Reconocimiento de genes en secuencias de ADN por medio de imágenes.

Luis A. Santamaría C. *Fac. Cs Computación. BUAP* Puebla, México ALULOC_KAPA@hotmail.com

Sarahí Zuñiga H. *Fac. Cs Computación. BUAP* Puebla, México sarahi.zuhe@gmail.com

Ivo H. Pineda T. *Fac. Cs Computación. BUAP* Puebla, México

María J. Somodevilla *Fac. Cs Computación. BUAP* Puebla, México mariasg@cs.buap.mx

Mario Rossainz L. *Fac. Cs Computación. BUAP* Puebla, México

***Resumen*—En los últimos años, el campo del aprendizaje automático ha progresado enormemente al abordar problemas difíciles de clasificación. El problema planteado en este artículo es reconocer secuencias de ADN, reconocer los límites entre exones e intrones utilizando una representación gráfica de secuencias de ADN y métodos recientes de aprendizaje profundo. El objetivo de este trabajo es clasificar secuencias de ADN utilizando una red neuronal convolucional (Convolutional Neural Network CNN). El conjunto de secuencias de ADN utilizado es la base de datos Molecular (Splice-junction Gene Sequences) Data Set que cuenta con 3190 secuencias, disponible en la página de la UCI, con tres clases de secuencias: límite exón-intrón, límite intrón-exón y ninguna. Se utilizó el conjunto de secuencias de ADN utilizado para el reconocimiento fueron 1847 secuencias de una base de datos con 4 tipos de virus de hepatitis C (tipo 1, 2, 3 y 6) tomada del repositorio disponible en la página de ViPR. Para la utilización de las secuencias de ADN se diseñó un método de representación donde cada base nitrogenada es representada en escala de grises para formar una imagen. Las imágenes generadas se utilizaron para entrenar la red neuronal convolucional. Los resultados muestran que una CNN puede hacer la clasificación de secuencias de ADN con un porcentaje de precisión de entrenamiento del 82 %, una precisión de validación del 75 % y una precisión de evaluación del 80.8 %. Se llega a la conclusión de que es posible clasificar las imágenes de secuencias de ADN de la base de datos empleada.**

***Palabras clave*—Reconocimiento de genes, Aprendizaje profundo, Redes neuronales convolucionales, Codificación de secuencias de ADN .**

I. Introducción

Los métodos de aprendizaje automático permiten identificar características que propician la clasificación, análisis y reconocimiento de patrones. En el área de la biología, el uso de métodos de aprendizaje automático, facilitan el reconocimiento de secuencias de ADN. Este trabajo reconoce los genes de ADN previamente procesados para ser representados por una imagen. Este artículo es divido en secciones, la primera sección es el estado del arte que es el conocimiento previo necesario para el reconocimiento de genes. La segunda sección detalla la metodología que fue utilizada para el análisis de las secuencias de ADN. La tercera y cuarta sección muestran los resultados y las conclusiones obtenidas.

II. Estado del Arte

Los mecanismos o procesos de predicción de genes son aquellos que, dentro del área de la biología computacional, se utilizan para la identificación algorítmica de trozos de secuencias, usualmente ADN genómico [8], y que son biológicamente funcionales. Esto, especialmente incluye los genes codificantes de proteínas y secuencias reguladoras. La identificación degenes es uno de los primeros y más importantes pasos para entender el genoma de una especie una vez ha sido secuenciado [11].

El ácido desoxirribonucleico (ADN) está compuesto por cuatro moléculas llamadas nucleótidos o bases nitrogenadas: adenina, timina, guanina y citosina [9]. Una molécula completa de ADN o, dicho de otro modo, una secuencia de ADN está compuesta por un alfabeto que contiene las letras de las cuatro bases nitrogenadas.

$$\begin{aligned} &\Sigma\{ATGC\} \\ \phi_i = (V_1, V_2, V_3, \ldots, V_n) \\ &V_i \in \Sigma \end{aligned} \tag{1}$$

Donde una cadena ϕ es una secuencia de ADN formada elementos del alfabeto Σ y puede definir las características de un organismo vivo, conteniendo toda la información genética en unidades de herencia llamadas genes. Los mecanismos o procesos de predicción de genes son aquellos que, dentro del área de la biología computacional, se utilizan para la identificación algorítmica de trozos de secuencias, usualmente ADN genómico [1], y que son biológicamente funcionales. Esto, especialmente incluye los genes codificantes de proteínas y secuencias reguladoras. La identificación de genes es uno de los primeros y más importantes pasos para entender el genoma de una especie una vez ha sido secuenciado [2].

Las uniones de empalme son puntos en una secuencia de ADN en la que se elimina ADN ‘’inútil” durante el proceso de creación de proteínas en organismos superiores. El problema planteado en este conjunto de datos es reconocer, dada una secuencia de ADN, los límites entre los exones (las partes de la secuencia de ADN retenidas después del corte y empalme) y los intrones (las partes de la secuencia de ADN que se cortan). Este problema consiste en dos subtareas: reconocimiento de límites de exón / intrón (denominados sitios EI) y reconocimiento de límites de intrón / exón (sitios IE). (En la comunidad biológica, los límites de IE se refieren a los ‘’aceptantes” mientras que los límites de EI se conocen

como ‘’donantes”) [7]. Ambas tareas son complicadas ya que no existe una secuencia estándar para reconocer intrones y exones, razón por la cual es interesante diseñar herramientas que nos ayuden a identificarlos y clasificarlos.

El número de proyectos de investigación sobre genomas actualmente vigentes aumenta a un ritmo acelerado, y proporcionar un catálogo de genes para estos nuevos genomas es un desafío clave. La obtención de un conjunto de genes bien caracterizados, es un requisito básico en los pasos iniciales de cualquier proceso de creación de un genoma. Los métodos de búsqueda de genes computacionales se pueden categorizar libremente como basados en la alineación y en la composición de secuencias o una combinación de ambos. Los métodos basados en la alineación de secuencias se pueden usar cuando se intenta predecir un gen que codifica una proteína para la cual existe un homólogo estrechamente relacionado, este es el enfoque en GeneWise [5] y PROCRUSTES [4].

Los algoritmos basados en composición de secuencias (también conocidos como métodos de búsqueda de genes) contienen un modelo probabilístico de estructura génica basado en señales biológicas (sitios de empalme y sitios de inicio / detención de traducción) y propiedades de composición de secuencias funcionales (exones como secuencias codificantes e intrones como secuencias intermedias entre exones e intrones). A diferencia de los métodos basados en la alineación, estos algoritmos se basan sólo en las propiedades intrínsecas de los genes para construir estructuras genéticas predichas. Genscan [10] y Geneid [3] son los dos ejemplos de este enfoque y pueden encontrar genes conocidos y genes nuevos siempre que los genes se ajusten al modelo probabilístico subyacente. Una cadena de ADN es una molécula caracterizada por cuatro bases nitrogenadas Adenina, Timina, Guanina y Citosina [12]. Para mejorar la representación de una cadena de ADN se utilizan secuencias que pueden ser transformadas a representación con valores numéricos o alfabéticos: A (adenina), T (timina), G (guanina) y C (citosina). [10]. Sin embargo, la representación de grandes cantidades de información como secuencias de ADN no hacen sencillo su análisis matemático, esto crea la necesidad de encontrar nuevas formas de representar la información.

En 1988 Lapedes [6] y su equipo de trabajo entrenaron una red neuronal para reconocer genes en secuencias de ADN, lograron una precisión del 91.2 % en las uniones de corte y empalme de intrón / exón y del 92.8 % en las uniones de empalme de exón / intrón. Lo que dio origen a plantear el uso de redes neuronales convolucionales para resolver este mismo problema de clasificación. Este trabajo consistió en buscar una nueva forma de representar secuencias de ADN para su análisis, como ya se ha hecho referencia, existen actualmente diferentes métodos para reconocer genes, pero estas representaciones complican su análisis. La propuesta que presentamos es generar imágenes a partir secuencias de ADN y someterlas a análisis por técnicas de aprendizaje profundo, en específico a redes neuronales convolucionales; utilizadas para la clasificación de imágenes. Actualmente no se ha encontrado un modelo matemático que resuelva el proceso de clasificación por redes neuronales, pero sus resultados llegan a ser tan altos que superan el 99 % en algunos casos [1].

II-A. Red Neuronal Convolucional (CNN)

En los últimos años, el campo del aprendizaje automático ha progresado enormemente al abordar problemas de clasificación, identificación y reconocimiento de patrones. En particular, se ha encontrado que un tipo de modelo llamado red neuronal convolucional CNN (Convolutional Neural Network) por sus siglas en inglés, que logra un rendimiento razonable en tareas de reconocimiento visual de hardware, igualando o superando el rendimiento humano en algunos dominios [11]. Una CNN es un algoritmo para el aprendizaje automático en el que un modelo aprende a realizar tareas de clasificación directamente a partir de imágenes, videos o sonidos. Las CNNs son especialmente útiles para localizar patrones en imágenes con el objetivo de reconocer objetos, caras y escenas. Aprenden directamente a partir de los datos de imágenes, utilizando patrones para clasificar las imágenes y eliminar la necesidad de una extracción manual de características.

Inception-v3 está diseñado para el desafío de Reconocimiento Visual, ésta es una tarea estándar en visión artificial, donde los modelos intentan clasificar imágenes completas en 1000 clases de ImageNet. TensorFlow es una herramienta para el aprendizaje automático. Si bien contiene una amplia gama de funcionalidades, TensorFlow está diseñado principalmente para modelos de redes neuronales profundas. Los modelos modernos de reconocimiento de imágenes tienen millones de parámetros; entrenarlos desde cero requiere una gran cantidad de datos de entrenamiento etiquetados y una gran cantidad de potencia de cálculo (cientos de horas de GPU o más). El aprendizaje de transferencia es una técnica que ataja mucho de esto tomando una pieza de un modelo que ya ha sido entrenado en una tarea relacionada y reutilizándola en un nuevo modelo, en la figura 1 se muestra un ejemplo de una CNN, los filtros se aplican a cada imagen de entrenamiento con diferentes resoluciones, y la salida de cada imagen convolucionada se usa como entrada para la capa siguiente [2]. Aunque no es igual de preciso en comparación a la capacitación del modelo completo, es sorprendentemente eficaz para muchas aplicaciones, funciona con cantidades moderadas de datos de capacitación (miles, no millones de imágenes etiquetadas) y se puede ejecutar en tan solo treinta minutos en una computadora portátil sin una GPU [11].

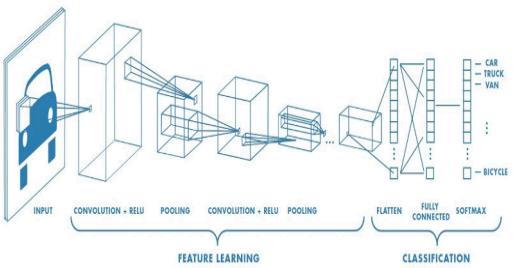

Figura 1. Ejemplo de una Red Neuronal Convolucional

III. Metodología

En esta sección se describe detalladamente como se generaron imágenes a partir de secuencias de ADN y su posterior uso en el entrenamiento de una red neuronal convolucional para clasificación de tres clases de secuencias. En esta etapa del proyecto consistió en convertir las secuencias de ADN a representaciones gráficas para entrenar una CNN. Un aspecto importante que se ha considerado este trabajo es que las CNN son utilizadas para el reconocimiento de patrones y clasificación de imágenes. Las secuencias de ADN de manera general son representadas por letras: A usada para la adenina, G para la guanina, C para la citosina y T para la timina, sin embargo, una CNN no está establecida para procesar información bajo este formato, por esta razón se diseñó una representación gráfica de las secuencias. El primer paso fue asignar un color en escala de grises a cada una de las letras como se muestra en el Cuadro I. Las escalas de grises va de 0 que representa negro, a 1 que represa el blanco, de tal manera que los colores intermedios resultantes son tonalidades de gris para mostrar un mejor contraste. Lo segundo fue hacer

Cuadro I
Representación por color de las bases nitrogenadas.

Base Nitrogenada	Valor de gris
A	0
C	0.3
G	0.7
T	1

que las secuencias pudieran ser representadas por una imagen específica a cada una. Para lograr esto se utilizó una matriz de dimensión 60 X 60, donde el valor 60 coincide con el número de bases nitrogenadas de todas las secuencias de la base de datos. Cada secuencia fue colocada en la primera fila y copiada en el resto de las filas hasta tener 60 en total, así el resultado final es una imagen con barras en la escala de grises como la que se muestra en la Figura 3, cada una de las imágenes obtenidas es específica para cada instancia de la base de datos como se observa en la Figura 2. En total se obtuvieron 3190 imágenes.

```
1  CCAGCTGCATCACAGGAGGCCAGCGAGCAGGTCTGTTCCAAGGGCCTTCGAGCCAGTCTG
2  CCAGCTGCATCACAGGAGGCCAGCGAGCAGGTCTGTTCCAAGGGCCTTCGAGCCAGTCTG
3  CCAGCTGCATCACAGGAGGCCAGCGAGCAGGTCTGTTCCAAGGGCCTTCGAGCCAGTCTG
.
.
.
60 CCAGCTGCATCACAGGAGGCCAGCGAGCAGGTCTGTTCCAAGGGCCTTCGAGCCAGTCTG
```

Figura 2. Secuencia ADN a codificar

III-A. Utilización de CNN a secuencias de ADN

En esta subsección se describe cómo se entrenó una CNN con las imágenes representativas de cada secuencia. Se utilizó

Figura 3. Imagen asociada

una CNN InceptionV3 a la que se le aplicó la transferencia de aprendizaje profundo para categorizar el reconocimiento de tres clases se secuencias de ADN: reconocimiento de límites de exón / intrón (denominados sitios EI), reconocimiento de límites de intrón / exón (sitios IE) y reconocimiento de ninguno de los dos anteriores (N).
Una vez que se logró representar a las secuencias de ADN como imágenes se utilizó una CNN y con la librería de software TensorFlow se construyó un modelo de clasificación basado en una red neuronal convolucional pre-entrenada. Se utilizaron CNNs InceptionV3 a las que se les aplicó la transferencia de aprendizaje profundo para categorizar el reconocimiento de una base de datos con cuatro clases de secuencias de ADN: virus de Hepatitis C tipo 1, 2, 3 y 6 y el reconocimiento de otra base de datos con tres clases de límites de exón / intrón (denominados sitios EI), reconocimiento de límites de intrón / exón (sitios IE) y reconocimiento de ninguno de los dos anteriores (N). Para ajustar el modelo a nuestro problema se entrenaron las últimas capas de las redes con instancias obtenidas de las bases de datos, ambas redes fueron entrenadas en 4000 pasos.

Primero se entrenó la CNN para hacer la clasificación de los 4 tipos de virus de Hepatitis, posteriormente se entrenó una CNN con solamente 2 clases: EI e IE y por último se entrenó otra CNN con todas las clases de la base de datos: EI, IE y N para comparar los resultados de las ultimas dos neuronas.

IV. Resultados

Los resultados de clasificación para la CNN entrenada con la base de datos de los cuatro tipos de virus de Hepatitis C muestran una precisión de evaluación 95 % con 145 imágenes probadas y al terminar el paso (k) 4000 la precisión de entrenamiento fue del 94.5 % y la precisión de validación del 95 % como se observa en la figura 4. El comportamiento decreciente de la entropia durante el entrenamiento, se aprecia en la figura 5 .
Al usar una CNN con las clases EI e IE se obtiene una precisión de evaluación del 80.8 % con 177 imágenes de prueba y al terminar el paso (k) 4000 la precisión de entrenamiento es del 82 % y la precisión de validación del 75 %. En la Figura 6 se muestra como la exactitud de entrenamiento (naranja) y validación (azul) va cambiando en cada paso y en la Figura 7 se muestra como la entropía disminuye con el incremento de los pasos durante el entrenamiento. Por otro lado, los resultados de la segunda CNN donde se utilizaron las tres clases de la base de datos muestran una precisión de evaluación

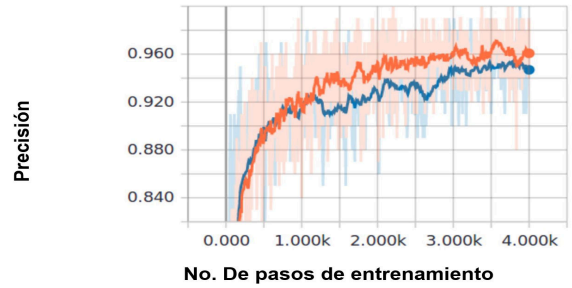

Figura 4. CNN con las clases de virus de Hepatitis C tipo 1, 2, 3 y 6. Naranja: precisión de entrenamiento. Azul: precisión de validación después de 4000 pasos (k).

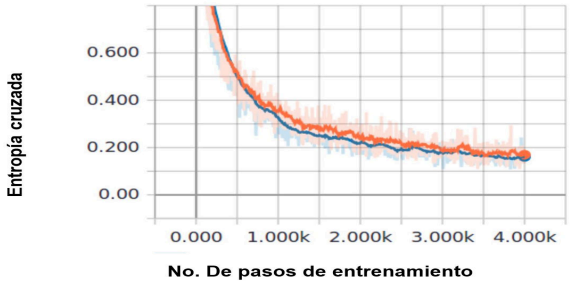

Figura 5. Entropía cruzada de la CNN con las clases de virus de Hepatitis C tipo 1, 2, 3 y 6 después de 4000 pasos (k). Trazo superior: entrenamiento. Trazo inferio: validación.

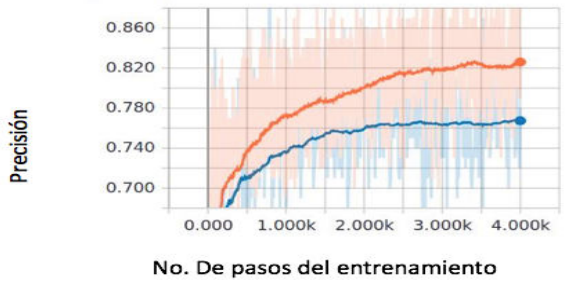

Figura 6. CNN con las clases IE y EI. Naranja: precisión de entrenamiento. Azul: precisión de validación después de 4000 pasos (k).

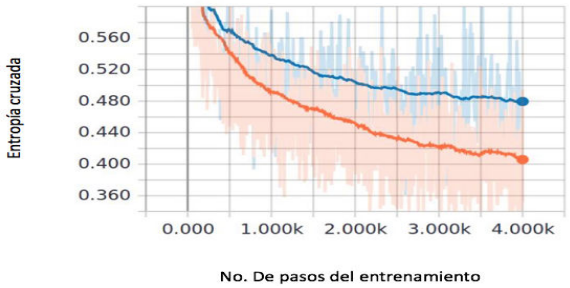

Figura 7. Entropía cruzada de la CNN con las clases IE y EI después de 4000 pasos (k). Naranja: entrenamiento. Azul: validación.

de 57.5 % con 301 imágenes y al terminar el paso 4000 la precisión de entrenamiento 69 % y la precisión de validación con un 56 % como se aprecia en la Figura 8. En la Figura 9 se muestra los cambios de la entropía en cada etapa del entrenamiento.

V. Conclusiones

Los resultados obtenidos de la CNN entrenada con la base de datos de virus de Hepatitis C sugieren que la metodología de aprendizaje automático empleada en este trabajo es adecuada para la clasificación de las imágenes generadas a partir de las secuencias de ADN, mostrando importantes y altos porcentajes de precisión de evaluación, precisión de entrenamiento y la precisión de validación. Estos resultados nos llevaron a realizar los siguientes experimentos para el reconocimento de exones e introns en la siguiente base de datos. Para este caso las CNN muestran que los porcentajes de precisión de validación son menores en comparación a los de una red neuronal tomando como referencia el trabajo de Lapedes [6]. La importancia del trabajo es que se presentan resultados favorables para seguir explorando el uso de las redes neuronales convolucionales utilizando la representación de las secuencias de ADN como imágenes, un método de codificación sencillo y práctico.

En este trabajo se ha logrado realizar clasificación de secuencias de ADN usando una CNN y los resultados demuestran que las CNN son capaces de realizar esta clasificación hasta con un 80.8 % de precisión de evaluación para el experimento con las clases IE e EI y el 57.5 % para el experimento con las clases IE, EI y N. Resultados similares se pueden observar en la precisión de entrenamiento y validación de las Figuras 6 y 8. En el caso de los cuatro tipos de hepatitis se logran resultados de hasta 94.5 % de precisión de evaluación .

La diferencia entre los resultados obtenidos para los experimentos con dos y tres clases se puede justificar que al incrementar el número de clases se incrementa la entropía Figuras 7 y 9. La entropía cruzada es una métrica que puede utilizarse para reflejar la precisión de los pronósticos probabilísticos y está estrechamente vinculada con la estimación por máxima verosimilitud. La entropía cruzada es una función que permite evaluar el resultado de la clasificación en vez de utilizar la métrica del error cuadrático medio, el valor de la entropia cruzada permite evaluar el progreso del proceso de aprendizaje de la información [1].

Por otro lado, se habla de que la transferencia de aprendizaje es buena cuando se disponen de pocas imágenes para entrenar la red y que permite llegar a resultados aceptables en la mayoría de los casos, sin embargo, todavía es posible mejorar aún más la precisión de validación y entrenamiento y disminuir la entropía si se entrena una red neuronal desde cero, es decir se debe contar con una base de datos de millones de instancias y un equipo de cómputo con GPU para entrenar esta red pero seguramente ofrecerá mejores resultados que la CNN preentrenada que utilizamos para este trabajo.

En conclusión, se puede afirmar que una red neuronal convolucional del modelo InceptionV3 es capaz de clasificar secuencias de ADN si la secuencia es procesada y transformada a una imagen, sin embargo, los porcentajes de exactitud se pueden mejorar si se entrena una CNN con una base de

Figura 8. CNN con las clases IE, EI y N. Naranja: precisión de entrenamiento. Azul: precisión de validación después de 4000 pasos (k).

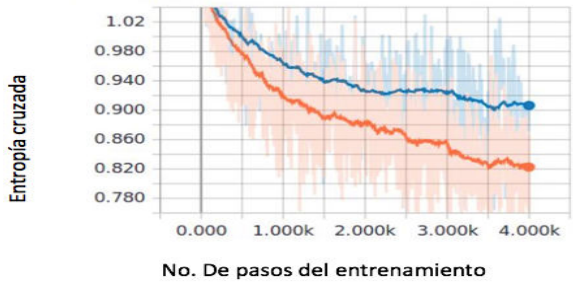

Figura 9. Entropía cruzada de la CNN con las clases IE, EI y N después de 4000 pasos (k). Naranja: entrenamiento. Azul: validación.

secuencias más grande.

Reconocimientos

Los autores agradecen al Consejo Nacional de Ciencia y Tecnología (CONACyT) de México , a la Benemérita Universidad Autónoma de Puebla la cual a través de la Facultad de Ciencias de la Computación han brindado el apoyo necesario para la realización y presentación del presente trabajo.

Referencias

[1] How to retrain an image classifier for new categories. https://www.tensorflow.org/tutorials/image_retraining. Accessed: 2018-05-28.

[2] Mathworks (2018). deep learning. https://la.mathworks.com/solutions/deep-learning/convolutional-neural-network.html. Accessed: 2018-05-27.

[3] Jonas S Almeida and Susana Vinga. Universal sequence map (usm) of arbitrary discrete sequences. *BMC bioinformatics*, 3(1):6, 2002.

[4] Chris Burge and Samuel Karlin. Prediction of complete gene structures in human genomic dna1. *Journal of molecular biology*, 268(1):78–94, 1997.

[5] Mikhail S Gelfand, Andrey A Mironov, and Pavel A Pevzner. Gene recognition via spliced sequence alignment. *Proceedings of the National Academy of Sciences*, 93(17):9061–9066, 1996.

[6] A Lapedes, Christopher Barnes, Christian Burks, R Farber, and K Sirotkin. Application of neural networks and other machine learning algorithms to dna sequence analysis. Technical report, Los Alamos National Lab., NM (USA), 1988.

[7] Michiel O Noordewier, Geoffrey G Towell, and Jude W Shavlik. Training knowledge-based neural networks to recognize genes in dna sequences. In *Advances in neural information processing systems*, pages 530–536, 1991.

[8] Christos A Ouzounis. Rise and demise of bioinformatics? promise and progress. *PLoS computational biology*, 8(4):e1002487, 2012.

[9] Arturo Panduro. *BiologÃa molecular en la clÃnica*. McGraw-Hill Interamericana, 2009.

[10] Genís Parra, Enrique Blanco, and Roderic Guigó. Geneid in drosophila. *Genome research*, 10(4):511–515, 2000.

[11] SL Salzberg, DB Searls, and S Kasif. Computational gene prediction using neural networks and similarity search. *Computational Methods in Molecular Biology*, 32:109, 1998.

[12] Zhu-Jin Zhang. Dv-curve: a novel intuitive tool for visualizing and analyzing dna sequences. *Bioinformatics*, 25(9):1112–1117, 2009.