Reportes Científicos de la FACEN

Reportes científicos de la FACEN, Enero-junio 2020, 11(1): 39-50
http://doi.org/10.18004/rcfacen.2020.11.1.39

Artículo de Revisión

Aplicaciones de computación en la nube para la ciencia biomédica

Cloud computing applications for biomedical science

Aleli Silva1, Graciela Riera[1] & Danilo Fernández Ríos[1,*]

[1] Universidad Nacional de Asunción - Facultad de Ciencias Exactas y Naturales - Departamento de Biotecnologia. San Lorenzo, Paraguay.

*Email: dfernandez@facen. una.py

Resumen: La investigación biomédica se ha convertido en un esfuerzo intensivo basado en una infraestructura de computación, almacenamiento y de red segura y escalable, que tradicionalmente se ha adquirido, apoyado y mantenido de forma local. Para ciertos tipos de aplicaciones biomédicas, el "*Cloud Computing*" o computación en la nube ha surgido como una alternativa a los enfoques de computación tradicionales mantenidos localmente. La computación en la nube ofrece a los usuarios acceso de pago por uso a servicios tales como infraestructura de hardware, plataformas y software para la solución de problemas computacionales biomédicos comunes. Los servicios de computación en la nube ofrecen almacenamiento y análisis seguros bajo demanda y se diferencian de la informática tradicional de alto rendimiento por su rápida disponibilidad y la escalabilidad de sus servicios. Como tales, los servicios en la nube están diseñados para abordar grandes problemas de datos y mejorar la probabilidad de compartir, reproducir y reutilizar datos y análisis. El presente trabajo es una adaptación del artículo "*Cloud computing applications for biomedical science: A perspective*", publicado bajo la licencia de CC0 1.0 Universal, la cual ofrece una lista de herramientas de computación en la nube útiles para académicos, investigadores y estudiantes de ciencias que trabajan con datos biológicos. Aquí, se proporciona una perspectiva introductoria sobre la computación en la nube para ayudar al lector de habla hispana a determinar su valor para su propia investigación.

Palabrasclave: *Análisis de genoma, Gestión de datos biológicos, Genómica del cáncer, Gestión de datos, Minería de datos, Procesamiento de datos, Visualización de datos.*

Abstract: Biomedical research has become an intensive effort, based on a secure and scalable computing, storage and network infrastructure, which has traditionally been acquired, supported and maintained locally. For certain types of biomedical applications, "Cloud Computing" has emerged as an alternative to traditional locally maintained computing approaches. Cloud computing provides users with usage-based access to services such as hardware infrastructure, platforms and software to solve common biomedical computing problems. Cloud computing services provide secure, on-demand storage and analysis and are differentiated from traditional high-performance computing by their rapid availability and service scalability. As such, cloud services are designed to address large data issues and improve the likelihood of sharing, replicating and reusing data and analysis. This paper is adapted from the article *"Cloud computing applications for biomedical science: A perspective"*, published under the CC0 1.0 Universal license, which provides a list of cloud computing tools useful for academics, researchers and science students working with biological data. Here, an introductory perspective on cloud computing is provided to help the Spanish-speaking reader determine its value for their own research.

Keywords: *Genome analysis, Biological data management, Cancer genomics, Data management, Data mining, Data processing, Data visualization.*

Introducción

El progreso en la investigación biomédica se encuentra cada vez más impulsado por la información obtenida a través del análisis y la interpretación de conjuntos de datos grandes y complejos. A medida que la capacidad de generar y probar hipótesis a través tecnologías de alto rendimiento se ha vuelto técnicamente más factible e incluso más común, el desafío de obtener conocimientos útiles ha pasado de la mesada de laboratorio a la inclusión de la informática.

El presente trabajo es una adaptación del artículo *"Cloud computing applications for biomedical science: A perspective"(Navale & Bourne, 2018)*, publicado bajo la licencia de CC0 1.0 Universal (Creative Commons, 2020), el cual ofrece una lista

Recibido: 11/02/2020 Aceptado: 28/05/2020

ISSN-L: 2078-399X **ISSN:** 2222-145X

de herramientas de computación en la nube útiles para académicos, investigadores y estudiantes de ciencias que trabajan con datos biológicos. Se realizó la presente adaptación con la intención de disponibilizar en habla hispana las principales posibilidades de la computación en la nube, actualmente accesible por la amplia adopción y el aumento de las capacidades de Internet, e impulsada por las necesidades del mercado. Estas herramientas han surgido como un enfoque poderoso, flexible y escalable para resolver problemas computacionales y de datos en las más diversas áreas de la investigación en biociencias.

El Instituto Nacional de Estándares y Tecnología (NIST, *National Institute of Standards and Technology* en inglés) clasifica las nubes en cuatro tipos: **públicas, privadas, comunitarias e híbridas**.

En una nube pública, la infraestructura existe en las instalaciones del proveedor de computación en la nube y es gestionada por éste, mientras que en una nube privada, la infraestructura puede existir dentro o fuera de las instalaciones del proveedor de computación en la nube, pero es gestionada por la organización privada. Ejemplos de nubes públicas incluyen Amazon Web Services (AWS), Google Cloud Platform (GCP) y Microsoft Azure.

Una comunidad en la nube es un esfuerzo de colaboración donde la infraestructura es compartida entre varias organizaciones —una comunidad específica— que tienen requisitos comunes de seguridad y cumplimiento de normas. La nube JetStream (Indiana University Pervasive Technology Institute, 2019) sirve como una nube comunitaria al servicio de la comunidad científica. Una nube híbrida es una composición de dos o más infraestructuras de nube distintas —privadas, comunitarias, públicas— que siguen siendo entidades únicas, pero que están unidas entre sí de forma que permiten la portabilidad de datos y aplicaciones de software (Mell & Grance, 2011).

Los tipos de nube mencionados anteriormente pueden utilizar uno o más servicios de nube: software como servicio (SaaS), plataforma como servicio (PaaS) e infraestructura como servicio (IaaS). SaaS permite al consumidor utilizar las aplicaciones del proveedor de computación en la nube (por ejemplo, Google Docs) que se ejecutan en la infraestructura de un proveedor de computación en la nube, mientras que PaaS permite a los consumidores crear o adquirir aplicaciones y herramientas e implementarlas en la infraestructura del proveedor de computación en la nube. IaaS permite al consumidor proporcionar procesamiento, almacenamiento, redes y otros recursos informáticos fundamentales.

Adopción de la nube para el trabajo biomédico

Considere ejemplos de cómo se han implementado las nubes y los servicios en la nube en biomedicina (Tabla 1) Sólo en genómica, el uso abarca desde aplicaciones individuales hasta máquinas virtuales completas con múltiples aplicaciones.

Herramientas individuales

BLAST (Camacho et al., 2009) es una de las herramientas más utilizadas en la investigación bioinformática. Una imagen del servidor BLAST se puede alojar en nubes públicas AWS, Azure y GCP para permitir quelos usuarios realicen búsquedas independientes con BLAST. Los usuarios también pueden enviar búsquedas utilizando BLAST a través de la interfaz de programación de aplicaciones o (API, *application programming interface* en inglés) del Centro Nacional de Información Biotecnológica (NCBI, *National Center for Biotechnology Information*) de los Estados Unidos para que se ejecuten en AWS y Google Compute Engine (NCBI, 2019b).

Además, la plataforma Microsoft Azure se puede aprovechar para ejecutar grandes tareas de correspondencia de secuencias BLAST dentro de límites de tiempo razonables. Azure permite a los usuarios descargar bases de datos de secuencias de NCBI, ejecutar diferentes programas BLAST en una entrada específica contra las bases de datos de secuencias, y generar visualizaciones de los resultados para facilitar el análisis. Azure también proporciona una manera de crear una interfaz de usuario basada en web, para programar y rastrear las tareas de BLAST, visualizar resultados, administrar usuarios y realizar tareas básicas (NCBI, 2019a).

CloudAligner es una herramienta rápida y completa, basada en MapReduce para mapeo de secuencias, diseñada para poder manejar secuencias largas (Nguyen, Shi, & Ruden, 2011), mientras que CloudBurst (Schatz, 2009) puede proporcionar un mapeo de lectura corta altamente

Tabla 1 (comienzo). Ejemplos de tipos de nube, modelos de servicio, flujos de trabajo y plataformas para aplicaciones biomédicas.

Uso biomédico	Tipo de nube	Modelos de servicios en la nube	Ejemplos de proveedores de nube	Notas adicionales
Herramientas individuales				
Alineación de secuencias	Nube pública	IaaS	AWS, Azure, Google	BLAST
Mapeo de secuencias largas	Nube pública	IaaS	AWS	CloudAligner, Elastic Map Reduce
Mapeo de secuencias cortas	Nube pública	IaaS	AWS	CloudBurst
Análisis secuencial de alto rendimiento	Nube pública	IaaS	AWS	Paquete Eoulsan, Elastic Map Reduce
Alineación de secuencias y genotipado	Nube pública	IaaS	AWS	Crossbow, Elastic Map Reduce
Asistencia médica				
Monitoreo de ECG en tiempo real	Nube híbrida	IaaS	AWS EC2	Uso combinado de recursos in situ con la nube pública
Servicios de telemedicina ECG de 12 derivaciones	Nube pública	PaaS	Microsoft Azure	Implementación de aplicaciones de ECG seguras, servicios de visualización y gestión de datos con base de datos en la nube
Almacenamiento y recuperación de imágenes de diagnóstico	Nube pública	PaaS	AWS, Microsoft Azure, Google Apps Engine	Hospedaje de los módulos centrales del Picture Archive Communication System para configurar repositorios de datos médicos
Herramientas de uso general				
Análisis de secuencia microbiana automatizado	Nube pública	IaaS	AWS EC2	cloVR
Computación bioinformática de alto rendimiento	Nube pública	IaaS	AWS	Cloud Biolinux
Big Data biomédica	Nube pública	PaaS	AWS, Azure, Google, IBM	Hadoop, MapReduces, BioQuery, Redshift

Abreviaciones: NGS) Secuenciación de nueva generación, **AWS)** Servicios web de Amazon, **EC2)** Nube de cálculo elástica, **S3)** Servicio simple de almacenamiento, **TCGA)** Atlas del genoma del cáncer, **GCP)** Plataformas de nube de google, **IaaS)** Infraestructura como servicio, **PaaS)** Plataforma como servicio, **SaaS)** Software como servicio.

Tabla 1 (final).

Uso biomédico	Tipo de nube	Modelos de servicios en la nube	Ejemplos de proveedores de nube	Notas adicionales
Flujos de trabajos y plataformas				
NGS y análisis de datos	Nube pública	IaaS	AWS	Galaxy, aplicaciones de código abierto
Análisis NGS	Nube privada	PaaS	Nube de datos protegidas Bionimbus	OpenStack, software para construir plataformas de nube
NGS para el trabajo de diagnóstico clínico	Nube pública	PaaS	AWS CloudMan	Cloud Biolinux, Cloud Biocentral
Patrón de mutación en miles de secuencias del genoma completo	Nube híbrida	IaaS	AWS ECS S3	Recursos universitarios combinados con la nube pública
Análisis de datos a gran escala (TCGA)	Nube pública	PaaS	Google Elastic Compute	Broad Institute FireCloud
Análisis de datos a gran escala (TCGA)	Nube pública	PaaS	GCP	Instituto de Biología de Sistemas
Análisis de datos a gran escala (TCGA)	Nube pública	PaaS, SaaS	AWS	Nube de Seven Bridges Genomics (SBG) interconectado con AWS y GCP
Análisis de datos genómicos	Nube pública	PaaS	AWS	Motor de conocimiento, Minería de datos y machine learning
Secuenciación a gran escala, análisis de datos e integración de datos fenotípicos y clínicos	Nube pública	PaaS, SaaS	AWS, Microsoft Azure	DNAnexus Herramientas informáticas Deep variant
Aplicaciones de flujos de trabajo para genómica	Nube pública	PaaS, SaaS	Plataforma Google cloud	DNAStack Pipelines de datos altamente curados

Abreviaciones: NGS) Secuenciación de nueva generación, **AWS)** Servicios web de Amazon, **EC2)** Nube de cálculo elástica, **S3)** Servicio simple de almacenamiento, **TCGA)** Atlas del genoma del cáncer, **GCP)** Plataformas de nube de google, **IaaS)** Infraestructura como servicio, **PaaS)** Plataforma como servicio, **SaaS)** Software como servicio.

sensible con MapReduce. El paquete Eoulsan integrado en un entorno IaaaS de nube permite realizar análisis de secuencias de alto rendimiento (Jourdren, Bernard, Dillies, & Le Crom, 2012). Para los análisis de resecuenciación de genoma completo, Crossbow(Langmead, Schatz, Lin, Pop, & Salzberg, 2009) es un *pipeline* de software escalable. Crossbow combina Bowtie, un alineador de lectura corta ultrarrápido y eficiente en memoria, y SoapSNP, un genotipador, en *pipeline* paralelo automático que puede correr en la nube.

Flujos de trabajo y plataformas

La integración de genotipo, fenotipo y datos clínicos es muy importante para la investigación biomédica. Las plataformas biomédicas pueden proporcionar un entorno para establecer un conducto de extremo a extremo para la adquisición, el almacenamiento y el análisis de datos.

Galaxy, una plataforma de código abierto basada en la web, se utiliza para la investigación biomédica con una gran cantidad de datos (Afgan et al., 2016). Para el análisis de datos a gran escala, Galaxy puede alojarse en la nube IaaS (Taylor, 2017). Se han logrado sistemas de flujo de trabajo fiables y altamente escalables basados en la nube para análisis de secuencias de próxima generación mediante la integración del sistema de flujo de trabajo Galaxy con GlobusProvision (Liu et al., 2014).

La Nube de Datos Protegida Bionimbus (BPDC, *Bionimbus Protected Data Cloud* en inglés) es una infraestructura privada basada en la nube para gestionar, analizar y compartir grandes cantidades de datos genómicos y fenotípicos en un entorno seguro, que se ha utilizado para estudios de fusión de genes (Heath et al., 2014). BPDC se basa principalmente en OpenStack, un software de código abierto que proporciona herramientas para construir plataformas en la nube (OpenStack, 2017), con un portal de servicio para un único punto de entrada y un único inicio de sesión para varios recursos disponibles de BPDC. Utilizando BPDC, el análisis de datos para el proyecto de resecuenciación de la leucemia mieloide aguda (LMA) se realizó rápidamente para identificar variantes somáticas expresadas en muestras primarias de alto riesgo de LMA (McNerney et al., 2013).

Se necesita una infraestructura escalable y robusta para los análisis de secuenciación de nueva generación o (NGS, *Next Generation Sequencing* en inglés) diagnóstico en los laboratorios clínicos. CloudMan está disponible en la infraestructura de nube de AWS (Afgan et al., 2010). Se ha utilizado como plataforma para la distribución de herramientas, datos y análisis de resultados. Las mejoras en el uso de CloudMan para el análisis de variantes genéticas se han realizado mediante la reducción de los costes de almacenamiento para análisis clínicos (Onsongo et al., 2014).

Como parte del *Pan Cancer Analysis of Whole Genomes* (PCAWG), se estudiaron patrones comunes de mutación en más de 2800 secuencias del genoma completo del cáncer, lo que requirió importantes recursos de computación científica para investigar el papel de las regiones no codificantes del genoma del cáncer y para comparar los genomas de células tumorales y normales (ICGC, 2017b). El centro de coordinación de datos del PCAWG cuenta actualmente con acuerdos de colaboración con el proveedor de nube AWS y el Cancer Collaboratory (ICGC, 2017a), un recurso académico de computación en nube mantenido por el Ontario Institute for Cancer Research y alojado en las instalaciones de Compute Canada.

Se utilizaron múltiples recursos académicos para completar el análisis de 1827 muestras en un período de 6 meses. Esto se complementó con el uso de recursos de la nube, donde 500 muestras fueron analizadas por AWS en 6 semanas (Stein, Knoppers, Campbell, Getz, & Korbel, 2015); lo cual demostró que los recursos públicos de la nube pueden ser utilizados para escalar de forma rápida un proyecto si se necesitan mayores recursos de computación.

En este caso, se utilizó el almacenamiento de datos AWS S3 para escalar de 600 terabytes a múltiples PBs. Las lecturas brutas, las alineaciones del genoma, los metadatos y los datos curados también se pueden cargar de forma incremental en AWS S3 para que la comunidad de investigadores del cáncer pueda acceder rápidamente a ellos. Las herramientas de búsqueda de datos y acceso también están disponibles para que otros investigadores las utilicen o reutilicen.

El Instituto Nacional del Cáncer (NCI, *National Cancer Institute* en inglés), ha financiado tres nubes piloto para proporcionar análisis genómicos, apoyo computacional y capacidades de acceso a los datos del Atlas del Genoma del Cáncer (TCGA, *The Cancer Genome Atlas* en inglés) (NCI, 2017). El objetivo de los proyectos piloto fue desarrollar una plataforma escalable para facilitar la colaboración

en la investigación y la reutilización de datos. Las tres nubes piloto han recibido conjuntos de datos de referencia armonizados del genoma del cáncer del Genomic Data Commons (GDC) (Grossman et al., 2016) que han sido analizados con un conjunto común de flujos de trabajo contra un genoma de referencia (por ejemplo, GRCh38).

El proyecto piloto del Broad Institute desarrolló FireCloud (Broad Institute, 2017a), utilizando la capacidad de cálculo elástica de Google Cloud para el análisis, la conservación, el almacenamiento y el uso compartido de datos a gran escala. Los usuarios también pueden cargar sus propios métodos de análisis y datos a espacios de trabajo y/o utilizar las herramientas del Broad Institute. FireCloud utiliza el Workflow Description Language (WDL) para permitir a los usuarios la ejecución de flujos de trabajo escalables y reproducibles (Broad Institute, 2017b).

El programa piloto del Instituto para la Biología de Sistemas o (ISB, *Institute for Systems Biology* en inglés) aprovecha varios servicios en la plataforma GCP. Los investigadores pueden utilizar aplicaciones de software basadas en la web para definir y comparar cohortes de forma interactiva, examinar los datos moleculares subyacentes para genes específicos o vías de interés, compartir puntos de vista con colaboradores y aplicar sus programas y scripts de software individuales a varios conjuntos de datos (ISB, 2017).

El ISB Cancer Genome Cloud (CGC) ha cargado datos procesados y metadatos del proyecto TCGA en el servicio de base de datos gestionado de BigQuery, lo que permite una fácil extracción de datos y enfoques de almacenamiento de datos que se pueden utilizar en datos genómicos a gran escala. El Seven Bridges Genomics (SBG) CGC ofrece tanto genómica SaaS como PaaS y utiliza AWS (SevenBridges Genomics, 2017). La plataforma también permite que los investigadores colaboren en el análisis de grandes conjuntos de datos de genómica del cáncer de forma segura, reproducible y escalable.

Las plataformas comerciales (AWS, Microsoft Azure) basadas en la nube (por ejemplo, DNAnexus) permiten el análisis de cantidades masivas de datos de secuenciación integrados con información fenotípica o clínica (Anderson, 2017). Otras plataformas bioinformáticas (por ejemplo, DNAstack) utilizan la GCP para proporcionar capacidad de procesamiento para más de un cuarto de millón de secuencias completas del genoma humano al año (DNAstack, 2017).

Asistencia de salud

Las aplicaciones de computación en la nube en la atención sanitaria incluyen la telemedicina/teleconsulta, las imágenes médicas, la salud pública, la autogestión del paciente, la gestión hospitalaria y los sistemas de información, la terapia y el uso secundario de los datos.

Un ejemplo conmovedor son los pacientes que sufren de arritmias cardíacas y requieren detección y monitoreo continuo de episodios. Los sensores portátiles se pueden utilizar para monitoreo de electrocardiograma (ECG) en tiempo real, detección de episodios de arritmia y clasificación. Utilizando AWS EC2, se integraron las tecnologías de computación móvil y se demostraron las capacidades de monitoreo de ECG para registrar, analizar y mostrar visualmente los datos de los pacientes en lugares remotos. Además, las herramientas de software que han monitorizado y analizado los datos de ECG se han puesto a disposición del público a través de la nube SaaS (Pandey, Voorsluys, Niu, Khandoker, & Buyya, 2012).

Herramientas de uso general

CloVR es una máquina virtual que emula un sistema informático, con bibliotecasy paquetes preinstalados para el análisis de datos biológicos (Angiuoli et al., 2011). De forma similar, Cloud BioLinux es un recurso público disponible con imágenes de máquinas virtuales y proporciona más de 100 paquetes de software para computación bioinformática de alto rendimiento (Krampis et al., 2012). Ambas imágenes de máquina virtual (CloVR y BioLinux) están disponibles para su uso en un entorno de IaaS en la nube.

La adopción de la nube también puede incluir

servicios gestionados diseñados para problemas generales de Big Data. Por ejemplo, cada uno de los principales proveedores de nube pública ofrece un conjunto de servicios para aprendizaje de máquina e inteligencia artificial, algunos de los cuales se encuentran pre-entrenados para resolver problemas comunes (por ejemplo, texto a voz).

Los sistemas de bases de datos como Google BigQuery (Google, 2012) y Amazon Redshift (Amazon, 2016) combinan la naturaleza escalable y elástica de la nube con soluciones de software y hardware ajustadas para ofrecer capacidades y rendimiento de bases de datos que de otro modo no serían fáciles de conseguir. Para conjuntos de datos biomédicos grandes y complejos, estas bases de datos pueden reducir los costes de gestión, facilitar la adopción de la base de datos y facilitar el análisis. Varias de las grandes aplicaciones de datos utilizadas en la investigación biomédica, como la biblioteca de software Apache Hadoop, están basadas en la nube (Luo, Wu, Gopukumar, & Zhao, 2016).

Desarrollo de un ecosistema digital basado en la nube para la investigación biomédica

Los ejemplos presentados más arriba, algunos todavía en proceso desde hace varios años, ilustran una desviación del enfoque tradicional a la computación biomédica. El enfoque tradicional ha sido descargar datos a sistemas informáticos locales desde sitios públicos y luego realizar el procesamiento, análisis y visualización de datos localmente. El tiempo de descarga, el costo y la redundancia necesarios para mejorar las capacidades informáticas locales a fin de satisfacer las necesidades de investigación biomédica de gran cantidad de datos (por ejemplo, en secuenciación e imágenes) hacen que este enfoque merezca reevaluación.

Los proyectos a gran escala, como el PCAWG presentado anteriormente, han demostrado la ventaja de utilizar recursos, tanto locales como públicos, de varias instituciones colaboradoras. Para las instituciones con infraestructura local establecida (por ejemplo, infraestructura de red de alta velocidad, repositorios de datos seguros), desarrollar un ecosistema digital basado en la nube con opciones para aprovechar cualquiera de los tipos de nube (pública, híbrida) puede ser ventajoso. Además, el desarrollo y la utilización de un ecosistema basado en la nube aumentan la probabilidad de una ciencia abierta.

Para promover el descubrimiento del conocimiento y la innovación, los datos y análisis abiertos deben ser localizables, accesibles, interoperables y reutilizables (*Findable, Accessible, Interoperable, and Reusable* o FAIR, por sus siglas en inglés). Los principios FAIR sirven de guía a los productores, administradores y difusores de datos para mejorar su reutilización, incluyendo algoritmos, herramientas y flujos de trabajo que son esenciales para una buena gestión del ciclo de vida de los datos (Wilkinson et al., 2016). Un ecosistema de datos biomédicos debe tener capacidad para indexar datos, metadatos, software y otros objetos digitales, un sello de la iniciativa Big Data to Knowledge (BD2K) del NIH (Jagodnik et al., 2017).

La adopción de los principios FAIR se ve facilitada por un paradigma emergente para la ejecución de conjuntos de herramientas de software complejos e interrelacionados, como los que se utilizan en el procesamiento de datos genómicos, e implican empaquetar software utilizando tecnologías de contenedores Linux, como Docker, y luego orquestar *pipelines* utilizando lenguajes de flujo de trabajo específicos del dominio, como WDL y Common Workflow Language (Amstutz et al., 2016). Los proveedores de computación en la nube también ofrecen funciones de procesamiento por lotes (por ejemplo, AWS Batch) que proporcionan automáticamente la cantidad óptima y el tipo de recursos de cálculo basado en el volumen y los requisitos específicos de recursos de los lotes de trabajo enviados, lo que facilita considerablemente el análisis a escala.

En la Figura 1 se ilustra la integración de productores, consumidores y repositorios de datos a través de una plataforma basada en la nube para apoyar los principios FAIR.

Es importante un programa de entrenamiento de forma regular para los usuarios de datos de la nube, especialmente para el manejo de datos sensibles (por ejemplo, información de identificación personal). El

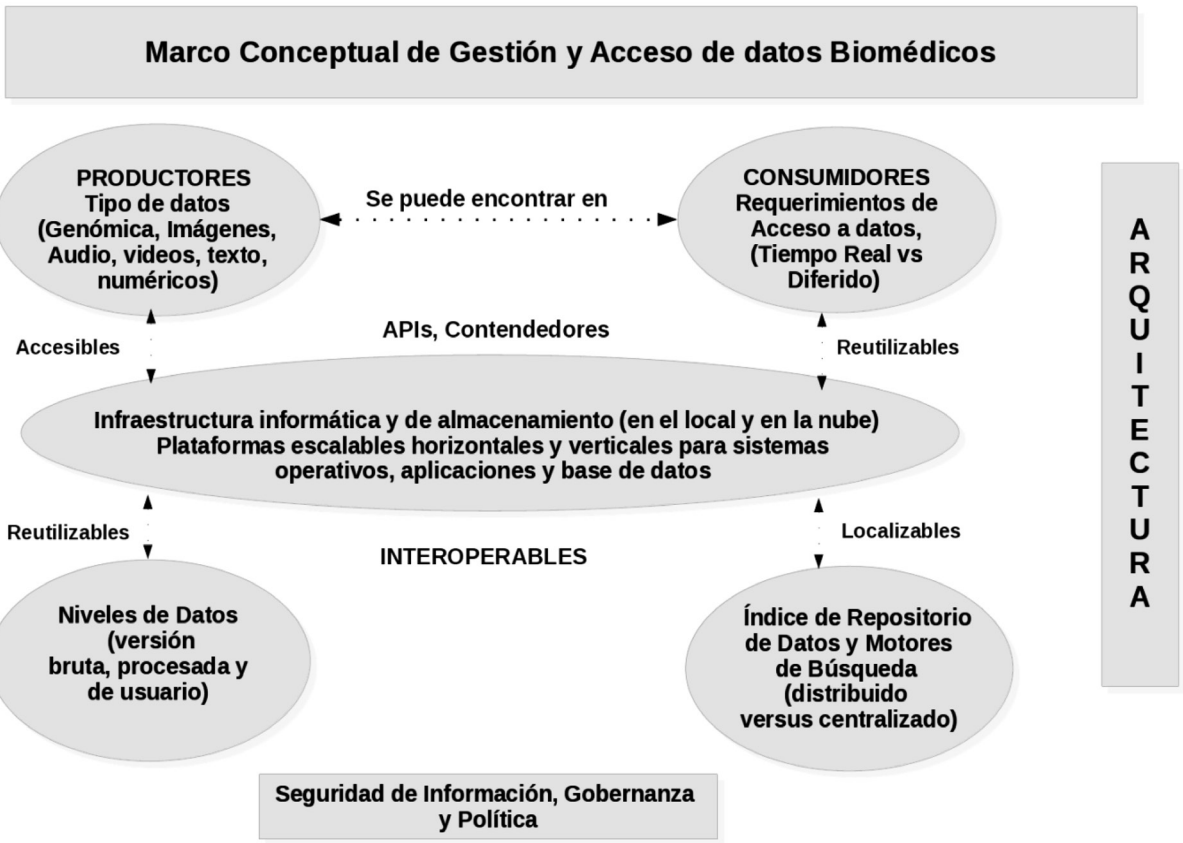

Figura 1. Plataforma conceptual basada en la nube con diferentes tipos de datos que fluyen entre productores y consumidores que requieren niveles de datos variables.

entrenamiento debe incluir métodos para proteger los datos que se mueven a la nube y controlar el acceso a los recursos de la nube, incluidas las máquinas virtuales, los contenedores y los servicios en la nube que intervienen en la gestión del ciclo de vida de los datos. Proteger las claves de acceso, utilizar autenticación multifactorial, crear listas de usuarios para la gestión de identidades y accesos con permisos controlados siguiendo el principio del mínimo privilegio —configurado para realizar las acciones necesarias para los usuarios—son algunas de las prácticas recomendadas que pueden minimizar las vulnerabilidades de seguridad que pueden surgir de usuarios inexpertos de la nube y/o de entidades externas maliciosas (Foster & Gannon, 2017).

La búsqueda de datos biomédicos abiertos ha motivado el interés en mejorar el acceso a los datos y, al mismo tiempo, mantener la seguridad y la privacidad. Por ejemplo, un concurso abierto a escala comunitaria para el desarrollo de nuevos métodos de protección de datos genómicos ha mostrado la viabilidad de la externalización segura de datos y la colaboración para el análisis de datos genómicos basados en la nube (Tang et al., 2016). Los resultados del trabajo demuestran que las técnicas criptográficas pueden apoyar el análisis comparativo público basado en la nube de los genomas humanos. Trabajos recientes han mostrado que utilizando un modelo híbrido de implementación de la nube, el 50%-70% de la tarea de mapeo de lectura puede llevarse a cabo de forma precisa y eficiente en una nube pública (Popic & Batzoglou, 2017).

En resumen, un ecosistema basado en la nube requiere capacidad para la interoperabilidad entre nubes, el desarrollo de herramientas que puedan operar en múltiples entornos de nube y que puedan

abordar los desafíos de la protección de datos, la privacidad y las restricciones legales impuestas por diferentes países (véase Molnár-Gábor, Lueck, Yakneen, & Korbel, 2017 para un análisis relacionado con los datos genómicos).

Conclusiones

El uso de las nubes, desde el análisis genómico a gran escala, pasando por la monitorización remota de pacientes, hasta el diagnóstico molecular en laboratorios clínicos, tiene ventajas pero también inconvenientes potenciales. Un primer paso es determinar qué tipo de entorno de nube se adapta mejor a la aplicación y, a continuación, si representa una solución rentable. Esta introducción intenta indicar qué se debe considerar, cuáles son las opciones y qué aplicaciones que pueden servir de referencia para tomar la mejor decisión sobre cómo proceder ya están en uso.

Literatura citada

Afgan, E., Baker, D., Coraor, N., Chapman, B., Nekrutenko, A., & Taylor, J. (2010). Galaxy CloudMan: delivering cloud compute clusters. *BMC Bioinformatics*, 11(Suppl 12): S4. https://doi.org/10.1186/1471-2105-11-S12-S4

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Grüning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A. & Goecks, J. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*, 44(W1): W3–W10. https://doi.org/10.1093/nar/gkw343

Amazon. (2016). *Enterprise Data Warehousing on Amazon Web Services*. Retrieved December 30, 2019, from: https://d0.awsstatic.com/whitepapers/enterprise-data-warehousing-on-aws.pdf

Amstutz, P., Crusoe, M.R., Tijanić, N., Chapman, B., Chilton, J., Heuer, M., Kartashov, A., Leehr, D., Ménager, H., Nedeljkovich, M., Scales, M., Soiland-Reyes, S. & Stojanovic, L. (2016). Common Workflow Language Specifications, v1.0. *Figshare*. https://doi.org/10.6084/m9.figshare.3115156.v2

Anderson, C. (2017). AZ Partners with DNAnexus for 2 Million Patient Sequencing Project. *Clinical OMICs*, 4(4): 32–32. https://doi.org/10.1089/clinomi.04.04.23

Angiuoli, S.V., Matalka, M., Gussman, A., Galens, K., Vangala, M., Riley, D.R., Arze, C., White, J.R., White, O. & Fricke, W.F. (2011). CloVR: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics*, *12*(1), 356. https://doi.org/10.1186/1471-2105-12-356

Broad Institute. (2017a). *FireCloud*. Retrieved December 31, 2019, from: https://software.broadinstitute.org/firecloud/

Broad Institute. (2017b). *Workflow Description Language*. Retrieved from: https://software.broadinstitute.org/wdl/

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T.L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1): 421. https://doi.org/10.1186/1471-2105-10-421

Creative Commons. (2020). *CC0 1.0 Universal*. Retrieved February 10, 2020, from: https://creativecommons.org/publicdomain/zero/1.0/

DNAstack. (2017). *Genomics made simple*. Retrieved December 30, 2019, from: https://dnastack.com/#/

Foster, I., & Gannon, D.B. (2017). *Cloud Computing for Science and Engineering*. Cambridge, MA: MIT Press.

Fusaro, V.A., Patil, P., Gafni, E., Wall, D.P., & Tonellato, P.J. (2011). Biomedical Cloud Computing With Amazon Web Services. *PLoS Computational Biology*, 7(8): e1002147. https://doi.org/10.1371/journal.pcbi.1002147

GEN. (2017). *DNAnexus Platform Offers Google-Developed DeepVariant*. Retrieved December 30, 2019, from: https://www.genengnews.com/topics/omics/dnanexus-platform-offers-google-developed-deepvariant/

Google. (2012). *An Inside Look at Google BigQuery*. Retrieved December 30, 2019, from: https://cloud.google.com/files/BigQueryTechnicalWP.pdf

Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., & Staudt, L.M. (2016). Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine*, 375(12): 1109–1112. https://doi.org/10.1056/NEJMp1607591

Heath, A.P., Greenway, M., Powell, R., Spring, J., Suarez, R., Hanley, D., Bandlamudi, C., McNerney, M.E., White, K.P. & Grossman, R.L. (2014). Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets. *Journal of the American Medical Informatics Association*, 21(6): 969–975. https://doi.org/10.1136/amiajnl-2013-002155

ICGC. (2017a). *Cancer Collaboratory*. Retrieved December 31, 2019, from: https://dcc.icgc.org/icgc-in-the-cloud/collaboratory

ICGC. (2017b). *PanCancer Analysis Working Group*. Retrieved December 30, 2019, from: https://dcc.icgc.org/pcawg

Indiana University Pervasive Technology Institute. (2019). *Jetstream: A National Science and Engineering Cloud*. Retrieved December 30, 2019, from: https://jetstream-cloud.org/

ISB. (2017). *Institute for Systems Biology: Cancer Genomics Cloud*. Retrieved December 31, 2019, from: https://isb-cgc.appspot.com/

Jagodnik, K.M., Koplev, S., Jenkins, S.L., Ohno-Machado, L., Paten, B., Schurer, S.C., Dumontier, M., Verborgh, R., Bui, A., Ping, P., McKenna, N.J., Madduri, R., Pillai, A. & Ma'ayan, A. (2017). Developing a framework for digital objects in the Big Data to Knowledge (BD2K) commons: Report from the Commons Framework Pilots workshop. *Journal of Biomedical Informatics*, 71: 49–57. https://doi.org/10.1016/j.jbi.2017.05.006

Jourdren, L., Bernard, M., Dillies, M.-A., & Le Crom, S. (2012). Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. *Bioinformatics*, 28(11): 1542–1543. https://doi.org/10.1093/bioinformatics/bts165

Knaus, J., Hieke, S., Binder, H., & Schwarzer, G. (2013). Costs of cloud computing for a biometry department. A case study. *Methods of Information in Medicine*, 52(1): 72–79. https://doi.org/10.3414/ME11-02-0048

Krampis, K., Booth, T., Chapman, B., Tiwari, B., Bicak, M., Field, D., & Nelson, K.E. (2012). Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. *BMC Bioinformatics*, 13(1): 42. https://doi.org/10.1186/1471-2105-13-42

Langmead, B., Schatz, M.C., Lin, J., Pop, M., & Salzberg, S.L. (2009). Searching for SNPs with cloud computing. *Genome Biology*, 10(11): R134. https://doi.org/10.1186/gb-2009-10-11-r134

Liu, B., Madduri, R.K., Sotomayor, B., Chard, K., Lacinski, L., Dave, U.J., Li, J., Liu, C. & Foster, I.T. (2014). Cloud-based bioinformatics workflow platform for large-scale next-generation sequencing analyses. *Journal of Biomedical Informatics*, 49: 119–133. https://doi.org/10.1016/j.jbi.2014.01.005

Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016). Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomedical Informatics Insights*, 8: BII.S31559. https://doi.org/10.4137/BII.S31559

McNerney, M.E., Brown, C.D., Wang, X., Bartom, E.T., Karmakar, S., Bandlamudi, C.,

Yu, S., Ko, J., Sandall, B.P., Stricker, T., Anastasi, J., Grossman, R.L. Cunningham, J.M., Le Beau M.M. & White, K.P. (2013). CUX1 is a haploinsufficient tumor suppressor gene on chromosome 7 frequently inactivated in acute myeloid leukemia. *Blood*, 121(6): 975–983. https://doi.org/10.1182/blood-2012-04-426965

Mell, P., & Grance, T. (2011). *The NIST definition of cloud computing*. Retrieved from: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

Molnár-Gábor, F., Lueck, R., Yakneen, S., & Korbel, J.O. (2017). Computing patient data in the cloud: practical and legal considerations for genetics and genomics research in Europe and internationally. *Genome Medicine*, 9(1): 58. https://doi.org/10.1186/s13073-017-0449-6

Navale, V., & Bourne, P.E. (2018). Cloud computing applications for biomedical science: A perspective. *PLOS Computational Biology*, 14(6): e1006144. https://doi.org/10.1371/journal.pcbi.1006144

NCBI. (2019a). *BLAST on Windows Azure*. Retrieved December 30, 2019, from: https://www.microsoft.com/en-us/download/details.aspx?id=52513

NCBI. (2019b). *Cloud BLAST*. Retrieved December 30, 2019, from: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=CloudBlast

NCI. (2017). *National Cancer Institute Cancer Genomics Cloud Pilots*. Retrieved December 31, 2019, from: https://cbiit.cancer.gov/sites/nci-cbiit/files/Cloud_Pilot_Handout_508compliant.pdf

Nguyen, T., Shi, W., & Ruden, D. (2011). CloudAligner: A fast and full-featured MapReduce based tool for sequence mapping. *BMC Research Notes*, 4(1): 171. https://doi.org/10.1186/1756-0500-4-171

Onsongo, G., Erdmann, J., Spears, M.D., Chilton, J., Beckman, K.B., Hauge, A., Yohe, S., Schomaker, M., Bower, M., Silverstein, K.A.T. & Thyagarajan, B. (2014). Implementation of Cloud based Next Generation Sequencing data analysis in a clinical laboratory. *BMC Research Notes*, 7(1): 314. https://doi.org/10.1186/1756-0500-7-314

OpenStack. (2017). *OpenStack Open Source Cloud Computing Software*. Retrieved December 30, 2019, from: https://www.openstack.org/

Pandey, S., Voorsluys, W., Niu, S., Khandoker, A., & Buyya, R. (2012). An autonomic cloud environment for hosting ECG data analysis services. *Future Generation Computer Systems*, 28(1): 147–154. https://doi.org/10.1016/j.future.2011.04.022

Popic, V., & Batzoglou, S. (2017). A hybrid cloud read aligner based on MinHash and kmer voting that preserves privacy. *Nature Communications*, 8(1): 15311. https://doi.org/10.1038/ncomms15311

Schatz, M.C. (2009). CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics*, 25(11): 1363–1369. https://doi.org/10.1093/bioinformatics/btp236

SevenBridges Genomics. (2017). Cancer Genomics Cloud. Retrieved December 31, 2019, from Cancer Genomics Cloud website: http://www.cancergenomicscloud.org/

Stein, L.D., Knoppers, B.M., Campbell, P., Getz, G., & Korbel, J.O. (2015). Data analysis: Create a cloud commons. *Nature*, 523(7559): 149–151. https://doi.org/10.1038/523149a

Tang, H., Jiang, X., Wang, X., Wang, S., Sofia, H., Fox, D., Lauter, K., Malin, B., Telenti, A., Xiong, L. & Ohno-Machado, L. (2016). Protecting genomic data analytics in the cloud: state of the art and opportunities. *BMC Medical Genomics*, 9(1): 63. https://doi.org/10.1186/s12920-016-0224-3

Taylor, J. (2017). *Galaxy on the Cloud*. Retrieved December 30, 2019, from: https://www.coursera.org/lecture/galaxy-project/galaxy-on-the-cloud-veQKq

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Santos, L.B.S., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao J. & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1): 160018. https://doi.org/10.1038/sdata.2016.18