UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

JBI2016

XIII Symposium on Bioinformatics

May, 10th -13th 2016

Valencia - Spain

iNB
Spanish National
Bioinformatics Institute

Main sponsor: FUJITSU intel

Sponsors:

FEDER
Fondo Europeo de Desarrollo Regional
"Una manera de hacer Europa"
UNIÓN EUROPEA

MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Instituto de Salud Carlos III

PRB2

etsinf
Escola Tècnica Superior d'Enginyeria Informàtica

ESCUELA TÉCNICA SUPERIOR INGENIEROS INDUSTRIALES VALENCIA

ITACA

IUMPA
Instituto Universitario de Matemática Pura y Aplicada

biobam
BIOINFORMATICS SOLUTIONS

SEQUENTIA

elixir

Affiliated Conference:

ISCB
INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Collaborators:

acm

BIB
BIOINFORMATICS

#106

Using Formal Language Theory to characterize biosequences.

Jose M. Sempere[1]

[1]*Departamento de Sistemas Informáticos y Computación. Universidad Politecnica de Valencia. Spain*

Formal Language Theory has been a good framework to build bioinformatic tools, especially for motif prediction tasks. for example, D. Searls [1] proposed formal grammars to characterize DNA and RNA biomolecules. The purpose of this work is to describe some motif prediction tools that we have developed by taking the formal language approach.

In this work, we describe tools that characterize the following motifs: (1) RNA/single stranded DNA hairpin structures, (2) nucleotide mutations (pairwise alignments), (3) RNA pseudoknots, and (4) DNA recombination through splicing sites. for the first task, we use non-regular languages based on the iterated and bounded hairpin languages [2], for the second task, we use classical operations over strings [3], for the third task we use stochastic context-free grammars [4] and Watson-Crick finite automata [5] and, finally, for the fourth task we use finite automata [6].

We discuss the (dis)advantages of specific (stochastic) language parsers, and its corresponding computational complexity. We establish some guidelines for a general strategy to apply formal languages in the development of good motif prediction tools. We propose grammatical inference as a good approach for the training of this tools, and we conclude by introducing BIOLANG a formal language approach to characterize biosequences, which actually is work in progress.

[1] D. Searls. The language of genes. Nature vol 420 (2002) 211-217.
[2] D. Cheptea et al. A new operation on words suggested by DNA biochemistry: hairpin completion, In: Proc. Transgressive Computing (2006) 216-228.
[3] M. Crochemore et al. Algorithms on strings. (2007)
[4] E. Rivas et al. The language of RNA; a formal grammar that includes pseudoknots. Bioinformatics vol 16 No 4 (2000) 334-340.
[5] G. Paun et al. DNA Computing. New Computing Paradigms. (1998)
[6] F. Wang at al. Recognition of Simple Splicing Systems using SH-Automaton. Journal of Fundamental Sciences vol 4 No 2 (2008) 337-342