IMPROVED COMPARISON OF PROTEIN SUBCELLULAR LOCATION PATTERNS

Ting Zhao[1], *Stalia Soto*[*] *and Robert F. Murphy*[1,2,3]

Departments of Biomedical Engineering[1] and Biological Sciences[2] and Center for Automated Learning and Discovery[3], Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.A.

ABSTRACT

A common task in cell and molecular biology is to evaluate the difference or similarity among location patterns under different circumstances. Our previous work has described an automated method to objectively compare any pair of subcellular location patterns using well-characterized numerical features. This paper describes an improved version of the previous comparison method by applying nonparametric testing and multiple testing methods. We show that the new approach has better performance for detecting differences, especially on small amounts of data. The new approach was also used to compare location patterns in a dataset containing 3D images. To demonstrate the application of the method, we collected images of drug-treated cells and then use the method to evaluate the effects of drugs on location patterns.

1. INTRODUCTION

Location proteomics, the systematic study of cellular protein localization, is critical to the thorough understanding of how cells work. Although there are many methods to determine protein locations, systematic description of location patterns had not been undertaken until automated, quantitative methods for interpreting fluorescence microscope images of cellular proteins were developed [1-3]. These methods have been shown to have higher sensitivity than human visual inspection [4], and are more flexible than the hand-tuned applications that are commonly used for drug screening by high-throughput microscopy. They can be expected to be widely used to study drug effects in a systematic manner [5]. Regardless of the application, the ability to reliably compare subcellular patterns between two conditions is critical to any automated microscopy system.

We have previously described a basic scheme for objective comparison of location patterns [6]. The approach is simple: each image is converted to a vector of features that describe the pattern in it, and then a feature matrix is formed from the vectors for all of the images of a given condition. Whether the patterns for two conditions are different or not can then be determined by testing the hypothesis that the two feature matrices are statistically the same. Such an approach has been shown to be able to distinguish any pair from a set of ten location patterns in HeLa cells [6]. In particular, the patterns of two Golgi proteins, giantin and gpp130, that cannot be distinguished by visual inspection [4] can be differentiated correctly.

However, the power of this approach suffers from the limitations of the testing method used, the Hotelling's T^2 test. It may have low power when the data are not normally distributed. Furthermore, the number of features that can be used in the test must be less than the number of samples.

Therefore we have explored other testing methods to overcome the disadvantages of the T^2 test. Since the distributions of the features of our data are unknown, a distribution free testing method would be more suitable. Usually there are two ways to get a distribution free test. One is to apply a χ^2 test on the binning of data, such as the power-divergence statistic test [7]. However, these tests do not work well on high dimensional data unless the sample sizes are very large. Another way is to do permutation tests. Some permutation tests such as the energy test [8] are computationally expensive because their rejection regions can only be determined by simulation. Therefore we prefer a permutation test whose test statistic has a known closed-form asymptotic distribution under the null hypothesis for any data distribution.

Here we implement two such testing methods and compare their results with those from the T^2 test. A multiple testing method based on false discovery rate theory was also applied to find differences between individual features.

2. DATA PREPARATION

2.1. 2DHeLa Dataset

The 2DHeLa dataset [1] was used in our initial study of location pattern comparison [6]. It includes images from all major subcellular patterns collected by immunofluorescence microscopy, including those of the two similar Golgi patterns mentioned above. For each pattern, there are 73-98 images of the protein channel. Parallel DNA channels were also collected to allow characterization of the distribution of each protein relative to a common frame of reference.

[*] Summer research program participant from the Universidad Metropolitana, San Juan, Puerto Rico.

0-7803-9577-8/06/$20.00 ©2006 IEEE

Table 1. List of cell lines and drugs for drug effects evaluation.

A) Cell lines

Tagged Gene	Tagged Protein
Rab21	GTP-binding protein RAB21 [Hs]
Glut1	Glucose transporter 1
Ap3b1	Adaptor-related protein complex AP-3 βl subunit
Cav1	Caveolin1

B) Drugs

Drug	Stock	Final Conc.
Bafilomycin A1[a]	DMSO[b], 0.5mM	200nM
Brefeldin A	200 proof EtOH, 0.5mg/ml	10μg/ml
Chloroquine	ddH_2O, 10mM	100μM
Nocodazole	DMSO, 10 mg/ml	10μg/ml

[a] A.G. Scientific, Inc., San Diego, CA, USA

[b] Dimethyl sulfoxide

2.2. 3D3T3 Dataset

We also used a 3D dataset from the CD-tagging project (http://cdtag.bio.cmu.edu/www/public/), which seeks to visualize all proteins in NIH 3T3 cells [9]. The CD-tagging technique randomly tags one protein in each cell with a fluorescent protein, GFP. Different lines of tagged cells were isolated, the tagged protein was identified by RT-PCR, and images of each line were collected by spinning disk confocal microscopy [3]. We used 90 clones from this dataset with 9-33 images for each clone.

2.3. 3D3T3 Drug Dataset

This dataset was collected to demonstrate the application of our approach to evaluating drug effects on cells. The images were acquired in the same manner as the 3D3T3 dataset, except that cells were treated with or without drugs 1 h before imaging. Four cell lines (Table 1A) and four drugs (Table 1B) were examined. All drugs were obtained from Sigma Chemical Co. (St. Louis, MO, USA) unless otherwise indicated. 10-20 images per cell-drug pair were collected.

3. METHODS

3.1. Feature Calculation

We have developed sets of Subcellular Location Features (SLFs) for describing either 2D or 3D images [4]. For 2D images, we used the 65 features of SLF6, the same features used previously for comparison of 2D image sets [6]. These features include 11 morphological features, 49 Zernike moment features and 5 edge features. For 3D images, 14 morphological features (SLF14) were calculated [10]. Both 2D and 3D features were designed to be insensitive to the translation and rotation of a cell within an image. Complete descriptions of the SLF can be found at http://murphylab.web.cmu.edu/services/SLF.

3.2. Multivariate Hypothesis Testing

To compare sets of images for two patterns, we assume that feature vectors of each pattern are drawn independently from a single distribution for that condition. If the cumulative distribution functions of the two distributions are denoted as F_1 and F_2 respectively, the null hypothesis H_0 can then be stated as: H_0: $F_1(x)=F_2(x)$, for every feature value x. The two patterns will be considered to be different if the null hypothesis is rejected.

In our previous work, the hypothesis test was conducted using the pooled Hotelling T^2 test. The test statistic, which is the Mahalanobis distance between the two groups, has an F-distribution. The F-distribution has two parameters, or degrees of freedom, which must be positive. In the case of our feature comparison, the second degree of freedom is the total number of images minus the number of features plus one. This results in a limitation that the number of images must be one more than the number of features. Another limitation of the test is that it works best for data sampled from multivariate normal distributions. While some SLFs show a normal distribution, some show more exponential or even bimodal distributions [11].

To overcome these limitations, two nonparametric approaches were investigated here. One is the Friedman-Rafsky (FR) test [12], which compares two groups by analyzing the structure of the minimal spanning tree (MST) of the pooled data. The test statistic is the number of edges that connect samples from different groups in the MST, and we reject the null hypothesis when the test statistic is small. The other approach is the k nearest neighbor (KNN) test [13]. Its test score is the number of occurrences when a k nearest neighbor of a sample belongs to the same group. The null hypothesis is rejected when the test statistic is large. Both the FR test and the KNN test were constructed using Euclidean distance and all features were normalized to zero mean and unit variance before distance calculation.

3.3. Multiple Testing

Once two patterns are found to be different, one may be interested in the source of the difference. Independent univariate tests for each feature are helpful for finding which features contribute to the difference, but the occurrence of false conclusions that a feature is different increases with the number of tested features. A better way to deal with this problem is the Benjamini-Hochberg (BH) method [14], in which the rejection threshold increases linearly with the ascent order of p-values of all the features. For example, if there are m features to test, we will get m p-values from univariate testing. We sort them in an ascending order and the sorted p-values are denoted as $p_1, p_2, \ldots, p_m$. Then the threshold at level α is $t\alpha/m$, where t is the largest value that satisfies $p_t< t\alpha/m$. The false discovery rate, which is defined as the number of false rejections divided by the number of rejections, can be well controlled if the p-values

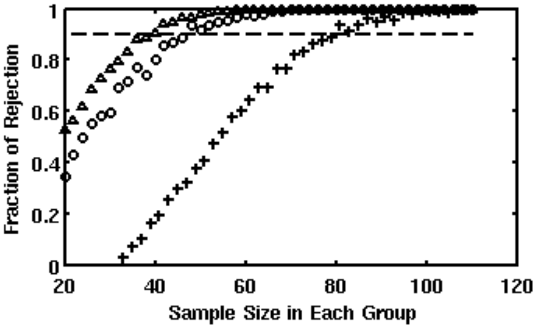

Fig. 1. The power of detecting difference of the testing methods upon different sample sizes for the three methods: the T^2 test (+), the FR test (○) and the KNN test (Δ). The dashed line indicates the level at which 90 percent of null hypotheses were rejected.

are independent or positively dependent. The p-values of SLFs turn out to be positively dependent because different images tend to generate different features. The BH method can also be used as a multivariate test by rejecting the null hypothesis if at least one feature is found to be different.

4. RESULTS

First, pairwise comparison was implemented for the 2DHeLa dataset by the three methods. We set the number of neighbors considered in the KNN test to 3, which has been used successfully for other data [15]. Test statistics were calculated for each pair of image groups. As observed before for the T^2 method [6], all pairs were found to be different at the level 0.05 by all methods. Giantin and gpp130 were the most similar patterns according to test statistic values (data not shown).

In order to compare the power of the three testing methods, we built 1000 pairs of groups by drawing samples from the feature distributions of gpp130 and giantin. Since the real feature distributions are unknown, they were estimated by the Gaussian kernel method (using a kernel density estimation toolbox for Matlab from http://ssg.mit.edu/~ihler/code/kde.shtml). (Examples of the estimated distributions are available as described below.) The more groups that a hypothesis testing method distinguishes at the same level, the more powerful the method. When there are 40 samples in each group, the KNN test is the most powerful method by distinguishing 89% of the pairs and the power of the FR test (79%) is close to the KNN test. However, the T^2 test only distinguished 15% of the pairs. This indicates that the KNN test and the FR test could be much more sensitive than the T^2 test when data size is small. This was also shown in fig. 1, which was obtained by calculating the power of the three methods with different number of samples. According to the figure, to distinguish 90 percent of the pairs, the KNN test required 38 images and the FR test required 46 images in each group.

The nonparametric methods were also shown to have higher power for the 3D3T3 dataset. In the pairwise comparison of the 3D images among 4005 pairs, the T^2 test failed to detect a difference between 226 pairs. The numbers of such undistinguished pairs are 214 and 158 for the FR test and the KNN test respectively.

Table 2. Drug effects evaluation. Each pair was compared by four methods, the T^2 test, the FR test, the KNN test, and the BH method. The values in each cell of the table show whether any difference was detected by the four methods (1 signifies a difference was found).

	Rab21	Glut1	Ap3b1	Cav1
Baf.	0/1/1/1	1/0/1/1	0/0/0/0	0/0/0/0
Bre.	0/0/0/0	0/0/0/0	0/0/0/0	0/1/1/0
Chl.	0/0/0/1	1/0/1/1	0/0/1/0	0/0/0/0
Noc.	0/1/1/1	0/0/0/0	0/0/0/0	0/1/1/1

Although the nonparametric methods showed higher sensitivity, it is possible that this is gained from a higher risk of false rejection. A valid method should only reject true null hypothesis lower than or close to the level. To validate the three testing methods, we randomly drew two groups of 40 images each from the same distribution and compared them. The FR test rejected 40 pairs as different out of 1000 trials, and the KNN test 56. This close match to the 5% rejection expected demonstrated the validity of the nonparametric methods. For 3D data, the false rejection rates were close to 5% for all of the three methods even if there were only 10 images in each group.

Univariate tests have also been tried in the previous work and 7 features were found to be different for the giantin and gpp130 pair [6]. However, no correction was done to reduce the rate of false positives. In the 1000 pairs that were from the same population, over 600 pairs have at least one different feature. When we used the BH method only three features instead of seven came out to be different (the histograms are available as described below) and only 16 out of the 1000 pairs had at least one different feature. Note that even if no individual feature is different it does not mean that the multivariate distributions are not different. The giantin and gpp130 patterns can still be distinguished from each other without the 7 most different features (p-value<0.01 for all the three tests, data not shown).

Finally we applied our testing methods to the small dataset of images in the presence and absence of various drugs. As shown in Table 2, no drug caused a change that was detected by all four methods. This means that the changes could be very minor and a powerful testing method is necessary. Among the testing methods, the KNN test has the highest sensitivity since it found seven different drug-control pairs; the FR test found four. In contrast, the T^2 test only found two affected pairs. It is not surprising that the nonparametric methods detected more changes because these methods have been shown to be more powerful than the T^2 test in the results above. While the correct answer to these tests cannot be known with certainty, they allow the possibility of screening drugs automatically at a specified confidence level.

5. DISCUSSION

This paper showed an improvement in comparing protein subcellular location patterns by using nonparametric testing methods. The improvement was especially significant for small amounts of data. This means that fewer images were required for correct detection of a difference. Although finding exactly the optimal number of images is impossible, the results on comparing giantin and gpp130 give us a clue to determine a reasonable sample size. From Fig. 1, we concluded that around 40 images in each group would allow us to detect minor changes.

The results from the drug evaluation experiments also showed that nonparametric methods were more sensitive. However, we may not be able to conclude that a drug changed a location pattern directly, since a difference may be caused by indirect effects such as induction of cell death or changes in cell shape or size. One way of alleviating this problem is to continually improve the feature set to remove dependency on changes in unwanted parameters.

In many cases, large numbers of cells can easily be acquired. However, in some cases the resulting increase in acquisition time per condition may be undesirable. To balance the tradeoff between speed and accuracy, coarse to fine methods can be used. First, powerful testing methods like the KNN test can be used to find candidates after acquiring a few images. Then more images are only taken for these identified candidates for further studies.

The comparison method we propose is not only helpful for drug screening, but also important for characterizing all proteins in location proteomics. With this method, we can relate one protein to another with a statistical significance level. Furthermore, each protein can be described in various states by specifying under what conditions the location pattern will change.

The data and code used for the work described here are available at http://murphylab.web.cmu.edu/data, along with supplementary figures. The image comparison methods described here are also available through the Protein Subcellular Location Image Database (http://murphylab.web.cmu.edu/services/PSLID).

6. ACKNOWLEDGEMENTS

This work was supported in part by NIH grant R01 GM068845, NSF grant EF-0331657, and Grant 017393 from the Commonwealth of Pennsylvania Tobacco Settlement Fund.

7. REFERENCES

[1] M. V. Boland and R. F. Murphy, "A Neural Network Classifier Capable of Recognizing the Patterns of all Major Subcellular Structures in Fluorescence Microscope Images of HeLa Cells," *Bioinformatics*, vol. 17, pp. 1213-1223, 2001.

[2] M. Velliste and R. F. Murphy, "Automated Determination of Protein Subcellular Locations from 3D Fluorescence Microscope Images," *Proc 2002 IEEE Intl Symp Biomedl Imag*, pp. 867-870, 2002.

[3] X. Chen, M. Velliste, S. Weinstein, J. W. Jarvik, and R. F. Murphy, "Location proteomics - Building subcellular location trees from high resolution 3D fluorescence microscope images of randomly-tagged proteins," *Proc SPIE*, vol. 4962, pp. 298-306, 2003.

[4] R. F. Murphy, M. Velliste, and G. Porreca, "Robust Numerical Features for Description and Classification of Subcellular Location Patterns in Fluorescence Microscope Images," *J VLSI Sig Proc*, vol. 35, pp. 311-321, 2003.

[5] Z. E. Perlman, M. D. Slack, Y. Feng, T. J. Mitchison, L. F. Wu, and S. J. Altschuler, "Multidimensional Drug Profiling by Automated Microscopy," *Science*, vol. 306, pp. 1194-1198, 2004.

[6] E. J. S. Roques and R. F. Murphy, "Objective Evaluation of Differences in Protein Subcellular Distribution," *Traffic*, vol. 3, pp. 61-65, 2002.

[7] T. R. C. Read and N. A. C. Cressie, *Goodness-of-fit statistics for discrete multivariate data.* New York: Springer-Verlag, 1998.

[8] G. Zech and B. Aslan, "A multivariate two-sample test based on the concept of minimum energy," *Proc. PHYSTAT2003*, pp. 97-100, 2003.

[9] J. W. Jarvik, G. W. Fisher, C. Shi, L. Hennen, C. Hauser, S. Adler, and P. B. Berget, "In vivo functional proteomics: Mammalian genome annotation using CD-tagging," *BioTechniques*, vol. 33, pp. 852-867, 2002.

[10] K. Huang and R. F. Murphy, "Boosting Accuracy of Automated Classification of Fluorescence Microscope Images for Location Proteomics," *BMC Bioinformatics*, vol. 5, pp. 78, 2004.

[11] X. Chen and R. F. Murphy, "Objective Clustering of Proteins Based on Subcellular Location Patterns," *J Biomed Biotechnol*, vol. 2005, pp. 87-95, 2005.

[12] J. H. Friedman and L. Rafsky, "Multivariate generations of the Wald-Wolfowitz and Smirnov two-sample tests," *Ann Stat*, vol. 7, pp. 697-717, 1979.

[13] Y. Henze, "A multivariate two-sample test based on the number of k nearest neighbor type coincidences," *Ann Stat*, vol. 15, pp. 772-783, 1988.

[14] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J R Stat Soc Ser B*, vol. 57, pp. 289-300, 1995.

[15] M. F. Schilling, "Multivariate Two-Sample Tests Based on Nearest Neighbors," *J Am Stat Assoc*, vol. 81, pp. 799-806, 1986.