AUTOMATED COMPARISON OF PROTEIN SUBCELLULAR LOCATION PATTERNS BETWEEN IMAGES OF NORMAL AND CANCEROUS TISSUES

Estelle Glory [1,2,3], Justin Newberg [1,2], and Robert F. Murphy [1,2,3,4,5]

Center for Bioimage Informatics[1], Department of Biomedical Engineering[2], Molecular Biosensor and Imaging Center[3], Ray and Stephanie Lane Center for Computational Biology[4], Departments of Biological Sciences and Machine Learning[5], Carnegie Mellon University, Pittsburgh, PA

ABSTRACT

Early cancer diagnosis and evaluation of cancer progression during treatment are two important factors for clinical therapy. In this study we propose a novel approach which automatically compares the subcellular location of proteins between normal and cancerous tissues in order to identify proteins whose distribution is modified by oncogenesis. This study analyzes 258 proteins in 14 different cancer tissues and their corresponding normal tissues using images provided by the tissue microarray collection of the Human Protein Atlas. Using texture features automatically extracted from the tissue images, 14 machine classifiers were trained to recognize the patterns of eight major organelles in each tissue. For each tissue-protein combination, the results of the classifier for normal and cancerous tissues were compared. Eleven proteins were identified as showing differences in location; these proteins may have potential as biomarkers.

Index Terms— Location proteomics, tissue micro-array, pattern recognition, cancer profiling, immuno-histochemistry, biomarker discovery

1. INTRODUCTION

Cancer therapeutics rely upon early detection of cancerous cells before they become metastatic or invasive for successful treatment. Moreover, early detection is important for studying disease progression, and in turn developing more effective therapeutics. Traditional approaches to measure early cancer stages are based on measurement of expression levels of biomarkers linked to specific types of cancer [1]. In addition to expression, protein location changes are also an important factor in cancer. For example, the reduction of cyclin D1-dependent kinase export from the nucleus leads to increased phosphorlyation and inactivation of the Rb protein (a tumor suppressing protein) [2]. In this paper, we present a new automated approach to identifying cancer biomarkers.

This approach involves using automated learning methods to compare subcellular location patterns between normal and cancerous tissues. Our group has successfully developed machine learning systems to recognize major subcellular organelle patterns in fluorescence micrographs of cell cultures [3-5]. With the recent development of tissue microarray technologies (TMA) that produce collections of histological sections of tissue, and the advent of the Human Protein Atlas image database of protein patterns in tissue obtained by TMA [6], we have recently extended our work to automatically classify subcellular location patterns in human tissue [7]. We showed that a single classifier can be trained to distinguish the patterns of 10 different proteins that localize to eight different organelles with an accuracy of 93% across 45 tissues.

Tissue samples that were imaged in the Human Protein Atlas were stained by a combination of two approaches. Indirect immunocytochemistry staining was performed using well-characterized antibodies against specific proteins, a horse radish peroxidase-conjugated secondary antibody, and diaminobenzidine. Oxidation of the diaminobenzidine by the peroxidase creates a brownish precipitate in cell regions expressing a specific protein. Hematoxylin was also used to stain nuclei and cell bodies a bluish color. The Atlas currently contains nearly 3,000,000 images for more than 3,000 different antibodies across 45 normal tissues and 20 cancerous tissues (http://www.proteinatlas.org). Up to three samples of each normal tissue are present, each coming from a different subject. One or two samples from 12 different patients are present for each cancerous tissue [6].

The method described here begins by training classifiers to distinguish eight major subcellular patterns in each tissue type present in the Atlas, through selection of features which are robust to the variability of tissue structures between normal and cancer samples. Then, for each protein, the tissue-specific classifier is applied to the cancerous and normal tissue images to identify proteins whose patterns change for a given cancer.

2. METHODS

We first grouped corresponding normal and cancer tissues using the nomenclature of the Human Protein Atlas

Table I. Subcellular locations of proteins in cancerous and normal tissues as determined from images in the Human Protein Atlas. Columns 1 and 2 show the nomenclature of the Atlas. The remaining columns show how many proteins were identified as belonging to the 8 major location classes ("normal"/"cancerous").

Cancerous tissue	Normal tissue	Endosome	Nucleolus	Nucleus	Mito.	Golgi	Cytos.	ER	Lyso.
Breast cancer	Breast	49/49	29/23	0/18	12/22	39/0	3/10	4/11	0/3
Endometrial cancer	Endometrium 1 & 2	35/103	8/45	8/25	63/2	105/15	5/18	8/19	1/6
Liver cancer	Liver	20/36	14/30	5/5	79/30	41/65	7/11	16/9	7/3
Lung cancer	Lung	6/48	69/68	0/29	87/9	63/56	12/10	0/11	0/6
Malignant lymphoma	Lymph node	16/10	9/7	5/92	36/5	138/71	0/15	14/10	0/8
Ovarian cancer	Ovary	20/31	2/54	7/8	55/8	56/29	1/9	3/3	0/2
Pancreatic cancer	Pancreas	25/43	26/42	1/8	45/14	80/68	1/12	9/3	6/3
Prostate cancer	Prostate	18/41	79/34	1/0	24/39	44/78	23/6	21/12	8/8
Renal cancer	Kidney	28/79	43/61	2/3	46/12	60/25	4/8	14/9	3/3
Skin cancer	Skin	31/24	61/20	0/24	47/6	46/96	4/11	1/3	0/6
Stomach cancer	Stomach 1 & 2	44/67	37/57	0/11	50/23	68/37	3/18	9/0	8/6
Testis cancer	Testis	14/39	61/84	3/19	56/8	49/21	6/11	7/8	0/6
Thyroid cancer	Thyroid gland	36/61	93/12	0/86	18/28	75/17	1/25	8/0	1/3
Urothelial cancer	Urinary bladder	10/53	30/31	0/17	57/3	47/29	3/7	2/6	1/4

(Table I). Among the 20 cancer tissue types, 14 are tissue specific while 6 have no single corresponding normal tissue in the database. For example "malignant_glioma" matched with "Cerebellum" and "Hippocampus", and was thus not used for further analysis.

2.1. Unmixing

The Human Protein Atlas images contain a mixture of bluish hematoxylin staining and brownish diaminobenzidine staining against a white background. For pattern analysis, we needed to distinguish between these signals. Thus, we performed color unmixing before applying automated subcellular recognition methods. Due to experimental variance, the spectra of the stains are not necessarily consistent across images, and the dye intensities are not always proportional to concentration [8]. For these reasons, we used a blind approach, which involves determining initial color bases, and then iteratively solving for both the color bases and the stain spectra. We initialized using the a priori knowledge that there should be two stains per image and each has a different hue [7]. We subsequently solved the variables with non-negative matrix factorization, which has been shown to be effective at unmixing immuno-histochemical stains [8, 9].

2.2. Dataset

Initially, a subset of 385 proteins from the Atlas was processed for this study. For each unmixed image, the level of staining was determined by total image intensity. Images with low diaminobenzidine staining were removed. Protein images for a specific tissue were not removed if at least two images with sufficient staining did remain. This left 258 proteins for further analysis.

2.3. Classification of major subcellular organelle patterns

We have recently trained a single classifier to recognize 10 proteins that localize to eight major organelles (nucleus, nucleoli, Golgi apparatus, endoplasmic reticulum, endosomes, lysosomes, mitochondria, and cytoskeleton) in blindly unmixed images with a resulting accuracy of 81.2% [7]. In this work we chose a set of 18 proteins, each known to localize to one of the eight major organelle classes and for which visual inspection indicated no change in location in cancerous tissues. We trained 14 separate classifiers to distinguish between the subcellular patterns in each of the 14 normal and cancerous tissues. Due to the limited number of normal tissue images, one protein image per normal tissue was added to the training set, while multiple corresponding cancer images were added to the sets. On average, each of the classes had approximately 23 images in each set (samples with no detectible protein were not considered).

A set of 839 features were extracted from the unmixed tissue images as described previously [7]. These "field level" features do not require segmentation of the image into single cell regions and are derived from both morphological image processing and multiresolution texture analysis. A discriminating subset of these features were chosen for the training images using Stepwise Discriminant Analysis feature selection (which we have previously shown outperforms many other methods for a similar subcellular pattern classification task [10]). Support vector machine (SVM) classifiers were then trained on the training set features (using the LIBSVM toolbox from http://www.csie.ntu.edu.tw/~cjlin/libsvm). A grid search with 10-fold cross-validation was used to determine SVM parameters. The classifiers were then applied to the corresponding testing sets for evaluation.

Table II. Number of identified proteins that change their location pattern between normal and cancer conditions in different tissue types. Columns 3, 4, and 5 show how many such proteins are found using three different evaluation methods. Column 6 shows agreement between the max. likelihood and plurality methods, in term of classes the protein have been assigned.

Cancerous tissue	Protein processed (# proteins)	Max. likelihood (# proteins)	Plurality (# proteins)	Majority (# proteins)	Agreement with plurality (# proteins)	Agreement with majority (# proteins)
Breast cancer	136	113	17	0	12	0
Endometrial cancer	233	195	18	3	14	3
Liver cancer	189	150	15	2	11	1
Lung cancer	237	188	18	1	11	1
Malignant lymphoma	218	154	29	3	23	2
Ovarian cancer	144	116	6	2	5	2
Pancreatic cancer	193	145	21	4	14	4
Prostate cancer	218	154	30	5	16	3
Renal cancer	200	138	22	4	13	4
Skin cancer	190	151	19	0	14	0
Stomach cancer	219	174	26	5	19	5
Testis cancer	196	143	21	3	13	2
Thyroid cancer	232	207	7	0	6	0
Urothelial cancer	150	107	13	2	8	1

For each test sample, a classifier determines the probabilities that the sample belongs to one of the eight classes. The sample is assigned a label corresponding to the class it is most likely to belong.

2.4. Application to normal and cancerous tissue

For each protein, up to three images of each normal tissue and up to twelve images of each cancerous tissue were available. Thus, multiple results were available for the subcellular location of each protein in each tissue. We used three approaches to combine these results and assign a final subcellular location for each combination protein-tissue: plurality voting, majority voting, and maximum likelihood. Plurality voting finds, for all images of a single protein, the classification label that occurs the most, and then assigns that location class to the protein. To raise the confidence of the protein classification and make the approach more robust to outliers, only votes for images that were classified with a probability higher than 0.75 were counted. Majority voting uses the same approach but requires that the plurality class contain more than 50% of the images. On the other hand, the maximum likelihood method defines the subcellular location pattern of a protein in a tissue as the class with the highest sum of probabilities for all images in each class. In the case of ties by either the plurality voting or maximum likelihood methods, the label was assigned as "undefined." A listing of the tissues and the corresponding number of proteins assigned to each subcellular location class by the maximum likelihood approach is shown in Table I.

3. RESULTS

The three methods tested to classify the subcellular location of protein in tissues give different levels of confidence in the results. As expected, plurality and majority votings which have been designed to take into account only reliable predicted labels are much more selective than the maximum likelihood voting approach. As a consequence, the counterpart of having a more dependable classification is the increase of the number of proteins with an undetermined location pattern. In contrast, the maximum likelihood method is able to classify every protein distribution in one of the 8 classes which provides a large amount of proteins with a different protein location pattern in normal and cancer tissues (Table II). Proteins whose location pattern has been found in agreement by the maximum likelihood and the plurality voting and with a difference between the normal and cancerous protein location should be considered with a higher interest to find potential biomarkers. This is even truer for the classification having an agreement between the maximum likelihood and the majority voting. We observe that a few such proteins are found for each tissue.

The proteins whose locations were observed to change between cancerous and normal tissue are all candidates for potential biomarkers that could be used to diagnose or monitor cancer. However, given the limited resolution of the tissue images and the limited number of samples in some cases, these comparisons varied extensively in confidence. Therefore, to identify a set of high-confidence potential biomarkers, we focused on classifiers for kidney, prostate and pancreas tissues (which showed the most reproducibility of protein pattern assignments for normal tissue). Using these 3 classifiers we identified 11 proteins with a different protein location pattern in cancer and normal tissues (Table III). An example of one of these proteins is shown in Figure 1.

Table III. Potential biomarkers identified by the majority method.

Ab ID	Protein name	Tissue	Normal	Cancer
13	Neprilysin	Kidney	Lyso.	Mito.
118	Cystatin-C	Prostate	Lyso.	Golgi
		Kidney	ER	Cytos.
135	Keratin-15	Pancreas	ER	Cytos.
		Prostate	ER	Golgi
239	WD repeat protein 13	Kidney	Endo.	Nucleolar
302	-	Pancreas	ER	Cytos.
508	Adlican	Kidney	ER	Cytos.
554	Mac-2 BP	Prostate	Cytos.	Lyso.
850	-	Pancreas	ER	Cytos.
851	GRIPE	Prostate	Cytos.	Golgi
911	Rho/Rac GEF	Pancreas	Golgi	Cytos.
1423	PACS-2	Prostate	ER	Golgi

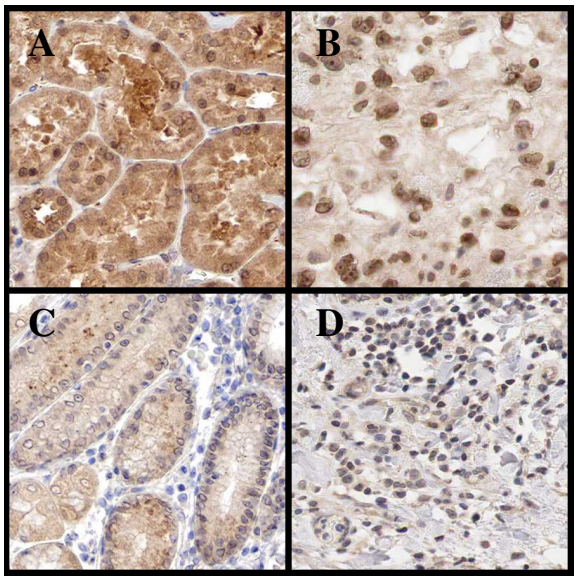

Figure 1. Example protein pattern change between normal and cancerous tissue. Protein is stained brown, while cell nuclei are stained blue. A WD repeat protein (Atlas ID# 239) shows a different pattern in normal tissue (A, C) than in cancerous tissue (B, D) in the kidney (A, B) and stomach (C, D). Classifiers assigned the normal tissue pattern as endosomal, while in cancerous tissue it was assigned as nucleolar.

4. CONCLUSION

We have presented a preliminary study that compares automatically the subcellular location of 258 proteins in cancer and normal tissues. The proposed method uses a classifier trained to recognize eight different subcellular patterns in histochemical stained tissue samples. Eleven proteins with a location pattern change have been identified and require further investigation to confirm their utility. This study illustrates the potential of machine learning systems to process large datasest of biological images produced by high-throughput technology. In future work we plan to refine the robustness of the classification approach (including adding more classes, such as soluble cytoplasmic proteins) and apply it to the full set of proteins provided by the Human Protein Atlas. We hope that creating a subcellular protein location change signature for each tissue-specific cancer will aid in both early diagnosis and monitoring of cancer.

ACKNOWLEDGMENTS

We thank our colleagues in the Center for Bioimage Informatics for helpful discussions. This work was done using images made publicly available by the HPR project, directed by Dr. Mathias Uhlen. This work was supported in part by NIH grant U54 DA0215 (Dr. Brian Athey, PI), by NIH grant U54 RR022241 (Dr. Alan Waggoner, PI), and by NSF ITR grant EF-0331657 (R.F.M.).

REFERENCES

[1] X.C. Song, G. Fu, X. Yang, Z. Jiang, Y. Wang, and G.W. Zhou, "Protein expression profiling of breast cancer cells by dissociable antibody microarray (DAMA) staining.," *Mol Cell Proteomics*, vol. 7, pp. 163-169, 2008.

[2] A.B. Gladden and J.A. Diehl, "Location, location, location: The role of cyclin D1 nuclear localization in cancer," *Journal of Cellular Biochemistry*, vol. 96, pp. 906-913, 2005.

[3] 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.

[4] K. Huang and R.F. Murphy, "Automated classification of subcellular patterns in multicell images without segmentation into single cells," *Proceedings of the 2004 IEEE International Symposium on Biomedical Imaging*, pp. 1139-1142, 2004.

[5] E. Glory and R.F. Murphy, "Automated Subcellular Location Determination and High Throughput Microscopy," *Developmental Cell* vol. 12, pp. 7-16, 2007.

[6] M. Uhlén and F. Ponten, "Antibody-based proteomics for human tissue profiling," *Mol Cell Proteomics*, vol. 4, pp. 384-393, 2005.

[7] J.Y. Newberg and R.F. Murphy, "A Framework for the Automated Analysis of Subcellular Patterns in Human Protein Atlas Images," *Journal of Proteome Research, in press*, 2007.

[8] A. Rabinovich, S. Agarwal, C.A. Laris, J.H. Price, and S. Belongie, "Unsupervised color decomposition of histologically stained tissue samples," in *Advancements in Neural Information Processing Systems*. Vancouver, BC, 2003, pp. 667-674.

[9] H. Seung and D. Lee, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788-791, 1999.

[10] K. Huang, M. Velliste, and R.F. Murphy, "Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images," *Proceedings of SPIE*, vol. 4962, pp. 307-318, 2003.