Automatic Analysis of Fluorescence In-Situ Hybridisation Images

William Clocksin and Boaz Lerner
Computer Laboratory
University of Cambridge
Cambridge CB2 3QG , UK
wfc@CL.cam.ac.uk

Abstract

Fast and accurate analysis of fluorescence in-situ hybridisation (FISH) images depends upon two components: a classifier to discriminate between artefacts and valid signal data, and well-discriminating features to represent the signals. After processing the image, we evaluate candidate feature sets by illustrating the probability density functions and scatter plots for the features. This analysis indicates the relative importance of members of a feature set, helps in identifying sources of potential classification errors and recommends several size, intensity and hue-based features for representing FISH signals. The recommendation is assessed by the probability of misclassification of a neural network based hierarchical strategy, and also by a feature selection technique making use of a class separability criterion. Represented by these features, nearly 90% of valid signals and artefacts within a set of 400 test images are correctly classified.

1 Introduction

Fluorescence in-situ hybridisation (FISH) allows the detection of specific DNA sequences in intact cells and chromosomes. It enables selective staining of various sequences in interphase nuclei and therefore the detection, analysis and quantification of specific numerical and structural chromosomal abnormalities within these nuclei. For example, trisomy (triplication) of chromosomes 13 and 21 is associated with Patau Syndrome and Down's Syndrome respectively. FISH is a widespread and diversely applied technology that is employed in many fields such as karyotype analysis, gene mapping, and clinical diagnosis of disease [2].

FISH images result from the fluorescence of three dyes: one for the cell nucleus and two for DNA hybridisation dots (for example, associated with chromosomes 13 and 21). To estimate the distribution of chromosomes per cell, it is necessary to inspect a large number of cells, particularly when the frequency of abnormal cells is low. Dot counting, the enumeration of signals (also called dots or spots) within the nuclei, is considered as one of the most important applications of FISH, yet there has been little progress in automating this task. Analysis of FISH imagery could be useful for the automation of this laborious and tedious screening task. Dot counting that relies

BMVC 2000 doi:10.5244/C.14.67

on the conventional approach of using an auto-focusing mechanism [8] suffers from a few shortcomings [6]. Therefore, we base FISH dot counting on images that are sampled at a fixed focal plane. This method is motivated by the assumption that nuclei are approximately uniformly distributed in the sample, so that translations at a fixed focal plane will provide a statistically equivalent sample as projections through different focal planes. This method overcomes most of the shortcomings of auto-focusing, but it relies on the acquisition of sufficient analysable images and more intensive image analysis. Dealing with many unfocused nuclei and signals, the system needs an improved discrimination capability between focused and unfocused signals. Therefore, the system described here is based on extracting well-discriminating characteristics of focused and unfocused signals and on a highly accurate classifier, trained using large numbers of examples of the two classes. In this paper, feature representations of signals are evaluated and highly accurate neural net classification strategies are developed to ensure an efficient automatic signal classification in FISH images. The system that implements the image processing and classification described here is written entirely in MATLAB, and the system has a graphical interface to permit use by a cytologist [7].

2 Image Acquisition

FISH slides were prepared, hybridised and viewed according to [5]. A total of 400 RGB images were collected from five slides and stored in TIFF format; a typical image is shown in Fig 1. During the preparation of the sample (in this case amniotic fluid) three different fluorophores are combined: chromosomes 13 and 21 are indicated by green and red signals respectively, whereas the nuclei are indicated by blue. An image contains one or more large blue blobs (cell nuclei), and a number of small spots or dots called signals. Dots can indicate the presence of chromosomes, or may be the result of artefacts such as overlapping signals, background fluorescence and contaminants.

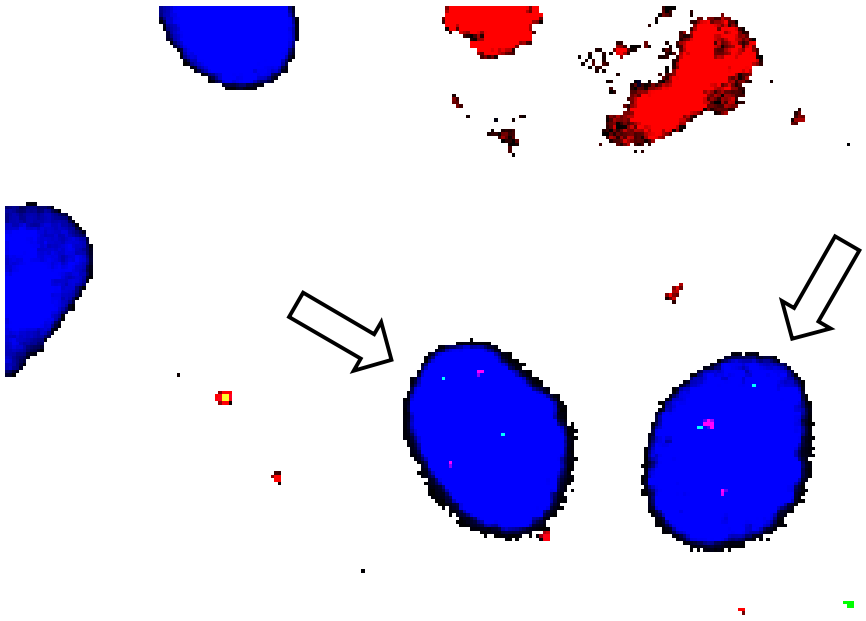


Fig 1. Monochrome version of a typical FISH image showing two complete nuclei (arrows) each with a pair of signals indicating chromosomes 13 and 21. Regions outside the nuclei are artefacts and nuclei partly outside the image. The normally black background is here rendered white to improve clarity.

Image acquisition is purely multichannel: each of the R, G, B channels is exposed separately, and there is negligible crosstalk between colour channels because the chromatic characteristics of the fluorophores are matched to the filters used during image capture. Therefore, nuclei can be segmented using the blue channel of the RGB image exclusively, whereas red and green regions can be segmented using the red and green channels, respectively. Multi-spectral FISH image analysis is better than the conventional monochrome-based analysis not only in facilitating pre-processing and segmentation, but also in yielding colour-based features that contribute to an efficient signal representation. To avoid the use of multiple features to distinguish individual colours, observe that signals of different fluorophores represented by the hue parameter of the HSI (hue, saturation, intensity) format can be easily resolved due to their different hues. Therefore, we use the RGB format for nucleus and signal pre-processing and segmentation and then characterise the signal hue using only the H channel of the HSI format. To convert RGB to HSI format, we follow standard methods such as given by Castleman [3].

Colour information is not used explicitly because in the FISH preparations there is no valid interpretation for colours other than R, G, B. For example, R and G dots can overlap owing to close proximity of chromosomes, possibly at different depths in the sample. Each of the overlapping signals can be determined from the R and G channel respectively; the yellow colour at the overlapped region is not informative about anything, apart from indicating that an overlap of signals has occurred.

3 Signal Measurement and Feature Selection

The first step is to find connected regions of pixels indicating the nucleus and signals. Each of the R, G, B channels is treated separately to find connected regions of pixels which define blobs for red signals, green signals and nuclei respectively. Because real signals have sharply rising edges, we use a global threshold of 90% of the maximum channel value for the criterion for including a pixel in a region. This unusually high threshold is effective at suppressing some artefacts such as background fluorescence, which are characterised by slowly varying channel values, while maintaining the integrity of signal boundaries.

Following this, a discriminating and compact representation for the signals is derived by measuring a set of features for each region. The set analysed here uses eight features as shown in Table 1. We compute, at the colour plane corresponding to the type of region, five RGB intensity-based measurements: area, eccentricity, the total and average intensities and the intensity standard deviation. We also compute two hue-based measurements: maximum and average hue. Finally, we measure the average intensity, essentially the monochrome brightness grey-level, which other colour image interpretation tasks [9] have favoured. This feature is similar to the projection of the image onto the eigenvector corresponding to the largest eigenvalue, but the requirements for its computation are negligible compared with those required in performing principal component analysis.

Number	Feature
1	Area of region *
2	Eccentricity of region *
3	Total Intensity of region *
4	Average intensity of region *
5	Intensity standard deviation of region *
6	Maximum Hue in region
7	Average Hue in region
8	Average Intensity (R+G+B)/3 of region

Table 1. Features extracted from regions of FISH images. Features marked with * relate to the single RGB plane corresponding to the region (blue=nucleus, red=chromosome 21, green=chromosome 13).

Once a sufficient set of features is measured, we can use each one or even all of them to classify signals into 'reals' (valid signals) and 'artefacts'. However, the best single feature may not be sufficiently discriminating for an accurate classification, and classification based on whole or most of the set may be complex, costly to compute, and inaccurate. Moreover, some of the features can be found to contribute very little to the classification accuracy. Therefore, the purpose of feature selection is to select a (small) subset of the feature set that yields an accurate classification in minimal computational cost. In practical problems and for a not very large feature set, we can search among all the possible feature subsets and evaluate each one of them using a criterion of class separability. The subset that achieves the highest value of the criterion is then selected to represent the patterns to the classifier.

The criterion of separability that is considered here is based on the within-class scatter matrix (Fukunaga, 1990)

$$S_w = \sum_{i=1}^{L} P_i E\left\{ (X - M_i)(X - M_i)^T \mid \omega_i \right\} = \sum_{i=1}^{L} P_i \Sigma_i$$

and the between-class scatter matrix

$$S_b = \sum_{i=1}^{L} P_i (M_i - M_0)(M_i - M_0)^T$$

where $M_0 = E\{X\} = \sum_{i=1}^{L} P_i M_i$ is the mean pattern of the mixture distribution. The $X \mid \omega_i$ are patterns of class ω_i (i = 1,L) with mean M_i, covariance matrix Σ_i, and prior probability P_i. The criterion $J_1 = tr\left(S_w^{-1} S_b\right)$, where *tr(A)* is the trace of matrix A, is expected to be larger when the between-class scatter matrix is larger and/or the within-class scatter matrix is smaller.

4 Signal Classification

Signals are classified into four classes – ‘real red’, ‘artefact red’, ‘real green’ and ‘artefact green’. Within the ‘artefact’ classes we expect to find mostly unfocused and overlap signals, and signals that are the result of background fluorescence. These signals will have patterns with different values of features than those of real signals, and hence will be classified as artefacts.

Before performing the experiments, the features are normalized to zero mean and unit variance. Patterns are divided randomly into training and test sets and classification into one of the four classes is implemented using cross-validation. The classifier is a two-layer neural net trained by the scaled conjugate gradient algorithm [1]. A validation set which is drawn from the training set assures that the classifier is not over-trained. It also allows the selection of a minimal network configuration based on only a few hidden units. Both factors ensure rapid training and improved generalization.

For the examples described in this paper, J_1 was computed for all combinations of three features, using all data, and feature sets with the highest J_1 values were used as input to various neural networks. In other work [5] we have chosen the optimal number of features automatically based on the probability of misclassification by the neural net.

Two of the three classification strategies of [5] are examined here. In the Simple classifier, patterns are classified into the four classes using a single neural network. The Hierarchical classifier uses three networks. Patterns are first classified into red and green classes using the colour network and then based on the results of this network they are classified by two other networks into real signals and artefacts of the two colours. The two classifiers are shown in Figure 2.

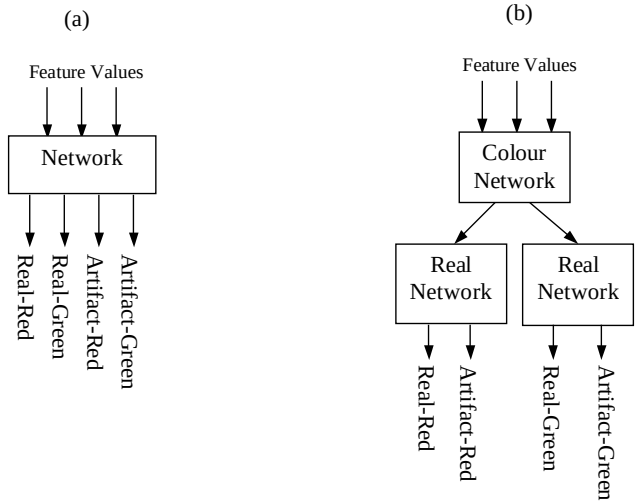
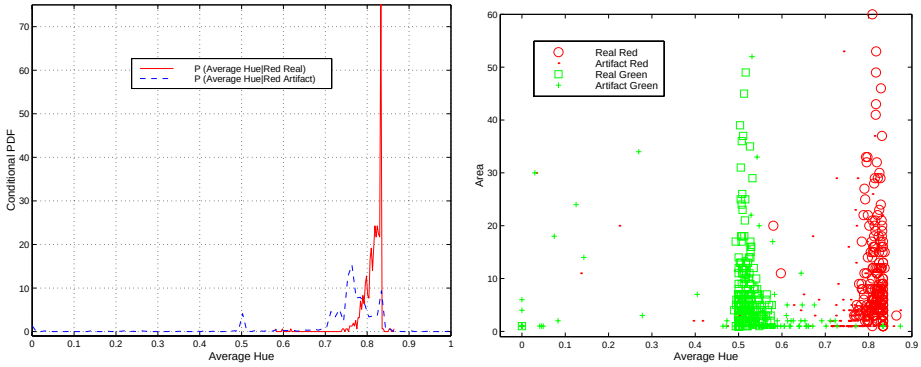


Figure 2. (a) The Simple classifier uses a single network; (b) The Hierarchical classifier uses a network to distinguish colours and two networks for distinguishing real signals from artefacts.

5 Experiments and Results

We created a database of 400 FISH images, which were captured from five slides. Following nuclei segmentation, the system identified 944 objects within these images as nuclei, of which 613 also contained signals. Following signal segmentation, 3,144 objects within the above nuclei were identified as potential signals and features were measured for them. Based on labels provided by expert inspection, 1,145 of the signals were considered as 'reals' (among them 551 were red) and 1,999 as 'artefacts' (among them 1,224 were red).



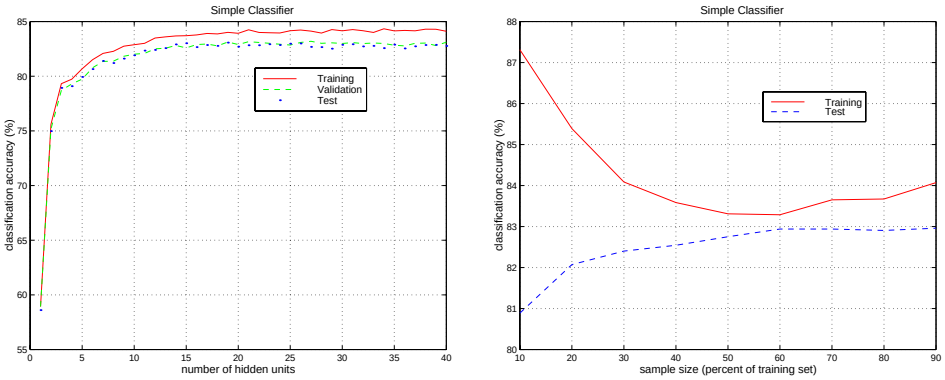
(a) (b)

Figure 3. (a) Histogram estimate of the one-dimensional pdf of the signal average hue for red signals. (b) A scatter plot for the signal average hue and area.

Features were first analysed visually using conditional probability density functions (pdfs) and scatter plots. Figure 3a shows an example of histogram estimate of one-dimensional conditional pdf of the average hue for red signals. The figure indicates moderate overlap between distributions of real signals and artefacts. Similar graphs derived for the other classes and features show different extents of overlap between distributions and demonstrate potential difficulty in classifying signals into reals and artefacts of two colours based on a single feature. The example of the scatter plot in Figure 3b strengthens the role of the average hue in colour discrimination. Based on the complete visual analysis, it was found that among the eight features of Table 1, average channel intensity, maximum (or average) hue and average intensity R+G+B/3 each provided reasonable discrimination capability between the four classes.

Before evaluating sets of features using the classification accuracy, we performed some experiments to find suitable configurations for each of the classification strategies. Input and output dimensions for each of the neural net classifiers are set by the feature space dimension and the number of classes, respectively. The number of hidden units is determined such that the network has the highest generalization capability. This is achieved by evaluating networks of different numbers of hidden units on an independent validation set drawn from the training set. The network which has the lowest error measured on the validation set is selected for training. Figure 4a shows an example of an experiment for determining the number of hidden units of the Simple classifier classifying signals represented by the area, average intensity and average hue (features 1, 4, 7 in Table 1, respectively). Finally, training of each of the

networks was continued for 100 epochs and the results were averaged for each network over three random initializations using the cross-validation (CV-5) technique.



(a) (b)

Figure 4. Success rates of the Simple classifier for features 1, 4, 7 for increasing (a) numbers of hidden units and (b) sample size.

We also examined the sensitivity of the success rate against the sample size by repeating the experiment for training sets of different sizes. The size of the training set is increased from 10% to 90% of the data, where the same unseen 10% of the data are used for the test. The results in Fig. 4b for the Simple classifier demonstrate that the success rate on the test set follows, as expected, the increase of the training sample size until its maximum level. However, the success rate on the training set has a minimum. The explanation is that for a very small sample size, training is very simple and classification of a few training patterns can be very accurate. It is more difficult to maintain this accuracy as the sample size increases and more variants of the training patterns are added. The success rate decreases until it reaches a minimum for a critical mass of learned patterns. After this point, as sample size continues to grow, the additional patterns are not so different from those of the critical mass. Thus, learning of the patterns of the (extended) critical mass is intensified, while at the same time the fraction of misclassified patterns becomes lower. The result of both trends is towards the improvement of the success rate on the training set as is shown in Fig. 4b.

Finally, we evaluated feature sets chosen from the features of Table 1 using the classification accuracy and criterion J_1. Each classifier, classifying signals represented by each of the feature sets, was checked using its own optimal configuration. The classifier configurations and accuracies for four feature sets are shown in Table 2. Here we see that unseen signals, represented by different combinations of features, can be classified as reals or artefacts of two colours with accuracies higher than 80%. In addition, the Hierarchical network is found to be a better classifier, even when inferior feature sets are employed.

Since only eight features are included in this investigation, we can allow exhaustive search for the best (according to criterion J_1) subset of, say three features. This search is done quickly since it involves the evaluation of only 56 subsets. Values for J_1 and the corresponding ranks for the four feature sets are also given in Table 2. These results and similar experiments with other feature sets demonstrate that feature sets consisting of the area (1), average intensity (4), average hue (7) and the average intensity (8) provide the best representations for FISH signals. The experiments also

show that hue-based features are crucial for separating signals of the two colours, while size and intensity-based features are essential for separating real signals from artefacts.

Features	J_1	Rank	Simple Neural Network			Hierarchical Neural Network		
			Hidden	Training	Testing	Hidden	Training	Testing
4, 7, 8	1.7543	1	7	79.0	78.2	3, 11	81.9	81.4
4, 5, 6	1.6789	12	16	79.0	77.3	12, 7	82.4	81.3
1, 4, 7	1.4958	10	15	84.3	83.0	1, 14	88.3	87.5
1, 4, 8	0.4342	42	9	56.8	54.9	1, 4	89.5	89.0

Table 2. Evaluation of feature combinations showing classification accuracy and criterion J_1. Features are defined by their numbers in Table 1. The 'Hidden' column is the number of hidden units in the network. For the hierarchical network, the two hidden values are the number of hidden units for the colour and real networks, respectively (see Figure 2b).

6 Discussion

This paper has explored suitable feature representations and classification methodologies for FISH signal classification. A family of features, consisting of measurements of size, shape, intensity, texture and colour, has been evaluated by different criteria. Histogram estimates of probability density functions and scatter plots provide preliminary visual insight into the relative importance of different features for the classification process. Feature selection enables the choice of feature sets of any type and number, which maximizes a class separability criterion J_1. The ultimate criterion for evaluating features for classification, however, is the probability of misclassification. Mismatches in selecting optimal feature sets according to the two criteria can be attributed to two factors: (a) the additional feature extraction stage performed by the hidden layer of each of the classifiers, and (b) the fact that J_1 is based on the Euclidian metric. This metric is useful for discrimination purposes only when the class patterns have equal covariance matrices.

Both the qualitative and quantitative analyses have demonstrated the superiority of size, hue and intensity-based features. When features of these families are combined together, even a single hue-based feature can completely separate signals of two fluorophores, leaving the task of discriminating real signals from artefacts to size and intensity-based features. Consequently, feature sets consisting of these features enable a hierarchical neural net based strategy to classify nearly 90% of the signals as reals or artefacts of two fluorophores.

Acknowledgements

This work was supported by EPSRC research contract GR/L51072. We thank Seema Dhanjal of Warwick University for preparation of the biological samples and FISH images, and Prof. Chris Bishop of Microsoft Research and Prof. Maj Hultén of Warwick University for discussions.

References

[1] Bishop, C.M., 1995. *Neural Networks for Pattern Recognition*. Clarendon Press.

[2] Carter, N.P., 1996. Fluorescence in situ hybridisation – state of the art. *Bioimaging* **4**, 41-51.

[3] Castleman, K.R., 1996. *Digital Image Processing*. Prentice-Hall.

[4] Fukunaga, K., 1990. *Introduction to Statistical Pattern Recognition* (2nd ed). Academic Press.

[5] Lerner, B., W.F. Clocksin, S. Dhanjal, M.A. Hulten and C.M. Bishop, 1999. Feature representation for the automatic signal classification in fluorescence in-situ hybridisation images. Technical Report 464, Computer Laboratory, University of Cambridge.

[6] Lerner, B., W.F. Clocksin, S. Dhanjal, M.A. Hulten and C.M. Bishop, 2000. Automatic signal classification in fluorescence in-situ hybridisation images. *Cytometry*, in press.

[7] Lerner, B., S. Dhanjal, and M.A. Hulten, 2000. GELFISH – Graphical Environment for Labelling FISH images. *Journal of Microscopy*, to appear.

[8] Netten, H., L.J. van Vliet, H. Vrolijk, W.C.R. Sloos, H.J. Tanke and I.T. Young, 1996. Fluorescent dot counting in interphase cell nuclei. *Bioimaging* **4**, 93-106.

[9] Ohta, Y., 1985. *Knowledge-based Interpretation of Outdoor Natural Color Scenes*. Pitman Publishing.