

UNIVERSIDAD NACIONAL DE INGENIERIA

FACULTAD DE CIENCIAS

CIENCIAS DE LA COMPUTACIÓN



PROYECTO DE BIOLOGIA COMPUTACIONAL

Título del Trabajo

Creación de un sistema de ayuda para el análisis de información Genética de Especies endémicas de las regiones del Perú

Autores

Lázaro Camasca, Edson Nicks
Leon Rios, Marco Naro

Profesor

Nuñez Iturri, Ciro Javier

Lima - Peru
(2019)

Contents

1	Objetivos	2
1.1	Objetivos Generales	2
1.2	Objetivos Específicos	2
2	Resumen Ejecutivo	2
3	Descripción del Proyecto	2
3.1	Determinar las especies y el material a utilizar	2
3.2	Elegir los marcadores moleculares	3
3.3	Realizar el alineamiento múltiple de genes homólogos	3
3.4	Modelo Evolutivo Kimura	4
3.4.1	Construcción de la matriz de distancias	4
3.5	Métodos para la construcción de árbol filogenético	4
3.5.1	UPGMA	4
3.5.2	Unión de Vecinos	5
3.6	Verificar la fiabilidad del árbol construido	5
3.7	Analizar el árbol filogenético	5
3.7.1	Modelamiento de la Estructura de Proteínas	5
3.8	Cronograma	5
4	Algoritmos e implementación computacional	6
5	Resultados	6
6	Conclusiones	6
7	Apéndice	6

1 Objetivos

1.1 Objetivos Generales

- Creación de una aplicación gráfico con una base de datos para el análisis de información genética.

1.2 Objetivos Específicos

- Recolectar información genética de especies endémicas.
- Desarrollar la aplicación para el análisis de secuencias.,
- Desarrollar algoritmos para obtener árboles filogenéticos
- Evaluar el árbol filogenético

2 Resumen Ejecutivo

Se pretender crear una aplicación gráfica conectada a una base de datos con la información genética de las especies endémicas del Perú, el software procesara las secuencias, creará el árbol filogenético, mostrará los resultados y analizará las relaciones evolutivas de las especies escogidas. Se escogió como marcaadores moleculares a la proteína NADH deshidrogenasa subunidad 2 debido la importante función respiratoria mitocondrial, hecho por el cual tiene presencia en todas las especies escogidas. Se realizó el alineamiento utilizando la herramienta en linea EMBL-EBI. Se utilizó el modelo evolutivo Kimura y el método para la construcción de árbol escogido fue un método basado en agrupamiento. Se implementó el UPGMA y el método Unión de vecinos.

3 Descripción del Proyecto

El proyecto sera implementado netamente en el lenguaje Python Las librerías utilizadas serán:

- BioPython para el procesamiento de secuencias
- Tkinter para el entorno gráfico.

Dentro de la GUI, se pobre escoger Especies para el análisis posterior.

Los datos recolectados serán reales de la base de datos de NCBI.

Se implementara algoritmos para el alineamiento de Genes homólogos.

Se implementara algoritmos para el alineamiento de Proteínas.

Se implementara la algoritmos para la Generación de Arboles Filogenéticos de acuerdo a un modelo.

Para el desarrollo del proyecto se seguirá la siguiente metodología:

3.1 Determinar las especies y el material a utilizar

Las especies se escogieron por ser especies endémicas del Perú, especias en peligro de extinción en el Perú y especies representativas del Perú. Además, se tuvo en cuenta la disponibilidad de material genético puesto que muchas de las especies endémicas poseen una base de datos genética incompleta y algunas no se encuentran codificadas en absoluto.

Se escogió las siguientes especies:

- Tremarctos ornatus
- Panthera onca
- Vicugna vicugna
- Aulacorhynchus huallagae
- Leopardus jacobita

- Inia geoffrensis
- Spheniscus humboldti
- Vultur gryphus
- Lama glama
- Cavia porcellus
- Platalea ajaja

3.2 Elegir los marcadores moleculares

La elección de los marcadores moleculares es una parte importante porque puede hacer una **gran diferencia** en la obtención de un árbol correcto.

Entre los marcadores moleculares, secuencias de nucleótidos o de proteínas, se optó por utilizar **secuencias de proteínas** por las siguientes razones:

- Como se va estudiar la evolución de grupos de **organismos ampliamente divergentes** se aconseja utilizar secuencias de proteínas.
- Las relaciones filogenéticas que se están analizando están en el **nivel más profundo - bacteriana**, por ello lo más adecuado es usar secuencias de proteínas conservadas.

Proteína NADH deshidrogenasa

La proteína a analizarse será el NADH deshidrogenasa, también conocido como Complejo I, subunidad 2 debido a que se encuentra presente en todas las especies y está codificado.

La proteína escogida cumple una importante **función en la respiración bacteriana y mitocondrial**. Por ende, es posible encontrarla en diversas especies y no es extraño que se haya codificado. Una importante observación es que no todas las especies se encuentran codificadas en la base de datos de NCBI, faltando genes importantes.

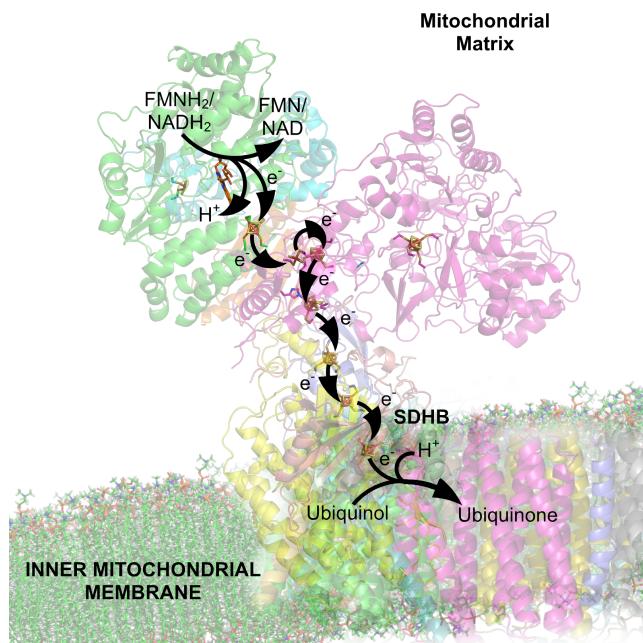


Fig. Proteína NADH deshidrogenasa

3.3 Realizar el alineamiento múltiple de genes homólogos

Este paso es el mas importante de todas, ya que éste establece las correspondencias posicionales en la evolución.

Sólo el alineamiento correcto produce inferencias filogenéticas correctas.

EMBL-EBI

Para la alineación de secuencias se utilizada la Web de EMBL-EBI: **Multiple Sequence Alignment** [click aquí para consultar la pagina](#)

Las secuencias se enviaran a la Web, esta realizara la alineación luego se descargara los resultados en formato clustal.

Cuando descargamos el archivo tiene el siguiente formato ".clwstrict" luego cambiamos a ".clustal".

Clustalo

Para usuarios de Ubuntu se utilizara el software de clustal.

3.4 Modelo Evolutivo Kimura

Ya que se quiere resultados lo más sofisticado (realista) se optara por el **Modelo Kimura**. Este modelo considera **diferentes las tasas de mutación** para las transiciones (substitución de una purina por otra o una pirimidina por otra) y para las transversiones (substitución de una purina por una pirimidina o vice versa)

De acuerdo a este modelo las transiciones ocurren más frecuentemente que las transversiones, lo cual provee mejores estimaciones de la distancia evolutiva

3.4.1 Construcción de la matriz de distancias

A partir del modelo Kimura tenemos:

$$d_{AB} = -(1/2)\ln(1 - 2p_{ti} - p_{tv}) - (1/4)\ln(1 - 2p_{tv})$$

Donde:

p_{ti} es la frecuencia observada de transición.

p_{tv} es la frecuencia de transversión.

Las distancias evolutivas calculadas pueden ser usadas para construir una matriz de distancias entre todos los pares de taxones

3.5 Métodos para la construcción de árbol filogenético

Los algoritmos basados en distancias para construir árboles filogenéticos pueden ser subdivididos:

Métodos basados en agrupamiento

Los algoritmos basados en agrupamiento calculan el árbol usando una matriz de distancias e iniciando por los pares de secuencias más similares.

Un gran ventaja es la habilidad para hacer uso de **diferentes modelos de substitución** para corregir las distancias evolutivas.

Métodos basados en optimalidad

Los algoritmos basados en optimalidad comparan muchas topologías alternativas de árboles y seleccionan el que tenga el mejor ajuste entre las distancias estimadas en el árbol y las distancias evolutivas reales

Los métodos basados en optimalidad requieren **muchas capacidades de computo** debido a la búsqueda exhaustiva que realizan, por ello se optó por escoger basados en agrupamiento.

El usuario podrá escoger dos métodos, a partir de la interfaz gráfica:

3.5.1 UPGMA

UPGMA (unweighted pair group method using arithmetic average) El método más simple basado en agrupamiento.

3.5.2 Unión de Vecinos

El método de "Unión de vecinos" parte de una matriz de distancias, que indica la distancia entre cada par de taxones. El algoritmo comienza con un árbol completamente sin resolver, cuya topología corresponde a la de una red en estrella, y aplica los siguientes pasos hasta que el árbol está completamente resuelto y las longitudes de sus ramas.

3.6 Verificar la fiabilidad del árbol construido

Para la fase beta se encontró el siguiente árbol.

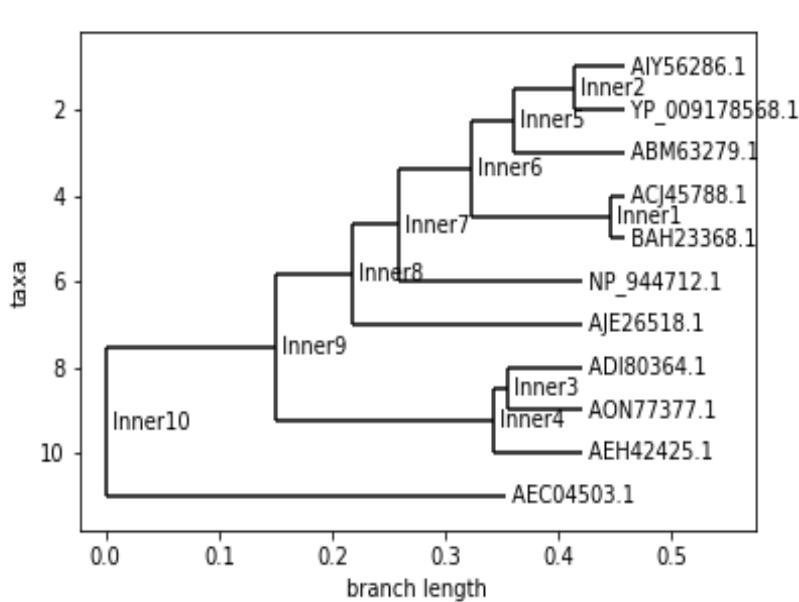


Fig. Árbol utilizando el método UPGMA

3.7 Analizar el árbol filogenético

Apartir del árbol filogenético se podrá descubrir/Mostrar/Analizar las relaciones evolutivas de las especies endémicas escogidas.

3.7.1 Modelamiento de la Estructura de Proteínas

En el análisis se encuentra el modelamiento de la Estructura de Proteínas

3.8 Cronograma

Para el desarrollo del proyecto se emplea un cronograma por semanas, las fechas del cronograma coinciden con las fechas propuestas para evaluaciones de práctica del proyecto. En dichas ocasiones se presentarán y evaluarán los avances del proyecto.

DESCRIPCION	SEMANA								
	7	8	9	10	11	12	13	14	15
Propuesta inicial									
Determinar los datos									
Elegir las especies endémicas									
Recolectar las secuencias y proteínas									
Alineamiento									
Alineamiento en genes homólogos y proteínas									
Ánalisis de los alineamientos									
Creación del árbol filogenético									
Elegir el modelo evolutivo y método									
Implementar el algoritmo y verificar su funcionamiento									
Analizar el árbol filogenético									
Modelamiento de la estructura de proteínas									
Avance del proyecto	5%	10%	15%	30%	40%	50%	60%	70%	100%

Fig: Cronograma por semanas

SEMANA	FECHA	ENTREGABLE	DESCRIPCION
7	29/04/19 al 06/05/19	Entregable 1	Propuesta inicial del proyecto
8	06/05/19 al 13/05/19		
9	13/05/19 al 20/05/19		
10	20/05/19 al 27/05/19	Entregable 2	Propuesta final y elección de especies, genes, proteínas
11	27/05/19 al 03/06/19		
12	03/06/19 al 10/06/19		
13	10/06/19 al 17/06/19		
14	17/06/19 al 24/06/19	Entregable 3	Realización y análisis de alineamiento de genes y proteínas
15	24/06/19 al 01/07/19	Entregable 4	Ánalisis y generación de árboles filogenéticos

Fig: Cronograma coincidente a los entregables

4 Algoritmos e implementación computacional

Una descripción de los algoritmos y herramientas que se [planean utilizar en caso de la propuesta] utilizados incluyendo pseudo código y código fuente

5 Resultados

Una descripción de los resultados [esperados en el caso de la propuesta]. Un reporte integrando los resultados proporcionados por la herramienta

6 Conclusiones

Incluye las ventajas y desventajas del enfoque utilizado, aspectos inesperados del proyecto, trabajo futuro, etc.

7 Apéndice