Máster en dirección de sistemas y TIC de la salud y en digitalización sanitaria

Curso académico 2023-2024

Trabajo de Fin de Máster

Reflexión estratégica para la aplicación de Inteligencia Artificial a las disciplinas Ómicas

Autores:

María Ana Batlle López

Francisco Javier Tombo Guerner

Enrique Maldonado Belmonte

Tutor:

Raúl Martínez Santiago

Octubre 2024

Trabajo de Fin de Máster

Tipo: Proyecto de investigación

Título: Propuesta de marco de trabajo para la aplicación de Inteligencia Artificial a disciplinas Ómicas.

Autoría: María Ana Batlle López, Francisco Javier Tombo Guerner, Enrique Maldonado Belmonte

Tutoría: Raúl Martínez Santiago

RAUL MARTINEZ SANTIAGO - 20202262M Fecha: 2024.10.29 14:14:17 +01'00'

Firmado por TOMBO GUERNER FRANCISCO JAVIER - ***1715** el día 28/10/2024 con un certificado emitido por AC FNMT Usuarios

Firmado por MALDONADO BELMONTE ENRIQUE - ***2964** el día 29/10/2024 con un certificado emitido por AC FNMT Usuarios

Nos gustaría agradecer, en primer lugar, a nuestros respectivos Servicios de Salud y a los organizadores y profesores el darnos la oportunidad de participar en este máster, que nos ha permitido adentrarnos en el apasionante mundo de la transformación digital en el ámbito de la Salud y establecer contactos con profesionales muy interesantes multidisciplinares con los que ya hemos entablado productivas colaboraciones y lazos de amistad. También agradecemos a nuestro tutor, Raúl Martínez, por guiarnos y asesorarnos en cómo afrontar este trabajo, así como darnos buenos consejos. Y por último a nuestras familias por apoyarnos incondicionalmente.

Índice

1 Resumen 4
2 Introducción 5
3 Fundamentos de Ómica 6
4 Fundamentos, Limitaciones y retos de la genómica 7
5 Conceptos básicos de Inteligencia Artificial 13
5.1 Conceptos generales 13
5.2 Machine Learning 13
5.3 Operativa de trabajo 14
6 Métodos, algoritmos y tecnologías 18
6.1 Métodos y algoritmos 18
6.2 Tecnologías 21
7 Validación y evaluación de modelos 22
7.1 Validación y evaluación general 22
7.2 Consideraciones Específicas para Datos Ómicos 24
7.3 Ejemplos 25
8 Implicaciones éticas 27
8.1 Consideraciones generales 27
8.2 Privacidad y Confidencialidad de los Datos 27
8.3 Equidad y Sesgo. 27
8.4 Transparencia e Interpretabilidad. 28
8.5 Responsabilidad y Gobernanza. 28
9 Marco regulatorio 30
9.1 Introducción 30
9.2 Regulación General de Protección de Datos (GDPR) 30
9.3 Ley de inteligencia artificial (AI Act) 31
9.4 Espacio Europeo de Datos de Salud (EHDS) 32
9.5 Iniciativas internacionales 33
10 Infraestructura tecnológica y de datos 34
11 Retos e incertidumbres de la Aplicación de la IA a los datos ómicos 39
12 Colaboración interdisciplinaria 42
13 Casos de estudio 44
13.1 Secuenciación del ADN 45
13.2 Identificación de variantes (Variant Calling) 45
13.3 Anotación del genoma: Filtrado de variantes y predicción de efectos. 46

13.4 Clasificación de variantes no codificantes 47
13.5 Mapeo fenotipo – genotipo 49
13.6 Análisis de datos multiómicos y multimodales integrados 51
13.7 Otros usos de la IA en genómica 52
14 Futuro de la Inteligencia Artificial en ómica 55
15 Conclusiones 61
Anexo I 63
Bibliografía 65
Figuras 71
Glosario 73

1 Resumen

El presente trabajo se justifica por la necesidad de establecer una reflexión estratégica sobre el abordaje de proyectos de inteligencia artificial en el ámbito de las ciencias ómicas, con el objetivo de abordar los desafíos inherentes a esta aplicación y maximizar su potencial en la investigación biomédica y en la aplicación de la Medicina Personalizada de Precisión.

Con la acelerada generación de datos complejos en ómica, la necesidad de herramientas avanzadas que faciliten el análisis eficiente y riguroso de grandes volúmenes de información es evidente. El estudio abarca el análisis del estado actual de la IA en ómica, junto con los principios básicos de la genómica y otras ciencias ómicas, y plantea una estructura que permita su implementación de forma ética y efectiva.

El trabajo se enfoca en:

- Investigar el estado del arte de la aplicación de la inteligencia artificial a las disciplinas ómicas.
- Analizar los distintos aspectos e implicaciones para afrontar proyectos de inteligencia artificial en este campo, revisando aspectos críticos como la privacidad de los datos genómicos, la equidad y representatividad en los algoritmos, y las implicaciones éticas de su uso. Evalúa asímismo la infraestructura tecnológica necesaria para el procesamiento de datos ómicos, proponiendo soluciones de alto rendimiento y escalabilidad que incluyan computación en la nube e infraestructura híbrida para gestionar el análisis bioinformático y la conservación de datos de manera segura y sostenible.
- Subraya la importancia de la colaboración interdisciplinaria, integrando a genetistas, bioinformáticos, ingenieros y expertos en ética para asegurar un enfoque holístico que fortalezca los resultados y la fiabilidad de las aplicaciones de IA en la medicina de precisión y la medicina personalizada.

Este enfoque permitirá avanzar en el uso de la IA en ciencias ómicas, promoviendo un desarrollo controlado y adaptativo que responda tanto a los desafíos tecnológicos como a los éticos y sociales del campo.

2 Introducción

La Medicina actual, que pone al paciente en el centro, siendo importante su participación en el proceso asistencial (Participativa) y designada como Medicina 6 P es una nueva manera de entender la práctica sanitaria en la que se busca una estrategia proactiva frente a reactiva (Preventiva), obteniendo una atención específica, atendiendo a características singulares de cada paciente o a características de grupos de pacientes (Personalizada). Para ello, deben tenerse en cuenta de forma integrada las características de las personas, según su genética, factores ambientales y estilo de vida. La incorporación del componente Poblacional añadió una perspectiva nueva a la MPP, que permite contextualizar y entender mejor las enfermedades, ya que el estudio de un individuo de forma aislada es limitado al no tener referencias que permitan entender las similitudes y diferencias relevantes en cada contexto. Además, la revolución digital en la medicina ofrece nuevas tecnologías de alto rendimiento (de Precisión), que analizan forma eficiente, eficaz desde distintas perspectivas (Ciencias Ómicas) la biología de las enfermedades, obteniendo datos cada vez más informativos, de alta calidad de cada persona, siendo la integración global de toda la información obtenida, aplicando algoritmos matemáticos e inteligencia artificial clave, para poder tener una visión global holística del paciente. Esto permite llegar a un mayor conocimiento de cómo tratar situaciones específicas de forma individualizada, reduciendo los efectos secundarios y pudiendo incluso adelantarse al desarrollo de la enfermedad (Preventiva). Además, estas metodologías, permiten estudiar simultáneamente múltiples pacientes, lo que además de recudir costes, permite asegurar tiempos de respuesta adecuados.

3 Fundamentos de Ómica

Existen diferentes disciplinas que se encargan del estudio y caracterización biológica del paciente. Frente a los modelos tradicionales donde se analizaban elementos específicos de forma individualizada, las ciencias ómicas se caracterizan realizar el estudio de la totalidad o el conjunto, que pueden ser genes, organismos, sistemas, ecosistemas, etc. Así, por ejemplo, frente al estudio puntual de uno o pocos genes concretos en un paciente, la genómica, aborda el estudio generalizado y simultaneo de varios genes e incluso de varios pacientes.

Podemos dividir las ciencias ómicas en base al objeto de estudio, pero también es interesante conocer las tecnologías que permiten el estudio de las mismas, ya que con una misma tecnología podemos en ocasiones abordar el estudio de varias disciplinas ómicas. Esto explica el diferente desarrollo de unas frente a otras, ya que no todas las tecnologías tienen el mismo grado de desarrollo e implantación y sólo algunas de las ómicas tienen una implantación asistencial significativa, siendo la genómica la que más extensión y aplicabilidad inmediata tiene en el momento actual en los centros de diagnóstico hospitalario. Por este motivo, nos centraremos en esta ciencia de datos como modelo para explicar cómo la IA puede ser de enorme ayuda para su abordaje, pero siendo necesario abordar algunos retos y siendo siempre conocedores de las limitaciones.

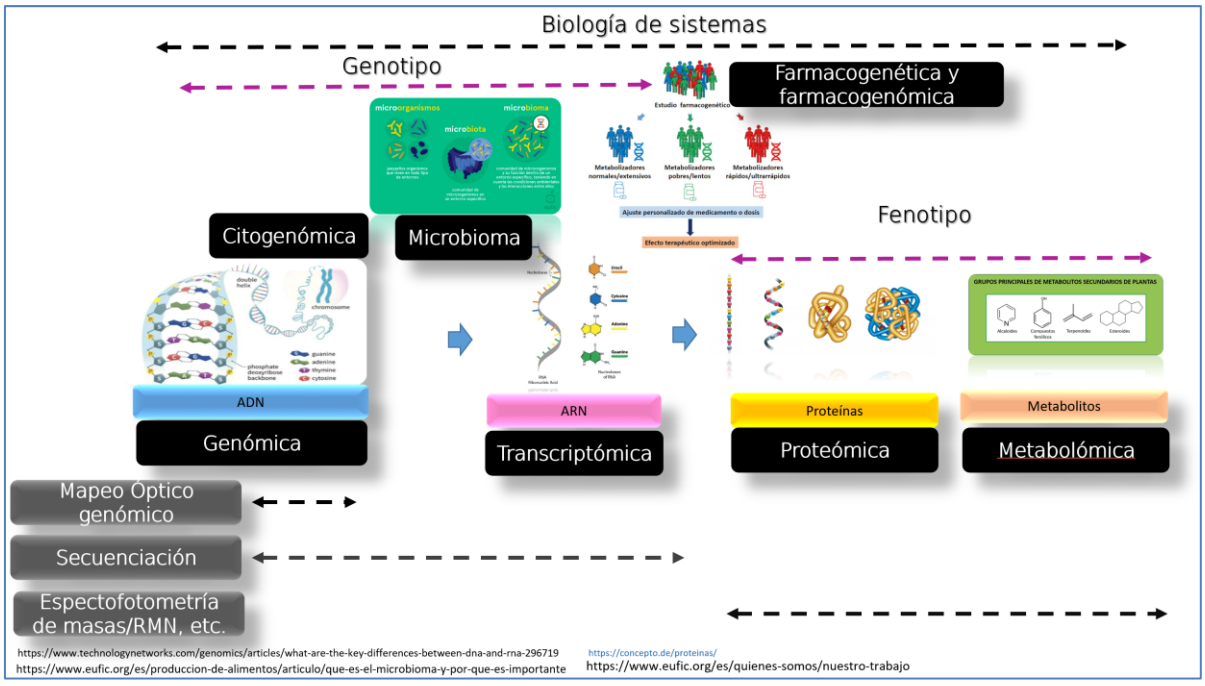

Figura 1. Ciencias ómicas. Gráfico mostrando algunas de las ciencias ómicas atendiendo al tipo de muestra, a las metodologías utilizadas para su estudio y a su vinculación con genotipo y fenotipo. La integración de toda la información requiere la utilización de aproximaciones basadas en Biología de sistemas.

4 Fundamentos, Limitaciones y retos de la genómica

Entendiendo los fundamentos, limitaciones y la aplicación de la genómica podemos ver el gran número de aplicaciones que tiene la IA para su mejor abordaje, pero debiendo considerar una serie de retos que supone su aplicación.

Como cualquier ciencia ómica, la genómica genera una cantidad enorme de datos, cuyo análisis requiere una gran capacidad de cómputo. Los equipos de secuenciación se adquirieron y/o se gestionaron en muchos centros dentro del ámbito de la investigación. En la mayoría de los casos los datos se han ido almacenando en silos, con formatos diversos, sin estar integrados con los sistemas hospitalarios y sin un control de accesos y de gestión de la información rigurosa. Conforme estos recursos han ido pasando a usarse en asistencia, y ante el un incremento exponencial de estudios por la repercusión clínica de la información obtenida, debiendo esta ser interpretada en el menor tiempo posible asegurando la consistencia e integridad de la información en todo momento, se ha ido viendo la necesidad de repensar dicha gestión. Para una gestión eficiente se precisa de hardware capaz de tratar y almacenar este tipo de datos y software capaz de automatizar muchos de los procesos que se realizan. Esto implica un elevado coste (por hacernos una idea, un solo genoma ocupa unos 100-200 GB). Es imposible predecir cuantos genomas se van a analizar en los próximos años. Por tanto, además de disponer del tipo de recursos suficientes, es preciso, tener un sistema fácilmente escalable.

Por otro lado, los datos se consideran altamente sensibles. El conocer si un individuo porta o no una determinada variante puede potencialmente tener enormes repercusiones asistenciales y sociales. En algunos países se habla que las compañías de seguros están planteando usar ciertos datos genómicos para avalar créditos por poner un ejemplo.

Otro reto es meramente metodológico y de limitación de recursos. Las ciencias ómicas, permiten obtener información múltiple compleja que requiere validación (para descartar los errores metodológicos) e interpretación (impacto de dicha información en contexto del paciente). En un exoma (que representa aproximadamente el 2% del genoma y contiene unos 20000 genes,) una vez descartados los errores técnicos, se obtendrán unas 100.000 variantes. Si aplicáramos un genoma completo obtendríamos unos 6 millones de variantes por paciente. De estas sólo una o dos tendrán un significado clínico evidenciado. Por lo tanto, es importante abordar las variantes genómicas identificadas de esta manera con un enfoque de “inocencia hasta que se demuestre lo contrario”.

Para llegar a estas variantes de significado clínico se aplican una serie de pasos complejos- conformando un *pipeline* de análisis- entre los que cabe destacar

- **Análisis primario:** refiriéndose al análisis inicial, en el que la información contenido en la muestra de ADN (información biológica) es transformada a datos digitales. En esta fase hay que

    - Para poder hacer un estudio de secuenciación de primera generación (secuenciación clásica o tipo sanger) o de segunda generación, lo primero que se requiere es descomponer el ADN en fragmentos pequeños. Esta tecnología no es capaz de estudiar las largas moleculares de DNA sin fragmentarlo. Las nuevas generaciones de secuenciación (conocidas como secuenciación de tercera generación) como Nanopore o PacBio son capaces de ir analizando fragmentos cada vez mayores, pero todavía requieren una fragmentación.

    - En la secuenciación clásica y en la de segunda generación, se requiere además que esos micro fragmentos sean enriquecidos (es decir necesitamos varias copias del mismo fragmento para poder analizarlo. Esto se puede hacer por varios métodos, pero fundamentalmente se hacen por PCR o por captura de las secuencias de interés con hibridación con sondas complementarias. En este enriquecimiento es posible también generar errores metodológicos, insertando variantes donde no las hay o puede ocurrir que falle la amplificación o el enriquecimiento de la región de estudio, siendo imposible su análisis. Se precisa conocer que existe esta limitación y existen en las diferentes herramientas en este sentido.

    - Existen varias metodologías y plataformas de secuenciación, tales como Ilumina, Ion Torrent Genexus, BGI, Oxford Nanopore, etc., que utilizan diferentes químicas y presentan unos abordajes diferentes.

- **Análisis Secundario.** Los datos ya secuenciados, son analizados y anotados basándonos en una secuencia considerada de referencia. Esta fase implica los siguientes pasos

    - Control de Calidad: Asegurar la precisión de los datos mediante filtros y controles que eliminan errores de secuenciación o variaciones poco confiables. Con cada plataforma nueva, es necesaria definir bien los errores inherentes a la metodología aplicada y hacer validaciones rigurosas con nuestras muestras. Las validaciones cruzadas, analizando la misma muestra en diferentes carreras con diferentes plataformas son interesantes para descartar este tipo de errores.

    - Alinear el genoma con respecto al genoma de referencia para poder detectar las variaciones de nuestro paciente con respecto a la población. Es decir, *recomponer el puzle*. Para poder determinar que variaciones tiene nuestra muestra, como primer paso, se debe recomponer el genoma y hacer una comparación con respecto a una referencia: El molde que permite la reconstrucción es lo que se denomina genoma de referencia. Las desviaciones

de nuestra secuencia de nucleótidos con respecto a la referencia son lo que conocemos como variantes. Al proceso de detección de estas desviaciones se le conoce como *Variant Calling* (Llamada de variantes). De éstas, algunas no tendrán repercusión alguna (son cambios normales fisiológicos) y otras, serán patológicos. En muchos casos no podremos establecer la repercusión y hablaremos de variantes de significado incierto. Uno de los aspectos que complican es que no se dispone de un genoma de referencia completo que represente bien a todos los seres humanos. Hasta hace no mucho había zonas que de hecho no estaban caracterizadas. ¿Cómo recomponemos un puzle si no tenemos un buen genoma de referencia? Este genoma de referencia ha ido evolucionando. Actualmente, la secuencia de referencia GCRCh37 es la más utilizada; sin embargo, ahora también está disponible GRCh38, que contiene menos secciones de secuencia de ADN desconocida. Puede darse la circunstancia que la recomposición y las variaciones que encontremos hoy con respecto a nuestro genoma de referencia, varíen cuando vaya cambiando nuestro genoma de referencia. Se necesitan sistemas que alerten de estas reclasificaciones y que permitan hacer estos análisis de forma eficiente.

  - Descartar las variantes de baja probabilidad de ser deletéreas por la elevada frecuencia poblacional. Si una variante existe en elevada frecuencia en la población es muy poco probable que constituya una causa de enfermedad. Para esto existen múltiples bases de datos que nos permiten hacer este análisis. Problema, existen discrepancias de unas bases a otras. No todas las bases de datos están curadas (es decir han sido depuradas) y además estas bases se actualizan casi a diario. Actualmente existen softwares que permiten y facilitan este análisis. Pero, de lo mencionado anteriormente, es esencial conocer de que fuentes se nutren nuestros softwares y con que cadencia hacen la consulta para ser fiables. Además, pocos softwares son capaces de analizar datos procedentes de cualquier analizador (muchos están especializados en tecnologías concretas y los pipelines de análisis son diferentes entre plataformas) lo que ha producido que según la patología se tengan que estar aplicando diferentes softwares. Se puede dar el caso de un paciente con dos patologías cuyo análisis se deba llevar a cabo con un software para una patología y con otro para otra patología. Es decir, con mucha frecuencia, en los laboratorios la información está tremendamente atomizada, y la curva de aprendizaje de los genetistas es muy elevada y la inversión de tiempo para el análisis, interpretación e integración de la información muy alta.

- Análisis terciario: Implica darles significado clínico a los hallazgos encontrados, proporcionando contexto a las variantes encontradas, identificando genes afectados, funciones conocidas, y posibles implicaciones biológicas. Esto conlleva varias consideraciones:

- Anotar, establecer asociaciones genotipo- genotipo y determinar la patogenicidad de las variantes (es decir establecer el impacto funcional de esa variante). Para esto existen diferentes predictores y algoritmos que nos permiten establecer la probabilidad de que una variación sea realmente deletérea.

- Con frecuencia sucede que encontramos una variante deletérea pero no explica la enfermedad original que motivó el estudio. Por otro lado, la implicación de una variante puede ser muy diferente en diferentes patologías. En esta fase, se pretende responder a preguntas como, ¿Qué evidencia científica hay de esta variante en la literatura? ¿existe algún tratamiento eficiente o en otras palabras es esta variante accionable? ¿Esta variante está relacionada con alguna enfermedad? Con frecuencia sucede que encontramos una variante deletérea pero no explica la enfermedad original que motivó el estudio. Por otro lado, la implicación de una variante puede ser muy diferente en diferentes patologías. Para poder establecer con certeza que una variante es causa de enfermedad, se precisan procesos de validación biológicos mediante estudios en modelos celulares y animales y estos resultados, deben ser contrastados y publicados: Esto es tremendamente laborioso, consume muchos recursos y es por ello por lo que muchas veces detectamos variantes para las cuales no hay validaciones funcionales. Otras veces ocurre que las haya, pero no están volcadas en las bases de datos que utilizamos o las validaciones realizadas no son rigurosos o diferentes grupos encuentran resultados contradictorios.

- Un aspecto importante, es que en general tendemos a analizar variante a variante y gen a gen de forma independiente y por separado. En cáncer es muy frecuente que encontremos varias mutaciones deletéreas en un mismo caso y muestra. Los nuevos modelos computacionales y biológicos nos están enseñando que el impacto de las mutaciones puede ser muy variable según se combinen con unas variantes u otras. Incluso el orden de adquisición de las variantes puede tener diferente impacto. La generación de información por tanto es exponencial, la capacidad de validación va aumentando a un ritmo constante y la distancia que separa ambas es cada vez mayor. Existen modelos de IA muy interesantes en este sentido que permiten hacer predicciones de una forma mucho más eficiente.

- Integrar la información con resto de sistemas. Ninguna ciencia ómica es capaz por si sola de brindarnos toda la información que precisamos para un abordaje personalizado. Se precisa una integración de toda la información de las diferentes ómicas junto con resto de información sociodemográfica, clínica y biológica del paciente, para realmente hacer un abordaje personalizado. Esta integración, con esta cantidad ingente de datos sólo es posible aplicando algoritmos complejos y biología de sistemas. Para poder hacer esta

interrelación, como hemos visto no sólo es necesaria la interoperabilidad local, sino que se precisa una interoperabilidad, semántica, sintáctica securizada, federada que permita de forma eficiente realizar un estudio personalizado. Existen ya varios estándares tanto semánticos (HUGO, HGNC, ACGM) como de interoperabilidad basados en HL7 FHIR y de persistencia basados en OpenEHR y GA4GH-Beacon.

Para el abordaje de los diferentes pasos, la IA tanto los modelos supervisados como los generativos es esencial. De no utilizarla, no podremos asegurar una asistencia de calidad ya que resulta materialmente imposible hacer un abordaje manual eficiente en el momento actual. Se requiere un nivel de recursos humanos cualificados y que, además deben de estar constantemente actualizados, es muy elevado.

Afortunadamente, empiezan a existir soluciones de mercado que permiten una gestión global de los datos, e incluso con alertas de reclasificación que, además, son capaces de interoperar según algunos de los estándares mencionados y que incluso permiten integrar datos de otras ciencias ómicas como cito genómica, etc., lo cual es necesario para un abordaje basado en medicina personalizada de precisión. Sin embargo, dichos sistemas no son perfectos y no siempre proporcionan toda la información esperada o correcta. Es importante ser conocedor de como obtiene esa información y que lógica aplican para poder conocer las limitaciones y aplicar criterios rigurosos de validación.

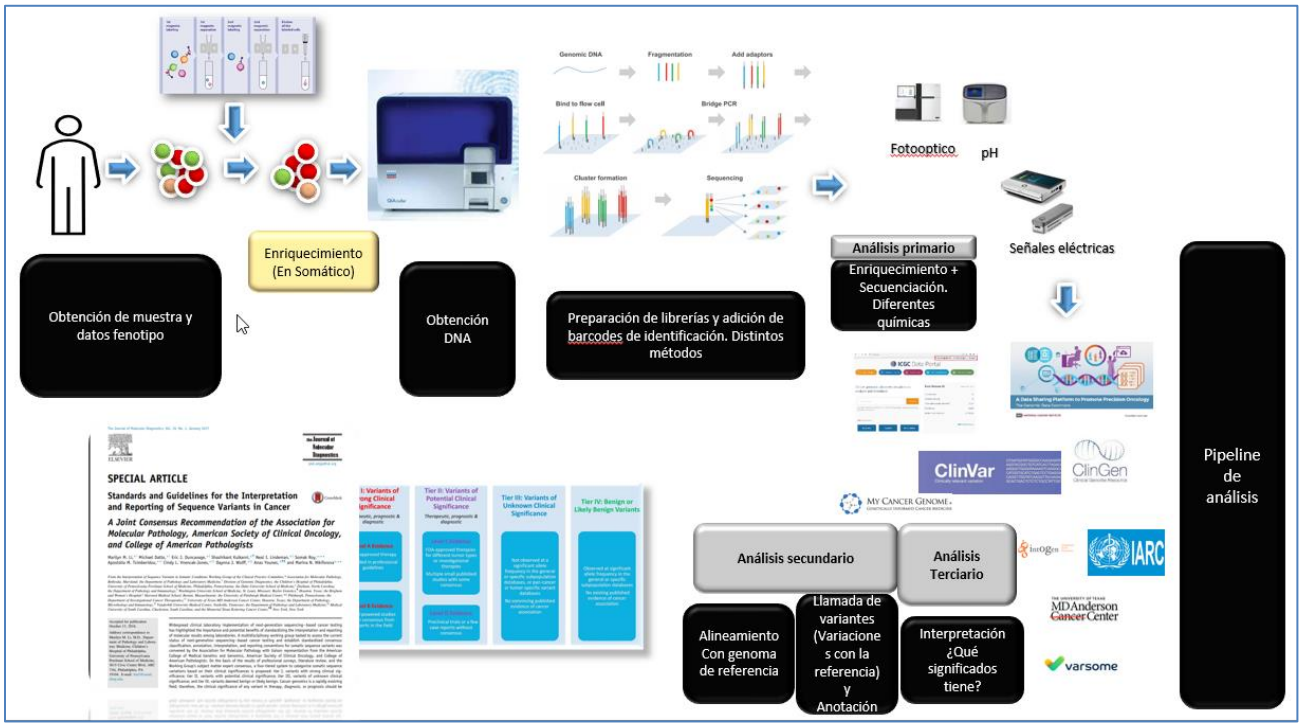

Figura 2. Pipeline de estudios genómico. Esquema de los diferentes pasos del proceso de análisis genómico. Existen varios métodos de enriquecer las células a estudio (esto es especialmente relevante en ámbito somático, donde no todas las células van a tener las mismas variantes). Así mismo, se contemplan diferentes formas de extraer el DNA y de enriquecer en las regiones de interés del genoma (este enriquecimiento es necesario, para poder llevar a cabo la secuenciación de primero o segunda generación. Para el análisis secundario (fase de alineamiento y llamada de variantes) y terciario (anotación interpretación de las variantes clínicas relacionándolo con datos fenotípicos del paciente y estableciendo la accionabilidad e impacto clínico) es preciso consultar varias bases de datos.

5 Conceptos básicos de Inteligencia Artificial

5.1 Conceptos generales

El origen del concepto de inteligencia artificial es difuso y discutible. Ya en la antigua Grecia se contemplaba la posibilidad de que la capacidad de pensar se simulara mediante un modelo formado por un conjunto reglas. En 1936 Alan Turing diseñó su máquina, capaz de interpretar instrucciones. Y finalmente en los años 50 se utiliza formalmente el término "inteligencia artificial" en una conferencia.

Parte del problema es saber qué es la Inteligencia Artificial. La RAE la define como "Disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico.". No es una definición incorrecta, pero peca de genérica y acepta como inteligencia artificial múltiples sistemas de información y algoritmos tradicionales. McKinsey la define como “La capacidad de una máquina para realizar funciones cognitivas que asociamos a la mente humana, como percibir, razonar, aprender, interactuar con el entorno y resolver problemas o incluso utilizar la creatividad”.

Un sistema informático que presenta inteligencia artificial es un sistema que recibe una entrada de información, la procesa, y obtiene una o más conclusiones. La diferencia principal entre un sistema algorítmico convencional y un sistema inteligente es que, el primer caso sigue una serie de reglas definidas que permite hacer un seguimiento de la operativa y entender cómo se llega a la conclusión, mientras que en el sistema inteligente el sistema obtiene conclusiones que un humano no es capaz de trazar y replicar, o al menos, no siguiendo la misma lógica que ha empleado el sistema

5.2 Machine Learning

La inteligencia artificial comprende un área de conocimiento muy amplia. Dentro de la misma, la principal tendencia es el Machine Learning o aprendizaje automático, que a su vez incluye el Deep Learning o aprendizaje profundo.

Machine Learning, por tanto, es un conjunto de técnicas y algoritmos que permiten detectar patrones y aprender cómo hacer predicciones y recomendaciones mediante el procesamiento de datos, en vez de recibir instrucciones explícitas. Los propios algoritmos pueden adaptarse en respuesta a nuevos datos para mejorar su eficacia progresivamente.

Los principales tipos de análisis que se realizan con Machine Learning son la descripción (saber qué ha ocurrido), la predicción (prever qué va a ocurrir) y la prescripción (sugerir qué acción tomar para maximizar la posibilidad de alcanzar un objetivo).

Deep Learning a su vez es un subtipo de Machine Learning, que requiere menor participación humana para su preparación, posee una complejidad técnica mayor, y puede obtener resultados más precisos, aunque esto último depende del problema que se quiera afrontar.

Las principales utilidades de Deep Learning son la clasificación de imágenes, el reconocimiento facial y el reconocimiento de voz.

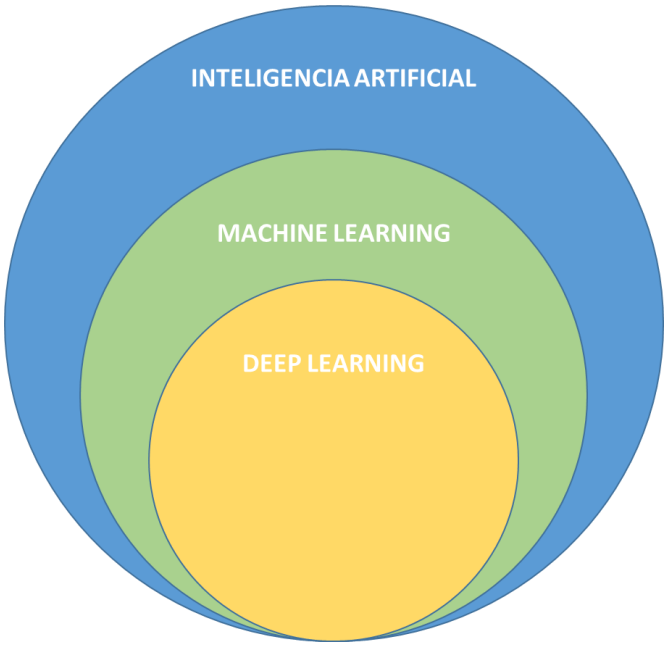

Figura 3. Esquema conceptual de inteligencia artificial y aprendizaje automático

Dentro del aprendizaje, podemos diferenciar dos tipos distintos dependiendo de la información explicativa (etiquetado) de los datos que se usan para el entrenamiento.

Por una parte, en el aprendizaje supervisado el modelo de Inteligencia Artificial se entrena con datos etiquetados. Esto significa que cada entrada tiene una salida esperada asociada, lo que permite que el modelo aprenda de los datos y haga predicciones más precisas. Este enfoque es útil cuando se tiene una gran cantidad de datos previamente clasificados, como la identificación de variantes genéticas que ya se conocen.

Por otra parte, en el aprendizaje no supervisado, el modelo trabaja con datos no etiquetados. El objetivo es identificar patrones ocultos o agrupaciones dentro de los datos sin la guía de resultados predefinidos. Es útil en ciencias ómicas cuando se quiere descubrir nuevas correlaciones o subgrupos dentro de un conjunto de datos, como en el caso de análisis de expresión génica sin un diagnóstico previo.

5.3 Operativa de trabajo

El proceso para llevar a cabo un sistema de aprendizaje automático es habitualmente el mismo, independientemente del tipo de datos que se manejen o la información que se

quiera obtener, y supone los siguientes pasos.

1. Análisis del objetivo. Debe analizarse qué información se quiere obtener. Parece un paso obvio, pero es fácil perderse más adelante en algoritmos y fórmulas si no se tiene claro exactamente qué se está buscando. Debe ser concreto, si es necesario más adelante se puede pivotar y adaptar objetivos, pero no debe empezarse a trabajar a ciegas.

2. Identificación de fuentes de datos. Pueden ser múltiples y heterogéneas, por ejemplo, obtener datos de imágenes de cámaras de video, de boletines oficiales o consultar bases de datos. Es necesario considerar las implicaciones legales del acceso y uso de esta información, por lo que se recomienda revisar la normativa de protección de datos en caso de duda.

3. Carga de los datos. Habitualmente en un dataset o entidad de almacenamiento temporal para poder trabajar con ellos.

4. Preprocesamiento de datos. Se realizan adaptaciones y normalizaciones en los datos obtenidos, se eliminan los campos que no ofrezcan utilidad para nuestro objetivo, y se eliminan aquellos registros que supongan un problema más que una ayuda (que se estime que tienen datos erróneos o incompletos, por ejemplo).

5. División de datos. El proceso normal es seleccionar un bloque de datos (entre el 70% y 80%) para generar el modelo, y el porcentaje restante como forma de validar la corrección del mismo.

6. Selección de algoritmo. Se decide qué técnica de Machine o Deep Learning es más adecuada para el problema.

7. Configuración de hiperparámetros. Dependiendo de la técnica, es necesario especificar valores técnicos para la aplicación del algoritmo: número de niveles, de iteraciones, peso asignado a campos de información, etc.

8. Aplicación del algoritmo. Se lleva a cabo sobre el bloque de datos de procesamiento que hemos dividido en el punto 5 de división de datos. Es la parte más compleja y a la vez la más sencilla. La más compleja por el peso matemático que hay detrás, pero la más sencilla porque es la invocación de los métodos que deben aplicarse, sin necesidad de intervención humana.

9. Obtención del modelo. Se genera un modelo, básicamente una serie de fórmulas que suponen una caja negra para llevar a cabo la predicción o clasificación de datos futuros.

10. Prueba del modelo. Se valida el nivel de éxito, aplicando el modelo de forma

automática sobre los datos de prueba obtenidos en el punto 5 de división de datos. El resultado dependerá de la calidad de los datos, del Preprocesamiento del mismo, de la configuración de hiperparámetros y del algoritmo aplicado.

11. Valoración. Debe valorarse si el resultado obtenido en la prueba del modelo es suficientemente bueno. La bondad del porcentaje de éxito obtenido dependerá del objetivo que se busque, debe ser realista, no será del 100% de aciertos, pero debe ser útil (si es del 50%, estamos realizando mucho esfuerzo cuando tirar una moneda al aire tiene la misma tasa de acierto). Aunque depende mucho del caso, un porcentaje inferior al 80% suele considerarse insuficiente, y del 95% muy bueno. Pero es una generalización. Si el resultado de la valoración no es positivo, debe realizarse el proceso de nuevo, "jugando con los datos" de forma iterativa. Para ello se realizan cambios en el preprocesamiento de datos, la división de datos, eligiendo otro algoritmo o cambiando los hiperparámetros.

12. Aplicación del modelo. El modelo obtenido se utiliza con nuevos sets de datos, obteniendo resultados nuevos.

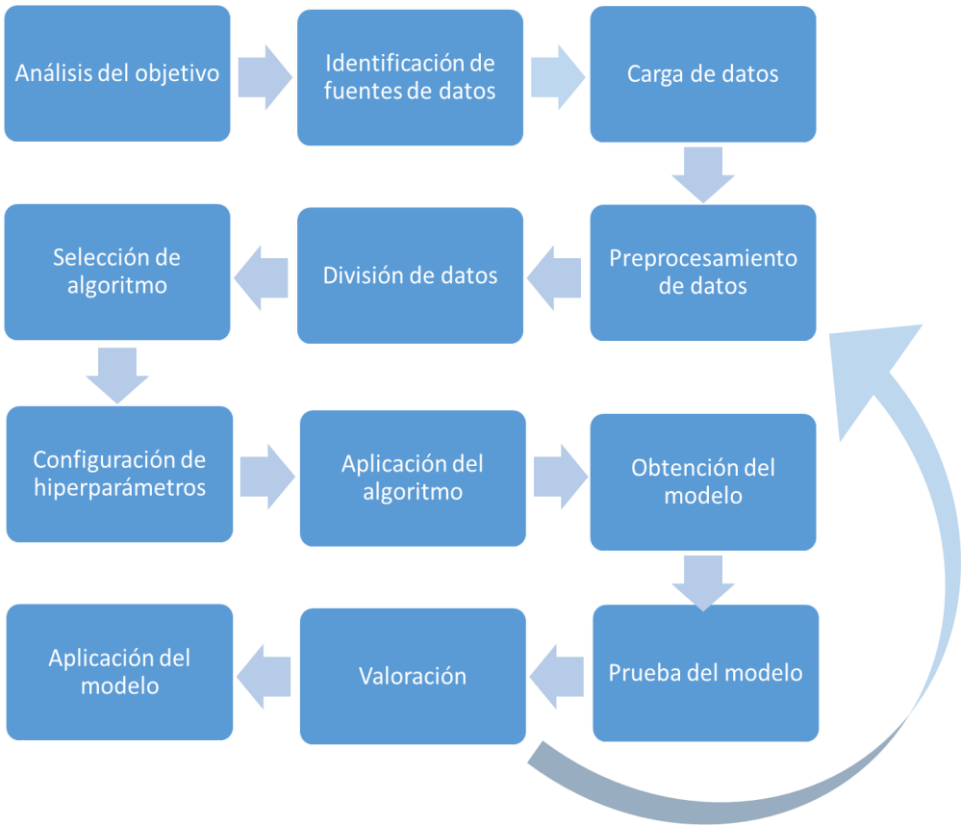

Figura 4. Operativa de trabajo

Como puede verse, en muchos aspectos es un proceso artesanal, y que requiere un conocimiento profundo del área de conocimiento sobre la que estamos trabajando.

Es necesario considerar también el considerable esfuerzo computacional que puede requerir realizar estos cálculos. Esto es, puede suponer la necesidad de utilizar varias

máquinas para su procesamiento (habitualmente con tarjetas de procesamiento de video de alta capacidad) y requerir un tiempo elevado. Dependerá de los algoritmos utilizados y del volumen de información.

6 Métodos, algoritmos y tecnologías

6.1 Métodos y algoritmos

La base matemática y estadística de los algoritmos de inteligencia artificial es fuerte, por lo que profundizar en ellos excede el alcance de este trabajo. Sin embargo, sí es conveniente conocer a grandes rasgos algunas de las principales técnicas y algoritmos de Machine y Deep Learning.

- Algoritmos de Regresión

    - Regresión Lineal. Es uno de los métodos más simples y utilizados en Machine Learning. Este algoritmo modela la relación entre una variable dependiente YYY y una o más variables independientes XXX. El modelo resultante se representa mediante una ecuación lineal:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

      donde β0 es la intersección, β1,β2,…,βn son los coeficientes de las variables independientes, y ϵ es el término de error.

    - Regresión Logística. Se utiliza principalmente para problemas de clasificación binaria. A diferencia de la regresión lineal, la salida de la regresión logística es una probabilidad que se mapea a clases binarias (0 o 1). El modelo se basa en la función logística (o sigmoide):

$$P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

  En las ciencias ómicas, la regresión lineal y logística se utilizan para modelar relaciones entre variables biológicas. Por ejemplo, la regresión lineal puede modelar la expresión génica en función de diversos factores ambientales o genéticos. La regresión logística se emplea para clasificar muestras en categorías, como sano o enfermo, basándose en perfiles ómicos.

- Árboles de Decisión y Random Forest

    - Árboles de Decisión. Son algoritmos de clasificación y regresión que segmentan repetidamente el espacio de características en subespacios más simples, evaluando una característica a la vez. Cada nodo del árbol representa una característica, y cada rama representa una regla de decisión. Este método es intuitivo y fácil de interpretar, aunque puede ser propenso al sobreajuste.

- Random Forest. Es un conjunto de métodos de aprendizaje que construyen múltiples árboles de decisión durante el entrenamiento y devuelven la clase que es el modo de las clases (clasificación) o la media de las predicciones (regresión) de los árboles individuales. Esta técnica mejora la precisión del modelo y reduce el sobreajuste combinando las predicciones de muchos árboles de decisión.

En aplicación a las ciencias ómicas, los árboles de decisión son útiles para identificar biomarcadores a partir de datos ómicos. Pueden ayudar a determinar qué genes, proteínas o metabolitos están más asociados con una condición particular, facilitando la interpretación biológica.

Random Forest puede aplicarse a las ciencias ómicas debido a su capacidad para manejar datos de grandes dimensiones y proporcionar interpretaciones sólidas. Este algoritmo se utiliza para la clasificación de enfermedades, predicción de respuestas a tratamientos e identificación de interacciones génicas.

- Agrupación o Clústering. Es una técnica de aprendizaje no supervisado que agrupa un conjunto de objetos de tal manera que los objetos en el mismo grupo (o clúster) son más similares entre sí que los de otros grupos. Uno de los algoritmos más utilizados para el clústering es K-means.

    - K-means. Este algoritmo particiona los datos en *k* clústeres, minimizando la suma de las distancias al cuadrado entre los puntos y el centro del clúster al que pertenecen. El procedimiento iterativo ajusta los centros de los clústeres y las asignaciones de puntos hasta que las posiciones de los centros no cambian significativamente.

El clústering es esencial para analizar datos ómicos, donde se agrupan genes, proteínas o metabolitos según patrones de expresión similares. K-means es popular, pero también se utilizan algoritmos más avanzados como el clústering jerárquico y el clústering basado en densidad (DBSCAN) para descubrir subtipos de enfermedades y relaciones funcionales entre moléculas.

- Redes Neuronales. Son una clase de algoritmos inspirados en la estructura del cerebro humano. Están compuestas por capas de neuronas artificiales que transforman la entrada mediante una serie de pesos y funciones de activación. Estas redes son especialmente efectivas en la detección de patrones complejos en datos, como en reconocimiento de voz, imágenes y texto. Destacan las redes neuronales convolucionales y recurrentes.

    - Redes Neuronales Convolucionales (CNN). Son un tipo de red neuronal diseñada para procesar datos que tienen una estructura de cuadrícula, como las imágenes. Las CNN utilizan capas convolucionales que aplican

filtros (o kernels) para extraer características relevantes de los datos. Estas redes son especialmente efectivas en tareas de reconocimiento de patrones y clasificación de imágenes, debido a su capacidad para capturar jerarquías de características desde bordes simples hasta formas complejas.

- Redes Neuronales Recurrentes (RNN). Están diseñadas para procesar datos secuenciales, como series temporales o secuencias de texto. Las RNN tienen conexiones cíclicas que permiten que la información persista, lo que las hace adecuadas para tareas donde el contexto previo es importante. Ejemplos de aplicaciones de RNN incluyen el modelado de lenguaje, traducción automática y predicción de secuencias en datos temporales.

En las ciencias ómicas, las CNN se utilizan para analizar datos de secuenciación y para identificar patrones en imágenes de tejido, facilitando el reconocimiento de estructuras biológicas complejas. Por otra parte, RNN son útiles para analizar series temporales de datos ómicos, como cambios en la expresión génica a lo largo del tiempo, permitiendo la modelización de dinámicas biológicas.

Mediante redes neuronales también es posible implementar Inteligencia Artificial Generativa. Este tipo de Inteligencia Artificial se refiere a modelos capaces de crear contenido nuevo y original. Estos modelos no solo analizan datos, sino que pueden generar imágenes, textos, o secuencias, como predecir nuevas secuencias genéticas o simular la evolución de una enfermedad. En ciencias ómicas, la IA generativa tiene un gran potencial en la creación de modelos de enfermedades o simulaciones moleculares.

- Procesamiento de Lenguaje Natural (PLN). El procesamiento de lenguaje natural es un campo de la inteligencia artificial que se ocupa de la interacción entre las computadoras y el lenguaje humano. Las técnicas de PLN permiten a las máquinas comprender, interpretar y generar lenguaje humano. Esto incluye tareas como análisis de sentimientos, traducción automática y generación de texto.

  En el contexto de las ciencias ómicas, el PLN se utiliza para extraer información relevante de la literatura científica y bases de datos biológicas. Herramientas de PLN pueden identificar relaciones entre genes, enfermedades y tratamientos a partir de textos científicos, acelerando el descubrimiento de conocimiento.

6.2 Tecnologías

Aunque lo realmente relevante es la base matemática y estadística que permite implementar y poner en marcha los algoritmos y técnicas de inteligencia artificial, sí hay una serie de tecnologías que, por su potencia, relativa sencillez y versatilidad se han posicionado como referentes.

Destacan especialmente:

- Python. Lenguaje de programación de propósito general, su carácter interpretado lo hace muy adecuado para la ejecución progresiva de acciones para la obtención de modelos de inteligencia artificial y la programación en tiempo real. Es muy versátil, con múltiples bibliotecas de visualización de datos, posibilidad de desarrollo web, utilización para scripting, y con una amplia comunidad de desarrollo detrás. Algunas de las bibliotecas más utilizadas en Python para IA incluyen NumPy, pandas, scikit-learn, y matplotlib.

- TensorFlow. Biblioteca de código abierto desarrollada por Alphabet, que incluye algoritmos y herramientas para la implementación de soluciones de inteligencia artificial. Es especialmente utilizada en modelos de redes neuronales debido a su capacidad para manejar grandes volúmenes de datos y optimizar el rendimiento mediante el uso de GPU.

- Keras. Interfaz de alto nivel que funciona como una capa superior de TensorFlow, facilitando el uso de dicha biblioteca. Permite a los desarrolladores construir y entrenar modelos de redes neuronales de manera más sencilla y rápida, gracias a su enfoque modular y amigable.

- PyTorch. Biblioteca de código abierto desarrollada por Meta Platforms, enfocada a Deep Learning. Ofrece una mayor flexibilidad y facilidad de uso para la investigación y el desarrollo de modelos de aprendizaje profundo. PyTorch es conocido por su uso dinámico de gráficos computacionales, lo que facilita la construcción y modificación de modelos sobre la marcha

7 Validación y evaluación de modelos

7.1 Validación y evaluación general

La validación y evaluación de los modelos generados mediante técnicas de Machine Learning es una etapa crítica para asegurar que los modelos funcionan correctamente con nuevos datos, no únicamente con los de entrenamiento.

Para la evaluación de los modelos, se detallan las principales métricas utilizadas en el análisis de datos:

1. Exactitud (Accuracy). Es la proporción de predicciones correctas sobre el total de predicciones realizadas. Se calcula mediante la siguiente fórmula:

$$\text{Exactitud} = \frac{\text{Número de predicciones correctas}}{\text{Número total de predicciones}}$$

   Es una métrica intuitiva y fácil de interpretar, y es útil para conjuntos de datos equilibrados donde las clases están representadas de manera similar.

   Por otra parte, no es adecuada para conjuntos de datos desbalanceados, ya que puede dar una falsa impresión de buen rendimiento. Por ejemplo, si una clase representa el 95% de los datos, un modelo que siempre predice esa clase tendrá una exactitud del 95%, pero no tendrá valor predictivo para la clase minoritaria.

2. Precisión, Recall y F1-Score. Estas métricas se utilizan para evaluar modelos en contextos de clasificación, especialmente cuando las clases están desbalanceadas.

   - Precisión (Precision): es la proporción de verdaderos positivos (TP) sobre el total de predicciones positivas (verdaderos positivos + falsos positivos). Indica la calidad de las predicciones positivas.

$$\text{Precisión} = \frac{\text{TP}}{\text{TP+FP}}$$

     Es útil cuando el costo de falsos positivos es alto y ayuda a entender la fiabilidad de las predicciones positivas.

   - Recall (Sensibilidad o Tasa de Verdaderos Positivos): es la proporción de verdaderos positivos sobre el total de positivos reales (verdaderos positivos + falsos negativos). Mide la capacidad del modelo para identificar correctamente todas las instancias positivas.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Es crucial cuando es importante capturar la mayoría de las instancias positivas, como en el diagnóstico de enfermedades, y es útil cuando el costo de falsos negativos es alto.

- F1-Score: es la media armónica entre precisión y recall. Proporciona un equilibrio entre las dos métricas y es útil para conjuntos de datos desbalanceados.

$$F1 = 2 \times \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}}$$

Combina tanto la precisión como el recall en una sola métrica. Es particularmente útil cuando se necesita un equilibrio entre precisión y recall. Por otra parte, no proporciona información sobre la clasificación de cada clase por separado, y puede ser menos intuitivo que otras métricas individuales.

3. Área Bajo la Curva (AUC-ROC)

La curva ROC (Receiver Operating Characteristic) traza la tasa de verdaderos positivos (TPR o sensibilidad) contra la tasa de falsos positivos (FPR) a diferentes umbrales de clasificación. El AUC (Área Bajo la Curva) proporciona una medida única del rendimiento del modelo:

$$\text{AUC-ROC} = \int_0^1 \text{ROC}(x)\, dx$$

Mide la capacidad del modelo para distinguir entre clases, independientemente del umbral de clasificación, y es robusta frente a desequilibrios de clase. Sin embargo, la interpretación puede ser menos intuitiva que otras métricas, y no proporciona información detallada sobre el rendimiento del modelo en diferentes umbrales.

4. Matriz de Confusión. Es una tabla que muestra el rendimiento del modelo, desglosando las predicciones en verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. En problemas de clasificación con múltiples clases, se extiende para mostrar el rendimiento en cada clase.

Estructura de la Matriz de Confusión para un problema binario:

	Predicción Positiva	**Predicción Negativa**
Clase Positiva	Verdaderos Positivos (TP)	Falsos Negativos (FN)
Clase Negativa	Falsos Positivos (FP)	Verdaderos Negativos (TN)

La matriz de confusión proporciona una visión detallada de cómo se comporta el modelo para cada clase, y facilita la identificación de errores específicos y patrones de clasificación incorrecta.

Además, permite calcular otras métricas derivadas como la tasa de falsos positivos (FPR) y la tasa de falsos negativos (FNR).

Es especialmente útil en problemas de clasificación con múltiples clases, proporcionando una matriz de tamaño n x n para *n* clases, facilitando el análisis detallado del rendimiento del modelo en cada clase.

7.2 Consideraciones Específicas para Datos Ómicos

Los datos ómicos presentan desafíos únicos que deben tenerse en cuenta al validar y evaluar modelos de Machine Learning.

- Alta Dimensionalidad y Escasez de Datos. Para abordar el problema se plantean las dos estrategias siguientes:
    - Reducción de Dimensionalidad. Se aplican técnicas como el Análisis de Componentes Principales (PCA) y t-SNE que pueden ayudar a reducir la dimensionalidad de los datos, manteniendo las características más relevantes. Esto facilita el entrenamiento de modelos y mejora la interpretabilidad.
    - Regularización: Métodos como la regresión Lasso y Ridge, así como las redes neuronales con regularización dropout, pueden prevenir el sobreajuste al imponer penalizaciones en los parámetros del modelo.

- Normalización y Escalado de los datos. Se propone el uso de los siguientes enfoques:

    - Normalización: Ajuste de los datos para que sigan una distribución normal puede mejorar el rendimiento de ciertos algoritmos de Machine Learning.

    - Escalado: La aplicación de técnicas como la estandarización (ajustar los datos para que tengan media 0 y desviación estándar 1) o el escalado min-max (ajustar los datos para que estén en un rango específico) son esenciales para algoritmos que dependen de la distancia entre puntos, como SVM y k-means.

- Selección de Características. La selección y priorización de características puede depurarse mediante las estrategias planteadas:

    - Filtrado de Características: Eliminar características irrelevantes o redundantes puede mejorar la eficiencia y precisión del modelo. Métodos estadísticos y técnicas basadas en información mutua son útiles para esta tarea.

    - Ensamblado de Modelos: Combinar múltiples modelos (bagging, boosting) puede mejorar el rendimiento general y la robustez de las predicciones.

7.3 Ejemplos

Algunos ejemplos de aplicación de técnicas de validación en Ciencias Ómicas:

- Clasificación de Cáncer: Los modelos de Machine Learning pueden predecir subtipos de cáncer basados en datos de expresión génica. La validación cruzada asegura que el modelo generaliza bien a nuevos pacientes.

- Identificación de Biomarcadores: Algoritmos como Random Forest pueden identificar genes o proteínas que actúan como biomarcadores para diagnosticar enfermedades. La matriz de confusión y el AUC-ROC son métricas clave para evaluar su rendimiento.

- Predicción de Interacciones Proteína-Proteína: Modelos de aprendizaje profundo pueden predecir interacciones basadas en secuencias de proteínas. La validación cruzada y el F1-Score son esenciales para evaluar estos modelos.

- Reconstrucción de Redes Metabólicas: Técnicas de Machine Learning pueden inferir redes metabólicas a partir de datos metabolómicos. La precisión y el recall ayudan a evaluar la calidad de las redes reconstruidas.

En cualquier caso, será necesario estudiar la situación particular para la selección de las técnicas más adecuadas al caso concreto, y mediante las métricas de valoración ir afinando hasta obtener el modelo que proporcione los mejores resultados.

8 Implicaciones éticas

8.1 Consideraciones generales

La aplicación de la inteligencia artificial (IA) a las ciencias ómicas permite revolucionar la investigación biomédica, proporcionando avances significativos en el entendimiento de procesos biológicos complejos y en el desarrollo de terapias personalizadas.

Sin embargo, el uso de esta tecnología también plantea importantes cuestiones éticas que deben ser cuidadosamente consideradas. A continuación, se exploran las principales implicaciones éticas del uso de IA en ciencias ómicas.

8.2 Privacidad y Confidencialidad de los Datos

La naturaleza de los datos ómicos, que incluye información genética detallada, plantea riesgos para la privacidad y la confidencialidad de los individuos.

- Reidentificación de Datos. Los datos genómicos son únicos para cada individuo, lo que aumenta el riesgo de reidentificación incluso cuando los datos están anonimizados. La combinación de datos genéticos con otros datos personales puede llevar a la identificación de individuos, poniendo en riesgo su privacidad.

- Protección de Datos. Es fundamental implementar medidas robustas de protección de datos, incluyendo el uso de técnicas avanzadas de anonimización y cifrado, para garantizar que la información personal no sea accesible sin autorización. La normativa de protección de datos, como el Reglamento General de Protección de Datos (GDPR) en la Unión Europea, establece requisitos estrictos para la recopilación, almacenamiento y uso de datos personales, incluyendo los datos ómicos.

- Consentimiento Informado. Es esencial la obtención del consentimiento informado de los participantes. Debe ser claro y específico sobre cómo se usarán sus datos, con qué propósito y quién tendrá acceso a ellos. Es importante asegurar que los participantes comprendan los riesgos asociados con la reidentificación y el uso potencial de sus datos en investigaciones futuras.

8.3 Equidad y Sesgo.

El sesgo en los datos y los algoritmos de IA puede llevar a resultados desiguales y perpetuar la discriminación. Debe contemplarse desde dos puntos de vista distintos:

- Representación de Datos. Los datos ómicos utilizados para entrenar modelos de IA deben representar adecuadamente la diversidad genética y ambiental de la población. La falta de representación puede llevar a sesgos que afecten negativamente a ciertos grupos, exacerbando disparidades en salud.

- Sesgo Algorítmico. Los algoritmos de IA pueden aprender y amplificar sesgos presentes en los datos de entrenamiento. Es crucial identificar y mitigar estos sesgos para asegurar que los modelos sean justos y equitativos. Las herramientas y técnicas para la detección y corrección de sesgos algorítmicos, como el análisis de equidad y los métodos de ajuste de sesgo, deben ser parte integral del desarrollo de modelos.

- Impacto en la Salud. Los sesgos en los modelos de IA pueden llevar a diagnósticos erróneos o tratamientos inadecuados para ciertos grupos poblacionales. Esto puede tener consecuencias graves en términos de acceso a cuidados y resultados de salud.

8.4 Transparencia e Interpretabilidad.

La transparencia y la interpretabilidad de los modelos de IA son cruciales para asegurar la confianza y la aceptación de estas tecnologías. Hay 3 aspectos clave relativos a la interpretación:

- Caja Negra de la IA. Muchos modelos de IA, especialmente los basados en redes neuronales profundas, se consideran "cajas negras" debido a su complejidad y falta de interpretabilidad. Esto puede dificultar la comprensión de cómo se toman las decisiones y la identificación de posibles errores.

- Aplicabilidad. Es esencial desarrollar técnicas que permitan explicar las decisiones de los modelos de IA de manera comprensible para los humanos. Métodos como SHAP (SHapley Additive exPlanations) y LIME (Local Interpretable Model-agnostic Explanations) pueden ayudar a proporcionar interpretaciones más claras.

- Transparencia en el Desarrollo. La transparencia en el desarrollo y la validación de modelos es fundamental. Esto incluye documentar y comunicar claramente los datos utilizados, los algoritmos implementados y los resultados de las evaluaciones de rendimiento y sesgo.

8.5 Responsabilidad y Gobernanza.

La implementación responsable de la IA en las ciencias ómicas requiere una gobernanza adecuada y la asignación clara de responsabilidades. Identificamos de ámbitos de trabajo en relación a la responsabilidad y la gobernanza:

- Responsabilidad Ética. Los desarrolladores e investigadores tienen la responsabilidad ética de garantizar que sus modelos de IA sean seguros, justos y respeten la privacidad. Esto incluye realizar evaluaciones éticas continuas y estar preparados para abordar cualquier impacto negativo.

- Marco de Gobernanza. Es esencial establecer un marco de gobernanza que supervise el uso de la IA en ciencias ómicas. Esto puede incluir comités de ética, regulaciones y directrices que aseguren el cumplimiento de principios éticos y normativos. La colaboración entre científicos, bioeticistas, legisladores y representantes de la comunidad es fundamental para desarrollar y mantener estos marcos de gobernanza.

- Responsabilidad Legal. Clarificar la responsabilidad legal en caso de errores o daños causados por modelos de IA es crucial. Las instituciones y los desarrolladores deben estar preparados para asumir la responsabilidad y tomar medidas correctivas cuando sea necesario.

9 Marco regulatorio

9.1 Introducción

El marco regulatorio del uso de la inteligencia artificial en ciencias ómicas es esencial para asegurar que los beneficios de esta tecnología se realicen de manera ética y responsable. La legislación actual proporciona una base sólida para el desarrollo y la implementación de IA en este campo. Sin embargo, es necesario un esfuerzo continuo para adaptar y fortalecer la regulación en respuesta a los rápidos avances tecnológicos, asegurando así una integración segura y beneficiosa de la IA en las ciencias ómicas.

Debe considerarse, por una parte, que las regulaciones deben adaptarse rápidamente a los avances tecnológicos para mantenerse relevantes, así como alinear las regulaciones a nivel global para facilitar la colaboración internacional y la transferencia de datos.

Por otra parte, es clave disponer de un marco regulatorio sólido puede fomentar la innovación responsable y la confianza pública en la IA, así como proteger los derechos individuales y empoderan a los participantes en la investigación científica.

9.2 Regulación General de Protección de Datos (GDPR)

La Regulación General de Protección de Datos (GDPR) de la Unión Europea es una de las legislaciones más completas y estrictas en términos de protección de datos personales. Sus principios y requisitos son especialmente relevantes para el uso de IA en ciencias ómicas debido a la naturaleza sensible de los datos involucrados.

Principios Clave del GDPR:

- Licitud, Lealtad y Transparencia: Los datos deben ser procesados de manera legal, justa y transparente para el interesado.
- Limitación de la Finalidad: Los datos deben ser recogidos con fines específicos, explícitos y legítimos, y no ser procesados de manera incompatible con esos fines.
- Minimización de Datos: Los datos recolectados deben ser adecuados, pertinentes y limitados a lo necesario en relación con los fines para los que son procesados.
- Exactitud: Los datos deben ser exactos y, cuando sea necesario, actualizados.
- Limitación del Plazo de Conservación: Los datos no deben ser conservados más tiempo del necesario para los fines del procesamiento.

- Integridad y Confidencialidad: Los datos deben ser procesados de manera que se garantice una seguridad adecuada.

Derechos de los Sujetos de Datos:

- Derecho de Acceso: Los individuos tienen derecho a saber si sus datos están siendo procesados y a acceder a esos datos.
- Derecho a la Rectificación: Los individuos pueden solicitar la corrección de datos inexactos.
- Derecho al Olvido: Los individuos pueden solicitar la eliminación de sus datos bajo ciertas circunstancias.
- Derecho a la Portabilidad de los Datos: Los individuos pueden recibir sus datos en un formato estructurado y comúnmente utilizado.

9.3 Ley de inteligencia artificial (AI Act)

La Ley de Inteligencia Artificial de la Unión Europea (AI Act), aprobada en 2024, establece un marco regulatorio integral para el uso y desarrollo de sistemas de inteligencia artificial (IA) en la UE. Su objetivo principal es garantizar que la IA se desarrolle y utilice de manera ética y segura, protegiendo los derechos fundamentales de los ciudadanos.

Esta ley representa un hito en la regulación de la inteligencia artificial, estableciendo un equilibrio entre la promoción de la innovación tecnológica y la protección de los derechos y la seguridad de los ciudadanos en la Unión Europea.

Puntos clave de la ley:

- Entrada en vigor y fases de implementación:
    - 1 de agosto de 2024: La ley entra en vigor.
    - 2 de febrero de 2025: Se prohíben ciertos usos de la IA, como sistemas de identificación biométrica y sistemas de "puntuación social" que puedan causar daño significativo.
    - 2 de agosto de 2026: Se implementan la mayoría de las obligaciones para los sistemas de IA de alto y bajo riesgo.
    - 2 de agosto de 2027 y más allá: Cumplimiento total de los sistemas de IA de uso general y aquellos regulados por leyes específicas de la UE.
- Clasificación de riesgos:

    - Riesgo inaceptable: Prohibidos, incluyendo sistemas de IA que manipulen el comportamiento humano de manera perjudicial.
    - Alto riesgo: Sujeto a requisitos estrictos, como sistemas usados en contratación laboral, infraestructura crítica, y evaluación de crédito.
    - Riesgo limitado y mínimo: Sujetos a menos restricciones, pero aún deben cumplir con ciertos estándares de transparencia y seguridad.

- Obligaciones para proveedores y usuarios de IA:
    - Proveedores: Deben realizar evaluaciones de riesgo, asegurar la calidad de los datos utilizados, y mantener una supervisión humana adecuada.
    - Desarrolladores: Necesitan registrar los sistemas de IA de alto riesgo y asegurar que cumplen con las normas de ciberseguridad y precisión.
    - Usuarios: Deben seguir las instrucciones proporcionadas, aplicar supervisión humana y monitorear el funcionamiento del sistema.

- Sanciones y cumplimiento:
    - Las multas por incumplimiento pueden alcanzar hasta el 7% de la facturación global anual o 35 millones de euros, lo que sea mayor. La ley tiene un alcance extraterritorial, aplicándose también a empresas no europeas que operan en la UE.

- Promoción de la innovación:
    - Se establecen "sandboxes" regulatorios donde las empresas pueden probar nuevas tecnologías de IA bajo supervisión controlada, fomentando así la innovación mientras se asegura el cumplimiento de las normas éticas y de seguridad.

9.4 Espacio Europeo de Datos de Salud (EHDS)

El Espacio Europeo de Datos de Salud (EHDS) es una iniciativa en el marco regulatorio de la Unión Europea que busca facilitar el intercambio seguro de datos de salud en toda Europa, promoviendo la investigación científica y la innovación en medicina, especialmente en campos como las ciencias ómicas. El Consejo y el Parlamento Europeo llegaron a un acuerdo provisional sobre esta regulación en marzo de 2024. De esta forma queda establecida una estructura legal y técnica que permite el acceso a datos sanitarios entre distintos países europeos.

Este marco legal tiene varios objetivos esenciales:

- Facilitar el acceso a los datos sanitarios por parte de profesionales de la salud y pacientes, garantizando al mismo tiempo la protección de la privacidad y la seguridad de los datos personales.

- Impulsar la investigación científica mediante el acceso seguro a datos a gran escala, lo que es clave para el avance en áreas como la medicina personalizada y el uso de IA en las ciencias ómicas.

- Proporcionar un entorno regulatorio armonizado que fomente la innovación y el desarrollo de tecnologías avanzadas, asegurando que los datos puedan ser utilizados de manera ética y efectiva en toda la UE.

El EHDS también incluye medidas para mejorar la interoperabilidad de los sistemas de salud y establece normas claras sobre el acceso y uso de los datos en investigación, desarrollo de políticas de salud pública y la creación de productos de IA más precisos y seguros.

9.5 Iniciativas internacionales

Adicionalmente, has surgido las siguientes iniciativas Internacionales que proporcionan directrices que se siguen en la elaboración de normativa legal.

- UNESCO: Recomendación sobre la Ética de la Inteligencia Artificial: La UNESCO ha desarrollado una guía global para la ética de la IA, promoviendo principios de equidad, transparencia y responsabilidad.

- Organización Mundial de la Salud (OMS): Ética y Gobernanza de la IA en Salud: La OMS ha establecido principios y recomendaciones para el uso de la IA en salud, incluyendo la protección de datos y la equidad. Foro Económico Mundial (WEF): Directrices para la IA: El WEF ha desarrollado marcos y herramientas para ayudar a las organizaciones a implementar IA de manera ética y responsable.

10 Infraestructura tecnológica y de datos

Las ciencias ómicas son unas disciplinas que se soportan en un enorme volumen de datos y requerimiento intensivo de capacidad computacional, por tanto, el éxito de estas ciencias depende en gran medida de una infraestructura informática adecuada que pueda hacer frente a las crecientes demandas de datos y permitir una amplia gama de actividades computacionales que consumen muchos recursos.

Los secuenciadores de ADN producen muchos datos, preprocesados por las herramientas software y algoritmos existentes en un dispositivo de secuenciación. Una vez obtenidos los datos genómicos, deben realizarse procesos de gestión y análisis de datos para obtener información de dichos datos. Los datos se procesan empleando bases de conocimientos (expertas en determinados dominios) disponibles en el dominio público, o a través de métodos de descubrimiento automatizados para inferir conclusiones a partir de los datos. Por último, los datos se pueden utilizar para crear modelos predictivos con métodos de aprendizaje automático para extrapolar nuestra comprensión de la información faltante en nuestra base de conocimientos, o para respaldar la inferencia si hay evidencias de apoyo inadecuadas en el conjunto de datos. La necesidad de realizar diferentes etapas de análisis depende de la pregunta científica que se investiga y requiere habilidades computacionales clasificadas en bioinformática, biología computacional y ciencias de datos. Por tanto, las etapas del procesamiento de datos genómicos se vinculan a distintos roles de usuarios finales, y según la naturaleza y el volumen de los datos cambian en cada etapa, también lo hace la naturaleza de los algoritmos y la necesidad de habilitar distintas herramientas o infraestructuras informáticas.

El procesamiento de datos genómicos que supone el análisis bioinformático de los datos se puede clasificar en tres grandes etapas: análisis primario, secundario y terciario (Figura 2).

Estas etapas requieren que una serie de herramientas se ejecuten de manera secuencial para procesar los datos, a menudo conocidas como flujos de trabajo, es decir, la salida de una herramienta va como entrada a la herramienta siguiente en el flujo de trabajo y así sucesivamente. Tradicionalmente, los flujos de trabajo se han escrito en scripts de shell, archivos por lotes o scripts de Perl/Python, aunque actualmente se está adoptando cada vez más especificaciones especializadas en la creación de flujos de trabajo, como el Lenguaje Común de Flujo de Trabajo (CWL) y el Lenguaje de Descripción de Flujo de Trabajo (WDL). Estas especificaciones se han desarrollado para describir flujos de trabajo y herramientas para entornos científicos con uso intensivo de datos, para hacerlos escalables en estaciones de trabajo, clústeres y entornos en la nube. El problema de la compatibilidad, la portabilidad y el intercambio de herramientas entre plataformas se está abordando mediante el uso de tecnologías de contenedores como Docker y Singularity. El uso de contenedores (p.e. kubernetes) se están convirtiendo en la opción

principal en la bioinformática y son un engranaje esencial en el desarrollo del flujo de trabajo basado en las especificaciones CWL y WDL.

Las especificaciones del flujo de trabajo, cuando se combinan con la tecnología de contenedores, han abierto las puertas para que la computación en la nube haga presencia en la genómica pudiendo adaptarse a tecnologías de nube hibrida integrando las tecnologías en la nube y la HPC (high-performance computing) para impulsar las demandas de datos y computación de la genómica.

La biología computacional comienza con conjuntos de datos que son mucho más pequeños en tamaño y son la esencia de los conjuntos de datos ya procesados a través de los flujos de trabajo bioinformáticos, por tanto, el volumen de datos no es un problema en la biología computacional, pero si lo es el volumen de cálculo computacional requerido ya que se emplean modelos matemáticos y estadísticos que se centran en paralelizar el cálculo usando habitualmente HPC tradicional. La tecnología de unidades de procesamiento gráfico (GPU) dispone de una capacidad paralela programable más barata, local y, a menudo, en el borde (edge), esto unido a que muchos clústeres de HPC admiten GPU como parte de su capacidad han desencadenado que el uso de GPU se esté empleando cada vez más en la genómica debido a sus capacidades de paralelismo de datos, desempeñando un papel cada vez más importante en la aceleración del análisis computacional siendo especialmente adecuado para operaciones de optimización utilizadas en biología de redes y técnicas de aprendizaje profundo.

En una última etapa, los científicos de datos deben lidiar con un gran volumen de datos, así como con un gran volumen de cómputo y lo hacen a través de dos vías técnicas principales: (i) la aplicación de un marco de reducción de para hacer frente a la gran cantidad de datos, y (ii) la aplicación de la IA y el aprendizaje automático a escala. Las tecnologías clave de la ciencia de datos que se aplican cada vez más para la genómica son las bases de datos NoSQL, Hadoop, Spark, el procesamiento del lenguaje natural (NLP) y el aprendizaje automático, incluidos los métodos de aprendizaje profundo. Los métodos de aprendizaje se profundo pueden explotar de manera eficiente en los clústeres de HPC tradicionales y granjas de GPU dedicadas.

Si agrupamos las diferentes actividades de las diferentes etapas enumeradas anteriormente, obtenemos dos paradigmas de procesamiento de software en genómica: (i) paradigma HPC tradicional centrado en la simulación, y (ii) análisis de datos de alto nivel (HDA), empezando a surgir la idea de una convergencia entre los dos donde ambos coexisten proporcionando interfaces de servicio. Esta infraestructura se despliega en recursos híbridos multinube, ya que la nube hibrida proporciona un escenario valido para el procesamiento de datos genómicos, donde la información que requiere privacidad (datos genómicos que se someten a un análisis bioinformático) se puede procesar en la infraestructura interna HPC (desplegada en nube privada o en

centro de proceso de datos privados (on-premise)) y archivar internamente, mientras que los conjuntos de datos que se procesan en las etapas posteriores pueden anonimizarse y compartirse con agentes/servicios externos se procesaran a través de la interfaz publica de la nube hibrida.

En la siguiente imagen se muestra un esquema de un diseño de multinube hibrida que proporciona servicios a la comunidad de usuarios de la información genómica. La infraestructura de esta propuesta combina diferentes elementos y soluciones tanto a nivel hardware (CPU, GPU, almacenamiento...) como a nivel software (máquinas virtuales, contenedores, middleware, marcos de trabajo, ...) que pueden proporcionarse desde nubes privadas como desde servicios en la nube pública.

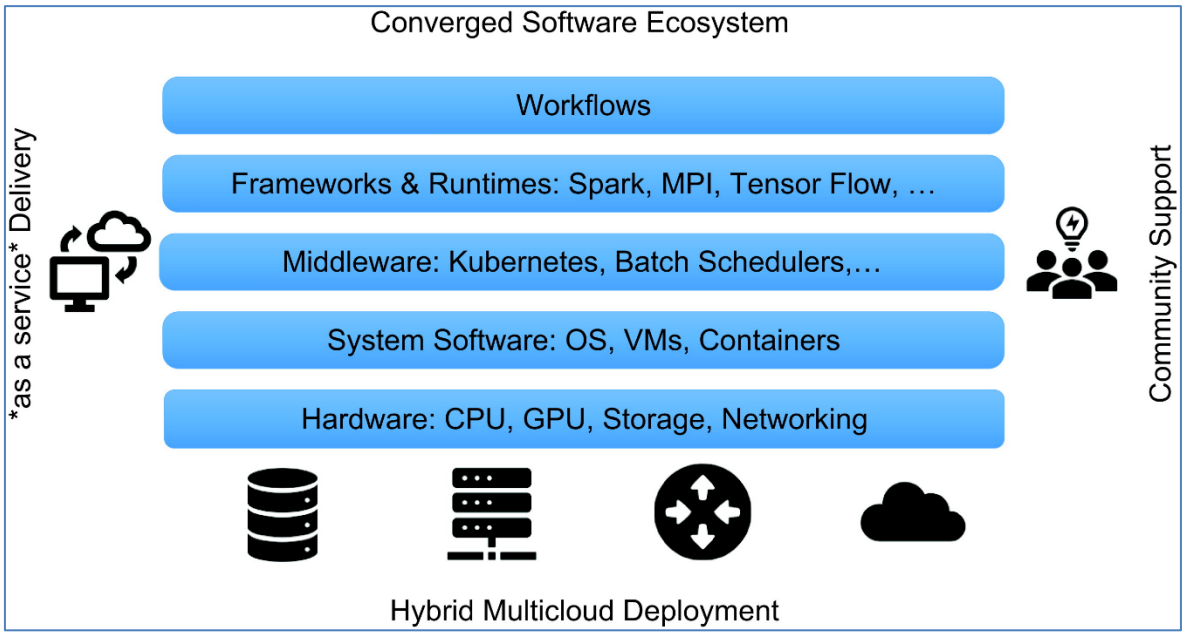

Figura 5. Esquema de despliegue en nube hibrida

La computación en la nube (con sus distintas combinaciones) se presenta como la mejor solución para el despliegue de una infraestructura tecnológica que de soporte al procesamiento de la información genómica, ya que en una gran proporción emplea las mismas herramientas que el big data. Dicho despliegue puede realizarse empleando las distintas modalidades de servicio que ofrecen los proveedores:

- Infraestructura como servicio (IaaS): proporciona la infraestructura hardware para la computación de alto rendimiento (HPC) como un servicio de “pago por lo que se necesita” que proporciona una escalabilidad de recursos a demanda. AWS (Amazon Web Services), Windows Azure o Google Compute Engine son los servicios de infraestructura más utilizados en la actualidad.

- Software como servicio (SaaS): proporciona la ejecución de software en entorno virtualizado en la nube, sin necesidad de que se instale ese software en el equipo del usuario. Un ejemplo de servicio SaaS en bioinformática es Biolinux que

proporciona una máquina virtual alojada en AWS con sistema operativo Ubuntu Linux y más de 100 herramientas bioinformáticas instaladas.

- Plataforma como servicio (Paas): proporciona entornos de desarrollo que disponen de bibliotecas de software y plataformas de desarrollo. MapReduce/Hadoop de Google es una de las soluciones más extendidas para el procesamiento de big data genómico en la nube existiendo en la actualidad decenas de implementaciones bioinformáticas sobre Hadoop como pueden ser GATK (kit de herramientas para análisis genéticos de datos de Resecuenciación) o FX (herramienta para análisis de secuencias ARN).

- Dato como servicios (DaaS): proporciona acceso a bases de datos o repositorios de información alojados en la nube. AWS proporciona acceso a varios conjuntos de datos genómicos públicos de gran tamaño como GenBank, NCBI o el proyecto 1000 Genome.

La infraestructura hardware y de datos necesaria para la gestión de datos genómicos se basa en las arquitecturas complejas empleadas para Big Data, como la Lambda o Kappa, junto almacenamiento de datos con soluciones SQL, No SQL o de sistemas de archivos distribuidos (Hadoop Distributed File System o Amazon S3) sobre discos duros, sistemas de almacenamiento en red /NAS) o memoria flash que pueden desplegarse on-premise (con la complejidad que conlleva) o en servicios de la nube.

Con la irrupción de la IA y la genómica surge una necesidad urgente de emplear soluciones de almacenamiento más sostenibles, ya que los enormes conjuntos de datos que se manejan pueden demandar elevadas densidades de potencia (40 o 50 kilovatios) por rack, por ello, la tendencia actual está moviéndose hacia el uso de almacenamiento en memoria flash en detrimento de los discos duros tradicionales (HDD). Estas memorias flash ofrecen significativas ventajas respecto a los discos duros en términos de rendimiento, eficiencia energética, confiabilidad y reducción de espacio físico y necesidades de refrigeración. Las soluciones basadas en memorias flash ofrecen una alta escalabilidad basada en servicios de suscripción de almacenamiento (STaaS), una elevada facilidad de administración, capacidades de seguridad nativas (encriptación, controles de acceso,...) y facilidad de integración con soluciones en la nube, existiendo actualmente en el mercado compañías (p.e. Pure Storage) que ofrecen soluciones específicas para el procesamiento de IA y el mercado de la salud (almacenamiento de imagen médica o patológica, datos de NGS, EHR,...).

Con respecto al software a emplear para el desarrollo y ejecución de procesos basados en IA, este es variado y dependerá de las necesidades de cada proyecto. A continuación, se muestra algunos ejemplos de las herramientas software con más implantación en este ámbito:

1. Lenguajes de programación:
    a. Python y r.
2. Bibliotecas de IA y Aprendizaje Automático:
    a. TensorFlow/PyTorch: Para el desarrollo de modelos de aprendizaje profundo.
    b. scikit-learn: Para modelos de aprendizaje automático más tradicionales.
    c. Bioconductor: En R, ofrece herramientas para el análisis de datos biológicos.
    d. Frameworks de Bioinformática: Herramientas específicas como Galaxy, GATK, o Bioconductor.
3. Gestión de Datos:
    a. Bases de Datos: Utiliza bases de datos relacionales (como PostgreSQL) o no relacionales (como MongoDB) para almacenar y gestionar datos ómicos.
    b. ETL (Extract, Transform, Load): Herramientas que permiten la integración de datos desde diferentes fuentes y su transformación para análisis posterior.
4. Entorno de Desarrollo
    a. Jupyter Notebooks: Ideal para el análisis interactivo y la visualización de datos.
    b. Entornos de Desarrollo Integrados (IDE): Como PyCharm o RStudio, para el desarrollo de scripts y modelos.
5. Visualización de Datos
    a. Herramientas de Visualización: Bibliotecas como Matplotlib, Seaborn (en Python) o ggplot2 (en R) para visualizar datos y resultados.
    b. Dashboards: Considera herramientas como Dash o Shiny para crear aplicaciones interactivas que presenten resultados de manera accesible.

En resumen, la infraestructura tecnológica y de datos que se debe dotar para el abordaje de proyectos de aplicación de IA en genómica es compleja y heterogénea ya que se debe dar servicio a procesos computacionales diversos de gran complejidad y que manejan elevados volúmenes de datos. El paradigma de la computación en la nube se perfila como el escenario más atractivo para este tipo de proyectos ya que nos permite hacer uso de los distintos tipos de servicios ofertados por los proveedores con la capacidad de personalizar los recursos que necesitemos en cada momento. Dentro de la computación en la nube, el esquema de nube hibrida será el más adecuado para este tipo de proyectos ya que permitirán restringir el procesamiento y almacenamientos de datos sensibles en la parte privada de la nube. En cuanto a la infraestructura necesaria para un despliegue On Premise, requerirá de una dotación que permita procesamiento de alto rendimiento (HPC) instalando servidores clásicos, granjas de GPU y soluciones de almacenamiento altamente escalables (como la memoria flash) donde se ejecuten tecnologías de virtualización de servidores, contenedores, middlewares y distintas soluciones software específicas para el procesamiento de los algoritmos de IA.

11 Retos e incertidumbres de la Aplicación de la IA a los datos ómicos

Aunque la utilidad de la IA para la gestión e interpretación de los datos ómicos, existen ciertos retos y/o incertidumbres que debemos afrontar:

- **¿Soluciones de gestión integral customizadas o comerciales? Resistencia a la incorporación de tecnología.** Existe controversia e incertidumbre sobre si el uso de soluciones comerciales puede conllevar a perder recursos humanos especializados en biotecnología e informática o genetistas, que impliquen perder capacidad de conocer y evaluar las nuevas tecnologías incorporadas, si nos centramos en soluciones comerciales cerradas.

  La respuesta a esta pregunta no es diferente a la que ya se ha vivido con la incorporación de otras tecnologías previas, como las hojas de cálculo, o las cadenas robóticas y los softwares de información de laboratorios con aplicación de reglas expertas. Estas soluciones, permiten una gestión más coste eficiente pero supervisada de los análisis.

  El utilizar soluciones más abiertas o más cerradas, dependerá mucho del número de recursos materiales (servidores, y otros requisitos de hardware y software) y perfiles de profesionales que dispongamos y cuánto tiempo es posible mantener dichos recursos. La colaboración publico privada en este campo parece indispensable, y apostar por soluciones ya existentes, siempre velando a que se adapten a las necesidades y requerimientos que aseguren la seguridad, consistencia, trazabilidad, parece una aproximación sensata pero no indiscutible. Soluciones mixtas también son interesantes

  Las soluciones que se diseñen o se implementen sean customizadas o comerciales cerradas, deben seguir un proceso de adaptación altamente demandante, por ser una ciencia enormemente cambiante y que requiere que dicha adaptación se haga en muy poco tiempo. En cualquier caso, es indispensable, que se sigan unos criterios de validación muy rigurosos, si queremos asegurar un abordaje de medicina personalizada de precisión eficaz, eficiente y segura y esto sólo es viable con grupos de trabajo multidisciplinares trabajando en colaboración.

  Además, tanto como los evaluadores como los usuarios finales deben ser siempre conocedores de las aplicaciones y limitaciones de las herramientas.

- **¿Dónde Deben estar los datos; On premise o in cloud?** Muchos de los servicios de IA requieren de plataformas de alto rendimiento y se ofrecen en sistemas SAAS. Una opción es mantener los datos on premise y sólo subir los datos imprescindibles y anonimizados para el análisis computacional. No obstante, ¿qué datos serían los indispensables? Independientemente de la decisión,

aunque hoy en día no es fácilmente identificable un paciente a partir de un estudio genómico, conforme vayamos incrementando el conocimiento es posible que pueda llegarse a producirse. Por ello, dada la especial sensibilidad de esta información, si se opta por esa solución, es indispensable ser propietarios de la información y exigir sistemas seguros que cumplan las normativas de seguridad

- **¿tiene sentido almacenar todos los ficheros que se extraen del secuenciador? ¿durante cuánto tiempo almacenamos la información?** los secuenciadores emiten un tipo de fichero, que luego pasa por un pipeline de análisis, que va generando ficheros subsiguientes, siendo el último (VCF), el que realmente tiene uso asistencial directo y además ocupa poco tamaño. Sin embargo, ese VCF puede ser diferente según el genoma de referencia que estemos utilizando para el alineamiento o según los pipelines que hayamos aplicado. Por tanto, si sólo mantenemos los VCF y quisiéramos reanalizarlo en un tiempo cuando haya cambiado el genoma de referencia o reanalizarlo con un pipeline diferente, no podremos hacerlo con garantía. Por otro lado, las tecnologías de secuenciación también evolucionan y su coste desciende: estamos ya por la tercera generación de secuenciadores y el coste de un genoma puede ya en la actualidad ser de 150 euros. El coste de almacenaje, por el contrario, está incrementándose, sobre todo cuando hablamos de memoria activa, pero incluso el almacenaje a largo plazo pasivo tiene un coste no desdeñable y no siempre abordable con las infraestructuras habituales. Además, preocupa que los datos caigan en las manos equivocadas. Incluso aunque los datos estén muy bien custodiados, el mero hecho de tener esta información tan sensible es un potencial riesgo hoy en día. Además, muchos modelos de IA no necesitan toda la información contenida en los ficheros para el entrenamiento de modelos. Por el contrario, si se quieren reanálisis precisamos disponer de los datos. Ante esto, ¿debemos mantener de forma indefinida la información, por si existen reclasificaciones?, o ¿tras un tiempo prudencial, una vez analizado e informado, se puede conservar sólo los logs de accesos y acciones, la muestra y si es preciso reanalizarla en futuro, volverla a secuenciar, con una tecnología más avanzada? En realidad, esta pregunta, ya es aplicable a otras áreas como la patología digital o la radiómica y la respuesta no está clara. Por ahora, en general, se están adoptando medidas muy conservadoras, almacenando las muestras originarias y los ficheros analizados todos los datos de forma local

Haciendo análisis DAFO de situación de una instalación nacional, se plantean las siguientes debilidades, amenazas, fortalezas y oportunidades.

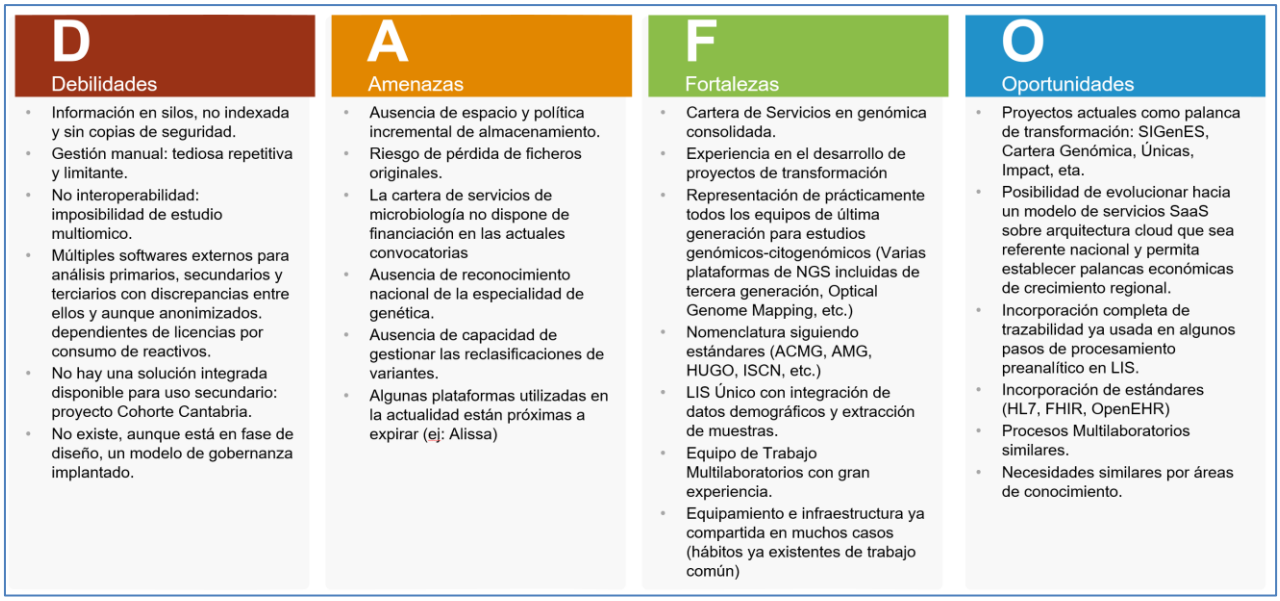

Figura 6. Esquema DAFO de situación de una instalación nacional.

12 Colaboración interdisciplinaria

Generalmente cuando se habla de multidisciplinariedad en medicina, con frecuencia nos evoca distintas especialidades médicas colaborando para hacer un abordaje más integral. Sin embargo, en el mundo de las ciencias ómicas, aunque esa colaboración es también importante, en este caso, lo verdaderamente imprescindible es conformar grupos de perfiles profesionales muy diferentes, trabajando juntos para poder desarrollar y evolucionar de forma exitosa. En industria es bastante común, que diferentes equipos o departamentos dentro de una organización colaboren y los que lo hacen suelen tener resultados muchos más coste-eficientes. Además, estos equipos, suelen ser más flexibles, están más motivados y suele ser más fácil capturar talento. En el ámbito de la medicina esto es mucho menos frecuente, aunque afortunadamente esto está cambiando

Las ómicas requieren de la participación de varios equipos, como son informáticos de sistemas, Bioinformáticos, Especialistas en IA, Técnicos de Laboratorio, Genetistas, Clínicos, Especialistas en Contratación, etc. Todos los roles son igualmente importantes y la aproximación integrada enormemente enriquecedora. Para que estos equipos funciones deben darse una serie de circunstancias:

- Los diferentes perfiles tienen que estar todos representados y deben ir formándose en adquirir unas nociones básicas de lo que hacen resto de perfiles, para poder intercambiar ideas. Un mínimo lenguaje común. Esta capacitación no es sencilla, y puede y debe ser progresiva, pero es esencial y tremendamente enriquecedora para el proyecto y para los propios profesionales. Las sinergias son mucho más eficientes que los equipos que trabajan de forma individualizada y luego intentan sumar las partes.

- Estos grupos deben disponer del tiempo necesario para hacer la customización del producto (si se opta por soluciones propias) y/o de validación y monitorización rigurosos (independientemente de si la solución es propia o externa).

- Los directivos y los gestores de proyectos deben ser expertos en coordinar la colaboración interdisciplinaria para obtener nuevos enfoques, impulsar planes ambiciosos y mantener al equipo enfocado hacia el objetivo común.

- Un desafío es la percepción de los riesgos. Por ejemplo, el personal de sistemas puede preferir evitar ciertas soluciones por miedo a la seguridad mientras que los genetistas pueden ser más partidarios de las mismas para poder hacer un análisis más eficiente. Deben encontrarse

soluciones que permitan ser innovadoras pero seguras. Las validaciones multidisciplinares en entornos controlados serán imprescindibles.

- Validaciones: Se deberán diseñar indicadores y medidas de control que permitan alertarnos de las desviaciones, siguiendo los esquemas de seguridad de alto nivel y las certificaciones de calidad. No obstante, no debemos dejar que el miedo al error nos impida avanzar. Con frecuencia pensamos en el miedo a que la IA cometa errores y se nos olvida que el error humano es una realidad en cualquier sistema.

13 Casos de estudio

Históricamente la mayoría de la actividad IA en genómica se centró en la fase de la investigación. Sin embargo, la mayoría de los aspectos del análisis genómico han utilizado de alguna manera el machine learning (ML) y el aprendizaje profundo (DL), desde la secuenciación, el fenotipado y la identificación de variantes, a la interpretación posterior (figura 7). De hecho, los algoritmos de aprendizaje automático se han usado en tareas bioinformáticas desde hace muchos años, (e.g. anotación del genoma y predicción del efecto variante). Ahora los avances en informática, aprendizaje profundo y el crecimiento de los conjuntos de datos biomédicos están permitiendo mejoras en las áreas de utilidad existentes.

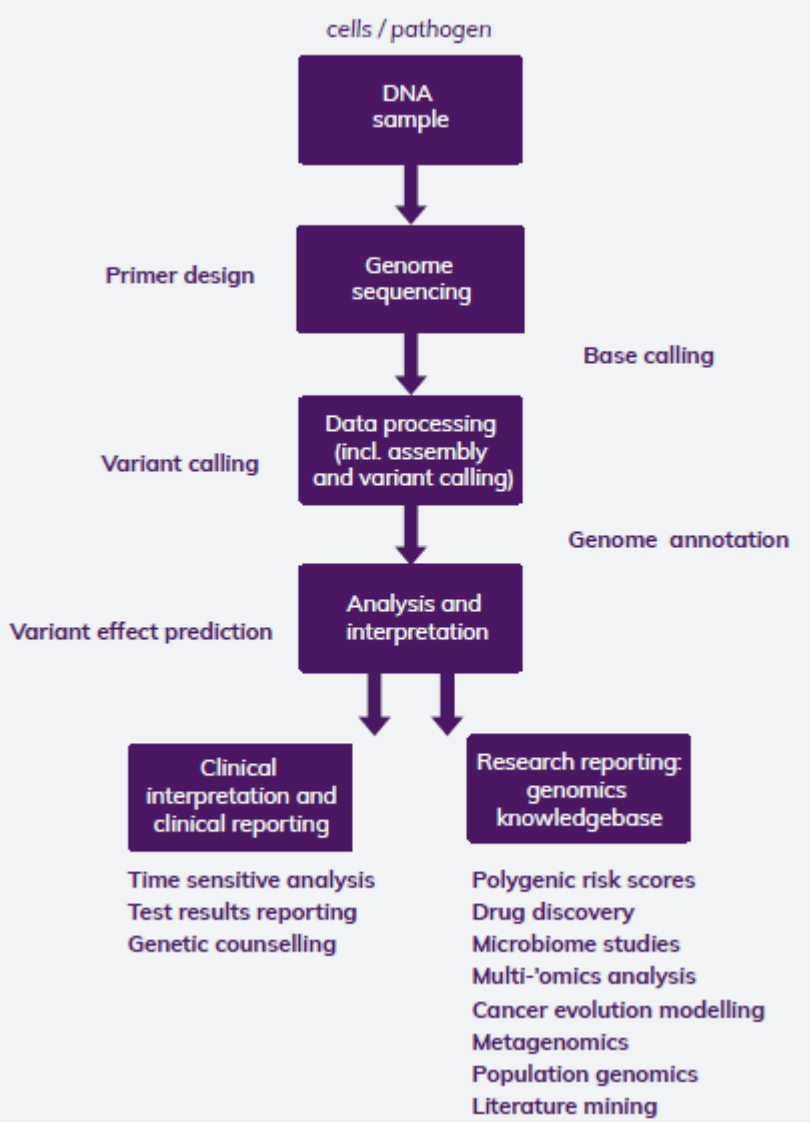

Figura 7. Etapas del análisis genómico y uso de la IA en cada etapa

Estos avances, junto con el aumento de las herramientas open-source y el acceso abierto a la investigación, están impulsando la expansión y el crecimiento del uso de la IA en diferentes tipos de análisis genómicos. Además de los recursos open-source, los

proveedores están incorporando en su software propietario algoritmos de aprendizaje automático en sus análisis genómicos. herramientas y servicios.

13.1 Secuenciación del ADN

La secuenciación del ADN, es decir el proceso de determinar el orden y tipo de las bases de nucleótidos en una molécula de ADN, está dinámicamente y rápidamente evolucionando por un lado por los cambios en las tecnologías de secuenciación, y por otro lado por la incorporación de la IA.

La secuenciación tradicional de Sanger, si bien es precisa, requiere mucho tiempo dado que se analiza gen a gen y muestra a muestra de forma individualizada, lo que la hace poco coste eficiente y poco sensible (no permite detectar variaciones que estén representadas por debajo del 10%, lo que se sabe que tiene importancia clínica en ciertos contextos). La introducción de la tecnología de la secuenciación de próxima generación (NGS) aceleró el proceso de secuenciación, permitiendo por un lado análisis simultáneo de varias muestras y de varios genes a la par que incrementaba la sensibilidad. Sin embargo, el análisis de estos datos requería métodos de análisis computacionales avanzados, que la IA ha podido proporcionar, siendo particularmente importantes los modelos de aprendizaje profundo, que permiten un análisis más rápido y preciso.

Un ejemplo, son los modelos de IA basados en redes neuronales profundas y las redes neuronales convolucionales (CNN) que permiten identificar las variaciones del genoma tanto para detectar errores de secuenciación y mejorar la calidad de los datos de la secuencia.
Dado que la secuenciación, incluso en el caso de las plataformas más avanzadas, requieren fragmentación del genoma, y su alineamiento con el genoma de referencia- Los algoritmos basados en IA son también capaces de realizar el ensamblaje virtual de los fragmentos cortos de ADN de forma rápida y precisa, una tarea que antes requería mucho tiempo y capacidad computacional.

13.2 Identificación de variantes (Variant Calling)

Para realizar el Variant Calling (variaciones de nuestro gen/genoma a estudio con respecto al genoma de referencia) requiere un procedimiento bioinformático laborioso. Aunque los métodos existentes están bien establecidos para la identificación de variantes, se están desarrollando una serie de herramientas basadas en DL con el objetivo de mejorar aún más la precisión de la identificación de variantes.

Una de estas herramientas, 'DeepVariant' de Google, ha superado a los métodos tradicionales, en ciertos conjuntos de datos, a pesar de que el modelo se entrenó sin conocimiento especializado sobre genómica o secuenciación de próxima generación (NGS). La herramienta trata la identificación de variantes como un problema de clasificación de imágenes, algo en lo que el aprendizaje profundo (DL) sobresale, convirtiendo los datos genómicos en imágenes y realizando un análisis de las imágenes para clasificar puntos en el genoma como variantes o no variantes.

Otro ejemplo de software basado en IA optimizado para la identificación de variantes es la herramienta GATK, diseñada y comercializada por el Broad Cancer Institute.

13.3 Anotación del genoma: Filtrado de variantes y predicción de efectos.

Un objetivo último de la anotación del genoma es identificar todos los tipos de elementos funcionales y todas sus apariciones en un genoma. Actualmente, es probable que aún haya clases de elementos genómicos sin descubrir dado el rápido descubrimiento de nuevas clases (como ARN no codificantes (ARNnc)) en los últimos años. Algunas clases de elementos también tienen hasta ahora muy pocas instancias descubiertas.

En términos de aprendizaje automático, estos dos hechos implican que actualmente es imposible realizar un aprendizaje puramente supervisado para todas las clases de elementos ya que este tipo de aprendizaje se base en que los algoritmos aprenden y luego detectan patrones específicos (p. Ej. secuencias de ADN). Los desarrollos con alto rendimiento en secuenciación están generando conjuntos de datos más grandes y detallados que pueden ayudar en el descubrimiento y predicción de estas características genómicas empleando DL para analizar estos conjuntos de datos y descubrir patrones ocultos en el detalle de estos datos.

En el proyecto ENCODE que tiene como objetivo delinear todos los elementos funcionales codificados en el genoma humano, se han adoptado varios enfoques diferentes para enfrentar este gran desafío, empleando métodos de aprendizaje supervisado y no supervisado.

Otra consideración importante es la necesidad de establecer la patogenicidad (probabilidad de causar enfermedades) y accionabilidad (búsqueda de tratamientos eficaces en base a la variante detectada). Esto es especialmente complejo en cáncer, debido a la naturaleza compleja y muchas veces desconocida de la biología del tumor, las muy numerosas, no siempre concordantes y dinámicas bases de datos. Se han usado métodos de machine learning, y están surgiendo enfoques de DL para optimizar la sensibilidad y especificidad de la detección de variantes somáticas.

La patogenicidad está muchas veces relacionada con el efecto de las variantes genéticas sobre las proteínas, ya que la forma de una proteína determina su función y disfunción en la enfermedad.

Una aproximación es mediante validaciones funcionales biológicas que han sido llevadas a cabo manualmente en laboratorios de investigación y publicadas en diversas plataformas. La consulta de dichas bases de datos es una herramienta aplicada a diario. Diversos softwares facilitan la búsqueda de información relativa a la variante de interés tales como franklin, Varsome, etc., extiendo incluso algunas especializadas para genes concretos como el TP53.

Otra forma de establecer la patogenicidad es aplicando los algoritmos incorporados por herramientas como Polyphen, Mutation Taster, CADD, Revel, AlphaMissense, Splice AI, etc, que permiten predecir el impacto del cambio proteico causada por una variante basándose en probabilidades aprendidas de datos genómicos etiquetados.

Al igual que con otros elementos del proceso de genómica, la popularidad del aprendizaje profundo está aumentando para el análisis funcional. Investigadores de Harvard han publicado un software open-source para predecir cómo se pliegan las proteínas en función de su secuencia de aminoácidos.

De manera similar, la herramienta 'AlphaFold' de DeepMind modela las propiedades de una proteína a partir de su secuencia genética.

De forma práctica, para llevar a cabo el análisis de patogenicidad y relación de enfermedad, se llevan a cabo simultáneamente varios de los análisis mencionados. Afortunadamente empiezan a existir plataformas como Franklin o Varsome que en la misma interfaz te presentan dicha información conjuntamente con posibilidad en muchas ocasiones de ir a la fuente de origen mediante hipervínculos, e indicándote como ha llegado al análisis

13.4 Clasificación de variantes no codificantes

El exoma, que abarca los genes que codifican para proteínas, en realidad constituye el 2% del genoma. Durante mucho tiempo se pensó que el resto del genoma era DNA basura o vestigial. Los avances en el conocimiento biológico han permitido ver que, en realidad, ese DNA no codificante tiene, entre otras cosas, funciones regulatorias muy importantes y su alteración puede ser causa importante de enfermedades.

La identificación computacional y la predicción de la variación patogénica de regiones no codificantes sigue siendo un desafío abierto para la genómica humana.

Los algoritmos de IA pueden mejorar nuestra capacidad para comprender la variación genética de las regiones no codificantes.

También es interesante, que el genoma humano tiene aproximadamente entre 20.000 y 25.000 genes. Y sin embargo el número de proteínas, gracias al splicing alternativo, que determina que un gen se transcriba con más o menos exones, se estima que es mayor de 100.000 proteínas.

Los defectos en el empalme génico, también conocidos como splicing, dependiendo de la patología pueden ser responsables de > del 10% de la variación genética patogénica, y con frecuencia difíciles de identificar debido a la complejidad de los potenciadores de empalme (splicing enhancers).

SpliceIA de Illumina (herramienta bioinformática basada en IA de código abierto), es una red neuronal profunda de 32 capas capaz de predecir empalmes canónicos y no canónicos directamente a partir de los datos de una secuencia de unión exón-intrón (Figura 8)., Los resultados de SpliceIA mejoran con la secuenciación de tercera generación usando la información de una secuencia de largo alcance (long-range sequence que además no requiere enriquecimiento previo de muestras, con técnicas que pueden inducir artefactos y variantes artificales), aumentando el porcentaje de precisión típico (57%) de muchas herramientas de predicción de empalme hasta un 95% usando su algoritmo de IA. De igual forma, el modelo de IA fue capaz de identificar variantes candidatas de empalme críptico (ocultas) subyacentes a los trastornos del neurodesarrollo.

Otro enfoque basado en aprendizaje profundo es DeepSEA que mejoró sustancialmente la capacidad para predecir la presencia de sitios hipersensibles a la ADNasa, diversas vías de transcripción y los cambios estructurales en las histonas. Varias extensiones del modelo DeepSEA usadas en secuencias genómicas de familias con trastornos del espectro autista han revelado mutaciones de novo en segmentos no codificantes.

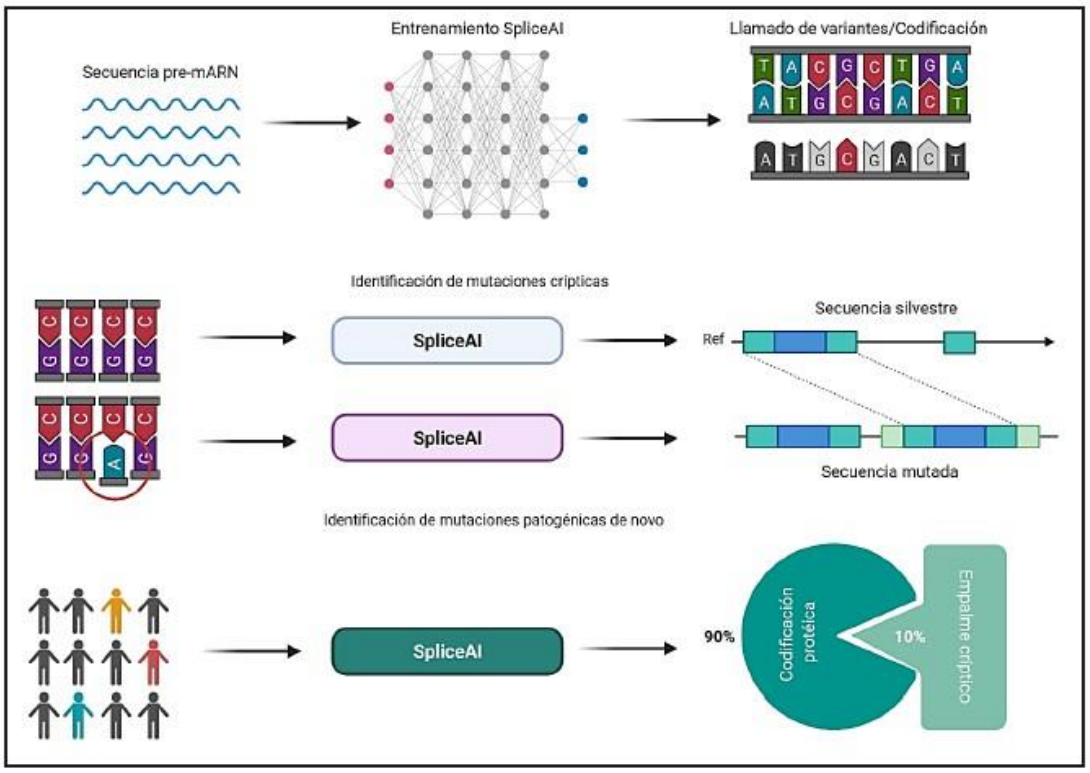

Figura 8. SpliceIA, una red neuronal profunda que modela con precisión el empalme de ARNm a partir de una secuencia genómica y predice la presencia de mutaciones de cripticas no codificantes en pacientes con enfermedades genéticas rara

13.5 Mapeo fenotipo – genotipo

El genoma humano contiene numerosas variantes genéticas patogénicas o potencialmente patogénicas, independientemente del estado de salud del individuo estudiado.

Dentro del análisis funcional es esencial mantener relación con fenotipo del paciente (correlación con fenotipo HPO en estudios germinales y con fenotipo tumoral en estudios somáticos por ejemplo).

Los algoritmos de IA han contribuido a mejorar el mapeo fenotipo-genotipo, combinando la extracción de información derivada del diagnóstico clínico, la integración de imágenes y el uso de datos derivados de registros electrónicos de la historia clínica (EHR).

Un ejemplo de la utilidad de las imágenes dentro del procesamiento del diagnóstico genético es el desarrollo de la estructura facial. La ontología del fenotipo humano enumera 1.007 términos para las anomalías faciales; estas alteraciones están asociadas con 4.526 enfermedades y 2.142 genes. Un médico experto en dismorfología podrá identificar estas anomalías de manera individual emitiendo un diagnóstico puntual y a partir del diagnóstico clínico se puede enfocar la secuenciación de genes específicos basado en el fenotipo dominante. Es habitual, que el diagnóstico clínico y los hallazgos moleculares no coincidan con precisión debido a la similitud fenotípica de múltiples alteraciones. Los modelos de IA pueden ser de enorme utilidad, especialmente en el ámbito asistencial, y en aplicaciones de exoma y genoma, donde se obtiene un enorme número de variantes, para establecer correlaciones entre las variantes obtenidas y el fenotipo.

En esa línea, DeepGestalt, es un algoritmo de análisis de imágenes faciales basado en redes neuronales convolucionales que supera la valoración clínica, siendo lo suficientemente preciso para distinguir entre diagnósticos moleculares asignados a un mismo diagnóstico (Figura 9). La integración de DeepGestalt con PEDIA (sistema de interpretación del genoma), permite que el modelo sea capaz de utilizar características fenotípicas extraídas de fotografías faciales para priorizar con precisión variantes patogénicas candidatas para 105 trastornos monogénicos diferentes en una población de 679 casos analizados. Un ejemplo, de aplicación del algoritmo es la identificación del síndrome de Cornelia de Lange, diagnosticado por IA con una exactitud del 98,6% (75-85% para la evaluación clínica), o el síndrome de Angelman determinado con una precisión del 92% (72% para el diagnóstico clínico).

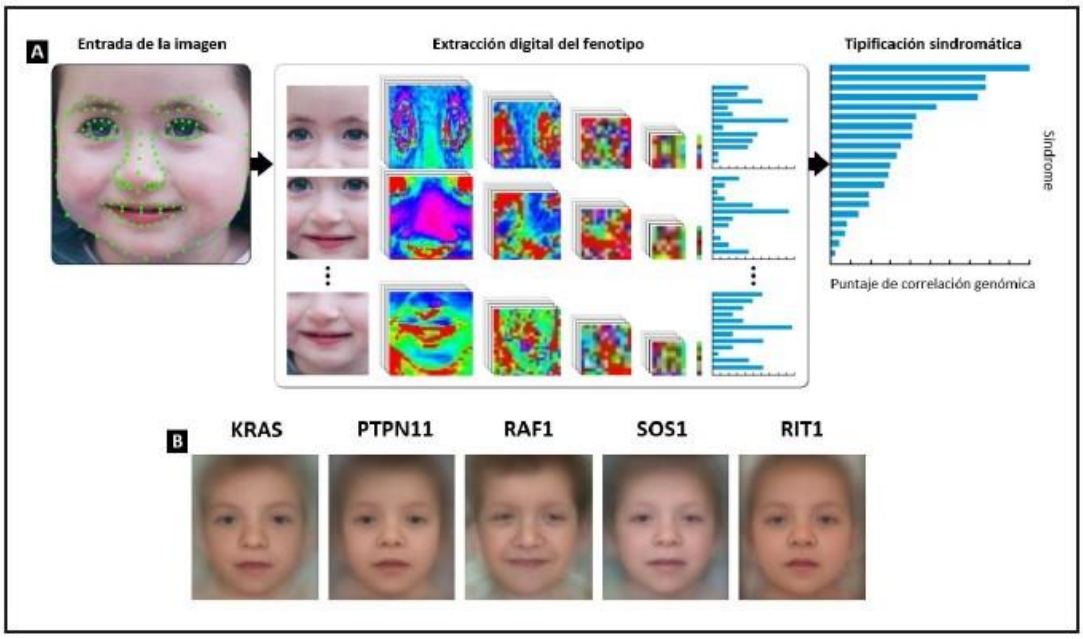

Figura 9. DeepGestalt, flujo de alto nivel. La imagen de entrada se procesa primero para lograr una detección del perfil facial, para la detección de puntos de referencia y su alineación. Después del preprocesamiento, la imagen de entrada se recorta en regiones. Luego, cada región alimenta una red neuronal convolucional para obtener un vector softmax que indica su correspondencia con cada síndrome en el modelo. Los vectores de salida de cada estrato de la red neuronal convolucional se agregan y clasifican para obtener la lista que se correlaciona con el patrón genómico. El histograma del lado derecho representa los síndromes de salida del DeepGestalt, ordenados por la puntuación de similitud. B. Fotografías compuestas de pacientes con síndrome de Noonan con diferentes genotipos y diferencias faciales sutiles caracterizados a través del DeepGestalt.

Otro ejemplo del mapeo es el uso para enfermedades genéticas raras, los datos fenotípicos pueden incluir patrones faciales específicos ya que muchas afecciones genéticas sindrómicas están asociadas con dismorfologías craneofaciales.

La empresa FDNA, ha creado un análisis de imágenes faciales, basado en el uso de teléfonos inteligente, llamado Face2Gene, para clasificar rasgos faciales distintivos en fotografías de personas con trastornos congénitos y del neurodesarrollo.

El sistema utiliza algoritmos de aprendizaje entrenados con decenas de miles de imágenes de pacientes para distinguir sutiles patrones faciales prediciendo el síndrome genético más probable y sugiriendo genes para priorización en la etapa de análisis en función de la asociación de ciertos genotipos con síndromes específicos.

Herramientas, como Exomiser y eXtasy, incorporan información de fenotipo y genotipo para puntuar y clasificar variantes que causan enfermedades. Otro ejemplo son el uso de extensiones del algoritmo ExPecto que revelaron una mejoría en la capacidad pronóstica de diferentes perfiles de expresión génica extraídos a partir de secuencias de ADN germinal y somático.

13.6 Análisis de datos multiómicos y multimodales integrados

La secuenciación del ADN es sólo una de las muchas herramientas para examinar la genómica funcional. La capacidad de medir y analizar otros componentes biológicos y sobre todo la relación entre los mismos puede ser igualmente importante para comprender la importancia de la variación genómica. De hecho, muchos pacientes remitidos para pruebas genéticas se enfrentarán a un resultado no concluyente debido a los límites de nuestro conocimiento científico, variaciones en los procesos celulares, cambios en el entorno y variantes no heredadas. Distintos factores pueden afectar a los procesos moleculares entre genotipo y fenotipo.

La biología de sistemas ha revolucionado la investigación biomédica permitiendo la generación de estudios multicapa que integran datos genómicos como información multiómica (proteoma, epigenoma, metaboloma, microbioma, ...). Esto proporciona una visión más completa de los procesos y sistemas biológicos, lo que conduce a una mejor comprensión de la enfermedad, especialmente en comparación con el análisis de una sola capa.

El número de dimensiones o características que se miden son muy altos (p.e. datos de expresión genética obtenido del análisis de miles de genes). Examinar estos grandes conjuntos de datos de alta dimensionalidad y múltiples requiere una elevada exigencia computacional, lo que has supuesto un desafío hasta la fecha. Sin embargo, la mejora de la potencia del hardware (e.g. con el empleo de GPU) en combinación con redes neuronales de aprendizaje profundo, que permiten procesar grandes conjuntos de datos y modelar relaciones complejas, está abriendo oportunidades para analizar y obtener nuevos conocimientos a partir de estos conjuntos de datos, ofreciendo la ventaja de ser capaces de examinar grandes volúmenes de tipos de datos dispares para descubrir patrones.

Para comprender mejor la complejidad biológica de las enfermedades, numerosos esfuerzos en investigación están intentando aplicar un enfoque ómico integrador.

Normalmente, los estudios combinan datos de múltiples tecnologías ómicas junto con registros médicos y, en algunos casos, incluso los monitores ambientales.

El despliegue clínico está actualmente limitado por los costos asociados con la recopilación de datos, almacenamiento y análisis, así como cuestiones de estandarización, reproducibilidad y utilidad. Aunque la medicina está muy lejos de los diagnósticos multiómicos de rutina, la creciente tendencia a combinar análisis avanzados con conjuntos de datos biomédicos detallados será clave para avanzar en la medicina personalizada y abordar las enfermedades multifactoriales.

A continuación, se describen algunos ejemplos de los análisis multimodales y multiómicos.

Las infecciones del tracto respiratorio inferior (IVRI) son la principal causa de muertes relacionadas con enfermedades infecciosas en todo el mundo y son difíciles de distinguir de los síndromes respiratorios no infecciosos. Los investigadores han implementado métodos de aprendizaje automático para el análisis integrado de datos derivados de tres elementos centrales de infecciones agudas de las vías respiratorias (el patógeno, el microbioma de las vías respiratorias y la respuesta del huésped) para obtener un diagnóstico preciso de LRTI en una cohorte prospectiva de pacientes críticos. Además de combinar datos tanto del huésped como del patógeno, el análisis incorporó datos de secuencias de ARN y ADN.

El análisis de imágenes histológicas (a nivel de tejido) ha sido una herramienta importante en el diagnóstico y pronóstico del cáncer durante más de un siglo y es otra área donde el aprendizaje automático puede respaldar los análisis fenotípicos. Existen numerosos trabajos que demuestran el potencial del aprendizaje profundo, en particular para facilitar flujos de trabajo en patología digital. Algunas de estas investigaciones buscan combinar el análisis de la imagen patológica con la genómica u otras mediciones basadas en ómicas para mejorar los modelos de predicción.

En un estudio, se aplicó un enfoque de aprendizaje profundo que integra imágenes histológicas y genómicas. Los datos predijeron la supervivencia general de los pacientes diagnosticados con tumores cerebrales con igual o mayor precisión que los expertos humanos. El enfoque utilizó el aprendizaje profundo para aprender patrones visuales (a partir de las imágenes) y biomarcadores moleculares asociados con los resultados de los pacientes.

En otro estudio, se utilizó un modelo integrador que combina datos ómicos con imágenes de histopatología que proporciono mejores resultados en el pronóstico en pacientes con adenocarcinoma de pulmón en estadio 1 en comparación con predicciones realizadas solo con imágenes o solo con análisis ómicos.

13.7 Otros usos de la IA en genómica

Además de los usos descritos anteriormente, el empleo de IA con genómica abarca más casos de usos que se describen a continuación:

- Genética poblacional: comprende el estudio de la variación genética dentro de las poblaciones y los factores que dan forma a esta variación en el espacio y el tiempo. El aprendizaje profundo podría alcanzar un mayor poder predictivo que la estimación estadística de los enfoques clásicos utilizados en genética poblacional. En comparación con los otros métodos, el aprendizaje automático hace menos suposiciones y es independiente de los procesos utilizados para crear conjuntos de

datos, por lo que podría reconocer mejor los fenómenos tal como son en la naturaleza, en lugar de como los científicos eligen representarlos en un modelo.

- Puntuaciones poligénicas (PGS): distinguen el efecto acumulativo del polimorfismo nucleótido simple (SNPs), que es un tipo de variante genómica, que individualmente tienen un pequeño efecto en un rasgo. Se han desarrollado como un medio para investigar la base genética de rasgos complejos, que están influenciados por múltiples SNP. Aunque los PGS aún no son ampliamente utilizados en la clínica, existe interés en utilizarlos para la predicción de enfermedades comunes. Se sugiere que ciertos métodos de aprendizaje automático podrían mejorar la potencia de la capacidad predictiva de estos modelos ya que hacen menos suposiciones y tienen mayor capacidad para reconocer patrones en datos fuertemente correlacionados. También podrían usarse para desarrollar más métodos dinámicos que explican mejor las interacciones complejas (p.e. entre genes y otros factores, como la influencia de los cambios de los factores genéticos a lo largo de la vida humana).

- Estudio del microbioma: los estudios del microbioma examinan todo el material genético dentro de una microbiota: la colección de microorganismos presentes en sitios particulares del cuerpo (p.e. el intestino, la piel). Se están aplicando métodos de aprendizaje automático en la investigación del microbioma para clasificar microbios específicos, secuencias en una muestra, e investigar el vínculo entre los cambios dinámicos en la microbioma y fenotipo y enfermedad del huésped.

- Análisis unicelular: el análisis de una sola célula (singel cell) es la aplicación de tecnologías ómicas a células individuales. Loa avances en las técnicas de secuenciación unicelular están ayudando a capturar la complejidad y diversidad de poblaciones celulares y proporcionar mayor detalle sobre las características moleculares de una enfermedad. Se están entrenando algoritmos de aprendizaje automático para analizar el volumen creciente de datos de una sola célula y abordar problemas de interpretación relacionados con la calidad de los datos, el ruido y heterogeneidad.

- Modelado de la evolución del cáncer: El modelado de la evolución del cáncer tiene como objetivo determinar el orden temporal de los cambios genéticos que ocurren en diferentes cánceres a medida que evolucionan y cambian. Esta información podría aportar nuevas estrategias para la detección temprana y para anticipar la progresión de la enfermedad. Hay varias investigaciones donde se están desarrollando métodos de aprendizaje automático para rastrear la evolución del cáncer y determinar qué cambios genéticos son impulsores del crecimiento del cáncer.

- Análisis urgentes y reanálisis periódicos: En términos generales, el valor del aprendizaje automático proviene de la oportunidad de

acelerar descubrimientos de importancia para la medicina genómica. Un grupo de investigadores en el Hospital RadyChildren de San Diego (USA), ha aplicado esta noción al utilizar aprendizaje automático para respaldar el análisis rápido de datos de secuenciación del genoma completo (WGS) para el diagnóstico de recién nacidos críticamente enfermos en menos de 24 horas. Los autores del estudio sugieren que un proceso de análisis automatizado tan rápido podría tener un elevado número de casos de uso, incluido el diagnóstico provisional inmediato y la reevaluación independiente en los casos en que la interpretación manual no proporcione un diagnóstico, y la revisión periódica de casos no resueltos.

- Farmacogenética y descubrimiento de medicamentos: Los mismos fármacos administrados a diferentes pacientes pueden tener diferente efecto entre otros motivos por las variaciones farmacogenéticas, que implican una diferente, absorción, metabolización, del fármaco en el individuo. La mortalidad, debida a efectos secundarios de los fármacos no ajustados en base a estas diferencias se sabe es elevada. Muchas empresas farmacéuticas (p.e. Astrazeneca con Benevolent AI) tienen en marcha iniciativas de I+D basadas en IA para utilizar el aprendizaje automático sobre datos genómicos con el propósito de identificar subenfermedades, biomarcadores, reutilización de fármacos y predicción sobre la respuesta de los fármacos entre otras.

- Edición del genoma: la edición del genoma, mediante la cual se eliminan, añaden o alteran secciones de ADN, es otra área de la investigación terapéutica que se ve facilitada por el aprendizaje automático. Las técnicas de edición del genoma se utilizan ampliamente en el ámbito de la investigación para averiguar el papel de los genes y las secuencias del ADN, pero también se utilizan cada vez más con fines terapéuticos, para reemplazar o alterar un gen defectuoso en pacientes. Se están entrenando algoritmos de aprendizaje automático y aprendizaje profundo para aumentar la eficiencia y precisión de la implementación de CRISPR (actualmente la más versátil, barato y sencilla herramienta para la manipulación genética). Se han desarrollado métodos algorítmicos para predecir la actividad del sistema de edición, los cambios exactos resultantes de las ediciones, y efectos fuera del objetivo: cambios no deseados en el ADN que pueden complicar o dificultar el uso de la tecnología.

14 Futuro de la Inteligencia Artificial en ómica

La aplicación de la Inteligencia Artificial sobre datos ómicos ha abierto una ventana de amplias oportunidades para el desarrollo de la medicina de precisión y la medicina personalizada (MP) así como para el desarrollo de medicamentos y la realización de ensayos clínicos.

La MP y la IA integradas en la atención médica tienen el potencial de producir diagnósticos más precisos, predecir el riesgo de enfermedad antes de que aparezcan los síntomas y diseñar planes de tratamiento personalizados que maximicen la seguridad y la eficiencia.

Como ya expuso David Ledbetter en un taller celebrado en marzo de 2023, el coste de la secuenciación y el análisis del genoma es cada vez más asequible. Las nuevas plataformas permiten realizar estudios en el día y obtener datos a tiempo real. Es en la interpretación donde existen más limitación, dado que existen todavía numerosas incógnitas, pero es sin duda en este área donde la IA está permitiendo un avance exponencial que permitirán no sólo unos tiempos de respuesta más rápidos para los médicos, sino que se pueda asegurar una interpretación eficaz, actualizada y personalizada, con recomendaciones terapéuticas más eficaces.

En numerosas ocasiones, conocer que genes son precisos evaluar en base a una sospecha clínica o un fenotipo puede ser complicado, dada entre otros motivos, a que el conocimiento biológico cambia a diario. Estar actualizado y establecer las asociaciones a tiempo real en todos los casos, no es sencillo para la mente humana.

Por otro lado, el estudio de unos genes concretos para la caracterización de una patología hoy puede ser insuficiente en seis meses, ante la rápida evolución del conocimiento y la aparición casi a diario de biomarcadores. Esto sumado al ya mencionado abaratamiento del coste de la secuenciación, ha hecho que sobre todo en patología hereditaria se esté optando por realizar exomas o genomas, tal y como ya predecía Ledbetter, y analizar solo virtualmente los genes que nos interesen dejando la posibilidad de ampliar el número de genes analizados, si surgen nuevos biomarcadores. Así mismo el disponer del genoma, permite, que, mediante el uso secundario de los datos, podamos encontrar más rápidamente esos nuevos biomarcadores que desconocemos a día de hoy. Esto hace pone de manifiesto, que el límite entre uso primario y secundario no debe ser tan rígido y distante. Los mismos datos, bien utilizados, con entornos seguros y controlados deben permitir uno uso primario y secundario más eficiente

Surge por tanto la necesidad de garantizar que se cuente con la infraestructura informática necesaria tanto para almacenar las grandes cantidades de datos genómicos como para permitir su reanálisis. Si bien la secuencia del genoma de un individuo es

estática, la interpretación es dinámica, y debe haber capacidad para volver a analizar los genomas.

Para potenciar el uso de la IA se recomienda dirigir las investigaciones futuras y áreas de innovación hacia:

- Aprendizaje profundo (DL) y descubrimiento de fármacos: explorar el uso de técnicas de DL, como redes neurales, para el descubrimiento de fármacos, centrando los esfuerzos en la detección virtual, diseño de nuevas moléculas y predicción de interacciones entre otras. A través de la IA, se podrá predecir cómo los tratamientos afectarán a diferentes individuos basándose en su perfil ómico, lo que permitirá terapias más eficaces y con menos efectos secundarios.

- Integración de datos multiómicos: investigando métodos para integrar datos multiómicos, como genómica, transcriptómica, proteómica, metabolómica y microbiómica, para obtener información completa sobre los mecanismos de las enfermedades y las respuestas a los medicamentos. Desarrollar enfoques basados en IA para analizar e interpretar conjuntos de datos biológicos complejos, identificar biomarcadores y personalizar estrategias de tratamiento en función de perfiles moleculares individuales. Esto facilitara el avance en la medicina personalizada.

- Optimización de ensayos clínicos con IA: desarrollo de algoritmos basados en IA para optimizar el diseño de ensayos clínicos, el reclutamiento de pacientes y la selección de criterios de valoración. Explorar métodos para aprovechar la evidencia del mundo real, los registros médicos electrónicos, los sensores portátiles y las tecnologías de salud móviles para agilizar las operaciones de ensayos clínicos, mejorar la participación de los pacientes y mejorar la calidad de los datos y el cumplimiento normativo.

- Monitorización en tiempo real y análisis predictivo: desarrollar sistemas basados en IA para monitorización en tiempo real, análisis predictivo y sistemas de alerta temprana en entornos de prestación de atención médica. Los esfuerzos de investigación pueden centrarse en el desarrollo de algoritmos para predecir el deterioro de los pacientes, los eventos adversos y las readmisiones hospitalarias, lo que permite intervenciones proactivas y una prestación de atención personalizada.

- Procesamiento del lenguaje natural (PLN) en el ámbito sanitario: promover la aplicación de técnicas de procesamiento del lenguaje natural (PLN) en el ámbito sanitario para extraer información de notas clínicas no estructuradas, literatura médica y contenido generado por los pacientes. Desarrollar modelos de PLN para el apoyo a la toma de decisiones clínicas, la codificación y documentación automatizadas y la gestión de la salud de la población, mejorando la recuperación de información y el descubrimiento de conocimientos en entornos sanitarios.

Dada la importancia de la IA en la genómica y la MP, un estudio reciente evaluó los factores que podrían mejorar la aplicación clínica de la IA en este campo, concluyendo que existe una necesidad significativa de investigación y desarrollo informático para aprovechar al máximo el potencial clínico de estas tecnologías, siendo esencial la creación de conjuntos de datos más grandes para replicar el éxito de la aplicación de IA en otros campos.

Igualmente se hace necesario establecer estándares de datos ómicos para conseguir la escalabilidad efectiva de tales tecnologías en todas las instituciones y organizaciones sanitarias. Esto es especialmente relevante en este ámbito, dado que los datos aislados no son interpretables

La disponibilidad de grandes conjuntos de datos de genómica funcional para el entrenamiento de algoritmos de IA, está impulsando el aumento de la productividad de la IA. Actualmente, las aplicaciones más prometedoras de la IA en la genómica clínica parecen ser la extracción de información fenotípica profunda de imágenes, registros médicos electrónicos (EHR) y otros dispositivos médicos para informar el análisis genético posterior.

Igualmente, los algoritmos de aprendizaje profundo (DL) están demostrado ser tremendamente prometedores en una variedad de tareas de genómica clínica, como la identificación de variantes, la anotación del genoma y la predicción del impacto funcional, ya que las redes neuronales profundas (convolucionales y recurrentes) son particularmente adecuadas para el análisis de datos genómicos.

La tendencia es que las herramientas de IA más generalizadas se conviertan en el estándar en estas áreas, especialmente para tareas de genómica clínica donde la inferencia a partir de datos complejos (es decir, la identificación de variantes) es una tarea recurrente con frecuencia.

Sin embargo, la utilidad de los algoritmos de IA como la herramienta definitiva de apoyo a la toma de decisiones clínicas para predecir fenotipos humanos complejos comunes no se ha demostrado de forma convincente. El aumento de los esfuerzos a escala de biobanco con recopilación longitudinal de datos de salud, como el Biobanco del Reino Unido y el Programa de Investigación All of Us, proporcionará potencialmente los conjuntos de datos de entrenamiento necesarios para hacer realidad este objetivo. Dada la dependencia de la IA de conjuntos de datos de entrenamiento a gran escala, es probable que la recopilación escalable de datos de fenotipos, y no de datos genómicos, sea la barrera más difícil de superar para hacer realidad esta ambición. La tecnología moderna de secuenciación de ADN permite la generación de datos genómicos de manera uniforme y a escala, pero la recopilación de datos de fenotipos requiere numerosos modos de recopilación de datos y tiende a ser lenta, costosa y muy variable en los sitios de recopilación. Por último, la interpretabilidad y la identificación del sesgo

de las máquinas son esenciales para la amplia aceptación de la tecnología de IA en cualquier modalidad de diagnóstico clínico.

Un desafío inherente a la genómica clínica es la comunicación de información compleja y resultados de pruebas a pacientes y médicos convencionales. Muchos sistemas de salud tienen un grupo limitado de expertos en genética (genetistas clínicos y asesores genéticos) que podrían verse abrumados por el creciente volumen de pruebas genéticas. El empleo de Chatbots de IA que permitan comunicar a los médicos y pacientes sus dudas sobre la información genética puede complementar y ampliar asesoramiento genético. El uso creciente de pruebas genéticas directas al consumidor (DTC) está acentuando aún más la necesidad de fortalecer la alfabetización genómica entre el público y los profesionales de la salud. En una época en la que muchas personas utilizan Internet para buscar información de salud, cuidadosamente Los chatbots diseñados y rigurosamente probados podrían desempeñar un papel constructivo en la difusión del conocimiento genómico. Algunos ejemplos de proyectos incipientes de chatbots son Genetic Information Assistant de Clare Genetics, GeneFax de OptraHealth u OptraGuru (que se puede consultar a través de las herramientas de asistente virtual Amazon Alexa y Microsoft Cortana).

Como se ha expuesto anteriormente, el paradigma de la medicina de precisión y personalizada, que está tomando impulso en todos los sistemas y organizaciones del cuidado de la salud, es un catalizador para el aumento en investigación y desarrollo impulsando el crecimiento de la IA en el ámbito de la genómica. Con los rápidos avances en la tecnología de IA, y su rentabilidad, se están desarrollando nuevas soluciones de software impulsadas por IA que están diseñadas específicamente para la industria de la genómica, emergiendo un mercado de soluciones software y servicios en la nube de IA sobre datos genómicos. Este mercado mundial se evaluó en 2023 en 560 millones US$, esperando que se alcance un valor de 16.431,16 millones US$ en 2033 con un crecimiento aproximado del 40%.

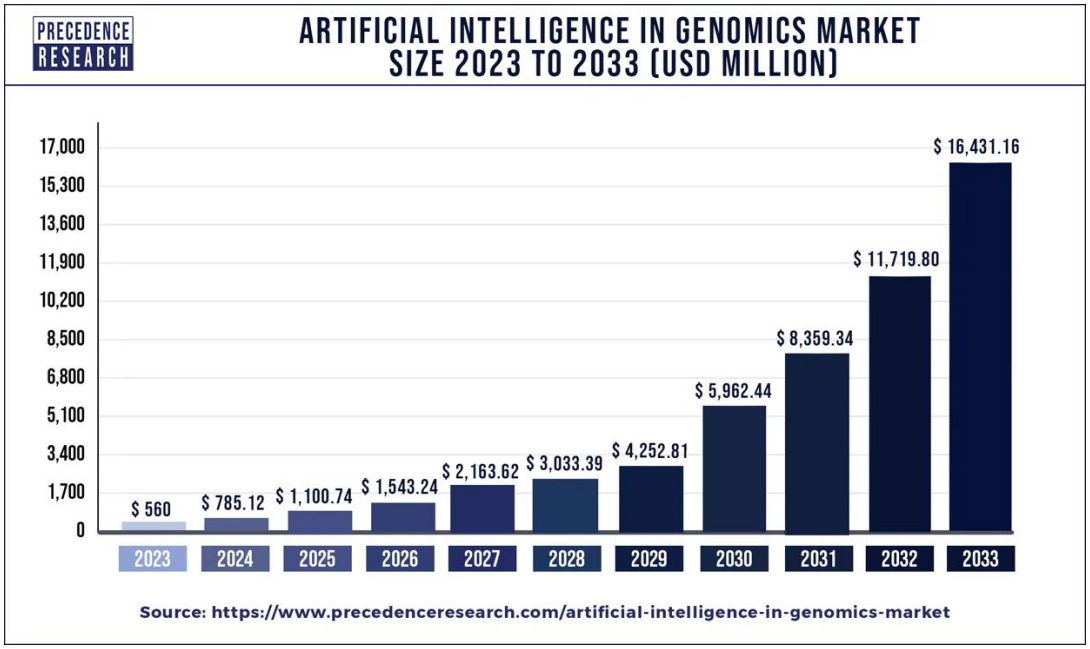

Figura 10. Previsión económica de crecimiento del mercado mundial en soluciones de IA en genómica

En los dos últimos años han surgido distintas iniciativas comerciales ofreciendo en la nube servicios y software basados en IA para procesamiento de datos genómicos, algunos ejemplos son:

- En octubre de 2022, Illumina puso en marcha una colaboración estratégica de investigación con AstraZeneca para acelerar el descubrimiento de dianas farmacológicas
- En mayo de 2023, Google Cloud lanzó soluciones impulsadas por IA para acelerar de forma segura el descubrimiento de fármacos y la medicina de precisión.
- En enero de 2024, QIAGEN anunció planes para acelerar las inversiones en el negocio de bioinformática QIAGEN Digital Insights
- En marzo de 2024, NVIDIA Healthcare lanzó microservicios de IA generativa para avanzar en el descubrimiento de fármacos, la tecnología médica y la salud digital

Las principales compañías del sector tecnológico (IBM, NVIDIA Corporation, Microsoft entre otras) están apostando por este mercado estando presentes en dicho mercado.

El futuro de las ciencias ómicas en el ámbito de los cuidados de la salud no se entiende sin la IA basada en el aprendizaje profundo, que permite automatizar muchas de las tareas manuales involucradas en la investigación genómica, reduciendo la necesidad de mano de obra humana y haciendo que el proceso sea más rentable. Muchas de las necesidades en el procesamiento de datos genómicos quedaran cubiertas por soluciones software o servicios que ofrecen las compañías que están dinamizando el mercado de la IA. Las instituciones y empresas que requieran de sus servicios tendrán

que velar porque las soluciones que contraten cumplan con las normativas regulatorias (FDA, Marcado CE, GDPR, HIPPA, ...), dispongan de una adecuada seguridad para asegurar la privacidad de los datos y dispongan de interfaces de interoperabilidad para la integración de la información obtenida con sus sistemas (EHR, PACS, ...).

15 Conclusiones

La inteligencia artificial (IA) ha tenido un impacto significativo en las ciencias ómicas. Las técnicas de machine learning y deep learning representan una ayuda extraordinaria para resolver problemas complejos, como la detección de patrones y la predicción de resultados.

El desarrollo de modelos de inteligencia artificial implica seguir una serie de pasos fundamentales, desde la recopilación y preparación de los datos hasta la validación de los resultados. Esta última etapa es crucial para asegurar la precisión y la aplicabilidad de los modelos a datos nuevos y no solo a los utilizados en el entrenamiento. Entre las técnicas de validación más destacadas se encuentran la validación cruzada, la matriz de confusión, el F1-Score y el AUC-ROC.

La aplicación de la IA en genómica no es reciente; se ha utilizado durante años y sigue evolucionando, especialmente en las fases de secuenciación del ADN y en el procesamiento de datos genéticos. La IA ha supuesto un avance decisivo, contribuyendo a mejoras significativas en la calidad y precisión de las predicciones en los análisis genómicos primarios, secundarios y terciarios, además de acelerar los procesos. Los modelos actuales han permitido mejorar la selección de genes a estudiar, establecer relaciones entre genotipo y fenotipo, y ofrecer recomendaciones terapéuticas basadas en la integración de datos ómicos.

Así como los autoanalizadores en los laboratorios permitieron eliminar los recuentos manuales, incrementando la cantidad y calidad de estudios realizados, las herramientas basadas en IA se han vuelto esenciales para aplicar las ciencias ómicas en un enfoque personalizado y de precisión. Los genetistas clínicos necesitan asistentes tecnológicos que les ayuden a analizar la información de los biomarcadores, dado que la mente humana no puede procesar rápidamente toda la información cambiante de los posibles casos clínicos que se presentan diariamente.

Es fundamental garantizar la correcta custodia de los datos en sistemas que aseguren la integridad y el acceso controlado. Estas soluciones deben ser escalables y adaptarse a las exigencias computacionales de la IA. Asimismo, es esencial que los datos utilizados en los análisis primarios sean validados y contrastados, lo que requiere un profundo conocimiento de los modelos y limitaciones de las soluciones de software empleadas. Además, dichas soluciones deben validarse con múltiples muestras y, cuando sea posible, mediante controles interlaboratorio para asegurar la consistencia de los resultados.

Persisten retos como la calidad de los datos y la necesidad de interpretarlos, así como la adopción de estándares que garanticen el éxito de la IA en los proyectos ómicos. El diseño de estos sistemas debe considerar la continua evolución tecnológica; es crucial

estar al día con las últimas investigaciones y técnicas para asegurar el éxito de estos proyectos. También es necesario reducir la brecha entre el uso primario y secundario de los datos. En muchos casos, los fracasos terapéuticos se deben a que no se están utilizando los biomarcadores o las dianas terapéuticas adecuadas. El esfuerzo conjunto y el trabajo colaborativo en entornos federados permiten avanzar más rápidamente en este campo.

En cuanto a las regulaciones, deben garantizar un uso racional, seguro y eficiente de los recursos, adaptándose a los avances tecnológicos para no frenar la innovación. La IA conlleva riesgos, como los relacionados con la privacidad de los datos, los sesgos en los algoritmos y la equidad en los resultados. Normativas como el GDPR y la Ley de Inteligencia Artificial de la Unión Europea buscan asegurar un uso seguro y responsable de esta tecnología.

La combinación de IA y ciencias ómicas tiene un gran potencial para revolucionar la medicina personalizada y de precisión, facilitando tratamientos más efectivos y con menos efectos secundarios a través del análisis de perfiles ómicos individuales. A medida que el aprendizaje automático acelera el ritmo del descubrimiento, el reto será cerrar la brecha entre la investigación y su aplicación clínica. Los sistemas y empresas de salud deben prepararse para incorporar a sus servicios la medicina personalizada y de precisión, abarcando desde la secuenciación de ADN hasta la medicina genómica o la farmacogenética.

Las organizaciones deben planificar estratégicamente el despliegue de la gestión de información ómica, considerando factores organizativos, funcionales, tecnológicos y legales. La aplicación de las técnicas ómicas requiere una colaboración interdisciplinaria, en la que científicos, bioinformáticos, expertos en IA y médicos trabajen de forma coordinada validando cada pase del proceso. La propia IA puede ayudar a reducir la carga de trabajo al automatizar el análisis e interpretación de datos. Además, es necesario implementar una infraestructura tecnológica adecuada, que incluya hardware de alto rendimiento, software especializado y soluciones que consideren los distintos paradigmas de computación, ya sea local o en la nube.

Finalmente, es imprescindible abordar los aspectos éticos y legales relacionados con la custodia y manipulación de la información genética y los datos personales, ya que estos factores pueden condicionar el alcance de los proyectos, la forma de abordarlos e incluso las soluciones técnicas a implementar.

Anexo I

Ejemplos de compañías que realizan algún tipo de actividad de IA con genómica.

Compañía	Caso de uso	Solución de IA
Ardigen	Descubrimiento de conocimiento	Algoritmos para descubrimiento de bioamarcadores y análisis del microbioma
BenevolentAI	Descubrimiento/desarrollo de fármacos	Descubrimiento de farmacias basado en datos biomédicos extraídos o inferidos
BostonGene	Apoyo a la decisión terapéutica	Obtener recomendaciones sobre terapias contra el cáncer
Cambridge Cancer Genomics	Apoyo a la decisión terapéutica	Obtener recomendaciones sobre tratamientos personalizados contra el cáncer
Clear Genetics	Asesoramiento/informes	Chatbot de IA para conversar con pacientes sobre genética
Congenica	Anotación de variantes	Sapientia: Algoritmos de anotación y priorización de variantes genómicas
Deep Genomics	Descubrimiento/desarrollo de fármacos	Algoritmos para descubrimiento de fármacos
Desktop Genetics	Edición de genoma	Deskgen AI: optimiza bibliotecas CRISPR de secuenciazión de genes
Fabric Genomics	Interpretación de variantes	Interpretación automatizada basada en fenotipos
FDNA	Fenotipado	Análisis facial basado en aprendizaje profundo para fenotipado de enfermedades raras y priorización de variantes basada en fenotipos
Freenome	Detección y tratamiento temprano de cáncer	Detección y el tratamiento tempranos del cáncer
Google (Brain)	Identificación de variantes	DeepVariant: Detección de variantes genéticas
Healx	Reutilización de medicamentos	Búsqueda de fármacos en enfermedades raras
IBM	Minería de literatura	Watson for Genomics: Extracción de información genómica de la literatura
Lantern Pharma	Descubrimiento/desarrollo de fármacos	Plataforma para terapias oncológicas de precisión
Literome	Minería de literatura	Sistema automatizado para extraer conocimiento genómico de PubMed
OptraHealth	Asesoramiento/informes	Asistente digital y chatbot de IA para conversar con pacientes sobre genética
Perthera	Apoyo a la decisión terapéutica	análisis de IA para búsqueda de terapias para el cáncer
Philips	Descubrimiento de conocimiento	La plataforma 'IntelliSpace Genomics' combina procesos personalizables con

Compañía	Caso de uso	Solución de IA
		aprendizaje profundo para obtener nuevos conocimientos
Sequana Health	Edición de genoma	Sistema IA de edición genómica CRISPR
SOPHiA Genetics	Interpretación de variantes	El explorador de genomas de Alamut Genova integra varias herramientas y algoritmos de predicción de patogenicidad de variantes sin sentido
Verge Genomics	Descubrimiento/desarrollo de fármacos	Empresa de descubrimiento de fármacos basada en aprendizaje automático centrada en enfermedades neurodegenerativas

Bibliografía

Adzhubei, I. A., Schmidt, S., Peshkin, L., et al. (2010). A method and server for predicting damaging missense mutations. Nature Methods, 7(4), 248-249.

Ainscough, B. J., Barnell, E. K., Ronning, P., et al. (2018). A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. Nature Genetics, 50(12), 1735-1743.

Al-Majzoub, A., Al-Ali, A., & Bou Hamdan, K. (2023). Healthcare Predictive Analytics Using Machine Learning and Deep Learning Techniques: A Survey. Journal of Electrical Systems and Information Technology. Disponible en: https://jesit.springeropen.com/articles/10.1186/s43067-023-00104-7

Alvarellos, M., Sheppard, H. E., Knarston, I., Davison, C., Raine, N., Seeger, T., Prieto Barja, P., & Chatzou Dunford, M. (2023). Democratizing clinical-genomic data: How federated platforms can promote benefits sharing in genomics. BMC Genomics, 24, 23. https://doi.org/10.1186/s12864-023-08901-3

Amstutz, P., Crusoe, M. R., TijaníC, N., et al. (2016). Especificaciones comunes del lenguaje de flujo de trabajo, v1.0. Disponible en: https://w3id.org/cwl/v1.0/

Auslander, N., Wolf, Y. I., & Koonin, E. V. (2019). In silico learning of tumor evolution through mutational time series. Proceedings of the National Academy of Sciences, 116(19), 9501-9510.

Bagheri, H., Muppirala, U., Masonbrink, R. E., Severin, A. J., & Rajan, H. (2019). Shared data science infrastructure for genomics data. BMC Bioinformatics, 20(436). https://doi.org/10.1186/s12859-019-2967-2

Beacon v2 Project Website. (n.d.). Disponible en: https://beacon-project.io/

Bera, K., Schalper, K. A., Rimm, D. L., et al. (2019). Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nature Reviews Clinical Oncology, 16(11), 703-715.

Bhatele, A., Koop, R., Middleton, D., Robey, S., & Sedlmair, M. (2020). Accelerated Genomics Data Processing using Memory-Driven Computing. ResearchGate. https://www.researchgate.net/publication/339094965_Accelerated_Genomics_Data_Processing_using_Memory-Driven_Computing

BioLinux: computación bioinformática preconfigurada y bajo demanda para la comunidad genómica. BMC Bioinformática, 13, 42.

BioSpace. (2023). Artificial Intelligence in Genomics Market Size to Reach US$ 16,431.16 Million by 2033. Disponible en: https://www.biospace.com/artificial-intelligence-in-genomics-market-size-us-16-431-16-million-by-2033

Boza, V., Brejova, B., & Vinar, T. (2017). DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. PLoS One, 12(6), e0178751.

Cardona, A. F., Ruíz-Patiño, A., Jaller, E., Rodríguez, J., & Pino, L. E. (2024). CAMINANDO A HOMBROS DE GIGANTES: INTERSECCIÓN ENTRE LA GENÓMICA Y LA IA. Medicina, 56(3), 1653-2148. Disponible en: https://revistamedicina.net/index.php/Medicina/article/download/1653/2148?inlin

Caravagna, G., Giarratano, Y., Ramazzotti, D., et al. (2018). Detecting repeated cancer evolution from multi-region tumor sequencing data. Nature Methods, 15(9), 707-714.

Carini, C., & Seyhan, A. A. (2024). Tribulations and future opportunities for artificial intelligence in precision medicine. Journal of Translational Medicine, 22, 411. https://doi.org/10.1186/s12967-024-05067-0

Clark, M. M., Hildreth, A., Batalov, S., et al. (2019). Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. Science Translational Medicine, 11(489).

Dias, R., & Torkamani, A. (2019). Artificial intelligence in clinical and genomic diagnostics. Genome Medicine, 11(70). https://doi.org/10.1186/s13073-019-0689-8

Dua, D., & Singh, H. (2023). A Comparative Analysis of Linear Regression, Neural Networks and Random Forests Regression. Nature. Disponible en: https://www.nature.com/articles/s41598-023-49899-0

Erikson, G. A., Bodian, D. L., Rueda, M., et al. (2016). Whole-genome sequencing of a healthy aging cohort. Cell, 165, 1002-1016.

FDA approves stroke-detecting AI software. (2018). Nature Biotechnology, 36, 290.

Grand View Research. (2023). AI in Genomics Market Size, Share & Trends Analysis Report. Disponible en: https://www.grandviewresearch.com/industry-analysis/ai-genomics-market-report

Gurovich, Y., Hanani, Y., Bar, O., et al. (2019). Identifying facial phenotypes of genetic disorders using deep learning. Nature Medicine, 25, 60-74.

Idhaya, T., Suruliandi, A., & Raja, S. P. (2024). A Comprehensive Review on Machine Learning Techniques for Protein Family Prediction. The Protein Journal, 43, 171–186. https://doi.org/10.1007/s10930-024-10181-5

Keras. (n.d.). Keras Documentation. Disponible en: https://keras.io/guides/

Kim, H. K., Min, S., Song, M., et al. (2018). Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. Nature Biotechnology, 36(3), 239-241.

Kircher, M., Witten, D. M., Jain, P., et al. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics, 46(3), 310-315.

König, H., Frank, D., Baumann, M., & Heil, R. (2021). AI models and the future of genomic research and medicine: True sons of knowledge? Artificial intelligence needs to be integrated with causal conceptions in biomedicine to harness its societal benefits for the field. BioEssays, 43(8), 2100025. https://doi.org/10.1002/bies.202100025

Kordos, M., Blachnik, M., & Wieczorek, T. (2023). Combining the Advantages of Neural Networks and Decision Trees for Regression Problems in a Steel Temperature Prediction System. Disponible en: https://link.springer.com/article/10.1007/s00500-023-07318-1

Krishna, R., Elisseev, V., & Antao, S. (2019). BaaS - bioinformática como servicio. En Euro-Par 2018: Talleres de Procesamiento Paralelo. Springer, Cham, págs. 601-612.

Krishna, R., & Elisseev, V. (2020). User-centric genomics infrastructure: Trends and technologies. Genome. https://doi.org/10.1139/gen-2020-0096

Krampis, K., Booth, T., Chapman, B., et al. (2012). Cloud BioLinux: Computación bioinformática preconfigurada y bajo demanda para la comunidad genómica. BMC Bioinformática, 13, 42.

Langelier, C., Kalantar, K. L., Moazed, F., et al. (2018). Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. Proceedings of the National Academy of Sciences, 115(52), E12353-E12362.

Let's Code AI. (2024). AI revolutionizes DNA analysis: From sequences to solutions. Medium. Disponible en: https://medium.com/@letscodeai/ai-revolutionizes-dna-analysis-from-sequences-to-solutions-discover-how-artificial-a4a3717c9488

Li, H., Handsaker, B., Wysoker, A., et al. (2009). El formato de Alineación/Mapa de secuencia y SAMtools. Bioinformática, 25(16), 2078-2079.

Madhukar, N. S., & Elemento, O. (2018). Bioinformatics Approaches to Predict Drug Responses from Genomic Sequencing. Methods in Molecular Biology, 1711, 277-296.

Maqsood, K., Hagras, H., & Zabet, N. R. (2024). An overview of artificial intelligence in the field of genomics. Discover Artificial Intelligence, 4(9). https://doi.org/10.1007/s44163-024-00103-w

McKenna, A., Hanna, M., Banks, E., et al. (2010). El kit de herramientas de análisis del genoma: un marco de mapreduce para analizar datos de secuenciación de ADN de próxima generación. Genome Research, 20(9), 1297-1303.

Muzzey, D., Evans, E. A., & Lieber, C. (2015). Understanding the Basics of NGS: From Mechanism to Variant Calling. Current Genetics Medicine Reports, 3(4), 158-165.

Navarro, F. C. P., Mohsen, H., Yan, C., et al. (2019). Genómica y ciencia de datos: una aplicación dentro de un paraguas. Genome Biology, 20(1), 109.

National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Board on Health Care Services, National Cancer Policy Forum, Board on Health Sciences Policy, & Roundtable on Genomics and Precision Health. (2023). Realizing the Potential of Genomics across the Continuum of Precision Health Care: Proceedings of a Workshop (S. Beachy, M. Hackmann, & K. Asalone, Eds.). National Academies Press (US). Disponible en: https://www.ncbi.nlm.nih.gov/books/NBK592662/

Pastor, Ó., León, A. P., Reyes, J. F. R., García, A. S., & Casamayor, J. C. R. (2021). Using conceptual modeling to improve genome data management. Briefings in Bioinformatics, 22(1), 45-54. https://doi.org/10.1093/bib/bbaa100

Pereira, R., Oliveira, J., & Sousa, M. (2020). Bioinformática y herramientas computacionales para el análisis de secuenciación de nueva generación en genética clínica. Journal of Clinical Medicine, 1(9), 132.

Powers, D. M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Disponible en: https://doi.org/10.48550/arXiv.2010.16061

PyTorch. (n.d.). PyTorch Documentation. Disponible en: https://pytorch.org/docs/stable/index.html

Python Software Foundation. (n.d.). Python Documentation. Disponible en: https://docs.python.org/3/

Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016. Diario Oficial de la Unión Europea. Disponible en: https://www.boe.es/doue/2016/119/L00001-00088.pdf

Reglamento de Inteligencia Artificial de la Unión Europea (AI Act). Diario Oficial de la Unión Europea. Disponible en: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065

Sahraeian, S. M. E., Liu, R., Lau, B., et al. (2019). Deep convolutional neural networks for

accurate somatic mutation detection. Nature Communications, 10(1), 1041.

Sankar, P. L., & Parker, L. S. (2017). The precision medicine initiative's all of us research program: an agenda for research on its ethical, legal, and social issues. Genetics in Medicine, 19, 743-750.

Schrider, D. R., & Kern, A. D. (2018). Supervised Machine Learning for Population Genetics: A New Paradigm. Trends in Genetics, 34(4), 301-312.

Stark, Z., Dolman, L., Manolio, T. A., Ozenberger, B., Hill, S. L., Caulfield, M. J., et al. (2019). Integrating genomics into healthcare: A global responsibility. American Journal of Human Genetics, 104(1), 13-20. https://doi.org/10.1016/j.ajhg.2018.11.014

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., et al. (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Medicine, 12, e1001779. https://doi.org/10.1371/journal.pmed.1001779

Telenti, A., Pierce, L. C. T., Biggs, W. H., Di Iulio, J., Wong, E. H. M., Fabani, M. M., et al. (2016). Deep sequencing of 10,000 human genomes. Proceedings of the National Academy of Sciences, 113, 11901–11906.

TensorFlow. (n.d.). TensorFlow Documentation. Disponible en: https://www.tensorflow.org/guide

Tiwari, R. K., & Etienne, M. (2024). Artificial Intelligence and Healthcare: A Journey through History, Present Innovations, and Future Possibilities. Life, 14(5), 557. https://doi.org/10.3390/life14050557

UNESCO. (n.d.). Recommendation on the Ethics of Artificial Intelligence. Disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000381137

Vakili, M., Ghamsari, M., & Rezaei, M. (2020). Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification. arXiv. Disponible en: https://doi.org/10.48550/arXiv.2001.09636

Walton, N. A., Nagarajan, R., Wang, C., Sincan, M., Freimuth, R. R., Everman, D. B., et al. (2023). Enabling the clinical application of artificial intelligence in genomics: A perspective of the AMIA genomics and translational bioinformatics workgroup. Journal of the American Medical Informatics Association.

Way, G. P., & Greene, C. S. (2018). Bayesian deep learning for single-cell analysis. Nature Methods, 15(12), 1009-1010.

Wekesa, J. S., & Kimwele, M. (2023). A review of multi-omics data integration through

deep learning approaches for disease diagnosis, prognosis, and treatment. Frontiers in Genetics, 14, 1199087. https://doi.org/10.3389/fgene.2023.1199087

Wick, R. R., Judd, L. M., & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biology, 20(1), 129.

World Economic Forum. (n.d.). AI Governance: A Holistic Approach to Implement Ethics into AI. Disponible en: https://www.weforum.org/reports/ai-governance-a-holistic-approach-to-implement-ethics-into-ai

World Health Organization. (n.d.). Ethics and Governance of Artificial Intelligence for Health. Disponible en: https://www.who.int/publications/i/item/9789240029200

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next generation sequencing data. Computational and Structural Biotechnology Journal, 16, 15-24.

Yu, K. H., Berry, G. J., Rubin, D. L., et al. (2017). Association of Omics Features with Histopathology Patterns in Lung Adenocarcinoma. Cell Systems, 5(6), 620-627 e3.

Zhang, J. X., Fang, J. Z., Duan, W., et al. (2018). Predicting DNA hybridization kinetics from sequence. Nature Chemistry, 10(1), 91-98.

Zhou, J., Park, C. Y., Theesfeld, C. L., Wong, A. K., Yuan, Y., Scheckel, C., et al. (2019). Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. Nature Genetics, 51, 973-980.

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nature Methods, 12, 931-934.

Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., & Troyanskaya, O. G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nature Genetics, 50, 1171-1179.

Figuras

Figura 1: Ciencias ómicas:
Fuente: Insipirado en Curso Online de Experto Universitario en Oncogenética (Universidad Politécncia de Valencia, Genotipia), https://www.technologynetworks.com/genomics/articles/what-are-the-key-differences-between-dna-and-rna-296719, https://www.eufic.org/es/produccion-de-alimentos/articulo/que-es-el-microbioma-y-por-que-es-importante, https://www.genome.gov/es/about-genomics/fact-sheets/Epigenomica, https://concepto.de/proteinas/, https://www.eufic.org/es/quienes-somos/nuestro-trabajo

Figura 2: Pipeline de estudios genómico
Fuente: Insipirado en Curso Online de Experto Universitario en Oncogenética (Universidad Politécncia de Valencia, Genotipia) y en Manual de Genética Hematológica (App for IOS and Android, ISBN 978-84-608-1494-8)

Figura 3: Esquema conceptual de inteligencia artificial y aprendizaje automático

Figura 4: Operativa de trabajo

Figura 5: Esquema de despliegue en nube hibrida
Fuente: Ritesh Krishna, Vadim Elisseev. User-centric genomics infrastructure: trends and technologies.
https://cdnsciencepub.com/doi/pdf/10.1139/gen-2020-0096

Figura 6: Esquema DAFO de situación de una instalación nacional.
Fuente: Inspirado en el documento de Reflexión estratégica de Genómica en Cantabria en el que Uno de los autores ha participado adpatándolo al marco global que aborda este proyecto

Figura 7: Etapas del análisis genómico y uso de la IA en cada etapa
Fuente: Sobia Raza, Artificial intelligence for genomic medicine. University of Cambridge.
https://www.phgfoundation.org/publications/reports/artificial-intelligence-for-genomic-medicine/

Figura 8: SpliceIA
Fuente: Cardona, A. F., Ruíz-Patiño, A., Jaller, E., Rodríguez, J., & Pino, L. E. (2024). CAMINANDO A HOMBROS DE GIGANTES: INTERSECCIÓN ENTRE LA GENÓMICA Y LA IA. Medicina.
https://revistamedicina.net/index.php/Medicina/article/download/1653/2148?inlin

Figura 9: DeepGestalt
Fuente: Cardona, A. F., Ruíz-Patiño, A., Jaller, E., Rodríguez, J., & Pino, L. E. (2024). CAMINANDO A HOMBROS DE GIGANTES: INTERSECCIÓN ENTRE LA GENÓMICA Y LA IA. Medicina.
https://revistamedicina.net/index.php/Medicina/article/download/1653/2148?inlin

Figura 10: Previsión económica de crecimiento del mercado mundial en soluciones de IA en genómica

Fuente: Artificial Intelligence in Genomics Market Size, Share and Trends 2024 to 2034. https://www.precedenceresearch.com/artificial-intelligence-in-genomics-market

Glosario

Pipeline de análisis	Conjunto de pasos secuenciales que permiten analizar datos genómicos, desde la preparación de muestras hasta la interpretación clínica.
IA (Inteligencia Artificial)	Disciplina que se ocupa de crear programas que ejecutan operaciones comparables a las de la mente humana, como el aprendizaje y la toma de decisiones.
Aprendizaje Automático (Machine Learning)	Subcampo de la IA que utiliza algoritmos para analizar grandes cantidades de datos y aprender de ellos sin ser explícitamente programado para ello.
Big Data	Conjunto masivo de datos que es tan grande que es difícil de procesar utilizando métodos tradicionales.
Bioinformática	Campo interdisciplinario que desarrolla y aplica métodos computacionales y estadísticos para analizar y gestionar datos biológicos.
Proteómica	Estudio del conjunto completo de proteínas expresadas en una célula o tejido.
Transcriptómica	Estudio de los transcriptos de ARN producidos por el genoma bajo condiciones específicas.
Metabolómica	Análisis de los metabolitos presentes en un organismo para entender su estado metabólico.
Metagenómica	Estudio de los genomas de comunidades microbianas presentes en un entorno específico.
Epigenómica	Estudio de las modificaciones en la expresión génica que no implican cambios en la secuencia del ADN.
Deep Learning	Técnica de aprendizaje profundo que se basa en redes neuronales con múltiples capas.
Redes Neuronales	Modelos computacionales inspirados en las conexiones neuronales del cerebro para el procesamiento de datos complejos.
Redes Neuronales Convolucionales (CNN)	Tipo de red neuronal especialmente efectiva para la clasificación de imágenes y el reconocimiento de patrones.
Redes Neuronales Recurrentes (RNN)	Modelo neuronal diseñado para procesar datos secuenciales como texto o series temporales.
Datos Ómicos	Conjunto de datos que incluye información de diferentes disciplinas ómicas (genómica, transcriptómica, etc.).
Algoritmos de Regresión	Algoritmos que modelan relaciones entre variables dependientes e independientes.
Regresión Lineal	Algoritmo que modela la relación entre una variable

	dependiente y una o más variables independientes.
Regresión Logística	Algoritmo utilizado para clasificación binaria que predice la probabilidad de un evento.
Árboles de Decisión	Modelo de toma de decisiones que segmenta los datos mediante reglas lógicas.
Random Forest	Conjunto de árboles de decisión que mejora la precisión mediante la agregación de múltiples predicciones.
Clustering (Agrupación)	Técnica de aprendizaje no supervisado que agrupa datos en clústeres basados en similitud.
K-means	Algoritmo de agrupación que minimiza la distancia entre puntos de datos y el centro de su clúster.
Validación Cruzada	Método de validación de modelos de IA en el que los datos se dividen en múltiples subconjuntos.
Regularización	Técnica que reduce la complejidad de un modelo para evitar el sobreajuste.
Overfitting (Sobreajuste)	Ocurre cuando un modelo es demasiado complejo y se adapta excesivamente a los datos de entrenamiento.
Algoritmo de Gradiente	Método que optimiza el rendimiento del modelo ajustando gradualmente sus parámetros.
Matriz de Confusión	Tabla que muestra el rendimiento de un algoritmo de clasificación con resultados verdaderos y predichos.
Precisión	Proporción de predicciones correctas sobre el total de predicciones positivas.
Exactitud	Proporción de predicciones correctas sobre el total de todas las predicciones.
Sensibilidad	Capacidad de un modelo para identificar correctamente instancias positivas.
Especificidad	Capacidad de un modelo para identificar correctamente instancias negativas.
F1-Score	Promedio armónico de la precisión y la sensibilidad en un modelo de clasificación.
AUC-ROC	Área bajo la curva ROC que mide la capacidad de un modelo para distinguir entre clases.
Interoperabilidad	Capacidad de sistemas y tecnologías para trabajar juntos e intercambiar datos de forma eficiente.
GDPR (Reglamento General de Protección de Datos)	Regulación de la UE sobre la protección de los datos personales de los ciudadanos.
AI Act	Ley de la UE que regula el desarrollo y uso ético de la inteligencia artificial.
Datos Multiómicos	Integración de datos de múltiples disciplinas ómicas para obtener una visión global del sistema biológico.
HPC (High Performance Computing)	Computación de alto rendimiento utilizada para procesar

	grandes volúmenes de datos genómicos.
TensorFlow	Biblioteca de código abierto para implementar modelos de inteligencia artificial, especialmente redes neuronales.
Keras	Interfaz de alto nivel para la construcción y entrenamiento de modelos de redes neuronales profundas.