Content-based progressive transmission of 2-DGE images

T-H Tsai[a], J Chen*[b], T-S Chen[c], S-F Chiou[d], T-S Wang[d], H-C Wu[d], H-F Tsai[e], M Hsieh[f], M-C Chou[a] and **S-Y Li[a]**

[a]Institute of Medicine, Chung Shan Medical University, Taichung, 402 Taiwan

[b]Graduate School of Computer Science and Information Technology, National Taichung Institute of Technology, Taichung, 404 Taiwan

[c]Department of Info. Mgt., National Taichung Institute of Technology, Taichung, 404 Taiwan

[d]School of Medical Technology, Chung Shan Medical University, Taichung, 402 Taiwan

[e]Department of Biology, Life Science Research Center, Tunghai University, Taichung, 433 Taiwan

Abstract: To a biologist, the protein spots and their locations on a 2-D gel electrophoresis (2-DGE) image are important information in protein studies. It is also very important that the 2DGE images can be transmitted and received quickly on the Internet. Currently, the progressive image transmission (PIT) method used for transmitting data on the Internet does not allow the important protein data on the 2-DGE image to be identified quickly after only a few transmissions. The content-based method proposed in the present paper applies the lossless bit-plane method (BPM) to important protein spots and the JPEG lossy compression to the unimportant non-protein background. Experimental results show that the protein spots can be quickly and completely identified after a few transmissions using BPM. Based on the partially transmitted image, the biologist can make a quick decision whether to accept or reject the sample. Once accepted, the remaining compressed background can be quickly transmitted using JPEG progressive transmission.

Keywords: 2-D gel electrophoresis (2-DGE), progressive image transmission (PIT), content-based, bit-plane

1 INTRODUCTION

In proteomics informatics, 2-D gel electrophoresis (2-DGE) is used to isolate protein information.[1] The 2-DGE with protein information can be digitized where the location, number and size of individual protein spots can be recorded and stored for use in bioinformatics researches. Already, a lot of time and money has been invested in research to separate the protein information[2,3] and to develop databases for 2-DGE images.[4,5] These ventures were ambitious and required larger storage capabilities and faster Internet transmissions for web-based databases.

Currently, the widely used progressive image transmission (PIT)[6–10] is one of the faster methods for image transmission on the Internet. In PIT, an image is partially and progressively transmitted over several transmission phases. However, it would be desirable if, after several phases, the image could be roughly perceived by the receiver—enough for the receiver to decide whether to discard or accept the image. A good PIT is one such that: the contour of

The MS was accepted for publication on 29 March 2006.
* *Corresponding author: Graduate School of Computer Science and Info. Tech., National Taichung Institute of Technology, Taichung, 404 Taiwan; e-mail: jeanne@ntit.edu.tw*

the whole image can be perceived after the first few phases; the recovered image is lossless; a lower total number of transmissions is required; faster transmissions are achieved; and last but not least, important data should be completely transmitted after the first few phases.

In the bit-plane method (BPM), a grey image is transmitted in eight transmission phases.[6] Each pixel value in the grey image is 8 bits on which the eight transmission bit-planes are based. Bit-plane slicing is performed from the most significant bit (MSB) to the least significant bit (LSB) planes. Each bit-plane makes a phase in transmission. This method is lossless, and the image can be recovered by combining the bit-planes in the order of transmission. Furthermore, the image can only be perceived after the final transmission.

The JPEG progressive compression method is lossy.[7] First, the image is divided into 8×8-sized blocks in the order from top to bottom and left to right. Next, each block is discrete cosine transformed (DCT) from the spatial domain to the frequency domain. Finally, the transformed coefficients in the blocks are divided into ten phases by zigzag scanned order and transmitted by its respective phases. The entire image can be recovered layer-by-layer by the different phases of the transmitted image. Although the image can only be perceived at the end of transmission, transmission speed is faster with the smaller amounts of data that have to be transmitted.

In bioinformatics searches on web-based databases, it would be desirable if important protein data could be identified after only a few transmissions. A quick and accurate decision could then be made whether to keep or discard the image. In BPM transmissions, the contour of the image cannot be perceived after a few transmissions, but this is compensated for by the lossless transmission. The whole image is transmitted uncompressed, which slows down the transmission speed. Meanwhile, the JPEG progressive compression method is lossy. Similarly, important protein data cannot be clearly defined after a few transmissions. Furthermore, some important protein information can be lost during transmission, which can be undesirable to a biologist. However, this is compensated for by faster transmission speed from transmitting the compressed data.

Neither of these two methods is suitable for the 2-DGE image transmission. However, the special features in these two methods can be combined into a new effective content-based method—suitable for 2-DGE image transmission. In the new content-based method, special care is taken that the important protein data on the 2-DGE will not suffer data loss after transmission. BPM is used to mark out the locations and contents of protein spots which are transmitted uncompressed in the first few phases of transmission—known as content based. The important protein spots can be identified in the first few transmission phases. The background is considered less important and is transmitted in later phases using the JPEG progressive compression method. The transmitted background will suffer some data loss but is considered acceptable. The compressed data will speed up transmission.

Section 2 discusses the PIT for the bit-plane and JPEG progressive compression methods. The experimental results the proposed content-based method are presented in section 3, and analysis in section four. Finally, section 5 concludes.

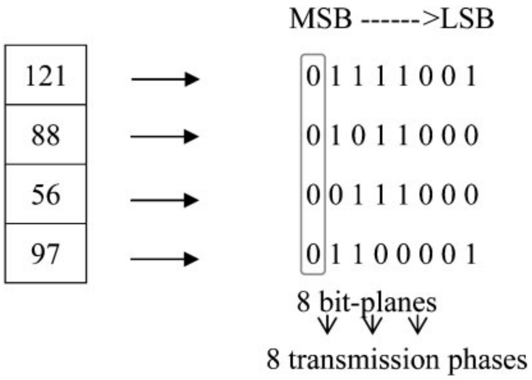
2 PROGRESSIVE IMAGE TRANSMISSIONS

There are many ways to transmit an image progressively—of which BPM is one of the simplest. In BPM, the image is not compressed, and each bit-plane constitutes a transmission phase. Therefore, in total there are the bit-plane number of transmission phases. Transmission is slow, but this is compensated for by being lossless. A second method is lossy JPEG progressive compression, which has a faster transmission speed owing to the smaller amount of data to transmit.

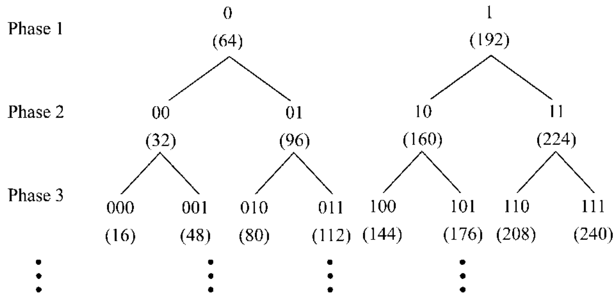
2.1 Bit-plane method

In BPM, bit-planes are sliced from the MSB to the LSB planes. In a grey image with a pixel value of 8 bits, there are eight bit-plane slices; each contributing to one transmission phase. Figure 1 illustrates an example of a grey image with four pixels and bit-plane slicing.

As the image is being transmitted progressively in eight phases, the image is restored by the phase-order received. As illustrated in Fig. 2, starting at phase 1, if the received bit is 0, the restored pixel value is $2^7/2=64$. Otherwise, if the bit is 1, the restored pixel value is $2^7/2+128=192$. In phase 2, bits received are added on to those from phase 1, making up two bits from each pixel. Altogether, there are four possible



1 Bit-planes vs transmission phases



2 Restoring pixel values in each phase

combinations for the two bits; {00, 01, 10, 11}. The pixel values are {32, 96, 160, 224}, respectively. This procedure is repeated iteratively until the eighth phase, where the final pixel values are restored.

2.2 JPEG progressive compression method

The JPEG compression standard was established by ISO and CCITT in 1992. It is a popularly used standard with functionalities such as progressive compression, baseline sequential, lossless coding and hierarchical coding. The progressive compression in JPEG allows an image to be transmitted progressively in several phases, where the image progressively appear from blurred to clear at the end of the last transmission.

In JPEG, the image is first divided into 8×8 blocks. The blocks are then individually DCT into their frequency domain. The eye is most sensitive to the low-frequency domain and least sensitive to the high-frequency domain. The coefficients after DCT are scanned in zigzag order from high to low frequency. The image is then transmitted progressively in the zigzag order scanned. Figure 3 illustrates the scan order for the 128×128-sized 'Lena'.

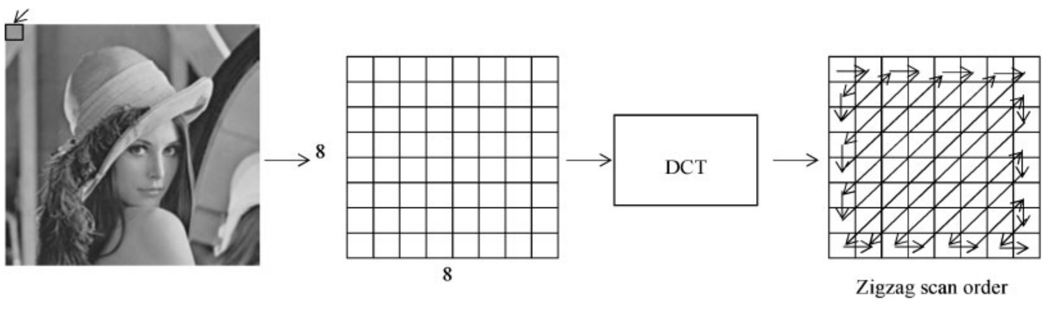
The JPEG progressive compression is a lossy method. Some important data might be lost during the compression process, but the transmission speed is faster.

3 PROPOSED CONTENT-BASED METHOD

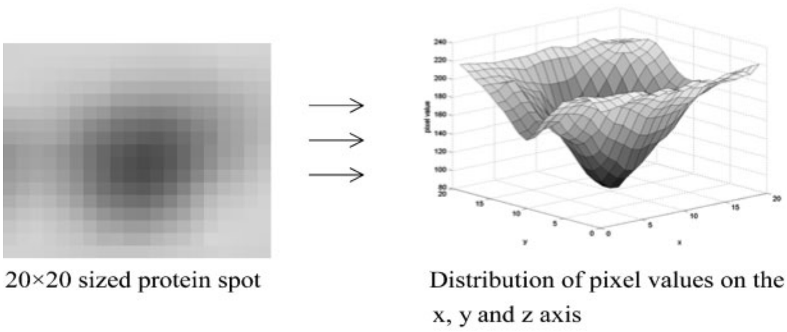
The 2-DGE image can be divided into two parts: one part consisting of important protein spots, while the other is the unimportant background. The lossless bit-plane method is suitable for transmitting the important protein spots, which are transmitted in the first few phases. Meanwhile, the lossy JPEG progressive compression method is used to compress the background for progressive transmissions. By compressing the background, less is transmitted, thereby increasing transmission speed. Lossy from JPEG compression is acceptable since the background is considered unimportant.

3.1 Detecting protein spots

From observation, a protein spot can be characterized as swirling from outside towards the inside, and its pixel value is from large to small. Visually, it is just a grey image with colours from light to dark. Figure 4 shows a 20×20 protein spot and the



3 Order of progressive transmission in JPEG



4 Distribution of pixel values for protein on x, y and z axes

distribution of its pixel values on the x, y and z coordinates.

Step 1. First, the average pixel value b for the $M \times N$ image is calculated with equation (1).

$$b = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} \text{pixel}}{M \times N} \quad (1)$$

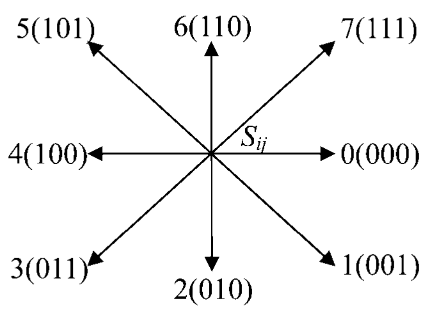
Step 2. In this step, the average background B is calculated. Let p_{ij} be a pixel in the image in the ith row and jth column, Tg be the total of $p_{ij}+b$ for all $p_{ij}>b$, and let Cg be the number of $p_{ij}>b$. The algorithm for calculating B is as follow.

a. Initialize $Tg=0$, $Cg=0$
b. For $i=1$ to M step $+1$
c. For $j=1$ to N step $+1$
d. If $p_{ij>}b$ then $Tg=Tg+(p_{ij}+b)$, $Cg=Cg+1$
e. Loop step (c)
f. Loop step (b)
g. $B=Tg/Cg$

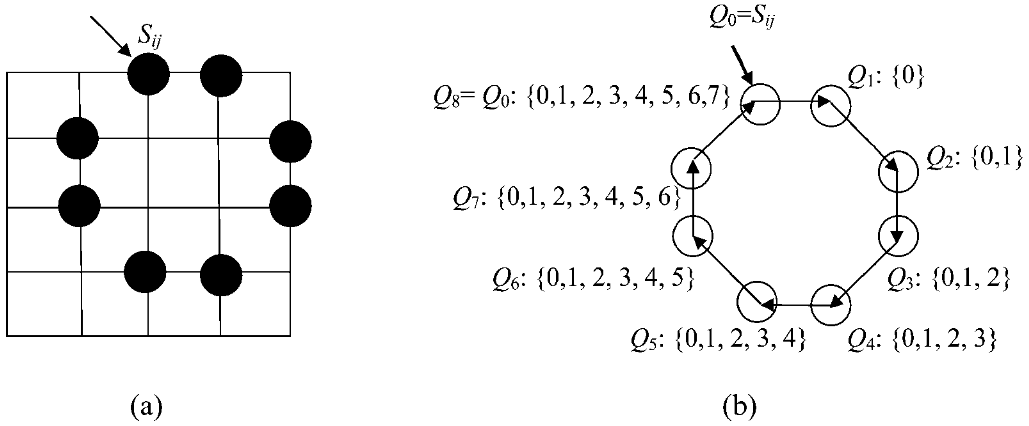
Step 3. Next, a threshold t is used to separate protein spots from the background. Greys in the image are collected and statistically analysed with Otsu's[11] statistical threshold method. The best threshold t for the whole image is determined here. The t threshold is the determinant for separating protein spots from the background. The 2-DGE image is scanned from row to column for the location of protein spots. If $p_{ij} \leqslant t$, the starting point S_{ij} of a protein spot is located at the (i, j) coordinate. S_{ij} can then be extrapolated (details in step 4) to find the outer ring perimeter enclosing the protein spot. Step 3 is repeated until all the protein spots are located.

Step 4. Starting from S_{ij}, a chain code process is used to locate the outer ring perimeter enclosing the protein spot. Eight directions can be extrapolated from S_{ij} (see Fig. 5) and are actually eight pixels surrounding S_{ij}. The eight directions are neighbouring pixels numbered {0, 1, …,7} in the sequence of search, clockwise, for the next point on the outer ring perimeter enclosing the protein spot. First, let S_{ij} be fixed at the top left location. Next, the search starts from direction 0 and moving clockwise towards 7 until the next point on the perimeter is located.

Figure 6 shows a 4×4 block for the protein spot and chain perimeter. Let an image have points $\{Q_0, Q_1, \ldots, Q_n\}$, where n is the number of points on the perimeter enclosing the protein spot, as shown in Fig. 6a. Let S_{ij} be point Q_0. At Q_0, the search starts from direction 0 to 7 for the next point on the perimeter; Q_1 is found and happens to be in direction 0; Q_1: {0}. Next, Q_1 is searched for, starting from 0 to 7; Q_2 is found and happens to be in direction 1 (Q_2 had to go through two comparisons before a successful hit; P_2: {0,1}). The process is repeated to find the rest of the points on the perimeter. This process can also be



5 Extrapolation of S_{ij} in eight directions and their search sequences



6 Protein spot: (a) protein spot in a 4×4 block; (b) chain code identifying perimeter enclosing protein spot

used successfully to locate irregularly shaped perimeters enclosing the protein spots.

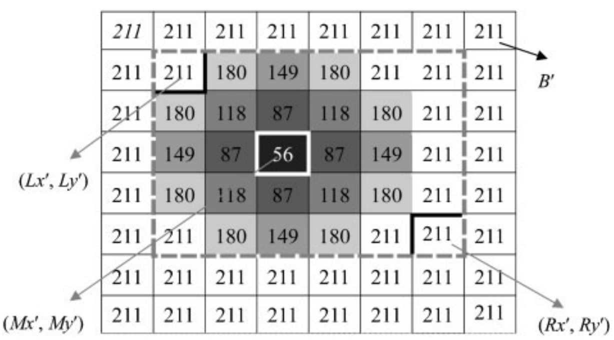
Step 5. Three coordinates, namely (Mx, My), (Lx, Ly) and (Rx, Ry) are collected for each protein spot. (Mx, My) is the middle coordinate of the topmost and the bottommost coordinates; (Lx, Ly) is the top-left coordinate; and (Rx, Ry) is the bottom-right coordinate. Steps 3–5 are repeated until all protein spots are located and relevant information recorded. The coordinate information for all protein spots together with the background average B are arithmetic-coded before being transmitted in the first phase.

3.2 Transmitting and receiving

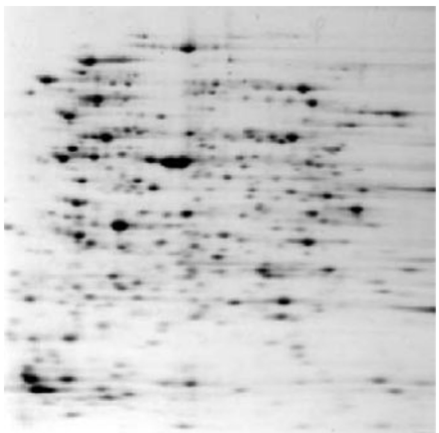
Once the protein spots have been separated from the background and all relevant data including the coordinates are collected, the next process is data transmission. Altogether, there are 19 phases of transmissions for the proposed content-based method.

Important information on the protein spots is transmitted in phase 1. Protein information including the middle, top-left and bottom-right coordinates for all protein spots and the background average B are all arithmetic-coded, compressed and transmitted. At the receiving end, the data are decompressed and the first phase for restoring the 2-DGE image begins. Of the three coordinates retrieved for each protein spot, the top-left coordinate (Lx, Ly) and bottom-left coordinate (Rx, Ry) are used to define an approximate rectangular realm for the protein spot. Using the middle coordinate (Mx, My) as centre, a circle can be defined to be the axiom for an approximation of a protein spot (see Fig. 7). The density of a protein spot is from inside out, with the smallest value pixel to the biggest and greyest to lightest. Taking advantage of this knowledge, the centre of the circle is set to a certain value, and the values of its surrounding pixels within the circle are progressively increased from 0 to B until all the pixels are completely restored. The process is repeated for restoring all the other protein spots. Pixels not within the circles are set to the background average B. Let the partially restored image be I_1.

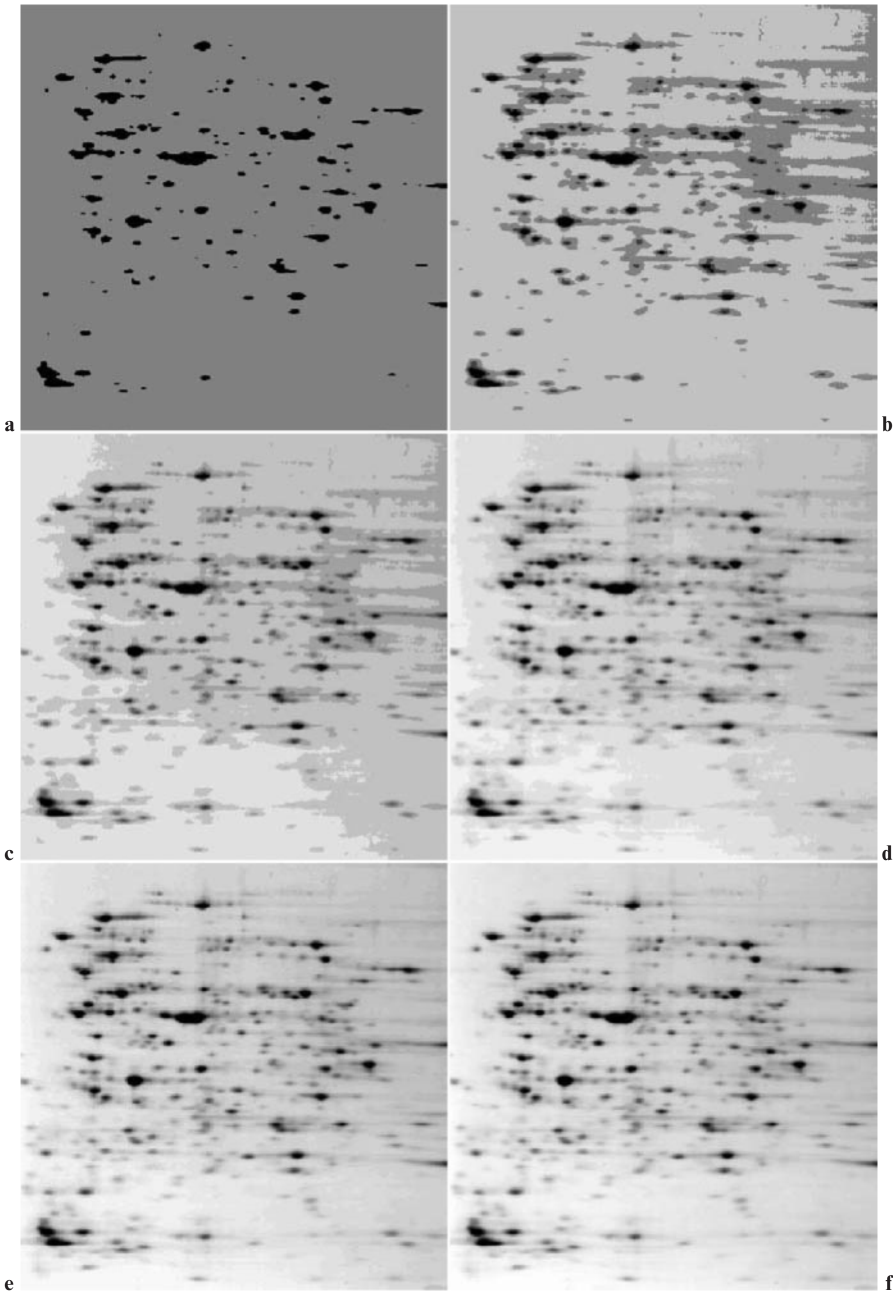
Phases 2–10 will be devoted mainly to restoring the full pixel values of the protein spots. Suppose the original image is O. Let the differenced values



7 Restored approximation of protein spot



8 Original 512×512 2-DGE image



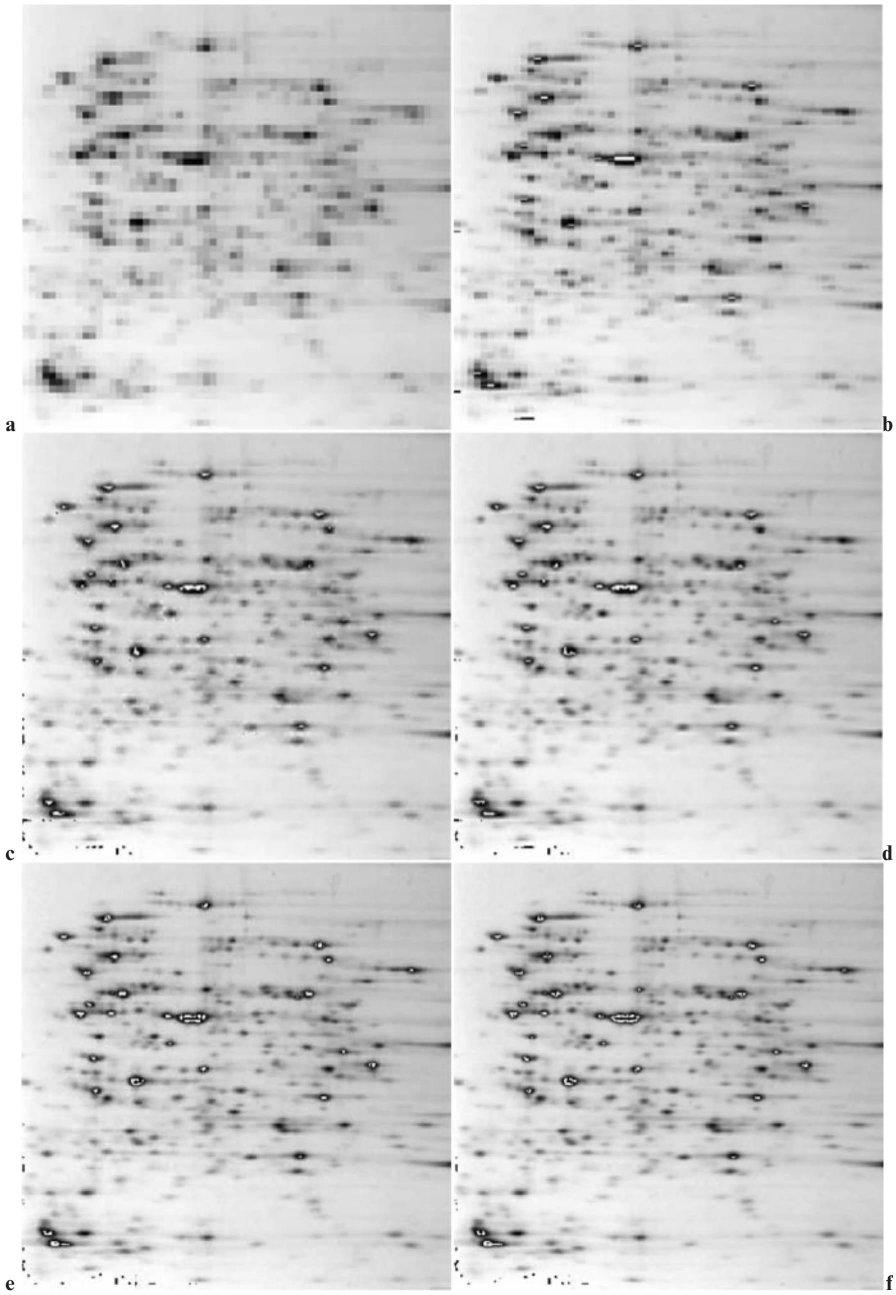
9 Partial images from bit-plane transmission method for phases 1–6: (a) phase 1: amount 32 kB; (b) phase 2: amount 64 kB; (c) phase 3: amount 96 kB; (d) phase 4: amount 128 kB; (e) phase 5: amount 160 kB; (f) phase 6: amount 192 kB

between O and I_1 be $\{T_1, T_2, \ldots, T_D\}$ for all the pixels in the protein spots and $\{G_1, G_2, \ldots, G_d\}$ for the background. D is the total number of pixels in the protein spots in image O, and d is the total number of background pixels. The differenced values are signed. Each negative value is represented by bit value 1 and positive value by bit value 0.

In phase 2, only bit streams of 0s and 1s representing the signs for all protein spots are transmitted. No translation is done at the receiver end.

The differenced values $\{T_1, T_2, \ldots, T_D\}$ for the protein spots are bit-plane sliced from the MSB to the LSB planes. Each bit-plane makes up one transmission phase. Since each pixel in the grey image is 8 bit, there are eight bit-planes altogether. The MSB plane is progressively transmitted in phase 3, followed by phase 4 for the next most significant, until phase 10, where the LSB plane is transmitted.

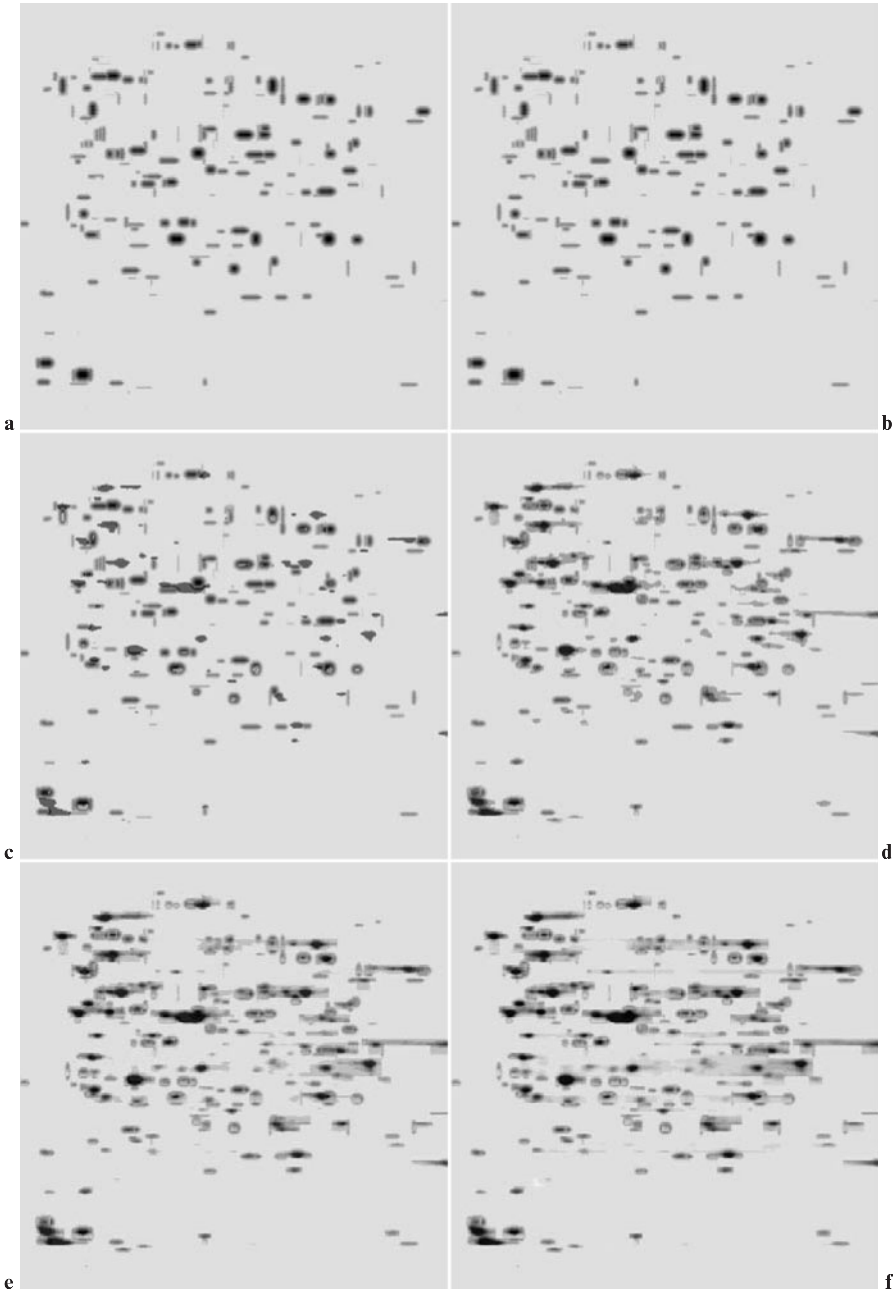
For example, if phase 3 is received, then bits from phase 2 are combined with those from phase 3 with



10 Partial images from JPEG compression method for phases 1–6: (a) phase 1: amount 8 kB; (b) phase 2: amount 16 kB; (c) phase 3: amount 24 kB; (d) phase 4: amount 36 kB; (e) phase 5: amount 48 kB; (f) phase 6: amount 56 kB

combinations {00, 01, 10, 11}. From these four types of combinations, the values represented are {+64, +192, −64, −192}, respectively. Progressively, the protein spots values are recovered with the completed values at phase 10.

The background difference values $\{G_1, G_2, \ldots, G_d\}$ are transmitted in phases 11–19. The signed information for the background differenced values is transmitted in phase 11. The actual background differenced values are to be transmitted in phases 12–19. $\{G_1, G_2, \ldots, G_d\}$ is a bit-plane sliced from the MSB to the LSB and transmitted in this order from phases 12–19, respectively. The process for combining the signs from phase 11 with phase 12 is the same as those in phase 2 with phase 3, respectively. Similarly, the progressive recovery of the background from phases 13–19 is similar to those from phases 3–10, respectively. Suppose image O' is completely restored at phase 19.

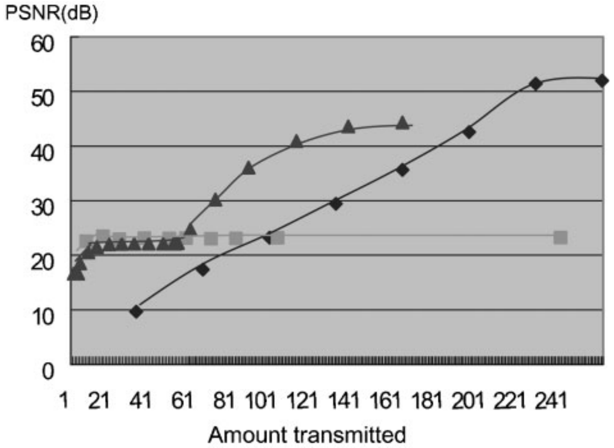


11 Partial images from proposed content-based method for phases 1–6: (a) phase 1: amount 2.19 kB; (b) phase 2: amount 3.98 kB; (c) phase 3: amount 5.36 kB; (d) phase 4: amount 8.86 kB; (e) phase 5: amount 13.36 kB; (f) phase 6: amount 18.77 kB

Image O' should be similar to O with some lossy distortions only in the background and lossless (no distortion) to the protein spots.

4 EXPERIMENTAL RESULTS AND ANALYSIS

In a 2-DGE image, it is desirable to be able to identify the protein spots quickly after a few transmissions. Figure 8 illustrates the original 512×512 2-DGE image. Figure 9a and b illustrates partial images recovered from each phase of the bit-plane transmission method for the different amount of data for each phase. Figure 10 illustrates the partial images recovered from the JPEG progressive compression method. Figure 11 illustrates partial images from the six transmissions, using the proposed content-based method.



12 Relation between PSNRs from different transmission methods: ♦ bit-plane method; ■ JPEG progressive coding; ▲ content-based

In BPM, the protein spots are seen as darker spots (see Fig. 9a). The protein spots are sliced from the MSB plane to the LSB plane and transmitted in this respective order. The restored image for each phase does not match the basic characteristic of the protein spots in the 2-DGE. In actual fact, the density of the original protein spots is distributed from inside out, with colours from dark to light. This means that, if the pixel values of the protein spots are smaller, their colours should be lighter. It takes a few more transmissions before the protein spots can be seen clearly. In the JPEG progressive compression method, the protein spots appeared rather distorted as block artefacts (see Fig. 10a). The size and locations of the protein spots are still distorted after six transmissions. In the proposed content-based method, however, the location, size and colour of the protein spots can be clearly identified at phase 1 (see Fig. 11a).

Generally speaking, a higher PSNR reflects higher similarity between the recovered image and its original. As shown in Fig. 12, with BPM, the PSNR is high for each transmission. However, the content-based method showed higher PSNRs then the other two methods at amounts $\geqslant$32 kBs. After the sixth transmission with 121 kB transmitted, BPM achieve a PSNR of 26 dB, JPEG progressive coding achieved a PSNR of 24 dB and the content-based method, a PSNR of 40 dB. As seen in the graph, after the third transmission, PSNRs for the content-based are higher. Visually, Fig. 12d–f shows closer similarities in the recovered images and the original one in Fig. 8. This proves that, in the content-based method, the protein spots can be identified at a lesser number of transmissions than the other two methods.

5 CONCLUSIONS

The experimental results show that the PSNRs are higher in the proposed content-based method after four transmissions. The recovered image also appears perceptually similar to the original. These results prove that the content-based method is better than either BPM or JPEG. With the increased need for transmitting larger 2-DGE images on the Internet, this could be a novel solution in situations where quick decisions are needed on whether to discard or accept an image. A good application would be for medical images, which are often very large and time-consuming to search on the web.

REFERENCES

1 Gibas, C. and Jambeck, P. *Developing Bioinformatics Computer Skills*, 2001 (O'Reilly).

2 Smilansky, Z. Automatic registration for images of two-dimensional protein gels. *Electrophoresis*, 2001, **22**, 1616–1626.

3 Veeser, S., Dunn, M. J. and Yang, G.-Z. Multiresolution image registration for two-dimensional gel electrophoresis. *Proteomics*, 2001, 856–870.

4 Hojte, P. B. and Wang, X. Registering electrophoresis images for bioinformatics study of protein. Proc. Int. Conf. on *Multimedia and Expo*, July 2003, Vol. 3/III, pp.465–468.

5 Wilson, R. Modelling of 2D gel electrophoresis images for proteomics databases. Proc. 16th Int. Conf. on *Pattern Recognition*, 2002, Vol. 1, pp. 767–770.

6 Chang, C. C., Chen, T. S. and Huang, C. F. *Electronic Imaging Techniques*, 2003 (Flag), pp.130–134.

7 Chen, T. S. and Lin, C. Y. A new improvement of JPEG progressive image transmission using weight table of quantized DCT coefficient bits. *Lect. Notes Comput. Sci.*, *LNCS*, 2002, **2532**, 720–728.

8 Chang, C. C., Hsia, H. C. and Chen, T. S. A new progressive image transmission scheme based on block truncation coding. In *Human Society @Internet*, 2001, pp. 383–397 (Springer, Berlin).

9 Dilmaghani, R. S., Ahmadian, A., Ghavami, M. and Aghvami, A. H. Progressive medical image transmission and compression. *IEEE Signal Processing Lett.*, 2004, **11**(10), 806–809.

10 Salous, M. N., Pycock, D. and Cruickshank, G. S. CBIT—context-based image transmission. *IEEE Trans. Inf. Technol. Biomed.*, 2001, **5**(2), 159–170.

11 Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybernet.*, 1979, **9**(1), 62–66.