

# Evolución de secuencias palindromicas en genomas de cianobacterias

Eduardo Padilla Mendoza

2023-09-01



# 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



# 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 *CCGATCCC*, 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).

En la tabla (Tabla ??) se muestra el conjunto **sel256** el cual contiene 9 palíndromos significativos.

## 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.

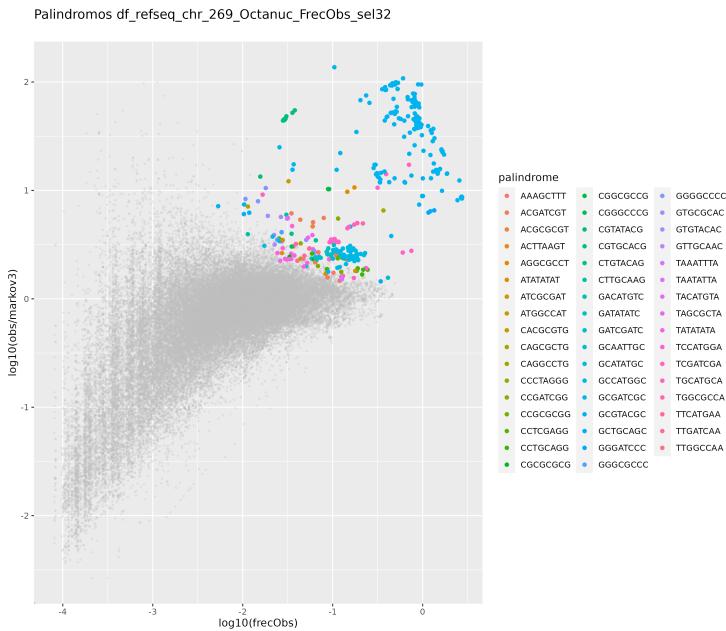


Figure 2.1: \*\*Enriquecimiento versus abundancia de palíndromos octámeros en el conjunto de genomas complete\_chr con un  $FDR \leq 1 \times 10^{-32}$ .\*\* Enrichment (\*\*O/E\*\*) in function of the frequency of the motif every 1000 nt (\*\*FreqObs\*\*). Each point represents a palindromic octamer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

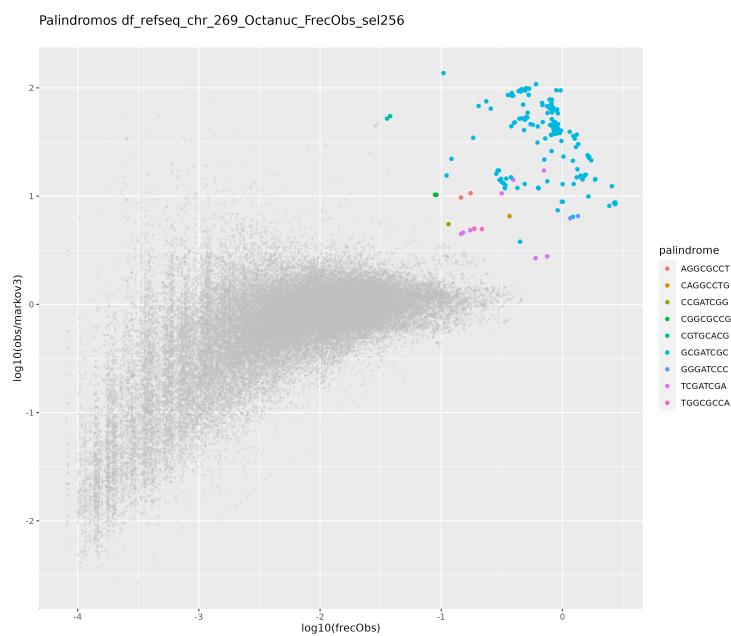


Figure 2.2: \*\*Enriquecimiento versus abundancia de palíndromos octámeros en el conjunto de genomas complete\_chr con un  $FDR \leq 1 \times 10^{-256}$ .\*\* Enriquecimiento (\*\*O/E\*\*) en función de la frecuencia del motivo cada 1000 nt (\*\*FrecObs\*\*). Cada punto representa un palíndromo octámero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

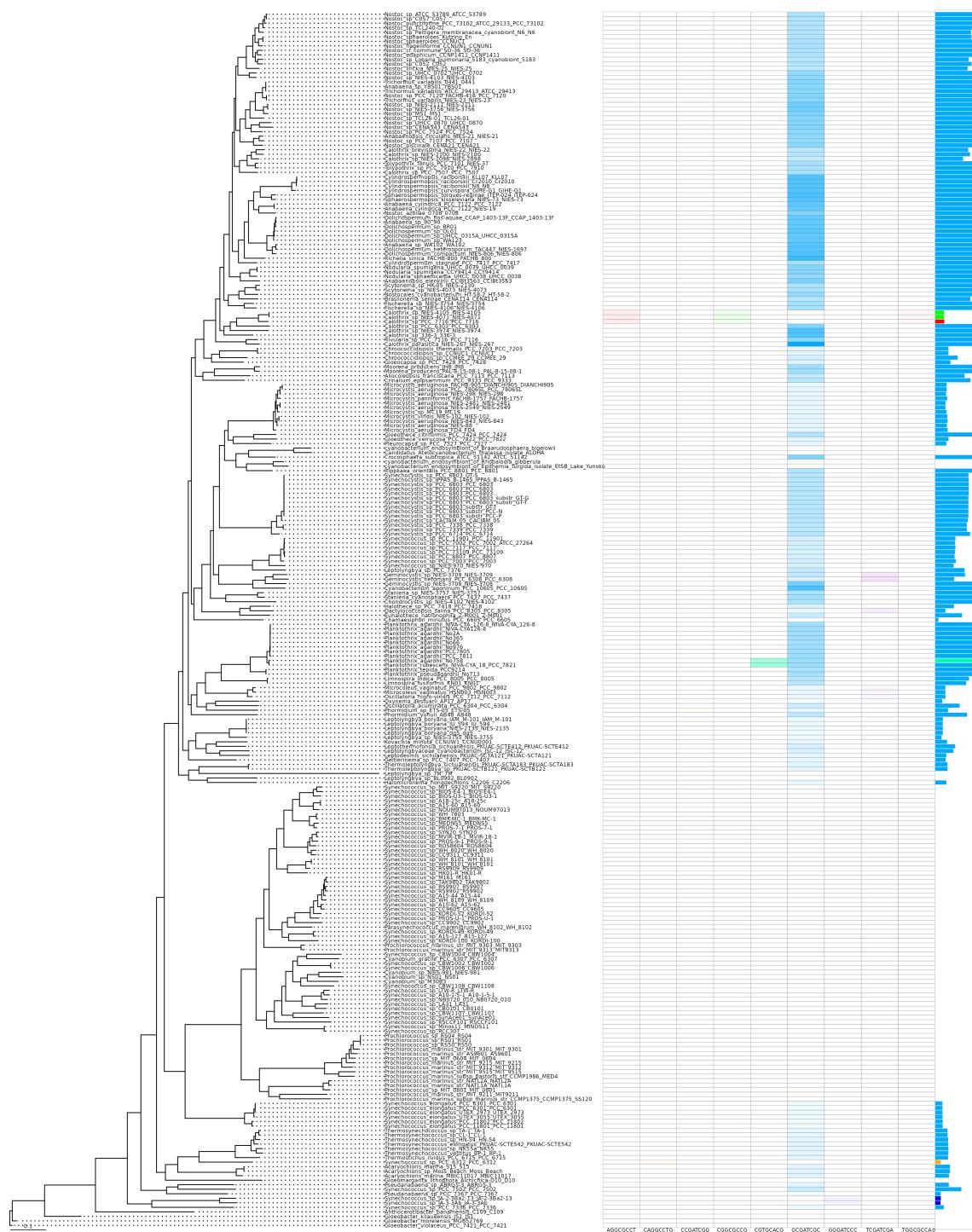


Figure 2.4: \*\*Filogenia del conjunto de genomas \*complete\_chr\* anotada de acuerdo a la tasa de observados sobre esperados (OE).\*\* La abundancia visualizada en esta filogenia es de acuerdo al conjunto \*\*sel256\*\*, es decir conteos con un  $FDR \leq 1 \times 10^{-256}$ . La filogenia muestra 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

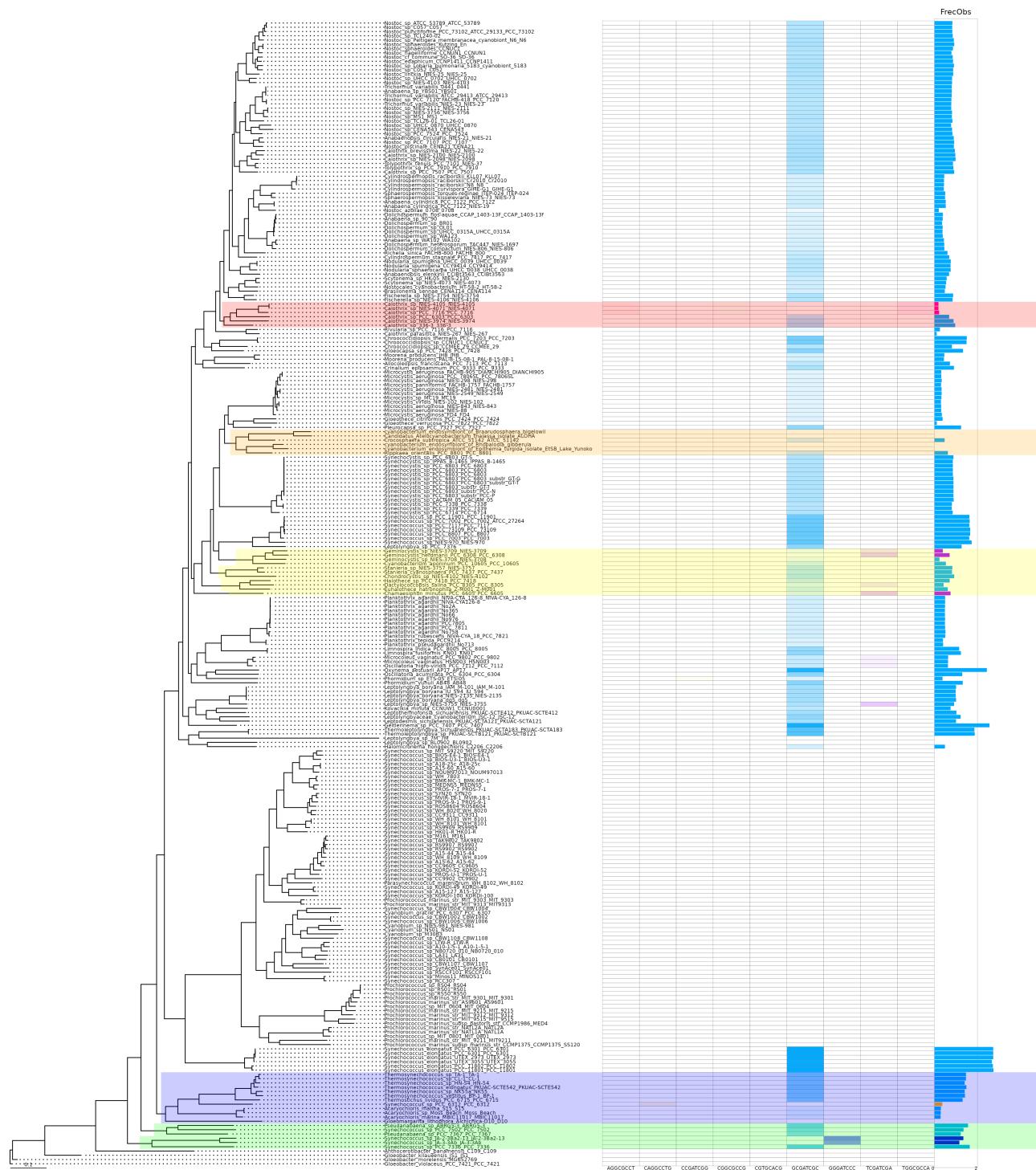


Figure 2.5: \*\*Casos de interés.\*\* En la figura se muestran remarcados los casos interesantes: \*\*clado calothrix\*\* (rojo), \*\*clado cyanobacterium\*\* (naranja), \*\*clado geminocystis\*\* (amarillo), \*\*clado thermosynechococcus\*\* (azul), \*\*clado pseudoanabaena\*\* (verde).

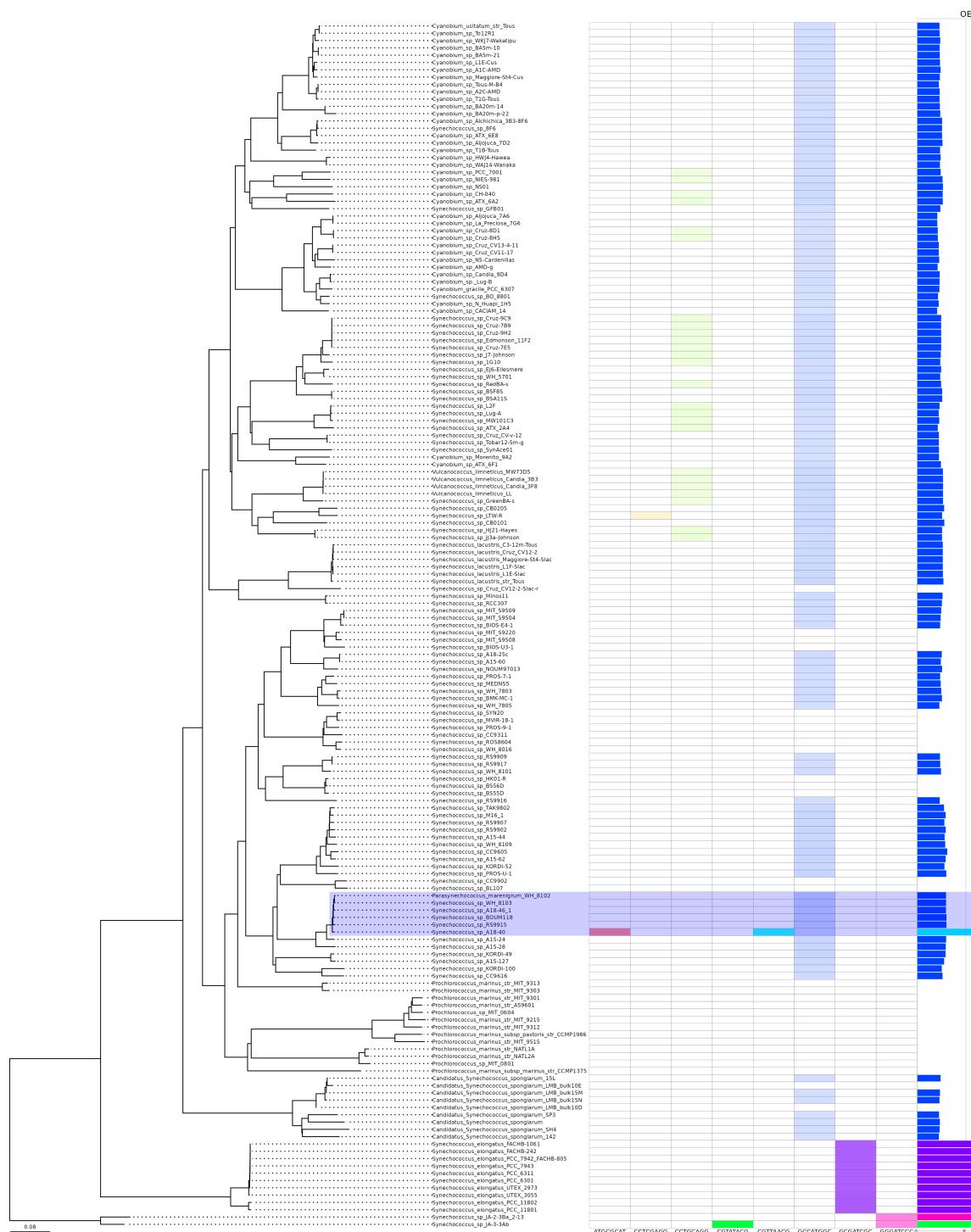


Figure 2.6: \*\*Casos de interés.\*\* En la figura se muestra remarcado el \*\*clado A18-40\*\* (azul).

## 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")
```



# 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.