

# spotFinding Suite v2.5 User Guide

A machine-learning-based automatic MATLAB package to analyze smFISH images.

By Allison Wu, 2013 July

## 1. Installation

### 1. Requirements

This software was developed in MATLAB 2011a and has been tested on both Mac and PC. Some functions might not work in earlier versions but the suite should be able to work on either OS platform.

- The user needs basic MATLAB knowledge to utilize the output results.
- TIFF or STK are two currently supported image formats.
- It relies on the MATLAB statistical toolbox

### 2. Installation

After downloading the spotFinding Suite v.2.5, fully extract it to a chosen directory. Then, go to File > Set Path in MATLAB. Press 'Add with Subfolders' to add the directory that saves the spotFinding suite (Fig. 1) and save. One should be able to utilize all the functions in the Rifkin lab spotFinding suite from any working directory.

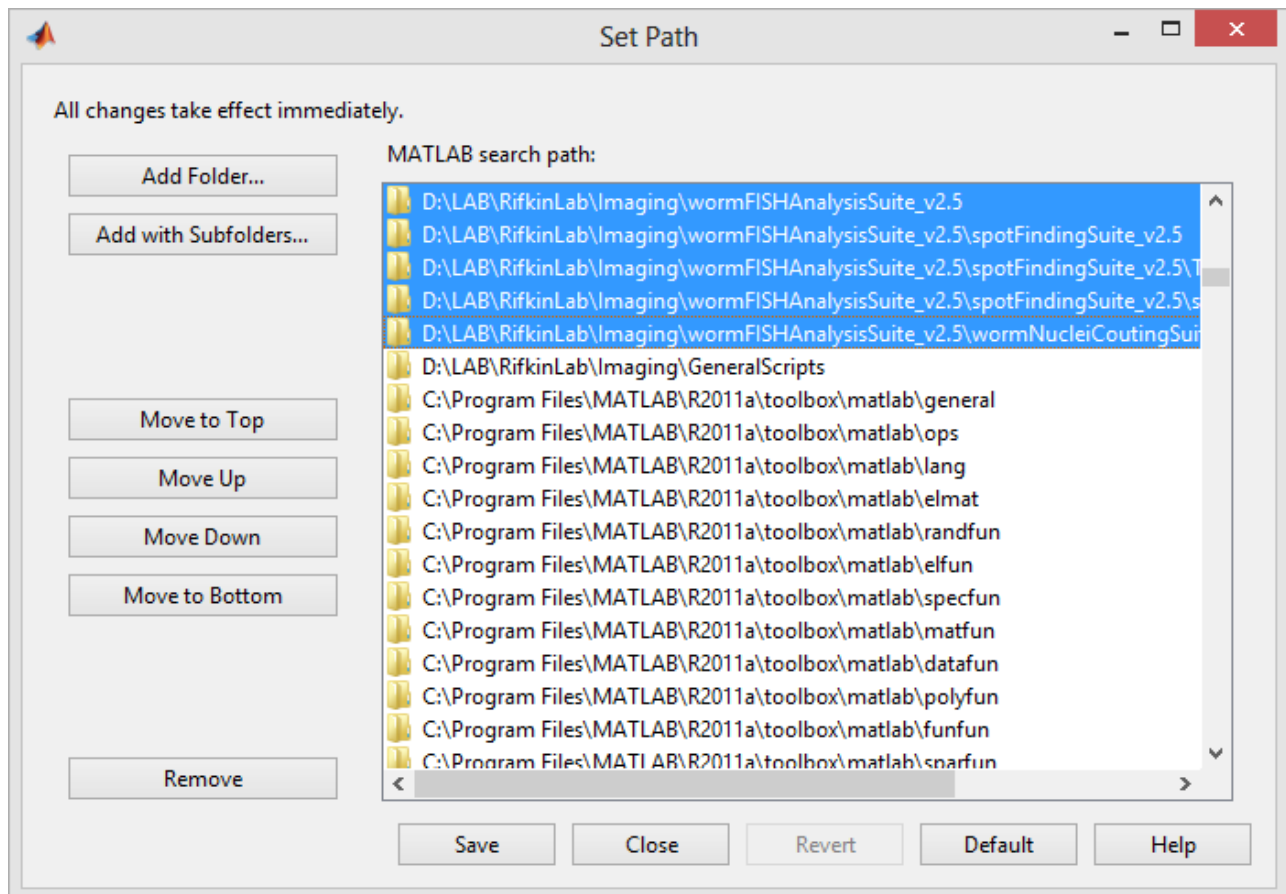


Fig. 1 Add the spotFinding Suite folder to MATLAB's set path.

- The following steps that are marked with '\*' take more than an hour for a batch of data with ~40 images but since these commands can operate automatically, no hands-on time is needed.
- Each function is annotated with detailed explanation. Please use 'help' for further details. e.g. `help createSegImages`
- All the files created by the function will show up in the working directory and all the functions will only search under the working directory.

## 2. Getting Started

spotFinding Suite is agnostic as to what actual biological specimens are being analyzed, whether cells or embryos or other things. Below we refer to the specimens being analyzed generically as *objects*. Note that *cell* is a specific type of data structure in Matlab and so this is what the word *cell* refers to below.

### 1. Create Masks for Your Images and Get the File Formats Correct

The masks are logical images that have entries of 1s where the objects are in the image and 0s denotes the space with no objects (Fig. 2). These masks are necessary for reducing the amount of memory needed for analyzing each image and can ensure proper scaling within the objects. However, the spotFinding Suite does not provide a way for the users to create masks for each image because there are currently many segmentation algorithms that can segment different kinds of images efficiently and automatically. The users should find their own ways to create the masks for their images. The Rifkin lab currently has another simple semi-automatic segmentation program for worm images, and we will happily share it with any other interested labs. However, segmentation is beyond the scope of this user guide. Here we will only discuss what one could do after all the segmentation masks are generated.



Fig.2 A mask that has one single object (left) and another mask that has multiple objects (right). Both of the masks are 1024 x 1024 pixels.

It is recommended that each image is segmented into no more than 5 masks for all the analyses. Each mask can have different number of objects in it as long as the exposure levels of all the

objects in the same mask can be scaled evenly. However, the total spot number estimate that the program will output is for each mask. To get a total spot number estimate per object, one still need masks with single objects but these can be applied after all the analysis is done. This will speed up the analysis and reduce the number of files generated. The example file contains an example of typical worm images that are best dealt with masks with single objects because worm embryos sometimes have different background exposure level in the same image so it would not be appropriate to scale their intensities and analyze them all together.

After creating masks for each image, the user has to make the file formats recognizable for the following steps. Note that the curly braces used below just designate the variable parts of the names. Do not include curly braces in your actual file names:

- ✓ Please make sure each of the tif or stk files has a 3-dimensional image stack at a single x,y position and that the z-axis order is the same as the real z-axis order.
- ✓ Make sure the mask files have the same x-y dimension as the image stack; that is, if the image stack has a size of 1024 x 1024 x 30, then each of the mask images should be 1024 x 1024.
- ✓ Please name the image file names as: {dye}\_{Position Identifier}.tif with no underscore within the text bounded by the curly braces.

For example:

- ✓ tmr\_Pos1.tif
- ✓ cy5\_001.tif
- ✗ tmr\_Pos\_1.tif
- ✗ Cy5001.tif
- ✗ TmrPos01.tif
- ✓ Make sure the mask files have the following naming pattern so that the suite can pair them with the correct image stacks: Mask\_{Position Identifier}\_{Mask Number}.tif, e.g. Mask\_Pos1\_1.tif is the mask for the first mask for the first image.
  - ✓ Mask\_Pos1\_1.tif
  - ✓ Mask\_001\_1.tif
  - ✗ Mask001\_1.tif
  - ✗ MaskPos1.1.tif
  - ✗ MaskPos1-1.tif

When all the above mentioned criteria are met, one can use the following MATLAB command to create the mask file format needed for the suite.

MATLAB Command :

```
>> createSegmenttrans(maskFileNameExample)
```

For example:

```
>> createSegmenttrans('Mask_Pos1_1.tif')
```

## 2. Getting Your Images Ready \*

MATLAB Command :

```
>> createSegImages('tif')
```

This command creates a {dye}\_{Position Identifier}\_segStacks.mat file for each image, e.g. cy5\_Pos10\_segStacks.mat is the segStacks file for position 10 in the cy5 channel. This mat file contains two cell variable: segStacks and segMasks. Each element of the cell variable, segStacks, contains a segmented image in a numerical matrix for each individual cells in this image and its counterpart in segMasks contains a logical matrix of the mask for the individual cell. These images are NOT the same size as the full image because to save on memory, the suite only saves the minimal rectangle (in x-y) necessary to contain the object (1s) indicated by the each mask (Fig.1). From this point on, all the analyses use these segStacks.mat files and not the original image files.

Note: Currently, the suite supports TIFF files and STK files. You can specify the file type as the input to createSegImages. Support for other file formats will be included in a future release. In the meantime, interested users could convert their images to TIFFs using other programs such as imreadBF() on the MATLAB file exchange.

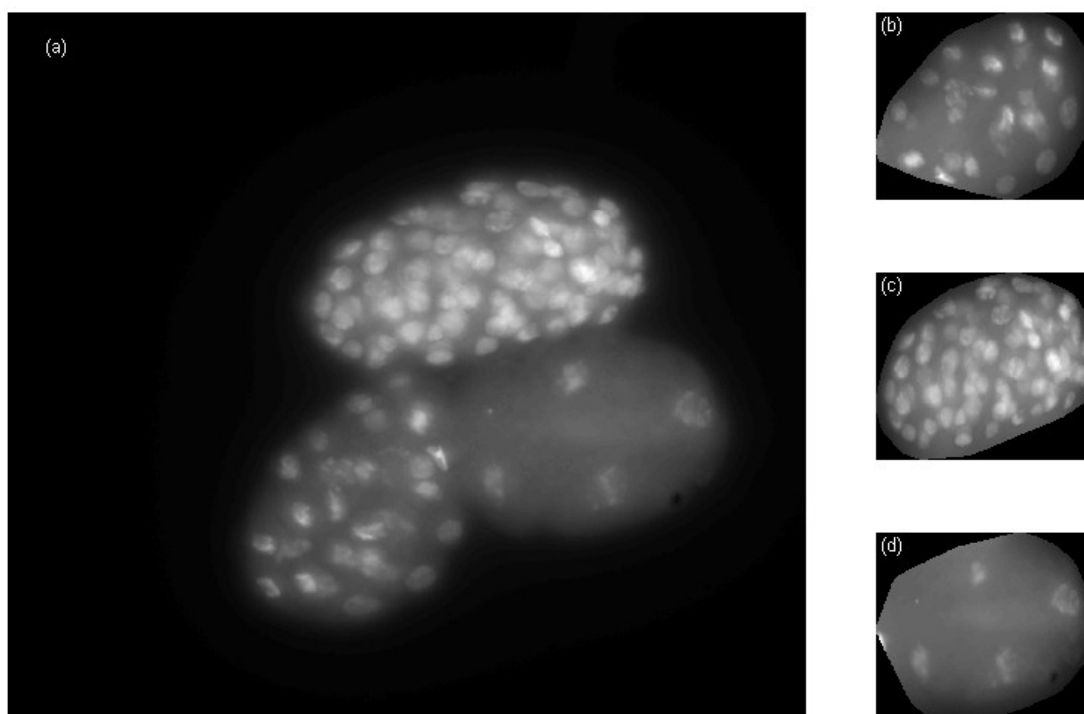


Fig. 3 (a) A full maximum projection DAPI image. (b-d) Three segmented individual cell image saved in the segStacks.mat file.

### 3. Find the Candidate Spots in Each Cell \*

MATLAB Command :

```
>> doEvalFISHStacksForALL
```

After getting the segStacks.mat files, the next step goes through each segStacks.mat image, except for DAPI images, finds the local maxima, and computes statistics that describe each local maximum. These statistics include features that describe how strong the shape feature is or how well a local maximum fits to a 2D Gaussian distribution, which reflects the fact that each spot is a diffraction-limited spot, etc. To see a full list of the features calculated, please refer to Appendix I.

If the suite used every local maximum for the following analysis, it would waste most of its time and memory on analyzing spots that are obviously bad. Therefore, the suite also filters out spots that are extremely unlikely to be a good spot by ignoring local maxima where one of the features, the scaled coefficient of determination from the fit to a 2D Gaussian, is below the specified threshold. The default setting is a very conservative setting that will not exclude any real spots based on our empirical explorations. All the statistics of each spot of each object are saved in the {dye}\_{Position Identifier}\_wormGaussianFit.mat files. Each file contains a cell array variable called 'worms', the elements of which save the spot information for each object in the image. To access the spot information for a particular object in a particular position, you need to first load in the wormGaussianFit.mat file for the specific image and type 'worm{object number in the cell array}' to view its statistics.

Example: To access the 2<sup>nd</sup> object in position 3 in the cy5 channel...

```
>> load cy5_Pos3_wormGaussianFit.mat
>> worm{2}
ans =

    version: 'v2.5'
    segStackFile: 'cy5_Pos3_SegStacks.mat'
    numberOfPlanes: 35
    cutoffStat: 'scd'
    cutoffStatisticValue: 0.7
    cutoffPercentile: 70
    bleachFactors: [35x1 double]
    regMaxSpots: [68246x5 double]
    spotDataVectors: [1x1 struct]
    goodWorm: 1
    functionVersion: {3x1 cell}

>> worm{2}.spotDataVectors

    locationStack: [758x3 double]
    rawValue: [758x1 double]
    filteredValue: [758x1 double]
    spotRank: [758x1 double]
    dataMat: [758x7x7 double]
    intensity: [758x1 double]
    rawIntensity: [758x1 double]
    totalHeight: [758x1 double]
    ...
    cumSumPrctile30RP: [758x1 double]
    cumSumPrctile90: [758x1 double]
    cumSumPrctile70: [758x1 double]
    cumSumPrctile50: [758x1 double]
    cumSumPrctile30: [758x1 double]
```

Note: In this object, there are 68246 regional maxima found but only 758 spots are left to be considered after using the cut-off value of 0.7 for the scd variable. By typing worm{2}.spotDataVectors, you can see a list of statistics or features calculated for the 758 spots.

### 3. Analyze the Spots Using the Random Forest Algorithm

## 1. Create a Training Set

After statistics for all the candidate spots are calculated, the user needs to prepare a training set to train the classifier. To create a good training set, here are some important points to follow:

- ✓ Because each channel and each batch of data may differ in quality and in the spot characteristics, we suggest that users create one training set for each channel in each batch of data independently so that the training set reflects the spots in each batch.
- ✓ The suite currently does not support using training sets from other batch of data. Using training sets from other batch of data will introduce errors in the subsequent functions such as `reviewFISHClassification()`. This feature will be implemented in a future release.
- ✓ A good training set should contain approximately the same amount of good spots and bad spots and should contain clearly good spots, clearly bad spots and some ambiguous spots for which the user will have to make some difficult classification. As with all supervised learning approaches, the algorithm is only as good as the quality of the training set.
- ✓ We suggest that the user first examines the max projection images of the particular channel and pick out 2-3 images for training so that the training spots will not come entirely from the same image and so it is assured that there will be a good representation of good spots and bad spots.
- ✓ It usually takes 3-4 rounds of training to get a robust classifier. In other words, the user trains an initial set, sees how it performs, either makes corrections and adds these corrections into the training set using the review GUI or adds more spots using the training GUI, retrains the classifier, and continues until the classifier does an acceptable job. It is better to increase the number of training spots at each round instead of starting with a huge training set since the training time needed is dependent on the number of spots. A training set of 300-400 spots will be a good start.

MATLAB Command :

```
>> createSpotTrainingSet('{dye}_{PositionNumber}','{Probe_name}')
```

Example: to pick out training spots from position 6 in cy5 channel for *C.elegans elt-2* probe, you can use ...

```
>> createSpotTrainingSet('cy5_Pos6','Cel_elt2')
```

**% Note: the probe name (2<sup>nd</sup> input) is entirely up to the users to decide. The 1<sup>st</sup> input should be in the same {dye}\_{position} format as described above.**

Before the GUI opens, the suite will search to see whether there exists any pre-established training set for this probe. If it finds a training set previously established, it will ask the user if he/she wants to overwrite the old training set or simply add new training spots to the training set.

When the GUI is started, the users should see a 16 x 16 pixel zoom-in window on the left, the original-sized image on the right. The 'Max. Merged Image' on the dlower-right corner is a maximum projection image of the neighboring slices, 2 slices above and 2 slices below and the current slice in the original-sized image.

This GUI allows the users to examine the candidate spots that are ordered by the spot rank, which uses one of the features as a crude quality score. The users can go down the spot rank and annotate each spot as good (Choose 'Next and Accept') or bad spot (Choose 'Next and Reject'), or they can pick out some good spots with high spot rank and use the 'Spot Rank' slider to jump to spots with low spot rank to add some bad spots to the training set. The users should keep in mind that this step is only meant to pick out a subset of examples of bad or good spots to train the training set.

In the panel on the right, the green rectangle specifies the area that is currently in the 16 x 16 panel. In the 16 x 16 window, candidate spots that are in the current frame are marked as blue. If the candidate spot is already in the training set, it will be marked as red. If there are multiple spots in the current frame, the user can click directly on the spot in the 16 x 16 zoom panel to reject the spot. If the user click 'Next and Accept' when there are multiple spots in the current frame, all the spots in this frame will be added to the training set as good spots.

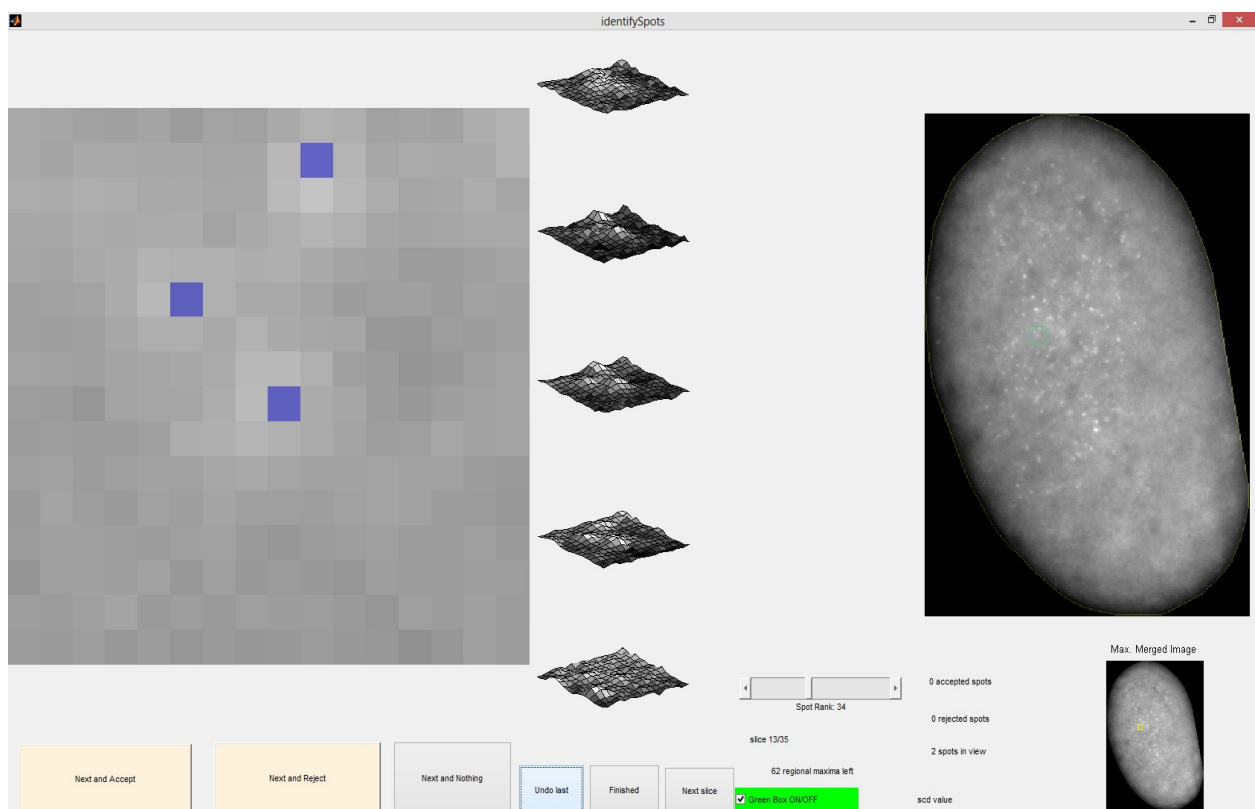


Fig. 4 The createSpotTrainingSet GUI is used to pick out spots for training set.

After the user has finished building the training set from a certain image and saves the training set, the user should see a new mat file called 'trainingSet\_{dye}\_{ProbeName}.mat', e.g. trainingSet\_tmr\_Cel\_end1.mat, in the working directory. This is the file that saves all the statistics of each spot in a structure variable called, 'trainingSet'. Later on, the user will use this file for training the classifier and the training results will also be saved in this file.

## 2. Train the Classifier : [Estimated time: 15-30 mins for 1000 training spots]

MATLAB Command :

```
>> load trainingSet_{dye}_{ProbeName}.mat
```



```
>> trainingSet=trainRFClassifier(trainingSet);
```

Example: to train the training set for *C.elegans end-1* tmr probe...

```
>> load trainingSet_tmr_Cel_end1.mat
```

```
>> trainingSet=trainRFClassifier(trainingSet);
```

In this step, this function will first determine which features are most invariant to the classification and will leave those out for further training. You will see the variables that are left out in the command window but you can always go back and check the list of variables that are left out after the training, which is saved in 'trainingSet.RF.VarLeftOut'. The second part of the function is to find the best number of variables sampled to construct the decision trees. Both of these parts will take a few minutes but these will ensure the robustness of the classifier.

When the training step is finished, you should see a new field called 'RF' in the trainingSet variable. This field saves all the statistics derived from training the random forest. In addition, you should see two new files added to the working directory. One is the 'probability estimates versus IQR' figure. The other is the {dye}\_{ProbeName}\_RF.mat file that saves all the trees.

The IQR in the 'Prob. Estimate v.s. IQR' figure is the interquartile range of probability estimates of being a good spot derived from all the trees in the forest. If the IQR is too large (the default cutoff is  $\text{IQR} > 0.3$ ), that means the trees in the forest cannot reach a consensus on the probability estimates. On the other hand, a small IQR means most trees agree on a certain probability estimate for this spot; these are the *concordant* spots. The 'Prob. Estimate v.s. IQR' figure shows how many training spots are in the *discordant* region ( $\text{IQR} > 0.3$ ). The number of *discordant* spots reflects how reliable the total spot number estimate is and it is further processed to give an estimate of the error range (See “Review the Spot Results” below).

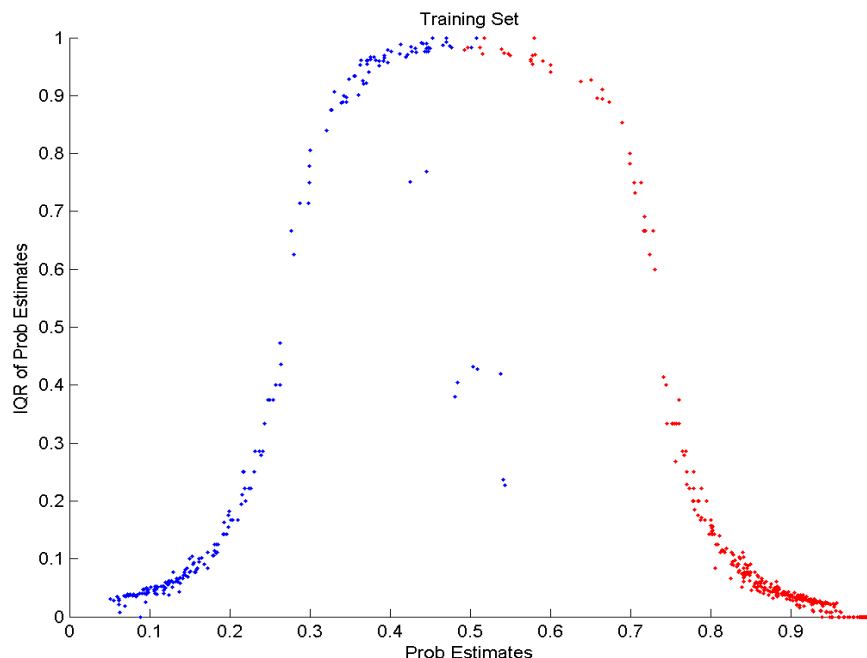


Fig. 5 A typical 'Probability estimate versus IQR' figure. The red dots are good spots and the blue dots are bad spots. Most spots fall in the concordant region. An approximately 10-20% *discordant* portion in the training set is common.



To interpret the training results, one can take a look into the RF field of the trainingSet variable:

Example:

```
>> trainingSet.RF

ans =

    Version: 'New method of estimating spot numbers, Apr. 2013'
    nTrees: 1000
    FBoot: 1
    VarLeftOut: {14x1 cell}
    statsUsed: {41x1 cell}
    VarImpThreshold: 0.21967
    VarImp: [1x55 double]
    dataMatrixUsed: [903x41 double]
    mTryOOBError: [32x2 double]
    NVarToSample: 6
    ProbEstimates: [903x1 double]
    spotTreeProbs: [903x1000 double]
    RFfileName: 'tmr_Cel_end1_RF.mat'
    ErrorRate: 0.016611
    MSE: 0.038268
    IQR: [903x1 double]
    IQRthreshold: 0.3
    DiscordantPortion: 0.19491
    SpotNumTrue: 560
    SpotNumEstimate: 563
    quantile: 75
    SpotNumRange: [543 610]
    Margin: [903x1 double]
    FileName: 'trainingSet_tmr_Cel_end1.mat'
    ResponseY: [903x1 logical]
    ConcordantErrorRate: 0.0013755
```

**<Interpretation>** In this training set, there are 903 training spots and 41 of the features, or statistics, are used. The 'dataMatrix' is an n-by-m numerical matrix that saves all the statistics for each spot, where n equals to the spot number and m is the number of statistics used. The field, 'dataMatrixUsed', saves the actual dataMatrix that is used for training the classifier. In the field of 'VarLeftOut', out can see the list of variables that have 'variable importance' in the lowest 25% percentile. The variable importance of a certain variable is defined by the change of error rate when the certain variable is permuted. The 'ProbEstimates' field has the average probability estimates among trees for each spot while the 'spotTreeProbs' saves the probability estimates derived from each individual tree for each spot. The training set error rate is 0.016611. The estimated total spot number is 563, which is close to the true spot number, 560. The 'spotNumRange' is the error range with a quantile of 75, which shows that in this set of spots, the estimate would fall between 543 and 610 75% of the time if the process were repeated. As for how the error range is estimated, please refer to the error estimation section in the Appendix II.

Note: While training the classifier, aside from visualizing the classification results in the reviewFISHClassification (discussed later), one should check if the 'ConcordantErrorRate' and the 'DiscordantPortion' is reduced in each round. These are the main two indicators that

indicate whether you could further optimize the training set.

### 3. Classify the Spots with a specified training set

To apply the classifier to a specified image, one needs to first load in the wormGaussianFit.mat file which saves all the spot information of each object in the image. Meanwhile, one also needs to load in the specific training set you would like to use to classify the spots.

**MATLAB Command :**

```
>> load trainingSet_{dye}_{ProbeName}.mat
>> load {dye}_{PositionNumber}_wormGaussianFit.mat
>> classifySpots(worms, trainingSet)
```

**Example:** To classify spots in the tmr image of position 6 with *C.elegnas end-1* tmr probe training set....

```
>> load trainingSet_tmr_Cel_end1.mat
>> load tmr_Pos6_wormGaussianFit.mat
>> classifySpots(worms, trainingSet)
```

One can also classify all the spots in the working directory all together with a specified training set. This function is basically a wrapper function for classifySpots. The first input 'toOverWrite' is a logical input that specifies whether the user would like to overwrite all current spot results in the directory. The 'dye' input is optional. If the user does not specify which channel this training set applies to, the program will ask the user in the command window so the user can enter it manually.

**MATLAB Command :**

```
>> load trainingSet_{dye}_{ProbeName}.mat
>> classifySpotsOnDirectory(toOverWrite, trainingSet, dye*)
```

**Example:** To classify tmr spots in the whole directory with the *C. elegnas* tmr probe training set....

```
>> load trainingSet_tmr_Cel_end1.mat
>> classifySpotsOnDirectory(1, trainingSet, 'tmr')
```

When spots in a certain image are classified, one should see a new file with the corresponding name of '{dye}\_{PositionNumber}\_spotStats.mat', which has a cell variable, spotStats, that has the spot analysis results for each object in the image in each entry.

**Example:** To examine the spot results in the 1<sup>st</sup> cell of image 6 in tmr channel....

```
>> load tmr_Pos6_spotStats.mat
>> spotStats{1}
```

ans =

```
      dataMatrix: [1099x41 double]
    spotTreeProbs: [1099x1000 double]
  ProbEstimates: [1099x1 double]
    SpotNumRange: [444 536]
SpotNumEstimate: 496
      quantile: 75
        Margin: [1099x1 double]
          IQR: [1099x1 double]
UnreliablePortion: 0.2475
```

```
classification: [1099x3 double]
trainingSetName: 'trainingSet_tmr_Cel_end1.mat'
locAndClass: [1099x4 double]
```

**<Interpretation>** There are 1099 candidate spots in this cell. The total spot number estimate is 496, with a 75% error range from 444 to 536. The 'locAndClass' field saves the relative spot location in this subimage in the first three column and the final classification of each spot in the last column.

#### 4. Review the Spot Classification Results (and Retrain).

This step is an important step for optimizing the training set. One can use this 'reviewFISHClassification' function to review spot results in some of images, curate the annotation, add some more spots into the training set and retrain the training set. Because random forests do not have an overfitting problem, one can keep repeating step 3-4 until the classification results in the several images no longer improve. It is common that the first result would not look very good (Fig. 6-1), which might due to some misclassified spots or simply not enough spots to allow the classifier to make good judgment. Usually, after 2-4 times of retraining, one should see a significant improvement of classification accuracy (Fig. 6-2).

To review spot classification for a particular image...

MATLAB Command :

```
>> reviewFISHClassification({dye}_ {PositionNumber})
```

Example: To review spot classification in first image in tmr channel...

```
>> reviewFISHClassification('tmr_Pos1')
```

The GUI starts up with the spot classification panel on the left. The candidate spots are ordered by the probability of being a good spot. The blue spots are classified as good spots while the yellow spots are classified as bad spots. The spot that is marked with red rectangle is the spot that is currently being curated. The user should see where the spot is in the cell, pointed by a small red arrow, in the panel on the right. The spots that have an X in their rectangles are spots that are manually curated and currently in the training set while the spots with slashes on them are manually curated but are not in the training set. These slashed spots may include some imaging anomalies, that are neither typical bad spots nor good spots so they might not be appropriate to be added into the training set.

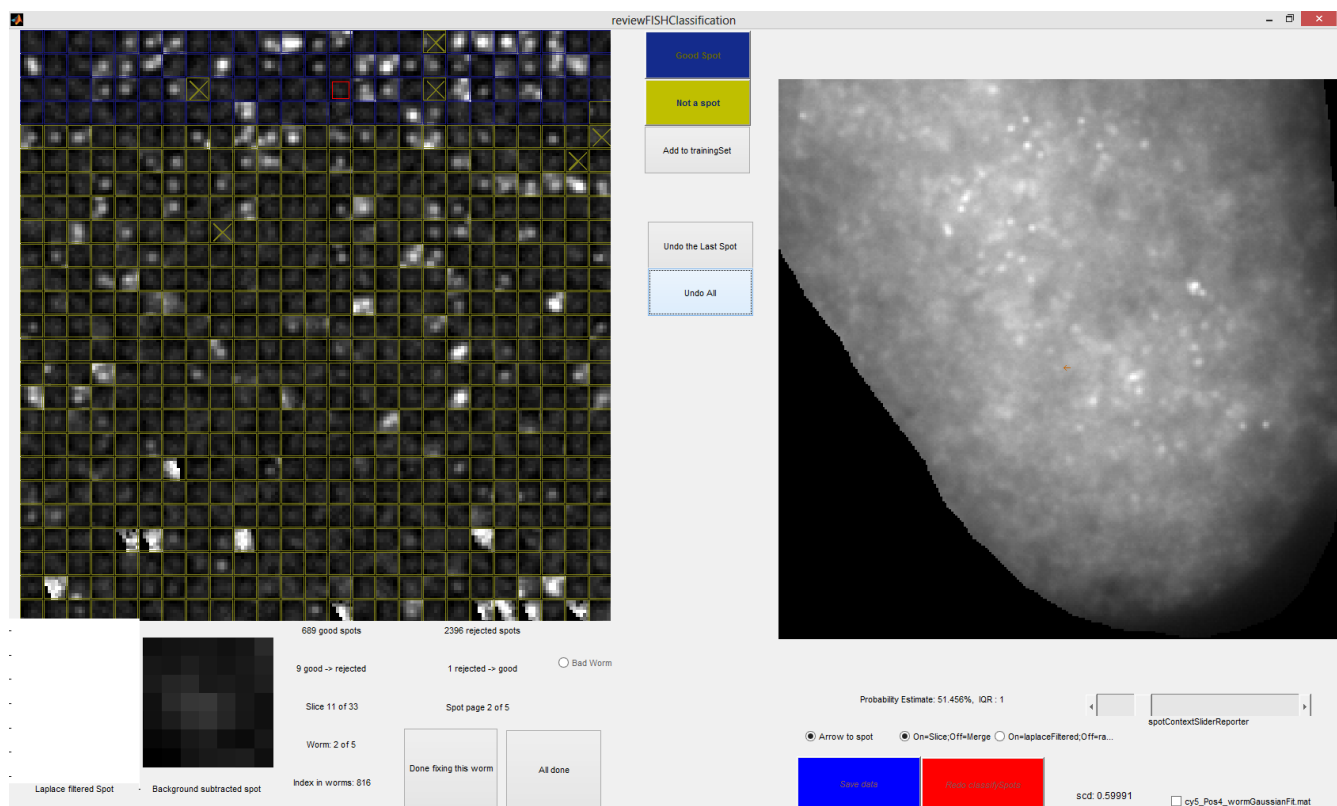


Fig. 6-1 The right panel shows spot classification results from a classifier that has only about 100 training spots. There are apparently too many false positives and false negatives in this classification result.

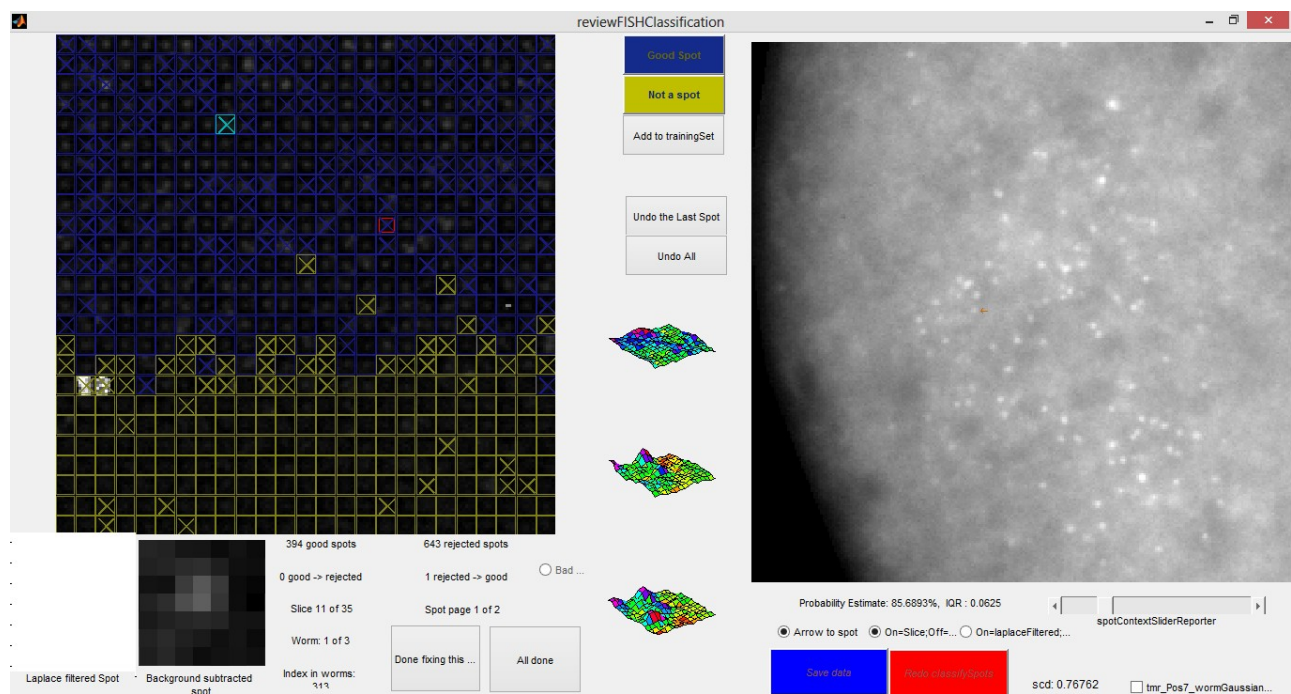


Fig. 6-2 shows spot results of the same embryo using a well-trained classifier which has about 1000 training spots.

After repeating step 3-4 several times on a few images, one should find the classifier's accuracy no longer improves. Then, one can classify all the spots in every image by using 'classifySpotsOnDirectory'.

## 5. Summarize and Interpret the Results

MATLAB Command :

```
>> spotStatsDataAligning(fileSuffix,alignDapi*)
```

Example:

```
>> spotStatsDataAligning('20130615',0)
% This command will create a file called, wormData_20130615.mat which saves
all the total spot number statistics.
```

After the user classifies all the spots, this command can be used to extract total spot number statistics from each position. The 'alignDapi' input is for worm users who would like to align the DAPI nuclei number as well. If this information is not available in the data set, one can just leave the input as '0' so that it will not try to align the DAPI nuclei number.

Two files should be generated after using this command. One is the wormData\_{fileSuffix}.mat file and the other is a figure called ErrorPercentagePlot\_{fileSuffix}.mat. The wormData MAT file has a wormData structure variable that saves the total spot number statistics extracted from all the images:

For example:

```
>> load wormData_20130524
>> wormData

wormData =

    spotNum: [201x6 double]
         U: [201x3 double]
         L: [201x3 double]
    meanRange: [0 80.022 79.721]
    errorPercentage: [201x3 double]

>> wormData.spotNum(1,:)

ans =

     1     0     1    -1    673    44

>> wormData.U(1,:)

ans =

     0    48    33

>> wormData.L(1,:)

ans =
```

<Interpretation> There are 201 objects in this whole batch. In the 'spotNum' field, the 6 columns are 'object index in the whole batch', 'position number', 'object index in the position', 'dye1', 'dye2', 'dye3' (alphabetical order, in this case, 'alexa','cy5','tmr'), and 'nuclei number' if 'alignDapi' input is '1'. A '-1' entry denotes any missing data. In this case, the first object in the whole batch is first object in the position 0 image. It has no alexa image found in the batch, therefore, -1. 673 cy5 spots and 44 tmr spots are found in this object. The U field has three columns that save the upper error bar of total spot number of each color for each object and the L field saves the lower error bar. Therefore, in this object, the upper bound of the total cy5 spot number is  $673+48=721$  while the lower bound equals to  $673-17=656$ . The 'meanRange' field saves the average error range of each channel to give the user a sense of how wide the error range is. The 'errorPercentage' is calculated by  $((U+L)/2)=\text{total spot number}$ . This is further visualized in the errorPercentage plot. Both the 'meanRange' and the 'errorPercentage' are meant to give the users a sense of how well the classifier does and whether it improves over several times of training.

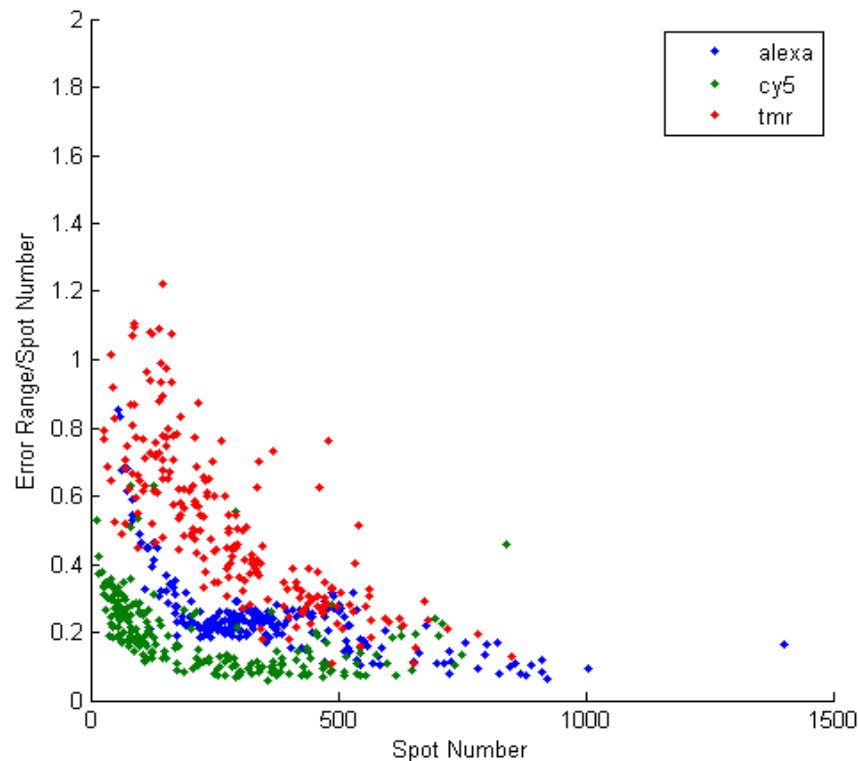


Fig. 7-1 A error percentage plot using spot results derived with an ill-trained training set. Note that error range is large in objects with different total spot numbers.

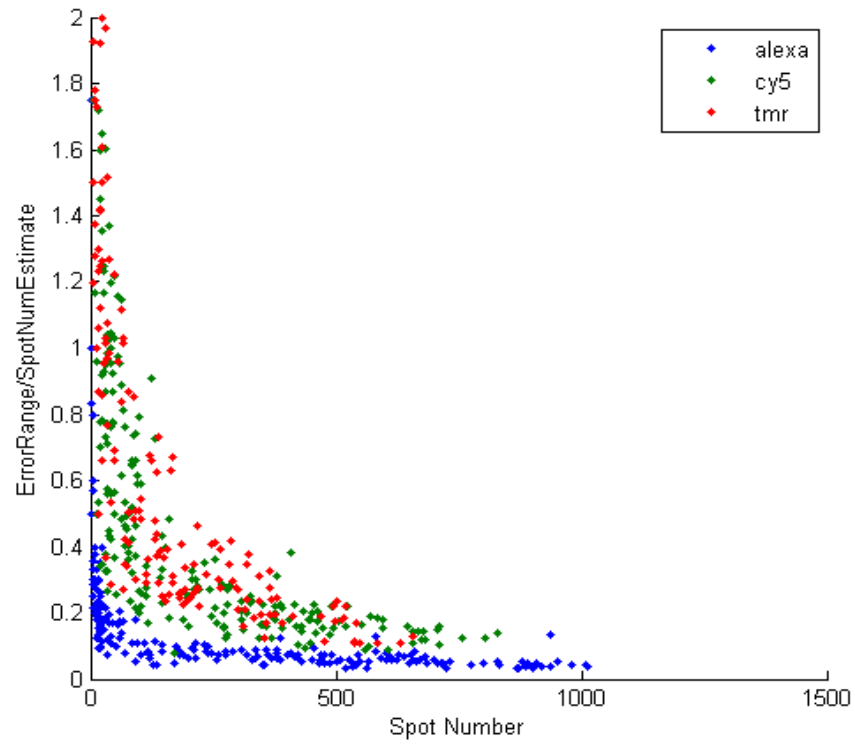


Fig. 7-2 A error percentage plot using spot results derived with an well-trained training set in the same data set. One should notice how the error percentage is reduced.