Bioinformatics, 31(7), 2015, 1111–1119
doi: 10.1093/bioinformatics/btu772
Advance Access Publication Date: 19 November 2014
Original Paper

OXFORD

Bioimage informatics

Bioimaging-based detection of mislocalized proteins in human cancers by semi-supervised learning

Ying-Ying Xu[1], Fan Yang[1], Yang Zhang[2] and Hong-Bin Shen[1,2,*]

[1]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China and [2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed.

Associate Editor: Robert F. Murphy

Received on June 30, 2014; revised on November 10, 2014; accepted on November 15, 2014

Abstract

Motivation: There is a long-term interest in the challenging task of finding translocated and mislocated cancer biomarker proteins. Bioimages of subcellular protein distribution are new data sources which have attracted much attention in recent years because of their intuitive and detailed descriptions of protein distribution. However, automated methods in large-scale biomarker screening suffer significantly from the lack of subcellular location annotations for bioimages from cancer tissues. The transfer prediction idea of applying models trained on normal tissue proteins to predict the subcellular locations of cancerous ones is arbitrary because the protein distribution patterns may differ in normal and cancerous states.
Results: We developed a new semi-supervised protocol that can use unlabeled cancer protein data in model construction by an iterative and incremental training strategy. Our approach enables us to selectively use the low-quality images in normal states to expand the training sample space and provides a general way for dealing with the small size of annotated images used together with large unannotated ones. Experiments demonstrate that the new semi-supervised protocol can result in improved accuracy and sensitivity of subcellular location difference detection.
Availability and implementation: The data and code are available at: www.csbio.sjtu.edu.cn/bioinf/SemiBiomarker/.
Contact: hbshen@sjtu.edu.cn
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Knowing the subcellular locations of proteins in human cancer tissues can improve the understanding of protein functions and cancer pathogenesis (Chou and Shen, 2008; Pierleoni *et al.*, 2006). It has been demonstrated that the translocation of protein might be a signal of cancer (Hanash *et al.*, 2008; Hung and Link, 2011). The cyclin D1 protein is an example: it shuttles between the nucleus and cytoplasm in a healthy cell and the reduction of exportation from the nucleus can lead to overexpression in the nucleus and the inactivation of the tumor-suppressing protein retinoblastoma (Benzeno *et al.*, 2006; Gladden and Diehl, 2005). Accurately detecting protein translocations in human cancer tissues can thus be of important help for clinical diagnosis and treatment. Because traditional wet lab experiments are expensive in time and costs (Eliceiri *et al.*, 2012; Winski *et al.*, 2002), automated methods are highly desired for handling the increasing amounts of biomedical data.

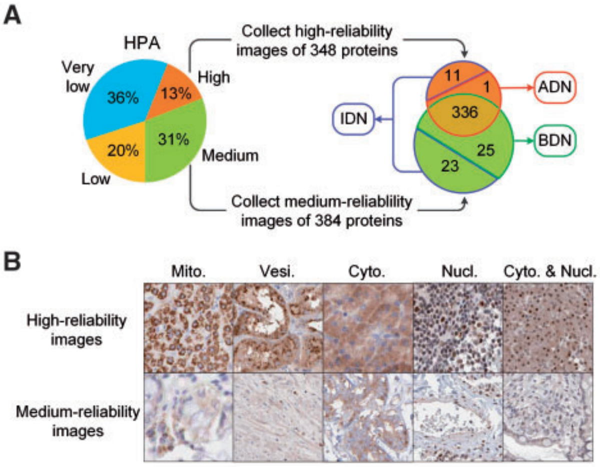
Despite its importance, only a few studies have reported automated methods to detect translocation details in cancerous tissues until now. One reason is that sequence-based analysis by itself is not sensitive enough for detection of protein translocation as translocation can be strongly effected by mutations outside the target sequence. For example, mutations in nucleoporin complexes can have dramatic effects on the nuclear localization of multiple other proteins (Hung and Link, 2011). Due to recent advances in microscopic imaging, image-based pattern analysis methods have gained popularity due to the intuitive and detailed information the images contain. For example, the Murphy group discussed the potential applications of their models based on automated analysis of fluorescence microscopy images to the analysis and classification of skin cancers (Glory and Murphy, 2007; Murphy, 2004). Rizzardi *et al.* (2012) compared the abilities of automated image analysis and pathologist visual examination in quantifying protein expression in ovarian cancer. Recently, our group developed a multilabel subcellular location predictor, *i*Locator, and identified several translocated proteins as potential cancer biomarkers (Xu *et al.*, 2013).

To compare the localization difference of a protein in normal and cancerous tissues, we have to know its subcellular locations in both normal and cancerous states first. This can be achieved either through wet-lab experiments or computational predictions. Since image data with experimentally annotated subcellular locations in cancerous states are rare, prediction models have been used instead, especially in the large-scale screening. Due to the lack of location labels for proteins in cancerous states, however, most of the existing methods employed an approach named transfer learning, where models are first trained on proteins in normal tissues and then used to predict the localization of proteins in cancerous tissues (Eliceiri *et al.*, 2012; Xu *et al.*, 2013). The performance of these approaches is poor, where one reason is the subtle differences in subcellular location patterns between cancer and normal states, which are influenced by cell mutations and morphological changes.

In fact, there are a large number of images of proteins with cancerous tissues. The Human Protein Atlas (HPA, version 11, http://www.proteinatlas.org/) (Uhlen *et al.*, 2010) database, for example, currently contains more than 1 million immunohistochemistry (IHC) microscopy images of proteins in cancerous tissues. But due to the lack of explicit subcellular annotations, no attempt has been made in using these images from cancerous tissues for constructing supervised models for cancer localization prediction.

To address the issues, we present a heuristic semi-supervised learning framework for subcellular location prediction by taking advantage of the unannotated cancer samples in developing predictors. The key advantage of the proposed semi-supervised method, in comparison to the traditional supervised learning algorithms, is that it can train prediction models with only a few labeled image samples and a large pool of unlabeled samples (Hady and Schwenker, 2013). An iterative and incremental strategy was designed to select unlabeled samples into the training set. To choose the most discriminative samples, we developed three different training modes: a single-training model consisting of only one classifier (McLachlan, 1975), a co-training model consisting of two classifiers (Cohen, 2002) and a tri-training model consisting of three classifiers (Zhou and Li, 2005). Also, as the incorporation of prior knowledge can improve the performance of semi-supervised methods (Liston and Stone, 2008), we took the location information from the corresponding normal tissues as prior knowledge to guide the selection process.

Another advantage of the proposed semi-supervised framework is that the training samples become typically much more enriched compared with the traditional supervised learning. First, it selected useful lower-quality images from normal tissues for training. In general, researchers prefer using well-stained images in the training set (Newberg and Murphy, 2008; Xu *et al.*, 2013). But selecting only high-quality images may introduce bias into modeling because the number of high expression level images in the HPA is relatively small (Fig. 1A). Therefore, instead of being discarded, some images of normal tissues with weak expression levels were selected for use in training by the semi-supervised strategy used in this study. Then, also the large cancer dataset was used for model construction by using the semi-supervised strategy of this study, which results in a much larger dataset useable for model construction. The final predictor by the semi-supervised training can be used for images from both normal and cancer tissues. We have tested the method on an independent cancer biomarker dataset composed of translocated or mislocated proteins, which have been confirmed by biological experiments. Comparing the prediction results from models trained with and without data from cancerous tissues shows that using the cancer data improves the sensitivity of detecting protein translocations or mislocations in human cancer tissues.


A
HPA
Very low
36%
High
13%
31%
Medium
20%
Low
Collect high-reliability images of 348 proteins
Collect medium-reliablility images of 384 proteins
11
1
336
25
23
ADN
IDN
BDN
B
Mito.
Vesi.
Cyto.
Nucl.
Cyto. & Nucl.
High-reliability images
Medium-reliability images

C

		Expression reliability	Number of proteins	Number of images	Annotation source
Normal	A dataset (ADN)	H	337	4,224	HPA
	B dataset (BDN)	M	361	5,531	HPA
	Independent dataset (IDN)	H or M	34	400	HPA
Cancer	C dataset (CDC)	H	348	21,920	None
Independent biomarker dataset (IBD)		H or M	10	147	Literatures

Fig. 1. Data collection. (**A**) Process of collecting normal datasets. The pie chart (left) shows the percentages of normal protein images with different levels of expression reliability in HPA version 11. Protein images with high and medium reliability corresponding to six subcellular locations in 11 tissues were collected (Supplementary Table S1). The overlapping part of two circles represents overlap on the protein level because some proteins have different reliability levels in different tissues. For example, ornithine carbamoyltransferase is one such protein because its reliability of expression in liver is high while in the colon it is medium. The IDN is randomly selected from the non-overlapping proteins and avoids protein overlap with the training set. The ADN and BDN are composed of the remaining images with high and medium reliability levels, respectively. Note that IDN has intersection with neither ADN nor BDN at the protein level. (**B**) Some examples of protein images with different reliability levels and subcellular locations. (**C**) Summary of all the datasets used in this study. The CDC is built by images of 348 proteins in cancerous tissues, where the 348 proteins are proteins whose images in corresponding normal tissues are of high reliability of protein expression. The IBD contains 10 proteins that were reported being translocated in human cancers by the literatures (Supplementary Table S2). In the column of expression reliability, H means high and M means medium

2 Methods

2.1 Datasets

Our image data were extracted from the HPA database, where the reliability of the annotated protein expression data is scored as high, medium, low and very low quality, depending on the consistency of the expression profile with the available literature (Uhlen *et al.*, 2010). To compromise between image quality and model generality, we used the top two categories of IHC images, i.e. high and medium reliability levels (Fig. 1). Three normal datasets with high and medium reliability levels were used, where the datasets ADN and BDN are for training and the independent dataset (IDN) is for testing. In the experiments, we evaluated different supervised and semi-supervised algorithms on the IDN, which is not contained in the training set for all the training stages. It should be noted that not all of the medium quality images in the BDN dataset were used. Only those that are capable of improving model performance were selected according to our semi-supervised strategy.

The cancer dataset (CDC) contains 21 920 images, which were selectively added into the training set to improve prediction performance for proteins in cancerous tissues. One hundred and forty-seven images corresponding to 10 biomarker proteins in normal and cancerous tissues were retrieved from the HPA database and composed the independent biomarker dataset (IBD) dataset. This dataset was used to validate whether the sensitivity of detecting the subcellular location difference between normal and cancer statuses is improved by incorporating the cancer data into training.

All these datasets are from 11 human tissues, i.e. breast, colon, liver, lung, lymph node, ovary, pancreas, prostate, kidney, thyroid gland and urinary bladder. They involve six major cellular organelles: cytoplasm, endoplasmic reticulum, Golgi apparatus, mitochondria, nucleus and vesicles. Among all the proteins in our datasets, 26% are multilabel proteins that belong to two or three organelles simultaneously. It should be noted that the label of each protein was obtained from the annotation of its immunofluorescence (IF) images with the same antibodies.

2.2 Image preprocessing and feature extraction

Because each original HPA image is the fusion of DNA and protein, the linear spectral separation method was used to separate DNA and protein channels (Xu *et al.*, 2013). Then we extracted the Haralick texture features, DNA distribution features and local binary patterns (LBP) features from these two channels (Nanni *et al.*, 2010; Tahir *et al.*, 2012; Xu *et al.*, 2013). Each of 10 Daubechies filters can generate 836 Haralick features. They are used to create separate feature sets referred to as db1 through db10. The dimensions of DNA distribution and LBP features are 4 and 256, respectively. A feature vector of 1096 components is used to represent the image in each Daubechies filter space. Many previous studies have demonstrated that feature selection from the high-dimensional vector is useful, so we used stepwise discriminant analysis as it has been demonstrated to work well in this field (Newberg and Murphy, 2008; Xu *et al.*, 2013).

2.3 Incremental semi-supervised learning

We prepared three datasets, i.e. ADN, BDN and CDC, to construct classifiers. Among them, ADN and BDN are normal datasets with different levels of reliability of protein expression, and CDC is a cancer dataset. All of the ADN dataset were used in our experiments because this dataset has the best quality. Then the samples in BDN and CDC datasets were selectively added to the training set by semi-supervised learning. A flow chart of the proposed method is shown in Figure 2A.

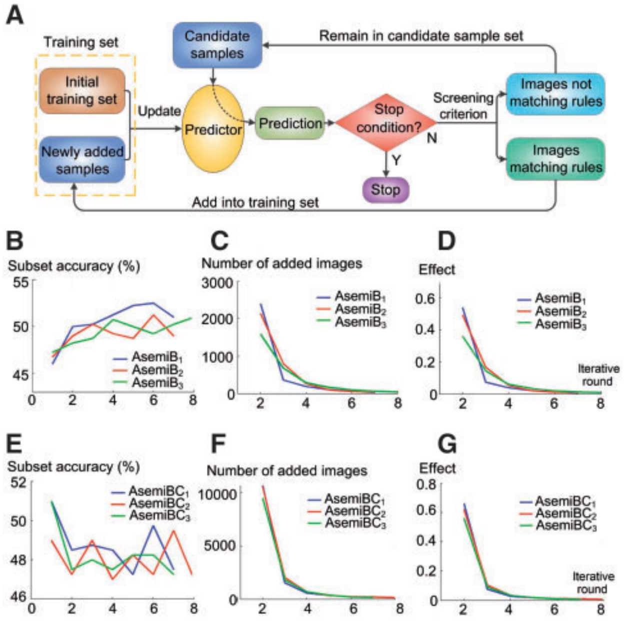


Fig. 2. Incremental process of iteratively adding candidate samples into the training set. (**A**) Flow chart of the iterative process. The initial training set is either the entire ADN dataset or the entire ADN dataset plus a selected subset of the BDN dataset, while the candidate samples to be included are images in either the BDN or CDC datasets, respectively. As the iterations proceed, the training set grows and the number of candidate samples decreases. (**B–D**) Results of iteratively adding BDN into the training set (initially ADN) using the proposed protocol with db7 features. (**E–G**) Results of iteratively adding CDC into the training set (initially ADN and a selected subset of BDN) using the proposed protocol with db7 features. (D, G) Shows effects caused by updating the training set in each round. The effect (eff_j) defined by Equation (1) is used for determining the stop condition of iterations. The model is considered stable when eff_j smaller than a threshold value (<0.01 in this study)

2.3.1 Incorporating new samples

The requirement for a sample to be added to the training set is that its predicted label set is the same as the annotation in HPA and the other classifier(s). This is because such images have more obvious discriminative features for a certain class and they can therefore help to enhance the classification boundary of the current model. Since there is no subcellular location annotation of proteins in cancerous tissues in the HPA, we compare the prediction output to the annotation of corresponding proteins in normal tissues to judge whether a sample in the CDC dataset should be selected or not. This is reliable when considering more than 95% proteins are actually not cancer biomarkers (Glory *et al.*, 2008). Note that when adding the samples from the CDC set, the initial classifier(s) are the resulting classifier(s) after adding the BDN set. This ensures the generality of final predictor for both normal and cancer proteins. To test different strategies, we have implemented three training modes, i.e. single-classifier mode, two-classifier mode and three-classifier mode. Details of their screening criteria to judge which samples need to be added are presented as follows.

The single-classifier mode just constructs one classifier, which will be iteratively updated until the stop condition is reached. Before the iteration process, an initial classier is trained using the entire ADN dataset. In each iteration round, the classifier is used to predict the subcellular locations of the images in the candidate sample set and those images whose predicted subcellular locations are the same as the annotations in HPA are selected and put into the training set. The classifier is then updated based on the new training set, which is ready for the next iteration.

According to the two-classifier mode, a predictor is composed of two classifiers, i.e. C_1 and C_2, where their initial models are trained on A_1 and A_2, which are generated from the ADN dataset via the bootstrap sampling method (Efron and Tibshirani, 1994). This sampling method randomly draws n independent samples with replacement from the original pooled set, where n is the number of samples in the pooled set. In this study, we sampled 4224 times with replacement from the ADN space and obtained approximately 63.2% of ADN images after discarding repeated images. This step can ensure $A_1 \subset \text{ADN}, \quad A_2 \subset \text{ADN}$ and $A_1 \neq A_2$, which guarantee the diversity of the initial models of C_1 and C_2. The candidate sample set was duplicated to two sets, B_1 and B_2, which were used for updating C_1 and C_2, respectively. In each iterative round of training, C_1 is firstly employed to predict the subcellular locations of the images in B_2, then those images whose predicted subcellular locations are exactly the same as the annotations in HPA were removed from B_2 and added to A_2 for updating the C_2 model. Analogously, A_1, the training set of C_1, was extended by predicting B_1 with C_2. As the iterations proceed, the size of B_1 and B_2 decreases while that of A_1 and A_2 increases.

The three-classifier mode trains three classifiers, i.e. C_1, C_2 and C_3, by three different training sets, i.e. A_1, A_2 and A_3, which are also initially constructed by using bootstrap sampling. Then the candidate sample set was duplicated to three sets, B_1, B_2 and B_3, for updating the three classifiers, respectively. In each round, C_1 and C_2 are used to predict the subcellular locations of the images in B_3. Those images whose label sets outputted from C_1 and C_2 are both the same as the HPA annotation were removed from B_3 and added to A_3 for updating C_3. A_1 and A_2 were updated in an analogous way based on the output from the other two classifiers.

2.3.2 Stopping condition

All the three modes are based on the iteration processes shown in Figure 2A. A critical question is when the iteration should terminate. The stopping condition of the iterations is determined by the effect of newly added samples to the classifier model. In this article, this effect is measured by the number of newly added samples and the change of the predicted scores for overlapping images in the current and previous rounds. To measure the change quantitatively, the t-test was used to compare the scores of two adjacent rounds and the average of the P-values was calculated. We thus define the effect as

$$\text{eff}_j = \frac{n_j}{N} \times \frac{1}{p_j}, \qquad (1)$$

where n_j is the number of samples newly added in the jth round, N is the total number of initial candidate samples, p_j is the average P-values between round j and $(j-1)$. The iteration stops when $\text{eff}_j < 0.01$, which is determined according to our experimental results. Details by varying eff_j are shown in Supplementary Figure S1.

2.4 Dynamic threshold criterion

Here, we used the support vector machine (SVM) as the classification model, and the LIBSVM-3.17 package is employed (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). The radial basis function was used as the kernel and its optimal width parameter was calculated by the data-driven calculator GFO (Lei *et al.*, 2012). To deal with multilabel proteins that can coexist in multiple subcellular locations, the binary relevance (BR) multilabel algorithm was used to deal with our datasets (Boutell *et al.*, 2004). According to BR, one binary SVM model was trained for predicting the relevance of test images to one class, so each BR classifier contains six SVM models (Xu *et al.*, 2013). A six-dimensional (6D) score vector $[s_1, s_2, \ldots, s_6]$ will be obtained per test image, where each score component represents the confidence of the input belonging to the corresponding class (six subcellular locations). Based on the outputted real-value confidence score vector, it is important to decide which class or classes should be assigned to a sample.

In a previous work, we investigated the top criterion (*T-criterion*) and the threshold criterion (*S-criterion*) to decide the label sets in multilabel classifications (Xu *et al.*, 2013). The *T-criterion* considers that the label set consists of the labels with positive scores, and if all the scores are negative, the label with the maximum score is considered as the unique label. The assumption of the *S-criterion* is that the score values corresponding to the real labels are the largest, and, in the case of a multiplex sample, its multiple labels will have similar scores. So in the *S-criterion*, a threshold is determined to measure whether a score is close enough to the largest one. However, it is a static threshold that is applied to all the images to be classified. A static unified threshold may not fit for all images because the scales of score vectors for different images can be variable, especially for the images in different classes.

To solve this problem, we proposed a dynamic threshold criterion (*D-criterion*) in this study, which can determine a specific threshold for each sample according to the scale and distribution of its score vector. For one image whose score vector is $[s_1, s_2, \ldots, s_6]$, the *D-criterion* can be presented as: if all the six scores are negative, then the label with the maximum score is considered as the unique label; if the maximum score is positive, then

$$y_i = \begin{cases} 1, \text{ if } \dfrac{s_{\max} - s_i}{s_{\max}} \le t \text{ or } s_i > \theta \\ -1, \text{ otherwise} \end{cases}, \qquad (2)$$
$$\text{where } s_{\max} = \max\{s_1, s_2, \ldots, s_6\}, s_{\max} > 0,$$

where y_i is the prediction of the sample's relevance to the ith class, and t and θ are two constant parameters that need to be determined. To derive these two parameters, the maximum a posteriori (MAP) principle is employed.

First, the degree of closeness between s_i and $s_{\max}$ is defined as

$$\text{tdif} = \frac{s_{\max} - s_i}{s_{\max}} \qquad (3)$$

Here, we define H_1 to denote 'yes' and H_2 to denote 'no' when deciding whether the ith class should be assigned to the predicted label set according to the tdif value. Supposing H_b is the final decision, the objective function with respect to tdif is

$$\begin{aligned} b = \underset{\varepsilon=1,2}{\arg\max}\, P(H_\varepsilon|\text{tdif}) = \underset{\varepsilon=1,2}{\arg\max} \frac{P(\text{tdif}|H_\varepsilon) \cdot P(H_\varepsilon)}{P(\text{tdif})}, \\ = \underset{\varepsilon=1,2}{\arg\max}\, P(\text{tdif}|H_\varepsilon) \cdot P(H_\varepsilon) \end{aligned} \qquad (4)$$

where $P(\text{tdif}|H_1) \cdot P(H_1)$ is the probability distribution of tdif for the positive samples in the ith class, and $P(\text{tdif}|H_2) \cdot P(H_2)$ is for negative samples not belonging to the ith class. So to distinguish H_1 and H_2 with a minimum error, the tdif value in the intersection of $P(\text{tdif}|H_1) \cdot P(H_1)$ and $P(\text{tdif}|H_2) \cdot P(H_2)$ is taken as the parameter t according to the MAP principle (Fig. 3).

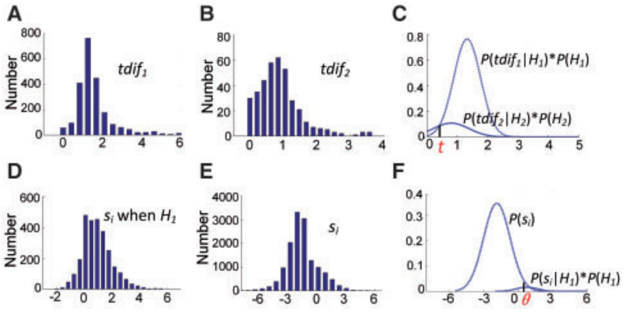


Fig. 3. Illustration of the process of determining parameters for *D-criterion*. Two constant parameters, t and θ, are needed in this criterion (Equation 2). Suppose the ith score of a sample outputted from classifier is s_i. When deciding whether the label i should be assigned to the predicted label set, we defined H_1 to denote yes and H_2 to denote no. t is set to distinguish H_1 and H_2, while θ is set to ensure that the labels with high scores are not missed. Both parameters are determined by maximizing posteriori principle, as well as score vectors of training set by 5-fold cross validation. (**A**) The histogram of tdif1. (**B**) The histogram of tdif2. tdif1 and tdif2 are tdif values corresponding to H_1 and H_2, respectively (Equation 3). (**C**) The fitting curves. The parameter t is obtained as the intersection point. (**D**) The histogram of s_i when H_1 happens. (**E**) The histogram of s_i. (**F**) The fitting curves. θ is set to ensure the ratio between the two regions of integration is 0.95. This figure is based on the model trained by ADN with db7 features

As for the other parameter θ, it is set to ensure all the high scores are not missed, and the confidence of the decision according to θ is

$$\alpha = P(H_1|s_i > \theta) = \frac{P(H_1)\cdot P(s_i > \theta|H_1)}{P(s_i > \theta)}$$
$$= \frac{\int_{\theta}^{+\infty} P(H_1)\cdot P(s_i|H_1)ds_i}{\int_{\theta}^{+\infty} P(s_i)ds_i} \quad (5)$$

Therefore, given a confidence score α, θ can then be calculated according to Equation (5). In this article, we set $\alpha = 0.95$, and its effects to θ and classification performance are shown in Supplementary Figure S2.

The statistics of $P(\mathrm{tdif}|H_1)$,$P(\mathrm{tdif}|H_2)$, $P(s_i|H_1)$, $P(H_1)$, $P(H_2)$ and $P(s_i)$ are based on the score vectors obtained by using 5-fold cross validation on the training set. The calculation process is given in Figure 3.

2.5 Evaluation metrics

Due to the fact that we are facing multilabel proteins, five multilabel classification metrics, i.e. subset accuracy, accuracy, recall, precision and average label accuracy were employed to evaluate the performance of the predictors (see Supplementary text for details). Among them, we mainly use the subset accuracy, which is the most stringent one since it requires the predicted label set to be exactly the same as the true label set. In addition, we also measured the sensitivity and AUC of each binary classifier in the models (see Supplementary text for detail).

3 Results

3.1 Baseline supervised model results

As a baseline, the most straightforward supervised method was used to train classifiers for comparison. We took the entire ADN, entire BDN and a combination of them (ADN + BDN), respectively, as training sets to construct classifiers. Then these classifiers were tested on the independent IDN dataset, and generated the results of simple supervised learning for comparison using the *T-criterion* (Fig. 4A) and the *D-criterion* (Fig. 4B), respectively.

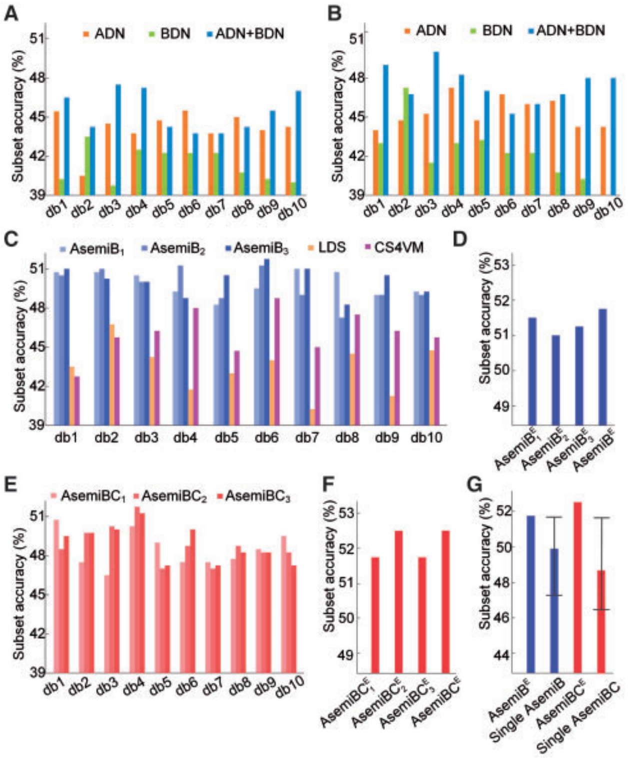


Fig. 4. Results of supervised learning and semi-supervised learning tested on the independent IDN dataset. (**A**) Results of baseline supervised classifiers trained on ADN, BDN and ADN + BDN datasets, respectively, using the *T-criterion*. (**B**) Results of supervised classifiers using the *D-criterion*. (**C**) Comparison results of our classifiers trained by adding BDN to ADN using semi-supervised strategy on three modes, with two other semi-supervised classifiers in literature. (**D**) Ensemble by fusing the classifiers after adding BDN. (**E**) Results of classifiers trained by subsequently adding CDC to the training set using the semi-supervised strategy on three modes. AsemiB$_1$ and AsemiBC$_1$ mean using one-classifier mode, AsemiB$_2$ and AsemiBC$_2$ mean using two-classifier co-training mode, and AsemiB$_3$ and AsemiBC$_3$ mean using three-classifier tri-training mode. (**F**) Ensemble by fusing the classifiers after adding CDC. AsemiB$_1^E$, AsemiB$_2^E$, AsemiB$_3^E$, AsemiBC$_1^E$, AsemiBC$_2^E$ and AsemiBC$_3^E$ are ensemble classifiers, and each of them is constructed by fusing 10 single classifiers of db1–db10. AsemiB^E is the ensemble of AsemiB$_1^E$, AsemiB$_2^E$ and AsemiB$_3^E$. AsemiBC^E is the ensemble of AsemiBC$_1^E$, AsemiBC$_2^E$ and AsemiBC$_3^E$. (**G**) Comparison of subset accuracies between ensemble classifiers and single classifiers

It can be seen from Figure 4A and B that: (1) *D-criterion* outperforms *T-criterion*, demonstrating the effectiveness of the *D-criterion*; (2) Overall, the subset accuracies of classifiers trained on ADN are better than those on BDN, indicating that the image quality can affect the model performance; (3) Interestingly, in some cases, the results of ADN + BDN are not better than those only using ADN, indicating that not all of the medium quality images in BDN have a positive effect on performance.

The first observation suggests that a dynamic threshold is better due to the specificity for testing samples, thus we will use the *D-criterion* in the following experiments. The second and third observations suggest that if we add all the BDN samples into ADN to train a supervised model, the performance does not improve sometimes. The reason could be that not all of the samples in the

BDN are complementary to the ADN; furthermore, some low-quality samples in the BDN will degenerate the model. This motivated us to explore a better way to take advantage of the candidate image samples rather than simply employing all of them.

3.2 Improvements by selectively adding medium-reliability data

The entire ADN was used as the initial training set, and then according to the semi-supervised iteration framework, not all of the BDN images, but only those which improve model performance were iteratively selected into the training set. The final results are three semi-supervised predictors, which are denoted as AsemiB_1 (one-classifier mode), AsemiB_2 (two-classifier mode) and AsemiB_3 (three-classifier mode), corresponding to the three training modes, respectively.

The classifier of each round is tested on IDN, and the changes of subset accuracies are shown in Figure 2B. The changes of number of added images, and effects on each iterative round are illustrated in Figure 2C and D. It can be seen that as the round increases, the subset accuracy tends to increase in all modes. All the final subset accuracies when these iterations terminate, i.e. 51, 49 and 51%, are higher than the result of directly adding the entire BDN, which is 46% as shown in Figure 4B. Besides, both the number of added images and effect value in the iteration decrease sharply. This indicates that the influence of the added images on classification decreases as the round increases. At the end of iterations of the db7 model, 56.75, 61.37 and 52.86% images in BDN were chosen and added to the training sets of AsemiB_1, AsemiB_2 and AsemiB_3, respectively. Compare Figure 4C and B, we can see that all the subset accuracies of three semi-supervised modes are higher than those of supervised learning. Adding medium-reliability data into training set not only expands the training sample space, but also validates the effectiveness of the proposed semi-supervised idea.

Considering that different semi-supervised learning methods have been widely used these years (Lee and Madabhushi, 2010; Luo *et al.*, 2013), we also compared our methods with two state-of-the-art semi-supervised algorithms, i.e. low-density separation (LDS) and cost-sensitive semi-supervised SVM (CS4VM). LDS is a graph-based method, which represents each labeled and unlabeled sample as a node and tries to place decision boundaries in regions where there are few data nodes (Chapelle and Zien, 2005). CS4VM incorporates the unlabeled data into the SVM by estimating their label means of misclassification costs (Li *et al.*, 2010). Figure 4C shows the results of LDS and CS4VM when taking ADN as labeled data, BDN as unlabeled data and IDN as testing set. The performances of our proposed methods are better than LDS and CS4VM on the multilabel dataset of this article. One reason can be the multilabel sample classification is much more comprehensive than the single-label case used by the two algorithms. For instance, the LDS might be unable to accurately find the boundaries in a graph built by multilabel data, because some multilabel samples are near the low-density areas and confuse the decisions.

3.3 Incorporating images from cancer tissues to the model

To enhance the performance of predicting subcellular locations of proteins in cancerous tissues, we consider adding some images from cancerous tissues into the training set to eliminate the transfer prediction error caused by the difference between the normal and cancer data. Actually, we conducted an experiment to quantify the differences of patterns between the two states. Based on the proteins in CDC set, we used the correlation coefficient (CC) to measure difference between normal and cancer images, where we assumed that proteins in the CDC set did not change their locations in cancer states. This is reasonable when considering that more than 95% protein images in current HPA database are actually not cancer biomarkers (Glory *et al.*, 2008). Each image was represented by its feature vector, and three CC matrixes were calculated: the first is the intra-CC in the normal images group, the second is the intra-CC in the cancer images group and the third is the inter-CC between normal and cancer sets. Figure 5 shows the averaged CC values based on six subcellular locations. It can be seen that the inter-CC values between normal and cancer images are lower than the intra-CC values in all cases. In addition, we also calculated the *P*-values with the student *t* test between normal and cancer dataset, and *P*-values of all the subcellular locations are <0.05. These results demonstrate that even for the same organelle, there is a difference between the normal and cancer data. This suggests that the transfer method of using normal data as the training set to predict the cancer data may miss some specific features of proteins in the cancer state.

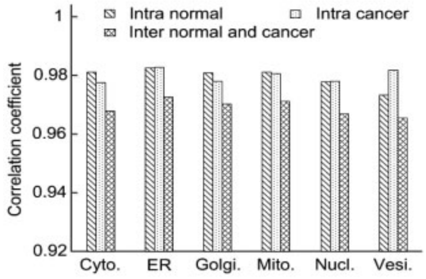


Fig. 5. Comparison of intranormal CC, intracancer CC and inter normal and cancer CC values. In the statistics, the high expression level dataset and CDC dataset are used as normal and cancer dataset, respectively. The db7 features are used and the feature dimension is 80

After adding BDN in above section, we obtained three classifiers, i.e. AsemiB_1, AsemiB_2 and AsemiB_3, by semi-supervised learning. Following the incremental selective learning protocol, images from CDC were subsequently added to these classifiers, and we got AsemiBC_1, AsemiBC_2 and AsemiBC_3 (Figs. 2E–G and 4E). It can be seen that the subset accuracies of classifiers on the independent IDN set fluctuate and decline slightly, which is because the added cancer data affected the prediction performance of normal data. This also highlights the difference between normal and cancer data. Nevertheless, the decline in performance is not significant, and the subset accuracies still outperform the baseline results from supervised models.

3.4 Performance of ensemble classifiers

Since an ensemble of multiple classifiers generally achieves better performance, we constructed ensemble classifiers by combining the 10 classifiers with db1–db10 features. The fusion method averages all the score vectors from the 10 single classifiers to get a final six-dimensional (6D) vector for each query image. These ensemble classifiers are tested on IDN to show their effectiveness on the normal dataset (Fig. 4D and F). By comparing the results between the ensemble classifier and the single classifiers, we find that the ensemble classifier outperforms the single classifier on IDN dataset. For example, a 2% improvement of the subset accuracy was observed for the $\text{AsemiBC}_2^{\text{E}}$ compared with the single classifier AsemiBC_2 on db7. The other merit of the ensemble strategy is that it can significantly reduce the negative bias by adding the cancer data to the training set. For instance, the subset accuracy of single AsemiBC$_1$ classifier on IDN with db7 feature is 47.5% (Fig. 4D), which is 4.25% lower than the AsemiBC$_1^E$.

One final ensemble predictor without-adding CDC and one final ensemble predictor adding CDC were created. All the classifiers without-adding CDC were fused to create AsemiB^E, and all the classifiers of adding CDC were fused to create AsemiBC^E. Both of them could achieve good performance on IDN testing set (Fig. 4G). It is worth pointing out that besides the most stringent metric in multilabel classification, subset accuracy (Fig. 4), we also used other indices to evaluate the AsemiB^E and AsemiBC^E and their results can be seen in Supplementary Tables S3 and S4. For example, the average label accuracy, which indicates the reliability of prediction for single locations, can achieve 87.04% for the final system (Supplementary Table S3), which implies the reliable detection of translocation from or to a specific location.

3.5 Detecting protein translocations of cancer biomarkers

The IBD set containing 10 reported biomarker proteins was used for validating whether the sensitivity of translocation detection can be enhanced by utilizing cancer data in the training phase. We compared the prediction results on the IBD set before and after adding the CDC dataset to see the effects of adding CDC data. The results from AsemiB^E and AsemiBC^E were compared, where the former did not incorporate the cancer data into training, whereas the latter did. To quantify the sensitivity of detecting the subcellular location changes, in addition to the predicted and reported location labels in the normal and cancer conditions, we also conducted independent sample *t* tests on the predicted score vectors to evaluate the significance of the location changes (Supplementary Fig. S3). The comparison results and *P*-values of the changes are shown in Table 1, from where we can see that:

1. The protein Bax and cyclin D1 prove that adding CDC dataset makes the classifiers more sensitive to detect the location changes occurring during cancer. In detail, protein Bax will partly translocate from the cytoplasm to the mitochondrion when lymphoma occurs (Nechushtan *et al.*, 1999). This translocation cannot be found by the predictors trained only on normal data, but can be picked out by AsemiBC^E, which was trained on both normal and cancer data. The protein cyclin D1 normally shuttles between cytoplasm and nucleus locations. However, in ovarian cancer cyclin D1 is found only in the nucleus (Gladden and Diehl, 2005). AsemiB^E predicts cyclin D1 its locations in cancer as both the nucleus and mitochondria, while AsemiBC^E correctly predicts its cancer location as the nucleus only.
2. The loss of nuclear localization of PTEN in pancreatic cancer is correctly predicted by both AsemiB^E and AsemiBC^E (Perren *et al.*, 2000), demonstrating that the machine-learning systems are effective for the detection of protein mislocalization.
3. AsemiBC^E is able to perform prediction better than AsemiB^E for the IBD proteins in their normal states. For example, the protein BAG-1 is reported to reside in the nucleus in normal conditions and translocate to the mitochondria during colorectal cancer (Takayama *et al.*, 1998). AsemiB^E predicted BAG-1 would localize in both the cytoplasm and nucleus in the normal state, whereas AsemiBC^E predicted only a nucleus location, which is experimentally correct. Other examples include NQO1 and GOLGA5.
4. The *P*-values also reveal the improved sensitivity for detecting protein translocations by the predictor of AsemiBC^E. The lower

Table 1. Comparison between literature descriptions and the results of predicting IBD by ensemble classifiers

Protein	Tissue	Protein translocations from normal to cancer condition		
		Reported by literature (normal → cancer)	Prediction by AsemiB^E (normal → cancer *P*-values of changed locations)[a,b]	Prediction by AsemiBC^E (normal → cancer *P*-values of changed locations)[a,b]
Bax	Lymph node	Cyto. → Cyto.& Mito.	Cyto. → Cyto. Mito.0.6336	Cyto. → Cyto.& Mito. **Mito.0.4402**
cyclin D1	Ovary	Cyto.& Nucl. → Nucl.	Cyto. → Nucl.& Mito. Cyto.0.0430	Cyto. → Nucl. **Cyto.0.0319**
PTEN	Pancreas	Cyto.& Nucl. → Cyto.	Cyto.& Nucl. → Cyto. **Nucl.0.3853**	Cyto.& Nucl. → Cyto. Nucl.0.5570
BAG-1	Colon	Nucl. → Mito.	Nucl.& Cyto. → Nucl.& Cyto. **Nucl.0.5001**, Mito.0.6513	Nucl. → Nucl.& Cyto. Nucl.0.5944, **Mito.0.3463**
GOLGA5	Thyroid gland	Gol. → Mito.	Gol.& Mito.& Nucl. → Gol. Gol.0.8560, **Nucl.0.0403**	Gol. → Cyto. **Gol.0.2699**, Nucl.0.5522
NQO1	Lung	Cyto. → Nucl.	Nucl. → Cyto. Cyto.0.0010, Nucl.0.0798	Cyto. → Cyto. **Cyto.0.0003, Nucl.0.0441**
SOX9	Breast	Nucl. → Cyto.	Nucl. → Nucl. Cyto.0.2628, Nucl.0.5170	Nucl. → Nucl. **Cyto.0.0741, Nucl.0.1143**
p53	Breast	Nucl. → Nucl.& Cyto.	Nucl. → Nucl. Cyto.0.1315	Nucl. → Nucl. **Cyto.0.0741**
TOP2A	Lung	Nucl. → Cyto.	Nucl. → Nucl. Cyto.0.2130, Nucl.0.7945	Nucl. → Nucl. **Cyto.0.1286, Nucl.0.5853**
IGFBP	Breast	Nucl. → Cyto.	Cyto. → Cyto. **Cyto.0.4517**, Nucl.0.7419	Cyto. → Cyto. Cyto.0.6124, **Nucl.0.6167**

[a] The results have two lines: the first line is the predicted subcellular location labels in normal and cancer conditions, respectively, by the classifier; the second line is the *P*-values measuring the subcellular location changes when cancer occurs (column 3), which are calculated by the independent sample *t* test on the predicted scores for normal and cancer images.

[b] Those translocations that have lower *P*-values are bold.

the *P*-value, the more significant the change. There are a total of 16 experimentally known changed locations for the 10 proteins. Twelve of them have lower *P*-values in AsemiBC^E with a *P*-value 0.0003–0.6167 compared with 0.001–0.8560 in AsemiB^E. These results suggest that the sensitivity of detecting protein subcellular location changes is enhanced by incorporating the cancer data into the model construction.

5. Although some improvements can be observed (with lower *P*-values) by incorporating the cancer images into the classification system construction, there are still considerable room for improvement. For instance, there are still some cases where none of the two predictors can get completely correct prediction. This suggests that tremendous future efforts are needed for further improvement.

4 Discussion and conclusions

In this article, we present a new automated bioimage analysis system for sensitively detecting translocated or mislocated proteins in human cancers. The new system is featured with a semi-supervised learning engine, which can help to enlarge the training space by incorporating lower-quality or unlabeled data key to the performance of a statistic model. The other merit of the new system is the capability of predicting proteins that shuttle among multiple subcellular locations, and a new dynamic *D-criterion* is proposed to deal with the multilabel set determination problem by considering the specificity of each protein. The new developed system has opened a new avenue for bioimage-based automated biomarker detection work, which suits large-scale data analysis and complement research from biological experiments.

We have shown that the strategy of selectively incorporating medium staining normal images with the developed semi-supervised framework is helpful for improving the classification accuracy on the normal images as demonstrated in the independent test dataset. On the other hand, some improvements were also observed when applying the semi-supervised algorithm for adding selected cancer images into training, but they have still considerable space for further improvement. For instance, some translocated or mislocated cancer biomarkers cannot be completely predicted, especially for those multi-label proteins.

To further improve the performance of our system, some efforts will be made in future studies. First, we will aim to improve the multilabel classification algorithm by taking the label correlations into account. Multiplex proteins that may shuttle among more than one subcellular location indicate a complex subcellular protein organization in the cell. The benchmark dataset of this study contains 26% multilabel proteins. This ratio is even much higher to reach approximately 60% according to a recent study of applying IF and fluorescent-protein tagging techniques on mammalian cells (Stadler, 2013). In this article, we transformed the multilabel problem into six binary classification problems, ignoring the correlation among different subcellular locations. It is expected that incorporating correlations, such as proteins coexisting at different locations due to spatial proximity or functional reasons, will be useful for further improving the performance.

Second, our imaging-based studies can be integrated with analysis of non-imaging data, such as proteomics and genomics analyses (Murphy, 2014). Amino acid sequence has been used for predicting protein subcellular locations for many years, and we have developed an efficient sequence-based subcellular location predictor called Cell-PLoc in previous studies (Chou and Shen, 2008; Shen and Chou, 2009). The Cell-PLoc can also deal with multilabel proteins and have wide coverage of subcellular components. Merging prediction results from different resources is a potential effective way for further enhancing the sensitivity for translocated proteins detection. The multiclassifier mode of this study also provides a feasible combination solution, which enables us to cotrain our image-based and sequence-based software to generate a better protein subcellular location prediction system.

Acknowledgements

We are grateful to Dr. Jeffrey Brender and Dr. Richard Jang for reading the manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China [Nos. 61222306, 91130033, 61175024], Shanghai Science and Technology Commission [No. 11JC1404800], a Foundation for the Author of National Excellent Doctoral Dissertation of People's Republic of China [No. 201048] and the National Institute of General Medical Sciences [GM083107].

Conflict of interest: none declared.

References

Benzeno,S. *et al.* (2006) Identification of mutations that disrupt phosphorylation-dependent nuclear export of cyclin D1. *Oncogene*, **25**, 6291–6303.

Boutell,M.R. *et al.* (2004) Learning multi-label scene classification. *Pattern Recogn.*, **37**, 1757–1771.

Chapelle,O. and Zien,A. (2005) Semi-supervised classification by low density separation. *Proc. AISTATS*, pp. 57–64.

Chou,K.C. and Shen,H.B. (2008) Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.*, **3**, 153–162.

Cohen,W.W. (2002) Improving a page classifier with anchor extraction and link analysis. *Advances in Neural Information Processing Systems*, **15**, 1481–1488.

Efron,B. and Tibshirani,R.J. (1994) *An Introduction to the Bootstrap*. CRC Press, Boca Raton, FL.

Eliceiri,K.W. *et al.* (2012) Biological imaging software tools. *Nat. Methods*, **9**, 697–710.

Gladden,A.B. and Diehl,J.A. (2005) Location, location, location: the role of cyclin D1 nuclear localization in cancer. *J. Cell Biochem.*, **96**, 906–913.

Glory,E. and Murphy,R.F. (2007) Automated subcellular location determination and high-throughput microscopy. *Dev. Cell*, **12**, 7–16.

Glory,E. *et al.* (2008) Automated comparison of protein subcellular location patterns between images of normal and cancerous tissues. In: *IEEE International Symposium on Biomedical Imaging 2008*, pp. 304–307.

Hady,M.F.A. and Schwenker,F. (2013) Semi-supervised learning. In: Bianchini,M. *et al.* (eds) *Handbook on Neural Information Processing*, Springer, Berlin, pp. 215–239.

Hanash,S.M. *et al.* (2008) Mining the plasma proteome for cancer biomarkers. *Nature*, **452**, 571–579.

Hung,M.C. and Link,W. (2011) Protein localization in disease and therapy. *J. Cell Sci.*, **124**, 3381–3392.

Lee,G. and Madabhushi,A. (2010) Semi-supervised graph embedding scheme with active learning (SSGEAL): classifying high dimensional biomedical data. In: Dijkstra,T.M.H. *et al.* (eds) *Pattern Recognition in Bioinformatics*, Springer, Berlin, pp. 207–218.

Lei,J.B. *et al.* (2012) GFO: a data driven approach for optimizing Gaussian function based similarity metric in computational biology. *Neurocomputing*, **99**, 307–315.

Li,Y.F. *et al.* (2010) Cost-sensitive semi-supervised support vector machine. In: *AAAI*, pp. 500–505.

Liston,D.B. and Stone,L.S. (2008) Effects of prior information and reward on oculomotor and perceptual choices. *J. Neurosci.*, **28**, 13866–13875.

Luo,Y. *et al.* (2013) Manifold regularized multitask learning for semi-supervised multilabel image classification. *IEEE Trans. Image Process.*, **22**, 523–536.

McLachlan,G.J. (1975) Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *J. Am. Stat. Assoc.*, **70**, 365–369.

Murphy,R.F. (2004) Automated interpretation of protein subcellular location patterns: implications for early cancer detection and assessment. *Ann. N. Y. Acad. Sci.*, **1020**, 124–131.

Murphy,R.F. (2014) A new era in bioimage informatics. *Bioinformatics*, **30**, 1353

Nanni,L. *et al.* (2010) Novel features for automated cell phenotype image classification. In: Arabnia,H.R. (ed) *Advances in Computational Biology*, Springer, Berlin, pp. 207–213.

Nechushtan,A. *et al.* (1999) Conformation of the Bax C-terminus regulates subcellular location and cell death. *EMBO J.*, **18**, 2330–2341.

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**, 2300–2308.

Perren,A. *et al.* (2000) Mutation and expression analyses reveal differential subcellular compartmentalization of PTEN in endocrine pancreatic tumors compared to normal islet cells. *Am. J. Pathol.*, **157**, 1097–1103.

Pierleoni,A. *et al.* (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, **22**, e408–e416.

Rizzardi,A.E. *et al.* (2012) Quantitative comparison of immunohistochemical staining measured by digital image analysis versus pathologist visual scoring. *Diagn. Pathol.*, **7**, 42.

Shen,H.B. and Chou,K.C. (2009) A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *Anal. Biochem.*, **394**, 269–274.

Stadler,C. *et al.* (2013) Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells. *Nat. Methods*, **10**, 315–323.

Tahir,M. *et al.* (2012) Protein subcellular localization of fluorescence imagery using spatial and transform domain features. *Bioinformatics*, **28**, 91–97.

Takayama,S. *et al.* (1998) Expression and location of Hsp70/Hsc-binding anti-apoptotic protein BAG-1 and its variants in normal tissues and tumor cell lines. *Cancer Res.*, **58**, 3116–3131.

Uhlen,M. *et al.* (2010) Towards a knowledge-based human protein atlas. *Nat. Biotechnol.*, **28**, 1248–1250.

Winski,S.L. *et al.* (2002) Subcellular localization of NAD (P) H: quinone oxidoreductase 1 in human cancer cells. *Cancer Res.*, **62**, 1420–1424.

Xu,Y.Y. *et al.* (2013) An image-based multi-label human protein subcellular localization predictor (iLocator) reveals protein mislocalizations in cancer tissues. *Bioinformatics*, **29**: 2032–2040.

Zhou,Z.H. and Li,M. (2005) Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Trans. Knowl. Data Eng.*, **17**, 1529–1541.