APRENDIZAJE POR TRANSFERENCIA PARA LA ANOTACIÓN FUNCIONAL DE PROTEÍNAS

Rosario Vitale

Instituto de investigación en señales, sistemas e inteligencia computacional, sinc(i), FICH-UNL, CONICET
Directora: Georgina Stegmayer

Área: Ingeniería

Palabras claves: Inteligencia artificial, Aprendizaje por transferencia, Aprendizaje automático

INTRODUCCIÓN

La anotación automática de proteínas sigue siendo un desafío sin resolver en bioinformática debido a la velocidad a la que se producen datos experimentales. Estos datos son anotados de forma manual, por lo tanto su realización es muy lenta (Bateman et al, 2023).

La anotación de proteínas incluye la clasificación de proteínas en familias y la base de datos de Pfam es ampliamente utilizada para este propósito. Los modelos clásicos de Pfam basados en Modelos ocultos de Markov (HMM) y BLASTp generan perfiles de familias a partir de las similitudes entre las secuencias de proteínas que pertenecen a una familia dada. De esta forma se pueden clasificar secuencias nuevas comparándolas con esos perfiles. Sin embargo aún queda alrededor del 25% de las proteínas que no han sido anotadas porque no coinciden con esas representaciones generadas. Esto ocurre ya que en algunos casos hay pocos ejemplos de una familia y no son suficientes para generar un perfil adecuado (Mistry et al, 2021).

Han aparecido recientemente modelos de aprendizaje profundo (DL, por sus siglas en inglés) capaces de inferir automáticamente patrones compartidos entre las secuencias de una familia, lo que permite la anotación computacional automática de secuencias completamente nuevas (Bileschi et al, 2022). Sin embargo, las técnicas de DL dependen de un gran número de datos para inferir patrones significativos en las secuencias y, como se dijo, muchas familias de Pfam constan de un número pequeño de secuencias de ejemplo. El aprendizaje por transferencia (TL, por sus siglas en inglés) ofrece una solución prometedora para este problema al ofrecer representaciones genéricas de secuencias de proteínas

Título del proyecto: Estimación de distancias semánticas y aprendizaje profundo para la predicción de nuevas funciones de genes

Instrumento: CAID 115

Año convocatoria: 2020

Organismo financiador: CONICET

Directora: Georgina Stegmayer

aprendidas previamente por grandes modelos de lenguaje y que luego pueden adaptarse para un problema específico (Unsal et al, 2022).

En este trabajo, se propone el uso de TL junto con modelos de ML para mejorar la clasificación de proteínas en familias. Resultados preliminares obtenidos han mostrado que incluso cuando TL se utiliza con modelos de aprendizaje automático clásico como k vecinos más cercanos (KNN, por sus siglas en inglés) y perceptrón multicapa (MLP, por sus siglas en inglés), estos modelos logran reducir significativamente el error de predicción en comparación con los métodos estándar y los modelos de DL con representaciones simples.

OBJETIVOS

- Desarrollar nuevos modelos y algoritmos basados en machine learning para clasificar proteínas en familias a partir de los vectores de características generadas con grandes modelos de lenguaje.
- Investigar el potencial del aprendizaje por transferencia (TL) en la tarea de clasificación de proteínas.

METODOLOGÍA

Para abordar este problema, se utilizó aprendizaje por transferencia (TL) que consta de dos etapas. Estas etapas se ilustran en la Figura 1. La primera etapa (izquierda) consiste en entrenar de forma auto-supervisada, con grandes bases de datos sin etiquetar, un modelo para realizar una tarea genérica, por ejemplo completar una secuencia dada. De esta manera se aprende una representación numérica que contiene información de todo el conjunto de datos y no solamente de los pocos que sí están anotados. La segunda etapa (derecha) utiliza la representación aprendida en la primera para entrenar un modelo, esta vez de forma supervisada y con datos etiquetados, para realizar otra tarea como puede ser la de clasificación en familias de Pfam.

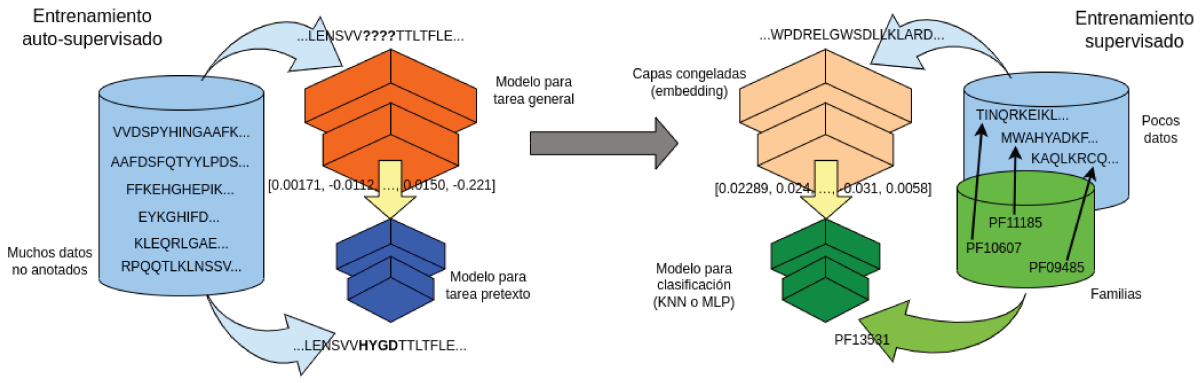

Figura 1: Aprendizaje por transferencia para la anotación funcional automática de proteínas

Hay varios grandes modelos de lenguaje de proteínas disponibles para la primera etapa. Los modelos seleccionados en este plan de trabajo fueron ESM1b, ProtTransBertBFD, ESM1v, ESM2 y ProtTransT5-XL-U50. Los dos primeros fueron elegidos porque resultaron ser los mejores en performance según una revisión del estado del arte realizada por el grupo de investigación (Fenoy et al, 2022). ESM1v y ESM2 se eligieron porque eran modelos más nuevos que ESM1b y sus autores los recomiendan debido a que obtienen mejores

resultados (Rives et al, 2021). De la misma manera, se optó por ProtTransT5-XL-U50 porque es la última versión disponible y es la que recomiendan sus autores por su mejor desempeño (Elnaggar et al, 2022).

Se entrenaron clasificadores muy utilizados en machine learning, como son MLP y KNN, recibiendo como entrada las proteínas codificadas con los modelos mencionados anteriormente. También se entrenaron ensambles. Un ensamble consiste en el entrenamiento de varios modelos que luego obtienen una única respuesta combinando por votación sus predicciones obtenidas individualmente. En este trabajo se ensamblaron 10 KNNs y 10 MLPs.

Para la validación cruzada del desempeño de los modelos se utilizó una partición de entrenamiento y prueba ya disponible de todo Pfam v32. Se compararon sus desempeños con otros modelos de aprendizaje profundo existentes como ProtCNN y el ensamble ProtENN (Bileschi et al, 2022) que utilizan representaciones más simples y están limitados, como se mencionó previamente, a entrenar únicamente con los pocos datos anotados.

RESULTADOS Y CONCLUSIONES

En la Tabla 1 se reproducen los resultados de modelos sin TL (mejor resultado en negrita). Los modelos HMM y BLASTp usan la secuencia cruda de la proteína como entrada. ProtCNN y ProtENN son modelos de DL reciben una representación simple de la proteína (de tipo one-hot). Se entrenó un MLP con la misma representación one-hot para facilitar la comparación de resultados.

Tabla 1: Resultados de modelos sin TL

Modelo	**HMM**	**BLASTp**	**ProtCNN**	**ProtENN**	**MLP**
Porcentaje de error (Número de errores)	18.10% (3844)	35.90% (7639)	27.60% (5882)	**12.20% (2590)**	41.57% (8852)

Por otro lado, entrenando modelos de aprendizaje automático con TL se obtuvieron los resultados mostrados en la Tabla 2. Comparando ambas tablas, el mejor resultado es el de MLP+ProtTransT5-XL-U50 con un 8,25% de error en clasificación, en comparación al ensable de modelos DL ProtENN con un 12,20% de error. Como se puede observar, con TL el error de predicción de la familia de cada proteína se puede reducir en un 55% con respecto a los métodos estándar de anotación funcional (HMM) y en un 32% con respecto a modelos de aprendizaje profundo con representaciones simples.

Tabla 2: Resultados de modelos con TL. Se muestran porcentajes de error y número de errores para cada combinación de modelo de aprendizaje y modelo de lenguaje aprendido por transferencia.

Método	**KNN**	**KNN Ensamble**	**MLP**	**MLP Ensamble**
ESM1b	15.16% (3229)	23.39% (4981)	14.33% (3052)	24.96% (5314)

ESM1v	21.33% (4541)	31.08% (6618)	18.59% (3959)	31.06% (6613)
ESM2	15.55% (3311)	26.90% (5728)	11.48% (2444)	21.63% (4605)
ProtTransBertBFD	39.49% (8408)	55.74% (11868)	21.63% (4606)	35.36% (7529)
ProtTransT5-XL-U50	8.63% (1838)	15.92% (3390)	**8.25% (1757)**	15.92% (3389)

Cabe destacar que aunque los modelos de aprendizaje automático utilizados en este estudio no son tan complejos como los de DL, los modelos profundos fueron efectivamente superados gracias al uso de TL, como se propuso en este plan de investigación. Se puede afirmar que la mejora en el desempeño se debe a TL ya que, usando la misma representación simple, el MLP es el que tiene más errores de clasificación, mientras que con un modelo de TL adecuado disminuye considerablemente su tasa de error. Como trabajo futuro, nos enfocaremos en probar TL con modelos de DL de forma de mejorar su performance en clasificación.

BIBLIOGRAFÍA

Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., Da Costa Gonzales, L. J., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., Zhang, J. 2023. Uniprot: the universal protein knowledgebase in 2023. Nucleic Acids Research, 51(D1), D523-D531.

Bileschi, M. L., Belanger, D., Bryant, D. H., Sanderson, T., Carter, B., Sculley, D., Bateman, A., DePristo, M. A., Colwell, L. J. 2022. Using deep learning to annotate the protein universe. Nature Biotechnology, 40(6), 932-937.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., Rost, B. 2022. Prottrans: Toward understanding the language of life through self-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 7112-7127.

Fenoy, E., Edera, A. A., Stegmayer, G. 2022. Transfer learning in proteins: Evaluating novel protein learned representations for bioinformatics tasks. Briefings in Bioinformatics, 23(4), bbac232.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., Bateman, A. 2021. Pfam: The protein families database in 2021. Nucleic Acids Research, 49(D1), D412-D419.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., Fergus, R. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15), e2016239118.

Unsal, S., Atas, H., Albayrak, M., Turhan, K., Acar, A. C., Doğan, T. 2022. Learning functional properties of proteins with language models. Nature Machine Intelligence, 4(3), 227-245.