Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 36 (2009) 57–64
Expert Systems with Applications
www.elsevier.com/locate/eswa

A new method to forecast of *Escherichia coli* promoter gene sequences: Integrating feature selection and Fuzzy-AIRS classifier system

Kemal Polat *, Salih Güneş

Selcuk University, Department of Electrical and Electronics Engineering, 42075 Konya, Turkey

Abstract

We have investigated the real-world task of recognizing biological concepts in DNA sequences in this work. Recognizing promoters in strings that represent nucleotides (one of A, G, T, or C) has been performed using a novel approach based on feature selection (FS) and Artificial Immune Recognition System (AIRS) with Fuzzy resource allocation mechanism (Fuzzy-AIRS), which is first proposed by us. The aim of this study is to improve the prediction accuracy of *Escherichia coli* promoter gene sequences using a novel system based on FS and Fuzzy-AIRS. The *E. coli* promoter gene sequences dataset has 57 attributes and 106 samples including 53 promoters and 53 non-promoters. The proposed system consists of two parts. Firstly, we have reduced the dimension of *E. coli* promoter gene sequences dataset from 57 attributes to 4 attributes by means of FS process. Second, Fuzzy-AIRS classifier algorithm has been run to predict the *E. coli* promoter gene sequences. The robustness of the proposed method is examined using prediction accuracy, sensitivity and specificity analysis, *k*-fold cross-validation method and confusion matrix. Whilst only Fuzzy-AIRS classifier has obtained 50% prediction accuracy using 10-fold cross-validation, the proposed system has obtained 90% prediction accuracy in the same conditions. These obtained results have indicated that the proposed system obtain the success rate in recognizing promoters in strings that represent nucleotides.

Keywords: *Escherichia coli* promoter gene sequences; Feature selection; Artificial immune system; AIRS classification system; Fuzzy resource allocation mechanism; 10-fold cross-validation; Prediction

1. Introduction

The estimation of *Escherichia coli* promoter gene sequences is an important issue in the field of molecular biology. Herein, we have first explained the promoters in molecular biology. Then a novel approach based on feature selection process and least Fuzzy-AIRS classifier has been explained.

Promoters are DNA sequences which affect the frequency and the 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 (Harley & Reynolds, 1987).

* Corresponding author. Tel.: +90 332 223 2056; fax: +90 332 241 0635.
E-mail addresses: kpolat@selcuk.edu.tr (K. Polat), sgunes@selcuk.edu.tr (S. Güneş).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.09.010

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 pin-point bases which interact with RNA polymerase, such data is unavailable for most genes (Harley & Reynolds, 1987).

In this study, we have proposed a novel approach based on two stages to predict the *E. coli* promoter gene sequences. First of all, we have reduced the dimensionality of *E. coli* promoter gene sequences dataset from 57 attributes to 4 attributes by means of FS process. Second, Fuzzy-AIRS classifier algorithm has been run to predict the *E. coli* promoter gene sequences. We obtained 90% success rate from the experiments made on the *E. coli* promoter gene sequences dataset taken from UCI (University of California Institute) machine learning database (UCI Machine Learning Repository, 2006) using 10-fold cross-validation. Also, we have used the 50–50% training-test split, 70–30% training-test split, and 80–20% training-test split from all *E. coli* promoter gene sequences dataset to test the proposed system. Using the proposed system, the obtained prediction accuracies are 86.54%, 87.50%, and 100% on the 50–50% training-test split, 70–30% training-test split, and 80–20% training-test split, respectively. Without feature selection (FS), Fuzzy-AIRS obtained 55.57%, 56.25%, and 50% prediction accuracies on the 50–50% training-test split, 70–30% training-test split, and 80–20% training-test split, respectively. As it can be seen in the above results, the proposed system has obtained better results than those of the Fuzzy-AIRS classifier.

2. Materials and methods

In this section, we have explained the *E. coli* promoter gene sequences dataset used and the proposed method in the following subsections.

2.1. E. coli promoter gene sequences dataset

In the 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 a genetic region which initiates the first step in the expression of an adjacent gene (transcription) (Geoffrey, Jude, & Michiel, 1990).

Table 1
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"

Table 1 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 (Geoffrey et al., 1990).

Dimensionality of *E. coli* promoter gene sequences dataset has 57 attributes and 106 samples including 53 promoters and 53 non-promoters. The input features are 57 sequential DNA nucleotides. 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 the transcription begins. Another "a" must be at −44, then any two nucleotides can appear, and finally there must be a "t" at location −41 (Geoffrey et al., 1990; UCI Machine Learning Repository, 2006). Fig. 1 presents the DNA sequences.

2.2. The proposed system

The proposed system consists of two main parts: Feature Selection and Fuzzy-AIRS classifier. We apply the former to the given *E. coli* promoter gene sequences dataset to reduce the dimensionality of this dataset. By means of this

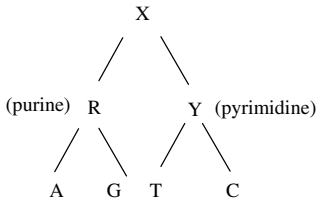

Fig. 1. DNA nucleotides.