

Automatic Seizure Detection Based on Star Graph Topological Indices

Enrique Fernandez-Blanco*, Daniel Rivero, Juan Rabuñal, Julián Dorado, Alejandro Pazos, and Cristian Robert Munteanu

*Department of Information and Communication Technologies, Computer Science Faculty,
University of A Coruña, 15071, A Coruña, Spain; efernandez@udc.es, rivero@udc.es, juan-ra@udc.es, julian@udc.es, apazos@udc.es, muntisa@gmail.com*

* **Corresponding author:** Enrique Fernández-Blanco, Computer Science Faculty, University of A Coruña, Campus de Elviña, S/N 15071 A Coruña, Spain; Fax: +34 981 167 160; Email: efernandez@udc.es.

Abstract

The recognition of seizures is very important for the diagnosis of patients with epilepsy. The seizure is a process of rhythmic discharge in brain and appears rarely and unpredictably. This behavior generates a need of an automatic detection of seizures by using the signals of the long-term electroencephalograph (EEG) recordings. Due to the non-stationary character of EEG signals, the conventional methods of frequency analysis are not the best alternative to obtain good results in diagnostic purpose. The present work proposes a method of EEG signal analysis based on star graph topological indices (SGTIs) for the first time. The signal

information such as amplitude and time occurrence is codified into invariant SGTIs which are the basis for the classification models that can discriminate the epileptic EEG records from the non-epileptic ones. The method with SGTIs and the simplest discriminant linear methods provide similar results to those previously published, which are based on the time-frequency analysis and artificial neural networks. Thus, this work is proposing a simpler and faster alternative for automatic detection of seizures from the EEG recordings.

Keywords: Automatic signal processing, epilepsy seizure detection, EEG signal, star graphs, linear discriminant analysis

1 INTRODUCTION

Epilepsy is one of the most common neurological disorders and it is also known as a seizure disorder which affects the nervous system. A seizure is a sudden surge of electrical activity in the brain that usually affects how a person feels or acts for a short time. Seizures are symptoms of many different disorders that may affect the brain. Depending on the brain regions involved during the course of seizure, epilepsies can be divided into two main classes such as the following: generalized seizures that involve almost the entire brain and focal (partial) seizures that remained restricted on a circumscribed brain region (epileptic focus). These seizures may be accompanied by psychic, autonomic, sensory or motor disturbances (Lehnertz et al., 2003; Litt and Echauz, 2002).

The usual procedure that can detect the suspected seizures is the electroencephalogram (EEG), typically a 20-minute brain-wave recording. This method yields numerous practical advantages. The duration of the procedure is too short for detecting the epileptic seizures and some of the registered seizures cannot be clearly linked to an etiology. One solution was the long-term EEG registration by using the portable recording systems in the normal environment of the patient (Waterhouse, 2003). Clinical neurophysiologists can analyze for several days recordings in a time-consuming process. Even if theoretically the patient can use the alarm button to warn about a seizure, in many cases they are not aware of the occurrence of the epileptic seizures. Thus, an automatic detection method of the seizure patterns is necessary. The identification of these patterns in the extracranial recordings is made difficult due to the obscuring of the real seizures by EMG, movements or eye-blink artifacts.

Another way to analyze the information of complex systems is represented by the graphical methods such as graph or complex network theory (Strogatz, 2001). Thus, the original information of any real system can be coded into graphs and their sets of non-variant numbers that characterize its topology, named topological indices (TIs) (Dehmer et al., 2008; Emmert-Streib and Dehmer, 2009b). These TIs are used as input to obtain the best classification models by using the statistical, Artificial Neural Network (ANN) or Machine Learning methods. In biological or biomedical systems, these graph approaches, can provide an intuitive vision and useful insights for helping to analyze complicated relations therein. Many previous

studies on a series of important biological topics have indicated that relation, for example, in protein folding kinetics(Chou, 1990), predicting protein folding rates(Chou and Shen, 2009a), enzyme-catalyzed reactions (Andraos, 2008; Chou, 1989; Chou and Forsen, 1980; Chou and Liu, 1981; Kuzmic et al., 1992; Myers and Palmer, 1985; Zhou, 1984), inhibition kinetics of processive nucleic acid polymerases and nucleases(Althaus et al., 1996), analyzing drug metabolism system(Chou, 2010), investigation into the slow conformational change with Chou's non-steady-state graphical rule (Lin and Neet, 1990), analyzing inhibition of HIV-1 reverse transcriptase(Althaus et al., 1993a; Althaus et al., 1993b; Althaus et al., 1993c), analyzing inhibition kinetics of processive nucleic acid polymerases and nucleases(Chou et al., 1994), DNA sequence analysis (Qi et al., 2007), analysis of codon usage(Zhang and Chou, 1994), microarray analysis (Emmert-Streib and Dehmer, 2009a) and analyze protein-protein interaction(Zhou, 2011a, b) with wengxiang diagrams(Chou et al., 2011). More information about apply graphs in complicated network system research can be obtained in (Diao et al., 2007; Gonzalez-Diaz et al., 2008; González-Díaz et al., 2007). In addition, several works have analyzed and classified information that is similar to EEG signals such as the blood proteome mass spectra. Thus, star and lattice graphs have been used to generate models that can discriminate whether a patient has or not several diseases such as prostate (Ferino et al., 2008), breast (Vilar et al., 2008a) or colon (Munteanu et al., 2009) cancers.

The present work proposes an automatic method to identify the epilepsy sei-

zures by using a classification model based on topological indices of the EEG recording star graphs and linear discriminant methods. Similar to the quantitative proteome-disease relationship (QPDR) for the proteome mass spectra (Ferino et al., 2008; Munteanu et al., 2009; Vilar et al., 2008a), this work is aimed at finding the best quantitative signal-disease relationship (QSDR). Previous works have used different graphs/networks for the simulation of epileptic seizures by a coupled ordinary differential equation lattice model (Larter et al., 1999), to study the effects of different connectivity topologies in Small World Networks on EEG-like activities (Min et al., 2006), to study brain complex networks (Bullmore and Sporns, 2009; Micheloyannis et al., 2009), to evaluate the neural efficiency hypothesis by using graph theoretical analysis of multi-channel EEG (Micheloyannis et al., 2006) or for data analysis for continuous EEG monitoring in the Intensive Care Unit (Scheuer and Wilson, 2004).

2 STATE OF THE ART

The processing of EEG signals is a topic that has had a great impact from its very beginning. Epileptic EEGs have been analyzed with many different tools and therefore their processing is very well documented (Kannathal et al., 2005; Mohseni et al., 2006b; Nigam and Graupe, 2004; Polat and Güneş, 2007; Srinivasan et al., 2005; Subasi, 2007).

The early observations of EEG recordings showed that their spectra contain some characteristic waveforms that fall primarily within four frequency bands.

This was the basis of the first methods of automatic EEG processing, which were based on a Fourier transformation. However, this method uses the fast Fourier-transform algorithm (FFT), which suffers from large noise sensitivity. A better frequency resolution which solves these spectral loss problems is obtained through power estimation by means of parametric methods. However, these methods are not suitable for frequency decomposition of these signals due to their non-stationarity nature.

The most common approach for EEG signal classification consists of two stages (in general for any pattern classification system): feature extraction using any signal processing tool and classification with any machine learning tool, such as ANNs or Support-Vector Machines (SVMs)(Rivero et al., 2011).

One of the most powerful tools for signal and image processing appeared in the late 1980s and it is called the wavelet transform (Addison, 2002). This technique has had a great impact and success, and now it is one of the most used tools in the field of signal processing with many scientific and engineering applications. It is also a very suitable tool in EEG processing, since it is appropriate for the analysis of non-stationary signals, which is a major advantage over spectral analysis (Mohseni et al., 2006b). This tool is suited for locating transient events. An example of these events refers to the spikes that occur during epileptic seizures in EEG signals. In a recent work by Subasi (Subasi, 2005; Subasi and Gursoy, 2010), EEG signals have been analyzed by means of the discrete wavelet transform. The signals were first decomposed into five levels using a DB4 filter and wavelet coeffi-

cients were selected as inputs to a dynamic fuzzy neural network and dynamic wavelet network, but they have obtained low performances on classification accuracy.

Other different analysis of the signal is carried out by the extraction of entropy-based features. Since the application of the concept of entropy to the information theory by Shannon, entropy is known as the amount of information stored in a more general probability distribution. Different entropy estimators have been applied to quantify the complexity of a signal. This has led to the extraction of different features that have been used as inputs of a classifier system (Kannathal et al., 2005).

Another approach refers to the study of EEG signals as chaotic systems. From this point of view, Lyapunov exponents are extracted from these signals. These exponents are a quantitative measure for distinguishing among the various types of orbits based upon their sensitive dependence on the initial conditions, and are used to determine the stability of any steady-state behavior (Abarbanel et al., 1991). Using Jacobi matrices, all these exponents can be estimated from a time series. Different features can be extracted from these exponents, which are subsequently used for the classification (Guler et al., 2005).

Time-frequency features are also powerful tools for this task. An epileptic signal has components in both time and frequency, but the conventional time and frequency representations present only one aspect. By means of computation of a time-frequency distribution, a signal is localized in both time and frequency do-

mains. This technique was used in some works, using the pseudo Wigner-Ville and the smoothed-pseudo Wigner-Ville distribution (Mohseni et al., 2006a; Tzallas et al., 2007). Features were extracted from this distribution of epileptic EEG signals, and they were used as inputs to a feed-forward back-propagation network.

The classification of the signals using these features can be carried out by different classification algorithms. SVMs and Linear Discriminant Analysis are two of the most used methods for this task. However, the ANNs are the most used tool to perform this classification. This paper used recurrent ANNs, which present an additional advantage: for the classification of a single window of the signal, the previous outputs of the neurons are taken into account. This means that the knowledge from the previous samples of the signal is used for the classification.

Recently, ANNs that apply Bayesian methods have been shown to be more robust compared with other techniques because they incorporate measures of confidence in their output for the Levenberg-Marquardt (LM) procedure (Vuckovic et al., 2002). In addition, standard Multilayer Perceptron (MLP) has been improved by using finite impulse response filters (FIR) instead of static weights for a temporal data processing (Tzallas et al., 2007).

3 MATERIALS AND METHODS

Figure 1 presents the flowchart of the new method based on star graph topological indices (SGTIs) of the epilepsy EEG recordings. The set of signal amplitudes of

each EEG segment is transformed into specific SGTIs that will be the input for the statistical methods that can find the best theoretical model and can discriminate between the epilepsy segments and the non-epilepsy ones. Additionally, these models based on the signal amplitudes along the time coordinate were compared with similar models by using the Fourier transformed data and transforming the frequency data into SGTIs. Details about each sub-step are presented in the paragraphs below.

Figure 1 comes about here

3.1 EEG segment sets

The EEG dataset includes recordings of both healthy and epileptic subjects (Andrzejak et al., 2001) and it is split into five subgroups (denoted Z, O, N, F and S) depending on the patient's diagnosis and sensor location. Each subgroup contains 100 single-channel EEG segments of 23.6 seconds. The non-seizure records are noted as Z for EEG registrations in patients with eyes open, O for eyes-closed EEG recordings, N for the case of hippocampal formation sensor location, and F for the epileptogenic zone. The only seizure segments are denoted as S. Z and O have been recorded extracranially whereas N, F and S have been recorded intracranially. Signals were selected after visual inspection for artifacts and have passed a week stationarity criterion.

Table 1 comes about here

The present calculations are based on three principal types of grouping of the Z,

O, N, F, S sets (see Table 1), in a similar way as Tzallas et al. (Tzallas et al., 2007). The classification problems depend on the presence or absence of epilepsy seizures and on the sensor location and recording as follows:

- Z-S (C1): normal eyes open (Z) versus seizures (S);
- O-S (C2): normal eyes closed (O) versus seizures (S);
- ZONF-S (C3): all non-seizure classes (ZONF) versus seizures (S).

Two of the classification problems from the Tzallas et al. work (Tzallas et al., 2007) (1 and 3) that involved more than two classes are solved indirectly as follows:

- Z-N-S (C4) using ZN-S (C4.1), NS-Z (C4.2) and ZS-N (C4.3) classifications;
- ZO-NF-S (C5) using ZONF-S (C5.1), NFS-ZO (C5.2) and ZOS-NF (C5.3) classifications.

Thus, the three-class model is evaluated by using the maximum probability of classification for each class in the individual model. An additional model based on the same Z-S classes (non-seizure open eyes versus seizures) and Fourier transformed data (C6) was generated in order to study the influence of this frequency transformation.

3.2 Star Graph Topological Indices

In previous works, several graphical representation have been used (lattice, spiral and star graphs) for characterize a diversity of complex systems such as drugs

(Dehmer et al., 2009; Molina et al., 2004), proteins (Chen et al., 2010a; Chen et al., 2010b; Hu et al., 2011; Huang et al., 2011; Munteanu et al., 2008), nucleic acids (González-Díaz et al., 2005), proteome mass spectra (Cruz-Monteagudo et al., 2008a) or drugs action on parasites (Prado-Prado et al., 2008). A graph is the abstract representation of a real complex network and it consists of nodes (vertices) and links between some of them with similar characteristics (Harary, 1969).

The star network/graph (SG) (Harary, 1969; Randić et al., 2007) is one of the most common computer network topologies and consists of one central switch, hub or computer, which acts as a conduit to transmit messages along the nodes (Lawrence and Barry). Thus, the hub and computer nodes, and the transmission lines between them, form a graph with the topology of a star. In the case of the proteins, the star graph can have 20 possible branches (“rays”) for each type of amino acid (star center is a non-amino acid vertex). Therefore, the information of the protein primary structure such as the type, the composition and the position of the amino acids in the protein chain are coded into SGTIs. The graph connectivity of the nodes is different compared with the original one inside the sequence. In addition, if the initial connectivity is added, the graph is an embedded one and the topological indices have the “e” suffix. In the case of the mass spectra, the signal amplitudes are grouped into intervals that correspond to different SG branches. In this way, the encoded information into the SGTIs is represented by the signal amplitude, the position in their recordings and the number of interval occurrences.

In the present work, each EEG segment was transformed into a series of characteristic SGTIs, similar to the proteome mass spectra (Cruz-Monteagudo et al., 2008b; Gonzalez-Diaz et al., 2008). The data were automatically transformed with S2SNet (Munteanu and Gonzáles-Diáz, 2008; Munteanu et al., 2012), a tool that can convert any string/sequence into star network/graph topological indices based on defined groups of string elements. In the first step, one S2SNet filter was used to transform the amplitude values into a string for each recording. In this step, the number of signal amplitude intervals (SG branches) was defined as 80 (the maximum types of characters from S2SNet), considering that this precision is more than sufficient if compared with previous papers on mass spectra. The transformation is illustrated by an example with one short fragment of a possible EEG record: “2 4 6 8 9 6 13 16 19 17 21 24 27 33 31” that will become the correspondent string, “AABBBBCDDDEEFGG”, where A, B, C, D, E, F, G letters correspond to different EEG amplitude signal intervals as follows: A = 0 to 5, B = 5 to 10, C = 10 to 15, D = 15 to 20, E = 20 to 25, F = 25 to 30 and G = 30 to 35 (see Figure 2).

Figure 2 comes about here

In the second step, the string is transformed into the embedded and non-embedded star graphs and their correspondent SGTIs by using the main module calculation of S2SNet. Due to the large number of amplitudes (4097) in every EEG recording, each case was divided into two parts on the time scale, the left (L) and the right (R) fragments (2048 and 2049 values). Thus, two groups of TIs are

obtained (${}_L\text{TI}/{}_L\text{Tle}$ and ${}_R\text{TI}/{}_R\text{Tle}$) and they are used to create the best classification model (Figure 1). The segments have been divided only into two parts in order to minimize the loss of information on the relations between consecutive signal amplitudes, and to avoid the computational complexity for the classification by having the minimum number of TIs as inputs.

The TIs calculation is based on several elements of the graph such as the initial and graph connectivity matrices, the distance matrix and the degree matrix (Bornholdt and Schuster, 2003). The connectivity matrices (M) in the sequence and in the star graph are combined in the case of the embedded graph. The dimension of this matrix is i by i , where i represents the number of nodes or the length of the sequence. The matrix connectivity of the embedded graph includes the initial sequence connectivity. The distance matrix (d) represents the number of nodes between two nodes. Due to the fact that the graph is undirected, M_{ij} is equal to M_{ji} and d_{ij} is equal with d_{ji} (the matrices M and d are symmetric). The degree vector (deg_i) is the number of connectivities of each node.

S2SNet (Sequence to Star Networks) tool was developed by our group, and it is based on wxPython (Rappin and Dunn, 2006) and it uses Graphviz (Koutsofios and North, 1993) as a graphics back-end. In this study, all the calculations are based on embedded (Emb) and non-embedded ($nEmb$) TIs, nodes without weights, Markov normalization and power of matrices/indices (n) up to 5. The results of the S2SNet calculation contain the following eight types of TIs (Todeschini and Consonni, 2002):

- Shannon entropy of the n powered Markov matrices (\mathbf{Sh}_n):

$$Sh = \sum_i p_i * \log(p_i), \quad (1)$$

where p_i are the n_i elements of the p vector, resulted from the matrix multiplication of the powered Markov normalized matrix ($n_i \times n_i$) and a vector ($n_i \times 1$) with each element equal to $1/n_i$;

- Trace of the n connectivity matrices (\mathbf{Tr}_n):

$$Tr_n = \sum_i (M^n)_{ii}, \quad (2)$$

where $n = 0 - 5$ (power limit), M = graph connectivity matrix ($i*i$ dimension); $ii = i^{\text{th}}$ diagonal element;

- Harary number (\mathbf{H}):

$$H = \sum_{i < j} \frac{M_{ij}}{d_{ij}}, \quad (3)$$

where d_{ij} are the elements of the distance matrix and M_{ij} are the elements of the M connectivity matrix;

- Wiener index (\mathbf{W}):

$$W = \sum_{i < j} d_{ij}, \quad (4)$$

- Gutman topological index (\mathbf{S}_6):

$$S_6 = \sum_{ij} \frac{\deg_i * \deg_j}{d_{ij}}, \quad (5)$$

where \deg_i are the elements of the degree matrix;

- Schultz topological index (non-trivial part) (\mathbf{S}):

$$S = \sum_{i < j} (\text{deg}_i * \text{deg}_j) * d_{ij}, \quad (6)$$

- Balaban distance connectivity index (**J**):

$$J = \frac{\text{edges}}{\text{edges} - \text{nodes} + 2} \sum_{i < j} M_{ij} \sqrt{\sum_k d_{ik} \sum_k d_{kj}}, \quad (7)$$

where $\text{nodes}+1=AA$ numbers/node number in the star graph+origin, $\sum_k d_{ik}$ is the node distance degree;

- Kier-Hall connectivity indices (${}^n\mathbf{X}$):

$${}^0X = \sum_i \frac{1}{\sqrt{\text{deg}_i}}, \quad (8)$$

$${}^2X = \sum_{i < j < k} \frac{M_{ij} * M_{jk}}{\sqrt{\text{deg}_i * \text{deg}_j * \text{deg}_k}}, \quad (9)$$

$${}^3X = \sum_{i < j < k < m} \frac{M_{ij} * M_{jk} * M_{km}}{\sqrt{\text{deg}_i * \text{deg}_j * \text{deg}_k * \text{deg}_m}}, \quad (10)$$

$${}^4X = \sum_{i < j < k < m < o} \frac{M_{ij} * M_{jk} * M_{km} * M_{mo}}{\sqrt{\text{deg}_i * \text{deg}_j * \text{deg}_k * \text{deg}_m * \text{deg}_o}}, \quad (11)$$

$${}^5X = \sum_{i < j < k < m < o < q} \frac{M_{ij} * M_{jk} * M_{km} * M_{mo} * M_{oq}}{\sqrt{\text{deg}_i * \text{deg}_j * \text{deg}_k * \text{deg}_m * \text{deg}_o * \text{deg}_q}}; \quad (12)$$

- Randic connectivity index (${}^1\mathbf{X}$):

$${}^1X = \sum_{ij} \frac{M_{ij}}{\sqrt{\text{deg}_i * \text{deg}_j}}. \quad (13)$$

92 indices ($n=5$) corresponding to the left and right parts of each recording, and to the embedded (23_{LTI} and 23_{RTI}) and non-embedded star graphs (23_{LTI} and 23_{RTI}) will be the input of the linear statistical methods in order to obtain the best

classification model that can predict whether an EEG recording contains an epilepsy seizure or not.

The models based on the SGTIs of the original signal amplitudes are compared with the Fourier transformed data. In the first step, the signal amplitudes in time are converted into frequencies with a Fourier transformation. In the second step, the scale is changed with a logarithmic one due to the large differences in the obtained frequencies. The resulted plot of the frequencies in logarithmic scale is symmetric and only half of the data is necessary leading to cases with 2049 frequency values. In the last step, these values are transformed into a string and finally into 46 SGTIs [23 TI(FT) and 23 Tle(FT)] by using S2SNet. The information codification of the signal amplitudes (or frequencies) involves a contraction of the variable number for the classification input from 4097 to 92 (or 2049 to 46 for the Fourier transformation case).

3.3 Statistical Analysis

Several classification models have been created with General/Linear Discriminant Analysis (GDA/LDA) method (Kowalski and Wold, 1982; Van Waterbeemd, 1995) from STATISTICA 6.0 package (StatSoft.Inc., 2002). These models can predict whether an EEG segment contains an epilepsy seizure and, consequently, if a patient has epilepsy. In the main models (C1-C3), there are only two groups of cases, one containing epilepsy seizure EEG segments and the other corresponding to non-epilepsy ones. The dependent variable (*EpilepsyOrNot*) takes 1 for epilep-

sy seizures and 0 for non-epilepsy cases (in the case of non-mixed classes), and the cross-validation variable (CV) takes two values ($train$ and val). The best cross-validation methods used to examine a predictor are the independent dataset test, the subsampling test, and the jackknife test (Chou and Zhang, 1995). The jackknife test has been increasingly used by researchers to examine the accuracy of various predictors (Chen and Li, 2007). In the present work, the independent data test is used by splitting the data at random in a training series ($train$, 50%) used for model construction and a prediction one (val , 50%) for model validation. A 10-fold cross-validation test was carried out for the entire model. Due to the big differences between the value ranges of the TIs, all the independent variables are standardized prior to model construction.

The general QSDR formula contains embedded and non-embedded TIs from the left and right parts of the segments:

$$EpSc = c_0 + \sum_{i=1}^n c_i *_{L} TI_i + \sum_{j=n}^m c_j *_{L} TLe_j + \sum_{k=m}^o c_k *_{R} TI_k + \sum_{l=o}^p c_l *_{R} Tle_l, \quad (14)$$

where $EpSc$ is the continue score value for the epilepsy/non-epilepsy classification, c_1 - c_p are the $_{L/R}TI(\mathbf{e})$ coefficients and c_0 is the independent term. We inspected the percentage of a good classification and the number of variables to be explored in order to avoid over-fitting or chance correlation. The Forward model type was tested for all the classifications and the simple linear mathematical form of the model has been chosen in absence of prior information.

In addition, the EEG segments are used to create classification models based on

the Fourier transformed signals for Z versus S sets. The obtained frequencies are used as input for S2SNet application and a similar series of 92 TIs was calculated.

The model is similar to (14):

$$\begin{aligned}
 EpSc(FT) = & d_0 + \sum_{i=1}^n d_i *_{L} TI_i(FT) + \sum_{j=n}^m d_j *_{L} TLe_j(FT) \\
 & + \sum_{k=m}^o d_k *_{R} TI_k(FT) + \sum_{l=o}^p d_l *_{R} TLe_l(FT)
 \end{aligned} \tag{15}$$

where $EpSc(FT)$ is the continue score value for the epilepsy/non-epilepsy classification using the Fourier transformation filter, d_1 - d_p are the $_{L,R}TI(e)(FT)$ coefficients and d_0 is the independent term. The results are compared in order to show the effect of the FT on the actual method based on SGTIs. The reported percentages are accuracy (Ac), sensitivity (Se) and specificity (Sp) that measure the ratio of the number of total, epilepsy (or class 1) or non-epilepsy (or class 0) recordings correctly classified by the model with respect to the real classification.

4 Results and discussion

Thirty classification models were tested by using five types of grouping the Z, O, N, F, S sets (see Table 1), depending whether they correspond to the epilepsy seizures or not in order to compare the results Tzallas et al. work (Tzallas et al., 2007). The aim is to find the best GDA equation able to discriminate between EEG recordings with and without epilepsy seizures. The attributes include 92 embedded and non-embedded SGTIs obtained with S2SNet application. The values obtained for the 10-fold accuracies (Ac), selectivity (Se) and specificity (Sp) with

the *Forward Stepwise* method are presented in Table 2. The models were based on non-embedded, embedded and both groups of topological indices.

Table 2 comes about here

The best model was considered Z-S (C1 in Table 1) based on non-embedded SGTIs due to the practical reason of extracranial measurements, the small number of the equation indices (only three) and the high values of correctly classified compounds of 99.00%, 98.21% and 100% for the test accuracy, selectivity and specificity, respectively (Eq. 15). Details about the 10-fold classification values are shown in Table 2.

$$EpSc = -0.18 + 120.13*_R Sh_0 + 114.38*_R^0 X - 21.05*_R^5 X \quad (16)$$

$$N = 200, Rc = 0.95, U = 0.10; Chi = 224.99, p < 0.001$$

where N is the number of cases (Z and S), Rc is the canonical regression coefficient, U is the Wilk's statistics, Chi is the Chi-Square statistics and p is the p-level (probability of error). Rc value shows a high level of correlation between the input variables and the classification of EEG recordings. Wilk's U is used to measure the statistical significance of the discriminatory power of the model and its values range from 1.0 (non-discriminatory power) to 0.0 (perfect discriminatory power). Chi shows the statistical significance in the discrimination between groups, a measure of the extent to which a variable makes a unique contribution to a prediction of group membership. Thus, this model presents a high level of correlation between the input variables and the EEG recording classification ($Rc=0.95$), and can discriminate very well between the epilepsy seizure recordings and the non-

seizure ones ($U=0.10$). The additional 10-fold cross-validation test showed an average of 99.30%, 98.61% and 100% for accuracy, selectivity and specificity, respectively. Another model based on almost triple TIs (eight) can be obtained with small differences in the test selectivity and specificity of 99.90% and 99.80% (with the same accuracy).

The quality of the best model for the C1 classification was tested by studying the receiver operating characteristic (ROC) curve (James A. Hanley, 1982) in order to show that the classifier is not random; the training cases against the full set residuals to search for patterns and the robustness of the GDA model to the interferences were also verified. The ROC curve can estimate a different and better threshold for the *a priori* classification probability. Since the area under the ROC curve (1.00) is significantly higher than the area under the random classifier curve (0.5, diagonal line), the C1 model is not random, but a truly statistically significant classifier. The mathematical form of the model has been chosen in absence of any prior information. Figure 3 shows that the training cases against the residuals did not present any characteristic pattern (Dillon and Goldstein, 1984). Due to the robustness of the GDA multivariate statistical techniques, the predictive ability and interference reached by using the proposed model should not be affected (see Figure 4).

Figure 3 comes about here

Figure 4 comes about here

The results shown in Table 2 are typically considered as excellent in the litera-

ture for LDA/GDA-QSAR/QSDR models (Castillo-Garit et al., 2008; Estrada and Molina, 2001; Marrero-Ponce et al., 2004; Morales et al., 2006; Vilar et al., 2008b). The main models based on two classes (C1-C3) provide excellent results with 10-fold test accuracy, selectivity and specificity between 92% and 100%, with only one exception of 88.83% for the test selectivity in the case of the C3 model based on non-embedded TIs. The grouping of all non-seizure EEG recordings (ZONF) versus the seizure set (S) gives rise to the possibility of a more general classification based on embedded TIs with 10-fold test accuracy, specificity and selectivity of 98.20%, 95.24% and 98.95% respectively.

In the case of indirect models, due to the mixture of seizure and non-seizure cases, there is a significant difference between all TI models and the non-embedded and embedded ones (maximum 14%), with 10-fold test accuracy values from 83.93 to 99.73% (see details in Table 2). The set consisting of all TIs has been used to generate three GDA models that can indirectly discriminate between three classes of EEG recordings. In the case of the Z-N-S classification (C4) the following equations can evaluate the probability of EEG recording to be classified as Z, N or S:

- ZN-S classification (C4.1) with test accuracy, selectivity and specificity values of 99.33%, 98.00% and 99.33% respectively:

$$\begin{aligned}
 EpSc = & -51.87 - 62.10*_L Sh_0e + 12.45*_L Tr_3e \\
 & - 46.15*_L Tr_3e - 179.00*_L Se + 54.13*_L Xe \\
 & + 537.41*_L^2 Xe - 280.13*_L^5 Xe + 16.79*_R^4 Xe \\
 N = 300, Rc = 0.99, U = 0.02; Chi = 563.63, p < 0.001
 \end{aligned}
 \tag{17}$$

- NS-Z classification (C4.2) with test accuracy, selectivity and specificity values of 94.67%, 98.00% and 93.00% respectively:

$$EpSc = -2.31 - 12.34*_R H + 0.33*_R S6 - 12.15*_R Je$$

$$N = 300, Rc = 0.84, U = 0.29; Chi = 179.05, p < 0.001' \quad (18)$$

- ZS-N classification (C4.3) with test accuracy, selectivity and specificity values of 96.00%, 92.00% and 98.00% respectively:

$$EpSc = -3.71 + 30.79*_L H + 9.51*_L J + 2.49*_L Tr_3e$$

$$+ 55.15*_L Tr_4e + 69.85*_L S6e - 36.64*_L Xe$$

$$+ 3.22*_R Tr_5e + 9.27*_R Je$$

$$N = 300, Rc = 0.90, U = 0.20; Chi = 233.42, p < 0.001 \quad (19)$$

The 10-fold values for the training, validation (test) and both sets are presented in detail in Table 2. The final evaluation is given by the maximum probability of classification for all three equations as Z, N or S.

Additional calculations have been made in order to study how the Fourier transformation of the EEG recordings has an influence on the case of the star graph based method. Thus, after the transformation of the original EEG signals in frequencies, the SGTIs have been calculated for the logarithmically transformed amplitudes. The best model with the Forward stepwise method is based only on the non-embedded SGTIs and it has accuracy, selectivity and specificity values of 90.00%, 92.45% and 87.23% respectively:

$$EpSc(FT) = -0.18 + 1.82 * Tr_2(FT) - 4.61 * H(FT)$$

$$- 13.81 * W(FT) - 27.19 * S(FT) + 38.37 * J(FT)$$

$$- 0.98 * X(FT)$$

$$N = 200, Rc = 0.82, U = 0.32; Chi = 107.33, p < 0.001 \quad (20)$$

These results are significantly lower than the models without Fourier transformation and demonstrate that there is no need to use it in the case of the star graph method applied to this set of EEG recordings.

The accuracy of our best models that can discriminate between the seizure and non-seizure EEG recordings (Z-S and ZONF-S) is at least equal to the previous ones in the literature and even better than other methods (see Table 3). In the case of the ZO-NF-S composed model, the accuracies are better than the previous ones (Tzallas et al., 2007) for the all the TI models and visibly poor for the models based only on the non-embedded or embedded TIs.

Table 3 comes about here

5 CONCLUSION

This study proposes epilepsy/non-epilepsy seizure classification models for EEG recordings by using the star graph topological indices of the signal sequences. The results prove an excellent predictive ability of using a small number of star graph TIs and the simplest and fastest GDA statistics linear models. In addition, it was demonstrated that for better results it is not necessary to Fourier transform the initial data in frequencies.

This method is completely deterministic, i.e., once the training and test sets have been defined, the model is built from the training set without any random process. Therefore, a training set will always generate always the same model. This is an advantage of other machine learning systems, such as ANNs, because a

validation set is not required, and there is no need to repeat the experiment many times in order to average the results due to the stochastic nature of the system.

This work can be useful in the automatic detection in practice of epilepsy seizures and breaks new ground in applying the same methodology to other signal classification in neurology or other fields.

6 FUTURE WORK

The next directions of graph application on medical spectra are the use of different type of graphs such as the lattice and spiral ones, the improvement of the input data with other spectra derived variables such as the mixed variables and solving other similar problems for spectra diseases such as schizophrenia or Alzheimer. The model presented here uses a Linear Discriminant Analysis method for classification. Following the same line, further work could be done by using different classification methods to test the performance of the features extracted from the signals. For instance, Artificial Neural Networks or Support-Vector Machines could be used to carry out the classification.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors(Chou and Shen, 2009b), we shall make efforts in our future work to provide a web-server for the method presented in this paper.

ACKNOWLEDGMENT

Cristian R. Munteanu acknowledges the Isidro Parga Pondal program of the Xunta de Galicia for the financial support of a research position at the University of A

Coruña (Spain). This work was supported in part by the grant (2007/127 and 2007/144) from the General Direction of Scientific Promotion, Technology and University System of Galicia and grant (PIO52048 and RD07/0067/0005) funded by the Carlos III Health Institute, Spain.

References

- Abarbanel HDI, Brown R, Kennel MB. Lyapunov exponents in chaotic systems: Their importance and their evaluation using observed data. *International Journal of Modern Physics*, 1991; 5: 1347-75.
- Addison PS. *The Illustrated Wavelet Transform Handbook*. Institute of Physics, 2002.
- Althaus IW, Chou JJ, Gonzales AJ, Deibel MR, Chou KC, Kezdy FJ, Romero DL, Palmer JR, Thomas RC. Kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-88204E. *Biochemistry*, 1993a; 32: 6548-54.
- Althaus IW, Chou JJ, Gonzales AJ, Diebel MR, Chou KC, Kezdy FJ, Romero DL, Aristoff PA, Tarpley WG, Reusser F. Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. *Journal of Biological Chemistry*, 1993b; 268: 6119-24.
- Althaus IW, Chou KC, Lemay RJ, Franks KM, Deibel MR, Kezdy FJ, Resnick L, Busso ME, So AG, Downey KM, Romero DL, Thomas RC, Aristoff PA, Tarpley WG, Reusser F. The benzylthio-pyrimidine U-31,355, a potent inhibitor of HIV-1 reverse transcriptase. *Biochemical pharmacology*, 1996; 51: 743-50.
- Althaus IW, Gonzales AJ, Chou JJ, Romero DL, Deibel MR, Chou KC, Kezdy FJ, Resnick L, Busso ME, So AG. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. *Journal of Biological Chemistry*, 1993c; 268: 14875-80.
- Andraos J. Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. *Canadian Journal of Chemistry*, 2008; 86: 342-57.
- Andrzejak RG, Lehnertz K, Mormann F, Rieke C, David P, Elger CE. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state. *Phys Rev E Stat Nonlin Soft Matter Phys*, 2001; 64: 061907.
- Bornholdt S, Schuster HG. *Handbook of Graphs and Complex Networks: From the Genome to the Internet*. WILEY-VCH GmbH & CO. KGa.: Weinheim, 2003.
- Bullmore E, Sporns O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 2009; 10: 186-98.
- Castillo-Garit JA, Marrero-Ponce Y, Torrens F, Garcia-Domenech R, Romero-Zaldivar V. Bond-based 3D-chiral linear indices: Theory and QSAR applications to central chirality codification. *Journal of computational chemistry*, 2008.
- Cruz-Monteagudo M, Gonzalez-Diaz H, Borges F, Dominguez ER, Cordeiro MN. 3D-MEDNEs: an alternative "in silico" technique for chemical research in toxicology. 2. quantitative proteome-toxicity relationships (QPTR) based on mass spectrum spiral entropy. *Chem Res Toxicol*, 2008a; 21: 619-32.
- Cruz-Monteagudo M, Munteanu CR, Borges F, Cordeiro MN, Uriarte E, Chou KC, González-Díaz H. Stochastic molecular descriptors for polymers. 4. Study of complex mixtures with topological indices of mass spectra spiral and star networks: The blood proteome case. *Polymer*, 2008b; doi:10.1016/j.polymer.2008.09.070.
- Chen L, Feng KY, Cai YD, Chou KC, Li HP. Predicting the network of substrate-enzyme-product triads by combining compound similarity and functional domain composition. *BMC Bioinformatics*, 2010a; 11: 293.
- Chen L, Huang T, Shi XH, Cai YD, Chou KC. Analysis of Protein Pathway Networks Using Hybrid Properties. *Molecules*, 2010b; 15: 8177-92.
- Chen YL, Li QZ. Prediction of the subcellular location of apoptosis proteins. *J Theor Biol*, 2007; 245: 775-83.
- Chou KC. Graphic rule for drug metabolism systems. *Current Drug Metabolism*, 2010; 11: 369-78.
- Chou KC. Graphical rules in steady and non-steady enzyme kinetics *Journal of Biological Chemistry*, 1989; 264: 12074-9.
- Chou KC. Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems *Biophysical Chemistry*, 1990; 35: 1-24.
- Chou KC, Forsen S. Graphical rules for enzyme-catalyzed rate laws. *Biochemical Journal*, 1980; 187: 829-35.
- Chou KC, Kezdy FJ, Reusser F. Review: Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. *Analytical Biochemistry*, 1994; 221: 217-30.
- Chou KC, Lin WZ, Xiao X. Wenxiang: a web-server for drawing wenxiang diagrams. *Natural Science*, 2011; 3.
- Chou KC, Liu WM. Graphical rules for non-steady state enzyme kinetics. *Journal of theoretical biology*, 1981; 91: 637-54.
- Chou KC, Shen HB. FoldRate: A web-server for predicting protein folding rates from primary sequence. *The Open Bioinformatics Journal*, 2009a; 3: 31-50.
- Chou KC, Shen HB. Review: recent advances in developing web-servers for predicting protein attributes. *Natural Science*, 2009b; 2: 63-92.
- Chou KC, Zhang CT. Prediction of protein structural classes. *Crit Rev Biochem Mol Biol*, 1995; 30: 275-349.
- Dehmer M, Borgert S, Emmert-Streib F. Entropy bounds for hierarchical molecular networks. *PloS one*, 2008; 3: e3079.
- Dehmer M, Varmuza K, Borgert S, Emmert-Streib F. On entropy-based molecular descriptors: statistical analysis of real and synthetic chemical structures. *J Chem Inf Model*, 2009; 49: 1655-63.
- Diao Y, Li M, Feng Z, Yin J, Pan Y. The community structure of human cellular signaling network. *Journal of theoretical biology*, 2007; 247: 608-15.
- Dillon WR, Goldstein M. *Multivariate analysis: Methods and applications*. Wiley: N. Y., 1984.

- Emmert-Streib F, Dehmer M. Analysis of Microarray Data: A Network-Based Approach. Wiley-VCH, 2009a.
- Emmert-Streib F, Dehmer M. Information Theory and Statistical Learning. Springer, 2009b.
- Estrada E, Molina E. 3D connectivity indices in QSPR/QSAR studies. *Journal of chemical information and computer sciences*, 2001; 41: 791-7.
- Ferino G, Gonzalez-Diaz H, Delogu G, Podda G, Uriarte E. Using spectral moments of spiral networks based on PSA/mass spectra outcomes to derive quantitative proteome-disease relationships (QPDRs) and predicting prostate cancer. *Biochemical & Biophysical Research Communications*, 2008; 372: 320-5.
- González-Díaz H, Agüero-Chapin G, Varona-Santos J, Molina R, de la Riva G, Uriarte E. 2D RNA-QSAR: assigning ACC oxidase family membership with stochastic molecular descriptors; isolation and prediction of a sequence from *Psidium guajava* L. *Bioorg Med Chem Lett*, 2005; 15: 2932-7.
- Gonzalez-Diaz H, Gonzalez-Diaz Y, Santana L, Ubeira FM, Uriarte E. Proteomics, networks and connectivity indices. *Proteomics*, 2008; 8: 750-78.
- González-Díaz H, Vilar S, Santana L, Uriarte E. Medicinal Chemistry and Bioinformatics – Current Trends in Drugs Discovery with Networks Topological Indices. *Current Topics in Medicinal Chemistry*, 2007; 7: 1025-39.
- Guler NF, Ubeylib I, Guler ED, I. G. Recurrent neural networks employing Lyapunov exponents for EEG signals classification. *Expert Systems with Applications*, 2005; 29: 506-14.
- Harary F. *Graph Theory*: MA, 1969.
- Hu LL, Huang T, Cai YD, Chou KC. Prediction of Body Fluids where Proteins are Secreted into Based on Protein Interaction Network. *Plos One*, 2011; 6: e22989.
- Huang T, Chen L, Cai YD, Chou KC. Classification and Analysis of Regulatory Pathways Using Graph Property, Biochemical and Physicochemical Property, and Functional Property. *PloS One*, 2011; 6: e25297.
- James A. Hanley BJM. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, 1982; 143: 29-36.
- Kannathal N, Choo ML, Acharya UR, Sadasivan PK. Entropies for detection of epilepsy in EEG. *Comput Methods Programs Biomed*, 2005; 80: 187-94.
- Koutsofios E, North SC. *Drawing Graphs with dot*. AT&T Bell Laboratories, Murray Hill: NJ, USA, 1993.
- Kowalski RD, Wold S. Pattern recognition in chemistry. In Krishnaiah PR, Kanal LN, editors. *Handbook of Statistic*. North Holland Publishing Company: Amsterdam 1982: 673-97.
- Kuzmic P, Ng KY, Heath TD. Mixtures of tight-binding enzyme inhibitors. Kinetic analysis by a recursive rate equation *Analytical Biochemistry*, 1992; 200 68-73.
- Larter R, Speelman B, Worth RM. A coupled ordinary differential equation lattice model for the simulation of epileptic seizures. *Chaos (Woodbury, N.Y.)*, 1999; 9: 795-804.
- Lawrence GR, Barry DW. Computer network development to achieve resource sharing. *Proceedings of the May 5-7, 1970, spring joint computer conference*, 543-549. ACM New York, NY, USA: Atlantic City, New Jersey.
- Lehnertz K, Mormann F, Kreuz T, Andrzejak RG, Rieke C, David P, Elger CE. Seizure prediction by nonlinear EEG analysis. *IEEE Engineering on Medicine and Biological Magazine*, 2003; 22: 57-63.
- Lin SX, Neet KE. Demonstration of a slow conformational change in liver glucokinase by fluorescence spectroscopy. *Journal of Biological Chemistry*, 1990; 265: 9670-5.
- Litt B, Echaz J. Prediction of epileptic seizures. *Lancet neurology*, 2002; 1: 22-30.
- Marrero-Ponce Y, Diaz HG, Zaldivar VR, Torrens F, Castro EA. 3D-chiral quadratic indices of the 'molecular pseudograph's atom adjacency matrix' and their application to central chirality codification: classification of ACE inhibitors and prediction of sigma-receptor antagonist activities. *Bioorg Med Chem*, 2004; 12: 5331-42.
- Micheloyannis S, Pachou E, Stam CJ, Vourkas M, Erimaki S, Tsirka V. Using graph theoretical analysis of multi channel EEG to evaluate the neural efficiency hypothesis. *Neuroscience letters*, 2006; 402: 273-7.
- Micheloyannis S, Vourkas M, Tsirka V, Karakonstantaki E, Kanatsouli K, Stam CJ. The influence of ageing on complex brain networks: a graph theoretical analysis. *Human brain mapping*, 2009; 30: 200-8.
- Min L, Gui-Qing Z, Tian-Lun C. Effects of Different Connectivity Topologies in Small World Networks on EEG-Like Activities. *Communications in Theoretical Physics*, 2006; 45: 373-8.
- Mohseni HR, Maghsoudi A, Sadati N. Automatic detection of epileptic seizure using Time-Frequency Distributios. *Advances in Medical, Signal and Information Processing (MEDSIP)*, 2006a.
- Mohseni HR, Maghsoudi A, Shamsollahi MB. Seizure detection in EEG signals: a comparison of different approaches. *Conf Proc IEEE Eng Med Biol Soc*, 2006b; Suppl: 6724-7.
- Molina E, Diaz HG, Gonzalez MP, Rodriguez E, Uriarte E. Designing antibacterial compounds through a topological substructural approach. *J Chem Inf Comput Sci*, 2004; 44: 515-21.
- Morales AH, Cabrera Perez MA, Gonzalez MP. A radial-distribution-function approach for predicting rodent carcinogenicity. *Journal of molecular modeling*, 2006; 12: 769-80.
- Munteanu CR, González-Díaz H. S2SNet - Sequence to Star Network, Reg. No. 03 / 2008 / 1338. Santiago de Compostela, Spain, 2008.
- Munteanu CR, Gonzalez-Diaz H, Borges F, de Magalhaes AL. Natural/random protein classification models based on star network topological indices. *J Theor Biol*, 2008; 254: 775-83.
- Munteanu CR, Magalhães AL, Sánchez AD, González-Díaz H. S2SNet: A Tool for Transforming Characters and Numeric Sequences into Star Network Topological Indices in Chemoinformatics, Bioinformatics, Biomedical, and Social-Legal sciences. *Current Bioinformatics*, 2012.

- Munteanu CR, Magalhaes AL, Uriarte E, Gonzalez-Diaz H. Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. *Journal of theoretical biology*, 2009; 257: 303-11.
- Myers D, Palmer G. Microcomputer tools for steady-state enzyme kinetics. *Computer Applied Bioscience*, 1985; 1: 105-10.
- Nigam VP, Graupe D. A neural-network-based detection of epilepsy. *Neurological research*, 2004; 26: 55-60.
- Polat K, Güneş S. Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform. *Applied Mathematics and Computation*, 2007; 187: 1017-26.
- Prado-Prado FJ, González-Díaz H, Martínez de la Vega O, Ubeira FM, Chou KC. Unified QSAR approach to antimicrobials. Part 3: First multi-tasking QSAR model for Input-Coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. *Bioorganic and Medicinal Chemistry*, 2008; 16: 5871-80.
- Qi XQ, Wen J, Qi ZH. New 3D graphical representation of DNA sequence based on dual nucleotides. *Journal of theoretical biology*, 2007; 249: 681-90.
- Randic M, Zupan J, Vikić-Topić D. On representation of proteins by star-like graphs. *J Mol Graph Model*, 2007; 290-305.
- Rappin N, Dunn R. *wxPython in Action*. Manning Publications Co.: Greenwich, CT, 2006.
- Rivero D, Fernandez-Blanco E, Dorado J, Pazos A. Using Recurrent ANNs for the Detection of Epileptic Seizures in EEG Signals. *IEEE Congress of Evolutionary Computation (CEC)*. IEEE Press: New Orleans (USA), 2011.
- Scheuer ML, Wilson SB. Data analysis for continuous EEG monitoring in the ICU: seeing the forest and the trees. *Journal of Clinical Neurophysiology*, 2004; 21: 353-78.
- Srinivasan V, Eswaran C, Sriraam N. Artificial neural network based epileptic detection using time-domain and frequency-domain features. *Journal of medical systems*, 2005; 29: 647-60.
- StatSoft.Inc. *STATISTICA*. 2002.
- Strogatz SH. Exploring complex networks. *Nature*, 2001; 410: 268-76.
- Subasi A. EEG signal classification using wavelet feature extraction and a mixture of expert model. *Expert Systems with Applications*, 2007; 32: 1084-93.
- Subasi A. Epileptic seizure detection using dynamic wavelet network. *Expert Systems with Applications*, 2005; 29: 343-55.
- Subasi A, Gursoy M. EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems with Applications*, 2010; 37: 8659-66.
- Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. Wiley-VCH, 2002.
- Tzallas AT, Tsipouras MG, Fotiadis DI. Automatic seizure detection based on time-frequency analysis and artificial neural networks. *Computational intelligence and neuroscience*, 2007: 80510.
- Van Waterbeemd H. Discriminant Analysis for Activity Prediction. In Van Waterbeemd H, editor. *Chemometric methods in molecular design*. Wiley-VCH: New York, 1995: 265-82.
- Vilar S, Gonzalez-Diaz H, Santana L, Uriarte E. QSAR model for alignment-free prediction of human breast cancer biomarkers based on electrostatic potentials of protein pseudofolding HP-lattice networks. *Journal of computational chemistry*, 2008a.
- Vilar S, Gonzalez-Diaz H, Santana L, Uriarte E. QSAR model for alignment-free prediction of human breast cancer biomarkers based on electrostatic potentials of protein pseudofolding HP-lattice networks. *J Comput Chem*, 2008b.
- Vuckovic A, Radivojevic V, Chen A, Popovic D. Automatic recognition of alertness and drowsiness from EEG by an artificial neural network. *Medical Engineering and Physics*, 2002; 24: 349-60.
- Waterhouse E. New horizons in ambulatory electroencephalography. *IEEE Engineering in Medicine Biology Magazine*, 2003; 22: 74-80.
- Zhang CT, Chou KC. Analysis of codon usage in 1562 E. Coli protein coding sequences. *Journal of Molecular Biology*, 1994; 238: 1-8.
- Zhou GP. The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism. *Journal of theoretical biology*, 2011a; 284: 142-8.
- Zhou GP. n extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. *Biochemical Journal*, 1984; 222: 169-76.
- Zhou GP. The Structural Determinations of the Leucine Zipper Coiled-Coil Domains of the cGMP-Dependent Protein Kinase I alpha and its Interaction with the Myosin Binding Subunit of the Myosin Light Chains Phosphase. *Proteins & Peptide Letters*, 2011b; 18: 966-78.

FIGURE LEGENDS

Figure 1. The flowchart of the proposal method based on star graph topological indices classification models. L and R are the left and right parts of the initial EEG recording.

Figure 2. Example of the S2SNet transformation of a small EEG recording fragment into embedded and non-embedded star graph topological indices by using only seven groups/value ranges.

Figure 3. Training cases against the residuals for the full set (C1 classification, Z vs. S, normal eyes open vs. seizure).

Figure 4. Residuals vs. deleted residuals plot for the GDA model in the case of the classification C1 (Z vs. S, normal eyes open vs. seizure).

TABLE LEGENDS

Table 1. The classes with the EEG segments of the five classification problems.

Table 2. 10-fold performance of the gda-based classification models on training, validation and total sets.

Table 3. The comparison of the results of the actual work and previous ones.