Agrupamiento no supervisado y redes convolucionales para el aprendizaje de estructuras en bioinformática

I.L. Fucksmann, L.A. Bugnon, and D.H. Milone

Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH-UNL/CONICET, Santa Fe, Argentina
{ifucksmann, lbugnon, dmilone}@sinc.unl.edu.ar

Resumen La predicción computacional del plegamiento de secuencias es un proceso fundamental para la determinación de los ácidos ribonucleicos (ARN), porque a partir de estas estructuras es posible estudiar las funciones que cumplen dichas secuencias. La inteligencia artificial, en particular el aprendizaje profundo, ha comenzado a ser utilizada para la predicción de estas estructuras a partir de la secuencia. Sin embargo, este problema se dificulta conforme se aumenta la longitud de la cadena, generando estructuras muy diferentes entre sí. Este trabajo busca mejorar el modelado y aprendizaje de estas estructuras con la incorporación de técnicas de agrupamiento de secuencias. Hemos encontrado una relación entre la estructura formada y la secuencia original mediante una medida basada en información mutua. Nuestros resultados son prometedores, alcanzando un valor de información mutua normalizado de 0.79 entre los agrupamientos generados entre las secuencias de entrada y los generados por la estructura a predecir. A partir de estos agrupamientos, se entrenaron modelos de predicción basados en redes convolucionales, independientes para cada grupo de secuencias, y se ensamblaron para obtener la predicción final, obteniendo en promedio un 5.20 % de mejora en F_1, comparando contra el modelo de referencia.

Palabras clave: Redes neuronales · Aprendizaje no supervisado · Bioinformática · Predicción de estructuras.

1. Introducción

La predicción de estructuras secundarias de ácidos ribonucleicos (ARN) es de gran importancia en la investigación biología molecular, debido a la amplia gama de roles que desempeñan estas moléculas en la célula. Las estructuras secundarias del ARN son fundamentales para determinar su función biológica, como su capacidad para catalizar reacciones químicas y para regular la expresión de los genes. Por lo tanto, la capacidad de predecir estructuras secundarias de ARN con alta precisión es una herramienta importante en la investigación y desarrollo tecnológico.

En los últimos años hubo un incremento en el uso de metodologías basadas en aprendizaje automático (ML, del inglés *machine learning*) que poseen rendimientos comparables a los métodos clásicos, basados en programación dinámica y restricciones termodinámicas para la predicción de estructuras a partir de secuencias de nucleótidos. En [1] se presenta una extensa validación experimental y un análisis detallado del desempeño de métodos clásicos y otros basados en ML, concluyendo que a pesar de las mejoras obtenidas, los nuevos modelos siguen sin obtener un desempeño significativamente superior. Queda claro que predecir computacionalmente la estructura de los ARN sigue siendo un gran desafío en bioinformática, y ha demostrado ser más difícil que la predicción de la estructura de las proteínas, debido entre otras razones a la riqueza estructural y la limitada cantidad secuencias con estructuras validadas que se disponen para el entrenamiento [2,3]. Es por esto que en este trabajo de investigación se plantea como objetivo mejorar la predicción de dichas estructuras mediante la combinación de aprendizaje no supervisado y profundo.

2. Métodos

La metodología que sugerimos comienza con una etapa de preprocesamiento de las secuencias de nucleótidos. Estas secuencias están formadas por 4 posibles elementos, que se identifican con los símbolos “A”, “C”, “G” y “U”. Para codificarlas cada elemento de la secuencia se convierte a un vector de dimensión 4, colocando un 1 en cada posición de acuerdo al símbolo (codificación más directa que se conoce en inglés como one-hot). Por otro lado, la estructura secundaria a predecir se codifica como matriz de conexiones binarias, cuyas filas y columnas representan la posición de cada nucleótido dentro de la secuencia. De esta manera se busca predecir una matriz cuyos valores iguales a uno representan las conexiones que forman la estructura del ARN.

Estas secuencias ya codificadas entran a una red neuronal convolucional (Figura 1), que dio buenos resultados en [4]. La primer etapa de esta red es una secuencia de capas convolucionales en una dimensión (1D). El uso de estas capas para cada dimensión de la codificación, convolucionando a lo largo de la secuencia, permite modelar adecuadamente relaciones de corto alcance. El número de capas y filtros utilizados fueron seleccionados previamente en base a la longitud de la secuencia y la cantidad de dimensiones de entrada, y su ajuste fino se hizo de manera empírica sobre otro conjunto de datos independiente [4]. A partir de la secuencia de largo L codificada en one-hot, las capas convolucionales 1D logran una primera extracción automática de características de bajo nivel, identificando patrones a de vecindad de cada posición de la secuencia.

La segunda etapa del modelo codificará relaciones entre elementos de la secuencia, incluyendo las conexiones distantes. Para esto se pasa de una codificación de $M \times L$, donde M es la dimensión del vector de características obtenido en la etapa anterior para cada nucleótido, a dos codificaciones de $C \times L$, que finalmente se llevará a la codificación $L \times L$ que es igual a la matriz de conexiones utilizada como referencia (etiqueta). Es importante destacar que dado que L

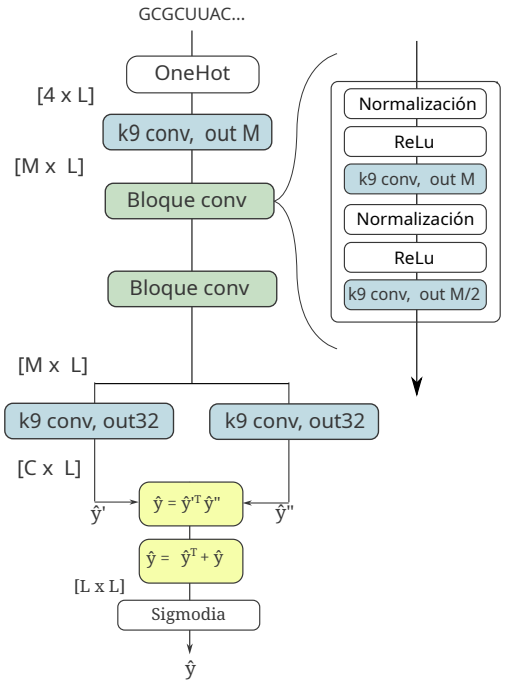

Figura 1. Esquema una red neuronal entrenable de extremo a extremo para predecir uniones entre elementos de la secuencia de entrada.

puede ser muy grande, la codificación que involucra L^2 se realiza solo al final de todo el proceso. Con una simple operación matricial entre las representaciones $C \times L$ se obtiene la primera matriz $\hat{y}$, que luego es sumada con su transpuesta para forzar la simetría (un requisito en las matrices de conexiones). Finalmente, una función sigmoidea transforma las posibles activaciones a un intervalo $[0, 1]$.

Las secuencias de ARN han sido tradicionalmente agrupadas en familias de acuerdo a sus estructuras y funciones biológicas [5]. Esta podría ser una información muy importante para el modelo de predicción, tanto para incorporarla a la entrada como para entrenar modelos de predicción independientes para cada familia de ARN. Sin embargo, esta información no está disponible cuando el modelo tiene que predecir la estructura de una secuencia desconocida, ya que la relación entre las secuencias y la familia correspondiente no es directa. Es decir, una vez que el modelo ha sido entrenado y se requiere predecir la estructura para una nueva secuencia, se supone que esa secuencia es desconocida y por lo tanto no tiene una familia asociada. Pero si las secuencias pudieran ser agrupadas de forma no supervisada, y esa agrupación tuviera un alto grado de correspondencia con la agrupación por familias o por similaridad de estructura, entonces se podrían entrenar modelos independientes para cada grupo y luego ensamblar las predicciones. Un esquema general de este método se describe en la Figura 2, en la que se puede observar que se generan grupos (representados mediante una elipse amarilla) utilizando las similaridades en secuencia. A continuación, se entrena un modelo individual por grupo. Una vez entrenado, a los datos de prueba se les asigna un grupo (caja de color rojo) según un criterio de similaridad con los grupos aprendidos desde los datos de entrenamiento, y luego cada secuencia sera predicha por su modelo correspondiente.

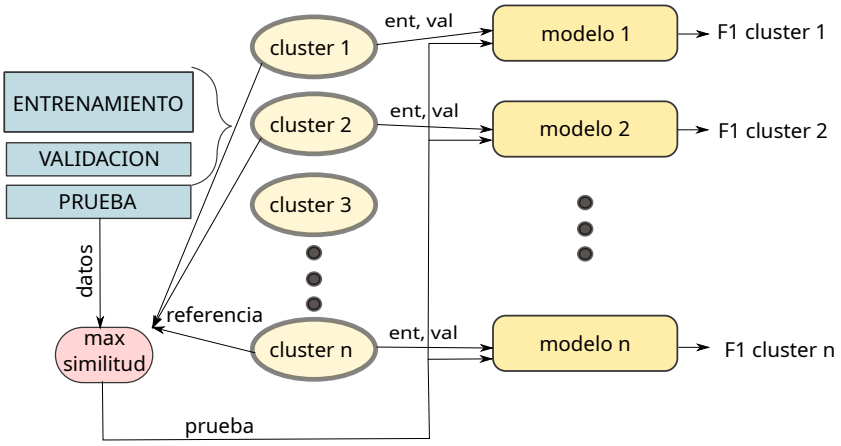

Figura 2. Diagrama del modelo de predicción basado en entrenamiento no supervisado y ensamble de modelos por grupo.

Para realizar el agrupamiento se puede utilizar la distancia entre secuencias, haciendo primero una alineación de a pares de nucleótidos[1] y luego contando las coincidiencias y normalizando por la longitud. De esta forma se puede generar una matriz de distancias entre de todas las secuencias, a partir de la cual se utiliza el método de agrupamiento espectral (SC, del inglés *spectral clustering*)[6]. Se optó por este algoritmo debido a su capacidad para trabajar directamente con matrices de distancias y que no requiere la definición y adaptación iterativa de centroides basados en vectores de características (que es este caso no se poseen). El mismo SC se puede aplicar de forma similar para realizar un agrupamiento en función de la distancia entre estructuras[2].

Sobre los datos de entrenamiento, se generaron agrupamientos para diferente número de grupos, $k \in [10, 30]$, particionando tanto por distancia entre secuencias como por distancia de estructuras. Luego se utilizó la información mutua normalizada (NMI, por sus siglas en inglés) para medir el nivel de correspondencia entre todos los agrupamientos para cada k,

$$NMI(Y, C) = \frac{2 \;\; I(Y; C)}{[H(Y) + H(C)]}, \tag{1}$$

donde Y son los grupos correspondientes a las estructuras, C los grupos según la distancia de secuencias, $H(\cdot)$ la entropía e $I(\cdot)$ la información mutua entre ambas particiones. De esta manera, el k óptimo queda determinado por el máximo NMI entre todas las particiones comparadas.

Una vez obtenido el agrupamiento óptimo a partir de los datos de entrenamiento, se dividen los datos por grupo para entrenar cada modelo (k predictores). Cada modelo utiliza la misma arquitectura que el modelo base (Figura 1). En el momento de la evaluación o la predicción de una nueva secuencia, es necesario identificar a qué grupo pertenece. Para esto se evaluaron dos metodologías:

- Selección por medoides: se encuentra el medoide de cada grupo, eligiendo la secuencia que tiene la menor distancia promedio a todas las de su grupo. Luego, dada una secuencia desconocida se mide la distancia a cada medoide y se asigna el grupo del medoide más cercano.
- Selección por la menor distancia: se calcula la distancia a cada una de las secuencias en el conjunto de datos de entrenamiento (todos los grupos) y se asigna a la secuencia de prueba el mismo grupo que la secuencia de entrenamiento más cercana. Si bien esta variante puede ser algo más costosa computacionalmente, al evitar el uso del medoide se asigna según una mejor representación de la distribución de los datos.

3. Resultados

En el presente estudio se empleó el conjunto de datas "Archive II"[7], el cual consta de 3974 secuencias de diversas familias y longitudes. Se realizo validación

[1] https://biopython.org/docs/1.75/api/Bio.pairwise2.html
[2] https://www.tbi.univie.ac.at/RNA/RNAdistance.1.html

cruzada, en donde el conjunto de datos fue dividido en grupos de entrenamiento, validación y prueba con un porcentaje del 80 %, 10 %, 10 % respectivamente. Para la fase de agrupamiento se precalcularon las dos matrices que contienen la distancia entre cada secuencia y entre cada estructura del conjunto de datos completo. Sobre estas matrices se utilizó el algoritmo de agrupamiento propuesto, siempre respetando las particiones de entrenamiento y prueba de forma de que ninguna información del conjunto de prueba sea utilizada tanto para la etapa no supervisada de agrupamiento como para el entrenamiento supervisado de los modelos profundos. Para llevar a cabo cada una de estas etapas, se emplearon las bibliotecas scikit-learn y PyTorch, respectivamente.

3.1. Agrupamientos según distancias de estructuras y de secuencias

Para comprobar la existencia de una coodependencia entre las estructuras y secuencias de nucleótidos, se realizó una serie de agrupamientos variando k usando ambas distancias, por secuencia y por estructura. Entre todas las particiones obtenidas se calculó la NMI, obteniendo 0.79 como valor máximo para un agrupamiento por distancia de estructura con $k = 22$ y un agrupamiento por distancia de secuencias con $k = 15$. En la Figura 3 se representa un mapa de calor que ilustra los valores de NMI de todas las combinaciones de k analizados. Los valores más altos de NMI se corresponden con colores más oscuros en la figura.

Una vez obtenido el k óptimo para cada agrupación se realizó la matriz de contingencia para tener una referencia más gráfica de dicha relación. En la Figura 4 las columnas corresponden a los grupos según la distancia de estructura y las filas con aquellos que se relacionan con las distancias de secuencia. Se puede observar como la distribución de los grupos se acomoda de manera que se corresponden uno o dos grupos con información estructural con uno que contiene información sobre la secuencia. Esto nos indica que para un grupo de los obtenidos por distancia de secuencias existen solo una o dos estructuras que lo definen bien. Cabe resaltar que muchos de estos clusters contienen solo datos que pertenecen a la misma familia de secuencias.

3.2. Prueba del modelo de predicción de estructuras

Se procedió a entrenar el modelo base utilizando los datos de entrenamiento, y se utilizó el criterio de finalización temprana con los datos de validación, con el fin de asegurar que el proceso de entrenamiento está avanzando correctamente y no haya sobre-ajuste. Se utilizó una velocidad de aprendizaje de 1E-3, un número máximo de 150 épocas y una paciencia de 20 épocas para la finalización temprana, ya que estos parámetros demostraron buen rendimiento en un conjunto de datos independiente [4]. El entrenamiento del modelo base llevo aproximadamente 87 min, mientras que el entrenamiento del modelo propuesto tomo ∼13 min para cada grupo, sumando un total de ∼78 min para todo el modelo[3]. En el proceso de agrupamiento se generaron 15 grupos utilizando los

[3] Experimentos realizados con una GPU Nvidia RX5000 con 24 GB de memoria

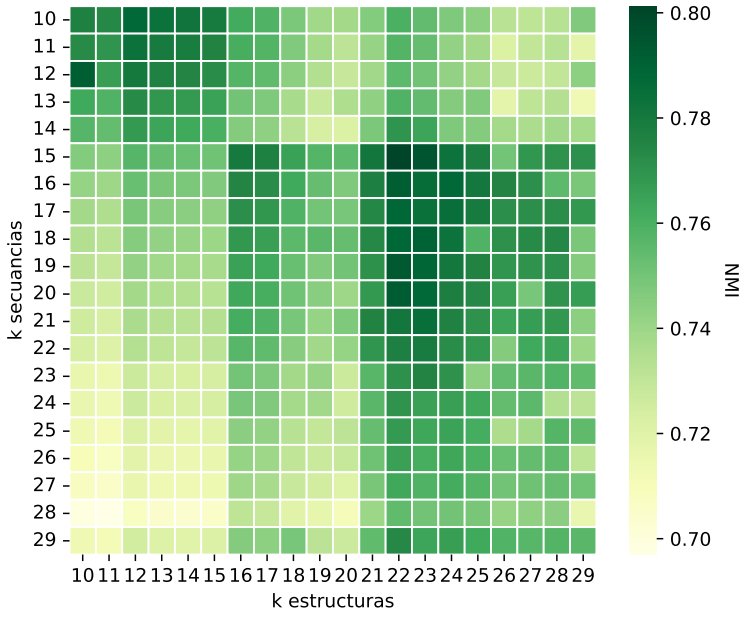

Figura 3. Valores de NMI para diferentes k, comparando los grupos formados con distancia entre secuencias (filas) y los formados por distancia de estructuras (columnas). Se puede observar que el máximo NMI se encuentra en torno a $k = 15$ según la distancia de secuencias.

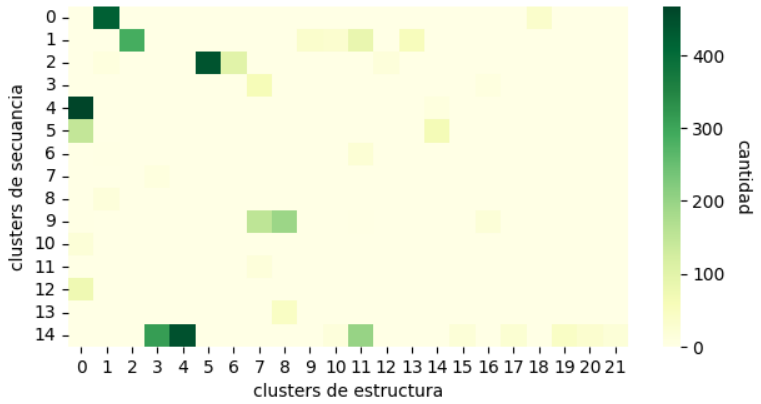

Figura 4. Matriz de contingencia para los grupos formados por distancia de secuencia (en las filas, $k = 15$) y por distancia de estructura (en las columnas, $k = 22$).

datos de entrenamiento y validación. Posteriormente, se entrenaron modelos por grupos correspondientes a los 6 clusters con más secuencias y que se dispersan en menos grupos de estructura. Luego se realizaron las pruebas en la correspondiente partición de test. En cada caso se midió el F_1 de las predicciones en relación a la referencia, es decir, la media armónica entre la sensibilidad y la precisión de las conexiones predichas para cada par de elementos de la secuencia.

Se comparó el rendimiento usando ambos métodos de selección de agrupamiento, por medoides y por el de menor distancia. En el caso de la selección del grupo por medoides los resultados medios obtenidos por el modelo propuesto y el modelo base fueron de $F_1 = 0{,}590$ y $F_1 = 0{,}769$, respectivamente. Esto muestra que el método por medoides no presenta un rendimiento aceptable, posiblemente porque los grupos tengan alta diversidad y no estén bien representados por el medoide. Para el caso de la selección por la mínima distancia los resultados fueron mucho más prometedores. En la Tabla 1 se pueden ver los valores de F_1 de prueba del modelo base y el modelo entrenado para cada grupo utilizando los mismos datos de prueba para cada comparación. Podemos observar que se obtienen mejores resultados en los modelos entrenados por cada grupo que utilizando el modelo base entrenado con el conjunto de de entrenamiento completo. En promedio se observa que el modelo propuesto logra un F_1 que supera al modelo base en un 5.20 % haciendo uso de técnicas no supervisadas en combinación con clasificadores profundos supervisados.

4. Conclusiones

En este trabajo se presenta un nuevo modelo de aprendizaje automático para la predicción de estructuras secundarias de ARN. Se propone realizar un preproceso con un agrupamiento basado en las características de las secuencias y

Tabla 1. Resultados en términos de F_1 para el modelo base contra el propuesto usando la selección por menor distancia.

	Modelo base	Propuesto
Cluster 1	0.580	0.587
Cluster 2	0.890	0.956
Cluster 3	0.779	0.854
Cluster 4	0.921	0.951
Cluster 5	0.877	0.958
Cluster 6	0.567	0.623
Promedio	0.769	0.821

sus estructuras conocidas, para dividir los datos en grupos y entrenar modelos independientes. Cada uno de estos modelos está basado en una red neuronal convolucional profunda que lleva a cabo la predicción. Los resultados obtenidos muestran que existe una relación entre las estructuras y secuencias, la cual se puede aprovechar mediante las técnicas de agrupamiento para separar el espacio de datos y predicciones en diferentes clasificadores y obtener mejores predicciones. Las secuencias de test se asignan a los grupos utilizando la distancia entre secuencias. Los experimentos muestran que esta estrategia de separación no supervisada y el posterior ensamble de modelos mejora significativamente los resultados de predicción. Como trabajos futuros se propone seguir investigando en cómo utilizar los grupos que poseen un número reducido de secuencias. En este sentido, una de las propuestas que tenemos es realizar transferencia de aprendizaje desde los modelos entrenados con más secuencias.

Referencias

1. L. A. Bugnon, A. A. Edera, S. Prochetto, M. Gerard, J. Raad, E. Fenoy, M. Rubiolo, U. Chorostecki, T. Gabaldón, F. Ariel, L. E. D. Persia, D. H. Milone, and G. Stegmayer, "Secondary structure prediction of long noncoding RNA: review and experimental comparison of existing approaches," *Briefings in Bioinformatics*, vol. 23, June 2022.
2. A. W. Senior and et al., "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, pp. 706–710, Jan. 2020.
3. H. Kamisetty, B. Ghosh, C. J. Langmead, and C. Bailey-Kellogg, "Learning sequence determinants of protein:protein interaction specificity with sparse graphical models," *Journal of Computational Biology*, vol. 22, pp. 474–486, June 2015.
4. L. Bugnon, L. D. Persia, A. M. Gerard, J. Edera, S. Raad, E. Prochetto, G. S. Fenoy, and D. Milone., "Improving the folding prediction of rna with deep learning," in *A2B2C conference*, 2022.
5. I. Kalvari, E. P. Nawrocki, N. Ontiveros-Palacios, J. Argasinska, K. Lamkiewicz, M. Marz, S. Griffiths-Jones, C. Toffano-Nioche, D. Gautheret, Z. Weinberg, E. Rivas, S. R. Eddy, R. D. Finn, A. Bateman, and A. I. Petrov, "Rfam 14: expanded coverage of metagenomic, viral and microRNA families," *Nucleic Acids Research*, vol. 49, pp. D192–D200, 11 2020.

6. A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems* (T. Dietterich, S. Becker, and Z. Ghahramani, eds.), vol. 14, MIT Press, 2001.
7. M. Sloma and D. Mathews, "Exact calculation of loop formation probability identifies folding motifs in rna secondary structures," *RNA*, vol. 22, 10 2016.