

GeneFlow

Diseño de una herramienta para el análisis de datos de expresión genética



Trabajo de Fin de Grado de Ingeniería Informática

Alba Casillas Rodríguez



Índice

- Estudio del problema
 - Motivación
 - Contexto biológico
- GeneFlow
 - ¿Qué es GeneFlow?
 - Descarga y lectura de datos
 - Preprocesamiento y análisis
 - Visualización de datos
 - Algoritmos de Aprendizaje Automático
 - Replicación del flujo de trabajo
- Demostración



Estudio del Problema

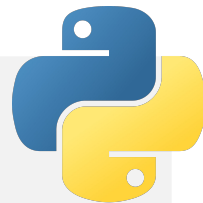
Motivación

Actualmente

- Gran avance en la investigación genética gracias a las nuevas tecnologías y el Big Data.
- Biólogos, estadísticos e investigadores con conocimiento en genómica pero falta de habilidades en programación.

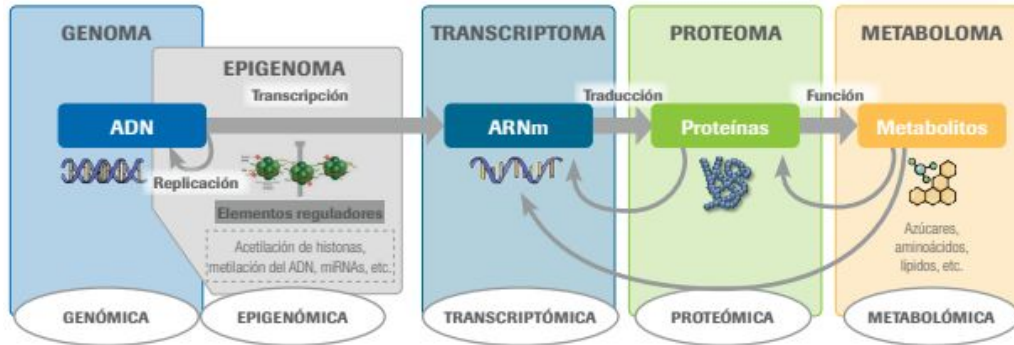
Problema

- Integración de distintos paquetes software en flujos de trabajo.
- Uso de diferentes lenguajes de programación y plataformas informáticas.



Contexto biológico

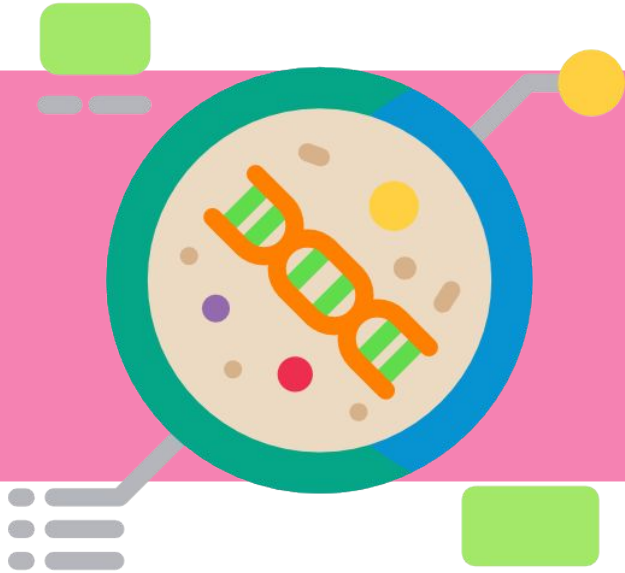
“Ómicas” son las ciencias que permiten estudiar un conjunto de moléculas.



El ADN se transcribe a ARNm (ARN mensajero).

La **transcriptómica** es la ciencia ómica que estudia la expresión de los transcritos que provienen de diferentes genes.

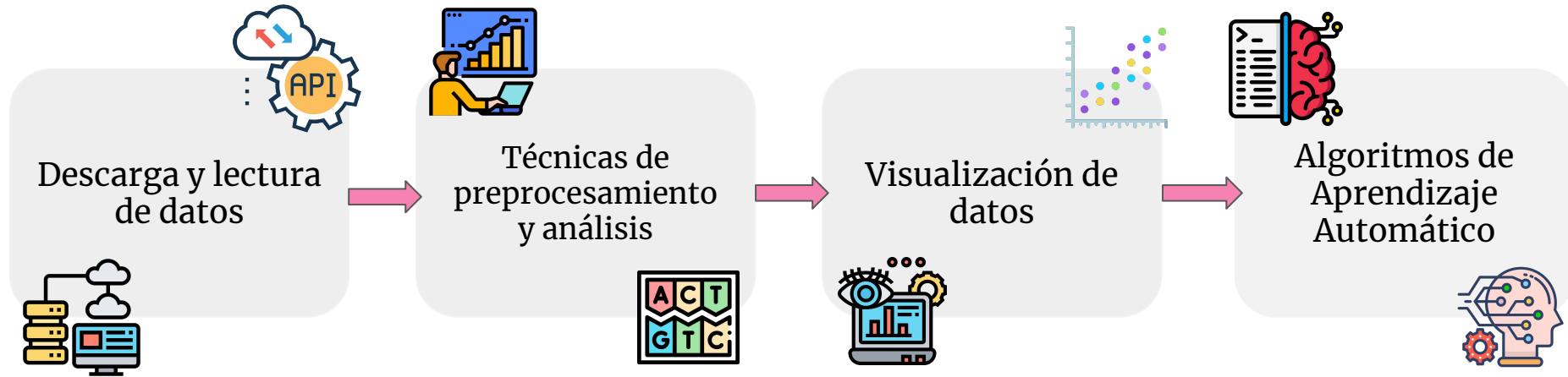
RNA-Seq (secuenciación del ARN) permite analizar cambios en el transcriptoma para revelar la presencia y cantidad de ARN en una muestra biológica en un momento dado.



GeneFlow

¿Qué ofrece GeneFlow?

GeneFlow es una biblioteca software escalable y flexible que permite trabajar con datos genómicos.



Descarga y lectura de datos

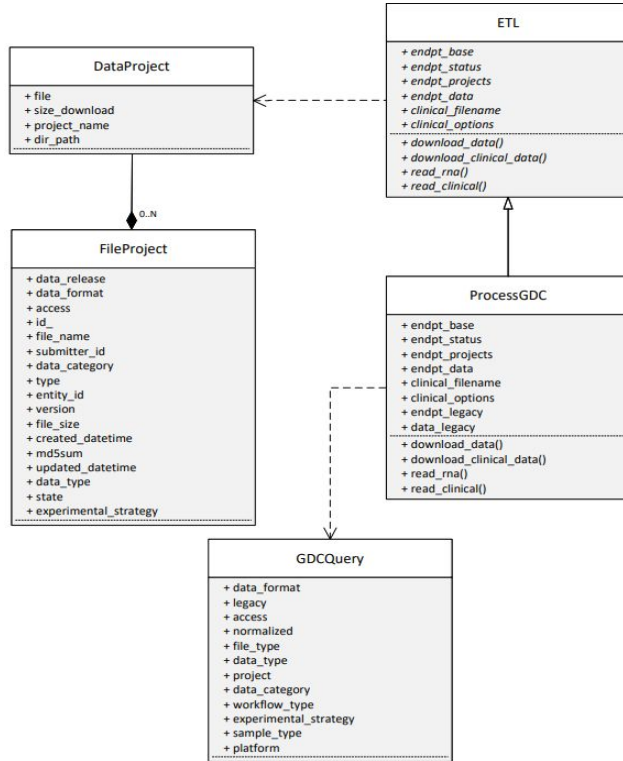


Diagrama de clases para la obtención de datos

Acceso a las fuentes de datos mediante peticiones a la API de:

- Genomic Data Commons Portal (GDC).

Programa:

- The Cancer Genome Atlas (TCGA).

The screenshot shows the NIH National Cancer Institute GDC Data Portal. The header includes the NIH logo and the text "NATIONAL CANCER INSTITUTE GDC Data Portal". Below the header, the text "Harmonized Cancer Datasets" and "Genomic Data Commons Data Portal" are displayed. A navigation bar contains buttons for "Projects", "Exploration", "Analysis", and "Repository". A search bar with the text "e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2" is present. The main content area features a "Data Portal Summary" section with the following data:

Data Release 19.0 - September 17, 2019	
PROJECTS	53
PRIMARY SITES	67
CASES	37,075
FILES	427,407
GENES	22,872
MUTATIONS	3,142,246

The bottom right corner of the screenshot shows the URL "portal.gdc.cancer.gov" and an illustration of two human figures with internal organs highlighted.

Preprocesamiento y análisis

Una **Tarea (Task)** engloba cualquier acción de preprocesamiento que se pueda ejecutar con GeneFlow.

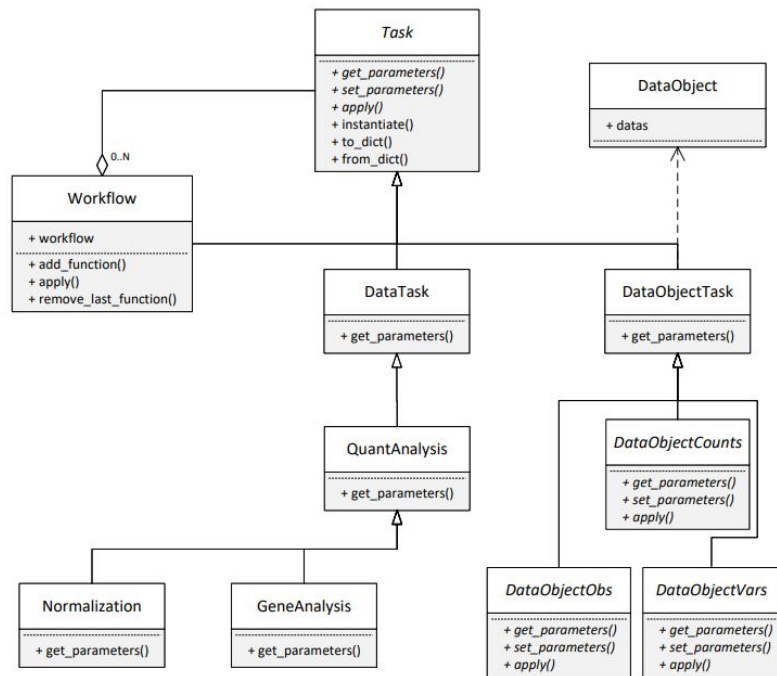
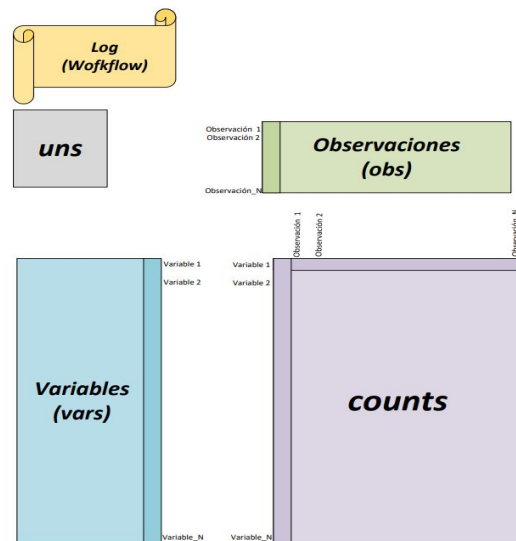


Diagrama de clases para el preprocesamiento de datos

- Diseño orientado a objetos.
- Representación en un formato semi-estructurado el proceso de trabajo.



Objeto DataObject

Visualización de datos

Boxplots de Log(CPM)

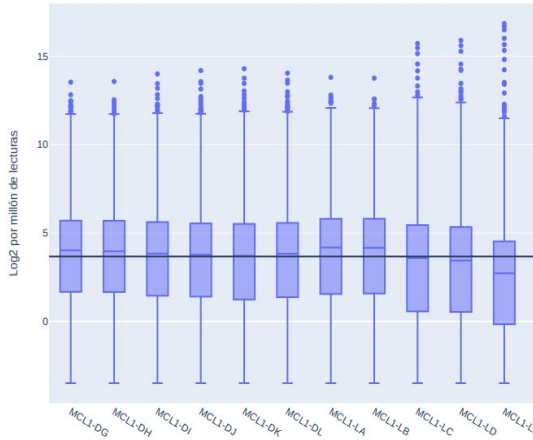
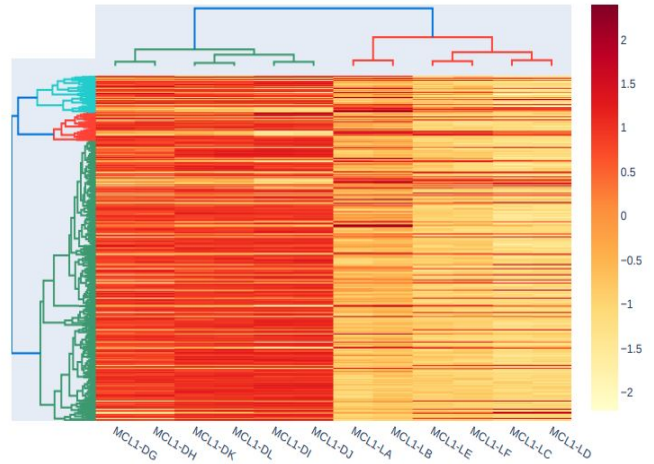


Diagrama de cajas

Clustermap



Clustermap

MDSplot según Estado Celular

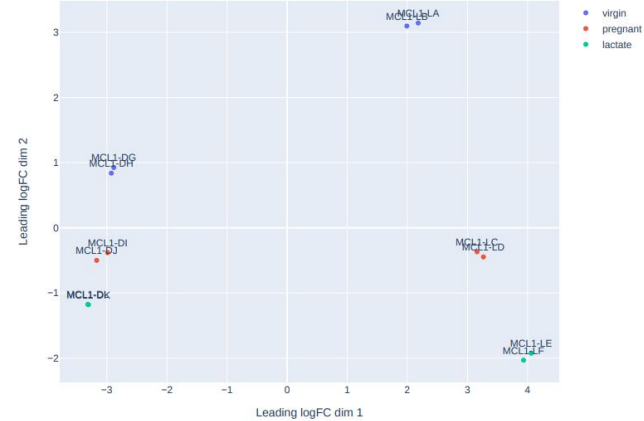


Gráfico de Escalamiento Multidimensional

Algoritmos de Aprendizaje Automático

Se han implementado métodos para:

- Partición del conjunto de datos.
- Normalización.
- Tratamiento de datos no balanceados.
- Validación cruzada.
- Algoritmos de Aprendizaje Automático.
- Cálculo de hiperparámetros.
- Cálculo de métricas.

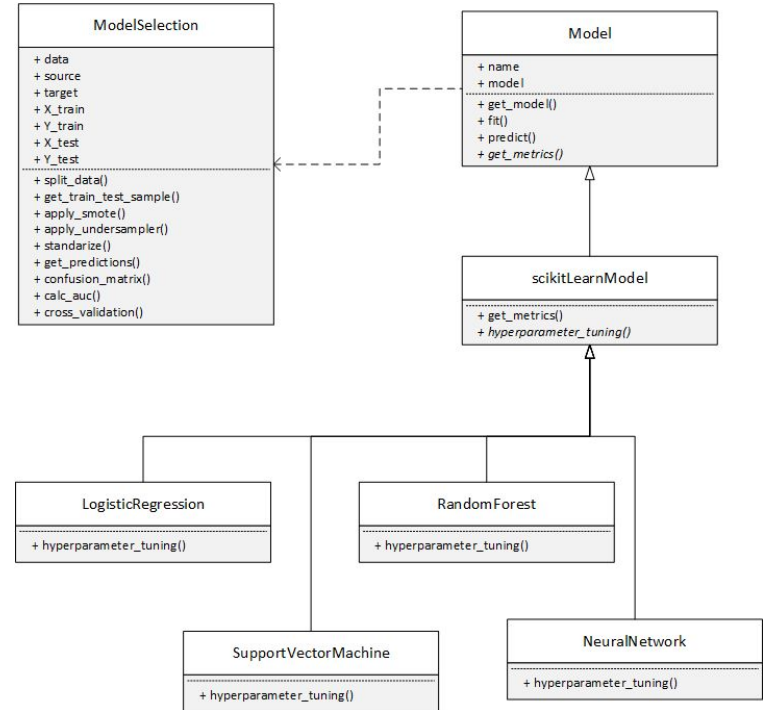


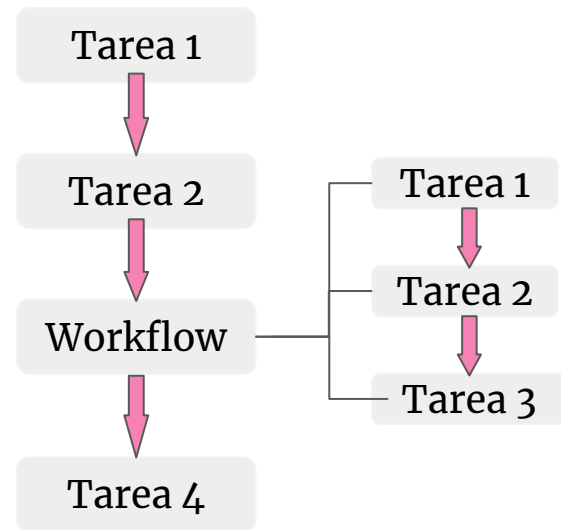
Diagrama de clases para el modelado de datos y algoritmos de Aprendizaje Automático

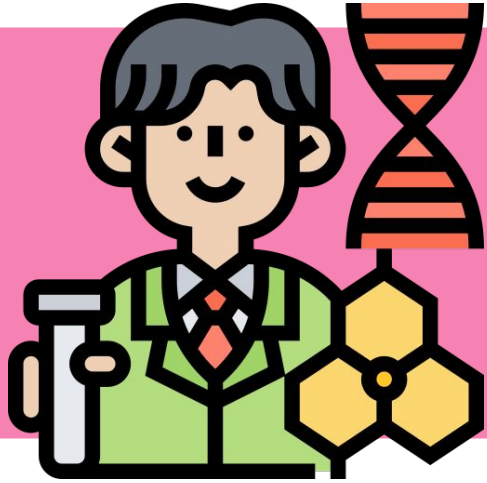
Replicación del flujo de trabajo

Reflexión es la capacidad de un programa para inspeccionar sus metadatos y modificar sus objetos en tiempo de ejecución.

La capacidad de reflexión en GeneFlow permite:

- Instanciación dinámica de las sub-Tareas.
- Reproducibilidad del flujo de trabajo realizado.
- Realización de experimentos más complejos.





Demostración

Conclusiones

Se ha desarrollado un paquete software en Python que permite:

- Analizar datos de expresión genética de una manera fácil y entendible para la comunidad científica.
- Integrar funciones de R y Python en un único paquete software.
- Obtención de datos mediante APIs.
- Análisis, preprocesamiento y visualización de datos.
- Entrenar modelos con algoritmos de Aprendizaje Automático.
- Reproducir experimentos replicando el flujo de trabajo.

