Unsupervised Unmixing of Subcellular Location Patterns

Luís Pedro Coelho LPC@CMU.EDU
Robert F. Murphy MURPHY@CMU.EDU
Lane Center for Computational Biology, Carnegie Mellon University and Joint Carnegie Mellon University-University of Pittsburgh Ph.D. Program in Computational Biology, 4400 Fifth Ave, Pittsburgh, PA 15217 USA

Abstract

With the advent of high-throughput microscopes, researchers can routinely image hundreds of different proteins per day, generating thousands of images. To be able to organize these images and extract meaningful information, we need automatic methods. The state-of-the-art in automated subcellular localization is classification in the space of image features. This approach is not suited, however, for handling mixture patterns (the pattern of a protein present in more than one location).

We have previously described methods for determining the fraction of fluorescence in various subcellular locations when the basic locations in which a protein can be present are given a priori. However, knowing all fundamental patterns a priori may be problematic. The alternative is unsupervised unmixing: given a set of images from different proteins, identify the basic patterns that best explain all the observed images as either examples of such basic patterns or combinations thereof.

We extend our previous work to handle this problem. Using a validation dataset, we show that this method can recover the underlying mixed patterns. It identifies meaningful basis patterns and mixture coefficients that correlate well with the probe concentrations that generated the dataset (the probe concentrations were kept hidden from the algorithm).

1. Introduction

The current best method for determining protein location is by fluorescence imaging of tagged cells. For proteome-wide studies of subcellular localization the amounts of data are so large that automated methods are required.

The traditional approach for automated protein subcellular localization from images is a discriminative feature-based approach. In this approach, a set of features is computed from each image and computation proceeds in this space of features, where machine learning methods can be applied. This approach has shown good results when used to identify images based on training data or for clustering unlabeled images (Boland & Murphy, 2001; Chen & Murphy, 2005; Hamilton et al., 2007).

The discriminative approach is not suited, however, for handling mixture patterns. A mixture of patterns occurs when a protein (or other marker) is present in more than one location. For example, one expects that some proteins are present in endosomes, others in lysosomes, while others will be present in both of these. With the discriminative approach, this will often result in three independent classes, a situation where the relationship between the pattern classes is neither represented nor discoverable from data. In general, there is no reason to expect that feature values of a mixture pattern will have any meaningful relationship with the feature values of the base patterns that compose it.

We have previously described methods for determining the fraction of fluorescence in various subcellular patterns (Zhao et al., 2005; Peng et al., 2009). However, this "pattern unmixing" approach requires that the user specify the fundamental patterns of which all patterns can be comprised.

Deciding in advance which are the fundamental patterns that occur in a large collection of images may be challenging and require a lot of guessing by the researcher. The alternative is unsupervised unmixing: given a set of images with different proteins tagged, identify the basic patterns that best explain all the observed images as either examples of such basic patterns or combinations thereof.

The main contribution of this paper is a method to solve the unsupervised subcellular pattern unmixing problem. This method is validated in a test dataset where two probes were mixed in known concentrations. These concentrations were, however, kept hidden from the algorithm and are used for validation only.

2. Methods

The methods presented here are an extension of the methods presented previously for supervised pattern unmixing (Zhao et al., 2005).

The algorithm starts by processing each image to extract salient objects. Objects are described by an 11 dimensional feature vector (using the same features as in (Zhao et al., 2005)). These features are used to cluster objects into object types. The intuition is that different patterns consist of different object types, while mixed patterns show objects from their constituent classes in proportion to the mixture coefficients.

Object Detection

Objects are defined as contiguous regions of non-background fluorescence. The previous work on pattern unmixing used a global threshold to find objects (Zhao et al., 2005). For the images used in the current study, we found that global thresholding methods correctly differentiate between the cell region and background, but do not differentiate between bright objects inside the cell and general cellular fluorescence. Local thresholds, on the other hand, work as intended inside the cell region, but pick up noise in background regions. Therefore, we employ a hybrid method and consider a pixel to be inside an object only if it is above both a global and a locally determined threshold. Once images have been binarized, objects are defined as above threshold contiguous regions. In this work, we used Ridler-Calvard (Ridler & Calvard, 1978) for global thresholding and the mean value of the 15×15 pixel window centred on the pixel of interest as the local threshold.

Clustering and Unmixing

Each object is characterized by its feature vector (normalized to z-scores). All the objects in the collection are then clustered using k-means clustering. For each value of k, 10 random restarts are performed. The final number of clusters is determined by the Bayesian information criterion.

After clustering, each object is assigned a cluster index. An image can then be summarised by simply counting the number of its constituent objects that belong to each class, i.e., for each image we obtain a vector of counts $\boldsymbol{x} \in \mathbb{R}^d$, where d is the number of object clusters.

In order to compensate for differing cell counts in each image, we normalise the vector $\boldsymbol{x}$ to be a vector of object fractions. Furthermore, some preliminary results show that the fit was dominated by frequent object types, which are present in great numbers in many images. Being so common, these objects were not discriminative. This led us to remove object types that appear in over 90% of images.

Our generative model for images is that a set of basis vectors $\mathcal{B} = \{\boldsymbol{b}_j\}_j$ combines with a vector $\boldsymbol{\alpha}$ of counts to form an image by

$$\boldsymbol{x} = \sum_j \alpha_j \boldsymbol{b}_j \tag{1}$$

In the unsupervised unmixing problem, we wish to invert the generative process. Given a set of images represented as vectors of counts $\mathcal{D} = \{\boldsymbol{x}_i\}_i$, we need to find a set of basis vectors $\mathcal{B} = \{\boldsymbol{b}_j\}_j$ and mixing coefficients α_{ij}, such that

$$\boldsymbol{x}_i = \sum_j \alpha_{ij} \boldsymbol{b}_j + \varepsilon_i, \tag{2}$$

where we wish to minimize $\sum_i \varepsilon_i^2$, subject to the constraint $\alpha_{ij} \geq 0$.

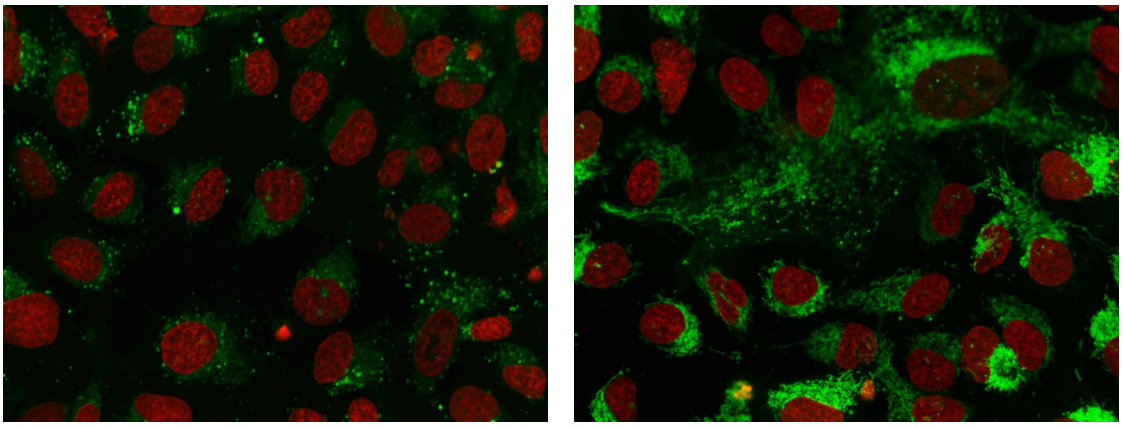
We add additional restrictions that guide the system towards a more meaningful answer. In particular, we restrict possible basis vectors to elements of the collection (i.e., $\boldsymbol{b}_j \in \mathcal{D}$) . This encodes the idea that our collection is broad enough to contain examples of the base classes and aids the interpretability of the final result. In the *Discussion* section, we reflect upon the implications of this choice.

Additionally, we consider that each mixture is a mixture of a small number of bases and therefore, bias the search towards sparser solutions. As the simplest implementation of this idea, we constrain the mixture vector $\boldsymbol{\alpha}_i$ corresponding to image i to have a small number of non-zero components.

3. Results

Dataset and Criteria

In order to validate this model, we used a test dataset of images created for the first validation of subcellular pattern unmixing in real images for cells (Peng et al.,



(a) Lyso-dominant (b) Mitotracker

Figure 1. Examples of selected bases. These are the first images corresponding to the selected samples. Images are false color panels, with red showing the nuclear channel and green the tracker channel. Images have been manually contrast stretched for publication.

2009). In this collection, cells were labeled with 8 different concentrations of lysotracker and mitotracker (for a total of 64 possible combinations). The emission spectra of both probes are similar, and thus the amount of each added cannot be distinguished using filters. This mimics the situation in which a protein is present in varying amounts in lysosomes and mitochondria.

Two criteria were used for evaluating the quality of the solution obtained:

1. The discovered basis sets should correspond to the basis patterns (one basis vector corresponding to the lysosomal pattern and the other to the mitochondrial pattern).

2. The inferred mixture coefficients α_{ij} for a given sample should correlate with the actual probe concentrations used to label that sample. We will compare our results to those obtained by supervised unmixing.

Unmixing Results

The algorithm returned two of the test samples as the bases for unmixing. One of the basis was a pure mitotracker sample, while in the other basis, lysotracker is dominant. Figure 1 shows the first image corresponding to each basis as examples.

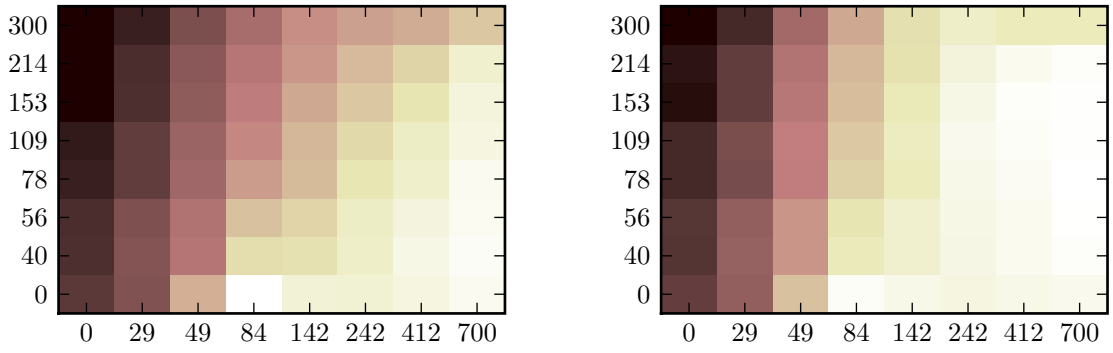
Given that $\{\boldsymbol{x}_i\}_i$ vectors were normalised to one, we compare the coefficients obtained from the algorithms with the underlying concentration fractions. For some images, the estimated fractions were zero. Discarding those images leads to correlations with the underlying fractions of 77% and 67%, respectively. This compares with 84% and 71% in the supervised case. The unsupervised algorithm does not do as well as the supervised version, but the difference is not very large. The full results are displayed as heat maps in Figure 2, where we can see that the unsupervised results are qualitatively very similar to the supervised results.

4. Discussion

We have presented an extension of the supervised unmixing problem to handle the unsupervised case. The model is based on automatic object detection and clustering. Images are then summarised by a simple vector in a low dimensional space and unmixing proceeds there. This extends our previously presented methods for supervised unmixing.

A validation dataset where conditions were known showed that the method can recover most of the structure of mixture patterns. One of the returned bases was a pure mitotracker pattern, while the other was a lysotracker dominant sample (Figure 1). The inferred mixture coefficients are highly correlated with the underlying concentrations that were used to generate the dataset.

The additive model we presented is a variation on traditional types of unsupervised dimensionality reduction. However, the restrictions we added, particularly the restriction that basis vectors be present in the source dataset, significantly change the nature of the problem and the interpretability of the solution. They rule out approaches based on principal component analysis or non-negative matrix approximation (Lee & Seung, 2000; Berry et al., 2007). We



(a) Supervised Unmixing (b) Unsupervised Unmixing

Figure 2. Estimation results of mixture fractions for supervised and unsupervised unmixing. In both the case of supervised and of unsupervised unmixing, we plot the estimate of the fraction of lysosomal pattern as a function of the hidden concentration of mitotracker and lysotracker. In both cases, dark brown corresponds to 0.0 and white corresponds to 1.0. Mitotracker concentration varies along the horizontal axis, while lysotracker varies along the vertical axis.

believe that the solutions obtained by our method are more appropriate when the goal is to organise a large collection of images in a meaningful way. That we can point to example images such as those presented in Figure 1 allows one to easily communicate the results of the algorithm. Other dimensionality reduction procedures often require a posteriori interpretation of the inferred basis, which might be appropriate in some domains (e.g., interpreting a set of words in a textual problem), but would be cumbersome for large sets of images.

The validation dataset used here was obtained using an automated microscope and used without any hand filtering of the data or special processing. This enables us to use this method in the large-data setting where human inspection of the images is impossible.

We are currently working on a graphical user interface to these methods. This will make our software implementation usable by a wider audience.

Acknowledgements

This work was funded by NIH grant GM075205. LPC was partially funded by the Fundação Para a Ciência e Tecnologia (grant SFRH/BD/37535/2007) as well as a fellowship from the Fulbright Program.

The authors thank Tao Peng, Estelle Glory, Ghislain Bonami, Sumit Chanda, and Daniel Rines for providing images as well as many helpful discusions.

References

Berry, M. W., Browne, M., Langville, A. N., Pauca, P. V., & Plemmons, R. J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, *52*, 155–173.

Boland, M. V., & 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.

Chen, X., & Murphy, R. F. (2005). Objective clustering of proteins based on subcellular location patterns. *Journal of Biomedicine and Biotechnology*, *2005*, 87–95. doi:10.1155/JBB.2005.87.

Hamilton, N., Pantelic, R., Hanson, K., & Teasdale, R. (2007). Fast automated cell phenotype image classification. *BMC Bioinformatics*, *8*, 110.

Lee, D. D., & Seung, S. H. (2000). Algorithms for nonnegative matrix factorization. *NIPS* (pp. 556–562).

Peng, T., Bonamy, G. M., Glory, E., Daniel Rines, S. K. C., & Murphy, R. F. (2009). Automated unmixing of subcellular patterns: Determining the distribution of probes between different subcellular locations. *Proceedings of the National Academy of Sciences (Submitted)*.

Ridler, T., & Calvard, S. (1978). Picture thresholding using an iterative selection method. *Systems, Man and Cybernetics, IEEE Transactions on*, *8*, 630–632.

Zhao, T., Velliste, M., Boland, M. V., & Murphy, R. F. (2005). Object type recognition for automated analysis of protein subcellular location. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, *14*, 1351–1359.