Automated Sub-Cellular Phenotype Classification: An Introduction and Recent Results

N. Hamilton[1,2,3] **R. Pantelic**[1,2] **K. Hanson**[1,2] **J.L. Fink**[1,2]
S. Karunaratne[1,2] **R.D. Teasdale**[1,2]

[1]Institute for Molecular Bioscience
[2]ARC Centre in Bioinformatics
[3]Advanced Computational Modelling Centre
The University of Queensland, Brisbane Qld 4072, Australia
Email: n.hamilton@imb.uq.edu.au

Abstract

The genomic sequencing revolution has led to rapid growth in sequencing of genes and proteins, and attention is now turning to the function of the encoded proteins. In this respect, microscope imaging of a protein's subcellular location is proving invaluable. High-throughput methods mean that it is now possible to capture images of hundreds of protein localisations quickly and relatively inexpensively, and hence genome-wide protein localisation studies are becoming feasible. However, to a large degree the analysis and localisation classification are still performed by the slow, coarse-grained and possibly biased process of manual inspection. As a step towards dealing with the fast growth in subcellular image data the Automated Sub-cellular Classification system (ASPiC) has been developed: a pipeline for taking cell images, generating statistics and classifying using SVMs. Here, the pipeline is described and correct classification rates of 93.5% and 86.5% on two 8-class subcellular localisation datasets are reported. In addition we present a survey of other important applications of cell image statistics. The complete image sets are being made available with the aim of encouraging further research into automated cell image analysis and classification.

Keywords: Subcellular phenotype, subcellular localisation, image statistics, image classification, machine learning.

1 Introduction

The advent of fast, automated and inexpensive sequencing technologies led to the completion of the human, mouse and many other genomes, and an exponential growth in genomic data. Sequence-based machine learning has played a pivotal role in automated annotation and prediction of structure and function of novel sequences and has become an essential tool. However, while sequence data are invaluable further information, such as experimentally-determined subcellular localisation (see Figure 1), trafficking and interaction partners is required to fully understand the functions of the tens of thousands of proteins that have been identified (Fink, Aturaliya, Davis, Zhang, Hanson, Teasdale & Teasdale 2006)(Stow & Teasdale 2005). Sequence-based approaches have been applied to predicting localisation (Yu, Chen, Lu & Hwang 2006) but tend to need high homology to proteins of known localisation, and so experimental verification is a necessity. Automated fluorescent microscope imaging technologies mean that it is now possible to capture hundreds of images per second including multiple fluorophores for cells under a variety of experimental conditions (Lang, Yeow, Nichols & Scheer 2006)(Bonetta 2005). Furthermore, cells may now be imaged in 3D, or indeed in 4D with a 3D stack captured over time to observe protein trafficking in real time. The desire and the ability to do high-throughput screenings of protein localisation and trafficking is leading to a rapid growth in cell images in need of analysis on a scale comparable to that of the genomic revolution. Automated image analysis and classification is essential.

Much of the reason for the rapid growth of machine learning techniques applied to genomic data is ubiquitousness of sequence information from publicly available databases. Until recently, dissemination of cell image data involved selection of a few "representative" images for publication in a paper. But a much richer range of data is becoming available with large-scale publicly accessible cell image databases such as the LOCATE mouse protein subcellular localisation database (Fink et al. 2006) (more databases are listed in (Matthiessen 2003)). Currently, the databases are largely human-curated, but the data becoming available offer many opportunities to train and apply machine learning techniques to experimental image classification and analysis. There is a real need to refine, discriminate and quantify to produce annotation of images in cell databases. Cells can exhibit a wide range of behaviours over the cell cycle that can potentially skew results, and techniques have been developed to automatically determine the phase of cell image sequences (Pham, Tran, Zhou & Wong 2006). Aberrant cell morphology also presents an interesting challenge to image classification. Atypical morphology may skew data when examining normal cells. Alternatively, atypical morphology may be the primary attribute which assists in the discrimination between, for instance, normal cells and cancerous cells (Thiran, Macq & Mairesse 1994). On the quantitative side, methods have been developed to select and count substructures, such as puncta (Pham, Crane, Tran & Nguyen 2004), from cell images. These automated techniques offer the opportunity to annotate at a much more refined level, thereby increasing the quality of data and allowing more subtle hypotheses to be tested.

As well as presenting the ASPiC pipeline, the aim here is to draw attention to the large image data sets that are now becoming available and to the great need for, and applications of, machine learning to these sets. In the following, we begin with an introduction to image statistics and their potential applications to

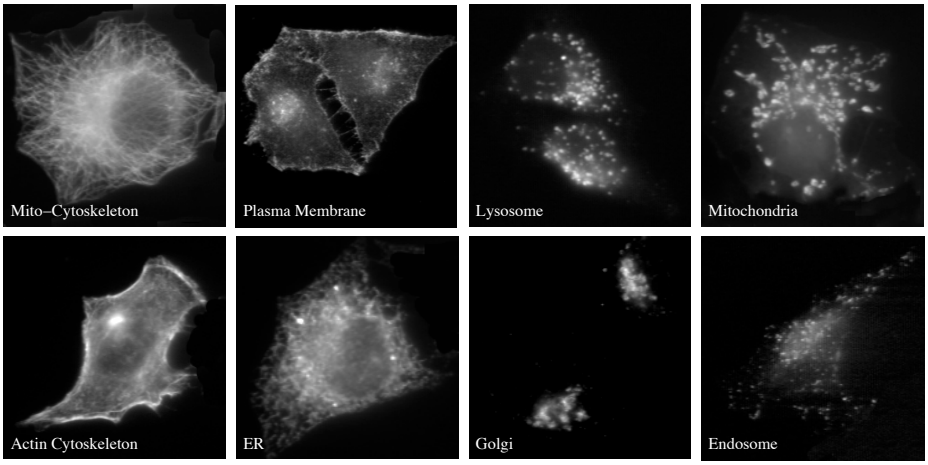

Figure 1: Samples of endogenously expressed proteins from our datasets

high-throughput cell imaging problems. The ASPiC system and the image sets being made available are then described, and we conclude with some remarks on future directions for analysis of subcellular imaging.

2 Image Statistics and their Uses

A common problem in cell biology is to determine the subcellular localisation of a given protein: does it localise to the nucleus or the cytoplasm? Has treatment of the cell modulated the localisation of an individual protein? Examples of fluorescently-tagged proteins exhibiting various subcellular localisations are shown in Figure 1. While applying learning algorithms with the image itself as input has proved quite successful (Danckaert, Gonzalez-Couto, Bollondi, Thompson & Hayes. 2002), generation of numeric image measures has a wider range of applications. The aim is to find measures that can differentiate between localisations in distinct classes when localisations within a given class can exhibit a very wide range of expression patterns and morphologies. To be applicable to as wide a range of images as possible, cell image measures should ideally be invariant under rotation, translation and scale changes. Here, some of the measures that have been applied to quantifying subcellular localisation images are described. More may be found in (Conrad, Erfle, Warnat, Daigle, Lorch, Ellenberg, Pepperkok & Eils 2004) and (Huang & Murphy 2004).

Area Intensity Measures

Typically, for subcellular localisation a pair of microscope images will be taken: one of the fluorescently-tagged protein of interest (POI); and one in which the nucleus of the cell is fluorescently-labelled. From these, image masks of the POI and the nucleus are created (see Figure 2). Area and intensity measures may then be calculated: the area of the region that the POI is expressed in; average intensity across the masked region; the ratio of the intensities of the POI in the nuclear to non-nuclear regions; area and intensity averages over various intersections and differences of POI and nuclear masks; and the standard deviation of the POI image intensity in the mask region. Statistics such as these will easily differentiate between proteins expressing in the cytoplasm and the nucleus. Generally, area and intensity *ratio* measures are better in that they are less affected by the image resolution or exposure.

Haralick Texture Measures

A more refined set of image measures that have been applied to a wide range of problems such as satellite imaging and computerized tomography are the Haralick texture measures (Haralick 1979). The idea is to find the correlation (and other measures) between pixel intensities at a given distance and angle. Hence, if an image contained a series of high-intensity bands at a given separation, a Haralick correlation measure (with the appropriate distance and angular separation) will return a high value. Suppose an image contains N gray tones, then for a given pixel pair separation d and angle θ a $N \times N$ *gray tone co-occurrence table* P is constructed. The entries P_{ij} are the relative frequency with which two pixels separated by distance d and angle θ have gray tone values i and j, respectively. Measures such as *uniformity*: $\sum_{ij} P_{ij}^2$; *entropy*: $\sum_{ij} P_{ij} \log P_{ij}$; and *correlation*: $\sum_{ij} (i-\mu)(j-\mu)P_{ij}/\sigma^2$, where μ and σ are the mean and standard deviation of the pixel intensities, are then applied to the occurrence matrix. The Haralick statistics may be applied to the whole of the mask region of the POI, or to subregions of it defined by intersections and differences of the POI and nuclear masks. Since there are many possible choices of d and θ, for a given d the occurrence matrix is sometimes averaged over a range of values of θ such as 0°, 90°, 180° and 270° degrees. This has the advantage of reducing rotational variance, though may lead to a reduced signal. More Haralick measures are given in the Appendix.

Zernike Moments

Another set of measures that are computationally relatively inexpensive and have proved useful in cell imaging are the Zernike moments (Khotanzad & Hong 1990)(Boland, Markey & Murphy 1998)(Zernike 1934). These are calculated using an orthogonal polynomial set, the Zernike Polynomials, on the unit circle. Given a complete (infinite) set of Zernike moments for a given image it is in theory possible to reconstruct the image perfectly. However, calculation of the first few moments will often give a general sense of the morphology of the imaged object, much as a small subset of Fourier coefficients will for a time series (Boland et al. 1998). The discrete equations for the Zernike moments are given in the Appendix.

Applications

In general, no one statistic is a good predictor of subcellular phenotype, and so machine learning techniques such as neural networks and support vector machines have been applied to classification based on image statistics. As shown in the next section, classification accuracies of greater than 90% may be obtained. In addition to allowing high-throughput classification of new images, an immediate application of a phenotype classifier is to *image database curation*. As the size and number of image databases expands, quality and uniformity of human classification becomes an issue. Experiments by Murphy lab on a set of images with 10 distinct known subcellular localisations (similar to those in Figure 1) found human classifiers had an accuracy of 83% compared to 92% for a machine classifier (Huang & Murphy 2004)(Murphy, Velliste & Porreca 2003). This may in part be explained by the inherent difficulty in providing accurate classifications for hundreds of images over a long time period, but it is worth noting that the human eye only registers a few tens of distinct gray scale values at a time, while a 8-bit cell image file may have close to 250, and hence there is potential for software to "see" much more. By flagging for re-examination the images for which the human and machine classifications disagree, there is the potential to significantly improve database quality.

Other applications of image statistics include *representative image selection* and *statistical comparison* of image sets. In the former, given a set of images of a particular protein, the aim is to select the image that best represents the variety of distributions observed. This may be done by finding the image that has statistics closest to the mean statistics vector of all the images (Roques & Murphy 2002). In the latter, there are two sets of experimental conditions for a protein where it is required to ascertain whether the two distributions are statistically significantly different. Using the Hotelling T^2-test on the image statistics, it has been shown (Roques & Murphy 2002)(Huang & Murphy 2004) that sets with the same localisation may be correctly identified, and that expression patterns that were known to be different can be distinguished, even to the extent of differentiating visually-indistinguishable images.

Finally, cell image statistics offer the possibility of searching image databases for similar images on the basis of image content rather than the (possibly biased) keywords supplied by the experimenter. Arguably the most powerful tool for genomic inference is the BLAST sequence matching algorithm (Altschul, Gish, Miller, Myers & Lipman 1990) that finds and quantifies similarity between sequences in a database. Once an "image BLAST" is developed for cell image databases, the ability to deduce biological inferences and associations will be greatly increased.

3 The ASPiC Pipeline

The Automated Subcellular Phenotype Classification system (ASPiC) is a fully-automated pipeline from experimental image to a subcellular classification suitable for direct database entry. The principle steps are outlined in Figure 2 and are described in detail below. ASPiC is currently being integrated into the LOCATE database (Fink et al. 2006) where it is providing classification on a 3-class nuclear or cytoplasm or nuclear and cytoplasm problem. Here, we describe its application to two 8-class subcellular localisation datasets. Other applications such as representative image selections are under development. The major parts of ASPiC are implemented in C++ using the ImageMagick++ image libraries.

3.0.1 Image Sets

An image collection was created for each of 8 subcellular organelles in two types of sets; one in which an *endogenous* protein or feature of the specific organelle was detected with a fluorescent antibody or other probe; and another in which an epitope- or fluorescently-tagged protein was transiently expressed (transfected) in the specific organelle and subsequently detected. Each set consisted of 50 images. Each image was accompanied by an additional image of the cells counterstained with the DNA specific dye 4,6-diamidino-2-phenylindole (DAPI), which highlights the location of the nucleus of every cell in the image. All images were of fixed HeLa cells, taken at 60X magnification under oil immersion. More details are available with the image sets.

Automated Cropping and Cell Selection

The first step is to select regions in which proteins are expressing in the POI and nuclear images using an automated grayscale thresholding scheme. A variety of schemes were tried, but the best was found to be to choose a minimum intensity and a maximum intensity, and take the average (μ) and standard deviation (σ) of the pixels with intensities in this range. For the POI images, a minimum intensity of 30 and a maximum of 250 is used, and 20 and 250 for the nuclear images. The threshold for the image is then set to $\mu - 0.9\sigma$, and above threshold pixels define the regions of interest. Using these minima and maxima in the calculation of μ and σ removes pixels that are either certainly background or overexposed, and gives results that were generally in agreement with the regions that the eye considers to be of interest. Contiguous regions are then selected and cropped in the POI image, together with the corresponding area in the nuclear image. To remove artefacts, any selected region that is small or faint is discarded. The circularity (perimeter squared over area) of the nuclear region mask is calculated, and nuclei with large circularity are discarded as these usually represent multiple or poorly imaged nuclei in a cell. In the case of multiple nuclei in the cropped region, the most central is selected. In some cases cells are not separable by thresholding. This is detected by multiple disjoint nuclei (in the nuclear mask) being contained within the POI mask, and is treated as a single cell by ASPiC. Using these criteria approximately 93% of source images are found to contain one or more valid cells.

Image Statistics

For each cropped cell a total of 95 statistics are generated composed of 25 area and intensity measures, 21 Haralick statistics and 49 Zernike moments of up to degree 12. Since Zernike moments are not rotationally invariant, the magnitudes of the moments are taken to minimise sensitivity to cell orientation. The area and intensity measures arise from taking intersections and differences of the POI and nuclear masks, and calculating average intensities, areas, intensity ratios and area ratios. Haralick measures were chosen from a list of those shown to be good for distinguishing subcellar localisation in (Conrad et al. 2004). Details of the measures used by ASPiC may be found in the Appendix.

Training and Testing

Support vector machine classifiers were created for the 8-class endogenous and 8-class transfected image

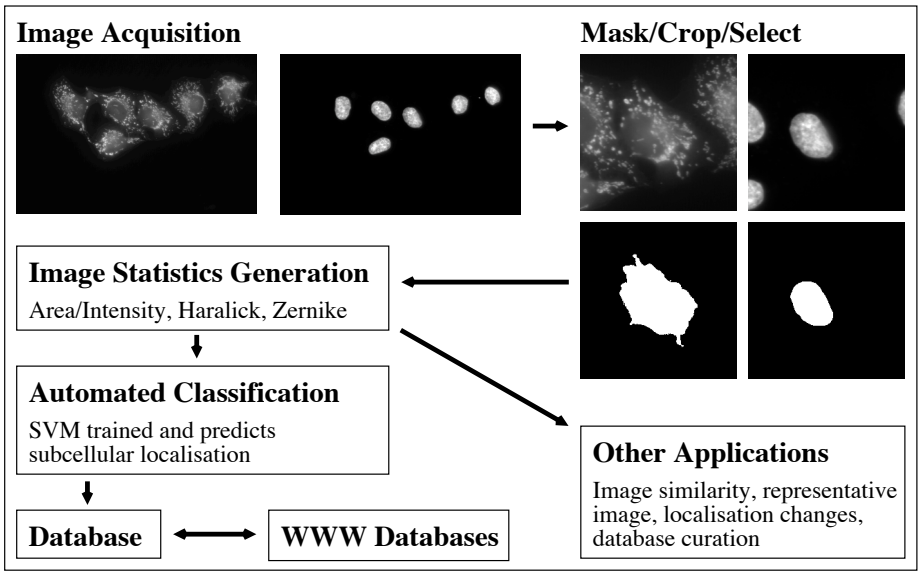

Figure 2: The Automated Subcellular Phenotype Classification System (ASPiC)

sets using the libsvm software with an RBF kernel (Chang & Lin 2001). An ANN was also briefly trained and tested on the same data sets, but found to give lower performance (data not shown). Two parameters are required to create and train the SVM: γ, the coefficient of the exponent for the RBF kernel and C, the penalty parameter of the error term. A grid search was performed to choose the values of γ and C that gave the best 5-fold cross validated performance on each data set. On the endogenous data set, the best cross validation accuracy was 94.3% using $\gamma = 0.03716$ and $C = 26.91$. For the tranfected data set, an accuracy of 89.8% was obtained using $\gamma = 0.03284$ and $C = 89.84$. Linear kernels were also tested with 5-fold cross validation on each data set, and gave 91.8% and 89.2% on the endogenous and transfected data sets, respectively. Polynomial kernels were also tested, but were also found not to perform as well as the RBF kernel. Once the RBF kernel and parameters were fixed as above, 100 random (class-balanced) splits of the data into 4/5 training and 1/5 testing set were performed and an SVM trained and tested. For each test set, the overall percentage of correct predictions was recorded, as well as the percentage of correct predictions for each class of the data in the test set.

Schemes for selecting subsets of the statistics ranked by F-score were also investigated using the *fselect* script available for libsvm. Selection by F-scores has been shown to significantly improve performance on some data sets (Chen & Lin 2006). The F-scores varied widely, from 0.02 to 9.2 for the endogenous data set, and from 0.04 to 6.2 on the transfected data set. There was no clear bias in ranking either subregion, Zernike or Haralick statistics highly, with all three types represented in the top ten. For each of the endogenous and transfected data sets, 5-fold cross validating with the best ranked subsets of 95, 72, 47, 23, 11, 5 and 2 features showed either no improvement or significantly degraded performance.

Post Classification Filtering

ASPiC includes a voting system for multiple classifications of the same protein in distinct cells where split votes are broken by the maximum confidence score output by the SVM. There are a variety of approaches to post-classification filtering (Chen, & Murphy 2006), however, here we report raw classification accuracies in order that the true accuracy may be seen.

3.0.2 Classification Accuracy and Comparing the Incomparable

Over 100 trials of splitting data sets 4/5 to 1/5 for training/testing the average correct classification rates were 93.1% for the endogenous test sets and 87.9% for the transfected, with standard deviations of 1.58 and 2.51, respectively. To test which classes were accurately or poorly classified, the classification accuracies for each image class were also recorded and the averages are given in Table 1. Generally, the classification accuracies are high, though certain classes such as the cytoskeleton classes are less well-predicted for transfected cells. Those that are poorly predicted tend to be those that are visually similar to other classes. ASPiC has also been tested using only those statistics that require a POI and not a nuclear image, and gave cross-validation results around 2.5% lower than those above. Before comparing these results with previous literature, it should be made clear that each group is testing their system on distinct image sets with different numbers of subcellular classes and varying degrees of automation, and hence are not necessarily directly comparable. Murphy lab have been developing and improving subcellular phenotype classifiers for a number of years and have contributed much in the area of subcellular image statistics and their uses. The most recent report (Huang & Murphy 2004) on subcellular classification gives 88% for a pure neural network classifier on a 10-class problem with manual cropping and curating. The image sets were prepared using protein antibodies of known localisation, and hence are comparable to our endogenous image sets. Using a majority voting system combining a number of learning algorithms, 92% was obtained on the same set. A wide range of statistics were used and various feature selection algorithms used to select the best for training including Zernike moments and Haralick measures, though their implementation of Haralick measures differs from ASPiC's in that ASPiC does not average over a range of angles. Also of interest is the work of Conrad et al. (Conrad et al. 2004) in which a wide variety of image statistics, feature selection and learning algorithms were

Endo.	Mito-Cyto.	Endosome	ER	Golgi	Actin-Cyto.	Lysosome	Mitochondria	PM
#	45	31	59	48	29	62	68	22
Acc.	96.7	93.5	93.9	98.9	83.6	94.9	91.3	88.8
Trans.	Cytoplasm	Lysosome	ER	Endosome	Peroxisome	Mito-Cyto.	Actin-Cyto.	Nucleus
#	43	16	59	30	34	37	27	23
Acc.	99.7	98.7	84.5	80.4	89.0	78.5	85.5	100

Table 1: Average classification percentages on Endogenous and Transfected test sets over 100 randomised splits of the data into 4/5 training, 1/5 testing.

tested on 11 classes of subcellular phenotype images. Of the methods tested, they found stepwise feature selection in conjunction with a SVM offered the best performance with an accuracy of 82.2%. While comparison is problematic, ASPiC is certainly competitive with a 93.1% accuracy, it is simple in that it is fully-automated and uses a single machine learning method, and has been shown to perform well on uncurated images.

4 Conclusions

It is clear that image statistics can differentiate subcellular localisation to a high degree of accuracy, and that automation offers many advantages in high-throughput, time saved, consistency and quantification.

Currently, statistics are relatively slow to compute. Cells need to be selected from images of plates, cropped, and then up to a hundred statistics calculated, all of which can take of the order of seconds on a standard PC. When faster statistics are developed the range of applications will grow. One application would be to flow cytometry where cells are imaged and sorted on the fly (Bonetta 2005). With current technology, cells are typically sorted according to whether a cell is expressing a protein (bright) or not (dark). A fast classifier would enable selection of, for instance, all those cells for which a given protein is expressing in the Golgi, and then perform further experiments on those. New statistics we are developing look promising as quick, relatively accurate measures with no cropping.

As the flood of cell image data begins, the need for new applications of classification and discrimination are greatly increasing. Certainly there is a need for automated classification, but cell image databases also need the ability to be *queried by image example* in a way that understands the content of the image rather than by matching researcher-supplied keywords. If a researcher was looking to see if a protein localised to the Golgi, they may not have noted that it was in fact localising to a subregion of the Golgi. However, an image content-based search might provide that level of discrimination. In the future, as biological databases become more integrated and queryable, it should be possible, for instance, with a few mouse clicks to start with a protein sequence, find images of its subcellular localisation, "image BLAST"to find proteins that exhibit similar expression or co-expression patterns, then read source literature on the proteins.

Experimental images described herein are available via the LOCATE web interface (Fink et al. 2006).

5 Appendix: Features used in ASPiC

Haralick Texture Measures

Suppose an image contains N gray tones, then for a given pixel pair separation d and angle θ a $N \times N$ *gray tone co-occurrence table* P is constructed. The entries P_{ij} are the relative frequency with which two pixels separated by distance d and angle θ have gray tone values i and j, respectively. This definition of the gray tone co-occurrence table is as in (Haralick 1979) with the minor variation that the matrix has been normalised to give relative frequencies rather than counts of pixel pairs. The following image statistics are then calculated in ASPiC.

Correlation: $\sum_{ij}(i-\mu)(j-\mu)P_{ij}/\sigma^2$ where μ and σ are the mean and standard deviation of the pixel intensities.
$d = 3$, $\theta = 0$; $d = 4$, $\theta = 45$; $d = 3$, $\theta = 135$.

Correlation2: Haralick's second information measure of correlation. See (Haralick, Shanmugam & Dinstein 1973).
$d = 2$, $\theta = 0$; $d = 3$, $\theta = 45$; $d = 1$, $\theta = 135$.

Contrast: $\sum_{ij}(i-j)^2 P_{ij}$
$d = 5$, $\theta = 0$; $d = 5$, $\theta = 135$.

Inverse difference moment: $\sum_{ij} P_{ij}/(1+(i-j)^2)$
$d = 1$, $\theta = 90$.

Unformity: $\sum_{ij} P_{ij}^2$
$d = 1$, $\theta = 0$; $d = 2$, $\theta = 0$; $d = 4$, $\theta = 45$.

Entropy: $\sum_{ij} P_{ij} \log P_{ij}$
$d = 4$, $\theta = 135$.

Sum entropy: $\sum_k((\sum_{i+j=k} P_{ij}) \log(\sum_{i+j=k} P_{ij}))$
$d = 1$, $\theta = 0$; $d = 4$, $\theta = 90$.

Difference entropy: $\sum_k((\sum_{|i-j|=k} P_{ij}) \log(\sum_{|i-j|=k} P_{ij}))$
$d = 4$, $\theta = 0$; $d = 3$, $\theta = 45$; $d = 1$, $\theta = 45$.

Sum variance: $\sum_k (k-S)^2 \sum_{i+j=k} P_{ij}$ where S is the sum entropy.
$d = 4$, $\theta = 90$.

Zernike Moments

The magnitudes of the first 12 Zernike moments are calculated exactly as described by the equations in (Boland et al. 1998) to give 49 features as follows. Let $I(x,y)$ be the pixel intensity at position (x,y). Define

$$Z_{nl} = \frac{n+1}{\pi} \sum_{x,y} V_{nl}^*(x,y) I(x,y)$$

where $x^2 + y^2 \leq 1$, $0 \leq l \leq n$, $n-l$ even, and V_{nl}^* is the complex conjugate of the Zernike polynomial of degree n and angular dependence l, given by

$$V_{nl}(x,y) = \sum_{m=0}^{(n-l)/2} \frac{(-1)^m (x^2+y^2)^{n/2-m} e^{il\theta} (n-m)!}{m!(\frac{n-2m+l}{2})!(\frac{n-2m-l}{2})!}$$

where $\theta = \tan^{-1}(y/x)$.

Cell images are centered when cropped. To scale each cell image into the unit circle, pixel coordinates are divided by 100 before calulation of the Zernike moments.

Subregion Statistics

Denoting the mask selected region of the POI by P, nuclear mask selected region by N, and the area of a region A by $|A|$, the following area statistics are calculated: $|P|$, $|N|$, $|P-N|$, $|N-P|$, $|P \cap N|$, $|N|/|P|$, $|P-N|/|N|$, $|P-N|/|P|$, $|N-P|/|P|$, $|N \cap P|/|P|$, $|N-P|/|N|$ and $|N \cap P|/|N|$. The variance over the POI mask region, as well as the ratio of the perimeter squared over the area of the POI mask region, are also calculated.

Denoting the average pixel intensity over a region A by $(A)_I$, the following intensity measures are calculated in the POI image: $(P)_I$, $(N)_I$, $(P-N)_I$, $(N-P)_I$, $(P \cap N)_I$, $(P-N)_I/(P)_I$, $(N-P)_I/(P)_I$, $(N \cap P)_I/(P)_I$, $(N-P)_I/(P)_I$, $(N \cap P)_I/(P)_I$, $(N)_I/(P)_I$ and $(P-N)_I/(N)_I$.

References

Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. (1990), 'Basic local alignment search tool', *J. Mol. Biol.* **215**, 403410.

Boland, M., Markey, M. & Murphy, R. (1998), 'Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images', *Cytometry* **33**(3), 366–375.

Bonetta, L. (2005), 'Flow cytometry smaller and better', *Nature Methods* **2**, 785 – 795.

Chang, C.-C. & Lin, C.-J. (2001), *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/ cjlin/libsvm`.

Chen, S.-C., & Murphy, R. F. (2006), 'A graphical model approach to automated classification of protein subcellular location patterns in multi-cell images', *BMC Bioinformatics* **7**(90).

Chen, Y.-W. & Lin, C.-J. (2006), *Feature extraction, foundations and applications*, Springer, chapter Combining SVMs with various feature selection strategies.

Conrad, C., Erfle, H., Warnat, P., Daigle, N., Lorch, T., Ellenberg, J., Pepperkok, R. & Eils, R. (2004), 'Automatic identification of subcellular phenotypes on human cell arrays', *Genome Research* **14**(6), 1130–6.

Danckaert, A., Gonzalez-Couto, E., Bollondi, L., Thompson, N. & Hayes., B. (2002), 'Automated recognition of intracellular organelles in confocal microscope images', *Traffic* **3**(1), 66.

Fink, J., Aturaliya, R., Davis, M., Zhang, F., Hanson, K., Teasdale, M. & Teasdale, R. (2006), 'Locate: A protein subcellular localization database', *Nucl. Acids Res.* **34**((database issue)).

Haralick, R. (1979), 'Statistical and structural approaches to texture', *Proceedings of the IEEE* **67**(5), 768–804.

Haralick, R., Shanmugam, K. & Dinstein, I. (1973), 'Textural features for image classification', *IEEE Trans. On SMC* **SMC-3**(6), 610 – 621.

Huang, K. & Murphy, R. (2004), 'From quantitative microscopy to automated image understanding', *J. Biomed. Opt.* **9**(5), 893–912.

Khotanzad, A. & Hong, Y. H. (1990), 'Invariant image recognition by zernike moments', *IEEE Trans. Pattern Anal. Mach. Intell.* **12**(5), 489–497.

Lang, P., Yeow, K., Nichols, A. & Scheer, A. (2006), 'Cellular imaging in drug discovery', *Nature Reviews Drug Discovery 5* **5**, 343–356.

Matthiessen, M. (2003), 'Biowaredb: the biomedical software and database search engine', *Bioinformatics* **19**(17), 2319.

Murphy, R., Velliste, M. & Porreca, G. (2003), 'Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images', *J. VSLI Sig. Proc.* **35**, 311–321.

Pham, T. D., Crane, D. I., Tran, T. H. & Nguyen, T. H. (2004), 'Extraction of uorescent cell puncta by adaptive fuzzy segmentation', *Bioinformatics* **20**(14), 2189–2196.

Pham, T., Tran, D., Zhou, X. & Wong, S. (2006), 'Integrated algorithms for image analysis and identification of nuclear division for high-content cell-cycle screening', *Int. J. Computational Intelligence and Applications* **6**(1), 21–43.

Roques, E. & Murphy, R. (2002), 'Objective evaluation of differences in protein subcellular localisation', *Traffic* **3**, 61–65.

Stow, J. & Teasdale, R. (2005), *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*, John Wiley and Sons, chapter Expression and localization of proteins in mammalian cells.

Thiran, J.-P., Macq, B. & Mairesse, J. (1994), Morphological classification of cancerous cells, *in* 'Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference', Vol. 3, pp. 706–710.

Yu, C., Chen, Y., Lu, C. & Hwang, J. (2006), 'Prediction of protein subcellular localization', *Proteins* **Epub ahead of print**.

Zernike, F. (1934), *Physica* **1**(689).