Jornada de Seguimiento de Proyectos, 2010
Programa Nacional de Tecnologías Informáticas

# Inferencia Gramatical: Técnicas y Aplicación al Procesamiento de Biosecuencias TIN2007-60769

Pedro García, Damián López, José M. Sempere, Manuel Vazquez de Parga,
Antonio Cano, José Ruiz and Tomás Pérez
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia

Gloria I. Álvarez
Departamento de Ciencias e Ingeniería de la Computación
Pontificia Universidad Javeriana (Cali, Colombia)

**Abstract**

This project proposal is related to the field of Artificial Intelligence. It considers automatic learning as a way to solve some problems that consider the prediction or analysis of biosequences (that is, coded sequences of DNA, RNA or proteins). The automatic learning techniques to be studied in this proposal consider the Inductive Inference of formal languages paradigm. This field is known as Grammatical Inference (GI). The main goal of this project proposal is to develop new GI algorithms for the learning of well-known language classes: regular, context free and context sensitive languages. The correctness of the algorithms to be developed within this project will be proved formally.

Other goal of this proposal is the solution of some bioinformatic tasks. In a more detailed way, we will apply GI techniques to the prediction and analysis of some characteristic features in biosequences, as functional domain prediction in protein sequences or detection of coding sequences in DNA sequences. The technology transference will be supported by the design and implementation of a set of multiplatform tools and of an interactive web site able to run the tools developed.
**Keywords**: Grammatical Inference; Automata Theory; Protein functional domain prediction.

## 1 Project Description

This project framework is Automatic Learning as a technique able to provide solution to some bioinformatic tasks related with the biosequences analysis. These tasks include, but are not limited to, the prediction of functional domains in protein sequences, or coding sequences in DNA sequences.

The aim of this project is the development of new Grammatical Inference (GI) algorithms for some families of formal languages (regular, context free or context sensitive) of Chomsky's

hierarchy. The theoretical efficiency of the developed algorithms have been proved both ways, formally as well as in practice, using synthetic testing grounds.

GI is a discipline that lies in between automata and formal language theory on the one hand, and pattern recognition and artificial intelligence on the other [1, 2, 3, 4]. This approach assumes that the objects can be represented using words over a certain alphabet and the class a certain object belongs to is established by determining if it belongs to a certain language. This language is represented by a grammar, a finite automaton or any other abstract device.

GI has been successfully applied in tasks where data are sequential in nature, for example in pattern recognition, speech or handwritten recognition tasks (for example [5, 6, 7, 8]). Some other problems where objects can be modelled in a hierarchical way are also candidates to be approached using these techniques [9]. Besides, an increasing number of works relate biological information with GI, for instance to propose new inference algorithms [10, 11]. Furthermore, previous work by the research team proved that a GI approach is useful to solve bio-processing problems [12].

As an advantage respect other techniques used to solve these tasks, it is worth to be noted that: GI techniques need less heuristics than other approaches; they can also be easily generalized (they are a more versatile approach to solve a priory non related tasks); and, that they can be easily formalized, allowing in that way as easy analysis of their complexity and correctness.

In this applied line of work, the project goals included the study of the behaviour of GI techniques when applied to bioinformatic problems. As a result of this study, another goal was the design and development of multi-platform tools, as well as a web server, that will be available to other scientific communities. These tools together with the remote server had to provide the technological transference of the project.

## 1.1 Project Goals

The main hypothesis that lead the research group to apply for the project were twofold. First, the broad experience of the team in the field of GI, as well as in the application of a GI approach to tackle applied problems, mainly in the pattern recognition field (handwritten digit recognition or automatic classification of documents). Thus, some members of the group had participated in several national projects that applied these techniques to applied problems; and second, the good previous results obtained in the design of GI-based solutions to bioprocessing tasks.

The goals the research team intended to achieve were:

- The design of new inference algorithms using positive presentation and also using NFAs as a way to represent the hypothesis.
- To propose new inference algorithms for subclasses of context free languages.
- The design of new inference algorithms for context sensitive languages using reductions.
- The application of GI techniques to biological information processing, mainly to be used in the prediction of functional domains in amino acid sequences.
- The development of multiplatform “stand-alone” tools and of a interactive web to support the above results.

These goals were scheduled into six tasks that are enumerated below together with the working plan for each one. Table 1 shows the proposed initial timing.

**Design of new GI algorithms:** In order to tackle the development of new learning algorithms, the project proposal considered to take advantage from finite semi-group theory results. Mainly, the framework of the inference of families of languages of the form V*LI [13] from positive presentation.

The identification of NFAs is a recent goal with growing interest in the GI community, under various approaches: Residual Finite State Automata (RFSAs) [14, 15, 16]; unambiguous Finite State Automata (UFAs) [17]; and the inference of regular languages by unrestricted NFAs [18]. The research team proposed to investigate the influence of different orders in the merging of states in the maximal automata of positive data, using the concept of Universal Automaton [19, 20, 21] as the theoretical result that should eventually lead to prove the convergence of those algorithms.

The most widely used corpus is due to Denis, Lemay and Terlute. Nevertheless, the experiments carried out show that this corpus neither represent in a proper way the NFAs, nor are sufficient to conclude anything from the experimentation. Thus, it was proposed to build a proper data corpus in order to compare both the algorithms already proposed and the ones to develop along this project as well.

**Context free and context sensitive language inference methods:** The growing interest on the inference of context-free languages is mainly because, among others, it has applications on the design of natural language processors, the design of compilers or the processing of biological sequences (concerning secondary and tertiary structure mainly).

The research group included the development of new GI algorithms for subclasses of the context-free language class. These algorithms were then to be used for the inference of context sensitive languages taking into account the reduction proposed in [22].

**GI based solutions to bioprocessing tasks:** Sequential nature of the DNA molecule and primary structure of proteins allow to use Language Theory theoretical results. Previous results of this research group show the viability of this approach. Thus, the detection and location of coiled-coil domains in amino acid sequences was successfully undertook by GI techniques [23, 24, 12].

The group proposed to tackle the prediction of other functional domains, as the transmembrane domain. To do this, the TMPDB y TMHMM datasets were to be taken into account. Concerning the gene prediction task, the group considered the dataset by Burset y Guigó [25]. The initial approach proposed consider inference of even linear languages approach and then a transformation of that language into a Mealy machine. This machine is then used to translate the test sequences.

Once the prediction tool were designed and tested, the project goals included the development of software able to run and easily use the solutions obtained. This goal allows to achieve the transference of technology. The implemented tools were to be easily exportable, and therefore capable to be run in whichever more extended platforms (Windows, Linux, etc.). The tools should allow to analyse, process, represent and show properly some features of biological sequences. The development of software to build a stand-alone platform and a web server was proposed.

| Task/activities | Leader and research team | First year | Second year | Third year |
|---|---|---|---|---|
| Design of GI methods from positive presentation | P.G., J.R., J.M.S. | xxxxxxxxxxxx | xxxxxxxxxxxx | xxxxxx |
| Design of new NFA inference methods | P.G., M.V., J.R., G.I.A. | xxxxxxxxxxxx | xxxxxxxxxxxx | |
| Experimental evaluation of NFA inference methods | G.I.A., M.V., Contr. 1 | xxxxxx | xxxxxxxxxxxx | xxxxxx |
| Context-free and sensitive language inference | J.M.S., D.L., T.P., A.C. | xxxxxxxxxxxx | xxxxxxxxxxxx | xxxxxx |
| GI-based solutions to bioprocessing tasks | D.L., A.C., T.P., Contr. 2 | xxxxxx | xxxxxxxxxxxx | xxxxxx |
| Tools and web design and implementation | T.P., Contr. 1, D.L., Contr. 2 | | xxxxxx | xxxxxxxxxxxx |

Table 1: Project tasks schedule. Task coordinator is underlined. P.G. stands for Pedro García; J.M.S. for J.M. Sempere; M.V. for M. Vázquez de Parga; A.C. for A. Cano; D.L. for D. López; J.R. for J. Ruiz; T.P. for T. Pérez and G.I.A. for G. Álvarez.

# 2 Summary of achievements

The research team has obtained good results in all the goals proposed. For the sake of brevity, let us to organize them into three sections: The first one summarizes all the results related with regular languages, including new grammatical inference algorithms and the characterization of subclasses of regular languages . The second section includes the results obtained on non-regular languages and the research carried out on several computational models; the third section explains how these results have been applied to protein functional domain prediction; finally, the fourth section summarizes the formation activity of the team.

## 2.1 Results on regular languages

The design of non-deterministic finite automata (NFA) identification algorithms was proposed as one of the goals of the project. In this line of work it is worth to emphasize the results included in the PhD thesis by Gloria I. Álvarez and Manuel Vazquez de Parga [26, 27]. On the one hand, Dr. Álvarez present several results on the identifiability of NFAS. On the other hand, Dr. Vázquez de Parga presents a family of algorithms that, on the one hand, improve the experimental behaviour of previous classic algorithms, and on the other hand, also improve the time complexity of those previous work.

Within this line of work, it is also to stress that the research team has proved that, when an inference algorithm by state-merging is considered, the merging order of states does not threaten the convergence of the algorithm [28]. This allows to use expert information to select those states to merge. The same result also allows to run the algorithm several times to, afterwards, select the best (under some criterion, for instance, the smaller) automaton obtained. This approach has been compared with the state-of-the-art regular language GI algorithms and proved competitive [29].

The research team has also obtained results on the efficient construction of quasi-reversible (non-deterministic) finite automata that characterizes the class of reversible languages [30]. The natural (non-trivial) extension of this work deals with the class of locally reversible languages. The team has tackled this problem [31], has characterized and algebraically studied the class of languages (the class is a positive variety of languages). This permitted the team to extend the previous model of locally $k$-reversible automata and to design an efficient algorithm to obtain, whichever locally $k$-reversible language is considered, a locally $k$-reversible automaton.

Another goal of the project was the design of new GI algorithms using positive presentation. In this sense, the family of commutative languages has been studied [32]. Despite the team prove that this class is not identifiable from positive sample, the research team propose an identification algorithm from complete sample. The experimentation carried out proves that the algorithm obtains high performance when compared with other classic algorithms.

In [33] the group studies the iterated superposition operation. The first result proves that this DNA-based operation is proved to be closed for the class of regular languages. Then, an edit-distance measure is defined using the defined operation. This edit-distance measure allows to use previous approaches to GI that successfully applied to pattern recognition or multiple ADN sequence alignment [34].

## 2.2 Results on context-free and context-sensitive languages

Concerning the proposed goals on non-regular languages, the team has studied two computational models: membrane systems (P-systems) and Watson-Crick automata (WKA). Some results have been obtained that: propose GI algorithms to identify context-sensitive languages from structural samples [35]; the learning of these language classes using the membrane paradigm has also been studied [36, 37]; and, the class of reversible WKA (languages) has been defined [38].

When context-free languages are taken into account, the universality of the Netwoks of Evolutionary Processors (NEP) model, allows the group to consider it as the base of study. In [39] constructive proofs are provided to show the equivalence between NEPs and NEPs with filtered connections. The paper also presents a new (easier than previous ones) construction technique. In [40] the classification power of NEPPs (Pictoric NEPs) with filtered connections is proven. This model can be used to tackle the inference of context-free languages using structural information. Another result obtained by the group [41] defines a new computational model of evolutionary P-systems and show how to simulate NEPs using this variation of P-systems.

In this section it is also worth to be noted the study on the inference of graph languages [42]. This work presents the first characterization of a class of graph languages that does not take into account the features of a generating machine but the structure of the graphs in the language. On the one hand, this result permits to use more powerful structural information to learn non-regular languages. On the other hand allows to model objects using

graph primitives in applied tasks without restrictions.

## 2.3 GI-based solution to protein functional domains prediction

The applied goals of the project were focused on the development of tools able to predict functional domains in protein sequences. The team has considered the transmembrane domain

due to the importance of such proteins playing the role of transporters or receptors in the cell.

The GI approach has proved to be suitable for the task [43]. The experimentation carried out has shown that GI techniques obtain competitive results when compared with other approaches to bioinformatic tasks. A software package that implements the tool is available in order to build a stand-alone platform. A web server available to the community has also been developed [44].

Another software package developed by the research team is a tool to simulate evolutionary processors [45]. This tool is one of the results of a PhD thesis coached by one group member. Another related result study the application of sticker systems to practical tasks [46].

### 2.4 Formation activities

From this research work, two PhD thesis have been defended. Both of them include original and competitive new GI algorithms on the identifiability of NFA. Besides, two predoctoral students have defended their Master Thesis on results directly related with the main lines of work of this project. Besides, a new predoctoral student has joined the group from Dr. Sakakibara group at Keio University. This student is also with Biotechvana SL, and it has allowed to start a collaboration that aims to consider GI techniques to process phylogenetic information.

## 3 Conclusions

We present in this section a summary of the results in order to show that achievement of the main goals proposed in the project.

The research team has published four articles in indexed journals [28, 30, 31, 43], and many papers in International Conferences some of them selected to a special issue on the Conference topics. Another goal proposed was the development of biosequence processing tools. In that line of work, a web server that implements the prediction tool developed is available to the community at the web page of the research group. The software package needed to build a stand alone tool is also available to be downloaded.

During the development of this tool, the group has initiated a collaboration with the Structural Genomics Group at Instituto de Investigación Príncipe Felipe. This collaboration has provided this team with a new, widely-annotated, dataset of biological information. This team has also initiated a collaboration with Dr. Calera-Rubio at the Universidad de Alicante in order to study new algorithms to infer multidimensional languages.

Biotechvana SL is a spin-off enterprise of the Universitat de València which is developing a toolkit to process phylogenetic information. Preliminary contact shows the common interest of integrate GI techniques within that toolkit.

The research team collaborates with Dr. Victor Mitrana as well. This researcher is an international reference in natural models of computation. Some results obtained can be used for the inference of non-regular languages.

The formation activities of the research team include two PhD Thesis [26, 27], as well as two Master Thesis, defended on subjects closely related with this project. Nowadays, three PhD students are with this group in several stages of their PhD work.

# References


[1] K.S. Fu and T.L. Booth. Grammatical inference: Introduction and survey - Part I. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(3):343–375, 1975. Part I and II.

[2] D. Angluin and C.H. Smith. Inductive inference:Theory and Methods. *Computing Surveys*, 15(3):237–269, 1983.

[3] L. Miclet. *Syntactic and Structural Pattern Recognition. Theory and Applications*, volume 7 of *Series in Computer Science*, chapter Grammatical inference, pages 237–290. World Scientific, 1990.

[4] Y. Sakakibara. Recent advances of grammatical inference. *Theoretical Computer Science*, 185(1):15–45, 1997.

[5] E Vidal, H Rulot, JM Valiente, and G Andreu. Application of the error-correcting grammatical inference algorithm (ecgi) to planar shape. In *Grammatical inference: theory, applications and alternatives*, volume IEE. Digest No: 1993/092, 1993.

[6] A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen. Pattern discovery in biosequences. *Lecture Notes in Artificial Intelligence*, 1433:257–270, 1998.

[7] I. Torres and A. Varona. $k$-tss language models in speech recognition systems. *Computer Speech and Language*, 15:127–149, 2001.

[8] P. Cruz-Alcazar, E. Vidal, and J.C. Perez-Cortes. Musical style identification using grammatical inference: The encoding problem. *Lecture Notes on Computer Science*, 2905:375–382, 2003.

[9] I. Perea and D. López. Syntactic modelling and recognition of document images. *Lecture Notes on Computer Science*, 3138:416–424, 2004.

[10] F. Coste and G. Kerbellec. A similar fragments merging approach to learn automata on proteins. In *Proc. of the ECML*, 2005.

[11] F. Coste and G. Kerbellec. Learning automata on protein sequences. In *Proc. of the JOBIM*, 2006.

[12] P. Peris, D. López, M. Campos, and J. M. Sempere. Protein motif prediction by grammatical inference. *Lecture Notes in Artificial Intelligence*, 4201:175–187, 2006. 8th International Colloquium, ICGI-06.

[13] P. García and J. Ruiz. Learning in varieties of the form v*li from positive data. *Theoretical Computer Science*, 362(1-3):100–114, 2006.

[14] F. Denis, A. Lemay, and A. Terlutte. Learning regular languages using rfsa. *Theoretical Computer Science*, 313(2):267–294, 2004.

[15] G. Álvarez, J. Ruiz, and P. García. Nondeterministic regular positive negative inference (nrpni). In *Proc. of XXXI Conferencia Latinoamericana de Informática*, pages 239–249, 2005.

[16] P. García G. Álvarez and J. Ruiz. A merging state algorithm for inference of rfsas. *Lecture Notes in Artificial Intelligence*, 4201:340–341, 2006. 8th International Colloquium, ICGI-06.

[17] F. Coste and D. Fredouille. Unambiguous automata inference by means of state-merging methods. In *Proc of the ECML*, 2003.

[18] M. Vazquez de Parga, P. García, and J. Ruiz. A family of algorithms for non-deterministic regular language inference. *Lecture Notes on Computer Science*, 4094:265–275, 2006. 11th International Conference on Implementation and Application of Automata (CIAA 06).

[19] C. Carrez. On the minimization of non-deterministic automata. Technical report, Laboratoire de Calcul de la Faculté des Sciences de L'Université de Lille, 1970.

[20] S. Lombardy. *Approache structurelle de quelques problémes de la théorie des automates.* PhD thesis, Ecole N.S. de Télécomunications, 2001.

[21] L. Polák. Minimalizations of nfa using the universal automaton. *Lecture Notes on Computer Science*, 3317:325–326, 2005.

[22] J. M. Sempere. A representation theorem for languages accepted by watson-crick finite automata. *Bulletin of the EATCS*, 83:187–191, 2004.

[23] D. Lopez, A. Cano, M. Vazquez de Parga, B. Calles, J.M. Sempere, T. Perez, J. Ruiz, and P. Garcia. Detection of functional motifs in biosequences: A grammatical inference approach. In *Proceedings of the 5th Annual Spanish Bioinformatics Conference*, pages 72–75. Univ. Politècnica de Catalunya, 2004. ISBN: 84-7653-863-4.

[24] D. López, A. Cano, M. Vázquez de Parga, B. Calles, J. M. Sempere, T. Pérez, M. Campos, J. Ruiz, and P. García. Motif discovery by $k$-tss grammatical inference. In G. Paliouras C. de la Higuera, T. Oates and M. Van Zaanen, editors, *IJCAI-05 Workshop on Grammatical Inference Applications: Successes and Future Challenges*, 2005. Working Notes.

[25] M. Burset and R. Guigó. Evaluation of gene structure prediction programs. *Genomics*, 34:353–367, 1996.

[26] G. I. Álvarez. *Estudio de la Mezcla de Estados Determinista y No Determinista en el Diseño de Algoritmos para Inferencia Gramatical de Lenguajes Regulares.* PhD thesis, Universidad Politécnica de Valencia, 2007.

[27] M. Vázquez de Parga. *Autómatas finitos: Irreducibilidad e Inferencia.* PhD thesis, Universidad Politécnica de Valencia, 2008.

[28] P. García, M. Váquez de Parga, G. I. Álvarez, and J. Ruiz. Universal automata and nfa learning. *Theoretical Computer Science*, 407(1-3):192–202, 2008.

[29] P. García, M. Váquez de Parga, G. I. Álvarez, and J. Ruiz. Learning regular languages using nondeterministic finite automata. *Lecture Notes on Computer Science*, 5148:92–102, 2008. 13th International Conference on Implementation of Automata (CIAA'08).

[30] P. García, M. Vázquez de Parga, and D. López. On the efficient construction of quasi-reversible automata for reversible languages. *Information Processing Letters*, 107(1):13–17, 2008.

[31] P. García, M. Vázquez de Parga, A. Cano, and D. López. On locally reversible languages. *Theoretical Computer Science*, 410(47-49):4961–4975, 2009.

[32] Gloria I. Álvarez Antonio Cano Gómez. Learning commutative regular languages. *Lecture Notes in Artificial Intelligence*, 5278:71–83, 2008. 9th International Colloquium on Grammatical Inference (ICGI'08).

[33] F. Manea, V. Mitrana, and J. M. Sempere. Some remarks on superposition based on watson-crick-like complementarity. *Lecture Notes on Computer Science*, 5583:372–383, 2009. 13th International Conference on Developments in Language Theory (DLT'09).

[34] M. Campos, D. López, and P. Peris. Incremental multiple sequence alignment. *Lecture Notes on Computer Science*, 4756:604–614, 2007. 12th Iberoamerican Congress on Pattern Recognition, CIARP 2005.

[35] J. M. Sempere. Learning context-sensitive languages from linear structural information. *Lecture Notes in Artificial Intelligence*, 5278:175–186, 2008. 9th International Colloquium on Grammatical Inference (ICGI'08).

[36] J. M. Sempere. Translating multiset tree automata into p systems. *Lecture Notes on Computer Science*, 5391:427–437, 2008. Selected and revised papers from 9th Workshop on Membrane Computing (WMC9).

[37] J. M. Sempere. Computing by carving with p systems. a first approach. In *Sixth Brainstorming Week on Membrane Computing*, pages 255–260. Fenix Editora, 2008.

[38] J. M. Sempere. Exploring regular reversibility in watson-crick finite automata. In *13th International Symposium on Artificial Life and Robotics (AROB'08)*, pages 505–509, 2008.

[39] P. Bottoni, A. Labella, F. Manea, V. Mitrana, and J. M. Sempere. Filter position in networks of evolutionary processors does not matter: A direct proof. *Lecture Notes on Computer Science*, 5877:1–11, 2009. 15th International Meeting on DNA Computing and Molecular Programming (DNA15).

[40] P. Bottoni, A. Labella, F. Manea, V. Mitrana, and J. M. Sempere. Networks of evolutionary picture processors with filtered connections. *Lecture Notes on Computer Science*, 5715:70–84, 2009. 8th International Conference on Unconventional Computation (UC'09).

[41] V. Mitrana and J. M. Sempere. Accepting evolutionary p systems. *Lecture Notes on Computer Science*, 2010. Selected and revised papers from 10th Workshop on Membrane Computing (WMC10).

[42] D. López, J. Calera-Rubio, and J. Gallego-Sánchez. Grammatical inference of graph languages with polynomial time complexity. *Journal of Machine Learning Research*, 2010. under review.

[43] P. Peris, D. López, and M. Campos. Igtm: an algorithm to predict transmembrane domains and topology in proteins. *BMC-Bioinformatics*, 9:367–378, 2008.

[44] P. Peris, D. López, and J. M. Sempere. igpred: grammatical inference tool for protein domain prediction. submitted.

[45] M. Campos, J. González, T.A. Pérez, and J. M. Sempere. Implementing evolutionary processors in java: A case study. In *13th International Symposium on Artificial Life and Robotics (AROB'08)*, pages 510–515, 2008.

[46] J. M. Sempere. Sticker expressions. In *14th International Meeting on DNA Computing (DNA 14th)*, pages 200–201, 2008.