Expert Systems with Applications 38 (2011) 12398–12408

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

PolyA-iEP: A data mining method for the effective prediction of polyadenylation sites

George Tzanis *, Ioannis Kavakiotis, Ioannis Vlahavas

Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

ARTICLE INFO

Keywords:
Data mining
Machine learning
Classification
Emerging pattern
Bioinformatics
Polyadenylation

ABSTRACT

This paper presents a study on polyadenylation site prediction, which is a very important problem in bioinformatics and medicine, promising to give a lot of answers especially in cancer research. We describe a method, called PolyA-iEP, that we developed for predicting polyadenylation sites and we present a systematic study of the problem of recognizing mRNA 3′ ends which contain a polyadenylation site using the proposed method. PolyA-iEP is a modular system consisting of two main components that both contribute substantially to the descriptive and predictive potential of the system. In specific, PolyA-iEP exploits the advantages of emerging patterns, namely high understandability and discriminating power and the strength of a distance-based scoring method that we propose. The extracted emerging patterns may span across many elements around the polyadenylation site and can provide novel and interesting biological insights. The outputs of these two components are finally combined by a classifier in a highly effective framework, which in our setup reaches 93.7% of sensitivity and 88.2% of specificity. PolyA-iEP can be parameterized and used for both descriptive and predictive analysis. We have experimented with Arabidopsis thaliana sequences for evaluating our method and we have drawn important conclusions.

1. Introduction

During the last decades two main scientific areas, namely biology and computer science have been characterized by major advances that have attracted the interest of all humanity. The growth of World Wide Web and the completion of Human Genome Project are two representative examples that reflect the extent of the development of these two scientific areas. However, biology and computer science have not grown separately. The need of the collaboration between biologists and computer scientists has been grown year by year as the two areas have been progressing and new scientific questions have been arising. Bioinformatics is a novel research area that has emerged as a solution to the aforementioned need. It is a very promising field that aims to provide the means to analyze and explain the vast amounts of biological data, contributing thereby to the development of other related areas like medicine.

Two relative subfields of computer science strongly related to artificial intelligence, namely data mining and machine learning, have provided biologists, as well as experts from other areas, a powerful set of tools to analyze new data types in order to extract various types of knowledge efficiently and effectively. These tools combine powerful techniques of artificial intelligence, statistics, mathematics, and database technology. This fusion of technologies aims to overcome the obstacles and constraints posed by the traditional statistical methods. A lot of interesting applications of artificial intelligence in bioinformatics is presented in Ezziane (2006).

In this paper we deal with polyadenylation site (or poly(A) site) prediction. Poly(A) site prediction is a challenging problem and the last years has attracted the attention of the scientific community, because the successful cure of this problem promises to provide a lot of answers in various fields of medicine, like cancer research. In many organisms, such as in Arabidopsis thaliana, which is a plant model organism, there are not many highly conserved signals or patterns around the poly(A) site and consequently the recognition of the poly(A) site is not trivial. The discrimination of mRNA 3′ ends that contain a poly(A) site from intronic or 5′ UTR sequences without a poly(A) site seems to be very difficult (mainly with intronic sequences) and the performance of the up to now proposed approaches is moderate. On the other hand, mRNA 3′ ends can be easily discriminated from coding sequences. This variability in the difficulty of discrimination has motivated our work and guided us to an effort to study this problem and define an approach that can improve prediction accuracy. Nowadays, the research in this field is focused on discovering new patterns around poly(A) site and on predicting the poly(A) site accurately. The method we propose can be used for both, pattern discovery and accurate prediction.

* Corresponding author. Tel.: +30 2310998433.
E-mail addresses: gtzanis@csd.auth.gr (G. Tzanis), ikavak@csd.auth.gr (I. Kavakiotis), vlahavas@csd.auth.gr (I. Vlahavas).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.04.019

The prediction of poly(A) sites can be divided into two sub-problems. The first sub-problem deals with the discrimination of the sequences that contain a poly(A) site from the ones that do not and the second deals with the prediction of the position of a poly(A) site inside a sequence. The advantage of this approach is double. Firstly, a large number of irrelevant sequences are filtered out before searching for the position of a poly(A) site inside a sequence increasing notably the prediction accuracy. Secondly, a more specific method for predicting the position of a poly(A) site inside a sequence that focuses only in sequences that contain a poly(A) site leading in better models can be used. This approach can provide an increased performance against a more general method that deals concurrently with the discrimination of sequences and the prediction of poly(A) sites inside a sequence. The first sub-problem of the approach described above has not been studied yet. In this paper we focus on this sub-problem.

Our contribution is an approach that combines the concept of emerging patterns (Dong & Li, 1999) and more specifically the interesting ones with a novel distance based scoring method. Our approach maintains the high interpretability of emerging patterns and offers a high prediction performance. The extracted emerging patterns may span across many elements around the polyadenylation site and can provide novel and interesting biological insights. Our method increases significantly the performance of poly(A) site prediction and reaches 93.7% of sensitivity and 88.2% of specificity. Moreover, The method we propose can be parameterized and re-trained in order to deal with poly(A) site prediction in any organism. Beyond the proposed method we draw important conclusions on the problem of discriminating mRNA 3′ ends with poly(A) sites from other sequences without a poly(A) site.

This paper is organized as follows. Section 2 provides the necessary background knowledge. Section 3 presents a concise review of the research area that is related to the problem dealt in this study. Section 4 provides some preliminary technical terminology and Section 5 is dedicated to the detailed description of our approach. The results of the experiments that were conducted in order to evaluate our method are presented in Section 6 and finally, the paper is concluded in Section 7.

2. Background knowledge

Two families of molecules are responsible for the structure and functioning of every living organism, as well as for the carriage of the genetic information. These are proteins and nucleic acids, which both are linear polymers of smaller molecules (monomers). The term "sequence" is used to refer to the order of monomers in a polymer. A sequence is represented as a string of different symbols, one for each monomer. There are 20 protein monomers called amino acids and five nucleic acid monomers called nucleotides. A nucleotide is characterized by the nitrogenous base it contains: adenine (A), cytosine (C), guanine (G), thymine (T), or uracil (U). The most common nucleic acids are *deoxyribonucleic acid* (*DNA*) and *ribonucleic acid* (*RNA*). DNA may contain a combination of A, C, G, and T. In RNA, U appears instead of T.

DNA contains the genetic instructions used in the development and functioning of all known living organisms and some viruses. The processes related with DNA are described by the central dogma of molecular biology, which deals with the detailed residue-by-residue transfer of sequential information (Fig. 1). It states

Fig. 1. The central dogma of molecular biology.

that information cannot be transferred back from protein to either protein or nucleic acid (Crick, 1970).

DNA replication, the basis for biological inheritance, is a fundamental process occurring in all living organisms to copy their DNA. *Transcription* is the process by which the information contained in a section of DNA is transferred to a newly assembled piece of *messenger RNA* (*mRNA*). In contrast, *reverse transcription* is the transfer of information from RNA to DNA (the reverse of normal transcription). This is known to occur in the case of retroviruses, such as HIV that causes acquired immunodeficiency syndrome (AIDS). *RNA replication* is the copying of one RNA to another. Many viruses replicate this way. Finally, *translation* is the production of proteins by decoding mRNA produced in transcription.

The process of *polyadenylation* occurs after transcription termination. It involves cleavage of the new transcript (mRNA), followed by template-independent addition of adenines at its newly synthesized 3′ end. The cleavage site is called *polyadenylation site* (*poly(A) site*). Polyadenylation is considered to be part of the larger process of producing mature mRNA for translation. The aim of the polyadenylation process is to protect the mRNA in order to reach intact the protein synthesis site.

The most important factors that are involved in the process of polyadenylation are the cis-regulatory elements and the trans-acting factors. The cis-regulatory elements are RNA sequences consisting of 2–10 nucleotides and their role is to help the trans-action factors define the poly(A) site. The most prominent cis-element is the hexamer AAUAAA or a close variant. This hexamer is located 10–35 nucleotides upstream of the cleavage site (poly(A)-site) and it can be found in about 50% of human genes (Hu, Lutz, Wilusz, & Tian, 2005) but only in 10% of Arabidopsis genes (Loke et al., 2005). The trans-acting factors are a protein complex which also includes a specificity factor (Cleavage and Polyadenylation Specificity Factor – CPSF), an endonuclease, and poly(A) Polymerase (PAP). The trans-acting factors are responsible for the cleavage at the appropriate site (poly(A) site) and the addition of the about 200 adenine residues (poly(A) tail) to the 3′ end (Lewin, 2004).

Nowadays, the research in this field is focused on discovering new cis-regulatory elements and on predicting the poly(A) site accurately. The accurate prediction of poly(A) site is a crucial step to define gene boundaries and get an insight in transcription termination in eukaryotes, which is a process less well understood.

3. Related work

An early approach to the problem of poly(A) site prediction was the work of Salamov and Solovyev (1997) who developed a software called POLYAH and an algorithm for the identification of 3′-processing sites of human mRNA precursors. The algorithm was based on a linear discriminant function (LDF) trained to discriminate real poly(A) signals from the other regions of human genes possessing the AATAAA sequence which is most likely non-functional. The accuracy of the method has been estimated on a set of 131 poly(A) regions and 1466 regions of human genes having the AATAAA sequence. When the threshold was set to predict 86% of poly(A) regions correctly, specificity of 51% and correlation coefficient of 0.62 had been achieved.

In 1999 Tabaska and Zhang developed polyadq, a program for detection of human polyadenylation signals. The program finds poly(A) signals using two discriminant functions: one specific for AATAAA type poly(A) sites and the other for ATTAAA type poly(A) sites. Polyadq predicts poly(A) signals with a correlation coefficient of 0.413 on whole genes and 0.512 in the last two exons of genes.