AUMENTACIÓN INTELIGENTE DE DATOS PARA MODELOS DE EXTREMO A EXTREMO Y DATOS SECUENCIALES EN BIOINFORMÁTICA

Matilde Garelik

Instituto de investigación en señales, sistemas e inteligencia computacional, sinc(i), FICH-UNL, CONICET
Director: Leandro Bugnon
Co-Director: Diego Milone

Área: Ingeniería

Palabras claves: Inteligencia artificial, Bioinformática, Secuencias RNA

INTRODUCCIÓN

El aprendizaje profundo ha potenciado la inteligencia artificial, generando además una gran transformación en diversos campos de aplicación. Sin embargo, quedan aún muchos desafíos para su desarrollo, como su capacidad para capturar relaciones entre elementos lejanos en datos secuenciales o aprender a partir de pocos ejemplos etiquetados. Este tipo de desafíos se encuentra frecuentemente en el área de la bioinformática. En esta investigación en particular trabajaremos con datos para la predicción de estructuras secundarias de ácidos ribonucleicos (RNA). El RNA es una molécula compuesta por una cadena de nucleótidos, que pueden ser de cuatro tipos: Adenina (A), Citosina (C), Guanina (G) o Uracilo (U). A diferencia del ADN, el RNA se forma como una sola molécula y no necesariamente adquiere una estructura de doble hélice. Durante su síntesis, el RNA es de cadena simple y se pliega sobre sí mismo (Mattick, 2023) (Zuker, 1981), (Zhang, 2019). La interacción de estas bases dentro de una molécula de RNA da lugar a su estructura secundaria, que es crucial para sus funciones y es fundamental para entender sus mecanismos de acción y su papel en los sistemas biológicos. Un ejemplo destacado son algunas de las vacunas contra el COVID-19, que utilizan RNA. Sin embargo, determinarla experimentalmente es costoso. Por eso, es importante desarrollar modelos computacionales de predicción.

Título del proyecto: Aumentación inteligente de datos para modelos de extremo a extremo y datos secuenciales en bioinformática.
Instrumento: PEIC I+D 2022-075

Año convocatoria: 2024

Organismo financiador: Agencia Santafesina de Ciencia Tecnología e Innovación

Director/a: Bugnon Leandro.

Es importante tener en cuenta las familias de RNA, que agrupan moléculas con estructuras y funciones similares. Cada familia presenta características estructurales específicas que pueden ser útiles para mejorar la precisión de los modelos de predicción. Reconocer y utilizar estas similitudes y diferencias es clave para desarrollar algoritmos capaces de predecir estructuras secundarias con mayor precisión.

Varios métodos han abordado la predicción de la estructura secundaria de RNA. Los enfoques clásicos, basados en la termodinámica, utilizan modelos de energía libres para estimar la estructura más estable. Sin embargo, estos métodos tienen limitaciones al no considerar adecuadamente interacciones complejas y estructuras no convencionales. Recientemente, los avances en aprendizaje profundo han ofrecido nuevas posibilidades para la predicción estructural de RNA. Modelos como SPOT-RNA (Singh, 2019) y E2Efold (Chen, 2020) han incorporado redes neuronales para mejorar la precisión de las predicciones. A pesar de sus logros, estos modelos a menudo requieren grandes volúmenes de datos de entrenamiento y pueden ser computacionalmente intensivos. Más recientemente estos métodos fueron superados por el sincFold (Bugnon, 2024): un modelo de aprendizaje profundo que predice la matriz de contactos nucleotídicos usando solo la secuencia de RNA y redes neuronales residuales (ResNet) para manejar dependencias a corto y largo alcance, permitiendo predicciones precisas con suposiciones mínimas.

Aumentar los datos de entrenamiento es crucial para mejorar el desempeño de los modelos de aprendizaje profundo. La cantidad y diversidad de los datos disponibles permiten al modelo aprender una mayor variedad de patrones, incrementando su capacidad de generalización. Por eso, en este trabajo haremos diferentes propuestas para la aumentación inteligente de datos para modelos de aprendizaje profundo como el sincFold.

OBJETIVOS

- Caracterizar y evaluar comparativamente los métodos más utilizados para aprender representaciones en secuencias biológicas.
- Proponer nuevos esquemas de aumentación de datos para modelos profundos de extremo a extremo para la predicción de estructuras a partir de las representaciones aprendidas.

METODOLOGÍA

Para llevar a cabo los experimentos de aumentación de datos, se utilizaron dos conjuntos de datos principales: RNAstralign y ArchiveII, ambos ampliamente usados por la comunidad científica.

- RNAstralign dataset (Weeks, 2010): 37.149 secuencias de 8 grandes familias de RNA. Es uno de los conjuntos de datos de estructura de RNA más completos disponibles.
- ArchiveII dataset (Tinoco, 1971): El conjunto de 3.975 secuencias más utilizado para métodos de plegamiento de RNA, que contiene estructuras de RNA de 9 familias.

Para evaluar el rendimiento de los modelos, se empleó la métrica F1, que es una medida que combina precisión y sensibilidad en una única métrica, proporcionando una visión equilibrada de la capacidad del modelo para predecir correctamente las estructuras secundarias de RNA.

Aumentaciones de datos propuestas:

1) Aumentación con distintos datasets: se entrenó el modelo con ambos datasets disponibles, los cuales contienen una amplia variedad de secuencias y estructuras secundarias de RNA. Para cualquier dataset de entrenamiento, dada la variabilidad de la precisión del modelo según la familia de RNA, no se usaron particiones de datos aleatorias, sino que se excluyeron dos familias de todo el dataset y lo que quedó se usó para el entrenamiento. Luego con las secuencias de una de las familias excluidas se confeccionó la partición de validación, y con la otra de test.
2) Aumentación con datos generados al azar: se generaron secuencias aleatorias de nucleótidos y predijo su estructura con un modelo termodinámico (Zhang, 2020) para resolver la escasez de datos experimentales. De este modo se puede crear un conjunto de datos adicional para agregarse al entrenamiento que aporta variabilidad y complejidad al modelo, mejorando así su capacidad de generalización.
3) Aumentación de datos inteligente: las secuencias aleatorias pueden ser muy distintas a las reales, y esto no ayuda en el aprendizaje del modelo. Por ello es que propusimos formas de producirlas de manera que consideren las características de las secuencias reales:
    a) Longitudes alrededor de mediana de cada familia: se generaron secuencias aleatorias de longitudes que están alrededor de la mediana de cada familia de RNA, para asegurar que el modelo se entrene con secuencias representativas de la distribución de longitudes observadas en los datos reales. Para esto, se hizo un análisis de las longitudes de secuencia de cada familia. Tomando algunas familias significativas, su mediana de longitudes es: 351 (16s), 117 (SRP), 76 (tRNA) y 363 (Telomerase).
    b) Secuencias parecidas: midiendo la distancia de edición entre una secuencia aleatoria y una real, se pueden generar secuencias que sean plausibles biológicamente. Se usó la distancia de edición para evaluar qué tan similares son las secuencias generadas a las secuencias reales en el dataset ArchiveII.

RESULTADOS Y CONCLUSIONES

Utilizando los datos de RNAstralign y ArchiveII, junto con las técnicas de aumentación de datos propuestas, se analizó cómo estas estrategias influyen en la precisión del modelo. Los resultados revelan mejoras significativas en F1 al aplicar métodos de aumentación, como la generación de secuencias aleatorias y la aumentación inteligente, optimizando así la capacidad del modelo para predecir estructuras secundarias de RNA con mayor exactitud.

Además, el modelo sincFold mostró un rendimiento superior en ciertas familias de RNA, particularmente en aquellas con mayor cantidad de datos o con secuencias más cortas. Para evaluar este aspecto, se entrenó el modelo con todos los datos disponibles, excluyendo dos familias: una se utilizó para validación y otra para prueba. Por cuestiones de espacio, se muestran solamente los resultados obtenidos con el tipo de aumentación inteligente.

Longitudes alrededor de mediana de cada familia: se generaron secuencias aleatorias con longitudes cercanas a la mediana de cada familia de RNA con características distintivas (16s, SRP, tRNA y Telomerase), asegurando que el modelo se entrene con secuencias representativas de la distribución real de longitudes. En la Tabla I se presentan los resultados: en cada fila se usó un dataset de entrenamiento diferente, primero sólo con ArchiveII (como referencia sin aumentación), luego sumándole 2000 secuencias aleatorias y luego sumándole

a ArchiveII 2000 secuencias aleatorias pero de longitud alrededor de la mediana de longitudes de secuencia de la familia con que se valida. En cada columna puede verse el F1 al testear con cada familia.

F1	16s	SRP	tRNA	telomerase
ArchiveII	0.3677	0.2985	0.6781	0.2221
ArchiveII + random2k	0.4202	**0.3591**	0.6699	0.2659
ArchiveII + random2k de len similar	**0.4101**	0.3156	**0.6954**	**0.2759**

Tabla I. Resultados al testear el modelo con distintas particiones y datos de entrenamiento, aumentados con secuencias aleatorias y con secuencias aleatorias de longitudes similares a la familia de Archive II con que se valida.

Se puede concluir que la propuesta de aumentación inteligente según la mediana de la longitud de cada familia aumenta el desempeño del modelo (con respeto a no usar aumentación) y en algunos casos mejora levemente respecto de usar datos puramente aleatorios
Como trabajo futuro se generará un conjunto de secuencias aleatorias cuyo puntaje de similaridad respecto a las reales sea mayor o igual al umbral (por ejemplo μ=0.5). En particular, que esta similaridad se de con la familia con que se va a testear y con estas secuencias construidas se completen los datos de entrenamiento. Es decir, sólo las secuencias generadas al azar con una similaridad biológica aceptable serán utilizadas para el proceso de entrenamiento del modelo.

BIBLIOGRAFÍA

Mattick, J. et al. (2023). Long non-coding RNAs: definitions, functions, challenges and recommendations. Nature Reviews Molecular Cell Biology, 24(6), 430–447.
Singh, J. et al. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nat Commun 10, 5407 (2019). https://doi.org/10.1038/s41467-019-13395-9
Tinoco, I. et al. (1971). Estimation of secondary structure in ribonucleic acids. Nature, 230(5293), 362–367.
Weeks, K. (2010). Advances in RNA structure analysis by chemical probing. Current Opinion in Structural Biology, 20(3), 295–304.
Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Research, 9(1), 133–148.
Zhang, H. et al. (2020). LinearPartition: linear-time approximation of RNA folding partition function and base-pairing probabilities. Bioinformatics, 36(Supplement 1), i258–i267.
Zhang, P., et al. (2019). Non-coding RNAs and their integrated networks. Journal of Integrative Bioinformatics, 16(3), 20190027.
Chen, X., et al.. (2020). RNA Secondary Structure Prediction By Learning Unrolled Algorithms. In Proceedings of the 8th International Conference on Learning Representations, 1(1), 1-10
Bugnon, L. et al. (2024). sincFold: end-to-end learning of short- and long-range interactions in RNA secondary structure. Briefings in Bioinformatics, 25(4), bbae271.