ELSEVIER

Knowledge-Based Systems 15 (2002) 217–223

Knowledge-Based SYSTEMS

www.elsevier.com/locate/knosys

A data mining approach based on machine learning techniques to classify biological sequences

M. Maddouri[a], M. Elloumi[b,*]

[a]Computer Science Department, National Institute of Applied Sciences and Technologies, Tunis-Carthage 2035 Tunis, Tunisia
[b]Computer Science Department, Faculty of Economic Sciences and Management of Tunis, El Manar 2092 Tunis, Tunisia

Received 15 March 2001; revised 16 April 2001; accepted 31 May 2001

Abstract

In molecular biology, biological macromolecules, like desoxyribonucleic acids (DNA) and proteins are coded by strings, called ‘primary structures’. For a long time, biologists gathered these primary structures in large databases. Now, they focus on analyzing these primary structures in order to extract useful knowledge. Data mining approaches can be helpful to reach this goal. In this paper, we present a data mining approach based on machine learning techniques to do classification of biological sequences. By using our approach, we use four steps as follows. (1) In the first step, we construct the set of the discriminant substrings, called discriminant descriptor (DD), associated with each family of primary structures. This construction is made thanks to an adaptation of the Karp, Miller and Rosenberg (KMR) algorithm. (2) In the second step, we use the DDs constructed during the first step to code the families of primary structures by a table of examples vs attributes, called ‘context’. (3) In the third step, we extract knowledge from the context constructed during the second step and represent it by production rules. This extraction is made by using an incremental production rules approach. (4) Finally, during the last step, we use the obtained production rules to do classification of primary structures. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Classification; Data mining; Machine learning; Production rules; Biological sequences; Discriminant substrings; Primary structures

1. Introduction

Over the last decades, biologists have sequenced a large number of biological macromolecules [1–12]. Mining biological data can help biologists to do classification in biological sequences. Indeed,

1. for proteins, mining data representing proteins can help biologists in the classification of a newly sequenced protein. The classification of a newly sequenced protein in a known family of proteins can be helpful in learning information, not only about the evolution of this protein, but also about its biological functions [13–16].
2. for desoxyribonucleic acids (DNA), mining data representing functional sites and non-functional sites [17], or data representing protein coding regions and protein non-coding regions [18,19], of DNA macromolecules can help biologists in the classification of a candidate site in the set of functional sites or the one of non-functional sites, or the classification of a candidate region in the set of protein coding regions or the one of protein non-coding regions. And hence, can help them to make a step towards gene recognition in DNA macromolecules [20–24].

Statistical approaches [20,25–35] used to do classification of biological sequences do not give explanations concerning the classification done. Whereas, machine learning approaches [18,36–39] give such explanations and, hence, can be helpful to improve biologists know-how and knowledge.

In this paper, we present a data mining approach based on machine learning techniques to do classification of biological sequences. Our approach uses four steps.

1. In the first step, we construct the set of the discriminant substrings, called ‘discriminant descriptor’ (DD) [40,41], associated with each family of primary structures. This construction is made thanks to an adaptation [40] of the Karp, Miller and Rosenberg (KMR) algorithm [42].
2. In the second step, we code the families of primary structures by a table of examples vs attributes. Indeed, in machine learning, data are coded by sets of attributes describing sets of examples, i.e. a table of examples vs attributes called ‘context’. On the other hand, in molecular

* Corresponding author. Tel.: +216-1233253; fax: +216-1712093.
E-mail address: Mourad.Elloumi@fsegt.rnu.tn (M. Elloumi).

0950-7051/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0950-7051(01)00143-5

biology, biological macromolecules are coded by strings. We build the context by representing each element of the DD by a binary attribute.

3. In the third step, we extract knowledge from the context and represent it by production rules, by using an incremental production rules approach [43,44].
4. Finally, during the last step, we use the obtained production rules to do classification of primary structures.

In Section 2 of this paper, we present some definitions and notations. In Section 3, we present our algorithm of construct DDs. In Section 4, we show how we build a context. In Section 5, we present our incremental production rules approach, to do classification of primary structures. In Section 6, we give an illustrative example. In Section 7, we present experimental results. Finally, in the last section, we present our conclusions and perspectives.

2. Definitions and notations

Let $\mathscr{A}$ be a finite alphabet, a string is a concatenation of elements of $\mathscr{A}$. The length of a string w, denoted by $|w|$, is the number of the characters that constitute this string. A portion of a string w, let us call it x, that begins at the position i and ends at the position j, $1 \leq i \leq j \leq n$, is called 'substring' of w, and we denote $x \subseteq w$.

Let $f_1, f_2, \ldots, f_n$be families of strings and x be a substring of a string of f_i, $1 \leq i \leq n$. The substring x is discriminant between f_i, $1 \leq i \leq n$, and $\cup_{j \neq i} f_j$ if, and only if, it appears in strings of f_i but does not appear in any string of $\cup_{j \neq i} f_j$. A discriminant substring x is minimum if, and only if, it contains no other discriminant substring. The set of discriminant and minimum substrings between f_i, $1 \leq i \leq n$, and $\cup_{j \neq i} f_j$ is called 'discriminant descriptor' (DD) and will be denoted by $\Delta_i(f_1,f_2,\ldots,f_n)$. The subset of $\Delta_i(f_1,f_2,\ldots,f_n)$ made-up by the discriminant and minimum substrings of length k is called 'k-discriminant descriptor' (k-DD) and will be denoted by k-$\Delta_i(f_1,f_2,\ldots,f_n)$. We have then:

$$\Delta_i(f_1,f_2,\ldots,f_n) = \bigcup_{k>0} k - \Delta_i(f_1,f_2,\ldots,f_n) \tag{1}$$

A substring x is ambiguous if, and only if, it appears both in strings of f_i, $1 \leq i \leq n$, and in strings of $\cup_{j \neq i} f_j$.

Let $f_1,f_2,\ldots,f_n$be families of strings and $\Delta_1,\Delta_2,\ldots,\Delta_n$be their respective DDs:

A context [45] is a triplet (F,Δ,R), where $F = \cup_{i=1}^{n} f_i$, $\Delta = \cup_{i=1}^{n} \Delta_i$ and R is a binary relation defined between F and Δ such that $(w,s) \in R$, $w \in F$ and $s \in \Delta$, when the string w contains the substring s.

Let A and B be two subsets, $A \subseteq F$ and $B \subseteq \Delta$, a Galois connection is a couple of operators (ϑ,ψ) such that $\vartheta(A) = \{s \in \Delta | (w,s) \in R$ for all $w \in A\}$ and $\psi(B) = \{w \in F | (w,s) \in R$ for all $s \in B\}$.

A concept of a context (F,Δ,R) is a maximum rectangle, it is a couple (A,B) with $A \in F$, $B \in \Delta$, $A = \vartheta(B)$ and $B = \psi(A)$. A is called the 'extend' and B is called the 'intend' of the concept (A,B).

A concept is pertinent, if and only if, it minimizes the Shannon entropy. The Shannon entropy of a concept $C_i = (A_i,B_i)$ is defined by:

$$h(C_i) = -\sum_{k=1}^{n} \frac{n_i^k}{n_i} * \log\left(\frac{n_i^k}{n_i}\right) \tag{2}$$

where n is the number of the families, n_i is the number of the strings of A_i and n_i^k, $1 \leq k \leq n$, is the number of the strings of A_i that belong to the family f_k.

The coverage of a context (F,Δ,R) [46] is defined as a set of concepts $CV = \{C_1,C_2,\ldots,C_n\}$, such that, each element of R is contained in at least one concept of CV. A coverage $CV = \{C_1,C_2,\ldots,C_n\}$ of a relation R is pertinent if, and only if, it is made-up by pertinent concepts.

3. Construction of DDs

Let $f_1,f_2,\ldots,f_n$ be families of strings, the construction of a DD between f_i, $1 \leq i \leq n$, and $\cup_{j \neq i} f_j$ is made by using an adaptation [40] of the Karp, Miller and Rosenberg (KMR) algorithm [42]: we concentrate, respectively, the strings of $f_1,f_2,\ldots,f_n$ into a single string t then, at each step, we filter the vector representing the repeats in t, such that, we get the substrings of t, of equal lengths, that appear in the i-th portion of t, i.e. in strings of f_i, but do not appear elsewhere, i.e. in any string of $\cup_{j \neq i} f_j$. The filtering of a vector V_k consists in assigning the value 0 to any component $V_{k,i}$ that corresponds to the position of the first character in t of an occurrence of a discriminant and minimum substring of length k. The positive components of this vector correspond then to the ambiguous substrings of length k. These substrings can generate longer discriminant and minimum substrings. The filtering is made after each construction of a vector V_k from a vector V_{k-1}. This filtering enables us to avoid that a discriminant and minimum substring generates other substrings that are discriminant but not minimum. The construction of the different DDs is done simultaneously.

4. Construction of a context

Let $f_1,f_2,\ldots,f_n$be families of strings and $\Delta_1,\Delta_2,\ldots,\Delta_n$be their respective DDs. The construction of the context (F,Δ,R), $F = \cup_{i=1}^{n} f_i$ and $\Delta = \cup_{i=1}^{n} \Delta_i$, is done by constructing a matrix R, of size $|F| \times |\Delta|$, such that $R[i,j] = 1$ if the i-th string of F contains the j-th substring of Δ else $R[i,j] = 0$.

5. Induction of rules

Within a context, each concept $C = (A,B)$ can be assigned a family label, since the strings constituting A belong to families. In general, the strings of A belong to different families, although they can contain the same discriminant