Expert Systems with Applications 39 (2012) 2330–2334

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Data mining for grammatical inference with bioinformatics criteria

Vivian F. López *, Ramiro Aguilar, Luis Alonso, María N. Moreno

Departament Informática y Automática, University of Salamanca, Plaza de la Merced S/N, 37008 Salamanca, Spain

ARTICLE INFO

Keywords:
Grammatical inference
Bioinformatic
Free Context Grammar
DNA
Sequential patterns

ABSTRACT

In this work a novel data mining process is described that combines hybrid techniques of association analysis and classical sequentiation algorithms of genomics, to generate grammatical structures of a specific language. Subsequently, these structures are converted to Context-Free Grammars. Initially the method applies to context-free languages with the possibility of being applied to other languages: structured programming, the language of the book of life expressed in the genome and proteome and even the natural languages. We used an application of a compilers generator system that allows the development of a practical application within the area of *grammarware*, where the concepts of the language analysis are applied to other disciplines, like bioinformatic. The tool allows measuring the complexity of the obtained grammar automatically from textual data.

1. Introduction

During recent years many approaches were introduced as data mining methods for pattern recognition in biological database. Bioinformatics employs computational and data processing technologies to develop methods, strategies and programs that permit to handle, order and study the immense quantity of biological data that have been generated and are currently generated.

To this aim, the computational linguistics has received considerable attention in bioinformatics. The study in Searls et al. (1999) indicated that a relation exists between formal languages theory and DNA. Being the linguistic view of DNA sequences, a rich source of ideas to model strings with correlated symbols. Most of the work (Jiménez-Montaño, 2009; Jiménez-Montaño, Feistel, & Diez-Martínez, 2010) has involved examinations of the occurrences of "words" in DNA. Searls and Dong (1993) found that such a linguistic approach proves useful not only in theoretical characterization of certain structural phenomena in sequences, but also in generalized pattern recognition in this domain, via parsing. The information represented on sequences involves grammatical inference for pattern recognition.

In this work a novel data mining process is described that combines hybrid techniques of association analysis and classical sequentiation algorithms of genomics to generate grammatical structures of a specific language. Subsequently, these structures are converted to Context-Free Grammars (CFG). Initially the method applies to context-free languages with the possibility of being applied to other languages: structured programming, the language of the book of life expressed in the genome and proteome and even the natural languages.

We used an application of the compiler generator so named GAS 1.0 system (López, Sánchez, Alonso, & Moreno, 2009), that represents an Integrated Development Environment (IDE) which allow the development of a practical application within the area for the automatic generation of language-based tools, that starts from the traditional solutions and facilitates the use of formal language theory in other disciplines: Grammar-Based Systems (GBSs) (Mernik, Crepinsek, Kosar, Rebernak, & Zumer, 2004). The tool allows measuring the complexity of the obtained grammar automatically from textual data.

2. Context-Free Grammar

A grammar G is defined like $G = (N,T,P,S)$, where N is the set of nonterminals symbols, T is the set of terminals symbols, P is the set of production rules and S is the initial symbol. The language of a grammar $L(G)$ is the set of all terminal strings w that have derivations from the initial symbol. This is: $L(G) = \{w \text{ is in } T^* | S \Rightarrow {}^*w\}$.

A CFG has production rules like $A \to \alpha$ where $A \in N$ and $\alpha \in (N \cup T)^*$. The substitution of A by α is carried out independently of the place in which appear A (Louden, 1997). The majority of the programming languages are generated by grammars of this type (enlarged with some contextual elements necessary for the language semantics).

2.1. Grammars and bioinformatics

The association analysis involves techniques that are different in its operations but all of them search relations among the attributes of a data set. Some techniques are:

* Corresponding author.
E-mail address: vivian@usal.es (V.F. López).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.08.058

- Association rules
- Discovery of sequential patterns (DSP), and
- Discovery of associations.

The association rules (AR) describe the relations of certain attributes with regard to others attributes in a database (DB). These rules identify cause-effect implications between the different attributes of the DB. Similar to the AR, the discovery of associations (DA) tries to find implications between different couples attribute-value so that the appearance of these determine a present association in a good quantity of the registers of the DB.

Discovery of sequential patterns (DSP) is very similar to the AR but search for patterns between transactions so that the presence of a set of items precede another set of items in a DB during a period of time. For example, if the data correspond to registers of articles purchased by clients, a description of what articles buys frequently a client can be obtained, and above all, which is the sequence of its purchase. Thus, the next time, the profile of the client would be known, and it will be able to predict the sequence of its purchase. This criteria can apply to another data control, for example, in the bioinformatics context, when the data to treat correspond to the chain of nucleotides of the genome and sequences are discovered as the patterns that codify genes conform some protein (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996).

Bioinformatics employs hybrid techniques to handle, order and study the immense quantity of biological data that have been generated and are currently generated. For example, for the human genome (HG), the bioinformatics seeks to find meaning to the language of the more than 37.000 million peers *A*, *C*, *T* and *G* that have been compiled and stored in the *book of life*.

They offer us the opportunity to understand the gigantic DB that contain the details of the circumstances of time and place in which the genes are activated, the conformation of the proteins that specify, the form in which they influence some proteins on others and the role that such influences can play in the diseases. Besides, what are the relations of the HG with the genomes of the model organisms, like the fly of the fruit, the mice and the bacteria? Will it be able to discover sequential patterns that show how are related between itself the fragments of information? and will it be able to conform a grammatical structure that show the interpretation of the resultant set? If we are able to infer that structure for this type of language we will contribute to understand the real function of the structure of the DNA and we will understand slightly more than the questions presented. One of the applications of the bioinformatics is the pharmacology, offering reviving solutions to the old model for the creation of new medicines. It is worth to note that, one of the more elementary bioinformatics operations consists of the search of resemblances between a fragment of DNA recently arranged and the already available segments of diverse organisms (remember and associate this with the DSP). The finding of approximate alignments permits to predict the type of protein that will specify such sequence. This not only provides trails on pharmacological designs in the initial phases of the development of medicines, but suppresses some that will constitute un resolving

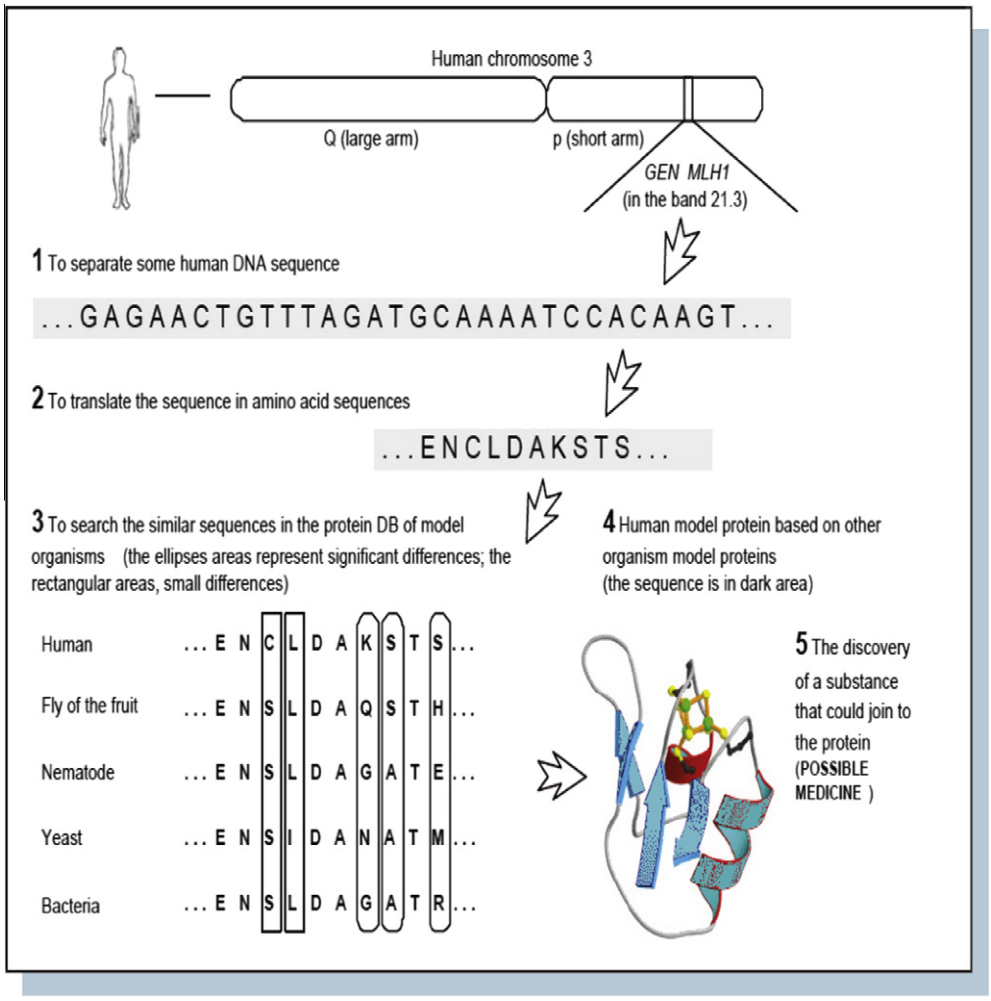

Fig. 1. Using of the bioinformatics in the pharmacology.