Plan de Gestión de Datos

INFORMACIÓN SOBRE EL PROYECTO

1. – Datos del Proyecto

- **Título del Proyecto (en castellano)**

Estimación de distancias semánticas y aprendizaje profundo para la predicción de nuevas funciones de genes

- **Título del Proyecto (en inglés)**

Semantic distance estimation and deep learning for the prediction of novel gene functions

- **Descripción del Proyecto (en castellano) Resumen**

La ciencia de datos ha experimentado un crecimiento exponencial en la última década. Cada día es más fácil adquirir y almacenar datos de todo tipo. Pero los desafíos ahora tienen que ver con la extracción de información útil de esos datos. La inteligencia artificial está proveyendo soluciones efectivas a gran cantidad de problemas de este tipo, especialmente desde el aprendizaje de máquina, que ha demostrado tener todo el potencial necesario para los desafíos actuales. En particular, el área de bioinformática presenta problemas en ciencia de datos cada vez más desafiantes. Por ejemplo, la predicción automática de la función de genes a partir de genomas completos y de mediciones experimentales de diferente naturaleza. Actualmente existen anotaciones semánticas con vocabulario controlado que describen a los genes en cualquier organismo en base a términos de la ontología de genes (GO). La curaduría (manual) de anotaciones para nuevos genes es un procedimiento muy costoso que requiere de conocimiento específico de parte del experto del dominio. Las herramientas computacionales basadas en aprendizaje de máquina pueden ayudar a encontrar rápidamente potenciales anotaciones para genes nuevos, e impulsar el descubrimiento de nuevo conocimiento en este dominio. Este proyecto se proponen nuevos modelos y algoritmos para predecir anotaciones de genes cuya potencial función es desconocida, es decir sin términos GO asociados, mediante el desarrollo de métodos novedosos de aprendizaje de máquina. En primer lugar se propone desarrollar un nuevo método a partir de factorización conjunta de matrices no negativas de distancias de expresión y distancias semánticas entre genes conocidos. Una vez realizada esta factorización, se propone utilizarla para reconstruir la información faltante en la matriz de distancia semántica a genes desconocidos. Una segunda etapa utilizará esta información semántica reconstruida para entrenar modelos probabilísticos y modelos de aprendizaje profundo que permitan predecir el conjunto de etiquetas GO que describen la función de cada gen desconocido.

- **Descripción del Proyecto (en inglés) Resumen**

Data science has experienced exponential growth in the last decade. Every day it is easier to acquire and store data of all kinds. But the challenges now have to do with extracting useful information from that data. Artificial intelligence is providing effective solutions to a large number of problems of this kind, especially machine learning, which has proven to have all the necessary potential for current challenges. In particular, the bioinformatics area presents problems in data science more challenging every time. For example, the automatic prediction of gene function from complete genomes and experimental measurements of different nature. There are currently semantic annotations with controlled vocabulary that describe genes in any organism based on terms of the ontology of genes (GO). Curation (manual) of annotations for new genes is a very expensive procedure that requires specific knowledge from the domain experts. The computer tools based on machine learning can help finding potential annotations for new genes, and drive the discovery of new knowledge in this domain. This project proposes new models and algorithms to predict gene annotations whose potential function is unknown, that is to say without GO terms, by developing novel machine learning methods. First, it is proposed to develop a new method with the non-negative matrix factorization of both expression distances and semantic distances between known genes. Once this factorization is done it is proposed to use it to reconstruct the missing information in the matrix of semantic distance of unknown genes. A second stage will use this information to train probabilistic models and deep learning models to predict the set of GO terms that could describe the function of each unknown gene

- Palabras Claves descriptivas del Proyecto (en castellano)

Bioinformática,
Aprendizaje maquinal,
Reconstrucción de matrices
Anotaciones funcionales
Ontología de genes
Distancia semántica
Función de genes

- Palabras Claves descriptivas del Proyecto (en inglés)

Bioinformatics
Machine learning,
Matrix reconstruction
Functional annotations,
Gene ontology,
Semantic distance
Gene function

2 – Datos del Director/ar del Proyecto

- Nombre y Apellido

GEORGINA SILVIA STEGMAYER

- Unidad Académica

FICH

- Teléfono oficial de contacto

+54 (342) 4575233/34, ext. 193

-Teléfono móvil de contacto

154288019

-E-mail del Director/a del Proyecto

gstegmayer@sinc.unl.edu.ar

DATOS RESULTANTES DE LA EJECUCIÓN DEL PROYECTO

-Describa la toma de muestras / datos a realizar

Los datos para la experimentación están disponibles en forma gratuita en Internet.

– Datos: ¿Existe alguna razón por la cual los datos declarados no deban

ser puestos a disposición de la comunidad/ser de acceso público? (marque X)	
	NO X
	SI. Elija una de las opciones: a) Se encuentra en evaluación de protección por medio de patentes b) No se inició el proceso de evaluación de patentabilidad, pero podría ser protegible c) Existe un contrato con un tercero que impide la divulgación d) Otro. Justifique.
– Período de Confidencialidad: Es el período durante el cual los datos no deberían ser publicados, contado a partir del momento de la toma de los mismos. El período máximo para la no publicación es de 5 (CINCO) años posteriores a su obtención. Luego de este periodo, los datos estarán disponibles para la comunidad/serán de acceso público. **Si Ud. considera que este tiempo es insuficiente, y necesita prorrogar el período de confidencialidad, indique sus motivos y la cantidad de años adicionales que considera necesarios. Marque su opción con "X".**	
	1 (UN) año
	2 (DOS) años
	3 (TRES) años
	4 (CUATRO) año
	5 (CINCO) años
	Otro. **Motivos:**