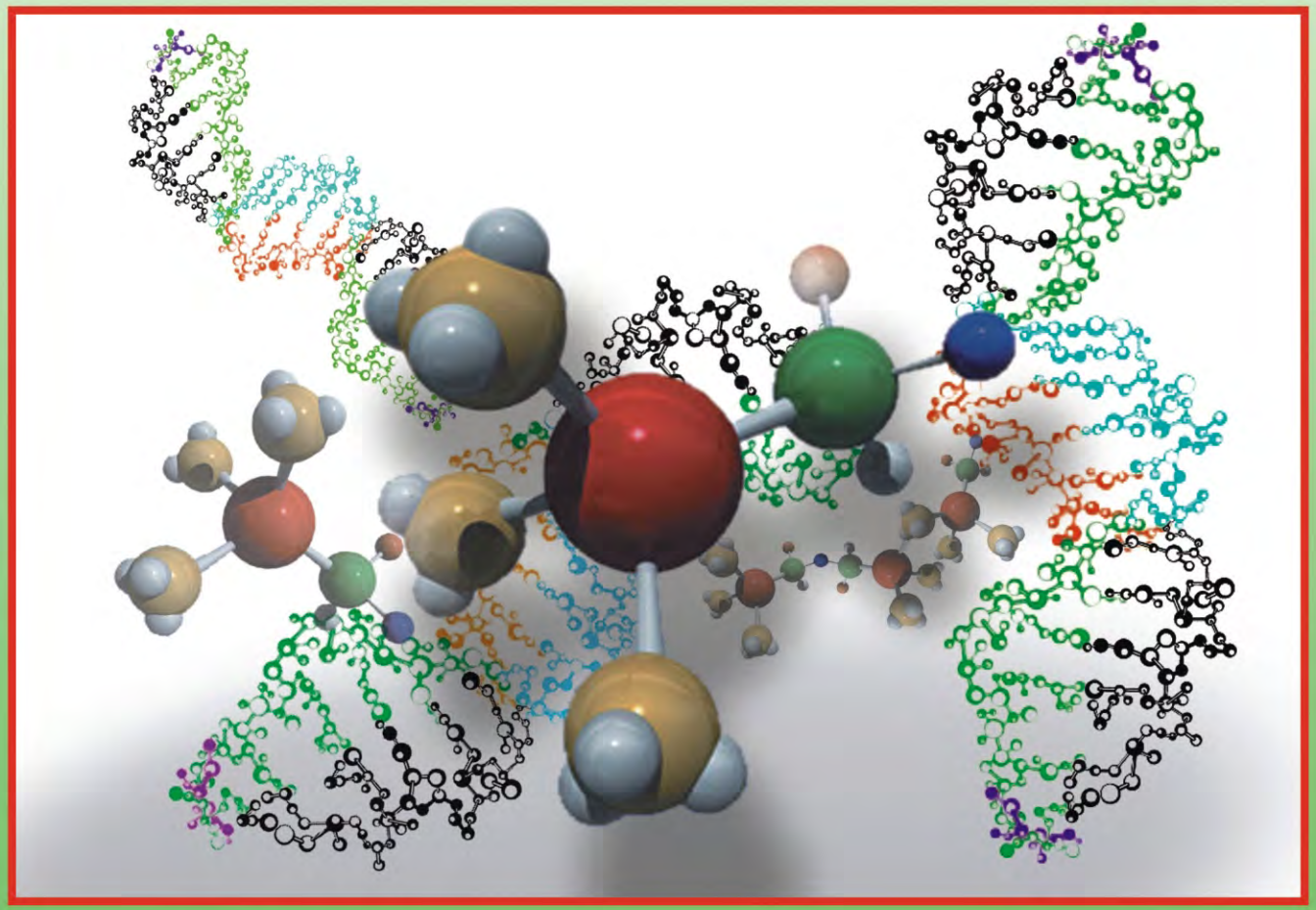
Scientific Research

JBiSE

ISSN: 1937-6871

Volume 2 Number 8 December 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 December 2009 in SciRes (http://www.scirp.org/journal/jbise).

TABLE OF CONTENTS

Volume 2, Number 8, December 2009

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 monthly by Scientific Research Publishing, Inc., USA.
E-mail: service@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 service@scirp.org.

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

SERVICES

Advertisements

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

Reprints (a minimum of 100 copies per order)

Contact the Reprints Co-ordinator, Scientific Research Publishing, Inc., USA.
E-mail: service@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, 575-581 **JBiSE**
doi: 10.4236/jbise.2009.28083 Published Online December 2009 (http://www.SciRP.org/journal/jbise/).

A novel vague set approach for selective contrast enhancement of mammograms using multiresolution

Arpita Das[1], Mahua Bhattacharya[2*]

[1]Institute of Radio Physics & Electronics, University of Calcutta, Kolkata, India;
[2]Indian Institute of Information Technology & Management, Gwalior, India.
Emali: arpita.rpe@caluniv.ac.in; *mb@iiitm.ac.in

Received 23 July 2009; revised 2 September 2009; accepted 3 September 2009

ABSTRACT

The proposed algorithm introduces a novel vague set approach to develop a selective but robust, flexible and intelligent contrast enhancement technique for mammograms. Wavelet based filtering analysis can produce Low Frequency (LF) and High Frequency (HF) subbands of the original input images. The extremely small size microcalcifications become visible under multiresolution techniques. LF subband is then fuzzified by conventional fuzzy c-means clustering (FCM) algorithm with justified number of clusters. HF components, representing the narrow protrusions and other fine details are also fuzzified by FCM with justified number of clusters. Vague set approach captures the hesitancies and uncertainties of truly affected masses/other breast abnormalities with normal glandular tissues. After highlighting the masses/microcalcifications accurately, both LF and HF subbands are transformed back to the original resolution by inverse wavelet transform. The results show that the proposed method can successfully enhance the selected regions of mammograms and provide better contrast images for visual interpretation.

Keywords: Multiresolution; Vague Set Approach; True-membership Functions; False-Membership Functions

1. INTRODUCTION

Breast cancer continues to be a significant public health problem. Primary prevention seems to be impossible since the causes of this disease still remain unknown. Thus early detection is the key to improving breast cancer prognosis. Screen/Film mammography is one of the most reliable, effective methods for early detection of breast carcinomas in women [1]. Screening of asymptomatic women using screen/film mammography has been shown a significantly reduction of breast cancer death rate.

Major advances in screen/film mammography have been occurred in the past few decades which in turns improved the image resolution and film contrast [2]. Despite of these advantages, screen/film mammography based image interpretation still remains very difficult. Breast mammograms are generally examined in presence of benign/malignant masses and other indirect signs of abnormalities like microcalcifications, skin thickening. The major reason of poor visibility is due to the slight differences of X-ray attenuation between normal glandular tissues and affected mass/microcalcification spots. The extreme small size of microcalcifications is also a reasonable cause of its low contrast appearance in mammograms. These facts create problem in the detection of breast cancer, especially in younger women with dense breasts. For this purpose development of computerized automated breast cancer diagnosis system attracts much attention of the researchers.

As it is described that contrast enhancement of mammographic features is critical but essential for breast cancer diagnosis. Many conventional contrast enhancement techniques adopt a global approach to enhance the images. However, it is quite difficult to enhance all features equally in the mammograms using those global approaches, because many local contrast information and details may be lost in the dark or bright regions of the breast [3].

As a result local-contrast enhancement techniques are developed to highlight the necessary local features. Adaptive neighborhood contrast enhancement (ANCE) method was implemented for improvement of medical image quality [4]. ANCE method provides the advantages of enhancing or preserving image contrast while suppressing the noise. However, it has a drawback. The performance of the ANCE method largely depends on how to determine the parameters used in the processing steps. In the article [5], contrast-to-noise ratio (CNR) of the low-contrast lesions is improved relative to the background. Moreover, adaptive contrast enhancement (ACE) algorithm is developed to adjust high frequency components of the images using contrast gains [6]. Incorporat-

ing non-linear function for computing ACE produces an adequate contrast gains resulting in little noise over enhancement. Mammographic feature enhancement algorithms are also key to the detection of breast abnormalities [7]. Lai *et al.* [8] compared several image enhancement methods for detecting circumscribed masses in mammograms. Authors compared an edge-preserving smoothing function [9], a half-neighborhood method [10], k-nearest neighborhood, directional smoothing [11], and median filtering [12]. In addition, authors also proposed an algorithm of selective median filtering. Among the techniques implemented, they concluded that selective median filtering with a particular size of mask performed best for image enhancement.

The fuzzy set theory [13] provides a suitable algorithm in analyzing complex systems and decision processes when the pattern indeterminacy is due to the inherent variability and vagueness. Image enhancement using smoothing with fuzzy sets [14] is developed for improving contrast of the pixels. Adaptive fuzzy logic based contrast enhancement method [15] is also developed to enhance the important mammographic features. In this technique, uncertain nature of mammograms is captured by the fuzzy membership functions. The index of fuzziness along with entropy of an image also reflects a kind of quantitative measure of its enhancement quality [16]. In this approach, fuzzy theory is adapted to the frequency content of each coefficient block in the DCT (Discrete Cosine Transform) encoded JPEG images. In decision making problems, particularly in a case of medical diagnosis, there is a fair chance of the existence of a non-null hesitation part at each moment of evaluation of any unknown object.

Vague Sets (VS) is intuitively straightforward extensions of Zadeh's fuzzy sets [17]. The drawback of using single membership value in fuzzy set theory is that the evidence for an element $u \in U$ and the evidence against $u \in U$ are mixed together (U is a classical set of objects, called the universe of discourse). The notation of VS, proposed by Gau *et al.* [18] allows more generalized interval based membership values instead of point based single membership as in fuzzy sets. To be more precise, a basic assumption of fuzzy set theory is that, if we specify the degree of membership of an element in a fuzzy set as a real number from [0,1], say a, then the degree of its non-membership is automatically determined as 1-a, need not hold for vague sets. In VS approach, it is assumed that non-membership should not be more than 1-a. The difference expresses the hesitancy concerning both membership and non-membership of an element to a set. This is mainly due to the fact that VS is more consistent with human behavior by reflecting the hesitancy present in real life situations.

In the proposed algorithm, VS approach is implemented to develop a selective but robust, flexible and intelligent contrast enhancement method for mammograms. Moreover, the hesitation of the breast mass boundaries and other abnormalities with the surrounding dense breast tissues is captured by VS. Multiresolution technique is incorporated to achieve more accurate enhancement of the detail features. Experimental results successfully enhance the selected mass regions of mammograms. Experimental results evaluate the proposed technique with conventional fuzzy based contrast enhancement methods.

2. METHODOLOGY FOR CONSTRAST ENHANCEMENT

The overall proposed selective contrast enhancement scheme implemented on mammogram has been demonstrated in **Figure 1** stepwise and in subsequent sections. The scheme involves a) decomposition of input images by wavelet transform for multiresolution analysis of coarse and fine objects b) fuzzification by standard FCM with justified number of clusters c) implementation of Vague Set approach for capturing the uncertainties between truly masses/microcalcifications from the surrounding dense glandular breast tissues d) defuzzification of membership functions to the spatial intensity domain and finally e) reconstruction of selectively enhanced image by inverse wavelet transform.

2.1. Multiresolution Approach for Decomposition of Input Image in Fine and Coarse Objects

In an image, if both fine and coarse objects or low and high contrast objects are present simultaneously, it is advan-

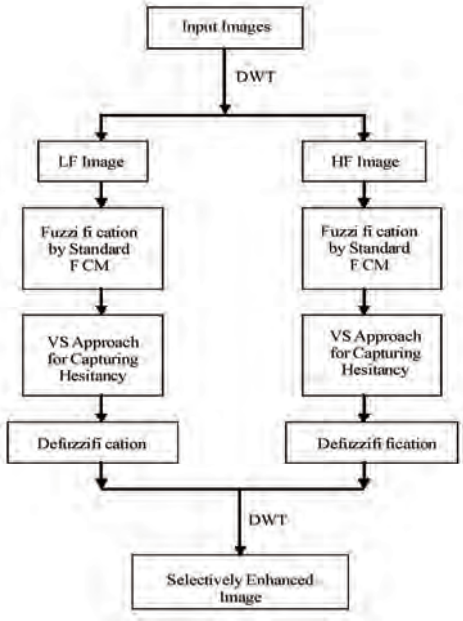


Figure 1. Block diagram for selective contrast enhancement scheme.

tageous to study them at several resolutions. This is the fundamental motivation for multiresolution processing. In the mammograms, very fine shades of gray-level intensities can be pointed out clearly by using multiresolution approach [18-19]. The mathematical concept behind this approach is described below.

Image pyramid: An image pyramid is a collection of decreasing resolution images arranged in the shape of a pyramid as shown in **Figure 2**. The base of the pyramid contains the highest resolution representation of the image. Moving up the pyramid both size and resolution of the images are decreased. The apex contains the lowest-resolution approximation. The base level J is size N × N when $N=2^J$, intermediate level j is size $2^j \times 2^j$, where $0 \le j \le J$. Fully populated pyramids are composed of J+1 resolution levels from $2^J \times 2^J$ to $2^0 \times 2^0$, but in practice most pyramids are truncated to P+1 levels, where $J-p \le j \le J$ and $1 < P \le J$, since a 1×1 pixel image is of little value.

As shown in **Figure 3**, any image of level j can be approximated to level j-1 by applying HAAR wavelet transform. It will contain only the gross structure of level j. The size of level j-1 image is just half of level j image.

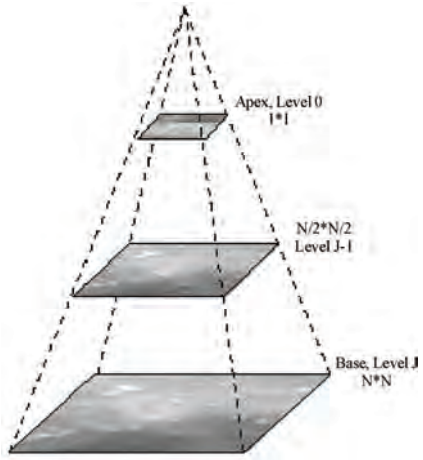


Figure 2. A pyramidal image structure.

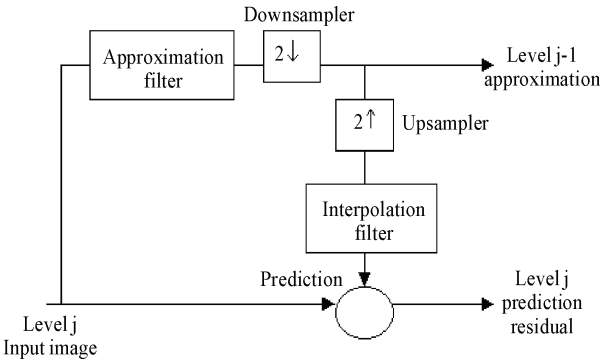


Figure 3. System for constructing image pyramids.

To determine the fine details of level j image, we first interpolate the level j-1 image to produce one of the same sizes as the level j image. This interpolated image is also known as *prediction* of level j image.

Subtracting the *prediction* of level j image from the original level j image produces level j *residual* image. This residual image contains only the fine details or the high frequency sub-band component of level j image. The technique for producing level j-1 approximation and level j prediction residual is the fundamental logic of wavelet based multiresolution processing.

In multiresolution analysis (MRA), scaling function is used to create a series of approximations of an image, each differing by a factor of 2 from its nearest neighboring approximations. Additional functions, called wavelets, are then used to encode the difference in information between adjacent approximations.

In the multiresolution approach, input image is decomposed into LF and HF subbands by forward wavelet transform whereas after defuzzification, LF and HF subbands are composed by inverse wavelet transform.

2.2. Intensity Based Clustering of LF and HF Details of Input Image Using Fuzzy C-Mean

Fuzzy c-means clustering is the most widely used algorithm of fuzzy classification. While considering the fuzzy set theory, the algorithm is developed based on k-means clustering. In this algorithm, each pixel does not belong exclusively to any single cluster but is represented by several memberships of each cluster. For a pixel, membership of each cluster is [0,1] and sum of those memberships is defined to be 1. The algorithm is performed with an iterative optimization of minimizing a fuzzy objective function.

In fuzzification step, the HF detail image is fuzzified by 3 clusters whereas LF approximate image fuzzified by 4 clusters. This particular choice of clusters represents the different gray shades of mammograms (clusters due to dark, gray, semi bright sets of pixels) along with the mass region (brighter sets of pixels).

Since FCM is intensity based clustering technique, it groups the dark, gray, bright pixels into separate clusters.

2.3. Vague Sets for Capturing Incompleteness/Hesitancy of Data

In this section the basic concepts related to Vague Sets (VS) and Fuzzy Sets are described. It is also illustrated that algebraic/graphical representation of VS is more intuitive for capturing the hesitancy or incompleteness of data.

2.3.1. Basics

Let U be a classical set of objects, called the universe of

discourse, where an element of U is denoted by u.

Fuzzy Set: A fuzzy set A = {< u, μA(u) > | u ∈ U} in a universe of discourse U is characterized by a membership function, μA, as follows: μA : U → [0,1].

Vague Set: A vague set V in a universe of discourse U is characterized by a true membership function, αV , and a false membership function, βV , as follows: αV: U → [0,1], βV: U → [0,1], and αV(u)+βV(u) ≤ 1, where αV(u) is a lower bound on the grade of membership of u derived from the evidence for u, and βV(u) is a lower bound on the grade of membership of the negation of u derived from the evidence against u.

It can be seen that the difference between VS and FS is due to the definition of membership values. In VS, the boundary (1-$\boldsymbol{\beta}_V$) is able to indicate the possible existence of a data value. This subtle difference gives rise to a simpler but meaningful graphical view of datasets. **Figure 4** and **Figure 5** depict a VS and a FS respectively. It can be seen that, the shaded part formed by the boundary in a given VS in **Figure 4** represents the possible "**hesitation region**" corresponds to the intuition of representing vague data.

In order to compare vague values, it is required to introduce two derived membership values for discussion. The first is called the **median membership**, $\mathbf{M_m} = (\boldsymbol{\alpha}_V + 1 - \boldsymbol{\beta}_V)/2$, which represents the overall evidence contained in a vague value. The second is called the **imprecision membership**, $\mathbf{M_i} = (1 - \boldsymbol{\beta}_V - \boldsymbol{\alpha}_V)$, which represents the overall imprecision/hesitation of a vague values.

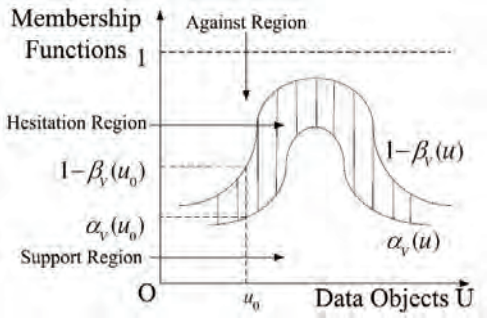


Figure 4. Membership functions of a VS.

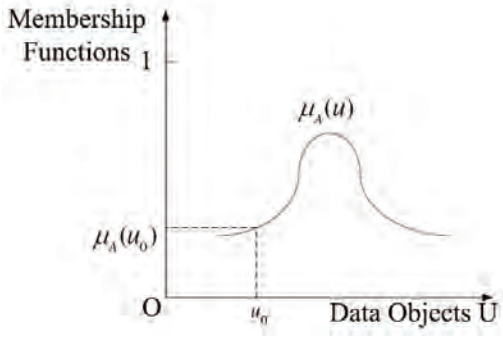


Figure 5. Membership functions of a FS.

2.4. Defuzzification and Reconstruction of Images Using LF and HF Subbands

In fuzzification step, each pixel-intensity of LF and HF subband images are transformed to intuitionistic fuzzy membership domain. After highlighting the appropriate cluster, both LF and HF subband images are transformed back to the spatial gray-level intensity domain. Then LF and HF subbands reconstruct the resulting image of original resolution by inverse wavelet transform.

3. RESULTS AND DISCUSSIONS

We have applied the proposed algorithm to a database consisting of 100 images obtained from Mammographic Image Analysis Society (MIAS), BIRADS and from EKO X-Ray & Imaging Institute, Kolkata. In the fuzzification step, the number of clusters (c) chosen by 4 and 3 for approximate images (low- frequency subband) and detail images (high frequency subband) respectively. These numbers of clusters are appropriate to represents the different gray shade intensities of the Mammographic features.

In approximate image, 4 clusters, denoted as A, B, C & D is shown graphically in **Figure 6**. Cluster A indicates the dark background of the images. Cluster B indicates the fatty breast tissues. Cluster C & D represents the diffused breast tissues and true mass regions respectively.

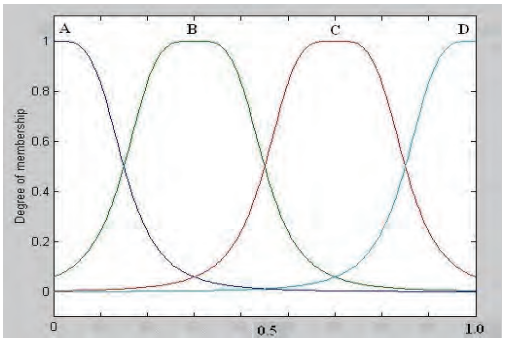


Figure 6. Schematic of fuzzy membership partition functions for approximate images.

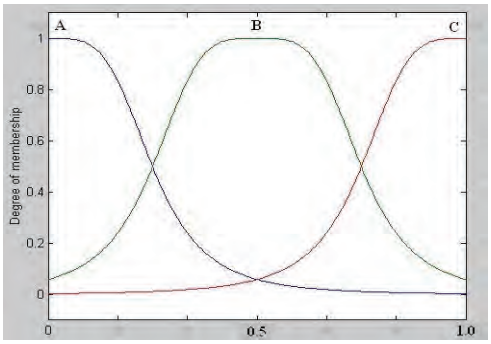


Figure 7. Schematic of fuzzy membership partition functions for detail images.

In detail image, 3 clusters, denoted as A, B & C is presented graphically in **Figure 7**. Cluster A indicates the drk background of the images. Cluster B indicates the irregular shades of breast due to dense parencymal tissues and Cluster C represents the fine boundaries of masses present in the mammograms.

After fuzzification with proper number of clusters, the vague set approach has been introduced for capturing the hesitancy and incompleteness of mammographic features. In the case of approximate images, only cluster D represents the membership function in evidence to true mass region (αV), and clusters A, B, C represent the evidence against the true mass region Thus the average membership function in evidence to false mass region (βV) is calculated by (A+B+C)/3.

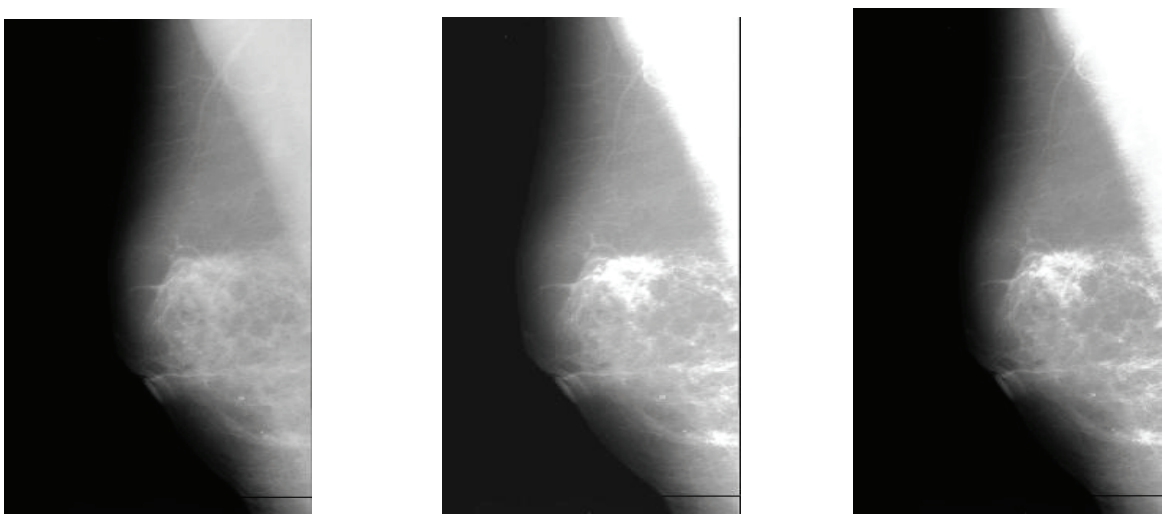
Similarly cluster C of detail images, represents the evidence of true mass boundaries (αV) whereas cluster A and B indicate the evidence against the true mass regions. The average membership function in evidence to false mass region (βV) is calculated by (A+B)/2.

The median membership value Mm = (αV+1−βV)/2, is set for the boundary value between the membership functions αV and 1−βV. In the present article, median membership value is considered as the limit of affected mass or other type of breast abnormalities.

The following mammograms are used to demonstrate the improved contrast of masses/microcalfications using VS approach in compare to the standard fuzzy set theory.

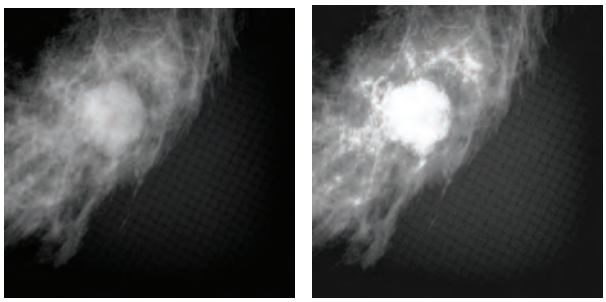
Figure 8(a) is the original mammogram to be enhanced. **Figure 8(b)** indicates the improved contrast between true mass region and normal breast tissues using the proposed algorithm, whereas **Figure 8(c)** highlight the same region using standard FCM. It is noted that **Figure 8(b)** obtained by VS approach, is capable of highlighting the true masses properly because of its efficiency to handle the uncertainty /hesitancy between true and false membership functions.

On the contrary, the enhanced result obtained by FCM skips to highlight some of the true mass regions that effect severely in proper diagnosis. **Figure 9(b)** and **Figure 10(b)** also represent improved contrast mammograms with circumscribed masses using the proposed methodology.



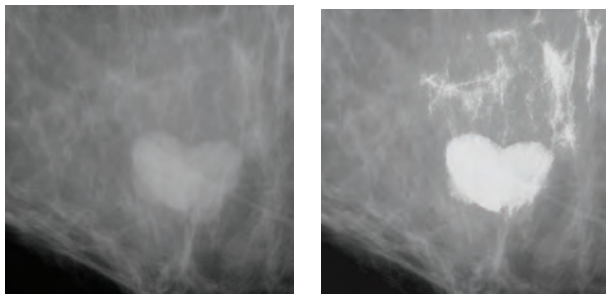
(a) (b) (c)

Figure 8. (a) Original Mammographic Image; (b) Enhanced image by VS; (c) Enhanced image by FCM.



(a) (b)

Figure 9. (a) Original Mammographic Image; (b) Enhanced image by proposed method.

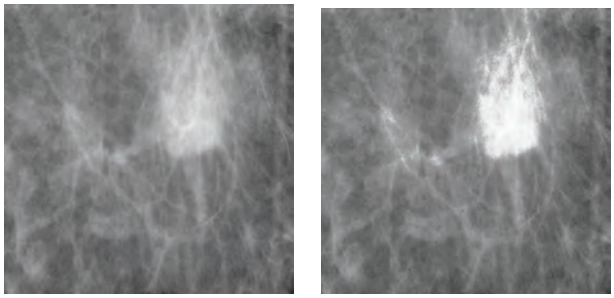


(a) (b)

Figure 10. (a) Original Mammographic Image; (b) Enhanced image by proposed method.

Figure 11(b) also shows the enhanced indistinct shaped mass region properly. In **Figure 12(b)** the proposed algorithm highlights comparatively large shaped mass boundary clearly along with the presence of few microcalcification spots surrounding the mass.

Table 1 exhibits the median and imprecision membership values of VS approach for a particular mammogram. The single membership grade of conventional fuzzy set theory for same mammographic image also listed in right most column. The larger values of median membership function, obtained from VS approach is capable of highlighting the true masses /microcalci fications with allowable hesitation margins ($\mathbf{M_m}$-μ_A) in compare to ordinary fuzzy set theory.

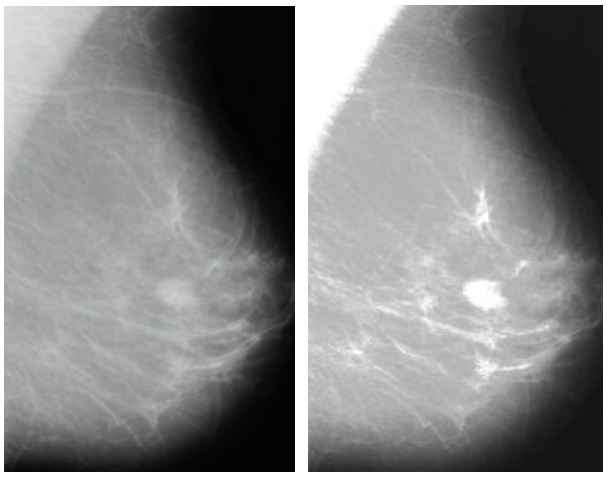


(a) (b)

Figure 11. (a) Original Mammographic Image; (b) Enhanced image by proposed method.

4. CONCLUSIONS

During the past two decades, interval based intuitionistic fuzzy sets have been used increasingly in the research areas focused on fuzzy sets and fuzzy logic. Goal of this paper is not to develop an interval based fuzzy set approach for handling the imprecise data but to apply more



(a) (b)

Figure 12. (a) Original Mammographic Image; (b) Enhanced image by proposed method.

Table 1. Membership values of vague set theory and fuzzy set approach.

Elements of clusters in Approx image	Median Membership Value (Mm) according to VS Approach	Imprecision /hesitation membership Value(Mi) according to VS Approach	Membership Value μA according to Fuzzy Set Approach
1st element of cluster D	1.0000	0.0000	1.0000
2nd element of cluster D	0.9999	0.0000	0.9999
3rd element of cluster D	0.9995	0.0004	0.9991
4th element of cluster D	0.9993	0.0013	0.9980
5th element of cluster D	0.9984	0.0006	0.9978
6th element of cluster D	0.9982	0.0023	0.9959
7th element of cluster D	0.9967	0.0032	0.9935
8th element of cluster D	0.9948	0.0019	0.9929
9th element of cluster D.	0.9942	0.0036	0.9906
10th element of cluster D	0.9926	0.0036	0.9890
11th element of cluster D	0.9910	0.0037	0.9873
12th element of cluster D	0.9901	0.0059	0.9842
13th element of cluster D	0.9874	0.0038	0.9836
14th element of cluster D	0.9869	0.0056	0.9813
15th element of cluster D	0.9844	0.0050	0.9794

robust Vague Set approach for uncer tainty management. Introducing two parameters, like median and imprecision membership values, VS approach becomes much easier to interpret and to visualize the vague data objects. Since medical diagnosis deals with the imprecise and incomplete information, accurate detection of truly affected region as well as degree of prognosis of the diseases is a difficult task. VS approach is then appropriate for the area concerning medical diagnosis. The authors have presented VS approach, which is efficient than the ordinary fuzzy sets for this purpose. The major advantage of VS over conventional fuzzy sets is that, VS includes both positive and negative evidences of an element in the universal set. As a result, the proposed method has the advantages of modelling and analyzing the uncertainties and hesitancies which are present in the diagnostic system in a more flexible and intelligent manner. Measurement of impreciseness in practice, it is found from the experimental results that VS is more natural than the conventional FS, especially for overlapping membership function domain. Introduction of multiresolution analysis makes the methodology more robust in that sense it is capable of enhancing very fine detail features by processing the high frequency subband components. For Mammographic images, ordinary contrast enhancement algorithms are unable to provide good contrast information in the selected region of interests. The concept of VS is found to be applied successfully to the problems of selective contrast enhancement. The resulted performance of the proposed algorithm shows improvements over the ordinary fuzzy based operations.

5. ACKNOWLEDGEMENMTS

The authors would like to thank to Dr. S. K. Sharma, Director, EKO Imaging and X-Ray Institute, Kolkata.

REFERENCES

[1] A. G. Haus and M. J. Yaffe, Eds (1993) A categorical course in physics, technical aspects of breast imaging, radiological society of North America, Presented at the 79 Scientific Assembly and Annual Meeting of RSNA.

[2] G. T. Barnes and G. D. Frey, Eds (1991) Screen film mammography, imaging considerations and medical physics responsibilities, Madison, WI: Medical Physics Publishing.

[3] R. C. Gonzalez and R. Woods, Digital image processing, Second Edition, Pearson Education.

[4] D-Y. Tsai, and Y. Lee, (2004) Improved adaptive neighborhood pre-processing for medical image enhancement, LNCS-3314, 576–581.

[5] P. F. Stetson, F. G. Sommer, and A. Macovski, (1997) Lesion contrast enhancement in medical ultrasound imaging, IEEE Trans. Medical Imaging, **16(4)**, 416–425.

[6] D-C. Chang and W-R Wu, (1998) Image contrast enhancement based on a histogram transformation of local standard deviation, IEEE Trans. Medical Imaging, **17(4)**, 518–531.

[7] A. F. Laine, S. Schuler, J. Fan, and W. Huda, (1994) Mammographic feature enhancement by multiscale analysis, IEEE Trans. Medical Imaging, **13(4)**, 725–740.

[8] S. Lai, X. Li, and W. F. Bischof, (1989) On techniques for detecting circumscribed masses in mammograms, IEEE Trans. Med. Imag., **8(4)**, 377–386.

[9] M. Nagao and T. Matsuyama, (1979) Edge preserving smoothing, computer graphics and image processing, **9 (4)**, 394–407.

[10] A. Scheer, F. R. D. Velasco, and A. Rosenfield, (1980) Some new image smoothing techniques, IEEE Trans. Syst., Man, Cyber., SMC-IO, **3**, 153–158.

[11] L. S. Davis and A. Rosenfield, (1978) Noise cleaning by iterated local averaging, IEEE Trans. Syst., Man, Cyber., SMC, **8**, 705–710.

[12] A. C. Bovik, T. S. Huang, and D. C. Munson, Jr., (1987) The effect of median filtering on edge estimation and detection, IEEE Trans. Pattern Anal. Machine Intell., PAMI, **9(2)**, 181–194.

[13] S. K. Pal and R. A. King, (1981) Image enhancement using smoothing with fuzzy set, IEEE Trans. Syst., Man, Cybern., SMC, **11**, 494–501.

[14] H. D. Cheng and H. Xu, (2002) A novel fuzzy logic approach to mammogram contrast enhancement, An Int. Journal on Information Sciences-Applications, **148(1-4)**, 167–184.

[15] S. K. Pal, (1982) A note on the quantitative measure of image enhancement through fuzziness, IEEE Trans. Pattern Analysis and Machine Intelligence, PAMI, **4(2)**.

[16] C. Popa, A. Vlaicu, M. Gordan and B. Orza, (2007) Fuzzy contrast enhancement for images in the compressed domain, Proc. of the Int. Multiconference on Computer Science and Information Technology, 161–170.

[17] L. A. Zadeh, (1965) Fuzzy sets, information and control, **8**, 338–353.

[18] W. L. Gau, D. J. Buehrer, (1993) Vague sets, IEEE Trans. Systems, Man, and Cybernetics, **23**, 610–614.

J. Biomedical Science and Engineering, 2009, 2, 582-586 **JBiSE**
doi: 10.4236/jbise.2009.28084 Published Online December 2009 (http://www.SciRP.org/journal/jbise/).

A novel method to reconstruct phylogeny tree based on the chaos game representation

Na-Na Li[1], Feng Shi[1], Xiao-Hui Niu[1,2*], Jing-Bo Xia[1]

[1]College of Science, Huazhong Agricultural University, Wuhan, Hubei, China;
[2]Tongji Medical College, School of Public Health, Wuhan, China.
Email: niuxiaoh@126.com

Received 5 September 2009; revised 20 September 2009; accepted 23 September 2009.

ABSTRACT

We developed a new approach for the reconstruction of phylogeny trees based on the chaos game representation (CGR) of biological sequences. The chaos game representation (CGR) method generates a picture from a biological sequence, which displays both local and global patterns. The quantitative index of the biological sequence is extracted from the picture. The Kullback-Leibler discrimination information is used as a diversity indicator to measure the dissimilarity of each pair of biological sequences. The new method is inspected by two data sets: the Eutherian orders using concatenated H-stranded amino acid sequences and the genome sequence of the SARS and coronavirus. The phylogeny trees constructed by the new method are consistent with the commonly accepted ones. These results are very promising and suggest more efforts for further developments.

Keywords: CGR (Chaos Game Representation); Discrimination Information; Phylogeny Tree

1. INTRODUCTION

Development of the nucleotide and protein sequencing technology have resulted in an explosive growth in the number of known DNA and protein sequences, it has raised many fundamental and challenging questions to modern biology. By analyzing a set of amino acid sequences (or proteins) of different species, reconstruction of the evolutionary history of genes and species is one of the most important subjects in the current study of molecular evolution. Although it is an important problem in bioinformatics, and, like many other problems, it is still an open subject for research. It is mainly due to the high degree of complexity of the problem [1] that leads to intractable search spaces when dealing with the phylogeny of a large number of species.

Current methods for the reconstruction of phylogeny trees can be roughly grouped into three kinds: maximum likelihood [2,3], maximum parsimony method [4] and distance-based methods. Maximum parsimony and maximum likelihood methods use previously aligned sequences of nucleotides as input, and they are less susceptible to errors. On the other hand, distance-based methods, such as UPGMA (unweighted pair group method using arithmetic averages) [5], Fitch-Margoliash [6] and neighbor-joining [7] use a matrix representing the distances between pairs of species, and they are based on the principle of similarity.

CGR [8,9] of biological sequences can investigate different hiding patterns of different biological sequences. It has been reported that for biological sequences at least 2000 bases are required to generate identifiable patterns [9], which do not depend on the order in which they are concatenated. In this paper, when the length of sequence is shorter than 2000 bases, we concatenate the sequence with itself repeatedly until the whole length has surpassed 2000. And we use the Kullback-Leibler discrimination information to measure the dissimilarity of each pair of biological sequences. The results proved the method is promising.

2. MATERIALS AND METHODS

2.1. Data Sets

In order to test our method, we have selected two test data, protein sequence and DNA sequence separately. The reconstruction of whole protein and nucleotide phylogenies using our new distance, all achieved very encouraging results.

2.1.1. Protein Data Set

It has been debated which two of three main groups of placental mammals are more closed related: Primates, Ferungulates, and Rodents. This is because by the maximum likelihood method, some proteins support the (Ferungulates, (primates, Rodents)) grouping while other proteins support the (Rodents, (Ferungulates, Primates)) grouping [10]. Cao *et al*. aligned 12 concatenated mito-

chondrial proteins from the following species (available in the EMBL database (release 61)): human (Homo sapiens, V00622), common chimpanzee (Pan troglodytes, D38116), pygmy chimpanzee (Pan paniscus, D38113), gorilla (Gorilla gorilla, D38114), orangutan (pongo pygmaeus, D38115), gibbon (Hylobates lar, X99256), Sumatran orangutan (pongo pygmaeus abelii, X97707), rat (Rattus norvegicus, X14848), house mouse (Mus musculus, V00711), grey seal (Halichoerus grypus, X72004), harbor seal (Phoca vitulina, X63726), cat (Felis catus, U20753), white rhino (Ceratotherium simum, Y07726), horse (Equus caballus, X79547), finback whale (Balaenoptera physalus, X61145), blue whale (Balaenoptera musculus, X72204), cow (Bos taurus, V00654), using opossum (Didelphis virginiana, Z29573), wallaroo (Macropus robustus, Y10524) and platypus (Ornithorhynchus anatinus, X83427) as the out-group, and built the maximum likelihood tree to confirm the grouping (Rodents, (Primates, Ferungulates)). So we select this controversial data set to test our method.

2.1.2. DNA Data Set

From NCBI (National center for biotechnology information), we download the 12 coronavirus sequences and 12 SARS virus sequences [11,12] that have been cultured isolating from the index case from all over the world. The 24 complete genome sequences' logo, accession, host, and location are listed in the **Table 1**.

2.2. Chaos Game of Representation of Proteins

It is known that the protein sequence is formed by 20 different kinds of amino acids. Basu. *et al*. [8] classify 20 kinds of amino acids to 12 different groups according to their different conservative substitutions such as alanine (A) and glycine (G), are considered as one vertex; serine (S) and threonine (T) represent a vertex; and so on. Furthermore, Basu. *et al*. claims that the following 12-vertex CGR algorithm is optimum for generation of distinct patterns for different protein families.

Following the chaos game algorithm, the first amino acid residue of the concatenated protein sequence is plotted halfway between the random initial point and the vertex labelled with the first residue. The second residue in the sequence is then plotted halfway between the first point and the vertex labelled with the second residue. The process must be repeated until the last residue in the sequence is plotted.

The 12-sided polygon is divided into 24 segments (grid) as shown in **Figure 1** and the segments are labelled serially with numbers 1-24. For each segment, says S_k, we count the number of points fall in S_k, says L_k. (The points falling on boundaries should be counted in any one of the neighboring segments). Then set $G_k = L_k/N$; k = 1; 2; …; 24; where N is the length of the protein sequence. From the above 12-vertex CGR algorithm, we can transform each protein sequence into a 24-dimensional vector (G_1; …; G_{24}).

Table 1. Coronaviruses and SARS virus sequences' information.

Logo	Accession	Host	Location
cAvian	NC_001451.1	Avian	
cBovine_1	AF391541.1	Bovine	
cBovine_2	AF391542.1	Bovine	
cBovine_3	U00735.2	Bovine	
cBovine_4	AF220295.1	Bovine	
cHuman	AF304460.1	Human	
cMouse	AF029248.1	Murine	
cMurine_1	AF208066.1	Murine	
cMurine_2	AF201929.1	Murine	
cMurine_3	AF208067.1	Murine	
cPig_1	NC_002306.2	Pig	
cPig_2	NC_003436.1	Pig	
SARS_BJ01	AY278488.2	Human	Beijing
SARS_HK_1	AY282752.1	Human	Hong Kong
SARS_HK_2	AY278491.2	Human	Hong Kong
SARS_HK_3	AY278554.2	Human	Hong Kong
SARS_SG_1	AY283794.1	Human	Singapore
SARS_SG_2	AY283795.1	Human	Singapore
SARS_SG_3	AY283796.1	Human	Singapore
SARS_SG_4	AY283797.1	Human	Singapore
SARS_SG_5	AY283798.1	Human	Singapore
SARS_TOR2	AY274119.3	Human	Toronto in Canada
SARS_TW1	AY291451.1	Human	Taiwan
SARS_Urban	AY278741.1	Human	United States

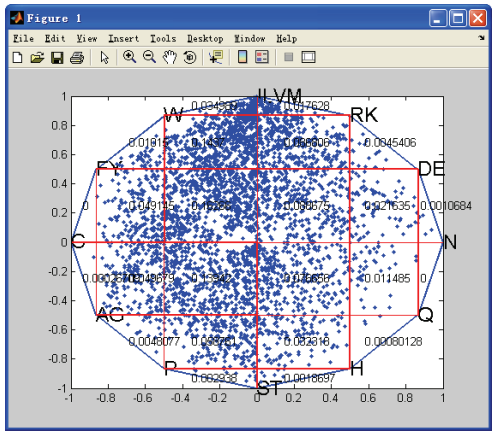


Figure 1. Chaos game representation of protein.

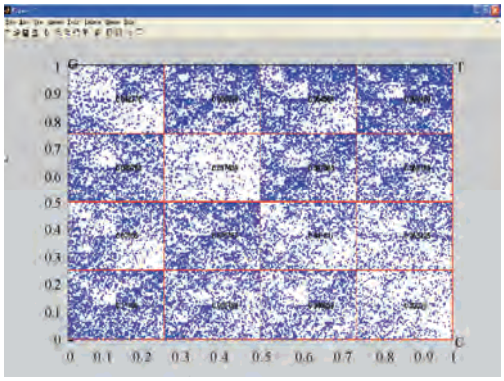


Figure 2. Chaos game of representation of DNA.

2.3. Chaos Game Representation of DNA Sequence

Similar to the chaos game representation of proteins, each of the four vertex of the square is labelled 'a', 'c', 'g', or 'u'. According to the DNA sequence [9], we plot half way between the random initial point and the vertex labeled with the first nucleotide acid. Then the second nucleotide acid in the sequence is plotted halfway between the first point and the vertex labelled with the second one. Following this method, it is repeated until the last nucleotide acid is plotted.

The square is divided into 16 segments (shown in **Figure 2**). Each of segments is labelled with the numbers 1-16.Then we can count the percent of the points that are fallen into each of segment. Following this algorithm, each DNA sequence will induce a 16-dimensional vector (G1; …; G16).

2.4. The Kullback-Leibler Discrimination Information

X is a discrete random variable. It has the different distribution laws under the different hypotheses. Such as, under hypothesis H1, its distribution law is defined as follow:

$$\begin{pmatrix} X \\ p_1(x) \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & \cdots & a_K \\ p_1(a_1) & p_1(a_2) & \cdots & p_1(a_K) \end{pmatrix}$$

By similarity, under hypothesis H2, its distribution is similar:

$$\begin{pmatrix} X \\ p_2(x) \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & \cdots & a_K \\ p_2(a_1) & p_2(a_2) & \cdots & p_2(a_K) \end{pmatrix}$$

The Kullback-Leibler discrimination information between the two distributions is defined as follow:

$$I(p_1, p_2) = \sum_{i=1}^{K} p_1(a_i) \log \frac{p_1(a_i)}{p_2(a_i)}$$

The detailed step to measure the dissimilarity using this concept is listed as follow.

For example, there are two sequences, X and Y. Following the CGR algorithm, they can transform into the vector of the percent, ($G_X(1)$; …; $G_X(k)$) and ($G_Y(1)$; …; $G_Y(k)$) k = 16 or 24, according to the kind of biological sequence. The two vectors can be seen as the two different distribution laws.

$$I(X, Y) = \sum_{i=1}^{k} G_X(i) \log \frac{G_X(i)}{G_Y(i)} \lim_{x \to \infty}$$

Then the Kullback-Leibler discrimination information of two frequencies distribution is defined as follow:

I(X, Y) denote the discrimination information between the X and Y. It is should be noted that maybe some $G_Y(i) = 0$, this make $G_X(i)/G_Y(i)$ no sense. In this case, we may treat $G_Y(i)$ as a very small positive real number, and this would not cause trouble, and make our discussion very conversional. At the same, we always note that $0 \cdot \log 0 = 0$.

Because the discrimination information has direction (also termed as directed divergence), it is I(X, Y) ≠ I(Y, X) in general, so we now introduce another measure J(X, Y) as the following:

$$J(X, Y) = I(X, Y) + I(Y, X)$$

Then J(X, Y) has the following properties:

(1) J(X, Y) ≥ 0

(2) J(X, Y) = 0 if and only if X = Y.

(3) J(X, Y) = J(Y, X).

At last, we introduce Distance (X, Y) to measure the diversity (dissimilarity) of the biological sequences, X and Y.

3. RESULTS

3.1. Protein Data Set

With the protein data set, firstly, the out-group species separate from other mammals. Secondly, the three classes grouped each other obviously. Above all, we computed the Distance (X, Y) for each pair of species X and Y and constructed a tree (shown in **Figure 3**) using the neighbor joining [7] program in the MOLPHY package. The tree is very close to the maximum likelihood tree of Cao *et al* [10]. We also support the collusion of the (Rodents, (Ferungulates, Primates)) grouping. And we try to connect the midpoint of every edge to divide the polygon into 84 segments. Then following the same routine, we get the similar phylogeny tree, there is one difference from the previous tree that the horse's position is different.

3.2. DNA Data Set

With the DNA data set, we reconstructed the phylogeny tree (shown in **Figure 4**), separated the coronavirus sequences and SARS sequences completely. And the SARS sequences are more resemble to the first group of coronavirus. These results are similar to the commonly accepted results [13]. The 12 SARS virus sequences are obviously separated from the 12 coronavirus sequences. It supports the conclusion that SARS virus belong to the coronavirus, but they are different from the conventional coronavirus. On the phylogeny tree, SARS viruses are closest to the c_pig1, c_pig2 and c_Human which belong to the first kind of the coronavirus according to the serotype. It shows that SARS virus is nearest to the first kind of coronavirus. This is different from the Rota *et al* [13]. But it supports the experiment result of the Ksiazek *et al* [14].

Then we further divide the every segment into four average parts. That is to say, we divide the square into 64 segments. With the same method, we get the completely same tree.

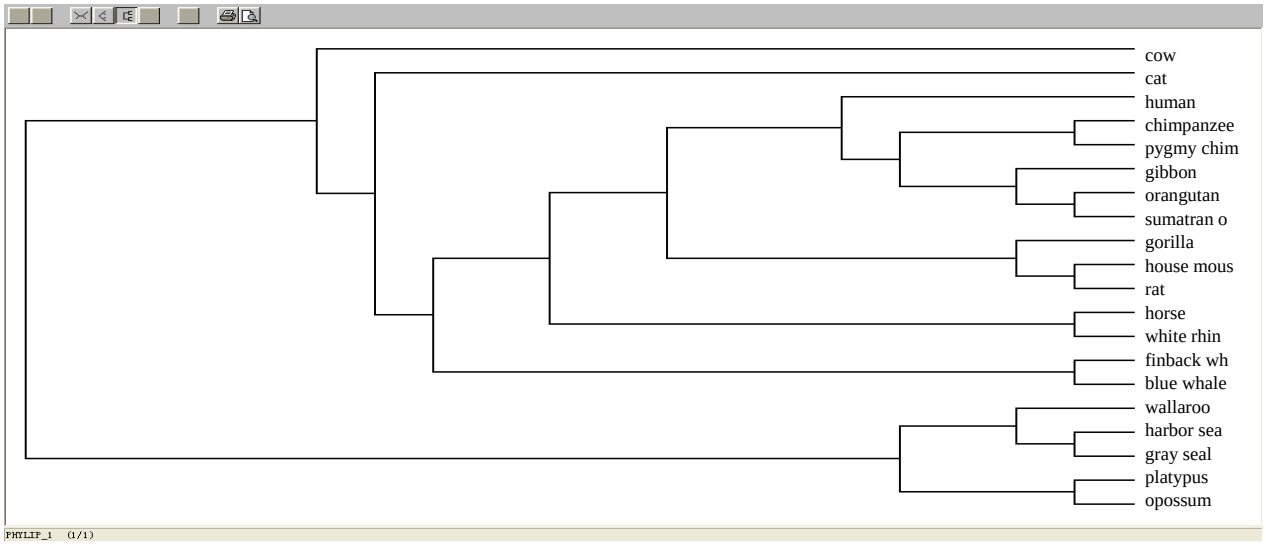


Figure 3. Phylogeny tree with the mitochondrial proteins from 20 species.

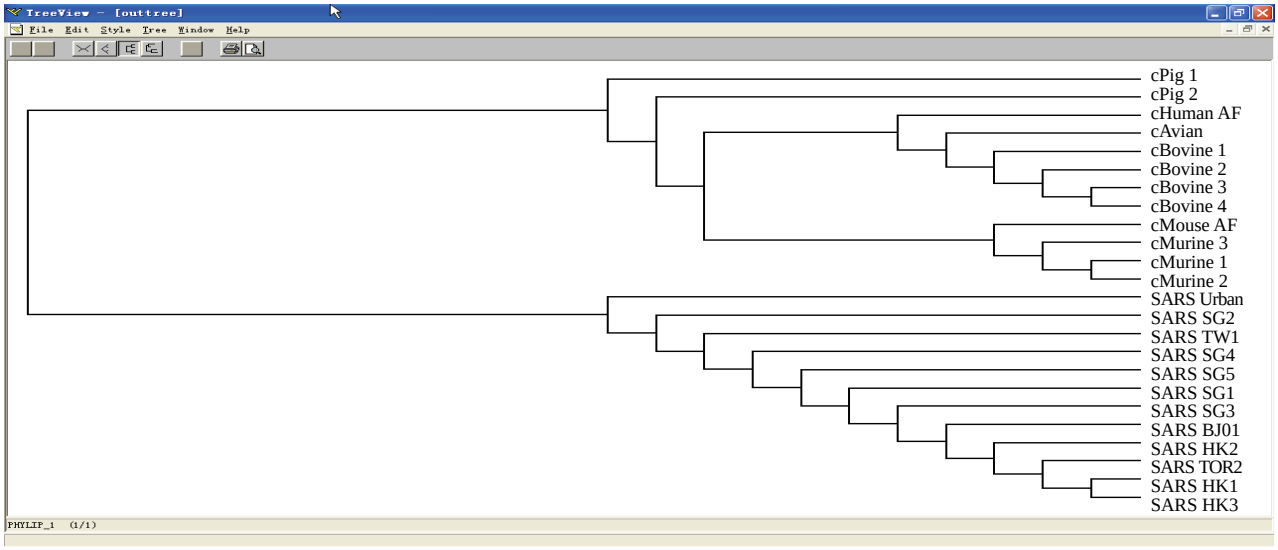


Figure 4. Phylogeny tree with coronavirus and SARS virus sequences.

4. CONCLUSIONS

We develop the new method based on the CGR of biological sequences. We achieved the promising results. This method is universal. It can reconstruct the phylogeny tree not only with the protein sequences data but also with the DNA (or RNA) sequences data. The numerical experiments show its stability. We tried to divide the square (or polygon) into more segments, and then we reconstruct the phylogeny tree in the similar way. We achieved the similar results. That is to say, the CGR method can show the distinct pattern for different proteins, no matter how to divide the pictures. And the Kullback-Leibler discrimination information can measure the dissimilarity of the proteins rightly.

The successful application to reconstruct the phylogeny tree means that this new measurement of the dissimilarity between the biological molecules can not only use to reconstruct the phylogeny tree, but also apply to other comparative genomics research communities.

REFERENCES

[1] G. H. Gonnet. (1994) New algorithms for the computation of evolutionary phylogenetic trees [M], ComputationalMethods in Genome Research (Suhai, S., ed.), Plenum, New York, 153–161.

[2] L. L. Cavalli Sforza and A. W. Edwards. (1967) Phylogenetic analysis: Models and estimation procedures [J], Genetics, **19(3)**, 233–257.

[3] J. Felesenstein. (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach [J], J Mol Evol, **17(6)**, 368–376.

[4] L. Jin and M. Nei. (1990) Limitation of the evolution parsimony method of phylogenetic analysis [J], Mol Biol Evol, **7(1)**, 82–102.

[5] R. R. Sokal and C. D. Michener. (1958). A statistical method for evaluating systematic relationships [J], Univ. Kans. Sci. Bull, **28**, 1409–1438.

[6] Chris. (2004) Fitch-Margoliash algorithm for calculating the branch lengths [EB/OL], http://www.bioinfo.rpi.edu/~bystrc/courses/biol4540/lecture12/sld002.htm.

[7] N. Saitou and M. Nei. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees [J], Molecular Biology and Evolution, **4(4)**, 406–425.

[8] S. Basu, A. Pan, C. Dutta and J. Das. (1997) Chaos game representation of protein, J. Mol. Graphics Model, **15**, 279–289.

[9] H. J. Jeffrey. (1990) Chaos game representation of gene structure [J], Nucleic AcidsRes., **18**, 2163–2170.

[10] Y. Cao, N. Okada, and M. Hasegawa. (1997) Phylogenetuc position of guinea pigs revisited [J], Mol. Biol. Evol., **14**, 461–464.

[11] M. A. Marra, S. J. Jones, C. R. Astell, *et al*. (2003) The genome sequence of the SARS-associated coronavirus [J], Science, **300(5624)**, 1399–1404.

[12] Y. J. Ruan, C. L. Wei, L. A. Ee, *et al*. (2003) Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection [J], The Lancet, **361(9371)**, 1779–1785.

[13] P. A. Rota, M. S. Oberste, S. S. Monroe, *et al*. (2003) Characterization of a novel coronavirus associated with severe acute respiratory syndrome [J], Science, **300 (5624)**, 1394–1399.

[14] T. G. Ksiazek, D. Erdman, C. Goldsmith, *et al*. (2003) A novel coronavirus associated with severe acute respiratory syndrome [J], N Engl J Med, **348(20)**, 1953–1966.

J. Biomedical Science and Engineering, 2009, 2, 587-593 **JBiSE**
doi: 10.4236/jbise.2009.28085 Published Online December 2009 (http://www.SciRP.org/journal/jbise/).

Fitting evolutionary process of matrix protein 2 family from influenza A virus using analytical solution of differential equation

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

[1]National Engineering Research Center for Non-food Biorefinery, Guangxi Academy of Sciences, 98 Daling Road, Nanning, Guangxi, China;
[2]Computational Mutation Project, DreamSciTech Consulting, 301, Building 12, Nanyou A-zone, Jiannan Road, Shenzhen, Guangdong, China.
Email: hongguanglishibahao@yahoo.com

Received 22 August 2009; revised 3 September 2009; accepted 4 September 2009.

ABSTRACT

The evolution of protein family is a process along the time course, thus any mathematical methods that can describe a process over time could be possible to describe an evolutionary process. In our previously concept-initiated study, we attempted to use the differential equation to describe the evolution of hemagglutinins from influenza A viruses, and to discuss various issues related to the building of differential equation. In this study, we attempted not only to use the differential equation to describe the evolution of matrix protein 2 family from influenza A virus, but also to use the analytical solution to fit its evolutionary process. The results showed that the fitting was possible and workable. The fitted model parameters provided a way to further determine the evolutionary dynamics and kinetics, a way to more precisely predict the time of occurrence of mutation, and a way to figure out the interaction between protein family and its environment.

Keywords: Amino-Acid Pair Predictability; Differential Equation; Evolution; Fitting; Influenza A Virus; Matrix Protein 2

1. INTRODUCTION

Very recently, we explored the possibility to use a differential equation to describe the evolution of hemagglutinin proteins from influenza A virus [1].

However, there are ten types of proteins from influenza A viruses, it is necessary to explore whether or not this differential description can be applied to other proteins from influenza A viruses. Also, it is intriguing to use the analytic solution of differential equation to fit the evolution of proteins from influenza A viruses.

This is so, because the mathematical modeling is generally the ending point of empiric experiments, and more importantly the mathematical modeling can provide us the tool for predicting the future of evolution.

As the evolution is a process along the time course, we at first needed to represent an evolutionary subject along the time course, and then we could consider how to apply the mathematical modeling to this process.

Now we are particularly interested in the evolution of proteins. However, a protein generally is a sequence of letters, which represent amino acids. Thus we need a method to represent a protein family along the time course before modeling [2-5].

In general, the evolution is a process of exchanging substances between a living subject and its environment. In this context, the differential equation is quite suitable because we defined $\frac{dy}{dt} = input - output$ for exchanging substance between a protein family and its environment along the time course [1].

As we know that the evolution of proteins goes through mutations, which bring in new mutating amino acids and take out mutated amino acids. This again requires the conversion of amino acids into numbers to represent the exchange [2-5].

Among ten types of proteins from influenza A virus, the matrix protein 2 (M2) is important because it constructs a proton channel in the virion and it is essential for infection [6]. Thus, the M2 protein was the target for anti-influenza drugs, and the M2 ion channel blockers was approved to treat influenza virus infections [7,8], but their use is limited by high frequencies of the resistance among currently circulating strains [9,10]. Also, a vaccine was designed basing on the conserved ectodomain of M2 protein, which could match multiple influenza virus strains including multiple subtypes [11].

In this study, our effort was made to apply the differential description to the evolution of M2 protein family from influenza A virus, and to use the analytical solution of differential equations to fit the evolutionary process of

M2 proteins.

2. MATERIALS AND METHODS

2.1. Data

5926 full-length M2 proteins of influenza A virus sampled from 1959 to 2008 were obtained from the influenza virus resources [12]. After excluded identical sequences, 1084 M2 proteins were actually used in this study.

2.2. Conversion of Proteins into Numbers

For two purposes, we needed to convert M2 proteins into numbers: 1) we needed a single number for a single M2 protein so that we could present the evolution of M2 protein family over time, and 2) we needed a number to present mutation, which resulted in numerical exchange between M2 protein and its environment. We used the amino-acid pair predictability to do this job [2-5].

For example, an M2 protein (accession number ABF01755) from an avian influenza virus, strain A/chicken/Magetan/BBVW/2005(H5N1), had 97 amino acids. The first and second amino acids could be counted as an amino-acid pair, the second and third as another amino-acid pair, the third and fourth, until the 96th and 97th, thus there were 96 amino-acid pairs.

This M2 protein had 10 glutamic acids (E) and 11 leucines (L): if the permutation could predict the appearance of amino-acid pair EL, it would appear once (10/97×11/96×96=1.13); actually it did appear once, so the pair EL was predictable. By contrast, this M2 protein had 7 isoleucines (I): if the permutation could predict the appearance of amino-acid pair IL, it would appear once (7/97×11/96×96=0.79); however, it appeared three times in realty, so the pair IL was unpredictable.

In this way, all amino-acid pairs in ABF01755 M2 protein were classified as predictable and unpredictable, which were 17.71% and 82.29%.

Taking another M2 protein (accession number ABF01771) as example, this M2 protein had only one amino acid different from ABF01755 M2 protein at position 65. However, its predictable and unpredictable portions were 20.83% and 79.17%. Thus, the amino-acid pair predictability distinguished the difference between M2 proteins in numbers rather than in letters that represented amino acids in proteins.

Based on the above computation, the difference in predictable portion between ABF01755 and ABF01771 M2 proteins was −3.12% (17.71%−20.83%), which was regarded as the exchange between M2 protein and its environment.

2.3. Differential Equation

If ABF01755 and ABF01771 M2 proteins would have a direct relation due to a single mutation, the difference between them was $\frac{dy}{dt} = input - output$, where y was the difference in predicable portion, t was the time required for mutation, *input* was the predictable portion brought in by mutating amino acid, *output* was the predictable portion taken away by mutated amino acid.

Unfortunately, we had no way to know if ABF01755 and ABF01771 M2 proteins had a direct mutation relationship although both were sampled in 2005.

As the predictable portion was determined using the permutation based on random principle, thus this exchange was in fact the exchange of randomness between M2 proteins and their environment, more accurately was the exchange of entropy between M2 proteins and their environment [1].

2.4. Statistics

The Student's *t*-test and Mann-Whitney *U*-test were used to compare the difference between uphill and downhill half-life, and $P<0.05$ was considered statistically significant. The SigmaPlot for Windows was used for fitting [13].

3. RESULTS AND DISCUSSION

Ideally, we would hope to have a direct mutation relationship for all 1084 M2 proteins from 1959 to 2008 involved in this study, because then we would have $\frac{dy}{dt} = input - output$ for each mutation relationship between any related two M2 proteins.

Although it was impossible to find out such a relationship among all of these M2 proteins sampled everywhere in the world, the real mutation relationship worked in this way no matter if we had sampled them or not. Thus, we had a system of differential equations for all M2 proteins.

On the other hand, our computations on predictable portions of 1084 M2 proteins provided us with the most update evolutionary process in **Figure 1**, which was read as follows. For example, the solid curve in the top panel presented the evolution of 1084 M2 proteins from 1959 to 2008, and each point was the mean value of predictable portions of all M2 proteins in given year with its standard deviation (vertically grey line). The similar reading was applied to other panels.

So the fluctuating solid curves in **Figure 1** presented the evolutionary process over time. If we could use the differential equation to describe these solid curves, it would mean that we were able to model the evolutionary process of M2 proteins.

3.1. Possibly Analytical Solution

These fluctuating solid curves suggested that the possi-

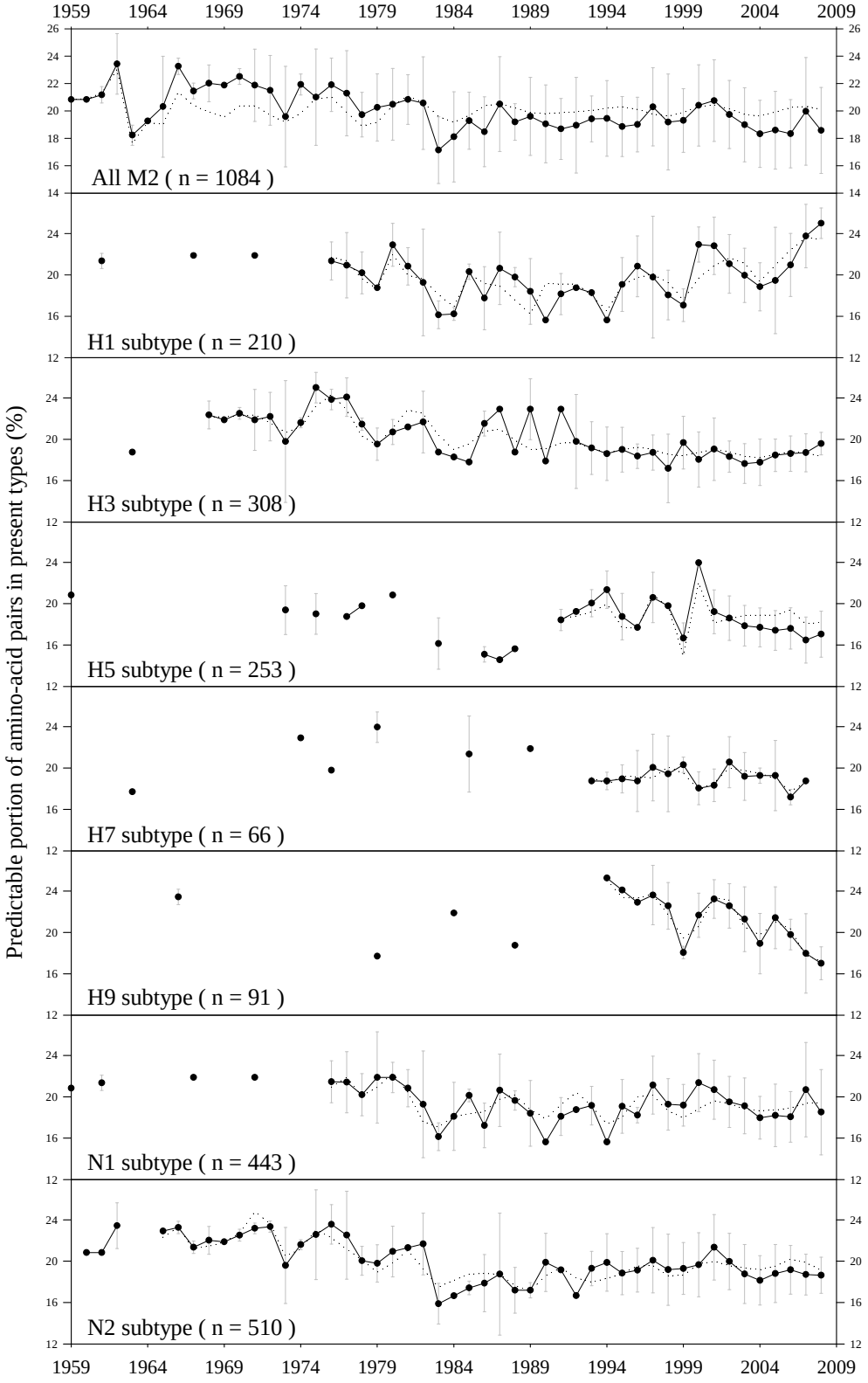


Figure 1. Evolutionary process of M2 proteins from influenza A viruses from 1959 to 2008 in terms of predictable portions with respect to different subtypes. The data are presented as mean±SD. The solid and dotted lines are the actual evolutionary process and fitted evolutionary process with analytical solution of differential equation.

bly analytical solution for n differential equations would be a sum of decaying exponential and sinusoidal functions $y(t)=\sum_{i=1}^{n} A_i e^{-k_i t}\cos(\alpha_i t+\varphi_i)+C$, where y was the fluctuating solid curve representing the predictable portion over time, A, α and k were parameters, t was time, φ was phase difference, and C was a constant [14].

3.2. Half-Life

This analytical solution governed a decaying trend with fluctuating solid curve because of negative exponential. Hence, we were able to determine the half-life of decaying phase of the fluctuating solid curve immediately. With decaying exponential, the half-life was $T_{1/2}=\frac{\ln(2)}{k}=\frac{0.696}{k}$, where $k=\frac{\ln(y_{peak})-\ln(y_{trough})}{t_{interval}}$, which was the downhill half-life. Symmetrically, we were able to compute the uphill half-life too because the uphill phase suggested that mutations led M2 proteins to become more predictable whereas the downhill phase suggested that mutations led M2 proteins to become less predictable.

Table 1 showed the computed half-life for all possible stratified peaks and troughs, and **Figure 2** compared the uphill half-life with the downhill one. As no statistical difference was found in **Figure 2**, it indicated that the uphill half-life was not different from the downhill half-life in principle.

These results suggested that we were able to use the analytical solution to fit the solid curve, because the unsolved problems in our previous study were that we were not able to determine the input function for this differential equation and were not able to determine if this evolutionary process was at steady-state.

Table 1. Half-life for all and each subtype of M2 proteins from influenza A viruses.

M2	Period	Length	Predictable portion (%)		Half-life (years)	
Subtype	Year	Years	Peak/Trough	Peak/Trough	Uphill	Downhill
All	1960-1962	3	20.83	23.44	18	
	1962-1963	2	23.44	18.23		6
	1963-1966	4	18.23	23.26	11	
	1966-1967	2	23.26	21.46		17
	1967-1970	4	21.46	22.51	58	
	1970-1973	4	22.51	19.59		20
	1973-1974	2	19.59	21.93	12	
	1974-1978	5	21.93	19.73		33
	1978-1981	4	19.73	20.83	51	
	1981-1983	3	20.83	17.14		11
	1983-1987	5	17.14	20.50	19	
	1987-1991	5	20.50	18.69		38
	1991-1997	7	18.69	20.30	59	
	1997-1998	2	20.30	19.19		25
	1998-2001	4	19.19	20.76	35	
	2001-2004	4	20.76	18.32		22
	2004-2007	4	18.32	19.97	32	
	2007-2008	2	19.97	18.58		19
H1	1976-1979	4	21.35	18.75		21
	1979-1980	2	18.75	22.92	7	
	1980-1983	4	22.92	16.15		8
	1983-1987	5	16.15	20.63	14	
	1987-1990	4	20.63	15.63		10
	1990-1992	3	15.63	18.75	11	
	1992-1994	3	18.75	15.63		11
	1994-1996	3	15.63	20.83	7	
	1996-1999	4	20.83	17.07		14
	1999-2000	2	17.07	22.94	5	
	2000-2004	5	22.94	18.85		18
	2004-2008	5	18.85	25.00	12	
H3	1972-1973	2	22.20	19.79		12
	1973-1975	3	19.79	25.00	9	
	1975-1979	5	25.00	19.53		14
	1979-1982	4	19.53	21.67	27	

	1982-1985	4	21.67	17.80		14
	1985-1987	3	17.80	22.92	8	
	1987-1988	2	22.92	18.75		7
	1988-1989	2	18.75	22.92	7	
	1989-1990	2	22.92	17.90		6
	1990-1991	2	17.90	22.92	6	
	1991-1998	8	22.92	17.19		19
	1998-1999	2	17.19	19.68	10	
	1999-2003	5	19.68	17.65		32
	2003-2008	6	17.65	19.59	40	
H5	1994-1996	3	21.35	17.71		11
	1996-1997	2	17.71	20.60	9	
	1997-1999	3	20.60	16.67		10
	1999-2000	2	16.67	23.96	4	
	2000-2007	8	23.96	16.48		15
H7	1996-1999	4	18.75	20.31	35	
	1999-2000	2	20.31	18.06		12
	2000-2002	3	18.06	20.57	16	
	2002-2006	5	20.57	17.19		19
	2006-2007	2	17.19	18.75	16	
H9	1997-1999	3	23.62	18.06		8
	1999-2001	3	18.06	23.24	8	
	2001-2004	4	23.24	18.93		14
	2004-2005	2	18.93	21.42	11	
	2005-2008	4	21.42	17.01		12
N1	1976-1978	3	21.46	20.21		35
	1978-1979	2	20.21	21.88	18	
	1980-1983	4	21.88	16.15		9
	1983-1985	3	16.15	20.14	9	
	1985-1986	2	20.14	17.23		9
	1986-1987	2	17.23	20.63	8	
	1987-1990	4	20.63	15.63		10
	1990-1993	4	15.63	19.15	14	
	1993-1994	2	19.15	15.63		7
	1994-1997	4	15.63	21.13	9	
	1997-1999	3	21.13	19.18		22
	1999-2000	2	19.18	21.37	13	
	2000-2004	5	21.37	17.98		20
	2006-2007	2	18.07	20.70	10	
	2007-2008	2	20.70	18.52		13
N2	1961-1962	2	20.83	23.44	12	
	1966-1967	2	23.26	21.35		16
	1967-1972	6	21.35	23.34	47	
	1972-1973	2	23.34	19.59		8
	1973-1976	4	19.59	23.57	15	
	1976-1979	4	23.57	19.79		16
	1982-1983	2	21.67	15.89		4
	1983-1987	5	15.89	18.75	21	
	1987-1988	2	18.75	17.19		16
	1989-1990	2	17.19	19.89	10	
	1990-1992	3	19.89	16.67		12
	1992-1994	3	16.67	19.87	12	
	1994-1995	2	19.87	18.85		26
	1995-2001	7	18.85	21.36	39	
	2001-2004	4	21.36	18.15		17
	2004-2006	3	18.15	19.17	38	

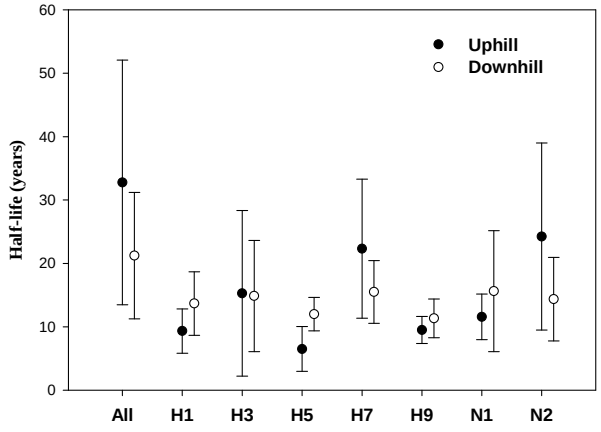


Figure 2. Comparison of uphill half-life with downhill half-life in all and different subtypes of M2 proteins from influenza A viruses. The data are presented as mean±SD.

3.3. Fitting

It was possible to use the analytical solution to fit the evolution of M2 proteins because the process of finding half-life provided the initial estimate for the parameter k_i in exponential terms.

The dotted lines in **Figure 1** were fitted lines using the analytical solution, and **Table 2** listed the fitted parameters for the analytical solution. As seen, the dotted lines generally were quite approximate to the evolutionary trend presented by the solid curve, indicating that the analytical solution was able to present the evolutionary process of M2 proteins from influenza A virus.

According to the general fitting principle, we were able to determine the goodness of our fitting through several ways, for example, 1) the Akaike's information criterion [15], 2) the plotting of residuals versus fitted predictable portion [16], 3) the plotting of residuals versus time [16], 4) the R or squared R value between fitted and actual data [13], etc. We mainly used the squared R value (**Table 2**) and Akaike's information criterion to determine if the difference between solid and dotted lines was acceptable. This was so because the sampled influenza A viruses were very unbalanced due to the practical difficulty in sampling, thus the solid lines could be biased on this account.

One possibility with this analytical solution was that the fluctuations would become less intensive as the time went on. This was possible because the evolutionary speed was becoming slower as less and less functional units needed to evolve. In fact, influenza viruses became more and more adapt to their environments after long-time evolution, thus they did not need to mutate a lot to suit for the changes in environments. This adaptation would lead the evolutionary speed of influenza A virus to be slower over time. For another example, the appendix in human could have very little speed for its evolution because its function is very much limited in general.

The use of differential equation to describe the evolution of proteins from influenza A viruses not only advanced our modeling ability in this field, but also provided us the tool to predict future mutations of influenza A viruses. For prevention of possible epidemic/pandemic, it is very important how to time mutations in proteins

Table 2. Parameters obtained after using the analytical solution to fit the evolutionary process of M2 proteins in **Figure 1.**

Subtype	All	H1	H3	H5	H7	H9	N1	N2
A_1	-1.0962	1.4086	-4.0515	-2.3573	1.5353	-3.9688	7.5911	-5.9992
k_1	0.0292	0.0000	0.1087	0.0762	0.0000	0.1360	0.2128	0.0677
a_1	0.9733	1.1999	1.0116	1.6356	1.2230	0.7309	0.6387	-0.1797
φ_1	-0.6025	-6.8726	1.1214	-1.9029	-2.5712	2.1098	3.8746	4.6835
A_2	60.1439	46.9288	-17.0142	-2.3072	-0.5694	-1.6362	2.1476	-2.0694
K_2	0.5698	0.0157	0.1206	0.0312	0.0000	0.0000	0.0557	0.0356
a_2	3.0209	6.3110	0.0685	-21.0433	1.9395	1.5054	1.5402	1.2495
φ_2	-10.7687	-4.0610	1.6425	-12.2725	3.0707	-3.1941	4.7661	0.7923
A_3	-2.5565	-1.4254	-2.6202	-1.2754	1.8712	3.3257	1.2562	1.1683
k_3	0.0687	0.0395	0.0589	0.0000	0.0000	0.0596	0.0198	0.0452
a_3	1.2499	2.4881	1.2671	2.8407	1.1729	0.2823	1.2257	4.2995
φ_3	-0.5466	3.2902	-2.6996	-0.1830	0.5379	4.8288	-1.5888	-10.0636
C	20.02	50.7341	18.4548	18.8510	19.1760	20.3824	18.9722	19.2776
R	0.5619	0.7951	0.8123	0.7793	0.8057	0.9601	0.7297	0.8772
R^2	0.3158	0.6322	0.6599	0.6074	0.6491	0.9217	0.5325	0.7695

from influenza A viruses. In the past we used the fast Fourier transform to do this job [3,5,17-20]. In the near future we are able to use the analytical solution with fitted parameters to time the mutations.

4. ACKNOWLEDGEMENTS

This study was partly supported by Guangxi Science Foundation (0907016 and 0991080) and Guangxi Academy of Sciences (0701 and 09YJ17SW07).

REFERENCES

[1] S. Yan and G. Wu. (2009) Describing evolution of hemagglutinins from influenza A viruses using a differential equation, Protein Pept. Lett., **16,** 794–804.

[2] G. Wu and S. Yan. (2002) Randomness in the primary structure of protein: Methods and implications, Mol. Biol. Today, **3,** 55–69.

[3] G. Wu and S. Yan. (2006) Mutation trend of hemagglutinin of influenza A virus: A review from computational mutation viewpoint, Acta Pharmacol. Sin., **27,** 513–526.

[4] G. Wu and S. Yan. (2006) Fate of influenza A virus proteins, Protein Pept. Lett., **13,** 377–384.

[5] G. Wu and S. Yan. (2008) Lecture notes on computational mutation, Nova Science Publishers, New York.

[6] T. Betakova. (2007) M2 protein-a proton channel of influenza A virus, Curr. Pharm. Des., **13,** 3231–3235.

[7] L. H. Pinto and R. A. Lamb. (2007) Controlling influenza virus replication by inhibiting its proton channel, Mol. Biosyst., **3,** 18–23

[8] J. Beigel and M. Bray. (2008) Current and future antiviral therapy of severe seasonal and avian influenza, Antiviral Res., **78,** 91–102.

[9] E. De Clercq and J. Neyts. (2007) Avian influenza A (H5N1) infection: Targets and strategies for chemotherapeutic intervention, Trends Pharmacol. Sci., **28,** 280–285.

[10] F. Hayden. (2009) Developing new antiviral agents for influenza treatment: What does the future hold? Clin. Infect. Dis., **48,** S3–S13.

[11] M. Schotsaert, M. De Filette, W. Fiers and X. Saelens. (2009) Universal M2 ectodomain-based influenza A vaccines: Preclinical and clinical developments, Expert Rev. Vaccines, **8,** 499–508.

[12] Influenza virus resources, (2009) http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/multiple.cgi.

[13] SPSS Inc., (2002) SigmaPlot for Windows, Version 8.02.

[14] G. Wu. (1996) Fit fluctuating blood drug concentration: A beginner's first note, Pharmacol. Res., **33,** 379–383.

[15] K. Yamaoka, T. Nakagawa and T. Uno. (1978) Application of Akaike's information criterion (AIC) in the evaluation of linear pharmacoketic equations, J. Pharmacokinet. Biopharm., **6**, 165–175.

[16] G. Wu, P. Cossettini and M. Furlanut. (1996) Prediction of blood cyclosporine concentrations in haematological patients with multidrug resistance by one-, two- and three-compartment models using Bayesian and nonlinear least squares methods, Pharmacol. Res., **34**, 47–57.

[17] G. Wu and S. Yan. (2006) Timing of mutation in hemagglutinins from influenza A virus by means of amino-acid distribution rank and fast Fourier transform, Protein Pept. Lett., **13**, 143–148.

[18] G. Wu and S. Yan. (2005) Timing of mutation in hemagglutinins from influenza A virus by means of unpredictable portion of amino-acid pair and fast Fourier transform, Biochem. Biophys. Res. Commun., **333,** 70–78.

[19] G. Wu and 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.

[20] S. Yan and G. Wu. (2009) Prediction of mutation position, mutated amino acid and timing in hemagglutinins from North America H1 influenza A virus, J. Biomed. Sci. Eng, **2,** 190–199.

J. Biomedical Science and Engineering, 2009, 2, 594-605 **JBiSE**
doi: 10.4236/jbise.2009.28086 Published Online December 2009 (http://www.SciRP.org/journal/jbise/).

Water activity and glass transition temperatures of disaccharide based buffers for desiccation preservation of biologics

Justin Reis[1], Ranjan Sitaula[2], Sankha Bhowmick[1,2,3]

[1]Department of Mechanical Engineering, University of Massachusetts Dartmouth, Massachusetts, USA;
[2]Biomedical Engineering and Biotechnology Program, University of Massachusetts Dartmouth, Massachusetts, USA;
[3]285 Old Westport Road Room # Textile 210 N. Dartmouth, MA 02747, USA.
Email: sbhowmick@umassd.edu

Received 22 August 2009, revised 27 September 2009; accepted 28 September 2009.

ABSTRACT

Studying the thermophysical properties of disaccharide based ternary solutions are gaining increasing importance because of their role as excepients in preservation protocols for biologics in general and mammalian cells in particular. Preservation strategies involve not only cryopreservation, but novel approaches like room temperature vitrification and lyophilization. In this study we investigate the water activity and glass transition temperature of citrate and tris buffers (widely used in the gamete preservation industry) with trehalose or sucrose after partial desiccation. After obtaining the water activity (a_w) through equilibration at different relative humidity environments, we measured the glass transition temperature (T_g) of these partially desiccated solutions using a differential scanning calorimetry (DSC). The experimental data was used in conjunction with the Gordon-Taylor equation to obtain 3-D contours of T_g as a function of water content and relative salt/sugar concentration. Results indicate that the glass transition behavior is a strong function of the excepient combination. Overall, that trehalose solutions yielded larger values for T_g than sucrose counterparts at low moisture contents in combination with the same buffer. We also saw that citrate solutions yielded larger glass transitions than their tris counterparts. Based on these results, a trehalose-citrate mixture can be picked as the preferred composition for storage applications. The 3-D contours which show a wide variation in slope depending on the salt-sugar concentration constitute important information for the desiccation preservation of biologics.

Keywords: Trehalose; Sucrose; TRIS; Citrate; DSC; Glass; Transition; Tenperature

1. INTRODUCTION

Desiccation preservation offers an attractive alternative to cryopreservation for the long term storage of mammalian cells and gametes. While cryopreservation has a stringent requirement of storage in liquid nitrogen at a temperature in the vicinity of-196°C, desiccation preservation offers the ability to store cells at or near ambient conditions. At the same time it eliminates the usage of toxic cryoprotectants such as glycerol and DMSO which require removal upon returning cells to ambient temperatures, severely affecting cell survival in the process [1].

One of the hypotheses behind the mechanism of desiccation preservation is the formation of glassy structure, a highly viscous state that minimizes molecular mobility of the matrix thereby suspending metabolic activities in the cells. Sugars, particularly disaccharides, have been effective in imparting cellular protection in the desiccated state. A number of studies have demonstrated the ability of different sugars such as trehalose, sucrose, raffinose and maltose to sustain a stable glassy state at low moisture content [2,3,4,5,6,8]. Such sugars form glasses at ambient temperature, thereby reducing molecular mobility and allowing a prolonged stable storage of biomaterials and cellular components [3,8,9,10,11].

The survival of mammalian cells in vitro requires a buffer or culture media generally consisting of various salt mixtures. In our study we chose to study ternary sugar-salt-water solutions. The interactions of ternary solutions can often be extremely difficult to predict without proper experimental studies of their thermophysics [6]. These interactions can produce results that may vary significantly even from similar studies of binary solutions [1,12].

Two key thermophysical parameters that will determine a desiccation preservation protocol include water activity (a_w) and glass transition temperature (T_g) [6,8,9,13,14]. Water activity (a_w) is defined as the ratio of the vapor pressure of water in a material (p) to the vapor pressure of pure water (po) at the same temperature [15]. It is an equilibrium state that is greatly responsible for a solution's ability to participate in physical, chemical and microbiological reactions [2,

16]. The glassy state is a non-equilibrium state at which substances exhibit an amorphous glass structure. Glass transition (T_g) is the temperature at which amorphous solids transition from solid to a less viscous state. The glass transition temperature is a function of the solution constituents and the moisture content, as well as a function of the water activity of the storage condition [17].

The objective of the current experimental study was to investigate the effect of water activity (given by the equilibrium relative humidity of the storage environment for room temperature conditions) on moisture contents and the subsequent effect of moisture content on T_g of various sugar-buffer-water ternary system. The particular buffers chosen for this study were the Tris and Citrate buffers whose composition can be seen in **Table 1**. These buffers are widely used bovine sperm extenders under a wide range of temperatures [18,19,20,21,22]. These buffers have an excellent buffering characteristics for biochemical studies, are non-toxic to living cells, and effective for maintaining osmotic pressure in cells. Trehalose and sucrose were chosen as the excipient sugars owing to their superior glass forming ability and their effective role in desiccation preservation which have been well documented in the studies of various biologics [4,5,8,11,23,24, 25]. The goal of our study was to create water activity tables for different concentrations of the sugars in each of the buffers. The T_g of samples were then plotted in as a 3D surface plot as a function of sugar concentration as well as moisture content. These 3D plots are extremely important for references in future work in determining which solutions will produce glass transitions at appropriate temperature levels.

2. MATERIALS AND METHODS

2.1. Sample Preparation

Trehalose dihydrate (Sigma Aldrich assay > 99%) and sucrose (Sigma Aldrich assay > 99.5%) were both purchased from Sigma Aldrich. The tris and citrate buffers were obtained from ABS global, Deforest, WI in concentrated forms and diluted with distilled water to 1X concentrations, which corresponds to an isotonic solution of 325 mosm. The composition of the tris and citrate buffers are presented in **Table 1**. Molar calculations were carried out to determine the appropriate quantity of trehalose or sucrose to be added to each individual volume of the buffer. Vials were then mixed thoroughly to ensure homogeneity. The range of each sugar in combination with respect to the buffer for the tris buffer were 0.713, 1.426, 2.139, 2.85, 5.705, 11.41 g sugar/g tris while solutions utilizing the citrate buffer used the range of 1.485, 2.97, 4.455, 5.94, 11.72, 23.76 g sugar/g citrate.

2.2. Generation of Humidity Environment and Drying Kinetics Curves

Stable relative humidity (RH) environments were generated by equilibrating samples in humidity boxes at room temperature (20°C). Humidity boxes consisted of a sealed plastic food container with a chosen desiccant inside, that was placed inside larger stackable desiccation cabinets (Sanplatec Polystyrene Mini Desiccator, Osaka, Japan) [26,27]. Our chosen desiccants were supersaturated solutions of magnesium nitrate, potassium acetate, lithium chloride, and lithium bromide that provided us with 53%, 21%, 11%, and 6% RH respectively. Drierite salts (Drierite Aldrich Chemical Company, St Louis, MO) was used to obtain 1.5% RH environment. A digital hygrometer (Oakton Thermohygrometer, Vernon Hills, IL) was used to determine the equilibrium RH values generated by the various desiccants. Dry samples (0% RH) were obtained by baking in a natural convective drying oven (Quincy Labs Model 10 Lab Oven, Chicago, IL) at 75°C for a minimum of 14 days or till no detectable variation in weight was observed. A 30μL volume of sample solution was carefully placed in a standard aluminum Differential Scanning Calorimeter (DSC) pan from TA Instruments (New Castle, DE). Pan weight and the sample weight measured using a digital scale (Mettler Toledo AB265S FACT, Columbus, OH) and recorded for gravimetric analysis and for the DSC experiment to determine the glass transition temperature (T_g). Pans were then carefully transported with tweezers to the appropriate relative humidity (RH) box where they were allowed to equilibrate. The samples were weighed periodically after they had been placed into the humidity box. Based on these weight measurements, a drying kinetics chart was generated. These charts were used to ascertain the time at which the samples had reached equilibrium moisture content.

2.3. DSC Experiments

After equilibration, the samples in the DSC pans were promptly hermetically crimped and sealed using a crimp from TA instruments to reduce any exposure to the room RH conditions. Samples were then ready for appropriate DSC experiments.

Table 1. Shows a breakdown of components of both the particular tris and citrate buffers which were used in this study. gm %=grams/100mL water.

Tris Buffer	Citrate buffer
2.42 gm % tris (hydroxymethyl aminomethane) 1.38 gm % citric acid mono-hydrate 1.0 gm % fructose	2.12 gm % sodium citrate dihydrate 0.183 gm % citric acid monohydrate

Table 2. Provides a generalized walkthrough of each portion of a DSC run. Each step is given a general description and the explanation for its use is found in the same row. Experiments were taken to at least 30˚C above the expected glass transition temperature while also considering degradation of samples.

Step #	Function	Description
1	low temperature equilibration	The low temperature equilibration is used to view for any crystallization. It also provides a constant starting point for each of the runs to begin for consistency.
2	First Heating Cycle	The purpose of this heating run is to erase any thermal history of the sample. During this run we want release any of the non-equilibrium properties of the sample such as a buildup of entropy and enthalpy which occurs due to the non equilibrium glassy state.
3	Holding Isothermal	The isothermal run is to equilibrate the sample at a temperature above the glass temperature. This makes sure that all we will have a sample in the equilibrium for our run.
4	Cooling	Now we need to cool our sample back down to our initial baseline. We have erased all the thermal history of the sample and are now ready to begin our actual experiment.
5	Second Heating Cycle	It is during this heating cycle where we will be able to determine our glass transition temperature. This is the cycle which we analyze and is the one we are interested in.
6	Cooling to ambient	This is merely to return the sample to ambient conditions where it can be safely replaced into the auto sampler.
	Note:	All heating and cooling runs were performed at a rate of 5 degrees Celsius per minute from -40˚C to 180˚C except for final cooling to ambient which was performed at 15 degrees Celsius per minute to ambient 25˚C.

Table 3. Shows the k values used in the gordon taylor equation when it was used for solutions containing moisture. k values were dependent upon sugar concentration in the buffer as well as the combination of sugar-buffer being modeled.

Table of k	Trehalose-Tris solutions	Trehalose-Citrate solutions	Sucrose-Tris solutions	Sucrose-Citrate solutions
Lowest sugar Concentration	k=1.4	k=0.37	k=0.5	k=0.57
Largest sugar concentration	k=0.17	k=0.34	k=0.35	k=0.4

A typical DSC run with a heat-cool-heat cycle is show-n in **Table 2**. All experiments for evaluating T_g were performed using a Q1000 DSC (TA Instruments, New Castle, DE) which is also equipped with a refrigerated cooling system (RCS). High purity nitrogen gas was used to purge at a flow rate of 50 mL/min for each run to ensure an inert experimental environment. Sample pans were placed in the auto sampler alongside a reference pan of known weight for comparison during the actual running of the DSC. After the DSC run had been concluded, the TA Universal Analysis software was used to analyze the graph of heat flux as a function of temperature. The glass transition was then located on the graph and calculated using the T_g software function.

2.4. Data Analysis

Initially large quantity of our sugar buffer concentration range samples were baked to determine a wet to dry weight ratio by weighing samples before entering the oven and then again after being baked at 75°C for at least 2 weeks. This ratio would then be multiplied to the initial weights of other samples prior to entering equilibration in humidity chambers in order to provide a weight for the sample if all moisture were removed which is required for the calculation of Dry Basis Moisture Content (DBMC). DBMC is a measure of residual moisture in samples in relation to their dry weight which is calculated to be void of moisture.

$$\%\ \text{Dry basis moisture content (DBMC)} = \frac{W_E - W_B}{W_B} * 100 \quad (1)$$

where W_E is the equilibrated weight and W_B is the baked weight.

All experimental T_g's were plotted as a function of DBMC to see the plasticizing effect of moisture on our ternary solutions. All experiments consisted of at least 3 repeats (n=3). The error bars in the figures represented the standard deviations of the repeats. The statistical significance of the experiment data were evaluated using the analysis of variance. Moisture contents as well as T_g's were tested for significance using Microsoft Excel's ANOVA Single Factor variance test. Statistical significance was assessed as $p<0.05$.

2.5. Modeling of T_g

Gordon and Taylor first developed a model for the prediction of glass transition in 1952 in their study of synthetic rubbers based upon individual components contribution to the glass transition of the overall homogenous uniformly packed mixture [10]. This **Eq.1** was used

for modeling our desiccated ternary solutions.

$$T_g = \frac{w_1 * T_{g1} + k w_2 * T_{g2}}{w_1 + k w_2} \quad (2)$$

where, w_1 represents the weight fraction and the subscript 2 designates the component with larger T_g. k is a model specific parameter. For baked samples (without any moisture) the value of k was given as the ratio of the smaller T_g over that of the larger T_g (k= T_{g1}/T_{g2}) commonly referred to as the Fox equation. However, samples containing moisture required the determination of a different k value than the dried samples. This was accomplished by using the T_g of water (136K) as T_{g1} in **Eq.2** and the baked salt-sugar mixture as T_{g2}. Best fit analysis using a minimization function for percent difference of analytical and experimental data was used in order to determine the most accurate value for k. All the values for k were then used for the creation of 3D plots.

3. RESULTS

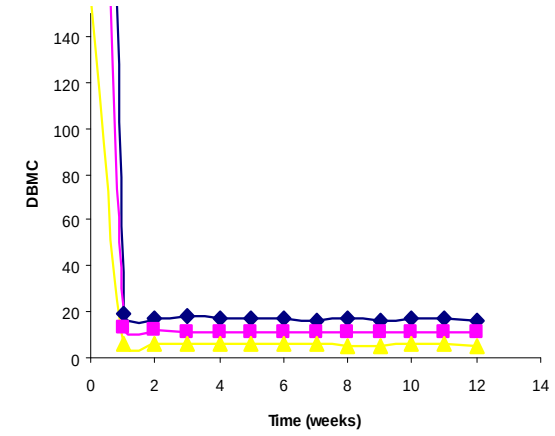
3.1. Drying Kinetics of Solutions

Figure 1 is a representative plot of the drying kinetics of different weight fraction sugar-buffer solution samples equilibrated at room temperature. Trehalose-tris solutions dried in a 7% RH environment and measured on a weekly basis for 12 weeks. **Figure 1** shows that most of the drying takes place within the first week of storage. The samples attain almost constant moisture contents after a period of three weeks. **Figure 1(b)** suggests a lower trehalose concentration resulted in a greater retention of moisture in the equilibrated state. While the equilibrium value of DBMC averaged 16.46% for samples with a trehalose-tris ratio of 0.713 g trehalose/ g tris, the corresponding value was 5.43% for the 11.41 g trehalose/ g tris concentration with $p<0.05$ between the sets. Similar results demonstrating a lower DBMC for higher sugar content solutions were observed for all buffer- sugar combinations in this study.

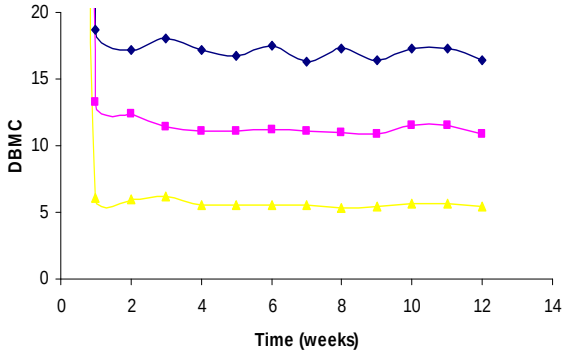
3.2. Effect of Sugar on the T_g of Baked Samples

3.2.1. Effect of Trehalose Concentration on Baked Samples T_g

Figure 2 shows the plot of T_g as a function of trehalose concentration. The trend shows that both trehalose- buffer solutions T_g asymptotically approach approximately 105˚C upon increasing trehalose concentration. The trehalose-tris samples increased their T_g too from a low starting point due to tris' low T_g value. **Figure 2** also shows that for trehalose-citrate samples fell towards the 105˚C value due to the elevated T_g of citrate. Trehalose tris buffer at 0.713g trehalose/g tris concentration showed



(a)



(b)

Figure 1. (a) Shows the drying kinetics of trehalose tris solutions in our 7% humidity environment. The blue diamonds (♦) represent the drying kinetics of 0.713g trehalose/g tris, the pink square (■) represents 2.85g trehalose/g tris, and the yellow triangle (▲) represents 11.41g trehalose/g tris. The second graph is a zoomed view showing the equilibration of samples over time.

an average T_g of approximately 30˚C. As trehalose con tent was then increased to 11.41g trehalose/ g tris the average glass transition increased to around 106˚ C. When using the citrate buffer the lowest trehalose concentration of 1.485 g trehalose/g citrate produced an average glass transition temperature of approximately 125˚C. When our trehalose mass ratio was increased to 23.76g trehalose/g citrate an average T_g of approximately 102˚C was observed. During the heating and cooling of samples no crystallization was present in any of the thermographs irrespective of the salt/sugar mixture content. The trendlines in the figure, which were fitted using the Gordon Taylor model, show that the values for T_g assimilate themselves with the majority mass fraction component of the solution. Samples with low sugar concentrations assimilated T_g's with the buffer involved (tris T_g ≈28.6°C citrate T_g ≈130°C) and move towards that of trehalose (T_g≈115°C).

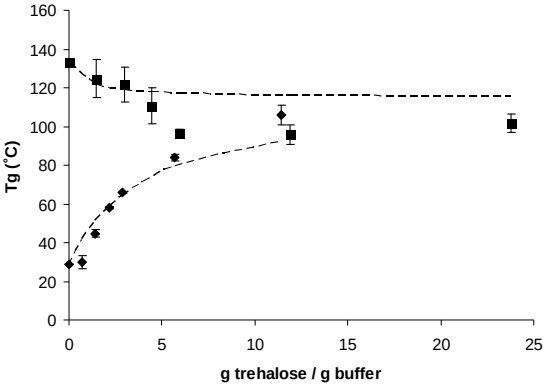


Figure 2. Shows glass transition as a function of trehalose concentration. The square blocks (■) represent the trehalose with citrate buffer while the diamond symbol (♦) show trehalose combined with tris buffer. The figure also depicts the Gordon Taylor depicted by the black dashed lines.

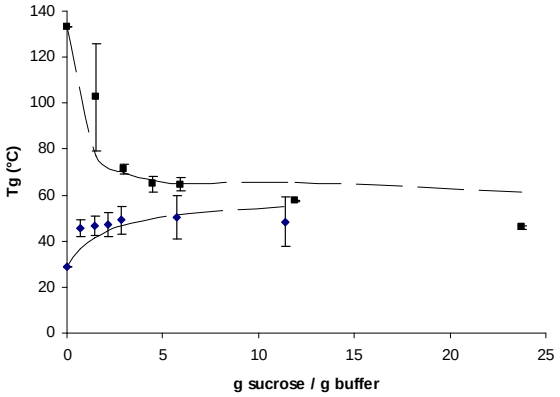


Figure 3. Shows glass transition as a function of sucrose concentration. The diamond symbol (♦) represents the sucrose with tris buffer while the square blocks (■) show sucrose combined with citrate buffer. Gordon Taylor modeling is depicted by the black dashed lines.

3.2.2. Effect of Sucrose Content on Baked Samples T_g

Figure 3 suggests the T_g of sucrose glasses as a function of sucrose concentration. Sucrose glasses demonstrated a similar asymptotic behavior with increasing sucrose content. Sucrose-tris samples showed an increase in their glass transition as sucrose concentration increased because pure sucrose has a larger T_g than tris. Sucrose-citrate samples showed a decrease in T_g as sucrose concentration was increased as the T_g of sucrose is below that of what we found for citrate. At mass ratio 0.713g sucrose/g tris, the average T_g was 45˚C. When sucrose concentration was increased 11.41 g sucrose/g tris the samples produced and average T_g of 48˚ C. The 1.485 g sucrose/g citrate concentration samples average glass transition temperature was approximately 103˚ C. As our mass ratio of sucrose increased all the way to a concentration of 23.76 g sucrose/g citrate the glass transition fell to approximately 46˚C. The trendlines fitted to the experiment data show that the samples follow the Gordon Taylor model of the two component models. Similar to the trehalose results, T_g assimilates itself with the majority fraction of the samples. The data sets exhibit a trend of approaching a T_g slightly below that of pure sucrose ($T_g \approx 60$˚C) as the concentration of sucrose increases.

3.3. Role of Moisture in Modulating Thermophysical Behavior of Sugar Based Buffers

First, water activity curves were generated from sample weight measurements taken after equilibration under different relative humidity environments used in the calculation of DBMC. These curves allowed us to determine effect of both sugars and buffers to retain moisture at equilibrium, and determine the T_g of the solution. In the corresponding sections we show sample plots of DBMC vs. a_w and T_g vs. DBMC for a high and low sugar-buffer combination. Finally, we show the 3-D surface contour of T_g as a function of moisture and sugar content by using the Gordon Taylor equation.

3.3.1. Effect of Moisture on Trehalose Tris Solutions

Figure 4 shows that over the range of relative humidity environments trehalose tris samples tended to equilibrate to specific moisture contents and then hold this in the range. The lower concentration of 0.713g trehalose / g tris show an initial jump in residual moisture content followed by a leveling in the 0.07 to 0.21 a_w where points are statistically the same ($p>0.05$). We then show a significant increase in moisture from a_w of 0.21 to 0.53 ($p<0.05$). The 11.41 g. trehalose/g. tris concentration solution shows a similar trend however after the initial jump in moisture the next successive four points are statistically the same ($p>0.05$). The figure also shows that residual moisture content was affected by trehalose concentration in the solution. Under similar equilibration environment, solutions containing higher concentrations

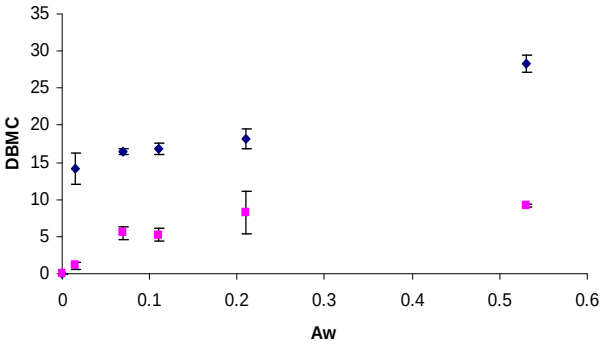
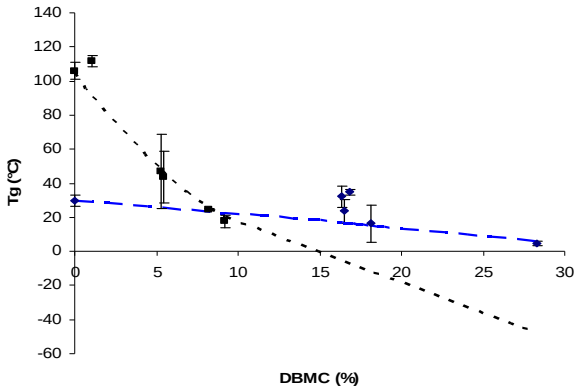


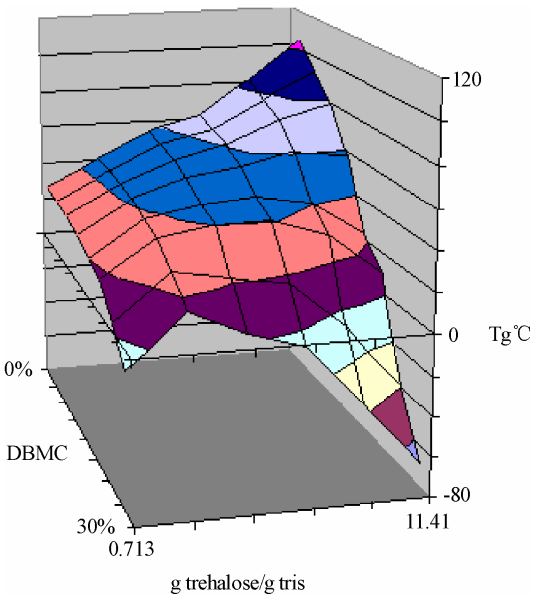
Figure 4. Shows trehalose tris solutions DBMC as a function of water activity. The diamond symbol (♦) represents 0.713g trehalose/g tris, while the square blocks (■) show 11.41 g trehalose/g tris.

of trehalose equilibrated to lower end moisture contents then lower trehalose concentration solutions.

Figure 5(a) shows trehalose-tris solutions exhibit a linear trend of decrease in T_g with increasing residual moisture. It also shows that samples with higher trehalose concentration undergoes a more drastic decrease in glass transition upon gaining moisture as compared to lower trehalose concentrations. The slopes illustrate that



(a)



(b)

Figure 5. (a) Shows glass transition of trehalose tris solutions as a function of dry basis moisture content. The diamond symbol (♦) represents 0.713 g trehalose/g tris and the triangle (■) represents 11.41 g trehalose/g tris. The figure also shows Gordon Taylor modeling of glass transition of trehalose tris solutions as a function of dry basis moisture content. The blue dashed line represents Gordon Taylor modeling of the 0.713g trehalose/g tris while the black dotted line shows Gordon Taylor modeling of 11.41g trehalose/g tris solutions; (b) shows a 3D surface plot of the Gordon Taylor models of our ternary trehalose tris moisture solution using our calculated k values. This surface plot includes all intermediary solution values for k plotted which were not shown in **Figure 5(a)**.

although low concentration of trehalose, such as the 0.713g trehalose/g tris level, has a much lower T_g than the 11.41g trehalose/g tris level, the lower concentration is far less affected by the addition of moisture into the ternary solution.

The Gordon Taylor modeling in **Figure 5(a)** shows that our 0.713g trehalose/g tris concentration prediction misses three of the experimental points; however fall directly upon the other three data points in the set. The three missing points seem to be outliers of the general linear trend of the decreasing glass transition as moisture increases. The largest sugar concentration of 11.41 is modeled quite effectively. The Gordon Taylor modeling was also used to determine values for k for all trehalose-tris concentrations to create the surface plot shown in **Figure 5(b)**. **Figure 5(b)** shows that the rate at which T_g decreases is a function of sugar concentration. The result also indicates the possibility for in intermediate maximum in larger moisture levels.

3.3.2. Effect of Moisture on Trehalose Citrate Solutions

Figure 6 shows equilibrated moisture content in trehalose-citrate solutions as a function of water activity (a_w). The lowest trehalose concentration of 1.485 g trehalose/g citrate seems to continue on a gradual increase as moisture content as a_w increases. We see significant difference between the 0.015 and 0.21 a_w environments ($p<0.05$) followed by a continued increase in the next successive points ($p<0.05$).The largest concentration of 23.76g trehalose/g citrate shows gradual gain before reaching a plateau around 12% DBMC. **Figure 6** also shows that especially at the largest a_w environment there is a large difference in the moisture capacities for the samples ($p<0.05$) where the samples containing larger trehalose concentrations equilibrate to lower DBMC's.

Figure 7(a) shows the predictable decrease in T_g of trehalose-citrate solutions as residual moisture increases. The lowest concentration of 1.485 g trehalose/g citrate shows a average T_g of 122°C at approximately 6% DBMC and it is reduced to 7°C at the largest DBMC of 31%. The 23.76g trehalose/g citrate yielded an average T_g of 98°C at a DBMC of 25% and 20°C at 12% DBMC. The plot also shows that the trehalose - citrate solutions show decreasing linear slopes of -5 indicating that the decrease in glass transition seems to be more affected by the moisture content. The figure also shows larger concentration of trehalose had slightly lower glass transition over the range of moisture content. The possibility of salts precipitating out when these solutions were equilibrated in the larger RH environments is a possibility for the large variation in the 23.76 concentration data at approximately 12% DBMC.

The Gordon Taylor model was then used to create the surface plot shown in **Figure 7(b)**. The plot was created from calculated k values for intermediary solutions and

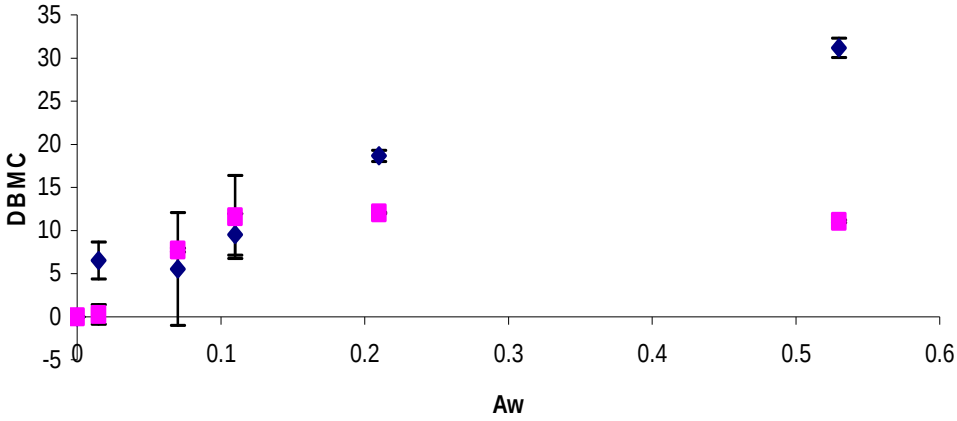
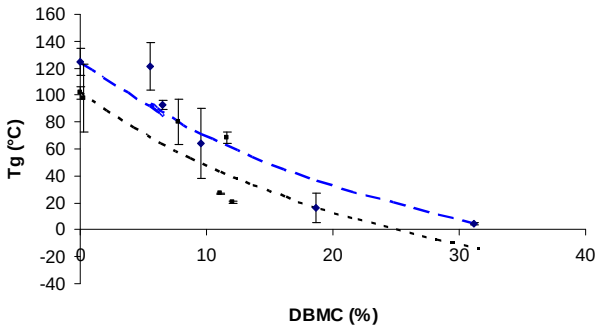
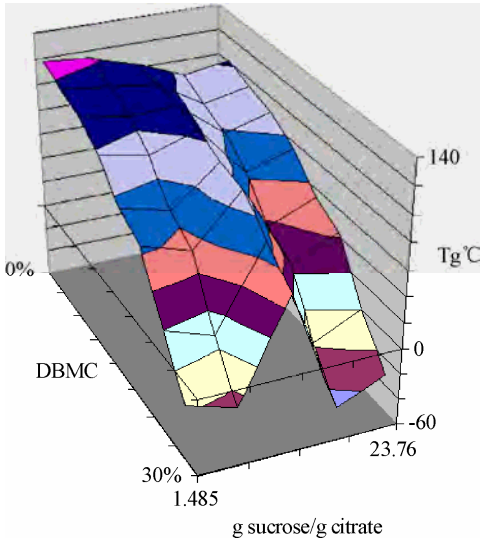


Figure 6. Shows trehalose citrate solutions DBMC as a function of water activity. The diamond symbol (♦) represents 1.485g trehalose/g citrate while the square blocks (■) show 23.76g trehalose/g citrate.



(a)



(b)

Figure 7. (a) Shows glass transition of trehalose citrate solutions as a function of dry basis moisture content. The diamond symbol (♦) represents 1.485g trehalose/g citrate and the square blocks (■) represents 23.76g. trehalose/g. citrate. The figure also shows Gordon Taylor modeling of glass transition of trehalose tris solutions as a function of dry basis moisture content. The blue dashed line represents Gordon Taylor modeling of the 1.485g trehalose/g citrate while the black dotted line shows Gordon Taylor modeling of 23.76g trehalose/g citrate solutions; (b) shows a 3D surface plot of the Gordon Taylor models of our ternary trehalose citrate moisture solution using our calculated k values. This surface plot includes all intermediary solution values for k plotted which were not shown in **Figure 7(a)**.

then the plotting of the Gordon Taylor equations in a 3D plot. We see that the profile of the surface is fairly flat due to the similarity T_gs of trehalose and citrate.

3.3.3. Effect of Moisture on Sucrose Tris Solutions

Figure 8 shows the sucrose-tris solutions equilibrated to different DBMC's as a function of a_w. Both concentration samples show a gain in moisture at 0.015 a_w environment. From this point forward we see that both sam-ples show a general trend of decrease in equilibrated DBMC ($p<0.05$ for both concentrations) before showing an increase at our highest a_w. **Figure 8** also shows that the larger sugar concentration solutions equilibrate to lower end moisture contents across the entire a_w range. While this trend is more prevalent in the 0.21 and 0.53 a_w environments; however even in the lower RH environments samples with lower sugar concentrations still equilibrated to larger average end moisture contents while being just barely significantly different at the 0.21 a_w environment ($p\approx0.05$).

Figure 9(a) shows the sucrose-tris samples' rapid decreases in T_g with even low levels of moisture. Our lowest sucrose concentration 0.713g sucrose/g tris yielded average T_g's of approximately 20°C in the vicinity of 10% DBMC and -4°C at a DBMC of 21%. Larger concentration of sucrose samples yielded slightly lower T_g's than the smaller sucrose concentration solution. The 11.41 g sucrose/g tris solutions experimental values fall above and below our predicted values in the larger

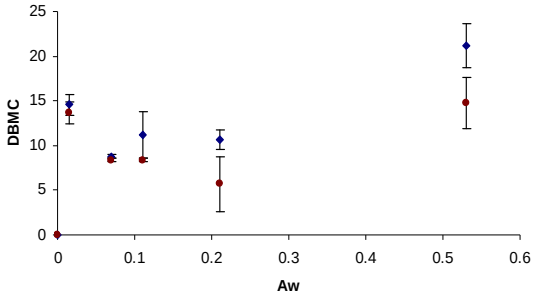


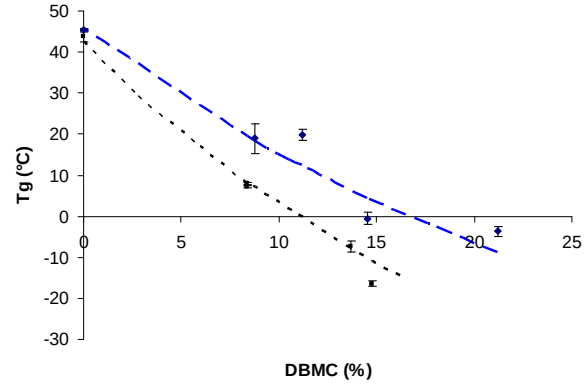
Figure 8. Shows sucrose tris solutions DBMC as a function of water activity. The diamond symbol (♦) represents 0.713g sucrose/g tris while the circles (●) show 11.41g sucrose / g tris.

DBMC's but effectively show the trend. The Gordon Taylor models were then used to create the surface plot shown in **Figure 9(b)**. The surface plot shows that the T_g for the range of sucrose tris concentrations varies far more drastically as a function of moisture content as opposed to sucrose concentration. At low moisture contents we see that T_g is almost invariant between the concentrations of sucrose.

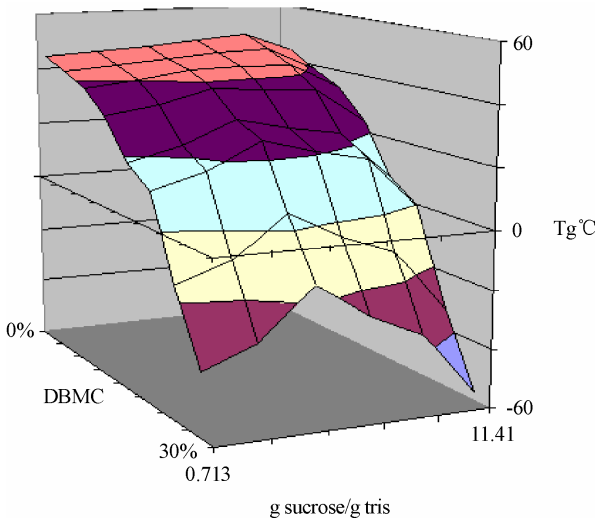
3.3.4. Moistures Effect on Sucrose Citrate Solutions

Figure 10 shows sucrose-ctrate samples equilibrated DBMC at the different relative humidity environments. We see that initially both the 1.485g sucrose/g citrate and the 23.76g sucrose/ g citrate gain significantly different moisture in the 1.5% RH environment ($p<0.05$). The next two a_w levels both solutions show a plateau ($p>0.05$ for both 23.76g sucrose/g citrate and 1.485g sucrose/g citrate). From this point the lower concentration shows the expected gain of moisture as a function of increasing relative humidity ($p>0.05$ between successive points), however the 23.76 g sucrose/ g citrate solution shows a decrease in it moisture contents ($p>0.05$ between successive points).

In **Figure 11(a)** we see that sucrose-citrate solutions decrease T_g as their residual moisture content increases. The 1.485 g sucrose/ g citrate solution yielded an average T_g of 64°C at 5% DBMC and 46°C at 18% DBMC. The 23.76 g sucrose/ g citrate yielded an average T_g of 37° at 5% DBMC and 8°C at 16% DBMC. **Figure 11(a)** also shows that T_g is affected by sucrose content. Lower concentrations of sucrose yielded higher glass T_g's when equilibrated to the same level as their higher concentration counterparts. The Gordon Taylor model slightly overestimates the glass transition of solutions. Baked solutions at this concentration had fairly large error bars showing large variation in T_g. This could cause an overestimation of T_g since we used this value for T_g 2 in the Gordon Taylor equation. Should T_g 1 be lower it would-flatten the entire curve and possibly hit all experimental points as well providing a better fit for our baked prediction as well.



(a)



(b)

Figure 9. (a) Shows glass transition of sucrose tris solutions as a function of dry basis moisture content. The diamond symbol (♦) represents 0.713 g sucrose/g tris and the square blocks (■) represent 11.41g sucrose/g tris. We can see here that as DBMC increases the trend is that our glass transition decreases in respect to a DBMC of 0%. The trend seems to be fairly flat over the general area of DBMC's we have mapped. The figure also shows Gordon Taylor modeling of glass transition of trehalose tris solutions as a function of dry basis moisture content. The blue dashed line represents Gordon Taylor modeling of the 0.713g sucrose/g tris while the black dotted line shows Gordon Taylor modeling of 11.41 g sucrose / g tris solutions; (b) shows a 3D surface plot of the Gordon Taylor models of our ternary sucrose tris moisture solution using our calculated k values. This surface plot includes all intermediary solution values for k plotted which were not shown in **Figure 9(a)**.

The Gordon Taylor models were then used to create the surface plot shown in **Figure 11(b)** from calculated k values for all solutions. The surface plot shows that T_g decreases at a similar rate for increasing moisture content regardless of sucrose concentration. The surface is

similar to a plane in which all solutions undergo the same rate of decreasing T_g with residual moisture with the difference in levels being a function of baked solutions T_g.

4. DISCUSSION

The current thermophysical study of the ternary solutions was driven by a larger goal of obtaining optimal excipient conditions for desiccation preservation of mammalian cells. Majority of studies showing the sugar stabilization effects are derived from the food preservation or pharmaceutical literature [28,29,30,31,32]. Disaccharides, particularly trehalose and sucrose, have shown to be important in the preservation of cells and biologics [2,3,4,5,6,7,9,10,11]. These sugars exhibit larger glass transition temperatures and possess excellent water replacement abilities. In combination with these sugars, tris and citrate salts were chosen for being industry standards for bovine sperm preservation [18,19,20, 33]. The complicated ternary solutions in this study were chosen based upon previous research showing their ability to sustain cellular life. The thermophysical properties of preservation medium were important in understanding which would be applicable as a desiccation medium. Solutions which undergo glass transition at less than ambient are clearly not appropriate since stability of medium is essential.

4.1. Baked Samples

The completely dried samples behaved exactly as the Gordon Taylor (Fox equation) two-component model predicted [12]. The variation in T_g of a given sugar-buffer mixture was a function of the glass transition of each component and its weight fraction. Similar trend for other mixtures have been observed in varying degrees by several studies [6,12].

The T_g of trehalose-tris and Trehalose-citrate mixtures converged to a limit of 105°C as trehalose concentration increased. The limit was approached both from above and below as the tris and citrate buffer were found to have a T_g at approximately 28.6°C and 133°C respectively. Though we would expect T_g to reach that of the pure sugar, extrapolation out to the pure limit may not be accurate for complex materials containing salts as described by Mazzobre *et al* [34]. Since we approach this limit from very different starting points, we assume that the lowered limit for T_g is caused by a common substance found in both biological buffers, in our case the citric acid monohydrate. This commonality between the buffers could be the most logical reason for both solutions to approach a common T_g roughly 10°C below that of pure trehalose. The deviation from the expected limit of the pure substance has been shown in previous studies in the literature. Jeong-Ah Seo and coworkers demonstrated that when different monosaccharides are combined with disaccharides, glass transition deviated from

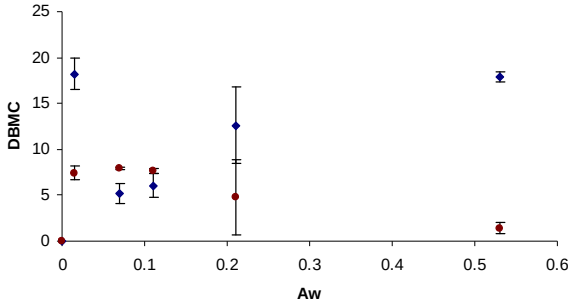
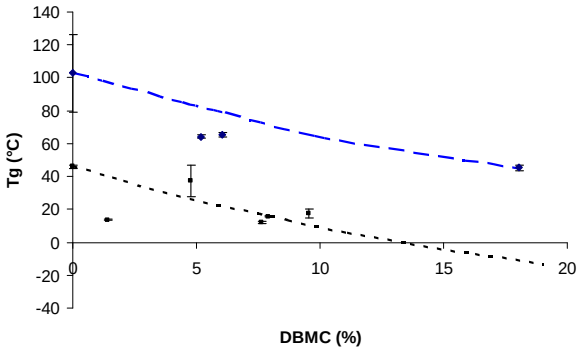
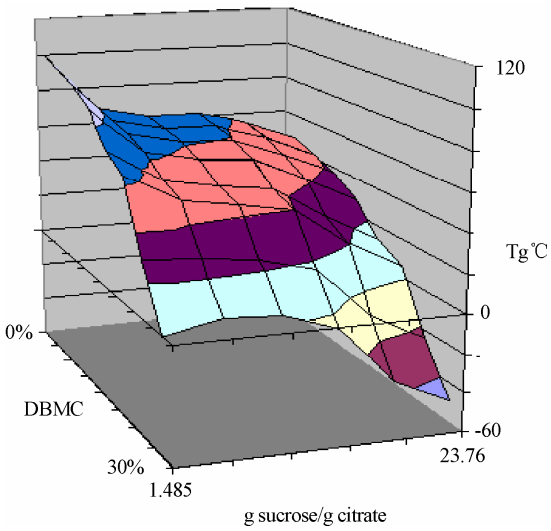


Figure 10. Shows sucrose citrate solutions DBMC as a function of water activity. The diamond symbol (♦) represents 1.485g sucrose/g citrate while the circles (●) show 23.76g sucrose/g citrate.



(a)



(b)

Figure 11. (a) Shows glass transition of sucrose citrate solutions as a function of dry basis moisture content. The diamond symbol (♦) represents 1.485g sucrose / g citrate and the square blocks (■) represent 23.76g sucrose/g citrate; (b) shows a 3D surface plot of the Gordon Taylor models of our ternary sucrose citrate moisture solution using our calculated k values. This surface plot includes all intermediary solution values for k plotted which were not shown in **Figure 11(a)**.

the expected value both on the high and low side of the Gordon Taylor prediction as a result of size and shape of molecules involved [35].

Similar to the trehalose results, the T_g of sucrose-buffer solutions also converged to a common limit at 47°C as the sucrose content increased. Again the limit of the solutions approached both from above and below due to the T_g's of our buffers involved and still they converge upon a common value which falls approximately 10°C below that of pure sucrose ($T_g \approx 60$°C). For similar reason as with trehalose, we again make the argument that since we approach a common limit found below the T_g of the sugar, the structure must be altered by a common component of the buffers, the citric acid monohydrate.

Trehalose based mixtures consistently showed larger T_gs than their sucrose equivalents. The glass transition temperature of trehalose is approximately two times larger than that of sucrose making it far superior in terms of the thermophysical property of glass transition. In comparing tris to citrate from a thermophysical standpoint, citrate clearly dominates with a roughly five times larger T_g than that of tris. However the T_g of the pure buffer is very weak and it is in combination with sugars that the glass transition becomes stronger and more prevalent. Based on Freeze dried results, Kets and coworkers have shown that the citrate was able to increase the glass transition of sucrose [6]. These numbers are slightly larger but comparable to our results. Even though the T_g of tris buffer could not be correlated to any literature value, our experimental results were quite clear and consistent. Hence, from a purely thermophysical standpoint, solutions containing larger fractions of citrate salt produce consistently larger T_g values than their tris counterparts.

4.2. Role of Moisture

4.2.1. Water Activity

Water activity curves are extremely important in this ternary study in order to predict a solution's ability to retain or release moisture under different relative humidity environments. Moisture content is also greatly responsible for a solution's glass transition temperature. Its role as a plasticizer has been shown in many similar studies of wide varieties of solution composition [2,12, 36,37].

The general trend observed from our water activity study is that larger sugar concentrations equilibrate to lower end moisture contents. At the largest a_w values, we consistently see that the lower sugar concentration solutions have a significantly larger DBMC, which is due to the more hygroscopic nature of salts compared to sugars.

Trehalose solution isotherms vary depending on sugar concentration. Solutions containing lower concentrations of trehalose equilibrated to larger end moisture contents. Solutions containing high concentrations of trehalose seem to plateau at constant moisture content after an initial increase in moisture content. The leveling shows the formation of stable trehalose dihydrate which results in the resilience of high trehalose concentration solutions to gain moisture [34]. On the other hand, sucrose water sorption isotherms show that the sucrose solutions contain large amounts of moisture at lower water activity. However the moisture content decreases when exposed to larger a_w environments. This is due to the fact that anhydrous sucrose crystallizes above this water activity [34]. Sucrose-tris solutions show more of a leveling than a decrease, possibly due to the fact that there is a lower sucrose concentration in relation to the buffer for our tris solutions. Mazzobre and coworkers also showed a very similar trend in there isotherms for sucrose-potassium chloride solutions. Their trend shows that as a_w increased moisture content decreased until levels which fall above our range of study. A comparison of our sucrose isotherms to trehalose isotherms show that sucrose tends to equilibrate to lower DBMC's in the upper a_w range whereas the reverse is true for lower a_w range.

On the other hand, it is difficult to dr a_w distinctions between tris and citrate salts in terms of water activity. Depending on the specific water activity examined it seems as if each sugar buffer solution show something slightly different.

4.2.2. Glass Transition

In comparing trehalose to sucrose one very important trend arises about there rate at which glass transition decreases as a function of moisture. Crowe and coworkers showed in their Stabilization of Dry Mammalian Cells study that at upon the gain of moisture sucrose and trehalose begin to assimilate T_g's. They show that at approximately 10% DBMC the T_g of trehalose and sucrose seem to be approximately 10°C different compared to dry states where trehalose has a T_g roughly two times larger than sucrose. This shows that the rate of change at which sucrose's glass transition decreases as a function of moisture content is less than that of trehalose, a trend which our data replicates. This was determined by fitting data points with a linear regression and comparing the slopes of the 3D surface plots. The surface plots show that in order to reach a similar T_g at 10% DBMC the trehalose graph shows a sharper decrease in T_g as a function of moisture content. When we compare solutions containing the largest sugar concentrations between sucrose and trehalose utilizing the same buffer the trehalose solutions slope is approximately twice as large as the sucrose samples. This may have far reaching implication in trying to stabilize mammalian cells near room temperature.

A notable difference between trehalose-tris and trehalose-citrate solutions lies on the rate at which T_g decreases as a function of moisture content. Comparing the 3-D contour plots, while sucrose samples exhibit similar rates for both the tris and citrate buffers trehalose does not. While the trehalose-citrate samples show

a rapid decrease in their glass transition upon the arrival of moisture, the trehalose tris-samples are far less affected.

When comparing tris and citrate buffers samples, the major difference is that the rate at which T_g decreases as a function of moisture content varies between the buffers. Comparing the 3D plots 7b and 11b, both sugars in combination with the citrate buffer show fairly constant rates of decreasing glass transition within that particular sugar buffer solutions range of sugar concentration. Tris samples on the other hand show a more varied rate which seems to increase as a function of sugar concentration. While sucrose-tris samples seem to present a very slight increase in the rate at which T_g decreases with increasing moisture, trehalose-tris samples show large variation in their slope of their 3D surface contour. The lowest concentration of trehalose exhibits a slope of approximately -1 in as a function of moisture content while our largest trehalose concentration produces a significantly larger slope of approximately -10 almost identical rates for T_g of trehalose water solutions. This is a clear difference between the tris and citrate buffer.

4.2.3. Modeling of Glass Transition

The use of the Gordon Taylor was chosen for its accuracy and overall simplicity. Other far more complex equations such as Millers equation and the Miller-Fox equation include specific component parameters such as excess thermal expansion coefficient and excess volume coefficients. Our buffer solution in itself is a complicated solution making these coefficients difficult to determine. Shah and Schall compared the Fox, Miller, and Miller Fox's equations ability to predict T_g [12]. When we look at the percent differences of the data which they compare to the experimental data we see that even for the least precise Fox equation the percent differences all fall below 9%. Without any given standard deviations of their experimental data it is difficult to even further comment on deviations between experimental and model data. When we examined the average percent difference for Shah and Schall work, we see that the Miller Fox equation is the most accurate with an average 1.86% percent difference while the least accurate Fox equation falls in at an average 2.98%. The additional 1% average accuracy hardly seems to warrant the usage of the far more complex equation.

5. OVERALL CONCLUSIONS

Increasing sugar concentration allows solutions to equilibrate at lower moisture contents.

Trehalose solutions yielded larger values for T_g at lower moisture contents than their sucrose counterparts.

Citrate solutions yielded larger glass transitions than their tris counterparts.

The rate of change of T_g with moisture content, d T_g /d (moisture content), had very different behavior depending on which sugar was present as the excepient. While sucrose content did not change that behavior, the presence of trehalose had a strong influence-an increasing trehalose concentration caused solutions to increase d T_g /d (moisture content) when compared to lower trehalose solutions.

Based on the current results, a combination of trehalose and citrate would be the preferred composition for storage applications.

REFERENCES

[1] M. Jochem and C. H. Kõrber. (1987) Extended phase diagrams for the ternary solutions H_2O-NaCl-glycerol and H_2O-NaCl-hydroxyethylstarch (HES) determined by DSC. Cryobiology, **24**, 513–536.

[2] A. H. Al-Muhtaseb, W. A. M. McMinn, and T. R. A. Magee (2004) Water sorption isotherms of starch powders Part 1: mathematical description of experimental data. Journal of Food Engineering, **61**, 297–307.

[3] J. H. Crowe, K. H. Nguyen, F. A. Hoekstra and L. M. Crowe. (1996) Is vitrification involved in depression of the phase transition temperature in dry phospholipids? Biochimica et Biophysica Acta, **1280**, 187–196.

[4] I. S. Davis, R. W. Bratton and R. H. Foote. (1993) Livability of bovine spermatozoa at 5, -25, and -85°C in tris-buffered and citrate-buffered yolk glycerol extenders. Journal of Dairy Scence, **46**, 333–336.

[5] Gordon and J. S. Taylor (1952) Ideal co-polymers and the second order transitions of synthetic rubbers. Journal of Applied Chemistry, **2,** 493–500.

[6] E. P. W. Kets, P. J. IJpelaar, F. A. Hoekstra and H. Vromans. (2004) Citrate increases glass transition temperature of vitrified sucrose preparations. Cryobiology, **48**, 46–54.

[7] S. B. Leslie, E. Israeli, B. Lighthart, L. M. Crowe and J. H. Crowe. (1995) Trehalose and sucrose protect both membranes and proteins in intact bacteria during drying. Applied and Environmental Microbiology, **61**, 3592–3597.

[8] T. Chen, S. Bhowmick, A. Sputtek, A. Fowler and M. Toner. (2002) The glass transition temperature of mixtures of trehalose and hydroxyethyl starch. Cryobiology, **44**, 301–306.

[9] L. M. Crowe, D. S. Reid and J. H. Crowe. (1996) Is trehalose special for preserving dry biomaterials? Biophysical Journal, **71**, 2087–2093.

[10] M. E. Elias and A. M. Elias. (1999) Trehalose and water fragile system: properties and glass transition. Journal of Molecular Liquids, **83**, 303–310.

[11] A. Eroglu, M. Russo, R. Bieganski, A. Fowler, S. Cheley, H. Bayley and M. Toner, (2000) Intracellular trehalose improves the survival of cryopreserved mammalian cells. Nature Biotechnology, **18**, 163–167.

[12] B. Shah and C. A. Schall. (2006) Measurement and modeling of the glass transition temperatures of multi-component solutions. Thermochimica Acta, **443**, 78–86.

[13] J. H. Crowe and L. M. Crowe. (2000) Preservation of mammalian cells-learning nature's tricks. Nature Biotechnology, **18**, 145.

[14] C. J. Roberts and F. Franks. (1996) Crystalline and amorphous phases in the binary system water-beta, beta-trehalose. Journal of Chemical Society, **92**, 1337–1343.
[15] H. A. Tajmir-Riahi, M. Naoui and S. Diamantoglou. (1994) DNA-carbohydreate interaction. The effects of mono-and disaccharides on the solution structure of calf thymus DNA. Journal of Biomolecular Structure and Dynamics, **12**, 217–234.
[16] S. L. Shamblin and G. Zografi. (1999) The effects of absorbed water on the properties of amorphous mixtures containing sucrose. Pharmaceutical Research, **16**, 1119–1124.
[17] W. F. Wolkers, H. Oldenhof, M. Alberda and F. A. Hoekstra. (1998) A fourier transform microspectroscopy study of sugar glasses: Application to anhydrobiotic higher plant cells. Biochimica et Biophysica Acta, **1379**, 83–96.
[18] J. H. Crowe, L. M. Crowe, W. E. Wolkers, A. E. Oliver, X. Ma, J. H. Auh, M. Tang, S. Zhu, J. Norris and F. Tablin. (2005) Stabilization of dry mammalian cells: lessons from nature. Integrative and Comparative Biology **45,** 810–820.
[19] R. H. Foote. (2002) The History of Artificial Insemination; Selected Notes and Notables. Journal Animal Scence, **80**, 1–10.
[20] R. H. Foote, I. S. Davis and R. W. Bratton. (1962) Survival of bovine spermatozoa at room temperature in citrate and cornell university and tris extenders containing whole and fractionated coconut milk. Journal of Dairy Science. **45, 1**383–1391.
[21] L. Ijaz, A. G. Hunter and E. F. Graham. (1989) Identification of the capacitating agent for bovine sperm in egg-yolk TEST semen extenders. Journal of Dairy Science, **72**, 2700–2706.
[22] L. S. Taylor and P. York. (1998) Characterization of the phase transitions of trehalose dihydrate on heating and subsequent dehydration. Journal of Pharmaceutical Sciences, **87**, 347–355.
[23] J. H. Crowe, J. F. Carpenter and L. M. Crowe. (1998) The role of vitrification in anhydrobiosis. Annual Review of Physiology, **60**, 73–103.
[24] W. Q. Sun, A. C. Leopold, L. M. Crowe and J. H. Crowe. (1996) Stability of dry liposomes in sugar glasses. Biophysical Journal, **70**, 1769–1776.
[25] F. Sussich, C. Skopec, J. Brady and A. Cesaro. (2001) Reversible dehydration of trehalose and anhydrobiosis: from solution state to an exotic crystal? Carbohydrate Research, **334**, 165–176.
[26] J. R. Green. (2005) Isothermal desiccation and thermophysical properties of trehalose-water mixtures, A thesis in mechanical engineering, University of Masachusetts at Dartmouth.
[27] R. Sitaula and S. Bhowmick. (2006) A study of the thermophysical properties and moisture sorption characteristics of trehalose-PBS glasses. Cryobiology, **52**, 369–385.
[28] T. P. Labuza, A. Kaanane and J. Y. Chen. (1985) Effect of temperature on the moisture sorption isotherms and water activity shift of two dehydrated foods, Journal of Food Science, **50**, 385–391.
[29] A. M. Lammart, S. J. Schmidt and G. A. Day. (1998) Water activity and solubility of trehalose. Food Chemistry, **61**, 139–144.
[30] D. P. Miller, J. J. de Pablo and H. R. Corti. (1997) Thermophysical properties of trehalose and its concentrated aqueous solutions. Pharmaceutical Research, **14**, 578–590.
[31] S. L. Shamblin and G. Zografi. (1998) Enthalpy relaxation in binary amorphous mixtures containing sucrose. Pharmaceutical Research, **15**, 1828–1834.
[32] S. L. Shamblin and G. Zografi. (1999) The effects of absorbed water on the properties of amorphous mixtures containing sucrose. Pharmaceutical Research, **16**, 1119–1124.
[33] R. W. Bratton and R. H. Foote. (1999) Fertility of bull sperm frozen and stored in clarified egg yolk-tris- glycerol extender. Journal of Dairy Science, **82,** 817–821.
[34] M. F. Mazzobre, M. P. Longinotti, H. R. Corti and M. P. Buera. (2001) Effect of salts on the properties of aqueous sugar systems, in relation to biomaterial stabilization. 1. water sorption behavior and ice crystallization/melting. Cryobiology, **43**, 199–210.
[35] J. A. Seo, S. J. Kim, H. J. Kwon, Y. S. Yang, H. K. Kim and Y. H. Hwang. (2006) The glass transition temperatures of sugar mixtures. Carbohydrate Research, **341**, 2516–2520.
[36] J. H. Crowe, L. M. Crowe, A. E. Oliver, N. Tsvetkova, W. Wolkers and F. Tablin. (2001) The trehalose myth revisited: Introduction to a symposium on stabilization of cells in the dry state. Cryobiology, **43**, 89–105.
[37] A. Simperler, A. Kornherr, R. Chopra, W. Jones, W. D. S. Motherwell and G. Zifferer. (2007) The glass transition temperatures of amorphous trehalose-water mixtures and the mobility of water: an experimental and in silico study. Carbohydrate Research, **342**, 1470–1479.

J. Biomedical Science and Engineering, 2009, 2, 606-611 **JBiSE**
doi: 10.4236/jbise.2009.25043 Published Online December 2009 (http://www.SciRP.org/journal/jbise/).

Influence of sampling on face measuring system based on composite structured light

Yang Shen[1,2], Hai-Rong Zheng[1,2*]

[1]Institute of Biomedical and Health Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China;
[2]Key Laboratory of Biomedical Informatics and Health Engineering, Chinese Academy of Sciences, Shenzhen, China.
Email: hr.zheng@siat.ac.cn

Received 7 August 2009; revised 4 September 2009; accepted 7 September 2009.

ABSTRACT

Human face can be rebuilt to a three-dimensional (3 D) digital profile based on an optical 3D sensing system named Composite Fourier-Transform Profilometry (CFTP) where a composite structured light will be used. To study the sampling effect during the digitization process in practical CFTP, the pectinate function and convolution theorem were introduced to discuss the potential phase errors caused by sampling the composite pattern along two orthogonal directions. The selecting criterions of sampling frequencies are derived and the results indicate that to avoid spectral aliasing, the sampling frequency along the phrase variation direction must be at least four times as the baseband and along the orthogonal direction it must be at least three times as the larger frequency of the two carrier frequencies. The practical experiment of a model face reconstruction verified the theories.

Keywords: Optical 3D Sensing; Composite Structured Light; Sampling; Spectral Aliasing

1. INTRODUCTION

Structured-light illumination is commonly used as an active optical 3D sensing technique for automated inspection and measuring surface topologies. The Fourier-Transform Profilometry (FTP) [1,2] is one of the classical 3D acquisition methods and it has been widely investigated [3,4,5] because of its advantages of obtaining data from only one frame and analyzing spectrum in whole-field as well as high resolution. Recently, an improved FTP method called Composite Fourier-Transform Profilometry (CFTP) was introduced [6,7]. This novel method prevents spectral aliasing between zero-frequency and baseband by using only one grating namely Composite Pattern (CP) that generated by integrating multiframe ordinary patterns, so that it allows for real-time implementations [8].

However, the data in both CFTP and FTP are digitally sampled to discrete signals during the digitization process in practice. The discrete images have periodical Fourier spectrum, and the fundamental spectrum including the useful information would be overlapped by the adjacent periodical weight [9]. Furthermore, the CP in CFTP is much more complexity than traditional sine grating, another kind of spectrum aliasing would be brought in. In this instance, choosing a proper sampling frequency is very important for the precise survey.

To study the influence caused by sampling, the knowledge of pectinate function and convolution theorem was employed in this article and the suggestion that how to select proper sampling frequencies was given. The ex periment verified the theories, and a beautiful 3D digital profile of a model face was acquired.

2. METHODS

2.1. CFTP Theory

A Composite Pattern (CP) in CFTP is generated as shown in **Figure 1**. The multiframe sine patterns to be modulated are as follows.

$$G_n = c + \cos(2\pi f_\phi y + \pi n) \quad (1)$$

where a constant c is used to offset G_n to be non-negative values, and f_ϕ is the baseband, y represents the

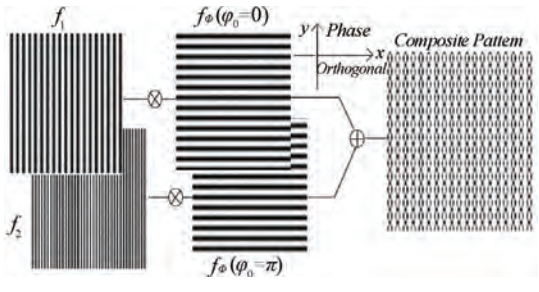


Figure 1. A composite pattern formed by simple strips.

depth distortion (i.e., phase dimension) direction, n represents the phase-shift index from 0 to 1. These signal

patterns are multiplied with different carrier frequencies respectively along the orthogonal direction. Accumulate all channels such that

$$I(x,y) = a + b\{[c + \cos(2\pi f_\phi y)]\cos(2\pi f_1 x) + [c + \cos(2\pi f_\phi y + \pi)]\cos(2\pi f_2 x)\} \quad (2)$$

where f_1 and f_2 are carrier frequencies along the orthogonal direction x, and the projection constants a and b are used to make sure the projection intensity of CP falls into the range of $I(x,y)$ [6,7]. Ideally, the reflected image of the specimen surface can be counted as follows

$$P(x,y) = ar(x,y) + br(x,y)\{[c + \cos(2\pi f_\phi y + \varphi(x,y))]\cos(2\pi f_1 x) + [c + \cos(2\pi f_\phi y + \varphi(x,y) + \pi)]\cos(2\pi f_2 x)\} \quad (3)$$

where $r(x,y)$ and $\varphi(x,y)$ represent the albedo radiation and distorted phase respectively. By means of 2D Fourier-transform and predigestion, the expression (3) will be translated into (4) shown as follows, where $F(\xi,\eta)$, $A(\xi,\eta)$, $B(\xi,\eta)$ and $\psi(\xi,\eta)$ represent the two-dimension Fourier spectrum of $P(x,y)$, $ar(x,y)$, $bcr(x,y)$ and $\frac{1}{2}br(x,y)\exp[j\varphi(x,y)]$ respectively.

$$\begin{aligned} F(\xi,\eta) = A(\xi,\eta) + (1/2)\{ & \\ & [B(\xi - f_1,\eta) + \psi(\xi - f_1,\eta - f_\phi) + \psi^*(\xi - f_1,\eta + f_\phi)] + \\ & [B(\xi + f_1,\eta) + \psi(\xi + f_1,\eta - f_\phi) + \psi^*(\xi + f_1,\eta + f_\phi)] + \\ & [B(\xi - f_2,\eta) - \psi(\xi - f_2,\eta - f_\phi) - \psi^*(\xi - f_2,\eta + f_\phi)] + \\ & [B(\xi + f_2,\eta) - \psi(\xi + f_2,\eta - f_\phi) - \psi^*(\xi + f_2,\eta + f_\phi)] \ \} \end{aligned} \quad (4)$$

Expression (4) suggests that the two carrier frequencies are evenly distributed and are separated by spectral frequency of background reflectance. Therefore, a smooth and flat background had better be selected to minimize the influence to the carrier spectrums. The distorted image is processed as a set of 1-D signal vectors by band-pass filters to separate out each channel. Cutoff frequencies of each band represent the individual patterns like that in traditional π Phase Shift FTP and are used to retrieve the depth of the measured object based on the traditional π Phase Shift FTP method [7] as follows:

$$h(x,y) = \varphi(x,y)L_0 / 2\pi f_\phi d \quad (5)$$

where d and L_0 are experimental setup parameters, $h(x,y)$ represents the reconstructed height.

2.2. Influence of Sampling on CFTP

Expression (3) indicates the continuous image, but in practical experiment it will be digitally sampled to discrete signals by projector and camera, and the discrete distorted pattern $S(x,y)$ is captured as

$$\begin{aligned} S(x,y) &= P(x,y)comb(x/\Delta x, y/\Delta y) \\ &= P(x,y)comb(x/\Delta x)comb(y/\Delta y) \end{aligned} \quad (6)$$

where $comb(x/\Delta x, y/\Delta y)$ is *Pectinate Function*, Δx and Δy are sampling spacing along phase direction and orthogonal direction respectively so that $f_x = 1/\Delta x$ and $f_y = 1/\Delta y$ represent the sampling frequency along the two directions respectively. Here suppose

$$\begin{cases} f_y = 1/\Delta y = mf_\phi \\ f_x = 1/\Delta x = nf \quad (f = \max\{f_1, f_2\}) \end{cases} \quad (7)$$

where m and n are multiple units, respectively represents the multiple relationship between sampling frequency and the selected experimental setup frequencies along the two orthogonal axes, and both them are positive numbers. These two introduced parameters enable us to calculate the proper sampling frequencies based on the known baseband and carrier frequencies, and the selecting criterions of sampling frequencies are determined as long as m and n are definitely.

Eq.4 shows that besides the frequency of background reflectance (i.e. $A(\xi,\eta)$), there are four peak values along the orthogonal direction, namely $\xi = \pm f_1, \pm f_2$; in each peak ξ there are three peak spectrums along the phase direction, namely $\eta = 0, \pm f_\phi$. To simplify the investigation, we will discuss the sampling effects along the two orthogonal directions respectively.

2.2.1. Sampling Analysis along Phase Direction

Any a peak value of ξ was selected, e.g. $\xi = f_1$, there are three peak spectrums in the channel along the phase direction:

$$F(f_1,\eta) = B(f_1,\eta) + \psi(f_1,\eta - f_\phi) + \psi^*(f_1,\eta + f_\phi) \quad (8)$$

The discrete spectrums of $F(f_1,\eta)$ can be calculated as

$$F_s(f_1,\eta) = F(f_1,\eta) * |\Delta y| comb(\Delta y\eta)$$
$$= \sum_{N=-\infty}^{\infty} [B(f_1,\eta - Nmf_\phi) + \psi(f_1,\eta - f_\phi - Nmf_\phi) + \psi^*(f_1,\eta + f_\phi - Nmf_\phi)\ ] \quad (9)$$

where $*$ is the convolution operator and N is integer. **Eq.9** indicates that the spectrums of $F(f_1,\eta)$ repeat periodically. To against the overlapping of spectrums, we must make sure ψ is separated from ψ^* in the same period, and also make sure $\psi(\psi^*)$ is separated from $\psi^*(\psi)$ comes from the adjacent periods, in another word, the sampling frequency is restricted, there must be at least four sampling dots in one period [9] so that

$$f_y = mf_\phi \geq 4f_\phi, \quad \text{or} \quad m \geq 4 \quad (10)$$

2.2.2. Sampling Analysis along the Orthogonal Direction

Any a peak value of η was selected to study the corresponding four peak spectrums of ξ along the orthogonal direction, e.g. $\eta = f_\phi$:

$$F(\xi, f_\phi) = \psi(\xi + f_2, f_\phi) + \psi(\xi + f_1, f_\phi) + \psi(\xi - f_1, f_\phi) + \psi(\xi - f_2, f_\phi) \quad (11)$$

Here suppose $f_2 > f_1$. Consider expression (7) and Pectinate Function $comb(x/\Delta x)$ along the orthogonal direction, the discrete spectrums of $F(\xi, f_\phi)$ can be calculated as (12):

$$F_s(\xi, f_\phi) = F(\xi, f_\phi) * |\Delta x| comb(\Delta x\ \xi)$$
$$= \sum_{N=-\infty}^{\infty} [\psi(\xi + f_2 - Nnf_2, f_\phi) + \psi(\xi + f_1 - Nnf_2, f_\phi) + \psi(\xi - f_1 - Nnf_2, f_\phi) + \psi(\xi - f_2 - Nnf_2, f_\phi)\ ] \quad (12)$$

Expression (12) indicates that the spectrums of $F(\xi, f_\phi)$ repeat periodically with period of nf_2. As shown in **Figure 2**, the real lines represent the starboard of $F(\xi, f_\phi)$, and the dashed represent the larboard spectrums of the adjoining period.

Figure 2 indicates that to escape the overlapping of spectrums, there must be have

$$\begin{cases} (f_1)_{\max} < (f_2)_{\min} \\ (f_2)_{\max} < nf_2 - (f_2)_{\max} \end{cases} \quad (13)$$

Imitating the definition of instantaneous frequency in domain of signal processing [5], we get

$$\begin{cases} (f_1)_{\max} = f_1 + (1/2\pi)|\partial\varphi/\partial x|_{\max} \\ (f_2)_{\min} = f_2 - (1/2\pi)|\partial\varphi/\partial x|_{\max} \\ (f_2)_{\max} = f_2 + (1/2\pi)|\partial\varphi/\partial x|_{\max} \end{cases} \quad (14)$$

Eq.13 can be modified by (5) and (14) such that

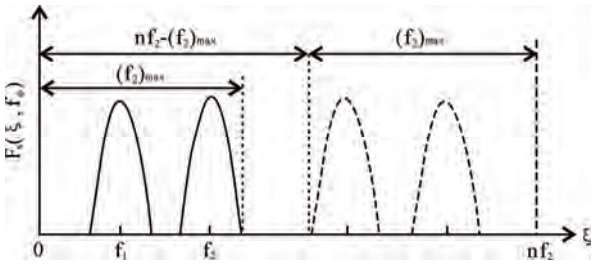


Figure 2. Replicated spectrum distribution.

$$\begin{cases} |\partial h/\partial x|_{\max} < (f_2 - f_1)(L_0/2f_\phi d) \\ |\partial h/\partial x|_{\max} < (n-2)f_2(L_0/2f_\phi d) \end{cases} \quad (15)$$

According to (15), we can get

$$n \geq 2 + (f_2 - f_1)/f_2 \quad (16)$$

If the condition $f_2 = 2f_1$ is selected in survey, there must be $n \geq 3$, so there has $f_x = nf_2 \geq 3f_2$.

2.3. Experiment

To support the analysis above, a model face was used as a tentative test. The projector used was a Panasonic (PT-P2500) digital projector with resolution of 1024×768. The image sensor used was a low-aberrance color CCD camera (Prosilica, EC1350C, made in Canada) with resolution of 1360×1024 and pixel size of 4.65um× 1.65um, and the maximum frame rate is 18fps. The focus of the camera lens (KOWA, LM12JCM, made in Japan) was 12mm. The image board was a 1394 card (KEC, 1582T, made in Taiwan). The reference plane as background was a piece of smooth and white board. **Figure 3** illustrates the experimental setup, in which the geometric parameters were set as L_0 =73mm and d =18mm, and the carrier frequencies f_1 and f_2 were set as 3/40 line/pixel and 6/40 line/pixel respectively and the baseband f_ϕ was given as 60/600 line/pixel. The lens of projector and camera must be at a same geometric plane surface and here they were setup coplanar at vertical curve. The horizontal beam contained CP illuminated over against model face, and the shooting angle of camera was setup as 45 degree which is

Figure 3. The diagram of the experimental setup.

the optimal angle value [4,5]. Because the specimen was a model face, the gesture and expression could be without consideration, however, with regard to a real human face, eyes exposure and shadows caused by gesture must be considered carefully.

According to the analysis above, to avoid spectral aliasing, there must be have $m \geq 4$ and $n \geq 3$. **Figure 4** shows the captured distortion composite pattern. **Figure 5** indicates the reconstructed profile. From the drawings, we can find that when sampling frequencies do not satisfy the sampling request, i.e. (a) when $m = 3$ and $n = 2$, the rebuilt errors were big and the details of face were lost. However, when $m \geq 4$ and $n \geq 3$, the details could be retrieved as shown in (b) and (c) with good resolution.

3. DISCUSSION

The accurate acquisition of 3D human face appearance characteristics is very important for designing a facial contouring surgery, and a good work is based on an exact 3-D face modeling [10]. People hope to find a non-ontact, rapid, precise way to acquire 3-D digital face depth, and then based on it to simulate and design an optimal plan for face surgery by modem technologies such as computer aided design etc [11].

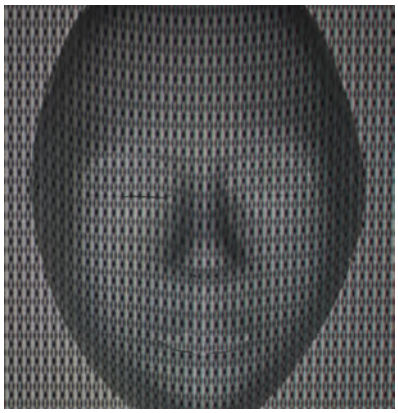
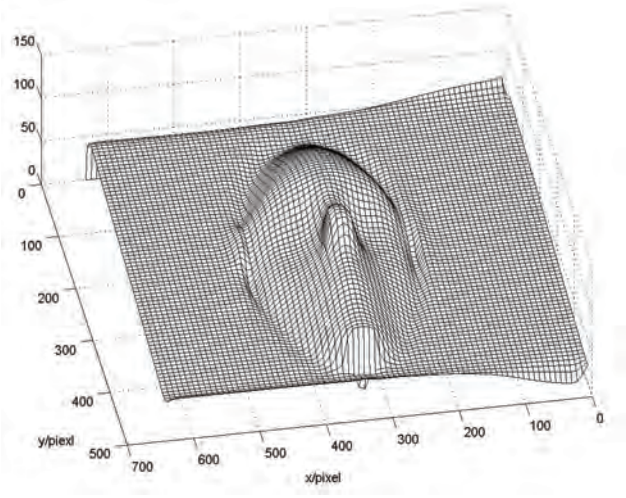
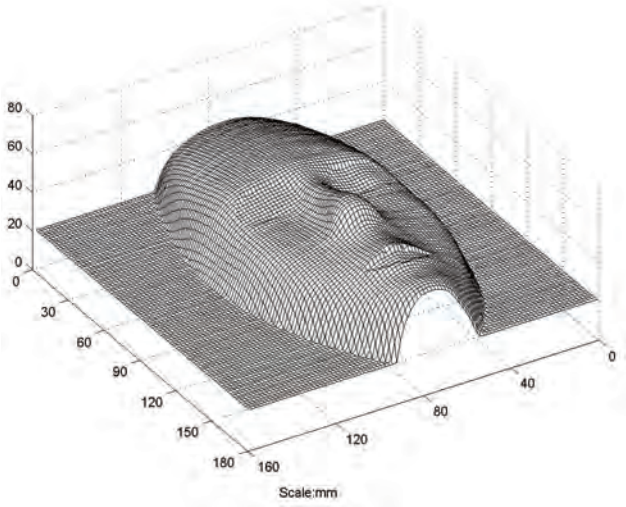


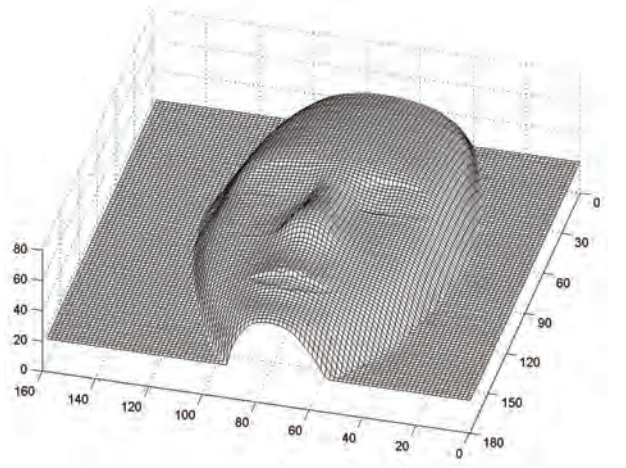
Figure 4. The captured distortion composite pattern modulated by height of the face model.



(a)



(b)



(c)

Figure 5. The rebuild shape, (a) when m=3, n=2; (b) when m=4, n=3; (c) m=5, n=5.

At present, there are about three types of 3D face modeling method to extract human face profile: one is the method based on computer tomography (CT) technology [12,13,14] and another one is based on passive optical 3D sensing technique [15,16,17] and the other is based on active optical 3D sensing technique [18,19,20].

The 3D reconstruction method based on CT technology is sensitively to skeleton and is convenient to be used for craniofacial plastics and oral and maxillofacial correction of abnormality and such fields, however, 3D profile of soft tissue is difficult to rebuilt by CT technology, especially the human face surface features.

The passive optical 3D sensing technique such as stereo vision uses two or more camera systems to capture the scene in ambient light from different viewpoints and to determine the height by matching image features of the corresponding surface features. In this method, a lot of factors need to be noticed, such as ambient light, background, vision angle and face gesture, expression and shading and so on, for they would influence the measuring accuracy directly. Besides, there always need to process a mass of data operation like correlation analysis and matching operation etc. Generally, the passive optical 3D sensing technique is more often used for 3D object recognition and understanding. Along with the development of computing technique, arithmetic speed is no longer a key limiting factor, and the passive optical 3D sensing technique is widely used in the field of machine vision.

The active optical 3D sensing technique employs structured light to illuminate the specimen. The time or space in structured light will be modulated by height, and then the 3D information can be extracted from the observation light by certain unwrapping algorithm [5]. For its feature of non-contact, high resolution and highly automated, the active optical 3D sensing technique is used in most 3D sensing systems with the purpose of 3D surface-shape measurement.

Phase Measuring Profilometry (PMP) is one important method of active optical 3D sensing technique [5]. In PMP, sinusoidal fringes and phase shifting technique are employed to acquire the height information that we wanted. A flaw of PMP is that it has to capture at least three continuous modulated phase shifting fringes corresponding to a static profile and therefore there will be some trouble for real-time dynamic measurement, and during the shooting process a little movement or facial expression changes of the target human face will potentially bring errors to the demodulated results. By using fast digital grating projection approach, a series of phase shifting fringes can be projected and shot within a short span of time. However, the images photographed by CCD camera would easily cause drawbacks such as trailing and distortion etc. due to rapid rotation of the phase shifting fringes, and then the inaccuracy of measurement will be raised. A one-shot technique, therefore, becomes a trend [19,20].

Here a novel one-shot approach for 3D human face profile measurement is introduced. A composite pattern (CP) is used in place of the series phase shifting fringes in PMP, and only a single frame of CP is needed to project and capture. The CP efficiently combines some phase shifting fringes and the same number of carrier gratings, and so that the phase shifting technique can be also utilized in this approach. This one-shot technique can avoid some unwanted troubles such as trailing and distortion etc. that happened in PMP for needing only one projection and corresponding one capture. Based on the proposed approach, 3D digital model of real human face could be acquired more conveniently and exactly.

Here we used this novel method to reconstruct a model face and acquired a good stereogram under the proper sampling frequencies which were the focus of our investigation. Because of the complexity of the composite pattern, another kind of spectrum overlapping would be brought in by the two modulating gratings during the digitization process. In this instance, choosing a proper sampling frequency is very important for the precise reconstruction. In the paper we discussed the sampling conditions along two directions and pointed out the rules, and then under the given sampling conditions we acquired a perfect digital 3D face profile.

4. CONCLUSIONS

Composite Fourier Transform Profilometry (CFTP) is an improved FTP method where a composite structured light is employed. To study the influence caused by sampling, the knowledge of pectinate function and convolution theorem was used and the suggestion that how to select proper sampling frequencies was given, that is, the sampling frequency along the phrase variation direction must be at least four times as the baseband and along the orthogonal direction it must be at least three times as the larger frequency of the two carrier frequencies.

5. ACKNOWLEDGEMENMTS

This study was supported by the National 973 Basic Research Program of China (No. 2010CB732600).

REFERENCES

[1] M. TAKEDA, K. MUTOH. (1983) Fourier transform profilometry for the automatic measurement 3-D object shapes. Appl. Opt., **24**, 3977–3982.

[2] M. Takeda, H. Ina. (1982) Fourier-transform method of fringe-pattern analysis for computer-based topography and interferometry. J. Opt. Soc. Am, **1**, 156–160.

[3] S. Fu, Y. S. Wang, G. B. Han. (2004) Fourier transform profilometry in 3-D measurement based on wavelet digital Filter. J.Optoelectronics·Laser, **2**, 205–207.

[4] W. J. Chen, X. Y. Su. (2000) Discussion on phase errors caused by frequency leakage in FTP. Acta Optica

Sinica, **10**, 1429–1433.

[5] X. Y. Su, J. T. Li. (1999) Information Optics. Publish House of Science, Beijing, 332–335 (in Chinese).

[6] C. Guan, L. G. Hassebrook. (2003) Composite structured light pattern for three-dimensional video. Optics Express, 5, 406–417.

[7] H. M. Yue, X. Y. Su. (2005) Improved fast fourier transform profilometry based on composite grating. Acta Optica Sinica, **6**, 767–770.

[8] Y. S. Xiao, X. Y. Su, Q. C. Zhang. (2006) 3-D surface shape restoration for the breaking surface of dynamic process. Laser Technology, **3**, 258–261.

[9] H. Yang, W. J. Chen. (1999) Influence of Sampling on Fourier-Transform Profilometry. Acta Optica Sinica, **7**, 929–934.

[10] G. L. Murrell, N. K. McIntyre, and B Trotter. (2003) Facial contouring. Facial Plast Surg Clin North Am, **3**, 391–397.

[11] R. Schmelzeisen and A. Schramm. (2003) Computer-assisted reconstruction of the facial skeleton. Arch Facial Plast Surg, **5**, 437–440.

[12] S. Prakoonwit, and R. Benjamin. (2007) Optimal 3D surface reconstruction from a small number of conventional 2D X-ray images. Journal of X-Ray Science and Technology, **4**, 197–222.

[13] M. Deling, W. Biao, F. Peng, and Y. Fuguang. (2007) Oral Implant Orientation of 3-D Imaging Based on X-Ray Computed Tomography (CT). Asian J. Inform. Techno, **11**, 1143–1147.

[14] Mahfouz, Badawi, A. Fatah, Kuhn, and Merkl. (2006) Reconstruction of 3D patient-specific bone models from biplanar xray images utilizing morphometric measurements. Proceedings of the 2006 International Conference on Image Processing, Computer Vision, and Pattern Recognition, 345–349.

[15] C. Zhang, F. S. Cohen, and H. RVSI. (2002) 3-D face structure extraction and recognition from images using 3-D morphing and distance mapping. IEEE Transactions on Image Processing, **11**, 1249–1259.

[16] H. Hirschmuller, P. R. Innocent, and J. Garibaldi. (2002) Real-time correlation-based stereo vision with reduced border errors. Int. J. Comput. Vis., **1/2/3**, 229–246.

[17] S. Huq, B Abidi, A. Goshtasby, and M. A. Abidi. (2004) Stereo matching with energy-minimizing snake grid for 3D face modeling. Proceedings of SPIE, **5404**, pp. 339–350.

[18] M. Takeda and K. Mutoh. (1983) Fourier transform profilometry for the automatic measurement 3-D object shapes. Appl. Opt., **24**, 3977–3982.

[19] H. M. Yue, X. Y. Su, and Z. R. Li. (2005) Improved fast fourier transform profilometry based on composite grating. Acta Optica Sinica, **6**, 767–770.

[20] C. Guan, L. G. Hassebrook, and D. L. Lau. (2003) Composite structured light pattern for three-dimensional video. Optics Express, **5**, 406–417.

J. Biomedical Science and Engineering, 2009, 2, 612-616 **JBiSE**
doi: 10.4236/jbise.2009.28088 Published Online December 2009 (http://www.SciRP.org/journal/jbise/).

Comparative analysis of current and magnetic multipole graphical models

Shi-Qin Jiang, Lu Bing, Jia-Ming Dong, Ming Chi, Wei-Yuan Wang, Lei Zhang

School of Electronics and Information Engineering, Tongji University, Shanghai, China.
Email: sqjiang@tongji.edu.cn

Received 28 July 2009; revised 1 September 2009; accepted 2 September 2009.

ABSTRACT

In recent year, a multipole graphical model, which is constructed by using individual MCG measurements based on the equivalent current dipole (ECD) or equivalent magnetic dipole (EMD) source model, has been developed with the aim of instead of the volume conductor model in the inverse solution of cardiac source estimation. In this paper, two graphical models known as the double magnetic dipole source model (DMD) and the dual current dipole source model (DCD) are introduced. The simulation results and the comparison of two evaluation criteria, i.e. average GOF (Goodness of Fit) and average RMSE (Root Mean Square Error), indicated that both multipole graphical models can provide a good representation of dynamic magnetic field from the noninvasively detected MCG-recordings, even when the heart is of the dilation. The time-averaged sources localization error and the RMSE for both models are demonstrated, and the characteristic of two multipole models is discussed.

Keywords: Biomagnetics; Inverse Problems; Dipole Source Localization; Modeling; Magnetocardiography

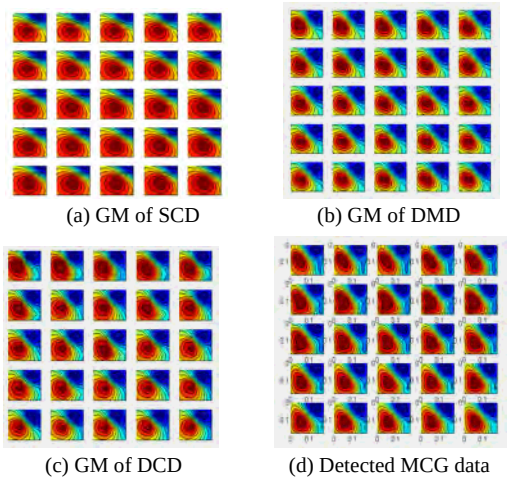
1. INTRODUCTION

In order to investigate cardiac electrical activity, the issue of reconstructing noninvasively the electrical or magnetic sources from detected MCG signals has received much attention since 1970'.

Due to the effect of human torso, the volume conductor model, i.e., heart-torso model, such as 3D finite element model (FEM), boundary element model (BEM), and ventricular propagated excitation model were developed [1,2,3]. Front two models are created with magnetic resonance imaging (MRI) data and conductivity values are assigned to each region. It is demonstrated that the use of torso models has brought significant improvements in results of dipole source localization.

In recent years, for both research and clinical applications, we developed a graphical model (GM) to describe the active magnetic field between detected MCG data and cardiac electrical sources [4]. The aim is to provide a simple model which can describe the varying magnetic field and conductivity properties of tissue. Furthermore, two multipole source models known as the double magnetic dipole source model (DMD) and the dual current dipole source model (DCD) are investigated. In **Figure 1**, there are three graphical models which are constructed based on different source models [5,6,7].

The graphical model consists of a set of magnetic field maps (MFM), which is constructed by using individual MCG measurements based on the equivalent current dipole (ECD) or equivalent magnetic dipole (EMD) source model. Each graphical model includes 25 magnetic field maps with a time interval of 4ms during 100 ms in ST-T segment. Each map of the graphical model corresponds on a set of model parameters, by which the space time pattern of the magnetic field over the body surface can be obtained with high-resolution. The procedure of model constructing is illustrated in **Figure 2**. It is implemented with three steps: initial values determination, source estimation, and GM construction in terms of optimized source parameters, which are estimated by applying the Levenberg-Marquart (LM) or the Nelder-Mead (NM) algorithm.



(a) GM of SCD (b) GM of DMD

(c) GM of DCD (d) Detected MCG data

Figure 1. Three graphical models and detected MCG data.

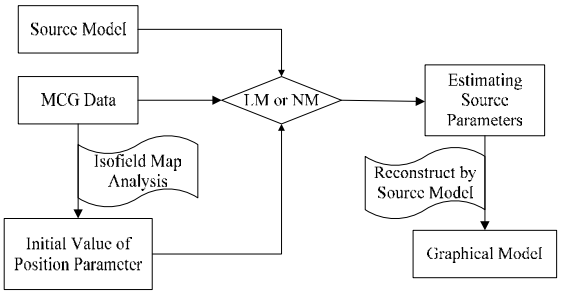


Figure 2. Schematic diagram of constructing a graphical model.

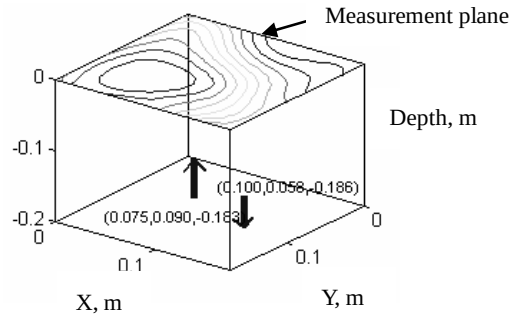


Figure 3. Sketch of a pair of magnetic dipole under the measurement plane.

Two multipole source models mentioned above, i.e., DMD and DCD, have different characteristic from current dipole and magnetic dipole. The double magnetic dipole is similar as a single current dipole (SCD), which is more suitable for experimental implementation. The dual current dipole is originally used to research the method of constructing a multipole source model. In this paper, sources localization results of both source models are demonstrated, and the characteristics of two different multipole graphical models are discussed.

2. METHODS

2.1. Double Magnetic Dipole Source Model

A simplified double magnetic dipole source model has developed with a pair of magnetic dipole as shown in **Figure 3**. The necessary conditions of the model are: one of magnetic moments should be positive, and another should be negative. Two magnetic dipoles under the surfing and the sinking of the detected magnetic field over the body surface at the same time. In other words, the magnetic moments are simplified as all vertical to measuring plane and opposite in direction. Therefore, the simplified double magnetic dipole is similar to an equivalent current dipole. However, the model has more parameters than that of SCD model, thus, it has more degree of freedom.

The magnetic field B_z detected along the Z-axis, which generated by a pair of magnetic dipole, is defined as:

$$B(x_j, y_j) = \sum_{i=1}^{2} \frac{m_i m_0 (2z_{0i}^2 - (x_j - x_{0i})^2 - (y_j - y_{0i})^2)}{4p((x_j - x_{0i})^2 + (y_j - y_{0i})^2 + z_{0i}^2)^{5/2}} \quad (1)$$

where $(x_{0i}, y_{0i}, z_{0i})(i = 1, 2)$ is the dipole position and $m_i(i = 1, 2)$ is the magnetic moment of two magnetic dipoles, respectively.

2.2. Dual Current Dipole Source Model

The dual current dipole source model was developed as the simplest multipole current source model. A magnetic field zero line (MFZL) method was proposed for determining the number of dipoles and the initial values in multiple source inverse solution. In general, the slope of the magnetic field zero line (MFZL) of magnetic field maps is mutative, which has been shown in **Figure 4(a)**. There are 25 varying MFZLs in the ST-T segment obtained from the detected magnetocardiograms in **Figure 1(d)**. As shown in **Figure 4(b)**, the MFZL can be roughly divided into several linear subsections, here are two subsections, according to its slope transition. Thus, we assume that there exist two current dipoles. Every one locates at the middle of a MFZL subsection. The strength of each current dipole depends on the length of the corresponding MFZL subsection [8]. Obviously, one current dipole is dominant; another is an accessorial equivalent dipole. Based on the above a priori assumptions, an advisable method of constructing a multipole source model is to locate a current dipole on a MFZL subsection. The method can be used to determine the initial values of dipoles in inverse solution. The multiple current dipoles source model is defined as:

$$B(x_j, y_j) = \sum_{i=1}^{n} \frac{(y_j - y_i) \cdot m_0 Q_{ix} - (x_j - x_i) \cdot m_0 Q_{iy}}{4p((x_j - x_i)^2 + (y_j - y_i)^2 + z_i^2)^{3/2}} \quad (2)$$

where (xi, yi, zi) is the position of the current dipole, Qx and Qy are the x and y components of current moments of the dipole, respectively. N is the number of dipoles.

The simulation results of the dual current dipole are presented as followings. It demonstrated that the accuracy of the graphical model was improved by means of the multipole current source model.

3. SIMULATION RESULTS

3.1. Accuracy of Graphical Models

It is necessary that the graphical model is of high accuracy for describing the detected MCG contour maps. The accuracy of graphical models depends on the accuracy of sources estimation. Thus, the source model and the algorithm play an important role in inverse solution. In terms of two evaluation criteria, i.e. GOF (Goodness of Fit) and RMSE (Root Mean Square Error), a comparison

performance among three graphical models is given in **Figure 5** and **Figure 6**, respectively. Two evaluation criteria are defined as followings:

$$GOF = \sqrt{1 - \frac{\sum_{i=1}^{N} (B_{zi} - B_{si})^2}{\sum_{i=1}^{N} B_{zi}^{\ 2}}} \quad (3)$$

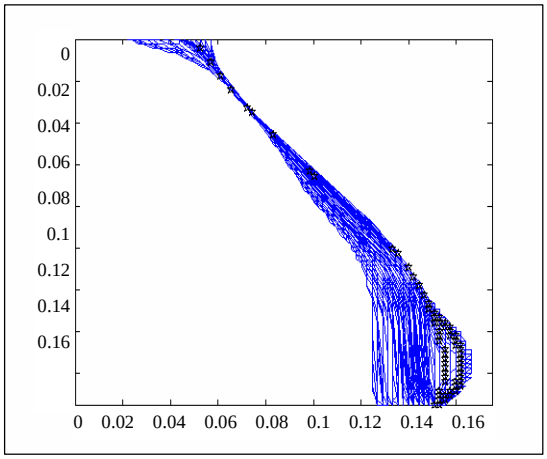
$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (B_{zi} - B_{si})^2} \quad (4)$$

where B_z is the detected magnetic field, B_s is the calculated magnetic field, and N is the number of measuring points. **Figure 5** and **Figure 6** demonstrated that two graphical models, i.e. DCD and DMD, all have higher GOF (more than 0.97) and lower RMSE (about 30 pT), compared with that of the SCD model.

3.2. Source Localization

The accuracy of equivalent dipole source localization is tested based on the double magnetic dipole and the dual current dipole graphical models by using the MCG measurements as shown in **Figure 1(d)**. The inverse problems are calculated with two non-linear local optimization algorithms, LM and NM algorithms, respectively. Two noise levels are considered in simulation, i.e. no noise case and 20dB signals-noise-ratio (SNR), respectively. The initial values are determined by means of a parameters calculation procedure. In simulation two cases were included: use of the calculated initial values and use of the calculated initial value adding a random number whose magnitude is 10% of the calculated initial values.

The time-averaged localization error and the time-averaged RMSE in 100ms ST-T segment with the DMD and DCD models are shown in **Tables 1-2** and **Tables 3-4**, respectively. In **Table 1**, the P expresses the positive magnetic dipole, and the N is the negative magnetic dipole. In **Table 3**, D and A express the dominant current dipole and the accessorial equivalent dipole, respectively. We can see that in most cases, the NM algorithm performs better than the LM algorithm regarding both of the averaged localization error and the averaged RMSE. By using NM algorithms, the sources localization results of the DCD model are 0.99 mm for the dominant dipole and 1.23 mm for the associate equivalent dipole when calculated initial values are



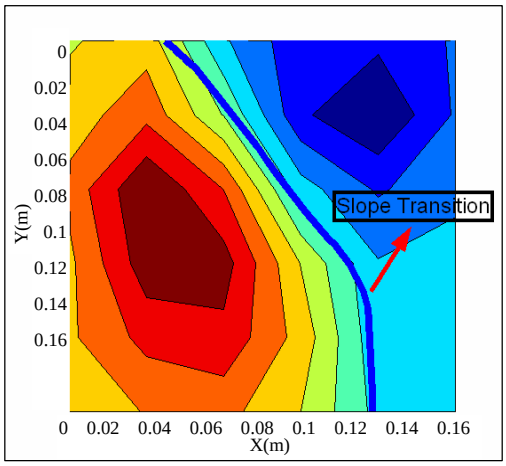
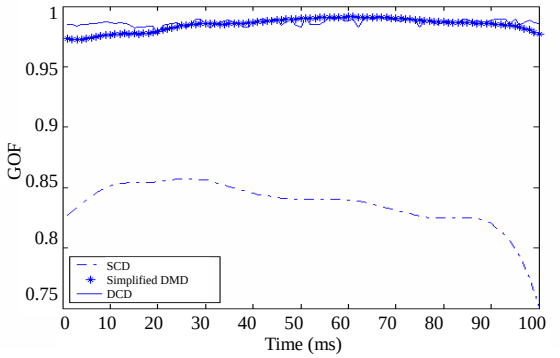


Figure 4. Schematic diagram of magnetic field zero line.



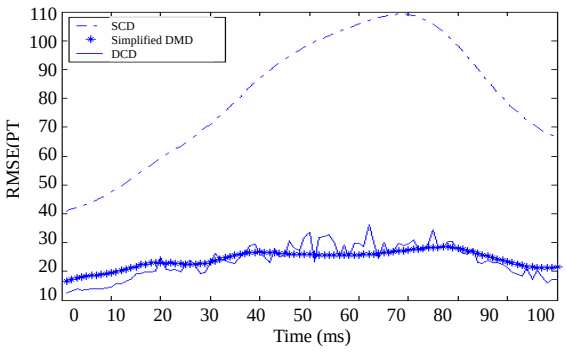


Figure 5. GOF curves of three GMs.

Figure 6. RMSE curves of three GMs.

Table 1. Averaged localization error of DMD model (mm).

Data noises \ Initial values		Calculated Initial Values		10% Random	
		LM	***NM***	***LM***	***NM***
No noise	P	0.0	0.0	0.34	0.41
	N	0.0	0.0	0.69	1.06
20db noise	P	1.34	1.86	1.39	1.40
	N	3.69	4.84	3.74	3.72

Table 2. Averaged RMSE of DMD model (pt).

Data noises \ Initial values	Calculated Initial Values		10% Random	
	LM	***NM***	***LM***	***NM***
No noise	0.0	0.0	189	155
20db noise	232	187	255	196

Table 3. Averaged localization error of DCD model (mm).

Data noises \ Initial vaues		Calculated Initial Values		10% Random	
		LM	***NM***	***LM***	***NM***
No noise	D	0.0	0.0	1.14	0.89
	A	0.0	0.00	1.43	0.96
20db noise	D	1.20	0.99	1.71	0.99
	A	1.79	1.22	1.88	1.23

Table 4. Averaged RMSE of DCD model (PT).

Data noises \ Initial values	Calculated Initial Values		10% Random	
	LM	***NM***	***LM***	***NM***
No noise	0.0	0.0	211	112
20db noise	228	137	236	155

used. In despite of the localization results of the DMD model have been improved, but they are still not as good as that of the DCD model.

Furthermore, **Figure 7** to **Figure 8** show the time-varying position and orientation curves of two magnetic dipoles in the 100 ms ST-T segment. The positive magnetic dipole, i.e. the dipole under the surfing of the magnetic field contour map, moved in a range of 0.11-0.12 m at Z direction, and the negative dipole moved in a distance about 0.106-0.116 m away from the measurement plane on the chest. In other words, the depth of two equivalent magnetic dipoles changed in a small region. Two moving magnetic dipoles are confined within the region of the heart. It is noticeable that x, y, z components curves of the location of two equivalent current dipoles are shown in **Figure 9** and **Figure 10**. The first current dipole, i.e. the dominant dipole, moved in a range of 0.06-0.07 m at Z direction, and the accessorial equivalent dipole, moved in a distance of 0.11-0.14 m away from the measurement plane on the chest. In other words, the accessorial equivalent dipole located a little deeper than the dominant dipole. Two moving current dipoles are confined within the region of the heart.

In sum, two equivalent magnetic dipoles are moved in almost the same depth as the accessorial equivalent current dipole. The x and y components curves of the location of the accessorial equivalent current dipole have obvious sudden variation, however, two equivalent magnetic dipoles have no such phenomena.

4. CONCLUSIONS

Two multipole graphical models are introduced for the investigation of the cardiac electrical activity in this paper. Because the accuracy of graphical models depends on the accuracy of sources estimation, thus, the initial value

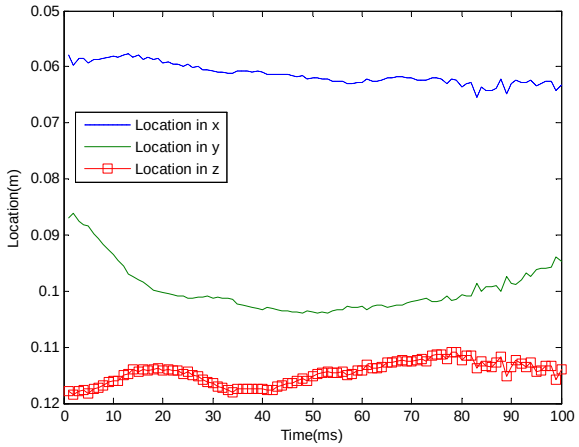


Figure 7. The time-varying location curves of the positive magnetic dipole.

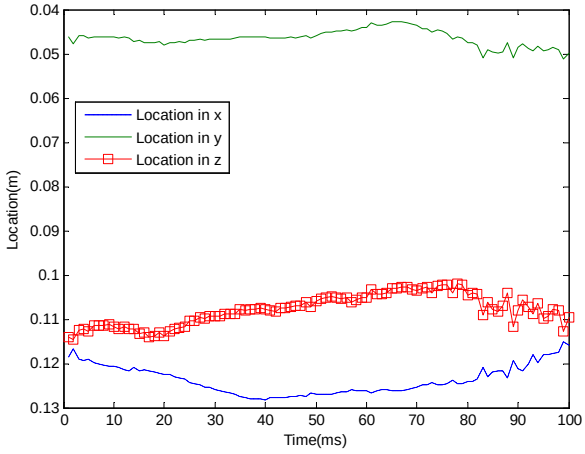


Figure 8. The time-varying location curves of the negative magnetic dipole.

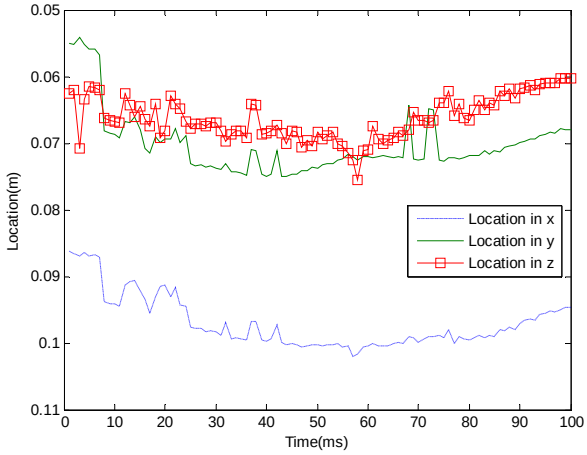


Figure 9. The time-varying location curves of the dominant current dipole.

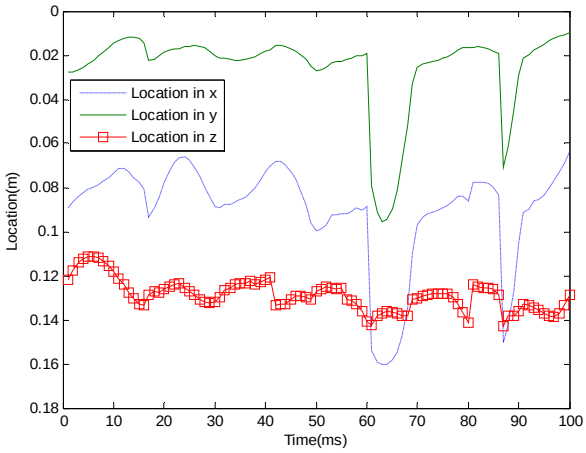


Figure 10. The time-varying location curves of the accessorial current dipole.

determining of source estimation is very important. The simulation results of the DCD model demonstrated that MFZL method mentioned above is an advisable approach to determine the initial values and the number of dipoles. Despite the simulation and the comparison results indicated that the both multipole graphical models have better GOF and RMSE, however, how to improve the accuracy of the graphical models still is a problem that remains to be solved. We have noted that two groups of reconstructed sources had some obvious difference on the accuracy of source localization, especially, the location variation of the accessorial current dipole in **Figure 10**. The time-varying location curves of the accessorial current dipole as shown in **Figure 10** revealed some useful information of electrophysiological characteristics and the function information of the heart during ventricular repolarization, which needs analyzing together with the professional doctors in the future.

5. ACKNOWLEDGMENTS

This work obtained support from the National Natural Science Foundation of China (60771030), National High-Technology Research and Development Program of China (2008AA02Z308), the Shanghai Science and Technology Development Foundation (08JC 1421800), and the Open Project of State Key Laboratory of Function Materials for Information.

REFERENCES

[1] D. Wu, H. C. Tsai, B. He. (1999) On the estimation of the Laplacian electrocardiogram during ventricular activation. Ann. Biomed. Eng., **27**, 731–745.

[2] J. Haueisen, J. Schreiber, H. Brauer, and T. R. Knösche, (2002) Dependence of the inverse solution accuracy in magnetocardiography on the boundary-element discretization. IEEE Transactions On Magnetics, **38(2)**.

[3] S. Ohyu, Y. Okamoto, S. Kuriki. (2000) Use of the ventricular propagated excitation model in the magnetocar-diographic inverse problem for reconstruc tion of electrophysiological properties. IEEE Transactions on Biomedical Engineering, **49(6)**.

[4] S. Q. Jiang, L. Zhang, M. Chi, M. Luo, L. M. Wang. (2008) Dipole source localization by means of simplified double magnetic dipole model. International Journal of Bioelectromagnetism (IJBEM), **10(2)**.

[5] S. Jiang, M. Chi, L. Zhang, M. Luo, L. Wang. (2007) Dipole source localization in magnetocar diography. Proceedings of 2007 Joint Meeting of the 6th International Symposium on Noninvasive Functional Source Imaging of the Brain and Heart.

[6] M. Chi, S. Jiang, L. Zhang. (2008) Graphical model of cardiac electromagnetic source. International Conference on Bioinformatics and Biomedical Engineering (ICBBE' 08).

[7] S. Jiang, J. Dong, M. Chi, and W. Wang. (2008) A graphical model for the cardiac multi-dipole sources. Proceedings of the 5th International conference on Information Technology and Application in Biomedicine (ITAB'08), 434–436.

[8] S. Jiang, W. Y. Wang, J. M. Dong, and A. L. Li. (2008) modeling of bioelectrical activity by means of measured mcg data. Biomagetism-Transdisciplinary Research and Exploration, 241–243.

[9] W. Andra and H. Nowak. (1998) Magnetism in medicine. Germany, WILEY-VCH.

[10] B. He. (2004) Modeling and imaging of bioelectrical activity principles and applications. Kluwer Academic/ Plenum Publishers, 161–162.

J. Biomedical Science and Engineering, 2009, 2, 617-620 **JBiSE**
doi: 10.4236/jbise.2009.28089 Published Online December 2009 (http://www.SciRP.org/journal/jbise/).

Effects of ultra-high hydrostatic pressure on foaming and physical-chemistry properties of egg white

Rui-Xiang Yang, Wen-Zhao Li, Chun-Qiu Zhu, Qiang Zhang

Key Laboratory of Food Nutrition and Safety, Tianjin University of Science & Technology, Tianjin, China.
Email: yrxsky@126.com

Received 11 July 2009; revised 1 September 2009; accepted 2 September 2009.

ABSTRACT

The influences of ultra-high hydrostatic pressure treatment on foaming and physical properties (solubility, hydrophobicity and sulfhydryl content) of egg white were investigated. A pressure range of 0-500 MPa, time range of 0-20 min and pH range of 7.5-8.5 were selected. The foaming property of egg white is improved by 350Mpa and 10min. The treatment resulted in increase of sulfhydryl content of egg white, while solubility and hydrophobicity were significantly decreased.

Keywords: Ultra-High Hydrostatic Pressure; Egg White; Foaming Property

1. INTRODUCTION

Egg white is well known for their high nutritional quality, foaming, gelling and emulsifying characteristics, which can give the foods unique color, flavor, and texture characteristics. Therefore, as an important ingredient, egg white has been wildly applied in the food industry, such as cakes, biscuits, breads, ice cream, and other protein products [1]. At present, for convenience and safety, liquid egg white products are widely used by home and producer. However, decrease of the foaming property after pasteurization is found in practice, which seriously affects the application and development of the egg white products.

Ultra-high static pressure technology is a new sterilization technology, which can be used to resolve the problems brought by pasteurization. Ultra-high pressure can improve the function of biological macromolecules by modifying their structure [2]. The high pressure does not affect the primary structure of protein molecules, but can cause agglutination of protein by changing the hydrogen bonds, disulfide bonds and hydrophobic groups among the protein molecules. In addition, ultra-high hydrostatic pressure can cause viscosity and surface tension of egg white increased [3]. The relationship between foaming property and solubility, hydrophobicity, sulfhydryl content under ultra-high hydrostatic pressure is also investigated in this paper, and only the irreversible changes in the properties are taken into account.

2. MATERIALS AND METHODS

2.1. Materials and Equipments

Fresh eggs were obtained from supermarket. Eggshell was washed by clean water and exposed to ultraviolet light for 30 min to disinfect. The egg white was separated from egg yolk and the chalazae were removed. The albumin was gently mixed and stored at 4 ℃ until use. The protein content of the egg white was determined to be 11.23 ± 0.56 % (w/v).

2.2. Ultra-High Hydrostatic Pressure Treatment

A volume of 200 mL of egg white was packed in polyethylene plastic packs (170×120mm) and sealed. The ultra-high hydrostatic pressure treatment was performed in high pressure equipment (Hpp-M1, Senmiao, China) and a pressure range of 0-600 MPa, time range of 0-20 min were selected in the experiment. All samples were analyzed at room temperature after 24-hour storage at 4℃.

2.3. Determination of Foaming Property

A volume of 35 mL (18～25℃) egg white was placed in a graduated cylinder of 250 mL (diameter =398 mm) and whipped for 5 min (which is an optimal time for whipping egg white in this measurement system in previous study) with a rotating anchor (275 mm diameter, rotor from the bottom of the graduated cylinder 15 mm) at 2100 rpm, using a laboratory stirrer with controlled constant speed (OJ-100, Onuo, China). The volume of foam (V_0) was recorded immediately after the whipping stopped. The foaming capacity (FC) [4] was defined by

$$FC(\%) = \frac{V_0}{35} \times 100\% \quad (1)$$

The volume of drainage of liquid (V_d) was recorded at 10 min, The foam stability was defined by

$$Drainage(\%) = \frac{V_d}{V_0} \times 100\% \quad (2)$$

Samples were determined in triplicate.

2.4. Determination of Solubility

Samples were centrifuged during 10 min at 10,000 r/min and 18℃. Protein content of the supernatant was determined.

2.5. Determination of Sulfhydryl Content

The concentration on sulfhydryl (SH) groups of the egg white solutions was determined using Ellman's reagent (5′, 5-dithiobis (2-nitrobenzoic acid), DTN-B). 1mL of sample add to 4mL of 0.05mL 0.01M DTNB in phosphate buffer (0.1M, pH 8.0), after 20min of mixing, the reaction mixture was centrifuged during 15 min at 10,000 r/min to remove precipitated protein. Finally, the absorbance of the supernatant was measured at 412 nm against a reagent blank. A blank sample in which DTNB was substituted by phosphate-buffer was carried through in parallel. The sulfhydryl content is expressed as a mole thiol/globulin grams = n mol / mg protein [5].

2.6. Determination of Surface Hydrop-Hobicity

The determination of the protein hydrophobic domains through Takagi [4] approach, using 8-aniline-based-1-Chennai acid (referred to as ANS). Each sample was diluted with the 0.1M phosphate buffer (pH6.8), until the protein concentration of 0.05%, 4.5mL of the diluted sample and 0.5mL 1.25×10^{-3}m ANS mixed phosphate buffer. Placing at room temperature for two hours, the fluoresceence intensity was measured (excitation wavelength 375nm, Kaloula 1.0nm, emission wavelength of 420-600nm (475nm), Kaloula 1.0nm, medium-speed scanning speed). Globulin in the hydrophobic region is expressed as fluorescence intensity and protein concentration of 0.05%.

3. RESULTS AND DISCUSSIONS

3.1. Effect of High Hydrostatic Pressure on Foaming C Property of Egg White

Foaming property is one of the most important surface characteristic of protein molecule. Albumin proteins are typical amphiphilic molecules, which are easy to extend on the water-air interface and adsorb gas in the whipping process. It has been verified that globulin and albumin of egg white play a major role in forming bubble; Ovomucin and egg lysozyme are favorable for foaming stability [1]. Iesel Plancken has shown that the foaming property of egg white is mainly relative to the content of sulfhydryl and the flexibility of protein molecule. The interaction between protein and protein can improve the foaming stability of egg white [2].

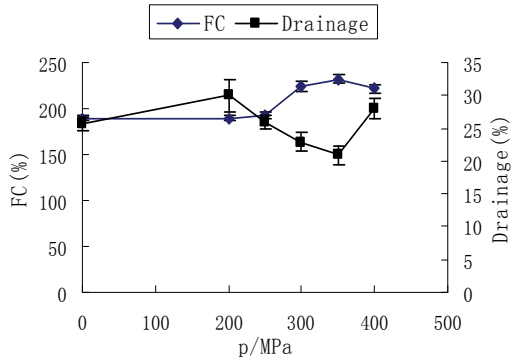


Figure 1. Effect of pressure on foaming property of egg white pH =8.1, t = 10 min, 18℃.

There are not significant changes for FC between 0-200MPa, as can be seen in **Figure 1**. But the drainage is increasing and reaches the maximum at 200MPa, which means the worst foam stability. With the pressure increasing, the FC is still rising until 350 MPa, but the drainage has been decreasing and reaches the minimum at 350 MPa. According to the relevant literature, the phenomenon above was analyzed as follows. It is known that practical aggregation of proteins can raise the protein foaming capability and stability [4]. In this process, aggregation affection can enhance the reaction among proteins and make them become precipitate or gel. While the protein gel appears, the foaming capability and stability will decline sharply.

3.2. Effect of Time on Foaming Property of Egg White

As can see from **Figure 2**, with the extension of processing time, the foaming capability increases and reaches the maximum at 10 min. While the drainage drops to the lowest, there is the best foaming stability. It indicates that in the aggregation of protein process, 10 min is suitable for foaming characteristic of egg white. According to the relevant report, the β-sheet structures of

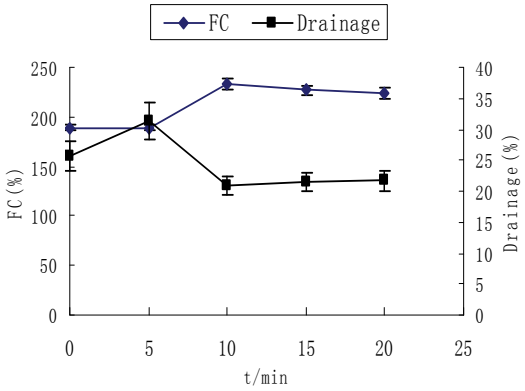


Figure 2. Effect of time on foaming property of egg white p = 350 MPa, pH = 8.2, 18℃.

protein treated with the pressure are unfolded, which can improve the ability of maintaining gas for egg white and make it easy to form bubble [4]. And also, it has enhanced the interaction among proteins, which is favorable for foaming stability. After 10min, the protein precipitate starts to appear, and the foaming capability is decreasing. The previous studies have shown that when egg white is treated by 350Mpa for 10min, the surface tension and viscosity is increasing [3]. From the phenomenon and the relevant conclusion above, we can draw a conclusion as follows: ultra-high pressure treatment can change the structure of protein molecules and cause aggregation of proteins, which can improve the foaming property of egg white.

Next, 10min was selected for different ultra-high pressure treatment to investigate the relationships between foaming property and other characteristic of egg white.

3.3 Effect of High Hydrostatic Pressure on Solubility of Egg White

Protein solubility is an important property of hydration. The protein and water molecules are connected through the interaction of the peptide bond or amino acid side chain (ionization, polar or non-polar group) [1]. As can be seen from **Figure 3**, the solubility of protein decreases with the pressure increasing. The tertiary and quaternary structures of protein treated by the ultra-high pressure were destroyed for that molecules of protein have been unfolded [4]. The reactions of aggregation among protein molecules have been enhanced, therefore, the solubility of egg white becomes decreasing. Comparing to **Figures 1,3**, there are not significant linear correlation between the foaming properties and solubility of egg white treated by ultra-high pressure. When the percent of soluble protein is 71%, the foaming property of egg white is better.

3.4. Effect of High Hydrostatic Pressure on Surface Sulfhydryl Content of Egg White

As can been seen from **Figure 4**, there are significant changes of surface sulfhydryl content with pressure reaching the maximum at 350MPa. When the egg white is treated by ultra-high hydrostatic pressure, the structure of the ovalbumin has been changed, which causes four internal sulfhydryl of c molecule being exposed and in creases the surface sulfhydryl content on the bases of relevant literature. Comparing with **Figures 1,5**, there is the same trend between the foaming properties and the surface sulfhydryl content of egg white. The foaming capability, stability and surface sulfhydryl content all reach the maximum at 350 Mpa.

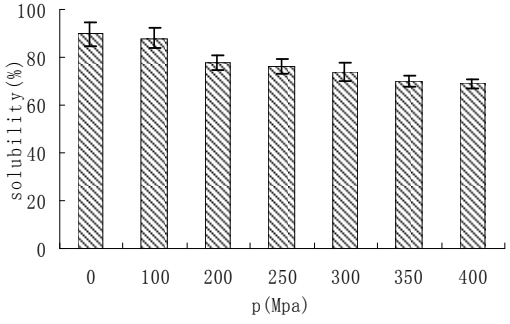


Figure 3. Effect of pressure on solubility of egg white pH =8.1, t = 10 min, 18℃.

3.5. Effect of High Hydrostatic Pressure on Hydrophobicity of Egg White

As can been seen from **Figure 5**, the hydrophobic of egg white has the maximum at 100MPa. With the pressure increasing, hydrophobicity of egg white decreases. It is known that hydrophobic interaction is the main force to maintain protein tertiary structure. So we can deduce that the tertiary and quarternary structures of proteins of egg white treated by the ultra-high hydrostatic pressure are destroyed, and more hydrophobic regions are exposed. With the pressure increasing, hydrophobic amino acids are buried and the hydrophobic region is reduced. Therefore, hydro-phobic property of protein decreased.

4. CONCLUSIONS

The foaming capability and stability of egg white are im-

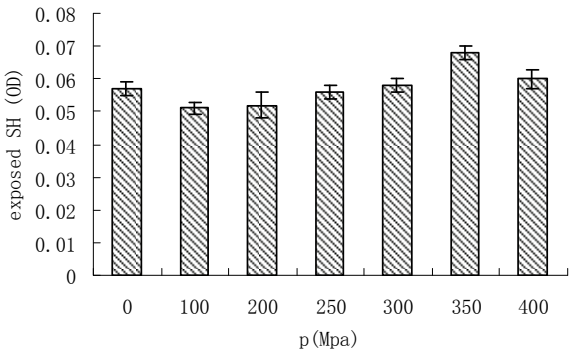


Figure 4. Effect of high hydrostatic pressure on sulfhydryl content of egg white pH =8.1, t = 10 min, 18℃.

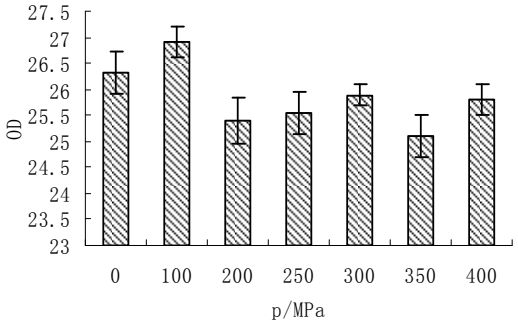


Figure 5. Effect of high hydrostatic pressure on hydrophobicity of egg white pH =8.1, t = 10 min, 18℃.

proved by the ultra-high hydrostatic pressure (100-400 MPa), and play the best performance by 350Mpa and 10min. The findings of the experiment of physical properties of egg white with ultra-high hydrostatic pressure show that with increasing pressure, the solubility of egg white will decrease. There is positive relation between the surface sulfhydryl content and the foaming properties of egg white. When the hydrophobicity of egg white is the lowest, the foaming capacity and foam stability of egg white will reach the best performance.

5. ACKNOWLEDGEMENTS

This research is support by the research fund of personnel of Tianjin University of Science & Technology.

REFERENCES

[1] X. D. Li and L. W. Zhang. (2005) Egg Science and Technology. Chemical Industry Press.

[2] E. V. Plancken, A. VanLoey, and M. E. Hendricks. (2007) Foaming properties of egg white affected by heat or high pressure treatment, Journal of Food Engineering, **78**, 1410–1426.

[3] W. Wang and W. Z. Li. (2009) Effect of ultra-high hydrostatic pressure on foaming and physical properties of the egg white. Journal of Tianjin University of Science and Technology, **24**, 35–38.

[4] H. Wang and Z. C. Tu. (2008) Egg white protein dynamic modification of ultrahigh-pressure micro jet and mechanism of, Nanchang University.

[5] K. Lomakina. (2005) A study of the factors affecting the foaming properties of egg white-a review. Food Science, **24**, 110–118.

[6] J. Liu and B. Jiang. (2007) Effects of ultra-high pressure on foaming property of chickpea protein isolated. Anhui Agricultural Science, **35**, 9012–9013.

J. Biomedical Science and Engineering, 2009, 2, 621-625 **JBiSE**
doi: 10.4236/jbise.2009.28090 Published Online December 2009 (http://www.SciRP.org/journal/jbise/).

FastCluster: a graph theory based algorithm for removing redundant sequences

Peng-Fei Liu[2], Yu-Dong Cai[1*], Zi-Liang Qian[3,4], Sheng-Yu Ni[5], Liu-Huan Dong[5], Chang-Hong Lu[6,7], Jin-Long Shu[6] , Zhen-Bing Zeng[2,5], Wen-Cong Lu[8*]

[1]Institute of Systems Biology, Shanghai University, Shanghai, China;
[2]Software Engineering Institute of East China Normal University, East China Normal University, Shanghai, China;
[3]Bioinformatics Center, Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China;
[4]Graduate School of the Chinese Academy of Sciences, Beijing, China;
[5]CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China;
[6]Department of Mathematics, East China Normal University, Shanghai, China;
[7]Institute of Theoretical Computing, East China Normal University, Shanghai, China;
[8]Department of Chemistry, Shanghai University, Shanghai, China.
Email: perphyliu@gmail.com; *cai_yud@yahoo.com.cn; zl_qian@yahoo.com.cn; sendru@gmail.com; dlh@picb.ac.cn; chluclear@gmail.com; jlshu@math.ecnu.edu.cn; zbzeng@sei.ecnu.edu.cn; *wclu@shu.edu.cn

Received 1 August 2009; revised 30 August 2009, accepted 3 September 2009.

ABSTRACT

In many cases, biological sequence databases contain redundant sequences that make it difficult to achieve reliable statistical analysis. Removing the redundant sequences to find all the real protein families and their representatives from a large sequences dataset is quite important in bioinformatics. The problem of removing redundant protein sequences can be modeled as finding the maximum independent set from a graph, which is a NP problem in Mathematics. This paper presents a novel program named FastCluster on the basis of mathematical graph theory. The algorithm makes an improvement to Hobohm and Sander's algorithm to generate non-redundant protein sequence sets. FastCluster uses BLAST to determine the similarity between two sequences in order to get better sequence similarity. The algorithm's performance is compared with Hobohm and Sander's algorithm and it shows that Fast- Cluster can produce a reasonable non-redundant pro- tein set and have a similarity cut-off from 0.0 to 1.0. The proposed algorithm shows its superiority in generating a larger maximal non-redundant (independent) protein set which is closer to the real result (the maximum independent set of a graph) that means all the protein families are clustered. This makes FastCluster a valuable tool for removing redundant protein sequences.

Keywords: Blast; Graph Theory; Redundant Sequences; CD-HIT

1. INTRODUCTION

Recently, with the explosion of biological sequence data, many biological sequence databases have redundant sequences which can cause problems for data analysis. These redundant sequences cannot provide valuable information for analysis but detracts from the statistical significance of interesting hits. Moreover, processing these redundant sequences often requires more time and computational resources. Removing redundant sequences is undoubtedly very helpful for performing statistical analysis and accelerating extensive database searching [1]. And it is also a way to obtain the real protein families and their representatives from a large sequences dataset. Therefore, it is necessary to develop an appropriate algorithm to remove redundant sequences from a biological sequence database.

Hobohm and Sander's algorithm is a widely used algorithm in many redundant sequence removing programs. Hobohm and Sander's algorithm was firstly introduced by U.Hobohm *et al.* of EMBL laboratory in 1992. In 1998, Lissa Holm and Chris Sander developed a program based on this algorithm to generate a non-redundant protein database NRDB90 [2]. After that, other researchers developed some programs for removing redundant sequences on the basis of Hobohm and Sander's algorithm, such as CD-HIT and PISCES.

CD-HIT [3,4] is a well-known program for processing large sequence databases efficiently. It is fast and flexible and can generate a representative set based on an incremental greedy algorithm introduced by Hobohm

and Sander [5,6]. It uses short word filtering to determine the similarity between two sequences rather than performing an actual sequence alignment. However, the results generated by short word filtering are not accurate to some degree. The lowest threshold of CD-HIT is around 40% and it is not suitable for removing redundancy on lower threshold. PISCES [7] is a public server for culling sets of protein sequences from the Protein Data Bank. It determines sequence similarity by PSI-BLAST [8] alignments which are more accurate, and it also uses a structural quality criterion to cull sequences from a sequence database.

Hobohm and Sander's algorithm has the advantage of being simple and fast. But the result set generated by this algorithm is not large enough since some non-redundant sequences may also be removed.

FastCluster, introduced in this paper, uses BLAST [9] to determine sequence similarity, which is a general sequence alignment tool and can provide better sequence similarity than word filtering. FastCluster makes improvements to Hobohm and Sander's algorithm and can get a larger non-redundant protein dataset, which means more protein families can be clustered.

2. METHODS

2.1. Hobohm and Sander's Algorithm

Hobohm and Sander's algorithm sorts all sequences by length in descending order to generate an ordered sequence set S. Then similar sequences will be put together into the same cluster. The longest sequence is added into the first cluster (initially empty), which is also the representative of the cluster, and then all the other sequences are compared with the representative. If the similarity between a sequence and the representative is above a threshold then it will be included into the same cluster as the representative's, otherwise a new cluster will be created with it as the representative. Every remaining sequence will be processed in the same way, either as the representative of a new cluster if the similarity between it and any representative is below the threshold, or included into some existing cluster if it is similar to the cluster's representative.

2.2. Graph Theory Based Algorithm

In order to make some improvements to Hobohm and Sander's algorithm, a new algorithm using maximum independent set of graph theory to generate a representative set is developed. Firstly all the sequences are clustered simply and the first sequence of each cluster is the temporary representative sequence of the cluster. Then the maximum independent set of each cluster (excluding the representative sequence) is figured out. Finally the maximum independent set of each cluster is processed and the final maximal independent set can be generated.

Based on the algorithm above, FastCluster was written in C++ and tested on Linux platform. The input to the program is a protein sequence set in FASTA format. Three output files can be generated by FastCluster. One is a FASTA file containing a list of representative proteins free from redundancy. Another output file lists the clusters and their members and the third output file contains clusters and the size of each cluster's maximal independent set. FastCluster can be downloaded from http://pcal.biosino.org/FastCluster.html.

2.3. Blast-Based Similarity Score (BSS)

FastCluster uses BLAST to make pair wise sequence alignments. The similarity score between two sequences is determined by the identical percentage of their hits (homologous sequence segments). When a sequence alignment has more than one hit, the percentage of the sum of all hits' identical is calculated to represent the overall similarity score between the two sequences.

A Python script was used to parse BLAST output and construct a BLAST-BASED SIMILARITY SCORE (BSS) matrix. An expectation value parameter 1e-3 is set to filter BLAST output in which the expectation value of each hit is smaller than 1e-3. When calculating BSS, three cases have to be considered: 1) there are no hits found; 2) there is one hit found; 3) there are more than one hit found. Formula 1 below shows how to calculate the BSS.

$$BSS = \begin{cases} 0 & (No\ hit\ \ found) \\ I/L & (One\ hit\ \ found) \\ \sum_1^n I / \sum_1^n L & (N\ \ hits\ \ found) \end{cases}$$

Formula 1: The formula to calculate BSS. 'I' is short for the length of a hit's identities, and 'L' stands for the length of a hit's length. In case of no hits found, the BSS is regarded as 0. When there is one hit in the BLAST output, the identical percentage is taken as the BSS. On the occasion of more than one hit found, the percentage of the sum of all hits' identities is taken as the BSS.

Figure 1 shows an example of how to calculate the BSS from BLAST output.

2.4. Graph Definition

A graph is a mathematical object which is composed of vertices and edges. It is usually used to represent relations between objects. Graphs can be categorized into four types: undirected graphs (or simple graphs), directed graphs, multigraphs and weighted graphs. FastCluster uses an undirected graph to represent relations between protein sequences.

2.5. Clique and Independent Set of a Graph

An undirected graph is denoted by $G = (V, E)$, in which V is the set of vertices and E are the set of edges and every edge is composed of two adjacent vertices in V.

```
# Case 1: no hits found
***** No hits found ******

# Case 2: one hit found
 Score =  180 bits (452), Expect = 1e-047
 Identities = 94/102 (92%), Positives = 94/102 (92%)

Query: 27  KTVVTSSISRFNHAETQTASATDVIGHXXXXXXXXETGNTKSLITSGLSTMSQQPRSTPA 86
           KTVVTSSISRFNHAETQTASATDVIGH        ETGNTKSLITSGLSTMSQQPRSTPA
Sbjct: 27  KTVVTSSISRFNHAETQTASATDVIGHSSSVVSVSETGNTKSLITSGLSTMSQQPRSTPA 86

Query: 87  SSIIGSSTASLEISTYVGIANGLLTNNGISVFISTVLLAIVW 128
           SSIIGSSTASLEISTYVGIANGLLTNNGISVFISTVLLAIVW
Sbjct: 87  SSIIGSSTASLEISTYVGIANGLLTNNGISVFISTVLLAIVW 128

# Case 3: more than one hit found (2 hits in this example)
 Score = 41.8 bits (96), Expect = 8e-005
 Identities = 35/163 (21%), Positives = 67/163 (40%), Gaps = 25/163 (15%)

Query: 20  PPRSEYQVLEEIGRGSFGSVRKVIHIPTKKLLVRKDIKYGHMNSKERQQLIAECSILSQL 79
           P   +Y +L+ I +G++GSV       T      K ++   M +K +   +     +  +
Sbjct: 789 PSIKDYDILKPISKGAYGSVYLARKKLTGDYFAIKVLRKSDMIAKNQVTNVKSERAIMMV 848

Query: 80  KHENIVEFYNWDFDEQKEVLYLYMEYCSRGDLSQMIKHYKQEHKYIPEKIVWGILAQLLT 139
           + +       +   + K+ L+L MEY   GDL+ +IK       Y+P++     L +++
Sbjct: 849 QSDKPYVARLFASFQNKDNLFLVMEYLPGGDLATLIKMM----GYLPDQWAKQYLTEIVV 904

Query: 140 ALYKCHYGVELPTLTTIYDRMKPPVKGKNIVIHRDLKPGNIFL 182
            +   H                    +N +IH DLKP N+ +
Sbjct: 905 GVNDMH--------------------QNGIIHHDLKPENLLI 926


 Score = 40.6 bits (93), Expect = 2e-004
 Identities = 20/57 (35%), Positives = 33/57 (57%), Gaps = 1/57 (1%)

Query: 248  YVGTPYYMSPEVLMDQPY-SPLSDIWSLGCVIFEMCSLHPPFQAKNYLELQTKIKNG 303
            + GTP Y++PE +   +   D WS+GC+ FE+   +PPF A+    +  KI +G
Sbjct: 1147 FFGTPDYLAPETIEGKGEDNKQCDWWSVGCIFFELLLGYPPFHAETPDAVFKKILSG 1203
```

Figure 1. An example of parsing BLAST output. Three cases are considered: (1) Case 1, the BSS is 0; (2) Case 2, the BSS is 94/102=0.922; (3) Case 3, the BSS is (35+20)/(163+57) =0.25.

In FastCluster, an undirected graph is defined as follows: any vertex represents a protein sequence; and if two protein sequences have a BSS above the given threshold, there is an edge between them.

A clique C of a graph G is a subset of V, and every vertex in C is adjacent to all the other vertices in C, while an independent set of a graph is a set of vertices, none of which are adjacent. A clique is said to be *maximal* if it is not the subset of any larger clique, and *maximum* if there are no larger cliques in the graph [10]. The complement of a graph G is the graph G' with the same vertex set but whose edge set consists of the edges not present in G. [http://mathworld.wolfram.com/GraphComplement.html]. By taking the complement of a graph, the maximum independent set problem is transformed into the maximum clique problem.

The concept of clique and independent set is illustrated in **Figure 2**.

2.6. Algorithm Procedure

The algorithm used in FastCluster is described in the following steps.

1) Run local BLAST for sequence set S to make pair wise alignments.
2) Parse BLAST output and construct a BSS matrix for all the sequences.
3) Sort all sequences in descending order by length and construct the first cluster with the longest sequence as its representative.
4) Align each remaining sequence in S with the existing representatives. If the BSS (from BSS matrix in step 2) between a sequence with any representative is above a given threshold, then it is included into that cluster, otherwise a new cluster starts with it as representative.
5) Repeat Step 4 until S is empty. Thus, a list of clusters for set S is generated and every cluster has a temporary representative sequence (the first sequence of that cluster).
6) Compute the *maximum* independent set for each cluster.
7) Construct an empty set R. If the maximum inde-

pendent set of the last cluster has more than or equal to 2 sequences, then every sequence in the maximum independent set is put into R, otherwise the representative of the cluster is put into R.

8) Process each remaining maximum independent set from the last one to the first. If a set has more than or equal to 2 sequences which have no edge with any sequence in R, then put these sequences in the maximum independent set into R, otherwise put the cluster's representative into R.

9) Output the final representative non-redundant set R to a file in Fasta format.

2.7. Algorithm Comparison

An algorithm comparison is shown in **Figure 3**. As shown below, S is an ordered sequence set and R is the result sequence dataset. $S(r_i)$ is composed ofsequences which are similar to representative sequence r_i. In **Figure 3**, $C(S(r_i))$ is the maximum independent set of $S(r_i)$, in which no sequence is similar to any other one, and f is a procedure to process the maximum independent set of every cluster.

Both of the algorithms share the same procedure to generate a list of clusters $S(r_1)$, $S(r_2)\ldots S(r_n)$. That is to say, both of them sort all sequences by length in decreasing order firstly to generate an ordered sequence set S and then partition them into different clusters. The difference lies in the way they generate result set R.

Hobohm and Sander's algorithm (**Figure 3(a)**) simply picks up the representative sequence of each cluster and put them together to compose result set R, while FastCluster behaves in a different way.

It searches the maximum independent set for each cluster and get n maximum independent set $C(S(r_1))$, $C(S(r_2))\ldots C(S(r_n))$. Then they are processed in a procedure f to generate the final result set R. f is such a procedure that if the number of sequences in $C(S(r_i))$ is less than 2, then r_i is added to R, otherwise every sequence in $C(S(r_i))$ is put to R on condition that the sequence has no edge with any sequence of current R which grows bigger in an incremental way.

Some $C(S(r_i))$ usually have more than 2 sequences, so after we replace r_i with $C(S(r_i))$, it is clear that R in FastCluster is larger than R in the algorithm of Hobohm and Sander. An example below is used to prove that.

2.8. An Example for Algorithm Comparison

As shown in **Figure 4**, vertices 1,2,3,4,5 and 6 represents 6 different protein sequences respectively. If two protein sequences have a BSS above the given threshold, there is an edge between them. Vertices 1,2,3,4,5 and 6

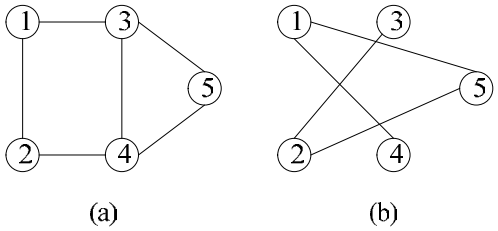


Figure 2. (a) is Graph G; (b) is its complementary graph G'. Vertices set $\{3,4,5\}$ is the maximum clique of G and the maximum independent set of graph G'. Vertices set $\{1,4\}$ or $\{1,5\}$ or $\{2,3\}$ or $\{2,5\}$ is the maximum clique of G' and is also the maximum independent set of graph G.

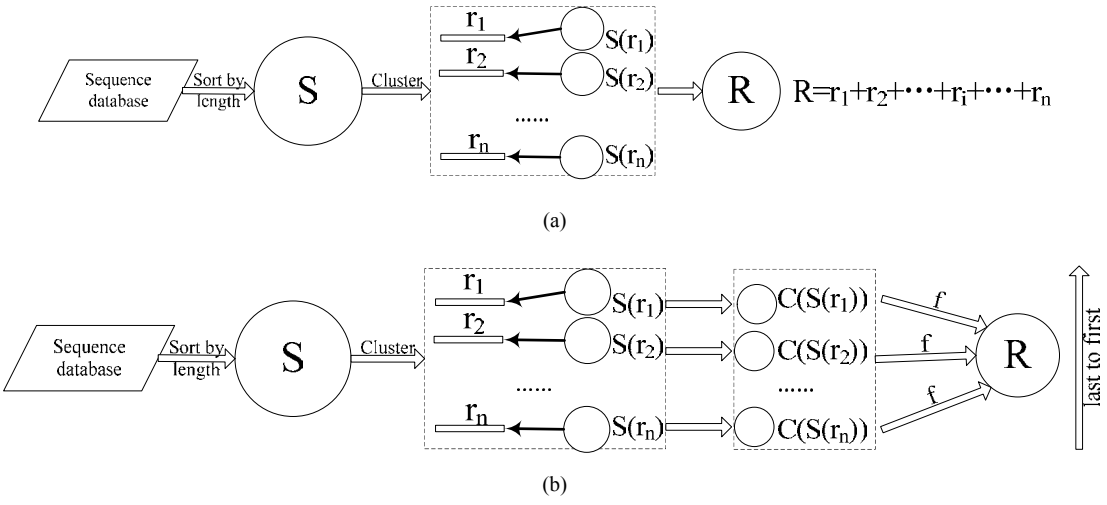


Figure 3. (a) Algorithm of Hobohm and sander; (b) Graph theory based algorithm by FastCluster.

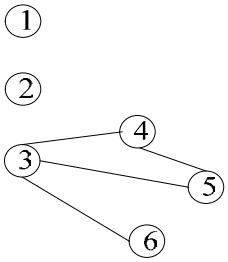


Figure 4. A graph example.

Table 1. Number of non-redundant sequences obtained using FastCluster and HSCluster.

Database	Threshold (%)	Number (FastCluster)	Number (HSCluster)
65718 Enzyme proteins	10	3167	2824
	20	3185	2845
	30	5256	4753
	40	12940	10058
	50	17275	17275
	60	25854	25020
	70	33400	32758
	80	41049	40555
	90	49986	49669

are assumed to be ordered in descending order by length. This example can be used to prove that FastCluster can produce more results than Hobohm and Sander's algorithm.

Hobohm and Sander's algorithm and FastCluster share the same way to generate three clusters, and they are $\{1\}$, $\{2\}$ and $\{3,4,5,6\}$ (**Figure 4**). But the way they generate the result set is different. According to Hobohm and Sander's algorithm, the representative vertex from each cluster is picked up to compose the result vertices $\{1,2,3\}$. While FastCluster computes the maximum independent set for each cluster and they are $\{1\}$, $\{2\}$ and $\{4,6\}$, and then it collects $\{4,6\}$, $\{2\}$ and $\{1\}$ together to compose the final result vertices $\{1,2,4,6\}$. Obviously, $\{1,2,4,6\}$ is larger than $\{1,2,3\}$. From this example, it is proved that FastCluster can produce more results than Hobohm and Sander's algorithm.

3. RESULTS AND DISCUSSION

FastCluster introduces a graph theory algorithm to remove redundant protein sequences and runs in a flexible and user-friendly way. It can also be changed to process other types of biological sequences easily.

A protein sequence set (65718 enzymes, from http://expasy.org/sprot/) was selected to test the performance of FastCluster. Another program HSCluster which implements the algorithm of Hobohm and Sander is also developed. The results generated by these two algorithms are shown in **Table 1**. It is clear that FastCluster can produce more results than HSCluster.

4. CONCLUSIONS

This paper makes an investigation on removing redundant biological sequences, which is modeled as a mathematical problem of finding a maximal independent set from a graph. Based on this model, FastCluster has made an improvement to Hobohm and Sander's algorithm in finding a larger independent set of a graph and thus generate more result sequences, which mean that more protein families can be clustered in a protein dataset. In a word, FastCluster provides an alternatively improved way to remove redundancy from a biological database and is also a computational tool to find more protein families and their representatives.

5. ACKNOWLEDGEMENTS

This research was supported by the grants from Shanghai Commission for Science and Technology (KSCX2-YW-R-112), and Shanghai Leading Academic Discipline Project (J50101).

REFERENCES

[1] G. Grillo, M. Attimonelli, S. Liuni, G. Pesole. (1996) CLEANUP: a fast computer program for removing redundancies from nucleotide sequence databases. CABIOS, **12,** 1–8.

[2] L. Holm and C. Sander. (1998) Removing near-neighbour redundancy from llarge protein sequence collections. Bioinformatics, **14,** 423–429.

[3] W. Li and A. Godzik. (2006) Cd-Hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, **22,** 1658–1659.

[4] W. Li, J L. aroszewski, A. Godzik. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics, **17,** 282–283.

[5] U. Hobohm, M. Scharf, R. Schneider, C. Sander. (1992) selection of representative protein data sets. Protein Sci, **1,** 409–417.

[6] U. Hobohm and C. Sander. (1994) Enlarged representative set of protein structures. Protein Sci, **3,** 522–524.

[7] G. Wang and R. L. Jr. Dunbrack. (2003) PISCES: a protein sequence culling server. Bioinformatics, **12,** 1589–1591.

[8] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, **25,** 3389–3402.

[9] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman. (1990) Basic local alignment search tool. J. Mol. Biol., **215,** 403–410.

[10] S. Niskanen and P. R. J. Östergård. (2003) Cliquer user's guide, Version 1.0, Communications Laboratory, Helsinki University of Technology, Espoo, Tech. Rep. T48. http://users.tkk.fi/~pat/cliquer.html.

J. Biomedical Science and Engineering, 2009, 2, 626-631 **JBiSE**
doi: 10.4236/jbise.2009.28091 Published Online December 2009 (http://www.SciRP.org/journal/jbise/).

Identification of microRNA precursors with new sequence-structure features

Ying-Jie Zhao, Qing-Shan Ni, Zheng-Zhi Wang

College of Mechatronics Engineering and Automation, National University of Defense Technology, Changsha, China.
Email: matriz@163.com; niqingshan@nudt.edu.cn

Received 7 August 2009; revised 2 September 2009; accepted 3 September 2009.

ABSTRACT

MicroRNAs are an important subclass of non-coding RNAs (ncRNA), and serve as main players into RNA interference (RNAi). Mature microRNA derived from stem-loop structure called precursor. Identification of precursor microRNA (pre-miRNA) is essential step to target microRNA in whole genome. The present work proposed 25 novel local features for identifying stem-loop structure of pre-miRNAs, which captures characteristics on both the sequence and structure. Firstly, we pulled the stem of hairpins and aligned the bases in bulges and internal loops used '—', and then counted 24 base-pairs ('AA', 'AU', …, '—G', except '——') in pulled stem (formalized by length of pulled stem) as features vector of Support Vector Machine (SVM). Performances of three classifiers with our features and different kernels trained on human data were all superior to Triplet-SVM-classifier's in positive and negative testing data sets. Moreover, we achieved higher prediction accuracy through combining 7 global sequence-structure. The result indicates validity of novel local features.

Keywords: MicroRNA; Precursor MicroRNA; Local Features; Pulled Stem; Stem-Loop; SVM

1. INTRODUCTION

MicroRNAs (miRNA) are small regulatory non-coding RNA molecule 17-25 bp long, and whose function is to down-regulate gene expression in a variety of manners, including translational repression, mRNA cleavage, and deadenylation [1,2]. More than one-third of human genes are thought to be regulated by miRNA, and these molecules represent the greatest number in eukaryotic genomes. The miRNA genes are initially transcribed as long primary transcripts (pri-miRNAs), which are then processed to the shorter, 60-120 bp stem-loop structures (called hairpin) known as miRNA precursor (pre-miRNA) [3]. Finally, the mature miRNA is separated from one of the two strands in pre-miRNA hairpin, and then by binding to a complementary target in the mRNA, which inhibits induces mRNA cleavage or translational repression [4].

Although the majority of the miRNA were identified through experimental way [5-7], computational prediction techniques become possible and necessary due to accumulation of information and data about miRNA properties [8]. All existing computational prediction methods can be classified two categories: the comparative sequence analysis approaches and the *de novo* (or *ab initio*) predictive approaches. Methods in the first category based on the assumption that miRNA genes are conserved in the primary sequences and secondary structure crossing species. Several algorithms have been developed and successfully been used for predicting miRNA in various species [9-17]. However, for a species that does not have a closely homologies species sequenced, the first category methods will not work [15]. For this reason, the secondary category methods, that are *de novo* prediction methods, have been developed to predict miRNA in single genome. Instead of evolutional information, those methods use characteristics of sequence and/or secondary structure of pre-miRNAs to achieve their purposes. The stem-loop hairpin structure is the most noticeable but not discriminative characteristic of pre-miRNAs, because a large amount of non-pre-miRNA sequences can fold themselves into pre-miRNA-like hairpins. To identify pre-miRNA hairpins, most existed methods use sets of features concerning sequence composition [17-19], topological properties of the stem-loop [17,19,20], thermodynamic stability [17,19,20], and sometimes other properties including entropy measures [19]. Xue [18] shown that local contiguous substructures of pre-miRNAs are significantly distinct with that of pseudo pre-miRNAs.

Moreover, most of *de novo* methods employed machine learning techniques to identify pre-miRNAs, such as Hidden Markov Models (HMM) [21,22], Support Vector Machine (SVM) [17-19,23], Naïve Bayes [24], Random Forest [25] and Random Walks [26]. SVM is a

supervised classification technique derived from the statistical learning theory of structural risk minimization principle, and first introduced by Vapnik [27]. It has been shown that SVM produce superior results than other supervised learning methods in a wide range of applications. Recently, they have been widely used in the bioinformatics field (include to learn the distinctive characteristics of miRNAs). SVMs have exhibited excellent generalization performance and less susceptible to over fitting than other techniques.

In this work, the novel local sequence-structure features of pre-miRNA based on "pulled" the stem-loop structure were introduced and SVM was employed as classifier to class real pre-miRNAs from pseudo ones. Those features contain information on both the sequence and structure of pre-miRNAs. Moreover, the new positive testing data set were built on updated miRNA registry database [28] with Xue's way [18]. The tests show that new method outperformed the Triplet-SVM-classifier.

2. METHODOLOGY

2.1. Features for Identify Pre-miRNA

The main difference in hairpins structure between pre-miRNA and pseudo pre-miRNAs are base pair composition in stem, the number of bulges and internal loops, and the size of bulges and internal loops. Simply, we can get sequence and structure information through counting base pair in "pulled" stem. Inspired by Xue's result, a novel local sequence-structures feature of pre-miRNAs are proposed, which based on "pulled" stem of hairpins. Firstly, the secondary structures of the pre-miRNA and the candidates are predicted with the RNAfold [29]. Then, the stems of hairpin are pulled, just as **Figure 1** shows. The bases in bulges and internal loops are aligned with '—'. Finally, counted the number of 24 base-pairs ('AA', 'AU', …, '—G', except '——', here '—' as fifth base) in pulled stem, such as **Table 1**, and normalized them with the length of pulled stem. It is noticeable that the base-pair 'AU' is different from 'UA' because of the direction of miRNA sequences (from 5' to 3'). The number of canonical base pair, that is 'AU', 'UA', 'GC', 'CG', 'GU' and 'UG', reveals the base pairs composition in stem. The number of non-canonical base pair (no gap) displays the information of internal loop. The number of gaped base pair shows the information of bulges. Another local feature is the length of pulled stem.

To improve the performance, the 7 global features used in other methods also are combined, which are numbers of base-pairs, GC content, length of sequences and central loop, free energy per nucleotide, 5' and 3' tail length.

The combined feature vector of **Figure 1** is shown as **Table 2**:

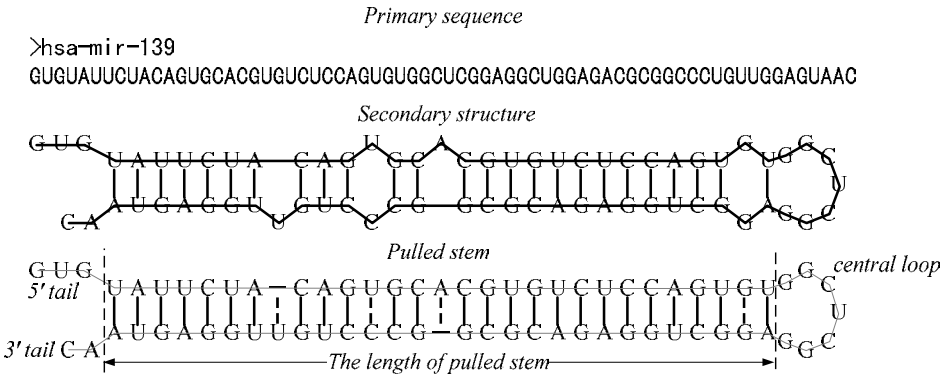


Figure 1. The example of pulled stem. The sequence is hsa-mir-139 of Homo sapiens from miRNA registry database [28].

Table 1. The statistic of 24 possible base pairs (except '——') in pulled stem in **Figure 1**.

The number of pair bases		3'				
		A	U	G	C	—
5'	A	0	4	0	0	1
	U	5	0	4	1	0
	G	0	0	1	5	0
	C	0	0	7	0	0
	—	0	1	0	0	×

Table 2. The composition of feature vector in our method.

Index	Type	Feature description	Value
1	Global	Length of central loops	7
2		Length of 5' tail	3
3		Length of 3' tail	2
4		Number of basepairs	25
5		GC content	40/68
6		Free energy of folding/length of sequence	-34.8976/68
7		Length of sequence	68
8	Local	Length of pulled stem	29
9~32		Proportion of AA/AU/…/—C pairs in pulled stem	0/0.1379/…/0

2.2. Data Sets

All verified pre-miRNAs hairpins (positive examples) come from miRNA registry database [28] in March 2009 (release 13.0), which contains 9539 reported pre-miRNA from 105 species, and 706 entries from *Homo sapiens*. The pseudo pre-miRNAs hairpins (negative examples) come from Xue' data sets [18], which contained 8494 pre-miRNA-like hairpins. SVM prediction model are trained on the same training data set of the Triplet-SVM-classifier [18], which contained 163 real human pre-miRNAs and 168 pseudo pre-miRNAs. The first testing data set (TE-C1) are 400 real human pre-miRNAs, which have no multiple loops and have low similarities each other (the sequence similarities are calculated using BLASTCLUST with S=80, L=0.5, W=16). Moreover, those sequences have low similarity with 163 training set. The CROSS-SPECIES testing set contains 3207 pre-miRNAs from 31 species. The selected criterion is same as Xue's [18] (Only the pre-miRNAs with no multiple loops are used. The pre-miRNAs that share high sequences similarities with the human pre-miRNAs are excluded to avoid biased evaluation of the SVM trained on human data. The similarity is calculated using BLASTCLUST with S=80, L=0.5, W=16). The negative testing data set (TE-C2 and TE-C3) are same as Xue's (including 1000 pseudo pre-miRNAs randomly picked up from the CODING data set, 2444 CONSERVED-HAIRPIN data set). The application of SVMs algorithms to every-day problems have been facilitated considerably by various easy-to-use software packages. Libsvm (version 2.87) [30] is used throughout this work.

2.3. Measures for Assessment

The prediction performance was evaluated by four indexes [31]: prediction accuracy (ACC), Matthews correlation coefficient (MCC), sensitivity (Sen) and selectivity (Sel).

$$ACC = \frac{tp + tn}{tp + tn + fp + fn} \times 100\% \quad (1)$$

$$MCC = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \times 100\% \quad (2)$$

$$Selectivity = \frac{tp}{tp + fp} \times 100\% \quad (3)$$

$$Sensitivity = \frac{tp}{tp + fn} \times 100\% \quad (4)$$

where, *tp* is true positive, *fp* is false positive, *tn* is true negative, and *fn* is false negative.

3. RESULTS AND DISCUSSION

To demonstrate the validity of novel local sequence-structures feature, firstly, SVM classifier are performed with only 24 novel features (not including the length of pulled stem) on all testing data sets. The feature vector of training sets are scaled to zero means and unit deviations, and the feature vector of testing sets are scaled according to the means and deviations of training sets. Three basic kernel functions (linear kernel, polynomial kernel and RBF kernel) have been tested on all testing data sets, and adjusted the parameters through grid way. The results were listed in **Table 3** (the detail results see supplemental). As a comparison, it also listed the result of Triplet-SVM-classifier (3SVM) [18]. The boldface in tables is the maximum in same row.

As shown in **Table 3**, the performance of three SVMs with 24 novel local features are better than Triplet-SVM-classifier's. The best SVM (RBF kernel) is able to predict 82% (2956 out of 3607) of all pre-miRNAs, and can identify 92% (3159 out of 3444) pseudo pre-miRNAs. In contrast, 3SVM reports 80% (2886 out of 3607) of all pre-miRNAs and 89% (3056 out of 3444) of all pseudo pre-miRNAs. This result demonstrates the validity of 24 novel local sequence-structure features for distinguishing real pre-miRNAs from pseudo ones.

To improve the performance of SVM classifier, SVM with appended 7 global features are test on all testing sets, and the result were listed in **Table 4**.

Table 3. Performance comparisons with three kernel (with 24 novel local features) and 3SVM [18].

Test set	Class	Result (true predicted/real)			
		Linear kernel	Polynomial kernel	RBF kernel	3SVM
TE-C1	pre-miRNA	**285/400**	273/400	278/400	269/400
TE-C2	pseudo	890/1000	896/1000	**904/1000**	881/1000
TE-C3	pseudo	2246/2444	2253/2444	**2255/2444**	2175/2444
CROSS-SPECIES	pre-miRNA	2668/3207	2670/3207	**2678/3207**	2597/3207
ACC		86.36	86.40	**86.73**	83.99
MCC		73.11	73.10	**73.64**	69.23
Sel		91.06	91.43	**91.72**	88.73
Sen		81.87	81.59	**81.95**	79.46

Table 4. Performance comparisons with three kernel (with 32 features) and 3SVM.

Test set	Class	Result (true predicted/real)			
		Linear kernel	Polynomial kernel	RBF kernel	3SVM
TE-C1	pre-miRNA	292/400	296/400	**303/400**	269/400
TE-C2	pseudo	953/1000	956/1000	**961/1000**	881/1000
TE-C3	pseudo	2244/2444	**2257/2444**	2240/2444	2175/2444
CROSS-SPECIES	pre-miRNA	2818/3207	2834/3207	**2850/3207**	2597/3207
ACC		89.45	89.96	**90.11**	83.99
MCC		79.12	79.94	**80.34**	69.91
Sel		92.83	**93.29**	92.94	88.73
Sen		86.22	86.78	**87.41**	79.46

Table 5. The prediction results of our method and 3SVM on cross species test sets.

Species (*ab.*)	Number	Methods (Accuracy/True predicted)		
		SVM+Our features		3SVM
		RBF+F24	RBF+F32	
Anopheles gambiae (aga)	57	93.0/53	**96.5/55**	91.2/52
Apis mellifera (ame)	53	96.2/51	**98.1/52**	94.3/50
Arabidopsis thaliana (ath)	102	96.1/98	**99.0/101**	89.2/91
Bombyx mori (bmo)	45	93.3/42	**95.6/43**	84.4/38
Bos taurus (bta)	153	82.4/126	**82.4/126**	82.4/126
Caenorhabditis briggsae (cbr)	87	93.1/81	93.1/81	**94.3/82**
Caenorhabditis elegans (cel)	144	**93.1/134**	91.0/131	86.1/124
Canis familiaris (cfa)	150	**88.7/133**	80.7/121	82.7/124
Chlamydomonas reinhardtii (cre)	37	91.9/34	**100.0/37**	94.6/35
Drosophila melanogaster (dme)	135	92.6/125	**94.8/128**	86.7/117
Drosophila pseudoobscura (dps)	66	**93.9/62**	90.9/60	87.9/58
Danio rerio (dre)	112	89.3/100	**96.4/108**	81.3/91
Epstein Barr virus (ebv)	24	**100/24**	95.8/23	91.7/22
Fugu rubripes (fru)	54	**100/54**	92.6/50	87.0/47
Gallus gallus (gga)	342	61.4/210	**81.6/279**	60.8/208
Human cytomegalovirus (hcmv)	11	72.7/8	**90.9/10**	63.6/7
Kaposi sarcoma-associated herpesvirus (kshv)	12	75.0/9	**75.0/9**	66.7/8
Monodelphis domestica (mdo)	33	**90.9/30**	84.8/28	87.9/29
Mouse gammaherpesvirus 68(mghv)	9	**88.9/8**	77.8/7	88.9/8
Macaca mulatta (mml)	211	79.1/167	**83.4/176**	82.5/174
Mus musculus (mmu)	306	73.9/226	**86.6/265**	75.8/232
Oryza sativa (osa)	189	86.2/163	**94.2/178**	88.9/168
Populus trichocarpa (ptc)	114	86.8/99	**96.5/110**	82.5/94
Pan troglodytes (ptr)	301	72.8/219	**78.4/236**	72.4/218

Rattus norvegicus (rno)	126	**95.2/120**	94.4/119	88.1/111
Schmidtea mediterranea (sme)	63	92.1/58	**95.2/60**	77.8/49
Triticum aestivum (tae)	16	93.8/15	**93.8/15**	93.8/15
Tetraodon nigroviridis (tni)	55	**96.4/53**	90.9/50	87.3/48
Vitis vinifera (vvi)	77	88.3/68	**96.1/74**	88.3/68
Xenopus tropicalis (xtr)	68	**95.6/65**	94.1/64	91.2/62
Zea mays (zma)	55	78.2/43	98.2/54	83.6/46
Total	3207	83.5/2678	**88.9/2850**	81.1/2597

We can see from **Table 4** that the performance of SVM classifier significantly increased by combining the 7 global features with 25 new local features (including the length of pulled stem). The ACC and MCC of the best SVM with 32 combined features are 90.11% and 80.34%, respectively. It indicated that the global features are important to identify real pre-miRNAs from pseudo ones.

Table 5 shows the SVM prediction on the CROSS-PECIES data sets, which contains 3207 known pre-iRNAs of 31 species. The SVM with new 24 local features and 32 combined features achieve overall accuracy of 83.5% and 88.9% on the CROSS-SPECIES data sets, respectively. The new 24 local features have better performance than Xue's local features in almost 31 species, especially for *Epstein Barr virus* (ebv) and *Fugu rubripes* (fru), our accuracy achieve 100% on those species, but Xue's accuracy is 91.7% and 87%, respectively.

4. CONCLUSIONS

In this paper, a novel local features different from Xue's [18] have been present for identifying real pre-miRNAs from pseudo ones. These features come from simply statistical on pulled stem of hairpin structure, and achieve higher accuracy than Triplet-SVM-classifier on updating testing data sets with SVM classifier. The results indicate that our method could be used as an alternative way for finding pre-miRNAs.

REFERENCES

[1] V. Ambros. (2004) The functions of animal microRNAs, Nature, **431**, 350–355.

[2] D. P. Bartel. (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function, Cell, **116**, 281–297.

[3] E. Lund, S. Guttinger, A. Calado, J. E. Dahlberg and U. Kutay. (2004) Nuclear export of microRNA precursors, Science, **303**, 95–98.

[4] L. He and G. Hannon. (2004) MicroRNAs: Small RNAs with a big role in gene regulation, Nat Rev Genet, **5**, 522–531.

[5] M. Lagos-Quintana, R. Rauhut, W. Lendeckel and T. Tuschl. (2001) Identification of novel genes coding for small expressed RNAs, Science, **294**, 853–858.

[6] N. C. Lau, L. P. Lim, E. G. Weinstein and D. P. Bartel. (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans, Science, **294**, 858–862.

[7] R. C. Lee and V. Ambros. (2001) An extensive class of small RNAs in Caenorhabditis elegans, Science, **294**, 862–864.

[8] E. Berezikov, E. Cuppen and R. H. A. Plasterk. (2006) Approaches to microRNA discovery, Nature genetics, **38**, s1–s7.

[9] L. P. Lim, M. E. Glasner, S. Yekta, C. B. Burge and D. P. Bartel. (2003) Vertebrate microRNA genes, Science, **299**, 1540.

[10] L. P. Lim, N. C. Lau, E. G. Weinstein, A. Abdelhakim, S. Yekta, M. W. Rhoades, C. B. Burge and D. P. Bartel. (2003) The microRNAs of Caenorhabditis elegans, Genes Dev, **17**, 991–1008.

[11] M. W. Jones-Rhoades and D. P. Bartel. (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA, Mol Cell, **14**, 787–799.

[12] E. Bonnet, J. Wuyts, P. Rouze and Van de Peer Y. (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences, Bioinformatics, **20**, 2911–2917.

[13] E. C. Lai, P. Tomancak, R. W. Williams and G. M. Rubin. (2003) Computational identification of Drosophila microRNA genes, Genome Biol, **4**, R42.

[14] A. Adai, C. Johnson, S. Mlotshwa, S. Archer-Evans and V. Manocha. (2005) Computational prediction of miRNAs in Arabidopsis thaliana, Genome Res, **15**, 78–91.

[15] I. Bentwich, A. Avniel, Y. Karov, R. Aharonov, S. Gilad, O. Barad, A. Barzilai, P. Einat, U. Einav, E. Meiri, E. Sharon, Y. Spector and Z. Bentwich. (2005) Identification of hundreds of conserved and nonconserved human microRNAs, Nat Genet, **37**, 766–770.

[16] X. Wang, J. Zhang, F. Li, J. Gu, T. He, X. Zhang and Y. Li. (2005) MicroRNA identification based on sequence and structure alignment, Bioinformatics, **21**, 3610–3614.

[17] J. Hertel and P. F. Stadler. (2006) Hairpins in a haystack: recognizing microRNA precursors in comparative genomics data. Bioinformatics, **22**, e197–e202.

[18] C. Xue, F. Li, T. He, G. P. Liu, Y. Li and X. Zhang. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine, BMC Bioinformatics, **6**, 310.

[19] K. L. S. Ng and S. K. Mishra. (2007) De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures, Bioinformatics, **23**, 1321–1330.

[20] T. Huang, B. Fan, M. Rothschild, Z. Hu, K. Li and S. Zhao. (2007) MiRFinder: An improved approach and

software implementation for genome-wide fast microRNA precursor scans, BMC Bioinformatics, **8**, 341.

[21] J. W. Nam, K. R. Shin, J. Han, Y. Lee, V. N. Kim and B. T. Zhang. (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure, Nucleic Acids Res, **33**, 3570–3581.

[22] S. Kadri, V. Hinman and P. V. Benos. (2009) HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models, BMC Bioinformatics, **10**, S35.

[23] A. Sewer, N. Paul, P. Landgraf, A. Aravin, S. Pfeffer, M. J. Brownstein, T. Tuschl, E. V. Nimwegen and M. Zavolan. (2005) Identification of clustered microRNAs using an ab initio prediction method, BMC Bioinformatics, **6**, 267.

[24] M. Yousef, M. Nebozhyn, H. Shatkay, S. Kanterakis, L. C. Showe and M. K. Showe. (2006) Combining multi-species genomic data for microRNA identification using a Naïve Bayes classifier, Bioinformatics, **22**, 1325–1334.

[25] P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun and Z. Lu. (2007) MiPred: Classification of real and pseudo microRNA precursors using random forest prediction model with combined features, Nucleic Acids Research, **35**, 339–344.

[26] Y. Xu, X. Zhou and W. Zhang. (2008) MicroRNA prediction with a novel ranking algorithm based on random walks, Bioinformatics, **24**, 50–58.

[27] V. N. Vapnik. (1995) The Nature of Statistical Learning Theory, Springer-Verlag, New York.

[28] S. Griffiths-Jones. (2004) The microRNA registry, Nucleic Acids Res, **32**, 109–111.

[29] I. L. Hofacker. (2003) Vienna RNA secondary structure server, Nucleic Acids Res, **31**, 3429–3431.

[30] C. C. Chang and C. J. Lin. (2001) LIBSVM: A library for support vector machines.

[31] P. P. Gardner and R. Giegerich. (2004) A comprehensive comparison of comparative RNA structure prediction approaches, BMC Bioinformatics, **5**, 140.

J. Biomedical Science and Engineering, 2009, 2, 632-636 **JBiSE**
doi: 10.4236/jbise.2009.28092 Published Online December 2009 (http://www.SciRP.org/journal/jbise/).

Normobaric hypoxia-induced brain damage in wistar rat

Ding-Yu Hu[1,2], Qin Li[1], Bo Li[3], Rong-Ji Dai[1], Li-Na Geng[1], Yu-Lin Deng[1*]

[1]School of Life Science and Technology, Beijing Institute of Technology, Beijing, China;
[2]Department of Fire Engineering, The Chinese People's Armed Police Force Academy, Langfang, Hebei, China;
[3]Beijing Vocational College of Electronic Science, Beijing, China.
Email: chem_hdy@yahoo.com.cn

Received 2 July 2009; revised 20 August 2009; accepted 28 August 2009.

ABSTRACT

The biochemical indicators of wistar rat under low oxygen concentration, such as brain water content, necrosis, lactic acid and Na^+-K^+-ATPase, was detected to evaluate normobaric hypoxia-induced brain damage and to investigate the mechanism of wistar rat brain injury. Histopathological changes in brain tissue induced by hypoxia were investigated via hematoxylin and eosin stain (HE). Hypoxia induced factor-1α (HIF-1α) expression in brain was confirmed using immunohistochemistry. The results showed that the level of lactic acid was positively correlated with the degree of hypoxia, while concentration-dependent decrease in total Na^+-K^+-ATPase activity was observed. Compared with the control group, hypoxia group had a significant difference on brain water content under severe hypoxic conditions, the rate of brain necrosis increased obviously, followed by the increase of lactic acid level and the decrease of Na^+-K^+-ATPase activity. Histopathological analysis of brain confirmed that there was **neuronal cell death in hippocampal gyrus. HIF-1α expression enhanced the hypoxia adaptation capability of the rat model through regulating the expressions of multiple genes. Lactic acid, Na^+-K^+-ATPase and HIF-1α played an important role in brain injury as a possible mechanism.**

Keywords: Hypoxia; Brain Damage; HIF-1α; Rat

1. INTRODUCTION

Hypoxia is an important pathobiological process in many diseases and causes changing of body functions easily [1,2]. Under airtight or demi-airtight environment, due to the effects of organism metabolic and impairment of gas exchange between the organization and environment, quality of the air in the cabin gets worse gradually, concentration of oxygen drops rapidly and concentration of carbon dioxide heightens rapidly. Hypoxia environment emerges quickly after appreciably long time. Hypoxia might lead to functional impairment, disturbance of consciousness, reaction dullness, retardation at action, damage of learning-memory function. Serious hypoxia might cause pathological damage or even death. Study on hypoxia mostly concentrated on hypoxic-ischemic encephalopathy (HIE) [3,4], plateau hypoxia [5,6], learning-memory [7]; therapy of various diseases induced by hypoxia and mechanisms [8,9], etc. Some studies had upgraded to cell and molecular level.

Hypoxia-induced brain damage is a hot research area of brain research. Brain damage may be induced by energy exhaustion in brain cell, overexpression of excitatory amino acids, oxygen free radical damage，apoptosis and inflammation, etc. The brain is susceptible to oxidative stress. This is due to the high content of polyunsaturated fatty acids, high rate of oxygen consumption, regional high concentrations of iron, and relatively low antioxidant capacity. These factors may predispose the premature infant, apoplexy patients to brain damage. Some of the mechanisms of hypoxia-induced brain damage were tried to be elucidated but not clearly completely nowadays. More experimental data would be needed. The investigation of the changes in energy metabolites and brain damage during hypoxia and brain hypoxic preconditioning might lead to the finding of an effective way to protect the brain from hypoxia injury.

The goal of this study was to investigate the biochemical effects of hypoxia on brain damage of rat model in the airtight cabin and provide more data for understanding the mechanism of brain damage. Brain water content, necrosis area, the levels of lactic acid and Na^+-K^+-ATPase activity were detected. HIF-1α (hypoxia induced factor-1α) expression was confirmed using immunohistochemistry method. Histopathological changes of brain in rat model induced by hypoxia were investigated via hematoxylin and eosin stain (HE). All of rat models were exposed to hypoxia for 2h at various concentrations of oxygen.

2. MATERIALS AND METHODS

2.1 Animals

Male wistar rats weighing 180–200g (provided by Institute of Laboratory Animal Science, Chinese Academy of Medical Science) were used in this study. Animals were

allowed to acclimatize for at least 7 days prior to experiment. Animals were housed at a room temperature of 22±2°C and a relative humidity of 50±10% with controlled light (12-h light/12-h dark cycle, with the light switched on at 7 a.m.). Food and water were available ad libitum. All animals received humane care in compliance with the Guide for the Care and Use of Laboratory Animals published by Beijing Administration Office of Laboratory Animal.

2.2. Normobaric Hypoxia Equipment

Animals were placed in a custom-made 16-liter plastic normobaric hypoxia chamber. Fresh soda lime was put on the bottom of chamber. O_2 and N_2 cylinders were linked with the chamber. The concentration of O_2 was controlled by infusing N_2 at flow rate of 7.5L/min. The concentrations of O_2 and CO_2 were monitored continuously respectively [10]. 18%, 15%, 12%, 10%, 8%, 6% O_2 were designed and used in the experiment respectively. Compared to hypoxia group, control group, which exposed to normobaric normoxia (21% O_2) without food and water, was set up.

2.3. Water Content of Brain Tissue

After exposed to hypoxia for 2h, rats were anesthetized with 1% pentobarbital (50 mg/kg of body weight, intraperitoneally) then killed by cervical dislocation. The brain of each rat was isolated and weighted. Water content of brain tissue detected by lyopyilization was calculated as a measure of hypoxia-induced brain damage, i.e. % water content = 100× ((wet brain weight-dry brain weight) / wet brain weight) %.

2.4. Estimation of Brain Necrosis

After exposed to hypoxia for 2h, rats were anesthetized with 1% pentobarbital (50 mg/kg of body weight, intraperitoneally), then killed by cervical dislocation. The brain of each rat was isolated and coronally sectioned into five slices (2 mm thick), and then those slices were placed in 3% 2, 3, 5-triphenyltetrazolium chloride (TTC) at 37 °C for 30 min. Those slices were dried on filter paper and weighted respectively. Total damage sections (grey section) were isolated and weighted. The relative damage percentage was estimated by calculating the brain damage area percentage by total slice (100×total damage section / total slice).

2.5. Analysis of Lactic Acid and Na^+-K^+-ATPase

The levels of lactic acid and Na^+-K^+-ATPase activity in rat model tissue were measured with kits (manufactured by Nanjing Jiancheng Bio-engineering Institute) according to the manufacturer's instruction. After exposed to hypoxia for 2h, rats were anesthetized with 1% pentobarbital (50 mg/kg of body weight, intraperitoneally), killed by cervical dislocation. The brain tissue was isolated for biochemical examinations over an ice cube. After weighting, the isolated tissue were collected in 0.1 M phosphate buffer (pH 7.4) and homogenized. The homogenates were centrifuged at 2000 r min^{-1} or 1000 r min^{-1} at 4℃ for 10 min. The supernatants were used for analysis of lactic acid and Na^+-K^+-ATPase activity respectively. The procedures of quantifying lactic acid and Na^+-K^+-ATPase activity were carried out according to the description of the kits. These indexes were evaluated by means of measurement of optical density at 530nm, 636nm with a UV spectrophotometer respectively.

2.6. Histopathological Examination

In histopathological examination, rats were exposed to 6% O_2 for 2h respectively and sacrificed by decapitation whose brains were taken out and transferred to 4% paraformaldehyde. Hippocampus sections were prepared (5 μm thick) and stained by hematoxylin and eosin. Stained sections were evaluated qualitatively (light microscopy) by an examiner blinded to experimental conditions.

2.7. HIF-1α Immunohistochemistry

After exposed to 6% O_2 for 2h, the rats were anesthetized with 1% pentobarbital (50 mg/kg of body weight, intraperitoneally) and perfused through the ascending aorta with 200 ml of 1% NaCl solution, followed with 200 ml of 4% paraformaldehyde. The brain of each rat was isolated and kept in the 4% paraformaldehyde solution until slicing. The brains were dehydrated in 10% sucrose for 1 day and then 30% sucrose solution for 2 days, till the brain sank to the bottom of the bottle. Hippocampus section were cut at 35μm thickness on a freezing microtome and processed for HIF-1α immunohistochemistry. The sections were rinsed in PBS-T (add 1ml of tween 20 to 2L of phosphate buffer saline), for three times. Then added with 3 ml of 1% H_2O_2 blocking solution at room temperature for 30 min. After reaction, the slices were rinsed and then added with 2 ml of 5% BSA solution for 20 min. Added 1:200 dilution of rabbit anti-HIF-1α antibody, and weaved in the refrigerator for 24h. The reaction was followed by adding biotin labeled monoclone mice anti-rabbit antibody. The slices were rinsed, soaked in the SABC solution for 30 min. Then DAB solution was used to stain for 10 min. The sections were dehydrated in ascending alcohol concentrations, cleared and covered in xylene. Rabbit anti-HIF-1α antibody, biotin labeled monoclone mice anti-rabbit antibody, SABC and DAB solution were purchased from Boster Biological Technology, LTD (Wuhan, Hubei, China).

2.8. Statistical Analysis

All results were expressed as mean±SEM. Statistical analysis of data was performed by applying one-way analysis of variance (ANOVA) followed by Tukey test. The p values less than 0.05 were considered as statistically significant difference.

3. RESULTS

3.1. Water Content of Brain Tissue

The increasing of water content of brain tissue was induced by hypoxia. While exposed to 8%, 6% O_2, water contents in the rat brain were the highest. The volumes of brain water content were 77.8% and 77.9% respectively. When exposed in 10% O_2 or more, the brain water contents were located near 76.8%. Compared with 21% O_2 group, there was significant difference in the brain water content of 8%, 6% O_2 group (**Figure 1**, $p<0.05$).

3.2. Brain Necrosis

Reduction reaction of TTC started under the effect of chondriosome succinate dehydrogenase in competent cell and then red stabile and indiffusible substance would be formed, while reduction reaction of TTC did not start in infarction section and the color of the section would be grey. The method might be used to evaluate the necrosis of brain. Results showed that brain infarction ratio increased obviously under serious hypoxia condition (6% O_2). For the brain infarction ratio, there was a significant difference between group 21% O_2 and group 6% O_2 (**Figure 2**, $p<0.05$).

3.3. Lactic Acid

When the rat was exposed to 10% O_2 or more, the level of lactic acid in 10% brain homogenate tissue changed from 1.23 mmol·L^{-1} to 1.26 mmol·L^{-1}. while the rat model group 6% O_2 was exposed to 8% O_2 or less, the level of lactic acid increased significantly, it changed from 1.26 mmol·L^{-1} to 4.2 mmol·L^{-1} (**Figure 3**). The more serious the degree of hypoxia was, the higher the lactic acid level was. It increased obviously at 10% O_2 or less. Lactic acid was accumulated in the brain sharply at 6% O_2.

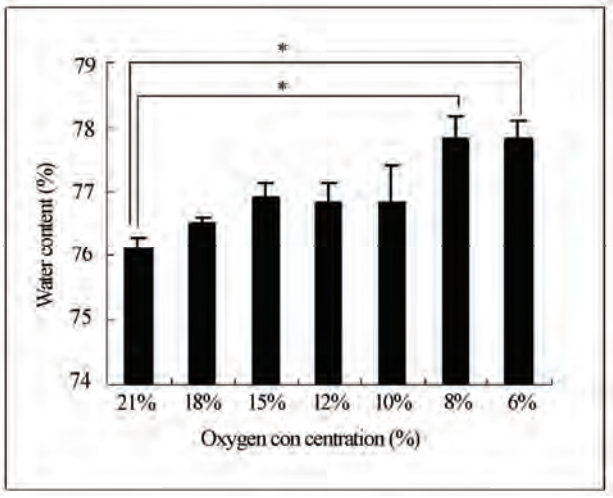


Figure 1. Brain water content at various concentrations of oxygen. (* $p<0.05$ compared with the group 21%).

3.4. Na^+-K^+-ATPase

Hyperactivity of Na^+-K^+-ATPase was large enough to maintain the ion homeostasis in the range of 21% ~12% O_2. The level of Na^+-K^+-ATPase activity decreased significantly at serious hypoxia (≤8% O_2) which induced the function disorder of cell due to cell oedema and atrophia (**Figure 4**).

3.5. Histopathological Examination

After exposed to 6% O_2 for 2h, the rat model was killed and the brain was taken out. Hippocampus sections were prepared for histopathological examination. Compared with the control group, histopathological analysis of brain confirmed that there was neuronal cell death in

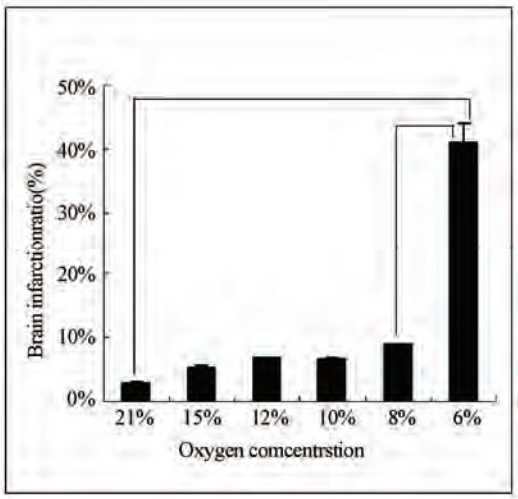


Figure 2. Brain necrosis at various concentrations of oxygen. (* $p<0.05$ compared with the group 6%).

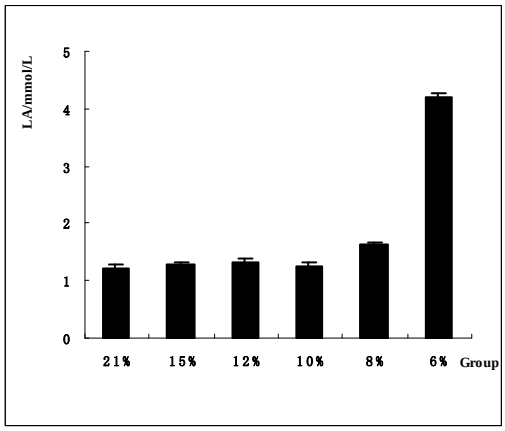


Figure 3. Lactic acid levels in rat models at various concentrations of oxygen.

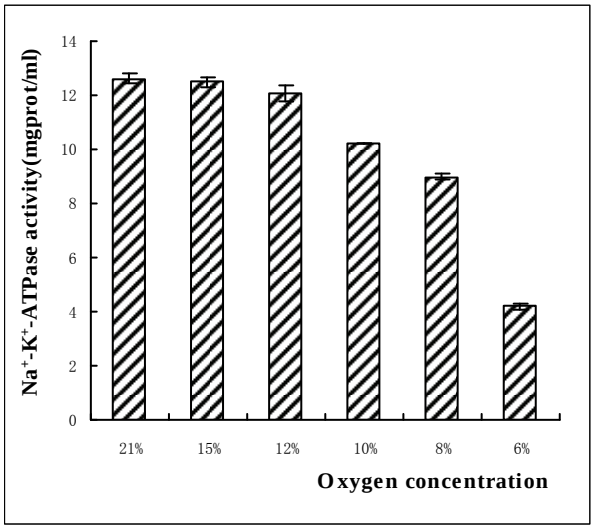


Figure 4. Na^+-K^+-ATPase activity in rat models at various concentrations of oxygen.

hippocampal gyrus of hypoxia group (**Figure 5A**). Previous study indicated that there were neuronal cells death in neuropile and cortex. When the concentration of oxygen was 8%, 10% separately, histopathological analysis of brain showed that there was no cell death in brain (**Figure 5B** and **Figure 5C**).

3.6. HIF-1α Immunohistochemistry

Compared to the control group, expression of HIF-1α in rat hippocampus section was obvious in **Figure 6**. Experiment confirmed that dilution ratio was an important factor to complete the HIF-1α immunohistochemistry successfully because of the instability and low abundance of HIF-1α.

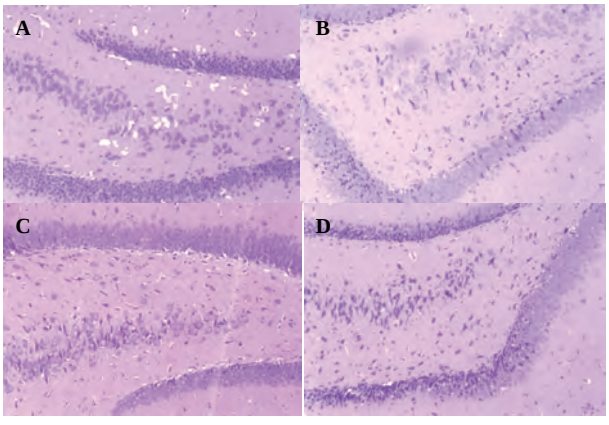


Figure 5. Representative photographs of histopathological examnation (HE) in the rat hippocampus under hypoxia conditions. (A: exposed to 6% O_2 for 2h; B: exposed to 8% O_2 for 2h; C: exposed to 10% O_2 for 2h; D: exposed to normobaric normoxia for 2h without food and water).

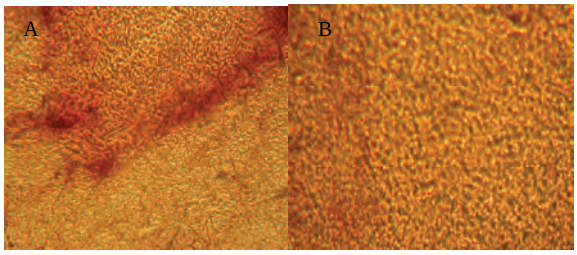


Figure 6. HIF-1α immunohistochemistry photomicrographs of rat hippocampus section (dilution 1:200, A: exposed to 6% O_2 for 2h; B: exposed to normobaric normoxia for 2h without food and water).

4. DISCUSSION

Extreme hypoxia would cause acidosis easily and lead to tissue oedema and cell death [11,12]. Results showed that the levels of lactic acid increased significantly under the severe hypoxia environment (6%~8% O_2). It indicated that hypoxia led to anaerobic metabolism and metabolic acidosis. Pyruvic acid from glycolysis was converted by anaerobic metabolism to lactic acid mostly.

Na^+-K^+-ATPase would transport Na^+ ions and K^+ ions against their concentration gradient. The decrease of Na^+-K^+-ATPase activity showed that the loss of ion homeostasis occurred. It is generally believed that loss of ion homeostasis played an important role in the pathogenesis of brain cell damage. Extreme hypoxia-induced perturbation of ion homeostasis led to the intracellular accumulation of sodium and calcium ions, followed by subsequent activation of proteases and phospholipases and the formation of oxygen and nitrogen free radicals [13]. Consequently, the events would cause changing of functional and structural including cerebral edema, eventually lead to cell death. Under hypoxia condition, energy exhaustion would induce inhibition of Na^+-K^+-ATPase activity and accumulation of lactic acid, followed by acidosis and cell apoptosis.

HIF-1 is a transcriptional activator that regulates the expression of multiple genes during continuous hypoxia [14]. HIF-1 is composed of a constitutively expressed HIF-1β and O_2 regulated HIF-1α subunit. Previous studies confirmed that HIF-1 played a general role in coordinating adaptive physiologic responses to hypoxia at the level transcription. HIF-1α has also been implicated in the coordinate transcriptional activation of genes encoding glycolytic enzymes in hypoxia cells, which provide an alternative means of energy product under conditions of limited oxygen availability [15,16]. The results of immunohistochemistry confirmed that HIF-1α was induced by hypoxia at 6% O_2. Expression of HIF-1α enhanced the hypoxia adaptation capability of the rat model through regulating the expression of multiple genes.

5. CONCLUSIONS

The values of lactic acid are positively correlated with the degree of hypoxia, while total Na^+-K^+-ATPase activity shows a concentration-dependent decrease. Compared with the control group; hypoxia group has a significant difference in brain water content under severe hypoxia condition. The area of brain necrosis increases sharply followed by the increase of lactic acid level and the decrease of Na^+-K^+-ATPase activity, neuronal cell death and HIF-1 expression appear in hippocampal gyrus obviously. Lactic acid, Na^+-K^+-ATPase and HIF-1α played an important role as a possible mechanism in brain injury.

6. ACKNOWLEDGEMENT

This work was supported by Commission of Science Technology and Industry for National Defense (Grant No. A2220060042) and the National Natural Science Foundation of China (Grant No. 20705005).

REFERENCES

[1] I. L. Kanstrup, T. D.Poulsen and J. M. Hansen. (1999) Blood pressured and plasma catecholamine in acute and prolonged hypoxia effects of local hypothermia, Apple Phys, **87(6)**, 2053–8.

[2] Q. H. Chen. (2001) The changes of function and morphology of pulmonary arterial vessels in the pika at high altitude, Chin. J. Appl Phys, **17(2)**, 178–81.

[3] W. J. Xia. (2005) The effects of hematopoietic growth factors and tanshinone ⅡA on neuro-protection, Doctor Dissertation, The Chinese University of Hong Kong, Hong kong, China.

[4] M. Christiane and H. Brahimi. (2007) Harnessing the hypoxia-inducible factor in cancer and ischemic disease, Biochem. Pharmacol, **73**, 450–457.

[5] S. D. Aramjit and K. Manoj. (2007) cDNA cloning, gene organization and variant specific expression of HIF-1α in high altitude yak (Bos grumiens), Gene, **386(1–2)**, 73–80.

[6] S. Fau, C. Po, B. Gillet, *et al*. (2007) Effect of the reperfusion after cerebral ischemia in neonatal rats using MRI monitoring, Experimental Neurology, **208(2)**, 297–304.

[7] L. Liu, T. van Groen, Inga Kadish, *et al*. (2009) DNA methylation impacts on learning and memory in aging, Neurobiology of Aging, **30(4)**, 549–560.

[8] G. D. Funk, A. G. Huxtable and A. R. Lorier. (2008) ATP in central respiratory control: A three-part signaling system, Respiratory Physiology & Neurobiology, **164(1–2)**, 131–142.

[9] M. L. Peter. (2008) Opioidergic and dopaminergic modulation of respiration, Respiratory Physiology & Neurobiology, **164(1–2)**, 160–167.

[10] J. E. Rice, R. C. Vannucci and J. B. Brierley. (1981) The influence of immaturity on hypoxia-ischemia brain damage in the rat, Ann. Neurol, **9**, 131–141.

[11] W. M. Bernhardt, C. Warnecke, C. Willam, *et al*. (2007) Organ protection by hypoxia and hypoxia-inducible factors, Methods Enzymol, **435**, 219, 221–245.

[12] O. Marta, S. Monika and A. Jan. (2008) Regulation of pH in the mammalian central nervous system under normal and pathological conditions: Facts and hypotheses, Neurochem Int, **52(6)**, 905–919.

[13] D. B. Kintner, Y. Wang and D. Sun. (2007) Role of membrane ion transport proteins in cerebral ischemic damage, Front. Biosci, **12**, 762–770.

[14] G. L. Semenza. (2004) O_2-regulated gene expression: Transcriptional control of cardiorespiratory physiology by HIF-1, J. Appl. Physiol, **96**, 1173–1177.

[15] M. W. Charles, B. Greg and L. S. Gregg. (1996) In vivo expression of mRNAs encoding hypoxia-inducible factor 1, Biochem Biophys Res Commun, **225**, 485–488.

[16] G. L. Semenza. (2001) HIF-1 and mechanisms of hypoxia sensing, Cell. Biology, **13**, 167–171.

J. Biomedical Science and Engineering, 2009, 2, 637-643 **JBiSE**
doi: 10.4236/jbise.2009.28093 Published Online December 2009 (http://www.SciRP.org/journal/jbise/).

Application of SOM neural network in clustering

Soroor Behbahani[1], Ali Moti Nasrabadi[2]

[1]Biomedical Engineering Department, Science and Research Branch, Islamic Azad University, Tehran, Iran;
[2]Biomedical Engineering Department, Faculty of Engineering, Shahed University, Tehran, Iran.
Email: soroor_behbahani@yahoo.com; a_m_nasrabadi@yahoo.com

Received 11 June 2009; revised 29 June 2009; accepted 27 July 2009.

ABSTRACT

The Self-Organizing Map (SOM) is an unsupervised neural network algorithm that projects high-dimensional data onto a two-dimensional map. The projection preserves the topology of the data so that similar data items will be mapped to nearby locations on the map. One of the SOM neural network's applications is clustering of animals due their features. In this paper we produce an experiment to analyze the SOM in clustering different species of animals.

Keywords: SOM Neural Network; Feature; Clustering; Animal

1. INTRODUCTION

The Self-Organizing Map (SOM) is a fairly well-known neural network and indeed one of the most popular unsupervised learning algorithms. Since its invention by Finnish Professor Teuvo Kohonen in the early 1980s, more than 4000 research articles have been published on the algorithm, its visualization and applications. The maps comprehensively visualize natural groupings and relationships in the data and have been successfully applied in a broad spectrum of research areas ranging from speech recognition to financial analysis. The Self-organizing Map performs a non-linear projection of multidimensional data onto a two-dimensional display. The mapping is topology-preserving, meaning that the more alike two data samples are in the input space, the closer they will appear together on the final map. The SOM belongs to the class of Neural Network algorithms. This is a group of algorithms based on analogies to the neural structures of the brain. The SOM in particular was inspired by an interesting phenomenon: as physicians have discovered, some areas of brain tissue can be ordered according to an input signal. Basically, the SOM is a computer program simulating this biological ordering process. Applied to electronic datasets, the algorithm is capable of producing a map that shows similar input data items appearing close to each other. There are numerous applications involving the SOM algorithm but the most widespread use is the identification and visualization of natural groupings in the data. The process of finding similar items is generally referred to as clustering. Compared to the *k-means* clustering algorithm, the SOM exemplifies a robust and structured self-organizing neural networks are based on the principle of transforming a set of p-variate observations into a spatial representation of smaller dimensionality, which may allow a more effective visualization of correlations in the original data [4].

2. SELF-ORGANIZING MAP

The Self-Organizing Map belongs to the class of unsupervised and competitive learning algorithms. It is a sheet-like neural network, with nodes arranged as a regular, usually two-dimensional grid. As explained in the previous section on Neural Networks, we usually think of the node connections as being associated with a vector of weights. In the case of Self-Organizing Maps, it is easier to think of each node as being directly associated with a weight vector.

The items in the input data set are assumed to be in a vector format. If n is the dimension of the input space, then every node on the map grid holds an n-dimensional vector of weights:

$$mi = [mi1, mi2, mi3, . , min] \quad (1)$$

The basic principle of the Self-Organizing Map is to adjust these weight vectors until the map represents a picture of the input data set. Since the number of map nodes is significantly smaller than the number of items in the dataset, it is needless to say that it is impossible to represent every input item from the data space on the map. Rather, the objective is to achieve a configuration in which the distribution of the data is reflected and the most important metric relationships are preserved. In particular, we are interested in obtaining a correlation between the similarity of items in the dataset and the distance of their most alike representatives on the map. In other words, items that are similar in the input space should map to nearby nodes on the grid [4].

2.1. Image's Characteristics

To represent the 3D image of 12 lead ECG, three axes for time, temporal and spatial are needed witch temporal axis represented the time domain of the cardiac signal and the spatial axis represented the locations of the limb and thoracic leads. The data axis is represented two extracted features of cardiac signal contains amplitude and wavelet coefficients. 6 leads are used to represent the image obtained by thoracic leads and 6 leads of 12 are used to represent the image obtained by limb leads.

In order to determine the information between consecutive leads in the spatial axis, an interpolation technique was used to witch could cause to homogeneousness of the image.

3. AN EXAMPLE OF SOM NEURAL NETWORK APPLICATION

More researches are performing in the field of SOM neural network applications in last two decades. One of the most important and famous examples of this application is clustering of animals due their features.

General features are using in this example based on the Kohonen animal data base (**Table 1**).

But the fact is that, these features are not sufficient for different species of animals.

In previous experiments, it had been assumed that there were only one species for each animal, whereas there may be exist more than 10 species for each special animal. So, for analyzing the ability of SOM neural network we perform a new experiment and assumed more than one species for them and increase the number of features to invent better separability. These features consist of geographical dispersion, nourishing and habitat, etc (**Table 2**).

The SOM size in this research is 7*7 and the initial weights are selected randomly.

Although there are 3 animals that are not settle in right location in SOM map, and selected wrong neurons, the results shows that extracted features could well separate the different species of animals. This result shows that the selected features for these 3 animals have not sufficient ability to separate them. This problem could be solved by adding extra features or choosing the features with more precise. One of the most important points in neural networks is the method of features extraction, but increasing the number of features could not always be the best solution for approving the results, because sometimes increasing the features lead to derangement in network. Another reason of bad result in neural networks relates to number of inputs. Increasing the number of inputs (animal species) leads to spreading the SOM size and could decrease the ability of it, because there would be more correlation between inputs, so the statistic of error will be increased.

4. RESULTS

Choosing suitable features for separating animal's species lead to good results of SOM neural network .There were some similarity between some of the animal's feature in Kohonen data base. For example the features of Goose and Owl, as well as, horse and zebra are exactly the same. And this similarity leads to wrong results in clustering of these animals. Although there are some errors, in this new experiment, these errors occurred between different species of one animal not between different animals. So the more similarity between animal's species, the more errors will occur.

Table 1. The animal data set.

Animal	Dove	Hen	Duck	Goose	Owl	Hawk	Eagle	Fox	Dog	Wolf	Cat	Tiger	Lion	Horse	Zebra	Cow
is																
Small	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0
Medium	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
Big	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
has																
Two kgs	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
Four legs	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
Hair	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
Hooves	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
Mane	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
Feathers	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
likes to																
Hunt	0	0	0	0	1	1	1	1	0	1	1	1	1	0	0	0
Run	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0
Fly	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0
Swim	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0

Figure 1. SOM neural network resule.

Table 2. Increasing number of features and animals species.

	Redfox	Afghanfox	EagleOw1	BrownfishOw1	LongearedOw1	ShortearedOw1	BarnOw1
Small	0	0	0	0	0	0.875	0.475
Medium	0.4	0.5	0.1	0.43	0.21	0	0
Big	0	0	0	0	0	0	0
2 leg	0	0	1	1	1	1	1
4 leg	1	1	0	0	0	0	0
Hair	1	1	0	0	0	0	0
Hoove	0	0	0	0	0	0	0
Mane	0	0	0	0	0	0	0
Feathers	0	0	1	1	1	1	1
Hunt	1	1	0.95	1	1	1	1
Run	1	1	0	0	0	0	0
Fly	0	0	1	1	1	1	1
Swim	0	0	0	0	0	0	0
Asia	1	1	1	1	1	1	1
Africa	1	0	0.4	0	0	0	0.7
Us	1	0	0.4	1	1	0	1
Europe	0	1	0	0	0	1	0
Mountainous	1	0	1	0	0	0	0.47
Plain	1	1	0	1	0	0	0
River	0	0	1	1	0	1	0
Jungle	1	1	0	0	1	0	0
Domestic	0	0	0	0	0	0	1
Carnivorous	1	1	1	1	1	1	1
Herbivorous	0	0	0	0	0	0	0
Frugivorous	1	0	0	0	0	0	0
Egg	0	0	1	1	1	1	1
Milk	1	1	0	0	0	0	0
Colour variation	0.5	0.45	0.2	0.25	0.2	0.13	0.3

	SakerFalcon	LannerFalcon	PeregrineFalcon	OspreyEagle	BootedEagle	BonelliEagle	SpottedEagle
Small	0	0	0.95	0	0	0	0
Medium	0.083	0.33	0.13	0.08	0.08	0.41	0.52
Big	0	0	0	0	0	0	0
2 leg	1	1	1	1	1	1	1
4 leg	0	0	0	0	0	0	0
Hair	0	0	0	0	0	0	0
Hoove	0	0	0	0	0	0	0
Mane	0	0	0	0	0.066	0	0
Feathers	1	1	1	1	1	1	1
Hunt	1	0.9	0.91	1	1	1	1
Run	0	0	0	0	0	0	0
Fly	0.8	1	1	1	1	1	1
Swim	0	0	0	0	0	0	0
Asia	1	1	0.98	1	1	1	0.91
Africa	0.86	0	1	0	0.243	0	1
Us	0	0.5	1	0.5	0.7	0.61	0
Europe	1	0.95	0	0	1	1	1
Mountainous	0.5	0.27	1	1	0	1	0.5
Plain	1	1	0	0	0	0.5	0
River	0	0	0.96	0	0	0	1
Jungle	0	0.032	0	0	1	0	1
Domestic	0	0	0	0	0	0	0
Carnivorous	1	1	1	1	1	0.95	0.89
Herbivorous	0	0	0	0	0	0	0
Frugivorous	0	0	0	0	0	0	0
Egg	1	1	1	1	1	0.92	1
Milk	0	0	0	0	0	0	0
Colour variation	0.24	0.54	0.35	0.24	0.4	0.33	0.21

	LesserspottedEagle	ImperialEagle	GoldenEagle	RedbreastedGoose	GreylagGoose	WithefrontedGoose	SantherbertDog
Small	0	0	0	0	0	0	0
Medium	0.58	0.61	0.71	0.23	0.68	0.31	0.76
Big	0	0	0	0	0	0	0
2 leg	1	1	1	1	1	1	0
4 leg	0	0	0	0	0	0	1
Hair	0	0	0	0	0	0	1
Hoove	0	0	0	0	0	0	0
Mane	0	0	0	0	0	0	0
Feathers	1	1	1	1	1	0.898	0
Hunt	1	0.99	1	0	0	0	0
Run	0	0	0	0	0.0363	0	1
Fly	0.95	1	1	1	1	1	0
Swim	0	0	0	1	0.99	0.85	0
Asia	1	0	0.87	1	0	0	0
Africa	0	1	1	0	0.9	0	1
Us	0.43	1	0	0	1	0	0
Europe	0.4	0	1	0.441	0	0	0.515
Mountainous	0	0.5	1	0	0.85	1	0
Plain	0.35	1	0	1	0	0.2	0
River	0	1	0	1	1	1	0.9
Jungle	0.9	0	0.03	0	0	0	0
Domestic	0	0	0	0	1	0.5	0
Carnivorous	1	1	1	0	0	0	0.858
Herbivorous	0	0	0	0	0	0	0.02
Frugivorous	0	0	0	0.98	1	1	0.04
Egg	1	1	1	1	1	1	0
Milk	0	0	0	0	0	0.021	0.88
Colour variation	0.4	0.16	0.13	0.5	0.1	0.15	0.12

	PointerDog	WoodPigeon	StockDove	RockDove	CollaredDave	Wolf	Kaiot
Small	0	1	0.8	0.8	0.675	0	0
Medium	0.31	0.211	0.02	0	0	0.48	0.825
Big	0	0	0	0	0	0	0
2 leg	0	1	1	1	1	0	0
4 leg	1	0	0	0	0	1	1
Hair	1	0	0	0	0	1	1
Hoove	0	0	0	0	0	0	0
Mane	0	0	0	0	0	0	0
Feathers	0	1	1	1	1	0	0
Hunt	0	0	0	0	0	1	1
Run	0.89	0	0	0	0	0.88	0.88
Fly	0	1	1	0.965	1	0	0
Swim	0.88	1	0	0	0	0	0
Asia	0	1	0.91	0.87	0.92	1	1
Africa	0.66	1	0.5	0.11	0.5	0	0
Us	1	0	1	1	0.19	0.17	1
Europe	0.79	1	1	0	1	1	0
Mountainous	0	0.2	0	0.95	0.21	0	0.4
Plain	0	1	1	0	0	1	1
River	0.78	0	1	1	0.2	0.11	0
Jungle	0.023	1	0	0	0	1	0
Domestic	0	1	1	1	0.95	0.032	0
Carnivorous	1	0	0	0	0	1	1
Herbivorous	0	0	0	0	0.1	0	0
Frugivorous	0.1	1	1	1	1	0	0
Egg	0	0	1	1	1	1	0
Milk	0	1	0	0	0	0	1
Colour variation	0.54	0.67	0.2	0.3	0.47	0.31	0.21

	Tiger	Lion	Horse	IranianZebra	Zebra	MarbledCat	ChinchilaCat
Small	0	0	0	0	0	0	0
Medium	0	0	0.85	0.62	0.967	0.45	0.16
Big	0.88	1	1	0	0	0	0
2 leg	0	0	0	0	0	0	0
4 leg	1	1	1	1	1	1	1
Hair	1	1	1	1	1	0	0
Hoove	0	0	1	1	1	0	0
Mane	0	1	1	1	1	0	0
Feathers	0	0	0	0	0	0	0
Hunt	1	0.89	0	0	0	1	0
Run	1	1	0.84	0.87	1	0.99	0.88
Fly	0	0	0	0	0	0	0.96
Swim	0	0	0	0	0	1	1
Asia	1	1	1	1	0.78	0	0
Africa	1	0	0.416	0	0	1	1
Us	0.9	0.3	0	0.91	1	0	0
Europe	1	0.78	0.445	0	0.93	0.35	0.032
Mountainous	1	0.95	0.985	0	0	1	1
Plain	0	1	0.95	0.35	0.64	0.033	0
River	0	0	0	0.019	0	0	1
Jungle	1	1	0.5	1	1	0	0
Domestic	0	0	1	0	0	0.91	1
Carnivorous	1	1	0	0	0	1	1
Herbivorous	0	0	1	1	0.8	0	0
Frugivorous	0.21	0.0033	0	0.1	0	0.2	0
Egg	0	0	0	0	0	0	0
Milk	1	1	1	1	1	1	1
Colour variation	0.3	0	0.92	0.54	0.75	0.21	0.3

	AsiangolodenCat	BlackearCat	CaucasianblackGrouse	MallardDuck	GadwellDuck	WigeonDuck	PanitalDuck
Small	0	0.818	1	0	0	0	0
Medium	0.55	0.8	0.03	0.28	0.16	0.083	0.25
Big	0.01	0	0	0	0	0	0
2leg	0	0	1	1	1	1	1
4leg	1	1	0	0	0	0	0
Hair	1	1	0	0	0	0	0
Hoove	0	0	0	0	0	0	0
Mane	0	0	0	0	0	0	0
Feathers	0	0	1	1	1	1	1
Hunt	1	1	0	0	0	0	0
Run	1	1	0	0	0	0	0
Fly	0	0	1	1	1	0.95	1
Swim	0	0	0	0	0	0	0
Asia	1	1	1	1	0.91	1	1
Africa	0	1	0	0.8	0.4	0	1
Us	0.33	0	0.5	1	0	1	0
Europe	0	1	0	0	1	0	0
Mountainous	0	1	0.99	1	0.45	0	0
Plain	1	0	0.5	0	0	1	0
River	1	0	0	1	0.3	1	1
Jungle	1	1	0.5	0	0	0	0
Domestic	1	0	0.95	0	0.5	0	0.75
Carnivorous	1	1	0	0	0	0	0
Herbivorous	0	0	0	0	0	0	0
Frugivorous	0	0	1	1	1	1	1
Egg	0	0	1	1	1	1	1
Milk	1	1	0	0	0	0	0
Colour variation	0.4	0.7	0.1	0.21	0.3	0.24	0.5

	GarganeyDuck	MarbledtealDuck	BlackBear	GrizliBear	PandaBear	OrangotanMonky	ShampainMonky
Small	0	0.95	0	0	0	0	0
Medium	0.191	0	0	0	0	0	0.65
Big	0	0	0.01	0.28	0.17	0.45	0.23
2leg	0	1	0	0	0	1	1
4leg	1	0	1	1	1	0	0
Hair	0	0	1	1	1	1	1
Hoove	0	0	0	0	0	0	0
Mane	0	0	0	0	0	0	0
Feathers	0	1	0	0	0	0	0
Hunt	1	0	1	1	1	1	1
Run	0	0	1	1	1	0	0
Fly	0	0.95	0	0	0	0	0
Swim	0	0	0	0	0	0	0
Asia	1	1	0	1	0.85	0.95	0
Africa	1	0.033	0	0	0	1	0.89
Us	0	0	1	1	0	0	0.045
Europe	1	1	0	1	1	1	0
Mountainous	0	0	0	0	1	0.21	0
Plain	0	0.5	0	1	0	0	0.9
River	1	1	1	1	0	0.89	0
Jungle	1	1	1	1	1	1	0.8
Domestic	0	0	0	0	0	0	0
Carnivorous	0	0	1	1	1	0	0.14
Herbivorous	0	0	0	1	1	1	0
Frugivorous	1	1	0	1	0	1	1
Egg	1	1	1	0	0	0	0
Milk	0	0	0	1	1	0.87	1
Colour variation	0.4	0.3	0.4	0.13	0.36	0.04	0.351

5. CONCLUSIONS

SOM is a highly useful multivariate visualization method that allows the multidimensional data to be displayed as a 2-dimensional map. This is the main advantage of SOM. The map units clustering makes it easy to observe similarities in the data. Through our experiment, we demonstrated that the possibility of quick observation of relationship between component (feature) and the class as well as the relationship among different component (feature) of the dataset from the visualization of a dataset. SOM is also capable of handling several types of classification problems while providing a useful, interactive, and intelligible summary of the data.

However, SOM also has some disadvantages. For example, adjacent map units point to adjacent input data vector, so sometimes distortions are possible because high dimensional topography can not always be represented in 2D. To avoid such phenomenon, training rate and the neighborhood radius should not be reduced too quickly.

Hence, SOM usually need many iterations of training. And SOM also does not provide an estimation of such map distortion. Alternatives to the SOM have been developed in order to overcome the theoretical problems and to enable probabilistic analysis.

Current research showed a simple application of SOM neural network in clustering. This method can be used in many applications that need classification and one of them could be disease clustering. As seen in current research, we used fuzzy method to determine the features of each animal.

Similar to this research, we can determine the features of diseases. This method could help the physician in their diagnosis. We can use the sign of diseases as the input of SOM neural network. As some classes of disease have similar symptoms the SOM neural network can show a limitation of neighbor diseases that have such symptoms, so the physician can focus on them to diagnose the patient's disease with more accuracy. Fuzzy features can increase the ability of SOM neural network if they choose carefully with more accuracy and of course it need some trail and error methods to find a rule to relate a membership function to each disease and its symptoms.

REFERENCES

[1] A. Forti, (2006) Growing hierarchical tree SOM: An unsupervised neural network with dynamic topology, , Gian Luca Foresti, Neural Networks, **19**, 1568–1580.

[2] S. Haykin, (1999) Neural networks a comprehensive foundation (2nd ed.), Prentice Hall.

[3] R. G. Adams, K. Butchart and N. Davey, (1999) Hierarchical classification with a competitive evolutionary neural tree, Neural Networks, **12**, 541–551.

[4] J. Li, Information visualization of self organizing maps.

J. Biomedical Science and Engineering, 2009, 2, 644-650 **JBiSE**
doi: 10.4236/jbise.2009.28094 Published Online December 2009 (http://www.SciRP.org/journal/jbise/).

Folding rate prediction using complex network analysis for proteins with two- and three-state folding kinetics

Hai-Yan Li[1], Ji-Hua Wang[1]

[1]Key Lab of Biophysics in Universities of Shandong, Dezhou University, Shandong, China.
Email: tianwaifeixian78@163.com , jhwyh@yahoo.com.cn

Received 5 September 2009; revised 9 October 2009; accepted 10 October 2009.

ABSTRACT

It is a challenging task to investigate the different influence of long-range and short-range interactions on two-state and three-state folding kinetics of protein. The networks of the 30 two-state proteins and 15 three-state proteins were constructed by complex networks analysis at three length scales: Protein Contact Networks, Long-range Interaction Networks and Short-range Interaction Networks. To uncover the relationship between structural properties and folding kinetics of the proteins, the correlations of protein network parameters with protein folding rate and topology parameters contact order were analyzed. The results show that Protein Contact Networks and Short-range Interaction Networks (for both two-state and three-state proteins) exhibit the "small-world" property and Long-range Interaction networks indicate "scale-free" behavior. Our results further indicate that all Protein Contact Networks and Short-range Interaction networks are assortative type. While some of Long-range Interaction Networks are of assortative type, the others are of disassortative type. For two-state proteins, the clustering coefficients of Short-range Interaction Networks show prominent correlation with folding rate and contact order. The assortativity coefficients of Short-range Interaction Networks also show remarkable correlation with folding rate and contact order. Similar correlations exist in Protein Contact Networks of three-state proteins. For two-state proteins, the correlation between contact order and folding rate is determined by the numbers of local contacts. Short-range interactions play a key role in determining the connecting trend among amino acids and they impact the folding rate of two-state proteins directly. For three-state proteins, the folding rate is determined by short-range and long-range interactions among residues together.

Keywords: Protein Contact Networks; Small-World; Scale-Free; Assortative Type; Folding Rate

1. INTRODUCTION

The network concept is increasingly used to describe the topology and dynamics of complex systems. As the essential matter of life, proteins are biological macromolecules made up of a linear chain of amino acids and fold into unique three-dimensional structures (native states). Despite the large degrees of freedom, proteins fold into their native states in a very short time. It is important to understand how proteins consistently fold into their native-state structures and the relationship between structures and function. A protein molecule can be treated as a complex network with each amino acid simplified as a node and the interaction between them as a link. Efforts have been made to model proteins as networks for studying protein topology, small world properties and examining the nucleation in protein folding [1-10]. Bagler and Sinha [11], in their recent protein network analysis, constructed Protein Contact Networks and Long-range Interaction Networks to analyze the assortative mixing of networks and folding kinetics of two-state proteins.

But there is a significant difference in the folding behavior of small proteins with simple two-state kinetics and of larger proteins having a three-state folding kinetics [12]. The two-state proteins have no visible intermediates in the course of folding, which therefore occur as an "all-or-none" process under all experimental conditions. However, the proteins with three-state folding kinetics fold via intermediates, which accumulate during the early stages of folding when it occurs in denaturant-free water [13-16]. Based on the work by Bagler and Sinha, two- and three-state proteins that belong to different structural classes were selected from protein crystal structure data bank to model the native-state protein structures as networks. To investigate various topological properties, the network models were constructed at three different length scales. Protein Contact Networks (PCNs) were built by considering the contacts between atoms in amino acid residues. There is a natural distinction of contacts into two types: long-range and short-

range interactions [7]. We considered the Long-range Interaction Networks (LINs) and Short-range Interaction Networks (SINs) of each protein, which are subsets of the corresponding PCNs. To investigate if the general network parameters can offer any clue to the biophysical properties of the existing three dimensional structure of a protein, these networks were analyzed to focus on their topology including clustering coefficients, shortest path length, average degree, degree distribution and assortative mixing behavior of the amino acid nodes. The determination of folding rate for two- and three-state folding kinetics has a significant difference. To uncover the relationship between the structural properties and the folding kinetics of the proteins, the correlation of protein network parameters with protein folding rate ($\ln k_f$) and topology parameters contact order (CO) was analyzed. The values of $\ln k_f$ and CO are available as given in Reference [12]. Through our coarse-grained complex network model of protein structures, it was found that short-range interactions play a key role in determining the connecting trend among amino acids and impact directly the folding rate of two-state proteins. For three-state proteins, the folding rate is determined by short-range and long- range interactions among residues together.

2. METHODS

2.1. Construction of PCNs, LINs, SINs and Their Random Networks

In this paper, 30 proteins with two-state kinetics and 15 proteins with three-state kinetics were studied and the dataset was taken from the paper [12]. The data of these protein structures were taken from the Protein Data Bank (PDB) to model them as Protein Contact Networks (PCNs) by setting the C_α atoms as the nodes, and established a link between two nodes, if the atoms were within a cut-off distance (0.8nm).

The Long-range Interaction Network (LIN) of a PCN was obtained by considering the interactions which occur between amino acids that were twelve or more amino acids apart in the primary sequence. A LIN was a subset of its PCN with same numbers of nodes (N) but fewer numbers of links due to removal of the short-range contacts. The Short-range Interaction Network (SIN) of a PCN was built with the amino acids separated within twelve. For compare, the random network was constructed with the same numbers of residues (N) and links as those of the PCNs, SINs, LINs.

2.2. Network Parameters

The degree of any node *i* is represented by $k_i = \sum_{j=1}^{N} a_{ij}$. Here a_{ij} is the element of the adjacency matrix, whose value is 1 if an edge connects a node "*I*" to another node "*j*" and 0 otherwise. *N* is the number of nodes. Average degree <*k*> of a network is defined as $< k >= \frac{1}{N}\sum_{i=1}^{N} k_i$.

The shortest path length is related to the link number of a pathway between two nodes and it is the least link number of all the pathways between two nodes. The average shortest path length is defined as $L = \frac{1}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} L_{ij}$, where L_{ij} is the shortest path length between nodes *i* and *j*.

The average clustering coefficient *C* is the average over all vertices of the fraction of the number of connected pairs of neighbours for each vertex. It is calculated as follows: $C = \frac{1}{N}\sum_{i=1}^{N} C_i$, where C_i is the clustering coefficient for a node *i* and defined as the fraction of links that exist among its nearest neighbours to the maximum number of possible links among them. It scales the cohesiveness of the neighbours of a certain node from the view of topology.

Many networks show "assortative mixing" on their degrees. The Assortativity Coefficient (*r*) measures the tendency of degree correlation. It is the Pearson Correlation Coefficient of the degrees at either ends of an edge. Its value was calculated using the function suggested by Newman [17] and was given as

$$r = \frac{M^{-1}\sum_i j_i k_i - \left[M^{-1}\sum_I \frac{1}{2}(j_i + k_i)\right]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}, -1 \le r \le 1$$

The networks having positive *r* values are assortative in nature and the negative value implies that the network is of disassortative type.

3. RESULTS AND DISCUSSION

3.1. Network Parameters of PCNs, LINs, SINs

3.1.1. Average Degree of the Networks

The average degree <*k*> was calculated for each of the three type networks (PCNs, SINs, LINs) of two- and three-state proteins. **Figure 1** shows the average degree <*k*> as a function of network size N. **Table 1** shows the average degree of three type networks for two- and three-state proteins. The values of <*k*> have no obvious difference between two- and three-state proteins. In other words, the average number of contacts per residue for three-state proteins is similar equal to that of two-state proteins. For two-state proteins, the average number of short-range contacts is smaller than that of

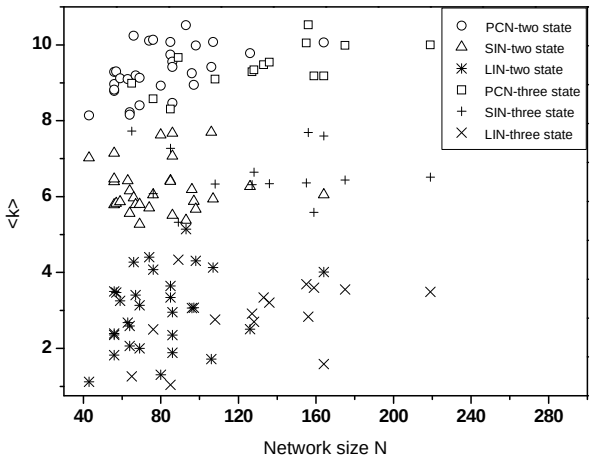


Figure 1. Average degree <*k*> as a function of the network size *N* for 30 two-state and 15 three-state proteins.

Table 1. Average degree <*k*> of PCNs, SINs, LINs.

Network type	$<k>_{two\text{-}state}$	$<k>_{three\text{-}state}$
PCNs	9.31±0.664	9.41±0.582
SINs	6.23±0.668	6.56±0.724
LINs	2.99±0.999	2.85±0.938

three-state proteins and the average number of long-range contacts is slightly higher than that of three-state proteins. In general, for coarse-grained complex network model of protein structures, it has been shown, for different folding kinetics, that the short-range interactions and long-range interactions are consistent with each other for a statistical equilibrium. It is observed that the average degree <*k*> for LINs shows lower values than that of SINs and PCNs regardless of their states. It indicates that long- range interactions exhibit a predominant lower average connectivity compared with short-range interactions. Protein structure has the strongest average connectivity by integrating both short-range and long-range interactions. To verify whether the observed trend depends on the network size (i.e., the number of amino acids of the protein), the correlation coefficient between <*k*> and *N* was calculated. Any significant relationship between <*k*> and *N* in SINs and LINs for two- and three-state proteins was not found. On the other hand, **Figure 1** indicates that the <*k*> of PCNs (both two- and three-state proteins) show a high positive correlation with N. The correlation coefficients of three-state proteins are higher than that of two-state proteins, and their values are 0.672 (p=0.006) and 0.511 (p=0.004), respectively.

3.1.2. "Small-World" Property

To examine whether the networks have the "small-world" property, the average clustering coefficient *C* and the average shortest path length *L* of each of the networks and their respective *Cr* and *Lr* for the random networks with the same size were calculated. According to Watts and Strongatz [18], a network has the "small-world" property if *C*>>*Cr* and *L*≥*Lr*. *Cr* and *Lr* can be calculated using the expressions $C_r \approx < k > / N$ and $L_r \approx \ln N / \ln < k >$. **Table 2** shows the <*C*> and <*L*> of 30 two-state proteins and 15 three-state proteins and the corresponding values of random networks. It is obviously found that PCNs and SINs (both two- and three-state proteins) are characterized by large values of <*C*> and <*L*> compared with the corresponding random networks, which have the typical property of small-world networks. It indicates that any two amino acids are connected with each other via only a few other amino acids in both two- and three-state proteins. Whereas LINs have similar <*C*> with their random networks and their <*L*> are smaller than those of the corresponding random networks. It indicates that LINs do not exhibit the "small-world" property. **Table 2** also shows that two-state proteins have similar values of <*C*> with three-state proteins for three types networks and LINs have remarkable lower <*C*> than those of PCNs and SINs. It suggests that long-range interactions have reduced congregating of amino acids, which may facilitate communication among distant residues in the native structure to some extent, but such a feature can also increase the folding time as it requires distant residues in the chain to come closer during the folding process. **Table 2** also shows that <*L*> of three-state proteins are more higher compared with corresponding two state proteins. It suggests that three-state proteins are packed more loosely than two-state proteins and it has a low global connectivity compared with two-state proteins.

3.1.3. Degree Distribution

The degree distribution is an important feature which characterizes the network topology. **Figure 2** shows the degree distribution of three types' networks for two- and three-state proteins. The shape of the degree distribution of small-world network is bell-shaped, Poisson-like. It has a pronounced peak at <*k*> and decays exponentially for large *k*. Thus the topology of the network is relatively homogeneous, all nodes having approximately the same number of edge. The shape of the degree distribution is Poisson distribution, which is another typical property of "small-world" networks. A network lacking a characteristic scale <*k*> and having degree distribution of a power-law form is known as "scale-free" network [19]. From **Figure 2(a)**, the long-range interaction distribution patterns (both two- and three-state), it is noticed that a large number of nodes with a small number of links and a small

Table 2. Values of <*C*> and <*L*> of three types networks of two- and three-state proteins as well as those for the corresponding random networks.

Network type	<*C*>	<*Cr*>	<*L*>	<*Lr*>
Two-state				
PCNs	0.589	0.122	2.97	1.947
SINs	0.649	0.085	7.269	2.413
LINs	0.039	0.04	3.796	5.677
Three-state				
PCNs	0.57	0.079	3.694	2.139
SINs	0.65	0.056	10.827	2.584
LINs	0.042	0.023	4.828	13.953

number of links with a large number of nodes in the distribution pattern, indicating the "scale-free" behaviour. But from **Figure 2(b)**, the short-range interaction distribution patterns as well as those of total interactions are of the Poisson type. This indicates again that PCNs and SINs exhibit "small-world" behaviour, while LINs indicate "scale-free" behaviour.

The scale-free degree distribution of LINs indicates that proteins contain hubs, i.e. central residues, which have a large number of long-range interactions with other residues. The kinetic mechanism of transitions from the denatured state to the native state is nucleation [20]. The nucleus is composed of a set of adjacent residues, and is stabilized by long-range interactions that are formed as the rest of the protein collapses around it. The Poisson degree distribution means that protein structures have a much smaller number of hubs than most self-organized networks including most cellular or social networks. The major reason for this deviation from the scale-free degree distribution lies in the limited simultaneous binding capacity of a given amino acid side-chain (also called as excluded volume effect). The limited amino acid side chain binding capacity contributes to the fact that each amino acid has a characteristic average degree. This depends on the interaction cut-off, which makes hydrophilic amino acids "strong hubs" (observed at high interaction cut-off allowing low overlaps), and hydrophobic amino acids "weak hubs" (at low interaction cut-off allowing high overlaps), respectively. Hubs are integrating various secondary structure elements, and, therefore, it is not surprising that they increase the thermodynamic stability of proteins.

3.1.4. Assortative Mixing Behavior of the Nodes

The assortative mixing concept has been used in social, technological and biological networks [17]. In social networks assortative mixing leads to homophily, i.e., the tendency of individuals to associate with similar partners. This quantity is also important to control epidemics since assortative has a profound impact on the percolation in networks. Contrary to social networks, which tend to be assortative, biological and technological networks tend to be disassortative. Concerning this aspect, the networks are classified as to show assortative mixing, if the degree correlation is positive, a preference for high-degree nodes to attach to other high-degree nodes, or disassortative mixing, otherwise. Assortativity Coefficient (*r*) for each of the networks was calculated, as shown in **Table 3**. It indicates that all the PCNs and SINs have positive *r* values regardless of two-state or three-state, while the LINs have both positive and negative *r* values. The ratio of negative r_L values for two-state is significantly higher than that for three-state. The former is 17/30, while the latter is 3/15.

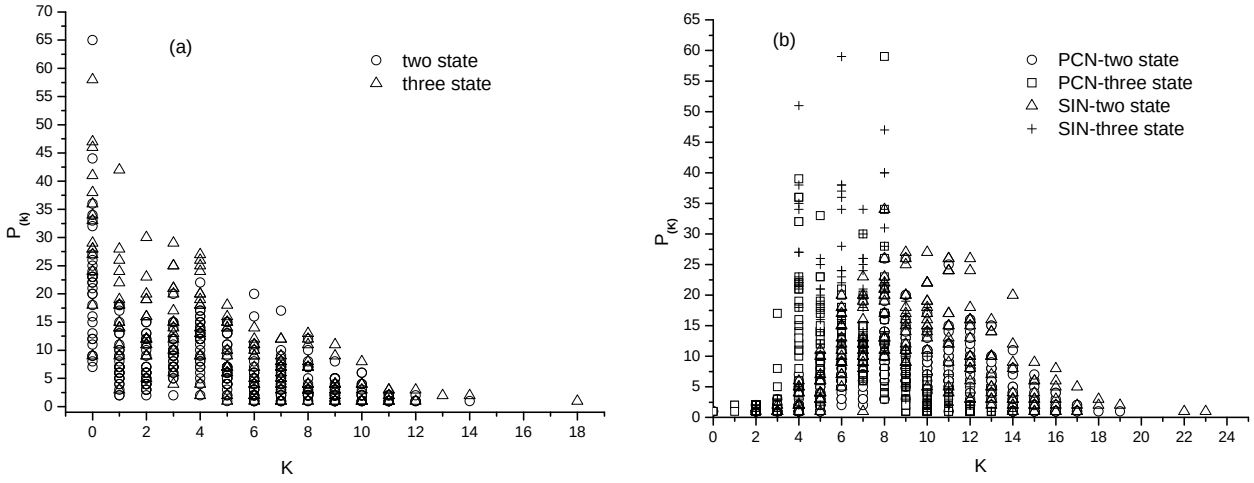


Figure 2. Degree distribution of three type networks for two- and three-state proteins. (a) LINs of two- and three-state proteins; (b) PCNs and SINs of two- and three-state proteins.

Table 3. Assortativity coefficient (r) of three type networks for two- and three-state proteins.

Network type	PCNs	SINs	LINs
r (Two-state)	0.106~0.424	0.106~0.562	-0.566~-0.006 0~0.35
r (Three-state)	0.074~0.562	0.047~0.54	-0.349~-0.117 0.008~0.369

The r values of different networks suggest that all PCNs and SINs are of assortative type, the LINs of three-state proteins (except three) are also of assortative type. While maximum of LINs of two-state have the characteristics of disassortative mixing, few others are of assortative type. Thus it may be said that in all of the PCNs and SINs the residues (nodes) with high degree have tendencies to be attached with the residues having high degree values. The result is consistent with previous study by S. Kundu [21] and Ganesh Bagler [11]. But in some LINs of two-state and three-state proteins having negative r values the mixing pattern of amino acid residues are different. Here the amino acids (nodes) having high degree values have a tendency to be attached with amino acids with smaller degree. This result is not consistent with Ganesh Bagler, who concluded that the assortative mixing in PCNs and LINs is a generic feature of protein structures. Recent research suggests that assortative mixing by degree reduces the stability of networks [22]. In almost all biological networks (e.g. protein interaction network, neural network etc.), nodes of high degree tend to avoid being connected to other highly connected nodes, i.e. these networks show disassortative mixing. This difference of assortative mixing between SINs and LINs may be a possible rea-son for the stability of native-state proteins and the research of assortative mixing in LINs may give interesting surprises in the future. However, the PCN is a composite network of SIN and LIN. When considering the protein structure networks, the r values had been obtained, which represent a cumulative effect of either all positive r values or a mixture of positive and negative r values. Thus it was find that protein structure networks always have positive r values and they are assortative.

3.2. Correlations of Protein Network Parameters with Folding Rate ($\ln k_f$) and Contact Order (CO)

To uncover the relationship between the structural properties and the folding kinetics of the proteins, the correlation of protein network parameters with protein folding rate ($\ln k_f$) and contact order (CO) was studied. The correlation coefficient between general network parameters (e.g., C, L, $<k>$, and r) and the folding rate logarithm ($\ln k_f$) were calculated out. And similar correlation between network parameters and CO was also discussed.

3.2.1. Correlation for Two-State Kinetics

For all the 30 two-state proteins, the clustering coefficients C of PCNs and LINs have not any significant relationship with the $\ln k_f$, and the correlation coefficients are 0.248 (p=0.186), −0.118 (p=0.534), respectively. However, SINs have high positive correlation between C and $\ln k_f$ (correlation coefficient are 0.602, p=0.000). From **Table 2**, the clustering coefficients of LINs are significant lower than those of PCNs and SINs, which show a low correlation with the folding rate of the proteins. It indicates that clustering of amino acids that participate in the long-range interactions, into "cliques" slows down the folding process of two-state proteins. SINs have the highest clustering coefficients among them and C of SINs have significant correlation with the folding rate, indicating that the short-range interactions may be playing a constructive and active role in determining the rate of the two state proteins folding process. The similar correlation occurs between r and $\ln k_f$. The correlation coefficient between r and $\ln k_f$ of SINs is −0.625 (p=0.000). For PCNs and LINs, the correlation coefficients are 0.295 (p=0.181) and 0.121 (p=0.753), respectively. It shows that short-range interactions play a key role in determining the connecting trend among amino acids and influence the folding rate of two-state proteins directly.

Previous studies have found that contact order (CO) has a significant correlation with folding rate of proteins (correlation coefficient of these 30 proteins is −0.72, p=0.000). As an experiential parameter based on 3D structure, though significant correlating with folding rate, the physical meanings of CO is ambiguity. In this study, it is found that C_{SIN} and r_{SIN} have a high correlation with contact order (CO). The correlation coefficients are −0.64 (p=0.000) and 0.817 (p=0.000), respectively. Since the clustering coefficients depend on the degree of the node, we calculated the correlation coefficients between $C_{SIN}*<k>_{SIN}$ and $\ln k_f$. It shows high positive correlation (correlation coefficients are 0.733, p=0.000) between them for these two-state proteins. A significant high correlation also exists between $C_{SIN}*<k>_{SIN}$ and CO, the value is −0.796 (p=0.000) (see **Figure 3**).

C_{SIN} measures the transitivity in the short-range interaction network and $<k>_{SIN}$ measures the average number of short-range contacts per residue. It indicates that the correlation between CO and $\ln k_f$ is determined by the number of local contacts for two-state proteins. It is consistent with the previous study by Mirny and Shakhnovich [23]. It is interesting to note that despite dissimilar quantities that CO and C_{SIN} measure, the similar correlation coefficients essentially indicate the important role of short-range contact formation in the rate of folding for two-state proteins.

3.2.2. Correlation for Three-State Kinetics

For three-state proteins, the clustering coefficients C of

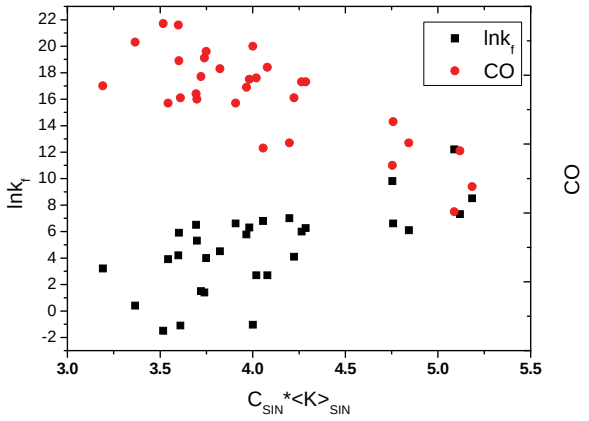


Figure 3. Correlation between C_{SIN}*<k>$_{SIN}$ with lnk_f and CO.

constructed to uncover the different influence of long-range and short-range interactions on two- and three-state folding kinetics. It was found that PCNs and SINs (both two- and three-state proteins) have the typical property of small-world networks, whereas LINs exhibit the "scale-free" property.

the PCNs show a high positive correlation with the folding rate (correlation coefficient is 0.652, p=0.001). However, C of LINs and SINs have not significant relationship with the lnk_f, and the correlation coefficients between C and lnk_f are −0.278 (p=0.315) and 0.405 (p=0.081), respectively. The similar correlation occurs between r and lnk_f. The correlation coefficient between r and lnk_f of PCNs is −0.603 (p=0.017). For LINs and SINs, the correlation coefficients are −0.394 (p=0.146) and −0.474 (p=0.075), respectively. It shows that, for three-state proteins, the folding rate is determined by short-range and long-range interactions among residues together.

4. CONCLUSIONS

The network concept is increasingly used to describe the topology and dynamics of complex systems. In this paper, the three type networks (PCNs, LINs, SINs) were-All of PCNs, SINs and nearly all LINs of three-state proteins are of assortative type. While maximum of LINs of two-state are of disassortative type. This different assortative mixing behaviour of LINs may be a possible reason for the stability of native-state proteins and the research of assortative mixing in LINs may give interesting surprises in the future.

For two-state proteins, C_{SIN} and r_{SIN} show high correlation with lnk_f and CO, which indicates the correlation between CO and lnk_f is determined by the numbers of local contacts. Short-range interactions play a key role in determining the connecting trend among amino acids and influence directly the folding rate of two-state proteins. For three-state proteins, C_{PCN} and r_{PCN} also show high correlation with lnk_f and CO, which shows that the folding rate is determined by short-range and long-range interactions among residues together.

5. ACKNOWLEDGMENTS

This work was supported by a grant from Chinese National Key Fundamental Research Project (No. 30970561) and Shandong Fundamental Research Project (NO. Y2005D12).

REFERENCES

[1] M. Vendruscolo, N. V. Dokholyan, E. Paci and M. Karplus. (2002) Small-world view of the amino acids that play a key role in protein folding, Physical Review E, **65,** 061910.

[2] N. V. Dokholyan, L. Li and F. Ding. (2002) Topological determinants of protein folding, Proc. Natl. Acad. Sci. (USA), **99,** 8637–8641.

[3] A. R. Atilgan, P. Akan and C. Baysal. (2004) Small-world communication of residues and significance for protein dynamics, Biophys. J., **86,** 85–91.

[4] G. Amitai, A. Shemesh and E. Sitbon. (2004) Network analysis of protein structures identifies functional residues, J. Mol. Biol., **344,** 1135–1146.

[5] D. J. Jacobs, A. J. Rader and L. A. Kuhn. (2001) Protein flexibility predictions using graph theory, Proteins: Structure, Function, and Genetics, **44,** 150–165.

[6] X. Jiao, S. Chang, C. H. Li, W. Z. Chen and C. X. Wan. (2007) Construction and application of the weighted amino acid network based on energy, Phys. Rev. E., **75,** 051903.

[7] L. H. Greene and V. A. Higman. (2003) Uncovering network systems within protein structures, J. Mol. Biol., **334,** 781–791.

[8] M. Aftabuddin and S. Kundu. (2006) Weighted and unweighted network of amino acids within protein, Physica A., **396,** 895–904.

[9] S. Kundu. (2005) Amino acids network within protein, Physica A., **346,** 104–109.

[10] K. V. Brinda and S. Vishveshwara. (2005) A network representation of protein structures:implications for protein stability, Biophys J., **89,** 4159–4170.

[11] G. Bagler and S. Sinha. (2007) Assortative mixing in protein contact networks and protein folding kinetics, Bioinformatics, **23(14),** 1760–1767.

[12] O. V. Galzitskaya, S. O. Garbuzynskiy, D. N. Ivankov and A. V. Finketics. (2003) Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics, Proteins: Structure, Function, and Genetics, **51,** 162–166.

[13] S. E. Jackson. (1998) How do small single-domain proteins fold? Fold Des., **3,** 81–91.

[14] A. R. Fersht. (1999) Kinetics of protein folding, In: Hadler GL. Editor: Structure and mechanism in protein science, W.H.Freeman & Co., New York, 40–572.

[15] A. R. Fersht. (2000) Transition-state structure as a unifying basis in protein-folding mechanisms: Contact order, chain topology, stability, and the extended nucleus mechanism, Proc Natl Acad Sci USA, **97,** 1525–1529.

[16] W. A. Eaton, V. Munoz, S. J. Hagen, G. S. Jas, L. J. Lapidus, E. R. Henry and J. Hofrichter. (2000) Fast ki-

netics and mechanisms in protein folding, Annu Rev Biophys Biomol Struct, **29,** 327–359.

[17] M. E. J. Newman. (2002) Assortative mixing in networks, Phys. Rev. Lett., **89,** 208701–208704.

[18] D. J. Watts and S. H. Strogatz. (1998) Collective dynamics of "small-world" networks, Nature, **393,** 440–442.

[19] M. E. J. Newman. (2003) The structure and function of complex networks, SIAM Rev., **45(2),** 167–256.

[20] A. R. Fersht. (1995) Optimization of rates of protein folding: The nucleation-condensation mechanism and its implications, Proc Natl Acad Sci USA, **92(24),** 10869–10873.

[21] M. Aftabuddin and S. Kundu. (2007) Hydrophobic, hydrophilic and charged amino acids' networks within protein, Biophysical Journal, **93(1),** 225–231.

[22] M. Brede and S. Sinha. (2005) Assortative mixing by degree makes a network more unstable, eprint arXiv: cond-mat/0507710.

[23] L. Mirny and E. Shalhnovich. (2001) Protein folding theory: From lattice to all-atom models, Annu Rev Biophys Biomol Struct, **30,** 361–396.

J. Biomedical Science and Engineering, 2009, 2, 651-655 **JBiSE**
doi: 10.4236/jbise.2009.28095 Published Online December 2009 (http://www.SciRP.org/journal/jbise/).

Transforming growth factor-β3 induced rat bone marrow-derived mesenchymal stem cells differentiation into smooth muscle cells by activating Myocardin

Lin-Lin Ma[1,2], Nan Wang[1,2], Zhen Zhou[1,2], Jun-Yun Zhang[1,2], Xue-Gang Luo[1,2], Yong Jiang[1,2], Tong-Cun Zhang[1,2*]

[1]Key Laboratory of Industrial Microbiology, Ministry of Education, Tianjin, China;
[2]College of Biotechnology, Tianjin University of Science and Technology, Tianjin, China.
Email: hippopotamus19851010@126.com; *tony@tust.edu.cn

Received 10 July 2009; revised 21 August 2009; accepted 31 August 2009.

ABSTRACT

Bone marrow mesenchymal stem cells (MSCs) can differentiate into smooth muscle cells (SMCs) and have tremendous potential for cell therapy and tissue engineering. In this study, to understand the effects of TGF-β3 on rat bone marrow-derived MSCs and the underlying molecular mechanism of this differentiation process, we investigated that the changes of myocardin-related transcription factors (MRTFs) at the transcriptional level after rat MSCs were treated with TGF-β3. The results showed that TGF-β3 increased the expression of contractile genes, such as SM22, smooth muscle-myosin heavy chain (SM- MHC), SM-α-actin in MSCs. When TGF-β3 induced MSCs differentiation into SMCs, myocardin and MRTF-A were activated. The data indicated that TGF-β3 induced rat bone marrow-derived MSCs differentiation into SMCs by activating mypcardin and MRTF-A.

Keywords: Mesenchymal Stem Cells; Smooth Muscle Cells; TGF-β; MRTFs

1. INTRODUCTION

In vivo, smooth muscle cells (SMCs) are found in the vascular system, as well as in visceral organs, notably the respiratory, genitourinary, and gastrointestinal systems. VSMCs were involved in atherosclerosis and hypertension, leading causes of heart failure. Thus there was the great interest in the field of cellular therapeutics involving these tissues. However, one major limitation to this approach has been that a reliable source of SMCs can be impractical and morbid. In addition, biopsies usually lead to limited amounts of cells. It has been shown that SMCs derived from diseased organs can lead to abnormal cells that are different from healthy SMCs [1]. Therefore, there is a great need for alternative sour ces of healthy SMCs.

Several groups have suggested the use of bone marrow-derived cells to repair smooth muscle tissues because of their stem cell-like properties [2]. Bone marrow-derived mesenchymal stem cells (MSCs) have a self-renewal capacity, long-term viability. It is important that MSCs can differentiate into a variety of cell types, such as osteogenic, adipogenic, chondrogenic, skeletal muscle cells, and SMCs in response to different microenviromental cues [3]. In vivo, MSCs transplanted into the heart can differentiate into SMCs and contribute to the remodeling of vasculature [4]. These findings indicated that MSCs might be as sources of healthy SMCs under some specific conditions, such as in the presence of some cytokine.

Transforming growth factor-β (TGF-β) proteins are multifunctional proteins that regulate cell growth, differentiation, migration, and extracellular matrix production. It has been shown recently that TGF-β increases smooth muscle (SM)-actin expression in MSCs [5]. Sphingosylphosphorylcholine (SPC) induces differentiation of human adipose-tissue-derived MSCs into smooth-muscle-like cells through a TGF-β-dependent mechanism [6]. These results indicate that the TGF-β was involved in mesenchymal lineage cell type differentiation into SMCs. Therefore, we investigated the role of TGF-β in bone marrow-derived MSC differentiation into SMCs.

Myocardin is expressed specifically in smooth and cardiac muscle cell lineages and activates smooth and cardiac muscle reporter genes by interacting with serum response factor (SRF). Myocardin shares homology with myocardin-related transcription factor-A (MRTF-A), and MRTF-B, which are expressed in a broad range of embryonic and adult tissues [7]. To understand whether MRTFs are implicated in the SMC differentiation of

This work was financially supported by National Natural Science Foundation of China (No.30800561), Tianjin Natural Science Foundation (09JCZDJC18100) and Scientific Research Foundation of Tianjin University of Science and Technology (20080409).

MSCs, we investigated the changes of MRTF family mRNA level after MSCs were induced by TGF-β3.

In this report, we show that TGF-β3 induced rat bone marrow-derived MSC differentiation into SMCs. TGF-β3 increased the expression of contractile genes, such as SM22, smooth muscle-myosin heavy chain (SM-MHC), SM-α-actin in MSCs. We also demonstrated that myocardin and MRTF-A play important roles in the TGF-β3-induced SMC differentiation in MSCs.

2. MATERIALS AND METHODS

2.1. Reagents

Dulbecco's modified Eagle's medium-low glucose (DMEM-LG) was purchased from Hyclone Co. TGF-β3 was purchased from Peoro Tech. Co. Fetal bovine serum (FBS) was obtained from Hycolon Lab, Inc. Fluorescein isothiocyanate (FITC)-conjugated or phycoerythrin (PE)-conjugated antibodies CD14, CD29, CD34, CD44, CD45, CD105 were purchased from BD Pharmingen, San Diego, CA. Mouse anti-rat SM-MHC and FITC-goat anti-mouse IgG were purchased from Santa Cruz. 3-(4, 5-dimethylthiazol-2-yl)-2, 5-diphenyltetrazolium (MTT) was purchased from Sigma. Co. Trizol reagent was purchased from Invitrogen, USA. All other reagents were ultrapure grade.

2.2. Cell Culture

Rat MSCs were isolated from the femurs and tibias of male Sprague-Dawley rats (90–100g) with a modified method originally described by Pittenger [3]. Briefly, bone marrow mononuclear cells were obtained by Percoll (1.073 g/ml) density gradient centrifugation. The cells were seeded in Dulbecco's modified Eagle's medium-low glucose (DMEM-LG) supplemented with 10% fetal bovine serum (FBS) and penicillin/streptomycin at 37℃ in humified air with 5% CO2. At 24 h after plating, nonadherent cells were removed by replacing medium. The antibiotic was removed after one media change. The medium was changed every 2–3 days and the cells were passaged in 0.05% trypsin-1 mM EDTA. All the cells were used between passages 4 and 6.

2.3. Flow Cytometric Analysis

Rat MSCs were phenotypically characterized by flow cytometry (Becton-Dickinson, San Jose, CA) by the method of Li [8]. The antibodies used in this study included FITC-conjugated or PE-conjugated antibodies CD14, CD29, CD34, CD44, CD45, CD105. To detect surface antigens, cells were collected and incubated (30 min at 48C) with the respective antibody at a concentration previously established by titration. At least 1×105 cells for each sample were acquired and analyzed.

2.4. Cell Differentiation Induction

When the cultures of MSCs reached subconfluence, cells were washed twice with the medium and divided into two groups. In control group, the cells were cultured in basal DMEM medium (without FBS); in TGF-β treatment group, the cells were incubated with 5, 10 ng/ml TGF-β in basal medium. Fresh TGF-β was dissolved in phosphate buffered saline (PBS) and applied to cells. The cells in the two groups were incubated for 24 hours. The morphological changes of the cells were observed under phase contrast microscope (Nikon, Japan).

2.5. Immunocytochemistry Assay

After MSCs were treated in the two ways mentioned above for 24 h, cells were fixed in 4% paraformaldehyde for 15 min, blocked with normal goat serum for 20 min at room temperature (RT). Then, primary antibodies (mouse anti-rat SM-MHC) were added and incubated in a humid chamber over night. After washing with 0.1 M phosphate buffered saline (PBS) three times, cells were incubated with appropriate secondary antibodies (FITC-goat anti-mouse IgG) for 30 min at 37℃. After washing with 0.1 M PBS, the samples were evaluated under inverted fluorescence microscope (Nikon, Japan).

2.6. Cell Viability Assay

Cells were seeded into 96-well plates and treated with or without TGF-β for 24 h, respectively. The viability of cells determined by using the method of MTT assay as described previously [9]. The light absorption was measured at 570 nm using Multiskan Spectrum (Thermo Labsystems). The viability (%) was calculated by the formula as follow. Viability (%)=(OD of control or treated group/OD of normal group)×100%. The viability of normal group was presumed as 100%.

2.7. RNA Isolation and Semi-Quantitative Reverse-Transcription Polymerase Chain Reaction (RT-PCR)

MSCs were treated in the two ways mentioned above for 24 h. Semi-quantitative RT-PCR analysis was carried out as described previously [6]. Briefly, total RNA was isolated from cells using Trizol reagent, two microgram of the sample was reverse-transcribed using M-MLV reverse transcriptase (Promega, USA) according to the manufacturer's instructions. The PCR primer sequences are listed in **Table 1**. Semiquantitative analysis of mRNA expression was performed by using the Biorad software, using human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) as a housekeeping gene.

2.8. Statistical Analysis

Data were expressed as mean±SE and accompanied by the number of experiments performed independently, and analyzed by t-test. Differences at $P<0.05$ were considered statistically significant.

3. RESULTS

3.1. Immunophenotypic Characterization of Rat MSCs

Rat MSCs isolated in this study were uniformly positive for CD29, CD44, CD105. In contrast, these cells were negative for other markers of the hematopoietic lineage CD14, CD34, the leukocyte common antigen CD45. Flow cytometry analyses showed that the MSC was a homogeneous cell population devoid of hematopoietic cells (**Figure 1**).

3.2. TGF-β Induced the SMC Differentiation of Rat MSCs

Rat MSCs were exposed to 5, 10 ng/ml TGF-β3 respectively in the absence of serum. The treatment of MSCs with TGF-β3 significantly changed the cell morphology. As show in **Figure 2A**, 24 hours after TGF-β3 treatment, MSCs have a more spread out and myoblast-like morphology, and intracellular fibrous structures were visible. There were no obvious morphological changes in the control group. To confirm the SMC differentiation of these MSCs, we analyzed the expression of SM-MHC by immunocytochemistry. The cells treated with 5 ng/ml TGF-β3 exhibited the positive SM-MHC (**Figure 2B**). The cells were treatment under the serum-free condition, thus we analyzed the cell viabilities. As shown in **Figure 3**, the viabilities in the control and D609 treatment groups were decreased obviously at 24 h, and are only about 60%. There were no significant between control group and each experiment group (P>0.05). These results shown that 10 ng/ml TGF-β3 did not regulate the cell growth, but could induce the differentiation of MSCs.

To confirm the characters of these differentiated MSCs, the expression of SM22, SM-α-actin, SM-MHC mRNA were examined. RT-PCR experiment results showed that at 24 h, the MSCs treated with 5 ng/ml TGF-β3 displayed weak expression of SM22, SM-α-actin, SM-MHC and 10 ng/ml TGF-β3 induced the intensive increase of SM22, SM-α-actin, SM-MHC in rat MSCs (**Figure 4**). In the control group, no expression of SM22, SM-α-actin, SM-MHC were detected (**Figure 4**). These results showed that TGF-β3 induced rat bone marrow-derived MSC differentiation into SMCs and increased the expression of contractile genes, such as SM22, SM-MHC, SM-α-actin.

Table 1. The PCR primer sequence.

SM22	sense	AGCCAGTGAAGGTGCCTGAGAAC
	antisense	TGCCCAAAGCCATTACAGTCCTC
SM-α-actin	sense	GAGAAGCCCAGCCAGTCG
	antisense	CTCTTGCTCTGCGCTTCG
SM-MHC	sense	TGAGTGACAGAGTCCGCAAG
	antisense	GCCGCAACAGTGGACTTAAG
myocardin	sense	TCACCGCCTTAGCTCATACC
	antisense	CTGTCCCTCTGACCATTCTG
MRTF-A	sense	CTGACCCGAATGCTCCAACA
	antisense	CCAGGGCCATCTGCACTCTT
MRTF-B	sense	GTAGCCAGACCCTTGTTGCC
	antisense	TGTTTGGTGCCGAGTTTGTG
GAPDH	sense	ATTCAACGGCACAGTCAAGG
	antisense	GCAGAAGGGGCGGAGATGA

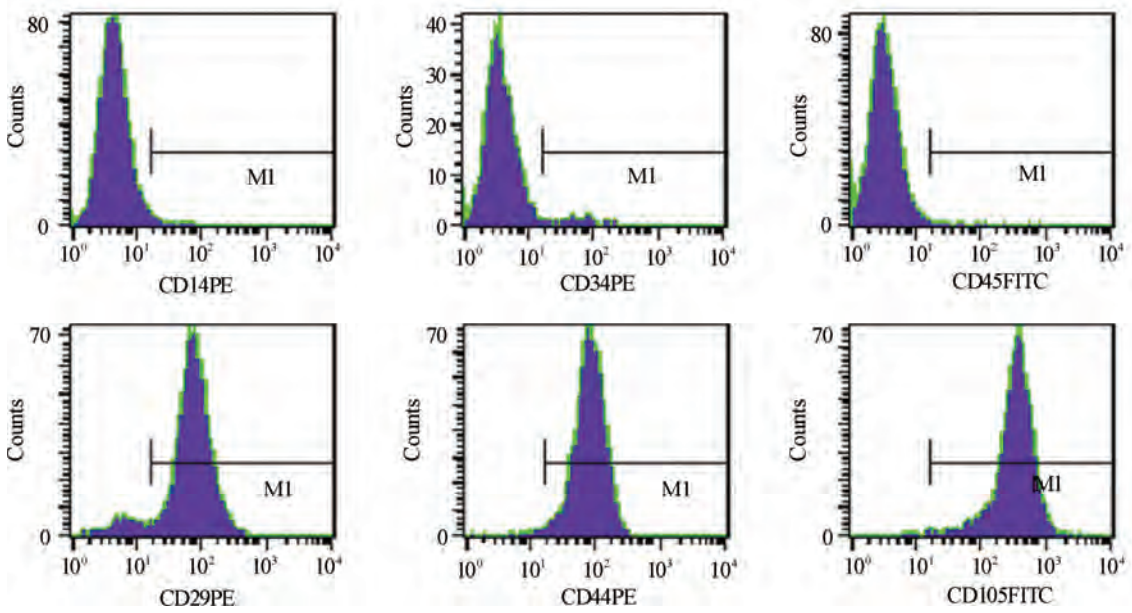


Figure 1. Immunophenotypic Characterixation of rat MSCs. the X-axis represents the cell number.

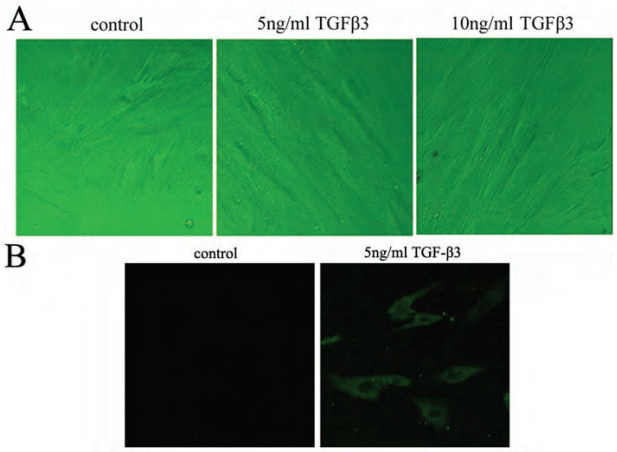


Figure 2. TGF-β3 induce SMC differentiation of MSCs. (a) The morphological changes of rat MSCs; (b) The expression of SM-MHC in undifferentiated and differentiated cells. The cells were cultured in the serum-free medium with/without TGF-β3 at 24 h.

3.3. Myocardin and MRTF-A were activated during the SMC differentiation of rat MSCs induced by TGF-β

To investigate the role of MRTF family in the SMC differentiation of rat MSC, we detected the expression of myocardin, MRTF-A, MRTF-B by RT-PCR. It was observed that myocardin, MRTF-A were activated (**Figure 5**) during the SMC differentiation of rat MSCs induced by TGF-β3, while no expression of MRTF-B was detected (data not shown). This result showed that myocardin, MRTF-A might contribute to the SMC differentiation of rat MSCs induced by TGF-β3, but MRTF-B might not be implicated in this differentiation process.

4. DISCUSSION

SMCs are critical in development and postnatal life. SMCs have been implicated in vascular development as well as in a variety of cardiovascular diseases, including hypertension and atherosclerosis. Hence, studies aimed at the SMCs differentiation of rat MSCs are of great importance and will provide evidence for tissue engineering and therapeutic applications. Cellular transplantation therapy

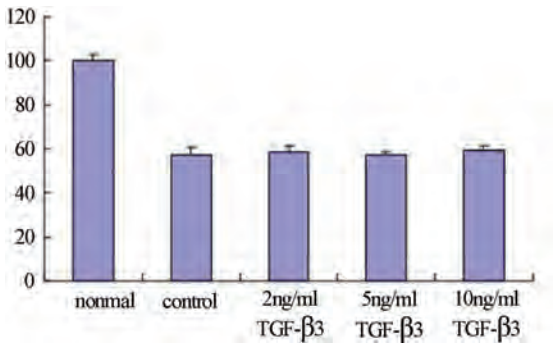


Figure 3. The viability of rat MSCs treated with TGF-β3. The cells were cultured in the serum-free medium with/ without TGF-β3 at 24 h.

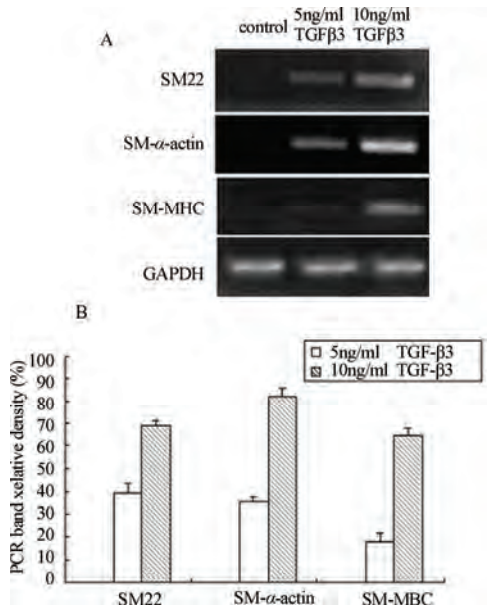


Figure 4. Expression of contractile genes in rat MSCs treated with TGF-β3. (a) During the SMC differentiation of MSCs, the mRNA levels of SM22, SM-α-actin, SM-MHC were elevated obviously. (b) Semiquantitative analysis of the contractile gene mRNA expression.

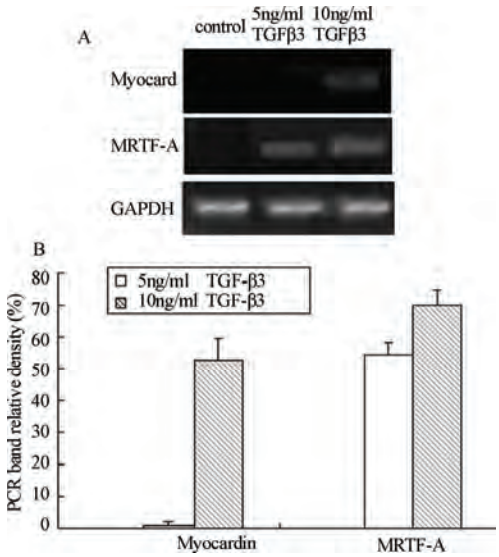


Figure 5. Expression of myocardin and MRTF-A in rat MSCs treated with TGF-β3. (a) During the SMC differentiation of MSCs, myocardin and MRTF-A were activated. (b) Semiquantitative analysis of myocardin and MRTF-A mRNA expression.

for the patients with cardiovascular diseases might possibly be achieved using the regenerated SMCs from autologous bone marrow cells in the near future. MSCs have great appeal for tissue engineering and therapeutic applications because of their general multipotentiality and relative ease of isolation. In vitro, MSCs have been shown to differentiate to SMCs in response to mechanical stress [10], growth factor, such as PDGF-BB and TGF-β [11], and direct contact with vascular endothelial cells [12]. TGF-β can contribute to development of SMCs from embryonic stem cells [13]. In this study, we found that TGF-β3 could induce SMC differentiation of rat bone marrow-derived MSCs and increased the expression of contractile genes, such as SM22, SM- MHC, SM-α-actin in MSCs.

However, the induction action of TGF-β3 and the underlying molecular mechanisms involved in the differentiation of MSCs into SMCs are not well known. Myocardin is known to be a potent serum response factor (SRF) cofactor that plays important roles in regulating smooth muscle and cardiac muscle gene transcription [7]. It is reported that the expression of myocardin at the transcriptional level were increased during the SMC differentiation of hATSCs induced by SPC [6]. Taken together with previous experimental results, our findings are consistent with the idea that myocardin and MRTF-A is activated during the process of SMC differentiation in MSCs.

In summary, our results in this study showed that TGF-β3 induced rat bone marrow-derived MSC differentiation into SMCs. TGF-β3 increased the expression of contractile genes, such as SM22, smooth muscle-myosin heavy chain (SM-MHC), SM-α-actin in MSCs. When TGF-β3 induced MSCs differentiation into SMCs, myocardin and MRTF-A were activated. The data indicated that TGF-β3 induced rat bone marrow-derived MSCs differentiation into SMCs by activating mypcardin and MRTF-A. Obviously, further studies are needed to clarify, for example, the interaction of myocardin and smad proteins as major intracellular mediators of TGF-β signal pathways.

REFERENCES

[1] J. Y. Lai, C. Y. Yoon, J. J. Yoo, T. Wulf, and A. Atala. (2002) Phenotypic and functional characterization of in bivo tissue engineered smooth muscle from normal and pathological bladders. J. Urol., Sugar Land, **168**, 1853–1857.

[2] D. Orlic, J. Kajstura, S. Chimenti, I. Jakoniuk, S. M. Anderson, B. Li, *et al*. (2001) Bone marrow cells regenerate infarcted myocardium. Nature, **410,** 701–705.

[3] M. F. Pittenger, A. M. Mackay, S. C. Beck, R. K. Jaiswal, R. Douglas, J. Mosca, *et al*., (1999) Multilineage potential of adult humanmesenchymal stem cells. Science, **284**, 143–147.

[4] S. Davani, A. Marandin, N. Mersin, B. Royer, B. Kantelip, P. Herve, *et al*. (2003) Mesenchymal progenitor cells differentiate into an endothelial phenotype, enhance vascular density, and improve heart function in a rat cellular cardiomyoplasty model. Circulation, **108**, II253–II258.

[5] B. Kinner, J. M. Zaleskas, M. Spector. (2002) Regulation of smooth muscle actin expression and contraction in adult human mesenchymal stem cells. Exp. Cell Res, **278**, 72–83.

[6] E. S. Jeon, H. J. Moon, M. J. Lee, H. Y. Song, Y. M. Kim, Y. C. Bae, *et al*., (2006) Sphingosylphosphorylcholine induces differentiation of human mesenchymal stem cells into smooth-muscle-like cells through A TGF-beta-dependent mechanism. J Cell Sci., **119**, 4994–5005.

[7] G. C. Pipes, E. E. Creemers, E. N. Olson. (2006) The myocardin family of transcriptional coactivators: versatile regulators of cell growth, migration, and myogenesis. Genes Dev., **20**, 1545–1556.

[8] C. D. Li, W. Y. Zhang, H. L. Li, X. X. Jiang, Y. Zhang, P. H. Tang, *et al*. (2005) Mesenchymal stem cells derived from human placenta suppress allogeneic umbilical cord blood lymphocyte proliferation. Cell Res., **15**, 539–547.

[9] P. Price and T. J. McMillan. (1990) Use of the tetrazolium assay in measuring the response of human tumor cells to ionizing radiation. Cancer Res., **50**, 1392–1396.

[10] N. Kobayashi, T. Yasu, H. Ueba, M. Sata, S. Hashimoto, M. Kuroki, *et al*. (2004) Mechanical stress promotes the epression of smooth muscle-like properties in marrow stromal cells. Exp. Hematol., **32**, 1238–1245.

[11] J. J. Ross, Z. Hong, B. Willenbring, L. Zeng, B. Isenberg, E. H. Lee, *et al*. (2006) Cytokine-induced differentiation of multipotent adult progenitor cells into functional smooth muscle cells, J Clin Invest., **116**, 3139–3149.

[12] S. G. Ball, A. C. Shuttleworth, C. M. Kielty. (2004) Direct cell contact influences bone marrow mesenchymal stem cell fate. Int. J. Biochem, Cell Biol., **36**, 714–727.

[13] S. Sinha, M. H. Hoofnagle, P. A. Kingston, M. E. McCanna, G. K. Owens. (2004) Transforming growth factor-β1 signaling contributes to development of smooth muscle cells from embryonic stem cells. Am. J. Physiol. Cell. Physiol., **287**, C1560–C1568.

J. Biomedical Science and Engineering, 2009, 2, 656-660 **JBiSE**
doi: 10.4236/jbise.2009.28096 Published Online December 2009 (http://www.SciRP.org/journal/jbise/).

Fingerprint image segmentation using modified fuzzy c-means algorithm

Jia-Yin Kang[1], Cheng-Long Gong[1], Wen-Juan Zhang[2]

[1]School of Electronics Engineering, Huaihai Institute of Technology, Lianyungang, China;
[2]School of Computer Engineering, Huaihai Institute of Technology, Lianyungang, China.
Email: jiayinkang@gmail.com

Received 16 July 2009; revised 31 August 2009; accepted 1 September 2009.

ABSTRACT

Fingerprint segmentation is a crucial step in fingerprint recognition system, and determines the results of fingerprint analysis and recognition. This paper proposes an efficient approach for fingerprint segmentation based on modified fuzzy c-means (FCM). The proposed method is realized by modifying the objective function in the Szilagyi's algorithm via introducing histogram-based weight. Experimental results show that the proposed approach has an efficient performance while segmenting both original fingerprint image and fingerprint images corrupted by different type of noises.

Keywords: Fingerprint; Segmentation; Fuzzy C-means; Histogram; Robustness

1. INTRODUCTION

Fingerprint segmentation is an important issue in fingerprint recognition system. A fingerprint image usually has to be segmented to remove uninterested regions before some other steps such as enhancement and minutiae detection so that the image processing will consume less CPU time. A fingerprint image generally consists of different regions: non-ridge regions, high quality ridge regions, and low quality ridge regions. Fingerprint segmentation is usually to identify non-ridge regions and unrecoverable low quality ridge regions and exclude them as background [1]. Most segmentation methods are block-wised ones which divide the fingerprint image into un-overlapped blocks and decide on the type (background and foreground) of each block. Some other methods are pixel-wised ones which determine the type of each pixel. Fingerprint segmentation typically computes the feature (or feature vector) of each element, block or pixel, and then determines the element's type based on the feature (vector). The features used in fingerprint segmentation mainly include statistical features of pixel intensity, directional image and ridge projection signal *et al*.

Fuzzy c-means (FCM) clustering algorithm, an unsupervised clustering technique, has been widely used in image segmentation since it was proposed [2,3]. Compared with hard c-means algorithm [4], FCM is able to preserve more information from the original image. However, for one thing, it is noise-sensitive for not taking into account the spatial information [5]; for another, it is supposed that each feature date has the same contribution to classifying results [6]. To solve the first problem, recently, many researchers proposed the algorithms accounting for spatial information via modifying the objective function of standard FCM algorithm [5,7,8]. To solve the second problem, Li *et al.* [6] proposed a modified clustering algorithm via introducing feature weight of the data.

Generally, fingerprint image is gray level image and is inevitably corrupted by noise during acquisition. Consequently, data feature of the fingerprint is the pixel's gray value. From the gray level histogram of the fingerprint image, it is easily known that the occurrence frequencies of the different gray levels are usually different. Therefore, different gray level pixel has different contribution to clustering results.

In this paper, we propose a modified algorithm for noisy fingerprint image segmentation. The proposed approach is based on modified fuzzy c-means which is robust to noise. Our method achieves more desirable performance compared to standard FCM and Szilagyi modified FCM in [8].

The paper is organized as follows. In Subsection 2.1, standard FCM clustering algorithm is described. Subsection 2.2 presents the proposed modified FCM algorithm to segment the fingerprint. In Section 3, the experimental results are presented. Finally, Section 4 gives our conclusions and discussions.

2. METHODOLOGY

2.1. Standard FCM Algorithm

The FCM algorithm assigns pixels to each category by using fuzzy memberships.

Let $X=\{x_i, i=1,2,\cdots,N \mid x_i \in \mathbb{R}^d\}$ denotes an image with N pixels to be partitioned into c classes, where x_i represents features data. The algorithm is an iterative optimization that minimizes the objective function defined as follows [3]:

$$J_m = \sum_{k=1}^{c}\sum_{i=1}^{N} u_{ki}^m \left\| x_i - v_k \right\|^2 \quad (1)$$

with the following constraints:

$$\{u_{ki} \in [0,1] \mid \sum_{k=1}^{c} u_{ki} = 1, \forall i, 0 < \sum_{i=1}^{N} u_{ki} < N, \forall k\} \quad (2)$$

where u_{ki} represents the membership of pixel x_i in the k^{th} cluster, v_k is the k^{th} class center; $\|\bullet\|$ denotes the Euclidean distance, $m>1$ is a weighting exponent on each fuzzy membership. The parameter m controls the fuzziness of the resulting partition. The membership functions and cluster centers are updated by the following expressions:

$$u_{ki} = \frac{1}{\sum_{l=1}^{c} \left(\frac{\left\| x_i - v_k \right\|}{\left\| x_i - v_l \right\|}\right)^{2/(m-1)}} \quad (3)$$

and

$$v_k = \frac{\sum_{i=1}^{N} u_{ki}^m x_i}{\sum_{i=1}^{N} u_{ki}^m} \quad (4)$$

In implementation, matrix V is randomly initialized, and then U and V are updated through an iterative process using **Eq.s 3 and 4** respectively.

2.2. Modified FCM Algorithm

Szilagyi *et al.*proposed a fast FCM clustering algorithm, EnFCM [8], which is used for gray level image segmentation. The algorithm accounts for pixel spatial information. Before the algorithm implementation, a linearly-weighted sum image ξ, composed by original image and local neighboring average of each pixel in original image , was calculated as follows:

$$\xi_i = \frac{1}{1+\alpha}\left(x_i + \frac{\alpha}{N_R}\sum_{j \in N_i} x_j\right) \quad (5)$$

where ξ_i is the gray value of the i^{th} pixel in the image ξ. N_i stands for the set of neighbors falling into a local window around x_i, and N_R is its cardinality. The parameter α in the second term controls the effect of the penalty. In essence, the addition of the second term in **Eq.5**, equivalently, formulates a spatial constraint and aims at keeping continuity on neighboring pixel values around x_i. Accordingly, the modified objective function was described as follows:

$$J_s = \sum_{k=1}^{c}\sum_{l=1}^{q} \gamma_l u_{kl}^m \left\| \xi_l - v_k \right\|^2 \quad (6)$$

where $\xi = \{\xi_l, l=1,2,\ldots,q\}$ is the data set rearranging from he image ξ defined in **Eq.5** according to gray level. $V=\{v_k\}(k=1,2,\cdots,c)$ represents the prototype of the k^{th} cluster, $U=\{u_{kl}\}(k=1,2,\cdots,c, l=1,2,\cdots,q)$ represents the fuzzy membership of gray value l with respect to cluster k. q denotes the number of the gray level of the given image which is generally much smaller than N. γ_l is the number of the pixels having the gray value equal to *l*, where $l=1,2,\cdots,q$. Naturally, $\sum_{l=1}^{q}\gamma_l = N$.

Similar to the standard FCM algorithm, under the constraints that $\sum_{k=1}^{c} u_{kl} = 1$ for any *l*, minimize J_s defined in **Eq.6**. Specifically, taking the first derivatives of J_s with respect to u_{kl} and v_k, and zeroing them, respectively, two necessary but not sufficient conditions for J_s will be obtained as follows:

$$u_{kl} = \frac{(\xi_l - v_k)^{-2/(m-1)}}{\sum_{r=1}^{c}(\xi_l - v_r)^{-2/(m-1)}} \quad (7)$$

$$v_k = \frac{\sum_{l=1}^{q}\gamma_l u_{kl}^m \xi_l}{\sum_{l=1}^{q}\gamma_l u_{kl}^m} \quad (8)$$

Obviously, in **Eq.6**, gray level was viewed as the classified data. Hence, the number of classified data only depends on gray level, and doesn't enlarge with the increasing of image size. However, **Eq.6** doesn't take different gray level which has different influence on classifying results into consideration, i.e., **Eq.6** considers that every gray level has the same contribution to the classifying results. Actually, according to the gray level histogram of the fingerprint image, it is clear that the occurrence frequencies of different gray level are different. Therefore, different gray level has different contribution to clustering results. Based on above analysis, we modified the objective function in **Eq.6** as follows:

$$J_s = \sum_{k=1}^{c}\sum_{l=1}^{q} w_l \gamma_l u_{kl}^m \left\| \xi_l - v_k \right\|^2 \quad (9)$$

where w_l is the weighting coefficient of $\zeta_l(l=1,2,\cdots,q)$, and can be computed via histogram as follows:

$$w_l = \frac{\gamma_l}{N}, \quad l=0,1,\cdots,q \quad (10)$$

where q denotes the number of the gray level of the given image. γ_l is the number of the pixels having the

gray value equal to l, where $l = 1, 2, \cdots, q$. Naturally, $\sum_{l=1}^{q} \gamma_l = N$, $\sum_{l=1}^{q} w_l = 1$, i.e., $w_l (l = 1, 2, \cdots, q)$ can be viewed as the occurrence probability of each gray level. Hence, from **Eq.10**, it is known that the weighting coefficient of each gray level can be given by the normalized image histogram.

Similarly, under the constraints that $\sum_{k=1}^{c} u_{kl} = 1$ for any l, minimize J_s defined in **Eq.9**. Specifically, taking the first derivatives of J_s with respect to u_{kl} and v_k, and zeroing them, respectively, two necessary but not sufficient conditions for J_s will be obtained as follows:

$$u_{kl} = \frac{(\xi_l - v_k)^{-2/(m-1)}}{\sum_{r=1}^{c} (\xi_l - v_r)^{-2/(m-1)}} \tag{11}$$

$$v_k = \frac{\sum_{l=1}^{q} w_l \gamma_l u_{kl}^m \xi_l}{\sum_{l=1}^{q} w_l \gamma_l u_{kl}^m} \tag{12}$$

From **Eq.12**, it is known that the function of weighting coefficient w_l lies in adjusting the clustering center. **Eq.9** will degenerated to **Eq.6** while $w_l = 1/q$.

The modified FCM algorithm (spatially weighting FCM clustering algorithm, called SWFCM) can be summarized as follows:

Step 1: Fix $m > 1$ and $2 \le c \le N - 1$; then select initial class prototypes $v_k (k = 1, 2, \cdots, c)$; set $\varepsilon > 0$ to a very small value.

Step 2: Compute the new image ζ in terms of **Eq.5** in advance.

Repeat:

Step 3: Compute/modify μ_{kl} with v_k by **Eq.s 11 and 12**.

Step 4: Update v_k with the modified μ_{kl} by **Eq. 12**.

Until ($|V_{new} - V_{old}| < \varepsilon$)

3. RESULTS AND DISCUSSIONS

In the following experiments, we first execute the three segmentation algorithms, FCM, EnFCM and SWFCM on an original fingerprint image in Subsection 3.1. Then perform the three segmentation algorithms on noisy fingerprint images corrupted, respectively, by Gaussian noise and salt and pepper noise to investigate the robustness of the algorithms in Subsection 3.2. Finally, the Subsection 3.3 gives the corresponding quantitative comparisons for the segmenting results presented in Subsection 3.1 and 3.2. In all the following experiments, we set the parameters $c = 2$, $m = 2$, $a = 5$, $\varepsilon = 10^{-5}$.

3.1. Results on Original Fingerprint Image

To compare the segmenting performances of the above three algorithms, FCM, EnFCM and SWFCM, we apply these algorithms to an original test fingerprint image as shown in **Figure 1(a)** with the size of 300×300 pixels, and the corresponding segmenting results are, respectively, displayed in **Figures 1(b,c,d)**.

As shown in **Figure 1**, it is clear that our proposed algorithm performs more visually significant than other two methods do. More detailed quantified comparison according to execution time and iteration step are given in the next subsection.

3.2. Results on Fingerprint Images Corrupted by Noises

To examine the above three algorithms' robustness to noise, we apply the three algorithms on fingerprint images corrupted by noises. **Figure 2(a)** is the original image with no noise and **Figures 2(b) and 2(c)** are the same images corrupted, respectively, by the Gaussian noise ($\mu = 0, \sigma = 0.05$) and the salt and pepper noise with noisy density $d = 0.02$. The segmenting results on **Figures 2(a) and 2(b)** are shown in **Figures 3 and 4**, respectively.

Figure 1. Fingerprint image segmentation. (a) Original fingerprint image; (b) Fingerprint segmentation result using FCM; (c) Fingerprint segmentation result using EnFCM; (d) Fingerprint segmentation result using SWFCM.

Figure 2. Fingerprint image. (a) Original image; (b) The image corrupted by Gaussian noise; (c) The image corrupted by salt and pepper noise.

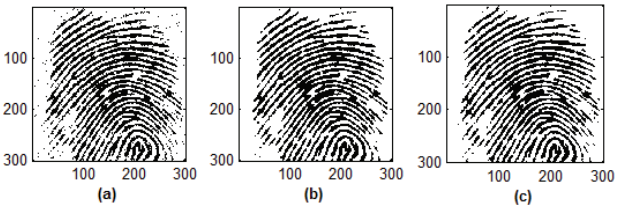


Figure 3. Segmenting results on fingerprint image corrupted by Gaussian noise. (a) Fingerprint segmentation result using FCM (b) Fingerprint segmentation result using EnFCM (c) Fingerprint segmentation result using SWFCM.

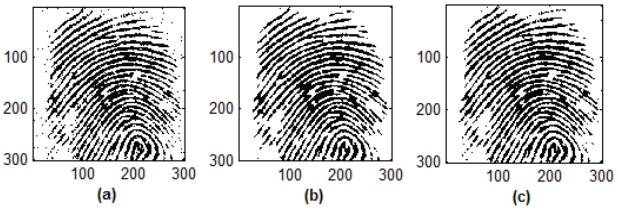


Figure 4. Segmenting results on fingerprint image corrupted by salt and pepper noise. (a) Fingerprint segmentation result using FCM; (b) Fingerprint segmentation result using EnFCM; (c) Fingerprint segmentation result using SWFCM.

Both from **Figures 3 and 4**, We can visually see that FCM is influenced by the Gaussian noise and the salt and pepper noise to different extents, respectively, which indicate that FCM algorithm lacks enough robustness to both the Gaussian noise and the salt and pepper noise, while EnFCM and SWFCM can basically eliminate the effect of the noises. Although the segmenting results using EnFCM and SWFCM are visually almost same, more detailed quantified comparison according to execution time, iteration step and signal-to-noise ratio (SNR) are needed to further investigate in the next subsection.

3.3. Quantitative Segmenting Results Comparisons

We tabulate quantitative segmenting comparisons in **Tables 1,2,3** of the above three algorithms for **Figures 1,3,4**, respectively.

From **Tables 1,2,3**, we can obviously see that the iteration step of SWFCM is less than that of FCM and EnFCM. Furthermore, the execution time (CPU: 2 GHz, Memory: 1GHz, Operating system: windows XP, Soft ware: Matlab 7.0) of SWFCM is reduced compared with FCM and EnFCM due to both the less iterations and only dependence on the number of the gray-level q (256) rather than the image size itself N (300×300).

Table 1. Comparison scores of three methods corresponding to **Figure 1**.

Algorithm	T	N
FCM	5.123	35
EnFCM	3.841	28
SWFCM	3.077	23

Table 2. Comparison scores of three methods corresponding to **Figure 3**.

Algorithm	T	N	SNR
FCM	3.198	17	28.075
EnFCM	2.573	13	28.591
SWFCM	1.967	11	31.308

Table 3. Comparison scores of three methods corresponding to **Figure 4**.

Algorithm	T	N	SNR
FCM	3.130	14	22.157
EnFCM	2.217	8	24.972
SWFCM	1.048	4	27.872

Note: In the above three tables, T stands for execution time with the unit of second; N stands for iteration step.

From **Tables 2 and 3**, we can also see that the SNR of SWFCM is less than that of FCM and EnFCM. From this results, it can be concluded that the newly-proposed algorithm give rise to better denoising performance than both FCM and EnFCM algorithms.

4. CONCLUSIONS

In this paper, an automatic modified FCM clustering algorithm for fingerprint segmentation was proposed. The proposed algorithm is realized by modifying the objective function in the Szilagyi's algorithm via introducing the gray histogram-based weighting. Experimental results show that proposed method can dramatically speed up FCM, and can achieve better denoising performance compared with EnFCM. Therefore, the proposed approach will be promising in real fingerprint recognition system, in which the fingerprint images are contaminated by noises heavily.

5. ACKNOWLEDGMENTS

The authors would like to address appreciations to anonymous reviewers for their valuable comments and suggestions to improve the presentation of this paper. This work was supported by the Jiangsu Postdoctoral Science Foundation (Grant No. 0901077C), the HHIT Science Foundation (Grant No. Z2008035) and Postdoctoral Science Foundation of China (Grant No. 20090451167).

REFERENCES

[1] J. P. Yin, E. Zhu and X. J. Yang. (2007) Two steps for fingerprint segmentation. Image and Vision Computing, **25**, 1391–1403.

[2] D. Zhang and Y. Wang. (2006) Medical image segmentation based on FCM clustering and rough set. Chinese Journal of Scientific Instrument, **27**, 1683–1687.

[3] W. J. Chen, M. L. Giger and U. Bick. (2006) A fuzzy c-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images. Academic Radiology, **13**, 63–72.

[4] J. M. Gorriz, J. Ramirez and E. W. Lang. (2006) Hard c-means clustering for voice activity detection. Speech Communication, **48**, 1638–1649.

[5] K. S. Chuang, H. L. Tzeng and S. W. Chen. (2006) Fuzzy

c-means clustering with spatial information for image segmentation. Computerized Medical Imaging and Graphics, **30**, 9–15.

[6] J. Li, X. B. Gao and L. C. Jiao, (2006) A new feature weighted fuzzy clustering algorithm, Acta Electronica Sinica, **34**, 89–92.

[7] S. C. Chen and D. Q. Zhang. (2004) Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure. IEEE Trans. Systems Man Cybernet, B, **34**, 1907–1916.

[8] L. Szilagyi, Z. Benyo and S. M. Szilagyii. (2003) MR brain image segmentation using an enhanced fuzzy c-means algorithm. 25th Annual International Conference of IEEE EMBS, 17–21.

J. Biomedical Science and Engineering, 2009, 2, 661-664 **JBiSE**
doi: 10.4236/jbise.2009.28097 Published Online December 2009 (http://www.SciRP.org/journal/jbise/).

Kolmogorov entropy changes and cortical lateralization during complex problem solving task measured with EEG

Lian-Yi Zhang

School of Electrical Engineering, Shanghai Dianji University, Shanghai, China.
Email: D310zlyi@sohu.com

Received 25 July 2009; revised 2 September 2009; accepted 3 September 2009.

ABSTRACT

The objective is to study changes in EEG time-domain Kolmogorov entropy and cortical lateralization of brain function areas during complex problem solving mental task in healthy human subjects. EEG data for healthy subjects are acquired during complex problem solving mental task using a net of 6 electrodes. The subject was given a nontrivial multiplication problem to solve and the signals were recorded for 10s during the task. Kolmogorov entropy values during the task were calculated. It was found that Kolmogorov entropy values were obviously greater in P4 channel (right) than ones in P3 channel (left) during complex problem solving task. It indicated that all subjects presented significant left parietal lateralization for the total frequency spectrum. These results suggest that it may be possible to noninvasively lateralize, and even eventually localize, cerebral regions essential for particular mental tasks from scalp EEG data.

Keywords: Wada Test; Cortical Lateralization; EEG; Brain Function Area; Complex Problem Solving

1. INTRODUCTION

Functional lateralization is an important organizing principle of the human brain. The left and right cortices have different specializations and each contributes to the performance of most cognitive tasks. The cortical lateralization can be confirmed through experiments such as injecting sodium amytal into a hemisphere (Wada test), measuring RT (Radioisotope Tracer), the split brain, hemispherectomy, and so on. Being invasive procedures, these traditional techniques are associated with some risk and discomfort. Because of this, in recent years, a burgeoning interest has developed in replacing traditional techniques with noninvasive measures. Some of these newer techniques, like the Wadatest, are based on "deactivation" of the cortex, such as repetitive transcranial magnetic stimulation [1], whereas other methods are based on structural imaging analyses [2]. However, the most promising novel noninvasive methods include direct measures of physiological activation. Some newer methods include event-related brain potentials, and whole-head magnetoencephalography. Neuroimaging studies (PET, FMRI, CT, SPECT) have also help localize lateralization effects in specific cortical areas [3,4]. All of these newer methods have variable limitations, and none have yet supplanted the Wadatest as the "old standard" for lateralizing cerebral dominance [5]. Among the aforementioned approaches, the most simple and least costly are based on scalp EEG measurements. Multichannel EEG devices are readily available in most clinical sites and could thus be used to replace the invasive traditional methods. A common practice in estimating human brain activity during performance of a mental task is also to process the electroencephalogram (EEG) in order to detect signal changes that could be related to mental processes.

The aim was to establish an EEG-based lateralization test. Here we presented initial results for four subjects examined in a complex problem solving task with 6-channel EEG time-domain Kolmogorov entropy computations. It was found that Kolmogorov entropy values were obviously greater in P4 channel (right) than ones in P3 channel (left) during complex problem solving task. It indicated that the subjects presented significant left parietal lateralization for the total frequency spectrum. These results were similar to those from the Wadatest. The notation KE is used to emphasize that it is the Kolmogorov entropy values of the time-domain EEG data.

The rest of the paper is organized as follows. Section 2 explains the methods proposed in this paper. Experiment task and data collection are described in Section 3. Rresults is in Section 4. Conclusions and Discussions are given in Section 5.

2. METHOD

2.1. Algorithm

Kolmogorov entropy (KE) describes the rate at which information about the state of the dynamic process is lost with time. Known as metric entropy, divide phase space

into D-dimensional hypercubes of content e^D. Let $P_{i_0,\cdots,i_n}$ be the probability that a trajectory is in hypercube i0 at t=0, i1 at t=T, i2 at t=2T, etc. Then define

$$K_n = -\sum_{i_0,\cdots i_n} P_{i_0,\cdots i_n} \ln P_{i_0,\cdots i_n} \quad (1)$$

where KN+1-KN is the information needed to predict which hypercube the trajectory will be in at (n+1)T given trajectories up to nT. The Kolmogorov entropy is then defined by

$$KE \equiv \lim_{T\to 0}\lim_{e\to 0^+}\lim_{N\to\infty}\frac{1}{NT}\sum_{n=0}^{N-1}(K_{n+1}-K_n) \quad (2)$$

The calculation of KE from a time series typically starts from reconstructing the system's trajectory in an embedding space. The EEG signals can reflect the state of brain activity. The EEG can be represented by projections of all variables in a multi-dimensional state space. Let $x_i, i=1,\cdots,N$ be a sample series of EEG. It is a discrete time series. Then, a m-dimensional time delay vector (in an N-dimensional space) X(n) can be constructed as follows:

$$X(n)=\{x(n),x(n+\tau),x(n+2\tau),\cdots,x(n+m-1)\tau\} \quad (3)$$

where τ is the time delay and m is the embedding dimension of the system. Then we can calculate the correlation sum Cm(e) introduced by Grassberger and Procaccia [6]:

$$C_m(e,N_m)=\frac{2}{N_m(N_m-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\Theta(e-\|x_i-x_j\|) \quad (4)$$

$$N_m = N-(m-1)\tau \quad (5)$$

where Θ is the Heaviside step function, $\Theta(x)=0$ if $x\le 0$ and $\Theta(x)=1$ for x>0. e is a given distance in a particular norm. If an attractor is present in the time series, the values $C_m(e,N_m)$ would satisfy $C_m(e,N_m)\propto e^D$, where D is the correlation dimension of the attractor and given by:

$$d_m(N_m,e)=\frac{\partial \ln C_m(e,N_m)}{\partial \ln e} \quad (6)$$

$$D=\lim_{e\to 0}\lim_{N_m\to\infty} d_m(N_{m,}e) \quad (7)$$

If e is small enough and d_m does not vary with m, Kolmogorov entropy (KE) can be calculated by the following equation:

$$KE=\lim_{e\to 0}\lim_{m\to\infty}\frac{1}{\tau}\log\left(\frac{C_m(e)}{C_{m+1}(e)}\right) \quad (8)$$

Here τ is the delay time. Higher and finite positive KE suggests chaos.

Actually, EEG signal is always changing with time. So the KE of EEG is not a constant over time. To measure the unorderly degree of EEG signal, mean Kolmogorov entropy within one second was introduced:

$$\text{Mean} \qquad KE=\frac{1}{N}\sum_{n=1}^{N}KE(n) \quad (9)$$

In the following of this paper the time series indicated EEG time series and KE refered the mean KE of EEG.

2.2. Complex Problem Solveing Task

EEG data were acquired during the task using a net of 6 electrodes. The electrodes were placed at O1, O2, P3, P4, C3 and C4 reference to the 10-20 system to record the EEG data in the experiment. Recordings were made with reference to the A1 and A2, electrode by using a high-pass filter of 0.1 Hz and a low-pass fillter of 100 Hz. **Figure 1** shows the placement of electrodes. The impedances of all electrodes were kept below 5 KΩ. The EEG was acquired with a sampling rate of 250 Hz, that is, 250 samples/second. The signals were recorded for 10s during the task, so each segment gave 2500 samples per channel. The data were recorded using an IBM-AT controlling a Lab Master analog to digital converter with 12 bits of accuracy.

Data for four subjects were used for this study. The subjects are mail. Subjects 1 and 2 were employees of a university. Subject 1 was left-handed and aged 48. Subject 2 was right-handed and aged 39. Subject 3 and subject 4 were right-handed college students. Subjects were placed in a dim, sound controlled room. The subject was given a nontrivial multiplication problem to solve and, as in all of the tasks, was instructed not to vocalize or make over movements while solving the problem.

3. RESULTS AND DISCUSSIONS

In order to obtain the data of spontaneous EEG signals, a FIR with bandpass filter 0.5-30 Hz was used. At the beginning of calculation KE of four subjects' EEG signal piece by piece, first 4 seconds (1000 samples) data are chosen as basic data and step length is 25 samples (the samples within 0.1s). To compare the effects of left hemisphere and right hemisphere during different mental

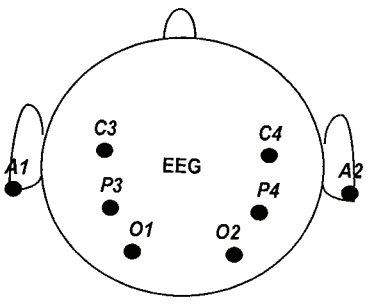


Figure 1. Electrode placement.

tasks, the mean KE of 5th, 6th, 7th, 8th, 9th, 10th second of every subject in different channel are calculated respectively. For the same time interval and the same subject, the left channel's mean KE and the right channel's mean KE of each cortical function area consist of a pair of data. So for each brain function area (Central, Parietal and Occipital) under the mental task, every subject has 6 pair of KE data.

The properties of KE for different types of dynamics are: KE=0 implies an ordered system, KE=∞ corresponds to a totally stochastic situation. The higher the KE, the closer to a stochastic the system is. So for the bilateral of the same cortical function, the small value of KE corresponds to the dominant hemisphere.

Figure 2 shows the 24 pair of KE data in parietal area for all subjects during the task. From **Figure 2**, it can be seen that the KEs in P4 channel (right) are obviously greater than that in P3 channel (left). It means that all subjects presented significant left parietal lateralization for the total frequency spectrum ($\bar{d}$ =2.0714, S_d=3.352, n=24, t=3.0274, 0.01<p<0.05). It can also be found from **Figure 2** that sometime the mean KE on the right (P4) is smaller than that on the left (P3) and this indicate that sometime the right half may be the dominant hemisphere.

There is no significant diference of KE changes in the central area and occipital area. It may mean that there is no significant lateralization in the central area and occipital area.

There is no significant diference of KE changes for right-handed and left-handed in the experiment. It may mean that there is no significant lateralization in central, parietal and occipital brain function area during the mental task for right-handed and left-handed.

Despite their low spatial resolution, electrophysiological measurements have succeeded in showing accurately the time-course of stimulus processing in the human brain. Inparticular, the early phases of cortical processing can be detected by EEG or MEG techniques alone [6]. The changes in EEG time-domain Kolmogorov entropy (KE) and localization of related cortical areas during the complex problem solving mental task in human subjects.

The mean KE on the left (P4) is not always greater than that on the right (P3). Sometime the mean KE of the dominant hemisphere is larger than that of the non-ominant. It means that sometime the right half may be the dominant hemisphere. That is to say, the dominant hemisphere is not always the same one during the same task. This indicates that the dominant hemisphere is not always the one that actuall controls performance on a particular task. This is consistent with previous known studies [7].

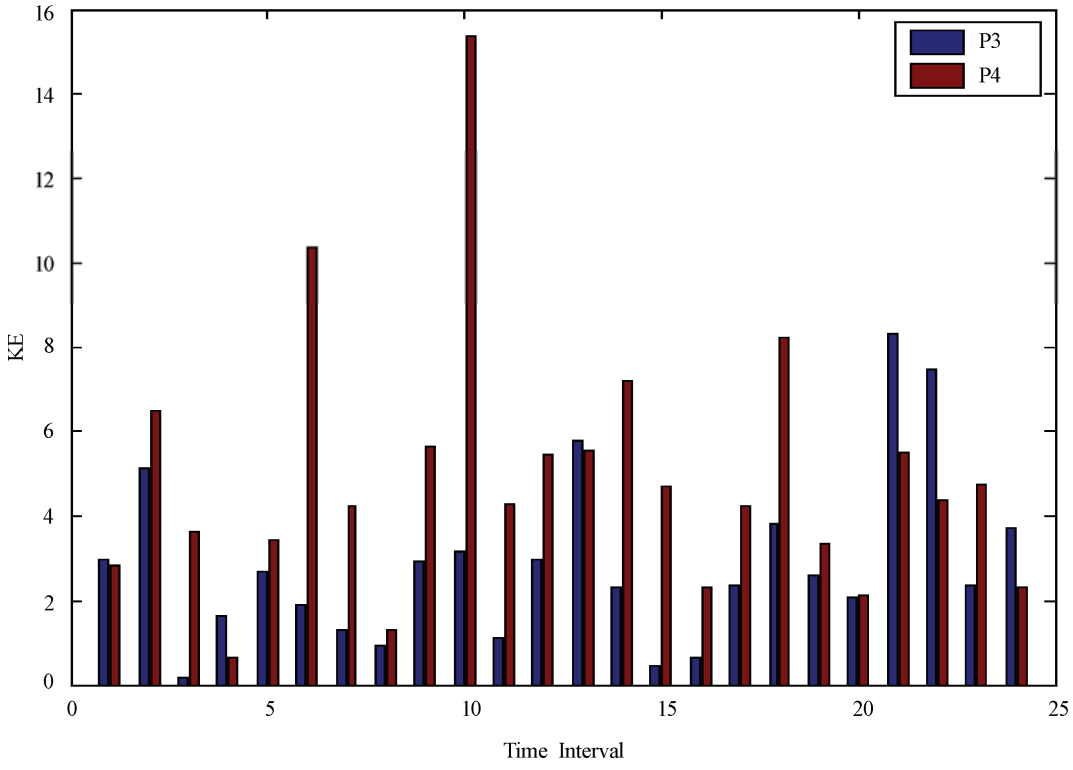


Figure 2. Comparison of KE between P3 and P4 during the task.

When the value mean KE on the left (P3) is contrasted with the one on the right (P4), the difference is very apparent and the advantage is not always the same. This may means that the advantage of dominant hemisphere is always changing. The greater the difference, the more advantageous the dominant hemisphere.

There may be no significant lateralization diference of KE changes for right-handed and left-handed in central, parietal and occipital brain function area during the task. It may indicate that the difference between right-handed and left-handed is not always existentent in different brain function areas and in different mental tasks.

4. CONCLUSIONS

During the complex problem solving mental task, the subjects presented significant left parietal lateralization for the total frequency spectrum. There is no significant lateralization in the central and occipital area during the task. The dominant hemisphere is not always the same one during the same task. The lateralization determined by Kolmogorov entropy of EEG proposed in this paper is consistent with previous known studies. The Klmogorov entropy changes of EEG can describe the cortical lateralization.

The lateralization for some particular mental task may involve in several brain areas synchronously. It is invasive, convenient and useful to analyze and to localize the different brain function area with Kolmogorov entropy of EEG.

Our results suggested lateralization of the complex problem solving task area to the left hemisphere. The results reported here do not replace the results obtained with the Wada test and other techniquies, but supplement them. In summary, these results suggest that it may be possible to noninvasively lateralize, and even eventually localize, cerebral regions essential for particular mental tasks from scalp EEG data. This could be very helpful in presurgical planning. These findings are preliminary and need to be further studied in a large population base.

REFERENCES

[1] A. Pascual-Leone, J. R. Gates, A. Dhuna. (1991) Induction speech arrest and counting errors with rapid-rate transcranial magnetic stimulation [J]. Neurology, **41**, 697–702.

[2] P. D. Charles, Abou-KhalilR, Abou-KhalilB, *et al*. (1994) MRI asymmetries and language dominance [J]. Neurology, **44**, 2050–2054.

[3] J. E. Adock, R. G. Wise, J. M. Oxbury *et al*. (2003) Quantitative fMRI assessment of the differences in lateralization of language-related brain activation in patients with temporal lobe epilepsy [J]. Neuroimage, **18(2)**, 423–438.

[4] P. Brockway and P. John. (2005) fMRI may replace the Wada test for language lateralization/localization [J]. Neuroimage, **11(5)**, S277.

[5] C. Ramon, M. Holmes, J. F. Walter, *et al*. (2009) Power spectral density changes and language lateralization during covert object naming tasks measured with high-density EEG recordings [J]. Epilepsy&Behavior, **14**, 54–59.

[6] P. Grassberger and I. Procaccia. (1983) Measuring the strangeness of strange attractors, Physica, **9D**, 189–209,.

[7] C. Spironelli and A. Angrilli. (2009) Developmental aspects of automatic word processing: Language lateralization of early ERP components in children, young adults and middle-aged subjects [J]. Biological Psychology, **80**, 35–45.

[8] J. B. Hellige. (1993) Hemispheric asymmetry: What's right and what's left [M]. Cambridge, MA: Harvard University Press.

Call for Papers

The 4th International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2010)

June 18-20, 2010 Chengdu, China

The 4th International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2010) will be held from June 18th to 20th, 2010 in Chengdu, China. You are welcome to share your recent advances and achievements in all aspects of bioinformatics and biomedical engineering on the conference. **And all accepted papers in iCBBE 2010 will be published by IEEE and indexed by Ei Compendex and ISTP.**

Topics

Bioinformatics and Computational Biology

- Protein structure, function and sequence analysis
- Protein interactions, docking and function
- Computational proteomics
- DNA and RNA structure, function and sequence analysis
- Gene regulation, expression, identification and network
- Structural, functional and comparative genomics
- Computer aided drug design
- Data acquisition, analysis and visualization
- Algorithms, software, and tools in Bioinformatics
- Any novel approaches to bioinformatics problems

Biomedical Engineering

- Biomedical imaging, image processing & visualization
- Bioelectrical and neural engineering
- Biomechanics and bio-transport
- Methods and biology effects of NMR/CT/ECG technology
- Biomedical devices, sensors and artificial organs
- Biochemical, cellular, molecular and tissue engineering
- Biomedical robotics and mechanics
- Rehabilitation engineering and clinical engineering
- Health monitoring systems and wearable system
- Bio-signal processing and analysis
- Biometric and bio-measurement
- Biomaterial and biomedical optics
- Other topics related to biomedical engineering

Special Sessions

Biomedical imaging
Biostatistics and biometry
The information technology in bioinformatics
Environmental pollution & public health

Sponsors

IEEE Eng. in Medicine and Biology Society, USA
Gordon Life Science Institute, USA
University of Lowa, USA
Wuhan University, China
Sichuan University, China
Journal of Biomedical Science and Engineering, USA

Important Dates

Paper Due: Oct.30, 2009
Acceptance Notification: Dec.31, 2009
Conference: June 18-20, 2010

Contact Information

Website:http://www.icbbe.org/2010/
E-mail: submit@icbbe.org

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 and Computational Biology
- Biomedical modeling
- Biomedical imaging, image processing and visualization
- Clinical engineering, wearable and real-time health monitoring systems
- Biomechanics and biotransport
- Software, tools and application in medical engineering
- Biomaterials
- Physiological signal processing
- Biomedical devices, sensors, artificial organs and nano technologies
- NMR/CT/ECG technologies and electromagnetic field simulation
- Structure-based drug design

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 December 2009 in SciRes (http://www.scirp.org/journal/jbise).

TABLE OF CONTENTS

Volume 2, Number 8, December 2009

9771937687001011