Análisis de redes de regulación transcripcional con Maximización de la Expectación

Santiago Toledano Gómez

Resumen—El objetivo de este Proyecto es mejorar el análisis de las redes de regulación transcripcional que se lleva a cabo con la plataforma CGB, mediante un algoritmo de Maximización de la expectación. A partir de factores de transcripción, CGB identifica genes regulados y el objetivo es mejorar esta identificación. Los resultados han sido mixtos, aunque se detectaron errores, también se obtuvieron resultados prometedores con motivos ajustados a la especie que se analiza, incrementando así la cantidad de genes regulados identificados. Es posible que tras mejorar el algoritmo se pueda utilizar para obtener más información sobre la regulación de los genes con consistencia.

Palabras clave— Redes de regulación transcripcional, Maximización de la expectación (EM), Factor de transcripción (TF), Regulación génica, Motivos, Genómica comparativa, Bioinformática.

Abstract— The objective of this project is to improve the analysis of transcriptional regulatory networks carried out with the CGB platform, using an Expectation Maximization (EM) algorithm. Based on transcription factors, CGB identifies regulated genes, and the goal is to improve this identification. The results have been mixed; although errors were detected, there were also promising outcomes with motifs adjusted to the analyzed species, increasing the number of identified regulated genes. It is possible that after improving the algorithm, it can be used consistently to obtain more information on gene regulation.

Index Terms— Transcriptional regulatory networks, Expectation Maximization (EM), Transcription factor (TF), Gene regulation, Motifs, Comparative genomics, Bioinformatics.

1 Introducción

La bioinformática consiste en el uso de la computación para procesar e interpretar los datos generados por la biología. El procesamiento de secuencias genéticas nos permite entender cómo se lleva a cabo el procesamiento de la información en el ADN. Uno de los ámbitos en investigación es el uso de métodos de genómica comparativa para analizar las redes de regulación transcripcional.

El objetivo principal de este proyecto es mejorar el análisis de las redes de regulación transcripcional mediante el algoritmo de Maximización de la Expectación.

* *E-mail de contacto: stoledano97@gmail.com*
* *Mención realizada: Computación*
* *Trabajo tutorizado por: Ivan Erill*
* *Curso: 2023/24*

2 PLANIFICACIÓN

2.1 Objetivos

Este proyecto pretende llegar a la mejora del análisis de las redes transcripcionales a través del cumplimiento de los siguientes objetivos:

1. Comprender el concepto de red de regulación transcripcional, que elementos lo componen y que función cumplen.
2. Entender cómo se procesan los patrones o motivos del ADN y cómo se calcula la información que contienen, utilizando herramientas como la matriz PSWM (Position-Specific Weight Matrix).
3. Comprender la genética comparativa entendiendo cómo se pueden conservar patrones funcionales en la evolución de la especie.
4. Aprender a usar la plataforma CGB e interpretar los resultados.
5. Entender y desarrollar un algoritmo de maximización de la expectación.
6. Utilizar Maximización de la expectación juntamente con la plataforma CGB para mejorar el análisis de las redes de regulación, mejorando los datos obtenidos de secuencias de enganche, genes regulados y probabilidades de regulación.
7. Analizar los resultados y evaluar si el análisis con el algoritmo de maximización de la expectación ha mejorado respecto al proceso anterior de análisis.

2.1 Metodología

Se trabajará preparando objetivos mensuales y haciendo pequeños ajustes semanales, se definirá una reunión semanalmente para aclarar las dudas y se usará un enfoque de trabajo iterativo repitiendo los siguientes pasos:

1. Entender el contexto biológico.
2. Obtener y entender los datos de las bases de datos biológicas como las bases de datos de ncbi.
3. Tratamiento y preparación de los datos, si es necesario.
4. Desarrollo del modelo.
5. Evaluación del modelo.
6. Análisis de los resultados.

2.1 Planificación

El proyecto se organizará en diferentes periodos, cada periodo durará aproximadamente un mes buscando realizar una serie de tareas específicas por cada periodo.

Semana 1-4:

- Definir la metodología, planificación, objetivos del proyecto y añadir el contexto necesario.

Semana 5-9:

- Utilizar y aprender el funcionamiento de la plataforma CGB.
- Preparar una primera instancia del modelo de maximización de la expectación.
- Mejorar el contexto realizado previamente.

Semana 10-14:

- Aplicar el algoritmo de maximización de la expectación a el marco probabilístico de la plataforma CGB.
- Iniciar el análisis de los resultados obtenidos.
- Realizar mejoras en el algoritmo de maximización de la expectación.

Semana 15-17:

- Realizar un análisis de los resultados obtenidos.
- Añadir las últimas mejoras al proyecto.
- Elaborar las conclusiones del proyecto.

3 CONTEXTO

3.1 Redes de regulación transcripcional

El ADN contiene la información fundamental para el funcionamiento y la creación de la vida. Se compone de una estructura de 2 hélices con una serie de nucleótidos o bases que codifican la información. Estos pueden tener 4 valores (A, T, G, C). Un gen es un conjunto de bases que suponen una unidad funcional. La mayoría de los genes contienen información para elaborar una proteína específica.

La transcripción del ADN transforma la información del ADN en ARN mensajero. Hay patrones promotores que marcan el inicio del gen y patrones terminadores que indican su finalización. Estos genes se pueden regular.

La regulación es llevada a cabo por proteínas llamadas factores de transcripción (TF), que se unen al ADN en regiones específicas. Los factores de transcripción activadores promueven la transcripción y los factores de transcripción represores inhiben la transcripción.

En un determinado momento, solo un subconjunto de todos los genes puede estar activo. El sistema de interacciones y relaciones entre los genes y las proteínas que controlan la transcripción es la red de regulación transcripcional.

3.2 Motivos

Los motivos de secuencia son patrones cortos recurrentes en el ADN que se presume tienen una función biológica. Un promotor sería un ejemplo de un motivo. [1]

Las proteínas de unión al ADN suelen aceptar un grado de tolerancia en su motivo, con algunas secuencias teniendo un nivel de actividad mejor que otras. La matriz de frecuencia de posición (PFM, Position Frequency Matrix) tiene en cuenta cuántas veces ocurre una base en una posición y proporciona una descripción más precisa que solo su secuencia con la base más común.

Por ejemplo, con una serie de 100 secuencias, se cuenta cada base en cada posición. Formando la siguiente matriz PFM:

	1	2	3	4	5	6	7	8
A	53	83	69	39	11	42	7	07
C	32	05	04	42	43	05	06	05
G	9	7	25	5	9	13	15	83
T	6	5	2	13	37	39	9	05

Las columnas representan las posiciones y las filas cada base del motivo.

A partir de esta matriz se puede obtener una matriz PSWM (position-specific weight matrix) con las probabilidades que una base se encuentre en determinada posición, donde cada columna es decir cada posición suma probabilidad de 1.

PSWM

	1	2	3	4	5	6	7	8
A	0.53	0.83	0.69	0.39	0.11	0.42	0.7	0.07
C	0.32	0.05	0.04	0.42	0.43	0.05	0.06	0.05
G	0.09	0.07	0.25	0.05	0.09	0.13	0.15	0.83
T	0.06	0.05	0.02	0.13	0.37	0.39	0.09	0.05

También se usa una representación del motivo gráfica llamada Logo, que muestra el motivo según la cantidad de información que contiene.

3.3 Genómica comparativa y plataforma CGB

La genómica comparativa se basa en comparar el genoma de diferentes especies con características comunes para identificar en que partes del genoma pueden estar estas características.

Un gen ortólogo es un gen presente en diferentes especies que evolucionó a partir de un gen ancestral común, que típicamente conserva funciones similares. Podemos utilizar la genética comparativa para encontrar esos genes, o sus promotores.

Solo los sitios de unión de los factores de transcripción que son funcionales deberían preservarse a lo largo de intervalos evolutivos sustanciales. Por lo tanto, la identificación de un sitio de unión en la región del promotor de dos o más operones ortólogos debería fortalecer nuestra confianza en su predicción como un elemento funcional. [2]

CGB es una plataforma para la genómica comparativa de las redes de regulación bacteriana. Puede utilizar y combinar datos experimentales y emplear un marco probabilístico formal para la integración e interpretación de los resultados de análisis. [2]

El objetivo principal de este proyecto es mejorar el análisis de las redes de regulación transcripcional que lleva a cabo la plataforma CGB con el algoritmo de maximización de la expectación.

4 MODELO

4.1 Expectation Maximization

EM es un algoritmo iterativo que se utiliza para estimar un resultado en función de unos parámetros. Se suele usar en casos en los que se tiene información faltante o incompleta y en contextos probabilísticos.

EM alterna entre dos pasos: Expectación (E) y Maximización (M):

- Expectación: Se estiman unos resultados a través de unos parámetros.
- Maximización: Se actualizan los parámetros para maximizar la expectación anterior.

Primero se utilizan los parámetros para estimar unos resultados (expectación) y, a partir de estos resultados, se actualizan los parámetros para maximizar los resultados (Maximización). Con estos parámetros actualizados, se estima un nuevo resultado (Expectación). Este proceso se repite iterativamente hasta que la mejora en la estimación sea mínima respecto la estimación anterior.

4.2 CGB

4.2.1. Vista general de CGB

CGB recibe un o más factores de transcripción, así como una lista de genomas en las bacterias a analizar, con unas secuencias de unión por cada TF. A partir de estas, ejecuta un análisis de la red de regulación transcripcional en esta bacteria y otras relacionadas con esta asignando la evidencia experimental en base a la distancia evolutiva. Se tiene en cuenta la conservación de estos sitios de unión a través de varias especies, para encontrar sitios funcionales. Entre otros datos, se encuentra la probabilidad de regulación de un gen dado este factor de transcripción y estas secuencias de unión.

4.2.2 Entorno

Este entorno se ha preparado en Linux.
La plataforma CGB se encuentra disponible en:
https://github.com/ErillLab/cgb3

Utiliza un entorno de miniconda, el entorno viene en preparado en un archivo yaml, para ejecutar el archivo solo tienes que ejecutar:
Conda env create -f conda_cgb_environment.yml

Depende de 3 programas externos Blast, Clustalo y BayesTraits.
Blast lo puedes instalar con el comando:
Sudo apt-get install ncbi-blast+

Clustalo lo puedes instalar con el comando:
Sudo apt-get install clustalo

BayesTraits está incluido en cgb en la carpeta bin.

4.2.2 Input

Recibe un archivo JSON. Este archivo JSON tiene 1 factor de transcripción (TF) con varias secuencias de unión conocidas de ese TF.
Este archivo también tiene varias opciones de configuración.

- Uno o varios TF
- Varios sitios de unión
- Opciones de configuración: En nuestro caso respecto la configuración por defecto usamos la opción site_printout: true

4.2.3 Output

Varias bacterias relacionadas por distancia filogenética.

1. Carpeta derived_PSWM
   1 archivo por bacteria.
   Archivo contiene una matriz PSWM que define 1 motivo para esta bacteria.
2. Carpeta posterior_probs
   1 archivo por bacteria.
   Archivo contiene una matriz con los diferentes genes de la bacteria y su probabilidad de regulación.
3. Carpeta identified_sites
   para obtenerla necesitas activar la opción site_printout en el Input.
   1 archivo por bacteria.
   Contiene una matriz con las secuencias de unión para cada gen.

4.3 Maximización

4.3.1. Resumen del modelo

A partir de los datos de CGB se crea este modelo.
Lo que se busca es con los genes, sus sitios de unión y su probabilidad de regulación, crear un nuevo motivo a partir del cual se generan nuevas secuencias de unión para volver a introducirlas en CGB y que busque nuevas probabilidades de regulación de los genes. Así ir iterando hasta que diferencia entre la probabilidad anterior y la nueva sea nula.
Así se construye un algoritmo de EM, donde el paso de expectación, con el que se buscan nuevas probabilidades de regulación, lo ejecuta CGB y el paso de maximización, con el que se actualiza el input de CGB, se define a continuación.

4.3.2. Paso de maximización

Nueva matriz

Con los datos de output de CGB: posterior-probabilities y identified-sites, se crea una nueva matriz que une posterior_probabilities y identified_sites, teniendo así la probabilidad de un gen enlazado a ese gen y a las secuencias de unión de ese gen.

Se selecciona los genes más probables con un umbral, por ejemplo, superior a 0.7.

Ejemplo:

Gen	Probabilidad	Secuencia de unión
Gen 1	0.95	ATGAT
Gen 2	0.87	TAGAT
Gen 3	0.7	AGGAT
Gen 4	0.5	CTTAT
Gen 5	0.5	GAGAA

A partir de esta matriz, se busca un nuevo motivo definido por la matriz updated_PSWM.

Para generar esta matriz, por cada secuencia de unión, en cada posición del motivo, se suma la probabilidad de regulación del gen a la base que coincide con la base de la secuencia de unión en aquella posición del motivo, por ejemplo (este proceso se visualiza mejor en el esquema de la página 5):

$$A_1 = 0.95 + 0.7 = 1.65$$
$$C_1 = 0.5$$
$$G_1 = 0.5$$
$$T_1 = 0.87$$

Updated_PSWM

	1	2	3	4	5
A	1.65				
C	0.5				
G	0.5				
T	0.87				

Pero para obtener una probabilidad dado que este resultado por posición excede la probabilidad máxima que es 100% se debe normalizar.

Normalizar

$$P = 1.65 + 0.5 + 0.5 + 0.87 = 3.52$$
$$A_1 = \frac{1.65}{3.52} = 0.47$$
$$C_1 = \frac{0.5}{3.52} = 0.14$$
$$G_1 = \frac{0.5}{3.52} = 0.14$$
$$T_1 = \frac{0.87}{3.52} = 0.25$$

Updated_PSWM

	1	2	3	4	5
A	0.47				
C	0.14				
G	0.14				
T	0.25				

$$0.47 + 0.14 + 0.14 + 0.25 = 1$$

Se repite este proceso para todas las posiciones del motivo.

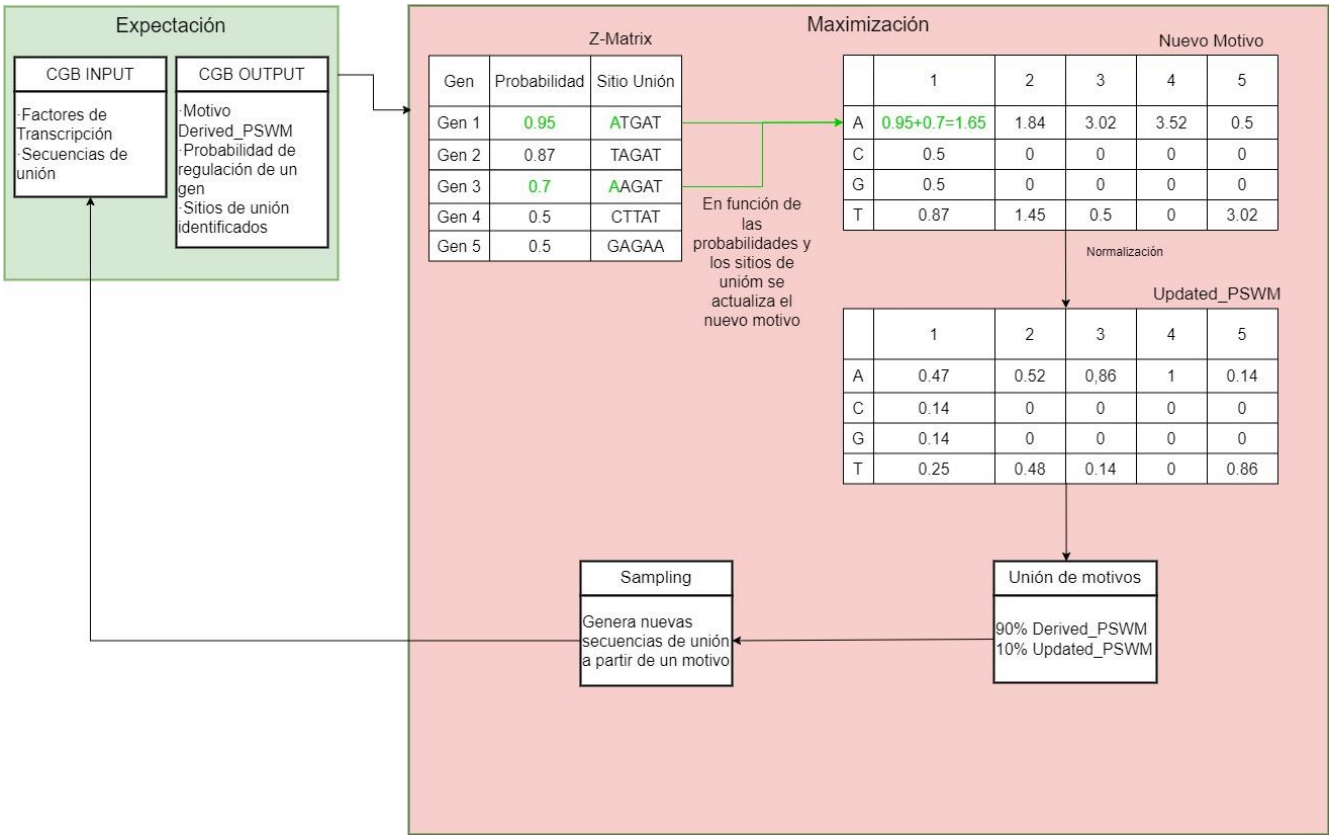
Dado que una bacteria puede tener pocos genes regulados, este método podría generar motivos poco precisos.
Por eso en vez de usar directamente updated_PSWM, lo que se hace es crear un motivo combinando updated_PSWM y derived_PSWM dónde el motivo original tiene una mayor parte del peso.

90% derived_PSWM
10% updated_PSWM

Con esto se actualiza el motivo, desde el que se hará un muestreo para generar nuevas secuencias de unión con las que se alimentará de nuevo CGB.

Este proceso se repetirá hasta que la diferencia entre la probabilidad anterior de regulación y la nueva probabilidad de regulación sea nula.

4.3.3. Esquema



5 Resultados

Análisis de las redes de regulación transcripcional para el factor de transcripción LexA, se prepara el siguiente input.


```
"TF": "LexA",
"motifs": [
    {
        "name": "LexA_Lmo",
        "genome_accessions": [
            "NC_003210.1"
        ],
        "protein_accession": "NP_464827.1",
        "sites": [
            "AAAAAGAATGTATGTTCGCTTT",
            "AAAAAGAATGTATGTTCGCTTT",
            "TGTACGAACGGTTGTTCTATAA",
            "AAAGCGAACATTTATTCGTATT",
            "ATATAGAACATACATTCGATTA",
            "AAAACGAACAAGCGTTCTTATT",
            "GTTGCGAACGTAGGTTCTGTGT",
            "AAAAAGAAAGTGTGTTCGTGTT",
            "TGATAAAACATATGTTCTGTTT",
            "CATACAAACATTTGTTCTTATT",
            "AAACCGAATATACGTTCTTATT",
            "CCACCGAACATATGTTTTTATT",
            "TTCAAGAACGTTTGTTCGTATA",
            "AAAAAGAACGTATGTGCGAAAG",
            "AAACCGAACATATTTTCGCATT",
            "AATAAGAACATTTGTTCGTATA",
            "TTTAAGAACGTTTGTTCGTATA"
        ]
    }
],
```


Donde se observa el TF LexA, y el TF LexA para el genoma de la bacteria Lmo (LexA_Lmo) con sus secuencias de unión conocidas (sites).

A partir de esta se analizará el TF LexA para el genoma de las siguientes bacterias: ace, Mtu, cgl, Cdi, Bsu, Sau, lxy. Estos genomas están especificados dentro del input, por ejemplo, a continuación, se observa cómo se incluye los genomas ace y Cdi en el input:


```
"genomes": [
    {
        "name": "ace",
        "accession_numbers": [
            "NC_008578.1"
        ]
    },
    {
        "name": "Cdi",
        "accession_numbers": [
            "NC_009089.1",
            "NC_008226.1"
        ]
    },
```


Con estos datos se inicia la ejecución.

En la página 7 podemos observar la figura de los motivos de las bacterias a analizar en la primera iteración y en la última iteración.

Se observan algunos resultados prometedores y algunos erróneos.

Podemos observar los resultados erróneos, en los genomas Cdi, Bsu, Sau y lxy. El mayor exponente del error es Bsu. En todas las posiciones, tiene un exceso de la base T porque en un algoritmo iterativo como EM se produce un riesgo de contaminación generando ruido, aumentando con cada subsecuente iteración.

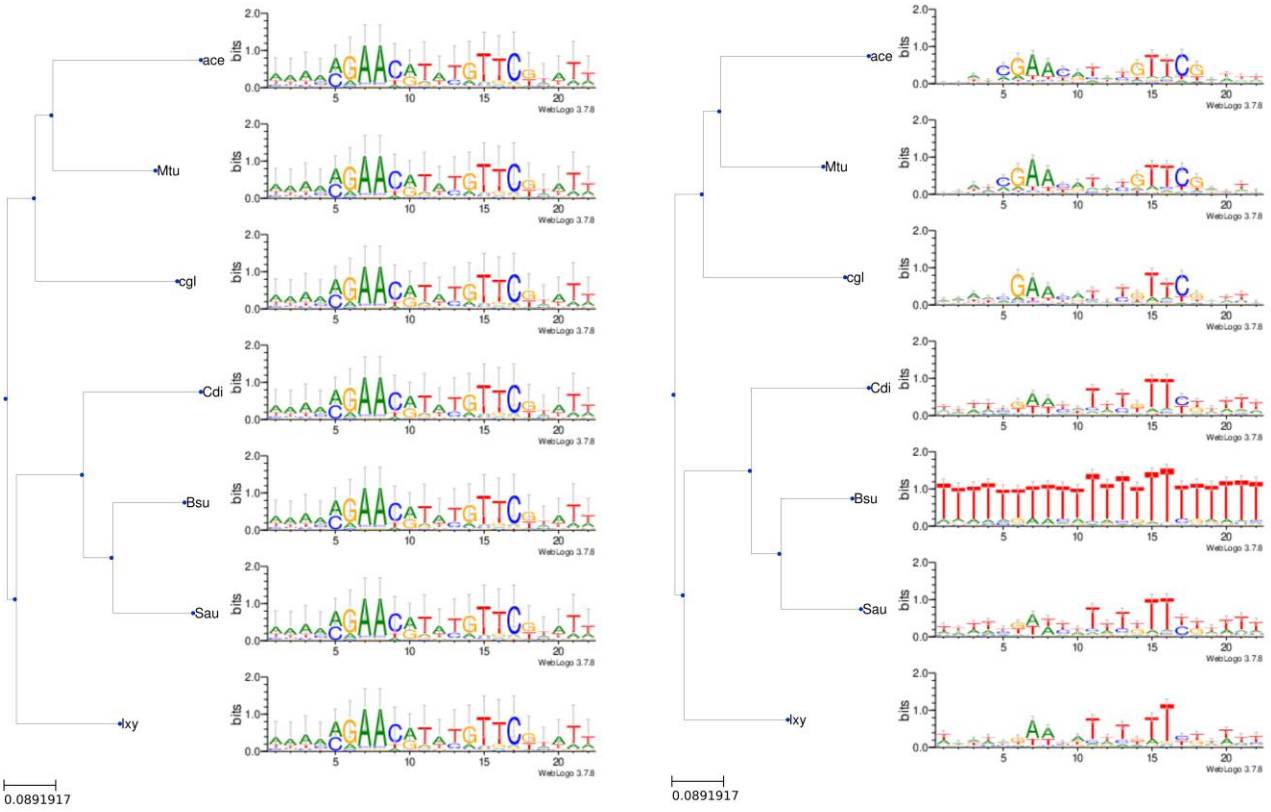
Lo que está pasando es que, en la primera iteración, solo unos pocos genes salen fuertemente regulados, cuyos sitios de unión tienen varias T. Esto hace que la matriz se actualice para buscar un patrón con T en distintas posiciones y deje de prestar atención al patrón real. Esto da lugar a que, en la siguiente iteración, se encuentren una serie de genes regulados con sitios de unión repletos de T. A partir de aquí, el sistema se estanca y se queda fijado con motivos formados completa o mayoritariamente por T.

Por otra parte, tenemos algunos resultados prometedores en las bacterias ace, Mtu, Cgl.

Observemos a continuación los genes regulados en la primera y última iteración para la bacteria ace:

Primera iteración:

Probabilidad	Genes Regulados
0.998	**ACEL_RS05115, ...**
0.998	**ACEL_RS05110**
0.976	**ACEL_RS07675, ...**
0.975	**ACEL_RS05130**
0.824	**ACEL_RS07440, ...**
0.824	**ACEL_RS07415, ...**



Izquierda primera iteración, derecha última iteración.

Última iteración:

Probabilidad	Genes Regulados
1.000	ACEL_RS07615
1.000	ACEL_RS11970
0.993	**ACEL_RS05115, ...**
0.992	**ACEL_RS05110**
0.990	**ACEL_RS05130**
0.980	**ACEL_RS07440, ...**
0.970	**ACEL_RS07675, ...**
0.940	**ACEL_RS07415, ...**
0.935	ACEL_RS12355, ...
0.916	ACEL_RS06785
0.907	ACEL_RS08210
0.904	ACEL_RS08220
0.901	ACEL_RS06265
0.898	ACEL_RS06270
0.863	ACEL_RS08215
0.850	ACEL_RS12600
0.825	ACEL_RS07395, ...
0.744	ACEL_RS10935, ...

Lo que se observa es que los genes que están regulados en la primera iteración se mantienen con probabilidades altas de regulación y se añaden nuevos genes regulados.

Es importante comentar que los genes que se encuentran en la primera pasada, así como los nuevos genes con alta probabilidad de regulación, están la mayoría relacionados con procesos de reparación del ADN, que es lo que controla la red de regulación SOS (coordinada por el factor de transcripción LexA), por lo tanto, los resultados obtenidos por EM son consistentes con lo que se conoce experimentalmente sobre esta red de regulación.

Lo que ocurre en este proceso es que el motivo se especializa para cada especie, obteniendo así unos mejores resultados, al conseguir una mayor cantidad de genes regulados con una alta probabilidad y mayores evidencias consistentes con los resultados conocidos de LexA.

6 Conclusión

Se concluye que los resultados obtenidos son prometedores, se obtiene una especialización del motivo para la especie, pero es necesario mejorar el algoritmo introduciendo un método para controlar el ruido.

Como paso a futuro, se podría introducir en el algoritmo un monitoreo del motivo, y cuando este se desvía excesivamente del motivo inicial, aplicar unas correcciones o reiniciarlo.

Si se mejora el algoritmo, se podrá utilizar para obtener más información al hacer el análisis sobre las redes de regulación transcripcional.

AGRADECIMIENTOS

A mi tutor Ivan Erill por tener tanta paciencia y estar disponible todas las veces que le pedí ayuda.

BIBLIOGRAFÍA

[1] D'haeseleer, P. What are DNA sequence motifs?. Nat Biotechnol 24, 423–425 (2006). https://doi.org/10.1038/nbt0406-423

[2] Kılıç, S., Sánchez-Osuna, M., Collado-Padilla, A. et al. Flexible comparative genomics of prokaryotic transcriptional regulatory networks. BMC Genomics 21 (Suppl 5), 466 (2020). https://doi.org/10.1186/s12864-020-06838-x

[3] Rye, C., Wise, R., Jurukovski, V., DeSaix, J., Choi, J., & Avissar, Y. (2016). Biology. Houston, Texas: OpenStax. https://openstax.org/books/biology/pages/1-introduction

[4] Bailey, T.L., Elkan, C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Mach Learn 21, 51–80 (1995). https://doi.org/10.1007/BF00993379

[5] Phillip Compeau & Pavel Pevzner. Bioinformatics Algorithms II. La Jolla, CA. Active Learning Publishers; 3rd edition (August 15, 2018). https://www.bioinformaticsalgorithms.org/

[6] Touchman, J. (2010) Comparative Genomics. Nature Education Knowledge 3(10):13

APÉNDICE

El código está disponible en:
https://github.com/ErillLab/cgb3EM