Challenge1

Alvarez C, Barrera M, Castro J, Ibañez C, Vasquez I, Villa J.

2/2/2021

CHALLENGE 1

El presente documento busca dar solución al siguiente problema propuesto.

- Ingrese a https://bit.ly/30LIowU
- (10 puntos) Tome la secuencia FASTA del gen *ADGRL3* y léala en R
- (15 puntos) Determine la longitud de la secuencia, además del número de bases A, C, G y T
- (5 puntos) Podemos afirmar que las bases siguen una distribución uniforme?
- (30 puntos) Calcule $P(s+1=j|s=i)$, donde s es el sitio de ocurrencia, e $i,j=\{A,C,G,T\}$

Lectura de datos

Para leer la secuencia de datos, previamente extraímos la secuencia de la dirección URL y la depositamos en un bloc de notas y haciendo uso de las funciones ***scan()*** y ***strsplit()*** hicimos la lectura del .txt

```
## Lectura de datos
datos <- scan("gen.txt", what = character())
splitdatos <- strsplit(datos, split = character())
```

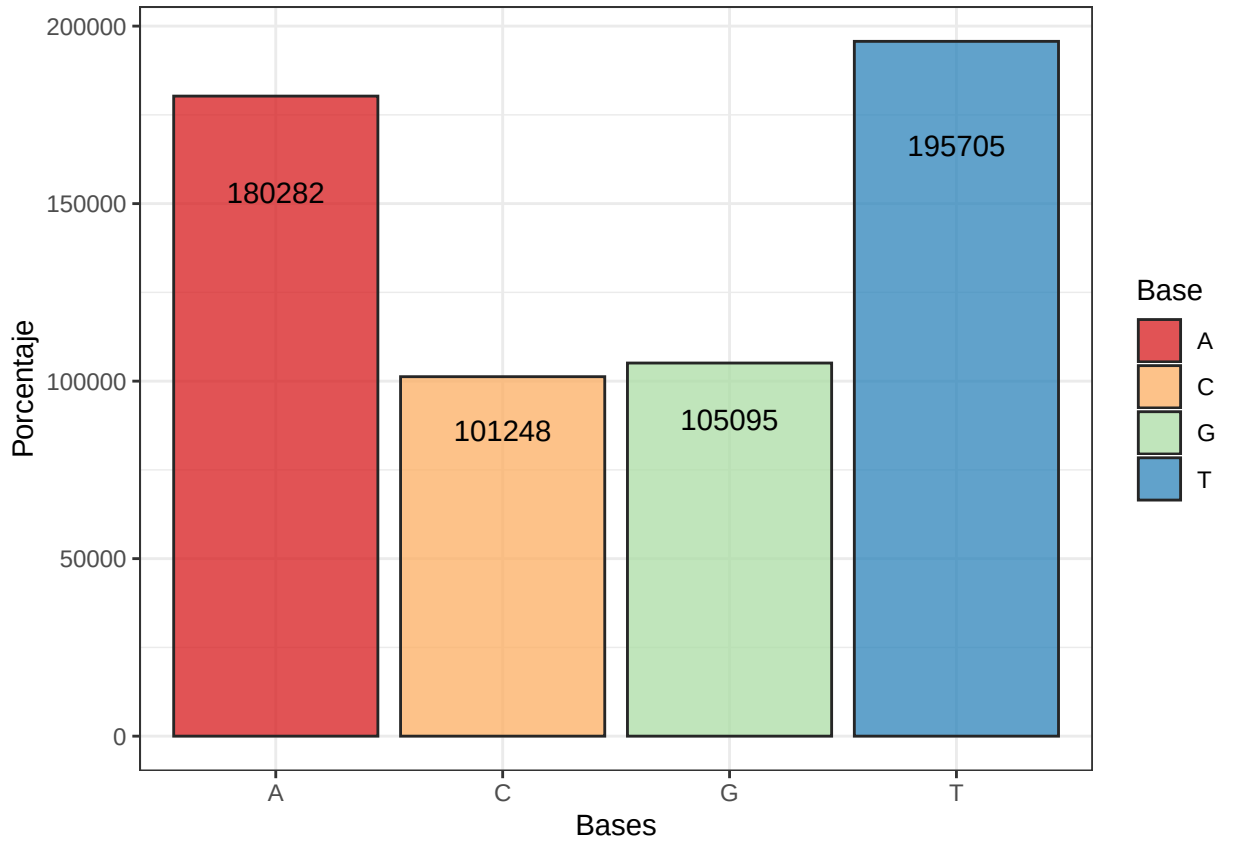
Longitud de datos y frecuencia de las bases

El número total de bases encontradas en el gen fue de 5.8233×10^5. La frecuencia de las bases presentes en la muestra del gen se muestra en la siguiente tabla

```
##   Bases Frecuencias
## 1     A      180282
## 2     T      195705
## 3     G      105095
## 4     C      101248
```

¿Sigue una distribución uniforme?

Si bien en casos de verificación de que un conjunto de datos sigue una distribución especifica lo más ortodoxo sea una prueba de bondad de ajuste, un buen primer indicativo puede ser un gráfico de los datos.



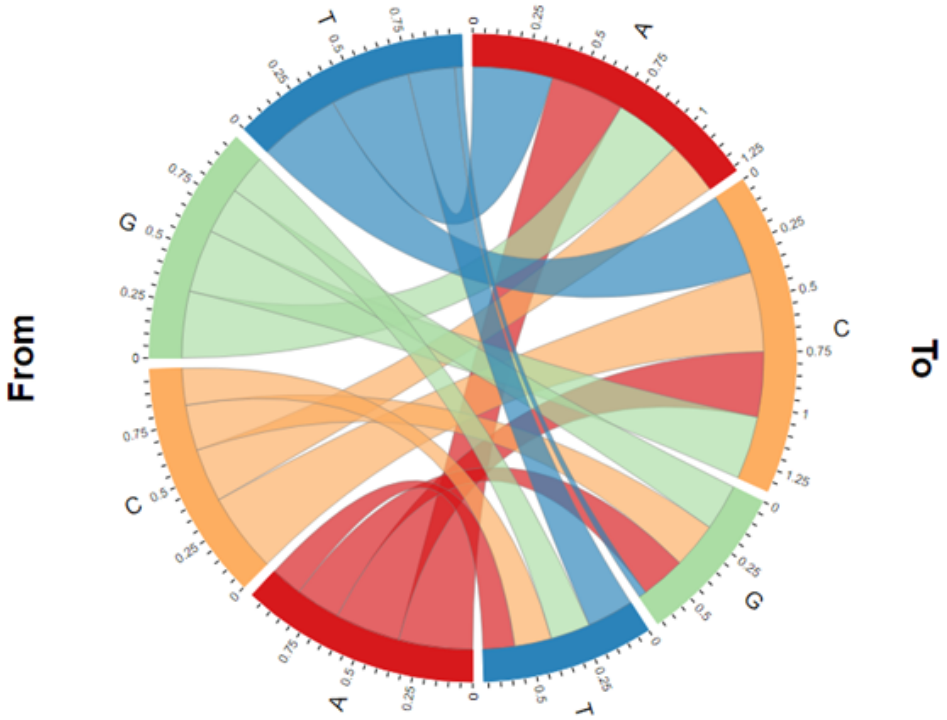
Según el gráfico presentado no podemos afirmar que los datos de la muestra del gen sigan una distribución uniforme.

Calculo de la probabilidad condicional

Las probabilidades de que dada una base i en la siguiente posición se encuentre una base j donde i,j = A,C,T,G se muestra en la siguente tabla, siendo i las filas y j las columnas.

```
##           A         T          G         C
## A 0.3428518 0.3024539 0.20690918 0.1477851
## T 0.2484811 0.3689839 0.21215605 0.1703789
## G 0.3085780 0.2917075 0.21532899 0.1843856
## C 0.3695122 0.3783717 0.03598131 0.2161348
```

También podemos presentar esta información de manera más gráfica con en el siguiente diagrama de Chord.



La fundamentación teorica de dicha tabla se basa escencialmente en la aplicación del teorema de Bayes, de la siguiente manera:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(s+1 = j \,|s = i) \; ; i,j = \{A, C, G, T\}$$

$$P(A) \rightarrow P(s+1 = j)$$

$$P(B) \rightarrow P(s = i)$$

$$P(s+1 = j \,|s = i) = \frac{P(s+1 = j \cap s = i)}{P(s = i)}$$

$$= \frac{\frac{\#\, Parejas\; i,j}{\#Parejas}}{\frac{\#Parejas\ donde\ s = i}{\#Parejas}}$$

$$= \frac{\#\, Parejas\; i,j}{\#Parejas\ donde\ s = i}$$