AGRUPAMIENTO NO SUPERVISADO Y REDES CONVOLUCIONALES PARA EL APRENDIZAJE DE ESTRUCTURAS EN BIOINFORMÁTICA

Fucksmann, Ignacio L.

Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional sinc(i)- UNL - CONICET
Director/a: Milone, Diego H.
Codirector/a: Bugnon, Leandro A.

Área: Ingenieria

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

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 el trabajo de Bugnon, Edera, et al. 2022 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 como se

Título del proyecto: Nuevas estrategias de aprendizaje profundo para predicción de estructuras secundarias en secuencias largas de RNA no codificantes que permitan el desarrollo de biotecnologías de bajo impacto ambiental,"
Instrumento: CONICET-AWS 2022012157000954
Año convocatoria: 2021
Organismo financiador: CONICET-Amazon
Director/a: Bugnon, L.

explica en Kamisetty et al. 2015; Senior y al 2020. 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.

OBJETIVOS

Los objetivos son:

- Implementar y evaluar la arquitectura neuronal profunda base para aprender las interacciones presentes en secuencias biológicas.
- Implementar y evaluar un método de agrupamiento no supervisado para mejorar la eficacia del modelo propuesto.
- Validar los resultados comparando el modelo propuesto con el base.

METODOLOGÍA

El método 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 numéricamente cada elemento de la secuencia se convierte a un vector de dimensión 4, colocando un 1 en cada posición de acuerdo al símbolo (codificación que se conoce en inglés como *one-hot encoding*). 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. La primera 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 (Bugnon, Persia, et al. 2022). 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 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 $M\ x\ L$, donde M es la dimensión del vector de características para cada nucleótido obtenido en la etapa anterior, a dos codificaciones de $C\ x\ L$, que finalmente se llevará a la codificación $L\ x\ L$ que es igual a la matriz de conexiones. Es importante destacar que dado que L 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 $C\ x\ L$ se obtiene la primera matriz, 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 según Kalvari et al. 2020. Esta podría ser una información muy importante para el modelo de predicción, tanto para incorporar a la entrada como para entrenar modelos de predicción independientes para cada familia de ARN, ya que la diferencia estructural

entre familias puede ser muy grande. 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.

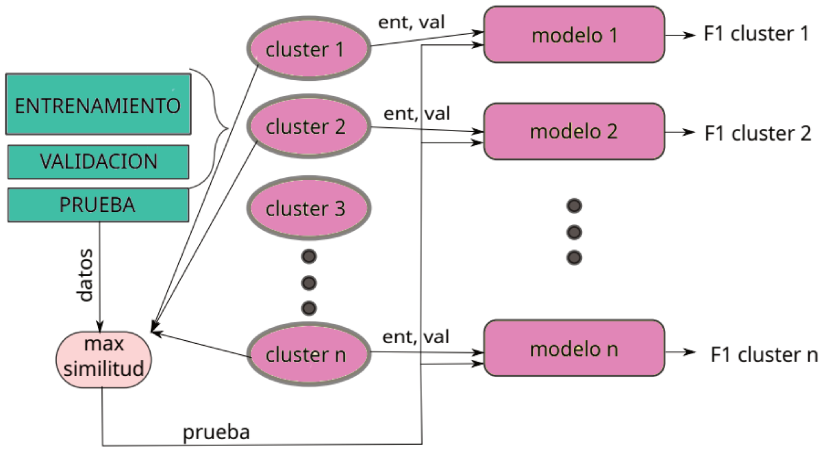

Figura 1: Diagrama del modelo de predicción basado en entrenamiento no supervisado y entrenamiento de modelos por grupo.

Un esquema general de este método se describe en la Figura 1, en la que se puede observar que se generan grupos (representados mediante una elipse rosada) utilizando las similaridades entre las secuencias. A continuación, se entrena un modelo convolucional como el antes descripto por grupo. Una vez entrenado, a los datos de prueba se les asigna un grupo (caja de color rojo claro) según un criterio de similaridad con los grupos aprendidos desde los datos de entrenamiento, y luego la estructura de cada secuencia sera predicha por el modelo correspondiente. Para medir la similaridad de una secuencia de test al grupo de entrenamiento se proponen dos alternativas: 1) midiendo la distancia al medoide de cada cluster, definido como la secuencia que tiene la menor distancia total a todas las demás secuencias de su cluster, y 2) usando como medida la menor distancia a todas las secuencias del cluster.

RESULTADOS Y CONCLUSIONES

Los resultados obtenidos muestran que existe una relación entre estructuras y secuencias, la cual se puede aprovechar mediante las técnicas de agrupamiento para separar el espacio de datos y predicciones en diferentes clasificadores. Las secuencias de test se asignan a los grupos utilizando la distancia entre secuencias.

Nº cluster	Modelo base	Modelo 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

Tabla 1. Resultados en términos de F1 para el modelo base contra el propuesto usando la selección por menor distancia.

Se comparó el rendimiento usando ambos métodos de selección de agrupamiento. En el caso de la selección del grupo por medoides los resultados medios obtenidos por el modelo propuesto y el modelo base fueron de $F1 = 0.590 \pm 0.14$ y $F1 = 0.769 \pm 0.10$, 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, en la Tabla 1 se pueden ver los valores de F1 de prueba del modelo base y el modelo entrenado para cada grupo utilizando los mismos datos de prueba. Podemos observar que se obtienen mejores resultados en los modelos entrenados por cada grupo que utilizando el modelo base entrenado con el conjunto de entrenamiento completo.

En promedio se observa que el modelo propuesto logra un F1 que supera al modelo base en un 5.20%, gracias al uso de técnicas no supervisadas en combinación con clasificadores profundos supervisados. Estos resultados confirman que esta estrategia de separación no supervisada y el posterior ensamble de modelos mejora significativamente los resultados de predicción.

BIBLIOGRAFÍA

Bugnon L. A., Edera A. A., Prochetto S., Gerard M., Raad J., Fenoy E., Rubiolo M., Chorostecki U., T. Gabaldón, Ariel F., Persia L. E. D., Milone D. H. y Stegmayer G.. 2022. “Secondary structure prediction of long noncoding RNA: review and experimental comparison of existing approaches”. *Briefings in Bioinformatics* 23(4). doi: 10.1093/bib/bbac205.

Bugnon L. A., Persia L. D., Gerard A. M., Edera J., Raad S., Prochetto E., Fenoy G. S., y Milone D. H. 2022. “Improving the folding prediction of RNA with deep learning”. en *A2B2C conference*.

Hetunandan K., Ghosh B., Langmead C. J., y Bailey-Kellogg C.. 2015. “Learning Sequence Determinants of Protein:Protein Interaction Specificity with Sparse Graphical Models”. *Journal of Computational Biology* 22(6):474–86. doi: 10.1089/cmb.2014.0289.

Kalvari I., Nawrocki E. P. , Ontiveros-Palacios N. , Argasinska J., Lamkiewicz K., Marz M., Griffiths-Jones S., Toffano-Nioche C., Gautheret D., Weinberg Z., Rivas E., Eddy S. R., Finn R. D., Bateman A., and Petrov A. I.. 2020. “Rfam 14: expanded coverage of metagenomic, viral and microRNA families”. *Nucleic Acids Research* 49(D1):D192–200. doi: 10.1093/nar/gkaa1047.

Senior A. W., y et al. 2020. “Improved protein structure prediction using potentials from deep learning”. *Nature* 577(7792):706–10. doi: 10.1038/s41586-019-1923-7