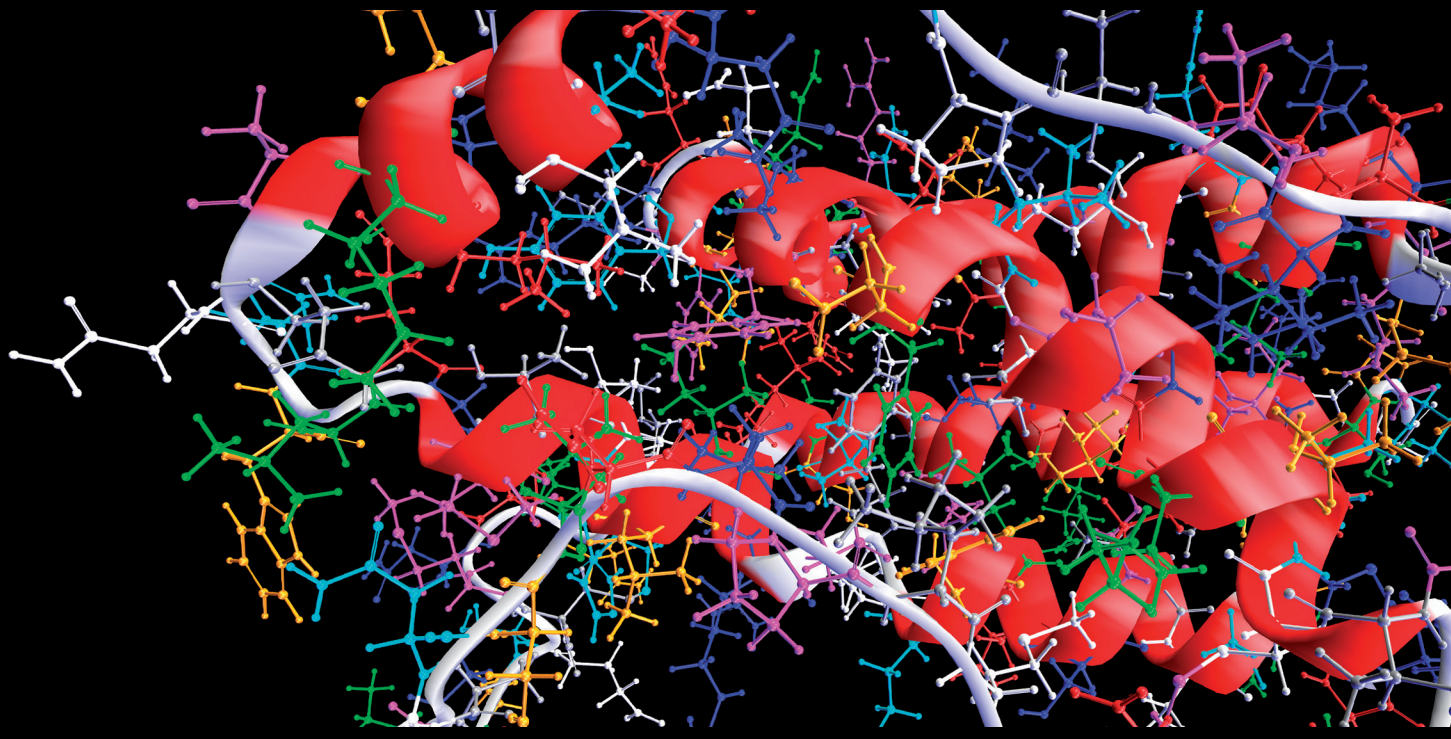
Computational and Mathematical Methods in Medicine

Novel Computational Methods and Tools in Biomedicine and Biopharmacy

Guest Editors: Yudong Cai, Tao Huang, Lei Chen, Shan Gao, and Ning Zhang



Novel Computational Methods and Tools in Biomedicine and Biopharmacy

Computational and Mathematical Methods in Medicine

Novel Computational Methods and Tools in Biomedicine and Biopharmacy

Guest Editors: Yudong Cai, Tao Huang, Lei Chen, Shan Gao, and Ning Zhang

Editorial Board

Contents

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 127515, 2 pages
http://dx.doi.org/10.1155/2014/127515

Editorial

Novel Computational Methods and Tools in Biomedicine and Biopharmacy

Yudong Cai,[1] Tao Huang,[2] Lei Chen,[3] Shan Gao,[4] and Ning Zhang[5]

[1] *Institute of Systems Biology, Shanghai University, Shanghai 200444, China*
[2] *Department of Genetics and Genomics Sciences, Mount Sinai School of Medicine, New York, NY 10029, USA*
[3] *College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China*
[4] *Boyce Thompson Institute for Plant Research, Cornell University, Ithaca, NY 14853, USA*
[5] *Department of Biomedical Engineering, Tianjin Key Laboratory of BME Measurement, Tianjin University, Tianjin 300072, China*

Correspondence should be addressed to Yudong Cai; cai_yud@126.com

Received 22 May 2014; Accepted 22 May 2014; Published 5 June 2014

With the rapid development of new biomedical and biopharmacy data collection technologies, it is desired to develop corresponding methods and tools for analyzing these big data with various structures. Such efforts can help deriving important information and knowledge from these data to promote the development of biomedicine and drug design.

In this special issue, eight interesting studies were included. Novel methods were proposed for analyzing unconventional data, such as biomedical annotation and annotator data, PET/CT Images. And useful tools were developed for interdisciplinary research, such as facial visualization system.

Q.-D. Fan et al. analyzed the cell apoptosis network at posttranslational level with time delay differential equations. Their results suggest that posttranslational modifications of p53 have different dynamics and functions. The article provides a dynamical insight of p53-induced cell repair and cell apoptosis.

A. Wu proposed a weighted and concept-extended resource description framework (RDF) model to rank the annotations by evaluating their correctness according to user's vote and the semantic relevancy between the annotator and the annotated entity. This approach is applicable and efficient even when data set is large.

W. Kong et al. reconstructed the signaling pathways of Alzheimer's disease by combining protein-protein interaction (PPI) data with gene expression data. They found that the genes on the reconstructed pathways play crucial roles in inflammatory response and APP (amyloid precursor protein).

W. Hongwei et al. described a new data visualization system by plotting the human face to observe the comprehensive effects of multivariate data. The graphics device interface (GDI+) in the Visual Studio.NET development platform was used to generate facial image according to Z values from sets of normal data.

Y. Guo et al. proposed a robust method for automatic lung tumor segmentation on PET/CT image. This method is based on fuzzy Markov random filed model and can achieve effective lung tumor segmentation even when tumors locate near other organs with similar intensities in PET and CT images.

H. Zhao et al. employed the hexaMplot to illustrate the continuous variation of the gene expressions of the embryonic cells treated with the different doses of tachyplesin I (TP I). The technology of hexaMplot was proved to be an intuitive and effective tool to illustrate the genetic interrelations in microarray analysis.

Y. Cui et al. proposed a novel case allocation system, MACT, using minimization method. This system employs a simplified database and has a unified interface that manages trials, participants, and allocations. Applications show that

MACT is stable, manageable, and easy-to-use. Its outstanding features are attracting more random clinical trials.

X. Qiu et al. proposed a hybrid method Quad-PRE to predict protein quaternary structure attributes using the properties of amino acid, predicted secondary structure, predicted relative solvent accessibility, position-specific scoring matrix profiles, and motifs. The overall accuracy of Quad-PRE is 81.7%. Quad-PRE can classify protein quaternary structure attributes effectively.

As more biomedical and biopharmacy data will be generated in the future and the data structure will be even more complex, the methods and tools in this special issue may become important and inspire other researchers.

Yudong Cai
Tao Huang
Lei Chen
Shan Gao
Ning Zhang

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 401201, 6 pages
http://dx.doi.org/10.1155/2014/401201

Research Article

Automatic Lung Tumor Segmentation on PET/CT Images Using Fuzzy Markov Random Field Model

Yu Guo,[1] Yuanming Feng,[1,2] Jian Sun,[2] Ning Zhang,[1] Wang Lin,[1] Yu Sa,[1] and Ping Wang[2]

[1] *Tianjin Key Lab of BME Measurement, Tianjin University, Tianjin 300072, China*
[2] *Department of Radiation Oncology, Tianjin Medical University Cancer Institute and Hospital, Tianjin 300060, China*

Correspondence should be addressed to Ping Wang; wangping@tjmuch.com

Received 28 March 2014; Accepted 12 May 2014; Published 29 May 2014

Academic Editor: Lei Chen

The combination of positron emission tomography (PET) and CT images provides complementary functional and anatomical information of human tissues and it has been used for better tumor volume definition of lung cancer. This paper proposed a robust method for automatic lung tumor segmentation on PET/CT images. The new method is based on fuzzy Markov random field (MRF) model. The combination of PET and CT image information is achieved by using a proper joint posterior probability distribution of observed features in the fuzzy MRF model which performs better than the commonly used Gaussian joint distribution. In this study, the PET and CT simulation images of 7 non-small cell lung cancer (NSCLC) patients were used to evaluate the proposed method. Tumor segmentations with the proposed method and manual method by an experienced radiation oncologist on the fused images were performed, respectively. Segmentation results obtained with the two methods were similar and Dice's similarity coefficient (DSC) was 0.85 ± 0.013. It has been shown that effective and automatic segmentations can be achieved with this method for lung tumors which locate near other organs with similar intensities in PET and CT images, such as when the tumors extend into chest wall or mediastinum.

1. Instruction

Combination of positron emission tomography (PET) and CT images provides complementary functional and anatomical information which has been used for tumor volume definition in radiation treatment (RT) planning for lung cancer patients [1]. Automatic methods for tumor segmentation on PET/CT images are highly desired to avoid the inter- and intraobserver variability caused by manual method.

Many automatic tumor segmentation techniques for identification and delineation of cancerous tissues have been reported such as for brain tumor [2], lung tumor [3], and prostate tumor [4]. The segmentation can be performed either on a single image set, such as CT [5], PET [6], or MRI images [7], or on the fused image set of different image modalities such as CT/PET [8–10] or multiparametric MRI images [2, 11]. Different types of tumors and image modalities have different image features, and thus different segmentation strategy should be developed for effective and accurate tumor segmentations.

CT images provide anatomical information with high spatial resolution. However, for the lung tumors abutting or involved in adjacent structures such as chest wall, mediastinum, or diaphragm which show intensities similar to those of tumors on the images, it is difficult to distinguish them from the adjacent tissues with commonly used autosegmentation algorithms. Lung tumors can be distinguished from the adjacent tissues on PET images, but the segmentation accuracy is still limited due to the coarser spatial resolution of the image data and motion artifacts as the result of time-consuming procedure of data acquisition. Therefore, one of the key points of lung tumor segmentation on PET/CT images is to combine the advantages of the two image modalities effectively.

Several methods [8–10] are proposed for lung tumor segmentation on PET/CT images. Most of the reported methods

fuse different features extracted from PET and CT images inside one single N-dimensional vector. In this paper, we propose a new strategy for fusing PET and CT information. The method is based on fuzzy Markov random field (MRF) model which has shown effective performance for unsupervised image segmentation [12, 13]. Different from traditional fuzzy MRF method, the proposed method designs a new joint posterior probabilistic model for effective combination of PET and CT image information. The new method was evaluated using image data of 7 patients with lung cancer in this study and experimental results showed its good performance in automatic tumor delineation.

This paper is organized as follows. We first present the basic theory about image segmentation using fuzzy MRF model in Section 2.1. The framework of lung tumor segmentation on CT/PET images using fuzzy MRF model is then described in Section 2.2. The evaluation of the proposed method and quantification results are shown in Section 3. Finally, discussion and conclusion are presented at the end.

2. Materials and Methods

2.1. Image Segmentation Based on Fuzzy MRF Model. The fuzzy MRF model is an unsupervised statistical methodology that takes place in Bayesian framework. Image segmentation based on fuzzy MRF model requires modeling two random fields [14]. For the set of pixels $S = \{1, \ldots, N\}$ of an image to be segmented, $Y = (y_s)_{s \in S}$ is the observed random field which represents the observed image and takes its value in the set of real numbers, while $X = (x_s)_{s \in S}$ is the unobserved random field, which corresponds to the final segmentation results and takes its value in the set of $\{1, 2, \ldots, k, \ldots, K\}$, with K being the number of classes. In comparison to the standard implementation where only a finite number of hard classes are considered, fuzzy segmentations allow each pixel to belong simultaneously to more than one class. From this point of view, x_s which is the realization of the random field X for the pixel at location s should be associated with a vector $[x_{s1}, x_{s2}, \ldots, x_{sk}, \ldots, x_{sK}]^T$ with $x_{s1} + x_{s2} + \cdots + x_{sk} + \cdots + x_{sK} = 1$, where x_{sk} is the membership degree of the pixel to class k. The segmentation problem consists of estimating $\mathbf{x}_s = [x_{s1}, \ldots, x_{sk}, \ldots, x_{sK}]^T$ $(s \in S)$ from the available noisy observation.

The relationship between X and Y can be modeled by the joint distribution $P(X, Y)$. According to Bayesian principle, we have $P(X, Y) = P(Y \mid X)P(X)$, where $P(X)$ is the prior distribution of X and $P(Y \mid X)$ is the posterior distribution. In the fuzzy MRF model, $P(X)$ is assumed to be stationary and Markovian.

Image segmentation problem is considered as a maximum a posteriori (MAP) problem. That is to estimate the membership degree matrix $\mathbf{x} = [\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_N]^T$ of the studied image which maximizes probability density function (PDF) $p(\mathbf{x} \mid \mathbf{y})$, where N is the number of pixels in the image and $\mathbf{y}$ represents the intensity or feature vector of the image. There is $\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{y})$. In Bayesian framework, $p(\mathbf{x} \mid \mathbf{y}) = p(\mathbf{y} \mid \mathbf{x})p(\mathbf{x})/p(\mathbf{y})$, where $p(\mathbf{y})$ can be considered independent of $p(\mathbf{x})$. Therefore, we have

$$\begin{aligned} \hat{\mathbf{x}} &= \arg\max_{\mathbf{x}} p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x}) \\ &= \arg\max_{\mathbf{x}} \left[\ln p(\mathbf{y} \mid \mathbf{x}) + \ln p(\mathbf{x})\right]. \end{aligned} \tag{1}$$

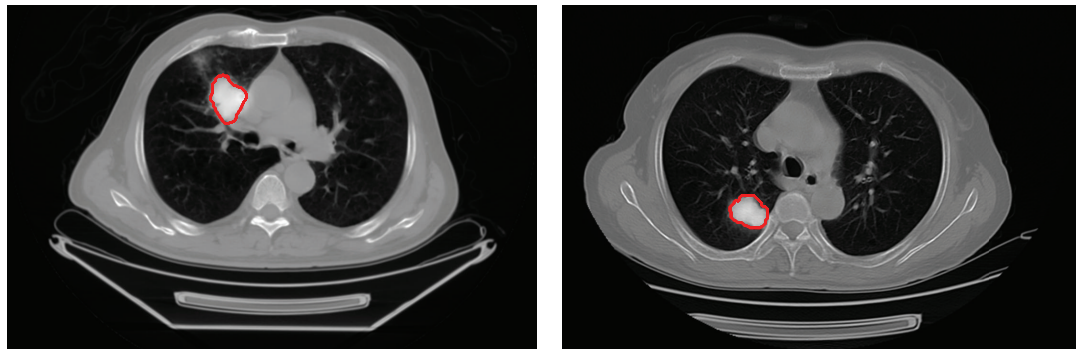
The probabilistic models used for $p(\mathbf{y} \mid \mathbf{x})$ and $p(\mathbf{x})$ are based on the prior knowledge of the studied image. In fuzzy MRF model, X is set as Gibbs distribution for the reason that pixels tend to belong to the same class with their neighbors. For the posterior distribution $p(\mathbf{y} \mid \mathbf{x})$, Gaussian distribution is usually used, since, for a region with common properties in images, it is reasonable to assume that the intensity or other feature values are distributed around the mean value of the class. As long as proper probabilistic models are determined, the membership degree matrix $\mathbf{x}$ can be estimated by solving the maximization problem in (2).

2.2. Tumor Segmentation on PET/CT Images Using Fuzzy MRF Model. In this subsection, the fuzzy MRF method for tumor segmentation on PET/CT images is described in detail. For segmenting tumor on PET/CT images, x_s is the membership degree of the voxel at location s to tumor class. y_s is related to some image features of the voxel at location s extracted from PET and CT images. Here, the CT image intensity y_{CT} and the standardized uptake value (SUV) y_{SUV} derived from PET images are used. Therefore, tumor segmentation on PET/CT image using fuzzy MRF model is actually done to estimate the membership degree to tumor class of each voxel by solving the maximization problem as shown in

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \left[\ln p(\mathbf{y}_{\mathrm{CT}}, \mathbf{y}_{\mathrm{SUV}} \mid \mathbf{x}) + \ln p(\mathbf{x})\right]. \tag{2}$$

One of the key points about the above maximization problem is to choose proper probabilistic models for $p(\mathbf{y}_{\mathrm{CT}}, \mathbf{y}_{\mathrm{SUV}} \mid \mathbf{x})$ and $p(\mathbf{x})$. Since regions which show tumor features on both PET and CT images may be tumor regions with higher possibility, we set $p(\mathbf{y}_{\mathrm{CT}}, \mathbf{y}_{\mathrm{SUV}} \mid \mathbf{x})$ as $\mathbf{x} \min[p_{11}(\mathbf{y}_{\mathrm{SUV}}), p_{21}(\mathbf{y}_{\mathrm{CT}})] + (1 - \mathbf{x}) \max[p_{10}(\mathbf{y}_{\mathrm{SUV}}), p_{20}(\mathbf{y}_{\mathrm{CT}})]$, where p_{11} and p_{21} are the PDFs of the two features given that the studied voxels belong to tumor class, while p_{10} and p_{20} are the PDFs of the two features given that the voxels belong to normal tissue class. p_{21} and p_{20} are set as Gaussian functions, since the CT intensities of tumor tissues and normal tissues around the tumor tissues are usually assumed to be distributed around the mean value of their own class. Besides, it is reasonable to assume that SUVs of normal tissues have a normal distribution, so p_{10} is also set as a Gaussian function. For $p_{11}(\mathbf{y}_{\mathrm{SUV}})$, a uniform distribution is used, which means that voxels with a SUV value greater than a threshold have the same possibility to be tumor class. The parameters of p_{11}, p_{21}, p_{10}, and p_{20} are estimated by fitting the histogram of the region obtained from C-means clustering with the selected distributions. The prior distribution of $p(\mathbf{x})$ is set as Gibbs distribution as in other fuzzy MRF methods [13]. As a result, the final maximization problem can be noted as

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} C(\mathbf{x}), \tag{3}$$



(a) (b)

Figure 1: Fused PET/CT images of two patients with lung tumors delineated by radiation oncologist shown in red lines.

with

$$C(\mathbf{x}) = \sum_i \left\{ \ln \left[x_i \min \left[p_{11}\left(y_{i\mathrm{SUV}}\right), p_{21}\left(y_{i\mathrm{CT}}\right) \right] + \left(1 - x_i\right) \max \left[p_{10}\left(y_{i\mathrm{SUV}}\right), p_{20}\left(y_{i\mathrm{CT}}\right) \right] \right] - \beta \sum_{j \in R_i} \left(x_i - x_j\right)^2 \right\}, \tag{4}$$

where i and j are the image indexes, x_i is the membership degree of the ith voxel belonging to tumor class, R_i is the neighborhood ($5 \times 5 \times 3$) of the ith voxel, and β is the smoothing parameter which affects the smoothness of segmentation results. Besides, $p_{10}(y_{i\mathrm{SUV}})$ can be computed with the following function:

$$p_{10}\left(y_{i\mathrm{SUV}}\right) = \frac{1}{\sqrt{2\pi\sigma_{10}^2}} \exp\left(-\frac{\left(y_{i\mathrm{SUV}} - \mu_{10}\right)^2}{2\sigma_{10}^2} \right), \tag{5}$$

where μ_{10} and σ_{10}^2 are, respectively, the mean and variance of the Gaussian function in (5). The function $p_{11}(y_{i\mathrm{SUV}})$ is shown as follows;

$$p_{11}\left(y_{i\mathrm{SUV}}\right) = \begin{cases} 0, & y_{i\mathrm{SUV}} < a, \\ \dfrac{1}{b-a}, & y_{i\mathrm{SUV}} \geq a, \end{cases} \tag{6}$$

where a is a SUV threshold to distinguish tumor tissues from normal tissues and b is the maximum SUV value of the PET image studied.

Gradient decent algorithm is used to solve the optimization problem in (3). The iterative form of the proposed method can be summarized as follows.

(1) Initialize the vector of membership degree $\mathbf{x}(0) = [x_1(0), x_2(0), \ldots, x_N(0)]$.

(2) Update $\mathbf{x}(n+1) = \mathbf{x}(n) - \alpha\Delta\mathbf{x}(n)$, where $\Delta\mathbf{x}(n) =$ and $\Delta x_i(n) = \partial[-C(\mathbf{x}(n))]/\partial x_i(n)$.

(3) Repeat step (2) until $\|\mathbf{x}(n+1) - \mathbf{x}(n)\|_2 \leq \varepsilon$, where $\|\cdot\|_2$ represent 2-norm and ε is the stopping threshold set as a small positive real number.

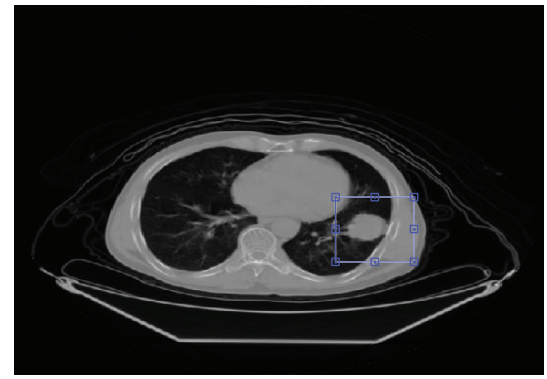


Figure 2: An axial CT slice of a patient with the region to be segmented marked.

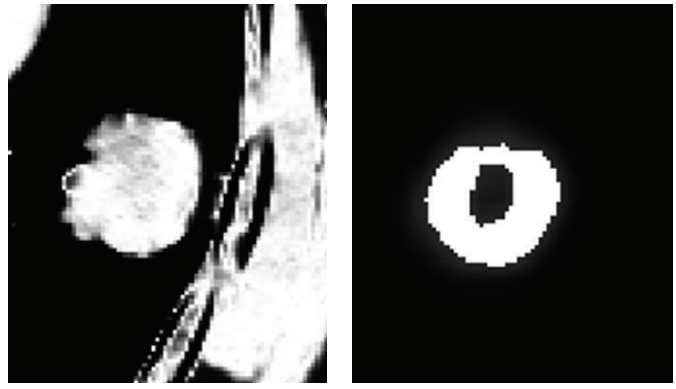
In the proposed method, the step size α and the smoothing parameter β need to be selected. A large step size may make the algorithm divergent, so using a small step size is common practice. β should be selected according to the smoothness of tumor tissues on images. Once the membership degree vector $\mathbf{x}$ is estimated, we determine the final tumor regions using a simple threshold method.

3. Experiments and Results

The PET and CT simulation images of 7 non-small cell lung cancer (NSCLC) patients were used to evaluate the proposed method. PET and CT fusion results of studied images using MIM 5.2 (MIM Software) were exported to in-house-developed software for the study. Gross tumor volume (GTV) delineation with the proposed MRF method and manual method by an experienced radiation oncologist on the fused images were performed, respectively. The manually contoured GTVs were checked and confirmed by another experienced radiation oncologist. The robustness of MRF method was evaluated by comparing the overlap of the two delineations using Dice's similarity coefficient (DSC) expressed as $2(v1 \cap v2)/(v1 + v2)$, where $v1$ is the manually delineated GTV volume and $v2$ is the GTV volume obtained with the new automatic method.

(a) (b)

Figure 3: Membership degrees to tumor class of the selected region in Figure 2 obtained with fuzzy C-means clustering; (a) only CT intensity feature is used; (b) only SUV feature is used.



(a) (b)

Figure 4: Probabilities of tumor class of the voxels within the selected region in Figure 2, (a) $p_{21}(y_{\mathrm{CT}})$; (b) $p_{11}(y_{\mathrm{SUV}})$.

Some lung tumors in the studied cases locate near other organs with similar intensities to tumor tissues in PET and CT images (such as the chest wall or the mediastinum), and it is difficult to segment these lung tumors. Figure 1(a) shows a fused PET/CT image of a patient with lung cancer; the tumor locates close to the mediastinum and has image features similar to those of the mediastinum. Figure 1(b) shows another case in which the tumor is close to the chest wall.

The proposed segmentation algorithm performs on a manually selected region rather than the whole volume enclosed by the image set. Lung tumors locate in the selected regions, as shown in Figure 2. There are much fewer tissue components in the selected region than in the whole volume which makes the segmentation easier. In order to obtain the probabilistic model parameters of p_{10}, p_{11}, p_{20}, and p_{21} in (3), the fuzzy C-means clustering [15] is applied to the regions of interest in the studied PET and CT images. Figure 3 shows the estimated membership degrees of the selected region in Figure 2 to tumor class with fuzzy C-means clustering. It is assumed that the voxels which have the membership degrees greater than 0.1 belong to tumor class. Then, we can have a rough segmentation of the selected region. According to the segmentation result, the desired probabilistic model parameters, such as the means and deviations of Gaussian models in (5), can be estimated. Figure 4 shows the probabilities of tumor class of the voxels in the selected region of Figure 2, which are computed based on the probabilistic model parameters estimated in the previous step.

Figure 5 shows the tumor segmentation results of 3 patients obtained, respectively, with the manual method, fuzzy MRF method using only PET images, and the proposed method using CT/PET images. As shown in Figure 5, lung tumors in these cases are close to other tissues with similar image features. We can also see that GTVs determined by using only PET images are bigger than the ones obtained with the other two methods in most cases, while GTVs obtained with the manual method and the proposed method are similar. For the other cases, we obtained the same results.

In our work, the traditional MRF model method in [13] and fuzzy C-means clustering are also used to segment lung tumors. CT/PET image based segmentation results obtained with the two methods are much worse than the results of the two methods using only PET images. The DSCs of the two methods for all the studied cases were 0.59 ± 0.034 and 0.62 ± 0.029, respectively. It means that the two methods

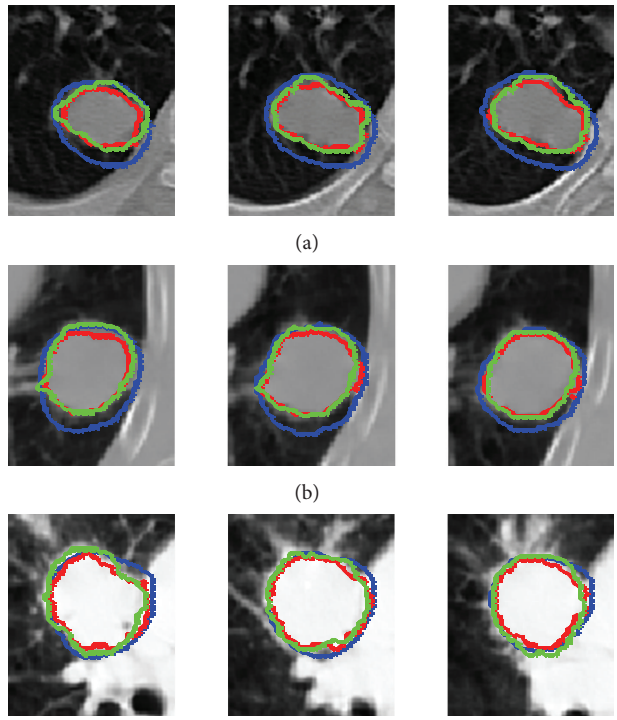


Figure 5: GTVs in axial CT slices of patient (a), (b), and (c). GTVs in blue are the results with fuzzy MRF method using only PET images and GTVs in red and green are the results with the new method and manual method, respectively, using both PET and CT images.

cannot effectively combine CT and PET image information to achieve accurate lung tumor segmentation. DSC of the proposed method for all the studied cases was 0.85 ± 0.013. Therefore, the proposed method is able to effectively utilize CT and PET image information and achieves good lung tumor segmentation.

4. Conclusions

CT/PET images provide complementary functional and anatomical information of human tissues and lead to better lung tumor definition. This paper proposes a fuzzy MRF model based method for automatic lung tumor segmentation on CT/PET images. Different from traditional fuzzy MRF model method, it utilizes a new joint posterior probabilistic model, which can effectively take advantage of both CT and PET image information for the identification and delineation of tumor volume. Experimental results show its good performance. For lung tumors which locate near other tissues with similar intensities in PET and CT images, such as when they extend into the chest wall or the mediastinum, this method was able to achieve more effective tumor segmentation. In future work, we will further test the reliability of this method with more clinical data.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by Grants from the National Natural Science Foundation of China (81171342, 81201148), Tianjin Research Program of Application Foundation and Advanced Technology (14JCQNJC09500), the National Research Foundation for the Doctoral Program of Higher Education of China (20130032120070, 20120032120073), and Independent Innovation Foundation of Tianjin University (60302064, 60302069).

References

[1] S. Katsuyuki, N. Akiko, A. Takanori et al., "Use of FDG-PET in radiation treatment planning for thoracic cancers," *International Journal of Molecular Imaging*, vol. 2012, Article ID 609545, 9 pages, 2012.

[2] W. Dou, S. Ruan, Y. Chen, D. Bloyet, and J.-M. Constans, "A framework of fuzzy information fusion for the segmentation of brain tumor tissues on MR images," *Image and Vision Computing*, vol. 25, no. 2, pp. 164–171, 2007.

[3] Y. Gu, V. Kumar, L. O. Hall et al., "Automated delineation of lung tumors from CT images using a single click ensemble segmentation approach," *Pattern Recognition*, vol. 46, no. 3, pp. 692–702, 2013.

[4] Y. Artan, M. A. Haider, D. L. Langer et al., "Prostate cancer localization with multispectral MRI using cost-sensitive support vector machines and conditional random fields," *IEEE Transactions on Image Processing*, vol. 19, no. 9, pp. 2444–2455, 2010.

[5] C. Li, X. Wang, S. Eberl et al., "A likelihood and local constraint level set model for liver tumor segmentation from CT volumes," *IEEE Transctions on Biomedical Engineering*, vol. 60, no. 10, pp. 2967–2977, 2013.

[6] M. Halt, C. C. le Rest, A. Turzo, C. Roux, and D. Visvikis, "A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET," *IEEE Transactions on Medical Imaging*, vol. 28, no. 6, pp. 881–893, 2009.

[7] H. Cui, X. Wang, and D. Feng, "Automated localization and segmentation of lung tumor from PET-CT thorax volumes based on image feature analysis," in *Proceedings of the 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS '12)*, pp. 5384–5387, San Diego, Calif, USA, September 2012.

[8] J. Wojak, E. D. Angelini, and I. Bloch, "Joint variational segmentation of CT-PET data for tumoral lesions," in *Proceedings of the 7th IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI '10)*, pp. 217–220, April 2010.

[9] H. Gribben, P. Miller, G. G. Hanna, K. J. Carson, and A. R. Hounsell, "Map-Mrf Segmentation of lung tumours in pet/ct images," in *Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI '09)*, pp. 290–293, July 2009.

[10] C. Ballangan, X. Wang, S. Eberl, M. Fulham, and D. Feng, "Automated detection and delineation of lung tumors in PET-CT volumes using a lung atlas and iterative mean-SUV threshold," in *International Society for Optics and Photonics, Image Processing*, vol. 72593 of *Proceedings of SPIE Medical Imaging*, February 2009.

[11] N. Zhang, S. Ruan, S. Lebonvallet, Q. Liao, and Y. Zhu, "Kernel feature selection to fuse multi-spectral MRI images for brain tumor segmentation," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 256–269, 2011.

[12] S. Ruan, S. Lebonvallet, A. Merabet, and J.-M. Constans, "Tumor segmentation from a multispectral MRI images by using support vector machine classification," in *Proceedings of the 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI '07)*, pp. 1236–1239, April 2007.

[13] X. Liu, D. L. Langer, M. A. Haider, Y. Yang, M. N. Wernick, and I. Š. Yetik, "Prostate cancer segmentation with simultaneous estimation of Markov random field parameters and class," *IEEE Transactions on Medical Imaging*, vol. 28, no. 6, pp. 906–915, 2009.

[14] H. Caillol, W. Pieczynski, and A. Hillon, "Estimation of fuzzy Gaussian mixture and unsupervised statistical image segmentation," *IEEE Transation on Image Processing*, vol. 6, no. 3, pp. 425–440, 1997.

[15] K.-S. Chuang, H.-L. Tzeng, S. Chen, J. Wu, and T.-J. Chen, "Fuzzy c-means clustering with spatial information for image segmentation," *Computerized Medical Imaging and Graphics*, vol. 30, no. 1, pp. 9–15, 2006.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 715494, 9 pages
http://dx.doi.org/10.1155/2014/715494

Research Article

Quad-PRE: A Hybrid Method to Predict Protein Quaternary Structure Attributes

Yajun Sheng,[1] Xingye Qiu,[1] Chen Zhang,[1] Jun Xu,[1] Yanping Zhang,[1] Wei Zheng,[1] and Ke Chen[2]

[1] *School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China*
[2] *School of Computer Science and Software Engineering, Tianjin Polytechnic University, No. 399 Binshui Road, Tianjin 300387, China*

Correspondence should be addressed to Ke Chen; kchen1.tj@gmail.com

Received 27 February 2014; Revised 24 April 2014; Accepted 27 April 2014; Published 18 May 2014

Academic Editor: Tao Huang

The protein quaternary structure is very important to the biological process. Predicting their attributes is an essential task in computational biology for the advancement of the proteomics. However, the existing methods did not consider sufficient properties of amino acid. To end this, we proposed a hybrid method Quad-PRE to predict protein quaternary structure attributes using the properties of amino acid, predicted secondary structure, predicted relative solvent accessibility, and position-specific scoring matrix profiles and motifs. Empirical evaluation on independent dataset shows that Quad-PRE achieved higher overall accuracy 81.7%, especially higher accuracy 92.8%, 93.3%, and 90.6% on discrimination for trimer, hexamer, and octamer, respectively. Our model also reveals that six features sets are all important to the prediction, and a hybrid method is an optimal strategy by now. The results indicate that the proposed method can classify protein quaternary structure attributes effectively.

1. Introduction

As is well known, the prediction of protein quaternary structure attributes (such as monomer, dimmer, trimer, tetramer, pentamer, hexamer, heptamer, and octamer) plays an important role in the structure bioinformatics. It can confirm how many subunits form the protein. It is the real requirement for the Anfinsen's dogma [1]. A variety of experimental techniques can determine protein quaternary structure. However, most methods are time-consuming and expensive. Moreover, the oligomers may be homooligomers or heterooligomers; the former consist of identical polypeptide chains, whereas the latter are nonidentical. Many computational methods are proposed.

As far as we know, the earliest work to study the quaternary structure type was in 2001 [2]. In this paper, Garian proposed a method named Quaternary Structure Explorer (QSE), which just judges whether or not a given protein is a homodimer. In 2003, Zhang et al. [3] first introduced support vector machine (SVM) to discriminate the differences of the primary sequences of both homodimer and nonhomodimer. Chou and Cai [4] solved the 2-state problem by using the pseudo amino acid composition. In 2006, Shi el al. [5] classified homooligomers based on amino acid composition distribution (AACD) and showed that the 2DPCA was an effective approach to decrease the high dimension of feature vector. In 2007, Carugo [6] proposed a method which is able to predict the quaternary structural type of hetero oligomeric proteins. Levy [7] proposed the PiQSi to get the annotations of about 15,000 proteins in PDB, which can be used as the benchmark dataset to test the quality of a method to predict the quaternary structure type. In 2009, Xiao and Lin introduced the grey incidence degree measure [8] to predict the protein quaternary structure attributes. The method is implemented as a web-server called Quat-2L [9], which firstly identifies the protein as homooligomer or

heterooligomer and secondly justifies how many subunits. In 2012, Sun et al. utilized discrete wavelet transform [10] based on Chou's PseAAC to identify the protein quaternary structure attribute. All these methods to predict the quaternary structure attributes are based on one set of features, and mostly for 2 states.

In this paper, we proposed a new method Quad-PRE to predict protein quaternary structures attributes among 6 states only based on the primary sequences, removing both pentamer and heptamer because of insufficient data. With 10 fold cross validation, our models achieved higher overall accuracy 81.7%, especially higher accuracy 92.8%, 93.3%, and 90.6% on discrimination for trimer, hexamer, and octamer, respectively. Our method could be an effective tool to predict the protein quaternary structure attributes.

2. Materials and Methods

2.1. Benchmark Dataset. The dataset is from the quaternary structure library PiQSi (http://www.PiQSi.org/) built by Levy [7]. Our original dataset was downloaded on December 12, 2011. Firstly, we download a whole annotated list including about 15,000 protein sequences and a nonredundant set including 1755 sequences (30% sequence id.) from the library and then remove sequences which are not in the nonredundant set from the whole annotated list. In order to use a set of "good" PDB files, we use the subset of those annotated as "NOT" or "PROBABLY NOT" being errors. In addition, the number of pentamer and heptamer is too little to analyze and we also removed them. Finally, we get a protein quaternary structure dataset with primary sequence as shown in Table 1.

2.2. Features. In this paper, we used three traditional methods and three tools (BLAST, GLAM2, and GIBBS) to select 632 features only based on unique primary sequences and denoted them as six terms: ART_1 feature, ART_2 feature, ART_3 feature, BLAST feature, GLAM2 feature, and GIBBS feature). The summary of the considered features is shown in Table 2 (See Tables S1–S3 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/715494 for more detailed information).

Firstly, we use three traditional methods to get the three feature sets, that is, the ART_1 feature by [12], ART_2 feature by [13], and ART_3 feature by [11], respectively. The sources of data used to generate the features from the original sequence include the protein sequence, the position-specific scoring matrix (PSSM) generated by PSI-BLAST [14], the secondary structure predicted by PSI-Pred [15], the solvent accessible surface area (ASA) values predicted using Real-SPINE [16], and the relative solvent accessibility (RSA) defined as the ratio of ASA of a residue observed in its three-dimensional structure to that observed in an extended (Gly-X-Gly or Ala-X-Ala) tripeptide conformation [17].

Secondly, we generate other three features sets by BLAST, GLAM2, and GIBBS, respectively. The three methods can describe the inherent properties of sequences. Primarily, we divide equally the feature set into 10 portions randomly, making sure that every portion contains at least one element

Table 1: The numbers of monomer, dimmer, trimer, tetramer, hexamer, and octamer in our benchmark dataset.

Total	Monomer	Dimer	Trimer	Tetramer	Hexamer	Octamer
1040	366	338	53	155	67	61

of each one of 6 states (monomer, dimmer, trimer, tetramer, hexamer, and octamer) so that we have 10 datasets

$$\{S_i \mid |S_i| = 104, i = 1, \ldots, 10\}. \quad (1)$$

Every S_i contains 6 subsets

$$S_i = \{p \mid p \in s_{ic}, c = 1, 2, 3, 4, 6, 8\}, \quad (2)$$

where each subset s_{ic} contains sequences which has c subunits in S_i. It is noted that the generated features depend on the original 10 fixed datasets.

For each sequence $P = a_1 a_2 \cdots a_L \in S_i$, we select the most similar five sequences in each one of 6 sets $\{p \mid p \in s_{kc}, k \neq i\}$, $c = 1, 2, 3, 4, 6, 8$ by PSI-Blastall. So we can get 30 features for each given sequence P based on the Evalue's index of the scientific notation from the results of the tool.

The sequence motifs can describe many properties of protein, such as transcription factor binding sites, splice junctions, and protein-protein interaction sites. Both GIBBS and GLAM2 are employed to find motifs from our datasets. In the same way, for each sequence $P \in S_i$, we get the motifs of each one of 6 sets $\{p \mid p \in s_{kc}, k \neq i\}$, $c = 1, 2, 3, 4, 6, 8$ by both GLAM2 and GIBBS, denoted as follows, respectively:

$$\begin{aligned} &M_c^{P_{\text{GLAM2}}}, \quad c = 1, 2, 3, 4, 6, 8, \\ &M_c^{P_{\text{GIBBS}}}, \quad c = 1, 2, 3, 4, 6, 8. \end{aligned} \quad (3)$$

In fact, there are many gaps in some motifs generated by GLAM2 so that we need to preprocess these motifs as follows.

(i) If a motif has more than five consecutive gaps, we delete those gaps and divide this motif into two new motifs.

(ii) If the AAs of a motif are less than five, we delete it.

Then we get updated

$$M_c^{P_{\text{GLAM2}}}, \quad c = 1, 2, 3, 4, 6, 8. \quad (4)$$

We use the modified Smith-Waterman dynamic programming (SW-DP) algorithm to make sequence alignment between the given sequence P and each one of $M_c^{P_{\text{GLAM2}}}$, $c = 1, 2, 3, 4, 6, 8$. The given sequence P acquires the five highest alignment scores from each of $M_c^{P_{\text{GLAM2}}}$, $c = 1, 2, 3, 4, 6, 8$, so that we can get 30 more features for the given sequence. The specific procedure is as follows. In fact, each position of each motif generated by GLAM2 possibly has more than one AA after preprocessing. We use

$$M_{\text{GLAM2}} = m_1 m_2 \cdots m_n \quad (5)$$

Table 2: Summary of the considered features, where y denotes one of the three secondary structure states and x denotes one of the 20 common AAs.

Feature sets	Description
Sequence-based (79)	Sequence length (1) Composition vector (20) The number of AAs in the sequence belonging to {R group, Electronic group, Hydrophobicity group, Exchange group} (18) First and second order composition moment vector (40)
PSSM-based (203)	From the PSSM matrix
Secondary structure (217)	Based on the features utilized in the PSI-Pred method (90) Based on the predicted secondary structure which describes collocation of helical and strand segments (127)
Average RSA based (23)	Average RSA of the residues with AA type x (20) Average RSA of the residues with secondary structure type y (3)
Average isoelectric point (1)	$pI = 1/N \sum_{i=1}^{N} pI_i$, the pI_i values in the paper [11]
Auto-correlation functions based on FH_i, EH_i, and Hp indices (25)	$A_n^a = 1/(N-n) \sum_{i=1}^{N-n} a_i a_{i+n}$, where a defines the corresponding physicochemical properties, such as two hydrophobicity indices (the Fauchere-Pliska's (FH) with $n = 1, 2, \ldots, 10$ and the Eisenberg's (EH) $n = 1, 2, \ldots, 6$), and hydropathy (HP) index with $n = 1, 2, \ldots, 9$.
Auto-correlation functions based on cumulative FH_i index (6)	$A_n^a = \sum_{i=1}^{N-n} \left(\sum_{j=1}^{i} a_j\right) \times \left(\sum_{j=1}^{i+n} a_j\right) / (N-n)$, where a is the FH index with $n = 1, 2, \ldots, 6$.
Sum of hydrophobicities based on FH_i and EH_i (2)	$H_{\text{sum}}^a = \sum_{i=1}^{N} a_i$, where a is the FH or the EH index.
R groups (5)	RG_i, where $i = 1$ corresponds to nonpolar aliphatic AAs (AVLIMG), $i = 2$ to polar uncharged AAs (SPTCNQ), $i = 3$ to positively charged AAs (KHR), $i = 4$ to negative AAs (DE), and $i = 5$ to aromatic AAs (FYW); the composition percentage of each group in the sequence is computed
Electronic groups (5)	EG_i, where $i = 1$corresponds to electron donor AAs (DEPA), $i = 2$ to weak electron donor AAs (LIV), $i = 3$ to electron acceptor AAs (KNR), $i = 4$ to weak electron acceptor AAs (FYMTQ), and $i = 5$ to neutral AAs (GHWS); the composition percentage of each group in the sequence is computed
Blast based (30)	Refer to subsection "*Features*"
GLAM2-based (30)	Refer to subsection "*Features*"
GIBBS-based (6)	Refer to subsection "*Features*"

to represent a motif with n length, where $m_i = \{b_{ij}\}$ and b_{ij} may be one of 20 common AAs or a gap. For the protein sequence $P = a_1 a_2 \cdots a_L$, the penalty function is defined as

$$d_{\text{GLAM2}}\left(m_i, a_j\right) = \begin{cases} 1 & \text{if } a_j \in m_i,\ a_j \neq \text{gap} \\ 0 & \text{if } a_j \in m_i,\ a_j = \text{gap} \\ -1 & \text{if } a_j \notin m_i,\ a_j \neq \text{gap} \\ -\dfrac{1}{3} & \text{if } a_j \notin m_i,\ a_j = \text{gap}. \end{cases} \tag{6}$$

Then we use the SW-DP algorithm to compute the alignment score between P and M_{GLAM2}.

In addition, GIBBS can find a motif like

$$M_{\text{GIBBS}} = t_1 t_2 \cdots t_n \tag{7}$$

for each one of $M_c^{P_{\text{GIBBS}}}$, $c = 1, 2, 3, 4, 6, 8$, where

$$t_i = \left(p_i^{b_1}, p_i^{b_2}, \ldots, p_i^{b_{21}}\right)^T, \quad i = 1, 2, 3, \ldots, n \tag{8}$$

represent probabilities of 20 common AAs and gap in the position i, and

$$\left\{b_j,\ j = 1, 2, \ldots, 21\right\} = \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V, -\}. \tag{9}$$

For the protein sequence $P = a_1 a_2 \cdots a_L$, the penalty function is defined as

$$d_{\text{GIBBS}}\left(t_i, a_j\right) = p_i^{a_j}. \tag{10}$$

We employ the SW-DP algorithm to calculate the alignment score between P and M_{GIBBS} again, and then we gain other 6 features for the sequence P by GIBBS.

2.3. The Overall Design. Gaining a protein quaternary structure dataset, we design our method Quad-PRE from primary sequence as below.

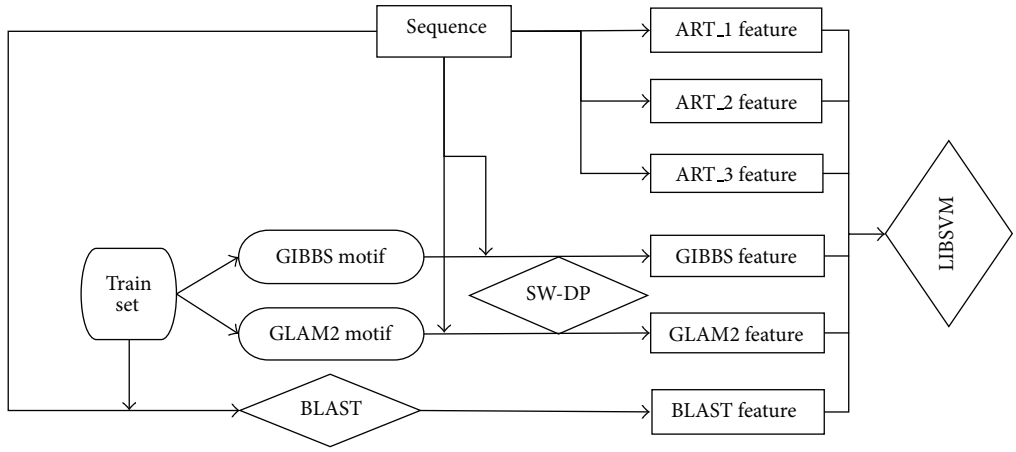


Figure 1: The diagram of Quad-PRE.

(1) Select the features based on properties of amino acid, PSSM, the secondary structure, the solvent accessible surface area, and the physicochemical property.

(2) In addition, we divide our dataset equally into ten portions randomly, but making sure that every portion contains at least one element of each one of 6 states. And then we obtain the new features of each sequence using BLAST, GIBBS, and GLAM2, respectively.

Our scheme is a hybrid method and we give a diagram for making it easy to follow, shown in Figure 1.

2.4. Classification. Support vector machine (SVM), which was shown to provide high quality predictions in classification, regression, and density estimation area, was implemented with LIBSVM [18] package. The support vector classification C-SVC is selected in this paper. There are several strategies to solve multiclass problem, such as one-versus-rest and one-versus-one. One-versus-rest strategy is used in this paper. The prediction performance was examined by n-fold cross validation, in which the training dataset is randomly divided into n subsets equally. The $n-1$ subsets are used to train the model and the remaining one subset is used to evaluate the model, repeated n times. If n is the number of the samples, it was named jackknife test (or leave-one-out cross validation).

We designed a predictor with 10-fold cross validation. First of all, the input sequence is converted into the feature space, and then the corresponding features are passed to the classifier. The prediction class of the sequence that corresponds to one has the highest probability. Overall accuracy (ACC), the sensitivity or true positive rate (TPR), the false positive rate (FPR), the specificity (SPC), the precision (PPV), and Matthew's correlation coefficient (MCC) for each class are used to measure the prediction performance; they are defined as follows:

$$\begin{aligned}
\mathrm{ACC} &= \frac{(\mathrm{TP}+\mathrm{TN})}{N}, \\
\mathrm{TPR} &= \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})}, \\
\mathrm{FPR} &= \frac{\mathrm{FP}}{(\mathrm{FP}+\mathrm{TN})}, \\
\mathrm{SPC} &= \frac{\mathrm{TN}}{(\mathrm{FP}+\mathrm{TN})} = 1-\mathrm{FPR}, \\
\mathrm{PPV} &= \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})}, \\
\mathrm{MCC} &= \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})\,(\mathrm{TP}+\mathrm{FP})\,(\mathrm{TN}+\mathrm{FP})\,(\mathrm{TN}+\mathrm{FN})}},
\end{aligned} \tag{11}$$

(12)

where TP is true positive number, TN is true negative, FP is false positive, FN is false negative, and N is total number of sequences. However, these metrics are not quite intuitive and easier-to-understand and we can adopt the formulation proposed recently to really understand them [19–21]. We also calculate the area under the ROC curve (AUC) to evaluate the predictions. Higher values of these measures indicate better quality of predictions.

3. Results and Discussion

3.1. Results and Comparison with Garian's QSE. The choice of the penalty factor C and the kernel function type is very important since SVM is sensitive to parameterization. In this paper, we consider the radial basis function (RBF) of kernel types following the Chang and lin [22]

$$K\left(x_i, x_j\right) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \tag{13}$$

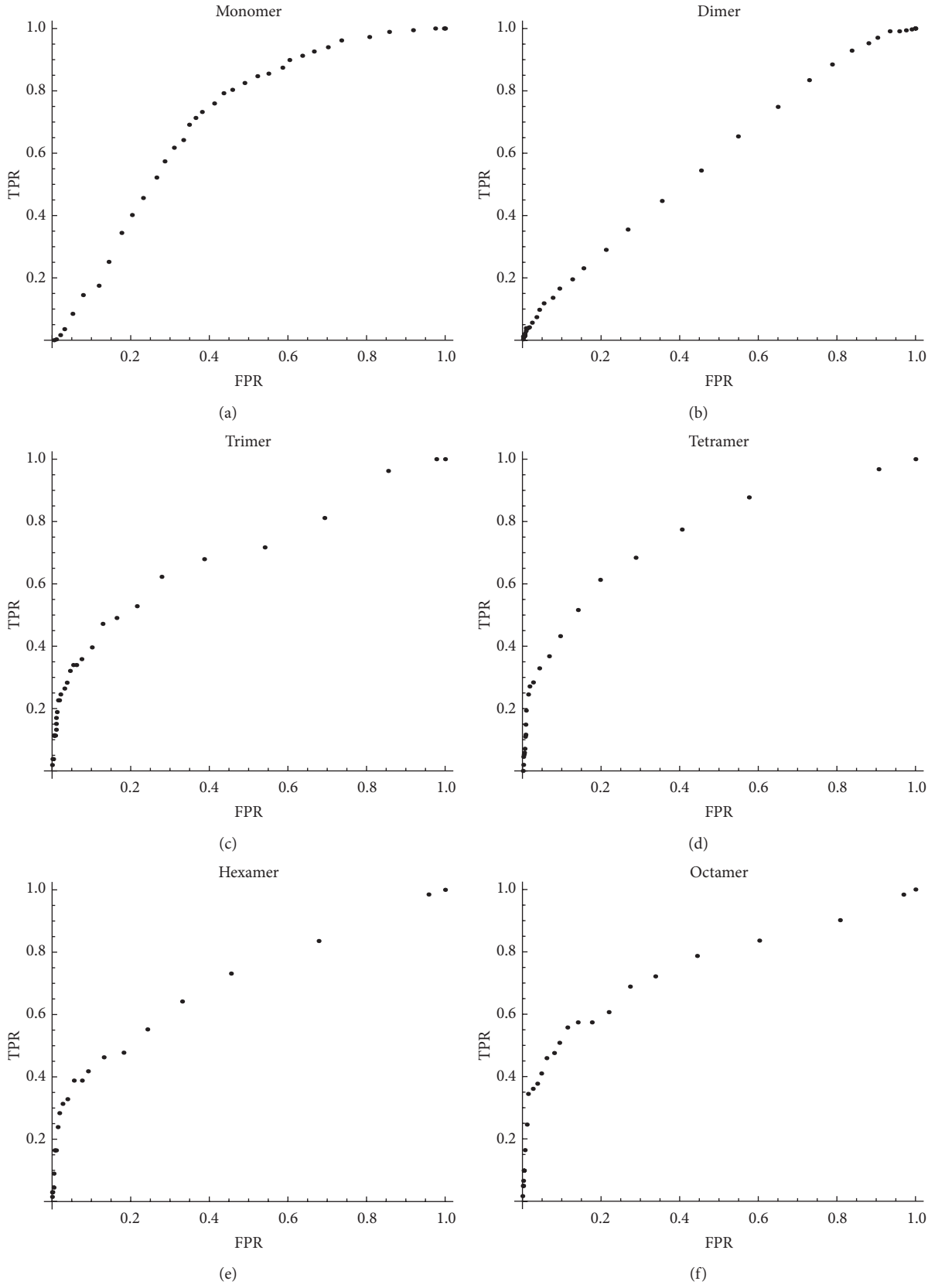


Figure 2: The ROC curves of six classes.

where γ is the width of the RBF function. To identify the optimal C and γ, a systematic grid search was conducted for

$$C = \{0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100\}\ \gamma = \{0.0025, 0.005, 0.01, 0.02, 0.04, 0.08, 0.16, 0.32, 0.64, 1.28, 2.56, 5.12, 10.24\} \quad (14)$$

by the 10-fold cross validation. Then we find the optimal C and γ are 0.1 and 0.01 with the average AUC value 0.704. With the best parameters, the average accuracy is 45.3% by 10-fold cross validation. The predicting matrix is as follows; the rw_{ij} is the number of the i class predicted as the j class

$$\mathrm{RW} = \begin{pmatrix} 285 & 55 & 12 & 1 & 5 & 8 \\ 185 & 104 & 9 & 8 & 7 & 25 \\ 24 & 9 & 13 & 0 & 1 & 6 \\ 55 & 50 & 3 & 28 & 4 & 15 \\ 18 & 20 & 6 & 0 & 16 & 7 \\ 22 & 8 & 5 & 0 & 2 & 24 \end{pmatrix}. \quad (15)$$

The TPR, SPC, PPV, MCC, and AUC of every class are shown in Table 3 and the ROC curves are shown in Figure 2. Following from Table 3, Quad-PRE achieved higher overall ACC 81.7%, especially higher accuracy 92.8%, 93.3%, and 90.6% on discrimination for trimer, hexamer, and octamer, respectively. And overall SPC is 87.0%, especially 96.5%, 99.0%, 98.0%, and 93.8% on discrimination for trimer, tetramer, hexamer, and octamer, respectively. These results show that our hybrid method has high accuracy and specificity.

In addition, we can see that it is a little more difficult to predict dimer from Figure 2, because the AUC for predicting dimer is smaller than other oligomers. More specifically, the AUC of dimer is 0.582, while those of monomer, trimer, tetramer, hexamer, and octamer are 0.703, 0.702, 0.765, 0.711, and 0.758, respectively (see Table 2). However, when comparing with the predicted results of Garian's QSE [2] of classifying homodimer and nonhomodimer, the ACC, SPC, PPV, MCC, and AUC of Quad-PRE are all larger than QSE's, other than the TPR (see Table 4). Apparently, Quad-PRE performs better than QSE's (ROC curves of two methods are shown in Figure 3).

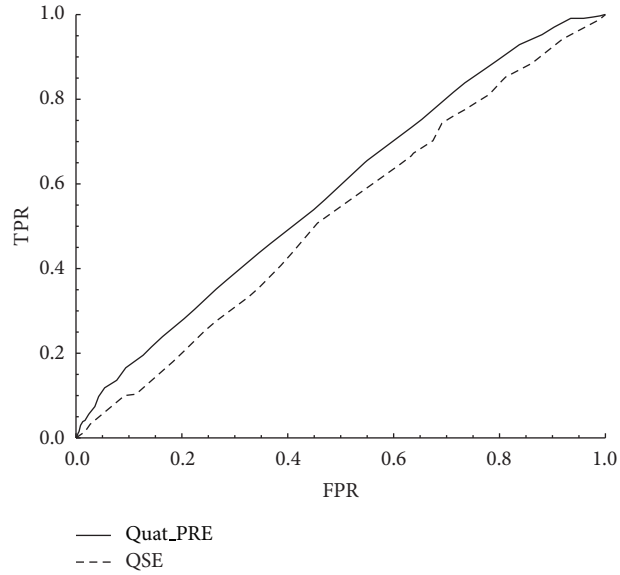


Figure 3: The ROC curves comparison Quad-PRE with Garian's QSE.

3.2. Discussion with Six Feature Groups. For confirming our generated new features (TOTAL) can improve the prediction of protein quaternary structure attributes, we compared the results from TOTAL features with those from each one of the six feature sets (ART_1, ART_2, ART_3, BLAST, GLAM2, and GIBBS), which are shown in Table 5. The ROC curves for predicting every attribute by six sets are shown in Figure 4, respectively.

From Figure 4, we can see that the average AUC, ACC, TPR, SPC,and MCC of any of 6 features sets are all smaller than TOTAL features except the PPV. In particular, there are almost the same average SPC values for all feature sets. And the two feature sets from both GIBBS and GLAM2 all do not perform well in every metric. From Table 5 we also know that ART_1, BLAST, ART_1, ART_1, BLAST, and ART_1 play key roles in improving average ACC, TPR, SPC, PPV, MCC, and AUC of our method, respectively, because the corresponding values of them are close to those of TOTAL. These results mean each feature set contributes to the improvement of our hybrid method, especially ART_1 because the average ACC, TPR, SPC, PPV, MCC, and AUC from which are almost superior to others (see Table 5).

From the view of the average AUC, the importance of the six feature sets from high to low is ART_1, ART_2, ART_3, BLAST, GLAM2, and GIBBS (see Table 5). And the AUC values of ART_1, ART_2, and ART_3 for every protein attribute are almost larger than those of BLAST, GIBBS, and GLAM2 (see Figure 4). We think that the possible reason should be that the ART_1, ART_2, and ART_3 have much more features than BLAST, GIBBS, and GLAM2. And because similar sequences should have similar structures and functions, the features from BLAST are superior to those from both GIBBS and GLAM2 in the performance of SVM.

4. Conclusions

To predict protein quaternary structure attribute is indeed a challenging problem. This paper presents a novel approach, that is, Quad-PRE, to solve the problem. Quad-PRE starts to consider the features about motifs generated by some tools. From analysis results, we know the number of these features is too little to play important roles in improving the performance of our method, so that we will attempt to find motif features more important in the future work. In addition, Quad-PRE is a multistate method classifying monomer, trimer, tetramer, hexamer, and octamer very well, while other previous methods to predict the quaternary structure attributes are mostly for 2 states.

In fact, the hybrid method Quad-PRE is high accuracy and specificity on discrimination for trimer, tetramer, hexamer, and octamer, respectively. But we compare the Garian's

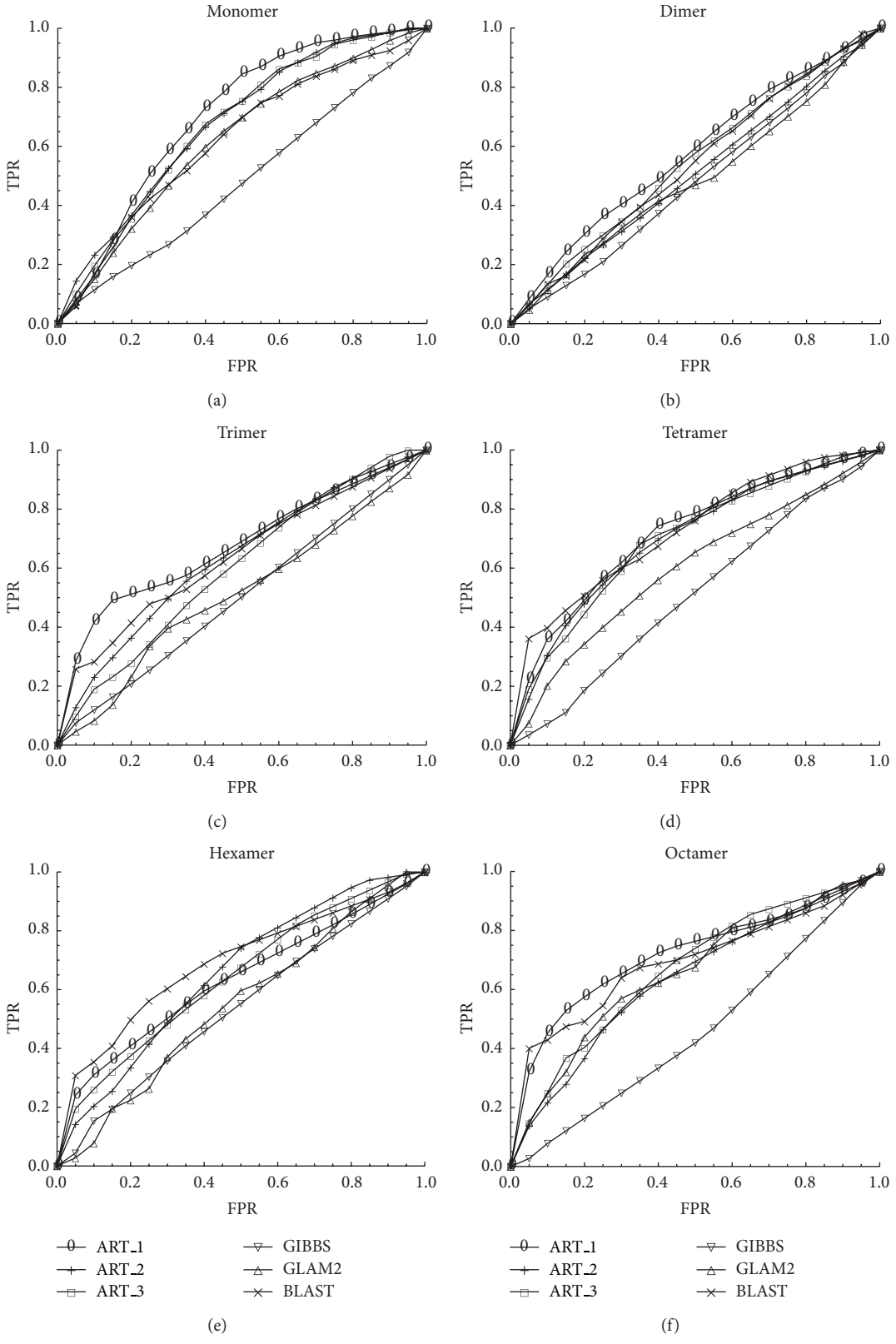


Figure 4: Comparison with the ROC curves of different classes for different feature groups.

Table 3: Predicted results with $C = 0.1$ and gamma = 0.01.

	Monomer	Dimer	Trimer	Tetramer	Hexamer	Octamer	Average
ACC	63.0%	63.8%	**92.8%**	87.0%	**93.3%**	**90.6%**	**81.7%**
TPR	77.9%	30.8%	24.5%	18.1%	23.9%	39.3%	35.7%
SPC	54.9%	79.8%	**96.5%**	**99.0%**	**98.0%**	**93.8%**	**87.0%**
PPV	48.4%	42.3%	27.1%	75.7%	45.8%	28.2%	44.6%
MCC	0.316	0.116	0.220	0.328	0.299	0.284	0.260
AUC	**0.703**	0.582	**0.702**	**0.765**	**0.711**	**0.758**	**0.704**

Table 4: Comparison with Garian's method.

	ACC	TPR	SPC	PPV	MCC	AUC
Quad-PRE	**63.8%**	30.8%	**79.8%**	**42.3%**	**0.116**	**0.582**
QSE	46.2%	73.8%	32.6%	34.7%	0.065	0.522

Table 5: Comparison with results are generated by different feature groups.

	ART_1	ART_2	ART_3	BLAST	GLAM2	GIBBS	Total
ave-ACC	**42.4%**	38.5%	39.9%	34.9%	23.7%	30.6%	**43.5%**
ave-TPR	23.9%	21.8%	23.2%	**33.0%**	22.7%	15.3%	**35.7%**
ave-SPC	**85.5%**	84.7%	85.1%	85.3%	84.5%	82.6%	**87.0%**
ave-PPV	**50.5%**	27.4%	28.9%	35.2%	20.5%	10.2%	**44.6%**
ave-MCC	0.153	0.090	0.111	**0.189**	0.051	−0.024	**0.260**
ave-AUC	**0.680**	0.662	0.661	0.660	0.573	0.510	**0.704**

QSE with our Quad-PRE using our dataset for confirming our method is effective. The results show that our hybrid method performs better than Garian's QSE in predicting the homodimmer or not from metrics ACC, SPC, PPV, MCC, and AUC. In addition, we analyze the importance of the six feature sets. The result clearly shows that each of six features sets contributes to the improvement in prediction, especially the ART_1 feature set. And three new feature sets gained by BLAST, GLAM2, and GIBBS are all effective, because these motif features describe the inherent properties of the sequence inherent and the motifs in protein sequences can help us to understand the structure and function of the molecules the sequences represent [23].

In this paper, we did not consider feature selection because we want to make full use of each feature as many as possible and analyze the importance of each one of six features sets. We believe that future improvements will be possible by designing better sequence representations rather than applying more complex classifiers.

Since user-friendly and publicly accessible web-servers [24] represent the future direction for developing practically more useful predictors, we shall make efforts in our future work to provide a web-server for the method presented in this paper.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 11201334) and Science and Technology Commission of Tianjin Municipality (Grant no. 12JCYBJC31900) to Ke Chen They gratefully acknowledge the help from Dr. Jianzhao Gao for valuable suggestions and comments.

References

[1] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White Jr., "The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 47, pp. 1309–1314, 1961.

[2] R. Garian, "Prediction of quaternary structure from primary structure," *Bioinformatics*, vol. 17, no. 6, pp. 551–556, 2001.

[3] S.-W. Zhang, Q. Pan, H.-C. Zhang, Y.-L. Zhang, and H.-Y. Wang, "Classification of protein quaternary structure with support vector machine," *Bioinformatics*, vol. 19, no. 18, pp. 2390–2396, 2003.

[4] K.-C. Chou and Y.-D. Cai, "Predicting protein quaternary structure by pseudo amino acid composition," *Proteins: Structure, Function and Genetics*, vol. 53, no. 2, pp. 282–289, 2003.

[5] J. Y. Shi, Q. Pan, S. W. Zhang et al., "Classification of protein homo-oligomers using amino acid composition distribution," *Acta Biophysica Sinica*, vol. 22, no. 1, pp. 49–55, 2006.

[6] O. Carugo, "A structural proteomics filter: prediction of the quaternary structural type of hetero-oligomeric proteins on the

basis of their sequences," *Journal of Applied Crystallography*, vol. 40, no. 6, pp. 986–989, 2007.

[7] E. D. Levy, "PiQSi: protein quaternary structure investigation," *Structure*, vol. 15, no. 11, pp. 1364–1367, 2007.

[8] X. Xiao and W.-Z. Lin, "Application of protein grey incidence degree measure to predict protein quaternary structural types," *Amino Acids*, vol. 37, no. 4, pp. 741–749, 2009.

[9] X. Xiao, P. Wang, and K.-C. Chou, "Quat-2L: a web-server for predicting protein quaternary structural attributes," *Molecular Diversity*, vol. 15, no. 1, pp. 149–155, 2011.

[10] X.-Y. Sun, S.-P. Shi, J.-D. Qiu, S.-B. Suo, S.-Y. Huang, and R.-P. Liang, "Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform," *Molecular BioSystems*, vol. 8, pp. 3178–3184, 2012.

[11] L. Kurgan and K. Chen, "Prediction of protein structural class for the twilight zone sequences," *Biochemical and Biophysical Research Communications*, vol. 357, no. 2, pp. 453–460, 2007.

[12] M. J. Mizianty and L. Kurgan, "Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences," *BMC Bioinformatics*, vol. 10, article 414, 2009.

[13] H. Zhang, T. Zhang, J. Gao, J. Ruan, S. Shen, and L. Kurgan, "Determination of protein folding kinetic types using sequence and predicted secondary structure and solvent accessibility," *Amino acids*, vol. 42, no. 1, pp. 271–283, 2012.

[14] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[15] K. Bryson, L. J. McGuffin, R. L. Marsden, J. J. Ward, J. S. Sodhi, and D. T. Jones, "Protein structure prediction servers at University College London," *Nucleic Acids Research*, vol. 33, no. 2, pp. W36–W38, 2005.

[16] J.-T. Huang, J.-P. Cheng, and H. Chen, "Secondary structure length as a determinant of folding rate of proteins with two- and three-state kinetics," *Proteins: Structure, Function and Genetics*, vol. 67, no. 1, pp. 12–17, 2007.

[17] S. Ahmad, M. M. Gromiha, and A. Sarai, "Real value prediction of solvent accessibility from amino acid sequence," *Proteins: Structure, Function and Genetics*, vol. 50, no. 4, pp. 629–635, 2003.

[18] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[19] W. Chen, P. M. Feng, H. Lin, and K.-C. Chou, "iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Research*, vol. 41, no. 6, article e68, 2013.

[20] W. R. Qiu, X. Xiao, and K. C. Chou, "iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components," *International Journal of Molecular Sciences*, vol. 15, no. 2, pp. 1746–1766, 2014.

[21] Y. Xu, J. Ding, L. Y. Wu, and K. C. Chou, "iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition," *PLoS ONE*, vol. 8, no. 2, article e55844, 2013.

[22] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.

[23] T. L. Bailey, "Discovering sequence motifs," *Methods in Molecular Biology*, vol. 395, pp. 271–292, 2007.

[24] X. Xiao, W. Z. Lin, and K. C. Chou, "Recent advances in predicting protein classification and their applications to drug development," *Current Topics in Medicinal Chemistry*, vol. 13, no. 10, pp. 1622–1635, 2013.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 245610, 8 pages
http://dx.doi.org/10.1155/2014/245610

Research Article

Dynamics of Posttranslational Modifications of p53

Qing-Duan Fan,[1,2] Guang Wu,[3,4] and Zeng-Rong Liu[1]

[1] *Institute of Systems Biology, Shanghai University, 99 Shangda Road, Shanghai 200444, China*
[2] *College of Fundamental Studies, Shanghai University of Engineering Science, 333 Longteng Road, Shanghai 201620, China*
[3] *Guangxi Academy of Sciences, 98 Daling Road, Nanning, Guangxi 530007, China*
[4] *DreamSciTech, Apartment 207, Zhencaili 26, Zhujiang Road, Hexi District, Tianjin 300222, China*

Correspondence should be addressed to Qing-Duan Fan; fanqingduan@163.com

Received 19 February 2014; Accepted 6 April 2014; Published 12 May 2014

Academic Editor: Lei Chen

The latest experimental evidence indicates that acetylation of p53 at K164 (lysine 164) and K120 may induce directly cell apoptosis under severe DNA damage. However, previous cell apoptosis models only studied the effects of active and/or inactive p53, that is, phosphorylation/dephosphorylation of p53. In the present paper, based partly on Geva-Zatorsky et al. (2006) and Batchelor et al. (2008), we propose a new cell apoptosis network, in which p53 has three statuses, that is, unphosphorylated p53, phosphorylated p53, and acetylated p53. The time delay differential equations (DDEs) are formulated based on our network to investigate the dynamical insights of p53-induced cell apoptosis. In agreement with experiments (Loewer et al. (2010)), our simulations indicate that acetylated p53 accumulates gradually and then induces the proapoptotic protein Bax under enough DNA damage. Moreover, phosphorylated p53 oscillates and initiates cell repair during DNA damage.

1. Introduction

The tumor suppressor p53, a multifunctional transcription factor, plays an essential role in regulating cellular processes including cell cycle arrest and apoptosis [1]. The importance of p53 also lies in its mutation in over 50% of human cancers. There are three main routes, namely, DNA double-strand break (DSB), ultraviolet light (UV), and oncogenes, which can result in an increase in p53 expression. Experiments demonstrate that cell outcomes depend on the extent of DNA damage, which probably decides the number of p53 pulses or p53 oscillation [2, 3]. Experimentally, p53 regulates cell cycle protein p21, PUMA, BCL-2, PTEN, Bax, Bak, and so on [4], and a cell survives when DNA damage is reparable or commits suicide when irreparable. That is to say, a cell has these two means of avoiding cancer under DNA damage [5]. It follows that the mechanism of repair or apoptosis is closely linked with p53.

There are large quantities of experimental and theoretical researches on p53 networks, in which hundreds of genes and their corresponding proteins are involved. Consequently, the networks include many positive and negative feedback loops acting upon p53. The most prominent of them is the feedback loop between p53 and Mdm2 (mouse double minute 2), which has been considered in many dynamical models. p53 and Mdm2 show nondecaying oscillations in an individual cell, while demonstrating decaying oscillations in cell population, which may be ascribed to aggregate behavior of cells [1]. The ATM- (Ataxia Telan-giesctasia Mutated-) p53-wip1 (wild-type p53-induced phosphatase 1) feedback loop plays an important role in the generation of p53 pulses [6]. These intricate positive and negative feedback loops display various dynamical behaviors [7]. The reliable and flexible mechanism can avoid the premature apoptosis resulting from fluctuations in p53 levels. It is indicated that p53 is modified in a progressive manner and that p53 is divided into p53-arrest and p53-apoptosis in the integrative model [8]. Moreover, high constant levels of active p53 may trigger apoptosis quickly once the decision favoring death is made in seriously damaged cells.

There are several "protein-protein" and "protein-mRNA" dynamical models which describe in detail intracellular signalling of the protein p53. Many theoretical results are obtained in determinate systems. Based on two compartments, nucleus and cytoplasm, an ODEs model is formulated to exhibit that the accumulation of p53 after triggering of ATM under DNA damage. The model also shows robustness of the protein oscillatory dynamics in response to different cellular environments [9]. A set of ordinary differential equations in single cell level shows p53 oscillations in each compartment, nucleus or cytoplasm, and between the two compartments [10]. Based on a sequence of precisely timed drug additions, the authors formulate a computational model, which shows that the dynamic of p53 changes from a pulse to a sustained response [11]. There are also quite a few stochastic systems, where the stochasticity of regulation on p53 shows high heterogeneity and stochastic character of single-cell response [12–15]. Up to now, only phosphorylation modification has been considered in dynamical models.

In fact, a cell is regulated accurately by many posttranslational modifications of p53, which can initiate a program of cell repair or apoptosis at different levels of DNA damage. Methylation of p53 facilitates its subsequent acetylation and protects p53 from ubiquitination [16]. Phosphorylation of p53 is important for inducing p21, a prime inhibitor of cell cycle. Recently, experiments find that a number of external and internal insults induce acetylation and accumulation of p53, via MYBBP1A, RPL5, and RPL11, without phosphorylation [17–19]. It is shown that acetylation of p53 at K164 and K120 may promote cell apoptosis rather than cell arrest [20, 21]. It is observed that p53 may fundamentally switch from pulsing under slight damage to monotonic increase under severe damage [22]. However, there is no corresponding theoretical result about the dynamics of acetylation of p53 up to now.

In the paper, we distinguish functionally the effect of acetylation from phosphorylation of p53 and develop the DDEs of p53 transcriptional regulatory networks based on new experiments [18, 22] and related researches [1, 2, 6, 8, 12]. We pay special attention to the effect of acetylation of p53 and the proapoptotic protein Bax in the case of DNA damage. In agreement with experiments, our simulations indicate that acetylated p53 accumulates gradually with serious DNA damage and induces Bax when p53 surpasses a level.

2. Methods and Models

Methylation allows p53 to be inactive, in normal circumstances p53 is samilarly inactive, so we may regard the initial status of p53 to be inactive without consideration of methylation. Dynamical models of two statuses of p53, that is, inactive p53 and active p53/phosphorylated p53, have been studied extensively [2, 3, 6, 8]. It is indicated that active forms of p53, such as phosphorylated p53 and acetylated p53, have different dynamics and functions experimentally [18]. Interestingly, p53 can be acetylated and accumulates without phosphorylation [17]. Acetylation of p53 on K120 is crucial to p53 dependent apoptosis but is dispensable for p53-mediated growth arrest [19]. In another experiment,

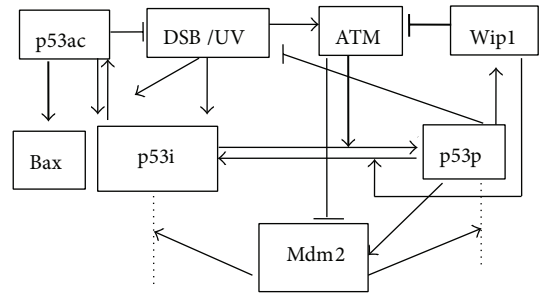


Figure 1: Schematic depiction of the model for the p53 networks in response to DNA damage. Transcription regulation is denoted in thick line. Degradation is denoted in dotted line. Sharp and blunt arrow denote activation and suppression, respectively. ATM's catalyzing phosphorylation of Mdm2, allowing Mdm2 easier to degrade, is regarded as repressive.

acetylation-defective of p53 at k120 can selectively block the transcription of proapoptotic target genes such as Bax and PUMA but has no obvious effect on cell cycle inhibitor p21 [16]. Accordingly, it is necessary that we explore the relation between the means of modifications and cell outcomes. It is shown that p53 may transcribe a few proapoptotic genes such as PUMA, Noxa, Bax, and Bid under excessive DNA damage. Moreover, Bax is very much related to cell apoptotic [23]. For simplicity, we choose Bax as a proapoptotic marker. Because both cell repair and cell apoptosis can lessen DSB; to put it in another way, p53ac and p53p can be regarded as inhibiting DSB. Based on preexisting researches [2, 3, 6], we obtain the following modular schematic depiction in Figure 1.

Considering the tetramer of p53 as a transcription factor, p53-induced Bax is characterized by a Hill function. When phosphorylation and dephosphorylation of p53 are considered separately, a near-optimal switch is possible via Hill equation, where the Hill coefficient equals the number of phosphorylation sites [24]. It is shown that p53 can be phosphorylated at sites S15, T18, S20, S37, S378, and S392 [25]; but Hill coefficient usually ranges between 2 and 4 [24]. Here we denote the ATM-induced phosphorylation of p53 and wip1-induced dephosphorylation of p53 via Hill equation. For simplicity, Hill coefficients take 4. Specially, wip1-induced dephosphorylated of ATM is denoted by Hill equation, where Hill coefficient takes 2 for ATM protein is a dimer. Active p53 has a weaker interaction with Mdm2 than inactive p53 and hence a lower degradation rate [1].

Phosphorylated p53 has a lower degradation rate than inactive p53 because the binding of it with Mdm2 is weaker than inactive p53 [1]. Additionally, the acetylated residues cannot be ubiquitylated by Mdm2 [26]. In the model, two delays are considered because the transcription of Mdm2 by p53 needs time [2] and wip1 expression with a delay would allow p53-induced cell repair [1]. Sustained damage leads to acetylation and accumulation of p53 [17]; however, no acetylation was detected in response to a transient and low-level damage (Figure 6(d) in [18]). In other words, p53 is acetylated only when DNA damage surpasses a certain level,

Table 1: The meaning and values of parameters in the systems (1)–(8).

Parameter	Biological meaning	Value
b_p	Inactive p53 production rate	3
b_{sp}	p53 saturating phosphorylate rate by ATM/ATR	10*/1
b_m	p53-dependent Mdm2 production rate	0.9
b_{mi}	p53-independent Mdm2 production rate	0.2*
b_i	Inhibitor Wip1 production rate	0.25*
b_s	Signal ATM production rate	10*
b_{pac}	Acetylation speed of p53 under DSB/UV damage	0.1/0.02
b_{Bax}	Saturating production rate of Bax	0.04
a_{mpi}	Mdm2-dependent p53 inactive degradation rate	5*
a_{pi}	Inactive p53 degradation rate	2*
a_{mpa}	Mdm2-dependent active p53 degradation rate	0.35
a_{sm}	Signal-dependent Mdm2 inactivation rate	0.5
a_{wpa}	Wip1-dependent dephosphorylation rate of p53	2.8
a_m	Mdm2 degradation rate	1*
a_i	Inhibitor Wip1 degradation rate	0.7*
a_{is}	Wip1-dependent Signal degradation rate	50*
a_s	ATM degradation rate	7.5*
a_{pac}	acetylation of p53 degradation rate	0.05
a_{depac}	deacetylation rate of p53	0.05
a_{Bax}	Bax degradation rate	0.04
τ_1	Time delay of Mdm2 transcription by p53	0.7*
τ_2	Time delay of Wip1 transcription by p53	1.25*
ta	Concentration of Bax for half maximal p53ac production	0.15
ts	Concentration of ATM for half-maximal p53 production	1*
ti	Concentration of wip1 for half-maximal Signal degradation	0.2*
θ_0	The threshold of DSBs for acetylating p53	0.1
k_{rep}	Repair rate of DSBs	0.1
n_1	Hill coefficients for phosphorylation and dephosphorylate of p53	4
n_2	Hill coefficients for dephosphorylation of ATM by wip	2
$[p53i]_0$	The initial condition of inactive p53	1
$[p53p]_0$	The initial condition of phosphorylated p53	0
$[Mdm2]_0$	The initial condition of Mdm2	0.2
$[wip1]_0$	The initial condition of wip1	0
$[ATM]_0$	The initial condition of ATM	0
$[p53ac]_0$	The initial condition of acetylated p53	0
$[Bax]_0$	The initial condition of Bax	0
$[DSBs]_0$	The initial conditions of two type of DSBs	3/0.3**

*denotes parameters from [2, 6], the others are estimated. **denotes serious/slight DSBs.

θ_0. It is shown that the transcription rate of p53 is independent of DNA damage; moreover, there exists an increase in translation rate of p53 following gamma irradiation, so signal strength, ATM, can be denoted by $\theta(x)$. Here θ(damage) equals 1 if damage exists, zero otherwise [2]. According to Figure 1 and the dynamical model in [2, 3, 6], we further formulate a set of DDEs:

$$\begin{aligned}\frac{d\mathrm{p53i}}{dt} &= b_p - a_{\mathrm{mpi}}\mathrm{Mdm2}\cdot\mathrm{p53i} - b_{\mathrm{sp}}\mathrm{ATM}\frac{\mathrm{p53i}^{n1}}{ts^n + \mathrm{p53i}^{n1}} \\ &\quad - a_{\mathrm{pi}}\mathrm{p53i} + a_{\mathrm{wpa}}\mathrm{wip1}\frac{\mathrm{p53p}^{n1}}{ti^{n1} + \mathrm{p53p}^{n1}} \\ &\quad - b_{\mathrm{pac}}\mathrm{p53i}\ \max\left(\mathrm{DSB} - \theta_0, 0\right) + a_{\mathrm{depac}}\mathrm{p53ac},\end{aligned} \tag{1}$$

$$\begin{aligned}\frac{d\mathrm{p53p}}{dt} &= b_{\mathrm{sp}}\mathrm{ATM}\frac{\mathrm{p53i}^{n1}}{ts^{n1} + \mathrm{P53i}^{n1}} - a_{\mathrm{mpa}}\mathrm{Mdm2}\cdot\mathrm{p53p} \\ &\quad - a_{\mathrm{wpa}}\mathrm{wip1}\frac{\mathrm{p53p}^{n1}}{ti^{n1} + \mathrm{p53p}^{n1}},\end{aligned} \tag{2}$$

$$\begin{aligned}\frac{d\mathrm{Mdm2}}{dt} &= b_{\mathrm{mi}} + b_m\mathrm{p53p}\left(t - \tau_1\right) - a_{\mathrm{sm}}\mathrm{ATM}\cdot\mathrm{Mdm2} \\ &\quad - a_m\mathrm{Mdm2},\end{aligned} \tag{3}$$

$$\frac{d\mathrm{wip1}}{dt} = b_i\mathrm{p53p}\left(t - \tau_2\right) - a_i\mathrm{wip1} \tag{4}$$

$$\frac{d\mathrm{ATM}}{dt} = b_s\theta\left(\mathrm{DSB}\right) - a_{\mathrm{is}}\mathrm{wip1}\frac{\mathrm{ATM}^{n2}}{ti^{n2} + \mathrm{ATM}^{n2}} - a_s\mathrm{ATM}, \tag{5}$$

$$\begin{aligned}\frac{d\mathrm{p53ac}}{dt} &= b_{\mathrm{pac}}\mathrm{p53i}\ \max\left(\mathrm{DSB} - \theta_0, 0\right) \\ &\quad - a_{\mathrm{pac}}\mathrm{p53ac} - a_{\mathrm{depac}}\mathrm{p53ac},\end{aligned} \tag{6}$$

$$\frac{d\mathrm{Bax}}{dt} = b_{\mathrm{Bax}}\frac{\mathrm{p53ac}^{n1}}{ta^{n1} + \mathrm{p53ac}^{n1}} - a_{\mathrm{Bax}}\mathrm{Bax}, \tag{7}$$

$$\frac{d\mathrm{DSB}}{dt} = -k_{\mathrm{rep}}\mathrm{DSB}\left(\mathrm{p53p} + \mathrm{p53ac}\right), \tag{8}$$

where p53i, p53p, and p53ac represent inactive p53, phosphorylated p53, and acetylated p53, respectively. In (1), the first term represents p53i synthesis; the second one, catalytic degradation of Mdm2; and the third one, phosphorylation; the fourth to the last one describe self-degradation, dephosphorylation, and acetylation, respectively. In (2), the first term denotes phosphorylation of p53 by ATM; the second one, ubiquitination by Mdm2; and the third one, wip-dependent dephosphorylation. In (3), the first term represents Mdm2 synthesis speed; the second one, activating Mdm2 via p53p, where the delay, τ_1, denotes the time for the transport of p53 from cytoplasm to nucleus and the transcription of Mdm2; the third one, catalyzation of ATM; and the last one, self-degradation. In (4), the first term represents activating wip1 via p53, where the delay, τ_2, is introduced owing to the time for the transport of p53 and the transcription of wip1 by p53 and the second one, self-degradation. In (5), the first term represents exciting ATM induced by DSB; the second one, dephosphorylation by wip1; and the third

one, self-degradation. In (6), the first term represents p53 acetylation caused by DSB, wherein $\max(x)$ refers to the maximum function and the second one, self-degradation of p53ac. In (7), the first term represents Bax induced by p53 and the second one, self-degradation of Bax. (8) denotes the reduction rate of DSB owning to cell repair and apoptosis. For simplicity, all the parameters in the system consisting of (1)–(8) are listed in Table 1.

3. Results

Various exogenous or endogenous stimuli can generate damaged DNA. DSBs and UV are the main types of stimuli, which can activate p53 and subsequently command cell outcomes. DSBs are discussed in two cases, that is, pulsing and repairable DSBs, so we simulate the model according to three cases of DNA damage.

Firstly, we consider the system equations ((1)–(7)) with DSBs at pulsing level. When $10 < t < 40$, DSBs take 3, otherwise DSBs take 0. θ_0, the threshold of DSB for acetylating p53, takes arbitrarily 0.1. The initial conditions of p53i and Mdm2 take 1 and 0.2, respectively; the others take 0. The numerical simulations with Matlab 7.10 (Mathworks) are shown in Figure 2(a), which indicates that p53p has several oscillations with a constant period of about 6 hs and that p53ac accumulates. When p53ac surpasses a level, 0.2 or so, it activates proapoptosis protein Bax. From a biological standpoint, it is reasonable that Bax is activated when p53ac reaches a sufficient level. It is better that the threshold can be testified by experiments. In our simulations, Bax is up to 0.6, which is a rather high level, and we think that cell apoptosis should occur. Experimentally, DSBs are basal in proliferating cell, but cell apoptosis does not occur. p53ac hardly expresses when DSBs are smaller than 0.3, and Bax is also the case (simulations are shown in Figure 2(b)).

Secondly, DSBs decrease under cell pair or cell apoptosis. We consider the system equations (1)–(8) with the parameters and initial conditions set in Table 1 and get numerical simulation shown in Figures 2(c) and 2(d). It is shown that p53p oscillates owing to decreasing DSBs, with a constant period similar to that in the system equations (1)–(7), while p53ac accumulates under serious DNA damage. p53ac ascends and then descends with the decrease of DSBs. When p53ac surpasses a certain level, 0.2 or so, it induces Bax. These simulation results, such as a sufficient level of p53ac activating Bax, are consistent with the biological fact that activation of some proteins need enough inducer. With parameters set in Table 1, simulations show that in the system p53p probably has 5 pulses when $[\mathrm{DSBs}]_0$ is 3 (Figure 2(c)). When $[\mathrm{DSBs}]_0$ drops to 0.3, p53ac and Bax are hardly expressed, and p53p has 2 pulses (Figure 2(d)). In the cases that $[\mathrm{DSBs}]_0$ is 1 or 20, cell has only 4 pulses, but cells have different outcomes, that is, cell repair and cell apoptosis (simulations not shown). The smaller $[\mathrm{DSBs}]_0$ is, the smaller p53ac and the number of pulse of p53p are. On the contrary, the greater $[\mathrm{DSBs}]_0$ is, the faster p53ac accumulates but the smaller the number of pulse of p53p is, which appear to show that cell apoptosis is faster.

Now we consider the robustness of the system parameters. Simulations of the system shown in Figures 2(c) and 2(d) with new parameters perturbation and with cited parameters are shown in Figure 3. When b_{pac}, b_{Bax}, and a_{depac} increase by 10%, and a_{pac} and a_{Bax} decrease by 10% in parameters of Figure 2(c), simulations of the p53 networks are shown in Figure 3(a). When b_{pac}, b_{Bax}, and a_{depac} decrease by 10% and a_{pac}, a_{Bax}, and k_{rep} increase by 10% in parameters of Figure 2(d), simulations are shown in Figure 3(b). Our new parameters increase or decrease by 10%, qualitative characteristics change little except for the position of the equilibrium point (Figure not shown). It is evident that steady points of p53i and Mdm2 do not change much comparing to that of using the initial parameters values. However, p53p and Mdm2 still oscillate while p53ac and Bax accumulate.

At last, when DNA is damaged by ultraviolet rays (UV), ATR, instead of ATM, phosphorylates p53 and Mdm2, except that dephosphorylation of wip1 on ATR is dispensable, the other pathways do not change. Consequently, we have the following differential equations in response to UV:

$$\begin{aligned}\frac{d\mathrm{p53i}}{dt} &= b_p - a_{\mathrm{mpi}}\mathrm{Mdm2}\,\mathrm{p53i} - b_{\mathrm{sp}}\mathrm{ATR}\frac{\mathrm{p53i}^{n1}}{ts^{n1} + \mathrm{p53i}^{n1}} \\ &\quad - a_{\mathrm{pi}}\mathrm{p53i} + a_{\mathrm{wpa}}\mathrm{wip1}\frac{\mathrm{p53p}^{n1}}{ti^{n1} + \mathrm{p53p}^{n1}} \\ &\quad - b_{\mathrm{pac}}\mathrm{p53i}\;\theta\left(\mathrm{UV} - \theta_0\right),\end{aligned} \tag{9}$$

$$\begin{aligned}\frac{d\mathrm{p53p}}{dt} &= b_{\mathrm{sp}}\mathrm{ATR}\frac{\mathrm{p53i}^{n1}}{ts^{n1} + \mathrm{p53i}^{n1}} - a_{\mathrm{mpa}}\mathrm{Mdm2} \\ &\quad \cdot \mathrm{p53p} - a_{\mathrm{wpa}}\mathrm{wip1}\frac{\mathrm{p53p}^{n1}}{ti^{n1} + \mathrm{p53p}^{n1}},\end{aligned} \tag{10}$$

$$\begin{aligned}\frac{d\mathrm{Mdm2}}{dt} &= b_m\mathrm{p53p}\left(t - \tau_1\right) + b_{\mathrm{mi}} \\ &\quad - a_{\mathrm{sm}}\mathrm{ATR}\,\mathrm{Mdm2} - a_m\mathrm{Mdm2},\end{aligned} \tag{11}$$

$$\frac{d\mathrm{ATR}}{dt} = b_s\mathrm{UV} - a_s\mathrm{ATR}, \tag{12}$$

$$\begin{aligned}\frac{d\mathrm{p53ac}}{dt} &= b_{\mathrm{pac}}\mathrm{p53i}\;\theta\left(\mathrm{UV} - \theta_0\right) \\ &\quad - a_{\mathrm{pac}}\mathrm{p53ac} - a_{\mathrm{depac}}\mathrm{p53ac},\end{aligned} \tag{13}$$

where (9)–(13) are based on [6].

We regard damage capability of UV as one-tenth of DSB. Paramters b_{pac}, a_{pac}, b_{Bax}, a_{Bax}, θ_0 and the initial conditions are set and shown in Table 1. Take $\mathrm{UV} = 8$ for $0 \le t < 15$, and take 0 for $t \ge 15$ as [6], the simulations of (4), (7), and (9)–(13) are shown in Figure 4. It is indicated that during slight DNA damage, p53p exhibits a small pulse so as to induce cell cycle arrest/cell repair, and then returns to the initial condition. It is worth noting that p53ac and cell apoptosis are little when $\mathrm{UV} \le 10$, which is consistent with experimental results [6]. Total p53 has a pulse as phosphorylated p53 does and then returns to the basal level.

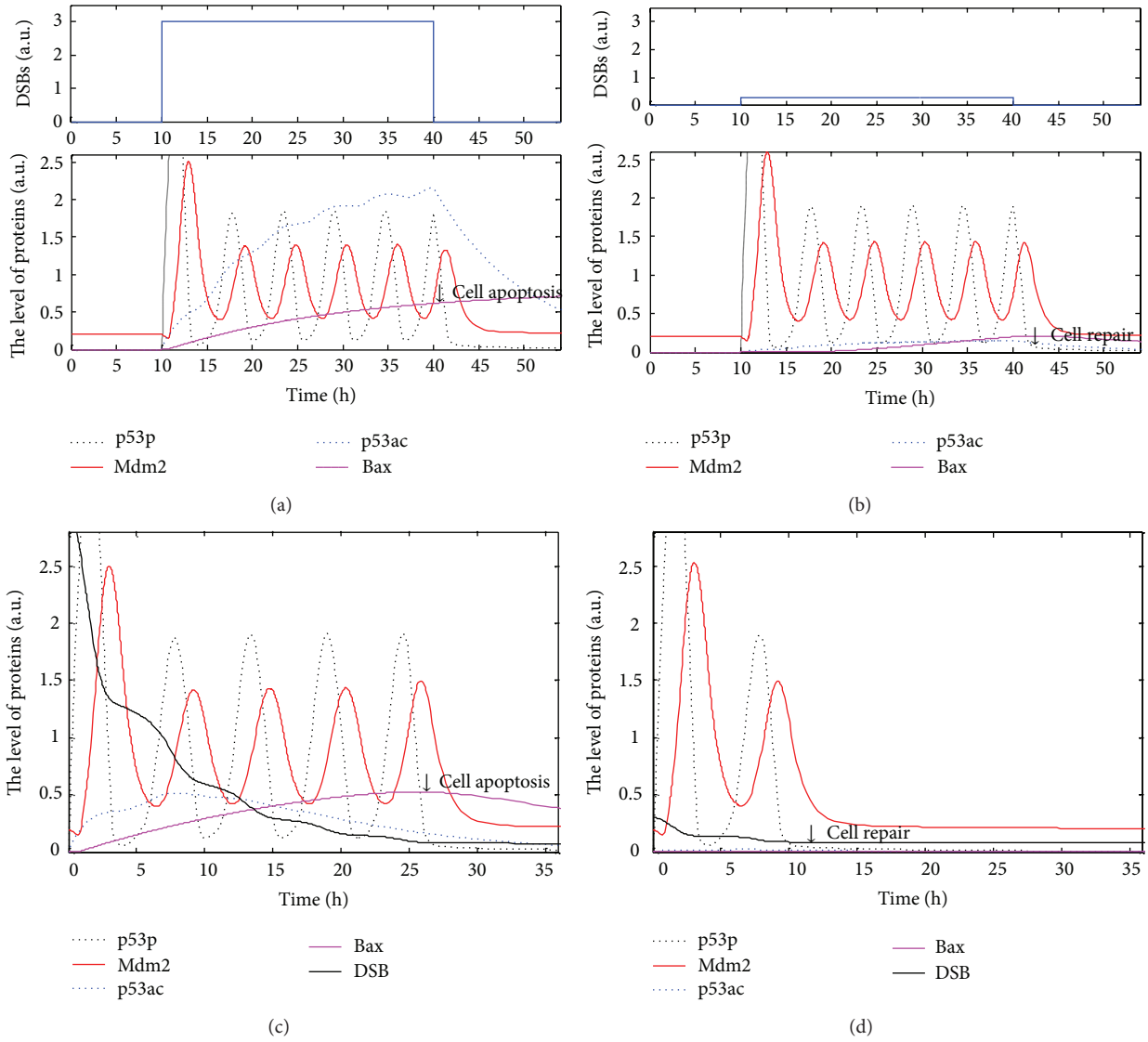


FIGURE 2: Simulations of the p53 regulatory networks under different types of DSBs. Black dotted lines indicates concentration of p53p; red solid lines Mdm2; blue dotted lines p53ac; carmine solid lines Bax; and black solid line DSB. (a, b) Pulsating DSB. (a) Pulse of DSBs takes 3 (a) and 0.3 (b). (c, d) DNA repair via p53 is considered. The initial DSBs take 3 (c) and 0.3 (d). (a, c) p53p oscillates while p53ac accumulates and induces cell apoptosis. (b, d) p53p oscillates while p53ac is hardly expressed during cell repair.

4. Conclusion

The tumor suppressor p53, a most frequently mutated protein in cancer cells, is a key regulator in cell cycle. The latest experiments [17, 18, 24] show that posttranslational modifications of p53, such as phosphorylation and acetylation, are closely linked with cell repair and cell apoptosis. In order to interpret the experimental phenomenon, we develop the regulatory networks and the DDEs model and discuss the dynamics of modifications of p53. Experimentally, acetylation of p53 at K120 and K164 plays an important role in regulating proapoptotic protein [16–19]. It is indicated exactly that p90 is critical to p53-mediated cell apoptosis through promoting acetylation of p53; moreover, p90 has no obvious effects on p53-mediated cell cycle arrest but it is specifically needed for p53-mediated apoptosis [27]. The phenomena that p53ac accumulates and activates proapoptotic protein Bax only under serious DNA damage (Figures 2(a) and 2(c)) and that the pulses of p53p and a little p53ac allow cell to reenter cell cycle under slight DNA damage (Figures 2(b) and 2(d)) are consistent with the latest experiments [2, 6, 18]. The robustness analysis of the model (Figure 3) shows that the accumulation of p53ac and the number of pulse of p53p depend on the extent of DNA damage. The number of pulses of p53p, which means cell repair, should lie on cell repair or cell apoptosis under different levels of DNA damage [28]. UV can lead to single strand break, a kind of slight DNA damage, which usually activates p53p and allows cell to reenter cell cycle (Figure 4). Accordingly, posttranslational modifications of p53 can help us know when and why a cell

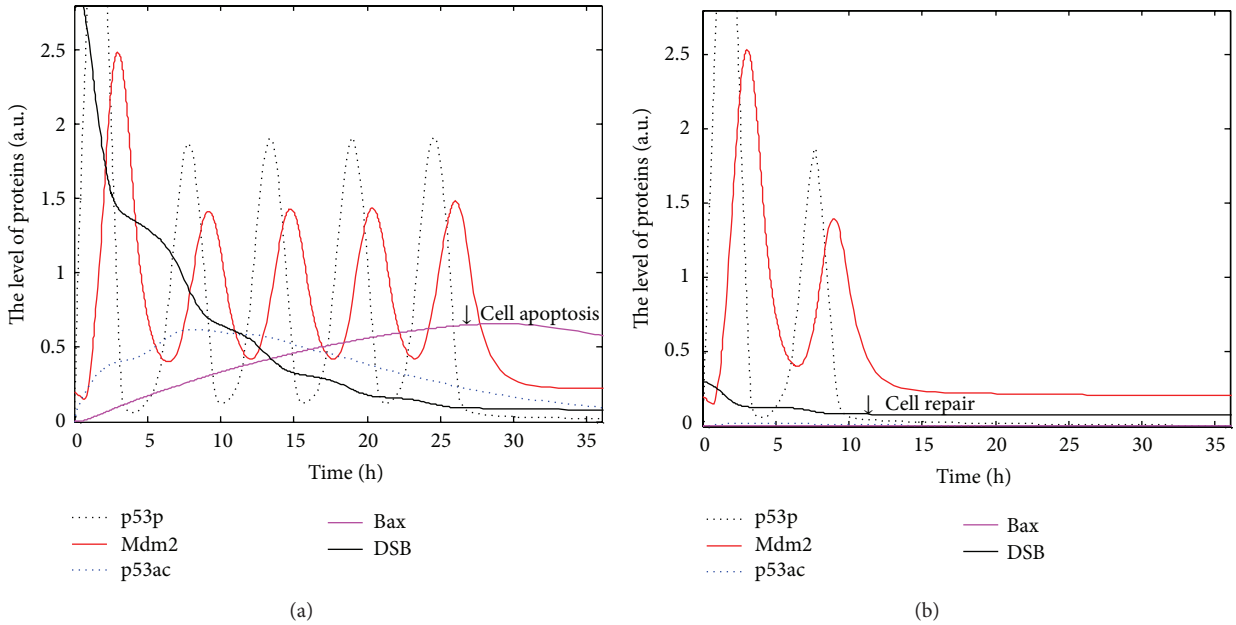


Figure 3: Parameters robustness of the p53 regulatory networks. Black dotted line indicates concentration of p53p; red solid line Mdm2; blue dotted line p53ac; carmine solid line Bax; and black solid line DSB. (a) Production rates decrease 10% and degradation rates increase 10% in Figure 2(c). (b) Production rates increase 10% and degradation rates decrease 10% in Figure 2(d).

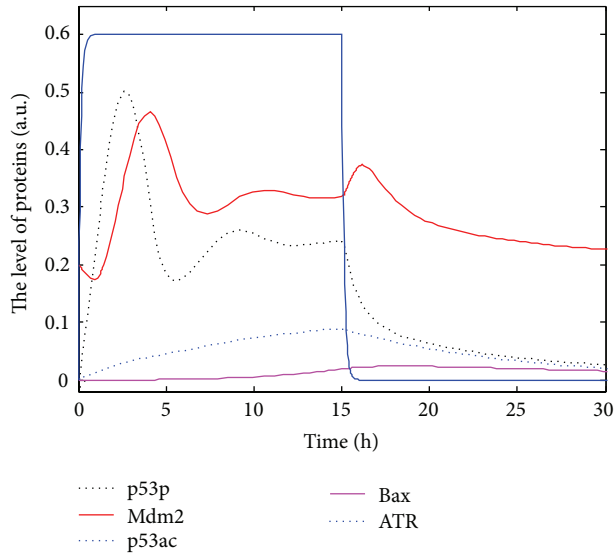


Figure 4: The dynamics of all kinds of p53 networks in response to UV = 8. Black dotted line indicates concentration of p53p; red solid line Mdm2; blue dotted line p53ac; carmine solid line Bax; and blue solid line ATR.

selects programmed cell death or cell repair. In other words, posttranslational modifications of p53 will be beneficial to the treatment of tumors as an innovative therapeutic strategy.

Since cell kinetics is complicated and precisely accurate, it is very difficult to consider all details of posttranslational modifications of p53 [25, 27]. Reference [29] shows phosphorylation may allow p53 stabilization, enhancement of DNA-binding, and activation of its cell-cycle arrest pathway, while acetylation may allow p53 to activate its apoptotic pathway. Yet there are exceptions. For example, phosphorylation at S46 is critical to the induction of proapoptotic genes p53AIP1 (p53-regulated Apoptosis-Inducing Protein 1), but it is not required for the induction of cell cycle inhibitor p21. Phosphorylation and acetylation of p53 probably have synergistic effects on cell cycle, such as acetylation of p53 at k320 may also induce p21 and repress apoptosis [30]. Methylation may allow p53 to be a transcriptionally inactive state; however, methylation of p53 facilitates its subsequent acetylation and protects p53 from ubiquitination [16]. Neddylation and sumoylation have not been demonstrated to affect p53 stability yet.

Additionally, p53 may regulate the transcription expression of many miRNAs, such as MiR-34 and MiR-200. On the other hand, the expression and modification of p53 are also regulated by quantities of MiRNAs [31, 32]. Accordingly, the mechanism of posttranslational modification needs exploration for a long time both in experiment and in theory, such as the mechanism of randomness of protein expression. Our further work will focus on more accurate posttranslational modifications of p53 in cell cycle.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors acknowledge support Grants from National Natural Science Foundation of China (Grants no. 10832006, 11172158, and 11101265).

References

[1] N. Geva-Zatorsky, N. Rosenfeld, S. Itzkovitz et al., "Oscillations and variability in the p53 system," *Molecular Systems Biology*, vol. 2, no. 1, 2006.

[2] E. Batchelor, C. S. Mock, I. Bhan, A. Loewer, and G. Lahav, "Recurrent initiation: a mechanismfor triggering p53 pulses in response to DNA damage," *Molecular Cell*, vol. 30, no. 3, pp. 277–289, 2008.

[3] L. Ma, J. Wagner, J. J. Rice, W. Hu, A. J. Levine, and G. A. Stolovitzky, "A plausible model for the digital response of p53 to DNA damage," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 40, pp. 14266–14271, 2005.

[4] C. Charvet, M. Wissler, P. Brauns-Schubert et al., "Phosphorylation of Tip60 by GSK-3 Determines the Induction of PUMA and Apoptosis by p53," *Molecular Cell*, vol. 42, no. 5, pp. 584–596, 2011.

[5] H. C. Reinhardt and B. Schumacher, "The p53 network: cellular and systemic DNA damage responses in aging and cancer," *Trends in Genetics*, vol. 28, no. 3, pp. 128–136, 2012.

[6] E. Batchelor, A. Loewer, C. Mock, and G. Lahav, "Stimulus-dependent dynamics of p53 in single cells," *Molecular Systems Biology*, vol. 7, article 488, 2011.

[7] S. L. Harris and A. J. Levine, "The p53 pathway: positive and negative feedback loops," *Oncogene*, vol. 24, no. 17, pp. 2899–2908, 2005.

[8] X.-P. Zhang, F. Liu, and W. Wang, "Two-phase dynamics of p53 in the DNA damage response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 22, pp. 8990–8995, 2011.

[9] L. Dimitrio, J. Clairambault, and R. Natalini, "A spatial physiological model for p53 intracellular dynamics," *Journal of Theoretical Biology*, vol. 316, pp. 9–24, 2013.

[10] J. Eliaš, L. Dimitrio, J. Clairam-bault, and R. Natalini, "The p53 protein and its molecular network: modelling a missing link between DNA damage and cell fate," *Biochimica et Biophysica Acta*, vol. 1844, no. 1, Part B, pp. 232–247, 2014.

[11] J. E. Purvis, K. W. Karhohs, C. Mock et al., "p53 dynamics control cell fate," *Science*, vol. 336, no. 6087, pp. 1440–1444, 2012.

[12] T. Sun, W. Yang, J. Liu, and P. Shen, "Modeling the basal dynamics of P53 system," *PLoS ONE*, vol. 6, no. 11, Article ID e27882, 2011.

[13] B. Liu, S. Yan, and Q. Wang, "Intrinsic noise and Hill dynamics in the p53 system," *Journal of Theoretical Biology*, vol. 269, no. 1, pp. 104–108, 2011.

[14] K. Puszyński, B. Hat, and T. Lipniacki, "Oscillations and bistability in the stochastic model of p53 regulation," *Journal of Theoretical Biology*, vol. 254, no. 2, pp. 452–465, 2008.

[15] Y.-C. Hung and C.-K. Hu, "Constructive role of noise in p53 regulatory network," *Computer Physics Communications*, vol. 182, no. 1, pp. 249–250, 2011.

[16] G. S. Ivanov, T. Ivanova, J. Kurash et al., "Methylation-acetylation interplay activates p53 in response to DNA damage," *Molecular and Cellular Biology*, vol. 27, no. 19, pp. 6756–6769, 2007.

[17] T. Kuroda, A. Murayama, N. Katagiri et al., "RNA content in the nucleolus alters p53 acetylation via MYBBP1A," *EMBO Journal*, vol. 30, no. 6, pp. 1054–1066, 2011.

[18] A. Loewer, E. Batchelor, G. Gaglia, and G. Lahav, "Basal dynamics of p53 reveal transcriptionally attenuated pulses in cycling cells," *Cell*, vol. 142, no. 1, pp. 89–100, 2010.

[19] Y. Tang, W. Zhao, Y. Chen, Y. Zhao, and W. Gu, "Acetylation is indispensable for p53 activation," *Cell*, vol. 133, no. 4, pp. 612–626, 2008.

[20] S. M. Sykes, T. J. Stanek, A. Frank, M. E. Murphy, and S. B. McMahon, "Acetylation of the DNA binding domain regulates transcription-independent apoptosis by p53," *Journal of Biological Chemistry*, vol. 284, no. 30, pp. 20197–20205, 2009.

[21] H. M. Collins, M. K. Abdelghany, M. Messmer et al., "Differential effects of garcinol and curcumin on histone and p53 modifications in tumor cells," *BMC Cancer*, vol. 13, no. 1, article 37, 2013.

[22] X. Chen, J. Chen, S. Gan et al., "DNA damage strength modulates a bimodal switch of p53 dynamics for cell-fate control," *BMC Biology*, vol. 11, article 73, 2013.

[23] J. K. Sax, P. Fei, M. E. Murphy, E. Bernhard, S. J. Korsmeyer, and W. S. El-Deiry, "BID regulation by p53 contributes to chemosensitivity," *Nature Cell Biology*, vol. 4, no. 11, pp. 842–849, 2002.

[24] X. Liu, L. Bardwell, and Q. Nie, "A combination of multisite phosphorylation and substrate sequestration produces switch-like responses," *Biophysical Journal*, vol. 98, no. 8, pp. 1396–1407, 2010.

[25] J.-P. Kruse and W. Gu, "Modes of p53 regulation," *Cell*, vol. 137, no. 4, pp. 609–622, 2009.

[26] D. W. Meek and C. W. Anderson, "Posttranslational modification of p53: cooperative integrators of function," *Cold Spring Harbor Perspectives in Biology*, vol. 1, no. 6, Article ID a000950, 2009.

[27] C. Dai, Y. Tang, S. Y. Jung, J. Qin, S. A. Aaronson, and W. Gu, "Differential effects on p53-mediated cell cycle arrest vs. apoptosis by p90," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 47, pp. 18937–18942, 2011.

[28] X.-P. Zhang, F. Liu, Z. Cheng, and W. Wang, "Cell fate decision mediated by p53 pulses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 30, pp. 12245–12250, 2009.

[29] Z. Jiang, R. Kamath, S. Jin, M. Balasubramani, T. K. Pandita, and B. Rajasekaran, "Tip60-mediated acetylation activates transcription independent apoptotic activity of Abl," *Molecular Cancer*, vol. 10, article 88, 2011.

[30] C. D. Knights, J. Catania, S. Di Giovanni et al., "Distinct p53 acetylation cassettes differentially influence gene-expression patterns and cell fate," *Journal of Cell Biology*, vol. 173, no. 4, pp. 533–544, 2006.

[31] Z. Feng, C. Zhang, R. Wu, and W. Hu, "Tumor suppressor p53 meets microRNAs," *Journal of Molecular Cell Biology*, vol. 3, no. 1, pp. 44–50, 2011.

[32] H. Hermeking, "MicroRNAs in the p53 network: micromanagement of tumour suppression," *Nature Reviews Cancer*, vol. 12, no. 9, pp. 613–626, 2012.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 258929, 11 pages
http://dx.doi.org/10.1155/2014/258929

Research Article

Ranking Biomedical Annotations with Annotator's Semantic Relevancy

Aihua Wu

Department of Computer Science, Shanghai Maritime University, Shanghai 201306, China

Correspondence should be addressed to Aihua Wu; ahwu@shmtu.edu.cn

Received 24 February 2014; Accepted 9 April 2014; Published 11 May 2014

Academic Editor: Tao Huang

Biomedical annotation is a common and affective artifact for researchers to discuss, show opinion, and share discoveries. It becomes increasing popular in many online research communities, and implies much useful information. Ranking biomedical annotations is a critical problem for data user to efficiently get information. As the annotator's knowledge about the annotated entity normally determines quality of the annotations, we evaluate the knowledge, that is, semantic relationship between them, in two ways. The first is extracting relational information from credible websites by mining association rules between an annotator and a biomedical entity. The second way is frequent pattern mining from historical annotations, which reveals common features of biomedical entities that an annotator can annotate with high quality. We propose a weighted and concept-extended RDF model to represent an annotator, a biomedical entity, and their background attributes and merge information from the two ways as the context of an annotator. Based on that, we present a method to rank the annotations by evaluating their correctness according to user's vote and the semantic relevancy between the annotator and the annotated entity. The experimental results show that the approach is applicable and efficient even when data set is large.

1. Introduction

Annotations are allowed in most online biomedical databases like NCBI (http://www.ncbi.nlm.nih.gov/), UCSC Gene Browser (http://genome.ucsc.edu/), GDB (http://www.gdb.org/), DDBJ (http://www.ddbj.nig.ac.jp/), and so forth. Shared annotations are becoming increasingly popular in online communities. It is a fundamental activity and plays an important role in the normal research community, with which researchers can explain and discuss the experimental data and share their discoveries [1–5]. As shared comments on documents, pictures, videos, and other annotations, it is also an important data source for biomedical researcher, because of its implying additional facts and annotator's opinions about the biomedical entity. As an example, researchers discovered information about a new protein family with annotations in Flybase [6] and UniProtKB/Swiss-Prot [7]. Now, more and more researchers recognize that it is important to attach and analyse annotations on biomedical entities.

As an open community, there may be many annotations attached with a single biomedical entity. Thus, a question of how to rank the annotations so that users can spend the least time to get the most useful information arises.

Ranking annotations is important and useful for an online biomedical community. As known, biomedical research is active and knowledge about the biomedical entity can be renewed every day. Many of the new discoveries appear in form of annotations. To follow the latest thinking and discovery, researchers will spend much time to view these annotations. A ranking module can help them to retrieve high quality annotations quickly and improve efficiency of the discussions. Rankings also encourage users to publish correct and validated opinion and materials about the biomedical data, so that the community will be more active and become a more important data center and discussion platform.

Ranking reviews, which can be viewed as a type of annotation, are a common problem in many e-commerce and news websites [8, 9]. Popular previous methods are mostly

based on voting or scoring. Unfortunately, voting and scoring cannot avoid spreading wrong opinions, because users would like to agree with the most popular reviews, even if they do not know whether it is right or not. As a result, useless, even spiteful, reviews constantly appear in the top position in many websites.

As we know, quality of a scientific annotation depends in part on how much the annotator learns about the biomedical entity. The more knowledge the annotator has, the more correct his annotations can be, thus, the more useful to the data user. For example, as for the H1N9 virus, annotations from an astrophysicists are normally with lower correctness than those submitted by a biologist who concentrates on bird flu. User's knowledge is indicated by his semantic background such as working experience, study, and research. If given user is viewed as an object, the semantic background will be a composite of all attributes that describe the user or his related objects and so the biomedical entity can be described. We say that a biomedical entity and a user are semantic related if their semantic backgrounds are partly matched. Obviously, the more they matched, the more the user may learn about the entity.

In the scientific community, an obvious fact is that the annotator's knowledge can be reflected in papers he published and approaches he focused on, which can be obtained from the Internet or other public data source. With such background data, how the annotator may learn about the entity he annotated can be deduced. Besides, accepted historical annotations do also reflect the annotator's knowledge about the annotated entity. If an annotator always contributes high quality annotation to entities with the same attributes, we can say he is familiar with other such entities. In this paper, we propose a weighted and concept-extended resource description framework (RDF) [10] to represent an annotator and a biomedical entity. For any given pair of annotator and biomedical entity, a RDF graph will be created, where the annotator is the root node, attributes of the entity and its one-step extended concepts are the leaf nodes, and each edge is assigned a weight denoting how much the root node learns about the target node. The weight will be evaluated by their cooccurrence in credible web data. On the other hand, frequent patterns of the biomedical entities that was historically annotated by given annotator will be mined. Suppose there is no malicious user, people only annotate biomedical entity that they know. Both the weight and the matching degree of the annotated entity to the frequent patterns are explained as the semantic relevancy. Accordingly, we present a method to rank the annotations by evaluating their correctness with the semantic relevancy between the annotator and the biomedical entity.

Organization. Section 2 is related works. Section 3 introduces the weighted RDF graph model and related concepts. Section 4 presents two main works of this paper. One is how to initialize RDF graph of an annotator and a biomedical entity by web information extraction, including details of computing weight for an annotator's RDF by association mining open credible web information. The other is the algorithm for mining frequent item of historical annotated biomedical entities. Section 5 shows formulas evaluating correctness of a new annotation. Section 6 states experimental results. And last section is the conclusion.

2. Related Work

Evaluating and ranking biomedical annotations are new problems. The most similar researches are ranking reviews, estimating quality of web content, and opinion strength analysis.

Ranking reviews or other web content has always been a complex problem and attracts renewed research interests in many fields, especially as web plays an increasing important role in delivering and achieving information for many people. Most previous methods are based on user's reputation, word-of-mouth, webpage links, and the other types of user's voting [8, 9, 11–14]. Ai and Meng proposed a method based on weighted fan-in page links and copies to recommend recruitment advertising [11]. It has a viewpoint that the more the users believe and the more dependable the websites are, the higher the quality of the advertisement will be. Largillier et al. present a voting system for news articles using statistical filter and a collusion detection mechanism in [8]. It is reasonable to rank web content according to author's reputation and user's voting in some applications. The former is unworkable when the user does not have enough historical annotations, while the latter cannot exclude propagation of rumors. In this paper, we try to evaluate annotation's quality from the new perspective of the semantic relation between annotators and the annotated biomedical entities, which is, to the best of our knowledge, scarcely considered by previous approaches. In biomedical domain, correctness of user's annotations largely depends on annotator's knowledge about the annotated entities. Semantic relevancy between them plays a critical role in the quality evaluation. Our method is more convincing.

Some prior works try also to discover inherent relationship between data and its users by data mining techniques [15–22]. They can be classified into three categories: statistical methods based on cooccurrence of terms [16], machine learning techniques [17], and hybrid approaches of them [18]. Staddon and Chow studied online book reviews of http://www.amazon.com/ and proposed a method of quality evaluation by mining the association rules between book authors and book reviewers [15]. In [22] the authors proposed three models to evaluate quality of Wikipedia articles by measuring the influence of author's authority, review behavior, and the edit history on quality of the article. These researches also try to discover semantic relationship between data and its users, but they did not consider textual content of reviews or other online opinions [18–20], and their criteria are simple; for example, association relationships are defined as the cooccurrence of the author's name and the annotator's name on web in [15]; as a result, they cannot reveal comprehensive semantic relevancy. We describe the entities by their entire semantic context with their attributes and related biomedical entities and based on that, we can analyze multidimensional

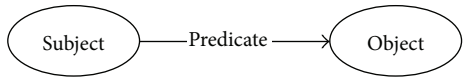


Figure 1: The atom triple of RDF.

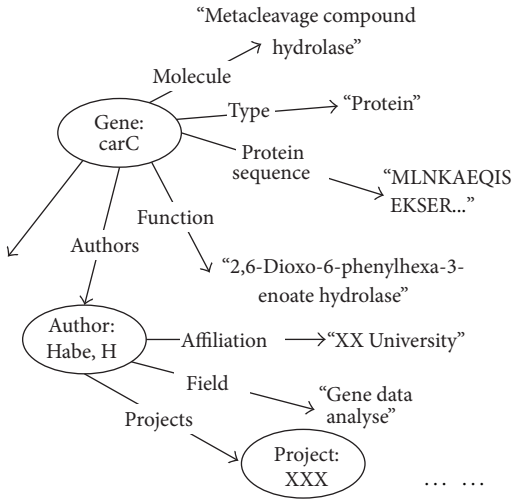


Figure 2: RDF graph of a biomedical entity.

semantic relationships between biomedical entities and their annotator. Still, we parse the textual content of the annotation and highlight attributes mentioned in it when matching patterns and evaluating its correctness.

Other related works are biomedical web information extraction, biomedical text mining, and biomedical entity recognition [23–28]. They are related but independent problems. We did not propose new algorithms for those problems and we did not develop a related tool, but we applied existing methods and applications. You can find some performance trials on the website of the Biocreative group (http://www.biocreative.org/) [23], ontology-driven term extraction service for biomedical text on the National Center for Biomedical Ontology (NCBO), and biomedical text mining applications developed by several academic groups and other organizations [24–29].

3. Weighted RDF Graph and Concepts

RDF is a graph based framework for representing concepts on the web by linking its concrete syntax to its formal semantics. In RDF, any expression is a triple of a subject, a predicate, and an object, which can be illustrated by a node-arc-node linked graph as shown in Figure 1. Node represents a subject or an object, and directed arc with a predicate represents relationship between them.

A biomedical entity can be viewed as a RDF subject; its attributes and concept field can be looked at as its objects. Figure 2 shows the RDF graph of protein structure 1J1I in RCSB, whose main features include molecule, protein sequence, function, and authors. Attribute nodes can be extracted from the online biomedical databases and their linked credible web sites. Here, we say that a node is an attribute node if its outdegree is 0, and the others are *entity nodes*. Tag of an *entity node* is composed of type name and ID of the entity in form of *typeName*:*entityID*. Attributes nodes will be extracted as more as we can so that an entity can be specified more exactly.

An annotator can also be viewed as a RDF subject, and biomedical entities he/she annotated can be its objects. Annotators may have many attributes, but we only consider those locally described and those related to the biomedical entity. We use two types of RDF graph to specify an annotator. One is named annotator's RDF graph whose composing details are present in Section 4.1, and the other is a set of frequent patterns of his/her historical annotated entities. In the RDF graph of annotator P, the annotator is the root node, the biomedical entity and its related concepts are the annotator's objects node, and weight on edge pointing to node A, which is marked as ω_P^A, is initialized as the correlation degree of P and A. Instead of weight, frequency and correctness are attached to each pattern, indicating their semantic relevancy.

Different from others, scientific data has complicated concept background. It can be a node in a complex relation network. There is a high possibility that people learning A will also learn about A's subconcepts, A's father concept, or A's related concepts. For example, an annotator who knows many of *Trichophyton tonsurans* and *Trichophyton schoenleini* may also know about *Trichophyton rubrum*, because they all a type of mycosis causing similar tinea. Intern weight will be calculated for such possibility.

Definition 1 (intent weight). Suppose annotator u learns about concept A_1 with weight of N, B is a father concept or a related concept of A, and there are $M - 1$ other concepts $A_2, A_3, \ldots, A_m$ who are also B's subconcept or related concept, but u does not indirectly know about them; then weight on edge pointing to $A_i (2 \le i \le m)$ in annotator graph of u is N/M. Such weight is called intent weight of A_i against A_1, marked as $\varpi_P^{A_1 \ldots A_i}$.

Total intent weight ϖ_u^A of a concept A in u's RDF graph is defined as follows:

$$\varpi_u^A = \sum_{i=1}^{i<=N} \left(\frac{\omega_u^{A_i}}{M_i} \right). \tag{1}$$

Here, A_i is father or related concept of A, M_i is number of concepts whose relationship with A_i is identical to that of A with A_i, and the relationships are defined in open biomedical databases such as FACTA+ and Go Terms.

Definition 2 (RDF path). (1) If there is an edge e between an entity node E and an attribute node A, we say that $E/e/A$ is a RDF path between E and A. (2) If there is a RDF path p between entity node E' and A and an edge between entity node E and E', we say that E/p is a RDF path between E and A. The first node is *root node* of a RDF path. And *pattern path* is a *RDF path* without entity node value.

In Figure 2, "Gene:carC\type\'protein'" is a RDF path, and "Gene\type\'protein'" is a pattern path.

Definition 3 (prefix path). Given a RDF path or a pattern path p, the subsequence from the root node to edge pointing to a nonroot node E is a prefix path of E in p.

Two RDF paths with identical prefix path are *conjugate*. Conjugate RDF paths can be merged into a sub-RDF graph and conjugate sub-RDF graphs can be merged into a bigger sub-RDF graph when merging the identical ancestor nodes.

Given two RDF paths p and g, if there is a RDF path p' in g, where $p' = p$, we say that $p \subset g$. Similarly, Given two sub-RDF graphs $g1$ and $g2$, if, for all $p \subset g1$ (p is a RDF path), $p \subset g2$, we say that $g1 \subset g2$, and if $g1 \subset g2$ and $g2 \subset g1$, we say that $g1 = g2$.

Likewise, two pattern paths with same prefix path are *conjugate*. Two conjugate pattern paths can be merged into a subpattern RDF graph. And a pattern path can belong to a pattern RDF graph g, if it is equal to a path in the graph. And for any two pattern RDF graphs $g1$ and $g2$, if, for all $p \subset g1$ (p is a pattern path), $p \subset g2$, we say that $g1 \subset g2$, and if $g1 \subset g2$ and $g2 \subset g1$, we say that $g1 = g2$.

Additionally, let us define some symbols used as follows.

(i) $pp_u^{\mathrm{cr},f}|^{N=n}$ is a frequent pattern path of user u from biomedical entity O to N with correctness cr and frequency f and$n$ is value of attribute N. Similarly, $pp_u^{\bar{\omega}}|^{N=n}$ is a path of user u pointing to N with weight $\bar{\omega}$ and n is value of attribute N.

(ii) P_u^{cr} is a frequent pattern of user u on attribute B with correctness of cr, which is composed of frequent pattern paths.

4. Building Annotator's RDF Graph

In the following, Section 4.1 states details of composing annotator's RDF graph and computing weights by association mining open credible web information. And Section 4.2 presents frequent mining algorithm.

4.1. Initializing Annotator's RDF Graph with Web Information. Too much information can be extracted from the huge Internet, but only those of the biomedical entity and the annotation are useful in this application.

Given an annotation $\langle u, o, r\rangle$ where u is the annotator in form of a RDF node or a RDF graph, o is RDF graph of the biomedical entity, and r is the annotation, complete RDF graph of u is comprised of the following:

(i) u,

(ii) o,

(iii) an edge from the root node of u pointing to o.

Here (1) u is initialized as an entity node when no local information can be used or a RDF graph generated according to the annotator's background data from the online database itself; (2) o is initialized as stated in the following.

Generating RDF Graph for a Biomedical Entity. RDF graph of a biomedical entity o is initialized according to what is described in the online database. In our experiments, we created o by the following steps.

(1) Recognize id (e.g., DOI) and type (protein, virus, etc.) of the biomedical entity with predefined keyword or normal structure and compose its entity node with tag of "Type:id."

(2) Extract each head item as an edge from predefined module such as "molecular description" and "experimental detail" and extract the value of the item as its attribute node or compose another level of entity nodes if the module contains several items and draw edges from the entity node to the attribute node.

(3) Extract family classification according to the linked database on the page like Go Terms, look one step more into the detail of the linked database, recognize relationships between entities (e.g., mapping a protein to an organism or finding protein of the same family), draw RDF graph for them, merge the RDF graphs of different linked databases, and eliminate duplicate RDF paths.

Figure 3 shows a segment of the information we will extract from the online database, and the circled items will be extracted as edge and their value will be extracted as attribute nodes. Figure 4 shows an example of one-step extension of the biomedical entity's related concept to FACTA+.

Annotation Analysis. Bioconcepts in the annotations can be extracted by biomedical text analysis tools like GENIA [29] and the others. These concepts are normally the annotation's topic. We extract bioconcepts and their attribute names in an annotation; here the attributes names can be recognized by patterns "XX of bioconcept" or "bioconcept's XX." For each concept, we draw an entity node and an edge for each of its attribute names even without attribute value. Merge and marked out the RDF graphs of the annotation into that of the biomedical entity o. If they cannot be merged, draw an edge from the annotator to its root nodes without weight.

Weight Calculating. We assign the weight on an edge will be assigned as the co-occurance of the annotator and the edge's target node in credible open data sources, such as news/talks/papers/personal pages published by predefined credible organizations, known proceedings, and websites. In the experiment, we use Google to search the news, talks, and personal pages, while Anne OTate [30] and PIE [31] to search papers on PubMED and MEDLine. At present, we did not consider the situation of different concepts inferring with the same biomedical entity, which is another scientific problem known as the biomedical text mining and clustering.

Suppose term of the annotator u is $t1$, term of the node A is $t3$, and term of the edge pointing to A is $t2$; then weight on the edge from web is defined as follows:

$$\omega 1_u^A = \begin{cases} \dfrac{(c(t1 \wedge t2) + c(t1 \wedge t3) - c(t1 \wedge t2 \wedge t3))}{c(t1)} \\ \qquad A \text{ is an attribute node} \\ \sum \omega 1_u^{Bi} \\ \qquad A \text{ is not an attribute node.} \end{cases} \tag{2}$$

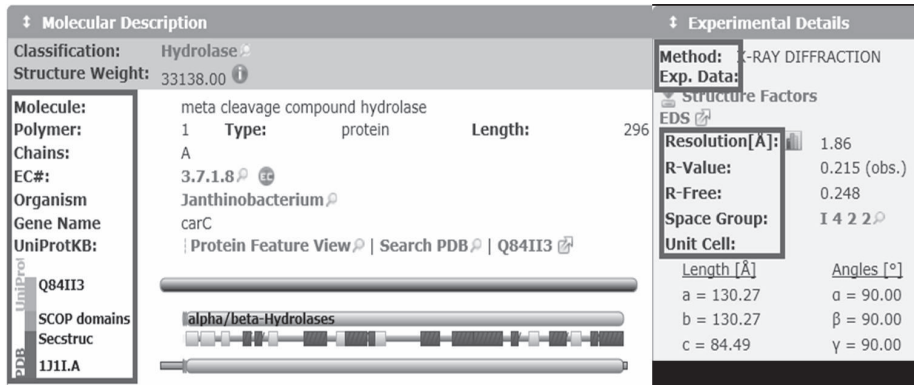


Figure 3: Examples of information extraction for annotated object.

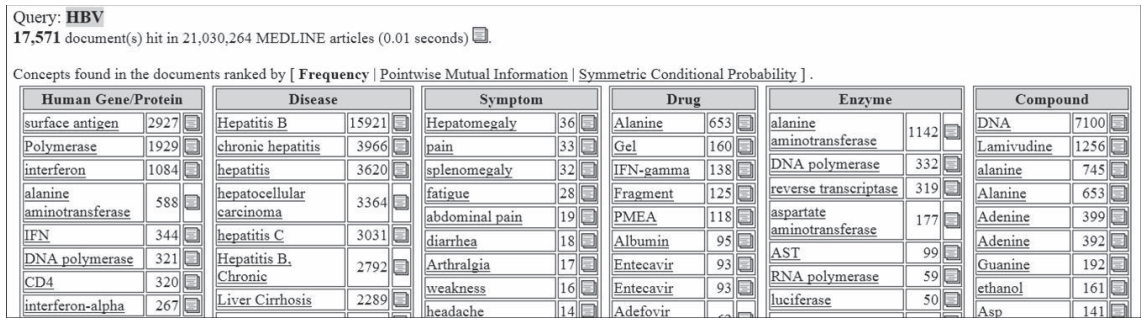


Figure 4: An example of one-step extension of the biomedical entity.

Here $c(t1 \wedge t2)$ is the count of web pages that include $t1$ and $t2$, and Bi is an object node that A points to.

Considering the fact indicated by intent weight $\varpi 1_u^A$, weight on the edge from web is finally defined as follows:

$$\omega 1_u^A = \omega 1_u^A + \varpi 1_u^A. \tag{3}$$

4.2. Mining the Frequent Entity Patterns. Annotator's knowledge about a biomedical entity can also be inferred by his historical annotations. In this section, we will present an algorithm to discover frequent features of the historical annotated entities with correctness larger than 0.6. The algorithm will consider not only direct attributes of the entity, but also that of its one-step extended related concepts.

As illustrated in Figure 5, firstly, the algorithms classify all annotations according to their annotator and then cluster each subset of annotations against their correctness with K-means. And correctness of each annotation in the cluster will be viewed as that of the cluster center. Lastly, frequent patterns are mined over biomedical entities in each cluster. Several questions arise here. First, because of the classification and cluster, the input data set can be too small to produce any patterns. The algorithms use *Laplacian* smoothing to solve it. Second, the algorithms can bring too much frequent patterns, while some of them can be included in or similar to another one. The algorithm uses Rule 1 to merge those that describe the same owner and the same attribute but with different attribute values and Rule 2 to merge the same patterns but with different correctness. Third, the data sets can be improperly clustered so that frequent pattern cannot be found. The algorithms use a new round of cluster and frequent pattern mining until mining results do not change.

Frequent sub-RDF graphs mining is the key step in the whole algorithm (step 2.3 of Algorithm 1). It takes the pattern paths of the entities as the items. Both the initial and final results are initialized as set of the frequent items obtained by the first round scan, and the result set is repeatedly refreshed by replacing each element with its one-item extension if the extension is also frequent. As shown in Figure 6, in the first round extension, each element in result set will conjunct with each element in initial set; for example, conjunctive of t_1 and t_2 is also frequent, so t_1 and t_2 will be replaced by $t_1 t_2$ in the result set.

Rule 1. Suppose that $p1_u^{\mathrm{cr}}, p2_u^{\mathrm{cr}}, \ldots, pn_u^{\mathrm{cr}}$ are a set of frequent patterns of user u with the same correctness rate cr and paths $pp1^{N1,f1} \in p1_u^{\mathrm{cr}}, pp2^{N2,f2} \in p2_u^{\mathrm{cr}}, \ldots, ppn^{Nn,fn} \in pn_u^{\mathrm{cr}}$ with the same or different frequency; if $N1, N2, \ldots, Nn$ are different attribute values of the same attribute node N, then $ppi^{Ni,fi} (1 \le i \le n)$ can be replaced by $pp^{\{N1,N2,\ldots Nn\},f}$. Specially, if $p1_u^{\mathrm{cr}}, p2_u^{\mathrm{cr}}, \ldots, pn_u^{\mathrm{cr}}$ are only different with each other on $ppi^{Ni,fi} (1 \le i \le n)$, then they can be merged into $p_u^{\mathrm{cr}}|_{N \text{ in}\{N1,N2,\ldots Nn\}}$ and can replace $ppi^{Ni} (1 \le i \le n)$ with $pp^{\{N1,N2,\ldots Nn\},f}$; furthermore, if domain $N = \{N1, N2, \ldots, Nn\}$, then they can be merged into $p_u^{\mathrm{cr}}|_{N=\mathrm{any}}$ by replacing $ppi^{Ni,fi} (1 \le i \le n)$ with $pp^{\mathrm{any},f}$. In each target path, frequency $f = \sum_{i=1}^{i=n} fi/n$.

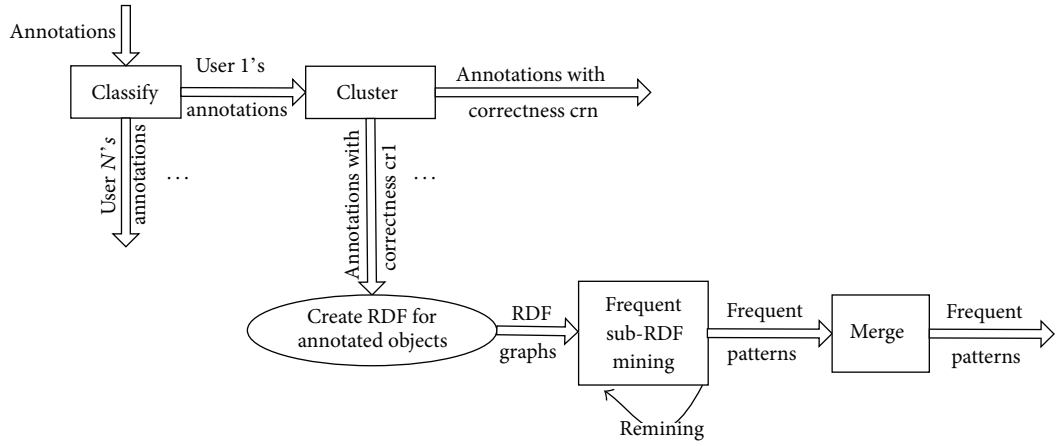


Figure 5: Illustration of the series of algorithms to mining frequent entity patterns.

input: $\mathfrak{R} = \{(o_i, u_i, \text{cr})(i = 1, \ldots, n,\ j = 1, \ldots, m)\}$, (o_i, u_i, cr) means that user u_i annotated biomedical entity o_i with correctness rate cr (cr > 0.6); o_j also refer to its RDF graphs;
ε, predefined threshold of frequency;
λ, group number of correctness rates defined by user
output: Ω, a set of frequent patterns.

(1) classify $\mathfrak{R}$ into different sets of $\mathfrak{R}_1 = \{(o_j, u, \text{cr})\}$, in each set, annotations are all submitted by user u;
(2) for each $\mathfrak{R}_1$
(2.1) $k = \lambda$; $\mathfrak{R}_1' = \mathfrak{R}_1$;
(2.2) cluster elements in $\mathfrak{R}_1'$ into a set of groups $\sum = \{\mathfrak{R}_{11}{}^{\text{cr1}}, \mathfrak{R}_{12}{}^{\text{cr2}}, \ldots, \mathfrak{R}_{1k}{}^{\text{crk}}\}$ according to cr with k-mean, and cluster center is the correctness of the group, for example, cr1 is correctness of $\mathfrak{R}_{11}$;
(2.3) for each $\mathfrak{R}_{1i}(1 \le i \le k)$ //find frequent patterns for given annotator with given correctness
(2.3.1) $\alpha'' = \alpha' = \{\text{Pattern Path } pp^f \mid pp \text{ belong to an entity } o \in \mathfrak{R}_{1i} \text{ and } f = |pp|/(|P| + 1) > \varepsilon\}$, here, $|pp|$ is count of pp in $\mathfrak{R}_{1i}$ and $|p|$ is count of all Pattern Paths in $\mathfrak{R}$) // set of frequent pattern paths
(2.3.2) for $\forall pp_i \in \alpha'(1 \le i \le |\alpha'|)$, pp_i can be a Pattern Path or a sub RDF graph. //find frequent conjugate items
{ for $\forall pp_j \in \alpha''(1 \le j \le |\alpha''|)$
{ If pp_i and pp_j are conjugate and $pp_j \not\subset pp_i \wedge pp_j \not\subset pp_i$ and $f = |pp_i \wedge pp_j|/(|P| + 1) > \varepsilon(|pp_i \wedge pp_j|$ is the conjunct appearance of pp_i and pp_j in Δ'). Then
{ merge pp_i and pp_j into a sub RDF graph g, and f is the frequency of g; $\alpha' = a' \cup g$;} }
If exists one graph $g \in \alpha'$ including pp_i, then remove pp_i from α';}
(2.3.3) Repeat Step (2.3.2) untill α' doesn't change;
(2.3.4) $P_u^{\text{cri}} = \alpha'$; $\omega_{1i} = \{o \mid o \in \mathfrak{R}_{1i} \wedge \neg\exists p(p \in a' \wedge p \text{ matches a RDF path of } o)\}$
(2.4) $p_u = p_u \cup P_u^{\text{cr1}} \cup \cdots \cup P_u^{\text{crk}}$; $\omega = \omega_{11} \cup \omega_{12} \cup \cdots \cup \omega_{1k}$;
(2.5) For any two pattern $g \in P_u^{\text{cri}}, g' \in P_u^{\text{crj}}(i \ne j)$, If $(g = g')$, then //merge same pattern with different cr
{ remove g, g' from p_u;
$p_u = p_u \cup g_u^{\text{crg}}(\text{cr}g = (n1 * \text{cr}g + n2 * \text{cr}j)/(n1 + n2); n1$ is number of entities matching g in $\mathfrak{R}_{1i}$; $n2$ is number of entities matching g' in $\mathfrak{R}_{1j}$) }
(2.6) $\Omega = \Omega \cup p_u$;
(2.7) if ($k > 1$ and $\omega \ne \phi$) { $\mathfrak{R}'_1 = \{(o, u, \text{cr}) \mid o \in \omega\}$; $k = \lambda \setminus 2$; go to (2.2);}
(3) circularly merge frequent patterns in Ω with Rule 1 and Rule 2 presented in this section until Ω doesn't change;
(4) return Ω;

Algorithm 1: Frequent pattern.

Rule 2. Suppose that $p1_u^{\text{cr1},f1}, p2_u^{\text{cr2},f2}, \ldots, pn_u^{\text{crn},fn}$ are a series of frequent patterns of user u but with different correctness and the same or different frequency; if $p1 = p2 = \cdots = pn$, then $p1_u^{\text{cr1},f1}, p2_u^{\text{cr2},f2}, \ldots, pn_u^{\text{crn},fn}$ can be merged into $p_u^{\text{cr},f}$, where $\text{cr} = \sum_{i=1}^{i=n}(\text{cr}i * fi)/\sum_{i=1}^{i=n} fi$ and $f = \sum_{i=1}^{i=n} fi/n$.

5. Ranking Annotation

In this section, we propose an algorithm to evaluate correctness (quality) for an annotation $r(u, o)$ of biomedical entity o from user u under different situations: (1) u isdirect semantically related to o; (2) o is an entity node in RDF graph

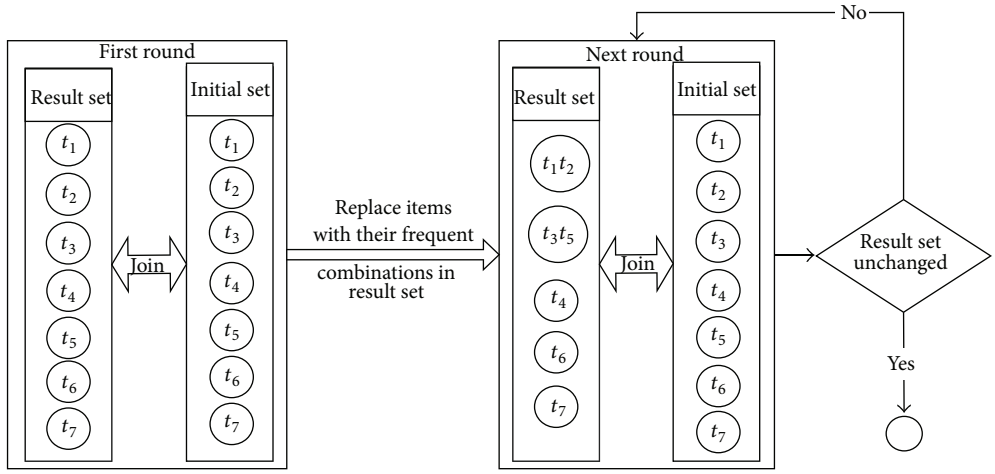


Figure 6: Illustration of frequent sub-RDF graphs mining.

of u or o matches at least one frequent entity pattern of u on o; (3) u has annotated another biomedical entity which is similar to o; (4) o has been annotated by other users who are similar to u; (5) u has never annotated any entity and o has never been annotated. Obviously, annotator is semantic related to the annotated biomedical entity in the first two situations, especially 100% semantic relevant in the first one. We will give formulas to evaluate correctness of annotations for the two situations in Section 5.1., while problem of computing correctness in the last three situations is called a "new user" problem, which will be solved by borrowing the credibility of its nearest neighbor. And details will be stated in Section 5.2. Totally, annotations will be ranked decreasing according to evaluating results of all annotations on the biomedical entity.

Besides the semantic relationship, we also consider user's voting and historical annotations on similar annotated biomedical entities from similar annotators when computing credibility of annotations. User's voting is a direct parameter for the agreement degree. And for new user problem where no semantic relationship exists, similar historical annotations can be borrowed to estimate the annotation's correctness.

5.1. Evaluating When Semantic Related. When annotator u is an attribute node in the RDF graph of the biomedical entity o or o is an attribute node of u, we say that they are semantic related to each other. More strictly, for an annotation $r(u, o)$, suppose $G1$ is RDF graph of annotation r, $G2$ is RDF graph of annotator u, $G3$ is RDF graph of biomedical entity o, and Ω is a set of frequent patterns of u, whose forming methods are all stated in Section 4; if $\exists$ a prefix path $\mathrm{pr1} \in G3$ and a prefix path $\mathrm{pr} \in G1$ that $\mathrm{pr1} = \mathrm{pr}$ and one of $G3$'sentity node is u, we say that u is direct semantically related to o. Normally, if (1) there is a prefix path $\mathrm{pr} \in G1$, where $\mathrm{pr} \in G2$, or (2) there is at least a path in $G3$ matching a frequent pattern in Ω, we say that u is semantically related to o.

Given an annotation $r(u, o)$, if user u is direct semantically related to biomedical entity o and supposing that V is a set of voting score on r, where only the max one of each user's voting will be kept, then correctness acr of r is

$$\mathrm{acr} = 1 + \left(\frac{\sum v\,(v \in V)}{|V|} \right). \tag{4}$$

Here, $|V|$ is the number of the element in set V. Furthermore, suppose $G1, G2, G3$ is RDF graph of r, u, and o corresponding, and Ω is a set of frequent patterns of u, if u is non-directly but semantically related to o, correctness of r is decided by the weight of $G1$ in $G3$ and the max matching degree of o to a frequent pattern in Ω. Supposing that P_u^{cr} isa frequent pattern of u with correctness cr andsupposing that P_u^{cr} has N RDF pattern paths, among which K pattern paths (suppose $pp1_u^{\mathrm{cr}1,f1}, \ldots, ppK_u^{\mathrm{cr}K,fK}$) match both a RDF path of o and a prefix path of $G1$, then the feature matching degree d_o^P of u and P_u^{cr} is defined as follows:

$$d_o^P = \sum_{i=1}^{K} (\mathrm{cr}_i * f_i), \tag{5}$$

cr_i, f_i is correctness of pattern path pp_i.

And supposing that there are M paths of $G1$ belonging to $G2$ with weight $\varpi 1, \ldots, \varpi M$ on each edge pointing to the attribute nodes, then correctness acr of r is defined as follows:

$$\mathrm{acr} = \max\left(d_o^P\right) + \sum_{i=1}^{M} \varpi i + \left(\frac{\sum v\,(v \in V)}{|V|} \right) \tag{6}$$

$$\left(p \in \Omega \text{ and } p \text{ match a prefix path of } G1\right).$$

5.2. Evaluating for "New User". When there is neither annotator's RDF graph nor frequent patterns indicating that the annotator u and the entity o are semantically related, but u has annotated other biomedical entities or o has been annotated by other user, we can use the nearest neighbor to evaluate correctness of annotation $r(u, o)$.

For a given biomedical entity o, its nearest neighbor is a set of biomedical entity in which each element o' satisfies the next condition:

$$\frac{|pp_2|}{|pp_o|} > \varepsilon, \qquad \frac{|pp_2|}{|ppo'|} > \varepsilon. \tag{7}$$

Here, $|pp_2|$ is number of RDF paths that belong to both o and o', $|pp_o|$ is number of paths that belong to o, $|ppo'|$ is number of paths that belong to o', and ε is threshold defined by user.

Similarly, nearest neighbor of a given user u is also a set of users among which each user u' satisfies the following conditions:

$$\frac{\left|\text{appear}\left(u, u'\right)\right|}{\left|\text{appear}\left(u'\right)\right|} > \varepsilon \quad \text{or} \quad \frac{\left|o^{\text{cr}>\theta}\right|}{\left|o'^{\text{cr}>\theta}\right|} > \varepsilon, \qquad \frac{\left|o^{\text{cr}>\theta}\right|}{\left|o''^{\text{cr}>\theta}\right|} > \varepsilon. \tag{8}$$

Here, $|\text{appear}(u')|$ is number of unique appearance of u' in papers, public talks, news, and so forth, especially papers in PubMED and MEDLine, while $|\text{appear}(u, u')|$ is the coappearance of u and u' in the above data sources. $|o^{\text{cr}>\theta}|$ is number of biomedical entities that was annotated by both u and $u'$with correctness larger than user defined threshold θ, $|o'^{\text{cr}>\theta}|$ is number of biomedical entities that was annotated by u with correctness larger than user defined threshold θ, $|o''^{\text{cr}>\theta}|$ is number of biomedical entities that was annotated by u' with correctness larger than user defined threshold θ, and ε is threshold defined by user.

Now, given an annotation $r(u, o)$, if user u is not semantically related to biomedical entity o, supposing that V is a set of unique user's voting score on r, supposing that U is a set of users who are the nearest neighbor of u, and O is a set of biomedical entity who are the nearest neighbor of o, then correctness acr of r is

$$\text{acr} = \begin{cases} \left(\dfrac{\sum \text{acr}_u^{oj}\,(oj \in O)}{|O|}\right) + \dfrac{\sum v\,(v \in V)}{|V|} \\ \qquad O \text{ is not empty} \\ \left(\dfrac{\sum \text{acr}_{ui}^{o}\,(ui \in U)}{|U|} + \dfrac{\sum v\,(v \in V)}{|V|}\right) \\ \qquad O \text{ is empty and } U \text{ is not empty.} \end{cases} \tag{9}$$

Here, $|V|$ is also the number of the elements in set V. acr_u^{oj} is correctness of annotation submitted by user u on biomedical entity o.

Lastly, given an annotation $r(u, o)$, if user u never submits any annotation and biomedical entity o has never been annotated and supposing that V is a set of voting score on r, where only the max one of each user's voting will be kept, then its correctness acr is defined as

$$\text{acr} = \left(\frac{\sum v\,(v \in V)}{|V|}\right). \tag{10}$$

6. Experimental Evaluation

There are three works in this paper: (1) extracting web information to compute relevancy of an annotator and a biomedical entity, (2) frequent pattern mining of the historical annotations, and (3) evaluating correctness of the annotations. We will state in this section how we use the existing tools to extract web information and get our experimental data and show performance of the frequent pattern mining and ranking evaluations.

6.1. Experimental Environment. Settings of the experiment are Intel Celeron 420 2.0 GHZ CPU, 1 GB memory, and windows XP+SP2. The local database is SQL Server 2000.

6.2. Data Preparation. As an example, we only use protein data in the experiments. But our approach can also be applied to other biomedical entities. We firstly get manually 500 protein structures and their scientific names from http://www.rcsb.org, download their files like FASTA sequence and PDB, crawl their web page, extract basic attributes from the files and webpage, and import them into SQL server. Then we search the Anne OTate [30] with scientific names of those protein structures and randomly get 1000 unique authors as our initial annotators. Although there are some annotations and ontology of biomedical entity in the online database, few of them are proper for the frequent pattern mining. Thus, we automatically generate 20000 historical annotations, of which 60 percent are designed as shown in Table 1 and the others are randomly generated: random annotator, random biomedical entity, and random annotation with random correctness.

As shown in Table 1 1000 of the annotators are classified as 9 types. Each type is designed to contribute certain number of annotations with correctness in certain range. To test the *cold-start* problem, several users are designed to contribute 5 or below annotations. On the other hand, to ensure the patterns can be found, at least five of each type of users will give annotations on 5 to 15 biomedical entities with common features.

As for the web information, we presearched and stored their weights in database for the 20000 pairs of users and biomedical entities. First, each biomedical entity will be one-step extended in FACTA+ to get its related concepts. Then, to evaluate the weight, we get information by two ways: searching Google for news, talks, and homepages and searching PIE the search [31] for papers and other documents. To search Google, we write a C# program which autosearches the predefined credible websites with Google service using keywords including name/affiliation of the annotator, scientific name of the biomedical entity, extended concept, or attribute name of the biomedical entity as a plus. On the other hand, we apply and evaluate PIE the search to count the documents that indicate their semantic relationship. The resulting corpus contains a set of medical articles in XML format. From each article we construct a text file by extracting relevant fields such as the title, the summary, and the body (if they are available).

6.3. Frequent Pattern Mining. We test 8 groups of data ($s1 \sim s8$ in Table 2), each of which only including annotations

Table 1: Annotator and annotation predefined in the experiments.

UserType/Num	Details of the designed annotations	Annotation ratio
U1/200	All annotation are 100% correct, and 5 of them only contribute 5 or below annotations	15%
U2/300	40% annotation with correctness 0.95~1; 50% annotations with correctness 0.9~0.95; 10% annotations with correctness 0.85~0.9; 22 of them only contribute 5 or below annotations	30%
U3/200	15% annotation with correctness 0.95~1; 55% annotations with correctness 0.9~0.95; 20% annotations with correctness 0.85~0.9; 10% annotations with correctness 0.8~0.85; 30 of them only contribute 5 or below annotations	15%
U4/80	10% annotations with correctness 0.9~0.95; 60% annotations with correctness 0.85~0.9; 30% annotations with correctness 0.8~0.85	10%
U5/80	30% annotations with correctness 0.85~0.9; 40% annotations with correctness 0.8~0.85; 30% annotation with correctness 0.75~0.8	10%
U6/40	5% annotations with correctness 0.9~0.95; 20% annotations with correctness 0.85~0.9; 30% annotations with correctness 0.8~0.85; 30% annotation with correctness 0.75~0.8; 25% annotations with correctness 0.7~0.75	8%
U7/40	5% annotations with correctness 0.8~0.85; 15% annotations with correctness 0.75~0.8; 50% annotation with correctness 0.7~0.75; 30% annotations with correctness 0.6~0.7	7%
U8/30	10% annotations with correctness 0.75~0.8; 30% annotations with correctness 0.7~0.75; 60% annotation with correctness 0.6~0.7	3%
U9/30	All annotation are below 60% correct; 5 of them only contribute 5 or below annotations	2%

Table 2: Data deployment in pattern mining.

	Entities A group	Fre. Attr.	100% fre. Attr.	Max degree fre. associate Attr.	Frequent threshold
$s1$	12	24	24	24	0.95
$s2$	15	5	0	5	0.7
$s3$	10	20	2	15	0.5
$s4$	20	0	0	0	0.85
$s5$	16	49	0	1	0.7
$s6$	28	22	3	3	0.7
$s7$	36	24	3	3	0.7
$s8$	18	5	3	3	0.7

published by one annotator and belonging to one correctness group. The max group ($s7$) has 700 annotations and about 36 biomedical entities but on different attribute sets, while the min group ($s3$) has 100 annotations and about 10 biomedical entities. Biomedical entities in each group have some common attributes, which can be recognized as frequent pattern paths (fre. Attr. column in the table) after the first round of computing in the algorithm. Some of the frequent pattern paths appear in every biomedical entity, we say that they are 100% fre. Attr. Association of such items is certainly frequent; thus, we put their association directly into the finial mining result set but ignore another round of computing. The experimental results (Figure 7) show that the main time consumer is recursively computing the associate frequent pattern paths. $s3$ takes the highest time, because the 18 frequent (frequency below 100%) items need 15 rounds of computing to judge whether any level of their associations is also frequent. $s4$ is carried out at minimal cost, because no frequent pattern path can be found and only the first round of computing will happen.

Table 3: Data deployment in ranking evaluation.

	Patterns	Annotations	Annotators	Entities
$c1$	49	5000	100	50
$c2$	100	10000	200	50
$c3$	196	20000	200	100
$c4$	285	30000	300	100
$c5$	400	40000	200	200

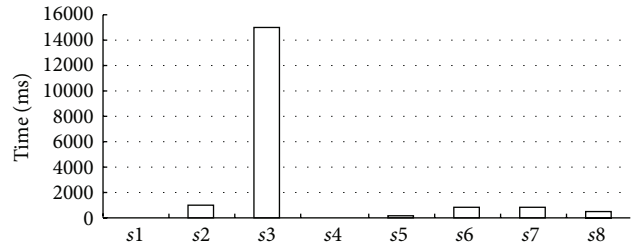


Figure 7: Time performance of frequent pattern mining.

6.4. Ranking. The experiments are executed over 5 sets of data. Different data sets contain different scales of annotations and frequent patterns. As shown in Table 3, $c1$ is the minimal data set, where 5000 annotations submitted by 100 annotators on 50 biomedical entities will be evaluated and ranked with 49 frequent patterns, while $c5$ is the maximal one including 40,000 annotations from 200 annotators on 200 biomedical entities, where it will be evaluated and ranked with 400 frequent patterns. For that weight on edge between each user and biomedical entity are precomputed

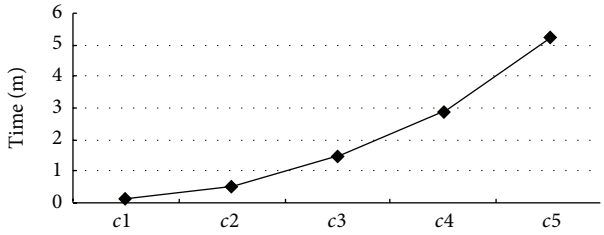


Figure 8: Time performance of correctness evaluation and ranking.

and stored in database, the most time-consuming is the pattern matching. As shown in Figure 8, time goes up as number of patterns or annotations goes up. But even for $c5$, 5 minutes is enough to rank 40,000 annotations, which show the efficiency and applicability of the algorithm.

7. Conclusion

In this paper, we propose an approach for ranking biomedical annotations according to user's voting and semantic relevancy between an annotator and the biomedical entity he annotated. Our idea is inspired by the fact that in a credible online scientific community, quality of web content is determined to some extent by the contributor's knowledge about the entity. People's knowledge can be discovered from his profile and his related historical behaviors, especially for the researchers who are deeply specialized in one scientific domain. Thus, our major work in this paper is to find out how much a given annotator may learn about a biomedical entity from his profile on the web and frequent patterns of entities that he annotated in history.

An entity can be semantically defined by its attributes and its related entities' attributes. And people's knowledge about an entity can be reflected by the annotator's knowledge about those attributes. To express such relation, we extend the RDF model by assigning weight on each edge, which denotes the degree of how the root node (the annotator) knows about the target node (an entity or one of its attributes). The weight can be evaluated with the cooccurrence of the annotator and the target node in credible web information. Besides, an intent weight can indicate that people who know concept A may also know A's related concept.

The second way to discover how the annotator semantically relates to the biomedical entity is frequent pattern mining over historical annotations, which revealed the common features of biomedical entities that an annotator may know. The pattern mining algorithm proposed in this paper can deal with problems caused by small example space, cold-start, and improper data source dividing.

In the future, we will go further on how to link record of a user and extract his profile information from the Internet when duplicate and uncertain data happen.

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This paper is supported by the National Natural Science Foundation of China under Grant no. 61202022.

References

[1] W. Gatterbauer, M. Balazinska, N. Khoussainova, and D. Suciu, "Believe it or not: adding belief annotations to databases," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 1–12, 2009.

[2] D. Bhagwat, L. Chiticariu, W. C. Tan, and G. Vijayvargiya, "An annotation management system for relational databases," *VLDB Journal*, vol. 14, no. 4, pp. 373–396, 2005.

[3] M. Y. Eltabakh, M. Ouzzani, W. G. Aref et al., "Managing biological data using bdbms," in *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering (ICDE '08)*, pp. 1600–1603, IEEE, April 2008.

[4] A. H. Wu, Z. J. Tan, and W. Wang, "Annotation based query answer over inconsistent database," *Journal of Computer Science and Technology*, vol. 25, no. 3, pp. 469–481, 2010.

[5] A. H. Wu, Z. J. Tan, and W. Wang, "Query answer over inconsistent database with credible annotations," *Journal of Software*, vol. 23, no. 5, pp. 1167–1182, 2012.

[6] S. Tweedie, M. Ashburner, K. Falls et al., "FlyBase: enhancing drosophila gene ontology annotations," *Nucleic Acids Research*, vol. 37, supplement 1, pp. D555–D559, 2009.

[7] M. Schneider, L. Lane, E. Boutet et al., "The UniProtKB/Swiss-Prot knowledgebase and its plant proteome annotation program," *Journal of Proteomics*, vol. 72, no. 3, pp. 567–573, 2009.

[8] T. Largillier, G. Peyronnet, and S. Peyronnet, "SpotRank: a robust voting system for social news websites," in *Proceedings of the 4th Workshop on Information Credibility on the Web (WICOW '10)*, pp. 59–66, ACM.

[9] N. Wanas, M. El-Saban, H. Ashour, and W. Ammar, "Automatic scoring of online discussion posts," in *Proceedings of the 2nd ACM Workshop on Information Credibility on the Web (WICOW '08)*, pp. 19–25, ACM, October 2008.

[10] http://www.w3.org/RDF/ .

[11] J. Ai and X. F. Meng, "C-Rank: a credibility evaluation method for deep web records," *Journal of Frontiers of Computer Science and Technology*, vol. 3, no. 6, pp. 585–593, 2009.

[12] J. Brown, A. J. Broderick, and N. Lee, "Word of mouth communication within online communities: conceptualizing the online social network," *Journal of Interactive Marketing*, vol. 21, no. 3, pp. 2–20, 2007.

[13] C. N. Ziegler and G. Lausen, "Propagation models for trust and distrust in social networks," *Information Systems Frontiers*, vol. 7, no. 4-5, pp. 337–358, 2005.

[14] M. Cheung, C. Luo, C. Sia, and H. Chen, "Credibility of electronic word-of-mouth: informational and normative determinants of on-line consumer recommendations," *International Journal of Electronic Commerce*, vol. 13, no. 4, pp. 9–38, 2009.

[15] J. Staddon and R. Chow, "Detecting reviewer bias through web-based association mining," in *Proceedings of the 2nd ACM Workshop on Information Credibility on the Web (WICOW '08)*, pp. 5–10, ACM, October 2008.

[16] A. Ghose, P. G. Ipeirotis, and A. Sundararajan, "Opinion mining using econometrics: a case study on reputation systems," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL '07)*, vol. 1, pp. 416–423, June 2007.

[17] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins, "The predictive power of online chatter," in *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '05)*, pp. 78–87, ACM, August 2005.

[18] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)*, pp. 168–177, ACM, August 2004.

[19] T. Lee and E. T. Bradlow, "Automatic construction of conjoint attributes and levels from online customer reviews," The Wharton School Working Paper, University of Pennsylvania, 2007.

[20] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," in *Proceedings of the 14th International Conference on World Wide Web*, pp. 342–351, ACM, 2005.

[21] D. Anthony, S. W. Smith, and T. Williamson, "The quality of open source production: zealots and good samaritans in the case of Wikipedia," Rationality and Society, 2007.

[22] M. Hu, E. P. Lim, A. Sun, H. W. Lauw, and B. Q. Vuong, "Measuring article quality in wikipedia: models and evaluation," in *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM '07)*, pp. 243–252, ACM, November 2007.

[23] C. N. Arighi, P. M. Roberts, S. Agarwal et al., "BioCreative III interactive task: an overview," *BMC Bioinformatics*, vol. 12, supplement 8, article S4, 2011.

[24] M. Krallinger, M. Vazquez, F. Leitner et al., "The protein-protein interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text," *BMC Bioinformatics*, vol. 12, supplement 8, article S3, 2011.

[25] The Annotation Ontology on Google Code, http://code.google.com/p/annotation-ontology/.

[26] D. Kwon, S. Kim, S. Y. Shin, and W. J. Wilbur, "BioQRator: a web-based interactive biomedical literature curating system," in *Proceedings of the BioCreative 4th Workshop*, vol. 1, pp. 241–246, Washington, DC, USA, 2013.

[27] Q. C. Bui, S. Katrenko, and P. M. A. Sloot, "A hybrid approach to extract protein-protein interactions," *Bioinformatics*, vol. 27, no. 2, Article ID btq620, pp. 259–265, 2011.

[28] A. B. Abacha and P. Zweigenbaum, "Automatic extraction of semantic relations between medical entities: a rule based approach," *Journal of Biomedical Semantics*, vol. 2, supplement 5, article S4, 2011.

[29] http://www.nactem.ac.uk/tsujii/GENIA/tagger/.

[30] http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/AneOTate.cgi

[31] http://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/PIE/index.html.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 454310, 6 pages
http://dx.doi.org/10.1155/2014/454310

Research Article

Genomic and Functional Analysis of the Toxic Effect of Tachyplesin I on the Embryonic Development of Zebrafish

Hongya Zhao,[1] Jianguo Dai,[2] and Gang Jin[2]

[1] *Industrial Center, Shenzhen Polytechnic, Shenzhen 518055, China*
[2] *School of Applied Chemistry and Biotechnology, Shenzhen Polytechnic, Shenzhen 518055, China*

Correspondence should be addressed to Gang Jin; jingang@szpt.edu.cn

Received 27 February 2014; Accepted 1 April 2014; Published 29 April 2014

Academic Editor: Tao Huang

Tachyplesin I (TP I) is an antimicrobial peptide isolated from the hemocytes of the horseshoe crab. With the developments of DNA microarray technology, the genetic analysis of the toxic effect of TP I on embryo was originally considered in our recent study. Based on our microarray data of the embryonic samples of zebrafish treated with the different doses of TP I, we performed a series of statistical data analyses to explore the toxic effect of TP I at the genomic level. In this paper, we first employed the hexaMplot to illustrate the continuous variation of the gene expressions of the embryonic cells treated with the different doses of TP I. The probabilistic model-based Hough transform was used to classify these differentially coexpressed genes of TP I on the zebrafish embryos. As a result, three line rays supported with the corresponding 174 genes were detected in our analysis. Some biological processes of the featured genes, such as antigen processing, nuclear chromatin, and structural constituent of eye lens, were significantly filtered with the smaller P values.

1. Introduction

Tachyplesin I (TP I) is originally isolated from the acid extracts of hemocytes of the horseshoe crab *Tachypleus tridentatus* in 1988 [1]. Since then, a series of biochemical analyses are made to gain sight into it. TP I consists of 17 amino acid residues, two disulfide bonds, and one unique α-arginine at the C terminal end and is characterized by a disulfide-stabilized β-sheet conformation. Among these 17 alpha amino acid residues, 4 cysteine residues constitute two disulfide bonds that contribute to the hemolytic ability of TP I in blood cells [2]. Furthermore, research has shown that TP I is active against both Gram-positive and Gram-negative bacteria [3], fungi [4], viruses [5], and cancer cells [6]. The peptide also interacts with DNA and inhibits the synthesis of macromolecules. Because of its potency and relatively small size, this peptide is a promising candidate of a novel alternative antibiotic in the pharmaceutical industry and animal and food industries. In comparison with the biochemical structure and mechanism of TP I, however, the literature on the toxicity of TP I is very limited, except its hemolytic ability.

In recent years the microarray technology is developed to be a powerful tool for investigation of functional genes that at present has become routine in many research laboratories [7–13]. To our best knowledge, microarray analysis for transcriptome changes of TP I has not been done. Therefore, the microarray experiment is originally designed to assess the toxic effect of TP I on embryo in our lab.

Despite a long tradition of using rats and mice to model human disease, several aspects of rodent biology limit their use in large-scale genetic and therapeutic drug screening programs. Currently, many researchers have sophisticatedly generated zebrafish (*Danio rerio*) models of a wide variety of human diseases. By contrast, zebrafish offers a model system that is ideally suited for large-scale microarray analysis. Zebrafish is a small vertebrate species, easily subjected to chemical mutagens and large numbers of mutant zebrafish. Given these advantages, zebrafish hold tremendous potential for the identification of toxic-causing and toxic-modifying genes [5, 6]. So the embryos of zebrafish were selected in our microarray experiment. The main aim of this study is to identify the functional coexpressed genes that show significant differential expression in the embryonic cells of

zebrafish treated with the different doses of TP I. Obviously, the functional analysis of the toxic effect of TP I at genome level may reveal more advantages and disadvantages as a novel alternative antibiotic in the future.

The modified hexaMplot is employed to illustrate the gene expression alterations in our micoarray data analysis. The original hexaMplot is a two-dimensional representation of three kinds of intensities for assessing the drug effect [13]. Its coordinates represent the log ratios of intensity pairs: $x_1 = \log E_2/E_1$ and $x_2 = \log E_1/E_0$, where E_0, E_1, and E_2 reflect the expression levels of genes in the normal, disease, and drug-treated samples, respectively. Note that genes appearing in the upper and lower half-plane of the hexaMplot are up- and downregulated, respectively, by the disease. Analogously, genes located in the left and right half-plane of the hexaMplot are up- and downregulated, respectively, by the drug treatment, compared with the disease sample. As a result, the slant axis $x_2 = -x_1$ is considered as $x_3 = \log E_2/E_0$. Obviously, along the axis the expression levels of genes in the normal and drug-treated samples are the same.

Naturally the three axes and six regions in hexaMplot are meaningful for assessing drug effect [13]. Some methodologies are proposed to assess the drug effect based on the hexaMplot [13–15]. In comparison with the previous algorithms to consider the differently expressed genes only, the probabilistic model-based HT is proposed to address noise and quantization with the contribution of all assayed genes by posterior probabilities [15]. The performance of the algorithm is proved to be more robust and powerful. So we analyze our microarray data with the model-based HT algorithm.

Note that experiment design to analyze the toxic effect of the different doses of TP I is different from that to assess the drug. The gradual alterations of the gene expressions with the different doses are of our interest. So we first modified the axes of the original hexaMplot with $x_1 = \log E_1/E_0$, $x_2 = \log E_2/E_0$, and $x_3 = \log E_2/E_1$, where E_0, E_1, and E_2 reflect expression levels of the genes in the normal and 1.5 ug/mL and 2.5 ug/mL treated samples, respectively. Analogously, the corresponding axes and quadrants also show the significant meanings to assess the toxic effect on assayed genes treated with the different doses of TP I.

Furthermore, the model-based HT in [15] was used to identify the functional genome groups along the line rays. Three line rays supported with the corresponding 174 genes are detected in our microarray data. The feature genes identified on the same line ray of hexaMplot may show the similar expression patterns in varying the doses of TP I. The functional groups of these genes, such as antigen processing, nuclear chromatin, and structural constituent of eye lens, show coherent biological function with high significance as detected by gene ontology analysis.

The paper has the following organization: after introducing our microarray experiment in Section 2, we apply the modified hexaMplot to demonstrate the gene alteration among the embryonic samples with the different doses and model-based HT to identify the function gene groups in Section 3. The conclusion and discussion are summarized in Section 4.

2. Methodology

In this section, the microarray experiment to assess the toxic effect of TP I on the embryonic development of zebrafish is described in Section 2.1. Then, the statistical tools to analyze the microarray data are mentioned in Section 2.2.

2.1. Microarray Gene Expression Experiment. Considering the great advantages of zebrafish in model system, we select the zebrafish model for exploring the potential developmental toxicity of TP I in our microarray experiment. TP I was synthesized by Shenzhen Han Yu Pharmaceutical Co., Ltd. (purity, >95.6%) according to the sequence reported by Nakamura et al. (1988) [1]. Wild-type TU of zebrafish were obtained from the Peking University College of Chemical Biology and Biotechnology.

For the acute toxicity test, zebrafish embryos of 3 hours postfertilization (hpf) at three stages of embryonic development were exposed to TP I in a 6-well plate up to 24 hpf. The embryonic development of zebrafish is commonly divided into seven stages, the zygote (0–0.75 hpf), cleavage (0.75–2.25 hpf), blastula (2.125–5.25 hpf), gastrula (5.25–10 hpf), segmentation (10–24 hpf), pharyngula (24–48 hpf), and hatching (48–72 hpf) periods. And we also found that the abnormal morphology of 3 hpf embryos treated with TP I appeared at 24 hpf. Therefore, we collected 20 hpf embryos to extract total RNA for gene microarray analysis.

The concentrations used were as follows: 0, 1.5, and 2.5 μg/mL. Four replicates were made for each concentration, and each replicate consisted of three wells. Each well contained 10 mL of treatment solution and 50 viable embryos. At 20 hpf, 50 abnormal embryos were selected from each replicate to extract total RNA for gene microarray. Four gene chips were used for each concentration. The NimbleGen array used interrogates 38,489 transcripts from Ensembl build Zv7 with 3 probes per transcript. The raw intensities were normalized in RMA method by NimbleScan v2.5 and the data quality was assessed by boxplot and scatter plot. The raw intensities were normalized in RMA method by NimbleScan v2.5 and the data quality was assessed by boxplot and scatter plot, as demonstrated in Figure 1. All in all, the expression values of 26272 transcript sequences in 12 samples of zebra embryo treated with the different doses of TP I are recorded.

2.2. HexaMplot and the Probabilistic Model-Based Hough Transform. Considering the different doses of TP I with 0, 1.5 μg/mL, and 2.5 μg/mL, we employed the probabilistic model-based HT in the modified hexaMplot to assess the toxic effect of TP I on the embryonic development of zebrafish. The original hexaMplot is a two-dimensional representation of R, G, and B intensities and proposed to assess the drug effect on assayed genes in [7]. HexaMplot provides a simple, intuitive, and efficient tool for assessing drug effect.

Similarly, we modified the raw hexaMplot with $x_1 = \log E_1/E_0$ and $x_2 = \log E_2/E_0$, where E_0, E_1, and E_2 reflect expression levels of the genes in the normal and 1.5 and 2.5 ug/mL treated samples, respectively. As demonstrated

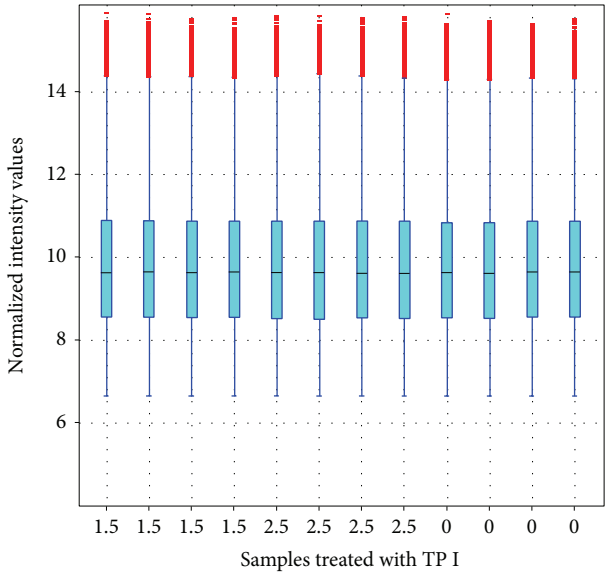


FIGURE 1: The boxplot of the normalized expression data of zebrafish embryonic samples treated with 1.5- and 2.5-μg/mL TP I and blank samples. Four replicates were used for each sample.

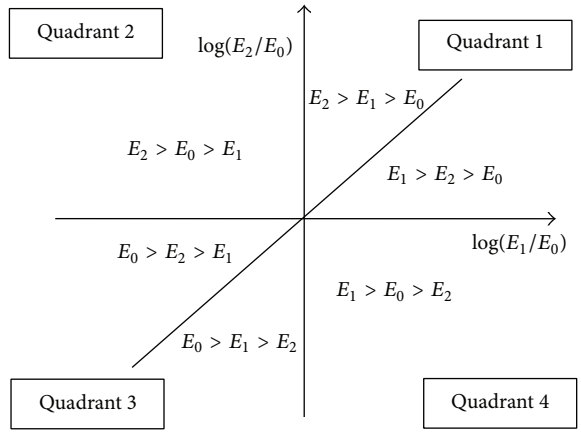


FIGURE 2: The modified layout of the hexaMplot to assess the toxic effect.

in Figure 2, the genes appearing in the upper and lower half-plane of the modified hexaMplot are up- and downregulated, respectively, by the 2.5 ug/mL TP I. Analogously, genes located in the left and right half-plane are up- and downregulated, respectively, by the treatment with 1.5 ug/mL TP I. Also note that along the slant axis $x_2 = x_1$, we have $\log E_2 = \log E_1$, meaning that the expression levels of genes in the 1.5 and 2.5 ug/mL treated samples are the same. Generally, the researchers would like to explore gradual alternation of assayed genes in varying the doses of TP I. The gene representation clustering the slant axis may imply that they are not sensitive to change of the doses of TP I.

Comparatively, the genes in quadrant 2 (Q2) and Q4 may reverse their expression patterns, either further enhancing upregulation or suppressing the downregulation of the gene by the adding dose. Generally, the three axes and six regions in the modified hexaMplot show intuitively the gradual change of gene expression to assess the toxic effect of TP I with the different doses.

Based on the intuitive illustration of hexaMplot, some statistical methods were proposed to identify the functional genes [13–15]. The probabilistic model-based HT was proposed in [15] to detect groups of genes with similar expression patterns with different doses of TP I. Each such group is aligned along a line ray starting in the hexaMplot or origin. The direction of the ray signifies whether the addition of TP I dose has positive or negative effect on expression of the group of genes, while the angle measures the effect level [7]. The lines were detected through HT. As discussed in [15], the model was improved by formulating the Bayesian probabilistic models with Gaussian kernel. Given an observation vector E, the posterior probability of parameter (α, r) is given by

$$p(\alpha, r \mid E) = \frac{p(E \mid \alpha, r)\, p(\alpha, r)}{\int_{\Omega} p(E \mid \alpha', r')\, p(\alpha', r')\, d\alpha' dr'}, \quad (1)$$

where α and r are the angle and distance of the point on the line from the origin and Ω is the range space of the two parameters. We are interested in the data points aligned along the line passing through the origin. So r can be integrated and then we obtain the probability of support for the angle α given the observation $p(\alpha \mid E)$. Given a set of observations and the corresponding Gaussian support kernel, the posterior probability of the set of selected points can be calculated to assess how much the set supports the line ray.

The probabilistic model showed some advantages in detecting the lines and their supporting gene points. First all assayed genes are considered in the algorithm instead of the differentially expressed genes only. Second the probabilistic model explicitly takes into account the size and the negatively correlated nature of the noise associated with hexaMplot gene representations. Third, both the strength of association of individual genes with a particular group (line ray in hexaMplot) and the support for the group by the selected genes can be quantified in a principled manner through posterior probabilities over the line angles [15].

3. Results

3.1. The Identification of Significant Genes. With the microarray data matrix of 26272 rows and 12 columns, the direct 2-fold change with $P < 0.05$ in t-test was used to identify the significant genes. Obviously, the traditional t-test method can only compare between the treatment and control groups. There are two treatment groups in our microarray experiment. So the three groups of t-test between any two groups were made. First we compared the samples treated with 1.5 μg/mL TP I and the blank groups. And 212 differentially expressed genes were identified, in which there were 102 upregulated genes and 110 downregulated.

Similarly, 307 significant genes were identified by comparing the expression data between 2.5 μg/mL TP I treatment group and blank group. And 147 genes were upregulated and 260 were downregulated. We also detected the differentially expressed genes between the two groups, respectively, treated

with 2.5 and 1.5 μg/mL TP I. 111 upregulated and 245 downregulated genes are identified.

Among all of the differentially expressed genes, 51%, 34%, and 11% were associated with the biological process, molecular function, and cellular components in the analysis of gene ontology, respectively. We found that most of the identified genes are related to the Jak/STAT signaling pathway, adherent junction signaling pathway, and tight junction signaling pathway.

Furthermore, we screened a series of significant genes related to the development of zebrafish including Ntl (no tail) and Tbx24 (T-box 24) (related to body axis), Pes (pescadillo) (related to liver), and vascular endothelial growth factor-(VEGF-) Ab1 (related to vasculature). Microarray analysis showed that the expressions of the development-related differential expression genes CYP11A1, Pes, Ntl, VEGF-Ab1, and Tbx24 significantly changed. The findings suggest that TP I interfered with the normal embryonic development of zebrafish.

VEGF is a major vasculogenic and angiogenic factor in embryonic vessel development. It is involved in many biological processes, including the growth and differentiation of vein cells, the growth of endothelial cells, and the permeability of vessels. For zebrafish, VEGF-A and Flk-1 play a crucial role in the angiogenesis of axes and intersegments. Morphological analysis has detected the significant absence of both axial and intersegmental vasculature in zebrafish with downregulated VEGF [16, 17].

CYP11A1 (cytochrome P450 subfamily XIA polypeptide 1), a member of the CYP450 family, which is related to the metabolism of xenobiotics, is expressed in the yolk syncytial layer during early embryogenesis [18, 19]. Therefore, significant expression of CYP11A1 may generate pericardial edema. In the present study, VEGF and CYP11A1 were significantly down- and upregulated, respectively. So it may be concluded that they are related to pericardial edema and abnormalities in the intersegmental vasculature.

T-box transcription factors are a large family of transcriptional regulators involved in many aspects of embryonic development, such as the development of neural tubes and somites [20–26]. Ntl and Tbx24, which are zebrafish T-box genes, are required for the development of the trunk, notochord and tail mesoderm, and medial floor plate [21–23]. During the embryogenesis of zebrafish, Ntl and Tbx24 are located in and derived from notochord cells and somites, respectively [25]. The research has shown that inhibiting Ntl and Tbx24 causes spinal flexion and somite defects [26]. Our microarray analysis also showed that Ntl and Tbx24 were significantly down- and upregulated. We conclude that these transcription factors may cause spinal flexion.

3.2. Genomic and Functional Analysis of the Toxic Effect of TP I. In this subsection, we applied the modified hexaMplot and probabilistic model-based HT to our microarray data to assess the toxic effect of TP I on the embryonic development of zebrafish. There are 26272 spots assayed in 12 microarrays obtained in four repeats of the normal and 1.5 ug/mL and 2.5 ug/mL treated embryonic samples. The original microarray data were normalized using LIMMA algorithms between and within microarrays [8]. The modified hexaMplot of the normalized data is shown in Figure 3. Three significant lines can be detected with the Bayesian probabilistic model. The optimal threshold is set to $\sigma = 0.04$ according to the algorithm [15]. And the corresponding gene points supporting the lines are marked with red, green, and blue in Figure 3.

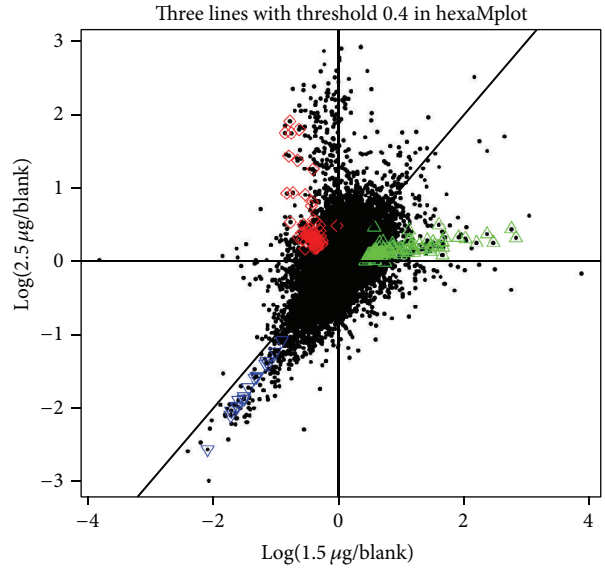


Figure 3: Detected line rays (solid bold lines) for sigma = 0.04 and selected points supporting the three lines.

In comparison with the layout of Figure 2, we can infer some conclusion about the toxic effect of TP I with different doses. In Q1 of Figure 2, 93 green points were identified to support one line ray. And the angle α of the line ray in Q1 is small. In other words, the 93 genes that supported the line are greatly suppressed from the upregulation to normal with the increase of dose of TP I. The group of 64 genes in Q2 shows the reverse trend of expression; that is, the downregulation in the samples treated with 1.5 ug/mL is enhanced to upregulation with 2.5 ug/mL. As to the 17 blue points in Q3, the downregulated genes are further suppressed with adding the dose of TP I. But the angle bias from axis $x_2 = x_1$ is very small. So the alternation of downregulation is not significant.

It is well known that coexpressed genes, not one gene, are involved in the biological function together. So we investigated the biological meaning of the three detected groups of genes (the points supporting the line rays in Figure 3) with gene ontology (GO) framework [27]. The results of GO analysis are summarized in Table 1. The table has the following organization: representative GO terms and its biological meaning for genes are listed in the first and second column. For each GO term, we report four terms of enrichment in the last column, including the total number of genes (N), the number of genes annotated to the GO term (B), the number of genes from our assayed gene annotated to it (n), and the number of genes in the intersection (b).

Table 1: The toxic effect on the embryonic development of zebrafish.

Line	GO term	Description	P value	Enrichment (N, B, n, b)
1	GO:0002483	Antigen processing and presentation of endogenous peptide antigen	$3.99E-5$	178.99 (6712, 3, 25, 2)
	GO:0019885	Antigen processing and presentation of endogenous peptide antigen via MHC class I	$3.99E-5$	178.99 (6712, 3, 25, 2)
	GO:0019883	Antigen processing and presentation of endogenous antigen	$7.96E-5$	134.24 (6712, 4, 25, 2)
	GO:0007218	Neuropeptide signaling pathway	$1.41E-4$	28.77 (6712, 28, 25, 3)
2	GO:0000790	Nuclear chromatin	$1.81E-4$	9.31 (6712, 68, 53, 5)
	GO:0000785	Chromatin	$4.71E-4$	5.94 (6712, 128, 53, 6)
	GO:0071778	WINAC complex	$6.03E-4$	50.66 (6712, 5, 53, 2)
	GO:0016591	DNA-directed RNA polymerase II, and holoenzyme	$6.03E-4$	50.66 (6712, 5, 53, 2)
3	GO:0005212	Structural constituent of eye lens	$7.99E-7$	1,118.67 (6712, 4, 3, 2)

The corresponding P values of the GO terms enriched were computed according to the following hypergeometric distribution and the values were listed in the third column of Table 1. Consider

$$P\text{ value} = \sum_{j=b}^{B} \frac{\binom{n}{j}\binom{N-n}{B-j}}{\binom{N}{B}}. \quad (2)$$

According to the biological function in Table 1, the 93 genes on the green line in Q1 are mainly involved in the antigen processing and presentation of endogenous peptide antigen. And their upregulation of the gene group is a little suppressed with adding dose of TP I from 1.5 to 2.5 ug/mL. We may conclude that the varying doses of TP I show little effect on the antigen process.

Similarly, the 64 genes of the red line in Q2 are related to the chromatin, WINAC complex, or DNA-directed RNA polymerase II, holoenzyme. The adding dose can reverse the gene expression pattern from downregulation to upregulation.

As mentioned in the previous literature of TP I, the peptide is very active against both Gram-positive and Gram-negative bacteria, fungi, viruses, and cancer cells and can interact with DNA and inhibit the synthesis of macromolecules. Our result may further show the biological mechanism in genome level. The 17 genes in Q3 are of interest.

We found that these genes may play an important role in the structural constituent of eye lens. And the expression of downregulated genes is further enhanced with the increasing dose of TP I. It may be concluded that TP I shows the toxic side effect on eye lens and this effect is strengthened with the adding dose. The result may discover some biological mechanism of TP I in genome level. Of course, more experiments are required to be performed to gain insight into the toxic effect of TP I.

4. Conclusion and Discussion

The advantages of tachyplesin I as a novel antimicrobial peptide are appearing with the insight of its biochemical mechanism of strong antimicrobial and anticancer activity. We originally focused on the genetic analysis of TP I. A series of microarray experiments are performed in our research. In this paper, the toxic effect of TP I on the embryonic development of zebrafish was assessed on the genome level. The hexaMplot was used to illustrate the gene expressions with the varying doses of TP I. The probabilistic model-based Hough transform (HT) was used to classify these coexpressed genes.

In our analysis, three line rays supported with the corresponding 174 genes were detected. The three groups of genes were classified into the coherent GO terms with the high significance as detected by gene ontology analysis. The GO functional groups of these genes, such as antigen processing, nuclear chromatin, and structural constituent of eye lens, may explore the biological mechanism of TP I in genome level. In particular we found that TP I shows the toxic side effect on eye lens and this effect was strengthened with the adding dose, which was of interest in the literature of TP I and provided a new direction for our further research.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This study was supported by the National Natural Science Funds of China (31100958 and 31272474), GDNSF (10151805501000007 and S2011020005160), GDPRSFS

(2012), GDUHTP (2011 and 2013), and Shenzhen Key Basic Research project (JC201005280534A, JC201105201191A, and JCYJ20130331151022276).

References

[1] T. Nakamura, H. Furunaka, T. Miyata et al., "Tachyplesin, a class of antimicrobial peptide from the hemocytes of the horseshoe crab (*Tachypleus tridentatus*). Isolation and chemical structure," *Journal of Biological Chemistry*, vol. 263, no. 32, pp. 16709–16713, 1988.

[2] A. Ramamoorthy, S. Thennarasu, A. Tan et al., "Deletion of all cysteines in tachyplesin I abolishes hemolytic activity and retains antimicrobial activity and lipopolysaccharide selective binding," *Biochemistry*, vol. 45, no. 20, pp. 6529–6540, 2006.

[3] K. Masuda, M. Ohta, M. Ito, S. Ohsuka, T. Kaneda, and N. Kato, "Bactericidal action of tachyplesin I against oral streptococci," *Oral Microbiology and Immunology*, vol. 9, no. 2, pp. 77–80, 1994.

[4] G.-L. Ouyang, Q.-F. Li, X.-X. Peng, Q.-R. Liu, and S.-G. Hong, "Effects of tachyplesin on proliferation and differentiation of human hepatocellular carcinoma SMMC-7721 cells," *World Journal of Gastroenterology*, vol. 8, no. 6, pp. 1053–1058, 2002.

[5] A. V. Hallare, M. Schirling, T. Luckenbach, H.-R. Köhler, and R. Triebskorn, "Combined effects of temperature and cadmium on developmental parameters and biomarker responses in zebrafish (Danio rerio) embryos," *Journal of Thermal Biology*, vol. 30, no. 1, pp. 7–17, 2005.

[6] C. Nathan, F. Stellabotte, and H. Stephen, "Tbx24 is required for proper dermomyotome formation in the posterior trunk of zebrafish," *Developmental Biology*, vol. 344, no. 1, p. 461, 2010.

[7] V. S. Tseng and C.-P. Kao, "Efficiently mining gene expression data via a novel parameterless clustering method," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 4, pp. 355–365, 2005.

[8] L. P. Wang, F. Chu, and W. Xie, "Accurate cancer classification using expressions of very few genes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 1, pp. 40–53, 2007.

[9] S. Mitra and Y. Hayashi, "Bioinformatics with soft computing," *IEEE Transactions on Systems, Man and Cybernetics C*, vol. 36, no. 5, pp. 616–635, 2006.

[10] B. Liu, C. Wan, and L. P. Wang, "An efficient semi-unsupervised gene selection method via spectral biclustering," *IEEE Transactions on Nanobioscience*, vol. 5, no. 2, pp. 110–114, 2006.

[11] S. A. Salem, L. B. Jack, and A. K. Nandi, "Investigation of self-organizing oscillator networks for use in clustering microarray data," *IEEE Transactions on Nanobioscience*, vol. 7, no. 1, pp. 65–79, 2008.

[12] F. Chu and L. P. Wang, "Applications of support vector machines to cancer classification with microarray data," *International Journal of Neural Systems*, vol. 15, no. 6, pp. 475–484, 2005.

[13] H. Zhao, R. N. S. Wong, K.-T. Fang, and P. Y. K. Yue, "Use of three-color cDNA microarray experiments to assess the therapeutic and side effect of drugs," *Chemometrics and Intelligent Laboratory Systems*, vol. 82, no. 1-2, pp. 31–36, 2006.

[14] H. Zhao and H. Yan, "HoughFeature, a novel method for assessing drug effects in three-color cDNA microarray experiments," *BMC Bioinformatics*, vol. 8, pp. 256–266, 2007.

[15] G. Tiňo, H. Zhao, and H. Yan, "Searching for coexpressed genes in three-color cDNA microarray data using a probabilistic model-based hough transform," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 4, pp. 1093–1106, 2011.

[16] J. E. Cannon, P. D. Upton, J. C. Smith, and N. W. Morrell, "Intersegmental vessel formation in zebrafish: requirement for VEGF but not BMP signalling revealed by selective and non-selective BMP antagonists," *British Journal of Pharmacology*, vol. 161, no. 1, pp. 140–149, 2010.

[17] Y. Gao, H. Zhao, X. Feng et al., "Expression of Recombinant Human Lysozyme-tachyplesin I (hLYZ-TP I) in pichia pastoris and analysis of antibacterial activity," *Biomedical and Environmental Sciences*, vol. 26, no. 4, pp. 319–322, 2013.

[18] H.-J. Hsu, P. Hsiao, M.-W. Kuo, and B.-C. Chung, "Expression of zebrafish cyp11a1 as a maternal transcript and in yolk syncytial layer," *Gene Expression Patterns*, vol. 2, no. 3-4, pp. 219–222, 2002.

[19] V. Teixeira, M. J. Feio, and M. Bastos, "Role of lipids in the interaction of antimicrobial peptides with membranes," *Progress in Lipid Research*, vol. 51, no. 2, pp. 149–177, 2012.

[20] D. L. Chapman and V. E. Papaioannou, "Three neural tubes in mouse embryos with mutations in T-box gene Tbx6," *Nature*, vol. 391, no. 6668, pp. 695–697, 1998.

[21] S. L. Amacher, B. W. Draper, B. R. Summers, and C. B. Kimmel, "The zebrafish T-box genes no tail and spadetail are required for development of trunk and tail mesoderm and medial floor plate," *Development*, vol. 129, no. 14, pp. 3311–3323, 2002.

[22] M. Nikaido, A. Kawakami, A. Sawada, M. Furutani-Seiki, H. Takeda, and K. Araki, "Tbx24, encoding a T-box protein, is mutated in the zebrafish somite-segmentation mutant fused somites," *Nature Genetics*, vol. 31, no. 2, pp. 195–199, 2002.

[23] I. Kushibiki, M. Kamiya, T. Aizawa et al., "Interaction between tachyplesin I, peptide derived from horseshoe crab, and lipopolysaccharide," *Biochimica et Biophysica Acta*, vol. 1844, no. 3, pp. 527–534, 2014.

[24] S. Schulte-Merker, M. Hammerschmidt, D. Beuchle, K. W. Cho, E. M. De Robertis, and C. Nüsslein-Volhard, "Expression of zebrafish goosecoid and no tail gene products in wild-type and mutant no tail embryos," *Development*, vol. 120, no. 4, pp. 843–852, 1994.

[25] B. W. Bisgrove, J. J. Essner, and H. J. Yost, "Multiple pathways in the midline regulate concordant brain, heart and gut left-right asymmetry," *Development*, vol. 127, no. 16, pp. 3567–3579, 2000.

[26] D. Qiu, X. Liu, J. Wang, and Y. Su, "Remove from marked Records Artificial synthesis of TAT PTD-tachyplesin fusion gene by overlap extension PCR," *Journal of Agricultural Biotechnology*, vol. 2, no. 3, pp. 1–4, 2013.

[27] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 340758, 7 pages
http://dx.doi.org/10.1155/2014/340758

Research Article

Integrating Gene Expression and Protein Interaction Data for Signaling Pathway Prediction of Alzheimer's Disease

Wei Kong,[1] Jingmao Zhang,[1] Xiaoyang Mou,[2] and Yang Yang[3]

[1] *Information Engineering College, Shanghai Maritime University, Shanghai 201306, China*
[2] *DNJ Pharma and Rowan University, Glassboro, NJ 08028, USA*
[3] *Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*

Correspondence should be addressed to Wei Kong; weikong@shmtu.edu.cn

Received 23 February 2014; Accepted 18 March 2014; Published 9 April 2014

Academic Editor: Tao Huang

Discovering the signaling pathway and regulatory network would provide significant advance in genome-wide understanding of pathogenesis of human diseases. Despite the rich transcriptome data, the limitation for microarray data is unable to detect changes beyond transcriptional level and insufficient in reconstructing pathways and regulatory networks. In our study, protein-protein interaction (PPI) data is introduced to add molecular biological information for predicting signaling pathway of Alzheimer's disease (AD). Combining PPI with gene expression data, significant genes are selected by modified linear regression model firstly. Then, according to the biological researches that inflammation reaction plays an important role in the generation and deterioration of AD, NF-κB (nuclear factor-kappa B), as a significant inflammatory factor, has been selected as the beginning gene of the predicting signaling pathway. Based on that, integer linear programming (ILP) model is proposed to reconstruct the signaling pathway between NF-κB and AD virulence gene APP (amyloid precursor protein). The results identify 6 AD virulence genes included in the predicted inflammatory signaling pathway, and a large amount of molecular biological analysis shows the great understanding of the underlying biological process of AD.

1. Introduction

Alzheimer's disease (AD) is a progressive and fatal neurodegenerative disorder manifested by cognitive and memory deterioration. The characteristic pathology changes in AD are fibrin deposition in cerebral cortex; it is the deposition of beta-amyloid (Aβ) in cell space and poly-Tau protein in cell. In pathomorphism, the expression is senile plaques (SP) and neurofibrillary tangles (NFT).

Many studies have investigated the mechanism of AD from various perspectives of its complexity. Recent researches show that a more accepted hallmark of AD is brain inflammation. Inflammation clearly occurs in pathologically vulnerable regions of AD brain and it does so with the full complexity of local peripheral inflammatory responses [1–3]. In the periphery, degenerating tissue and the deposition of highly insoluble abnormal materials are classical stimulants of inflammation. Likewise, in the AD brain damaged neurons and neurites and highly insoluble Aβ peptide deposits and neurofibrillary tangles provide obvious stimuli for inflammation [4–7].

To give insight to the AD mechanisms, high-throughput gene expression data has received extensive attention and made substantial progress in reconstructing the gene regulatory network. However, due to the underlying shortcomings of microarray technology such as small sample size, measurement error, and information insufficiency, unveiling disease mechanism has remained a major challenge to the AD research community. To overcome these problems, pathway information and network-based approaches [8] have been applied and become more informative and powerful for discovering disease mechanism.

Protein-protein interaction (PPI) networks are reconstructed from protein domain characteristics, gene expression data, and structure-based information with other evidence, for example, gene homology, function annotations,

and sequence motifs [9]. PPI data contain structure information among different genes while gene expression data do not. In our study, PPI network data as a priori pathway information is introduced for predicting the inflammatory signaling pathway in AD. Many literatures have given outstanding achievements by integrating gene expression data and PPI data, such as identification of protein complexes [10], small subnetworks [11], and biomarkers [12]. Zhao et al. presented an integer linear programming (ILP) method to uncover pathways among the given starting proteins, ending proteins, and some transduction factor proteins [13]. However, how to select the transduction factor proteins is a great problem. In our study, a modified network-constrained regularization analysis method [14] is proposed for linear regression analysis to select appropriate number of significant genes. Simulation results show that this method can lead to an efficiently global smoothness of regression coefficients.

Based on that, ILP model is presented to reconstruct the inflammatory signaling pathway by integrating PPI data with the AD gene expression data. In the ILP model, the starting and ending proteins of the predicting pathway need to be arranged in advance. Nuclear transcription factor NF-κB (nuclear factor-kappa B) as one of the most important inflammatory factors is selected as the starting gene of the signaling pathway. As we know that NF-κB plays a key role in regulating the immune response to infection, therefore incorrect regulation of NF-κB has been linked to cancer, inflammatory and autoimmune diseases, septic shock, viral infection, and improper immune development. NF-κB has also been implicated in processes of synaptic plasticity and memory [15]. On the other hand, APP (amyloid precursor protein) as the most important AD virulence gene and precursor protein of Aβ is arranged as the ending protein of the predicting pathway.

The experiment results show that 6 AD virulence genes are identified being included in the predicted inflammatory signaling pathway, and a large amount of inflammation related genes and pathways has been found by molecular biological analysis and they show the great understanding of the pathogenesis of AD.

2. Methods

2.1. Linear Regression Model. Linear regression model is widely used in estimation and variable selection. In our study, the model is applied to selected subset of significant genes which are important for AD and are going to be the transduction factors of reconstructing pathway. In the next prediction step, gene expression data and PPI data will be integrated by ILP model. After all of the above, a pathway could be identified between NF-κB and APP. The usual linear regression model can be expressed as

$$\mu = \sum_{j=1}^{p} x_j \beta_j = x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p, \tag{1}$$

where $\mu = (\mu_1, \mu_2, \ldots, \mu_n)^T$ is response vector, n is the sample number, $x_j = (x_1, x_2, \ldots, x_n)^T$, $j = 1, 2, \ldots, p$, is the predictor, represented by the jth gene's expression data in all samples, and β_j is the jth gene's weight vector. Assume that the predictors are standardized and the response is centered, we get

$$\sum_{i=1}^{n} \mu_i = 0, \qquad \sum_{i=1}^{n} x_{ij} = 0, \qquad \sum_{i=1}^{n} x_{ij}^2 = 0, \quad \text{for } j = 1, 2, \ldots, p. \tag{2}$$

Gene expression data has the characteristic of less sample and great noise. As a simple model, linear regression model has significant performance in handling less sample and great noise data. The significant genes will get a larger coefficient while the nonsignificant genes will get a smaller coefficient.

2.2. Network-Constrained Regularization for the Linear Regression Model. Before using linear regression model, coefficient β needs to be estimated. Many methods have been proposed which focused on addressing high-dimensionality genomic data such as LASSO, LA-SEN, and LARS. Here, a modified network-constrain regularization analysis by C. Li and H. Li [14] is applied to estimate the coefficient since it has been proved to perform better than other methods. This method is a lasso-type problem. It defines a normalized Laplacian matrix L as

$$L = \begin{cases} \dfrac{1 - w(u,v)}{d_u} & \text{if } u = v \text{ and } d_u \neq 0, \\ \dfrac{-w(u,v)}{\sqrt{d_u d_v}} & \text{if } u \text{ and } v \text{ are adjacent,} \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $w(u, v)$ represents the weight of edge between linked genes u and v. $d_v = \sum_{u \sim v} w(u, v)$ represents all the adjacent genes of v on the network. Then the definition of the network-constrained regularization criterion is

$$L(\lambda_1, \lambda_2, \beta) = (\mu - X\beta)^T (\mu - X\beta) + \lambda_1 |\beta|_1 + \lambda_2 \beta^T L \beta, \tag{4}$$

where $X = (x_1|K|x_p)$; $|\beta_1| = \sum_{j=1}^{p} |\beta_j|$; λ_1, λ_2 are nonnegative turning parameters. And then we estimate β by minimizing (4):

$$\beta = \underset{\beta}{\operatorname{argmin}} \{L(\lambda_1, \lambda_2, \beta)\}. \tag{5}$$

Minimizing (4) is equivalent to solving a lasso-type optimization problem. Turning parameters are estimated by 10-fold cross-validation (CV). Genes in gene interaction network are selected by PubGene; we chose genes related to Alzheimer.

2.3. Integer Linear Programming (ILP). The ILP model formulates signaling network detection as an optimization problem and treats a signaling network as a whole entity as described in its original publication [13]. PPI network is a weighted undirected graph, that can be described as

$\mathbf{G}(\mathbf{V}, \mathbf{E}, \mathbf{W})$, where $\mathbf{V}$ is vertices in the graph, representing protein; $\mathbf{E}$ is edge between proteins; and $\mathbf{W}$ represents the weight of edges. $\mathbf{W}$ can be calculated by gene expression data. The ILP model can be described as follows:

$$\begin{aligned}
&\underset{\{x_i, y_{ij}\}}{\text{Minmize}} \quad S = -\sum_{i=1}^{|V|}\sum_{j=1}^{|V|} w_{ij} y_{ij} + \lambda \sum_{i=1}^{|V|}\sum_{j=1}^{|V|} y_{ij} \\
&\text{Subject to:} \quad y_{ij} \le x_i, \\
&\qquad y_{ij} \le x_j, \\
&\qquad \sum_{j=1}^{|V|} y_{ij} \ge 1, \quad \text{if } i \text{ is either a starting} \\
&\qquad \text{or ending protein,} \\
&\qquad \sum_{j=1}^{|V|} y_{ij} \ge 2x_i, \quad \text{if } i \text{ is not a starting} \\
&\qquad \text{or ending protein,} \\
&\qquad x_i = 1, \quad \text{if } i \text{ is a protein known in STN,} \\
&\qquad x_i \in \{0, 1\}, \quad i = 1, 2, \ldots, |V|, \\
&\qquad y_{ij} \in \{0, 1\}, \quad i, j = 1, 2, \ldots, |V|,
\end{aligned} \tag{6}$$

where w_{ij} is the weight between proteins i and j in weighted undirected graph G; y_{ij} is a binary variable to denote whether the edge $E(i, j)$ is a part of the STN. x_i is also a binary variable to denote whether protein i is a component of the STN. λ is a positive penalty parameter. $|V|$ includes all proteins in the PPI network. $y_{ij} \le x_i$ and $y_{ij} \le x_j$ mean that only if proteins i and j are both components of STN, the edge $E(i, j)$ should be considered. $\sum_{j=1}^{|V|} y_{ij} \ge 1$ represents at least one protein contact with starting protein or ending protein. $\sum_{j=1}^{|V|} y_{ij} \ge 2x_i$ makes sure that if x_i is selected as a component of STN, there are at least two proteins link to the vertex.

The starting protein and ending protein have confirmed above that the genes selected by linear model were treated as transduction factors. The method of chosen parameter λ can be found in its original publication. Then we detected protein pathway by the ILP model.

3. Results and Discussion

To evaluate ILP model, AD dataset, series GSE1297, was used which were human hippocampal gene expression downloaded from GEO DataSets from the National Center for Biotechnology Information (NCBI) offered by Blalock et al. [16]. The hippocampal specimens they used are obtained through the Brain Bank of the Alzheimer's Disease Research Center at the University of Kentucky. The human Gene Chips (HG-U133A) of Affymetrix and Microarray Suite 5 are used in analyzing the microarray data. There are a total of 9 control, 7 incipient, 8 moderate, and 7 severe AD samples included in this dataset with 22283 gene expressions in each sample. The PPI data we used is downloaded from website BioGRID (http://thebiogrid.org/) with 12466 proteins and 40323 interactions in total.

The file format of microarray data downloaded from NCBI is CEL. The probe data needs data processing like background correct, normalization, probe correct, and so on. Then ANOVAs were used on preliminary select genes and removed all genes whose P value was less than 0.05. After processing, 7030 genes for each sample were left. Then taking linear regression model with modified network-constrained regularization and AD biological information, the coefficients $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ and $p = 7030$ were obtained. Among them, 7017 values of β were zeroes and the other 13 βs were nonzero values. So these 13 genes with nonzero values of β were considered as significant genes to AD phenotype and they are denoted in green circles with "(1)" and gene names in Figure 1.

The 13 selected genes can be mapped to PPI network to get the corresponding proteins and the interactions between them and other proteins. Each selected gene was connected with some other genes in PPI network by the edges. In ILP algorithm, edge between i and j was represented by y_{ij}. When ILP chose this edge, $y_{ij} = 1$, otherwise $y_{ij} = 0$.

ILP tries to assign 0 or 1 for y_{ij} to ensure the result network has the largest weight. For the weight of edge, w_{ij}, here we use the Pearson coefficient of the gene expression values to represent the weight between proteins i and j.

Then using NF-κB as starting protein and APP as ending protein ILP model was applied to formulate the signaling network. In the ILP algorithm, penalty parameter λ is a size control parameter that needs to be adapted manually. If its value is too large, the predicted signaling network will be enormous, otherwise it will be too small to catch the useful biological information. In our simulation experiment, after adapting from small value to large value, λ was determined as 0.65.

We finally got a signaling pathway with several small subnetworks. This network is reconstructed by 45 genes including 13 selected significant genes and is shown in Figure 1.

In Figure 1, "(0) NF-κB" represents the starting protein NF-κB, "(3) APP" represents the ending protein APP, "(1)" with the protein names denote the corresponding selected genes by the regress model, and "(2)" with protein names are selected by ILP to reconstruct the signaling pathways between NF-κB and APP. In order to analyze the biological functions of the pathways and subnetworks, the predicted result was mapped into its coding gene pathway network, and the online analysis website DAVID (http://david.abcc.ncifcrf.gov/home.jsp) was utilized to further understand their molecular biological functions to AD. Table 1 shows the KEGG pathway analysis result.

First of all, among the prediction results, there are 5 genes that have been confirmed as the AD virulence genes such as SNCA, CALM1, GSK3B, PSEN1, and APP which have been biologically demonstrated playing crucial roles in AD. Based

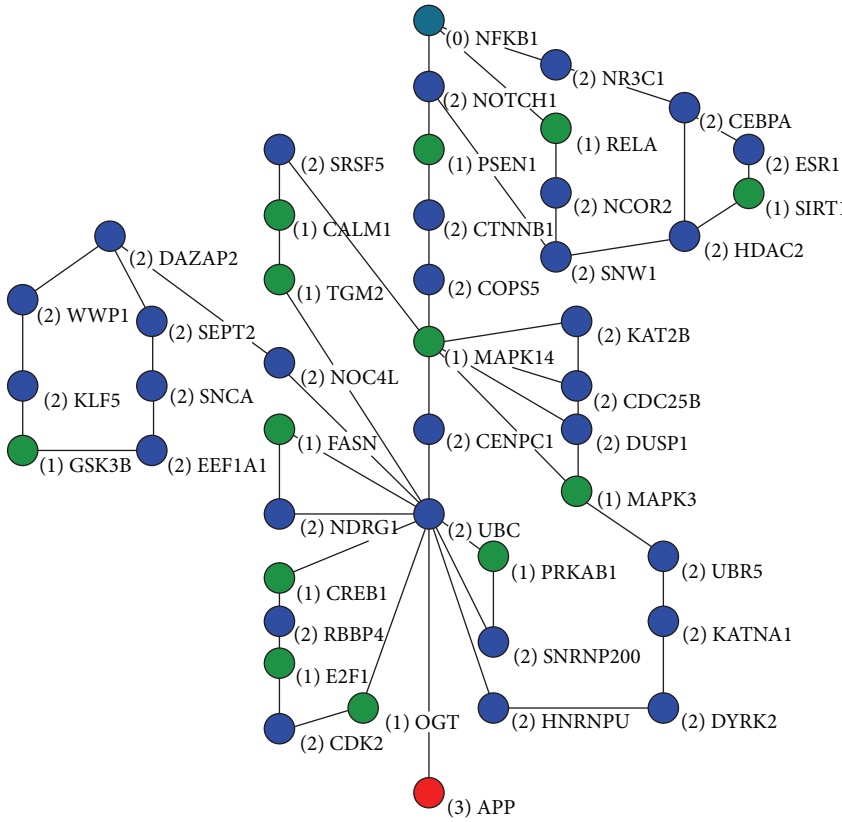


Figure 1: Protein signaling pathway predicting result between NF-κB and APP.

on Table 1, T cell receptor signaling pathway, B cell receptor signaling pathway, the Notch signaling pathway, NOD-like receptor signaling pathway, Toll-like receptor signaling pathway, MAPK signaling pathway, neurotrophin signaling pathway, insulin signaling pathway, and so on were found to include a major part of important genes derived from the regression model. Specially, the main predicted pathway in Figure 1 includes NFKB1, NOTCH1, PSEN1, CTNNB, COPS5, MAPK14, CENPC1, UBC, and APP; the molecular biological analysis shows that they have close correlation between inflammatory response and AD.

It was found that inflammation is a major mechanism of acute brain injury and chronic neurodegeneration [17]. During the onset of an inflammatory response, signaling pathways are activated for translating extracellular signals into intracellular responses converging to the activation of NF-κB, the central transcription factor in driving the inflammatory response [18]. NF-κB has long been considered a prototypical proinflammatory signaling pathway, largely based on the activation of NF-κB by proinflammatory cytokines, such as interleukin-1 (IL-1) and tumor necrosis factor α (TNFα), and the role of NF-κB on the expression of other proinflammatory genes including cytokines, chemokines, and adhesion molecules, which has been extensively reviewed elsewhere [9].

Recent studies have also found that Notch receptors in Notch signaling pathway regulate cell differentiation and function, and Notch1 has been shown to induce glia in the peripheral nervous system [19, 20]. NF-κB, Notch, MAKP, and PSEN1 included in the main pathway of Figure 1 were observed to have strong regulating functions between each other, since interleukin-1 (IL-1) activates NF-κB via interleukin-1 receptor-associated kinase (IRAK) and mitogen-activated protein kinase (MEKK1(MAP3 K)) dependent inhibition of NF-κB inhibitor (I-κB) [21, 22]. V-rel reticuloendotheliosis viral oncogene homolog (c-Rel (NF-κB subunit)) can trigger Notch homolog 1 translocation-associated (NOTCH1 receptor) signaling pathway by inducing expression of Jagged1, ligand for Notch receptors [23, 24]. NOTCH1 receptor activated by Jagged1 or Delta-like 1 (DLL1) is cleaved by ADAM metallopeptidase domain 17 (ADAM17) and PSEN1 to intracellular domain of NOTCH1. NOTCH1 is transported to nucleus and participates in recombination signal binding protein for immunoglobulin kappa J region (RBP-J kappa (CBF1)) mediated transcription [24, 25].

It was also found that β-catenin- (CTNNB1-) dependent WNT signaling pathways have crucial roles in the regulation of diverse cell behaviours, including cell fate, proliferation, survival, differentiation, migration, and polarity [26, 27]. It is interesting to note that loss of TNFα function would inhibit Wnt/β-catenin signaling [28]. Recently studies show that Wnt/β-catenin and NF-κB are independent pathways; cross-regulation between the Wnt and NF-κB signaling pathways has emerged as an important area for the regulation of a diverse array of genes and pathways active in chronic inflammation, immunity, development, and tumorigenesis.

Table 1: KEGG pathway analysis of the predicted pathways and subnetworks in Figure 1.

Pathway	Number of genes
Prostate cancer	9
Notch signaling pathway	8
Neurotrophin signaling pathway	6
Pathways in cancer	6
Chronic myeloid leukemia	6
Alzheimer's disease	5
Melanogenesis	9
T cell receptor signaling pathway	5
Acute myeloid leukemia	5
NOD-like receptor signaling pathway	5
Cell cycle	5
Insulin signaling pathway	6
Pancreatic cancer	4
B cell receptor signaling pathway	4
Small cell lung cancer	4
Progesterone-mediated oocyte maturation	4
MAPK signaling pathway	4
Toll-like receptor signaling pathway	5
Endometrial cancer	6
Spliceosome	4
Glioma	5
Adipocytokine signaling pathway	3
Epithelial cell signaling in *Helicobacter pylori* infection	3
RIG-I-like receptor signaling pathway	3
Colorectal cancer	3

Both β-catenin and NF-κB activate inducible nitric oxide synthase (iNOS) gene expression [29].

In addition, the regulatory network between COPS5 and CENPC1 has been extracted from our algorithm which is also observed to be implicated in the pathogenesis of AD. The COP9 (constitutive photomorphogenesis 9) signalosome (COPS), a large multiprotein complex that resembles the 19S lid of the 26S proteasome, plays a central role in the regulation of the E3-cullin RING ubiquitin ligases (CRLs). The catalytic activity of the COPS complex, carried by subunit 5 (COPS 5/Jab1), COPS-dependent COPS 5, displays isopeptidase activity; it is intrinsically inactive in other physiologically relevant forms [30]. Increased APP and accumulation of neurotoxic Aβ in the brain are central to the pathogenesis of AD. COPS5 is found to be a novel RanBP9-binding protein that increases APP processing and Aβ generation [31]. COPS5 regulates the stability of the inner kinetochore components CENP-T and CENP-W, providing the first direct link between COPS5 and the mitotic apparatus and highlighting the role of COPS5 as a multifunctional cell cycle regulator [32]. CENP-T interacts with both centromeric chromatin and microtubule binding kinetochore complexes. Transient targeting of CENP-C to a noncentromere LacO locus induces the recruitment of some outer kinetochore proteins, similar to CENP-T [33]. Our result exhibited that COPS5 regulates CENP-C in the main pathway and ubiquitin and NF-κB were found to be associated with them. Ubiquitin can degrade the IκB which is the inhibitor of NF-κB, processing of precursors, and activation of the IκB kinase (IKK) through a degradation-independent mechanism [34]. On the other hand, COPS5 functions through CDK2 to control premature senescence in a novel way, depending on cyclin E in the cytoplasm [35].

4. Conclusions

Although many efforts have been done several decades of AD, it is still difficult to uncover its phenotype-pathway relationship and pathogenesis. Recent studies show that the pathology of AD has an inflammatory component that is characterized by upregulation of proinflammatory cytokines, particularly in response to Aβ. However, the signaling pathways and regulatory networks of the inflammation in AD pathogenesis are very difficult to reconstruct due to the complexity.

To discover the inflammation signaling pathway and regulatory network of AD, in our study, protein interactive network data, PPI was introduced to overcome the information insufficiencies of DNA microarray gene expression data by integer linear programming (ILP) method. Two stages had been used in predicting inflammatory pathway for AD. Firstly, significant genes had been selected by linear regression analysis with the modified network-constrained regularization analysis. Then ILP model was applied to reconstruct the signaling pathway between NF-κB and AD virulence gene APP since NF-κB has long been considered a prototypical proinflammatory signaling pathway. From the molecular biology analysis, we found that genes on the main pathway of the reconstruction results play crucial roles in inflammatory response and APP which give more biological insight for AD pathogenesis, such as NF-κB, NOTCH1, CTNNB1, COPS5, and their signaling pathways. Even more, the pathogenic contribution of the inflammatory response in AD is supported by our finding of the regulating and functions of the genes and subnetworks in the predicted signaling pathways. In general, our studies on combining PPI and gene expression data discover the signaling pathways of inflammatory response on AD and help for deeply understanding the pathogenesis of AD.

Conflict of Interests

The authors declare no conflict of interests.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (nos. 61271446 and 61003093).

References

[1] A. Haruhiko, S. Barger, S. Barnum et al., "Inflammation and Alzheimer's disease," *Neurobiology of Aging*, vol. 21, pp. 383–421, 2000.

[2] M. Meyer-Luehmann, T. L. Spires-Jones, C. Prada et al., "Rapid appearance and local toxicity of amyloid-β plaques in a mouse model of Alzheimer's disease," *Nature*, vol. 451, no. 7179, pp. 720–724, 2008.

[3] Y.-J. Lee, S. B. Han, S.-Y. Nam, K.-W. Oh, and J. T. Hong, "Inflammation and Alzheimer's disease," *Archives of Pharmacal Research*, vol. 33, no. 10, pp. 1539–1556, 2010.

[4] D. Galimberti and E. Scarpini, "Inflammation and oxidative damage in Alzheimer's disease: friend or foe?" *Frontiers in Bioscience*, vol. 3, pp. 252–266, 2011.

[5] H. Johnston, H. Boutin, and S. M. Allan, "Assessing the contribution of inflammation in models of Alzheimer's disease," *Biochemical Society Transactions*, vol. 39, no. 4, pp. 886–890, 2011.

[6] C. Holmes and J. Butchart, "Systemic inflammation and Alzheimer's disease," *Biochemical Society Transactions*, vol. 39, no. 4, pp. 898–901, 2011.

[7] X. Lan, R. Liu, L. Sun, T. Zhang, and G. Du, "Methyl salicylate 2-O-β-D-lactoside, a novel salicylic acid analogue, acts as an anti-inflammatory agent on microglia and astrocytes," *Journal of Neuroinflammation*, vol. 8, article 98, 2011.

[8] N. Bhardwaj and H. Lu, "Correlation between gene expression profiles and protein-protein interactions within and across genomes," *Bioinformatics*, vol. 21, no. 11, pp. 2730–2738, 2005.

[9] Y. Huang, X. Sun, and G. Hu, "An integrated genetics approach for identifying protein signal pathways of Alzheimer's disease," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 14, no. 4, pp. 371–378, 2011.

[10] J. Feng, R. Jiang, and T. Jiang, "A max-flow-based approach to the identification of protein complexes using protein interaction and microarray data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 3, pp. 621–634, 2011.

[11] Y. Huang, J. Zhang, and Y. Huang, "Computational identification of proteins sub-network in Parkinson's disease study," in *Proceedings of the International Conference on Anti-Counterfeiting, Security and Identification (ASID '12)*, pp. 1–4, 2012.

[12] M. J. Jahid and J. Ruan, "Identification of biomarkers in breast cancer metastasis by integrating protein-protein interaction network and gene expression data," in *Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS '11)*, pp. 60–63, 2011.

[13] X.-M. Zhao, R.-S. Wang, L. Chen, and K. Aihara, "Uncovering signal transduction networks from high-throughput data by integer linear programming," *Nucleic Acids Research*, vol. 36, no. 9, article e48, 2008.

[14] C. Li and H. Li, "Network-constrained regularization and variable selection for analysis of genomic data," *Bioinformatics*, vol. 24, no. 9, pp. 1175–1182, 2008.

[15] X. Zhou, L. Yuan, X. Zhao et al., "Genistein antagonizes inflammatory damage induced by β-amyloid peptide in microglia through TLR4 and NF-κB," *Nutrition*, vol. 30, no. 1, pp. 90–95, 2014.

[16] E. M. Blalock, J. W. Geddes, K. C. Chen, N. M. Porter, W. R. Markesbery, and P. W. Landfield, "Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 7, pp. 2173–2178, 2004.

[17] T. Rolova, L. Puli, J. Magga et al., "Complex regulation of acute and chronic neuroinflammatory responses in mouse models deficient for nuclear factor kappa B p50 subunit," *Neurobiology of Disease*, vol. 15, no. 64, pp. 16–29, 2013.

[18] P. T. Foteinou, S. E. Calvano, S. F. Lowry, and I. P. Androulakis, "In silico simulation of corticosteroids effect on an NFkB-dependent physicochemical model of systemic inflammation," *PLoS ONE*, vol. 4, no. 3, Article ID e4706, 2009.

[19] S. Zanotti and C. Ernesto, "*Notch1* and *Notch2* expression in osteoblast precursors regulates femoral microarchitecture," *Bone*, vol. 62, pp. 22–28, 2014.

[20] K. Tanigaki, F. Nogaki, J. Takahashi, K. Tashiro, H. Kurooka, and T. Honjo, "*Notch1* and *Notch3* instructively restrict bFGF-responsive multipotent neural progenitor cells to an astroglial fate," *Neuron*, vol. 29, no. 1, pp. 45–55, 2001.

[21] P. Renard and M. Raes, "The proinflammatory transcription factor NFκB: a potential target for novel therapeutical strategies," *Cell Biology and Toxicology*, vol. 15, no. 6, pp. 341–344, 1999.

[22] J. Yao, W. K. Tae, J. Qin et al., "Interleukin-1 (IL-1)-induced TAK1-dependent Versus MEKK3-dependent NFκB activation pathways bifurcate at IL-1 receptor-associated kinase modification," *The Journal of Biological Chemistry*, vol. 282, no. 9, pp. 6075–6089, 2007.

[23] J. Bash, W.-X. Zong, S. Banga et al., "Rel/NF-κB can trigger the Notch signaling pathway by inducing the expression of Jagged1, a ligand for Notch receptors," *The EMBO Journal*, vol. 18, no. 10, pp. 2803–2811, 1999.

[24] C. Osipo, T. E. Golde, B. A. Osborne, and L. A. Miele, "Off the beaten pathway: the complex cross talk between Notch and NF-κB," *Laboratory Investigation*, vol. 88, no. 1, pp. 11–17, 2008.

[25] V. Bolós, J. Grego-Bessa, and J. L. De La Pompa, "Notch signaling in development and cancer," *Endocrine Reviews*, vol. 28, no. 3, pp. 339–363, 2007.

[26] J. N. Anastas and R. T. Moon, "WNT signalling pathways as therapeutic targets in cancer," *Nature Reviews Cancer*, vol. 13, no. 1, pp. 11–26, 2013.

[27] K. N. Nejak-Bowen and S. P. S. Monga, "Beta-catenin signaling, liver regeneration and hepatocellular cancer: sorting the good from the bad," *Seminars in Cancer Biology*, vol. 21, no. 1, pp. 44–58, 2011.

[28] M. Gong, C. Liu, L. Zhang et al., "Loss of the TNFα function inhibits Wnt/β-catenin signaling, exacerbates obesity development in adolescent spontaneous obese mice," *Molecular and Cellular Biochemistry*, 2014.

[29] Q. Du and D. A. Geller, "Cross-regulation between Wnt and NF-κB signaling pathways," *Forum on Immunopathological Disease Therapeutics*, vol. 1, no. 3, pp. 155–181, 2010.

[30] A. Echalier, Y. Pan, M. Birol et al., "Insights into the regulation of the human COP9 signalosome catalytic subunit, CSN5/Jab1," , *Proceeding of National Academy of Science of USA*, vol. 110, no. 4, pp. 1273–1278, 2013.

[31] H. Wang, D. Dey, I. Carrera et al., "COPS5 (Jab1) protein increases β site processing of amyloid precursor protein and amyloid β peptide generation by stabilizing RanBP9 protein levels," *The Journal of Biological Chemistry*, vol. 288, no. 37, pp. 26668–26677, 2013.

[32] Y. Chun, M. Lee, and B. Park, "CSN5/JAB1 interacts with the centromeric components CENP-T and CENP-W and regulates

their proteasome-mediated degradation," *The Journal of Biological Chemistry*, vol. 288, no. 38, pp. 27208–27219, 2013.

[33] K. E. Gascoigne, K. Takeuchi, A. Suzuki, T. Hori, T. Fukagawa, and I. M. Cheeseman, "Induced ectopic kinetochore assembly bypasses the requirement for CENP-A nucleosomes," *Cell*, vol. 145, no. 3, pp. 410–422, 2011.

[34] Z. J. Chen, "Ubiquitin signalling in the NF-κB pathway," *Nature Cell Biology*, vol. 7, no. 8, pp. 758–765, 2005.

[35] A. Yoshida, N. Yoneda-Kato, and J. Y. Kato, "CSN5 specifically interacts with CDK2 and controls senescence in a cytoplasmic cyclin E-mediated manner," *Science Reports*, vol. 3, p. 1054, 2013.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 645064, 8 pages
http://dx.doi.org/10.1155/2014/645064

Research Article

MACT: A Manageable Minimization Allocation System

Yan Cui,[1,2] Huaien Bu,[3] Hongwu Wang,[3] and Shizhong Liao[1]

[1] *School of Computer Science and Technology, Tianjin University, 92 Weijin Road, Nankai District, Tianjin 300072, China*
[2] *Department of Common Required Courses, Tianjin University of Traditional Chinese Medicine, 312 Anshanxi Road, Nankai District, Tianjin 300193, China*
[3] *College of Traditional Chinese Medicine, Tianjin University of Traditional Chinese Medicine, 312 Anshanxi Road, Nankai District, Tianjin 300193, China*

Correspondence should be addressed to Hongwu Wang; cms_tjutcm@126.com

Received 31 October 2013; Revised 5 January 2014; Accepted 16 January 2014; Published 23 February 2014

Academic Editor: Lei Chen

Background. Minimization is a case allocation method for randomized controlled trials (RCT). Evidence suggests that the minimization method achieves balanced groups with respect to numbers and participant characteristics, and can incorporate more prognostic factors compared to other randomization methods. Although several automatic allocation systems exist (e.g., randoWeb, and MagMin), the minimization method is still difficult to implement, and RCTs seldom employ minimization. Therefore, we developed the minimization allocation controlled trials (MACT) system, a generic manageable minimization allocation system. *System Outline*. The MACT system implements minimization allocation by Web and email. It has a unified interface that manages trials, participants, and allocation. It simultaneously supports multitrials, multicenters, multigrouping, multiple prognostic factors, and multilevels. *Methods*. Unlike previous systems, MACT utilizes an optimized database that greatly improves manageability. *Simulations and Results*. MACT was assessed in a series of experiments and evaluations. Relative to simple randomization, minimization produces better balance among groups and similar unpredictability. *Applications*. MACT has been employed in two RCTs that lasted three years. During this period, MACT steadily and simultaneously satisfied the requirements of the trial. *Conclusions*. MACT is a manageable, easy-to-use case allocation system. Its outstanding features are attracting more RCTs to use the minimization allocation method.

1. Background

The blind randomized control trial (RCT) is commonly accepted as the gold standard research method for evaluating medical innovations [1]. When correctly performed, the RCT ensures that the same sort of participants receive intervention and control, thus eliminating selection and confounding biases [2]. When the study size is large enough, unbiased allocation of participants into either treatment groups or control groups maintains the balance of numbers and prognostic factors of participants among groups. This is essential to establish the internal validity of an RCT and ensure that the results are objective and scientific. Unbiased allocation allows RCTs to evaluate the effectiveness of medical innovations, such as a new surgical operation or a drug, in a sample population. When the number of participants is small, unbiased allocation is not appropriate, and alternative allocation methods are required.

Methods for allocation of trial participants can be divided into two categories: randomized allocation that includes simple, stratified, and blocked randomizations and dynamic allocation that includes biased coin, urn designs, and minimization techniques. Although permuted block methods are increasingly popular, simple randomization is more widely used than dynamic methods. In fact, a 2001 review indicated that only 4% of 150 RCTs employed minimization methods [3].

The simple randomization method allocates participants according to a pregenerated random number table, random number generator, or methods similar to a coin toss. This represents a completely unpredictable approach. However,

simple randomization may result in an imbalance of prognostic factors among groups when trial sizes are small [4].

Minimization, which was first proposed by Taves in 1974 [5], employs a deterministic algorithm to allocate participants into groups. This deterministic algorithm minimizes differences among groups with respect to prognostic factors associated with the entire trial. Literature indicates that minimization method achieves good group balance [4]. However, the nonrandom nature of the method may introduce selection bias, as it may be possible to predict which group the next subject will be enrolled in, provided the factor levels of the new subject are known [6]. To enhance the unpredictability of the minimization method, a random element has been introduced. A random element is a probability (p) value ranging from 0.5 to 1 [7]. After the initial minimization allocation, a participant is assigned to a group based on the probabilities for p. Other methods may be used to reduce predictability, such as excluding exact details of the algorithm in the protocol, collecting data that is not used as a prognostic factor at the time of randomization, or introducing site as a factor in a multisite trial.

Evidence suggests that minimization allocation outperforms simple randomization, resulting in less chance of an imbalance of prognostic factors and treatment factors among groups. As minimization is increasingly used to allocate participants in RCTs, these observations strengthen the credibility of trial results. However, the allocation of future patients to a trial is less predictable in simple randomization allocation compared to minimization [8]. Previously, a simulation study showed that the performance of simple randomization was similar to minimization when random element was applied [9].

Currently, there are few random allocation systems available. Kenjo et al. created an easily customized and multi-institutional minimization allocation system [10]. This system is based on the Practical Extraction and Report Language (PERL) for writing common gateway interface (CGI) script. This system balances prognostic factors among groups. However, details describing the system are limited. Cai et al. developed a generic minimization random allocation and blinding system coded with Microsoft Visual Basic and Active Server Pages (ASP) programming languages [11, 12]. System details, usage, and a portion of the code are available, but the design of the database is complex, resulting in difficulties associated with programming and maintenance. Furthermore, trial management is confusing, and system administration is difficult. Morice developed randoWeb, an online randomization tool for clinical trials [13]. This system provides simple, stratified, and dynamic randomization methods, but system management details are limited.

We sought to facilitate case allocation and RCT management through the development of a minimization allocation controlled trial (MACT) system. In this system, multicenter, multigrouping, multiple prognostic factors, and multilevel RCTs with simple randomization or minimization can be achieved. Benefits of this system include the use of an optimized common relational database based on just 4 tables, rather than an entity-attribute-value (EAV) model [14].

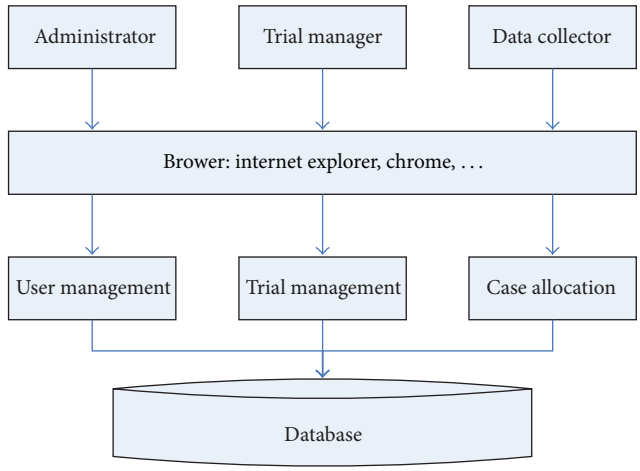


Figure 1: System architecture.

The common relational database provides a unified programming and management interface that makes MACT easy to generalize. The MACT codes are concise and easy to understand, which makes the final system operation, maintenance, and management very convenient. In theory, the maximum number of prognostic factors in MACT is limited by the capability of the hardware. In reality, the maximum number of prognostic factors is limited by the study design, as increasing numbers of prognostic factors make it difficult to control balance and predictability. In practice, the actual number of prognostic factors is much smaller than MACT's capability.

2. System Outline

2.1. System Architecture

2.1.1. Software and Hardware. MACT is designed to run on a Windows 2000 platform and microcomputers that are able to support the MACT system. In practice, our platform is a DELL SC440 Server with 1 GB RAM, 160 GB hard disk, and a Pentium dual-core CPU E2180 @ 2.0 GHz.

User and trial management employs the internal IIS server of Windows 2000. E-mail requires a SMTP server and the DBMS is SQL Server 7.0.

Users can access MACT through any internet browser application; tested browsers include Internet Explorer and Chrome.

2.1.2. System Architecture Diagram. A schematic of MACT system architecture is shown in Figure 1. All users, including administrators, trial managers, and data collectors, utilize browsers. They access certain modules to complete their respective user management, trial management, and participant allocation tasks (Figure 1).

2.2. User Management. MACT has three tiers of management: administrator, trial manager, and data collector. In addition, there are two other classes of users: registered users

Table 1: User permissions coding.

Coding	Permissions
128	Administrator
64	Trial manager
32	Data collector
16	Reserved
8	Reserved
4	Reserved
2	Banned user
1	Registered user

and banned users. Registered and banned users are unable to complete any operations in MACT; however, registered users can be changed into data collectors or trial managers, but banned users cannot. Preassigned permissions are encoded as shown in Table 1.

The administrator has the most privileges. The administrator has the ability to access all modules and manages users through user management modules. Administration tasks include assigning data collectors to trials; adjusting permissions (e.g., changing users permissions to data collector or trial manager); and banning offending users. New users can register by telephone, short message service (SMS; text), or email and administrators can adjust permissions.

2.3. Trial Management. Trial managers design studies, register participants, and assign prognostic factors, numbers of groups, and p. Initialization of new trials involves

(i) entering the name of the trial in the data collection interface,

(ii) setting the p to enhance the unpredictability of the minimization method,

(iii) adding the important prognostic factors and identifying the associated levels in each factor to determine inclusion and exclusion criteria,

(iv) assigning group names to treatment regimens (e.g., A, B, C, etc.).

Trial managers can view the total number of participants included in a trial, the number of participants recruited in each subcenter, and the number of participants in each group. Trial managers can communicate with data collectors via email.

2.4. Participant Allocation. Data collectors at each trial subcenter are responsible for recruiting participants. Data collectors enter participant information into MACT, and the system automatically allocates the participant to a group. Details of the allocation process are communicated to the data collectors and the trial manger via email. Data collectors are blinded to the specifics of the treatment regimens.

Patient privacy is protected by hardware and software firewalls. Access to patient data is restricted through appropriate MACT management procedures.

3. Methods

3.1. Database Design. Unlike previous systems, MACT employs a traditional relational database instead of an EAV database. EAV databases are often used in instances where the amounts of attributes, properties, or parameters that can be used to define an entity are potentially limitless, but sparse. EAV databases are especially applicable to RCTs, where they take into account different prognostic factors, levels, and allocation bias. Such differences result in heterogeneity. This heterogeneity is intensified if several RCTs exist within one system. The EAV design transforms the heterogeneity into one table that has three columns: entity, attribute, and value. Entity is the patient event and includes the patient ID; attribute or parameter is a foreign key into a table of attribute definitions; attribute definitions include an attribute ID, attribute name, description, data type, units of measurement, and input validation; value is the value of the attribute. The EAV database model has many advantages. First, it stores heterogeneous data through a unified logical structure that provides a convenient interface for programming. Second, it has a simple structure that is easily adaptable to accommodate multitrials. When new data arrives, it is appended directly to existing data without changing the structure of the table. Third, it efficiently utilizes storage space. However, the EAV model has two insurmountable shortcomings. First, the program has no ability to manipulate all the attributes within a whole row using one operation. Accessing a whole row in the EAV model requires many operations and multiple table joins. The process of querying an n-column row requires n SELECTs and $n - 1$ JOIN statements. In contrast, SQL data definition and query language used in most relational databases uses one SELECT statement. Second, EAV tables save storage space when the original table is sparse, but, when the occupancy rate is greater than 1/3, the EAV tables occupy a significantly higher amount of storage due to data redundancy.

In order to improve the EAV model, MACT created a common relational database with a rule column (RTWR) to store trial data. MACT's database is illustrated in Figure 2.

The following is an explanation of the database and its various components.

CENTER. Stores information of all users (administrator, trial managers, data collectors, registered users, and banned users), including username, password, email addresses, and permission levels. Operations include registering users, changing permissions, and removing users. The latter two operations can only be performed by an administrator.

TRIAL. Stores information on all trials, including trial manager's ID, allocation bias, number of prognostic factors, the pattern string of every prognostic factor, and the number of groups. The pattern string contains the name, cutoff value, and weight of prognostic factors. The rule column in TRIAL contains the pattern string. The rule column in TRIAL and RULEDATA constitute the RTWR structure. This structure enables MACT to store heterogeneous data in RULEDATA.

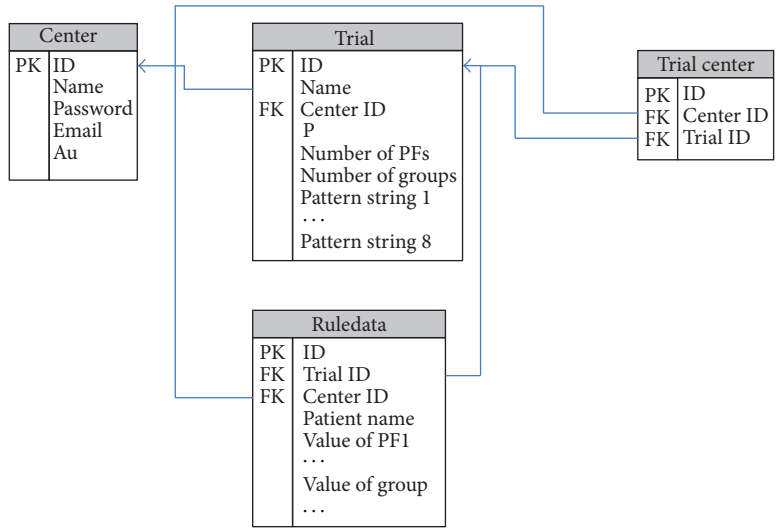


Figure 2: Database structure.

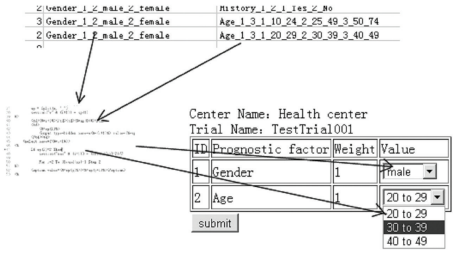


Figure 3: Rule translator.

RULEDATA. It stores all trial data, including trial ID, manager ID, participant names, and all values of prognostic factors. Data transformation uses a pattern string as follows.

(i) Information about prognostic factors can be acquired from TRIAL through trial ID.

(ii) A preprogrammed conversion module acquires the name, cutoff values, and weight of prognostic factors from TRIAL and shows them on screen.

For example, in Figure 3, the pattern string Age_1_3_1_20_29_2_30_39_3_40_49 is transformed: the name of the prognostic factor is Age; its weight is 1; it is divided into several levels, 20 to 29, 30 to 39, and 40 to 49. When data is needed, all of the information can be retrieved through the conversion module by one operation.

The common relational database has the ability to store heterogeneous data that can be retrieved by a single operation. However, the maximum number of prognostic factors is limited in this database, as the number of columns is unalterable; the maximum number of prognostic factors of all trials is equivalent to the number of columns in RULEDATA and storage space is wasted. As trials have different numbers of prognostic factors, there are empty columns in RULEDATA (Figure 3).

Table 2: The setup of multitrial simulations.

Number	Number of prognostic factors	Number of maximum levels	Number of groups	Allocation bias
1	2	3	2	0.9
2	3	4	3	0.8
3	6	5	3	0.7

3.2. Minimization Allocation. The minimization method involves

(i) allocation of the first participant into an arbitrary group with the probability 1/number of groups,

(ii) allocation of subsequent participants into group with probability p, such that $\arg\min_h G_h = \sum_{i=1}^{M} w_i D_{ij}$, where D_{ij} is defined as $D_{ij} = \mathrm{Var}(x_{1ij}, x_{2ij}, \ldots, x_{kij})$, where k is the number of groups, i is the number of prognostic factors, j is the number of levels, w is the weight of prognostic factors, p is the allocation bias, and x is the number of case

Because minimization allocation and p allocation are two independent procedures, the minimization allocation transforms into simple random allocation when $p = 0.5$. In this situation, MACT performs simple randomization.

4. Simulations and Results

Single and multitrial simulations were performed to ensure that MACT would achieve the desired results. The single-trial simulation included 300 participants where $i = 2$, $j_{\max} = 2$, $k = 2$, $p = 0.8$, where $j_{\max}$ represents the maximum value of prognostic factors. In the multitrial simulation, 3 RCTs were run simultaneously (Table 2). Balance of prognostic factors between intervention groups and the unpredictability of the

Table 3: The standard deviation of cases and prognostic factors among groups.

Number of cases	100		200		300	
Method	Minimization	Simple	Minimization	Simple	Minimization	Simple
Number of cases	***1.41***	7.07	***1.41***	7.07	***0.00***	1.41
PF1						
Level 1	***0 .705***	4.95	**0.705**	2.12	***0.00***	1.41
Level 2	2.12	2.12	***0.705***	4.95	***0.00***	2.83
PF2						
Level 1	2.12	2.12	***2.12***	9.19	4.95	4.95
Level 2	4.95	***0.705***	***4.95***	6.36	8.49	8.49
Level 3	***4.24***	5.56	4.24	4.24	***3.54***	4.95

Table 4: The unpredictability of a single trial.

Cases	100		200		300	
Method	Minimization	Simple	Minimization	Simple	Minimization	Simple
Correct	29	21	77	75	131	120
Incorrect	21	29	73	75	119	130
Correct%	42.0	58.0	51.3	50.0	52.4	48.0
SD	5.66	5.66	2.83	***0.00***	8.49	***7.07***

minimization and simple randomization allocation methods were compared. Unpredictability was evaluated using Support Vector Machine (SVM) [15], a technique for bioinformatics classification [16–18]. In this task, SVM considered the allocations of the first 50 participants to predict allocation of the 51st participant. Subsequently, the 51st participant was added into the simulation and the above process was repeated until there were no new participants. The total numbers of correctly and incorrectly predicted results were recorded. The unpredictability is evaluated by the standard deviation of the number of correct and incorrect predicted allocation results. When the standard deviation is 0, the allocation is perfectly unpredictable.

4.1. Single Trial Simulation. Table 3 shows the balance of prognostic factors (PF) between intervention groups achieved by minimization and simple randomization in the single trial simulation. The values are standard deviations of participants among groups according to levels of prognostic factors. Smaller values indicate a smaller difference between groups; bold and italicized values indicate better performance (Tables 3–6). When 100 participants were allocated, minimization outperformed simple randomization in most cases. When 200 participants were allocated, minimization was not inferior to simple randomization in all cases and outperformed simple randomization in most cases. When 300 participants were allocated, the performance of minimization and simple randomization was similar, although minimization tended to outperform simple randomization.

We demonstrated the predictability of minimization and simple randomization in a single trial simulation in Table 4. The total numbers of correctly and incorrectly predicted results were also recorded (Table 4). Note that more incorrect predictions do not result in higher unpredictability because smart forecasters can use the opposite result of predictive algorithms. For this reason, we took the standard deviation from the number of correct and incorrect predictions as the measurement of unpredictability. When the first 100 participants were allocated, the unpredictability of minimization and simple randomization was the same in our simulation as indicated by the standard deviations. As the number of participants increased to 200 and 300, the unpredictability of minimization was greater than simple randomization.

4.2. Multitrial Simulations. The multitrial simulations confirmed that as the number of participants allocated increased, minimization achieved a better balance of prognostic factors among groups, while simple randomization had better unpredictability (Tables 5 and 6). Only the result of the largest trial, which is number 3 in Table 2, is recorded in Tables 5 and 6, because all the results are similar to the single trial.

5. Applications

MACT was initially employed for RCT 2006BAI08B02-01, which had four prognostic factors, including gender (2 levels), age (4 levels), primary disease (4 levels), and cardiac function (3 levels). In total, 340 participants were allocated into two groups and the allocation bias was 0.8. The MACT allocated participants are shown in Table 7. MACT achieved an ideal balance of prognostic factors between the two groups and 126 correct and 164 incorrect predictions.

Subsequently, MACT was employed for RCT 2008BAI53B04, which had three prognostic factors including gender (2 levels), age (7 levels), and medical score (10 levels). In total, 370 cases were allocated into two groups and the allocation bias was 0.8. There were 154 correct and 166 incorrect predictions (Table 8).

Table 5: The standard deviation of cases and prognostic factors among groups.

Number of Cases	300		400		500	
Method	Minimization	Simple	Minimization	Simple	Minimization	Simple
Number of cases	***0.816***	1.41	***0.816***	4.97	***0.816***	1.41
PF1						
Level 1	***0.816***	2.16	***0.577***	3.70	***0.500***	4.92
Level 2	***0.00***	1.83	***0.577***	2.65	***0.957***	3.59
Level 3	***0.500***	3.20	***0.500***	3.30	***1.41***	4.97
Level 4	***1.15***	2.58	***0.577***	2.38	***0.577***	2.65
Level 5	***0.957***	1.71	***0.500***	3.77	***0.00***	5.48
PF2						
Level 1	***0.957***	4.27	***0.816***	3.65	***0.577***	3.70
Level 2	***0.816***	3.92	***0.500***	6.99	***0.957***	11.9
Level 3	***0.957***	4.86	***0.500***	4.99	***0.500***	8.62
PF3						
Level 1	***0.816***	5.23	***0.00***	2.94	***0.500***	3.59
Level 2	***0.816***	6.32	***0.816***	6.38	***0.500***	9.57
PF4						
Level 1	***0.500***	2.99	***0.957***	4.50	***0.577***	3.70
Level 2	***0.500***	4.03	***0.957***	4.19	***0.816***	5.42
Level 3	***0.500***	2.87	***0.957***	2.22	***0.577***	4.43
Level 4	***0.957***	4.65	***0.00***	5.48	***0.577***	4.43
Level 5	***0.816***	2.94	***0.500***	2.63	***0.577***	2.52
PF5						
Level 1	***0.500***	5.50	***0.577***	7.54	***0.500***	6.80
Level 2	***0.957***	5.12	***1.00***	5.26	***0.816***	8.29
Level 3	***0.577***	2.38	***0.500***	3.70	***0.500***	5.85
Level 4	***0.816***	4.16	***0.710***	4.79	***0.816***	4.00
PF6						
Level 1	***0.577***	5.07	***2.06***	6.18	***4.11***	8.62
Level 2	***1.26***	2.36	***1.91***	4.43	***1.71***	5.32
Level 3	***1.71***	2.50	***2.50***	5.74	***4.12***	6.95

Table 6: The unpredictability of multitrials.

Cases	300		400		500	
Method	Minimization	Simple	Minimization	Simple	Minimization	Simple
Correct	42	63	67	89	86	113
Incorrect	208	187	283	261	364	337
Correct %	16.8	25.2	19.1	25.4	19.1	25.1
SD	117	***87.7***	153	***122***	197	***158***

In practical applications, MACT achieved a good balance of prognostic factors among groups. This greatly improved the internal validity of the RCTs and yielded more robust conclusions.

6. Conclusions

MACT is an easy-to-manage allocation system. Currently, 11 hospitals in northern China are registered as subcenters. The trial managers and data collectors in these subcenters became familiar with the system within one hour's training. Trials are managed and participants are allocated without further programming. MACT has excellent stability. So far, MACT has run continuously for three years. With the exception of regular hardware maintenance, the system has never failed. As an easy-to-expand, easy-to-manage, and stable system, MACT facilitates the use of the minimization method in the practice of clinical trials.

Conflict of Interests

The authors declared that they have no conflict of interests regarding this work.

Table 7: The allocation results of project 2006BAI08B02-01.

PF	Level	Group 1	Group 2
1	1	91	94
	2	78	77
2	1	10	9
	2	25	27
	3	45	46
	4	89	89
3	1	137	136
	2	7	9
	3	7	7
	4	18	19
4	1	21	20
	2	96	97
	3	52	54
Total		169	171

Table 8: The allocation results of project 2008BAI53B04.

PF	Level	Group 1	Group 2
1	1	121	122
	2	64	63
2	1	10	10
	2	48	46
	3	34	37
	4	33	34
	5	37	39
	6	18	14
	7	5	5
3	1	0	0
	2	3	3
	3	17	17
	4	78	78
	5	36	36
	6	23	24
	7	12	12
	8	8	7
	9	1	2
	10	7	6
Total		185	185

Authors' Contribution

Yan Cui designed database, programmed MACT, and wrote this paper. Huaien Bu was responsible for routine maintenance of MACT. Hongwu Wang provided mathematical models and methods of this system. Shizhong Liao gave critical revision of this paper and suggestion on database design.

Acknowledgment

This work is supported by the National Basic Research Program of China (973 Program, Grant 2011CB505406).

References

[1] S. L. Silverman, "From randomized controlled trials to observational studies," *American Journal of Medicine*, vol. 122, no. 2, pp. 114–120, 2009.

[2] T. J. Kaptchuk, "The double-blind, randomized, placebo-controlled trial: gold standard or golden calf?" *Journal of Clinical Epidemiology*, vol. 54, no. 6, pp. 541–549, 2001.

[3] C. J. Weir and K. R. Lees, "Comparison of stratification and adaptive methods for treatment allocation in an acute stroke clinical trial," *Statistics in Medicine*, vol. 22, no. 5, pp. 705–726, 2003.

[4] J. M. Lachin, "Properties of simple randomization in clinical trials," *Controlled Clinical Trials*, vol. 9, no. 4, pp. 312–326, 1988.

[5] D. R. Taves, "Minimization: a new method of assigning patients to treatment and control groups," *Clinical Pharmacology and Therapeutics*, vol. 15, no. 5, pp. 443–453, 1974.

[6] N. W. Scott, G. C. McPherson, C. R. Ramsay, and M. K. Campbell, "The method of minimization for allocation to clinical trials: a review," *Controlled Clinical Trials*, vol. 23, no. 6, pp. 662–674, 2002.

[7] S. J. Pocock and R. Simon, "Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial," *Biometrics*, vol. 31, no. 1, pp. 103–115, 1975.

[8] A. Hagino, C. Hamada, I. Yoshimura, Y. Ohashi, J. Sakamoto, and H. Nakazato, "Statistical comparison of random allocation methods in cancer clinical trials," *Controlled Clinical Trials*, vol. 25, no. 6, pp. 572–584, 2004.

[9] L. Xiao, P. W. Lavori, S. R. Wilson, and J. Ma, "Comparison of dynamic block randomization and minimization in randomized trials: a simulation study," *Clinical Trials*, vol. 8, no. 1, pp. 59–69, 2011.

[10] Y. Kenjo, Y. Antoku, K. Akazawa, E. Hanada, N. Kinukawa, and Y. Nose, "An easily customized, random allocation system using the minimization method for multi-Institutional clinical trials," *Computer Methods and Programs in Biomedicine*, vol. 62, no. 1, pp. 45–49, 2000.

[11] H. W. Cai, H. L. Xia, D. Z. Xu, D. H. Gao, and Y. Yan, "A generic minimization random allocation and blinding system on web," *Journal of Biomedical Informatics*, vol. 39, no. 6, pp. 706–719, 2006.

[12] H. W. Cai, J. L. Xia, D. H. Gao, and X. M. Cao, "Implementation and experience of a web-based allocation system with Pocock and Simon's minimization methods," *Contemporary Clinical Trials*, vol. 31, no. 6, pp. 510–513, 2010.

[13] V. Morice, "RandoWeb, an online randomization tool for clinical trials," *Computer Methods and Programs in Biomedicine*, vol. 107, no. 2, pp. 308–314, 2012.

[14] H. Y. Kim and H. A. Park, "Development and evaluation of data entry templates based on the entity-attribute-value model for clinical decision support of pressure ulcer wound management," *International Journal of Medical Informatics*, vol. 81, no. 7, pp. 485–492, 2012.

[15] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[16] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[17] M. Abbasi, M. A. Abduli, B. Omidvar, and A. Baghvand, "Forecasting municipal solid waste generation by hybrid support

vector machine and partial least square model," *International Journal of Environmental Research*, vol. 7, no. 1, pp. 27–38, 2013.

[18] H. Wang and G. Huang, "Application of support vector machine in cancer diagnosis," *Medical Oncology*, vol. 28, supplement 1, pp. S613–S618, 2011.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 390212, 5 pages
http://dx.doi.org/10.1155/2013/390212

Research Article

Quantitatively Plotting the Human Face for Multivariate Data Visualisation Illustrated by Health Assessments Using Laboratory Parameters

Wang Hongwei[1] and Liu Hui[2]

[1] *Department of Computer Science, Dalian Medical University, Dalian 116044, China*
[2] *College of Medical Laboratory, Dalian Medical University, Dalian 116044, China*

Correspondence should be addressed to Liu Hui; liuhui60@sina.com

Received 18 October 2013; Revised 28 November 2013; Accepted 9 December 2013

Academic Editor: Lei Chen

Objective. The purpose of this study was to describe a new data visualisation system by plotting the human face to observe the comprehensive effects of multivariate data. *Methods*. The Graphics Device Interface (GDI+) in the Visual Studio.NET development platform was used to write a program that enables facial image parameters to be recorded, such as cropping and rotation, and can generate a new facial image according to Z values from sets of normal data ($Z > 3$ was still counted as 3). The measured clinical laboratory parameters related to health status were obtained from senile people, glaucoma patients, and fatty liver patients to illustrate the facial data visualisation system. *Results*. When the eyes, nose, and mouth were rotated around their own axes at the same angle, the deformation effects were similar. The deformation effects for any abnormality of the eyes, nose, or mouth should be slightly higher than those for simultaneous abnormalities. The facial changes in the populations with different health statuses were significant compared with a control population. *Conclusions*. The comprehensive effects of multivariate may not equal the sum of each variable. The 3Z facial data visualisation system can effectively distinguish people with poor health status from healthy people.

1. Introduction

Data visualisation is the study of visual representation of data to communicate information clearly and effectively through graphical means [1–3]. The medical sciences have a uniquely intertwined relationship with bioinformatics. The rapidly expanding field of biology creates enormous challenges to enable researchers to gain insights from large and highly complex data sets. Although researchers and practitioners often create patterns that can be visually identified, such as charts, graphs, and interactive displays, when solving a large range of problems, there are no definite accepted methods to identify these complex relationships [4–6].

Traditional data visualisation systems are mostly based on mathematical models, but complex bioinformatic correlations may not follow previously known statistical rules. Therefore, it is essential to explore methods for data visualisation that do not completely rely on mathematical models. We established the 3Z facial data visualisation system based on changes in human facial features. Certain specific bioinformatics rules for correlations may be elucidated with the use of the data visualisation system.

Chernoff first developed the idea of using human facial characteristics as a means to visualise data [7, 8]. The idea behind using faces is that humans easily recognise faces and notice small changes without difficulty. Chernoff faces display multivariate data in the shape of a human face. However, the effectiveness of this form of visualisation is still open to speculation. When assigning several variables to various facial features when drawing Chernoff faces, different drawers may make different choices. Accordingly, different grouping results may be obtained. Therefore, Chernoff faces may not work well to detect the comprehensive effects of multivariate data, and updates of facial data visualisation are required for this purpose.

Displaying data quantitatively is required for data visualisation systems. Thus, it is important to quantitatively plot

the human face based on multivariate data. The correct relations of multivariate data should not be obtained without quantitative data display. We developed a program that enables facial image parameters to be recorded quantitatively according to Z values from sets of normal data and can generate a new facial image for observing comprehensive effects of multivariate data. Thus, a new analytic platform may elucidate complex bioinformatic correlations.

From a public point of view, the health status of an individual is influenced by social, psychological, and biological factors [9–11]. Thus, any quantitative evaluation of health status is a challenging task. Deteriorations in health status for any reason will induce changes in somatic biological factors, with a consequently increased risk of disease. This process is accompanied with changes in clinical laboratory parameters [12–14]. The present study was performed to elucidate the parameters of a facial data visualisation system based on changes in the facial features of individuals and their role in health assessment by assessing correlations with laboratory parameters.

2. Methods

2.1. Construction of the Facial Data Visualisation System. The data visualisation system was constructed on the basis of human facial images, including the eyes, nose, and mouth. These three features were rotated clockwise around their own axes, with the original value of 0° and the terminating value of 45°. Data from different groups were normalised as 0–45, and the larger the value, the larger the angle of rotation and the more significant the facial changes. The present study randomly inputted the three groups of data from both eyes, the nose, and the mouth to form the data visualisation system.

We used object-oriented technology to develop a graphics processing program with the technical support of the Microsoft.NET Framework and Graphics Device Interface (GDI+) of the Visual Studio.NET development platform. The program enabled us to develop parameters that could process graphics or images, such as cropping, rotation, and the generation of a new image, to achieve data visualisation.

2.2. Laboratory Parameters for Health Assessment. The related experimental indicators of organ function, blood lipid levels, and stress levels were divided into three groups to evaluate the health status from three dimensions to reflect the impact of different factors on health. Evaluation indicators of organ function included albumin (Alb, 41 ± 3.5 g/L). Evaluation indicators of blood lipids included cholesterol (Chol, 5.0 ± 0.9 mmol/L), and evaluation indicators of stress included the neutrophil count (Neut, $3.5 \pm 1.2 \times 10^9$/L). Laboratory indices were measured with an automatic analyser in the clinical laboratory of our university hospital using standard commercial reagent kits.

2.3. Standardisation of Measurement Values. Based on the reference ranges, we obtained the mean values and the standard deviation (SD) for the normally distributed data. Z-values were calculated according to the following formula for Neut and Chol, where X was the measured value and Mean was the mean value:

$$Z = \frac{(X - \text{Mean})}{\text{SD}}. \quad (1)$$

For Alb, a lower measured value represented a worse health status; therefore, the Z values were calculated according to the following formula:

$$Z = \frac{(\text{Mean} - X)}{\text{SD}}. \quad (2)$$

Thus, the measured values for the parameters as mentioned were comparable after transformation, where $Z < 0$ indicated that it was normal. The larger the value is, the poorer the health status is.

2.4. The Angle Transformation. $Z > 3$ indicated that the health status was very poor or the patient was ill. Therefore, emphasis was placed on Z-score changes between 0 and 3 in assessing health status. Therefore, $Z < 0$ was counted as 0. $Z > 3$ was counted as 3. The transformed value after the multiplication of Z by 15 was between 0 and 45, which was suitable for the facial data visualisation system. The measured values for different observed subjects were normalised (Z value transformation), and the transformation was made for the angle 0–45°($Z \times 15$).

The transformed values of angles were input into the facial data visualisation system, and thus, the correlations between the values from different dimensions and the possible integrative effects were observed.

2.5. Subjects. Fifteen patients with glaucoma requiring ophthalmologic surgery (57.8 ± 7.2 years old, 7 male and 8 female) were randomly selected as the glaucoma group.

Forty individuals with nonalcoholic fatty liver as indicated by ultrasonography (31.4 ± 4.3 years old, 20 male and 20 female) were selected as the fatty liver group. Subjects in the fatty liver group who met the following criteria were excluded from the study: (1) those suffering from other liver diseases, such as viral hepatitis; (2) subjects who had a dependence on alcohol; and (3) individuals aged >40 years old.

Forty elderly subjects without organic disease as detected by imaging examinations (82.5 ± 2.4 years old, 20 male and 20 female) were randomly selected as the elderly group.

In the control group, there were 40 individuals (20 male and 20 female) aged between 20 and 30 years old without organic disease as detected by imaging examinations, and their average age was 26.7 years old.

2.6. Statistical Analysis. The null hypothesis was that means in different populations were equal. One-way ANOVA was used to analyse the differences of each indicator among the various populations. The degrees of freedom were 3 and 156, respectively, in our sample; $F_{0.05(3,150)}$ was 2.66. A difference was considered to be statistically significant when the P value was less than 0.05. Statistical analyses were performed with SPSS statistical analysis software for Windows (SPSS, Chicago, IL, USA).

Table 1: Measured data and angular transformation for the populations with different health statuses.

Group	Measured value (mean ± SD)			Angular transformation (median)		
	Neut	Alb	Chol	Neut	Alb	Chol
Glaucoma	5.09 ± 1.86	39.11 ± 4.57	6.11 ± 1.51	11.0	4.0	18.5
Fatty liver	3.73 ± 1.35	40.29 ± 3.06	5.86 ± 1.13	0.1	1.1	12.9
Elderly	3.67 ± 1.32	38.62 ± 2.28	5.61 ± 1.09	0.0	9.5	7.9
Control	3.33 ± 1.04	40.99 ± 2.48	5.31 ± 1.39	0.0	1.8	0.5
F value	6.513	5.116	2.102	—	—	—
P value	<0.0001	0.002	0.103	—	—	—

$F_{0.05(3,150)} = 2.66$.

3. Results

The intentionally preset data (normal value; the organs were rotated 10°, the eyes were rotated 30°, the nose was rotated 30°, and the mouth was rotated 30°) were input into the facial data visualisation system. The consequent facial changes are shown in Figure 1. The deformation extent for (b) was relatively mild, whereas the deformation extents for (c)–(e) were relatively major, and the deformation extents for (c), (d), and (e) were identical.

The measured data for the populations with different health statuses are shown in Table 1. The measured values were converted into the angle transformation using (1), (2), and the formula described at the angle transformation section in Method. The median angle values in different groups are also shown in Table 1 and were input into the facial data visualisation system (rotation of eye depends on data of Neut; that of nose depends on data of Alb, and that of mouth depends on data of Chol). The resulting facial changes are shown in Figure 2. The facial changes in the populations with different health statuses were significant compared with those in the control populations.

4. Discussion

To assess the comparability of change in measured values, the data transformation into Z values was conducted using the average value and standard deviation. Because the maximal Z value preset was 3 in our system, the data visualisation system was defined as a 3Z facial data visualisation system. The positions of human facial features (eyes, nose, and mouth) constitute the major facial characteristics, and the main mode of recognition is in the brain. Thus, observers are sensitive to facial deformation. The present study rotated the eyes, nose, and mouth around their own respective axes, and significant deformation effects should be found when $Z = 2$ according to statistical theory. The observations ($Z = 2$) made in Figures 1(c)–1(e) [(c), eyes rotated 30°; (d), nose rotated 30°; and (e), mouth rotated 30°] revealed that the deformation effects from the rotating angles of the eyes, nose, and mouth on the face were similar, which indicated that the rotation of these features can be used to observe the comprehensive effects of multivariate data.

As seen in Figure 1, the sum of the rotations for eyes, nose, and mouth was 30° when they were rotated 10° separately (Figure 1(b)), and the deformation effects should be identical to those shown in Figures 1(c), 1(d), or 1(e). However, the results reveal that the deformation extent in Figure 1(b) was significantly lower than those shown in Figures 1(c), 1(d), or 1(e), which indicates that the deformation effects for any abnormality in the eyes, nose, or mouth should be higher than that for small simultaneous abnormalities. Therefore, the display of data with the use of facial characteristics may unravel new data patterns. Concerning the complex objects affected by multiple factors, the influence on the subject from significant changes in a certain factor may be higher than that from small changes in several factors; therefore, the facial data visualisation system may be more suitable for the analysis of certain complex systems.

Changes in health status usually do not accompany specific clinical signs. Routinely measured clinical laboratory parameters may be the only useful information for the assessment of healthy status of individuals [14–16]. However, the diagnostic thresholds usually provided are used for diagnosing diseases rather than assessing healthy status. Our results indicate that the P values were all less than or near 0.05 for the 3 indicators for the populations with different health statuses, indicating that these 3 indicators can be used for health assessment in distinguishing people with good and poor health. The problem was that we could not define the degree of poor health status using these 3 indicators. The health status of individuals may be determined by the quantity of the dimension with the poorest indication for health or the average of different dimensions. Therefore, the objective and accurate evaluation of health status becomes difficult.

As the experimental data in this study were represented by values, they were suitable for data visualisation. The facial data visualisation system may be more suitable for displaying and analysing health status parameters. We selected patients with glaucoma as a representative stress population, patients with fatty liver as a representative overnutrition population and elderly subjects as a representative population with functional insufficiency, because the health status of individuals is determined by multiple factors. The present study used the parameters of Neut, Chol, and Alb to represent the stress-response level, blood lipid level, and organ function level of the subjects, respectively [15–17], in three populations with different health statuses, compared with normal controls. Any item from these three indices among the populations

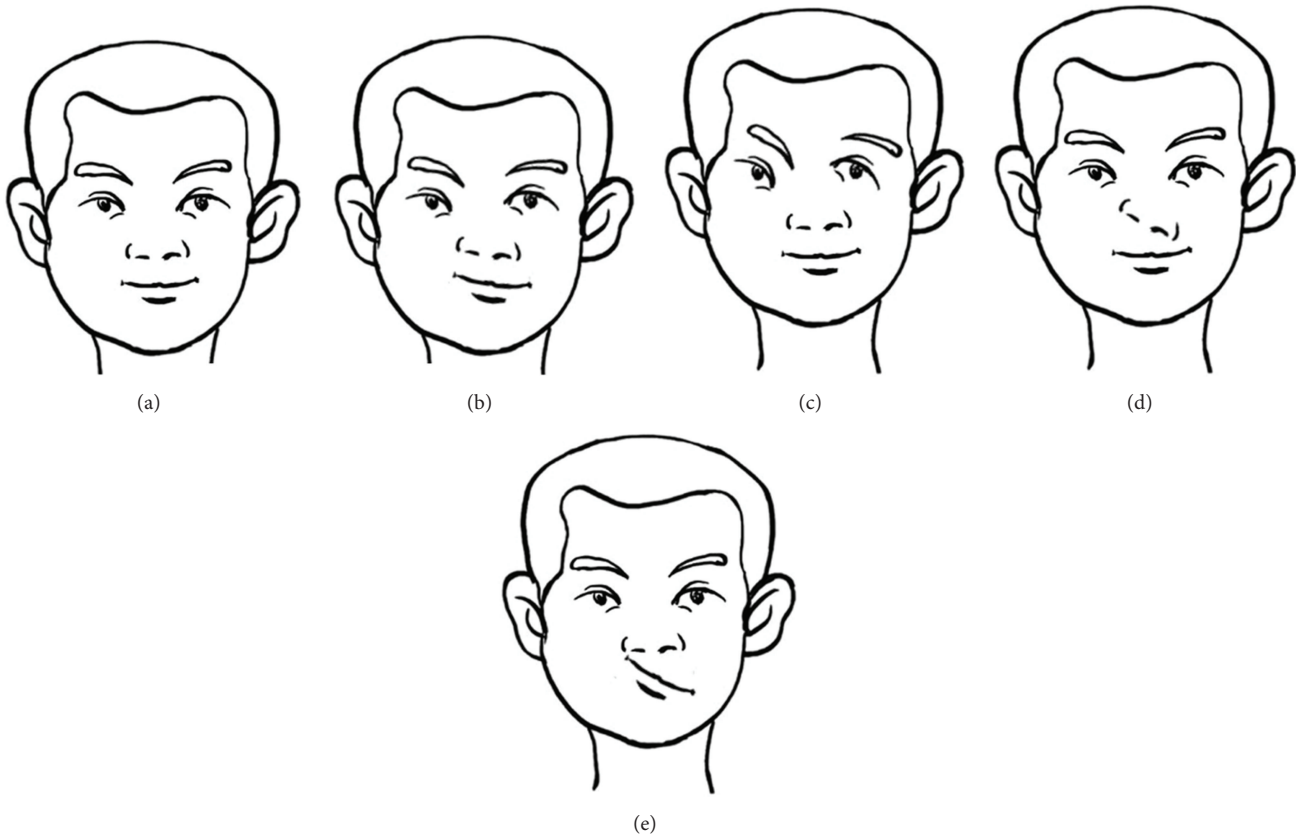


Figure 1: Effectiveness of the data display using the facial data visualisation system. (a) Normal value (eye = 0; nose = 0; mouth = 0); (b) facial features were rotated 10° (eye = 10; nose = 10; mouth = 10); (c) the eyes were rotated 30° (eye = 30; nose = 0; mouth = 0); (d) the nose was rotated 30° (eye = 0; nose = 30; mouth = 0); (e) the mouth was rotated 30° (eye = 0; nose = 0; mouth = 30).

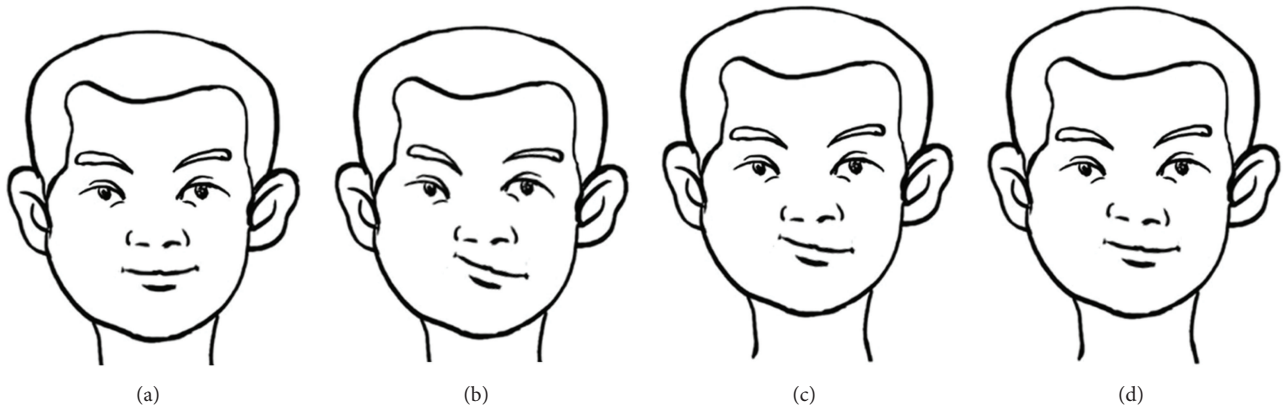


Figure 2: Effectiveness of the display of measured values for the populations with different health statuses using the facial data visualisation system. (a) Normal control; (b) glaucoma group; (c) fatty liver group; (d) elderly people.

may be different from that of normal population, although the degrees were not the same. The measured values for the populations as mentioned above were displayed by using the facial data visualisation system. The results indicated that the facial changes in the control population were not significant, whereas the facial changes in the populations with different health statuses were significant (Figure 2). The facial data visualisation system could effectively distinguish people with a poor health status from normal people, which indicates that this system may be a valid tool for the analysis of complex systems.

This facial data visualisation system will primarily be used for individual evaluation, rather than population evaluation. Therefore, the relatively small sample sizes in our study were acceptable for obtaining features of different health statuses. The limitation of our facial data visualisation system was that

the system has only three rotational variables and was not suitable for displaying more parameters. We suggest selecting three factors that explain most of the variance observed in a much larger number of manifest variables with the factor analysis method first and then, using our facial data visualisation system, displaying these three factors. Although the factor analysis technique could solve the above problem, the development of a facial data visualisation system with more variables is still an important future research direction.

5. Conclusions

In this work, we explore the way for data visualization that does not completely rely on mathematical models. The facial data visualization system based on quantitative changes in human facial features has been established. A deeper understanding of multivariate data could be obtained by plotting facial image system through intuitive experience. Certain specific bioinformatics rules for correlations could also be unraveled with using this data visualization system. The facial data visualization system may effectively distinguish people with a poor health status.

Conflict of Interests

The authors declare that they have no conflict of interests.

References

[1] S. Durinck, J. Bullard, P. T. Spellman, and S. Dudoit, "GenomeGraphs: integrated genomic data visualization with R," *BMC Bioinformatics*, vol. 10, article 2, 2009.

[2] A. P. Francisco, C. Vaz, P. T. Monteiro, J. Melo-Cristino, M. Ramirez, and J. A. Carrio, "PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods," *BMC Bioinformatics*, vol. 13, article 87, 2012.

[3] F. Jourdan, L. Cottret, L. Huc et al., "Use of reconstituted metabolic networks to assist in metabolomic data visualization and mining," *Metabolomics*, vol. 6, no. 2, pp. 312–321, 2010.

[4] J. Kennedy and J. Roerdink, "Highlights of the 1st IEEE symposium on biological data visualization," *BMC Bioinformatics*, vol. 13, Supplement 8, article S1, 2012.

[5] C. W. Bartlett, S. Y. Cheong, L. Hou et al., "An eQTL biological data visualization challenge and approaches from the visualization community," *BMC Bioinformatics*, vol. 13, supplement 8, article S8, 2012.

[6] J. H. Ostroff and D. C. Trost, "MDV: a multivariate data visualization tool for clinical laboratory data and other time-varying continuous measurements," *AMIA Annual Symposium Proceedings*, vol. 1068, 2005.

[7] H. Chernoff, "The use of faces to represent points in K-dimensional space graphically," *Journal of the American Statistical Association*, vol. 68, no. 342, pp. 361–368.

[8] B. Flury and H. Riedwyl, "Graphical representation of multivariate data by means of asymmetrical faces," *Journal of the American Statistical Association*, vol. 76, no. 376, pp. 757–765, 1981.

[9] Ó. Kristjánsdóttir, A. M. Unruh, L. McAlpine, and P. J. McGrath, "A systematic review of cross-cultural comparison studies of child, parent, and health professional outcomes associated with pediatric medical procedures," *Journal of Pain*, vol. 13, no. 3, pp. 207–219, 2012.

[10] A. Chalmers, S. Harrison, K. Mollison, N. Molloy, and K. Gray, "Establishing sensory-based approaches in mental health inpatient care: a multidisciplinary approach," *Australasian Psychiatry*, vol. 20, no. 1, pp. 35–39, 2012.

[11] R. Lopez and B. Goldoftas, "The urban elderly in the United States: health status and the environment," *Reviews on Environmental Health*, vol. 24, no. 1, pp. 47–57, 2009.

[12] H. Liu, G. Wang, G. Luan, and Q. Liu, "Effects of sleep and sleep deprivation on blood cell count and hemostasis parameters in healthy humans," *Journal of Thrombosis and Thrombolysis*, vol. 28, no. 1, pp. 46–49, 2009.

[13] W.-F. Teng, W.-M. Sun, L.-F. Shi, D.-D. Hou, and H. Liu, "Effects of restraint stress on iron, zinc, calcium, and magnesium whole blood levels in mice," *Biological Trace Element Research*, vol. 121, no. 3, pp. 243–248, 2008.

[14] H. Liu, Y. Wang, X. Qi, and H. Yuan, "Serum glucose- and C-reactive protein-based assessment of stress status in a healthy population," *Clinical Laboratory*, vol. 56, no. 5-6, pp. 227–230, 2010.

[15] W. Hongwei, Z. Xinyu, L. Guihong, L. Xiliang, and L. Hui, "Nonspecific biochemical changes under different health statuses and a quantitative model based on biological markers to evaluate systemic function in humans," *Clinical Laboratory*, vol. 56, no. 5-6, pp. 223–225, 2010.

[16] L. Hui, L. Shijun, Z. Xinyu, W. Yuai, and X. Xiaoting, "Objective assessment of stress levels and health status using routinely measured clinical laboratory parameters as biomarkers," *Biomarkers*, vol. 1, no. 6, pp. 525–529, 2011.

[17] G. Lippi, G. Targher, M. Franchini, and G. C. Guidi, "Biochemical correlates of lipoprotein(a) in a general adult population. Possible implications for cardiovascular risk assessment," *Journal of Thrombosis and Thrombolysis*, vol. 27, no. 1, pp. 44–47, 2009.