

```

---
title: "Descubrimiento de un motivo a partir del alineamiento de secuencias
de ADN "
output: html_notebook
---

```

Codigo para descubrir la secuencia motivo a partir de un alineamiento  
Se crea la matrix de frecuencia, la de probabilidades (PWM), y se crea  
el logo de secuencia a partir del cual creamos la expresion regular  
para buscar el posible motivo en prosite (scanprosite)

```

# Load libraries
```{r}
require(ggplot2)
require(ggseqlogo)
```

```

```

# Ejemplo1: 10x9 Secuencias de nucleótidos iniciales en formato de cadenas
```{r}

```

```

secuencias = c (
  "GAGGTAAAT",
  "GCCGTAAGT",
  "GAGGTTGGT",
  "GCCGTCAGT",
  "GAGGTCATT",
  "GACGTA CTT",
  "GTGGTAACT",
  "GACGTATAT",
  "GGCGTGAGT",
  "GAGGTAAGT")
N = length (secuencias)      # Numero de secuencias (Filas)
M = nchar (secuencias [1]) # Tamaño de cada secuencia (Columnas)
message ("Vector de cadenas de secuencias:")
message (paste ("Dimension NxM: ", N, M))
print (secuencias)
```

```

```

# Convertimos las cadenas a matriz de caracteres
```{r}
seqMatrix = matrix (unlist (strsplit (secuencias, split="")),nrow=N,byrow =
T)
message ("Secuencias como Matrix de caracteres: ")
print (seqMatrix)
```

```

```

# Calculamos la matrix de frecuencias (PCM)
```{r}
freqMatrix = NULL
bases = c("A","C","G","T")
# aminos = ....
for (j in 1:M) {
  conteos = c()
  for (B in bases) {
    nB = length(which (seqMatrix[,j]==B)) # Conteo de la base B en la
posicion j
    conteos = c(conteos, nB)
  }
}

```

```

    }
    freqMatrix = cbind (freqMatrix, conteos) # Adicionamos los conteos de la
posicion j
}
# Asignamos nombres a las filas y columnas
rownames (freqMatrix) = bases      # Para nucleotidos
#rownames (freqMatrix) = aminos    # Para aminoacidos

colnames (freqMatrix) = paste0 ("p", 1:M)
freqMatrix = freqMatrix # TRABAJAR SIN PSEUDOCONEOS, ya que vamos solo a
obtener el motivo
message ("Matrix de frecuencias:")
print (freqMatrix)
```

# Calculamos la matrix de posición de probabilidades o matrix ponderada de
posiciones (PWM)
```{r}
probMatrix = freqMatrix/(N) # Dividimos entre N+4 porque el pseudoconteo
adiciona 1 por cada base
PWM = probMatrix
message ("Matrix de probabilidades (PWM): ")
print (round (probMatrix, 2))
```

# Creamos el logo de la matriz de probabilidades
```{r}
ggseqlogo (PWM)
```

# Creamos manualmente la expresion regular para ingresarla en scanprosite
```{r}
print ("G-x-[CG]-G-T-x-x-x-T")
```

```