Using SVM to Predict the Protein Spots in 2D-GE Images

Kuan-Yi Tung[1,2], Shu-Fen Chiou[3], Cheng-Ying Yang[4,5], Min-Shiang Hwang[6,7]*

[1] Department of Industrial Education, National Taiwan Normal University, Taiwan
[2] National Taiwan Science Education Center, Taiwan
[3] Dept. of Information Management, National Taichung University of Science and Technology, Taiwan
[4] Department of Computer Science, University of Taipei, Taiwan
[5] Department of Information Technology, Overseas Chinese University, Taiwan
[6] AI Research Center, Dept. of C.S. and Info. Eng., Asia University, Taiwan
[7] Department of Medical Research, China Medical University Hospital, China Medical University, Taiwan
z50522@gmail.com, sfchiou@nutc.edu.tw, cyang@utaipei.edu.tw, mshwang@asia.edu.tw

Abstract

Proteomics has emerged as a prominent field within bioinformatics in recent years. A proteome is defined as the complete set of proteins expressed by a cell in a given state, and it can vary under different conditions. In proteomic analysis, the initial and essential step is the separation of proteins, for which two-dimensional gel electrophoresis (2D-GE) serves as a crucial and effective technique. With the increasing availability of online image databases, biologists are particularly interested in the information that can be extracted from protein spots. Thus, the rapid and accurate identification of these spots is of great importance. In this work, it proposes a method to separate and detect protein spots in 2D-GE images using the Support Vector Machines (SVMs) technique, with image contrast as the primary feature. First, the image contrasts in 2D-GE images are calculated. Then, the protein spots and background regions are manually labeled within the image. Their contrast values and pixel intensities are subsequently used as training data. A predictive model is then constructed using LIBSVM to identify protein spots in the images. Experimental results demonstrate that the proposed approach achieves promising performance in detecting protein spots from 2D-GE images.

Keywords: 2D-GE image, Protein spots, SVMs, Two-dimensional gel electrophoresis

1 Introduction

High-quality medical images are widely used to enhance disease detection and improve treatment outcomes [6-10]. In recent years, proteomics has become a central topic in bioinformatics, with many tools developed to support diverse research applications. Two-dimensional gel electrophoresis (2D-GE) [2, 15] is a powerful and widely adopted technique in proteomics for separating complex protein mixtures. It combines two classes of polyacrylamide gel electrophoresis (PAGE): proteins are first separated according to their isoelectric point (pI) through isoelectric focusing (IEF) in the horizontal dimension, followed by separation based on molecular weight using sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) in the vertical dimension. The resulting gel is stained and scanned to generate a 2D-GE image, where each visible spot typically corresponds to a unique protein or isoform [20].

These 2D-GE images are essential for downstream analyses, including the identification of protein expression differences, comparisons between diseased and healthy samples, and the detection of post-translational modifications. Protein spots can be excised from gels and analyzed via mass spectrometry to determine protein mass, peptide sequence tags, and other attributes [2, 11]. This information enables database-based protein identification and provides insights into disease mechanisms, ultimately contributing to the development of targeted therapies [14].

Despite its utility, analyzing protein spots in 2D-GE images remains challenging due to variability in gel patterns, noise, and overlapping spots. Over the past decades, a variety of computational methods and software tools have been proposed to address these issues. Well-known examples include PDQuest [19], Melanie, and Delta2D [4], which apply image processing techniques to detect, match, and quantify spots across gels. However, these approaches often struggle with limited accuracy in complex or low-quality images [17].

More recently, machine learning and deep learning methods have been applied to improve protein spot detection. Techniques such as convolutional neural networks (CNNs) and advanced image segmentation have shown promise in achieving higher precision. Nevertheless, these approaches still face challenges, including the requirement for large annotated datasets, variability among biological samples, and the difficulty of distinguishing genuine protein spots from artifacts [12].

Given these limitations, there remains a strong demand for more robust and automated methods for accurate protein spot detection in 2D-GE images. This study seeks to address these challenges by applying LIBSVM to predict protein spots, thereby enhancing the reliability of proteomic analysis and contributing to biomedical research.

*Corresponding Author: Min-Shiang Hwang; Email: mshwang@asia.edu.tw
DOI: https://doi.org/10.70003/160792642026092705016

Protein spot detection in 2D-GE images has long been a critical issue, and numerous methods have been proposed to tackle it [1, 12-13]. Many approaches rely on general image processing techniques without fully exploiting the unique characteristics of 2D-GE images. For example, Bettens et al. [1] employed the Watershed technique, while Persson and Bigun [13] utilized symmetry features for spot detection. Kostopoulou et al. proposed a multi-thresholding scheme; however, determining optimal thresholds remains a difficult task.

In this work, it proposes a method that leverages image contrast as a discriminative feature for protein spot detection. First, the contrasts of 2D-GE images are calculated. Protein spots and background regions are then manually labeled, and their contrast values, along with pixel intensities, are collected as training data. Using LIBSVM, it constructs a predictive model capable of distinguishing protein spots from background regions in the processed images. The rest of this paper is organized as follows: Section 2 introduces 2D-GE and related studies. Section 3 describes the proposed method. Section 4 presents experimental results and discussions. Finally, Section 5 concludes the paper.

2 Related Works

2.1 2D-GE Images

2D-GE [2] is based on Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). It combines two distinct separation methods to resolve proteins in the first and second dimensions based on two independent properties: isoelectric point (PI) and molecular weight.

After the two-dimensional separation, the resulting 2D-GE gel is stained to visualize the proteins. The stained gel is then scanned or photographed to create a digital image, which is referred to as the 2D-GE image. Figure 1 is an 8-bit grayscale 2D gel electrophoresis (2D-GE) image. Every 2D-GE image includes protein spots, background, and noise.

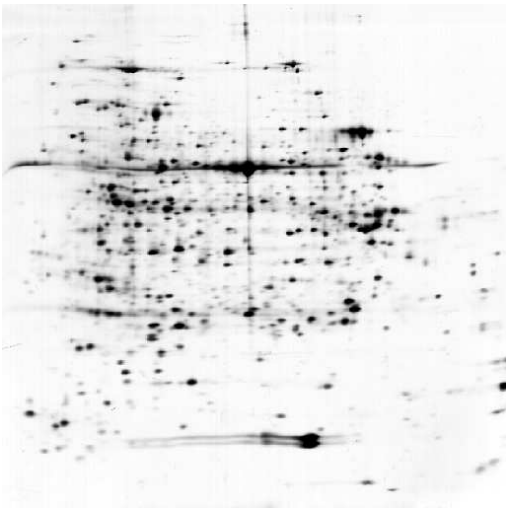

Figure 1. A Sample of 512 × 512 2D-GE image

Protein spots. Here are the key pieces of information found in 2D-GE images. A protein spot in an image generally appears as a roughly dark circle. It is not uniform in color or intensity. The color intensity decreases from the center of the spot to its edge. Figure 2 illustrates a protein spot. Each spot in Figure 1 represents a specific type of protein.

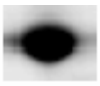

Figure 2. A protein spot

Noises. 2D-GE images typically contain noise from inhomogeneous gels, scanning errors, and other issues. Sometimes, this noise can be easily mistaken for protein spots. It is essential, yet challenging, to distinguish between protein spots and noise.

Background. The non-protein areas of the 2D-GE images, which, in principle, should be uniform and minimal to enhance the contrast of the protein spots.

2.2 The Symmetry Features

According to the method proposed by Persson and Bigun [12], symmetry features within 2D-GE images are identified and used to find protein spots. For each pixel, a feature vector is generated, which includes the local Laplacian of Gaussian (LoG) transform and the local symmetry derivative response. Then, these vectors are classified to achieve the proposed detection of protein spots.

2.3 Image Contrast

In their article [16], Tamura, Mori, and Yamawaki define texture images as having six features: coarseness, contrast, directionality, lifelikeness, regularity, and roughness. According to their definition, most researchers process the image for image retrieval or other tasks [9]. However, it could believe these features can be applied to other image-processing domains. Specifically, the contrast could help differentiate the spots in 2D-GE images for image segmentation.

By analyzing the color distribution in 2D-GE images, it could assess the degrees of chromatic differentiation, referred to as the degree of contrast. Each pixel in the image has a unique contrast value that can be calculated relative to its neighboring pixels. The contrast for the pixel (i,j) is defined in Eq. (1).

$$F_{con}(i,j) = \frac{\sigma}{\left(u_4 / \sigma^4\right)^{1/4}} \tag{1}$$

where σ is the standard deviation of all pixel values in the image, and u_4 is the fourth moment of the image. This contracr feature is used for the prediction with SVM.

2.4 SVM

SVM is a standard tool for data classification [3, 5,

18-19]. This work represents two sets of data as vectors in an n-dimensional space within SVM, aiming to maximize the margin between the datasets and create a separating hyperplane in this space. There are many types of research based on SVM for image classification, text classification, face recognition, and other applications. One of these researches is LIBSVM [5]. It provides a simple interface supporting vector classification, regression, and distribution estimation. In this work, LIBSVM is used to train the protein spot patterns and predict the protein spots in 2D-GE images.

3 The Proposed Method

In the proposed approach, 2D-GE images are processed to separate protein spots from the background. For the clarity, it defines the several parameters. Let N and M denote the width and height of the image, respectively. The contrast of a pixel located at the *i*-th row and *j*-th column is represented as $F_{con}(i, j)$. The method consists of three main stages: (1) Image preprocessing, (2) Model training, and (3) Protein spot prediction.

3.1 Image Preprocessing

The image preprocessing for contrast calculation generally involves the 2D-GE images by evaluating the intensity differences between each pixel and its surrounding neighborhood. In this section, it calculates the contrasts in the testing 2D-GE images for the next steps.

Let the 2D gel electrophoresis (2D-GE) image be represented as *I*. For each pixel in the image, it computes the contrast with its eight neighboring pixels using Eq. (1). After calculating the contrast for all pixels in *I*, it identifies the maximum and minimum contrast values. Then, it normalizes all the contrast values using Eq. (2), where the "max" refers to the maximum contrast and "min" refers to the minimum contrast.

$$C_{i,j} = \frac{F_{con}(i, j)}{(\max - \min)} \quad \textbf{(2)}$$

$C_{i,j}$ represents the advanced contrast of pixel, where *i* and *j* denote the pixel in the *i*-th row and *j*-th column. Therefore, it treats each pixel and the contrast as input values for LIBSVM. This means that for each pixel, its contrast C_(i,j) could be extracted and the input alongs with the corresponding label (e.g., whether it belongs to a protein spot or background) into LIBSVM for training. With this way, LIBSVM learns how to distinguish protein spots from the background based on the contrast of each pixel, and builds a classification model to predict the class of pixels in new images.

3.2 Training Model

In this step, it uses 1/α *I* as the training set and (1-1/α) *I* as the predicting set. Manually, it circles the protein spot patterns in 1/α *I*. Here, α is an integer value. Then it divides the data into two separate datasets: one containing the protein spot patterns set $P\{C_{i,j}\}$, and the other containing the background set $B\{C_{i,j}\}$. These datasets are the input to LIBSVM to generate a model *S*.

3.3 Predicting The Protein Spots

After training data, LiBSVM generates a model *S*. *S* is used to predict the (1-1/α)*I*. Next, all of $C_{i,j}$ and $G_{i,j}$ in (1-1/α)*I* are put into LIBSVM. The LIBSVM will predict the protein spots and mark them up with color by model *S*.

4 Experimental Results

With the previous mentioned operation, it used the original 2D-gel grayscale images in Figure 1, Figure 3, and Figure 4, which named g_1, g_2, and g_3, respectively. The image size is 512×512, and the pixel values of the image span from 0 to 255. In the first step, it computes the contrast of all pixels.

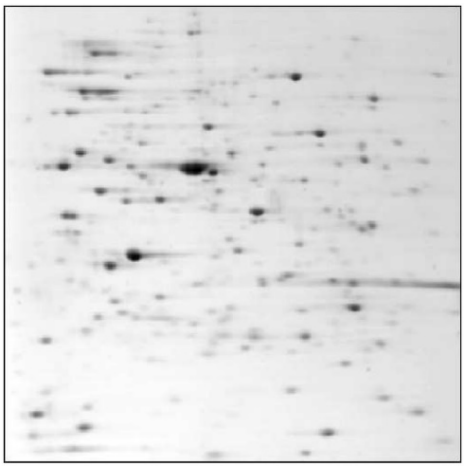

Figure 3. The original image g_2

It sets α=4 and circle the protein spots manually in1/4g_1 as protein spot patterns, and others are background. With circleing the spots, it examines the dark areas closely. Then, it puts these data to train the model S_1 by contrasts of 1/4g_1 using LIBSVM. Repeatedly, it makes the same process in 1/4g_2 and 1/4g_3 to get the models S_2 and S_3.

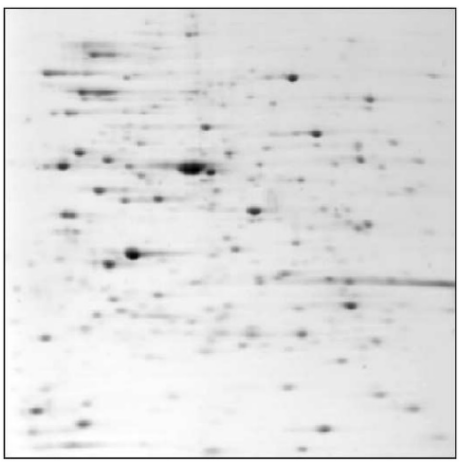

Figure 4. The original image g_3

After obtaining S_1, it could predict the protein spots in 3/4 g_1 using LIBSVM with model S_1. Figure 5 shows the result. The red parts are the ones those are circled as protein spot edges, and the green parts are the ones with the prediction as protein spot edges. It follows the same steps for $3/4g_2$ and $3/4g_3$. Figure 6 and Figure 7 are the experimental results.

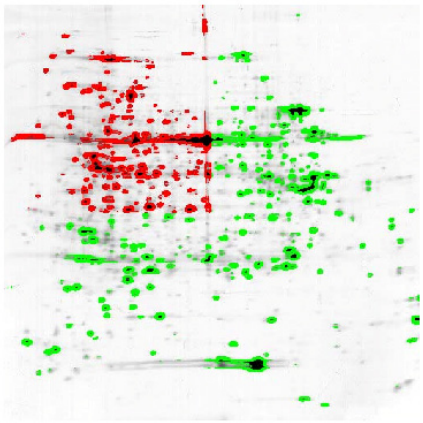

Figure 5. The predicted spot image for g_1

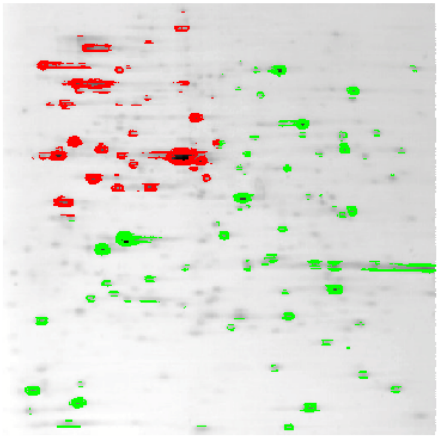

Figure 6. The predicted spot image for g_2

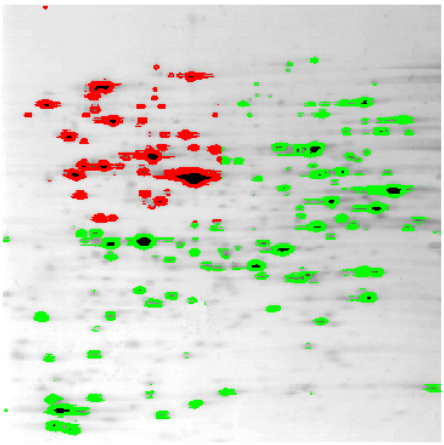

Figure 7. The predicted spot image for g_3

In the experimental results, it could visually identify the most promising protein spots in the 2D-GE images. When marking these possible spots, also, it requires the biologists to verify whether the true protein spots are. According to the experimental results, one image needs its specific model. Hence, in the future, the work could study a general model to predict the protein spots for all images.

5 Conclusion

In this study, the problem of protein spot detection in 2D-GE images is described. Leveraging the chromatic features of the images, it proposed a simple yet effective scheme based on Support Vector Machines (SVMs) to separate protein spots from the background. Specifically, after that the image contrasts were first calculated from 2D-GE images, the protein spots and the background regions were manually annotated in the 1/α image. The contrast and pixel intensity values were used as training data for LIBSVM, which generated a predictive model to identify protein spots in the (1−1/α) image. Experimental results confirmed the feasibility of the approach, though each image required a model tailored to its specific characteristics.

The main limitation of this work lies on the size and the diversity of dataset. The performance of models is strongly influenced by the quality and variability of the images used for training and testing. Since the dataset employed in this study was relatively small, it may not adequately reflect the broad variability present in real-world 2D-GE images. Consequently, the generalizability of models to different experimental conditions—such as variations in sample type, imaging settings, or noise levels—remains limitation

The future work will focus on developing a more generalized model capable of predicting protein spots across diverse datasets. Furthermore, since the current method still relies on expert annotation for initial protein spot identification, future extensions could incorporate additional machine learning algorithms or external databases to achieve a fully automated analysis pipeline. Also, it plans to investigate approaches for scaling the method to larger datasets, thereby enhancing its applicability to high-throughput proteomics. Those possible strategies include optimizing the algorithm for computational efficiency and adopting parallel processing techniques.

Acknowledgement

This work was supported by the Ministry of Science and Technology, Taiwan, under grants NSTC 113-2221-E-845 -005 -MY2.

References

[1] E. Bettens, P. Scheunders, D. V. Dyck, L. Moens, P. V. Osta, Computer analysis of two-dimensional electrophoresis gels: A new segmentation and modelling algorithm, *Electrophoresis*, Vol. 18, No. 5, pp. 792–798, 1997. https://doi.org/10.1002/elps.1150180523

[2] T. S. Chen, P. I. Chen, M. C. Hsiao, Automatic high dynamic range and resolution imaging for 2dge images,

IEEE Proceedings of the 6th World Congress on Intelligent Control and Automation, 2006, Dalian, China, pp. 10507–10511.
https://doi.org/10.1109/WCICA.2006.1714063

[3] Y. Cui, Z. Jin, J. Hu, Research on Intrusion Detection Method Based on Hierarchical Self-convergence PCA-OCSVM Algorithm, *International Journal of Network Security*, Vol. 22, No. 6, pp. 916–924, November, 2020.
https://doi.org/10.6633/IJNS.202011_22(6).04

[4] A. W. Dowsey, J. S. Morris, H. B. Gutstein, G. Z. Yang, Informatics and statistics for analyzing 2-d gel electrophoresis images, in: S. Hubbard, A. Jones (Eds.), *Proteome Bioinformatics*, *Methods in Molecular Biology*, Vol. 604, pp. 239–255, 2010.
https://doi.org/10.1007/978-1-60761-444-9_16

[5] R. E. Fan, P. H. Chen, C. J. Lin, Working set selection using second order information for training Support Vector Machines, *Journal of Machine Learning Research*, Vol. 6, pp. 1889–1918, 2005.
https://www.jmlr.org/papers/v6/fan05a.html

[6] C. Feng, L. Gong, R. Du, Achieving efficient and privacy-preserving online medical primary diagnosis, *International Journal of Network Security*, Vol. 28, No. 4, pp. 929–939, July, 2026.
https://doi.org/10.6633/IJNS.202607_28(4).17

[7] L.-C. Huang, S.-F. Chiou, M.-S. Hwang, A Reversible Data Hiding Based on Histogram Shifting of Prediction Errors for Two-Tier Medical Images, *Informatica*, Vol. 32, No. 1, pp. 69–84, March, 2021.
https://doi.org/10.15388/20-INFOR422

[8] L.-C. Huang, M.-S. Hwang, L.-Y. Tseng, Reversible and high-capacity data hiding in high quality medical images, *KSII Transactions on Internet and Information Systems*, Vol. 7, No. 1, pp. 132–148, January, 2013.
https://doi.org/10.3837/tiis.2013.01.009

[9] L.-C. Huang, L.-Y. Tseng, M.-S. Hwang, A Reversible Data Hiding Method by Histogram Shifting in High Quality Medical Images, *Journal of Systems and Software*, Vol. 86, No. 3, pp. 716–727, March, 2013.
https://doi.org/10.1016/j.jss.2012.11.024

[10] L.-C. Huang, L.-Y. Tseng, M.-S. Hwang, The study of data hiding in medical images, *International Journal of Network Security*, Vol. 14, No. 6, pp. 301–309, November, 2012.
https://doi.org/10.6633/IJNS.201211.14(6).01

[11] K. Marshall, Y. Twum, Y. Li, W. Gao, Chapter One - Spotting targets with 2D-DIGE proteomics, *Advances in Clinical Chemistry*, Vol. 125, pp. 1–22, 2025.
https://doi.org/10.1016/bs.acc.2024.11.003

[12] D. Matuzevičius, Synthetic data generation for the development of 2D gel electrophoresis protein spot models, *Applied Sciences*, Vol. 12, No. 9, Article No. 4393, May, 2022.
https://doi.org/10.3390/app12094393

[13] M. Persson, J. Bigun, Detection of spots in 2-d electrophoresis gels by symmetry features, in: S. Singh, M. Singh, C. Apte, P. Perner (Eds.), *Pattern Recognition and Data Mining*, vol. LNCS 3686, Springer, Berlin, Heidelberg, 2005, pp. 436–445.
https://doi.org/10.1007/11551188_47

[14] A. Punetha, D. Kotiya, Advancements in oncoproteomics technologies: Treading toward translation into clinical practice, *Proteomes*, Vol. 11, No. 1, Article No. 2, March, 2023.
https://doi.org/10.3390/proteomes11010002

[15] S. Saha, S. Banerjee, D. Banerjee, S. Chandra, A. Chakrabarti, 2dge and dige based proteomic study of malignant b-cells in b-cell acute lymphoblastic leukemia, *EuPA Open Proteomics*, Vol. 3, pp. 13–26, June, 2014.
https://doi.org/10.1016/j.euprot.2014.01.002

[16] H. Tamura, S. Mori, T. Yamawaki, Textural features corresponding to visual perception, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 8, No. 6, pp. 460–473, June, 1978.
https://doi.org/10.1109/TSMC.1978.4309999

[17] P. Tsakanikas, E. S. Manolakos, A fully automated 2-DE gel image analysis pipeline for high throughput proteomics, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'11)*, Prague, Czech Republic, 2011, pp. 605–608.
https://doi.org/10.1109/ICASSP.2011.5946476

[18] Z. Wang, W. Wang, J. Zhang, T. Yang, NFC-Defender: SVM-based Attack Detection on NFC-enabled Mobile Device, *International Journal of Network Security*, Vol. 23, No. 3, pp. 379–385, May, 2021.
https://doi.org/10.6633/IJNS.202105_23(3).02

[19] P.-S. Xie, J.-L. Wang, H. Wang, Y.-C. Pan, X.-Y. Li, T. Feng, Industrial Internet Vulnerability Detection Method Based on CBAM-CNN-SVM, *International Journal of Network Security*, Vol. 25, No. 3, pp. 385–393, May, 2023.
https://doi.org/10.6633/IJNS.202305_25(3).01

[20] X. Zhan, B. Li, X. Zhan, H. Schlüter, P. R. Jungblut, J. R. Coorssen, Innovating the concept and practice of two-dimensional gel electrophoresis in the analysis of proteomes at the proteoform level, *Proteomes*, Vol. 7, No. 4, Article No. 36, December, 2019.
https://doi.org/10.3390/proteomes7040036

Biographies

Kuan-Yi Tung received her M.S. degree from University of Taipei, Taiwan. Currently, she is a doctoral candidate in the Department of Industrial Education at National Taiwan Normal University. Also, she is employed as a staff at the National Taiwan Science Education Center and serves as Executive Editor of science journal published by the National Taiwan Science Education. Her research interests include science education, learning assessment and image processing.

Shu-Fen Chiou received a B.B.A degree in Information Management from National Taichung Institute of Technology, Taichung, Taiwan, ROC, in 2004. She studied M.S. degree in Computer Science and Engineering from National Chung Hsing University for one year, and she started to pursue the Ph.D. degree. She received a Ph. D. from Computer Science and Engineering from National Chung Hsing University in 2012. She is currently an assistant professor of department of Information Management, National Taichung University of Science and Technology. Her current research interests include information security, network security, data hiding, text mining and big data analysis.

Cheng-Ying Yang received the M.S. degree in Electronic Engineering from Monmouth University, New Jersey, in 1991, and Ph.D. degree from The University of Toledo, Ohio, in 1999. He is a member of IEEE Satellite & Space Communication Society. Currently, he is employed as a Professor at The University of Taipei, Taiwan. His research interests are performance analysis of digital communication systems, signal processing, and computer security.

Min-Shiang Hwang received M.S. in industrial engineering from National Tsing Hua University, Taiwan in 1988, and a Ph.D. degree in computer and information science from National Chiao Tung University, Taiwan, in 1995. He was a professor and Chairman of the Department of Management Information Systems, NCHU, during 2003-2009. He was also a visiting professor at the University of California (UC), Riverside, and UC. Davis (USA) during 2009-2010. He was a distinguished professor of the Department of Management Information Systems, NCHU, during 2007-2011. He obtained the 1997, 1998, 1999, 2000, and 2001 Excellent Research Award of National Science Council (Taiwan). Dr. Hwang was a dean of the College of Computer Science, Asia University (AU), Taichung, Taiwan. He is currently a chair professor with the Department of Computer Science and Information Engineering, AU. His current research interests include cryptography, Steganography, and mobile computing. Dr. Hwang has published over 300+ articles on the above research fields in international journals.