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

En la era de la genómica, el análisis eficiente y preciso de secuencias genómicas es fundamental. La tecnología de secuenciación de nueva generación (NGS) ha revolucionado el campo de la genómica al proporcionar un volumen masivo de datos a una escala sin precedentes. Uno de los pasos críticos en el análisis de estos datos es el "variant calling", donde se identifican variaciones genéticas a partir de secuencias de ADN. En este contexto, hemos explorado el uso del Deep Symbolic Learning (DSL) como una innovadora aproximación computacional que combina el aprendizaje profundo con representaciones simbólicas. En este artículo, discutimos los principios del DSL y su aplicabilidad en genómica. Examinamos las ventajas y desafíos de su uso en el contexto del variant calling y destacamos la importancia de una validación meticulosa. Para garantizar la calidad de los resultados, es esencial adoptar técnicas de validación apropiadas y herramientas de software específicas. Ofrecemos una visión detallada de estas técnicas y herramientas, con el objetivo de establecer estándares claros para la implementación y validación de algoritmos de DSL en pipelines genómicos. Esta investigación subraya el potencial del DSL para mejorar la precisión en el descubrimiento de variantes, ofreciendo perspectivas prometedoras para la genómica del futuro.

Palabras Clave: Next-Generation sequencing, Validación, Deep Symbolic Learning

Introducción

El vasto universo del genoma humano ha sido objeto de fascinación y estudio desde la concepción de la biología molecular. Con la aparición de la tecnología de secuenciación de nueva generación (NGS, por sus siglas en inglés), se ha abierto una ventana sin precedentes al estudio detallado de nuestra información genética, permitiendo descifrar misterios relacionados con enfermedades, evolución y otros aspectos fundamentales de la biología. La NGS ha impulsado un crecimiento exponencial en la generación de datos genómicos, convirtiendo el análisis de estos datos en una tarea compleja pero crucial.

Dentro de este análisis, una de las fases más relevantes es el "variant calling", o detección de variantes. Esta etapa consiste en identificar las diferencias entre una muestra genómica y una referencia, permitiendo así reconocer mutaciones, polimorfismos y otras variantes genéticas. Estas variantes son esenciales para entender la diversidad genética, la predisposición a enfermedades y, en muchos casos, la respuesta a tratamientos. Una identificación precisa y confiable de variantes es, por tanto, esencial para traducir la vasta cantidad de datos generados por NGS en información biológicamente relevante y aplicable.

Sin embargo, la tarea no es sencilla. Las variaciones genéticas pueden ser sutiles o estar en regiones del genoma de difícil secuenciación. Además, los errores intrínsecos de las técnicas de secuenciación, así como la complejidad inherente de la biología molecular, presentan desafíos en la detección precisa de variantes. Es aquí donde las técnicas computacionales avanzadas, como el Deep Symbolic Learning (DSL), se presentan como herramientas prometedoras, capaces de integrar y aprender de grandes volúmenes de datos genómicos, y ofrecer resultados precisos y confiables en la tarea de variant calling.

A través de este artículo, exploraremos el potencial del DSL en el contexto del variant calling, delineando su integración en un pipeline de análisis de datos genómicos y discutiendo las técnicas de validación necesarias para asegurar la calidad de sus resultados.

La tecnología de secuenciación de nueva generación, comúnmente conocida como NGS (del inglés "Next-Generation Sequencing"), representa una revolución en la manera de analizar y comprender la información genética. Antes de la llegada de NGS, la secuenciación de ADN se realizaba principalmente mediante la técnica de Sanger, un proceso costoso y relativamente lento que permitía secuenciar fragmentos pequeños de ADN de manera individual. Si bien esta técnica fue pionera y fundamental para proyectos emblemáticos como la secuenciación del genoma humano, la creciente necesidad de analizar múltiples muestras y grandes regiones genómicas requería de una tecnología más eficiente.

NGS surge como respuesta a esta necesidad. A diferencia de la secuenciación de Sanger, NGS permite secuenciar millones de fragmentos de ADN de manera simultánea, en un solo experimento. Esta capacidad de "alto rendimiento" ha reducido drásticamente los costos y tiempos de secuenciación, permitiendo que la genómica se vuelva más accesible y que se generen enormes volúmenes de datos en tiempos relativamente cortos.

El proceso general de NGS comienza con la preparación de una biblioteca de fragmentos de ADN que se desean secuenciar. Estos fragmentos se adhieren a una superficie y se amplifican para generar colonias de ADN clonal, cada una correspondiente a una única molécula de ADN inicial. Posteriormente, la secuenciación se realiza ciclo tras ciclo, incorporando nucleótidos marcados fluorescentemente. Al incorporarse, cada nucleótido emite una señal lumínica específica, que es detectada y traducida para determinar la secuencia del fragmento.

Existen diversas plataformas y metodologías dentro del espectro de NGS, como Illumina, 454 de Roche y Ion Torrent, cada una con sus particularidades, ventajas y limitaciones. Sin embargo, todas comparten el principio básico de secuenciar múltiples fragmentos en paralelo, lo que otorga a NGS su poder y versatilidad.

El advenimiento de NGS ha impulsado avances significativos en áreas como la genómica personalizada, la metagenómica, la genómica evolutiva, entre otras. Sin embargo, el aluvión de datos que produce ha planteado retos computacionales y analíticos, haciendo esencial la integración de herramientas avanzadas como el Deep Symbolic Learning para extraer el máximo valor de la información genética obtenida.

El Deep Symbolic Learning (DSL) se ha posicionado como un enfoque innovador en el mundo del aprendizaje automático y la inteligencia artificial, fusionando dos paradigmas tradicionalmente distintos: el aprendizaje profundo y el aprendizaje simbólico.

El aprendizaje profundo, representado principalmente por las redes neuronales profundas, es conocido por su habilidad para manejar grandes cantidades de datos, aprender características de forma automática y realizar tareas de predicción con alta precisión. Sin embargo, sus modelos son a menudo vistos como "cajas negras", debido a que la interpretación de sus procesos internos puede resultar compleja.

Por otro lado, el aprendizaje simbólico, que tiene raíces en la lógica y en la inteligencia artificial basada en conocimientos, se centra en la manipulación y razonamiento de símbolos y reglas explícitas. Esta aproximación ofrece transparencia y explicabilidad, pero puede ser limitada en cuanto a la adaptabilidad y generalización ante grandes volúmenes de datos o tareas complejas.

DSL combina lo mejor de ambos mundos. A través de esta fusión, es posible entrenar modelos que no solo sean potentes y precisos, sino también interpretables y basados en reglas y estructuras simbólicas. Esto significa que el DSL puede aprovechar la capacidad del aprendizaje profundo para aprender automáticamente de los datos, al tiempo que retiene una estructura lógica y simbólica que facilita la interpretación y la toma de decisiones basada en reglas claras.

En el contexto de la genómica, esta combinación es particularmente valiosa. Los datos genómicos son inherentemente complejos, con estructuras jerárquicas, relaciones no lineales y patrones sutiles que pueden ser cruciales para tareas como el variant calling. El DSL, al ser capaz de modelar de forma efectiva tanto patrones explícitos como implícitos, se posiciona como una herramienta prometedora para la genómica.

Por ejemplo, mientras que un modelo de aprendizaje profundo podría identificar patrones complejos en secuencias genómicas que indican la presencia de una variante, la componente simbólica podría representar reglas biológicas conocidas, como las implicaciones funcionales de una mutación en una región particular del genoma. Esta combinación permite que el proceso de análisis sea a la vez adaptable y basado en conocimientos previamente

establecidos, optimizando así la precisión y la interpretabilidad de los resultados.

Dada la creciente necesidad de métodos computacionales robustos y explicables en el análisis genómico, el DSL se presenta como una solución potencialmente revolucionaria, ofreciendo un equilibrio entre potencia analítica y transparencia en el proceso de toma de decisiones.

Fundamentos del DSL

El Deep Symbolic Learning (DSL) surge como respuesta a una necesidad palpable en el ámbito de la inteligencia artificial: combinar la eficiencia y capacidad de generalización del aprendizaje profundo con la explicabilidad y estructura del aprendizaje simbólico.

El aprendizaje profundo se basa en redes neuronales con múltiples capas, conocidas como redes neuronales profundas. Estas redes son capaces de modelar relaciones complejas y no lineales entre las entradas y las salidas, y se han destacado, especialmente en los últimos años, por su rendimiento en tareas de procesamiento de imágenes, lenguaje natural, entre otras. Su fuerza radica en la capacidad de aprender características y representaciones de los datos de forma automática y jerárquica, es decir, sin requerir una ingeniería manual de características. No obstante, la interpretación de lo que estas redes "aprenden" puede ser opaca, dando lugar al término "caja negra".

El aprendizaje simbólico, por su parte, es una rama de la inteligencia artificial que utiliza símbolos para representar el conocimiento. A diferencia del aprendizaje profundo, que suele ser inductivo (aprende de los datos), el aprendizaje simbólico es principalmente deductivo (aprende de reglas preestablecidas y hechos). En este paradigma, el conocimiento se representa mediante reglas, hechos y símbolos, permitiendo un razonamiento lógico y una fácil interpretación. El aprendizaje simbólico ha sido fundamental en áreas donde la explicabilidad es esencial, como en sistemas expertos y en la toma de decisiones basada en reglas.

DSL busca integrar estos dos enfoques para aprovechar sus fortalezas complementarias. En un sistema DSL típico, las representaciones aprendidas por las redes neuronales profundas son traducidas o transformadas en estructuras simbólicas. Estas estructuras pueden ser reglas, árboles de decisión, o cualquier otra representación que permita una interpretación clara. A la inversa, el conocimiento simbólico también puede ser incorporado en el aprendizaje profundo, guiando o restringiendo el proceso de aprendizaje de la red neuronal.

La esencia del DSL radica en su capacidad para aprender de forma end-to-end, es decir, desde los datos crudos hasta reglas y representaciones simbólicas, sin perder la potencia del aprendizaje profundo ni la claridad del aprendizaje simbólico. Esta integración se logra mediante diversos mecanismos, como la regularización basada en conocimiento, la conversión de redes neuronales en árboles de decisión, entre otros.

En resumen, el DSL representa una fusión prometedora de dos mundos, permitiendo modelar la complejidad inherente de los datos reales, al tiempo que proporciona interpretaciones claras y basadas en reglas, facilitando la toma de decisiones y la confiabilidad en diversas aplicaciones.

Ventajas y desventajas del DSL en el contexto de la genómica

La genómica es una disciplina que, por su naturaleza, genera volúmenes masivos de datos. Estos datos son intrincados y poseen estructuras complejas, patrones sutiles y relaciones a múltiples niveles. Al abordar la genómica desde el prisma del Deep Symbolic Learning (DSL), se obtienen una serie de ventajas y desventajas notables:

Ventajas:

1. Modelado de Relaciones Complejas: Los datos genómicos a menudo contienen relaciones no lineales, jerárquicas y multifactoriales. El componente de aprendizaje profundo en el DSL permite capturar y modelar estas relaciones intrincadas con gran precisión.

2. Interpretabilidad: Una de las críticas más comunes a los modelos de aprendizaje profundo en genómica es su falta de explicabilidad. Con DSL, se integra el razonamiento simbólico, ofreciendo interpretaciones claras en términos de reglas o estructuras simbólicas, lo cual es crucial para entender implicaciones biológicas o genéticas.

3. Incorporación de Conocimiento Previo: La genómica tiene décadas de investigación y conocimiento acumulado. DSL permite la integración de este conocimiento en forma de reglas o hechos, lo que puede guiar o mejorar el proceso de aprendizaje y predicción.

4. Flexibilidad: DSL es adaptable. Puede sintonizarse para priorizar precisión (aprendizaje profundo) o explicabilidad (aprendizaje simbólico) según las necesidades del análisis genómico en cuestión.

Desventajas:

1. Complejidad Computacional: Fusionar aprendizaje profundo con aprendizaje simbólico puede incrementar la complejidad computacional, requiriendo más recursos y tiempo, especialmente con datasets genómicos de gran tamaño.

2. Desafíos en la Integración: No siempre es sencillo combinar conocimientos simbólicos con modelos profundos. Esto puede llevar a situaciones en las que el modelo no converge o no logra representar adecuadamente el conocimiento previo.

3. Sobreajuste a Reglas: Al integrar reglas y hechos conocidos, existe el riesgo de que el modelo se ajuste demasiado a estas reglas y pierda capacidad de generalizar a nuevos datos o situaciones no contempladas en el conocimiento previo.

4. Necesidad de Expertise Doble: Implementar DSL en genómica requiere una comprensión profunda tanto del aprendizaje automático como de la genética y biología molecular. Esto puede hacer que la barrera de entrada sea más alta en comparación con usar técnicas más convencionales.

En conclusión, mientras que el DSL ofrece un potencial significativo para abordar los desafíos del análisis genómico, también presenta desafíos inherentes a su naturaleza híbrida. Como con cualquier herramienta, su eficacia dependerá en gran medida de la correcta implementación y adaptación al problema específico en cuestión.

Aplicación del DSL en la fase de variant calling en NGS

El variant calling es un proceso esencial en el análisis de datos genómicos, que identifica variantes (como SNPs y indels) a partir de datos de secuenciación. Dada la naturaleza crucial de esta tarea, es imperativo contar con herramientas precisas, robustas y explicables. El Deep Symbolic Learning (DSL) ofrece un enfoque prometedor en este ámbito, y a continuación, detallamos su aplicación en la fase de variant calling.

1. Modelado de Características Genómicas:

Las secuencias genómicas presentan una diversidad de características que pueden ser cruciales para identificar variantes. Mientras que el aprendizaje profundo en DSL puede extraer automáticamente características de los reads y su contexto, el componente simbólico puede representar conocimientos biológicos previos, como regiones conservadas, sitios de splicing o zonas propensas a mutaciones.

2. Incorporación de Conocimiento Previo:

Existen reglas y hechos en genética que son ampliamente conocidos. Por ejemplo, ciertas variantes en regiones específicas pueden ser patogénicas o tener un efecto funcional determinado. DSL permite la integración de este conocimiento en el proceso de variant calling, dando prioridad o asignando confiabilidad a variantes basadas en hechos biológicos previamente establecidos.

3. Razonamiento Probabilístico y Determinación de Variantes:

El proceso de variant calling no es siempre binario. A veces, la evidencia puede ser ambigua o insuficiente para hacer una determinación clara. DSL, al combinar el razonamiento simbólico con redes neuronales, puede proporcionar un marco de decisión basado en probabilidades, ponderando la evidencia en los datos y el conocimiento previo.

4. Evaluación y Validación de Llamados:

Una vez que se realiza el variant calling, es vital validar la precisión y confiabilidad de los llamados. Con DSL, esta validación no solo puede basarse en métricas de desempeño, sino también en la interpretación y coherencia de las reglas y representaciones simbólicas derivadas.

5. Explicabilidad y Reporte:

Finalmente, una vez identificadas las variantes, los investigadores o profesionales de la salud pueden requerir explicaciones claras sobre cómo y por qué se determinó una variante específica. El componente simbólico de DSL facilita esta tarea, ofreciendo justificaciones basadas en reglas o hechos que son fácilmente comprensibles.

En resumen, la aplicación de DSL en el variant calling en NGS promete mejorar no solo la precisión y robustez del proceso, sino también su interpretabilidad y confiabilidad. Al integrar el aprendizaje automático avanzado con el razonamiento lógico y simbólico, se abre la puerta a un análisis genómico más informado y confiable.

Técnicas de validación

Métodos de validación tradicionales en variant calling

La validación es un paso crucial en el proceso de variant calling, ya que garantiza que las variantes identificadas son auténticas y no artefactos o errores del proceso de secuenciación o análisis. En el contexto del variant calling, existen varios métodos de validación tradicionales que han sido ampliamente utilizados:

1. Validación por Replicación: Consiste en secuenciar la misma muestra varias veces y comparar los resultados. Si una variante se identifica consistentemente en replicados independientes, es probable que sea genuina. Esta técnica es directa pero puede ser costosa en términos de tiempo y recursos.

2. Comparación con Bases de Datos de Referencia: Se compara el conjunto de variantes identificadas con bases de datos de variantes conocidas, como dbSNP o gnomAD. Si una variante ya ha sido reportada y validada previamente, hay una alta probabilidad de que sea real. Sin embargo, esto no garantiza la identificación de variantes nuevas o raras.

3. Análisis de Homocigosidad: Las variantes que ocurren en regiones de homocigosidad (regiones del genoma donde un individuo tiene dos copias idénticas) pueden ser más fáciles de validar porque cualquier variante heterocigota en estas regiones es probablemente un artefacto.

4. Técnicas de Sanger: La secuenciación de Sanger se considera el "estándar de oro" para la validación de variantes. Consiste en resecuenciar una región específica usando la metodología de Sanger y comparar los resultados con los obtenidos por NGS. Si ambas técnicas identifican la misma variante, se considera validada.

5. Análisis de Calidad y Cobertura: Evaluar la calidad de las bases y la cobertura en las posiciones donde se detectaron variantes. Las variantes identificadas en regiones con baja calidad o cobertura deben tratarse con precaución, ya que tienen una mayor probabilidad de ser artefactos.

6. Herramientas Bioinformáticas: Existen diversas herramientas y software diseñados específicamente para la validación de variantes, como GATK's VariantFiltration, que aplican una serie de filtros y criterios para discernir entre variantes reales y artefactos.

Estos métodos tradicionales han sido pilares en la validación de variant calling. Sin embargo, con la incorporación de técnicas avanzadas como el Deep Symbolic Learning, es posible que se requieran o se puedan desarrollar nuevas estrategias de validación, que combinen la robustez de los métodos tradicionales con las ventajas del aprendizaje automático y el razonamiento simbólico.

Particularidades de la validación cuando se usa DSL

Al incorporar el Deep Symbolic Learning (DSL) en el proceso de variant calling, emergen nuevas dimensiones y desafíos en la validación. A continuación, se describen algunas de las particularidades y consideraciones especiales al validar variantes identificadas mediante DSL:

1. Validación de Reglas Simbólicas: DSL, al generar reglas o estructuras simbólicas, requiere una validación de estas representaciones. Es fundamental asegurarse de que las reglas generadas tengan sentido biológico y no sean producto de sobreajuste o artefactos del modelo.

2. Interacción entre Componentes Profundos y Simbólicos: En DSL, los componentes de aprendizaje profundo y simbólico interactúan. Al validar, es crucial entender cómo estas interacciones afectan la identificación de variantes y si introducen sesgos o errores.

3. Sensibilidad a Datos de Entrenamiento: Como cualquier modelo de aprendizaje automático, los modelos basados en DSL son sensibles a los datos con los que se entrenan. Se deben realizar validaciones cruzadas y pruebas con conjuntos de datos independientes para garantizar la generalización del modelo.

4. Explicabilidad versus Precisión: Puede surgir un equilibrio entre la precisión del modelo y su capacidad de explicación. En la validación, es esencial considerar ambos aspectos y determinar si se está sacrificando uno en detrimento del otro.

5. Validación Iterativa: Dado que DSL combina técnicas de aprendizaje automático con razonamiento simbólico, puede ser beneficioso adoptar un enfoque iterativo de validación. Por ejemplo, tras una primera ronda de variant calling, se pueden refinar las reglas simbólicas y reentrenar el modelo para una segunda ronda, mejorando así la precisión y confiabilidad.

6. Evaluación de Ambigüedades: El razonamiento simbólico en DSL puede introducir o resolver ambigüedades en el proceso de variant calling. Durante la validación, es crucial identificar y tratar estas ambigüedades de manera adecuada.

7. Uso de Métricas de Validación Complementarias: Además de las métricas tradicionales de validación, como la precisión, recall y F1-score, puede ser útil introducir métricas que evalúen la calidad y coherencia de las representaciones simbólicas generadas.

En resumen, aunque el DSL presenta un potencial prometedor para el variant calling, también introduce particularidades en la validación que deben abordarse con cuidado. Estas consideraciones garantizan que los resultados sean no solo precisos sino también biológicamente significativos y explicables.

Consideraciones sobre el conjunto de datos de validación

Al introducir técnicas avanzadas como el Deep Symbolic Learning (DSL) en el proceso de variant calling, el conjunto de datos de validación adquiere una importancia aún mayor.

Estas son algunas consideraciones clave al seleccionar y trabajar con conjuntos de datos de validación en este contexto:

1. Diversidad Genómica: Es esencial que el conjunto de datos de validación refleje la diversidad genómica de la población estudiada. Esto garantiza que el modelo puede identificar variantes en una amplia variedad de contextos genómicos y no solo en aquellos presentes en el conjunto de entrenamiento.

2. Tamaño del Conjunto de Datos: Aunque es tentador optar por conjuntos de datos más grandes debido a la percepción de que proporcionan resultados más robustos, es crucial que estos datos sean de alta calidad. A veces, es preferible un conjunto de datos más pequeño pero meticulosamente curado.

3. Conjuntos de Datos Independientes: Para evitar el sobreajuste y garantizar la generalización del modelo, es vital utilizar conjuntos de datos de validación que sean completamente independientes del conjunto de entrenamiento.

4. Variantes Conocidas versus Desconocidas: El conjunto de validación debe contener tanto variantes previamente identificadas y validadas (para evaluar la precisión del modelo) como nuevas variantes desconocidas (para evaluar la capacidad del modelo de identificar variantes no vistas previamente).

5. Representatividad Biológica: Además de las variantes, es importante que el conjunto de datos de validación refleje otros aspectos biológicos del genoma, como regiones de alta conservación, zonas propensas a mutaciones o regiones reguladoras.

6. Metadatos Asociados: Siempre que sea posible, es útil tener acceso a metadatos asociados con las muestras en el conjunto de validación. Estos pueden incluir información sobre la condición clínica del individuo, su fenotipo o cualquier tratamiento recibido. Estos metadatos pueden ayudar a interpretar los resultados y proporcionar contextos adicionales.

7. Errores y Artefactos Conocidos: Es beneficioso estar al tanto de los errores y artefactos asociados con las tecnologías de secuenciación utilizadas en el conjunto de datos de validación. Estos pueden influir en la interpretación de los resultados y deben ser considerados al validar las variantes identificadas.

8. Flexibilidad en la Evaluación: Debido a las particularidades del DSL, puede ser necesario adaptar o desarrollar nuevos protocolos de validación. El conjunto de datos de validación debe ser lo suficientemente flexible para permitir estos enfoques adaptativos.

En definitiva, el conjunto de datos de validación no es solo una herramienta para evaluar la eficacia de un modelo, sino una pieza clave que, si se selecciona y utiliza adecuadamente, puede mejorar significativamente la precisión y confiabilidad del proceso de variant calling en un contexto DSL.

Métricas y herramientas de evaluación

La validación cuantitativa de las llamadas de variantes es crucial para garantizar la calidad y confiabilidad de los resultados. Las métricas de evaluación proporcionan una forma de medir la eficacia del proceso de variant calling. Aquí, describimos algunas de las métricas clave utilizadas en este contexto:

1. Precisión (Accuracy): Esta métrica evalúa la proporción de predicciones correctas (tanto variantes verdaderas como no variantes) en relación con todas las predicciones realizadas. Es una métrica global que considera tanto los positivos como los negativos verdaderos.

$$\mathrm{P} = \frac{\text{Positivos Verdaderos} + \text{Negativos Verdaderos}}{\text{Total de Predicciones}}$$

2. Sensibilidad (Recall o Tasa Verdadera Positiva): Mide la proporción de variantes reales que fueron correctamente identificadas por el modelo en relación con todas las variantes reales presentes en los datos.

$$\text{Sensibilidad} = \frac{\text{Positivos Verdaderos}}{\text{Positivos Verdaderos} + \text{Falsos Negativos}}$$

3. Especificidad (Tasa Verdadera Negativa): Evalúa la proporción de no variantes que fueron correctamente identificadas en relación con todas las no variantes presentes.

$$\text{Especificidad} = \frac{\text{Negativos Verdaderos}}{\text{Negativos Verdaderos} + \text{Falsos Positivos}}$$

4. Valor Predictivo Positivo (VPP): Esta métrica indica la proporción de variantes identificadas que son verdaderas variantes en relación con todas las variantes identificadas.

$$\mathrm{VPP} = \frac{\text{Positivos Verdaderos}}{\text{Positivos Verdaderos} + \text{Falsos Positivos}}$$

5. Valor Predictivo Negativo (VPN): Indica la proporción de no variantes identificadas que son verdaderas no variantes en relación con todas las no variantes identificadas.

$$\text{VPN} = \frac{\text{Negativos Verdaderos}}{\text{Negativos Verdaderos} + \text{Falsos Negativos}}$$

6. F1-Score: Es una métrica que combina la precisión y la sensibilidad en una única cifra. Es especialmente útil cuando las clases están desequilibradas.

$$\text{F1-Score} = 2 \times \frac{\text{Precisión} \times \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}}$$

Estas métricas proporcionan una visión cuantitativa de la calidad de las llamadas de variantes. Sin embargo, es importante tener en cuenta que no hay una única métrica "perfecta" y, a menudo, se deben considerar en conjunto para obtener una visión completa de la eficacia del proceso de variant calling, especialmente en el contexto del Deep Symbolic Learning.

Herramientas de software recomendadas

A medida que el análisis genómico ha avanzado, han surgido diversas herramientas de software diseñadas para facilitar la evaluación y validación de llamadas de variantes. Estas herramientas varían en sus capacidades y aplicaciones, pero muchas de ellas son ampliamente reconocidas en la comunidad genómica. Aquí presentamos algunas de las más recomendadas para validar variantes, especialmente en el contexto de modelos basados en Deep Symbolic Learning:

1. GATK (Genome Analysis Toolkit): Desarrollado por el Broad Institute, GATK es quizás uno de los paquetes de software más populares para el análisis de genomas. Aunque es ampliamente conocido por sus herramientas de variant calling, GATK también ofrece herramientas como `VariantEval` para la evaluación y validación de llamadas de variantes.

2. VCFtools: Esta es una suite de programas para trabajar con archivos VCF (Variant Call Format) y BCF (Binary Call Format). Ofrece utilidades para comparar variantes, analizar sitios y poblaciones de variantes, y evaluar la calidad de las llamadas.

3. bcftools: Similar a VCFtools, bcftools es una colección de utilidades para el análisis de variantes genéticas presentes en el formato VCF/BCF. Es especialmente útil para filtrar y estadísticas.

4. BEDTools: Aunque no se limita estrictamente a la validación de variantes, BEDTools es una poderosa herramienta para trabajar con datos genómicos en varios formatos, incluido BED. Puede ser útil para operaciones como la intersección y comparación de conjuntos de variantes.

5. RTG Tools: Real Time Genomics (RTG) ofrece herramientas que permiten la evaluación rápida y precisa de variantes contra un conjunto de referencia. Su función `vcfeval` es especialmente útil para la evaluación de llamadas de variantes.

6. hap.py: Es una herramienta que compara conjuntos de variantes genómicas para encontrar diferencias en llamadas y genotipos. Se basa en la re-genotipificación de un conjunto de verdad de referencia y produce métricas de concordancia detalladas.

7. DeepVariant: Aunque es una herramienta de variant calling por sí misma, basada en aprendizaje profundo, DeepVariant ofrece capacidades integradas para evaluar la calidad de las llamadas. Su aproximación basada en imágenes puede ser de interés al considerar métodos alternativos de evaluación en el contexto del DSL.

Al seleccionar herramientas para la validación, es crucial considerar la naturaleza de los datos, los requerimientos específicos del proyecto y la experiencia del equipo de investigación. La combinación adecuada de software y métricas puede mejorar significativamente la precisión y confiabilidad de las llamadas de variantes en análisis genómicos basados en Deep Symbolic Learning.

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