ESCUELA POLITÉCNICA NACIONAL
VICERECTORADO DE
INVESTIGACIÓN Y PROYECCIÓN SOCIAL

PROYECTOS DE INVESTIGACIÓN (Internos, Semilla, Inter y Multidisciplinarios, Externos):

Área del proyecto: Ciencias Básicas ❑ Ciencias Aplicadas x

FACULTAD: CIENCIAS + INSTITUTO BIOLOGÍA + DEP. BIOLOGÍA PUCE + CENTRO DE MODELIZACIÓN MATEMÁTICA MODEMAT

DEPARTAMENTO: MATEMÁTICA

LINEA DE INVESTIGACIÓN: Modelización Matemática y Cálculo Científico
(verificable en el SAEW)

1 **Proyecto de Investigación**

Título: DESARROLLO E IMPLEMENTACIÓN DE ALGORITMOS PARA LA RECONSTRUCCIÓN DE ÁRBOLES FILOGENÉTICOS.

Resumen del proyecto (máximo 200 palabras)

En el presente proyecto pretendemos constituir un equipo interdisciplinario con investigadores del Centro de Modelización Matemática ModeMat, del Instituto de Biología de la EPN y del Departamento de Biología de la Universidad Católica del Ecuador (PUCE) con el objetivo de desarrollar algoritmos para la reconstrucción de árboles filogenéticos, implementarlos computacionalmente, y evaluar su rendimiento. La investigación de la evolución es un campo multidisciplinario que ha experimentado un notable desarrollo durante las últimas tres décadas, y que involucra principalmente a la biología, las matemáticas, la estadística y las ciencias de la computación. Uno de los grandes objetivos perseguidos en este contexto es la inferencia de los procesos de evolución de los organismos a partir de patrones observables en su ADN. Por ejemplo, conocidos segmentos de la secuencia de ADN para determinadas especies, se busca reconstruir su historia evolutiva y entender los procesos que la gobiernan. La historia evolutiva suele expresarse mediante un *árbol filogenético,* que revela las relaciones de parentesco entre las especies. La tarea de reconstrucción de árboles filogenéticos es compleja, debido a la gran cantidad de datos a procesar y a la enorme potencia de cálculo requerida.

Palabras clave (3-5): evolución, análisis filogenético, algoritmos, cálculo científico a gran escala

4	**Objetivos, hipótesis y resultados esperados de esta propuesta de investigación**

- **Objetivos**
  - Desarrollar e implementar algoritmos para la reconstrucción de árboles filogenéticos.
  - Probar el desempeño de algoritmos nuevos y existentes para la reconstrucción de árboles filogenéticos de cuatro grupos de especies: hormigas, roedores, ranas y lagartijas.
  - Construir un repositorio computacional local con programas para el análisis filogenético.
  - Aumentar las colecciones del Instituto de Biología de la EPN y la secuenciación de los especímenes allí resguardados.
- **Hipótesis**

  Los métodos de paralelización y las técnicas de optimización combinatoria pueden ser empleados para desarrollar algoritmos de reconstrucción de árboles filogenéticos con una mayor precisión y una mejor eficiencia computacional.
- **Resultados esperados**
  - Un repositorio local con programas para el análisis filogenético.
  - Al menos dos ponencias en eventos científicos internacionales.
  - Al menos una publicación en una revista internacional indexada.
  - Dos tesis de grado en el contexto del desarrollo y análisis de algoritmos para la reconstrucción de árboles filogenéticos, y de la modelización matemática procesos evolutivos.
- **Potenciales Usuarios**

  Biólogos, matemáticos, bioinformáticos y otros investigadores interesados en el estudio de la filogenética.

5	**Relevancia de esta propuesta de investigación con los objetivos científicos del departamento y su Línea de Investigación.**

La Modelización Matemática y el Cálculo Científico constituyen una de las cinco líneas de investigación identificadas en el Programa de Investigación del Departamento de Matemática actualmente vigente. Su propósito es el estudio y desarrollo de técnicas computacionales eficientes para la solución de problemas prácticos y la investigación de fenómenos de otras disciplinas. La bioinformática es una de las sublíneas principales, y está explícitamente mencionada en el documento del programa. En este contexto, fue desarrollada en el 2006 una tesis de grado dedicada al estudio de ciertos métodos heurísticos para la construcción de árboles filogenéticos.

6	**Descripción del proyecto, metodología, cronograma de trabajo y justificación del equipo requerido**

- Descripción del proyecto (Máximo una carilla)

La investigación de la evolución es un campo multidisciplinario que ha experimentado un notable desarrollo durante las últimas tres décadas, y en el que convergen la biología, las matemáticas, la estadística y las ciencias de la computación [1,2]. En particular, el estudio de varios problemas concernientes a la evolución ha motivado el desarrollo de áreas de la matemática tan diversas como la combinatoria [3], la geometría [4] y la teoría de probabilidades [5].

Uno de los grandes objetivos generales perseguidos en este contexto es la inferencia de los procesos de evolución de los organismos a partir de patrones observables de variación genómica y del ADN. Por ejemplo, conocidos segmentos de las secuencias de ADN para determinadas especies, se busca reconstruir su historia evolutiva y entender los procesos que gobiernan su evolución. La historia evolutiva es expresada generalmente por medio de un *árbol filogenético,* que revela las relaciones de parentesco entre las especies, así como los procesos por los cuales las diferentes poblaciones son alteradas a lo largo del tiempo: especialización, hibridización o extinción. Los métodos y algoritmos utilizados para la reconstrucción de filogenias se basan en el empleo de diversos modelos (de naturaleza probabilística o combinatoria) que representan el proceso evolutivo en diferentes escalas, desde segmentos de secuencias de ADN hasta genomas completos. Los análisis filogenéticos juegan el rol principal en uno de los más complejos y ambiciosos proyectos que ha abordado la biología en la actualidad: la reconstrucción del Árbol de la Vida [6].

Entre los primeros algoritmos propuestos para la reconstrucción de árboles filogenéticos se encuentran métodos estadísticos de clasificación basados en criterios de distancias, tales como el UPGMA [7] y el método *Neighbour-Joining* [8]. Ambos son aún ampliamente utilizados en algunos contextos, debido a que pueden alcanzar un nivel razonable de precisión sin requerir de una potencia de cálculo elevada. Una mayor exactitud es conseguida con los métodos basados en estimación por máxima verosimilitud, cuya aplicación al análisis filogenético molecular fue propuesta inicialmente por Felsenstein [9]. Al hacer suposiciones explícitas sobre el modelo de evolución subyacente, estos métodos pueden inferir con mayor detalle las relaciones de parentesco entre las especies. Por último, entre los algoritmos más empleados en la actualidad están aquellos basados en el esquema de inferencia bayesiana, propuestos originalmente por Rannala y Yang [10]. Su principal ventaja es la capacidad para considerar incertidumbre dentro del modelo, lo que generalmente se consigue calculando varios árboles filogenéticos alternativos, y estimando sus respectivas probabilidades *a posteriori*. El costo por este aumento en la capacidad de modelamiento es una mayor demanda de recursos computacionales; usualmente los métodos de inferencia bayesiana se implementan como algoritmos de simulación del tipo Cadenas de Markov Monte Carlo (MCMC, por sus siglas en inglés). El paquete computacional *MrBayes* contiene varios modelos basados en este esquema y constituye una de las herramientas informáticas más frecuentemente utilizadas para el análisis filogenético [11].

El desarrollo de nuevos algoritmos para la reconstrucción de árboles filogenéticos, así como su implementación computacional eficiente (a través de técnicas como la paralelización) es un foco activo de investigación en la actualidad. Varias implementaciones están disponibles a través de portales públicos de Internet. Entre ellos, cabe mencionar al *Cyberstructure for Phylogenetic Research – CIPRES*, que permite el acceso a los recursos computacionales de la red XSEDE, los mismos que incluyen a 16 supercomputadores localizados en diferentes laboratorios y centros de investigación en los Estados Unidos de América.

En este proyecto nos proponemos estudiar el comportamiento de diferentes algoritmos para la reconstrucción de árboles filogenéticos, al aplicarlos sobre datos correspondientes a algunos grupos de especies estudiadas por investigadores del Instituto de Biología de la EPN y del Departamento de Biología de la PUCE. En base a sus observaciones, desarrollaremos en el Centro de Modelización Matemática (ModeMat) nuevos métodos que permitan obtener mejores filogenias y optimizar el uso de los recursos computacionales disponibles. Crearemos un repositorio con las herramientas informáticas desarrolladas, así como aquellas libremente disponibles por parte de otros autores, para facilitar su acceso local por parte de los investigadores. Este repositorio podría eventualmente servir como punto de partida para la creación, en el marco de un proyecto posterior de mayor alcance, de un portal web para análisis filogenéticos a nivel de la región andina.

- Metodología y diseño de la investigación (Máximo una carilla)

En el presente proyecto pretendemos desarrollar algoritmos para la reconstrucción de árboles filogenéticos, implementarlos computacionalmente, y evaluar su desempeño comparándolos con otros algoritmos y programas actualmente disponibles. Para las pruebas computacionales, utilizaremos datos correspondientes al ADN de varias especies de hormigas, roedores, ranas y lagartijas actualmente estudiadas en el Instituto de Biología de la EPN y el Departamento de Biología de la PUCE. Elegimos estos grupos de organismos porque contamos en ambas instituciones con los especialistas requeridos para poder evaluar la calidad de los resultados obtenidos por nuestros algoritmos en la reconstrucción de las filogenias.

En una primera fase del proyecto, luego de conformado el grupo multidisciplinario de investigadores, estudiaremos el estado del arte en lo concerniente al análisis filogenético: las necesidades y la problemática desde el lado de la biología, las herramientas computacionales más utilizadas en la actualidad, sus ventajas y limitaciones, así como los modelos matemáticos subyacentes y sus algoritmos de solución.

En una fase subsiguiente, el trabajo se desarrollará en paralelo a lo largo de tres líneas de acción: los investigadores del Instituto de Biología de la EPN, con el apoyo de un auxiliar estudiantil, completarán colecciones de especímenes de hormigas, roedores y ranas, y prepararán las correspondientes muestras para su secuenciación. Se requerirán para esta última actividad Kits de extracción de ADN y polimerasa TAQ. Estos especímenes pasarán posteriormente a enriquecer las colecciones del Instituto de Biología (EPN). Por su parte, los investigadores del Departamento de Biología de la PUCE pondrán a punto las secuencias de ADN de muestras de ranas y lagartijas actualmente disponibles en el museo QCAZ de esa institución. Simultáneamente, los investigadores del ModeMat (EPN), conjuntamente con dos auxiliares estudiantiles contratados para el proyecto, prepararán un servidor de cálculo con los programas actualmente disponibles para análisis filogenéticos, y trabajarán en el desarrollo e implementación de prototipos de nuevos algoritmos. Para facilitar el trabajo de los investigadores será necesaria la adquisición de dos computadores portátiles para el Centro de Modelización, y un computador portátil para el Instituto de Biología de la EPN, así como de software especializado para dar soporte a las tareas de secuenciación.

La tercera fase del proyecto comprenderá la ejecución de pruebas computacionales. El ModeMat pondrá a disposición de los investigadores la infraestructura del Laboratorio Nacional de Cálculo Científico, pero será necesaria la adquisición de un servidor blade original para tener potencia de cálculo que pueda ser dedicada exclusivamente al proyecto. Se generarán árboles filogenéticos para los cuatro grupos de especies (hormigas, roedores, ranas y lagartijas) en base a las secuencias de ADN obtenidas. Estos árboles serán analizados por los correspondientes especialistas, quienes compararán la calidad de las soluciones, con aquellas obtenidas por los programas actualmente disponibles. Asimismo, se evaluará el rendimiento de los algoritmos desde el punto de vista computacional.

Este ciclo se repetirá por una ocasión para permitir a los investigadores del ModeMat afinar los modelos y algoritmos en base a la retroalimentación obtenida por sus colegas biólogos, y para permitir a estos últimos preparar y seleccionar nuevas secuencias de ADN que pudieran ser de interés en una segunda prueba. Finalizado el proyecto, se construirá un repositorio local con todos los algoritmos estudiados, con el objetivo de que el mismo esté a disponibilidad de otros biólogos, matemáticos e informáticos interesados en desarrollar investigación en filogenética.

Justificación del equipo requerido

Como se indicó arriba, el servidor blade será necesario para garantizar la dedicación exclusiva de una determinada potencia de cálculo para las actividades del proyecto. Adicionalmente, el proyecto tendrá acceso a la infraestructura del Laboratorio Nacional de Cálculo Científico, conforme a las capacidades disponibles.

Las tres computadoras portátiles son requeridas para que los auxiliares estudiantiles contratados y los investigadores del Instituto de Biología de la EPN puedan llevar a cabo las tareas de implementación de algoritmos y preparación de secuencias de ADN. Estos últimos necesitarán además de software especializado para poder agilitar su trabajo. Por último, se requiere de dos monitores de alta resolución (uno en el ModeMat y otro en el Instituto de Biología de la EPN) para poder visualizar y comparar adecuadamente los resultados de los algoritmos de reconstrucción de árboles filogenéticos.

Referencias bibliográficas

[1] Gascuel, O. (ed.) (2005). *Mathematics of evolution and phylogeny*. Oxford University Press.
[2] Semple, C. and Steel, M. (2003). *Phylogenetics*. Oxford University Press.
[3] Bandelt, H.-J. and Dress, A.W.M. (1992). A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics*, 92, 47-105.
[4] Billera, L., Holmes, S., and Vogtmann, K. (2001). The geometry of tree space. *Advances in Applied Mathematics*, 28, 771-801.
[5] Steel, M. (1994). Recovering a tree from the leaf colourations it generates under Markov model. *Applied Mathematics Letters*, 7, 19-23.
[6] Tree of Life (2003). Science, 300 (special issue).
[7] Sneath, P.H.A. and Sokal R.R. (1973). *Numerical Taxonomy*. pp. 230-234.
[8] Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4), 406-425.
[9] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6), 368-376.
[10] Rannala, B. y Yang, Z. (1996). Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution*, 43(3), 304-311.
[11] Ronquist, F. Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics, 19(12), 1572-1574.

Cronograma de trabajo anual:
Año 1

Actividad	MESES					
	1-2	3-4	5-6	7-8	9-10	11-12
Investigación del estado del arte	X					
Adquisición de equipos	X					
Colección de muestras		X				
Obtención de secuencias de ADN			X	X		
Recopilación e instalación del software para análisis filogenético		X				
Desarrollo e implementación de nuevos algoritmos			X	X		
Pruebas computacionales sobre las secuencias de ADN obtenidas					X	
Análisis y presentación de resultados					X	X

Año 2

Actividad	MESES					
	1-2	3-4	5-6	7-8	9-10	11-12
Colección de muestras adicionales	X					
Obtención de secuencias de ADN		X	X			
Ajuste de modelos y algoritmos		X	X			
Pruebas computacionales con los modelos calibrados				X		
Análisis y presentación de resultados					X	
Construcción del repositorio para cálculo de árboles filogenéticos					X	
Preparación de informes y artículos con los resultados					X	X

7	**Fecha de inicio**
	1ro de septiembre de 2014
8	**Tiempo dedicación docentes, infraestructura, equipamientos y fondos adicionales.**
	- Tiempos de dedicación semestral del Director de proyecto, de los docentes participantes y otros colaboradores. (Máximo 300 horas por semestre para el Director y 210 horas por semestre para los docentes colaboradores) o Luis Miguel Torres Carvajal (Director): 300 horas o Adrián Esteban Troya Proaño: 210 horas o Pablo Alejandro Moreno Cárdenas: 210 horas o Santiago Rafael Ron Melo: 100 horas - Infraestructura y equipos disponibles para la ejecución del proyecto Recursos computacionales del Laboratorio Nacional de Cálculo Científico administrado por el Centro de Modelización Matemática ModeMat. - Otros fondos de otros organismos (si los hubiere)

9 **Presupuesto estimado para la ejecución del presente proyecto**

Se recomienda que los costos de los equipos, reactivos y materiales de laboratorio, **estén sustentados con proformas actuales:**

Año 1

Lista de ítems (por favor especifique)	**Cantidad solicitada (US $)**
1. Contratación de pasantes	
Contratación de 3 pasantes	
Subtotal	7.500
2. Equipos	
3 computadores portátiles MacBook Air 13" (incluye teclado, trackball y garantías), US$ 2589 c/u	
2 Monitores Thunderbolt de alta resolución US$ 1230	
Subtotal	10.227
3. Reactivos y materiales de laboratorio	
4 Kits de extracción de ADN Genomic x 250rxs. (US$ 821 c/u)	
12 TAQ Platinum Polimerasa (US$ 140 c/u)	
Subtotal	4.964
4. Literatura especializada	
Literatura especializada	
Licencias de software	
Subtotal	4.000
5. Viajes técnicos y de muestreo	
Colección de nuevos especímenes para el Instituto de Biología EPN	
Subtotal	4.000
6. Presentación de ponencias en congresos internacionales	
Subtotal	6.000
TOTAL AÑO 1 (Proyectos Semilla hasta US$ 10.000,00 más IVA) (Proyectos Inter y Multidisciplinarios US$ 40.000,00 más IVA)	**36.691**

ESCUELA POLITÉCNICA NACIONAL
VICERECTORADO DE
INVESTIGACIÓN Y PROYECCIÓN SOCIAL

	Año 2	
	Lista de ítems (por favor especifique)	**Cantidad solicitada (US $)**
	7. Contratación de pasantes	
	Contratación de 3 pasantes	
	Subtotal	7.500
	8. Equipos	
	1 servidor blade (más dispositivos de conexión necesarios)	
	Subtotal	16.000
	9. Reactivos y materiales de laboratorio	
	4 Kits de extracción de ADN Genomic x 250rxs. (US$ 821 c/u)	
	12 TAQ Platinum Polimerasa (US$ 140 c/u)	
	Subtotal	4.964
	10. Literatura especializada	
	Subtotal	
	11. Viajes técnicos y de muestreo	
	Colección de nuevos especímenes para el Instituto de Biología EPN	
	Subtotal	4.000
	12. Presentación de ponencias en congresos internacionales	
	Subtotal	6.000
	TOTAL AÑO 2 (Proyectos Inter y Multidisciplinarios US$ 40.000,00 más IVA)	**38.464**
	TOTAL	**75.155**
10	Nombre: Luis Miguel Torres Carvajal CC: 171212164-7	

DECLARACION DEL JEFE DE DEPARTAMENTO

Esta propuesta ha sido aprobada por el Consejo del Departamento Matemática en Sesión del 20 junio 14 mediante Resolución No. 042. y las instalaciones, incluyendo personal, edificios, equipo y recursos financieros están a disposición del aplicante de acuerdo con las especificaciones que se encuentran en esta aplicación.

JEFE DEL DEPARTAMENTO
Nombre: Dr. Luis Horna Huaraca
CC: 1500110059

_Quito, 20 de junio de 2014_____
(lugar y fecha)