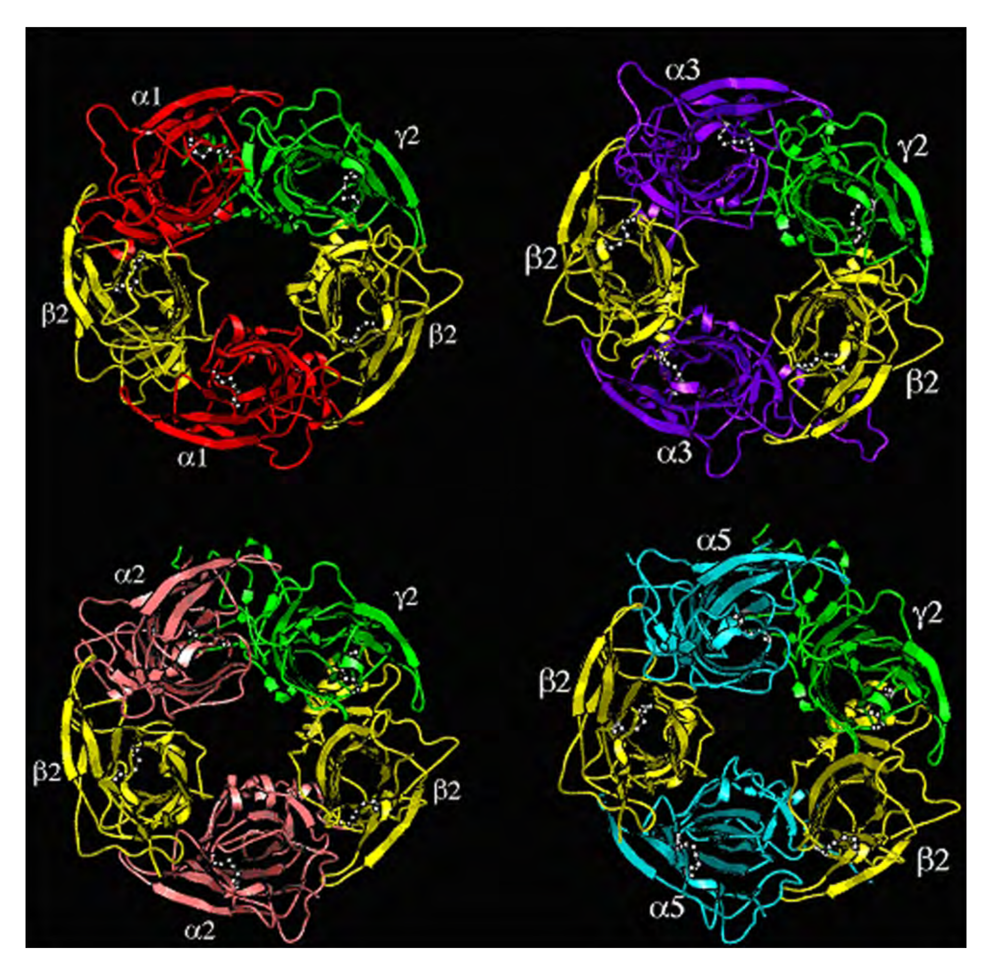
Scientific
Research
Publishing

# JBiSE

ISSN: 1937-6871

Vol. 1 No. 1 May 2008

## Journal of Biomedical Science & Engineering



Editor-in-Chief
Kuo-Chen Chou
Huai-Bei Zhou

www. SRPublishing. org/journal/jbise

# Journal Editorial Board

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

Scientific Research Publishing

# Sequence based prediction of relative solvent accessibility using two-stage support vector regression with confidence values

Ke Chen[1], Michal Kurgan[1] & Lukasz Kurgan*[1]

[1]Department of Electrical and Computer Engineering, University of Alberta, T6G 2V4, Edmonton, CANADA. * Correspondence should be addressed to Lukasz Kurgan (lkurgan@ece.ualberta.ca).

## ABSTRACT

**Predicted relative solvent accessibility (RSA) provides useful information for prediction of binding sites and reconstruction of the 3D-structure based on a protein sequence. Recent years observed development of several RSA prediction methods including those that generate real values and those that predict discrete states (buried vs. exposed). We propose a novel method for real value prediction that aims at minimizing the prediction error when compared with six existing methods. The proposed method is based on a two-stage Support Vector Regression (SVR) predictor. The improved prediction quality is a result of the developed composite sequence representation, which includes a custom-selected subset of features from the PSI-BLAST profile, secondary structure predicted with PSI-PRED, and binary code that indicates position of a given residue with respect to sequence termini. Cross validation tests on a benchmark dataset show that our method achieves 14.3 mean absolute error and 0.68 correlation. We also propose a confidence value that is associated with each predicted RSA values. The confidence is computed based on the difference in predictions from the two-stage SVR and a second two-stage Linear Regression (LR) predictor. The confidence values can be used to indicate the quality of the output RSA predictions.**

**Keywords: Relative solvent accessibility; Support vector regression; PSI-BLAST; PSI-PRED; Secondary protein structure**

## 1. INTRODUCTION

The knowledge of three dimensional protein structure plays the key role in understanding protein's function. Computational prediction of the tertiary protein structure is one of the central topics in structural biology due to the large and exponentially growing gap between the number of known protein sequences and the number of known structures. Despite several decades of extensive research in tertiary structure prediction, this task is still a big challenge, especially for sequences that do not have a significant sequence similarity with known structures [1]. As a result, the predictions of the solvent accessibility [2] and the secondary structure [3] are addressed as an intermediate step towards the prediction of the tertiary structure. The relative solvent accessibility (RSA) reflects the degree to which a residue interacts with the solvent molecules. Since protein-protein and protein-ligand interactions occur at the protein surface, only the residues that have a large surface area exposed to the solvent can possibly bind to the ligands and other proteins. As a result, prediction of solvent accessibility provides useful information for prediction of binding sites [4] and is vitally important for understanding the binding mechanism of proteins [5]. Chan and Dill pointed that the burial of core residues is the driving force in protein folding, which suggests that knowledge of localization of individual residues (surface vs. buried) provides useful information to reconstruct the 3D-structure of proteins [6-8].

The existing solvent accessibility prediction methods use the protein sequence, which is converted into a fixed-size feature-based representation, as an input to predict the RSA for each of the residues. These methods can be divided into two main groups:

– *Real valued* predictors predict RSA value (the definition is given in the Materials section). The representative existing methods are based on linear regression [9], neural network based regression [11], neural networks [12], support vector regression [10, 13, 15], and look up table [14]. In Ahmad's study, binary coding of the sequence was taken as the input features [12], while all other studies used the evolutionary information in the form of the PSSM profile derived with PSI-BLAST as the input features [9-11, 13-15].

– *discrete valued* predictors classify each residue into a predefined set classes. The classes are usually defined based on a threshold and include buried, intermediate, and exposed classes (in most cases the predictions concern only two classes, i.e., buried vs. exposed). The corresponding prediction methods apply fuzzy-nearest neighbor [17], neural network [16, 20, 22], support vector machine [19, 21], two stage support vector machine [18], information theory [23], and probability profile [24]. Early studies only use sequence to generate features [20, 23], while recent studies use the evolutionary information in the form of the PSSM profile to generate features [18, 19].

The PSI-BLAST profile [25] was recently introduced as an efficient sequence representation that improves classification accuracy [16]. Subsequently, researchers have found that secondary structure predicted using the PSI-PRED method [3] improves the real value RSA predictions [2].

This paper investigates whether improved sequence representation, which is based on the information harvested from the sequence, the PSI-BLAST profile and the predicted secondary structure, could lead to improving the RSA predictions. We also investigate whether it would be possible to build an index that would indicate the quality of the predicted RSA value. The above hypotheses translate into the two following goals: (1) we aim at proposing a prediction method that minimizes the RSA prediction error; (2) the method should provide a confidence value that indicates the quality of the predicted RSA values.

The first goal is achieved by designing a custom-selected set of features, which is based on performing feature selection, to represent the input sequence. As suggested in previous studies, the PSI-BLAST profile, PSI-PRED predicted secondary structure and additional features that indicate termini of the sequence were adopted to represent the input sequence. In contrast to prior works, we do not use all features from the PSI-BLAST profile, but instead we use two feature selection methods to select a subset of best-performing features. This results in a simplified prediction model, reduced computational time, and optimized predictive quality.

To address the second goal, the confidence values are computed based on the difference in predictions of RSA by two predictors: a support vector regression and a linear regression. These values can be used to indicate the quality of the output RSA predictions.

## 2. MATERIALS

### 2.1. Dataset

The dataset used in this paper is referred to as the Manesh dataset [23] and consists of 215 low-similarity, i.e., < 25%, proteins. The sequences are available online at http://gibk21.bse.kyutech.ac.jp/rvp-net/all-data.tar.gz. The Manesh dataset was widely used by researchers to benchmark prediction methods [2, 12-15, 20, 24], and this motivated us to use it to design and validate our method.

### 2.2. Relative solvet accessibility

RSA reflects the percentage of the surface area of a given residue that is accessible to the solvent. RSA value, which is normalized to [0, 1] interval, is defined as the ratio between the solvent accessible surface area (ASA) of a residue within a three-dimensional structure and ASA of its extended tripeptide (Ala-X-Ala) conformation

$$RSA = \frac{ASA \text{ in a three-dimensional structure}}{ASA \text{ in an extended tripepetide}} \qquad (1)$$

### 2.3. Feature representation

*PSI-BLAST profile.* PSI-BLAST is used to compare different protein sequences to find similar sequences and to discover evolutionary relationships [25]. PSI-BLAST generates a profile representing a set of similar protein sequences in the form of a 20×$N$ position-specific scoring matrix, where $N$ is the length of the sequence (window) and where each amino acid in the sequence (window) is described by 20 features. We used PSI-BLAST with the default parameters and the BLOSUM62 substitution matrix. The profile was computed for a 15 residues wide window centered on a target residue and thus it consists of 300 features. The selected size is motivated by previous studies that adopted this window size [18] and obtained good secondary structure prediction results [3].

*Secondary structure predicted with PSI-PRED.* The quality of secondary structure prediction has significantly improved in the last decade and nowadays it is successfully used in prediction of tertiary structure. Recently, secondary structure predicted with the PSI-PRED algorithm was shown to improve prediction of solvent accessibility [2]. We used PSI-PRED25 with default parameters to predict secondary structure from the protein sequences. PSI-PRED assigns three probabilities for each residue, which correspond to the probability of assuming helix, strand, and coil conformation, respectively. These probabilities were taken as features for the proposed RSA prediction method.

*Binary code.* The amino acids that are located at the two termini of the sequence have larger probability of being exposed to the solvent. This fact is implemented during RSA prediction by using a binary code that indicates position of a given residue that is located close to either terminus. The following binary vector

$$(a_1, a_2, a_3, a_4, a_5, b_1, b_2, b_3, b_4, b_5)$$

is used to encode the first five positions at the N terminus (denoted by $a_i$) and the last five position at the C terminus (denoted by $b_i$). For instance, the third residue in the sequence is encoded as (0,0,1,0,0,0,0,0,0,0), while a residue that is outside of the first and the last five residues in the sequence is encoded as (0,0,0,0,0,0,0,0,0,0).

### 2.4. Feature selection

PSI-BLAST profile includes 300 features, and thus feature selection methods were used to reduce the dimensionality. We applied the *correlation-based feature selection* (CBFS), and another feature selection method, namely correlation-based method for relevance and redundancy analysis (CBRR), which selects a subset of features based on filtering redundancy within the feature set. The CBFS method is based on Pearson correlation coefficient *r* computed for a pair of variables (*X*, *Y*) as

$$r = \frac{\sum_i (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_i (x_i - \bar{x}_i)^2}\sqrt{\sum_i (y_i - \bar{y}_i)^2}} \qquad (2)$$

where $\bar{x}_i$ is the mean of *X* and $\bar{y}_i$ is the mean of *Y*. The value of *r* is bounded within [-1, 1] interval. Higher absolute value of *r* corresponds to stronger correlation between *X* and *Y*. This method ranks individual features based on the correlation coefficient between each feature and the actual RSA values. A subset of features with the highest absolute *r* value is selected.

The CBRR feature selection method considers both the relevance of the features with respect to the target (RSA values), and the redundancy between the features. It involves two steps: (1) selecting a subset of relevant features, and (2) selecting predominant features from among the relevant features. The details can be found in [26].

The 300 features corresponding to the PSI-BLAST profile, 3 features corresponding to the predicted secondary structure and 10 binary code values were processed with both feature selection methods. The feature selection was processed using the training set of Manesh dataset, which includes 30 sequences [14, 20].

The CBRR method automatically filters the redundancy among the features and selects the final number of selected features, which in our case was 15. The selected features include 13 features from the PSI-BLAST profile, and 2 predicted secondary structure features, see **Table 1**. In case of CBFS, the number of selected features should be specified by the user. Hence, we tested the performance of different number of selected features using support vector regression model with default parameters to predict RSA values for the test set of the Monash dataset. The mean absolute error (MAE) steadily decreases to 15.6% by adding up to 70 features, and it saturates when adding additional features, see **Figure 1**. As a result, the 70 features with the highest Pearson correlation were selected when using CBFS. The selected features include 65 features from the PSI-BLAST profile, all 3 predicted secondary structure features, and 2 binary code values that correspond to the first and last position in the sequence, see **Table 1**.

The two feature sets selected by CBRR and CBFS and the full feature set (313 features) were compared by predicting RSA values for the test set of the Manesh dataset using support vector regression with default parameters. The 15 features selected by CBRR obtain 16.7% MAE, while the 70 features selected by CBFS and the full feature set both result in 15.6% MAE, see **Figure 2**. The features selected by CBFS provide lower MAE than the features selected by CBRR, and they cover only 23% of the full feature set. As a result, the 70 features selected By CBFS were used to design the proposed prediction model. The selected features are summarized in **Table 2**.

The feature selection shows that most of the 300 features generated by PSI-BLAST are either redundant and have little or no impact on the RSA Predictions. **Table 2** shows that when predicting RSA for the residue $A_i$ that is located in the center of the window:

– the features to encode the two leftmost positions ($A_{i-7}$, $A_{i-6}$) and the rightmost position ($A_{i+7}$) were not selected, i.e., these amino acids have no impact on the prediction of the central amino acid. Therefore, a sliding window of size 13 would be sufficient for the RSA prediction. The two amino acids that are adjacent to $A_i$, i.e., $A_{i-1}$ and $A_{i+1}$, have the most significant impact on the prediction since they correspond to the largest number of the selected features. Interestingly,

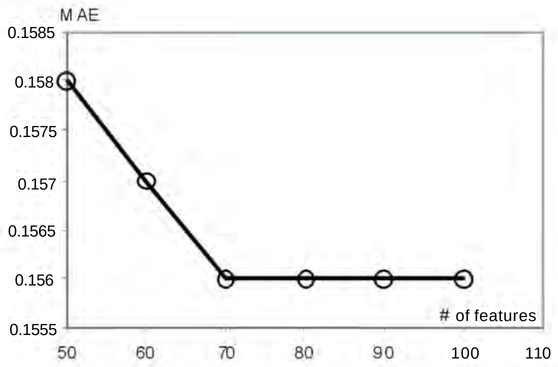


**Figure 1.** The MAE values against the number of selected features. The MAE is obtained by using support vector regression with default parameters to predict test set of the Monash dataset.

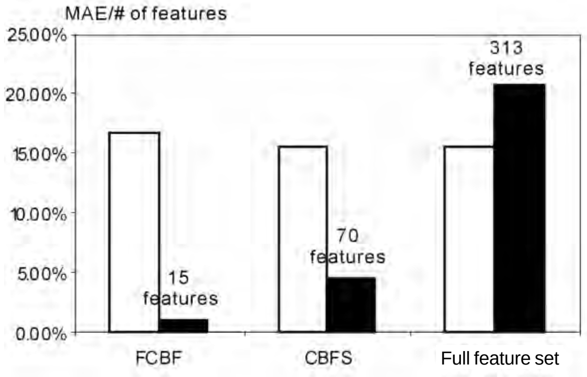


**Figure 2.** Bar chart of MAE values (white) and number of features (gray) for features selected by CBRR, CBFS, and the full feature set.

**Table 1.** Summary of the feature selection results.

| Features set | Total # features | # selected features by CBFS | # selected features by CBRR |
|---|---|---|---|
| PSI-BLAST profile | 300 | 65 | 13 |
| Binary code | 10 | 2 | 0 |
| Predicted second. structure | 3 | 3 | 2 |
| Total | 313 | 70 | 15 |

residues at *i*-2 and *i*+2 positions have relatively small influence on the prediction.

– The selected features are almost symmetrically distributed around $A_i$, e.g., amino acids E, K, Q, R, and D have similar impact on the solvent accessibility of the central residue at the third left position ($A_{i-3}$) and the third right position ($A_{i+3}$).

– Hydrophilic residues, which include E, K, Q, R, and D, may have impact on the solvent accessibility of $A_i$ residue which is 3 or 4 positions away from the these residues. This pattern covers 19 of the selected features and we hypothesize that this is related to the α-helical structures due to the following two reasons. Firstly, these 5 hydrophilic residues have larger probability (above 0.5) to form helical structure than strand and coil structures [27]. Secondly, α-helix consists of 3.6 residues per turn, and hence if two residues in a helix are separated by 2 or 3 residues in the sequence then they are spatially close to each other, which in turn may induce some interactions between them. For instance, the hydrogen bond that maintains the helical structure occurs between two residues that are separated in a sequence by three other residues, i.e., $A_i$ and $A_{i+4}$.

## 3. METHODS

### 3.1. Prediction method

Linear Regression (LR) and Support Vector Regression (SVR) were already applied in the RSA prediction [10,13,15]. In this paper, we propose an improved two-stage model, which not only aims at reducing the prediction error, but we also propose and test a confidence value that is associated with each predicted RSA value.

The proposed two-stage prediction model works as follows:

STAGE 1. The input sequences is inputted into PSIPRED to compute predicted secondary structure and into PSI-BLAST to compute the PSI-BLAST profile. Next, the input sequence, the predicted secondary structure, and the PSI-BLAST profile are used to compute the selected 70 features using a 15 residues wide window centered over the being predicted residue, and for each residue in the input sequence. The 70 features are used as an input to the LR model and SVR model that predict a real value (predicted RSA value) for the central residue in a given window.

STAGE 2. The aim of the stage two is to refine the predictions from stage one. Similarly to other two-stage designs [13,18], the second stage "smoothes" the predictions. It takes the three predicted secondary structure features (computed in stage one by PSIPRED) and a 7 residues wide window from the first stage predictions centered over the predicted residue as the input to provide the refined real value predictions.

Since the prediction quality of SVR is better than the quality of LR (results are discussed in the following), the predictions from SVR are taken as the final prediction outcome. The LR results serve as a reference to evaluate quality of SVR predictions. This means that if predictions from SVR and LR are similar then SVR predictions are assumed to be of high quality. On the other hand, if the two predictions are different then the SVR prediction is assumed to be of lower quality. The corresponding confidence value is defined as

$$C = 1 - \left| R_i - T_i \right| \tag{3}$$

where $R_i$ is the predicted RSA from SVR, and $T_i$ is the predicted RSA from LR. A detailed overview of the prediction procedure is shown in **Figure 3**.

The optimization of the prediction, through adjustment of internal parameters of the predictors and selection of the window size for the second stage, was performed by dividing the Manesh dataset into

**Table 2.** Summary of feature selection results for the PSI-BLAST profile by correlation-based feature selection method.

| 15-wide window | $A_{i-7}$ | $A_{i-6}$ | $A_{i-5}$ | $A_{i-4}$ | $A_{i-3}$ | $A_{i-2}$ | $A_{i-1}$ | $A_i$ | $A_{i+1}$ | $A_{i+2}$ | $A_{i+3}$ | $A_{i+4}$ | $A_{i+5}$ | $A_{i+6}$ | $A_{i+7}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total # of features | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| # of selected features | 0 | 0 | 2 | 4 | 5 | 0 | 8 | 19 | 7 | 1 | 6 | 6 | 4 | 3 | 0 |
| The selected features | | | I<br>L | E<br>K<br>Q<br>R | E<br>K<br>Q<br>R<br>D | | E<br>K<br>Q<br>R<br>D<br>N<br>P<br>S | C D<br>E F<br>G H<br>I K<br>L M<br>N P<br>Q R<br>S T<br>V W<br>Y | E<br>K<br>Q<br>H<br>D<br>N<br>G | P | E<br>K<br>Q<br>R<br>D<br>P | E<br>K<br>Q<br>R<br>D<br>P | I<br>L<br>V<br>F | I<br>L<br>V | |

**Table 3**. Optimization of parameters for two-stage SVR.

| First stage | | | Second stage | | |
|---|---|---|---|---|---|
| Parameter C | Parameter γ | MAE | Parameter C | Parameter γ | MAE |
| 1 | 0.001 | 0.157 | 1 | 0.01 | 0.150 |
| 1 | 0.005 | 0.153 | 1 | 0.08 | 0.149 |
| 1 | 0.01 | 0.151 | 1 | 0.15 | 0.148 |
| 1 | 0.02 | 0.151 | 1 | 0.2 | 0.148 |
| 1 | 0.03 | 0.152 | 1 | 0.3 | 0.148 |
| 1 | 0.05 | 0.155 | 1 | 0.4 | 0.149 |
| 0.5 | 0.01 | 0.152 | 0.5 | 0.15 | 0.148 |
| 0.8 | 0.01 | 0.151 | 0.8 | 0.15 | 0.148 |
| 1 | 0.01 | 0.151 | 1 | 0.15 | 0.148 |
| 2 | 0.01 | 0.151 | 2 | 0.15 | 0.148 |
| 3 | 0.01 | 0.151 | 3 | 0.15 | 0.148 |
| 5 | 0.01 | 0.152 | 5 | 0.15 | 0.148 |

two subsets, one used to compute the prediction model and the other to perform test. Similarly to [14], 30 sequences were used for training and the remaining 185 as the test set. The linear regression is parameterless and thus it does not require optimization. For SVR, RBF kernel was used for both stages. The parameters for the first stage SVR are $\gamma$=0.01 and $C$=1, and for the second stage $\gamma$=0.15 and $C$=1. These parameters, which were based on experiments summarized in **Table 3**, provide the lowest MAE. We note that the adjustment of C has little impact in the quality of predictions. The MAE of the final prediction for the second stage windows sizes of 5, 7, 9, 11, 15, and 21 equal 0.149, 0.148, 0.148, 0.148, 0.148, and 0.148, respectively. This shows that the window size of 7 is the best choice to provide accurate predictions.

### 3.2. Linear regreesion

A linear regression with $p$ coefficients and $n$ data points (number of samples), assuming that $n>p$, corresponds to the construction of the following expression:

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{11} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{pmatrix} \quad (4)$$

where $y_i$ is the predicted RSA value, $x_i = (x_{i1}, x_{i2},\ldots, x_{ip})$ is the vector of $p$ features representing $i^{th}$ protein sequence, $\beta_i$ (constant) is parameter to be estimated, and $\varepsilon_i$ is the standard error. The above formula can be written in vector-matrix form as:

$$y = X.\beta + \varepsilon \quad (5)$$

The solution to minimize the mean square error $||\varepsilon_i||$ is

$$\beta = (X^T X)^{-1} X^T \vec{y} \qquad \vec{\varepsilon} = \vec{y} - X.\beta \quad (6)$$

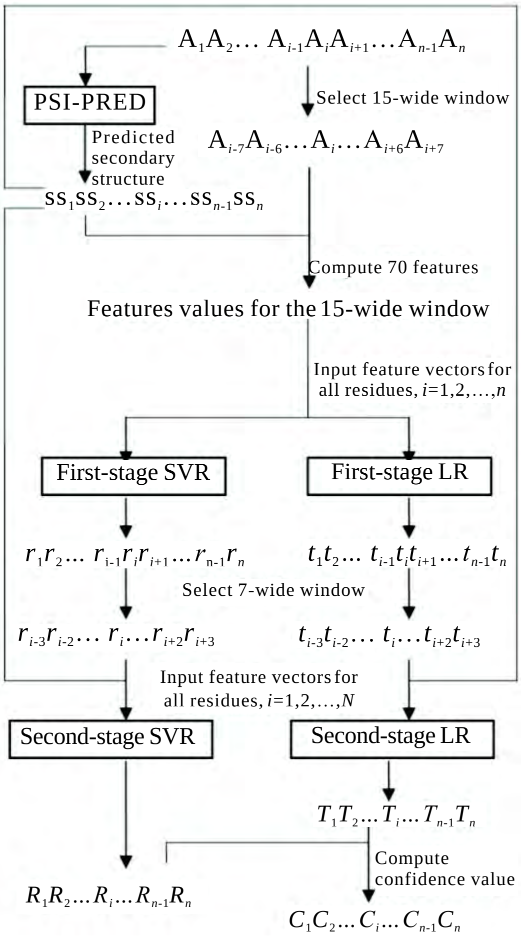

$A_1A_2\ldots A_{i-1}A_iA_{i+1}\ldots A_{n-1}A_n$
PSI-PRED
Select 15-wide window
Predicted secondary structure
$A_{i-7}A_{i-6}\ldots A_i\ldots A_{i+6}A_{i+7}$
$SS_1SS_2\ldots SS_i\ldots SS_{n-1}SS_n$
Compute 70 features
Features values for the 15-wide window
Input feature vectors for all residues, $i$=1,2,…,$n$
First-stage SVR
First-stage LR
$r_1r_2\ldots r_{i-1}r_ir_{i+1}\ldots r_{n-1}r_n$
$t_1t_2\ldots t_{i-1}t_it_{i+1}\ldots t_{n-1}t_n$
Select 7-wide window
$r_{i-3}r_{i-2}\ldots r_i\ldots r_{i+2}r_{i+3}$
$t_{i-3}t_{i-2}\ldots t_i\ldots t_{i+2}t_{i+3}$
Input feature vectors for all residues, $i$=1,2,…,$N$
Second-stage SVR
Second-stage LR
$T_1T_2\ldots T_i\ldots T_{n-1}T_n$
Compute confidence value
$R_1R_2\ldots R_i\ldots R_{n-1}R_n$
$C_1C_2\ldots C_i\ldots C_{n-1}C_n$

**Figure 3**. RSA prediction with the proposed system; the RSA value for the $i^{th}$ residue is predicted based on the 70 feature values (see **Table 1**) that are computed over a 15 residues wide window centered on $i^{th}$ residue; the feature values are inputted into the first-stage predictor (LR and SVR); next, the first-stage predictions are aggregated into 7 residue wide windows and inputted, together with the predicted secondary structure of the central residue, into the second-stage predictor that provides the RSA values. Finally, compare the predictions from SVR and LR, and calculate the confidence value *C*.

### 3.3. Support vector regression

Given a training set of $n$ data point pairs $(x_i, y_i)$, $i$ = 1, 2,…, $n$, where $x_i$ denotes the vector of $p$ features representing $i^{th}$ protein sequence, $y_i$ denotes the predicted RSA value, finding the optimal SVR is achieved by solving:

$$\min \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i + \xi_i^*) \quad (7)$$

such that

$$\begin{aligned} & y_i - w\cdot x_i - b \le \varepsilon + \xi_i \\ & w\cdot x_i + b - y_i \le \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \ge 0 \end{aligned} \quad (8)$$

where $w$ is a vector perpendicular to $wx$-$b$=0 hyperplane, $C$ is a user defined complexity constant, $\xi_i$ and $\xi_i^*$ are slack variables that measure the degree of prediction error of $x_i$ for a given hyperplane, and $z=\phi(x)$ where $k(x,x')=\phi(x)\cdot\phi(x')$ is a user defined kernel function.

The SVR was trained using sequential minimal optimization algorithm [28] that was further optimized by Shevade and colleagues [29]. The proposed SVR uses RBF kernel

$$k(x_i, x_i') = e^{-\gamma\|x-x'\|^2} \qquad (9)$$

for both stages.

## 4. RESULTS AND DISCUSSION

The SVR and LR predictors were implemented in Weka [30], which is a comprehensive open-source library of machine learning methods. The Manesh dataset consists of 50682 instances (individual residues). The evaluation was performed using two test types to allow for a comprehensive comparison with previous studies. To compare with [2] and [12], 5-folds cross validation was executed. On the other hand, following several other prior studies [14, 20, 24], Manesh dataset was divided into two subsets, 30 sequences were used for training and the remaining 185 as independent test set. The results of both tests, i.e., 5 folds cross-validation and independent test, were reported in **Tables 4 and 5**. In total, the proposed method was compared with six real value RSA prediction methods [2, 12-15, 24] and one method that aims at prediction of discrete states [20].

We note that in statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, sub-sampling (such as 5-fold and 7-fold) test, and jackknife test [31]. However, as elucidated by [32] and demonstrated in [33], among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used by investigators to examine the accuracy of various predictors [34-42].

### 4.1. Comparison with competing prediction methods

For the 5 folds cross-validation test, the mean absolute error (MAE) value of the first stage of the proposed method equals 14.6 and the corresponding Pearson's correlation coefficient (*r*) equals 0.67. After the second stage, the MAE value is reduced to 14.3 and *r* is improved to 0.68. **Table 4** compares the proposed two-stage SVR with recent methods for RSA prediction, which include neural network and support vector regression models [2, 12, 13, 15]. The proposed method obtains 0.6 to 3.7 lower MAE when compared with the abovementioned methods. This translates into 4% to 20% error reduction, respectively. Since some methods predict discrete valued classes (exposed vs. buried), we also examined the performance of our method by converting the real value prediction into the two states prediction. We followed the standard approach, in which the state is defined based on the predicted RSA value and a predefined threshold. For instance, a 5% threshold means that the residues having an RSA value (%) greater or equal 5 are defined as exposed, and otherwise they are classified as buried. The threshold's value is usually adjusted between 5% and 50%. We note that for all thresholds, our method provides the highest accuracy, see **Table 4**. The proposed two-stage model provides 0.3%-0.6% higher accuracies than the prediction coming from the first stage for various thresholds. When compared to the best performing, existing two-stage SVR method [13], our predictions are characterized by lower MAE and more accurate two states predictions.

For the independent test, the MAE value for the first stage of the proposed method equals 15.0 and the corresponding Pearson's correlation coefficient *r* equals 0.66. After the second stage, the MAE value is reduced to 14.8 and *r* is improved to 0.67. **Table 5** compares the proposed two-stage SVR with recent methods for RSA prediction, which include neural network and look-up table based methods [14, 20, 24]. The proposed method obtains 1.5 to 4.0 lower MAE when compared with the above three methods. This translates into 9% to 21% error reduction, respectively. Similarly to the 5-folds cross validation test, we also examined the performance of our method by converting the real value prediction into the two states prediction. The threshold's value was adjusted between 5 and 50%.

For all thresholds our method consistently provides the highest accuracy, see **Table 5**. The two-

**Table 4.** Experimental comparison between the proposed two-stage SVR and other reported methods; the results were reported based on 3 or 5-folds cross validation test; the real valued predictions were converted to two state prediction (buried vs. exposed) with different threshold (5%~50%); unreported results are denoted by "-"; best results are shown in bold.

| Reference | Prediction method | MAE (%) | Correlation coefficient *r* | Accuracy for two-states (buried vs. exposed) prediction | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | 5% | 10% | 20% | 30% | 40% | 50% |
| [2] | Neural Network | 15.2 | 0.67 | 74.9% | 77.2% | 77.7% | 77.8% | 78.1% | 80.5% |
| [11] | Neural Network | 18.0 | 0.50 | - | - | - | - | - | - |
| [12] | Two-stage SVR | 14.9 | 0.68 | 81.1% | 78.5% | 77.6% | - | - | 79.5% |
| [14] | SVR | 16.3 | 0.58 | - | - | - | - | - | - |
| This paper | One-stage SVR | 14.6 | 0.67 | 80.5% | 79.1% | 78.3% | 78.3% | 78.3% | 80.5% |
| This paper | Two-stage SVR | 14.3 | 0.68 | 81.1% | 79.7% | 78.8% | 78.6% | 78.8% | 80.8% |

**Table 5.** Experimental comparison between the proposed two-stage SVR and other reported methods; the results were reported based on a test on the independent dataset (30 sequences for training and 185 sequences for test); the real valued predictions were converted to two state prediction (buried vs. exposed) with different threshold (5%~50%); unreported results are denoted by "-"; best results are shown in bold.

| Reference | Prediction method | MAE (%) | Correlation coefficient r | Accuracy for two-states (buried vs. exposed) prediction | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | 5% | 10% | 20% | 30% | 40% | 50% |
| [13] | Look-up table | 18.8 | 0.48 | - | - | - | - | - | - |
| [19] | Neural Network | - | - | 74.6% | 71.2% | - | - | - | 75.9% |
| [23] | Neural Network | 16.3 | 0.58 | 75.7% | 73.4% | - | - | - | 76.2% |
| This paper | One-stage SVR | 15.0 | 0.66 | 79.8% | 78.7% | 77.7% | 77.7% | 77.5% | 79.8% |
| This paper | Two-stage SVR | 14.8 | 0.67 | 80.3% | 79.2% | 78.1% | 78.0% | 78.0% | 80.2% |

stage model provides 0.3%-0.5% higher accuracies than the one-stage model for various thresholds. When compared with the best-performing, competing method based on neural network [24], our predictions result in higher accuracies over all thresholds, i.e., the differences range between 4% and 5.8%, and better MAE and correlation coefficient value.

The three main observations based on the performed empirical evaluation include: (1) the proposed two-state predictor obtains favorable (lower) error rates when compared with six competing methods; (2) the improvements are obtained for both real value and two-state predictions; and (3) the introduction of the second stage in our design allows for obtaining improved predictions when compared with a one stage design.

### 4.2. Confidence value for RSA prediction

As one of the goals of this work, we defined confidence values to measure the quality of the predicted RSA. The confidence values are based on the difference of predictions made by the two-stage SVR and the two-stage LR. The following discussion is based on results of five folds cross-validation tests.

The MAE for two-stage SVR is 0.143 and for two-stage LR is 0.155. The difference between the predictions from SVR and LR for the same residues ranges between 0 and 0.294. As a result, the confidence value *C* distributed in the interval [0.706, 1] for the Manesh dataset. Higher *C* values indicate that the predictions from SVR and LR are more consistent, and therefore the corresponding predictions from the two-stage SVR are assumed to be more accurate.

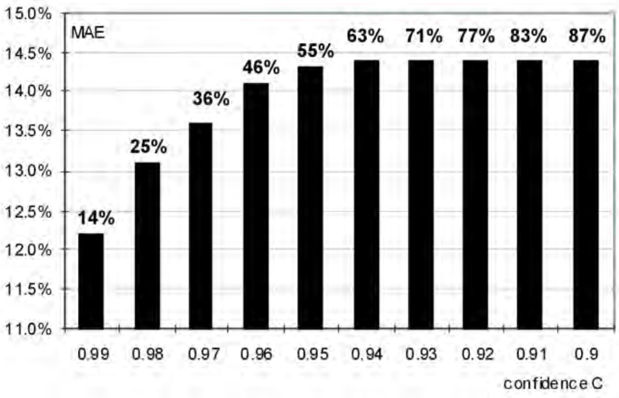


**Figure 4.** Bar chart of MAE values for the corresponding thresholds of confidence value *C*. The numbers above the bar show the corresponding coverage, i.e., number of residues for which the predictions had confidence value above the threshold. For example, for residues predicted with which *C* > 0.99 the MAE equals 12.2, and these residues cover 14% of the dataset.

The *C* value of 7101 samples, which covers 7101/50682= 14% of the dataset, are greater than 0.99, and the corresponding MAE of these samples equals 0.122, see **Figure 4**. The *C* value of 12846 samples, which covers 12846/50682= 25.3% of the dataset, are greater than 0.98, and the corresponding MAE of these samples equals 0.131. The *C* value of 18174 samples, which covers 18174/50682= 35.9% of the dataset, are greater than 0.97, and the MAE of these samples is 0.136. When the threshold for *C* value is set equal or lower than 0.96, the MAE saturates at 0.143, see **Figure 4**, which is equal to the MAE for the entire dataset (without using the confidence values). This shows that the confidence values can be used to identify a subset of the predictions which on average have better quality than the remaining predictions. This way, the user could select a desired fraction of best performing predictions. Additionally, the user could inspect quality of prediction for specific amino acids or groupings of amino acids that share certain properties such as hydrophobicity, charge, size, etc.

## 5. CONCLUSIONS

This paper proposes a novel method for the real value RSA prediction. The proposed method addresses two goals, which include improving the quality of RSA prediction, and development of a confidence value that allows for selection of better performing RSA predictions.

Empirical tests with the Manesh dataset show that the proposed method is characterized by lower prediction error when compared with six competing real value RSA prediction methods. We also show that the PSI-BLAST profile that is commonly used to represent sequences can by largely reduced by using feature selection, which results a simpler, interpretable model and in reduction of the computational time required to develop the prediction model. Our model indicates that window size of 13 is sufficient and only about 22% of the PSI-BLAST features are useful for the RSA prediction. The selected features are symmetrically distributed around the predicted residue and include hydrophilic resides when considering the distance of 3 or 4 positions from the predicted residue. The confidence value *C* allows the user to select a subset of the predictions which on average are characterized by better quality than the remaining predictions.

The knowledge of the surface residues, which are predicted by the proposed method and which are directly involved in the interaction with other biological molecules, was used, for instance, for identifying protein function and stability [43, 44], for prediction of binding sites [4], understanding the binding mechanism of proteins [5], reconstruction of the 3D-structure of proteins [6-8], and to aid fold recognition [45, 46]. Therefore, improved prediction of the surface residues would have impact on improving quality of solutions for these associated tasks.

## ACKNOWLEDGMENTS

This work was supported in part by NSERC Canada. K.C. also acknowledges support provided through scholarship sponsored by Alberta Ingenuity Fund.

## REFERENCE

[1] Ginalski, K. & Rychlewski, L. Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment. *Proteins* 2003, 53(Suppl. 6):410-417.

[2] Garg, A., Kaur, H. & Raghava, G. P. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. *Proteins* 2005, 61(2):318-24.

[3] Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol.* 1999, 292(2):195-202.

[4] Huang, B. & Schroeder, M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol.* 2006, 6:19.

[5] Chou, K. C. Review: Low-frequency collective motion in biomacromolecules and its biological functions. *Biophysical Chemistry* 1988, 30: 3-48

[6] Chan, H. S. & Dill, K. A. Origins of structures in globular proteins. *Proc Natl Acad Sci USA* 1990, 87: 6388-92.

[7] Wang, J. Y., Lee, H. M. & Ahmad, S. Prediction and evolutionary information analysis of protein solvent accessibility using multiple linear regression. *Proteins* 2005, 61(3):481-91.

[8] Arauzo-Bravo, M. J., Ahmad, S. & Sarai, A. Dimensionality of amino acid space and solvent accessibility prediction with neural networks. *Comput Biol Chem.* 2006, (2):160-8.

[9] Wagner, M., Adamczak, R., Porollo, A. & Meller, J. Linear regression models for solvent accessibility prediction in proteins. *J Comput Biol.* 2005, 12(3):355-69.

[10] Yuan, Z. & Huang, B. Prediction of protein accessible surface areas by support vector regression. *Proteins* 2004, 57(3):558-64.

[11] Adamczak, R., Porollo, A. & Meller, J. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins* 2004, 56(4):753-67.

[12] Ahmad, S., Gromiha, M. M. & Sarai, A. Real value prediction of solvent accessibility from amino acid sequence. *Proteins* 2003, 50(4):629-35.

[13] Nguyen, M. N.& Rajapakse, J. C. Two-stage support vector regression approach for predicting accessible surface areas of amino acids. *Proteins* 2006, 63(3):542-50.

[14] Wang, J. Y., Ahmad, S., Gromiha, M. M. & Sarai, A. Look-up tables for protein solvent accessibility prediction and nearest neighbor effect analysis. *Biopolymers* 2004, 75(3):209-16.

[15] Xu, W. L., Li, A., Wang, X., Jiang, Z. H. & Feng, H. Q. Improving Prediction of Residue Solvent Accessibility with SVR and Multiple Sequence Alignment Profile. *Proceedings of the 27th IEEE Annual Conference on Engineering in Medicine and Biology,* Shanghai, China, 2005.

[16] Cuff, J. A. & Barton, G. J. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 2000, 40(3):502-11.

[17] Sim, J., Kim, S. Y. & Lee, J. Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method. *Bioinformatics* 2005, 21(12):2844-9.

[18] Nguyen, M. N. & Rajapakse, J. C. Prediction of protein relative solvent accessibility with a two-stage SVM approach. *Proteins* 2005, 59(1):30-7.

[19] Kim, H. & Park, H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. *Proteins* 2004, 54(3):557-62.

[20] Ahmad, S. & Gromiha, M. M. NETASA: neural network based prediction of solvent accessibility. *Bioinformatics* 2002, 18(6):819-24.

[21] Yuan, Z., Burrage, K. & Mattick, J. S. Prediction of protein solvent accessibility using support vector machines. *Proteins* 2002, 48(3):566-70.

[22] Gianese, G. & Pascarella, S. A consensus procedure improving solvent accessibility prediction. *J Comput Chem.* 2006, 27(5):621-6.

[23] Naderi-Manesh, H., Sadeghi, M., Araf, S.& Movahedi, A. A. M. Predicting of protein surface accessibility with information theory. *Proteins* 2001, 42:452-459.

[24] Gianese, G., Bossa, F. & Pascarella, S. Improvement in prediction of solvent accessibility by probability profiles. *Protein Eng.* 2003, 16(12):987-92.

[25] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. & Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* 1997, 17:3389-402.

[26] Yu, L. & Liu, H. Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research.* 2004, 5:1205-24.

[27] Chen, K., Kurgan, L. & Ruan, J. Optimization of the Sliding Window Size for Protein Structure Prediction, *IEEE Symposium on Comp Intelligence in Bioinformatics and Computational Biology,* 2006, 366-72.

[28] Smola, A. J. & Scholkopf, Bernhard. *A Tutorial on Support Vector Regression.* NeuroCOLT2 Technical Report Series, 1998.

[29] Shevade, S. K., Keerthi, S. S., Bhattacharyya, C. & Murthy, K., *Improvements to SMO Algorithm for SVM Regression.* Technical Report CD-99-16, Control Division Dept of Mechanical and Production Engineering, National University of Singapore, 1999.

[30] Witten, I. & Frank, E. *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, 2005.

[31] Chou, K. C. & Zhang, C. T. Review: Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology* 1995, 30:275-349.

[32] Chou, K. C. & Shen, H. B. Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. *Nature Protocols* 2008, 3:153-162.

[33] Chou, K. C. & Shen, H. B. Review: Recent progresses in protein subcellular location prediction. *Analytical Biochemistry* 2007, 370:1-16.

[34] Diao, Y., Ma, D., Wen, Z., Yin, J., Xiang, J. & Li, M. Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel-Ziv complexity. *Amino Acids* 2008, 34:111-117.

[35] Tan, F., Feng, X., Fang, Z., Li, M., Guo, Y. & Jiang, L. Prediction of mitochondrial proteins based on genetic algorithm partial least squares and support vector machine. *Amino Acids* 2007, 33:669-675.

[36] Li, F. M. & Li, Q. Z. Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. *Amino Acids* 2008, 34:119-125.

[37] Fang, Y., Guo, Y., Feng, Y. & Li, M. Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids 2008, 34:103-109.

[38] Zhang, S. W., Zhang, Y. L., Yang, H. F., Zhao, C. H. & Pan, Q. Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. *Amino Acids* 2007, DOI 10.1007/s00726-00007-00010-00729.

[39] Shi, J. Y., Zhang, S. W., Pan, Q. & Zhou, G. P. Using Pseudo Amino Acid Composition to Predict Protein Subcellular Location: Approached with Amino Acid Composition Distribution. *Amino Acids* 2007, DOI 10.1007/s00726-00007-00623-z.

[40] Zhou, X. B., Chen, C., Li, Z. C. & Zou, X. Y. Improved prediction of subcellular location for apoptosis proteins by the dual-layer support vector machine. *Amino Acids* 2007, DOI 10.1007/s00726-00007-00608-y.

[41] Nanni, L. & Lumini, A. Combing Ontologies and Dipeptide composition for predicting DNA-binding proteins. *Amino Acids* 2008, DOI 10.1007/s00726-00007-00018-00721.

[42] Nanni, L. & Lumini, A. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids* 2008, DOI 10.1007/s00726-00007-00016-00723.

[43] Eisenberg, D. & McLachlan, A. D. Solvation energy in protein folding and binding. *Nature* 1986, 319:199-203.

[44] Gromiha, M. M., Motohisa, O., Hidetoshi, K., Hatsuho, U. & Akinori, S. Role of structural and sequence information in the prediction of protein stability changes, comparison between buried and partially buried mutations. *Protein Engineering* 1999, 12(7):549-555.

[45] Cheng, J. & Baldi, P. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics* 2006, 22(12):1456-63.

[46] Liu, S., Zhang, C., Liang, S. & Zhou, Y. Fold recognition by concurrent use of solvent accessibility and residue depth. *Proteins* 2007, 68:636-645.

J. Biomedical Science and Engineering, 2008, 1, 10-14

Scientific
Research
Publishing

# Comparative analysis of internal and external-hex crown connection systems - a finite element study

**Rudi C. Van Staden *** [1], **Hong Guan**[1], **Yew-Chaye Loo**[1], **Newell W. Johnson**[1] **& Meredith Nell**[2]

[1]Griffith School of Engineering/Dentistry and Oral Health, Griffith University, Gold Coast Campus, Queensland 4222, Australia. [2]Neoss Pty Ltd,Harrogate HG1 2PW, United Kingdom.*Correspondence should be addressed to Rudi C. Van Staden (r.vanstaden@griffith.edu.au).

## ABSTRACT

**Objectives: The abutment connection with the crown is fundamental to the structural stability of the implant system and to the prevention of mechanical exertion that can compromise the success of the implant treatment. The aim of this study is to clarify the difference in the stress distribution patterns between implants with internal and external-hex connections with the crown using the Finite Element Method (FEM). Material and Methods: The internal and external-hex connections of the Neoss and 3i implant systems respectively, are considered. The geometrical properties of the implant systems are modeled using three-dimensional (3D) brick elements. Loading conditions include a masticatory force of 200, 500 and 1000N applied to the occlusal surface of the crown along with an abutment screw torque of 110, 320 and 550Nmm. The von Mises stress distribution in the crown is examined for all loading conditions. Assumptions made in the modeling include: 1. half of the implant system is modeled and symmetrical boundary conditions applied; 2. temperature sensitive elements are used to replicate the torque within the abutment screw. Results: The connection type strongly influences the resulting stress characteristics within the crown. The magnitude of stress produced by the internal-hex implant system is generally lower than that of the external-hex system. The internal-hex system held an advantage by including the use of an abutment between the abutment screw and the crown. Conclusions: The geometrical design of the external-hex system tends to induce stress concentrations in the crown at a distance of 2.89mm from the apex. At this location the torque applied to the abutment screw also affects the stresses, so that the compressive stresses on the right hand side of the crown are increased. The internal-hex system has reduced stress concentrations in the crown. However, because the torque is transferred through the abutment screw to the abutment contact, changing the torque has greater effect on this hex system than the masticatory force. Overall the masticatory force is more influential on the stress within the crown for the external-hex system and the torque is more influential on the internal-hex system.**

**Keywords: Component; Biomedical modelling; Dental implant; Finite element technique**

## 1. INTRODUCTION

Dental implants are a consistently accepted form of dental treatment. Clinical research in oral implantology has led to advancements in the biomechanical aspects of implants, implant surface features and implant componentry. These advancements in implant componentry include the modification of the external-hex connection between the abutment and crown to the currently used internal-hex (**Figure 1b**)). Although both internal and external-hex connected implant systems are extensively used, distinctly different performances are on offer in terms of the stress characteristics produced within the crown. Observations by practitioners have aided the identification of implant components which lead to mechanical failure of the crown and implant [1-3]. Failure may be defined as the point at which the material exceeds the fracture stress, as indicated by its stress strain relationship. There are two major factors which can cause the crown and implant to fail. These are described below;

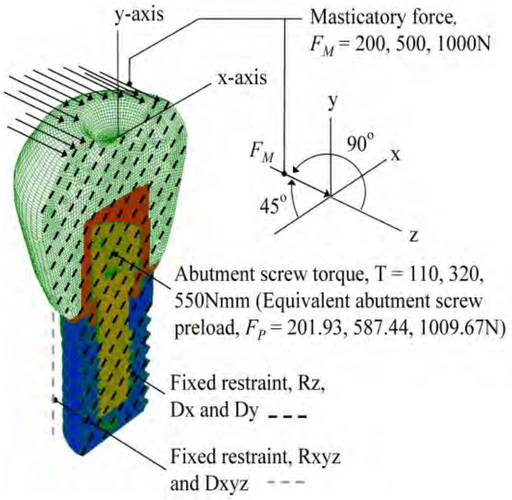
- typically, over tightening of the abutment screw causes failure of the crown for internal and external-hex systems.

- failure of the implant may also be a result of over tightening of the abutment screw or excessive masticatory loads being transferred from the occlusal plane of the crown to an area of stress concentration at the interface between the abutment and implant body.

Using theoretical techniques, such as the FEM, all mechanical aspects that could affect the implant success can be evaluated. FEM has been used extensively to evaluate the performance of dental implant prosthesis [4-15]. Studies by Maeda *et al.* (2006), Merz *et al.* (2000) and Khraisat *et al.* (2002) have all considered the behavior of the stress within the abutment screw however disregarding the stress within the crown. To date no published research appears to have investigated the stress characteristics in the crown due to an internal or external-hex system. Ultimately, the outcome of this study will facilitate dental practitioners to identify locations within the implant system that are susceptible to stress concentrations.

## 2. METHODOLOGY

The modeling and simulation herein are performed using the Strand7 Finite Element Analysis (FEA) System (2004). The first step of the modeling is to define the geometry of the implant system. This is then followed by specifying the material behavior in terms of the Young's modulus, Poisson's ratio and density for the implant and componentry. After applying the appropriate loading and restraint conditions, the internal and external-hex systems can be evaluated for their contributions to the stress characteristics within the crown.


y-axis
Masticatory force,
$F_M$ = 200, 500, 1000N
x-axis
y
90°
x
$F_M$
45°
z
Abutment screw torque, T = 110, 320, 550Nmm (Equivalent abutment screw preload, $F_P$ = 201.93, 587.44, 1009.67N)
Fixed restraint, Rz, Dx and Dy
Fixed restraint, Rxyz and Dxyz

a) Loading and restraint conditions (with detailed variables)

### 2.1. Modelling

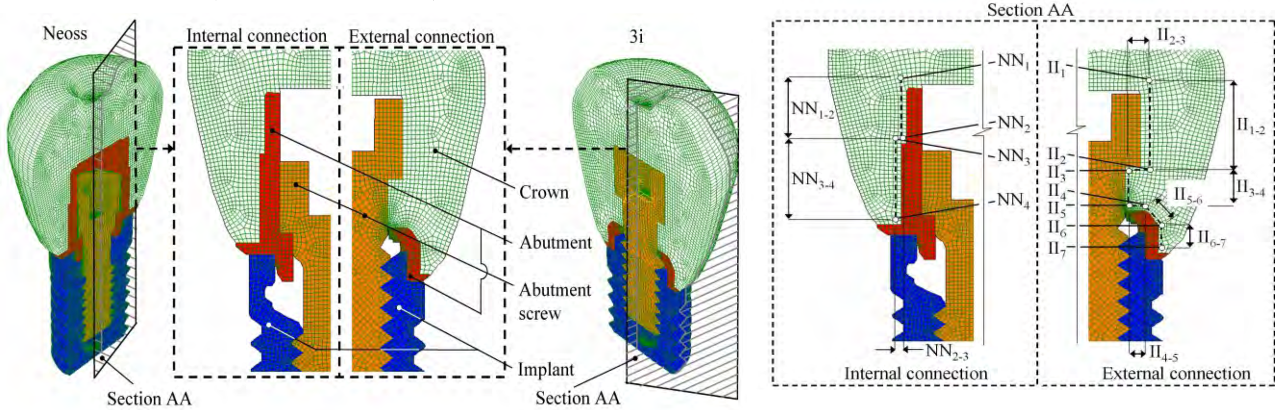
Data acquisition for the internal and external-hex systems are obtained from the manufacturer's data. Shown in **Figure 1b**) are details of the Neoss (2006) and 3i (2006) systems.

Shown in **Figure 1a**) are the detailed variables considered in this study. The implant is conical with 2 degrees of taperage, a helical thread, diameter of 4.5mm, and length of 11mm. Different fixed restraints are applied to the symmetrical edge of the implant system as compared to the outer edge of implant thread. The symmetrical edge is restrained from rotating around the z-axis and translating through the x- and y-axis. The outer edge of the implant thread is restrained from deforming in any direction. Note that these loading and restraint conditions are the same for both internal and external-hex systems.

For the Neoss and 3i finite element models, the total numbers of elements are respectively 13464 and 30420 for the implant, 3564 and 9108 for the abutment, 17424 and 25956 for the abutment screw, 38484 and 47052 for the crown. The total number of nodal points for the entire Neoss and 3i models are 82547 and 122688 respectively.

### 2.2. Stress Measuring

As indicated in **Figure 1c**) the von Mises stresses along the lines NN ($NN_{1-2}$, $NN_{2-3}$ and $NN_{3-4}$) and II ($II_{1-2}$, $II_{2-3}$, $II_{3-4}$, $II_{4-5}$, $II_{5-6}$ and $II_{6-7}$) for the Neoss and 3i systems respectively, are measured for all possible combinations of loading. Note that, for example, along the line $II_{1-2}$ the beginning location of the line is identified as $II_1$ and the end as $II_2$. These locations are believed by clinicians to be critical for examining the stress levels in the crown. Note that both lines NN and II are chosen on Section AA because the highest stress magnitudes (compressive is prominent over ten-



b) Implant systems

c) Locations for measuring stress profile and contour

**Figure 1.** Finite element model of internal and external-hex systems.

sile) occur on this plane due to the masticatory loading characteristics.

### 2.3. Loading Conditions

Masticatory force, $F_M$, is applied to the occlusal surface of the crown at 100, 250 or 500N, inclined at 45° along the x- and y-axis (**Figure 1a**). The preload, $F_P$, of 100.97, 293.72 or 504.84N is applied to the abutment screw through the use of temperature sensitive elements (**Figure 1a**)). Note that $F_M$ and $F_P$ are set to half of the total magnitude because only half of the implant system is modelled. Therefore the total $F_M$ modelled is 200, 500, 1000N and $F_P$ is 201.93, 587.44, 1009.67N. The manner of modelling the masticatory forces and the preload applied to the abutment screw is described by van Staden *et al.* (2008). In this study both the abutment screw preload, $F_P$, and surface area between abutment and abutment screw are halved when compared with that used by van Staden *et al.* (2008) due to the modelling assumption aforementioned. Calculations for the abutment screw surface pressure, $q$, confer identical results than that found by van Staden *et al.* (2008).

For the present study a negative temperature (-10 Kelvin, K) is applied to all the nodal points within the abutment screw, causing each element to shrink. A trial and error process is applied to determine the temperature coefficient, $C$, for both the Neoss and 3i systems (i.e. $C_{Neoss}$ and $C_{3i}$) that can yield an equivalent $q$. It is found that when $F_P$=201.93, 587.44 and 1009.67N then $C_{Neoss}=-3.51\times10^{-4}$, $-9.28\times10^{-4}$ and $-15.60\times10^{-4}$ /K, and $C_{3i}=-0.98\times10^{-4}$, $-1.80\times10^{-4}$ and $-2.68\times10^{-4}$ /K, respectively.

**Table 1.** Material propertles.

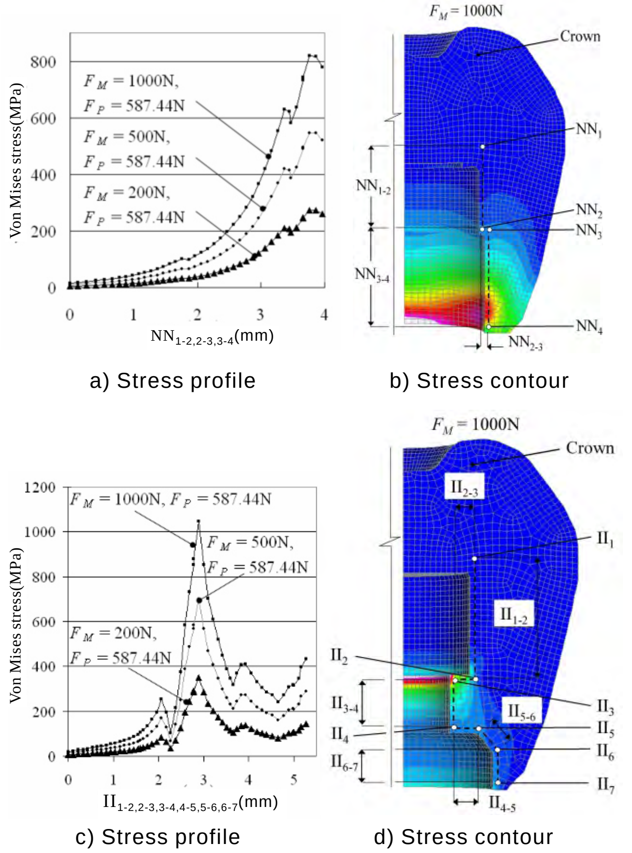
| Component | Description | Young's modulus,E(Gpa) | Poisons ratio,v | Density,p (g/cm3) |
|---|---|---|---|---|
| Implantand abutment | Titanium (grade4) | 105.00 | 0.37 | 4.51 |
| Abutment screw | Gold(precisionalloy) | 93.00 | 0.30 | 16.30 |
| Crown | Zirconia(Y-TZP) | 172.00 | 0.33 | 6.05 |

### 2.4. Material Properties

The material properties used are specified in terms of Young's modulus, Poisson's ratio and the density for the implant and all associated components (**Table 1**). All material properties are assumed to be linear, homogeneous and elastic in behavior.

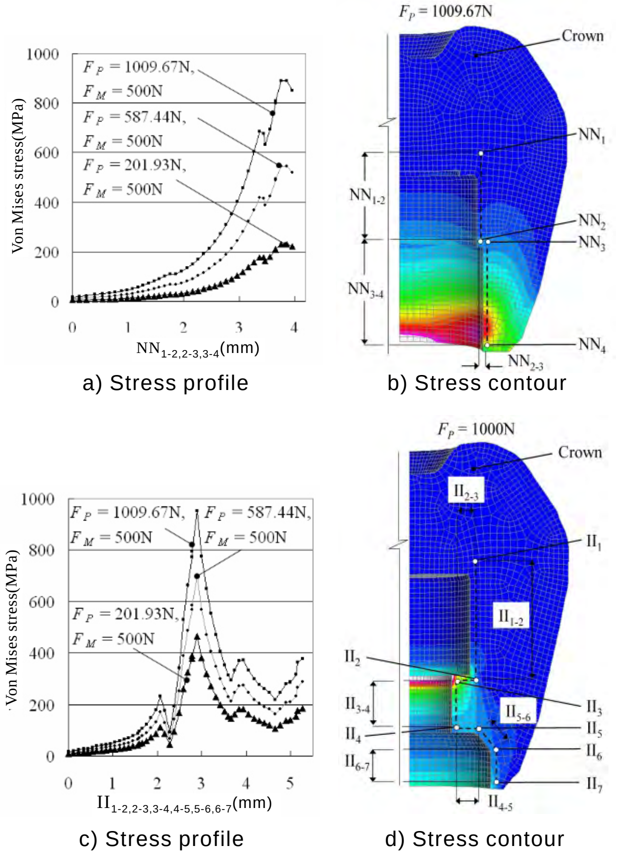
## 3. RESULTS DISCUSSION

Zirconia typically used as a dielectric material has proven adequate for application in dentistry. With its typical white appearance and high Young's moduli it is ideal to be used in the manufacturing of sub frames



a) Stress profile
b) Stress contour
c) Stress profile
d) Stress contour

**Figure 2.** Stress characteristics when varying $F_M$.



a) Stress profile
b) Stress contour
c) Stress profile
d) Stress contour

**Figure 3.** Stress characteristics when varying $F_P$.

for the construction of dental restorations such as crowns and bridges, which are then veneered with conventional feldspathic porcelain. Zirconia has a fracture strength that exceeds that of Titanium therefore it may be considered as a high strength material. However with cyclic preload and masticatory loads the compressive strength of 2.1GPa (Curtis *et al.* 2005) can easily be exceeded especially for implant systems with external-hex connections, as confirmed during this study.

The distribution of von Mises stresses in the crown is discussed for both the internal and external-hex systems for all combinations of masticatory and preload forces. Shown in **Figure 1c**), are the von Mises stresses measured between locations $NN_{1-2}$ (0-1.76mm), $NN_{2-3}$ (1.76-1.87mm) and $NN_{3-4}$ (1.87-3.96mm) for the Neoss system. For the 3i system the von Mises stresses are measured between locations $II_{1-2}$ (0-2.38mm), $II_{2-3}$ (2.38-2.78mm), $II_{3-4}$ (2.78-3.67mm), $II_{4-5}$ (3.67-4.06mm), $II_{5-6}$ (4.06-4.65mm) and $II_{6-7}$(4.65-5.27mm), as shown in **Figure 1c**).

### 3.1. Masticatory Force, $F_M$

The distributions of von Mises stresses along the lines NN and II for all values of $F_M$ are shown in **Figure 2**. Note that the preload, $F_p$, is set at its medium value, i.e. 587.44N.

In general, when the applied masticatory force, $F_M$, is increased, the von Mises stresses also increase proportionally, because the system being analysed is linear elastic. When $F_M$ increases the stress along the line NN increases showing two peaks along the line $NN_{3-4}$ (refer to **Figure 2a**)). The larger of these two peaks occurs at a distance of ±3.8mm in length from $NN_1$. This stress peak (as can be identified in **Figure 2b**)) is caused by a sharp corner and sudden change in section at this point.

Elevated stress concentrations are identified at the beginning of the line $II_{3-4}$ (**Figure 2c**) and **Figure 2d**)). This stress peak, as can be identified in **Figure 2c**), is caused by a sharp corner at this point. For the 3i system, the volume of the crown exceeds that of the Neoss system, thereby suggesting that the 3i crown may endeavor greater resistance to the applied masticatory forces. However, even though the Neoss crown has a thinner wall thickness along the line $NN_{3-4}$, reduced stresses are still evident due to the abutments high Young's modulus. Overall, the design differences between the Neoss and 3i systems ultimately results in the 3i system having higher stresses when $F_M$ is increased.

### 3.2. Preload Force, $F_P$

To investigate the effect of different preload $F_P$, $F_M$ is kept as a constant and its medium value, i.e. 500N is considered herein. The distributions of von Mises stresses along the lines NN and II for all values of $F_P$ are shown in **Figure 3**.

As found for $F_M$, when $F_P$ increases the stresses calculated along the line NN increase, showing two peaks along the line $NN_{3-4}$ (refer to **Figure 3a**) and **Figure 3b**)). Also, as found for $F_M$, elevated stress peaks are identified at the beginning of the line $II_{3-4}$ (**Figure 3c**) and **Figure 3d**)). Overall, all values of $F_M$ cause greater stresses along lines NN and II, than do varying values of $F_P$.

## 4. DISCUSSION

FEA has been used extensively to predict the biomechanical performance of the jawbone surrounding a dental implant [21, 22]. Previous research considered the influence of the implant dimensions and the bone-implant bond on the stress in the surrounding bone. However, to date no research has been conducted to evaluate the stress produced by different implant to crown connections (i.e. internal and external-hex). The analysis completed in this paper uses the FEM to replicate internal and external-hex systems when subjected to both $F_M$ and $F_P$ loading conditions. As shown in **Table 2**, two stress peaks were revealed along the lines NN and II at locations 3.76 and 2.89mm from the top. The stress values shown were calculated with the other variables (i.e. $F_M$ or $F_P$) set to its average.

The mastication force $F_M$ is applied on the occlusal surface of the crown, evenly distributed along 378 nodal locations (**Figure 1a**), and orientated at 45° in the x-y plane. This induces compressive stresses in the right hand side of the crown and tensile in the left. Varying $F_M$ from 200 to 1000N for the internal and external-hex systems results in a change in von Mises stress of 545.64 (818.47-272.82MPa) and 698.09MPa (1047.14-349.05MPa) respectively. The geometrical design of the external-hex system tends to induce stress concentrations, located 2.89mm from the apex in this study. For this system, a stress concentration at this point is also induced by $F_P$, increasing the compressive stresses on the right hand side of the crown. Increasing $F_P$ from its minimum to maximum values, for the external-hex system, increases the stress by 485.46MPa (951.67-466.21MPa).

The internal-hex system has reduced stress concen-

**Table 2**. Von Mises stress (MPa) in crown (location of stress recording in brackets).

| Variables | $F_M$ (N) | | | $F_P$ (N) | | |
|---|---|---|---|---|---|---|
| Line | 200 | 500 | 1000 | 201.93 | 587.44 | 1009.67 |
| NN (3.76mm) | 272.82 | 545.64 | 818.47 | 231.55 | 545.64 | 891.83 |
| II (2.89mm) | 349.05 | 698.09 | 1047.14 | 466.21 | 698.09 | 951.67 |

trations, demonstrating that this design is less susceptible to stress concentrations within the crown. However, because of the transfer of the preload through the abutment screw to abutment contact, changing $F_P$ is more influential on this hex system than $F_M$. Overall $F_M$ is more influential on the stress within the crown for the external-hex system and $F_P$ is more influential on the internal-hex system.

## 5. CONCLUSION

This research is a pilot study aimed at offering an initial understanding of the stress distribution characteristics in the crown under different loading conditions. Realistic geometries, material properties, loading and support conditions for the implant system were considered in this study. The geometrical design of the external-hex system tends to induce stress concentrations in the crown at a distance of 2.89mm from the apex. At this location, $F_P$ also affects the stresses, so that the compressive stresses on the right hand side of the crown are increased. The internal-hex system has reduced stress concentrations in the crown. However, because the preload is transferred through the abutment screw to the abutment contact, changing $F_P$ has greater effect on this hex system than $F_M$. Overall $F_M$ is more influential on the stress within the crown for the external-hex system and $F_P$ is more influential on the internal-hex system.

Future recommendations include the evaluation of other implant variables such as the implant wall thickness and thread design. Ultimately, all implant components can be understood in terms of their influence on the stress produced within the implant itself.

## ACKNOWLEDGEMENTS

This work was made possible by the collaborative support from griffith's school of engineering and school of dentistry and oral health. A special thank you goes to messer john divitini and fredrik engman from neoss pty ltd for their continual contribution.

## REFERENCE

[1]Y. Maeda, T. Satoh & M. Sogo. In vitro differences of stress concentrations for internal and external hex implantabutment connections: a short communication. *Journal of Oral Rehabilitation* 2006, 33:75-78.

[2]B. R. Merz, S. Hunenbart & U. C. Belser. Mechanics of the implantabutment connection: an 8-degree taper compared to a butt joint connection. *International Journal of Oral and Maxillofacial Implants* 2000, 15:519-526.

[3]A. Khraisat, R. Stegaroiu, S. Nomura & O. Miyakawa. Fatigue resistance of two implant/abutment joint designs. *Journal of Prosthetic Dentistry* 2002, 88:604-610.

[4]S. Capodiferro, G. Favia, M. Scivetti, G. De Frenza & R. Grassi. Clinical management and microscopic characterisation of fatique-induced failure of a dental implant. *Case report. Head and Face Medicine* 2006, 22, 2:18.

[5]P. Gehrke, G. Dhom, J. Brunner, D. Wolf, M. Degidi & A. Piattelli. Zirconium implant abutments: fracture strength and influence of cyclic loading on retaining-screw loosening. *Quintessence International* 2006, 37, 1:19-26.

[6]A. Khraisat. Stability of implant-abutment interface with a hexagon-mediated butt joint: failure mode and bending resistance. *Clinical Implant Dentistry and Related Research* 2005, 7(4):221-228.

[7]F. H. G. Butz, M. Okutan & J. R. Strub. Survival rate, fracture strength and failure mode of ceramic implant abutments after chewing simulation. *Journal of Oral Rehabilitation* 2005, 32(11):838-843.

[8]K. J. Anusavice & P. H. Dehoff. Influence of metal thickness on stress distribution in metal-ceramic crowns. *Journal of Dental Research* 1986, 65(9):1173-1178.

[9]K. J. Anusavice. Stress distribution in metal-ceramic crowns with a facial porcelain margin. *Journal of Dental Research* 1987, 66(9):1493-1498.

[10]K. J. Anusavice. Influence of incisal length of ceramic and loading orientation on stress distribution in ceramic crowns. *Journal of Dental Research* 1988, 67(11):1371-1375.

[11]H. Y. Suzuki. Finite element stress analysis of ceramics crown on premolar. Relation between ceramics materials and abutment materials. *Nippon Hotetsu Shika Gakkai Zasshi* 1989, 33(2):283-293.

[12]T. Hino. A mechanical study on new ceramic crowns and bridges for clinical use. *Osaka Daigaku Shigaku Zasshi* 1990, 35(1):240-267.

[13]Zhang, B. & Wang, H. Three-dimensional finite element analysis of all-ceramic crowns of the posterior teeth. *Hua Xi Yi Ke Da Xue Xue Bao* 2000, 31(2):147-148.

[14]K. A. Proos, J. Ironside & G. P. Steven. Finite element analysis studies of an all-ceramic crown on a first premolar. International *Journal of Prosthodontics* 2002, 15(4):404-412.

[15]A. Imanishi, T. Ohyama & T. Nakamura. 3-D Finite element analysis of all-ceramic posterior crowns. *Journal of Oral Rehabilitation* 2003, 30(8):818-822.

[16]Strand7 Pty Ltd. *Strand7 Theoretical Manual* 2004. Sydney, Australia.

[17]Neoss Pty Ltd, *Neoss Implant System Surgical Guidelines* 2006. United Kingdom.

[18]http://www.3i-online.com.htm (accessed 12th July 2006).

[19]R. C. van Staden, H. Guan, Y. C. Loo, N. W. Johnson & N. Meredith. Stress Evaluation of Dental Implant Wall Thickness using Numerical Techniques.*Applied Osseointegration Research* 2008, (In Press).

[20]A. R. Curtis, A. J. Wright & G. J. Fleming. The influence of simulated masticatory loading regimes on the bi-axial flexure strength and reliability of a Y-TZP dental ceramic. *Journal of Dentistry* 2005, 34(5):317-325.

[21]D. H. DeTolla, S. Andreana, A. Patra, R. Buhite & B. Comella. Role of the finite element model in dental implants. *Journal of Oral Implantology* 2000, 26(2):77-81.

[22]J. P. Geng, K.B. Tan & G.R. Liu. Application of finite element analysis in implant dentistry: a review of the literature. *Journal of Prosthetic Dentistry* 2001, 85(6):585-598.

J. Biomedical Science and Engineering, 2008, 1, 15-21

Scientific Research Publishing

# Construction and control of genetic regulatory networks:a multivariate Markov chain approach

Shu-Qin Zhang[1], Wai-Ki Ching[2], Yue Jiao[2], Ling-Yun Wu[3] & Raymond H. Chan[4]

[1]School of Mathematical Sciences, Fudan University, Shanghai, 200433, China. [2]Advanced Modeling and Applied Computing Laboratory, Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong. [3]Institute of Applied Mathematics, Academy of Mathematics and System Sciences, Chinese Academy of Sciences. [4]Department of Mathematics, the Chinese University of Hong Kong, Shatin, N.T., Hong Kong. *Correspondence should be addressed to Shu-Qin Zhang (zhangs@fudan.edu.cn).

## ABSTRACT

**In the post-genomic era, the construction and control of genetic regulatory networks using gene expression data is a hot research topic. Boolean networks (BNs) and its extension Probabilistic Boolean Networks (PBNs) have been served as an effective tool for this purpose. However, PBNs are difficult to be used in practice when the number of genes is large because of the huge computational cost.In this paper, we propose a simplified multivariate Markov model for approximating a PBN. The new model can preserve the strength of PBNs, the ability to capture the inter-dependence of the genes in the network, and at the same time reduce the complexity of the network and therefore the computational cost. We then present an optimal control model with hard constraints for the purpose of control/intervention of a genetic regulatory network. Numerical experimental examples based on the yeast data are given to demonstrate the effectiveness of our proposed model and control policy.**

**Keywords: Gene expression sequences; Multivariate Markov chain; Optimal control policy; Probabilistic Boolean networks**

## 1. INTRODUCTION

An important issue in systems biology is to understand the mechanism in which cells execute and control a huge number of operations for normal functions, and also the way in which the cellular systems fail in disease, eventually to design some control strategy to avoid the undesirable state/situation. Many mathematical models such as neural networks, linear model, Bayesian networks, non-linear ordinary differential equations, Petri nets, Boolean Networks (BNs) and its generalization Probabilistic Boolean Networks (PBNs), multivariate Markov chain model etc. [1,2,4,11,15,16,17,21] have been proposed. Among all the models, BNs and PBNs have received much attention. The approach is to model the genetic regulatory system by a Boolean network and infer the network structure from real gene expression data. Then by using the inferred network model, the underlying gene regulatory mechanisms can be uncovered. This is particularly useful as it helps to make useful predictions by computer simulations. We refer readers to the survey paper by Shmulevich *et al.* [18, 19] and the book by Shmulevich and Dougherty [20].

The BN model was first introduced by Kauffman [12, 13, 14]. The advantages of this model can be found in Akutsu *et al.* [1], Kauffman [14] and Shmulevich *et al.* [17]. Since genes exhibit switching behavior [10], BN models have received much attention. In a BN, each gene is regarded as a vertex of the network and is quantized into two levels only (expressed (1) or unexpressed (0)). We remark that the idea and the model can be extended easily to the case of more than two states. The target gene is predicted by several genes called its input genes through a Boolean function. If the input genes and the Boolean functions are given, a BN is defined. The only randomness involved here is the initial system state. However, the biological system has its stochastic nature and the microarray data sets used to infer the network structure are usually not accurate because of the experimental noise in the complex measurement process. Thus stochastic models are more reasonable choices. To overcome the deterministic nature of a BN, Akutsu *et al.* [1] proposed the noisy Boolean networks together with an identification algorithm. In their model, they relax the requirement of consistency imposed by the Boolean functions. Regarding the effectiveness of a Boolean formalism, Shmulevich *et al.* [17] proposed a PBN that can share the appealing rule-based properties of Boolean networks and it is robust in the presence of uncertainty. The model parameters can be estimated by using Coefficient of Determination (COD) [8].

The dynamics of the PBN can be studied in the context of standard Markov chain [17, 18, 19]. This makes the analysis of the network easy. However, the number of parameters (state of the system) grows exponentially with respect to the number of genes *n*. Therefore it is natural to develop heuristic methods for model training or to consider other approximate model. Here we propose a simplified multivariate Markov model, which can capture both the intra- and inter-associations (transition probabilities) among the gene expression sequences. The number of parameters in the model is only $O(n^2)$ where *n* is the number of genes in a captured network. We remark that this order is already minimal. We then develop efficient model parameters estimation methods based on linear programming. We further propose an optimal control formulation for regulating the network so as to avoid some undesirable states which may correspond to some disease like cancer.

The rest of the paper is structured as follows. In section 2, we present the simplified multivariate Markov model. In section 3, the estimation method for model parameters is given. In section 4, an optimal control formulation is proposed. In section 5, we apply the proposed model and method to some synthetic examples and also the gene expression dataset of yeast. Concluding remarks are then given to address further research issues in section 6.

## 2. THE MULTIVARIATE MARKOV CHAIN MODEL

In this section, we first review a multivariate Markov chain model proposed in Ching, *et al.* [3] for modeling categorical time series data. We remark that the model has been first applied to predicting demand of inventory of correlated products. Later the model was applied to the building of genetic regulatory networks [4] from gene expression data. However, the number of parameters is still large and further reduction of the model parameters is necessary and a simplified model was proposed in [5]. In the remainder of this section, we present the simplified multivariate Markov chain model.

Given *n* categorical time sequences, we assume they share the same state space *M*. We denote the state probability distribution of Sequence *j* at time *t* by $\boldsymbol{V}_t^{(j)}$, $j=1,2,\cdots,n$ . In Ching, *et al.* [3], the following first-order model was proposed to model the relationships among the sequences:

$$\mathbf{V}_{t+1}^{(i)} = \sum_{j=1}^{n} \lambda_{ij} P^{(ij)} \mathbf{V}_t^{(j)}, \quad i = 1,2,\ldots,n \qquad (1)$$

Where

$$\lambda_{ij} \geq 0 \quad \text{for} \quad 1 \leq i,j \leq n \quad \text{and} \quad \sum_{j=1}^{n} \lambda_{ij} = 1. \qquad (2)$$

Here $\lambda_{ij}$ is the non-negative weighting of Gene *j* to Gene *i*. The matrix $P^{(ij)}$ is a transition probability matrix for the transitions of states in Sequence *j* to states in Sequence *i* in one step, see for instance [3]. In matrix form we have

$$\mathbf{V}_{t+1} \equiv \begin{pmatrix} \mathbf{V}_{t+1}^{(1)} \\ \mathbf{V}_{t+1}^{(2)} \\ \vdots \\ \mathbf{V}_{t+1}^{(n)} \end{pmatrix} = Q \begin{pmatrix} \mathbf{V}_{t}^{(1)} \\ \mathbf{V}_{t}^{(2)} \\ \vdots \\ \mathbf{V}_{t}^{(n)} \end{pmatrix} \equiv Q\mathbf{V}_t$$

where

$$Q = \begin{pmatrix} \lambda_{11}P^{(11)} & \lambda_{12}P^{(12)} & \cdots & \lambda_{1n}P^{(1n)} \\ \lambda_{21}P^{(21)} & \lambda_{22}P^{(22)} & \cdots & \lambda_{2n}P^{(2n)} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{n1}P^{(n1)} & \lambda_{n2}P^{(n2)} & \cdots & \lambda_{nn}P^{(nn)} \end{pmatrix}.$$

We note that the column sum of *Q* is not equal to one (the column sum of each $P^{(ij)}$ is equal to one). The followings are two propositions [3] related to some properties of the model.

**Proposition 2.1** If $\lambda_{ij} > 0$ for $1 \leqslant i,j \leqslant n$ , then the matrix *Q* has an eigenvalue equal to 1 and the eigenvalues of *Q* have modulus less than or equal to 1.

**Proposition 2.2** Suppose that $P^{(ij)}$ $(1 \leqslant i, j \leqslant n)$ are irreducible and $\lambda_{ij} > 0$ for $1 \leqslant i,j \leqslant n$ . Then there is a vector

$$\bar{\mathbf{V}} = [\bar{\mathbf{V}}^{(1)}, \bar{\mathbf{V}}^{(2)}, \cdots, \bar{\mathbf{V}}^{(n)}]^T$$

such that

$$\bar{\mathbf{V}} = Q\bar{\mathbf{V}}$$

and

$$\sum_{i=1}^{m} [\bar{\mathbf{V}}^{(j)}]_i = 1, 1 \leq j \leq n$$

where *m* is the number of states.

In Proposition 2.2, we require all $P^{(ij)}$ are irreducible. But actually, if *Q* is irreducible, we can get the same conclusion. If the model is applied to gene expression data sequences, one may take *M*={0,1} and $\boldsymbol{V}_t^{(i)}$ to be the expression level of the *i*-th gene at the time *t*. From (1), the expression probability distribution of the *i*-th gene at time (*t*+1) depends on the weighted average of $P^{(ij)}\boldsymbol{V}_t^{(j)}$. We remark that this is a first-order model and $\lambda_{ij}$ actually give the weighting of how much Gene *i* depends on Gene *j*. In Ching, *et al.* [4], this model has been used to find cell cycles. The most proper parent genes for the *i*-th gene (i.e., $\boldsymbol{V}_{t+1}^{(i)}$ ) can be retrieved from the corresponding

$\lambda_{ij}$ . The higher the value of $\lambda_{ij}$, the stronger the parent and child relationship between *i*-th and *j*-th gene will be. When this process is repeated for each *j*, the whole genetic network can be constructed. Given a set of genes

$$\{V^{(j_h)} : h = 1,2,\ldots,w \quad \text{and} \quad j_h \in (1,2,\ldots,n)\}$$

If for any gene in this set, the rest genes are the only candidates being a corresponding parent gene, then this set of genes forms a cycle.

A simplified model was proposed in Ching *et al.* [5] by assuming

$$P^{(ij)} = I \quad \text{if} \quad i \neq j. \qquad (3)$$

The simplified model has smaller number of parameters and it has been shown to be statistically better in terms of BIC, see for instance [5]. Moreover, Propositions 1 and 2 still hold for the simplified model.

## 3. ESTIMATION OF MODEL PARAMETERS

In this section, we present methods to estimate $P^{(ij)}$ and $\lambda_{ij}$ . We estimate the transition probability matrix $P^{(ij)}$ by the following method. First we count the transition frequency of the states in the *i*-th sequence. After making a normalization, we obtain an estimate of the transition probability matrix. We have to estimate *n* such *m*-by-*m* transition probability matrices to get the estimate for $P^{(ij)}$ as follows:

$$F^{(ii)} = \begin{pmatrix} f_{11}^{(ii)} & \cdots & f_{1m}^{(ii)} \\ \vdots & \ddots & \vdots \\ f_{m1}^{(ii)} & \cdots & f_{mm}^{(ii)} \end{pmatrix},$$

From $F^{(ij)}$, one can obtain the estimate for $P^{(ij)}$ as follows:

$$\hat{P}^{(ii)} = \begin{pmatrix} \hat{p}_{11}^{(ii)} & \cdots & \hat{p}_{1m}^{(ii)} \\ \vdots & \ddots & \vdots \\ \hat{p}_{m1}^{(ii)} & \cdots & \hat{p}_{mm}^{(ii)} \end{pmatrix},$$

Where

$$\hat{p}_{ab}^{(ii)} = \begin{cases} \dfrac{f_{ab}^{(ii)}}{\sum_{a=1}^{m} f_{ab}^{(ii)}}, \text{if } \sum_{a=1}^{m} f_{ab}^{(ii)} \neq 0 \\ \dfrac{1}{m}, \text{otherwise.} \end{cases}$$

Besides $\hat{p}^{(ii)}$, we need to estimate the parameters $\lambda_{ij}$ . It can be shown that the multivariate Markov model has a "stationary vector" $\overline{\mathbf{V}}$ in Proposition 2. The vector $\overline{\mathbf{V}}$ can be estimated from the gene expression sequences by computing the proportion of the occurrence of each gene and we denote it by

$$\hat{\mathbf{V}} = (\hat{\mathbf{V}}^{(1)}, \hat{\mathbf{V}}^{(2)}, \ldots, \hat{\mathbf{V}}^{(n)})^T.$$

We therefore expect that

$$Q\hat{\mathbf{V}} \approx \hat{\mathbf{V}}.$$

$$\begin{pmatrix} \lambda_{11}\hat{P}^{(11)} & \lambda_{12}I & \cdots & \lambda_{1n}I \\ \lambda_{21}I & \lambda_{22}\hat{P}^{(22)} & \cdots & \lambda_{2n}I \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{n1}I & \lambda_{n2}I & \cdots & \lambda_{nn}\hat{P}^{(nn)} \end{pmatrix} \hat{\mathbf{V}} \approx \hat{\mathbf{V}}.$$

From the above equation, it suggests one possible way to estimate the parameters $\Lambda = \{\lambda_{ij}\}$ as follows:

$$\min_{\lambda} \max_{k} \left\| \left[ \lambda_{ii}\hat{P}^{(ii)}\hat{\mathbf{V}}^{(i)} + \sum_{j=1, i\neq j}^{n} \lambda_{ij}\hat{\mathbf{V}}^{(j)} - \hat{\mathbf{V}}^{(i)} \right]_k \right\| \qquad (4)$$

subject to

$$\sum_{j=1}^{n} \lambda_{ij} = 1, \quad \text{and} \quad \lambda_{ij} \geq 0, \quad \forall j.$$

We note that the following formulation of *n* linear programming problems can give the necessary solutions of Problem (4). For each *i*:

$$\min_{\lambda} w_i$$

Subject to

$$\begin{cases} w_i\mathbf{e} & \geq & \hat{\mathbf{V}}^{(i)} - B_i\lambda_{i,} \\ w_i\mathbf{e} & \geq & -\hat{\mathbf{V}}^{(i)} + B_i\lambda_{i,} \end{cases} \qquad (5)$$

Where

$$B_i = [\hat{\mathbf{V}}^{(1)} \mid \hat{\mathbf{V}}^{(2)} \mid \cdots \mid P^{ii}\hat{\mathbf{V}}^{(i)} \mid \cdots \mid \hat{\mathbf{V}}^{(n)}],$$

and

$$\mathbf{e} = (1,1,\ldots,1)^T.$$

Here $\lambda_{ij}$ is the *i*-th row of $\Lambda$ .

We remark that the estimation method can be applied to the simplified model (3). We remark that other vector norms such as $\|.\|_2$ and $\|.\|_1$ can also be used but they have different characteristics. The former will result in a quadratic programming problem while $\|.\|_1$ will still result in a linear programming problem. The main computation cost comes from solving the linear programming problem. In the estimation of $\hat{P}_{ii}$, it involves only counting frequencies of transitions and therefore the cost is minimal. Once the model parameters are available, one can then construct the underlying genetic network easily. We will demonstrate this in the section of numerical examples. The model can also be further modified to include extra conditions such as some $\lambda_{ij}$ are known to be zero. Such information can be included by adding the constraints $\lambda_{ij}=0$ . Furthermore, for large network, it is known that the in-degree follows the Poisson distribution while the out-degree follows the power-law, i.e., the number of out-degree to some negative power. These important properties can also be easily included in our proposed model [24].

## 4. THE OPTIMAL CONTROL FORMULATION

In this section, we present the optimal control problem based on the simplified multivariate Markov model (3) and formulate it based on the principle of dynamic programming. In the simplified model (3) we proposed above, the matrix $Q$ can be regarded as a "transition probability matrix" for the multivariate Markov chain in certain sense, and $\boldsymbol{V}_t$ can be regarded as a joint state distribution vector. We then present a control model based on the paper by Ching, *et al.*[6]. Beginning with an initial joint probability distribution $\boldsymbol{V}_0$ the gene regulatory network (or the multivariate Markov chain) evolves according to two possible transition probability matrices $Q_0$ and $Q_1$. Without any external control, we assume that the multivariate Markov chain evolves according to a fixed transition probability matrix $Q_0$ ($\equiv Q$). When a control is applied to the network at one time step, the Markov chain will evolve according to another transition probability $Q_1$ (with more favorable steady states or a more favorable state distribution). It will then return back to $Q_0$ again if there is no control. We note that one can have more than one type of controls, i.e., more than one transition probability matrix $Q_1$ to choose in each time step. For instance, in order to suppress the expression of a particular gene, one can directly toggle off this gene. One may achieve the goal indirectly by means of controlling its parent genes which have a primary impact on its expression too. But for the simplicity of discussion, we assume that there is only one direct possible control here. We then suppose that the maximum number of controls that can be applied to the network during a finite investigation period $T$ (finite-horizon) is $K$ where $K \leqslant T$. The objective here is to find an optimal control policy such that the state of the network is close to a target state vector $\mathbf{Z}$. Without loss of generality, here we focus on the first gene among all the genes. Accordingly, we remark that the sub-vector $\mathbf{Z}^{(1)}$ denotes the vector containing the first two entries in $\mathbf{Z}$. It can be a unit vector (a desirable state) or a probability distribution (a weighted average of desirable states). The control system is modeled as:

$$\mathbf{v}(i_t i_{t-1} \ldots i_1) = Q_{i_t} \cdots Q_{i_1} \mathbf{v}_0,$$
$$i_1, \ldots, i_t \in \{0,1\} \quad \text{and} \quad \sum_{j=1}^{t} i_j \le K,$$

where $\mathbf{v}(i_t\, i_{t-1} \cdots i_1)$ represents all the possible network state probability distribution vectors up to time $t$. We define

$$U(t) = \{\mathbf{v}(i_t i_{t-1} \ldots i_1) : i_1, \ldots, i_t \in \{0,1\}$$
$$\text{and} \quad \sum_{j=1}^{t} i_j \le K\}$$

to be the set which contains all the possible state probability vectors up to time $t$. We note that one can conduct a forward calculation to compute all the possible state vectors in the sets $U(1), U(2), \ldots U(\mathrm{T})$ recursively. Here the main computational cost is the matrix-vector multiplication and the cost is $O((2n)^2)$ where $n$ is the number of genes in the network. We note that some state probability distribution actually does not exist because the maximum number of controls is $K$, the total number of vectors involved is only

$$\sum_{j=0}^{K} \frac{T!}{j!(T-j)!}.$$

For example if $K$=1, the complexity of the above algorithm is $O(T(2n)^2)$.

Returning to our original problem, our purpose is to make the system go to the desirable states. The objective here is to minimize the overall average of the distances of the state vectors $\mathbf{v}(i_t \ldots i_1)$ ($t$=1,2,…,$T$) to the target vector $\mathbf{z}$, i.e.,

$$\min_{\mathbf{v}(i_T i_{T-1} \ldots i_1) \in U(T)} \frac{1}{T} \sum_{t=1}^{T} \| \mathbf{v}(i_t \ldots i_1) - \mathbf{z} \|_2 \,. \qquad (6)$$

To solve (6), we have to define the following cost function

$$D(\mathbf{v}(\mathbf{w}_t), t, k), \quad 1 \le t \le T, \quad 0 \le k \le K$$

as the minimum total distance to the terminal state at time $T$ when beginning with state distribution vector $\mathbf{v}(\mathbf{w}_t)$ at time $t$ and that the number of controls used is $k$. Here $W_t$ is a Boolean string of length $t$. Given the initial state of the system, the optimization problem can be formulated as:

$$\min_{0 \le k \le K} \{D(\mathbf{v}_0, 0, k)\} \qquad (7)$$

subject to:

$$D(\mathbf{v}(\mathbf{w}_t), t, K+1) = \infty, \quad \text{for all } \mathbf{w}_t \text{ and } t,$$
$$D(\mathbf{v}(\mathbf{w}_T), T, k) = \| \mathbf{v}(\mathbf{w}_T) - \mathbf{z} \|_2,$$
$$\text{for} \quad \mathbf{w}_T = i_T \ldots i_1, \sum_{j=1}^{T} i_j \le K, k = 0, 1, \ldots, K.$$

To solve the optimization problem, one may consider the following dynamic programming formulation:

$$D(\mathbf{v}(\mathbf{w}_{t-1}),t-1,k) =$$
$$\min\{\| \mathbf{v}(0\mathbf{w}_{t-1})-\mathbf{z}\|_2 + D(\mathbf{v}(0\mathbf{w}_{t-1}),t,k),$$
$$\| \mathbf{v}(1\mathbf{w}_{t-1})-\mathbf{z}\|_2 + D(\mathbf{v}(1\mathbf{w}_{t-1}),t,k+1)\}. \quad (8)$$

Here $0\mathbf{w}_{t-1}$ and $1\mathbf{w}_{t-1}$ are Boolean strings of size $t$. The first term in the right-hand-side of (8) is the cost (distance) when no control is applied at time $t$ while the second term is the cost when a control is applied. The optimal control policy can be obtained during the process of solving (8). We remark that instead of considering the objective (6), one can consider

$$\min_{V(i_T i_{T-1},\ldots,i_1)\in U\ T} \sum_{t=1}^{T} \alpha_t \| v(i_t \ldots i_1) - z \|_l$$

With $\{\alpha_i\}$ a new weighting and a different vector norm $\|.\|_l$. Furthermore, it is interesting to study the case of infinite horizon. In this case $\alpha_t$ is chosen to be $(1-\alpha)\alpha^{t-1}$ for some discount factor $\alpha \in (0,1)$.

## 5. NUMERICAL EXPERIMENTS

### 5.1. A Simple Example

In this subsection, we consider a small five-gene network whose gene expression series can be found in the Appendix. **Figure 1** shows the five-gene network. We note that Gene 1 and Gene 4 depends on all the other genes, Gene 2 depends on Gene 1 and Gene 3 only, Gene 3 depends on Gene 1 and Gene 2 only, while Gene 5 depends on itself only.

To solve the linear programming problem in equation (5), infinity norm is chosen for all numerical experiments. The matrices $\mathbf{\Lambda}$, $P$, and $Q_0$ (without control) are obtained from the proposed model as follow:

$$P = \begin{pmatrix} P_1 & I_2 & I_2 & I_2 & I_2 \\ I_2 & P_2 & I_2 & I_2 & I_2 \\ I_2 & I_2 & P_3 & I_2 & I_2 \\ I_2 & I_2 & I_2 & P_4 & I_2 \\ I_2 & I_2 & I_2 & I_2 & P_5 \end{pmatrix}$$

Where

$$P_1 = \begin{pmatrix} 0.6000 & 0.4286 \\ 0.4000 & 0.5714 \end{pmatrix} \quad P_2 = \begin{pmatrix} 0.2857 & 0.6667 \\ 0.7143 & 0.3333 \end{pmatrix}$$

$$P_3 = \begin{pmatrix} 0.5000 & 0.4467 \\ 0.5000 & 0.5333 \end{pmatrix} \quad P_4 = \begin{pmatrix} 0.3571 & 0.6000 \\ 0.6429 & 0.4000 \end{pmatrix}$$

$$P_5 = \begin{pmatrix} 0.4000 & 0.3158 \\ 0.6000 & 0.6842 \end{pmatrix} \quad I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 0.3369 & 0.2604 & 0.2604 & 0.1417 & 0.0005 \\ 0.5000 & 0.0000 & 0.5000 & 0.0000 & 0.0000 \\ 0.5000 & 0.5000 & 0.0000 & 0.0000 & 0.0000 \\ 0.2045 & 0.2045 & 0.2045 & 0.2028 & 0.1838 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{pmatrix}$$

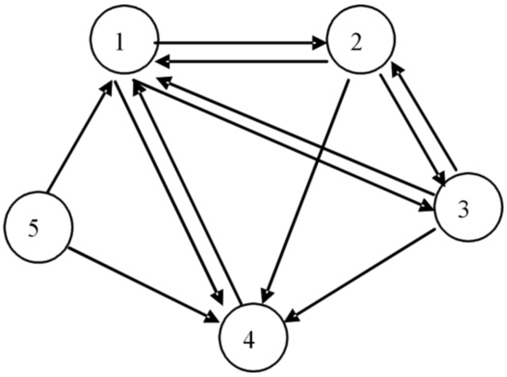


**Figure 1.** The Five-gene Network.

and

$$Q_0 = \begin{pmatrix} 0.3369P_1 & 0.2604I_2 & 0.2604I_2 & 0.1417I_2 & 0.0005I_2 \\ 0.5000I_2 & 0.0000P_2 & 0.5000I_2 & 0.0000I_2 & 0.0000I_2 \\ 0.5000I_2 & 0.5000I_2 & 0.0000P_3 & 0.0000I_2 & 0.0000I_2 \\ 0.2045I_2 & 0.2045I_2 & 0.2045I_2 & 0.2028P_4 & 0.1838I_2 \\ 0.0000I_2 & 0.0000I_2 & 0.0000I_2 & 0.0000I_2 & 1.0000P_5 \end{pmatrix}.$$

The target here is to suppress the first gene but no preference on other genes. The control we used is to suppress the first gene directly. Thus the control matrix is as follows:

$$Q_1 = \mathbf{Diag}\left(\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, I_2, I_2, I_2, I_2\right).$$

Without loss of generality, we assume that the initial state vector is the uniform distribution vector (for each gene), that is

$$\mathbf{v}_0 = \frac{1}{2}(1,1,1,1,1,1,1,1,1,1)^T.$$

Moreover, we assume that the total time $T$ is 12 and we try several different numbers of controls $K$=1,2,3,4,5. **Table 1** shows the numerical results. All the computations were done in a PC with Pentium D and Memory 1GB with MATLAB 7.0. In **Table 1**, "Policy" represents the optimal time step at the end of which a control should be applied. For instance, means that the optimal control policy is to apply the control at the end of the $t$=1,2,3-th time step. From **Table 1**, observable improvements of the optimal value is obtained when $K$ increases from 1 to 5.

### 5.2. The Yeast Example

**Table 1.** Numerical results for the 5-gene network.

| $K$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Control Policy | [1] | [2] | [1,2,3] | [1,2,3,7] | [1,2,3,7,8] |
| Objective Value | 0.5628 | 0.4277 | 0.3379 | 0.2717 | 0.2090 |
| Time in Seconds | 0.02 | 0.02 | 0.06 | 0.15 | 0.23 |

In this subsection, we apply our proposed simplified multivariate Markov models to the yeast data sequences [23]. Genome transcriptional analysis is an important analysis in medicine, etiology and bioinformatics. One of the applications of genome transcriptional analysis is used for eukaryotic cell cycle in yeast. The fundamental periodicity in eukaryotic cell cycle includes the events of DNA replication, chromosome segregation and mitosis. It is suggested that improper cell cycle regulation leads to genomic instability, especially in the etiology of both hereditary and spontaneous cancers [9, 22]. Eventually, it is believed to play one of the important roles in the etiology of both hereditary and spontaneous cancers. The dataset used in our study is the selected set from Yeung and Ruzzo (2001) [23]. In the discretization, if an expression level is above (below) a certain standard deviation from the average expression of the gene, it is over-expressed (under-expressed) and the corresponding state is 1 (0) [4].

To solve the linear programming problem in (5), infinity norm is chosen for all numerical experiments. The matrices $\Lambda$, $P$, and $Q_0$ (without control) are obtained from the proposed model. The initial state vector is assumed to be the uniform distribution (for each gene) vector

$$\mathbf{v}_0 = \frac{1}{2}(1,1,\cdots,1)^T.$$

In addition, we assume that the total time $T$ is 12 and several different maximum numbers of controls $K$=1,2,3,4,5 are tried in our numerical experiments. The target is to suppress the first gene but no preference on other genes. That is the target state vector $\mathbf{Z}^{(1)}$ is $(1,0)^T$. The control we used is to suppress the first gene directly. Thus the control matrix $Q_1$ takes the same form as the following:

$$Q_1 = \mathbf{Diag}(\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, I_2, I_2, \ldots, I_2, I_2).$$

It means that we want to control the first gene such that it will be unexpressed with more probabilities. The transitions of all the other genes will not be changed. **Table 2** reports the numerical results and the computational time for different numbers of controls $K$. From **Table 2**, observable improvements of the optimal value is obtained when $K$ increases from 1 to 5. For example, if we will conduct 4 controls totally in the 12 time steps, we need to suppress the first gene in the first 4 steps, and will not control it in other steps. These experiments show that even the number of genes (384 genes in this data set) is comparatively large, the method still can find the control policies fast.

**Table 2.** Numerical results for the yeast data set.

| $K$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Control Policy | [1] | [2] | [1,2,3] | [1,2,3,4] | [1,2,3,4,5] |
| Objective Value | 0.6430 | 0.5751 | 0.5165 | 0.4582 | 0.4000 |
| Time in Seconds | 4.00 | 20.60 | 67.90 | 152.88 | 245.95 |

## 6. CONCLUDING REMARKS

In this paper, we proposed a simplified multivariate Markov model for approximating PBNs. Efficient estimation methods based on linear programming method are presented to obtain the model parameters. Methods for recovering the structure and rules of a PBN are also illustrated in details. We then give an optimal control formulation for control the network. Numerical experiments on synthetic data and gene expression data of yeast are given to demonstrate the effectiveness of our proposed model and formulation.

For future research, we will extend the control problem to the case of having multiple control policy. We will develop efficient heuristic methods for solving the control problem and genetic algorithm is a possible approach [7]. Extension of the study to the case of infinite horizon is also interesting. Finally, we will also apply our model to more real world datasets.

## APPENDIX

The five gene expression sequences.

$$\begin{aligned} Gene1 &: 110000111110000010100011110101 \\ Gene2 &: 010101110100110010011011010100 \\ Gene3 &: 011011001100011000011111010100 \\ Gene4 &: 110101010101111001001000111001 \\ Gene5 &: 111111011011101110001010001111 \end{aligned}$$

## ACKNOWLEDGEMENTS

Research supported in part by RGC Grants 7017/07P and HKU CRCG Grants. The preliminary version of the paper has been published in the proceedings of the 2nd International Conference on Bioinformatics and Biomedical Engineering, Shanghai, China, 2008.

## REFERENCE

[1] T. Akutsu, S. Miyano & S. Kuhara. Inferring Qualitative Relations in Genetic Networks and Metabolic Arrays. *Bioinformatics* 2000, 16: 727-734.

[2] J. Bower. *Computational Modeling of Genetic and Biochemical Networks. MIT Press, Cambridge* 2001, M.A.

[3] Ching, W., E. Fung & M. Ng. A multivariate Markov Chain Model for Categorical Data Sequences and Its Applications in Demand Predictions. *IMA Journal of Management Mathematics* 2002, 13: 187-199.

[4] Ching, W., E. Fung, M. Ng & T. Akutsu. On Construction of Stochastic Genetic Networks Based on Gene Expression Sequences. *International Journal of Neural Systems* 2005, 15: 297-310.

[5] Ching, W., Zhang, S., & M. Ng. On Multi-dimensional Markov Chain Models. *Pacific Journal of Optimization* 2007, 3: 235-243.

[6] Ching, W., Zhang, S., Jiao, Y., T. Akutsu & Wong, A. *Optimal Finite-Horizon Control for Probabilistic Boolean Networks*

*with Hard Constraints*. *The International Symposium on Optimization and Systems Biology* 2007.

[7]Ching, W., H. Leung, Tsing, N. & Zhang, S. *Optimal Control for Probabilistic Boolean Networks : Genetic Algorithm Approach,* 2008.

[8]E. Dougherty, S. Kim & Chen, Y. Coefficient of Determination in Nonlinear Signal Processing. *Signal Processing* 2000, 80: 2219-2235.

[9]M. Hall, & G. Peters. Genetic Alterations of Cyclins, Cyclin-dependent Kinases, and Cdk Inhibitors in Human Cancer. *Adv. Cancer Res.* 1996, 68: 67-108.

[10]Huang, S. & D.E. Ingber. Shape-dependent Control of Cell Growth, Differentiation, and Apoptosis: Switching Between Attractors in Cell Regulatory Networks. *Exp. Cell Res.* 2000, 261: 91-103.

[11]H. de Jong. Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. *J. Comput. Biol.* 2002, 9: 69-103.

[12]S. Kauffman. Metabolic Stability and Epigenesis in Randomly Constructed Gene Nets. *J. Theoret. Biol.* 1969, 22: 437-467.

[13]S. Kauffman. Homeostasis and Differentiation in Random Genetic Control Networks. *Nature* 1969, 224: 177-178.

[14]S. Kauffman. *The Origin of Orders*, Oxford University Press, New York, 1993.

[15]S. Kim, S. Imoto & S. Miyano. Dynamic Bayesian Network and Nonparametric Regression for Nonlinear Modeling of Gene Networks from time Series Gene Expression Data. *Proc. 1st Computational Methods in Systems Biology, Lecture Note in Computer Science* 2003, 2602: 104-113.

[16]F. Nir, L. Michal, N. Iftach & P. Dana. Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology* 2000, 7(3-4): 601-620.

[17]I. Shmulevich, E. Dougherty, S. Kim & W. Zhang. Probabilistic Boolean Networks: A Rule-based Uncertainty Model for Gene Regulatory Networks. *Bioinformatics* 2002, 18: 261-274.

[18]I. Shmulevich, E. Dougherty, S. Kim & Zhang, W. Control of Stationary Behavior in Probabilistic Boolean Networks by Means of Structural Intervention. *Journal of Biological Systems* 2002, 10: 431-445.

[19]I. Shmulevich, E. Dougherty, S. Kim & W. Zhang. From Boolean to Probabilistic Boolean Networks as Models of Genetic Regulatory Networks. *Proceedings of the IEEE* 2002, 90: 1778-1792.

[20]I. Shmulevich & E. Dougherty. *Genomic Signal Processing*, Princeton University Press, USA, 2007.

[21]P. Smolen, D. Baxter & J. Byrne. Mathematical Modeling of Gene Network. *Neuron* 2002, 26: 567-580.

[22]Wang, T. C., R.D. Cardiff, L. Zukerberg, E. Lees, A. Amold & E.V. Schmidt. Mammary Hyerplasia and Carcinoma in MMTV-cyclin D1 Transgenic Mice. *Nature* 1994, 369: 669-671.

[23]K. Yeung & W. Ruzzo. An Empirical Study on Principal Component Analysis for Clustering Gene Expression Data. *Bioinformatics* 2001, 17: 763-774.

[24]Zhang, S., Ching, W., Tsing, N., H. Leung & Guo, D. *A. Multiple Regression Approach for Building Genetic Networks. The Proceedings of the International Conference on BioMedical Engineering and Informatics* 2008, Sanya, China.

J. Biomedical Science and Engineering, 2008, 1, 22-26

Scientific
Research
Publishing

# Compression of ECG signal using video codec technology-like scheme

**Di-Hu Chen * & Sheng Yang**

Department of Precision Machinery & Instrumentation, University of Science and Technology of China, Hefei 230027, China. * Correspondence should be addressed to Di-Hu Chen (dhchen@mail.ustc.edu.cn).

## ABSTRACT

**In this paper, we present a method using video codec technology to compress ECG signals. This method exploits both intra-beat and inter-beat correlations of the ECG signals to achieve high compression ratios (CR) and a low percent root mean square difference (PRD). Since ECG signals have both intra-beat and inter-beat redundancies like video signals, which have both intra-frame and inter-frame correlation, video codec technology can be used for ECG compression. In order to do this, some pre-process will be needed. The ECG signals should firstly be segmented and normalized to a sequence of beat cycles with the same length, and then these beat cycles can be treated as picture frames and compressed with video codec technology. We have used records from MIT-BIH arrhythmia database to evaluate our algorithm. Results show that, besides compression efficiently, this algorithm has the advantages of resolution adjustable, random access and flexibility for irregular period and QRS false detection.**

**Keywords: ECG compression; Video CODEC; QRS detection; Arithmetic coding**

## 1. INTRODUCTION

The electrocardiogram (ECG) is an important tool for diagnosis of heart diseases. The volume of ECG data produced by modern monitoring system can be quite large over a long period of time and data compression is often needed for efficient process, store and transmit of such data. In the past, many ECG compression methods were proposed and could be classified into three major categories [1]: a) Parameter extraction techniques. b) Transform-domain techniques. c) Direct time-domain techniques.

In this paper, we present a method for compression of ECG data using video codec technology. Since ECG signals have both intra-beat and inter-beat correlations like video signals with intra-frame and inter-frame correlations, video codec technology can be used for ECG compression. For ECG signals, there is a little difference, so some pre-process will be needed: ECG signals should be segmented and period normalized to a sequence of beat cycles with the same size. Then these beat cycles can be treated as 'picture frames' and compressed with a video codec.

In this work, we present a method using video codec technology to compress ECG signals. This method exploits both intra-beat and inter-beat correlations of the ECG signals to achieve high compression ratios (CR) and a low percent root mean square difference (PRD). Although video codec technology was developed to compress video signals, it can be used to compress other signals as well, and we illustrate how video codec technology can be used to compress ECG signals. In Section II, we take a brief overview of video codec technology. Section III presents the coding algorithm. Experimental results and comparisons with other algorithm are presented in Section IV. At last, we provide conclusions.

## 2. OVERVIEW OF VIDEO CODEC TECHNOLOGY

Representing video material in a digital form requires a large number of bits. The volume of data generated by digitizing a video signal is too large for most storage and transmission systems. This means that compression is essential for most digital video applications. Statistical analysis of video signals indicates that there is a strong correlation both between successive picture frames and within the picture elements themselves. Theoretically decorrelation of these signals can lead to bandwidth compression without significantly affecting image resolution. A video signal consists of a sequence of individual frames. Each frame may be compressed individually using an image CODEC, such as JPEG. This is described as intra-frame coding for each frame is intra coded without any reference to other frames. However, better compression performance may be achieved by exploiting the temporal redundancy in a video sequence or the similarities between successive

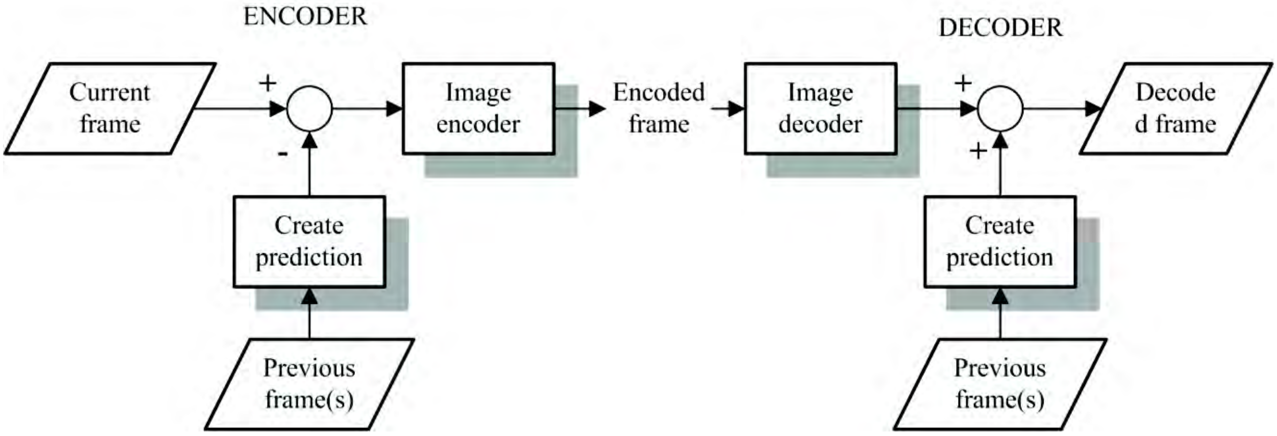


**Figure 1.** Video CODEC with prediction.

video frames. This may be achieved by introducing two functions: 1. Prediction: create a prediction of the current frame based on one or more previously transmitted frames. 2. Compensation: subtract the prediction from the current frame to produce a residual frame. Then the residual frame is compressed by an image CODEC. In order to decode the frame the decoder adds the prediction to the decoded residual frame. This is described as inter-frame coding for frames are coded based on some relationship with other video frames. **Figure 1** shows the process above.

## 3. METHOD

### 3.1. System overview

The redundancies in ECG signals can be broadly classified into two types: The redundancies in a single ECG cycle and the redundancies across ECG cycles. These redundancies are sometimes referred to as intra-beat and inter-beat redundancies [2]. These are the same with redundancies in video signals. On the other hand, there is a little difference between video signals and ECG signals: A video signal consists of a sequence of individual frames and these frames are of the same size. But for ECG signals, these 'frames' or beat cycles are jointed together, and even the sizes of them are not the same. The comparability of the ECG signals and video signals motivates us to design a novel ECG compression scheme using video codec technology, in which the scheme employs the arithmetic coding for intra-beat redundancies, and a predictor using cross correlation for inter-beat redundancies.

The functional block diagram of the proposed coding scheme is shown in **Figure 2**. The encoder system consists mainly four parts: segmentation, period normalization, predictor and residual coding. The proposed encoding algorithm is briefly described as follows. Since ECG signals are continuous and in order to use compress them using a video codec scheme, firstly we should segment them to a sequence of cycles, by noting that the length of each beat cycle may be varying, a period normalization process is then proceeded to ensure that the size of each beat cycle is adjusted to be the same. Initially, the counter is set to zero and we select the first cycle as the pre-

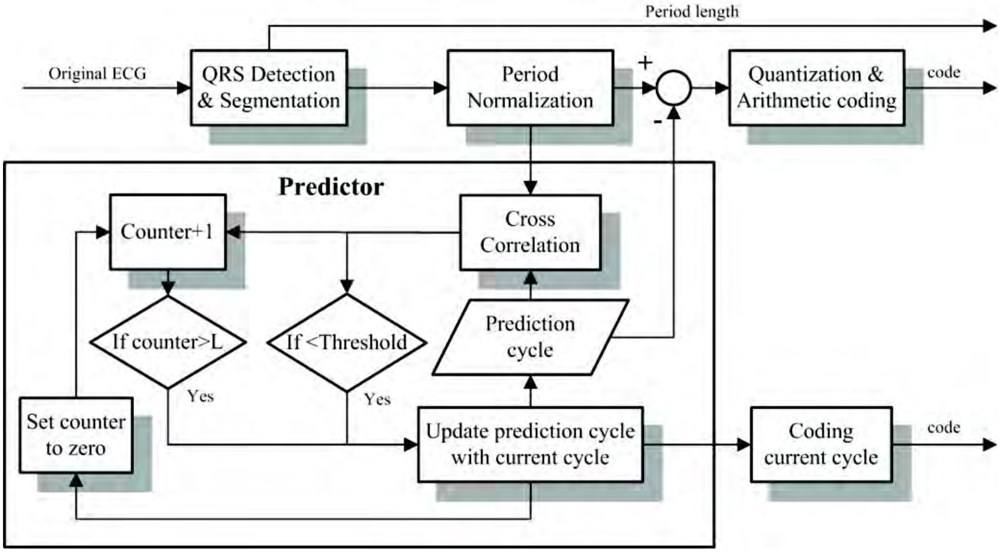


**Figure 2.** Functional block diagram of the encoder.

diction cycle and compress this cycle with no prediction, then any time when there is a new cycle, the counter is added by one and the cross correlation coefficient of the new cycle and the prediction cycle is calculated. If the result is less than the threshold, which indicates that this new cycle and the prediction cycle have little similarity, or the count is larger than *L* (used for random access), we set the counter to zero and set this new cycle as the prediction cycle and compress it with no prediction, else the prediction cycle is subtracted from this new cycle, and the residual cycle is then quantized and compressed with the arithmetic coding.

### 3.2. QRS detection and segmentation

To cut continuous ECG signals to individual beats, the peaks of QRS waves should be detected firstly to identify each heartbeat. We use a different method to do this: Let $x(i)$ denote the ECG signal, and a corresponding different signal $x'(i)$ is given by

$$x'(i) = 2x(i) - x(i+n) - x(i-n) \qquad (1)$$

where $n$ is a small integer determined by the sampling frequency (typically a value between $0.01f$ and $0.02f$ is used, where $f$ is the sampling frequency). Several zero points are added to the front and the end of the ECG signals for calculation of the first and last few points of $x'(i)$. When select proper $n$ for different sample frequency, (1) is like a band pass filter. It makes the QRS waves be amplified and the other waves be weaken. **Figure 3** shows a typical ECG signal and its corresponding difference signal generating by (1). The sample frequency is 360Hz with $n$ equals to 5.

For the different signal $x'(i)$, we can use a similar scan algorithm in [3] for the QRS detection. Results show that, our method has a higher detection rate.

After each QRS peak of heartbeat cycles is identified, the original ECG signal is cut at every QRS peak.

### 3.3. Period normalization

Since each ECG period can have a different duration, and in order to compress them using video codec technology, we normalize each ECG period to the same length. We implement this using a method similar to the one described in [4]. Let $x_k=[x_k(1)x_k(2)\cdots x_k(N_k)]$ denote the $k$-th ECG cycle. Then the period-normalized ECG cycle $y_k=[y_k(1)y_k(2)\cdots y_k(N)]$ is computing using

$$y_k(n) = \tilde{x}_k(t') \qquad (2)$$

Where $\tilde{x}_k(t')$ is an interpolate version of the samples $x_k(n)$, and $t'=\frac{(n-1)(N_k-1)}{N-1}+1$, $N_k$ is the period of the $k$-th ECG cycle, and $N$ is the normalized period. We utilize cubic-spline interpolation [5] to determine $\tilde{x}_k(t')$.

The $N$ above can be thought as the resolution, like the spatial resolution (typically 352×288 or 352×240 pixels in MPEG-1) in a video encoder. The value of $N$ is predefined in consideration of the sample frequency and it can affect the CR and the PRD.

After period normalization, each ECG period will be with the same length like video frames with the same size. Then we can use similar video CODEC technology to compress them.

### 3.4. Prediction

In part 2 we know that, in order to exploit the similarities between successive video frames, two functions prediction and compensation are introduced. The key to this approach is the prediction function: if the prediction is accurate, the residual frame will be containing little data and will hence be compressed to a very small size by the image CODEC.

For video compression, the simplest predictor is just the previous transmitted frame. We can utilize this in ECG compression. Since successive ECG cycles are very similar all the times, we make a small change and introduce the cross correlation coefficient. Cross correlation coefficient is a standard method of estimating the degree to which two series are correlated. Consider two series $x_i$ and $y_i$ where $i$=0,1,2... $N-1$, the cross correlation coefficient is defined as

$$r = \frac{\sum_i [(x_i-\bar{x})(y_i-\bar{y})]}{\sqrt{\sum_i (x_i-\bar{x})^2}\sqrt{\sum_i (y_i-\bar{y})^2}} \qquad (3)$$

Where $\bar{x}$ and $\bar{y}$ are the means of the corresponding series.

Prediction with cross correlation is shown in **Figure 2**. Initially we set the counter to zero. The first ECG beat cycle is set as the prediction beat cycle and compressed with no prediction. Any time when there is a new beat cycle, the counter is added by one and the cross correlation coefficient of the new beat cycle and the prediction beat cycle is calculated. If the counter is smaller than $L$ (predefined for random access) and the correlation result is higher than the threshold (typically 0.95 or more), which indicates that the prediction beat is similar with the current beat to a great extent, then we use it as the prediction

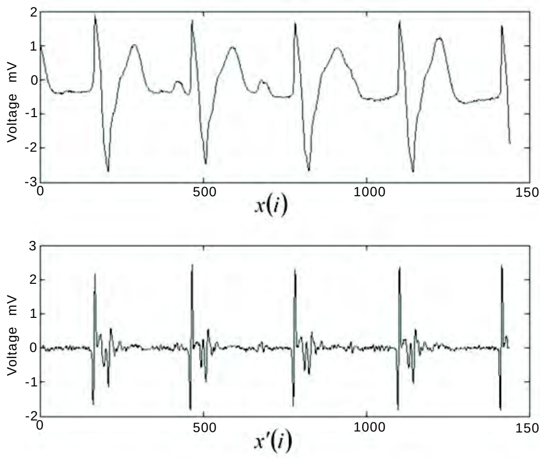


**Figure 3.** ECG signal and corresponding different signal.

of the current beat. Otherwise, we use the current beat to replace the prediction beat and compress it with no prediction and set the counter to zero again.

### 3.5. Quantization and Coding

The quantization stage removes less important information, such as information that does not have a significant influence on the appearance of the reconstructed ECG signals, making it possible to compress the remaining data.

In this paper, we use the arithmetic coding [6] for compression of the residual signal and the period information. An arithmetic encoder converts a sequence of data symbols in to a single fractional number. The longer the sequence of the symbols, the greater the precision required to represent the fractional number. Arithmetic coding provides a practical alternative to Huffman coding and can more closely approach the theoretical maximum compression [7].

### 3.6. Coding of beat cycles

In the video coding standard MPEG-1, each frame of video is encoded to produce a coded picture. There are three main types: I-pictures, P-pictures and B-pictures. I-pictures are intra-coded without any motion-compensated prediction. An I-picture is used as a reference for further predicted pictures. P-pictures are inter-coded using motion-compensated prediction from a reference picture. B-pictures are inter-coded using motion-compensated prediction from two reference pictures, the P- and/or I-pictures before and after the current B-picture. However, in our proposed scheme for ECG compression, we only introduced two types: I-cycles and P-cycles.

I-cycles are useful resynchronization points in the coded bit stream: because it is coded without prediction, an I-cycle may be decoded independently of any other coded cycles. This support random access by a decoder in some degree (a decoder may start decoding the bit stream at any I-cycle position). However, an I-cycle has poor compression efficiency because no prediction is used.

In MPEG-1 due to the existence of several picture types, a group of pictures (GOP) is the highest level of the hierarchy. A GOP is a series of one or more picture to assist randomly access into the picture sequence. The first coded picture in the group is an I-picture. It is followed by an arrangement for P- and B-pictures. Likewise, we introduce the group of cycles in our scheme to assist random access into the ECG data. The group of cycles length is defined as the distance between I-cycles, which is represented by parameter *L* in **Figure 2**. A short group of cycles may support random access well at the cost reducing the compression ratio. **Figure 4** shows a typical group of cycles.

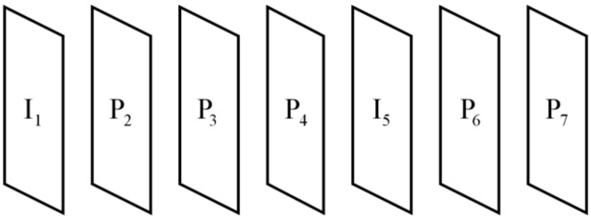


**Figure 4.** Group of cycles in coded bit stream.

## 4. RESULT

We used the MIT-BIH arrhythmia database to evaluate the performance of the proposed scheme. The ECG data used in our experiments are sampled at 360 Hz and each sample has a resolution of 12 bit per sample. Through period normalization, we have made the number of samples in each beat cycle equal 240. Although for a typical hart rate of 75 beat per minute, 288 samples in each beat cycle will be good, but a relative small samples will increase compression ratio without obviously affecting the reconstruction quality.

We use two widely used measures, the compression ratio (CR) and the percent root mean square difference (PRD) to evaluate our scheme. The CR and PRD are defined as

$$CR = B_{ori} / B_{cp} \qquad (4)$$

Where $B_{ori}$ is the total bits of the original ECG signal, $B_{cp}$ is the total bits of the ECG signal after compression.

$$PRD = 100 \times \sqrt{\sum_{i=1}^{n} [x_{ori}(i) - x_{rec}(i)]^2 \Big/ \sum_{i=1}^{n} x_{ori}^2(i)} \qquad (5)$$

Where $\boldsymbol{x}_{ori}$ and $\boldsymbol{x}_{rec}$ are the original and the reconstructed ECG signals, and *n* denotes the length of the signals.

**Figure 5** and **Figure 6** show example of ECG data from record 117 and record 119 with irregular period before and after compression.

In **Table 1**, the proposed method is compared with other methods in literature for record 117 and 119.

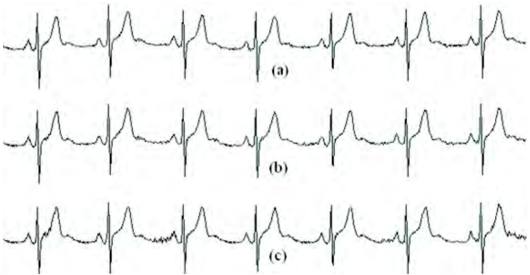


**Figure 5.** Reconstruction example of MIT-BIH record 117 with quantization level of 10 $\mu V$ and 20 $\mu V$ : (a) original signal of channel 1, (b) reconstruction signal of channel 1 with quantization level of 10$\mu V$ , CR=16 and PRD=2.87, (c) reconstruction signal of channel 1 with quantization level of 20 $\mu V$ ,CR=30.79 and PRD=5.50.

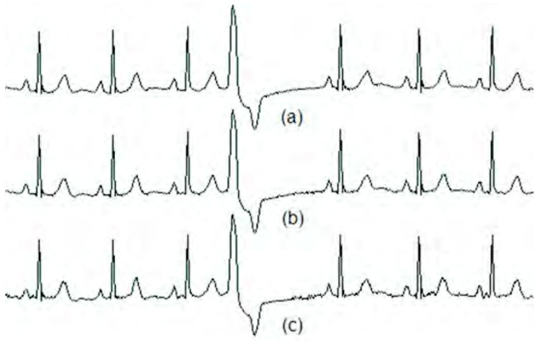


**Figure 6.** Reconstruction example of MIT-BIH record 119 with quantization level of 10 $\mu V$ and 20 $\mu V$ : (a) original signal of channel 1, (b) reconstruction signal of channel 1 with quantization level of 10 $\mu V$ , CR=14.2 and PRD=3.03, (c) reconstruction signal of channel 1 with quantization level of 20 $\mu V$ ,CR=24.2 and PRD=6.25.

**Table1.** PRD comparison of different algorithms for record 117 and 119.

| Algorithm | Record | CR | PRD (%) |
|---|---|---|---|
| Lu *et. al*[8] | 117 | 8:1 | 1.18 |
| Hilton[9] | 117 | 8:1 | 2.6 |
| Djohan *et. al*[10] | 117 | 8:1 | 3.9 |
| Proposed | 117 | 8.1:1 | 1.13 |
| Proposed | 117 | 16:1 | 2.87 |
| Proposed | 117 | 30.8:1 | 5.5 |
| Lee *et.al*[1] | 119 | 24 | 10.5 |
| Lu *et. al*[8] | 119 | 21.6 | 5.5 |
| Proposed | 119 | 14.2 | 3.03 |
| Proposed | 119 | 24.2 | 6.25 |

## 5. CONCLUSION

The main contribution of this paper is to provide an effective and efficient ECG compression scheme using video codec technology. We have tested the performance of the proposed scheme by compressing record from the MIT-BIH arrhythmia database and compared the results with other methods. The results show that the proposed algorithm compares favorable to other methods in literature. Besides compression efficiently, the proposed algorithm benefits from characteristics of the video codec and has the following advantages: a) Resolution adjustable. By changing the length *N* in section 3.3, we can achieve different resolution just like spatial resolution in a video codec; b) Random accessible. In coding stream of the ECG data, the I-cycles are intra-coded without any prediction, thus we can access the ECG data from every I-cycle. c) Flexibility for irregular period and QRS false detection. In our scheme, the irregular periods or the QRS false detection beat cycles will be treated as the new prediction cycles and compressed with no prediction if they don't have enough similarity with the formal prediction cycle.

## REFERENCE

[1] H. Lee & K. M. Buckley. ECG data compression using cut and align beats approach and 2-D transforms. *IEEE Trans-Biomed. Eng*. 1999, (46):556-565.

[2]Ali Bilgin & W. Marcellin. Compression of electrocardiogram signals using JPEG2000. *IEEE Transaction on Consumer Electronics*. 2003, 49(4).

[3] Engelse, W.A.H. & Zeelenberg, C. (). A single scan algorithm for QRS detection and feature extraction. *IEEE Computers in Cardiology* 1979, pages 37-42.

[4] Wei, J. J., Chang, C. J., Chou, N. K. & Jan, G. J. ECG data compression using truncated singular value decomposition. *IEEE Trans. on Information Technology in Biomedicine* 2001, 5:290-299.

[5] T. M. Lehman, C. Gonner, & K. Spitzer. Survey: interpolation methods in medical image processing. *IEEE Trans. on Medical Imaging* 1999, 18:1049-1075.

[6] James A. Storer, *ed*. Practical implementations of arithmetic coding. *Image and text compression*, MA, 1992 pages 85-112.

[7] I. Witten, R. Neal & J. Cleary. Arithmetic coding for data compression. *Communications of the ACM* 1987, 30(6).

[8] Lu, Z., D. Y. Kim, & W. A. Pearlman. Wavelet compression of ECG signals by the set partitioning in hierarchical trees algorithm. *IEEE Trans. on Biomedical Engineering* 2000, 47:849-856.

[9] M. L. Hilton. Wavelet and wavelet packet compression of electrocardiograms. *IEEE Trans. on Biomedical Engineering* 1997, 44:394-402.

[10] A. Djohan, T. Q. Nguyen & W. J. Tompkins. ECG compression using discrete symmetric wavelet transform. *Proc. of 17th Int. IEEE Med. Biol. Conf*. 1995.

# Possible roles of electrical synapse in temporal information processing: a computational study

**Xu-Long Wang, Xiao-Dong Jiang & Pei-Ji Liang** *

Department of Biomedical Engineering, Shanghai Jiao Tong University. *Correspondence should be addressed to Pei-Ji Liang (pjliang@sjtu.edu.cn).

## ABSTRACT

**Temporal information processing in the range of tens to hundreds of milliseconds is critical in many forms of sensory and motor tasks. However, little has been known about the neural mechanisms of temporal information processing. Experimental observations indicate that sensory neurons of the nervous system do not show selective response to temporal properties of external stimuli. On the other hand, temporal selective neurons in the cortex have been reported in many species. Thus, processes which realize the temporal-to-spatial transformation of neuronal activities might be required for temporal information processing. In the present study, we propose a computational model to explore possible roles of electrical synapses in processing the duration of external stimuli. Firstly, we construct a small-scale network with neurons interconnected by electrical synapses in addition to chemical synapses. Basic properties of this small-scale neural network in processing duration information are analyzed. Secondly, a large-scale neural network which is more biologically realistic is further explored. Our results suggest that neural networks with electrical synapses functioning together with chemical synapses can effectively work for the temporal-to-spatial transformation of neuronal activities, and the spatially distributed sequential neural activities can potentially represent temporal information.**

**Keywords: Model; Neural network; Electrical synapse; Temporal information processing**

## 1. INTRODUCTION

Biological neural systems are endowed with the ability to process temporal information given the inherent temporal nature of sensory environments and motor tasks. Neuroscientists roughly categorize temporal information processing in the neural system into four different time scales: microseconds, milliseconds, seconds and circadian rhythm, which serve for different physiological functions and rely on different neural mechanisms. The process within the scale of millisecond is perhaps the most sophisticated and the least well understood one among these categories. Behavioral tasks with temporal information processing falling within this scale include speech discrimination in the auditory system, motion information processing in the visual systems, and movement coordination in the motor system [1-3].

Information processing in neural systems normally consists of a number of successive stages. Neural activities in a certain stage are mostly determined by neural activities of the preceding stages and our perception of the world in the brain is based on the spatio-temporal patterns of neuronal activities produced at sensory stages [4-5]. Physiological observations indicate that neurons in the sensory levels do not respond selectively to the temporal properties of external stimuli. Temporal information is thus suggested to be contained in the temporal patterns of neuronal activities in the sensory layer. On the other hand, neurons which show selective response to specific temporal properties, especially the duration content, have been reported in the cortex of many species [6-10]. Temporal information is therefore suggested to be transformed into the spatially distributed neuronal activities in the cortex and neural mechanisms which contribute to the temporalto-spatial transformation of neuronal activities are required.

Electrical synapse is another type of widely distributed neuronal connection in the neural systems in addition to chemical synapse [11-12]. Functional role of electrical synapse has been identified in fine motor coordination which requires temporal information processing in milliseconds scale [13]. In the present work, we try to explore possible neural mechanisms of electrical synapse in processing the duration content of external stimuli via computational approach. Briefly, we construct neural networks containing both electrical and chemical synapses, which are activated by stimuli with various durations. Computational results show that electrical synapse can substantially contribute to the temporal-to-spatial transformation of neuronal activities, and the neuronal activities in such networks can potentially represent information about stimulus durations.

## 2. MODELS AND METHODS

### 2.1. Model structure

Two types of computational models are constructed. One is a small-scale neural network which contains only several tens of neurons. Another is a large-scale one which is more biologically realistic. We use the simple model to clarify the basic properties of neural networks with electrical synapses functioning together with chemical synapse in temporal information processing. The overall behavior is further tested in the large-scale model which is more biologically realistic.

The schematic structures of the small- and large-scale neural networks are illustrated in **Figure 1**, A and B respectively. Stimuli with various durations are applied, as represented by various durations of the input currents. The input current is injected to an input neuron (S) and then transformed into spike trains of this neuron.

The input neuron is connected to some of the ten excitatory neurons (E) in the small-scale model. Electrical synapses are presented among assigned neurons, as indicated in the figure. Excitatory neurons are connected to each other recurrently by chemical synapses and each excitatory neuron is further coupled with an inhibitory neuron (I) to ensure its stability. Parameters used in the small-scale model are listed as follows:

$IS_{ip}$: Intensity of the input current;

$CS_{se}$: Strength of chemical synapse from input neuron to excitatory neurons;

$CS_{ee}$: Strength of chemical synapse between excitatory neurons;

$ES_{ee}$: Strength of electrical synapse between excitatory neurons;

$CS_{ei}$: Strength of chemical synapse from excitatory neurons to inhibitory neurons;

$CS_{ie}$: Strength of chemical synapse from inhibitory neurons to excitatory neurons.

The large-scale neural network model contains 400 excitatory neurons and 100 inhibitory neurons. The ratio between the excitatory and inhibitory neurons follows the experimental observations from neocortical area [14]. The neural network is further divided into 100 subgroups with each subgroup consisting of 4 excitatory neurons and 1 inhibitory neuron. Excitatory and inhibitory neurons in each individual subgroup are connected recurrently. Input neuron is connected to excitatory and inhibitory neurons on a random basis. All excitatory neurons are further connected with each other probabilistically in a recurrent way, and the synaptic strengths are variables which follow normal distributions. Parameters used for synaptic connections in the extended model are listed as follows:

$CP_{se}$: Probability of chemical synapse from input to excitatory neurons;

$CM_{se}$ and $CD_{se}$: Mean and standard deviation of strength of chemical synapse from input to excitatory neurons;

$CP_{si}$: Probability of chemical synapse from input

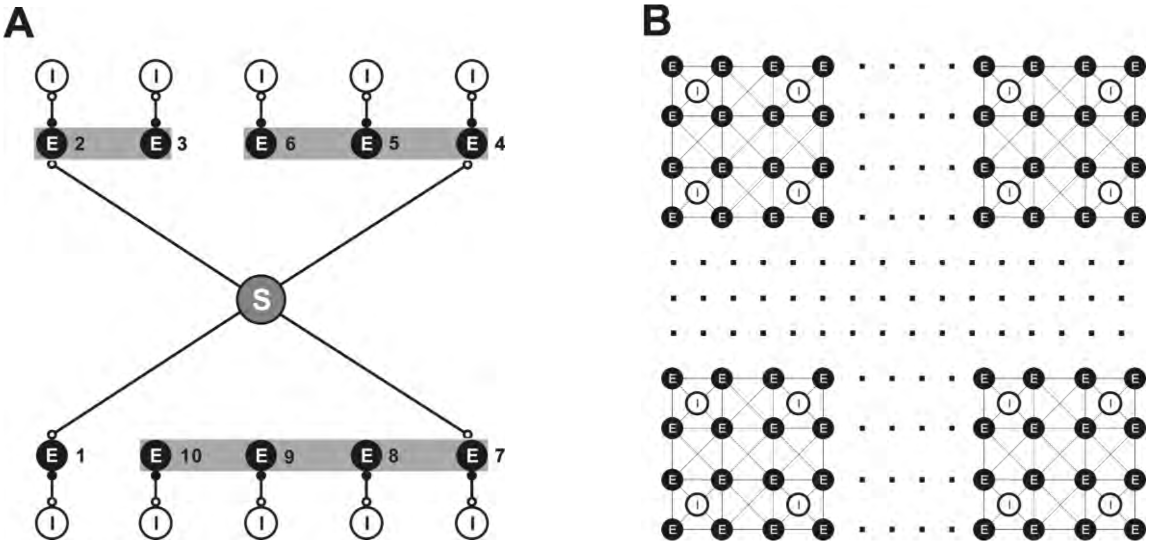


**Figure 1.** A. Schematic structure of the small-scale neural network model. The input neuron (S) is connected to 4 of the 10 the excitatory neurons (E). All excitatory neurons are connected to each other in a recurrent way and each excitatory neuron is coupled with an inhibitory neuron (I). Excitatory and inhibitory synapses are represented by open and solid circles, respectively. Neurons in grey shadow are electrically coupled together recurrently. B. Schematic structure of the large-scale neural network model. Input neuron is connected to excitatory (E) and inhibitory (I) neurons in the network on a random basis. All excitatory neurons are further connected with each other probabilistically in a recurrent way. Electrical synapses are formed between some of the excitatory neurons randomly.

to inhibitory neurons;

$CM_{si}$ and $CD_{si}$: Mean and standard deviation of strength of chemical synapse from input to inhibitory neurons;

$CP_{ee}$: Probability of chemical synapse between excitatory neurons;

$CM_{ee}$ and $CD_{ee}$: Mean and standard deviation of strength of chemical synapse between excitatory neurons;

$CM_{ei}$ and $CD_{ei}$: Mean and standard deviation of strength of chemical synapse from excitatory to inhibitory neurons;

$CM_{ie}$ and $CD_{ie}$: Mean and standard deviation of strength of chemical synapse from inhibitory to excitatory neurons;

$EP_{ee1}$: Probability of electrical connection between excitatory neurons within one subgroup;

$EP_{ee2}$: Probability of electrical connection between excitatory neurons in different subgroups;

$EM_{ee}$ and $ED_{ee}$: Mean and standard deviation of strength of electrical synapse between excitatory neurons.

## 2.2. Mathematical description of neurons and synapses

### 2.2.1 Description of integrate-and-fire neuron

Neurons are described in an integrate-and-fire manner (I-F neuron) [5]. Membrane potential of the input neuron ($V_s$), excitatory neuron ($V_{Ex}$), and inhibitory neuron ($V_{In}$) can be determined as follows:

$$C\cdot\frac{dV_S}{dt}=g_{leak}\cdot(V_{eq}-V_S)+I_S \qquad (1)$$

$$\begin{aligned}C\cdot\frac{dV_{Ex}}{dt}&=g_{leak}\cdot(V_{eq}-V_{Ex})+[g_{ex}(t)\cdot(E_{ex}-V_{Ex})\\&\quad+g_{in}(t)\cdot(E_{in}-V_{Ex})]+I_{esyn}\end{aligned} \qquad (2)$$

$$\begin{aligned}C\cdot\frac{dV_{In}}{dt}&=g_{leak}\cdot(V_{eq}-V_{In})+[g_{ex}(t)\cdot(E_{ex}-V_{In})\\&\quad+g_{in}(t)\cdot(E_{in}-V_{In})]\end{aligned} \qquad (3)$$

where

$C$ represents the membrane capacitance;

$V_{eq}$ denotes the equilibrium membrane potential;

$g_{leak}$ is the leak conductance;

$g_{ex}$ and $g_{in}$ represent the conductance of excitatory and inhibitory synapses, respectively;

$E_{ex}$ and $E_{in}$ represent the reversal membrane potentials of excitatory and inhibitory synapses, respectively;

$I_{esyn}$ represents the current passing through electrical synapses.

In addition, when the membrane potential reaches a threshold ($V_{th}$), the neuron fires an action potential, and the membrane potential is immediately reset to the equilibrium potential ($V_{eq}$) after a firing lasting time ($T_{fire}$).

Parameter values chosen for the I-F neuron model are listed in **Table 1**. These values are mostly adopted from Troyer and Miller (1997) [15], except that the firing lasting time of inhibitory neurons is chosen as 4 to ensure the neurons' inhibitory effect on the activities of excitatory neurons.

### 2.2.2 Description of synaptic current

The chemical synapses are modeled as follows [16-17]:

$$I_{csyn}=g_{csyn}\cdot g(t)\cdot(E-V_{post}) \qquad (4)$$

where $g_{ex}(t)$ and $g_{in}(t)$ in eqns (2) and (3) are presented by $g_{csyn}(t)\cdot g(t)$ here, with $g_{csyn}$ representing synaptic strength which is modified by a factor of $g(t)$:

$$\frac{dg(t)}{dt}=\frac{1}{\tau_{syn}}\cdot[f(t)-g(t)] \qquad (5)$$

where

$$\frac{df(t)}{dt}=\frac{1}{\tau_{syn}}\cdot[\Theta[V_{pre}-E_{thr}]-f(t)] \qquad (6)$$

in which $\tau_{syn}$=15 *ms* , $E_{thr}$= - 40 *mV*, and $\Theta(u)$ follows a step function:

$$\begin{cases}\Theta(u)=0 & u\le 0\\ \Theta(u)=1 & u>0\end{cases}$$

The electrical synapses are described as follows:

$$I_{esyn}=g_{esyn}\cdot(V_{pre}-V_{post}) \qquad (7)$$

Where $g_{csyn}$ represents the synaptic strength. We adopt this abstract function which simply depicts that the current passing through the electrical synapses is generally dependent on the membrane potential difference between the pre-synaptic and post-synaptic neurons [18].

## 3. RESULTS

**Table 1.** Parameter values for the I-F neuron model. The firing lasting time ($T_{fire}$) for sensory and excitatory neurons is set as 1.75 *ms* whereas that for inhibitory neuron is set as 4 *ms*.

| C<br>(*pF*) | $V_{eq}$<br>(*mV*) | $V_{th}$<br>(*mV*) | $g_{leak}$<br>($\mu S$) | $E_{ex}$<br>(*mV*) | $E_{in}$<br>*mV*) | $T_{fire}$<br>(*ms*) |
|---|---|---|---|---|---|---|
| 0.5 | -74 | -54 | 0.025 | 0 | -74 | 1.75/4 |

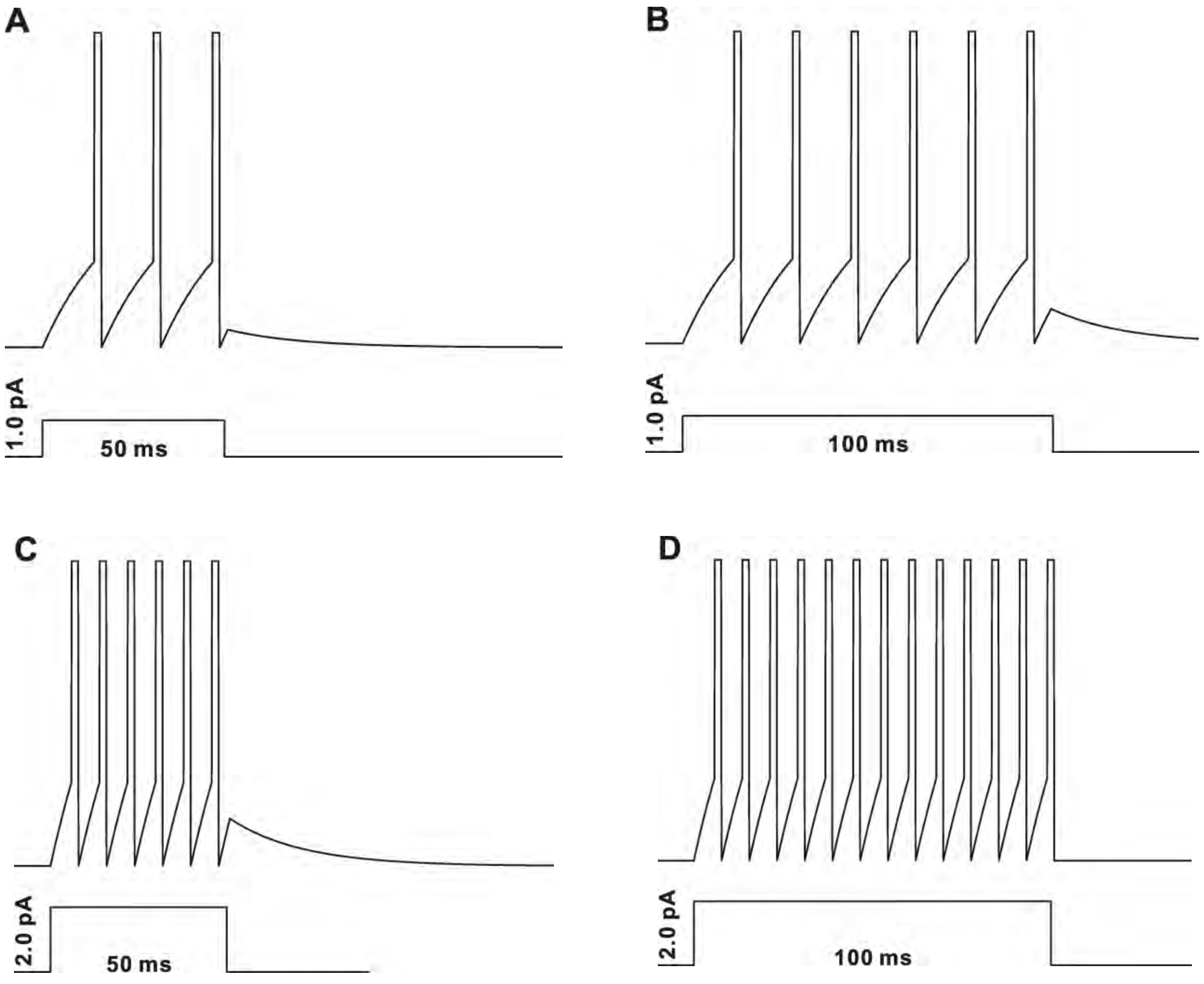


**Figure 2.** Spike activities of the input neuron (S) in response to constant injected currents with various intensities and magnitudes.

### 3.1 Stimulus duration is represented by spike trains of input neuron

The injected current is first transformed into a spike train of the input neuron. Spiking properties of the input neuron (S) are shown in **Figure 2**. Injected currents with different magnitudes and durations are applied to the input neuron to test its performance. A sustained current elicits periodic spikes from the input neuron and the duration of the spike train is determined by the stimulus duration. Input neuron can therefore mimic the function of sensory neuron in neural system.

### 3.2 Performance of the small-scale neural network model

#### 3.2.1 Temporal information can be represented by the spatially distributed activities of a group of neurons

Representative firing patterns of the simple model are given in **Figure 3**. Parameters used for **Figure 3** are listed in **Table 2** and the synaptic connection follows that illustrated in **Figure 1A**. Input neuron is connected to four of the ten excitatory neurons. Three neuronal groups are electrically coupled together which contain 2, 3 and 4 neurons, respectively. Raster plots of the firing performances of the model neurons in absence and presence of electrical synapses are compared with stimulus duration being 50 *ms* (**Figure 3A&B**) and 100 *ms* (**Figure 3C&D**), respectively.

Results given in **Figure 3B&D** suggest that electrical synapses in a neural network can effectively transform the temporal domain spike train of the input neuron into the spatial-temporal firing pattern of a group of neurons. Each activated neuron in the group fires within a specific time window, which is determined by the configuration of the synaptic connection of the neural network. Furthermore, stimulus with longer duration can evoke spikes from more neurons and therefore the stimulus durations can be represented by the spatial and temporal structure of the sequential neuronal activities.

#### 3.2.2 The output pattern is closely related to the electrical coupling configuration

Electrical synapses between excitatory neurons and

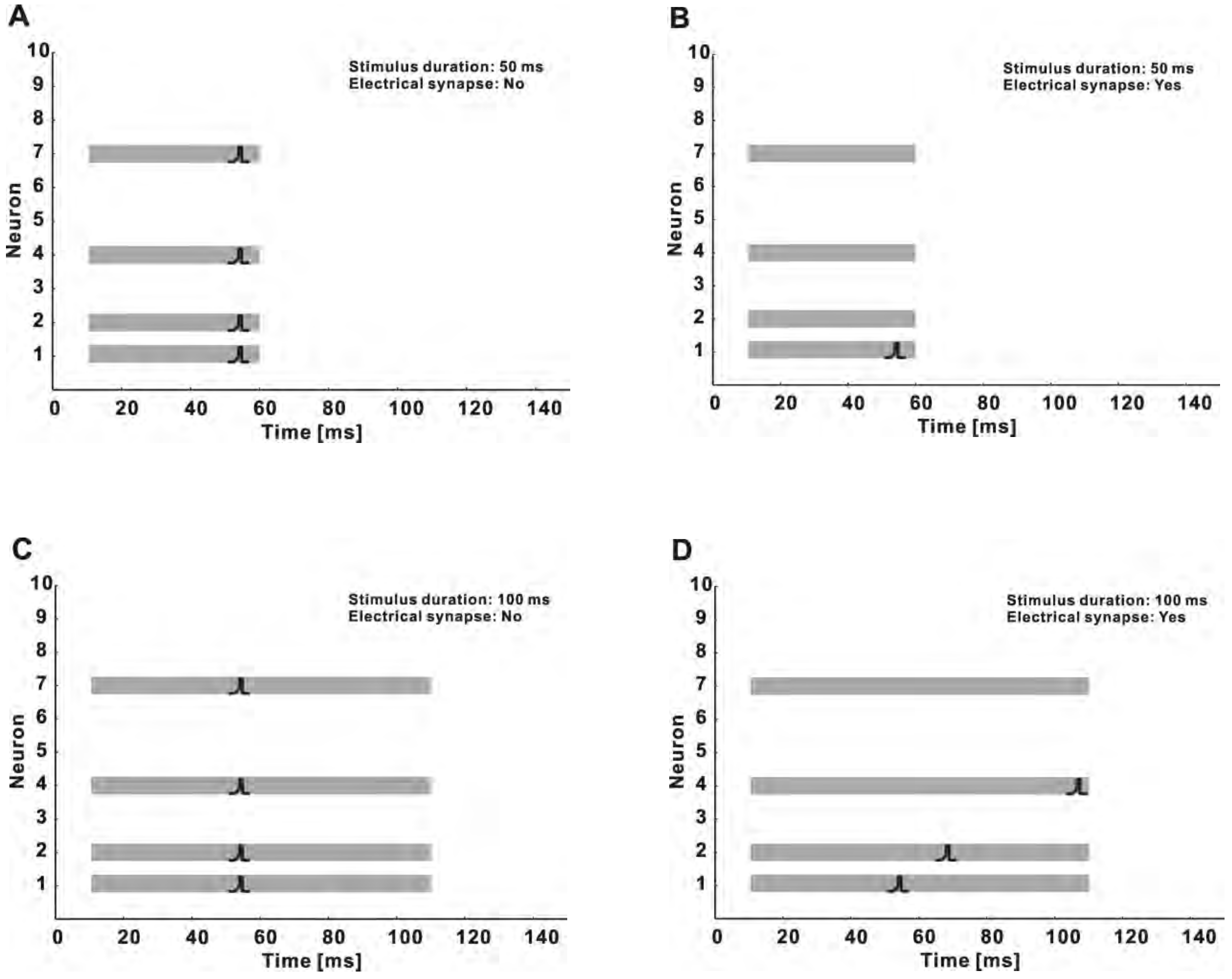


**Figure 3.** Raster plots for neuronal activities of the small-scale model elicited by 50 and 100 *ms* stimulus durations. Stimuli are indicated by grey shadows. A, 50 ms duration, without electrical synapses; B, 50 ms duration, with electrical synapses; C, 100 ms duration, without electrical synapses; D, 100 ms, with electrical synapses.

**Table 2.** Parameter values used in the small-scale neural network model.

| $IS_{ip}(pA)$ | $CS_{se}(\mu S)$ | $CS_{ee}(\mu S)$ | $ES_{ee}(\mu S)$ | $CS_{ei}(\mu S)$ | $CS_{ie}(\mu S)$ |
|---|---|---|---|---|---|
| 2.0 | 0.075 | 0.0001 | 0.02 | 1.0 | 2.0 |

synaptic connections from input neuron to the network are important factors that influence the model's performance. There are three groups of neurons electrically coupled together in the small-scale model presented in **Figure 1A**. Neurons within each group are all electrically coupled in a recurrent manner. Furthermore, only one neuron in each group is connected to the input neuron. The model outputs in response to stimuli with different durations are presented in **Figure 3**. However, any change in the configurations of the electrical coupling and input neuron connection may also cause relevant changes in the results. Take the 3-neuron group in **Figure 1A** (E4, E5 and E6) for an example, relevant possibilities of the electrical coupling within this group as well as the chemical synapses between these neurons and the input neuron are tested, with the rest structure of the neural network kept unaltered. The spiking activities of these three neurons under the test conditions are plotted in **Figure 4**. The firing activities are quite different with different synaptic configurations. Generally, spikes can be elicited from the neurons that are chemically connected to the input neuron, and longer delay is produced when the chemically activated neuron is electrically coupled with more neurons that do not receive chemical input from the input neuron (e.g. **A&F** vs **B&D**).

### 3.3 Performance of the large-scale neural network model

Performance of the small-scale model suggests a mechanism for temporal information processing in a neural network containing electrical synapses. In real neural network, the synaptic strengths as well as the

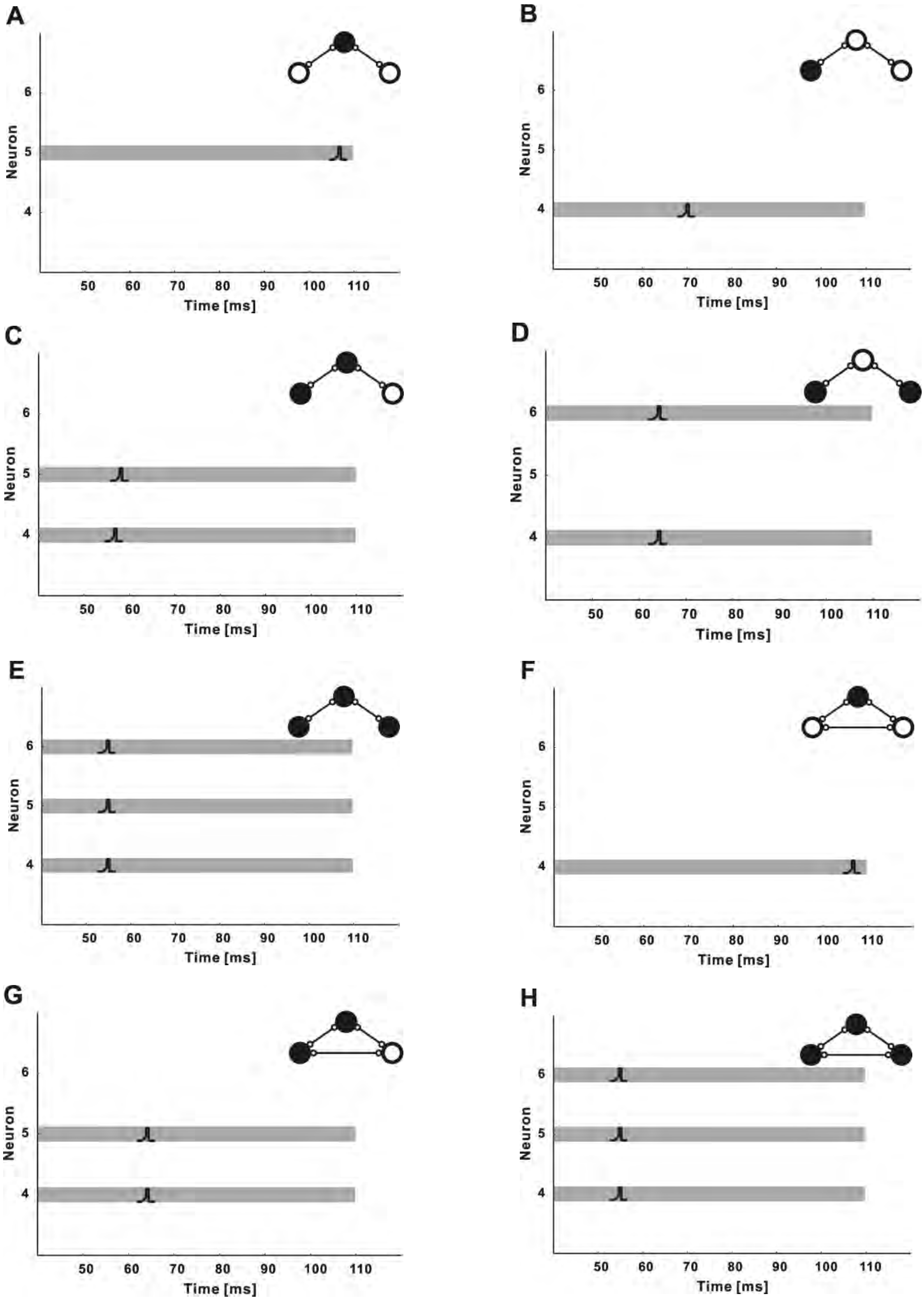


**Figure 4.** Raster plots for spike activities of threeneuron group with different synaptic configurations. Neurons receive synaptic input from input neuron are represented by solid circle. Electrical synapses are represented by solid lines. The stimulus duration is 100 *ms* with the current intensity to input neuron being 2.0 *pA*.

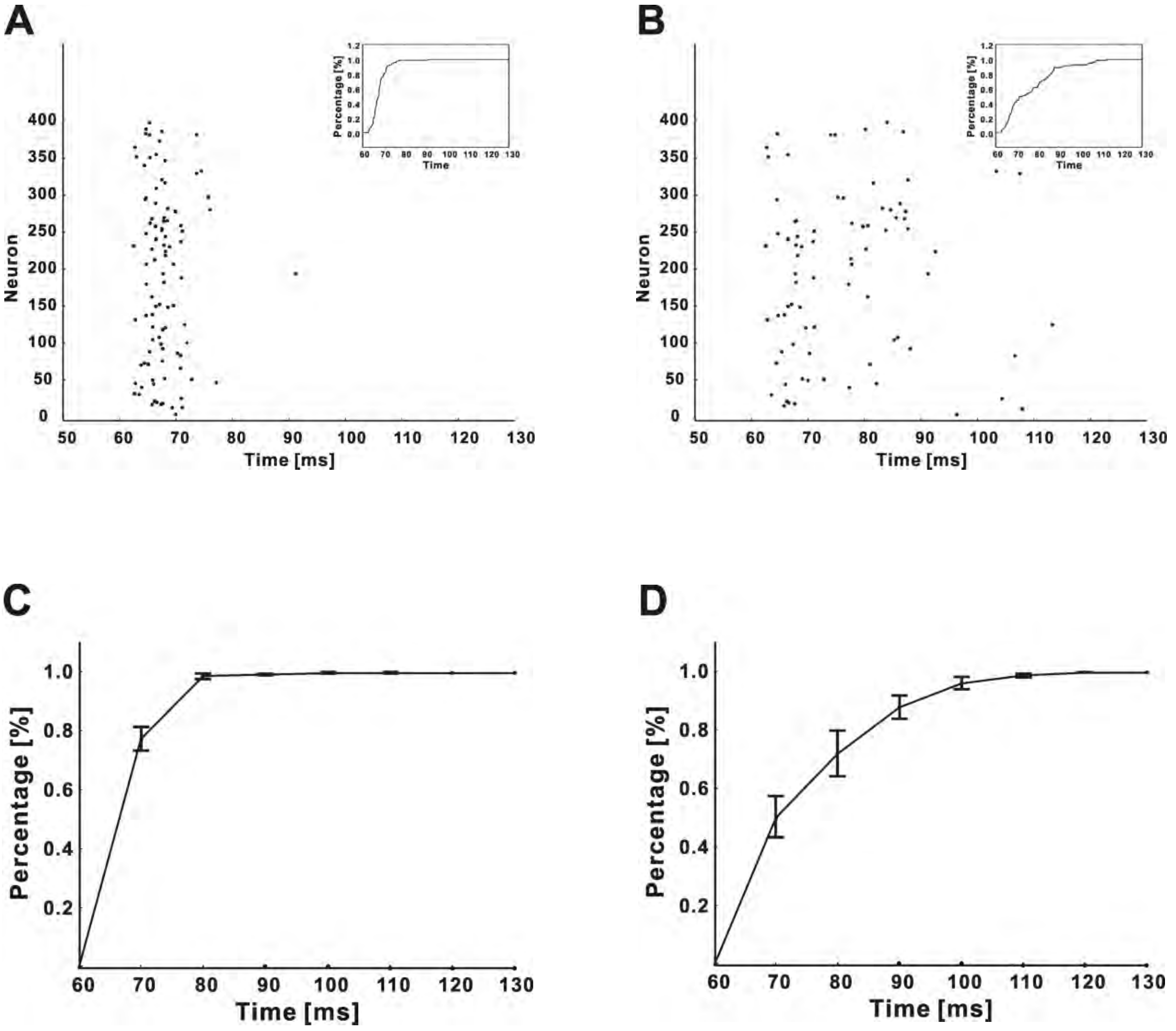


**Figure 5.** A and B are representative raster plots of the neuronal activities of the large-scale model in absence and presence of electrical synapses, respectively. The stimulus duration is 100 *ms*. Inset graphs represent the processes of spike activity recruitment. C and D show the recruitment processes in absence and presence of electrical synapses, respectively. Data are averaged based on 10 independent trails (Mean±S.D.).

**Table 3.** Parameter values used in the large-scale neural network model.

| | | | |
|---|---|---|---|
| $CP_{se}$ | 0.25 | $CM_{se}/CD_{se}$ ($\mu S$) | 0.055/0.003 |
| $CP_{si}$ | 0.98 | $CM_{si}/CD_{si}$ ($\mu S$) | 0.03/ 0.01 |
| $CP_{ee}$ | 0.005 | $CM_{ee}/CD_{ee}$ ($\mu S$) | 0.001/0.001 |
| / | / | $CM_{ei}/CD_{ei}$ ($\mu S$) | 0.2/0.01 |
| / | / | $CM_{ie}/CD_{ie}$ ($\mu S$) | 0.7/0.01 |
| $EP_{ee1}$ | 0.25 | $EM_{ee}/ED_{ee}$ ($\mu S$) | 0.01/0.001 |
| $EP_{ee2}$ | 0.0002 | $EM_{ee}/ED_{ee}$ ($\mu S$) | 0.01/0.001 |

electrical coupling configuration are not fixed but variable. A large-scale model which is more biologically realistic is constructed with parameter variations, and its performance is tested.

Representative firing patterns of the large-scale model in absence and presence of electrical synapses are shown in **Figure 5A** and **B**, respectively. The stimulus duration time is 100 *ms*. Neural network parameters used for **Figure 5** are listed in **Table 3**. The inset graphs represent the recruitment process of the neuronal spiking activities. The temporal distribution of the neuronal activities under these two conditions is compared by analyzing the recruitment process in ten independent trials. The results are shown in **Figure 5C** and **D**. It is clear that the presence of electrical synapses results in a broader temporal distribution of the sequential spike activities of the neurons (**B & D**), while the neuronal firing

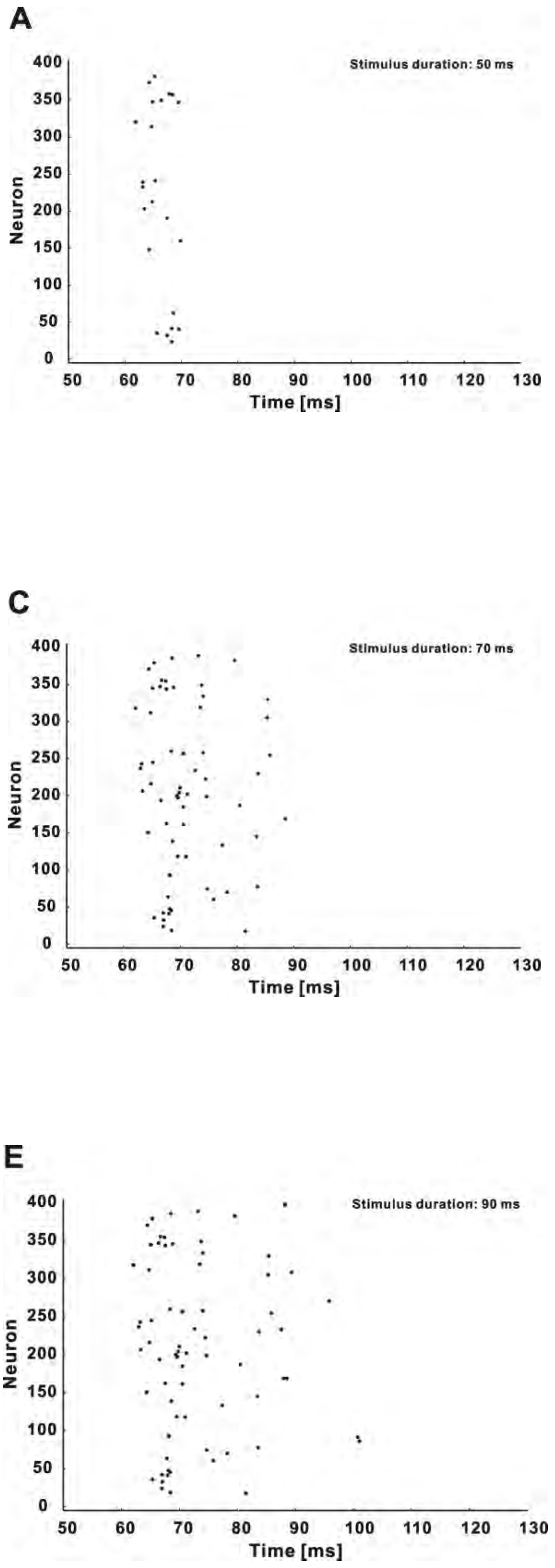


**Figure 6.** Raster plots of the large-scale neural network in response to stimuli with different durations. The configuration of the model is identical for Figure A to F.

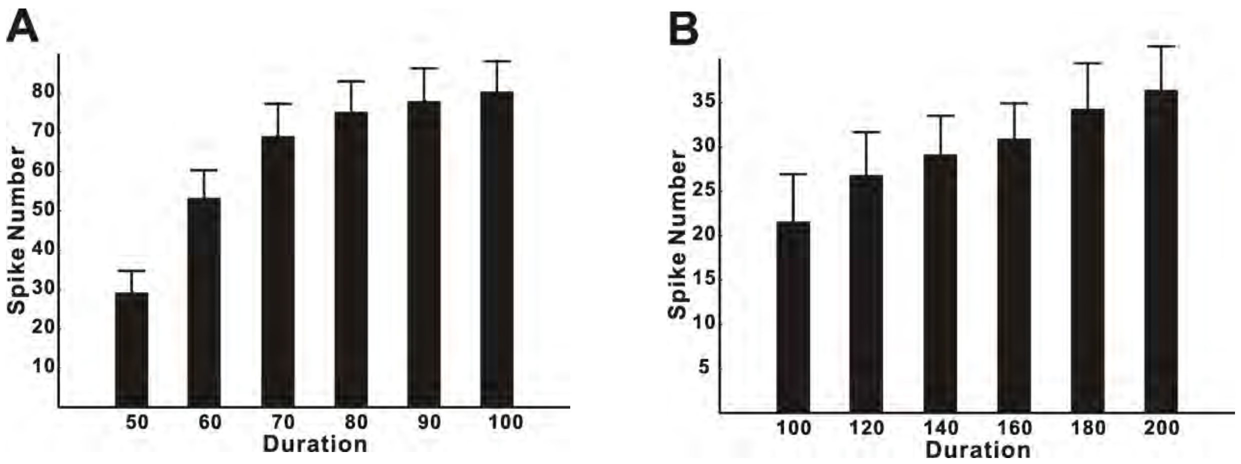


**Figure 7.** Recruitment of neuronal activities (activated numbers) for the large-scale model in response to stimuli with durations ranging from 50 to 100 *ms* (A, step 10 ms) and 100 to 200 *ms* (B, step 20 ms). The mean values of synaptic strength from input to excitatory neurons are 0.055 and 0.038 for results in Figure A and B, respectively. Data are analyzed from 10 independent trials in the form of (Mean±S.D.).

activities are limited within a narrow temporal window in absence of electrical coupling (**A & C**).

The firing patterns of the large-scale model in response to stimuli with various durations are further tested. Stimuli with durations varying from 50 *ms* to 100 *ms* are applied to the network, with steps being 10 ms. Raster plots of typical spike activities of the network are given in **Figure 6**, **A** to **F**. It is revealed that the model neurons fire in a sequential pattern, with more neurons being sequentially recruited in response to longer duration. Such recruitment process in response to durations ranging from 50 *ms* to 100 *ms* is averaged based on ten independent trials and the result is shown in **Figure 7A**.

Stimuli with durations varying from 50 *ms* to 100 *ms* are applied and relevant results are given in **Figure 6** and **Figure 7A**. However, models with this structure can effectively represent durations in other ranges while relevant parameters are changed. These parameters include the capacitance value of the I-F neuronal model, the time constant for chemical synaptic strength, the synaptic strengths from input neuron to the network *et al.* Stimuli with durations ranging from 100 *ms* to 200 *ms* are applied to the network, in which the mean value of synaptic strength from input neuron to the neural network ($Cm_{se}$) are changed (from 0.055 $\mu S$ to 0.038 $\mu S$ ). The performance of the model (averaged across ten independent trials) is plotted in **Figure 7B**.

## 4. DISCUSSION

Temporal information processing in neural system is critical for animal behavior. Neuroscientists have tried a lot in understanding the neural basis of relevant processes via both experimental [6-10] and computational approaches [19-24].

In the present study, the computational results demonstrate that electrical synapses could effectively contribute to the formation of a spatio-temporal firing pattern of neuronal ensembles while each neuron within the ensemble fires within different time windows, and the spatio-temporal pattern of the neuronal activities is capable of representing stimulus duration in the form of sequential firing activities of the spatially distributed neurons.

The contribution of electrical synapses in the formation of spatio-temporal firing pattern is particularly examined in the present study. However, it is necessary to mention that other factors can also contribute to this process. For example, membrane capacitance of specific neurons can be variable because of variation in surface area as well as the membrane capacitance value per unit area [25-28]. These changes can function in parallel to electrical synapses in influencing the sequential firing patterns of neuronal ensembles.

Special role of electrical synapse is proposed in our models and there are also experimental clues which indicated possible roles of electrical synapse in temporal information processing. Data demonstrated that gap junction coupling within inferior olive mediated by connexin 36 could add 10-20 of precision to the fine temporal coordination of muscle firing during movement [13].

Neurons in the present work are modeled following the classic I-F neuron fashion without any specific properties for temporal information processing. These neurons can be tuned to response to any non-temporal properties of natural stimulus and thereby function for the corresponding behavioral tasks. For example, these neurons could be tone selective neuron which function for auditory behavior, or mechanosensory neurons which function for mechanosensation. While both electrical and chemical synapses are universal in the central nervous system, the model results suggest that both the spatial and temporal neuronal activities produced at the sensory layer of neural system could be processed together by sharing the same neural circuit. Temporal content of external stimulus could be read out from spike patterns of neuronal ensembles in the brain.

## ACKNOWLEDGEMENTS

The authors would like to thank Dr. Shi-Yong Huang for helpful discussion. This work was supported by grants from the Hi-Tech Research and Development Program of China (No. 2006AA01Z125).

## REFERENCE

[1]Buonomano DV& Karmarkar UR. How do we tell time? Neuroscientist 2002, 8:42-51.
[2]Mauk MD & Buonomano DV. The neural basis of temporal processing. Annu Rev Neurosci 2004, 27:307-340.
[3]Ivry RB & Spencer RM. The neural representation of time. Curr Opin Neurobiol 2004, 14:225-232.
[4]deCharms RC & Zador A.Neural representation and the cortical code. Annu Rev Neurosci 2000, 23:613-647.
[5]Dayan, P. & Abbott, LF () Theoretical Neuroscience, MIT Press 2001.
[6]Casseday JH, Ehrlich D & Covey E. Neural tuning for sound duration: role of inhibitory mechanisms in the inferior colliculus. Science 1994, 264:847-850.
[7]Galazyuk AV & Feng AS. Encoding of sound duration by neurons in the auditory cortex of the little brown bat, Myotis lucifugus. J Comp Physiol 1997, [A] 180:301-311.
[8]He J., Hashikawa T., Ojima H. & Kinouchi Y. () Temporal integration and duration tuning in the dorsal zone of cat auditory cortex. J Neurosci 1997, 17:2615-2625.
[9]Ehrlich D, Casseday JH & Covey E. Neural tuning to sound duration in the inferior colliculus of the big brown bat, Eptesicus fuscus. J Neurophysiol 1997, 77:2360-2372.
[10]Fremouw T, Faure PA, Casseday JH & Covey E. Duration selectivity of neurons in the inferior colliculus of the big brown bat: tolerance to changes in sound level. J Neurophysiol 2005, 94:1869-1878.
[11]Connors BW & Long MA. Electrical synapses in the mammalian brain. Annu Rev Neurosci 2004, 27:393-418.
[12]Sohl G., Maxeiner S. & Willecke K. Expression and functions of neuronal gap junctions. Nat Rev Neurosci 2005, 6:191-200.
[13]Placantonakis DG, Bukovsky AA, Zeng XH, Kiem HP & Welsh JP. Fundamental role of inferior olive connexin 36 in muscle coherence during tremor. Proc Natl Acad Sci U S A 2004, 101:7164-7169.
[14]Beaulieu C., Kisvarday Z., Somogyi P., Cynader M. & Cowey A. Quantitative distribution of GABA-immunopositive and -immunonegative neurons and synapses in the monkey striate cortex (area 17). Cereb Cortex 1992, 2:295-309.
[15]Troyer TW & Miller KD. Physiological gain leads to high ISI variability in a simple model of a cortical regular spiking cell. Neural Comput 1997, 9:971-983.
[16]Rall W. Distinguishing theoretical synaptic potentials computed for different soma-dendritic distributions of synaptic input. J Neurophysiol 1967, 30:1138-1168.
[17]Nowotny T., Rabinovich MI, Huerta R. & Abarbanel HD. Decoding temporal information through slow lateral excitation in the olfactory system of insects. J Comput Neurosci 2003, 15:271-281.
[18]Kopell N. & Ermentrout B. Chemical and electrical synapses perform complementary roles in the synchronization of interneuronal networks. Proc Natl Acad Sci U S A 2004, 01:15482-15487.
[19]Hooper SL, Buchman E. & Hobbs KH. A computational role for slow conductances: single-neuron models that measure duration. Nat Neurosci 2002, 5:552-556.
[20]Buonomano DV. Decoding temporal information: A model based on short-term synaptic plasticity. J Neurosci 2000, 20:1129-1141.
[21]Nowotny T, Rabinovich MI & Abarbanel HD. Spatial representation of temporal information through spike-timing-dependent plasticity. Phys Rev E Stat Nonlin Soft Matter Phys 2003, 68:011908.
[22]Buonomano DV & Merzenich MM. Temporal information transformed into a spatial code by a neural network with realistic properties. Science 1995, 267:1028-1030.
[23]Mauk MD & Donegan NH. A model of Pavlovian eyelid conditioning based on the synaptic organization of the cerebellum. Learn Mem 1997, 4:130-158.
[24]Medina JF, Garcia KS, Nores WL, Taylor NM & Mauk MD. Timing mechanisms in the cerebellum: testing predictions of a large-scale computer simulation. J Neurosci 2000, 20:5516-5525.
[25]Chitwood RA, Hubbard A. & Jaffe DB. Passive electrotonic properties of rat hippocampal CA3 interneurones. J Physiol 1999, 515 (Pt 3):743-756.
[26]Gentet LJ, Stuart GJ & Clements JD. Direct measurement of specific membrane capacitance in neurons. Biophys J 2000, 79:314-320.
[27]Major G., Larkman AU, Jonas P., Sakmann B. & Jack JJ. Detailed passive cable models of whole-cell recorded CA3 pyramidal neurons in rat hippocampal slices. J Neurosci 1994, 14:4613-4638.
[28]Thurbon D., Luscher HR, Hofstetter T. & Redman SJ. Passive electrical properties of ventral horn neurons in rat spinal cord slices. J Neurophysiol 1998, 80:2485-2502.

Scientific Research Publishing

# Graft copolymerization of N,N-Dimethylacrylamide to cellulose in homogeneous media using atom transfer radical polymerization for hemocompatibility

Li-Feng Yan *, Wei Tao

Hefei National Laboratory for Physical Science at Microscale, and Department of Chemical Physics, University of Science and Technology of China, Hefei, 230026, P.R.China. * Correspondence should be addressed to Li-Feng Yan (lfyan@ustc.edu.cn).

## ABSTRACT

**In homogeneous media, N,N -Dimethylacrylamide (DMA) was grafted copolymerization to cellulose by a metal-catalyzed atom transfer radical polymerization (ATRP) process. First, cellulose was dissolved in DMAc/LiCl system, and it reacted with 2-bromoisobutyloyl bromide (BiBBr) to produce macroinitiator (cell-BiB). Then DMA was polymerized to the cellulose backbone in a homogeneous DMSO solution in presence of the cell-BiB. Characterization with FT-IR, NMR, and GPC measurements showed that there obtained a graft copolymer with cellulose backbone and PDMA side chains (cell-PDMA) in well-defined structure. The proteins adsorption studies showed that the cellulose membranes modified by the as-prepared cell-PDMA copolymer own good protein adsorption resistancet.**

**Keywords: Cellulose; Atom transfer radical polymerization (ATRP); Homogeneous; Graft copolymerization; Hemocompatibility**

## 1. INTRODUCTION

Cellulose is the most fluent feedstock in the world that could be used to prepare new kinds of materials, and cellulose derivatives have potential application as functional polymers. Graft copolymers are the important topic for their novel properties. Today, “grafting from” method has been widely used to prepare cellulose copolymers. Ceric ion initiation, Fenton's reagent and γ -radiation are the widely used methods to graft monomers to cellulose [1,2]. However, there are some drawbacks of these methods, such as the production of unwanted homopolymer together with the graft copolymer, chain degradation of the cellulose backbone during the formation of free radical grafting sites, and the presence of a considerable amount of ungrafted cellulose in the product. In addition, these techniques usually results in the graft copolymer with poor control over the composition, such as molecular weight and the polydispersity of the grafted chains [3]. Recently, controlled/“living” radical polymerization methods have been developed [4], which is able to minimize chain transfer and to control the molecular weight and polydispersity. Among them atom transfer radical polymerization (ATRP) and reversible addition fragmentation transfer polymerization (RAFT) are the two convenient methods to prepare well-defined polymers. Using living free radical polymerization methods to prepare cellulose graft copolymer is an attractive topic and some investigations had been carried out. Perrier, *et al.* reported a preparation of polystyrene graft cellulose by a RAFT process [5]. Carlmark and Malmstrom synthesized a poly(2-hydroxyethyl methacrylate) graft cellulose using an ATRP process [6]. However, in both the studies, the graft copolymerization occurs only on the surface of cellulose fiber due to the heterogeneous process. Huang, *et al.* reported a homogeneous ATRP process to prepare cellulose graft copolymers with different monomers; the reason why ethyl cellulose was selected as the feedstock is its easily dissolving ability in many solvents[8, 9, 25, 26, 27]. By now there are still less reports to synthesize cellulose graft copolymer through a living radical polymerization directly from cellulose in its homogeneous solution, and it is important to prepare well-defined structures of the graft copolymer.

Poly(N,N-dimethylacrylamide) (PDMA) is well-known for its remarkable water solubility and biocompatibility [10]. Recently, well-defined PDMA has been prepared by both RAFT [11] and ATRP processes [12]. Also PDMA has been grafting polymerization to polystyrene colloid by ATRP method [13].

Hemodialysis is one of the most important methods for blood purification [14], and cellulose membranes, especial cellulose acetate (CA) membranes, are still the major materials for hemodialysis [15]. The cellulose membranes could take the porous and asymmetrical structure and have both good permeability and mechanical strength. However thrombus formation on the blood-contact surface could not suppressed by the membrane. Thus, its hemocompatibility must be further improved for better hemodialysis [16]. Several efforts had been carried out to solve these problems, such as modification of the surface of the membrane with low-molecular-weight compounds, hydrophilic polymers and biologically active heparin [17,18].

In this paper, synthesis of the graft copolymer composed of PDMA chains and cellulose backbone (cell-PDMA) in homogeneous solution have been studied via an ATRP. Moreover, the protein adsorption resistivity on the cellulose membrane surface modified with the cell-PDMA was evaluated to understand hemocompatibility of the cell-PDMA.

## 2. EXPERIMENTAL SECTION

### 2.1. Materials

The chemical formula of the DMA is shown in **Scheme 1**. Commercial product of microcrystalline (Sigma, DP = 121) was used without further purification. 2,2'-Bipyridine (bpy) purchased from Aldrich was recrystallized from ethanol to remove impurities. DMA, CuBr with purity of 99.999% and 2-bromoisobutyloyl bromide (Br*Bi*B) were purchased from Aldrich and used without further purification. Other solvents and reagents were extra-pure grade reagents and used without further purification.

### 2.2. Dissolution of cellulose in N,N-dimethyl acetoamide (DMAc)/LiCl

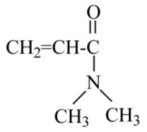
After dried in vacuum at 35°C overnight microcrystalline cellulose (5.167 g) was put into a 250 ml three-necked round-bottom flask, and adding 100 ml of distilled water for 30 min to swell it, then water was removed and fresh water was added again, and the process was repeated for three times. Then removing the water and adding 100 ml of methanol to swell again for 30 min for three times. After removing methanol the cotton was dried in vacuum at 50 °C for 3 h. Then cooling down the solution and adding 120 ml of DMAc and heated at 160 °C for 1.5h, and removing 20 ml of DMAc under reduced press by a rotary evaporator. At the same time, about 10.22 g of LiCl was dried in baker at 60 °C. After the removing process of DMAc finish, adding the dried LiCl into the system, and stirring at 80 °C for 13 h, and the cellulose solution was obtained at the end [19].

### 2.3. Synthesis macroinitiator for ATRP

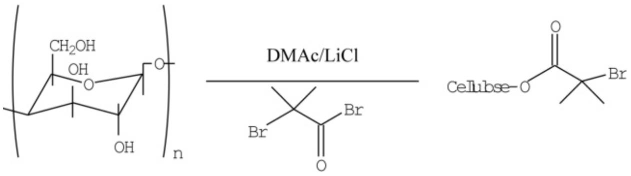
Cellulose was acylated with Br*Bi*B in the presence of pyridine as shown in **Scheme 2** [25, 26, 27]. In a 250 ml three-necked round-bottom flask, 60 ml of the cellulose solution in DMAc/LiCl and 5 ml of pyridine were added and mixed, then 6.3541 g of Br*Bi*B was slowly dropped into the solution at 0 °C in an ice/water bath. The reaction mixture was further stirred at room temperature overnight. Then the mixture was added with de-ionized water and plenty of precipitate appeared, and after washed by plenty of de-ionized water, the precipitate was dried at 50 °C in vacuum overnight. Finally, there obtained white powder product of macroinitiator(cell-B*i*B) with weight of 4.81 g. The cell-B*i*B can be well dissolved in dimetyl sulfoxied (DMSO).

### 2.4. Grafting copolymerization of DMA by the cell-B*i*B

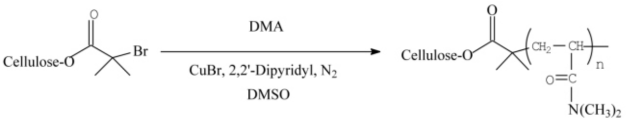
The cell-B*i*B(0.1737 g, 0.9 mmol) was dissolved in 30 ml of DMSO in a 100 ml of flask. Then 7.92 g (0.08 mol) of DMA was added, and the solution was evacuated and flushed with nitrogen for 30 min. Finally, 0.1021 g of bpy (0.7 mmol) and 0.0444 g of CuBr (0.31 mmol) were added, and the polymerization was carried out at room temperature under the protect of nitrogen. A few milliliter of samples were withdrawn from the flask at different reaction time using degassed syringes to determine monomer conversion and molecular weight.



**Scheme 1**. Chemical structure of DMA.



**Scheme 2**. Synthesis route for the macroinitiator (cell-B*i*B).



**Scheme 3**. Graft copolymerization of DMA on cellulose backbone in homogeneous solution via the ATRP route.

The samples were diluted with DMSO and filtering the solution through a silicon gel column to remove the Cu ions catalyst, and then plenty of hexane was added to produce the precipitate of the products. The products were dried at 40 °C in vacuum overnight.

### 2.5. Isolation of the grafted PDMA chains by hydrolysis

The copolymers were hydrolyzed by 70% $H_2SO_4$ for 8h at boiling point. At the end, the residual polymer was participated into plenty of hexane and was dried by freeze drying, then the products were analyzed by GPC.

### 2.6. Characterization

The chemical structure was confirmed using an FT-IR (FT/IR-615, JASCO, Tokyo, Japan). $^1$H- and $^{13}$C-NMR spectra were obtained on a NMR spectrometer (α -300, JEOL, Tokyo, Japan) with $D_2O$ as the solvent. The molecular weights of these polymers were determined by gel permeation chromatography (GPC). The mixture of methanol/water = 7/3 containing 10 mmol/L of lithium bromide was used as an eluent for the GPC measurement at a flow rate of 0.4 ml/min (Column: SB-804 HQ, Shodex, Tokyo, Japan). The number-averaged molecular weight ($M_n$) and weight-averaged molecular weight ($M_w$) were calculated using poly(ethylene glycol) standards.

X-ray photoelectron spectroscopy (XPS) was conducted on an AXIS-HSi (Shimadzu/KRATOS, Kyoto, Japan) employing Mg $K_\alpha$ excitation radiation (1253.6 eV). The take-off angle of the photoelectron for each atom was fixed at 90 deg.

For Atomic force microscopy (AFM) measurement, the sample was dissolved in DMF at a concentration of 8×$10^{-6}$ g/m. Then a droplet (20 µl) of the solution was deposited onto freshly cleaved mica, and it was spin-coated at speed of 900 rpm for 8 s and then 4000 rpm for 30s. The height image of the copolymer on mica were measured by an AFM (Nanoscope IIIa, D.I.) in tapping mode with silicon TESP cantilevers. The scanning rate ranged from 0.5 Hz to 1.0 Hz, and 512×512 pixels images were record.

### 2.7. Coating of the cell-PDMA on cellulose membrane

The regenerated cellulose membrane, Cuprophan$^{(TM)}$, was obtained from Enka, A. G. (Wappertal-Barmen, Germany). The thickness of the membranes was 20 µm. First the cellulose membranes were cut into pieces with diameter of 1.5cm, and they were immersed into deionized water for 30 min, and then were dried at 35 °C in vacuum for 15h. Then the cellulose membranes were immersed into the 0.5 wt% aqueous solution of the cell-PDMA for 3 min, and the membranes were took out and dried under atmospheric conditions for 2h, and then was dried at 35 °C in vacuum for 15 h. The structure of the grafted DMA on the cellulose membranes were confirmed using XPS and FT-IR. The ratio of nitrogen atom (N) in the DMA unit versus carbon atom (C) was determined from the XPS elemental analysis.

### 2.8. Protein adsorption on the membrane surface

Amount of proteins adsorbed on the membrane was measured by almost the same method reported previously [20]. The round (diameter: 1.5 cm) cellulose membranes were placed into a 24-well plate. To equilibrate the membrane surface, phosphate buffer solution (PBS, pH 7.4, ionic strength : 0.15 mol/l) was added into each well and allowed to remain for 15 h at room temperature. Protein solutions were prepared in the concentration of 4.5 mg/ml of albumin, 1.6 mg/ml of γ -globulin, and 0.3 mg/ml of fibrinogen, which are 10% of the concentration of the human plasma level. After removing the PBS, 1.0 ml of each protein solution was poured onto each membrane and allowed to remain at 37 °C for 3 h. After rinsing the membrane three times with PBS, the membrane was taken out of the 24-well plate, and was rinsed again sufficiently with the 50 ml of PBS. The membrane was placed into a glass bottle with a 1 wt% aqueous solution of sodium dodecyl sulfate (SDS) and shaken (150 rpm) in a shaking bath for 3 h at room temperature to detach the adsorbed protein on the surface. A protein analysis kit (Micro BCA protein assay reagent kit, #23235, Pierce, Rockford, IL, USA) based on the bicinchoninic acid method was used to determine the protein concentration in the SDS solution.

## 3. RESULTS AND DISCUSSION

The cell-B*i*B was prepared by partial esterification of the hydroxyl groups of the glucose units of cellulose with B*i*BBr in the presence of pyridine. The reaction was carried out homogeneously in DMAc/LiCl solution at room temperature for 23 h. The formation of the ester bond resulted in the appearance of the characteristic peaks at 1743 cm$^{-1}$ for the C=O stretching band in the FTIR spectrum, as shown in **Figure 1**.

The substitution of the hydroxyl groups on the cellulose backbone with B*i*BBr was also confirmed by

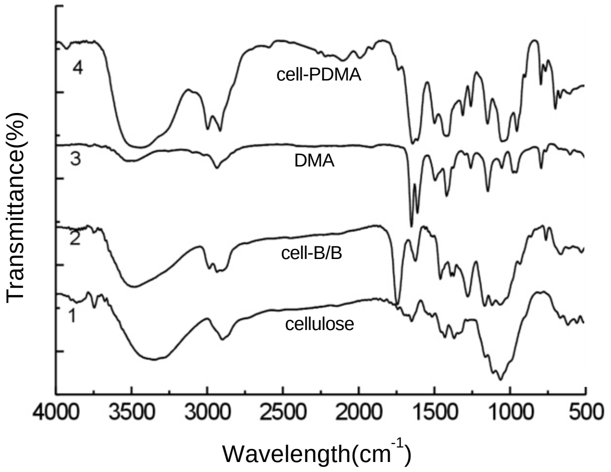


**Figure 1**. FT-IR spectra of cotton (1), cell-B*i*B (2), DMA (3) and cell-PDMA (4).

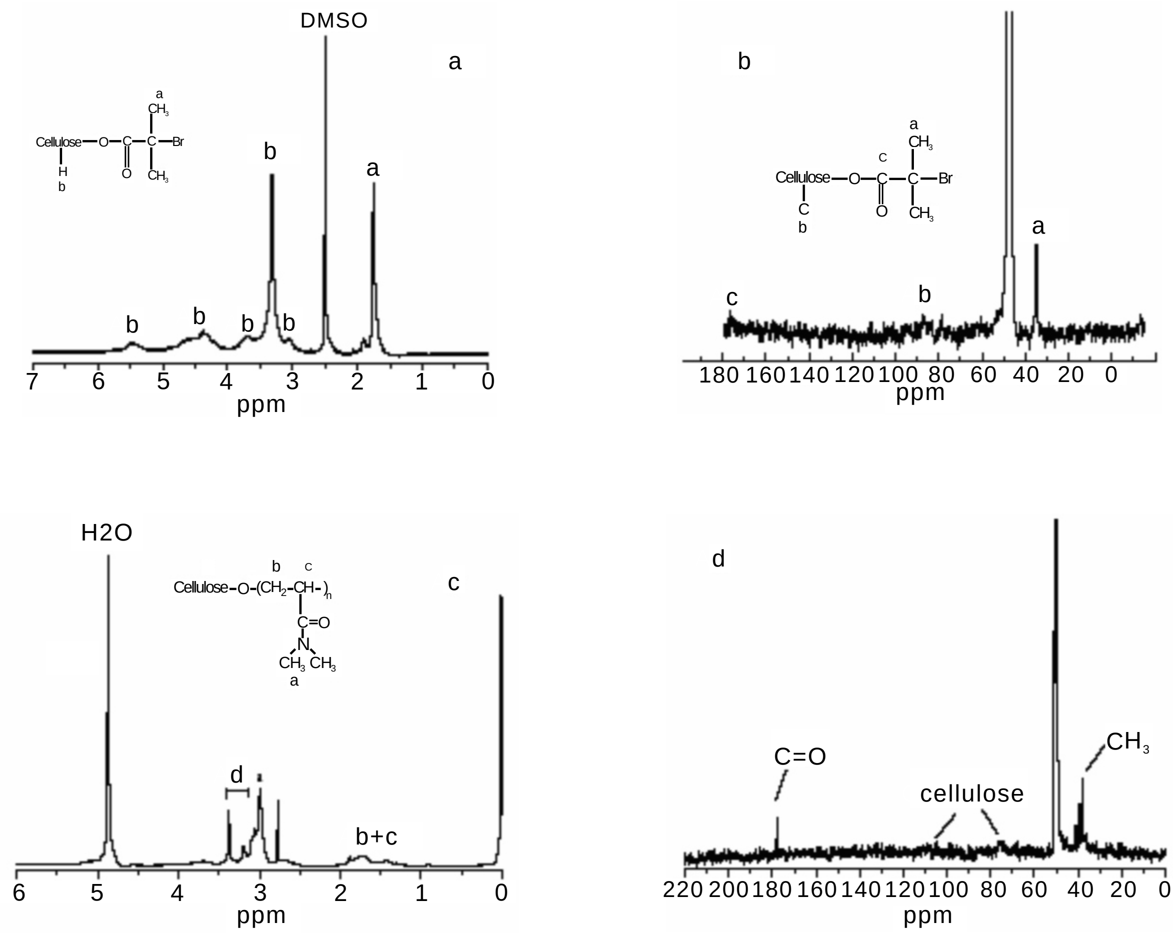


**Figure 2.** $^1$H-NMR and $^{13}$C-NMR spectra of cell-B*i*B (a, b) and cell-PDMA (c, d).

both the $^1$H-NMR and $^{13}$C-NMR. As shown in **Figure 2a**, there appears a new single peak at 1.8 ppm (peak a) for methyl protons in the ester group of B*i*B, and the peaks at δ = 2.8-5.6 ppm (peak b) for the methylene protons and hydroxyl protons in the glucose units of cellulose [21]. The total substitution degree (DS) of B*i*B is obtained by the ratio of the integral of the methyl groups to the integral of protons of glucose, and the DS is 0.2. **Figure 2b** shows the $^{13}$C-NMR of the cell-B*i*B, and clearly both the methyl carbon from B*i*B (peak a) and the carbon in glucose (peak b) appear, and the peak c at 176 ppm attributed to the C=O carbon of B*i*B [22].

The as-prepared cell-B*i*B can be dissolved well in DMSO. The graft copolymerization of DMA to cellulose was carried out in DMSO at 100 °C, [DMA]:[cell-B*i*B]:[CuBr]:[bpy] = 88:1:2.9:1.3, and $[DMA]_0$ = 2.7 M. **Figure 3** shows the kinetic plot of the reaction, and the variation of $\ln([M]_0/[M])$ is linear with time, indicating a constant concentration of propagating radicals which is the characteristic of the controlled/"living" radical polymerization.

The chemical structure of the cell-PDMA was identified by FT-IR spectroscopy, NMR and GPC. As shown in **Figure 1**, when the FT-IR spectrum of cell-PDMA was compared with that of the cell-B*i*B and DMA monomer, the absorptions at 1642 cm$^{-1}$ appeared after grafting, which was assigned to the free C=O of PDMA, and the peaks at about 3100-3500 cm$^{-1}$ was assigned to the OH group of cellulose [23].

**Figure 2c** shows a $^1$H-NMR spectrum for the cell-PDMA in methanol-$d_4$ at 25 °C, the spectra is about the same as that of PDMA. The resonance bands observed at 2.9-3.1 ppm are attributed to the dimethyl group, and those observed at 1.3~1.8 ppm is attributed to the methyl amd methylene protons of PDMA [24]. Part of the resonance bands of cellulose protons are overlapped with that of PDMA while there appear peaks at 2.9-4.0 ppm for the characteristics of cellulose. **Figure 2d** shows a $^{13}$C-NMR spectrum for cell-PDMA in $D_2O$ at 25 °C. The characteristic of the resonance peak for PDMA was observed at 35 ppm, which is attributed to the dimethyl moiety [25]. The weak peaks appear at 75-85 ppm are attributed to the carbon for cellulose back bone, and the peak appear at 182 ppm is attributed to the carbon for the carbonyl groups.

The grafted PDMA chains were converted to individual molecules through hydrolysis of the backbone

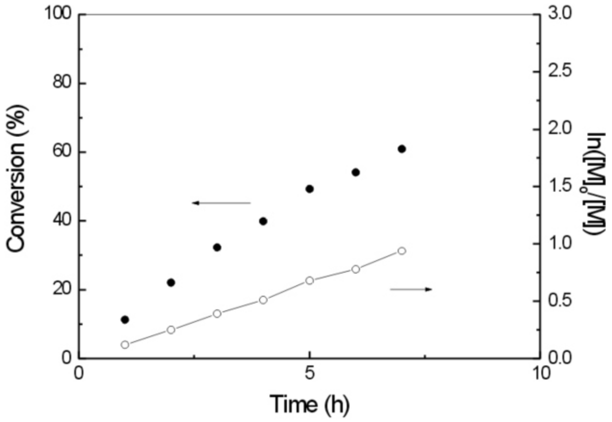


**Figure 3**. Time-conversion and the first-order kinetic plot for the polymerization of DMA initiated by the cell-B*i*B in the homogeneous solution of DMSO at 100 °C. [$M_0$] and [M] are concentrations of monomer at polymerization time = 0 and at corresponding time, respectively.

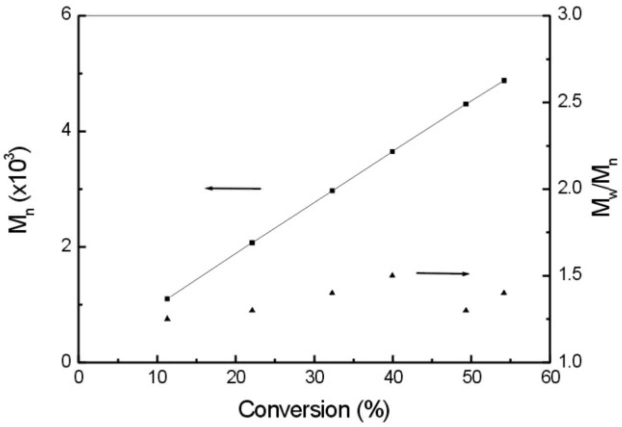


**Figure 4**. Dependence of $M_n$ and $M_w/M_n$ on monomer conversion in the graft polymerization of DMA in DMSO, the PDMA was hydrolyzed from the side chain of the copolymer before the GPC measurements.

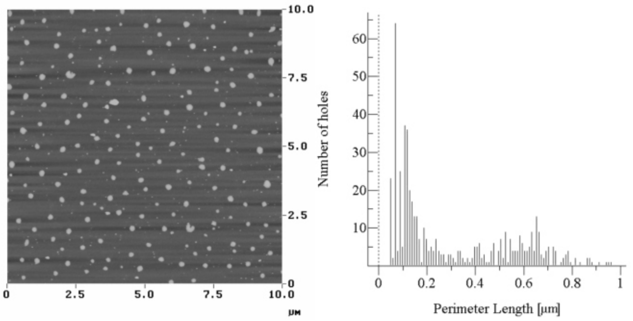


**Figure 5.** Typical AFM image of the cell-PDMA (a) and the perimeter distribution of the particles (b).

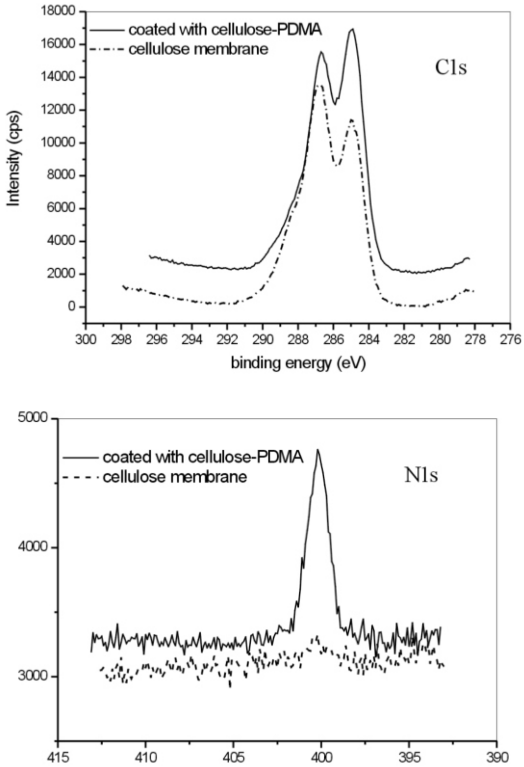


**Figure 6**. XPS spectra of $P_{2p}$, $N_{1s}$, $C_{1s}$, and $O_{1s}$ observed on the original cellulose membrane (down row) and that coated with the cell-PDMA (upper row).

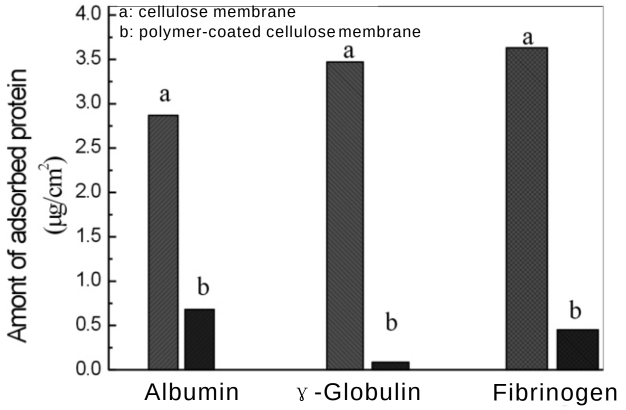


**Figure 7**. Amount of proteins adsorbed on original cellulose membrane (a) and cellulose membrane coated with cell-PDMA (b).

to determine their molecular weight. **Figure 4** shows the plot of $M_n$ and the $M_w/M_n$ versus the monomer conversion during the polymerization. The molecular weight of the graft copolymer is increased linearly with the monomer conversion, and the polydispersity is decreased with the monomer conversion. The results also confirmed that the graft copolymerization is a controlled/"living" radical polymerization.

**Figure 5** shows the AFM image of the cell-PDMA copolymer deposited on surface of the new cleaved mica. Many nanoparticles appear with a homogeneous size, and **Figure 5b** gives the perimeter distribution of the particles. Clearely, there are two kinds of particles exist, one with the diameter about 200 nm, and the other about 38 nm in diameter. Huang also reported similar result when they measrued the size of celluose-PS graft copolymer by AFM, and they concluded that the smaller particles are the graft copolymer and the bigger one are the micelles of the graft copolymer when comparing the AFM data to dynamic laser light scattering results. Here, we believe that the smaller particles result from the cell-PDMA copolymer while the bigger one is the aggregates or micelle of the graft copolymer.

The as-prepared cell-PDMA was a water-soluble polymer having both affinities to the cellulose base membrane, and its potential blood compatibility could improve the surface blood compatibility of the cellulose membrane by a convenient technique, such as coating by its aqueous solution. Coating of cellulose membrane with the cell-PDMA was carried out by immersing the membrane into its aqueous solution following a dry process under vacuum. The amount of the copolymer immobilized on the membrane was measured by XPS. **Figure 6** shows the XPS chart of both the original cellulose membrane and the copolymer coated cellulose membrane (upper row). The peaks attributed to nitrogen (400 eV) atoms was observed on the surface of membrane coated with the cell-PDMA. For the membrane coated with the copolymer, the atomic concentrations of nitrogen and carbon are 1.82% and 66.16%, respectively. The mole fraction of DMA on the membrane surface is 0.028 by calculation, which defined as [number of DMA unit (mol)]/[number of DMA unit (mol) + number of cellulose unit (mol)] was calculated from the value of N/C.

The adsorption of proteins during contact with blood on artificial surface is the initials step in a sequence of events which cause activation of several cascades of proteolysis systems in the plasma, e.g., complement, coagulation pathway, etc. therefore, the amount of proteins adsorbed on the surface is one of the important factors for evaluating the hemocompatibility of materials. Here, the adsorption of three typical plasma proteins such as albumin, γ -globulin, and fibrinogen on the cell-PDMA coating cellulose membranes and original cellulose membranes were measured. As shown in **Figure 7**, the amounts of each absorbed protein on the membrane coated with cell-PDMA was 70-80% reduced by comparison with those on the original cellulose membrane for all of the proteins examined in this study. That is, grafting of PDMA chains on the cellulose plays an important role to reduce protein adsorption.

## 4. CONCLUSION

In this study, we have successfully synthesized the cell-PDMA in homogeneous media using an ATRP controlled/"living" radical polymerization. The characterizations indicate that the graft copolymerization is efficient and the obtained copolymer owns well-defined structures.After coated the cell-PDMA onto the surface of commercial cellulose membrane, there obtained membrane with good hemocompatibility, which was confirmed by the protein adsorption experiments. This provides a new chance to modify the surface of polysaccharide materials to improve their hemocompatibility. The cell-PDMA has a strong potential application on surface treatment to enhance separation ability and selectivity on every cellulose membrane including CA and nitrocellulose, which are applied in biotechnology research and bioengineering field.

## ACKNOWLEDGEMENT

This work is supported by the National Basic Research Program of China (No. 2007CB210201) and the National Key Technology R&D Program (No. 2006BAF02A09).

## REFERENCE

[1] D.W.Jenkins & S.M.Hudson. Review of vinyl graft copolymerization featuring recent advances toward controlled radical-based reactions and illustrated with chitin/chitosan trunk polymers. *Chem Rev*. 2001, 101: 3245-3273.

[2]Hu, Z.H. & Zhang, L.M. Water-soluble ampholytic grafted polysaccharides. II. Synthesis and characterization of graft terpolymers of starch with acrylamide and [2-(methacryloylox)] ethyl dimethyl (3-sulfopropyl) ammonium hydroxide. *J Macromol Sci Pure Appl Chem* 2002, A39: 419-430.

[3]D. Roy, J. T. Guthrie & S. Perrier. Graft polymerization: grafting poly(styrene) from cellulose via RAFT polymerization. *Macromolecules* 2005, 38:10363-10372.

[4]Qiu, J., B. Charleux & K. Matyjaszewski. Controlled/living radical polymerization in aqueous media: homogeneous and heterogeneous system. *Prog Polym Sci* 2001, 26:2083-2134.

[5]D. Roy, J.T. Guthrie & S. Perrier. RAFT graft polymerization of 2-(Dimethylaminoethyl) Methacrylate onto cellulose fibre. *Aust J Chem* 2006, 59: 737-741.

[6]A. Carlmark & E.E. Malmstrom. ATRP grafting from cellulose fibers to create block-copolymer grafts. *Biomacromolecules* 2003, 4: 1740-1745.

[7]T. Aoki, H.K. Kawashima, H. Katono, K. Sanui, N. Igata, T. Okano & Y. Sakurai. Temperature-Responsive Interpentrating Polymer Networks Constructed with Poly (Acrylic Acid) and Poly (N,N-Dimethylacrylamide). *Macromolecules* 1994, 27: 947-952.

[8]M.S. Donovan, T.A. Sanford, A.B. Lowe, B.S. Sumerlin, Y. Mitsukami & C.L. McCormick. RAFT polymerization of N,N-dimethylacrylamide in water. *Macromolecules* 2002, 35: 4570-4572.

[9]Ding, S.J., M. Radosz & Shen, Y.Q. Atom transfer radical polymerization of N,N-dimethylacrylamide. *Macromole Rapid Commun* 2004, 25: 632-636.

[10]J. N. Kizhakkedathu & D. E. Brooks. Synthesis of poly (N,N-dimethylacrylamide) brushes from charged polymeric surfaces by aqueous ATRP: Effect of surface initiator concentration. *Macromolecules* 2003, 36: 591-598.

[11]T. Nishimura. *Biomedical Applications of Polymeric Materials*, Eds., CRC Press, Boca Raton, 1993.

[12]P. Delanaye, B. Lambermount, J. M.Dongne, B.Dubois, A.Ghuysen, N. Janssen, T. Desaive, P. Kolh, V.D'Drio & J. M. Krzesinki. Confirmation of high cytokine clearance by hemofiltration with a cellulose triacetate membrane with large pores: an in vivo study. *Int. J. Artif. Organs* 2006, 29:944.

[13]H.D. Humes, W.H. Fissell & K. Tiranathanagul. The future of hemodialysis membranes. *Kidney Int* 2006, 69:1115-1119.

[14]Kung, F.C., Chou, W.L. & Yang, M.C. In vitro evaluation of

cellulose acetate hemodialyzer immobilized with heparin. *Polym Adv Tech* 2006, 17: 453-462.

[15]Yuan, J., Zhang, J., Zang, X.P., Shen, J. & Lin, S. Improvement of blood compatibility on cellulose membrane surface by grafting betaines. *Colloids Surf B Biointerf* 2003, 30:147-155.

[16]T. Furuzono, K. Ishihara, N. Nakabayashi & Y. Tamada. Chemical modification of silk fibroin with 2-methacryloyloxyethyl phosphorylcholine. II. Craft-polymerization onto fabric through 2-methacryloyloxyethyl isocyanate and interaction between fabric and platelets. *Biomaterials* 2000, 21:327-333.

[17]K. Ishiahra, R. Aragaki, T. Ueda, A. Watanabe & N. Nakabayashi. Reduced thrombogenicity of polymers having phospholipid polar groups. *J Biomed Mater Res* 1990, 24:1069-1077.

[18]Qi, Z. H. *Synthesis of CA by solid acid catalyst. BS Thesis*, University of Science and Technology of China, 2001.

[19]K. Fukumoto, K. Ishihara, R. Takayama, J. Aoki & N. Nakabayashi. Improvement of blood compatibility on cellulose dialysis membrane.2.blood compatibility of phospholipid polymer grafted cellulose membrane. *Biomaterials* 1992, 13: 235-239.

[20]P. Vlcek, M. Janata, P. Latalova, J. Kriz, E. Cadova, L. Toman. Controlled grafting of cellulose diacetate. *Polymer* 2006, 47:2587-2595.

[21]D. Bontempo, G. Masci, P.D. Leonardis, L. Mannina, D. Capitani & V. Crescenzi. Versatile grafting of polysaccharides in homogeneous mild conditions by using atom transfer radical polymerization. *Biomacromolecules* 2006, 7:2154-2161.

[22]E. Meaurio, L.C. Cesteros & I. Katime. FTIR study of hydrogen bonding of blends of poly(mono n-alkyl itaconates) with poly(N,N-dimethylacrylamide) and poly(ethyloxazoline). *Macromolecules* 1997, 30:4567-4573.

[23]Huynh-Ba-Gia & J.E. McGraph. High resolution NMR spectra of poly n,n-dimethylacrylamide in $CDCl_3$ solution. *Polym Bull* 1980, 2:837-840.

[24]B.L.Rivas, S.A. Pooley, M. Soto, H.A. Maturana & K.E. Geckeler. Poly(N,N'-dimethylacrylamide-co-acrylic acid): Synthesis, characterization, and application for the removal and separation of inorganic ions in aqueous solution. *J Appl Polym Sci* 1998, 67:93-100.

[25]Shen, D.W., Yu, H. & Huang, Y. Densely grafting copolymers of ethyl cellulose through atom transfer radical polymerization. *J Polym Sci Part A Polym Chem* 2005, 43:4099-4108.

[26]Shen, D., Yu, H. & Huang, Y. Synthesis of graft copolymer of ethyl cellulose through living polymerization and its self-assembly. *Cellulose* 2006, 13:235-244.

[27]Kang, H., Liu, W., He, B., Shen, D., Ma, L. & Huang, Y. Synthesis of amphiphilic ethyl cellulose grafting poly (acrylic acid) copolymers and their self-assembly morphologies in water. *Polymer* 2006, 47:7927-7934.

Scientific
Research
Publishing

# Simulation for chaos game representation of genomes by recurrent iterated function systems

Zu-Guo Yu [1,2,*], Long Shi [1], Qian-Jun Xiao [1] & Vo Anh [2]

[1]School of Mathematics and Computational Science, Xiangtan University, Hunan 411105, China. [2]School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Q 4001, Australia.* Correspondence should be addressed to Zu-Guo Yu (yuzg1970@yahoo.com).

## ABSTRACT

**Chaos game representation (CGR) of DNA sequences and linked protein sequences from genomes was proposed by Jeffrey (1990) and Yu *et al.* (2004), respectively. In this paper, we consider the CGR of three kinds of sequences from complete genomes: whole genome DNA sequences, linked coding DNA sequences and linked protein sequences. Some fractal patterns are found in these CGRs. A recurrent iterated function systems (RIFS) model is proposed to simulate the CGRs of these sequences from genomes and their induced measures. Numerical results on 50 genomes show that the RIFS model can simulate very well the CGRs and their induced measures. The parameters estimated in the RIFS model reflect information on species classification.**

**Keywords**: **Genomes; Chaos game representation; Recurrent iterated function systems**

## 1. INTRODUCTION

The hereditary information of organisms (except for RNA-viruses) is encoded in their DNA sequences which are one-dimensional unbranched polymers made up from four different kinds of monomers (nucleotides): adenine (*a*), cytosine (*c*), guanine (*g*), and thymine (*t*). Based on a technique from chaotic dynamics, Jeffrey (1990) proposed a chaos game representation (CGR) of DNA sequences by using the four vertices of a square in the plane to represent *a*,*c*,*g* and *t*. The method produces a plot of a DNA sequence which displays both local and global patterns. Self-similarity or fractal structures were found in these plots. Some open questions from the biological point of view based on the CGR were proposed (Jeffrey 1990).

If the DNA sequences were a random collection of bases, the CGR would be a uniformly filled square, conversely, any patterns visible in the CGR represent some pattern (information) in the DNA sequence (Goldman 1993). Goldman (1993) interpreted the CGRs in a biologically meaningful way. All points plotted within a quadrant must corresponding to subsequences of the DNA sequence that end with the base labelling the corner of that quadrant. He also proposed a discrete time Markov Chain model to simulate the CGR of DNA sequences and use the sequence's dinucleotide and trinucleotide frequencies to calculate the probabilities in these models. Goldman's Markov model can be calculated directly and easily from the raw DNA sequences, without reference to the CGR.

Deschavanne *et al.* (1999) used CGR of genomes to discuss the classification of species. Almeida *et al.* (2001) showed the distribution of positions in the CGR plane is a generalization of Markov Chain probability tables that accommodates non-integer orders. Joseph and Sasikumar (2006) proposed a fast algorithm for identifying all local alignments between two genome sequences using the sequence information contained in their CGR.

Twenty different kinds of amino acids are found in proteins. The idea of CGR of DNA sequences proposed by Jeffrey (1990) was generalized and applied for visualizing and analyzing protein databases by Fiser *et al.* (1994). Generalization of CGR of DNA may take place in several ways. In the simplest case, the square in CGR of DNA is replaced by an *n*-sided regular polygon (*n*-gon), where *n* is the number of different elements in the sequence to be represented. As proteins consist of 20 kinds of amino acids, a 20-sided regular polygon (regular 20-gon) is the most adequate for protein sequence representation. A few thousand points result in an 'attractor' which gives a visualization of the rare or frequent residues and sequence motifs. Fiser *et al.* (1994) pointed out that the chaos game representation can also be used to study 3D structures of proteins.

Published Online May 2008 in SciRes. http://www.srpublishing.org/journal/jbise

Basu *et al.* (1998) proposed a new method for the chaos game representation of different families of proteins. Using concatenated amino acid sequences of proteins belonging to a particular family and a 12-sided regular polygon, each vertex of which represents a group of amino acid residues leading to conservative substitutions, the method generates the CGR of the family and allows pictorial representation of the pattern characterizing the family. Basu *et al*. (1998) found that the CGRs of different protein families exhibit distinct visually identifiable patterns. This implies that different functional classes of proteins produce specific statistical biases in the distribution of different mono-, di-, tri-, or higher order peptides along their primary sequences.

A well-known model of protein sequence analysis is the HP model proposed by Dill *et al.* (1985). In this model 20 kinds of amino acids are divided into two types, hydrophobic (H) (or non-polar) and polar (P) (or hydrophilic). But the HP model may be too simple and lacks sufficient information on the heterogeneity and the complexity of the natural set of residues (Wang and Wang 2000). According to Brown (1998), one can divide the polar class in the HP model into three classes: positive polar, uncharged polar and negative polar. So 20 different kinds of amino acids can be divided into four classes: non-polar, negative polar, uncharged polar and positive polar. In this model, one considers more details than in the HP model. We call this model a *detailed HP model* (Yu *et al.*2004a). Based on the detailed HP model, we proposed a CGR for the linked protein sequences from the genomes (Yu *et al.* 2004b).

The recurrent iterated function system in fractal theory (Barnsley and Demko, 1985; Falconer, 1997) has been applied successfully to fractal image construction (Barnsley and Demko, 1985; Vrscay, 1991), one dimensional measure representation of genomes (Anh *et al.* 2002; Yu *et al.* 2001, 2003) and magnetic field data (Wanliss *et al.* 2005; Anh *et al.* 2005) for example. Yu *et al.* (2007) proposed a CGR for the magnetic field data and used the RIFS model to simulate the CGR.

Although we proposed the CGR for linked protein sequences from genomes (Yu *et al.* 2004b), we did not consider how to simulate the CGRs. In this paper, we extend the CGR to the study of whole-genome DNA sequences and linked coding DNA sequences from genomes. Then we use the RIFS model to simulate the CGR of these 3 kinds of data from genomes and their induced measures. The probability matrix in our RIFS model is similar to the one in Markov model used by Goldman (1993), but the way to estimate this matrix is different.

## 2. CHAOS GAME REPRESENTATION OF GENOMES

Three kinds of sequences from complete genomes are considered, namely, whole-genome DNA sequences (including protein-coding and non-coding regions), linked sequences of all protein-coding DNA sequences and linked sequences of all protein sequences from complete genomes.

For DNA sequences, the CGR is obtained by using the four vertices of a square in the plane to represent *a*,*c*,*g* and *t* (Jeffrey 1990). The first point of the plot is placed half way between the center of the square and the vertex corresponding to the first letter, the *i*th point of the plot is placed half way between the (*i*-1)th point and the vertex corresponding to the *i*th letter in the DNA sequence.

For linked protein sequences, we outline here the way to get the CGR from Yu *et al.* (2004b). The protein sequence is formed by twenty different kinds of amino acids, namely Alanine (*A*), Arginine (*R*), Asparagine (*N*), Aspartic acid (*D*), Cysteine (*C*), Glutamic acid (*E*), Glutamine (*Q*), Glycine (*G*), Histidine (*H*), Isoleucine (*I*), Leucine (*L*), Lysine (*K*), Methionine (*M*), Phenylalanine (*F*), Proline (*P*), Serine (*S*), Threonine (*T*), Tryptophan (*W*), Tyrosine (*Y*) and Valine (*V*) (Brown 1998, page 109). In the detailed HP model, they can be divided into four classes: non-polar, negative polar, uncharged polar and positive polar. The eight residues *A, I, L, M, F, P, W, V* designate the non-polar class; the two residues *D, E* designate the negative polar class; the seven residues *N, C, Q, G, S, T, Y* designate the uncharged polar class; and the remaining three residues *R, H, K* designate the positive polar class.

For a given protein sequence $s=s_{1}...s_{l}$ with length $l$, where $s_i$ is one of the twenty kinds of amino acids for $i=1,...,l$ ,we define

$$a_i=\begin{cases}0, & if \quad s_i \quad is \quad non-polar,\\ 1, & if \quad s_i \quad is \quad negative-polar,\\ 2, & if \quad s_i \quad is \quad uncharged-polar,\\ 3, & if \quad s_i \quad is \quad positive-polar,\end{cases} \qquad (1)$$

We then obtain a sequence $X(s)=a_1\cdots a_l$, where $a_i$ is a letter of the alphabet {0,1,2,3}. We next define the CRG for a sequence $X(s)$ in a square [0,1]× [ 0,1], where the four vertices correspond to the four letters 0,1,2,3. The first point of the plot is placed half way between the center of the square and the vertex corresponding to the first letter of the sequence $X(s)$; the *i*th point of the plot is then placed half way between the (*i*-1)th point and the vertex corresponding to the *i*th letter. We then call the obtained plot the CGR of the protein sequence *s* based on the detailed HP model.

Usually whole-genome DNA sequences and linked coding DNA sequences are relatively long, hence the resulting CGRs are too dense to visualize any pattern directly. The linked protein sequences are 3 times shorter than the linked coding DNA sequences, and their CGRs produce clearer self-similar patterns. For example, we show the CGR of the linked protein sequence of the bacterium *Mycobacterium tuberculosis* CDC1551 (MtubC) in **Figure 1**.

Considering the points in a CGR of an organism,

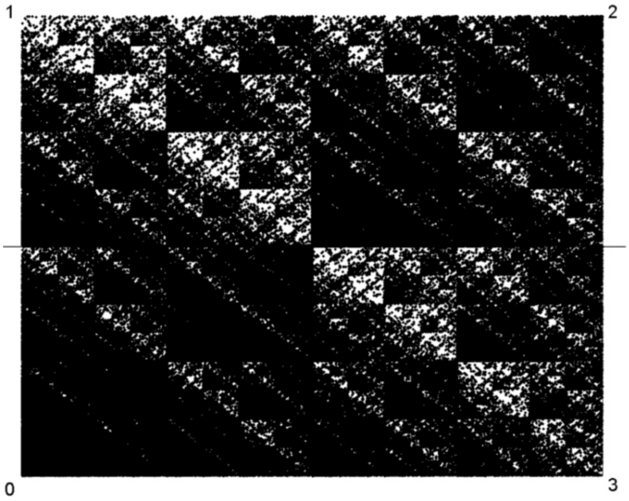


**Figure 1.** Chaos game representation of the linked protein sequence from genome of Mycobacterium tuberculosis CDC1551(MtubC) (with 1325681amino acids).

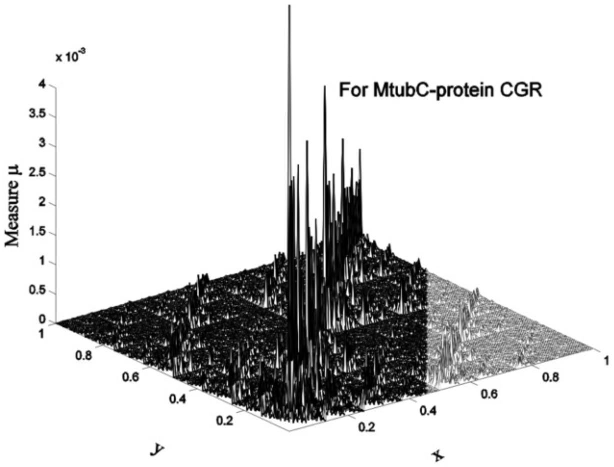


**Figure 2.** The measure $\mu$ based on a 128×128 mesh of the CGR in Figure 1.

we define a measure $\mu$ by $\mu(B)=\#(B)/N_l$, where $\#(B)$ is the number of points lying in a subset $B$ of the CGR and $N_l$ is the length of the sequence. We divide the square [0,1]×[0,1] into meshes of sizes 64×64, 128×128, 512×512 or 1024×1024. This results in a measure for each mesh. We then obtain a 64×64, 128×128, 512×512 or 1024×1024 matrix **A**$=(\mu_{kl})_{J\times J}$, where $J$=64,128,512 or 1024, each element $\mu_{kl}$ is the measure value on the corresponding mesh. We call **A** the *measure matrix* of the organism. The measure $\mu$ based on a 128×128 mesh on the CGRs are considered in this paper. For example, the measure $\mu$ based on a 128×128 mesh of the CGR in **Figure 1** is shown in **Figure 2**.

## 3. RECURRENT ITERATED FUNCTION SYSTEM FOR A MEASURE

Consider a system of contractive maps $S=\{S_1,S_2,\cdots,S_N\}$ and the associated matrix of probabilities $\boldsymbol{P}=(p_{ij})$ such that $\sum_j p_{ij}=1, i=1,2,\cdots,N$. We consider a random sequence generated by a dynamical system

$$x_{n+1}=S_{\sigma_n}(x_n), n=0,1,2,..., \qquad (2)$$

where $x_0$ is any starting point and $\sigma_n$ is chosen among the set {1,2,⋯,N}with a probability that depends on the previous index $\sigma_{n-1}$ : $P(\sigma_n=i)=p_{\sigma_{i-1},i}$. Then $(S, \boldsymbol{P})$ is called a *recurrent iterated function system*. Then there exist compact sets $A, A_i, i=1,2,\cdots, N$ such that

$$A=\bigcup_{i=1}^{N}A_i \qquad A_i=\bigcup_{j:p_{ji}>0}^{N}S_i(A_j)$$

where set $A$ is called the attractor of the RIFS $(S, \boldsymbol{P})$. A major result for RIFS is that there exists a unique invariant measure $\mu$ of the random walk (Eq. 2) whose support is $A$ (Barnsley *et al.*, 1989).

The coefficients in the contractive maps and the probabilities in the RIFS are the parameters to be estimated for the measure that we want to simulate. We now describe the method of moments to perform this task. In the two-dimensional case of our CGRs, we consider a system of $N$ contractive maps

$$S_i=s_i\begin{pmatrix}x\\y\end{pmatrix}+\begin{pmatrix}b_1(i)\\b_2(i)\end{pmatrix}, i=1,2,...,N$$

If $\mu$ is the invariant measure and $A$ the attractor of the RIFS in $\boldsymbol{R}^2$, the moments of $\mu$ are

$$g_{mn}=\int_A x^m y^n d\mu=\sum_{i=1}^{N}\int_{A_i}x^m y^n d\mu_i=\sum_{i=1}^{N}g_{mn}^{(i)}$$

Using the properties of the Markov operator defined by $(S, \boldsymbol{P})$ (Vrscay, 1991), we get

$$\begin{aligned}g_{mn}^{(i)}&=\int_{A_i}x^m y^n d\mu_i\\&=\sum_{j=1}^{N}p_{ji}\int_{A_j}(s_j x+b_1(j))^m(s_j y+b_2(j))^n d\mu_j\\&=\sum_{j=1}^{N}p_{ji}\sum_{k=0}^{m}\sum_{l=0}^{n}\begin{pmatrix}m\\n\end{pmatrix}\begin{pmatrix}n\\l\end{pmatrix}s_j^{k+l}b_1(j)^{m-k}b_2(j)^{n-l}g_{kl}^{j}\end{aligned} \qquad (3)$$

When $n$=0, $m$=0, from $\sum_{j=1}^{N}g_{00}^{(j)}=1$ we have

$$g_{00}^{(i)}=\sum_{j=1}^{N}p_{ji}g_{00}^{(j)}\Rightarrow\sum_{j=1}^{N}(p_{ji}-\delta_{ji})g_{00}^{(j)}=0 \qquad (4)$$

for $i=1,2,\cdots,N$.

Then we can get the values for $g_{00}^{(j)}$, $j=1, 2,\cdots,N$ by solving the above linear equations.

When $m$=0, n⩾1

$$g_{0n}^{(i)} = \sum_{j=1}^{N} p_{ji} \sum_{l=0}^{n} \binom{n}{l} s_j^l b_2(j)^{n-l} g_{0l}^{(j)}$$

hence the moments are given by the solution of the linear equations

$$\sum_{j=1}^{N} (s_j^n p_{ji} - \delta_{ji}) g_{0n}^{(j)}$$

$$= -\sum_{l=0}^{n-1} \binom{n}{l} \sum_{j=1}^{N} s_j^l b_2(j)^{n-l} g_{0l}^{(j)}, i = 1, \ldots N. \qquad (5)$$

When $n$=0,$m$⩾1

$$g_{m0}^{(i)} = \sum_{j=1}^{N} p_{ji} \sum_{k=0}^{m} \binom{m}{k} s_j^k b_1(j)^{m-k} g_{k0}^{(j)}$$

hence the moments are given by the solution of the linear equations

$$\sum_{j=1}^{N} (s_j^m p_{ji} - \delta_{ji}) g_{m0}^{(j)}$$

$$= -\sum_{k=0}^{m-1} \binom{m}{k} \sum_{j=1}^{N} s_j^k b_1(j)^{m-k} g_{k0}^{(j)}, i = 1, \ldots N. \qquad (6)$$

When m,n⩾1

$$g_{mn}^{(i)} =$$

$$\sum_{j=1}^{N} p_{ji} \sum_{k=0}^{m-1} \sum_{l=0}^{n} \binom{m}{n} \binom{n}{l} s_j^{k+l} b_1(j)^{m-k} b_2(j)^{n-l} g_{kl}^{(j)}$$

$$+\sum_{j=1}^{N} p_{ji} \sum_{l=0}^{n-1} \binom{n}{l} s_j^{m+l} b_2(j)^{n-l} g_{ml}^{(j)} + \sum_{j=1}^{N} p_{ji} s_j^{m+n} g_{mn}^{(j)},$$

hence the moments are given by the solution of the linear equations

$$\sum_{j=1}^{N} (s_j^{m+n} p_{ji} - \delta_{ji}) g_{mn}^{(j)} =$$

$$\sum_{k=0}^{m-1} \sum_{l=0}^{n-1} \binom{m}{n} \binom{n}{l} \sum_{j=1}^{N} p_{ji} s_j^{k+l} b_1(j)^{m-k} b_2(j)^{n-l} g_{kl}^{(j)}$$

$$-\sum_{l=0}^{n-1} \binom{n}{l} \sum_{j-1}^{N} p_{ji} s_j^{m+l} b_2(j)^{n-l} g_{ml}^{(j)}$$

$$-\sum_{k=0}^{m-1} \binom{m}{k} \sum_{j-1}^{N} p_{ji} s_j^{k+n} b_1(j)^{m-k} g_{kn}^{(j)} \qquad (7)$$

for $i$=1,2,…,$N$.

If we denote by $G_{mn}$ the moments obtained directly from a given measure, and $g_{mn}$ the formal expression of moments obtained from the above formulae, then solving the optimization problem

$$\min_{s_i, b_1(i), b_2(i), p_{ij}} \sum_{m,n} (g_{mn} - G_{mn})^2$$

will provide the estimates of the parameters of the RIFS.

Once the RIFS ($S_i(x)$,$p_{ij}$,$i$,$j$=1,2,⋯,$N$ ) has been estimated, its invariant measure can be simulated in the following way: Generate the attractor of the RIFS via the random walk (Eq. 2). Let $\chi_B$ be the indicator function of a subset $B$ of the attractor $A$. From the ergodic theorem for RIFS (Barnsley *et al.*, 1989), the invariant measure is then given by

$$\mu(B) = \lim_{n\to\infty} \left[ \frac{1}{n+1} \sum_{k=0}^{n} \chi_B(x_k) \right]$$

By definition, an RIFS describes the scale invariance of a measure. Hence a comparison of the given measure with the invariant measure simulated from the RIFS will confirm whether the given measure has this scaling behaviour. This comparison can be undertaken by computing the cumulative walk of a measure visualized as intensity values on a $J\times J$ mesh; here $J$=128 in this paper.

If we convert the two-dimensional matrix $\mathbf{A}=(\mu_{kl})_{J\times J}$ to an one dimensional vector by concatenate every row in $\mathbf{A}$ at the end of previous row. We denote the one-dimensional vector as $f=(f_1,f_2,\cdots,f_{J\times J})$. The cumulative walk is defined as

$$F_j = \sum_{i=1}^{j} (f_i - \overline{f}), \; j = 1,2,\ldots, J\times J$$

Where $\overline{f}$ is the average value of all element in vector $f$.

Returning to the CGR, an RIFS with 4 contractive maps $\{S_1,S_2,S_3,S_4\}$ is fitted to the measure obtained from the CGR using the method of moments. Here we can fix

$$S_1 = \frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix} \qquad S_2 = \frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0 \\ 0.5 \end{pmatrix}$$

$$S_3 = \frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} \qquad S_4 = \frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0.5 \\ 0 \end{pmatrix}$$

Hence the parameters needed to be estimated are the probabilities in the matrix ***P***. Once we have estimated the probability matrix in the RIFS, we can start from the point (0.5, 0.5) and use the chaos game algorithm Eq. (2) to generate a random point sequence $\{x_i\}$ with the same length $N_l$ of the whole- genome DNA sequence, linked coding DNA sequence or the linked

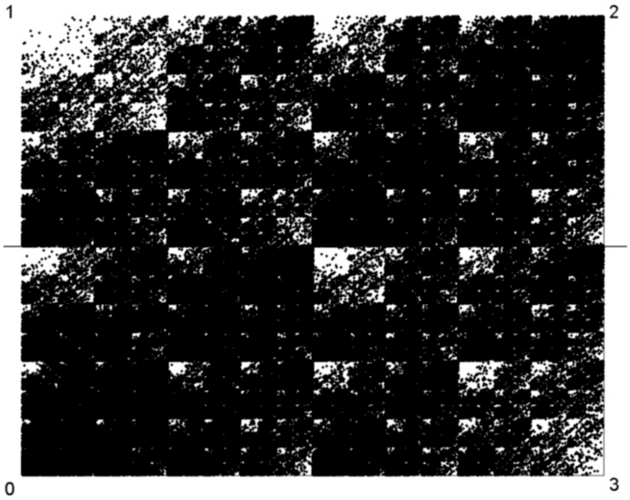


**Figure 3.** The RIFS simulated CGR for the CGR in Figure 1.

protein sequence. Then the plot of the random point sequences is the RIFS simulation of the original CGR of the data. For example the RIFS simulated CGR of the CGR in **Figure 1** is shown in **Figure 3**. Comparing the RIFS simulation in **Figure 3** with the original CGR in **Figure 1**, it is apparent that they are quite similar. We then obtain the 128× 128 mesh measure $\mu'$ based on the simulated CGR. The measure $\mu'$ can be regarded as a simulation of the measure $\mu$ induced from the original CGR. For example, we show the 128× 128 mesh measure $\mu'$ based on the simulated CGR of **Figure 3** in **Figure 4**. The cumulative walks of these two measures can then be obtained to show the performance of the simulation.

We determine the goodness of fit of the measure simulated from the RIFS model relative to the original measure based on the following *relative standard error* (*RSE*) (Anh *et al*. 2002):

$$e = \frac{e_1}{e_2}$$

Where

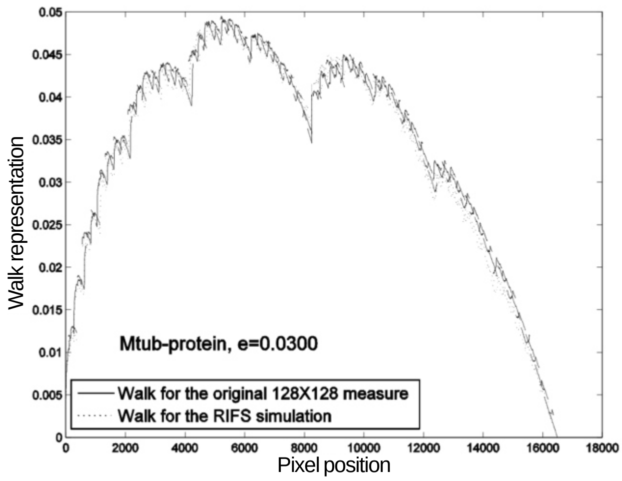


**Figure 5.** The walk representation of measures induced by the CGR in Figure 1 and its RIFS simulation in Figure 4.

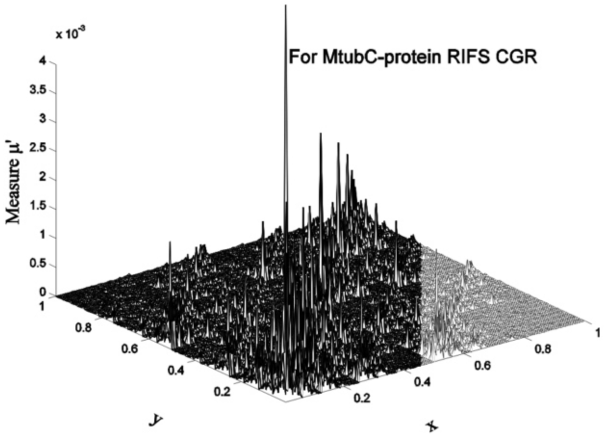


**Figure 4.** The measure $\mu'$ based on a 128× 128 mesh of the RIFS simulated CGR in Figure 3.

$$e_1 = \sqrt{\frac{1}{M}\sum_{j=1}^{M}(F_j - \widehat{F}_j)^2}$$

and

$$e_2 = \sqrt{\frac{1}{M}\sum_{j=1}^{M}(F_j - F_{ave})^2}$$

Here $M$=128×128, $(F_j)_{j=1}^{M}$ and $(\widehat{F}_j)_{j=1}^{M}$ are the walks of the original measure and the RIFS simulated measure respectively, $F_{ave}$ is the mean value of $(F_j)_{j=1}^{M}$.

The goodness $e < 1.0$ indicates the simulation is very well (Anh *et al*. 2002). For example, the cumulative walks for the measure induced by the CGR in **Figure 1** and its RIFS simulation in **Figure 4** are given in **Figure 5**. It is seen that the two walks are almost identical. This indicates that RIFS fits very well the measure induced by the original CGR. The RSE $e$=0.0300 is very small, which also indicates excellent fitting.

## 4. DATA, DISCUSSION AND CONCLUSION

We downloaded whole-genome DNA sequences, coding DNA sequences and protein sequences from 50 complete genomes of Archaea and Eubacteria from the public database Genbank at the web site http://www.ncbi.nlm.nih.gov/Genbank/. We list the name of the 50 bacteria in Appendix.

We then produce the CGRs of the data from the 50 genomes as described in **Section 2**. For more examples, we plot the chaos game representation of the linked coding sequence from genome of *Mycoplasma pulmonis* UAB CTIP (Mpul) in **Figure 6**. Fractal (self-similarity) patterns can be seen in these CGRs. We only use the moments of 128× 128 mesh measure $\mu$ based on the CGRs to estimate the parameters (probability matrix) in the RIFS model. Then the RIFS simulation of the original CGRs is performed using the chaos game algorithm. We then get

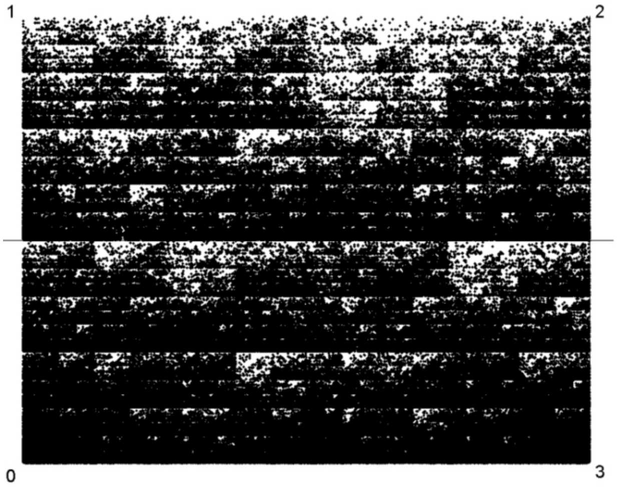


**Figure 6.** Chaos game representation of the linked coding sequence from genome of *Mycoplasma pulmonis* UAB CTIP (Mpul) (with 873,651 bps).

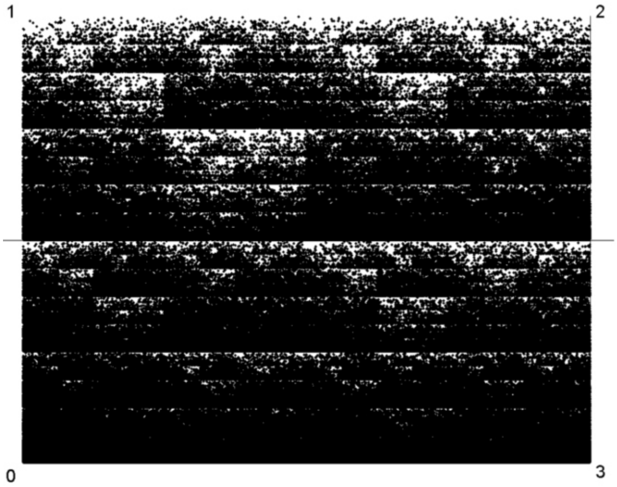


**Figure 7.** The RIFS simulated CGR for the CGR in Figure 6.

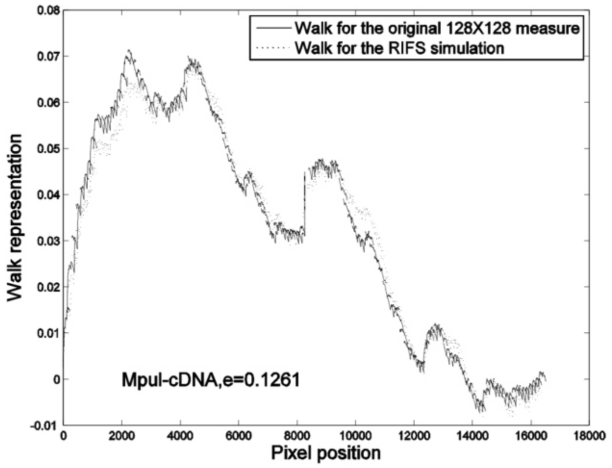


**Figure 8.** The walk representation of measures induced by the CGR in Figure 6 and its RIFS simulation in Figure 7.

the 128×128 mesh measure $\mu'$ based on the simulated CGR. To show the performance of the simulation, we compare the cumulative walks of the original measure and its simulation $\mu'$. For example, the RIFS simulated CGR of the linked coding sequence from genome of *Mycoplasma pulmonis* UAB CTIP (Mpul) based on the 128×128 mesh measure $\mu$ from **Figure 6** is shown in **Figure 7**, while the walk representation of measures induced by the CGR in **Figure 6** and its RIFS simulation in **Figure 7** are shown in **Figure 8**.

Goldman (1993) interpreted the patterns in CGRs of DNA sequences by the dinucleotide and trinucleotide frequencies in the original sequence. The probability matrix in our RIFS model characterizes the dinucleotide or di-amino acid frequencies (information) which is similar to the one in Markov model used by Goldman (1993), but the way to estimate this matrix is different.

The values of the RSE of the simulation for 50

**Table 1.** The goodness of fit for the walk representations of three kinds of data from 50 genomes.

| Species (abbrev.) | e for whole DNA | e for coding DNA | e for linked proteins |
|---|---|---|---|
| Aful | 0.5797 | 0.2669 | 0.0366 |
| Paby | 0.3502 | 0.3214 | 0.0333 |
| Pyro | 0.4324 | 0.3411 | 0.0361 |
| Mjan | 0.2136 | 0.2675 | 0.0647 |
| haloNRC | 0.3728 | 0.3569 | 0.0297 |
| Taci | 0.2707 | 0.2735 | 0.1030 |
| Tvol | 0.3126 | 0.2716 | 0.1308 |
| Mthe | 0.5188 | 0.5676 | 0.0299 |
| Aero | 0.6213 | 0.2222 | 0.0452 |
| Ssol | 0.3798 | 0.3612 | 0.1098 |
| MtubH | 1.3037 | 0.5862 | 0.0333 |
| MtubC | 1.3010 | 0.5711 | 0.0300 |
| Mlep & | 0.4271 | 0.3332 | 0.0404 |
| Mpneu | 0.0484 | 0.0589 | 0.1686 |
| Mgen | 0.0731 | 0.2305 | 0.2617 |
| Mpul | 0.0639 | 0.1261 | 0.2267 |
| Uure | 0.0783 | 0.2064 | 0.4058 |
| Bsub | 0.4051 | 0.8012 | 0.0684 |
| Bhal | 0.1198 | 0.2652 | 0.0489 |
| Llac | 0.1032 | 0.1879 | 0.0500 |
| Spyo | 0.1049 | 0.1759 | 0.0678 |
| Spne | 0.1125 | 0.1358 | 0.0932 |
| SaurN | 0.1264 | 0.2728 | 0.1020 |
| SaurM | 0.1229 | 0.2680 | 0.1054 |
| CaceA | 0.1887 | 0.1693 | 0.1859 |
| Aqua | 0.4825 | 0.3457 | 0.0661 |
| Tmar | 0.4470 | 0.6674 | 0.0597 |
| Ctra | 0.8986 | 0.4769 | 0.1066 |
| Cpneu | 0.7786 | 0.7170 | 0.1312 |
| CpneuA | 0.7593 | 0.7093 | 0.1044 |
| CpneuJ | 0.7899 | 0.7352 | 0.1290 |
| Syne | 0.0521 | 0.0396 | 0.0667 |
| Nost | 0.1411 | 0.1439 | 0.0931 |
| Bbur | 0.1466 | 0.1255 | 0.2008 |
| Tpal | 0.3068 | 0.1212 | 0.0908 |
| Atum | 0.2614 | 0.2655 | 0.0403 |
| smel | 0.1739 | 0.1957 | 0.0380 |
| Ccre | 0.1171 | 0.1558 | 0.0259 |
| Rpro | 0.3887 | 0.7126 | 0.2132 |
| Nmen | 0.1973 | 0.1933 | 0.0430 |
| NmenA | 0.2039 | 0.1993 | 0.0559 |
| EcoliKM | 0.3225 | 0.3472 | 0.0714 |
| EcoliOH | 0.3222 | 0.3810 | 0.0868 |
| Hinf | 0.0677 | 0.2388 | 0.0883 |
| Xfas | 0.1246 | 0.1460 | 0.0324 |
| Paer | 0.2149 | 0.1823 | 0.0470 |
| Pmul | 0.1032 | 0.2087 | 0.0911 |
| Buch | 0.1954 | 0.2598 | 0.3911 |
| Hpyl | 0.2567 | 0.2615 | 0.1161 |
| Cjej | 0.1540 | 0.1797 | 0.0802 |

genomes are listed in **Table 1**.

It is seen that all the values of the RES except two are much less than 1.0, confirming that the RIFS model can simulate very well the measures of three kinds of data. The values of *e* for whole-genome DNA data are generally larger than those for linked coding DNA data, which in turn are larger than those for linked protein data. In other words, the RIFS model can simulate the measures for linked protein data better than the measures for linked coding DNA data, and can simulate measures for linked coding DNA data better than the measures for whole-genome DNA data. We notice that the linked protein sequence is shorter than the corresponding linked coding DNA sequence, while the linked coding DNA is shorter than the whole-genome sequence. We guess the length of the data reflects the information complexity of the data and the RIFS model is still simple model which simulates simpler data better. This result indicates that we can use the estimated parameters in the RIFS model for linked protein data from genomes to characterize the genomes. We find that the estimated probability matrices in the RIFS model for species from the same category are similar to each other. For example, the estimated probability matrices for the measures of linked protein sequences from the three **Gram-positive Eubacteria (high G+C)** *Mycobacterium tuberculosis* H37Rv (MtubH), *Mycobacterium tuberculosis* CDC1551 (MtubC) and *Mycobacterium leprae* TN (Mlep) are:

$$P_{MtubH}=\begin{pmatrix} 0.495551 & 0.149496 & 0.215737 & 0.139217 \\ 0.410094 & 0.027692 & 0.286638 & 0.275576 \\ 0.421544 & 0.096754 & 0.354118 & 0.127584 \\ 0.386300 & 0.263546 & 0.266087 & 0.084086 \end{pmatrix}$$

$$P_{MtubC}=\begin{pmatrix} 0.496060 & 0.146193 & 0.218983 & 0.138764 \\ 0.413542 & 0.028024 & 0.282788 & 0.275647 \\ 0.419026 & 0.101162 & 0.344503 & 0.135310 \\ 0.388569 & 0.259119 & 0.267148 & 0.085164 \end{pmatrix}$$

$$P_{Mlep}=\begin{pmatrix} 0.490039 & 0.143671 & 0.226108 & 0.140182 \\ 0.414127 & 0.038055 & 0.272109 & 0.275709 \\ 0.406399 & 0.123836 & 0.313224 & 0.156541 \\ 0.399737 & 0.260004 & 0.293543 & 0.046717 \end{pmatrix}$$

Hence we can use the RIFS estimated probability matrices of the linked protein sequences from genomes to define a distance metric between two species for the purpose of construction of phylogenetic tree. This work is being undertaken.

We can now draw some conclusions. First, the chaos game representation of the three kinds of data from genomes can give a visualization of the genomes and produce some fractal patterns. Second, the RIFS model can be used to simulate CGRs of genomes and their induced measures. Third, the RIFS simulation of measures for linked protein data is better than that of measures for whole-genome DNA data and linked coding DNA data. Finally, the estimated parameters in the RIFS models for the linked protein data from genomes can be used to characterize the phylogenetic relationships of the genomes.

## ACKNOWLEDGEMENTS

Financial support was provided by the Chinese National Natural Science Foundation (grant no. 30570426), Fok Ying Tung Education Foundation (grant no. 101004) and the Youth foundation of Educational Department of Hunan province in China (grant no. 05B007) (Z.-G. Yu), and the Australian Research Council (grant no. DP0559807) (V.V. Anh).

## REFERENCE

[1]J.S. Almeida, J.A. Carrico, A. Maretzek, P.A. Noble & M. Fletcher. Analysis of genomic sequences by Chaos Game Representation. *Bioinformatics* 2001,17:429-437.

[2]V.V. Anh, K.S. Lau, & Z.G. Yu. Recognition of an organism from fragments of its complete genome. *Phys. Rev. E* 2002, 66(031910):1-9.

[3]V.V. Anh, Z.G. Yu, J.A. Wanliss, & S.M. Watson. Prediction of magnetic storm events using the Dst index. *Nonlin. Processes Geophys*. 2005 , 12:799-806.

[4]M.F. Barnley, J.H. Elton & D.P. Hardin. Recurrent iterated function systems. *Constr. Approx. B* 1989, 5: 3-31.

[5]M.F. Barnsley & S. Demko. Iterated function systems and the global construction of fractals. *Proc. R. Soc. London, Ser. A* 1985, 399:243-275.

[6]S. Basu, A. Pan, C. Dutta & J. Das. Chaos game representation of proteins. *J. Mol. Graphics and Modelling* 1998, 15:279-289.

[7]T.A. Brown. *Genetics* (3rd Edition) 1998. CHAPMAN & HALL, London.

[8]P.J. Deschavanne, A Giron, J. Vilain, G. Fagot & B. Fertil. Genomics signature: Characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol* 1999, 16:1391-1399.

[9]K.A.Dill. Theory for the folding and stability of globular Proteins. *Biochemistry* 1985, 24:1501-1509.

[10]K. Falconer. *Techniques in Fractal Geometry* 1997, Wiley.

[11]A. Fiser, GE Tusnady & I. Simon. Chaos game representation of protein structures. *J. Mol. Graphics* 1994, 12:302-304.

[12]N. Goldman. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences.

[13]H.J. Jeffrey. Chaos game representation of gene structure. *Nucleic Acids Research* 1990, 18(8): 2163-2170.

[14]J.Joseph & R. Sasikumar. Chaos game representation for comparision of whole genomes. *BMC Bioinformatics* 2006, 7(243): 1-10.

[15]E.R. Vrscay. Iterated function systems: theory, applications and inverse problem. *Fractal Geometry and Analysis* 1991, pages 405-468.

[16]J. Wang & W. Wang. Modeling study on the validity of a possibly simplified representation of proteins. *Phys. Rev. E* 2000, 61:6981-6986.

[17]J.A. Wanliss, V.V. Anh, Z.G. Yu & S. Watson. Multifractal modelling of magnetic storms via symbolic dynamics analysis. *J. Geophys. Res.* 2005, 110(A08214):1-11.

[18]Z.G. Yu, V.V. Anh & K.S. Lau. Measure representation and multifractal analysis of complete genomes. *Phys. Rev. E* 2001, 64(031903):1-9.

[19]Z.G. Yu, V.V. Anh & K.S. Lau. Iterated function system and multifractal analysis of biological sequences. *International J. Modern Physics B* 2003, 17: 4367-4375.

[20]Z.G. Yu, V.V. Anh, and K.S. Lau, "Fractal analysis of large proteins based on the Detailed HP model", *Physica A*, 337 (2004a), pp. 171-184.

[21]Z.G. Yu, V.V. Anh & K.S. Lau. Chaos game representation, and

multifractal and correlation analysis of protein sequences from complete genome based on detailed HP model. *J. Theor. Biol.* 2004, 226(3): 341-348.

[22]Z.G. Yu, V.V. Anh, J.A. Wanliss & S.M. Watson. Chaos game representation of the Dst index and prediction of geomagnetic storm events. *Chaos, Solitons & Fractals* 2007, 31:736-746.

## APPENDIX

These 50 bacteria include eight **Archae Euryarchaeota**: *Archaeoglobus fulgidus* DSM 4304 (Aful, NC000917), *Pyrococcus abyssi* GE5 (Paby, NC000868), *Pyrococcus horikoshii* OT3 (Pyro, NC000961), *Methanococcus jannaschii* DSM 2661 (Mjan, NC000909), *Halobacterium* sp. NRC-1 (haloNRC, NC002607), *Thermoplasma acidophilum* DSM 1728 (Taci, NC002578), *Thermoplasma volcanium* GSS1 (Tvol,NC002689), and *Methanobacterium thermoautotrophicum* deltaH (Mthe, NC000916); two **Archae Crenarchaeota**: *Aeropyrum pernix* K1 (Aero, NC000854) and *Sulfolobus solfataricus* P2 (Ssol, NC002754); three **Gram-positive Eubacteria (high G+C)**: *Mycobacterium tuberculosis* H37Rv (MtubH, NC000962), *Mycobacterium tuberculosis* CDC1551 (MtubC, NC002755) and *Mycobacterium leprae* TN (Mlep, NC002677); twelve **Gram-positive Eubacteria (low G+C)**: *Mycoplasma pneumoniae* M129 (Mpneu, NC000912), *Mycoplasma genitalium* G37 (Mgen, NC000908), *Mycoplasma pulmonis* UAB CTIP (Mpul, NC002771), *Ureaplasma urealytiaum* serovar 3 str. ATCC 700970 (Uure, NC002162), *Bacillus subtilis* subsp. subtilis str. 168 (Bsub, NC000964), *Bacillus halodurans* C-125 (Bhal, NC002570), *Lactococcus lactis* subsp. lactis Il1403 (Llac, NC002662), *Streptococcus pyogenes* M1 GAS (Spyo, NC002737), *Streptococcus pneumoniae* TIGR4 (Spne, NC003028), *Staphylococcus aureus* subsp. aureus N315 (SaurN, NC002745), *Staphylococcus aureus* subsp. aureus Mu50 (SaurM, NC002758), and *Clostridium acetobutylicum* ATCC 824 (CaceA, NC003030). The others are **Gram-negative Eubacteria**, which consist of two **hyperthermophilic bacteria**: *Aquifex aeolicus* VF5 (Aqua, NC000918) and *Thermotoga maritima* MSB8 (Tmar, NC000853); four **Chlamydia**: *Chlamydia trachomatis* D/UW-3/CX (Ctra, NC000117), *Chlamydia pneumoniae* CWL029 (Cpneu, NC000922), *Chlamydia pneumoniae* AR39 (CpneuA, NC002179) and *Chlamydia pneumoniae* J138 (CpneuJ, NC002491); two **Cyanobacterium**: *Synechocystis* sp. PCC6803 (Syne, NC000911) and *Nostoc sp. PCC7120* (Nost, NC003272); two **Spirochaete**: *Borrelia burgdorferi* B31 (Bbur, NC001318) and *Treponema pallidum* Nichols (Tpal, NC000919); and fifteen **Proteobacteria**. The fifteen Proteobacteria are divided into four subdivisions, namely **alpha subdivision**: *Agrobacterium tumefaciens* strain C58 (Atum, NC003062), *Sinorhizobium meliloti* 1021 (smel, NC003047), *Caulobacter crescentus* CB15 (Ccre, NC002696) and *Rickettsia prowazekii* Madrid (Rpro, NC000963); **beta subdivision**: *Neisseria meningitidis* MC58 (Nmen, NC003112) and *Neisseria meningitidis* Z2491 (NmenA, NC003116); **gamma subdivision**: *Escherichia coli* K-12 MG1655 (EcoliKM, NC000913), *Escherichia coli* O157:H7 EDL933 (EcoliOH, NC002695), *Haemophilus influenzae* Rd (Hinf, NC000907), *Xylella fastidiosa* 9a5c (Xfas, NC002488), *Pseudomonas aeruginosa* PA01 (Paer, NC002516), *Pasteurella multocida* subsp. multocida str. Pm70 (Pmul, NC002663) and *Buchnera* str. APS (Buch, NC002528); and **epsilon subdivision**: *Helicobacter pylori* 26695 (Hpyl, NC000915) and *Campylobacter jejuni* subsp. jejuni NCTC 11168 (Cjej, NC002163). The abbreviations in the brackets stand for the names of these species and their NCBI accession numbers.

J. Biomedical Science and Engineering, 2008, 1, 52-58

Scientific Research Publishing

# A combinatorial analysis of genetic data for Crohn's disease

**Weidong Mao**[1] **& Jeonghwa Lee**[2]

[1]Department of Mathematics & Computer Science, Virginia State University, Petersburg, VA 23806, USA. [2]Department of Computer Science, Shippensburg University, Shippensburg, PA 17257, USA. Correspondence should be addressed to Weidong Mao (wmao@vsu.edu) or Jeonghwa Lee(jlee@ship.edu).

## ABSTRACT

**The both environmental and genetic factors have roles inthe developmentof somediseases. Complex diseases, suchas Crohn'sdisease orTypeII diabetes, are caused by a combination of environmental factors and mutations in multiple genes. Patients who have beendiagnosed withsuch diseasescannot easily be treated. However, many diseases can be avoided ifpeople athigh riskchange theirliving style, one examplebeing theirdiet. Buthow canwe telltheir susceptibility to diseases before symptoms are found andhelp themmake informeddecisions about their health? With the development of DNA microarray technique, it is possible to access the human genetic information related to specific diseases. This paper uses a combinatorial method to analyze the genetic data for Crohn's disease and search disease-associated factors for given case/control samples. An optimum random forest based methodhas beenapplied topublicly available genotype data on Crohn's disease for association study andachieved apromising result.**

**Keywords:** Genetic factor; Crohn's disease; Random forest

## 1. INTRODUCTION

Crohn's disease (also known as regional enteritis) is a chronic, episodic, inflammatory condition of the gastrointestinal tract characterized by transmural inflammation (affecting the entire wall of the involved bowel) and skip lesions (areas of inflammation with areas of normal lining in between). Crohn's disease is a type of inflammatory bowel disease (IBD) and can affect any part of the gastrointestinal tract from mouth to anus. As a result, the symptoms of Crohn's disease can vary among affected individuals. The exact cause of Crohn's disease is unknown. However, research shows that the inflammation seen in the people with Crohn's disease involves several factors: the genes the patient has inherited, the immune system itself, and the environment [1]. In other words, genetic factor has been invoked in the pathogenesis of the disease.

Although the Crohn's disease cannot easily be treated, it can be avoided if people at high risk change their living style, such as their diet. But how can we tell the susceptibility of people to the disease before symptoms are found and help them make informed decisions about their health? With the development of DNA microarray technique, it is possible to access the human genetic information related to specific diseases. Assessing the association between DNA variants and disease has been used widely to identify regions of the genome and candidate genes that contribute to disease [2].

99.9% of one individual's DNA sequences are identical to that of another person. Over 80% of this 0.1% difference will be Single Nucleotide Polymorphisms (SNP) and they promise to significantly advance our ability to understand and treat human disease. A SNP is a single base substitution of one nucleotide with another. Each individual has many single nucleotide polymorphisms that together create a unique DNA pattern for that person. It is important to study SNPs because they represent genetic differences among human beings. Genome-wide association studies require knowledge about common genetic variations and the ability to genotype a sufficiently comprehensive set of variants in a large patient sample [3]. High-throughput SNP genotyping technologies make massive genotype data, with a large number of individuals, publicly available. Accessibility of genetic data makes genome-wide association studies for complex diseases possible.

Success stories when dealing with diseases caused by a single SNP or gene, sometimes called monogenic diseases have been reported [4]. However, most complex diseases, such as psychiatric disorders, are characterized by a non-mendelian, multifactorial genetic contribution with a number of susceptible genes interacting with each other [5]. A fundamental issue in the analysis of SNP data is to define the unit of genetic function that influences disease risk. Is it a single SNP, a regulatory motif, an encoded protein subunit, a combination of SNPs in a combination of genes, an interacting protein complex, a metabolic or a physiological pathway [6]? In general, it may be impossible to associate a single SNP or gene with a disease because a disease may be caused by completely different modifications of alternative pathways, and each gene only makes a small contribution. This makes the identification of genetic factors difficult. Multi-SNP interaction analysis is more reliable but it is computationally infeasible. An exhaustive search among multi-SNP combination is computationally infeasible even for a small number of SNPs. Furthermore, there are no reliable tools applicable to large genome ranges that could rule out or confirm association with a disease.

It is important to search for informative SNPs among a huge number of SNPs. These informative SNPs are assumed to be associated with genetic diseases. Tag SNPs generated by the multiple linear regression based method [7] are good informative SNPs, but they are reconstruction-oriented instead of disease-oriented. Although the combinatorial search method [8] for finding disease-associated multi-SNP combinations has a better result, the exhaustive search is still very slow.

Multivariate adaptive regression spline models [9, 10] are used to detect associations between diseases and SNPs with some degree of success. However, the number of selected predictors is limited, and the type of possible interactions must be specified in advance. Multifactor dimensionality reduction methods [11, 12] are developed specifically to find gene-gene interactions among SNPs, but they are not applicable to a large set of SNPs.

Random forest model has been explored in disease association studies [13], but it was applied on simulated case-control data in which the interacting model among SNPs and the number of associated SNPs are specified, thus making the association model simple and the association is relatively easier to detect. For real data, such as Crohn's disease [14], multi-SNP interaction is much more complex , which involves more SNPs.

In Section 2 of this paper, we propose an optimum random forest model for searching the disease-associated multi-SNP combination for given case-control data. In the optimum random forest model, we generate a forest for each variable (e.g. SNP) instead of randomly selecting some variables to grow the classification tree. We can find the best classifier (a combination of SNPs which includes the SNP) for each SNP, and then we may have *M* classifiers if the length of the genotype is *M*. We rank classifiers according to their prediction rate, and the SNP with a higher prediction rates is more disease-associated.

The association of multi-SNP combination can be measured by the disease susceptibility prediction rate. In Section 3 we address the disease susceptibility prediction problem [15, 16, 17, 18]. The goal of disease susceptibility prediction is to assess accumulated information targeted to predicting susceptibility to complex diseases with significantly high accuracy and statistical power. The problem is based on the association study we described above. The Disease-associated multi-SNP combination found in association studies can be used to predict the susceptibility to diseases. On the other side, the prediction results can be used to evaluate the accuracy of the association studies. A higher prediction rate means the higher reliability of the association studies.

The proposed method is applied to analyze the genetic data of the Crohn's disease. We find the disease-associated multi-SNP combination and apply it to predict the susceptibility. The accuracy of the prediction is higher than that of all previously known methods. It can be also applied in disease prevention and control in the near future. For example, after training the available case-control genome data, we can find those significant SNPs which are well associated with the disease. When a patient comes, and we obtain his/her genetic data, we don't need to check the whole sequence, but only disease-associated SNPs instead. This will save much money and time for diagnosis and can be done before the onset of diseases. Therefore, treatment could start earlier to prevent or delay the occurrence of the disease.

## 2. DISEASE ASSOCIATION SEARCH FOR CROHN'S DISEASE

In this section we first give an overview of the random forest tree and classification tree, then we will describe the genetic model. Next we propose the optimum random forest algorithm to search Tag SNPs.

### 2.1. Classification Trees and Random Forest

In machine learning, a Random Forest is a classifier that consists of many classification trees. Each tree is grown as follows:

1. If the number of cases in the training set is *N*, sample *N* cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are *M* input variables, a number $m<<M$ is specified such that at each node, *m* variables are selected randomly out of the *M* and the best split on these m is used to split the node. The value of *m* is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning [19].

A different bootstrap sample from the original data is used to construct a tree. Therefore, about one-third of the cases are left out of the bootstrap sample and not used in the construction of the tree. Cross-validation is not required because the one-third **oob (out-of-bag)** data is used to get an unbiased estimate of the classification error as trees are added to the forest. It is also used to get estimates of variable importance. After each tree is built, we compute the proximities of each terminal node.

In every classification tree in the forest, put down the **oob** samples and make prediction the classification of the **oob** samples. In such way we can compute the importance score for variables in each tree based on the number of votes cast for the correct class. All variables can be ranked and those important variables can be found in this way.

Random forest is a sophisticated method in data mining to solve classification problems, and it can be used efficiently in disease association studies to find most disease-associated variables such as SNPs that may be responsible for diseases.

## 2.2. Genetic Model

Recent work has suggested that SNPs in human population are not inherited independently; rather, sets of adjacent SNPs are present on alleles in a block pattern, so called **haplotype**. Many haplotype blocks in human have been transmitted through many generations without recombination. This means although a block may contain many SNPs, it takes only a few SNPs to identify or to tag each haplotype in the block. A genome-wide haplotype would comprise half of a diploid genome, including one allele from each allelic gene pair. The **genotype** is the descriptor of the genome which is the set of physical DNA molecules inherited from the organism's parents. A pair of haplotype consists of a genotype.

SNPs are bi-allelic and can be referred as 0 for majority allele and 1, otherwise. If alleles on both haplotypes are the same, then the corresponding genotype is homogeneous, and can be represented as 0 or 1. If the two alleles on the two haplotypes are different, the genotype is heterozygous, represented as 2.

In **Figure 1**, there are four chromosomes, we assume the first two chromosomes belong to one person and the other two chromosomes belong to another person. We can find on most sites the four chromosomes are identical, but on some sites they are different, nucleotides on these sites are SNP. The haplotype is the concatenation of SNPs and a genotype is composed of two haplotypes.

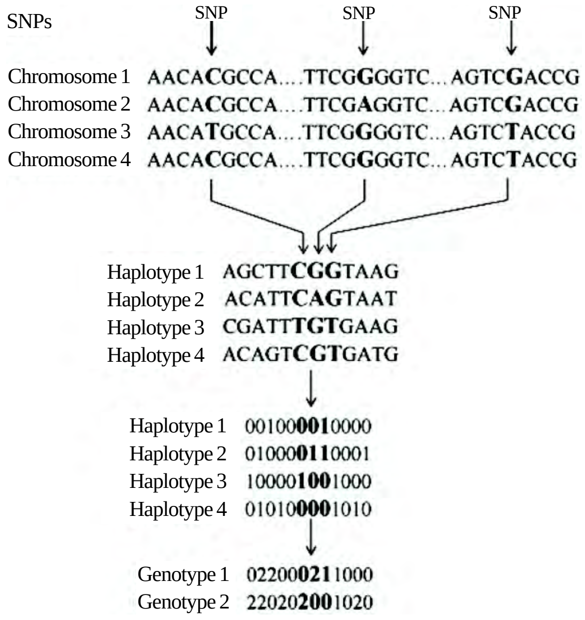


**Figure 1.** SNP, haplotype and genotype.

The case-control sample populations consist of *N* individuals who are represented in genotype with *M* SNPs. Each SNP attains one of the three values 0, 1, or 2. The sample *G* is an (0, 1, 2)-valued *N* x *M* matrix, where each row corresponds to an individual, each column corresponds to a SNP.

The sample *G* has 2 classes, case and control, and *M* variables, and each of them represents a SNP. To construct a classification tree, we split the sample *S* into 3 child sub-samples, depending on the value (0, 1, 2) of the variable (SNP) on the splitting site (loci). In fact we can construct a binary tree (split sample according to homozygous or heterozygous), but there is no way to tell the difference between major allele (1) and minor allele (0). In order to distinguish them we split the sample into 3 sub-samples instead of 2. We grow the tree to the largest possible extent. The construction of the classification tree for case-control sample is illustrated in **Figure 2**. In the first level, we split the sample (30 genotypes, 14 cases and 16 controls) into 3 sub-samples (17, 8, 5) at loci 5 (the 5th SNP). In the second level, the first sub-sample splits at loci 9 and the second sub-sample splits at loci 7. No splitting is required for the third sub-sample because it is a terminal node with only one class. In the third level, the only split node splits at loci 3. The relationship of a leaf to the tree on which it grows can be described by the hierarchy of splits of branches (starting from the trunk) leading to the last branch from which the leaf hangs. The collection of split site is a Multi-SNPs combination (*MSC*), which can be viewed as a classification tree. In this example, *MSC* = {5, 9, 7, 3}and *m* = 4, which is a collection of 4 SNPs, represented as their loci.

## 2.3. Searching for Disease Associated Multi-SNPs

To fully understand the basis of complex diseases, it

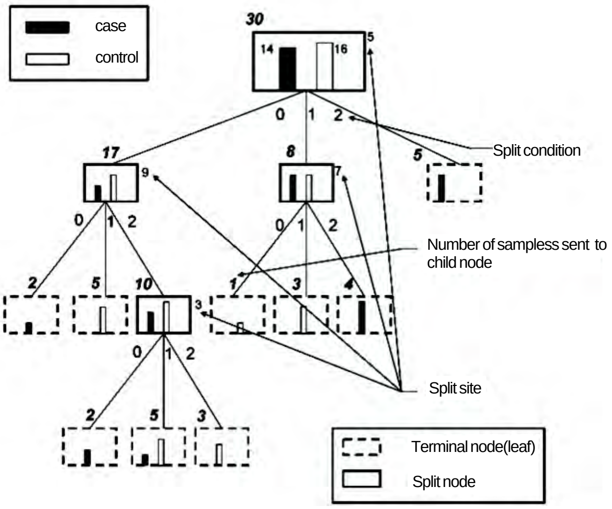


**Figure 2.** Classification tree for case-control sample.

is important to identify the critical genetic factors involved, which is a combination of multiple SNPs. For a given sample *G*, *S* is the set of all SNPs (denoted by loci) for the sample, and a multi-SNPs combination (*MSC*) is a subset of *S*. In disease associations, we need to find a *MSC* which consists of a combination of SNPs that are well associated with the disease. To find such *MSC*, we need first rank all SNPs according to their association degree (measured as weight) with diseases. Based on the sorting, we can find the *n* most disease associated SNPs for a given threshold *n*.

Although there are many statistical methods to detect the most disease associated SNPs, such as odds ratio or risk rates, the result is not satisfactory. We decide to use the random forest to find them.

### 2.4. Optimum Random Forest

We randomly generate a group of *MSCs* for each SNP. The size of the *MSC* should be much less than the size of set *S* ($m << M$). Each *MSC* can be represented as a tree and all trees make the forest *F*. All trees (or *MSCs*) of the forest $F_i$ ($i=1, 2, ..., M$) must include the $i^{th}$ SNP and the other ($m$-1) SNPs can be randomly chosen from *S* except the $i^{th}$ SNP. In this way, the *M* forests cover all SNPs in *S*.

We grow a classification tree for every *MSC* in each forest $F_i$. We run all the testing samples down these trees to get the classifier for each sample in the training set, then we can get a classification rate for each tree in $F_i$. The $MSC_i$ is the representative for the forest $F_i$ and the $MSC_i$ has the highest classification rate among all trees in $F_i$. Each member (SNP) of the $MSC_i$ is assigned a weight $w_{i,j}$ ($j \in MSC$) based on the classification rate. The weights for SNPs in the same *MSC* are the same. We can find *M MSCs* for the *M* forests. If a SNP is not a member of $MSC_i$, then $w_{i,j} = 0$.

The weight for each SNP $W_j$ (j = 1, 2, …, *M*) in *M* is the sum of weights from all *MSCs*.

$$W_j = \sum_{i=1}^{M} w_{i,j} \qquad (1)$$

In the general random forest (GRF) algorithm, the *MSC* is selected completely randomly and $m << M$. It may miss some important SNPs if they are not chosen for any *MSC*. In our optimum random forest (ORF) algorithm, this scenario is avoided because we generate at least one *MSC* for each SNP. On the other hand, in GRF, the classifier (forest) consists of trees where there is a correlation between any two trees in the forest, and the correlation will decrease the rate of the classifier. But in ORF, we generate a forest by randomly choosing *MSC* and samples for each tree and the prediction for testing samples is in this forest only, which is completely independent from the other trees. In this way, we extinguish the correlation among trees.

All SNPs are sorted according to their cumulative weights. The most disease-associated SNP is the one with the highest weight. The contribution to diseases of each SNP is quantified by its weight, but in GRF there is no way tell the difference of contribution among SNPs. The GRF can only tell the difference among classifiers (trees).

## 3. DISEASE SUSCEPTIBILITY PREDICTION

In this section we first describe the input and the output of prediction algorithms and then show how to apply the optimum random forest to the disease susceptibility prediction.

Data sets have *n* genotypes and each has *m* SNPs. The input for a prediction algorithm includes:

(G1) Training genotype set $g_i = (g_{i,j}), i = 0, 1, ..., n, j = 1, ... m, g_{i,j} \in \{0,1,2\}$

(G2) Disease status $s(g_i) \in \{0,1\}$, indicating if $g_i$, $i = 0, 1, ..., n$, is in case (*1*) or in control (*0*) , and

(G3) Testing genotype $g_t$ without any disease status.

We will refer to the parts (G1-G2) of the input as the training set and to the part (G3) as the test set. The output of prediction algorithms is the disease status of the genotype $s(g_t)$.

We use leave-one-out cross-validation to measure the quality of the algorithm. In the leave-one-out cross-validation, the disease status of each genotype in the data set is predicted while the rest of the data is regarded as the training set.

We describe several universal prediction methods below. These methods are adaptations of general computer-intelligence classifying techniques.

**Closest Genotype Neighbor (CN).** For the test genotype $g_t$, find the closest (with respect to Hamming distance) genotype $g_i$ in the training set, and set the status $s(g_t)$ equals to $s(g_i)$.

**Support Vector Machine Algorithm (SVM).** Support Vector Machine (SVM) is a generation learning system based on recent advances in statistical learning theory. SVMs deliver a state-of-the-art performance in real-world applications and have been used in case/control studies [18, 20]. There are some SVM softwares available and we decide to use libsvm-2.71 [19] with the following radial basis function:

$$exp(- \tau ※ | u-v |^{2})$$

**General Random Forest (GRF).** We use Leo Breiman and Adele Cutler's original implementation of RF version [19]. This version of RF handles unbalanced data to predict accurately. RF tries to perform a regression on the specified variables to produce the suitable model. RF uses bootstrapping to produce random trees and it has its own cross-validation technique to validate the model for prediction/classification.

**Most Reliable 2 SNP Prediction (MR2) [17].** This method chooses a pair of adjacent SNPs (site of $s_i$ and $s_{i+1}$) to predict the disease status of the test genotype $g_t$ by voting among genotypes from the training set which have the same SNP values as $g_t$ at the chosen sites $s_i$ and $s_{i+1}$. They choose the 2 adjacent SNPs with the highest prediction rate in the training set.

**LP-based Prediction Algorithm (LP).** This method assumes that certain haplotypes are susceptible to the disease while others are resistant to the disease. The genotype susceptibility is then assumed to be a sum of susceptibilities of its two haplotypes.

We want to assign a positive weight to susceptible haplotypes and a negative weight to resistant haplotypes such that for any control genotype the sum of weights of its haplotypes is negative and for any case genotype it is positive. We would also like to maximize the confidence of our weight assignment which can be measured by the absolute values of the genotype weights. In other words, we would like to maximize the sum of absolute values of weights over all genotypes.

This method is based on a graph $X = \{H, G\}$, where the vertices $H$ correspond to distinct haplotypes and the edges $G$ correspond to genotypes connecting its two haplotypes. The density of $X$ is increased by dropping SNPs which do not collapse edges with an opposite status. The linear program assigns weights to haplotypes that, for any non-diseased genotype, the sum of weights of its haplotypes is less than 0.5 and greater than 0.5 otherwise. We maximize the sum of absolute values of weights over all genotypes. The status of the testing genotype is predicted as sum of its endpoints [15].

**Optimum Random Forest (ORF).** In the training set, the optimum random forest algorithm we described above is used to sort all SNPs, and find out the *m* most disease associated SNPs for a given threshold *m*. The *m* most disease associated SNPs (Tag SNPs) are used to build the optimum random forest to test the left-out sample. In leave-one-out test, since the training set is different after leaving one sample out, we may have different Tag SNPs for different training sets. The *m* variables (SNPs) are used to grow many different classification trees by permuting the order of the splitting site (Note that the tree {3, 9, 5}is different from the tree {5, 9, 3}). We may use the *m* Tag SNPs to grow many (say, 500) trees and choose the best tree (classifier) to predict the disease status of the testing genotype. The best tree has the highest average prediction rate (over 1000 trials) in the training set. Then we run the testing genotype down the best tree to get its disease status. The Optimum Random Forest algorithm is illustrated in Figure 3.

## 4. RESULTS & DISCUSSION

In this section we first describe the genetic data of the Crohn's disease and then discuss our experimental results.

### 4.1. Data Set

The genetic data is derived from the 616 kilobase region of human Chromosome 5q31 that may contain a genetic variant responsible for Crohn's disease by genotyping 103 SNPs for 129 trios [14]. All offspring belong to the case population, while almost all parents belong to the control population. In the entire data, there are 144 case and 243 control individuals. The missing genotype data and haplotypes have been inferred using the 2SNP phasing method [21].

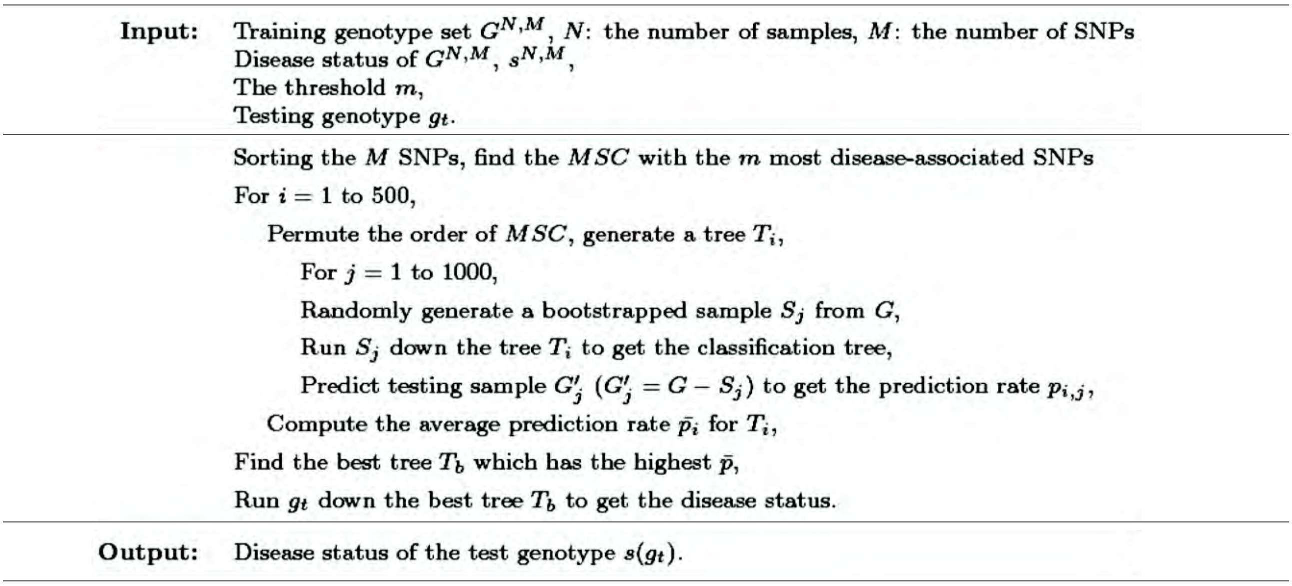
### 4.2. Measures of Prediction Quality

To measure the quality of prediction methods, we need to measure the deviation between the true disease status and the result of predicted susceptibility, which can be regarded as measurement error. We will present the basic measures used in epidemiology to quantify the accuracy of our methods.

The basic measures are:

**Sensitivity:** the proportion of persons who have the disease and who are correctly identified as cases.

**Specificity:** the proportion of people who do not



```
Input:   Training genotype set $G^{N,M}$, $N$: the number of samples, $M$: the number of SNPs
         Disease status of $G^{N,M}$, $s^{N,M}$,
         The threshold $m$,
         Testing genotype $g_t$.

         Sorting the $M$ SNPs, find the $MSC$ with the $m$ most disease-associated SNPs
         For $i = 1$ to 500,
           Permute the order of $MSC$, generate a tree $T_i$,
             For $j = 1$ to 1000,
             Randomly generate a bootstrapped sample $S_j$ from $G$,
             Run $S_j$ down the tree $T_i$ to get the classification tree,
             Predict testing sample $G'_j$ ($G'_j = G - S_j$) to get the prediction rate $p_{i,j}$,
           Compute the average prediction rate $\bar{p}_i$ for $T_i$,
         Find the best tree $T_b$ which has the highest $\bar{p}$,
         Run $g_t$ down the best tree $T_b$ to get the disease status.

Output:  Disease status of the test genotype $s(g_t)$.
```

**Figure 3.** Optimum Random Forest Algorithm.

have the disease and who are correctly classified as controls.

The definitions of these two measures of validity are illustrated in Table1.

In this table:

*a* = True positive, people with the disease who test positive

*b* = False positive, people without the disease who test positive

*c* = False negative, people with the disease who test negative

*d* = True negative, people without the disease who test negative

From Table1, we can compute Sensitivity (accuracy in classification of cases, Specificity (accuracy in classification of controls) and accuracy:

$$Sensitivity = \frac{a}{a+c} \qquad (2)$$

$$Specificity = \frac{d}{b+d} \qquad (3)$$

$$Accuracy = \frac{a+d}{a+b+c+d} \qquad (4)$$

Sensitivity is the ability to correctly detect a disease. Specificity is the ability to avoid calling normal as disease. Accuracy is the percent of the population that are correctly predicted.

### 4.3. Results and Discussion

The normalized weights of 103 SNPs are shown in Figure 4. SNPs with higher weights are more associated with the disease.

In Table 2 we compare the optimum random forest (ORF) method with the other 5 methods we described in Section 3. The best accuracy is achieved by ORF - 74.4%. From the results we can find that the ORF has the best result since we select the most disease-associated multi-SNPs to build the random forest for prediction. Because these SNPs are well associated with the disease, the random forest may produce a good classifier to reflect the association.

**Table1.** Classification contingency table.

| | | True Status | |
|---|---|---|---|
| | | + | - |
| Classified | + | a | b |
| Status | - | c | d |

**Table 2.** The comparison of the prediction rates of 6 prediction methods.

| Measures | Prediction Methods | | | | | |
|---|---|---|---|---|---|---|
| | CN | SVM | GRF | MR2 | LP | ORF |
| Sensitivity | 45.5 | 20.8 | 34.0 | 30.6 | 37.5 | 70.1 |
| Specificity | 63.3 | 88.8 | 85.2 | 85.2 | 88.5 | 76.9 |
| Accuracy | 54.6 | 63.6 | 66.1 | 65.5 | 69.5 | 74.4 |

Figure 5 shows the receiver operating characteristics (ROC) curve for 6 methods. A ROC curve represents the tradeoffs between sensitivity and specificity. The ROC curve also illustrates the advantage of ORF over all previous methods.

If the size of *MSC* is *m*, and the total number of SNPs is *M*, to get a good classifier, then *m* should be much less than *M*. The prediction rate depends on the size of *MSC*, as shown in Figure 6. In our experiment, we found that the best size of *MSC* is 19.

## 5. CONCLUSION

In this paper, we discuss the potential of applying ran-

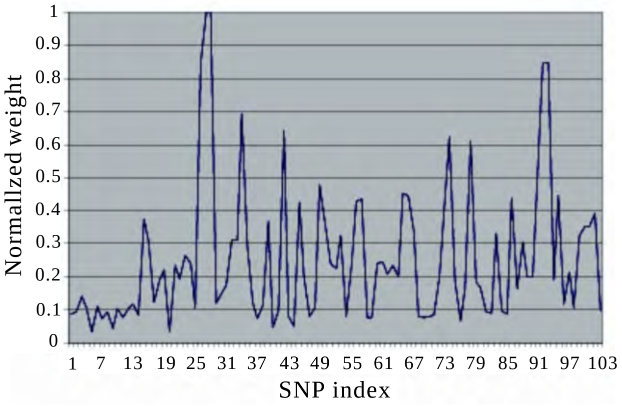


**Figure 4.** Normalized weights for 103 SNPs.

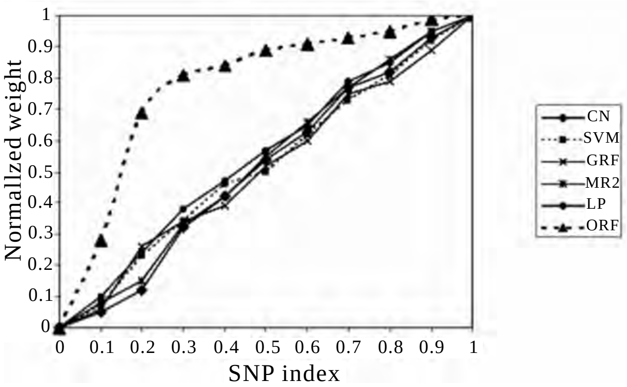


**Figure 5.** ROC curve for 6 prediction methods.

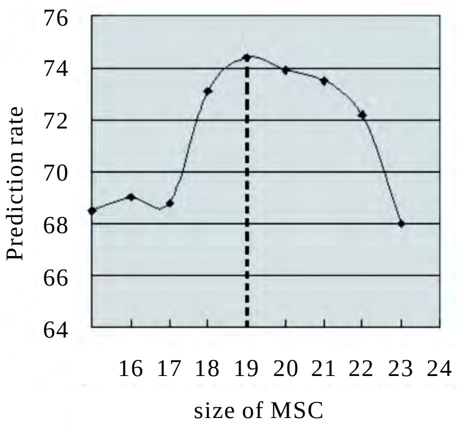


**Figure 6.** Best *MSC* size.

dom forest on disease association studies. The proposed genetic susceptibility prediction method based on the optimum random forest is shown to have a high prediction rate and the multi-SNPs being selected to build the random forest are well associated with diseases. Actually the cause of complex diseases is the combination of the environmental, genetic factors and some other factors such as infection and races. In our future work we are going to analyze the interactive contribution of these factors for the development of complex diseases. Our next project is going to find the relationship between the genetic factor and race in the development of Type 2 Diabetes. The integrated software will be available soon for public use.

## REFERENCE

[1]National Digestive Diseases InformationClearinghouse (NDDIC), http://digestive.niddk.nih.gov/ddiseases/pubs/crohns.

[2]Cardon, L.R. & Bell, J.I. Association Study Designs for Complex Diseases. *Nature Reviews: Genetics* 2001, 2:91-98.

[3]Hirschhorn, J.N. & Daly, M.J. Genome-wide Association Studies for Common Diseases and Complex Diseases. *Nature Reviews: Genetics* 2005, 6:95-108.

[4]Merikangas, KR. & Risch, N. Will the Genomics Revolution Revolutionize Psychiatry. *American Journal of Psychiatry*, 2003, 160: 625-635.

[5]Botstein, D. & Risch, N. Discovering Genotypes Underlying Human Phenotypes: Past Successes for Mendelian Disease, Future Approaches for Complex Disease. *Nature Genetics* 2003, 33: 228-237.

[6]Clark, A.G., Boerwinkle E., Hixson J. & Sing C.F. Determinants of the success of whole-genome association testing. *Genome Res.* 2005, 15:1463-1467.

[7]He, J. & Zelikovsky, A. Tag SNP Selection Based on Multivariate Linear Regression. *Proc. of International Conference on Computational Science* 2006, LNCS 3992:750-757.

[8]Brinza, D., He, J. & Zelikovsky, A. Combinatorial Search Methods for Multi-SNP Disease Association. *Proc. of International Conference of the IEEE Engineering in Medicine and Biology* 2006, pages 5802-5805.

[9]Cook N.R., Zee R.Y. & Ridker P.M. Tree and Spline Based Association Analysis of gene-gene interaction models for ischemic stroke. *Stat Med* 2004, 23(9):439-453.

[10]York T.P. & Eaves L.J. Common Disease Analysis using Multivariate Adaptive Regression Splines (MARS): Genetic Analysis Workshop 12 simulated sequence data. *Genetic Epidemiology* 2001, 21 (S1):649-654.

[11]Ritchie M.D., Hahn L.W., Roodi N., Bailey L.R., Dupont W.D., Parl F.F. & Moore J.H. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet.* 2001, 69: 138-147.

[12]Hahn L.W., Ritchie M.D. & Moore J.H. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 2003, 19:376-382.

[13]Lunetta, K., Hayward, L., Segal, J. & Van Eerdewegh, P. Screening Large-scale Association Study Data: Exploiting Interactions Using Random Forests" , *BMC Genetics* 2004, pages 5:32.

[14]Daly, M., Rioux, J., Schaffner, S., Hudson, T. & Lander, E. High resolution haplotype structure in the human genome. *Nature Genetics* 2001, 29:229-232.

[15]Mao, W., He, J., Brinza, D. & Zelikovsky, A. A Combinatorial Method for Predicting Genetic Susceptibility to Complex Diseases. *Proc. International Conference of the IEEE Engineering In Medicine and Biology Society* 2005, pages 224-227.

[16]Mao, W., Brinza, D., Hundewale, N., Gremalschi, S. & Zelikovsky, A. Genotype Susceptibility and Integrated Risk Factors for Complex Diseases. *Proc. IEEE International Conference on Granular Computing* 2006, pages 754-757.

[17]Kimmel, G. & Shamir R. A Block-Free Hidden Markov Model for Genotypes and Its Application to Disease Association. *J. of Computational Biology* 2005, 12(10): 1243-1260.

[18]Listgarten, J., Damaraju, S., Poulin B., Cook, L., Dufour, J., Driga, A., Mackey, J., Wishart, D., Greiner,R. & Zanke, B. Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms. *Clinical Cancer Research* 2004, 10:2725-2737.

[19]Breiman, L. & Cutler, A. *http://stat.berkeley.edu/breiman.*

[20]Waddell, M., Page,D., Zhan, F., Barlogie, B. & Shaughnessy, J., Predicting Cancer Susceptibility from SingleNucleotide Polymorphism Data: A Case Study in Multiple Myeloma. *Proc. of the 5th international workshop on Bioinformatics* 2005, pages 21-28.

[20]Chang, C. and Lin, C. *http://www.csie.ntu.edu.tw/libsvm.*

[21]Brinza, D. & Zelikovsky, A. 2SNP: Scalable Phasing Based on 2-SNP Haplotypes. *Bioinformatics* 2006, 22(3):371-373.

J. Biomedical Science and Engineering, 2008, 1, 59-63

Scientific Research Publishing

# Hilbert Huang transform for predicting proteins subcellular location

**Feng Shi * , Qiu-Jian Chen & Na-na Li**

School of Science, Huazhong Agricultural University, Wuhan, Hubei, China. Correspondence should be addressed to Feng Shi (shifeng@mail.hzau.edu.cn).

## ABSTRACT

**Apoptosis proteins have a central role in the development and homeostasis of an organism. These proteins are very important for understanding the mechanism of programmed cell death, and their function is related to their types. The apoptosis proteins are categorized into the following four types: (1) Cytoplasmic protein; (2) Plasma membrane-bound protein; (3) Mitochondrial inner and outer proteins; (4) Other proteins. A novel method, the Hilbert-Huang transform, is applied for predicting the type of a given apoptosis protein with support vector machine. High success rates were obtained by the re-substitute test (98/98=100%) and jackknife test (91/98 = 92.9%).**

**Keywords:Hilbert Huang transform; Support vector machine; Subcellular location predict**

## 1. INTRODUCTION

Apoptosis, or programmed cell death, is a fundamental process controlling normal tissue homeostasis by regulating a balance between cell proliferation and death [1]. This process entails the autolytic degradation of cellular components, and is characterized by blebbing of cell membranes, shrinkage of cell volumes, and condensation of nuclei [2], and is currently an area of intense investigation. Cell death and renewal are responsible for maintaining the proper turnover of cells, which ensures a constant controlled flux of fresh cells. Programmed cell death and cell proliferation are tightly coupled. When apoptosis malfunctions, a variety of formidable diseases can ensue: blocking apoptosis is associated with cancer and autoimmune disease, whereas unwanted apoptosis can possibly lead to ischemic damage or neurodegenerative disease [3]. Apoptosis is considered to have a key role in these several devastating diseases and, in principle, provides many targets for therapeutic intervention [4]. To understand the apoptosis mechanism and functions of various apoptosis proteins, it will be helpful to obtain information about their subcellular location. This is because the subcellular location of apoptosis proteins is closely related to their function [5,6]. It has been known that there are 732 archetypical proteins with “apoptosis” domains [7], and only 98 of these proteins are known to be the apoptosis protein (for more details, one can visit: http://www.apoptosis-db.org). Scientists usually deal with a number of protein sequences already known belonging to apoptosis proteins. However, it is both time-consuming and costly to determine which specific subcellular location a given apoptosis protein belongs to. Confronted with such a situation, can we develop a fast and effective way to predict the subcellular location for a given apoptosis protein based on its amino acid sequence? Recently, Guo-ping Zhou [7] attempted to identify the subcellular location of apoptosis proteins according to their sequences by means of the covariant discriminant function, which was established on the basis of the Mahalanobis distance and Chou's invariance theorem [7,8,9].The results were quite promising, indicating that the subcellular location of apoptosis proteins are predictable to a considerably accurate extent if a good vector representation of protein can be established. It is expected that, with a continuous improvement of vector representation methods by incorporating amino acid properties, and by using more powerful mathematics methods, some theory predicting method might eventually become a useful tool in this area because the function of an apoptosis protein is closely related to its subcellular location. The present study was initiated in an attempt to address this problem.

Chou and Elrod made an extensive research in predicting subcellular location mainly based on the amino acid composition. Subsequently, in order to take into account the sequence-order effects and improved the prediction quality, Chou has further incorporated the quasi-sequence order effect [5] and introduced the concept of “pseudo-amino-acid composition” [9]. For example, Chou [10] classified membrane proteins into five different types and proposed a covariant discriminant algorithm to predict the types of membrane proteins. Recently, Cai *et al.* [11] applied neural network to this problem. To improve the prediction quality, Chou [5] proposed a new method in which the covariant discriminate algorithm was augmented to incorporate the quasi-sequence-order effect. This method uses the amino acid composition and the sequence-order-coupling numbers (reflecting the sequence order effect) in order to improve the prediction quality. Feng [12] proposed a new representation of unified attribute vector, that each protein can be represented by a vector, which is 20-D vector in Hilbert space with unified length. Hence, all of proteins have their representative points on the surface of the 20-D globe. The representative points of the proteins in the same family or with the higher sequence identity are closer on the surface. The overall predictive accuracy could be improved from 3% to 5% for different databases [12] with this simply modification of the usage of the amino acid composition. Recently, a series of new powerful approaches have been developed by Chou and his co-workers [13]. Encouraged by the great successes of the previous invertigators in the area, here we would like to use a different strategy, the support vector machines, to approach this very important but also very difficult problem in the hope that our approach can play a complementary role to the existing methods.

## 2. HILBERT HUANG TRANSFORM

The HHT consists of two parts: empirical mode decomposition (EMD) and Hilbert spectral analysis (HSA). This method is potentially viable for nonlinear and nonstationary data analysis, especially for time-frequency-energy representations. It has been tested and validated exhaustively, but only empirically. In all the cases studied, the HHT gave results much sharper than those from any of the traditional analysis methods in time-frequency-energy representations. Additionally, the HHT revealed true physical meanings in many of the data examined. Powerful as it is, the method is entirely empirical. In order to make the method more robust and rigorous, many outstanding mathematical problems related to the HHT method need to be resolved. In this section, a brief introduction to the methodology of the HHT will be given. Readers interested in the complete details should consult [14].

### 2.1. The empirical mode decomposition method (the sifting process)

In this method any time series, including non-linear and non-stationary series, can be decomposed into a finite number of intrinsic mode functions (IMFs) through empirical mode decomposition (EMD) process. An IMF is a function which must follow two conditions: (1) the difference between the numbers of extrema and zero-crossings is of ⩽1 ; and (2) the mean of the upper envelop (linked by local maxima) and the lower envelop (linked by local minima) are zero at every point.

The EMD process is as follows. According to Hilbert-Huang transform(HHT)[14], once the extrema of a time series $x(t)$ are identified, all the local maxima and minima are connected by two special lines as the upper and lower envelopes respectively. Their mean is designated as $m_1$, and the difference between $x(t)$ and $m_1$ is $x(t)$-$m_1$=$h_1$ . If $h_1$ is not an IMF, $h_1$ is treated as the data and undergoes the procedure above, then $h_1$-$m_{11}$=$h_{11}$ . Repeat this sifting procedure $k$ times until $h_{1k}$ is an IMF, that is $h_{1(k-1)}$-$m_{1k}$=$h_{1k}$, thus the first IMF component is obtained, i.e. . Then separate $IMF_1$ from the original time series by $x(t)$- $IMF_1$=$r_1$. Treat $r_1$ as the new data and subject it to the same sifting process above. Repeat this procedure on all the subsequent $r_j$ , i.e. $r_1$-$IMF_2$=$r_2$ , $r_2$-$IMF_3$=$r_3$,⋯, $r_{n-1}$-$IMF_n$=$r_n$ .

So the result is:

$$x(t)=\sum_{j=1}^{n} IMF_j(t)+r_n(t)$$

### 2.2. Hilbert transform

Having obtained the intrinsic mode function components $IMF_i$ (denoted as $c_i$), one will have no difficulty in applying the Hilbert transform to each IMF component,

$$H(c_i(t))=\frac{1}{\pi}PV\int_{-\infty}^{\infty}\frac{c_i(t)}{t-\tau}d\tau$$

in which the PV indicates the principal value of the singular integral. With the hilbert transform, the analytic signal is defined as

$$A(c_i(t))=c_i(t)+jH(c_i(t))=a_i(t)e^{j\theta_i(t)}$$

Here, $a_i(t)$ is the instantaneous amplitude, and $\theta_i(t)$ is the phase function,

$$a_i(t)=\sqrt{c_i^2(t)+H^2(c_i(t))}$$

$$\theta_i(t)=\arctan\frac{H(c_i(t))}{c_i(t)}$$

and the instantaneous frequency is simply

$$\varpi_i(t)=\frac{d\theta_i(t)}{dt}$$

With the Hilbert Spectrum defined, we can also define the marginal spectrum $h(\varpi)$ as

$$h(\varpi)=\int_0^T H(\varpi,t)dt$$

The marginal spectrum offers a measure of the total amplitude (or energy) contribution from each

**Table 1.** Comparative summary of Fourier, Wavelet and HHT analyses.

| | Fourier | Wavelet | Hilbert |
|---|---|---|---|
| Basis | A priori | a priori | adaptive |
| Frequency | Convolution: global Uncertainty | convolution: regional uncertainty | differentiation local, certainty |
| Presentation | energy -frequency | energy-time-frequency | energy-time-frequency |
| Nonlinear | no | no | yes |
| Nonstationary | no | yes | yes |
| Feature Extraction | no | discrete: no continuous: yes | yes |
| Theoretical base | theory complete | theory complete | empirical |

frequency value. This spectrum represents the accumulated amplitude over the entire data span in a probabilistic sense.

The combination of the empirical mode decomposition and the Hilbert spectral analysis is also known as the "Hilbert-Huang transform" (HHT) for short. Empirically, all tests indicate that HHT is a superior tool for time-frequency analysis of nonlinear and nonstationary data. It is based on an adaptive basis, and the frequency is defined through the Hilbert transform. Consequently, there is no need for the spurious harmonics to represent nonlinear waveform deformations as in any of the priori basis methods, and there is no uncertainty principle limitation on time or frequency resolution from the convolution pairs based also on a priori basis.

A comparative summary of Fourier, wavelet and HHT analyses is given in the **Table1**:

This table shows that the HHT is indeed a powerful method for analyzing data from nonlinear and nonstationary processes: it is based on an adaptive basis; the frequency is derived by differentiation rather than convolution; therefore, it is not limited by the uncertainty principle; it is applicable to nonlinear and nonstationary data and presents the results in time-frequency-energy space for feature extraction.

Support Vector Machine (SVM) is one type of learning machines based on statistical learning theory. A complete description to the theory of SVMs for pattern recognition is in Vapnik's book.[15]. SVMs have been used in a range of bioinformatics problems including protein fold recognition [16]; proteinprote in interactions prediction [17]; prediction of protein subcellular location [17, 18], protein secondary structure prediction,T-cell epitopes prediction, Classification of protein quaternary structure [19].

In this paper, we apply Vapnik's support vector machine for predicting the types of apoptosis proteins. We have used the OSU_SVM, a Matlab SVM toolbox (http://www.ece.osu.edu/~maj/osu_svm), which is an implementation of SVM for the problem of pattern recognition.

## 3. TRAINING AND PREDICTION

According to their subcellular location [12], apoptosis proteins are classified into the following four types: (1) type I: Cytoplasmic protein; (2) type II: Plasma membrane-bound protein; (3) type Ⅲ: Mitochondrial inner and outer proteins; (4) type Ⅳ: Other proteins (see **Table 2**).

In this research, we first translate every aminoacid sequence *s* into a numerical sequence *f* by hydrophobicity index, then, decompose it into a finite number of intrinsic mode functions (IMFs) through empirical mode decomposition (EMD) process, we just select the 2nd to 4th components (IMF2, IMF3, IMF4), because first IMF just reflects the rand composition and the last is just the trendences composition of the numerical sequence *f*. Then applying the Hilbert

**Table 2.** List of the acession numbers for the 98 apoptosis proteins classified into four categories according to their subcellular locations. (Type I: 43 Cytoplasmic proteins; Type II: 30 Plasma membrane-bound proteins; Type III: Mitochondrial inner and outer proteins ; Type IV: 12 Other proteins).

| Type | Type I | Type II | Type III | Type IV |
|---|---|---|---|---|
| proteins | NP_033941, NP_033940, NP_033939, NP_031637, NP_031570, NP_031563, NP_031490, NP_033447, , NP_036246, NP_001218, NP_004041, NP_065209, NP_001151, NP_071610, NP_071567, NP_066961, NP_037054, NP_036894, NP_005649, NP_004392, NP_004315, NP_001187, NP_001159, NP_001157, NP_001156, P55212, P42574, P39429, P55867, P22366, P55866, P55214, P55269, P29466, P55865, P29452, Q02357, O54786, Q60989, Q62210, Q60431, O70201, XP_013050, | NP_037223, NP_037275, NP_032013, NP_032612, NP_037315, NP_005916, NP_005579, NP_000034, NP_001056, NP_003781, NP_002498, NP_036742, NP_031553, NP_031549, P50555, P25118, P18519, P51867, O19131, Q63199, O77736, , O02703, Q13014, Q63690, Q07820, Q91828, Q91827, Q07812, P28825, NP_001179 | P10417, P53563, Q07816, P49950, Q07817, O95831, Q9OX1, Q9JM53, Q9VQ79, O77737, Q00709, XP_008738, NP_033873, | Q63369, Q90660, Q00653, Q04861, P19838, NP_032715, P98150, Q15121, Q62048, NP_033872, NP_004040, NP_005736 |

a.Derived from SWISS-PROT data bank.
b.Of the 12 other apoptosis proteins, five are located in nucleus, two in endoplasmic reticulum, one in microtubule, and one in lysosome [7].

transform to each IMF component, we get the instantaneous amplitude $a_i(t)$, then get the energy value $e_i = \sum_t a_i^2(t)$ , (t=2, 3, 4). Next, get its energy ratio $g_i = \frac{e_i}{e_2 + e_3 + e_4}$ .Last every protein was represented as a point or a vector in a 23-D space. The first 20 components of its vector were supposed to be the occurrence frequencies of the 20 amino acids in the protein concerned, the last three components were its energy ratio times a weight, there, we set the weight is 0.2.

The computations were carried out on a PC. Also for the SVM, the width of the Gaussian RBFs is selected as that which minimized an estimate of the VC-dimension. After being trained, the hyper-plane output by the SVM was obtained. The SVM method is applied to two-class problems. In this paper, for the four-class problems, we have used a simple and effective method: "one-against-others" method [16] to transfer it into two-class problems. We first test the selfconsistency and leave-one-out cross-validation (jackknife test) of the method, followed by testing the method by prediction of an independent dataset. As a result, the rates of self-consistency, cross-validation of prediction were quite high.

In addition to the prediction algorithm, we also need to construct a training data set to complete the establishment of a statistical prediction method. To realize this, based on the SWISS-PROT data bank, 98 apoptosis proteins (the date were taken from Zhou [7]) were classified into the following four subcellular locations: (1) cytoplasmic, (2) plasma membrane-bound, (3) mitochondrial, and (4) other (**Table 1**).

## 4. RESULTS AND DISCUSSION

By means of the SVM algorithm described in the last section, a statistical prediction was performed for the 98 apoptosis proteins listed in **Table 2**. The prediction was conducted by two different approaches, the re-substitution test and the jackknife test. The results are given in **Table 3**.

### 4.1. Re-substitution test

The so-called re-substitution test is an examination for the self-consistency of a prediction method[7]. When the re-substitution test was performed for the current study, the type of each apoptosis protein in a data set was in turn identified using the rule parameters derived from the same data set, the so-called training data set. As shown in **Table 3**, the overall success rate thus obtained for the 98 apoptosis proteins in **Table 1** was 100%, indicating an excellent self-consistency.

However, during the process of the re-substitution test, the rule parameters derived from the training data set include the information of the query protein later plugged back in the test. This will certainly underestimate the error and enhance the success rate because the same proteins are used to derive the rule parameters and to test themselves. Nevertheless, the re-substitution test is absolutely necessary because it reflects the self-consistency of a prediction method, especially for its algorithm part. A prediction algorithm certainly cannot be deemed as a good one if its self-consistency is poor. In other words, the re-substitution test is necessary but not sufficient for evaluating a prediction method. As a complement, a cross-validation test for an independent testing data set is needed because it can reflect the effectiveness of a prediction method in practical application. This is important especially for checking the validity of a training data set-whether it contains sufficient information to reflect all the important features concerned so as to field a high success rate in application.

### 4.2. Jackknife test

As is well known, the independent data set test, subsampling test, and jackknife test are the three methods often used for cross-validation in statistical prediction. Among these three, however, the jackknife test is deemed as the most effective and objective one for a comprehensive discussion about this). During jackknifing, each protein in the data set is in turn singled out as a tested protein and all the rule parameters are calculated based on the remaining proteins. In other words, the subcellular location of each apoptosis protein is identified by the rule parameters derived using all the other apoptosis proteins except the one that is being identified. During the process of

**Table 3.** Tested results for the 98 apoptosis prtoeins in Table 2 by both Re-substitution test and Jackknife test. All use Gauss RBF kernel function, while the value C =15, and the gama= 80.

| Test method | | Success Rate | | | | |
|---|---|---|---|---|---|---|
| | | **Type** Ⅰ | **Type** Ⅱ | **Type** Ⅲ | **Type** Ⅳ | **Overall** |
| Re-substitute | covariant | 43/43=100% | 30/30=100% | 9/13=60.2% | 7/12=58.3% | 89/98=90.8% |
| | SVM | 42/43=97.70% | 30/30=100% | 13/13=100% | 12/12=100% | 97/98=99.0% |
| | HHT | 43/43=100% | 30/30=100% | 13/13=100% | 12/12=100% | 98/98=100% |
| Jack-knife | covariant | 42/43=97.7% | 22/30=73.3% | 4/13=30.8% | 3/12=25.0% | 71/98=72.5% |
| | SVM | 39/43=91.4% | 28/30=93.3% | 12/13=92.5% | 9/12=75.0% | 88/98=89.8% |
| | HHT | 41/43=95.3% | 29/30=96.7% | 12/13=96.7% | 9/12=75.7 | 91/98=92.9% |

jackknifing, both the training data set and testing data set are actually open, and a protein will in turn move from one to the other. As expected, the success prediction rates by jackknife test were decreased in comparison with those by the re-substitution test. Such a decrement is particularly more remarkable for small subsets. This is because the cluster-tolerant capacity for small subsets is usually low. And hence the information loss resulting from jackknifing will have a greater impact on the small subsets than the large ones. Nevertheless, as shown in **Table 2**, the overall jackknife rate for the data set of the 98 apoptosis proteins could still reach 93%. It is expected that the success rate for identifying the subcellular location of apoptosis proteins can be further enhanced by improving the training data of small subsets by adding into them more new proteins that have been found belonging to the subcellular location defined by these subsets.

## 5. CONCLUSIONS

The above results, together with those obtained by the covariant discriminant prediction algorithm [7], have indicated that the types of apoptosis proteins are predictable with a considerable accuracy. It is anticipated that the HHT, and the SVM, if effectively complemented with each other, will become a powerful tool for predicting the types of apoptosis proteins. The current study has further demonstrated that the datasets originally constructed by Zhou[7] will be very useful for the area of apoptosis study. It is expected that the prediction quality can be further improved if the current HHT can be properly combined with pseudoamino acid composition[9] and function domine composition and with other amino acid properties.

## ACKNOWLEDGMENTS

The authors thank Dr. Guo-Ping Zhou for providing the amino acid sequences of the apoptosis proteins and some helpful discussions. The work was partly supported by Huazhong Agricultural University, P. R. China

## REFERENCE

[1]Zhou, P., Chou, J. J., Olea RS, Yuan, J. & G. Wagner. Solution structure of Apaf-1 CARD and its interaction with caspase-9 CARD: a structural basis for specific adaptor/caspase interaction. Proc Natl Acad Sci USA 1999, 96:11265-11270.

[2]Kerr J.F., Wyllie A. H. & A. R. Currie. Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. Br J Cancer 1972, 26:239-257.

[3]Schulz J. B., Weller M. & M. A. Moskowitz. Caspases as treatment targets in stroke and neurodegenerative diseases. Ann Neurol 1999, 45:421-429.

[4]Barinaga M. Stroke-damaged neurons may commit cellular suicide. Science 1998, 281:1302-1303.

[5]Chou, K. C. A new branch of proteomics: prediction of protein cellular attributes. Gene Cloning and Expression Technologies 2002, 4:57-70,

[6]Huang, J. & Shi, F. Support vector machines for predicting apoptosis proteins types. Acta bioinformatics 2005, 53:39-47.

[7]Zhou, G. P. & Doctor. K. Subcelluar location of Apoptosis proteins. Proteins:Structure, Function, and Genetic 2003, 50:44-48.

[8]Chou, K C. A. novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins:Structure, Function and Genetics 1995, 21:319-344.

[9]Chou, K. C. Prediction of protein cellular attributes using pseudo-amino-acid-composition. Proteins :Structure, Function, and Genetics 2001, 43:246-255 (Erratum:ibid., 2001, vol. 44, 60).

[10]Chou, J. J., Li, H., Salvesen G.S., Yuan, J. & G. Wagner. Solution structure of BID, an intracellular amplifier of apoptotic signaling. Cell 1999, 96:615-624.

[11]Cai, Y. D., Liu, X. J. & Chou, K. C. Artificial neural network model for predicting membrane protein types. J. Biomol. Struct. Dyn. 2001, 18:607-610.

[12]Feng, Z. P. Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers 2001, 58:491-499.

[13]Cai, Y. D. & Chou, C. Nearest neighbour algorithm for predicting protein subcellular location by combing functional domain composition and pseudo-amino acid composition. Biochem Biophys Res Comm. 2003, 305:407-411.

[14]Huang, N. E., Shen, Z., Long, S. R., Wu, M. L., H.H. Shih, Zheng, Q., N.C. Yen, C.C. Tung & Liu, H. H. The empirical mode decomposition and Hilbert spectrum for nonlinear and nonstationary time series analysis. Proc. Roy. Soc. London A 1998, 454:903-995.

[15]Vapnik V. Statistical Learning Theory. WileyInterscience 1998.

[16]Ding, C. H. & I. Dubchak. Multiclass protein fold recognition using support vector machines and neural networks. Bioinformatics 2001, 17:349-358.

[17]Cai, Y. D., Liu, X. J., Xu, X. B.& Chou, K. C. Support vector machines for prediction of protein subcellular location by incorporating quasisequenceorder effect. J. Cell. Biochem. 2002, 84:343-348.

[18]Hua, S. J. & Sun, Z. R. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 2001, 17:721-728.

[19]Hua, S. J. & Sun, Z. R. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J. Mol. Biol. 2001, 308:397-407.

J. Biomedical Science and Engineering, 2008, 1, 64-67

Scientific
Research
Publishing

# Pattern recognition of motor imagery EEG using wavelet transform

**Bao-Guo Xu & Ai-Guo Song**

School of Instrument Science and Engineering Southeast University, Nanjing 210096, China.

## ABSTRACT

**Brain-computer interface (BCI) provides new communication and control channels that do not depend on the brain's normal output of peripheral nerves and muscles. In this paper, we report on results of developing a single trial online motor imagery feature extraction method for BCI. The wavelet coefficients and autoregressive parameter model was used to extract the features from the motor imagery EEG and the linear discriminant analysis based on mahalanobis distance was utilized to classify the pattern of left and right hand movement imagery. The performance was tested by the Graz dataset for BCI competition 2003 and the satisfactory results are obtained with an error rate as low as 10.0%.**

**Keywords**: **Brain-computer interface (BCI); Motor imagery; Wavelet coefficients; Autoregressive model**

## 1. INTRODUCTION

Left and right hand movement imagery can modify the neuronal activity in the primary sensorimotor areas, leading to the changes of the mu rhythm and beta rhythm. BCI requires effective online processing method to classify these EEG signals in order to construct a system enabling severely physically disabled patients to communication with their surroundings [1-4].

This paper presents a novel effective method for feature extraction of motor imaginary. We combine the discrete wavelet transform (DWT) with autoregressive model (AR) to extract more useful information for non-stationary EEG signals. Applying this method to analyze the Graz dataset for BCI competition 2003, we achieved the classification accuracy of 90.0%.

## 2. METHODOLOGY

### 2.1. Experimental paradigm

The data set was provided by department of medical informatics, institute for biomedical engineering, university of technology Graz [5]. It was recorded from a normal subject (female, 25y) during a feedback session. The subject sat in a relaxing chair with armrests. The task was to control a feedback bar by means of imagery left or right hand movements. The order of left and right cues was random.

**Figure 1** shows the timing of the experiment. The first 2s was quite; at t=2s an acoustic stimulus indicated the beginning of the trial; the trigger channel (#4) went from low to high, and a cross "+" was displayed for 1s; then at t=3s, an arrow (left or right) was displayed as cue. At the same time the subject was asked to move a bar into the direction of the cue. The feedback was based on AAR parameters of channel #1 (C3) and #3 (C4), the AAR parameters were combined with a discriminant analysis into one output parameter.

The recording was made using a G.tec amplifier and a Ag/AgCl electrodes. Three bipolar EEG chan-

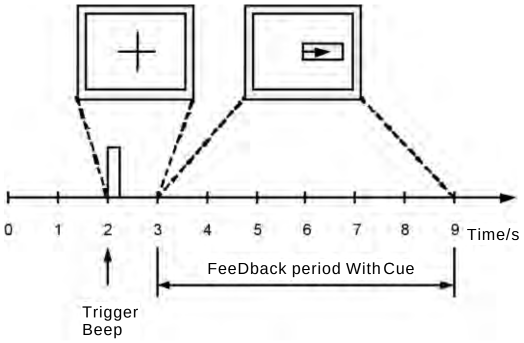


**Figure1.** Timing scheme.

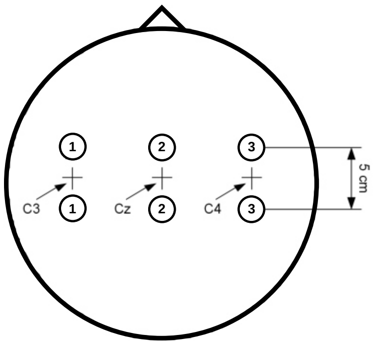


**Figure 2.** Electrode positions.

Published Online May 2008 in SciRes. http://www.srpublishing.org/journal/jbise

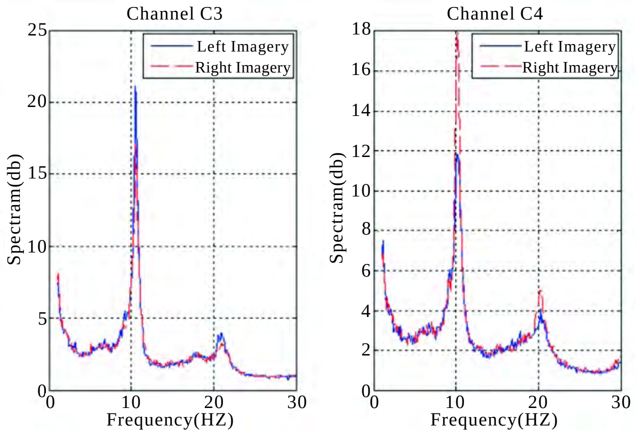


Figure 3. Average power spectrums on channel C3 and C4.

nels (anterior '+', posterior '-') were measured over C3, Cz and C4 [Figure 2]. The EEG was sampled with 128Hz, it was filtered between 0.5 and 30Hz. Similar experiments are described in [6].

The experiment consists of 7 runs with 40 trials each. All runs were conducted on the same day with several minutes break durrng experiment. One half of the datasets are provided for training; others are for evaluating the performance of the system.

## 2.2. Feature consideration

Central brain oscillations in the mu rhythm in the range of 7-12Hz and beta above 13Hz bands are strongly related to sensorimotor tasks. Sensory stimulation, motor behavior, mental imagery can change the functional connectivity cortex which results in an amplitude suppression or in an amplitude enhancement .This phenomenon was also called event-related desynchronization (ERD) and event-related synchronization (ERS) [7, 8]. Left and right hand movement imagery is typically accompanied with ERD in the mu and beta rhythms and has the characteristic of contralateral dominance.

The power spectrums on C3 and C4 of the training set are shown in Figure 3. It indicates that the power spectrums mainly distribute in the range of 8-13Hz and 19-24Hz.In addition, the power of mu and beta rhythms evoked by right hand movement imagery is lower than that of left hand movement imagery for channel C3, and it is contrary for channel C4 which is consistent with the principle of contralateral dominance. This led us to use wavelet decomposition to extract the differences between the two motor imagery tasks.

## 2.3. Procedure

The flow chart of processing single-trial motor imagery EEG is shown as in Figure 4. First, the time window was used to filter the data in temporal domain in order to get the segment that contained the most obvious difference between the two motor imagery tasks. Then EEG signals were decomposed into the frequency sub-bands using DWT and a set for statistical features was extracted from the sub-bands to represent the distribution of wavelet coefficients according to the characteristics of motor imagery EEG signals. Also the sixth-order AR coefficients of segmentation EEG signals were estimated using Burg's algorithm. Next, the combination features of wavelet coefficients and the AR coefficients were used as an input vector. Finally linear discriminant analysis (LDA) based on mahalanobis distance was utilized to classify computed features into different categories that represent the left or right hand movement imagery.

## 2.4. Feature extraction using discrete wavelet transforms

Classic Fourier transform has succeeded in stationary signals processing. However, EEG signal contains non-stationary or transitory characteristics. Thus it is not suitable to directly apply Fourier transform to such signals. The wavelet transform decomposes a signal into a set of functions obtained by shifting and dilating one single function called mother wavelet [10, 11]. Continuous wavelet transform is given by

$$W_f(a,\tau) = \frac{1}{\sqrt{a}} \int_R f(t)\Psi^*\left(\frac{t-\tau}{a}\right)dt \qquad (1)$$

Where $\Psi(t)$ is the mother wavelet, $a$ is the scale parameter and $\tau$ is the shift parameter. In principle the CWT produced an infinite number of coefficients, thus it provides a redundant representation of the signal.

The DWT provides a highly efficient wavelet representation that can be implemented with a simple recursive filter scheme and the original signal reconstruction can be obtained by an inverse filter. The pro-

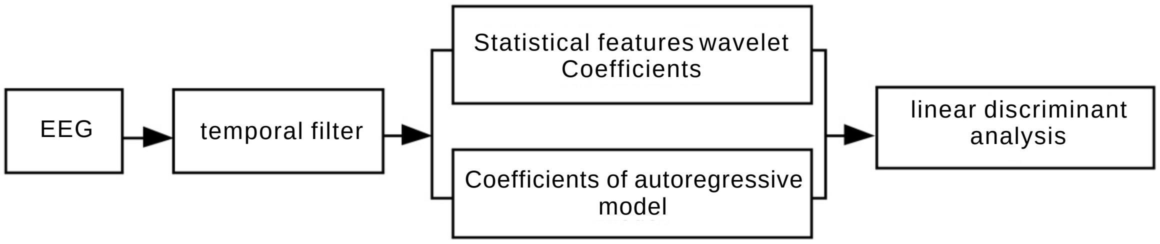

EEG
temporal filter
Statistical features wavelet Coefficients
Coefficients of autoregressive model
linear discriminant analysis

Figure 4. Flow chart of the data processing.

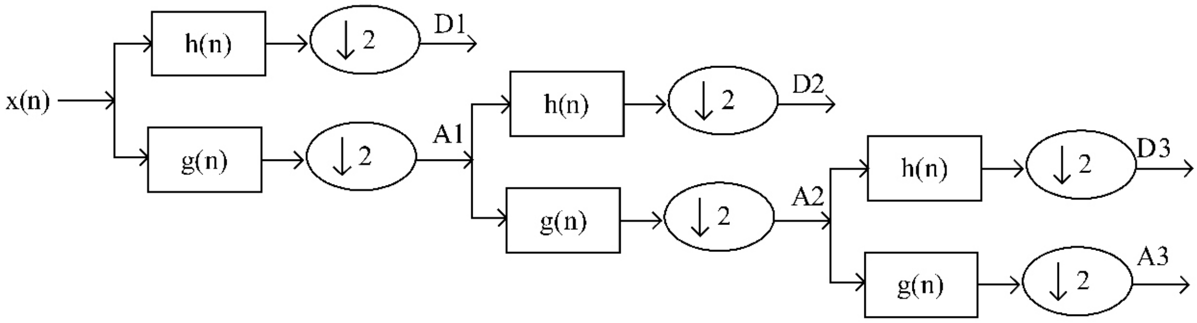


**Figure 5.** Decomposition of DWT; h[n] is the high-pass filter; g[n] is the low-pass filter.

cedure of multi-resolution decomposition of a signal x[n] is schematically shown in **Figure 5.**

The number of levels of decomposition is chosen on the basis of the dominant frequency components of the signal. According to the motor imagery EEG signals itself, we chose the level of 4 and the wavelet of Daubechies order 10.As a result, the EEG signal is decomposed into the details D1-D3 and approximation A3. The ranges of different frequency band are shown in **Table 1**.

The extracted wavelet coefficients show the distribution of the motor imagery signal in time and frequency. It can be seen from the table that the component D3 decomposition is within the mu rhythm, D2 is within the beta rhythm. Statistics over the set of wavelet coefficients were computed so as to reduce the total dimension of the feature vectors. The statistical features of each sub-band are as follows:

(1) Mean of the absolute values of the coefficients.

(2) Standard deviation of the coefficients.

(3) Average power of the wavelet coefficients.

These features represent the frequency distribution and the amount of changes in frequency distribution. Thus 12 statistical features of wavelet coefficients are obtained for two channels.

## 2.5. Feature extraction using autoregressive model

EEG signal can be considered as the output of a linear filter driven by a white noise. This filter, referred to as AR, is a linear combination of the previous output itself. A zero-mean, stationary autoregressive process of order *p* is given by

$$x(n) = -\sum_{i=1}^{p} a_p(i)x(n-i) + \varepsilon(n) \qquad (2)$$

Where *p* is the model order, *x(n)* is the signal at the sampled point *n*, $a_p(i)$ *is* the AR coefficients and ε (*n*) is a zero-mean white noise. In application, the values of the $a_p(i)$ have to be estimated from the finite samples of data $x(1),x(2),x(3),\ldots,x(N)$.

The first important things involved in using AR model is determining the optimal AR model order since too low a model order tends to smooth the spectrum and too high tends to introduce spurious peaks. Here order six was used based on the suggestions [9].

Then the Burg's method was used to estimate the AR coefficients. This method is more accurate and yields better resolution without the problem of spectral 'leakage' as compared to other methods such as Levison-Durbin as it uses the data points directly. In addition, the Burg's method can minimize both forward and backward error.

Next the AR coefficients were computed and we got six coefficients for each channel, giving a total of 12 AR coefficients features for each EEG segment for a motor imagery task.

## 2.6. Linear discriminant analysis (LDA)

LDA is one of the most effective linear classification methods for brain-computer interface, and it requires fewer examples for obtaining a reliable classifier output [12].

As to the LDA method, assume that each data element $s_i$ has *m* features. Then, an element $s_i$ is one point in a dimensional feature space. The number of examples is *n* , each example is assigned to one of two classes *C*={0,1}; Then, *S* is a matrix of size *n*×*m*, and *C* is a vector of size *n*.$N_0$. And $N_1$ are the number of elements for class 0 and 1, respectively.

The mean $\mu_c$ of each class *c* is the mean over all $s_i$ with *i* being all elements with in class *c* . The total mean of the data is

$$\mu_c = \left(\frac{N_0\mu_0 + N_1\mu_1}{N_0 + N_1}\right) \qquad (3)$$

**Table 1.** Frequencies correspond to different levels of deposition for daubechies order 10 wavelet with a sample rate 128HZ.

| Decomposed signal | Frequency range (Hz) | Level |
|---|---|---|
| D1 | 32-64 | 1 |
| D2 | 16-32 | 2 |
| D3 | 8-16 | 3 |
| A3 | 0-8 | 3 |

**Table 2 .** Dirrerent wavelet used for extracting features.

| Wavelet | Recognition rate |
|---|---|
| Daubechies order 10 | 90% |
| Discrete Meyer | 90% |
| Coiflets order 5 | 89.29% |
| Rbio1.3 | 87.86% |

The covariance matrix *C* of the data is the expectation value for

$$C=E<(s-\mu)^T(s-\mu)> \qquad (4)$$

Then, the weight vector *w* and the offset $w_0$ are

$$w=C^{-1}(\mu_1-\mu_0)^T \qquad (5)$$

$$w_0=-\mu w \qquad (6)$$

The weight vector *w* determines a separating hyperplane in the *m* -dimensional feature space. The normal distance *D*(*x*) of any element *x* is

$$D(x)=xw+w_0=(x-\mu)C^{-1}(\mu_1-\mu_0)^T \quad (7)$$

If *D*(*x*) is larger than 0, *x* is assigned to class 1, while if *D*(*x*) is smaller than 0, *x* is assigned to class 0. However, *D*(*x*)=0 indicates that all elements *x* are part of the separating hyperplane.

## 3. EXPERIMENT RESULTS

Here, we have had 6 statistical wavelet coefficients and 6 AR coefficients for each channel, giving a total of 24 features for a motor imagery task. These parameters were selected as inputs of LDA classifier. Table 2 compared the classification performances among four different wavelets. The results show the Daubechies order 10 gave the best performance and the recognition rate is as high as 90.0%. Also the results indicate that method of combining DWT with AR model are capable of extracting more useful information from the simultaneously acquired motor imagery EEG. Furthermore, when the window of 384 samples with a shift of 1 sample was used, maximum classification accuracy of 92.1% is achieved.

## 4. CONCLUSION AND FUTURE WORK

In this paper, a novel single-trial motor imagery EEG classification method is proposed. The pattern classification techniques as described in this work make possible the development of a fully automated motor imagery EEG signals analysis system which is accurate, simple and reliable enough to use in brain-computer interface. Future work will utilize the algorithms developed in this study to directly control the embedded rehabilitation robot so as to help the patient with severed paralysis to solve the problem of environment control and provide a new communication and channel to outside world.

## ACKNOWLEDGEMENTS

The work was founded by the National Basic Research Program of China (973 Program) (No.2002CB321/02), Natural Science Foundation of China (No.60475034,No.60643007) and 863 High-Tech project (No.2006AA04Z246).

## REFERENCE

[1]J. Virts. The Third International Meeting on Brain-Computer Interface Technology: Making a Difference. *IEEE Trans .Neural. Syst. Rehabil. Eng.* 2006, 14:126-127.

[2]T. M. Vaughan. Brain-computer Interface Technology: A Review of the Second International Meeting. *IEEE Trans .Neural. Syst. Rehabil. Eng.* 2003, 11:94-109.

[3]J. R. Wolpaw, N. Birbaumer, and W. Heetderks, *et al.* Brain-computer Interface Technology: A Review of the First International Meeting. *IEEE Trans. Rehabil. Eng.* 2000, 8:164-173 .

[4]J. R. Wolpaw, N. Birbaumer & D. J. McFarland, *et al.* Brain-computer interface for communication and control . *Clinical Neurophysiology* 2002 , 113:767-791.

[5]B. Blankertz, K. R. Muller & G. Curio, *et al.* BCI Competition 2003-Progress and Perspectives in Detection and Discrimination of EEG Single Trials. *IEEE Trans. Rehabil. Eng.* 2004, 51:1044-1051.

[6]A. Schlögl, C. Neuper & G. Pfurtscheller. Estimating the mutual information of an EEG-based Brain-Computer-Interface. *Biomedizinische. Technik.* 2002, 47:3-8.

[7]E. Houdayer, E. Labyt & J. Cassim, *et al.* Relationship between event-related beta synchronization and afferent inputs: analysis of finger movement and peripheral nerve stimulations. *Clinical Neurophysiology* 2006, 117:628-636.

[8]G. Pfurstcheller & F. H. Lopes da Silva. Event-related EEG/MEG synchronization and desynchronizaiton: basic principles. *Clinical Neurophysiology* 1999, 110:1842-1857.

[9]G. Pfurstcheller & C. Neuper. Motor imagery and Direct Brain-Computer Communication. *Proc. IEEE* 2001, 89:1123-1134.

[10]A. Subasi. EEG signal classification using wavelet feature extraction and a mixture of expert model. *Expert System with Application*, in press.

[11]A. Subasi. Automatic recognition of alertness level from EEG by using neural network and wavelet coefficients. *Expert System with Application* 2005, 28:701-711.

[12]A. Schlogl. A new linear classification method for an EEG-based brain-computer interface. unpublished.

J. Biomedical Science and Engineering, 2008, 1, 68-74

Scientific Research Publishing

# Design and control of a novel hydraulically/pneumatically actuated robotic system for MRI-guided neurosurgery

[1]Cyrus Raoufi, [2]Andrew A. Goldenberg & [3]Walter Kucharczyk

[1]Department of Applied Technology, California State University, Humboldt. [2]Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Canada. [3]Department of Medical Imaging, University of Toronto, Toronto, Canada. Correspondence should be addressed to Cyrus Raoufi (raoufi@humboldt.edu), Andrew A. Goldenberg (golden@mie.utoronto.ca), and Walter Kucharczyk (w.kucharczyk@utoronto.ca).

## ABSTRACT

**In this paper the design of a novel modular hydraulic/pneumatic actuated tele-robotic system and a new infrastructure for MRI-guided intervention for closed-bore MRI-guided neurosurgery are presented. Candidate neurosurgical procedures enabled by this system would include thermal ablation, radiofrequency ablation, deep brain stimulators, and targeted drug delivery. The major focus is the application of the designed MR-compatible robotic system to MRI-guided brain biopsy. Navigation and operating modules were designed to undertake the alignment and advancement of the surgical needle respectively. The mechanical design and control paradigm are reported.**

**Keywords: MR-compatible robot; Tele-surgery; Tele-robotics; Medical robot**

## 1. INTRODUCTION

The common requirement for most neurosurgical procedures is to manipulate a surgical tool relative to an anatomic target. This includes aligning, orienting, and advancing the tool to a specific anatomic target in the brain. The advantages of robotic-based neurosurgical procedures are well recognized in the clinical and technical community due to both the locating accuracy and the tele-surgery potential of the robotic systems. A neurosurgical procedure is a highly interactive process and the goal of neurosurgical robotic system is to provide the neurosurgeon with a reliable tool that augments his or her ability during the operation. Any surgical robotic system has to meet specific design considerations for its intended use such as safety, capability of being sterilized, fault-tolerancy, accuracy, stability, and dexterity. MRI-guided applications impose additional demands such as remote control, reduced size, lightweight structure, and ability to operate in the MRI bore. Primarily, there is the issue of MR-compatibility of materials and devices. Conventional robotic systems are not suitable for use inside the MRI scanner because they contain ferromagnetic materials and electrical circuits. These components cause spatial distortions and impart noise to the MR images, while conversely the magnetic field of the MRI system interferes with the electrical circuits. The strong magnetic field dictates that only non-ferromagnetic materials can be used for the mechanical parts.

The major shortcoming in the use of conventional MRI systems for neurosurgery is their reliance on preoperative MR images. As surgery progresses and anatomic tissue are removed or distorted, the intracranial anatomic positional relationship of the brain and surrounding structures change. This is commonly referred to as “brain shift”. Intra-operative changes due to tumor resection, brain swelling, and cerebrospinal fluid (CSF) leakage further increase brain shift [1, 2, 3]. As these processes are unavoidable in most neurosurgical procedures, they decrease the accuracy in all surgery that is based on preoperative MR images [3]. These intra-operative changes make it difficult or impossible to accurately determine the true intra-operative anatomic position of the anatomic target based on the preoperative images. Accurate localization during surgery thus requires the acquisition of intra-operative images. In recent years, advances in computer technology, robotics, and a significant increase in the accuracy of imaging have helped the clinicians in planning and executing surgical procedures in MRI environments. The advantages of surgical robotics are well known in clinical environments due to their precisions, accuracy, repeatability, and capability for tele-surgery [4].

In the area of MRI-guided tele-surgery, there are currently several systems under development. Tajima *et al.* [5] designed and built a prototype of an MRI-compatible manipulator for treatment and diagnosis of heart diseases. Larson *et al.* [6] developed a device to perform minimally invasive interventions in the breast with real time MRI guidance for the early detection and treatment of breast cancer. Engineering Services Inc. (Ontario, Canada) has also developed an MR-compatibletele-robotic system for prostate surgery [7]. Krieger *et al.* [8] designed and developed a novel remotely actuated manipulator (APT-MRI) to access prostate tissue under MRI guidance. Fischer *et al* [9] designed a robotic assistant system using pneumatic components aimed to be used for prostate needle placement in a closed-bore MRI scanner. Kim *et al* [10] designed and developed a new master-slave MR-compatible surgical manipulator for minimally invasive liver surgery. Chinzei *et al.* [11] designed and developed a novel MR-compatible manipulator used to position and direct an axi-symmetric tool such as laser pointer or a biopsy catheter. Moser *et al.* [12] designed and developed a one DOF MR-compatible master-slave robotics system and a haptic interface using hydraulic transmission. Koseki *et al.* [13] designed and developed an endoscope manipulator for trans-nasal neurosurgery capable of being used inside the gantry of vertical field open MRI. Flueckiger *et al.* [14] proposed a haptic interface compatible with MR scanner for neuroscience studies. Miyata *et al.* [15] designed and developed an MR-compatible forceps manipulator using a new cam mechanism for the multi-function micromanipulator system for neurosurgery procedures. Engineering Services Inc. has also developed an MR-compatibletele-robotic system using water hydraulic and pneumatic actuators for neurosurgery [7]. The Calgary Health Region and University of Calgary are developing the world's first image guided neurosurgical robot (NeuroArm$^{TM}$) in collaboration with MD Robotics for micro-neurosurgery. The robot is under design and construction stage now [16]. Nakamura *et al.* [17] developed and manufactured the 6 DOF manipulator using non ferromagnetic materials (aluminum) and actuated by ultrasonic motors.

The goal of our research project is to design, fabricate, and test a hydraulic/pneumatic actuated MR-compatibletele-robotic system for MRI-guided neurosurgery, in particular, the brain biopsy. The mechanical design and related infrastructure are reported.

## 2. ROBOT DESIGN

### 2.1. MR-compatible robotic system infrastructure

MRI-guided tele-robotic system requires surgical planning, MR-image acquisition, human-machine interface, navigation, and sensing. To address those components required for MRI-guided intervention, an infrastructure is needed regardless of the type of surgical operation. A schematic diagram of the proposed infrastructure is illustrated in **Figure 1**. The entire system consists of three main subsystems as follows: (i) operating unit; (ii) power/control unit; and (iii) surgeon-machine interface unit. The operating and surgeon-machine interface units are communicating through MR images and related information using an image processing device. The image processing device is used to provide information required by both the surgeon-machine interface unit and power/control unit. The operating unit and power/control unit are communicating through power transmission and sensory information systems. Also, the surgeon-machine interface and power/control units are communicating through operation inputs created by operator input device (master).

As shown, all three units communicate through

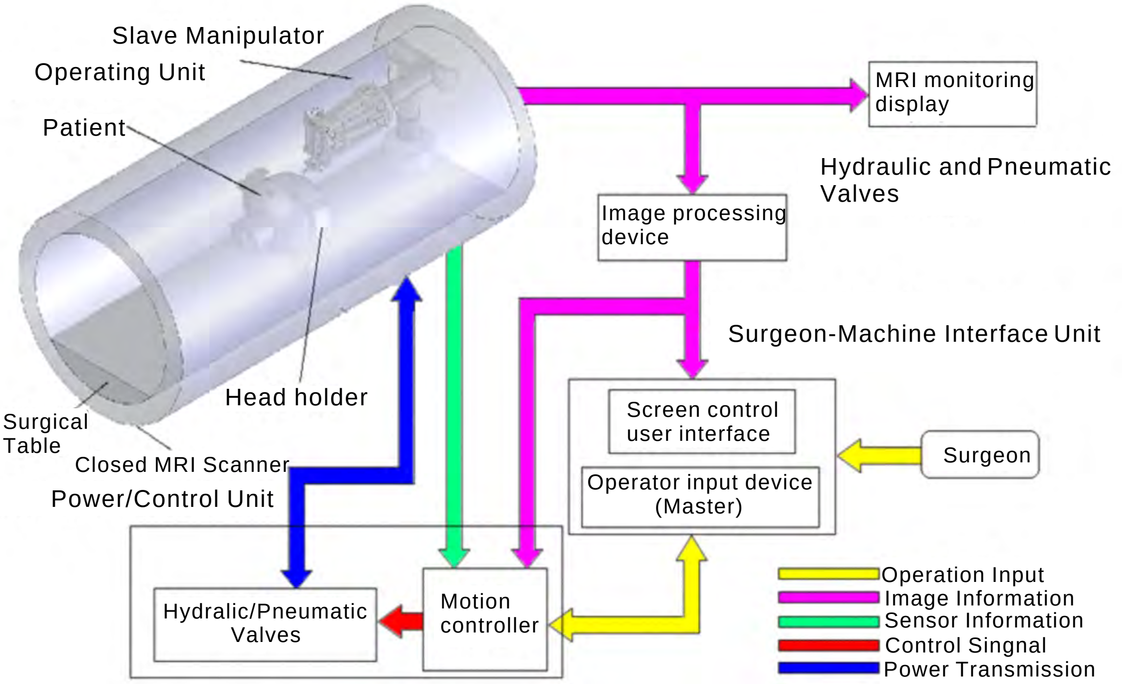


**Figure 1**. A schematic of the entire system.

image information, sensory information, control signals, and power transmission. As illustrated, the visualization of the surgical tool and the target as well as surgical planning based on intra-operative MR images are completed on a display monitor in front of the surgeon in the surgeon-machine interface unit. One should note that the proposed infrastructure is based on a fundamental principle which is both the surgeon and power/control unit share the control of the tele-robotic system such that the surgeon will use his/her judgment and expertise to control the entire procedure. In other words, it is almost impossible to eliminate the surgeon from the control system and have the entire tele-robotic system performed the required task autonomously.

## 2.2. Operating unit

Operating unit comprises the slave manipulator, head holder, surgical table, and MRI scanner located in MR operating room. The patient's head and the slave manipulator are fixed to the surgical table in order to avoid any relative displacement during the surgical operation. The patient's head needs to be secured and fixed in all surgical operations to avoid unexpected motions caused by disorderly reaction of the patient's body.

Due to the presence of strong magnetic field and switching gradients both the head holder and the slave manipulator are required to be constructed from MR-compatible materials and devices. The slave manipulator must perform the required tasks in a confined space between the patient and the bore of the MR scanner. Therefore, the slave manipulator is needed to be designed in a very compact size. In addition, the slave manipulator required to be registered with respect to the MR scanner such that the position and orientation of the surgical tool with respect to the target could be determined based on data obtained from the MR images. One should note that the patient's head must be secured during the operation as the desired position and orientation of the surgical tool with respect to the target will be obtained while the surgical device is outside the patient's skull. Thus, the head holder is considered as a major component in the proposed infrastructure for application of the tele-robotic system in MR-guided neurosurgery procedures.

## 2.3. Manipulator power/control unit

The manipulator power/control unit is located in an adjacent control room at a proper distance away from the MR scanner due to electrical/electronic devices and circuits as well as non-MR-compatiblematerials used in its structure. The major function of the manipulator power/control unit is to provide required power to the slave manipulator. The power/control unit consists of two major sub-units: (i) hydraulic power units, hydraulic valves, and pneumatic valves; and (ii) motion controller devices such as computer and electrical/electronic components and circuits. The surgeon could manipulate the slave manipulator inside the MR scanner through a master manipulator located in the surgeon-machine interface unit. The motion controller in the power/control unit is also communicating with the master manipulator in the surgeon-machine interface unit to provide appropriate control signals to hydraulic and pneumatic valves. The motion controller also receives the sensory data feedback from the slave manipulator. In addition, the motion controller is also provided with the MR images data originated from the image processing device as shown in **Figure 1**.

## 2.4. Surgeon-machine interface unit

The major function of the surgeon-machine interface unit is to provide an interface between the entire tele-robotic system and the surgeon as the end user. The goal of using tele-robotic system for MR-guided neurosurgery is not to replace the surgeon with the robot, but to provide him/her with advanced tools for remote execution of neurosurgical procedures. The unit is located in the adjacent control room to avoid magnetic interference due to use of electrical devices and non-MRI-compatible materials used in its structure. A master and a screen control user interface are the major subsystems of this unit. The images of the

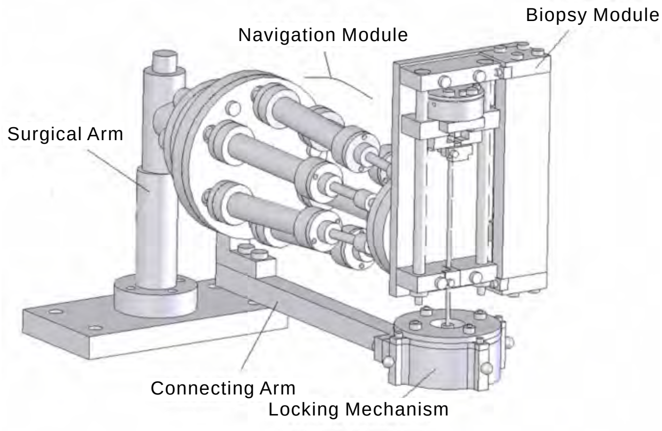


**Figure 2**. 3D model of the slave manipulator.

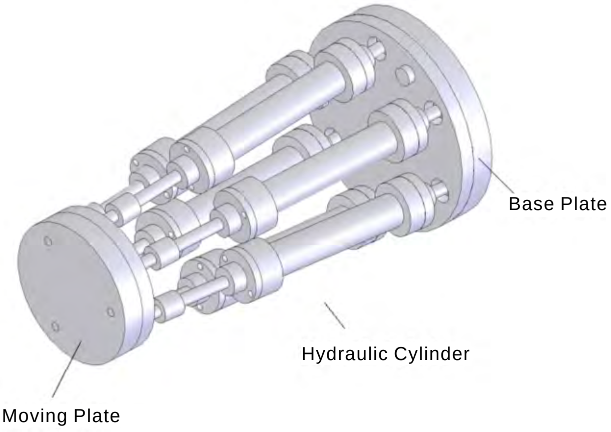


**Figure 3**. 3D model of the navigation module.

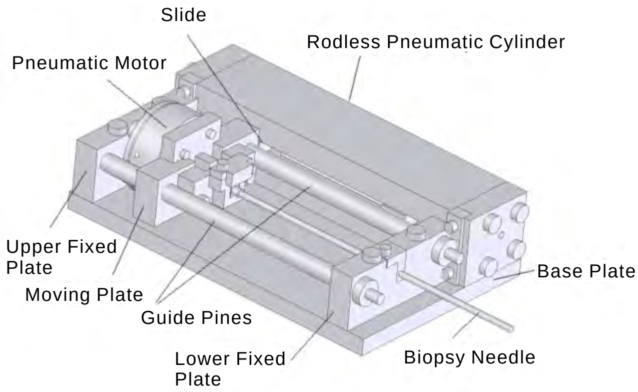


Figure 4. 3D model of the biopsy module.

slave and surrounding environment are projected on the screen to allow visualization of the target and surgical tools movements. The surgeon would manipulate the position and orientation of the surgical devices via the master controller. Surgeons strongly rely on the visual MR images as they are only reliable source of information during the operation. The screen control user interface is the unit that provides the visualization of the tissue and surgical tool while the operation progresses. There are several important challenging issues that one must consider in designing the screen control user interface including [18]: (i) integration of navigation and display with robot systems; (ii) updating the MR images in real time; (iii) providing the surgeon with means of controlling the information displayed; and (iv) finding ways to communicate useful information without overwhelming the surgeon by pointless details. The master manipulator is the unit with which surgeons could communicate their control commands. Any commonly used interfaces for human-machine interactions such as mice, joystick, touch screens, push buttons, and foot switches could be used.

### 2.5. Mechanical design for the slave manipulator

A 3D model of the slave manipulator is shown in Figure 2. The surgical needle is held and advanced by the biopsy module. The biopsy module is attached to the navigation module.

The navigation module is a six degrees of freedom parallel mechanism consisting of a base and a platform interconnected through 6 legs (or struts). Six linear hydraulic actuators are used to provide required linear displacement for each leg. A locking mechanism is used to guide the needle as well as lock the robot at desired orientation. It is fixed to the base of the parallel mechanism through a connecting arm by screws. All three units (the navigation module, biopsy module, and the locking mechanism) are held by a surgical arm. The surgical arm is attached to a surgical table through a set of screws.

### 2.6. Navigation module

A 3D model of the navigation module is shown in Figure 3. It consists of a base plate and a moving plate interconnected through 6 links. Each link consists of a hydraulic linear actuator, a spherical joint, and a universal joint.

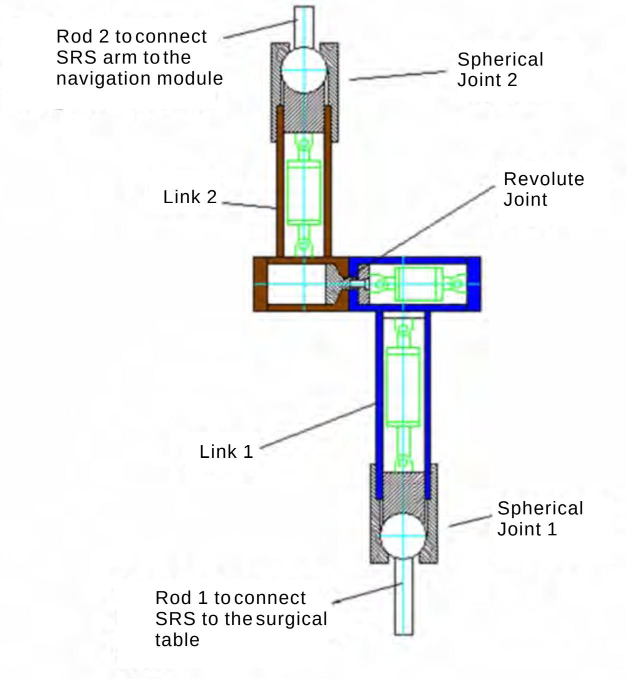

Rod 2 to connect SRS arm to the navigation module
Spherical Joint 2
Link 2
Revolute Joint
Link 1
Spherical Joint 1
Rod 1 to connect SRS to the surgical table

Figure 5. A schematic diagram of the surgical arm.

### 2.7. Biopsy module and locking mechanism

A 3D model of the biopsy module is presented in Figure 4. It is basically a three-plate mechanism including: (i) a lower fixed plate, (ii) anupper fixed plate, and (iii) a moving plate. Both lower and upper fixed plates are attached to the base plate by two sets of screws. Two guide pins are used to support the moving plate. The moving plate is moved up and down using a pneumatic rodless cylinder. The moving plate is attached to the slide of the pneumatic cylinder. A 3D model of the locking system is shown in Figure 2. The locking system consists of a connecting arm and locking mechanism. As shown, the locking mechanism is attached to the base plate of the parallel mechanism through the connecting arm. All mechanical parts are constructed from MR-compatible materials.

### 2.8. Surgical arm

The surgical arm supports both the navigation and biopsy modules during the operation. The surgical arm has to be easily maneuvered by the clinician to be located at the entry point on the patient's skull. The design of the surgical arm is shown in Figure 5. It consists of two links and three joints as follows: (i) a spherical joint 1; (ii) a revolute joint 2; and (iii) a spherical joint 2. The Spherical-Revolute-Spherical (SRS) arm is illustrated in fully deployed configuration in order to show its components and corresponding function of each component. As shown, rod 1 connects the SRS arm to the surgical table and rod 2, at the other end, connects the navigation module to the

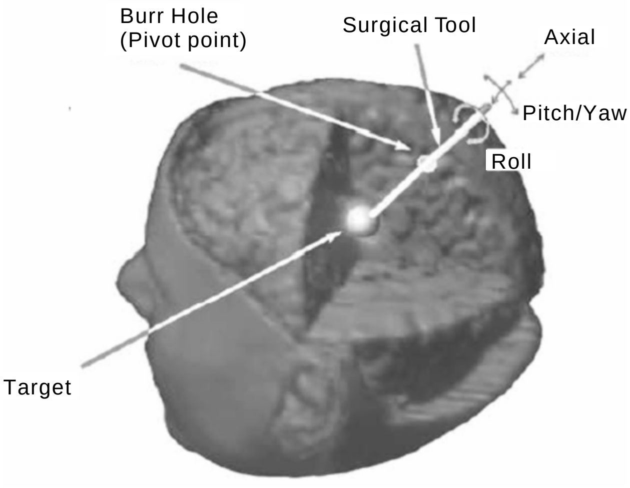


**Figure 6**. The target, entry point, and the needle.

SRS arm.

## 3. ROBOT CONTROL

### 3.1. Needle alignment

An entry point, a surgical tool and a target are depicted in **Figure 6**. Required motions to align and advance the surgical tool with respect to the target are also shown. The surgical tool is rotated about the burr-hole by Yaw and Pitch angles. This point is also called the pivot point. The conventional surgical tool placement at an entry point includes the following three tasks: (i) move the needle tip to the entry point using 3 DOFs; (ii) orient the needle by pivoting around the entry point using 2DOF (Yaw and Pitch angles); and (iii) insert the needle into the body using 1 DOF (translation along a straight trajectory). Using the proposed tele-robotic system shown in **Figure 1**, the brain biopsy procedure would be carried out as follows:

**(1)Preoperative imaging stage.** The patient is placed inside the MRI scanner and preoperative images are obtained.

**(2)Surgical planning stage.** Based on the preoperative images, an entry point is determined and the incision is made by a surgeon.

**(3)Pre-alignment stage.** The slave manipulator is attached to the surgical table, and the navigation module and biopsy needle are manually located at the entry point. Although this stage doesn't require high accuracy in positioning, the slave has to be locked such that the surgical tool is positioned at the entry point. Accurate alignment with respect to target will be done in the next stage;

**(4)Real time navigation stage.** The patient is moved into the bore of MRI scanner. The navigation module is maneuvered remotely in order to align the surgical tool with the desired direction based on intra-operative images.

**(5)Intra-operative operation stage.** The operation is carried out by advancing the needle using intra-operative images as visual feedback. When the needle reaches the target, it is rotated by 180 degrees in order to cut the tissue specimen (tumor). Then the needle is pulled out completing the operation.

**(6)Final stage**. The MRI table is moved out the MRI bore. The slave manipulator and head holder are detached from the table and patient's skull respectively.

### 3.2. Robot control architecture

As mentioned, the surgeon adjusts the orientation of the surgical tool (yaw and pitch angles) based on visual MR images through the master. The inverse kinematics of the navigation module is used to obtain the desired length of each strut related to the desired position and orientation of the needle biopsy.

The hydraulic/pneumatic circuit of the system and overall control system are shown in **Figure 7** and **Figure 8** respectively. Six MR-compatible hydraulic cylinders are equipped with six fiber optic encoders to feedback the actual length of each strut. Using inverse kinematic of the navigation module, the desired length of each strut of the navigation module is determined. A PID controller provides a control signal that drives a hydraulic proportional valve in each servo control loop. The hydraulic valve controls the length of the strut by regulating the flow from/to each hydraulic actuator. In addition, a pneumatic valve (V7) is used to control the tip position of the biopsy needle. The semi-rotary pneumatic motor is also actuated by an on/off pneumatic valve (V8).

A block diagram of the control algorithm used in the controller is shown in **Figure 9**. The inputs are six feedback displacement signals from the slave side (LA1, LA2, LA3, LA4, LA5, and LA6), two signals form master side including desired Yaw and Pitch angels, and desired length of each strut (LD1, LD2, LD3, LD4, LD5, and LD6). The outputs are control signals (S1, S2, S3, S4, S5, and S6) to control the proportional valves.

A PC-based supervisory controller is designed to control entire system as illustrated in **Figure 10**. The trajectory of each joint is calculated based on the inverse kinematics in a PC-based supervisory controller and fed to each joint controller RS485 Bus. As shown in **Figure 10**, six optical encoders are used to feedback the position signals to six microprocessors. Each actuator has individual microprocessor to control its proportional valve.

## 4. CONCLUSION AND FUTURE WORK

We have designed an MR-compatible tele-robotic system that can be used for orientation and advancement of a biopsy needle on the brain biopsy procedure. The robot has been designed such that it will perform desired tasks inside MR scanner GE Signa 1.5T. To date, design and analysis of the entire system have been completed. Material selection and the controller architecture and its component have been finalized. A physical prototype of the slave manipulator is in the process of being constructed. Current and future

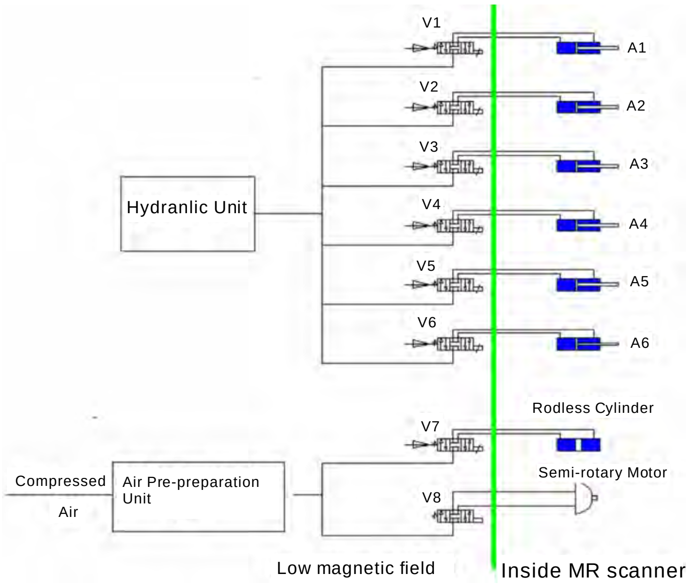


**Figure 7.** A schematic of hydraulic/pneumatic circuit.

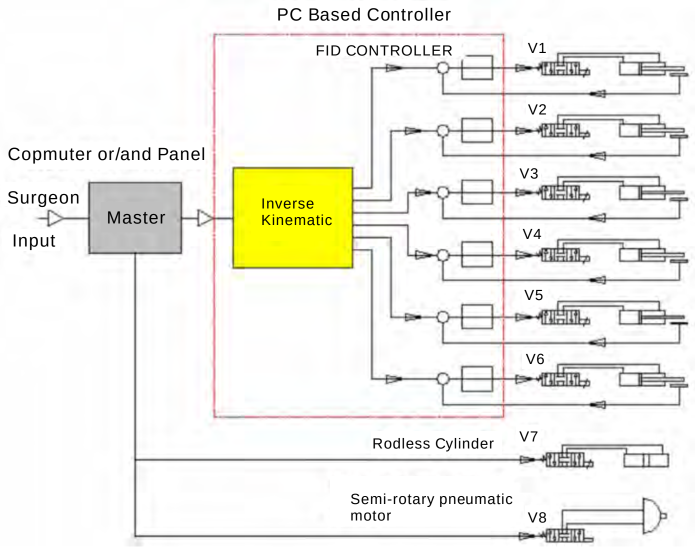


**Figure 8.** A schematic of overall control architecture.

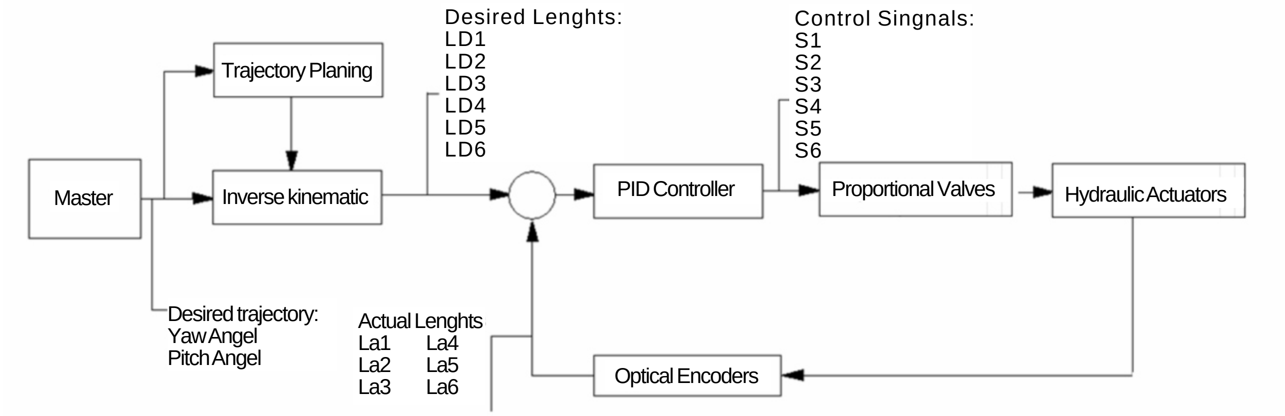


**Figure 9.** A schematic of overall control architecture.

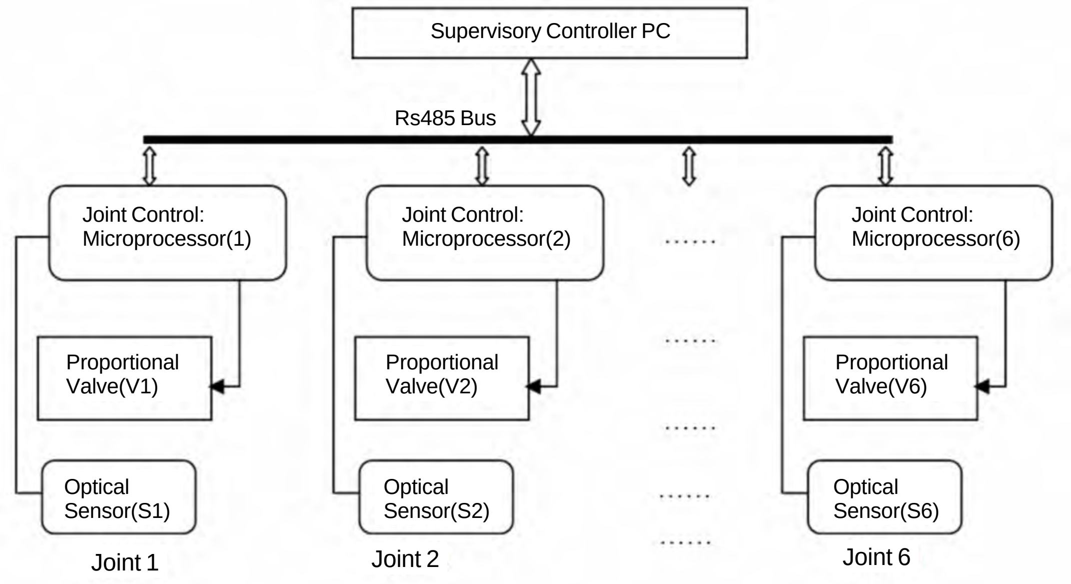


**Figure 10**. Supervisory control configuration.

work includes the development of the slave manipulator and performance of series of experimental tests inside the MR scanner using the first physical prototype.

## ACKNOWLEDGEMENTS

This work was partially supported by Natural Sciences and Engineering Research Council of Canada (NSERC), grant held by Professor Andrew A. Goldenberg and Ontario Research and Development Challenge Fund (ORDCF), grant held by Professor W. Kucharczyk. The authors would like to thank Engineering Services Inc., Canada, and MRI department in Toronto General Hospital, Canada, for providing help and equipment.

## REFERENCE

[1]R. D. Howe, & Y. Matsuoka. Robotics for Surgery. *Annual Review of Biomedical Engineering* 1999, 1:211-240.

[2]J. Kettenbach, D. F. Kacher, S.K. Koskinen, S. Silverman, A. Nabavi, D. Geringt, C. Tempany, R. B. Schwartz, R.Kikinis, P. K. Black & F.A. Jolesz. Interventional and Intra-operative Magnetic Resonance Imaging. *Annual Review Biomedical Engineering* 2000, 2:661-90.

[3]A. Nabavi, D. F. Kacher, D. T. Gering *et al.* Neurosurgical procedure in 0.5 Tesla, open-configuration intraoperative MRI: planning, visualization, and navigation. *Automedia* 2001, 00:1-35.

[4]K. Chinzei, N. Hata, F.A. Jolesz, & R. Kikinis, Medical Image Computing and Computer-Assisted Intervention - MICCAI 2000. *Third International Conference Proceedings* 2000, pages 921-30.

[5]F. Tajima, K. Kishi, K. Nishizawa, K. Kan, Y. Nemoto, H. Takeda *et al.* Development of MR-compatible Surgical Manipulator toward a Unified Support System for Diagnosis and Treatment of Heart Disease. *Proc .of MICCA O2* 2002, pages 83-90.

[6]B. Larson, N. Tsekos & A. g. Erdman. A Robotic Device for Minimally Invasive Breast Intervention with Real-Time MRI Guidance. *Proc. Of the Third IEEE Symposium on Bioinformatics and Bioengineering* 2003.

[7]Engineering Services Inc., http://www.esit.com, *Internal report.*

[8]A. Krieger, R. Susil, C. Menard, J. Coleman, G. Fichtinger, E. Atalar, & L. Whitcomb. Design of a Novel MRI Compatible Manipulator for Image Guided Prostate Intervention. *IEEE Trans. On Biomedical Engineering* 2005, 52(2):306-313.

[9]G.S. Fischer, I. Iordachita, S. P. DiMaio & G. Fichtiger. Design of a Robot for Transperineal Prostate Needle Placement in MRI scanner. *IEEE International Conference on Mechatronics 2006,* page 6.

[10]K. Daeyoung *et al.* A New, Compact MR-Compatible Surgical Manipulator for Minimally Invasive Liver Surgery. *MICCAI* 2002, pages 99-106.

[11]K. Chinzei & K. Miller. MR Guided Surgical Robot. *Proc. 2001 Australian Conference on Robotics and Automation* 2001, Sydney.

[12]R. Moser, R. Gassert, E. Burdet, L. Sache, H. Woodtli, J. Erni, W. Maeder & H. Bleuler. An MR-compatible Robot Technology. *Proc. Of the IEEE, International Conference on Robotics & Automation* 2003.

[13]Y. Koseki, T. Washio, K. Chinzei & H. Iseki. Endoscope Manipulator for Trans-nasal Neurosurgery, Optimized for and Compatible to Vertical Field Open MRI. *Proc. of MICCAI* 2002, pages 114-121.

[14]M. Flueckiger, M. M. Bullo *et al.* FMRI compatible haptic interface actuated with traveling wave ultrasonic motor. *IAS Annual Meeting (IEEE Industry Applications Society)* 2005, 3:2075-2082.

[15]N. Miyata , E. Kobayashi, D. Kim, K. Masamue *et al.* Micro-grasping Forceps Manipulator for MR-Guided Neurosurgery. *MICCAI* 2002, pages 107-113.

[16]Calgary HealthTrust, www.cbi.ucalgary.ca/CHT, 2004.

[17]R. Nakamura, K. Masamune, Y. Nishikawa, E. Koboayashi, I. Sakuma,T. Dohi, H. Iseki & K. Takakura. Development of a sterilizable MRI-compatible manipulator for stereotactic neurosurgery. *Proc. Of Computer Assisted Radio Surgery (CAR'99) 1999.*

[18]*R. Taylor & D. Stoianovici. Medical Robotics in Computer-Integrated Surgery. IEEE Transaction on Robotics and Automation 2003, 32(5):765-781.*

Call for Papers

# Journal of Biomedical Science and Engineering (JBiSE)

www. srpublishing. org/journal/jbise

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

## Subject Coverage

Bioelectrical and neural engineering

Bioinformatics

Medical applications of computer modeling

Biomedical modeling

Biomedical image processing & visualization

Real-time health monitoring systems

Biomechanics and bio-transport

Patten recognition and medical diagnosis

Biomedical effects of electromagnetic radiation

Safety of wireless communication devices

Biomedical devices, sensors, and nano technologies

NMR/CT/ECG technologies and EM field simulation

Physiological signal processing

Medical data mining

Scientific Research Publishing

JBiSE

ISSN:123-456-78910

Vol. 1 No. 1 May 2008

Journal of Biomedical Science & Engineering

www. SRPublishing. org/journal/jbise

Editor-in-Chief

Kuo-Chen Chou

Huai-Bei Zhou

## Editors–in–Chief

**Kuo-Chen Chou**
Gordon Life Science Institute, San Diego,California, USA

**Huai-Bei Zhou**
Wuhan University, Wuhan,Hubei, China

## Notes for Intending Authors

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

## Website and E-Mail

www. srpublishing. org/journal/jbise
Email: jbise@srpublishing.org

## Editorial Board

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

**J. Biomedical Science and Engineering, 2008, 1.**
Published Online May 2008 in SciRes (http://www.srpublishing.org/journal/jbise).

# TABLE OF CONTENTS

**Volume 1** **May 2008**