BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 18 2013, pages 2343–2349
doi:10.1093/bioinformatics/btt392

Bioimage informatics

Advance Access publication July 8, 2013

Determining the subcellular location of new proteins from microscope images using local features

Luis Pedro Coelho[1,2,†], Joshua D. Kangas[1,2], Armaghan W. Naik[1,2], Elvira Osuna-Highley[3], Estelle Glory-Afshar[3], Margaret Fuhrman[4], Ramanuja Simha[5], Peter B. Berget[4,‡], Jonathan W. Jarvik[4] and Robert F. Murphy[1,2,3,4,6,*]

[1]Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA, [2]Joint Carnegie Mellon University-University of Pittsburgh Ph.D. Program in Computational Biology, Pittsburgh, PA 15213, USA, [3]Department of Biomedical Engineering and [4]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA, [5]Department of Computer and Information Sciences, University of Delaware, Newark, NJ 19716, USA and [6]Department of Machine Learning, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Evaluation of previous systems for automated determination of subcellular location from microscope images has been done using datasets in which each location class consisted of multiple images of the same representative protein. Here, we frame a more challenging and useful problem where previously unseen proteins are to be classified.

Results: Using CD-tagging, we generated two new image datasets for evaluation of this problem, which contain several different proteins for each location class. Evaluation of previous methods on these new datasets showed that it is much harder to train a classifier that generalizes across different proteins than one that simply recognizes a protein it was trained on.

We therefore developed and evaluated additional approaches, incorporating novel modifications of local features techniques. These extended the notion of local features to exploit both the protein image and any reference markers that were imaged in parallel. With these, we obtained a large accuracy improvement in our new datasets over existing methods. Additionally, these features help achieve classification improvements for other previously studied datasets.

Availability: The datasets are available for download at http://murphylab.web.cmu.edu/data/. The software was written in Python and C++ and is available under an open-source license at http://murphylab.web.cmu.edu/software/. The code is split into a library, which can be easily reused for other data and a small driver script for reproducing all results presented here. A step-by-step tutorial on applying the methods to new datasets is also available at that address.

Contact: murphy@cmu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 14, 2012; revised on June 12, 2013; accepted on July 3, 2013

*To whom correspondence should be addressed.
†Present address: European Molecular Biology Laboratory (EMBL), Heidelberg, Germany.
‡Present address: Department of Biological Sciences, University of the Sciences, Philadelphia, USA.

1 INTRODUCTION

Generation of images of cells and tissues is increasingly easy. With the advent of automated microscopes, the capability for data generation has out-stripped the capability for visual data analysis. This has led to extensive work on automated methods for interpreting microscope images.

The problem of classification of subcellular patterns has received particular attention, and a number of datasets and classifiers have been described. These datasets typically feature one different protein for each class of interest, with multiple images for the same tagged protein. On these datasets, better than human performance has been reported (Murphy *et al.*, 2003; Nattkemper *et al.*, 2003).

This previous work implicitly assumed that results obtained in those datasets can be generalized to the problem of classifying previously unseen proteins. In this work, we test this assumption using two new datasets where there are multiple proteins in each location class (and multiple images per protein). These datasets were created using NIH 3T3 cell lines expressing green fluorescent protein (GFP)-tagged proteins created by CD-tagging (García Osuna *et al.*, 2007; Jarvik *et al.*, 2002).

We tested classifiers using a cross-validation protocol whereby images from the same protein are never present in both the training and testing sets. This is a stricter proxy for cross-protein generalization than randomizing by image, and guards against the possibility that learning is based on properties of the tagging method (e.g. intensity) or too specific to the protein in question (e.g. a particular subpattern of an organelle). With this protocol and existing methods, generalization accuracy was only 60% for our new datasets.

We therefore investigated whether improved generalization could be obtained using alternative feature representations of the images. Many previous systems use image-level features such as texture features (Chebira *et al.*, 2007; Huang *et al.*, 2003; Nanni and Lumini, 2008; Nanni *et al.*, 2010; Shamir *et al.*, 2008b), but some specialized features for cell images have also been proposed (Boland and Murphy, 2001), including features for single-cell regions [in fact, historically, classification on cell-segmented images was reported first (Boland *et al.*, 1998)].

In the computer vision literature, local features, such as the scale-invariant feature transform, introduced by Lowe (1999), have shown good results in many settings. They have not been widely used in bioimage analysis [there are a few uses of patch-based methods, a basic form of these features (Huh *et al.*, 2009; Marée *et al.*, 2007)]. Object-level features, which can be seen as a form of local features, were used for subcellular location unmixing, both in supervised and unsupervised modes (Coelho *et al.*, 2010a; Peng *et al.*, 2010; Zhao *et al.*, 2005).

Local features, as presented in the literature, are generally defined on a gray-scale image and do not take advantage of the multiple image channels frequently acquired by fluorescent microscopy. There is some work on natural scene color images (van de Sande *et al.*, 2010), but it does not directly apply to fluorescence microscopy images for analyzing subcellular patterns where one channel is privileged (depicting the protein distribution of interest) and others serve as references. Naturally, the simplest protocol is to ignore all but the primary channel. However, the use of a reference channel can provide additional important information, particularly at the local level. For example, we could distinguish between two vesicle classes that appear similar in the primary (protein) channel but differ in distance from the nucleus because the region containing vesicles will appear differently in the reference nuclear channel. We present a simple protocol to take these reference channels into account. Using these features, we obtain a large accuracy gain on our datasets. We also use other datasets to further validate the value of the features and find that they lead to good results in all tested datasets.

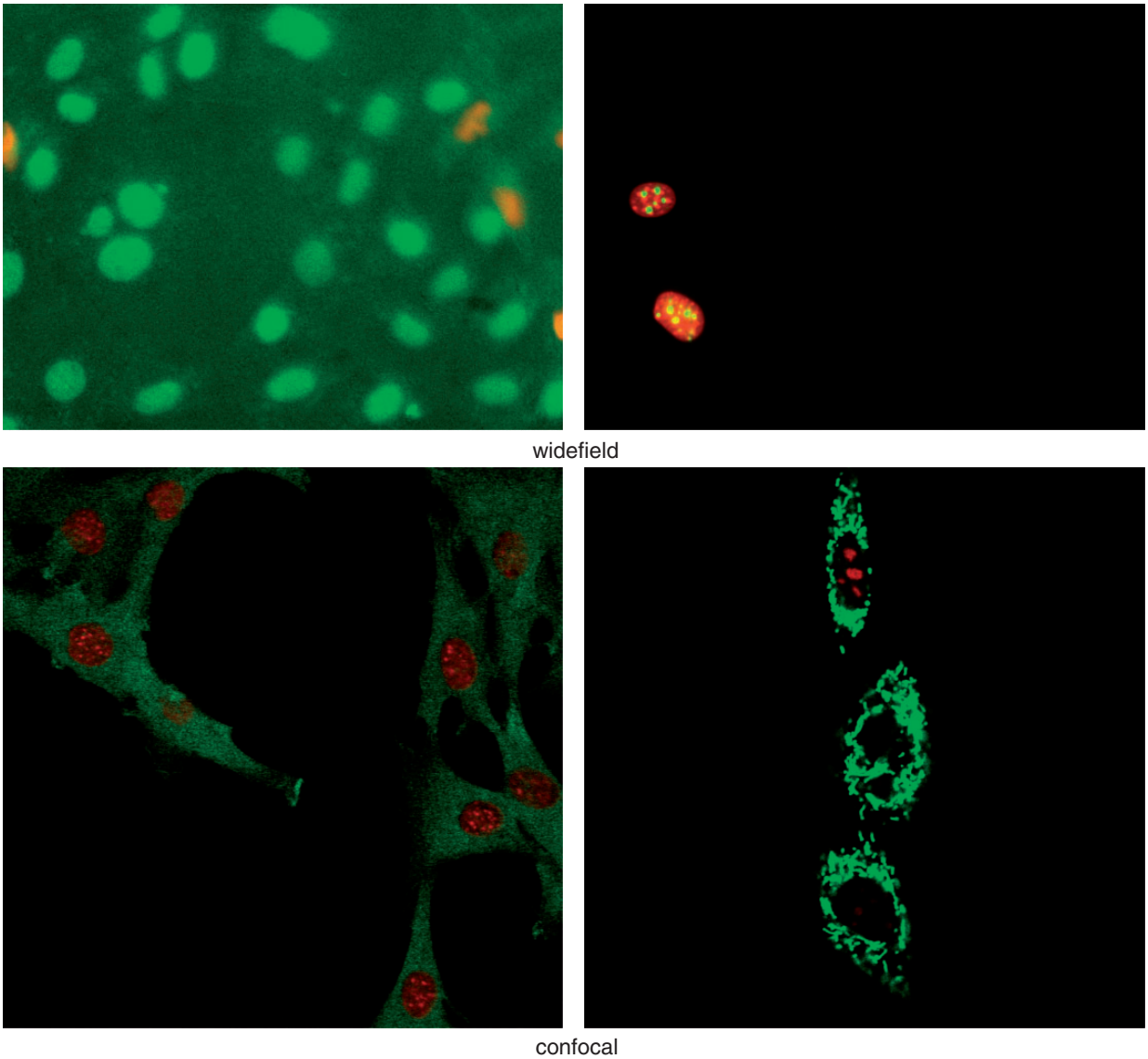

Fig. 1. Examples of RandTag datasets. Top row consists of widefield images (nuclear pattern on the left, and nucleoli on the right), second row of confocal images (cytoplasmic pattern on the left, mitochondrial pattern on the right). Images are false color: the red channel shows the nuclear marker Hoechst, the green channel is the GFP-tagged protein. Images shown are the first image in their classes and have not been manually chosen. The widefield images were automatically acquired and the quality is lower than if they had been manually acquired. Images have been contrast stretched for publication

2 DATASETS

2.1 RandTag datasets

Two datasets are introduced in this article, both from the RandTag (RT) project (García Osuna *et al*., 2007). The first dataset consists of widefield images, the second of confocal images.

The widefield images were collected with an automated microscope. Therefore, the quality of the images is variable. As a preprocessing step, images that are completely out-of-focus or empty of cells were removed. The confocal images were acquired manually and are of higher quality. Examples are shown in Figure 1. These examples were not chosen as particularly pleasant looking, but are representative of the images in the dataset.

The images were labeled by three experts (The experts were L.P.C., E.O.H. and E.G.A. for the widefield dataset, and L.P.C., E.G.A. and A.N. for the confocal dataset.) using a protocol where the experts first labeled the images independently and were then given an opportunity to change their minds given the other labelings. Only images where all experts agreed after this second step were retained. Table 1 shows summary statistics for these two datasets.

The two datasets contain multiple images of the same protein, and multiple proteins per location class. Most other subcellular location datasets contain multiple images per protein but only one protein for each location class (the exception is the Locate database).

2.2 Other datasets

We present the main properties of the datasets in Table 2. All are publicly available.

2.2.1 Murphy Lab 2D HeLa Dataset The Murphy Lab 2D HeLa dataset is by now a benchmark in the field, used by many researchers (Boland and Murphy, 2001; Chebira *et al*., 2007; Huang and Murphy, 2004; Lin *et al*., 2007; Marée *et al*., 2007; Nanni *et al*., 2010; Rajapakse, 2008; Shamir *et al*., 2008b). The dataset contains approximately 100 images collected by widefield fluorescence microscopy (with nearest neighbor deconvolution) for each of 10 subcellular patterns. Nanni *et al*. (2010) obtained the best reported results on this dataset, 96% accuracy, using a combination of texture and other features.

2.2.2 Locate endogenous and Locate transfected These images were collected by widefield microscopy to detect 10 *endogenous* proteins or 11 *transfected* proteins (Hamilton *et al*., 2007). Each dataset contains approximately 50 images for each subcellular patterns.

2.2.3 Locate Confocal Aturaliya *et al*. (2006) presented a collection of mouse membrane-bound proteins imaged with confocal microscopy. The images are available online in the locate database (Available at http://locate.imb.uq.edu.au/). It consists of 6985 images of 2047 different mouse proteins expressed in HeLa cells. The images were manually annotated and most proteins are labeled with more than one location. We are not aware of previous work in automatic classification of these images.

2.2.4 Image Informatics and Computational Biology Unit (IICBU) 2008 Benchmark The IICBU 2008 collection of datasets includes several collections of bioimages with different properties, which was intended for testing computer vision algorithms (Shamir *et al*., 2008a) (The datasets are available at http://ome.grc.nia.nih.gov/iicbu2008.).

We used the fluorescent microscopy datasets (the collection includes other modalities). This collection includes the HeLa 2D dataset, but it includes a version without dna channel. Our experiments were on the original, two channel, dataset.

2.2.5 Human Protein Atlas The Human Protein Atlas (HPA) contains a collection of confocal images of immuno-stained proteins in human cells, with visual annotation (Barbe *et al*., 2008). We used those images where the visual annotation is to a single location class (Li *et al*., 2012a, b).

3 MATERIALS AND METHODS

3.1 SURF-Ref

Speeded-Up Robust Features (SURF) are calculated by a two pass algorithm. The first pass detects interest points by using an approximate Gaussian blob detector. These interest points are localized in both space (i.e. at a specific pixel location) and scale (i.e. they have an automatically determined size). The second pass computes 64 descriptors at each interest point.

Table 1. Properties of RandTag datasets

	UL	NO	N	M	G	Cyto	PM	Lyso	Cytosk	ER
Widefield										
Number of proteins	12	5	14	8	3	10	3	4	9	3
Number of images	254	113	255	175	63	155	51	69	197	50
Confocal										
Number of proteins	5	3	8	12	4	8	3	3	18	3
Number of images	20	17	40	60	16	34	12	12	80	13

Note: UL, unlabeled; NO, nucleoli; N, nuclear; M, mitochondria; G, Golgi; Cyto, cytoplasmic; PM, plasma membrane; Lyso, lysosome; Cytosk, cytoskeleton; ER, endoplasmic reticulum.
Shown are number of proteins (first line) and images (second line) per class.

Table 2. Dataset statistics

Name	Number of images	Number of classes	Reference
RT-widefield	1382	10	
RT-confocal	304	10	
HeLa2D	862	10	Boland and Murphy, 2001
LOCATE-transfected	553	11	Hamilton *et al*., 2007
LOCATE-endogenous	502	10	Hamilton *et al*., 2007
Binucleate	41	2	Shamir *et al*., 2008a
CHO	327	5	Shamir *et al*., 2008a
Terminalbulb	970	7	Shamir *et al*., 2008a
RNAi	200	10	Shamir *et al*., 2008a
HPA	1842	13	Barbe *et al*., 2008

Note: The first two datasets, from the RandTag project, are introduced in this article; the others were previously described elsewhere.

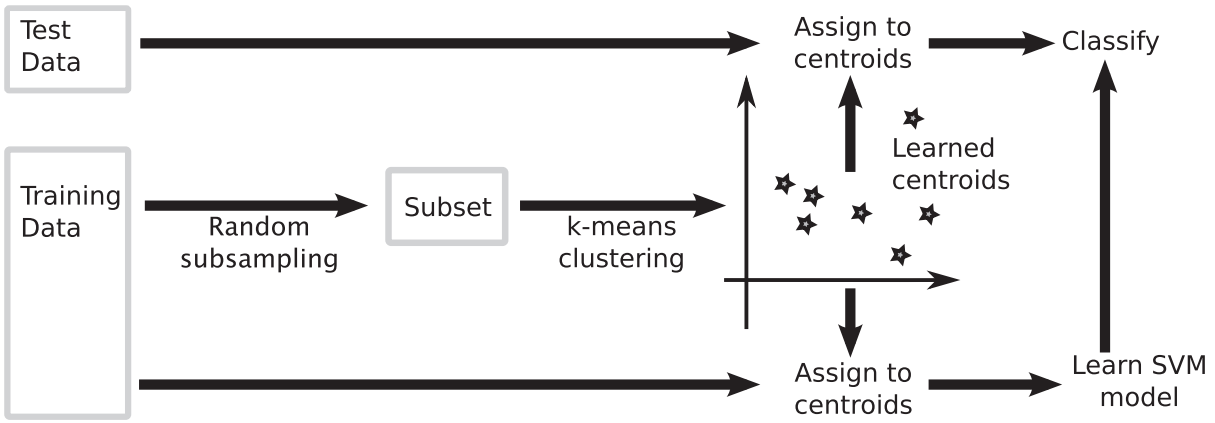

Fig. 2. Overview of local feature-based classification. The training set is subsampled and centroids are obtained from this smaller dataset. All the training points are then projected to their nearest centroid and a SVM is trained to classify the per-image histograms. At testing time, the same centroids and SVM are used

SURF works on a single channel (a gray-scale image), while bioimages are frequently multichannel: in addition to the primary channel, one or more reference channels are often acquired in parallel. Typically the primary channel is a protein image and a nuclear marker is used as a reference. SURF as presented in the literature can only be applied to the primary channel, discarding valuable information.

The protocol to incorporate the reference information is as follows: run the *point detection on the primary channel* and compute feature descriptors *on both channels* independently. The feature descriptor for each point is then the concatenation of both descriptors.

3.1.1 Baseline feature sets As a baseline feature set, we used a global feature set, which includes Haralick texture features (Haralick *et al.*, 1973), parameter-free Threshold Adjacency Statistics (Coelho *et al.*, 2010b; Hamilton *et al.*, 2007), object and skeleton features (Boland and Murphy, 2001), and overlap features (Newberg and Murphy, 2008).

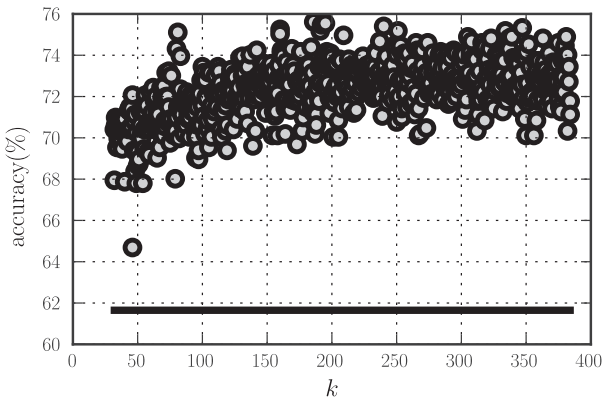

Fig. 3. Results of classification as a function of the number of clusters k. Each dot is the result of one clustering of the data (differing by a different number of clusters and a different initial set of clusters). The solid line is the baseline accuracy (using global instead of local features)

3.2 Classification

Computing local features leads to several hundred descriptor vectors per image. To use these in classification, we clustered the descriptor vectors. This process assigns each descriptor to a cluster index. We represent an image as a normalized histogram of membership in the various clusters (Willamowski *et al.*, 2004). This is known as the "bag of visual words" model.

The first step is to obtain a set of k centroids, using k-means. This algorithm takes two parameters: k, the number of clusters; and an initial set of centroids. This is implemented by setting the random number generator seed to different values and randomly selecting elements. For efficiency, centroids were obtained from a fraction (1/16th) of the data. All feature vectors are then assigned to the closest centroid. The resulting histogram can then be used with a standard support vector machine (SVM) classifier. Figure 2 provides an overview of the method.

As Figure 3 shows, there is a large variation in accuracy for different choices of the random seed even for the same value of k: the difference between the highest scoring and the lowest can be as high as six percentage points. Furthermore, as Figure 4 shows, the typical solution of minimizing the value of the Akaike information criterion (AIC) introduced by Akaike (1974), will not necessarily lead to a high accuracy. In fact, high AIC leads to high accuracy.

Given the results in Figures 3 and 4, we used $k = n/4$ clusters, where n is the number of images in the training set. For the RT-widefield dataset shown in the Figure, this corresponds to circa 310 clusters. Supplemental Figure S1 repeats the calculation for the other datasets and confirms the value of this rule. We used a different random initialization for each point.

The models learned are SVM based after feature normalization and selection using stepwise discriminant analysis (Jennrich, 1977a, b). A radial basis function kernel is used for the SVM, and an inner loop of cross-validation is used to select the hyperparameters. For the Locate database, which is a multilabel dataset, we used a separate classifier per label; for all other datasets, we used the "one versus one" strategy to convert binary classification into multiclass learning (These are the default settings

for the milk Python machine learning library used in this work, no settings were changed or tuned).

3.3 Significance computation

For the measurement of statistical significance, we used a Bayesian approach. Given a dataset of size n, on which two algorithms correctly classify c_0 and c_1 elements, respectively, we assume that each algorithm has an underlying accuracy of r_i and compute the $P(r_0 > r_1 | c_0, c_1, n)$, the probability that the first algorithm is better than the second one. We also assume that the performance of the algorithms is independent,

$$p(c_0, c_1, n | r_0, r_1) = p(c_0, n | r_0) p(c_1, n | r_1), \quad (1)$$

and compute

$$\frac{\int_0^1 \int_0^1 [\![r_0 > r_1]\!] p(c_0, n | r_0) p(c_1, n | r_1) dr_0 dr_1}{\int_0^1 \int_0^1 p(c_0, n | r_0) p(c_1, n | r_1) dr_0 dr_1}. \quad (2)$$

To be able to numerically obtain a value for (2), we model the accuracy of each classifier with a binomial distribution:

$$p(c, n | r) = r^c (1 - r)^{n-c}. \quad (3)$$

In this framework, higher values are better, which is the opposite of the traditional statistical practice. Therefore, we report $1 - P(r_0 > r_1 | c_0, c_1, n)$ as a significance value. If the assumptions (1) and (3) are accepted, this significance value is the probability of making a Type I error (i.e. erroneously rejecting the null hypothesis that $r_0 \leq r_1$).

The Locate database needs to be handled differently as its proteins are annotated with multiple labels. The system we built learns a binary classifier for each label and, at evaluation time, outputs all the labels whose corresponding binary classifier returned a positive label. Each binary classifier was learned independently. For evaluation, the above framework is not directly applicable and we measured and report the F_1 score.

3.4 Cross-validation

All results were obtained using cross-validation. Ten-folds were used, except in the cases where the smallest class had less than 10 objects. In that case, the number of folds was set to the minimum class size.

When handling the RandTag datasets with multiple images of the same protein, we can perform cross-validation in two ways:

(1) **Per image**, in which we group the images into 10-folds without taking the depicted protein into account.

(2) **Per protein**, in which we group the proteins into 10-folds. In this setting, there were never any images in training and testing from the same protein. Accuracy is still reported on a per-image level (the fraction of images that were correctly classified).

Software All software presented was developed in Python and C++ and incorporates code from dlib (Dlib's webpage is at http://www.dlib.net) by David King and LIBSVM by Chang and Lin (2001) for feature computation and classification, respectively. The software is designed to be easily reused in new datasets (Coelho, 2013).

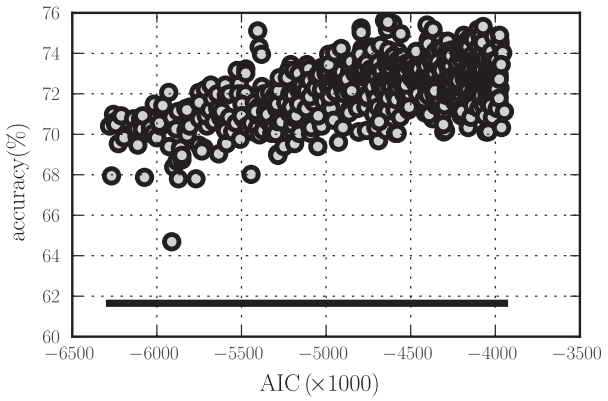

Fig. 4. Results of classification as a function of the Aikake information criterion. Each dot is the result of one clustering of the data (differing by a different number of clusters and a different initial set of clusters). Note that AIC is typically *minimized*

4 RESULTS

4.1 Generalization to new proteins

As described above, the RandTag datasets have images from several proteins in each dataset. Cross-validation over proteins is a stricter test of generalization capabilities and it was expected that it would lead to lower accuracies than the cross-validation over images (where training and test sets have different images of the same labeled protein). Table 3 shows that the resulting difference in accuracy is large: a drop of 22 percentage points (84–62%).

Even with multiple proteins per class, having examples from the same protein in training and testing results in high measured accuracies. These values (91–88%) are close to what is typically reported in subcellular location problems.

However, when results are evaluated using the stricter cross-validation protocol, accuracy values are much lower, circa 60% for the baseline results. The differences in results with the two forms of cross-validation are statistically significant (at the 10^{-51} and 10^{-10} levels, for the widefield and confocal datasets, respectively).

The HPA dataset also contains multiple proteins per class and a small number of images of each protein, often only two.

Table 3. Comparison of per-protein and per-image cross-validation

Dataset	Method	Baseline	SURF-ref+ baseline
RT-widefield	Per image	84.2	91.1
RT-widefield	Per protein	61.6	70.3
RT-confocal	Per image	83.9	88.8
RT-confocal	Per protein	59.9	65.5
HPA	Per image	76.1	82.8
HPA	Per protein	68.2	78.3

Note: Shown are accuracies, in percentage, obtained either with per-protein or per-image cross-validation on two feature sets.

Therefore, we used 2-fold cross-validation. The results in this dataset confirm that performing cross-validation per protein results in lower accuracy than cross-validating per image. The difference of 12 percentage points is significant at the 10^{-7} level.

4.2 SURF-ref

Table 4 summarizes the results obtained. On five of the datasets, using SURF variations shows a statistically significant improvement over the baselines used. On the other datasets, the results are not statistically distinguishable from the baseline.

The worst results are obtained in the *RNAi* dataset, where local features alone perform much worse (significant at the 2.5×10^{-8} level). However, once the baseline is added, the results are indistinguishable from the baseline. Therefore, we recommend the use of all features combined.

4.2.1 Computational Costs SURF-ref is efficient in terms of computational time. On average, our implementation requires 7 s per image for both interest point detection and feature descriptor computation. Images in this case are 768×1024 pixels large.

As part of interest point detection, each point is ranked according to a metric of how strongly it matches the approximate filter used—see the original SURF article for details (Bay *et al.*, 2008). For large datasets, the computed feature data can be overwhelming. Therefore, we limited the number of interest points per image to 1024 (which are the 1024 highest matches according to the metric alluded to above). The traditional SURF consists of 64 descriptor values. In addition, we save the location, scale, angle and match strength for 70 floating point values per interest point.

Table 4. Summary of results for all datasets

Dataset	Baseline	SURF	SURF-ref	Local[a] + baseline	Significance
HeLa2D	86.0	88.7	90.4	**94.4**	2.4×10^{-9}
RT-widefield (per image)	84.2	79.6	85.7	**91.1**	4.0×10^{-8}
RT-widefield	61.6	67.5	**71.9**	70.3	1.1×10^{-6}
RT-confocal (per image)	83.9	80.9	84.2	**88.8**	0.04
RT-confocal	59.9	**72.0**	62.8	65.5	0.08
LOCATE-transfected	75.4	84.8	84.8	**88.1**	3.1×10^{-8}
LOCATE-endogenous	74.5	91.2	91.2	**95.6**	1.8×10^{-22}
Binucleate	85.4	**95.1**		**95.1**	0.08
CHO	96.6	96.9		**98.5**	0.08
Terminal bulb	**45.8**	32.4		44.6	0.31
RNAi	**72.0**	43.0		67.5	0.17
HPA	69.9	67.8	78.0	**78.9**	5.0×10^{-10}
LOCATE[b]	66	62		**69**	[c]

Shown are accuracy (as a percentage) and significance as defined in Section 3.3. The baseline is cell-level features for the HeLa 2D dataset and field-level features for all other sets. Significance is on the difference between the baseline and the bolded column.

[a]For datasets with a dna channel, SURF-ref was used, for those without a dna channel, SURF features were used.

[b]As discussed in the text, F_1 score is shown.

[c]The significance calculation is not directly applicable to F_1 scores.

5 DISCUSSION

This work frames the subcellular location problem as recognizing *different* proteins in the same class. While this may have been implicit in previous work, it was not directly tested by datasets with a single representative protein per location class.

We introduce two new datasets, which contain multiple proteins per class (and multiple images per protein). We observed that when cross-validation was performed over proteins, the resulting accuracy was much lower than when it was performed over images (where it is comparable with other datasets). This is intuitive as it is an easier problem to recognize proteins that are in the training set than proteins that are only in the same class (in particular, in the first case, it is possible that the system distinguishes the proteins by artifacts of the tagging or variation in subpatterns).

Our data show that it is incorrect to assume that the high accuracy values obtained in datasets composed of multiple images of the same protein imply that the system would generalize well to other proteins in the same location class. Our datasets are publicly available. There is still a lot of room for improvement in accuracy and we hope that other researchers will test their methods on this harder problem.

We also introduce a new methodology for classification of subcellular location patterns, which is based on interest point detection and local feature analysis. We developed a protocol to integrate the information in reference channels (which are typically acquired in parallel to the protein of interest). We implemented this method based on SURF, but the protocol is a generic method and could be applied to other local feature sets, such as scale-invariant feature transform (Lowe, 1999) or any combination of detector and descriptor. On our new datasets, these methods performed better than the traditional whole-field features by 10 percentage points (a difference that is highly statistically significant).

We tested these features on traditional datasets as well. On these, the baseline methods already perform well and there was less room for improvement. In four cases, the results are statistically indistinguishable from the baseline. It should be noted, though, that in no dataset did we observe that adding the local features lead to a statistically distinguishable worse outcome. These features have the further advantage that they are computed on the raw images without any pre-processing such as background correction or contrast enhancement. No tuning is necessary for adapting to new datasets and it is flexible for application to large datasets. Therefore, we recommend that local features with reference information be added to the standard toolkit for bioimage classification.

ACKNOWLEDGEMENTS

We thank the HPA project team, especially Emma Lundberg, for providing the high-resolution confocal microscopy images used for the HPA dataset, and Jieyue Li for preparing this dataset for computational analysis.

Funding: National Institute of General Medical Sciences (R01 GM075205 to R.F.M.); National Institutes of Biological Imaging and Bioengineering (T32 EB009403-01 to A.W.N. by

training grant); Fundação para a Ciência e Tecnologia (SFRH/BD/37535/2007 to L.P.C.); Siebel Scholars Foundation.

Conflict of Interest: none declared.

REFERENCES

Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. Automatic Control*, **19**, 716–723.

Aturaliya,R.N. *et al.* (2006) Subcellular localization of mammalian type II membrane proteins. *Traffic*, **7**, 613–25.

Barbe,L. *et al.* (2008) Toward a confocal subcellular atlas of the human proteome. *Mol. Cell. Proteomics*, **7**, 499–508.

Bay,H. *et al.* (2008) Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.*, **110**, 346–359.

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

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

Chang,C.C. and Lin,C.J. (2001) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Sys. Technol.*, **3**, 1–30.

Chebira,A. *et al.* (2007) A multiresolution approach to automated classification of protein subcellular location images. *BMC Bioinformatics*, **8**, 210.

Coelho,L.P. *et al.* (2010a) Quantifying the distribution of probes between subcellular locations using unsupervised pattern unmixing. *Bioinformatics*, **26**, i7–i12.

Coelho,L.P. *et al.* (2010b) Structured literature image finder: extracting information from text and images in biomedical literature. *Lect. Notes Comput. Sci.*, **6004**, 23–32.

Coelho,L.P. (2013) Mahotas: open source software for scriptable computer vision. *J. Open Res. Softw.*, **1**.

García Osuna,E. *et al.* (2007) Large-scale automated analysis of location patterns in randomly tagged 3T3 cells. *Ann. Biomed. Eng.*, **35**, 1081–1087.

Hamilton,N.A. *et al.* (2007) Fast automated cell phenotype image classification. *BMC Bioinformatics*, **8**, 110.

Haralick,R.M. *et al.* (1973) Textural features for image classification. *IEEE Trans. Syst. Man Cybern.*, **3**, 610–621.

Huang,K. and Murphy,R.F. (2004) Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinformatics*, **5**, 78.

Huang,K. *et al.* (2003) Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images. In *Proceedings of SPIE 4962, Manipulation and Analysis of Biomolecules, Cells, and Tissues*. SPIE, United States, pp. 307–318.

Huh,S. *et al.* (2009) Efficient framework for automated classification of subcellular patterns in budding yeast. *Cytometry*, **75**, 934–40.

Jarvik,J.W. *et al.* (2002) *In vivo* functional proteomics: mammalian genome annotation using CD-tagging. *Biotechniques*, **33**, 852–854, 856, 858–60 passim.

Jennrich,R.I. (1977a) Stepwise discriminant analysis. In Enslein,K. *et al.* (ed.) *Statistical Methods for Digital Computers*. John Wiley & Sons, New York.

Jennrich,R.I. (1977b) Stepwise Regression. In Enslein,K. *et al.* (ed.) *Statistical Methods for Digital Computerss*. John Wiley & Sons, New York.

Li,J. *et al.* (2012a) Automated analysis and reannotation of subcellular locations in confocal images from the human protein atlas. *PLoS One*, **7**, e50514.

Li,J. *et al.* (2012b) Protein subcellular location pattern classification in cellular images using latent discriminative models. *Bioinformatics*, **28**, i32–i39.

Lin,C.C. *et al.* (2007) Boosting multiclass learning with repeating codes and weak detectors for protein subcellular localization. *Bioinformatics*, **23**, 3374–3381.

Lowe,D.G. (1999) Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*. Vol. **2**, IEEE, New York, pp. 1150–1157.

Marée,R. *et al.* (2007) Random subwindows and extremely randomized trees for image classification in cell biology. *BMC Cell Biol.*, **8(Suppl. 1)**, S2.

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

Nanni,L. and Lumini,A. (2008) A reliable method for cell phenotype image classification. *Artif. Intell. Med.*, **43**, 87–97.

Nanni,L. *et al.* (2010) Novel features for automated cell phenotype image classification. *Adv. Exp. Med. Biol.*, **680**, 207–13.

Nattkemper,T.W. *et al.* (2003) Human vs machine: evaluation of fluorescence micrographs. *Comput. Biol. Med.*, **33**(1), 31–43.

Newberg,J. and Murphy,R.F. (2008) A framework for the automated analysis of subcellular patterns in human protein atlas images. *J. Proteome Res.*, **7**, 2300–8.

Peng,T. *et al.* (2010) Determining the distribution of probes between different subcellular locations through automated unmixing of subcellular patterns. *Proc. Natl. Acad. Sci. USA*, **107**, 2944–2949.

Rajapakse,J.C. (2008) Protein localization on cellular images with Markov random fields. *IEEE Int. Joint Conf. Neural Netw.*, 2127–2132.

Shamir,L. *et al.* (2008a) IICBU 2008: a proposed benchmark suite for biological image analysis. *Med. Biol. Eng. Comput.*, **46**, 943–947.

Shamir,L. *et al.* (2008b) Wndchrm—an open source utility for biological image analysis. *Source Code Biol. Med.*, **3**, 13.

van de Sande,K.E.A. *et al.* (2010) Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, **32**, 1582–1596.

Willamowski,J. *et al.* (2004) Categorizing nine visual classes using local appearance descriptors. In *ICPR 2004 Workshop Learning for Adaptable Visual Systems*. Cambridge, UK.

Zhao,T. *et al.* (2005) Object type recognition for automated analysis of protein subcellular location. *IEEE Trans. Image Process.*, **14**, 1351–9.