



## Becas colaboración curso 2020/2021

Fecha: 19 Junio 2020

### Vicerrectorado de Investigación, Innovación y Transferencia

Subcomisión de I+D+i

Propuesta del departamento *SISTEMAS INFORMÁTICOS Y COMPUTACIÓN*

**Núm Proyecto: 2020/32/00014**

#### Responsable

Sempere Luna, José María

#### E-mail

jsempere@dsic.upv.es

#### Ext.

73532

#### Título proyecto

Diseño e implementación de sistemas de anotación genómico basados en computación biomolecular y biocelular y técnicas de machine learning

#### Valoración proyecto

4

#### Descripción proyecto

En este proyecto se aborda el diseño e implementación de un sistema de anotación de información de naturaleza genómica (ARN estructural, exomas, proteínas,...). Fundamentalmente, se elegirá un dominio de aplicación (mutaciones genéticas, localizaciones de "splicing alternativos", motivos funcionales y/o estructurales, etc.) y se diseñarán e implementarán los algoritmos que permitan anotar secuencias de acuerdo con el dominio seleccionado. Para la obtención de esos algoritmos, se aplicarán técnicas de machine learning bajo el paradigma de Inferencia Gramatical. Se definirán modelos de computación que permitan realizar la tarea bajo estudio (i.e. autómatas de Watson-Crick, sistemas de splicing, sistemas P, ...) y se definirán y probarán los algoritmos de inducción que, a partir de un conjunto finito de datos anotados, permitan obtener los modelos de anotación previamente definidos. Posteriormente, se diseñarán e implementarán los algoritmos de análisis (parsing) que permitan utilizar los modelos sobre nuevas secuencias que no se hayan utilizado durante la fase de machine learning. El objetivo final del proyecto es obtener una prueba de concepto que valide estas nuevas técnicas como alternativa a las técnicas de mapping que habitualmente se utilizan en bioinformática.

#### Actividades a realizar por el alumno

El alumno deberá cubrir una serie de fases durante el proyecto que pasamos a detallar:

(1) Fase de formación y aprendizaje de nuevos conceptos y técnicas: El alumno se deberá familiarizar con los conceptos básicos de la computación con ADN y la computación con membranas. También con las técnicas y resultados básicos de la inferencia gramatical. Los conceptos básicos sobre biología molecular de la célula y genética también se deberán adquirir en esta fase, así como el conocimiento de los formatos y repositorios que se suelen utilizar en el ámbito de la bioinformática.

(2) Fase de diseño y prueba de los algoritmos: Se deberán adaptar algoritmos clásicos de inferencia gramatical (fundamentalmente, algoritmos sobre lenguajes regulares, k-testables y lineales) a los modelos de computación predefinidos (básicamente autómatas de Watson-Crick, sistemas de splicing y sistemas P). Estos algoritmos se probarán sobre conjuntos de datos del dominio de aplicación

(3) Fase de comunicación de resultados: Al final del proyecto, el alumno debe participar activamente en la redacción de una comunicación científica sobre los resultados del proyecto. Esta comunicación se enviará a alguna revista o congreso científico donde el alumno aparecerá como coautor.

#### Horario

El horario es flexible (mañanas y/o tarde) y está marcado por objetivos. Todas las semanas habrá una sesión de seguimiento que se realizará preferiblemente por la mañana.