



UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería de la Salud



INGENIERÍA
DE LA SALUD

**TFG del Grado en Ingeniería de la
Salud**

Clinical Report Scraping

Presentado por Samuel González Martín
en Universidad de Burgos

11 de junio de 2024

Tutores: Antonio Jesús Canepa Oneto – Patricia Saiz
López



UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería de la Salud



D. Antonio Jesús Canepa Oneto, profesor del departamento de Ingeniería Informática, área de Lenguajes y Sistemas Informáticos.

Dra. Patricia Saiz López, doctora de anatomía patológica del Hospital Universitario de Burgos.

Expone:

Que el alumno D. Samuel González Martín, con DNI 71307513S, ha realizado el Trabajo final de Grado en Ingeniería de la Salud titulado: Clinical Report Scraping.

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección del que suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Burgos, 11 de junio de 2024

Vº. Bº. del Tutor:

Vº. Bº. del Tutor:

D. Antonio Jesús Canepa Oneto

Dra. Patricia Saiz López

Resumen

El presente trabajo de fin de grado (TFG) se enmarca en el desarrollo de una aplicación informática para el departamento de anatomía patológica del Hospital Universitario de Burgos. El objetivo principal es automatizar la extracción de información relevante a partir de informes PDF generados por el software Oncomine Reporter, optimizando el flujo de trabajo y minimizando errores humanos.

La aplicación desarrollada permite extraer de forma precisa y eficiente la información clínica de los informes PDF, reduciendo significativamente el tiempo dedicado a la lectura manual y mejorando la calidad de los datos recopilados. Además, se ha creado una base de datos para almacenar la información extraída, facilitando su consulta y análisis posterior.

Descriptores

Shiny app, Rstudio, anatomía patológica, Oncomine Reporter, PDFscraping, ClinVar, MongoDB.

Abstract

The present undergraduate thesis (TFG) involves the development of a software application for the Department of Pathological Anatomy at the University Hospital of Burgos. The primary objective is to automate the extraction of relevant information from PDF reports generated by the Oncomine Reporter software, optimizing the workflow and minimizing human errors.

The developed application accurately and efficiently extracts clinical information from the PDF reports, significantly reducing the time required for manual reading and improving the quality of the collected data. Additionally, a database has been created to store the extracted information, facilitating subsequent consultation and analysis.

Keywords

Shiny app, Rstudio, Pathological anatomy, Oncomine Reporter, PDFscraping, ClinVar, MongoDB.

Índice general

Índice general	iii
Índice de figuras	iv
Índice de tablas	v
Introducción	1
Objetivos	3
Conceptos teóricos	5
3.1. Anatomía patológica	5
3.2. Estado del arte y trabajos relacionados.	11
Metodología	13
4.1. Descripción de los datos.	13
4.2. Técnicas y herramientas.	13
Resultados	25
5.1. Resumen de resultados.	25
5.2. Discusión.	31
Conclusiones	33
6.1. Aspectos relevantes.	33
Lineas de trabajo futuras	35
Bibliografía	37

Índice de figuras

3.1. Historia de una biopsia	6
3.2. Secuenciación por Síntesis	9
4.1. Ejemplo tabla	21
5.1. Tabla de genes mutados	27
5.2. Tabla de genes patogénicos	27
5.3. Tabla de datos en MongoDBCompass	29
5.4. Acceso a tabla anidada	29
5.5. Tabla anidada	30
5.6. Ejemplo de búsqueda	30

Índice de tablas

5.1. Comparativa entre las dos tablas mostradas en la app	27
---	----

Introducción

La anatomía patológica es la rama de la medicina que estudia el efecto de las enfermedades en los órganos del cuerpo, tanto en su conjunto como en el ámbito microscópico. Su objetivo principal es diagnosticar y entender las alteraciones estructurales y funcionales que las enfermedades provocan en el cuerpo humano.

En el ámbito de la anatomía patológica, la gestión eficiente de los datos genéticos es fundamental para el diagnóstico y tratamiento de los pacientes. El software Oncomine Reporter [[ThermoFisher, 2022](#)] desempeña un papel crucial al documentar las mutaciones encontradas en las genotecas de cada paciente y generar informes en formato PDF. Sin embargo, el procesamiento manual de estos documentos es una tarea laboriosa y propensa a errores, lo que resalta la necesidad de una solución más automatizada y eficiente.

Este proyecto se centra en el desarrollo de una aplicación Shiny [[Shiny, 2020](#)], programada en el lenguaje R, diseñada para cubrir esta necesidad específica del servicio de anatomía patológica del Hospital Universitario de Burgos. La aplicación se encarga de extraer y procesar automáticamente la información contenida en los documentos PDF generados por Oncomine Reporter.

Los informes no solo detallan las mutaciones detectadas, sino que también incluyen información relevante de los pacientes, como el número de historia clínica y el sexo. La aplicación Shiny ha sido programada para identificar y organizar estos datos en tablas claras y accesibles. Estos datos son de gran utilidad para el posterior análisis de los llamados datos en vida real (RWD), información sobre el estado de salud de los pacientes y la prestación de servicios sanitarios. El objetivo del análisis de RWD es generar evidencia en vida real (RWE) [[FarmaIndustria, 2024](#)].

Además, la aplicación mejora significativamente el valor de los informes al integrar mediante API datos adicionales provenientes de la base de datos ClinVar, proporcionando el significado clínico asociado a cada mutación detectada. Esta funcionalidad no solo facilita la lectura y comprensión de los informes, sino que también aporta un contexto clínico crucial para los profesionales de la salud.

La interfaz de la aplicación es diseñada para ser intuitiva y fácil de usar, permitiendo a los usuarios visualizar los datos en tablas organizadas. Asimismo, ofrece la opción de mantener un registro con los detalles de cada paciente en una base de datos MongoDB, lo que mejora la accesibilidad y gestión de la información. Los usuarios también tienen la posibilidad de descargar los datos en el formato deseado, proporcionando una flexibilidad adicional para la documentación y el análisis.

Objetivos

Este trabajo se ha pensado como una continuación a un TFG previo y su principal objetivo es la creación de una aplicación Shiny con la que extraer los datos relevantes de los informes clínicos obtenidos del software Oncomine Reporter empleado por el servicio de anatomía patológica del HUBU. A su vez, se ha buscado desarrollar una base de datos donde mantener un historial de los pacientes estudiados que facilite un posterior análisis de los llamados datos de mundo real.

Dentro de este objetivo se encuentran otros más específicos ordenados en tres subgrupos:

1. Los objetivos marcados por los requisitos del software/hardware/análisis
 - Implementar el código en R, favoreciendo su comprensión a los patólogos.
 - Desarrollar de una aplicación Shiny con una interfaz clara e intuitiva.
 - Alojar los datos obtenidos en la app en una base de datos NoSQL (MongoDB).
 - Consultar con API genes de patogenicidad desconocida en la base de datos de ClinVar para cumplimentar la información de Oncomine Reporter.
2. Los objetivos de carácter técnico, relativos a la calidad de los resultados, velocidad de ejecución, fiabilidad o similares.
 - Adaptar del código a los nuevos modelos de informe. generados por Oncomine Reporter.

- Implementar los cambios solicitados por el HUBU.
 - Optimizar y limpiar el código para minimizar los tiempos de espera.
3. Los objetivos de aprendizaje, relativos a aprender técnicas o herramientas de interés.
- Mejorar mis conocimientos a la hora de programar en R.
 - Buscar paquetes de R que fueran de utilidad para llevar a cabo la aplicación.
 - Estudiar y poner en práctica las bases sobre programación de aplicaciones Shiny.
 - Analizar las ventajas y desventajas del uso las bases de datos SQL y noSQL.
 - Aprender a trabajar con conjuntos de datos no estructurados.
 - Buscar documentación y aprender a trabajar con API clínicas.

Conceptos teóricos

Este proyecto ha sido desarrollado en conjunto con el departamento de anatomía patológica del Hospital Universitario de Burgos. Por ello se va a comenzar contextualizando sobre el origen de los datos empleados en el trabajo.

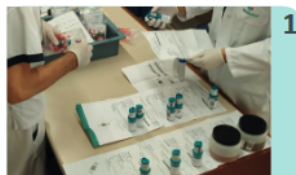
3.1. Anatomía patológica

La anatomía patológica es la rama de la medicina que estudia el efecto de las enfermedades en los órganos del cuerpo, tanto en su conjunto como en el ámbito microscópico. Su objetivo principal es diagnosticar y entender las alteraciones estructurales y funcionales que las enfermedades provocan en el cuerpo humano.

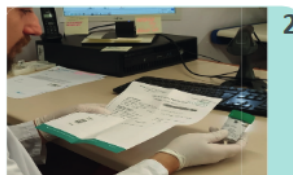
El camino que sigue una muestra desde que es recibida hasta que se consigue un informe clínico como los usados en la aplicación es el siguiente 3.1 [Alonso et al., 2023]:

HISTORIA DE UNA BIOPSIA. VIAJE POR ANATOMÍA PATOLÓGICA

La Anatomía Patológica suele ser la gran desconocida entre los procesos analíticos para el paciente. Cuando explicas que trabajas en este campo dentro de la sanidad, las personas que no están familiarizadas en este ámbito, suelen tener gran curiosidad porque les expliques en qué consiste. El equipo de Anatomía Patológica de Hospital Universitari General de Catalunya deseamos mostrar el camino y proceso que siguen todas las Biopsias que llegan a nuestro departamento y que pueda ayudar a entender esta labor minuciosa y de equipo.



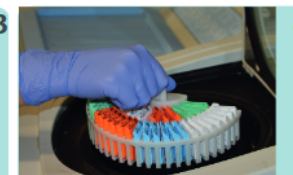
1 La biopsia, que siempre irá acompañada de su petición, puede llegar desde quirófano, Pequeña Cirugía o consultas transportada por el personal sanitario hasta Anatomía Patológica.



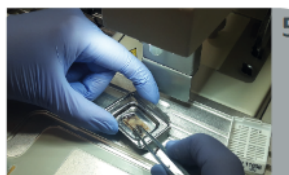
2 En Secretaría se verifican los datos del paciente, petición y muestra. Se le da un código bidimensional (QR) que le acompañará durante toda su trazabilidad.



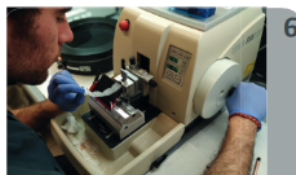
3 La muestra es analizada macroscópicamente, se toman secciones y se colocan en los cassettes impresos con código bidimensional (QR). Si la muestra es pequeña, se analiza todo.



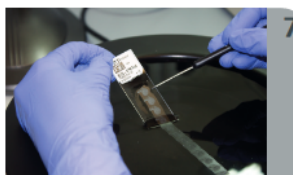
4 El tejido se procesa durante unas 14 horas para substituir el agua que contiene por parafina líquida a 60°.



5 Las secciones del tejido se colocan en moldes para confeccionar un bloque de parafina sólida.



6 Los bloques de parafina se cortan en secciones seriadas de 3µ con una herramienta de corte que se llama microtomo



7 Las secciones de tejido en parafina se estirarán en un baño de flotación de agua destilada a 37°C - 40°C y se pescan en un cristal (portaobjetos) identificado con su código bidimensional (QR)



8 Las secciones de tejido pescadas en el baño de flotación se colocan en bandejas y se tiñen de forma automatizada con Hematoxilina&Eosina



9 Los portaobjetos teñidos se etiquetan y se lee el código bidimensional (QR) para validar y asegurar su trazabilidad.



10 La imagen del portaobjeto puede ser digitalizada en un escáner.



11 El patólogo recupera el caso con el código bidimensional, lo estudia con pantalla o el microscopio y lo correlaciona con la historia clínica. El diagnóstico se redacta siguiendo unos protocolos internacionales.



12 El patólogo puede necesitar de estudios Inmunohistoquímicos y de biología molecular para completar el diagnóstico.



13 El informe se introduce en el sistema informático, donde lo podrán consultar los médicos peticionarios.



14 Una vez diagnosticados los casos, los cristales (portaobjetos) se archivan cuidadosamente y los bloques se guardan indefinidamente escaneados para poder ser recuperados en cualquier momento

Figura 3.1: Historia de una biopsia

Fijación y preparación de muestras

Cuando una biopsia llega a anatomía patológica, su preparación varía según su tipo y origen. Las muestras tisulares son fijación y posteriormente criogenizadas o, en la gran mayoría de casos, embebidas en parafina formando bloques. Una vez se tiene un bloque de parafina con la muestra, se corta en microtomos que son llevados a baños a 40 °C/50 °C para estirar la parafina.

Por otro lado, las muestras citológicas son conservadas en CytoLyt (mezcla de metanol y etanol) y sometidas a dos centrifugaciones para conseguir sedimentar el contenido celular.

Tras su respectivo baño o centrifugación, la muestra es recogida con un portaobjetos con el que se realizará la tinción pertinente. En el caso de las citologías, las más comunes son la tinción nuclear y la tinción de Papanicolau. Las muestras procedentes de bloques de parafina son tintadas según las indicaciones del profesional solicitante.

La tinción está automatizada gracias a dos máquinas, Leica ST5020-CV5030 de Leica Biosystems [Biosystems, 2024] y Ventana HE 600 de Roche [Roche, 2024]. Esta segunda realiza la tinción primaria hematoxilina eosina, no obstante, también desparafina y monta las muestras en un portaobjetos.

Celularidad Tumoral y Extracción de Ácidos Nucleicos

Una vez realizadas las tinciones demandadas, las muestras son enviadas al patólogo, el cual se encargará de determinar el porcentaje de celularidad tumoral de la muestra. La muestra deberá cumplir unos determinados estándares, en caso contrario se podrán llevar a cabo métodos de macro o microdisección con el fin de aumentar este porcentaje.

Las secciones de tejido son desparafinadas, hidrolizadas y centrifugadas para obtener el ADN y el ARN de las muestras. La cuantificación es realizada mediante fluorometría con Qubit [ThermoFisher, 2024a] y en función de las medidas obtenidas, se diluyen las muestras para obtener una concentración de 10 ng/ml.

Ion Chef

Tras la dilución se crean las librerías de ADN y ADNc, se retrotranscribe el ARN. Estas librerías son creadas gracias a Ion Chef [ThermoFisher, 2024b], un robot desarrollado por la empresa ThermoFisher capaz de preparar de forma automática bibliotecas, preparar plantillas y cargar chips.

Ion Chef es preparado con una placa que contiene un código de barras molecular para cada paciente (suelen ser 8 pacientes por carrera) y un panel con las regiones a amplificar. En este caso se está trabajando con un panel de 52 genes:

- 35 genes para SNV (*Single Nucleotide Variant*), entre las que se incluyen variantes de múltiples nucleótidos y las inserciones y deleciones.
- 19 genes para el estudio de variantes en el número de copias.
- 23 genes para el estudio de fusiones.

Los dos primeros grupos de genes son detectados en la secuencia de ADN, mientras que las fusiones son detectadas en la secuencia de ADNc.

En la fase de templado se combinan ambas muestras en una proporción de 4:1 (4 veces más de ADN que de ARN). Esta mezcla se lleva a cabo mediante pipeteo, asegurando una concentración adecuada de ADN y ARN para la secuenciación.

Las bibliotecas combinadas se transfieren nuevamente al Ion Chef, donde se distribuyen en placas. Los fragmentos de los genes a estudiar se unen a esferas magnéticas que se encuentran en los reactivos proporcionados por el fabricante. Estas esferas, por estequiometría, deberían ser monoclonales, lo que significa que cada esfera debe contener múltiples copias de un solo fragmento de ADN. A continuación, las esferas se someten a un proceso de amplificación para aumentar la cantidad de ADN disponible para la secuenciación.

Ion Torrent

Las esferas amplificadas se cargan en un chip especializado que contiene pocillos diseñados para alojar una única esfera con la secuencia de ADN a estudiar. El chip se introduce en el secuenciador y se procede a realizar la secuenciación por síntesis [Goodwin et al., 2016]. Este método se basa en la incorporación secuencial de desoxinucleótidos (dNTPs) a una cadena en crecimiento, donde cada dNTP debe ser añadido uno a la vez para asegurar que solo un tipo de nucleótido genere la señal. La plataforma Ion Torrent detecta los iones de hidrógeno liberados durante la incorporación de los dNTPs, provocando un cambio en el pH reflejado en un pico de voltaje 3.2 [Abdulkader et al., 2021].

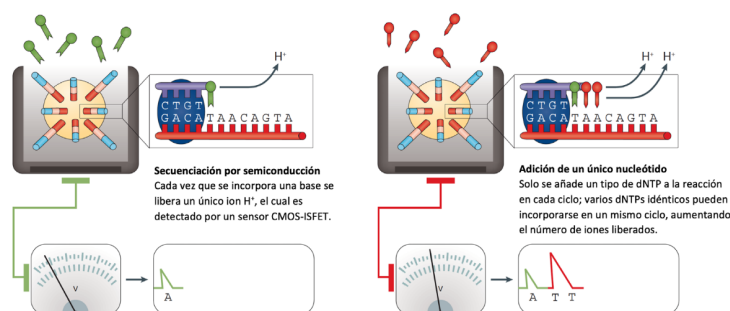


Figura 3.2: Secuenciación por Síntesis

Torrent Suite Software

Torrent Suite Software [ThermoFisher, 2024d] interpreta de forma automatizada los datos de la secuenciación. Gracias a este programa se puede:

- Rastrear y organizar los datos del estudio.
- Observar el progreso en tiempo real y monitorear las métricas clave.
- Consultar estadísticas de la ejecución mediante gráficos y métricas y así comprobar si la carrera cumple los parámetros de calidad.
- Exportar los datos en archivos BAM (*Binary Alignment Map*) a Ion Reporter Software [ThermoFisher, 2024c].

Un archivo BAM es un formato de archivo binario empleado en bioinformática para almacenar datos de alineación de secuencias.

Ion Reporter Software

Recoge los datos provenientes de Torrent Suite Software, analiza las variantes, filtra los resultados, produce anotaciones y genera informes. Ion Reporter ofrece control sobre las versiones y permite actualizaciones según las necesidades del servicio. Está diseñado con flujos de trabajo personalizados y automatizados. Finalmente, el producto que nos devuelve son archivos VCF en el que aparecen las muestras de calidad con anotaciones sobre las variantes genéticas, como variantes de nucleótidos únicos (SNVs), inserciones, deleciones y otras variaciones genéticas encontradas en secuencias genómicas. Estos archivos son subidos al software OncoPrint Reporter.

Oncomine Reporter

Oncomine Reporter es una herramienta web desarrollada por la empresa ThermoFisher para el análisis y visualización de datos en el campo de la oncología y la genómica del cáncer.

Este software produce informes con documentación relevante a las muestras de ADN de los pacientes. Por cada carrera de pacientes devuelve una carpeta con tantos PDF como pacientes haya en la carrera. A su vez, ofrece diversas plantillas flexibles para los informes con las que se puede seleccionar los datos que se quieren obtener y como va a ser su visualización [ThermoFisher, 2022].

Entre los posibles datos se encuentran los genes mutados y etiquetas descriptivas sobre los mismos (patogenicidad, frecuencia del alelo, exón, etc.), los biomarcadores relevantes o comentarios sobre la muestra.

Toda la información de los informes es contrastada y vinculada con fuentes públicas sobre terapias dirigidas, directrices y ensayos clínicos globales [of Veterans Affairs, 2024].

Datos en vida real y evidencia en vida real

Según la Agencia del Medicamento Estadounidense (FDA), los datos en vida real (RWD) se refieren al estado de salud de los pacientes y a la prestación de servicios sanitarios, recopilados rutinariamente a partir de diversas fuentes. Entre estas fuentes se incluyen las historias clínicas electrónicas, los procedimientos de reclamaciones y facturación, los registros de medicamentos y de pacientes, y los datos generados por los propios pacientes, como los obtenidos de aparatos de uso doméstico y dispositivos móviles. Estos datos proporcionan información valiosa que puede informar sobre diferentes estados de salud, ofreciendo una visión más completa y práctica del impacto de las intervenciones médicas en la vida cotidiana de los pacientes. Es por ello que se busca extraer la información relevante de los informes clínicos.

Los principales usos de los RWD incluyen complementar la información de los ensayos clínicos, reducir la incertidumbre en la evaluación de nuevos medicamentos, definir subpoblaciones de pacientes y áreas de efectividad de los fármacos, y favorecer el desarrollo y definición de guías de práctica clínica. Además, los RWD refuerzan el concepto de medicina basada en valor, aportan información relevante en el procedimiento de precio y reembolso, mejoran el acceso a nuevos medicamentos, potencian la investigación de nuevos

fármacos y reducen el coste de la I+D, así como mejoran los procedimientos de farmacovigilancia.

La evidencia en vida real, por otro lado, es el conocimiento derivado del análisis de los datos en vida real [FarmaIndustria, 2024].

3.2. Estado del arte y trabajos relacionados.

Actualmente, Oncomine Reporter ha sido remplazado en el mercado por softwares con mayor capacidad, precisión y funcionalidad para la interpretación de datos genómicos en oncología. Entre las más destacadas se encuentran:

- Illumina TruSight Oncology 500 [Illumina, 2024]. Plataforma de secuenciación de próxima generación desarrollada por Illumina. Es capaz de cubrir más de 500 genes, detecta variantes somáticas (SNV, inserciones/deleciones y fusiones) y analiza la carga mutacional tumoral y la inestabilidad de microsatélites.
- Tempus xT [Tempus, 2024]. Herramienta desarrollada por Tempus que es capaz de analizar 595 genes relacionados con el cáncer. Proporciona informes con información detallada sobre mutaciones, copias, variantes y fusiones. Además, incluye recomendaciones terapéuticas basadas en literatura científica y ensayos actuales e integra datos de expresión génica y otros biomarcadores relevantes para una visión más completa del tumor.

Respecto al avance en *PDFscraping*, actualmente este campo se ha desarrollado en gran medida con el uso de inteligencias artificiales. Nos podemos encontrar tecnologías como:

- Adobe PDF Extract API [Developer, 2023]. Servicio proporcionado por Adobe en la nube. Emplea inteligencia artificial y aprendizaje automático para extraer texto, tablas, imágenes y datos de archivos PDF. Es muy preciso y versátil.
- Textract [Services, 2023a]. Es un servicio de Amazon Web Services que utiliza IA para extraer texto y datos de documentos escaneados, incluidas tablas y formularios.

- Camelot [[Camelot, 2023](#)]. Biblioteca de Python diseñada específicamente para extraer tablas de PDF. Es eficaz para trabajar con tablas bien formateadas.

Metodología

4.1. Descripción de los datos.

En este TFG los datos usados para desarrollarlo han sido informes obtenidos del software Oncomine Reporter y proporcionados por el servicio de anatomía patológica del Hospital Universitario de Burgos (HUBU).

El modelo de estos informes ha ido cambiando a lo largo de estos meses con el fin de intentar establecer un modelo único en todo el país y con él, también han cambiado los datos que aparecen en los PDF y la forma de extraerlos.

Con el fin de evitar la difusión de cualquier tipo de información sensible, estos informes han sido pseudoanonimizados desde el HUBU previamente a su uso para este proyecto.

4.2. Técnicas y herramientas.

En este apartado se explicarán las diversas metodologías que se han empleado a lo largo de la creación de esta aplicación Shiny para *PDFscraping*.

aplicación Shiny

Shiny es un paquete de R que simplifica significativamente el proceso de creación de aplicaciones web interactivas, permitiendo a los usuarios generar interfaces dinámicas para la exploración y visualización de datos sin requerir un conocimiento profundo en programación web.

Una aplicación Shiny consta de tres componentes fundamentales:

- **Interfaz de Usuario (UI):** Esta capa corresponde a la interfaz visible de la aplicación, donde se disponen los elementos gráficos y de entrada de datos (widgets) que permiten la interacción del usuario. La UI define cómo se presentan los datos y qué acciones pueden llevar a cabo los usuarios.
- **Servidor (*Server*):** El servidor es responsable de procesar las acciones del usuario y generar las respuestas correspondientes. Recibe los datos de entrada provenientes de la interfaz de usuario, realiza los cálculos necesarios y actualiza la interfaz de usuario con los resultados obtenidos. Es el componente que coordina la lógica de la aplicación.
- **Llamada a la función shinyApp:** una función llamada `runApp()` con el nombre de la aplicación Shiny es la encargada de ejecutarla.

Los widgets, elementos fundamentales de la interfaz de usuario, son los encargados de capturar la interacción del usuario. Estos elementos, como botones, deslizadores o cuadros de texto, permiten al usuario enviar información que luego es procesada por la aplicación [Shiny, 2020].

PDFscraping

El *scraping* de datos es referido al conjunto de técnicas mediante las cuales un programa informático es capaz de extraer datos del resultado generado por otro programa, bien sea una página web o cualquier documento digital. Se basa en analizar la estructura del documento identificando los datos relevantes para extraerlos de manera sistemática [Cloudflare, 2024].

Si hablamos específicamente de *PDFscraping* nos estamos refiriendo al uso de *scraping* de datos de documentos PDF. El proceso de extracción de datos de PDF puede ser tedioso de forma manual, pero gracias al uso de software y *scripts* especializados, este proceso puede automatizarse para devolvernos los datos en forma de tablas, gráficas, documentos de texto, hojas de cálculo o bases de datos.

En esta nueva versión de *PDFscraping* se ha decidido dejar de lado Python y empezar a trabajar con R, lenguaje de programación con el que el equipo de anatomía patológica del HUBU se encuentra más familiarizado. Además, se ha decidido dejar el código abierto y accesible a cualquiera que use la app. De esta manera, en el caso de que sea necesario por los patólogos modificar cualquier variable del código, van a poder hacerlo de forma fácil y eficaz.

Programación en R

R es un lenguaje de programación estadística y un entorno de software utilizado principalmente para el análisis de datos y la generación de gráficos. Entre sus ventajas se incluyen [Lizana, 2020]:

- Es gratis
- Es libre
- Posee una amplia gama de paquetes
- Flexible y potente
- Ofrece potentes capacidades de visualización de datos
- Gran riqueza en métodos y técnicas estadísticas

Rstudio

RStudio es un entorno de desarrollo integrado (IDE) que te permite trabajar con R. Actúa como un centro de control, proporcionando todas las herramientas necesarias en un espacio de trabajo personalizable, evitando la necesidad de saltar entre diferentes programas [Software, 2024].

Funcionalidades clave de RStudio:

- **Consola y Editor:** Ejecuta código R directamente desde el editor, con resaltado de sintaxis y ayudante de código que te ayuda a programar más rápido y sin errores.
- **Visualización:** Crea gráficos de alta calidad para explorar y comunicar tus datos de manera efectiva.
- **Explorador de espacio de trabajo:** Gestiona tus objetos y variables creados durante el análisis, manteniendo tu entorno organizado.
- **Historial de comandos:** Revisa y vuelve a ejecutar fácilmente comandos anteriores para una mayor productividad.
- **Depuración:** Identifica y corrige errores en tu código con mayor facilidad.

RStudio está disponible en ediciones gratuitas y comerciales. Funciona en los sistemas operativos más comunes (Windows, Mac y Linux), e incluso cuenta con versiones para otros menos comunes como FreeBSD lo que subraya su flexibilidad. Además, RStudio Server permite que varios usuarios accedan al entorno de forma remota a través de un navegador web, ideal para la colaboración en equipo [RStudio, 2024].

Más allá de estas características se encuentra:

1. La capacidad de integración con Git para gestionar el control de versiones de tu código para un desarrollo colaborativo, seguro y eficiente
2. Gran cantidad de paquetes disponibles para R que amplían las funcionalidades del software y te permiten abordar tareas específicas de análisis de datos.
3. R Markdown permite crear documentos que combinan código, texto y resultados, ideales para la generación de informes y presentaciones.

Si bien no vamos a entrar en detalles técnicos de R Markdown, ya que en su lugar se ha empleado LaTeX, es importante destacar el papel fundamental que han jugado diversos paquetes de R en el desarrollo del software. Además, también se va a hacer especial énfasis en Git y GitHub para el control de actualizaciones del proyecto.

Capacidad de integración con Git

Git

Git es un sistema de control de versiones distribuido que te permite mantener un registro detallado de cada modificación que realizas, incluyendo quién la hizo, cuándo y por qué. Esto te permite rastrear el historial del proyecto y volver a versiones anteriores si es necesario. Además, Cada miembro del equipo puede trabajar en su propia copia del proyecto sin necesidad de un servidor central. No obstante, proporciona herramientas para impulsar la comunicación y la colaboración entre los miembros del equipo. Puedes compartir tu código, realizar revisiones y recibir comentarios de tus compañeros.

Facilita la fusión de los cambios realizados por diferentes personas, incluso si han trabajado en las mismas partes del código. Esto evita conflictos y garantiza que todos estén trabajando con la última versión del código. El registro de versiones nos tiende la opción de retroceder fácilmente a

una versión anterior del código y restaurarlo en caso de errores. Esto te permite solucionar problemas y experimentar sin miedo a romper el código [Microsoft, 2023].

Git funciona almacenando “capturas de imagen” de tu código en diferentes puntos del tiempo. Estas “capturas de imagen” se denominan *commits* y contienen información sobre los cambios realizados, quién los realizó y cuándo [Astigarraga and Cruz-Alonso, 2022]. Cada desarrollador tiene su propia copia del repositorio de Git, lo que les permite trabajar de forma independiente sin necesidad de estar conectados a un servidor central. Esto facilita el trabajo sin conexión y la colaboración en equipo [Microsoft, 2023].

Cuando un desarrollador desea compartir sus cambios con el resto del equipo, puede subir sus *commits* a un repositorio remoto. El repositorio remoto puede estar alojado en un servicio como GitHub o GitLab, o en un servidor privado. Una vez que los *commits* se han subido al repositorio remoto, otros desarrolladores pueden descargarlos y fusionarlos con su propia copia del código. Git se encarga de resolver automáticamente cualquier conflicto que pueda surgir durante la fusión [Astigarraga and Cruz-Alonso, 2022].

GitHub

Plataforma de creación de proyectos abiertos. Es ampliamente usado en proyectos de programación gracias a sus funciones colaborativas que permiten trabajar a la vez a todos sus participantes. Entre sus repositorios podemos encontrar herramientas y aplicaciones de libre acceso que cualquiera puede descargar y permitiendo aportar conocimiento a todos los miembros de la comunidad. Si prefieres que tus proyectos no sean visibles para cualquier usuario de la red, puedes crear repositorios privados.

Entre sus cualidades se encuentra el sistema de control de versiones que permite mantener los estados posteriores a la última actualización del proyecto. Gracias a este sistema podemos comparar los cambios realizados en cada versión, cargar otras versiones si ocurre algún problema en la actual o fusionar versiones.

Al ser los proyectos abiertos, permite a otros usuarios comentar, colaborar y contribuir a la mejora del código de tu repositorio. Crear un documento de presentación en el que se explique tu trabajo también es posible en GitHub, lo que facilita la comprensión del mismo para cualquiera que visite tu página [Fernández, 2019].

Gran cantidad de paquetes disponibles

Entre los paquetes con los que se han trabajado se encuentran:

Pdftools

Paquete que nos permite realizar operaciones sobre documentos PDF [Ooms, 2023]. Entre estas funcionalidades se encuentran:

- Extracción de texto de PDF. Esto nos facilita la automatización de la lectura y análisis de documentos.
- Extracción de metadatos. Como, por ejemplo, el número de historia clínica, el sexo del paciente o los genes mutados.
- Conversión a otros formatos. Podemos pasar los datos a texto plano, imágenes o documentos CSV.
- Manipulación de páginas. Nos permite recortar, girar o reorganizar las páginas.
- Extracción de imágenes. Pdftools puede extraer imágenes de los archivos PDF.

Stringr

El paquete Stringr nos permite la manipulación de cadenas de texto [LVaudor, 2017]. Con Stringr, puedes realizar una amplia variedad de operaciones en cadenas de caracteres, como:

- Detectar coincidencias
- Obtener subconjuntos de cadenas
- Gestionar longitudes
- Transformar cadenas
- Juntar y separar
- Ordenar cadenas

Este paquete está diseñado para ser fácil de usar y consistente, con nombres de funciones y argumentos que siguen una convención clara y coherente.

Una de las características destacadas de Stringr es su integración con expresiones regulares, lo que permite realizar operaciones avanzadas en texto con facilidad. Además, Stringr permite manejar de manera consistente los valores nulos y vectores de longitud cero, lo que simplifica el trabajo con datos incompletos o vacíos [Posit Software, 2023].

Tidyverse

Grupo de paquetes que propone un conjunto de estándares para organizar los datos de manera que sean fáciles de manipular, visualizar y modelar.

Algunos de estos paquetes esenciales para el análisis de datos en R son:

- **dplyr**: Proporciona un conjunto de funciones para manipular datos de forma eficiente, permitiendo filtrar, seleccionar, ordenar, agrupar y resumir conjuntos de datos de manera intuitiva [Wickham et al., 2023].
- **tidyr**: Se utiliza para reorganizar, transformar y limpiar conjuntos de datos. Proporciona un conjunto de funciones útiles para manipular la estructura de los datos, como agregar, eliminar, modificar y reordenar filas y columnas [Wickham et al., 2024b].
- **readr**: Ofrece herramientas para importar y exportar datos desde y hacia diferentes formatos de archivos, como archivos delimitados por comas (CSV) y archivos separados por tabulación (TSV) [Wickham et al., 2024a].
- **purrr**: Proporciona funciones para trabajar con funciones y listas de manera más eficiente. Facilita la aplicación de funciones a elementos individuales o a grupos de datos [Wickham and Henry, 2023].
- **ggplot2**: Es un paquete para crear gráficos estadísticos, basado en la gramática de gráficos de Wilkinson (The Grammar of Graphics). Permite crear gráficos complejos de manera sencilla y flexible [Wickham, 2016].

Readxl

Este paquete pertenece a la categoría de importación de datos y simplifica la tarea de extraer información de Excel hacia R. Readxl no tiene

dependencias externas, lo que facilita su instalación y uso en todos los sistemas operativos. Está diseñado específicamente para manejar datos tabulares almacenados en una sola hoja de Excel.

Readxl es compatible tanto con el formato antiguo .xls como con el formato moderno basado en XML .xlsx. Para leer archivos .xls, utiliza la biblioteca C Libxls, que simplifica muchas de las complejidades del formato binario subyacente. Mientras que para analizar archivos .xlsx, se vale de la biblioteca C++ RapidXML.

Este paquete forma parte del conjunto de herramientas conocido como Tidyverse, que en conjunto ofrecen diversas formas de representar y manipular datos. A diferencia de Readxl, que requiere una instalación independiente, los demás paquetes de Tidyverse se descargan automáticamente al instalarlo [Hernández, 2020].

Shiny

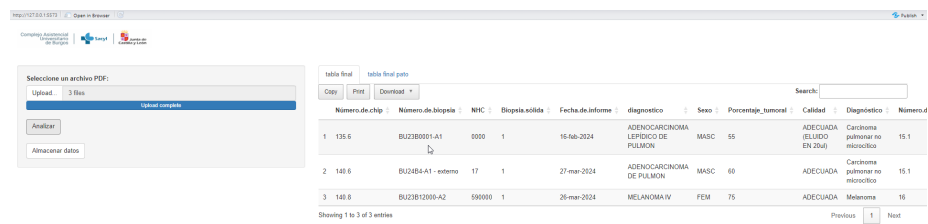
Paquete que nos facilita el desarrollo de aplicaciones web interactivas desde R sin necesidad de tener amplios conocimientos en la programación web.

Shiny se caracteriza por su enfoque en reactividad, es decir, los cambios realizados por el usuario en la interfaz de la aplicación provocan actualizaciones automáticas en los resultados mostrados, lo que proporciona una experiencia interactiva fluida. Esto permite a los usuarios explorar y analizar datos de manera intuitiva, modificando parámetros y visualizando resultados de forma instantánea.

Shiny facilita la creación de aplicaciones web mediante un conjunto de funciones y herramientas que simplifican tareas como la creación de interfaces de usuario, la gestión de eventos de entrada, la generación de gráficos y la integración de datos. Además, ofrece flexibilidad para personalizar el aspecto y la funcionalidad de las aplicaciones mediante HTML, CSS y JavaScript, lo que permite crear aplicaciones web con un diseño atractivo y adaptado a las necesidades específicas del usuario [Gómez et al., 2016].

DT

El paquete DT de R ofrece una interfaz para la visualización de tablas de datos como la de la figura 4.1. Permite representar objetos de datos de R, ya sean matrices o *dataframes*, como tablas en páginas HTML. Esto incluye funcionalidades como filtrado, paginación, clasificación y muchas otras características útiles para trabajar con tablas de datos [Gutiérrez, 2020].



Número de chip	Número de biopsia	NHC	Biopsia acilida	Fecha de informe	diagnostico	Sexo	Porcentaje_tumoral	Calidad	Diagnóstico	Número.d
1	135.5	BU23B0001-A1	0000	16-feb-2024	ADENOCARCINOMA LEPÍDICO DE PULMON	MASC	55	ADECUADA (ELUIDO EN 25u)	Carcinoma pulmonar no microscópico	15.1
2	140.6	BU24B4-A1 - externo	17	27-mar-2024	ADENOCARCINOMA DE PULMON	MASC	60	ADECUADA	Carcinoma pulmonar no microscópico	15.1
3	140.8	BU23B12000-A2	590000	26-mar-2024	MELANOMA/IV	FEM	75	ADECUADA	Melanoma	16

Figura 4.1: Ejemplo tabla

Mongolite

Una versión más reciente del controlador de MongoDB para R, ofrece una gama de operaciones que incluyen indexación, tuberías de agregación, encriptación TLS y autenticación SASL, entre otras. Este controlador se basa en los paquetes Jsonlite y mongo-c-driver para R. La instalación de Mongolite puede realizarse desde CRAN o directamente desde RStudio, como se detalla en una sección posterior [MongoDB, 2024].

Processx

Processx es una herramienta diseñada para facilitar la gestión de procesos del sistema desde el entorno de R, gracias a ella podemos iniciar la base de datos de MongoDBCompass sin necesidad de usar la central de comandos de la computadora. Este paquete permite ejecutar y controlar procesos en segundo plano, proporcionando una interfaz eficiente para trabajar con la salida estándar y el error estándar de estos procesos mediante conexiones no bloqueantes.

Con este paquete se puede verificar si un proceso en segundo plano está activo, esperar a que termine, obtener el estado de salida de procesos finalizados y finalizar procesos cuando sea necesario. Además, ofrece la capacidad de sondear la salida y los errores de un proceso con un tiempo de espera configurable, así como de gestionar varios procesos simultáneamente [Chang et al., 2024].

Rentrez

Este paquete ofrece una interfaz en R para acceder a 50 bases de datos de NCBI. Está bien documentado, cuenta con una amplia gama de pruebas unitarias y una base de usuarios activa. Esta base de datos contiene una gran cantidad de información sobre genética, biología molecular y medicina,

incluidos datos como secuencias genéticas, anotaciones de genes, artículos científicos y mucho más.

La interfaz proporcionada por Rentrez permite a los investigadores realizar consultas en bases de datos y descargar o importar registros específicos en sesiones de R para análisis posteriores.

La naturaleza completa del paquete, su extensa cantidad de pruebas y el hecho de que implementa las políticas de uso de NCBI hacen de Rentrez una herramienta poderosa para los desarrolladores de nuevos paquetes que realizan tareas más específicas [Winter, 2017].

Otras tecnologías aplicadas

Como ya se ha mencionado en el paquete Mongolite, se ha empleado MongoDB como base de datos para almacenar la información extraída de la aplicación. En el paquete Rentrez también se ha hecho hincapié en la biblioteca bioinformática NCBI, en especial se ha usado ClinVar, uno de sus bancos de datos genéticos. ClinVar ha jugado un papel importante para la extracción de datos relevantes a los genes con mutaciones detectadas por el software Oncomine Reporter.

MongoDB

MongoDB es una base de datos NoSQL de código abierto que ofrece alto rendimiento, alta disponibilidad y escalado automático. En lugar de almacenar datos en tablas rígidas como las bases de datos tradicionales, MongoDB utiliza un enfoque basado en documentos, similar a JSON.

Un documento en MongoDB es similar a un objeto en lenguajes de programación. Almacena información como par clave-valor, y puede incluir otros documentos, listas e incluso listas de documentos. Esto permite modelar estructuras de datos complejas de forma natural.

Sus principales ventajas son:

- **Alto rendimiento:** MongoDB está optimizado para la velocidad y la eficiencia, permitiendo consultas rápidas incluso en grandes conjuntos de datos.
- **Rico lenguaje de consulta:** Soporta consultas complejas para lectura, escritura, agregación de datos y búsqueda de texto.

- **Alta disponibilidad:** La replicación automática de datos garantiza la continuidad del servicio en caso de fallos.
- **Escalabilidad horizontal:** MongoDB permite distribuir los datos a través de múltiples servidores para manejar grandes volúmenes de información.

A diferencia de las bases de datos relacionales, MongoDB no impone un esquema rígido a los datos, permitiendo que los documentos tengan estructuras distintas. Además, su facilidad de escalado es notable, ya que a medida que los datos crecen, es posible agregar más servidores al clúster de MongoDB para satisfacer la demanda creciente. Por último, permite realizar consultas profundas y complejas sobre documentos completos sin la necesidad de emplear uniones complejas, lo que simplifica significativamente la gestión y el análisis de datos [Chauhan, 2019].

ClinVar

ClinVar es uno de los bancos de datos genéticos más importantes, reúne información de cerca de más de 700.000 mutaciones. Opera como uno de los recursos del National Center for Biotechnology Information (NCBI) [NIH, 2019].

Las mutaciones datadas en ClinVar son acompañadas de información sobre el remitente, enfermedad del portador (en caso de que se desconozca se incluyen signos y síntomas), significado clínico de la mutación y artículos científicos en los que esta mutación haya sido reportada con anterioridad.

ClinVar permite a cualquier persona interesada en enfermedades genéticas informarse de forma gratuita sobre cualquier mutación compartida. Compartir información en esta base de datos es un proceso complejo, por lo que la colaboración es voluntaria y de libre elección para cada laboratorio [Taniguti, 2022].

API del NCBI

Por otro lado, una interfaz de programación de aplicaciones (API) es un conjunto de protocolos y definiciones que permiten a dos componentes de software comunicarse entre sí. Una API podría considerarse como un contrato de servicios en el que se establece cómo los desarrolladores deben estructurar la comunicación, solicitudes y respuestas [Services, 2023b].

En el caso de la API de ClinVar es proporcionada por el NCBI y permite hacer consultas en su base de datos mediante un software ajeno a ClinVar.

En el caso de este TFG es usada para consultar genes que Oncomine Reporter no es capaz de asegurar su significado clínico y determinar si son patogénicos, benignos o significado incierto. Finalmente, este significado clínico es devuelto al patólogo evitando su búsqueda manual.

Resultados

Todas las instrucciones necesarias para llevar a cabo cualquier operación mencionada en este apartado aparecen descritas con mayor detalle en el apéndice F de los anexos.

5.1. Resumen de resultados.

Con esta aplicación web cualquiera puede seleccionar uno o varios documentos PDF obtenidos a través del software Oncomine Reporter. De estos documentos, son extraídos y mostrados en pantalla, distintos campos relevantes al paciente del que sea el informe, sistematizando y evitando posibles errores humanos. Estos datos son mostrados en tablas exportables a nuestro ordenador en diversos formatos.

Más allá de una visualización, también ofrece conexión con una base de datos MongoDB con la que podemos mantener un historial de todos los pacientes cuyos informes han sido procesados por la aplicación.

Los datos mostrados en tablas son enriquecidos gracias al uso de la API del NCBI con la idea de aportar un punto de vista extra al patólogo a la hora de tomar una decisión en el diagnóstico.

Extracción y visualización

Al pulsar el botón “*Upload...*” se nos abrirá nuestro explorador de archivos para seleccionar los documentos deseados. Una vez seleccionados los documentos, si clicamos en el botón “analizar” la app procederá a procesar estos PDF seleccionados.

El output obtenido en pantalla serán dos tablas, una con los genes patológicos y la otra con el total de genes mutados. La tabla mostrada dependerá de la pestaña seleccionada en la app. La pestaña “Genes mutados” nos mostrará la tabla con todos los genes mutados y la pestaña “Genes patogénicos” contiene la tabla con solo los genes patogénicos.

Entre los datos de las tablas se encuentran:

- El número de chip correspondiente a la librería de ADN del paciente.
- El número de biopsia.
- El número de historial clínico.
- Un número representativo del tipo de biopsia. El 1 indicaría que la muestra se ha extraído por biopsia, el 2 que se ha extraído por punción y el 3 por citología.
- La fecha en la que se obtuvo el informe.
- El tipo de cáncer de la muestra.
- El diagnóstico final del patólogo.
- El sexo del paciente.
- El porcentaje de células tumorales en la muestra.
- La calidad de la muestra y las limitaciones que pueda tener para su análisis.
- El número asociado al diagnóstico.
- El porcentaje de frecuencia alélica.
- El número de ensayos clínicos
- El número de fármacos aprobados y disponibles.
- Columna con dos valores, 1 o 0. En el caso de que el valor sea 1 indicará que hay ensayos, por otro lado, en caso de 0 se indicará que no hay ensayos.
- Columna con dos valores, 1 o 0. En el caso de que el valor sea 1 indicará que hay fármacos, por otro lado, en caso de 0 se indicará que no hay fármacos.

Luego hay variables específicas de cada tabla. En la siguiente tabla 5.1 se hace una comparativa entre los valores específicos para cada una de las tablas de la aplicación 5.2 y 5.1.

Genes mutados	Genes patogénicos
Los genes mutados en la muestra.	Los genes patogénicos.
El recuento de genes mutados.	El número correspondiente a la mutación correspondiente a cada gen.
Las fusiones (en el caso de que existan).	Los cambios sufridos por la mutación.
La afectación que produce la mutación. Si es benigna, patogénica o no se sabe.	
El significado clínico de la mutación tras la consulta en la API de CLinVar	

Tabla 5.1: Comparativa entre las dos tablas mostradas en la app

	Número del diagnóstico	Mutaciones detectadas	Número de la mutación específica	Total del número de mutaciones	Porcentaje de frecuencia alélica, ACN	Fusiones ID	Patogenicidad	Patogenicidad Clinvar	Ensayos clínicos	SLMO ensayo	Fir
#	15.1	KRAS S80D	35.52	2	22.51%,51.20%		Pathogenic/Pathogenic	Likely pathogenic/Conflicting classifications of pathogenicity	16	1	
#	15.1	ALK F359R	4.25	2	32.73%,38.42%		Sin resultados/Sin resultados	Sin resultados/Sin resultados	0	0	
#	15.1	CTNMB1 BRAF	11.7	2	11.90%,27.81%		Pathogenic/Pathogenic	Pathogenic/Likely pathogenic/other/Sin resultados	2	1	
14		CCND1	8	1			Sin resultados	Uncertain significance	0	0	
		KRAS G12V S80D	35.29.52	3	42.34%,53.11%,51.44%		Pathogenic/Sin resultados/Pathogenic	Pathogenic/Sin resultados/Conflicting classifications of pathogenicity	0	0	
#	15.1	EGFR ARAS L858R	13.35.40	3	16.05%,18.05%		Sin resultados/Pathogenic/Sin resultados	Other/Pathogenic/Conflicting classifications of pathogenicity	13	1	
#				0					0	0	
Next											

Figura 5.1: Tabla de genes mutados

Cellul	Diagnóstico	Número del diagnóstico	Genes patogénicos	Número de la mutación específica	X frecuencia alélica	Cambios	Total de mutaciones patogénicas	Ensayos clínicos	SLMO ensayo	Farmacogenético	SLMO farmacogenético
ADECUADA	Cáncer de pulmón no neuroendocrino	15.1	KRAS S80D	35.52	22.51%,51.20%,32.51%,51.20%,22.51%	2	2	16	1	7	1
ADECUADA	Cáncer de pulmón no neuroendocrino	15.1				0	0	0	0	0	0
ADECUADA	Melanoma	16	CTNMB1 BRAF	11.7	11.90%,27.81%	(S49F), (S497F)	2	2	1	0	0
RNA NO VALORABLE	Cáncer de hígado	14				0	0	0	0	0	0
ADECUADA	Cáncer de páncreas		KRAS S80D	35.52	42.34%,51.44%,42.34%,51.44%,42.34%	(G12V), (G12V), (G12V), (G12V), (G12V)	2	0	0	1	1
ADECUADA	Cáncer de pulmón no neuroendocrino	15.1	KRAS	35	16.05%,18.05%,16.05%	(G12S), (G12S), (G12S)	1	13	1	15	1
ADECUADA	Glioblastoma (IDH-wildtype Grade II)					0	0	0	0	0	0
Previous											Next

Figura 5.2: Tabla de genes patogénicos

Más allá de la simple visualización de datos, esta herramienta permite modificar la información mostrada. Esto es de especial relevancia para variables como el significado clínico de mutaciones. En este campo, OncoMine

Reporter no siempre es capaz de determinar la patogenicidad, por lo que el patólogo es obligado a contrastar esta información en bases de datos clínicas como OncoKB, Varsome, Cbioportal o Franklin. Gracias a la flexibilidad de las celdas es posible dejar constancia de esta búsqueda en la tabla y por consiguiente en la base de datos de los pacientes.

Además, la aplicación ofrece la posibilidad de ordenar los datos de forma ascendente o descendente, simplemente seleccionando la columna que se desee organizar.

Encima de cada tabla podemos encontrar tres botones:

1. Botón “*Copy*”: nos permite copiar en el portapapeles los datos de la tabla.
2. Botón “*Print*”: al pulsarlo nos mostrará una previsualización de impresión del documento y la opción de imprimirlo.
3. Botón “*Download*”: es una lista desplegable que nos proporciona dos opciones de extensión. Podremos elegir entre descargar un documento con la tabla en formato CSV o XLSX.

A la derecha de los botones encontramos un motor de búsqueda en el que podremos escribir el valor que se desee encontrar en la tabla. Por ejemplo, si queremos buscar un paciente, introducimos su número de historial clínico y nos aparecerá este único paciente en la tabla. Esta búsqueda afecta de igual manera a las operaciones de los botones, nos imprimirá, copiará o descargará únicamente el o los pacientes que se muestren en pantalla.

Historial de pacientes

Por último, hay un botón más a la izquierda de la web. Este botón está etiquetado como “Almacenar datos”, su función consiste en almacenar los datos extraídos en una base de datos MongoDB. Este proceso está programado para que no se produzcan duplicados, de tal forma que si se mantiene registro de un paciente previamente al intento de guardado, estos datos no serán almacenados de nuevo.

La base de datos de MongoDB es visualizable desde la aplicación MongoDBCompass [5.3](#).

_id	Número_de_biopsia	NNC	Número_de_biopsia_1	Biopsia_adida	Fecha_de_inicio
1	ObjectID('666b057a67f13dc...')	"12880023-A1"	"12880023-A1"	"1"	"14-mar-2023"
2	ObjectID('666b057a67f13dc...')	"12880023-A2"	"12880023-A2"	"1"	"14-mar-2023"
3	ObjectID('666b057a67f13dc...')	"12880023-A3"	"12880023-A3"	"1"	"14-mar-2023"
4	ObjectID('666b057a67f13dc...')	"12880023-A4"	"12880023-A4"	"1"	"14-mar-2023"
5	ObjectID('666b057a67f13dc...')	"12880023-A5"	"12880023-A5"	"1"	"14-mar-2023"
6	ObjectID('666b057a67f13dc...')	"12880023-A6"	"12880023-A6"	"1"	"14-mar-2023"
7	ObjectID('666b057a67f13dc...')	"12880023-A7"	"12880023-A7"	"1"	"14-mar-2023"
8	ObjectID('666b057a67f13dc...')	"12880023-A8"	"12880023-A8"	"1"	"14-mar-2023"
9	ObjectID('666b057a67f13dc...')	"12880023-A9"	"12880023-A9"	"1"	"14-mar-2023"
10	ObjectID('666b057a67f13dc...')	"12880023-A10"	"12880023-A10"	"1"	"14-mar-2023"
11	ObjectID('666b057a67f13dc...')	"12880023-A11"	"12880023-A11"	"1"	"14-mar-2023"
12	ObjectID('666b057a67f13dc...')	"12880023-A12"	"12880023-A12"	"1"	"14-mar-2023"

Figura 5.3: Tabla de datos en MongoDBCompass

MongoDB, como ya hemos mencionado, es una base de datos NoSQL. Esto significa que podemos almacenar datos en formato de listas dentro de las columnas de nuestro *dataFrame*. En MongoDB Compass, estos datos se presentan como tablas anidadas 5.5 dentro de nuestra tabla principal. Estas tablas secundarias se pueden acceder desde cualquier celda de las columnas que contengan *arrays* 5.4. Además, están vinculadas a la tabla principal mediante un ID específico y único para cada fila, el cual es generado automáticamente por MongoDB cuando se ingresa un nuevo dato.

Número_de diagnóstico	Mutaciones detectadas	Número de la mutación esp...
1 15.1"	[] 1 elements	[] 1 elements
2 15.1"	[] 1 elements	[] 1 elements
3 15.1"	[] 1 elements	[] 1 elements
4 15.1"	[] 2 elements	[] 2 elements
5 15.1"	[] 2 elements	[] 2 elements
6 15.1"	[] 2 elements	[] 2 elements
7 15.1"	[] 2 elements	[] 2 elements
8 15.1"	[] 2 elements	[] 2 elements
9 34"	[] 1 elements	[] 1 elements
10 15.1"	[] 2 elements	[] 2 elements
11 15.1"	[] 1 elements	[] 1 elements
12 16"	[] 2 elements	[] 2 elements

Figura 5.4: Acceso a tabla anidada

TFG > test > TFG

Documents 12 Aggregations Schema Indexes 1 Validation

ADD DATA EXPORT DATA 1 - 12 of 12

	_id	Mutaciones_detectadas	ObjectID
1	ObjectID('660eb5d7e87f15...')	null	No field
2	ObjectID('660eb5d7e87f15...')	"KRAS"	No field
3	ObjectID('660eb5d7e87f15...')	"MYC"	No field
4	ObjectID('660eb5d7e87f15...')	"EGFR"	"PIK3CA"
5	ObjectID('660eb5d7e87f15...')	"EGFR"	"FGFR1"
6	ObjectID('660eb686e87f15...')	"FGFR1"	"PIK3CA"
7	ObjectID('660eb686e87f15...')	"CDK6"	"MET"
8	ObjectID('660eb686e87f15...')	"CTNNB1"	"EGFR"
9	ObjectID('660eb686e87f15...')	"MYC"	No field
10	ObjectID('6643c748f9d2e4...')	"KRAS"	"SMO"
11	ObjectID('6643c748f9d2e4...')	"ALK"	No field
12	ObjectID('6643c748f9d2e4...')	"CTNNB1"	"BRAF"

Figura 5.5: Tabla anidada

Dentro de la base de datos también podemos hacer búsqueda de datos como en la imagen 5.6. Las búsquedas básicas deben hacerse siguiendo el patrón de {nombre_de_la_columna: "dato a buscar"}. Estas consultas pueden ser tan complejas y concretas como deseemos mediante el uso de variables como, las columnas, el proyecto, el orden de aparición, o incluso empleando comandos.

TFG > test > TFG

Documents 3 Aggregations Schema Indexes 1 Validation

{ "Número_de_chip": "135,6" } Generate query Explain Reset Find Options

ADD DATA EXPORT DATA UPDATE DELETE 1 - 1 of 1

	_id	Número_de_chip	Número_de_biopsia	MMC	Biopsia_válida	Fecha_de_informe
1	ObjectID('66477e93af3d835...')	"135,6"	"B02388001-A1"	"8008"	"1"	"16-feb-2024"

Figura 5.6: Ejemplo de búsqueda

Al igual que en la aplicación Shiny que podíamos descargar documentos con los datos analizados, en MongoDBCompass también podemos descargar los datos de la base de datos en formato CSV o JSON. Además, podemos hacer otras operaciones como eliminar datos concretos o subir documentos CSV o JSON para almacenar la información contenida en ellos.

API del NCBI

Gracias a la integración con la API del NCBI, la herramienta también puede buscar en ClinVar la mutación y su correspondiente cambio en la secuencia de ADN para determinar si es patológica o benigna, siempre y cuando esta información esté disponible en la base de datos de ClinVar. En caso contrario se mostrará como una cadena de texto vacía.

El significado clínico de los genes mutados consultados en la biblioteca de ClinVar será añadido a la tabla “Genes mutados” como una columna extra llamada “Patogenicidad Buscada”. Con esta columna se pretende enriquecer la información aportada por los informes para facilitar al patólogo la decisión de un diagnóstico.

5.2. Discusión.

Gracias a este programa, la extracción de datos de informes clínicos producidos por el software OncoPrint Report genera de ser una tediosa labor manual a un proceso rápido y automatizado. Todos los campos relevantes para el Hospital Universitario de Burgos (HUBU) son obtenidos con una alta precisión. Además, se proporciona a los profesionales de la salud un programa entendible y modificable gracias al código abierto y a la migración del mismo de Python, lenguaje de la versión anterior, a R.

El hecho de enriquecer los resultados con búsquedas en bases de datos clínicas y el registro de pacientes manteniendo un historial de los datos obtenidos son otros dos aspectos relevantes que destacan dentro de la aplicación y que facilitan la labor de los patólogos.

En resumen, la aplicación ha resultado muy intuitiva y fácil de usar, toda información es fácilmente exportable en el formato deseado y se ha logrado una mayor flexibilidad al seleccionar documentos de nuestro ordenador, permitiendo analizar informes de cualquier directorio. Los datos obtenidos de ClinVar complementan la información de los PDF y la base de datos permite tener un historial actualizado de pacientes que favorezca un posterior análisis de los llamados datos en vida real.

En un futuro, si el modelo de los informes comienza a ser actualizado nuevamente, lo óptimo quizá sería plantear un sistema de aprendizaje automático NLP (Procesamiento de Lenguaje Natural) que se entrenara con todos los modelos de PDF probados a lo largo de los años para fortalecer la aplicación. Actualmente, esto es inviable debido a la escasa cantidad

de informes disponibles, ya que se generan semanalmente una media de 8 pacientes por semana y el software no ha sido utilizado por más de 4 años.

Sin embargo, si fuera de vital importancia aumentar la robustez ante los cambios en el modelo, se podría investigar la aplicación de BERT [Meijomil, 2022], un modelo preentrenado para manejar secuencias de datos. BERT es capaz de comprender el contexto de una palabra basándose en las palabras que la rodean utilizando un enfoque bidireccional.

Conclusiones

En este apartado se van a detallar puntos significativos en el desarrollo del proyecto.

6.1. Aspectos relevantes.

Se han llevado a cabo varias actualizaciones sobre el proyecto anterior. Entre las más destacables se encuentran:

- **Código abierto.** El programa desarrollado cuenta con un código abierto y accesible por el usuario, permitiéndole hacer modificaciones en el mismo si así lo desea. Anteriormente, este código no era visible por el usuario, simplemente era un elemento ejecutable que se encargaba de hacer todo el proceso sin mediación del usuario.
- **Interfaz intuitiva.** Gracias a la biblioteca Shiny en esta versión se cuenta con una interfaz sencilla e intuitiva que proporciona a usuario una mayor interacción con el software.
- **Flexibilidad de directorios.** En la versión anterior, los documentos PDF debían encontrarse en una ruta específica y dada por el desarrollador del código. Sin embargo, ahora es posible seleccionar cualquier informe clínico de nuestra computadora sin importar el directorio en el que se encuentre. Lo mismo ocurre con las descargas de documentos procedentes de la aplicación.
- **Versatilidad de formatos.** En esta versión podemos exportar nuestros datos en una amplia variedad de formatos. Mediante Shiny podemos seleccionar formatos CSV, Excel o incluso PDF y TXT. Si lo

deseáramos, a través de MongoDB, podemos además seleccionar entre CSV y JSON.

- **Personalización de los datos.** Una vez obtenidas las tablas, el patólogo dispone de la opción de manipular los datos de las mismas a su gusto antes de exportarlos o registrarlos en la base de datos. El anterior software extraía y exportaba los datos directamente sin poder mediar en el proceso.
- **Registro de pacientes.** Gracias a la incorporación de bases de datos de MongoDB se proporciona la opción de mantener un historial con todos los pacientes y sus respectivos datos asociados procedentes de los informes clínicos.
- **Uso de API clínicas.** Todos los informes son enriquecidos con información contrastada en la base de datos de ClinVar proporcionando un soporte extra al patólogo.

Además de estas actualizaciones, también cabe destacar que durante este proyecto también se trabajó en otra aplicación Shiny similar que fue presentada en conjunto con un compañero, Rodrigo Pascual García, a la Hackathon Automoción organizada por la Universidad de Burgos en la que obtuvimos el primer puesto.

Lineas de trabajo futuras

Aunque se han alcanzado todos los objetivos propuestos inicialmente, incluyendo algunas mejoras adicionales, todavía quedan aspectos por perfeccionar e ideas por incorporar. Estos puntos abren la posibilidad de desarrollar una versión 3.0 del proyecto original.

Consulta en más bases de datos genómicas

Como se mencionó anteriormente, el software que genera los documentos PDF tiene limitaciones para determinar la afectación de ciertas mutaciones. Estas mutaciones deben ser buscadas posteriormente por los patólogos responsables en diferentes bases de datos genómicas para completar la información faltante. Entre las más utilizadas se encuentran OncoKB, VarSome, ClinVar y Cbioportal. De estas, ClinVar es la única que se ha podido incorporar hasta el momento.

Gestión de usuarios

Tanto la aplicación Shiny como la base de datos MongoDB son accesibles para cualquier persona con acceso al ordenador donde están instaladas o a cualquier otro ordenador que comparta la conexión a la base de datos. Para evitar problemas y proteger la información en caso de accesos no autorizados, sería recomendable desarrollar un sistema de gestión de usuarios.

Mostrar datos estadísticos

Sería una muy buena idea añadir una utilidad a la web que nos muestre datos estadísticos relevantes para los patólogos. Por ejemplo, que mutaciones

son más frecuentes en un tipo de cáncer en función del sexo o una gráfica que muestre la frecuencia de aparición de mutaciones en cada gen.

Bibliografía

- [Abdulkader et al., 2021] Abdulkader, I., Arriola, E., Bellosillo, B., Biscuola, M., Palanca, S., and Rojo, F. (2021). Abordaje del diagnóstico molecular del cáncer de pulmón no microcítico mediante ngs: Opinión de expertos. Acceso realizado el 9 de junio de 2024.
- [Alonso et al., 2023] Alonso, L. F., Cuevas, L. A., García, R. R., Barrera, J. M., del Río Ignacio, J. J., Espejo, A. S., Florindo, I. B., Fernández, I. C., and Rojo, M. G. (2023). Estudio de variantes genéticas en 169 pacientes de cáncer de pulmón no microcítico. *Revista Española de Patología*, 56:233–242. Acceso realizado el 6 de junio de 2024.
- [Astigarraga and Cruz-Alonso, 2022] Astigarraga, J. and Cruz-Alonso, V. (2022). ¿se puede entender cómo funcionan git y github! *Ecosistemas*, 31. Acceso realizado el 22 de mayo de 2024.
- [Biosystems, 2024] Biosystems, L. (2024). Leica st5020-cv5030. <https://www.leicabiosystems.com/es-es/equipo-histologia/tincion-rutinaria-y-aplicacion-de-cubreobjetos/leica-st5020-cv5030/>. Acceso realizado el 6 de junio de 2024.
- [Camelot, 2023] Camelot (2023). camelot-dev/camelot: A python library to extract tabular data from pdfs. <https://github.com/camelot-dev/camelot>. Acceso realizado el 6 de junio de 2024.
- [Chang et al., 2024] Chang, W., Software, P., PBC, and Csárdi, A. D. S. G. (2024). Cran - package processx. <https://cran.r-project.org/web/packages/processx/index.html>. Acceso realizado el 5 de junio de 2024.
- [Chauhan, 2019] Chauhan, A. (2019). A review on various aspects of mongodb databases. Acceso realizado el 22 de mayo de 2024.

- [Cloudflare, 2024] Cloudflare (2024). ¿qué es el scraping de datos? | cloudflare. <https://www.cloudflare.com/es-es/learning/bots/what-is-data-scraping/>. Acceso realizado el 22 de mayo de 2024.
- [Developer, 2023] Developer, A. (2023). Pdf extract api | adobe pdf services. <https://developer.adobe.com/document-services/docs/overview/pdf-extract-api/>. Acceso realizado el 6 de junio de 2024.
- [FarmaIndustria, 2024] FarmaIndustria (2024). Real world evidence, el uso de datos que cambiará la vida de los pacientes. <https://www.farmaindustria.es/web/reportaje/evidencia-en-vida-real-rwe-la-nueva-era-de-datos-sanitarios-que-revolucionara-la-vida-de-los-pacientes/>. Acceso realizado el 10 de junio de 2024.
- [Fernández, 2019] Fernández, Y. (2019). Qué es github y qué es lo que le ofrece a los desarrolladores. <https://www.xataka.com/basics/que-github-que-que-le-ofrece-a-desarrolladores>. Acceso realizado el 22 de mayo de 2024.
- [Goodwin et al., 2016] Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. Acceso realizado el 9 de junio de 2024.
- [Gutiérrez, 2020] Gutiérrez, I. P. (2020). Documento 7 dt package - analisis | laboratorio de computación - bookdown. https://bookdown.org/ignacio_p_g/ipg-labcomp/dt-package-analisis.html. Acceso realizado el 20 de mayo de 2024.
- [Gómez et al., 2016] Gómez, D., Mulero, J., Nueda, M., and Pascual, A. (2016). Aplicaciones diseñadas con shiny: un recurso docente para la enseñanza de la estadística. Acceso realizado el 22 de mayo de 2024.
- [Hernández, 2020] Hernández, S. K. P. (2020). Rpubs - trabajo final. https://rpubs.com/Sonia_Ponce/625541. Acceso realizado el 22 de mayo de 2024.
- [Illumina, 2024] Illumina (2024). Trusight oncology 500 | enable cgp utilizing dna and rna from ffpe. <https://www.illumina.com/products/by-type/clinical-research-products/trusight-oncology-500.html>. Acceso realizado el 6 de junio de 2024.
- [Lizana, 2020] Lizana, M. I. F. (2020). Ventajas de r como herramienta para el análisis y visualización de datos en ciencias sociales. *Revista Científica de la UCSA*, 7:97–111. Acceso realizado el 21 de mayo de 2024.

- [LVaudor, 2017] LVaudor (2017). strings_es. Acceso realizado el 20 de mayo de 2024.
- [Meijomil, 2022] Meijomil, S. (2022). Google bert: qué es, cómo funciona y cómo te afecta - inboundcycle. <https://www.inboundcycle.com/blog-de-inbound-marketing/google-bert-que-es-como-funciona>. Acceso realizado el 6 de junio de 2024.
- [Microsoft, 2023] Microsoft (2023). ¿qué es git? - azure devops | microsoft learn. <https://learn.microsoft.com/es-es/devops/develop/git/what-is-git>. Acceso realizado el 22 de mayo de 2024.
- [MongoDB, 2024] MongoDB (2024). Data analysis with mongodb and r | mongodb. <https://www.mongodb.com/resources/languages/mongodb-and-r-example>. Acceso realizado el 20 de mayo de 2024.
- [NIH, 2019] NIH (2019). Hoja informativa de pruebas genéticas. <https://www.cancer.gov/espanol/cancer/causas-prevencion/genetica/hoja-informativa-pruebas-geneticas>. Acceso realizado el 20 de mayo de 2024.
- [of Veterans Affairs, 2024] of Veterans Affairs, U. (2024). Thermo fisher oncomine reporter (okr). <https://www.oit.va.gov/Services/TRM/ToolPage.aspx?tid=16608#>. Acceso realizado el 20 de mayo de 2024.
- [Ooms, 2023] Ooms, J. (2023). *pdftools: Text Extraction, Rendering and Converting of PDF Documents*. R package version 3.4.0.
- [Posit Software, 2023] Posit Software, PBC, H. W. (2023). Cran - package stringr. <https://cran.r-project.org/web/packages/stringr/index.html>. Acceso realizado el 20 de mayo de 2024.
- [Roche, 2024] Roche (2024). Ventana® he 600. <https://diagnostics.roche.com/es/es/products/instruments/ventana-he-600-ins-4090.html>. Acceso realizado el 6 de junio de 2024.
- [RStudio, 2024] RStudio (2024). rstudio/rstudio: Rstudio is an integrated development environment (ide) for r. <https://github.com/rstudio/rstudio/>. Acceso realizado el 21 de mayo de 2024.
- [Services, 2023a] Services, A. W. (2023a). Extracción inteligente de texto y datos con ocr - amazon textract - amazon web services. <https://aws.amazon.com/es/textract/>. Acceso realizado el 20 de mayo de 2024.

- [Services, 2023b] Services, A. W. (2023b). ¿qué es una api? - explicación de interfaz de programación de aplicaciones - aws. <https://aws.amazon.com/es/what-is/api/>. Acceso realizado el 6 de junio de 2024.
- [Shiny, 2020] Shiny (2020). Shiny. <https://shiny.posit.co/>. Acceso realizado el 19 de mayo de 2024.
- [Software, 2024] Software, P. (2024). Download rstudio | the popular open-source ide from posit. <https://posit.co/products/open-source/rstudio/>. Acceso realizado el 21 de mayo de 2024.
- [Taniguti, 2022] Taniguti, N. (2022). Bancos de datos genéticos – clinvar - blog mendelics. <https://blog.mendelics.com.br/es/bancos-de-datos-geneticos-clinvar/>. Acceso realizado el 20 de mayo de 2024.
- [Tempus, 2024] Tempus (2024). Tempus xt & xr - tempus. <https://www.tempus.com/oncology/genomic-profiling/xt-xr/>. Acceso realizado el 6 de junio de 2024.
- [ThermoFisher, 2022] ThermoFisher (2022). Oncomine™ reporter. <https://www.thermofisher.com/order/catalog/product/es/es/A34298>. Acceso realizado el 20 de mayo de 2024.
- [ThermoFisher, 2024a] ThermoFisher (2024a). Cuantificación fluorimétrica qubit | thermo fisher scientific - es. <https://www.thermofisher.com/es/es/home/industrial/spectroscopy-elemental-isotope-analysis/molecular-spectroscopy/fluorometers/qubit.html>. Acceso realizado el 6 de junio de 2024.
- [ThermoFisher, 2024b] ThermoFisher (2024b). Ion chef™ instrument. <https://www.thermofisher.com/order/catalog/product/4484177>. Acceso realizado el 6 de junio de 2024.
- [ThermoFisher, 2024c] ThermoFisher (2024c). Ion reporter software | thermo fisher scientific - es. <https://www.thermofisher.com/es/es/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-data-analysis-workflow/ion-reporter-software.html>. Acceso realizado el 6 de junio de 2024.
- [ThermoFisher, 2024d] ThermoFisher (2024d). Torrent suite software | thermo fisher scientific - es. <https://www.thermofisher.com/es/es/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent->

[next-generation-sequencing-data-analysis-workflow/ion-torrent-suite-software.html](#). Acceso realizado el 6 de junio de 2024.

- [Wickham, 2016] Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Acceso realizado el 20 del mayo de 2024.
- [Wickham et al., 2023] Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4, <https://github.com/tidyverse/dplyr>.
- [Wickham and Henry, 2023] Wickham, H. and Henry, L. (2023). *purrr: Functional Programming Tools*. R package version 1.0.2, <https://github.com/tidyverse/purrr>.
- [Wickham et al., 2024a] Wickham, H., Hester, J., and Bryan, J. (2024a). *readr: Read Rectangular Text Data*. R package version 2.1.5, <https://github.com/tidyverse/readr>.
- [Wickham et al., 2024b] Wickham, H., Vaughan, D., and Girlich, M. (2024b). *tidyr: Tidy Messy Data*. R package version 1.3.1, <https://github.com/tidyverse/tidyr>.
- [Winter, 2017] Winter, D. (2017). *rentrez: An r package for the ncbi eutils api*. https://www.researchgate.net/publication/323081538_rentrez_An_R_package_for_the_NCBI_eUtils_API. Acceso realizado el 20 de mayo de 2024.