

UNIVERSIDAD AUTONOMA DE AGUASCALIENTES

CENTRO DE CIENCIAS BÁSICAS

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

Tesina

“Metodología para el cálculo de distancias entre arboles filogenómicos, un enfoque con los mejores aciertos bidireccionales”

**Presenta**

“Emilio Vitali Padilla Socconini”

”

Para Obtener el Grado de Ingeniería en

Computación Inteligente

Director:

Dra. Eunice Ponce de León Sentí

Comité Tutoral:

Dra. Aurora Torres Soto

Dra. Eunice Esther Ponce de León Sentí

Dr. Francisco Javier Álvarez Rodríguez

Aguascalientes, Ags., junio 2020



Aguascalientes, Ags., 10 de junio de 2020

A quien corresponda:

Por medio de la presente me permito informar que el alumno(a) **Emilio Vitali Padilla Socconini**  la carrera de Ingeniería en Computación Inteligente con **ID: 180653**, ha terminado satisfactoriamente su tesina titulada: “**Algoritmo para el cálculo de distancias entre arboles filogenéticos, un enfoque con los mejores aciertos bidireccionales.”**, correspondiente a la materia de Seminario de Investigación II.

2

Para los fines que al interesado convengan.

**ATENTAMENTE**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Dra. Eunice Esther Ponce de León**

Director



Aguascalientes, Ags., 10 de junio de 2020

A quien corresponda:

Por medio de la presente me permito informar que el alumno(a) **Emilio Vitali Padilla Socconini** de la carrera de Ingeniería en Computación Inteligente con **ID: 180653**, ha terminado satisfactoriamente su tesina titulada: **”Metodología para el cálculo de distancias entre arboles filogenéticos, un enfoque con los mejores aciertos bidireccionales”**, correspondiente a la materia de Seminario de Investigación II.

Para los fines que al interesado convengan.

**ATENTAMENTE**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Dra. Eunice Esther Ponce de León**

**Director**

índice

[Resumen. 6](#_Toc42761496)

[Introducción 7](#_Toc42761497)

[Planteamiento del problema de investigación 7](#_Toc42761498)

[Justificación 7](#_Toc42761499)

[Objetivos 8](#_Toc42761500)

[Objetivo General 8](#_Toc42761501)

[Objetivos específicos 8](#_Toc42761502)

[Hipótesis 8](#_Toc42761503)

[Preguntas De Investigación (Opcional si se agregan hipótesis, de no hacerlo se deben incluir) 8](#_Toc42761504)

[Marco Teórico 8](#_Toc42761505)

[Metodología 8](#_Toc42761506)

[Desarrollo 9](#_Toc42761507)

[Experimentación y pruebas 9](#_Toc42761508)

[Análisis e interpretación de resultados 9](#_Toc42761509)

[Conclusiones 9](#_Toc42761510)

[Referencias 9](#_Toc42761511)

[Anexos 9](#_Toc42761512)

[Notas Importantes 9](#_Toc42761513)

# Resumen.

Hay antecedentes en una tesina anterior de la necesidad de generar automáticamente árboles filogenomicos para diferentes valores de exigencia de mejores aciertos bidireccional entre proteínas inter-especie, por esta razón y para implementar una metodología para encontrar un árbol filogenético límite se construye una sucesión de árboles y sus distancias para detectar en qué momento las distancias entre los árboles sucesivos está muy cercana a cero. en ese momento podríamos hablar de uña árbol límite al que teniendo en cuenta la cada vez más exigencia de los aciertos bidireccionales

# Introducción

Los humanos como grupo somos muy buenos para organizar cosas. No necesariamente cosas como armarios o habitaciones, que personalmente confieso me cuestan mucho trabajo. En cambio, a la gente a menudo le gusta agrupar y ordenar las cosas que ve en el mundo que le rodea. Empezando por el filósofo griego Aristóteles, este deseo de clasificar se ha extendido a los muchos y diversos seres vivos de la tierra.

Los sistemas de clasificación más modernos se basan en las relaciones evolutivas entre organismos, esto es, en su filogenia. Los sistemas de clasificación basados en la filogenia organizan las especies u otros grupos de manera que reflejen nuestra comprensión de su proceso evolutivo a partir de sus ancestros comunes.

# Planteamiento del problema de investigación

consiste en analizar una sucesión de árboles filogenomicos construidos a partir del e-value del software blast referente a la exigencia de calidad de los mejores aciertos bidireccionales entre parejas de proteínas de organismos en estudio, mediante la implementación de una distancia entre árboles que nos permita hallar un árbol límite.

# Justificación

Cuando se trabaja con árboles filogenómicos existe la necesidad de tener la certeza de que se obtenga un árbol que está construido con una buena calidad, una forma de lograrlo es utilizando un dato de entrada de calidad a partir del cual se pueda construir un árbol filogenómico. En la presente tesina utilizaremos métodos basados en distancias para construir árboles filogenómicos, por lo que la calidad del dato está en la matriz de distancias que refleja cuán cercanos o cuán lejanos están los organismos en estudio. Si para crear los árboles usamos la calidad de los mejores aciertos bidireccionales (los e-valores en el software BLAST) entre proteinas, tendría que elegirse un valor elevado de exigencia de la calidad en las comparaciones , para asegurarnos que las similitudes no sean debidas al azar y tener en cuenta además una mayor exigencia al considerar las similitudes de cualquier par de proteínas de los organismos en estudio. Las matrices distancias que utilizaremos están soportadas por el conteo de los mejores aciertos bidireccionales entre organismos y se obtendrán tantas matrices distancias como valores de calidad de los mejores aciertos bidireccionales tengamos. En la metodología que se propone se utiliza el concepto de árbol filogenómico límite, el cual es un árbol que se obtiene de una sucesión de árboles obtenidos por aplicar sucesivamente una mayor exigencia a los mejores aciertos bidireccionales. Se espera que aumentando la exigencia de los mejores aciertos bidireccionales pero no permitiendo que el árbol filogenómico deje de ser un árbol, la distancia entre los árboles sucesivos previos a el árbol límite empiece a ser muy pequeña hasta que ésta tienda a cero. Esto indica que se ha llegado a un árbol que, aunque se exija una mayor calidad en los mejores aciertos bidireccionales este no deja de ser el mismo o muy parecido.

# Objetivos

## Objetivo General

Analizar una sucesión de árboles filogenomicos construida a través de los e-values del programa blast utilizando una distancia del tipo diferencia simétrica que nos permita definir en esa sucesión un árbol límite.

## Objetivos específicos

Seleccionar una distancia del tipo diferencia simétrica adecuada para detectar las diferencias entre árboles.

Implementar un algoritmo que calcule esa distancia entre dos árboles usando el formato Newick para cada par de árboles consecutivos en la sucesión de árboles basada en el e-value de blast.

Encontrar el árbol limite, entendiéndose por aquel que a partir de un e-value de exigencia de los mejores aciertos bidireccionales entre parejas de proteínas de organismos en estudio, la distancia entre árboles cercanos al árbol limite tiende a cero.

# Hipótesis

Se piensa que en la sucesión de árboles filogenética se converja a un árbol cuya distancia que a pesar de que la exigencia aumente no cambie su topología (Que su distancia converja a cero)

# Preguntas De Investigación (Opcional si se agregan hipótesis, de no hacerlo se deben incluir)

# Marco Teórico

Este apartado describe los fundamentos y teorías, enfoques teóricos, postulados, conceptos, que sustentan el abordaje y/o construcción de la solución del problema y toda la investigación en general, esto proporciona un carácter de seriedad académica de una verdadera investigación.

# Metodología

Se especifica el tipo de investigación, nivel es una investigación experimental

Describir la base de datos que se tiene y describir a grandes rasgos

* Elegir un tipo de distancia para comparar árboles filogenómicos.
* Establecer correctamente los parámetros en el programa que llama a MEGA para ejecutar el árbol filogenómico correspondiente al método seleccionado y a la matriz dada por datos.
* Elegir un lenguaje de programación
* Diseñar e implementar un algoritmo para el cálculo de las distancias entre dos árboles filogenomicos
* Diseñar e implementar la metodología de encontrar el árbol filogenómico límite en una sucesión de árboles adecuada para encontrar el e-valor en BLAST de los mejores aciertos bidireccionales entre proteínas.
* Diseñar e implementar un experimento para el análisis.

# Desarrollo

# Experimentación y pruebas

# Análisis e interpretación de resultados

# Conclusiones

# Referencias

* Academy, K. (2015). Árboles filogenéticos. Retrieved from https://es.khanacademy.org/science/biology/her/tree-of-life/a/phylogenetic-trees
* Estructura de un árbol filogenético. (Khan Academy, 2015)
* **MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms (Kumar, Stecher, Li, Knyaz, and Tamura 2018).**
* Kerstetter, J. E., O'Brien, K. O., Caseria, D.M, Wall, D. E. & Insogna, K. L (2005) *The impact of dietary protein on calcium absorption and kinetic measures of bone turnover in women*. J Clin Endocrinol Metab (2005) Vol. 90, p. 26-31
* Wilkins, Marc (December 2009). «Proteomics data mining»

# Anexos

Estos deben estar identificados con una numeración (Anexo 1, Anexo 2….. Anexo n) o el alfabeto Anexo A, Anexo B…… Anexo Z.

Los Anexos pueden ser: Diccionario de Términos, Test, Exámenes, encuestas, Calendario de actividades, código, etc.

# Notas Importantes

* *La Tesina se realizará de manera individual.*
* *Referencias bibliográficas (mínimo de 40 que estén citadas dentro del texto de la tesina, incluyendo las de figuras y tablas no propias)*
* *El trabajo deberá tener el aval de algún profesor del Departamento de Ciencias de la Computación que garantice el desarrollo del trabajo hasta su finalización.*
* *Considerar en el documento las convenciones definidas en clase:* 
  + *Tipo de letra 10 puntos y 11 puntos títulos.*
  + *Enumerar secciones y subsecciones del texto. (1, 1.1, 1.1.1, etc.).*
  + *Interlineado 1.5*
  + *Referencias tipo APA o IEEE en el cuerpo del texto, figuras y tablas.*
  + *Nombrar todas las figuras y tablas.*
  + *Márgenes 2 cm. a los lados, arriba y abajo.*
  + *Extensión aproximada de 60 cuartillas considerando todo el documento.*

* *La tesina debe ser liberada por todo el comité tutorial, el cual incluye al (los) directores y los profesores de la materia de Seminario de Investigación II.*
* *Entregar un documento por cada miembro del comité tutoral y uno para el departamento, estos documentos deben estar encuadernados o engargolados.*
* *El alumno debe defender su trabajo ante el jurado, teniendo como fecha límite el periodo de exámenes finales a nivel licenciatura, tomando en cuenta que este debe ser programado con anterioridad.*

**Documento elaborado por la academia de Fundamentos y Teorías Computacionales del Departamento de Ciencias de la Computación, 2019**

**Universidad Autónoma de Aguascalientes**