Master en Ciencia de Datos

Estado del Arte del TFM

Expresión Génica en Cáncer de Mama Asociados al Grado Histológico mediante Análisis de Microarrays

Gonzalo Canales Nuñez

Index

1. ESTADO DEL ARTE

1.1 Necesidad técnica del presente TFM derivada del problema actual en el diagnóstico del cáncer de mama

El diagnóstico del cáncer de mama ha avanzado considerablemente en las últimas décadas gracias a la incorporación de nuevas tecnologías de imagen, mejoras en los estudios patológicos y el desarrollo de herramientas moleculares. Sin embargo, persisten retos clínicos y tecnológicos que limitan la detección temprana y la precisión diagnóstica, con un impacto directo en la supervivencia y las posiblidades de personalización del tratamiento [1].

En particular, el grado histológico tumoral sigue siendo un marcador pronóstico fundamental, ya que se asocia directamente con la agresividad y la evolución clínica del tumor. No obstante, su evaluación mediante métodos morfológicos tradicionales, como el sistema de *Nottingham o Elston-Ellis*, presenta cierta variabilidad interobservador, lo que puede comprometer la reproducibilidad diagnóstica. Además, estos métodos no siempre reflejan la heterogeneidad molecular del cáncer de mama, una enfermedad reconocida por su diversidad biológica y clínica [2].

En este contexto, los perfiles de expresión génica ofrecen una vía prometedora para caracterizar los subtipos tumorales y explorar su relación con el grado histológico. El análisis transcriptómico permite capturar diferencias biológicas que no son evidentes en la morfología tisular y puede contribuir a una clasificación tumoral más objetiva mediante patrones de expresión génica.

Los recientes avances en biología computacional y aprendizaje automático permiten procesar grandes volúmenes de datos ómicos e identificar genes relevantes para la clasificación histológica y pronóstica del cáncer [3].

Por tanto, el presente trabajo se enmarca en la necesidad técnica de aplicar metodologías de ciencia de datos a la transcriptómica del cáncer de mama, con el objetivo de determinar si existen genes de expresión diferencial asociados al grado histológico. Este enfoque pretende aportar una herramienta complementaria y más objetiva para la clasificación tumoral dentro del paradigma de la medicina personalizada.

1.2 La aparición de la Bioinformática y los Datos Ómicos

Con la introducción de tecnologías de alto rendimiento como los microarrays y, posteriormente, la secuenciación masiva de ARN (RNA-Seq), la biomedicina ha entrado en la era de los datos ómicos. A continuación, se detallan los fundamentos de la tecnología utilizada en el presente trabajo.

1.2.1 Microarrays y expresión génica: fundamentos biológicos y tecnológicos

La expresión génica es el proceso mediante el cual la información contenida en un gen se traduce en una molécula funcional, habitualmente una proteína. Este proceso implica la transcripción del ADN en ARN mensajero (mRNA), cuya cantidad refleja el nivel de actividad del gen en un determinado tejido o condición biológica. Por ello, medir la expresión génica permite determinar qué genes están activos y en qué grado, proporcionando una instantánea del estado funcional de una célula.

Los microarrays de ADN fueron una de las primeras tecnologías que permitieron cuantificar de forma simultánea la expresión de miles de genes. Estos dispositivos consisten en una lámina de vidrio o silicio que contiene miles de sondas, que son pequeñas secuencias de ADN sintético, dispuestas en posiciones fijas, cada una diseñada para hibridar con el ARN de un gen específico. Durante el experimento, el ARN extraído de las muestras biológicas se marca con fluoróforos y se deposita sobre el chip, donde se une (hibrida) con sus sondas complementarias. Posteriormente, un escáner láser detecta la intensidad de fluorescencia en cada punto del microarray, la cual es proporcional a la cantidad de ARN presente y, por tanto, al nivel de expresión del gen correspondiente.

El resultado es una matriz donde cada fila corresponde a un gen y cada columna a una muestra. Para el análisis de Machine Learning, esta matriz se transpone para tener el formato *Muestras x Genes*. A poseriori, estos valores se transforman y normalizan para eliminar sesgos experimentales y permitir comparaciones entre muestras y experimentos. En la mayoría de los estudios los valores se expresan en escala $\log_2$, de manera que un incremento de una unidad equivale aproximadamente a una duplicación en la cantidad de ARN detectado. Esta transformación estabiliza la varianza y facilita los análisis estadísticos y el entrenamiento de modelos de aprendizaje automático.

Los microarrays constituyen, por tanto, una herramienta fundamental para generar datos transcriptómicos de alta dimensión, al proporcionar un panorama global de la actividad génica en una muestra. Su desarrollo marcó el inicio de la era de los datos ómicos, que permitió integrar la biología molecular con la estadística y la ciencia de datos, abriendo nuevas vías para el diagnóstico y la clasificación de enfermedades complejas como el cáncer.

1.2.2 Variabilidad y ruido en los datos de expresión génica

Los datos derivados de microarrays presentan una alta variabilidad, originada tanto por factores técnicos (ruido del escáner, calidad del ARN, condiciones de hibridación) como por factores biológicos (heterogeneidad entre pacientes, diferencias en la composición celular o en el estado fisiológico de las células). Estas fuentes de variación pueden introducir ruido significativo y afectar la reproducibilidad de los resultados.

Estas técnicas no solo corrigen las diferencias técnicas entre muestras, sino que también permiten identificar los patrones de expresión más consistentes y representativos del fenómeno biológico subyacente.

El manejo de estos datasets requiere, además del modelado predictivo, un riguroso preprocesamiento de datos. Esto incluye tareas esenciales como la imputación de valores faltantes, la transformación logarítmica de los niveles de expresión y la normalización para asegurar la comparabilidad entre muestras y la corrección de sesgos introducidos en el proceso de obtención y procesamiento, que cobran una vital importancia para garantizar la calidad del input del modelo e identificar los patrones de expresión más consistentes y representativos del fenómeno biológico subyacente.

1.2.3 Estructura y dimensionalidad de los datos

En términos cuantitativos, un dataset de microarrays típico contiene entre 100 y 500 muestras biológicas y entre 10 000 y 25 000 variables correspondientes a genes o sondas. Este tipo de estructura constituye un caso de alta dimensionalidad o paradigma $p \gg n$, en el que el número de variables es muy superior al número de observaciones [4].

Esta relación impone desafíos específicos al análisis computacional: el riesgo de sobreajuste, la redundancia entre genes y la pérdida de interpretabilidad del modelo. En consecuencia, resulta esencial aplicar métodos de selección o reducción de características, que permitan conservar únicamente los genes más relevantes para la tarea de clasificación o predicción.

En este contexto, los algoritmos de Machine Learning (ML) y las técnicas de reducción de dimensionalidad se han convertido en herramientas fundamentales para:

1. Extraer información relevante de *datasets* masivos.
2. Construir modelos predictivos que faciliten la toma de decisiones clínicas.
3. Identificar subconjuntos mínimos de genes con valor diagnóstico o pronóstico (biomarcadores) reduciendo, al mismo tiempo, la dimensionalidad del espacio de características [5].

Estas transformaciones tecnológicas y metodológicas abren la puerta a enfoques de clasificación molecular, como se discute en la siguiente sección.

1.3 De los Subtipos Moleculares a la Predicción del Grado Histológico

Los primeros estudios de análisis de expresión génica, como los realizados por Perou y Sørlie (2000), demostraron que los cánceres de mama pueden agruparse en subtipos moleculares con perfiles transcriptómicos distintos [6].

Esto abrió la puerta a una posible clasificación basada en datos, donde las diferencias en expresión génica reflejan los mecanismos biológicos subyacentes mejor que la morfología tradicional .

Posteriormente, otros estudios han demostrado que los perfiles de expresión génica reflejan con notable fidelidad las características histológicas y biológicas del tumor, incluyendo el grado de diferenciación celular, lo que respalda la existencia de una huella molecular asociada al grado histológico (GH) [7].

Esta correlación entre datos histológicos y transcriptómicos sustentaría la hipótesis que el presente trabajo plantea contrastar: la búsqueda de un modelo capaz de inferir el grado histológico del tumor a partir de su perfil de expresión génica, contribuyendo a la mejora en la precisión y la interpretabilidad del modelo.

Además, este tipo de enfoques podría contribuir al desarrollo de herramientas computacionales complementarias al diagnóstico histopatológico, aportando una estimación objetiva y reproducible del grado tumoral basada en datos moleculares.

1.4 El problema la Dimensionalidad y la Selección de Características

Los datasets de microarrays plantean un desafío clásico en la Ciencia de Datos: el número de variables (genes) es varios órdenes de magnitud superior al número de muestras disponibles. Este fenómeno, conocido como la maldición de la dimensionalidad (curse of dimensionality), fue introducido por Bellman en 1957 para describir las dificultades computacionales derivadas del aumento de dimensiones, y está estrechamente relacionado con problemas más contemporáneos

como el overfitting, la pérdida de interpretabilidad y la reducción de la capacidad de generalización de los modelos [8].

Para abordar este problema, la Selección de Características (Feature Selection, FS) se ha demostrado esencial. Los métodos de FS se agrupan en tres grandes categorías [9]:

- ***Filter*:** Evalúan la relevancia estadística de cada variable de forma independiente (e.g., *t-test*, ANOVA). Estos algoritmos son rápidos y **sirven como una referencia metodológica básica en la evaluación de genes individuales**, pero no consideran las interacciones entre genes.
- ***Wrapper*:** Utilizan un modelo de ML (p. ej., SVM, Random Forest) para evaluar iterativamente subconjuntos de genes según su rendimiento predictivo.
- ***Embedded*:** Integran la selección de características dentro del propio algoritmo de aprendizaje. Éste es el caso de la Regularización L1 (usada en LASSO), que aplica una penalización que fuerza los coeficientes de las variables menos relevantes a cero, realizando simultáneamente la reducción de dimensionalidad y el ajuste del modelo [10].

En general, los métodos de selección de características tipo *Wrapper* tienden a ofrecer subconjuntos más eficaces y adaptados al modelo, mejorando la precisión y la interpretabilidad a costa de mayor coste computacional, mientras que la reducción de redundancia y ruido contribuye a optimizar el rendimiento global del sistema [11]. Es por ello que en bioinformática, los métodos *Wrapper* junto con los *Embedded* son especialmente útiles ya que permiten la identificación de subconjuntos de genes más compactos y biológicamente interpretables, maximizando el rendimiento y reduciendo la redundancia.

1.5 *Machine Learning* en Expresión Génica: Modelado y Benchmarking

El uso de *Machine Learning* (ML) en la clasificación de tumores a partir de datos de expresión génica está ampliamente documentado [12,13]. Modelos como *Support Vector Machines* (SVM) o *Random Forest* (RF) se han aplicado con éxito en tareas de diagnóstico y pronóstico, normalmente precedidos por una reducción de dimensionalidad [14]. Estos algoritmos se emplean con frecuencia en entornos de alta dimensionalidad debido a su robustez frente al ruido y su capacidad para manejar un gran número de variables predictoras.

Pese a ello, los resultados de los modelos presentan una fuerte dependencia del método de selección de características (*Feature Selection, FS*) utilizado. En la práctica, el proceso de selección de genes tiene tanto impacto en la precisión del modelo como el propio algoritmo de clasificación, lo que justifica la necesidad de realizar un *benchmarking* sistemático de distintas estrategias de FS [15].

En este sentido, resulta especialmente relevante comparar dos estrategias contrastantes de selección de características:

- Por un lado, RFE (*Recursive Feature Elimination*) es un método *Wrapper* basado en eliminación recursiva, que utiliza un clasificador (como SVM) para evaluar iterativamente subconjuntos de genes y eliminar los menos relevantes en cada paso. Este enfoque suele

proporcionar una alta precisión y una selección adaptada al modelo, aunque presenta un elevado coste computacional [16].
- Por otro lado, LASSO (*Least Absolute Shrinkage and Selection Operator*) es un método *Embedded* con regularización L1, que penaliza los coeficientes de menor relevancia hasta llevarlos a cero. De esta forma, realiza simultáneamente la reducción de dimensionalidad y el ajuste del modelo, produciendo modelos estables con un coste computacional más bajo [10].

La comparación entre ambos enfoques permitirá determinar cuál de los dos logra un mejor equilibrio entre precisión, estabilidad e interpretabilidad biológica en el contexto de la clasificación del grado histológico del cáncer de mama.

El presente Trabajo Final de Máster se centrará, por tanto, en evaluar comparativamente los métodos RFE y LASSO mediante datos reales de microarrays, con el objetivo de identificar la firma génica mínima, robusta y biológicamente coherente capaz de discriminar de forma precisa los grados histológicos I, II y III del cáncer de mama.

A pesar de los avances logrados, aún no se han comparado de forma sistemática estos enfoques en el contexto específico de la clasificación del grado histológico, lo que plantea una oportunidad de investigación que se aborda en la siguiente sección.

1.6 Brecha de Investigación y Contribución del TFM

A pesar de los avances en el uso de datos transcriptómicos para la clasificación del cáncer de mama, la literatura actual presenta una carencia significativa de estudios comparativos que evalúen diferentes estrategias de selección de características desde una perspectiva integral. La mayoría de los trabajos se centran en demostrar la eficacia de un único método o en optimizar el rendimiento predictivo, sin analizar de manera conjunta otros factores igualmente determinantes como la estabilidad de las firmas génicas o su interpretabilidad biológica.

En este contexto, el presente TFM aborda esta brecha metodológica mediante un benchmarking comparativo entre dos enfoques representativos de selección de características:

- **LASSO (Least Absolute Shrinkage and Selection Operator)**, un método *embedded* basado en regularización L1 que combina ajuste de modelo y reducción de dimensionalidad en un solo proceso.
- **RFE (Recursive Feature Elimination)**, un método *wrapper* iterativo que evalúa y elimina características de manera recursiva en función del rendimiento del modelo subyacente.

El objetivo central no es únicamente determinar cuál de los dos métodos alcanza una mayor precisión predictiva, sino también comparar su estabilidad, eficiencia computacional y reproducibilidad biológica. Este enfoque pretende aportar evidencia sobre las ventajas relativas de cada familia metodológica en escenarios de alta dimensionalidad ($p \gg n$), típicos de los estudios de expresión génica.

Así, el trabajo contribuye a:

1. **Evaluar empíricamente** el desempeño de LASSO y RFE en la clasificación del grado histológico (I, II, III) del cáncer de mama.

2. **Analizar la estabilidad** de las firmas génicas resultantes frente a la variabilidad de las particiones de datos.
3. **Explorar la interpretabilidad biológica** de los genes seleccionados mediante análisis funcional (enriquecimiento GO o KEGG).
4. **Proponer recomendaciones metodológicas** para la selección de características en estudios ómicos de alta dimensionalidad.

Con ello, el TFM busca ir más allá del simple rendimiento numérico, promoviendo la adopción de buenas prácticas en la comparación, validación e interpretación de modelos de Machine Learning aplicados a datos biomédicos, contribuyendo a una mayor reproducibilidad y transparencia en el campo de la bioinformática traslacional.

1.7 Bibliografía capítulo

1. World Health Organization, *Breast Cancer Fact Sheet*, 2023. Disponible en: https://www.who.int/news-room/fact-sheets/detail/breast-cancer
2. C. M. Perou, T. Sørlie, M. B. Eisen *et al.*, “Molecular Portraits of Human Breast Tumours,” *Nature*, vol. 406, pp. 747–752, 2000.
3. K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine Learning Applications in Cancer Prognosis and Prediction,” *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.
4. Raza, M., et al. (2023). Dealing with dimensionality: the application of machine learning to multi-omics data. Bioinformatics, 39(2), btad021.
5. Shou, Q., et al. (2025). Cancer classification in high dimensional microarray gene expressions by feature selection using eagle prey optimization. Frontiers in Genetics, 15.
6. Perou, C. M., Sørlie, T., et al. (2000). Molecular portraits of human breast tumours. Nature, 406(6797), 747-752.
7. Sørlie, T., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences (PNAS), 98(19), 10869-10874.
8. Bellman, R. E. (1957). Dynamic Programming. Princeton University Press.
9. Li, J., Cheng, K., Wang, S., et al. (2017). Feature Selection for Classification: A Review. IJCSE, 14(3), 11-20.
10. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.
11. Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering, 17(4), 491-502.
12. Chen, X., & Liu, Y. (2009). The application of machine learning methods in cancer prediction. The Cancer Journal, 15(4), 269-276.
13. Sharma, H., & Rai, R. K. (2024). ML Classification of Cancer Types Using High Dimensional Gene Expression Microarray Data. ResearchGate.
14. Bhandari, K., & Agrawal, A. (2024). View of Class Prediction of High-Dimensional Data with Class Imbalance: Breast Cancer Gene Expression Data. ijasre, 10(2), 1-15.
15. Shou, Q., et al. (2025). Cancer classification in high dimensional microarray gene expressions by feature selection using eagle prey optimization. Frontiers in Genetics, 15.
16. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1), 389-422.