RECONSTRUCCIÓN FILOGENÉTICA

CEFIRE 2011

Antes de empezar

El objetivo de este tutorial es proporcionar a los alumnos un esquema básico en la reconstrucción filogenética. Vamos a basar este ejemplo en la utilización de una gran base de datos (Ensembl) y la herramienta web Phylemon.

Introducción

El objetivo de la filogenia es reconstruir la historia de la vida y poder explicar así la diversidad de seres vivos que observamos. La forma de representar esta información es mediante un gran árbol genealógico (el árbol de la vida). **El principio que se esconde detrás de la filogenia es que podemos agrupar criaturas en función de su nivel de similaridad**, de forma que asumimos que cuanto más parecidas sean dos especies, más cerca estarán de su antecesor común.

La **filogenética** es un tipo de filogenia que se basa en la comparación de genes equivalentes procedentes de diferentes especies para crear el árbol genealógico de esas especies y poder determinar cual es el nivel de parentesco que hay entre ellas. Los métodos filogenéticos también son aplicables a la hora de estudiar la evolución dentro de una familia génica. La filogenética nos va a permitir descubrir qué se esconde detrás de cada familia de proteínas: mutaciones, deleciones, duplicaciones, especiación, pérdida o ganancia de función, inactivación, etc.

Existen tres grandes razones por las que es necesario el uso de la filogenética:

1. Determinar los parientes más cercanos de un organismo en el que estamos interesados.
2. Determinar la función de un gen mediante sus ortólogos.
3. Trazar el origen de un gen.

Dos genes serán **ortólogos** si provienen de diferentes organismos, derivan de un mismo ancestro común y han sido separados únicamente por procesos de especiación (y no duplicaciones génicas). Teóricamente, los genes ortólogos, a pesar de estar en diferentes organismos, conservan la misma función. Por el contrario, si dos genes han sido generados por un proceso de duplicación génica, diremos que son **parálogos**. Los genes parálogos provienen de la misma familia génica pero es más probable que tengan funciones distintas.

1. Búsqueda de información sobre la proteina de interes

Estamos interesados en descubrir la historia que se esconde detrás del gen de la hemoglobina, más concretamente la subunidad α_1 (HBA1).

El primer paso va a ser obtener cierta información acerca del gen que codifica para esta proteína. Vamos a la página principal de Ensembl (http://www.ensembl.org/index.html) y seleccionamos "Human" como especie a consultar en el recuadro de búsqueda. A continuación, introducimos el nombre de nuestro gen "HBA1" y clicamos en "Go" (Figura 1).

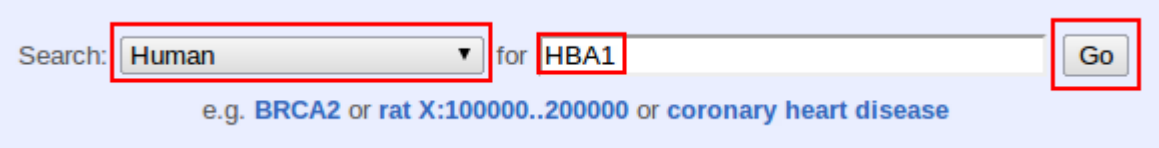

Figura 1: Búsqueda de HBA1 en Ensembl

El resultado de la búsqueda nos nuestra dos genes relacionados (Figura 2):

Figura 2

Nosotros escogeremos "HBA1" (Figura 3).

Figura 3

Directamente se nos redirige a la página de resultados. Con la información que proporcionan, responde a las siguientes preguntas:

BLOQUE DE PREGUNTAS 1

1. ¿Podrías indicar el identificador que Ensembl da a este gen?
2. ¿Para cuántos tránscritos codifica esta secuencia génica?
3. ¿Y para cuántas proteínas?
4. ¿Podrías indicar el número de ortólogos conocidos? ¿Y de parálogos?
5. Indica el valor de presión selectiva bajo la se encuentra el Macaco (*Macaca mulatta*).

2. Alineamiento

Para poder identificar cómo ha evolucionado el gen de la hemoglobina en distintas especies, hemos seleccionado una serie de secuencias homólogas a nuestro gen. En el material del curso para esta sesión encontrarás un archivo llamado "homolog_proteins_hba1.fasta", este documento contiene las secuencias proteicas seleccionadas en formato FASTA. Si echas un vistazo al documento podrás observar que es un formato muy sencillo formado por un encabezado (en amarillo) y la secuencia de aminoácidos (en azul).

```
>HBA1_Humano Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1 PE=1 SV=2
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR
```

En encabezado nos aporta información sobre la especie y la proteína a las que corresponde la secuencia de aminoácidos. Si te fijas en el archivo podrá ver que hemos escogido las secuencias de HBA1 y HBA2 para humano, rata, orangután, rana, chimpancé, lemur, ganso, ratón y un pez.

Tal y como hemos comentado anteriormente, la filogenética se basa en la comparación de genes equivalentes. La forma utilizada para comparar estas secuencias es su alineamiento. La calidad del alineamiento de todas nuestras secuencias es, realmente, lo que acabará definiendo el perfil de nuestro árbol filogenético, y suele ser el paso más complicado. Por suerte, existen herramientas que nos van a facilitar este trabajo.

Phylemon es una herramienta web que permite realizar análisis filogenéticos de una forma sencilla. Nos vamos a ayudar de ella de aquí en adelante. Para empezar, dirígete a su página web (http://phylemon.bioinfo.cipf.es). A la derecha encontrarás un cuadro que te permitirá ingresar si ya eres un usuario registrado o entrar de forma anónima.

Una vez dentro, desde el menú principal, puedes ver las diferentes secciones de la aplicación: "Alignment", "Phylogeny", "Evolutionary tests", "Pipeliner", "Utilities". Selecciona "Alignment".

Phylemon dispone de diferentes herramientas para realizar el alineamiento, nosotros

vamos a utilizar **Muscle**. Lo primero que nos va a pedir el programa son las secuencias que queremos alinear. Tal y como se muestra en la figura 4, desde el apartado “Select your input data” haz clic en “browse server”.

Figura 4

A continuación se abrirá un cuadro de diálogo donde vamos a seleccionar el archivo que queremos y a especificar de qué tipo de información se trata. Para eso, seleccionamos “upload new file” (figura 5) y en “choose file” selecciona el archivo de secuencias “homolog_proteins_hba1.fasta”. A continuación, en “Select format”, vamos a decirle que nuestro archivo contiene secuencias no alineadas, “non-aligned sequences”. Finalmente, indicamos cómo vamos a llamar a nuestro archivo y seleccionamos “Upload”.

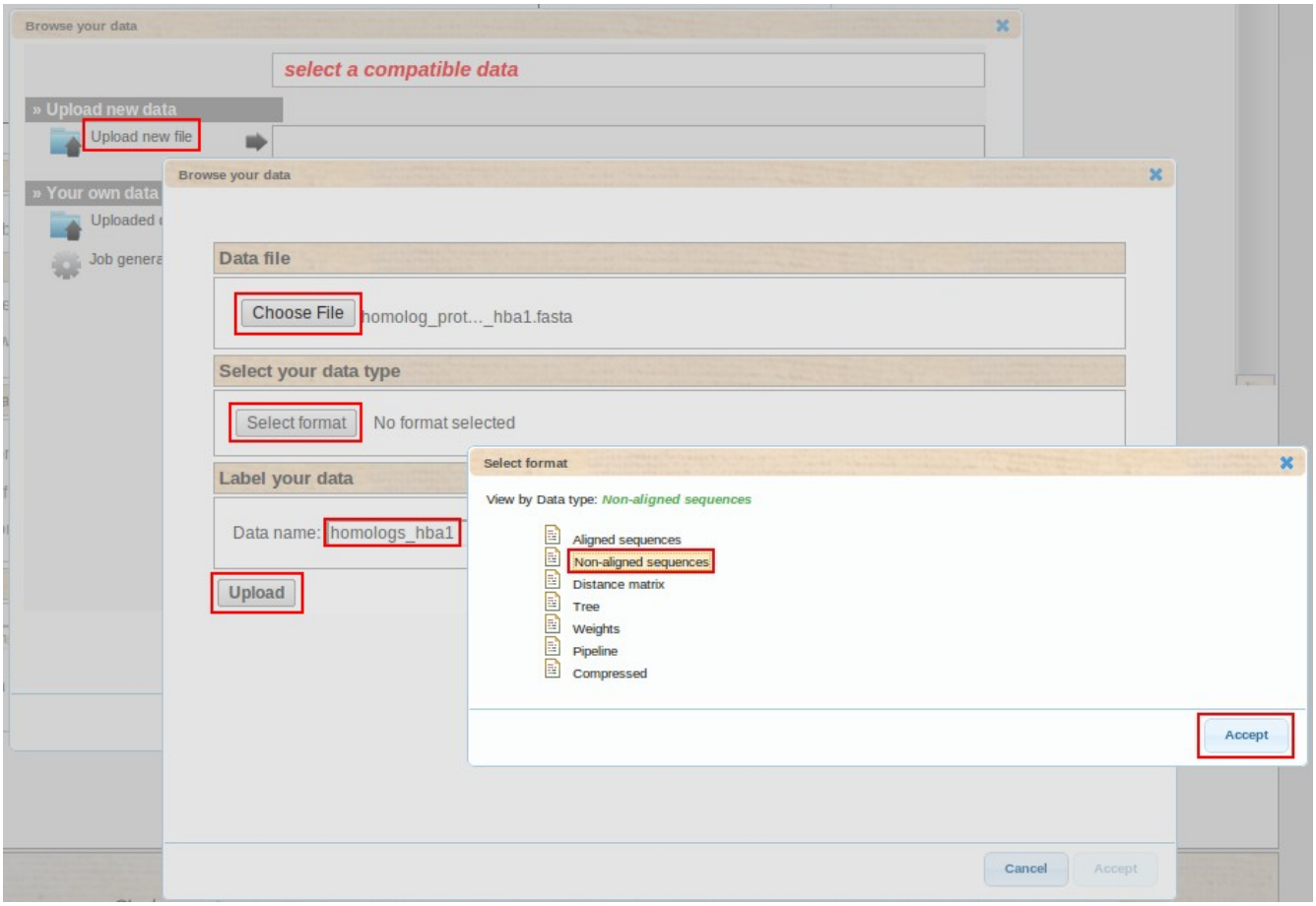

Figura 5: Carga de las secuencias en Phylemon2.0

Ahora que ya tenemos nuestras secuencias en la web vamos a ejecutar el alineamiento.

Vamos a dejar los parámetros por defecto, únicamente daremos un nombre con el que podamos identificar este trabajo y clicaremos en “Run” (Figura 6).

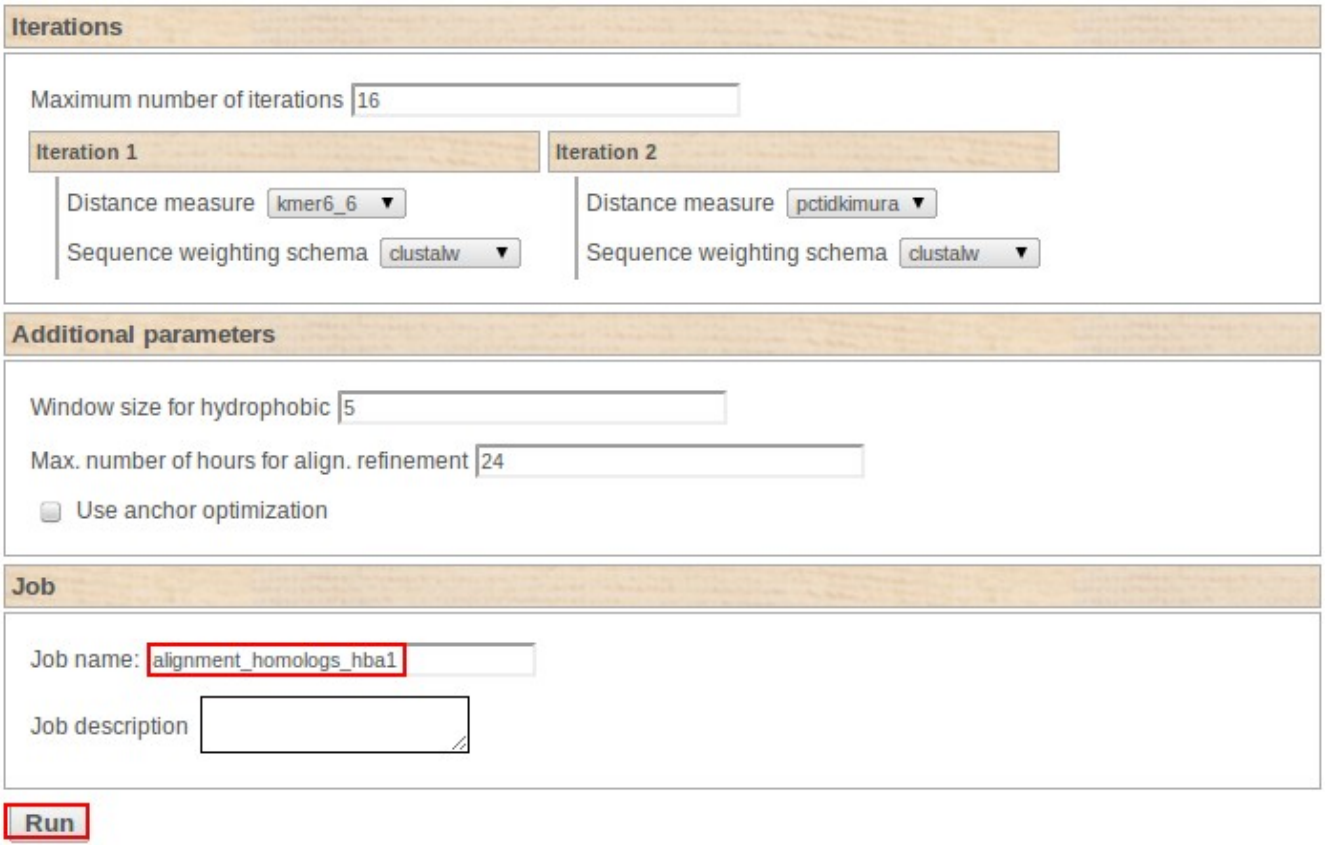

Figura 6: Alineamiento

Si esperamos unos segundos podemos observar el resultado de nuestro alineamiento a la derecha, en el menú “job list”. Cuando haya acabado, ábrelo y clica en “view” tal y como indica la figura 7.

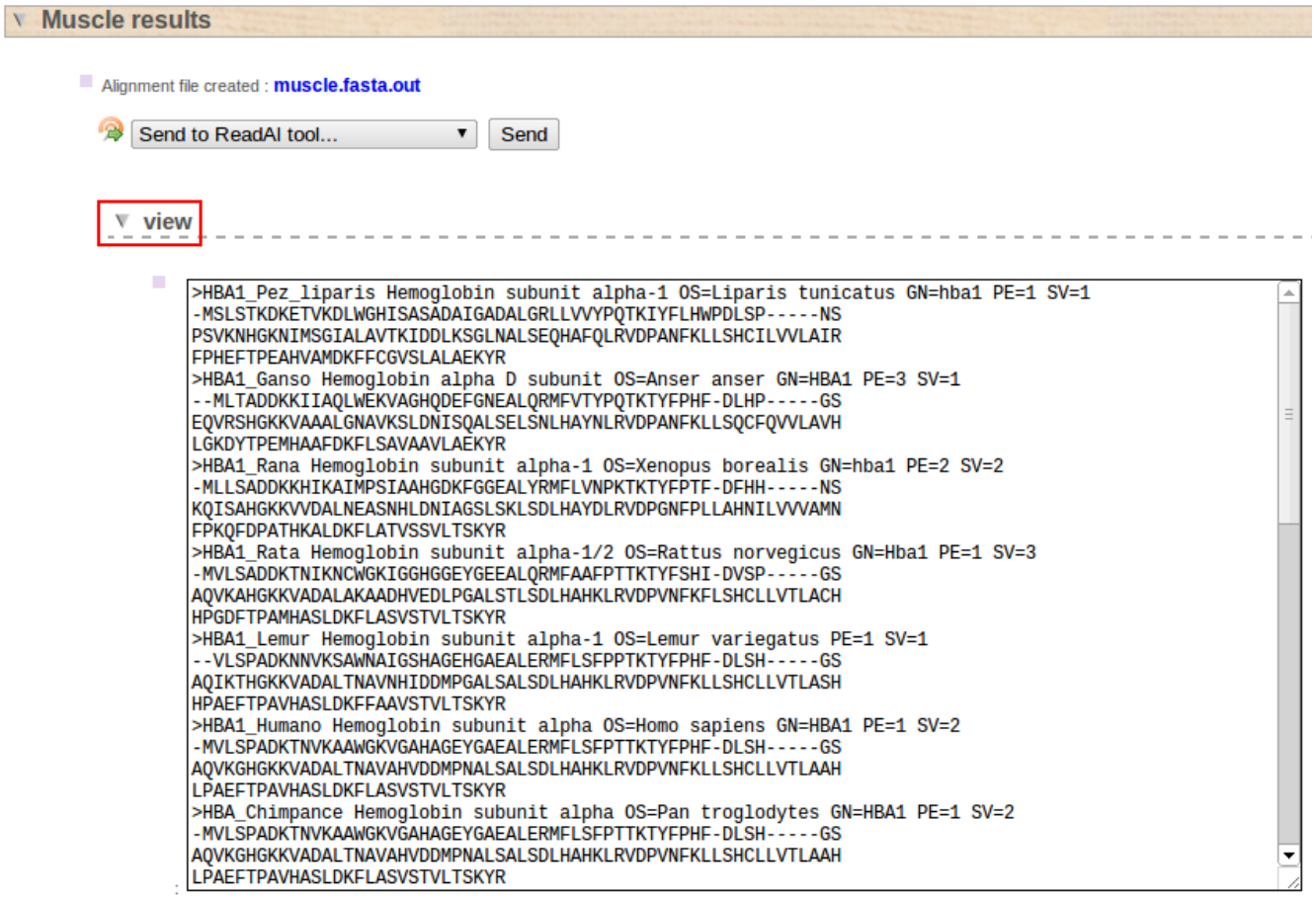

Figura 7: Resultados del alineamiento

Trata de responder a las siguientes preguntas:

BLOQUE DE PREGUNTAS 2

1. El archivo resultante del alineamiento es muy parecido a nuestro archivo original, aun así se trata de un alineamiento. Indica las diferencias que observas.
2. ¿Qué crees que significan?
3. ¿Crees que es importante tener un buen alinemiento? ¿Por qué?

3. Construcción del árbol filogenético

Una vez tenemos los datos alineados, podemos realizar el análisis filogenético. Distinguimos tres grupos de métodos que podemos utilizar para construir un árbol:

- métodos basados en la distancia
- métodos basados de parsimonia
- métodos de máxima verosimilitud *(maximum likelihood)*

Para nuestro ejemplo, vamos a utilizar uno de los métodos de *maximum likelihood* (ML). Los árboles construidos por ML se consideran más precisos que los árboles generados por las otras metodologías, ya que producen la construcción más probable (estadísticamente hablando) que puede explicar nuestro alineamiento. Es decir, que nuestro alineamiento va a ser la historia que explique cómo, desde una secuencia ancestral, una serie de mutaciones ha conducido a toda la diversidad que observamos. El objetivo de ML es intentar que el alineamiento y el árbol expliquen exactamente la misma historia.

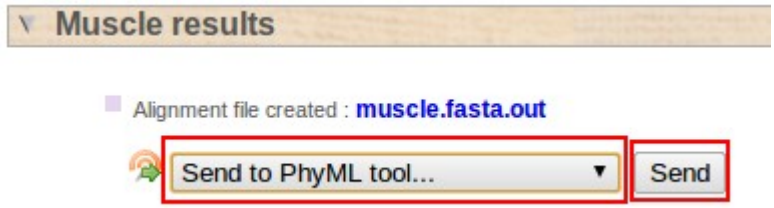

Figura 8: Enviando el alineamiento a PhyML

Para reconstruir nuestro árbol vamos a utilzar PhyML, una herramienta integrada en Phylemon. En la página de resultados obtenidos en el alineamiento anterior puedes observar una pestaña desplegable. Si seleccionamos “Send to PhyML tool...” y clicamos en “Send”, Phylemon se va a encargar de enviar nuestro alineamiento a la herramienta PhyML. Rápidamente, la herramienta nos redirige a PhyML donde podemos ver que ha cargado correctamente el alineamiento (Figura 9).

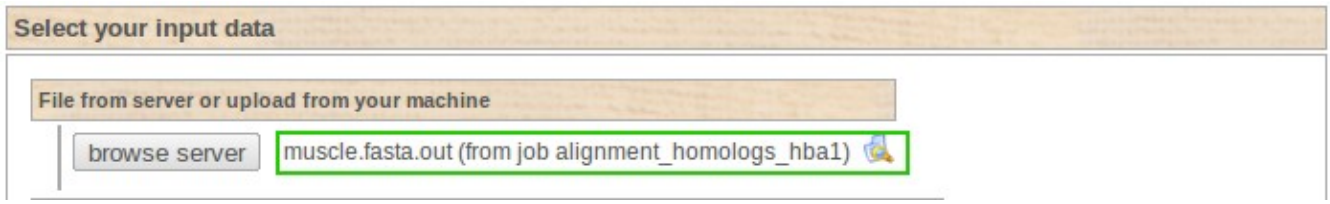

Figura 9

Observa que podemos configurar una gran cantidad de parámetros. Nosotros vamos a trabajar con la configuración por defecto. Únicamente debemos seleccionar “Amino Acid” en el apartado “Data type”, pues nuestras secuencias son de proteínas. Indicamos el “Job name” que queramos y ejecutamos con “Run”.

Figura 9

Espera unos segundos y podrás ver el resultado. Si haces click en “view” dentro del apartado “PhyML Tree” podrás ver nuestro árbol. Como podrás observar, el formato no es muy agradable ni sencillo de entender. Esta forma de anotar un árbol se denomina Newick y es famosa por sus numerosos paréntesis. El formato Newick es la manera más estandarizada para almacenar árboles filogenéticos.

Afortunadamente, existen herramientas que permiten tener una visualización más gráfica de nuestro árbol, en Phylemon podemos encontrar dos: ETE y Archaeopteryx. Así que, de la misma manera que anteriormente hemos enviado los resultados del alineamiento a PhyML, vamos a decirle a Phylemon que envíe nuestro árbol a uno de los visualizadores, para este ejemplo utilizaremos ETE.

Figura 10

4. Visualización del árbol filogenético

Cuando se haya completado el proceso, en la página de resultados podrás observar el árbol filogenético. Vamos a jugar un poco con él.

La raíz del árbol filogenético corresponde al antecesor común más antiguo, lamentablemente, la reconstrucción filogenética no es capaz de decirnos hacia dónde se dirige la evolución, de manera que no podemos determinar dónde se localiza el nodo raíz. Para poder determinar el principio del árbol se utilizan los llamados ***outgroups***. Los *outgroups* son organismos tan distantes a nuestro grupo de interés que podemos utilizarlos como raíz. En nuestro caso, el organismo más lejano a los primates es, probablemente, el pez.

En el árbol que hemos construido sitúate sobre el nodo existente entre HBB_Rana y el resto de HBBs (Figura 11). Haz click sobre ese nodo y selecciona “set as root” (Figura 12). Verás que la topología del árbol ha cambiado. Lo que hemos hecho ha sido fijar ese nodo como raíz del árbol.

Llegados a este punto, trata de responder a las siguientes preguntas:

BLOQUE DE PREGUNTAS 3

1. Recuerda los conceptos de paralogía y ortología. ¿Podrías identificar en el árbol alguno de ellos?
2. ¿Cuál crees que es la evolución que ha seguido el gen de la Hemoglobina?
3. ¿Crees que tiene sentido la agrupación que ha hecho de las especies?

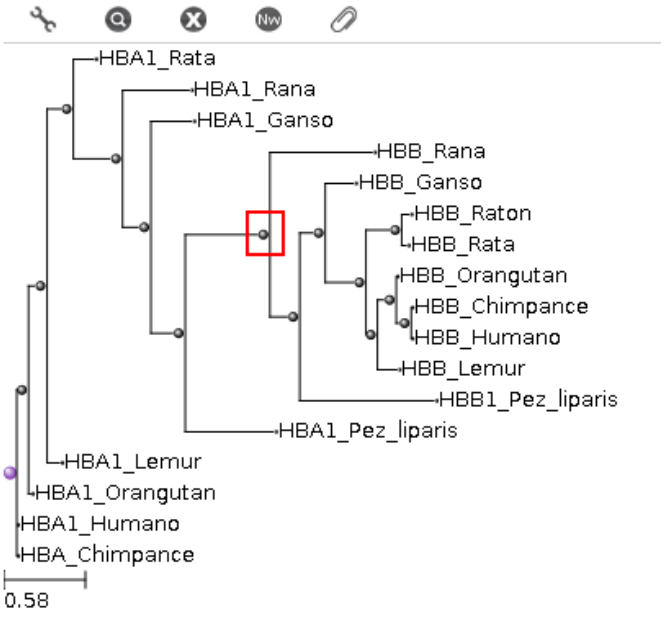

Figura 11

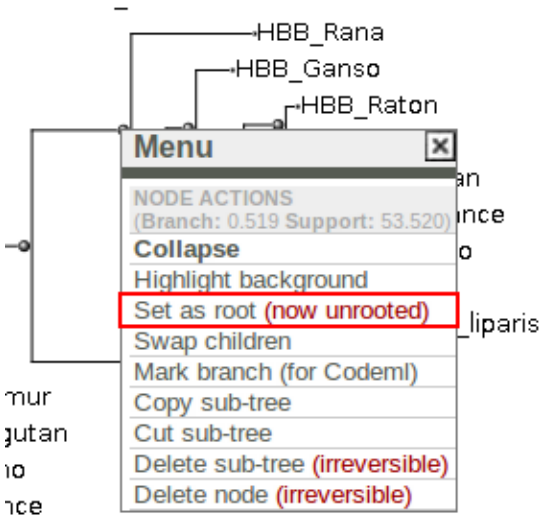

Figura 12