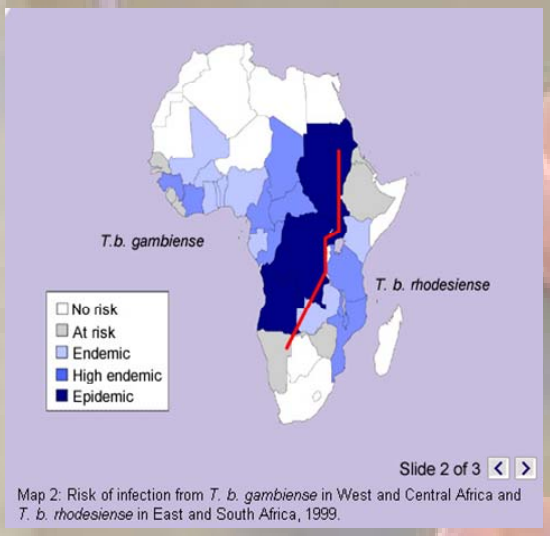
Conversión génica entre genes VSGs y su relación con la arquitectura del núcleo interfásico de *Trypanosoma brucei*



Jenifer García-Montejo
Tutor responable: Dr. Fernando Alvarez-Valin
Sección Biomatemática.
Facultad de Ciencias-UdelaR

Indice

Resumen

Trypanosoma brucei es un protozoario parásito de mamíferos que tiene como vector a la mosca Tse-Tse (genero Glossina), causante de la Trypanosomiasis africana (también conocida como Enfermedad del Sueño en humanos y Nagana en el ganado doméstico) (Stich et al., 2002). Existen varias subespecies de *T. brucei* entre ellas *T. brucei brucei*, *T. brucei rodhesiense* y *T. brucei gambiense*, las cuales difieren en el hospedero y en el tipo de virulencia que tienen en humanos. (Gastellu-Etchegorry, 2001).

Para evadir el sistema inmune de los mamíferos hospederos, estos parásitos han desarrollado una estrategia que consiste en el cambio de sus glicoproteínas variables de superficie (VSGs) de forma tal que el sistema inmune no pueda reconocer al parásito, mecanismo que es conocido como variación antigénica (Berriman. et al.,2005)

Estas proteínas VSGs tienen un porcentaje de identidad menor al 30% entre si a nivel de secuencia y llama la atención que a pesar de ello no varía significativamente su estructura tridimensional. (Zitzmann et al., 1999) (Chattopadhyay et al. 2004). Los mecanismos para lograr la variación antigénica son varios y entre ellos, la conversión génica (Jakson, 2007).

En el presente trabajo, a través de análisis computacionales de las secuencias de genes VSGs, se presentan evidencias que indican que la conversión génica ocurre entre copias inactivas de estos genes. Además, se aprecia una alta frecuencia de eventos de conversión génica entre ciertos *clusters* de genes VSG contenidos en algunos segmentos cromosómicos. Estas frecuencias pueden ser consideradas como una indicacion de la distancia entre los segmentos cromosómicos en el núcleo interfásico. Dicha inferencia se basa en la premisa de que para que pueda existir conversión génica, los genes involucrados deben estar cerca en el espacio, y los segmentos en cuestion presentan en una cierta proporción de segmentos génicos candidatos a haber sufrido eventos de conversión.

Introdución

Trypanosoma brucei es un parásito flagelado perteneciente al orden Kinetoplástida, que se caracteriza por tener kinetoplasto (organelo que se encuentra dentro de la mitocondria asociada al flagelo, el cual posee ADN extracelular que codifica proteínas funcionales para la mitocondria que lo contiene), un único núcleo, citoesqueleto constituído por microtúbulos y fagelos que facilitan su locomoción y adhesión celular.
Fue descrito por Sir David Bruce (Bruce et al. 1914), quien aisló por primera vez los distintos estadíos del ciclo de vida de *T. brucei* de muestras de sangre, así como de glándulas salivales y sistema digestivo de la mosca Tse-Tse, descubriendo así el agente causante de la enfermedad y su vector.

Existen 3 subespecies de *T. brucei*:
T.b. brucei: no infecta humanos, solo infecta ganado y animales de laboratorio, por lo que se ha convertido en modelo para investigación.
T.b. rhodesiense: Se distribuye por la zona este de África (Figura 1) y afecta tanto al hombre como a ungulados salvajes y domésticos. Su vector de transmisión es la mosca tse-tse de los grupos *Glossina morsitans* y *Glossina palpalis.* El reservorio primario son los animales

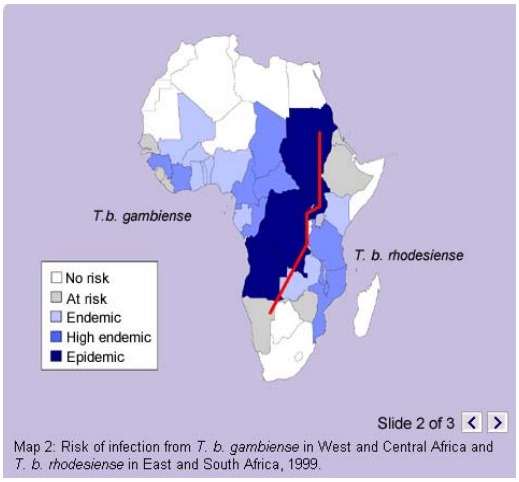


Figura 1: Mapa de zonas de riesgo a contraer tripanosomiasis africana por *T. brucei gambiense* y *T. brucei rhodesiense* (World Health Organization, 2000)

salvajes y el ganado. Este causa la trypanosomiasis aguda, la cual puede llevar a la muerte de las personas infectadas en pocas semanas o meses si no es tratada (Gastellu-Etchegorry M, 2001)

T.b. gambiense: Se encuentra en la zona oeste de África (Figura 1). Sus hospederos son el hombre y ungulados domésticos. A diferencia de *T. b. rhodesisense*, su reservorio primario son los humanos y su vector son las moscas del grupo *Glossina palpalis.* Esta sub-especie causa trypanosomiasis crónica, pudiendo transcurrir años desde la infección hasta la manifestación de los primeros síntomas.

Si bien las tres subespecies de *T. brucei* difieren en los aspectos antes mencionados, el ciclo de vida es el mismo: va variando sus estados y adquiriendo diferentes morfologías que varían en tamaño, forma, localización del flagelo y ubicación del kinetoplasto según el entorno en el cual se encuentran. Cuando una mosca infectada pica a un mamífero hospedero, inyecta trypomastigotas en su forma metacíclica en la piel. Una vez allí, los trypomastigotas se transforman y pasan al torrente sanguíneo, mediante el cual son llevados a diferentes partes del cuerpo. De ésta forma alcanzan el sistema linfático y el fluido espinal, donde continúan replicándose por fisión binaria. El ciclo continúa cuando una mosca se alimenta de un mamífero infectado y los parásitos pasan al intestino medio de la mosca, donde se transforman en trypomastigotas procíclicos, allí se multiplica por fisión binaria. Luego abandonan el intestino, se transforman en epimastigotes y pasan a las glándulas salivales, donde continúan multiplicándose. Una vez allí pasan a la etapa metacíclica, siendo en ésta etapa donde comienzan a expresarse los VSGs (Variant Surface Glycoproteins) metacíclicos. En la Figura 3 se puede ver el ciclo de vida de *T. brucei*.

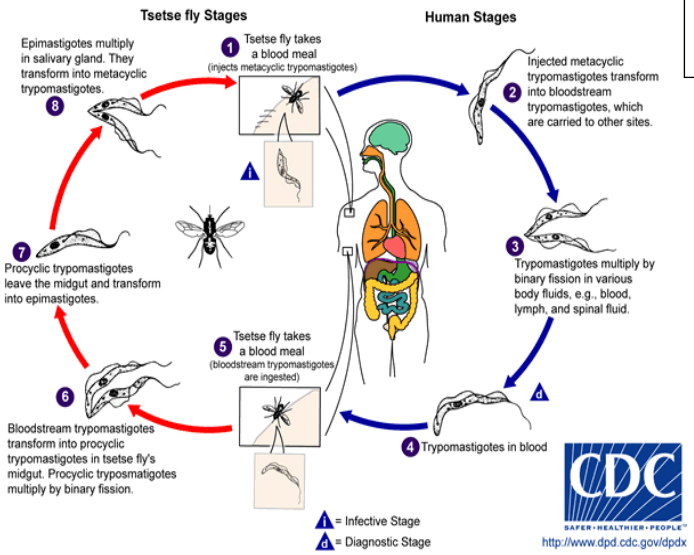


Figura 3: Ciclo de vide da Tripanosoma brucei (Extraido de www.dpd.cdc.gov/dpdx/HTML/ImageLibrary/TrypanosomiasisAfrican_il.htm)

Estos parásitos emplean una estrategia para evadir al sistema inmune de sus hopederos mamíferos que se basa en la variación antigénica (Berriman et al., 2005). La variación antigénica es un mecanismo por el cual los parásitos cambian cíclicamente sus proteínas de superficie que son blanco de los anticuerpos. En el caso de *T. brucei,* estas proteínas son los VSGs. Éstas son codificadas por un gran repertorio de genes que se expresan de a uno por vez, recubriendo toda la membrana de forma tal que solo estas proteínas sean vistas como antígenos. La estrategia de supervivencia frente al sistema immune consiste en que algunos tripanosomas cambian el gen VSG que se está expresando, cubriéndose así con una capa compuesta por un nuevo tipo de VSG para el cual el sistema inmune no ha generado anticuerpos. Cuando el sistema inmune produce los anticuerpos capaces de reconocer a ese tipo particular de VSG, es capaz de eliminar al 99% de los parásitos La tasa de cambio de VSGs es menor al 0.01 por generación celular (Barry et al., 2001), pero esta tasa de cambio es suficiente para que algunos parásitos ya hayan cambiado el VSG que expresaban, el cual no es reconocido por el sistema inmune y una nueva población prolifera, repitiéndose el ciclo una y otra vez hasta que el hospedero muere.

Los VSGs (Variant Surface Glycoproteins) son proteínas de membrana, que recubren toda la superficie de *T. brucei*, anclándose a la misma por un enlace covalente entre el C-terminal de la proteína con entre 3 y 5 azucares, y estos azucares a su vez, con ácido fosfatidil-inositol (fosfolípido de membrana).

El extremo C- terminal., que es más conservado respecto al extremo N-terminal, tiene aproximadamente 100 aminoácidos y forma 4 alfa-hélices. Por su parte el extremo N-terminal es de 300 a 350 aminoácidos, siendo extremadamente variable y se dispone alrededor de las hélices. Todo esto forma una estructura tridimensional que se caracteriza por ser muy conservada, lo cual llama la atención dada la altísima variabilidad de la estructura primaria de los VSGs en general (tienen un porcentaje de identidad aminoacídica menor del 30%) y del extremo N-terminal en particular (Zitzmann et al., 1999) (Chattopadhyay, et al. 2004)

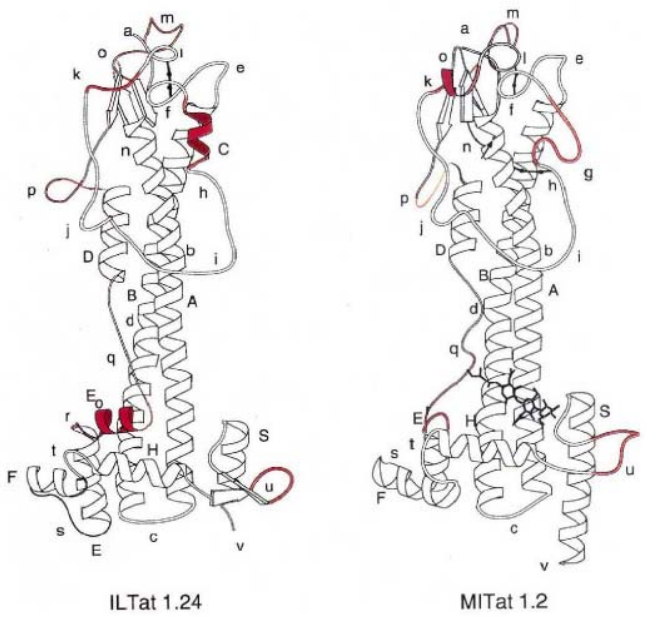


Figura 4: A la izquierda: Estructura tridimensional (inferida por cristalografía) de dos variantes de proteínas VSGs de *T. brucei* (ILTat 1.24 y Mitad 1.2) cuya identidad aminoacídica es inferior al 20%, sin embargo, como puede apreciarse su estructura tridimensional es muy similar.
Abajo: Fragmento de un alineamiento de diferentes proteínas VSGs, ilustrativo del grado de variabilidad que estas proteínas presentan.

```
Tb927.1.5060 1                  --DQQSPPGGLCYEGNTKKNSRGGWNNAGAHRNIWISLKSECIKENGEGLLAESEFQEMKKQLKEKLKEKKDIKISAREGTFY
Tb927.5.110/Tb05.25N21.420      --NRQPDLKDLCYTGNVRKVNSQEWPTTQKHRSTWDDLRSRCITGSGKGVPSETEFHENKVQFRMRIKKRK--NSDGRQ-HFY
Tb09.v1.0290 1                  --DGTSPPGDLCYTGNKERYISQTWRSSSNHKNTWSELQNKCESGRIGVPPTRTEFQDKRKRLETGIKKRK--GSSGRR--YY
Tb09.160.5440 1                 --DGTSPPGDLCYTGNKERYISQTWRSSSNHKNTWSELQNKCESGRIGVPPTRTEFQDKRKRLETGIKKRK--GSSGRR--YY
Tb927.5.130/Tb05.25N21.360      --DQKSTIKIICYAGNVKQTGSARWRNGEGHEKLWKTMQSKCDTGLREGMPTVAEFQEKKKQLKERVRRYT--DTSGRE-HHY
Tb927.3.1470/Tb03.1J15.350      --QNGAGAEDLCFSGNKWYKGGEVWSNKEQAKQHWEKIRLHCNQLPKREQHIQNQLYHLKEKVQAVLRTATE-KKGRESDQNI
Tb927.3.1520/Tb03.1J15.190      --QNGAGAEDLCFSGNKWYKGGEVWSNKEQAKQHWEKIRLHCNQLPKREQHIQNQLYHLKEKVQAVLRTATE-KKGRESDQNI
Tb927.3.2540/Tb03.4808.310      --GNGDHQKNLCYEANAWYDPAESWGNKDRAEESWGKIKQKCTEQINIESIGFESLTTTVQNLRKKLNEMTE-KTT--ERETK
Tb09.v1.0300 1                  --NGKGEMVNLCYESNIWQNGEESWSNGTQAEQHWKSVMEKCSTAIYAARPDAKTILKTMKENLSTKLKERER-GTG--NSETK
Tb927.3.5680/Tb03.2H15.460      --ERRAATNEGCCTGCETGGNGVAWDPRVHSEERWNLLRSKCEKVSSGTPLSSRALSVAESTFFNQLKISK----GSTSGRPT
Tb09.160.5350 1                 --GSAVESDQGCCAECEKGGNSDTWIPSQAGEERWGYLKHQCEKIGSNAELSRKSLSDAMNLLLKRLRPS-----STRGMDLT
Tb927.2.2060/25N14.235 1        --PMPNDRQQACCEGCSTGANEDPWKPVEQSNTRWKHLAQRCKSIGTNEKLSSSSLSRATKVFLEVLKRST----AQKEGTKT
Tb927.3.1500/Tb03.1J15.250      --MNLGRETQLCCSRCARGDNRAVWRPDEDAGERWYFLKEQCSSMEGPQEGFNKLVNDFRDMINHGVDGCT----GTTYVFGD
Tb927.3.1510/Tb03.1J15.200      --MNLGWKAHLCCKECVRNQNEGMWETNKNATGRWHFLKKQCSSMEGPQEGFNKLVDNFRKATYGGEDECR----GTAYTLGG
```

A pesar de la alta variabilidad que presentan, se pueden clasificar los VSGs según el tipo de N y C-terminal que poseen, siendo el criterio de clasificación, la posición de determinadas cisteinas y del anclaje GPI. (Carrington et al., 1991) (Blum et al., 1993).

El repertorio de genes VSGs consta de 2 subsets: el primero constituido por más de 200 copias que se encuentran en los telomeros de mas de 100 microcromosomas especializados; el segundo subset consta de genes VSGs ubicados en tandem arrays en las zonas telomericas y subtelomericas dentro de los 11 cromosomas de tamaño superior a 1 Mb que compone el genoma de esta especie. Dentro de éste último grupo, se estima que son más de 1000 el número de copias, de las cuales, un 5% son genes completos, 9% codifican para VSGs con plegamiento atípico, el 17% son fragmentos de genes, y el 70% restante son pseudogenes

Con relación a la regulación de la expresión de los genes VSG y la variación antigénica cabe decir que se transcriben aquellos genes codificantes de VSG, que están ubicados en los llamados sitios de expresión telomércos activos (Figura 5) Bloodstream Expression Site – BES (en el caso de que el sitio se activa en el hospedero) o Metacyclic Expression Site – MES (en el caso de que se active en la mosca Tse-Tse). Existen al menos 20 BESs, con un gen VSG cada uno, pero solo es activo un BESs por célula (y por ende, un solo gen VSG). En las etapas tempranas de la infección, el mecanismo de variación antigenica consiste en cambiar el BES activo, siendo los mecanismos de activación y desactivación de los BESs

diversos y no todos completamente conocidos**.** (Vanhamme et al. 2001). En la figura 6, se muestra el mecanismo por el cual es regulada la actividad de los BES.

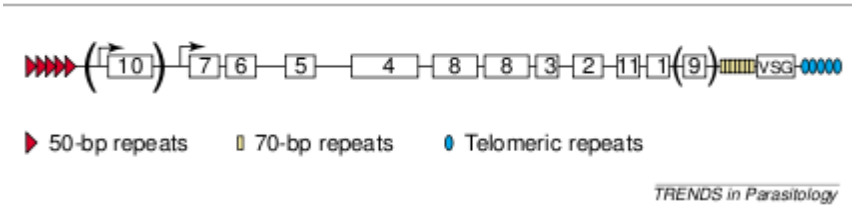


Figura 5: Esquema de un Bloodstream expresión Site. Consta de secuencias repetidas de 50pb seguidas de unidades policistrónicas bajo control de un único promotor, compuesta por un VSG flanqueado por repetidos de 70pb y repetidos telomericos, y hacia el extremo 5' del VSG hay al menos, 8 genes adicionales (Expresión site associated genes – ESAGs). (Vanhamme et al. 2001)

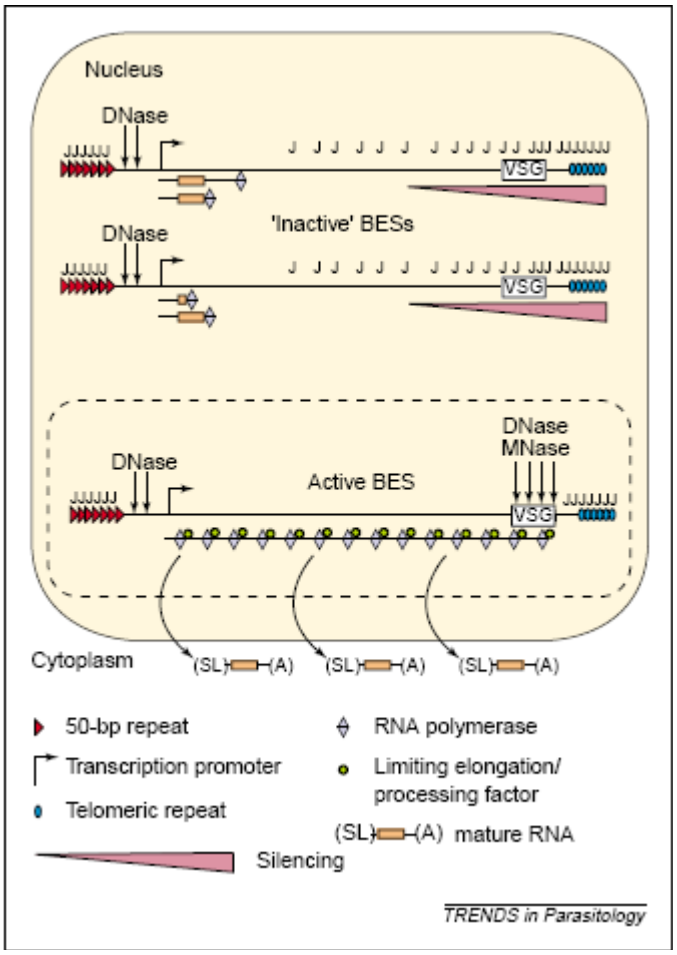


Figura 6: Regulación del BES activo.
Solamente un Bloodstream Expression Sites (BES), se encuentra activo por vez. La regulación de la expresión se realiza reprimiendo los restantes BES de forma reversible, por medio de silenciamiento telomérico.
Este mecanismo consiste en que si bien en todos los sitios BES comienza la síntesis de ARN, solo se continúa en la copia activa. En el caso de los BES inactivos, los transcriptos quedan atrapados en la cromatina. El cambio de BES que se está expresando se daría cuando existe trascripción efectiva de un BES inactivo, por lo que, mediante un pequeño período de tiempo, se da una expresión intermedia, para luego, activar el BES inactivo y desactivar el BES activo anteriormente (Chaves et al., 1999). En el dibujo, J indica la presencia de la base J- [β-D-(glucosil) hydroxymethyluracil], las flechas indican sitios sensibles a nucleasas. (Vanhamme, et al., 2001)

En las etapas tardías de la infección, además del cambio de actividad de los BESs, existen diferentes mecanismos, entre ellos la conversión génica (Jakson, 2007), por los cuales un gen activo de VSGs (lógicamente dentro de un BES activo) es remplazado por otro gen VSG que hasta el momento era inactivo. Éste es un proceso mediante el cual un gen "convierte" al otro, lo que tiene como consecuencia que el segundo se transforme en una copia del primero. Cabe resaltar que la conversión pueda abarcar todo el gen, o que la copia sea parcial, o sea, que un fragmento del primero convierta a un fragmento del segundo

("segmental gene conversion"), generando así, un gen nuevo de tipo mosaico. (McCulloch et al., 1996) (Figura 7).

Éste mecanismo de conversión génica ha sido bien documentado para el caso en el cual una o varias copias inactivas de genes VSG convierten al gen activo que se encuentra localizado en el BES (Barry et al., 2005)

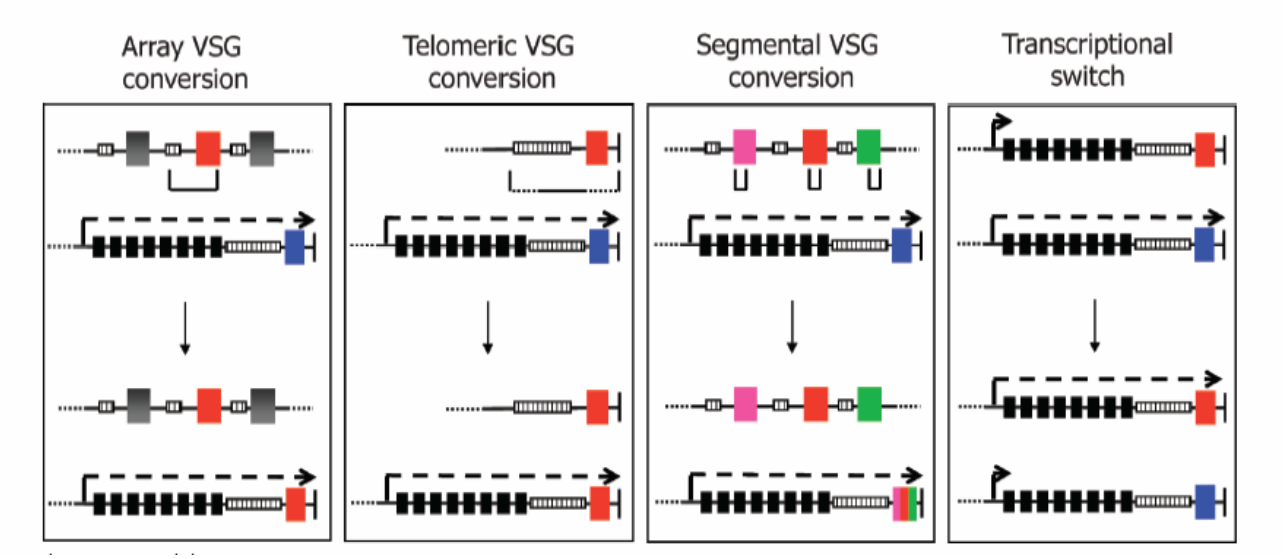


Figura 7: Diferentes mecanismos de cambio de VSG durante la variación antigenica. En las primeras etapas de la infección, el mecanismo consiste en cambiar el BES activo (Transcriptional Switch) así como sustituir un VSG por otro mediante recombinación telomérica (Telomeric VSG conversion). Ya en las etapas tardías, se puede dar que una o varias copias inactivas conviertan a una copia activa (Segmental VSG convertion), asi como también se da por transposición duplicativa (Array VSG convertion) (Stcokdale et al., 2008)

Hipótesis y Objetivos

En este trabajo pusimos a prueba la hipótesis sobre si existe conversión génica entre las copias inactivas, en otras palabras, si las copias inactivas del repertorio antigénico también interactúan “convirtiéndose” unas a otras o fragmentos de las mismas (Figura 8).

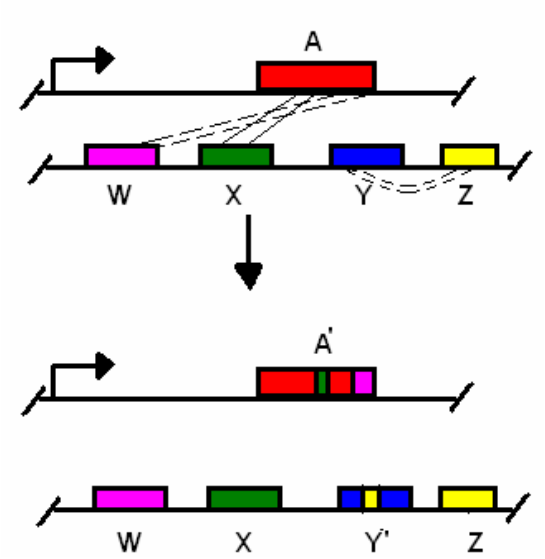


Figura 8: Tipos de conversión génica que se está buscando evidencia en el presente trabajo. En la figura se muestra como un gen A, que se encuentra en un expression site, es convertido por los genes W y X, lo que se denota como un cambio de A a A’. Ademas se muestra un como un gen Y es convertido por un gen X, siendo Y’ el gen Y modificado. En el presente trabajo se busca evidencia de eventos de conversión génica entre copias inactivas.

Nuestra estrategia consiste en identificar regiones candidatas a haber sufrido conversión génica sobre la base de que éstos serían segmentos con alta identidad rodeados de regiones génicas con baja identidad.

Por otra parte, los segmentos génicos que están sufriendo conversión génica deben interactuar físicamente, lo que a su vez implica que éstos deben encontrarse cercanos en el espacio. En caso de que se encuentre evidencia de segmentos que hayan participado en eventos de conversión génica, permitiría hasta cierto punto, inferir la ubicación relativa que tienen éstos genes (y los segmentos cromosómicos que los contienen) en el núcleo interfásico. Nuestro análisis de las interacciones “físicas” de las copias de los genes VSGs arrojaría luz no solo en la dinámica evolutiva de éstas, sino también en la organización espacial del núcleo interfásico de los tripanosomas africanos.

Materiales y métodos

Base de Datos

Las secuencias de genes VSG fueron extraídas de tres bases de datos: GeneBank, GeneDB y VSGDB, las cuales difieren entre si dado que es un genoma que está en proceso de secuenciado y anotación, lo cual se está llevando a cabo, por un lado, Barry, Marcello y colaboradores en el Instituto Sanger, y por otro The Institute for Genomic Research (TIGR). Para realizar el presente trabajo, se armó una base de datos local con aquellos genes VSG, que fueron anotadas antes del 13/10/2007 provenientes del VSGDB y se completó la información con datos del GeneDB y GeneBank.

Para este fin, se hicieron consultas on-line en la base de datos VSGDb, pidiendo las secuencias de proteínas de todos los VSGs anotados hasta el momento, como se ve en la Figura 9.
La página que se presentó como resultado de dicha consulta se guardó entera dado que el sistema no permite guardar el resultado como un archivo de texto.

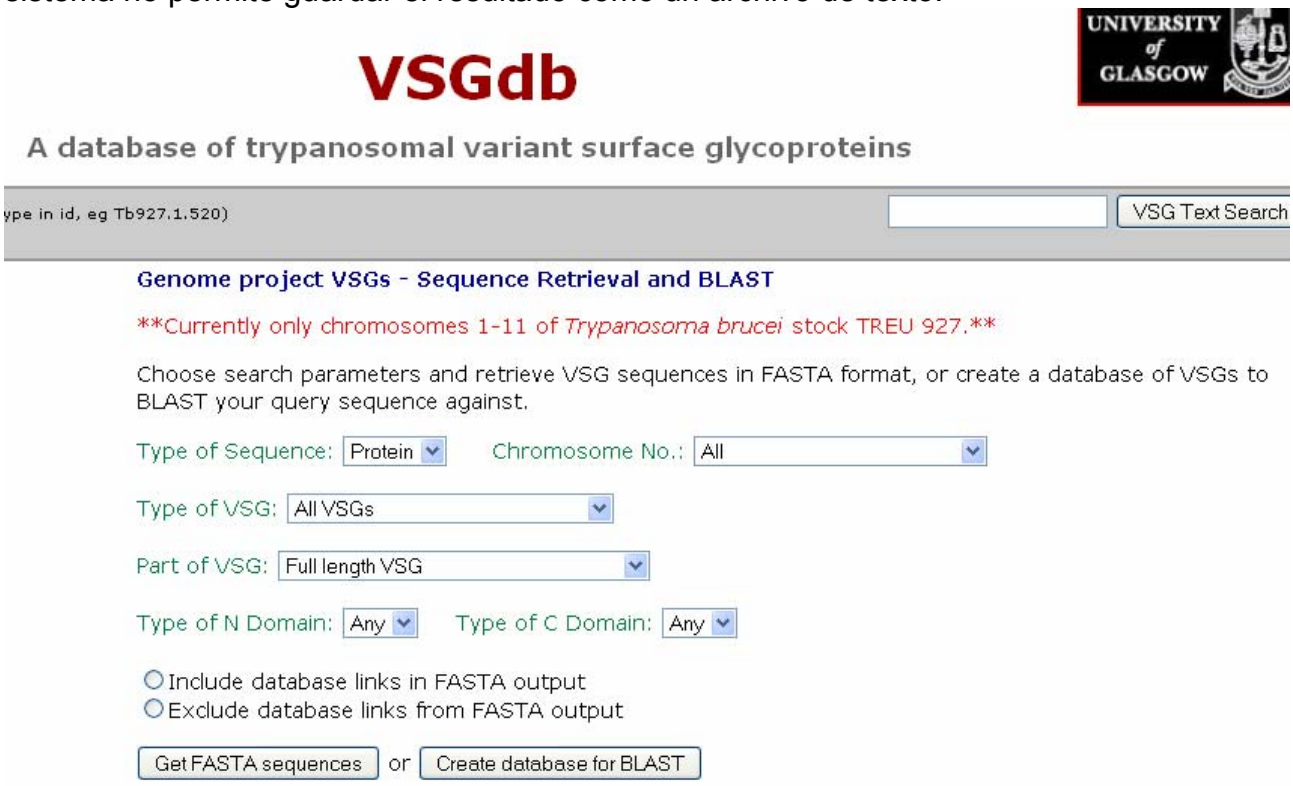


Figura 9: Página de VSGdb en donde se clickean las opciones para obtener los datos deseados. Para el trabajo, se seleccionaron en el menú desplegable, que diera como resultado las secuencias enteras de todos los tipos de VSGs, de todos los cromosomas y luego se pulsó “Get FASTA sequences”.
Por otra parte, se pidió lo mismo, pero que diera los resultados para cada uno de los cromosomas.

Luego se extrajeron los caracteres particulares de los archivos .xml (que son los que hacen que un navegador de internet cualquiera los pueda abrir) así como también los detalles de los dibujos y links de la página, para extraer las secuencias en texto plano con formato multifasta, siendo éste archivo modificado, el que se utilizó para hacer las consultas de Blast.

Software empleado

FormatDB

Este programa permite convertir un archivo fasta o ASN.1 en una base de datos que podrá ser utilizada por Blast instalado localmente.

Blast (Basic Local Alignment Sequence Tool)

Éste es una familia de programas desarrollados por Altschul (Altschul et al., 1990) en base a un algoritmo cuya finalidad es cotejar una secuencia o un conjunto de secuencias aminoacídicas o de ADN con una base de datos y encontrar secuencias homólogas mediante métodos heurísticos de búsqueda.

En términos genéricos el algoritmo que usa Blast es como sigue. En primer lugar, busca coincidencias exactas de "palabras" cortas entre la/s secuencia/s *query* y la base de datos. El tamaño de estas palabras (W) por defecto es 11 letras para secuencias de ADN y 4 para aminoácidos. Luego, trata de extender el *match* en ambas direcciones con el fin de aumentar el *score*del alineamiento. En este paso no se consideran deleciones o insersiones.

El *score*del alineamiento es determinado por una matriz de puntaje, la cual es una medida de la de la probabilidad relativa de que ambos aminoácidos (o un nucleótido según el caso) enfrentados porque provienen de un ancestro común (y se haya dado el cambio de un aminoácido a otro o se hayan mantenido conservados) versus la probabilidad que estos dos aminoácidos estén enfrentados al azar. Existen varias matrices de puntaje que se pueden utilizar en Blast, entre ellas las matrices PAM y las BLOSUM para proteínas.

Si se encuentra un alineamiento así generado, con un *score*alto, se hace otro alineamiento entre la secuencia *query* y la secuencia de la base de datos con la que dio un resultado positivo. Esto se hace usando una variante del algoritmo de Smith-Waterman. Los alineamientos estadísticamente significativos, son los únicos que se muestran en el resultado final.

Las matrices PAM, desarrolladas por Margarett Dayhoff (Dayhoff et al., 1978), fueron construidas basándose en alineamiento de secuencias de proteínas estrechamente relacionadas y reconstrucción filogenética a partir de dichos alineamiento. Usando estos datos se calculó la probabilidad de cambio de un aminoácido a otro. Luego dicha matriz de probabilidad se re-escala de forma tal que corresponda a 1% de divergencia por iteración. Esta matriz, conocida como PAM1, es una unidad de divergencia. Las restantes matrices PAM se obtienen con potencias de la matriz PAM1, por ejemplo la PAM250 es PAM1 a la 250. Un segundo tipo de matrices, conocidas como matrices BLOSUM, fueron desarrolladas por Steven Henikoff y Jorja G. Henikoff (Henikoff et al., 1992). Estos autores usaron regiones conservadas de familias de proteínas (sin gaps en los alineamientos), a partir de los cuales se calculan las frecuencias relativas de aminoácidos y las probabilidades de sustitución entre los distintos aminoácidos.

En el caso de alineamientos de secuencias de ADN, si bien se puede especificar una matriz de puntaje, por defecto lo que el programa hace es asignar un valor de puntaje igual a 1 cuando se presenta un *match* o 0 en el caso de miss*match*es.

Existen variantes de Blast dependiendo de que sea lo que se quiere comparar.

- Blastn: compara secuencias nucleotidicas.
- Blastp: compara secuencias proteicas.
- Blastx: usa una secuencia nucleotidica como *Query*, la cual traduce en sus 6 marcos de lectura, y luego compara con una base de datos de proteínas.
- TBlastx: compara proteínas traduciendo las secuencias de ADN *query* en los 6

marcos de lectura y compara esto con con una base de datos nucleotidica, previa traducción de las mismas también.
- TBlastn: compara secuencias proteicas con una base de datos traducida en sus 6 marcos de lectura.
- Psi-Blast: Lo que hace es buscar dominios en una secuencia aminoacidica *query* en una base de datos de dominios de proteinas. La diferencia de ésto con Blastn es que en vez de usar las matices de puntaje antes mencionadas, emplea otro tipo de matrices armadas teniendo en cuenta los aminoácidos que se encuentran en los dominios proteicos más comunmente.

Una vez obtenidos los alineamientos, BLAST calcula el *expected value* (e-value) para cada uno de ellos, siendo el e-value de una alineamiento de largo L con un *score*S, la cantidad de alineamientos que pueden encontrarse con el mismo largo L y un *score*igual o mayor que S si se toman dos secuencias de la base de datos al azar.
Se puede correr Blast de 2 maneras: a través de un servidor web, o en forma local.
En la primer forma, se brinda a la página web una secuencia de ADN o aminoacidíca, y esta es comparada con la base de datos del servidor, siendo el ejemplo más conocido, la página el NCBI.
En el segundo caso, se corre Blast utilizando una base de datos local. En éste trabajo se emplea ésta segunda modalidad. Esto se puede hacer tanto en windows como en cualquier sistema operativo tipo Unix (Linux, Mac o Solaris) y se puede bajar del sitio de NCBI www.ncbi.nlm.nih.gov/Blast/download.shtml, incluso algunas de las distribuciones de Linux (Debian, por ejemplo) ya lo incluyen en sus paquetes. En este caso se usó en Linux (distribución de Fedora 7 y Debian Etch).

Para éste trabajo se alinearon las secuencias de ADN de VSGs usando como base de datos las mismas secuencias (o sea se hizo Blast de una base de datos contra si misma) con Blastn, empleando un *valor umbral* de e-value de 3,0 x e^{-10} y usando opciones específicas para que no enmascare sitios de baja complejidad.

BioParser

Es un programa desarrollado en BioPerl capaz de transformar los datos brindados en una salida de Blast en una base de datos MySQL, y una vez hecho esto puede usarse para filtrar dicha información según consultas que se hagan a dicha base de datos on-line, y presenta los resultados ordenados en forma de tabla, usando para eso un servidor Apache.
Luego, esta tabla fue procesada por un script en Perl que lo que hace es eliminar los autohits y aquellos datos que se encuentran repetidos, dado que se comparó una base de datos contra si misma.

Para éste trabajo, los datos que se usaron fueron aquellos en que el porcentaje de identidad entre las secuencias fuera mayor a un umbral dado, digamos 70%. En la tabla se muestran los nombres de las secuencias que presentaron *match*, sus longitudes, la hebra involucrada, el porcentaje de identidad entre las secciones que alinearon, el porcentaje que representa las secciones que alinearon en sus respectivas secuencias, y las coordenadas de inicio y fin de las secciones que presentaron *match*.

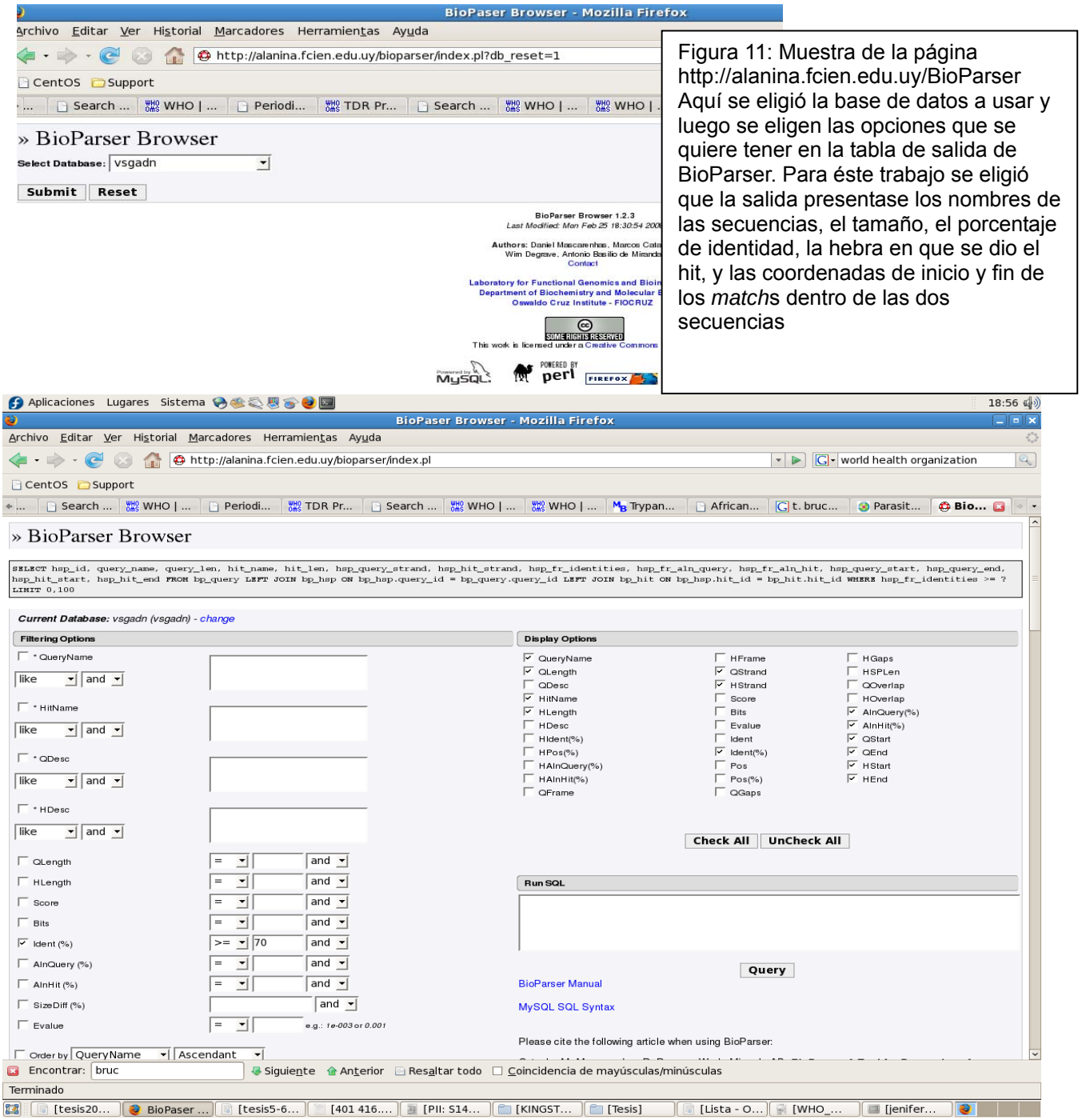


Figura 11: Muestra de la página http://alanina.fcien.edu.uy/BioParser Aquí se eligió la base de datos a usar y luego se eligen las opciones que se quiere tener en la tabla de salida de BioParser. Para éste trabajo se eligió que la salida presentase los nombres de las secuencias, el tamaño, el porcentaje de identidad, la hebra en que se dio el hit, y las coordenadas de inicio y fin de los *match*s dentro de las dos secuencias

Scripts desarrollados en Perl y Bash

Como la página de internet de la cual se obtuvieron las bases de datos no permitía descargar la información en texto plano, se tuvieron que guardar las páginas enteras, para luego transformarlas en un archivo multifasta. Para transformar estos archivos lo que se hizo fue eliminar los caracteres especiales del formato .xml con sed. Sed es un programa de UNIX que permite, entre otras cosas, la búsqueda patrones dentro de un archivo de texto y los sustituye por lo que el usuario define, en este caso, por nada. Con ésta misma función, y partiendo de los archivos obtenidos por cromosoma, se armó un archivo único que contenía los nombres de los genes y el cromosoma al cual pertenece.

Por otra parte, como en Blast se analizó un genoma contra sí mismo, el resultado de Blast, y por consiguiente, el de BioParser presentaba autohits (genes que pegaban contra si mismos) y múltiples hits, los cuales son causados mayoritariamente cuando un gen es una copia de otro, o sea, son los casos de familias génicas con varias copias parálogas.
Para manejar éstos datos, se empleó un script en Perl el que analiza el archivo por filas y compara las filas entre sí, cuando encuentra algunas de las comparaciones en las que el nombre del gen *query* es igual al del hit o que se encuentren relacionados dos genes más de una vez, los descarta.

De esta forma se obtuvo un archivo para su visualización en Pajeck (ver más adelante) y este archivo en formato Pajek se lo procesó en un script en Bash que lo que hace es tomar la lista de los genes y los cromosomas a los que pertenecen, de forma tal que al hacer la presentación gráfica de las relaciones entre genes en forma de red se vieran distintos colores según en que cromosoma están los genes que aparecen en la red.

Por último, se empleó un script en Bash con el cual, a partir del archivo generado por Pajek, con los grupos compuestos por más de tres genes, y del archivo multifasta, crea archivos con la secuencia de cada gen por separado y detallando en que coordenadas de dicho gen, da alineamientos significativos y con que genes, de forma tal que Artemis lo pudiese leer. Esto último se hizo para obtener una representación a escala de cuanto abarcan los fragmentos candidatos a haber participado en eventos de conversión génica respecto al tamaño total de los genes.
En la Figura 12 se puede observar el orden en que fueron ejecutados los scripts detallados anteriormente, así como también el orden en que se ejecutaron el resto de los programas.

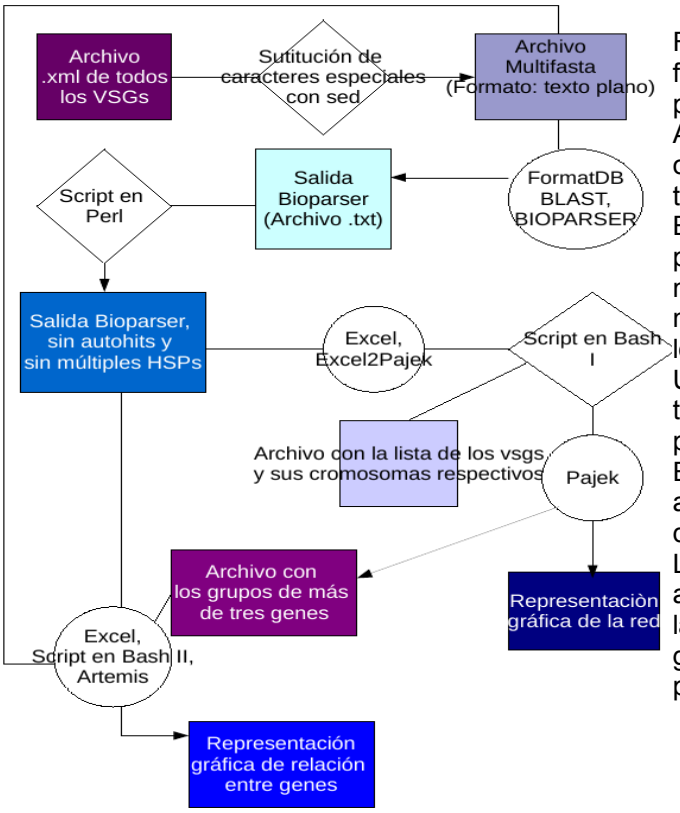


Figura 12: En este diagrama se muestra el orden en que fueron ejecutados los programas y los scripts diseñados para el análisis realizado en éste trabajo.
A partir de un archivo .xml con todos los genes VSGs, se obtuvo un archivo multifasta, el cual luego fue empleado tanto como base de datos como secuencias *query* para Blast (se comparó la base de datos contra si misma) y procesado con BioParser, y un script en Perl para mostrar los resultados de Blast en formato tabla, con los nombres de los genes, las coordenadas en que se dieron los hits, y el porcentaje de identidad entre secuencias.
Una vez hecho esto, se seleccionaron aquellos hits que tuviesen un largo de entre 40 y 500 bases, empleando para ello Excel, y luego se modificó ésta tabla con Excel2pajek y un script en Bash de forma tal que Pajek armara una red de relaciones entre genes según los datos de la tabla.
Luego se extrajeron los genes que formaban grupos del archivo multifasta y con los datos de las coordenadas de la salida de BioParser se generaron archivos por cada gen para Artemis con anotaciones de los genes que pegaron.

Pajek

Para visualizar los resultados obtenidos en Blast y BioParser, modificados de la forma antes descrita, se lo convirtió en un archivo para ser abierto en el programa Pajek.
Éste es un programa desarrollado por Batagelj y Mrvar (Batagelj et. al., 1998) que se utiliza para visualizar redes, analizarlas y puede calcular las propiedades de las mismas entre otras utilidades.
El archivo de entrada para éste programa es un texto plano y el formato depende de si se quiere una red en modo 1 o 2, en la Figura 13 se pueden ver ambos modos.
Para éste trabajo se empleó el formato en el modo 1, especificando además de qué color se querían representar a cada uno de los nodos (en éste caso, los genes) para poder ver de forma gráfica una red de relaciones entre genes, siendo el criterio de relación, los eventos de conversión génica.
Una vez visualizados los resultados, se empleó Pajek para obtener los datos de las secuencias que formaban grupos de relaciones que involucraran más de 3 genes. Los grupos de 2, fueron descartados para evitar ruido en la información.

Excel2pajek

Es un programa desarrollado por Jürger Peffer, el cual se encuentra disponible en http://vlado.fmf.uni-lj.si/pub/networks/pajek/howto/excel2Pajek.htm, que lo que hace es tomar dos columnas previamente definidas de un archivo Excel cualquiera, lo convierta en un archivo para Pajek.
Para este trabajo se tomó la salida de BioParser modificada con un script en Perl, detallado más abajo, y del que luego usando Excel se descartaran aquellos resultados en los cuales, el tamaño del hit estuviera por fuera 40 y 500 bases.

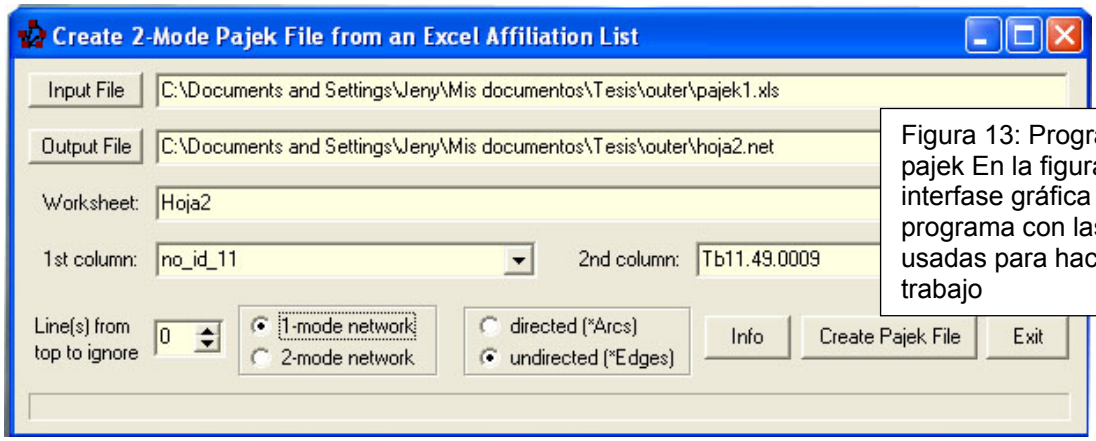


Figura 13: Programa Excel2 pajek En la figura se muestra la interfase gráfica de dicho programa con las opciones usadas para hacer el presente trabajo

ClustalW

Por otra parte, se repitieron los alineamientos y la construcción del árbol de los grupos, con ClustalW, usando las opciones que vienen por defecto.
Éste consiste en un programa de alineamiento global diseñado por Higgins y colaboradores (Thompson et. al 1994) que primero toma las secuencias de a pares, las alinea, calcula un *score*de distancia para cada alineamiento. Para alinear y calcular la similitud en éste primer paso, se puede hacer por el algoritmo de Wilbur y Lipman (Wilbur et al., 1983) o por método de programación dinámica.
El algoritmo de Wilbur y Lipman (Wilbur et al., 1983), realiza algo equivalente a hacer un alineaminto por dot-plot, donde el usuario define el tamaño de complete *match* mínimo (k-tuple), el número mínimo de k-tuples que debe ser considerado al elegir las mejores diagonales (Top Diag), el número de diagonales cercanas a las diagonales mejores al alinear (Window length), y el penalty por gap, siendo el score, el número de k-tuples en el mejor alineamiento restando los *penalties* por cada gap.
Alinea usando una matriz de sustituciones a elección del usuario, y restado los *penalties* por gaps definidos por el usuario, para luego calcular el *score*como el número de *match*es entre las secuencias alineadas, sobre el total de residuos analizados, sin contar los gaps.

Una vez calculados los scores, se construye una matriz de distancia y se clusteriza usando el método de reconstrucción filogenética llamado Neighbour-joining (Saitou et al., 1987) para formar un árbol guía, y a partir de éste, se alinea de forma progresiva considerando el árbol guía. Para ello emplea, una matriz de sustitución definida por el usuario, y *penalties* por gaps tanto incluir un nuevo gap de cualquier tamaño (Gap Opening) y el costo por cada item en el gap (*gap extention penalty*) que el usuario elige como condiciones iniciales, y que luego varían según si, en la posición en que debería ir un gap al agregar una secuencia al alineamiento, ya hay gaps o no, y además el *score*es ponderando por la distancia de las secuencas a la raíz del árbol dado que cada valor de la matriz de sustitución es multiplicado por el peso de las secuencias a alinear.
El peso de las secuencias es asignado según el árbol guía, siendo normalizado el peso de forma tal que el que tiene más peso es el que está menos relacionado con el resto de las secuencias y seteado como 1.0 y el resto, con una cifra menor. Esto se basa en el hecho de que los grupos relacionados contienen mayor cantidad de información duplicada.

Por otra parte, si bien los *penalties* son fijados por el usuario al principio, ClustalW corrige éstos números de la siguiente manera:

- Dependiendo de la matriz de sustitución que se esté usando, el *gap open penalty* (GOP) es normalizado según el *score*promedio de dos residuos con mis*match*.
- El porcentaje de identidad entre dos secuencias (o dos grupos de secuencias) es usado para aumentar el GOP cuando las secuencias están muy relacionadas o disminuirlo cuando no.
- Considera el largo de las secuencias, en el hecho de que cuanto más larga una secuencia, crece el *score*por una cuestión de que son más residuos y entonces aumenta la probabilidad de *match* por azar. Entonces, lo que se hace es incrementar GOP con el largo de la secuencia, siendo el muevo GOP, el GOP normalizado por la matriz de sustitución, sumado al logaritmo del largo de la secuencia mas corta.
- La diferencia entre largos de la secuencia, influye en el *gap extention penalty* (GEP), y éste es calculado de forma tal que cuanto más sea la diferencia en el largo de dos secuencias, mas grande es el GEP, de forma tal que no fomente gaps muy largos solo por el hecho de la diferencia de largo.

Entonces, tomando en cuenta éstas modificaciones, los *penalties* iniciales son:
GOP= [GOP definido por el usuario + log (min N, M)]*(*score*promedio de los mis*match*s)*(factor normalizador por el porcentaje de identidad) donde M y N son los largos de las secuencias.
GEP= GEP defindo por el usuario + [1.0 - |log (N/M)|]

Si bien éstos son los *penalties* iniciales, a medida que va alineando, los va manipulando dependiendo de diferentes situaciones en las cuales se deba colocar gaps. Lo que hace es generar una tabla de gap open *penalties* por cada posición para las dos secuencias a alinear o para cada set, según corresponda según las siguientes reglas:

- Los gaps terminales, son usualmente sin costo.
- Si ya existe un gap en una determinada posición, GOP es disminuido según la cantidad de secuencias que ya tengan gaps en esa posición, y el GEP baja a la mitad.
  GOP_nuevo=GOP_anterior para esa posición * 0,3 * (nro. de secuencias sin gaps /nro. total de secuencias)

Si no hay ningún gap en esa posición, entonces GOP es aumentado si esa posición está a 8 o menos residuos de una posición con gaps.
GOP_nuevo=GOP_anterior para esa posición * {[2+(8 – distancia desde el gap a ésa posición) * 2]/8}

- Si, en el caso de ailnear aminoácidos, en cualquier posición donde se deba introducir un gap, tiene al lado 5 residuos hidrofílicos o más - lo cual es considerado un bloque hidrofílico - y no hay gaps, entonces GOP se disminuye a un tercio. Los aminoácidos que pueden constituir este tipo de bloque lo puede definir el usuario, pero D, E, G, K, N, Q, P y R son los que vienen por defecto.
- Cuando se alinea aminoácidos, si no hay un bloque hidrofilico y no hay gaps en esa posición, entonces GOP es multiplicado por el valor en la tabla 1 correspondiente al aminoácido que se halla alineado en esa posición, y si hay varios tipos de aminoácidos en esa posición, GOP es multiplicado por el promedio de todas las contribuciones hechas por cada secuencia.

Tabla 1: Valores que ClustalW emplea para modificar el *gap open penalty* (GOP) cuando se debe insertar un gap en una posición donde no hay gaps ni bloques hidrofílicos (Thompson et. al 1994)

A	1.13	M	1.29
C	1.13	N	0.63
D	0.96	P	0.74
E	1.31	Q	1.07
F	1.20	R	0.72
G	0.61	S	0.76
H	1.00	T	0.89
I	1.32	V	1.25
K	0.96	Y	1.00
L	1.21	W	1.23

Luego de realizado el alineamiento, construye un nuevo árbol por Neighbour-joining y ése es el que presenta como resultado.

Muscle

Luego de obtenidos los nombres de los genes que formaban grupos de relaciones, se alinearon las secuencias de éstos genes, resultando en un alineamiento por grupo.
Por otra parte, se realizó un alineamiento empleando como entrada a todas las secuencias que formaron grupos.
Asimismo, se analizó el árbol que Muscle empleó para hacer los alineamientos. Cabe destacar que usualmente no se usan estos árboles para hacer estudios filogenéticos, pero para este caso de estudio particular sirve.
Muscle consiste en un programa de alineamiento múltiple diseñado por Robert Edgar (Edgar, 2004) que lo que busca es optimizar el tiempo que lleva hacer un alineamiento y a su vez, que éste sea de buena calidad.
Para obtener los alineamientos, lo primero que hace es contar la frecuencia de k-meros que tienen las secuencias a ser alineadas, calculando así los índices de similitud entre secuencias y con esto construye un árbol guía.
Una vez, hecho este árbol guía va alineando las secuencias siguiendo el orden de las ramas del árbol. Cando dicho alineamiento se obtiene, se calculan las distancias Kimura y luego clusteriza por UPGMA (Sneatch et al., 1973) y construye un segundo árbol. Éste segundo árbol es comparado con el primero, identificando aquellos nodos en el cual difiere el orden de las ramas.
A partir de éste segundo árbol hace un nuevo alineamiento, siguiendo el orden de las ramas del árbol y calcula el SP *score* (Carillo et al., 1988).
El SP *score* (sum-of pair-score) se calcula de la siguiente manera:
Se toman de a pares todas las secuencias alineadas y se calcula para cada par la suma de los scores de la matriz de sustitución empleada para cada residuo alineado, considerando gap *penalties*.
Los gap *penalties* se calculan descartando las columnas en las cuales las dos secuencias comparadas presentan gaps (Figura 15) y luego se calcula por cada gap $g+\lambda e$, siendo g el *gap open penalty*, λ la longitud del gap y e, el *gap extention penalty*. Los *gap open penalty* y *gap extention penalty* son parámetros fijados previamente; y, en el presente trabajo, se emplearon los valores por defecto.

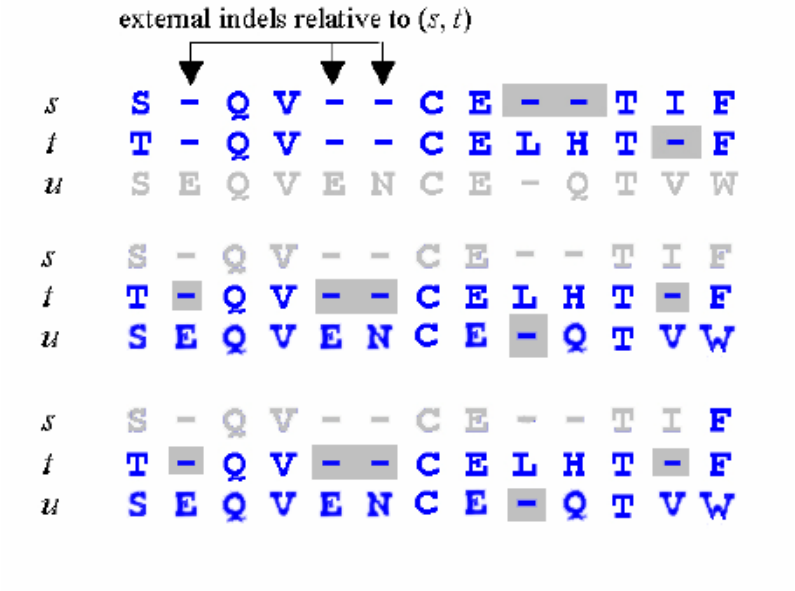


Figura 15: Cálculo del SP score. En esta figura se señalan con flechas aquellos gaps que se descartan en el cálculo del SP score, mientas que los que se emplean en el cálculo se resaltan con bloques grises.(Modificado de Edgar., 2004)

Una vez calculado el SP-score para el primer alineamiento, borra una rama del árbol, separando los profiles, siendo éstos los alineamientos dentro de cada rama, realinea y calcula el SP *score*del nuevo alineamiento. Este SP *score*es comparado con el SP *score*del alineamiento anterior y toma en cuenta el alineamiento con mejor score. Ésta parte del procedimiento es iterada hasta que fueron visitadas todas las ramas y no hay cambios o hasta un número máximo de iteraciones. Las ramas son visitadas en orden decreciendo la distancia a la raíz.

En la Figura 16 se ve un esquema del diagrama de flujo de Muscle.

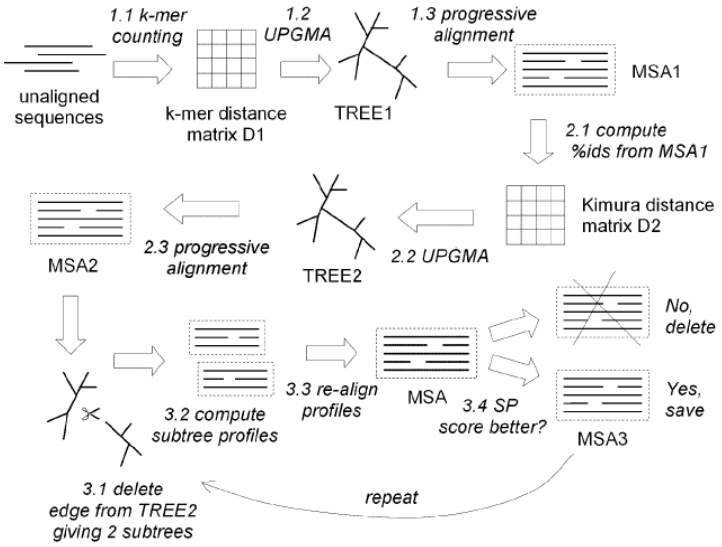


Figura 16: Esquema de cómo trabaja Muscle
(Edgar, 2004)

Mega

Una vez obtenidos los árboles, se los visualizó con MEGA
Es un programa diseñado por el grupo de Masatoshi Nei (Tamura et. al., 2007) que se emplea para análisis de secuencias como por ejemplo, hacer alineamientos, construcción y visualización de árboles filogenéticos, cálculo de índices de similaridad, entre otras funciones.

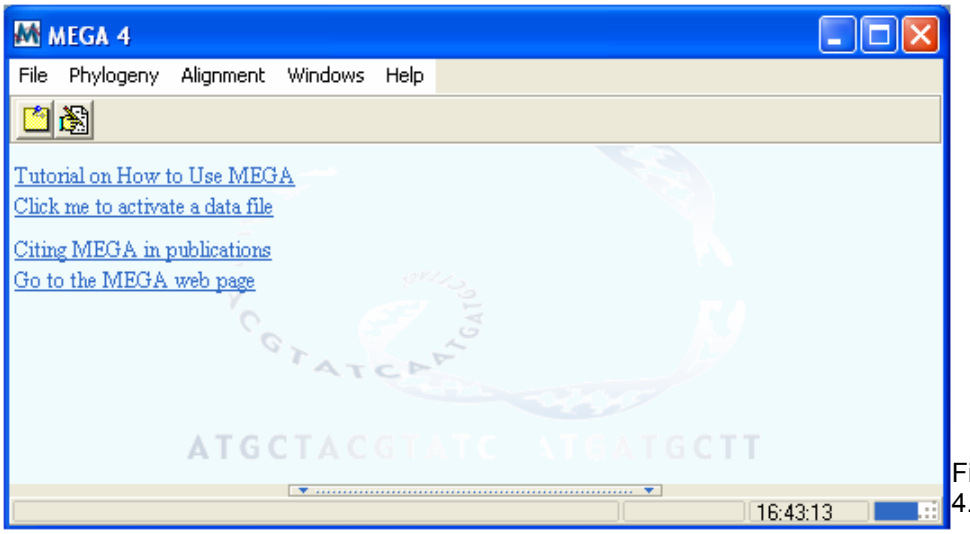


Figura 17: Interfase grafica de MEGA 4.0

Artemis

Es un programa desarrollado en Java por el Sanger Centre, el cual es utilizado para visualización y anotación de genomas. En la Figura 18 se puede ver la anotación del contig NT_032968.7, perteneciente al cromosoma 1 humano como ejemplo de cómo trabaja Artemis.
En éste trabajo se empleó para visualizar en que regiones de un gen seleccionado presenta homologías con otros genes.

Resultados

En el presente trabajo se analizaron 931 secuencias de genes VSGs, de las cuales, 47 son funcionales, 91 tienen folding atípico y el resto son pseudogenes.
Dado que nos interesa comparar genes codificantes de VSGs entre si, se armó una base de datos local con dichas secuencias y se alinearon usando como base de datos las mismas secuencias (o sea se hizo Blast de una base de datos contra si misma) con Blastn, empleando un *valor umbral* de e-value de 3,0 x e^{-10} y usando opciones específicas para que no enmascare sitios de baja complejidad. Luego el resultado de Blast fue representado en BioParser y posteriormente fue depurado con excel y los scripts antes descriptos con la finalidad de encontrar aquellas secuencias que comparten segmentos de entre 40 y 500 bases con un porcentaje de identidad mayor al 70%.

Al observar la salida de BioParser sin modificar (y por ende, de Blast), 872 de las secuencias analizadas presentaron hits con otras secuencias. De los 872 hits, el que tuvo un menor porcentaje de identidad fue de 78.90%.
Al observar la salida de BioParser modificada, se pudo constatar que muchos de los alineamientos con alto porcentaje de identidad (mayor al 70%) se localizan en la región 3', lo

cual se puede ver en la tabla 2, donde se muestra parte del archivo de salida del BioParser.

Tabla 2: Parte de la salida de BioParser. Los datos que presenta son: Los nombres de los genes, los largos, las cadenas en la cual se encuentran los genes, el porcentaje de identidad del *match*, el porcentaje de las secuencias implicado en el alineamiento y las coordenadas de las bases implicadas. En la tabla se observa que, si bien hay algunos hits localizados en los expremos 5', los hits se dan principalmente en los extremos 3'.

.

Query	Query_L	Hit	Hit_L	Q_Str	H_Str	%Id	%Aln_Q	%aln_H	Q_Start	Q_End	H_Start	H_End
Tb09.v4.0001	1491	Tb11.21.0011	1172	1	-1	98.28	3.89	4.95	6	63	1115	1172
Tb09.244.0170	681	Tb927.4.5680	1722	1	1	94.23	7.64	2.96	4	55	1253	1303
Tb10.v4.0027	1554	Tb11.30.0014	1527	1	1	83.42	12.81	13.03	1	199	1	199
Tb927.8.180	378	Tb11.14.0026	1722	1	1	88.33	15.87	3.48	16	75	1360	1419
no_id_18	324	Tb10.v4.0167	1575	1	1	93.02	13.27	2.73	25	67	1261	1303
Tb10.v4.0015	1518	Tb08.27P2.630	1518	1	1	90.97	30.63	30.63	1	465	1	465
Tb09.v2.0290	339	Tb10.v4.0145	1476	1	1	96.08	15.04	3.46	58	108	1324	1374
Tb927.4.5490	315	Tb927.5.5170	1542	1	1	89.09	17.46	3.57	48	102	1326	1380
Tb09.v2.0290	339	Tb09.244.1390	1554	1	1	90.20	15.04	3.28	60	110	1326	1376
Tb09.v2.0290	339	Tb927.5.4840	1458	1	1	94.87	11.50	2.67	79	117	1291	1329
Tb09.v2.0290	339	Tb09.v4.0005	1482	1	1	81.98	32.74	7.49	24	134	1284	1394
Tb09.v2.0110	318	Tb09.354.0180	1527	1	1	85.90	23.58	5.11	52	126	1255	1332
Tb927.5.4890	321	Tb09.354.0180	1527	1	1	86.96	20.56	4.52	64	129	1260	1328
no_id_13	432	Tb11.15.0004	1536	1	1	91.67	11.11	3.12	133	180	1252	1299
Tb09.v2.0110	318	Tb09.v4.0121	1618	1	1	87.50	20.13	3.96	75	138	1369	1432
Tb927.4.5800	399	Tb927.5.4650	1569	1	1	86.11	18.05	4.59	109	180	1285	1356
Tb927.5.4890	321	Tb09.244.0710	1533	1	1	86.08	24.30	5.15	68	145	1261	1339
Tb09.v2.0290	339	Tb927.6.5280	1599	1	1	89.09	16.22	3.44	101	155	1370	1424
Tb08.27P2.565	408	Tb11.44.0004	1545	1	1	84.09	21.57	5.70	114	201	1284	1371
Tb09.244.1890	489	Tb11.14.0009	1530	1	1	91.49	9.61	3.07	216	262	1317	1363
Tb09.244.1940	537	Tb10.v4.0168	1491	1	1	91.94	10.99	4.16	233	291	1304	1365
Tb927.3.260	468	Tb09.v4.0137	1550	1	1	85.90	16.67	4.84	187	264	1266	1340
no_id_24	263	Tb11.49.0003	1461	1	1	89.39	24.71	4.52	87	151	1263	1328
Tb927.3.200	315	Tb09.v4.0102	1518	1	1	90.48	20.00	4.15	121	183	1330	1392
Tb09.v2.0150	534	Tb927.4.5780	1590	1	1	81.60	23.41	7.86	193	317	1246	1370
Tb11.24.0010	582	Tb09.v4.0103	1545	1	1	92.86	9.62	3.62	296	351	1289	1344
Tb09.v2.0150	534	Tb11.57.0021	1494	1	1	89.29	10.49	3.75	268	323	1270	1325
Tb09.v2.0150	534	Tb927.3.120	1557	1	1	87.50	11.99	4.11	264	327	1251	1314

A partir de aquellas secuencias que cumplen con las condiciones establecidas anteriormente, se construyó una red de relaciones, la cual se analizó con el programa Pajek. Esta aproximación nos permitió detectar la existencia de 36 grupos de genes (de 3 o mas genes), involucrando a 266 genes (de un total de 931).

El análisis de dichos grupos permite ver que estas no son interacciones aleatorias de todos los genes contra todos, sino que se forman grupos separados. Por otra parte las relaciones dentro de cada grupo, son de la siguiente forma: si un gen A se relaciona con un gen B, y a su vez el gen B se relaciona con un gen C, esto no implica necesariamente que A se relacione con C de forma directa. Esto es precisamente lo esperable cuando los eventos de conversión génica entre el gen en cuestión y distintos genes afectan distintas coordenadas. Veamos por ejemplo el caso del grupo que contiene al gen Tb08.27P2.490, que se señala como ejemplo en la Figura 19. Este gen presenta hits con los genes Tb11.35.0001,

Tb11.43.0003 y Tb09.v4.0103 en diferentes regiones y por eso Blast no encuentra relaciones directas entre estos últimos genes, como se muestra en la Figura 20.

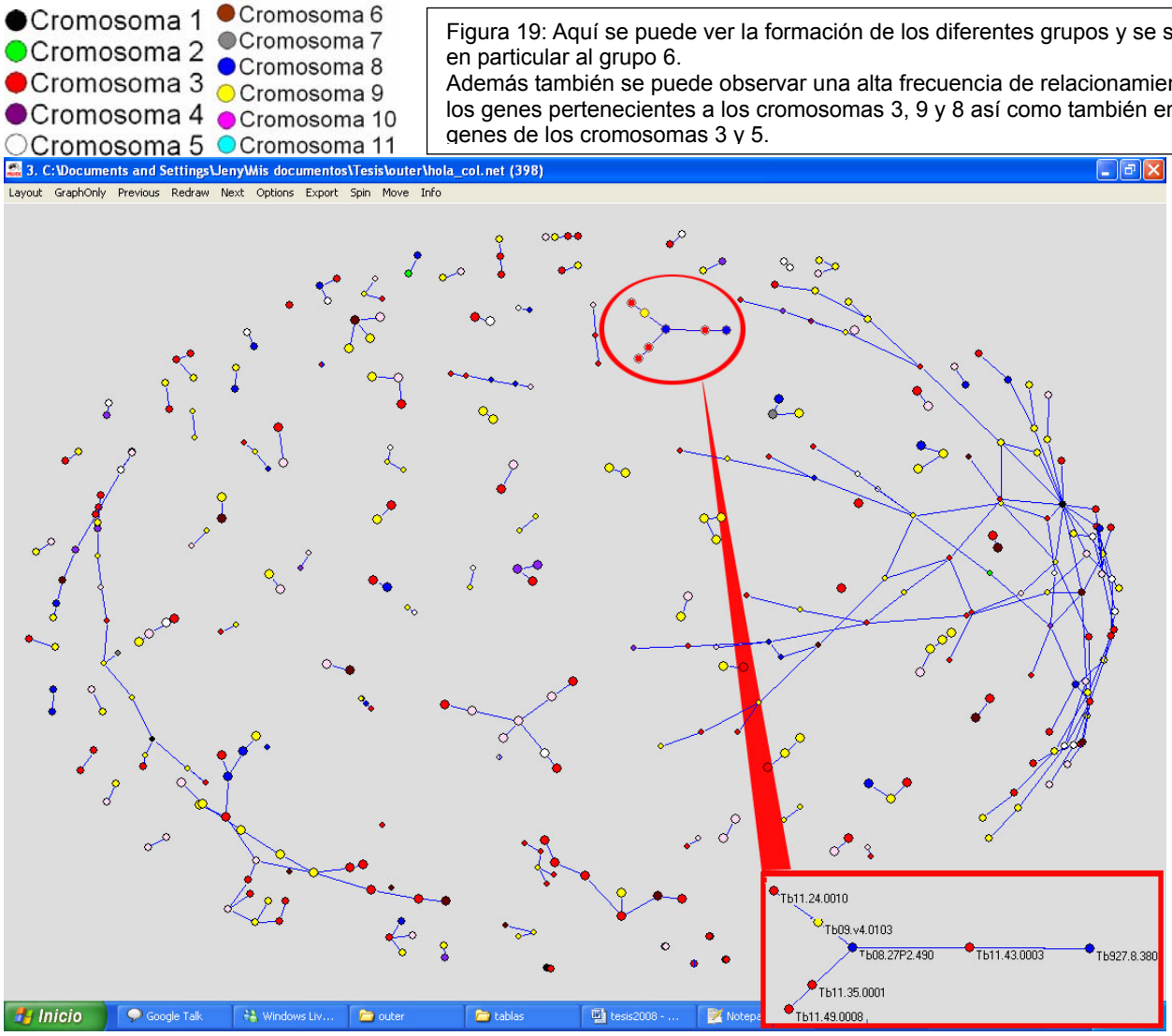


Figura 19: Aquí se puede ver la formación de los diferentes grupos y se señala en particular al grupo 6.
Además también se puede observar una alta frecuencia de relacionamiento entre los genes pertenecientes a los cromosomas 3, 9 y 8 así como también entre los genes de los cromosomas 3 y 5.

Tabla 3: Fragmento de la salida de BioParser que contiene los datos del grupo 6 que contiene al gen VSG Tb08.27P2.490. Los datos que presenta son: Los nombres de los genes, los largos, el porcentaje de identidad del *match*, el porcentaje de las secuencias implicado en el alineamiento y las coordenadas de las bases implicadas. Estas coordenadas son las empleadas para construir los archivos en Artemis.

Query	Query_length	Hit	Hit length	%id	%aln query	%aln_hit	Query start	Query end	Hit start	Hit end
Tb08.27P2.490	1590	Tb11.35.0001	1545	93.65	3.96	4.08	1358	1420	1327	1389
Tb08.27P2.490	1590	Tb09.v4.0103	1545	93.48	2.89	2.98	1473	1518	1431	1476
Tb08.27P2.490	1590	Tb11.43.0003	1521	92.19	4.03	4.21	1513	1576	1444	1507
Tb11.35.0001	1545	Tb11.49.0008	1458	94.12	3.30	3.50	1459	1509	1360	1410
Tb11.43.0003	1521	Tb927.8.380	1491	90.14	4.67	4.76	1272	1342	1245	1315
Tb09.v4.0103	1545	Tb11.24.0010	582	92.86	3.62	9.62	1289	1344	296	351

Por otra parte, se tomaron las secuencias dentro de cada grupo y se visualizaron con Artemis analizando la salida de Blast para ver los alineamientos correspondientes. De esta forma es posible ver para cada secuencia de los grupos, los segmentos que presentan un alto porcentaje de identidad con segmentos de otras secuencias del grupo al que pertenecen. Con esto, se pudo observar que en la mayoría de los casos se da que un gen tiene fragmentos con un alto porcentaje de identidad con secuencias pertenecientes a distintos genes, formandos mosaicos como los de la Figura 20. También se pudo observar que habían genes que presentaban fragmentos que presentaban *match*es con múltiples genes a la vez (Figura 21).

Figura 20: Imagen en Artemis del gen Tb08.27P2.490, en la cual se puede ver que presenta hits con tres genes diferentes en tres lugares distintos formando un mosaico.
Además se muestran los respectivos alineamientos de estos segmentos obtenidos con Blast.

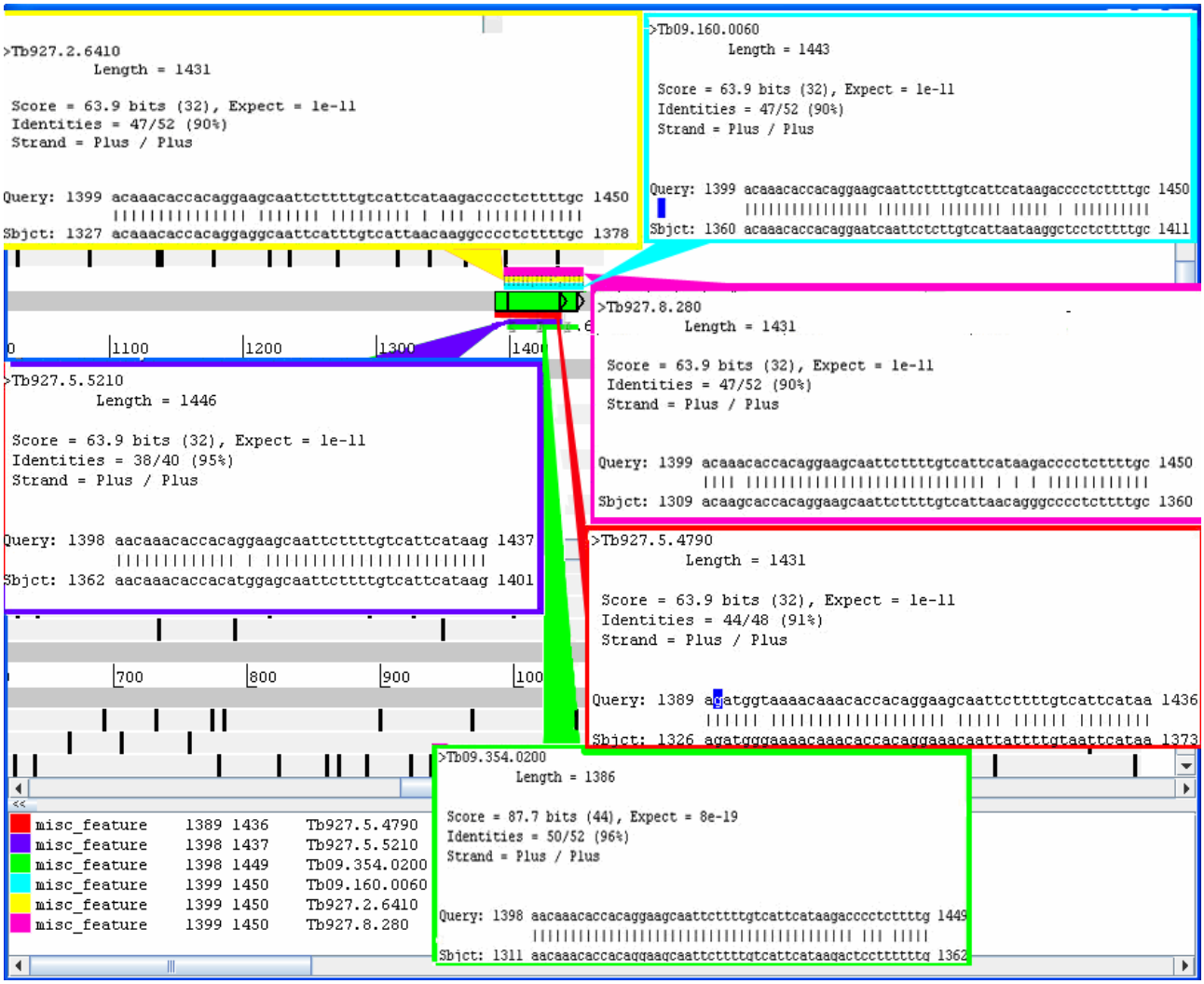


Figura 21: Imagen en Artemis del gen Tb09.277.0250, en el cual se ve que en una sección pequeña en el extremo 3' pega con varios genes, a la cual se le superpusieron los respectivos alineamientos hechos por Blast correspondientes y se le resaltaron los sitios donde se da el alineamiento dentro del gen.FIGURA 20

Posteriormente se realizaron alineamientos múltiples con los programas ClustalW y Muscle de las secuencias completas con los genes que formaron grupos compuestos por 3 o más genes. Esto permitió ver que a pesar del hecho que estos genes presentaban regiones de alta identidad con otros genes, los alineamientos involucrando a los genes completos eran malos. Esta observación también se puede apreciar al observar los árboles filogenéticos derivados de dichos alineamientos (Figuras 21 y 22). La forma de verlo en los árboles es que los genes que pertenecen a un mismo grupo de los presentados en la figura 19 no aparecen todos juntos en el árbol filogenético (lo esperable si la similitud se debiera a ancestría común), sino que aparecen dispersos por todo el árbol, y a su vez, las distancias entre estos genes son muy grandes.

Esta observación indica que aquellos genes que presentan segmentos con alta identidad, presentan identidades muy bajas por fuera de los mismos, de los contrario aparecerían agrupados en los árboles filogenéticos, y los largos de las ramas (que son proporcionales a el nivel de disimilitud de secuencias) serían considerablemente menores.

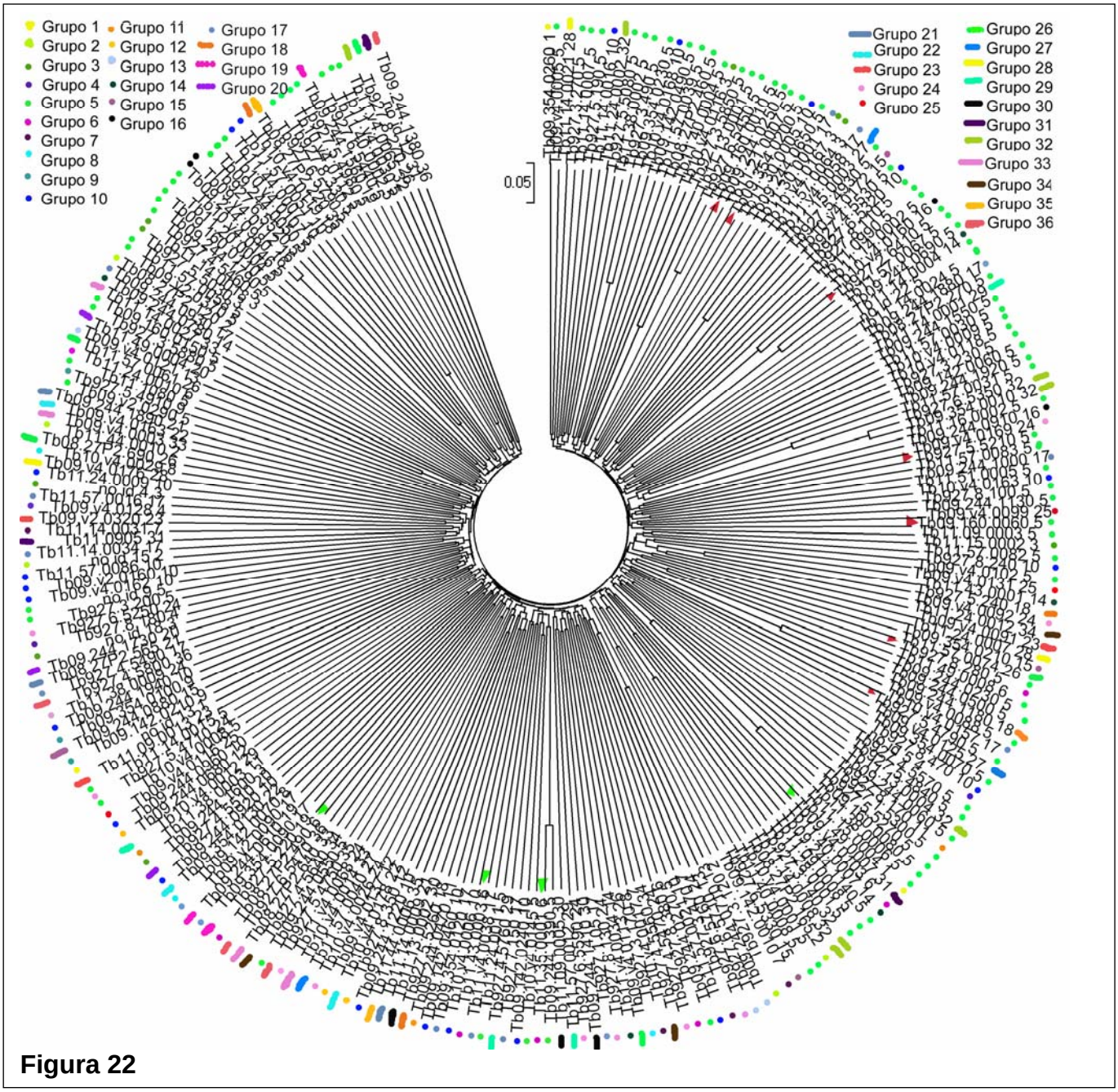


Figura 22

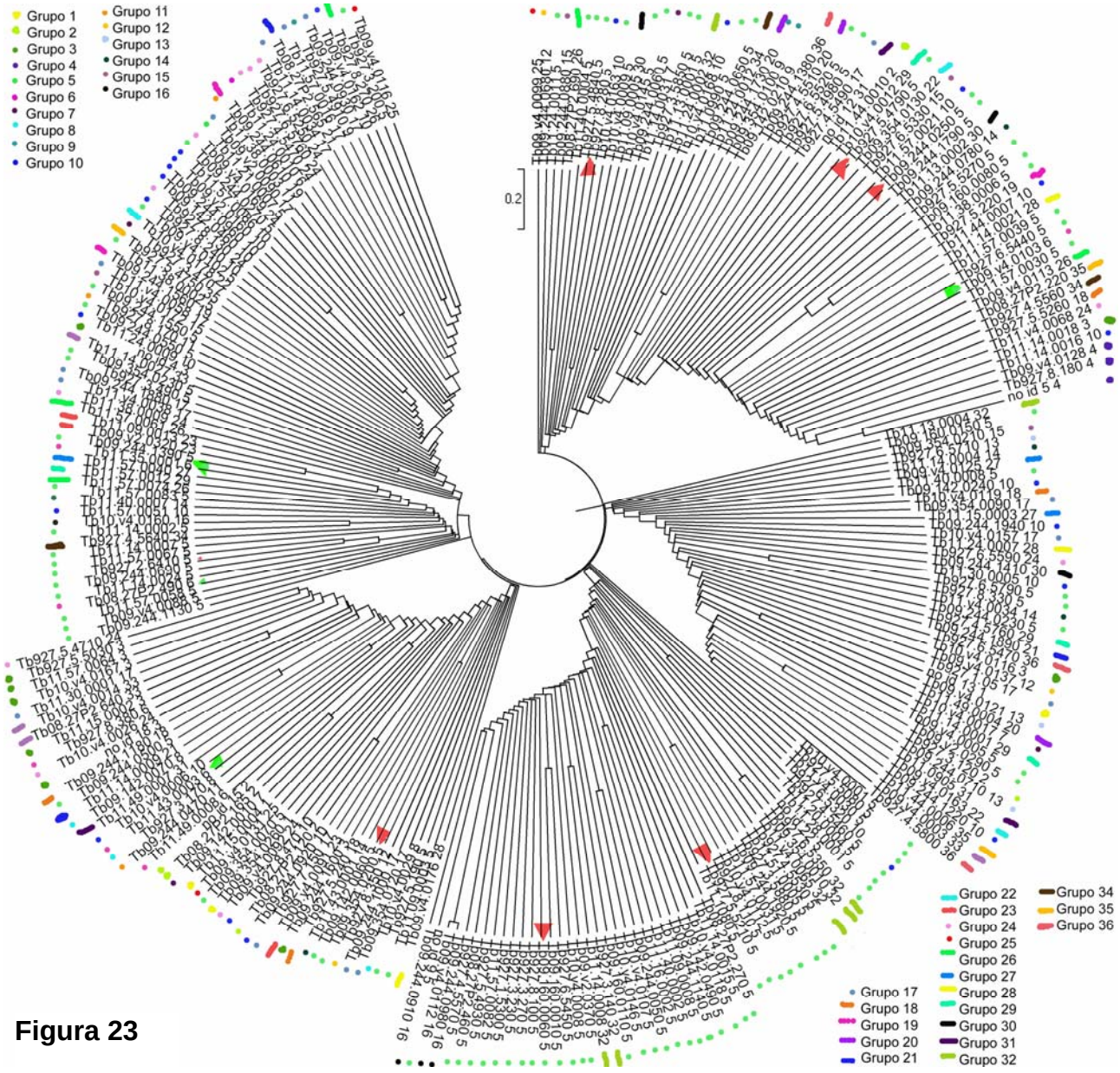


Figura 23

Figura 22: Árbol filogenético obtenido usando ClustalW
Figura 23: Árbol filogenético obtenido usando Muscle.
Si bien las ramas se disponen en forma diferente, en ambos árboles se puede observar que la distribución de los genes de los diferentes grupos encontrados por Pajek es bastante dispersa. Con la flecha roja, se indica la posición de los genes Tb09.244.0250, Tb9275.4790, Tb927.5.5210, Tb09.160.0060 y Tb927.8.280 (ver Figura 19). Con flechas verdes están señalados los genes Tb08.27P2.490, Tb09.v4.0103, Tb11.35.0001 y Tb11.43.0003 (ver Figura 20).
En ambas figuras el largo de las ramas es proporcional a la cantidad de divergencia nucleotídica. Para referencia se colocan una regla que corresponden respectivamente a 5% y 20% de divergencia nucleotídica.

Por último, se reclasificaron todos los VSG en grupos tomando en cuenta las posiciones de los mismos en los diferentes cromosomas (Tabla 4) y a partir de las relaciones entre genes encontradas con pajek, se calculó la frecuencia de relaciones entre grupos (Tabla 5). Cabe destacar que en el cálculo de dichas frecuencias, se excluyeron algunos genes VSGs de los cromosomas 8, 9, 10 y 11 debido a que estos solo están registrados en el VSGdb pero no estan anotados en el Genbank. Puesto que el VSGdb no dispone las coordenadas de los genes en los cromosomas no fue posible localizar a los mismos.

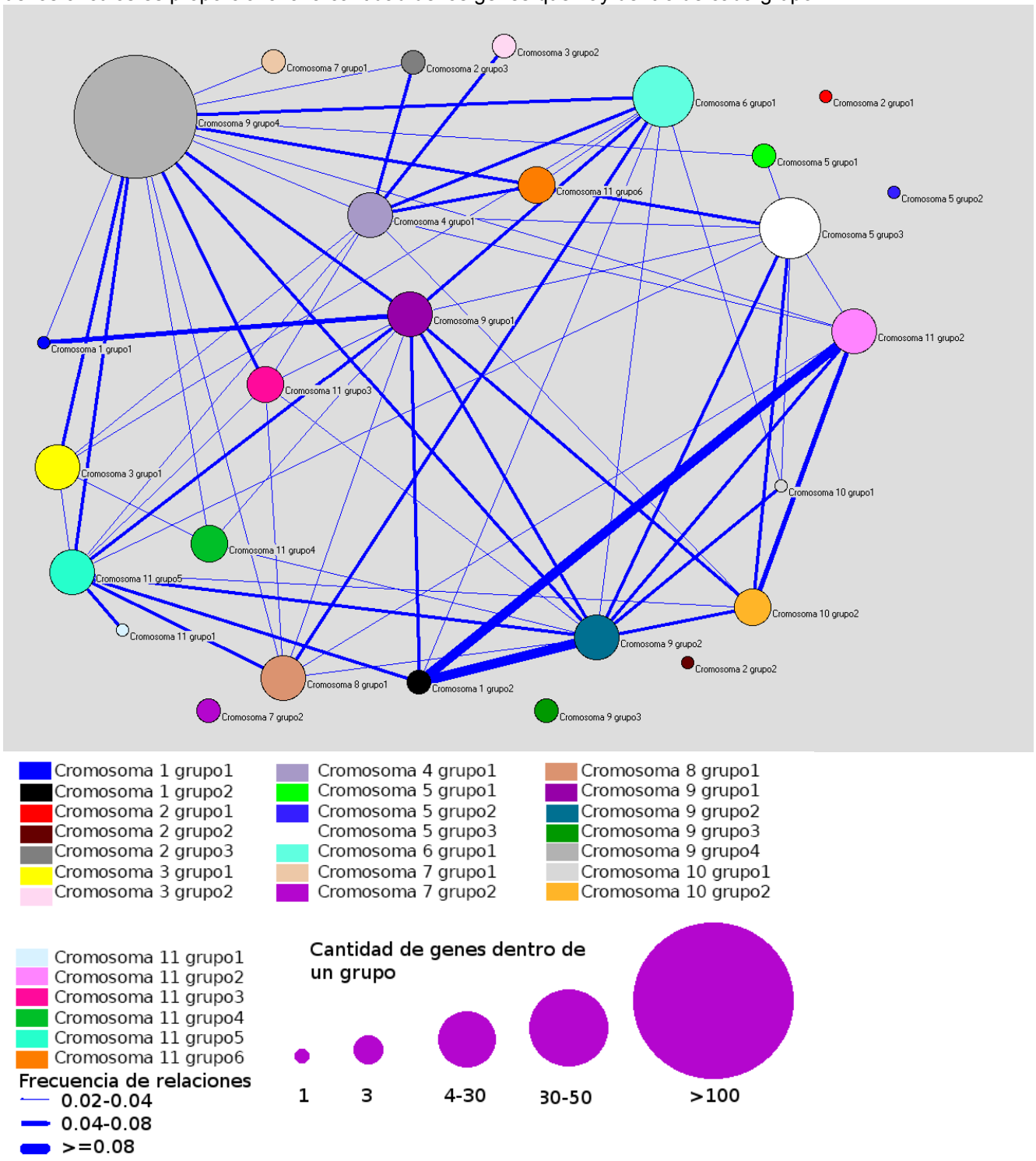
Tabla 4: Muestra las coordenadas de los nuevos grupos en los diferentes cromosomas. Para determinarlas, se tomó la coordenada de inicio del gen que aparece en el grupo y la coordenada fin del último gen

	Inicio	fin	cant de genes dentro del grupo
Cromosoma 1 grupo1	381	1961	1
Cromosoma 1 grupo2	1050169	1057855	3
Cromosoma 2 grupo1	393056	394390	1
Cromosoma 2 grupo2	1166451	1167905	1
Cromosoma 2 grupo3	1178158	1193947	5
Cromosoma 3 grupo1	2731	109885	33
Cromosoma 3 grupo2	1637900	1653225	6
Cromosoma 4 grupo1	1477030	1590261	35
Cromosoma 5 grupo1	36196	59811	8
Cromosoma 5 grupo2	1236331	1237764	1
Cromosoma 5 grupo3	1386066	1602637	62
Cromosoma 6 grupo1	1444310	1616066	53
Cromosoma 7 grupo1	5	11733	5
Cromosoma 7 grupo2	1768836	1782885	5
Cromosoma 8 grupo1	2	134249	32
Cromosoma 9 grupo1	1871	151258	44
Cromosoma 9 grupo2	188040	319439	36
Cromosoma 9 grupo3	2408372	2443654	5
Cromosoma 9 grupo4	2504099	3055206	151
Cromosoma 10 grupo1	1712	3088	1
Cromosoma 10 grupo2	3930077	4053616	27
Cromosoma 11 grupo1	2912	4348	1
Cromosoma 11 grupo2	4543426	4656874	33
Cromosoma 11 grupo3	4707288	4782248	23
Cromosoma 11 grupo4	4805763	4924231	25
Cromosoma 11 grupo5	4966531	5169071	47
Cromosoma 11 grupo6	5208028	5230339	9

Para el cálculo de dichas frecuencias, se emplearon las siguientes fórmulas dependiendo de si se está contabilizando relaciones entre genes de un mismo grupo o si se trata relaciones entre genes de dos grupos distintos:
Las frecuencias de relaciones entre genes de un mismo grupo se calcularon como el cociente entre la cantidad de genes que se relacionan y la cantidad de genes dentro del grupo por dos menos 1 (se resta uno porque no es posible un evento de conversión génica entre dos segmentos de un mismo gen):
Frecuencia de relaciones dentro de un grupo= cantidad de relaciones/ (cantidad de genes dentro del grupo*2 -1)
En cuanto al caso de grupos distintos, la frecuencia se calculó como el cociente entre la cantidad de genes que se relacionan y la cantidad de genes totales de los grupos, o sea, si

se tienen dos grupos A y B, la frecuencia de relaciones es:
Frecuencia de relaciones entre los grupos A y B = cantidad de relaciones entre A y B/ (cantidad de genes dentro del grupo A + cantidad de genes dentro del grupo B).
Una vez calculadas las frecuencias, se diseñó un gráfico el cual se representaron las relaciones entre los diferentes grupos, siendo el tamaño de los vértices la cantidad de genes que se encuentran dentro del grupo representado y el grosor de las líneas es proporcional a la frecuencia de recombinaciones (Figura 24).

Figura 24: Gráfico de frecuencia de relaciones entre los grupos definidos por la ubicación espacial de los genes en los diferentes cromosomas. El grosor de las lineas es proporcional a la frecuencia de relaciones y el tamaño de los círculos es proporcionál a la cantidad de los genes que hay dentro de cada grupo.



Si se observa la tabla 5 y la figura 24, se puede ver que el segmento correspondiente al grupo 4 del cromosoma 9 se relaciona con varios segmentos de diferentes cromosomas aunque las frecuencias de las mismas son bajas. Por otra parte, no se observan relaciones que involucren a los genes pertenecientes al grupo 3 del cromosoma 9 así como tampoco a los genes que están en el grupo 2 del cromosoma 2. También se puede observar que existe una alta frecuencia de relaciones entre el el grupo 2 del cromosoma 11, el grupo 2 del cromosoma 1 y el grupo 3 del cromosoma 9.
Más alla de las relaciones entre un segmento cromosómico con otros, lo que cabe destacar de esta figura y de la tabla 5, es que las frecuencias de relaciones entre grupos son variables.

Discusión

Al analizar los alineamientos, las representaciones en Artemis, la red de relaciones en Pajek y los árboles filogenéticos, se puede afirmar que estas evidencias conjuntas indican que efectivamente ocurren procesos de conversión génica entre genes VSGs inactivas. En resumen de los 931 genes VSGs analizados en este trabajo, 266 presentaron evidencia de conversión génica. Esta estimación se basa fundamentalmente en el hecho de que genes identificados como pertenecientes a un mismo grupo en la figura 19, por presentar segmentos con identidades superiores al 90%, se encuentran completamente dispersos en los árboles filogenéticos. Esto indica por una lado que los mismos presentan identidades de secuencias muy bajas cuando las secuencias son tomadas enteras, mientras que estos mismos genes tienen segmentos con porcentajes de identidad mayores al 90% (los segmentos que encuentra Blast y por ende hacen que pajek encuentre la existencia de éstos grupos). Esto puede ser tomado como evidencia de conversión génica dado que, por un lado, la alta identidad observada en los mencionados segmentos no puede ser atribuida a separación reciente, porque si esto fuera así, lo esperable es que también encontráramos alta identidad a lo largo del gen completo. Por otra parte, tampoco podría atribuirse esta identidad a conservación debido a restricciones funcionales, si lo fuera, uno esperaría que el patrón de conservación fuera repetitivo. Sin embargo las zonas de alta identidad cambian de una comparación a otra.

Un aspecto interesante es que algunos genes, presentan fragmentos con altos porcentajes de identidad con fragmentos de varias secuencias pero las secuencias enteras de los genes involucrados están separadas en el árbol filogenético. Un ejemplo es lo que se observa desde la base 1389 a la 1450 del gen Tb09.244.0250 (Figura 21). Ésto puede deberse a que los segmentos de los genes que presentan dicha particularidad, sean hot spots de recombinación, o puede también deberse a que dichos segmentos se hallan conservados en los genes involucrados debido a restricciones funcionales. Los resultados disponibles por el momento no permiten discernir entre ambas posibilidades.

Si bien, las observaciones descritas anteriormente se pueden apreciar en todas las representaciones de Artemis y en los correspondientes resultados de Blast, en el presente trabajo solo se muestran para el caso de los genes Tb08.27P2.480 y Tb09.244.0250 por cuestiones de practicidad y representatividad, dado que en el primero se pueden apreciar mosaicos y el segundo, un segmento que presenta un alto porcentaje de identidad con fragmentos de varios genes VSGs a la vez.

Por otra parte, tal como surge de la simple observación del archivo de salida de BioParser modificada, se puede apreciar que en la mayoría de los genes, los hits se localizan en la porción 3'. La sobreabundancia de posible eventos de conversión en esta región del gen, asi como la casi total ausencia en la región 5' puede interpretarse como resultado de que los sitios de recombinación se localizan preferentemente en esta zona del gen. Una interpretación alternativa es que estos segmentos de alta identidad no se deben a eventos de conversión, sino a mayor restricción funcionales tal como ya fue indicado anteriormente. Sin embargo el hecho que el patrón de fragmentos de alta identidad no sea repetitivo es un fuerte indicador de que no se debe exclusivamente a restricciones.

A su vez, en la Figura 24 y la tabla 5 no se observan relaciones que involucren a los genes pertenecientes al grupo 3 del cromosoma 9 así como tampoco a los genes que están en el grupo 2 del cromosoma 2. Esto puede deberse a que efectivamente no hayan existido eventos de conversión génica para dichos segmentos, o que no se hayan podido detectar debido a las exigencias impuestas para el tamaño y el porcentaje de identidad al seleccionar solo aquellas relaciones entre genes que comparten segmentos de entre 40 y 500 bases con un porcentaje de identidad mayor al 70%.

Además, a partir de la figura 24 y de la tabla 5 se puede ver que las frecuencias de relaciones entre los diferentes segmentos cromosómicos son variables. La variabilidad de frecuencia de relaciones entre los diferentes segmentos cromosómicos podría atribuirse a estas dos razones:

- Distancia entre los cromosomas: Si las relaciones entre diferentes segmentos cromosómicos se deben a eventos de conversión génica, se podria decir que la frecuencias de relaciones encontradas son una aproximación a las frecuencias de eventos de conversión génica entre los diferentes segmentos cromosómicos. Dado que los eventos de conversión génica son consecuencia de fenómenos de recombinación y que dos hebras de ADN deben estar cercanas en el espacio para que exista recombinación, cuanto más lejos se encuentren dos segmentos cromosómicos en el núcleo interfásico, menor es la probabilidad de que ocurran eventos de conversión génica entre los mismos. Entonces, las frecuencias de eventos de conversión génica, sería inversamente proporcionales a la distancias entre segmentos cromosómicos y que, cuanto más alta la frecuencia de relaciones entre dos grupos, más cerca se encuentran dichos segmentos cromosómicos en el núcleo interfásico. Con esto se podría considerar a la Figura 24 como una aproximación a un mapa de posiciones relativas de los diferentes segmentos cromosómicos en el núcleo interfásico de *T. brucei*.

- Tiempo en que ha transcurrido desde la ocurrencia de los eventos y el momento en el que fueron detectados: dado que cuanto más tiempo haya pasado desde la ocurrencia de un eventos de conversión génica, más difícil es detectarlo, entonces, en lugar de considerar a las frecuencias de relaciones encontradas como aproximación a las frecuencias de eventos de conversión génica entre los diferentes segmentos cromosomicos, podría decirse que es una aproximación al tiempo en que ha transcurrido desde los eventos de conversión génica hasta ahora. Si en lugar de considerar a la probabilidad de eventos de conversión génica como una aproximación de distancia entre segmentos, se postula que solo ocurren eventos de conversión

génica si dos segmentos están muy juntos, se podría decir que las diferencias entre frecuencias de relaciones se deben a que las posiciones de los diferentes segmentos cromosómicos ha variado en el tiempo, por lo cual la figura 24 se podría considerar como un esquema de posiciones relativas de los diferentes segmentos en el tiempo.

Conclusiones

En el presente trabajo, se presentan evidencias que indican que la conversión génica ocurre entre copias inactivas de estos genes. Por otra parte, se aprecia una alta frecuencia de eventos de conversión génica entre ciertos clusters de genes VSG contenidos en algunos segmentos cromosómicos. Estas frecuencias pueden ser consideradas como una indicación de la distancia entre los segmentos cromosómicos en el núcleo interfásico. Dicha inferencia se basa en que para que pueda existir conversión génica, los genes involucrados deben estar cerca en el espacio, y los segmentos en cuestión presentan una cierta proporción de segmentos génicos candidatos a haber sufrido eventos de conversión.

Perspectivas

Dadas las conclusiones del presente trabajo, cabe preguntarse cuales son las secuencias que actúan para el reconocimiento de la maquinaria de recombinación en estos hot spots, y donde se encuentran.

Para responder esa pregunta, se deberían analizar las secuencias que presentan evidencia de haber participado en eventos de conversión génica y sus secuencias flanqueantes para ver si tienen alguna propiedad y así identificar a las posibles secuencias involucradas en el reconocimiento por la maquinaria de recombinación, y una vez confirmadas, se podrían buscar en el resto de los VSGs que no mostraron evidencia de participar en eventos de conversión génica para poder ubicar posibles secuencias que hayan participado en eventos de conversión genica y que por la metodología implementada, no se hallan podido detectar.

Bibliografia

Altschul, S.; Gish, W.; Miller, W.; Myers, E. Lipman, D. (1990). “Basic local alignment search tool”. Journal of Molecular Biology. Vol. 215; 403-410

Barry, J.D.; McCulloch, R. (2001). “Antigenic variation in trypanosomes: enhanced phenotypic variation in an eukaryotic parasite”. Advances in Parasitology , Vol 49 :1-70.

Barry, J.D.; Marcello, L.; Morrison, L.J.; Read, A.F.; Lythgoe, K.; Jones, N.; Carrington, M.; Blandin, G.; Böhme, U.; Caler, E.; Hertz-Flowler, C.; Renauld, H.; El-Sayed, N.; Berriman, N.(2005) “What the genome sequence is revealing about trypanosome antigenic variation” Biochemical Society Transactions Vol 33, 986-989.

Batagelj, V., Mrvar, A. (1998) “Pajek – Program for Large Network Analysis” Connections Vol. 21, Nro 2, 47-57

Berriman, M. et al. (2005) “The Genome of the African Trypanosome: *Trypanosoma brucei*” Science Vol 309, 409-422

Blum, M.; Down,J.; Gurnett, A.; Carrington, M.; Turner, M.; Wiley, D. (1993) “A structural motif in the variant surface glycoproteins of *Trypanosoma brucei*” Nature Vol. 392, 603-609

Bruce, D., Hamerton, A., Watson D.; Bruce R., (1914) “Description of a Strain of Trypanosoma brucei from Zululand. Part II.--Susceptibility of Animals” Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character, Vol. 87, No. 598, pp. 511-516

Catanho, M.; Mascarenhas, D.; Degrave W; de Miranda, AB (2006). “BioParser: A tool for processing of sequence similarity analysis reports”. Applied Bioinformatics. Vol.5, Nro.1, 49-53.

Carillo, H. and Lipman, D. (1988) “The multiple sequence alignment problem in biology”. SIAM Journal of Applied Mathematics, Vol. 48., Nro. 5, 1073-1082.

Carrington, M.; Miller, N.; Blum, M; Roditi, I; Wiley, D.; Turner, M. (1991) “Variant specific glycoprotein of Trypanosoma brucei consists of two domains each having an independently conserved pattern of cysteine residues”. Journal of Molecular Biology Vol. 221, Nro. 3, 823-835.

Chaves, I.; Rudenko, I; Dirks-Mulder, A; Cross, M.; Borst, P. (1999) “Control of variant surface glycoprotein gene-expression sites in *Trypanosoma brucei*” The EMBO Journal Vol.18, Nro. 17, 4846–4855. http://www.nature.com/emboj/journal/v18/n17/pdf/7591894a.pdf

Chattopadhyay, A., Jones,N. Nietlispach, D., Nielsen, P., Voorheis, P., Mott, H., Carrington, M.“Structure of the C-terminal Domain from *Trypanosoma brucei* Variant Surface Glycoprotein MITat1.2” Journal of Biochimestry. http://www.jbc.org

Dayhoff, M., Schwartz, R., Orcutt, B. (1978) “A model of evolutionary change in proteins” Atlas of Protein Sequence and Structure Vol 5, Dayhoff (ed.). 345-352.

Edgar, R. (2004) “MUSCLE: multiple sequence alignment with high accuracy and high throughput” Nucleic Acids Research, Vol. 32, Nro. 5

Gastellu-Etchegorry, M., Helenport, J, Pecoul, B., Jannin, J., Legros, D. (2001) “Availability and affordability of treatment for Human African Trypanosomiasis” Tropical Medicine and International Health, Vol 6, Nro 11, 957-959. http://www.blackwell-synergy.com/doi/pdf/10.1046/j.1365-3156.2001.00764.x

Henikoff, S., Henikoff, J., (1992). “Amino acid substitution matrices from protein blocks”. Proceedings in Natural. Academy of Science. Nro. 89, 10915-10919

Jackson, A. (2007) “Tandem gene arrays in *Trypanosoma brucei*: Comparative phylogenomic analysis of duplicate sequence variation” BioMed Central. http://www.biomedcentral.com/1471-2148/7/54

Kumar S, Tamura K & Nei M (2004) “MEGA3: Integrated Software for Molecular Evolutionary Genetics Analysis and Sequence Alignment” Briefings in Bioinformatics Nro5,150-163.

McCulloch, R et al (1997) “Gene Conversions Mediating Antigenic Variation in Trypanosoma brucei Can Occur in Variant Surface Glycoprotein Expression Sites Lacking 70-Base-Pair Repeat Sequences” Molecular and Celular Biology, Vol 17, Nro 2, 833-843.

Saitou, N. and Nei., M. (1987). “The neighbor-joining method: a new method for reconstructing phylogenetic trees.” Molecular Biology and Evolution, Vol. 4, Nro. 4, 406-425

Sneath, P., Sokal., R. (1973) “Numerical Taxonomy: The Principles and Practice of Numerical Classification”. Ed. Freeman. San Francisco, 230-234

Stich, A.; Abel, P.M.; Krishna, S. (2002) “Human African trypanosomiasis”. BMJ, Nro 325, 203-206. http://www.bmj.com/cgi/reprint/325/7357/203.

Stockdale, C.; Swiderski, M.; Barry, D.; McCulloch, R.(2008) “Antigenic Variation in Trypanosoma brucei: Joining the DOTs ” Plos Biology Vol.6, Nro. 7, 1386-1391

Tamura, K., Dudley, J., Nei, M., Kumar, S., (2007) “MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0” Molecular and Biology Evolution Vol. 24, Nro. 8, 1596-1599.

Thompson, J., Higgins, D., Gibson, T. (1994) “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap *penalties* and weight matrix choice” Nucleic Acids Research Vol. 22, No. 22, 4673-4680

Vanhamme, L.; Pays, E.; McCulloch, R.; Barry,J. D. (2001) “An update on antigenic variation in African trypanosomes”.TRENDS in parasitology Vol 17, 338-343.

Wilbur, W., Lipman, D. (1983) “Rapid similarity searches of nucleic acid and protein data banks” Vol. 80, 726-730,. National Academy of Science USA http://www.pnas.org/cgi/content/abstract/80/3/726

World Health Organization ,(1998) “Control and surveillance of African Trypanosomiasis” WHO Tecnichal Reports Series Nro 881. http://whqlibdoc.who.int/trs/WHO_TRS_881.pdf

World Health Organization (2000) “Report on Global Surveillance of Epidemic-prone Infectious Diseases” World Health Organization Department of Communicable Disease Surveillance and Response

http://www.who.int/csr/resources/publications/surveillance/a_tryps.pdf

Zitzmann, N,Mehlert, A, Carroué, S, Rudd, M, Ferguson, M. (1999) “Protein structure controls the processing of the N-linked oligosaccharides and glycosylphosphatidylinositol glycans of variant surface glycoproteins expressed in bloodstream form *Trypanosoma brucei*” Glycobiology Vol. 10,Nro 3,. 243-249.