Trabajo de Fin de Máster

Análisis computacional de periodicidades en cadenas de ADN nucleosomales

Trabajo de Fin de Máster en Ingeniería Informática

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

Autor: Ángel Becedas Mesa

Tutor: Rodrigo Santamaría Vicente

Certificado del tutor

D. Rodrigo Santamaría Vicente, profesor/a del Departamento de Informática y Automática de la Universidad de Salamanca.

HACEN CONSTAR:

Que el trabajo titulado “Análisis computacional de periodicidades en cadenas de ADN nucleosomales” ha sido realizado por D. Ángel Becedas Mesa, con DNI 45135476-T y constituye la memoria del trabajo realizado para la superación de la asignatura Trabajo de Fin de Máster de la Titulación Máster en Ingeniería Informática de la Universidad de Salamanca.

Y para que así conste a todos los efectos oportunos.

En Salamanca, a 03 de septiembre de 2021

D. Rodrigo Santamaría Vicente

Firmado por SANTAMARIA VICENTE RODRIGO - 70879303L el día 03/09/2021 con un certificado emitido por AC FNMT Usuarios

Tabla de contenido

Índice de ilustraciones

Índice de tablas

1. Introducción

El objetivo de este trabajo es llevar a cabo un análisis de periodicidades sobre una cadena de ADN de una levadura de la familia *S. Pombe.* Varios autores han realizado estudios similares con el fin de encontrar patrones de repetición en secuencias de ADN. Por ejemplo, podemos ver este tipo de análisis en trabajos como "*Visible periodicity of strong nucleosome DNA sequences*" de Bilal Saliha, Vijay Tripathia y Edward N. Trifonova o en "*The 3-Base Periodicity and Codon Usage of Coding Sequences Are Correlated with Gene Expression at the Level of Transcription Elongation*" de Edoardo Trotta. Estos artículos y algunos similares, sirven como referencia a la hora de realizar este trabajo y se incluyen las referencias al final de este documento. En caso de darse, este tipo de repeticiones podrían indicar algunas características del ADN que pueden ser de utilidad en tareas como el estudio de la replicación del material genético de una célula al dividirse o como en este caso, la forma en la que el ADN se empaqueta dentro del núcleo de la célula.

Con este trabajo, se pretende también mostrar que la informática no consiste solamente en el puro desarrollo de aplicaciones web o móviles, sino que también puede ser una herramienta de gran ayuda para muchas otras ramas de la ciencia como la biología, la física, las matemáticas, la sociología o la economía.

La tecnología toma cada vez más presencia en todos los ámbitos y el avance tecnológico es clave para que se produzca también un avance científico.

1.1. Conceptos fundamentales

En este trabajo se combinan conceptos de dos ramas distintas de la ciencia, la informática y la biología. Esta combinación se conoce como bioinformática, un campo interdisciplinar que se centra en el desarrollo y aplicación de herramientas software para analizar datos biológicos. Estos algoritmos y herramientas informáticas permiten analizar conjuntos de datos de gran tamaño y complejidad.

A continuación, se definen algunos conceptos de biología que serán importantes a lo largo de este trabajo.

- ADN. El Ácido Desoxirribonucleico (ADN) es una molécula que contiene el material genético de los seres vivos. Esta molécula está compuesta por dos cadenas que se enrollan entre sí creando una estructura de doble hélice. Las dos cadenas se unen a través de enlaces entre bases. Hay cuatro bases posibles, cada una identificada por su inicial, Adenina (A), Timina (T), Citosina (C) y Guanina (G). Los enlaces son siempre entre A y T o entre C y G.

- Nucleótido. Las bases nitrogenadas (A, T, C y G) no son las únicas que conforman el ADN. Estas bases se combinan con otros compuestos para formar moléculas orgánicas que se conocen como nucleótidos. Cada una de las dos cadenas que conforman el ADN está compuesta por una serie de nucleótidos. Los nucleótidos se unen siempre por parejas y las bases siempre se unen con otra base fija, por lo que se obtienen parejas Adenina – Timina y Citosina – Guanina. Estas parejas de nucleótidos se conocen como dinucleótidos.

- Nucleosoma. Las cadenas de ADN pueden llegar a medir varios metros en algunas especies como la humana, y toda esa información debe empaquetarse para que pueda ser almacenada dentro del núcleo de las células. Alrededor de 150 pares de bases se enrollan alrededor de un núcleo de histonas (proteínas) dando lugar a un nucleosoma. Los nucleosomas se organizan como "cuentas" en un collar que, a su vez, se pliegan sobre sí mismas para formar finalmente los cromosomas. Estos grupos de 150 pares de bases que forman los nucleosomas serán los fragmentos de ADN sobre los que se realiza el estudio que se presenta en este documento.

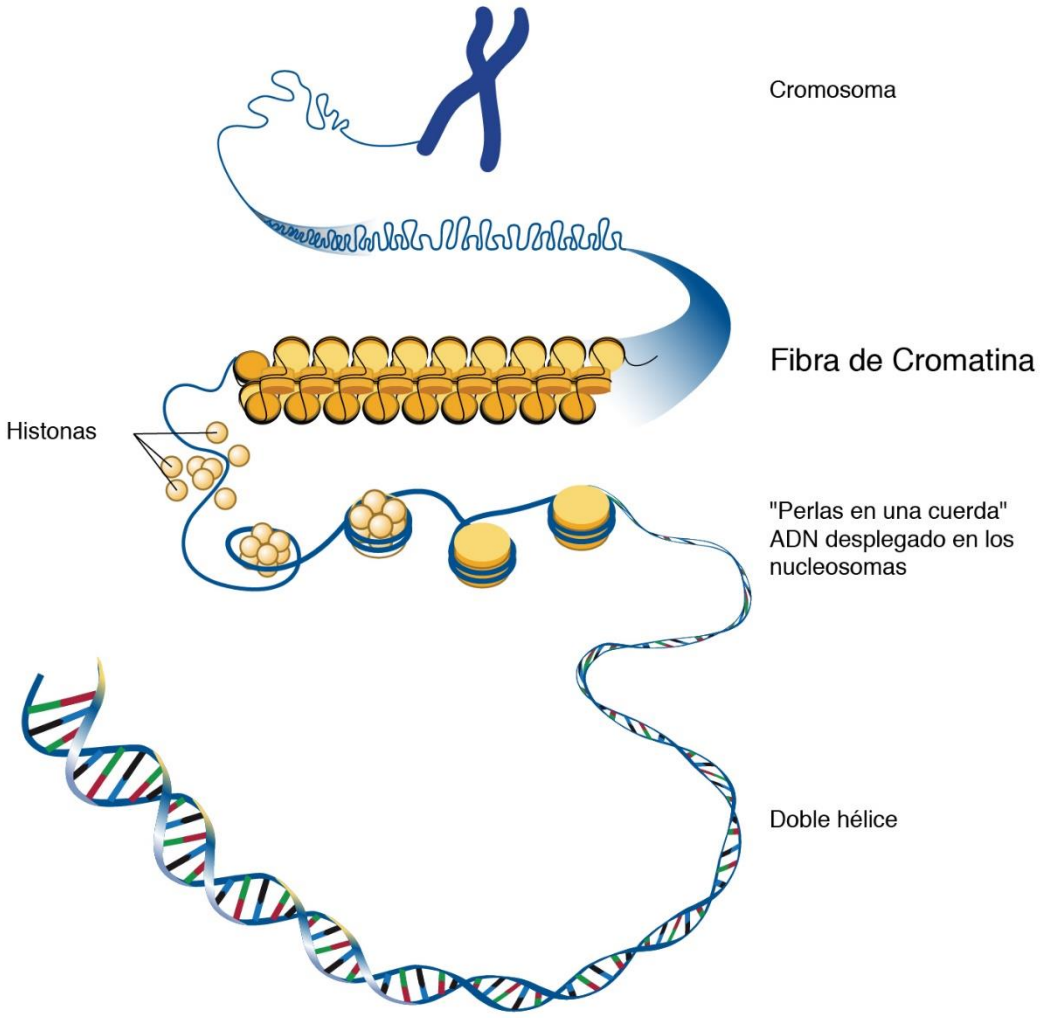

Ilustración 1: Esquema de la compactación del ADN (National Human Genome Research Institute)

- Secuenciación de ADN. Se conoce como secuenciación de ADN al proceso de determinación de la estructura de las cadenas de ADN. Una secuencia de ADN es una serie ordenada de nucleótidos representados por su base.

También un par de conceptos relacionados con el algoritmo a implementar:

- Ventana (window). Longitud de la sección que se va a analizar dentro de una cadena.

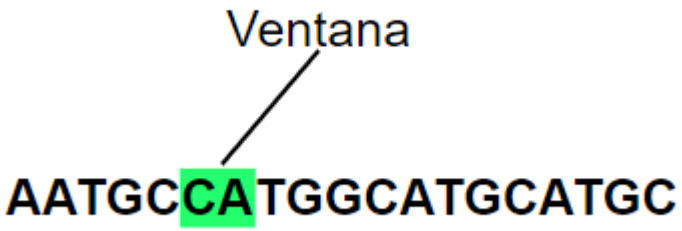

Ilustración 2: Esquema del concepto de ventana

- Paso (step). Número de posiciones que avanza la ventana en cada iteración del análisis de la cadena completa.

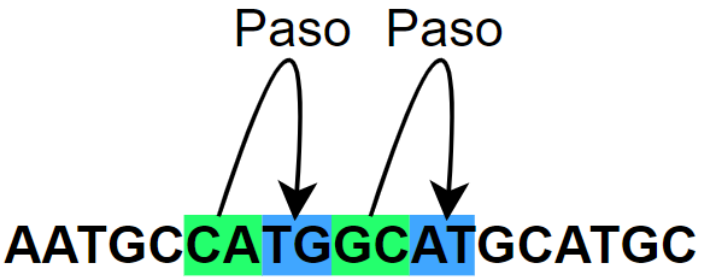

Ilustración 3: Esquema del concepto de paso

1.2. Caso de estudio

Como se expone en la introducción, el objetivo de este trabajo es realizar un análisis sobre el ADN secuenciado de una levadura para buscar periodicidades dentro de la secuencia. Hay varias formas de buscar este tipo de periodicidades en la secuencia, pero en este caso particular, el método utilizado se basa en el cálculo de distancias de los diferentes dinucleótidos dentro de secciones de la secuencia completa.

Desde el punto de vista biológico, estos patrones podrían indicar o caracterizar las zonas de la cadena de ADN en las que se colocan las proteínas para formar los nucleosomas.

Se entiende como distancia al número de pasos (steps) desde la primera aparición del dinucleótido, la ventana a analizar, hasta que vuelve a aparecer. En este caso, el step es de dos posiciones (las ventanas no se solapan) pero también se podría realizar el estudio con un step de una posición (con ventanas solapantes). Por ejemplo, si tenemos la siguiente sección de ADN:

ATCGTAACCGAT

Si estudiamos el dinucleótido AT, aparecería otro a distancia 5.

AT	CG	TA	AC	CG	**AT**
-	1	2	3	4	**5**

Se repite este proceso para todos los dinucleótidos de la cadena hasta un valor de distancia máximo de 10, es decir, se analizan diez ventanas en diez pasos consecutivos.

2. Metodología

La parte de desarrollo software de este trabajo consiste en la creación de métodos de lectura, procesado y escritura de datos que permitan tratar la información del caso de estudio. También se creará un método que permita realizar el cálculo de la distancia de los dinucleótidos siguiendo el procedimiento expuesto anteriormente.

Todo el software se desarrolla con lenguaje Python, en su versión 3.8.3. Se ha elegido este lenguaje de programación por su sencillez y por su creciente popularidad en tareas del campo de la ciencia de datos o "Data Science" como la minería de datos utilizada en el "Machine Learning" y el "Big Data". También cuenta con librerías y herramientas desarrolladas para tareas de bioinformática en proyectos como BioPython o Bioconda. Se han utilizado algunas librerías como "os" y "csv" para la escritura de los resultados y otras para el tratamiento de la información como "itertools", "pandas" o "numpy".

El código desarrollado se ha publicado en GitHub bajo la licencia GPL 3. El enlace al repositorio es el siguiente: https://github.com/angelUsal/Analisis-periodicidades-ADN-TFM

2.1. Lectura y escritura de datos

Para el tratamiento de los datos, principalmente de las secuencias de ADN, se utilizan dos formatos de documento de texto muy extendidos en el ámbito de la bioinformática, el formato FASTA y el BED.

Un documento en formato FASTA se caracteriza por tener cabeceras que identifican a una secuencia determinada (por ejemplo, identifica un cromosoma) y una serie de caracteres (A, T, C y G) para representar los nucleótidos de las diferentes secuencias. Las cabeceras comienzan con el carácter especial ">". En este caso, contamos con la secuencia de la levadura en formato FASTA. En el documento podemos ver tres cabeceras con sus tres secuencias correspondientes, cada una para un cromosoma.

```
1  >seq1
2  ATGTTGCCGTATAA
3  >seq2
4  ACCTCTGTAGCCAC
5  >seq3
6  GCCTACGATCTAGT
```

Ilustración 4: Ejemplo de un archivo con formato FASTA

El formato BED, se utiliza para almacenar secciones de ADN en forma de coordenadas. Suelen ser archivos tabulados y en cada columna se almacena una información distinta. El mínimo es de tres columnas, una para identificar el cromosoma, otra para la coordenada de inicio y otra para la coordenada de fin del fragmento. Además, se pueden incluir otras columnas con información extra como un nombre, una puntuación o la orientación de la hebra

de ADN. Para este caso de estudio se utiliza un archivo BED en el que se detallan las coordenadas de los fragmentos de secuencia sobre los que se llevará a cabo el análisis de distancias.

```
chromosome1 1   150
chromosome1 250 400
chromosome2 50  300
chromosome2 100 500
chromosome3 10  90
chromosome3 70  125
```

Ilustración 5: Ejemplo de un archivo con formato BED

Existen muchas librerías que proporcionan métodos para tratar este tipo de archivos. En este caso, se utiliza la librería bioio.py proporcionada por el tutor como material para este trabajo (https://github.com/rodrigoSantamaria/ibfg/blob/master/bioio.py). También se proporcionan como material el FASTA y el BED de la levadura *S. Pombe*.

2.2. Cálculo de distancias

Obtención de datos

El primer paso para realizar el análisis es combinar la información de la secuencia completa del documento FASTA con las secciones que aparecen en el BED. Para esto, se implementa un método que lee ambos documentos y para cada par de coordenadas del BED (más un posible margen que se pasa como argumento a la función), obtiene las secciones correspondientes del FASTA y escribe el resultado en un nuevo documento de tipo FASTA.

Ilustración 6: Esquema del procesamiento de los datos de entrada

En este archivo de resultados, el identificador de cada secuencia del FASTA se compone del cromosoma y las coordenadas de origen y fin de la sección. A continuación, se muestra un esquema de este proceso.

Algoritmo de cálculo de distancias

Una vez obtenido el nuevo FASTA, se procede al cálculo de las distancias. Para ello, se obtienen las secciones almacenadas en el documento y sobre cada una se aplica el algoritmo del cálculo de distancias.

El algoritmo funciona de la siguiente manera:

1. Se obtiene el primer dinucleótido de la cadena a analizar, es decir, los dos primeros caracteres.
2. Se lee el siguiente dinucleótido y se comparan para comprobar si se trata del mismo. Se van leyendo dinucleótidos diferentes por lo que el paso es de 2 posiciones (las ventanas no se solapan).
    - Si son iguales, se ha encontrado una repetición y se aumenta el contador en el que se almacenan las ocurrencias para ese dinucleótido y esa distancia.
3. Luego, se obtiene la siguiente pareja y se vuelve a comparar. Esto se repite hasta llegar al dinucleótido a distancia 10, que es la máxima que se va a analizar.
4. Una vez se ha hecho el análisis del primer dinucleótido, se repite con el siguiente, y así hasta llegar al final de la sección.
5. Se toma la siguiente sección del FASTA y se repite el proceso.

Todas las ocurrencias se almacenan en un diccionario Python en el que las claves son los diferentes dinucleótidos y para cada clave, un vector de diez posiciones. En cada posición del vector se almacena el contador de ocurrencias para cada distancia (de 1 a 10).

3. Análisis de resultados

El diccionario obtenido se puede ver como una matriz que contiene la información del número de ocurrencias para cada dinucleótido para distancias de entre 1 y 10 posiciones. La matriz cuenta con 10 columnas, una para cada valor de las posibles distancias, y una fila para cada uno de los posibles dinucleótidos (las 16 posibles combinaciones de los nucleótidos Adenina, Timina, Citosina y Guanina).

Los autores Bilal Salihab, Vijay Tripathia & Edward N. Trifonova, denominan a estas estructuras matriciales como “matrices de doblabilidad” bajo la hipótesis de que los puntos donde se observan picos en la periodicidad se corresponden con los lugares donde la cadena de ADN es más flexible y, por lo tanto, donde se puede doblar para formar los nucleosomas en el empaquetado.

Buscar los patrones en las matrices de datos puede ser complicado, sin embargo, si se representan en forma de gráfica, se pueden analizar los datos con mayor facilidad.

3.1. Resultados por grupos

Para facilitar la visualización de los datos en las tablas y para evitar que se produzca mucho solapamiento en las gráficas, se separan los resultados separados en cuatro grupos, con cuatro dinucleótidos cada uno. En este caso, se agrupan los dinucleótidos según su primera base.

Para cada grupo se genera una tabla a partir de los resultados obtenidos en el algoritmo y con los datos de dicha tabla, se generan las diferentes gráficas que permitirán detectar si existen o no patrones claros.

Primer grupo

A continuación, se muestra la "matriz de doblabilidad" para los cuatro primeros dinucleótidos.

	1	2	3	4	5	6	7	8	9	10
AA	60025	54684	52926	54200	53614	54087	50385	49341	49730	49818
AT	26005	31324	31180	29801	28307	30840	27778	27410	28485	27014
AC	9295	10022	10075	9397	9136	9780	8699	8762	9091	8580
AG	12812	12542	14386	11907	11613	13863	11209	10916	13085	10639

Tabla 1: Tabla de resultados del primer grupo

A partir de estos datos, se genera el siguiente gráfico:

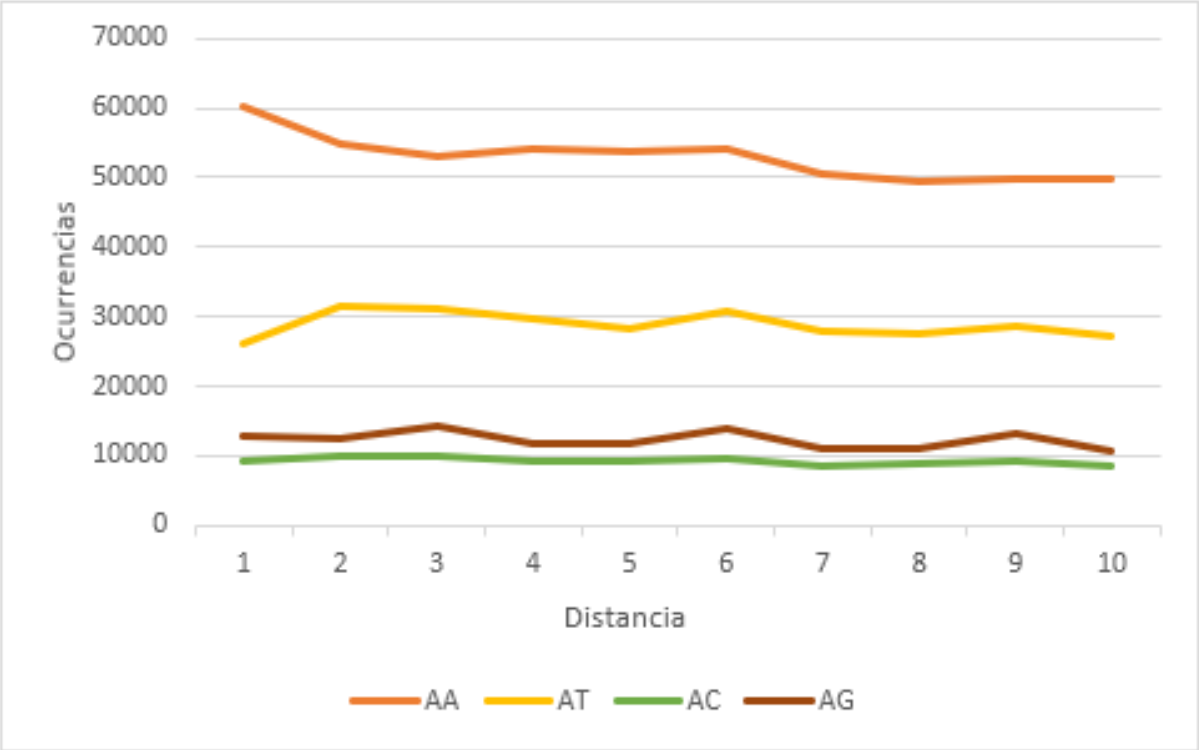

Ilustración 7: Gráfico de resultados del primer grupo

Analizando el gráfico obtenido, se puede observar que, para algunos dinucleótidos, parece que se ve un pico de ocurrencias en la distancia 6. Además, para el dinucleótido AG, se observan picos relativamente claros en las distancias 3, 6 y 9. Este es el caso más claro dentro del grupo.

Segundo grupo

A continuación, se muestra el análisis de los siguientes cuatro dinucleótidos (TA, TT, TC y TG).

	1	2	3	4	5	6	7	8	9	10
TA	20324	21276	22612	20688	19898	22180	19590	19036	20484	18712
TT	60234	55261	53463	54570	53432	54491	50860	49728	49654	49904
TC	13624	14255	17591	13695	12883	17013	12342	12249	15611	11830
TG	10000	14220	17741	13763	12930	16878	12581	12717	15554	12439

Tabla 2: Tabla de resultados del segundo grupo

La gráfica correspondiente a los datos obtenidos es la siguiente:

Ilustración 8: Gráfico de resultados del segundo grupo

Para tres de los cuatro dinucleótidos, TA, TC y TG, se observa un patrón similar al que se observaba para el dinucleótido AG, esto es, ligeros picos de ocurrencias para las distancias 3, 6 y 9.

Tercer grupo

A continuación, se muestra el análisis de los siguientes cuatro dinucleótidos (CA, CT, CC y CG).

	1	2	3	4	5	6	7	8	9	10
CA	10060	14373	17687	13544	12864	16664	12743	12633	15584	12284
CT	12886	12467	14316	11975	11637	14101	11037	11064	13219	10806
CC	2761	4199	5226	4318	4168	4877	3956	3883	4314	3770
CG	2545	2831	3822	3043	2767	3604	2878	2909	3203	2712

Tabla 3: Tabla de resultados del tercer grupo

Y la gráfica resultante.

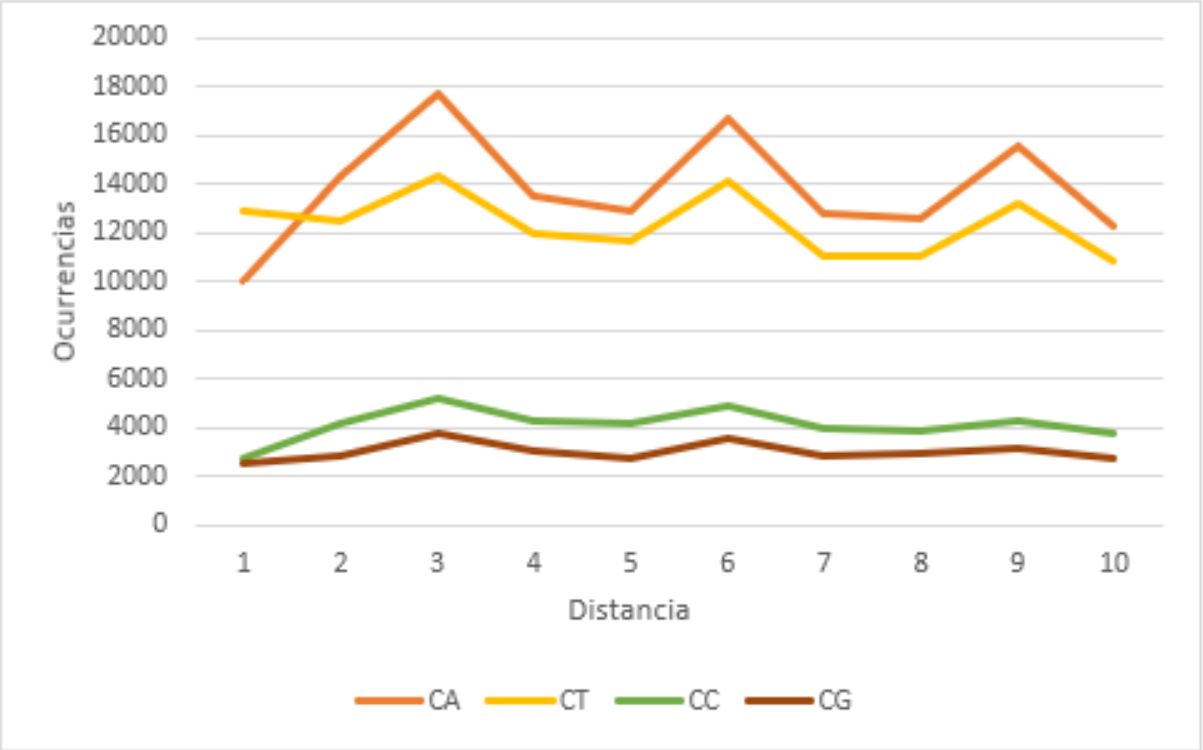

Ilustración 9: Gráfico de resultados del tercer grupo

En este grupo también se puede observar el mismo patrón para todos los dinucleótidos, especialmente, para los dinucleótidos CA y CT.

Cuarto grupo

A continuación, se muestra el análisis de los últimos cuatro dinucleótidos (GA, GT, GC y GG).

	1	2	3	4	5	6	7	8	9	10
GA	13151	14160	17312	13634	12659	16379	12125	12141	15327	11802
GT	9362	10155	10196	9550	9049	9800	8798	8706	9208	8483
GC	3352	5062	6574	4846	4837	6250	4741	4524	5585	4511
GG	2865	4209	5115	4436	4132	4880	3975	3856	4434	3858

Tabla 4: Tabla de resultados del cuarto grupo

Y la gráfica resultante.

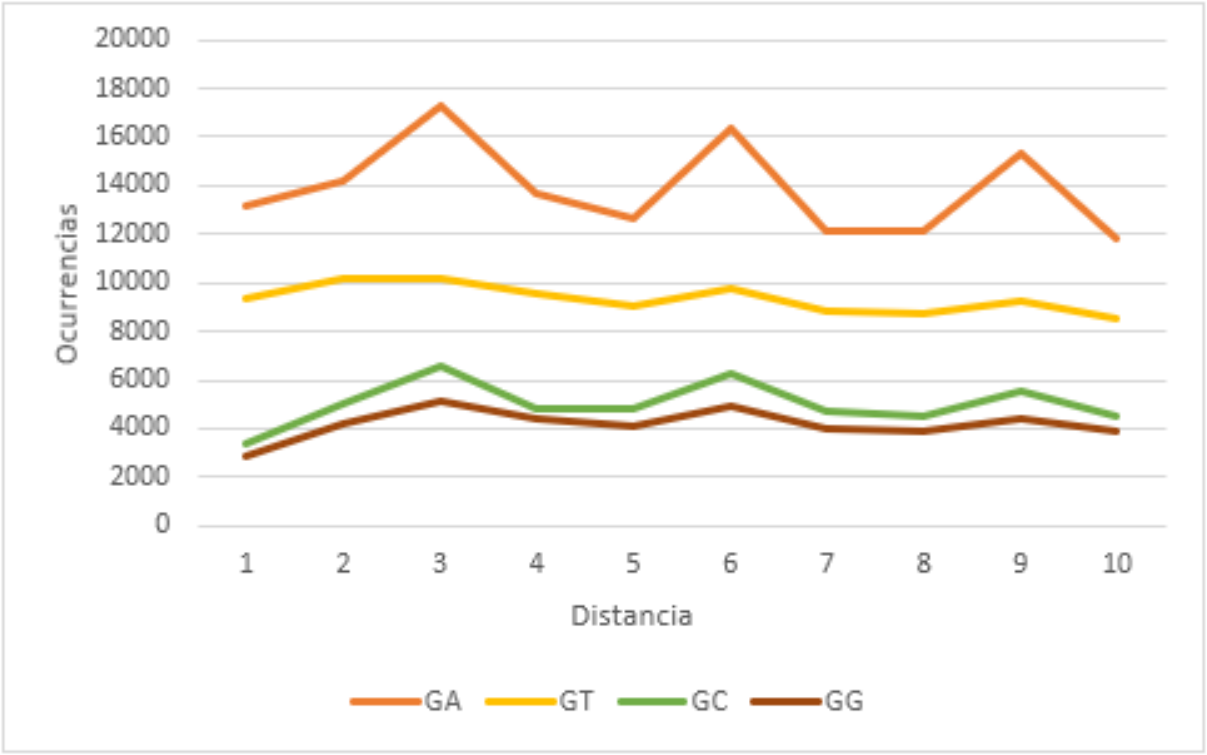

Ilustración 10: Gráfico de resultados del cuarto grupo

Como ocurría en el segundo grupo, se aprecia el patrón claramente en tres de los cuatro dinucleótidos, GA, GC y GG, siendo especialmente notable en el caso de GA. Para el dinucleótido se observan ligeramente dos de los picos, para las distancias 6 y 9.

3.2. Resultados globales

A continuación, se muestran la matriz y la gráfica para los dieciséis dinucleótidos.

	1	2	3	4	5	6	7	8	9	10
AA	60025	54684	52926	54200	53614	54087	50385	49341	49730	49818
AT	26005	31324	31180	29801	28307	30840	27778	27410	28485	27014
AC	9295	10022	10075	9397	9136	9780	8699	8762	9091	8580
AG	12812	12542	14386	11907	11613	13863	11209	10916	13085	10639
TA	20324	21276	22612	20688	19898	22180	19590	19036	20484	18712
TT	60234	55261	53463	54570	53432	54491	50860	49728	49654	49904
TC	13624	14255	17591	13695	12883	17013	12342	12249	15611	11830
TG	10000	14220	17741	13763	12930	16878	12581	12717	15554	12439
CA	10060	14373	17687	13544	12864	16664	12743	12633	15584	12284
CT	12886	12467	14316	11975	11637	14101	11037	11064	13219	10806
CC	2761	4199	5226	4318	4168	4877	3956	3883	4314	3770
CG	2545	2831	3822	3043	2767	3604	2878	2909	3203	2712
GA	13151	14160	17312	13634	12659	16379	12125	12141	15327	11802
GT	9362	10155	10196	9550	9049	9800	8798	8706	9208	8483
GC	3352	5062	6574	4846	4837	6250	4741	4524	5585	4511
GG	2865	4209	5115	4436	4132	4880	3975	3856	4434	3858

Tabla 5: Tabla de resultados totales

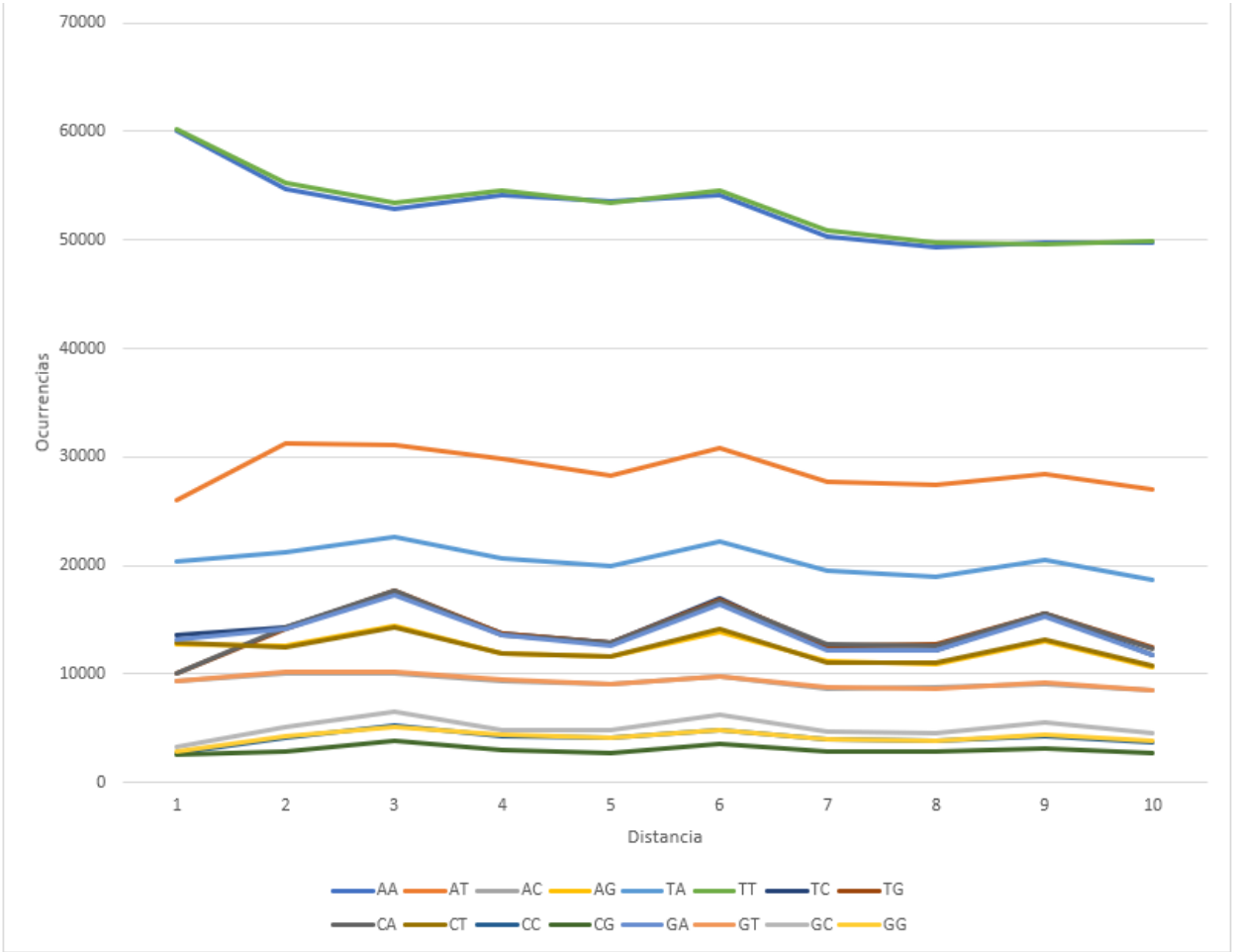

Ilustración 11: Gráfico de resultados totales

El único punto claro para todos los dinucleótidos es en la distancia 6 y para la mayoría, menos las excepciones expuestas anteriormente, se observan también picos en las distancias 3 y 9.

3.3. Análisis específico

Teniendo en cuenta los resultados del análisis inicial, se toma uno de los casos más claros, el de la pareja GA. Para este análisis específico, se aplica el mismo algoritmo, pero, en vez de generar una matriz global, se obtiene una para cada entrada del FASTA, es decir, para cada fragmento de longitud de 150 nucleótidos, que representa la sección de ADN a la que se une un nucleosoma.

Una vez hecho esto, se filtran aquellas listas en las que al menos una de las posiciones tenga un número de ocurrencias superior a 6.

A continuación, se muestra la tabla con los resultados obtenidos.

1	2	3	4	5	6	7	8	9	10
4	4	10	4	6	9	4	4	6	5
0	0	6	1	0	7	0	0	7	0
0	0	10	0	0	10	0	0	8	0
0	1	8	1	1	8	0	1	8	0
1	1	7	0	1	6	0	1	7	0
0	1	7	1	1	7	1	1	8	1
0	0	8	0	0	10	0	0	10	0
0	0	8	0	0	9	0	0	8	0
1	2	7	0	1	6	2	2	7	1
6	9	5	7	6	6	5	5	5	6
7	7	8	6	5	5	5	4	3	4

Tabla 6: Tabla de resultados del análisis específico

Analizando los datos de la tabla, se aprecia nuevamente el patrón anterior, con el mayor número de ocurrencias en las distancias 3, 6 y 9 para la mayoría de los casos.

4. Conclusiones

Como conclusiones desde el punto de vista biológico, se han obtenido resultados que parecen claros, sobre todo, para alguno de los dinucleótidos. Sin embargo, el hecho de encontrar este tipo de patrones no garantiza que en esos puntos concretos de la cadena sea donde se unen las proteínas para formar los nucleosomas en la compactación del ADN dentro del núcleo celular.

Por lo tanto, se debería realizar un análisis más completo y tener en cuenta otros factores para llegar a una conclusión. Hasta el momento, ningún autor ha conseguido descifrar el por qué los nucleosomas se forman en el lugar en el que lo hacen.

Desde el punto de vista del proyecto como Trabajo de Fin de Máster en Ingeniería Informática, se ha cumplido el objetivo de utilizar la informática no solo como una herramienta de creación de aplicaciones, sino como una parte fundamental para cualquier otra disciplina científica.

5. Líneas de trabajo futuras

A partir de este punto, se abren varias posibles líneas de trabajo futuras para este proyecto.

En primer lugar, en cuanto al proceso de análisis, se podría realizar el mismo análisis, pero utilizando trinucleótidos en lugar de dinucleótidos, o utilizando, por ejemplo, ventanas solapantes (con un paso de una posición).

Además, se podrían realizar más análisis individuales de los casos más claros, y comparar los resultados con los mapas de nucleosomas de esta levadura. También se podría analizar si esos casos más claros se deben a algún otro factor.

También, para asegurar que los datos obtenidos son válidos, se podría realizar el análisis mezclando las cadenas de ADN de forma aleatoria, es decir, “barajando” los datos para comprobar que solo se observan estos patrones en el ADN sin modificar.

Por último, desde el punto de vista de la programación, se podría mejorar la herramienta de varias formas. Por ejemplo, se podría internacionalizar o crear un instalador que descargue también las bibliotecas que se utilizan en el código. También sería recomendable implementar algún tipo de control de errores que muestre mensajes adecuados cuando las funciones no se estén utilizando de forma correcta.

6. Referencias

1. Ilya Ioshikhes, Alex Bolshoy, Konstantin Derenshtey, Mark Borodovsky and Edward N. Trifonov (1996), “Nucleosome DNA Sequence Pattern Revealed by Multiple Alignment of Experimentally Mapped Sequences”.

2. Edward N. Trifonov, Joel L. Sussman (1980), “The pitch of chromatin DNA is reflected in its nucleotide sequence”.

3. Edoardo Trotta (2011), “The 3-Base Periodicity and Codon Usage of Coding Sequences Are Correlated with Gene Expression at the Level of Transcription Elongation”.

4. Bilal Salihab, Vijay Tripathia & Edward N. Trifonova (2015), “Visible periodicity of strong nucleosome DNA sequences”.

5. Yulia M.Suvorova, Maria A.Korotkova, Eugene V.Korotkov (2014), “Comparative analysis of periodicity search methods in DNA sequences”, https://www.sciencedirect.com/science/article/pii/S1476927114000875

6. Hu Jin, H Tomas Rube, Jun S Song (2016), “Categorical spectral analysis of periodicity in nucleosomal DNA”. https://pubmed.ncbi.nlm.nih.gov/26893354/

7. National Human Genome Research Institute, “Nucleosoma”, https://www.genome.gov/es/genetics-glossary/Nucleosoma

8. Bilal Salihab, Vijay Tripathia & Edward N. Trifonova (2015), “Visible periodicity of strong nucleosome DNA sequences”.

9. Rodrigo Santamaría, librería bioio.py, https://github.com/rodrigoSantamaria/