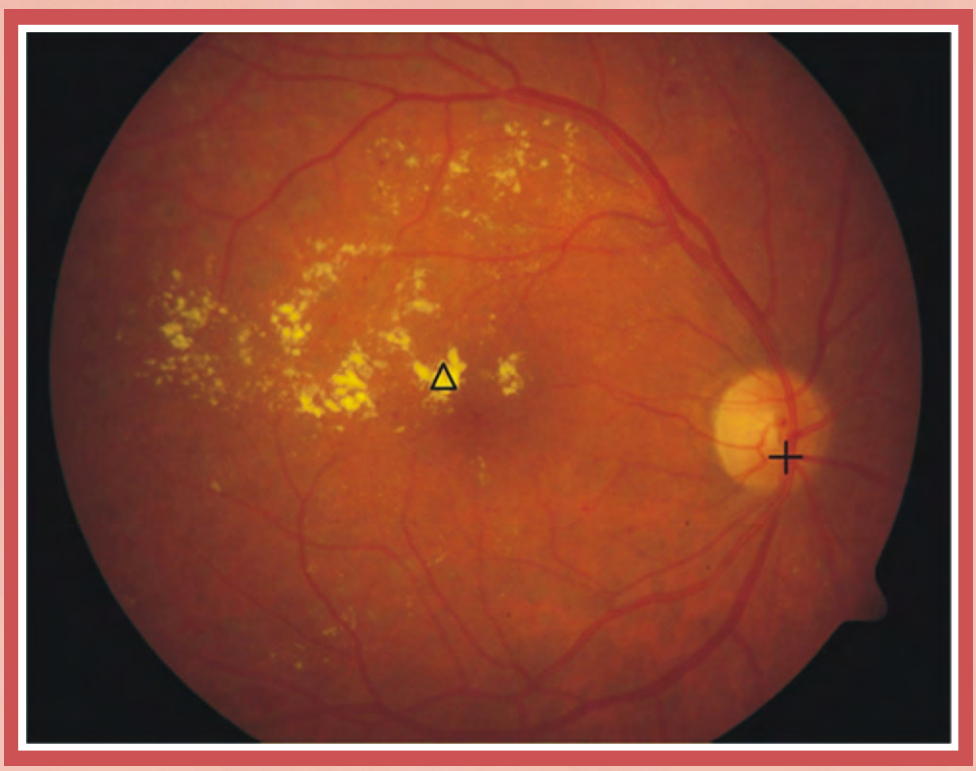
Scientific Research Publishing

JBiSE

ISSN: 1937-6871

Volume 2 Number 2 April 2009

Journal of Biomedical Science and Engineering



www. scirp. org/journal/jbise

Editor-in-Chief
Kuo-Chen Chou

Journal Editorial Board

ISSN 1937-6871 (Print) ISSN 1937-688X (Online)
Http://www.scirp.org/journal/jbise

Journal of Biomedical Science and Engineering (JBiSE)
Published Online April 2009 in SciRes (http://www.scirp.org/journal/jbise).

TABLE OF CONTENTS

Volume 2, Number 2, April 2009

News and Announcement
JBiSE Editorial Office……77
Systems Biology: The take, input, vision, concerns and hopes
G. Tucker……78
A novel approach in ECG beat recognition using adaptive neural fuzzy filter
G. N. Golpayegani, A. H. Jafari……80
Effects of lead exposure on alpha-synuclein and p53 transcription
P. J. Zuo, A. B. M. Rabie……86
Automatic detection and boundary estimation of optic disk in fundus images using geometric active contours
G. B. Kande, T. S. Savithri, P. V. Subbaiah, M. R. N. Tagore……90
The effect of different number of diffusion gradients on SNR of diffusion tensor-derived measurement maps
N. Zhang, Z. S. Deng, F. Wang, X. Y. Wang……96
The impact of frequency aliasing on spectral method of measuring T wave alternans
D. H. Chen, S. Yang……102
Micropath - A pathway-based pipeline for the comparison of multiple gene expression profiles to identify common biological signatures
M. Khan, C. B. Gorle, P. Wang, X. H. Liu, S. L. Li……106
Prediction of mutation position, mutated amino acid and timing in hemagglutinins from North America H1 influenza A virus
S. M. Yan, G. Wu……117
Bioinformatics analysis and characteristics of envelop glycoprotein E epitopes of dengue virus
H. Zhong, W. Zhao, L. Peng, S. F. Li, H. Cao……123
Analysis and expression of the polyhedrin gene of antheraea pernyi nucleopolyhedrovirus (AnpeNPV)
J. X. Huang, H. L. Wu, Y. Wu, S. Y. Zhu, W. B. Wang……128

The figure shown on the front cover illustrates the Comparison of optic disk localization with maximum local variation method. (From Giri Babu Kande, Journal of Biomedical Science and Engineering, 2009, 2, 90-95)

Journal of Biomedical Science and Engineering (JBiSE)

SUBSCRIPTIONS

The ***Journal of Biomedical Science and Engineering*** (Online at Scientific Research Publishing, www.scirp.org) is published biomonthly by Scientific Research Publishing, Inc. 5005 Paseo Segovia, Irvine, CA 92603-3334, USA.
E-mail: jbise@scirp.org

Subscription Rates: Volume 2 2009

Printed: $50 per copy.

Electronic: freely available at www.scirp.org.

To subscribe, please contact Journals Subscriptions Department at jbise@scirp.org.

Sample Copies: If you are interested in obtaining a free sample copy, please contact Scientific Research Publishing, Inc. at jbise@scirp.org.

SERVICES

Advertisements

Contact the Advertisement Sales Department at jbise@scirp.org.

Reprints (a minimum of 100 copies per order)

Contact the Reprints Co-ordinator, Scientific Research Publishing, Inc. 5005 Paseo Segovia, Irvine, CA 92603-3334, USA.
E-mail: jbise@scirp.org

PRODUCTION INFORMATION

For manuscripts that have been accepted for publication, please contact:
E-mail: jbise@scirp.org

J. Biomedical Science and Engineering, 2009, 2, 77-77

News and announcement

We are pleased to announce that two of the Editorial Board Members of **JBiSE**, Kuo-Chen Chou and Hong-Bin Shen, have been identified by Science Watch (http://sciencewatch.com/ana/fea/09maraprFea/) as the authors with the highest numbers of Hot Papers published over the preceding two years (2007 and 2008). Among the 13 authors listed in the table of **"**Scientists with Multiple Hot Papers" by Science Watch, Professor Dr. Kuo-Chen Chou of Gordon Life Science institute and Shanghai Jiaotong University ranks No.1 with 17 hot papers, and Associated Professor Hong-Bin Shen of Shanghai Jiaotong University ranks No.4 with 13 hot papers.

Meanwhile, the review article by Kuo-Chen Chou and Hong-Bin Shen, entitled "Recent Progresses in Protein Subcellular Location Prediction" published in Analytical Biochemistry, has been identified by Science Watch as the New Hot Paper in the field of Biology & Biochemistry (http://sciencewatch.com/dr/nhp/2009/09marnhp/09marnhpChou/).

For more information about the hot research and hot papers, go to visit the web-sites at http://www.sciencenet.cn/htmlnews/2009/3/216833.html; http://sciencewatch.com/ana/fea/pdf/09maraprFea.pdf; and http://sciencewatch.com/dr/nhp/2009/pdf/09marnhpChou.pdf.

Please join us to send our sincere and warm congratulations to our fellow board members, Kuo-Chen Chou and Hong-Bin Shen, for their prominent contributions in science. Meanwhile, we hope this announcement can attract more researchers to submit their best papers to **JBiSE**, the journal that publishes the highest quality of research and review articles in all important aspects of biology, medicine, engineering, and their intersection.

We would also like to take this opportunity to announce that, owing to the large number of manuscripts that we are receiving, **JBiSE** will increase publication frequency from quarterly to bi-monthly in 2009.

JBiSE Editorial Office

J. Biomedical Science and Engineering, 2009, 2, 78-79

Scientific Research Publishing

Systems Biology: The take, input, vision, concerns and hopes

Graeme Tucker[1]

[1]Lighthouse Laboratories, SABC Loneragan Building, Murdoch University, South Street, Murdoch, Western Australia 6150. Correspondence should be addressed to Graeme Tucker (gtucker@lighthouselabs.org.au).

Received Jul. 4th, 2008; revised Mar. 1st, 2009; accepted Mar. 9th, 2009

ABSTRACT

Systems Biology is a relatively new branch of biology that brings together an interdisciplinary team of scientist, computer engineers and mathematicians. Biomedicine can gain much from the input of Systems Biology. The object and aims of this article centre on clarification and direction for Systems Biology, notably in regard to human health and disease.

The take: Many consider today's Systems Biology as an incarnation that was instigated by the biotechnological revolution of high throughput data acquiring instruments. Data gathering and processing, emerged with two streams: refinement during the gathering stage or, refinement at the point of analysis. However, the rapid development of this high throughput hardware was missing the resources to decipher the vast amounts of data produced. Handling of such data required the adaptation of computational platforms and users. The user(s) emerged as an interdisciplinary cross breed of researchers: mathematicians, computer specialist and to a lesser extent the biologist, because of skill displacement. An immediate challenge for this new breed of researcher was defining Systems Biology. Initial workers reflected upon their activity for inspiration: informatics /computational biology demonstrates the 'relativity' of biological composite by use of 'networking' illustrations supported by software programs and mathematics. Consideration with regard to this form of data representation closely resembled their engineering heritage; in particular Control Systems. Thus, the word 'System' had relevance. Because the theme related to Biology the phrase Systems Biology was coined.

However, a decade on there is still controversy surrounding the definition of Systems Biology and what it means to a broad spectrum of researchers in their daily activities. Rightly or wrongly the biologist and philosopher is at risk of being displaced from the driving seat of biological research. Computology could function as a new word to describe the computer and mathematical specialist engaged in biology. Aside from those resisting change, many traditional biologist have reacted to this branch of biology as a branch lacking turgidity. Others argue that their wet medium for research is now waters muddied by Systems Biology. Advocates of Systems Biology from all camps continue to seek clarification and as such new definitions are regularly floated as vessels to carry Systems Biologists on their voyage of discovery.

Input: In a diffuse setting Systems Biology in one sentence could be described as follows: 'Systems Biology is a branch of biology that is an interjection of fields and disciplines, with applications collectively integrating broadly acquired data from all research levels for the purpose of indicating or modelling phenotype. '

A definition of Systems Biology in one word could possibly be, 'Relativity'. Nature complies with the law of physics and physical things are superimposed on mathematical principles. This can include stimuli such as the fight or flight response or, sexual arousal.

Of particular interest and relevance, is the notion that the Systems Biologist will consider and address their challenge beyond the immediate viewpoint of the biological composites of interest.

The term 'Model' is frequently used in context of Systems Biology. A model is a representation of a feature/function; a condition that is simulated or, a composite that is mimicked. That representation need not necessarily be a biological attribute. For example an intrinsic or extrinsic stressor (infection or sun stroke) can lead directly to a physical affect (fever or hyperthermia).. In an indirect setting such as arachnophobia, consideration is given to memory, psychological predisposition and reflex. An example of a traditional model for a biologist is an organism such as yeast or, the mouse. A model for the computer engineer is a software program; a model for a mathematician is an equation. Thus, computation serves to 'simulate,' mathematics serves to 'mimic' and biology serves to 'feature.' Together, they serve Systems Biology. From a retrospective perspective the phrase is perceived as an encompassing mode or stage. From another perspective, Systems Biology is a language, a script and an instrument. In many respects it is the mirrored reflection of Biological Systems in the literal sense.

Vision: In the not too distant future it is suspected that

Systems Biology will be a prerequisite of project planning for wet based laboratory research, much in the same way that consideration of ethics is a necessity for research endorsement and support.

Concerns: Wet laboratory research has had the *in vitro* model for many years. For the researcher it has proved unambiguous in distinguishing the significance and relativity of outcomes to that of the *in vivo* setting. However, with the emergence of the *in silico* model the risk of ambiguity is clearly present. Of particular concern is that wet lab 'fact' is fiction-ated (*in silico*) and that the subsequent fiction outcomes, after model manipulation / extrapolation, are inferred as fact in the bigger picture of true fact. In that setting the audience of communicated *in silico* science is at risk of having their own works misguided, notably the wet laboratory researcher. Therefore is the informatics camp muddying the waters of wet lab science?

In addition, is *in silico* science operating with aseptic techniques when fact-ors are introduced in the extrapolation process? Is there a risk of fashioning fact to the point of becoming arte-fact?

Equally, is there a risk of fashioning wet laboratory experiments to fit the predicted *in silico* assumptions?

Hopes: Systems Biology needs some standardisation policies and techniques to ensure that the concerns outlined above are not encountered.

The incorporation of Systems Biology principles in project planning will hopefully have the benefit of minimising and reducing calculated aspects such as cost and risk. However, in the spirit of far reaching science (true philosophy), which is usually accompanied by risk, it is hoped that this reach is not compromised; if anything, it is hoped that Systems Biology will gather the pace and direction of research.

Turnaround benefits shall hopefully be seen in medicine (facilitation of modes for the prevention, diagnosis, treatment and cure of disease), agriculture (increased crop yield, crop resistance and crop tolerance) and in the development of sustainable energy through bio-fuels.

Beyond Systems Biology is Systems Life.
Can it factor in consciousness and meaning?

J. Biomedical Science and Engineering, 2009, 2, 80-85

A novel approach in ECG beat recognition using adaptive neural fuzzy filter

Glayol Nazari Golpayegani[1], Amir Homayoun Jafari[1]

[1]Biomedical Engineering Department, Islamic Azad University, Science and Research Branch, Tehran, Iran.
Email: Gelayol777@yahoo.com, Amir_j73@yahoo.com

Received Jan. 5th, 2009; revised Jan. 16th, 2009; accepted Feb. 10th, 2009

ABSTRACT

Accurate and computationally efficient means of electrocardiography (ECG) arrhythmia detection has been the subject of considerable research efforts in recent years. Intelligent computing tools such as artificial neural network (ANN) and fuzzy logic approaches are demonstrated to be competent when applied individually to a variety of problems. Recently, there has been a growing interest in combining both of these approaches, and as a result, adaptive neural fuzzy filters (ANFF) [1] have been evolved. This study presents a comparative study of the classification accuracy of ECG signals using (MLP) with back propagation training algorithm, and a new adaptive neural fuzzy filter architecture (ANFF) for early diagnosis of ECG arrhythmia. ANFF is inherently a feed forward multilayered connectionist network which can learn by itself according to numerical training data or expert knowledge represented by fuzzy if-then rules [1]. In this paper we used an adaptive neural fuzzy filter as an ECG beat classifier. We combined 3 famous wavelet transforms and used them mid 4 the order AR model coefficient as features. Our results suggest that a new proposed classifier (ANFF) with these features can generalize better than ordinary MLP architecture and also learn better and faster. The results of proposed method show high accuracy in ECG beat classification (97.6%) with 100% specificity and high sensitivity.

Keywords: Adaptive Neural Fuzzy Filter, ECG Arrhythmia Classification, Pattern Recognition, Multilayer Perceptron.

1. INTRODUCTION

Electrocardiography deals with the electrical activity of the heart. Bio-signals being non-stationary signals, the reflection may occur at random in the time-scale. Therefore, for effective diagnostic, ECG pattern and heart rate variability may have to be observed over several hours. Thus the volume of the data being enormous, the study is tedious and time consuming. Therefore, computer-based analysis and classification of cardiac diseases can be very helpful in diagnostic [1]. Several algorithms have been developed in the literature for detection and classification of ECG beats. One of ECG beat recognition is neural network classification method [2-9]. Multilayer perceptron (MLP), has been shown to be able to recognize and classify ECG signals more accurately. However conventional Neural Networks with Back Propagation algorithm (BPNN) suffers from slow convergence to local and global minima and from random settings of initial of weights, which may make the neural networks have very poor mappings from inputs to output. More over in conventional neural networks, users have to determine the structure of network such as the numbers of hidden layer before training and it is too hard to select proper structure. Another ECG classification method is Fuzzy Hybrid Neural Network [10]. In this structure a FCM algorithm is used to clustering the features and the center of clusters will be used as the input of MLP neural network. This structure made the results conventional MLP better but because of using MLP as final classifier, the shortcomings of MLP are still exist.

To overcome the shortcomings encountered in neural networks, while still keeping their advantages, an adaptive neural fuzzy filter, (ANFF) has been developed in [1]. The ANFF is a feed forward multilayer network that integrates the basic elements and functions of a traditional fuzzy system into a connectionist structure. An important feature of this adaptive filter is that it can dynamically partition the input space and output space using irregular fuzzy hyper box [11]. For the adaptation of membership functions in the ANFF, the back propagation algorithm is used to find the optimal parameters under the mean square error (MSE) criterion. Hence, in the ANFF, the Fuzzy ART is used for structure learning and the back propagation algorithm for parameter learning. The ANFF can thus on-line partition the input-output spaces, tune membership functions, and find proper fuzzy logic rules dynamically on the fly. Users need not give the initial fuzzy partitions, membership functions, or fuzzy logic rules except for the case that

expert knowledge is available and is used as the initial fuzzy rules. Hence, there are no hidden nodes in the beginning of learning; they are created and begin to grow as the training signal arrives. Since the structure of the ANFF is constructed from fuzzy if-then rules, once the input-output relationship is constructed, it will not be destroyed and, thus, no knowledge forgetting may happen while in conventional neural network we might had this event. These properties can make the ANFF more suitable for on-line classification than neural networks. Therefore in this paper we decided to use this filter as a classifier instead of a filter. More over, in this paper we used the fuzzy combination of 3 wavelets which were: Daubechies, Symlet, Biorthogonal mid 4th order AR model coefficients as features. These features beside ANFF had more efficient results than MLP. This algorithm was faster and more reliable than MLP and it has high mean accuracy 97.6% and also high sensitivity and specificity.

This paper organized as follows. Section 1 describes the basic structure and functions of ANFF in brief. The on-line structure/parameter learning algorithm of the ANFF, which combines fuzzy ART and back propagation learning algorithm under the MSE criterion is presented in Section 3. In Section 4 we describe the feature extraction method. In Section 6 we will show the results of this method and compare this method with conventional neural network (MLP). Finally conclusions are summarized in the last section.

2. THE STRUCTURE OF ADAPTIVE NEURAL FUZZY FILTER

2.1. Adaptive Neural Fuzzy Filters

In this section, we will describe the structure and functions of ANFF briefly. The ANFF (see **Figure 1**) has five layers with node and link numbering defined by the brackets on the left-hand side of the figure.

Layer-1 nodes are input nodes representing input variables. Layer-5 nodes are output nodes representing output variables. Layer-2 and layer-4 nodes are term nodes that act as membership functions representing the terms of respective input and output variables. Each layer-3 node is a rule node representing one fuzzy logic rule. Thus, together all the layer-3 nodes will be as a fuzzy rule base. The links between layers 3 and 4 function as a connectionist inference engine. Layer-3 links define the preconditions of rule nodes, and layer-4 links define the consequents of the rule nodes. Therefore, each rule node has at most one link to some term node of a linguistic node, and may have no such links. This is true both for precondition links (link in layer 3) and consequent links (links in layer 4). The links in layers 2 and 5 are fully connected between linguistic nodes and their corresponding term nodes. The arrows indicate the normal signal flow directions when the network is in operation (after it has been built and trained). When we are in structure learning step, ANFF operates in Up-Down mode and when we want to obtain estimated output, ANFF operates in Down-Up mode.

The ANFF uses the technique of complement coding from fuzzy ART [12] to normalize the input-output training vectors. Complement coding is a normalization process that rescales an n-dimensional vector, **x=(**x1, x2… xn**),** to its 2n-dimensional complement coding such that

$$\begin{aligned} x' &\equiv \left(\bar{x}_1, \bar{x}_1^c, \bar{x}_2, \bar{x}_2^c, \cdots, \bar{x}_n, \bar{x}_n^c\right) \\ &\equiv \left(\bar{x}_1, 1-\bar{x}_1, \bar{x}_2, 1, -\bar{x}_2, \cdots, \bar{x}_n, 1-\bar{x}_n\right) \end{aligned} \quad (1)$$

where $x' \equiv (\bar{x}_1, \bar{x}_2, \ldots \bar{x}_n) = \bar{x} = \frac{x}{\|x\|}$ and $\bar{x}_i^c$ is the complement of $\bar{x}_i$, i.e., $\bar{x}_i^c = 1 - \bar{x}_i$. As mentioned in [11], complement coding helps avoid the problem of category proliferation when using fuzzy ART clustering. It also preserves training vector amplitude information. In applying the complement coding technique to the ANFF, all training vectors (either input state vectors or desired output vectors) are transformed to their complement coded form in the preprocessing process, and the transformed vectors are then used for training.

A typical network consists of nodes with some finite number of fan-in connectionist from other nodes represented by weight values, and fan-out connectionists to other nodes. Associated with the fan-in of a node is an integration function f which combines information, activation, or evidence from other nodes, and provides the net input, i.e.,

$$net - input = f\ (z_1^{(k)}, z_2^{(k)}, \ldots z_p^{(k)}; w_1^{(k)}, w_2^{(k)}, \ldots w_p^{(k)}) \quad (2)$$

where $Z_i^{(k)}(i), \ldots p$ is the ith input to a node in layer k, and w(i) is the weight of the associated link. The superscript in the above equation indicates the layer number. This notation will be also used in the following equations. Each node also outputs an activation value as a function of its net-input.

$$output = a(f) \quad (3)$$

The description of the functions of nodes in each of five layers of the ANFF can be obtained from Reference [1].

2.2. The Structure-Learning Step

The structure learning step consists of three learning processes: input fuzzy clustering process, output fuzzy clustering process, and mapping process. The first two processes are performed simultaneously on both sides of network, and are described below.

2.2.1. Input Fuzzy Clustering Process

For input fuzzy clustering, ANFF uses the fuzzy ART fast learning algorithm [11,12] to find the input membership function parameters $V_{ij}^{(2)}, U_{ij}^{(2)}$. For this purpose, first, the values of choice functions, Tj, for each complement coded vector are computed by

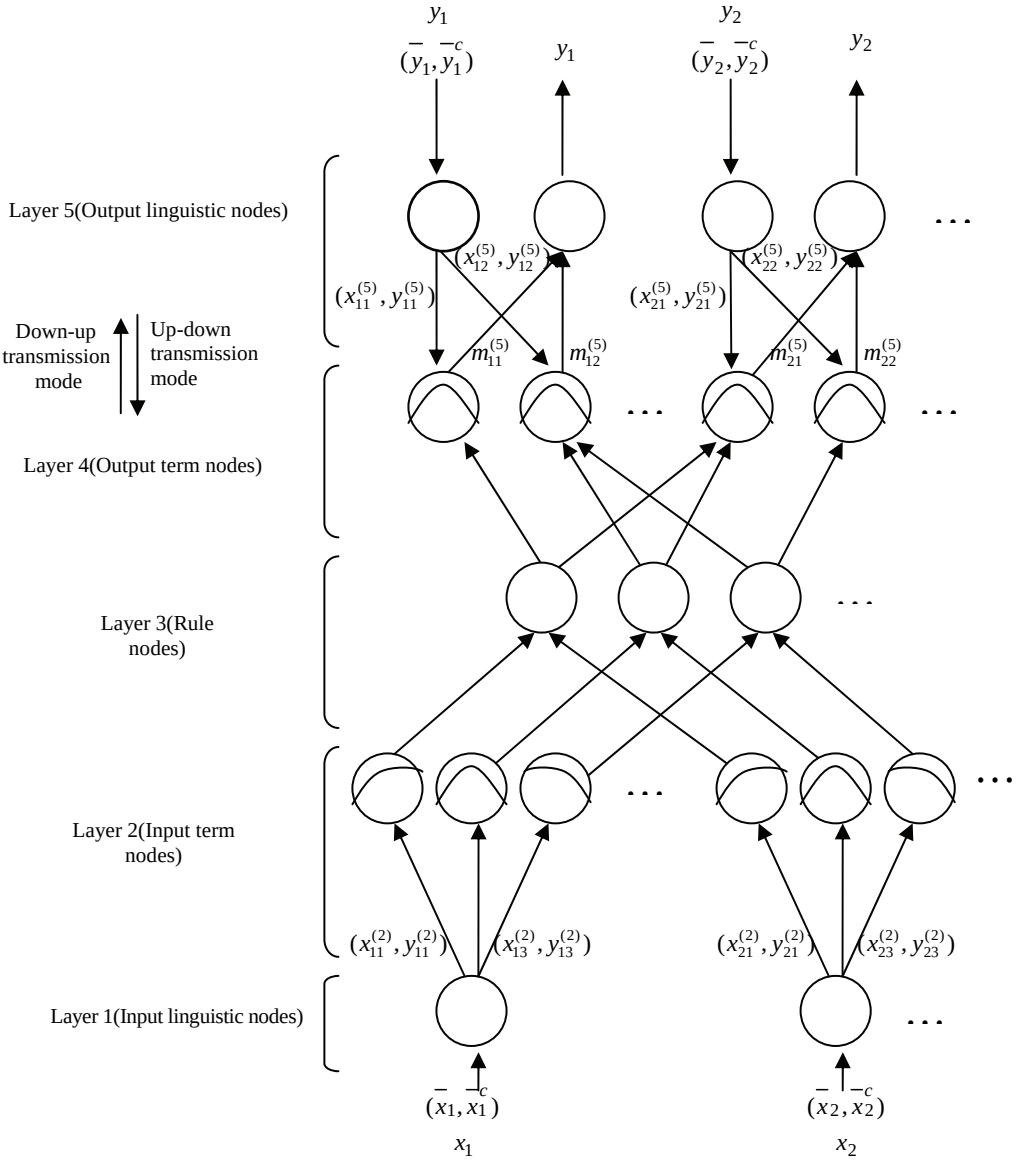


Figure 1. Structure of Adaptive Neural Fuzzy Filter (ANFF).

$$T_j(x') = \frac{|x' \wedge \omega_j|}{a + |\omega_j|} \qquad j = 1,....,N \tag{4}$$

where "Λ" is the minimum operator performed for the pairwise elements of two vectors, α ≥ 0 is a constant, N is the current number of rule nodes, and wj is the complement weight vector for each rule node.

Note that the choice function value indicates the similarity between the input vector **x′** and the complement weight vector **w**j. We then need to find the complement weight vector closest to **x′.** The chosen category is indexed by **J**, where

$$\frac{|x' \wedge \omega_j|}{|x^t|} \geq \rho \tag{5}$$

where ρ is between 0 and 1 is a vigilance parameter. If the vigilance criterion is not met, we say *mismatch reset* occurs. In this case, the choice function value TJ is set to zero for the duration of the input presentation to prevent persistent selection of the same category during search (we call this action "disabling J"). A new index J is then chosen using (20). The search process continues until the chosen J satisfies (21). If no such J is found, then a new input hype box is created by adding a set of n new input term nodes, one for each input linguistic variable, and setting up links between the newly added input term nodes and the input linguistic nodes. The complement weight vectors on these new layer-2 links are simply given as the current input vector, x′. These newly added input term nodes and links define a new hyper box, and thus a new category, in the input space. We denote this newly added hyper box as J.

2.2.2. Output Fuzzy Clustering Process

The output fuzzy clustering process is exactly the same as the input fuzzy clustering process except that it is performed between layers 4 & 5 which are working in the up-down transmission mode.

2.2.3. Mapping Process

After the two hyper boxes in the input and output spaces are chosen in the input and output fuzzy clustering processes, the next step is to perform the mapping process which decides the connections between layer-3 and layer-z4 nodes. This mapping process is described by the following algorithm, wherein connecting rule node J output hyper box K we means connecting the rule node J to the output term nodes that constitutes the hyper box K in the output space.

Step 1: IF rule node J is a newly added node THEN connect rule node J to output hyper box K.

Step 2: ELSE IF rule node J not connected to output hyper box K originally THEN disable J and perform Input Fuzzy Clustering Process to find the next qualified J.

Step 3: ELSE no structure change is necessary. In the mapping process, hyper boxes J and K are resized according to the *fast learning rule* [40] by updating weights, WJ and WK, as

$$W_J^{(new)} = x' \wedge W_J^{(old)} \quad , W_k^{(new)} = x' \wedge W_k^{(old)} \tag{6}$$

2.3. The Parameter-Learning Step

After the network structure has been adjusted according to the current training pattern in the structure-learning step, it is then necessary to fine tune the network parameters using the same training pattern. Basically, the back propagation algorithm is used to find node output errors, which are then analyzed to guide parameter adjustment. As mentioned above, the goal of training the ANFF is to minimize the error function

$$E = \frac{1}{2}\left(S(K) - \hat{S}(K) \right)^2 \tag{7}$$

where s(k) is the desired signal, and s^(k) is the filtered signal. Based upon this MSE criterion and in analogy to the back propagation algorithm, we can derive the following

$$W(K+1) = W(K) + \Delta W(K) = W(K) + \eta\left(-\frac{\partial E}{\partial w} \right)$$
$$-\frac{\partial E}{\partial W} = -\frac{\partial E}{\partial F} \cdot \frac{\partial F}{\partial W} = -\frac{\partial E}{\partial \alpha} \cdot \frac{\partial \alpha}{\partial f} \cdot \frac{\partial f}{\partial w} \tag{8}$$

where W is the adjustable parameter in the filter.

The adapting equations of the parameters of each layer can be obtained from [1] in more details.

3. DATA ACQUISITION AND FEATURE EXTRACTION METHOD

We had 4 kinds of ECG beats. There were: Atrial Fibrillation beats (AF), Ventricular Tachycardia beats (VT), Super Ventricular Tachycardia Arrhythmia beats (SVA) and Normal beats. We selected 18 signals from each kind of beat from MIT-BIH Arrhythmia database. So we had 72 signals totally. The arrhythmia database of MIT-BIH, have been preprocessed in such away that they are clean from any corrupting noise. Therefore we did not need to do any additional preprocessing on our ECG signals. The sampling frequency of MIT-BIH Arrhythmia database is 250Hz. We used a 1000 point moving window with 200 point overlap between each two windows for each ECG signal. So each ECG signal converted to 24 segments which each segment had 1000 point. Therefore we had 1728 segment totally. We used 80% of these segments for training ANFF and 20% of that to test the performance of ANFF.

We used a fuzzy combination of 3 wavelet transform which were: Daubechies (db4), Symlets (sym8) and Biorthoginal (bior4.4). A 4 level decomposition was done with each wavelet transform. And then we selected the maximum value between these 3 wavelets in each level. Therefore, 4 wavelet coefficients were obtained for each segment with this fuzzy combination if wavelet transforms. We also used the 4th order AR model coefficients as another feature extraction method so 4 features were obtained for each segment by this method. Therefore finally we had an 8 dimensional feature vector for each segment. Then we used these features as input for ANFF. We also defined the target for each segment and used them as output variable for ANFF.

4. RESULTS

We considered following values for the constants in ANFF: learning parameter (η)=0.1, fuzziness parameter (γ)=0.6, choice function parameter (α)=0.7.

In the first step of process we used only the Daubechies wavelet coefficients as ECG features. In the second step of process we used Daubechies wavelet mid 4th order AR model coefficient as ECG features. In third step of process we used only the fuzzy combination of three wavelets as ECG features. And finally we used fuzzy combination of wavelets mid 4th order AR model coefficient as ECG features. **Table 1** shows the results of using these features with ANFF separately. Then also did these steps with an MLP with 40 hidden nodes. **Table 2** shows the results of using these features with MLP. In both methods (ANFF and MLP) we calculated specificity and sensitivity and accuracy for each kind of features. We selected 1382 segments for training and 346 segments for test randomly.

As we can see from **Table 1** and **Table 2**, among 4 kinds of features extracted, the fuzzy combination of 3 pre nominate wavelet transforms mid 4th order AR model coefficients provided best results for both ANFF and MLP structures. Also we can see that in compare with MLP, ANFF had much better results. ANFF had high accuracy about 97.6% and also it had higher specificity and sensitivity than MLP. Furthermore unlike the MLP, ANFF did not need to have pre determined struc-

Table 1. Results of ANFF in ECG beat recognition.

Classifier	Total test segments	SE(%)	SP(%)	Accuracy(%)
ANFF with wavelet features	346	92.7%	98.1%	90.33%
ANFF with wavelet and AR model coefficients features	346	95.7%	98.4%	94.2%
ANFF with fuzzy combination of wavelets	346	96.6%	98.6%	92.1%
ANFF with fuzzy combination of wavelets and AR model coefficient and features	346	97.6%	100%	97.6%

Table 2. Results multi layer perceptron meural network (MLP) in ECG beat recognition.

Classifier	Total test segments	SE(%)	SP(%)	Accuracy(%)
MLP with wavelet features	346	91.4%	96.7%	86.7%
MLP with wavelet and AR model coefficients features	346	93.49%	96.93%	91.3%
MLP with fuzzy combination of wavelets	346	95.3%	97.47%	88.7%
MLP with fuzzy combination of wavelets and AR model coefficient features	346	94.76%	99.6%	94.2%

ture and it will find the best structure during training by itself.

5. CONCLUSION

In this paper an ANFF has been developed to classify ECG signals by using 4 different features set. Results show that among these features set, a fuzzy combination of Daubeches, Symlets and Biorthogonal wavelet transforms mid 4^{th} order AR model coefficients, had best results.

As we know, several algorithms have been developed for ECG beat recognition in literatures. Most of these algorithms have used Neural Networks as their final classifier. There are several kinds of Neural Networks. One of the most useful neural networks in ECG classification is Multi Layer Perceptron (MLP) neural network. It is easy to use and it has been shown reliable results in ECG beat classification. Therefore we decided to compare the performance of our purposed method with MLPNeural Networks, such as MLP, are capable in classification. They have strong learning and generalization ability but they have some disadvantages. For example: (1) we need iterative training cycles to encode the relation between inputs and outputs into the neural networks, so neural networks have long training time. (2) users have to predetermine the size and structure of neural networks (such as the numbers of nodes of hidden layers) initially and it is hard for users. (3) we can not use linguistic rules in neural networks directly.

Adaptive Neural Fuzzy Filters (ANFF) overcomes these shortcomings. In ANFF, users do not need to predetermine the numbers of nodes in hidden layer and these nodes are generated automatically during the training process. Since the structure of ANFF is optimum, it takes much shorter to train in compare with neural networks. More over, we can use linguistic rules, which are obtained from expert knowledge, in ANFF directly. ANFF was introduced by Chin-Teng Lin and Chia-Feng Juang in 1997 [1]. They introduced the structure of ANFF in details and they used ANFF as an adaptive filter for noise cancellation. We decided to use this useful filter as a classifier. Therefore in this paper we used ANFF as an ECG beat classifier and we compared its performance with most commonly used classifier (MLP).

On the other hand, wavelet transforms (WT) are must commonly feature extraction method in ECG beat recognition algorithms. An over view on previous works in this context shows that AR model beside wavelet transform has showed better results in ECG beat classification. More over, there are several mother wavelets which are used for ECG feature extraction. Therefore we decided to combine three mother wavelets which are most common used for ECG feature extraction and we used AR model coefficients beside this combination as our final feature extraction method.

A comparative assessment of performance of ANFF with MLP neural networks show that more reliable results are obtained with the ANFF for the classification of ECG signals.

MLP neural networks are still able to generalize with good recognition accuracy. However, they take longer to train and users have to predetermine the size and structure of network such as the number of hidden layers and it is truly difficult. The aim in developing ANFF was to achieve more optimum results with relatively few signal features. It has been demonstrated that the training time of ANFF was much shorter than time required by MLP and the accuracy (97.6%), specificity (100%) and sensitivity (97.6%) of ANFF were much

better than those of MLP. Furthermore the ANFF can be trained by numerical data and linguistic information expressed by fuzzy if-then rules. Another key feature of ANFF is that, without any given initial structure, the ANFF can construct itself automatically from numerical training data.

The proposed method, which incorporates the techniques of Adaptive Neural Fuzzy Filter and back propagation learning and combines their advantages, can be said to be more capable of recognizing other biological signals than conventional Neural Networks with Back Propagation algorithm (BPNN) such as Multi Layer Perceptron (MLP).

REFERENCES

[1] C.T. Lin, C.F. Juang, (2001) "An adaptive neural fuzzy filter and its applications," IEEE Transactions On Systems, MAN, And Cybernetics, VOL. 27, NO. 4, 1103−1110.

[2] S. Osowaki, T.H. Linh, (2001) "ECG beat recognition using fuzzy hybrid neural network," IEEE Trans. Biomed. Eng. 48 (11) 1265−1271.

[3] Y. Ozbay, B. Karlik, (2001) "A reconition of ECG arrhythmias using artificial neyral network," Proceedings of the 23rd Annual Conference, IEEE/EMBS, Istanbul, Turkey, pp. 76−80.

[4] Y. Ozbay, "Fast recognition of ECG arrhythmias," (1999) Ph.D. Yhesis, Institute of Natural and Applied Science, Selcuk University,.

[5] S.Y. Foo, G. Harvey, A. Meyer-Baese, (2002) "Neural network-based ECG pattern recognition", Eng. Appl. Artif. Intell. 15, 353−360.

[6] V. Pilla, H.S. Lopes, (1999) "Evolutionary training of a neuro-fuzzy network for detection of a P wave of the ECG," Proceeding of the third international conference on computational intelligence and multimedia applications, New Dehli, India, 102−106.

[7] M. Engin, S. Demirag, (2003) "Fuzzy-hybrid neural network based ECG beat recognition using three different types of feature sets," Cardiovasc. Eng. Int. J. 3 (2) 71−80.

[8] S. Hykin, (1994) Neural Networks: A comperhensive Foundation, Macmillan, New York.

[9] B. Karlik, m.o. Tokhi, M. Alci, (2003) "A fuzzy clustering neural network architecture for multifunction upper-limb prosthesis," IEEE Trans. Biomed. Eng. 50 (11), 1255−1261.

[10] R. Acharya, P.S. Bhat, S.S. Iyengar, A. Roo, S. Dua, (2001) "Classification of heart rate data using artificial neural network and fuzzy equivalence relation," J. Pattern Recognition Soc, 4, 238−244.

[11] G. A. Carpenter, S. Grossberg, and D. B. Rosen, (2001) "Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive adaptive resonance system," Neural Networks, 4, 759−771.

[12] G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen, (2002) "Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps," IEEE Trans. Neural Networks, vol. 3, pp. 698−712.

J. Biomedical Science and Engineering, 2009, 2, 86-89

Effects of lead exposure on alpha-synuclein and p53 transcription

Pei-Jun Zuo[1], A. Bakr M. Rabie[1]

[1]Faculty of Dentistry, the University of Hong Kong, Hong Kong SAR. Correspondence should be addressed to Peijun Zuo (pzuo@hkucc.hku.hk).

Received Nov. 1st, 2008; revised Feb. 19th, 2009; accepted Feb. 23rd, 2009

ABSTRACT

Objective: Epidemiological studies have found that lead exposure increases the risk for Parkinson's disease and patients with Parkinson's disease have lower odds of developing non-smoking-related cancer (1). It would be interesting therefore to find the molecular links between Parkinson's disease and cancer. To do this, we studied mRNA expression of alpha-synuclein gene, a promising genetic marker for Parkinson's disease, and expression of the tumor suppressor gene p53 after oxidative stress induced by lead. Methods: We used ATDC5 cell line as a model of tumor and treated by lead nitrate for 0, 2, 4, 16, 24 and 48 hours. The mRNAs of alpha-synuclein and p53 were quantified by reverse transcriptase polymerase chain reaction and expressed as mean (±SD) for 3 samples at each time point. Results: Expression of both of alpha-synuclein and p53 mRNA increased with increasing exposure of lead treatment. The levels of alpha-synuclein and p53 mRNA were correlated with each other (r=0.9830; P<0.001). Conclusion: We propose that lead's neurotoxicity in PD is caused by alpha-synuclein expression and aggregation, which releases the inhibitory influence of alpha-synuclein on p53 expression, thereby allowing p53 to act as the cell's guardian of the genome and reduce tumorigenic potential. Treatments that reduce alpha-synuclein aggregation may need to account for a concomitant reduction in p53's protective effect.

Keywords: Alpha-synuclein, p53, Real-time PCR, ATDC5, Aging, Cancer

1. INTRODUCTION

Parkinson's disease (PD) typically affects people aged 50 years and older, and the risk of disease increases with age. A promising diagnostic marker for PD is alpha-synuclein (2), which is the primary structural component of inclusion bodies (Lewy bodies) that are found in the neurons of patients with PD. Interestingly, epidemiological evidence shows that individuals with PD have reduced odds for many common types of non-smoking-related cancers (1). However, it is not known if this finding indicates a direct association between the two diseases, such as a reduced risk of non-smoking-related cancer among patients who develop PD, or a reduced risk of developing PD or other age-related neurodegenerative diseases among individuals with non-smoking-related cancer. It would therefore be interesting to study the possible biological and molecular links between age-related neurodegenerative diseases such as PD and cancer, especially because cancer is often also regarded as a disease of aging. The findings would provide important information on the mechanisms underlying normal and abnormal developmental and ageing processes. The p53 tumor suppressor protein may be one such link because hyperactivation of p53 in mice has been shown to increase resistance to spontaneous tumorigenesis while apparently accelerating aging (3). Cancer and neurodegenerative diseases might also be interrelated at the etiologic or environmental level: occupational exposure to lead is a risk factor for PD (4), while lead can be carcinogenic in rodents and genotoxic in fish (5), and it can also induce cell apoptosis via p53 (6).

To investigate the possible link between lead exposure and expression of alpha-synuclein and p53, we used a lead-sensitive cell culture model. We selected the ATDC5 cell line because it is an established mouse embryonic carcinoma-derived cell line that has both carcinogenic and chondrogenic properties (7). Not only does this cell line have chondroprogenitor potential and produce chondrocyte-specific extracellular matrix when stimulated by insulin, but it can also proliferate rapidly in the presence of fetal bovine serum (8). Expression levels of alpha-synuclein and p53 were thus measured at the mRNA level after treatment of cells with lead nitrate.

2. EXPERIMENTAL PROCEDURES

Cell Culture -The ATDC5 cells were cultured in a 1:1 mixture of Dulbecco's modified Eagle's medium and Ham's F-12 medium (Flow Laboratories, Irvine, UK) containing 5% fetal bovine serum (GIBCO BRL, Gaithersburg, MD), 100 U/mL penicillin, and 100 µg/mL streptomycin (Biofluids Inc., Rockville, MD, USA)

and then incubated at 37ºC in a humidified atmosphere containing 5% carbon dioxide. An inoculum of cells (10^4 per mL in 30 mL) was transferred to each of 7 Petri dishes. Lead nitrate was added to the cells. The final concentration of the lead nitrate per dish was 200 µmol/L. For increasing the solubility of lead nitrate, glutamic acid was added to medium in equimolar amounts of the lead nitrate. The cells were harvested at times of 0, 2, 4, 16, 24 and 48 hours.

Reverse Transcriptase Polymerase Chain Reaction Analysis- Total RNA was isolated from cells by using an RNeasy Mini Kit (Qiagen Sciences, Germantown, MD) according to the manufacturer's instructions. Reverse transcription of the mRNA was performed with oligo-(dT) primers and MulV reverse transcriptase (Applied Biosystems, Foster City, CA). Polymerase chain reaction (PCR) was then done in a StepOne Real-Time PCR System (Applied Biosystems); each 20-µL sample contained Power SYBR Green PCR Master Mix (Applied Biosystems), 40 ng of complementary DNA, and the pairs of primers listed in Table 1. The cycling profiles were as follows: 95ºC for 10 minutes, 95ºC for 15 seconds and 60ºC for 1 minute (for 40 cycles). Samples were run concurrently with standard curves derived from PCR products, and serial dilutions were performed to obtain appropriate template concentrations.

Mouse β-actin was used as an example of a lead-insensitive house-keeping protein and thus its primer acted as a negative control to correct for RNA recovery and reverse transcription efficiency. The mRNA concentrations were determined by by optical density at λ260/280 and were standardized at each time point to mouse β-actin mRNA concentrations.

Statistical Analysis- Data were expressed as mean ± SD for 3 or more replicates per sample, in arbitrary units relative to the mRNA level of β-actin. The student t test was used to evaluate differences between groups. Differences were considered significant at a level of $p < 0.05$.

3. RESULTS

Alpha-synuclein mRNA- Lead exposure induced a 50% increase in alpha-synuclein mRNA expression in ATDC5 cells at 2 hours (**Figure 1**). The mRNA level further increased by about 20% between 2 hours and 16 hours, remained the same between 16 and 24 hours and then increased rapidly to 3 times the 24-hour level at 48 hours. The results thus show that the expression of alpha-synuclein was significantly increased by the exposure time.

Table 1. List of primers

Primer	Sequence (sense/antisense)
Alpha-synuclein	5'- AGT GGA GGG AGC TGG GAA TA TAG-3'
	5'-TCC TCA CCC TTG CCC ATC T-3'
p53	5'-AGC GCT GCT CCG ATG GT-3'
	5'-TTC CTT CCA CCC GGA TAA GA-3'
Mouse β-actin	5'-GGC CAA CCG TGA AAA GAT GA-3'
	5'-CAG CCT GGA TGG CTA CGT ACA-3'

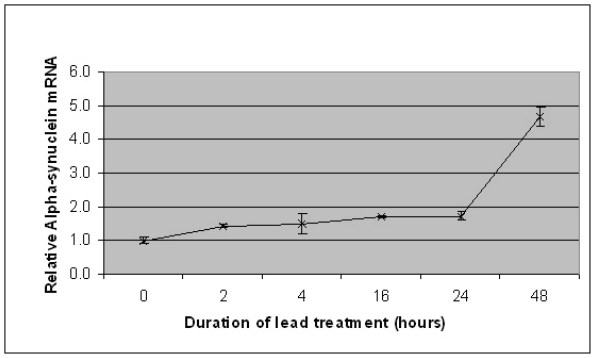


Figure 1. Alpha-synuclein mRNA in ATDC5 cells. ATDC5 cells were treated with 200 µM lead nitrate for duration of 0, 2, 4, 16, 24 and 48 hours. The results as expressed as mean ± SD Psmc3 mRNA amount relative to beta-actin for three samples.

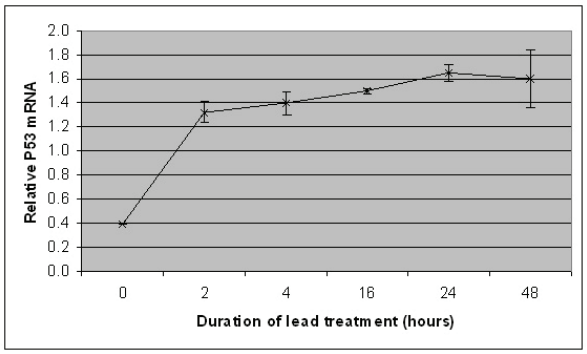


Figure 2. p53 mRNA in ATDC5 cells. ATDC5 cells were treated with 200 µM lead nitrate for duration of 0, 2, 4, 16, 24 and 48 hours. The results as expressed as mean ± SD p53 mRNA amount relative to beta-actin for three samples.

p53 mRNA- Expression of p53 mRNA in ATDC5 cells was also significantly increased by the exposure time. By 2 hours, lead exposure had rapidly increased the p53 mRNA level to more than 3.4 times the basal amount. (**Figure 2**) There was a moderate further increase in mRNA expression between 2 and 24 hours, by about 25%, and then expression decreased very slightly at 48 hours, but the decrease was not statistically significant. The rising trend in p53 mRNA level with an increase in exposure time was similar to that of alpha-synuclein mRNA. The levels of alpha-synuclein and p53 mRNA were correlated with each other ($r=0.9830$; $P<0.001$).

4. DISCUSSION

We have demonstrated that expression of alpha-synuclein and p53 mRNA increased with increasing duration of lead exposure in ATDC5 cells and that there was a correlation between alpha-synuclein and p53 mRNA expression after lead treatment. Although we did not examine the expression or localization of the corresponding proteins, our findings suggest that alpha-synuclein and p53 expression are stimulated by lead treatment in a coordinated way, which may reflect a shared cellular response to biological effects of lead exposure such as neurotoxicity, genotoxicity and oxida-

tive stress. In this study, alpha-synuclein was assumed to be a marker of increased likelihood of PD development and p53 was a marker of decreased cancer potential. Our results thus suggest a molecular mechanism for the inverse epidemiological association observed between PD and cancer (1).

Alpha-synuclein aggregation appears to play a major role in Lewy body formation and PD (9) (10), and occupational exposure to lead is a risk factor for PD (4). In addition, lead can form inclusion bodies in renal cells of poisoned humans or animals (11) and lead is a known potent neurotoxicant. Evidence indicates that lead exposure early in life may later on cause neurodegenerative disease such as Alzheimer's disease (12). Indeed, Alzheimer's disease is similar to PD, in that both are caused by a progression of amyloidogenesis in the brain. The link between alpha-synuclein and p53 expression on lead exposure demonstrated in this study suggests that alpha-synuclein protein expression and aggregation in cells can be stimulated by lead and contribute to inclusion body formation seen in cells of PD patients, while p53 protein accumulates to prevent cellular and genetic damage due to increased oxidative stress.

Because the p53 transcription factor regulates the cell cycle and apoptosis, it has been described as "the guardian of the genome," referring to its tumor suppressor role in conserving genomic stability by preventing the accumulation of mutations (13). An increase in p53 expression during lead exposure would result in an increased likelihood of cell cycle arrest to allow for genomic repair, or apoptosis if the genome is irreparable, and a similar mechanism might explain the possible lowered risk of tumorigenesis among PD patients. Indeed, lead can induce cell apoptosis via p53 in culture cells (6), and doing so would decrease the genotoxic potential of lead and hence the likelihood of cancer transformation. Another known role of p53 is to induce differentiation, which is another potential mechanism for preventing tumorigenesis. Interestingly, lead ion (Pb^{2+})-induced neurotoxicity may also be partially mediated through p53-independent apoptosis that is enhanced by glutamate (14), thereby again lowering tumorigenic potential.

A link between alpha-synuclein and p53 function has been previously demonstrated. In neuronal cell cultures, alpha-synuclein reduced the ability of cells to apoptose with and without the apoptotic trigger of staurosporine, and also reduced p53 expression and transcriptional activity (15). Both of the p53 expression and transcriptional activity was tested 48 hours after transfection. However, the dopamine-derived drug 6-hydroxydopamine reversed these effects and increased alpha-synuclein aggregation. The present data indicated p53 transcriptional activity was increased in 48 hours as well as alpha-synuclein transcriptional activity after lead treatment. Although we did not test for apoptotic status and immunoreactivity in our study, it is conceivable that alpha-synuclein has an effect on p53 expression and function and that lead treatment promotes alpha-synuclein aggregation, thereby resulting in p53 expression and related downstream cellular events. This series of events may explain lead's neurotoxicity and parallels the proposed role of the natural toxin 6-hydroxydopamine in alpha-synuclein aggregation in the etiology of PD (15).

The biological mechanism connecting alpha-synuclein, p53, lead exposure, PD and cancer does not appear to be simple. Epidemiological studies have shown that the odds of only non-smoking-related, but not smoking-related, cancer were lowered among patients with PD (1). The paradox is that smoking increases a human's intake of lead. Yet, there is a weak association between stomach and lung cancer frequency and an individual's exposure to lead (16), and a small but statistically significant increase in mortality has been found among employees at lead battery plants and lead smelters (17). In contrast, these employees' mortality from kidney cancer, bladder cancer, cancer of the central nervous system, lymphatic cancer and hematopoietic cancer was not increased (17). Finally, although both PD and cancer can be viewed as diseases of aging, laboratory studies have reported that p53 can uncouple cancer and aging by regulating both in a mutually exclusive way-namely, p53 hyperactivation in mice reduced the risk of spontaneous cancer but accelerated organismal aging (3). Organismal aging may be related by p53's ability to induce apoptosis and, on prolongued activation, to induce cellular terminal cell cycle block or "senescence," and may be controlled by subcellular localization (18). We will use of a neuronal cell line to do assays for protein expression, phenotypes such as apoptosis, differentiation, cell division, cell cycle, etc give better information.

Our findings provide some insight into the association between PD and cancer, and the behavior of cells in response to lead exposure. We propose that lead's neurotoxicity in PD is caused by alpha-synuclein expression and aggregation, which releases the inhibitory influence of alpha-synuclein on p53 expression and allows p53 to act as the cell's guardian of the genome, thereby reducing tumorigenic potential. It is possible that in the absence of lead, other triggers of alpha-synuclein expression and aggregation are involved in promoting p53 expression in the etiology of PD and other age-related neurodegenerative diseases. The results also suggest that treatments for PD based on preventing alpha-synuclein aggregation need to take into account the possible side effect of reducing p53 expression and function, and increasing tumorigenic potential. Treatments for cancer based on p53 expression also need to take into account the side effect of aging if p53 is allowed to be hyperactivated.

5. DISCLOSURE STATEMENT

The authors have declared that no competing interests exist.

ACKNOWLEDGMENTS

The authors thank Dr. Trevor Lane for critical evaluation of this manuscript. This research was supported by research grant of Professor Dr. A. Bakr M. Rabie.

REFERENCES

[1] A. B. West, V. L. Dawson and T. M. Dawson, (2005) *Trends Neurosci* 28, 348−352.
[2] O. M. El−Agnaf, S. A. Salem, K. E. Paleologou, M. D. Curran, M. J. Gibson, J. A. Court, M. G. Schlossmacher and D. Allsop, (2006) *FASEB J* 20, 419−425.
[3] S. D. Tyner, S. Venkatachalam, J. Choi, S. Jones, N. Ghebranious, H. Igelmann, X. Lu, G. Soron, B.Cooper, C. Brayton, S. Hee Park, T.Thompson, G. Karsenty, A. Bradley, and L. A. Donehower, (2002) *Nature* 415, 45−53.
[4] S. Coon, A. Stark, E. Peterson, A. Gloi, G. Kortsha, J. Pounds, D. Chettle and J. Gorell, (2006) *Environ Health Perspect* 114, 1872−1876.
[5] A. G. Osman, I. A. Mekkawy, J. Verreth, S. Wuertz, W. Kloas, and F. Kirschbaum, (2008) *Environ Toxicol*.
[6] J. Xu, L. D. Ji and L. H. Xu, (2006) *Toxicol Lett* 166, 160−167.
[7] T. Atsumi, Y. Miwa, K. Kimata and Y. Ikawa, (1990) *Cell Differ Dev* 30, 109−116.
[8] C. Shukunami, C. Shigeno, T. Atsumi, K. Ishizeki, F. Suzuki and Y. Hiraki, (1996) *J Cell Biol* 133, 457−468.
[9] P. Jenner and C. W. Olanow, (1998) *Ann Neurol* 44, S72−84.
[10] W. Zhou and C. R. Freed, (2004) *J Biol Chem* 279, 10128−10135.
[11] W. Qu, B. A. Diwan, J. Liu, R. A. Goyer, T. Dawson, J. L. Horton, M. G. Cherian and M. P. Waalkes, (2002) *Am J Pathol* 160, 1047−1056.
[12] L. D. White, D. A. Cory−Slechta, M. E. Gilbert, E. Tiffany−Castiglioni, N. H. Zawia, M. Virgolini, A. Rossi−George, S. M. Lasley, Y. C. Qian and M. R. Basha, (2007) *Toxicol Appl Pharmacol* 225, 1−27.
[13] S. Bates, A. C. Phillips, P. A. Clark, F. Stott, G. Peters, R. L. Ludwig and K. H. Vousden, (1998) *Nature* 395, 124−125.
[14] J. Loikkanen, K. Chvalova, J. Naarala, K. H. Vahakangas and K. M. Savolainen, (2003) *Toxicol Lett* 144, 235−246.
[15] C. Alves Da Costa, E. Paitel, B. Vincent and F. Checler, (2002) *J Biol Chem* 277, 50980−50984.
[16] H. Fu and P. Boffetta, (1995) *Occup Environ Med* 52, 73−81.
[17] O. Wong and F. Harris, (2000) *Am J Ind Med* 38, 255−270.
[18] J. Wesierska−Gadek and G.Schmid, (2005) *Cell Mol Biol Lett* 10, 439−453.

J. Biomedical Science and Engineering, 2009, 2, 90-95

Automatic detection and boundary estimation of optic disk in fundus images using geometric active contours

Giri Babu Kande[1], T. Satya Savithri[2], P. Venkata Subbaiah[3], M. R. N. Tagore[4]

[1]Vasireddy Venkatadri Institute of Technology. [2]Jawaharlal Nehru Technological University, Hyderabad, India. [3]Amrita Sai Institute of Science & Technology, Paritala, India. [4]Vasireddy Venkatadri Institute of Technology, Nambur, India; Correspondence should be addressed to Giri Babu Kande (kgiribabu @yahoo.com).

Received Nov. 14th, 2008; revised Dec. 29th, 2008; accepted Jan. 18th, 2009.

ABSTRACT

This paper proposes two efficient approaches for automatic detection and boundary estimation of optic disk in ocular fundus images. The proposed approach for optic disk detection uses the vessel branch with the most vessels to localize the optic disk. The boundary detection algorithm involves two steps; first, the color morphology in Lab space is used to have homogeneous optic disk region, then the boundary of the optic disk is estimated by using geometric active contour with new variational formulation. The success rates of disk localization and disk boundary detection are 99.7% and 96.95% respectively.

Keywords: Fundus, Geometric Active Contour, Optic disk, Retina.

1. INTRODUCTION

Optic disk (OD) is characterized as bright yellowish disk, from which, blood vessels and optic nerves emerge. The location of the optic disk is an important issue in retinal image analysis as it is a significant landmark feature and its diameter is usually used as a reference length for measuring distances and sizes. Precise localization of optic disc boundary is very useful in proliferative diabetic retinopathy, where fragile vessels develop in the retina, largely in the OD region, in response to circulation problems created during earlier stages of the disease. If the optic disc is identified, the position of areas of clinical importance such as the fovea may be determined.

Many schemes have been proposed to detect OD. Early detection schemes [1] and [2] aim simply to find the largest clusters of pixels with the highest intensities, while meets their difficulty when large hard exudates coexist in retinal image. Differentiating the two became a challenge. The area with the highest intensity variation of adjacent pixels was identified as optic disk in [3]. These methods could obtain satisfactory result in normal retinal images where optic disk is obvious and brightest. Only in [4], images with small lesions were considered. These methods will lead to the wrong disk localization when there are large areas of bright lesions similar to optic disk in an image. Principal component analysis [5] can locate optic disk even for retinal images having bright lesions. But this method is very slow. The algorithm in [6] uses a cost function, which is based on a combination of both global and local cues, to find the location of optic disc. A hybrid approach which uses properties of both appearance and model-based approaches is proposed to locate optic disk in [7] and in [8] local fractal analysis is used to find optic disk.

The contour of optic disk was estimated as a circle or an ellipse in [1, 2] and [4], because the shape of optic disk is round or vertically slightly oval. In one approach, Hough transform was employed to obtain the estimated circle of optic disk based on the result of edge detection [1, 2]. In another approach, optic disk contours were estimated by the Hausdorff based matching between the detected edges and the template of circle with different sizes [4]. Estimating the shape of optic disk as a circle or an ellipse cannot provide enough information to the ophthalmologists. As the shape of optic disk is important to diagnose eye diseases, the exact boundary detection of optic disk has been investigated. “Snakes” was applied to detect the exact contour of optic disk in [9, 10, 11]. The major advantage of these algorithms is their ability to bridge discontinuities in the image feature being located. However, the algorithms were sensitive to the preprocessing and the methods proposed in the above papers were not fully automatic due to the requirement of manual initialization. The main difficulty to apply these methods to disk boundary detection is how to remove the influence of blood vessels. A modified active shape model is proposed in [5] to detect the shape of optic disk. The algorithms in [12,13] uses gradient vector flows to detect the boundary of optic disk.

In this paper, we propose a new algorithm to effi-

ciently localize the optic disk in ocular fundus images. The proposed algorithm is based on finding the vessel branch having more number of blood vessels. Our method achieves more accurate results compared to [1], [2] and [3] with an accuracy of 99.7%. The proposed optic disk boundary detection algorithm involves two steps. First, colour mathematical morphology in Lab space is used to have homogenous optic disk region for the snake to lock onto. Then geometric active contours are used to detect the boundary of optic disk. The proposed method achieves more accurate results compared to [6] and [13] with an accuracy of 96.95%.

The paper is organized as follows. In Section 2.1, localization of optic disk based on the vessel branch with most vessels is described. Section 2.2 presents the proposed geometric active contour with variational formulation to detect the boundary of the optic disk. In Section 3, the experimental results of the proposed algorithms are presented and compared to existing methods. Discussion and conclusions are in Section 4.

2. METHODOLOGY

2.1. Localization of Optic Disk Based on the Branch with the Most Vessels

Optic disk is the entrance region of blood vessels and optic nerves to the retina. It is a significant anatomic landmark for the detection of other features and its dimensions are often studied for a clue of some diseases as well. The method of optic disk localization by finding the largest cluster of brightest pixels is simple, fast and works well in the normal retinal images, but it can not locate optic disk correctly in the images where the area of bright lesions is large or optic disk is obscured by blood vessels. The method based on finding the branch with the most vessels is proposed to localize optic disk in this paper.

The vasculature in the retinal image consists of many vessels of various lengths and various thicknesses. The proposed method turns the vessel probability map into a network of vessels and branches. In this method network information is stored about the connections between vessels and branches. By means of this network, the branch with most vessels connected to it can be selected. The selected branch is used to determine the optic disc.

We obtained the binary blood vessel skeleton map from [14]. For each skeleton pixel the amount of neighbor skeleton pixels is determined. If the amount of neighbors is smaller than three, then the pixel is added to the vessel-image. Otherwise the pixel is added to the branch-image. A vessel is a collection of points, starting at a point with only one neighbor and ending at a point with only one neighbor. The vessels are detected as follows: If a begin point of a vessel is detected within the vessel-image, then the vessel is traced towards its endpoint. The begin point and the endpoint are marked to avoid tracing a vessel twice. The branches are detected by applying eight-connected component analysis on the branch-image.

The constructed vessel-branch network can be used to find the optic disc in many ways. A very simple algorithm is the selection of the branch with most vessels. For each branch of the network the amount of vessels connected to it is stored. The optic disc contains the optic nerve from which a few main vessels split up into many smaller vessels which spread around the retina. Vessel segments in this area of the retina are often small and are therefore often combined into one large branch of the network with many vessel objects connected to it. An increasing amount of vessel connections of a branch also increases the probability of the branch being located in the optic disc area. The algorithm is performed as follows:

1. Select the branch with the most vessel connections. If there are several branches with the highest number of vessel connections, then the branch with the most branch pixels is selected.
2. Take the bounding box of the branch with the most vessel connections.
3. Select the center of the bounding box as the OD-center.

For example, given a color retinal image in **Figure 1(a)**, the binary vasculature and the skeleton of the vasculature overlaid on the retina image are shown in **Figure 1(b)** and **Figure 1(c)**. The detected center of the optic disk is as shown in **Figure 1(e)** where the vessel branch with the most vessels is used to localize the optic disk.

2.2. Boundary Detection of Optic Disk by Geometric Active Contour Model

(a) (b) (c) (d) (e)

Figure 1. Localization of optic disk (a) original Color retina image (b) The binary vessel map (c) Overlay of the vessel branch network and the original image (d) The bounding box of the best branch and the determination of the optic disk center and (e) Detected optic disk center.

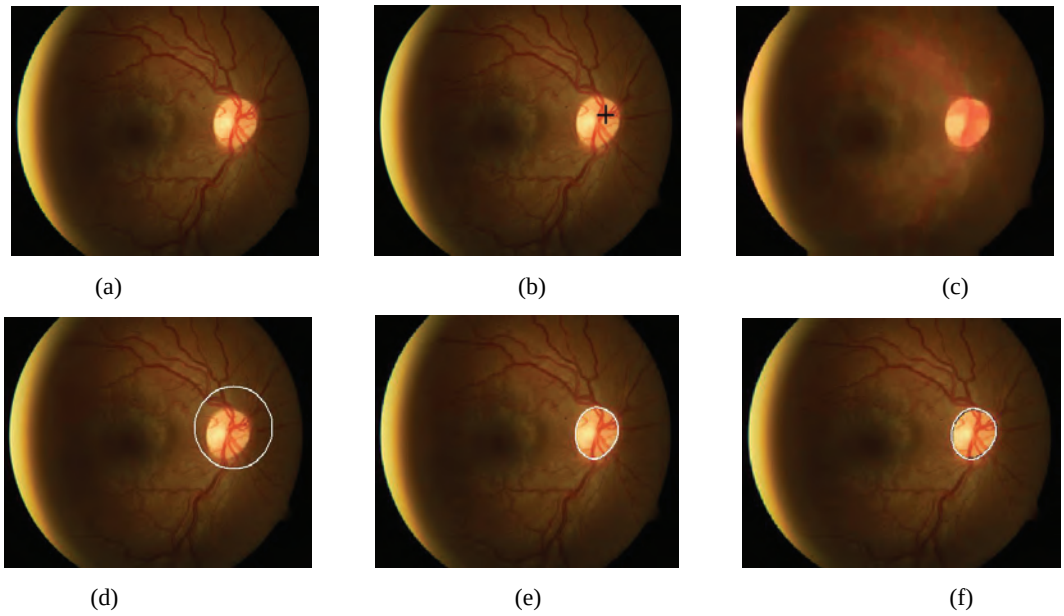


Figure 2. Boundary detection of optic disk (a) Color retina image (b) located optic disk (c) Result after applying color morphology in L*ab* space (d) Initial snake (e) Detected optic disk boundary and (f) Overlay of the ground-truth and the result of proposed method.

Change in the shape, color or depth of optic disk is an indicator of various ophthalmic pathologies especially for glaucoma. The accurate detection of the optic disk boundary can be used to assess the progress of eye disease and the treatment results. Some parts of the disk boundary are not well defined and some parts are partly obscured by the blood vessels in retinal images, which make the detection of disk shape complicated. A geometric active contour model is proposed to detect the disk boundary in retinal images.

Firstly, the original color retinal image is preprocessed using color mathematical morphology in L*ab* space. This helps to remove blood vessels more cleanly and provides a more homogeneous optic disc region for the geometric active contour to lock onto. We performed dilation first to remove the blood vessels in optic disk region and then an erosion to restore the boundaries to their former position. The morphology in L*ab* space is performed using the method described in [15].

For each arbitrary point x in the color space, the definitions for dilation (I_d) and erosion (I_e) by structuring element K is defined as:

$$I_d(x)=\{I(y):I(y)=\max[I(z)],Z\in K_x\}$$
$$I_e(x)=\{I(y):I(y)=\min[I(z)],Z\in K_x\} \quad (1)$$

We always used a symmetrical disc structuring element of size 13, since the blood vessels were determined to be not wider than 11 pixels. For the color retinal image shown in **Figure 2(a)**, the L*ab* morphology output is as shown in **Figure 2(c)** where the blood vessels are removed cleanly to have homogenous optic disk region.

The optic disk boundary is determined by fitting a geometric active contour model with variational formulation. The initial contour for a snake must be close to the desired boundary otherwise it can converge to the wrong resting place. We used the method described for localizing the optic disc to automatically position an initial snake. The initial contour for the image shown in **Figure 2(a)** is as given in **Figure 2(d)**. In general, a snake is a set of points initially placed near the contour of interest, which are gradually brought closer to the exact shape of the desired region in the image. This is carried out through iterative minimization of an energy function comprising an internal and an external term:

$$\varepsilon(\phi)=\mu P(\phi)+\varepsilon_{g,\lambda,v}(\phi) \quad (2)$$

where $P(\phi)=\int_\Omega \frac{1}{2}(|\nabla\phi|-1)^2\,dxdy$ a metric to characterize how close a function Φ is to a signed distance function in $\Omega\subset\Re^2$, μ>0 is a parameter controlling the effect of penalizing the deviation of Φ from assigned distance function. $\varepsilon_{g,\lambda,v}(\phi)$ is the external energy for a function Φ(x,y) and it is defined in [18] as

$$\varepsilon_{g,\lambda,v}(\phi)=\lambda L_g(\phi)+vA_g(\phi) \quad (3)$$

where $\lambda > 0$ and v are constants, g is an edge indicator function defined for an image I as

$$g=\frac{1}{1+|\nabla G_\sigma * I|^2} \quad (4)$$

where G_σ is the Gaussian kernel with standard deviation σ. The terms $L_g(\phi)$ and $A_g(\phi)$ are defined as

$$L_g(\phi) = \int_\Omega g\delta(\phi)|\nabla\phi|\,dxdy \quad (5)$$

and

$$A_g(\phi) = \int_\Omega gH(-\phi)|\nabla\phi|\,dxdy, \quad (6)$$

respectively, where δ is the univariate Dirac function, and H is the Heaviside function. The external energy $\varepsilon_{g,\lambda,v}$ drives the zero level set toward the object boundaries, while the internal energy $\mu P(\phi)$ penalizes the deviation of ϕ from a signed distance function during

its evolution. By calculus of variations, the Gateaux derivative (first variation) of the functional ε in (2) can be written as

$$\frac{\partial \varepsilon}{\partial \phi} = -\mu[\nabla \phi - div\left(\frac{\nabla \phi}{|\nabla \phi|}\right) - \lambda\delta(\phi) div\left(g\frac{\nabla \phi}{|\nabla \phi|}\right) - vg\delta(\phi) \quad (7)$$

where ∇ is the Laplacian operator. Therefore, the function ϕ that minimizes this functional satisfies the Euler – Lagrange equation $\frac{\partial \varepsilon}{\partial \phi} = 0$. The steepest descent process for minimization of the functional ε is the following gradient flow:

$$\frac{\partial \phi}{\partial t} = \mu\left[\nabla \phi - div\left(\frac{\nabla \phi}{|\nabla \phi|}\right)\right] + \lambda\delta(\phi) div\left(g\frac{\nabla \phi}{|\nabla \phi|}\right) + vg\delta(\phi) \quad (8)$$

This gradient flow is the evolution equation of the level set function in the proposed method. The second and the third term in the right hand side of (8) correspond to the gradient flows of the energy functional $\lambda L_g(\phi)$ and $vA_g(\phi)$, respectively, and are responsible of driving the zero level curve towards the object boundaries. The detected optic disk boundary by the proposed method for the image in **Figure 2(a)** is as shown in **Figure 2(e)** and **Figure 2(f)** shows the superimposition of the result of proposed method on the hand labeled ground-truth image.

3. EXPERIMENTAL RESULTS

The proposed algorithms are tested and evaluated on four publicly available databases of color retinal images: STARE [17], DRIVE [18], DIARETDB0 [19], and DIARETDB1 [20] databases. The DIARETDB0 database consists of 130 color retinal images of size 1500 × 1152. The DIARETDB1 database consists of 89 color retinal images of size 1500 × 1152. The DRIVE database contains 40 color images divided into 20 training and 20 test images. The downloaded images were of size 565×584. The STARE database consists of 81 slides which were digitized to 700 × 605 pixels, 8 bits per color channel.

3.1. Localization of Optic Disk

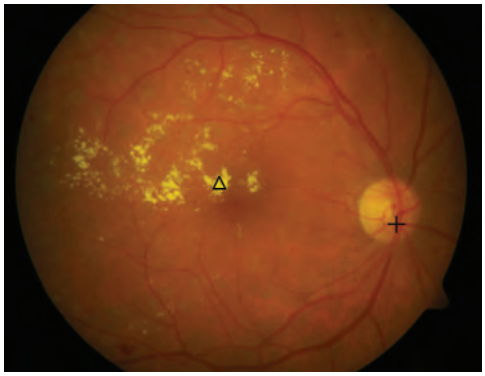


Figure 3. Comparison of optic disk localization with maximum local variation method [3].

Table 1. Performance comparison of maximum local variation method [3] and the proposed method.

Database	Number of Images	Success rate in %	
		maximum local variation method	Proposed Method
STARE	81	79	99
DRIVE	40	100	100
DIARETDB0	130	89	98
DIARETDB1	89	84	100

Compared with the localization of optic disk by the centroid of the largest cluster of the brightest pixels [1], [2] the proposed algorithm achieves more accurate results. An example is shown in **Figure 3**, where "+" indicates the localization by the proposed algorithm of finding the branch with most vessels, "Δ" represents the localization by the centroid of the largest cluster of the brightest pixels [3]. The method in [3] gives the wrong localization when processing the retinal images with large areas of lesions, while the proposed method can obtain the correct localization. The proposed algorithm works pretty well even though the input retinal image is in a low-contrast condition. The success rate of optic disk locating process is 99.7% based on the images tested.

3.2. Boundary Detection of Optic Disk

In order to evaluate the performance of our algorithm for detecting optic disk boundary, we compare results of the proposed algorithm with the state-of-the-art results obtained from GVF snake method [13], 2D Circular Hough Transform method and hand-labeled ground-truth segmentations. In GVF-snake, the images are preprocessed by morphological operation; and the parameters in the energy functions were carefully set to make a balance between the smoothness and the accuracy on the resulted boundary. In 2D Circular Hough Transform method, the dimensions of the normal circular Hough Transform histogram are reduced from 3 to 2 dimensions by assuming that the approximate OD radius is known. Only the first few circles are evaluated by using the maximum point from Hough space. The disk boundary manually marked by the experienced ophthalmologist is set to be the ground-truth. Then we use a simple and effective overlap measure to evaluate the accuracy of the detected boundary.

$$M = \frac{n(R \cap T)}{n(R \cup T)} \quad (9)$$

where R and T correspond to the ground-truth and the detected optic disk region respectively and n(.) is the number of pixels in a region.

Specifically, we classify retinal images into three categories, normal retinal images, abnormal retinal images with ill-defined optic disk, and retinal images with fuzzy elliptic optic disk. From column by column, the first column of **Figure 4** presents the results obtained from normal retinal images. Results of abnormal retinal images with ill-defined optic disk are shown on the sec-

ond column. The last column presents results from retinal images with fuzzy elliptic optic disk. From row by row, the first row shows original retinal images. The second row shows the results of GVF snake. The third row presents simulation results of Hough transform [6]. The last row presents our simulation results.

An example of normal retinal image having elliptic optic disk is illustrated in first column of **Figure 4**, where both the proposed method and GVF snake give the successful results; Hough transform gives the failed result. The measured accuracies for the GVF snake, Hough transform and the proposed method are 99.3%, 78% and 99.5% respectively. The example given in the second column is an optic disk with ill-defined boundary and noises from the surrounding tissue. The proposed method correctly located the disk boundary, while the GVF snake and Hough transform methods failed. The accuracies for the GVF snake and Hough transform are 72% and 74% respectively. For the same image the proposed method has an accuracy of 99.1%. One more example of fuzzy elliptic optic disk is illustrated in the third column. The accuracies for the GVF snake and Hough transform are 81% and 82% respectively. For the same image the proposed method has an accuracy of 98.2%. **Table 2** compares our approach with the GVF nake and Hough transform methods in terms of accu racy for STARE, DRIVE, DIARETDB0, DIARETDB1 databases. It can be obviously seen that the proposed method provides better result.

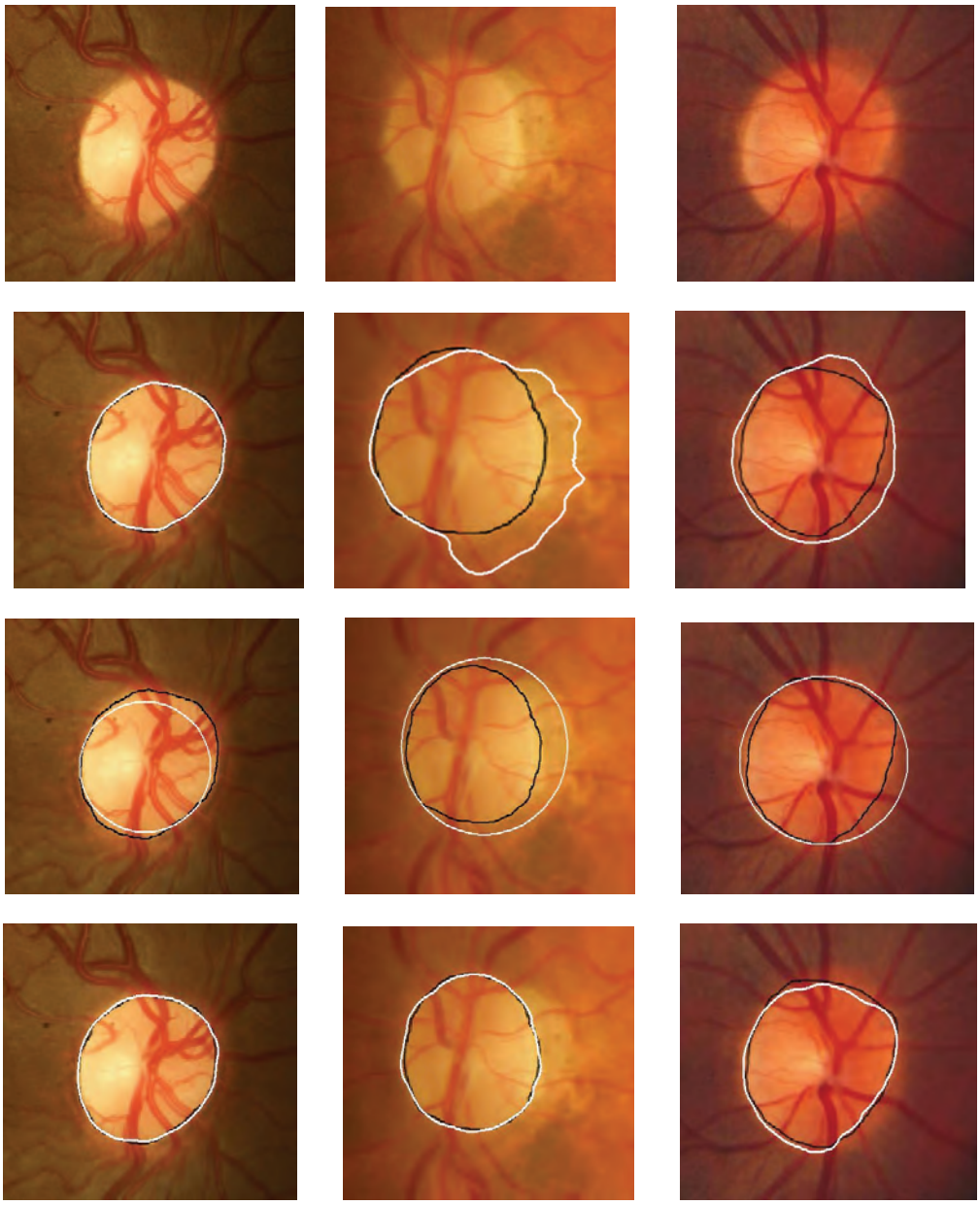


Figure 4. First row: Example images with closer view of optic disk; Second row: Results from GVF Snake; Third row: Results from Hough Transform; Last row: Results from our method.

Table 2. Comparison of Average Accuracy for detecting the boundary of optic disk.

Database	Number of Images	Average Accuracy in%		
		Hough Transform	GVF Snake	Proposed Method
STARE	81	83.8	89.8	94.2
DRIVE	40	93.5	98.6	99.3
DIARETDB0	130	85.4	95.1	96.8
DIARETDB1	89	89.2	95.4	97.5

4. CONCLUSION

In this paper, two algorithms are presented for localizing the optic disk and detecting the boundary of optic disk. The proposed algorithms are implemented using MATLAB 7.0 on a core 2 Duo 1.8 GHz PC with 1GB memory and the time taken for detection and boundary extraction is approximately 25 sec/image. The algorithm for optic disk localization is based on finding the vessel branch with most vessels. When compared to other methods [1] [2] [5] [3] the proposed algorithm is fast and can locate the optic disk accurately even though the retina image contains large areas of bright lesions. In our experiments, the PCA based method [3] failed to locate optic disk in retinal images that contains large area of lesions around the optic disk as there is no such case in the training set. The proposed boundary detection algorithm uses color morphology and geometric active contour with new variational formulation to estimate the contour of optic disk. The color mathematical morphology in L*ab* space is used to have homogenous optic disk region. Then the boundary of the optic disk is located by using geometric active contour with new variational formulation. The results were compared with the results from Hough transform method, GVF snake method and validated against experienced ophthalmologist's hand drawn ground truth. The average accuracy result of the proposed method for STARE, DRIVE, DIARETDB0 and DIARETDB1 databases is quite successful with accuracy of 96.95% compared to the accuracy results of Hough transform method and GVF snake method which are 87.97% and 94.72% respectively . One visible advantage of this method is that the ODs are detected even though the boundary of the OD is not continuous or blurred.

REFERENCES

[1] S. Tamura, Y. Okamoto, and K. Yanashima (1988) Zero crossing interval correction in tracking eye-fundus blood vessels, Pattern Recogn., 21, 227–233.

[2] Z. Liu, O. Chutatape, and S. M. Krishnan (1997) Automatic image analysis of fundus photograph, Proc. 19th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society, 2, 524–525.

[3] C. Sinthanayothin, J. F. Boyce, H. L. Cook, and T. H.Williamson (1999) Automated location of the optic disk, fovea, and retinal blood vessels from digital color fundus images, Br. J. Ophthalmol., 83, 902–910.

[4] M. Lalonde, M. Beaulieu, and L. Gagnon (2001) Fast and robust optic disk detection using pyramidal decomposition and Hausdorff-based template matching, IEEE Trans. Med. Imag., 20, 1193–1200.

[5] H. Li, O.Chutatape, (2004) Automated Feature Extraction in Color Retinal Images by a Model Based Approach IEEE trans. In Biomedical Engineering, 51, 246–254.

[6] M. Niemeijer, M. D. Abràmoff, and B. van Ginneken (2007) Segmentation of the Optic Disc, Macula and Vascular Arch in Fundus Photographs, IEEE trans. on medical imaging, 26, 116–127.

[7] G.D.Joshi, G.Vidhyadhari, J. Sivaswamy (2008) Optic disk detection using topographical features Proc. International EURASIP conference (BIOSIGNAL).

[8] H.Ying, M.Zhang, J.C.Liu, (2007) Fractal-based Automatic Localization and Segmentation of Optic Disc in Retinal Images, Proc. of the international conference of IEEE Engineering in Medicine and Biology Society, 29, 4139–4141.

[9] S. Lee and L. M. Brady. (1991) Integrating stereo and photometric stereo to monitor the development of glaucoma, in Image and Vision computing, 9, 39–44.

[10] D. T. Morris and C. Donnison. (1999) Identifying the neuroretinal rim boundary using dynamic contours, Image Vis. Computing, 17, 169–174.

[11] F. Mendels, C. Heneghan, and J. P. Thiran. (1999) Identification of the optic disk boundary in retinal images using active contours, in Proc. Irish Machine Vision and Image Processing Conf., 103–115.

[12] V. Thongnuch and B. Uyyanonvara (2007) Automatic optic disk detection from low contrast retinal images of ROP infant using GVF snake, Suranaree Journal. Sci. Technol. 14,223–226.

[13] A. Osareh, M. Mirmehdi, B. Thomas, and R. Markham (2002) Comparison of colour spaces for optic disc localisation in retinal images,. IEEE 16th International Conference on Pattern Recognition, 1, 743.746.

[14] G. B. Kande, T. S. Savithri, P. V. Subbaiah (2008) Retinal Vessel Segmentation using Local Relative Entropy Thresholding , Proc. IEEE Int. Conf. on Systems, Man and Cybernetics, 8, 3448-3453.

[15] A. Hanbury, J. Serra (2001) Mathematical Morphology in the Lab Colour Space , Proc. Int. Conf. on Stereology

[16] C. M. Li, C. Y. Xu, C. F. Gui, and M. D. Fox (2005) Level Set Evolution Without Reinitialization: A New Variational Formulation, Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, 1, 430-436.

[17] A. Hoover and M. Goldbaum (2003) Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels, IEEE Trans. on Medical Imaging, 22, 951- 958.

[18] M. Niemeijer, J. J. Staal, B. van Ginneken, M. Loog, M. D. Abramoff (2004) Comparative study of retinal vessel segmentation methods on a new publicly available database, in: SPIE Medical Imaging, 5370, 648-656.

[19] T. Kauppi, V. Kalesnykiene, J. K. Kamarainen, L. Lensu, I. Sorri, H. Uusitalo, H. Kälviäinen, J. Pietilä, DIARETDB0: Evaluation Database and Methodology for Diabetic Retinopathy Algorithms, Technical report.

[20] T. Kauppi, V. Kalesnykiene, J. K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kälviäinen, J. Pietilä, DIARETDB1 diabetic retinopathy database and evaluation protocol, Technical report.

J. Biomedical Science and Engineering, 2009, 2, 96-101

The effect of different number of diffusion gradients on SNR of diffusion tensor-derived measurement maps

Na Zhang[1], Zhen-Sheng Deng[1*], Fang Wang[1], Xiao-Yi Wang[2]

[1]Institute of Biomedical Engineering, School of Info-physics and Geomatics Engineering, Central South University, Changsha, Hunan, P.R. China (410083);[2]Department of Radiology, XiangYa Hospital of School of Medicine, Central South University, Changsha, Hunan, P.R. China (410008);*Corresponding author: Zhensheng Deng (bmedzs@csu.edu.cn or dengzhensheng@hotmail.com)

Received Dec. 12th, 2008; revised Feb. 12th, 2009; accepted Feb. 16th, 2009

ABSTRACT

Diffusion tensor imaging (DTI) is mainly applied to white matter fiber tracking in human brain, but there is still a debate on how many diffusion gradient directions should be used to get the best results. In this paper, the performance of 7 protocols corresponding to 6, 9, 12, 15, 20, 25, and 30 noncollinear number of diffusion gradient directions (NDGD) were discussed by comparing signal-noise ratio (SNR) of tensor- derived measurement maps and fractional anisotropy (FA) values.

All DTI data (eight healthy volunteers) were downloaded from the website of Johns Hopkins Medical Institute Laboratory of Brain Anatomical MRI with permission. FA, apparent diffusion constant mean (ADC-mean), the largest eigenvalue (LEV), and eigenvector orientation (EVO) maps associated with LEV of all subjects were calculated derived from tensor in the 7 protocols via DTI Studio. A method to estimate the variance was presented to calculate SNR of these tensor-derived maps. Mean ± standard deviation of the SNR and FA values within region of interest (ROI) selected in the white matter were compared among the 7 protocols.

The SNR were improved significantly with NDGD increasing from 6 to 20 (P<0.05). From 20 to 30, SNR were improved significantly for LEV and EVO maps (P<0.05), but no significant differences for FA and ADC-mean maps (P>0.05). There were no significant variances in FA values within ROI between any two protocols (P> 0.05).

The SNR could be improved with NDGD increasing, but an optimum protocol is needed because of clinical limitations.

Keywords: Diffusion Tensor Imaging, Diffusion Gradient, Signal Noise Ratio, Estimating Variance

1. INTRODUCTION

Diffusion tensor imaging (DTI) has emerged as a noninvasive magnetic resonance imaging (MRI) modality capable of providing in vivo fundamental information of the white matter structure, which is required for viewing structural connectivity in the human brain [1,2]. It is commonly used to demonstrate subtle abnormalities in a variety of diseases (including stroke, multiple sclerosis, dyslexia, and schizophrenia) and is currently becoming part of many routine clinical protocols [3]. The principle of DTI is based on diffusion anisotropy of water molecular. By acquiring diffusion weighted (DW) images with diffusion gradients oriented in at least six noncollinear directions (The tensor has 6 independent parameters, that is why the minimal number of diffusion gradient directions (NDGD) for DTI measurement is 6), it is possible to measure the diffusion tensor modeled by a 3 dimension (3D) ellipsoid in each voxel [4,5]. The tensor-derived matrices, like diffusion anisotropy maps and color-coded orientation maps, which could characterize specific features of the diffusion process, can be calculated from tensor via DTI Studio [6]. DTI always operates under the assumption of a single ellipsoid. For estimation of more complex geometries, high angular resolution diffusion imaging (HARDI) or Q-Ball Imaging needs to be used [7].

NDGD is one of the most important factors for DW images acquisition. As NDGD increasing, more DW images are used to calculate the diffusion tensor, resulting in more accurate tensor estimation but much longer imaging time. Considering signal-noise ratio (SNR), if the same amount of time is used to acquire DW images, one 6-diffusion gradients is used, and the other 12-diffusion gradients, in the former, more images are acquired in each direction, which results in more averages, but in the latter, less averages. So which is better for single 3D ellipsoid estimation is still open to debate.

Some researchers [8,9] claimed that more than 6 diffusion gradients can provide better measurements of the tensor than the conventional 6-diffusion gradients. A

recent study with Monte Carlo simulations[10] concluded that at least 20 NDGD were necessary for a robust estimation of diffusion anisotropy, whereas at least 30 NDGD were required for a robust estimation of tensor orientation and mean diffusivity. Ni *et al* [11] found that NDGD = 6 and number of excitations (NEX) = 10 were sufficient for estimation of FA values from region of interest (ROI) calculations. All these researchers mentioned above did their studies with the constraint of constant imaging time. Previous work by Poonawalla [12] concluded that when the acquisition time was held constant, the sum of the diffusion tensor variances decreased as NDGD increased, and signal averaging may not be as effective as increasing NDGD, especially when NDGD is small (e.g., NDGD < 13).

In this paper, the SNR of fractional anisotropy (FA) maps, apparent diffusion constant mean (ADC-mean) maps, the largest eigenvalue λ_1 (LEV) maps, and eigenvector orientation (EVO) maps associated with λ_1 derived from diffusion tensor and the FA values calculated from ROI were compared in 7 protocols corresponding to different NDGD (6, 9, 12, 15, 20, 25, and 30 noncollinear). Unlike in previous work where the NEX varied to keep imaging time constant, the number of images averaged (NEX = 3) is fixed in this work so as to all the 7 protocols have the same original SNR. So the imaging time for the 7 protocols was not held constant and the higher NDGD protocols would be expected to perform better given that the imaging time was greater.

The purpose of this work is to independently determine the effect of NDGD on SNR of these tensor-derived measurement maps mentioned above with the fixed NEX.

2. MATERIALS AND METHODS

All DTI data used in this paper were downloaded from the website of Johns Hopkins Medical Institute Laboratory of Brain Anatomical MRI with permission.

2.1. Subjects

All images were acquired in eight healthy volunteers (three females, five males; range, 21–29 years). The subjects did not have any history of neurological diseases. Institutional review board approval was obtained for the study, and informed consent was obtained from all subjects.

2.2. Data Acquisition

A 1.5T MR scanner (Gyroscan NT; Philips Medical Systems, Best, the Netherlands) was used. DTI data were acquired by using a single-shot echo-planar imaging sequence with 7 protocols corresponding to different NDGD (6, 9, 12, 15, 20, 25, and 30 noncollinear), and the *b* value was 700 smm^{-2}. The image matrix was 256 × 256 pixels, with a field of view of 246 × 220 mm (nominal resolution, 2.2 mm). Transverse sections of 2.2 mm thickness were acquired parallel to the anterior commissure–posterior commissure line. A total of 55 sections covered the entire hemisphere and brainstem without gaps. The acquisition time per dataset was approximately 6 minutes. All DW imaging were repeated 3 times, so they have the same original SNR. Five additional images for each slice with minimal DW (b_0=33 smm^{-2}) were also acquired, and all 7 DTI acquisitions have the same b_0 images.

2.3. Definitions of DTI Measurements

The ADC, which is used to characterize the water diffusion, can be calculated from the following Equation (1) [6].

$$ADC_k = \frac{\ln(S_k / S_0)}{b}, \text{ (k = 1, 2… K; K} \geq 6) \quad (1)$$

where, the constant *b* is the diffusion-weighting factor, S_0 is the signal obtained without diffusion gradient, and S_k is the signals corresponding to the different gradient directions (k=1, 2…K; K ≥ 6). ADC-mean can be calculated by averaging the set of ADC_k.

From the diffusion tensor, three eigenvalues, $\lambda_1>\lambda_2>\lambda_3$, which define the diffusion magnitude, can be determined by diagonalizing the tensor for each voxel. Three eigenvectors (associated with three eigenvalues), which describe the diffusivity in the three directions**,** can be calculated. Based on these three diffusivities, the FA commonly used for anisotropy definitions is calculated to yield values between 0 and 1 by the following Equation (2) [6].

$$FA = \sqrt{\frac{3}{2}} \frac{\sqrt{(\lambda_1 - \overline{D})^2 + (\lambda_2 - \overline{D})^2 + (\lambda_3 - \overline{D})^2}}{\sqrt{\lambda_1^{\,2} + \lambda_2^{\,2} + \lambda_3^{\,2}}} \quad (2)$$

where,

$$\overline{D} = \frac{\lambda_1 + \lambda_2 + \lambda_3}{3}$$

2.4. SNR Calculation

SNR measures the roughness or granularity of diffusion tensor-derived measurement maps, it should be equal to the ratio of power spectrum of signal to that of noise. But in general, spectrum analyze is not recommended to estimate the SNR of magnetic resonance (MR) images because it is actually re-created from frequency-signal or k-space. In addition, spectrum analyze is good for random signal (include random noise) analyze, this is not the case for the measurements derived from tensor calculation because the only resource of the random error during this tensor calculation comes from the finite bit length of the computers.

The method presented by Mouyan Zou [13] can be used to estimate approximately SNR of an image, which is the variance of signal divided by that of noise. According to the theory, local variance of all pixels of an image should be calculated, the maximum of the local variance which stands for the signal variance is divided by the minimum which stands for the noise variance, and the result (see Equation (3)) as the approximate SNR should be amended by empirical formula.

$$SNR=\frac{\sigma_{max}^{2}}{\sigma_{min}^{2}} \quad (3)$$

where, σ^2 is an estimated value of the local variance.

Since "local variance" affects the SNR measurement, the local neighborhood included 10×10 pixels as an kernel was used to calculate the local variance in this study. The kernel of larger or smaller than 10×10 pixels was not suggested because the former resulted in lower SNR and the latter resulted in higher SNR.

The SNR calculated for the base images, which were averaged thrice, is the same for all 7 protocols (SNR= 67.85) by using this method.

3. DATA PROCESSING

The DW images were transferred to a workstation and processed by employing DTI Studio [6] developed for diffusion tensor images calculating and fiber tracking. For each DTI dataset, the six independent elements of the 3×3 diffusion tensor were calculated for each voxel. After diagonalization, three eigenvalues, $\lambda_1>\lambda_2>\lambda_3$, and three eigenvectors were calculated for each voxel, and in turn, LEV maps and EVO maps associated with λ_1 were obtained. Then FA maps and ADC-mean maps were also obtained by using Equations (2) and (1). EVO maps associated with λ_1 were used as an indicator of fiber orientation. On the EVO maps, red, green, and blue colors were assigned to right-left, anterior-posterior, and superior- inferior orientations, respectively [14].

Based on MatLab platform, estimating variance as an approach was used to provide a global estimate of SNR for these tensor-derived measurement maps that characterizes the uncertainty of the DTI measurements mentioned above in the 7 protocols corresponding to 6, 9, 12, 15, 20, 25 and 30 noncollinear NDGD with a 700 mm^2/sec *b* value. To illustrate the effect of NDGD on FA values, we also calculated FA values within ROI (about 30 pixels) selected in the white matter from FA maps in

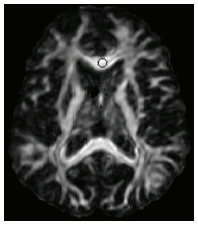


Figure 1. An FA map corresponding to 30 noncollinear diffusion gradient directions, with an ROI (about 30 pixels) in the white matter marked with a circle.

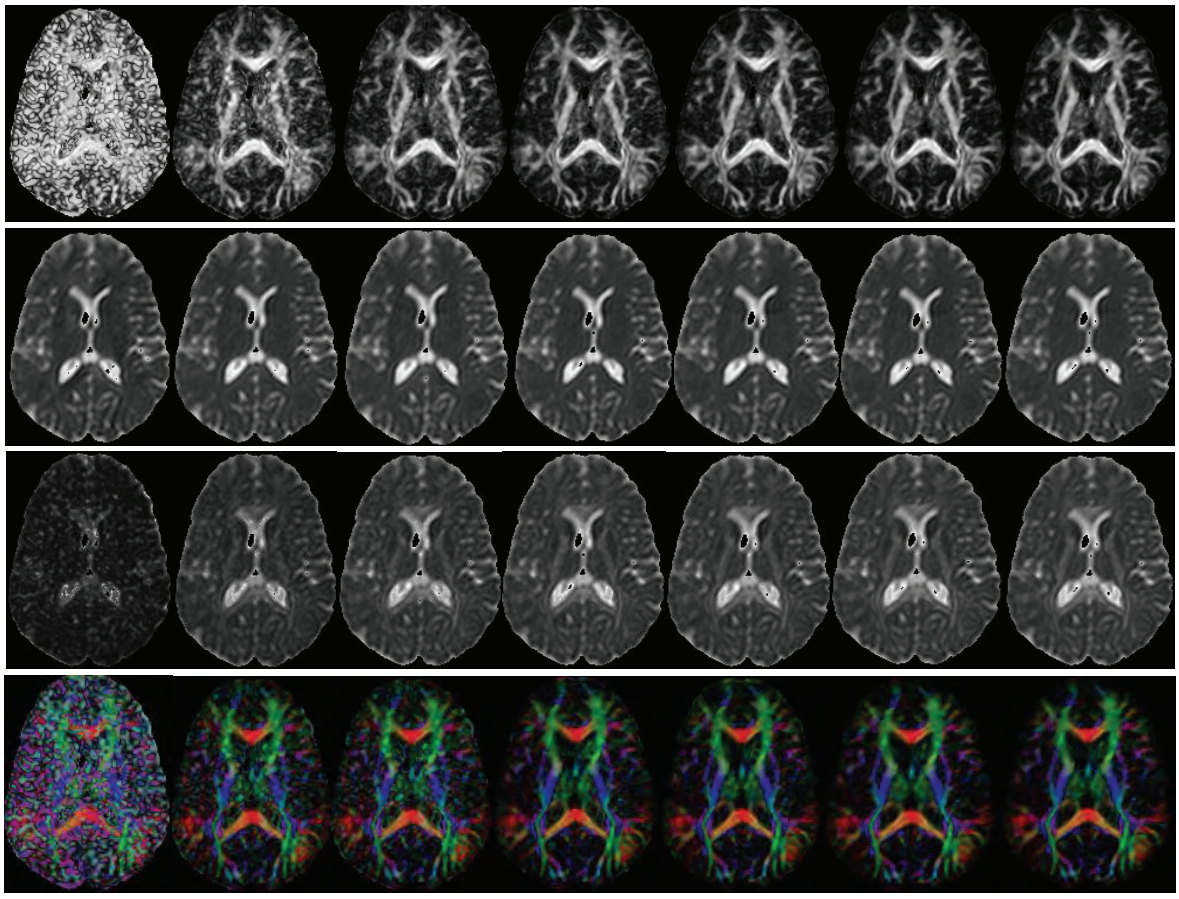


Figure 2. Tensor-derived measurement maps. The first, second, third, and forth rows are FA maps, ADC-mean maps, LEV maps, and color-coded maps for the eigenvectors associated with λ1, respectively, which corresponding to 6, 9, 12, 15, 20, 25, and 30 noncollinear NDGD arranged from left to right.

all 7 protocols. An FA map corresponding to 30 noncollinear NDGD with an ROI was shown in **Figure 1**.

An unpaired T-test for the SNR of these tensor-derived measurement maps and the FA values within ROI in the 7 protocols was performed by using SPSS11.5 software, a *P* value less than 0.05 for a measurement was considered as statistically significant.

4. RESULTS

The tensor-derived measurement maps of one subject acquired from tensor calculation in the 7 protocols were shown in **Figure 2**. The improvement of the SNR with the NDGD increasing could be observed visually. The mean ± standard deviation for the SNR of tensor-derived measurement maps and the FA values within ROI in the 7 protocols were listed in **Table 1**. The SNR of these tensor-derived measurement maps varied with the NDGD increasing were shown in **Figure 3**. The correlations between FA values within ROI and NDGD were fitted by linear lines and shown in **Figure 4**.

In order to predigest the results, the P values for the SNR of tensor-derived measurement maps and FA values within ROI in the 4 protocols corresponding to different NDGD (6, 12, 20, and 30 noncollinear) instead of all the 7 protocols were listed in **Table 2**.

From the curves in **Figure 3**, which demonstrate the SNR as a function of the NDGD for tensor-derived measurement maps, we note that the SNR of the tensor-derived measurement maps could be improved with more NDGD. Also it is obvious that the more NDGD were used, the higher SNR could be obtained from **Figure 2**. This is perfectly accordant with the findings of D.K. Jones *et al* (i.e., when the NDGD is more than 6, which happens frequently in practice in order to improve the SNR and reduce the bias of tensor estimation) [15].

Both **Figure 4** and all *P* values (>0.05) of FA values within ROI listed in **Table 2** demonstrate that there are no significant variances in the FA values within ROI between any 2 protocols with the NDGD increasing, which are accordant with the conclusion of the previous original research by Ni *et al* [11].

5. DISCUSSION

Ni *et al* [11] found that NDGD = 6 and NEX = 10 were sufficient for estimation of FA from ROI calculations. But in this paper, the FA maps shown in the first column in Fig 2 look bad, it is considered that the NDGD is too low (NDGD=6), which results in low SNR for these tensor-derived measurement maps. Because our results were based on the same original SNR, i.e. all the DW images used in this study were averaged thrice, so the SNR of these tensor-derived measurement maps varied in the 7 protocols only is dependent on the NDGD.

The functional dependence of the SNR of each tensor-derived measurement on the NDGD is the most interesting result for this paper. The curves in **Figure 3** and **Figure 4** indicated that the SNR of FA maps, ADC-mean maps, LEV maps, and EVO maps associated with λ_1 were improved with the NDGD increasing, but there were no significant variances in the FA values within ROI among these 7 protocols (refer to the last column in **Table 1**).

Table 1. Mean ± standard deviation for the SNR of tensor-derived measurement maps and the FA values within ROI in the 7 protocols. *NDGD represents number of diffusion gradient directions.

NDGD*	SNR of Tensor-derived Measurement				FA values
	FA Maps	ADC-mean	LEV Maps	EVO Maps	
6	35.78 ± 1.39	49.08 ± 3.39	34.39 ± 1.20	40.30 ± 0.53	0.79 ± 0.06
9	51.65 ± 1.43	60.96 ± 1.87	39.96 ± 1.54	51.08 ± 0.54	0.73 ± 0.08
12	61.62 ± 2.24	66.55 ± 1.64	44.09 ± 1.73	58.22 ± 0.23	0.75 ± 0.08
15	63.98 ± 1.67	68.67 ± 1.44	45.95 ± 1.38	59.65 ± 0.46	0.76 ± 0.08
20	69.39 ± 1.66	72.63 ± 1.29	48.22 ± 1.57	61.27 ± 0.43	0.74 ± 0.09
25	71.15 ± 2.46	73.28 ± 1.47	53.16 ± 1.26	64.28 ± 0.41	0.74 ± 0.08
30	72.45 ± 7.52	73.68 ± 1.77	57.44 ± 1.65	69.40 ± 0.49	0.74 ± 0.08

Table 2. P values for the SNR of tensor-derived measurement maps and the FA values within ROI in the 4 protocols corresponding to different NDGD (6, 12, 20, and 30 noncollinear). *NDGD represents number of diffusion gradient directions.

NDGD*	P values for the SNR of tensor-derived measurement maps and the FA values within ROI				
	FA values	ADC-mean Maps	LEV Maps	EVO Maps	FA values
6 vs 12	<<0.005	<<0.005	0.001	<<0.005	0.287
6 vs 20	<<0.005	<<0.005	<<0.005	<<0.005	0.177
6 vs 30	<<0.005	<<0.005	<<0.005	<<0.005	0.150
12 vs 20	0.010	0.011	0.046	0.050	0.725
12 vs 30	0.011	0.009	0.001	<<0.005	0.693
20 vs 30	0.089	0.376	0.006	0.011	0.977

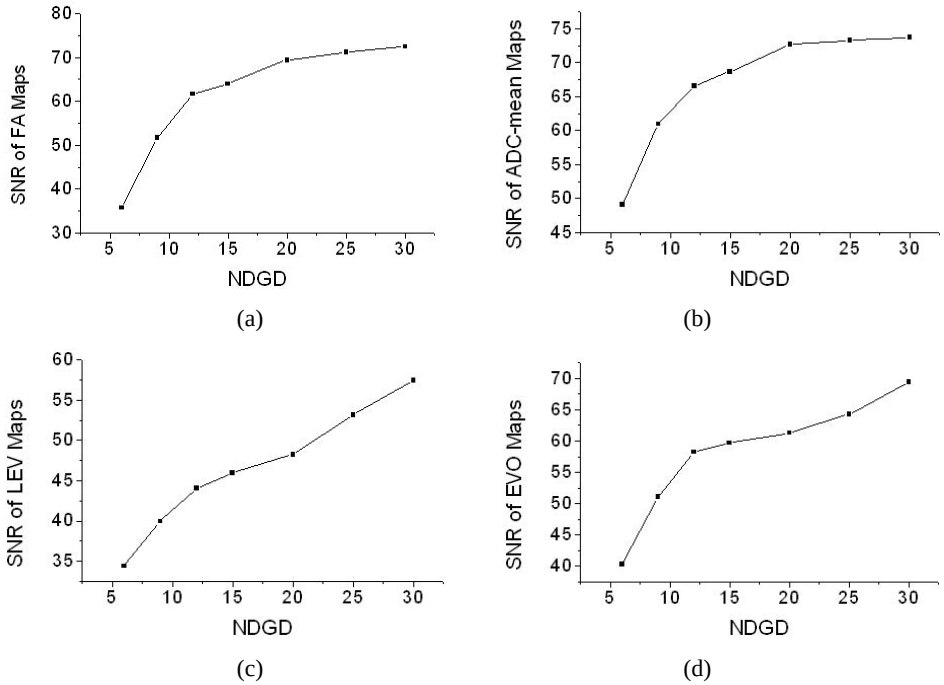


Figure 3. SNR of tensor-derived measurement maps vs NDGD.

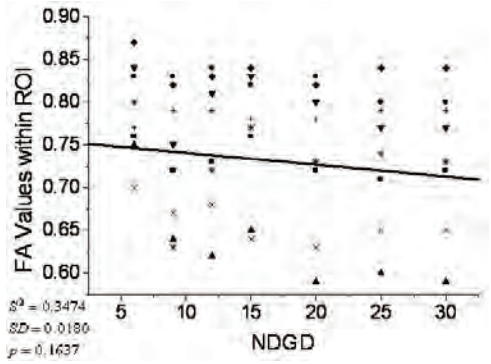


Figure 4. FA values within ROI vs NDGD.

In the previous studies, with the constant imaging time, some researchers investigated various protocols (different NDGD) in terms of the variance of FA measurements and demonstrated that a protocol employing 24 or 30 NDGD outperformed a protocol with only 6 NDGD [8, 9]. Papadakis NG *et al* considering three diffusion anisotropy maps concluded that the minimum NDGD required for robust anisotropy estimation was between 18 and 21[16].

In our study, we mainly focused on the effect of different NDGD on the SNR of tensor-derived measurement maps (FA maps, ADC-mean maps, LEV maps, and EVO maps associated with λ_1) in the 7 protocols with fixed NEX=3. The curves in **Figure 3 (a)** and **(b)** show that, for FA maps and ADC maps, there is a remarkable and linear improvement in the SNR when the NDGD increases from 6 to 12. Further improvement in the SNR (albeit less remarkable) is observed when the NDGD is further increased. Especially, there is no considerable difference occurred in the SNR when the NDGD increases from 20 to 30, which means that it has little contribution to SNR of FA and ADC-mean maps when NDGD is more than 20. *P* values (20 vs 30) = 0.089 for FA maps and 0.376 for ADC maps in the **Table 2** are much more than 0.05, which also demonstrates that there is no significant improvement in SNR by increasing the NDGD from 20 to 30. This is consistent with the conclusion suggesting that 20-diffusion gradients be probably sufficient for in vivo human study of diffusion anisotropy [10].

For LEV maps and EVO maps associated with λ_1, seen from the curve in **Figure 3 (c)** and **(d)**, there is a significant improvement in SNR by increasing the NDGD from 6 to 12 and from 20 to 30. The curve increases almost linearly when NDGD increases from 6 to 30 for SNR of LEV maps. *P* values (20 vs 30) = 0.006 for LEV maps and 0.011 for EVO maps associated with λ_1 in the **Table 2**, which are less than 0.05, also demonstrate that the SNR of LEV maps and EVO maps associated with λ_1 are significant different between 20 and 30 diffusion gradients. This suggests that the NDGD less than 30 be not enough for tensor-orientation estimation, which also is consistent with the results from Monte Carlo study [10].

Therefore, the more NDGD are used, the higher SNR would be obtained. But in clinical applications of DTI, the total scanning time could not be too long because of the artifacts caused by patients' motion. So an optimum NDGD for DTI data acquisition is needed due to both the requirements and limitations mentioned above. There would be a trade-off between the NDGD and clinical limitations.

6. CONCLUSION

7 different types of results derived from 6, 9, 12, 15, 20, 25, and 30 noncollinear NDGD, respectively, have been compared in terms of SNR of tensor-derived measurement maps and FA values within ROI based on Matlab platform. The SNR of FA maps and ADC-mean maps increased linearly as the NDGD increased from 6 to 20. And the curves were almost level as the NDGD increasing from 20 to 30. For SNR of LEV and EVO maps, the curves were linearly direct ratio to the NDGD. FA values within ROI were independent of NDGD. This study provides insight into the effect of NDGD on SNR and may be useful in understanding the tradeoffs involved in DTI acquisition design.

ACKNOWLEDGMENTS

We are grateful to Dr. Hangyi Jiang who works in Johns Hopkins University School of Medicine, Dr. Maolin Qiu who works in Yale University, and Dr. Bob L. Hou from Memorial Sloan Kettering Cancer Center for their helpful discussions and encouraging comments during the course of this study. In addition, we thank Dr. Hangyi Jiang again for supplying the data and the software (DTI Studio) for our study. Finally, we would like to thank the reviewers for their valuable remarks.

REFERENCES

[1] M. Jackowski, C. Y. Kao, M. L. Qiu, *et al*. (2005) White matter tractography by anisotropic wavefront evolution and diffusion tensor imaging**.** Medical Image Analysis, 9, 427–440.

[2] T. McGraw, B.C. Vemuri, Y. Chen, *et al*. (2004) DT-MRI denoising and neuronal fiber tracking**.** Medical Image Analysis, 8, 95–111.

[3] D. L. Bihan, J. F. Mangin, C. Poupon, *et al*. (2001) Diffusion Tensor Imaging: Concepts and Applications. Journal of Magnetic Resonance Imaging, 13, 534–546.

[4] P. J. Basser, J. Mattiello, D. LeBihan. (1994) Estimation of the effective selfdiffusion tensor from theNMRspin echo. JMagn Reson B, 103, 247–254.

[5] P. J. Basser, J. Mattiello, D. LeBihan. (1994) MR diffusion tensor spectroscopy and imaging. Biophys J, 66, 259–267.

[6] H. Jiang, P. C.M. van Zijl, *et al*. (2006) DtiStudio: Resource program for diffusion tensor computation and fiber bundle tracking. computer methods and programs in biomedicine, 8 1,106–116.

[7] D. S. Tuch. (2004) Q-Ball Imaging. Magnetic Resonance in Medicine, 52, 1358 – 1372

[8] N. G. Papadakis, D. Xing, G. C. Houston, *et al*. (1999) A study of rotational invariant and symmetric indices of diffusion anisotropy**.** Magn Reson Imaging, 17, 881–92.

[9] S. Skare, M. Hedehus, M.E. Moseley, *et al.* (2000) Condition number as a measure of noise performance of diffusion tensor data acquisition schemes with MRI. J Magn Reson, 147, 340–52.

[10] D. K. Jones. (2004) The effect of gradient sampling schemes on measures derived from diffusion tensor MRI: a Monte Carlo study**.** Magn Reson Med, 51, 807–15.

[11] H.Ni, V.Kavcic. T. Zhu, *et al*. (2006) Effects of Number of Diffusion Gradient Directions on Derived Diffusion Tensor Imaging Indices in Human Brain. AJNR Am J Neuroradiol, 27,1776-81

[12] A. H. Poonawalla, MS, X. H Joe Zhou, PhD *. (2004) Analytical error propagation in diffusion anisotropy calculations. JMRI, 19, 489–498

[13] M. Zou (2001) Deconvolution and Signal recovery. Publishing Company of National Defence and Industry (Chinese book).

[14] S Pajevic, C Pierpaoli. (1999) Color schemes to represent the orientation of anisotropic tissues from diffusion tensor data: application to white matter fiber tract mapping in the human brain. Magn Reson Med, 42, 526–540.

[15] D.K. Jones, M.A. Horsfield. (1999) A. Simmons. Optimal strategies for measuring diffusion in anisotropic systems by magnetic resonance imaging**.** Magn. Reson. Med, 42 (3), 515–525.

[16] N.G. Papadakis, C. D. Murrills, L. D. Hall, *et al*. (2000) Minimal gradient encoding for robust estimation of diffusion anisotropy. Magn Reson Imaging, 18, 671–679.

J. Biomedical Science and Engineering, 2009, 2, 102-105

The impact of frequency aliasing on spectral method of measuring T wave alternans*

Di-Hu Chen[1], Sheng Yang[1]

[1]Department of Precision Machinery & Instrumentation, University of Science and Technology of China, Hefei 230027, China.
Email: yangs@ustc.edu.cn.

Received Jan. 5th, 2009; revised Feb. 19th, 2009; accepted Feb. 20th, 2009.

ABSTRACT

In this paper we investigate frequency aliasing in spectral method of measuring T wave alternans, which may lead a high false positive rate. Microvolt T wave alternans(TWA) has been evaluated as a means of predicting occurrence of ventricular tachyarrhythmia events and its association with the genesis of ventricular arrhythmias has been demonstrated. Nowadays, spectral method is one of the most widely used procedures for measurement of microvolt TWA. In our study, based on the sampling theory, the alternans frequency 0.5 cycles/beat, at which the power of the spectrum is used to calculated the V_{alt} and K score (these two parameters indicate the TWA), is equal to the nyquist frequency. Thus this generates frequency aliasing which will make the power at the alternans frequency ($P_{0.5}$) be two times of the real magnitude of the original spectrum amplitude. With the assumption that the noise spectrum follows the normal distribution, in spectral method of measuring T wave alternans, the measuring standard K score>3 to consider the T wave alternans significant is only with a p<0.133. By change the standard to K score>6 can solve this problem and make the p value to p<0.0027.

Keywords: TWA, Sampling, Spectral method, Frequency Aliasing

1. INTRODUCTION

Sudden cardiac death (SCD) is the leading cause of cardiovascular mortality in the developed countries [1]. There is no an effective diagnostic method to identify patients at high risk for SCD. Though many non-invasive tests related to high-risk of SCD such as frequent and complex ventricular arrhythmias in 24-hour Holter monitoring, ventricular late potentials in signal-average ECG, low heart rate variability, and increased dispersion of repolarization are introduced, the positive predictive value of these tests is too low to consider them as sufficient to make a decision about specific treatment, especial defibrillator implantation. Recently risk stratification research has been focused on microvolt T wave alternans, which is considered as a promising clinical marker of arrhythmic events [2].

Microvolt T wave alternans (TWA), also called repolarization alternans, is a phenomenon appearing in the electrocardiogram (ECG) as a consistent fluctuation in the repolarization morphology on an every-other-beat basis [3]. Microvolt TWA has been evaluated prospectively in a variety of patient populations as a means of predicting occurrence of ventricular tachyarrhythmia events and its association with the genesis of ventricular arrhythmias has been demonstrated [4].

In measuring of TWA, spectral method is a widely used method. This method uses a certain measurements taken on corresponding points of some consecutive T wave to compute a spectrum. And then two parameters: the alternans voltage (V_{alt}) and alternans ratio (K score) are calculated from this spectrum. These two parameters indicate whether the TWA is significant.

This paper investigates frequency aliasing (also called aliasing in short) in spectral method of measuring T wave alternans, which may lead a high false positive rate. In section II we take a brief view of the spectral method of measuring T wave alternans. In section III, we introduce the sampling theory and frequency aliasing and in section IV, we investigate frequency aliasing in the spectral method.

2. SPECTRAL METHOD OF MEASURING T WAVE ALTERNANS

Until now, two main techniques have been applied for measurement of microvolt TWA in clinical setting: fast Fourier Transform (FFT) spectral method and modified moving average (MMA) analysis method [4]. The FFT spectral method which was developed at Massachusetts Institute of Technology by Dr. Richard J. Cohen [5, 6, 7] is the most widely used procedure. This technique uses 128 measurements taken on corresponding points of 128 consecutive T waves to compute a spectrum. Each T wave is measured at the same time relative to the QRS complex [8]. For the spectrum is created by measure-

*This work was support by the National Natural Science Foundation of China (60571034)

ments taken once per beat, its frequencies are in the units of cycles/beat. The point on the spectrum corresponding to exactly 0.5 cycles/beat indicates the level of alternation of T wave waveform [8].

Two measurements are obtained form the analysis: the alternans voltage (V_{alt}) and alternans ratio (K score). The V_{alt} measured in uV, represents the square root of alternans power which is defined as the difference between the power at the alternans frequency (0.5 cycles/beat) and the power at the noise frequency (0.44 and 0.49 cycles/beat). And also, it corresponds to the root mean square difference in the voltage between the overall mean beat and either the odd-numbered or even- numbered beats. The alternans ratio K score is calculated as the ratio of the alternans power divided by the standard deviation of the noise in the reference frequency band. See bellow

$$V_{alt} = P_{0.5} - \mu \tag{1}$$

$$K\ score = \frac{P_{0.5} - \mu}{\sigma} \tag{2}$$

where $P_{0.5}$ is the power at 0.5 cycles/beat, μ and σ are the mean and standard deviation. When the alternans power is >3 SD above the noise level (K score>3), alternans is considered significant in statistic.

3. NYQUIST FREQUENCY AND FREQUENCY ALIASING IN MEASURING T WAVE ALTERNANS

Consider a continuous-time signal $f(t)$. We define sampling as the generation of an ordered number sequence by taking values of $f(t)$ at specified instants of time [9]. In most cases continuous-time signals are sampled at equal increments of time. The sample increment, called the sample period, is usually denoted as T_s. Therefore, the sampled signal values available in the computer are $f(nT_s)$, where n is an integer.

Figure 1 shows the ideal impulse sampling operation. This is seen to be a modulation process, in which the carrier signal $\delta_T(t)$ is defined as the train of impulse function:

$$\delta_T(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT_S) \tag{3}$$

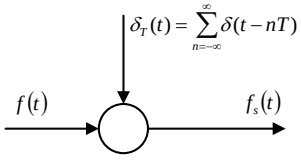


Figure 1. Impulse sampling.

The output of the modulator, denoted by $f_s(t)$ is given by

$$f_s(t) = f(t)\delta_T(t) = \sum_{n=-\infty}^{\infty} f(nT_s)\delta(t - nT_s) \tag{4}$$

We begin to investigate the characteristics of the sampling operation in **Figure 1** by taking the Fourier transform of $f_s(t)$. For Fourier transform, we can easy get

$$\delta_T(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT_s) \xleftrightarrow{F} \omega_s \sum_{k=-\infty}^{\infty} \delta(\omega - k\omega_s) \tag{5}$$

where $\omega_s = \frac{2\pi}{T_s}$ is the sampling frequency in radians/second. The sampling frequency in hertz is giver by $f_s = \frac{1}{T_S}$; therefore, $\omega_s = 2\pi f_s$. For multiplication in the time domain results in convolution in the frequency domain. Then from (2) and (3),

$$\begin{aligned} F_s(\omega) &= \frac{1}{2\pi} F(\omega) * \left[\omega_s \sum_{k=-\infty}^{\infty} \delta(\omega - k\omega_s) \right] \\ &= \frac{1}{T_s} \sum_{k=-\infty}^{\infty} F(\omega) * \delta(\omega - k\omega_s) \end{aligned} \tag{6}$$

where $F(\omega)$ is the Fourier transform of $f(t)$ and $F_s(\omega)$ is the Fourier transform of $f_s(t)$. Because of the convolution property of the impulse function,

$$F(\omega) * \delta(\omega - k\omega_s) = F(\omega - k\omega_s) \tag{7}$$

Thus the Fourier transform of the impulse-modulated signal (2) is given by

$$F_s(\omega) = \frac{1}{T_s} \sum_{k=-\infty}^{\infty} F(\omega - k\omega_s) \tag{8}$$

Frequency domain characteristics of the sampling operation can be derived from this result.

From (6) we see that the effect of sampling $f(t)$ is to replicate the frequency spectrum of $F(\omega)$ about the frequencies $k\omega_s, k = \pm 1, \pm 2, \pm 3\ldots$. This result is show in **Figure 2(b)** for the signal of **Figure 2(a)**.

The frequency $\omega_s / 2$ is called the Nyquist frequency and the Shannon sampling frequency. One of the requirements for sampling is that the sampling frequency

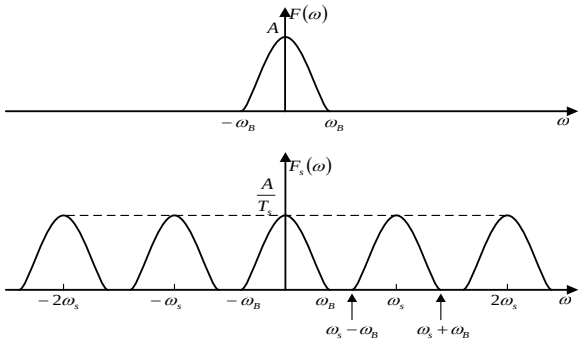


Figure 2. The frequency spectrum of a sampled signal.

must be chosen such that $\omega_s > 2\omega_M$, where ω_M is the highest frequency in the frequency spectrum of the signal to be sampled.

From part 2 we know, in spectral analysis of microvolt T wave alternans, the sampling frequency is 1 cycles/beat and the alternans frequency is 0.5 cycles/beat, which is exactly 0.5 of the sampling frequency. This is also the Nyquist frequency. In sampling theory, input-signal frequencies that exceed the Nyquist frequency are aliased. That is, they are folded back or replicated at other positions in the spectrum above and below the Nyquist frequency. (See **Figure 3**)

So in spectral method of measuring T wave alternans, power at the alternans frequency ($P_{0.5}$) which is used to indicate the level of alternation of T wave waveform is two times of the real magnitude of the original spectrum at 0.5 cycles/beat.

4. DISCUSSION

In spectral method of measuring T wave alternans, as mentioned in part 2, the T wave alternans is considered significant when the K score is higher than 3, while the K score represents the ratio of the alternans power and the standard noise power deviation. In order to explain why this rule works, let's take a look at the normal distribution.

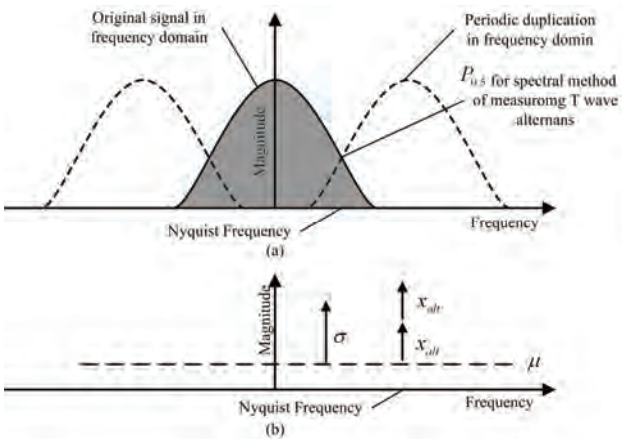


Figure 3. (a) Input signal frequencies exceed the Nyquist frequency are aliased.(b) With the frequency aliasing, P0.5 is two times of the real magnitude.

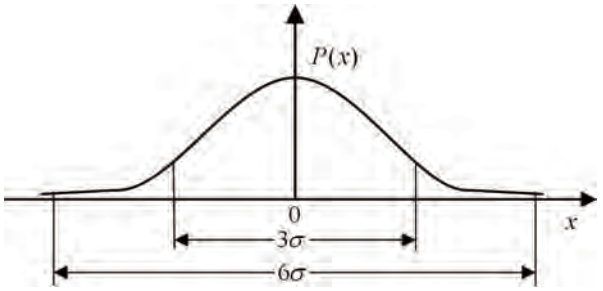


Figure 4. Probability density function of the normal distribution.

A normal distribution in a variate X with mean μ and variance σ^2 is a statistic distribution with probability function

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \tag{9}$$

on the domain $x \in (-\infty, \infty)$. The importance of the normal distribution as a model of quantitative phenomena in the natural and behavioral sciences is due in part to the central limit theorem. Many measurements, ranging from psychological to physical phenomena can be approximated, to varying degrees, by the normal distribution. In the spectral method of measuring T wave alternans, the noise of the power spectrum can be assumed to be normal. In normal distribution, if $(x-u)/\sigma > 3$, then the x is statistically significant with the probability ($p < 0.0027$) due to chance. This can be easily calculated via (9) and this probability is also called the false positive rate. When there is frequency aliasing, let x_{alt} be the real alternans power, and assume that the distribution of the noise spectrum is normal no matter whether there is T wave alternans, then $P_{0.5} = 2x_{alt} + \mu$ and (2) can be written as

$$K_{score} = \frac{2x_{alt} + \mu - \mu}{\sigma} = \frac{2x_{alt}}{\sigma} > 3 \tag{10}$$

This is also $x_{alt}/\sigma > 1.5$, from the table of the standard normal distribution we can get that in this condition the p value is only less than 0.133 ($p < 0.133$), which means a high false positive rate.

In order to solve this problem, we can change the standard from $K_{score} > 3$ to $K_{score} > 6$. From (10) we can get $p < 0.0027$ if consider T wave alteransn significant when $K_{score} > 6$.

5. CONCLUSION

In this paper study, based on the sampling theory, in spectral method of measuring T wave alternans, the alternans frequency is equal to the nyquist frequency, and this makes frequency aliasing in the power spectrum, which will lead the increase of the false positive rate, from $p < 0.0027$ to $p < 0.133$. By changing the standard from $K_{score} > 3$ to $K_{score} > 6$ can effectively solve this problem.

REFERENCES

[1] A. Bay& and J. Guindo, (1989) Sudden Cardiac Death. Spain: MCR.

[2] J. P. Martinez, S. Olmos and P. Laguna, (2000) Simulation Study and Performance Evaluation ofT−Wave Alternans Detec-

tor. Proceedings of the 22nd Annual EMBS International Conference, July 23−28, Chicago IL.

[3] J. P. Martínez and S. Olmos, (2005) Methodological Principles of T Wave Alternans Analysis: A Unified Framework. IEEE Transactions On Biomedical Engineering, vol. 52, NO. 4.

[4] B. D. Nearing, R. L. Verrier. (2002) Modified moving average method for T−wave alternans analysis with high accuracy to predict ventricular fibrillation. J Appl Physiol, 92, 541−49.

[5] D. R. Adam, J. M. Smith, S. Akselrod, S. Nyberg, A. O. Powell, R. J. Cohen. (1984) Fluctuations in T−wave morphology and susceptibility to ventricular fibrillation. J Electrocardiol, 17, 209–18.

[6] A. L. Ritzenberg, D. R. Adam, R. J. Cohen. (1984) Period multiplying−evidence for nonlinear behavior of the canine heart. Nature, 307, 159– 61.

[7] J. M. Smith, E. A. Clancy, C. R. Valeri, J. N. Ruskin, R. J. Cohen. (1988) Electricalalternans and cardiac electrical instability. Circulation, 77, 110– 21.

[8] D. M. Bloomfield, S. H. Hohnloser, R. J. Cohen. (2002) Interpretation and classification of microvolt T−wave alternans tests. J Cardiovasc Electrophysiol, 13:502– 12.

[9] C. L. Phillips, J. M. Parr and E. A. Riskin. (2004) Signal, System and Transform. China Machine Press, Beijing.

J. Biomedical Science and Engineering, 2009, 2, 106-116

MicroPath-A pathway-based pipeline for the comparison of multiple gene expression profiles to identify common biological signatures

Mohsin Khan[1], Chandrasekhar Babu Gorle[1], Ping Wang[3], Xiao-Hui Liu[2], Su-Ling Li[1]

[1]Molecular Immunology & bioinformatics Group, Microarray Facility, Division of Bio-Sciences, Brunel University, Uxbridge, UB8 3PH, UK; [2]Intelligent Data Analysis Group, Department of Information Systems and Computing, Brunel University, Uxbridge, UB8 3PH, UK; [3]Immunology Group, Institute of Cell and Molecular Sciences, Barts and London School of Medicine, London, UK. Correspondence should be addressed to Su-Ling Li (su-ling.li@brunel.ac.uk)

Received Jan. 2nd, 2009; revised Feb. 15th, 2009; accepted Mar. 4th, 2009.

ABSTRACT

High throughput gene expression analysis is swiftly becoming the focal point for deciphering molecular mechanisms underlying various different biological questions. Testament to this is the fact that vast volumes of expression profiles are being generated rapidly by scientists worldwide and subsequently stored in publicly available data repositories such as ArrayExpress and the Gene Expression Omnibus (GEO). Such wealth of biological data has motivated biologists to compare expression profiles generated from biologically-related microarray experiments in order to unravel biological mechanisms underlying various states of diseases. However, without the availability of appropriate software and tools, they are compelled to use manual or labour-intensive methods of comparisons. A scrutiny of current literature makes it apparent that there is a soaring need for such bioinformatics tools that cater for the multiple analyses of expression profiles.

In order to contribute towards this need, we have developed an efficient software pipeline for the analysis of multiple gene expression datasets, called Micropath, which implements three principal functions; 1) it searches for common genes amongst n number of datasets using a number crunching method of comparison as well as applying the principle of permutations and combinations in the form of a search strategy, 2) it extracts gene expression patterns both graphically and statistically, and 3) it streams co-expressed genes to all molecular pathways belonging to KEGG in a live fashion. We subjected MicroPath to several expression datasets generated from our tolerance-related in-house microarray experiments as well as published data and identified a set of 31 candidate genes that were found to be co-expressed across all interesting datasets. Pathway analysis revealed their putative roles in regulating immune tolerance. MicroPath is freely available to download from: www.1066technologies.co.uk/micropath.

Keywords: Co-Expression Analysis, Microarray, Permutations and Combinations, Multiple Gene Expression Analysis

1. INTRODUCTION

There is a general consensus amongst scientists and researchers that the fundamental asset of microarray technology lies in its inherent ability to produce a global snapshot of the cellular state in the milieu of any given biological question. It is therefore not surprising that microarrays have revolutionised the field of molecular biology by offering an efficient and cost effective medium for biologists to quantify mRNA transcript levels of several thousands of genes concurrently in order to observe specific states of the transcriptome (in response to a particular treatment or specific time point). Owing to this innate faculty to decipher the transcriptome, gene expression profiles pertaining to a wide variety of biological questions are being rapidly generated by scientists worldwide and are deposited and subsequently made accessible through public repositories such as ArrayExpress [1] and the Gene Expression Omnibus [2]. With so much wealth of high throughput biological data made available, biologists have become motivated to utilise these sets of data in an attempt to investigate common regulatory signatures, which may be implicating the transcriptome state across multiple gene expression profiles sharing a similar biological theme. One of the most widely accepted methodologies of comparing expression profiles is based on the assumption that genes across different biological conditions sharing similar expression patterns are likely to be involved in the same biological processes [2], and therefore, may share common regulatory signatures. By using this method of comparison, which is one of the most successful methods to date, coupled with the availability of publicly available data

repositories offering gene expression profiles, biologists have been granted the opportunity to answer complex biological questions pertinent to biological phenomena underlying various different disease states.

To this end, we have developed a novel bioinformatics software pipeline called MicroPath, which specialises in the cross comparison of multiple gene expression datasets and attempts to identify common regulatory signatures from the standpoint of molecular pathway analysis. When one scrutinises current literature relevant to automated solutions of gene expression analysis, it becomes apparent that there is an increasing demand for software applications that offer an efficient pipeline to the analysis of multiple gene expression profiles. Although current meta-analyses studies have been conducted with the purpose of employing statistical techniques to compare cDNA and affymetrix gene expression profiles [3,4,5,6], it cannot be denied that there is a mounting need for this process to be automated. Nevertheless, various approaches/algorithms of statistical nature have already been implemented with the purpose of identifying the most relevant pathways in a given experiment [7,8,9] together with methods such as Gene Set Enrichment Analysis (GSEA), which ranks genes based on the correlations between their expressions and observed phenotypes in the context of biological pathway discoveries [10]. There are also tools available that functionally annotate gene expression data [11,12]. Albeit, it remains infeasible for biologists to cross compare several expression profiles without an automated solution, and hence, they are faced with the labour-intensive task of employing manual methods to carry out their comparisons. MicroPath uses the meta-analytic standard and has been specifically developed to: compare several significantly expressed sets of genes in order to find the intersection of common genes using both number crunching methods as well as the classical permutation and combination principle, extract putative regulatory signatures using both statistical and graph-based approaches and finally, mapping these sub-sets of co-expressed genes to molecular pathways all in the form of a high throughput pipeline.

2. IMPLEMENTATION

The front-end of MicroPath was developed in Visual Basic.Net and Perl, and the database back-end was developed in MySQL. Upon analysing the users input files (gene expression profiles), processed data is displayed intuitively on the graphical user interface, which is equipped with various interactive objects such as charting facilities, buttons, drop-down menus and user input/output dialogues. The interface is also equipped with a function to export processed data into Microsoft excel for further scrutiny and use.

2.1. System Architecture

MicroPath carries out meta-profiling of multiple gene expression datasets using two different approaches. Firstly, the intersection of common genes is identified across *n* number of expression profiles, which is then plotted graphically using a simple number crunching exercise. The second approach applies to a situation where an attempt to identify common genes across *n* number of expression profiles using the aforementioned approach fails due to the absence of common genes across all datasets (this situation is especially common when a large number of expression profiles are compared, which reduces the probability of finding a common gene amongst them). Consequently, MicroPath applies the permutations and combinations mathematical principle to solve this problem (refer to *implementation of meta-analysis strategy* below for details). Once the intersection of a set of common genes has been identified and subsequently displayed on the interface (using either of the above methods), the next stage in the analysis is to extract patterns from the intersection in order to identify common genes that are being expressed in accordance with the biological question. MicroPath offers a semi-automated graph-based approach to achieve this as well as classical statistics to identify the overall correlation of gene expression. Finally, co-expressed genes (common genes that are expressed in accordance to the relevant biological question) are mapped to all molecular pathways known to date in order to reveal their molecular dependencies (refer to **Figure 1** for the complete system architecture).

2.2. Implementation of Meta-analysis Strategy

In theory, an intersection of a sub-set of common genes across multiple gene expression profiles should be easily attainable using simple number crunching methods of comparison. In practice, this is not always the case since the likelihood of identifying genes sharing common accession identifiers decreases as the number of profiles to compare increases. This inverse relationship makes sense both mathematically and biologically. From a biological perspective, regulatory signatures tend to be diluted over entire datasets and as a result, only a proportion of the total number of profiles to compare may actually share common genes. In such a scenario, using a simple method of comparison would break down at some point and no common genes would be reported to the user, although common genes may be present within *n*-1 expression profiles. To prevent potentially interesting biological findings to be hampered at this point in the analysis, we have applied the principle of mathematical combinations to the comparison of multiple gene expression profiles. All possible combinations of comparing *n* number of datasets with each other are firstly computed using the combination equation:

$$ {}^{n}C_{r} = \frac{n!}{R!(n-r)!} $$

Where *n*! Is the factorial of the total number of datasets *n*, and r! Is the factorial of the selected number of datasets to compare when comparing *n* datasets results in zero common genes, r.

This generates the total number of permutations of comparing datasets (Cr) for given values of *n* (total number of datasets imported by user) and *r* (number of intended datasets used to search for common genes when zero common genes are reported across *n* datasets) (**Table 1**).

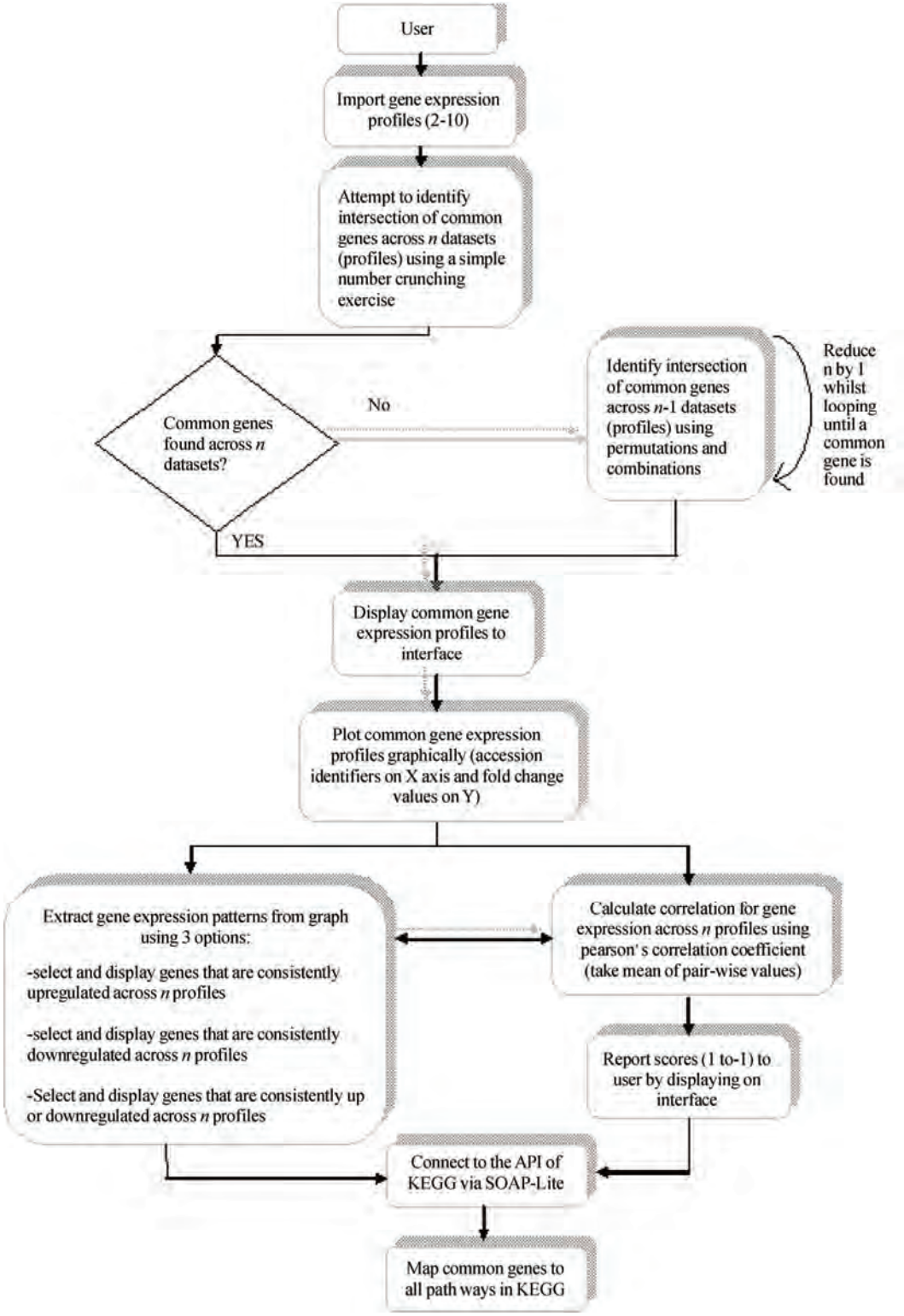


Figure 1. Functions of MicroPath. Users are prompted to import up to 10 gene expression profiles, which are then compared using a direct comparison method. If this method yields zero common genes, MicroPath automatically attempts to identify an intersection of common genes by reducing the search space to *n*-1 datasets using permutations and combinations. This process is continued until at least 1 common gene is reported. Following this, users are provided with a function to search for expression patterns graphically and gene expression correlations are calculated statistically using the pearson's correlation coefficient algorithm. Finally, co-expressed genes are mapped to all molecular pathways of KEGG in a high throughput fashion by automatically accessing its API via SOAP-Lite.

Table 1. Multiple gene expression profile search strategy generated from applying the principle of permutations and combinations. The first column represents the total number of expression datasets, n, that users may import (this is the search space). The second column represents, *r*, the number of expression datasets to compare if zero common genes are reported to be matched across n datasets. The final column represents the total number of mathematical combinations possible for each given value of n and *r*.

Total number of expression datasets (***n***)	Number of intended expression datasets to compare when comparing *n* datasets yields no results (***r***)	***n - r***	Total number of combinations of *r* (***Cr***)
10	9	1	10
10	8	2	45
10	7	3	120
10	6	4	210
10	5	5	252
10	4	6	210
10	3	7	120
10	2	8	45
9	8	1	9
9	7	2	36
9	6	3	84
9	5	4	126
9	4	5	126
9	3	6	84
9	2	7	36
8	7	1	8
8	6	2	28
8	5	3	56
8	4	4	70
8	3	5	56
8	2	6	28
7	6	1	7
7	5	2	21
7	4	3	35
7	3	4	35
7	2	5	21
6	5	1	6
6	4	2	15
6	3	3	20
6	2	4	15
5	4	1	5
5	3	2	10
5	2	3	10
4	3	1	4
4	2	2	6
3	2	1	3

These combinations of datasets (*Cr*) are then used as a criterion to search for common genes across *r* number of gene expression profiles when comparing *n* number of datasets fail to yield any common genes. However in this scenario, *n* number of datasets are still used as the search space from which all possible combinations (*Cr*) of *r* datasets are compared to each other in order to increase the probability of finding a common gene. Once common genes have been identified using this method, MicroPath will report the results to the interface.

2.3. Extracting Gene Expression Patterns Graphically and Statistically

Following the identification of common genes across *n* datasets using either of the methods described earlier, the next stage in the analysis is to generate a graphical representation of this expression data from which biologically meaningful patterns can be extracted. Because signals pertaining to transcriptome states tend to be diluted over entire profiles, a specific criterion is required to narrow down the common genes of interest to include only those genes that are consistently regulated according to the biological question. The assumption we have made is that any given common gene across *n* datasets can exhibit one of three specific behaviours. It can either be consistently upregulated across all datasets, downregulated across all datasets and up or downregulated across all datasets. Based on the nature of the specific biological question, users can select the appropriate pattern from the options, which will result in a graphical display of those genes which satisfy the search criteria. Together with this faculty to graphically extract patterns for individual gene expression data points, MicroPath also implements the pearsons correlation coefficient statistical test in order to extract a global gene expression pattern existing between common genes pertaining to two individual expression profiles. The correlations are calculated in a pair-wise manner until each expression data has been statistically compared to all other datasets within *n*, according to the pearsons correlation coefficient equation:

$$r = \frac{\sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right)\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}\right)}}$$

Each pair-wise score is then finally averaged in order to provide a global measure of correlation existing between *n* expression profiles. Scores are reported from-1 (perfect negative correlation) to+1 (perfect positive correlation).

2.4. High Throughput Molecular Pathway Analysis

To decipher molecular mechanisms fundamental to the researcher's biological question, it is necessary to map common gene expression profiles of co-expressed genes to molecular pathways. This is because biological pathways reveal molecular dependencies that exist between genes by illustrating how they collaborate with one another when they participate in specific biological functions. Furthermore, pathways reveal various signalling cascades that play imperative roles in dictating these gene associations. In light of this, we have implemented Micropath to access the Application Programming Interface (API) of the molecular pathway database belonging to KEGG [13] using SOAP-Lite in order to dynamically interact with the static pathway maps. Perl scripts were written for MicroPath to specifically 1) search for user's co-expressed genes in all biological pathways, 2) highlight genes on to pathways, and 3) return the results of the search to Micropath's interface (i.e. URL's of colour coded pathway maps) (**Figure 2**). Once MicroPath has searched for all of the user's co-expressed genes in all of the molecular pathways, the URL of each pathway is displayed on the sub-interface. In order to avoid redundancy issues, the URL for each pathway will highlight all co-expressed genes that participate in a given pathway. To help users identify biologically meaningful pathways relevant to their specific biological question, MicroPath will calculate the number of genes identified in a given pathway and 1) express this as a percentage in relation to the total number of common genes from the intersection and 2) express this as a percentage in relation to the total number of genes belonging to that pathway.

Clicking on these links will generate the specific KEGG pathway in HTML on which users co-expressed genes will be highlighted.

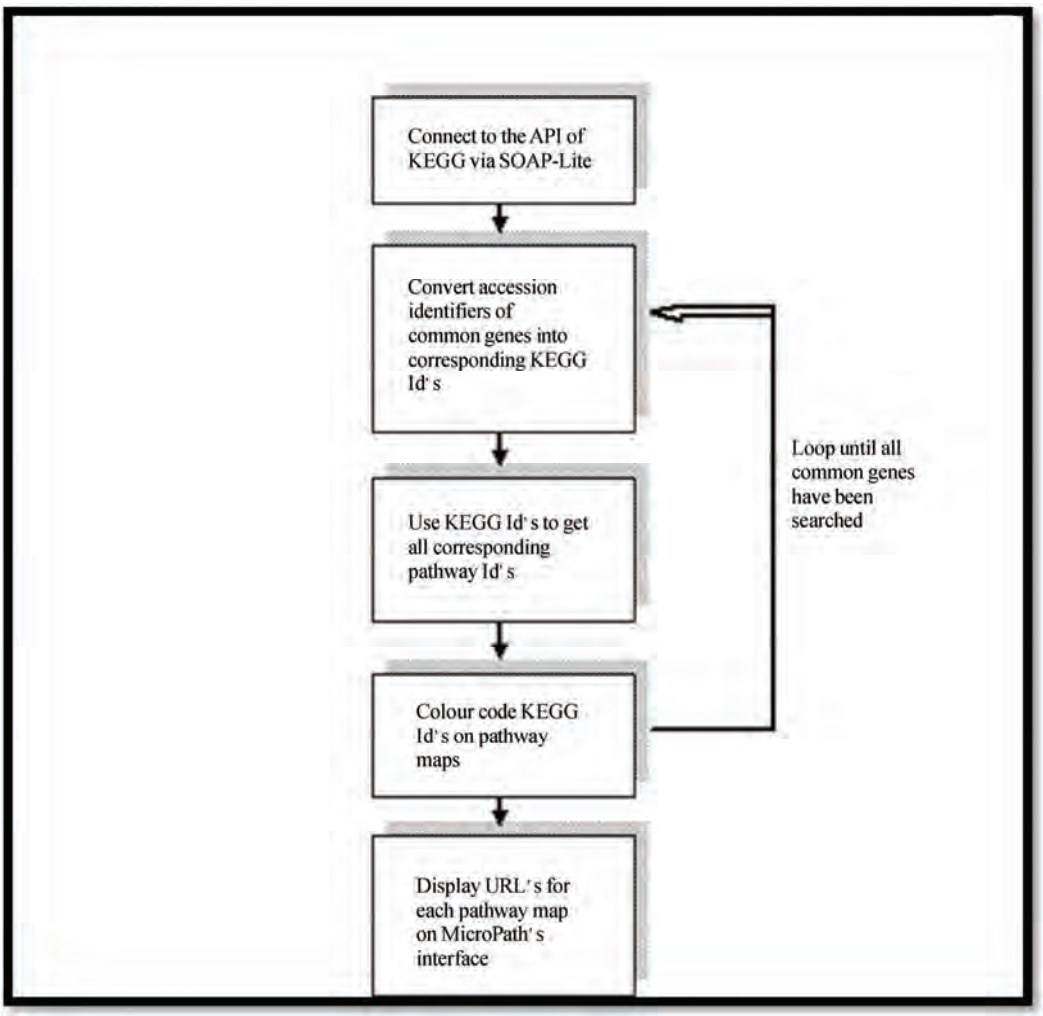


Figure 2. Flow diagram of how MicroPath carries out high throughput molecular pathway analysis by connecting to the API of KEGG.

2.5. Generating and Processing Gene Expression Datasets

Gene expression datasets used for the purpose of this article were generated from our in-house microarray experiments as well as published datasets, where the fold change approach was used to select a set of differentially expressed genes from pre-processed data. Matchminer [14] and the Synergizer [15] tools were used to convert gene Hugo identifiers and long names into Genbank accession Id's in order to ensure that the gene identifiers were of the same type across all datasets prior to comparison. Raw expression data was generated, filtered and normalised using GenePix pro 4.1 [16] and Acuity 4.0 [17] software. Although we used cDNA microarray data for the purpose of demonstrating MicroPath's capabilities, other data types generated from different platforms such as affymetrix can also be analysed provided Genbank ac--cession identifiers are used to represent the genes.

3. RESULTS AND DISCUSSION

Regardless of the biological question, a typical microarray experiment almost always results in the generation of a set of differentially expressed genes, which represents genes of most importance to the biologist. Therefore, by carrying out several biologically related microarray experiments, several sets of differentially expressed genes would be generated, which would need to be compared and mined efficiently in order to help answer the biological questions asked by the investigators from different research laboratories around the world. Employing manual methods of comparison in this situation would be very inefficient and infeasible. In light of this, to demonstrate the benefits that can be derived from analysing multiple gene expression profiles using MicroPath, we employed datasets generated from our in-house microarray experiments as well as published data. The biological question related to these studies focussed on unravelling the underlying molecular mechanisms dictating immune tolerance by analysing the role of Egr-2 in implicating T-cell tolerance. Although the Early Growth Response gene (Egr-2) has been recently characterised as a candidate tolerance-inducing transcription factor, which interacts with specific genes in order to induce the state of T-cell tolerance [18,19], the possibility of further putative unknown target genes exists that may be vital to the mechanism of tolerance. Hence, the biological purpose of our experiments was to attempt to identify such potentially important genes via the comparison of biologically related expression datasets using MicroPath.

Data consisting of a set of differentially expressed genes generated from the comparison of tolerance Vs activated mice CD4+ T cells was obtained from the ArrayExpress website (accession number: e-mexp-283). The first in-house experiment aimed to generate differentially expressed genes from the comparison of an un-stimulated T cell line from which the Egr-2 gene had been knocked out and a wild type un-stimulated cell line. The second in-house experiment focussed on the comparison between an Egr-2 knock-out T cell line activated with CD3/CD28 for 6 hours and a wild type cell line also activated with CD3/CD28 for 6 hours. Results generated from these experiments were then compared with the aforementioned published tolerance data using MiNer in order to understand the molecular mechanisms controlling immune tolerance.

3.1. Comparison of Gene Expression Profiles Pertaining to Immune Tolerance

The first step in the analysis was to subject the above-mentioned expression profiles to MicroPath in order to identify genes amongst them that had the same accession identifiers. Having done this, MicroPath identified 31 differentially expressed genes that were common to all three expression datasets and generated a graph to delineate their expression values (**Table 2**, **Figure 3**). A simple number crunching exercise was used to perform this task since its use generated a reasonable number of common genes, which did not warrant the use of permutations and combinations to perform the search. The next step was to use these 31 differentially expressed genes as a search space to determine those genes that have the potential to be co-expressed. In order to do this, we employed MicroPath's graphical utility to extract gene expression patterns, which led to the identification of 6/31 genes that were found to be upregulated in tolerance Vs activated CD4+T-cells and downregulated in both p-KOA0 Vs WTA0 and p-KOA6 Vs WTA6 datasets (**Table 2**). The remaining 25 common differentially expressed genes were found to be highly and lowly expressed in tolerance and knock-out datasets respectively. Statistical analysis revealed an overall pearson's correlation score of 0.109 from the pair-wise comparison of tolerance data with p-KOA0 Vs WTA0 and a score of -0.123 from the comparison of tolerance with p-KOA6 Vs WTA6. Furthermore, Reverse Transcriptase PCR experiments confirmed that 15 genes from our tolerance Vs activated data were found to be highly expressed in immune tolerance and from these 15 genes, 8 were found to be common amongst all three expression profiles (**Table 2**).

Because Egr-2 has been previously characterised and found to be highly upregulated in immune tolerance, these results generated from MicroPath are biologically significant because as expected, those genes that were highly expressed in our tolerance Vs activated datasets were found to be insignificantly expressed in our p-KOA6 Vs WTA6 and p-KOA0 Vs WTA0 datasets (from which the Egr-2 gene was knocked out of the cell lines). Amongst these genes, Ap1s1, Shd, Surf6, Vil2, Lilrb4, Tbx21 and Pdcd1lg2 (**Table 2**) have been confirmed to be upregulated in the process of immune tolerance [20], all of which were found to exhibit low expression values in our knock-out expression datasets. This consistent gene expression pattern can be seen graphically in **Figure 3**. However, from the 31 interesting common genes, 16 were not confirmed to be involved in

Table 2. Tabulated overview of gene accession ids, Hugo ids and fold change values belonging to 31 common genes identified from the comparison of tolerant Vs activated CD4+T cells, p-KOA0 Vs WTA0 and p-KOA6 Vs WTA6 expression datasets. Entries highlighted in bold represent genes that were found to be up-regulated in tolerance Vs activated CD4+ T cells and down-regulated in both p-KOA0 Vs WTA0 and p-KOA6 Vs WTA6 datasets. Entries with * represent genes that have been confirmed to be highly expressed in tolerance by RT-PCR.

Gene ID	HUGO ID	Fold Change (p-KOA0 Vs WTA0)	Fold Change (p-KOA6 Vs WTA6)	Fold Change (Tolerance Vs activated)
NM_007381	Acadl	0.371336	0.624525	6.373
NM_007457	Ap1s1 *	0.542474	0.31525	4.965
NM_007664	Cdh2	0.243646	-0.7999	1.658
NM_008205	H2-M9	-0.08048	0.116434	2.857
NM_008972	**Ptma**	**-1.31334**	**-0.46688**	**5.42**
NM_009128	**Scd2**	**-0.18816**	**-0.39366**	**4.552**
NM_009168	**Shd ***	**-0.17495**	**-0.53582**	**2.838**
NM_009298	Surf6 *	0.272072	0.126301	4.365
NM_009465	Axl	0.149539	1.475806	3.836
NM_009510	Vil2 *	-0.49824	0.319645	3.151
NM_010102	Edg6	0.313489	0.132689	1.573
NM_010413	**Hdac6**	**-0.90335**	**-0.8226**	**4.745**
NM_010548	Il10 *	3.083863	1.660739	3.521
NM_010638	Bteb1	0.024803	-0.42533	1.613
NM_011125	**Pltp**	**-0.5354**	**-0.71558**	**4.363**
NM_011620	Tnnt3	-0.61646	0.035844	1.665
NM_011696	Vdac3	-0.98084	0.191964	4.701
NM_011705	Vrk1	0.466922	-0.34601	2.032
NM_013488	Cd4	0.584494	0.420277	4.905
NM_013490	**Chka**	**-2.13728**	**-0.69458**	**5.677**
NM_013532	Lilrb4 *	0.792335	1.110898	2.111
NM_013615	Odf2	2.776384	3.004449	4.809
NM_013814	Galnt1	-0.47752	0.500297	2.246
NM_013866	Zfp385	0.118995	0.428591	1.664
NM_016772	Ech1	-0.0666	0.053081	4.284
NM_019507	Tbx21 *	0.124767	-0.32731	1.595
NM_019561	Ensa	0.778767	-0.44703	1.718
NM_019777	Ikbke	0.291602	-0.00772	1.609
NM_020027	Bat2	0.291219	-0.23966	5.091
NM_021396	Pdcd1lg2 *	1.140087	0.079182	3.921
NM_021538	Cope	0.154049	0.264541	2.035

tolerance by RT-PCR yet some of them also exhibited a coherent pattern of gene expression. For example, Ptma, Scd2, Hdac6, Pltp and Chka were all highly expressed in tolerance and conversely downregulated in both knock out datasets. There is a possibility that these genes may also be insignificantly expressed due to the absence of Egr-2. However, conducting RT-PCR for these specific genes would be required in order to confirm that their over-expression results in T-cell tolerance.

3.2. Deciphering Gene Regulatory Networks of Co-Xpressed Genes Via High Throug -Hput Molecular Pathway Analysis

The final stage of the analysis entails using MicroPath's function to connect to the Application Programming Interface (API) of KEGG via SOAP-Lite in order to carry out high throughput molecular pathway analysis. Therefore, for this stage in the analysis, we used MicroPath to map 31 of our co-expressed interesting genes to KEGG pathways and from these 31 genes, 14/31 were identified in a total of 31 molecular pathways (**Table 3**). Interestingly, several of these pathways were related to the study of immunology and illustrated biological networks such as MapKinase, Jak-Stat, T-cell receptor signalling and Cytokine-cytokine interactions. More specifically, the Pdcd1lg2 gene (accession id: NM_021396) was identified in the Cell Adhesion Molecules (CAM) pathway (Table 3) and studies have confirmed that the over-expression of Pdcd1lg2 has resulted in consistently low levels of Interleukin-2 (IL-2) in naive CD4(+) T-cells [21]. Further studies have correlated the over-expression of this gene to the negative regulation of T-cell activation. In one particular study, PDL2 (Pdcd1lg2) deficient mice were created in order to characterise the function of this gene in T-cell activation and tolerance, and results generated from this study suggested that Antigen-presenting cells from PDL2-deficient mice were found to be more potent in activating T-cells in vitro when compared to the wild-type counterparts [22]. These findings are conclusive and correlate well with the results generated from our in-house microarray experiments because using MicroPath to compare all three of our datasets followed by extracting gene expression patterns from them resulted in an important finding that Pdcd1lg2 was not only found to be over-expressed in tolerance (fold change of 3.921), but it was also under-expressed in our KOA0 Vs WTA0 and KOA6 Vs WTA6 knock-out datasets (with a fold change of 1.140 and 0.079 respectively) (**Table 2**). This

particular finding is in agreement with the aforementioned studies, concluding that Pdcd1lg2 has a negative inhibitory role towards the process of T-cell activation. In addition, molecular pathway analysis of the Interleukin-10 (IL-10) gene using MicroPath, identified its role in the Cytokine-cytokine interaction, Jak-STAT and T-cell receptor signalling pathways; all three of which are important immunological pathways. IL-10 is a well known cytokine, which has previously been shown to successfully induce immune tolerance in Dendritic Cells [23]. Results generated from MicroPath revealed that IL-10 was highly expressed in our tolerance data with a fold change of 3.521, which was found to be expressed lower in our KOA0 Vs WTA0 profile (fold change: 3.084). Interestingly, following activated with CD3/CD28 for 6 hours, its expression dropped significantly to 1.66, perhaps attributable to the absence of Egr-2. Likewise, other genes from the 31 co-expressed interesting genes show similar patterns of expression and perhaps may be candidate genes for Egr-2 mediated T-cell tolerance. However, this is yet to be confirmed by publications. Finally, the pathway analysis function of MicroPath was used to calculate the percentage of genes identified in each pathway in relation to 1) the intersection of common genes and 2) the total number of genes comprising each pathway. From the results, the Cell Adhesion Molecules (CAM) pathway was particularly significant since 12.91% of the overall pathway was affected by 6.84% of genes common to all 3 expression profiles (**Table 4**).

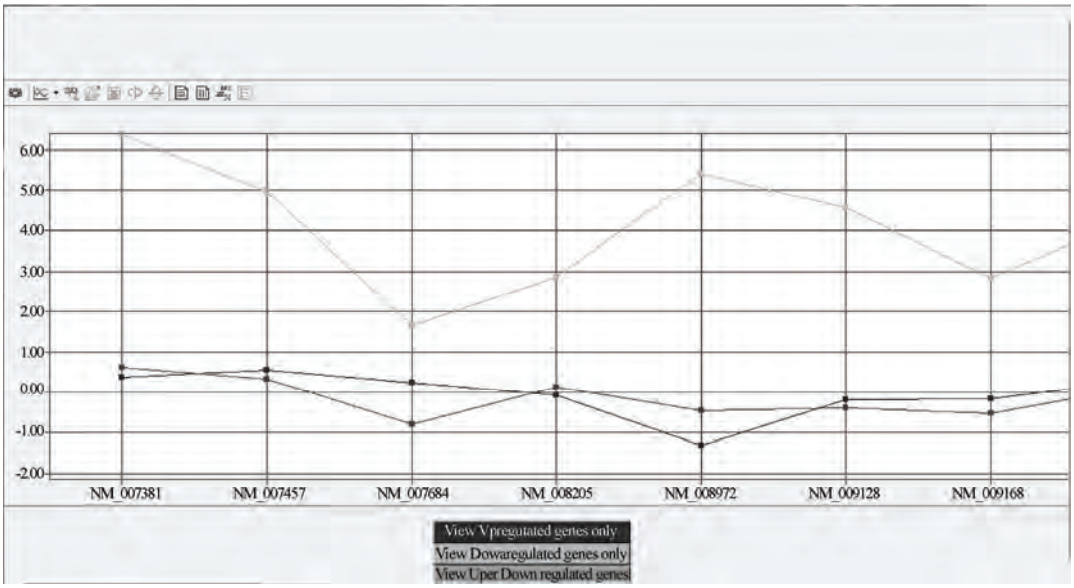


Figure 3. A preliminary graphical overview of common interesting genes generated from the comparison of tolerant Vs activated CD4+ T cells (green), p-KOA0 Vs WTA0 (red) and p-KOA6 Vs WTA6 (blue) expression datasets. It can be seen that genes that are highly expressed in tolerance appear to be expressed poorly in the knock-out datasets. This pattern is consistent throughout the 31 gene expression data points.

Table 3. Tabulated data generated from high throughput molecular pathway analysis of co-regulated genes. 14/31 common interesting genes were identified in a total of 31 molecular pathway maps of KEGG.

GenBank Accession ID	HUGO ID	Pathway ID	Total No of pathways	GenBank Accession ID	HUGO ID	Pathway ID	Total No of pathways
NM_007381	Acadl	mmu00071 mmu00280 mmu00410 mmu00640 mmu03320	5	NM_009510	Vil2	mmu04670 mmu04810	2
NM_007664	Cdh2	mmu04514	1	NM_008205	H2-M9	mmu04514 mmu04612 mmu04940	3
NM_013488	Cd4	mmu04514 mmu04612 mmu04640 mmu04660	4	NM_013814	Galnt1	mmu00512 mmu01030	2
NM_011696	Vdac3	mmu04020	1	NM_019777	Ikbke	mmu04010 mmu04620	2
NM_011125	Pltp	mmu03320	1	NM_010102	Edg6	mmu04080	1
NM_016772	Ech1	mmu00350 mmu00362 mmu00628	3	NM_021396	Pdcd1lg2	mmu04514	1
NM_010548	Il10	mmu04060 mmu04630 mmu04660	3	NM_013652	Ccl4	mmu04060 mmu04620	2

The fundamental strength of MicroPath stems from the implementation of a novel search strategy for the comparison of multiple gene expression profiles. Although there are a few software that cater for multiple gene expression comparison, there is currently no software that searches for common genes beyond simple number crunching methods of comparison (**Table 5**). Just because a direct comparison of a given number of datasets may not yield any common genes, it does not mean that the analysis should end here since there is a potential to identify common genes across *n*-1 profiles. MicroPath ensures that such genes are identified, which current software would overlook. When coupled with other important functions such as pattern extraction and pathway analysis, it becomes apparent that MicroPath would offer valuable assistance to biologists wanting to decipher their high throughput data.

Table 4. Results generated from pathway analysis showing the extent to which each pathway is affected by common genes from the intersection. The percentages reflect the proportion of common genes that contribute towards controlling the proportion of each pathway.

Pathway ID	Pathway Name	GenBank Accession ID	Result from Analysis
mmu00071	Fatty Acid Metabolism	NM_007381	3.26% of genes contribute 8.45% role in pathway
mmu00280	Valine, leucine and isoleucine degradation	NM_007381	3.26% of genes contribute 2.73% role in pathway
mmu00410	Beta Alanine Metabolism	NM_007381	3.26% of genes contribute 7.14% role in pathway
mmu00640	Propanoate Metabolism	NM_007381	3.26% of genes contribute 5.88% role in pathway
mmu03320	PPAR Signalling Pathway	NM_007381	3.26% of genes contribute 1.92% role in pathway
mmu04514	Cell Adhesion Molecules	NM_007664 NM_008205 NM_013488 NM_021396	12.91% of genes contribute 6.84 % role in pathway
mmu04612	Antigen Processing & Presentation	NM_013488	3.26% of genes contribute 2.44% role in pathway
mmu04640	Hematopoietic Cell Lineage	NM_013488	3.26% of genes contribute 0.76 % role in pathway
mmu04660	T Cell Receptor Signalling Pathway	NM_013488 NM_010548	6.45 % of genes contribute 3.33 % role in pathway
mmu04020	Calcium Signalling Pathway	NM_011696	3.26% of genes contribute 2.33 % role in pathway
mmu00350	Tyrosine Metabolism	NM_016772	3.26% of genes contribute 2.17 % role in pathway
mmu04060	Cytokine-cytokine receptor interaction	NM_010548 NM_013652	6.45 % of genes contribute 0.73 % role in pathway
mmu04630	JAK-STAT Signalling Pathway	NM_010548	3.26% of genes contribute 3.85 % role in pathway
mmu04670	Leukocyte Transendothelial Migration	NM_009510	3.26% of genes contribute 1.25 % role in pathway
mmu04810	Regulation of Actin Cytoskeleton	NM_009510	3.26% of genes contribute 1.47 % role in pathway
mmu04940	Type I Diabetes Mellitus	NM_008205	3.26% of genes contribute 4.35 % role in pathway
mmu00512	O-Glycan Biosynthesis	NM_013814	3.26% of genes contribute 10 % role in pathway
mmu04010	MAPK Signalling Pathway	NM_019777	3.26% of genes contribute 0.83 % role in pathway
mmu04620	Toll-Like Receptor Signalling Pathway	NM_019777 NM_013652	6.45% of genes contribute 1.32 % role in pathway
mmu04080	Neuroactive Ligand-Receptor Interaction	NM_010102	3.26% of genes contribute 1.15 % role in pathway

Table 5. Functional comparison of MicroPath to similar software packages and applications.

Function	MicroPath	EXPANDER [24]	INCLUSIVE [25]	Pathway Studio [26]	KEGG [13]	BioCarta [27]	MaXlab [28]
Suitable for high throughput data analysis	YES	YES	YES	YES	NO	NO	YES
Suitable for comparing multiple gene expression profiles	YES	YES	NO	YES	NO	NO	YES
Implementation of efficient algorithm to search for common genes from *n*-1 datasets	YES	NO	NO	NO	NO	NO	NO
Graphical representation of gene expression values from multiple datasets	YES	NO	NO	NO	NO	NO	YES
Pattern extraction from Graph data	YES	NO	NO	NO	NO	NO	NO
Construction of pathway maps	YES	NO	NO	YES	YES	YES	NO
Mapping gene expression data to pathway maps	YES	NO	NO	YES	NO	NO	NO
User interactive software (S) or Database (D)	S	S	S	S	D	D	S

4. Conclusion

In this article, we have illustrated the potential benefits that can be derived from using MicroPath for the analysis of multiple gene expression profiles. Each function of the software has been developed to streamline the overall analysis pipeline, providing users with a walkthrough of how their data is biologically deciphered. Here, we have applied to our software, microarray datasets generated from different laboratories pertaining to the molecular mechanisms underlying immune tolerance. However, MicroPath is capable of analysing data for any given biological question, whether the datasets are taken from public repositories such as ArrayExpress or generated from in-house microarray experiments. We believe that its faculty to use both number crunching and permutations and combinations as the search strategy to identify the intersection of common genes, coupled with its function to extract gene expression patterns graphically and statistically makes this an attractive software for biologists to use. Finally, its ability to carry out live streaming of mapping genes to biological pathways makes it a useful tool for the automation of multiple gene expression analysis.

Availability and requirements
Project name: MicroPath
Project home page: www.1066technologies.co.uk/micropath
Operating system(s): MicroPath has been tested on Windows 2000, XP and Vista
Programming language: Visual Basic.Net, Perl
Other requirements: None
License: N/A
Any restrictions to use by non-academics: No

Acknowledgements

This study was supported by grants from the UK Medical Research Council (MRC) (Grant number: G0300520).

REFERENCES

[1] U. Sarkans, H. Parkinson, G. G. Lara, A. Oezcimen, A. Sharma, N. Abeygunawardena, S. Contrino, E. Holloway, P. Rocca-Serra, G. Mukherjee, M. Shojatalab, M. Kapushesky, S. A. Sansone, A. Farne, T. Rayner and A. Brazma. (2005) The ArrayExpress gene expression database: a software engineering and implementation perspective. Bioinformatics 21(8): 1495– 1501.

[2] T. Barrett and R. Edgar. (2006) Mining Microarray Data at NCBI's Gene Expression Omnibus (GEO). Methods Mol Biol 338: 175–190.

[3] D. Ghosh, Barette, T. R., Rhodes, D. and Chinnaiyan, A. M. (2003). Statistical issues and methods for meta-analysis of microarray data, A case study in prostate cancer. Funct. Integr. Genomics 3, 180–188.

[4] D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh and A. M. Chinnaiyan, (2002). Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res. 62, 4427–4433.

[5] D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey and A. M. Chinnaiyan (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc. Natl. Acad. Sci. USA 101, 9309–9314.

[6] J. Wang, K. R. Coombes, W. E. Highsmith, M. J. Keating and L. V. Abruzzo (2004). Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: A meta-analysis of three microarray studies. Bioinformatics 20, 3166–3178.

[7] S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu and Romero, R. (2007). A systems biology approach for pathway level analysis. Genome Res. 17, 1537–1545.

[8] J. Stelling, (2004). Mathematical models in microbial systems biology. Curr. Opin. Microbiol. 7, 513–518.

[9] G. Joshi-Tope, M. Gillespie, I. Vasrik, P. D'Eustachio, E. Schmidt, B. de Bone, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, *et al*. (2005). A knowledgebase of biological pathways. Nucleic Acids Res. 33, D428–D432.

[10] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, *et al*. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. 102: 15545–15550.

[11] S. Khalid, F. Fraser, M. Khan, P. Wang, X. Liu and S. Li, (2006a). Analysing Microarray Data using the Multi-functional Immune Ontologiser. J. Integrative Bioinformatics 3, 25.

[12] S. Khalid, M. Khan, P. Wang, X. Liu and S. -L. Li, (2006b). Application of bioinformatics in the design of gene expression microarrays. Second International Symposium on Leveraging Applications of Formal Methods, Verification and Validation (isola 2006), pp. 146–160.

[13] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno and M. Hattori, (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res, 32.

[14] K. J. Bussey, D. Kane, M. Sunshine, S. Narasimhan, S. Nishizuka, W. C. Reinhold, B. Zeeberg, W. Ajay and J. N. Weinstein, (2003) MatchMiner: a tool for batch navigation among gene and gene product identifiers. Genome Biology. 4, R27.

[15] G. F. Berriz and F. P. Roth, The Synergizer service for translating gene, protein, and other biological identifiers. (2008). Bioinformatics. [Epub ahead of print].

[16] GenePix pro 4.1: http://www.axon.com

[17] Acuity 4.0: http://www.moleculardevices.com/pages/software/gn_acuity.html

[18] M. Safford, S. Collins, M. A. Lutz, A. Allen, C. Huang, J. Kowalski, A. Blackford, M. R. Horton, C. Drake, R. H. Schwartz and J. D. Powell, (2005) Egr-2 and Egr-3 are negative regulators of T cell activation. Nature Immunology 6 472–480.

[19] L. E. Warner, J. Svaren, J. Milbrandt and J. R. Lupski, (1999) Functional consequences of mutations in the early growth response 2 gene (EGR2) correlate with severity of human myelinopathies. Hum. Mol. Genet. 8 1245–1251.

[20] P. O. Anderson, B. A. Manzo, A. Sundstedt, S. Minaee, A. Symonds, S. Khalid, M. E. Rodriguez-Cabezas, K. Nicolson, S. Li, D. C. Wraith and P. Wang, (2006) Persistent antigenic stimulation alters the transcription program in T cells, resulting in antigen-specific tolerance. European Journal of Immunology. 36, 1374–85.

[21] H. Kuipers, F. Muskens, M. Willart, D. Hijdra, F. B. van Assema, A. J. Coyle, H. C. Hoogsteden and B. N. Lambrecht (2006). Contribution of the PD-1 ligands/PD-1 signaling pathway to dendritic cell-mediated CD4 (+) T cell activation. European Journal of Immunology. 36 (9), 2472–82.

[22] Y. Zhang, Y. Chung, C. Bishop, B. Daugherty, H. Chute, P. Holst, C. Kurahara, F. Lott, N. Sun, A. A. Welcher and C. Dong, (2006). Regulation of T cell activation and tolerance by PDL2. Proc Natl Acad Sci U S A, 103(31), 11695–11700.

[23] X. Li, K. Dou, H. Liu, F. Zhang and L. Cai, (2007). Immune tolerance induced by IL-10 and methylprednisolone modified dendritic cells in vitro. Chinese Journal of cellular and molecular Immunol. 23 (5), 436−8.

[24] R. Shamir, A. Maron-Katz, A. Tanay, C. Linhart, I. Steinfeld, R. Sharan, Y. Shiloh and R. Elkon, (2005) EXPANDER-an integrative program suite for microarray data analysis. BMC Bioinformatics, 6: 232.

[25] G. Thijs, Y. Moreau, F. D. Smet, J. Mathys, M. Lescot, S. Rombauts, P. Rouze, B. D. Moor and K. Marchal, (2002) INCLUSive: Integrated Clustering, Upstream sequence retrieval and motif Sampling. Bioinformatics, 18, 331−332.

[26] A. Nikitin, S. Egorov, N. Daraselia and I. Mazo,. (2003) Pathway studio-the analysis and navigation of molecular networks. Bioinformatics, 19, 2155−2157.

[27] BioCarta, Charting pathways of life. http://www.biocarta.com.

[28] S. Khalid, M. Khan, C. B. Gorle, K. Fraser, P. Wang, X. Liu and S. Li, MaXlab: A novel application for the cross comparison and integration of biological signatures from microarray studies. In Silico Biology 8, 0029: 2008.

J. Biomedical Science and Engineering, 2009, 2, 117-122

Prediction of mutation position, mutated amino acid and timing in hemagglutinins from North America H1 influenza A virus

Shao-Min Yan[1], Guang Wu[2*]

[1]Guangxi Academy of Sciences, 98 Daling Road, Nanning, Province Guangxi 530007, China. [2]DreamSciTech Consulting, 301, Building 12, Nanyou A-zone, Jiannan Road, Shenzhen, Province Guangdong 518054, China. Correspondence should be addressed to Guang Wu (hongguanglishibahao@yahoo.com)

Received Jan. 8th, 2009; revised Feb. 9th; 2009; accepted Feb. 10th, 2009

ABSTRACT

This study was trying to predict the mutations in H1 hemagglutinins of influenza A virus from North America including the predictions of mutation position, the predictions of would-be-mutated amino acids and the predictions of time of occurrence of mutations. The results paved a possible way for accurate, precise and reliable prediction of mutation in proteins from influenza A virus.

Keywords: Hemagglutinin, Influenza, Mutation, Neural Network, Prediction

1. INTRODUCTION

Mathematical modelling provides a promising hope to predict the mutation in proteins from influenza A virus, not only because the history shows that accurate, precise and reliable predictions are mainly based on mathematical modelling, but also the prediction of mutations at protein can be classified as prediction of mutation position, prediction of mutated amino acid and timing of mutation [1]. All these require the use of more sophisticated mathematical tools.

Perhaps, the best way to predict the mutation is to find its cause, thus a mutation could occur if the same cause appears again. However, the causes, which led to historical mutations, might not leave any sign to trace, and the evolved proteins from influenza A virus may no longer be sensitive to the causes, which led to mutations in the past. All these mean that the mutation causes would be poor predictors for prediction of mutations, while the preparedness for possible pandemic/epidemic of influenza would lag behind the appearance of influenza without prediction. On the other hand, no matter what mutation cause is, any cause would leave signs in a protein, otherwise, no mutation would be recorded. These signs can be arguably used for prediction. This is the basic consideration for prediction of mutations using modelling.

Generally, the amino acids in a protein is represented as alphabet, thus a number of models use amino-acid symbols as operating units, for example, sequence alignment, phylogenetics, and multi-sequence comparison, by which the history of proteins of interests can be traced [2,3]. Unfortunately, these symbol-based approaches cannot accurately and precisely answer the predictions proposed, because they cannot operate in sophisticated mathematical models, whose operating units are values.

In this view, the protein science actually is at the historical phase of searching for the ways to represent a protein sequence as a numeric sequence, and it is hoped that the numeric sequence is sensitive to mutations, positions of amino acids in protein sequence, composition of protein sequence, length of protein sequence, neighbouring amino acids.

In fact, currently there are several ways to transfer a protein sequence into a numeric sequence, and the most profound one would be the use of the physicochemical property to represent a protein sequence [4] as well as related approaches [5-10].

On the other hand, other approaches are also developed, for example, the approaches based on random mechanism to quantify each amino acid in a protein as well as a protein in whole [1].

This study was designed to predict the mutation positions, the mutated amino acids and the time of occurrence of mutations in the hemagglutinins from North America H1 influenza A viruses using neural network, because the hemagglutinin is the major surface antigen of influenza viruses, against which neutralizing antibodies are elicited during virus infection and vaccination [11-15]. Among various types, the H1 influenza virus is the cause for several historical disasters, such as 1918 Spanish flu, 1977 Russian flu, 1950 and 1988 epidemics [16-18].

2. MATERIALS AND METHODS

The amino acid sequences and corresponding RNA sequences of 494 hemagglutinins from North America influenza A/H1 viruses isolated from 1918 to 2008 were obtained from the influenza virus resources [19]. Forty-

six identical hemagglutinins were excluded, thus the remaining 448 hemagglutinins were used in this study.

2.1 Amino-Acid Pair Predictability

According to the permutation [1, 20, 21], for example, there are 47 asparagines "N" and 37 valines "V" in the hemagglutinin, strain A/swine/Ontario/53518/03(H1N1), accession number DQ280219, the frequency of amino-acid pair NV is 3 (47/566×37/565×565=3.072), that is, NV would appear three times in this hemagglutinin. Actually 3 NVs can be found in this hemagglutinin, so NV is predictable and the difference between its predicted and actual frequency is 0. Again, there are 48 leucines "L" in DQ280219 hemagglutinin, and the frequency of random presence of LL is 4 (48/566×47/565×565=3.986), i.e. there would be four LLs in the hemagglutinin. But LL appears nine times in reality, so the difference between its predicted and actual frequency is -5. After such calculations [22], each amino-acid pair had its difference between predicted and actual frequency. As a point mutation is relevant to a single amino acid, which connects with two neighbouring amino acids except for the terminal one and constructs two amino-acid pairs, so each amino acid has the sum of difference between predicted and actual frequency in two neighbouring amino-acid pairs, which is the first quantification for each amino acid in a hemagglutinin. Nevertheless, any hemagglutinin must have a certain amount of predictable amino-acid pairs, by which the percentage of how many amino-acid pairs predictable can be found. This predictable portion is the quantification for a whole hemagglutinin.

2.2 Amino-Acid Distribution Probability

According to the occupancy of subpopulations and partitions, the positions of any type of amino acids in hemagglutinin can be viewed as a certain distribution [10, 23-31], whose probability is $\frac{r!}{q_0!\times q_1!\times\ldots\times q_n!}\times\frac{r!}{r_1!\times r_2!\times\ldots\times r_n!}\times n^{-r}$ [32], where r is the number of amino acids, n is the number of partitions, r_n is the number of amino acids in the n-th partition, q_n is the number of partitions with the same number of amino acids, and ! is the factorial function. For instance, there are 36 lysines "K" in DQ280219 hemagglutinin. Their predicted and actual distribution probabilities are 0.0419 and 0.0020 [33], so the ratio of predicted versus actual distribution probabilities is 20.95, whose natural logarithm is 3.0421, which is the second quantification for each amino acid in a hemagglutinin.

2.3 Future Composition of Amino Acids

The relationship between 64 RNA codons and translated amino acids is governed by translation probability [1, 34-36], based on which the amino acid mutating probability can be determined. For example, alanine "A" has the 12/36 chance of mutating to "A", but cysteine "C" has no chance of mutating to "A", then both aspartic acid "D" and glutamic acid "E" have the 2/18 chance of mutating to "A", and so on. In total, the future composition of amino acid "A" is 6.1271% in DQ280219 hemagglutinin, whereas its current composition is only 5.1146% (29/567), and the ratio is 1.1980 (6.1271%/5.1146%), thus the future composition of amino acids is got [1], and assigned the ratio of predicted versus actual compositions to each amino acid [37], which is the third quantification for each amino acid in hemagglutinin [1].

Although there are countless mutation causes impacting a parent protein, these causes should leave their traces in the protein, which should be measured out using these three quantifications, which in fact represent the countless mutation causes.

2.4 Prediction of Mutation Position

Any mutation cause can lead to occurrence or non-occurrence of mutation, which can be classified as unity and zero after comparing a parent protein with its daughter protein. In this way, the occurrence or non-occurrence of mutation in a parent protein becomes a binary sequence. Thus, two datasets can be got, the mutation cause dataset, which are three quantifications, and the mutation consequence dataset, which is a binary sequence. Moreover, these two datasets have the position-to-position relationship (**Table 1**), which is the cause-mutation relationship. Mathematically this relationship is the problem of classification, which can be solved either using the logistic regression in statistics or neural network. The feed forward backpropagation neural network

Table 1. Inputs and target of DQ280219 hemagglutinin sequence.

Position	Amino acid	Quantified hemagglutinin sequence			Mutation sequence
		I	II	III	
1	M	−2	0.0000	1.2569	0
…	…	…	…	…	…
276	R	−2	1.2809	1.9392	0
277	G	−1	2.3790	0.7887	0
278	H	0	0.0000	1.2396	1
279	G	0	2.3790	0.7887	1
280	S	1	4.0008	1.1081	0
…	…	…	…	…	…
566	I	−2	1.1285	0. 9590	0

would be applied to this relationship to predict the mutation position.

2.5 Prediction of would-be-mutated Amino Acid

The prediction was made using the amino-acid mutating probability, which was based on the relationship between RNA codons and translated amino acids [1].

2.6 Timing of Mutation

As each hemagglutinin is different one from another due to mutation, each hemagglutinin would be quantified differently one from another. Along the time axis, all hemagglutinins would construct their evolutionary process, and the timing of the mutation would be possible by detailed analysis of this evolutionary process.

2.7 Software and Statistics

The MatLab software [38] was used for prediction. The prediction sensitivity, specificity and total correct rate were calculated according to the published method [39].

3. RESULTS AND DISCUSSION

The performance of modelling was measured using the prediction sensitivity (42.9%±31.4%), specificity (99.5% ±0.4%) and total correct rate (99.0%±0.4%), because the predicted mutation positions can be classified as the positives, false positives, negatives and false negatives when comparing the predicted with the actual mutation positions. As seen, the prediction sensitivity was still low although the prediction specificity and total correct rate were quite high.

After the prediction of possible mutation positions using the neural network, the would-be-mutated amino acids at predicted positions can be predicted using the amino-acid mutating probability [1]. **Figure 1** illustrates the prediction of possible mutation positions and mutated amino acids at the predicted positions, where the solid line in the lower panel is the predicted mutation probability by the neural network with respect to each amino acid in ABX58602 hemagglutinin and the dotted line is the cut-off mutation probability of 0.5, that is, the amino acid whose mutation probability is larger than 0.5 risks mutation. For this hemagglutinin, there were two positions (98 and 507) whose mutation probability was larger than 0.5, so the amino acid E at these positions would have a larger chance of mutation. Meanwhile, the would-be-mutated amino acid can be determined using the amino-acid mutating probability (upper panel), where the amino acid "D" has the largest chance to appear. So the lower panel indicates the possible mutation positions with probability, and the upper panel displays the would-be-mutated amino acids with probability.

Figure 2 displays the evolution of North America H1 hemagglutinins. This predictable portion fluctuated over time, which represented the mutation process. With fast Fourier transform, which is suitable to find the periodicity in chaotic dataset, the mutation periodicity can be found from **Figure 2**, where (i) the evolutionary process of influenza A virus hemagglutinins from 1978 to 2008 contained many periodicities; (ii) each periodicity suggested different number of mutations along the time course; (iii) the periodicity with the biggest number of mutations was about 5.6 years, thus the time when mutations would occur in future can be estimated; and (iv) the hemagglutinin periodicity provides the chance to trace the possible mutation cause in nature, because each periodicity may correspond to a natural phenomenon.

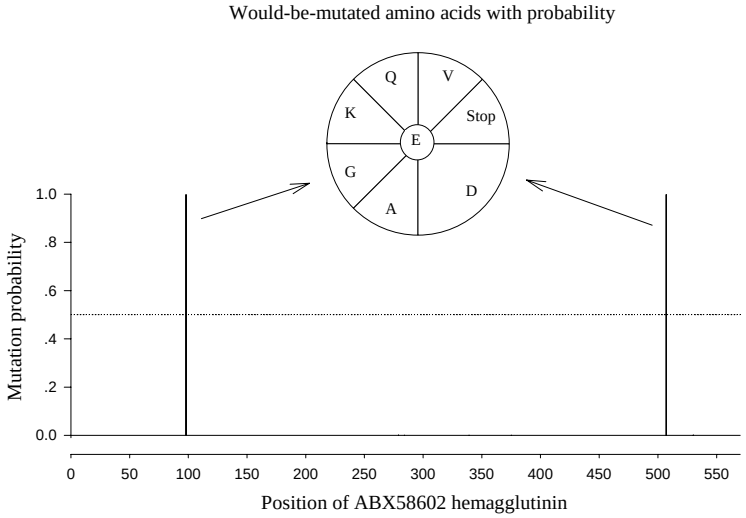


Figure 1. Prediction of mutation positions and would-be-mutated amino acids. On the lower panel: the x-axis represents the position of ABX58602 hemagglutinin from 1 to 565, because ABX58602 hemagglutinin is composed of 565 amino acids; the y-axis represents the mutation probability predicted using neural network model, where there are two probabilities larger than 0.5 at positions 98 and 507. On the upper panel, the centre of pie is labelled as "E" glutamic acid, which is the amino acid at positions 98 and 507 of ABX58602 hemagglutinin. The other letters represent the would-be-mutated amino acids, and the area occupied by letter represents the probability to mutate to this amino acid based on the amino-acid mutating probability, for example, "E" has the largest chance to mutate to "D".

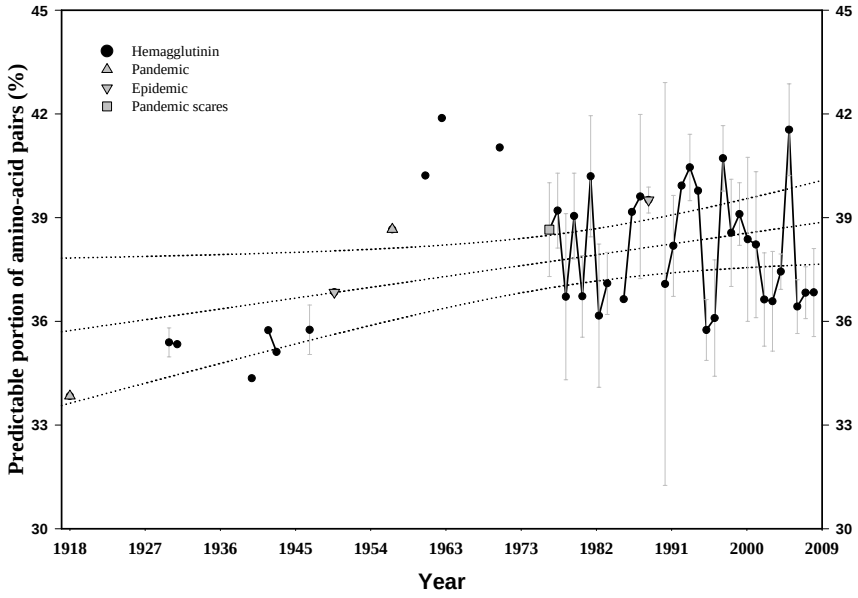


Figure 2. Evolution of 448 hemagglutinins of North America H1 influenza viruses. The data are presented as mean±SD. The dotted lines are regressed lines 95% confidence intervals.

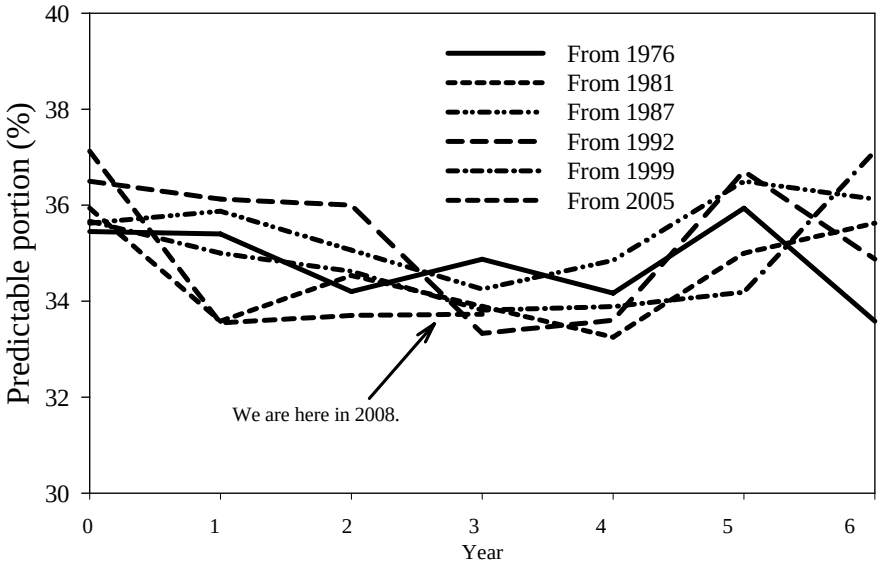


Figure 3. Stratification of hemagglutinin evolution after finding the periodicity using fast Fourier transform.

Furthermore, an attempt was made to time the mutation by stratifying the hemagglutinin evolution in **Figure. 2** according to its periodicity, and **Figure 3** shows such an example, where the hemagglutinin evolution in **Figure 2** is stratified according to 6-year periodicity because it was the periodicity with biggest number of mutations. **Figure 3** shows that there would be a 2-year stable period before possible more mutations would occur in 2010.

At this stage of development, it is yet to verify the predictions made in this study. However, this is not uncommon phenomenon in science, because the first step is to find a way to transfer the measurements into the domain, where a mathematical model can be applied, the second step is build a model, and the third step is to make the predictions. These three steps are more related to theoretical work. Thereafter the last step would be the verification experimentally, which is certainly beyond the scope of this paper. On the other hand, the science advances so much, it is impossible to verity each hypothesis and prediction, for example, the humans cannot create another earth without global warming to compare the effects on subjects of interests. With respect to the predictions in this study, the verifications can be done by using the same method in man-made mutations in industrial enzymes, where each mutation can be recorded and compared with prediction.

The frequency of mutations is not identical along a hemagglutinin sequence, namely, the different position has different chance of mutations. In fact, the prediction made in this study is consistent with this observation as seen in **Figure 1**, where the predicted mutation probability is not identical along the ABX58602 hemagglutinin.

To the best of knowledge, there are several models conducted at different levels for the prediction of possible pandemic/epidemic of influenza. At epidemiological level, the predictions were made using early indicators [40], time series analysis [41,42], etc. At clinical level, the prediction was made using medical visit [43], outbreak signatures [44], etc. At social level, the prediction was made using sales of computer printers, elections, and the Federal Reserve's decisions about interest rates [45]. At seroarcheological level, the prediction was made using accumulation of mutations or true recombinational events [46]. At protein level, the prediction was made with epitope [47-49], conformation [50]. However, no similar prediction was made with respect to the approached used in this study. The difference includes: (1) the quantification of protein sequences in this study was based on the random principle, (2) the occurrence and non-occurrence of mutation was quantified as yes-no event, (3) the cause-mutation relationship was defined using neural network, (4) the would-be-mutated amino acid was determined using the amino-acid mutating probability, and (5) the time of mutation was determined using the fast Fourier transform to stratify the time interval between outbreak of influenza.

This study paved a possible way for accurate, precise and reliable prediction of mutation in proteins from influenza A virus, because the model in prediction was the cause-mutation model, which was helpful for understanding of underlined mutation mechanism.

ACKNOWLEDGEMENTS

This study was supported in part by National Natural Science Foundation No. 20666002 (Guangxi Assignment No. 0728001).

REFERENCES

[1] G. Wu & S. Yan. (2008) Lecture Notes on Computational Mutation. Nova Science Publishers, New York.

[2] E. Ghedin, N. A. Sengamalay, M. Shumway, J. Zaborsky, T. Feldblyum, V. Subbu, D.J. Spiro, J. Sitz, H. Koo, P. Bolotov, D. Dernovoy, T. Tatusova, Y. Bao, K. St George, J. Taylor, D.J. Lipman, C.M. Fraser, J.K. Taubenberger & S.L. Salzberg. (2005) Large−scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. Nature, 437, 1162−1166.

[3] J. C. Obenauer, J. Denson, P. K. Mehta, X. Su, S. Mukatira, D. B. Finkelstein, X. Xu, J. Wang, J. Ma, Y. Fan, K.M. Rakestraw, R.G. Webster, E. Hoffmann, S. Krauss, J. Zheng, Z. Zhang & C.W. Naeve. (2006) Large−scale sequence analysis of avian influenza isolates. Science, 311, 1576−1580.

[4] K. C. Chou. (2004) Structure bioinformatics and its impact to biomedical science. Curr. Med. Chem, 11, 2105−2134.

[5] K. C. Chou. (2004) Modelling extracellular domains of GABA−A receptors: subtypes 1, 2, 3, and 5. Biochem. Biophys. Res. Commun, 316, 636−642.

[6] K. C. Chou. Insights from modelling the 3D structure of extracellular domain of alpha 7 nicotinic acetylcholine receptor. Biochem. Biophys. Res. Commun, 319, 433−438.

[7] K. C. Chou. (2004) Coupling interaction between thromboxane A2 receptor and alpha−13 subunit of guanine nucleotide−binding protein. J. Proteome Res. 2005, 4, 1681−1686.

[8] X. Xiao, S. Shao, Y. Ding, Z. Huang, X. Chen & K.C. Chou. (2005) An application of gene comparative image for predicting the effect on replication radio by HBV virus gene missense mutation. J. Theo. Biol, 235, 555−565.

[9] X. Xiao, S. H. Shao & K. C. Chou. (2006) A probability cellular automation model for hepatitis B viral infections. Biochem. Biophys. Res. Commun, 342, 605−610.

[10] S. Yan & G. Wu. (2008) Quantitative relationship between mutated amino−acid sequence of human copper−transporting ATPases and their related diseases. Mol. Divers, 12, 119−129.

[11] Q. S. Du, S.Q. Wang & K. C. Chou. (2007) Analogue inhibitors by modifying oseltamivir based on the crystal neuraminidase strcutre for trating drug−resistant H5N1 virus. Biochem. Biophys. Res. Commun, 363, 525−531.

[12] J. R. Schnell & J. J. Chou. (2008) Structure and mechanism of the M2 proton channel of influenza A virus. Nature, 451, 591−595.

[13] S. Q. Wang, Q. S. Du & K. C. Chou. (2007) Study of drug resistance of chicken influenza A virus (H5N1) from homology−modeled 3D structure of neuraminidases. Biochem. Biophys. Res. Commun, 354, 634−640.

[14] D.Q. Wei, Q.S. Du, H. Sun & K.C. Chou. (2006) Insights from modeling the 3D structure of H5N1 influenza virus neuraminidase and its binding interactions with ligands. Biochem. Biophys. Res. Commun, 344, 1048−1055.

[15] D. C. Wiley & J. J. Skehel. (1987) The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. Annu. Rev. Biochem, 56, 365−394.

[16] Y. Kanegae, S. Sugita, K. F. Shortridge, Y. Yoshioka & K. Nerome. (1994) Origin and evolutionary pathways of the H1 hemagglutinin gene of avian, swine and human influenza viruses: cocirculation of two distinct lineages of swine virus. Arch. Virol, 134: 17−28.

[17] A.H. Reid, T.G. Fanning, J.V. Hultin & J.K. Taubenberger. (1999) Origin and evolution of the 1918 "Spanish" influenza virus hemagglutinin gene. Proc. Natl. Acad. Sci. USA , 96, 1651−1656.

[18] J. K. Taubenberger, A. H. Reid, A. E. Krafft, K. E. Bijwaard & T. G. Fanning. (1997) Initial genetic characterization of the 1918 “Spanish” influenza virus. Science , 275, 1793−1796.

[19] Influenza virus resources. (2008) http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/multiple.cgi.

[20] G. Wu & S. Yan. Randomness in the primary structure of protein: methods and implications. Mol. Biol. Today 2002, 3: 55−69.

[21] G. Wu & S. Yan. (2006) Mutation trend of hemagglutinin of influenza A virus: a review from computational mutation viewpoint. Acta Pharmacol. Sin, 27: 513−526.

[22] Amino-acid pair predictability. (2008) http://www.dreamscitech.com/Service/rationale.htm.

[23] G. Wu & S. Yan. (2000) Prediction of distributions of amino acids and amino acid pairs in human haemoglobin α−chain and its seven variants causing −thalassemia from their occurrences according to the random mechanism. Comp. Haematol. Int, 10, 80−84.

[24] G. Wu & S. Yan. (2001) Analysis of distributions of amino acids, amino acid pairs and triplets in human insulin precursor and four variants from their occurrences according to the random mechanism. J. Biochem. Mol. Biol. Biophys, 5, 293−300.

[25] G. Wu & S. Yan. (2001) Analysis of distributions of amino acids and amino acid pairs in human tumor necrosis factor precursor and its eight variants according to random mechanism. J. Mol. Model, 7, 318−323.

[26] G. Wu & S. Yan. (2002) Random analysis of presence and absence of two-and three-amino-acid sequences and distributions of amino acids, two-and three-amino-acid sequences in bovine p53 protein. Mol. Biol. Today, 3: 31−37.

[27] G. Wu & S. Yan. (2002) Analysis of distributions of amino acids in the primary structure of apoptosis regulator Bcl−2 family according to the random mechanism. J. Biochem. Mol. Biol. Biophys, 6, 407−414.
[28] G. Wu & S. Yan. (2002) Analysis of distributions of amino acids in the primary structure of tumor suppressor p53 family according to the random mechanism. J. Mol. Model, 8, 191−198.
[29] G. Wu & S. Yan. (2004) Determination of sensitive positions to mutations in human p53 protein. Biochem. Biophys. Res. Commun, 321, 313−319.
[30] G. Wu & S. Yan. (2005) Searching of main cause leading to severe influenza A virus mutations and consequently to influenza pandemics/epidemics. Am. J. Infect. Dis., 1, 116−123.
[31] G. Wu & S. Yan. (2005) Prediction of mutation trend in hemagglutinins and neuraminidases from influenza A viruses by means of cross-impact analysis. Biochem. Biophys. Res. Commun., 326, 475−482.
[32] W. Feller. (1968) An Introduction to Probability Theory and Its Applications. 3rd ed, Vol, I. Wiley, New York, p. 34−40.
[33] Amino−acid distribution probability. (2008) http://www.dreamscitech. com/Service/timing.htm.
[34] G. Wu & S. Yan. (2005) Determination of mutation trend in proteins by means of translation probability between RNA codes and mutated amino acids. Biochem. Biophys. Res. Commun, 337, 692−700.
[35] G. Wu & S. Yan. (2006) Determination of mutation trend in hemagglutinins by means of translation probability between RNA codons and mutated amino acids. Protein Pept. Lett, 13, 601−609.
[36] G. Wu & S. Yan. (2007) Translation probability between RNA codons and translated amino acids, and its applications to protein mutations. In: Leading-Edge Messenger RNA Research Communications. ed. Ostrovskiy M. H. Nova Science Publishers, New York, Chapter 3, 47−65.
[37] Amino-acid mutating probability. (2008). http://www.dreamscitech. com/Service/lag.htm.
[38] MathWorks Inc. (2001) MatLab-The Language of Technical Computing (version 6.1.0.450, release 12.1), 1984−2001.
[39] Systat Software Inc. Systat for Windows, version 11.00.01. 2004.
[40] E. Andersson, S. Kühlmann−Berenzon, A. Linde, L. Schiöler, S. Rubinova & M. Frisén. (2008) Predictions by early indicators of the time and height of the peaks of yearly influenza outbreaks in Sweden. Scand. J. Public Health, 36, 475−482.
[41] Y. T. Li, H. W. Zhang, H. Ren, J. Chen & Y. Wang. (2007) Application of time series analysis in the prediction of incidence trend of influenza−like illness in Shanghai. Zhonghua Yu Fang Yi Xue Za Zhi, 41, 496−498.
[42] J. Saltyte Benth & D. Hofoss. (2008) Modelling and prediction of weekly incidence of influenza A specimens in England and Wales. Epidemiol. Infect, 136, 1658−1666.
[43] R. Sebastian, D. M. Skowronski, M. Chong, J. Dhaliwal & J. S. Brownstein. (2008) Age−related trends in the timeliness and prediction of medical visits, hospitalizations and deaths due to pneumonia and influenza, British Columbia, Canada, 1998−2004. Vaccine, 4, 1397−1403.
[44] P. F. Craigmile, N. Kim, S. A. Fernandez & B. K. Bonsu. (2007) Modeling and detection of respiratory-related outbreak signatures. BMC Med. Inform. Decis. Mak, 7: 28.
[45] P. M. Polgreen, F. D. Nelson & G. R. Neumann. (2007) Use of prediction markets to forecast infectious disease activity. Clin. Infect. Dis, 44: 272−279.
[46] R. G. Webster. (1997) Predictions for future human influenza pandemics. J. Infect. Dis, 176 (Suppl 1): S14−S19.A. Suhrbier, C. Schmidt & A. Fernan. (1993) Prediction of an HLA B8−restricted influenza epitope by motif. Immunology, 79, 171−173.
[47] A. Suhrbier, C. Schmidt & A. Fernan. Prediction of an HLA B8-restricted influenza epitope by motif. *Immunology.* 1993, 79: 171-173.
[48] P. Somvanshi, V Singh & P. K. Seth. (2008) Prediction of epitopes in hemagglutinin and neuraminidase proteins of influenza A virus H5N1 strain: a clue for diagnostic and vaccine development. OMICS, 12, 61−69.
[49] P. Gogolák, A. Simon, A. Horváth, B. Réthi, I. Simon, K. Berkics, E. Rajnavölgyi & G.K. Tóth. (2000) Mapping of a protective helper T cell epitope of human influenza A virus hemagglutinin. Biochem. Biophys. Res. Commun, 270: 190−198.
[50] M. Young, K. Kirshenbaum, K. A. Dill & S. Highsmith. (1999) Predicting conformational switches in proteins. Protein Sci, 8, 1752−1764.

J. Biomedical Science and Engineering, 2009, 2, 123-127

Bioinformatics analysis and characteristics of envelop glycoprotein E epitopes of dengue virus

Hua Zhong[1], Wei Zhao[1], Liang Peng[1], Shan-Feng Li[1], Hong Cao[1]

[1]Department of Microbiology, School of Public Health and Tropical Medicine, Southern Medical University, Guangzhou, Guangdong 510515, China. Correspondence should be addressed to Hong Cao (gzhcao@fimmu.com), Tel.:+86 20 61648723.

Received Sep. 4th, 2008; revised Dec. 22nd, 2008; accepted Jan. 8th, 2009

ABSTRACT

The major envelope glycoprotein E of dengue (DEN) virus plays a central role in the biology of flaviviruses. It is capable of inducing a protective immune response *in vivo* and responsible for the viral binding to the cellular receptor. The crystal structures of glycoprotein E ectodomains have already been determined. However, it is still unclear where the well-defined B-cell epitopes for glycoprotein E which induce the neutralizing antibodies locates. Thus, in order to characterize the role of glycoprotein E in the pathogenesis of dengue virus infection, we first used network servers (http://bio.dfci. harvard.edu/Tools/ & http://www. imtech. res. in) to predict and analyze the well defined B-cell and T-cell epitopes of the glycoprotein of the DEN-1 HAWAII strain. Then based on the highly conserved envelop glycoprotein amino acids, the hydrophilicity, antigenicity, accessibility and flexibility of envelop glycoprotein E were further predicted by using Biotic softwares (DNASTAR) and network servers (http://bio. dfci.harvard.edu/Tools/), the secondary structure was putatively obtained. In our study, the sequence at 281-295 amino acid (aa) for dengue virus type 1 HAWAII strain and the sequence at 345-359, 383-397 for dengue virus type 2 NGC strain were predicted as the more prevalent epitopes by using multiple parameters and different analysis softwares, respectively. Two epitopes of DEN-2 and one of DEN-1 locate on the domain Ⅲ and domain Ⅱ of the protein E, respectively. Subsequently, further studies will be carried out to examine the antigenicity and protection of the synthetic peptides with higher scores in the average antigen index (AI) and better hydrophilic properties determined by our data.

Keywords: Dengue Virus, Glycoprotein E, Epitope, Bioinformatics

1. INTRODUCTION

Dengue virus, a flavivirus belonging to the *flaviviridae* family, is a mosquito-borne human pathogen that causes dengue and dengue hemorrhagic fever which is currently one of the serious public health threats throughout the tropical and subtropical regions of the world [1]. Four serotypes of DEN virus have been identified (DEN-1, 2, 3 and 4), and each of these serotypes can infect humans and cause disease. This virus shares many characteristics with other flaviviruses, having a single-stranded RNA genome surrounded by an icosahedral scaffold and covered by a lipid envelope. The complete virion is 50 nm in diameter and contains an 11-kb plus-sensed RNA genome that is composed of seven nonstructural (NS) protein genes and three structural protein genes, core (C, 100 amino acids), membrane (M, 75 amino acids), and envelope (E, 495 amino acids) [2,3]. The order of proteins encoded is 5'-CprM(M)-E-NS1-NS2A-NS2B-NS3-NS4A-NS4B-NS5-3'[4].The 495-amino-acid (aa) envelop (E) glycoprotein, one of the three structural proteins, is the principal component of the external surface of the virion [5], and it is responsible for a wide range of biological activities, including binding to host cell receptors, fusion to and entry into host cells, therefore, this protein directly affects host range, cellular tropism, and, in part, the virulence of the virus [2,5]. Furthermore, the E protein also stimulates host immunity by inducing protective and neutralizing antibodies [6]. It is a main target and important antigen for vaccine development, and many attempts have been made to elucidate the structure-function relationships of the dengue virus glycoprotein E. The crystal structures of protein E ectodomains have already been determined. However, the location of well-defined B-cell and T-cell epitopes for glycoprotein E is largely unknown. Mapping of the B-cell and T-cell epitopes should be important for immunoinfectomic studies of dengue virus infection. Random peptide display has been applied in antigenic epitope determination. However, a combination of computational methods (e.g., bioinformatics) and experimental approaches of conventional biology should be a holistic way to determine the rigorous B-cell and T-cell epitopes. Thus, in order to characterize the role of glycoprotein E in the pathogenesis of dengue virus infection, we used bioinformatics and molecular approaches to predict and analyze its B-cell and T-cell antigen epitopes. Parameters

such as hydrophilicity, flexibility, accessibility, turns, exposed surface, polarity and antegenic propensity of polypeptide chains have been correlated with the location of continuous epitopes in a few well-characterized proteins. Net servers and the software DNA star are applied in our study.

2. MATERIALS AND METHODS

2.1. Antigenic Peptide Prediction

The online web server (http://bio.dfci. harvard. edu/Tools/) give us a pathway to predict sequences of peptides within a protein that are likely to be antigenic by eliciting an antibody response. Antigenic peptides are determined using the method of Kolaskar and Tongaonkar [7]. Predictions are based on a table that reflects the occurrence of amino acid residues in experimentally known segmental epitopes. We enter the amino acids of dengue virus type 1 HAWAII strain as well as dengue virus type 2 NGC (New Guinea C) strain, both of whom are standard strains, then operate the applet.

2.2. Hydrophilicity Estimation

The website locating at (http://us.expasy.org/tools/protscale.html) can give us a hydrophilicity prediction of the envelop glycoprotein E, it is based on the method of Hopp & woods.

2.3. Secondary Structure Presumption

Logging in the same web server mentioned in the third step, the β-turn and and coil of the E protein can be obtained, using the algorithm [8,9] of Levvit as well as Deleage & Roux [10,11].

2.4. Surface Accessibility and Average Flexibility Assumption

The Protean procedure of The DNASTAR software can supply us with the E protein's Surface accessibility and flexibility using methods of Emini [12] and Karplus-Schulz [13].

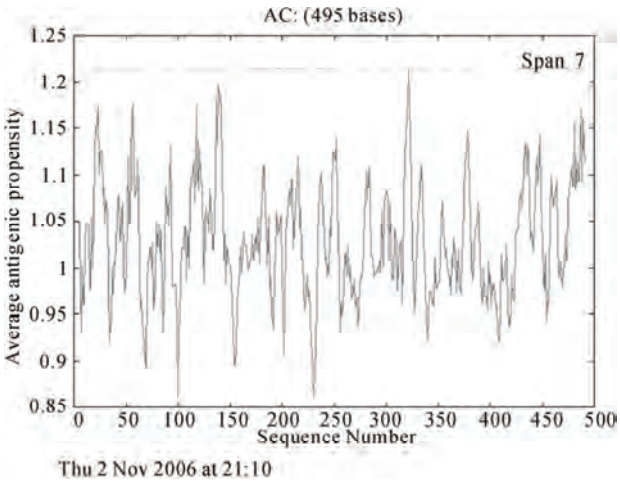
2.5. Tertiary Structure Prediction

After the process of secondary structure prediction, we can use the method of JPRED to link in the PDB ID, since the sequence of the protein E has 39 hits to the PDB database with E values of less than 0.0001. Thus, we make an entry in the PDB in the ID "IP58" to have a look at the structure by using the viewer Software named RASWIN32b2a.

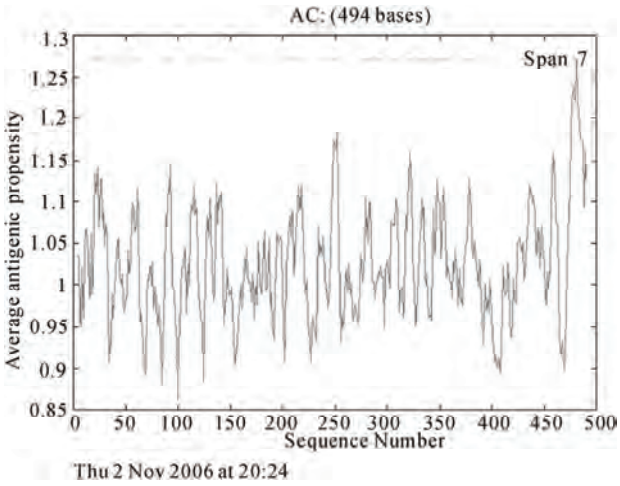
3. RESULTS

3.1. Antigenic Peptide Prediction of the Glycoprotein E

The prediction results for antigenic peptides of the glycoprotein E for DEN-1 and DEN-2 are shown in **Figure 1**.



(a) DEN-1 HAWAII strain



(b) DEN-2 NGC strain

Figure 1. Peptides predicted as antigenic epitope sites of the protein E.

Table 1. Peptides predicted as B-cell epitope sites of the protein E.

No	Start position	Sequence	End position
1	17	GATWVDVVLEHGSCVT	32
2	39	PTLDIELLKT	48
3	51	TNPAVLRKLCIE	62
4	87	DANFVCRR	94
5	109	GKGSLITCAKFKCVTK	124
6	126	EGKIVQYENLKYSVIVTVHT	145
7	169	PTSEIQLTDYGALTLDCSP	187
8	193	FNEMVLL	199
9	204	KSWLVHKQWFLDLPLPW	220
10	234	EDLLVTFKT	242
11	246	KKQEVAVLG	254
12	278	IFAGHLKCRL	287
13	296	GMSYVMCTG	304
14	316	QHGTVLVQVK	325
15	329	TDAPCKIPF	337
16	352	ITANPIVT	359
17	373	FGESYIVVGA	382
18	424	SIGGVFTSVGKLVHQIFGTAYGVLFSG	450
19	458	GIGILLTW	465
20	472	SASLSMTCIAVGMVTLYLGV	491

The B-cell (**Table 1**) and T-cell epitopes (**Table 2**) of the glycoprotein of DEN-1 HAWAII standard strain were predicted by the means of the position of amino acids with the online server respectively.

3.2. Hydrophilicity Prediction of the Glycoprotein E

The prediction results for hydrophilicity of the protein E of DEN-1 and DEN-2 are diagramed in **Figure 2**.

Table 2. Peptides predicted as T-cell epitope sites of the protein E.

HLA Sites	Peptides Position		
HLA-A2	206-215	117-126	483-492
HLA-A11	299-308	238-247	50-59
HLA-A24	298-307	439-446	325-334
HLA-B51	216-225	420-446	206-215
HLA-B60	48-57	313-322	256-265
HLA-B62	414-423	291-300	124-133

3.3. Secondary Structure Prediction

The prediction of protein E's secondary structure is shown in **Figure 3**.

3.4. Surface Probability of the Glycoprotein E

The surface probability assumption of the protein E of DEN-1 and DEN-2 strains is diagramed in **Figure 4**.

3.5. Flexibility Presumption of the Glycoprotein E

The Flexibility presumption of the glycoprotein E of DEN-1 and DEN-2 strains is shown in **Figure 5**.

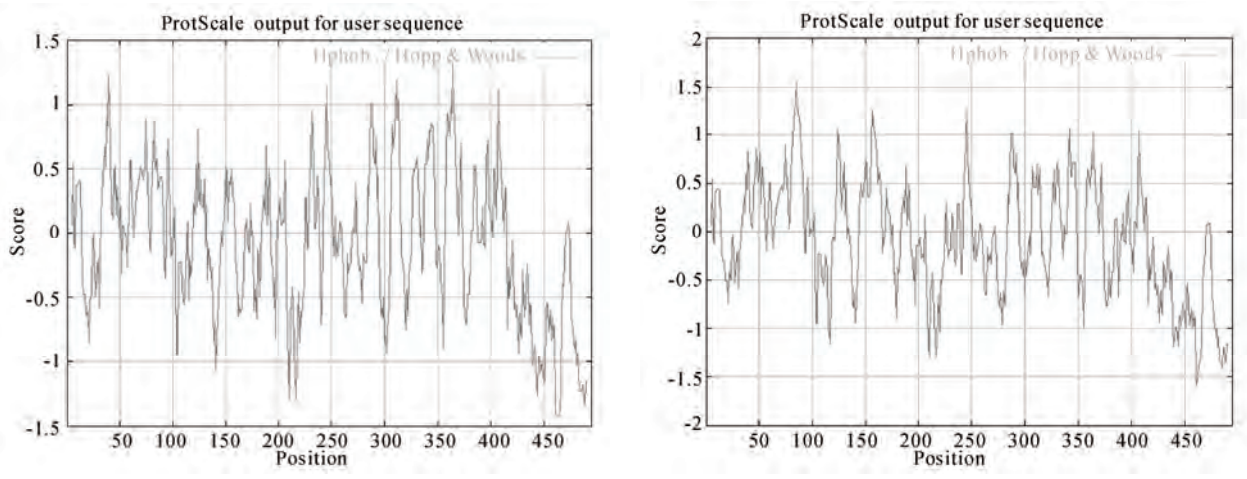
3.6. Putative Tertiary Structure of the Glycoprotein E

The putative tertiary structure of the glycoprotein E is shown in **Figure 6** (protein data bank). The ID is IP58.

4. DISCUSSION

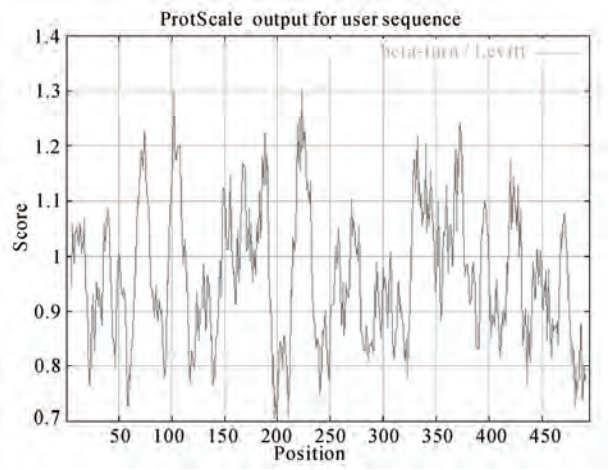
It is well known that although urgently needed for dengue virus infection, specific drugs for treatment and effective vaccination for prevention are currently unavailable. This is the main reason why we have focused on the so-called Antibody-dependent enhancement (ADE). Studies suggest that during a secondary infection with a different serotype, the presence of cross-reactive, non-neutralizing antibodies enhances the efficiency with which dengue virus infects susceptible cells. A molecular understanding of the events that lead to antibody neutralization, enhancement, or escape will be critical to the improvement of vaccines It is therefore important to determine which surface features on the dengue virion are responsible for inducing protective or enhancing immune response in the different serotypes. Thus, the structural and functional organizations of the dengue virus proteins are of central interests for the understanding of the biology of dengue virus and the mechanisms of virus-cell interactions.

The dengue E ectodomain consists of structurally distinct domains: Ⅰ, Ⅱ and Ⅲ [14]. The domain Ⅲ appears to play an important part in host cell receptor binding for viral entry and in inducing protective immunity. The rigorous B-cell and T-cell epitopes were not identified yet. In our study, we focused on the characterizing the B-cell and T-cell epitopes of dengue virus envelop E glycoprotein by deploying the bioinformatics approaches, the sequence at 281-295 amino acid (aa) for dengue virus type 1 HAWAII strain and the sequence at 345-359, 383-397 for dengue virus type 2 NGC strain were predicted as the more prevalent epitopes by using multiple parameters and different analysis softwares, respectively. The sequences selected not only have higher scores in the average antigen index (AI), which could predict the antigen epitope of envelop glycoprotein E, but also showed better hydrophilic properties. Two epitopes of DEN-2 and one of DEN-1 locate on

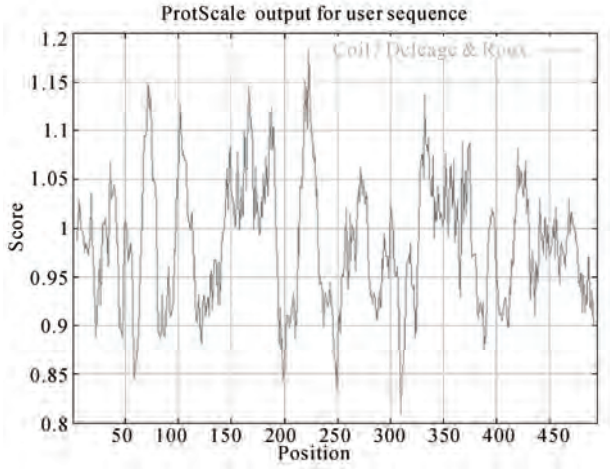


(a) DEN-1 HAWAII strain **(b) DEN-2 NGC strain**

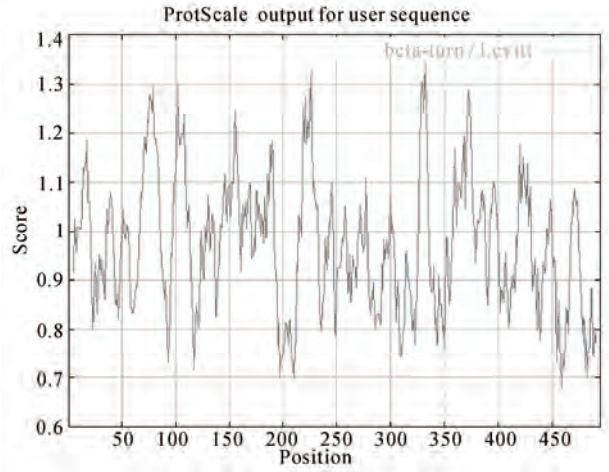
Figure 2. The hydrophilicity of the protein E.



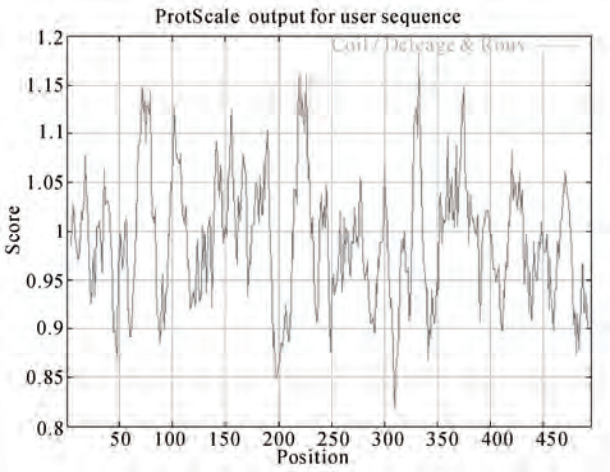
(a1) β-turn the coiled region of DEN-1 HAWAII strain



(a2) β-turn the coiled region of DEN-1 HAWAII strain

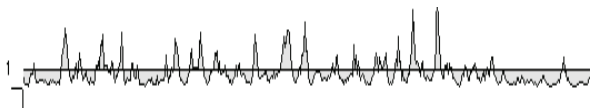


(b1) β-turn the coiled region of DEN-2 NGC strain

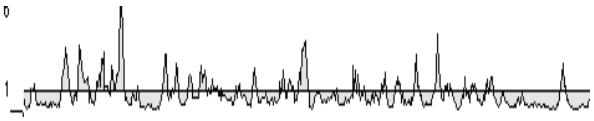


(b2) β-turn the coiled region of DEN-2 NGC strain

Figure 3. The prediction of the secondary structure of the protein E.



(a) DEN-1 HAWAII strain



(b) DEN-2 NGC strain

Figure 4. Surface probability of the glycoprotein E.

(a) DEN-1HAWAII strain

(b) DEN-2 NGC strain

Figure 5. The flexibility presumption of the glycoprotein E.

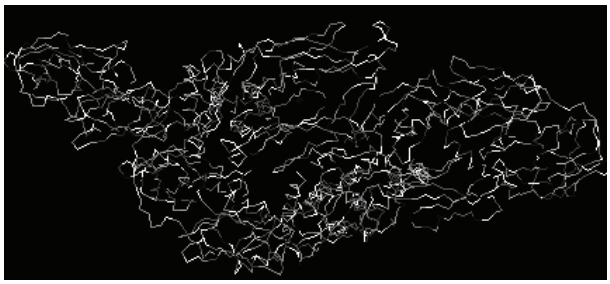


Figure 6. Putative tertiary structure of the glycoprotein E.

the domain III and domain Ⅱ of the protein E, respectively. The domain III has been hypothesized to contain multiple type- and subtype-specific epitopes eliciting only virus-neutralizing monoclonal antibodies while the domain Ⅱ is involved in virus-mediated membrane fusion, and contains many cross-reactive epitopes eliciting neutralizing and non-neutralizing monoclonal antibodies. The predicted epitopes can be used for the devel-

opment of vaccine and the dissection of the ADE effect. The further experimental studies will be performed to determine the immunogenicity and protection effect of peptides with higher scores in the average antigen index (AI) and better hydrophilic properties, and to identify vaccine candidates.

ACKNOWLEDGEMENTS

This work was supported by Guangzhou Key Technology R&D Program (No.2008Z1-E401 to H. Cao).

REFERENCES

[1] T. Monath.(1994) Dengue: the risk to developed and developing countries. Proc Natl Acad Sci USA 91, 2395−2400.

[2] R. J. Kuhn, W. Zhang, M. G. Rossmann. (2002) Structure of dengue virus: implications for flavivirus organization, matu--ration, and fusion. Cell 108, 717−725.

[3] E. A. Henchal, J. R. Putnak. (1990) The dengue viruses. Clin Microbiol Rev 3, 376−396.

[4] C. M. Rice, B. N. Fields, D. M. Knipe, P. M. Howley. (1996)*Flaviviridae*: the viruses and their replication. Virology 3, 931−959.

[5] J. T. Roehrig. Immunochemistry of the dengue viruses, 199−219 *In:* D. J. Gubler and G. Kuno (eds), Dengue and dengue hemorrhagic fever. CAB International, New York, N. Y. 1997.

[6] G. J. Chang, Molecular biology of dengue viruses, 175−198 *In*: D. J. Gubler, G. Kuno (eds.), Dengue and Dengue Hemorrhagic Fever. CAB International, London, 1997.

[7] A. S. Kolaskar, P. C. Tongaonkar. (1990) A semi−empirical method for prediction of antigenic determinants on protein antigens, FEBS Lett 276, 172−174.

[8] M. Levitt, J. Greer. (1977) Automatic identification of secondary structure in globular proteins. J Mol Biol 114, 181−239.

[9] J Cheng. (2008) A multi-template combination algorithm for protein comparative modeling. BMC Structural Biology 8, 18−36.

[10] G. Deleage, B. Roux. (1987) An algorithm for protein secondary structure prediction based on class prediction. Protein Eng 1, 289−294.

[11] J. Martin, J. F. Gibrat, F. Rodolphe. (2006) Analysis of an optimal hidden Markov model for secondary structure prediction. BMC Structural Biology 6, 25−45.

[12] E. A. Emini, J. V. Hughes, D. S. Peflow. (1985) Induction of hepa-titis A vires-neutralizing antibody by a vires-specific synthetic peptide. J Virol 55,836−839.

[13] P. A. Karplus, G. Schultz. (1985) Prediction of chain flexibility in proteins. Naturwissenschaften 72, 212−213.

[14] Y. Modis, S. Ogata, D. Clements. (2003) A ligand−binding pockets in the dengue virus envelop glycoprotein. Science 100, 6986−6991.

J. Biomedical Science and Engineering, 2009, 2, 128-134

Analysis and expression of the polyhedrin gene of Antheraea pernyi nucleopolyhedrovirus (AnpeNPV)

Jia-Xi Huang[1], Hui-Ling Wu[1], Yan Wu[1], Shan-Ying Zhu[1], Wen-Bing Wang[1]

Institute of Life Sciences, Jiangsu University, 301 Xuefu Road, Zhenjiang 212013, P. R. China

Received Dec. 3rd, 2008; revised Jan. 1st, 2009; accepted Jan. 5th, 2009

ABSTRACT

The polyhedrin (polh) gene is often used to analyse evolution of baculovirus. In this report, the polh of Antheraea pernyi nucleopolyhedrovirus (AnpeNPV) was cloned and sequenced. The Open reading frame (ORF) of the AnpeNPV consists of 738 nucleotides encoding 245 amino acids with molecular masses of 29 kDa. The deduced amino acids were significant homology with other baculoviruses, such as Attacus ricini NPV (ArNPV) and Autographa californica NPV (AcNPV). A strongly hydrophilic region was predicted at positions from 30 to 50 of the AnpeNPV Polh protein by bioinformatics analysis. Expression of the polh gene of AnpeNPV in E. coli was examined by SDS–PAGE, Western blot and Mass-spectrum analysis. The result showed that the bacterium expression system was suitable for the virus gene expression. It indicated that the products of the polh gene expressed in this system can be easier to use for raising antibodies.

Keywords: Antheraea Pernyi, Insect, Baculovirus, NPV, Polyhedrin, Prokaryotic Expression

1. INTRODUCTION

The Chinese oak silkworm Antheraea pernyi (Lepidoptera: Saturniidae) is an economically important insect primarily for the production of tussah silk. In recent years, consumption of the silkworm pupae as food delicacies has also gained tremendous popularity. The jaundice disease of the oak silkworm caused by the infection of A.pernyi nucleopolyhedrovirus (AnpeNPV) is a major threat to the tussah industry [1]. AnpeNPV is a member of the Baculoviridae with large, enveloped, double-stranded DNA. Baculoviridae are widely known to the scientific community in the form of commercial baculovirus expression vectors (BEVs) [2,3]. Baculoviruses also have an established application as insecticides against agricultural and forestry pests [4,5]. Currently, the Baculoviridae comprises two genera, Nucleopolyhedrovirus (NPV) and Granulovirus (GV) [1]. During the infection cycle, NPVs produce two structurally and functionally distinct virion phenotypes: occlusion- derived virus (ODV) and budded virus (BV) [6]. The occluded viruses of the NPV are referred to as polyhedra. Polyhedrin is the major protein component of the polyhedra [7]. The polh gene is not essential for viral development, and normally deletion of the polh gene is not interfering with viral replication in cultured cells. However, in per os infectivity, the polyhedra or occlusion bodies are required for the oral infection of insects [8]. Baculovirus entry into host cells involves that ODVs are released from the occlusion body by the alkaline environment within the midgut lumen of the larva and subsequently initiate primary infection of the mature columnar epithelial cells of the midgut [6].

In order to explore effective propagation and infectivity of the polyhedra, this paper analysised the nucleotide sequence and promoter (prmoter-Ap) of the polh gene of AnpeNPV by bioinformatics tools, and further prokaryotic expression for AnpeNPV polyhedrin (polh-Ap).

2. MATERIALS AND METHODS

2.1 Materials

The Wild-type AnpeNPV strain was maintained in our laboratory. Restriction Enzymes, T4 DNA ligase, PCR reagents pMD18-T and DNA purification kit were purchased from TaKaRa Company (China, Dalian); primers and other reagents were bought from Shanghai Sangon Bio-technology Corpotation. The vectors for expression, and *Escherichia coli* strain DH5α and BL21 were kept in our laboratory.

2.2. Aplification of the AnpeNPV polh Gene

AnpeNPV genomic DNA was isolated using the method described by previously [9,10] and about 15-20 ng DNA was used as template for standard PCR. The specific primers were designed based on the sequence of ORF (GenBank: EU195295). The polh-Ap forward primer (5′ CCG GAA TTC ATG CCA GAT TAC TCA TAC CGG 3′)containing an *Eco*R I restriction site (underlined), and the reverse primer (5′ CCC AAG CTT CTA GTA CGC GGG GCC AGT 3′) containing a *Hin*d III restriction site (underlined). The PCR conditions were 1 cycle at 94 °C for 5 min; 30 cycles at 94 °C for 45 s, 62°C for 45 s, and 72 °C for 1 min; and 1 cycle at 72 °C for 10 min. The PCR product was examined by electrophoresis in 1% agarose gel with the ethidium bromide staining.

2.3. Cloning and Construction of Expresion Plasmid

The PCR products were ligated into pMD18-T vector using T4 DNA ligase and then transformed into *E. coli* (DH5a), and sequenced, respectively.

The recombinant plasmid pMD-polh-Ap was digested with *Eco*R I and *Hin*d III, and was purified to ligate with the Pet28a vector digested with *Eco*R I and *Hin*d III, and transformed into *E. coli* (BL21).

2.4. Analysis of the *polh* Gene

The amino acid sequence was deduced with Expasy Translate tool (http://au. expasy. org/tools/dna. html) according to the AnpeNPV *polh* gene sequence. Align using DNAstar CLUSTAL W program. Phylogenetic tree was made by MEGA 3.1 software.

In order to explore regulatory sequence in the putative promoter region, NNPP (Promoter Predication by Neural Network http://www.fruitfly.org/seq_tools/promoter.html), promoter scan and transcription factor binding sites (http://www-bimas.cit.nih.gov/molbio/proscan/) were applied together to make a comprehensive prediction.

2.5. Expression of the *polh* Gene in *E.coli*

A positive clone was cultured in LB medium supplement with Kanamycin (final concentration of 50μg/ ml) ovenight at 37 ℃ with shaking, then the culture was added into 100mL fresh LB medium and cultured at 37℃ with shaking to A600 about 0.6. The culture was induced with IPTG (final concentration of 8 μg/ mL) and shaked at 30 ℃ for 10 hours. SDS polyacrylamide gel was used to analyze the expression in the Mini-Protein system (Bio-Rad, USA). After electrophoresis, the gel was stained with Coomassie Brilliant Blue R250 to visualize the protein bands. Protein samples were separated on SDS-10% polyacrylamide gels and transferred to PVDF membranes. Blots were soaked in TBST buffer (10 mmol/L Tris-HCl, pH 7.6, 0.15 mol/L NaCl, 0.1% Tween 20) with 5% nonfat dried milk. The antiserum against the His-Polh fusion protein (His antibody) at a dilution of 1:2,000 monoclonal antibody was added as the first antibody, followed by addition of 1:5,000 dilution horseradish peroxidase-conjugated goat anti-mouse immunoglobulin G as the secondary antibody. Blots were visualized with the Enhanced chemiluminescence Western blot kit (Amersham). The predicted Polh protein band was cut out for Mass-spectrum analysis.

3. RESULTS

3.1. Nucleotide and Amino Acid Sequence Analysis

The ORF of cloned gene has two different nucleotides from the published sequence (DQ486030), but no amino acid residues were changed. The 738 nucleotides (including the stop codon TAG) encoded a putative peptide of 235 amino acids by an Expasy Translate tool.

```
ATGCCAGATTACTCATACCGGCCGACCATTGGTCGCACCTATGTGTACGACAACAAGTAT
M  P  D  Y  S  Y  R  P  T  I  G  R  T  Y  V  Y  D  N  K  Y
TACAAAAACTTAGGGTCCGTCATTAAAAACGCCAAGCGCAAGAAGCATTTAGTCGAACAT
Y  K  N  L  G  S  V  I  K  N  A  K  R  K  K  H  L  V  E  H
GAAGAGGAAGAAAAGCATTGGGATCCTTTAGACAATTACATGGTCGCGGAAGACCCTTTC
E  E  E  E  K  H  W  D  P  L  D  N  Y  M  V  A  E  D  P  F
CTGGGGCCGGGTAAAAACCAAAAACTGACACTTTTCAAGGAAATCCGCAACGTTAAACCC
L  G  P  G  K  N  Q  K  L  T  L  F  K  E  I  R  N  V  K  P
GACACAATGAAACTTATTGTCAACTGGAGCGGTAAAGAATTTCTGCGCGAAACTTGGACC
D  T  M  K  L  I  V  N  W  S  G  K  E  F  L  R  E  T  W  T
CGTTTTGTTGAGGATAGCTTTCCGATTGTAAACGACCAAGAGGTCATGGATGTGTTCCTC
R  F  V  E  D  S  F  P  I  V  N  D  Q  E  V  M  D  V  F  L
GTCATTAACCTGCGCCCCACGCGCCCCAACAGGTGCTACAAGTTCCTGGCGCAGCACGCG
V  I  N  L  R  P  T  R  P  N  R  C  Y  K  F  L  A  Q  H  A
CTCAGATGGGACTGCGACTACGTGCCGCACGAGGTAATCCGCATTGTGGAGCCATCCTAC
L  R  W  D  C  D  Y  V  P  H  E  V  I  R  I  V  E  P  S  Y
GTGGGCATGAACAACGAGTACAGAATTAGCCTCGCCAAGAAAGGCGGCGGCTGCCCCATC
V  G  M  N  N  E  Y  R  I  S  L  A  K  K  G  G  G  C  P  I
ATGAACATTCACAGCGAGTACACCAACTCGTTTGAATCGTTTGTAAACCGCGTAATCTGG
M  N  I  H  S  E  Y  T  N  S  F  E  S  F  V  N  R  V  I  W
GAGAACTTTTACAAGCCCATTGTGTACATTGGCACGGACTCGGGTGAGGAGGAGGAAATT
E  N  F  Y  K  P  I  V  Y  I  G  T  D  S  G  E  E  E  E  I
CTCATCGAGGTTTCGCTTGTGTTCAAGGTCAAGGAGTTTGCGCCCGACGCGCCACTGTTT
L  I  E  V  S  L  V  F  K  V  K  E  F  A  P  D  A  P  L  F
ACTGGCCCCGCGTACTAG
T  G  P  A  Y
```

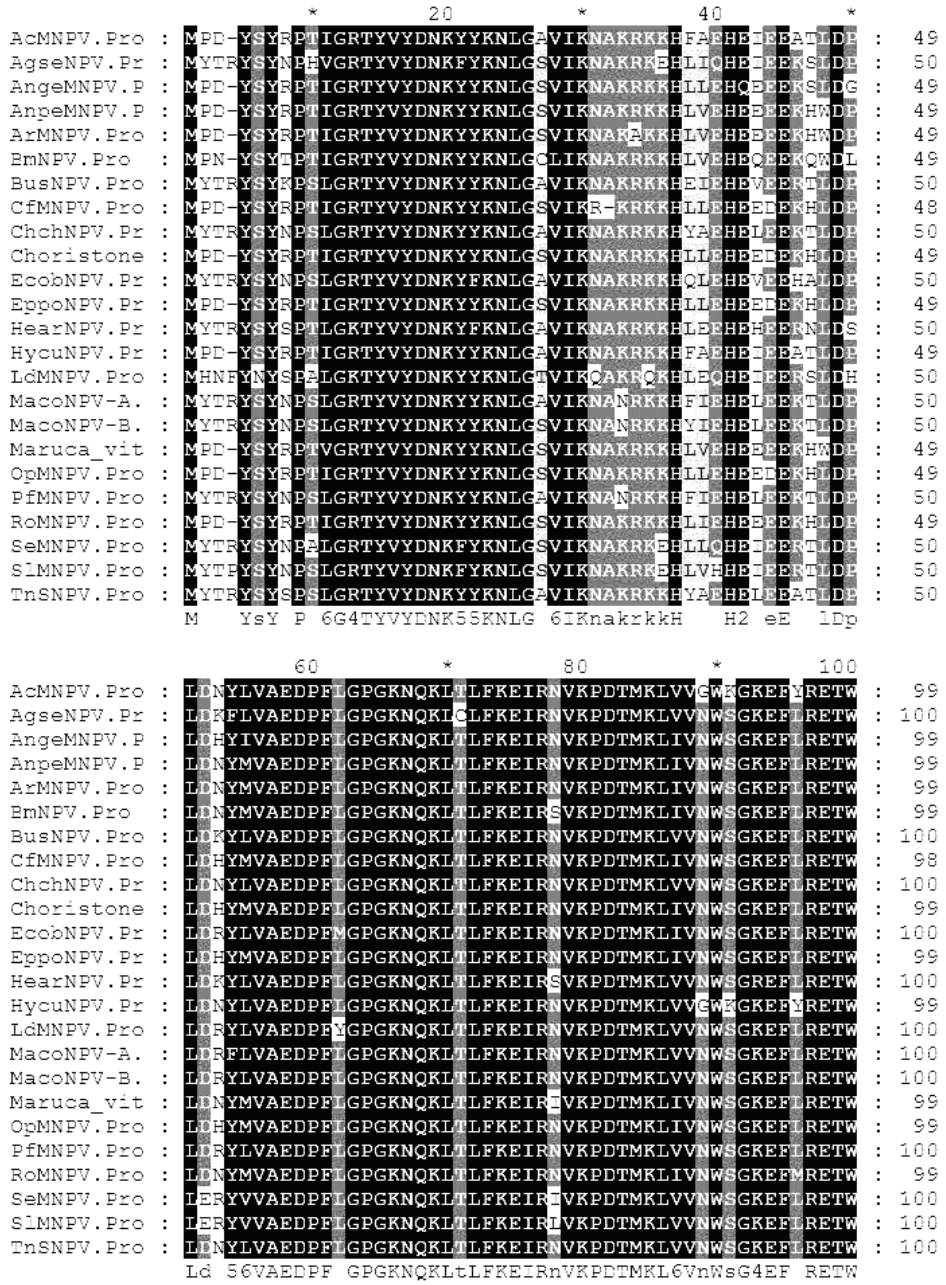
Figure 1. Nucleotide sequence and deduced amino acid sequence of the polyhedrin gene. The predicted amino acid is represented by the one letter code designation below the nucleotide sequence. The initiate and the stop codes are framed.

The nucleotide sequence of Polh-Ap and its deduced amino acid sequence are shown in **Figure 1**. This deduced polypeptide contains 16 strongly basic, 16 strongly acidic, 113 hydrophobic and 58 hydrophilic amino acids with the calculated molecular mass of 29 kDa, and the isoelectric point was of 6.1.

3.2. Protein and Homology Analysis

Using BLAST software of NCBI to search for homology in the GenBank database, the deduced amino acid sequence showed an identity of 97%, 98%, 98%, 97%, 93% and 89% to the corresponding genes of *Attacus ricini* NPV(ArNPV, AAP16625), *Epiphyas postvittana* NPV (EppoNPV, NP_203170), *Maruca vitrata* MNPV (YP_950731), *Rachiplusia ou* MNPV (RoMNPV, NP_702998) [11], *Bombyx mori* NPV (BmNPV, AAA 46734) [12] and *Autographa californica* NPV (AcNPV, NP_054037) [13], respectively. Comparison of the deduced amino acid sequence with that of the corresponding genes of many species is shown in **Figure 2**. This protein was demonstrated to be highly conserved in baculoviruses.

The predict of secondary structure for polh-Ap by CLC Protein Workbench 3.0.3.(**Figure 4**). There are 4 regions



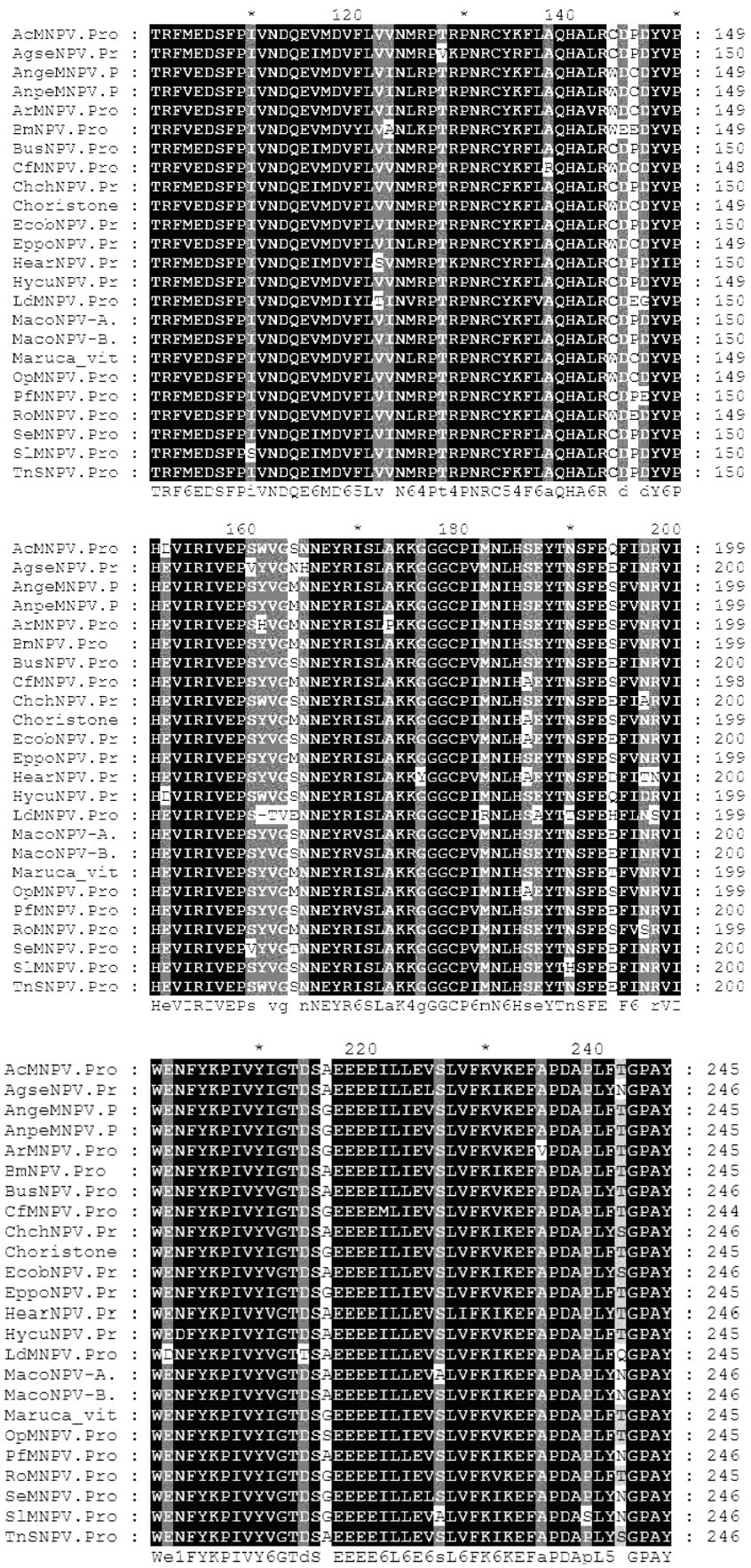


Figure 2. Alignment of the polyhedrin genes of baculoviruses, The sequences were aligned using DNAstar CLUSTAL W program.

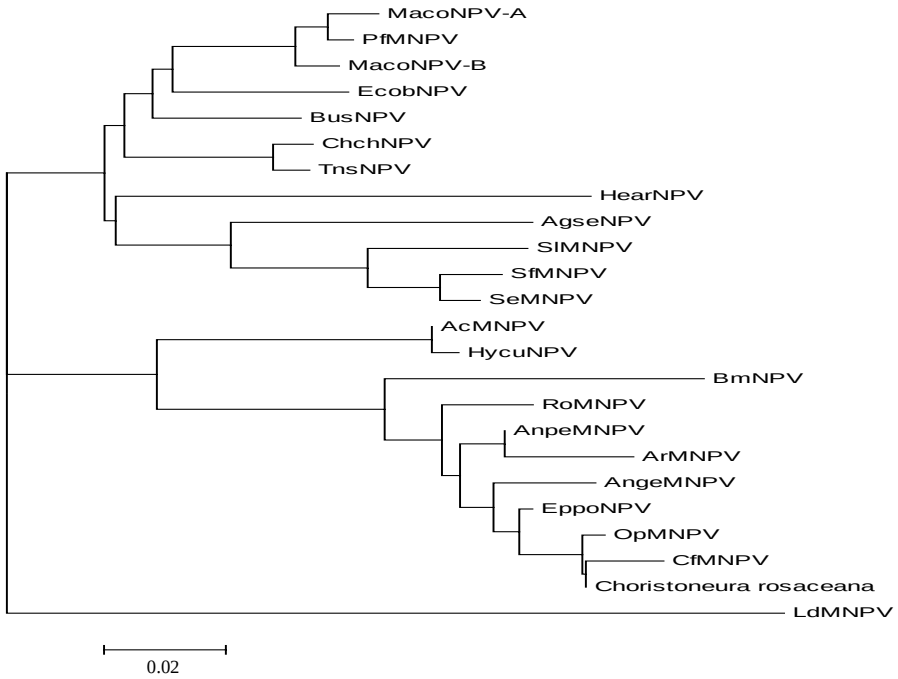


Figure 3. Phylogeny of the polyhedrin protein. Phylogenetic tree of polyhedrin gene was constructed by MEGA version 3.1 from CLUSTAL W alignments. The neighbor-joining method was used to construct the tree. From the phylogenetic tree, the polh gene of AnpeNPV was closest to that of ArNPV.

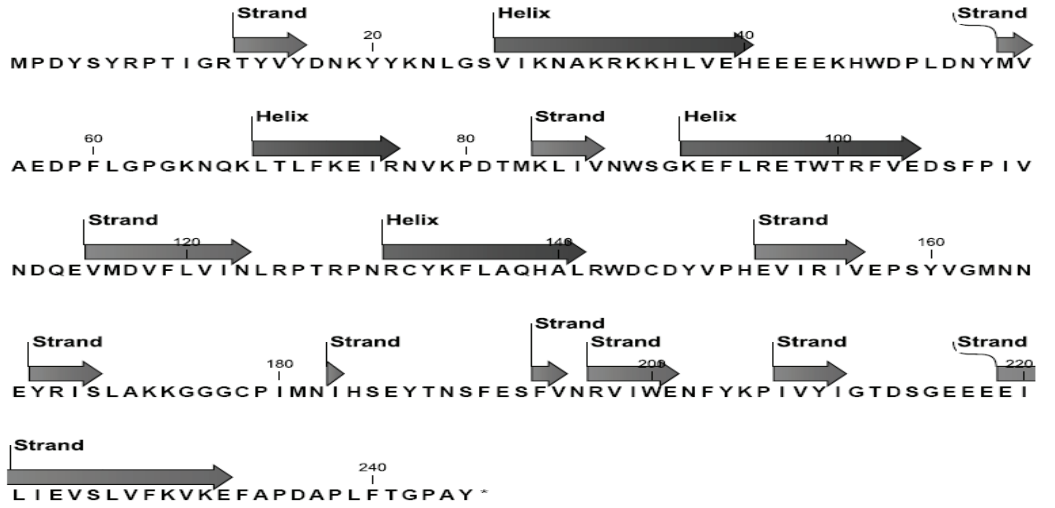


Figure 4. The secondary structure of the Polh protein of AnpeNPV. It contains 4 regions of alpha helix and 11 pieces of ß-sheet.

of helical and 11 pieces of ß-sheet in the sequence.

3.3. Construction of Expression Plasmid

The fragment of polh-Ap was sequenced to be sure containing a correct ORF, and was inserted into the expression pET28a vector and then was expressed in *E col* (BL21) The recombinant plasmid was identified by digestion with *Eco*R I and *Hin*d III. The result of electrophoresis indicated the recombinant plasmid was successfully constructed (**Figure 6**).

3.4. Expression of the AnpeNPV *polh* Gene in *E. coli*

The *E. coli* BL21 transformed with the pET28a/polh-Ap plasmid to express the His-6PGL fusion protein of about 34 kDa, which was consistent with the expected mo lecular mass of the fusion protein of pET28a/polh-Ap (**Figure 7**). The result showed that the AnpeNPV *polh* gene was highly expressed in *E. coli*. The expression products can be used as antigen to raise the antibody of the Polh protein.

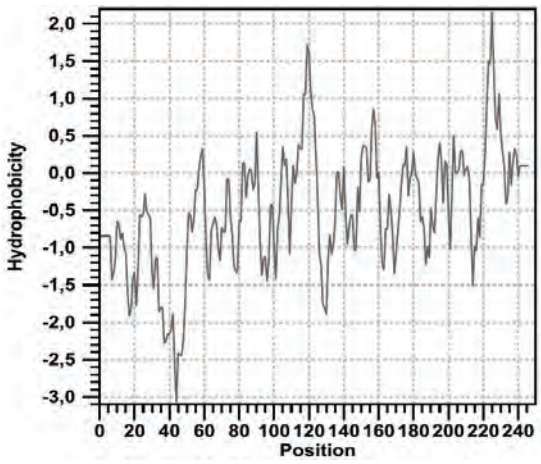


Figure 5. The hydrophobicity profile of AnpeNPV Polh protein. The X-axis contains 245 increments, each representing an amino acid in the sequence of AnpeNPV polyhedrin. The Y-axis represents the range of hydrophilicity values (from 2.2 to -3.1) with employ of Kyte-Doolittle scale. One region of strongly hydrophilicity exists at positions from 30 to 50 of the AnpeNPV polyhedrin protein.

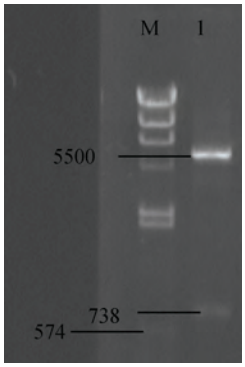


Figure 6. Identification of the recombinant plasmid pET28a/polh-Ap by electrophoresis in agarose gel 1, pET28a/polh-Ap; M, DNA molecular mass marker.

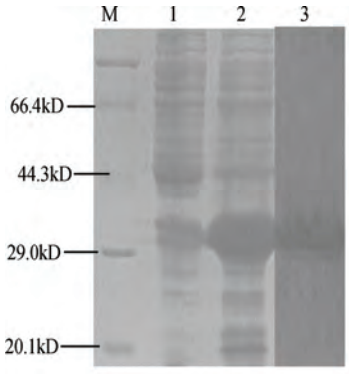


Figure 7. The expression products of AnpeNPV polh gene were analyzed by SDS-PAGE and Western-blot 1, Protein of E. coli BL21 contained pET28a induced by IPTG; 2, Protein of E. coli BL21 contained pET28a /polh-Ap induced by IPTG; 3, Western-blot of the fusion protein; M, Protein marker.

4. DISCUSSION

In this report the AnpeNPV *poly* gene was cloned and compared with other baculoviruses. Polyhedrin genes

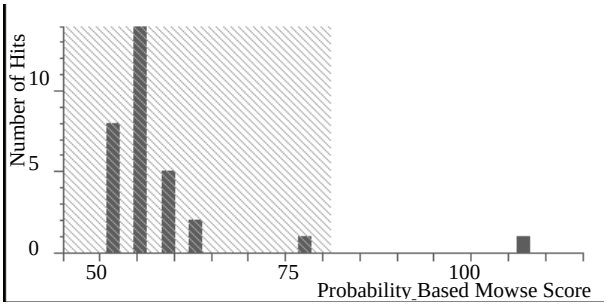


Figure 8. Mass-spectrum with Mascot analysis (Mass: 29003; Score: 107; Expect: 0.00013) Protein score is -10*Log (P), where P is the probability that the observed match is a random event. Protein scores greater than 81 are significant (p<0.05).

are highly conserved among many baculoviruses. The AnpeNPV polyhedrin gene was closest to that of ArNPV from the Phylogeny tree (**Figure 3**), differing by only five amino acids (**Figure 2**). DNA sequence comparison polyhedra containing low numbers of virions [16]. of AnpeNPV and ArNPV polyhedrins showed that a difference in identity to 97%, of which only thirteen dif ferences (**Figure 2**). The result suggests that the *poly* genes of AnpeNPV and ArNPV have evolved from a common ancestor distinct from the other NPVs. The AnpeNPV *polh* gene is very closely related to NPV group I than that of group II (**Figure 3**). Availability of polyhedrin protein sequences of other baculoviruses may aid in their classification and may help define baculovirus species [14].

Alignment results showed that the variability regions occur at the beginning of N-terminus (position 2 to 4) and the domain from position 31 to 52 (**Figure 2**). Even in this region, some positions are conserved, such as, H37, H41, E45 and D49. In contrast, the C-terminus (from 198 to 245) is highly conserved. The cysteine positons (at 133 and 179) and the prolines (at 60，64，81，109，127，130，150，159，180，207，236，and 244) of AnpeNPV polyhedrin appear to be very important (**Figure 2**). Cysteines often form disulphide bonds critical for protein structure; proline breaks helical and ß-sheet regions and is often associated with turns in the secondary structure of proteins [15]. Therefore, both these amino acids could be crucial in determining the conformation of these proteins. These conserved regions may be necessary to give the proteins their characteristic common properties: namely crystal formation and alkali solubility. Indeed, a mutant of AcNPV with a single pro tein changed to Leu at position 62 resulted in cubic polyhedra containing low numbers of virions [16].

Hydrophilic regions are exposed on surface of the protein and are highly polar. They have a tendency to be antigenic sites [17]. There is one region of strongly hydrophilicity at positions from 30 to 50 in the AnpeNPV polyhedrin (**Figure 5**). Comparison of the baculovirus polyhedrin sequences indicates that although they vary in amino acid sequence in this region, their basic pattern of hydrophilicity is preserved (date no shown). Therefore, much of the variation in amino acid sequence is

neutral and does not alter the overall nature of the proteins. This region therefore presents a potential antigenic site which may be useful for production of antibodies capable of differentiating or identifying different baculoviruses [18]. Ultimately predicted antigenic determinants from proteins of pathogenic organisms might also be useful in the production of synthetic vaccines [17].

The *polh* gene of baculovirus is a very late gene which expressed in late stage of virus infection. It is not an essential gene in virion development and could be deleted for foreign gene expression [19,20]. Some evidences showed that the level of the foreign gene expression was related to genetic codes of the gene. To test the *polh* gene expression in another system, we constructed a bacterium expression system to express the Polh protein. The result indicates that this gene is suitable for *E. coli* expression system. It might be helpful to produce the virus proteins to raise antibodies.

ACKNOWLEDGEMENTS

This work was supported by the 973 National Basic Research Program of China (2005CB121005); The Six-Field Top programs of Jiangsu Provice; National Natural Science Foundation of Jiangsu Education Communitte(06KJD180043); Innovation Foundation for Graduate Students of Jiangsu Province.

REFERENCES

[1] Q. Fan, S. Li, L. Wang, B. Zhang, B. Ye, Z. Zhao, Cui, L. (2007). The genome sequence of the multinucleocapsid nucleopolyhedrovirus of the Chinese oak silkworm Antheraea pernyi. Virology 366(2), 304−315.

[2] O.A. Lihoradova, I. D. Ogay, A. A. Abdukarimov, S. S. Azimova, D. E. Lynn, Slack, J. M. (2007). The Homingbac baculovirus cloning system: An alternative way to introduce foreign DNA into baculovirus genomes. J Virol Methods 140 (1-2), 59−65.

[3] Z. M. Nie, Z. F. Zhang, D. Wang, P. A. He, C. Y. Jiang, L. Song, F. Chen, J. Xu, L. Yang, L. L. Yu, J.Chen, Z. B. Lv, J. J. Lu, X. F. Wu, Zhang Y. Z. (2007) Complete sequence and organization of Antheraea pernyi nucleopolyhedrovirus, a dr-rich baculovirus. BMC Genomics 8, 248−261.

[4] S. P. Cook, R. E. Webb, J. D. Podgwaite, Reardon, R. C. (2003) Increased mortality of gypsy moth Lymantria dispar (L.) (Lepidoptera: Lymantriidae) exposed to gypsy moth nuclear polyhedrosis virus in combination with the phenolic gycoside salicin. J Econ Entomol 96(6), 1662−1667.

[5] Moscardi, F. (1999) Assessment of the application of baculoviruses for control of Lepidoptera. Annu Rev Entomol 44, 257−289.

[6] X. Dai, T. M. Stewart, J. A. Pathakamuri, Q. Li, Theilmann, D. A. (2004) Autographa californica multiple nucleopolyhedrovirus exon0 (orf141), which encodes a RING finger protein, is required for efficient production of budded virus. J Virol 78(18), 9633−9644.

[7] S. G. Kamita, S. Maeda, Hammock, B. D. (2003) High-frequency homologous recombination between baculoviruses involves DNA replication. J Virol 77(24), 13053−13061.

[8] A. M. Khurad, A. Mahulikar, M. K. Rathod, M. M. Rai, S. Kanginakudru, Nagaraju J. (2004) Vertical transmission of nucleopolyhedrovirus in the silkworm, Bombyx mori L. Journal of Invertebrate Pathology 87, 8–15.

[9] S. Gomi, C. E. Zhou, W. Y. Yih, K. Majima, Maeda S. (1997) Deletion analysis of four of eighteen late gene expression factor gene homologues of the baculovirus, BmNPV. Virology 230, 35−47.

[10] W. B. Wang, S. Y. Zhu, L. Q. Wang, F. Yu, Shen W. D. (2005) Cloning and sequence analysis of the Antheraea pernyi nucleopolyhedrovirus gp64 gene. J Biosci 30, 605−610.

[11] L. H. Robert, Bonning B.C. (2003) Comparative analysis of the genomes of Rachiplusia ou and Autographa californica multiple nucleopolyhedroviruses Journal of General Virology 84, 1827–1842.

[12] S. Gomi, K. Majima, Maeda S. (1999) Sequence analysis of the genome of Bombyx mori nucleopolyhedrovirus. J Gen. Virol. 80, 1323−1337.

[13] M. D. Ayres, S. C. Howard, J. Kuzio, M. Lopez-Ferber, Possee R.D. (1994) The complete DNA sequence of Autographa californica nuclear polyhedrosis virus. Virology 202, 586−605.

[14] E. A. van Strien, D. Zuidema, R.W. Goldbach, Vlak, J. M. (1992) Nucleotide sequence and transcriptional analysis of the polyhedrin gene of Spodoptera exigua nuclear polyhedrosis virus. J Gen Virol 73 (Pt 11), 2813−2821.

[15] P. Y. Chou, Fasman, G. D. (1977) Beta-turns in proteins. J Mol Biol 115(2), 135−175.

[16] E. B. Carstens, A. Krebs, Gallerneault, C. E. (1986) Identification of an amino acid essential to the normal assembly of Autographa californica nuclear polyhedrosis virus polyhedra. J Virol 58(2), 684−688.

[17] T. P. Hopp, Woods, K. R. (1981). Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci U S A 78(6), 3824-3828.

[18] Rohrmann, G. F. (1986) Polyhedrin structure. J Gen Virol 67(8), 1499−1513.

[19] R. D. Possee, S. C. Howard, (1987) Analysis of the polyhedrin gene promoter of the Autographa californica nuclear polyhedrosis virus. Nucleic Acids Res 15(24), 10233-10248.

[20] G. E. Smith, M. J. Fraser, Summers, M. D. (1983) Molecular engineering of the Autographa californica nuclear polyhedrosis virus genome:deletion mutations within the polyhedrin gene. J Virol 46, 584−593.

Call for Papers

Journal of Biomedical Science and Engineering (JBiSE)

www. scirp. org/journal/jbise

JBiSE, an international journal, publishes research and review articles in all important aspects of biology, medicine, engineering, and their intersection. Both experimental and theoretical papers are acceptable provided they report important findings, novel insights, or useful techniques in these areas. All manuscripts must be prepared in English, and are subject to a rigorous and fair peer-review process. Accepted papers will immediately appear online followed by printed in hard copy.

Subject Coverage

Bioelectrical and neural engineering

Bioinformatics

Medical applications of computer modeling

Biomedical modeling

Biomedical image processing & visualization

Real-time health monitoring systems

Biomechanics and bio-transport

Patten recognition and medical diagnosis

Biomedical effects of electromagnetic radiation

Safety of wireless communication devices

Biomedical devices, sensors, and nano technologies

NMR/CT/ECG technologies and EM field simulation

Physiological signal processing

Medical data mining

Other related topics

Editor–in–Chief

Editorial Board

Notes for Intending Authors

Submitted papers should not have been previously published nor be currently under consideration for publication elsewhere. Paper submission will be handled electronically through the website. All papers are refereed through a peer review process. For more details about the submissions, please access the website.

Website and E-Mail

www. scirp. org/journal/jbise

Email: jbise@scirp.org

ISSN 1937-6871 (Print), 1937-688X (Online)

Journal of Biomedical Science and Engineering, 2009, 2.
Published Online April 2009 in SciRes (http://www.scirp.org/journal/jbise).

TABLE OF CONTENTS

Volume 2, Number 2, April 2009

News and Announcement
JBiSE Editorial Office..... 77

Systems Biology: The take, input, vision, concerns and hopes
G. Tucker..... 78

A novel approach in ECG beat recognition using adaptive neural fuzzy filter
G. N. Golpayegani, A. H. Jafari..... 80

Effects of lead exposure on alpha-synuclein and p53 transcription
P. J. Zuo, A. B. M. Rabie..... 86

Automatic detection and boundary estimation of optic disk in fundus images using geometric active contours
G. B. Kande, T. S. Savithri, P.V. Subbaiah, M. R. N. Tagore..... 90

The effect of different number of diffusion gradients on SNR of diffusion Tensor-derived measurement maps
N. Zhang, Z. S. Deng, F. Wang, X. Y. Wang..... 96

The impact of frequency aliasing on spectral method of measuring T wave alternans
D. H. Chen, S. Yang..... 102

Micropath - A pathway-based pipeline for the comparison of multiple gene expression profiles to identify common biological signatures
M. Khan, C. B. Gorle, P. Wang, X. H. Liu, S. L. Li..... 106

Prediction of mutation position, mutated amino acid and timing in hemagglutinins from North America H1 influenza A virus
S. M. Yan, G. Wu..... 117

Bioinformatics analysis and characteristics of envelop glycoprotein E epitopes of dengue virus
H. Zhong, W. Zhao, L. Peng, S. F. Li, H. Cao..... 123

Analysis and expression of the polyhedrin gene of antheraea pernyi Nucleopolyhedrovirus (AnpeNPV)
J. X. Huang, H. L. Wu, Y. Wu, S. Y. Zhu, W. B. Wang..... 128

977193768700105