Representación Dinámica de Árboles Filogenéticos y Aplicaciones

Pedro H, Stephane K, Walter S, Alfredo M

18/5/2017

Extracto

En el presente documento se revisará de manera general la **Teoría de la Filogénesis**, su concepción y su uso en la biología, posteriormente se detalla una de las aplicaciones más importantes a nivel científico en el mundo de representación Filogenética de enfermedades epidemiológicas, Finalmente se documentará las diferentes aplicaciones que con este enfoque puede aplicarse en estudios de caso.

1. Introducción

1.1 Un Poco de Historia

La clasificación del mundo natural en categorías significativas y útiles ha sido durante mucho tiempo un impulso humano básico y es sistemáticamente evidente al menos desde el tiempo de la antigua Grecia. Durante cerca de 2,000 años la noción de una "Gran Cadena del Ser", o mejor conocida como *"Scala Naturae"* era una noción dominante, que enfatizó un punto de vista estático de la realidad y representó una jerarquía que partió de la materia y la naturaleza (como las rocas) y se movió hacia arriba a los seres humanos, a los ángeles, y finalmente y más alto de todos, Dios.[1]

Fue hasta la llegada de Charles Darwin, que los esquemas clasificatorios dejaron rápidamente atrás nociones de la *"Scala Naturae"*, además los estudiosos se alejaron lentamente de postular relaciones entre especies basadas en rasgos esenciales presuntos o basadas en similitud física general.

Finalmente el campo de la filogenética toma un giro funcional y más científico en sus intentos de construir una descripción objetiva de las relaciones evolutivas entre organismos basadas en estudios genéticos, moleculares, arqueológicos, históricos y con el propósito específico de explicar, predecir y probar similitudes y diferencias entre organismos.

1.2 Filogenia Biológica

Este término refiere a la historia evolutiva de un grupo taxonómico de organismos. Es conocido que es parte esencial en el estudio científico de la identificación, clasificación, ecología e historias evolutivas de los organismos, la cual es sistemática. La filogenia muestra las relaciones entre grupos de organismos en particular las diferencias y similitudes entre ellos. [2]

En las ciencias biológicas, el modo en cómo se representa dicha história evolutiva es a través de un árbol llamado filogenético, por lo que utilizando la lógica de los diagramas de árbol se buscan estudiar historias evolutivas. La relación entre taxones suele demostrarse mediante datos de secuenciación molecular y matrices de datos morfológicos. Por lo tanto, la filogenia es esencial para comprender la biodiversidad, la genética, las evoluciones y la ecología entre grupos de organismos.

Es importante mencionar los tipos de árboles, ya que pueden ser enraizado o no enraizado. Un árbol filogenético enraizado implica un ancestro común en el que descendían taxones estrechamente relacionados. Un árbol filogenético no enraizado, en contraste, no muestra un antepasado común, sino que la hipótesis sobre el grado de relación evolutiva entre taxones o clasificaciones.[3]

Lo que se entiende es que laa noción de que toda la vida está genéticamente conectada a través de un vasto árbol filogenético es una de las nociones más mencionadas en la ciencia. Qué maravilloso pensar en el antepasado común de los seres humanos y los escarabajos. Este organismo probablemente era una especie de gusano. En algún momento, esta especie de gusano ancestral se dividió en dos especies de gusanos separadas, que luego se dividieron una y otra vez, cada división (o especiación) resultando en nuevos linajes que evolucionan independientemente. Poco sabían estos gusanos, esos cientos de millones de años atrás, que algunos de ellos acabarían evolucionando como escarabajos, mientras que sus hermanos y hermanas acabarían siendo humanos o jirafas.

1.3 Filogenia Computacional

La noción central del presente análisis es el árbol filogenético. Un árbol es un grafo, una estructura matemática que describe cómo se relacionan entre sí, por cualesquiera tipos de relación, los diferentes elementos de un conjunto. Un árbol filogenético es un grafo que describe las relaciones de parentesco, ancestría y descendencia, entre un conjunto de genes, de organismos, de especies superiores.

Formalmente, un árbol consiste en nodos conectados por ramas (llamados bordes o aristas). Los nodos terminales, (hojas o taxones terminales) representan secuencias u organismos de los que tenemos datos, puede tratarse de entidades biológicas extintas o vivientes. Los nodos internos representan a los ancestros hipotéticos y el ancestro de todas las secuencias comprendidas en el árbol es la raíz del árbol, existen también tipos de árboles específicos como el árbol aditivo, de hecho los árboles ultramétricos, que son los que usualmente se usan para representar filogenia biológica.

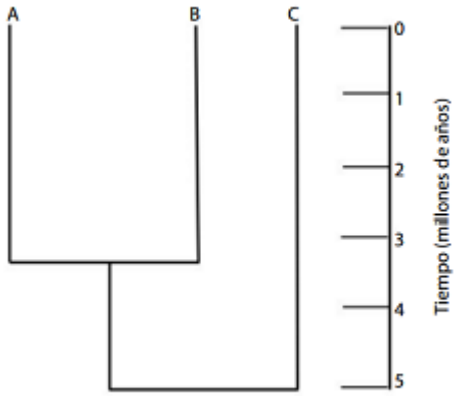

Figura 1. Un árbol ultramétrico que muestra el tiempo de divergencia en millones de años entre las secuencias A, B y C

Los árboles pueden estar enraizados (o polarizados) o no. Un árbol enraizado contiene un nodo del cual descienden, en última instancia, todos los demás nodos, de manera que tiene una direccionalidad que corresponde al tiempo evolutivo: entre más cerca esté un nodo del nodo raíz, más viejo será. En un árbol que no está enraizado no podemos definir el orden de aparición de los eventos evolutivos, por lo que su utilidad es más limitada en el estudio de la evolución. Algunos de los métodos de reconstrucción filogenética generan árboles que no están enraizados y por lo tanto no hacen ninguna diferencia entre los árboles.[4]

Hay toda una serie de softwares y paquetes computacionales que ayudan a la reconstrucción de árboles filogenéticos, fuchos basados en Java, o sencillamente son paquetes de análisis para lenguajes como R, Python, Matlab, etc. Dichos programas están basados principalmente en la alineación de secuencia genética, por lo que podría decirse que “comen” secuencias, de este modo, si se pretende replicar el uso abstracto se deberán tener secuencias de strings de algún modo, para después reconstruir el árbol.[5]

2. Nextrain, Monitoreo de Virus en tiempo real

2.1 Open Science

Antes que nada habrá que entender la importancia de la ciencia reproducible (Open Science), actualmente es formalmente un proyecto conjunto entre investigadores, diarios científicos, los organismos de financiamiento científico e implica al público en general, siendo la idea fundamental la reproducibilidad y la comparabilidad de los trabajos científicos, y bajo esta misma temática se busca hacerse lo más simple posible para participar en debates públicos sobre cuestiones científicas.

Los medios para ello son las visualizaciones interactivas, promociones del discurso a través de una documentación transparente y clara de los resultados, así como el uso de un lenguaje directo y de fácil comprensión, sin mencionar las fuentes claras de información. La provisión de infraestructura de los estudios son beneficiosos para la comunidad y que se le permite participar.[6]

Hoy día existen organizaciones como "The Open Science Prize" que premian precisamente a los mejores proyectos de este tipo anualmente. Nextstrain y sus creadores ganaron en la edición 2017 y esto gracias a la belleza de la aplicación, el impacto de la herramienta y la fácil reproducibilidad.

Seguramente, NextStrain todavía no puede ayudar en la predicción de brotes. A pesar de que ahora podemos observar los virus que mutan, no sabemos qué mutación le dará al virus las cualidades que necesita para iniciar una epidemia. Sin embargo, NextStrain es útil para el seguimiento de la propagación de una enfermedad una vez que ha comenzado.

2.2 Filogenia Dinámica Aplicado a la Epidemiología Genética

El objetivo de este proyecto es promover el intercambio abierto de datos genómicos virales y aprovechar estos datos para hacer inferencias epidemiológicamente accionables. Se desarrolló un marco integrado para la epidemiología molecular en tiempo real y el análisis evolutivo de las epidemias emergentes, como el virus Ebola, y el virus Zika. El proyecto utilizará una plataforma de visualización en línea donde los resultados de los análisis estadísticos pueden ser utilizados por los funcionarios de salud pública para los datos epidemiológicos dentro de los días de las muestras que se toman de los pacientes.[7]

La aplicación se basa en diferentes métodos de análisis distribuidos en módulos: Integración de Datos, Filtrado, Alineamiento y Construcción del Árbol.

3. Otras Aplicaciones Propuestas

3.1 Árboles Genealógicos de Idiomas

De modo general la diversificación y difusión de las lenguas opera de forma similar a la diversificación y difusión de los organismos biológicos. Ambos procesos implican la replicación de códigos que cambian continuamente, dando lugar a nuevas variedades que se diferencian cada vez más de sus progenitores con el tiempo. Como resultado, los "árboles filogenéticos", que muestran descendencia de antepasados comunes, son una característica común tanto de la biología evolutiva como de la lingüística.[8]

Pero hay que aclarar que la lingüística y representar la evolución de los idiomas es un reto difícil de atacar, ciertas teorías hablan que un gran enfoque es visualizar la herencia y la evolución de los lenguajes a través de árboles, ya que tienden a agruparse de forma arbórea de modo natural. Únicamente que se deben hacer una serie de suposiciones para que funcione el estudio aplicado, ya que si bien la noción de "heredar" y conexiones en un grafo es similar, algunas otras ideas no son tan parecidas[9].

De cualquier modo lo que se debería buscar es la definición correcta de las relaciones del grafo, ya que para los árboles aditivos sólo se tiene las relaciones: "es padre de" "es hijo de". Es importante mencionar esto ya que las relaciones en la evolución de los lenguajes es igual de compleja pero en otro sentido, suceden ciertos fenómenos particulares en la lingüística por su propia naturaleza.

- Dos lenguas, no importa cuán separadas, pueden ser combinadas por una generación más joven en una lengua criolla que se convierte en un rico lenguaje y en algunas ocasiones incomprensible a sus 'padres'.
- Un idioma puede convertirse en la fuente del vocabulario y el otro idioma en la gramática.
- Los lenguajes que parten de fuentes muy diferentes pueden acercarse por proximidad, como las lenguas balcánicas.
- No hay nada como la recombinación del ADN en el lenguaje.
- No existe una selección "natural"
- El sexo y la muerte, dos partes primarias del mecanismo de la evolución biológica, difícilmente corresponden con las versiones de lenguaje obvias (superposición de las comunidades de lenguaje del habla y extinción, respectivamente).
- No hay competencia análoga para recursos o compañeros en el lenguaje
- No hay nichos llenos de falta de competencia en el lenguaje.
- Por otra parte una analogía que hace el cambio de la fuente es el concepto "mutación", que para la especiación biológica viene de los errores moleculares de la transcripción en el ADN; En el lenguaje proviene de distorsiones auditivas, errores de producción y errores de procesamiento mental.
- Las mutaciones se dan por selección natural, mientras que la variación lingüística (vista en términos de propiedades gramaticales profundas) está limitada por un sistema de parámetros pero no está sujeto a selección natural

Como resultado, los árboles ramificados de descendencia lingüística son simplemente análogos a los diagramas filogenéticos de la evolución biológica, y no indican el mismo tipo de relaciones. En la actualidad es increiblemente dificil crear un solo árbol de todos los idiomas existentes debido a las numerosas extinciones de idiomas indocumentados.

Las lenguas creadas por el hombre forman una clase distinta y ampliamente independiente de replicadores culturales con comportamiento y fidelidad que puede rivalizar con la de los genes. No obstante lo común entre la evolución biológica y lingüística han sido los métodos estadísticos inspirados por la filogenia, los árboles filogenéticos construidos a partir de elementos lingüísticos muestran la historia de las culturas humanas, los estudios comparativos revelan características sorprendentes y generales de cómo evolucionan las lenguas, incluyendo patrones en las tasas de evolución de los elementos lingüísticos y factores sociales que influyen en las tendencias temporales de la evolución del lenguaje.[10]

La evolución lingüística, a diferencia de la del reino biológico, se mueve rápidamente. En las sociedades no alfabetizadas, las palabras cambian tan rápidamente que después de unos cinco a ocho mil años no se pueden rastrear los cognados suficientes para establecer una relación lingüística. En el mismo lapso de tiempo, las estructuras gramaticales pueden sufrir transformaciones al por mayor, y los inventarios de sonido pueden cambiar drásticamente también.

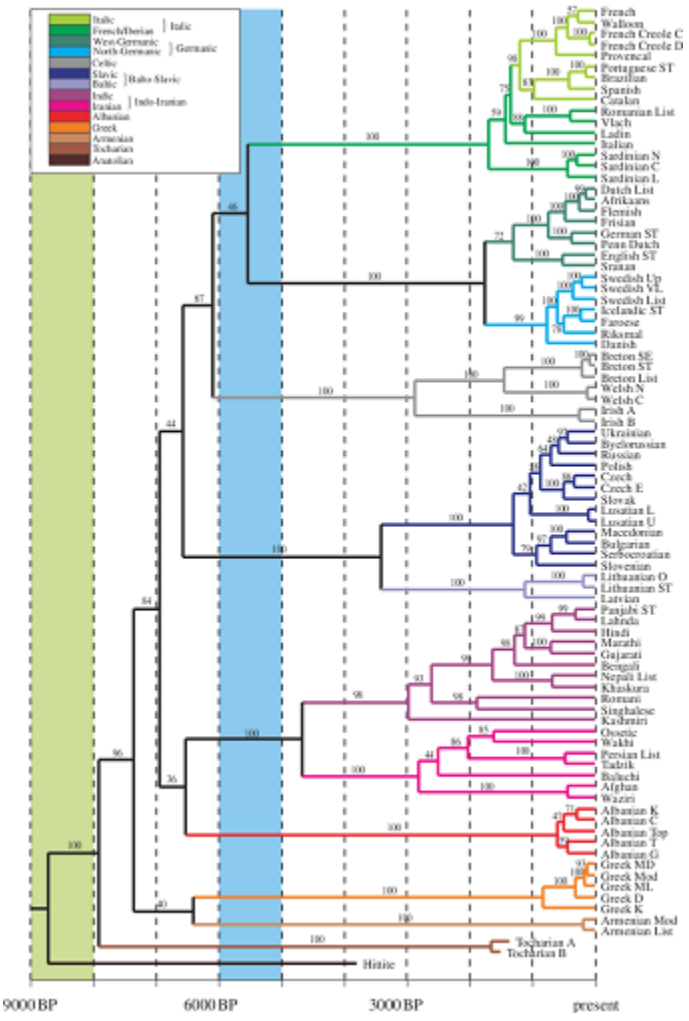

Figura 2. Representación en árbol multimétrico de la evolución de las lenguas indo-europeas

Atacar el problema de la diferenciación lungüistica con un enfoque como el de Nextstrain es de gra reto. La manera de atacarlo sería definiendo las relaciones en el grafo y tener diferentes tipos de "cepas". Los algoritmos para medir cercanía no solo se deben fijar en cadenas de palabras, si no también en estructuras gramaticales, fonética, etc.

3.2 Filogenia y agricultura

Es también conocido que con el advenimiento de las herramientas de la bioinformática en los últimos años, ha sido posible empezar a aplicar las técnicas de secuenciamiento genético aplicado a cosechas agrícolas inclusive. Por ejemplo, los genomas del Berro, el Arroz, el Trigo y el Maíz han sido ya secuenciados y es posible obtener información pública. La realización de análisis comparativo genético ha permitido detectar que la variación genética de este tipo de plantas es menor de lo que se estimaba. A lo largo de milenios el ser humano ha utilizado técnicas de hibridación y selección para resaltar las características de estas cosechas para hacer las plantas más resistentes o incrementar su productividad. Creemos que el aplicar una metodología similar al proyecto Nextstrain sobre el genoma alguna de estas plantas permitiría divulgar la gran diversidad biológica que existe entre especies y en particular resaltar las características fenotípicas de las plantas que mejor se adaptan a cada ecosistema.

Dado que el Maíz fue domesticado en México y que representa el 32% de las cosechas mundiales de grano, creemos que es de particular interés para trabajar con ella. Desde 2009 se cuenta con el mapa completo del genoma del Maíz y existe un gran acervo de información genética pública sobre el tema. Existen diferentes

bases de datos genómicas del Maíz (GRAMENE y MaizeDB) que son actualizadas regularmente y pueden ser utilizadas para nuestro análisis.

En particular existe un conjunto de datos para la realización de análisis comparativo genético del Maíz que contiene información genómica pública de más de 17,000 muestras de la planta. Este conjunto de datos se llama ZeaGBSv2.7. Se cuenta con información genómica con imputación (para datos faltantes) en formato HDF5 con coordenadas AGPv2 y se incluyen también AGPv3 y AGPv4.[11]

3.3 Virus Cibernéticos y su Herencia

Con la gran similitud entre los virus que forman parte de la naturaleza, encontramos en la era moderna los virus informáticos o cibernéticos. Éstos se caracterizan por ser una pieza de software que puede "infectar" a otros programas modificándose. Estos cambios pueden ser injertar en el programa original una rutina que haga copias (replicación) del mismo virus que propicie la propagación del mismo infectando a otros programas ad infinitum. Teniendo esto en cuenta lo que proponemos es mapear el fenómeno de la propagación de un virus como el Zika o el Ébola, a la propagación de un virus informático.

La naturaleza del problema es análoga al problema con los virus de la naturaleza. Los virus informáticos también se propagan, van infectando otros programas o archivos al heredar su información que le permite replicarse, dañar y mutar como a los virus normales. Por eso podemos construir un árbol filogenético que nos permita ver la evolución de un determinado virus en una determinada localidad y cómo ha ido mutando para tratar de evadir los diferentes anti-virus y poder tomar medidas necesarias para evitar su propagación según los intereses de una empresa, una región, un gobierno, un país o hasta in interés de orden global.

4. Reproducibilidad Científica

4.1 Directrices de Reproducibilidad

Primero que nada el científico de datos debe de minimizar la complejidad de la reproducibilidad de un producto de datos, es decir, debe homogeneizar los procesos, debe de considerar herramientas de desarrollo e implementación multiplataforma, en la medida de lo posible usar herramientas open-source y que sean lo menos específicas posible para no limitar el proceso de reproducción.

Planteamos tres fases: **Abastecimiento de datos**, **Transformación de datos**, **Visualización de datos**

En la fase de abastecimiento el científico define con claridad lo relacionado con los datos (fuentes, transformaciones, tipos,etc.). Posibles escenarios:
- Definir qué tipo de datos alimentarán el pipeline - Tipos de fuentes de datos se puede empezar alimentar el pipeline - Debe hacer explícito si existe una fase de transformación de datos de diferentes fuentes hacia una forma homogénea (una estructura de datos común) a partir de la cual se inicializará el análisis de los mismos.

Posterior a esto se define cuáles serán el conjunto de herramientas computacionales que usa para llevar a cabo cualquier preprocesamiento de los datos para su posterior análisis. Por ejemplo:
- El tipo de base de datos usa (Relacionales,NoSQL) (Oracle,MongoDB,Mysql, RethinkDB) - El tipo de archivos (.txt,.csv,.db, fasta, etc.)

En la fase de transformación el científico puede optar por dos opciones:
- Dar a conocer que existe un modelo que realiza un proceso para llevar a cabo un determinado fin (clasificar,predecir,calcular,aproximar,etc.) - Explicar de manera explícita qué objetivo tiene, qué modelo usa para alcanzarlo, si es nuevo o está basado en modelos conocidos.

Al final de esta etapa deberá explicar cuál es la salida, en qué formato está, y si no es explícita su interpretación, abundar sobre cómo se debe de interpretar o inclusive explicar si se debe de llevar a cabo una transformación posterior para su correcta interpretación.

En la etapa de visualización el científico de datos adecua de la mejor manera los datos obtenidos en la etapa de transformación para poder plasmar la interpretación de los datos en un lenguaje visual que haga sentido dentro de un contexto específico.

En este sentido el científico debe delinear muy bien la tecnología que mejor le sirva para su propósito final que es comunicar bien sus resultados.[12]

Lo que se recomienda es usar Docker (más ligero que una máquina virtual y un laboratorio aislado) que proveerá el científico de datos (Dockerfile) con las herramientas y tecnologías necesarias para proveerse sin que los interesados pierdan tiempo en la instalación de las mismas.

4.2 Docker, Ambientes Computacionales Aislados

Al construir una aplicación, o producto científico replicable, algo de lo más importante a considerar es el hecho que pueda ser utilizada prácticamente en cualquier máquina. Lo anterior implica un sin fin de problemáticas ya que ningún usuario tiene exactamente los mismos programas, herramientas, plug-in's, paqueterías, librerías, etc, que ningún otro. Una forma de ver esto es que cada ordenador en el mundo es un ambiente totalmente diferente a cualquier otro que exista, por lo que es imposible que un programa pueda considerar en su instalación, por sí solo, adaptarse a toda la infinidad de ambientes que existen. Es por esto que una herramienta que facilita la noción de Open Science es *Docker*.

Docker no es más que un proveedor de ambientes aislados. Esta herramienta automatiza el despliegue de aplicaciones dentro de contenedores de software. La eficiencia del recurso es algo muy importante ya que permite que contenedores independientes se ejecuten dentro de una sola instancia evitando la sobrecarga que generarían máquinas virtuales convencionales.

Bajo este esquema, docker es capaz de generar una infinidad (hasta donde la capacidad de la máquina local lo permita) de capas de abstracción donde en cada una se instalen los requerimientos con los que funcionará la aplicación respectiva. Todo esto se vacía en un *Dockerfile* para construir una imagen de docker, de tal forma que garantiza la replicabilidad del producto que se esté generando ya que se instalará en el ambiente aislado que se está creando a modo para su correcto despliegue. Las imágenes de docker pueden irse actualizando a través de versiones conforme nuevos datos van llegando o nuevas implementaciones en los algoritmos se van generando quedando disponibles para los usuarios que tengan acceso a las imágenes.

El repositorio por naturaleza para las imágenes de docker es *Docker Hub*. En éste es donde existen las versiones que se van actualizando. Es, a través de esto, que *Nextrain* funciona. El aplicativo funciona con cuatro módulos: **fauna**, **augur**, **auspice**, **janus**.

Cada uno de estos módulos tiene una imagen de docker que se encuentra en *Docker Hub*. La idea con la que trabaja la aplicación es que quien quiera contribuir a ella con nuevos datos o actualizaciones de algún tipo, deberá bajar la última versión del módulo al que se hará alguna contribución, incorporar la aportación y subir una versión actualizada de la imagen de docker en turno. De esta forma cada vez que alguien descargue la aplicación, tendrá la versión más actualizada y podrá visualizarla y utilizarla desde la comodidad de su propio ordenador.[13]

Como podemos observar en el ejemplo de Nextstrain, la ciencia de datos no tiene fronteras y saber explotar sus herramientas es un aspecto vital en el aprovechamiento de la disciplina. El hecho de que un producto de datos de esta naturaleza nos muestra los alcances que tiene esta rama en un ambiente como es la biología, nos confirma que las aplicaciones no están acotadas pero definitivamente un aspecto vital a considerar es el tema de la escalabilidad y poder replicar el ejercicio desde cualquier plataforma para lo cual, generar ambientes aislados es de total importancia, por lo que saber cómo utilizar docker se volverá en uno de los mayores *plus* que cualquier producto de datos puede tener.

5. Conclusiones

Es evidente que la definición y construcción de un pipeline para una aplicación es producto de un correcto y minucioso análisis. En estos momentos han surgido diferentes ideas en diferentes ámbitos que tratan de aportar a la comunidad con propuestas que pueden coadyuvar con soluciones para problemas que importan e impactan a la gran mayoría. Este es el camino que han iniciado y al cual se han sumado personas propositivas

y que marcará el rumbo de una tendencia global. Por otro lado, muchas de estas propuestas pueden inspirar a personas para aplicar la abstracción de esos modelos de solución hacia otros campos del conocimiento.

El crear un producto de datos es un reto que pocos asumen pero que definitivamente puede aportar muchísimo cuando se crea con sentido y con objetivo para solucionar alguna problemática que se encuentre desatendida. Una vez arrancado el proceso de construcción, algo que se debe tomar mucho en cuenta es el trabajar en un ambiente aislado para poder garantizar la replicabilidad del producto. Docker es hoy por hoy una herramienta no sólo es recomendable sino contagiosa y adictiva para este tipo de tareas y más aún si se trata de un proyecto de contribución continua y multilateral, además de un buen control de versiones.

Por último es importante recalcar que combinar la Teoría de Grafos, la lógica que sigue la Teoría de la Filogénesis y las herramientas Computacionales de Ciencia de Datos crean un arma completa muy poderosa que tiene la capacidad de atacar una gran diversidad de problemas como se revisó en el presente documento, es por ello que se deben seguir impulsando este tipo de estudios y seguir haciendo preguntas en diversos campos de la ciencia.

Bibliografía

1 Sultan Haque, O. (August 04, 2016). *"Phylogenetics"*. Retrieved from *https://www.britannica.com*

2 Biology-Online.org (September 18, 2016). *"Phylogeny Concepts"*

3 Maddison, D. R., K.-S. Schulz, and W. P. Maddison. 2007. *The Tree of Life Web Project*. Pages 19-40 Linnaeus Tercentenary: Progress in Invertebrate Taxonomy. Zootaxa 1668:1-766

4 Martinez C., León 2010 *"Reconstrucción de la Historia de Cambio de los Caracteres"*. Pages 89-103

5 Omictools.com *"Phylogenetic network building software tools"* (2017). Retrieved from *https://omictools.com/phylogenetic-network-construction-category*

6 Kasberger, Stefan. *"Was ist Open Science?"* (March 7, 2016). Retrieved from *http://openscienceasap.org/open-science/*

7 Demeure, Y. *"La data-visualisation pour combattre les pandémies"** (March 8, 2017). Retrieved from *http://sciencepost.fr*

8 Baker, Mark C. (2001) *"The Natures of Nonconfigurationality"*. In Mark Baltin and Chris Collins (eds.) Retrieved from *http://www.geocurrents.info/cultural-geography/linguistic-geography/linguistic-phylogenies-are-not-the-same-as-biological-phylogenies*

9 linguistics.stackexchange.com *"What evolution framework best describes the change between languages over time?"** (September 22, 2011)

10 Pagel, Mark. *"Human language as a culturally transmitted replicator"* (2009) Pages 4-14

11 Singh et al. *"Role of bioinformatics in agriculture and sustainable development"* Maize genome mapped. (2009) Nature

12 Stalling,W. *"Network Security Essentials"*. (2015) Pearson

13 *"Docker Overview"* Retrieve from *https://www.docker.com/what-docker*