

# Evolución de secuencias palindromicas en genomas de cianobacterias

Eduardo Padilla Mendoza

2023-08-01



# Contents

<b>Resumen</b>	<b>5</b>
<b>1 Introducción</b>	<b>7</b>
<b>2 Métodos</b>	<b>9</b>
2.1 Abundancia de palíndromos . . . . .	9
2.2 Significancia de los conteos observados . . . . .	12
2.3 Visualización de la abundancia: OE vs Frecuencia Observada cada 1000nt . . . . .	14
2.4 Filogenia . . . . .	14
2.5 Identificación de casos relevantes . . . . .	16
2.6 Reconstrucción Ancestral de sitios palindrómicos en ortólogos . .	16
<b>3 Resultados</b>	<b>23</b>
3.1 Clado Calothrix . . . . .	23
3.2 Clado A18-40 . . . . .	36



# Resumen

El palíndromo altamente iterado 1 (HIP1 por sus siglas en inglés) cuya secuencia es 5'-GCGATCGC-3', está ampliamente representado en las cianobacterias con excepción de las pico-cianobacterias marinas y otros linajes. El origen de HIP1 y su función (si es que tiene alguna) permanecen desconocidos. Se ha observado que el sitio de reconocimiento (5'-Gm6ATC-3') de la enzima Dam metiltransferasa específica para adenina N6 de clase D12 (Dam-met) y el sitio de reconocimiento de DmtC (5'-m5CGATCG-3') están contenidos en HIP1, lo que sugiere una posible relación. Sin embargo, la asociación funcional de otros genes con HIP1 no se ha reportado.



# Chapter 1

## Introducción



# Chapter 2

## Métodos

Se descargaron 2 conjuntos de genomas de cianobacterias del servidor del NCBI (<https://www.ncbi.nlm.nih.gov/genome/browse/#!/prokaryotes>).

Estos conjuntos corresponden a:

- 269 genomas completos y aquellos que solo contenian el cromosoma (**complete\_chr**)
- 165 genomas nuevos usados en Cabello-Yeves et al. (2022) (**pico**)

Dichos genomas fueron descargados en formato Genebank (.gbk o gbff).

### 2.1 Abundancia de palíndromos.

Una vez descargados los genomas, el siguiente paso fue calcular el valor observado y esperado de repeticiones de todos los posibles octámeros palindrómicos de 8 nucleótidos.

El valor observado es el número de veces que cada octámero palindrómico se repite a lo largo de cada genoma. El valor esperado se calculó mediante un **modelo de markov de 3er orden**.

#### 2.1.1 Modelos de Markov

En una cadena de Markov, el valor tomado por una variable aleatoria depende de los valores tomados por la variable aleatoria en un estado anterior. El número de estados históricos que influyen en el valor de la variable aleatoria en un lugar dado a lo largo de la secuencia también se conoce como el **grado del proceso de Markov**. El modelo de cadena de Markov de **primer grado** tiene parámetros  $|\Sigma| + |\Sigma|^2$ , correspondientes a las frecuencias de nucleótidos individuales así como a las frecuencias de dinucleótidos. De esta manera, este modelo permite que una posición sea dependiente de la posición anterior. Sin embargo, las frecuencias

se modelan de manera invariable en la posición y, por lo tanto, pueden no ser adecuadas para modelar señales. Este modelo de secuencia  $M$  se define sobre el espacio muestral  $\Sigma^*$  y asigna una probabilidad a cada secuencia  $x$  de longitud  $n(x)$  sobre  $\Sigma^*$ :

$$P(x|M) = P_1(x_1) \prod_{i=2, \dots, n(x)} P_2(x_i|x_{i-1}, \dots, x_{i-n}) \quad (2.1)$$

donde  $P_1$  es una función de probabilidad en  $\Sigma$  que modela la distribución de  $\alpha$ 's en la primera posición de la secuencia y  $P_2$  es la función de probabilidad condicional en  $\Sigma \times \Sigma$  que modela la distribución de  $\beta$ 's en la posición  $i > 1$  en el símbolo alfabético  $\alpha$  en la posición  $i - 1$ . La estimación de parámetros se hace utilizando el estimador de **probabilidad máxima**. Las probabilidades de transición se estiman utilizando el teorema de Bayes, como se muestra a continuación:

$$P_2(\beta|\alpha) = \frac{P(\alpha\beta)}{P(\alpha)} \quad (2.2)$$

De esta manera, las probabilidades transicionales condicionales de encontrar una base  $\beta$  en la posición  $(i)$  dado que la base  $\alpha$  se encontró en la posición  $(i - 1)$  se calculan encontrando la abundancia del dinucleótido  $\alpha\beta$  como una fracción de la abundancia del nucleótido  $\alpha$ .

### Ejemplo:

Considerando una la secuencia de 25 nucleótidos.

*Seq = AACGTCTCTATCATGCCAGGATCTG*

Al considerar los modelos de cadena de Markov de **primer grado**, es necesario calcular los  $4 - \text{parmetros}$  correspondientes a las **frecuencias de nucleótidos individuales** y los  $4^2$  parámetros correspondientes a las **frecuencias de dinucleótidos**. Los parámetros de  $\Sigma$  son:

$$\begin{aligned} \Sigma &= \{ \text{freq}(A), \text{freq}(C), \text{freq}(G), \text{freq}(T), \} \\ &= \left\{ \frac{6}{25}, \frac{7}{25}, \frac{7}{25}, \frac{5}{25} \right\} \end{aligned} \quad (2.3)$$

Para calcular  $P_2$ , los valores de probabilidad condicional  $\Sigma \times \Sigma$ , las frecuencias de dinucleótidos y las probabilidades se calculan a partir de los datos de secuencia. Las frecuencias de los dinucleótidos y las probabilidades se muestran a continuación (con los números entre paréntesis que representan las probabilidades):

$$\Sigma \times \Sigma = \left\{ \begin{array}{llll} freq(AA) = \frac{1}{24} & freq(AC) = \frac{1}{24} & freq(AT) = \frac{3}{24} & freq(AG) = \frac{1}{24} \\ freq(CA) = \frac{2}{24} & freq(CC) = \frac{1}{24} & freq(CT) = \frac{3}{24} & freq(CG) = \frac{1}{24} \\ freq(TA) = \frac{1}{24} & freq(TC) = \frac{4}{24} & freq(TT) = \frac{0}{24} & freq(TG) = \frac{1}{24} \\ freq(GA) = \frac{1}{24} & freq(GC) = \frac{1}{24} & freq(GT) = \frac{1}{24} & freq(GG) = \frac{1}{24} \end{array} \right\} \quad (2.4)$$

A continuación, las probabilidades condicionales se calculan utilizando el teorema de Bayes (consulte la Ecuación (2.2)). Por ejemplo, la probabilidad de encontrar  $C$  en la posición  $i+1$  dado que se ha encontrado una  $A$  en la posición  $(i)$  es:

$$P(C|A) = \frac{P_{AC}}{P_A} = \frac{\frac{1}{24}}{\frac{6}{25}} \quad (2.5)$$

Para secuencias grandes, la probabilidad condicional  $P(S_i|S_{i-1})$  se aproxima a:

$$P(S_i|S_{i-1}) = \frac{freq(S_i S_{i-1})}{freq(S_{i-1})} \quad (2.6)$$

Las probabilidades condicionales para la secuencia de ejemplo se muestran en (2.4). Usando estos parámetros del modelo, la probabilidad de encontrar el patrón *CAAT* en esta secuencia usando el **modelo de Markov de primer orden** de la secuencia subyacente sería igual a:

$$\begin{aligned} P(C)P(A|C)P(A|A)P(T|A) &= P(C) \cdot \frac{P(CA)}{P(C)} \cdot \frac{P(AA)}{P(A)} \cdot \frac{P(AT)}{P(A)} \\ &= \left(\frac{7}{25}\right) \cdot \left(\frac{50}{168}\right) \cdot \left(\frac{25}{144}\right) \cdot \left(\frac{75}{144}\right) \\ &= 0.0075 \end{aligned} \quad (2.7)$$

### 2.1.2 Modelo de Markov de orden 1 para hallar octanucleótidos

Por ejemplo, para una octanucleótido de 8 letras, digamos HIP1:

$$W = GCGATCGC$$

Los parametros de  $\Sigma$  corresponden a:

$$\Sigma = \{freq(A), freq(C), freq(G), freq(T)\} \quad (2.8)$$

Los valores de probabilidad condicional de  $\Sigma \times \Sigma$  son:

$$\Sigma \times \Sigma = \begin{Bmatrix} freq(AA) & freq(AC) & freq(AT) & freq(AG) \\ freq(CA) & freq(CC) & freq(CT) & freq(CG) \\ freq(TA) & freq(TC) & freq(TT) & freq(TG) \\ freq(GA) & freq(GC) & freq(GT) & freq(GG) \end{Bmatrix} \quad (2.9)$$

Si queremos usar un **modelo de orden 1**, la probabilidad de hallar  $W$  segun las ecuaciones (2.1) y(2.2) es:

$$\begin{aligned} P(W) &= P(G) \cdot P(C|G) \cdot P(G|C) \cdot P(A|G) \cdot P(T|A) \cdot P(C|T) \cdot P(G|C) \cdot P(C|G) \\ &= P(G) \cdot \frac{P(GC)}{P(G)} \cdot \frac{P(CG)}{P(C)} \cdot \frac{P(GA)}{P(G)} \cdot \frac{P(AT)}{P(A)} \cdot \frac{P(TC)}{P(T)} \cdot \frac{P(CG)}{P(C)} \cdot \frac{P(GC)}{P(G)} \\ &= P(GC) \cdot \frac{P(CG)}{P(C)} \cdot \frac{P(GA)}{P(G)} \cdot \frac{P(AT)}{P(A)} \cdot \frac{P(TC)}{P(T)} \cdot \frac{P(CG)}{P(C)} \cdot \frac{P(GC)}{P(G)} \end{aligned} \quad (2.10)$$

finalmente:

$$P(W) = \frac{P(GC) \cdot P(CG) \cdot P(GA) \cdot P(AT) \cdot P(TC) \cdot P(CG) \cdot P(GC)}{P(C) \cdot P(G) \cdot P(A) \cdot P(T) \cdot P(C) \cdot P(G)} \quad (2.11)$$

### 2.1.3 Abundancia de acuerdo a la frecuencia observada y tasa OE

Adicionalmente se calculó una abundancia de acuerdo a la frecuencia observada cada 1000 nucleótidos (**FrecObs**) y otra en base a la tasa de sitios observados sobre esperados (**OE**).

## 2.2 Significancia de los conteos observados

Para darle una significancia estadística al conteo se usó una **prueba binomial** y un test **FDR**.

### 2.2.1 Prueba binomial.

Para calcular la probabilidad de que el **conteo esperado**, el cual sigue una distribución binomial, tome valores MAYORES O IGUALES al **conteo observado**, usamos la función ***pbinom***

```
pbinom(q, size, prob, lower.tail = FALSE)
```

Donde:

- **q**: Cuantil o vector de cuantiles
- **size**: Numero de experimentos ( $n \geq 0$ )

- **prob**: Probabilidad de éxito en cada experimento
- **lower.tail**: si es TRUE, las probabilidades son  $P(X \leq x)$ , o  $P(X > x)$  en otro caso.

Tomemos un caso particular del conteo:

Spp	Palindrome	Observed	Markov (Expected)	GenomeSize
336-3	GCGATCGC	6202	65.396286071305	6420126

La probabilidad de que se observen **6202** sitios *GCGATCGC*, O MAS, si el número de sitios posibles en el genoma es **6420119** ( $6420126 - 8 + 1$ , es decir  $GenomeSize - k + 1$ ) y la probabilidad de observar dicho sitio es de: **1.018615e-05** ( $\frac{65.3962860713054}{6420126 - 8 + 1}$ , es decir  $\frac{Expected}{GenomeSize - k + 1}$ ), es casi **0**.

En otras palabras, la probabilidad de que suceda lo que estoy observando es muy baja.

### 2.2.2 FDR

Para estudios en los que se realizan miles de test de forma simultánea, el resultado de estos métodos es demasiado conservativo e impide que se detecten diferencias reales. Una alternativa es controlar el false discovery rate o FDR.

Para nuestros datos el FDR se calculó en R de usando los valores obtenidos de la prueba binomial:

```
p.adjust(pval, method="fdr")
```

Donde **pval** es la probabilidad obtenida de la prueba binomial.

### 2.2.3 Conjuntos de conteos de acuerdo a la significancia

Se crearon 4 conjuntos de resultados de acuerdo a 4 valores mínimos de significancia de acuerdo al FDR:

- **sel32** ( $1 \times 10^{-32}$ )
- **sel64** ( $1 \times 10^{-64}$ )
- **sel128** ( $1 \times 10^{-128}$ )
- **sel256** ( $1 \times 10^{-256}$ )

El conjunto más laxo corresponde a **sel32** ya que su valor de corte de FDR es  $1 \times 10^{-32}$ , debido a esto, es el conjunto con más palíndromos (Figura 2.1). Por otro lado, el conjunto **sel256** es el conjunto más restrictivo ya que su valor de corte de FDR es de  $1 \times 10^{-256}$ , y por lo tanto tiene menos palíndromos (Figura 2.2).

## 2.3 Visualización de la abundancia: OE vs Frecuencia Observada cada 1000nt

Para visualizar la abundancia creamos un gráfico que muestra el enriquecimiento OE vs la abundancia por cada 1000 nucleótidos. Esto se hizo para cada conjunto de significancia y para cada conjunto de genomas.

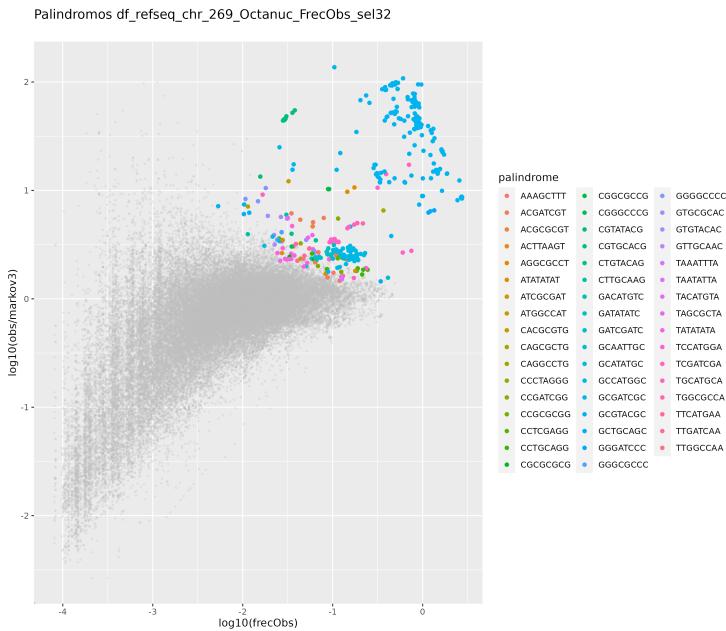


Figure 2.1: \*\*Enriquecimiento versus abundancia de palíndromos octámeros en el conjunto de genomas complete\_chr con un  $FDR \leq 1 \times 10^{-32}$ .\*\* Enrichment (\*\*O/E\*\*) in function of the frequency of the motif every 1000 nt (\*\*FreqObs\*\*). Each point represents a palindromic octamer of a genome.

## 2.4 Filogenia

Se infirieron filogenias para los dos conjuntos de genomas. Para esto usamos el software **Orthofinder** (Emms and Kelly (2019)), el cual utiliza **FastME** para inferir la filogenia (Lefort et al. (2015)). **FastME** proporciona algoritmos de distancia para inferir filogenias. FastME se basa en una evolución mínima equilibrada, que es el principio mismo de Neighbor Joining (NJ).

El software se corrió en la línea de comandos de la siguiente manera:

```
orthofinder -f genomas/
```

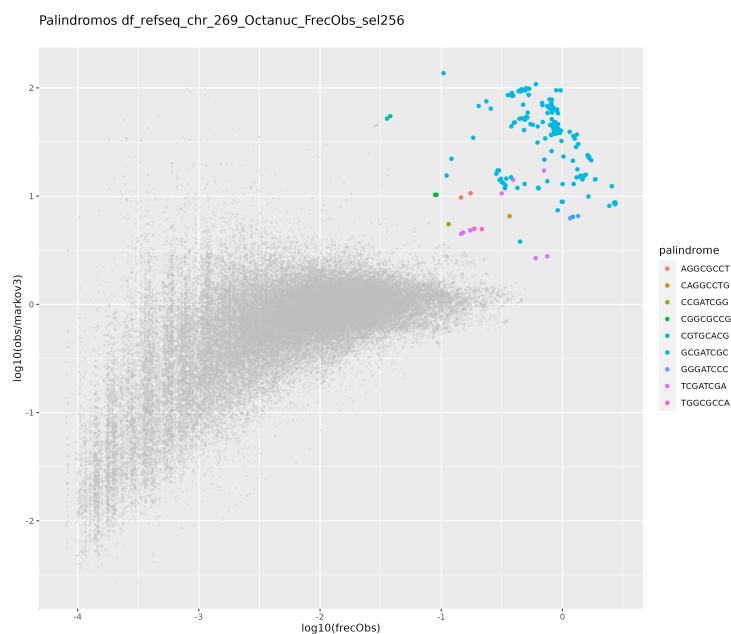


Figure 2.2: \*\*Enriquecimiento versus abundancia de palíndromos octámeros en el conjunto de genomas complete\_chr con un  $FDR \leq 1 \times 10^{-256}$ .\*\* Enriquecimiento (\*\*O/E\*\*) en función de la frecuencia del motivo cada 1000 nt (\*\*FrecObs\*\*). Cada punto representa un palíndromo octámero de un genoma.

### 2.4.1 Anotación de la filogenia

Para tener una forma de más visual de entender la distribución de los palíndromos en los genomas, anotamos las filogenias de acuerdo a su abundancia. Se anotaron 4 filogenias según la significancia (**sel32**, **sel64**, **sel128** y **sel256**) para los 2 conjuntos de genomas. Además, esta anotación se hizo para la abundancia de acuerdo a la Frecuencia Observada por cada 1000 nucleotidos (*FrecObs*) (Figura 2.3) y a la tasa de Observados sobre esperados (*OE*) (Figura 2.4).

La anotación de las filogenias consistió en agregarles un heatmap que mostrara la abundancia de cada palíndromo y un diagrama de barras que indicara aquel palíndromo con mayor abundancia.

## 2.5 Identificación de casos relevantes

De acuerdo a las filogenias anotadas, se buscaron aquellos casos en los que HIP1 o algún otro palíndromo se hubiera ganado o perdido abruptamente y en su lugar hubiese otro palíndromo abundante. Además, se buscó que en aquellos casos, las ramas en la filogenia no fueran tan largas. Esto se hizo de manera visual revisando el diagrama de barras que mostraba el palíndromo más abundante para cada especie. En total hubo 6 subclados que mostraban cambios abruptos en la abundancia de sus palíndromos (Figura 2.5).

También se hallo un caso interesante en el conjunto **pico** (**clado A18-40**) el cual sirvió como punto de partida para análisis posteriores. En este caso se muestra que la especie *Synechococcus* A18-40 muestra una tasa OE mucho mayor comparada con las demás especies del clado (Figura 2.6).

## 2.6 Reconstrucción Ancestral de sitios palindrómicos en ortólogos

Para tratar de entender como es que los sitios HIP1 han ido evolucionando, hicimos una reconstrucción de sitios ancestrales y posteriormente construimos varios conjuntos de redes para visualizar dicha evolución.

### 2.6.1 Ortólogos

Para simplificar la reconstrucción de secuencias ancestrales usamos únicamente los ortólogos. Para obtener esto usamos el pipeline `get_homologues`:

```
get_homologues.pl -d gbff -t 0 -M -n PPN
```

Después de obtener los ortólogos filtramos:

- aquellos que no estuvieran en las 6 especies del clado
- aquellos que tuvieran mas de una copia (parálogos)
- aquellos sin sitios HIP1

## 2.6. RECONSTRUCCIÓN ANCESTRAL DE SITIOS PALINDRÓMICOS EN ORTÓLOGOS17



Figure 2.3: \*\*Filogenia del conjunto de genomas \*complete\_chr\* anotada de acuerdo a la Frecuencia observada cada 1000 nt (FreqObs).\*\* La abundancia visualizada en esta filogenia es de acuerdo al conjunto \*\*sel256\*\*, es decir conteos con un  $FDR \leq 1 \times 10^{-256}$ . La filogenia muestra 269 especies, frente a la filogenia se muestra un heatmap que indica la abundancia de cada palíndromo. Frente al Heatmap se muestra un Diagrama de barras el cual indica el palíndromo mas abundante de entre todos.

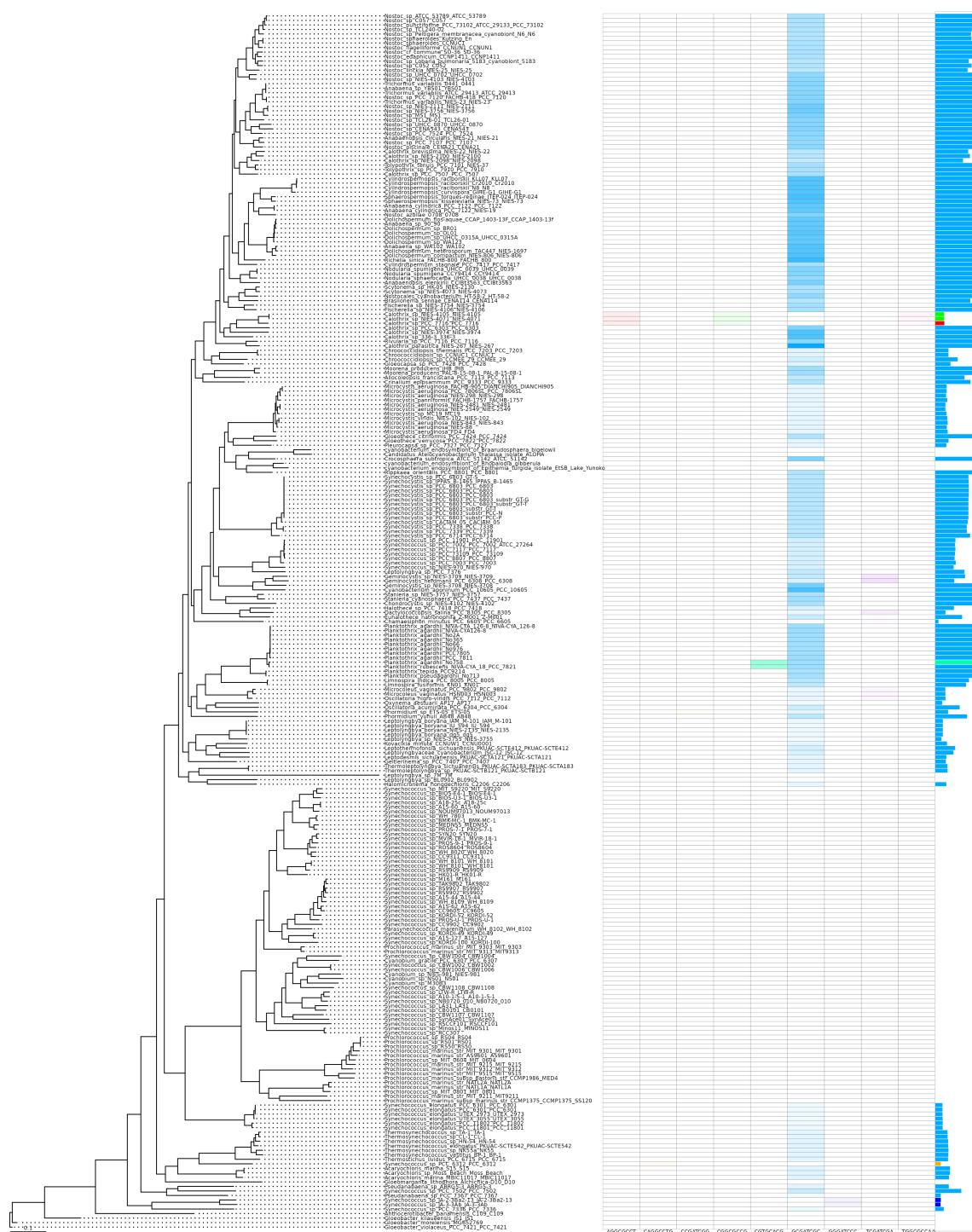


Figure 2.4: \*\*Filogenia del conjunto de genomas \*complete\_chr\* anotada de acuerdo a la tasa de observados sobre esperados (OE).\*\* La abundancia visualizada en esta filogenia es de acuerdo al conjunto \*\*sel256\*\*, es decir conteos con un  $FDR \leq 1 \times 10^{-256}$ . La filogenia muestra 269 especies, frente a la filogenia se muestra un heatmap que indica la abundancia de cada palíndromo. Frente al Heatmap se muestra un Diagrama de barras el cual indica el palindromo mas abundante de entre todos.

## 2.6. RECONSTRUCCIÓN ANCESTRAL DE SITIOS PALINDRÓMICOS EN ORTÓLOGOS19

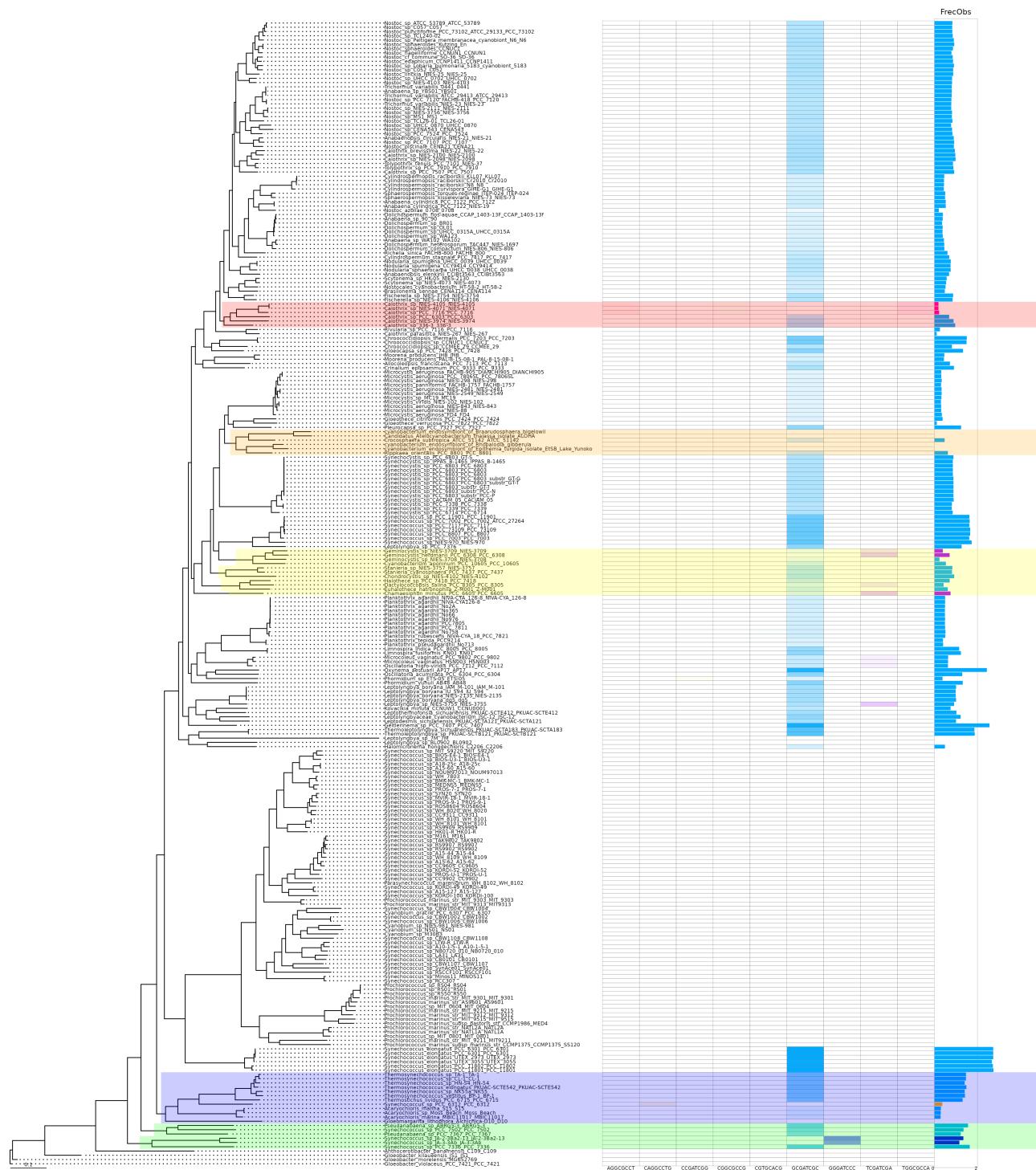


Figure 2.5: \*\*Casos de interés.\*\* En la figura se muestran remarcados los casos interesantes: \*\*clado calothrix\*\* (rojo), \*\*clado cyanobacterium\*\* (naranja), \*\*clado geminocystis\*\* (amarillo), \*\*clado thermosynechococcus\*\* (azul), \*\*clado pseudoanabaena\*\* (verde).

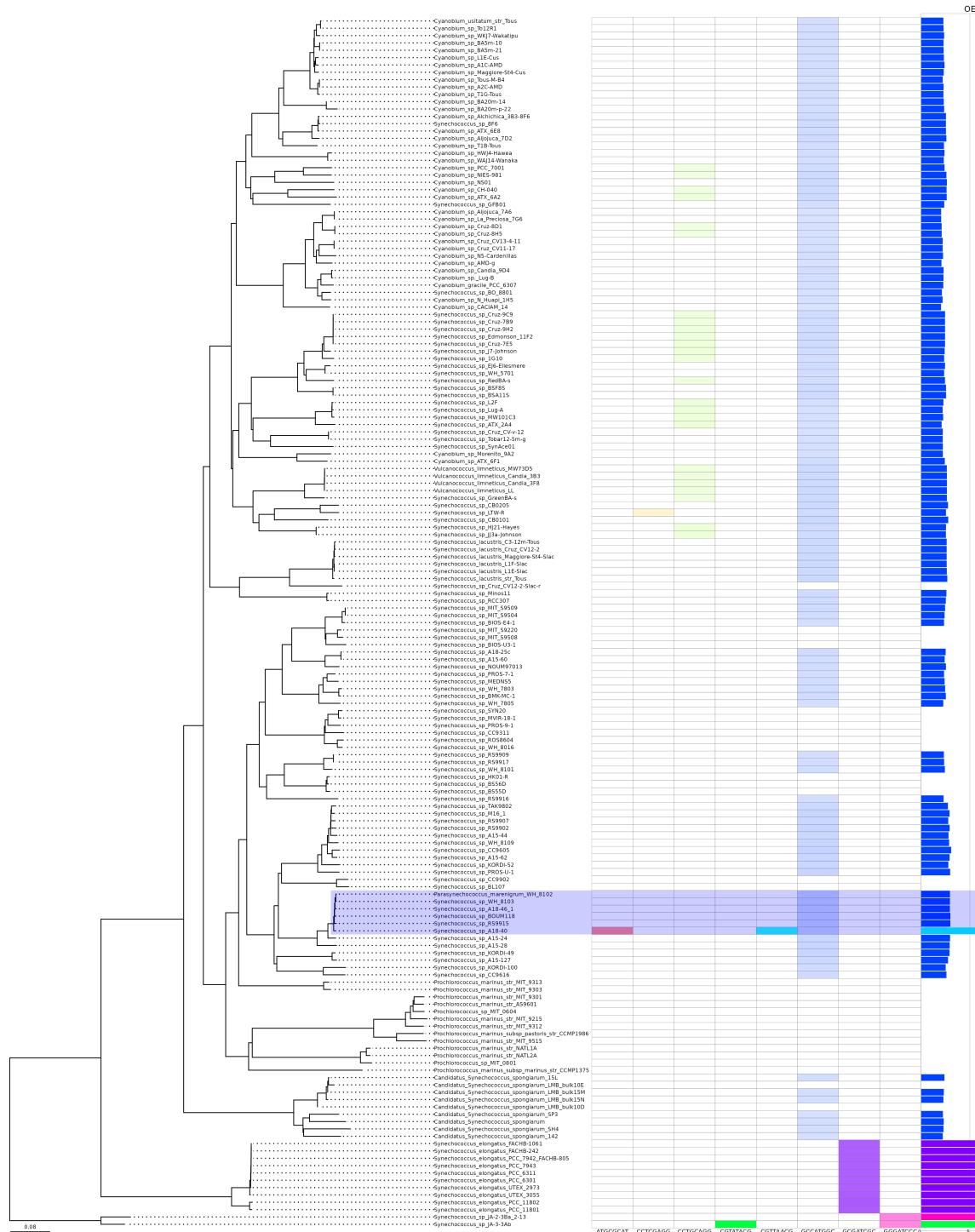


Figure 2.6: \*\*Casos de interés.\*\* En la figura se muestra remarcado el \*\*clado A18-40\*\* (azul).

## 2.6.2 Reconstrucción

Para hacer la reconstrucción usamos la paguetería de R `phangorn`, la cual proporciona varios métodos para estimar estados de caracteres ancestrales con Máxima Parsimonia (MP) o Máxima Verosimilitud (ML). En este caso usamos ML. Adicionalmente podemos asignar los estados ancestrales según la máxima verosimilitud (“ml”):

$$P(x_r = A) = \frac{L(x_r = A)}{\sum_{k \in \{A, C, G, T\}} L(x_r = k)}$$

y el criterio de mayor probabilidad posterior (“bayes”):

$$P(x_r = A) = \frac{\pi_A L(x_r = A)}{\sum_{k \in \{A, C, G, T\}} \pi_k L(x_r = k)}$$

dónde  $L(x_r)$  es la probabilidad conjunta de los estados en las puntas y el estado en la raíz  $x_r$  y  $\pi_i$  son las frecuencias base estimadas del estado  $i$ .

Toda la información de la reconstrucción fue guardada en dos tablas las cuales contienen listas de cada transición entre cada estado. Estas tablas fueron creadas con la siguiente función:

```
source("ASR_Orth_Functions/NodeAndEdges.R")

Create_Transition_Table (SitesTable = "Clados/Callothrix_clade/PALINDROMES/GCGATCGC/Orthologues_PALINDROMES",
                           EvolutionModel = "F81",
                           Method = "bayes",
                           Phylogeny = "Clados/Callothrix_clade/SpeciesTree_rooted.txt",
                           OrthoPath = "Clados/Callothrix_clade/PALINDROMES/GCGATCGC/Only_OPALINDROMES")
```



# Chapter 3

## Resultados

### 3.1 Clado Calothrix

El clado calothrix contiene 6 especies y es de interes ya que segun la filogenia estan estrechamente relacionadas y muestra un cambio en el palindromo mas abundante, pasando de **GCGATCGC** a **TGGCGCCA** (Figure 3.1).

#### 3.1.1 GCGATCGC

##### 3.1.1.1 Red de transiciones

Para hacer mas visual la reconstrucción, construimos una red de las transiciones entre los estados ancestrales. Esto lo hicimos en r usando la función `Create_Transition_Table()`:

```
source("ASR_Orth_Functions/NodeAndEdges.R")
Nodes.Edges <- Create_Transition_Table(SitesTable = "Clados/Calothrix_clade/PALINDROMES/GCGATCGC",
                                         EvolutionModel = "F81",
                                         Method = "bayes",
                                         Phylogeny = "Clados/Calothrix_clade/SpeciesTree_rooted.txt",
                                         OrthoPath = "Clados/Calothrix_clade/PALINDROMES/GCGATCGC/336-3/0")
```

Posteriormente creamos la red usando la función `Create_Network()`:

y visualizamos dicha red .

Para visualizar la red usamos la paqueteria `networkD3`. Hicimos 2 figuras, la (Figura ??) muestra la red como una conexión de nodos a través de vertices con un grosor proporcional al numero de veces que ocurrió cada transición. En dicha red podemos ver algunos nodos con bordes muy gruesos como **GCAATTGC**, **GCAATCGC**, **GCAATAGC**, **GCGATTGC** (Tabla ??).

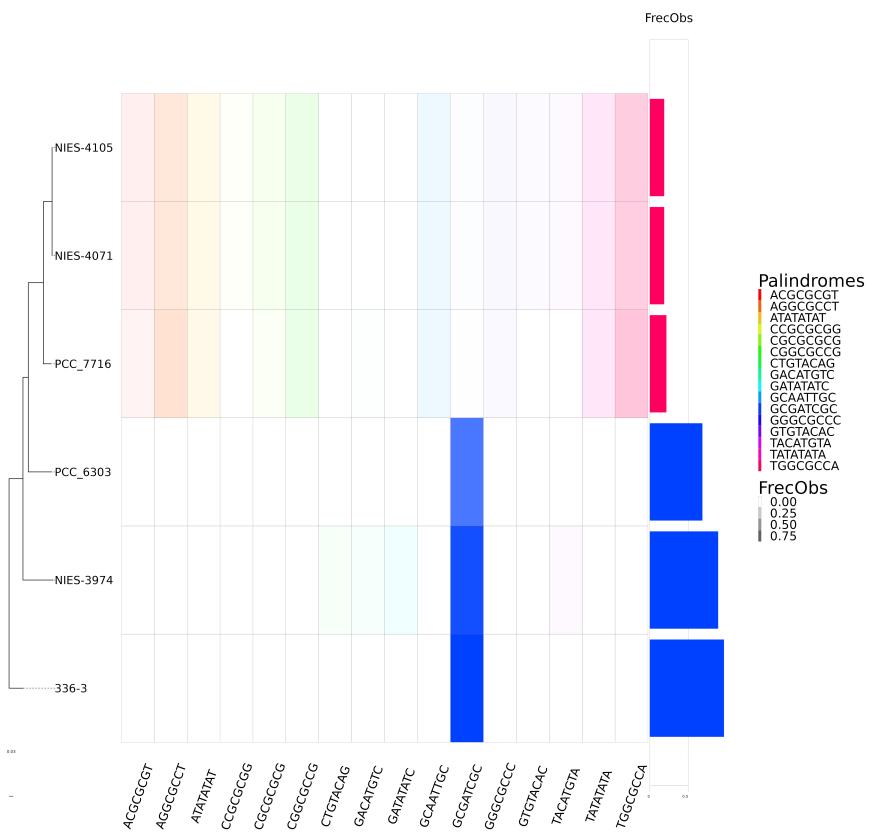


Figure 3.1: \*\*Filogenia anotada del clado *Calothrix*.\*\* En esta imagen se muestra un cambio abrupto en la Frecuencia observada de \*\*GCGCATCGC\*\* en las especies PCC\_6303, NIES-3974 y 336-3.

En la (Figura ??) podemos ver las transiciones de una forma mas ordenada, con el numero de ocurrencias y la dirección en la que ocurrieron.

### 3.1.1.2 Transiciones entre Nodo 9 y Nodo 10

Para entender más como es que se gana o se pierden los sitios palindrómicos revisamos la transición en tre los nodos 9 y 10. Esto es porque es esta transicion de nodos la que separa a los dos subclados entre los que hay una repentino cambio de abundancia de sitios palindrómicos (Figura 3.2).

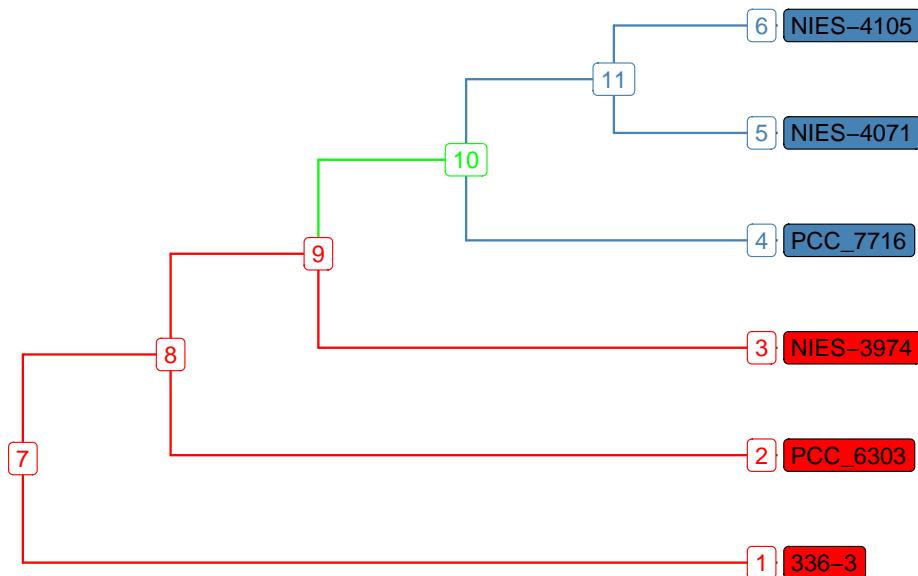


Figure 3.2: **Filogenia del clado Calotrix.** En rojo y azul se muestran los subclados unidos (en verde) por la transición entre los nodos 9 y 10.

Para hacer esto filtramos los datos de la red para mostrar únicamente las transiciones que se dieron entre los nodos 9 y 10 e hicimos las mismas figuras. En la (Figura ??) se muestra la red como una conexión de nodos a través de vértices con un grosor proporcional al número de veces que ocurrió cada transición. En la (Figura ??) podemos ver las transiciones de una forma mas ordenada, con el número de ocurrencias y la dirección en la que ocurrieron.

### 3.1.1.3 Mutaciones en los codones

Para entender como es que se van ganando o perdiendo los sitios palindrómicos hicimos un análisis del tipo mutaciones de los sitios. Esto lo hicimos viendo en que marco de lectura se encontraba cada nodo y revisando la secuencia de aminoácidos que codificaban. En la (Figura 3.3) mostramos 3 gráficos que indican la abundancia de los péptidos codificados por los sitios palindrómicos de acuerdo al marco de lectura en el que se encuentran.

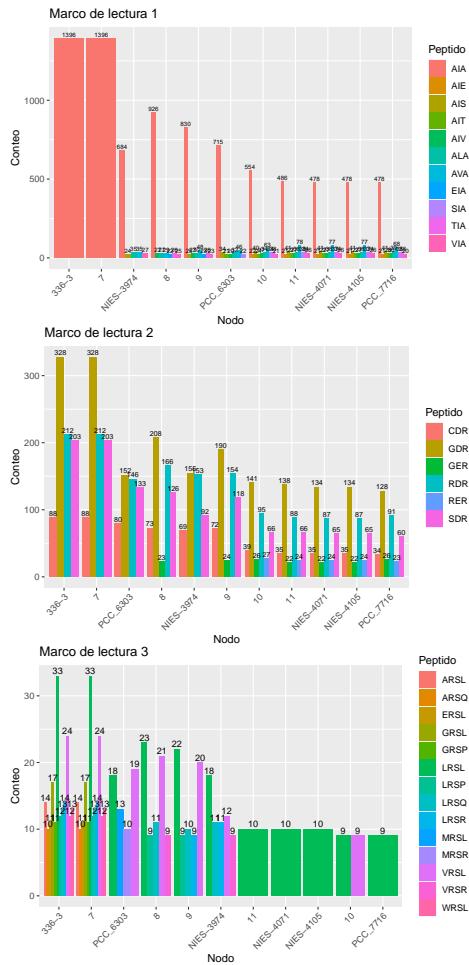


Figure 3.3: \*\*Abundancia de peptidos por cada nodo segun el marco de lectura.\*\*.

En la (Figura 3.4) mostramos 3 gráficos que indican la abundancia del tipo de mutaciones que hay en cada nodo de acuerdo al marco de lectura. Los sitios de mutaciones mostrados pueden ser de los siguientes tipos:

- Conservative (la secuencia de AA cambió pero tiene similitud de acuerdo al score de BLOSUM62)
- ConservativeNoSiteMut (la secuencia de AA cambió pero tiene similitud de acuerdo al score de BLOSUM62. Sin embargo, el sitio no sufrió mutaciones)
- Deletion (La secuencia de AA tiene sufrió 1 o mas delecciones)
- NoMutation (La secuencia de AA no sufrió mutaciones)
- NoSynonym (La secuencia de AA cambió)
- NoSynonymNoSiteMut (La secuencia de AA cambió. Sin embargo, el sitio no sufrió mutaciones.)
- Synonym (El sitio sufrió mutaciones. Sin embargo, la secuencia de AA no cambió.)

#### **3.1.1.4 Análisis de sitios en los cuales su ancestro era HIP1**

Para tratar de entender como es que los sitios HIP1 se pierden hicimos un análisis únicamente en las transiciones en las que el nodo ancestral tenía un sitio HIP1.

En la (Figura 3.5) mostramos 3 gráficos que indican la frecuencia del tipo de sustituciones que hubo para estos casos para cada nodo en cada uno de los marcos de lectura.

En la (Figura 3.6) mostramos 3 gráficos (uno por cada marco de lectura) que indican la frecuencia de las mutaciones en cada uno de los 8 nucleótidos del sitio HIP.

En la (Figura 3.7) mostramos 3 gráficos (uno por cada marco de lectura) que indican la frecuencia del tipo sustitución de bases.

#### **3.1.1.5 Análisis de sitios en los cuales solo el nodo actual tiene HIP1**

Para tratar de entender como es que los sitios HIP se ganan, hicimos un análisis únicamente en las transiciones en las que el nodo actual tenía un sitio HIP1.

En la Figura 3.8 mostramos 3 gráficos (uno por cada marco de lectura) que indican la frecuencia del tipo de sustituciones que hubo para estos casos para cada nodo en cada uno de los marcos de lectura.

### **3.1.2 TGGCGCCA**

#### **3.1.2.1 Red de transiciones**

Para hacer mas visual la reconstrucción, construimos una red de las transiciones entre los estados ancestrales. Esto lo hicimos en r usando la función

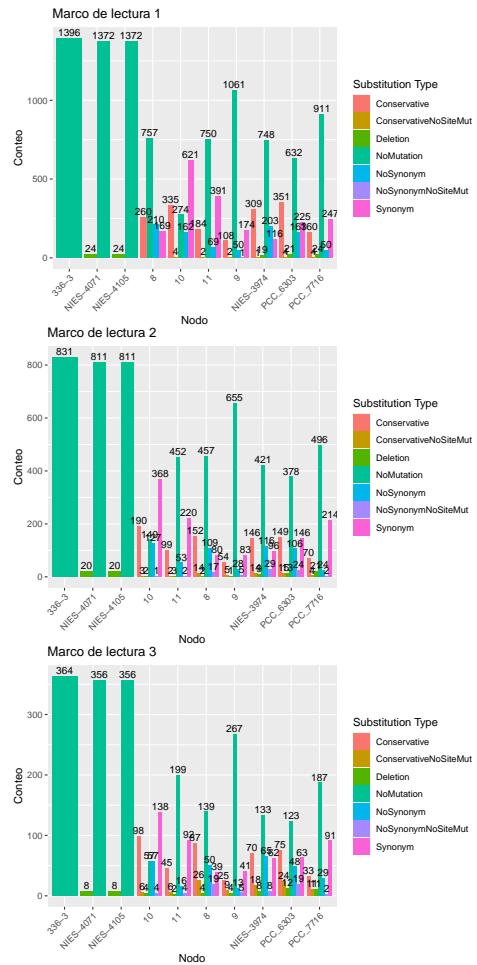


Figure 3.4: \*\*Abundancia del tipo de sustitución por cada nodo segun el marco de lectura.\*\*

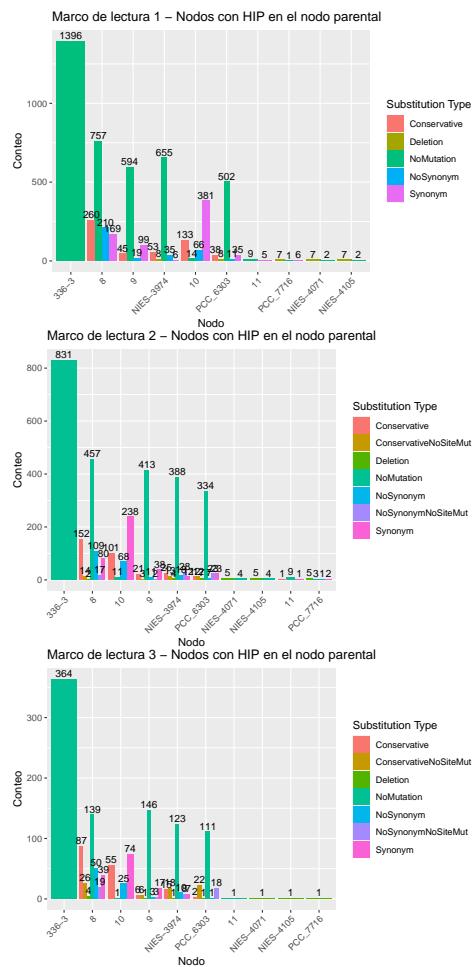


Figure 3.5: \*\*Abundancia del tipo de sustitución por cada nodo segun el marco de lectura. Unicamente para transiciones en los que el nodo ancestral era un sitio HIP1.\*\*

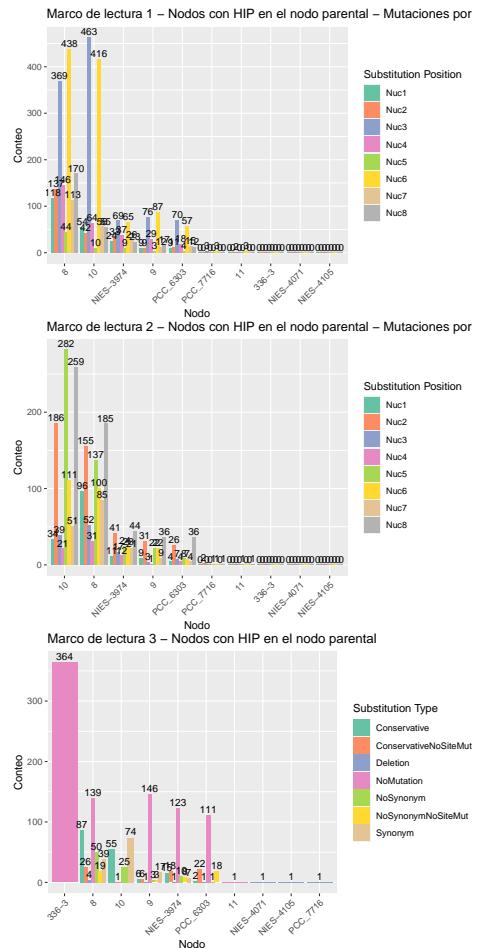


Figure 3.6: \*\*Frecuencia de las mutaciones de cada nucleótido del sitio HIP para cada nodo segun el marco de lectura.\*\*.

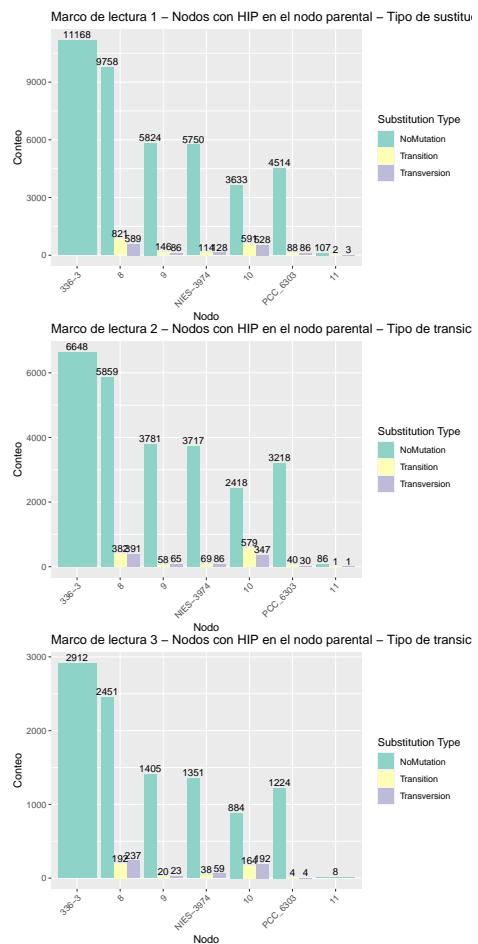


Figure 3.7: \*\*Frecuencia del tipo de sustituciones de base en los sitios HIP para cada marco de lectura\*\*.

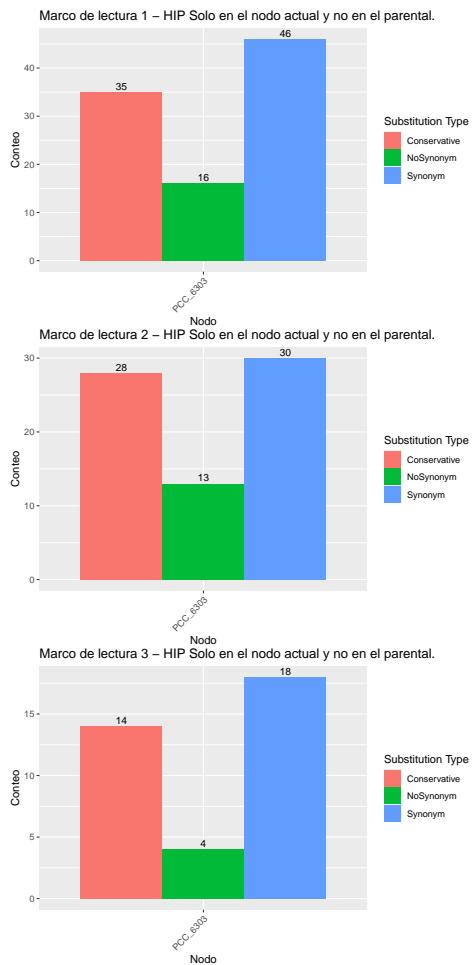


Figure 3.8: \*\*Abundancia del tipo de sustitución por cada nodo segun el marco de lectura. Unicamente para transiciones en los que el nodo actual era un sitio HIP1.\*\*.

```
Create_Transition_Table():
  source("ASR_Orth_Functions/NodeAndEdges.R")
  Nodes.Edges <- Create_Transition_Table(SitesTable = "Clados/Callothrix_clade/PALINDROMES/TGGCGCCA",
                                           EvolutionModel = "F81",
                                           Method = "bayes",
                                           Phylogeny = "Clados/Callothrix_clade/SpeciesTree_rooted.txt",
                                           OrthoPath = "Clados/Callothrix_clade/PALINDROMES/TGGCGCCA/PCC_771")
```

Posteriormente creamos la red usando la función `Create_Network()`:

y visualizamos dicha red .

Para visualizar la red usamos la paquetería `networkD3`. Hicimos 2 figuras, la (Figura ??) muestra la red como una conexión de nodos a través de vértices con un grosor proporcional al número de veces que ocurrió cada transición. En dicha red podemos ver algunos nodos con bordes muy gruesos como **GCAATTGC**, **GCAATCGC**, **GCAATAGC**, **GCGATTGC** (Tabla ??).

En la (Figura ??) podemos ver las transiciones de una forma más ordenada, con el número de ocurrencias y la dirección en la que ocurrieron.

### 3.1.2.2 Transiciones entre Nodo 9 y Nodo 10

Para entender más como es que se gana o se pierden los sitios palindrómicos revisamos la transición entre los nodos 9 y 10. Esto es porque es esta transición de nodos la que separa a los dos subclados entre los que hay una repentino cambio de abundancia de sitios palindrómicos (Figura 3.2).

Para hacer esto filtramos los datos de la red para mostrar únicamente las transiciones que se dieron entre los nodos 9 y 10 e hicimos las mismas figuras. En la (Figura ??) se muestra la red como una conexión de nodos a través de vértices con un grosor proporcional al número de veces que ocurrió cada transición. En la (Figura ??) podemos ver las transiciones de una forma más ordenada, con el número de ocurrencias y la dirección en la que ocurrieron.

### 3.1.2.3 Mutaciones en los codones

Para entender como es que se ganan o pierden los sitios palindrómicos hicimos un análisis del tipo mutaciones de los sitios. Esto lo hicimos viendo en qué marco de lectura se encontraba cada nodo y revisando la secuencia de aminoácidos que codificaban. En la (Figura 3.9) mostramos 3 gráficos que indican la abundancia de los péptidos codificados por los sitios palindrómicos de acuerdo al marco de lectura en el que se encuentran.

En la (Figura 3.10) mostramos 3 gráficos que indican la abundancia del tipo de mutaciones que hay en cada nodo de acuerdo al marco de lectura. Los sitios de mutaciones mostrados pueden ser de los siguientes tipos:

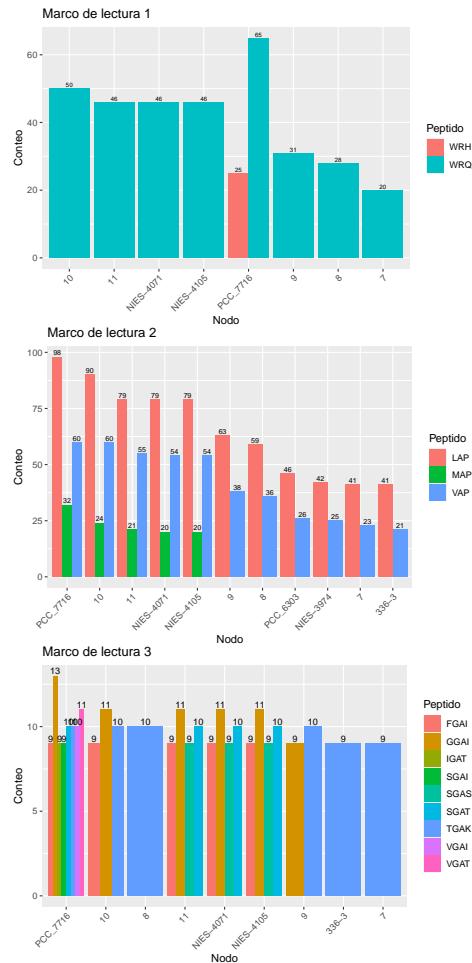


Figure 3.9: \*\*Abundancia de peptidos por cada nodo segun el marco de lectura.\*\*.

- Conservative (la secuencia de AA cambió pero tiene similitud de acuerdo al score de BLOSUM62)
- ConservativeNoSiteMut (la secuencia de AA cambió pero tiene similitud de acuerdo al score de BLOSUM62. Sin embargo, el sitio no sufrió mutaciones)
- Deletion (La secuencia de AA tiene sufrió 1 o mas delecciones)
- NoMutation (La secuencia de AA no sufrió mutaciones)
- NoSynonym (La secuencia de AA cambió)
- NoSynonymNoSiteMut (La secuencia de AA cambió. Sin embargo, el sitio no sufrió mutaciones.)
- Synonym (El sitio sufrió mutaciones. Sin embargo, la secuencia de AA no cambió.)

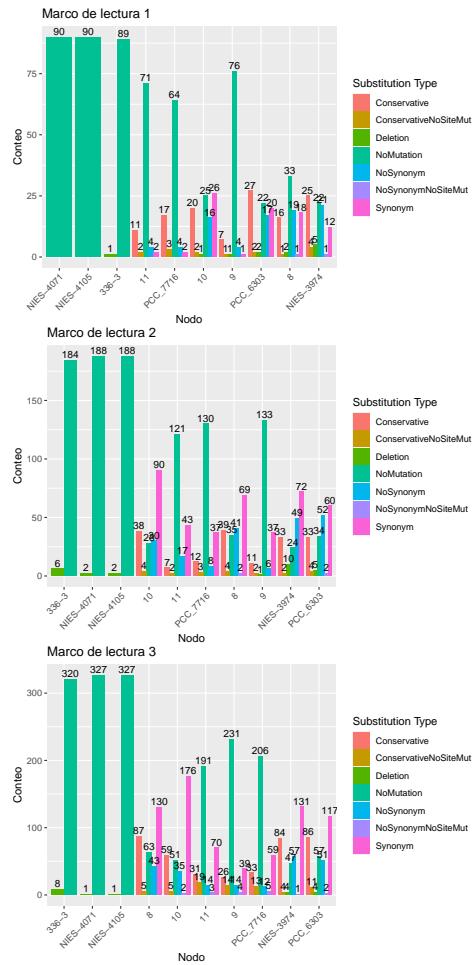


Figure 3.10: \*\*Abundancia del tipo de sustitución por cada nodo segun el marco de lectura.\*\*.

### 3.1.2.4 Análisis de sitios en los cuales su ancestro era TGGCGCCA

Para tratar de entender como es que los sitios HIP1 se pierden hicimos un análisis únicamente en las transiciones en las que el nodo ancestral tenía un sitio TGGCGCCA.

En la (Figura 3.11) mostramos 3 gráficos que indican la frecuencia del tipo de sustituciones que hubo para estos casos para cada nodo en cada uno de los marcos de lectura.

En la (Figura 3.12) mostramos 3 gráficos (uno por cada marco de lectura) que indican la frecuencia de las mutaciones en cada uno de los 8 nucleótidos del sitio HIP.

En la (Figura 3.13) mostramos 3 gráficos (uno por cada marco de lectura) que indican la frecuencia del tipo sustitución de bases.

### 3.1.2.5 Análisis de sitios en los cuales solo el nodo actual tiene TG-GCGCCA

Para tratar de entender como es que los sitios TGGCGCCA se ganan, hicimos un análisis únicamente en las transiciones en las que el nodo actual tenía un sitio TGGCGCCA.

En la Figura 3.14 mostramos 3 gráficos (uno por cada marco de lectura) que indican la frecuencia del tipo de sustituciones que hubo para estos casos para cada nodo en cada uno de los marcos de lectura.

## 3.2 Clado A18-40

El clado A18-40 fue de particular interés ya que entre las especies **Synechococcus\_sp\_A18-40** y **Synechococcus sp RS9915** hay un cambio abrupto del palíndromo con mayor OE. Más aun, los genomas de ambas especies son muy parecidos y ambas especies son cercanas según la filogenia (Figura 3.15).

Para saber que tan parecidos eran los genomas hicimos dos análisis de sintenia, uno de enfocado en los ortólogos (Figura 3.16) y otro enfocado en el genoma (Figura 3.17).

### 3.2.1 CGTTAACG

El palíndromo con la tasa OE más alta es **CGTTAACG** y lo tiene la especie **Synechococcus sp A18-40** (Tabla 3.1) con un conteo de 112 sitios para dicho palíndromo mientras que en las demás especies oscila entre 3 y 15 sitios.

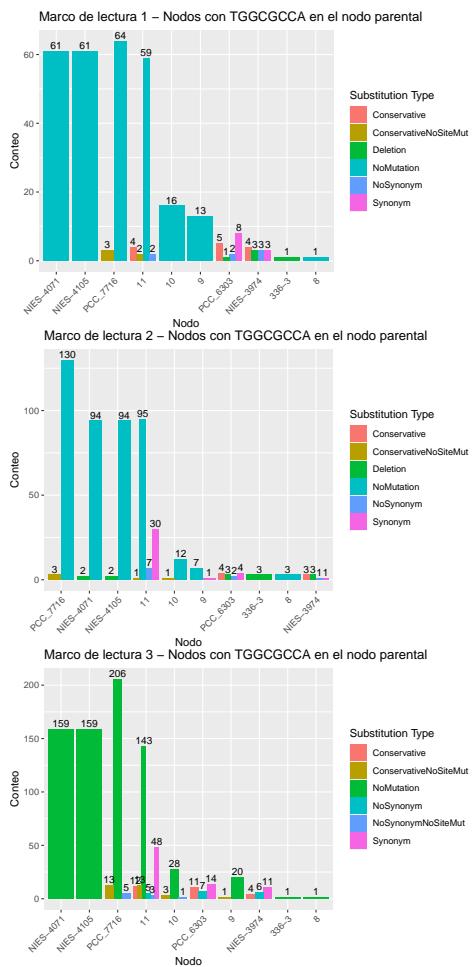


Figure 3.11: \*\*Abundancia del tipo de sustitución por cada nodo segun el marco de lectura. Unicamente para transiciones en los que el nodo ancestral era un sitio TGGCGCCA.\*\*

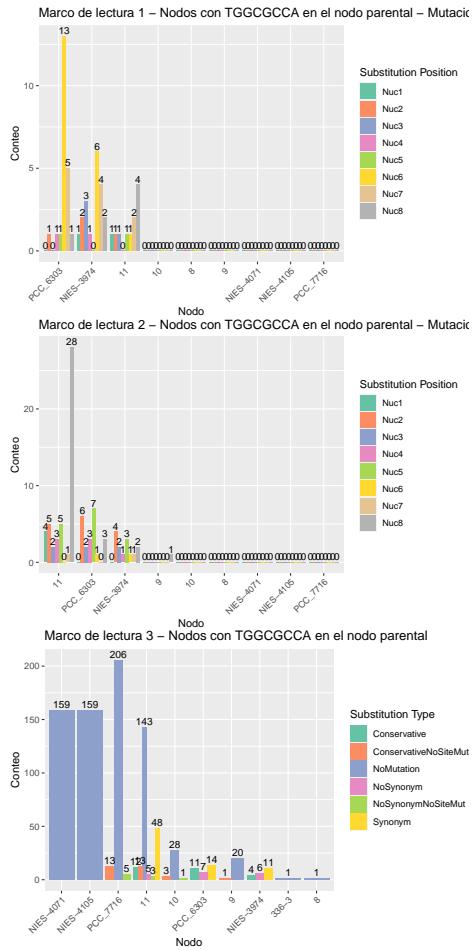


Figure 3.12: \*\*Frecuencia de las mutaciones de cada nucleótido del sitio TG-GCGCCA para cada nodo segun el marco de lectura.\*\*.

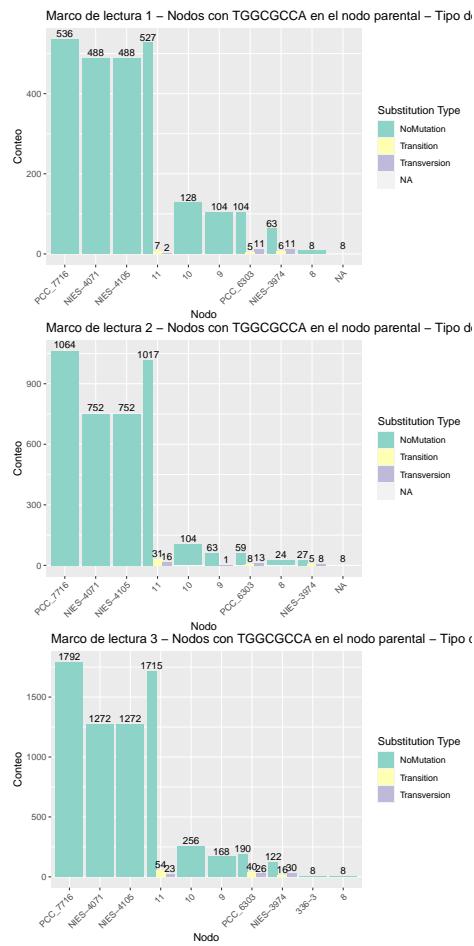


Figure 3.13: \*\*Frecuencia del tipo de sustituciones de base en los sitios TGCGCCA para cada marco de lectura\*\*.

Table 3.1: Conteo del palíndromo \*\*CGTTAACG\*\* en el clado \*\*A18-40\*\*

spp	palindrome	obs	markov3
Synechococcus_sp_A18-46_1	CGTTAACG	10	9.19
Synechococcus_sp_A18-40	CGTTAACG	112	9.71
Synechococcus_sp_WH_8103	CGTTAACG	9	9.14
Synechococcus_sp_RS9915	CGTTAACG	10	8.58
Synechococcus_sp_A15-28	CGTTAACG	3	7.58
Parasynechococcus_marenigrum_WH_8102	CGTTAACG	15	9.35
Synechococcus_sp_A15-24	CGTTAACG	11	7.99
Synechococcus_sp_BOUM118	CGTTAACG	11	8.24

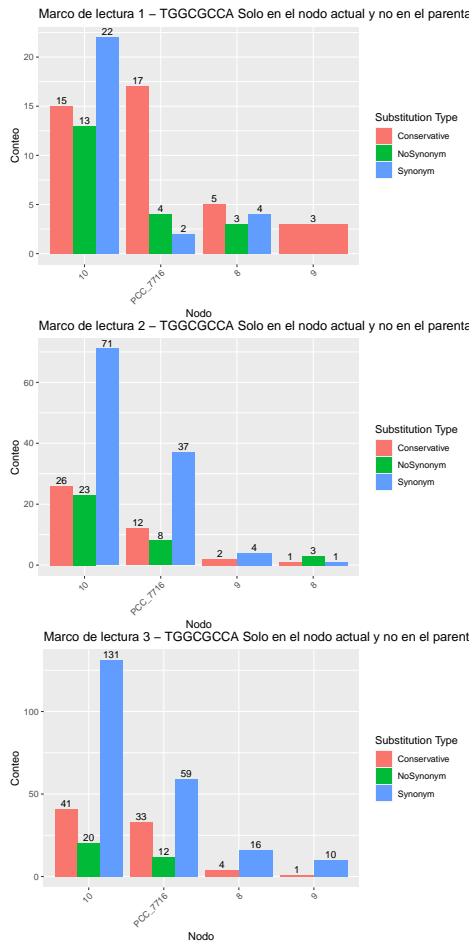


Figure 3.14: \*\*Abundancia del tipo de sustitución por cada nodo segun el marco de lectura. Unicamente para transiciones en los que el nodo actual era un sitio TGGCGCCA.\*\*.

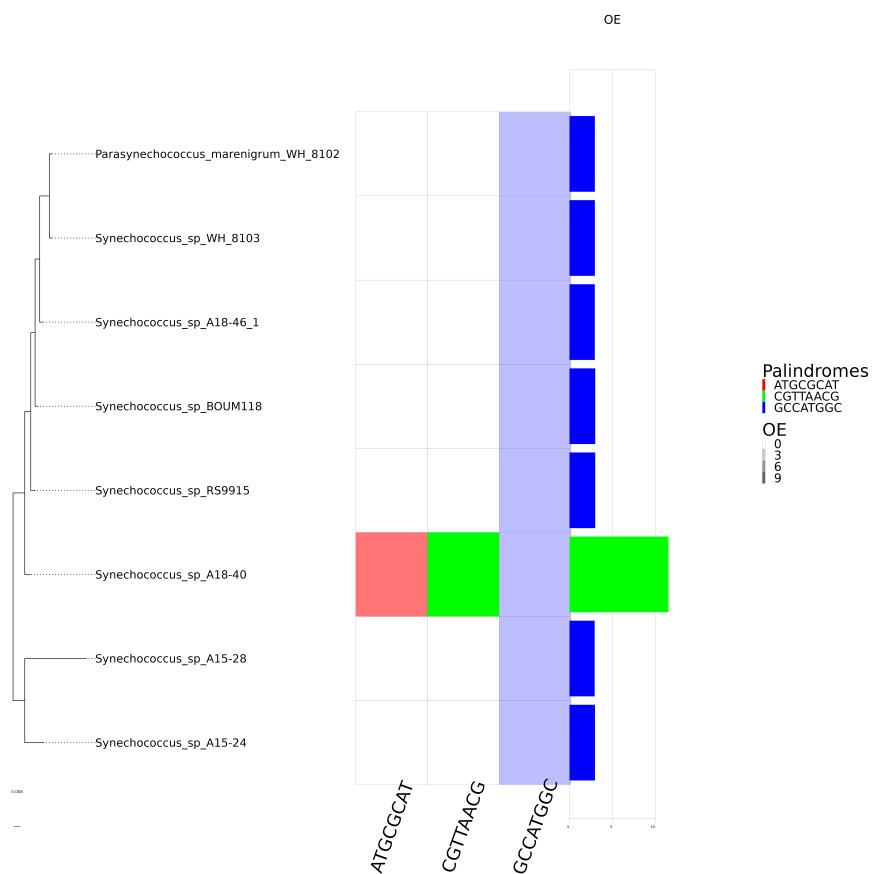


Figure 3.15: \*\*Filogenia anotada del clado A18-40.\*\* En esta imagen se muestra un cambio abrupto en la tasa OE de la especie *Synechococcus* sp A18-40.

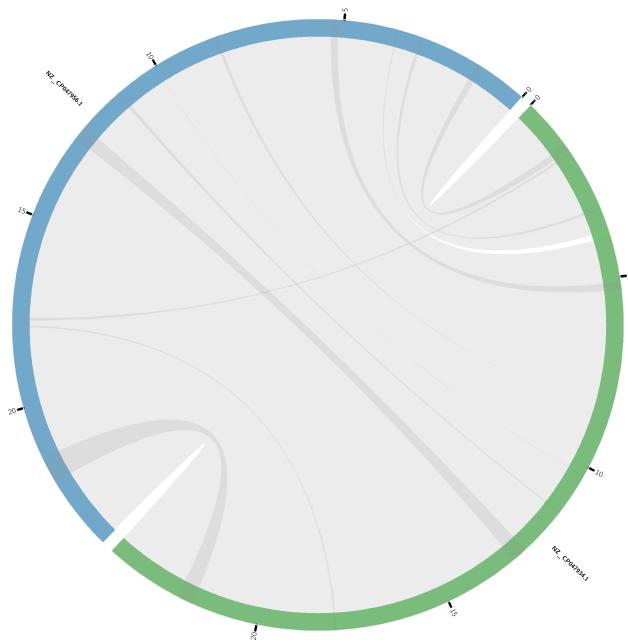


Figure 3.16: \*\*Sintenia de ortólogos entre las especies *Synechococcus* sp A18-40 y *Synechococcus* sp RS99150.\*\* En esta imagen se hizo un análisis de sintenia de ortólogos para ver que tan parecidos eran los genomas. En azul se muestra la especie *Synechococcus* sp A18-40 y en verde *Synechococcus* sp RS99150.

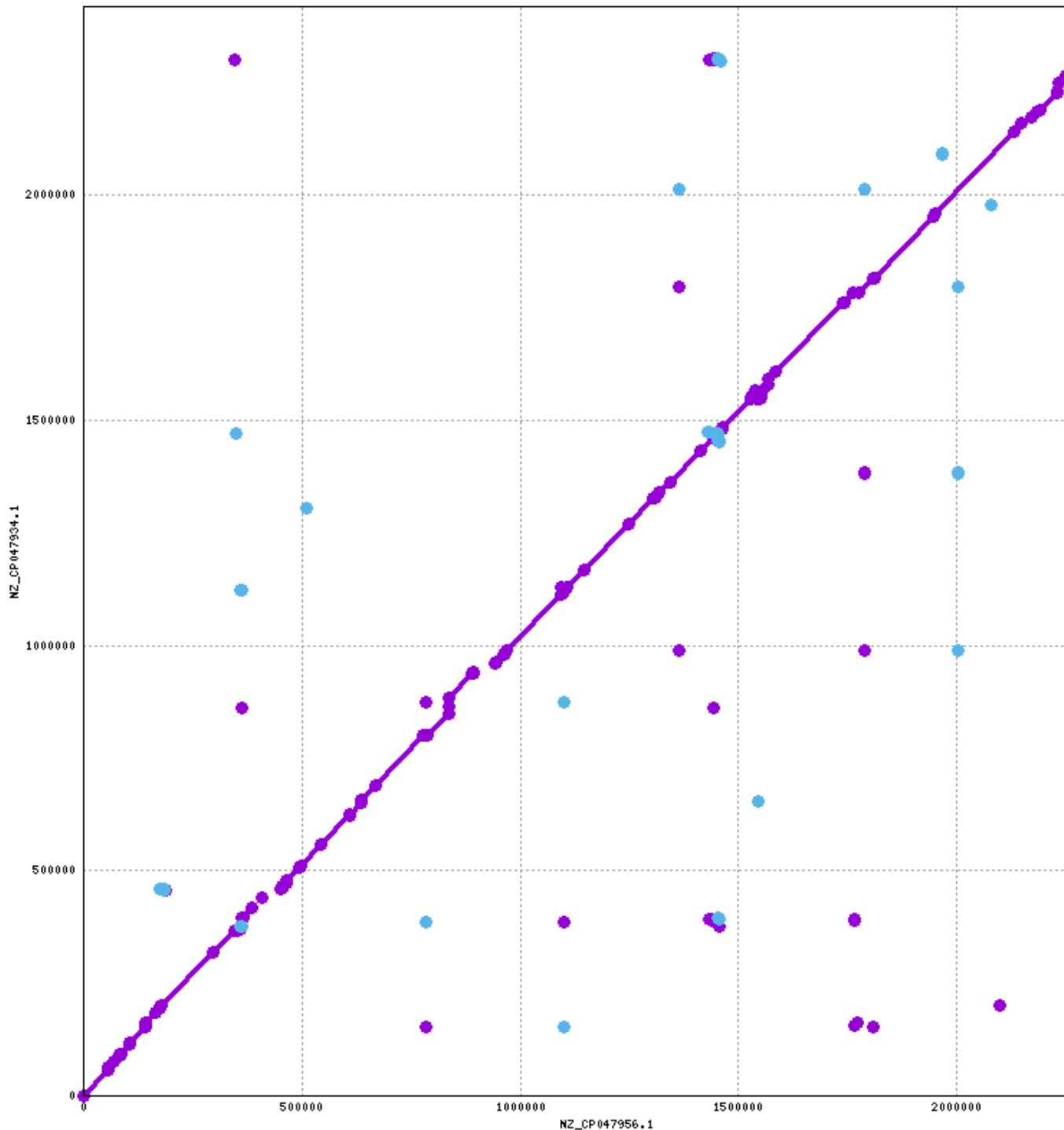


Figure 3.17: \*\*Sintenia de DNA entre especies *Synechococcus* sp A18–40 y *Synechococcus* sp RS99150.\*\* En esta imagen se hizo un análisis de sintenia de ortólogos para ver que tan parecidos eran los genomas. En el eje X se muestra la especie *Synechococcus* sp A18–40 y en el eje Y *Synechococcus* sp RS99150.

Table 3.2: \*\*Ubicación de los sitios CGTTACG.\*\* La tabla muestra las primeras 15 líneas de la tabla. La primera columna se muestra el número de sitio. La segunda columna muestra el intervalo en el que se encuentra el palíndromo. La tercera columna muestra a cuantos nucleótidos se encuentra el último sitio. La cuarta columna indica la diferencia entre la distancia del último palindromo y la distancia del siguiente.

SiteNumber	Interval	Dist2NextPal	DifBetDist
1	315323:315331	315323	-315323
2	522737:522745	207406	107917
3	594946:594954	72201	135205
4	683860:683868	88906	-16705
5	893774:893782	209906	-121000
6	894122:894130	340	209566
7	894470:894478	340	0
8	894818:894826	340	0
9	895166:895174	340	0
10	895514:895522	340	0
11	895862:895870	340	0
12	896210:896218	340	0
13	896558:896566	340	0
14	896906:896914	340	0
15	897254:897262	340	0

### 3.2.1.1 Ubicación de sitios CGTTAACG en los genomas del clado A18-40

Para tratar de entender la distribución del palíndromo en el genoma buscamos la ubicación de cada sitio y analizamos la distancia entre cada uno de ellos (Tabla 3.2). En dicho análisis pudimos observar que había 101 sitios que se encontraban entre repeticiones de 340 nucleótidos (columna 3 de la Tabla 3.2).

Para entender un poco más esta secuencia hicimos un blast el cual arrojó que dicha repetición de 340 nucleótidos era un motivo **SWM\_repeat** el cual se encuentra altamente repetido en una proteína de la superficie celular requerida para la movilidad (**QNJ16559.1**). Además, según la secuencia de aminoácidos, el palíndromo está segmentado en 2 partes. La primera mitad corresponde a **TTA ACG** en el primer y segundo codón y **CG** en el último codón del motivo **SWM\_repeat** (Figure 3.18).

Debido a que estos 101 sitios solo estaban presentes en la especie **Synechococcus sp A18-40** se concluyó que la tasa elevada de OE solo se debía a que tenía presente dicha proteína, ya que si se omitía del conteo, las tasas OE eran homogéneas en todo el clado.

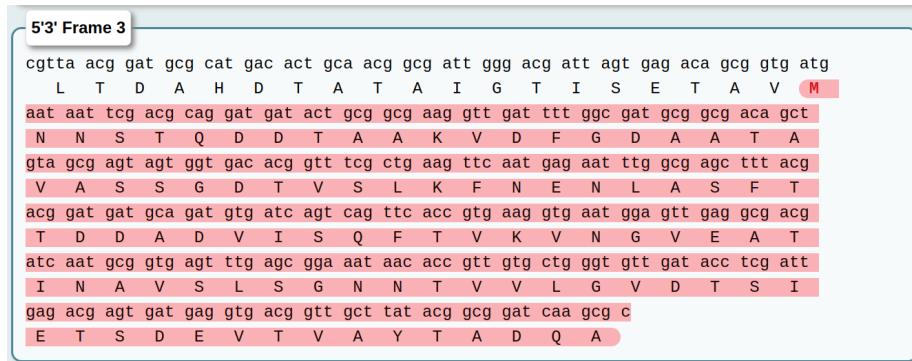


Figure 3.18: \*\*Traducción del motivo SWM\_repeat.\*\* En esta imagen se muestra la secuencia traducida del motivo SWM\_repeat la cual se encuentra en la especie *Synechococcus* sp A18-40.

Table 3.3: Conteo del palíndromo ATGCGCAT en el clado A18-40

spp	palindrome	obs	markov3
<i>Synechococcus</i> sp_A18-46_1	ATGCGCAT	31	21.14
<i>Synechococcus</i> sp_A18-40	ATGCGCAT	135	21.67
<i>Synechococcus</i> sp_WH_8103	ATGCGCAT	28	20.73
<i>Synechococcus</i> sp_RS9915	ATGCGCAT	31	20.50
<i>Synechococcus</i> sp_A15-28	ATGCGCAT	22	19.56
<i>Parasynechococcus</i> marenigrum_WH_8102	ATGCGCAT	28	20.60
<i>Synechococcus</i> sp_A15-24	ATGCGCAT	30	19.93
<i>Synechococcus</i> sp_BOUM118	ATGCGCAT	29	19.87

### 3.2.2 ATGCGCAT

El segundo palíndromo con la tasa OE mas alta es **ATGCGCAT** en la misma especie **Synechococcus** sp **A18-40** (Tabla 3.3) con un conteo de 135 sitios para dicho palíndromo mientras que en las demás especies oscila entre 22 y 31 sitios.

#### 3.2.2.1 Ubicación de sitios ATGCGCAT en los genomas del clado A18-40

Al analizar la distribución del palíndromo en el genoma pudimos observar que, al igual que CGTTAACG, había 101 sitios que se encontraban entre repeticiones de 340 nuclótidos (columna 5 de la Tabla 3.4)

Al hacer un blast el cual arrojó que dicha repetición de 340 nucleotidos era el mismo motivo **SWM\_repeat** de la misma proteína de la superficie celular requerida para la movilidad (**QNJ16559.1**). Al revisar la ubicación del

Table 3.4: \*\*Ubicación de los sitios ATGCGCAT.\*\* La primera columna se muestra el numero de sitio. La segunda columna muestra el intervalo en el que se encuentra el palíndromo. La tercera columna muestra a cuantos nucléotidos se encuentra el ultimo sitio. La cuarta columna indica la diferencia entre la distancia del ultimo palindromo y la distancia del siguiente.

SiteNumber	Interval	Dist2NextPal	DifBetDist
1	24777:24785	24777	-24777
2	91754:91762	66969	-42192
3	106816:106824	15054	51915
4	303729:303737	196905	-181851
5	415357:415365	111620	85285
6	420623:420631	5258	106362
7	540975:540983	120344	-115086
8	540987:540995	4	120340
9	571743:571751	30748	-30744
10	694673:694681	122922	-92174
11	708637:708645	13956	108966
12	760389:760397	51744	-37788
13	867192:867200	106795	-55051
14	893783:893791	26583	80212
15	894131:894139	340	26243

palíndromo pudimos notar que las mismas cantidades de los palíndromos **AT-GCGCAT** y **CGTTAACG** se debe a que estan a un nucleótido de distancia el uno del otro (Figure 3.19).

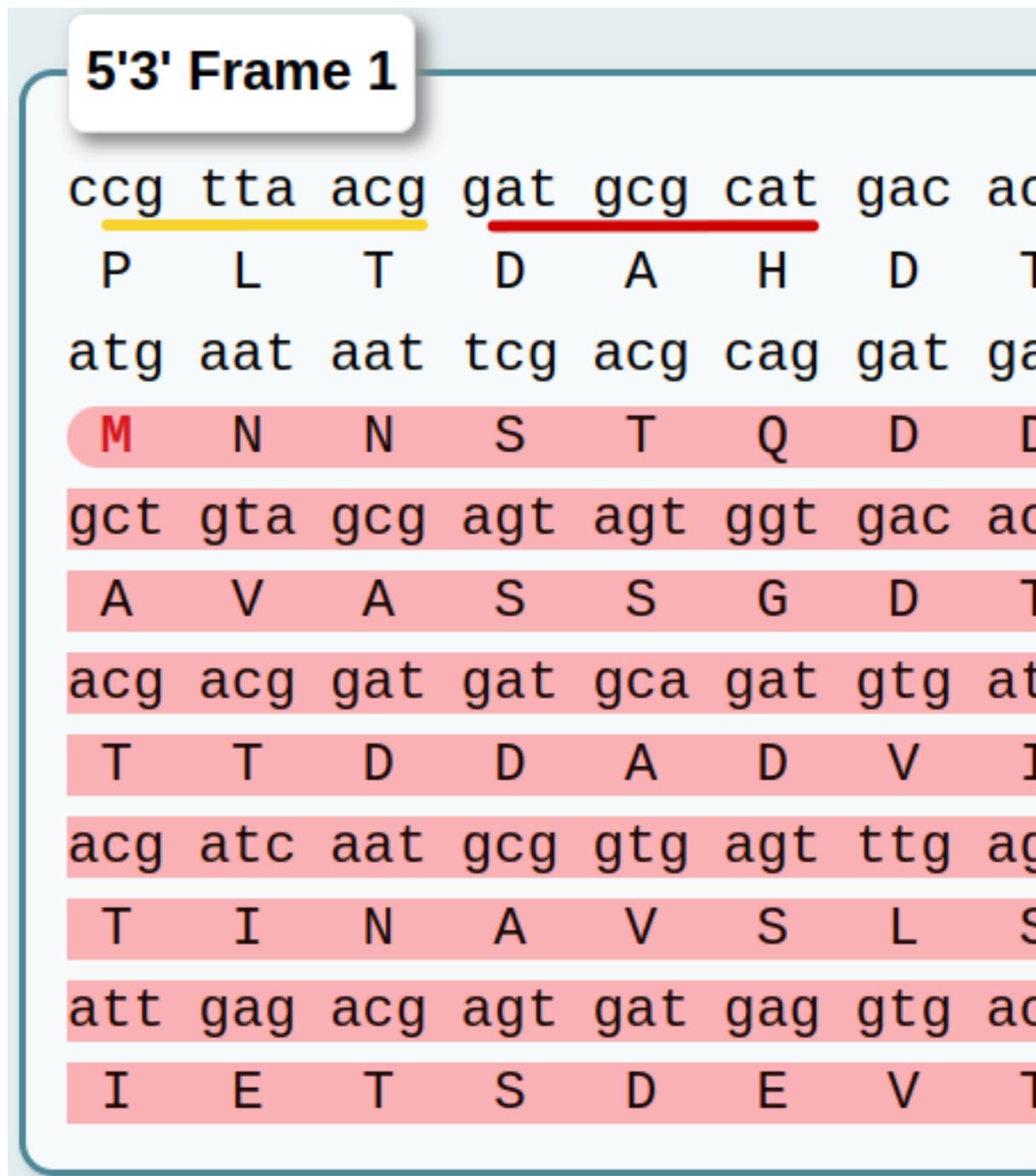


Figure 3.19: \*\*Traducción del motivo SWM\_repeat.\*\* En esta imagen se muestra la secuencia traducida del motivo SWM\_repeat. Subrayado en amarillo y rojo se señalan los palíndromos \*\*CGTTAACG\*\* y \*\*ATGCCGCAT\*\* respectivamente, los cuales se encuentran a un nucleotido de distancia.

# Bibliography

- Cabello-Yeves, P. J., Callieri, C., Picazo, A., Schallenberg, L., Huber, P., Roda-Garcia, J. J., Bartosiewicz, M., Belykh, O. I., Tikhonova, I. V., Torcello-Requena, A., et al. (2022). Elucidating the picocyanobacteria salinity divide through ecogenomics of new freshwater isolates. *BMC biology*, 20(1):1–24.
- Emms, D. M. and Kelly, S. (2019). Orthofinder: phylogenetic orthology inference for comparative genomics. *Genome biology*, 20:1–14.
- Lefort, V., Desper, R., and Gascuel, O. (2015). Fastme 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Molecular biology and evolution*, 32(10):2798–2800.