Trabajo Práctico Final

Introducción a la Bioinformática - UNQ

Cursada: 1er Cuatrimestre 2020
Docentes: Ana Julia Velez Rueda y Patricio Barletta

1. Objetivo **1**
1.1 Objetivo Específico 1
2. Contexto **1**
3 Requerimientos detallados **2**
3.1 Carga de secuencias 2
3.2 Datación y localización 3
3.3 Inferencia evolutiva 3
3.3 Visualización 3
3.4 Tecnologías 3
4. Forma de entrega **4**
5. Bibliografía **4**

1. Objetivo

El objetivo del presente trabajo práctico es integrar los conceptos biológicos y bioinformáticos desarrollados durante la materia, junto con los conocimientos, prácticas y habilidades propias de la informática y la programación adquiridas en otras materias, a través de la construcción de un software simple pero innovador que sirva de asistencia al proceso del análisis bioinformático y/o herramienta educativa para la enseñanza de la Biología.

1.1 Objetivo Específico

El objetivo es desarrollar un software que permita la visualización para estudios filogenéticos y filodinámicos (la geolocalización) de secuencias. El software podrá ser desarrollado bajo cualquier arquitectura y empleando cualquier tecnología de **código libre**.

2. Contexto

La historia de las especies y cómo estas han cambiado desde que se desarrollara la vida en la tierra, ha quedado registrada en los genomas de las especies actuales. Los estudios evolutivos permiten realizar inferencias estructurales y funcionales donde el conocimiento

sobre los sistemas aún es insuficiente. Sin embargo, este tipo de inferencia, requiere de la estimación de distancias evolutivas entre especies, basadas en las diferencias entre genes ortólogos. Como bien ya explicamos dos secuencias que comparten un ancestro común se denominan secuencias homólogas (Reeck et al., 1987). La predicción de homología se realiza extrayendo de las secuencias la información conservada durante la evolución, para lo que resulta necesario la comparación de las secuencias para identificar los residuos que tienen en común.
Los árboles filogenéticos son algo así como el árbol genealógico de las especies, e implican una hipótesis sobre las relaciones que existen entre los organismos. Su confección requiere de ciertas métricas que tengan en cuenta el tiempo requerido para poder observar la divergencia de las secuencias (Kalyaanamoorthy et al. 2017). Una vez determinadas las distancias entre las secuencias, los distintos organismos pueden ser agrupados, utilizando distintos algoritmos de clustering. Debemos tener en cuenta que la construcción de árboles filogenéticos requiere de la utilización de caracteres que sean indicadores fiables de una ascendencia común. Genes distintos tienen tasas evolutivas distintas que dependen de la estructura y función de las proteínas que codifican (Bromham 2009), por lo que las secuencias que se seleccionen para el análisis deben ser informativas y representar la evolución de dichas especies.

> **Para pensar**: ¿qué factores pueden hacer que las mutaciones ocurran?¿Cómo se relacionan estas con la diversidad biológica?

Entender una filogenia es sencillo cuando lo comparamos con los árboles genealógicos, en los que la raíz representa el linaje ancestral y los extremos de las ramas representan los descendientes de ese antepasado. Las filogenias siguen la pista a los patrones de ascendencia compartidos por los linajes, donde cada linaje tiene una parte de historia que es única y otras partes que son compartidas con otros linajes. Las inferencias evolutivas nos proponen, entonces, relaciones entre las especies, que comparten características en común (Iantorno et al. 2014; Ashkenazy et al. 2019). En el presente trabajo, construiremos una herramienta que nos permita visualizar dichas relaciones de secuencias homólogas, trazando a su vez su origen geográfico y su datación, a modo de facilitar el análisis epidemiológicos y migratorios de especies.

3 Requerimientos detallados

Como bien explicitamos, el objetivo del presente trabajo es desarrollar un software que permita la visualización para estudios filogenéticos y filodinámicos (la geolocalización) de secuencias, y deberá dar respuesta a los casos de uso expresados a continuación:

3.1 Carga de secuencias

El sistema debe permitir la carga de secuencias a analizar o un alineamiento múltiple de secuencias en el formato FASTA[1], para su posterior procesamiento. Si la secuencia

[1] https://es.wikipedia.org/wiki/Formato_FASTA

ingresada es inválida, el software deberá notificarlo apropiadamente, mostrando a el/la usuario/a un mensaje claro y útil. El programa debe explicitar en su documentación el formato del input requerido.

3.2 Datación y localización

Las secuencias debe tener contar con un código de accesión de GenBank, además pueden tener los datos de datación y de localización separado por tabs. Esta localización deben reescribirse automáticamente en latitud y longitudes para su geolocalización. El programa debe habilitar la posibilidad al usuario de cargar de forma complementaria una tabla que correlacione estos accessions, la localización y los datos de relativos a la datación.

> **BONUS: Como forma adicional, se considerará como un mejor programa aquel que también realicen una notación automática de las secuencias, tomando los accession numbers y obteniendo los datos correspondientes a dicha secuencia de la base de datos de NCBI (https://www.ncbi.nlm.nih.gov/home/about/policies/#accessibility).**

3.3 Inferencia evolutiva

El programa deberá establecer las relaciones evolutivas entre las secuencias, utilizando los algoritmos optimizados que ya existen. Si las secuencias cargadas no se encuentran alineadas, deberá realizar este paso previo requerido para las inferencias evolutivas.

> **BONUS: Como forma adicional, se considerará como un mejor programa aquel que permita a los usuarios customizar los parámetros de corrida para cada paso.**

3.3 Visualización

Finalmente, el sistema deberá permitir la visualización del árbol filogenético para dichas secuencias y su ubicación en un mapa interactivo, siguiendo las localizaciones cargadas por el usuario, de una forma que sea fácil de relacionar una visualización con la otra.

> **BONUS: Como forma adicional, se considerará como un mejor programa aquel que incorpore la temporalidad de las dataciones en la visualización. Mostrando la migración de las poblaciones.**

3.4 Tecnologías

El software podrá estar implementado utilizando cualquier tecnología. Sin embargo, se recomienda fuertemente utilizar alguna de las siguientes:

- Python. Librerías recomendadas: Biopython, Geopy, Folium, Pandas.
- JavaScript

Herramientas complementarias:

- IQtree: http://www.iqtree.org/
- Clustal: http://www.clustal.org/omega/

4. Forma de entrega

El trabajo práctico se realizará en equipos de hasta 4 integrantes, con **fecha de entrega del proyecto los días 17 de Julio**. El mismo deberá además estar o a un repositorio público en Github creado para la materia (https://github.com/BioinformaticaUNQ), el cual deberá contener un archivo README.md con los datos de l@s integrantes del equipo. Los trabajos serán presentados por los miembros del equipo en una **exposición oral (con DEMO)**, los días **17 y 24 de Julio.**

5. Bibliografía

Ashkenazy, H., Sela, I., Levy Karin, E., Landan, G. and Pupko, T. 2019. Multiple sequence alignment averaging improves phylogeny reconstruction. *Systematic Biology* 68(1), pp. 117–130.

Bromham, L. 2009. Why do species vary in their rate of molecular evolution? *Biology Letters* 5(3), pp. 401–404.

Iantorno, S., Gori, K., Goldman, N., Gil, M. and Dessimoz, C. 2014. Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. *Methods in Molecular Biology* 1079, pp. 59–73.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. and Jermiin, L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* 14(6), pp. 587–589.