

Grammatical inference applied to linguistic modeling of biological regulation networks¹

DOI: 10.3395/reciis.v1i2.Sup.104en



*Elias
Bareinboim*

Programa de Engenharia
de Sistemas e Computação,
COPPE, UFRJ / Laboratório
de Bioinformática, LNCC,
Rio de Janeiro, Brazil
eliasb@ufrj.br



*Ana Tereza R.
Vasconcelos*

Laboratório de Bioinformática,
LNCC, Rio de Janeiro,
Brazil
atrv@lncc.br

João C. P. da Silva

Departamento de Ciência da Computação, Instituto de Matemática, UFRJ, Rio de Janeiro, Brazil
jcps@dcc.ufrj.br

Abstract

We present a methodology based on grammatical inference algorithms applied to the linguistic modeling of biological regulation networks. The linguistic approach to the problem of regulation networks was proposed by COLLADO-VIDES, who proved and formalized the need for use of context sensitive languages to represent such networks. The learning of context sensitive languages is a difficult task, our proposed methodology describes this class from language with a simpler nature that can be learned by already consolidated grammars inference algorithms. In addition to the proposed methodology, we suggest promising directions for this research.

Keywords

Gene regulation, linguistic modeling, context sensitive language, augmented regular expression, grammatical inference

The goal of this work is to present a proposal for refining the linguistic approach defined by Collado-Vides (COLLADO - VIDES, 1989; 1991; 1992; 1993a; 1993b) for modeling of genetic regulation networks using algorithms of grammatical inference. After an extensive survey on the promoters sigma 70 of bacteria E. Coli (COLLADO - VIDES, 1993c), Collado-Vides formulated a grammar that generated a language containing all the regulatory sequences known until then.

The resulting grammar satisfied some properties that were established as relevant (COLLADO-VIDES, 1993b; 1993c). For example, the nucleotides individually, in pairs or in triplets should not be used as the smallest unit of information subject to manipulation in the system, or should not compose the language alphabet. For this role, three types of categories were considered relevant: the promoters (Pr), the operators (Op) and the activators (I).

Since these categories were established, COLLADO-VIDES (1993b; 1993c) tried to select simple and relevant properties that better characterized each category, and allowing the definition of the grammar. The considered properties were: (i) the existence of a proximal site (within the range of -60 to +20 nucleotides) related with a promoter, (ii) obedience to the *proximal precedence principle*, implicating that the proximal operators are represented on the right of the promoter, and proximal activators are represented on the left of the promoter, (iii) obedience to the *positional precedence principle* establishing that, given the sites A and B, whose position on the DNA strand is, respectively, c_1 and c_2 , and if $c_1 < c_2$ then we say that A precedes B, (iv) other sites that are not classified as proximal are considered optional and remote sites, (v) identification through an attribute of which proteins can bind to proximal or remote sites; (vi) the identification of two types of coordinates: the *c coordinates* to explain the absolute distances of sites to the promoter, and the *d coordinate* to define the relative distance among remote sites and the homologous proximal sites.

The language generated according to such grammar belongs to a class of languages denoted as context sensitive languages (COLLADO-VIDES, 1991). This class is one of the most complex classes from a computational perspective considering the Chomsky hierarchy (CHOMSKY, 1959; ULLMAN et al., 2001), being above of context-free languages classes and the regular languages classes, the latter being the simplest in the hierarchy.

In his work, COLLADO-VIDES manually generated the grammar from examples of regulatory sequences and some of the properties identified as representative. Our goal is to generate such grammar automatically applying machine learning algorithms to the range of examples of regulatory sequences.

Initially, we search the literature for machine learning algorithms applied to the class of languages which were context sensitive. In general, the learning of context sensitive language is an arduous task. This fact basically results from the vast complexity of this class of languages, from a theoretical point of view as well as computational. Based on the technical difficulty of working with such languages, we seek alternative ways to address this problem.

ALQUÉZAR et al. (1995; 1997) proposed an approach to learn context sensitive language through approximation by regular languages. That task is done by initially learning a simpler language (regular), and then refining itself, adding restrictions and generating an associated context sensitive language. The class of context sensitive languages is not complete, but goes beyond the languages considered as trivial (FU, 1982).

The augmented regular expressions (AREs) (ALQUÉZAR et al., 1995; 1997) are used to describe, to recognize and to learn classes of context sensitive languages. They combine the descriptive power of regular expressions, used to denote regular language, with a set of linear restrictions that correlates the symbols of the expression.

In short, the handling of AREs is done in two stages, namely, learning and recognition. The learning method that will induce ARE can be described by three basic steps:

Step 1. Inference of regular grammar

Input: Set of positive and negative examples.

Output: Finite automaton that recognizes a regular language which contains all the positive examples and none of the negative ones.

Step 2. Translation

Input automaton obtained in the previous step.

Output: Regular expression equivalent to the entry automaton.

Step 3. Process of induction of the restrictions

Input: Regular expression obtained in the previous step.

Output: Context sensitive language that approximate the desired target set, accepting the positive examples and rejecting the negative examples.

The recognition method can be performed in two steps:

Step 1. Preliminary examination of the expression, which evaluates if, regardless of the restrictions, the expression is consistent in relation to the original regular expression (considering an expression following the step three of the learning process).

Step 2. Verification of restrictions, in which, after evaluation of the expression and the success in its confirmation, one must determine if the set of restrictions are not being violated.

Given the above, our proposal is to adapt the methodology of AREs to the problem of refining the linguistic representation of regulatory expressions proposals by (COLLADO-VIDES, 1991; 1992; ROSENBLUETH, 1996). Ultimately, we hope to get as a result a set with the largest possible number of regulatory sequences, and a new set of candidates that can be proved experimentally.

The main steps of the proposed methodology can be described as follows:

- algorithm A_0 : It does the automatic translation of the information (the transcription factor, promoter, starting and final positions of the sequence, type and central position) present in the database RegulonDB (HUERTA et al., 1998) for Prolog clauses (BRATKO, 2000) that will be used to calculate the potential candidates to regulatory sequences.

The best candidate sequences are selected (ROSENBLUETH, 1996; HERTZ, 1999), and these will compose the set of positive examples. This set is used as input to the algorithm for grammatical inference A_1 ;

- algorithm A_1 : Performs the grammatical inference. It uses an algorithm of grammatical inference for regular languages, generating a finite automaton. There is no restriction on such original algorithm, and it is possible

to use any algorithm for grammatical inference already available in the literature. If the generated automaton is not deterministic, the algorithm A_2 is used. Otherwise, the algorithm A_3 is used;

- algorithm A_2 : Transforms a non-deterministic finite automaton in a deterministic finite automaton (ULLMAN et al., 2001);

- algorithm A_3 : Transforms a deterministic finite automaton in the corresponding regular expression (ULLMAN et al., 2001);

- algorithm A_4 : It builds a set of linear restrictions from the regular expression generated by the algorithm A_3 and the set of examples generated by the algorithm A_1 , (ALQUÉZAR et al., 1995).

The resulting expression of the algorithm A_4 together with the regular expression of A_3 characterizes the desired context sensitive language and is called Augmented Regular Expression (ARE).

We observed that for a proper functioning of our proposal, we should make an adjustment in the representation of regulatory sequences. Unlike the method proposed by COLLADO-VIDES (1993b), the distance should be represented as a new category and not as a property of the existing categories, allowing the distances to be expressed through the restrictions generated by the ARE.

All the above described algorithms have been implemented. We are using the K-tail (FU, 1982) procedure for the learning algorithm for regular languages A_1 . Currently we are in the process of testing and the results seem promising. Thus, the phase of result analysis can be started.

After obtaining conclusive results, the development of the work should move in the following directions, namely, using other algorithms for grammar learning, proposing a non-linear ARE, using classification tools based on context, using probabilistic grammars and other proposals.

Specifically, we are researching for other algorithms for regular language grammar learning to replace the K-tail, and test them. The K-tail algorithm allows the restriction of the regular language to a certain limit based on a parameter k , which is the size of the tail of a word. These restrictions also comprise a fairly significant amount of words that we would like to reduce. The search for other learning algorithms aims to circumvent this type of restriction.

The AREs approach currently uses as restrictions only linear functions, which compel some rigidity to the treatment of the distances between the categories (Pr, Op, I). To bypass this limitation, we assign probability distributions to the distances among the categories and then model a structure of dependency between the linear constraints variables. Based on this modeling, we get the probability distributions, possibly conditional, for the distance between the proximal sites.

In the area of text classification, the work of COHEN et al. (1996) and FREUND et al. (1997) present two interesting algorithms that consider the issue of

context, called Sleeping Experts and RIPPER. Both algorithms are for classification and are not generative (do not create new words).

These algorithms are attractive for problems of text classification of large scale, being efficient and robust, tolerant to noise and with performance in linear time or almost linear. Both algorithms allow the context of a word to influence how the classification of the sentence as a whole will be made. This classification is handled internally by linear and non-linear classifiers, respectively.

We study the possibility to use these algorithms in two situations: (i) replacing the entire framework of AREs, (ii) use of a combined form when it is concatenated with the ARE, classifying the words that are generated by ARE according to the positive and negative examples used in the learning process.

As for probabilistic grammars, in a typical construction of a deterministic grammar, it generates various rules of transition, and some of them are applied many times while others only a few in isolated examples. If we remove the rules that are less used, the grammar would not be complete. This distinction between rules that are very or little used is not included in the description of a typical grammar.

The probabilistic grammars are defined as grammars that have probabilities associated with each rule (STERGOS et al., 2001). The result is that when the syntactic generator returns a new word, it associates the probability of this word being generated. This probability can later be used to determine which, among all the words generated, is more likely to be obtained. With this, we can ignore peculiar words that are unlikely to be achieved in reality. Another possibility is to use Hidden Markov Models (HMMs), a specific type of automaton, with probabilities associated with their states.

We can pinpoint some other proposals:

- to compare the performance between the approaches: ARE, Collado-Vides, text categorization, not linear ARE, probabilistic grammars;

- to use implicit knowledge acquired with the bacterium *E. Coli* through the implementation of this methodology to infer regulatory sequences in other phylogenetically similar organisms (knowledge transfer);

- to use an approximation through context-free grammars based on the statistical model of syntactic analyzer (Collins et al., 1997) rather than regular, to then bring the target grammar. This approach is also used in textual classification problems.

We can consider the proposed methodology as a good starting point for a new framework to a better understanding of the regulation in biological networks based on the information exposed in this work and in more detail in BAREINBOIM (2005). Moreover, such theoretical advances in conjunction with the detailed proposals of action and how to develop such work are very challenging, interesting as subject for future study.

Notes

1. This project was funded through support of specialized multi-tasks projects of the Laboratory of Bioinformatics (LABINFO) of the National Laboratory for Scientific Computation / Ministry of Science and Technology [#506321/2004-5], National Council of Technological and Scientific Development (CNPq) and the state of Rio de Janeiro research Agency (Faperj).

Bibliographic references

ALQUÉZAR, R.; SANFELIU, A. **Augmented Regular Expressions: A formalism to describe, recognize and learn a class of context-sensitive languages**. Relatório Técnico (12-06) – Politécnica de Cataluna, 1995. Available at: <<http://www.lsi.upc.es/dept/techreps/techreps.html>>. Accessed: 11 Oct. de 2007.

ALQUÉZAR, R.; SANFELIU, A. Recognition and learning of a class of context-sensitive language described by augmented regular expressions. **Pattern Recognition**, v.30, p.163-182, 1997.

ANNUAL ACM Symposium on Theory of Computing, El Paso, Texas, USA. Proceedings of the Twenty Ninth Annual ACM Symposium on Theory of Computing, 1997, p.334-343.

BAREINBOIM, E. **Artificial intelligence techniques applied to the biological regulation networks problem**. 2005. 75p. Monograph (Undergraduate in Computer Science) - Computer Science Department, Mathematics Institute, Federal University of Rio de Janeiro.

BRATKO, I. **Prolog: Programming for Artificial Intelligence**. 3rd ed. Addison Wesley, 2000.

CHOMSKY, N. On certain formal properties of grammars. **Information and Control**, v.2, p. 91-112, 1959.

COHEN, W.; SINGER Y. Context-sensitive learning methods for text categorization. In: SIGIR-96. ACM International Conference on Research and Development in Information Retrieval, 19, 1996, Zurich, **Proceedings**. New York, USA: ACM, 1996, p.307-315.

COLLADO-VIDES, J. A Transformational-Grammar Approach to the Study of The Regulation of Gene Expression. **Journal of Theoretical Biology**, v.136, p.403-425, 1989.

COLLADO-VIDES J. The Search of a Grammatical Theory of Gene Regulation is Formally Justified by Showing the Inadequacy of Context-free Grammars.

Computer Applications in the Bioscience (*CABIOS*) v.7, p.321-326, 1991.

COLLADO-VIDES, J. Grammatical model of the regulation of gene expression. Proceedings of National Academy of Science, USA, v.89, p.9405-9409, 1992.

COLLADO-VIDES, J. The Elements for a Classification of Units of Genetic Information with a Combinatorial Component. **Journal of Theoretical Biology**, v.163, p.527-548, 1993.

COLLADO-VIDES, J. A Linguistic Representation of the Regulation of Transcription Initiation: I. An Ordered Array of Complex Symbols with Distinctive Features. **Biosystems**, v.29, p.87-104, 1993.

COLLADO-VIDES, J. A Linguistic Representation of the Regulation of Transcription Initiation: II. Distinctive Features of Sigma 70 Promoters and their Regulatory Binding Sites. **Biosystems**, v.29, p.105-128, 1993.

COLLINS, M. Three Generative, Lexicalised Models for Statistical Parsing. In: Annual Meeting of the Association for Computational Linguistics, 35. Conference of the European Chapter of the Association for Computational Linguistics, 8, 1997, Madrid. **Proceedings**. Madrid. 1997, p.16-23. Available at: <www.aclweb.org/anthology/P97-1003.pdf>. Accessed: 11 Oct. 2007.

FREUND, Y. et al. **Using and Combining Predictors That Specialize**.


FU, K.S. **Syntactic pattern recognition and applications**. New Jersey: Prentice Hall, 1982.

HERTZ, G.Z.; STORMO, G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. **Bioinformatics**, v.15, n.7-8, p.563-77, 1999.

HUERTA, A.M. et al. RegulonDB: A Database on Transcription Regulation in Escherichia coli. **Nucleic Acids Research**, v.26, p.55-60, 1998.

ROSENBLUETH, D.A. et al. Syntactic recognition of regulatory regions in Escherichia coli. **Computer Applications in the Bioscience**, v.12, p.415-422, 1996.

STERGOS, A. D. **On Grammars – The Chomsky Hierarchy and Probabilistic Grammars**. Available at: <<http://citeseer.ist.psu.edu/443342.html>>. Accessed: 11 Oct. 2007.

ULLMAN, J. D.; HOPCROFT, R. **Introduction to Automata Theory, Languages, and Computation**. Boston: Addison-Wesley, 2001. 

About the authors

Elias Bareinboim

Holds a degree in Computer Sciences from the Instituto de Matemática/ Universidade Federal do Rio de Janeiro (UFRJ) and a Masters Degree in Systems and Computer Engineering from COPPE/Universidade Federal do Rio de Janeiro (UFRJ). His masters thesis deals with modeling of theoretical properties, and computer simulations of complex systems, with main results in the characterization and interpretation of an important characteristic of complex networks. He has experience in computer sciences, with emphasis in modeling of complex systems and artificial intelligence. Currently he collaborates with the Laboratory for Bioinformatics of the Laboratório Nacional de Computação Científica (LNCC) and the Program “Engenharia de Sistemas e Computação” at COPPE/UFRJ.

Ana Tereza Ribeiro de Vasconcelos

Graduated in Biological Sciences at the Universidade do Estado do Rio de Janeiro (1983), and obtained a masters degree in Biological Sciences (Biophysics) at the Universidade Federal do Rio de Janeiro (1995) and a Ph.D. degree in Biological Sciences (Genetics) from the Universidade Federal do Rio de Janeiro (2000). Currently she is a researcher at the Laboratório Nacional de Computação Científica, with experience in the fields of genetics and emphasis in Bioinformatics and research activities mainly in the following themes: genomics, bioinformatics, annotation of bacterial genomes. She has also been president of the Associação Brasileira de bioinformática e Biologia Computacional - AB3C.