Identificación de biomarcadores con poder pronóstico en cáncer: una perspectiva desde la ciencia de datos biomédicos y la bioinformática

» Sebastián Menazzi[2 4 *], Hernán Chanfreau[1 4], David Nastasi[1], Juan Martín Lichowski[1], Diego Martinez[1], Genaro Camele[3], Matías Butti[1 4 **]

[1] CAETI , Universidad Abierta Interamericana, Argentina; [2] Hospital de Clínicas "José de San Martín", Argentina; [3] III-LIDI, Facultad de Informática, Universidad Nacional de La Plata, Argentina; [4] GenomIT S.A. - Research Unit
matias.butti@uai.edu.ar

Fecha de recepción: 1 de febrero de 2019.
Fecha de aceptación: 1 de junio de 2019.

Resumen

En el estudio del cáncer, los perfiles de expresión génica tienen gran relevancia dado que permiten conocer la actividad de genes de interés en el tejido en análisis. El avance biotecnológico y la disminución de costos de secuenciación han permitido producir grandes volúmenes de datos moleculares incluidos los perfiles de expresión génica, que pueden ser analizados junto con los datos de supervivencia (recidiva de un tumor u óbito) para obtener información valiosa sobre el pronóstico del paciente. El objetivo es identificar perfiles de expresión que muestren asociación con características clínicamente accionables, como respuesta a un tratamiento o capacidad de recidiva del tumor.

El análisis de estos grandes volúmenes de datos biomédicos requiere de conocimiento computacional, bioinformático y bioestadístico. La plataforma Bioplat permite democratizar estos análisis y es especialmente útil para equipos que tienen la experiencia biológica pero no la computacional/bioestadística. Además integra múltiples fuentes de datasets, permite incorporar datos propios y provee una base de datos curada. Ofrece puntos de extensión para que científicos de la computación puedan incorporar fácilmente nuevos algoritmos, herramientas o técnicas de machine learning.

PALABRAS CLAVE: GRANDES VOLÚMENES DE DATOS, BIOINFORMÁTICA, BIOESTADÍSTICA, BIOPLAT

* Sebastián Menazzi es Médico Genetista. Ex residente y jefe de residentes del Centro Nacional de Genética Médica. Actualmente médico de planta División Genética - Hospital de Clínicas "José de San Martín". Maestría en Biología Molecular Médica (UBA). Miembro de la Comisión Directiva de la Sociedad Argentina de Genética Médica. Docente de Genética en Facultad de Medicina - Universidad del Salvador. Director médico de genomIT.

** Matías Butti es director del grupo de investigación de Bioinformática en Oncogenómica funcional del CAETI. Es cofundador de GenomIT -empresa de bioinformática y ciencia de datos biomédicos- y Zoigen -laboratorio de estudios genómicos-. Licenciado en Informática de UNLP, Graduado de la Maestría en Explotación de Datos y Descubrimiento de Conocimiento (UBA) y está culminando su carrera de Medicina. Profesor de UAI, UTN y docente de UNLP.

Identification of biomarkers with predictive power in cancer: a perspective from the science of biomedical data and bioinformatics

Abstract

In the study of cancer, gene expression profiles have great relevance since they show the activity of genes of interest in the tissue under analysis. The biotechnological advances and the sequencing cost reduction have allowed to produce large volumes of molecular data including gene expression profiles, which can be analyzed together with survival data (recurrence of a tumor or death) to obtain valuable information on the prognosis of the patient. The objective is to identify expression profiles that show association with clinically actionable characteristics, in response to a treatment or recurrence capacity of the tumor.

The analysis of these large volumes of biomedical data requires computational, bioinformatic and biostatistical knowledge. The Bioplat platform allows to democratize these analyses and is especially useful for teams that have biological experience but not computational / biostatistical. It also integrates multiple sources of datasets, allows to incorporate the user's own data and provides a curated database. It offers extension points so that computer scientists can easily incorporate new machine learning algorithms, tools or techniques.

KEYWORDS: BIG DATA, BIOINFORMATICS, BIOSTATISTICAL KNOWLEDGE, BIOPLAT.

1. Introducción

El cáncer es, en esencia, una enfermedad genética (Vogelstein and Kinzler, 2004). La aparición y el progreso del tumor se caracterizan por la acumulación de mutaciones y la desregulación en la expresión de múltiples genes (Stratton and Rahman, 2008) (Hanahan y Weinberg, 2011). Esto indica que los procesos celulares afectados durante el desarrollo del cáncer se llevan a cabo de manera modular, donde raramente las alteraciones biológicas observadas puedan ser atribuidas a una modificación génica discreta (Stuart *et al.*, 2003).

Algunas de las alteraciones genéticas que predisponen a un alto riesgo de aparición de la enfermedad se heredan (están presentes en el ADN germinal, y son las llamadas formas hereditarias del cáncer), y pueden ser transmitidos a la descendencia, generalmente con un patrón de tipo autosómico dominante; otros se producen sobre tejidos específicos en algún momento de la vida del individuo (alteraciones somáticas, constituyen las formas más frecuentes del cáncer, de tipo esporádico), fenómeno que puede ser favorecido por factores nocivos internos o externos del ambiente en el que se desarrolla la célula (radiaciones, tóxicos ambientales o fenómenos de estrés oxidativo) . En la teoría del doble golpe de Knudson (Knudson, 1971) se postuló que para la aparición del tumor era necesario que hubiera una falta de función de las dos copias (los dos alelos) de ciertos genes denominados supresores tumorales; en ocasiones una de estas mutaciones es heredada en el ADN germinal y la otra se agrega a nivel del tejido.

Las mutaciones en genes asociados al tumor, junto a otros procesos químicos y moleculares (como la metilación, la regulación mediante microARN), tienen como repercusión la aparición de **patrones de transcripción** (patrones de actividad) de los diversos genes involucrados. Estos

patrones son diferentes a los que se observan en los tejidos libres de enfermedad, con algunos sobreexpresados (más actividad que lo esperado en ese tejido) y otros silenciados (menos actividad que lo esperado en ese tejido). Se denomina **transcriptoma o perfil de transcripción** a los niveles de expresión que presenta cada gen en el tejido en estudio, y fundamentalmente se refiere a la concentración de mensajeros de ARN, que son sensibles a señales de activación o supresión.

Biomarcadores

El NIH estadounidense define el concepto de biomarcador como un conjunto de signos medibles y evaluables objetivamente que permiten una asociación con procesos biológicos normales o patológicos , o bien con la respuesta a una intervención terapéutica (Biomarkers Definitions Working Group, 2001).

Si bien en la práctica clínica estos signos pueden ser variados (incluyen desde el examen físico del paciente hasta el diagnóstico por imagen o pruebas bioquímicas clásicas), la genómica ha ampliado la posibilidad de identificar biomarcadores basados en información molecular que aporta mayor precisión sobre el funcionamiento biológico del tejido. En particular los biomarcadores tumorales, definidos como un conjunto de elementos moleculares del paciente y del tumor que permiten mejorar el diagnóstico o la elección de la estrategia terapéutica, son de gran utilidad ya no solo a nivel académico sino también clínico.

Cuando estos biomarcadores se utilizan para inferir si un paciente es elegible o no para un tratamiento según la biología de su tumor, se los conoce como companion diagnostics biomarkers (Jørgensen and Hersom, 2016) (Agarwal *et al*., 2015) (Ocana *et al*., 2015). Su objetivo, en el marco de la denominada "medicina de precisión", es evitar tratamientos invasivos con los cuales no se verá beneficiado el paciente, por falta de eficacia o aparición de efectos adversos, o bien la selección de fármacos específicos que serán de gran eficacia para tratar ese tumor debido a su mecanismo de acción dirigido al defecto molecular identificable en el companion test.

Los primeros biomarcadores moleculares estudiaban alteraciones cromosómicas, y posteriormente variantes genéticas en el ADN. Con el avance biotecnológico y las mejoras en materia de capacidad de análisis de datos, surgió un gran interés en la búsqueda y aplicación de biomarcadores basados en el transcriptoma del tejido enfermo, es decir en los patrones de actividad de los genes en ese tejido puntual; esto se aplica en la actualidad en particular para el estudio del cáncer. (Casamassimi *et al*., 2017). Este interés se debe a que el transcritpoma es una representación dinámica del estado celular que resulta de mutaciones y alteraciones bioquímicas y moleculares.

Identificación de biomarcadores

En el ser humano se han descripto alrededor de 23.000 genes y unas 250.000 proteínas, y en la actualidad los métodos de lectura de la expresión del ADN (mediante el uso de chips de expresión o secuenciación masiva basada en técnicas de nueva generación) hacen posible analizar en un mismo experimento en paralelo múltiples señales de transcripción. Es por esto que **es indispensable el uso de herramientas bioinformáticas y ciencia de datos biomédicos** para encontrar asociaciones entre perfiles transcriptómicos y determinados resultado clínicos, generalmente la recidiva del tumor o el óbito del paciente. El estudio de estas asociaciones permitirá postular potenciales biomarcadores con poder pronóstico/predictivo. Un marcador pronóstico es el que tiene la capacidad de inferir cómo sería la progresión de la enfermedad en caso de ausencia de tratamiento o de solo tratamiento con terapias estándares no dirigidas; el objetivo es evitar o reducir tratamientos innecesarios.

Un caso de éxito de un biomarcador pronóstico basado en expresión génica, de uso frecuente en la práctica clínica de la oncología, es el OncoTypeDX, que mediante la evaluación de la expresión de 21 genes es capaz de inferir el riesgo de recidiva, dato de utilidad para decidir el uso o no de quimioterápicos adyuvantes a la cirugía en el cáncer de mama invasivo en estadío temprano con receptor de estrógeno positivo (McVeigh and Kerin, 2017).

2. Problema

Con el rápido avance biotecnológico, la disminución de costos de secuenciación y el interés académico y clínico en el análisis de datos genómicos, existen cada vez mayores volúmenes de datos de transcriptomas de pacientes diagnosticados con distintos tipos de cáncer, que podrían resultar de gran utilidad para encontrar más y mejores biomarcadores con poder pronóstico. Sin embargo, su análisis por parte de los especialistas en cada tumor resulta difícil por la complejidad técnica, computacional y bioestadística asociada, tanto para trabajar con datos públicos como propios de la institución a la que pertenecen.

3.Solución

3.1. Plataforma

Luego de varios años de trabajo en la identificación de marcadores pronóstico en cáncer con herramientas disponibles en la comunidad y desarrollos aislados e independientes (Abba, Butti *et al*., 2015) (Lara , Butti, Abba and Gutiérrez, 2013) (Lacunza, Butti, Abba 2013) (Abba, Lacunza, Butti and Aldaz, 2010), organizamos y automatizamos los pasos de análisis en una solución a la que llamamos Bioplat.

Bioplat es una plataforma de software para identificar, validar estadísticamente y optimizar (reducir dimensionalidad con la misma fuerza de asociación) biomarcadores con poder pronóstico en cáncer, basados en el transcriptoma del tumor en estudio.

Se trata de una herramienta de investigación básica y no tiene por objetivo identificar el marcador final sino acelerar el proceso de búsqueda de hipotéticos marcadores que pasen a una siguiente fase de validación experimental y estudios adicionales.

Si bien la plataforma presenta un conjunto de algoritmos y metahaeurísticas propios, fue diseñada proveyendo puntos de extensión para invitar a científicos de la computación, de la bioinformática y la bioestadística a implementar e incorporar nuevos algoritmos y técnicas de machine learning.

3.2. Pipeline de la plataforma

El proceso comienza con la identificación de un potencial biomarcador con poder pronóstico. Este paso inicial se realiza a partir de una lista de genes que postula el usuario, o importando marcadores publicados en bases de datos secundarias con las cuales se integra Bioplat o mediante una lista de genes con potencial poder pronóstico sugerida por la plataforma luego de aplicar distintos algoritmos de selección de genes (ver imagen 1). Éste es uno de los principales puntos de extensión para el trabajo futuro: incorporar nuevas técnicas de machine learning para que los potenciales biomarcadores sugeridos sean cada vez más precisos.

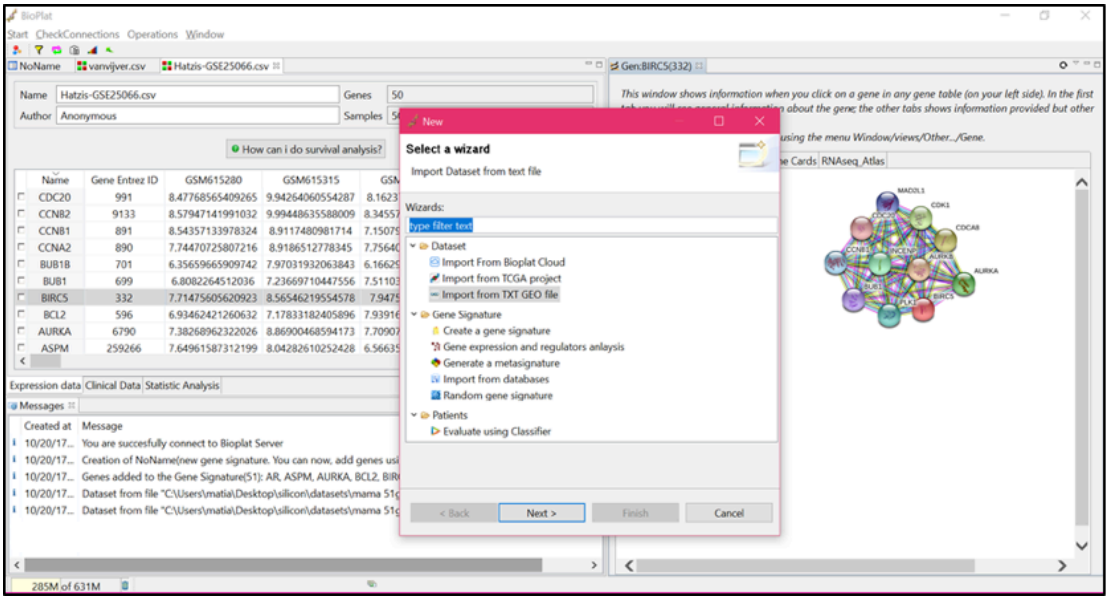

Imagen 1. Paso 1 en Bioplat: Identificación de biomarcadores

El segundo paso del proceso es la validación bioestadística basada en análisis de supervivencia. Se utilizan técnicas de logRankTest (Bland and Altman 2004), Concordance Index, curvas ROC y visualizaciones de heatmaps y Kaplan-Meier (imagen 2). Para esta validación es necesario disponer de datasets que integren los datos del transcriptoma del tumor y los datos de supervivencia del paciente (recidiva u óbito). Los datasets pueden ser provistos por el usuario o pueden ser importados desde bases de datos públicas como NCBI (Barret, 2019), molSigDB (Subramanian *et al.*, 2005) (Liberzon *et al.*, 2011), TCGA (Cancer Genome Atlas Research Network, 2013) con las cuales tiene integración la plataforma. También está disponible la "Bioplat Cloud", base de datos propia curada por nuestro equipo de biólogos.

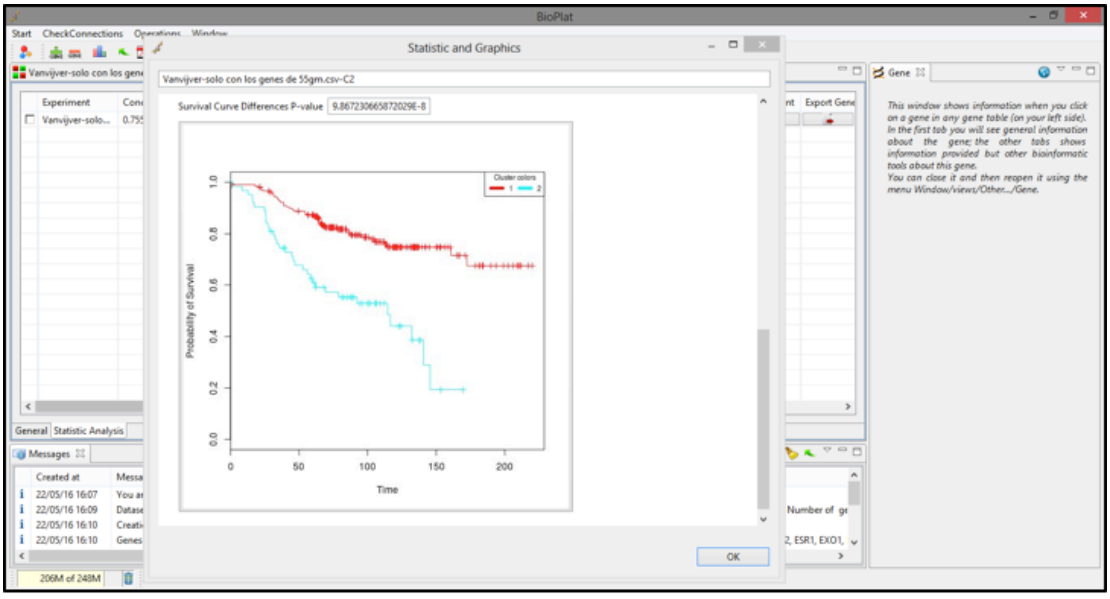

Imagen 2. Paso 2 en Bioplat: Validación estadística

El tercer paso es la optimización de biomarcadores: Habiendo obtenido un conjunto de genes que muestren poder pronóstico in-silico, puede ser útil realizar un paso adicional de "feature selection" para reducir la dimensionalidad, es decir identificar sólo aquellos genes que realmente están aportando a ese pronóstico. Si bien está disponible la "búsqueda ciega" que evalúa todas las combinaciones de genes, también se implementaron metaheurísticas, como Particle Swarm Optimization (Kennedy and Eberhart, 1995), para mejorar la performance de la búsqueda siendo que reduce el espacio de soluciones.

Una vez identificado un potencial biomarcador, la plataforma permite convertirlo en un clasificador para evaluar nuevos pacientes a partir de los niveles de expresión de sólo los genes seleccionados (imagen 3 e imagen 4).

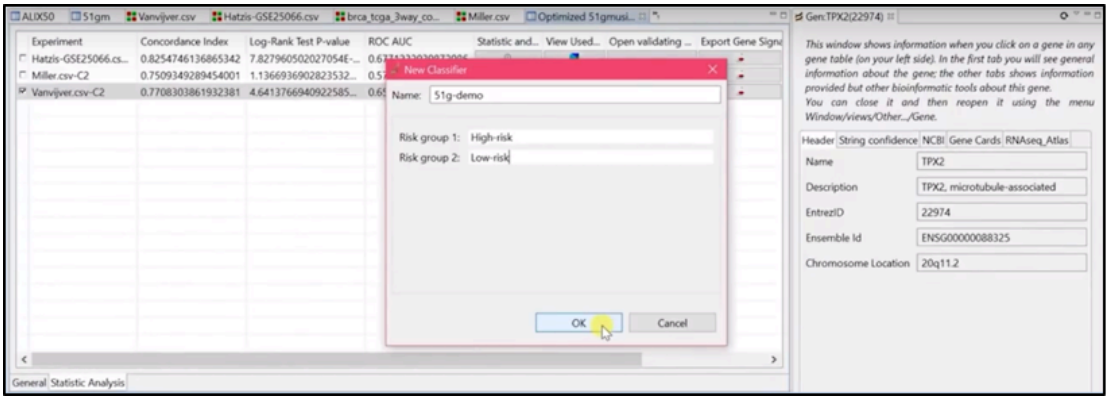

Imagen 3. Creación de clasificador a partir de un potencial biomarcador

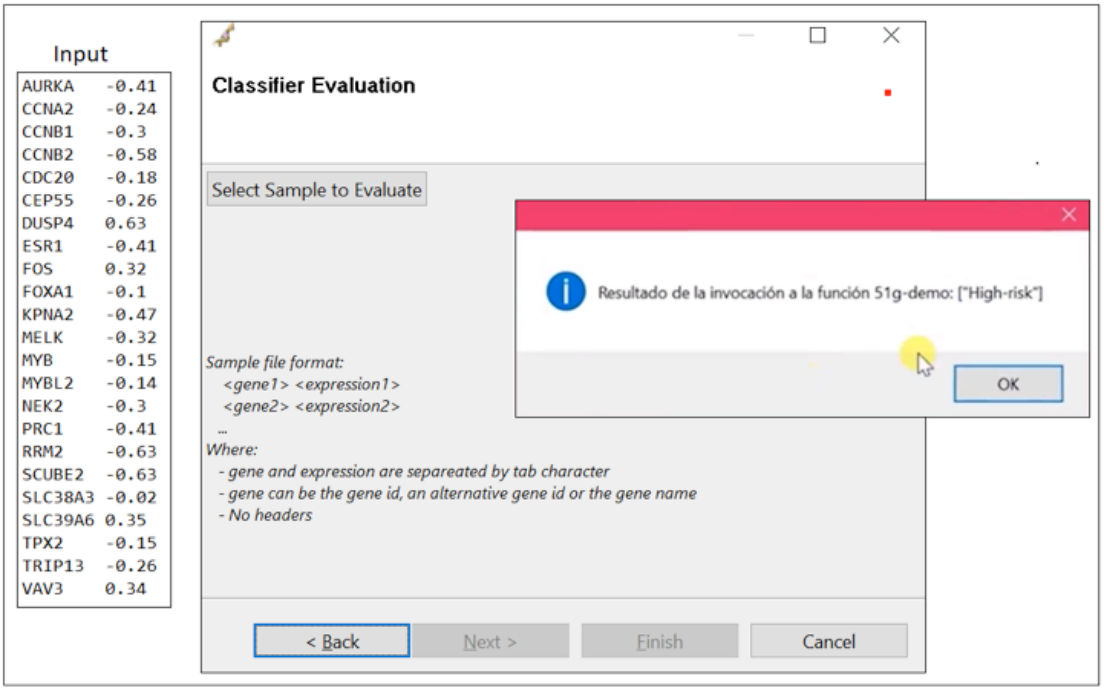

Imagen 4. Evaluación de un nuevo caso, a partir de los datos de expresión de sólo los genes seleccionados en el biomarcador

La plataforma adicionalmente ofrece una serie de herramientas, a las que denominamos "Companion tools" que permiten importación y exportación de datos, integración con diversas herramientas bioinformáticas de utilidad para el trabajo en la búsqueda de biomarcadores, normalización de datasets, particionamiento de datasets para validación, entre otras.

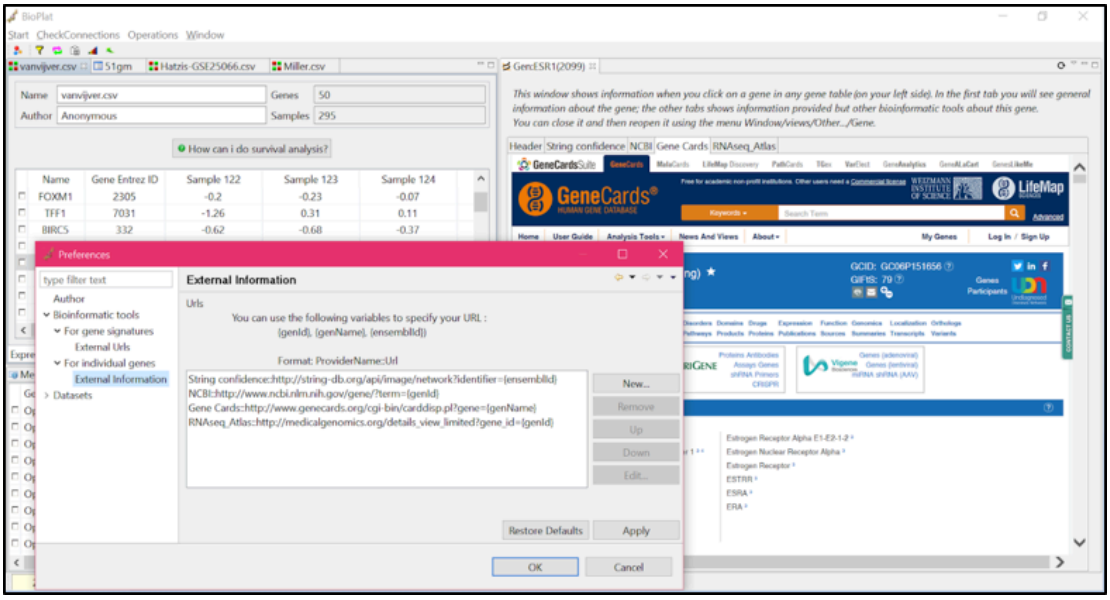

Imagen 5. "Companion tools" de Bioplat

Se puede observar el funcionamiento de la plataforma en este video resumido o en este video completo.

4. Materiales y métodos

La figura 6 muestra los componentes de la arquitectura y su forma de interacción.

Todas las operaciones estadísticas están implementadas utilizando R y Bioconductor (Gentleman, 2004) y se exponen al backend de Bioplat utilizando RServer. Debido a que el Backend está implementado en Java, se diseñó un bridge Java-R al que denominamos R4J. El cliente también está desarrollado en Java.

La Bioplat Cloud está implementada utilizando PostgreSQL y el sistema de gestión de la misma utiliza Spring Boot y Apache Wicket.

La predicción de clases para convertir en clasificador un hipotético biomarcador se base en el paquete pamr (Prediction Analysis for Microarrays) de R.

La plataforma es de uso libre y se puede pedir su acceso completando el siguiente formulario.

Toda la arquitectura está montada en máquinas virtuales sobre Docker y corriendo en la solución de nube de IBM (Bluemix).

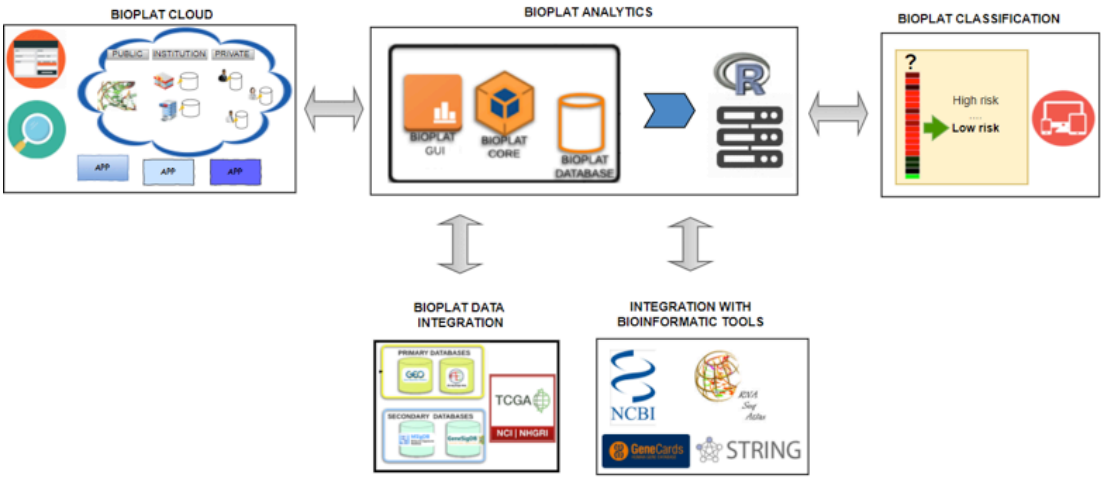

Imagen 6. Arquitectura de a plataforma Bioplat

5. Conclusiones

La identificación de biomarcadores es un área de gran interés para el avance biomédico en general y para el cáncer en particular. Plataformas como Bioplat contribuyen con la democratización de la identificación de potenciales biomarcadores, habilitando a investigadores especializados en un cáncer particular -pero no en bioestadística ni en computación- a analizar sus datos genómicos. Por lo tanto este tipo de plataformas constituyen una herramienta muy importante para el avance en la investigación del cáncer.

Bioplat presenta claros puntos de extensión para que investigadores de ciencias de la computación puedan integrar fácilmente sus ideas sobre análisis de datos transcriptómicos.

6. Trabajo futuro

» Probar y adaptar nuevas técnicas de machine learning para sugerir, de forma más precisa, lista de genes con potencial poder pronóstico.
» Implementar nuevas estrategias para reducir la dimensionalidad y compararlas con Particle Swarm Optimisation (Jones, 2005).
» Migración de la interfaz gráfica de RCP a web para simplificar su acceso.
» Integración con la plataforma multiomics desarrollada por nuestro equipo que integra, además de información transcriptómica, información de microARN y metilación.

Bibliografía

» Abba, M.C., Gong, T., Lu, Y., Lee, J., Zhong, Y., Lacunza, E., Butti, M., Takata, Y., Gaddis, S., Shen, J., Estecio, M.R., Sahin, A.A., Aldaz, C.M. *The molecular landscape of breast ductal carcinoma in situ* (DCIS).

» Abba, M.C., Lacunza, E., Butti, M.D., Aldaz, C.M. (2010). Breast cancer biomarker discovery in the functional genomic age: asystematic review of 42 gene expression signatures. *Biomarker Insights* 5: 103-118. ISSN: 1177-2719.

» Agarwal, A., Ressler, D., Snyder, G. (2015). The current and future state of companion diagnostics.

Pharmgenomics Pers Med. 8:99–110. doi:10.2147/PGPM.S49493

» Barrett, T. *et al*. (2013). *NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res*.;41:D991–D995.,

» Biomarkers Definitions Working Group, Atkinson Jr., A. J., Colburn, W. A., DeGruttola, V. G., DeMets, D. L., Downing, G. J. & Spilker, B. A. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical pharmacology & therapeutics* 69(3), 89-95.

» Bland, J.M., Altman, D.G. (2004). The logrank test. *BMJ*.;328(7447):1073. doi:10.1136/bmj.328.7447.1073

» Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A. *et al*. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet.* 45(10):1113–1120. doi:10.1038/ng.2764

» Casamassimi, A., Federico, A., Rienzo, M., Esposito, S., Ciccodicola, A. (2017). Transcriptome Profiling in Human Diseases: New Advances and Perspectives. *Int J Mol Sci.* 18(8):1652. Jul 29. doi:10.3390/ijms18081652

» Gentleman, R.C. *et al*. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 5, R80.

» Hanahan, D., Weinberg, R. (2011). "Hallmarks of cancer: The next generation". *Cell* 2011;144:646-674.

» Jones, K. O. (2005). *Comparison of genetic algorithm and particle swarm optimization. In Proceedings of the International Conference on Computer Systems and Technologies,* pp. 1-6.

» Jørgensen, J.T., Hersom, M. (2016). Companion diagnostics-a tool to improve pharmacotherapy. *Ann Transl Med.* 4(24):482. doi:10.21037/atm.2016.12.26

» Kennedy, J., Eberhart, R. (1995). Particle swarm optimization*. IEEE Neural Netw. Proc.*, 4, 1942-1948.

» Knudson, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences* 68(4), 820-823.

» Lacunza, E., Butti, M.D., Abba, M.C. (2011). SIlico identification of gene expression meta-Signature that predicts breast cancer prognosis – 1er. Congreso Argentino de bioinformática.

» Lara, S. J. P., Butti, M. D., Abba, M. C., Gutiérrez, F. A. A. (2013). *Nuevos biomarcadores moleculares para la detección del cáncer de pulmón: una aproximación desde la oncogenómica funcional*. ISBN 978-958-761-659. Bogotá: Instituto de Biotecnología-IBUN, Universidad Nacional de Colombia.

» Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27(12), 1739-1740.

» McVeigh, T. P., Kerin, M. J. (2017). Clinical use of the Oncotype DX genomic test to guide treatment decisions for patients with invasive breast cancer. *Breast Cancer: Targets and Therapy* 9, 393.

» Ocana, A., Ethier, J.L., Díez-González, L. *et al*. (2015). Influence of companion diagnostics on efficacy and safety of targeted anti-cancer drugs: systematic review and meta-analyses. *Oncotarget* 6 (37):39538–39549. doi:10.18632/oncotarget.5946

» Stratton, R. (2008). The emerging landscape of breast cancer susceptibility. *Nat Genet* 40(1):17-22.

» Stuart, J.M., Segal, E., Koller, D., Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302:249-255.

» Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 102(43), 15545-15550.

» Vogelstein, B., Kinzler, K.W. (2004). Cancer genes and the pathways they control. *Nat Med* 10:789- 799.