Arquitectura de Deep Symbolic Learning para Variant Calling en NGS

Ángel Canal-Alonso[1], Pedro Jiménez[1] and Noelia Egido[1], Juan Manuel Corchado[1]

[1] Departamento de Bioinformática y Biología Computacional, AIR Institute, Carbajosa de la Sagrada, España

E-mail: acanal@air-institute.com

Resumen

El proceso de Detección de Variantes (Variant Calling) es fundamental en la bioinformática, demandando una precisión y confiabilidad máximas. Este estudio examina una estrategia innovadora de integración entre un pipeline tradicional desarrollado in-house y un avanzado Sistema Inteligente (SI). Si bien el pipeline original ya contaba con herramientas basadas en algoritmos tradicionales, presentaba limitaciones, particularmente en la detección de variantes raras o desconocidas. Por tanto, se introdujo el SI con el objetivo de proporcionar una capa adicional de análisis, capitalizando las técnicas de aprendizaje profundo y simbólico para mejorar y potenciar las detecciones previas.

El principal desafío técnico residía en la interoperabilidad. Para superar esto, se empleó NextFlow, un lenguaje de scripting diseñado para gestionar flujos de trabajo bioinformáticos complejos. Mediante NextFlow, se facilitó la comunicación y el traspaso eficiente de datos entre el pipeline original y el SI, garantizando así la compatibilidad y reproducibilidad.

Posterior al proceso de Variant Calling del sistema original, se transmitían los resultados al SI, donde se implementaba una secuencia meticulosa de análisis, desde el preprocesamiento hasta la fusión de datos. Como resultado, se generaba un conjunto optimizado de variantes que se integraban con los resultados previos. Las variantes corroboradas por ambas herramientas se consideraban de alta fiabilidad, mientras que las discrepancias indicaban áreas para investigaciones detalladas.

El producto de esta integración avanzaba a etapas subsiguientes del pipeline, usualmente de anotación o interpretación, contextualizando las variantes desde perspectivas biológicas y clínicas. Esta adaptación no solo mantuvo las funcionalidades originales del pipeline, sino que también se potenció con el SI, estableciendo un nuevo estándar en el proceso de Variant Calling. Esta investigación ofrece un modelo robusto y eficiente para la detección y análisis de variantes genómicas, destacando la promesa y aplicabilidad del aprendizaje combinado en la bioinformática..

Palabras Clave: Next-Generation sequencing, Explainable Artificial Intelligence, Deep Symbolic Learning

---

Introducción

El Aprendizaje Simbólico Profundo (ASP) representa una innovadora corriente en el ámbito de la inteligencia artificial, enfocándose en fusionar las ventajas del Aprendizaje Profundo con las del aprendizaje simbólico. Este método nace en respuesta a una de las principales limitaciones del Aprendizaje Profundo: la ausencia de claridad y explicabilidad en sus modelos. Lo que el ASP busca es crear sistemas que, además de ser altamente eficientes en su rendimiento, sean igualmente claros y justificados en sus operaciones.

El Aprendizaje Simbólico Profundo funciona mediante la combinación de redes neuronales, que son adeptas a aprender representaciones detalladas y estratificadas de la información, con esquemas simbólicos que favorecen la elaboración de modelos lógicos y con una semántica bien definida. A diferencia del enfoque netamente numérico o estadístico del

Aprendizaje Profundo tradicional, el ASP integra elementos como símbolos, directrices y vínculos lógicos en su proceso, otorgando así una dimensión adicional de contexto y organización al saber generado.

Uno de los puntos sobresalientes del ASP es su habilidad para capitalizar el conocimiento previamente codificado en formatos simbólicos, utilizándolo como una base para orientar y enriquecer el aprendizaje de las redes neuronales. Esta característica resulta esencial en áreas donde hay un vasto cuerpo de saberes, como sucede en campos como la biología y la genómica.

Dentro de la fase de Detección de Variantes en flujos de trabajo de Secuenciación de Nueva Generación (NGS), el ASP surge como una estrategia con mucho potencial. Las redes neuronales tienen la capacidad de discernir patrones intrincados y matices en los datos de secuenciación. Paralelamente, el componente simbólico puede incluir directrices y datos previamente conocidos sobre variantes genéticas, mutaciones, y su importancia biológica. Esto podría no solo aumentar la exactitud en la detección de variantes, sino también brindar fundamentos lógicos y basados en conocimientos previos sobre las razones de clasificar ciertas secuencias como variantes.

La dualidad inherente al Aprendizaje Simbólico Profundo también introduce una versatilidad superior en el modelado. Mientras que las redes neuronales se adaptan a las especificidades y variaciones en la información, el segmento simbólico puede funcionar como un baluarte, asegurando que las inferencias y resultados estén en sintonía con la base de conocimientos biológicos previamente establecida. De este modo, el ASP se consolida como una herramienta puntera y confiable para abordar los retos específicos de la detección de variantes y del ámbito genómico en su totalidad.

En el ámbito de la bioinformática y la genómica, la precisión, eficiencia y explicabilidad son imperativos esenciales. Respondiendo a estas necesidades, este estudio presenta la concepción y estructuración de un Sistema Inteligente basado en Aprendizaje Simbólico Profundo (ASP) con el objetivo primordial de optimizar la fase de Detección de Variantes en flujos de trabajo de Secuenciación de Nueva Generación (NGS). Esta innovación es el fruto de un extenso y meticuloso proceso de diseño y experimentación.

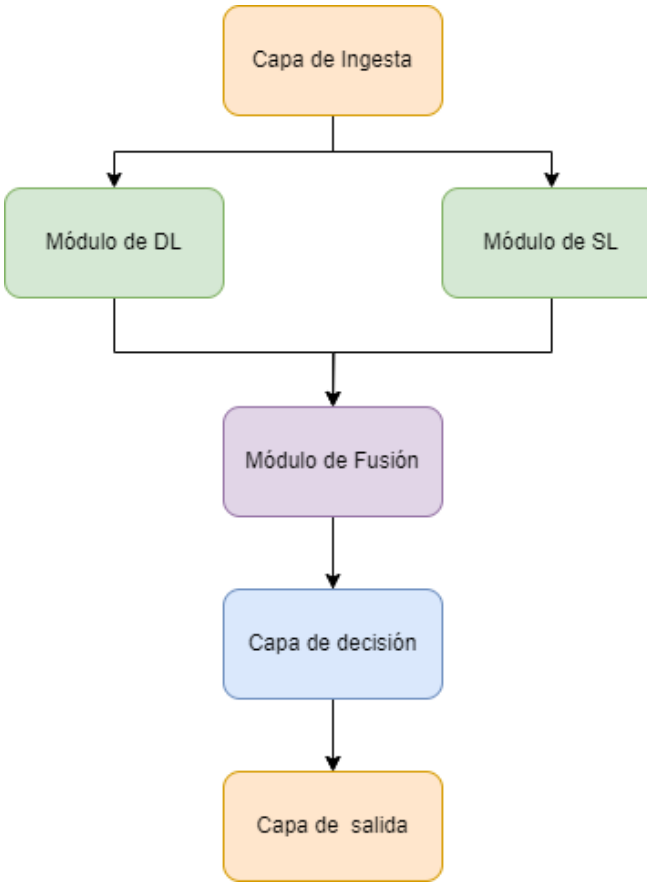
La motivación subyacente para desarrollar tal sistema radica en la convergencia de dos paradigmas: la potencia y capacidad de adaptación del aprendizaje profundo y la claridad semántica y estructural del aprendizaje simbólico. Al amalgamar estas dos perspectivas, aspiramos a configurar una herramienta que no solo ostente una precisión vanguardista sino que, adicionalmente, proporcione a los científicos y profesionales de la salud una perspectiva transparente e interpretable de los resultados derivados.

La arquitectura del sistema se despliega a través de distintas etapas cruciales. Inicialmente, se centra en la adquisición y un preprocesamiento riguroso de los datos de secuenciación, garantizando la calidad y pertinencia de la información con la que se alimentará el modelo. Posteriormente, nos embarcamos en el diseño, construcción y entrenamiento de modelos híbridos, que aprovechan la profundidad de las redes neuronales y la semántica de las representaciones simbólicas. Finalmente, se focaliza en la interpretación de los resultados, no solo desde una perspectiva numérica o estadística, sino también desde un marco lógico y semántico.

Cada uno de estos segmentos ha sido finamente calibrado para garantizar una cohesión máxima entre los componentes de aprendizaje profundo y simbólico. Esta integración asegura que, independientemente de las variaciones y ruidos intrínsecos a los datos de secuenciación, el sistema no solo retiene su robustez sino que también se mantiene alineado con el corpus de conocimiento genómico consolidado. En conjunto, este trabajo aspira a establecer un nuevo estándar en el ámbito de la detección de variantes genómicas, abogando por un enfoque integrador que combina precisión con explicabilidad.

Arquitectura del modelo

La arquitectura del modelo representa el núcleo del sistema y ha sido diseñada para encapsular de manera eficiente las capacidades de Deep Symbolic Learning. Esta arquitectura se compone de múltiples capas y módulos, cada uno con una función específica, que trabajan de manera conjunta para interpretar los datos de secuenciación



1. Capa de Ingesta y Preprocesamiento: Antes de que cualquier modelo pueda trabajar con los datos, estos deben ser adecuadamente preparados. Esta capa se encarga de la ingestión de datos brutos de NGS y realiza operaciones esenciales de preprocesamiento como la normalización, la corrección de errores y la identificación preliminar de regiones de interés. Además, transforma los datos en representaciones que pueden ser fácilmente consumidas por las capas subsiguientes.
2. Módulo de Aprendizaje Profundo: Una vez preprocesados, los datos son alimentados a este módulo, que consiste en redes neuronales convolucionales y recurrentes diseñadas específicamente para secuencias genómicas. Estas redes detectan patrones y características en los datos, como secuencias conservadas, regiones repetitivas o potenciales sitios de variantes. La naturaleza profunda de este módulo permite al sistema aprender representaciones jerárquicas de los datos, desde características de bajo nivel, como tríos de nucleótidos, hasta abstracciones de alto nivel, como estructuras genómicas.
3. Módulo de Aprendizaje Simbólico: Paralelamente al módulo de aprendizaje profundo, el módulo simbólico opera sobre los mismos datos. Utilizando un conjunto predefinido de reglas, heurísticas y conocimientos previos sobre genómica, este módulo identifica y valida hallazgos. Por ejemplo, puede confirmar la identificación de una variante si coincide con reglas conocidas sobre patrones de mutación o contextos genómicos específicos.
4. Módulo de Fusión: Este es uno de los componentes más críticos de la arquitectura. Aquí, las salidas de los módulos de aprendizaje profundo y simbólico se combinan y se integran en una representación unificada. Se emplean técnicas de fusión de datos para combinar las fortalezas de ambos enfoques, asegurando que el sistema no sólo detecte patrones en los datos sino que también los valide y los interprete de acuerdo con el conocimiento genómico existente.
5. Capa de Decisión: Con la información unificada proveniente del módulo de fusión, esta capa se encarga de tomar decisiones finales sobre la identificación y clasificación de variantes. Utiliza tanto información cuantitativa (como la probabilidad asignada por el módulo de aprendizaje profundo) como cualitativa (como la validación proporcionada por el módulo simbólico) para emitir juicios finales.
6. Capa de Interpretación y Salida: Una vez tomadas las decisiones, esta capa presenta los resultados de una manera interpretable para el usuario. Las variantes detectadas se anotan con información relevante, como su potencial impacto funcional o clínico. Además, gracias al componente simbólico, el sistema puede proporcionar explicaciones claras sobre por qué se identificó una variante particular.

En su conjunto, la arquitectura está diseñada para ser modular, permitiendo que componentes individuales sean actualizados o reemplazados según avance la investigación o cambien las necesidades. Esta flexibilidad garantiza que el sistema pueda mantenerse al día con los rápidos avances en genómica y bioinformática.

Preprocesamiento de los datos

El proceso de preprocesamiento en el contexto de secuenciación de nueva generación (NGS) es crucial, pues determina en gran medida la calidad y precisión de los resultados posteriores. A pesar de los avances en las tecnologías de secuenciación, los datos obtenidos directamente de los instrumentos suelen estar plagados de imperfecciones, artefactos y ruido. La Capa de Ingesta y Preprocesamiento actúa como el primer filtro y transformador de estos datos brutos, preparándolos para el análisis subsiguiente.

Inicialmente, cuando los datos son ingeridos, la capa comienza identificando y separando las diferentes muestras y runs, garantizando que cada conjunto de secuencias esté claramente demarcado y catalogado. Esta organización inicial

es fundamental para evitar confundir o mezclar datos provenientes de distintas fuentes o condiciones experimentales. Una vez catalogados, los datos pasan por un proceso de filtrado de calidad. Aquí, las secuencias que no cumplen con un umbral de calidad específico, determinado por la calidad de lectura, son descartadas o corregidas. Esta corrección se hace usando información redundante en los datos, como reads superpuestos, y con la ayuda de técnicas especializadas que estiman la secuencia correcta basándose en patrones observados en los datos.

A continuación, se lleva a cabo una etapa de alineación, donde las secuencias son mapeadas contra un genoma de referencia. Es en esta fase donde las variaciones, como las SNPs y las indels, comienzan a destacarse. Sin embargo, debido a la naturaleza inherentemente ruidosa de los datos de secuenciación, no todas las discrepancias observadas con respecto al genoma de referencia se deben a variantes reales. Errores sistemáticos en la secuenciación, artefactos y ruido pueden crear falsas señales. Por tanto, es esencial un paso de refinamiento post-alineación que recalibre y optimice la calidad de la alineación, ajustando y corrigiendo posibles errores.

Una vez que los datos han sido alineados y refinados, se realiza una operación de identificación de regiones de interés. Estas regiones son segmentos del genoma que son particularmente relevantes para el análisis subsiguiente, ya sea porque muestran signos de variación, porque están asociadas con regiones genómicas conocidas de interés clínico o funcional, o porque presentan patrones que indican la presencia de elementos regulatorios u otras características genómicas.

En paralelo con la identificación de estas regiones, la capa también se dedica a normalizar los datos. La normalización es esencial para garantizar que los datos de diferentes fuentes, tecnologías o condiciones experimentales sean comparables entre sí. Esta tarea puede involucra ajustar la profundidad de cobertura, corregir sesgos sistemáticos o transformar las secuencias en representaciones que sean más amigables para el análisis posterior.

El resultado de esta capa de preprocesamiento es un conjunto de datos limpio, estructurado y listo para ser consumido por los módulos de aprendizaje profundo y simbólico. Es esencial entender que, a pesar de ser solo el primer paso en el pipeline, las decisiones y operaciones realizadas en esta fase tienen un impacto directo en la precisión, eficiencia y calidad de los resultados finales del sistema.

Parámetros del modelo

Módulo de Deep Learning

El Módulo de Aprendizaje Profundo, como parte integral de nuestro sistema, juega un papel esencial en la detección y comprensión de patrones complejos y sutiles en los datos de secuenciación. Esta detección se basa en la capacidad del aprendizaje profundo para extraer características jerárquicas de los datos, desde aspectos fundamentales, como las secuencias de nucleótidos individuales, hasta interpretaciones de nivel superior, como la identificación de estructuras genómicas y regiones reguladoras.

Una de las piedras angulares del módulo es la incorporación de redes neuronales convolucionales (CNNs). Las CNNs son particularmente adecuadas para tratar datos de secuenciación, ya que pueden reconocer patrones locales dentro de las secuencias y son invariantes a desplazamientos. Es decir, si una particular secuencia de nucleótidos es indicativa de una variante o de un fenómeno genómico, las CNNs pueden detectarla independientemente de su posición en la secuencia de entrada. Las capas convolucionales en la red examinan segmentos superpuestos de la secuencia genómica, identificando patrones y características relevantes a través de filtros, y transformándolos en representaciones de alto nivel que son más fáciles de interpretar por las capas subsiguientes.

Por otro lado, también se integran redes neuronales recurrentes (RNNs), y más específicamente las LSTM (Long Short-Term Memory) dada la naturaleza secuencial de los datos genómicos. Estas redes son expertas en manejar dependencias a largo plazo y secuencias de longitud variable, lo que las hace ideales para comprender contextos más amplios en los datos de secuenciación. Por ejemplo, la influencia de una secuencia distante en la expresión o función de un gen en particular puede ser captada por estas unidades recurrentes.

Mientras que las CNNs se centran en los patrones espaciales y las RNNs captan la dinámica temporal, la combinación de ambas permite al módulo obtener una vista completa y detallada de los datos. Esta sinergia entre las redes convolucionales y recurrentes es crucial para abordar las complejidades y variabilidades inherentes a los datos genómicos.

Adicionalmente, en el proceso de entrenamiento de este módulo, es esencial definir y ajustar diversos parámetros. El ratio de aprendizaje, el tamaño del batch, la función de pérdida y la regularización son aspectos críticos que determinan la eficacia del modelo en aprender de los datos sin caer en sobreajuste. El modelo se entrena utilizando grandes conjuntos de datos etiquetados, donde las secuencias genómicas vienen acompañadas de información sobre las variantes conocidas, estructuras genómicas y otras anotaciones relevantes. Con el tiempo, el modelo ajusta sus pesos y parámetros internos para minimizar la discrepancia entre sus predicciones y los datos reales, llegando a una representación óptima que puede generalizarse a nuevos datos no vistos anteriormente.

En conjunto, el Módulo de Aprendizaje Profundo es un amalgama de técnicas avanzadas de aprendizaje automático

que, al trabajar en conjunto, permiten al sistema identificar, clasificar y comprender las variantes y estructuras presentes en los datos de secuenciación de nueva generación con una precisión y eficiencia sin precedentes. La naturaleza adaptativa y evolutiva de este módulo asegura que, a medida que se disponga de más datos y conocimientos, el sistema pueda continuar mejorando y refinando su capacidad de interpretación genómica.

Módulo de aprendizaje simbólico

El Módulo de Aprendizaje Simbólico representa una faceta fundamental de nuestro sistema que contrasta y complementa al módulo de aprendizaje profundo. Mientras que el aprendizaje profundo excava en los datos para descubrir patrones implícitos y relaciones complejas, el aprendizaje simbólico se centra en representar y utilizar el conocimiento explícito y estructurado sobre el dominio en cuestión, en este caso, la genómica.

La base de este módulo radica en la creación y manipulación de representaciones simbólicas de la información. En el contexto de secuenciación genómica, estas representaciones abordan estructuras genómicas conocidas, reglas heredadas de estudios previos, patrones genéticos asociados con fenotipos específicos, entre otros. Estos símbolos y reglas se organizan en estructuras de conocimiento, a menudo referidas como bases de conocimiento, que son esencialmente sistemas de reglas o lógicas diseñadas para razonar sobre los datos.

Uno de los enfoques centrales dentro del aprendizaje simbólico es el sistema basado en reglas. Estas reglas son derivadas de conocimientos previos, literatura científica o incluso a través de expertos en el campo. Estas reglas son aplicadas a los datos para filtrar, clasificar y predecir la presencia de fenómenos genómicos particulares. Las reglas definidas para el presente modelo son:

- Regla de Secuencia Motif: Si se detecta un motif específico de nucleótidos en una región promotor, predice la unión de un factor de transcripción conocido.
- Regla de Variante Patogénica: Si se identifica una variante en un exón de un gen asociado con una enfermedad hereditaria y esa variante ha sido previamente catalogada como patogénica, clasifícala como de alto riesgo.
- Regla de Splicing: Si se detecta una variante en las dos primeras o últimas posiciones de un intrón, considera la posibilidad de que afecte a los sitios de splicing y, por tanto, a la formación del ARNm.
- Regla de Conservación: Si una variante se encuentra en una región altamente conservada a través de diferentes especies, es probable que esa región tenga una función biológica importante.
- Regla de Repetición: Si una secuencia tiene múltiples repeticiones de un trinucleótido específico, considera la posibilidad de que esté relacionada con enfermedades de repetición trinucleotídica.
- Regla de Silenciador: Si una variante se encuentra en una región conocida por contener elementos silenciadores, evalúa el potencial de dicha variante para afectar la regulación génica.
- Regla de Interacción Proteica: Si una variante se identifica en un dominio de interacción proteica, investiga su potencial impacto en la formación de complejos proteicos.
- Regla de Sinónimos: Si una variante no cambia el aminoácido resultante en una proteína, generalmente se clasifica como sinónima, pero aún puede ser revisada para potenciales efectos en el splicing o la estabilidad del ARNm.
- Regla de Efecto Fundador: Si una variante específica es común en una población o grupo étnico particular y se asocia con una enfermedad, considera la posibilidad de un efecto fundador.
- Regla de Compensación: Si se detectan múltiples variantes en un mismo gen o vía y una es patogénica, investiga las otras variantes para ver si tienen un potencial efecto compensatorio.

Una ventaja clave del aprendizaje simbólico es su capacidad de interpretabilidad. A diferencia de los modelos de aprendizaje profundo, que a menudo son considerados como cajas negras, los sistemas basados en reglas ofrecen un razonamiento claro y transparente detrás de cada decisión o predicción. Esta claridad es invaluable en campos como la genómica, donde la interpretación y justificación de los resultados puede tener implicaciones significativas en áreas como el diagnóstico clínico o la investigación biomédica.

Sin embargo, no es suficiente simplemente codificar el conocimiento existente. El módulo también es capaz de "aprender" o refinar sus reglas y representaciones basándose en nuevos datos. Mediante técnicas como la inducción de reglas, el módulo puede examinar los datos, compararlos con su base de conocimiento actual y ajustar, eliminar o crear nuevas reglas para reflejar mejor la realidad de los datos. Esto es especialmente útil en un campo en constante evolución como la genómica, donde nuevos descubrimientos pueden cambiar nuestra comprensión de los sistemas biológicos.

En combinación con el Módulo de Aprendizaje Profundo, el Módulo de Aprendizaje Simbólico ofrece una comprensión holística y profunda de los datos genómicos. Mientras que el primero se centra en descubrir patrones y relaciones no evidentes en los datos, el segundo proporciona un marco estructurado y justificado para interpretar y razonar sobre estos descubrimientos. Juntos, ofrecen una potente combinación de intuición basada en datos y razonamiento basado en conocimientos, permitiendo al sistema operar con

una precisión, eficiencia y transparencia sin parangón en el análisis de secuenciación de nueva generación.

Módulo de Fusión

El Módulo de Fusión se erige como el componente integrador esencial en nuestro sistema, encargado de amalgamar los resultados y las intuiciones obtenidos tanto del Módulo de Aprendizaje Profundo como del Módulo de Aprendizaje Simbólico. Esta tarea es esencial ya que, si bien ambos módulos por separado son poderosos, es su colaboración coordinada la que da origen a la verdadera sinergia y potenciación del análisis.

Desde una perspectiva técnica, el Módulo de Fusión opera en varias etapas. Inicialmente, recopila las salidas del Módulo de Aprendizaje Profundo. Estas salidas, en forma de vectores de características, representaciones latentes o clasificaciones directas, encapsulan patrones complejos y relaciones no lineales descubiertas en los datos. Estas representaciones son extremadamente valiosas pero pueden carecer de interpretabilidad directa o de conexiones con el conocimiento biológico explícito.

Simultáneamente, el Módulo de Fusión accede a la base de conocimientos del Módulo de Aprendizaje Simbólico. Las reglas, estructuras y representaciones simbólicas proporcionan un marco estructurado y una contextualización para los datos, basados en décadas de investigación y comprensión en genómica.

Con ambas fuentes de información a su disposición, el Módulo de Fusión comienza el proceso de integración. Emplea técnicas avanzadas, como el razonamiento basado en lógica difusa y las redes neuronales de atención, para ponderar adecuadamente la información de ambos módulos. En esencia, se trata de determinar dónde confiar más en las predicciones basadas en datos y dónde aplicar el conocimiento simbólico para corregir, guiar o complementar esas predicciones.

Por ejemplo, si el Módulo de Aprendizaje Profundo detecta una posible variante genética de interés pero esa variante está en contradicción con una regla simbólica bien establecida, el Módulo de Fusión puede optar por dar prioridad a la regla o, al menos, enviar una alerta para una revisión más detallada.

Una consideración crucial en este proceso es la retroalimentación. El Módulo de Fusión no solo integra, sino que también aprende. A medida que recibe más datos y se enfrenta a más escenarios, refina su capacidad para equilibrar y combinar información de los otros módulos. Esto es esencial para garantizar que el sistema, en su conjunto, se mantenga adaptativo y evolutivo, ajustándose a los nuevos desafíos y descubrimientos en el campo de la genómica.

Finalmente, el Módulo de Fusión culmina su operación produciendo una serie de resultados unificados que incorporan tanto el aprendizaje profundo como el simbólico. Estos resultados pueden ser clasificaciones, predicciones, anotaciones o cualquier otro formato de salida relevante para el análisis genómico, pero lo que es seguro es que reflejan una visión integrada y holística del problema, aprovechando lo mejor de ambos mundos.

Implementación final

En el entorno en constante evolución de la bioinformática, el proceso de Detección de Variantes (Variant Calling) representa un pilar crítico, donde la precisión y confiabilidad son esenciales. En este contexto, el pipeline original, desarrollado in-house, ya contaba con herramientas de avanzada para dicho proceso. Si bien estas herramientas, fundamentadas en algoritmos convencionales y repositorios de referencia, demostraban una eficiencia razonable, presentaban ciertas limitaciones, particularmente en términos de falsos positivos y en la detección de variantes atípicas o aún no catalogadas.

Con la finalidad de reforzar y no suplantar el sistema existente, se introdujo un Sistema Inteligente (SI). Esta estrategia fue motivada por la aspiración de que el SI ofreciera un nivel adicional de análisis, empleando técnicas de aprendizaje profundo y simbólico, con la finalidad de revisar y potencialmente mejorar las detecciones previamente identificadas por el pipeline original.

El primer desafío tecnológico enfrentado fue garantizar una comunicación fluida y eficaz entre el SI y las herramientas del pipeline original, dadas las particularidades de ambos sistemas. En este escenario, recurrimos a NextFlow, un lenguaje de scripting diseñado específicamente para gestionar workflows bioinformáticos complejos. Gracias a sus capacidades innatas para coordinar, monitorizar y asegurar la reproducibilidad de tareas, NextFlow emergió como la plataforma ideal para nuestra integración.

El proceso de integración se inició con la creación de una función específica en NextFlow que invocaba al SI. Posteriormente, tras el proceso de Variant Calling del sistema original, se facilitaba el traspaso de resultados al SI utilizando canales de NextFlow. Esta transición se gestionó garantizando que los datos fueran compatibles, adoptando generalmente el formato VCF.

Una vez dentro del SI, se instauró una secuencia de análisis que comenzaba con un módulo de preprocesamiento, seguido por módulos de aprendizaje profundo y simbólico, culminando con un módulo de fusión. La culminación de este proceso generaba un conjunto de variantes optimizadas que se combinaban con los resultados previos.

Esta convergencia ofreció notables beneficios. Las variantes identificadas de forma uniforme por ambas herramientas eran consideradas de alta fiabilidad. Las

discrepancias, por contraste, proporcionaban un indicativo para futuras investigaciones o revisiones detalladas.

El producto final de esta integración se transmitía a la siguiente etapa del pipeline, usualmente orientada a la anotación o interpretación, donde las variantes eran analizadas desde una perspectiva biológica y clínica.

En resumen, esta adaptación permitió que el pipeline original no solo retuviera sus funcionalidades iniciales, sino que también se enriqueciera con la profundidad y precisión del SI, potenciando de esta manera el proceso integral de Variant Calling. La integración descrita en este estudio puede servir como un paradigma para futuras investigaciones y desarrollos en el ámbito bioinformático.

Conclusiones

A pesar del éxito y la notable mejora que el Sistema Inteligente (SI) ha aportado al proceso de Variant Calling, como con cualquier tecnología emergente, existen ciertas limitaciones que deben ser abordadas para alcanzar su máximo potencial.

Una de las principales restricciones ha sido la capacidad computacional. Aunque el SI es altamente eficiente en su operación, los módulos de aprendizaje profundo y simbólico, por su naturaleza, requieren un alto grado de poder computacional, especialmente cuando se manejan grandes volúmenes de datos. En la etapa actual, se ha enfrentado ocasionalmente a cuellos de botella en términos de velocidad de procesamiento. Con una inversión adicional en infraestructura, como la adquisición de hardware más potente o la implementación en entornos de computación en la nube con mayores recursos, estos desafíos podrían mitigarse fácilmente.

Otra limitación ha sido el tamaño y la diversidad de los datos de entrenamiento disponibles. Si bien se han utilizado vastos conjuntos de datos para entrenar el SI, siempre existe el riesgo de sesgos inadvertidos. Ampliar la diversidad de los datos, tanto en términos geográficos como étnicos, permitiría mejorar la generalización del sistema a diferentes poblaciones. Una financiación adicional podría destinarse a la adquisición de más datos y al establecimiento de colaboraciones con instituciones que posean repositorios genómicos variados.

La integración con bases de datos externas, aunque ha sido una fortaleza, también presenta limitaciones. Dependiendo de la disponibilidad y la actualización de estas bases de datos, podría haber lagunas en el conocimiento que el SI puede acceder. En el futuro, sería ideal considerar la creación de una base de datos interna, constantemente actualizada, que compendie la información más reciente en genómica y que pueda ser alimentada tanto por resultados internos como externos.

Por último, mientras que el reentrenamiento incremental ha sido una ventaja, la periodicidad y la eficacia del mismo podrían mejorar con la implementación de un sistema más automatizado que monitoree continuamente la aparición de nuevos datos y ajuste el modelo en tiempo real. Esto podría requerir, nuevamente, una inversión en desarrollo y en sistemas de monitoreo de vanguardia.

References

Garcia-Retuerta D, Canal-Alonso A, Casado-Vara R, Rey AM, Panuccio G, Corchado JM. Bidirectional-Pass Algorithm for Interictal Event Detection. In Practical Applications of Computational Biology & Bioinformatics, 14th International Conference (PACBB 2020). PACBB 2020. Advances in Intelligent Systems and Computing, vol 1240. Springer, Cham. https://doi.org/10.1007/978-3-030-54568-0_20

Castillo Ossa LF, Chamoso P, Arango-López J, Pinto-Santos F, Isaza GA, Santa-Cruz-González C, Ceballos-Marquez A, Hernández G, Corchado JM. A Hybrid Model for COVID-19 Monitoring and Prediction. Electronics. 2021; 10(7):799.
https://doi.org/10.3390/electronics10070799

Intelligent Platform Based on Smart PPE for Safety in Workplaces. Márquez-Sánchez S, Campero-Jurado I, Herrera-Santos J, Rodríguez S, Corchado JM. Sensors (Basel). 2021 Jul 7;21(14):4652
https://doi.org/10.3390/s21144652

A. Canal-Alonso, R. Casado-Vara and J. Manuel Corchado, "An affordable implantable VNS for use in animal research," 2020 27th IEEE International Conference on Electronics, Circuits and Systems (ICECS), 2020, pp. 1-4,
doi: 10.1109/ICECS49266.2020.9294958

An Agent-Based Clustering Approach for Gene Selection in Gene Expression Microarray. Ramos J, Castellanos-Garzón JA, González-Briones A, de Paz JF, Corchado JM. Interdiscip Sci. 2017 Mar;9(1):1-13
DOI 10.1007/s12539-017-0219-6

Agradecimientos

El presente estudio ha sido financiado por el proyecto AIR Genomics (con número de expediente CCTT3/20/SA/0003), mediante la convocatoria 2020 PROYECTOS I+D ORIENTADOS A LA EXCELENCIA Y MEJORA COMPETITIVA DE LOS CCTT por el Instituto de Competitividad Empresarial de Castilla y León y fondos FEDER