Military Technical College
Kobry El-Kobbah,
Cairo, Egypt

6th International Conference
on Electrical Engineering
ICEENG 2008

A robust approach for improved prediction of E.coli promoter gene sequences: combining feature selection, fuzzy weighted pre-processing and AIRS

By

Kemal Polat* Salih Güneş*

Abstract:

In this paper, a different hybrid approach based on combining Feature Selection, Fuzzy Weighted Pre-processing and Artificial Immune Recognition System is proposed to forecast the E.coli Promoter Gene Sequences, which has promoters in strings that represent nucleotides (one of A, G, T, or C). The proposed approach comprises three stages. In the first stage, the dimensionality of this dataset has been reduced to 4 attributes from 57 attributes by means of feature selection process by C4.5 decision tree rules. In the second stage, fuzzy weighted pre-processing has been used to weight E.coli Promoter Gene Sequences dataset that has 4 attributes in interval of [0,1]. Finally, AIRS classifier, is inspried from immune system, has been run to forecast the E.coli Promoter Gene Sequences. While only the AIRS algorithm obtained 53.85% prediction accuracy on the prediction of E.coli Promoter Gene Sequences using 50-50% training-test split, the proposed method obtained 90.38% prediction accuracy on the same conditions. This success shows that the proposed system is a robust and effective system in the prediction of E.coli Promoter Gene Sequences.

Keywords:

E.coli Promoter Gene Sequences; Prediction; AIRS; Feature Selection; Fuzzy Weighted Pre-processing; Hybrid System

---

* Selcuk University, Electrical and Electronics Engineering Department, 42035, Konya, TURKEY

1. Introduction:

The estimation of E.coli promoter gene sequences is important issue in molecular biology field. Here, we first have explained the promoters in molecular biology. Then a different hybrid approach based on combining Feature Selection, Fuzzy Weighted Pre-processing and Artificial Immune Recognition System has been explained for predicting the E.coli promoter gene sequences.

Promoters are DNA sequences which affect the frequency and location of transcription initiation though interaction with RNA polymerase. Two conserved regions about 35 and 10 base pairs (bp) upstream from the transcription start (-35 and -10 regions, respectively) were identified by comparison of relatively few promoters. More extensive compilations and comparisons of prompters for genes of E.coli and its phage and plasmids supported and extended the concept of a "consensus" promoter sequence: a-35 (TTGACA) and -10 (TATAAT) region separated by 17 bp with transcription initiating at a purine about 7 bp downstream from the '3' end of the -10 region. While the -35 and -10 regions show the greatest conservation across promoters and are also the sites of nearly all mutations which affect transcriptional strength, other bases flanking the -35 and -10 regions, in addition to the start point also occur at greater than random frequencies and sometimes affect promoter activity. In addition, variation in spacing between the -35 and -10 regions plays a role in promoter strength [1].

Promoter compilations and analysis have led to computer programs which predict the location of promoter sequences on the basis of homology either to the consensus sequence or to a reference list of promoters. Such programs are of practical significance in searching new sequences; thus promoter compilations are important beyond proving data regarding promoter structure. However, current compilations are based on sequences aligned by eye in attempts to maximize homology to a consensus sequence. Unfortunately, sequences closer to the consensus sequence may be missed thus weakening the homology between promoters and consequently reducing the predictive power of algorithms. Although promoter elements evidence that pin-point bases which interact with RNA polymerase, such data is unavailable for most genes [1].

In this work, we have proposed a novel hybrid combination based on three parts to predict the E.coli promoter gene sequences. The proposed approach comprises three stages. In the first stage, the dimensionality of E.coli Promoter Gene Sequences dataset has been reduced to 4 attributes from 57 attributes by means of feature selection process by C4.5 decision tree rules. In the second stage, fuzzy weighted pre-processing has been used to weight E.coli Promoter Gene Sequences dataset that has 4 attributes in interval of [0,1]. Finally, artificial immune recognition system (AIRS) classifier algorithm, is inspired from immune system, has been run to forecast the E.coli Promoter Gene Sequences. In order to show the performance of the proposed system, we have used the prediction accuracy, sensitivity and specifity analysis, and confusion matrix. While only

the AIRS algorithm obtained 53.85% prediction accuracy on the prediction of E.coli Promoter Gene Sequences using 50-50% training-test split, the proposed method obtained 90.38% prediction accuracy on the prediction of E.coli Promoter Gene Sequences using 50-50% training-test split.

2. The Proposed Approach:

The proposed method consists of three main parts: (1) Feature selection process, (2) Fuzzy weighted pre-processing and (3) AIRS classifier. In the first stage, the dimensionality of E.coli Promoter Gene Sequences dataset has been reduced to 4 attributes from 57 attributes by means of feature selection process by C4.5 decision tree rules. In the second stage, fuzzy weighted pre-processing has been used to weight E.coli Promoter Gene Sequences dataset that has 4 attributes in interval of [0,1]. Finally, artificial immune recognition system (AIRS) classifier algorithm, is inspried from immune system, has been run to forecast the E.coli Promoter Gene Sequences. We explain the details of the feature selection, preprocessing and classification steps in the following subsections.

2.1 Feature selection process:

The number of features (attributes) and number of instances in the raw dataset can be enormously large. This enormity may cause serious problems to many data mining systems. Feature selection is one of the long existing methods that deal with these problems. Its objective is to select a minimal subset of features according to some reasonable criteria so that the original task can be achieved equally well, if not better. By choosing a minimal subset of features, irrelevant and redundant features are removed according to the criterion. Simpler data can lead to more concise results and their better comprehensibility. Since feature selection can only deal with discrete features (attributes), you need to run Discretion first if there are continuous features in the dataset. FS process with C4.5 decision tree rules cannot directly use a data file with continuous attributes in mining. To solve this problem, a discrimination system has to be used. You might pass your data file with continuous attributes to this system, which will discrete the continuous attributes and output a file that contains the discredited data. Upon return, discredited data can then be used in the mining [2].

Feature selection (FS) was done according to the feature distributions over the E.coli promoter gene sequences dataset. After it was detected for classes to be mixed, it was concluded with the consultations of experts that these features were not critical in classification. So the number of features was reduced to 4 by removing these features with the use of C4.5 decision tree rules. We have used the 6^{th}, 15^{th}, 16^{th}, and 17^{th} attributes in prediction of E.coli promoter gene sequences by means of FS process.

2.2. Fuzzy weighted pre-processing:

In the fuzzy weighted pre-processing, we extract new feature values for each feature. To do this, we first apply two membership functions, known as the input and output membership functions, and we then use weighting procedure to the given data. The membership functions are selected as triangular membership functions as shown in Fig. 1(a) and Fig. 1(b) respectively.

The formation of these membership functions is realized as follows: First, the arithmetic mean values of each feature for corresponding samples are calculated using Eq. (1):

$$m_i = \frac{1}{N}\sum_{k=1}^{N} x_{k,i} \qquad (1)$$

In this expression, $x_{k,i}$ represents the i^{th} feature value of sample x_k, k=1,2,…N.

After the calculation of the sample means for each feature, the input membership function is formed by triangles as shown in Fig 1(a). The supports of these triangles are determined by $m_i/8$, $m_i/4$, $m_i/2$, m_i, $2m_i$, $4m_i$, $8m_i$. The lines of the input membership functions are named as mf1, mf2,…mf8. To form the output membership function, the interval [0,1] is divided into 8 equal parts and the corresponding lines are named as mf1', mf2',…mf8' as shown in Fig 1(b). Note here that these input and output membership functions are formed for each feature so that there will exist different input-output membership function configurations for each feature since the sample means of each feature differs.

After determining the input and output membership functions for each feature, the weighted procedure can start. In this procedure, a feature value $x_{k,i}$ is assumed to be in the x axis of input membership function and y axis values for the points at which this value cuts the input membership functions are determined. For example, if a feature value is between 0 and $m_i/8$, then this point will cut both lines mf1 and mf2. The y values at these intersection points, say y_1 and y_2, are known as membership values (μ) and they will then be used in a fuzzy rule base in the following manner: First, the input membership value, $\mu(i)$, is determined by using the above mentioned intersection points as follows:

$$\mu(i)=\mu_{A\cap B}(x_{k,i}) = \text{MIN}\,(\mu_A(x_{k,i}), \mu_B(x_{k,i})),\ x\in X \qquad (2)$$

In this expression, $\mu_A(x_{k,i})$ and $\mu_B(x_{k,i})$ membership values correspond to the intersection points. The rule base for our system is used as presented in Table 1. After $\mu(i)$ value is determined using Eq. (2) for the $x_{k,i}$ feature value, the output weight value

is determined by using output membership functions and the rules in Table 1. To do this, first, the input membership value, μ(i), is presented to the output membership function to determine the corresponding weighted value of the original feature value. This membership value is then taken as a point in y-axis of the output membership functions and, the intersection points, which are cut by this membership value, are determined. As the output membership functions show there will be more than one intersection points. Which of them will be used is decided by the rules given in Table-1. For example, if the input feature value cuts mf1 and mf2 lines in input membership functions then the output value for this feature will be the arithmetic mean of two points that μ(i) cuts mf1' and mf2' at the output membership functions [3].

2.3 AIRS Classification Algorithm: Classification stage:

AIRS is a resource limited supervised learning algorithm inspired from immune metaphors. In this algorithm, the used immune mechanisms are resource competition, clonal selection, affinity maturation and memory cell formation. The feature vectors presented for training and test are named as Antigens while the system units are called as B cells. Similar B cells are represented with Artificial Recognition Balls (ARBs) and these ARBs compete with each other for a fixed resource number. This provides ARBs, which have higher affinities to the training Antigen to improve. The memory cells formed after the whole training Antigens were presented are used to classify test Antigens. The algorithm is composed of four main stages, which are initialization, memory cell identification and ARB generation, competition for resources and development of a candidate memory cell, and memory cell introduction. We give the details of our algorithm below.

1. Initialization: Create a set of cells called the memory pool (M) and the ARB pool (P) from randomly selected training data.
2. Antigenic Presentation: for each antigenic pattern do:

(a) Clonal Expansion: For each element of M, determine its affinity to the antigenic pattern, which resides in the same class. Select the highest affinity memory cell (mc) and clone mc in proportion to its antigenic affinity to add to the set of ARBs (P).

(b) Affinity Maturation: Mutate each ARB descendant of the highest affinity mc. Place each mutated ARB into P.

(c) Metadynamics of ARBs: Process each ARB using the resource allocation mechanism. This process will result in some ARB death, and ultimately controls the population. Calculate the average stimulation for each ARB, and check for termination condition.

(d) Clonal Expansion and Affinity Maturation: Clone and mutate the randomly selected subset of the ARBs left in P based on their stimulation level.

(e) Cycle: While the average stimulation value of each ARB class group is less than

a given stimulation threshold go to step 2.c.

(f) Metadynamics of Memory Cells: Select the highest affinity ARB of the same class as the antigen from the last antigenic interaction. If the affinity of this ARB with the antigenic pattern is better than that of the previously identified best memory cell mc then add the candidate (mc-candidate) to memory set M. If the affinity of mc and mc-candidate are below the affinity threshold, remove mc from M.

3. Classify: Classify data items using the memory set M. Classification is performed in a k-Nearest Neighbor fashion with a vote being made among the k closest memory cells to the given data item being classified.

These steps are repeated for each training antigen. After training, test data are presented only to memory cells. k-NN algorithm is used to determine the classes in test phase. For more detailed information about AIRS, the reader is referred to [3], [4].

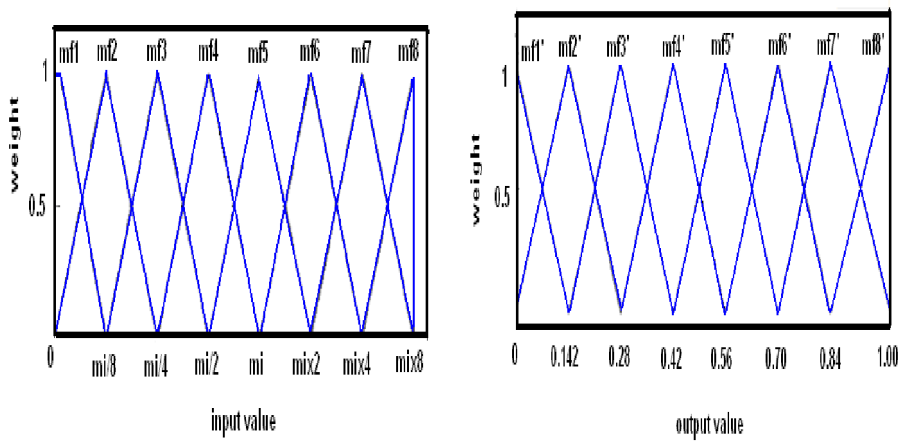

Figure (1): *(a) Input Membership Function, (b) Output Membership Function*

Table (1): *Fuzzy rule base for our system.*

k value	*Rules*
1	**If Input_value cuts mf(k) and mf(k+1) then Output_value=(mf(k)'(y)+mf(k+1)'(y))/2**
2	**if Input_value cuts mf(k) and mf(k+1) then Output_value= mf(k)'(y)+mf(k+1)'(y))/2**
3	**if Input_value cuts mf(k) and mf(k+1) then Output_value= mf(k)'(y)+mf(k+1)'(y))/2**
4	**if Input_value cuts mf(k) and mf(k+1) then Output_value= mf(k)'(y)+mf(k+1)'(y))/2**
5	**if Input_value cuts mf(k) and mf(k+1) then Output_value= mf(k)'(y)+mf(k+1)'(y))/2**
6	**if Input_value cuts mf(k) and mf(k+1) then Output_value= mf(k)'(y)+mf(k+1)'(y))/2**
7	**if Input_value cuts mf(k) and mf(k+1) then Output_value= mf(k)'(y)+mf(k+1)'(y))/2**

3. The Application Results and Discussion:

In this section, we first explain the E.Coli Promoter Gene Sequences Dataset we used in our experiments. Finally, we give the experimental results and discuss our observations from the obtained results.

3.1 E.Coli Promoter Gene Sequences Dataset:

In present study, the real-world task of recognizing biological concepts in DNA sequences has been investigated. In particular, the task is to recognize promoters in strings that represent nucleotides (one of A, G, T, or C). A promoter is genetic region which initiates the first step in the expression of an adjacent gene (transcription) [1].

Table 2 presents the initial domain theory used in the promoter recognition task. The first rule says that the promoter includes two subcategories: a contact and a conformation region. The second rule states that a contact involves two regions, while subsequent rules define alternative ways these regions can appear [1].

Dimensionality of E.coli Promoter Gene Sequences dataset has 57 sequential DNA nucleotides and 106 samples including 53 promoters and 53 non-promoters. A special notation is used to simplify specifying locations in the DNA sequence. The biological literature counts locations relative to the site where transcription begins. Fifty nucleotides before and six following this location constitute an example. When a rule's antecedents refer to input features, they first state the starting location, and then list the sequence that must follow. In these specifications, "x" indicates that any nucleotide will suffice. Hence, the first rule for conformation says that there must an "a" 45 nucleotides before the site where transcription begins. Another "a" must be at -44, then any two nucleotides can appear, and finally there must be a "t" at location -41 [1, 5, 6]. Figure 2 presents the DNA sequences.

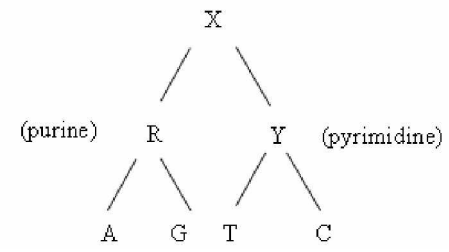

Figure (2): *DNA Nucleotides.*

Table (2): *A Domain Theory for Promoters.*

Promoter	-contact, conformation.
contact	- minus_35, minus_10.
minus_35	-@_37 “cttgac”.
minus_35	-@_36 “ttgxca”.
minus_35	-@_36 “ttgaca”.
minus_35	-@_36 “ttgac”.
minus_10	-@_14 “tataat”.
minus_10	-@_13 “taxaxt”.
minus_10	-@_13 “tataat”.
minus_10	-@_12 “taxxxt”.
conformation	-@_45 “aaxxa”.
conformation	-@_45 “axxxa”, @_4 “t”, @_28 “txxxtxaaxxtx”.
conformation	@_49 “axxxxt”, @_1 “a”, @_27 “txxxxaxxtxtg”.
conformation	@_47 “caaxttxac”, @_22 “gxxxtxc”, @_8 “gcgccxcc”.

3.2 Results and Discussion:

In this study, we have conducted the experiments on the predicting of E.coli Promoter Gene Sequences using a mixed approach based on feature selection process, fuzzy weighted pre-processing, and AIRS classifier. We have used the 50-50% training-test split of all the E.coli Promoter Gene Sequences dataset to test the proposed method. The efficiency of proposed method has been measured using prediction accuracy, sensitivity and specifity analysis, and confusion matrix. On the conducted applications, only the AIRS classifier obtained 53.85% prediction accuracy on the forecasting the E.coli Promoter Gene Sequences using 50-50% training-test split. The combination of AIRS and fuzzy weighted pre-processing obtained 57.59% prediction accuracy. As for proposed approach obtained 90.38% prediction accuracy on the forecasting the E.coli Promoter Gene Sequences using 50-50% training-test split. Also, we have used the C4.5 decision tree classifier to compare the obtained prediction accuracies in applications. Table 3 presents the obtained results including number of attributes, prediction accuracy, sensitivity, and specifity values of methods used.

Table (3): *The obtained prediction accuracies, sensitivity and specifity values for AIRS classifier, AIRS with fuzzy weighted pre.proce, proposed approach, and C4.5 decision tree classifier using 50-50% training-testing split.*

Method	Number of Attributes	Prediction Accuracy (%)	Sensitivity (%)	Specifity (%)
LSSVM Classifier by Polat et al. (2007) [8]	57	65.38	70	62.50
FS-LSSVM Classifier by Polat et al. (2007) [8]	4	84.62	90.90	80
AIRS classifier (2008)	57	53.85	52.00	100
AIRS with fuzzy weighted pre.proce. (2008)	57	57.69	56.09	69.23
C4.5 decision tree classifier (2008)	57	70.08	68.90	75.00
The proposed approach (2008)	***4***	***90.38***	***92***	***88.88***

Table (4): *The confusion matrixes of AIRS classifier, AIRS with fuzzy weighted pre-processing, and proposed approach using 50-50% training-testing split.*

Output/Desired	*Result (promoter)*	*Result (non-promoter)*	Method
Result (promoter)	2	24	*AIRS*
Result (non-promoter)	0	26	
Result (promoter)	9	18	*AIRS with fuzzy weighted pre-processing*
Result (non-promoter)	4	23	

Result (promoter)	24	2	*Proposed Approach*
Result (non-promoter)	3	24	

In this study, there were two prediction classes: promoter and non-promoter. Classification results of the system were displayed by using a confusion matrix. In a confusion matrix, each cell contains the raw number of exemplars classified for the corresponding combination of desired and actual network outputs. Table 4 gives the confusion matrixes showing the classification results of AIRS classifier, AIRS with fuzzy weighted pre-processing, proposed approach, and C4.5 decision tree classifier using 50-50% training-testing split.

From the above results, we conclude that the mixed system combining the Feature selection process, Fuzzy Weighting Pre-Processing and AIRS obtains very promising results in forecasting E.coli Promoter Gene Sequences.

4. Conclusions:

In this paper, a novel mixed approach based on feature selection, fuzzy weighted pre-processing, and AIRS classifier has been applied on the task of predicting E.coli promoter gene sequences and the most accurate learning methods have been evaluated. Experiments have been conducted on the E.coli promoter gene sequences dataset to decide whether any sequence belong to promoters or not. While only the AIRS algorithm obtained 53.85% prediction accuracy on the prediction of E.coli Promoter Gene Sequences using 50-50% training-test split, the proposed method obtained 90.38% prediction accuracy on the prediction of E.coli Promoter Gene Sequences using 50-50% training-test split. This success shows that the proposed system is a robust and effective system in the prediction of E.coli Promoter Gene Sequences. It is hoped that more interesting results will follow on further exploration of data.

5. Acknowledgments:

This study has been supported by Scientific Research Project of Selcuk University.

References:

[1] Harley, C. and Reynolds, R. (1987). *Analysis of E.coli promoters sequences*. Nucl. Acids Res., 15:2343-2361.

[2] Polat, K., Şahan., S, Kodaz, H., and Günes., S., (2005). *A New Classification Method for Breast Cancer Diagnosis: Feature Selection Artificial Immune Recognition System*, Lecture Notes in Computer Science (LNCS), Vol. 3611, 830-838.

[3] Polat K., Şahan S. and Güneş S. *A New Method to Medical Diagnosis: Artificial Immune Recognition System (AIRS) with Fuzzy Weighted Pre-processing and Application to ECG Arrhythmia*. Expert Systems with Applications;2006, 31(2), 264-269.

[4] Watkins, A. and Timmis, J. *Artificial Immune Recognition System (AIRS): Revisions and Refinements*. In Proc. of 1st International Conference on Artificial Immune Systems, pages 173-181, University of Kent at Canterbury, September 2002.

[5] Geoffrey, G.T., Jude, W.S., and Michiel, O.N. (1990). *Refinement of Approximate Domain Theories by Knowledge-Based Neural Networks*, Proc. of the eight national conf. on Artificial Intelligence, 861-866.

[6] UCI Machine Learning Repository, http://www.ics.uci.edu/~mlearn/MLRepository.html (last arrived: 2008).

[7] Confusion Matrix, http://www.gepsoft.com/Gepsoft/APS3KB/Chapter09/Section2/SS03.htm (last arrived: 2008).

[8] Polat, K., and Güneş, S., *A novel approach to estimation of E. coli promoter gene sequences: Combining feature selection and least square support vector machine (FS_LSSVM)*, Applied Mathematics and Computation, 2007, 190(2), 1574-1582