FACULTAD DE INGENIERÍA DE COMPUTACIÓN Y SISTEMAS
UNIDAD DE POSGRADO

ARQUITECTURA EMPRESARIAL EN UN CENTRO DE INVESTIGACIÓN CIENTÍFICA PARA EL ANÁLISIS BIOINFORMÁTICO DE GRANDES VOLÚMENES DE DATOS BASADO EN EL *CLOUD COMPUTING*

PRESENTADO POR
MIRELLA ROCÍO FLORES GONZÁLEZ

ASESOR
CARLOS ARTURO RAYMUNDO IBÁÑEZ

TESIS
PARA OPTAR EL GRADO ACADÉMICO
DE MAESTRA EN INGENIERÍA DE COMPUTACIÓN Y SISTEMAS CON MENCIÓN EN GESTIÓN DE TECNOLOGÍAS DE INFORMACIÓN

LIMA, PERÚ
2018

USMP
UNIVERSIDAD DE SAN MARTIN DE PORRES

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE COMPUTACIÓN Y SISTEMAS

ARQUITECTURA EMPRESARIAL EN UN CENTRO DE INVESTIGACIÓN CIENTÍFICA PARA EL ANÁLISIS BIOINFORMATICO DE GRANDES VOLÚMENES DE DATOS BASADO EN EL *CLOUD COMPUTING*

TESIS

PARA OPTAR EL GRADO ACADÉMICO DE:

MAESTRA EN INGENIERÍA DE COMPUTACIÓN Y SISTEMAS CON MENCIÓN EN GESTIÓN DE TECNOLOGÍAS DE INFORMACIÓN

PRESENTADO POR:

FLORES GONZÁLEZ, MIRELLA ROCÍO

ASESOR: RAYMUNDO IBÁÑEZ, CARLOS ARTURO, PhD.

LIMA – PERÚ

2018

Con el orgullo viene el oprobio; con la humildad, la sabiduría. Proverbios 11:2

La presente tesis está dedicada a DIOS todopoderoso, a mi madre, a mi padre que está en el cielo, a mis hermanos y sobrinos, por sus esfuerzos y por ser mi inspiración para culminar esta etapa profesional.

Agradezco a los asesores que me apoyaron en la culminación de esta tesis y la institución donde se realizó la tesis.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

RESUMEN

El presente trabajo de investigación surgió como una necesidad de gestionar los procesos y el análisis los grandes volúmenes de datos generados en el ámbito científico, especialmente biología molecular.

Las tecnologías y procesos generalmente están desarrollados ampliamente para el sector comercial. Se plantea un modelo de arquitectura empresarial enfocado en el análisis bioinformático dentro de un centro de investigación científica, que permite realizar los análisis de grandes volúmenes de datos biológicos en el *cloud computing*. Se utilizó la metodología del ADM de TOGAF, que brindó un marco de trabajo para crear una arquitectura en capas que construya la estrategia y el modelo de coordinación de los diferentes procesos, tecnología, aplicaciones y datos, cada uno con alto nivel de detalle y alto nivel de flexibilidad. Esta arquitectura está orientada a modelar tanto sistemas bioinformáticos actuales como nuevos diseños de sistemas de mediana y pequeña escala. Se presenta dos casos de estudio reales que validan esta arquitectura

La investigación propone una arquitectura empresarial para el análisis bioinformático de grandes volúmenes de datos integrando todos los procesos y componentes tecnológicos en un modelo basado en el *cloud computing*.

Palabras clave: bioinformática, *cloud computing*, grandes volumenes de datos, TOGAF.

ABSTRACT

The current research project emerged as a need to manage the processes and *big data* analysis of data generated in science, especially molecular biology.

Technologies and processes are generally widely developed for the commercial sector. A business architecture model is proposed that focuses on bioinformatics analysis within a scientific research center, which allows the analysis of biological *big data* based on the *cloud computing*. TOGAF ADM methodology provides a *framework* to create a layered architecture that builds the strategy and model of coordination of different processes, technology, applications and data, each with a high detail level and a high flexibility level. This architecture is oriented to model both current bioinformatics systems and new designs of medium and small scale bioinformatics systems.

Therefore, this research proposes a business architecture for the bioinformatics analysis of big data. It integrates all process and technological components in a single mode based in the *cloud computing*.

Keywords: bioinformatics, *cloud computing*, big data, TOGAF.

INTRODUCCIÓN

En las investigaciones biológicas, cada día se crean millones de *gigabytes* de datos provenientes de biología molecular, tanto que la mayoría de los datos se han creado tan solo en los últimos 5 años. Las tecnologías de *Next Generation Sequencing* han generado que haya una explosión de grandes volúmenes de datos provenientes de biología molecular, estas tecnologías pueden generar rápidamente la secuencia de todo un genoma y hacen que sea posible comparar rápidamente el contenido genético de los seres vivos. La complejidad aumenta cuando se aborda los problemas concernientes a la generación de estos grandes volúmenes de datos, almacenamiento y mantenimiento, integración y análisis de datos, el uso de herramientas para el análisis a gran escala, la traducción de los datos de información a conocimiento, entre otros, donde todos estos se apoyan en la informática. Es así que, el uso de técnicas de cómputo para analizar datos biológicos se refiere como bioinformática, que es la base en muchas investigaciones en las ciencias de la vida.

En este sentido, surgió la necesidad de los investigadores que hacen uso de la bioinformática de buscar soluciones para el manejo de procesos, la gestión y análisis de los grandes volúmenes de datos generados de la explosión de datos de biología molecular, ya que las tecnologías y procesos convencionales no se adaptan a sus necesidades. En consecuencia, esta investigación se centró en diseñar un modelo de arquitectura empresarial que optimice la articulación entre los procesos, las aplicaciones y la infraestructura

tecnológica para los análisis bioinformáticos de grandes volúmenes de datos. Para ello, se ha planteado una arquitectura dentro del marco metodológico del ADM de TOGAF como un modelo de estrategia y coordinación de los diferentes componentes tecnológicos que permita el uso de sistemas paralelos y distribuidos, basado en la integración del modelo de los cuatro sistemas básicos formados por el sistema de almacenamiento, el sistema de memoria, el sistema de procesamiento (CPU) y el sistema de red, en una única plataforma del *cloud computing.*

La investigación pretende contribuir en la generación de conocimiento en las investigaciones a corto plazo, a menor costo y generar resultados confiables para los investigadores científicos, mediante una arquitectura empresarial para la gestión de procesos y la infraestructura tecnológica en el *cloud computing* desarrollando estándares de almacenamiento, alineamiento de procesos, conectividad de bases de datos y la definición de una arquitectura adecuada para satisfacer los requerimientos de las aplicaciones para el tratamiento de este tipo de información. El enfoque permite a los investigadores científicos continuar investigando sobre la comprensión de funciones biológicas, enfermedades u otros, que se pueden ensayar sólo a través de la secuenciación. Adicionalmente involucra mejoras en el tiempo de cálculo, cambiando así fundamentalmente la economía de gran procesamiento de datos en las ciencias biológicas y la genómica.

La investigación ha sido desarrollada de la siguiente manera. El primer capítulo describe el problema y detalla los objetivos planteados. El segundo capítulo está constituido por la teoría y conceptos básicos. El tercer capítulo presenta la metodología utilizada en la investigación. El cuarto capítulo desarrolla la arquitectura propuesta. El quinto capítulo describe los resultados obtenidos. El sexto capítulo presenta las discusiones.

CAPÍTULO I
PLANTEAMIENTO DEL PROBLEMA

1.1 Determinación del problema

La biología es una ciencia muy amplia que estudia a los seres vivos, esta disciplina tiene diversos campos de estudio como la biología atómica y molecular, biología celular, fisiología, anatomía e histología. La biología molecular estudia cómo funcionan los componentes moleculares, genes, proteínas y otros, como interaccionan, encuentran defectos y secuencian el genoma de cada individuo, etc. La bioinformática es un componente esencial para conseguir estos objetivos, debido a que con las investigaciones permiten ampliar, descubrir y transformar los conocimientos de la biología (principalmente biología molecular), que pasa de ser ciencia analítica a una ciencia de síntesis por excelencia, de una evaluación "*in vitro*" (método por excelencia de la biología) a una evaluación "*in silico*" por el papel de la computación en el nuevo laboratorio biológico (Alemán, Báez, Fernández, & Morales, 2003). Así, la bioinformática es un área de investigación multidisciplinaria, la cual puede ser ampliamente definida como la intersección principalmente entre dos ciencias: biología y computación (Campos, Campos, & Ortiz, 2011), la cual se fundamenta en la aplicación de programas de cómputo, basados principalmente en métodos de algoritmos matemáticos que ofrecen información relevante desde el punto de vista genético, evolución, propiedades antigénicas, etc.

La secuenciación genómica a gran escala se usa ampliamente para responder preguntas en varios ámbitos como la función biológica, enfermedades, evolución, agricultura, etc. Las nuevas técnicas de investigación en biología molecular incluyen el secuenciamiento a gran escala (*Next Generation Sequencing*, NGS). Las tecnologías NGS actualmente disponibles son *Illumina Genome Analyzer* (GA), *Applied Biosystems* (ABI) Sólido, *Roche* de 454, máquinas de secuenciación *Helicos 'Heliscope*, PacBio, entre otros, en la Figura 1 se muestran las más utilizadas (Lee, et al., 2013).

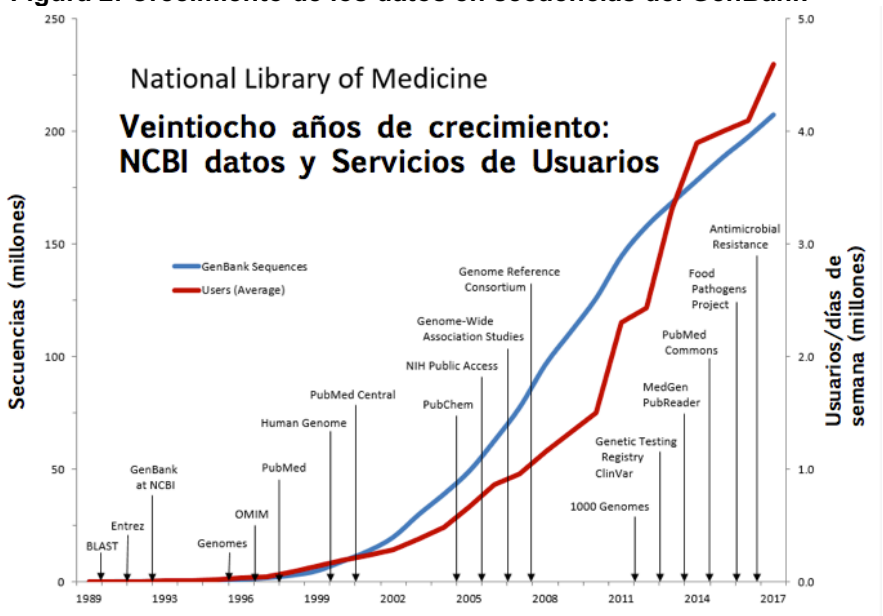
Figura 1: Generación de datos por las tecnologías de NGS

Fuente: (Chien-Yueh Lee, 2013)

Estas tecnologías proporcionan de manera rápida y económica un volumen sin precedentes de datos de secuencias biológicas de ADN/ARN de alta calidad, llamados comúnmente datos “crudos” o datos sin procesar. Estos grandes volúmenes de datos (GVD) se encuentran en ficheros del orden de gigabytes, por ejemplo, en la Figura 1 podemos observar que el volumen de datos de los archivos de salida oscila entre 1GB and 600GB. Los problemas biológicos actuales requieren soluciones que requieren métodos computacionales avanzados y sofisticados para manejar esta explosión de datos provenientes de estas investigaciones.

Adicionalmente a los datos provenientes de NGS, es necesario trabajar con datos provenientes de la extracción de una gran variedad de repositorios de datos públicos y privados, cada uno con un vocabulario único y esquema diferente. En los repositorios de datos públicos del GenBank se calcularon 207’265,929 registros de secuencias genómicas en el 2017. La Figura 2 muestra el número de secuencias genómicas en cada lanzamiento de GenBank, a partir de la Versión 3 en 1982 hasta la actualidad, donde el número de bases en GenBank se ha duplicado aproximadamente cada 18 meses, ver Figura 2. A modo de ejemplo, la cantidad de entradas de una base de datos de secuencias genéticas en GenBank ha creciendo en 38’930,533 en los últimos cinco años (National Library of Medicine, 2018), estas entradas tienden a multiplicarse, teniendo así un crecimiento exponencial. A pesar del desarrollo en las altas velocidades de conexión de las redes, cuando se comparte y accede geográficamente a la fuente de data distribuida, el periodo de latencia se convierte en un problema debido a la localización de la data, de esta manera es necesario incrementar los recursos para la obtención de estos datos, como es el ancho de banda, lo cual implica mayores gastos. Así mismo y desafortunadamente los sistemas para gestión de base de datos tradicionales (DBMS) no son capaces de manejar estos grandes volúmenes de datos a una escala tan grande.

Figura 2: Crecimiento de los datos en secuencias del GenBank



Fuente: Adaptado de (National Library of Medicine, 2018)

Los intentos de consolidar estos datos con diversos almacenes de datos o consultas añaden complejidad importante y han demostrado limitaciones en la flexibilidad, lo cual requiere un esfuerzo masivo y sostenido de los tipos de datos, así como la infraestructura tecnológica necesaria para dar soporte a estos grandes proyectos de investigación. Todos los datos por separado, se deben normalizar a través de un proceso tedioso y largo para resolver las diferencias de nomenclatura y la recopilación de información en una sola vista. Por ello, los actuales enfoques de integración de datos deben manejar la complejidad y diversidad de los grandes volúmenes de datos de los que disponemos, comprender dichos datos nos permite obtener información que es de gran utilidad. Así, nace la necesidad de implementar procesos que sean capaces de resumir de manera comprensible dicha información.

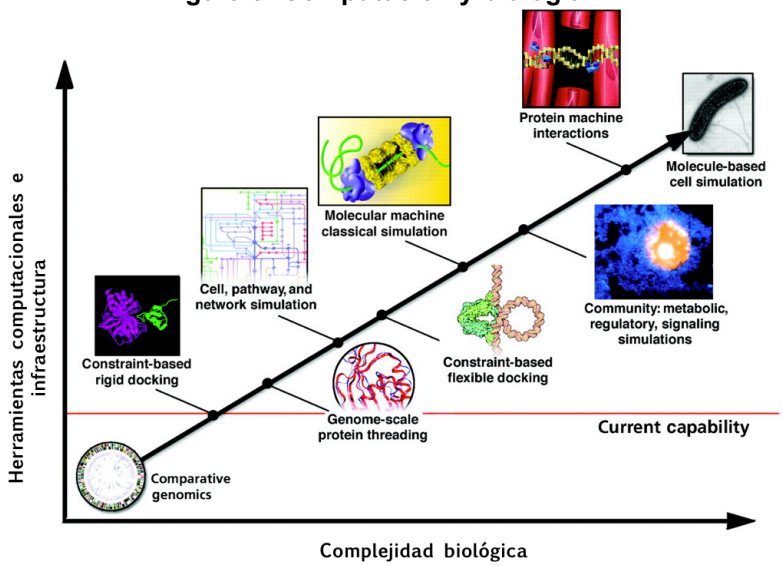
Por otro lado, como se ha saturado la capacidad de almacenamiento local, los datos se encuentran en diferentes unidades de almacenamiento personales y/o en servidores externos, causando desorden con los datos y consecuentemente problemas con la recuperación de los datos, esta diversificación da lugar a un uso inadecuado de los datos, dificultad para centralizar y compartir información, por ende se torna difícil encontrar la información porque hay demasiados datos y es difícil generar conocimiento porque hay demasiada información en diversas fuentes. Los investigadores comúnmente analizan los datos “crudos” y terminan utilizando 5 veces el volumen de almacenamiento de los datos "crudos", debido a que los datos se dejan en el lugar equivocado, encontrar datos es imposible, y manejan muchas copias de un solo ejemplar, además mover datos es difícil y lento.

Se han generado muchos datos a nivel genómico, un genoma tiene miles de genes, donde un gen puede representar miles de bases y una proteína cientos de aminoácidos. Las computadoras personales son capaces de analizar pocos miles de proteínas, en este escenario el análisis es complicado. Por tanto, el tiempo de computación ha crecido enormemente. La mayor parte de las aplicaciones en bioinformática cargan los datos en memoria principal para acelerar los cálculos, no utilizan varios procesadores para distribuir la carga de trabajo, es decir no están preparadas para un cambio de escenario.

Además, pocos programadores tienen algún entrenamiento formal en computación de alto rendimiento, y precisamente aquí se torna complicado pasar de la computación secuencial a la paralela. Así mismo, existe un desarrollo de herramientas heterogéneas y aisladas, sin interfaces o con interfaces básicas debido al mercadeo *open source* amplio; esto confunde a los investigadores que recién se inician o tienen poco conocimiento en bioinformática. Los investigadores conocedores de los análisis bioinformáticos no comparten su expertise formalmente en repositorios ya que la formación que tienen en su mayoría viene del lado de las ciencias. Por ello, se necesita un inventario y clasificación de software más utilizado para cada tipo de análisis, así mismo se requiere incluir tecnologías de punta que permitan lidiar con las necesidades de estos softwares. En tal sentido, la demanda de mejores infraestructuras y más robustas es una constante. Las aplicaciones que se utilizan procesan gran cantidad de datos biológicos y necesitan un registro de base de datos de secuencias genómicas no estructurada, las cuales se deben actualizar de los principales repositorios de datos genómicos en el mundo en este caso del Genebank, como se vio anteriormente.

Hace muchos años se pensaba que con las tasas de crecimiento, capacidad y velocidad de los ordenadores sería suficiente, pero esto ha cambiado, la demanda se ha incrementado, tal como Frazier (2003) lo manifiesta y describe en la Figura 3, hay una correlación entre la complejidad de la biología con los requerimientos computacionales y en la actualidad esta aseveración no ha cambiado. Es así que los sistemas tienen altos requerimientos para procesamiento de datos; para poder analizar conjuntos genómicos comparándolos mutuamente es necesario recurrir a los sistemas distribuidos, ya que una única computadora no puede realizar el cómputo. El uso del software especializado para bioinformática consume muchos recursos de hardware, debido a la gran cantidad de datos que procesa y a la complejidad del procesamiento. Así mismo, en caso de concurrencia de análisis causa la sobrecarga de procesamiento ya que varios análisis pueden ejecutarse en simultáneo, lo cual puede generar interrupciones en caso el servicio colapse, perjudicando así la continuidad del servicio.

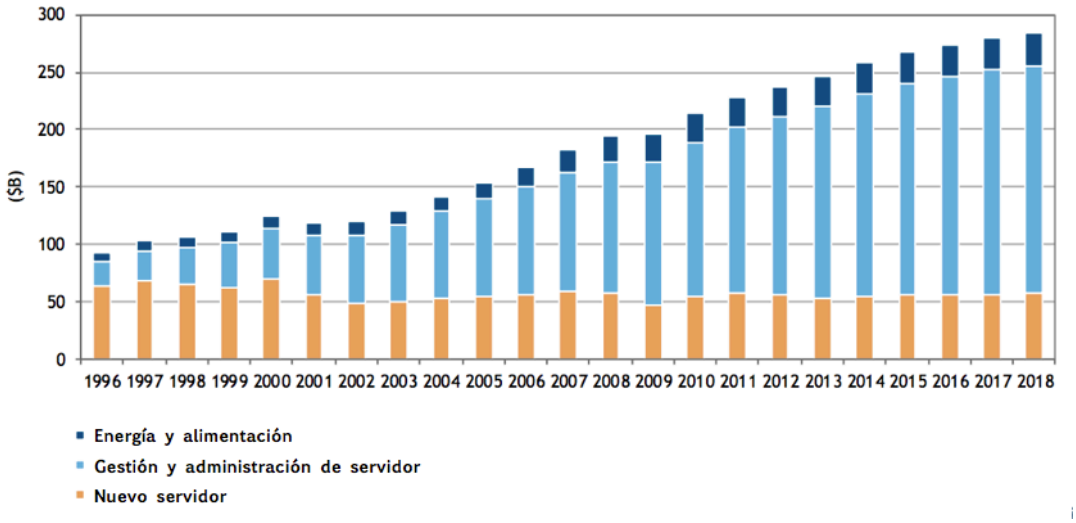
Figura 3: Computación y biología



Fuente: (Frazier, Johnson, Thomassen, Oliver, & Patrinos, 2003)

En este sentido, los requerimientos de mejores prestaciones de hardware se hacen evidentes, si bien es cierto el costo de adquisición real de los servidores han descendido en los últimos años, lo que realmente influye en la economía del centro de datos son los gastos operativos, principalmente la administración, (Bailey, 2011) muestran en la Figura 4. El gasto de adquisición del servidor y el gasto asociado de alimentación y enfriamiento son planos. Mientras el costo de administración de servidores físicos se mantiene, el costo de administración de servidores virtuales es muy elevado y asciende.

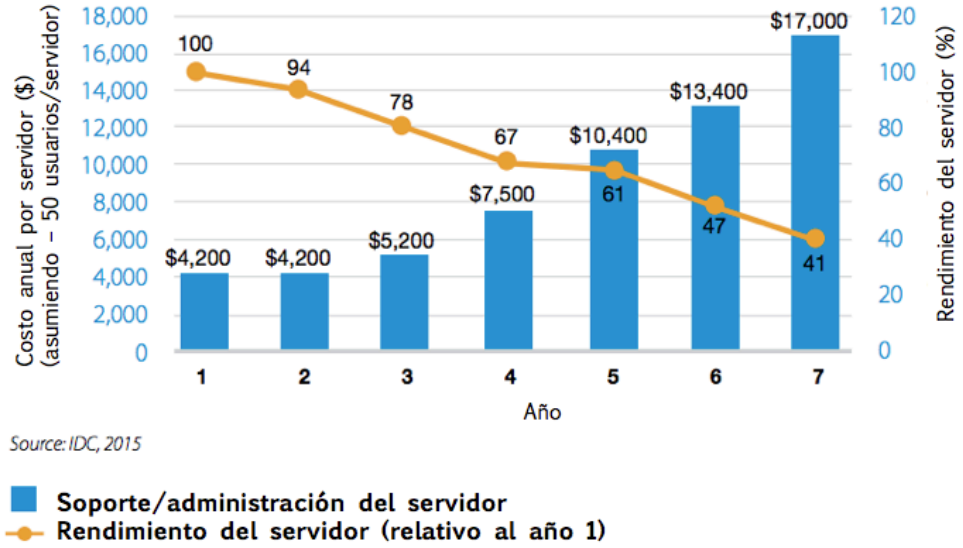
Figura 4: Nuevo modelo económico para centro de datos



Fuente: Adaptado de (Rutten, 2016)

La investigación de IDC (Scaramella, 2016), muestra como los costos y rendimiento cambian en el ciclo de vida del servidor, es así que el rendimiento del servidor disminuye cada año en un aproximado de 14% incrementando el número de fallas, lo cual genera el incremento de costos de soporte (Figura 5). Así mismo, los inesperados fallos de electricidad, tiempo de baja, además del bajo rendimiento de las aplicaciones debido al desgaste del hardware reduce la productividad de los investigadores bioinformáticos, debido a que la interrupción de los análisis genera un tiempo extra al volver a repetirlos.

Figura 5: Rendimiento de los servidores frente a los costos invertidos



Fuente: Adaptado (Scaramella, 2016)

Por otro lado, es importante analizar es la situación de producción científica de Bioinformática en Latinoamérica y especialmente en Perú. Según el estudio realizado (De Las Rivas, Bonavides-Martínez, & Campos-Laborie, 2017), muestra un total de 2,119 artículos científicos publicados entre los años 1991 y 2016 en el campo de la bioinformática en los 19 países de Latinoamérica (en el cual no considera Portugal, España y Puerto Rico), siendo este un 3.9% de un total de 53,439 publicaciones en este rubro a nivel mundial en estos mismos años. Brasil es el líder con 1,068 artículos, seguido de México con 345, Argentina con 270, etc. como se aprecia en la Figura 6. Perú se encuentra en el noveno puesto con 20 artículos científicos publicados, lo cual representaba 0.63 artículos por millón de personas en el 2016. Por otro lado, la Figura 7 muestra el crecimiento de artículos científicos para este

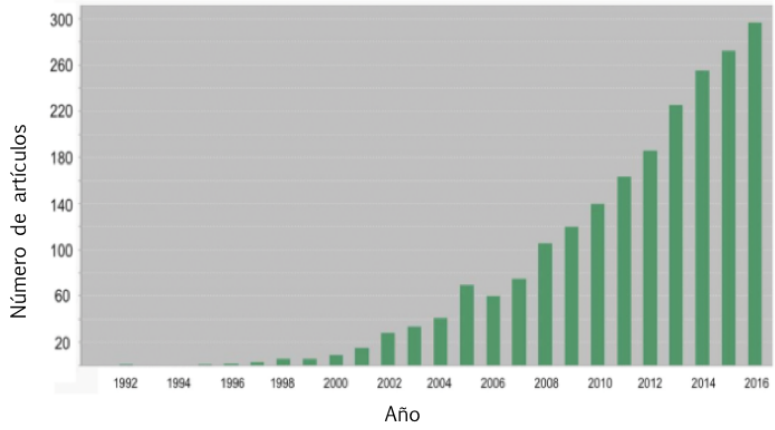
campo, sobre todo en los últimos 5 años en Latinoamérica, lo cual conduce a premisa que bioinformática está en crecimiento en Latinoamérica y de la misma forma ira creciendo en Perú.

Figura 6: Artículos publicados por país entre los años 1991 y 2016

Países	N ARTÍCULOS en Bioinformática (1991 - 2016)
Brazil	1,068
Mexico	345
Argentina	270
Chile	236
Colombia	162
Cuba	61
Uruguay	35
Peru	20
Venezuela	17
CostaRica	16
Ecuador	15
Panama	8
Bolivia	6
Guatemala	3
Honduras	3
Paraguay	3
Dominican Republic	2
El Salvador	2
Nicaragua	2
TOTAL	2,119

Fuente: Adaptado de (De Las Rivas, Bonavides-Martínez, 2017).

Figura 7: Artículos publicados en Latinoamérica entre 1991 y 2016



Fuente: Adaptado de (De Las Rivas, Bonavides-Martínez, 2017).

De todo lo expuesto anteriormente, podemos deducir que no se cuenta con un proceso formal bioinformático ni manejo de datos, el difícil tratamiento de datos, la ralentización de los análisis, la incompatibilidad con aplicaciones actuales, la gestión manual de los recursos informáticos y los altos costos generan una necesidad de mejora en la gestión de recursos y procesos en los investigadores en su afán de mejorar la productividad de la investigación. En tal sentido, es necesario afrontar toda la problemática antes mencionada

desde una vista de negocio, vista de aplicaciones. de datos y vista de infraestructura, de manera que sea viable en nuestro contexto peruano especialmente en investigación donde las tendencias indican que bioinformática está incursionando con mayor vigorosidad.

1.2 Formulación del problema

1.2.1 Problema general

¿Cómo la arquitectura empresarial impacta en el análisis bioinformático de grandes volúmenes de datos?

1.2.2 Problemas específicos

- ¿Cómo una vista de negocio impacta en los procesos bioinformáticos?
- ¿Cómo una vista de aplicaciones apoya el trabajo de los investigadores?
- ¿Cómo una vista de datos impacta en el análisis de grandes volúmenes de datos?
- ¿Cómo una vista de infraestructura apoya los análisis bioinformáticos?

1.3 Objetivos

1.3.1 Objetivo general

Mejorar los análisis bioinformáticos de grandes volúmenes de datos a través de una arquitectura empresarial.

1.3.2 Objetivos específicos

- Optimizar los procesos bioinformáticos que se llevan a cabo en las investigaciones mediante una vista de negocio.
- Implementar una vista de aplicaciones que permita responder satisfactoriamente al trabajo y necesidades de los investigadores.
- Mejorar la gestión y análisis de los grandes volúmenes de datos provenientes de las investigaciones, a través de una vista de datos.
- Proveer una infraestructura robusta que soporte los análisis bioinformáticos mediante una vista de infraestructura.

1.4 Justificación de la investigación

La genética, genómica y en general la biología molecular han presentado grandes avances en las últimas décadas. “Este vertiginoso crecimiento se debe en parte al avance de las ciencias de la computación y a su aplicación directa al estudio de la genética alcanzando así la primera anotación del Genoma Humano publicado” (Myers, 2001). Con el uso de la bioinformática ha sido posible empezar la anotación de los genomas y esta ha resultado en la elaboración de mapas genéticos y físicos cada vez más complejos, pero esta área carece de procesos formales en lugares que recién se empieza a utilizar bioinformática.

Así, la presente investigación pretende modelar una arquitectura empresarial para bioinformática que es un área relativamente nueva de las ciencias y que no tiene precedentes en implementación de arquitecturas empresariales, brindando así los lineamientos en cuanto al desarrollo de los procesos, el manejo de los datos, el uso de las aplicaciones y la tecnología, en centros de investigación que recién comienzan a hacer el uso de la bioinformática en sus investigaciones.

El alcance de este proyecto abarca el diseño de un modelo de arquitectura empresarial para un área de bioinformática que podrá ser implementado dentro de pequeños o medianos centros de investigación científica como por ejemplo agricultura, con campo de acción en las biociencias moleculares: biología molecular, bioquímica, genómica y genética. El producto a entregar consta del diseño de la arquitectura empresarial, el modelo de negocio, aplicaciones, datos y la infraestructura tecnológica que interactúan, así como el mapeo e integración de todas las actividades que soportan todos estos procesos.

1.5 Limitaciones de la investigación

El proyecto contempla la validación del modelo por parte de los investigadores quienes realizan los análisis bioinformáticos, mas no contempla la validación de los resultados de estos análisis, debido a que no

es objetivo de esta investigación y está sujeto a la interpretación, experiencia y conocimiento de los investigadores científicos.

Los algoritmos de bioinformática y la potencia de cálculo son las principales limitantes para el análisis de la generados por las tecnologías actuales de secuenciamiento a gran escala; no obstante, se considera que el uso del *cloud computing* proporciona diferentes tipos de tecnologías escalables y por ende una mejora para el funcionamiento de las aplicaciones.

La investigación no contempla el respaldo, seguridad de información y garantía del servicio en el *cloud computing*, debido a que estos son asumidos por el proveedor en el Contrato a Nivel de Servicio (SLA). El SLA tiene una disponibilidad del 99,95% anual, en el tiempo restante pueden existir errores de acceso o conexión. El acceso físico a la infraestructura tecnológica no está contemplado, debido a que el uso del *cloud computing* es remoto.

1.6 Viabilidad de la investigación

Los resultados de los estudios son a menudo generalizables hasta cierto punto, pero según (Routio, 2007) un método más formal para presentar conocimiento general es hacer un modelo. Una técnica usual para lograr un modelo científico generalizable es detectar invariantes, quitar la variación y abordar el problema. Por lo tanto, la investigación es técnicamente viable.

La investigación es económicamente, socialmente y operativamente viable debido a que este estudio se desarrolló en un centro de investigación piloto en agricultura, donde el modelo fue evaluado a través de la experiencia de los investigadores (los expertos), quienes luego del uso de los componentes de la implementación determinaron la fiabilidad y conformidad de los resultados.

CAPÍTULO II
MARCO TEÓRICO

2.1 Antecedentes del problema

La bioinformática es la base para investigaciones en biología, donde en muchos casos se vuelve indispensable, por ejemplo, en medicina, el diagnóstico genético en personas puede detectar enfermedades genéticas que de otra manera no pueden ser pero generalmente las personas no tienen esta información.

> "La tecnología proporciona un elemento teórico y proporciona las herramientas prácticas, para que los científicos puedan explorar las proteínas, RNA y DNA. Desde el principio de los años 90, muchos laboratorios han estado analizando el genoma completo de varias especies tales como bacterias, levaduras, ratones y seres humanos. Durante estos esfuerzos de colaboración, se han generado cantidades enormes de datos los cuales se recogen y se almacenan en grandes bases de datos, la mayoría de las cuales son publicadas y accesibles" (Valdiosera, 2006).

La comparación de secuencias, ya sea nucleótidos o aminoácidos, para buscar semejanzas o diferencias se realizan de manera automatizada debido a que naturalmente no es posible comparar cientos de miles de pares de bases de manera manual, ya que conlleva a una pérdida de tiempo y al aumento de errores en la comparación. Para esto, muchas técnicas computacionales se han desarrollado para abordar este problema, lo cual se llama comúnmente bioinformática. Así, con el incremento de la capacidad computacional y la complejidad de los sistemas y las técnicas de

investigación, ha surgido la necesidad de tener expertos en ambas áreas, biología y computación.

El uso de computadoras para afrontar los problemas biológicos se inició con el desarrollo de algoritmos que permitan la comprensión de las interacciones en los procesos biológicos y sus relaciones con otros organismos. Luego, el aumento desmesurado de secuencias biológicas disponibles y la complejidad de las técnicas usadas por las computadoras para la obtención y análisis de los datos, hacen de la bioinformática un área en demanda. La bioinformática y otra multitud de herramientas de la biología de sistemas desarrollados se ejecutan en línea de comandos ya que no han adoptado un enfoque basado en la web o una interfaz más amigable para la interacción con los investigadores.

El desarrollo de la bioinformática en Perú es incipiente en investigaciones agrícolas llevada a cabo por las instituciones académicas y de desarrollo tecnológico, en contraste a otros países, donde la bioinformática es el eje principal para este tipo de investigación; sin embargo, en cualquier contexto las investigaciones generan una demanda de servicios informáticos en un tiempo relativamente corto para los análisis bioinformáticos.

Por otro lado, según (Zhang, Big Services Era: Global Trends of Cloud computing and Big Data, 2012), las próximas innovaciones son alrededor de arquitectura de la tecnología, hay tres tendencias principales. La primera tendencia de innovación es integrar hardware, software y servicio en una caja. Este tipo de innovación puede ocultar la complejidad de los sistemas de TI de las empresas y hacerlos listo para el *cloud computing*. Lo más importante, el aspecto de servicios de la caja puede capturar las mejores prácticas de operaciones en diversos escenarios. La segunda tendencia innovación es proporcionar plataformas de *cloud computing* abierta por los principales proveedores de servicios de Internet. Los algoritmos de análisis de los grandes volúmenes de datos, APIs almacenamiento en el *cloud computing* y las API de predicción y de traducción se están convirtiendo gradualmente en instrumentos y activos importantes para las innovaciones abiertas. La tercera

innovación de arquitectura tecnológica viene de Internet móvil y los servicios de redes sociales. Arquitectura tecnológica se usa para realizar la arquitectura de aplicaciones y arquitectura de datos, lo que apoya aún más la arquitectura de negocios.

En tal sentido, la amplia disponibilidad de instrumentos de secuenciación de ADN de alto rendimiento, y el desarrollo de nuevas técnicas basadas en la secuenciación, proporciona un recurso potencialmente muy valioso para los investigadores bioinformáticos. La adopción generalizada de secuenciación de alto rendimiento ha aumentado considerablemente la escala y la complejidad de la infraestructura computacional necesaria para realizar la investigación genómica. Es así que la transformación de los datos de secuencia en información biológicamente significativa requiere una infraestructura computacional sofisticada y de apoyo. El tamaño de la infraestructura computacional requerido supera las expectativas de la infraestructura local tradicional. Así mismo, la configuración y el mantenimiento asociado con una infraestructura computacional local presentan problemas significativos para los investigadores individuales y representan esfuerzos importantes para el soporte informático necesario. Una alternativa a la construcción y mantenimiento de la infraestructura local es el *cloud computing*, que ofrece acceso bajo demanda a la infraestructura computacional flexible.

2.1.1 Investigaciones relacionadas

a) Registro de herramientas y servicios de datos: un esfuerzo de la comunidad para documentar los recursos bioinformáticos (Ison J, 2016).

En el artículo científico publicado por Ison, se presenta una infraestructura europea de información biológica, para un registro integral y consistente de información sobre recursos bioinformáticos. El mantenimiento sustentable de este registro de herramientas y servicios de datos está garantizado por un esfuerzo de curación dirigido y adaptado a las necesidades locales, y compartido entre una red de socios comprometidos. En noviembre de 2015, el registro tenía 1785 recursos, de 126 registros individuales,

incluidos 52 proveedores institucionales y 74 personas. Su fin es el de crear un estándar para la diseminación de información sobre recursos bioinformáticos. El registro debería ayudar al descubrimiento y uso eficiente de herramientas y, por lo tanto, proporcionar un soporte útil para proyectos de ciencias de la vida (Ison J, 2016).

En general este repositorio alberga una lista de herramientas utilizadas y permite su filtro de acuerdo a ciertos parámetros como fecha de publicación, cantidad de citaciones, actualizaciones, etc.

b) Propuesta de *framework* de arquitectura empresarial para pymes basado en un análisis comparativo de los *framework* de Zachman y TOGAF (Matute, 2014)

El estudio de (Matute, 2014) sugiere que se debe considerer las complicaciones existentes al momento de aplicar un *framework* en una organizacion, es importante tomar en cuenta las areas de negocio a ser cubiertas, asi como las necesidades dentro del entorno y la tecnología desarrollada existente para ser incluida en la optima integración.

La investigación aborda los conceptos básicos para poder comprender los términos de arquitectura empresarial y framework, asi mismo desarrolla el análisis exhaustivo de los 2 *frameworks* más utilizados, Zachman y TOGAF, con el objetivo de proporcionar herramientas que permitan evaluar las ventajas y desventajas de estos frameworks, y asi tener los criterios suficientes para poder dimensionar las fortalezas, debilidades, oportunidades y amenazas de cada uno de estos. Además, se estudian los conceptos e implicaciones para comprender las PYMES y sus alances y más específicamente de las PYMES en Ecuador, haciendo uso del análisis FODA y de la aplicación directa de un *framework* en una PYME. Finalmente, describe una propuesta de framework hibrido de acuerdo a los *frameworks* descritos previamente, tanto Zachman y TOGAF, con el objetivo de proponer un método sencillo, que facilite un procedimiento para poder ser implementado en una PYME (Matute, 2014).

Por último, Matute concluye que una arquitectura empresarial es clave para facilitar los procesos de negocio y mejorar los servicios de TI en una organización, debido a que se realiza el alineamiento de la estrategia de negocio, los servicios de información y la infraestructura dentro de las organizaciones.

c) Servicio de selección de *cloud* de calidad-conscientes para modeladores computacionales (Nizamani, 2012).

Esta tesis doctoral ayuda a los modeladores de cómputo, a seleccionar el proveedor de servicios cloud más rentable. Esto es cuando se opta por utilizar el *cloud computing* en lugar de usar las instalaciones High Performance Computing (HPC) in house. Se propone y evalúa un nuevo servicio de selección computacional de calidad consciente (QAComPS). Esto selecciona el más barato proveedor de servicios de la nube. Después de la selección se pone en marcha de forma automática y se ejecuta el servicio seleccionado. Después de la selección se pone en marcha de forma automática y se ejecuta el servicio seleccionado. QaComPS incluye una ontología integrada que hace uso de OWL (Ontology Web Languaje) características. La ontología proporciona una especificación estándar y un vocabulario común para describir diferentes proveedores de servicios de la nube. Las descripciones semánticas son procesadas por el servicio de gestión de la información QaComPS. Estas descripciones de proveedores son utilizadas por un filtro para seleccionar automáticamente el servicio de más alto rango que cumple con los requisitos del usuario. Una interfaz se utiliza para transferir información semántica a o desde el servicio de Gestión de la Información QAComPS y la selección no semántica y servicios administrados (Nizamani, 2012).

El servicio también se evaluó cualitativamente por un grupo de modelistas computacionales. Los resultados de la evaluación fueron muy prometedores y demostraron el potencial de QaComPS de hacer computación en nube más accesible y rentable para los modelistas computacionales.

2.1.2 Plataformas bioinformáticas

Existen varios proyectos y soluciones en el contexto de alto rendimiento de secuenciación y bioinformática en general que utiliza *cloud computing* para generar la capacidad computacional y la escalabilidad bajo demanda. Sin embargo, estos proyectos se refieren principalmente a problemas específicos y proporcionan soluciones personalizadas para una herramienta o metodología determinada. Aunque es muy valiosa, tales herramientas proporcionan una ayuda mínima para los investigadores que desean componer una variedad de herramientas en un análisis o para los investigadores que desean utilizar sus propias herramientas cuando se utilizan los recursos de *cloud computing*. Por lo tanto, hay una necesidad de una solución flexible que se puede utilizar en una variedad de escenarios y que proporciona soporte para la personalización.

Galaxy Cloudman es una plataforma basada en la web para los datos de investigación en biomédica y apoya a los usuarios a configurar y trabajar en clouds. Funciona en combinación con Bio-Linux y configura en el arranque Oracle, así como Galaxy. Desafortunadamente, Cloudman no soporta MapReduce y por lo tanto una ejecución gráfica del seguimiento de los trabajos no es posible, así como los beneficios de la computación paralela.

Eoulsan es una estructura modular que permite la configuración de clusters de computación en nube. Un sistema de conexión modular permite la integración de algoritmos disponibles predefinidos, lo cual no satisface a los científicos quienes desean un entorno flexible y dinámico. Eoulsan solo tiene que ser ejecutado en la línea de comandos. Este sistema mejora el uso de programas para los científicos, pero no se centran en una ejecución gráfica de puestos de trabajo en grupos públicos y privados.

Todas estas herramientas se enfocan principalmente en las herramientas de análisis, mas no proporcionan una visión global del proceso de negocio, ni en el tratamiento de los GVD. Por otro lado, el no tener el control absoluto del repositorio de los datos es otro punto en contra para los investigadores quienes son cuidadosos de los datos.

2.2 Bases teóricas

2.2.1 Bioinformática

La informática es actualmente una herramienta básica para la biología, de tal forma que el proyecto más ambicioso de la historia de la biología, el secuenciamiento del genoma humano, no hubiera sido posible. La biología ha cambiado mucho a través de los años, según (Barraza, 2014) ha habido un avance notable en las técnicas de laboratorio, pero además se convirtió en una ciencia que está generando muchos datos. Por ejemplo, la secuenciación del genoma humano, nuevas técnicas para el estudio simultaneo de cómo funcionan miles de genes. Las áreas biológicas (Barraza, 2014) se centran principalmente en genómica (caracterización y predicción de genes), evolución (análisis filogenético), proteómica (modelamiento de proteínas) y metabolómica.

La bioinformática tiene el gran objetivo de almacenar la gran cantidad de datos y aumentar la velocidad de los análisis de estos grandes volumnes de datos que la ciencia está manejando desde la publicación del genoma humano en el año 2001. Este avance con el genoma es fundamental, ya que ha servido para la crear nuevas medicinas, transgénicos, etc, a su vez también ha traído consigo el aumento de información. La situación se complica cuando cada 6 mese los datos se duplican según algunas predicciones pesimistas. Por su lado, las grandes empresas como IBM, Seagate, Hewlett-Packard, Hitachi, etc tratan en conseguir memoria de alta densidad que con lleva a la ampliación e implementación de la biotecnología (Joyanes, 2003).

El término bioinformática se refiere a la aplicación de las técnicas informáticas al estudio de la información genética (RAE, av. 23.a Ed). Según la definición del *National Center for Biotechnology Information* (NCBI): "Bioinformática es un campo de la ciencia en el cual confluyen varias disciplinas tales como: biología, computación y tecnología de la información" (National Library of Medicine, 2018).

El fin último de este campo es facilitar el descubrimiento de nuevas ideas biológicas, así como crear perspectivas globales a partir de las cuales se puedan discernir principios unificadores en biología. Al comienzo de la "revolución genómica", el concepto de bioinformática se relacionaba solamente con el desarrollo y mantenimiento de la base de datos donde tenían los datos biológicos, los cuales son las secuencias de aminoácidos y nucleótidos. La creación de esta base de datos no solo implica su diseño, además también las interfaces en las cuales se puedan administrar y acceder a los datos.

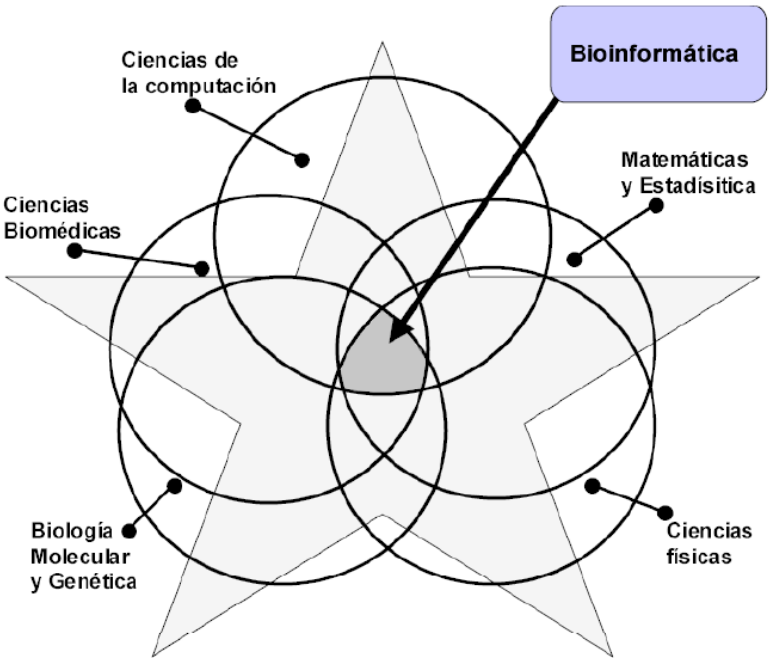
La bioinformática para (Castro, 2009), se refiere a la gestión y análisis de datos biológicos utilizando la computadora y las tecnologías disponibles, este campo abarca matemática, inteligencia artificial, estadística, química, informática y computación. Además, lo define como una ciencia que se dedica a estudiar la biología y sus fenómenos, como la biológica molecular desde un alcance computacional, con el objetivo de generar métodos que faciliten el comprender, simular y predecir los sistemas biológicos en estudio, mediante las herramientas de computación y generar respuesta o soluciones en el análisis de datos, simulación de sistemas, generalmente a nivel de biológica molecular.

La bioinformática es una disciplina académica y de investigación pero que ha demostrado las posibles grandes implicaciones para la futura comprensión de una gama grande de problemas médicos, moleculares y biológicos que busca capturar, almacenar, analizar e integrar información biológica y genética, aplicando técnicas de computación a la gestión de la datos biológicos y métodos provenientes de disciplinas tales como matemáticas aplicadas, informática, estadística, inteligencia artificial, ciencias de la salud, etc. Esta nueva disciplina se dedica fundamentalmente a los problemas de almacenamiento y análisis de la información que se produce a partir de un desarrollo de las tecnologías genómicas y aporta en numerosas áreas de investigación que van desde el modelado de ciclos de células, hasta biología computacional y genética, biofísica, genómica, proteómica, análisis de secuencias y biología de sistemas, la bioinformática permite:

- El desarrollo de fórmulas matemáticas y estadísticas, que permitan relacionar diferentes componentes, por ejemplo, encontrar un gen en una secuencia y predecir su estructura, crear árboles genealógicos, etc. (Universidad de Valencia, 2016).
- El estudio de los algoritmos que den soporte a los proyectos genéticos.
- Identificar agentes causantes de enfermedades a las plantas, como los virus, que de otra forma serian complicados de detectar.
- El uso de herramientas informáticas tales como bases de datos, almacenes (data *warehouse*), minería de datos (data *mining*), minería de datos en la Web (web *mining*), tecnologías colaborativas y lenguajes de programación fundamentalmente para Internet, facilitan el proceso de esos inmensos mares de datos que supone la biología.

La integración de herramientas y áreas de I+D+i propias de informática (Figura 8), tales como Interacción hombre-computador, computación científica y paralela, entornos distribuidos para resolución de problemas, bibliotecas digitales, Ingeniería de Software e Ingeniería Web (Coltell,2002).

Figura 8: Bioinformática, centro de otras ciencias



Fuente: (Coltell, 2002)

La generación de secuencias, así como el subsiguiente almacenamiento, interpretación y análisis son tareas hoy en día enteramente dependientes de computadores. Pero el primer reto de la comunidad bioinformática actual es el propio almacenaje de esta cantidad masiva de datos, ya que debe hacerse de manera inteligente y eficiente. En ella recae la responsabilidad de proveer un acceso fácil y fiable a estos datos. Porque los datos en sí mismos carecen de significado antes de ser analizados, y el volumen de datos en el presente hace imposible la interpretación manual. Es por ello que herramientas computacionales muy incisivas han de ser desarrolladas para permitir la extracción de información biológica útil. Según (Lechuga, 2012) existen tres procesos biológicos centrales alrededor de los cuales se han de crear herramientas bioinformáticas:

- La secuencia de ADN precede a la secuencia de la proteína.
- La secuencia de la proteína determina su estructura.
- La secuencia de la proteína determina su función.

La integración de la información aprendida acerca de estos procesos biológicos fundamentales nos debería permitir alcanzar el fin último tan deseable de comprender completamente los organismos biológicos.

2.2.1.1 Áreas de investigación bioinformática

El objetivo principal de la bioinformática es comprender mejor los procesos biológicos. Para lograr este objetivo, la bioinformática se enfoca en el desarrollo de técnicas de cálculo complejas, tales como el reconocimiento de patrones, la minería de datos, el algoritmo de aprendizaje automático, algoritmos de HPC y visualización. Se presentan las principales áreas (Orobitg, 2013), donde la informática, matemáticas y estadística se han utilizado para definir los métodos y sistemas capaces de encontrar soluciones y facilitar la investigación de algoritmos.

a) Análisis de secuencias

Consiste en comparar el ADN, ARN o secuencias de proteínas para entender sus características, función, estructura y evolución. La comparación de secuencias nuevas con aquellos con funciones conocidas es

una forma clave de la comprensión de la biología de un organismo del que viene la nueva secuencia. Las metodologías de análisis de secuencia más comunes son la alineación de secuencias y búsquedas en bases de datos biológicas. En general, existen muchas herramientas y técnicas que proporcionan la secuencia de alineaciones y analizan los alineamientos producidos para entender su biología.

b) Anotación de genomas

En el contexto de la genómica, anotación es el proceso de definición de los genes y otras características biológicas de la secuencia de ADN. Este proceso se divide en tres pasos: En primer lugar, se identifican las porciones del genoma que no codifican para proteínas. Luego, la predicción de genes o la búsqueda de genes, consiste en la identificación de elementos en el genoma. Por último, la información biológica se adjunta a las secuencias.

c) Biología evolutiva computacional

Se refiere al estudio del origen y la descendencia de especies, así como sus cambios en el tiempo. Consiste en la construcción del árbol de la vida que determina las relaciones evolutivas entre las especies.

d) Análisis de la expresión génica

Los genes son expresados de acuerdo a la medición de mARN presente al momento de hacer experimento mediante diferentes técnicas de laboratorio. Estas técnicas generalmente pueden contener ruido, contaminaciones o no ser muy precisas. Esto precisa tener herramientas estadísticas que puedan excluir el ruido presente en la muestra para medir adecuadamente la expresión de los genes.

e) Análisis de regulación génica

Conjunto de mecanismos utilizados por las células para aumentar o disminuir la producción de productos génicos específicos (proteínas o ARN). Técnicas bioinformáticas han sido aplicadas a la exploración de varios pasos en este proceso, que permitan reconocer

regiones de regulación genética. Los datos de expresión pueden usarse para inferir la regulación génica.

f) Análisis de expresión de proteínas

Este es un subcomponente de la expresión génica que consta de las etapas después de ADN ha sido transcrita en ARN mensajero (mARN). La expresión de proteína es comúnmente utilizada por investigadores para denotar la medición de la presencia y abundancia de una o más proteínas en una célula o tejido particular.

g) Genómica comparativa

Las herramientas bioinformáticas pueden usarse en el estudio de las relaciones entre números, localizaciones y funciones bioquímicas de los genes de especies diferentes, debido a que luego de ser secuenciados, se comparan las estructuras. La genómica comparativa supone que los componentes del genoma que son similares entre especies diferentes son conservados en el tiempo, de igual manera a los elementos clave para la supervivencia.

h) Modelado de sistemas biológicos

Consiste en la elaboración y utilización de algoritmos eficientes, estructuras de datos, visualización y herramientas de comunicación con el objetivo de modelar sistemas biológicos con las computadoras. Las simulaciones artificiales tratan de describir los procesos evolutivos por medio de la computadora de sencillas formas de vida (artificial).

i) Análisis de imágenes de alto rendimiento

El análisis de imágenes de alto rendimiento consiste en el uso de herramientas computacionales para automatizar el análisis de grandes cantidades de imágenes a través del procesamiento y cuantificación. De lograrse un sistema de alta calidad y con mínimos errores, se podría prescindir del observador.

j) Estructura de la proteína de predicción

Haciendo uso de la computación se trata de predecir o moldear la estructura tridimensional de unas proteínas partiendo de su composición en aminoácidos, la cual es su estructura secundaria y terciaria a partir de la estructura primaria. Esta predicción es muy importante y difiere del diseño de proteínas.

k) Análisis de mutaciones en el cáncer

El reordenamiento de las células afectadas por el cáncer es complejas e impredecibles. En muchos genes en el cáncer, se trata de identificar si existen sustituciones de bases individuales aun no conocidas, mediante secuenciación exhaustiva. Los sistemas automatizados que se producen en la bioinformática sirven para administrar la cantidad de información y secuencias obtenidas y generar nuevos algoritmos y software que comparen estas secuencias con las secuencias del genoma humano y los polimorfismos.

l) Interacción molecular

Estudio de las interacciones entre proteínas, ligandos y péptidos. Para el estudio de las interacciones moleculares, algoritmos de acoplamiento son algoritmos eficientes, que consisten en la simulación molecular dinámica del movimiento de los átomos alrededor de enlaces rotativos.

m) Medición de la biodiversidad

Un ecosistema tiene la biodiversidad de un conjunto completo de especies de aquel medio ambiente y sus genomas.

Las bases de datos almacenan los nombres de las especies, asi como su información básica, tamaño de la población, distribución, estados, información genética, relación con otras especies, hábitat entre otros. El software deberá analizar, predecir, compartir y visualizar esta información

2.2.1.2 Aplicaciones de la bioinformática

La ciencia de la bioinformática tiene usos muy dispares a la par que beneficiosos en el mundo actual, incluyendo la medicina molecular, genómica micro biótica, agricultura y ganadería, y estudios comparativos. Es más, la evolución de estas y otras ramas afines del saber es inimaginable hoy en día sin la aportación que la bioinformática supone, ya que en última instancia se encarga de tender el puente entre el vasto conocimiento anterior y la ingente cantidad de datos y recursos disponibles hoy en día.

a) Medicina molecular

El genoma humano tiene una gran importancia en los campos de la investigación biomédica y de la genética clínica ya que las enfermedades tienen, en mayor o menor medida, un componente genético. Éste podrá ser heredado (como en el caso de muchas enfermedades hereditarias, incluyendo la fibrosis quística) o ser el resultado de la respuesta del cuerpo a un estrés ambiental que causa alteraciones en el genoma (como el cáncer, diabetes, enfermedades cardíacas, etc.). El conocimiento completo del genoma humano nos permite buscar los genes directamente asociados con diferentes enfermedades, tratando de descubrir o mejorar nuestro entendimiento acerca de las bases moleculares de estas enfermedades. Este nuevo conocimiento permitirá mejorar tratamientos, curas e incluso desarrollar test predictivos.

La cantidad de datos y la complejidad de los problemas que hacen que la computadora y herramientas estadísticas esenciales para estudios exitosos. Con las recientes mejoras en las tecnologías que ahora permiten la determinación simultánea de muchos miles de marcadores, el análisis de datos se está convirtiendo en el cuello de botella de los estudios, y por lo tanto cada vez es más importante el desarrollo de métodos de análisis más rápidos y mejores.

b) Genómica microbiana

Los microorganismos son ubicuos, es decir, que se encuentran en todas partes. Han sido encontrados sobreviviendo prósperamente en

ambientes extremos de calor, frío, radiación, salinidad, acidez y presión. Están presentes en el medioambiente, en nuestros cuerpos, en el aire, en la comida y en el agua. Y tradicionalmente se han usado diversas propiedades de los microorganismos en la panadería, destilería y en la industria alimenticia en general. La llegada del genoma completo de estos microorganismos podría aumentar sus aplicaciones en el medioambiente, salud o industria.

c) Agricultura y ganadería

El secuenciado de los genomas de plantas y animales pueden reportar enormes beneficios para la comunidad agrícola. Las herramientas bioinformáticas pueden usarse para encontrar los genes en dichos genomas y elucidar sus funciones. El conocimiento genético específico podría entonces usarse para producir cultivos más fuertes, resistentes a sequías, enfermedades e insectos, así como para mejorar la calidad del ganado haciéndolo más saludable, resistente a enfermedades y más productivo.

2.2.2 Arquitectura empresarial

Definamos primero los términos por separado, primero se define empresa y luego arquitectura. “Una empresa es un grupo de departamentos u organizaciones que cuentan con recursos y que tienen un conjunto común de metas, fines y principios” (Maya, 2010), por lo tanto, una empresa puede ser una organización, una corporación, una sucursal, una división o solo un área. Por otro lado, “una arquitectura es la organización fundamental de un sistema, que incorpora sus componentes, las relaciones entre ellos y con el entorno, y los principios que gobiernan su diseño y evolución” (Maya, 2010).

Entonces, una arquitectura empresarial estudia los componente de una empresa, por lo tanto “una arquitectura empresarial puede considerarse como una colección de procesos de negocio, de sistemas o aplicaciones, de tecnologías y de datos que soportan las estrategias de negocio de una empresa. Así, una arquitectura empresarial captura información detallada acerca de los cuatro dominios o áreas, realiza una descripción completa de la empresa desde diferentes perspectivas y logra una visión” (Maya, 2010).

> Una arquitectura empresarial puede compararse con un plano que permite la ubicación óptima de los recursos de TI (Tecnologías de la Información), los cuales son el soporte de la función del negocio. El objetivo de una arquitectura empresarial es crear un entorno de TI unificado, es decir sistemas de hardware y software estandarizados, que se enlacen con el negocio de la organización y con su estrategia (Maya, 2010).

2.2.2.1 *Frameworks*

> Un framework es un lenguaje para permitir la comunicación de los stakeholders de una arquitectura empresarial. Este tiene un método detallado, un conjunto de herramientas de soporte, que proporciona directrices sobre cómo describir o documentar arquitecturas, el cual generalmente no proporciona lineamientos sobre cómo construir o implementar una arquitectura específica o sobre cómo desarrollar o adquirir sistemas (Maya, 2010).

Existen diferentes frameworks para desarrollar una arquitectura empresarial, pero todos tienen cuatro elementos básicos (arquitectura de negocio, de información, de sistemas y de tecnologías de infraestructura).

Los modelos de arquitecturas empresariales o frameworks en capas son muy usados, ya que proponen modelos de divisiones en la empresa que no se mezclan. Pero en líneas generales no se tiene un consenso de los lineamientos de una arquitectura en capas, es así que se puede utilizar varios modelos en una empresa. En la siguiente sección se describe los *frameworks* más utilizados actualmente.

a) TOGAF

TOGAF es un marco de referencia de arquitectura (The Open Group, 2013), y herramienta para asistir en la aceptación, creación, uso y mantenimiento de arquitecturas. Está basado en un modelo de procesos apoyado por las mejores prácticas y conjunto reutilizable de activos arquitectónicos existentes.

TOGAF se enfoca en el desarrollo de una arquitectura y gobierno como una herramienta ágil. Específicamente define los modelos a

usarse como representación de la arquitectura, de manera que proporcione una dirección para el desarrollo del proceso. TOGAF puede ser usando en diferentes tipos de organizaciones pequeñas, medianas o grandes, inclusive en el gobierno, debido al modelo escalable que presenta. Debido a todos los niveles que este framework manejar, TOGAF se hace cargo de la arquitectura de negocios hasta la arquitectura de datos y tecnología. Además, este *framework* es fácilmente customizable de acuerdo a las necesidades de la empresa, así como los productos *open source*, sin dejar de lado que la retroalimentación e información son procesos reales (Bustamante, 2016).

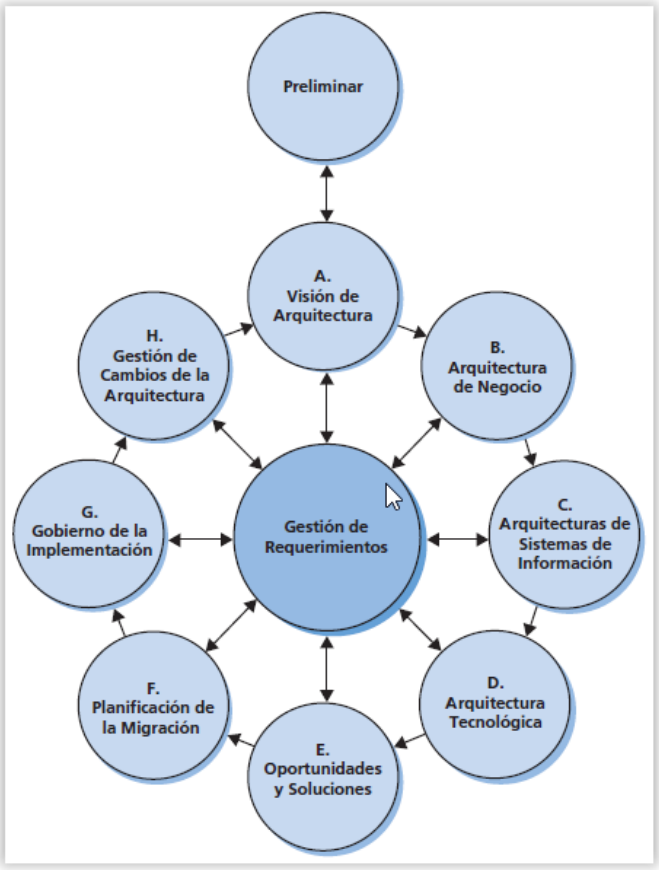
Los principales beneficios de TOGAF (Bustamante, 2016) son:

- Fue desarrollado con años de investigación por arquitectos reconocidos.
- El vocabulario usado es simple, de manera que todos puedan comprender la información dentro de las organizaciones.
- Define el sistema de información como bloques que luego interactúan unos con otros.
- Posee estándares recomendados.
- Posee productos que se utilizan para complementar los bloques.

Según (The Open Group, 2013), “el Método de Desarrollo de Arquitectura (ADM) es un método para obtener arquitecturas empresariales que son específicas para la organización y está especialmente diseñado para responder a los requerimientos del negocio”. El ADM describe:

- Un método estándar seguro que permite desarrollar una arquitectura empresarial.
- Un método que define la arquitectura al nivel de negocio, aplicaciones, datos y tecnología, asegurando que todos los requerimientos de la empresa se cumplan.
- Posee una guía con técnicas para el desarrollo de la arquitectura empresarial.

Figura 9: TOGAF Arquitectura método de desarrollo



Fuente: (The Open Group, 2013)

El ADM consta de ocho fases de desarrollo, complementados con la fase preliminar en la que se definen el marco y los principios. El ADM está configurado como un modelo de proceso iterativo, donde la etapa final indica el comienzo de una nueva iteración. Las descripciones actuales de arquitectura disponibles son consideradas como la arquitectura de línea de base. El TOGAF ADM se describe en la Figura 9.

En las siguientes secciones, se describen las distintas fases de este método. TOGAF (Standard O. G., TOGAF Versión 9.1, 2011) en su *framework* de alto nivel reconoce tres dominios o vistas que son: arquitectura de negocios, arquitectura de datos e Información, arquitectura de aplicaciones y arquitectura tecnológica. Para cada una de estas se entrará a identificar las que tienen mayor relevancia en el contexto de la organización.

- **Visión de la arquitectura**

La fase A se inicia con un requerimiento de trabajo de arquitectura de la organización patrocinadora a la organización de arquitectura. Define lo que es y lo que está fuera del ámbito de aplicación, que deben tomar las decisiones sobre la base de la evaluación practica de los recursos y la disponibilidad de la competencia.

Entrada para esta fase es, por ejemplo, pero no se limita a: el modelo de organización de la arquitectura empresarial, con las necesidades presupuestarias y las solicitudes de cambio, y el repositorio de arquitectura poblada que contiene documentación arquitectónica existente.

Salida para la fase A puede incluir una declaración de trabajo arquitectura, estados refinados de los principios de negocio, los conductores y las metas, los principios de la arquitectura y de un documento de definición de arquitectura proyecto aprobado.

- **Arquitectura de negocio**

La arquitectura de negocio describe la estrategia de producto o servicio y el organizacional, funcional, procesos, información y áreas geográficas del entorno empresarial. El alcance de esta fase dependerá sobre todo en el entorno empresarial. Las vistas que se contemplaran para este dominio, son las siguientes:

  - Mapa de objetivos empresariales
  - Estructura administrativa

La entrada para la fase B incluye descripciones existentes arquitectura, los objetivos de negocio y principios, arquitectura repositorio.

Output consiste en una versión refinada y actualizada de los entregables de la fase arquitectura de visión, un documento para la definición del proyecto de arquitectura y una especificación de requisitos de proyecto de arquitectura. Ademas, los componentes de la arquitectura de negocio en una

hoja de ruta de la arquitectura es parte de la salida de esta fase, que puede incluir catálogos, matrices y diagramas.

- **Arquitecturas de Sistemas de Información**

Los objetivos de la fase de Sistemas de Información Arquitecturas son desarrollar la arquitectura de sistemas de información de destino, para los datos y el nivel de aplicación, que describe cómo la arquitectura de sistemas de información de la empresa permitirá a la arquitectura empresarial y la visión de arquitectura, de manera que se ocupa de la solicitud de arquitectura de trabajo y preocupaciones de los interesados. Además, se identifican también las diferencias entre información de línea base y de destino de los sistemas de arquitecturas.

A partir del modelo se definen vistas diferentes de la arquitectura de software:

- Vista lógica, corresponde a la abstracción del sistema.
- Vista de procesos, procesos de ejecución.
- Vista física, mapeo del software sobre el hardware.

A partir de lo anterior las vistas que se desarrollaran son las vistas de aplicaciones por procesos La entrada para esta fase es la salida de la fase de arquitectura de negocio, acompañada de los principios de datos y aplicaciones.

La salida para esta fase son refinados y versiones actualizadas de las prestaciones anteriores, los resultados del análisis de la brecha entre la línea de base y la arquitectura de destino y los componentes de los sistemas de información de una hoja de ruta de la arquitectura.

- **Arquitectura tecnológica**

Objetivos de la fase de Tecnología son desarrolla la arquitectura tecnológica de destino que permite a las aplicaciones y datos de los elementos físicos y lógicos, y la visión de arquitectura desplegarse, e identificar los componentes de candidatos en base a diferencias entre la línea

de base y la arquitectura objetivo. Las vistas que se utilizaran para este dominio son:

- Redes y telecomunicaciones
- Middleware
- Infraestructura de Servidores

- **Oportunidades y soluciones**

Fase E se centra en la generación de la versión inicial completa de la arquitectura de hoja de ruta, en base a los elementos de la hoja de ruta de análisis de brecha y arquitectura candidato de fases B, C y D.

- **Planeamiento de migración**

Planificación de la migración es considerada con la finalización de la arquitectura hoja de ruta y el plan de implementación y migración de soporte. Estos planes también tienen que ser coordinadas con el enfoque de la empresa para la gestión y aplicación de los cambios en la cartera total.

- **Implementación de gobernanza**

Esta fase garantiza la conformidad con la arquitectura de destino por los proyectos de implementación y realiza funciones de gobernanza de la arquitectura de la solución y las solicitudes de cambio de aplicación impulsada.

- **Administración de arquitectura del cambio**

La fase de cierre del bucle se asegura de que se mantiene la arquitectura del ciclo de vida, que se ejecuta el marco de gobierno arquitectura y que las capacidades cumplen con los requisitos.

Dentro de TOGAF, se reconoce el papel de la seguridad, pero no se ha completado. El trabajo de la arquitectura empresarial y profesional de la seguridad a menudo se separa mientras que necesita ser integrado plenamente en ella. Una arquitectura de seguridad tiene su propia metodología de seguridad discreta, puntos de vista y opiniones, a menudo se

hace cargo de los flujos no normativas a través de sistemas y entre las aplicaciones y pide su propio conjunto único de habilidades y competencias de la empresa y arquitectos de TI. TOGAF ADM menciona algunos aspectos de seguridad por fase, que serán discutidos en la sección de integración (The Open Group, 2013).

b) ZACHMAN

Este framework fue creado por John Zachman, basado en principios de arquitectura e ingeniería comunes. El *framework* Zachman es un framework muy conocido para elaborar y definir una arquitectura empresarial. Sirve como apoyo para desarrollar los sistemas organizacionales actuales que son muy complejos.

Este framework posee una guía que permite, esta guía bidimensional permite la clasificación y organización de los elementos de una empresa, los cuales son importantes para la gestión de la empresa como para el soporte de los sistemas de información. “El eje vertical proporciona múltiples perspectivas de toda la arquitectura y el eje horizontal una clasificación de los diferentes artefactos de la misma empresa” (Maya, 2010).

El framework Zachman no tiene un proceso definido de implementación, debido a que se centran en garantizar que todos los elementos de la empresa este organizados adecuadamente, que sus relaciones permitan un sistema completo dejando de lado el orden del cual proceden. Además, el framework comprende aspectos particulares de los sistemas durante su evolución, lo cual puede ayudar en la toma de decisiones en los cambios como en las modificaciones.

> El *framework* contiene seis filas y seis columnas que producen 36 celdas o aspectos. Las filas son: alcance y planeador, modelo de negocio y propietario, modelo de sistema y diseñador, modelo de tecnología y constructor, componentes y contratista, sistema trabajando. Las columnas son: quién, cuándo, por qué, qué, cómo, dónde (Maya, 2010).

2.2.3 Grandes volúmenes de datos

En esta nueva era tecnológica, se están creando millones de bytes correspondientes a datos cada día, es así que más del 90% de los datos existentes fueron creados en la última década. Existen estudios que manifiestan que diariamente están siendo generados tantos datos digitales como toda la información escrita durante el curso de la historia humana, pues es así que presenciamos una era de información digital. Este avance tecnológico ha generado una tendencia hacia un enfoque nuevo hacia las decisiones y el entendimiento, que está siendo usada para generar los grandes volúmenes de datos, tanto estructurados, semiestructurados o no estructurados, el cual sería muy lento y caro al cargar dentro las bases de datos tradicionales (relacionales) para el análisis (Avila, 2011). Generalmente el termino *big data* o grandes volúmenes de datos es un conjunto de software, hardware y técnicas que se dedican al procesamiento, administración, clasificación y análisis de GVD de diversa procedencia. El termino GVD se utiliza para los datos que no pueden ser procesados o analizados por las herramientas y procesos convencionales.

> "El objetivo fundamental de los GVD es dotar de una infraestructura tecnológica a las empresas y organizaciones con la finalidad de poder almacenar, tratar y analizar de manera económica, rápida y flexible la gran cantidad de datos que se generan diariamente, para ello es necesario el desarrollo y la implantación tanto de hardware como de software específicos que gestionen esta explosión de datos con el objetivo de extraer valor para obtener información útil para nuestros objetivos o negocios" (Aguilar, 2018).

Los GVD tienen grandes retos que enfrentar, tales son la captura, almacén, búsqueda, divulgación, análisis y visualización.

2.2.3.1 Características de GVD

GVD genera complejidad y crecimiento con respecto a diferentes aspectos de la manipulación de datos. Stonebraker introduce grandes datos de acuerdo con el modelo de "3Vs" (Stonebraker, 2012). *Big data* se centra en los conjuntos de datos que son de un gran volumen, necesitan gran

velocidad, o exhiben gran variedad. Se puede definir la tecnología GVD con sus 3 dimensiones generalmente conocidas como las 3v:

a) Volumen

La tecnología de GVD puede hacerse cargo de la gestión de una vasta cantidad de datos generados cada dia por las organizaciones y empresas a nivel mundial. Hoy en día, los tamaños de datos están aumentando de manera exponencial, varias organizaciones están experimentando el fenómeno de diluvio de datos debido a múltiples factores, como los datos almacenados recopilados a lo largo de los años, la transmisión e intercambio de datos por medios de comunicación social, el aumento de los sensores de los datos recogidos, etc. Stonebraker considera el gran volumen tan importante y desafiante para dos tipos de análisis: "pequeños análisis" y "grandes análisis"; los análisis pequeños incluyen las operaciones más pequeñas, como ejecutar análisis SQL (recuento, suma, máx., min y promedio), mientras que los grandes análisis son operaciones más complejas que pueden ser muy costosos en grandes bases de datos, tales como clustering, regresiones y aprendizaje automático.

b) Variedad

Los datos vienen en diversos formatos: desde datos de texto a vídeo y audio, de DBMS tradicionales (sistemas de gestión de base de datos), de formatos de datos semi-estructurados (por ejemplo, XML), de los objetos binarios grandes. El principal desafío es la integración de todos estos tipos de datos, y gestionarlos de una manera eficiente. SAS estima que el 80% de los datos de la organización no son numéricos. Sin embargo, esos datos deben ser incluidos en el análisis y la toma de decisiones.

> Los GVD han de tener la capacidad de combinar una gran variedad de información digital en los diferentes formatos en las que se puedan presentar ya sean en formato video, audio o texto. Diferentes fuentes de información como las nuevas tecnologías wearables que monitorizaran nuestra actividad física, el internet de las cosas que conectará los dispositivos y máquinas entre sí, millones de mensajes escritos en redes sociales como Facebook, Instagram o Twitter, millones de videos que son cargados en

> Youtube diariamente, son algunos ejemplos de la generación de diversos tipos y generos de datos (Aguilar, 2018).

c) Velocidad

Para muchas organizaciones, este aspecto es el más difícil de GVD no es exclusivamente el volumen grande. Es más bien, qué tan rápido se procesan los datos para satisfacer las demandas. Una amplia gama de aplicaciones requiere un procesamiento rápido de datos en tiempo real. Por ejemplo, el comercio electrónico, las etiquetas RFID y los contadores inteligentes, la colocación de anuncios en las páginas Web, en forma en esta clase de aplicaciones.

> "La tecnología GVD ha de ser capaz de almacenar y trabajar en tiempo real con las fuentes generadoras de información como sensores, cámaras de videos, redes sociales, blogs, páginas webs; son fuentes que generan millones y millones de datos al segundo, por otro lado, la capacidad de análisis de dichos datos han de ser rápidos reduciendo los largos tiempos de procesamiento que presentaban las herramientas tradicionales de análisis" (Aguilar, 2018).

Además de los anteriormente mencionados 3 dimensiones de grandes volúmenes de datos en forma de volumen, velocidad y variedad, SAS define otras dimensiones complementarias.

- **Veracidad:** GVD debe realizar los análisis de manera inteligente a esta cantidad de datos con el objetivo de obtener información confiable que colabore en la toma de decisiones en las empresas.

- **Gran variabilidad:** Además de la velocidad y la variedad, los datos pueden exhibir grande Variabilidad en la forma de datos inconsistentes o no regular los flujos con picos periódicos. En particular, este es un comportamiento típico para aplicaciones como las redes sociales que de pronto podría experimentar cargas elevadas durante actualización de noticias o con eventos de temporada. En este contexto, la variabilidad del *Big data* puede ser particularmente difícil de manejar, por ejemplo, cuando los medios de comunicación social participan.

- **Gran complejidad:** La gestión de grandes volúmenes de datos que provienen de fuentes y sistemas múltiples es la mayoría de las veces difícil. Aspectos de gestión pueden ser particularmente desafiante para la vinculación, a juego, la limpieza y la transformación de datos a través de sistemas y plataformas.

Muchas organizaciones y empresas a nivel mundial diariamente alamacenan y gestionan este vasta cantidad de datos e informacion generada, los GVD es aplicado indistintamente en varios dominios, tanto en el sector privado, centros de investigación científica, tanto el sector estatal a nivel de ciudad o de pais, tanto asi que consideran que esta tecnologia es lamejor opcion en el momento de tomar las decisiones e influye positivamene en los resultados (Aguilar, 2018).

El gobierno utiliza los GVD para brindar mejores servicios a la población, por ejemplo, usando aplicaciones como Google maps, se analizan el tráfico y se localizan las congestiones de modo que se pueda predecir cuándo habrá una congestión, luego como medida se cambiaran las señales de tránsito de modo que disminuya la congestión.

En el sector industrial, los GVD son muy útiles en la mejora de procesos, por ejemplo, con el uso de sensores, se pueden registrar datos del estado de todas las maquinas, de modo que el mantenimiento se realice de una manera más efectiva, en menos tiempo y rediciendo costos. Además, las aplicaciones pueden variar desde análisis del clima, bolsa de valores, predecir tendencias en el comercio, predecir poblaciones, etc siendo todos estos una muestra de grandes avances en tecnología.

En el área biológica, el deseos de encontrar explicaciones a estudios de biología molecular y secuencias de ADN, se encuentra con muchos retos al lidiar con estos GVD, que crecer exponencialmente y complejamente.

2.2.3.2 Datos biológicos: ADN y ARN

El ácido desoxirribonucleico (ADN), ácido ribonucleico (ARN) y proteínas son tres moléculas constituyentes de la vida. Cada una se especializa en realizar una función específica. La información codificada dentro del ADN otorga las características hereditarias del ser tales como el color del cabello, el color de los ojos, así como la predisposición a ciertas enfermedades. El ADN está compuesto por cuatro moléculas: Adenina, Timina, Citosina y Guanina, denominadas en forma general nucleótidos y codificadas con las letras A T C y G. Por otro lado, el ARN tiene una gran variedad de roles, son los transmisores de la información entre el ADN y las proteínas. A su vez, las proteínas son cadenas formadas por 20 aminoácidos diferentes. La traducción en mecanismo biológico que permite relacionar la secuencia de nucleótidos y la secuencia de aminoácidos de la proteína, conocido de forma general como código genético. En el código genético se encuentran las instrucciones contenidas en un gen - segmento de ADN - que le dicen a la célula cómo hacer una proteína específica.

ARN es un ácido nucleico que está compuesto por una cadena de ribo nucleótidos. Existe en las células procariotas como en las eucariotas, y es el único material genético de ciertos virus (virus ARN). El ARN celular es lineal y de hebra sencilla, pero en el genoma de algunos virus es de doble hebra.

Las proteínas son las responsables de casi todas las funciones de un ser vivo y de la formación de muchas estructuras vivientes. De hecho, cada proteína tiene un ordenamiento o secuencia de aminoácidos que es exclusivo en ella y diferente a todas las demás. La secuencia de aminos que caracteriza a una proteína se denomina estructura primaria. Esta estructura es muy importante ya que de ella depende la función que desempeña en un organismo. Sin embargo, la complejidad de las moléculas de proteínas no termina aquí. Una vez establecida la cadena de aminoácidos, su estructura primaria, otras interacciones entre los aminoácidos hacen que adopten configuraciones espaciales diversas denominadas estructuras secundarias.

Además, existen plegamientos sobre sí misma que constituye la estructura terciaria que le da un aspecto de ovillo. Varios ovillos a su vez pueden asociarse y formar una única molécula, la estructura cuaternaria.

2.2.3.3 Generación de grandes volúmenes de datos biológicos

> Al empezar este nuevo milenio, secuenciar 1 Mbp (millón de pares de bases), que son las "letras" de un genoma de ADN, la molécula de la vida, costaba alrededor de 10 mil dólares. Pocos años más tarde, en 2008, secuenciar el genoma humano, con algo más de 3.200 millones de letras, ya era posible en unas seis semanas y su coste total caía hasta los 60 mil dólares. Por aquel tiempo los proyectos especulaban con alcanzar el genoma humano a unos mil dólares en los siguientes tres años. En efecto, en octubre de 2009 varias compañías apuntaron al genoma de los mil dólares con un nuevo avance tecnológico, la secuenciación con nano poro y ya en marzo de 2011, el coste había caído a menos de medio dólar por Mbp. Como referencia, en 1990, el Congreso de los Estados Unidos de América aceptó financiar con 3 billones de dólares el proyecto Genoma Humano, un dólar por base. El genoma se completó en 2003, dos años antes de lo previsto, costando "solo" 2,7 billones de dólares, lo que representa 2,7 millones de veces el coste actual. En febrero de 2012 se lanzó un dispositivo en miniatura del tamaño de una memoria USB, diseñado para hacer de la secuenciación de ADN una tecnología universalmente accesible por menos de 900 dólares (Trelles, 2013).

Actualmente, los investigadores siguen tratando de desarrollar nuevos procedimientos más rápidos y menos costos que faciliten la secuenciación del genoma de cada persona, donde países como USA esto es algo normal, es así que esto generaría el inicio de la medicina personalizada basada en la genética. El genoma humano no solo es el objetivo de los científicos. Lo expuesto anteriormente se puede aplicar a todas las demás especies, plantas, animales, etc. Actualmente, existen muchas especies en estudio de todos los reinos, los cuales tienen una secuencia genómica, también, existen cientos de proyectos en marcha para obtener estos datos. La genómica comparativa llama la atención para la creación de nuevos métodos que estudien las relaciones existentes de estas especies en la generación e interpretación de todos estos datos.

Los grandes volúmenes de datos y proyección de que estos datos seguirán creciendo y además enfrentan varios retos. La compartición de estos GVD y su movimiento es el primero. Las redes de comunicación mueven pequeñas cantidades de datos, pero cuando los datos son voluminosos se experimenta muchas demoras cuando los ficheros descargador por internet que demoran horas en transferir. Esto hablando en el uso de internet, pero asi mismo en la intranet, mover de un servidor a otro es lento y caro.

La transferencia de datos es importante, pero la parte más importante es su procesamiento y las tecnologías necesarias para ayudar este importante aspecto. “Pensar en grande” es un nuevo reto, ya que al formular proyectos en las ciencias biológicas, se tiene que pensar tanto en la secuenciación tradicional como el comportamiento de los organismo en su ambiente, o también pensar el caso de los meta genomas, el cual incluye la colección de muchos datos a nivel genético, celular, etc,

Los datos provienen de las nuevas técnicas de investigación en biología molecular como el secuenciamiento a gran escala (*Next Generation Sequencing*, NGS). Estas tecnologías NGS actualmente incluyen Illumina Genome Analyzer (GA), Applied Biosystems (ABI) Sólido, Roche de 454, máquinas de secuenciación Helicos ‘Heliscope’ (Lee, et al., 2013). Y la reciente tecnología PacBio y 10X Genomics.

2.2.3.4 Formatos típicos de GVD

Las descripciones de estos formatos de archivos contienen información con respecto a la salida típica de plataformas de la nueva generación de secuenciación y pueden no incluir todas las informaciones u opciones para cada formato de archivo.

a) FASTA

Es un archivo de texto que contiene secuencias de nucleótidos representadas por símbolos de la IUPAC (por ejemplo, A, C, G, T, N). Cada *read* proveniente de una plataforma de secuenciación será representada en

este archivo por dos líneas de texto. La primera línea es la cabecera que identifica de forma única del *read*. Cada única entrada siempre comienza con un ">" directamente seguido de un identificador único (Figura 10). Además, esta línea puede contener coordenadas específicas para una ejecución única en una determinada plataforma de secuenciación y la longitud de la *read*. Hay típicamente una línea que sigue la línea de cabecera, que contiene la secuencia de nucleótidos de dicho *read*. En algunos casos, puede haber más de una línea de secuencia debido a la forma en que ciertos programas escriben estos archivos. Debido al tamaño de los archivos (números de *reads* que contienen), no es recomendable para abrirlos con una interfaz. Otra extensión de archivo similar es *fna*, un archivo de salida desde la plataforma de Roche GS FLX / 454. Un ejemplo de dos cortos *reads* en un archivo fasta puede verse a continuación.

Figura 10: Formato FASTA

```
> HWSW5CG01CQQSC rank=0005822 x=1007.0 y=2442.5 length=41
TGCTATATTCTATATGCTATATGCTATATGATATATGCTAA

> HWSW5CG01ESA06 rank=0005835 x=1845.0 y=2120.0 length=43
GATATCGTGATTGCCAATCAAGTCTTGTACGTATGATCAAACC
```

Fuente: Elaboración del autor

b) FASTQ

Al igual que fasta, este archivo contiene un encabezado seguido por una línea de secuencia a continuación por una cabecera de calidad y otra línea de métricas de calidad. La línea de encabezamiento de secuencia siempre comienza con una "@" seguido directamente por un identificador único y puede contener información adicional. La cabecera de la calidad siempre empieza con "+" y puede contener información adicional. Al igual que fasta, normalmente hay una línea para cada cabecera, la secuencia y las métricas de calidad, pero más de una línea puede aparecer por secuencia y métricas de calidad. La línea (s) que contiene las métricas de calidad puede parecer diferente de cada secuenciador. Métricas típicas de calidad son de carácter codificado y parecen bastante complejo. Un ejemplo de dos cortos *reads* en un archivo con extensión Fastq se muestra en la Figura 11.

Figura 11: Formato FASTQ

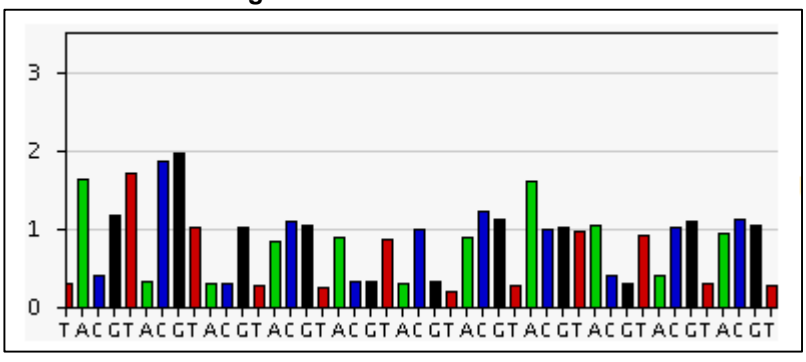
```
@ K5HV3: 00029: 00029
AAAAAGGGTAAAAACGATCGCTCACAGGTACGTATGATCAAACC
+
AB >> (44 * 44 ::: /:? 447444C765 @ - * 44 ::: /:? 447444C765 * *

@ K5HV3: 00030: 00032
TAAAAAAGGTACGTATGATCAAACCTACGTATGATCAAACCTAT
+
77EECB * @ 4 * 44 ::: /:? 447444C765 * 44 ::: /:? 447444C765 /
```

Fuente: Elaboración del autor

c) SFF

Standard Flowgram Format es un archivo binario que contiene información de la secuencia, la calidad y el flujo de nucleótidos. Este archivo proviene de la salida de Roche GS FLX/454 y las plataformas de Ion Torrent PGM. La secuencia de nucleótidos que contiene es "ligeramente depurada" lo que significa que la secuencia recortada (por ejemplo, adaptadores y secuencia de mala calidad) todavía está anotada. Las métricas de calidad para cada nucleótido en la secuencia también se anotan. Información sobre el flujo de nucleótidos se almacena como la señal relativa dada por la incorporación de nucleótidos. Una representación gráfica de flujo de nucleótidos se puede ver en la Figura 12. Donde el eje x son los nucleótidos secuenciados y el eje y es la intensidad de la señal.

Figura 12: Formato SFF



Fuente: Elaboración del autor

d) BAM

Una versión binaria del formato SAM (*Sequence Alignment Map*) que contiene la secuencia y los niveles de calidad de nucleótidos. Los archivos

pueden contener alineados (basado en una referencia) o no alineados *reads*. La compresión de este formato permite una gran cantidad de datos que se almacena en un archivo que ocupa relativamente poco espacio en el disco duro. Además, el archivo también puede almacenar metadatos (por ejemplo, *flags* para *paired reads*, QC), que ofrece una clara ventaja sobre fastq.

2.2.3.5 Infraestructuras para grandes volúmenes de datos

El fenomenal crecimiento de grandes volúmenes de datos requiere infraestructuras eficientes y paradigmas nuevos de computación con el fin de cumplir con los enormes volúmenes de datos y su gran velocidad. De aquí en adelante, se presentan tres modelos comunes de infraestructura que ofrecen excelentes medios para gestionar grandes volúmenes de datos.

a) Clusters

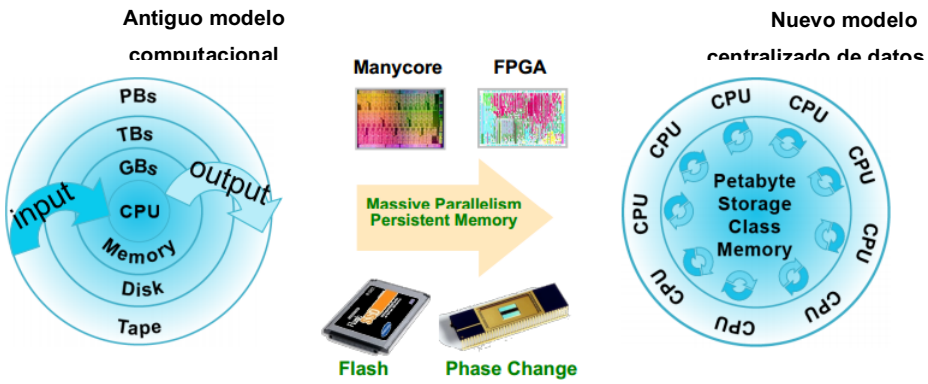
Un clúster de computación consiste en conectar un conjunto de equipos con un software regular (sistemas operativos), así como a través de las redes de área local (LAN) para construir un sistema global. El clúster de computación fue una de las inspiraciones principales para la computación distribuida eficiente. El objetivo principal de este paradigma de computación es proporcionar un mejor rendimiento y disponibilidad a un bajo costo.

Con los años, las herramientas informáticas de clúster y paradigmas evolucionaron hasta llegar a una alta eficiencia. Hoy en día, la construcción de un clúster se ha convertido en fácil con la ayuda para la planificación eficiente de tareas, y balanceo de carga. Por otra parte, los usuarios del clúster pueden confiar en herramientas estandarizadas y dedicadas, como los sistemas operativos Linux, *Message Passing Interface* (MPI) herramientas, y diversas herramientas de monitoreo para ejecutar una amplia gama de aplicaciones. En tal sentido, se ha pasado de un antiguo modelo de computación central a un nuevo modelo de paralelización en clúster para datos centralizados (Figura 13). Estos tamaños de clúster pueden variar desde decenas de nodos a decenas de miles de nodos.

En la era de los grandes datos, es común que las organizaciones creen clústeres dedicados que por lo general consisten en hardware comercial. Estos clústeres ejecutan las plataformas de *Big data* (como los sistemas de Hadoop y No SQL), para alojar de manera espectacular los grandes volúmenes de datos de manera eficiente. Permiten mientras tanto, el procesamiento en tiempo real y de alto grado de disponibilidad.

En el ámbito de la informática de alto rendimiento (HPC), supercomputadoras, que consisten en cientos de miles de núcleos pueden ejecutar aplicaciones de alto rendimiento que exhiben altas demandas de E/S. Estas supercomputadoras han dedicado arquitecturas masivamente paralelas para satisfacer las altas exigencias de cálculo y las crecientes necesidades de almacenamiento de HPC.

Figura 13: Nuevo modelo de computación en clúster



Fuente: Adaptado de (Foster,2003)

b) Grids

El *Grid Computing* es una infraestructura de hardware y software que proporciona acceso fiable, coherente, penetrante, y de bajo costo a las capacidades computacionales de alta gama. Ian Foster reformula su definición como una lista de control. Asi, un grid es un sistema que:

- Proporciona la coordinación de los recursos para los usuarios que no tienen un control centralizado.
- Proporciona, protocolos e interfaces abiertos de uso general.
- Entregar cualidades no triviales de servicio.

La coordinación de la sub-organizaciones grid computing y recursos heterogéneos podrían estar dispersos en áreas geográficamente remotas. Su objetivo es ocultar la complejidad de los sub-sistemas de usuarios y los expone a la ilusión de un sistema global en su lugar.

c) Clouds

La computación paralela y distribuida ha ido evolucionando en los últimos años con el fin de proporcionar una mejor calidad de servicio a los usuarios al tiempo que asegura la eficiencia en términos de rendimiento, el costo, y la tolerancia a fallos. Un paradigma de computación recién emergida es el *cloud computing*. En este paradigma, el software y hardware se entregan como un servicio a través de una red informática, típicamente Internet. Servicios en la nube difieren en su nivel de abstracción. Se pueden clasificar en tres niveles: infraestructura, plataforma y software.

El *cloud computing* de inmediato ha sido adoptado por las aplicaciones de *Big data*. Para muchas organizaciones, la adquisición de recursos de una manera *Pay-as-You-Go* con el fin de ampliar sus plataformas de grandes volúmenes de datos cuando sea necesario, fue una necesidad de mucho tiempo, y los proveedores de nube ofrecen precisamente eso. Por otra parte, la mayoría de los proveedores de nube ofrecen las plataformas de grandes datos como un servicio. Ellos permiten a los clientes ejecutar sus aplicaciones sin tener que preocuparse acerca de la gestión de la infraestructura y su mantenimiento costoso. Hoy en día, es natural para las empresas emergentes, como Netflix que ofrece vídeo a la carta de servicios, alojando y gestionando todos sus datos en la nube (Amazon EC2). Más adelante se destaca aún más este paradigma de *cloud computing* como un medio natural y excelente para GVD.

2.2.4 Cloud computing

La informática en la nube o *cloud computing*, es un tema bastante trillado en estos días y con mucha razón, representa un nuevo punto de inflexión comparado con la computación en red, ofrece mayor eficiencia,

escalabilidad masiva y desarrollo más rápido y más fácil del software. Se trata de nuevos modelos de programación, nueva infraestructura de TI y la habilitación de nuevos modelos de negocio. Según el Instituto Estadounidense de Estándares y Tecnología (NIST, por sus siglas en inglés), *cloud computing* se define como un "modelo de acceso por petición a un conjunto compartido de recursos informáticos configurables (por ejemplo, redes, servidores, soportes de almacenamiento, aplicaciones y servicios) que puede activarse y desactivarse con una gestión e interacción del proveedor de servicios mínimas".

Cloud computing (Gartner Group, n.d.) "es un estilo de computación en el que las capacidades de TI escalables y elásticas se entregan como un servicio utilizando las tecnologías de Internet". Existe una modificación con respecto a la definición original de 2008: se sustituye la cláusula "escalables masivamente" por "flexibles y escalables" para calificar las capacidades entregadas. Una empresa puede requerir recursos masivos de computación en un momento dado, también puede requerir reducir sus requisitos; luego, la flexibilidad es un punto escencial de la computación en la nube.

La definición (Froilan, 2010), describe el *cloud computing* en cinco características que, al combinarse de varias maneras, resultan en diferentes modelos de computación en la nube. Gartner sugiere que no se debe enfocar en una característica solamente. Estas características son las siguientes:

- Basado en servicios. Lo cual requiere que el proveedor brinde una completa automatización en sus servicios. Esto requiere que tenga una interfaz sencilla que opaque la complejidad de la implementación.
- Escalabilidad y flexibilidad. El servicio debe ser escalable tanto horizontalmente como verticalmente, esta escalabilidad debe realizarse rápidamente (inclusive segundos es requerido).
- Compartido, El servicio debe permitir la compartición de recursos, que requieren ser el uso con la máxima eficiencia.
- De acuerdo al uso. El servicio debe tener varios planes de acuerdo a las necesidades del usuario, donde cual el pago depende del consumo.

- Internet. Los servicios deben estar siempre disponible a través de internet.

Los servicios de *cloud computing* aumentaron su importancia como soluciones económicas para llevar a cabo análisis a gran escala en una función de las necesidades, en particular para los laboratorios de medianos y pequeños de centros de investigación que no pueden pagar el costo de comprar y mantener una infraestructura suficientemente poderosa. Un modelo computacional en el *cloud computing* es ideal para el análisis de datos de secuencias a gran escala. En este modelo, la computación y almacenamiento existen como recursos virtuales en los centros de datos remotos y pueden ser asignados y se liberan de forma dinámica según sea necesario.

Los avances en biología y las tecnologías de la información traen profundas influencias en la bioinformática, debido a su naturaleza interdisciplinaria. Por esta razón, la bioinformática está experimentando una nueva tendencia en el análisis de manera que la computación está pasando de la infraestructura informática in house hacia el *cloud computing*. Esto ha sido necesario con el fin de manejar las enormes cantidades de datos biológicos generados por las tecnologías experimentales de alto rendimiento, donde el *cloud computing* aborda el almacenamiento de datos grandes y temas de análisis en muchos campos de la bioinformática. Los datos que son demasiado grandes para ser procesadas convencionalmente también son demasiado grandes para ser transportados en cualquier lugar. Entonces según una estrategia de prioridades, el código para realizar el análisis tiene que ser movido, no los datos. El *cloud computing* posee las facilidades de una tecnología para lograr este resultado.

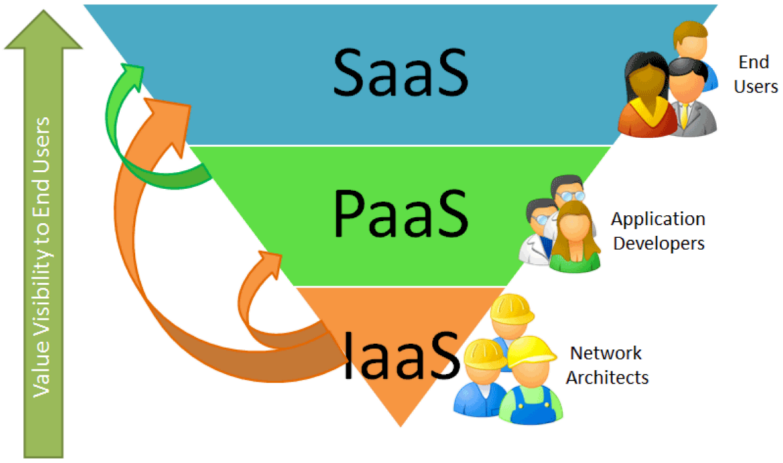
Los investigadores y los laboratorios experimentan la disponibilidad esporádica de datos, donde pueden tener una necesidad de, por un período de tiempo, procesar enormes cantidades variables de datos. Dicha variabilidad del volumen de datos impone requisitos variables sobre disponibilidad de recursos informáticos que se utilizan para procesar los datos dados. En lugar de tener que comprar y mantener los recursos informáticos

deseados o tener que esperar mucho tiempo para el procesamiento de datos se plantea ejecutar en infraestructuras de *cloud computing*.

2.2.4.1 Arquitectura del *cloud computing*

La arquitectura del *cloud computing* esta basado en el hardware, plataforma y aplicaciones, como en las siguientes capas (Figura 14).

Figura 14: Arquitectura del cloud computing



Fuente: (Lindsey, 2010)

a) Software como servicio (SaaS)

(Avila, 2011) “SaaS se encuentra en la capa más alta y consiste en la entrega de aplicaciones completas como un servicio”. El proveedor del servicio ofrece SaaS como una aplicación encargada de procesamiento y mantenimiento. Esta aplicación tiene un modelo para distribuir de uno a muchos, donde el cliente recibe las salidas de las aplicaciones. El proveedor del servicio se encarga de que el servicio siempre esté disponible y en funcionamiento, de acuerdo a las necesidades de los clientes.

b) Plataforma como servicio (PaaS)

PaaS (Avila, 2011) “se centra en un modelo en el que se proporciona un servicio de plataforma con todo lo necesario para dar soporte al ciclo de planteamiento, desarrollo y puesta en marcha de aplicaciones y servicios web a través de la misma”. El servicio del proveedor se encarga del

escalamiento de los recursos de acuerdo a la demanda lo requiera para tener un buen rendimiento en el servicio brindado, así también sea seguro, etc.

El cliente está enfocado en el desarrollo, depuración y pruebas de las herramientas en uso para el software funcione adecuadamente a través de internet, con lo cual el cliente no se preocupa del hardware, solo se preocupa de aumentar la productividad de su producto. Por ejemplo, Google Apps permite a sus usuarios utilizar su hardware y permitir que ellos desarrollen, alojen y compartan sus aplicaciones web. En el futuro probablemente se reemplace los alojamientos de datos de las empresas, los administradores de sistemas, ya que esta responsabilidad pasara a ser parte de la responsabilidad del proveedor.

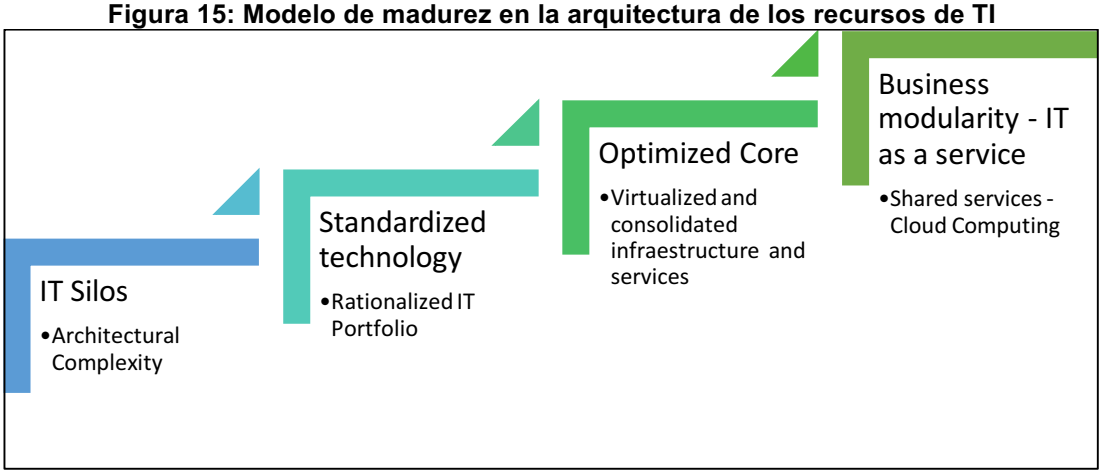
c) Infraestructura como servicio (IaaS)

IaaS corresponde a la capa más baja según (Avila, 2011). “IaaS hace uso externo de servidores para espacio en disco, base de datos, ruteadores, switches, así como tiempo de cómputo evitando de esta manera tener un servidor local y toda la infraestructura necesaria para la conectividad y mantenimiento dentro de una organización”. En este modelo el pago corresponde solo al uso de la infraestructura como espacio usado en disco, CPUs, transferencia de datos, etc.

IaaS posee una mayor flexibilidad que PaaS, debido a que brinda mayor flexibilidad, y el cliente se hará cargo de instalar, configurar y mantener el software de la infraestructura. Si los proyectos que salen del alcance de los PaaS, y se quiere ir mas allá, la opción a seguir es contratar una IaaS (Figura 15).

Además, IaaS permite a los sistemas ser escalables automáticamente o semiautomáticamente de manera que los servicios se ofrecen según los requerimientos del cliente. Un ejemplo de almacenamiento a media es DropBox y skyDrive, en los cuales solo se paga de acuerdo al espacio requerido.

Figura 15: Modelo de madurez en la arquitectura de los recursos de TI



Fuente: Adaptado de (Kuthanur, 2015)

2.2.4.2 Modelos de despliegue del *cloud computing*

Existen diferentes tipos o modelos de *cloud computing*, de acuerdo al autor se puede tener algunos modelos intermedios o híbridos (Tabla 1), pero en general se mencionan 3 diferentes modelos:

Tabla 1: Modelos de cloud computing

	Nube pública	**Nube alojada**	**Nube privada**
Ubicación	Centro de datos conectados por internet	Centro de datos conectados por internet	Centro de datos interno
Tenencia	Múltiples clientes	Múltiples clientes	Una empresa
Infraestructura	Compartida	Alojada en un centro de datos (totalmente dedicada o compartida parcialmente)	Dedicada
Seguridad	Común a todos los clientes, con limitadas configuraciones	Común a todos los clientes, con mayor configuración	Única del cliente
Gestión de la nube	Proveedor	Proveedor o departamento de TI	Departamento de TI
Gestión de la infraestructura	Proveedor	Proveedor	Departamento de TI
Facturación	Por consumo	Pago mensual por infraestructura dedicada, acceso facturado por consumo	Medidas por consumo a cada unidad

Fuente: (Forrester Research, 2009)

Algunos autores mencionan también el modelo comunitario, que es cuando la infraestructura es compartida entre varias empresas o comunidades de modo que tienen objetivos en común (requisitos de seguridad, infraestructura requerida, políticas, necesidades, etc). Este modelo puede ser gestionado por ellos mismo o contratar a alguien externo.

2.2.4.3 Ventajas del *cloud computing*

Cuando los recursos de infraestructura son más abundantes, la informática trabaja de diferente forma, se pueden desplegar soluciones más rápidas, permiten una mejor gestión, son más flexibles y están más disponibles, Esta accesibilidad es importante que esté al alcance de los recursos de TI, y esto es lo que ofrece el cloud computing. Se han definido las siguientes ventajas para el cloud computing (Rosales Rosero, 2010):

- Agilidad en el despliegue de las aplicaciones con menos recursos de la empresa.
- Menores inversiones en TI.
- Flexibilidad.
- Satisfacción de las necesidades de la empresa mediante estas aplicaciones o servicios.
- Compartición de datos y procesos que mejora la participación de los involucrados.
- Eficiencia al desplegar servicios procesos de negocio.

> A pesar de que el mercado de cloud computing para las empresas se está desarrollando a gran velocidad. Los retos que plantean la seguridad de los datos, la privacidad y la normativa retrasarán la implementación en algunos casos. Este tipo de tecnología se ha desplegado en aquellas áreas con un bajo riesgo percibido, como el correo electrónico, las copias de seguridad y recuperación, el procesamiento por lotes, y el análisis y control de calidad. Otras aplicaciones cloud emergentes incluyen la gestión de las relaciones con los clientes, los flujos de trabajo, las comunicaciones unificadas y la planificación sencilla de recursos empresariales. Las aplicaciones más sofisticadas se irán desplegando a medida que el mercado madure. Por ejemplo, en el caso de la investigación y el desarrollo en las industrias farmacéuticas, de automoción y en otras se está desplazando cada vez en mayor medida hacia la nube (Rosales Rosero, 2010).

2.2.4.4 *Cloud computing* y los GVD

GVD y *cloud*, dos de las tendencias que están definiendo la informática emergente, muestran un gran potencial para una nueva era de aplicaciones combinadas. La provisión de capacidades analíticas de grandes volúmenes de datos utiliza modelos de prestación de *cloud computing* que podrían facilitar la adopción de muchas empresas, y que además de los ahorros de costes importantes, podría simplificar ideas útiles que podría proporcionarles diferentes tipos de ventaja competitiva.

2.3 Definiciones de términos básicos

- ADN: El ácido desoxirribonucleico, es un ácido nucleico que contiene instrucciones genéticas usadas en el desarrollo y funcionamiento de todos los organismos vivos conocidos y algunos virus, y es responsable de su transmisión hereditaria.
- Arquitectura informática: Es el diseño conceptual y la estructura operacional fundamental de un sistema de computadoras. Es decir, es un modelo y una descripción funcional de los requerimientos y las implementaciones de diseño para varias partes de una computadora.
- Bioinformática: Es la aplicación de tecnología de computadores a la gestión y análisis de datos biológicos.
- Bioinformático: Es una persona que utiliza y desarrolla herramientas de software bioinformáticas para analizar los datos de secuencias y estructuras moleculares y así responder preguntas de tipo biológico y/o encontrar nuevo conocimiento.
- Biología computacional: Sinónimo de bioinformática solo para efectos de esta tesis, dado que sus diferencias no alteran el curso de la misma.
- Computación de alto rendimiento (*High performance Computing* o HPC en inglés): Es una tecnología que se apoya en el uso de los clústeres, supercomputadores o mediante el uso de la computación paralela.
- *Cloud computing*: Es un conjunto de hardware, redes, almacenamiento, servicios e interfaces que permiten el suministro de la informática como un servicio. Los servicios en la nube se basan en la entrega de software, de infraestructura y de almacenamiento a través de internet.

- Datos crudos o sin procesar: *Raw data* en inglés. Es un término para los datos provenientes directamente de una fuente. Los datos crudos no han sido objeto de procesamiento o cualquier otra manipulación, y también se hace referencia a los datos primarios.
- Datos estructurados: Datos que tienen bien definidos su longitud y su formato, como las fechas, los números o las cadenas de caracteres. Se almacenan en tablas.
- Datos no estructurados: Datos en el formato tal y como fueron recolectados, carecen de un formato específico. No se pueden almacenar dentro de una tabla ya que no se puede desgranar su información a tipos básicos de datos.
- Genebank: Es la base de datos de secuencias genéticas del NIH (*National Institutes of Health* de Estados Unidos), una colección de disponibilidad pública de secuencias de ADN.
- Genoma: Es el conjunto de genes contenidos en los cromosomas, lo que puede interpretarse como la totalidad de la información genética que posee un organismo o una especie en particular.
- Infraestructura tecnológica: Es el conjunto de hardware (servidores, puestos de trabajo, redes, enlaces de telecomunicaciones, etc.) y software (sistemas operativos, bases de datos, lenguajes de programación, herramientas de administración, etc.) sobre el que se asientan los diferentes servicios que en conjunto dan soporte a las aplicaciones.
- *In silico*: Es una expresión que significa hecho por computadora o vía simulación computacional.
- *Next generation sequencing* (NGS): Término general utilizado para describir una serie de diferentes tecnologías de secuenciación modernas que permiten secuenciar ADN y ARN más rápida y barata anteriores tipo de secuenciación.
- Performance: Desempeño con respecto al rendimiento de una computadora, un dispositivo, un sistema operativo, un programa o una conexión a una red.
- Plataforma: Es un sistema que sirve como base para hacer funcionar determinados módulos de hardware o de software con los que es compatible.

Al definir plataformas se establecen los tipos de arquitectura, sistema operativo, lenguaje de programación o interfaz de usuario compatibles.

2.4 Hipótesis

2.4.1 Hipótesis general

Una adecuada arquitectura empresarial mejora el análisis bioinformático de grandes volúmenes de datos.

2.4.2 Hipótesis específicas

- Una adecuada vista de negocio brinda los lineamientos para mejorar los procesos de negocio bioinformáticos.
- Una adecuada vista de aplicaciones define los recursos disponibles para mejorar los trabajos de los investigadores.
- Una adecuada vista de datos sirve como base para mejorar la gestión de los GVD.
- Una adecuada vista de infraestructura permite identificar los recursos tecnológicos que optimizan los análisis bioinformáticos.

2.5 Variables

Variable dependiente:

- Optimización de los análisis de los grandes volúmenes de datos.

Variables independientes:

- Desarrollo de una arquitectura empresarial,
- Uso del *cloud computing*

2.6 Matriz de consistencia

En la Tabla 2 se observa la matriz de consistencia, cabe resaltar que en todos los indicadores el instrumento de medición es un cuestionario de encuesta.

Tabla 2: Matriz de consistencia

PROBLEMA GENERAL	OBJETIVO GENERAL	HIPÓTESIS GENERAL	VARIABLE	DIMENSIONES	INDICADORES
PP: ¿Cómo la arquitectura empresarial impacta en el análisis bioinformático de grandes volúmenes de datos?	OG: Mejorar los análisis bioinformáticos de grandes volúmenes de datos a través de una arquitectura empresarial.	HG: Una adecuada arquitectura empresarial mejora el análisis bioinformático de grandes volúmenes de datos.	Desarrollo de una arquitectura empresarial	Organización	Adaptabilidad a cambios Nivel de participación de los interesados
PROBLEMAS ESPECÍFICOS	**OBJETIVOS ESPECÍFICOS**	**HIPÓTESIS ESPECÍFICAS**			
PE1: ¿Cómo una vista de negocio impacta en los procesos bioinformáticos?	OE1: Mejorar los procesos bioinformáticos que se llevan a cabo en las investigaciones mediante una vista de negocio.	HE1: Una adecuada vista de negocio brinda los lineamientos para mejorar los procesos de negocio.		Procesos	Alineamiento de procesos al negocio. Nivel de comunicación. Alineamiento de TI al
PE2: ¿Cómo una vista de aplicaciones apoya el trabajo de los investigadores?	OE2: Implementar una vista de aplicaciones que permita responder satisfactoriamente al trabajo y necesidades de los investigadores.	HE2: Una adecuada vista de aplicaciones define los recursos disponibles para mejorar los trabajos de los investigadores.		Aplicaciones	Nivel de satisfacción con las aplicaciones. Disponibilidad de los sistemas. Utilidad
PE3: ¿Cómo una vista de datos impacta en el análisis de GVD?	OE3: Mejorar la gestión y análisis de los GVD provenientes de las investigaciones, a través de una	HE3: Una adecuada vista de datos sirve como base para mejorar la gestión de los GVD.		Datos	Disponibilidad de los datos. Satisfacción con la gestión de datos.
PE4: ¿Cómo una vista de infraestructura apoya los análisis bioinformáticos?	OE4: Proveer una infraestructura robusta que soporte los análisis bioinformáticos mediante una vista de infraestructura.	HE4: Una adecuada vista de infraestructura permite identificar los recursos tecnológicos que optimizan los análisis bioinformáticos.	Uso de *cloud computing*	Infraestructura	Continuidad del servicio Costos Nivel de desempeño

Fuente: Elaboración del autor

CAPÍTULO III
METODOLOGÍA DE DESARROLLO

3.1 Tipo de investigación

3.1.1 Tipo de investigación

En la presente investigación se utilizó un tipo de investigación proyectiva. Según Hurtado (2000), “consiste en la elaboración de una propuesta o de un modelo, como solución a un problema o necesidad de tipo práctico, en un área particular del conocimiento, a partir de un diagnóstico preciso de las necesidades del momento, los procesos explicativos o generadores involucrados y las tendencias futuras”. Efectivamente, la propuesta de un modelo de línea de investigación en el área de bioinformática de las ciencias biológicas permitirá orientar y organizar los procesos investigativos en la mencionada área.

3.1.2 Método de la investigación

El método utilizado en la investigación es el método lógico deductivo, el cual consiste en encontrar principios desconocidos, a partir de los conocidos. La modelación es el método que opera en forma práctica o teórica con un objeto, no en forma directa; en el modelo se revela la unidad de lo objetivo y lo subjetivo. Como primer paso se define el sistema que hay que modelar y los objetivos para el modelo, de manera que la definición del sistema supone generalmente dibujar las fronteras alrededor de lo que se modela, luego determinar las variables claves y la relación entre estas.

3.2 Diseño de investigación

Esta investigación es de carácter cuasi experimental, debido a que “se manipulan las variables independientes para ver su efecto y relación con la variable dependiente” (Hernández, Fernández, & Baptista, 2010). En este diseño los sujetos no son asignados al azar a los grupos; sino que son grupos intencionales. El diseño contempla la administración de la pre y post prueba al grupo, para evidenciar que se cumplan los objetivos de la investigación.

3.3 Población y muestra

En esta investigación, la población estará conformada por los investigadores bioinformáticos del Centro Internacional de la Papa. “Para determinar la muestra el procedimiento no es mecánico, ni con base en fórmulas de probabilidad, sino que dependen del proceso de toma de decisiones de una persona o de un grupo de personas” (Hernández, Fernández, & Baptista, 2010). Se realizó la elección de sujetos con ciertas características que se desean estudiar de acuerdo al problema, es por ello que la muestra es no probabilística intencional conformada por 9 expertos representativos que se ven beneficiados con la arquitectura empresarial para los análisis de datos genómicos, quienes validaran los resultados de esta investigación. En este caso particular la muestra es igual a la población.

3.4 Técnicas de recolección de datos

Para realizar la presente investigación se recurrirá a las técnicas de carácter primario, como secundario, siendo éstos los siguientes:

- Encuesta: Con el fin de obtener información de los implicados y especialistas, captar y comprender la situación problema.
- Observación científica: Para obtener la percepción directa del objeto de investigación, permite conocer la realidad mediante la percepción directa de la situación.
- Documentación: se procederá a la consulta bibliografía de textos relacionados al tema, revistas, tesis y otros documentos vinculados a la investigación.
- Modelamiento: Archimate software.

3.5 Técnicas para el procesamiento de la información

3.5.1 Tratamiento estadístico

"Las muestras no probabilísticas se caracterizan por el hecho de que no es posible determinar la probabilidad de inclusión de cada unidad elemental de la población en la muestra dado que la confiabilidad de los resultados de estas muestras no puede medirse, las muestras no probabilísticas no se prestan para tratamiento y análisis estadísticos" (Sampieri, 2010). Se realizará un procesamiento de datos informático básico para el tratamiento de datos, los tipos de análisis son del tipo descriptivos.

- Análisis descriptivos. Este tipo de análisis servirá para la recolección, orden, análisis y representación de grupos de datos, usando graficas que describan las características de este conjunto, este análisis suele ser básico, los cuales implican medidas de tendencia central y de dispersión.

3.5.2 Validación del instrumento de recolección de datos

El Coeficiente Alfa de Cronbach (Wikipedia, n.d.), "se trata de un índice de consistencia interna que toma valores entre 0 y 1 y que sirve para comprobar si el instrumento que se está evaluando recopila información defectuosa y por tanto nos llevaría a conclusiones equivocadas o si se trata de un instrumento fiable que hace mediciones estables y consistentes". Su interpretación será que, cuanto más se acerque el índice al extremo 1, mejor es la fiabilidad, considerando una fiabilidad respetable a partir de 0,80. La fórmula es la siguiente:

$$\alpha = \frac{K}{K-1}\left[1 - \frac{\Sigma S_i^2}{S_T^2}\right]$$

Donde:

K: El número de ítems

S_i^2: Sumatoria de Varianzas de los Ítems

S_T^2: Varianza de la suma de los Ítems

α: Coeficiente de Alfa de Cronbach

3.6 Prueba de hipótesis

Una hipótesis es aceptada a través del aporte de evidencia en su favor, que acredite la validez de la misma. Es decir, cuantas más investigaciones apoyen una hipótesis, más credibilidad tendrá. La presente investigación se valida con la prueba de hipótesis, así mismo con los resultados obtenidos en los casos de estudio descrito en el capítulo 5.

La prueba de hipótesis se realizó con la prueba de Wilcoxon. “Wilcoxon es una prueba no paramétrica para realizar análisis de dos muestras relacionadas con datos ordinales y determinar si existen diferencias entre ellas, por lo tanto, no necesita una distribución específica”. Además, el número de individuos de la muestra es inferior a treinta (Wikipedia, n.d.).

CAPÍTULO IV
ARQUITECTURA EMPRESARIAL PROPUESTA

Este capítulo describe el modelo de arquitectura empresarial. Se adoptó como marco metodológico de TOGAF que contempla la fase preliminar y las 4 primeras fases del ADM.

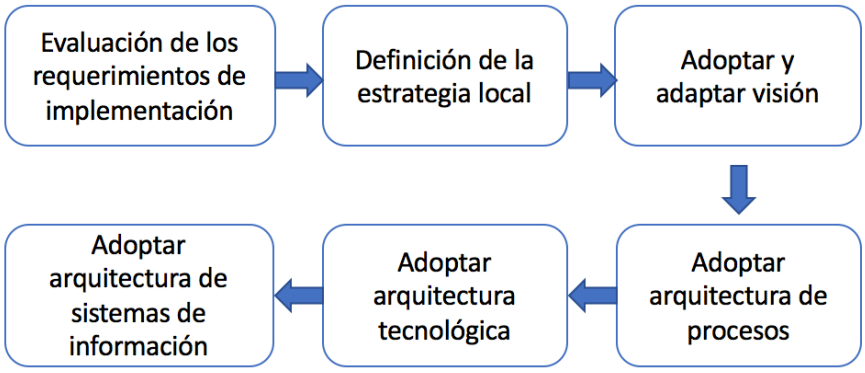
4.1 Descripción del modelo de arquitectura empresarial

Un modelo se refiere a la abstracción de una realidad especifica en un contexto definido. Por lo tanto, este capítulo describe un modelo con las especificaciones de la arquitectura empresarial para los análisis de grandes volúmenes de datos. El modelo describe las consideraciones preliminares a la implementación de una arquitectura empresarial como son las cuestiones organizacionales y quienes deben estar directamente involucrados. Así mismo, se define la visión general del trabajo de arquitectura y posteriormente se detalla cada componente del modelo que son la arquitectura de negocio, arquitectura de sistemas de información y la arquitectura tecnológica.

La adopción del modelo se deberá llevar a cabo mediante una evaluación previa de los prerrequisitos del modelo, de cumplir con los requisitos previos, el centro de investigación deberá formular una estrategia local de implementación del modelo (ver Figura 16). En primera instancia deberá adoptar y adaptar la visión planteada en el modelo para su contexto especifico. Posteriormente se deberá adoptar la arquitectura de procesos, la arquitectura

tecnológica y por último la arquitectura de sistemas de información que se implantará en base a la tecnología antes adoptada.

Figura 16: Adopción del modelo



Fuente: Elaboración del autor

En este sentido, como se mencionó anteriormente, es necesario considerar primero si este modelo es conveniente para el centro de investigación en particular.

4.1.1 Evaluación de los requisitos del modelo

La evaluación de cumplimiento de los pre requerimientos, permite hacer una revisión de cuáles son los escenarios en los cuales este modelo será más conveniente implementarlo.

- El modelo se desarrolló para pequeños y medianos centros de investigación.
- Los centros de investigación deben realizar investigaciones haciendo uso de bioinformática.
- Los involucrados deben estar dispuestos a modificar procesos, servicios, entre otros en el momento de la implementación, de acuerdo a la arquitectura planteada.
- El centro debe poseer infraestructura para copias de respaldo locales.
- El centro debe asegurarse que la arquitectura sugerida es compatible con sus necesidades.

A continuación, sin más preámbulos se presenta el modelo.

4.2 Consideraciones preliminares

Esta sección describe la fase preliminar al ADM que comprende una vista general para proveer un contexto para la propuesta de las cuatro fases de arquitecturas de la presente investigación. Tanto esta fase como la fase de visión de arquitectura es un modelo genérico, que debe adaptarse a un contexto concreto de organización.

La principal motivación de esta propuesta es cubrir la necesidad de una arquitectura que soporte de forma sencilla, flexible y fiable las investigaciones para los científicos.

4.2.1 Modelo organizacional

El modelo organizacional describe el modelo organizacional de arquitectura empresarial, el cual define la organización, roles y responsabilidades para un desarrollo exitosos de la arquitectura empresarial dentro de la organización. Cabe resaltar que los objetivos de arquitectura, *framework* de arquitectura personalizado y el repositorio de arquitectura se encuentran desarrollados en la siguiente sección de visión de arquitectura.

4.2.1.1 Alcance de las organizaciones impactadas

Los principales procesos de negocio se beneficiarán del diseño de un modelo de arquitectura empresarial que tiene como alcance el proceso de análisis bioinformático de los grandes volúmenes de datos, las personas involucradas en dicho proceso, los sistemas y la tecnología que soporta dicho macro proceso.

4.2.1.2 Evaluación de madurez

Las organizaciones objetivo generalmente no tienen una arquitectura empresarial definida, especialmente en el área científica las metodologías y modelos de madurez no han sido desarrollados. En este sentido, estas organizaciones y especialmente el rubro de bioinformática tienen un grado de madurez muy básico.

4.2.1.3 Roles y responsabilidades para el equipo de arquitectura

En la Tabla 3 se describe los roles, también las responsabilidades de los involucrados en el proceso de análisis bioinformático de los grandes volúmenes de datos.

Tabla 3: Roles y responsabilidades del equipo

Stakeholder	Descripción	Responsabilidades
Director general	Es la máxima autoridad que gestiona y lidera la organización.	Gestionar en base los objetivos estratégicos.
Investigador	Es el actor principal del negocio, ya que da inicio a al principal proceso a través de los requerimientos.	Tener claras las preguntas de investigación. Obtener datos de calidad.
Líder de TI	Es el responsable de atender los requerimientos para diseñar un modelo de servicios para el análisis de GVD.	Brindar servicios que satisfagan las necesidades de los investigadores.
Administrador del sistema	Es el responsable de administrar los accesos y servicios informáticos.	Velar que los sistemas funcionen correctamente.
Bioinformático	Es el actor principal del negocio, ya que es el que ejecuta los principales procesos con requerimientos ya establecidos.	Analizar los datos con herramientas y servicios. Realizar el control de calidad de sus resultados.

Fuente: Elaboración del autor

4.2.1.4 Restricciones para el trabajo de arquitectura

Las restricciones pueden variar de una organización a otra. Se presenta una lista de restricciones a tener en cuenta en la implementación de una arquitectura empresarial.

4.2.1.5 Gobierno y Estrategia de Soporte

El comité de arquitectura encargado del gobierno de las actividades de arquitectura está conformado por un investigador principal, el líder de TI y el bioinformáticos. Es necesario, pero no imprescindible contar con un arquitecto empresarial quien pueda guiar la implementación de la arquitectura empresarial. De acuerdo a la experiencia de cada integrante del comité, se debe tener un responsable para la arquitectura de negocio, arquitectura de datos, aplicaciones y para la arquitectura de infraestructura.

Tabla 4: Restricciones y mitigaciones del modelo

Restricción	Severidad	Probabilidad	Mitigación
La alta gerencia no esté de acuerdo con la arquitectura empresarial en el ámbito científico.	Alta	Baja	La correcta comunicación de los beneficios evitará que suceda.
Las tecnologías a utilizar van cambiando constantemente.	Media	Media	El modelo contempla tecnologías escalables como el *cloud computing*.
El especialista de TI no está familiarizado con las necesidades científicas.	Media	Media	El profesional bioinformáticos se encarga de crear lazos.
Los investigadores científicos no están familiarizados con TI.	Media	Baja	El profesional bioinformáticos se encarga de crear lazos.

Fuente: Elaboración del autor

4.2.2 *Framework* de arquitectura adaptado

Según el modelo TOGAF, se debe adoptar el modelo para su integración en la organización. La adaptación y personalización del modelo es responsabilidad de cada organización, y no se encuentra en el alcance de esta investigación. Esta adopción integrara los marcos de gestión de proyectos y procesos, personalización de los detalles como la inclusión de la cultura organizacional, los interesados y la actual capacidad de la arquitectura. La personalización en este nivel seleccionará entregas y adecuados artefactos para satisfacer las necesidades de los involucrados y del proyecto en general.

En esta sección, se debe incluir la planificación de negocios corporativos; la cartera, programa y gestión de proyectos, desarrollo del sistema; operaciones y servicios.

4.2.3 Requerimiento de trabajo de Arquitectura

Este documento inicia el ciclo de arquitectura empresarial, este documento presenta los motivos que conllevan a arquitectura empresarial, el alcance del compromiso de negocio para apoyar el trabajo de arquitectura empresarial y la estrategia que se debe seguir para realizar el trabajo de arquitectura empresarial. Particularmente en este modelo, se asume que la

alta dirección es quien patrocina el trabajo de arquitectura para beneficiar las investigaciones desde una perspectiva de TI.

4.2.3.1 Patrocinadores organizacionales

En esta sección se debe listar los patrocinadores que llevaran a cabo la implantación de la arquitectura empresaria. Listar el nombre, cargo y responsabilidades.

4.2.3.2 Misión de la organización

Describe la misión de la empresa. Cabe resaltar, que se presenta una misión genérica del centro piloto CIP en esta propuesta, la misión se debe adaptar para un contexto concreto de la organización particular.

Misión	*"Nuestra misión es trabajar con nuestros socios para alcanzar la seguridad alimentaria, el bienestar y la igualdad de género de las personas pobres que dependen de los cultivos y sistemas alimentarios de raíces y tubérculos en el mundo en desarrollo. Lo hacemos mediante investigación y la innovación en ciencia, tecnología y el fortalecimiento de las capacidades".* **(Centro internacional de la papa, 2015)**

4.2.3.3 Objetivos de Negocio

Describe los objetivos de negocio de las organizaciones. Listar los objetivos de negocio organizacionales. Por ejemplo, para nuestra institución piloto CIP, en la siguiente tabla se describe los seis objetivos estratégicos de negocio.

4.2.3.4 Plan estratégico de negocio

En este apartado se debe detallar el plan estratégico de negocio de la organización. De igual manera, se presenta un corto ejemplo en el centro piloto de la investigación.

Investigación y desarrollo (I+D) con el fin de entregar en plazos más cortos soluciones para la seguridad alimentaria con nuestros productos y en las geografías seleccionadas yendo a escala con las tecnologías emblemáticas.

Investigación básica para el desarrollo que intente entregar resultados de la investigación en el futuro, a través de descubrimientos emblemáticos, que representen soluciones de más largo plazo para el desarrollo.

Sobre conservación y uso de la biodiversidad, constante compromiso de proteger y utilizar las colecciones mundiales de papa y camote. (Centro internacional de la papa, 2016)

Tabla 5: Objetivos de negocio

Objetivo	Objetivo detalle
Camote resistente y nutritivo	Se pretende llegar a al menos 15 millones de personas con pocos recursos en África subsahariana, Asia y Haití, permitiéndoles mejorar la calidad de su dieta en un 20% y aumentar sus ingresos por cultivos en un 15%.
Papas ágiles para Asia	Mejorar la productividad de los sistemas e ingresos agrícolas de al menos siete millones de personas en los países asiáticos de China, Bangladesh, India, Vietnam, Pakistán, Nepal y Asia Central durante los próximos 10 años.
Semillas de papa para Africa	en un plazo de 10 años, al menos 600,000 personas aumentarán sus rendimientos de papa en un 50% y los ingresos en al menos US$ 800/ha por temporada.
Soluciones prácticas	Establecer productos innovadores para el nivel experimental para al menos una solución de cambio de juego
Sistemas alimentarios resistentes	Desarrollar un conjunto de marcos, métodos y herramientas basados en sistemas y ciencias sociales y se concluirán el análisis y la mitigación de la vulnerabilidad alimentaria
Conservando la biodiversidad para el futuro.	Acortar a seis meses o menos la limpieza fitosanitaria de cultivos de rutina.

Fuente: Elaboración del autor

4.2.3.5 Cambios en el ambiente organizacional

La presente propuesta de arquitectura empresarial no pretende realizar cambios organizacionales, debido a que se aplicara solo un ciclo del ADM de TOGAF.

4.2.3.6 Restricciones de la organización

Se presenta a continuación una lista de restricciones de la organización a tener en cuenta en la implementación:

- La alta gerencia no siempre está consciente de los detalles de los problemas del ámbito científico.

- Las tecnologías a utilizar van cambiando constantemente.
- Las organizaciones del rubro de investigación no siempre se pueden adaptar plenamente a los procesos de mejora continua y metodologías implementadas en organizaciones convencionales.
- Las organizaciones sin fines de lucro del ámbito científico generalmente poseen objetivos a largo plazo.

4.2.3.7 Descripción del negocio actual

En esta sección se detalla la problemática del negocio y específicamente del área de estudio. Se presenta la problemática en bioinformática de nuestra investigación.

Los problemas biológicos actuales requieren soluciones que usen métodos avanzados a nivel informático para datos provenientes de las investigaciones, debido a que los volúmenes en estas áreas van en aumento acelerado en cantidad y complejidad, es así que existen archivos muy grandes provenientes de las nuevas tecnologías de secuenciamiento y cada uno que deben ser almacenado y analizado. Por otro lado, debido a estos grandes volúmenes de datos y la capacidad para almacenar los que necesita, los usuarios guardan la información en diferentes unidades de almacenamiento personales y/o en servidores externos, causando desorden con la información y consecuentemente problemas con la recuperación de la información, esta diversificación da lugar a un uso inadecuado de la información, dificultad para centralizar y compartir información, por ende se torna difícil encontrar la información porque hay demasiados datos y es difícil generar conocimiento porque hay demasiada información en diversas fuentes.

Así mismo, este gran volumen de datos debe ser procesado y analizado con supercomputadoras que tengan las suficientes de prestaciones informáticas, lo cual genera una demanda creciente de hardware conforme los datos van creciendo en tamaño, es por ello que la infraestructura realiza los análisis en un plazo más prolongado de tiempo. Por todo ello, la ralentización de los análisis, el difícil tratamiento de datos, la incompatibilidad con aplicaciones actuales y los gastos incurridos han generado una necesidad

debido a que los usuarios optan por el uso de infraestructura de otras organizaciones en su afán de continuar con las investigaciones. Esta situación, genera la necesidad de proponer un modelo que pueda abarcar una solución a toda esta problemática antes mencionada, el cual contempla el diseño de una arquitectura del sistema que permite apalancar la gestión de recursos y la penetración suficiente para manejar la complejidad creciente en que se desenvuelven las investigaciones.

4.2.3.8 Propósito del trabajo de arquitectura

Una arquitectura busca resolver de una manera coherente e integrada, al mismo tiempo que ofrece un camino para poder entender y conceptuar entre todos los interesados. A partir de esto, se genera la definición de arquitectura empresa, que intentar definir y describir todos los componentes; tecnológicos y no tecnológicos, así como las relaciones entre ellos y con el entorno.

La expectativa que se tiene de realizar un trabajo de arquitectura empresarial es llegar a diseñar adecuadamente y poner en marcha el proceso de análisis bioinformático de modo que pueda contribuir a los objetivos organizacionales, disminuir los costos operativos y alcanzar los objetivos de investigación esperado.

4.2.3.9 Aplicabilidad

La aplicabilidad del modelo de arquitectura empresarial se extiende a pequeñas y medianas organizaciones dedicadas a las investigaciones científicas principalmente del sector agrícola, que posean proyectos con el componente bioinformático como parte de la investigación. Se recomienda la implantación de este modelo tomando en cuenta los siguientes aspectos:

- La cooperación de las áreas de apoyo como requerimiento puede beneficiar grandemente la implementación.
- La organización debe tener proyectos definidos en el mediano y largo plazo, de esta manera la implementación de una arquitectura empresarial TOGAF desarrollará un proceso una mejora continua. Así mismo,

estos proyectos ayudan a manejar las inversiones en infraestructura a largo plazo.

- Incluir expertos y personal con experiencia, es un factor clave de éxito que permite que la implementación de arquitectura sea productiva.

4.2.4 *Framework* de gobernabilidad de la arquitectura

La implementación de una arquitectura empresarial genera una gran cantidad de salidas y estos se incrementan de acuerdo al ciclo de ADM en la que se encuentre, en esta investigación solo se considera el primer ciclo y queda a responsabilidad de la organización como manejar y almacenar los documentos de gestión arquitectónica.

4.3 Definición de visión de la arquitectura

Esta sección describe la capacidad y los lineamientos que se entregarán como salida de la arquitectura empresarial propuesta. En esta sección se presenta la declaración de trabajo de arquitectura con la que se da inicio al trabajo.

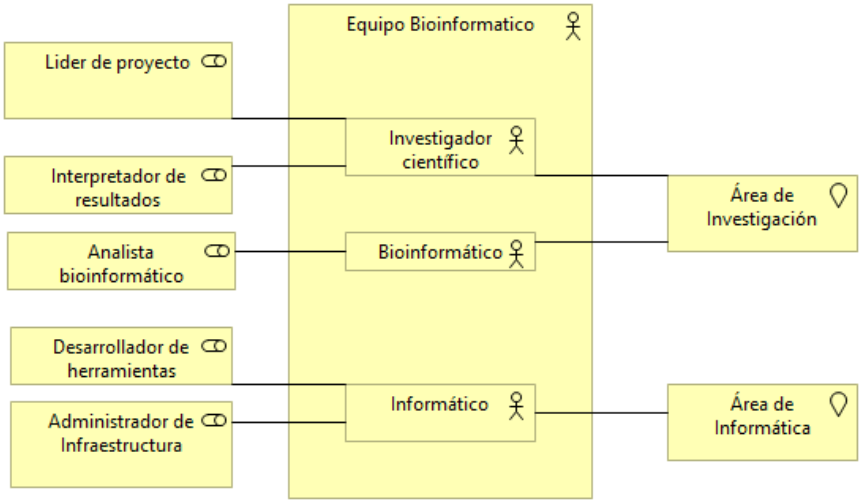
4.3.1 Visión general de la arquitectura

La visión de la arquitectura permite presentar de una manera amplia toda la arquitectura en su conjunto. En esta sección se detalla la lista de los interesados y sus preocupaciones. A partir de este análisis se crea una lista de requerimientos que necesitan ser satisfechos, asi mismo se presenta el alcance y como el modelo resultante.

4.3.1.1 Interesados y preocupaciones

El equipo bioinformático está compuesto de personas del área de investigación como al área de informática, este equipo está conformado por bioinformáticos, informáticos e investigadores científicos, quienes cumplen roles de líder de proyecto analista bioinformático, interpretación de resultados, desarrollador de herramientas y administrador de infraestructura. La Figura 17 muestra el diagrama de *stakeholder*s de los principales involucrados, el área y sus roles en el negocio.

Figura 17: Stakeholders y sus roles



Fuente: Elaboración del autor

A partir de la definición de los interesados y sus roles, en la Tabla 6 se muestra el mapa de *stakeholders*, se detalla las funciones de cada rol y se define sus preocupaciones, así también se describe el tipo de reporte para cada uno de ellos. Este mapeo nos permitirá saber en detalle cuales son las preocupaciones y necedades que la arquitectura de negocio debe cubrir.

Tabla 6: Mapa stakeholders

Stakeholder	Descripción	Principal Preocupación	Clase
Director general	Es la persona investida de máxima autoridad en la gestión y dirección administrativa.	Que las metas y los objetivos estratégicos, se interprete de una manera adecuada.	Mantener Satisfecho
Investigador	Es el actor principal del negocio, ya que da inicio a los principales procesos con requerimientos ya establecidos.	Que las herramientas y servicios estén disponibles cuando lo necesite. Que los datos de investigación no se pierdan	Persona Clave
Líder de TI	Es el responsable de atender los requerimientos para diseñar un modelo de servicios para el análisis de grandes volúmenes de datos.	Que se diseñe un modelo de servicios que cumpla con las expectativas de los investigadores.	Mantener informado
Administrador del sistema	Es el responsable de administrar los accesos y servicios informáticos.	Que el modelo se implemente correctamente	Persona Clave
Bioinformático	Es el actor principal del negocio, ya que es el que ejecuta los principales procesos con requerimientos ya establecidos.	Que los datos, herramientas y servicios estén disponibles cuando lo necesite.	Persona Clave

Fuente: Elaboración del autor

4.3.1.2 Objetivos de la declaración del trabajo de arquitectura

Los objetivos del trabajo de arquitectura son los siguientes:

- Definir el alcance y el enfoque que se utilizará.
- Presentar las metas y los objetivos estratégicos.
- Plantear correctamente el alcance y las restricciones.
- Proponer un modelo que sea factible de implementar.

4.3.1.3 Alcance

La arquitectura empresarial propuesta se enfoca en la unidad de análisis objetivo que pretende servir como modelo de arquitectura para áreas de bioinformática existentes o en su defecto para nuevas. Los principales procesos de negocio se beneficiarán del diseño de un modelo de arquitectura empresarial enfocado en grandes volúmenes de datos provenientes de las investigaciones en biología molecular.

4.3.1.4 Descripción de procesos

El ámbito de este estudio, tiene como alcance dos procesos de negocio el Proceso de análisis bioinformáticos de grandes volúmenes de datos y el proceso de interpretación, mientras que el primero representa el interés principal de esta investigación por ser trascendental

a) Proceso de análisis bioinformáticos de grandes volúmenes de datos

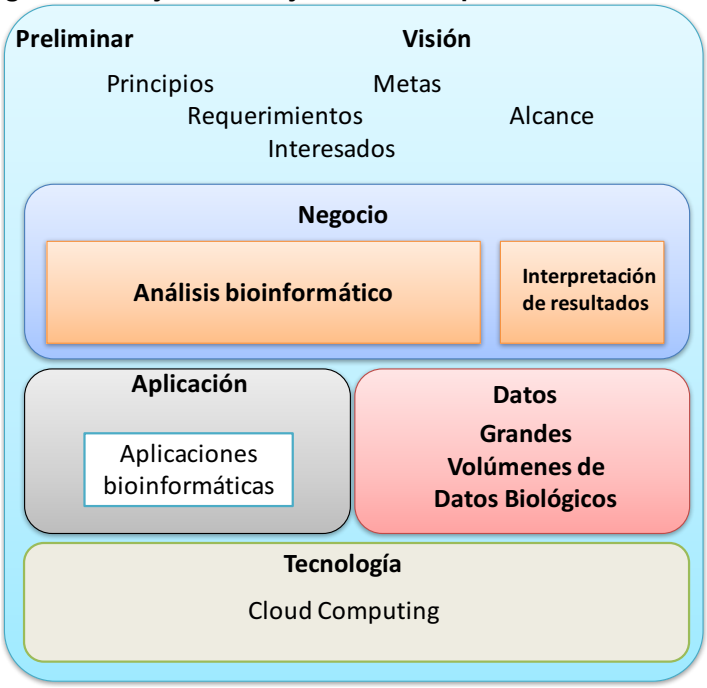
El proceso de análisis bioinformático de grandes volúmenes de datos es el proceso principal en este estudio. Se centra en los análisis bioinformáticos de los grandes volúmenes de datos, el cual consiste de análisis computacional de datos. La entrada consta de los GVD generados a través de las tecnologías de *Next Generation Sequencing* - NGS (descrito en la parte teórica de ésta tesis), los cuales constan de datos de ADN, ARN, proteínas, etc. Estos datos pasan por procesos de procesamiento y análisis según la naturaleza del estudio con software especializado para tales funciones. Cabe resaltar que estos procesos pueden resultar repetitivos de acuerdo al análisis que se realice y en términos de computación y datos puede tomar muchos días de procesamiento.

b) Proceso de interpretación

A partir de los resultados generados de los análisis de GVD, el interpretador de resultados, en este caso un investigador científico, utilizando herramientas específicas para tal función, genera el producto final que es el resultado final del análisis bioinformático.

Los datos recibidos como resultados deben ser analizados para verificar que sean significativos, así mismo deben ser validados a través de estadígrafos, contrastación con literatura existente, así como por la experiencia de los científicos, entre otros. En esta fase los resultados pueden analizarse tantas veces como haga falta y lo requiera el investigador científico.

Figura 18: Flujo de trabajo de los componentes del modelo



Fuente: Elaboración del autor

4.3.1.5 Modelo resultante de la arquitectura

Dentro de esta visión de arquitectura se presenta el modelo resultante de la arquitectura en un alto nivel, el cual nos dará una idea de la arquitectura resultante. Cabe resaltar que el proceso para obtener este modelo de arquitectura se detalla en las siguientes fases del ADM de TOGAF. En la Figura 18 se muestra la arquitectura con la parte preliminar, visión, capas de negocio, aplicación, datos y tecnología.

4.3.2 Objetivos de negocio

En la sección 4.3.1.1 se describe los objetivos de negocio.

4.3.3 Principios para la arquitectura

Como marco de desarrollo del modelo, el cual se encuentra direccionado al cumplimiento de los objetivos, se contempla los siguientes principios, sin antes mencionar las políticas que lo acompañan.

4.3.3.1 Políticas

En la Tabla 7 se detalla las políticas que se deben identificar al momento de proponer una arquitectura que soporten todos los procesos de negocio que se citan en la presente investigación.

Tabla 7: Políticas de TI

Política	Detalle
Procesos y proyectos	Las áreas de gestión y operativas deben interaccionar con los procesos, donde cada proceso debe tener un dueño quien es el responsable de la eficiencia y eficacia del proceso.
Planes y programas	Mediante un proceso de planeación integral se debe desarrollar planes y proyectos que garanticen la continuación.
Regulación interna	Las políticas, normativas, reglas, procesos, estructuras y procesos deben ser compartidos entre los involucrados, de esta forma ellos tendrán identificados sus responsabilidades en la organización y el impacto en otras áreas.
Desarrollo de tecnología	La infraestructura tecnológica es uno de los pilares, los cuales facilitaran la creación de valor agregado con nuevos servicios. Estos sistemas deberán ser de última generación que puedan automatizar los procesos y sea de fácil acceso para los investigadores.
Tecnología de información	Los estándares para el uso de la infraestructura deben ser definidos dentro de la organización.
Operaciones	Las actividades técnicas se programarán según el plan operativo de los cuales son parte los investigadores.
Negocios	Los procesos deben ser parte de una mejora continua para que se puedan adaptar a los desarrollo tecnológicos.

Fuente: Adaptado de (PACIFICTEL S.A., 2006)

4.3.3.1 Principios de negocio

Para asegurar el logro de estas metas de la investigación se deben cumplir con los siguientes principios básicos mostrados en la Tabla 8.

Tabla 8: Principios de arquitectura

Principios	Declaración
Procedimiento de conexión entre SI.	Definir los mecanismos de integración que se deben utilizar entre los sistemas de información existentes y los nuevos, teniendo en cuenta las mejores prácticas
Acuerdos de Nivel de Servicios (SLA)	Revisar la prestación de servicios de los Acuerdos de Nivel de Servicio (SLA), con el propósito de conocer las condiciones del servicio y el valor que aporta el servicio de Tecnología de Información (TI) al cumplimiento de los servicios brindados.
Provisión de Tecnología	El desarrollo de la bioinformática debe contar con tecnologías de información que permitan el apoyo hacia el logro de los objetivos, promuevan ventajas competitivas, contemplen las mejores prácticas y sean flexibles a los cambios que se produzcan.
Planeación Estratégica	Los servicios bioinformáticos deben estar definidos y priorizados. Igualmente, los servicios de TI deben cumplir con los principios, políticas y arquitecturas, para garantizar la integración de servicios.
Negociaciones tecnológicas	En la negociación se deben obtener los productos y servicios objetivos, de manera que las relaciones de trabajo sean efectivas y permitan flexibilidad y la actualización tecnológica.
Gestión de los datos	La única manera de proporcionar un nivel coherente y medible de investigación de calidad es la gestión de los datos que debe estar alineada con el negocio.

Fuente: Elaboración del autor

4.3.3.2 Principios para los datos

Los datos son el centro de esta investigación y particularmente los GVD, deben tener en cuenta los principios de la Tabla 9.

Tabla 9: Principios de datos

Principio	Declaración
Datos como activo	Los datos son activos de gran valor de conocimiento, por tanto, se deben gestionar de acuerdo a su importancia.
Compartir datos	Los investigadores tienen acceso a los datos puedan realizar sus investigaciones; por lo tanto, los datos se comparten.
Datos accesibles	Los datos deben ser accesibles cuando realizan sus funciones.
Definición común de datos	Cada elemento de dato tiene un responsable de la calidad de datos
Seguridad de datos	Los datos están protegidos de uso y divulgación no autorizada.

Fuente: Elaboración del autor

4.3.3.3 Principios para las aplicaciones

Se deben tomar en cuenta los principios de la Tabla 10.

4.3.3.4 Principios para las tecnologías de información

Las estructuras de la tecnología de la información, se están volviendo más dinámicas con el fin de proporcionarle flexibilidad e incluso mantener un enfoque estable para alcanzar las necesidades futuras. De esta manera los principios son extensivos para la bioinformática, ver Tabla 11.

Tabla 10: Principios de aplicaciones

Principio	Declaración
Independencia de la tecnología	Las aplicaciones son independientes de opciones tecnológicas específicas y por lo tanto pueden funcionar en una variedad de plataformas tecnológicas
Uso fácil	Las aplicaciones son fáciles de usar. La tecnología subyacente es transparente para los investigadores.
Basados en el cambio	Sólo en respuesta a las necesidades de cambios en las investigaciones las aplicaciones cambiarán.

Fuente: Elaboración del autor

Tabla 11: Principios de TI

Principios	Declaración
Gestión de recursos y servicios informáticos	El encargado de TI debe gestionar servicios e infraestructura tecnológica según las necesidades específicas de los investigadores, siendo responsable por su uso.
Asignación de recursos y servicios informáticos	El proceso infraestructura verifica y asigna los servicios e infraestructura tecnológica que se ajusten a las necesidades específicas de los investigadores y su rol.
Licenciamiento de Software	El software que se utilice debe ser licenciado legalmente o debe ser de libre distribución.
Utilización de los recursos y servicios informáticos	Los servicios informáticos en este punto serán de uso directo para las actividades relacionadas con la investigación y su utilización puede ser monitoreada por los encargados de TI.
Responsabilidades del proceso infraestructura	La administración, mantenimiento, operación y expansión de la infraestructura informática es responsabilidad directa del proceso de infraestructura
Especificación de infraestructura	Las especificaciones de infraestructura tecnológica deben ser de acuerdo al presupuesto, las funciones y los criterios técnicos requeridos en el proceso.
Conservación de datos	El almacenamiento de los datos en repositorios y su respectivo respaldo debe ser según su estado crítico y el ciclo de vida, lo cual determinara el recurso a usar.
Control de la diversidad técnica	La diversidad tecnológica es controlada para minimizar el costo de mantenimiento, la experiencia en el procesamiento y la conectividad entre múltiples ambientes.
Interoperabilidad	La interoperabilidad, y escalabilidad son factores claves al momento de elegir el software y hardware.

Fuente: Elaboración del autor

4.3.4 Declaración de trabajo arquitectura

El proyecto se centra en desarrollar una arquitectura que permita la gestión de los análisis bioinformáticos para los grandes volúmenes de datos. El marco de trabajo se apoya en la Metodología de Desarrollo de Arquitectura (ADM) de TOGAF, que propone como derivar una arquitectura empresarial que dé respuesta a los requerimientos del proceso a través de cada una de las etapas del ciclo con un conjunto de recursos que ayude en la aplicación de la metodología, partiendo de las cuatro dimensiones las cuales son: negocio, aplicaciones, datos y tecnología.

4.4 Modelo de negocio

El modelo de negocio describe la arquitectura del negocio propuesta y constituye la base sobre la cual se cimientan las demás arquitecturas, sirviendo como punto de partida para el diseño de la arquitectura del sistema de información y la arquitectura tecnológica. Este modelo propuesto tiene como punto de partida el análisis de línea base de arquitectura, así mismo se realizó un análisis comparativo de las unidades de bioinformáticas con otros centros de investigación. Este análisis no está dentro de la metodología TOGAF, pero se detalla a continuación con el propósito de crear paso previo a la propuesta de arquitectura de negocio.

4.4.1 Definiciones principales

4.4.1.1 Objetivos y limitaciones

a) Objetivos

La arquitectura empresarial en el ciclo de trabajo se presenta de acuerdo a los cuatro componentes del modelo:

- **Organización y procesos:**
  - Crear una arquitectura que se adapte a los cambios.
  - Incrementar el nivel de participación de los interesados.
  - Alinear los procesos al negocio.
  - Incrementar la comunicación entre todos los involucrados.
  - Alinear las tecnologías de información al proceso.
- **Aplicaciones**
  - Incrementar el nivel de satisfacción de los usuarios con las aplicaciones.
  - Mejorar la disponibilidad de los sistemas.
  - Brindar mayor utilidad a los sistemas y aplicaciones.
- **Datos**
  - Mejorar la disponibilidad de los datos.
  - Incrementar el nivel de satisfacción con la gestión de datos.
- **Infraestructura**
  - Proveer recursos tecnológicos desde dentro como fuera del ambiente físico de trabajo.

- Mejorar el nivel de continuidad del servicio.
- Disminuir costos.
- Mejorar el nivel de desempeño de la tecnología.

b) Limitaciones

A continuación, se presenta una lista de limitantes a tener en cuenta en la implementación de una arquitectura empresarial:

- La alta gerencia no siempre está de acuerdo con adoptar la arquitectura empresarial en un solo área del ámbito científico.
- Las tecnologías a utilizar van cambiando constantemente.
- El especialista de TI no está familiarizado con las necesidades científicas.
- Los investigadores científicos no están familiarizados con TI.

4.4.1.2 Linea base

a) Organización estructural

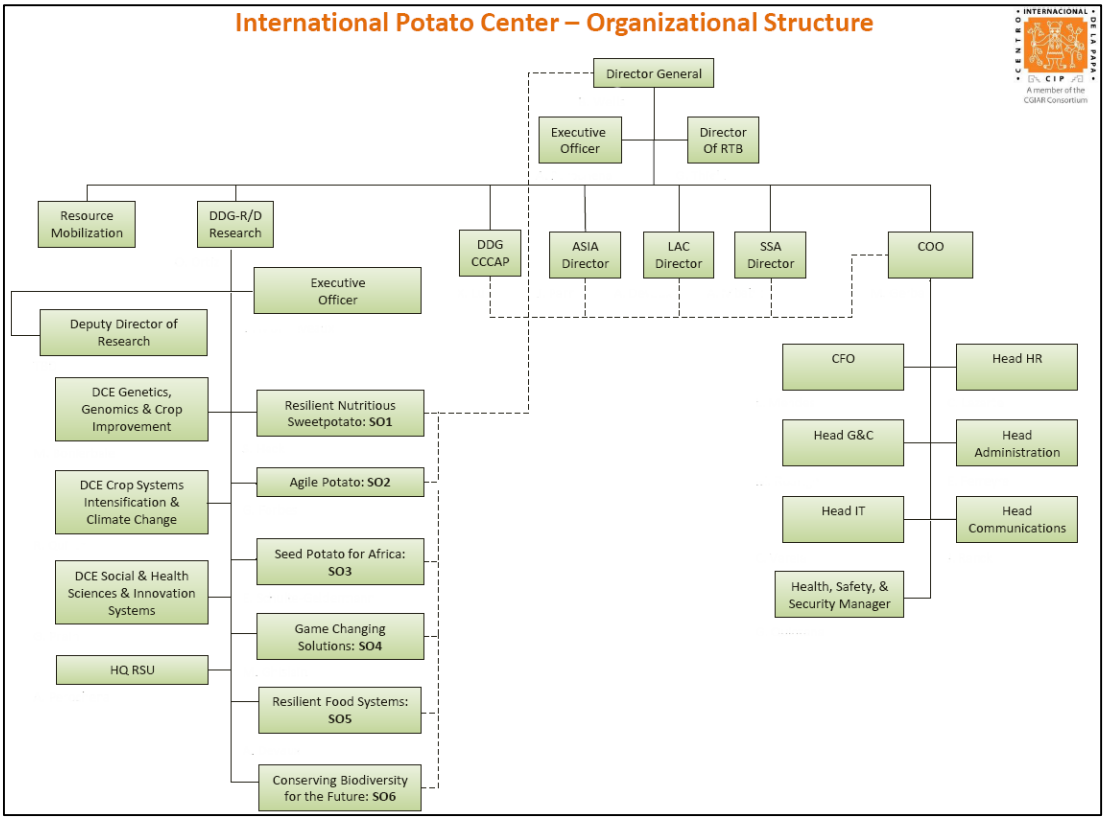
Las prioridades de los centros de investigación globales incluyen el sostenimiento de la biodiversidad de cultivos; mejoramiento de nutrientes, variedades resistentes a las plagas y enfermedades, y la construcción de sistemas agro-económico-sociales flexibles para las poblaciones marginales en los países en desarrollo. Los objetivos y metas de cada centro de investigación pueden diferir ampliamente.

En la presente investigación el Centro Internacional de la Papa (CIP) fue elegido como centro piloto. “CIP es una organización de investigación agrícola internacional sin fines de lucro con un mandato mundial de realizar investigación en papa, camote, yuca, otros tubérculos y raíces, la gestión sostenible de los recursos naturales y la agricultura urbana” (Centro internacional de la papa, 2015).

La investigación en el CIP “está dirigida a incrementar la producción de alimentos y fortalecer los sistemas agrícolas, para mejorar la calidad de vida de los países en desarrollo beneficiando a aquellos que no tienen ni capital, ni recursos, ni la semilla de calidad que el CIP les ofrece”

(Centro internacional de la papa, 2015). El CIP, al momento que se realizó la investigación, tenía tres programas científicos globales, como se aprecia en el organigrama de la Figura 19, los cuales son transversales entre los cultivos:

Figura 19: Organigrama CIP



Fuente: Elaboración del autor

- **Genética, genómica y mejoramiento de cultivos**

Recursos genéticos investiga el rasgo de abastecimiento, la filogenia, la estructura de genes, capacidad de cruce de rasgo y la introgresión de genes. Mejoramiento de cultivos desarrolla papa resistente y variedades de camote. Así mismo, mejora los niveles de rasgo objetivo, revierte la genética y genómica funcional de genes y la detección de redes.

- **Sistemas de cultivos integrados y cambio climático**

Desarrolla el conocimiento científico, estrategias, modelos, métodos de evaluación de riesgos básicos y herramientas de soporte de decisiones relacionadas con las principales restricciones de papa y camote.

- **Nutrición social y ciencias de la innovación**

Esta área se dedica a la realización de la caracterización de poblaciones objetivo, perspectivas de los usuarios, modelado del impacto económico/salud, encuestas de nutrición y consumo, el cambio de comportamiento y estudia el entorno político.

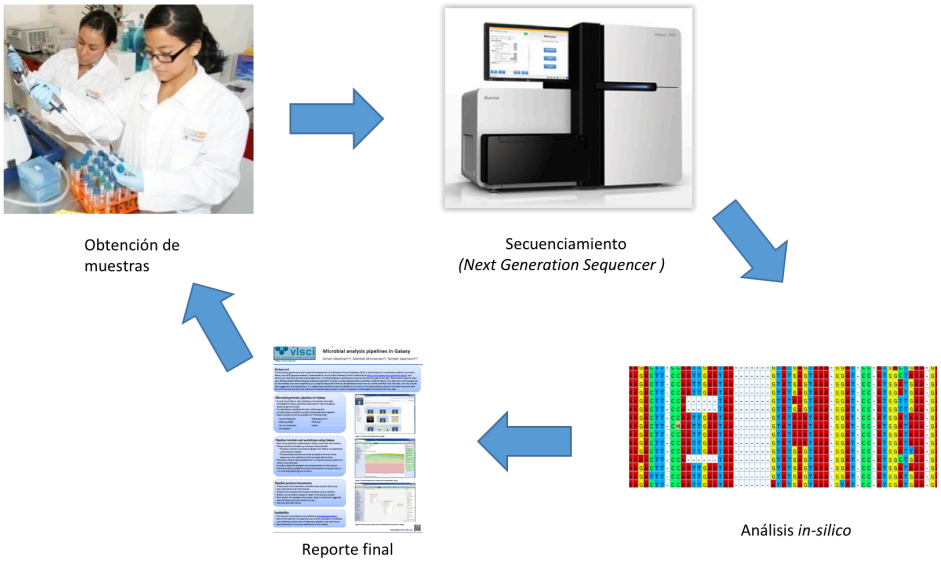
Como se observa en el organigrama, el CIP no tiene un área formal de bioinformática debido a que es un área pequeña en términos de significancia a nivel global, con las oficinas a nivel internacional. A nivel organizacional bioinformática trabaja dentro el marco del área de genética, genómica y mejoramiento de cultivos.

b) Línea base lógica de la arquitectura

En el marco de trabajo del área de investigación de científica especialmente en los análisis de datos genéticos y genómicos en plantas, se tiene un proceso genérico no formal que nos sirve de punto de partida para la propuesta. El proceso se describe a continuación (Figura 20).

Las muestras de las plantas de campo como las plantas de invernadero son colectadas y sometidas a un proceso de laboratorio para la extracción del ADN/ARN de la planta, cuyas extracciones pueden ser de diferentes partes como del tallo, fruto, hojas, raíz o una combinación que dependerá del tipo de investigación en cuestión. Estas muestras se envían a empresas especializadas en secuenciamiento, que utilizando diversas técnicas de *Next Generation Sequencing* (como se detalló en la parte teórica), entregan las muestras en formato de archivos de texto que contienen las secuencias de nucleótidos en formato FASTQ (llamado GVD para este estudio). Posteriormente se realiza el análisis *in-silico* de estas muestras. El análisis i*n-silico* consiste en el análisis de las muestras biológicas haciendo uso de las computadoras, en este estudio conocido como análisis bioinformático. Luego del análisis *in-silico* se obtienen reportes finales, los cuales sirven de base para profundizar o realizar investigaciones. Este puede ser iterativo y muy complejo, y si el investigador obtenga alguna conclusión significativa, el resultado se publica en artículos científicos en revistas.

Figura 20: Proceso genérico de investigación del análisis *in silico*



Fuente: Elaboración del autor

4.4.1.3 Benchmarking en centros de investigación

El modelo de negocio propuesto está basado en un análisis comparativo de 5 centros de investigación e institutos con sus respectivas unidades de bioinformática. La información obtenida es de acceso público a través de internet.

a) CIMMIT

El Centro Internacional de Mejoramiento de Maíz y Trigo, o CIMMYT (*International Maize and Wheat Improvement Center*, en inglés), (CIMMYT, n.d.) colabora con instituciones nacionales de investigación agrícola, otros centros del CGIAR, empresas privadas y las instituciones de investigación avanzada para abordar el problema a escala global al proporcionar a los agricultores la mejor semilla, agronomía, y la información necesaria para aumentar los rendimientos de los cultivos de maíz y trigo. Cuenta con más de 175.000 accesiones, ya que el CIMMYT mantiene el banco de semillas de maíz y trigo más grande del mundo. Es un "Arca de Noé" de los recursos genéticos para dos de los cultivos más importantes del planeta. (CIMMYT, n.d.) tiene los siguientes programas:

- Programa nacional de maíz
- Programa global del trigo

- Conservación de agricultura
- Investigación de socioeconómica
- Programa de recursos genéticos

Los recursos genéticos y la bioinformática son la responsabilidad del Programa de Recursos Genéticos (GRP). GRP contribuye a la misión general del CIMMYT de aumentar la productividad de los cultivos para mejorar la seguridad alimentaria y mejorar los medios de vida mediante el almacenamiento, análisis y difusión de la mayor colección del mundo de los recursos genéticos de maíz y trigo, que están contenidas en el Centro de Recursos Genéticos Wellhausen-Andersen (CIMMYT, n.d.).

b) CIAT

El Centro Internacional de Agricultura Tropical (CIAT) "contribuye a reducir el hambre y la pobreza, así como mejorar la nutrición humana en los trópicos mediante una investigación que aumente la eco-eficiencia de la agricultura. CIAT tiene la responsabilidad global para la mejora de dos alimentos básicos, yuca y frijol común, junto con forrajes tropicales para el ganado". En América Latina y el Caribe, llevan a cabo la investigación sobre el arroz también. En representación de los diversos grupos de alimentos y componente clave de la biodiversidad agrícola del mundo, esos cultivos son vitales para el abastecimiento mundial de alimentos y la seguridad nutricional (CIAT, n.d.). CIAT tiene una participación importante en los programas de investigación del CGIAR, que abordan los principales desafíos de la agricultura de nuestro tiempo. "El CIAT es el centro principal para el programa sobre el Cambio Climático, Agricultura y Seguridad Alimentaria (CCAFS), que ayuda a los pequeños agricultores a adaptarse y mitigar los efectos del aumento de las temperaturas y las lluvias cada vez más impredecibles" (CIAT, n.d.). CIAT tiene los siguientes programas:

- Agro biodiversidad
- Conservación de suelos
- Decisión y análisis de políticas

El área de bioinformática forma para del área de agro biodiversidad, el cual se encarga de las investigaciones a nivel genético a través del uso de herramientas de análisis y almacenamiento de datos.

c) IRRI

El Instituto Internacional de Investigación del Arroz (IRRI) es la organización de investigación más importante del mundo dedicada a la reducción de la pobreza y el hambre a través de la ciencia de arroz; mejorar la salud y el bienestar de los productores de arroz y los consumidores; y proteger el medio ambiente de cultivo de arroz para las generaciones futuras. IRRI es una organización independiente, sin fines de lucro, la investigación y la institución educativa. El instituto, con sede en Los Baños, Filipinas, cuenta con oficinas en 17 países productores de arroz en Asia y África, y cerca de 1.400 miembros del personal que representan 36 nacionalidades. (IRRI, n.d.) "tiene como objetivo reducir la pobreza y el hambre, mejorar la salud de los productores de arroz y los consumidores, y asegurar la sostenibilidad ambiental del cultivo del arroz". Los programas de investigación que IRRI desarrolla son:

- Diversidad genética
- Mejora de variedades
- Sistema de producción sostenible
- Adicionar valor
- División de Ciencias Sociales
- Desarrollo del sector del arroz

El área de biometría y bioinformática se encarga de los servicios biométricos, gestión de datos, la biología computacional y bioinformática a través del servicio de consulta, colaboración en la investigación y la capacitación. Se apoya en los siguientes tópicos:

- Desarrollo de modelos científicos de datos, estándares de redes y el desarrollo de software.
- Plataforma para los recursos genéticos, genómica y mejoramiento de los cultivos.
- Desarrollo de una red de información de arroz.

- Análisis bioinformático para la preparación de las últimas secuencias genómicas públicas.

d) SANGER

Los investigadores del *Trust Sanger Institute* (Sanger, 2016) participan en proyectos que buscan una mayor comprensión de la función de genes en la salud y la enfermedad y generar datos y recursos de valor duradero a la investigación biomédica. La investigación se divide en seis áreas principales de investigación. Al trabajar dentro de estas áreas, los científicos líderes pueden explorar y desarrollar sus propias hipótesis.

- La genética humana
- Ratón y genética del pez cebra
- la genética de patógenos
- malaria
- Genética celular
- Bioinformática

El programa de Bioinformática desarrolla y aplica métodos para procesar, almacenar y analizar los datos generados por proyectos de alto rendimiento. Sus principales objetivos son inferir conocimiento genómico a través del análisis e integración de datos computacionales y generar recursos de valor duradero a la investigación biomédica.

e) PASTEUR

La estrategia científica del Instituto Pasteur (Pasteur, 2016) se basa en un fuerte apoyo a la investigación básica que está principalmente impulsado por la curiosidad. Los objetivos de los científicos del Instituto Pasteur es acelerar nuestra comprensión de diversos procesos fisiológicos y patológicos y desarrollar nuevas estrategias para el diagnóstico, prevención y tratamiento de la enfermedad. Su trabajo abarca una amplia gama de campos de investigación, incluyendo la microbiología y enfermedades infecciosas, inmunología, la neurociencia, la biología del desarrollo, genética y cáncer. La ciencia y la investigación biológica han entrado en la era de grandes volúmenes de datos. El Instituto Pasteur ha abrazado con entusiasmo esta

revolución mediante la creación del Centro de Bioinformática, Bioestadística y Biología Integrativa. El Centro tiene como objetivo facilitar colaboraciones en bioinformática e impulsar la interacción.

El Centro tiene como objetivo desarrollar las fortalezas existentes en el Instituto Pasteur que alberga un gran número de expertos e investigadores en sus diversas disciplinas, mientras que construcción de otras nuevas. Los siguientes equipos están afiliados al Centro Pasteur Instituto de Bioinformática, Bioestadística y Biología Integrativa:

- Unidad de genómica evolutiva microbiana.
- Unidad de bioinformática estructural.
- Unidad de enfermedades infecciosas.
- Unidad de imágenes y modelamiento
- Unidad de análisis de imagen cuantitativa

El Centro de Informática para Biología (CIB) satisface las necesidades de TI de los científicos del campus a través del desarrollo, suministro y puesta en marcha de herramientas y recursos bioinformáticos. Desarrolla enfoques, métodos, algoritmos y herramientas computacionales y de integración para la visualización post-genómica de datos y el análisis.

A partir de la descripción de cada uno de estos centros de investigación e institutos con sus respectivas áreas de bioinformática, se procedió a realizar el siguiente cuadro comparativo (Tabla 12).

Tabla 12: Bioinformática en otros centros de investigación

Procesos	IRRI	CIAT	CIMMYT	PASTEUR	SANGER
Desarrollo de software y herramientas para bioinformática.	x	x		x	
Prestación de servicios de infraestructura informática.	x				
Análisis de datos genómicos	x	x	x	x	x
Interpretación de resultados				x	x
Actualización de repositorio de datos genómicos	x		x		x

Fuente: Elaboración del autor

Como se puede observar de la tabla anterior, el principal proceso, en los centros que hacen uso de bioinformática, es el análisis de los datos. Del cuadro anterior se definió los siguientes procesos de negocio:

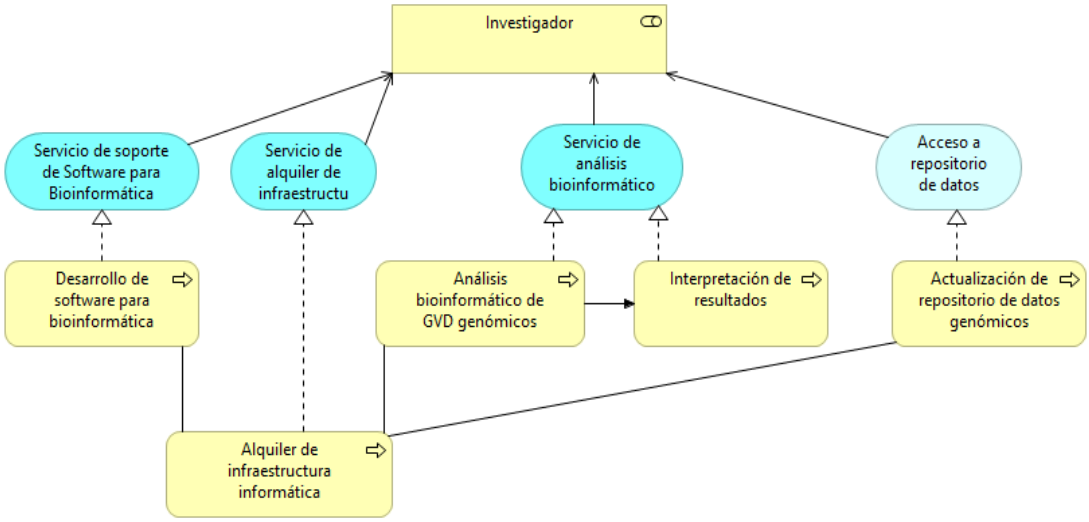
- Desarrollo de software y herramientas para bioinformática.
- Prestación de servicios de infraestructura informática.
- Análisis de datos genómicos.
- Interpretación de resultados.
- Actualización de repositorio de datos genómicos.

4.4.1.4 Identificación de los procesos y servicios

Para poder estudiar ampliamente el proceso de análisis bioinformático de datos, es necesario conocer primero el procesos y servicios relacionados en el área de bioinformática.

A partir del proceso genérico de investigación, así como de la revisión de los procesos de los centros de investigación que trabajan con bioinformática, se identificaron los servicios. En la Figura 21 se muestran los procesos de negocio, así como los servicios que se generan.

Figura 21: Identificación de procesos y servicios



Fuente: Elaboración del autor

De la figura anterior, se definen los procesos de negocio:

- **Desarrollo de software para bioinformática**

Este proceso involucra el desarrollo de software y herramientas basados en modelos científicos de datos, estándares de redes, como plataforma para los recursos genéticos, genómica y mejoramiento de los cultivos. Aplicando métodos y algoritmos para procesar, almacenar y analizar los datos generados por proyectos de alto rendimiento. Así mismo herramientas de visualización post-genómica de datos y el análisis.

- **Prestación de servicios de infraestructura informática**

Este proceso soporta los servicios de alquiler de infraestructura con plataformas bioinformáticas disponibles para el uso de computación intensiva, para mejores prestaciones de cálculo.

- **Análisis de datos genómicos**

Este análisis se detalla en los siguientes puntos.

- **Interpretación de resultados**

El objetivo es inferir conocimiento a través del análisis e integración de datos computacionales, se detalla en los siguientes puntos.

- **Actualización de repositorio de datos genómicos**

El repositorio consiste en el almacenamiento y difusión de la mayor colección del mundo de los recursos genéticos, a través de redes de información y una serie de recursos disponibles de las investigaciones por ejemplo NCBI.

A partir de estos procesos, se definieron los servicios que se tienen disponibles como soporte para las investigaciones y para los investigadores. Estos son los siguientes:

- Servicio de software para bioinformática.
- Servicio de alquiler de infraestructura.
- Servicio de análisis de datos.
- Servicio de acceso a repositorio actualizado de datos genómicos.

4.4.2 Especificación de requerimientos

En la Tabla 13 se definen los requerimientos que se deben satisfacer para cumplir con cada uno de los procesos de negocio. En la tabla de requerimientos de negocio están definidos los requerimientos técnicos que deben ser satisfechos, donde cada requerimiento técnico puede ser funcional o no funcional. Los procesos esta interrelacionados con los requerimientos técnicos los cuales deben ser cumplidos, así, los nuevos procedimientos que deben satisfacer los servicios.

Tabla 13: Detalle de los requerimientos de los análisis bioinformáticos

Requerimiento	Dificultad	Prioridad
Sustento bibliográfico	Media	Media
Análisis y comprobación con experimentos anteriores y similares	Media	Baja
Validación en laboratorio	Baja	Alta
Integración de herramientas de análisis	Alta	Media
Centralizar en un solo ambiente	Media	Alta
Disponible en cualquier momento y lugar	Baja	Alta
Recuperar datos en distintos formatos.	Media	Media
Permitir gestión de reportes	Baja	Media
Múltiples conexiones al sistema simultáneamente	Media	Media
Permite la interpretación de resultados	Baja	Media
Eficiente comunicación de datos	Media	Alta
Gestionar copias de respaldo	Media	Media
Software actualizado	Baja	Alta
Infraestructura robusta	Alta	Alta

Fuente: Elaboración del autor

Teniendo en cuenta que el objetivo de la investigación es el análisis bioinformático, se muestran en la Tabla 14 los requerimientos de este proceso, así como el grado de dificultad que representa y la prioridad asignada, obteniendo es así que el requerimiento critico es obtener una infraestructura robusta.

Tabla 14: Requerimientos por proceso

Proceso	Requerimiento
Prestación de servicios de infraestructura informática	Inserción y descarga de archivos de fuentes externas
	Permite múltiples protocolos de transferencia
	Disponible en cualquier momento y lugar
	Infraestructura con las mejores prestaciones
	Software actualizado
	Permitir almacenar datos no estructurados provenientes de GVD
	Rapidez en los análisis de GVD
	Rapidez en la transferencia de archivos
Actualización de repositorio de datos genómicos	Privado y seguro
	Inserción de gran cantidad de GVD
	Integridad, disponibilidad y durabilidad de los datos
	Búsqueda de información y búsqueda de secuencias
	Validación de datos antes de actualización.
	Permitir clasificación de datos
Desarrollo de software para bioinformática	Procesamiento rápido, eficiente, flexible y escalable
	Protocolos de seguridad
	Acceso al servicio a través de un navegador web
	Análisis en modo background y análisis paralelos.
	Control de errores
	Uso de algoritmos rápidos
	Gestión de accesos
Interpretación de resultados	Sustentar bibliográficamente
	Análisis y comprobación con experimentos anteriores y similares
	Validación en laboratorio
Análisis bioinformático de GV	integración de herramientas de análisis
	Centralizar en un solo ambiente
	Disponible en cualquier momento y lugar
	Recuperar datos en distintos formatos.
	Permitir gestión de reportes
	Múltiples conexiones al sistema simultáneamente
	Permite la interpretación de resultados
	Eficiente comunicación de datos
	Gestionar copias de respaldo
	Infraestructura robusta
	Software actualizado

Fuente: Elaboración del autor

4.4.3 Procesos de negocio seleccionados

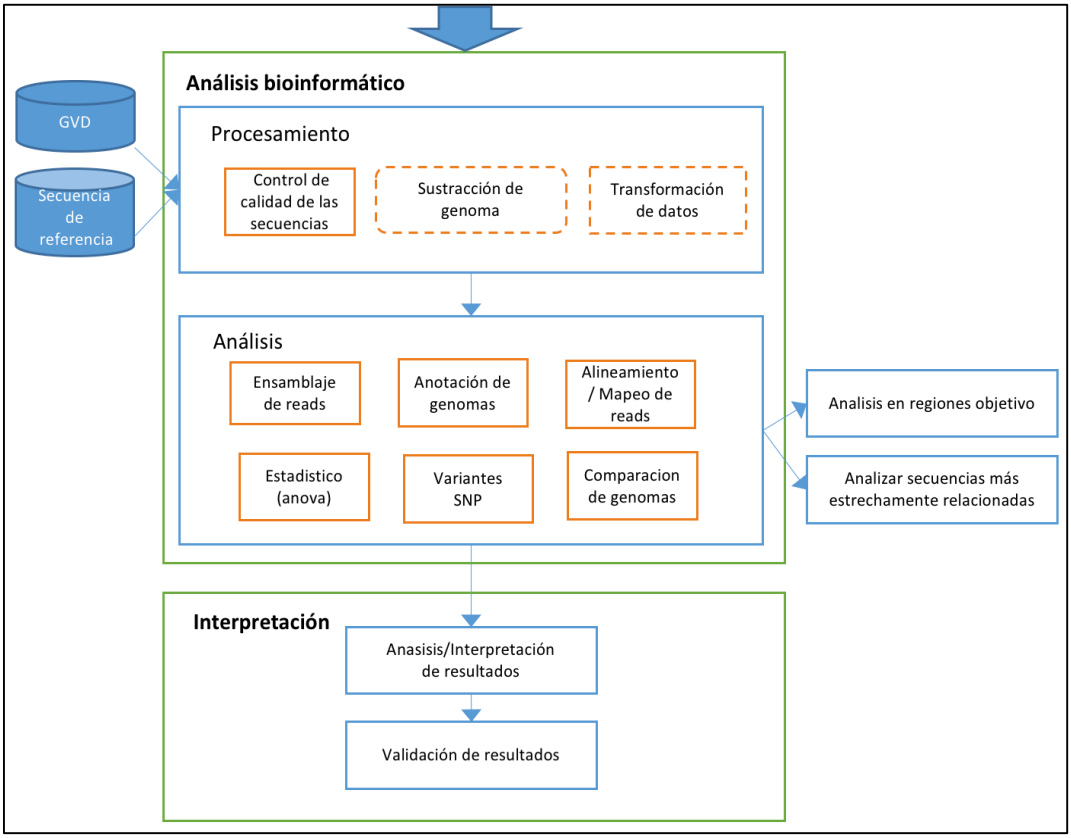
Los centros de investigación y en especial el área de bioinformática, de acuerdo a la naturaleza de su investigación se especializan en cualquiera de los procesos de negocio antes mencionados, pero debido a que el core se

centra en la investigación, se han seleccionado los procesos de negocio más relevantes. Dentro del alcance del trabajo de arquitectura de negocio se incluirán dos de los cinco procesos de soporte a bioinformática, estos son:

- Proceso de análisis bioinformáticos de GVD.
- Proceso de interpretación.

Estos procesos se encuentran directamente relacionados uno con otro. A continuación, se presenta el flujograma general generado como propuesta de arquitectura de negocio para los dos procesos antes mencionados, el cual se muestra en la Figura 22.

Figura 22: Análisis bioinformático e interpretación de resultados



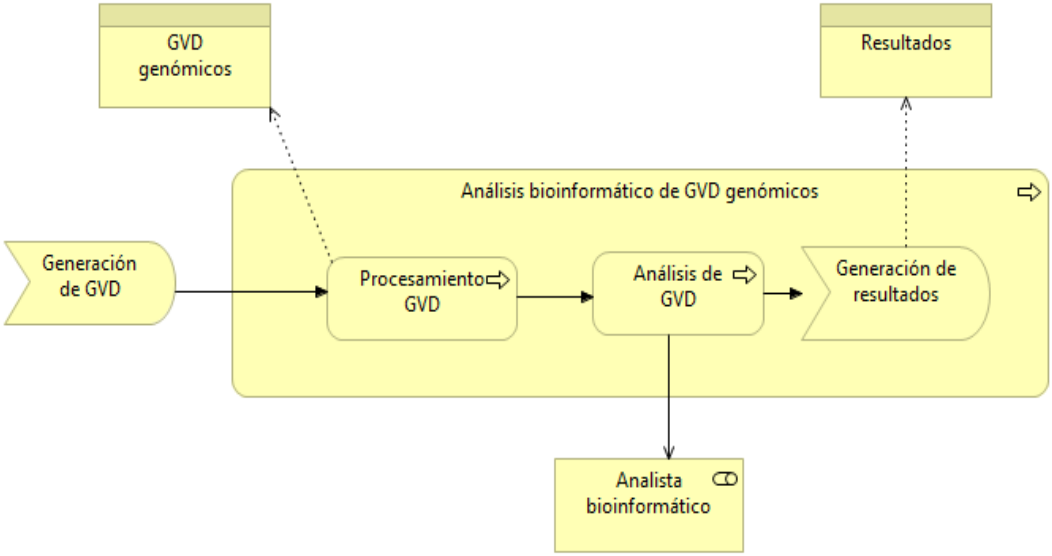
Fuente: Elaboración del autor

En las siguientes secciones se desarrollan en detalle estos principales procesos de negocio para un área de bioinformática, siendo el análisis bioinformático es el objetivo de la investigación, el cual se centra en el análisis bioinformático de GVD y la interpretación de resultados.

4.4.3.1 Proceso de análisis bioinformáticos de GVD

Este es el proceso principal en este estudio, que se centra en los análisis bioinformáticos de los grandes volúmenes de datos, dentro del cual se detallan los servicios bioinformáticos, en la Figura 23 se muestra el diagrama de actividades de este proceso. La entrada consta de los GVD generados a través de las tecnologías NGS, los cuales pasan por los procesos de procesamiento y análisis para generar los resultados. Cabe resaltar que estos procesos pueden resultar repetitivos de acuerdo al estudio que se realice.

Figura 23: Diagrama del proceso de análisis de GVD



Fuente: Elaboración del autor

Las fases del proceso de análisis bioinformático son los siguientes:

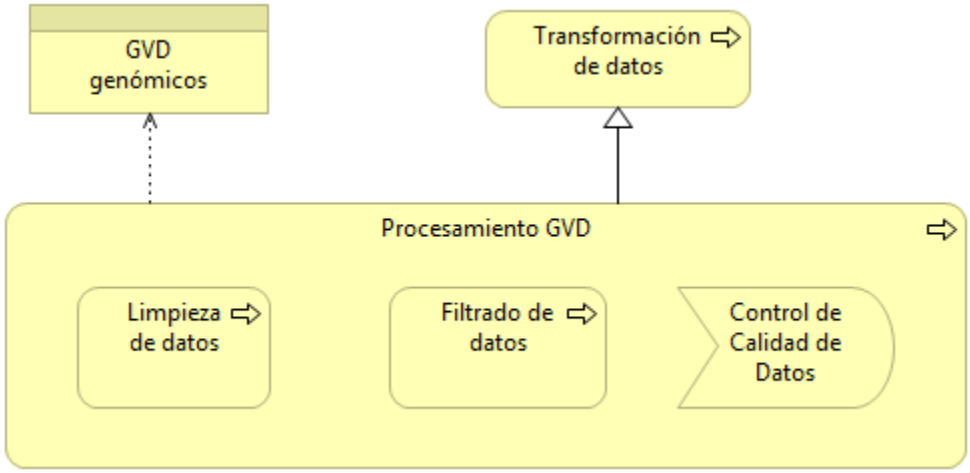
a) Procesamiento

Una vez almacenados los datos, el siguiente paso en un sistema para GVD es explotar estos GVD. El avance tecnológico también trajo consigo el avance en las herramientas de análisis y procesamiento de la información, sobre todo a las que se refieren a los datos no estructurados. Por esto, se vuelve esencial considerar algún tipo de procesamiento distribuido, con el fin de ser capaz de procesar grandes volúmenes de datos, de modo que los

algoritmos de procesamiento pueden ser implementados y ejecutados en paralelo.

El procesamiento actúa como un paso entre el almacenamiento de datos y los análisis (Figura 24). Este paso incluye tanto la limpieza de datos como los pasos opcionales de filtrado de datos y cambio de formato. La limpieza de datos, es un proceso de comprobación y validación de los datos, donde los datos que no son relevantes son eliminados mediante un recorte de las secuencias de baja calidad, otro tipo de limpieza incluye la eliminación de adaptadores y cebadores del PCR si la secuencia los tiene presente, con el objetivo de analizar los datos con mayor precisión.

Figura 24: Procesamiento de datos



Fuente: Elaboración del autor

Los GVD de datos también pueden ser procesados para filtrar características de interés para el investigador o para seleccionar datos objetivos, usualmente los investigadores utilizan *k-mer* que es un valor estadístico como valor para filtro. También es posible adicionar datos históricos, lo cual favorecerá su posterior análisis.

El control de calidad es una medida importante y efectiva para determinar la calidad de las muestras, y también sirve para indicar si las secuencias están correctamente o no, de esta manera se detectan datos corruptos de ser el caso. La transformación de estos GVD crudos en otro

formato o en otro tipo de dato comprimido es opcional, este proceso no implica la perdida de datos, pero es un paso para incrementar la efectividad de los análisis. De ser correcto el proceso, el siguiente paso son los análisis de los grandes volúmenes de datos que se detallan en el siguiente punto.

b) Análisis de GVD

Luego de los pasos previos de procesamiento, se realiza la parte más interesante que es el análisis de datos para extraer el conocimiento contenido en los datos, este proceso usualmente es iterativo hasta que se realice una adecuada validación de resultados. En cierto sentido el análisis bioinformático es la parte principal del modelo, donde todos los procesos analíticos residen, así como los servicios bioinformáticos. El análisis de los GVD hace uso de algoritmos especializados y altamente precisos y cálculos estadísticos que se observan con menos frecuencia en el entorno típico de negocios en general, para deducir leyes (modelos matemáticos, regresiones, las relaciones no lineales, y los efectos causales) de grandes conjuntos de datos para revelar relaciones, dependencias, y para realizar predicciones de los resultados y comportamientos.

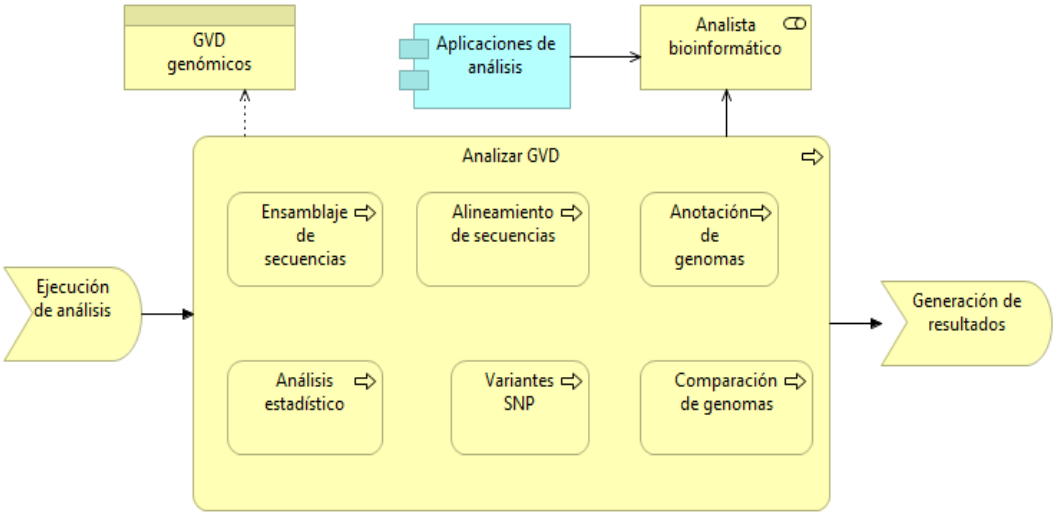
Debido a la complejidad de los análisis y del tiempo que estos tardan, generalmente los investigadores usan flujos de trabajo automáticos, los cuales son a menudo de entrada y salida intensiva, debido al tamaño de los datos de entrada típicamente muchos gigabytes y al incremento de cinco veces del tamaño de los datos en los archivos intermedios.

Para automatizar los análisis es necesario la elección de las aplicaciones que se utilizaran estas varían de acuerdo a los resultados que se deseen obtener. Existen 2 tipos de flujos automáticos, los micro flujos ayudan a extraer partes de un largo análisis como realizar pruebas, verificar datos, etc. Los macro flujos definen un flujo de trabajo completo el cual permite tener el control completo de los elementos de análisis y los resultados a obtener.

Un macro flujo de trabajo de análisis se muestra en la Figura 25 e incluye los siguientes pasos:

- Dentro del paso previo se considera el control de calidad de los GVD, adicionalmente como pasos alternativos es posible realizar la sustracción de algún genoma específico o transformar los datos a un formato diferente del original, realizado este paso se procede a la lectura de los GVD.
- Dentro del análisis se considera los servicios de mapeo de los *reads* al genoma de referencia, en caso de tener una secuencia de referencia se provee como entrada para los análisis, lo cual genera un beneficio adicional al análisis debido a que facilita el análisis. De otro modo, el servicio de ensamblaje de *reads* se realiza sin secuencias de referencia.
- Los otros servicios se realizan utilizando un protocolo específico (por ejemplo, el análisis de expresión de RNA-Seq, variante pidiendo re-secuenciación). En este paso se elige el programa o herramienta a utilizar de acuerdo al tipo de servicio que se desea realizar, para ello hay protocolos establecidos por los investigadores por cada análisis.
- Generación de resultados. Luego de varios ciclos de análisis de ser el caso se llega a resultados que tendrá una nueva verificación de calidad y serán presentados a través de informes.

Figura 25: Flujo de análisis bioinformático



Fuente: Elaboración del autor

Los procesos identificados siguen este macro flujo de trabajo, cuyo resultado de estos análisis pueden crear modelos basados en los datos,

los cuales deben ser analizados por un investigador o un experto, quien es capaz de interpretar y validar los modelos y/o resultados.

En la Tabla 15, se detalla las actividades realizadas de los procesos inmersos en el análisis de grandes volúmenes de datos: Están listadas los detalles en orden jerárquico de acuerdo a la priorización de los procesos, una lista de todas las operaciones suministradas por la especificación, así mismo se detalla las entradas y los datos de salida.

Tabla 15: Actividades de los análisis de GVD

Análisis	Actividad	Entrada	Salida
Mapeo a genomas	Utilizar programa de mapeo Identificación de las características intrínsecas de la secuencia.	GVD Secuencia de referencia	Datos procesados
Análisis de variantes SNP	Identificación de las diferencias de secuencia y variaciones Identificación de polimorfismos de un solo nucleótido (SNP).	Datos procesados Secuencia de referencia	Resultados de regiones a estudiar.
Anotación de genomas	Identifican las porciones del genoma que no codifican para proteínas búsqueda de genes Información biológica se adjunta a las secuencias.	GVD Secuencia de referencia	Secuencias Genomas
Ensamblaje de genomas	Reconstruir un genoma de una especie. Construir secuencias de datos,	GVD	Secuencias Genomas
Análisis estadístico	Análisis ANOVA Análisis de confiabilidad	Resultados parciales	Resultados
Comparación de genomas	Comparan sus estructuras. Verificar funciones.	GVD	Resultados

Fuente: Elaboración del autor

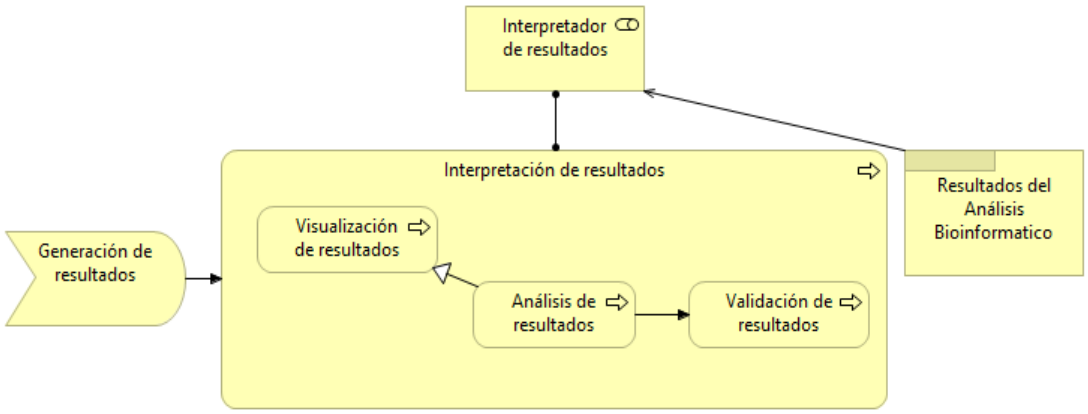
4.4.3.2 Proceso de interpretación

En la Figura 26, se puede observar el diagrama del proceso de interpretación de pre-resultados a partir de los resultados generados de los análisis de GVD realizado por un interpretador de resultados, en este caso un investigador científico utilizando herramientas específicas para tal función, generando así el producto final que es el resultado final del análisis bioinformático.

Los datos recibidos como resultados deben ser analizado para verificar que sean significativos, así mismo deben ser validados a través de estadígrafos, contrastación con literatura existente, así como por la experiencia de los científicos, entre otros. En esta fase los resultados pueden

analizarse y/o visualizarse tantas veces como haga falta y lo requiera el investigador.

Figura 26: Diagrama del proceso de interpretación de resultados

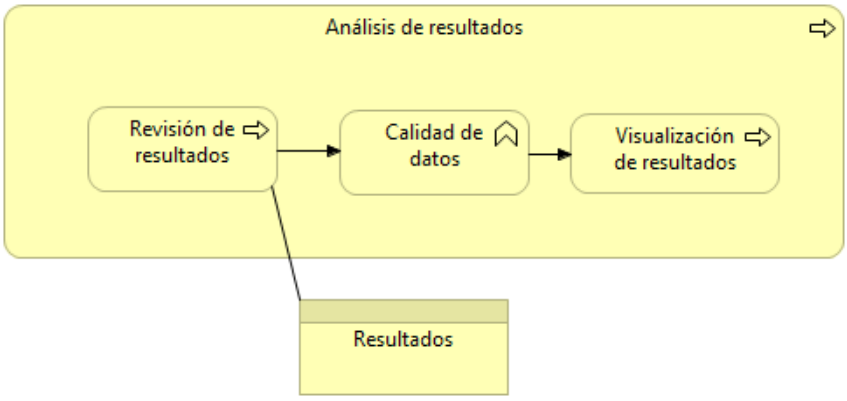


Fuente: Elaboración del autor

a) Análisis de resultados

Para el proceso se tiene una lista de operaciones que responden a las actividades en los procesos. Las operaciones en este nivel son muy generales e implican el uso de más operaciones que aparecerán a medida que se desarrolle la investigación, ver Figura 27.

Figura 27: Análisis de resultados



Fuente: Elaboración del autor

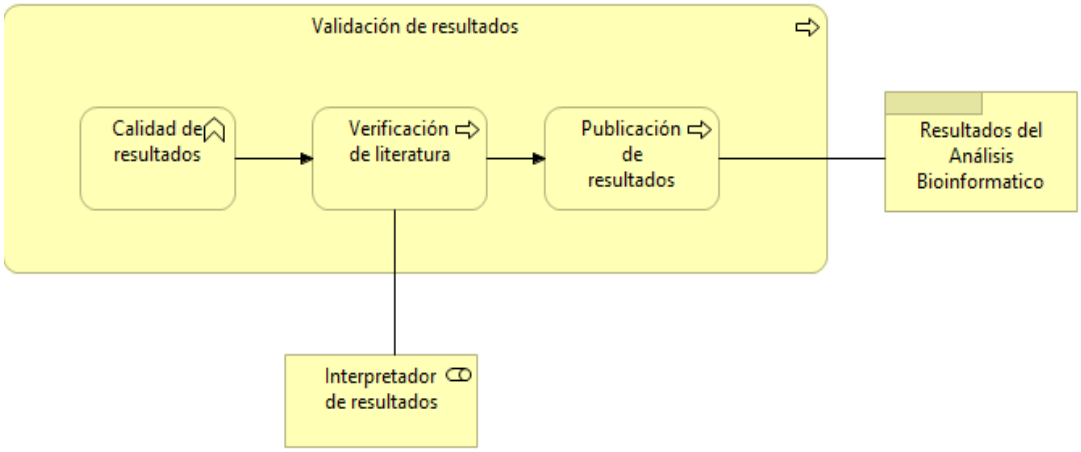
La primera fase consiste en la revisión de los resultados, lo cual incluye la verificación de que los análisis se han realizado y finalizado correctamente, verificación que los resultados se han generado acorde con el tipo de análisis. Posteriormente se realiza la función de revisión de calidad de

datos, donde se verifica que los datos no estén corruptos, estén completos, los resultados deben estar dentro de los límites de cada análisis, así mismo se debe verificar que se hayan ejecutado correctamente los parámetros de análisis. Adicionalmente se realiza la visualización de resultados para un mejor control de calidad de datos.

b) Validación de resultados

Luego se realiza la verificación de la calidad de resultados mediante la aplicación de estadígrafos y la verificación con la experiencia de los investigadores científicos, además se revisa publicaciones e investigaciones relacionadas para la alineación de resultados (ver Figura 28).

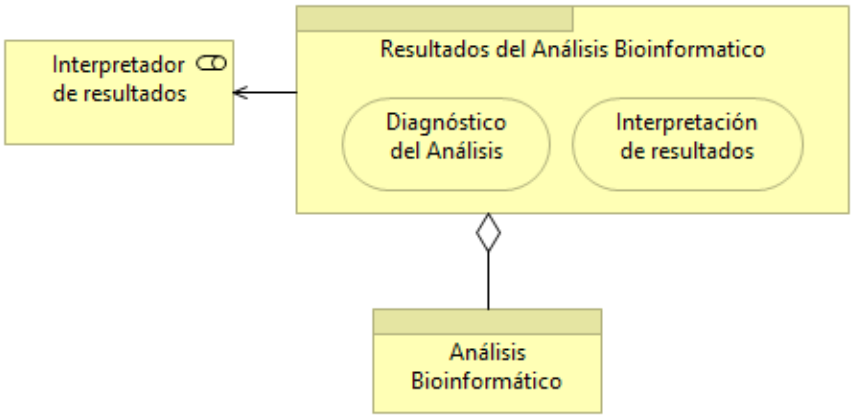
Figura 28: Validación de resultados



Fuente: Elaboración del autor

Finalmente, luego de todos estos pasos se realiza la publicación de resultados finales, realizado por el interpretador de resultados, cuyo producto es el resultado del análisis bioinformático. Luego de los análisis bioinformáticos que se señalaron anteriormente, se entregan los siguientes productos de la Figura 29, cuyos resultados dependiendo del análisis realizado constan de un diagnóstico y la explicación parcial realizada por el bioinformático responsable, para ser entregado al experto del estudio.

Figura 29: Producto de los análisis bioinformáticos



Fuente: Elaboración del autor

4.4.4 *Roadmap* de arquitectura

El *roadmap* de la arquitectura enumera paquetes de trabajo individuales (Tabla 16) y los establece en una línea de tiempo (Tabla 18) para mostrar la progresión de la arquitectura a lo largo del proceso. El roadmap de la arquitectura se detalla a lo largo de las Fases E y F del ADM, cuyo desarrollo está fuera del alcance de esta investigación.

4.4.4.1 Lista de paquetes de trabajo

La lista de paquetes de trabajo se detalla en la Tabla 16.

Tabla 16: Lista de paquetes de trabajo

ID	Paquetes de trabajo	Descripción	Dependencias
A1	Planeación estratégica	Analizar la posición de la organización, identifica su contexto y establecer la visión. Establece un plan operativo customizado para la implementación.	A4
A2	Negocio y procesos.	Definir los procesos y adaptar dentro del marco del modelo propuesto. Definir el flujo de datos a través de los procesos.	A1, A4
A3	Gestión tecnológica	Adopción del modelo tecnológico. Revisión y alineamiento de las aplicaciones.	A2, A1
A4	Recursos humanos	Capacitación de los participantes, para una satisfactoria implementación de la AE.	A1

Fuente: Elaboración del autor

Tabla 17: Gestión de riesgos

Riesgo	Prob. de ocurrencia	Impacto	Estrategia de mitigación	Responsable
El modelo no se adapta a las necesidades	Media	Alto	Realizar la validación del modelo con los investigadores.	Líder de TI
Los datos no se encuentran disponibles	Baja	Alto	Realizar copias de respaldo de los datos.	Administrador del sistema
No se cuentan con recursos económicos.	Baja	Alto	El uso del *cloud computing* minimiza el costo de los recursos.	Director general
No hay aplicaciones disponibles para los tipos de datos	Media	Medio	Las aplicaciones se guardarán en un repositorio común para su uso.	Administrador del sistema
Fallas técnicas	Alta	Bajo	Un pipeline tiene secuencias de operaciones, éstas se pueden volver a repetir en caso de fallo.	Administrador del sistema
Excepciones Temporales	Alta	Bajo	Manejo de excepciones, se resuelven con el tiempo	Administrador del sistema

Fuente: Elaboración del autor

4.4.4.2 Gestión de riesgos

En la Tabla 17 se detalla los riesgos asociados al desarrollo del modelo propuesto, también se muestra la probabilidad de ocurrencia y el impacto que ocasionaría, así mismo se definen las estrategias de mitigación para cada riesgo enunciado. El riesgo con mayor probabilidad de ocurrencia y de mayor impacto es el riesgo de que el modelo no se adapte a las necesidades de los investigadores, por este motivo en la presente investigación el modelo se valida con la participación de los investigadores, siendo el responsable de esta situación el líder de TI.

4.4.4.3 Tiempos de implementación y migración

En la siguiente Tabla 18 se muestra una línea de tiempo para cada paquete de trabajo, cabe resaltar que esta estimación puede variar de acuerdo a la adaptación del modelo de arquitectura empresarial.

Tabla 18: Lista de paquetes de trabajo

ID	Paquetes de trabajo	Año 1 Semestre 1	Año 1 Semestre 2	Año 2 Semestre 1	Año 2 Semestre 2
A1	Planeación estratégica	X			
A2	Negocio y procesos	X	X	X	
A3	Gestión tecnológica		X	X	X
A4	Recursos humanos	X	X		

Fuente: Elaboración del autor

4.5 Sistemas de información

La vista funcional de servicios está representada por el uso de modelo de capas, esto también permite la vista de la perspectiva de la solución. En esta sección se presentará el modelo de datos y el modelo de aplicaciones dentro del marco de una arquitectura de sistemas de información.

4.5.1 Modelo para los datos

Los datos son considerados ahora como el cuarto paradigma de la ciencia, después de los tres primeros paradigmas que son experimentales, teóricos y la ciencia computacional. Este cambio está siendo impulsado por el

rápido crecimiento de los datos y de las mejoras en los instrumentos científicos, los cuales están generando grandes volúmenes de datos con mayor velocidad que en el pasado. Las arquitecturas garantizan accesibilidad, facilidad de uso, uso ético y el intercambio de datos entre las personas, datos compartidos como apoyo para los investigadores, asegurar la integridad de datos, adhesión a las políticas de almacenamiento de datos. La seguridad de los datos incluye, pero no está limitado a políticas de seguridad, administración de vulnerabilidades, cumplimiento de política de datos, el cual no está cubierto al detalle. Aspectos como la confidencialidad, integridad y disponibilidad son variables que se garantiza en la presente investigación.

Los grandes volúmenes de datos, se destacan por el almacenamiento de "datos no estructurados", estos datos simplifican el problema de modelado mediante la eliminación de la falta de correspondencia entre la manera de almacenar los datos y la forma en que se usan. El objetivo de modelado de datos es identificar las aplicaciones o los objetos de negocio y modelos de documentos que se adapten naturalmente a esos objetos.

Por lo tanto, los grandes volúmenes de datos que se utilizan en bioinformática, pueden adquirir varios formatos (Tabla 19) dependiendo del tipo de tecnología que se utiliza para el secuenciamiento, en nuestro caso los GVD generados por Illumina Technologies los formatos obtenidos son:

Tabla 19: Formato de los grandes volúmenes de datos

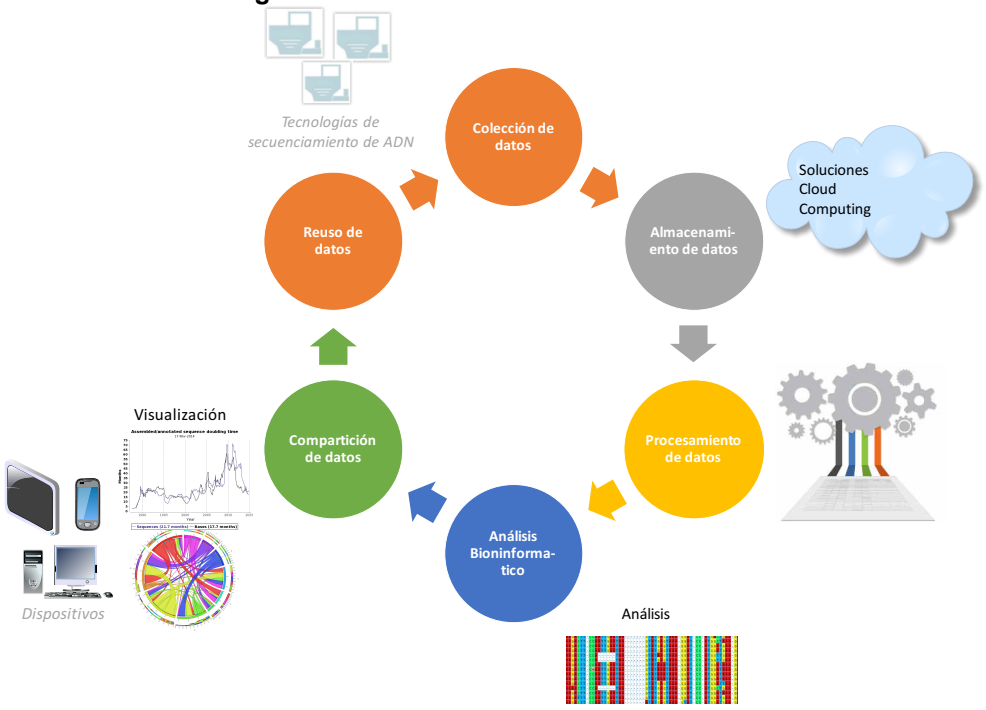
Formato	Descripción	Ejemplo
FASTA	Archivo de texto que contiene secuencias de nucleótidos representadas por símbolos de la IUPAC (A, C, G, T, N).	> HWSW5CG01ESA06 rank=0005835 length=43 GATATCGTGATTGCCAATCAAGTCTTGT
FASTQ	Archivo de texto que contiene secuencias de nucleótidos junto a la calidad de secuenciamiento.	@ K5HV3: 00029: 00029 AAAAAGGGTAAAAACGATCGCTCACAGG + AB>>(44*44:::/:?447444C765@-

Fuente: Elaboración del autor

4.5.1.1 Gestión de datos

La gestión de los grandes volúmenes de datos, se realiza a través de las fases del ciclo de vida de los datos, desde la obtención hasta la publicación y reúso. Los datos se consideran como una unidad en sí misma, separada de los procesos y actividades de negocio. Cada cambio de estado se representa en la Figura 30. La separación de datos del proceso permite identificar requisitos de datos comunes.

Figura 30: Ciclo de vida de los GVD



Fuente: Elaboración del autor

La generación de los grandes volúmenes de datos se realiza con las tecnologías de *Next Generation Sequencing*, esta tecnología nos brinda los datos "crudos", generalmente son obtenido a partir de sitios FTP con los accesos correspondientes para la descarga. Posteriormente se realiza el almacenamiento de estos datos, como parte de la propuesta se incluye el uso del *cloud computing* que se detalla en la sección 4.5.

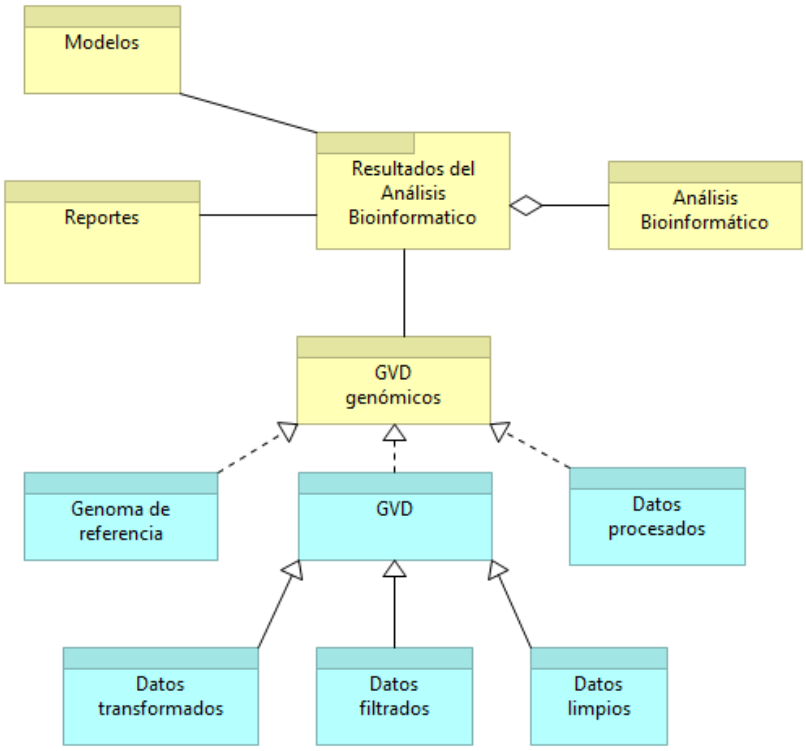
Los datos almacenados, serán procesados y analizados como se describió en la fase de arquitectura de negocio (ver sección 4.3.3). Los resultados serán interpretados y compartidos a través de publicaciones de los

resultados obtenidos. Estos datos pueden ser reusados y usados conjuntamente como otros datos anteriores. El ciclo de vida de los grandes volúmenes de datos es repetitivo y ciertas fases como el procesamiento y análisis pueden realizarse tantas veces como sea necesario.

4.5.1.2 Modelo conceptual de datos

A lo largo de los procesos de negocio (Figura 31) se generan nuevos datos generados de los análisis bioinformáticos, así como de las aplicaciones usadas, a partir de los Grandes Volúmenes de Datos, creando así nuevos tipos de datos, no necesariamente GVD, que se convertirán en los resultados, cuya representación se muestra en la Figura 31.

Figura 31: Diagrama de datos



Fuente: Elaboración del autor

Los datos genómicos que se utilizan en bioinformática son:

- Los genomas de referencia de los individuos en estudio, así como las secuencias de referencia de genes específicos.

- Los grandes volúmenes de datos de los *reads* provenientes del secuenciamiento de ADN/RN de los individuos en estudio, también llamados datos crudos. Los grandes volúmenes de datos pasan por un proceso de control de calidad, en el caso de ser necesario se filtran secuencias de interés. Por otro lado, se pueden extraer secuencias defectuosas o quitar los primers (secuencias de identificación). En otros casos, es necesario cambiar el formato de los GVD para uso para programas específicos, por ejemplo, de FASTQ a FASTA.
- Los datos procesados, que son los generados después del procesamiento y análisis de los GVD. Estos datos se clasifican en datos intermedios y datos finales.

Los resultados generados en bioinformática son los siguientes:

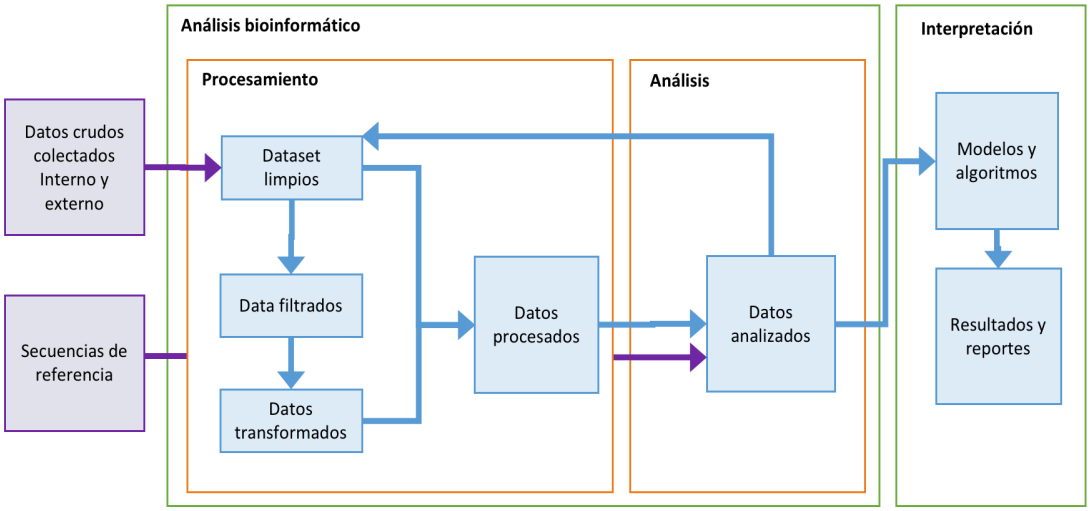
- Resultados de los análisis bioinformáticos, que son los diagnósticos finales luego de la validación por los expertos.
- Reportes que son generados por los programas de análisis, los cuales serán revisados y verificados.
- Los análisis bioestadísticos generan modelos de comportamiento entre otros.

4.5.1.3 Modelo lógico de datos

La capa de datos se beneficia de la propiedad de extensibilidad de la arquitectura, haciendo factible el uso de cualquier almacenamiento de datos. En la Figura 32 se aprecia el flujo de los datos a través de los procesos en estudio. Las acciones realizadas por la capa de datos dependen del tamaño de los datos de entrada, donde la idea es mejorar la eficiencia en la transferencia de datos.

La gestión de datos no estructurados de automatización es clave para proyectos que hace uso de GVD, para nuestro propósito específico y debido a que los datos no estructurados no tienen bases de datos relacionales, se introdujo el uso de metadata, el cual nos permite la identificación de los GVD.

Figura 32: Flujo de los datos a través de los procesos



Fuente: Elaboración del autor

4.5.1.4 Modelo de metadatos

El modelo de metadatos es una definición de estructura de datos genérica e independiente de la plataforma de implementación, establecida por un patrón de diseño para la integración con las aplicaciones. La adopción de un formato de metadato común y estándar para la búsqueda, recuperación, el intercambio de datos y los resultados de análisis es un paso crítico y de gran alcance, que beneficia a los investigadores y no especialistas, permitiendo estandarizadas y transparentes prácticas de análisis reproducibles. Para una mejor gestión de los datos, los GVD deben poseer metadata que incluye el tipo de dato: datos crudos, datos intermedios y datos finales para poder realizar una mejor identificación de cada uno. En la Tabla 20 se describe y detalla la metadata que se emplea en los GVD.

Esta información de metadata debe ir junto con los datos GVD, para poder identificar la información ligada a cada conjunto de datos. Así mismo se crea una metadata para los datos resultantes, como se muestra en la Tabla 21, estos resultados dependen de los proyectos, los análisis y el tipo de resultado que se obtuvo.

Tabla 20: Metadata para GVD

Campo	Valores	Descripción
id	número	Este campo sirve como un identificador único para el conjunto de huellas que se agrupan en esta estructura de datos.
desc	Cadena	Este campo sirve para proporcionar una sencilla descripción del conjunto de datos contenido en el archivo.
attr	cadena	Los atributos de los campos relacionados con el grupo. Esta información podría ser: el día que se completó el experimento, condiciones experimentales exactas, o cualquier otra información que se relaciona con el conjunto de datos.
ubi	cadena	Contiene la ubicación física exacta dentro del bloque de almacenamiento.
md5	cadena	Mantiene un identificador específico. Por defecto se utiliza un hash MD5.
type	cadena	Específica el valor particular sin son datos crudos, datos intermedios o datos finales.
values	file	Contiene los datos reales de un solo set de datos.

Fuente: Elaboración del autor

Tabla 21: Metadata para los resultados

Campo	Descripción
Proyecto	Grupos de las muestras en los proyectos.
Muestra	Representa la muestra biológica de un experimento de secuenciación. Una nueva muestra se crea en cada nueva solicitud de un experimento de secuenciación.
Datos crudos	Registro de archivos FASTQ producidos. Almacena las rutas de acceso a archivos.
Datos alineados	Almacena el software y el genoma de referencia utilizado, así como la ruta de acceso a los datos resultantes alineados (en formato BAM)
Referencia	Secuencia de referencia.

Fuente: Elaboración del autor

4.5.2 Aplicaciones

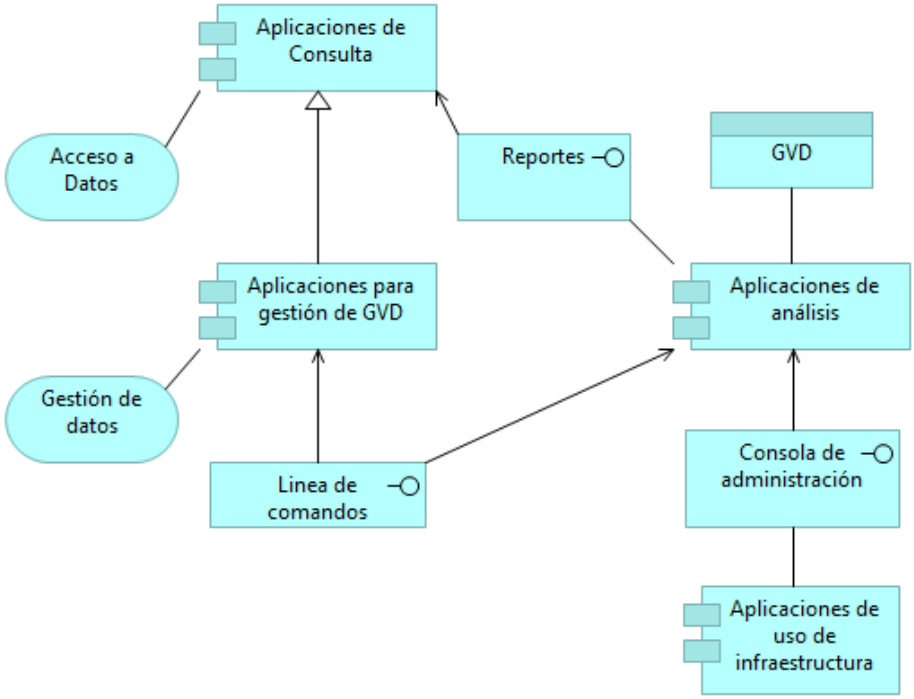
El software bioinformatico es muy diverso, ya que existen herramientas que funcionan por linea de comando, herramientas con interfaz grafica y servicios web para el uso publico. Por ejemplo, BLAST es una herramienta para biologia computacional que permite comparar las secuencias con otras secuencias de una base de datos, los cuales pueden ser secuencias de proteina o de nucleotidos, tanto por linea de comandos, como por interfaz web. El NCBI (*National Center for Biotechnology Information*) tiene una web site muy frecuentado con interfaz para BLAST y que posee las bases de datos de muchos organismos a nivel mundial. “Existen multitud de software bioinformático con otros objetivos: alineamiento estructural de proteínas, predicción de genes y otros motivos, predicción de estructura de proteínas,

predicción de acoplamiento proteína-proteína, o modelado de sistemas biológicos, entre otros. En el software para alineamiento de secuencias y software para alineamiento estructural pueden encontrarse diferentes programas adecuados para cada objetivo en particular que parecen similares" (National Library of Medicine, 2018).

4.5.2.1 Modelo conceptual de aplicaciones

El modelo conceptual de la arquitectura de aplicaciones esta basada en el uso de los datos, debido a que en cada parte del proceso consiste en el uso de estos datos, se contruye el diagrama de las aplicaciones que se utilizan para investigacion en bioinformatica, en la Figura 33.

Figura 33: Aplicaciones para el negocio de bioinformática

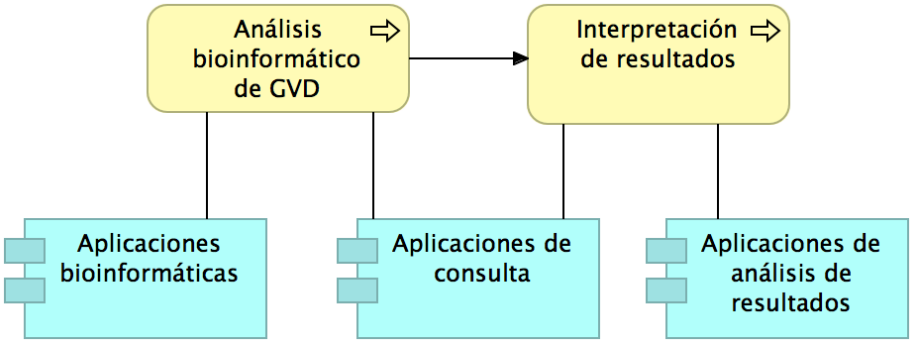


Fuente: Elaboración del autor

Los investigadores o biólogos pueden usar ciertas aplicaciones para la gestión de los grandes volúmenes de datos debido a que es un paso previo e importante para el proceso principal de análisis bioinformático de GVD. Todas las aplicaciones de bioinformática se llevan a cabo en el nivel de usuario. Los investigadores usan ciertas aplicaciones y luego utilizan el resultado en su investigación. Estas aplicaciones se definieron apartir de los

procesos, por ende se muestra la relación entre los procesos en estudio y las aplicaciones que se detallan en la arquitectura, ver Figura 34.

Figura 34: Aplicaciones a partir de los procesos



Fuente: Elaboración del autor

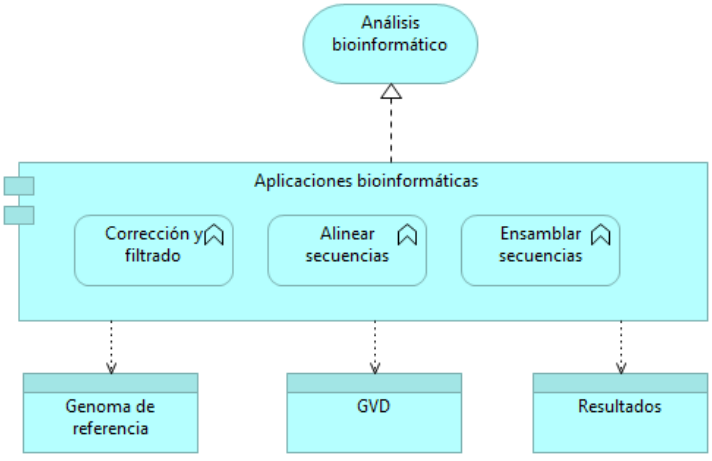
4.5.2.2 Modelo lógico de aplicaciones

El modelo lógico de aplicaciones se muestra para los dos procesos de negocio el análisis bioinformático de GVD y el análisis de los resultados.

a) Aplicaciones para análisis bioinformático de GVD

Las aplicaciones para el análisis bioinformático se utilizan para la corrección de los GVD, el filtrado de datos y secuencias objetivo, así mismo para el alineamiento y el ensamblaje de secuencias. Para tales tareas se utilizan genomas de referencia y los grandes volúmenes de datos como datos de entrada para obtener los resultados, ver Figura 35.

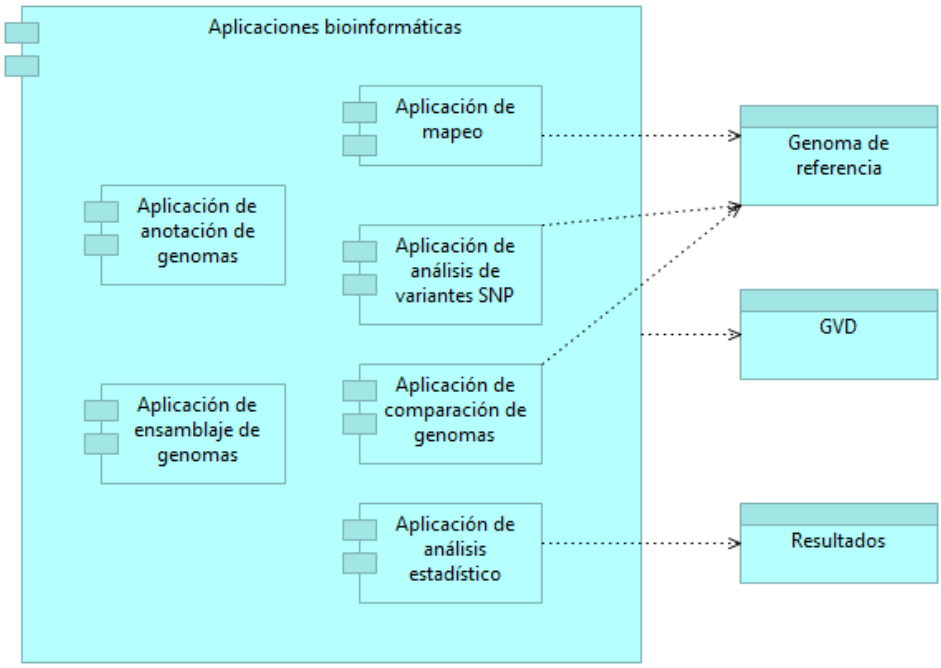
Figura 35: Comportamiento de las aplicaciones



Fuente: Elaboración del autor

Estas aplicaciones de análisis bioinformático proporcionan a los investigadores formas fáciles de construcción de *pipelines* para la resolución de problemas. Los flujos de datos antes estudiados, se pueden componer en *pipelines* de análisis de datos, y también puede manejar el procesamiento paralelo basado en la partición de datos. Por esto, los investigadores pueden utilizar *pipelines* de bioinformática con el fin de desarrollar un programa paralelo de manera eficiente y sencilla a partir de estas aplicaciones. Estas aplicaciones bioinformáticas están compuestas de diferentes aplicaciones de acuerdo a la naturaleza del análisis como se muestra en la Figura 36.

Figura 36: Estructura de aplicaciones



Fuente: Elaboración del autor

Las secuencias de referencias son necesarias para el caso de aplicaciones de mapeo, análisis de variantes SNP y comparación de genomas. Las aplicaciones que se utilizan según el tipo de análisis de los grandes volúmenes de datos se muestran en la Tabla 22, estos pueden ser análisis de secuencias, análisis evolutivo, análisis estructural y análisis de alto rendimiento, descritos en la parte teórica. La escala define la complejidad del análisis y se listan las aplicaciones por los tipos de aplicaciones.

Tabla 22: Herramientas para análisis bioinformático

Aplicación	Funciones	Escala	Ejemplo
Aplicación de mapeo a genomas	Múltiples alineamientos de secuencias. Mapeo a genomas.	Alta	BWA, SOAP, Bowtie, MAQ Blast
Aplicación de análisis de variantes SNP	Análisis de SNPs pequeñas inserciones y deleciones e indels.	Pequeña	BWA, Picard VCFtools, GATK, SAMtools
Aplicación de anotación	Análisis estructural de proteínas y ARN.	Pequeña	m5nr Trinity
Aplicación de ensamblaje de genomas	Ensamblaje de secuencias cortas.	Alta	Velvet Soap-de novo Abyss, Trinity
Aplicación de análisis estadístico	Análisis bioestadístico de datos crudos y resultados.	Pequeña	BioPerl BioConductor R projects
Aplicación de comparación de genomas	Detección de secuencia homólogos de distancia / ortólogos	Mediana	LGA STRALCP Garnier EMBOSS

Fuente: Elaboración del autor

Tabla 23: Catálogo de aplicaciones bioinformáticas

Aplicación	Descripción	Input	Output
Cufflinks	Ensamblaje transcripción, la expresión diferencial y la regulación diferencial de ARN-Seq.	FASTQ FASTA GFF	SAM BAM
TopHat	Empalme rápido para unión de los *reads* de ARN-Seq. Fusión de un algoritmo para el descubrimiento de nuevas transcripciones.	FASTQ FASTA	BAM BED
Bowtie	Alineador ultrarrápido de memoria-eficiente de *reads* cortos.	FASTQ FASTA	SAM
MAQ	MAQ significa mapeo y ensamblaje con calidad que construye ensamblaje por mapeo de secuencias cortas con referencia de otras secuencias.	FASTQ FASTA	MAQ
GATK	Kit de herramientas con un enfoque principal en descubrimiento de variantes y genotipo, así como garantía de calidad de los datos.	VCF	SAM BAM
BWA	Programa eficiente que alinea las secuencias de nucleótidos relativamente cortos contra una secuencia de referencia larga, tales como el genoma humano que implementa dos algoritmos, BWA-corto y BWA-SW.	FASTQ FASTA	SAM
Picard	Un conjunto de herramientas (en Java) para trabajar con datos de *Next Generation Sequencing* en BAM.	SAM BAM	SAM
Trinity	Posee un método para la reconstrucción eficiente y robusta de transcriptomas a partir de datos ARN-Seq.	FASTQ FASTA	FASTA
Blast	Basic Local Alignment Search Tool encuentra las regiones de similitud local entre secuencias.	FASTQ FASTA	TXT
Velvet	Velvet es un ensamblador de genómico especialmente diseñado para las tecnologías de secuenciación de secuencias cortas.	FASTQ FASTA	FASTA

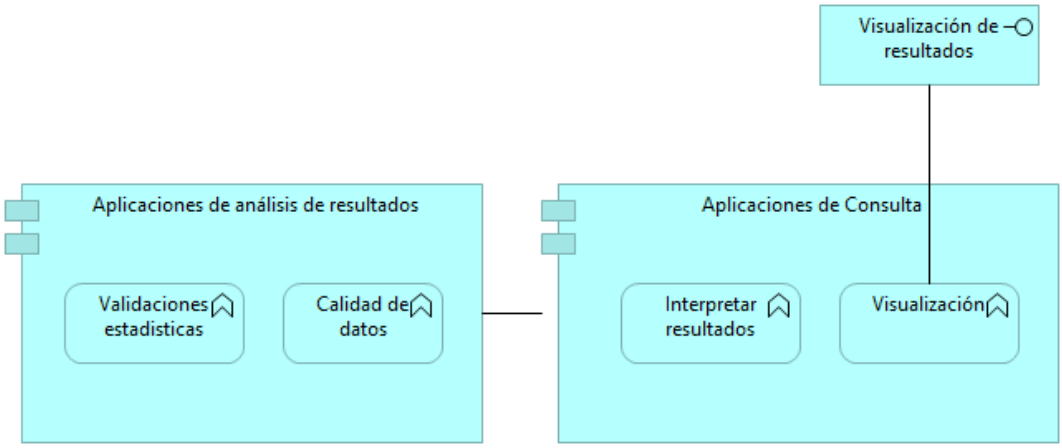
Fuente: Elaboración del autor

En la Tabla 23 se lista algunas de las aplicaciones para bioinformática más utilizadas, así mismo como los requisitos de entrada, que incluye los GVD y los tipos de archivos generados. Se presenta un paquete de herramientas que proporciona la solución completa para el análisis de datos de secuenciación de última generación NGS y también se aplica una variedad de software de código abierto, con el fin de proporcionar un análisis bioinformático integral para los servicios de secuenciación.

b) Aplicaciones para interpretación de resultados

Las aplicaciones para interpretación son herramientas auxiliares que permiten la correcta validación e interpretación de resultados. Las aplicaciones de análisis de resultados y las aplicaciones de consulta se usan a través de interfaces que permiten la visualización de resultados (Figura 37).

Figura 37: Interpretación de resultados



Fuente: Elaboración del autor

c) Aplicaciones de consulta

Estas herramientas permiten ver el listado de los datos almacenados, así como las características principales de los datos como el tamaño de archivo, el formato, la fecha de creación, el estado se encuentra, permisos, etc. Por otro lado, también permiten guardar un registro de las acciones realizadas por el investigador y permiten la interpretación de resultados. El modelo usa s3cmd para consultas del almacenamiento, así como monitor de Amazon EMR.

d) Aplicaciones de análisis de resultados

Estas herramientas permiten verificar el resultado según el tipo de análisis bioinformático realizado, estas aplicaciones participan directamente durante la limpieza y el filtrado, las validaciones estadísticas, la verificación de la generación y conversión de resultados finales de los análisis de calidad. Generalmente son scripts o funciones customizados para cada tipo de análisis.

e) Aplicaciones de visualización

Estas herramientas permiten visualizar estadísticas y datos pre procesados, filtrados, limpios y ver gráficamente los resultados obtenidos luego de los análisis de los GVD, lo cual nos permite la extracción de conocimiento, estas aplicaciones son muy diversas y son usadas a criterio del investigador en la fase de interpretación. Tal es el caso de Tablet o DNAstar.

4.6 Arquitectura tecnológica a utilizar

Para nuestro propósito, la arquitectura tecnológica tiene como preocupación el volumen de información de los GVD que los investigadores tienen en bioinformática. En el nivel inferior de la arquitectura se trata de un problema de TI de los sistemas de archivos y fiabilidad de almacenamiento de gran cantidad de datos, así mismo como los recursos para el procesamiento de estos grandes datos, cuya solución no es obvia y no es única. El verdadero desafío es tener acceso a dichos GVD de manera eficiente y rápida, aplicando paralelismo masivo para disminuir los tiempos de cálculo.

4.6.1 Requerimientos

En tal caso para el desarrollo de esta tesis, se parte de varios requerimientos críticos que se han asumido como requisitos de diseño arquitectónico, tales como escalabilidad, rendimiento, flexibilidad, soporte para añadir nuevas funcionalidades entre otros. En la Tabla 24 muestra la relación entre los requerimientos y los atributos de calidad. Donde, los atributos de calidad críticos son la disponibilidad y la integridad, por lo cual la tecnología utilizada tiene como requerimientos principales estos atributos.

Tabla 24: Árbol de calidad

Atributo	Refinamiento	Requerimiento	Importancia	Riesgo
Performance	Latencia de data	El acceso y procesamiento de los GVD debe ser lo más rápido posible.	A	M
Flexibilidad	Cambios en las reglas de servicios	La tecnología podrá cambiarse en el caso de nuevas funcionalidades o servicios.	A	M
Portabilidad	Despliegue en otro SO.	El sistema tiene que poder ser desplegado en otro servidor Linux.	M	M
Escalabilidad	Crecimiento del Sistema	La infraestructura debe soportar el aumento de requerimientos para nuevas funcionalidades y/o servicios.	A	M
Seguridad	Confidencialidad	El sistema tiene que implementar mecanismo de seguridad.	A	M
	Disponibilidad	La infraestructura debe estar siempre disponible.	A	A
	Integridad	Datos originales no se modifican.	A	A

Fuente: Elaboración del autor

4.6.1.1 Decisiones a partir de los requerimientos

La Tabla 25 muestra las decisiones en el diseño de la arquitectura que se tomó a partir de los requerimientos no funcionales.

Tabla 25: Decisiones a partir de requerimientos

Atributo de Calidad	Requerimientos	Decisión
Performance	El procesamiento debe ser en el menor tiempo posible.	Soporte para computación distribuida: Clústeres HPC.
	El acceso a los archivos debe realizarse en el menor tiempo posible.	Soporte para almacenamiento de datos S3
Flexibilidad	Permite variantes de diseño y la modificación en la estructura del modelo.	Amazon Web Service como proveedor de Infraestructura como servicio
Portabilidad	El flujo de trabajo puede ser desplegado en cualquier plataforma.	Pipelines para el flujo de trabajo. Imágenes del SO para acelerar el aprovisionamiento.
Escalabilidad	Debe soportar el aumento de nuevas funcionalidades y/o requerimientos para cada uno de los servicios.	Implementación del modelo en capas. Uso de servidores escalables: Amazon EC2.
	El modelo es funcional para sistemas a gran escala y permite el aumento del número de recursos que componen la arquitectura.	Uso de *framework* que administre la escalabilidad: -Amazon EMR, Auto-scaling Soporte para sistemas distribuidos como Hadoop
Seguridad	El sistema tiene que implementar mecanismos de seguridad para otorgar confidencialidad.	Componente de seguridad propio del *cloud computing*: Single Sign On. Certificación de seguridad del proveedor ISO 27001

	La infraestructura debe estar siempre disponible.	Acuerdo del Nivel de Servicio del Proveedor (SLA).
	Los datos originales no se modifican.	Alto nivel de redundancia en el almacenamiento. Soporte para almacenamiento.
Disponibilidad	Manejar procesos simultáneos en cualquier momento.	Herramientas de monitoreo: Monitor Amazon EMR.

Fuente: Elaboración del autor

4.6.1.2 Definición de las características evaluadas

En la Tabla 26 se presentan las características utilizadas de AWS (AWS, n.d.) a partir de las decisiones, y sus descripciones correspondientes.

Tabla 26: Servicios de AWS a utilizar

Característica	Descripción	AWS
Escalabilidad automática (*auto-scaling*)	AWS posee un monitor de recursos que permite aumentar o disminuir la cantidad de recursos asignados de acuerdo a los requerimientos.	*Auto-scaling*
Imágenes para acelerar el aprovisionamiento	Las imágenes son máquinas virtuales pre configuradas con sistemas operativos y aplicaciones o con marcos de trabajo instalados y pre configurados, permite focalizar en la construcción o despliegue de las aplicaciones.	AMI
Soporte para almacenamiento de datos	Ofrece un espacio en las plataformas que permiten guardar la información y presérvalas en un largo periodo de tiempo.	S3
Soporte para sistemas distribuidos	Software que permite distribuir el procesamiento de grandes volúmenes de datos con el uso de un grupo de servidores básicos.	Hadoop EMR
Soporte para tecnologías de GVD	La tecnología para GVD hace referencia a los sistemas que manipulan data sets. Las dificultades se centran en buscar, almacenar, compartir, análisis y visualizar.	S3 EMR

Fuente: Elaboración del autor

Una vez definidos las características y servicios de la arquitectura, es posible crear el diseño de la arquitectura tecnológica. Los riesgos asociados a las decisiones arquitectónicas de tecnología son los siguientes:

- Integridad en la transmisión de los GVD.
- Difícil manejo los servicios en el *cloud computing*.
- El componente para la seguridad no cubre las expectativas.

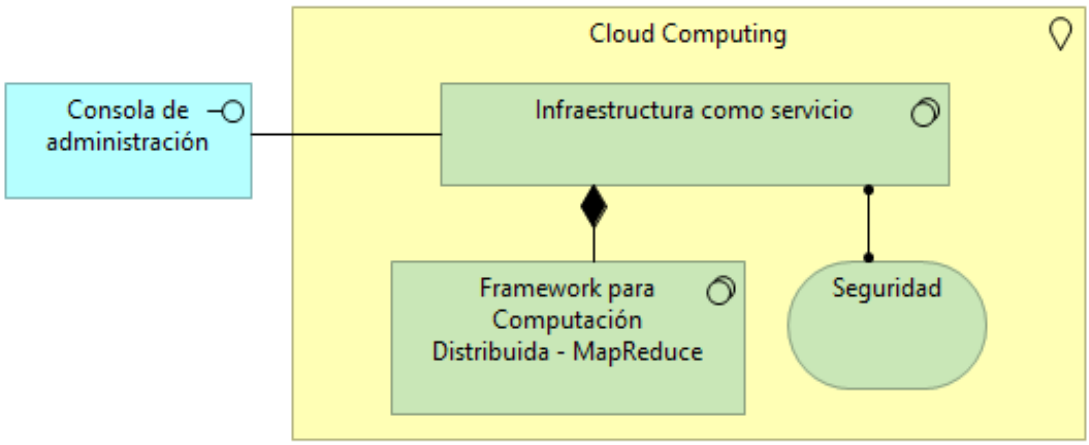
4.6.2 Modelo conceptual de la arquitectura

La segunda tendencia de innovación (Zhang, Big Services Era: Global Trends of Cloud Computing and Big Data, 2012), es proporcionar plataformas de nube abierta por los principales proveedores de servicios. En tal sentido la

propuesta se alinea con esta tendencia y utiliza la arquitectura tecnológica para soportar la arquitectura de aplicaciones y la arquitectura de datos, que apoyan la arquitectura de negocios. El reto consiste en aprovechar los datos con infraestructura elástica y flexible para derivar ideas sin pasar tiempo desarrollando infraestructura compleja.

La propuesta intenta encontrar un ajuste entre los requerimientos de portabilidad, performance, escalabilidad, disponibilidad, flexibilidad, puntos de sensibilidad, seguridad, compensaciones y riesgos descritos previamente siendo los más críticos disponibilidad y escalabilidad. Un sistema completo se puede modelar mediante la integración de los sistemas básicos en el modelo, cada uno con el correspondiente nivel de detalle que proporciona un alto nivel de flexibilidad. El detalle define el nivel de abstracción usado para implementar los componentes del modelo. En esta investigación, el nivel de detalle en la arquitectura tecnológica es lo suficientemente rápido como para modelar grandes sistemas y aplicaciones para el análisis bioinformático de GVD. En la Figura 38 muestra el flujo de trabajo de los componentes principales.

Figura 38: Componentes de la arquitectura tecnológica



Fuente: Elaboración del autor

La propiedad de flexibilidad de la arquitectura permite el soporte para cualquier tipo de computación distribuida de Clúster en el *cloud*. La capa tecnológica recibe de la capa superior las tareas de los pipelines que se ejecutan y una referencia a los datos de entrada.

$$\text{Pipelines de análisis para datos NGS} \quad + \textit{Cloud}\ \text{IaaS} = \quad \text{Flexible, escalable y fácil análisis bioinformático}$$

4.6.2.1 Cloud computing

La adopción generalizada de secuenciación de alto rendimiento ha aumentado considerablemente la escala y la complejidad de la infraestructura computacional necesaria para realizar la investigación genómica. Asi, la transformación de las secuencias en información biológicamente significativa requiere una infraestructura computacional. El tamaño de la infraestructura computacional requerido supera las expectativas de la infraestructura local tradicional. Así mismo, la configuración y el mantenimiento asociado con una infraestructura computacional local presentan problemas significativos para los investigadores individuales. Una alternativa a la construcción y mantenimiento de la infraestructura local es el *cloud computing*, que ofrece acceso bajo demanda a la infraestructura computacional flexible.

Cloud computing trabaja con servidores virtuales llamados instancias. Una instancia de la nube se comporta como un servidor local, excepto que ofrece los beneficios de la disponibilidad de recursos de *cloud computing* y el modelo de *pay as you go* de propiedad de los recursos. Tener un acceso sencillo a la nube permite tener tantas instancias como sea necesario adquirir y comenzar a procesar los GVD. Una vez que la necesidad desaparece, esas instancias pueden ser liberadas tan simples como se adquirieron. Con este paradigma, se paga sólo por los recursos que necesitan y su uso, mientras que todas las otras preocupaciones y gastos son eliminados. *cloud computing* puede también ser una solución para los recursos permanentes, donde los conjuntos de datos grandes pueden ser archivados y de fácil acceso sin necesidad de copiar a otros recursos locales. Para el modelo propuesto, se plantea el uso del *cloud computing* como un recurso no permanente, debido a que su ciclo de vida tiene la duración del tiempo de un análisis bioinformático.

A continuación, se detalla la comparación de proveedores, debido a que la propuesta está basada en esta elección:

a) Comparación de proveedores *cloud computing*

Incentivado por los posibles grandes beneficios de *pay-as-you-go*, hay varios proveedores de *cloud computing* y se prevé que en los próximos

años habrá más proveedores. Actualmente, el proveedor más grande es Amazon, que ofrece las nubes comerciales para el procesamiento de grandes volúmenes de datos. Además, Google también proporciona una plataforma en la nube que permite a los usuarios desarrollar aplicaciones web, y almacenar y analizar los datos. Sin embargo, los proveedores comerciales actualmente no están aún en condiciones de proporcionar gran cantidad de datos y software para el análisis de la bioinformática. Por otra parte, es muy difícil para los proveedores comerciales mantener el ritmo de las necesidades emergentes de la investigación académica, en consecuencia, se requiere *cloud computing* que sean flexibles y permitan realizar los estudios de bioinformática. La mayoría de los proveedores están ofreciendo "ampliación" y funciones de "reducción", es decir, la capacidad de aumentar dinámicamente almacenamiento, memoria RAM y CPU en un solo servidor, así como la capacidad de clonar fácilmente los servidores basados en imágenes predeterminadas.

Se realizó una tabla comparativa entre los principales proveedores de *cloud computing*. La Tabla 27 es una comparación de algunas de las características y capacidades actuales, en función de los requerimientos. Se basó en la información disponible al público en el sitio web de los proveedores al momento de la investigación. Para efectos de comparaciones en costo se tomó como referencia un servidor Linux de 30 GB de memoria y 8 CPUs. Hay una amplia variación en los rangos de precios, por debajo de US $50 a más de US $500, donde en general hay una reducción generalizada de los precios. Esto es evidencia del aumento de la competencia en este espacio, así como las economías de escala que los grandes proveedores de IaaS disfrutan y que pueden dar a sus clientes.

Amazon Web Services (AWS) es en un líder del mercado y ofrece valores interesantes como demanda; *pay-as-you-go;* elástico; programable; y en muchos casos mejor seguridad que sus competidores.

Tabla 27: Comparación de proveedores cloud computing

Proveedores	Características Cloud			Preocupaciones de los usuarios		
	Reducciones de costes / optimizaciones	Escalabilidad y Automatización	Seguridad	Elección y Flexibilidad	Portabilidad	Confiabilidad y disponibilidad
	Variedad de Planes de Precios	Precio Promedio Mensual	Escalabilidad	Monitoreo	Sistemas distribuidos	Certificación	Centros de datos	Tipos de instancia	Sistemas operativos compatibles	Estándares	VM Upload (AMI)	Tiempo de servicio	SLA
Amazon (EC2)	3	$0.33	Si	Extensivo	Si	Si	9	37	4	Propietario	Si	+5 años	99.95%
Rackspace	1	$0.36	Si	Extensivo	Si	Si	6	22	3	OpenStack	Si	+10 años	99.90%
Microsoft	1	$0.44	Si	Promedio	Si	Si	9	16	3	Hyper V	Si	+4 años	99.95%
Google	3	$0.38	Si	Promedio	Si	Si	3	15	3	Propietario	No	+4 años	99.90%
HP	1	$0.90	Si	Promedio	Si	Si	1	11	3	OpenStack	No	+2 años	99.95%

Fuente: Elaboración del autor

Amazon Web Services ofrece una gama completa de ofertas de almacenamiento y computación, incluyendo instancias bajo demanda y

servicios especializados, tales como el almacenamiento persistente escalable como Amazon Elastic Block Store (EBS); almacenamiento escalable como Amazon S3; recursos elásticos bajo demanda como Amazon Elastic Compute Cloud (Amazon EC2); y herramientas como Amazon Elastic MapReduce que junto con Hadoop permite la distribución de grandes cantidades de carga de trabajo. Además, AWS ofrece servicios de infraestructura como flujos de trabajo, el paso de mensajes, almacenamiento prolongado de archivos, los servicios de búsqueda, bases de datos relacionales y No SQL y mucho más. AWS posee Data Center en Estados Unidos, Canadá, Reino Unido, China, Alemania, Francia, Corea del Sur, Irlanda, India, Brasil, Australia, Singapur.

En general, AWS cumple con los requisitos de portabilidad, seguridad, performance, escalabilidad, disponibilidad, flexibilidad de manera igual o superior que sus competidores y además de ofrecer un precio muy inferior a los demás en cuanto al uso de los servidores virtuales, naturalmente los costos de almacenamiento no se incluyen en esta estimación debido a que depende a la cantidad de GVD.

Finalmente, en el Magic Quadrant de Gartner (Gartner, Inc., 2017) en referencia a la infraestructura como servicio en el *cloud computing* de 2018, Gartner situó a Amazon Web Services entre los líderes del cuadrante y clasificó a AWS como poseedor de la mayor integridad de visión y capacidad de ejecución por algunos años consecutivos.

b) Componente de seguridad y contingencia

Los sistemas funcionan sobre la infraestructura del *cloud computing*, en tal sentido la responsabilidad se comparte al momento de definir la seguridad. Por un lado, AWS se hace responsable de la infraestructura, y por otro el usuario se hace cargo de todo lo que se utilice o conecte con la infraestructura. Se debe configurar de acuerdo a la cantidad de los datos y la importancia de los servicios que fueron seleccionados.

En los servicios IaaS como Amazon EC2 y Amazon S3, se tiene mas trabajo de configuración de seguridad, ya que se tiene más control sobre

estas. En Amazon S3 se tiene que crear normativas quienes tiene acceso al almacenamiento e incorporar técnicas de cifrado de datos para la transferencia y crear *backups*.

Los proveedores de *cloud computing* como AWS tienen la certificación ISO 27001, ya que tienen medidas que les ayudan a proteger su infraestructura, además se validó con éxito como proveedor de servicio de Nivel 1 conforme al Estándar de seguridad de datos (Data Security Standard, DSS) del sector de tarjetas de pago (Payment Card Industry, PCI).

4.6.2.2 *Framework* para computación distribuida

En lugar de utilizar un servidor superpotente, se opta por distribuir el trabajo en pequeños paquetes de tareas que se van enviando a nodos utilizando computación distribuida, de modo que se aprovecha el tiempo de reposo de todos los ordenadores conectados a una red. Y luego se recopila los resultados de todos los nodos y genera un informe final del análisis.

El modelo propuesto se centra en el uso de programas y herramientas que utilicen computación distribuida debido a que favorece:

- Disponibilidad. Permite el reemplazo de un nodo averiado.
- Costo-efectividad. La reducción del tiempo de análisis y la optimización de los recursos bioinformáticos de un centro son las principales ventajas de esta tecnología, destinada a cálculos que pueden tardar varios días en completarse y que implican enormes cantidades de datos.
- Escalabilidad. Clúster de computación puede escalar a sistemas muy grandes. cientos o incluso miles de máquinas pueden conectarse en red para satisfacer las necesidades de la aplicación.

AWS permite aumentar la velocidad de investigación ejecutando la informática de alto rendimiento en el *cloud computing* y reducir costes al ofrecer informática en clúster bajo demanda sin grandes inversiones de capital. Accede a una red con gran ancho de banda para cargas de trabajo intensivas de E/S y altamente acoplada, permitiendo escalar en miles de

núcleos para aplicaciones orientadas al rendimiento, lo cual favorece los análisis de GVD. El paralelismo y la distribución de datos mejora mediante:

- Procesamiento de datos a través de un clúster de servidores a través de un planificador de trabajo.
- Gestión de datos a través de un flujo de trabajo optimizado que se detalla en el pipeline.

Un buen planificador de trabajo no causa sobrecarga de virtualización. Sin embargo, cada nodo tiene que gestionar el número disponible de sus núcleos, ya que las aplicaciones necesitan compartir recursos al azar. Por esta razón, el desarrollo del planificador de trabajo es complicado y requiere habilidades de programación de sistemas muy avanzados. Por lo tanto, la tesis se desarrolla en la parte superior de la infraestructura virtual. En tal sentido, el modelo contempla el uso de Amazon Elastic MapReduce (EMR).

a) *Map reduce*

El marco *MapReduce* permite un procesamiento escalable y análisis de grandes conjuntos de datos mediante la distribución de la carga computacional en los nodos informáticos conectados, referido como un clúster. En Bioinformática, MapReduce se adopta a escenarios de la próxima generación de secuenciación de los datos (NGS) como el mapeo de un genoma de referencia, la búsqueda de SNPs en *reads* cortos o cadenas similares en los archivos de genotipo. Sin embargo, las tareas como la instalación y el mantenimiento de MapReduce en un sistema de clúster, la importación de datos en su sistema de archivos distribuido o ejecutar programas MapReduce requieren conocimientos en informática avanzada y por lo tanto se utiliza Hadoop para el fácil manejo de MapReduce.

Clústeres de máquinas independientes se usan para escalar más arriba en la cartografía de lectura (*read mapping*), como una reacción al cada vez mayor grado de paralelismo en la tecnología de secuenciación. La mayoría de ellas aprovechan el paradigma de programación MapReduce para procesar problemas paralelos con conjuntos de GVD. La característica

principal de MapReduce es que administra de manera eficaz y robusta la paralelización y distribución de datos de entrada y el código de usuario en un gran número de nodos de computación. Claramente, la alineación de *reads* es un problema paralelo, ya que cada *reads* se puede alinear de forma individual a la referencia sin dependencias de otro *reads*.

"Amazon EMR es un servicio web que facilita el procesamiento rápido y rentable de grandes cantidades de datos. Amazon EMR utiliza Hadoop, un *framework* de código abierto, para distribuir los datos y el procesamiento en un clúster de instancias de Amazon EC2 de tamaño variable. El rendimiento que entrega Amazon EMR va a depender del tipo de instancias de EC2 que se elijan para formar el clúster. Amazon EMR da la capacidad de mover grandes cantidades de datos de manera rápida de Amazon S3 hacia HDFS (Hadoop Distributed File System) y viceversa" (AWS, n.d.).

El enfoque principal de usar EMR radica en la facilidad de uso de los trabajos de MapReduce en el *cloud computing* para aplicaciones bioinformáticas, el cual usa flujos de trabajo personalizados, en la plataforma altamente optimizada de Hadoop MapReduce en combinación con Amazon S3 y puede ser ejecutado por una interfaz de usuario básica.

4.6.2.3 Instancias de trabajo

En un sistema informático, el nodo es un bloque de construcción para la creación de entornos distribuidos. La infraestructura provista por el *cloud computing* consta de cuatro sistemas básicos en cada nodo que conforman una sola plataforma de computación. Estos sistemas constan de sistemas de almacenamiento, sistema de memoria, sistema de procesamiento y sistema de red. El *cloud computing* en este caso Amazon EMR facilita el acceso y demanda a recursos alojados en estos nodos, llamados instancias.

a) CPU

Los algoritmos bioinformáticos y la potencia de cálculo son los principales cuellos de botella para el análisis de GVD generados por las tecnologías actuales, como las metodologías de NGS. Mientras que los

procesadores multinúcleo con algunos núcleos proporcionan una mejora relativa en las aplicaciones, procesadores de múltiples núcleos (chips con decenas de núcleos) representan el siguiente gran paso en asequible computación de alto rendimiento. Los procesadores son indispensables tanto para hardware y software del sistema y para evaluar la funcionalidad y el rendimiento del sistema que soportara los procesos de análisis. AWS ofrece los servicios de instancias optimizadas para informática, que disponen de los procesadores de mejor rendimiento y ofrecen la mejor relación precio/rendimiento informático de EC2.

b) Memoria

El rendimiento del sistema de memoria es sensible a un gran número de parámetros e influye directamente en la realización de los análisis bioinformáticos. Cada uno de estos parámetros tiene un número de valores dentro de la ejecución de los programas en los análisis. La cantidad de memoria necesaria para asignar a las instancias, se configura en la solicitud de las instancias en la configuración del nuevo clúster, mediante el establecimiento de los parámetros con los valores correspondientes. Esos valores no cambiarán durante la ejecución. Entonces, es responsabilidad del investigador decidir si esas cantidades de memoria son importantes para el análisis en el entorno correspondiente.

AWS ofrece la posibilidad de utilizar instancias para el uso optimizado de memoria, las instancias de esta familia ofrecen grandes tamaños de memoria para aplicaciones de alto rendimiento, como en este caso informática científica y otras aplicaciones con uso intensivo de memoria.

c) *Network*

Los proveedores de la nube IaaS equipan su infraestructura con redes de alta velocidad, no sólo dentro de los centros de datos, sino también a través de los centros distribuidos geográficamente. Esto tiene un costo monetario, aunque los servicios no reflejan actualmente el uso de la red y el costo. El coste de la red por lo general se inserta dentro del costo de otros servicios como el servicio de cómputo y servicios de almacenamiento, y varía

según el tipo de servicio y ubicación de los centros de datos. En AWS, el costo de la transferencia de datos entre instancias de es cero si los servicios están situados en la misma zona de disponibilidad, mientras que la transferencia de datos salientes a internet siempre tiene un costo. Por otra parte, en AWS la velocidad de transferencia depende del tipo de instancia elegida.

d) Tipos de instancias

Los perfiles de instancias y especificaciones varían significativamente entre los vendedores. A medida que la carga de trabajo es principalmente intensiva de la CPU, es importante mantener el número de núcleos de CPU consistentes en la elección, así mismo en la cantidad de memoria de cada instancia. Los proveedores de la nube IaaS ofrecen diferentes instancias que varían de acuerdo a la capacidad del recurso y en consecuencia los precios.

Amazon EC2 ofrece un conjunto de instancias con diferentes configuraciones y precios, desde el más barato (un servidor con una capacidad de CPU de un tamaño 1.0-1.2GHz y la memoria de 1.7GiB) llega en costo de 0.026 dólares por hora, y una de las instancias más caras (High E/S Cuádruple Extra Large Instancia, lo que equivale a un servidor con capacidad de CPU de 32 × 1,0 - 1,2 GHz y la memoria de 244 GiB) viene a un costo de 2.8 dólares por hora.

Se recomiendan las instancias optimizadas para computo o para memoria de AWS para aplicaciones que requieren un rendimiento de memoria de alta en el mejor punto de precio por GiB de RAM. Las instancias incluyen las siguientes características:

- Procesadores "Ivy Bridge" Intel Xeon E5-2670 v2.
- La virtualización de hardware (HVM).
- Almacenamiento de instancia SSD respaldados, incluyendo soporte TRIM.
- Mejoras en las redes con menor latencia y alto rendimiento de paquetes por segundo.

4.6.2.4 Sistema de almacenamientos I/O

Los servicios requieren un conjunto de instancias para ejecutar las aplicaciones de los servicios; estos servidores están conectados a un grupo de dispositivos de almacenamiento, que almacenan los datos de los servicios. Los proveedores de la nube IaaS ofrecen dos tipos de servicios de almacenamiento que son diferentes en su política de precios y la usabilidad. Amazon Web Services tiene dos servicios de almacenamiento representativos: Amazon Simple Storage Service (Amazon S3) y Amazon Elastic Block Store (Amazon EBS).

Los servicios de almacenamiento se facturan normalmente de acuerdo con los GB utilizados por mes. Teniendo en cuenta los GVD que se gestionan y la necesidad de reducir la latencia de movimiento de datos al procesar datos, Amazon S3 es la primera opción para lograr no sólo rendimiento escalable sino también servicios fiables y predecibles. Amazon S3 crea una copia adicional dentro de una zona de disponibilidad pesto protege los datos frente a los fallos de componentes, ofreciendo una alta disponibilidad y durabilidad. Amazon S3 también interactúa con Amazon EBS en el procesamiento local dentro de las instancias. Debido a que Amazon EBS se puede conectar a cualquier instancia de Amazon EC2 y puede ser expuesto como un dispositivo dentro de la instancia. En consecuencia, en este estudio hemos adoptado el esquema de trabajo de Amazon S3 debido a que es un componente de almacenamiento permanente dentro del flujo de trabajo propuesto.

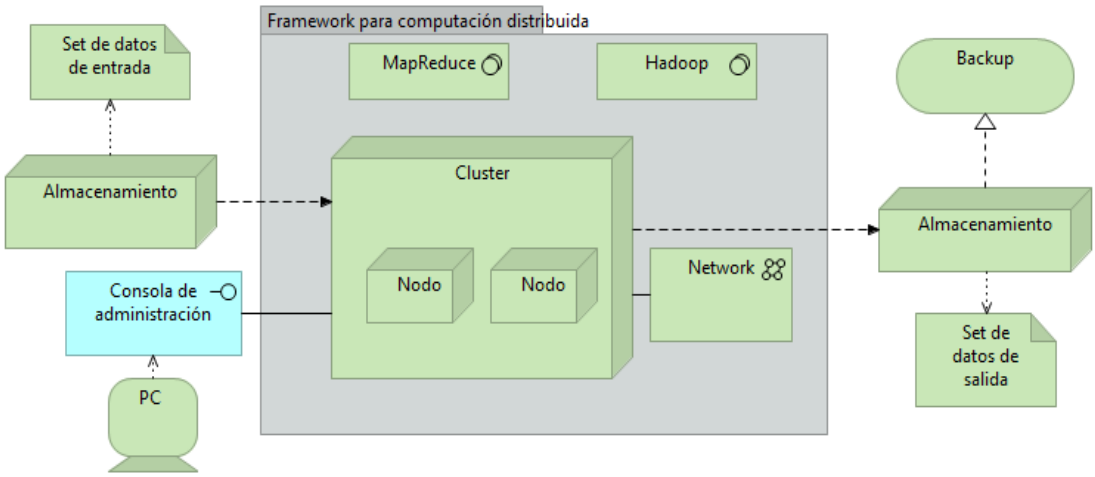
En el modelo el uso de Amazon S3 se plantea para el almacenamiento de los datos provenientes del secuenciamiento de los GVD proveniente de *Next Generation Sequencing*, usualmente los archivos fastq, fasta y también para los pipelines de análisis.

4.6.3 Modelo lógico de los componentes del modelo

El despliegue del modelo en el *cloud*, como Amazon Web Service (Figura 39), proporciona beneficios obvios como la demanda de recursos a medida, la fijación de precios basada en el uso, una mejor utilización de los

recursos, el aumento de la velocidad de procesamiento y una mejor experiencia para el investigador. Sin embargo, la creación de un clúster es una tarea no trivial que implica una serie de pasos de instalación y configuración, tanto para la plataforma y los paquetes de software dependientes que pueden ser propenso a errores y consumir tiempo. Estos pasos no requieren que los investigadores se conviertan en expertos de TI o dependan de los recursos de TI potencialmente dispersos.

Figura 39: Flujo de trabajo de los componentes del modelo



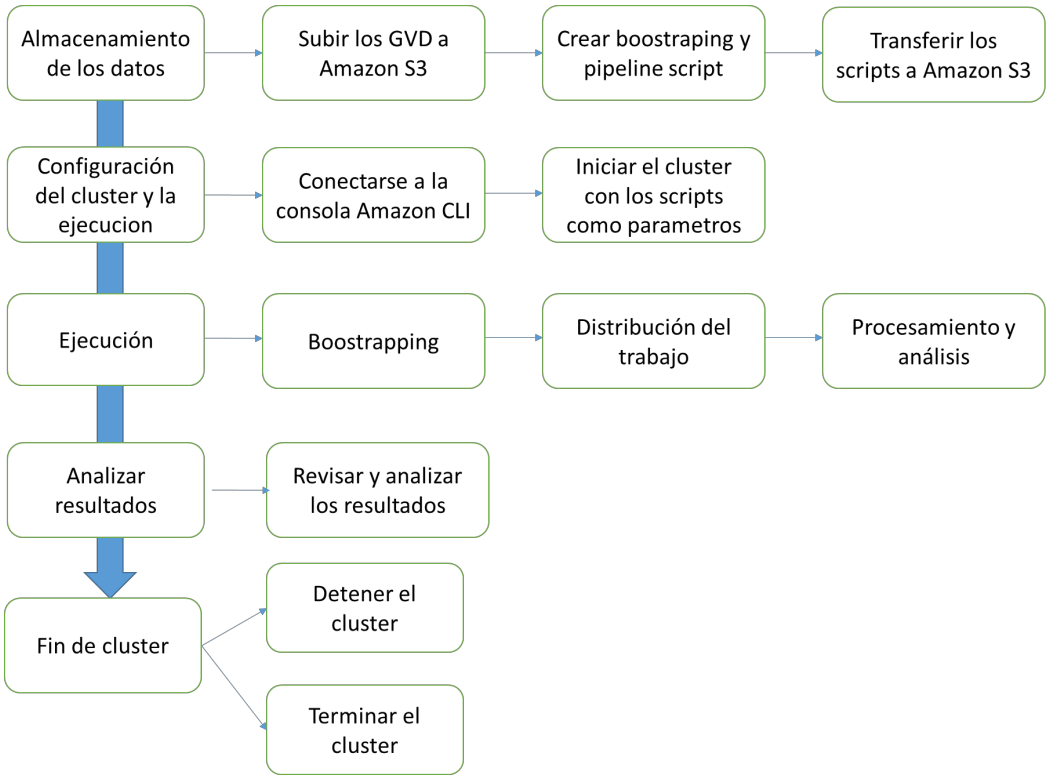
Fuente: Elaboración del autor

Para hacer frente a estos problemas, se presenta un modelo basado en el *cloud computing* para automatizar el proceso de implementación y los análisis bioinformáticos en Amazon EC2 a través de Amazon EMR.

En la Figura 39 se presentan el método para el despliegue del modelo en EC2 del cluster utilizando Amazon EMR (MapReduce). En concreto, tenemos el dominio de herramientas específicas separadas de los componentes básicos de Amazon S3 (Almacenamiento). En el contexto de EC2, esto se logra por tener una imagen de máquinas que contiene sólo las secuencias de comandos bootstrapping esenciales y un proceso de arranque bien definido de las herramientas de S3, para que las herramientas deseadas residan en los volúmenes de datos externos (es decir, volúmenes de EBS) luego del despliegue del bootstrapping. Este enfoque permite utilizar cualquier imagen de la máquina para una variedad de escenarios y evolucionando con

el tiempo, ya que puede ser contextualizada en tiempo de ejecución (es decir, preparar e implementar una instancia de otro modo genérico).

Figura 40: Marco de trabajo de la infraestructura



Fuente: Elaboración del autor

Al contar con herramientas específicas en los volúmenes de datos externos de S3, es fácil de modificar el contenido de esos volúmenes, y por lo tanto la disponibilidad de instrumentos subyacentes, sin la necesidad de modificar y reconstruir la imagen de la máquina. Esto es conveniente para los usuarios finales, ya que pueden centrarse en el uso de las herramientas, mientras que los desarrolladores pueden centrarse en asegurar que la infraestructura se comporta como se desee. En concreto en la Figura 40 se muestra en detalle el flujo de trabajo para los análisis bioinformáticos.

Primero es necesario la recolección de GVD en Amazon S3, así como las secuencias de referencia y las herramientas a utilizar, el investigador modifica el contenido del archivo de bootstrapping para la instalación de las herramientas deseadas (exhibido como un script de implementación de herramientas) en el clúster. Por otro lado, crea un pipeline de análisis con los

pasos del workflow de análisis definidos en la Arquitectura del negocio de la Sección 4.3.1.2, este pipeline debe hacer referencia a los GVD, las secuencias de referencia a través del uso de las aplicaciones que se definieron en la Arquitectura de aplicaciones de la Sección 4.4.2.2. Estos scripts de bootstrapping y pipeline también deben ser almacenados en el mismo repositorio de Amazon S3. Esto hace posible que los investigadores puedan añadir sus propias herramientas para las instancias de clúster.

Luego, a través de línea de comandos de AWS CLI, que es una herramienta de conexión a los servicios de AWS, se crea un nuevo clúster y se indica como parámetros de configuración del clúster la referencia al bootstrapping y pipeline que se encuentran en Amazon S3.

Posteriormente, luego de crearse el clúster se comenzará con el bootstrapping de las aplicaciones y luego la ejecución del pipeline, que usará la referencia a los GVD de entrada en el *framework* para computación distribuida EMR, donde Hadoop a través de un proceso llamado MapReduce distribuirá la carga de trabajo al clúster de recursos de procesamiento para el análisis respectivo. El set de datos generados es transferido y almacenado a la misma ubicación del repositorio de almacenamiento de S3. Finalmente, los resultados logrados del análisis bioinformático podrán ser obtenidos desde Amazon S3 para su interpretación y análisis por parte de los investigadores.

Toda la información sobre el estado de los análisis (instancias de las tareas) se monitorean como sistema de flujo a través del monitor de la herramienta EMR para mantener informados al investigador sobre la evolución de la ejecución. Como resultado, la funcionalidad y la arquitectura fácilmente se pueden ajustar a las necesidades informáticas de los investigadores individuales.

4.6.4 Análisis de costo beneficio de la propuesta

Los costos y beneficios asociado a la implementación de la propuesta vienes dad por los cuantificables y los no cuantificables. Entre los costos no cuantificables se tiene la resistencia al cambio de algunos investigadores,

riesgos de transición y riesgos de perdida de datos, los cuales son mitigados con las medidas descritas en la sección de componente de seguridad. Por otro lado, se tiene los beneficios no cuantificables entre los cuales se resalta la simplificación de la administración de los sistemas de información, incremento de flexibilidad de la infraestructura, además del incremento de la disponibilidad de los sistemas.

En cuanto a los costos cuantificables se realizó los cálculos correspondientes a un año, el cual se detalla en los ítems y los costos de la Tabla 28. Para lo cual se consideró un servidor estándar cuyo costo de uso por hora es de S/. 1.09, considerando que se tendrá un almacenamiento para resultados de 500GB, con costos de transferencia de S/ 0.3 por 1 millón de transacciones, además se contempla un pequeño costo de administración que se refiere al mantenimiento de la cuenta en el *cloud computing* asi como asistencia a los usuarios. Así mismo, se propuso como el costo de implementación de un servidor in house un costo inicial de S/. 32,800, costos de energía, mantenimiento y alimentación anuales de S/. 8,790, mientras que los costos de administración del servidor ascienden a S/. 13,776 el cual se va incrementando año a año.

Tabla 28: Costos cloud computing

Item	Monto S/.
AWS EC2/año	9562
AWS S3 almacenamiento/año	3936
Transferencia de datos/año	1968
Administración/año	3608

Fuente: Elaboración del autor

Las siguientes formulas se utilizaron para el cálculo del retorno de la inversión (ROI), donde las operaciones se realizan tomando en cuenta el ahorro obtenido en 5 años comparado a la adquisición de un nuevo servidor. El detalle de los cálculos se presenta en el Anexo 7. El ahorro promedio anual es de S/. 53,059. Mientras el costo total de inversión es de S/. 211,560 para el periodo mencionado y el tiempo de retorno es en el primer año.

Ahorro = Costo fijo de la inversión en TI – costo del *cloud computing*

ROI = Ahorro / costo total *cloud computing* * 100%

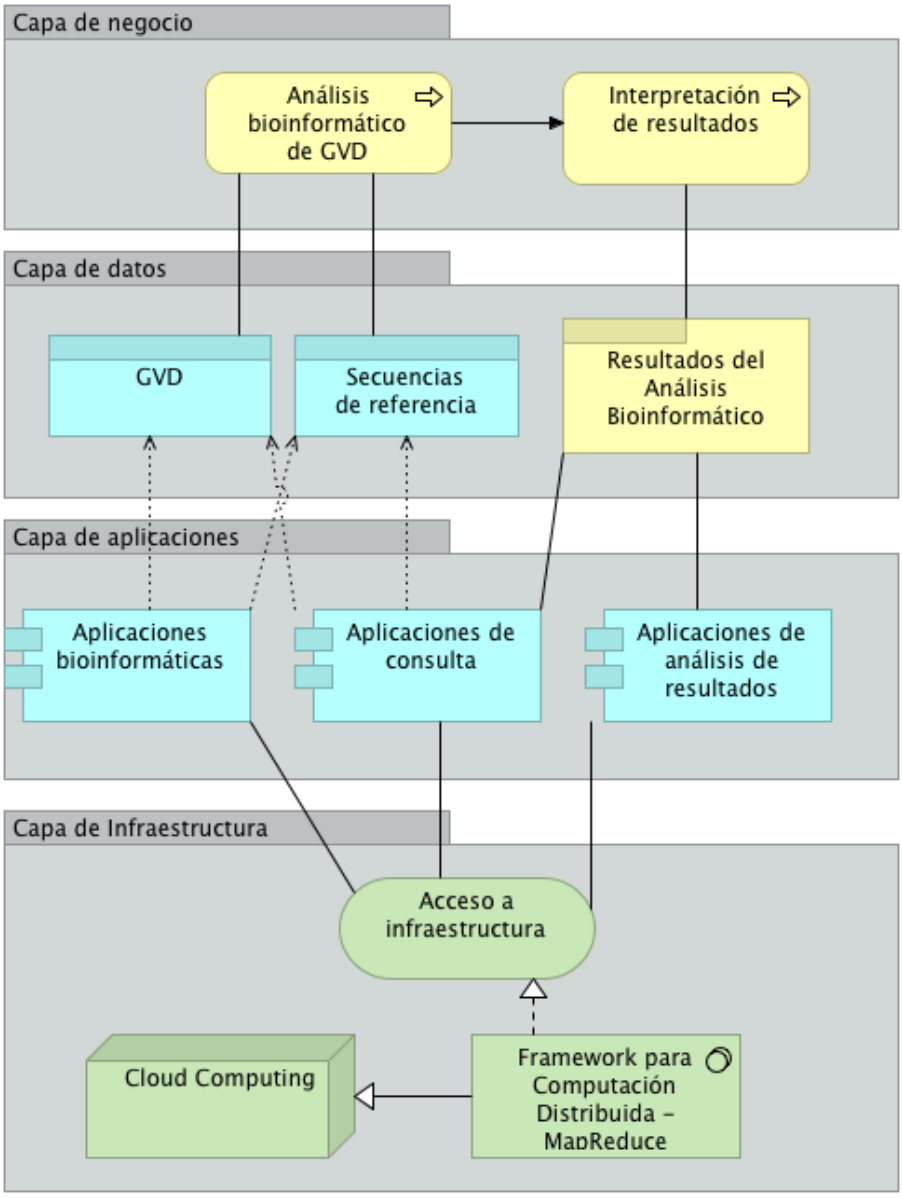
El costo total de implementación de la inversión es de S/. 211,560, así mismo se considera el ahorro como rendimiento del proyecto, que se logra gracias al uso del *cloud computing*. La tasa de descuento es del 4%. Tomando en cuenta 5 años de la implementación de la arquitectura tecnológica y se calcula que el valor actual neto del proyecto es 38,800 mientras que la tasa de retorno (TIR) es de 23%. Con lo cual se demuestra que la arquitectura presentada también es viable financieramente.

4.6.4.1 Vista general del modelo de arquitectura

En la actualidad el término arquitectura a nivel de los sistemas de información se usa para indicar que se dispone de una estrategia y modelo de coordinación de los diferentes componentes tecnológicos que soportan los procesos y servicios, y por consecuencia las necesidades de negocio, en la Figura 41, se muestra un resumen del modelo en capas de las arquitecturas desarrolladas anteriormente.

En la arquitectura de negocio de los procesos de análisis in silico de las muestras biológicas se identificó la estrategia de que la arquitectura se relaciona con los recursos bidireccionalmente, así, la estrategia general y los procesos influyen en la definición de los recursos, del mismo modo, las características o limitaciones de los recursos influyen en la conceptualización de la estrategia y diseño de procesos, y por lo tanto en la definición de la arquitectura de negocio. Así mismo, se diseñó la arquitectura de sistemas de información, con los requerimientos de información, y las necesidades antes definidas, que luego dio paso a la arquitectura de datos en el cual se definieron la naturaleza de los datos y sus fuentes necesarios para soportar los procesos de análisis bioinformático. En la arquitectura de aplicación se identificó los requisitos para la presentación de os servicios bioinformáticos.

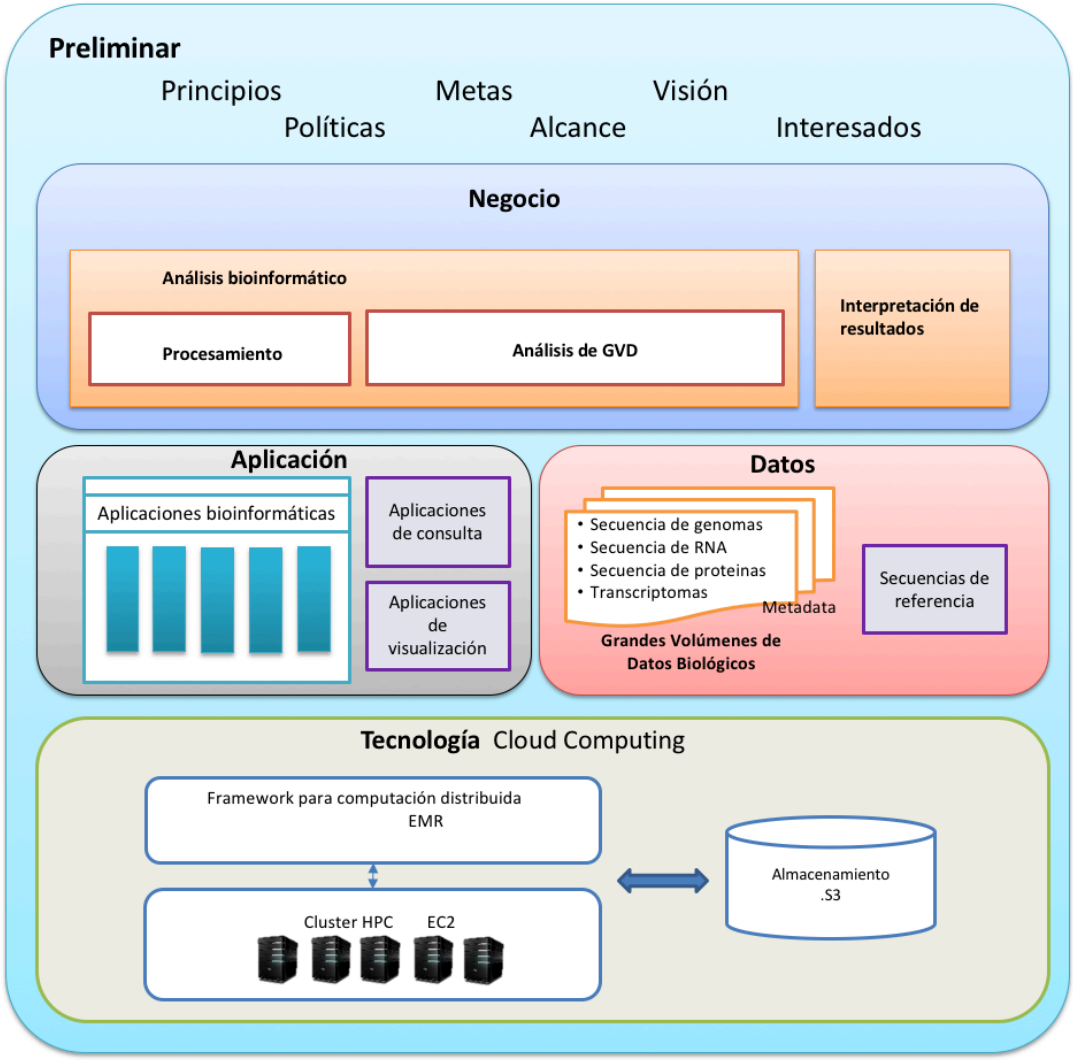
Figura 41: Visión en capas de la arquitectura



Fuente: Elaboración del autor

Por último, en la arquitectura tecnológica se determinó la tecnología a utilizar, el uso del *cloud computing* y los servicios a utilizar, así mismo se definen especificaciones para el aprovechamiento de este tipo de tecnología. La figura 42 muestra una visión general de la arquitectura empresarial para el análisis bioinformático de grandes volúmenes de datos, que se desarrolló a lo largo de este capítulo.

Figura 42: Arquitectura empresarial de análisis de GVD



Fuente: Elaboración del autor

CAPÍTULO V
RESULTADOS

El desarrollo de la presente investigación permitió obtener una propuesta de arquitectura empresarial para el análisis bioinformático de grandes volúmenes de datos. Para ello, se analizó la problemática, se planteó el problema, objetivos, hipótesis y variables de investigación y se realizó toda la lectura necesaria para recopilar toda la información teórica que sustenta la investigación. Luego, se desarrolló la arquitectura empresarial, el modelo de negocio, modelo de infraestructura, modelo de aplicaciones y modelo de datos, utilizando la metodología del ADM de TOGAF como línea base.

En este capítulo, se muestra como la arquitectura empresarial propuesta se validó interna y externamente. Se validó internamente con la aplicación en situaciones reales en el centro de investigación piloto, estas aplicaciones corresponden a dos casos de estudio donde la autora estuvo involucrada directamente. Para validar externamente ésta arquitectura, se diseñó un cuestionario para la recolección de datos basada en la operacionalización de las variables en estudio, y se realizó la encuesta en el centro piloto a 9 personas, quienes son los expertos. En esta sección se analizan e interpretan los resultados por medio de un análisis estadístico descriptivo, formado por frecuencias relativas porcentuales y medidas de tendencia central, cuyas dimensiones corresponden a las hipótesis especificas planteadas. Por último, se realiza la prueba de hipótesis para validar la investigación.

5.1 Casos de estudio

5.1.1 Caso1: Completa secuencia genómica de la nueva variante de Potato Virus X (Kutnjak, et al., 2014)

5.1.1.1 Descripción del caso

Las plantas constantemente son afectadas con diferentes patógenos que afectan seriamente su crecimiento, rendimiento y cultivo. Los estudios de los virus en las plantas permiten prevenir las infecciones y estudiar los factores que permiten las contaminaciones con estos. En los cultivos de papa, Potato virus X es uno de los virus más difundidos y estudiados intensivamente debido a sus nefastos efectos en estos cultivos. Es así que es importante conocer profundamente este virus, y más específicamente conocer su estructura de ADN y su mecanismo de acción, que permitirá desarrollar nuevos tratamientos que permitan a las plantas ser inmunes a estos virus, protegiendo así los cultivos. En este sentido, la estructura completa del ADN se refiere como genoma y el proceso de obtención de este genoma se realiza a través de la secuenciación en este caso.

En el artículo científico publicado por Kutnjak (2014), se detalla el procedimiento para obtener el genoma de Potato Virus X (PVX), partiendo por la parte de extracción de ARN de las muestras en el laboratorio y los resultados de los análisis bioinformáticos. Esta investigación se realizó en CIP, para la parte bioinformática se utilizó la infraestructura basada en la arquitectura presentada en la presente investigación. Para la validación esta tesis se presenta la parte bioinformática y se detalla como el uso de metodología propuesta contribuyó en el ensamblaje de este virus.

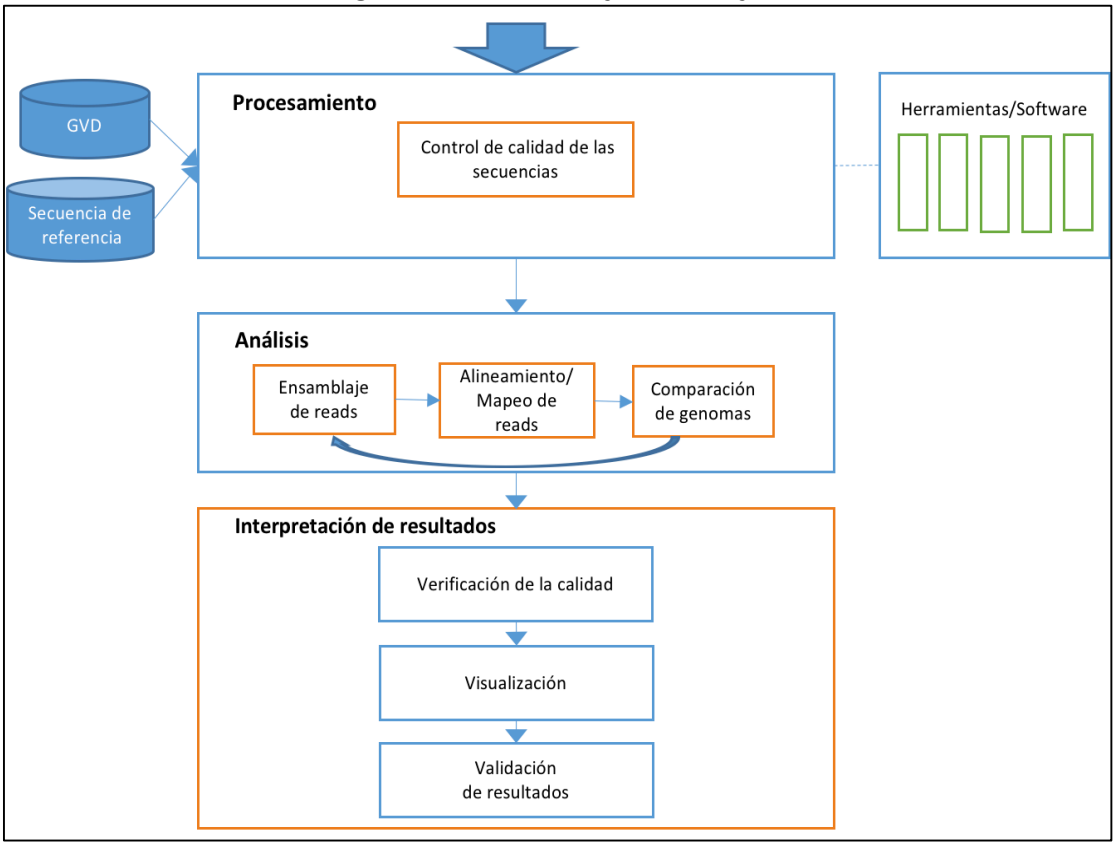
5.1.1.2 Desarrollo, materiales y métodos

Pequeños RNAs aislados de papa fueron enviados a Fasteris Life Sciences SA (Suiza) para la preparación de las secuencias en la plataforma Illumina Hiseq2000 (una máquina que lee el ARN de las plantas y los transforma en datos), cuyo resultado fueron los GAF214, GAF215, GAF216, GAF217, GAF218.

Se realizó el análisis bioinformático de los GVD (5 GAFs) utilizando el siguiente flujo de trabajo, basado en la arquitectura de negocio propuesta, ver Figura 43. El control de calidad de las secuencias del archivo FASTQ, se realizó con la herramienta fastQC. Este es un paso previo a cualquier tipo de análisis. Se ensamblo las secuencias de longitud 21-24 pb utilizando Velvet v0.6.04 (Zerbino & Birney, 2008) y el script AssemblyAssembler 1.4 (Jacob Crawford). Las secuencias ensambladas se compararon con todas las secuencias virales depositados en la base de datos NCBI GenBank utilizando BLASTN y BLASTx.

Se utilizó la secuencia completa del genoma de PVX, que mostró la más alta similitud con la más larga secuencia coincidente, como referencia para el mapeo de secuencias (21 a 24 nt) utilizando MAQ (http://maq.sourceforge.net); posteriormente se verifico la calidad del ensamblaje.

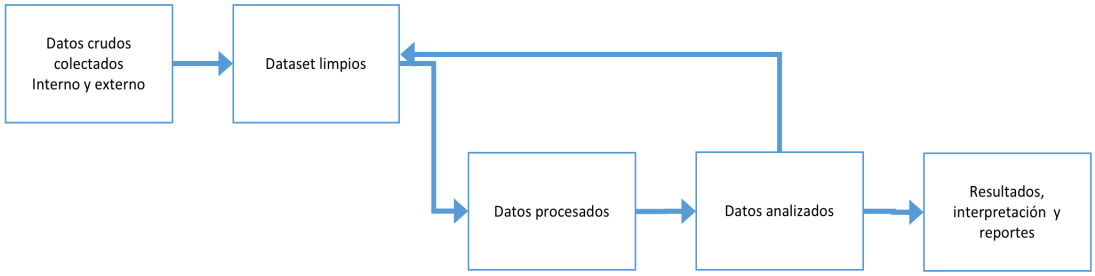
Figura 43: Vista de flujo de trabajo



Fuente: Elaboración del autor

Se usó Seqman para ensamblar y visualizar la secuencia de consenso de salida MAQ y los contigs (secuencias) ensamblados de Velvet y se obtuvo una nueva secuencia de consenso con mejor cobertura del genoma viral. Con el fin de perfeccionar la secuencia de consenso del genoma viral, los 2 últimos pasos (usando la nueva secuencia de consenso como referencia) fueron realizados varias veces. Dentro de este procedimiento descrito, los GVD pasan por diferentes etapas pasando de ser datos crudos a ser datos limpios luego del control de calidad realizado, así mismo son procesados como se mencionó anteriormente y luego se realizó el análisis respectivo. Una vista del flujo de datos se muestra en la Figura 44.

Figura 44: Flujo de los datos



Fuente: Elaboración del autor

Se muestran en la Tabla 29 los metadatos de las muestras de PVX en estudio (grandes volúmenes de datos), son datos no estructurados, por lo cual solo presentan metadatos.

Tabla 29: Metadatos de los GVD

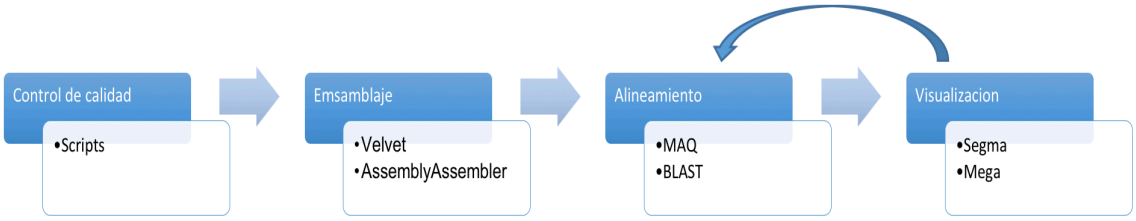
Campo	Valor
ids	GAF214 GAF215 GAF216 GAF217 GAF218
desc	Samples from CIP greenhouse infection, Lima
attr	Potato
data	Small RNA
.hash	Valor HASH
.state	*Raw* data
.type	FASTQ

Fuente: Elaboración del autor

Luego de mostrarse el flujo de datos, en la Figura 45 se muestra las herramientas que se utilizaron, basada en la arquitectura de aplicaciones. Para el control de calidad de las secuencias se utilizó FASTQC con el cual se verifico que todas las secuencias del archivo FASTQ tengan el formato correcto, así mismo entre las aplicaciones de ensamblaje de secuencias cortas se utilizó Velvet y AssemblyAssembler. Los archivos generados sirvieron de entrada para el uso de BLAST y posteriormente para el uso de la aplicación de mapeo a genomas con múltiples alineamientos de secuencias y mapeo a genomas (MAQ) como se muestra en la Figura 45.

Se utilizó *s3cmd* como aplicación de consultas de almacenamiento, así mismo se utilizó el monitor de Amazon EMR, para la revisión del estado de los procesos durante el análisis. Segman y MEGA son las aplicaciones de visualización de alineamientos usadas.

Figura 45: Uso de aplicaciones



Fuente: Elaboración del autor

Para realizar los análisis de las secuencias antes señalados, se utilizó un clúster de 4 instancias m1.xlarge, los cuales distribuyeron la carga de trabajo a través del uso de Hadoop y MapReduce para los análisis en el menor tiempo posible y con los menores recursos, el manejo del cluster es transparente al usuario. En la Tabla 30 se puede observar el uso de los componentes para obtener los resultados de los análisis. Cabe resaltar que la visualización de los resultados se realiza en computadoras de escritorio, ya que este tipo de datos procesados no son GVD, por ello no se detalla en la Tabla 30. Estos GVD fueron almacenados en Amazon S3 juntos con los resultados finales.

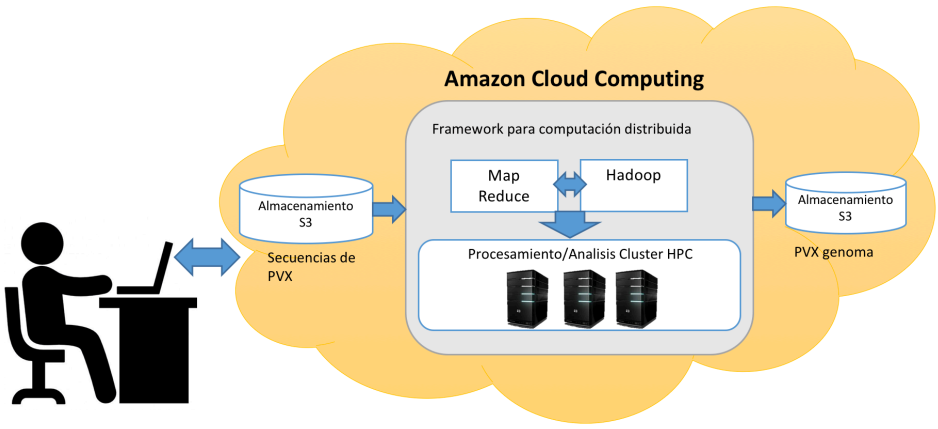
Tabla 30: Recursos de infraestructura en el cloud computing

Aplicación	Numero de pares de bases	Numero de contigs	% de *reads* mapeados	Tiempo (min)	RAM (GB)	CPU
Velvet	1'432'104	2000	-	812	60	36
MAQ	1'432'104	-	80%	523	51	30
Assembly Assembler	1'432'104	1000	-	1000	60	36

Fuente: Elaboración del autor

Del cuadro anterior nos da una información sobre los recursos que se utilizó en los análisis llevados a cabo, cabe resaltar que esta información corresponde al tiempo y recursos para cada ejecución del programa en mención, cabe resaltar que el uso de MAQ como se mencionó anteriormente tuvo ejecuciones repetitivas. Por otro lado, debido a que se ejecutó en un clúster el tiempo de análisis se redujo considerablemente ya que el uso de una sola instancia hubiera cuadruplicado el tiempo. Por otro lado, usar el *cloud computing* permitió realizar el análisis completo sin problemas de faltas de recursos, al utilizar Hadoop permitió el balance de la carga de trabajo.

Figura 46: Componentes del análisis en el cloud computing



Fuente: Elaboración del autor

Todo este análisis se realizó con el *cloud computing* con el uso del modelo propuesto en la siguiente investigación, en la Figura 46 se muestra un diagrama de los componentes utilizados.

5.1.1.3 Resultados

La validación de resultados a través de la verificación de la calidad de resultados, la validación con literatura y la publicación de resultados se realizó por parte del autor del artículo (Kutnak, 2014). El autor del artículo menciona que las secuencias de consenso de PVX obtenidos por análisis de las 5 muestras de invernadero fueron significativamente diferentes de secuencias de cualquier otro PVX disponible en NCBI GenBank. Por lo tanto, se generó una nueva secuencia de consenso que abarca el genoma completo de PVX, de las 5 muestras, como parte de una nueva variante de PVX. Como resultado de la investigación del caso de estudio se muestra un árbol filogenético de la nueva variante del PVX, donde se puede apreciar que las 5 muestras en estudio se encuentran en un grupo diferente de los otros PVX.

En este sentido, el análisis bioinformático de PVX permitió ensamblar una nueva variante de este virus como hallazgo para la comunidad interesada de patógenos en plantas. Este resultado se realizó satisfactoriamente usando la arquitectura propuesta en la presente tesis a través del modelo de negocio, de infraestructura, aplicaciones y datos.

5.1.2 Caso 2: Los virus del camote africano: la cartografía y la diversidad de los virus de camote. (Kreuze, et al., 2014).

5.1.2.1 Descripción

La investigación en mención corresponde a un poster de presentación (ver Anexo 5) durante la reunión anual en CIP, por motivos de propiedad intelectual solo se presenta en este caso de estudio una breve introducción de la investigación, debido a que esta investigación aún se encuentra en desarrollo y no tiene una publicación formal en el mundo científico.

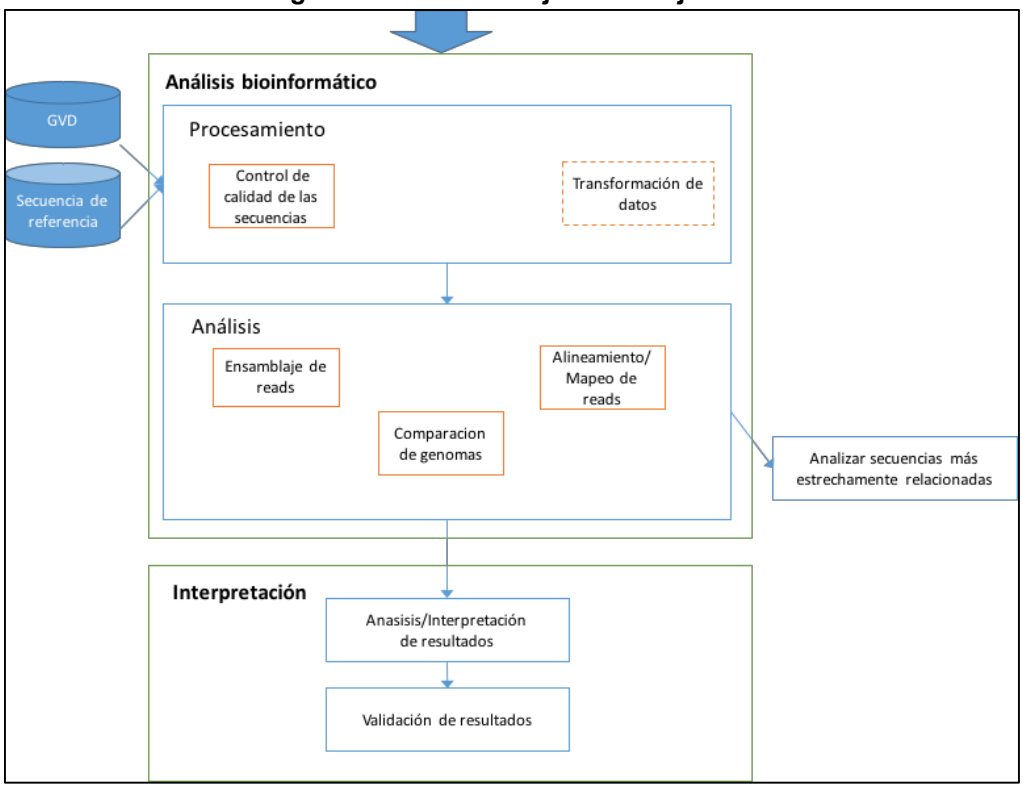
Las enfermedades virales de las plantas se consideran un obstáculo importante para la producción de camote en todo el mundo, particularmente en África subsahariana (SSA). Con pocas excepciones, no hay información clara sobre la prevalencia y distribución de los diferentes virus de camote y sus respectivas variantes. Esta información básica es esencial no sólo para

gestionar la propagación sino también para evaluar el impacto de estas enfermedades virales. Es así que, aprovechando la tecnología de secuenciación de próxima generación (NGS) y los análisis bioinformáticos. Este poster es parte de un proyecto cuyos esfuerzos se centran en determinar los virus de camote en África. La arquitectura propuesta en la presente tesis mediante la parte bioinformática contribuyo grandemente en los esfuerzos de la detección de virus de camote.

5.1.2.2 Desarrollo

Se colectaron muestras de camote cultivados en el campo en 13 países, los cuales son Ghana, Mozambique, Etiopía, Tanzania, Guinea, Nigeria, Benín, Malawi, Zambia, Zimbabwe, Angola, Kenia y Uganda. Los virus han sido identificados por un método genérico para el diagnóstico virus desarrollado en el CIP llamado "pequeña secuenciación del ARN y ensamblaje (SRSA)", dentro del marco de trabajo de la propuesta de la arquitectura empresarial implementada en CIP. Las identidades de los virus fueron confirmadas por PCR (un método de laboratorio).

Figura 47: Vista de flujo de trabajo



Fuente: Elaboración del autor

El desarrollo de los análisis siguió el flujo de trabajo que se muestra en la Figura 47 donde las muestras de cada campo de los países pasaron una por una la misma metodología del modelo de negocio.

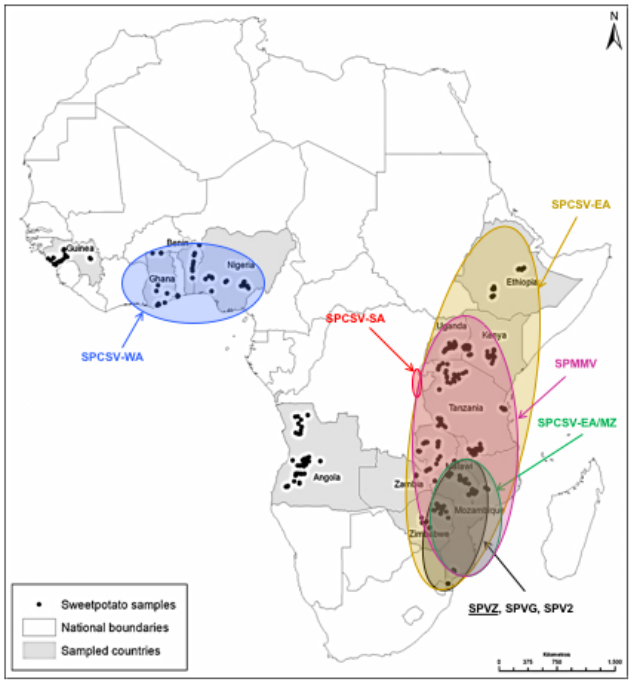
En análisis bioinformático comprendió principalmente el procesamiento y el análisis de los datos. En el procesamiento, se realizó el control de calidad de las secuencias, lo que consiste en verificar si los datos no están corruptos, así mismo que el formato interno sea el adecuado. En algunos casos los datos estuvieron comprimidos y se procedió a descomprimir, verificando nuevamente que en este proceso no se alterara la data original. El análisis estuvo comprendido por el alineamiento y mapeo de las secuencias, a referencias de virus y patógenos ya existentes, así mismo se procedió a encontrar variantes del mismo virus a través del ensamblaje de las secuencias, para finalmente realizar una comparación de genomas. Este procedimiento también contempla el análisis detallado de secuencias especificas las cuales fueron usadas para determinar secuencias estrechamente relacionadas.

Por último, la interpretación y validación de los resultados se lleva a cargo por parte de los científicos involucrados. No es posible desplegar detalles de los modelos de datos, aplicaciones realizados en este caso de estudio, debido a que esta investigación aún carece de publicación formal.

5.1.2.3 Resultados

La arquitectura empresarial que se desarrolló, sirvió de base para llevar a cabo este proyecto que ayuda en el análisis de la infección de virus en las muestras de los camotes africanos, así la comprensión de la distribución de virus y la variabilidad permitirá un adecuado manejo de las enfermedades virales de camote en todo el continente africano. Como se observa en la Figura 48.

Figura 48: Distribución de los virus de camote en África



Fuente: (Kreuze, et al., 2014).

5.2 Análisis de la encuesta

Con el fin de determinar la medición del impacto de la arquitectura empresarial en la organización, se realizó una encuesta antes y después de la implementación de la arquitectura empresarial basada en el *cloud computing*, cuya muestra es no probabilística debido a que la elección de los miembros se determinó debido al criterio de expertise; las muestras no probabilísticas no se prestan para tratamiento y análisis estadísticos. Sin embargo, ya que las muestras basadas en la opinión son seleccionadas en base a la opinión de expertos que son representativas de la población, se realizó el análisis descriptivo. Aunque se generaliza a toda la población, las conclusiones obtenidas tras el análisis descriptivo, calcula las medidas de tendencia central, para ver en qué medida los datos se agrupan o dispersan en torno a un valor central y de dispersión.

La validación interna de la encuesta está representada por la consistencia y la fiabilidad de la encuesta, mientras que la validez del contenido se evidencia con la revisión de la literatura realizada y la consulta sistemática a expertos. Se utilizó el índice de consistencia interna Alfa de Cronbach, que es

un instrumento fiable que hace mediciones estables y consistentes, cuyo valor más cercano al extremo 1 mejora la fiabilidad, considerando que los mejores resultados son a partir de 0,80. Los resultados obtenidos usando el índice de consistencia interna Alfa de Cronbach fue de 0.7972 para la encuesta de la pre prueba, mientras que de 0.8328 para la encuesta de post prueba, este resultado permite señalar que el instrumento posee alta consistencia interna, por tanto se puede afirmar que mide con un alto grado de fiabilidad el impacto de una arquitectura empresarial en bioinformática.

La encuesta abarca las variables independientes organizadas en 5 dimensiones (organización, procesos, datos, aplicaciones e infraestructura) de la investigación y agrupan catorce indicadores, correspondientes a las 14 preguntas del Anexo 1, se encuestó a 9 investigadores que trabajan con bioinformática y hacen uso de los análisis bioinformáticos de grandes volúmenes de datos, llevado a cabo en el Centro Internacional de la Papa. Para esta investigación en particular la muestra es igual a la población.

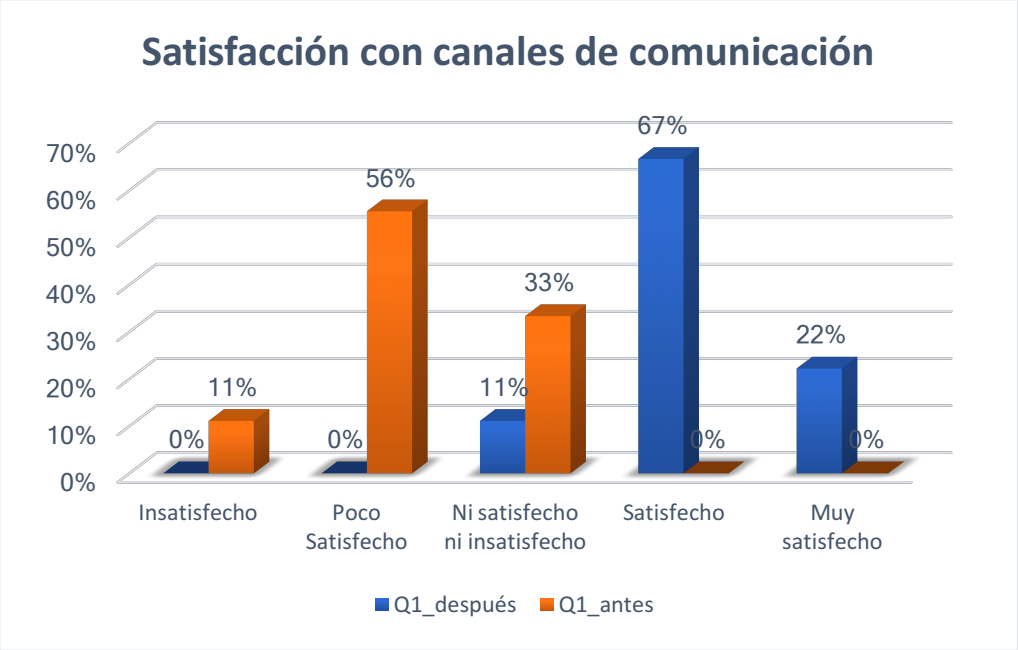
5.2.1 Análisis descriptivo de los datos de la encuesta

Los datos recolectados se muestran en el Anexo 2, mientras que la tabulación de los resultados se encuentra en el Anexo 3. A continuación, se muestran los análisis de los resultados obtenidos en el Anexo 3 de acuerdo a las dimensiones en estudio.

5.2.1.1 Procesos

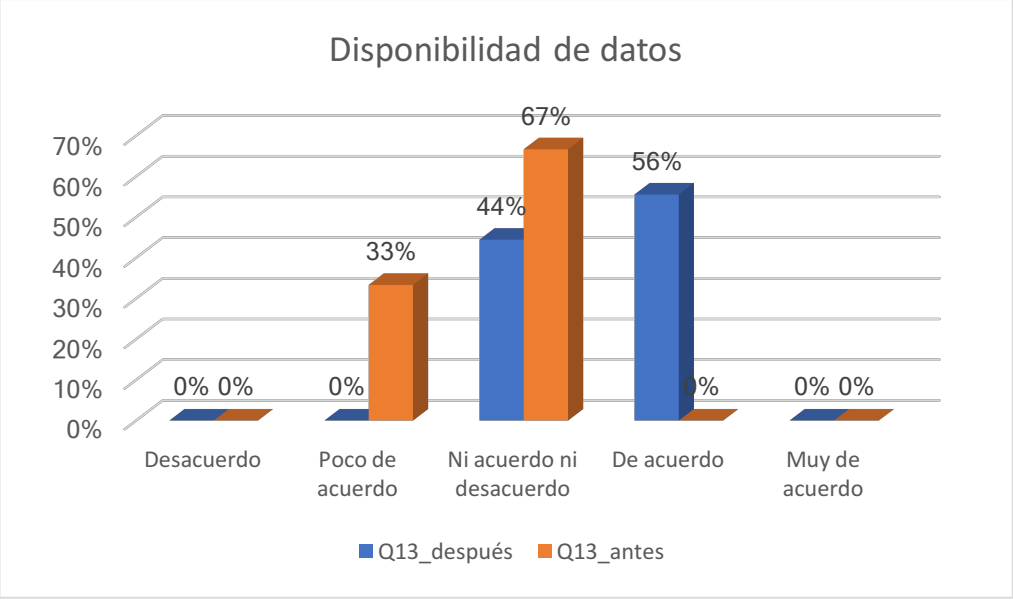
Posteriormente a la implementación de la arquitectura empresarial, la mayoría de los encuestados responden favorablemente referente al tema de alineamiento entre el negocio y la tecnología, el 88% de los encuestados considera que los procesos desarrollados apoyan los objetivos de la investigación, comparado con un 56% que anteriormente no se encontraba en de acuerdo ni desacuerdo. Desde el punto de vista de comunicación, más del 67% se encuentra satisfecho con la comunicación con el personal de TI disponibles en el centro, mientras que antes de la arquitectura el 67% se encontraba poco satisfecho e insatisfecho (Figura 49).

Figura 49: Satisfacción con canales de comunicación



Fuente: Elaboración del autor

Figura 50: Disponibilidad de datos



Fuente: Elaboración del autor

5.2.1.2 Datos

En la encuesta realizada luego de la investigación, el 56% de los encuestados se encuentra de acuerdo que los grandes volúmenes de datos siempre están disponibles para su uso (Figura 50), cuya situación ha cambiado referente al 67% de los encuestados antes de la arquitectura

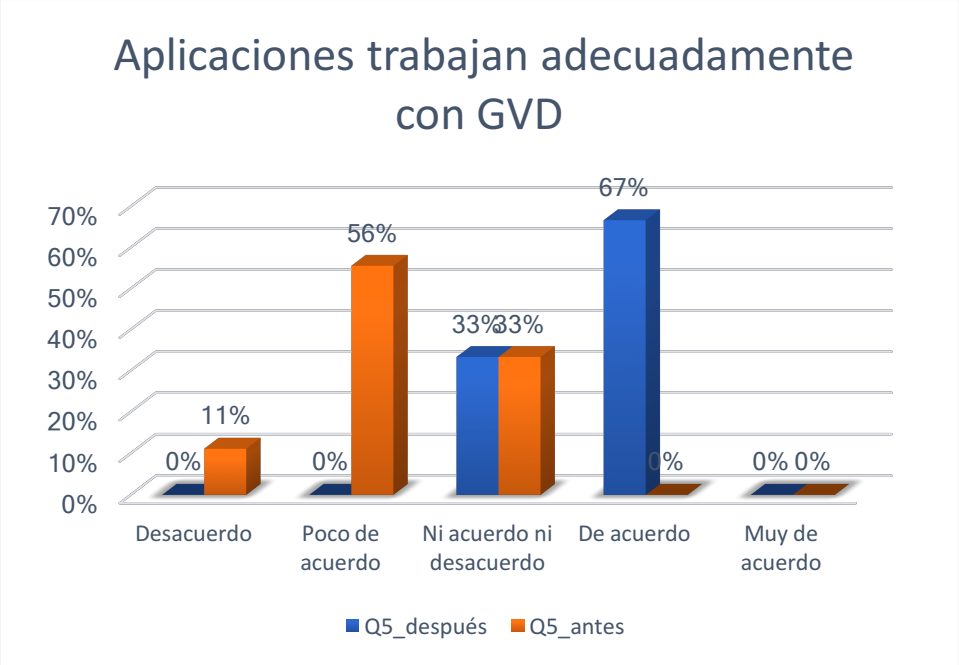
quienes no tenían una idea clara sobre la disponibilidad de los datos. Así mismo, posterior a la implementación más del 70% de los encuestados se encuentra de acuerdo con la gestión de los grandes volúmenes de datos de sus respectivos proyectos de investigación.

5.2.1.3 Aplicaciones

Al evaluar la satisfacción de los investigadores, inicialmente el 11% de los encuestados estaba satisfecho con el nivel de disponibilidad de los sistemas, ahora se determinó que el 89% de los investigadores se encuentran satisfechos o muy satisfechos con el nivel de disponibilidad de los sistemas para su uso. Además, ahora 89% los encuestados se encuentran de satisfechos a muy satisfechos con la facilidad de uso de los sistemas de información que utilizan.

Por otro lado, anteriormente el 33% no precisaba si estaba de acuerdo que las aplicaciones trabajaban adecuadamente con los grandes volúmenes de datos. Luego de la arquitectura, más del 66% considera que las diversas aplicaciones trabajan adecuadamente con los grandes volúmenes de datos y solo el 33% no tiene una postura clara (Figura 51).

Figura 51: Aplicaciones trabajan adecuadamente con GVD

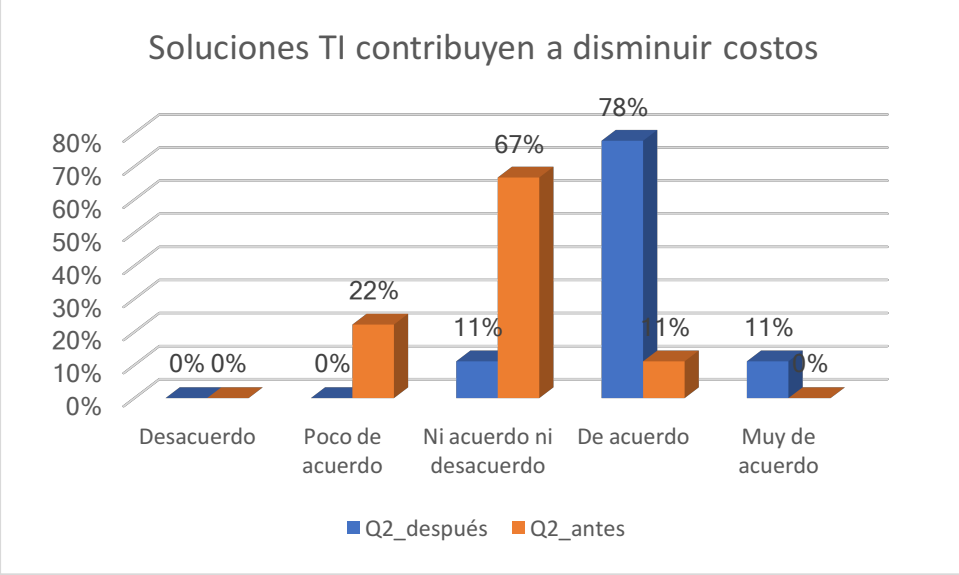


Fuente: Elaboración del autor

5.2.1.4 Infraestructura

En la encuesta realizada luego de la arquitectura el 78% de las personas están de acuerdo que los sistemas de información (*cloud computing*) rápidamente se adaptan a cambios y el 56% está de acuerdo que las inversiones en TI aportan al objetivo científico. Así mismo, anteriormente el 67% de los investigadores no tenían una opinión clara si un cambio en las soluciones de TI disminuiría los costos, ahora el 77% está de acuerdo y 11% muy de acuerdo que la implementación de soluciones en TI contribuye a la disminución de costos (Figura 52) y el 66% de ellos se muestra de acuerdo al considerar que el sistema de información que se encuentra en el *cloud computing* reduce el tiempo de ejecución de sus actividades en comparación al 78% quienes no tenían una postura clara previamente.

Figura 52: Soluciones de TI contribuyen a disminuir costos



Fuente: Elaboración del autor

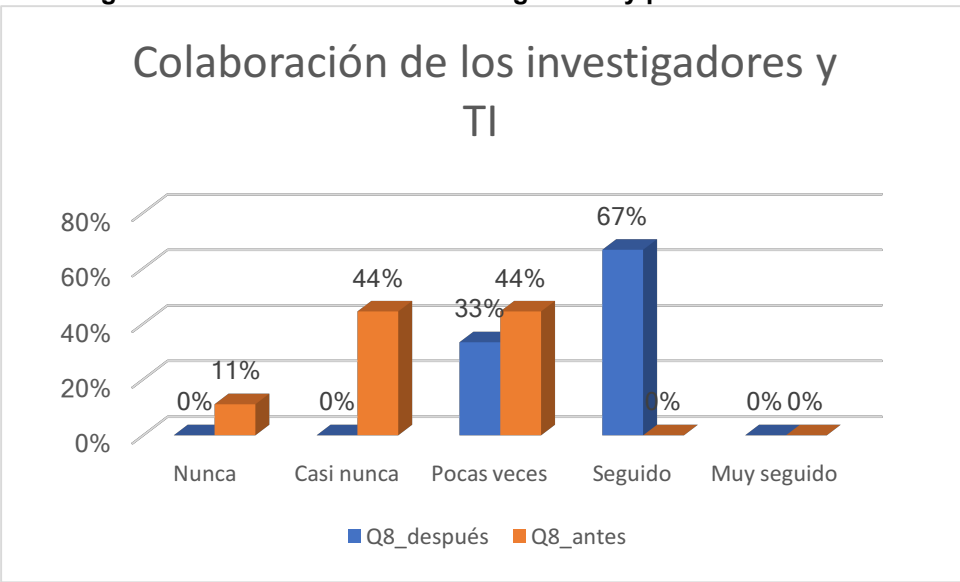
5.2.1.5 Organización

En líneas generales el 89% de los investigadores bioinformáticos se encuentran satisfechos con respecto a los servicios de TI en el *cloud computing*, mientras que antes de la arquitectura solo el 11% estaba de acuerdo. Referido al tema de cultura organizacional, el 67% de los investigadores pocas veces o casi nunca recibe/formula quejas o reclamos por cambios realizados en sistemas, debido a algún proyecto de sistemas/TI,

mientras que anteriormente el 100% de los encuestados recibía quejas/reclamos.

Así mismo el 66% de los encuestados manifiesta que existe colaboración de los investigadores científicos con el desarrollo de proyectos en TI (Figura 53), mientras anteriormente expresaba que más del 55% no tenía colaboración con los proyectos de TI o pocas veces el 44%.

Figura 53: Colaboración de investigadores y personal de TI



Fuente: Elaboración del autor

A partir de estas encuestas se calculó que la mediana global de la respuesta es que los científicos se encuentran satisfechos con las condiciones asociadas al impacto de una arquitectura empresarial para el área de bioinformática que trabajan, comparado a la situación previa donde la mediana global de la respuesta era que los científicos se encontraban poco satisfechos con la situación y condiciones para los grandes análisis de datos (Anexo 2). Por lo cual, se concluye que existe satisfacción del usuario final, alineamiento entre negocio y tecnología e influencia positiva en la cultura organizacional, comunicación, sistemas de información, servicios tecnológicos; por lo tanto, mejora de las investigaciones.

5.3 Comprobación de la hipótesis

La comprobación de la hipótesis se realizó a partir del planteamiento de la hipótesis nula y alternativa. Se formuló una regla decisión y se usó un estadístico que se utilizó para la hipótesis general y para las hipótesis específicas de la investigación.

- **Regla de decisión:**

Se utilizó la misma regla de decisión para la hipótesis general, así como para las hipótesis especifica. Si el p-valor resultante de la prueba es menor al nivel de significancia 0.05 se rechaza la hipótesis nula y se acepta la hipótesis alterna, en caso contrario se acepta la hipótesis nula.

- **Estadístico de prueba de hipótesis:**

La prueba de hipótesis consiste en el análisis de diferencias de grupos, para esto se utilizó la prueba de Wilcoxon, esta prueba se realizó tanto a la hipótesis general como a las hipótesis específicas. Wilcoxon es una prueba no paramétrica para realizar análisis de dos muestras relacionadas con datos ordinales y determinar si existen diferencias entre ellas, por lo tanto, no necesita una distribución específica. Además, el número de individuos de la muestra es inferior a treinta. Para el cálculo de esta prueba se utilizó R Projects v3.2.4 para el análisis de los datos en estudio.

5.3.1 Prueba de la Hipótesis General:

Ho: Una arquitectura empresarial no mejora los análisis bioinformáticos de grandes volúmenes de datos.

Ha: Una arquitectura empresarial mejora los análisis bioinformáticos de grandes volúmenes de datos.

Los resultados del cálculo de la prueba de Wilcoxon usando R projects (Figura 54). Debido a que el p-valor obtenido 0.001513 es menor al nivel de significancia 0.05 (5%), se rechaza la hipótesis nula y se acepta la hipótesis alterna; por lo tanto, se afirma con un 99% de confianza que una arquitectura empresarial basada en el *cloud computing* contribuye positiva y significativamente en optimizar los análisis bioinformáticos de GVD. Así, de

este resultado se infiere que la implementación de una arquitectura empresarial beneficia al centro de investigación.

Figura 54: Wilcoxon - arquitectura empresarial

```
        Wilcoxon rank sum test with continuity correction

data:  arq_despues and arq_antes
W = 75, p-value = 0.001513
alternative hypothesis: true location shift is not equal to 0
```

Fuente: Elaboración del autor

5.3.2 Prueba de las Hipótesis Específicas:

H0: Una vista de negocio no mejora los procesos bioinformáticos.

Ha: Una vista de negocio mejora los procesos bioinformáticos.

Los resultados obtenidos después del análisis Wilcoxon realizado son los siguientes:

Figura 55: Wilcoxon – vista de negocio

```
        Wilcoxon rank sum test with continuity correction

data:  v_negocio_despues and v_negocio_antes
W = 75, p-value = 0.00166
alternative hypothesis: true location shift is not equal to 0
```

Fuente: Elaboración del autor

En el análisis se obtuvo el p-valor igual a 0.00166, debido a que este resultado es menor que 0.05 se rechaza la hipótesis nula y se acepta la hipótesis alterna. Es decir, se afirma con más del 99% de seguridad que poseer una vista de negocio mejora los procesos bioinformáticos, en efecto un mejor diseño de negocio basado en la arquitectura TOGAF contribuyo positivamente en mejorar los procesos y análisis bioinformáticos.

H0: Una vista de aplicaciones no mejora los trabajos de los investigadores.

Ha: Una vista de aplicaciones mejora los trabajos de los investigadores.

Los resultados obtenidos del análisis realizado en esta hipótesis específica, son los siguientes:

Figura 56: Wilcoxon - vista de aplicaciones

```
        Wilcoxon rank sum test with continuity correction

data:  v_app_despues and v_app_antes
W = 76.5, p-value = 0.0009973
alternative hypothesis: true location shift is not equal to 0
```

Fuente: Elaboración del autor

Debido a que el p-valor 0.0009973 es menor que 0.05 se rechaza la hipótesis nula y se acepta la hipótesis alterna. Es decir, se afirma con más del 99% de certeza que poseer una vista de aplicaciones contribuye positivamente en la realización en los trabajos de los investigadores, es así que esta arquitectura basada en TOGAF aporta beneficiosamente en los análisis que realizan los investigadores.

H0: Una vista de datos no permite mejorar la gestión de los GVD.

Ha: Una vista de datos permite mejorar la gestión de los GVD.

Los resultados obtenidos del análisis realizado en esta hipótesis específica, son los siguientes:

Figura 57: Wilcoxon - vista de datos

```
        Wilcoxon rank sum test with continuity correction

data:  v_datos_despues and v_datos_antes
W = 66.5, p-value = 0.01396
alternative hypothesis: true location shift is not equal to 0
```

Fuente: Elaboración del autor

En este análisis se obtuvo un p-valor 0.01396 es menor que 0.05 se rechaza la hipótesis nula y se acepta la hipótesis alterna. Se afirma con más del 98% que poseer una vista de datos permite mejorar la gestión de los grandes volúmenes de datos. En efecto, la gestión de este tipo de datos ahora

se realiza de manera más sencilla, proveyendo más accesibilidad, usabilidad y disminuyendo tiempo en las actividades de los investigadores.

H0: Una vista de infraestructura no permite optimizar los análisis bioinformáticos.

Ha: Una vista de infraestructura permite optimizar los análisis bioinformáticos.

Los resultados obtenidos del análisis realizado en esta hipótesis específica, son los siguientes:

Figura 58: Wilcoxon - análisis bioinformático

```
        Wilcoxon rank sum test with continuity correction

data:  cloud_comp_despues and cloud_comp_antes
W = 62.5, p-value = 0.03296
alternative hypothesis: true location shift is not equal to 0
```

Fuente: Elaboración del autor

En esta prueba se obtuvo el p-valor 0.03296 que es menor que 0.05 se rechaza la hipótesis nula y se acepta la hipótesis alterna. En este sentido, se afirma con más del 96% que poseer una vista de infraestructura basada en el *cloud computing* permite optimizar los análisis bioinformáticos. En efecto una arquitectura basada en el *cloud computing* ofrece muchos beneficios que apoyan ampliamente a los objetivos de las investigaciones.

Luego del análisis de los resultados de este capítulo, se concluye que la arquitectura empresarial basado en TOGAF permitió optimizar los análisis bioinformáticos de grandes volúmenes de datos, mediante un marco de trabajo que diseñó, planificó e implementó los diferentes componentes tecnológicos, aplicaciones y datos, cuyos resultados tienen soporte en los dos casos de estudio expuestos, el análisis de la encuesta realizada, así como en la validación de la prueba de hipótesis de la presente investigación.

CAPÍTULO VI
DISCUSIÓN

6.1 Discusión

En la presente investigación se diseñó y modeló la arquitectura empresarial para brindar servicios que colaboren con bioinformática, así mismo se presentó el modelo de aplicaciones o sistemas de información que soportan tecnológicamente ésta arquitectura.

En el diseño de una arquitectura empresarial el objetivo más importante es relacionar correctamente el objetivo de negocio con la parte tecnológica de modo que estos trabajen efectivamente hacia las metas organizacionales. Cabe resaltar, que para (Gomez, 2015) la definición de las diferencias entre arquitectura empresarial y plan estratégico de tecnologías de información se hace cada vez más imperceptible, pero hay diferencias que permiten definir claramente los límites de cada campo, la arquitectura empresarial determina la parte lógica de la organización de toda la empresa y el área de TI de la lógica de organización del área de sistemas. Las arquitecturas son generalmente constantes en el tiempo mientras que las tecnologías tienen un cambio más frecuente en el mercado, de este modo los problemas se tornan a problemas de negocio. Pero, cuando se genera una arquitectura empresarial, tal y como se desarrolló en esta investigación, se establece un mapa de ruta con los procesos que deben ser ejecutados, priorizados y dependiente entre ellos.

Generalmente los primeros resultados en la implementación de una arquitectura empresarial para (Gomez, 2015) están en “la categoría de ahorros, (en licenciamiento, hosting, personal, etc.) pero además en las decisiones de tecnología, datos o aplicaciones, incluso de inversión, éstos generan resultados de mayor impacto a nivel de negocio”, por ejemplo, mayor satisfacción en los científicos, tiempos de análisis optimizados o incluso nuevos y servicios de calidad, todo esto se ve reflejado en un área de bioinformática más productivo. La arquitectura empresarial se centra en la inversión y en los objetivos como resultados de la organización que agreguen valor óptimamente y coherentemente hacia los interesados, obteniendo asi la inversión en menos tiempo y mayor margen.

El *framework* de arquitectura empresarial “TOGAF es un método comprobado con años de investigación el cual fue desarrollado por arquitectos de talla mundial” (Bustamante, 2016). En este estudio se realizó una arquitectura empresarial para un campo específico e innovador como bioinformática, pero no se han realizados pruebas en áreas de bioinformática específicas y más profundas. Además, se demostró que TOGAF crea un enlace entre negocio y TI en el área de bioinformática, aportando beneficios como la minimización de costos y riesgos, identificando oportunidades y flexibilidad con un lenguaje común.

Durante la elaboración de la arquitectura, utilizando TOGAF, se definieron los conceptos necesarios para establecer la visión de la arquitectura, lo que permitió conocer el contexto de la organización, objetivos a futuro y los requerimientos del negocio. Se especificaron los interesados que participan pasiva o activamente, de esta forma se comprende el funcionamiento básico, y los procesos base que tenía el grupo antes de contar con una estructura organizacional establecida. Al entender los conceptos se hizo una investigación de la lógica del negocio en otros centros de investigación y una comparación de los principales procesos de negocio y el desarrollo de sus actividades evidenciando la necesidad de una estructura organizacional definida para alcanzar sus metas. De esta manera, se creó una visión a alto nivel que facilita el entendimiento de la complejidad de bioinformática

mediante la organización y presentación de artefactos que se conceptualizan y se describen.

La arquitectura propuesta define las actividades necesarias para el cumplimiento de los objetivos de los investigadores científicos, y establece flujos de procesos que permiten realizar de forma controlada y ordenada los análisis bioinformáticos. La investigación permite obtener y profundizar nuevos conocimientos y habilidades en el campo de la arquitectura empresarial con el *framework* TOGAF y sus diferentes técnicas. La investigación realizada se considera un buen punto de partida para dar paso a nuevas investigaciones que complementen el estudio realizado, se resalta la importancia de realizar investigaciones futuras de este tipo.

Tomando en consideración todo lo analizado, se puede determinar que una arquitectura empresarial es fundamental para lograr que TI soporte y facilite los procesos de negocio de una organización, ya que permite alinear la estrategia de negocio con la infraestructura de comunicación y los servicios de información de un centro de investigación.

CONCLUSIONES

1. La arquitectura empresarial basado en TOGAF permitió optimizar los análisis bioinformáticos de grandes volúmenes de datos mediante el diseño, planificación e implementación de componentes tecnológicos, aplicaciones y datos, siendo validado por dos casos de estudio y la prueba de hipótesis.

2. La arquitectura de negocio permitió definir, esquematizar, articular y optimizar los procesos bioinformáticos en las investigaciones mediante una estrategia de alineamiento a las necesidades y objetivos científicos.

3. La arquitectura de aplicaciones brindó sistemas de fácil acceso, disponibles y flexibles, un catálogo de software, que alinean las aplicaciones y las necesidades de las investigaciones científicas, mejorando la integración y la reutilización de los sistemas.

4. La arquitectura de datos permitió mejorar la gestión, manipulación y análisis de los grandes volúmenes de datos provenientes de las investigaciones, para tenerlos disponibles, accesibles e íntegros.

5. La arquitectura de la infraestructura tecnológica basada en el *cloud computing* permitió el mejor desarrollo de los proyectos bioinformáticos en términos de facilidad de uso, reducción de tiempo y costos, flexibilidad, escalabilidad y adaptabilidad a las necesidades cambiantes.

RECOMENDACIONES

1. Contar con el apoyo de la alta gerencia para la adopción de la arquitectura empresarial, debido a que el proyecto requiere disponer del tiempo de los directivos y de la información gerencial. Aún más, cuando las decisiones en inversiones en tecnología juegan un rol importante para el apoyo a las investigaciones.

2. Considerar la adecuada relación entre procesos, personas, servicios y aplicaciones en la arquitectura de negocio son clave, así como otros un modelo gerencial claro, estrategias bien definidas, el conocimiento de la organización y un adecuado desarrollo de las iniciativas de la organización. Así mismo, incluir un framework adecuado para maximizar los beneficios de la arquitectura empresarial y cumplir los objetivos y metas de la organización.

3. Impulsar las relaciones de trabajo de diferentes campos: la informática, estadística y la biología para crear herramientas útiles y significativas para la investigación y mejorar la arquitectura de aplicaciones. Por tanto, la comunicación es el factor clave para la mejor interacción entre profesionales multidisciplinarios, tales como biólogos e informáticos.

4. Optimizar los resultados procedentes de NGS en el contexto del diagnóstico genético y discriminar solo la información de utilidad en la arquitectura de datos. La estrategia de selección y captura de determinadas regiones genómicas de interés en el laboratorio permitiría limitar el análisis de

datos útil, lo cual disminuiría la complejidad del análisis, ahorros en tiempo y dinero.

5. Considerar un nuevo modelo de portabilidad que implique disponibilidad por parte de los proveedores. Actualmente, no existe un estándar ni normativas para la interoperabilidad y portabilidad de aplicaciones con respecto a los proveedores de *cloud*, lo cual implica dependencia absoluta de un solo proveedor.

FUENTES DE INFORMACIÓN

BIBLIOGRÁFICAS

Avila, O. (2011). Computación en la nube. ContactoS, 80, 45–52.

Bailey, M. (2011). *The Value of Smarter Datacenter Services.* IDC.

Castro, O. (2009). *Bioinformática: Interfaz gráfica para comparación de genomas vía web.* Barcelona, España.

De Las Rivas, J., Bonavides-Martínez, C., & Campos-Laborie, F. (2017). Bioinformatics in Latin America and SoIBio impact, a tale of spin-off and expansion around genomes and protein structures. Briefings in Bioinformatics, 1091.

Forrester Research. (2009). Which cloud computing platform is right for you? (F. Research, Ed.)

Frazier, M. E., Johnson, G. M., Thomassen, D. G., Oliver, C. E., & Patrinos, A. (2003). Realizing the potential of the genome revolution: the genomes to life program. *Science , 300*(5617), 290-293.

Hernández, R., Fernández, C., & Baptista, P. (2010). *Metodología de la investigación.* La Habana: Editorial Félix Varela.

Ison J, R. K. (2016). Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Research, 44*, D38-D47.

Kreuze, J. F., Fuentes, S., Perez, A., Galvez, M., Gutierrez, D., ..., & Flores, M. (2014). *The African sweetpotato virome: mapping and diversity of sweetpotato viruses.* International Potato Center, Lima.

Kutnjak, D. &. (2014). Complete genome sequences of new divergent potato virus X isolates and discrimination between strains in a mixed infection using small RNAs sequencing approach. *Virus Research, 191*, 45-50.

Lechuga, J. (2012). La bioinformática al servicio de la genómica. España: Univeridad de Santiago de Compostela.

Lee, C.-Y., Chiu, Y.-C., Wang, L.-B., Kuo, Y.-L., Chuang, E., Lai, L.-C., & Tsai, M.-H. (2013). Common applications of next-generation sequencing technologies in genomic research. *Transl Cancer Res, 2(1)*, 33-45.

Maya, J. (2010). Arquitectura empresarial: Un nuevo reto para las empresas de hoy.

Liang-Jie Zhang, F. (2012). Editorial: Big Services Era: Global Trends of *cloud computing* and *Big data*. *IEEE Vol. 5, No. 4*, 467.

Matute, M. M. (2014). Propuesta de *framework* de arquitectura empresarial para pymes basado en un análisis comparativo de los frameworks de Zachman y TOGAF . *Tesis de maestria*. España.

Myers, E. (2001). The sequence of the human genome. *Science*.

Nizamani, S. A. (2012). A Quality-aware Cloud Selection Service for Computational Modellers.

Orobitg, M. (2013). *High performance computing on biological sequence aligment.* Lleida: Universitat de Lleida.

Rosales Rosero, E. (2010). Una Cloud: Infraestructura como Servicio para Cloud . (U. d. Andes., Ed.).

Scaramella, J. (2016). *Why Upgrade Your Server Infrastructure Now?* IDC Document #US41292016.

Rutten, P. (2016). *HPE's Superdome X: The Mission-Critical Scale-Up x86 Platform for SAP, Oracle, and SQL Server* (IDC #US41302316 ed.). International Data Corporation (IDC).

The Open Group. (2013). *TOGAF.* Amersfoort: Van Haren Publishing.

Zerbino, D., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. 18(5), 821-9.

ELECTRÓNICAS

Aguilar, S. (2018). Ques es big data. Obtenido de https://www.academia.edu/11432044/QU%C3%89_ES_BIG_DATA

Alemán, M., Báez, E., Fernández, J., & Morales, J. (2003). Bioinformática: Algunos retos filosóficos, sociales, éticos y jurídicos. Recuperado el 2019, de http://www.revmatanzas.sld.cu/revista%20medica /ano%202003/vol2%202003/tema07.htm

AWS. (s.f.). Amazon Web Services. Recuperado el 01 de 02 de 2015, de https://aws.amazon.com/

Barraza, F. (2014). Introducción a la Bioinformática. Recuperado el 2016, de http://cic.javerianacali.edu.co/wiki/lib/exe/fetch.php?media=materias:bioinfo_sesion1.pdf

Bustamante, J. (2016). Arquitectura Empresarial basado en TOGAF. Recuperado el 2016, de http://www.ucipfg.com/Repositorio/MATI/MATI-04/BLOQUE-ACADEMICO/Unidad-1/lecturas/Resumen-01-Introduccion.pdf

Campos, R., Campos, G., & Ortiz, M. (2011). Concepto, aplicación y tendencia de la Bioinformática como Tecnología de Información. Recuperado el 2019, de http://www.iiis.org/CDs2011/ CD2011CSC/CISCI_2011/PapersPdf/CA633GF.pdf

Centro internacional de la papa. (2015). *Misión del CIP.* Obtenido de https://cipotato.org/

CIMMYT. (s.f.). CIMMYT. Recuperado el 20 de 02 de 2016, de https://www.cimmyt.org/

CIAT. (s.f.). CIAT. Recuperado el 21 de 02 de 2016, de https://ciat.cgiar.org/

Froilan, F. (2010). Un salto a la nube. Obtenido de http://virtual.iesa.edu.ve/servicios/wordpress/wp-content/uploads/2013/09/e-10froiolan.pdf

Gartner Group. (s.f.). *cloud computing* . Recuperado el 2 de Abril de 2015, de http://www.gartner.com/it-glossary/cloud-computing/

Gartner, Inc. (2017). *Magic Quadrant for Cloud Infrastructure as a Service, Worldwide*. Obtenido de https://www.gartner.com/doc/3738058

Gomez. (2015). Arquitectura empresarial. Recuperado el 01 de 08 de 2016, de https://chae201511700921759.wordpress.com/2015/02/10/que-es-arquitectura-empresarial/

Gonzalez, J. (2016). Análisis de secuencias. Obtenido de http://rpubs.com/jagonzalez/seqAnalysis

IRRI. (s.f.). IRRI. Recuperado el 21 de 02 de 2016, de https://irri.org/

Joyanes, L. (2003). La bioinformática como convergencia de la biotecnología y la informática . Recuperado el 2017, de http://files.bioinformaticos.webnode.es/200000061-76a4d77a20/Articulo%20bioinformatica%20y%20biotecnologia.pdf.pdf

Kuthanur. (2015). *What is an Enterprise Architecture Maturity Model? .* Obtenido de https://blogs.informatica.com/2015/04/06/enterprise-architecture-maturity-model/#fbid=P1DBfGuoDIt

Lindsey, M. (2010). *Best Practices for Sourcing cloud computing Services.* Obtenido de https://www.eiseverywhere.com/file_uploads/acfebd89b8db4ffaae36858b3f18d066_Lindsey_Wednesday_1140_SNWF10.pdf

National Library of Medicine. (10 de Abril de 2018). *National Center for Biotechnology Information*. Obtenido de https://www.nlm.nih.gov/about/2019CJ.html

PACIFICTEL S.A. (2006). Manual de políticas general. Recuperado el 05 de 2016, de https://www.uv.mx/personal/fcastaneda/files/2010/10/manual_politicas_pacifictel.pdf

Pasteur. (21 de febrero de 2016). Pasteur Inttitute. Obtenido de https://www.pasteur.fr/en

Real Academia Española. (s. f.). Bioinformática. En Diccionario de la lengua española (avance de la 23.a ed.). Recuperado de http://buscon.rae.es/draeI/SrvltConsulta?TIPO_BUS=3&LEMA=bioinformatica

Routio. (2007). Arte como análisis. Recuperado el 2019, de http://www2.uiah.fi/projects/metodi/276.htm

Sanger. (21 de febrero de 2016). Sanger. Obtenido de https://www.sanger.ac.uk/

Stonebraker, M. (21 de September de 2012). *Communications of the ACM*. Recuperado de: http://cacm.acm.org/blogs/blog-cacm/155468-what-does-big-data-mean/fulltext

University of Essex. (2015). UK Data Archive. Obtenido de http://www.data-archive.ac.uk/create-manage/life-cycle

Universidad de Valencia. (2016). Las ventajas reales que nos ofrece la bioinformática. Recuperado el 2016, de https://www.uv.es/master-id-biotecnologia-biomedicina/es/blog-1285955365462/GasetaRecerca.html?id=1285964672071

Valdiosera, C. (2006). Biocomputación, un nuevo enfoque . Recuperado el 2018, de https://www.jornada.com.mx/2006/03/09/index.php?section=ciencias&article=a05n1cie

Wikipedia. (s.f.). Alfa de Cronbach. Recuperado el 12 de 04 de 2016, de https://es.wikipedia.org/wiki/Alfa_de_Cronbach

Zhang, L.-J. (. (2012). Big Services Era: Global Trends of *cloud computing* and *Big data*. *IEEE, v5*, n4. Obtenido de http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6365210

ÍNDICE DE ANEXOS

ANEXOS

Anexo 1: Encuesta

¿Se encuentra satisfecho con el nivel de disponibilidad de los sistemas para su uso? 5. Muy satisfecho 4. Satisfecho 3. Normal 2. Poco Satisfecho 1. Insatisfecho
¿Considera que la implementación de soluciones en TI contribuye a la disminución de costos? 5. Muy de acuerdo 4. De acuerdo 3. Ni de acuerdo ni en desacuerdo 2. Poco de acuerdo 1. Desacuerdo
¿Considera que los sistemas de información rápidamente se adaptan a cambios? 5. Muy de acuerdo 4. De acuerdo 3. Ni de acuerdo ni en desacuerdo 2. Poco de acuerdo 1. Desacuerdo
¿Considera que los procesos desarrollados en su trabajo apoyan a los objetivos de la investigación? 5. Muy de acuerdo 4. De acuerdo 3. Ni de acuerdo ni en desacuerdo 2. Poco de acuerdo 1. Desacuerdo
¿Considera que las aplicaciones trabajan adecuadamente con los grandes volúmenes de datos? 5. Muy de acuerdo 4. De acuerdo 3. Ni de acuerdo ni en desacuerdo 2. Poco de acuerdo 1. Desacuerdo
Cree Ud. que las inversiones en TI aportan al objetivo científico. 5. Muy de acuerdo 4. De acuerdo 3. Ni de acuerdo ni en desacuerdo 2. Poco de acuerdo 1. Desacuerdo
¿Recibe/formula quejas o reclamos por cambios realizados en sistemas, debido a algún proyecto de sistemas/TI? 5. Nunca 4. Casi nunca 3. Pocas veces 2. Seguido 1. Muy seguido
Existe colaboración de los investigadores científicos en el desarrollo de proyectos en TI 5. Muy seguido 4. Seguido 3. Pocas veces 2. Casi nunca 1. Nunca
¿Se encuentra satisfecho con los canales de comunicación del personal de TI disponibles en el centro? 5. Muy satisfecho 4. Satisfecho 3. Normal 2. Poco Satisfecho 1. Insatisfecho
¿Se encuentra satisfecho con la gestión de los grandes volúmenes de datos relacionados a su trabajo? 5. Muy satisfecho 4. Satisfecho 3. Normal 2. Poco Satisfecho 1. Insatisfecho
¿Se encuentra satisfecho con la facilidad de uso de los sistemas de información que utiliza? 5. Muy satisfecho 4. Satisfecho 3. Normal 2. Poco Satisfecho 1. Insatisfecho
¿Cuál es el nivel de satisfacción con respecto a los servicios de TI servidores / *cloud computing*? 5. Muy satisfecho 4. Satisfecho 3. Normal 2. Poco Satisfecho 1. Insatisfecho
¿Considera que es los grandes volúmenes de datos siempre se encuentran disponibles? 5. Muy de acuerdo 4. De acuerdo 3. Ni de acuerdo ni en desacuerdo 2. Poco de acuerdo 1. Desacuerdo
¿Considera que el sistema de información reduce el tiempo de ejecución de sus actividades? 5. Muy seguido 4. Seguido 3. Pocas veces 2. Casi nunca 1. Nunca

Anexo 2: Datos obtenidos en las encuestas

Preguntas/ Encuestados	P1	P2	P3	P4	P5	P6	P7	P8	P9	Varianza	Mediana
Q1	1	3	2	3	2	2	2	3	2	0.444	2
Q2	3	2	3	3	2	3	3	4	3	0.361	3
Q3	3	3	3	4	3	3	2	3	3	0.250	3
Q4	2	3	3	2	2	2	3	3	3	0.278	3
Q5	1	2	2	3	3	2	2	3	2	0.444	3
Q6	2	3	2	3	2	3	2	3	2	0.278	2
Q7	2	3	2	2	3	2	3	3	3	0.278	3
Q8	1	2	3	3	2	2	2	3	3	0.500	2
Q9	1	2	2	3	3	3	2	3	2	0.500	2
Q10	2	3	3	3	3	3	2	4	3	0.361	3
Q11	2	2	2	3	3	3	2	3	2	0.278	2
Q12	3	2	3	4	3	3	3	3	2	0.361	3
Q13	3	3	2	3	3	3	2	3	2	0.250	3
Q14	3	2	3	3	3	3	3	3	2	0.194	3

Las preguntas (Qi) se refieren a las preguntas del Anexo 1 aplicadas como pre prueba.

Preguntas/ Encuestados	P1	P2	P3	P4	P5	P6	P7	P8	P9	Varianza	Mediana
Q1	4	4	3	5	4	4	4	4	5	0.321	4
Q2	4	3	4	4	4	4	4	4	5	0.222	4
Q3	4	3	3	4	4	3	4	4	5	0.395	4
Q4	5	4	4	5	3	5	4	4	5	0.444	4
Q5	4	4	3	4	3	4	4	3	4	0. 222	3
Q6	4	4	3	5	3	3	4	4	4	0.395	4
Q7	3	4	3	4	4	3	4	4	5	0.395	4
Q8	3	4	3	4	4	4	3	4	4	0.222	4
Q9	4	3	4	4	5	3	4	3	4	0.395	3
Q10	4	3	3	5	4	4	3	4	4	0.395	3
Q11	4	4	4	4	4	4	5	3	5	0.321	4
Q12	4	3	3	5	4	4	5	4	4	0.444	4
Q13	4	3	3	4	3	4	3	4	4	0.247	4
Q14	4	5	3	4	4	4	4	5	4	0.321	4

Las preguntas (Qi) se refieren a las preguntas del Anexo 1 aplicadas como post prueba.

Anexo 3: Tabulaciones de la encuesta

Preguntas/ Respuestas	1. Desacuerdo	2. Poco de acuerdo	3. Ni de acuerdo ni desacuerdo	4. De Acuerdo	5. Muy de acuerdo	Total
Q1	11%	56%	33%	0%	0%	100%
Q2	0%	22%	67%	11%	0%	100%
Q3	0%	11%	78%	11%	0%	100%
Q4	0%	44%	56%	0%	0%	100%
Q5	11%	56%	33%	0%	0%	100%
Q6	0%	56%	44%	0%	0%	100%
Q7	0%	44%	56%	0%	0%	100%
Q8	11%	44%	44%	0%	0%	100%
Q9	11%	44%	44%	0%	0%	100%
Q10	0%	22%	67%	11%	0%	100%
Q11	0%	56%	44%	0%	0%	100%
Q12	0%	22%	67%	11%	0%	100%
Q13	0%	33%	67%	0%	0%	100%
Q14	0%	22%	78%	0%	0%	100%

Las preguntas (Qi) se refieren a las preguntas del Anexo 1 como pre prueba.

Preguntas/ Respuestas	1. Desacuerdo	2. Poco de acuerdo	3. Ni de acuerdo ni desacuerdo	4. De Acuerdo	5. Muy de acuerdo	Total
Q1	0%	0%	11%	67%	22%	100%
Q2	0%	0%	11%	78%	11%	100%
Q3	0%	0%	33%	56%	11%	100%
Q4	0%	0%	11%	44%	44%	100%
Q5	0%	0%	56%	44%	0%	100%
Q6	0%	0%	33%	56%	11%	100%
Q7	0%	0%	33%	56%	11%	100%
Q8	0%	0%	33%	67%	0%	100%
Q9	0%	0%	33%	56%	11%	100%
Q10	0%	0%	33%	56%	11%	100%
Q11	0%	0%	11%	67%	22%	100%
Q12	0%	0%	22%	56%	22%	100%
Q13	0%	0%	44%	56%	0%	100%
Q14	0%	0%	11%	67%	22%	100%

Las preguntas (Qi) se refieren a las preguntas del Anexo 1 como post prueba.

Anexo 4: Comparación de plataformas NGS (Chien-Yueh Lee, 2013)

Table 1 Comparison of next generation sequencing platforms

Company	Sequencing Principle	Detection	System platform	Read length (bp)	Number of Reads	Time/run	Throughput/run	Accuracy	Machine cost ($)	Advantage	Disadvantage
Illumina	Reversible terminator sequencing by synthesis	Fluorescence/ Optical	HiSeq 2500/1500	36/50/100	3 billion (SE)	2~11 days	600 GB	> 99%	740,000	Very high throughput; Cost-effectiveness; Steadily improving read lengths; Massive throughput	Long run time; Short read lengths; Expensive instrument; Lower error rate
			Genome Analyzer IIx	35/50/75/100	320 million (SE)	2~14 days	95 GB	> 99%	250,000	High throughput; The most widely used platform	Low multiplexing capability of samples
			MiSeq	25/36/100/ 150/250	17 million (SE)	4~27 hours	8.5 GB	> 99%	125,000	High throughput; Cost-effectiveness; Short run times; Appropriate throughput for microbial applications; Minimal hands-on time; High coverage	Short read lengths
Roche	Pyrosequencing	Optical	454 GS FLX+	700	1 million	23 hours	0.7 GB	99.997%	450,000	High throughput; Longer read lengths; Short run times; High coverage	Appreciable hands-on time; High reagent costs; Higher error rate in homopolymers regions
			454 GS Junior	400	1 million	10 hours	0.035 GB	> 99%	108,000	Longer read lengths; Short run times	
Helicos Biosciences	Single molecule sequencing	Fluorescence/ Optical	Heliscope	25~55 (average: 32)	600~800 million	8 days	37 GB	99.99%	999,000	Single-molecule nature of technology; Non-bias representation of templates for genome	Expensive instrument; Very short read lengths (increase cost and difficulty of assembly); Higher error rate
ABI Life Technologies	Ligation	Fluorescence/ Optical	5500 SOLiD	75+35	1.4 billion	7 days	90 GB	99.99%	350,000	High throughput; Lowest reagent cost	Long run times; Very short read lengths (increase cost and difficulty of assembly)
			5500xl SOLiD	75+35	2.8 billion	7 days	180 GB	99.99%	595,000	Very high throughput; Low error rate; Massive throughput	
	Proton detection	Change in pH detected by Ion-Sensitive Field Effect Transistors (ISFETs)	Ion Personal Genome Machine (PGM)	35/200/400	12 million	2 hours	2 GB	> 99%	80,000	Short run times; Low cost per sample; Appropriate throughput for microbial applications; Direct measurement of nucleobase incorporation events	Appreciable hands-on time; High reagent costs; Higher error rate in homopolymers (sequential washing steps)
			Ion Proton Chip I/II	Up to 200	60-80 million	2 hours	10 GB / 100 GB	> 99%	243,000	Short run times; Flexible chip reagents	Instrument not available at time of writing
Pacific Bioscience	Real-time, single molecule DNA sequencing	Fluorescence/ Optical	PacBio RS	Average: 3000	~50 K	2 hours	13 GB	84~85%	750,000	Short run times; Very long read lengths; Low reagent costs; Simple sample preparation	No paired reads; Highest error rates; Expensive instrument; Difficult installation
Oxford Nanopore	Nanopore exonuclease sequencing	Electrical Conductivity	gridION	Tens of Kb	4~10 million	According to experiment	Tens of GB	96%	According to experiment	Extremely long read lengths; Low cost of α-HL nanopore production; Customization; No fluorescent labeling; No optics	4% error rates; Cleaved nucleotide may be read in the wrong order; Difficult to fabricate a device with multiple parallel pores

Anexo 5: Completa secuencia genómica de Potato Virus X (Kutnjak, 2014)

Virus Research 191 (2014) 45–50

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

Short communication

Complete genome sequences of new divergent potato virus X isolates and discrimination between strains in a mixed infection using small RNAs sequencing approach

Denis Kutnjak[a], Rocio Silvestre[b], Wilmer Cuellar[b,1], Wilmer Perez[b], Giovanna Müller[b], Maja Ravnikar[a], Jan Kreuze[b,*]

[a] Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia
[b] International Potato Center (CIP), Lima, Peru

ARTICLE INFO

Article history:
Received 25 April 2014
Received in revised form 9 July 2014
Accepted 12 July 2014
Available online 19 July 2014

Keywords:
Andes
Next generation sequencing
Phylogeny
Potato virus X
Small RNAs

ABSTRACT

Potato virus X (PVX; genus *Potexvirus*, family *Alphaflexiviridae*, order *Tymovirales*) is one of the most widespread and intensively studied viruses of potato. However, little is known about its diversity in its likely center of radiation, the Andean region of South America. To fill this gap, the strategy of Illumina deep sequencing of small RNAs was used to obtain complete or near complete genome sequence of PVX from 5 symptomatically infected greenhouse and 3 field samples (*Solanum tuberosum*) from Peru. PVX sequences determined in this study were assigned into three different phylogenetic groups of isolates. Notably, a complete genome sequence of a representative of a new PVX phylogenetic lineage was obtained, which shows a high level of sequence dissimilarity to other completely sequenced isolates (~17%). The new PVX genotype was detected in greenhouse and field samples. One of the field samples was infected with the mixture of two PVX strains, which were efficiently discriminated using small RNA sequencing approach. The study confirms the utility of small RNAs deep sequencing for successful viral strain differentiation and discovery of new viral strains and indicates a high diversity of PVX in the Andean region of South America, a pattern which may be expected also for other potato pathogens.

Isolates of *Potato virus X* (PVX), the type member of the genus *Potexvirus* (family: *Alphaflexiviridae*), are present world-wide in potato growing areas, infecting a wide range of hosts, notably within the *Solanaceae* family (King et al., 2011). It has been investigated intensively and thus become one of the most important models for studies of plant–virus interactions (Scholthof et al., 2011). Alone it is usually moderately pathogenic but it can cause a significant economic loss in synergistic co-infection with some potyviruses, especially potato virus Y (Khurana and Singh, 1988; Vance, 1991). PVX virions are flexuous filaments, 470–580 nm long. They contain a single positive sense ssRNA molecule (~6400 bp) with five open reading frames (ORFs) coding for RNA dependent RNA polymerase (RdRp), three proteins involved in cell-to-cell movement – the triple gene block (TGB) – and a coat protein (CP) (King et al., 2011).

According to the hypersensitive reaction induced by N_b and N_x resistance genes on potato (*Solanum tuberosum* L.) cultivars carrying them, PVX strains can be divided into four strain groups (Cockerham, 1955; Santa Cruz and Baulcombe, 1993). Group 1 cannot overcome either gene, group 2 overcomes N_x, group 3 N_b, and group 4 overcomes both genes. Furthermore, some isolates (PVX-HB, PVX-MS) overcome both N_b and N_x and also the extreme resistance gene R_x (Moreira et al., 1980; Jones, 1985; Querci et al., 1995; Feigelstock et al., 1995). "Gene for gene" interactions of these three host resistance genes with specific determinants in virus genomes have been determined (reviewed in Malcuit et al., 2000) in coat protein (for N_x and R_x) and the 25 kDa movement protein region (for N_b). There is no evidence for recombination between different PVX strains in nature. Thus, the acquisition of virulence determinants (or loss of virulence ones) can only be explained as a result of independent evolution (Malcuit et al., 2000).

Within PVX there is a phylogeographic split separating majority Eurasian (clade I) and South American (clade II) groups of isolates (Yu et al., 2008, 2010; Cox and Jones, 2010). However, some

* Corresponding author at: International Potato Center (CIP), Avenida La Molina 1895, La Molina Apartado Postal 1558, Lima, Peru. Tel.: +51 1 349 6017x3054; fax: +51 1 317 5326.
E-mail address: j.kreuze@cgiar.org (J. Kreuze).
[1] Current affiliation: International Center for Tropical Agriculture (CIAT), Cali, Colombia.

http://dx.doi.org/10.1016/j.virusres.2014.07.012
0168-1702/

Anexo 6: Poster *The African sweetpotato virome* (Kreuze & Col., 2014)

Game-changing Solutio

The African sweetpotato virome:

mapping and diversity of sweetpotato viruses

Jan F. Kreuze[1]
Segundo Fuentes[1]
Ana Perez[1]
Joao De Souza[1]
Marco Galvez[1]
Dina L. Gutierrez[1]
Mirella Flores[1]
Shan Gao[2]
Yi Zheng[2]
Zangjun Fei[2]

[1] Virology Laboratory
International Potato Center
P.O. Box 1558
Lima 12, Peru

[2] Bioinformatics Laboratory
Boyce Thompson Institute
533 Tower Road Ithaca
Ithaca, NY 14853, USA

Introduction

Viral diseases are considered a major constraint for sweetpotato production worldwide, particularly in Sub-Saharan Africa (SSA). With few exceptions, there is no clear information about the prevalence and distribution of the different sweetpotato viruses and their respective strains. This basic information is essential not only to manage the spread but also to evaluate the impact of these viral diseases. Taking advantage of the next-generation sequencing technology, this project has concentrated its efforts to determine the sweetpotato virome in Africa.

Methodology

Collections of field-grown sweetpotato samples (Fig. 1) have been carried out in 13 countries (Ghana, Mozambique, Ethiopia, Tanzania, Guinea, Nigeria, Benin, Malawi, Zambia, Zimbabwe, Angola, Kenya, and Uganda) across SSA.

Viruses have been identified by a generic method for virus diagnosis developed at CIP (1) called "small RNA sequencing and assembly (sRSA)". Virus identities were confirmed by PCR, and new primers were designed to fill the gaps on the genome sequences of the new viruses and variants. PCR amplified products were confirmed by Sanger sequencing.

Figure 1. Symptoms induced by virus infection in sweetpotato plants collected in Ghana (A,B) and Mozambique (C,D).

Results

While many samples are still being processed, results from Ghana (Fig. 2) and Mozambique have already led to the characterization of near complete genomes of new sweetpotato viruses (SPVZ [Fig. 3], alphasatellite). This also revealed the underlying genetic variation of known viruses in distinct geographic regions of Africa (Fig. 4), including new strains of SPCSV and SPMMV.

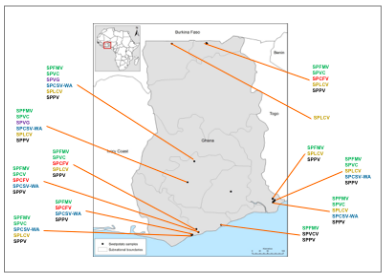


Figure 2. Sweetpotato viruses in Ghana. **SPFMV**: *Sweet potato feathery mottle virus*, **SPVC**: *Sweet potato virus C*, **SPVG**: *Sweet potato virus G*, **SPCSV-WA**: *Sweet potato chlorotic stunt virus* – West Africa, **SPLCV**: *Sweet potato leaf curl virus*, **SPCFV**: *Sweet potato chlorotic fleck virus*, **SPPV**: Sweet potato pakakuy virus, **SPVCV**: Sweet potato vein clearing virus.

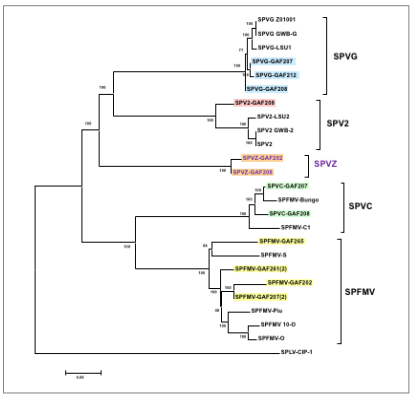


Figure 3. Variability of sweetpotato potyviruses. GAF202, GAF207, GAF208, and GAF212: samples from Mozambique. GAF261 and GAF265: samples from Ghana. **SPV2**: *Sweet potato virus 2*, **SPVZ**: Sweet potato virus Z, **SPLV**: *Sweet potato latent virus*.

Conclusions

Nearly all plants collected in Ghana and Mozambique were infected by at least one virus; however, multiple infections were most common.

About 12 sweetpotato viruses and their variants have been identified, including a new potyvirus (SPVZ) and an alphasatellite.

Understanding virus distribution and variability will enable an adequate management of the sweetpotato viral diseases across the African continent.

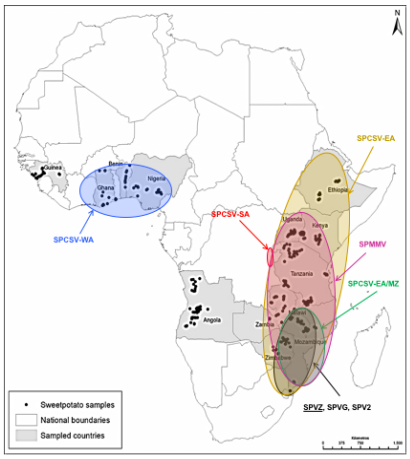


Figure 4. Distribution of sweetpotato virus populations across Africa. **SPMMV**: *Sweet potato mild mottle virus*, **SPCSV**: *Sweet potato chlorotic stunt virus*, **WA**: WA-strain, **SA**: SA-strain, **EA**: EA-strain **MZ**: MZ-strain.

Bibliography

1. Kreuze, J.F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., Simon, R. (2009). Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388: 1-7

Acknowledgements

We thank our collaborators for their support during the sweetpotato collections in Africa: Martine Zandjanakou-Tachin, Steffen Schulz, Maria Andrade, Douglas Miano, Joseph Ndunguru, Settumba Mukassa, Putri E. Abidin, Elizabeth Ngadze, Martin Chiona, Emily Mueller, Britta Kowalski, Joseph Nii Lamptey, Edward Carrey.

Anexo 7: Cálculos financieros

Cálculo de costos de servidor *in house:*

Año	0	1	2	3	4	5
Costo de servidor	-32800	0	0	0	0	0
Energía y alimentación	0	8790.4	8790.4	8790.4	8790.4	8790.4
Gestión y administración del servidor	0	13776	18368	24600	34112	43952
Costo Total	0	22566.4	27158.4	33390.4	42902.4	52742.4

Cálculos de VPN y TIR

Flujo de efectivo	-32800.00	3492.12	8084.12	14316.12	23828.12	33668.1
VP	-32800.00	3357.81	7474.23	12726.98	20368.38	27672.7

	S/.
VPN	38,800.15
TIR	23%