Integración de Datos Proteómicos en Cáncer de Mama: Aplicación de Deep Learning y Autoencoders

I. Contexto de Aplicación

Las enfermedades crónicas como el cáncer representan una de las principales causas de muerte en el mundo. En el área clínica es muy común, que parte del diagnóstico y clasificación molecular se apoyen en biomarcadores obtenidos de ADN, ARN o proteínas. Siendo un caso particular, los estudios proteómicos (asociados a las proteínas expresados por el organismo de estudio) han permitido analizar la expresión diferencial de proteínas en en diferentes enfermedades, lo que ofrece información más cercana al fenotipo celular y que esta ligado a las características de estas enfermedades en el cuerpo humano [1], [2].

El dataset que se usará en este estudio proviene del Clinical Proteomic Tumor Analysis Consortium (CPC) y contiene perfiles proteómicos cuantificados mediante iTRAQ (Isobaric Tags for Relative and Absolute Quantification) la cual es una técnica de proteómica cuantitativa basada en espectrometría de masas (MS) que permite comparar los niveles de proteínas entre varias muestras en un mismo experimento. Para este estudio en general se analizaron 77 muestras de cáncer de mama [2]. Este conjunto de datos ha sido empleado previamente para identificar patrones de expresión proteica y clasificarlos en subtipos moleculares, en concordancia con el sistema PAM50, que constituye un estándar clínico de clasificación basado originalmente en la expresión de ARN [1], [3].

El interés de este trabajo es explorar cómo métodos de aprendizaje profundo, en particular autoencoders, pueden aprender representaciones latentes de los perfiles proteómicos que permitan mejorar la agrupación de pacientes en subtipos, y evaluar la integración de estos embeddings en esquemas de fusión de redes como Similarity Network Fusion (SNF) [6].

II. Objetivo de Machine Learning

El objetivo de machine learning en este proyecto es:

- Predecir o identificar subtipos moleculares de cáncer de mama a partir de datos proteómicos de alta dimensionalidad, usando arquitecturas de deep learning basadas en autoencoders.
- Reducir el ruido y la dimensionalidad de los datos para obtener una representación latente que capture patrones relevantes en la expresión proteica.
- Validar si estas representaciones permiten mejorar la clasificación respecto al sistema PAM50 o generar agrupamientos alternativos con significado biológico.

En términos prácticos, se busca que el modelo pueda aprender una representación robusta del proteoma y que estas representaciones sean útiles para el clustering de pacientes, abriendo la posibilidad de mejorar los sistemas de estratificación en oncología.

III. Dataset

El proyecto utiliza tres archivos de datos principales:

1. **77_cancer_proteomes_CPTAC_itraq.csv**
    - Contiene la expresión de aproximadamente 12.000 proteínas.
    - Incluye 77 muestras tumorales y 3 muestras de control de individuos sanos.
    - Cada fila corresponde a un identificador de proteína (RefSeq Accession), junto con su símbolo y nombre génico, seguido de las columnas de valores log2 de expresión.
    - Tamaño aproximado: ~15 MB en disco.
2. **clinical_data_breast_cancer.csv**
    - Contiene información clínica asociada a cada muestra, incluyendo clasificaciones moleculares y subtipos asignados según PAM50.
    - La primera columna permite enlazar con las muestras del dataset principal.
3. **PAM50_proteins.csv**
    - Lista de las 50 proteínas utilizadas en la clasificación PAM50, con sus identificadores RefSeq.
    - Permite contrastar el desempeño de los modelos con el estándar clínico de subtipos.

Distribución de las clases:
De acuerdo con la clasificación PAM50, los tumores de mama se dividen en **cinco subtipos**: Luminal A, Luminal B, HER2-enriched, Basal-like y Normal-like [1]. En el dataset clínico asociado, estos subtipos se encuentran de forma no balanceada, lo que constituye un reto para el modelado.

IV. Métricas de Desempeño

Para evaluar los modelos de autoencoder y las técnicas de clustering, se utilizarán métricas de desempeño tanto en la fase de reconstrucción como en la fase de agrupamiento:

1. **Errores de reconstrucción (fase de autoencoder):**
    - **Mean Squared Error (MSE)**: para medir qué tan bien el autoencoder reconstruye los perfiles proteómicos originales.
    - **Mean Absolute Error (MAE):** como alternativa más robusta frente a outliers.
2. **Métricas de clustering y clasificación (fase de subtipos):**
    - **Accuracy y F1-score**: si se comparan las etiquetas predichas con los subtipos PAM50 como referencia.
    - **Adjusted Rand Index (ARI)** y **Normalized Mutual Information (NMI):** métricas no supervisadas que cuantifican la similitud entre las agrupaciones obtenidas y las de referencia.
    - **Silhouette Score:** para evaluar la cohesión y separación de los clusters de manera intrínseca, sin referencia a etiquetas.

Estas métricas permiten validar tanto la capacidad del modelo de aprender representaciones significativas como su valor práctico en la estratificación de pacientes.

V. Referencias

[1] J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, ... and P. S. Bernard, “Supervised risk predictor of breast cancer based on intrinsic subtypes,” *Journal of Clinical Oncology*, vol. 27, no. 8, pp. 1160–1167, 2009, doi: 10.1200/JCO.2008.18.1370.

[2] P. Mertins, D. R. Mani, K. V. Ruggles, M. A. Gillette, K. R. Clauser, P. Wang, ... and S. A. Carr, “Proteogenomics connects somatic mutations to signalling in breast cancer,” *Nature*, vol. 534, no. 7605, pp. 55–62, 2016, doi: 10.1038/nature18003.

[3] K. A. Hoadley, C. Yau, T. Hinoue, D. M. Wolf, A. J. Lazar, E. Drill, ... and C. M. Perou, “Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer,” *Cell*, vol. 173, no. 2, pp. 291–304.e6, 2018, doi: 10.1016/j.cell.2018.03.022.

[4] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[5] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, “Deep learning–based multi-omics integration robustly predicts survival in liver cancer,” *Clinical Cancer Research*, vol. 24, no. 6, pp. 1248–1259, 2018, doi: 10.1158/1078-0432.CCR-17-0853.

[6] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, ... and A. Goldenberg, “Similarity network fusion for aggregating data types on a genomic scale,” *Nature Methods*, vol. 11, no. 3, pp. 333–337, 2014, doi: 10.1038/nmeth.2810.

[7] J. Tan, M. Ung, C. Cheng, and C. S. Greene, “Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders,” in *Proceedings of the Pacific Symposium on Biocomputing (PSB)*, vol. 20, pp. 132–143, 2015, doi: 10.1142/9789814644730_0014.