

# Proyecto de Informática para Biología: Búsqueda de regiones reguladoras

Prof. Luis Garreta  
Pontificia Universidad Javeriana

May 12, 2015

## 1 Descripción

El objetivo del proyecto es descubrir posibles regiones reguladoras en secuencias no-codificantes de un metagenoma comparándolo con genomas de bacterias. Para esto se propone un enfoque evolutivo que busque estas regiones a través de comparaciones con secuencias de genomas de bacterias y los posibles motivos que se encuentren se verifiquen con una herramienta o base de datos pública de motivos. El siguiente *pipeline* muestra el flujo del enfoque propuesto:

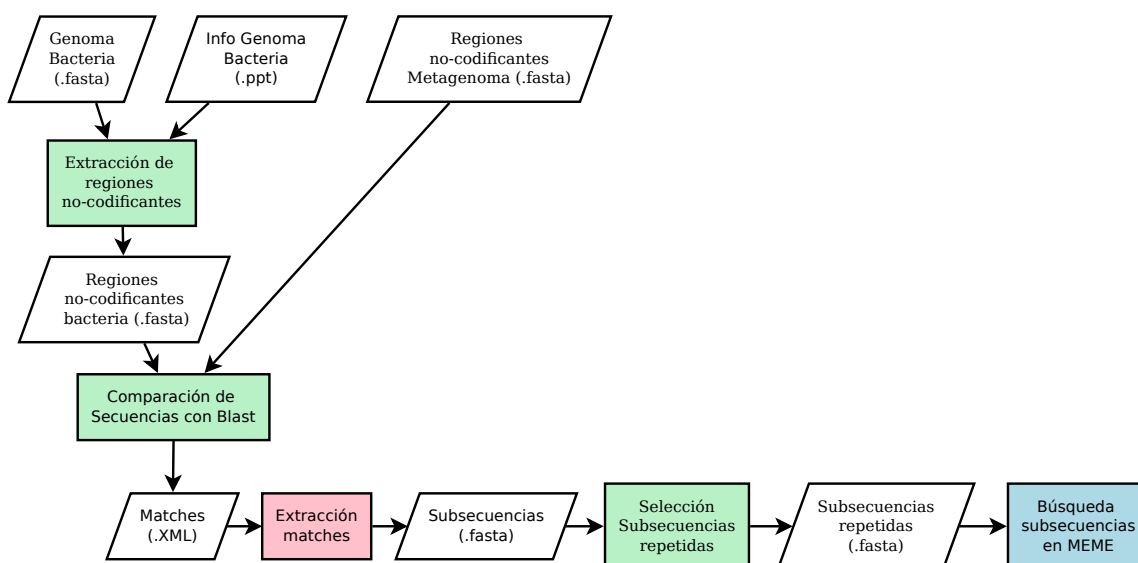


Figure 1: Pipeline propuesto búsqueda regiones reguladoras

Los rombos significan datos de entrada o salida, mientras que los cuadros significan procesos sobre esos datos. Los cuadros verdes son los módulos que usted debe desarrollar, mientras que el cuadro rojo es un módulo externo listo para usar, y el cuadro azul es un proceso que puede realizarlo de forma manual o creando un módulo en python.

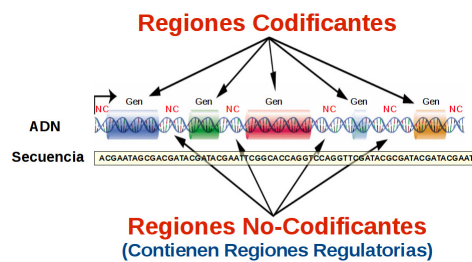
A continuación la descripción de cada elemento:

1. **Genoma Bacteria (.fasta):** Archivo con la secuencia de ADN del genoma de una bacteria en formato fasta.
2. **Info Genoma Bacteria (.ppt):** Archivo con la información de las coordenadas de los genes de la bacteria.
3. **Regiones no-codificantes Metagenoma (.fasta):** Archivo con la secuencia de ADN de un metagenoma en formato fasta.
4. **Extracción de regiones no-codificantes:** Módulo que recibe como entrada dos archivos y genera uno nuevo. Los archivos de entrada son el genoma de la bacteria y la información de las posiciones de los genes. Este módulo obtiene las coordenadas de los genes del archivo .ppt y extrae las subsecuencias de las regiones no-codificantes y las escribe a un nuevo archivo.

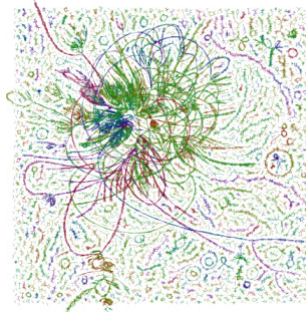
5. **Regiones no-codificantes bacteria (.fasta):** archivo resultante del proceso anterior con las regiones no-codificantes de la bacteria en formato fasta.
6. **Comparación de secuencias con Blast:** Módulo que recibe como entrada dos archivos y genera uno nuevo. Recibe como entrada los archivos de secuencias no-codificantes de la bacteria y el del metagenoma y ejecuta internamente Blast con la opción de generar un archivo en formato XML.
7. **Matches (.XML):** archivo con los resultados del proceso anterior (blast) en formato XML
8. **Extraccion matches:** Módulo que procesa el archivo XML anterior y extrae las subsecuencias que emparejaron (match) y las escribe en un nuevo archivo en formato .fasta
9. **Selección de subsecuencias repetidas:** módulo que recibe como entrada las subsecuencias y genera un nuevo archivo en formato fasta con las que se repiten más de una vez.
10. **Subsecuencias repetidas (.fasta):** archivo resultante del proceso anterior
11. **Búsqueda subsecuencias en MEME:** Proceso manual (o un módulo) que busca las subsecuencias que más se repiten en la herramienta de búsqueda de motivos MEME.

## 2 Antecedentes

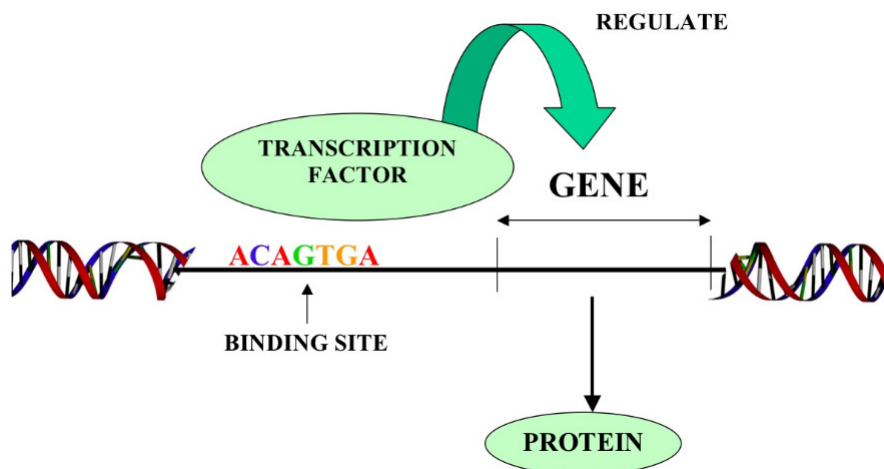
### 2.1 Regiones no-codificantes



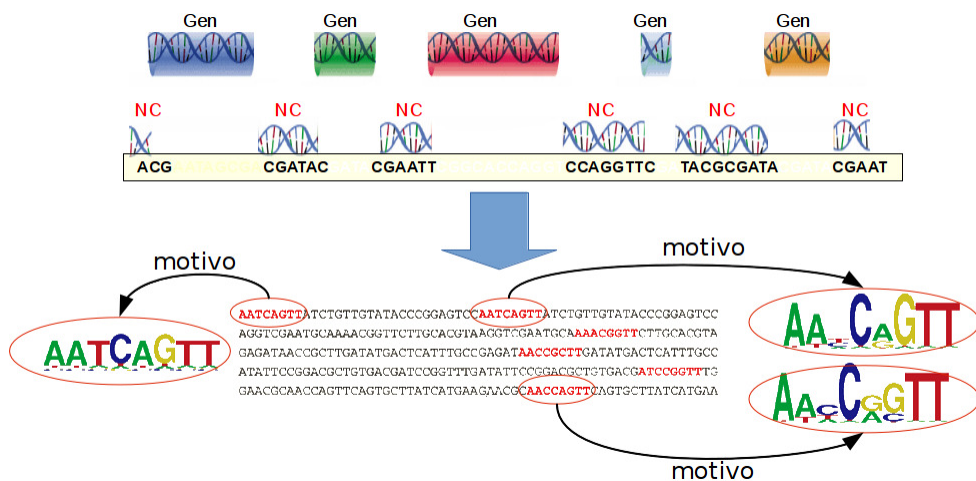
### 2.2 Metagenomas



## 2.3 Factores de Transcripción y Motivos en la Vida



## 2.4 Motivos en Regiones no-codificantes de Metagenomas



## 2.5 Búsqueda de Motivos Experimentales

