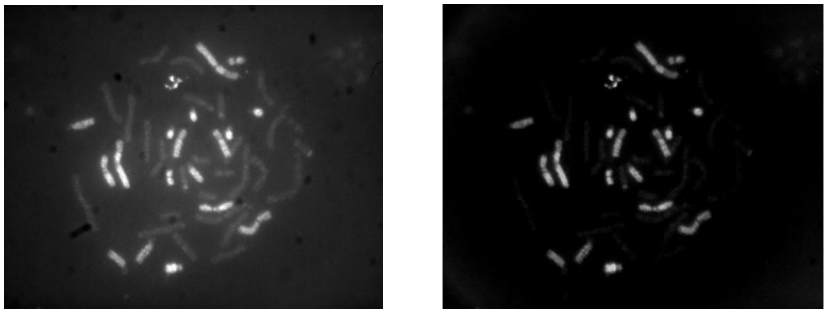
## 4 Results and Discussion

The standard M-FISH database used for experimentation is obtained from Advanced Digital Imaging Research (ADIR) Laboratory. The image set contains 200 M-FISH images. The images are divided into directories based on Karyotype and slides. The first character in the directory name represents the probe set (A, ASI; P, PSI; V, Vysis).

We are working on Vysis probe dataset. One MFISH image set consists of seven images. In addition to six fluorophore images there is another file that contains ground truth of correct Karyotype. Images are of size 517 × 645 and the format is .png.
The experimentation is performed on number of M-FISH chromosome image sets which include male and female chromosomes, normal and abnormal as well as touching and overlapping cases. Seven datasets, each containing six images is used for training of classifier and testing is done on nineteen datasets totaling 114 images.

**Result of Pre-Processing**

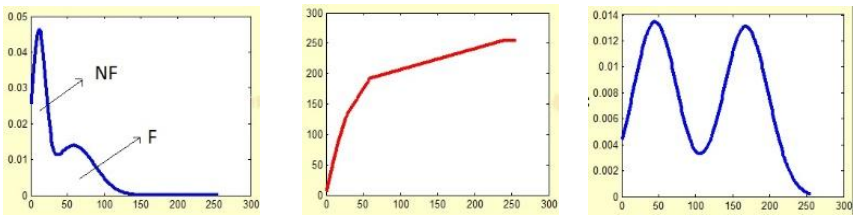
As explained in earlier section, the elevated background surface is modeled as 2D cubic surface and is subtracted from the original image to achieve significant noise reduction and uniform intensity background. Fig. 2(a) shows the input image. As we can see, chromosome pixels are bright but background also shows some non-uniform non-zero intensities. This non-flat background surface is different for each image and hence separately estimated for each image. Fig. 2(b) is background corrected image i.e. image with uniform background.
After Background correction, next step is feature normalization. As we can see in fig. 2(b), along with the chromosomes which are expected to be bright in a



(a) Before BC (b) After BC

**Fig.2.** Result of Background Correction

channel image, other chromosomes also show some sensitivity and hence some non-zero intensities. Intensity distribution of every image is different and so the transfer function has to be estimated separately for each image. Fig. 3 shows an example. Original data distribution is shown in fig. 3(a). First Gaussian corresponds to non-fluorophore class and second to fluorophore class. As it can be observed, the boundary between these two classes is not clear. They have quite overlap which leads to classification error. So, this original distribution is normalized using an image specific piecewise linear mapping function in fig. 3 (b). Fig. 3 (c) shows normalized data in which non-fluorophore and fluorophore classes are distinct and hence probability of misclassification is reduced. It can be observed that, intensity values in initial distribution were ranging from 0 to 150. After normalization, they range from 0 to 255 with a clearer boundary between two classes. These results are of Gold channel of V270259 image set.




(a) Original Distribution (b) Transfer Function (c) Normalized Distribution

**Fig.3.** Result of Feature Normalization

**Classifier Accuracy**

Accuracy of the Fuzzy and Statistical classifier can be calculated by comparing our automatic karyotype image pixel by pixel with the reference karyotype image provided in the MFISH database. Table 2 presents the accuracy comparison for eleven commonly cited datasets in the literature. It can be observed that overall classification accuracies of the fuzzy and statistical classifier with univariate approach are higher than that of the multivariate classifiers by Hyohoon Choi et al. [8].

**Table 2.** Comparison of Overall Accuracies for Commonly Cited Datasets

| Image Set | ML Classifier | MD Classifier | Fuzzy Classifier | Statistical Classifier |
|---|---|---|---|---|
| V1301XY | - | 90.81 | 95.91 | 97.36 |
| V1302XY | - | 92.99 | 97.20 | 98.04 |
| V1303XY | - | 92.77 | 96.66 | 97.69 |
| V1305XY | - | 94.72 | 97.08 | 98.25 |
| V1306XY | - | 94.66 | 96.42 | 97.49 |
| V1308XY | 96.49 | 96.28 | 97.36 | 98.19 |
| V1309XY | 86.29 | 84.50 | 96.63 | 97.76 |
| V1310XY | 86.90 | 84.19 | 97.19 | 98.02 |
| V1311XY | 94.54 | 94.50 | 97.69 | 98.63 |
| V1312XY | 95.20 | 94.83 | 96.58 | 97.49 |
| V1313XY | 94.88 | 94.53 | 96.96 | 97.97 |
| **Average** | **92.38** | **92.25** | **96.88** | **97.89** |

ML: Maximum Likelihood and MD: Minimum Distance

Also, the approaches from literature survey use multivariate analysis technique for classification which gives rise to higher computational complexity as simultaneous processing on five images is complex in space and time. In multivariate analysis, a total of 24 additions and 30 multiplications are required for each pixel and class. The image size being 517 × 645 and number of chromosome classes is 24, leads to almost 240 million multiplications/additions for each image and takes around 2.5 minutes on a 167 MHz Sun workstation to accomplish both segmentation and classification [5]. With our univariate approach, only 3 additions and 10 multiplications are needed per pixel and thus number of computations reduces to just 1.3 million. It takes around 1.2 minutes to run on a 2.4GHz machine having 3GB RAM, Intel i3 processor. Thus we have achieved better classification accuracy with lower computational complexity.

## 5 Conclusion

We have proposed an automated chromosome classification technique and results are tested on various cases that include male and female, normal and abnormal as well as touching and overlapping chromosomes. Removal of elevated background and normalization of feature distribution significantly increase the classifier accuracy.
We have proposed univariate analysis technique of the M-FISH data which exploits advantage of multivariate data as well as univariate analysis using Fuzzy and Bayes classifier. It not only reduces the computation and time complexity but also achieves better classification accuracy. There is approximately 4 % to 5% increase in overall classification accuracy with fuzzy and statistical classifiers respectively compared to other approaches. Also, drastic reduction in computational and time complexity is achieved.
Currently, classification accuracy for overlapping chromosome cases is less than that of normal cases. In future, we will work towards handling overlapping chromosome cases with increased accuracy.

## Acknowledgement

We are thankful to Department of E&TC, Pune Institute of Computer Technology, Pune for the all-round support and cooperation during the project work. We would also like to extend our sincere thanks to Dr. Sanjay N. Talbar from SGGSIE&T, Nanded for his valuable encouragement and support.

## References

[1] George, I., Arisaka,O.: Chromosome Analysis Using Spectral Karyotyping (SKY), Cell Biochemistry and Biophysics. 62.1 (2012)

[2] Moradi, M., Setarehdan, S.: New features for automatic classification of human chromosomes: A feasibility study, Pattern Recognition Letters, 27, 1, 19-28, (2006)

[3] Sampat, M., Castleman, K., Bovik, A.: Pixel-by-pixel classification of MFISH images, IEEE International Conference on EMBS/ BMES, 2, 999-1000, (2002)

[4] Sampat, M., Bovik, A., Aggarwal, J., Castleman, K.: Supervised parametric and non-parametric classification of chromosome images, The Journal of Pattern Recognition Society on Pattern Recognition, 38, 8, (2005)

[5] Schwartzkopf, W., Bovik, A., Evans, B.: Maximum-likelihood techniques for joint segmentation-classification of multispectral chromosome images, IEEE Transactions on Medical Imaging, 24, 12, 1593-1610, (2005)

[6] Choi, H., Castleman, K., Bovik, A.: Segmentation and Fuzzy-Logic Classification of M-FISH Chromosome Images, International Conference on Image Processing, Atlanta, GA, 69-72 (2006)

[7] Karvelis, P., Tzallas, A., Fotiadis D., Georgiou, I.: A Multichannel Watershed-Based Segmentation Method for Multispectral Chromosome Classification, in IEEE Transactions on Medical Imaging, 27, 5, 697-708, (2008)

[8] H. Choi, A. C. Bovik and K. R. Castleman: Feature Normalization via Expectation Maximization and Unsupervised Nonparametric Classification for M-FISH Chromosome Images, IEEE Transactions on Medical Imaging, 27, 8, 1107-1119, (2008)

[9] Choi, H., Castleman, K., Bovik, A.: Color Compensation of Multicolor FISH Images, in IEEE Transactions on Medical Imaging, 28, 1, 129-136, (2009)

[10] Li, J., Lin, D., Cao, H., Wang, Y.: An improved sparse representation model with structural information for Multicolor Fluorescence InSitu Hybridization (M-FISH) image classification, BMC Systems Biology, (2013)