International Journal of Biometrics and Bioinformatics (IJBB)

ISSN : 1985-2347

VOLUME 4, ISSUE 3

PUBLICATION FREQUENCY: 6 ISSUES PER YEAR

International Journal of Biometrics and Bioinformatics (IJBB)

Volume 4, Issue 3, 2010

Edited By
Computer Science Journals
www.cscjournals.org

Editor in Chief Professor João Manuel R. S. Tavares

International Journal of Biometrics and Bioinformatics (IJBB)

Book: 2010 Volume 4, Issue 3

Publishing Date: 31-07-2010

Proceedings

ISSN (Online): 1985-2347

IJBB Journal is a part of CSC Publishers

http://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers

Editorial Preface

This is the third issue of volume four of International Journal of Biometric and Bioinformatics (IJBB). The Journal is published bi-monthly, with papers being peer reviewed to high international standards. The International Journal of Biometric and Bioinformatics is not limited to a specific aspect of Biology but it is devoted to the publication of high quality papers on all division of Bio in general. IJBB intends to disseminate knowledge in the various disciplines of the Biometric field from theoretical, practical and analytical research to physical implications and theoretical or quantitative discussion intended for academic and industrial progress. In order to position IJBB as one of the good journal on Bio-sciences, a group of highly valuable scholars are serving on the editorial board. The International Editorial Board ensures that significant developments in Biometrics from around the world are reflected in the Journal. Some important topics covers by journal are Bio-grid, biomedical image processing (fusion), Computational structural biology, Molecular sequence analysis, Genetic algorithms etc.

The coverage of the journal includes all new theoretical and experimental findings in the fields of Biometrics which enhance the knowledge of scientist, industrials, researchers and all those persons who are coupled with Bioscience field. IJBB objective is to publish articles that are not only technically proficient but also contains information and ideas of fresh interest for International readership. IJBB aims to handle submissions courteously and promptly. IJBB objectives are to promote and extend the use of all methods in the principal disciplines of Bioscience.

IJBB editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJBB. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can

provide to our prospective authors is the mentoring nature of our review process. IJBB provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

Editorial Board Members
International Journal of Biometrics and Bioinformatics (IJBB)

Editorial Board

Editor-in-Chief (EiC)

Associate Editors (AEiCs)

Editorial Board Members (EBMs)

Table of Content

Volume 4, Issue 3, July 2010

Pages

Performance Comparison Of 2-D DCT On Full/Block Spectrogram And 1-D DCT On Row Mean Of Spectrogram For Speaker Identification

Dr. H. B. Kekre hbkekre@yahoo.com
Senior Professor,
MPSTME, SVKM's NMIMS University
Mumbai, 400-056, India.

Tanuja K. Sarode tanuja_0123@yahoo.com
Ph.D. Scholar, MPSTME, SVKM's NMIMS University
Assistant Professor, TSEC, Bandra (W),
Mumbai, 400-050, India.

Shachi J. Natu shachi_natu@yahoo.com
Lecturer, TSEC, Bandra (W),
Mumbai, 400-050, India.

Prachi J. Natu prachi.natu@yahoo.com
Lecturer, TSEC, Bandra (W),
Mumbai, 400-050, India.

Abstract

The goal of this paper is to present a very simple approach to text dependent speaker identification using a combination of spectrograms and well known Discrete Cosine Transform (DCT). This approach is based on use of DCT to find similarities between spectrograms obtained from speech samples. The set of spectrograms forms the database for our experiments rather than raw speech samples. Performance of this approach is compared for different number of coefficients of DCT when DCT is applied on entire spectrogram, when DCT is applied to spectrogram divided into blocks and when DCT is applied to the Row Mean of a spectrogram. Performance comparison shows that, number of mathematical computations required for DCT on Row Mean of spectrogram method is drastically less as compared to other two methods with almost equal identification rate.

Keywords: Speaker identification, Speaker Recognition, Spectrograms, DCT, Row Mean

1. INTRODUCTION

With an extensive use of internet technology and a switch over from single user applications to multi-user applications, security has become a major issue. To provide security, it has become crucial to identify users and to grant access only to those users who are authorized. Problem of identifying users can be handled using various approaches either separately or in combination with each other. More and more sophisticated techniques are used with the increase in need of security. Uses of login and password, retinal blood vessel patterns, face recognition, fingerprint recognition are some of the widely used techniques. Login and password technique is not secure enough. This is because attackers can easily steal the password using sophisticated electronic

eavesdropping techniques [1]. Techniques like face recognition, fingerprint recognition and retinal blood vessel patterns also have their own drawbacks. To identify an individual by these methods, he/she should be willing to undergo the tests and should not get upset by these procedures. Speaker identification allows non-intrusive monitoring and also achieves high accuracy rates which conform to most security requirements. Speaker recognition is the process of automatically recognizing who is speaking based on some unique characteristics present in speaker's voice [2]. For this recognition purpose, preserving the speaker specific characteristics present in the speech signal is important. Speaker recognition can be classified into two main categories, namely speaker identification and speaker verification. Speaker identification deals with distinguishing a speaker from a group of speakers. In contrast, speaker verification aims to determine if a person is the one who he/she claims to be from a speech sample. Speaker identification problem basically consists of two stages: feature extraction stage and pattern classification stage. For the given test utterance, classifier finds out which speaker has pronounced this utterance. To perform this job, models are constructed for each speaker using training data. Speaker specific information from the test utterance is then compared with these models to generate similarity measure so that test utterances can be related to each speaker. These classifiers are of various types and can be grouped into template based and stochastic based classifiers [3]. Template based classifiers are the simplest one. Examples of template based classifiers are: Dynamic Time Warping and Vector Quantization. Stochastic models provide better flexibility and more meaningful results in the form of probabilistic scores [4]. Gaussian Mixture Model, Hidden Markov Model, Neural Networks are the examples of stochastic models.

Speaker identification can be further categorized into text-dependent and text independent speaker identification based on the relevance to speech contents [2]. The text dependent speaker identification can be either a 'closed set' or an 'open set' speaker identification [2]. In closed set problem, from N known speakers, the speaker whose reference template has the maximum degree of similarity with the template of input speech sample of unknown speaker is obtained. This unknown speaker is assumed to be one of the given set of speakers. In the open set text dependent speaker identification, matching reference template for an unknown speaker's speech sample may not exist. In this paper, closed set text dependent speaker identification is considered. In the proposed method, speaker identification is carried out with spectrograms and DCT [15-18]. Thus an attempt is made to formulate a digital signal processing problem into pattern recognition of images.

The rest of the paper is organized as follows: in section 2 we present related work carried out in the field of speaker identification. In section 3 we discuss spectrograms. In section 4 we present our proposed approach. Section 5 elaborates the experiment conducted. Results are tabulated in section 6. Conclusion has been outlined in section 7.

2. RELATED WORK

Many approaches are available in literature for speaker identification process based on various approaches for feature extraction. Feature extraction is the process of extracting subset of features from the entire feature set. The basic idea behind the feature extraction is that the entire feature set is not always necessary for the identification process.

The Mel Frequency Cepstrum Coefficients (MFCC) is one of the popular techniques of feature extraction. The MFCC parameter as proposed by Davis and Mermelstein [5] describes the energy distribution of speech signal in a frequency field. Wang Yutai et. al. [6] has proposed a speaker recognition system based on dynamic MFCC parameters. This technique combines the speaker information obtained by MFCC with the pitch to dynamically construct a set of the Mel-filters. These Mel-filters are further used to extract the dynamic MFCC parameters which represent characteristics of speaker's identity.

Sleit, Serhan and Nemir [7] have proposed a histogram based speaker identification technique which uses a reduced set of features generated using MFCC method. For these features, histograms are created using predefined interval length. These histograms are generated first for

all data in feature set for every speaker. In second approach, histograms are generated for each feature column in feature set of each speaker.

Another widely used method for feature extraction is use of linear Prediction Coefficients (LPC). LPCs capture the information about short time spectral envelope of speech. LPCs represent important speech characteristics such as formant speech frequency and bandwidth [8].

Vector Quantization (VQ) is yet another approach of feature extraction [19-22, 25]. In Vector Quantization based speaker recognition systems; each speaker is characterized with several prototypes known as code vectors [9]. Speaker recognition based on non-parametric vector quantization was proposed by Pati and Prasanna [10]. Speech is produced due to excitation of vocal tract. Therefore in this approach, excitation information can be captured using LP analysis of speech signal and is called as LP residual. This LP residual is further subjected to non-parametric Vector Quantization to generate codebooks of sufficiently large size. Combining nonparametric Vector Quantization on excitation information with vocal tract information obtained by MFCC was also introduced by them.

3. SPECTROGRAMS [11]

A spectrogram is an image that shows how the spectral density of a signal varies with time. Spectral density describes how the energy of a signal is distributed with frequency. If f(t) is a finite energy signal, its spectral density is the square of the magnitude of continuous Fourier transform of the signal. The most common format of showing a Spectrogram is a graph with two geometric dimensions. The horizontal axis represents time, whereas the vertical axis represents frequency. A third dimension indicating amplitude of a particular frequency is represented by the intensity or color of each point in the image.

Spectrograms can be created in one of the two ways: using a series of bandpass filters or by calculating Short Time Fourier Transform (STFT) for the signal. The first approach usually uses analog processing, while the second one is a digital process. In the approach using STFT, digitally sampled data are divided into chunks of specific size say 128, 256 etc. which usually overlap. Fourier transform is then obtained to calculate the magnitude of the frequency spectrum for each chunk. Each chunk then corresponds to a vertical line in the image, which is a measurement of magnitude versus frequency for a specific moment in time.

4. PROPOSED APPROACH

In the proposed approach, first we converted the speech samples collected from various speakers into spectrograms. This was done using the second approach of creating spectrogram as mentioned in section 3. Thus we converted the speech sample database into image database. These spectrogram images are then resized to 256 x 256 sizes. The Discrete Cosine Transform [12, 23, 24] is then applied to these images in three different ways to obtain their feature vectors. In the first one, DCT is applied to entire image. Out of total database, 60% of images were used as trainee images and 40% images were used for testing purpose. Euclidean distance between test image and trainee image is used as a measure of similarity. Euclidean distance between the points X(X_1, X_2, etc.) and point Y(Y_1, Y_2, etc.) is calculated using the formula shown in equation (1).

$$D = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \quad \text{.....(1)}$$

Smaller is the Euclidean distance between test image and trainee image, more accurate speaker identification is achieved. Fig. 1 shows the flowchart for the first method using DCT.

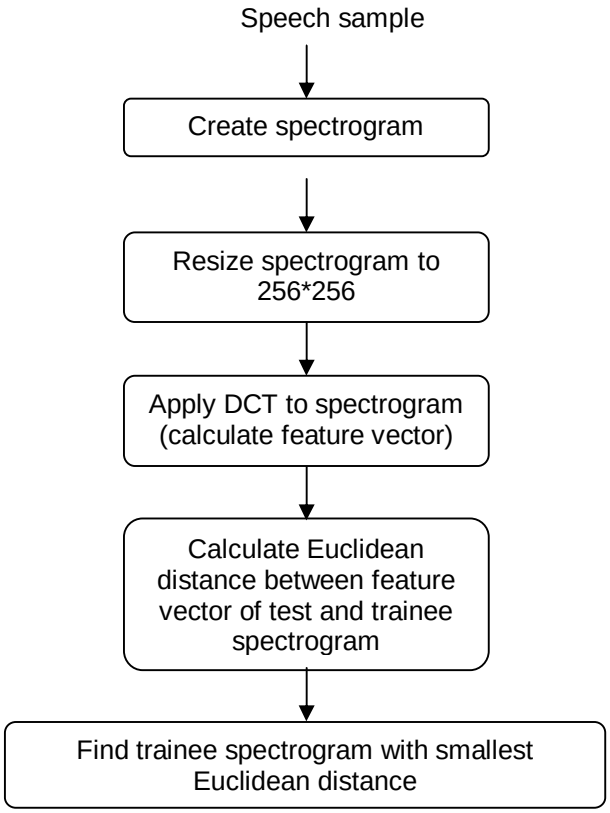

Fig.1: Flowchart for the proposed approach 1

In the second method, resized image is divided into four equal parts as shown in Fig.2 and then DCT is applied to each part. DCT for each block when appended as columns forms a feature vector for an image. Again Euclidean distance is used as a measure of similarity. Fig. 3 shows the flowchart for second method.

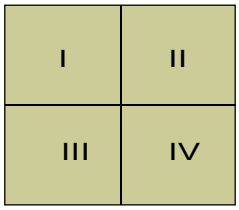

Fig.2: Image divided into four equal nonoverlapping parts

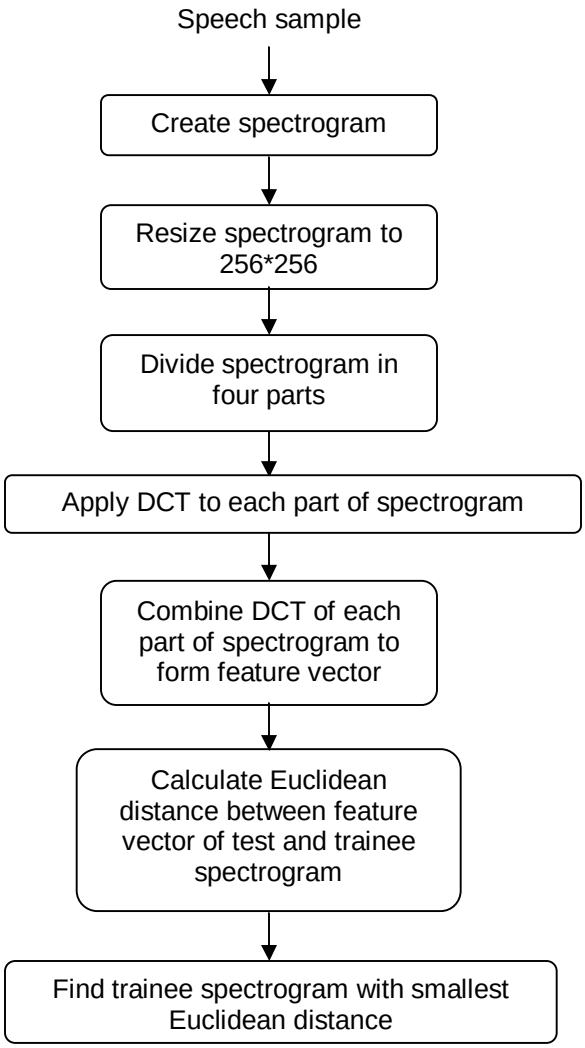

Fig.3: Flowchart for the proposed approach 2

In the third method, Row Mean of an image is calculated [26]. Row mean is nothing but an average of pixel values of an image along each row. Fig. 4 shows how the Row Mean of an image is obtained. DCT is then calculated for this Row mean of an image and Euclidean distance is used to identify speaker.

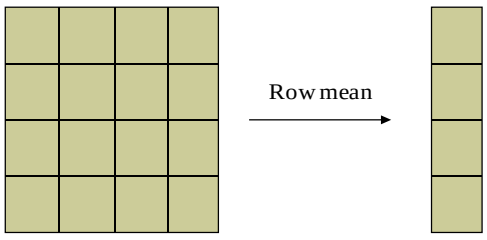

Fig.4: Row Mean of an image

5. EXPERIMENTS

To study the proposed approach we recorded six distinct sentences from 30 speakers: 11 males and 19 females. These sentences are taken from VidTIMIT database [13] and ELSDSR database [14]. For every speaker 10 occurrences of each sentence were recorded. Recording was done at varying times. This forms the closed set for our experiment. From these speech samples spectrograms were created. Before creation of spectrograms, DC offset present in speech samples was removed so that signals are vertically centered at 0. After removal of DC offset, speech samples were normalized with respect to amplitude to -3 dB and also with respect to time. Spectrograms generated from these speech samples form the image database for our experiment. In all we had 1800 spectrograms in our database.

For every speaker 6 spectrograms were used as trainee images and 4 spectrograms were used as test images per sentence, i.e. we had 1080 spectrograms for training purpose and 720 spectrograms for testing purpose. DCT was then applied to the trainee images and result was stored as feature vectors for trainee images.

Similarly, feature vectors for test images were obtained by applying DCT to test images. Euclidean distance between the test image and trainee images was calculated to determine the most probable match i.e. to identify speaker.

Being a text dependent approach, Euclidean distance for a test image of speaker say ‘x’ for a particular sentence say ‘s1’ is obtained by comparing the feature vector of that test image with the feature vectors of all the trainee images corresponding to sentence ‘s1’. Results are calculated for set of test images corresponding to each sentence.

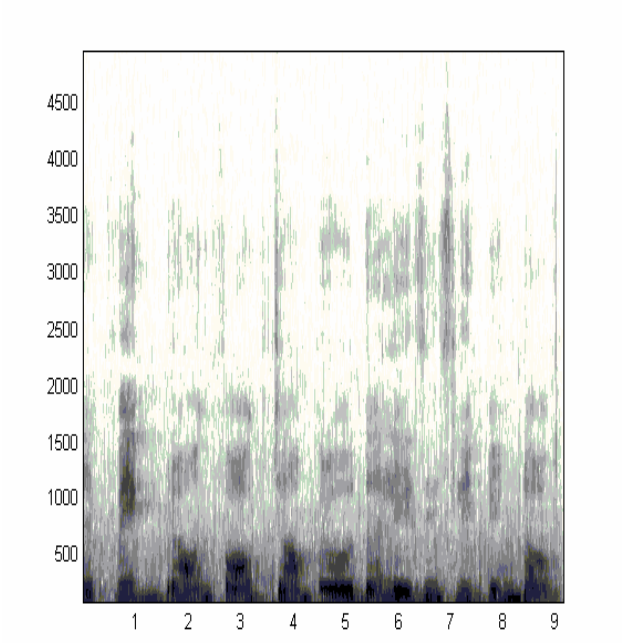

Fig.5: Spectrogram of sentence s1 for speaker 1

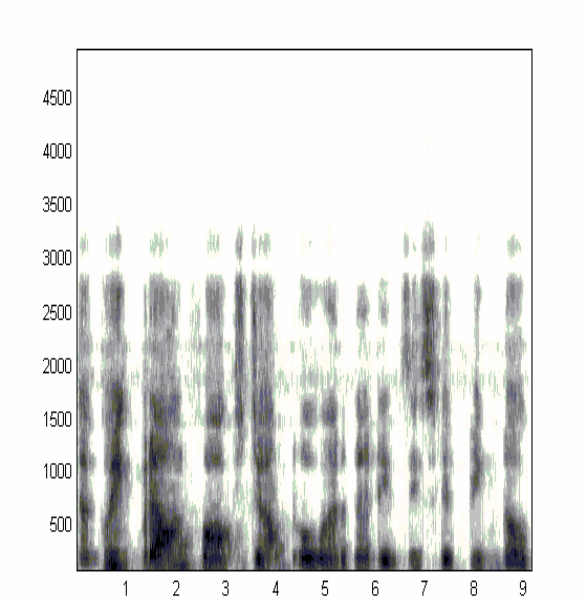

Fig.6: Spectrogram of sentence s1 for speaker 5

Fig.5 and Fig.6 show that the spectrogram for the same sentence, uttered by different speakers is different. The three approaches/methods that were carried out are described in the following subsections.

5.1. Method 1: DCT On Entire Image:

i) As shown in Fig.1, feature vector is obtained by applying DCT on full image.
ii) Euclidean distance between feature vector of test image and trainee image is calculated
iii) Trainee Image with the smallest Euclidean distance is declared as identified speaker.
iv) Steps ii) and iii) are repeated for selected portion of feature vector.

This selection of feature vector is illustrated in following Fig.7. and is based on the number of rows and columns that we selected from the feature vector of an image. For example, we had selected full feature vector (i.e. 256*256), then portion of size 192*192, 128*128, 64*64, 32*32, 25*25, 20*20, 18*18 and 16*16 was selected from the feature vector. For these different sizes, identification rate was obtained.

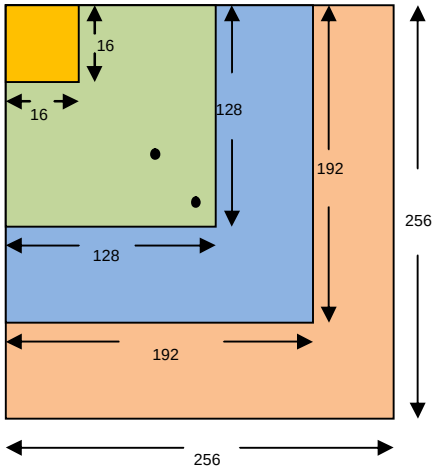

Fig.7: Selection of varying size portion from feature vector

5.2. Method 2: DCT on image block:

i) As shown in Fig.3, feature vector is obtained by taking DCT of image blocks. These image blocks are obtained by dividing image into four parts as shown in Fig.2.
ii) Euclidean distance between feature vector of test image and trainee image is calculated
iii) Trainee Image with the smallest Euclidean distance is declared as identified speaker
iv) Steps ii) and iii) are repeated for selected portion of feature vector.

Selection of feature vector is similar to the one shown in Fig.7. But in this method, size of feature vector is 128*512, 96*384, 64*256, 32*128, 16*64 and 8*32.

5.3. Method 3: DCT on Row Mean of an image:

a) Row mean of full image -

i) Row Mean of an image is obtained.
ii) DCT is applied to this Row Mean to obtain the feature vector.
iii) Euclidean distance between feature vectors of test image and trainee image is calculated.
iv) Trainee image with the smallest Euclidean distance is declared as identified speaker.

b) Row Mean of image blocks -

i) Image is divided into blocks of size 128*128.
ii) Calculate Row Mean of each block.
iii) Apply DCT on Row Mean of each block to form the feature vector of image.
iv) Euclidean distance between feature vector of test image and trainee image is calculated.
v) Trainee image with the smallest Euclidean distance is declared as identified speaker.

Steps ii) to v) in Row Mean of image blocks are repeated for the block size 64, 32, 16 and 8.

6. RESULTS AND COMPLEXITY ANALYSIS

6.1. Results

Following tables show the identification rate obtained for different number of coefficients. These different numbers of coefficients are based on selection of varying sized feature vector portion as shown in Fig. 7.

Table 1 shows the identification rate for sentences s1 to s6 when different numbers of DCT coefficients are taken to find the matching spectrogram i.e. to identify speaker using DCT on full image. Portion from feature vector selected for these coefficients is 256*256, 192*192, 128*128, 64*64, 32*32, 25*25, 20*20, 18*18 and 16*16 respectively.

Portion of feature vector selected	Sentence					
	S1	S2	S3	S4	S5	S6
256*256	54.16	59.16	56.66	56.66	68.33	62.50
192*192	58.33	65	67.5	65	73.33	69.16
128*128	65.83	64.16	71.66	67.5	74.16	72.5
64*64	70.83	70.83	71.66	72.50	77.50	75.83

32*32	75	73.33	74.16	75	80	77.5
25*25	75.83	75	75.83	73.33	80	81.66
20*20	78.33	75.33	78.33	71.66	81.66	80
18*18	75.83	75	77.5	75	80.83	78.33
16*16	72.5	76.66	74.16	74.16	76.66	79.16

Table 1: Identification rate for sentences s1 to s6 for varying portion of feature vector when DCT is applied to full image

Table 2 shows the overall identification rate considering all sentences, for various percentages of DCT coefficients i.e. for portions of different sizes from the feature vector in first approach.

Portion of feature vector selected	Number of DCT coefficients	Identification rate (%)
256*256	65536	60
192*192	36864	66.38
128*128	16384	69.30
64*64	4096	73.19
32*32	1024	75.83
25*25	625	76.94
20*20	400	**77.63**
18*18	324	77.08
16*16	256	76.66

Table 2: Overall Identification rate for varying number of DCT coefficients when DCT is applied to full image

Similarly Table 3 shows the identification rate for sentences s1 to s6 when different numbers of DCT coefficients are taken to identify speaker using DCT on image blocks, whereas, Table 4 shows the overall identification rate considering all sentences, for various number of DCT coefficients using the same approach.

Portion of feature vector selected	Sentence					
	S1	S2	S3	S4	S5	S6
128*512	54.16	59.16	57.5	57.5	68.33	63.33
96*384	60	63.33	65.33	65	73.33	68.33
64*256	65	65	70.83	66.66	74.16	71.16
32*128	70.83	70.83	70.83	71.66	76.66	75
16*64	75.83	74.16	75	75.83	81.66	77.5
8*32	69.16	76.66	75	75.83	75	75.83

Table 3: Identification rate for sentences s1 to s6 for varying portion of feature vector using DCT on image blocks

Portion of feature vector selected	Number of DCT coefficients	Identification rate (%)
128*512	65536	60
96*384	36864	65.97
64*256	16384	68.88

32*128	4096	72.63
16*64	1024	**76.66**
8*32	256	74.58

Table 4: Identification rate for varying size of feature vector portion using DCT on image blocks

Table 5 and Table 6 show the sentence wise identification rate and overall identification rate when DCT of Row Mean is taken by dividing an image into different number of non-overlapping blocks.

No. of blocks for image split	Sentence					
	S1	**S2**	**S3**	**S4**	**S5**	**S6**
Full image (256*256)	57.5	66.66	64.16	60.83	60.83	62.5
4 Blocks (128*128)	60.83	70.83	63.33	65.83	70	65.83
16 Blocks (64*64)	69.16	75.83	70.83	65.83	73.33	71.66
64 Blocks (32*32)	75	76.66	75.83	70	78.83	75.83
256 Blocks (16*16)	76.66	75	75.83	72.5	80	82.5
1024 Blocks (8*8)	74.16	72.5	75	72.5	80.83	78.33

Table 5: Identification rate for sentences s1 to s6 for DCT on Row mean of an image when image is divided into different number of nonoverlapping blocks

No. of blocks for image split	Number of DCT coefficients	Identification rate (%)
Full image (256*256)	256	62.08
4 Blocks (128*128)	512	66.11
16 Blocks (64*64)	1024	71.11
64 Blocks (32*32)	2048	75.27
256 Blocks (16*16)	4096	**77.08**
1024 Blocks (8*8)	8192	75.55

Table 6: Overall Identification rate for DCT on Row mean of an image when image is divided into different number of nonoverlapping blocks

6.2. Complexity Analysis

For 2-D DCT on N*N image, $2N^3$ multiplications are required and $2N^2(N-1)$ additions are required. For 2-D DCT on four blocks of size N/2*N/2, N^3 multiplications are required and $N^2(N-2)$ additions are required. For 1-D DCT on N*1 image, N^2 multiplications are needed and N(N-1) additions are needed. Further for the calculation of Euclidean distance between the feature vectors of size

M*N, number of multiplications required are M*N and number of additions required are 2MN-1. These computational details are summarized in Table 7.

	No. of Multiplications	No. of Additions
2-D DCT on N*N image	$2N^3$	$2N^2(N-1)$
2-D DCT on four blocks of size N/2*N/2 each	N^3	$N^2(N-2)$
1-D DCT on N*1 image	N^2	$N(N-1)$

Table 7: Computational details for 2-D DCT on N*N image, 2-D DCT on N/2*N/2 image and 1-DCT on N*1 image respectively

Considering the above facts, we compare the number of DCT coefficients used and number of computations in terms of multiplications and additions including DCT calculation and Euclidean distance calculation, for the highest identification rate obtained using our three methods. The comparisons are given in Table 8.

Parameter	DCT on Full image	DCT on image blocks	DCT on Row Mean of image
Number of DCT coefficients used	400	1024	4096
Number of multiplications required	33554832	16778240	69632
Number of additions required	33424159	16648191	69631
Identification Rate	77.63	76.66	77.08

Table 8: Number of DCT coefficients used, number of multiplications and number of additions for DCT on full image, DCT on image blocks and DCT on Row Mean of 256*256 image

7. CONCLUSION

In this paper we considered closed set text dependent speaker identification rate using three different ways of applying DCT on spectrograms. For each method, Identification rates obtained for various numbers of DCT coefficients are compared. It has been observed that as the number of DCT coefficients chosen is smaller up to a certain limit; better identification rate is achieved in all three methods. Further it has been observed that DCT on full image gives its maximum identification rate of 77.63% for only 20*20 portion of feature vector i.e. by using only 400 DCT coefficients. DCT on image blocks gives maximum identification rate of 76.66% when 16*64 portion of its feature vector is considered which has 1024 DCT coefficients. Finally DCT on Row Mean gives maximum identification rate of 77.08% for Row Mean of 8*8 size image blocks i.e. for 4096 DCT coefficients.

Further when these maximum identification rates in all three methods are compared, it has been observed that though number of coefficients used in Row Mean method is higher, number of multiplications and additions reduce drastically as compared to other two methods. Number of multiplications in DCT on full image method is 482 times more than the number of multiplications in Row Mean method whereas for DCT on image blocks it is 241 times more. Number of additions needed in DCT on full image and DCT on image blocks are also 480 times and 239

times more than the additions required in Row mean method respectively. Identification rate obtained by Row Mean method is very much closer to the one obtained by applying DCT on full spectrogram and with considerably lesser number of mathematical computations.

8. REFERENCES

[1] Evgeniy Gabrilovich, Alberto D. Berstin: "Speaker recognition: using a vector quantization approach for robust text-independent speaker identification", Technical report DSPG-95-9-001', September 1995.

[2] Tridibesh Dutta, "Text dependent speaker identification based on spectrograms", Proceedings of Image and vision computing, pp. 238-243, New Zealand 2007,.

[3] J.P.Campbell, "Speaker recognition: a tutorial", Proc. IEEE, vol. 85, no. 9, pp. 1437-1462, 1997.

[4] D. O'Shaughnessy, "Speech communications- Man and Machine", New York, IEEE Press, 2nd Ed., pp. 199, pp. 437-458, 2000.

[5] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transaction Acoustics Speech and Signal Processing, vol. 4, pp. 375-366, 1980.

[6] Wang Yutai, Li Bo, Jiang Xiaoqing, Liu Feng, Wang Lihao, "Speaker Recognition Based on Dynamic MFCC Parameters", International Conference on Image Analysis and Signal Processing, pp. 406-409, 2009

[7] Azzam Sleit, Sami Serhan, and Loai Nemir, "A histogram based speaker identification technique", International Conference on ICADIWT, pp. 384-388, May 2008.

[8] B. S. Atal, "Automatic Recognition of speakers from their voices", Proc. IEEE, vol. 64, pp. 460-475, 1976.

[9] Jialong He, Li Liu, and G¨unther Palm, "A discriminative training algorithm for VQ-based speaker Identification", IEEE Transactions on speech and audio processing, vol. 7, No. 3, pp. 353-356, May 1999.

[10] Debadatta Pati, S. R. Mahadeva Prasanna, "Non-Parametric Vector Quantization of Excitation Source Information for Speaker Recognition", IEEE Region 10 Conference, pp. 1-4, Nov. 2008.

[11] Tridibesh Dutta and Gopal K. Basak, "Text dependent speaker identification using similar patterns in spectrograms", PRIP'2007 Proceedings, Volume 1, pp. 87-92, Minsk, 2007.

[12] Andrew B. Watson, "Image compression using the Discrete Cosine Transform", Mathematica journal, 4(1), pp. 81-88, 1994,.

[13] http://www.itee.uq.edu.au/~conrad/vidtimit/

[14] http://www2.imm.dtu.dk/~lf/elsdsr/

[15] H.B.Kekre, Sudeep D. Thepade, "Improving the Performance of Image Retrieval using Partial Coefficients of Transformed Image", International Journal of Information Retrieval (IJIR), Serials Publications, Volume 2, Issue 1, pp. 72-79 (ISSN: 0974-6285), 2009.

[16] H.B.Kekre, Tanuja Sarode, Sudeep D. Thepade, “DCT Applied to Row Mean and Column Vectors in Fingerprint Identification”, In Proceedings of International Conference on Computer Networks and Security (ICCNS), 27-28 Sept. 2008, VIT, Pune.

[17] H.B.Kekre, Sudeep D. Thepade, Archana Athawale, Anant Shah, Prathmesh Verlekar, Suraj Shirke,“Energy Compaction and Image Splitting for Image Retrieval using Kekre Transform over Row and Column Feature Vectors”, International Journal of Computer Science and Network Security (IJCSNS),Volume:10, Number 1, January 2010, (ISSN: 1738-7906) Available at www.IJCSNS.org.

[18] H.B.Kekre, Sudeep D. Thepade, Archana Athawale, Anant Shah, Prathmesh Verlekar, Suraj Shirke, “Performance Evaluation of Image Retrieval using Energy Compaction and Image Tiling over DCT Row Mean and DCT Column Mean”, Springer-International Conference on Contours of Computing Technology (Thinkquest-2010), Babasaheb Gawde Institute of Technology, Mumbai, 13-14 March 2010, The paper will be uploaded on online Springerlink.

[19] H.B.Kekre, Tanuja K. Sarode, Sudeep D. Thepade, Vaishali Suryavanshi, “Improved Texture Feature Based Image Retrieval using Kekre’s Fast Codebook Generation Algorithm”, Springer-International Conference on Contours of Computing Technology (Thinkquest-2010), Babasaheb Gawde Institute of Technology, Mumbai, 13-14 March 2010, The paper will be uploaded on online Springerlink.

[20] H. B. Kekre, Tanuja K. Sarode, Sudeep D. Thepade, “Image Retrieval by Kekre’s Transform Applied on Each Row of Walsh Transformed VQ Codebook”, (Invited), ACM-International Conference and Workshop on Emerging Trends in Technology (ICWET 2010),Thakur College of Engg. And Tech., Mumbai, 26-27 Feb 2010, The paper is invited at ICWET 2010. Also will be uploaded on online ACM Portal.

[21] H. B. Kekre, Tanuja Sarode, Sudeep D. Thepade, “Color-Texture Feature based Image Retrieval using DCT applied on Kekre’s Median Codebook”, International Journal on Imaging (IJI), Volume 2, Number A09, Autumn 2009,pp. 55-65. Available online at www.ceser.res.in/iji.html (ISSN: 0974-0627).

[22] H. B. Kekre, Ms. Tanuja K. Sarode, Sudeep D. Thepade, "Image Retrieval using Color-Texture Features from DCT on VQ Codevectors obtained by Kekre’s Fast Codebook Generation", ICGST-International Journal on Graphics, Vision and Image Processing (GVIP), Volume 9, Issue 5, pp.: 1-8, September 2009. Available online at http://www.icgst.com/gvip/Volume9/Issue5/P1150921752.html.

[23] H. B. Kekre, Sudeep Thepade, Akshay Maloo, “Image Retrieval using Fractional Coefficients of Transformed Image using DCT and Walsh Transform”, International Journal of Engineering Science and Technology, Vol.. 2, No. 4, 2010, 362-371

[24] H. B. Kekre, Sudeep Thepade, Akshay Maloo,”Performance Comparison of Image Retrieval Using Fractional Coefficients of Transformed Image Using DCT, Walsh, Haar and Kekre’s Transform”, CSC-International Journal of Image processing (IJIP), Vol.. 4, No.2, pp.:142-155, May 2010.

[25] H. B. Kekre, Tanuja Sarode “Two Level Vector Quantization Method for Codebook Generation using Kekre’s Proportionate Error Algorithm” , CSC-International Journal of Image Processing, Vol.4, Issue 1, pp.1-10, January-February 2010

[26] H. B. Kekre, Sudeep Thepade, Akshay Maloo, “Eigenvectors of Covariance Matrix using Row Mean and Column Mean Sequences for Face Recognition”, CSC-International Journal of Biometrics and Bioinformatics (IJBB), Volume (4): Issue (2), pp. 42-50, May 2010.

Drug target identification using gene expression microarray data of *Toxoplasma gondii*

Budhayash Gautam budhayashgautam@gmail.com
Department of Computational Biology and Bioinformatics,
Sam Higginbottom Institute of Agricultural, Technology and Sciences,
Allahabad-211007, U.P., India

Pramod Katara katarapramod@gmail.com
Department of Bioscience and Biotechnology,
Banasthali University,
Rajasthan-304022, India

Satendra Singh satendralike@gmail.com
Department of Computational Biology and Bioinformatics,
Sam Higginbottom Institute of Agricultural, Technology and Sciences,
Allahabad-211007, U.P., India

Rohit Farmer rohit.farmer@gmail.com
Department of Computational Biology and Bioinformatics,
Sam Higginbottom Institute of Agricultural, Technology and Sciences,
Allahabad-211007, U.P., India

Abstract

Toxoplasma gondii is an obligate intracellular Apicomplexan parasite that can infect a wide range of warmblooded animals including humans. This pathogen is one of the most common in humans due to many contributing factors that include: (1) its complex life cycle allows it to be transmitted both sexually via felid fecal matter and asexually via carnivorism. (2) *Toxoplasma* has an extremely wide host cell tropism that includes most nucleated cells. (3) In humans and other intermediate hosts, *Toxoplasma* develops into a chronic infection that cannot be eliminated by the host's immune response or by currently used drugs. In most cases, chronic infections are largely asymptomatic unless the host becomes immune compromised and often suffer from life threatening encephalitis. Unfortunately, owing to the toxic side effects and general low efficacy of all known drugs for toxoplasmosis, new chemotherapeutic agents are urgently required. The mechanisms by which *Toxoplasma* grows within its host cell, encysts, and interacts with the host's immune system are important questions. The use of DNA microarrays in transcriptional profiling, genotyping, and epigenetic experiments has impacted our understanding of these processes. In the past years, many existing data analysis methods from other fields have been applied to gene expression data. Among these, clustering methods are a large family of commonly used data analysis methods. Clustering methods are particularly useful in the analysis of gene expression data and organization of gene expression data. In present work Hierarchical and kmeans clustering methods have been used for the analysis of microarray gene expression data of

Toxoplasma gondii with the help of Genesis tool. On the basis of coexpression, some probable gene targets in *Toxoplasma gondii* have been found (IL-16 and Inhibitor of kappa light chain Epsilon etc.). Their promoters have been predicted with Neural Network Promoter Predictor, of BDGP. Predicted promoters were aligned using ClustalW to SAG family members, (which are related to the invasion or surface attachment of the parasite to the host cell) to see common regulation. They showed good amount of sequence similarity with the already known targeted genes of SAG family. These probable targets have important functions in the case of encephalitis in AIDS patient caused by *Toxoplasma gondii*, as these targets play major role in prevention of AIDS.

Keywords: *Toxoplasma gondii*, Toxoplasmosis, Microarray Data Analysis, Hierarchical Clustering, K-means Clustering, Co-expression, Encephalitis.

1. INTRODUCTION

Toxoplasma gondii, represent a major source of human and animal disease worldwide. *T. gondii* is the cause of significant morbidity and mortality in patients who have AIDS (acquired immunodeficiency syndrome) and serious congenital birth defects in both humans and livestock. **[1]** Although infection is usually asymptomatic in healthy individuals, immune-compromised patients often suffer from life-threatening encephalitis. A common requirement for intracellular pathogens is they must scavenge nutrients from their hosts while avoiding innate host defense mechanisms **[2]**. *Toxoplasma* is no different and how it replicates within a host cell has been the focus of intense investigation by several laboratories. Biochemical- and cell-biological-based assays demonstrated that parasites modify host microtubule and intermediate filament organization **[3, 4, 5]**, inhibit host cell apoptosis **[6, 7]**, upregulate pro-inflammatory cytokines **[8, 9, 10, 11]**, and scavenge purine nucleosides, cholesterol, and other nutrients from their host cells **[12, 13]**. To examine the molecular basis for these changes, DNA microarrays have been used to analyze changes in host gene expression following infection **[14, 15, 16]**. These studies indicated that changes in host transcription were extremely widespread. These changes came in at least two distinct waves with the first wave being induced within 2 hours and included a large number of pro-inflammatory response genes **[15]**. The second wave of gene expression included genes that encode proteins that function in a diverse set of cellular processes. Most striking from these studies was the finding that glucose, mevalonate, and iron metabolic genes were upregulated specifically by *Toxoplasma* **[15]**.

An important advantage of DNA microarrays is that they can readily examine multiple time points and conditions **[17].** As a first step, microarrays spotted with the cDNAs used for the bradyzoite EST sequencing project **[18]** were generated and used to compare the transcriptional responses that take place at various time points following induction of differentiation **[19]**. Although these first generation microarrays were spotted with fewer than 650 unique genes, they demonstrated that the microarrays could be used to discover additional bradyzoite-specific genes. Besides gene discovery, DNA microarrays can also be used to map transcriptional pathways. As an example, the transcriptional response of wild-type parasites and bradyzoite differentiation mutants were compared after stimulating the parasites to undergo differentiation. The resulting microarray data demonstrated that the transcriptional pathways induced during development were hierarchal **[20, 21]**. The full complexity associated with differentiation was demonstrated using full-genome *Toxoplasma* microarrays that compared the transcriptional responses of three distinct *Toxoplasma* strains to a drug that induces bradyzoite development **[22].** GRA (dense granule proteins) and SAG (Surface antigens) are known to have some importance in the cavenging of nutrients from their hosts and invasion to their host's cell respectively **[21]**.

As a parasite with a potentially devastating clinical outcome, an important goal of toxoplasmosis research is the development of new drugs and treatments. There are two major reasons that new drugs are needed to treat *Toxoplasma* infections. First, the drugs currently used to treat *Toxoplasma* infections are poorly tolerated, have severe side effects, and cannot act against bradyzoites **[23]**. Second, there are reports that *Toxoplasma* is developing resistance to the current generation of drugs **[24, 25]**. How resistance to these drugs has developed is not known but is critical to understand because it will lead to improved drug design and will increase our understanding of the biological functions of these drug targets. One way to understanding mechanisms of resistance is to compare the transcriptional profiles of wild-type and resistant parasites grown in the absence or presence of the drug. Such studies in bacterial resistance have demonstrated that pathogen responses to antibiotics are multifactorial and complex **[26].** Whether the same will be true in *Toxoplasma* is unclear, but data from these types of experiments will likely impact new anti-*Toxoplasma* drug design.

Possible drug targets can be detected with the help of microarray data analysis **[27, 28]**. Microarray technology can either be used to investigate the functions of genes, or be used in the diagnosis of diseases **[29]**. In the past years, many existing data analysis methods from other fields have been applied to gene expression data; also many novel methods are developed or under developing particularly for gene expression data analysis **[30, 31]**. Among these, clustering methods are a large family of commonly used data analysis methods. Clustering methods are particularly useful in the analysis of gene expression data and organization of gene expression data and the results from clustering can be used as a starting point for further analysis. A natural basis for organizing gene expression data is to group together genes with similar patterns of expression i.e. co-expressed genes. Co-expressed genes may reveal much about coregulatory mechanisms. For example, if a single regulatory system controls two genes, then the genes are expected to be co-expressed. In general there is likely to be a relationship between co-expression and co-regulation. In co-expression analysis we look for previously-uncharacterized genes that mimic the expression patterns of known genes (with or without the types of treatment used in differential expression). The assumption in co-expression is that if the expression of one gene is very similar to the expression of another gene, then it is likely that they are related in their function. Highly similar expression, like highly similar sequence, suggests similar function **[32, 33]**.

Our main aim in the study is to do analysis of the expression pattern of the gene expression microarray data of *Toxoplasma gondii* to find out new probable drug targets with the help of Hierarchical and k-means clustering methods and to do comparison of the results of both methods to confirm the expression pattern. In the next step, the genes co-expressed with the SAG and GRA family members have to identify. Than the promoter sequences of *Toxoplasma gondii* have to be predicted, because there is no prior information about the promoters of *Toxoplasma gondii*. Then sequence alignment has been done to find out the similarity among SAG and GRA family member's promoters and probable target genes promoter's if any.

2. MATERIALS AND METHODS

2.1 Data retrieval and Preprocessing:

Microarray data of *Toxoplasma gondii* was obtained from “The Stanford Microarray Database” (SMD) **[34]**, and all nucleotide sequences were obtained from NCBI (National Centre for Biotechnology Information) **[35]**. After getting microarray data, filtering has to be done and this has been done in the following manner: First, the value of regression correlation was taken to 0.6, this leads to the 4371 genes to pass filters. Then next is deviation filter, this only select genes whose Log (base2) of R/G Normalized Ratio (Mean) is more than 2 standard deviation(s) away from the mean in at least 1 array. This filter removes 3364 genes, leaving 1007 genes. Then only those genes selected who’s Log (base2) of R/G Normalized Ratio (Mean) is absolute value > 2

for at least 1 array(s).This filter removed 431genes, leaving 576.By using genes with > 80% good data ,249 genes were removed and 327 genes were left for further clustering method. Filtering is done using Genesis tool **[36].**

2.2 Clustering:

Hierarchical clustering and k-means clustering were done using Genesis tool. In hierarchical clustering complete linkage clustering method is used. The aim is generally to define clusters that minimize intra-cluster variability while maximizing inter-cluster distances, i.e. finding clusters, in which members are similar to each other, but distant to members of other clusters in terms of gene expression based on the used similarity measurement. Eucledian distance was used to calculate the distance within and among clusters.

2.3 Comparison of expression patterns:

Results obtained by both clustering methods were compared. To validate results clusters were compared to find out similar expression patterns. For a better visibility of the similarity of clusters from different algorithms, a set of clusters is chosen for a comparison.

2.4 Promoter prediction:

Because there is no prior knowledge about the promoters of *Toxoplasma gondii*, promoters have been predicted with Neural Network Promoter Predictor, of BDGP **[37]**. Promoters of the SAG family members were also predicted.

2.5 Promoter analysis:

Promoters of the genes which are present in the same cluster (in which SAG genes are present) are compared by sequence alignment using ClustalW **[38]**.

3. RESULTS AND DISCUSSION

The heat map is generated in the terms of “differential experimental condition”, along with differential expression. Expression image showing 327 genes in sixteen different experiments conducting results from microarray gene expression data of *Toxoplasma gondii*. The color scale ranges from saturated green for log ratios -3.0 and below to saturated red for log ratios 3.0 and above. Each gene is represented by a single row of colored boxes; each column represents an expression value from particular experiment. From these 327 genes, whose expression changed substantially is chosen for cluster analysis using (1) Hierarchical Clustering, (2) k-means. In hierarchical clustering complete linkage clustering method is used. The tree is well distributed and clusters are easily visible by inspecting the ordered expression image. For the *Toxoplasma gondii* dataset, five main clusters marked from one to five and their several sub-clusters were also marked. The clusters have been interactively specified by looking on the patterns in the expression image and then selecting the corresponding sub tree on the left side **(Fig. 1)**.

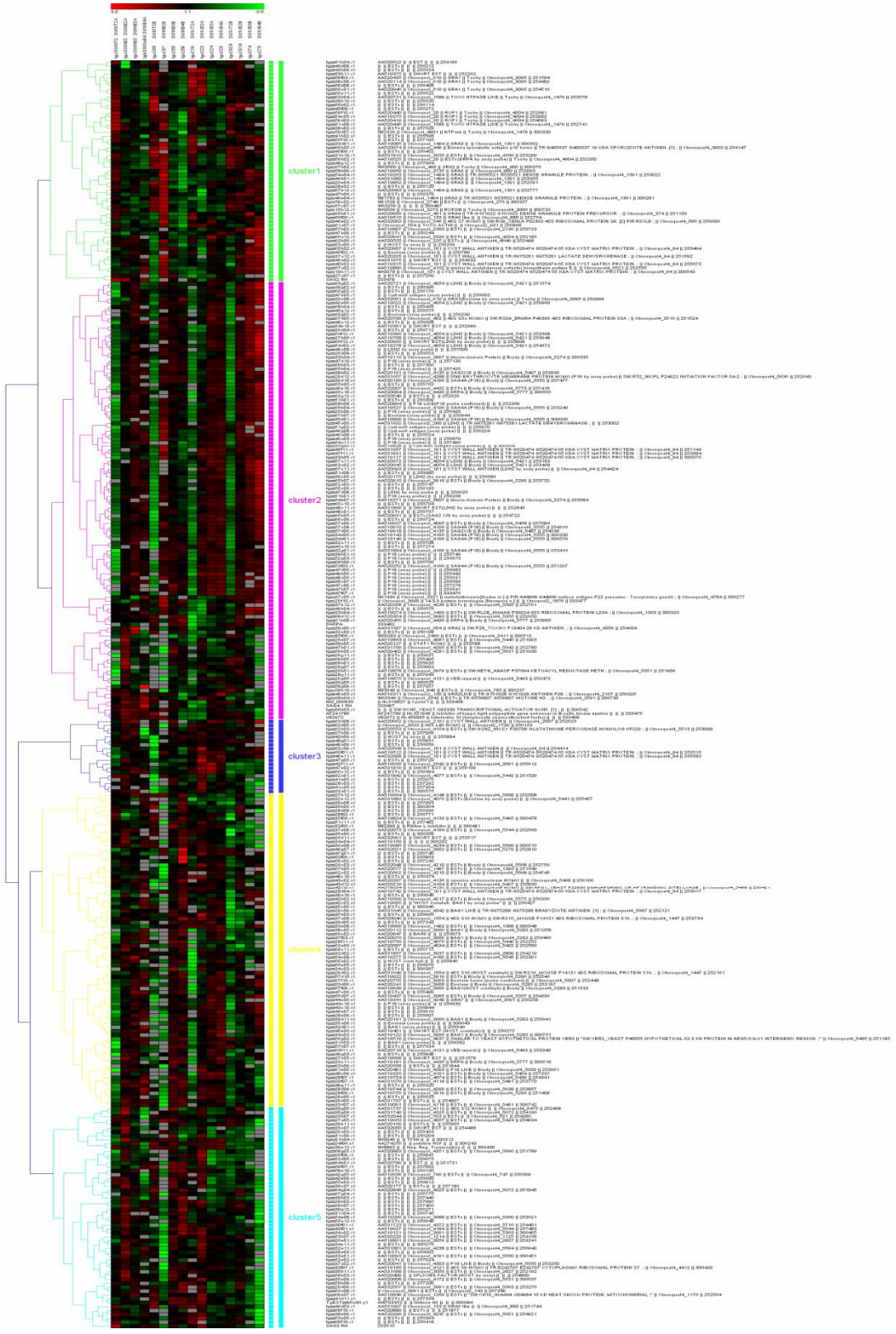

Figure 1: Tree showing Hierarchical clustering using complete linkage. Five main clusters can be clearly seen in the above tree. Each cluster is presented in a different colour to differentiate from other cluster.

By seeing various clusters it can be clearly find out that cluster 2 is the important one for the present work as it has SAG family members **(Fig. 2).**

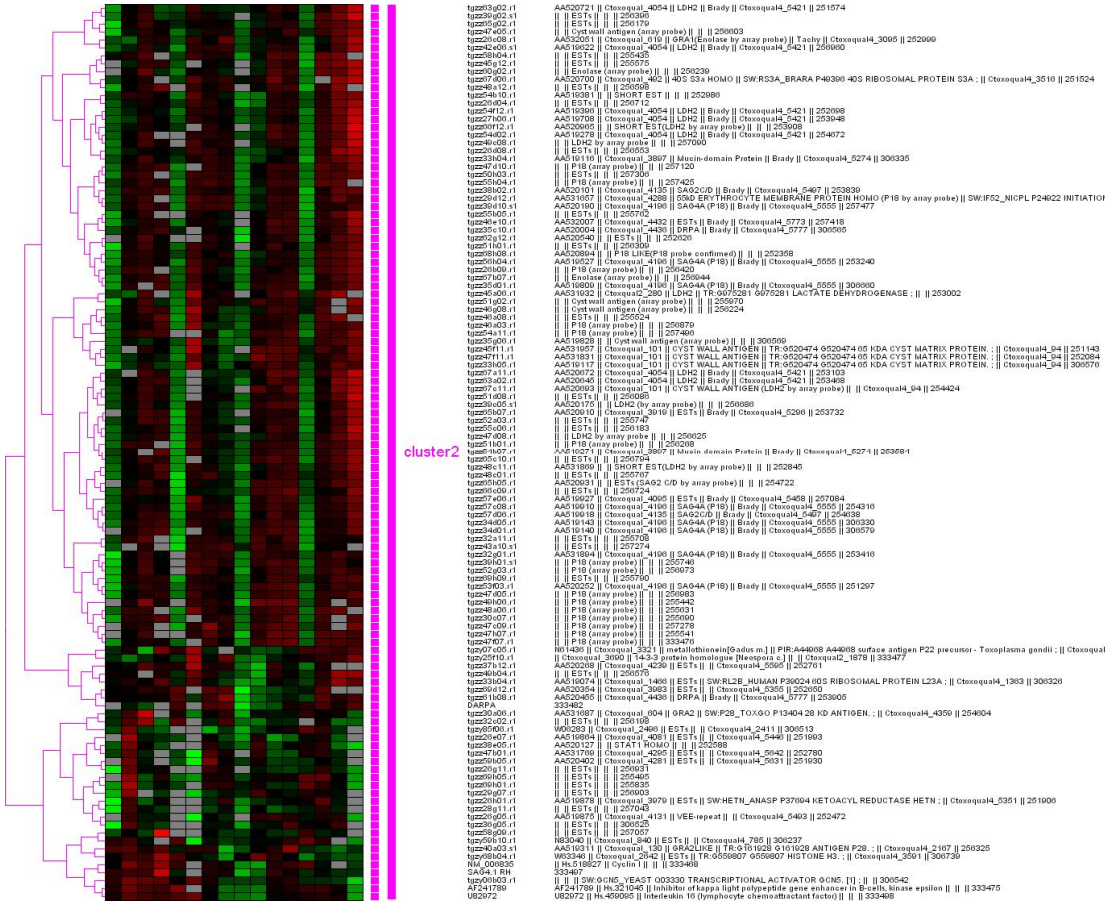

Figure 2 : Expression image of cluster 2. Several subclusters can also be seen in the above image.

In K-means clustering the number of clusters has to be predetermined. Additionally, the maximum number of iteration cycles has to be specified. Usually the algorithm converges (no more reallocations between the clusters) before 50 cycles, but it can occur that the convergence criterion is not reached in a specific time. Therefore, to prevent infinite calculation, the maximum number of cycles has to be declared. For this analysis, k=5 was used for the number of clusters, since 5 basic patterns have been found by hierarchical clustering. The clusters of k-means are very similar to the clusters found with hierarchical clustering. Cluster sizes and gene distribution are of course little different, but both algorithms have found the same basic patterns as the means of the clusters are very similar to each other.

Microarray analysis reveals expression patterns of both previously known and developmentally regulated genes. Stage specific genes e.g. *SRS9,* Ctoxoqual_2199, “mucin domain” protein etc. were clearly identified as such by microarray analysis. Analysis of the microarray data clearly indicates developmental changes in the transcript abundance of *BAG1, SAG4A, LDH2, SAG1* and *NTP1,* which is similar to as published **[19].** Ctoxoqual_819 and Ctoxoqual_4140 were identified as constitutively expressed, as suggested **[19].** Microarray analysis identified several genes whose transcript levels decrease during bradyzoite development. This set includes several previously described genes that were not known to be developmentally regulated: ROP1, ROP2,

ROP4, GRA1, GRA5, GRA8, and MIC1, which is similar to as published **[19].** Constitutively expressed genes include many that encode housekeeping proteins, such as actin and ribosomal proteins. The group of genes that is repressed in bradyzoites includes several that encode proteins that are targeted to the unique secretory organelles of apicomplexan parasites. Of this set, only *NTP*1 has previously been shown to be downregulated during bradyzoite development **[39]**. For the GRA1, GRA5, and ROP1 proteins, previous studies have shown only that these are qualitatively present in bradyzoites. Thus, microarray analysis may be a more sensitive and quantitative method for detecting subtle changes compared to immunofluorescence staining, although changes in transcript and protein abundance will not always parallel each other. The finding that transcript levels of a subset of *GRA* genes decrease during bradyzoite development, while the transcripts for *GRA2, GRA3, GRA4, GRA6,* and *GRA7* are constitutively expressed, shows that selective regulation of *GRA* genes occurs during differentiation, which is similar to as published **[19]**.

The regulated expression of genes encoding rhoptry proteins, if reflected in the protein levels, may explain structural differences in the rhoptries of tachyzoites and bradyzoites. The rhoptries of tachyzoites appear mottled by electron microscopy, while bradyzoite rhoptries are extremely electron dense **[40]**, a phenotype that was also observed in ROP1-knockout tachyzoites. This developmental change could, therefore, be due to decreased expression of ROP1 in bradyzoites. The expression profiles obtained by microarray analysis allowed us to identify distinct classes of temporally regulated genes. Predictions regarding the role of these genes in Toxoplasma development and physiology can be made based on these profiles. A crucial role for the surface proteins SAG2C/D and SAG4A and the unknown protein encoded by the Ctoxoqual_3905 4432 contig is suggested by the early induction and continued expression of these genes at high levels (all-high class), as suggested **[19].**

Microarray data are usually presented as clusters of genes that are grouped based on fold change in transcript levels over time. In such cases, the data reveal changes relative to the starting time or condition, but they do not provide information on whether a given gene's transcripts are abundant or rare relative to other genes. We have employed a technique that determines such relative transcript abundance at each time point, thus allowing distinct expression patterns to emerge from clusters of genes that may otherwise appear to be coordinately regulated using the fold change criteria. Differences in the amount of transcript present for each gene suggest distinct regulatory mechanisms due, for example, to different promoter strengths or different mRNA stabilities. Relative abundance data can also be used to distinguish between "stagespecific" genes (i.e., significant transcript levels in only one stage) and "up- (or down-) regulated" genes (substantial transcript levels in both stages but higher in one than the other).

The identification of distinct classes of regulated genes will make it possible to search for common regulatory sequences **[19]**. For example, the analysis of genes with very high transcript levels and coordinate expression over time may reveal conserved sequence elements in their promoters or untranslated regions, as has been shown for groups of coordinately regulated genes identified by microarray analysis of sporulation in yeast **[41]**. Ongoing *T. gondii* genome sequencing efforts will facilitate such analysis as the sequences needed to identify consensus motifs become available.

To validate clustering results, clusters obtained by both algorithms were compared to find out similar expression patterns, because if different clusters of the above used clustering methods will have similar expression patterns alongwith similar genes then the results will be more useful and reliable. For a better visibility of the similarity of clusters from different algorithms, a set of clusters is chosen for a comparison. Cluster one of hierarchical clustering is similar to cluster five of k-means clustering and cluster two of hierarchical clustering is similar to cluster four of k-means clustering **(Fig. 3 and 4).** Most of the GRA and ROP gene family members are present in the cluster one of hierarchical clustering, are also present in the cluster five of k-means clustering. Similarly, most of the SAG gene family members are present in the cluster two of hierarchical

clustering, are also present in the cluster four of k-means clustering. These results are actually validating our clustering outputs.

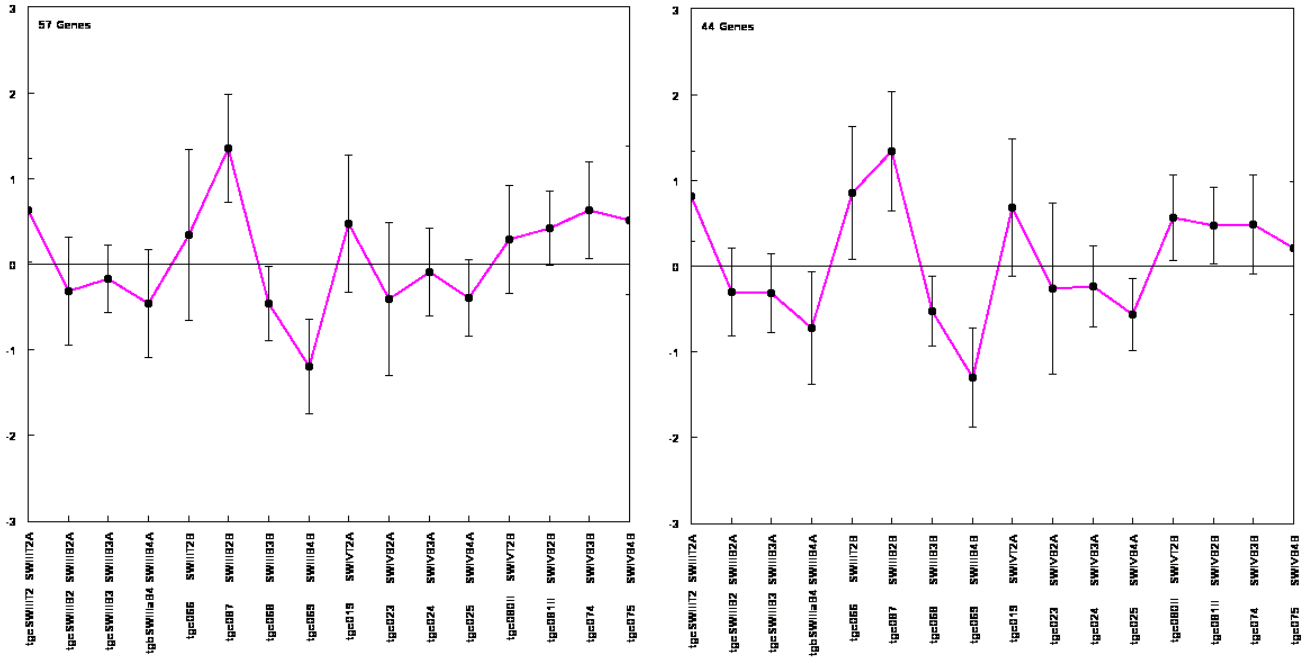

Figure 3 : Expression image of **(A)** cluster 1 of Hierarchical clustering **(B)** cluster 5 of K-means clustering. On the left of the both image expression level is shown into colour ranges from saturated green for log ratio -3 and saturated red for log ratio +3. Both clusters have almost similar expression pattern as both clusters have almost similar average line for expression. Also total numbers of genes are quite similar.

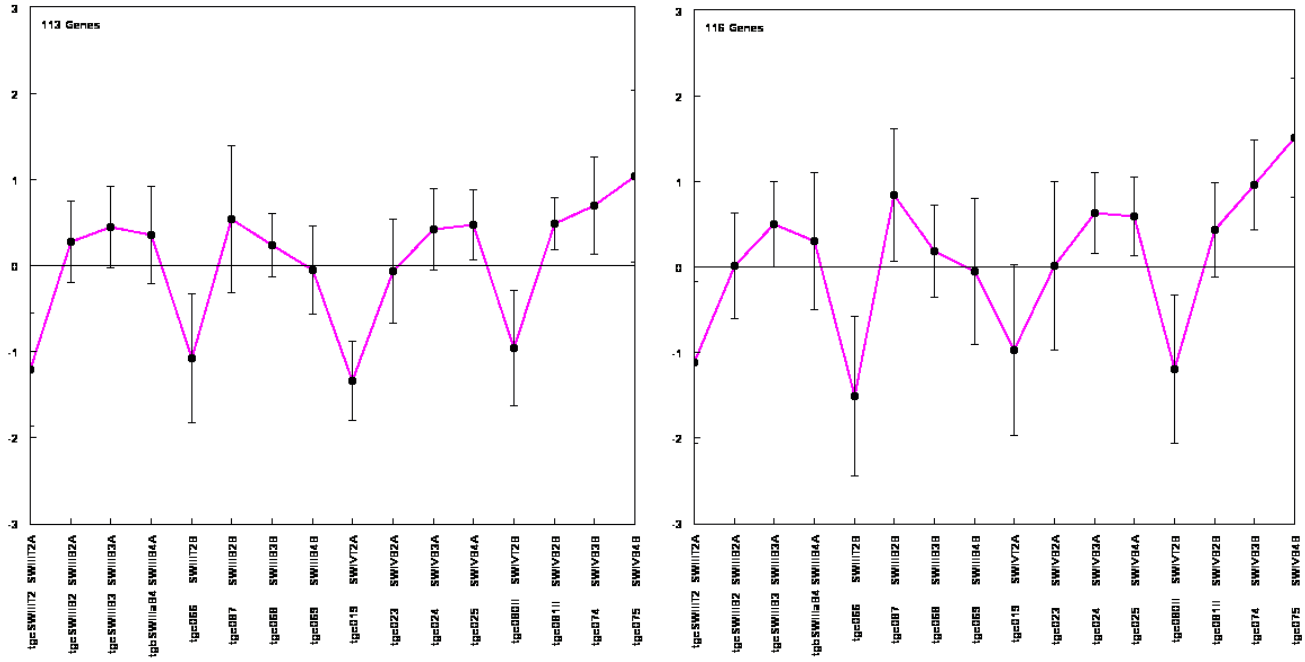

Figure 4 : Expression image of **(A)** cluster 2 of Hierarchical clustering **(B)** cluster 4 of K-means clustering. On the left of the both image expression level is shown into colour ranges from saturated green for log ratio -3 and saturated red for log ratio +3. Both clusters have almost similer expression pattern as both clusters have almost similar average line for expression. Also total numbers of genes are almost similar.

In the promoter analysis, promoters of SAG family and some GRA found in cluster 2 are predicted **(Tab. 1 A)**. Some of the genes in cluster two have been identified as probable targets, these are: U82972 (Interleukin 16), AF241789 (Inhibitor of kappa light chain epsilon), NM_006835cyclin 1(CCNI), W63346 (Histone H3). There promoters are also predicted **(Tab.1 B)**.

Table 1 : Predicted promoters sequences of **(A)** SAG and GRA family members and **(B)** Probable drug targets.

Accession id.	Predicted promoter
(A)	
AA520101 (SAG2C/D)	ACATCTTTTACAAATGTGTCTCACCAGCGTCGGCGCGCGAACAGGTGGGA
AA520190 (SAG4AP18)	GAAGGGGGCAGACCACTGCAACGATGAGCCCGTCGAGCTCGCTGCATTGT
AA519527 (SAG4A(P18))	ATGTGTATCATGCTGCGAACGCATAAAGTACAGTCGAGTGATGCGTGTTA
AA519809 (SAG4A(P18))	TTCTCAAGGTTGAAAAGGGAGACCACTGCAAGGATGAGCCCGTCGAGCTC
AA520931 (SAG2C/D)	GCGAAGAGGAAATAAATGCAGATGTCTTCCACAAAGATGCAACAGCATTG
AA519910 (SAG4A(P18))	TTCTCAAGGTTGAAAAGGGAGACCACTGCAAGGATGAGCCCGTCGAGCTC
AA51991 (SAG2c/D)	ACATCTTTTACAAATGTGTCTCACCAGCGTCGGCGCGCGAGACAGGTGGG
AA519143 (SAG4A(P18))	TTCTCAAGGTTGAAAAGGGAGACCACTGCAAGGATGAGCCCGTCGAGCTC
AA531894 (SAG4A(P18))	ATGTGTATCATGCTGCGAACGCATAAAGTACAGTCGAGTGATGCGTGTTA
AA532051 (GRA1)	CTGCTGTCGCATATGTTTGGGGGGAAATTGCTCGGATATCTTCATTTGGT
AA519311 (GRA2LIKE)	GGCTGCAGATTTGTATAACACAACATGATGTAGCCGCCACGGTTTTTTTT
AA531687 (GRA2)	TGAAGTTCGCTGAAAACGTCGGACAGCACAGTGGGGGGGCGATTATGTTG
(B)	
U82972 (IL-16)	GGGCGAGGGGCTGCACCCACTCTTGTGCCCCAGCAGCCTGAGCAAGTACT
AF241789 (Inhibitor of kappa light chain)	CAGGAGGCCGTGCACAAGCAGACCAGTGTGGCCCCCCGACACCAGGAGTA
NM_006835 (CYCLIN)	CAGCCGTGCGTCCCGCTCGAGCGCCAGCGCCCGCGCCCGCGCCCCCCGAT
W63346 (HISTONE-H3)	TCCATTTTTAAAACGCAAACTCGATTACGCCACCGCTGTCGCACTCTGAC

On comparing these promoters good amount of similarities have been found among these. On analysis of these results we can conclude that there is a good amount of seuqence similarity among the promoters of (1) AA520101_SAG2C/D and U82972_IL-16, AF241789 _Inhibitor of kappa light Chain Epsilon, AA532051_GRA1, AA531687_GRA2, W63346_HISTONE-H3, NM_006835_CYCLIN. (2) AA519809_SAG4A_P18 and U82972_IL-16, AF241789_Inhibitor of kappa light Chain, AA532051_GRA1, W63346_HISTONE-H3, & NM_006835_CYCLIN. But, found low similarity between AA519809_SAG4A_P18 with A531687GRA2. Results are showing similarities in the SAG family with IL-16 and Inhibitor of kappa light Chain Epsilon genes expression pattern. IL-16 gene is an important genes. IL-16 is a multi-functional cytokine that uses CD4 as a receptor to signal diverse biological activities by target cells including T-lymphocytes, monocytes and eosinophils. This cytokine is a functionally significant endogenous antiviral factor. The antiviral activity of IL-16 may be of therapeutic benefit in HIV/AIDS, but its greatest potential is for immune reconstitution. Stimulation of CD4+ T-cells with IL-16 primes cells to respond to IL-2, by up-regulating the expression of IL-2 receptor p75 (CD25). Co-treatment of peripheral blood mononuclear cells (PBMC) with IL-16 plus IL-2 (or IL-15) in vitro selectively expands the population of CD4+ T-cells. In combination with IL-16, the beneficial effects of IL-2 may be augmented and specifically targeted to CD4+ T-cells. Thus, IL-16 shows considerable promise as an agent for the biological therapy of HIV/AIDS and other infectious diseases like toxoplasmosis. Thus IL-16 gene can be a probable drug target in case of the toxoplasmosis. Inhibitor of kappa light Chain Epsilon can also be a probable drug target as it has important functions in immune system. As it is known that, NK and T-cell-derived IFNγ is the critical cytokine in protection against infections with all *Toxoplasma* strains **[42]**. This cytokine protects against *Toxoplasma* infections by upregulating the expression of inducible nitric oxide synthase,

indoleamine dioxygenase, and a family of IFNγ-regulated GTPases that degrade the parasitophorous vacuole **[43]**. Regardless of its effectiveness, some parasites can evade IFNγ-mediated killing and develop into bradyzoites. One possible mechanism by which the parasite avoids IFNγ is to disable IFNγ-induced signaling. Indeed, microarray and cell biological assays demonstrated that IFNγ-induced transcription is abrogated in cells previously infected with *Toxoplasma* **[44, 45]**. In contrast to the polymorphic ROP16 and ROP18 virulence factors, *Toxoplasma*'s effects on IFNγ-dependent transcription are strain independent **[43]**. The mechanism underlying parasite abrogation of IFNγ-stimulated transcription is still unclear but does not appear to involve blocking nuclear localization of STAT1, which is a key IFNγ-regulated transcription factor **[46]**.

4. CONCLUSIONS

Despite enormous development in the technology of gene expression analysis, there are different aspects of the microarray data analyis of the *T. Gondii,* that have been poorly addressed or very less known. Recent evidences strongly improve our understanding about this organism **[46].** Our study proposes some probable drug targets for the treatment of toxoplamosis. As these genes (IL-16 and Inhibitor of kappa light chain) have very important role in immune system in protecting body from any antigenic activities thus, these genes can act as possible drug targets for the treatment of toxoplasmosis. But further study is needed to develop more useful non-resistant drugs for toxoplasmosis.

5. Acknowledgments

The authors would like to acknowledge the support and facilities provided by the Department of Computational Biology and Bioinformatics, Sam Higginbottom Institute of Agricultural, Technology and Sciences, Allahabad, U.P., India.

6. REFERENCES

[1] J. P. Dubey, D. S. Lindsay and C.A.Speer. "*Structures of Toxoplasma gondii tachyzoites, bradyzoites, and sporozoites and biology and development of tissue cysts".*Clinical Microbiology Review, 11, 267-299,(1998).

[2] A. P. Sinai and K. A.Joiner. *"Safe haven: the cell biology of nonfusogenic pathogen vacuoles"*. Annual Review of Microbiology.51:415–62, (1997).

[3] I. Coppens. "*Toxoplasma gondii sequesters lysosomes from mammalian hosts in the vacuolar space"*. Cell. 125(2):261–74 (2006).

[4] M. E. Walker. *"Toxoplasma gondii actively remodels the microtubule network in host cells".* Microbes and Infection. 10(14–15):1440–9, (2008).

[5] S. K. Halonen and E. Weidner. "*Overcoating of Toxoplasma parasitophorous vacuoles with host cell vimentin type intermediate filaments"*. Journal of Eukaryotic Microbiology. 41(1):65–71 (1994).

[6] P. B. Nash. "*Toxoplasma gondii-infected cells are resistant to multiple inducers of apoptosis"*. Journal of Immunology. 160(4):1824–30, (1998).

[7] S. Goebel, C. G. Luder and U. Gross. "*Invasion by Toxoplasma gondii protects human-derived HL-60 cells from actinomycin Dinduced apoptosis"*. Medical Microbiology and Immunology. 187(4):221–6, (1999).

[8] Z. Y. Li. *"Toxoplasma gondii soluble antigen induces a subset of lipopolysaccharide-inducible genes and tyrosine phosphoproteins in peritoneal macrophages".* Infection and Immunity 62(8):3434–40, (1994).

[9] M. P. Brenier-Pinchart. "*Toxoplasma gondii induces the secretion of monocyte chemotactic protein-1 in human fibroblasts, in vitro".* Molecular and Cellular Biochemistry, 209(1–2):79–87, (2000)..

[10] G. S. Yap and A. Sher *"Cell-mediated immunity to Toxoplasma gondii: initiation, regulation and effector function"*. Immunobiology. 201(2):240–7,(1999).
[11] C. F. Denney, L. Eckmann. and S. L. Reed. "*Chemokine secretion of human cells in response to Toxoplasma gondii infection".* Infection and Immunity 67(4):1547–52, (1999).
[12] J. D. Schwartzman and E. R. Pfefferkorn. "*Toxoplasma gondii: purine synthesis and salvage in mutant host cells and parasites".* Experimental Parasitology. 53(1):77–86, (1982).
[13] I. Coppens, A. P. Sinai. and K. A. Joiner. *"Toxoplasma gondii exploits host low-density lipoprotein receptor-mediated endocytosis for cholesterol acquisition".* Journal Cell Biology. 149(1):167–80 (2000).
[14] M. Gail, U. Gross and W. Bohne. "*Transcriptional profile of Toxoplasma gondii-infected human fibroblasts as revealed by gene-array hybridization".* Molecular Genetics and Genomics. 265(5):905–12, (2001).
[15] I. J. Blader, I. D. Manger and J. C. Boothroyd. *"Microarray analysis reveals previously unknown changes in Toxoplasma gondii-infected human cells".* Journal of Biological Chemistry. 276(26):24223–31 (2001).
[16] D. Chaussabel. *"Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites".* Blood. 102(2):672–81, (2003).
[17] J. C. Boothroyd. *"DNA microarrays in parasitology: strengths and limitations".* Trends in Parasitology. 19(10):470–6, (2003).
[18] I. D. Manger. "*Expressed sequence tag analysis of the bradyzoite stage of Toxoplasma gondii: identification of developmentally regulated genes".* Infections and Immunity. 66(4):1632–7, (1998).
[19] M. D. Cleary. "*Toxoplasma gondii asexual development: identification of developmentally regulated genes and distinct patterns of gene expression".* Eukaryotic Cell. 1(3):329–40, (2002). [20] M. Matrajt. *"Identification and characterization of differentiation mutants in the protozoan parasite Toxoplasma gondii".* Molecular Microbiology.44(3):735–47, (2002).
[21] U. Singh, J. L. Brewer and J. C. Boothroyd. "*Genetic analysis of tachyzoite to bradyzoite differentiation mutants in Toxoplasma gondii reveals a hierarchy of gene induction".* Molecular Microbiology.44(3):721–33, (2002).
[22] M. S. Behnke. "*The transcription of bradyzoite genes in Toxoplasma gondii is controlled by autonomous promoter elements".* Molecular Microbiology.68(6):1502–18, (2008).
[23] B. Dannemann. "*Treatment of toxoplasmic encephalitis in patients with AIDS. A randomized trial comparing pyrimethamine plus clindamycin to pyrimethamine plus sulfadiazine. The California Collaborative Treatment Group".* Annals of Internal Medicine. 16(1):33–43, (1992).
[24] H. Baatz. "*Reactivation of toxoplasma retinochoroiditis under atovaquone therapy in an immunocompetent patient".* Ocular Immunology & Inflammation. 14(3):185–7, (2006).
[25] T. V. Aspinall. "*The molecular basis of sulfonamide resistance in Toxoplasma gondii and implications for the clinical management of toxoplasmosis".* Journal of Infectious Disease. 185(11):1637–43, (2002).
[26] M. D. Brazas and R. E. Hancock. *"Using microarray gene signatures to elucidate mechanisms of antibiotic action and resistance".* Drug Discovery Today. 10(18):1245–52, (2005).
[27] B. Luft. and J. S. Remington. *"AIDS commentary: toxoplasmic encephalitis in AIDS".* Clinical Infectious Diseases. 15:211–222. (1992).
[28] D. E. Jr Bassett, M. B. Eisen and M. S. Boguski. "*Gene expression informatics--it's all in your mine".* Nature Genetics. Jan;21(1 Suppl):51-5, (1999).
[29] Debouck, C. and Goodfellow, P. N. (1999). DNA microarrays in drug discovery and development. *Nat* Genet. 21(1): 48–50 [PMID: 9915501]
[30] A. Brazma, A. Robinson, G. Cameron and M. Ashburner. *"One-stop shop for microarray data".* Nature. Feb 17;403(6771):699-700, (2000).
[31] A. Brazma and J. Vilo. *"Gene expression data analysis".* FEBS Letters, Aug 25; 480(1):17-24 (2000).
[32] M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. **8**;95(25):14863-8.[PMID: 9843981]

[33] Heyer, L. J., Kruglyak, S. and Yooseph, S. (1999). "Exploring expression data: identification and analysis of coexpressed genes". Genome Research. Nov;9(11):1106-15, (1998).
[34] http://genome-www5.stanford.edu/
[35] http://ncbi.nlm.nih.gov/
[36] A. Sturn, J. Quackenbush, and Z. Trajanoski. *"Genesis: cluster analysis of microarray data".* Bioinformatics. 18: 207-208. (2002).
[37] A. H. Waibel *et al. "IEEE Transactions on Acoustic, Speech, and Signal Processing".* 37(3):328-339, (1989).
[38] http://www.ebi.ac.uk/Tools/clustalw2/index.html.
[39] V., D. Nakaar, K. R. Bermudes and K. A. Joiner. *"Upstream elements required for expression of nucleoside triphosphate hydrolase genes of Toxoplasma gondii".* Molecular and Biochemical Parasitology. 92:229–239, (1998).
[40] D. J. Ferguson and W. M. Hutchison. *"An ultrastructural study of the early development and tissue cyst formation of Toxoplasma gondii in the brains of mice.*" Parasitology Research. 73:483–491 (1987).
[41] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown and I. Herskowitz. *"The transcriptional program of sporulation in budding yeast".* Science 282:699–705, (1998).
[42] P. J. Gaddi and G. S. Yap. *"Cytokine regulation of immunopathology in toxoplasmosis".* Immunology & Cell Biology. 85(2):155–9, (2007).
[43] I. J. Blader and J. P. Saeij. *"Communication between Toxoplasma gondii and its host: impact on parasite growth, development, immune evasion, and virulence".* APMIS. 117(5–6):458–76 (2009).
[44] S. K. Kim, A. E. Fouts and J. C. Boothroyd. *"Toxoplasma gondii dysregulates IFN-gamma-inducible gene expression in human fibroblasts: insights from a genome-wide transcriptional profiling".* Journal of Immunology. 178(8):5154–65. (2007).
[45] C. G. Luder. *"Toxoplasma gondii down-regulates MHC class II gene expression andantigen presentation by murine macrophages via interference with nuclear translocation of STAT1alpha".* European Journal of Immunology. 31(5):1475–84, (2001).
[46] K. M. Brown and J. Blader. *"The role of DNA microarrays in Toxoplasma gondii reserch the causative agent of ocular toxoplasmosis".* Journal of Ocular biology disease & Informatics; 2:214-222, (2009).

Protecting Identity Using Biometrics Protection Systems

Fathimath Sabena sabenaadam@hotmail.com
University College of Technology and Innovation (UCTI)
Kuala Lumpor, 5700, Malaysia

Ali Dehghantanha ali_dehqan@ucti.edu.my
University College of Technology and Innovation (UCTI)
Kuala Lumpor, 5700, Malaysia

Andy Seddon andy@ucti.edu.my
University College of Technology and Innovation (UCTI)
Kuala Lumpor, 5700, Malaysia

Abstract

Biometrics Identity Management (BIdM) is a newly rising and developing discipline which could be expressed as the study of verification and validation methods for the next generation. The two key terms enclosed in the title of this paper are– "Biometrics Vulnerabilities" and "Identity Management". Every one of us has an identity. By utilizing this identity along with distinctive characteristics we distinguish ourselves from one another. By cross referencing the data from both sources, a guideline that would adapt the best practices to maintain the sequence of BIdM and identity theft integrity was designed. Based on the findings a guideline is proposed to the experts and end-users to use. A walk through with the BIdM consultant was done to identify areas of improvement to fine tune the artifact.

For proper identity management this guideline can be used as the processes in data collection and data maintenance procedures are included. The procedures include extracting the data from data collection for proofs, data matching and handling the data in an appropriate way. The guideline will have its proper BIdM techniques by having the best practices of tackling its vulnerabilities.

Databases having biometric data are themselves a threat to privacy. While distinguishing gaps in BIdM and discovering new approaches to tackle the vulnerabilities, issues and protect such databases and increasing the awareness programs, this research can be further extended.

Keywords: Biometrics, Identity Management, Biometric Systems, Biometric Technologies, Technical Vulnerabilities, Social Vulnerabilities, Privacy vulnerabilities, Identity Fraud, Review of Biometrics Vulnerabilities.

1. INTRODUCTION

Biometric technology can be described as the computerized detection of a person's behavioral and bodily attributes. Physically individuals to records of identity, creating a one-to-one communication among records and people, it can associate limiting individuals to single record or

single person to proceedings. While identity administration, although biometric technology is a usual tool, many believe that it intrudes privacy. According to the Computer Crime survey 2006, published by Computer Security Institute and the FBI, *"Unauthorized access continues to be the second greatest source of financial loss*" [1].

Biometric information is a sensitive form of data. It may have a much larger effect on the individual in certain ways if it gets infringed or stolen. The data is unique and of high quality. It concerns with the person's physical characteristics hence it should be treated with the utmost importance to make sure it is firmly saved, encrypted and only reachable to authorized individuals and when no more required destroyed under the intentions in which it was gathered.

For improving privacy and guaranteeing a one-to-one communication amid people and records, biometric technologies have been suggested as a natural tool in identity management procedures. But regarding their values some commentators raised questions, is biometrics an effective tool for regulating individuals?

Monitoring of an identification technique of biometric contribution, the user's biometric transparently may not be agreeable for the user, or gathering biometric templates in a crucial database since a biometric of user's could be utilized for offensive reasons if the biometric is acquired by an unofficial person. Users biometric can give information which a user willingly may not want to provide. For example, reasons for law enforcement, a fingerprint interpretation can be utilized while medical information could be obtained from an eye scan.

Upon which a biometric technique can be used, there are eight points, as pointed out by [2]. Although the greatest publicity was received for fake biometric attack, all further usage needed some kind of entrance to the biometric techniques and methods possibly signifying additional serious threat.

Cost is constantly a vital issue when implementing a new technology. Regarding the cost of a biometric system people frequently concentrate specially on the cost of hardware, sensor and linked software, although the definite cost of the fundamental elements usually goes much further while executing any biometric system. Furthermore, there will be extra operating expenses linked with administration, integration, user education, installation, and system maintenance and data collection.

This paper will explore identity management. It will specifically address biometrics vulnerabilities, why we should care about biometrics vulnerabilities with identity management, what are the best options which are available to solve the vulnerabilities, what are the solutions and why some are more effective than others.

2. REVIEWING of RELATED WORKS

BIdM system must abide with Protection of Data and by means of the Convention on Human Rights in Europe. Data is copied by cloning the Hard Disk Drive. The interviewee's pointed out that the vulnerabilities and issues controlled should be protected according to their respective countries data security laws. The Data Protection Act controls the means of organizations dealing with data that recognizes users. In BIdM technologies situations, transparency and proportionality are both standards overriding in each situation which will pertain. Transparency or clearness denotes putting it to understand how and why data is utilized and with no preceding agreement not moving ahead. Proportionality necessitates the use of individual's biometric data, or an individual's private life is not interfered, with the advantages of the system should be reasonable. It commonly denotes matching organization's privileges or public at large with the rights of the individual. In legal terms personal data is more important than any other data BIdM "private" information its dealing with.

By closing off, admission to the system is vital in preventing the suspect from sabotaging the system. When the data gathering is conducted, the intruder's log on to the system should be completely blocked. Through data gathering phase, the most skillfully required items are logs as they enclose information about the intruder's actions. Moreover, to make certain no data is written onto the intruder's computer the BIdM administrator's team requires to directly pulling off the computer plug from the wall socket. Additionally, this will preserve the files such as temporary files and swap files. The data has to be kept for further planning (how it was performed, should it

be reserved as a data security and handling exercise and the hard disk drive required to be cloned).
The data duplicated by the inspector is same as per the intruder's original hard disk drive is ensured by the imperative factor called Hashing. The precise copy of the hard disk should be done ensuring all the information duplicated is same as the original when duplicating the intruder's hard disk. To ensure that the data is not changed it should be done at the intruder's location. The images must not be general copies. It should be genuine bit by bit or close images of the originals.
Inspectors may at times accidentally delete the proof critical for the investigation whilst recuperating proof from the system. By applying the write blocker technique this can be prevented from the intruder's workstation writing or deleting proof. Extreme consideration should be given when extracting data from the intruder's workstation as it is a very important aspect during data gathering.
The items collected should be labeled and documented in order to keep those as sequence of custody. These will come to assistance during the management in internal, civil and criminal inspection. During the reporting stage where the inspector requires rebuilding the BIdM based on the proof in these documentations will be of benefit.
All magnetic data is collected in antistatic wrapping, for example, paper bags, no scratching or folding of computer media is allowed, etc. To prevent loss, demolition of data, modification and physical harm all proof is labeled. Therefore, the guideline distinguishes that computer equipment are "fragile" electronic devices that are receptive to physical shock, static electricity and magnetic sources, temperature, humidity in wrapping and while transportation of proof. Additionally, during transportation, all proof should be kept at a distance from magnetic sources.
In common the interviewees revealed the information maintenance methods for BIdM from different policies and standards. It should be stored in a safe and secured place where intruders cannot reach and only authorized persons can access the room to prevent the proof being tempered. The people who access the storeroom should be logged.
The use of BIdM software to analyze the data to prevent accidental tampering of the data is recommended by the interviewees. To analyze the information the interviewees agree that a secure place or a laboratory should be utilized. Utilizing bootable diskette Hard disk drive is investigated to avoid the operating system. The lab that is secured and regimented to perform the analysis work is suggested. From all the computer systems such as applications, network and log files more proof are searched for. Selection of the right tool to acquire the data will secure the proof.
The interviewees agreed that BIdM activities are recorded proof to the legal and domestic inquiry teams and reporting is vital. The reports should be free from technical jargons and should be in a straightforward way to understand in any language it may present to the authorized concerned people. Also the report will enable to tackle similar problems in the future.
Based on the survey questions the main areas using the BIdM technologies in the three countries are government, health, financial, travel, consumer market and private areas. Among these areas government area embodies the leading BIdM technologies for end users. BIdM technologies are used in government organizations, offices, departments, federal agencies and military areas.

3. PROPOSED MODEL

In this research much more will be discussed on how and what should be done when an intruder tries to attack the system, what are the vulnerabilities detected, classified, handled and the procedures used for handling such cases. As the topic is very wide, data protection aspects will be highlighted in this research.
Nevertheless, to sustain these principles in practice can be challenging. For instance, lack of reliability in BIdM systems is one of the vulnerabilities either from their vulnerability to interference or from error. It is not always "accurate", is the second vulnerability of BIdM, for instance, as additional data accrue in huge databases not only of comparative consistency, but also the fingerprint checks are credible to reduce over time. "Function creep" is a vulnerability to BIdM systems. At the time the data was gathered data is utilized in a way not permitted nor foreseen.

The BIdM guideline will form part of the incident response plan that deals with any types of BIdM related incidents as shown in Figure 1.

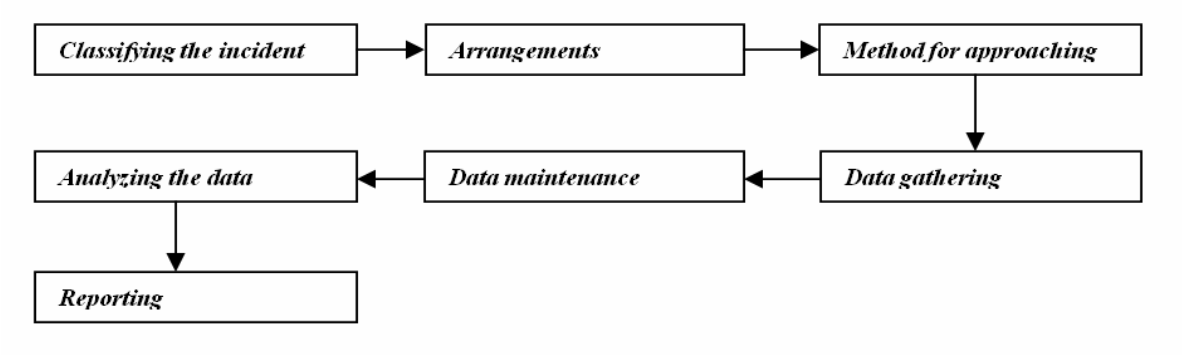

FIGURE 1: An Overview of the BIdM guideline

The three parties involved in this guideline abide by the computer law in their respective countries on computer crimes. These organizations have their own policies on BIdM and computer crime. BIdM will normally execute the monitoring, identification and preliminary evaluation of the vulnerabilities, issues and data maintenance as shown in Figure 2. When a concern is raised the issue will be taken to the top management and to perform the processes in the BIdM it will be treated by the help of the guideline.

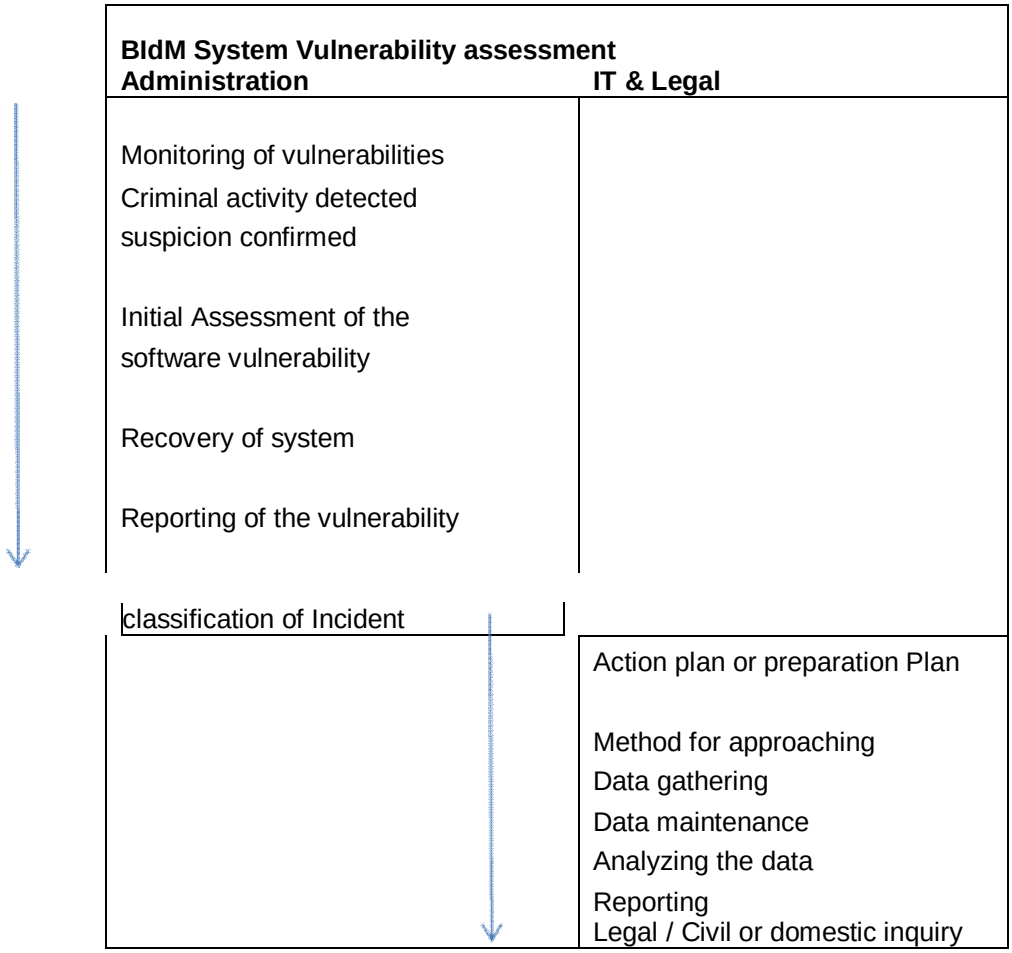

FIGURE 2: BIdM Process paths

The organization's structure will also be maintained by the BIdM team as that will provide an indication of the association of other personnel and sections that will be supporting through the BIdM vulnerabilities, issues controlling and administration with concerned computer crimes as shown in Figure 3.
The sections will be Information Technology (IT), Human Resource (HR) and Legal department. Linking with third party will be carried out by the legal department such as the regulator of legal matters or law enforcement agencies. It indicates that IT personnel can support gathering and investigating of data of proof as they possess expertise in the matters, supported on the chart for BIdM including experts in technical, security administration, matters with administration.

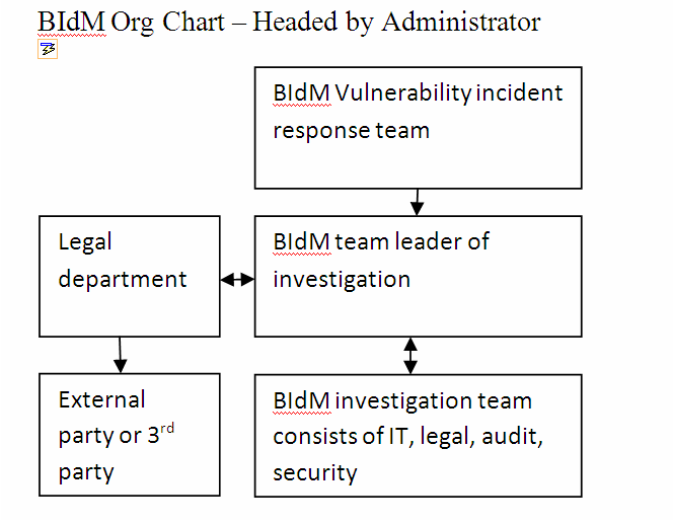

FIGURE 4: BIdM Org Chart

The above phases are considered from the interviews and from the literature review conducted. BIdM is a key global facilitator for managing digital identities while establishing trust and protecting personal information. BIdM is a hub of infrastructure protection capability and continuously evolving and growing cyber security. Operating networks comprises of controlling access to a network or service, complying with local legal and regulatory requirements, and performing online e-transactions.
There are many activities that relate to BIdM standardization as BIdM is becoming an important issue throughout the world as shown in Figure 5.

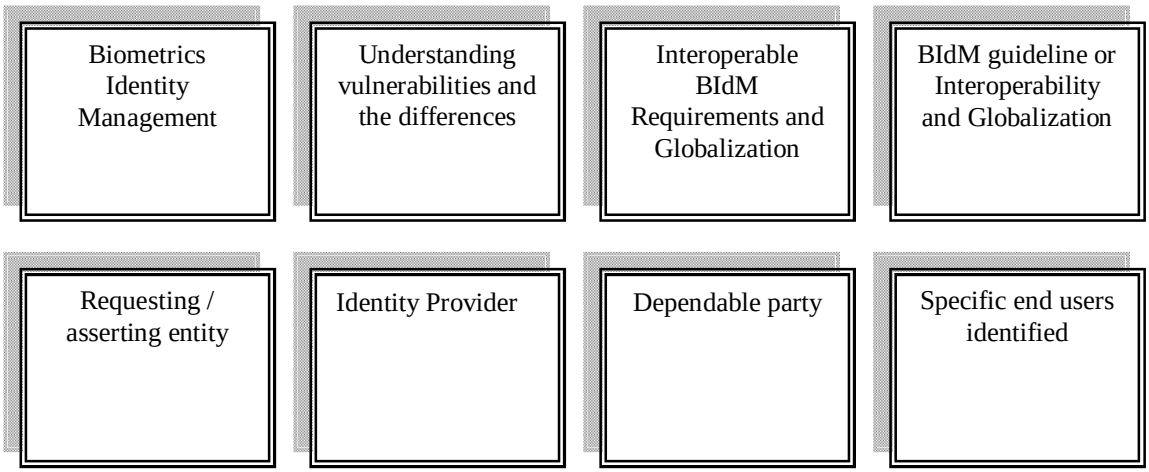

FIGURE 5: BIdM guideline for global interoperability

At present there are BIdM competence and resources across the globe often vague private and public sector limitations which employ and manifest enormous global arrangement of communications networks, Information and Communications Technology systems and services. Among various environments the demanding requirements have resulted in a rapid expansion of BIdM proprietary standardization work and solutions.
Nevertheless, if the discussions, standards, recommendations, guidance etc., are to be comprehensible, easily understood and unambiguous, it is imperative that the terms are described as accurately as possible and a mutual terminology is utilized.
At a higher level the vulnerabilities expressed in the BIdM cause and effect diagram, failure due to an intrinsic and adversary attack are the two methods of failure in a biometric technique.
In adversary attacks for personal benefits a capable intruder or perhaps a group arranged tries to get around BIdM technique. Adversary attacks can be further classified into three types established on factors such as system administration, biometric obviousness, vulnerable infrastructure that permit to negotiate security of the system with an adversary.
Intrinsic failures take place due to the limited discrimination of the specific biometric trait, corresponding technologies or trait extraction as well as intrinsic limitations in the sensing.
The analysis and recommendations are determined on the findings from the real life experiences, interviews, questionnaires and the survey conducted.
The administrator will commence working cautiously with the IT personnel to figure out the vulnerability when BIdM system vulnerability is brought up to him. Subsequently to perform the event classifying phase he will next inform the management and work with them. The administrator will discuss with human resource, legal department and internal audit accordingly when conducting a vulnerability evaluation involving a computer crime. The departments will provide extensive assistance to the system administrator when the vulnerability is classified major or minor. Urgent attention is required if it is a major vulnerability while a minor one needs to be concentrated on basis of time frame prearranged. The administrator can officially commence the mission of inspecting the vulnerability assessment phase called the “arrangements phase” once the event has been classified.
The administrator must make sure before the investigation commences that he has the appropriate tools and software that will be required for the following steps that the BIdM guideline indicates such as during the stages of data gathering and data analysis.
The administrator has to make certain that through the data gathering stage, the application of necessary medium to obtain the information is handy. For maintaining the sequence of legal measures, administrator also requires having notebooks, sequence of custody forms, tags, labels and logs.
As stated by the predefined policies all media utilized should be handled cautiously.
According to Privacy Impact Assessment for the DHS / UK visas Project, as the data is saved at U.S. Citizenship and Immigration Services (USCIS) for a brief period, the risks concerning privacy are less. Although the risks are less, USCIS assures the security control and access are ascertained to alleviate risks concerning privacy with authoritative and illegal users, explicitly abusive and improper distribution of information. Consecutively tracking and identifying illegal uses of system data audit trails will be kept. Additionally, the ASCs abide with security guidelines of the Department of Homeland Security (DHS), which give strengthened measures for securing computers, computer services, networks and against incidents and illegal information distribution [4].
The administrator requires developing a strategy when executing the data gathering phase to make sure the following are taken care of accordingly:

- To assist in the vulnerability evaluation the resources that will be required.
- In the intruder’s working area probable location of data source.
- With minimum interruption to the organizations’ systems particularly the critical servers and networks when the essential data is gathered.
- Measures to make certain the proof is not damaged or corrupted.
- Perception of technical information about the intruder’s linked and attached devices including network.

- To make sure proof gathering events are legally permissible and stick to respective country's law and organization policy.

1) Entire possible information sources must be contemplated by the administrator. From various data sources that may derive from various applications or devices the proceedings related to the vulnerability can be acquired.

- The logs may enclose precious information such as when the document was initially created and when it was last opened or changed, these logs are required to be given importance over non volatile data. Before they vanish the administrator must seize the logs.
- The intruder's system that was potentially utilized to execute the computer vulnerability for instance network details, the place of the computer system, information on operating systems and related application system.
- Volatile information could be possible sources of proof. Mostly, if a computer is shut down it passes the information to system time that is missing.

2) Distinguishing important steps to be taken to prevent maintaining information integrity and tampering for instance hashing techniques, image copy of hard disk and write blocker.

3) While handling the data gathering stage it is necessary to avoid bringing down crucial equipment such as the systems, servers and networks to minimize disruption to the organization.

4) The administrator has to make certain that the process he conducts does not interrupt the organization's IT and legal policies and for this purpose resource assignment is necessary for providing assistance and guidance on the inspection procedure. Consequently, it is vital to seek advice from HR personnel, legal department and internal audit. Additionally, support is required from the victim of the computer crime such as the IT department, management, line manager, system owner and application owner and business process owner.

5) In data gathering and analysis phase the personnel should be skilled on the significance of protecting the data against any unintentional tampering and conserving its integrity. It is also vital that all personnel at the site be sufficiently briefed on their functions and responsibilities.

6) The investigation should be conducted cautiously so that the intruder is uninformed that he is being scrutinized and the persons engaged in the inspection must know this. The intruder may try concealing his tracks and eradicate the proof if he is conscious about the inquiry.

7) Any advancement and revising of the vulnerability will be scrutinized by the administrator. Overall picture of the vulnerability crime case can be reviewed with a "vulnerability assessment board" that can provide details. To perform the event construction phase during the data analysis phase would be effortless for the administrator.

3. MODEL ANALYSIS

Laboratory is suitable for the analysis phase. While ensuring data integrity is conserved the data analysis can now be performed on the collected data.

a) Using the toolkit that endow with easy analysis of data at the lowest level the inspector analyses the data accumulated, searching for proof in the data that was gathered by investigating across a wide span of areas.

b) To make certain that the data investigated is not modified; write blockers can be utilized throughout the process to avoid writing to the said hard disk drive. Access to the image copy should be only as read-only. Solely on the replicated copy of the hard disk drive the analysis phase must be performed.

c) The intruder may have intentionally deleted or accidentally damaged some files. To reconstruct these lost files a toolkit can be utilized.

d) By comparing proof to activities and sources the inspector will then try to reconstruct the case. In reconstructing the case the case board will be helpful and if required, respect personal history or information of the intruder and other proofs (notes, photographs etc).

e) Only authorized people must be permitted into the laboratory. It should have an entrance control list having who had access to the collected data in the laboratory with dates and times.

Reporting is vital as it gives information of proceedings taken by the inspector to reconstruct the proof and inspecting. During the guideline process it should be emphasized that the sequence of protection and integrity of proof was sustained.

This research is utilized to explain to the people who use BIdM in any angle that the analysis was done in a fair and dependable manner and reporting is not utilized only for the internal, legal or criminal investigation. To demonstrate to the workers in the organization that the investigation was conducted in a reasonable and consistent manner according to the country's respective law, organization's policy and there should be transparency presented in the investigation.
The designed guideline was presented to the Assistant Controller on paper and on slides to enhance the things which were reviewed in the guideline. The Assistant Controller was selected as he is available all the time rather than the administrator and the consultant. The user was not opted as the Assistant Controller has vast experience in the field of BIdM systems management. The guideline has mentioned numerous vulnerabilities in the field of BIdM. However it concentrated more on one single area instead of all the areas as for this study it would be a vast topic if all the areas were covered. The vulnerabilities from user's side are one important fact that needs to be tackled. Business rules, standards, policies, frameworks and guidelines are very important for proper access control requirements.
For the guideline there are some areas that need to be improved.
It requires coordination among the concerned people such as the victim, IT department, the line manager, management, and any possessor of the system or the application that were involved by computer crime sequentially for the guideline to be effective.
Assistance will be provided to the inspector by these employees on an understanding of the effect of the computer crime to their application system. While preserving data integrity they can also provide information on ways to obtain critical proof from the applications.
During the progression the service of a case board will be beneficial as this can represent the entire sequence of events in a diagrammatical perspective:

- The place of the target system and intruder
- Organization chart / Network map
- How the computer crime was performed
- To identify the profile the intruder's history
- Evolution of the computer crime
- Congregated proof and additional proof that requires to be gathered or situated

To make certain that the authentic incident is captured before the computer investigation team does their work at the intruder's location during the data gathering phase. Hence prior to the proof collection or shifting from the intruder's place, the recording of incident such as obtaining photographs should be done.
Thorough method for approaching profiling of the intruder is vital, since it is important to know the individuality of the intruder particularly through data gathering and data analysis stage.
Personal history of the intruder is imperative for the inspector. The inspector could possibly build a profile of the intruder and his patterns in perpetrating the computer crime if he knows the intruder's criminal history. For instance, when dealing with a technically competent intruder the inspector has to be very cautious as he may have committed the computer crime utilizing convoluted technical skills.

- DNA Matching - Proof definite identification of an individual can be formed with this supreme biometric technology.
- Body Odor - For identification body odor can be digitally verified. Mastiff Electronic System Ltd is laboring on a similar system which is a British company.
- Vein pattern credentials - In that it utilizes infrared light creating the copy of pattern of vein in individual's wrist, hand, or face similar to retinal credentials.
- Keystroke Dynamics - Keystroke dynamics, is an innovative biometric technology and also referred to as typing rhythms.
- Ear shape credentials - The ear shape (geometry) is measured for this.
- Body salinity (salt) credentials - Progress in this part has been conducted by Massachusetts Institute of Technology (MIT) and IBM.

The inspector and the BIdM team at this point will be at the intruder's work area or in the physical site. The following must be conducted once the inspector reaches the site:

a) Avoid closing down any programs which are running while unplugging the power supply from the rear of the computer. To protect losing the basis of the proof a normal or graceful shut down should not be performed as temporary files, swap files or any malicious program will be erased.
b) Send unwanted people away from the computer by the security personnel and secure the place. This will prevent the proof from been damaged by anyone especially the intruder who would want to erase all proof.
c) The inspector should inspect to make certain that no self damaging program has been executed if the computer was running upon arrival. If such a program is operational the inspector should instantly remove the power cord from the wall socket.
d) To avoid other individuals who are accomplices from entering the intruder's computer via the network or internet to manipulate the data, the inspector should unplug the network cable / modem.
e) At the scene, to be documented the following should be accomplished:

- With accurate time and the proceedings carried at the site,
- Scene should be sketched by means of a network map,
- Place where the computer is situated, details of individuals nearby,
- Specifications of the computer such as the model number, serial number, operating system, IP address, system dates and other applications should be noted,
- Peripherals attached to the system, monitor details, etc.,
- If required, video tape or photographs should be taken.

f) To make certain it does not modify information on the primary medium where the backup or copy is, the inspector must utilize write-blocker specifically software support or hardware.
g) Preliminary interview should be performed by requesting constructive information relevant to the user identification, password information and about the computer system affected.
h) The items required to be printed out immediately are the things in volatile medium in the current memory such as logs, temporary files and printer buffer.
i) To ensure that the data is not tampered and permissible for legal explanation the inspector needs to authenticate the integrity of the data. Example of such techniques is hashing (This technique is used for authenticating biometric parameters via biometric hashing).
j) Make sure that completed exhibit labels are attached to the system, connection modes, peripherals and all objects have been signed.
k) The computer connectivity maybe restructured at a later date when the cables and the parts are tagged.
l) Prior to leaving the scenario let the intruder sign the proof, media and documentation gathered from the incident. The computed hashing of the hard disk drive must also be signed by the intruder.
The moving of detained media, computers and peripherals must be dealt with cautiously to prevent heat, damage or jostling for example, the place where the intruder maybe situated in a distinct physical location for instance in a branch office. Media devices must be kept in anti-static bags and be secluded from magnetic fields. To preserve the sequence of protection the carrying details should also be documented. Before moving the proof the inspector should note the names of title of all handlers of proof, the time of leaving from the computer crime scene, the advent time to the laboratory or store location.
Identity theft is one of the fastest growing crimes in the present day. Hence if industries require fighting this rising prevailing, data confidentiality and integrity is vital. This guideline will give ways on how to handle the crime scenes of identity theft, to gather information on such events, revealing trouble by existing verification systems, to provide robust verification, demonstrating how a resolution of biometric can be utilized, and considering additional advantages of utilizing multifactor authentication performances.

- Cloning Identity - To establish another life in this offense the imposter utilizes the victim's information. They live and work as a different person. For instance, criminals running away from warrants, unlawful aliens, turning into a "new life" to abandon a poor work and financial history or people disappearing from cruel situations.

- Criminal Identity theft - When blocked by law enforcement, the imposter in this crime gives the victim's information as an alternative to the imposter's own information. Ultimately it is in the identification of victim's when the warrant for arrest is released.
- Financial Identity theft – In this case normally the imposter focuses on the victim's identification and Social Security number (SSN). The imposter may request for credit cards, personal loan, buying merchandise, leasing cars or apartments and telephone service.
- Commercial or Business Identity theft - businesses are victims of identity theft as well. In the name of the business usually the imposter acquires checking accounts or credit cards. The business only finds out when the dejected suppliers send collection notification or when their business rating results are affected.
- Web spoofing - In this case the imposter monitor and change all web pages forwarded to the victim's machine and scrutinize all information the victim key in into the forms.

Instead of tokens or passwords, BIdM devices are able to offer improved security however they can give extra issues in engineering if proper standards, policies, frameworks and guidelines are not followed. Organizations planning to exploit BIdM strategies must seek recommendation from various aspects.

There are five major weaknesses the systems are prone to such as, user's workstation, communication paths facing interference, public network connections such as internet, dial-in-lines and interfaces with other systems in the network.

BIdM system's vulnerabilities can be reduced with well designed security architecture with principles of defense incorporated. Communication paths require encryption and other techniques, unattended workstations need physical security, conjunction with firewall intrusion detection system, connections should be permitted to public networks with approved firewalls, to any operating system accessing the network strong controls must be applied, firewall and standard circuits, local area networks (wireless), border controls and proper application of access control. The sensitive systems should be isolated while monitoring systems are accessed and used.

3. CONCLUSION & FUTURE WORK

Biometrics Identity Management (BIdM) is an intricate combination of the technological, cultural, legal and social aspects. Although standards, policies, frameworks, guidelines, methods and procedures for authenticating, securing and maintaining principle's confidential information is highly recommended. It is not simply about it all, even if this does have a vital component in its comprehension. Security with access control plays a major role. From unauthorized access, the repository which stores BIdM objects must be sheltered. To secure critical objects, cryptography and one-way hash functions may be vital. A prudent balance between Privacy and Security needs to be accomplished.

Due to revolutionary dynamics of organization, BIdM crime is on the rise, unsuccessful security system, upsurge in employee misconduct and lesser hacking expertise needed. BIdM worldwide is continually faced with threats and incidents of computer crime. BIdM standards, policies and guidelines are important processes that provide methods for organizations to take proper action against these employees, which will prevent future occurrences of computer crime. BIdM may offer organizations a stronger method of authentication as it is uniquely bound to individuals, with reliable user authentication and must ensure that an attacker cannot masquerade as a legitimate user. The BIdM was created based on literature review, primary and secondary research. Using appropriate technologies the planned BIdM contains seven phase tactic to extract proof centered on the intruder's workstation. The data gathering approach was very helpful as people from three different countries consisting of BIdM experts and the potential users of the computer BIdM participated.

In conclusion the BIdM which comprises of the stages and procedures were formed. A walk through with the BIdM manager who has vast knowledge and extensive experience in dealing with BIdM vulnerabilities especially in the area of network environments, revealed the areas for improvement. For instance, profiling the intruder the participation of the victim in the BIdM process and linking a case board. These were later incorporated into the guideline to make it more effective.

The data gathered must be secured and well protected ensuring that the contents cannot be overwritten in any way and it should be kept in a place where it would be free from any possible tampering, for instance, a protected location to keep the data.

Some organizations may not have standard procedures for assessment and backup for system and applications. Moreover there are also challenges in the proof gathered. Since it may not be permissible, organizations that possess these procedures may not be able to manipulate the data gathered as proof.

As stipulated by the regulator the usage of BIdM is more extensive where it is utilized for monitoring, evaluation of product, assessment of system, system auditing, data acquisition and data recovery.

IT Security must align with business needs and accept the importance of team work as engagement from other departments is crucial in order to make security work for the organization. Collaboration is best opted for any type of business.

In raising levels of quality most people are still usually unaware of the role played by the policies and standards, safety, efficiency, reliability and interchangeability as well as in providing such benefits at cost-effective manner. In general, these policies and standards can be made available from experts of a technical committee or working groups who meet to discuss and debate until they reach consensus on a draft agreement, or some new policies and standards or revision of existing policies and standards to bring to effect to follow up in the organization.

Based on the experts' views and present conditions, two critical areas of standardization with utilizing appropriate policies and the use of hybrid technologies are heavily dependent on the future of BIdM systems. Additional efforts and research on BIdM needs to be scrutinized and exploited in the manipulation of the user's private information such as what is exposed to whom and its influence guiding their awareness of BIdM in the world of e-commerce and the outline of the user's behaviors. The importance is being understood by users, designers and implementers of BIdM systems, managers, legislators, which will not be a minor task. It is also vital for national biometric associations, international biometric associations and regional biometric associations to collaborate and enhance their cooperation to establish international quality standards in biometric industry.

4. REFERENCES

1. Gordon, A., Loeb, M., Lucyshyn, W., Richardson, R., 2006, Computer Crime and Security Survey, [Online], Computer Security Institute & FBI, [Online] Available from: http://pdf.textfiles.com/security/fbi2006.pdf [Accessed on 19th July 2009]
2. Roberts, C., 2006, Biometric Technologies – Fingerprints, [Online] Available from: http://www.ccip.govt.nz/newsroom/information-notes/2006/biometrics-technologies-fingerprints.pdf [Accessed on 22nd July 2009]
3. Managing Australia's Borders, "The Department of Immigration and Citizenship (DIAC) [Online], Available from: http://www.immi.gov.au/managing-australias-borders/border-security/systems/identity.htm [Accessed on 12th September 2009]
4. Patentdocs, 2009, Systems and Methods for Accessing a Tamperproof Storage Device in a Wireless Communication Device Using Biometric Data [Online], Available from: http://www.faqs.org/patents/imgfull/20090193519_07 [Accessed on 10st October 2009]
5. Home Office – Identity and passport service, "Identity cards for airside workers", [Online], Available from: http://www.ips.gov.uk/cps/rde/xchg/ips_live/hs.xsl/964.htm [Accessed on 30th September 2009]

CALL FOR PAPERS

Journal: International Journal of Biometrics and Bioinformatics (IJBB)
Volume: 4 **Issue:** 3
ISSN:1985-2347
URL: http://www.cscjournals.org/csc/description.php?JCode=IJBB:

About IJBB

The International Journal of Biometric and Bioinformatics (IJBB) brings together both of these aspects of biology and creates a platform for exploration and progress of these, relatively new disciplines by facilitating the exchange of information in the fields of computational molecular biology and post-genome bioinformatics and the role of statistics and mathematics in the biological sciences. Bioinformatics and Biometrics are expected to have a substantial impact on the scientific, engineering and economic development of the world. Together they are a comprehensive application of mathematics, statistics, science and computer science with an aim to understand living systems.

We invite specialists, researchers and scientists from the fields of biology, computer science, mathematics, statistics, physics and such related sciences to share their understanding and contributions towards scientific applications that set scientific or policy objectives, motivate method development and demonstrate the operation of new methods in the fields of Biometrics and Bioinformatics.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJBB.

IJBB LIST OF TOPICS

The realm of International Journal of Biometrics and Bioinformatics (IJBB) extends, but not limited, to the following:

- Bio-grid
- Bioinformatic databases
- Biomedical image processing (registration)
- Biomedical modelling and computer simulation
- Computational intelligence
- Computational structural biology
- DNA assembly, clustering, and
- Bio-ontology and data mining
- Biomedical image processing (fusion)
- Biomedical image processing (segmentation)
- Computational genomics
- Computational proteomics
- Data visualisation
- E-health

mapping

- Fuzzy logic
- Gene identification and annotation
- Hidden Markov models
- Molecular evolution and phylogeny
- Molecular sequence analysis
- Gene expression and microarrays
- Genetic algorithms
- High performance computing
- Molecular modelling and simulation
- Neural networks

CFP SCHEDULE

Volume: 4
Issue: 4
Paper Submission: July 31 2010
Author Notification: September 01 2010
Issue Publication: September 2010

CALL FOR EDITORS / REVIEWERS

CSC Journals is in process of appointing Editorial Board Members for ***International Journal of Biometrics and Bioinformatics***. CSC Journals would like to invite interested candidates to join **IJBB** network of professionals/researchers for the positions of Editor-in-Chief, Associate Editor-in-Chief, Editorial Board Members and Reviewers.

The invitation encourages interested professionals to contribute into CSC research network by joining as a part of editorial board members and reviewers for scientific peer-reviewed journals. All journals use an online, electronic submission process. The Editor is responsible for the timely and substantive output of the journal, including the solicitation of manuscripts, supervision of the peer review process and the final selection of articles for publication. Responsibilities also include implementing the journal's editorial policies, maintaining high professional standards for published content, ensuring the integrity of the journal, guiding manuscripts through the review process, overseeing revisions, and planning special issues along with the editorial team.

A complete list of journals can be found at http://www.cscjournals.org/csc/byjournal.php. Interested candidates may apply for the following positions through http://www.cscjournals.org/csc/login.php.

Please remember that it is through the effort of volunteers such as yourself that CSC Journals continues to grow and flourish. Your help with reviewing the issues written by prospective authors would be very much appreciated.

Feel free to contact us at coordinator@cscjournals.org if you have any queries.

Contact Information

Computer Science Journals Sdn BhD
M-3-19, Plaza Damas Sri Hartamas
50480, Kuala Lumpur MALAYSIA

Phone: +603 6207 1607
+603 2782 6991
Fax: +603 6207 1697

BRANCH OFFICE 1
Suite 5.04 Level 5, 365 Little Collins Street,
MELBOURNE 3000, Victoria, AUSTRALIA

Fax: +613 8677 1132

BRANCH OFFICE 2
Office no. 8, Saad Arcad, DHA Main Bulevard
Lahore, PAKISTAN

EMAIL SUPPORT
Head CSC Press: coordinator@cscjournals.org
CSC Press: cscpress@cscjournals.org
Info: info@cscjournals.org

COMPUTER SCIENCE JOURNALS SDN BHD
M-3-19, PLAZA DAMAS
SRI HARTAMAS
50480, KUALA LUMPUR
MALAYSIA