



UNIVERSIDAD TÉCNICA  
FEDERICO SANTA MARÍA

DEPARTAMENTO  
DE INFORMÁTICA

# Desarrollo de una interfaz de usuario para la identificación de contextos genómicos en genomas bacterianos

Vicente José Saona Urmeneta

Memoria para optar al título de Ingeniero Civil en Informática

**Profesor Guía: Carlos Buil Aranda**

**Profesor Correferente: Roberto Durán Vargas**

**Noviembre - 2021**

## **AGRADECIMIENTOS**

A Roberto por motivarse a hacer esto, por motivarme a hacer esto y por su tremenda disposición y valiosos aportes al proyecto, esenciales para lograr llevarlo a cabo.

A Andrea, Rodrigo y Javier, por sus aportes en la concepción de este proyecto.

Al profesor Piérola, por poner a disposición su estudio de usuarios de software de visualización bioinformática, muy útil como punto de partida para la investigación aquí presentada.

Al profesor Buil, por la tremenda eficiencia de sus reuniones y sus aportes certeros.

A todos los usuarios que me ayudaron a evaluar la aplicación, cuyos comentarios han sido muy valiosos para el perfeccionamiento de esta herramienta.

A Nicolás Acevedo, genio del diseño de interfaces. Sin él, la presentada en este trabajo no tendría el aspecto moderno, profesional y pulido que tiene hoy.

A todos los amigos y familiares que me apoyaron en la realización de este trabajo.

## Resumen

Grandes han sido los avances en la secuenciación de genomas bacterianos en los últimos años, lo cual ha facilitado una inmensa cantidad de datos para los análisis en genómica comparativa, filogenómica y caracterización de nuevas cepas bacterianas, mas las herramientas de visualización de secuencias genómicas no se han desarrollado con la misma rapidez, encontrando en la actualidad herramientas poco amigables debido a su falta de versatilidad y facilidad de uso. En este trabajo se presenta el proceso de desarrollo de una plataforma basada en tecnología web utilizando Node.js, para apoyar el análisis comparativo de contexto genómico por medio de la generación interactiva de gráficos de visualización. La plataforma web utiliza como archivos de entrada secuencias de aminoácidos en formato FASTA o archivos de anotación de genomas en formato Genbank (GBFF). La comparación con otros genomas se realiza utilizando datos genómicos de la base de datos pública NCBI (RefSeq) utilizando la herramienta BLAST y TaxonKit para obtener los grupos taxonómicos de filtrado. La aplicación pasó por una evaluación experta de usabilidad y una medición desde los usuarios, siendo validada como herramienta eficiente para el trabajo y fácil de usar, además de identificar oportunidades de mejora. La aplicación se encuentra disponible para su uso y su código está en un repositorio público.

**Palabras clave:** Contexto genómico, visualización de datos, experiencia de usuario, interfaz usuaria, bioinformática.

## Abstract

Great advances have been made in the sequencing of bacterial genomes in recent years, resulting in an enormous growth of databases used in comparative genomics, phylogenomics and new bacterial strain characterization, but sequence visualization tools have not kept up with this advancement, as they lack usability and customizability. Here we present the development process of a web-based platform on Node.js to support genomic context analysis through the interactive generation of visualization graphics. The platform needs as input aminoacid sequence files in FASTA format or annotated genomes in GenBank file format (GBFF). Inter-genomic comparison uses genomic data from NCBI public database, using BLAST for gene lookup and TaxonKit for taxonomic groups filtering. The application has gone through an usability expert evaluation and an user-centered evaluation, being validated in both its pragmatic and hedonic quality, and identifying potential improvement. It is available to use, and its code is in a public repository.

**Keywords:** Genomic context, data visualization, user experience, user interface, bioinformatics.

## Glosario

ADN / DNA: Ácido Desoxirribonucleico

ARN / RNA: Ácido Ribonucleico

BLAST: *Basic Local Alignment Search Tool*

CSS: *Cascading Style Sheets*

DDBJ: *DNA Data Bank of Japan*

EggNOG: *Evolutionary genealogy of genes: Non-supervised Orthologous Groups*

EMBL: *European Molecular Biology Laboratory*

ENA: *European Nucleotide Archive*

HTML: *HyperText Markup Language*

iTOL: *interactive Tree Of Life*

LabMMBA: Laboratorio de Microbiología Molecular y Biotecnología Ambiental

LTS: *Long Term Support*

mRNA: Ácido Ribonucleico mensajero

NCBI: *National Center for Biotechnology Information*

NLM: *National Library of Medicine*

PFGE: *Pulsed-field Gel Electrophoresis*

PGAP: *Prokaryotic Genome Annotation Pipeline*

RAST: *Rapid Annotation using Subsystem Technology*

RefSeq: *Reference Sequence*

RGB: *Red, Green and Blue*

rRNA: Ácido Ribonucleico ribosomal

SVG: *Scalable Vector Graphics*

tRNA: Ácido Ribonucleico de transferencia

WGS: *Whole Genome Shotgun Sequencing*

## Índice de contenidos

|                                                                                |           |
|--------------------------------------------------------------------------------|-----------|
| <b>Resumen</b>                                                                 | <b>3</b>  |
| <b>Abstract</b>                                                                | <b>3</b>  |
| <b>Glosario</b>                                                                | <b>4</b>  |
| <b>Índice de contenidos</b>                                                    | <b>6</b>  |
| <b>Índice de figuras</b>                                                       | <b>8</b>  |
| <b>Índice de tablas</b>                                                        | <b>12</b> |
| <b>Introducción</b>                                                            | <b>13</b> |
| <b>1. Definición del problema</b>                                              | <b>15</b> |
| 1.1. Objetivo General                                                          | 15        |
| 1.2. Objetivos específicos                                                     | 16        |
| <b>2. Marco conceptual</b>                                                     | <b>17</b> |
| 2.1 Herramientas bioinformáticas complementarias                               | 22        |
| 2.2. Principales herramientas de visualización de contexto genómico existentes | 25        |
| 2.2.1. MGcV                                                                    | 25        |
| 2.2.2. DNAFeaturesViewer                                                       | 26        |
| 2.2.3. GeneSpy                                                                 | 27        |
| 2.2.4. Integrative Genomics Viewer                                             | 29        |
| 2.2.5. FlaGs                                                                   | 29        |
| 2.2.6 Gene Graphics                                                            | 29        |
| 2.2.7. Otras herramientas                                                      | 30        |
| 2.2.8. Apreciaciones generales                                                 | 30        |
| <b>3. Proceso de diseño</b>                                                    | <b>32</b> |

|                                                      |           |
|------------------------------------------------------|-----------|
| 3.1 Definición del problema y alcance de la solución | 32        |
| 3.2 Caracterización preliminar del sistema           | 34        |
| 3.3 Prototipado preliminar                           | 34        |
| 3.4 Desarrollo iterativo                             | 36        |
| 3.5 Pruebas de usuario                               | 40        |
| 3.5.1 Tareas de usuario                              | 41        |
| 3.5.2 Observaciones realizadas                       | 43        |
| 3.5.3 Comentarios recibidos                          | 43        |
| 3.5.4 Mejoras implementadas                          | 44        |
| <b>4. Resultados</b>                                 | <b>46</b> |
| 4.1 Tecnologías utilizadas                           | 46        |
| 4.2 Descripción exhaustiva de la interfaz            | 47        |
| 4.2.1 Tutorial                                       | 47        |
| 4.2.2 Ingreso de datos                               | 49        |
| Búsqueda de homólogos                                | 50        |
| Selección directa de contextos                       | 54        |
| 4.2.3 Visualización y edición                        | 56        |
| 4.3 Flujo de trabajo                                 | 64        |
| 4.3.1 Búsqueda de homólogos                          | 64        |
| 4.3.2 Contextos especificados de forma directa       | 65        |
| 4.3.3 Generación del gráfico                         | 65        |
| <b>5. Validación de la solución</b>                  | <b>69</b> |
| 5.1. Validación experta                              | 69        |
| 5.2. Validación con usuarios                         | 70        |
| <b>Conclusiones</b>                                  | <b>76</b> |

|                                                    |           |
|----------------------------------------------------|-----------|
| Trabajo futuro                                     | 77        |
| Eficiencia y confiabilidad                         | 77        |
| Escalabilidad                                      | 77        |
| Usabilidad                                         | 77        |
| Funcionalidad                                      | 78        |
| <b>Referencias bibliográficas</b>                  | <b>79</b> |
| Otra bibliografía relacionada                      | 84        |
| <b>Anexo: Resultados de la encuesta a usuarios</b> | <b>85</b> |

## Índice de figuras

|                                                                        |    |
|------------------------------------------------------------------------|----|
| 1. Representación de un fragmento de genoma                            | 17 |
| 2. Ejemplo de análisis de sintenia en bacterias                        | 18 |
| 3. Secuenciación mediante Sanger                                       | 20 |
| 4. Proceso de secuenciación de genoma completo WGS                     | 21 |
| 5. Cantidad histórica de registros de secuencia almacenadas en GenBank | 22 |
| 6. Interfaz de MGcV                                                    | 26 |
| 7. Diagrama generado usando DNAFeaturesViewer                          | 27 |
| 8. Flujo de carga de datos en GeneSpy                                  | 28 |
| 9. Ejemplo de interfaz de GeneSpy                                      | 28 |
| 10. Edición de imagen de Gene Graphics                                 | 30 |
| 11. <i>Abstraction laddering</i>                                       | 32 |
| 12. <i>Storyboard</i> mostrando contexto de uso de la aplicación       | 33 |
| 13. Prototipo de papel                                                 | 35 |
| 14. Primer prototipo funcional                                         | 36 |
| 15. Primera parte del tutorial                                         | 37 |
| 16. Segunda parte del tutorial                                         | 37 |
| 17. Tercera parte del tutorial                                         | 38 |
| 18. Interfaz de definición de la búsqueda BLAST                        | 38 |
| 19. Interfaz de selección de contextos a graficar                      | 39 |
| 20. Interfaz de visualización de datos y edición del diagrama          | 40 |

|                                                                                               |    |
|-----------------------------------------------------------------------------------------------|----|
| 21. Ejemplo de archivo FASTA utilizado en pruebas de usuario                                  | 42 |
| 22. Vista principal definitiva                                                                | 45 |
| 23. Campo numérico para edición del gráfico                                                   | 45 |
| 24. Cuadro de diálogo mostrando el tutorial                                                   | 48 |
| 25. Vista principal definitiva                                                                | 49 |
| 26. Selección del origen de la secuencia de origen en la búsqueda de homólogos                | 50 |
| 27. Menú de selección del origen de la secuencia de origen en la búsqueda de homólogos        | 51 |
| 28. Especificación de secuencia <i>query</i> desde archivo de genoma anotado                  | 51 |
| 29. Especificación directa de secuencia <i>query</i>                                          | 52 |
| 30. Especificación de secuencia <i>query</i> a través de número de acceso                     | 52 |
| 31. Archivo siendo arrastrado sobre la sección de especificación de la secuencia <i>query</i> | 53 |
| 32. Sección de personalización de la búsqueda de genes homólogos                              | 53 |
| 33. Presentación inicial de la sección de especificación directa de contextos a graficar      | 55 |
| 34. Interfaz de selección de contextos al arrastrar un archivo sobre ella                     | 56 |
| 35. Interfaz inicial de visualización y edición del diagrama                                  | 57 |
| 36. Cambio de la apariencia del cursor al ser posado sobre un texto arrastrable               | 58 |
| 37. Tipos de selección en el diagrama generado                                                | 59 |
| 38. Indicador de reacción al seleccionar un elemento                                          | 59 |
| 39. Interfaz de selección de colores                                                          | 60 |
| 40. Interfaz de selección del grosor del trazo de las flechas                                 | 61 |
| 41. Alternativas de forma para representar gráficamente cada gen                              | 61 |
| 42. Comportamiento al esconder un gen                                                         | 62 |

|                                                                                   |    |
|-----------------------------------------------------------------------------------|----|
| 43. Ejemplo de alineación de genes de interés                                     | 62 |
| 44. Personalización de etiquetas de gen                                           | 63 |
| 45. Tabla de información adicional de gen                                         | 63 |
| 46. Personalización de línea central y etiqueta de contexto genómico              | 64 |
| 47. Tabla de información adicional de contexto genómico                           | 64 |
| 48. Alineamiento automático de contextos genómicos con respecto a genes homólogos | 66 |
| 49. Alineamiento de contextos genómicos con respecto a su gen de inicio           | 67 |
| 50. Espaciado entre los contextos graficados                                      | 67 |
| 51. Proporciones de las representaciones de los genes                             | 68 |
| 52. Resultados de evaluación según lineamientos de usabilidad web                 | 69 |
| 53. Resultados generales AttrakDiff                                               | 71 |
| 54. Resultados por dimensión AttrakDiff                                           | 72 |
| 55. Resultados por cada punto AttrakDiff                                          | 74 |

## Índice de tablas

|                                                                            |    |
|----------------------------------------------------------------------------|----|
| 1. Datos utilizados para las tarea de usuario de graficar un solo contexto | 41 |
| 2. Datos utilizados para las tarea de usuario de graficar varios contextos | 42 |

## Introducción

En la última década, los costos y accesibilidad asociados a la secuenciación, ensamblaje y anotación de genomas bacterianos han disminuido considerablemente llevando a un aumento explosivo en la cantidad de genomas (983.291 genomas bacterianos en agosto 2021, GenBank-NCBI), además de convertirse en un paso clave en la caracterización de nuevas bacterias con potencial biotecnológico. La extracción de información desde estos genomas, también conocido como "*genome mining*", ha sido una metodología muy útil para revelar determinantes intrínsecos relacionados al catabolismo (degradación de xenobióticos y contaminantes), anabolismo (producción de antibióticos, osmoprotectores y bioplásticos), resistencia y tolerancia a estresores ambientales (pH, salinidad, temperatura, metales pesados) y mecanismos moleculares asociados a procesos subcelulares (transcripción, traducción, quimiotaxis) de bacterias aisladas desde ambientes naturales.

La cantidad de información obtenida al secuenciar genomas bacterianos (como también en el caso de metagenomas) puede ser abrumadora y difícil de trabajar manualmente, por lo que la utilización de herramientas bioinformáticas se ha vuelto indispensable. Los genomas se obtienen como millones de lecturas crudas (secuencias cortas o largas dependiendo del tipo de secuenciación), que deben ser ensambladas como un verdadero puzzle con tal de obtener el panorama completo utilizando ensambladores basados en grafos de De Bruijn (e.g. SPAdes, Canu, HGAP, Flye, IDBA, Falcon, etc) [Segerman (2020)].

Luego de ser ensamblada, esta información debe ser interpretada en términos de funcionalidad mediante la anotación, proceso en el cual se explora el genoma para la búsqueda de secuencias codificantes, aquellas que hipotéticamente poseen funciones dentro de cada microorganismo, basado en comparación y homología de secuencias con genes reportados con anterioridad en microorganismos ya estudiados a nivel experimental obtenidas de bases de datos. Actualmente, este proceso se realiza de forma automática, utilizando diversas herramientas que han sido desarrolladas para el proceso de anotación (RAST, Prokka, PGAP, EggNOG, Orthofinder). Sin embargo, hasta el día de hoy suelen ser poco precisas al analizar proteínas sin función reportada, encontrándose muchos errores en algunas de estas predicciones.

En este proceso, la identificación de agrupaciones genéticas en los genomas obtenidos y su comparación con otros microorganismos mediante homología de secuencias y análisis de sintenia o contexto genómico, permiten reconocer determinantes genéticos únicos que podrían ser de gran importancia para la función metabólica o fisiológica en estudio e identificar nuevos patrones en la evolución de estos genomas. Además, el reconocimiento de secuencias

codificantes hipotéticas dentro de agrupaciones génicas asociadas a una función conocida son de gran ayuda para inferir su función fisiológica desconocida.

La identificación de secuencias homólogas y el posterior análisis de su contexto genómico en genomas bacterianos, así como también de otros organismos, son uno de los primeros análisis que permiten la identificación de nuevos genes y enzimas importantes para el desarrollo científico y tecnológico de las áreas biomédicas, ambientales y biotecnológicas. Nuevos blancos farmacológicos asociados al control de bacterias patógenas humanas, la comprensión de los mecanismos moleculares asociados a enfermedades humanas para el desarrollo de nuevas terapias, nuevos mecanismos de resistencia a ambientes contaminados y el descubrimiento de nuevas enzimas desde bacterias ambientales para la industria son algunos ejemplos de su importancia.

Aunque actualmente existen software dedicados a la visualización de análisis de contexto genómico, como por ejemplo MGcV [Overmars, L. et al., 2013], Mauve [Darling, A. et al. (2004)] o CGViewer [Stothard, P., Wishart, D., (2005)], estos no tienen la flexibilidad necesaria en cuanto a los datos que permiten ocupar para las visualizaciones ni en cuanto a la forma de presentación, lo cual es necesario para una buena comprensión de la información. Por su parte, las herramientas que sí proveen esta flexibilidad, como DNAFeaturesViewer [Zulko et al. (2020)] requieren de habilidades avanzadas de programación para su uso, lo cual es una importante barrera en su adopción, impidiendo el uso cotidiano para usuarios que no poseen este tipo de herramientas y habilidades.

Este proyecto de memoria surge desde la necesidad de una herramienta de fácil uso que posibilite la visualización de contextos genómicos, con la capacidad de presentar datos de distinto origen, generando gráficos personalizables para resaltar la información relevante al estudio en cuestión.

Por medio de diversas técnicas de diseño centrado en el usuario se ha definido el problema a abordar y por medio de un proceso de prototipado rápido y desarrollo iterativo se ha logrado una interfaz funcional para la generación de diagramas de contexto genómico, los cuales facilitan la visualización de datos de genomas bacterianos. Esta aplicación se ha desarrollado en base a tecnología web y ha sido validada por integrantes del Laboratorio de Microbiología Molecular y Biotecnología Ambiental (LMMBA) y otros científicos del área.

En el capítulo 1 se define el problema a abordar y el alcance y objetivos de este trabajo. En el capítulo 2 se presenta un marco teórico que permita comprender el desarrollo e importancia de este proyecto. En el capítulo 3 se detalla el proceso de diseño y desarrollo de este. En el capítulo 4 se describe minuciosamente la aplicación desarrollada y en el capítulo 5 se presentan los estudios realizados para validar dicho producto.

## 1. Definición del problema

Los análisis de contexto genómico consisten en la comparación de uno o más genes homólogos entre dos o más organismos de interés, enfocándose en la presencia y ausencia de genes, direccionalidad de cada uno de estos, tamaño de cada gen y tamaño intergénico entre ellos. Esto permite identificar genes únicos en algún organismo que podrían brindar una ventaja adaptativa al proceso en estudio, genes ubicuos en todos los organismos que aún no han sido caracterizados o cambios en la composición de la estructura general de la agrupación génica.

Si bien existen aplicaciones que facilitan la construcción de diagramas de contexto genómico, y algunos estudiosos del área las ocupan, también hay una gran parte que no siente satisfechas sus necesidades por estas al considerarlas de difícil usabilidad y proveer estas poca flexibilidad en cuanto a sus resultados entregados y por lo tanto los potenciales usuarios utilizan otras herramientas para este fin, siendo uno de los procedimientos más comunes calcular, usando algún programa de manejo de datos como *Microsoft Excel*, las posiciones que le corresponde a cada gen en un lienzo, para luego utilizarlas en algún programa de edición gráfica como *Adobe Illustrator* o *Microsoft PowerPoint*. Este tipo de procedimientos consumen una gran cantidad de tiempo de mano obra altamente calificada, siendo fácilmente automatizables, además de limitar el número de genomas a comparar, lo que restringe la profundidad del análisis y las conclusiones obtenidas.

Si bien esta situación es enfrentada por usuarios sin expertise de programación de áreas biológicas, aparentemente la situación se vive de manera semejante para muchos investigadores del área. Es difícil hacer una estimación numérica de los potenciales usuarios, pero estos fácilmente superan el millón, considerando que sólo entre junio y abril del 2021 se cargaron 765.688 secuencias a GenBank, el principal repositorio de secuencias genómicas a nivel mundial [NCBI (2021)].

Dado que el análisis de contexto genómico es un paso fundamental en los estudios de genómica comparativa y la identificación funcional de genes y que las herramientas existentes son insuficientes por su dificultad de uso o carencia de opciones de personalización, se necesita el desarrollo de una nueva herramienta de visualización de contextos genómico que cumpla con los requerimientos de sus usuarios.

### 1.1. Objetivo General

Construir una interfaz de usuario para la visualización de datos de contexto genómico en genomas bacterianos.

## **1.2. Objetivos específicos**

- Determinar las tareas principales de usuario que se requieren para la visualización de datos genómicos necesarios para la identificación de contextos genómicos de genomas bacterianos.
- Realizar un estudio del estado del arte en cuanto a las herramientas bioinformáticas que sirven al análisis de contexto genómico.
- Desarrollar un prototipo de interfaz funcional para una interfaz que permita a los usuarios realizar dichas tareas.
- Validar el sistema con los usuarios.

## 2. Marco conceptual

Desde tiempos ancestrales se ha distinguido la existencia de ciertas cualidades que caracterizan a cada individuo y que se replican por herencia, a las que llamamos genes, noción que fue formalizada a mediados del siglo XIX por Gregor Mendel. Si bien en lenguaje común se habla de “el gen de los ojos azules” y otras cualidades semejantes, en realidad estas características se expresan como producto de la presencia de diversas proteínas cuya producción está determinada por una secuencia específica de bases nitrogenadas en la larga cadena de ácido desoxirribonucleico (ADN) presente en nuestras células [Dahm, R. (2015)]. El ADN es una doble cadena de nucleótidos que contienen una molécula de azúcar (en particular, una desoxirribosa), un fosfato y una base nitrogenada que puede ser timina (T), adenina (A), citosina (C) o guanina (G). Las dos cadenas están enlazadas entre ellas, de tal manera que cada adenina siempre se corresponde con una timina, y una citosina siempre va con una guanina en la cadena opuesta, unidas mediante puentes de hidrógeno. Esto nos permite describir toda la información genética anotando solamente la secuencia de una de las cadenas [Dale, J., von Schantz, M. (2002)].

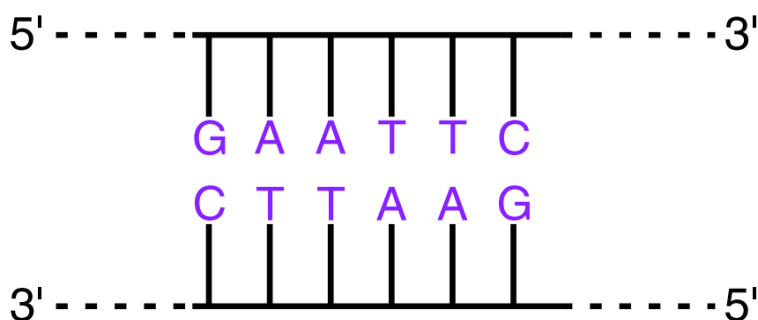


**Figura 1:** Representación de un fragmento de genoma. Nótese que cada base nitrogenada está acompañada de su complementaria en la cadena opuesta.

**Fuente:** Dale, J., von Schantz, M. (2002)

La traducción de una secuencia de nucleótidos a una proteína se realiza mediante codones: Secuencias de tres nucleótidos específicos que determinan la producción de cada uno de los veintiún aminoácidos conocidos. Dado que la unión de múltiples aminoácidos genera proteínas complejas, a las secuencias de codones que forman estas se les llama gen y a la secuencia total de ADN de una especie se le llama genoma [Villée, C. (1996) p. 657], y es de sumo interés el estudio de los de cada organismo y de las semejanzas y diferencias entre ellos, especialmente para conocer su potencial de producción de diversas sustancias, que pueden ser explotadas a nivel industrial y comercial. Dado que es posible descifrar la secuencia de nucleótidos de los genomas, se pueden buscar ciertas características a través de los análisis de sintenia, que es la búsqueda de semejanzas en los genes de distintos organismos y cómo estos genes se organizan.

Durante el proceso de síntesis de proteínas en la célula se produce ácido ribonucleico (ARN), molécula semejante al ADN e igualmente capaz de transmitir información. Para su síntesis, llamada transcripción, se utiliza el ADN como plantilla, generando tres tipos de ARN: El mensajero (mRNA), que posee la información para la producción de proteínas, el ARN ribosomal (rRNA), que constituye parte estructural de los organelos intracelulares encargados del proceso de traducción, llamadas ribosomas, y el ARN de transferencia (tRNA), que actúa como adaptador entre cada codón y el aminoácido correspondiente en el proceso de formación de la cadena polipeptídica. Por este motivo, podemos distinguir tres tipos de genes en la cadena de ADN, según el tipo de ARN que permiten sintetizar.

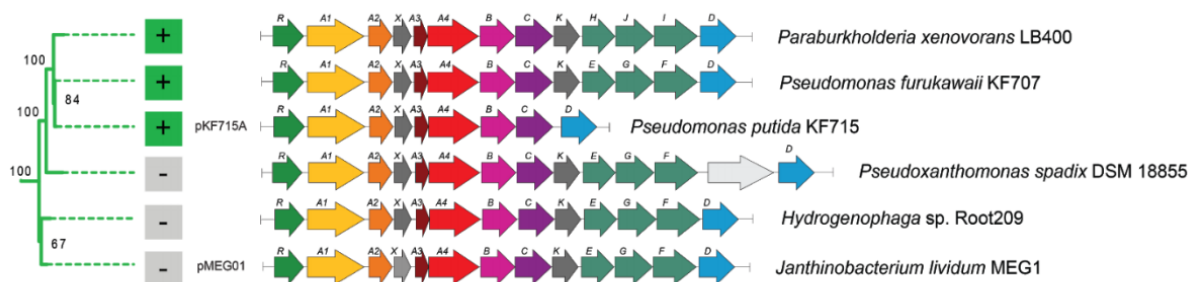


**Figura 2:** Análisis de sintenia en la ruta de degradación de bifenilo en bacterias. Cada flecha representa un gen, y cada fila corresponde a un operón extraído de una cepa distinta, especificada al lado derecho.

**Fuente:** Durán, R., trabajo sin publicar

Al analizar las organizaciones génicas, para indicar la posición relativa entre dos secuencias de interés se dice que una está “río arriba” con respecto a otra si se halla en dirección 5' de la hebra de ADN codificante que contiene el gen de referencia y “río abajo” si se encuentra en dirección 3' de la hebra de ADN codificante que contiene el gen en cuestión (ver figura 1).

Durante los análisis de sintenia, se deben comparar genomas de organismos que pueden ser parte del mismo o diferentes taxones. Un taxón es un conjunto de organismos con un antepasado común al que se le asigna una categoría taxonómica, que puede ser dominio, reino, filo, clase, orden, familia, género o especie. A cada taxón se le otorga un nombre en latín y una descripción y se le asocia un “ejemplar tipo” o “cepa tipo”. Por ejemplo, el ser humano pertenece al dominio *Eukarya*, reino *Animalia*, filo *Chordata*, clase *Mammalia*, orden *Primates*, familia *Hominidae*, género *Homo*, especie *Homo sapiens*. Tener presentes las relaciones taxonómicas de los organismos durante su estudio es esencial para comprender los mecanismos de la evolución y generación de nuevos rasgos genéticos.

Si bien es posible que la presencia de genes homólogos en distintas especies se deba a la herencia desde un ancestro común, también es posible que se deba a una transferencia horizontal en el árbol evolutivo de las especies, es decir, que un gen se transfiera de un organismo a otro de un taxón completamente distinto. Este tipo de transferencia de información genética es común en organismos unicelulares como bacterias y arqueas, las cuales

adquieren rápidamente fragmentos de ADN del medio ambiente o de otros organismos y los mantienen si estos permiten una mejora adaptativa frente a algún estresor ambiental. Algunos de estos genes son transferidos mediante elementos genéticos autotransferibles llamados “transposones”, y su identificación es de sumo interés para comprender los mecanismos de generación de nuevas especies [McClintock (1930)].

Los genes dentro de una célula normalmente no operan para producir el número máximo de enzimas en todo momento, sino que se ven reguladas por la presencia de ciertas sustancias en el medio, estadios de crecimiento o factores ambientales externos. Esto permite que la célula pueda enfrentar adecuadamente cambios del medio como el frío, la deshidratación, entre otros. A los genes que codifican proteínas reguladoras de la expresión transcripcional (síntesis de mRNA) de otros se les llama “genes reguladores” y generalmente se ubican próximos a los genes que regulan. De esto se puede comprender que un grupo de genes próximos entre sí que comparten el mismo regulador comparten una relación funcional, al expresarse juntos en las mismas condiciones ambientales. A dichas agrupaciones génicas se les llama “operones” [Villem, C. (1996) pp. 672 - 675].

En *Bacteria* y *Archaea*, la recombinación y reordenamiento de los genes es muy común durante su rápido proceso de evolución, lo que ha resultado en una baja conservación del orden de los genes entre genomas filogenéticamente distantes, por lo tanto, una secuencia de genes contiguas que se repita en dos o más especies distintas tiene una alta probabilidad de corresponder a un operón. Si no se conoce la función de uno o más de estos genes, su inclusión en dicho operón posibilita una predicción de esta [Wolf, Y. et al. (2001)].

Debido a lo anterior, la conservación del orden de los genes se considera como información muy importante para la predicción de la función de estos, especialmente en organismos de los dominios *Bacteria* y *Archaea*. Por esto, el análisis del contexto de los genes se ha vuelto una estrategia de popularidad creciente, especialmente por el crecimiento de la cantidad de genomas secuenciados a la fecha.

Para conocer la secuencia de bases nitrogenadas que componen el genoma de un organismo existen diversos métodos, que para muchos estudios se utilizan de manera complementaria. El estándar de oro en la industria es la secuenciación de Sanger, que se basa en la polimerización del ADN mediante el uso de dideoxinucleótidos marcados que sirven como terminadores de la reacción. Estos dideoxinucleótidos están unidos a distintos fluoróforos, moléculas que emiten un color determinado al ser excitadas a una longitud de onda específica, de manera que el largo de la cadena en que están insertos permite visualizar la posición de cada base nitrogenada, como se ve en la figura 3.

Si bien este método destaca por su precisión, está siendo rápidamente desplazado en la práctica por el método de Secuenciación de Genoma Completo (WGS), que consiste en la replicación del ADN aislado y la secuenciación de segmentos cortos, que luego son procesados por programas especializados para rearmar la secuencia completa mediante un proceso de ensamblaje (figura

4). Este proceso, a pesar de no dar total certeza en el orden original de las secuencias, es mucho menos costoso y por lo tanto ha permitido incrementar la cantidad de genomas secuenciados de manera explosiva en los últimos años, como se puede ver en la figura 5. No siempre es posible obtener un ensamblaje completo, por lo que el genoma de un organismo queda en fragmentos registrados como *contigs*: Secuencias no necesariamente identificadas como cromosoma o plásmido, sino como fragmento de alguno de estos.

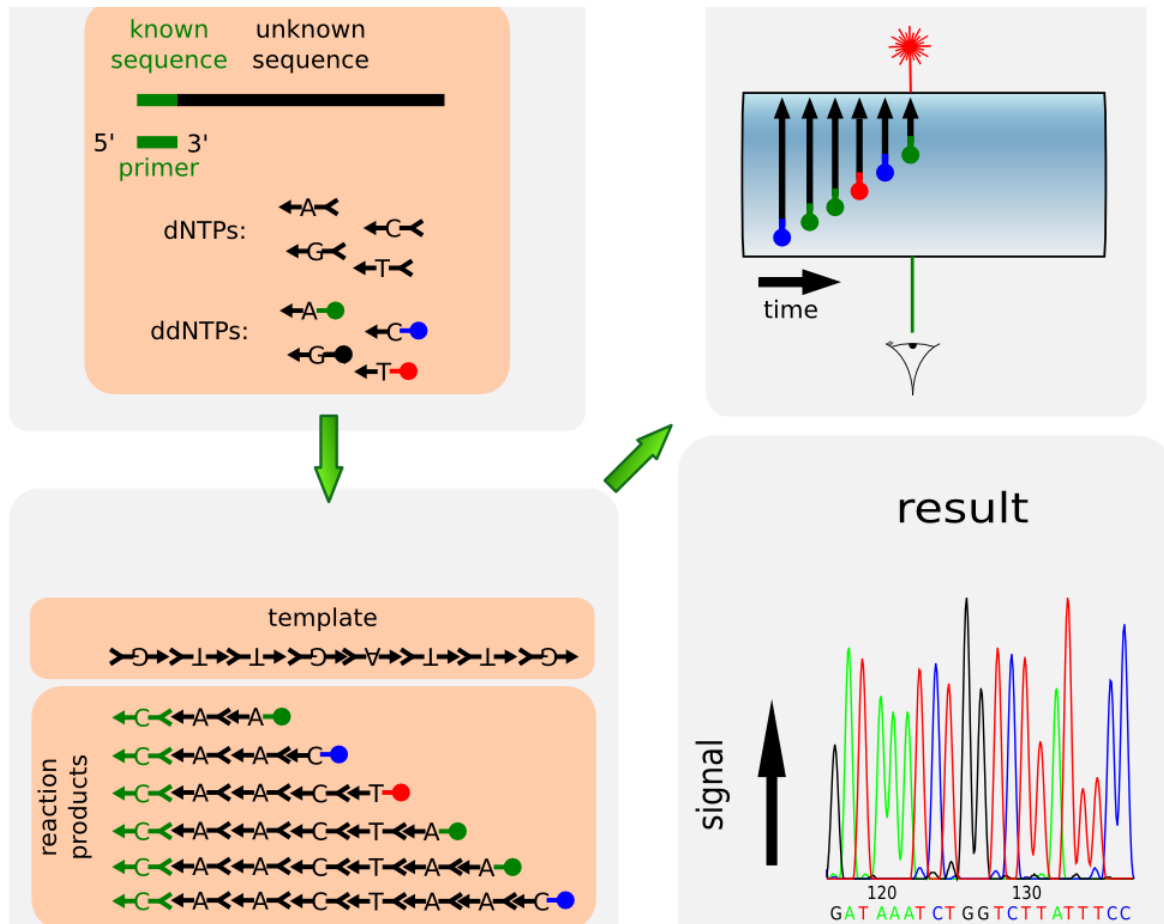


**Figura 3:** Secuenciación mediante Sanger.

**Fuente:** Instituto Universitario de Conservación y Mejora de la Agrodiversidad Valenciana (2021) "Secuenciación de Sanger"

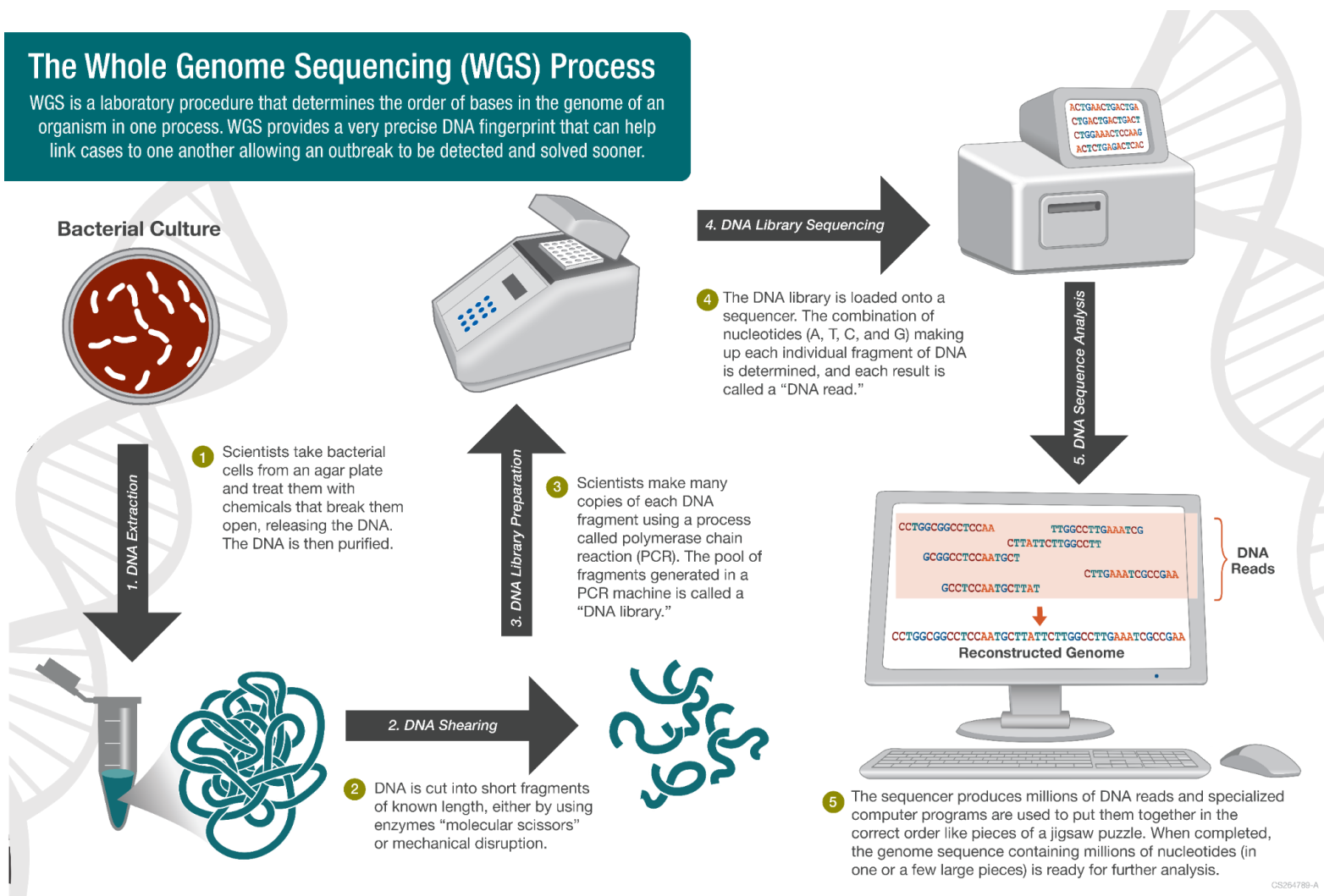


**Figura 4:** Proceso de secuenciación de genoma completo WGS.

**Fuente:** CDC (2016) "Whole Genome Sequencing (WGS)"

## 2.1 Herramientas bioinformáticas complementarias

La bioinformática es un campo interdisciplinario que involucra las áreas de genómica, química, ciencias computacionales, matemática y estadística para analizar datos biológicos y generar nuevo conocimiento, incluyendo la construcción de múltiples herramientas para analizar, organizar, entender y visualizar la información asociada a macromoléculas biológicas [Bilotta, M. et al. (2019)]. El producto desarrollado en el presente trabajo espera abordar la visualización de datos genómicos que pueden tener distintas fuentes de origen y ser manejados por otros sistemas, por lo que se ha considerado esencial que la aplicación desarrollada en el presente trabajo sea compatible con el uso de dichas herramientas bioinformáticas. Esta sección presenta las principales bases de datos, software y herramientas a considerar como “complementarias” de la aplicación desarrollada.

En pos de proveer a los investigadores con un mejor acceso a herramientas de análisis computacional, el Centro Nacional de Información Biotecnológica (NCBI) de la Biblioteca Nacional de Salud (NLM) de los Estados Unidos de América presta una serie de servicios, entre los que destaca el almacenamiento y curación de la más grande base de secuencias genómicas a nivel mundial, GenBank, mantenida en conjunto con el Banco de Datos de ADN de Japón (DDBJ) y el Archivo Europeo de Nucleótidos (ENA).

Esta base de datos contiene una inmensa cantidad de secuencias FASTA, formato de texto plano donde cada secuencia es representada por una descripción en su primera línea comenzando con un símbolo mayor que “>”, para luego continuar con líneas que contienen la información de la secuencia en letras que representan un nucleótido o aminoácido específico, según sea FASTA de nucleótidos o de aminoácidos (este último también llamado “FASTA de proteínas”), junto con la información de su origen y medio de obtención de la secuencia.

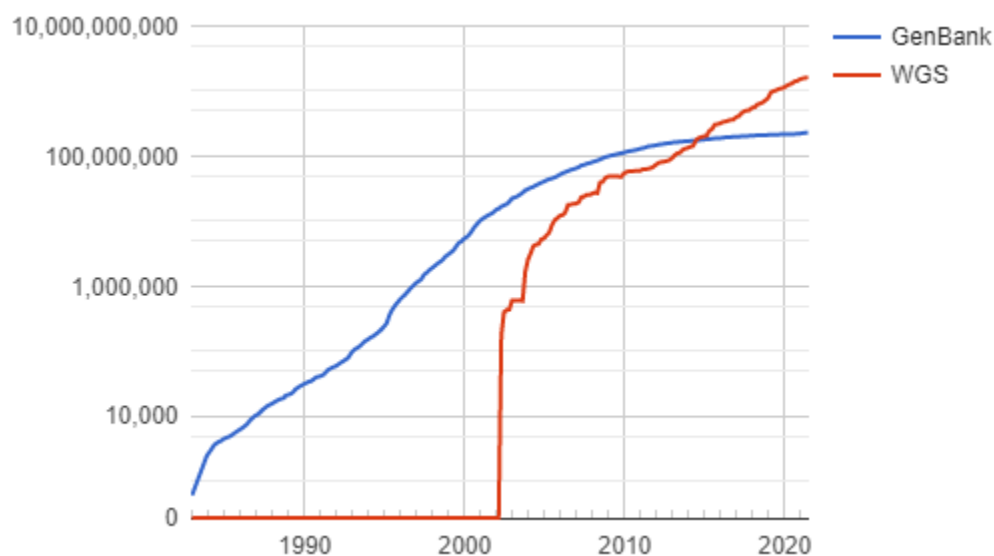


**Figura 5:** Cantidad de registros de secuencia almacenadas en GenBank desde diciembre de 1982 hasta junio de

2021. En **rojo**, las secuencias obtenidas a través de WGS y en **azul** las obtenidas por métodos más tradicionales.

**Fuente:** NCBI (2021) “GenBank and WGS Statistics”

La identificación de distintas lecturas de una misma región genómica (locus) se realiza mediante etiquetas de locus o “*locus tag*”, que corresponden a identificadores únicos de cada región del código genético de cada genoma, únicos a cada cepa de cada especie que haya sido ingresada a esta base de datos, y que se comparte entre todas las lecturas realizadas de la misma región.

Varios genes presentes en distintos organismos pueden haber evolucionado a partir de un ancestro común, por lo que se les llama “genes homólogos” y suelen tener relación funcional o estructural [Pearson, W. (2013)]. El estudio e identificación de genes homólogos es importante para develar las relaciones evolutivas entre distintas especies, comprender los mecanismos de transferencia horizontal de genes e identificar funciones enzimáticas en nuevos organismos caracterizados. Genes que presentan una alta similitud (una similitud superior a la esperable por coincidencia) tienen una alta probabilidad de ser homólogos entre sí.

Por este motivo el NCBI brinda un servicio web de búsqueda de genes similares dentro de sus datos utilizando el algoritmo de alineación de secuencias FASTA “Basic Local Alignment Search Tool” (BLAST) [Altschul, S. (1990)], que es el estándar en la industria para realizar este tipo de búsquedas. Este compara una secuencia denominada consulta o “*query*” con cada segmento de cada una de las secuencias almacenadas en la base de datos y calcula la semejanza, identidad y cobertura de estas.

La identidad entre dos secuencias indica el porcentaje de bases nitrogenadas (o aminoácidos, según corresponda al tipo de secuencia) idénticas entre ambas y la cobertura indica qué fracción de la *query* fue utilizada para calcular dicha identidad, dado que BLAST utiliza las subsecuencias que maximicen la semejanza entre ambas. La semejanza, por su lado, es un valor numérico que se calcula considerando diversos factores. Por lo general, genes con una identidad superior al 30% y una cobertura por sobre el 70% tienen una alta probabilidad de ser homólogos.

Dentro de la inmensa cantidad de datos públicamente disponibles, se distinguen algunos subconjuntos que destacan por la importancia o calidad de sus datos. Uno de estos es la colección de secuencias de referencia (*RefSeq*), que corresponde a un conjunto comprehensivo, integrado, no redundante y bien anotado de los datos de GenBank. RefSeq provee una referencia estable para la anotación de genomas, identificación y caracterización de genes, análisis de mutación y polimorfismo, estudios de expresividad y análisis comparativos [NCBI (2021)].

El conjunto de datos “RefSeq Select” consiste de transcritos representativos (seleccionados) por cada gen que determina la codificación de una proteína. Estos transcritos son elegidos a través de un proceso automático basado en múltiples criterios de selección, que incluyen el uso previo de la secuencia en bases de datos clínicas, la expresión de la transcripción, la conservación de la región codificante, el largo de la secuencia y la proteína y su concordancia con el isomorfo

canónico definido en *SwissProt* (ver más abajo). En *Bacteria* y *Archaea*, es definido como todas las secuencias de proteínas anotadas en las bases de datos RefSeq reference y genomas representativos (representative genomes), en general incluyendo un genoma por especie. RefSeq Select se suele utilizar como referencia en la literatura debido a que los transcritos seleccionados suelen estar bien expresados, conservados y representan en general la biología del gen.

Otro conjunto de datos aún más reconocido por su calidad es el *Universal Protein Resource* (UniProt), mantenida por el Laboratorio Europeo de Biología Molecular (EMBL), que a su vez se divide en el conjunto de datos anotados automáticamente, *Translated EMBL* (TrEMBL) y el de datos manualmente anotados, revisados por expertos, complementados con información extraída de literatura previa y otras herramientas de análisis computacional, llamado *Swiss-Prot*. Este último en particular es el principal referente mundial de calidad de datos genómicos. Las secuencias presentes en UniProt provienen en un 95% de GenBank, pero también se incluyen secuencias de otras fuentes [UniProt Consortium (2021)].

Una etapa importante durante el procesamiento de datos genómicos es la anotación automática, proceso que permite obtener todas las características o “features” de un genoma utilizando comparaciones. El servicio de “*Rapid Annotations using Subsystems Technology*” (RAST) [Aziz, R., et al. (2008)] permite anotar secuencias genómicas, identificando regiones codificantes, rRNA y tRNA, asignando funciones a los genes, prediciendo además los subsistemas representados en el genoma y reconstruyendo la red metabólica que posee, generando un archivo descargable para posterior utilización por parte del usuario.

Prokka es una aplicación de línea de comandos implementada en Perl que permite anotar genomas bacterianos en un tiempo que rodea los diez minutos en un computador de escritorio típico, para lo cual compara las secuencias ingresadas con bases de datos de proteínas bien caracterizadas de distintas fuentes, incluyendo al NCBI y Swiss-Prot [Seemann, T. (2014)].

El “*Prokaryotic Genome Annotation Pipeline*” (PGAP), anotador oficial del NCBI que utiliza el método de comparación con secuencias similares cuando se hallan disponibles datos confiables y suficientes para predecir la función de los genes. En caso contrario, utiliza métodos de predicción estadística analizando las propiedades de la secuencia misma [Tatusova, T. et al. (2016)]. Esta pieza de *software* tiene un funcionamiento limitado a genomas de *Bacteria* y *Archaea*.

EggNOG-mapper es otra herramienta para anotación funcional basada en asignaciones precomputadas de ortología, es decir, que aprovecha la información taxonómica disponible para predecir con mayor precisión la función de los genes [Cantalapiedra, C. et al. (2021)]. RAST, Prokka, PGAP y EggNOG-mapper son ampliamente utilizadas para la anotación de genomas y por lo tanto han sido identificados como las principales fuentes de los archivos de entrada de la aplicación presentada en este trabajo.

Un árbol filogenético es un grafo que representa la relación evolutiva entre distintos taxones utilizando información de las secuencias de nucleótidos, aminoácidos o inclusive de genomas completos. “interactive Tree Of Life” (iTOL) es una aplicación web de disponibilidad pública que permite visualizar, manipular y anotar árboles filogenéticos y es utilizada por decenas de miles de usuarios a lo ancho del mundo. Está construida en base a Javascript vainilla y HTML5, cargando al cliente con el procesamiento, lo cual permite una alta responsividad a las interacciones del usuario y un fino control en varios parámetros de visualización de datos [Letunic I. (2016)].

## **2.2. Principales herramientas de visualización de contexto genómico existentes**

Actualmente existen en la industria múltiples sistemas que permiten generar diagramas de contexto genómico, que tienen distintas ventajas y dificultades según el conjunto de usuarios en el que está enfocado cada uno [Piérola, M. (2020)]. Para este trabajo se ha realizado un exhaustivo análisis de ellos, sus alcances y limitaciones. A continuación se enumeran y describen las principales herramientas de visualización de contexto genómico y análisis de sintenia disponibles.

### **2.2.1. MGcV**

Iniciales de “Microbial Genomic context Viewer” (o “visualizador de contexto genómico microbial”) [Overmars, L. et al. (2013)], es una aplicación desarrollada en Python y disponible a través de una plataforma web, con conexión directa con el servicio BLAST del NCBI, que es la herramienta más frecuentemente usada para hallar similitudes entre secuencias genómicas [Madden, T. (2002)] [Pertsemlidis, A., Fondon, J. (2001)], permitiendo utilizar los datos que esta herramienta entrega como entrada para la visualización, además de permitir la búsqueda de genes directamente por su nombre. También pudiéndose utilizar datos subidos por el usuario, en un formato específico de esta aplicación, conteniendo sólo la información necesaria y bien ordenada. MGcV posee algunas opciones de personalización de la imagen, incluyendo el rango de nucleótidos a mostrar, el tamaño de la imagen, criterio usado para colorear, y forma de etiquetar los genes.

Su principal problema es que sólo acepta archivos en un formato propio, sin dar la opción de procesar archivos en algún formato más estándar, como podría ser el de GenBank Flat File (GBFF), formato general obtenido de la anotación automática que comprende toda la información de este proceso.

Por otro lado, la funcionalidad de descargar las imágenes generadas no estaba funcionando al inicio de este estudio (diciembre 2020), junto con varios enlaces del sitio y al finalizar este texto (agosto 2021) la herramienta derechamente no estaba disponible.

Además, no permite la personalización de colores individuales, tamaños ni formas, de manera que la imagen resultante es totalmente estándar, sin tener la capacidad de adecuarse a los requerimientos particulares del usuario.

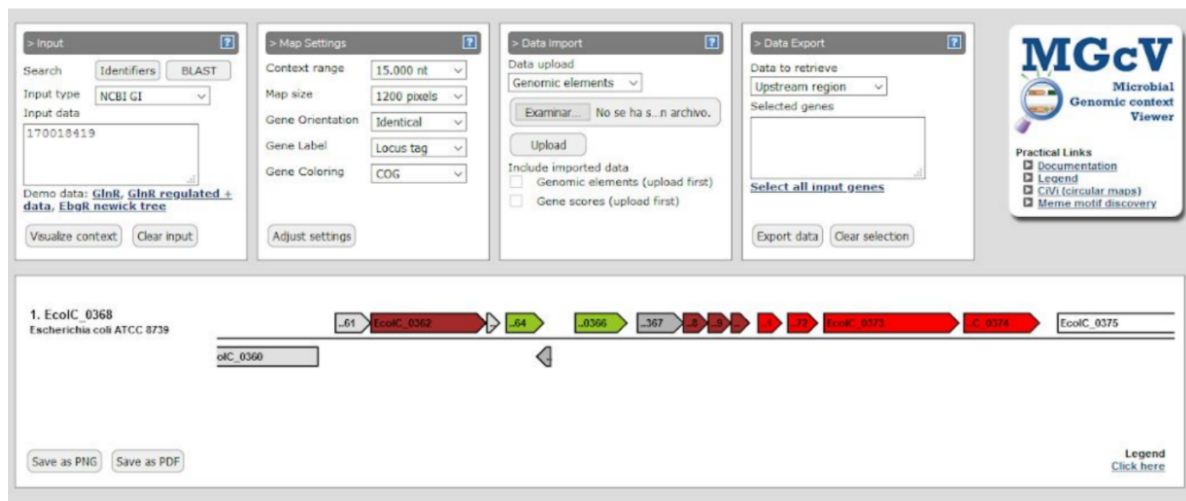


**Figura 6:** Interfaz de MGcV mostrando un fragmento del genoma de *Escherichia Coli*.

**Fuente:** Overmars, L. et al.

## 2.2.2. DNAFeaturesViewer

DNAFeaturesViewer, en español “visor de características de ADN” [Zulko et al. (2020)], es una biblioteca de Python que permite, a través del código, generar archivos de imagen en diversos formatos, a partir de datos de cualquier fuente, incluyendo la lectura directa desde registros de GenBank (archivos en formato .gb o .gff). Es una herramienta de código abierto disponible en GitHub. Esta biblioteca considera muchas funcionalidades, incluyendo la personalización de los colores uno por uno (aunque la coloración siguiendo un criterio debe ser programada por el usuario), la inserción de datos adicionales en la imagen, la posibilidad de separar la imagen en varias líneas, para insertarla en un documento e imprimirla.

Si bien esta es una herramienta muy poderosa, su uso requiere conocimientos de programación, lo cual limita bastante los usuarios que puede tener.

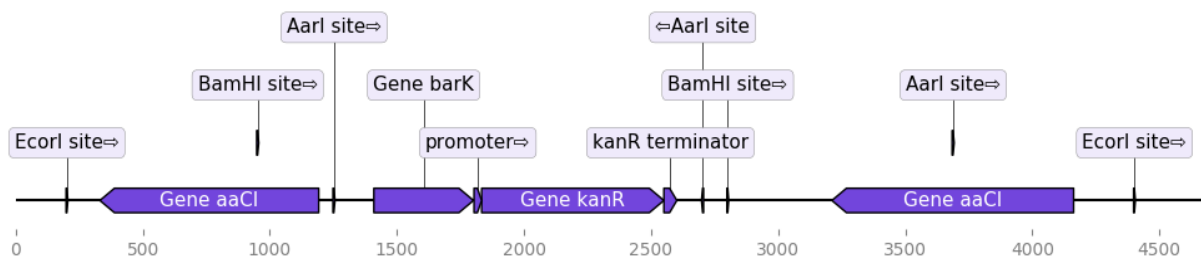


**Figura 7:** Diagrama generado usando DNAFeaturesViewer.  
**Fuente:** Archivo README de DNAFeaturesViewer

### 2.2.3. GeneSpy

GeneSpy [Garcia, P. et al. (2019)] es una aplicación de escritorio que permite importar archivos .gff, con flexibilidad para la forma en que se selecciona un gen de interés y la manera en que se muestra cierta información. La imagen es personalizable en cuanto a tamaño de la región genómica a mostrar, la escala, la información que se muestra, los colores y el orden de los genomas. En particular en cuanto a los colores, tiene cierto grado de predicción, asignando colores iguales para genes con la misma función, como comportamiento por defecto, pero esto también es personalizable. Facilita la integración con iTOL por medio de la exportación de archivos adecuados, aunque no lo incluye directamente en la interfaz.

Lamentablemente, este programa está escrito en Python y requiere de la instalación de múltiples bibliotecas aparte del mismo Python versión 2. Además, su funcionamiento depende de bases de datos tipo SQL que el mismo usuario debe construir y utilizar para seleccionar las secuencias a graficar, lo que implica una engorrosa serie de pasos, como se puede ver en la figura 8.

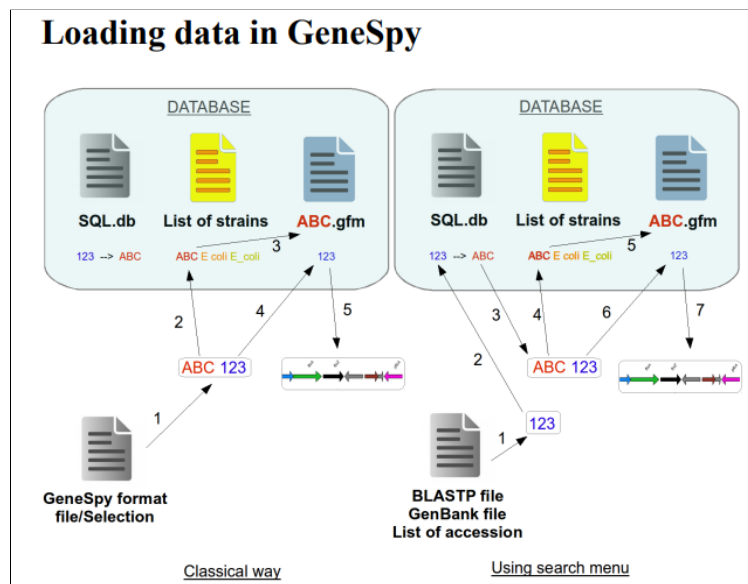


**Figura 8:** Flujo de carga de datos en GeneSpy.

**Fuente:** Manual de usuario de GeneSpy

Por otro lado, su interfaz tiene un aspecto anticuado y antiestético, como se puede ver en la figura 9.



**Figura 9:** Una de las interfaces que forman parte del proceso de selección de los genomas a graficar en GeneSpy.

**Fuente:** Manual de usuario de GeneSpy

#### **2.2.4. Integrative Genomics Viewer**

El “visor genómico integrativo” (IGV) es una aplicación disponible tanto en su versión de escritorio como web, que se caracteriza por la facilidad de navegar visualmente un genoma mostrando distinta granularidad en los datos a distintas escalas, lo cual disminuye la carga para el computador, facilitando su manejo. Su código está disponible libremente en GitHub.

Lamentablemente, no admite registros de genoma anotado, aunque sí recibe archivos del usuario en formato fasta. También permite la inserción de anotaciones por parte del usuario, pero no entrega una forma gráfica de hacerlo, si no que debe importarse esta información a través de archivos en un formato propio de la aplicación, que deben ser hechos “a mano” por el usuario.

Una gran desventaja de esta aplicación es la falta de opciones para personalizar la imagen misma, sin ser posible cambiar formas, colores ni estilo en general.

#### **2.2.5. FlaGs**

Flanking Genes (“genes circundantes” en español), es una herramienta programada en Python para Mac y Linux que también está disponible a través de un servidor web en <http://www.webflags.se/> , que permite utilizar genomas tanto de la base de datos del NCBI como ingresados por el usuario. Posee la característica de que incluye la posibilidad de hacer un análisis de parentesco evolutivo, incluyendo el árbol filogenético en los gráficos que entrega como resultado [Saha, C. K. et al. (2020)].

Sin embargo, no tiene opciones de personalización de la imagen resultante. De hecho, en la versión web, la imagen es entregada en formato SVG por medio de correo electrónico cuando se termina de procesar la solicitud del usuario, sin ninguna instancia de retroalimentación por parte de este.

#### **2.2.6 Gene Graphics**

Esta aplicación web permite graficar contextos genómicos de GenBank y editar la imagen mientras se visualizan [Harrison, K. et al. (2018)].

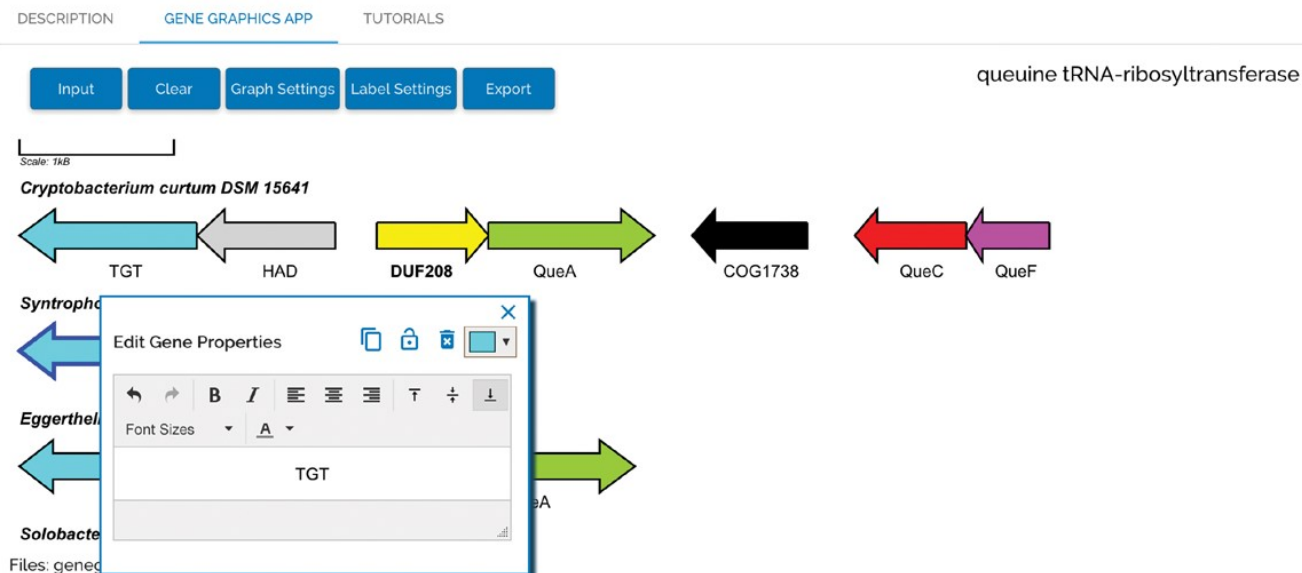


**Figura 10:** Edición de imagen de Gene Graphics.

**Fuente:** Captura de pantalla de <https://katlabs.cc/genegraphics/app>

Si bien tiene amplio rango de personalización del diagrama resultante, se restringe a graficar genomas anotados y almacenados en las bases del NCBI, no permitiendo graficar genomas que el usuario posea, anotados por él mismo o en estado de borrador.

### 2.2.7. Otras herramientas

Existen varias otras herramientas más centradas en facilitar la comparación de genomas, aunque con pocas opciones para las imágenes generadas, como Gene context tool NG [Martinez-Guerrero, C. E. (2008)] o SynTax [Oberto, J. (2013)]. También existen otras bibliotecas de código, para usar por medio de comandos en código, como genoPlotR [Guy, L. et al. (2010)].

### 2.2.8. Apreciaciones generales

Las distintas herramientas existentes tienen enfoque en distintos aspectos, variando en ventajas y desventajas. Los dos factores que más llaman la atención son la falta de facilidad de uso e interactividad en las aplicaciones y la falta de disponibilidad o dificultad de instalación.

Aspectos positivos que debería considerar el desarrollo de una aplicación de este tipo incluyen: La integración con otras herramientas que tienen MGcV y GeneSpy y su coloración según criterios dados por los mismos genomas, la flexibilidad y personalizabilidad que ofrecen DNAFeaturesViewer y Gene Graphics y la eficiencia de IGV. La disponibilidad pública del código de algunas de estas aplicaciones facilita su estudio e incorporación de estas características en la aplicación a construir.

Aspectos negativos que deben evitarse incluyen el uso de formatos de archivo propios o no estándar, la suposición de un gran manejo computacional por parte del usuario y la posibilidad de perder el trabajo avanzado durante una sesión al cerrarse la aplicación por cualquier motivo.

### 3. Proceso de diseño

En este capítulo se describe el procedimiento y técnicas utilizadas en el diseño de una interfaz que facilita la generación de diagramas de contexto genómico, desde el acotamiento del problema hasta las últimas pruebas de usuario. Este desarrollo se ha realizado con un alto involucramiento de potenciales usuarios para obtener una mayor facilidad de uso y capacidad de personalización con respecto a las herramientas existentes en el mercado.

#### 3.1 Definición del problema y alcance de la solución

La exploración del problema desde distintas perspectivas y la especificidad de su definición es un factor determinante en la originalidad y efectividad de las soluciones propuestas [McKilligan, S., Creeger, S (2018)].

Para tener un mejor entendimiento del problema se realizó una serie de entrevistas con Roberto Durán, investigador del Laboratorio de Microbiología Molecular y Biotecnología Ambiental (en adelante, “el cliente”), que permitieron una mayor comprensión de las necesidades presentes, en qué circunstancias se dan y cómo se suelen suplir.

Para asegurar que se estaba enfrentando el problema adecuado se utilizó la técnica de *abstraction laddering*, consistente en formular el problema como una pregunta que empiece con “¿Cómo podemos nosotros...”, lo que nos hace reflexionar no sólo en si la pregunta es necesaria de responder, sino que además reafirma que la respuesta esté limitada a los recursos de los que se dispone. A partir de esta formulación inicial, se pregunta de manera sucesiva “¿Por qué queremos hacer esto?”, para hallar versiones más abstractas del problema y “¿Cómo lo haremos?” para hallar versiones más concretas del problema [LUMA Institute (2012) pp. 50-53] [Pearson, T. (2020)], con lo que se llegó a lo presentado en la figura 11.

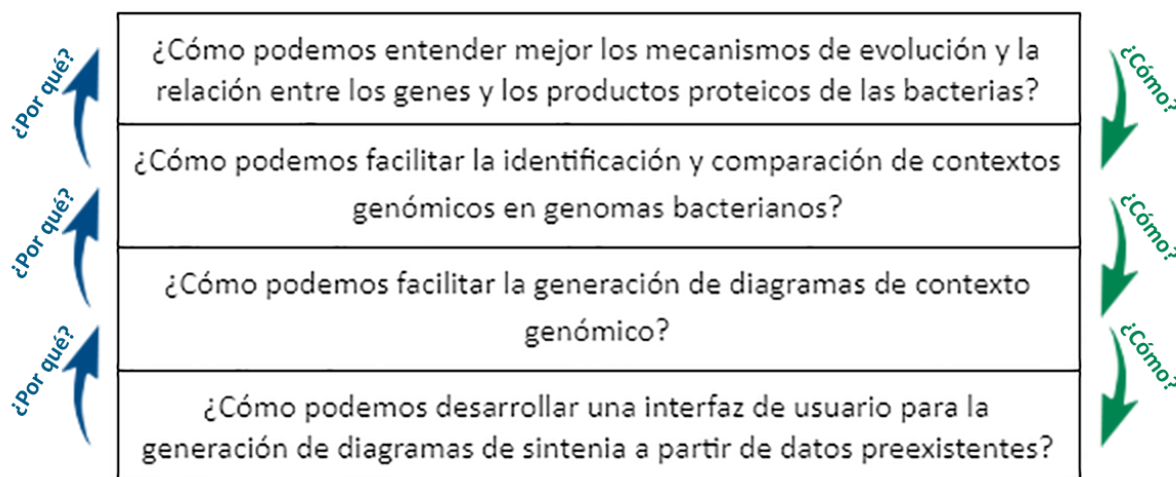


**Figura 11:** *Abstraction laddering*.

**Fuente:** Elaboración propia

La versión más abstracta del problema responde a un trabajo destinado a expertos del área biológica y, de hecho, el motor de la genómica comparativa durante años, por lo que se ha considerado muy amplio enfrentarlo de manera completa. La segunda formulación es posible de afrontar desde el ámbito de la informática con un estudio acotado del tema, pero comprende un flujo de trabajo aún muy amplio para ser abarcado en un trabajo de título de pregrado.

La formulación más específica del problema ya ha sido resuelto por las aplicaciones existentes y utilizadas en la industria, por lo que no es de interés. Sin embargo, el dibujo de diagramas de contexto genómico sigue siendo una dificultad para los académicos y es un problema que puede ser abordado desde un trabajo de título de Ingeniería Civil Informática, por lo que se ha planteado el problema desde esta perspectiva para llegar a la conclusión de que se debía desarrollar una aplicación que facilitara este trabajo.

Para comprender mejor el contexto de uso de la aplicación y los resultados esperados, se realizó un *storyboard* [LUMA Institute (2012) pp. 68-69] representando un caso típico de uso. Con esto se vislumbra la estrecha relación entre la herramienta BLAST (*Basic Local Alignment Search Tool*) y los diagramas de sintenia, ambos herramientas centrales en genómica comparativa.

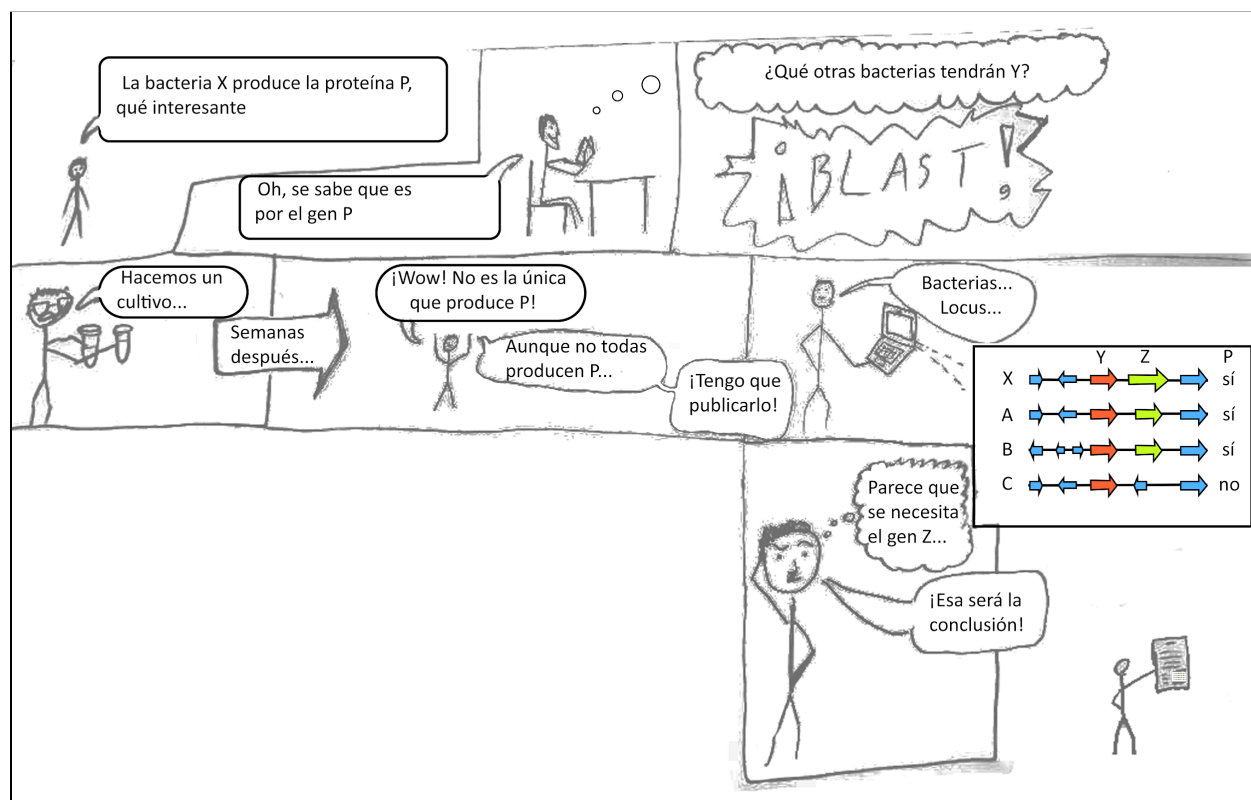


**Figura 12:** *Storyboard* mostrando contexto de uso de la aplicación.

**Fuente:** Elaboración propia

A pesar de las facilidades existentes previamente, los expertos del área siguen considerando muy engorroso el proceso de identificación y comparación de contextos genómicos, al tener que utilizar múltiples herramientas que pueden tener una importante curva de aprendizaje, además de considerar el procedimiento muy largo y tedioso.

Por este motivo se ha propuesto desarrollar un sistema que facilite este trabajo, integrando varios de sus pasos, incluyendo la búsqueda de homólogos y la diagramación de contextos genómicos a partir de genomas anotados.

### **3.2 Caracterización preliminar del sistema**

En base a las alternativas existentes y las necesidades de los usuarios, ha sido posible describir las características deseadas para el sistema a construir:

- Debe prestar gran libertad al usuario en cuanto a la imagen a entregar, permitiendo controlar colores, tipografía, información a mostrar, forma de las flechas, entre otros.
- Debe incluir una interfaz gráfica que sea fácil de usar para el público objetivo, sin requerir mayores conocimientos de programación para su instalación ni uso.
- Debe incluir integración con la mayor cantidad posible de bases de datos genómicos, o al menos GenBank, NCBI, DDBJ y EMBL, además de la posibilidad de que el usuario utilice sus propios genomas, para permitir el uso de archivos no publicados e incluso genomas en borrador (no completamente anotados), desde formatos estándar, sin requerir ningún proceso de transformación previo de los datos por parte del usuario.
- No debe forzar al usuario a tener abierta una conexión de internet durante largo tiempo, de manera que el trabajo no se pierda en caso de ocurrir algún imprevisto técnico en la máquina del usuario.

### **3.3 Prototipado preliminar**

A partir de estas definiciones, se pudo realizar un bosquejo de la interfaz de selección de los contextos a graficar, para validar con el cliente. En dicho prototipo de papel se incluyó la posibilidad de graficar tanto archivos anotados por el propio usuario como genomas presentes en la base de datos del NCBI, con varias formas de especificar la región de interés, además de presentar la posibilidad de graficar directamente los resultados de una búsqueda BLAST en el contexto correspondiente a su genoma de origen, con varias opciones de filtro.

Full-screen uploader

Alt 1

Genoma inicial (Selecciona) ☒ gi/acc.Nº/fasta\*

Desde - hasta  -  codones

☒ Añadir genoma

☒ Buscar semejantes

*g/acc.Nº* Incluir  genes antes  
y  genes después

Los  más parecidos BD:  ☒

☒ Sólo el más parecido de cada género ☒ Sólo del mismo reino

Covertura mínima:  ☒ Sólo del mismo género

Identidad mínima:

Paleta de colores ☒

Dale

Al pasar el mouse, que salgan tooltips mostrando los colores

Figura 13: Prototipo de papel.

Fuente: Elaboración propia

Para familiarizarse con el problema y tener la posibilidad de hacer pruebas de concepto, se ha desarrollado un prototipo empleando tecnología web, en particular, Node.js, Angular, Bootstrap y D3 (figura 14). Este prototipo ha sido utilizado por algunos investigadores del Laboratorio de Microbiología Molecular y Biotecnología Ambiental en sus estudios [Durán, R et al. (2021)] [González Rivera, E. et al. (2021)].

Este sistema permite la lectura de pequeños segmentos del genoma, pudiendo incluir varios operones para graficar sus secuencias codificantes. También ofrece al usuario la posibilidad de

alinear los distintos contextos genómicos según un gen de interés, además de proporcionar una mínima interfaz para la personalización de la imagen, pudiéndose modificar la posición de las etiquetas, el tamaño de letra, colores y grosor de trazo.

Los archivos deben ser subidos desde la máquina del usuario usando el formato .gbff, estándar de genoma anotado, y las imágenes pueden ser descargadas en formato .svg.

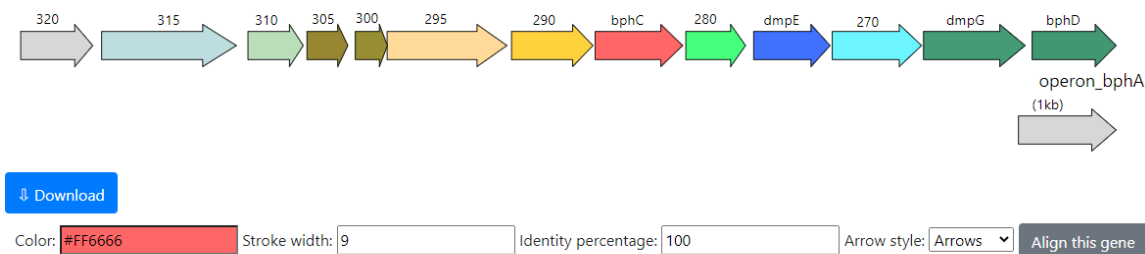


**Figura 14:** Captura de pantalla del prototipo preliminar. Aquí se muestra seleccionado el gen bphC (marcado en color rojo), mostrándose las opciones de edición visual de la imagen generada.

**Fuente:** Elaboración propia

### 3.4 Desarrollo iterativo

A continuación se procedió a desarrollar una versión más completa de la aplicación, con reuniones periódicas para validar el avance con el cliente y plantear alternativas de implementación. A lo largo de este proceso se fueron puliendo o modificando algunas definiciones iniciales, restringiendo ciertas opciones y flexibilizando otras. Por ejemplo, que la búsqueda de homólogos debía ser alternativa excluyente del dibujo de varios contextos definidos por el usuario, que la opción de “sólo del mismo...” debería permitir elegir grupos de cualquier nivel taxonómico, entre otras cosas.

Después de varios meses, se llegó a una versión funcional con seis vistas: Tres a modo de tutorial, para explicar a un usuario nuevo qué es capaz de hacer la aplicación y cómo usarla, una para definir los contextos a graficar utilizando una búsqueda tipo BLAST a partir de un gen de interés, una para definir completamente los contextos a graficar, ya sea desde las bases del NCBI o desde archivos del usuario, y una última de visualización de los datos y edición del gráfico.

El tutorial es mostrado a los usuarios que ingresan por primera vez a utilizar la aplicación. La primera parte explica en general el objetivo de esta y su forma de uso.

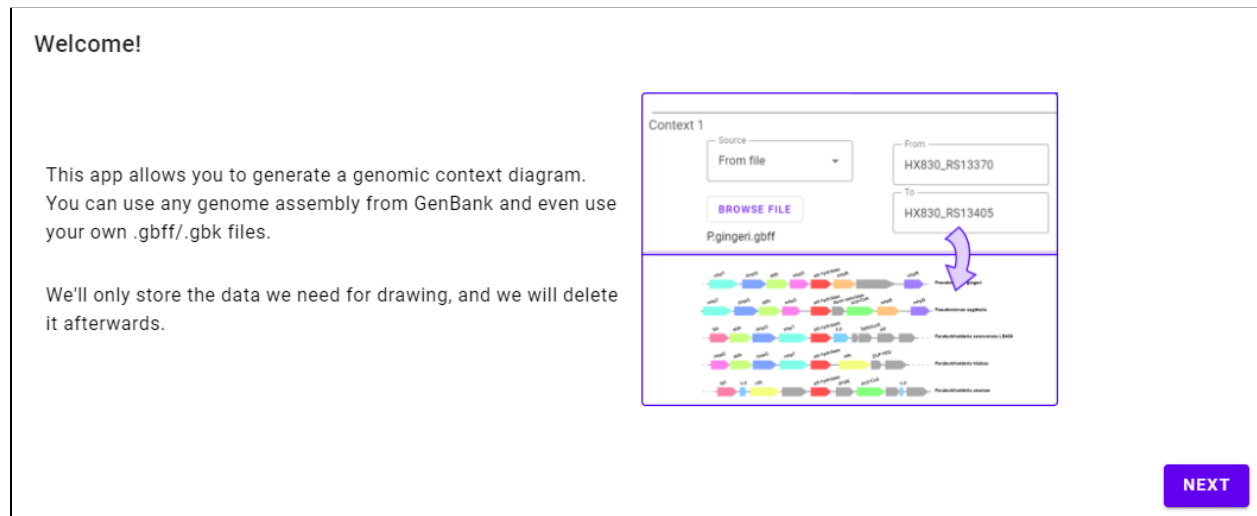


**Figura 15:** Primera parte del tutorial, indicando la utilidad de la aplicación.  
**Fuente:** Elaboración propia

La segunda parte del tutorial muestra la posibilidad de integrar la búsqueda tipo BLAST y graficar los resultados en su contexto directamente.

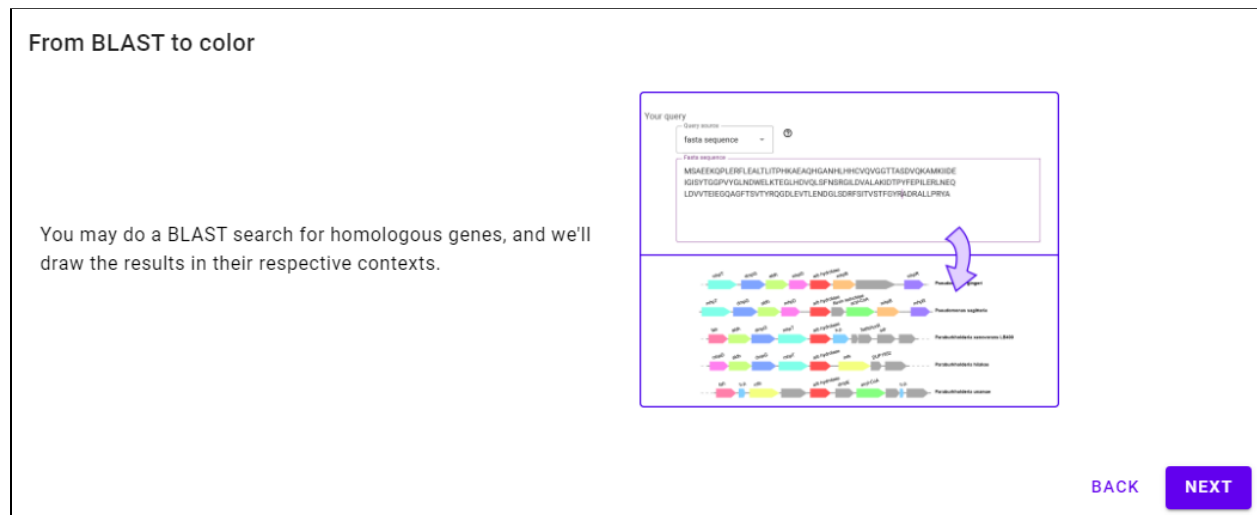


**Figura 16:** Segunda parte del tutorial, mostrando la integración con BLAST.  
**Fuente:** Elaboración propia

La última parte del tutorial muestra la posibilidad de analizar el gráfico visualizando datos del contexto que normalmente no serían visibles en el gráfico, y de editar distintos aspectos visuales de la imagen.

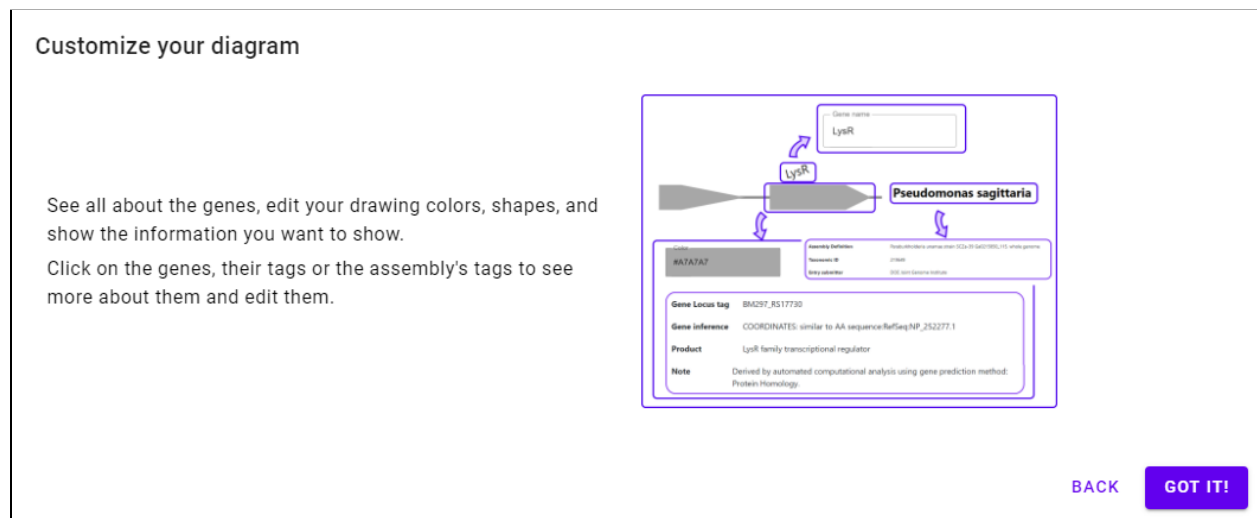


**Figura 17:** Tercera parte del tutorial, mostrando las opciones de personalización del diagrama.  
**Fuente:** Elaboración propia

SEARCH HOMOLOGOUS

DRAW MY CONTEXTS

---

Your query

Query source

?

---

Customize your search

Best

10

matches

?

☐

One of each

Species

?

Minimum coverage:

70

%

?

Database

Reference proteins

?

Include only

?

Minimum identity:

30

%

?

DRAW IT!

?

**Figura 18:** Interfaz de definición de la búsqueda BLAST.  
**Fuente:** Elaboración propia

Una vez finalizado el tutorial, se accede a la interfaz de elección de los contextos a graficar, ya sea a través de una búsqueda tipo BLAST, o a través de la selección directa de estos. Se definió

la personalización de la búsqueda como la vista por defecto debido a que este corresponde al caso de uso más típico y que incluye mayor trabajo por parte de la herramienta.

Esta permite consultar homólogos de una secuencia FASTA ingresada como texto, o extraída de un genoma anotado, ya sea como archivo del usuario o extraído desde NCBI. Al tomar la decisión de extraer la secuencia de un genoma anotado, esta se grafica en su contexto junto con los resultados.

El usuario debe especificar en qué base de datos le interesa buscar homólogos, y puede filtrar los resultados hasta una cierta cantidad, por cobertura e identidad de los homólogos con respecto a la secuencia de consulta, o por criterios taxonómicos de los genomas en que se encuentren los resultados.

The interface is titled "DRAW MY CONTEXTS" and is divided into two main sections: "Context 1" and "Context 2".

**Context 1:**

- Source:** A dropdown menu set to "From file" with a help icon (?).
- Action:** A "BROWSE FILE" button.
- File Name:** A text field containing "P\_xenovorans\_LB400.gbff".
- From:** A text field containing "BXE\_RS42260".
- To:** A text field containing "BXE\_RS42320" with a help icon (?).

**Context 2:**

- Source:** A dropdown menu set to "Assembly accessic" with a help icon (?).
- Accession number:** A text field containing "GCF\_000807755.1".
- From:** A text field containing "QX33\_RS06180".
- To:** A text field containing "QX33\_RS06225" with a help icon (?).

**Bottom Controls:**

- A "+ ADD CONTEXT" button with a help icon (?).
- A "DRAW IT!" button with a help icon (?).

**Figura 19:** Interfaz de selección de contextos a graficar.

**Fuente:** Elaboración propia

Por otro lado, la pestaña de selección directa de contextos a graficar, permite subir genomas anotados desde la máquina del usuario o extraerlos desde la base de datos del NCBI. En ambos casos, se puede especificar la región de interés a través del primer y último *locus tag* de dicha región. Por defecto, en esta vista hay espacio para insertar un contexto y en la parte de abajo un botón permite incrementar la cantidad de contextos a graficar.

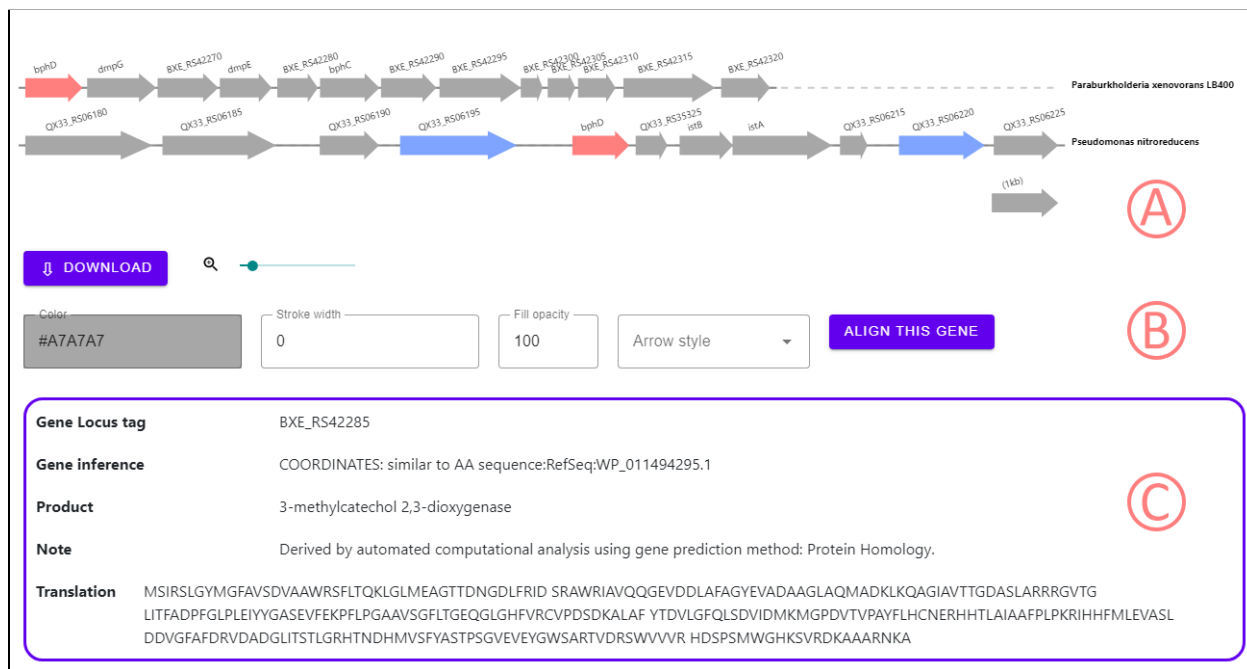


**Figura 20:** Interfaz de visualización de datos y edición del diagrama. (A) Gráfico descargable, (B) Barra contextual de herramientas de edición y (C) Datos adicionales del elemento seleccionado, que puede ser un contexto completo o un gen en particular.

**Fuente:** Elaboración propia

Finalmente, la interfaz de visualización de datos y edición del gráfico muestra en todo momento el diagrama tal cual como se descargaría, junto con un botón para descargar la imagen en formato svg y una barra deslizable para ver el diagrama con mayor o menor aumento.

A medida que el usuario hace clic en cada elemento del diagrama se abre una barra contextual de edición y una sección de información adicional. Las opciones de edición varían según la naturaleza del elemento seleccionado, que puede ser una flecha (es decir, la representación de un gen), una etiqueta de flecha, o una etiqueta de contexto. La información adicional puede referirse a un gen en particular (en caso de seleccionar una flecha o su respectiva etiqueta) o al genoma del que fue extraído el contexto (en caso de seleccionar la etiqueta a la derecha o la línea central). Todos los textos son móviles por medio de arrastrarlos con el cursor.

### 3.5 Pruebas de usuario

16 investigadores, profesionales y estudiantes de doctorado de áreas biológicas (e.g. microbiología, bioquímica, biotecnología) que necesitan realizar análisis de contextos genómicos como parte de su trabajo probaron este prototipo a través de la técnica del *think-aloud testing*: Todos los sujetos realizaron tareas predefinidas utilizando la aplicación, expresando sus sentimientos y pensamientos en voz alta durante su uso para evaluar la facilidad de uso de la interfaz e identificar elementos que puedan ser malinterpretados, con el fin de comprender la raíz de estas confusiones y poder corregirlas rápidamente [LUMA Institute (2012)]

p. 20]. Además, entre estas pruebas se realizaron pequeñas modificaciones para experimentar con distintas posibilidades e ir validando las observaciones realizadas en las entrevistas iniciales al verificar que los usuarios posteriores no se encontraban con las mismas dificultades.

### 3.5.1 Tareas de usuario

Las tareas de usuario asignadas a cada sujeto corresponden a dos o tres de cada una de las siguientes, con ligeras modificaciones y en orden variable:

- **Tarea definida por el usuario**

En muchos casos, el usuario mostró mucho interés en probar la aplicación con los contextos o genes que estaba estudiando en el momento. Este tipo de pruebas permitió más que ningún otro comprobar la utilidad funcional de la aplicación.

- **Graficar un contexto**

La tarea más sencilla consistió en graficar una región predefinida de un genoma especificado por el entrevistador, destacando un gen de interés central en esta región y escondiendo dos flechas correspondientes a genes con características particulares, donde no incluirlas en el gráfico tiene sentido biológico. Los datos utilizados fueron los mostrados en la tabla 1.

**Tabla 1: Datos utilizados para la tarea de graficar un solo contexto**

**Fuente:** Elaboración propia

| Nombre del organismo                     | Región de interés   |                     |
|------------------------------------------|---------------------|---------------------|
|                                          | Desde               | Hasta               |
| <i>Paraburkholderia xenovorans</i> LB400 |                     |                     |
| Identificador de acceso                  | BXE_RS25530         | BXE_RS25585         |
| GCF_000013645.1                          | Cromosoma 2, 805429 | Cromosoma 2, 805429 |

- **Graficar varios contextos**

Similar a la anterior, pero con múltiples contextos en la misma imagen. Importante para probar la facilidad de la especificación de múltiples regiones de interés y las opciones de edición que solo tienen sentido para la comparación entre contextos, como la opción de alinear el gen de interés en cada contexto. Los datos utilizados fueron los que se presentan en la tabla 2.

**Tabla 2: Datos utilizados para la tarea de graficar varios contextos**  
Fuente: Elaboración propia

| Nombre                                              | Identificador de acceso | Desde         | Hasta         |
|-----------------------------------------------------|-------------------------|---------------|---------------|
| <i>Paraburkholderia xenovorans</i><br>LB400         | GCF_000013645.1         | BXE_RS25530   | BXE_RS25585   |
| <i>Paraburkholderia hiiakae</i><br>strain LMG 27952 | GCF_904848665.1         | JMY28_RS11405 | JMY28_RS11460 |
| <i>Paraburkholderia unamae</i><br>strain SCZa-39    | GCF_003096875.1         | C7402_RS28280 | C7402_RS28335 |

Tanto estos datos como los usados en la tarea anterior fueron entregados a los usuarios de distintas formas, por ejemplo, entregando sólo el identificador de acceso, sólo el archivo con el genoma anotado o entregando tanto el archivo como el identificador. También se variaron ligeramente las regiones de interés, agregando algunos genes río abajo o río arriba, para forzar a los usuarios a enfrentarse a distintas problemáticas

- **Hacer una búsqueda tipo BLAST**

Tarea en que se entregó un gen de interés extraído desde las bases del NCBI, para que el usuario graficara en su contexto homólogos presentes en RefSeq que cumplieran distintos requisitos. A todos los usuarios a los que se les solicitó esta tarea, recibieron el archivo mostrado en la figura 21.

```
>BXE_RS42260 from GCF_000013645.1
MTALTESSTSKFVKINEKGFSDFNHYNEAGNGETVIMLHGGGPGAGGWSNYYRNVGP
FVDAGYRVILKDSPGFNKSDAVVMDEQRGLVNARAVKGLMDALDIDRAHLVGNSMGGA
TALNFALEYPDRIKGLILMGPGLGSPMFAPMPMEGIKLLFKLYAEPSYETLKQMLQV
FLYDQSLITEELLQGRWEAIQRQPEHLKNFLISAQKAPLSTWDVTARLGEIKAKTFIT
WGRDDRFVPLDHGLKLLWNIDDARLHVFSKCGHWAQWEHADEFNRLVIDFLRHA
```

**Figura 21:** “BXE\_RS42260.faa”, archivo enviado para búsqueda y diagramación de homólogos presentes en RefSeq y su contexto.

Fuente: Elaboración propia

A los usuarios se les solicitó filtrar los resultados de distinta manera, graficando distinta cantidad de contextos, restringiendo al género *Paraburkholderia* o graficando solo uno de cada género.

### 3.5.2 Observaciones realizadas

A algunos sujetos experimentales les confundió la visión de las opciones de búsqueda de homólogos cuando dicha sección no formaba parte de su tarea, llevándolos en muchos casos a intentar utilizarlas, empezando a llenar los campos, antes de decidir presionar en la pestaña “Draw my contexts”. Un par de usuarios comentó que “Search homologous” parecía tener una relación de secuencialidad con la otra pestaña. En palabras de un usuario: “Debería buscar primero mis genes y después graficarlos, ¿no?”

Varios usuarios intentaron utilizar archivos arrastrándolos desde el explorador de su sistema operativo a la aplicación, lo cual gatilló el comportamiento por defecto del navegador, visualizar dicho archivo en una pestaña nueva, que no era lo esperado por el usuario.

En la interfaz de selección de los contextos a graficar, a muchos usuarios se les hizo difícil percibir la función del botón “+ Add context” (ver figura 19), como se les dificultó percibir la posibilidad de graficar varios contextos en una sola imagen.

A algunos usuarios les parecía obvio que la región de interés debía ser especificada en cantidad de pares de bases nitrogenadas, y no en *locus tag* de inicio y término, como había sido definido para el prototipo, intentando usarlo así sin dudar y sin revisar el texto de ayuda desplegable.

Por defecto, el gráfico se ajustaba al ancho de la ventana, por lo que al graficar un contexto muy amplio (por ejemplo, un genoma completo) el diagrama se veía muy pequeño, llegando a ser difícil de percibir.

Entre las opciones de edición del diagrama, a algunos usuarios les confundía el comportamiento del botón “align this gene” (ver figura 20), pues era claro que movía algún contexto dentro del diagrama, pero era difícil visualizar cuál y cómo. En palabras de un usuario “Okay, alineó mi gen, pero... ¿Con respecto a qué?”.

Dentro de la misma interfaz, era difícil notar el efecto de cambiar los campos numéricos, tales como tamaño de fuente o grosor del trazo, debido a que estos números estaban en unidades proporcionales a los pares de bases graficadas, donde cada par de bases es una distancia muy pequeña en la escala del diagrama que tiene un largo de miles o millones de estas, por lo que la diferencia entre una tipografía de tamaño 140 y una de 143 podía ser nula, dependiendo de la resolución de la pantalla, y si había diferencia, era difícil de percibir.

### 3.5.3 Comentarios recibidos

En esta área de estudio existen convenciones bien establecidas en cuanto al uso de cursivas: Los nombres científicos de especie deben ir con cursivas, a diferencia del nombre de la cepa. De la misma forma, un nombre de gen debe ir con cursiva, a diferencia de un *locus tag*. El prototipo usado para estas pruebas no consideraba la posibilidad de estilizar los textos de esta forma,

pero muchos usuarios lo mencionaron como un aspecto indispensable en este tipo de diagramas para su publicación en revistas científicas.

Algunos usuarios indicaron que les sería cómodo que al seleccionar un gen se mostrara no sólo la información contenida en el archivo de origen, sino que se complementara con información adicional, por ejemplo, de los dominios de la proteína que se sintetiza a partir de dicho gen.

Los usuarios estaban de acuerdo en que SVG es el mejor formato para descargar el diagrama, sin embargo, que sería cómodo que se pudiera transformar a algún otro formato de imagen más tradicional, que tenga mayor compatibilidad con otros programas y que sea más ampliamente conocido como “formato de imagen”, con enfoque en las personas que no están familiarizadas con los gráficos vectoriales.

Algunos usuarios indicaron que la presentación de la escala como una flecha de referencia no era ideal, porque no corresponde a un gen y por lo tanto no debería ser indicada como una flecha. Algo más cercano a la escala en planos arquitectónicos o geográficos permitiría mayor claridad de lo que indica dicho elemento.

Un par de usuarios indicaron que les interesaba la visualización de genes homólogos entre sí que están presentes en el mismo genoma, cada uno en su respectivo contexto, para lo cual sería razonable utilizar sólo un archivo de entrada para varios contextos en la misma imagen. Si bien la interfaz actual lo permite, es a través de la repetición del archivo, lo cual no es ideal para este caso.

### **3.5.4 Mejoras implementadas**

En base a las observaciones realizadas por el entrevistador y los comentarios recibidos, se hicieron algunos cambios entre las pruebas, con lo que se pudo verificar que estos cambios facilitaban la comprensión de la interfaz. Otros cambios se realizaron después de finalizadas las pruebas, debido a que requerían mayor tiempo de desarrollo y estaban relacionados con aspectos funcionales de la aplicación, por lo que no modificaban la forma de uso de esta.

Se transformó la primera vista (después del tutorial), escondiendo las opciones de búsqueda, con el objetivo de obligar al usuario a tomar la decisión de qué alternativa era la que deseaba utilizar, como se ve en la figura 22, y no sobrecargarlo cognitivamente con las opciones de búsqueda, que sólo causarían confusión en el caso de un usuario que desea definir directamente los contextos a graficar. Además se cambió la formulación de la primera alternativa a “BLAST & Draw” para expresar la completitud de la tarea subyacente, y distinguirla de un paso dentro de un flujo más grande que pudiera abarcar ambas opciones. Se comprobó en más de diez pruebas de usuario que esta disposición facilita la comprensión.

En la interfaz de selección de los contextos a graficar, se eliminó el botón “+ Add context”, de manera que apareciera automáticamente el espacio para ingresar otro contexto al especificar la fuente del último contexto presente en ese momento en pantalla. Es decir, al indicar el primer

contexto a graficar aparece el espacio para un segundo contexto, al indicar el segundo se muestra el espacio para un tercero, y así sucesivamente. Se comprobó en múltiples entrevistas que esto automáticamente permitía al usuario prever el comportamiento del sistema al ingresar más contextos, y no confundía al usuario dejar un espacio vacío al final del listado, satisfaciendo el objetivo de esta modificación.

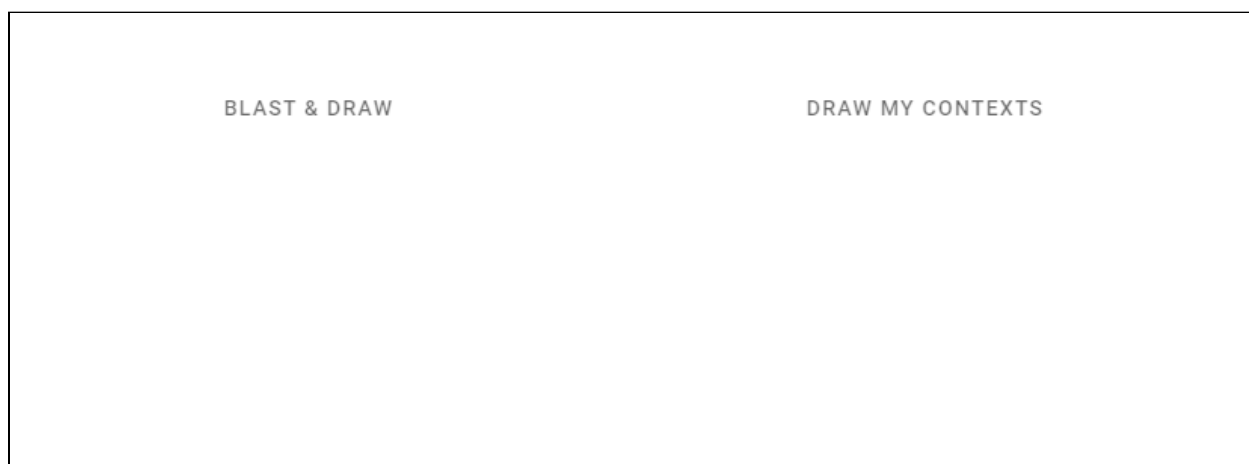


**Figura 22:** Vista principal definitiva. Nótese que se centra la atención del usuario en leer las dos opciones y tomar una decisión antes de presentarle las opciones correspondientes a cada una.

**Fuente:** Elaboración propia

Para todos los campos numéricos que permiten editar aspectos gráficos del diagrama, se definió un incremento acorde a las proporciones del gráfico para los botones de incrementar y decrementar dicho número (ver figura 23), de manera que fuera fácil percibir el efecto de modificar estos valores con sólo uno o dos clics.



**Figura 23:** Campo numérico para edición del gráfico. En rojo, a la derecha, están destacados los botones para incrementar (arriba) y decrementar (abajo) la cantidad numérica indicada.

**Fuente:** Elaboración propia

Se implementó la posibilidad de arrastrar los archivos para enviarlos, especificar la región de interés en pares de bases en vez de ingresar *locus tag* de inicio y término de dicha región, y la escala inicial del gráfico se ajustó considerando un mínimo para los casos en que la razón entre el ancho y el alto de la figura fuere excesivo, para asegurar que siempre sea visible, independientemente de la longitud total, a pesar de que pueda quedar oculta una parte del gráfico, resultando en que este no se pueda visualizar completo a primera vista.

## 4. Resultados

La aplicación resultante está disponible para su uso en la dirección web <http://200.1.19.244:3030/> y su código está en el repositorio público <https://github.com/vsaona/InUsIConGeB.git>

### 4.1 Tecnologías utilizadas

Alojada en una máquina con un procesador de veinticuatro núcleos corriendo Ubuntu 20.04.2 LTS, la aplicación utiliza un servidor web Node.js con el *framework* Express. Las solicitudes que implican alto uso del procesador son trabajadas en paralelo al hilo principal, lo cual le brinda una mayor responsividad al sistema para servir a múltiples clientes de forma concurrente.

Se ha tomado esta elección de tecnologías debido a que tiene un uso más eficiente de la memoria que otras alternativas de la industria, mayor escalabilidad y tiene un mejor desempeño computacional para las peticiones de contenido dinámico y con alto uso de procesador, al utilizar todo el poder de procesamiento disponible [Chaniotis, I. et al. (2015)]. Además, el autor tenía experiencia previa utilizando este *software*, elemento esencial para los plazos de desarrollo.

Para diagramar la información se utilizó la biblioteca de *Javascript* D3 que representa la frase en inglés “Data-Driven Documents”, que en español se podría traducir como “Documentos Derivados de Datos”. Esta biblioteca facilita la representación de conjuntos de datos en gráficos vectoriales. Fue elegida porque muchas de las alternativas utilizan elementos no estándar para construir y representar las imágenes, a diferencia de D3 que se basa en la manipulación de elementos SVG, lo que la hace compatible con cualquier navegador [Van Dierendonck, R. et al. (2015)]

Para la búsqueda de homólogos se evaluó la posibilidad de utilizar DIAMOND, algoritmo de alineación de secuencias homólogas que utiliza un método heurístico para agilizar la búsqueda encontrando resultados cercanos a los óptimos [Buchfink, B. et al. (2021)], pero después de realizar un estudio comparativo con BLAST (*Basic Local Alignment Search Tool*) y una exhaustiva revisión de la bibliografía disponible, se determinó que BLAST era mucho más eficiente para este caso de uso, tanto espacial como temporalmente, debido a que DIAMOND está diseñado para optimizar la consulta de múltiples secuencias y no de un solo gen [Buchfink, B. (2020)].

Para filtrar los resultados de la búsqueda por criterios taxonómicos se utilizó Taxonkit, herramienta de línea de comandos escrita en Go destinada a la manipulación eficiente de datos taxonómicos del NCBI [Shen, W., Ren, H. (2021)].

Para la presentación se han utilizado componentes web de *Material Design*, sistema de diseño creado por Google para ayudar a los equipos de desarrollo a construir experiencias digitales de alta calidad para Android, iOS y web [Google (2021)]. Estos componentes son ampliamente

utilizados en dichas plataformas, por lo que se espera que para la mayor parte de los usuarios sean familiares, lo que aplanará la curva de aprendizaje de la aplicación. Estos componentes se han dispuesto en conjunto con elementos estándar de HTML estilizados con CSS siguiendo los principios de diseño Material.

En particular para la ventana de selección de colores se ha utilizado la biblioteca `jscolor` [Odvárko, J. (2021)], debido a que el campo de ingreso de color por defecto de HTML aún no es soportado de manera estándar por todos los navegadores, teniendo una distinta forma de presentación en cada uno de los que lo implementa.

## **4.2 Descripción exhaustiva de la interfaz**

A continuación se procederá a describir en detalle cada uno de los componentes de la interfaz en su versión presente, posterior a las modificaciones realizadas en base a las pruebas de usuario realizadas.

### **4.2.1 Tutorial**

La versión actual de la aplicación empieza mostrando el tutorial para los usuarios primerizos en un cuadro de diálogo, tal como se muestra en la figura 24. Esto se logra a través del uso de *cookies*. El tutorial es descartable al clicar fuera del cuadro de diálogo, pero se ha decidido no ofrecer una opción explícita para saltarse esta parte debido a que en general es muy útil para el usuario ver esta presentación, incluso cuando crea que no la necesita. El tutorial incluye las tres partes presentadas en las figuras 15, 16 y 17, mostrando las principales vistas de la aplicación y las tareas que se pueden llevar a cabo utilizando esta.

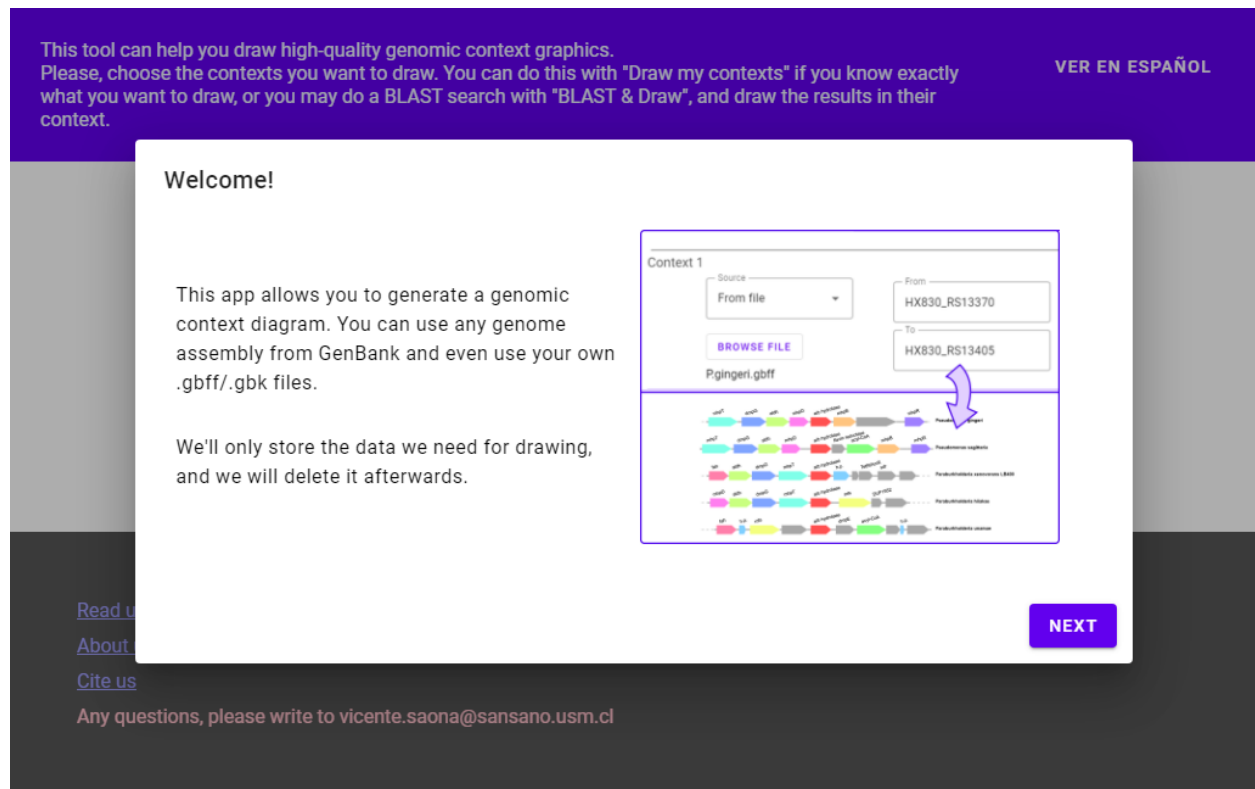


**Figura 24:** Cuadro de diálogo que muestra el tutorial en caso de ser la primera vez que se ingresa a la aplicación.  
**Fuente:** Captura de pantalla de <http://200.1.19.244:3030/>

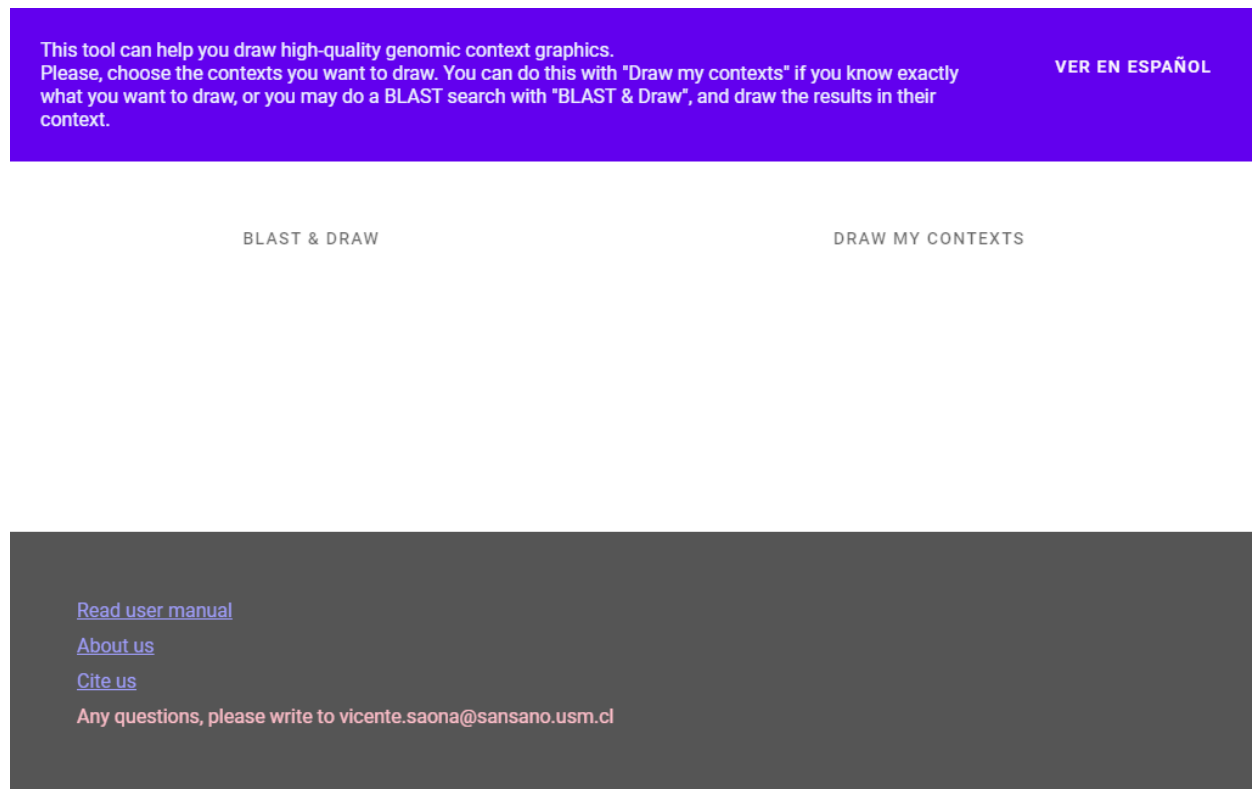


**Figura 25:** Vista principal definitiva donde se elige qué se desea visualizar.  
**Fuente:** Captura de pantalla de <http://200.1.19.244:3030/>

#### 4.2.2 Ingreso de datos

Al cerrar el tutorial, o al abrir la aplicación en caso de no ser el primer ingreso, se muestran al centro las dos alternativas de flujo de trabajo entre las cuales el usuario debe elegir. Con fondos más oscuros, para no llamar la atención, se presentan un encabezado y pie de página, el primero con una pequeña explicación del objetivo de la aplicación y de la decisión que el usuario debe tomar ("BLAST & Draw" o "Draw my contexts"), además de un botón para ver la versión en español de la aplicación, como se ve en la figura 25.

Al pie de página se muestra en primer lugar el enlace al manual de usuario, que empieza con una reiteración del tutorial inicial e incluye una descripción exhaustiva de cada una de las opciones que provee la aplicación, ordenada por secciones y con enlaces a documentos complementarios en caso de ser pertinente.

Se incluye el enlace a una página que presenta a los autores, con un breve resumen de su carrera académica y su rol en el desarrollo de la herramienta, para visibilizar a las personas que respaldan la aplicación y así hacerla más confiable. Adicionalmente, se incluye el enlace al presente documento, para facilitar el acceso a los investigadores y desarrolladores a quienes interese comprender mejor el proceso y motivaciones de la construcción de esta pieza de

*software*, y bajo el entendido de que esto no es un complemento directo del manual de usuario, al tener mucha información que no es útil para facilitar el uso de la aplicación.

Por otro lado, el pie de página incluye los créditos correspondientes, para facilitar al usuario citar a los autores en caso de insertar figuras generadas usando esta aplicación en sus publicaciones científicas. Para esto, se presenta un texto clicable que abre una ventana de diálogo con la referencia en un formato estándar.

Por último, se presenta el contacto del desarrollador, para ser utilizado en caso de que la aplicación tenga algún comportamiento inesperado o *bug*. Se ha considerado que esto es esencial para la mejora continua y el uso sostenido de la aplicación.

### Búsqueda de homólogos

La pestaña de búsqueda se separa en dos secciones: La selección de la secuencia de origen (*query*) en la parte superior y la definición del espacio de búsqueda y filtro de los resultados en la parte inferior, con el título “Customize your search”.

La primera contiene inicialmente sólo un menú desplegable para seleccionar de dónde desea el usuario extraer la secuencia de origen, junto con un ícono de ayuda que tiene la propiedad de desplegar un cuadro de texto al posar el cursor sobre él. Este cuadro de texto explica brevemente para qué se utilizará la *query* y cuáles son las fuentes alternativas para especificarla. El ícono mismo forma parte del conjunto de íconos diseñados por Google y es utilizado consistentemente para el mismo fin en diversas aplicaciones.



**Figura 26:** Selección del origen de la secuencia de origen en la búsqueda de homólogos.

**Fuente:** Captura de pantalla de <http://200.1.19.244:3030/>

Al hacer click en el campo “*Query source*” se despliega un menú con tres opciones, donde la primera corresponde a un archivo de genoma anotado cargado desde la máquina del usuario, la segunda corresponde a un identificador de genoma anotado de las bases del NCBI y la tercera corresponde a la inserción directa de la secuencia FASTA a utilizar para la búsqueda. Al seleccionar tanto la primera como la segunda opción, se graficará en la parte superior del diagrama el contexto de la secuencia de interés en el archivo de origen, además de los resultados de la búsqueda.

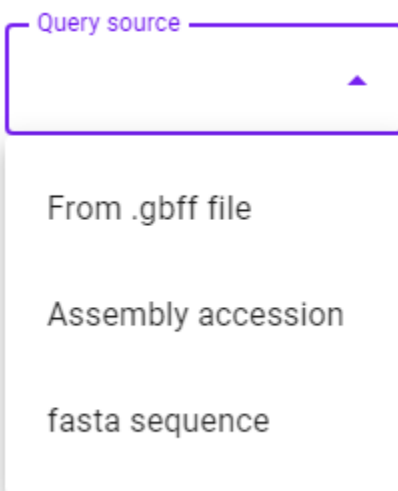


**Figura 27:** Menú de selección del origen de la secuencia de origen en la búsqueda de homólogos.

**Fuente:** Captura de pantalla de <http://200.1.19.244:3030/>

Para el caso en que la secuencia se deba extraer de un genoma anotado completo, se requiere que el usuario especifique además el gen correspondiente, para lo cual aparece un campo de texto al lado derecho donde el usuario puede especificar el *locus tag* que corresponda a su gen de interés, como se aprecia en la figura 28.

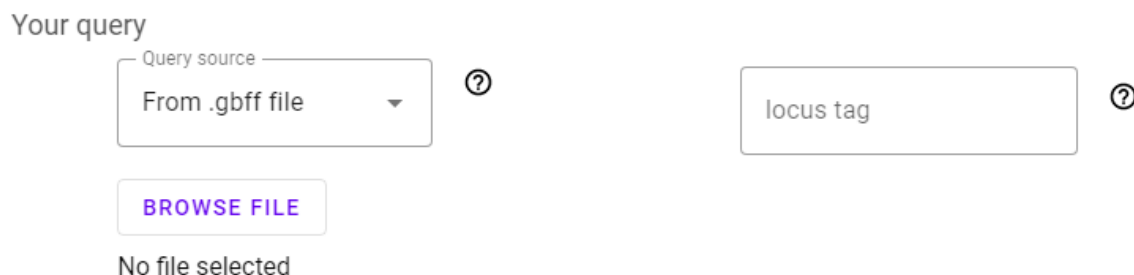


**Figura 28:** Especificación de secuencia *query* desde archivo. Nótese la aparición de un botón para examinar los archivos del sistema y elegir el deseado y del campo de texto a la derecha para especificar el *locus tag* de interés.

**Fuente:** Captura de pantalla de <http://200.1.19.244:3030/>

Para cualquiera de las tres alternativas, aparece un campo debajo del menú desplegable para indicar la fuente de datos. En caso de que sea un archivo, esto corresponde a un botón para examinar los archivos del sistema usando el explorador predeterminado por el sistema operativo y un texto indicando el nombre del archivo seleccionado. En caso de que corresponda a un identificador de acceso del NCBI, corresponde a un campo de texto donde se puede ingresar dicho identificador. Si la elección corresponde a la inserción directa de la secuencia FASTA, aparece un área de texto que ocupa la mayor parte del ancho de la ventana, correspondiente al tamaño esperado del texto a ingresar, bajo el entendido de que este tipo de

datos suele tener una longitud de entre cien y seiscientos caracteres, además de la línea de cabecera.

#### Your query

Query source ?

fasta sequence ▼

Protein fasta sequence

```
>CAI_00002 HTH-type transcriptional activator RamA
MDSQRMKDDDEAVRAALTSLKTATGIPVTMYALLLPDNLQVTQWVGLRTPALQNLIEA
GSGVGGRVSTRRPVGVSDYARANIISHENDRPVQDEGLHSIVAVPVIVSREIRGVLYAG
VHSPVRLGDKVIEEVTMTARTLEQDLAVNSALRRVEGGKVGTAksgrvmngaeweqvrst
HskLRMLANRVDDDLRKELEQLCDQMVPVRVKQSTKLSARELDVLSVALGHTNVEAA
EEMGIGAETVKSylRSVMRKLGAHTRYEAUNAARRIGALP
```

**Figura 29:** Seleccionada la opción de secuencia FASTA, con el campo lleno con un gen de *Corynebacterium alimapuensis*, cepa VA37.

**Fuente:** Captura de pantalla de <http://200.1.19.244:3030/>

Dado que la línea de cabecera no es utilizada para el proceso de búsqueda, que el usuario la incluya o no no afecta a los resultados obtenidos. Sin embargo, se consideró como un punto importante la flexibilidad en su inclusión, debido a que algunos usuarios consideran que la línea de encabezado forma una parte íntegra del formato, y otros no la consideran como tal.

#### Your query

Query source ?

Assembly accessic ▼ ?

locus tag ?

accession number

**Figura 30:** Seleccionada la opción de identificador del NCBI. Nótese la aparición de un campo de texto de largo acorde al esperado.

**Fuente:** Captura de pantalla de <http://200.1.19.244:3030/>

La sección de especificación de la *query* permite que el usuario arrastre archivos. En caso de identificarlo como un archivo de tipo FASTA, se selecciona automáticamente la alternativa de especificar la secuencia directamente en el menú desplegable y se copia el contenido del archivo en el área de texto. En caso de identificarlo como un archivo de genoma anotado, se muestra la opción correspondiente y se carga dicho archivo. Para mostrar la posibilidad de soltar archivos en esta sección se ha seguido la convención de estilizarla con un borde discontinuo mientras se sostiene el archivo sobre esta sección, como se muestra en la figura 31.

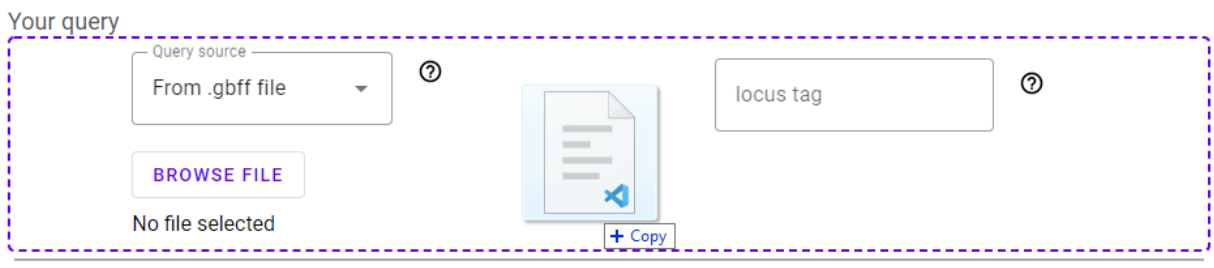


**Figura 31:** Archivo siendo arrastrado sobre la sección de especificación de la secuencia *query*.

**Fuente:** Captura de pantalla de <http://200.1.19.244:3030/> en una máquina con Windows 10

Debajo de la sección de especificación de la secuencia *query*, se extiende la sección de configuración de la búsqueda y filtro de los resultados. En la primera fila están las configuraciones esenciales, de cuántos contextos se desea visualizar y en qué base de datos se deberían buscar estos homólogos.

En la práctica el primer campo actúa como un número máximo de contextos a graficar, pues es posible que no se encuentren suficientes resultados para alcanzar la cantidad especificada, especialmente si se filtran los resultados de manera muy restrictiva.

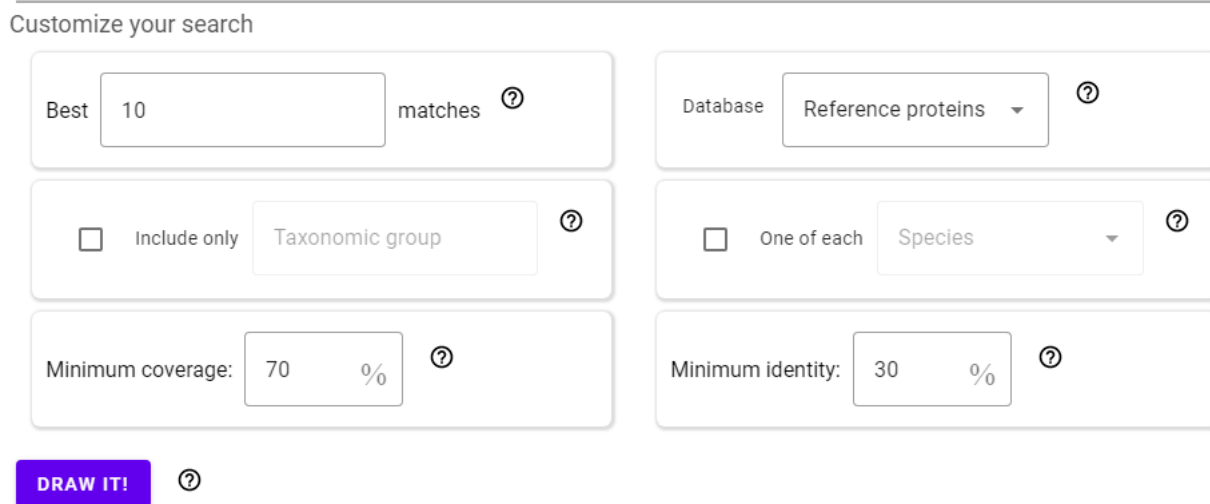


**Figura 32:** Sección de personalización de la búsqueda de genes homólogos, con sus valores por defecto.

**Fuente:** Captura de pantalla de <http://200.1.19.244:3030/>

El segundo campo permite elegir la base de datos dentro de la cual se realiza esta búsqueda de homólogos. Debido a la cantidad de memoria secundaria que se tiene disponible en el servidor en que está implementada esta aplicación, esto se ha limitado a las proteínas de referencia (*RefSeq*), el subconjunto seleccionado “*RefSeq Select*” y *Swissprot*, todas bases de datos con las que suelen trabajar los potenciales usuarios que fueron entrevistados, debido a la alta calidad de los datos que estas contienen. En el eventual caso de que esta aplicación fuera desplegada

con mayores recursos físicos, sería deseable tener más opciones, como lo tiene la herramienta en línea que dispone el NCBI para realizar búsquedas tipo BLAST.

El tercer y cuarto campo permiten filtrar los resultados mostrados según criterios taxonómicos. El de la izquierda, para restringir los resultados mostrados a un grupo específico, ya sea un reino (ej. *Bacteria*), un género (ej. *Paraburkholderia*), etcétera. El campo de la derecha permite mostrar sólo el más cercano de cada grupo nivel taxonómico de un nivel especificado.

Los últimos dos campos permiten restringir los resultados de la búsqueda a aquellos que poseen un mínimo de identidad con y cobertura de la secuencia *query*. Entiéndase que la identidad es un porcentaje de cuántos de los aminoácidos de la secuencia encontrada se presentan también en una sección de la secuencia *query*, mientras que la cobertura indica qué porcentaje de la *query* se utilizó para obtener dicho porcentaje de identidad (recordando que BLAST escoge la sección del *query* que maximice la semejanza).

Si bien esta información es visible posteriormente en la interfaz de edición del gráfico, y es posible esconder los contextos que no sean interesantes para el usuario, a partir de las entrevistas se llegó a la conclusión de que resultados por debajo de cierta cobertura e identidad pueden no tener ninguna significancia, dependiendo del contexto del estudio en ejecución, por lo que es importante dejar esto como filtro previo, para evitar mostrar información irrelevante y evitar trabajo innecesario por parte del usuario.

En el fondo de esta sección aparece el botón “Draw it!”, cuyo objetivo es gatillar el envío de estos datos al servidor para empezar el proceso que culmina en la exhibición del diagrama editable. Al ser presionado este botón, en caso de que haya algún problema con los datos, se destaca el campo en cuestión. Por ejemplo, si se eligió extraer la *query* desde un archivo pero no se especificó el *locus tag*. En caso de no haber problema aparente, se despliega un diálogo que indica que se está procesando la solicitud, que se puede demorar varios minutos, y que la página se actualizará automáticamente cuando el proceso finalice.

### **Selección directa de contextos**

Al presionar la segunda pestaña se ve una interfaz más sencilla inicialmente, sólo con un menú desplegable para elegir el origen del primer contexto, que puede ser un archivo de genoma anotado cargado desde el dispositivo del usuario o un número de acceso del NCBI.

The screenshot shows a web interface with two tabs: 'BLAST & DRAW' and 'DRAW MY CONTEXTS'. The 'DRAW MY CONTEXTS' tab is active. Below the tab, there is a section for 'Context 1'. It contains a dropdown menu labeled 'Source' with a question mark icon to its right. Below the dropdown is a purple button labeled 'DRAW IT!' with a question mark icon to its right.

**Figura 33:** Presentación inicial de la sección de especificación directa de contextos a graficar.

**Fuente:** Captura de pantalla de <http://200.1.19.244:3030/>

Al seleccionar cualquiera de los dos se muestra, debajo del menú desplegable, el campo que corresponda a la alternativa seleccionada, que corresponde a un botón en el caso del fichero y un campo de texto en el caso del identificador. En cualquier caso se muestra al lado derecho dos campos para especificar el inicio y término de la región genómica que se desea graficar y se genera una nueva sección debajo para especificar el siguiente contexto.

Los campos de limitación de la región genómica se pueden ocupar de tres formas:

- Con *locus tags*, en cuyo caso se grafican los genes que estén en el archivo si corresponden a los especificados o si se encuentran entre ellos.
- Con números, que son interpretados como región genómica (en cantidad de pares de bases nitrogenadas) que interesa graficar, con lo que se grafican todos los genes que empiecen en el número menor o río abajo y terminen en el número mayor o río arriba.
- Dejándolos vacíos. Esto se entiende como que la región de interés no tiene límites, por lo que se grafica todo el cromosoma (Cromosoma, péptido o sección de una de estas, según lo que corresponda al contenido del archivo).

Cada sección correspondiente a un contexto permite que el usuario arrastre un archivo de genoma anotado para facilitar la selección de estos, lo cual se marca a través del convencional borde discontinuo, como se aprecia en la figura 34.

The interface is titled "DRAW MY CONTEXTS" and is divided into two main sections: "Context 1" and "Context 2".

**Context 1:** This section is highlighted with a dashed purple border. It contains a "Source" dropdown menu currently set to "From file", accompanied by a help icon (?). Below this is a "BROWSE FILE" button and the text "No file selected". To the right is a file icon with a blue checkmark and a "+ Copy" button. Further right are "From" and "To" input fields, with a help icon (?) next to the "To" field.

**Context 2:** This section is a solid purple block containing a "Source" dropdown menu with a help icon (?) next to it.

At the bottom left of the interface is a prominent purple button labeled "DRAW IT!" with a help icon (?) next to it.

**Figura 34:** Interfaz de selección de contextos al arrastrar un archivo sobre ella.

**Fuente:** Captura de pantalla de <http://200.1.19.244:3030/> en una máquina ejecutando Windows 10.

En la parte inferior de esta pestaña se presenta el botón principal para enviar los datos y generar el diagrama. Al presionarse este, se destacan los datos que hayan sido ingresados de forma aparentemente incorrecta (es decir, que tengan un formato distinto al esperado) en caso de presentarse. En caso contrario, se abre un cuadro de diálogo indicando que la información está siendo enviada al servidor y que se debe esperar algunos segundos. La duración de este periodo depende principalmente del tamaño total de los archivos cargados y el ancho de banda de la conexión que disponga el usuario.

Todos los campos de ingreso de datos de ambas pestañas están acompañados de un ícono de interrogación que permite desplegar un texto de ayuda al posar el cursor sobre este. Estos textos están para ayudar a los usuarios en caso de duda y para hacerlos sentirse más seguros a lo largo del proceso.

#### 4.2.3 Visualización y edición

Después del periodo de carga e independientemente del método utilizado para especificar los contextos, se muestra una ventana con tres secciones: La sección superior contiene el diagrama generado, debajo de este una barra con herramientas generales y finalmente una invitación a utilizar las capacidades interactivas de la aplicación.

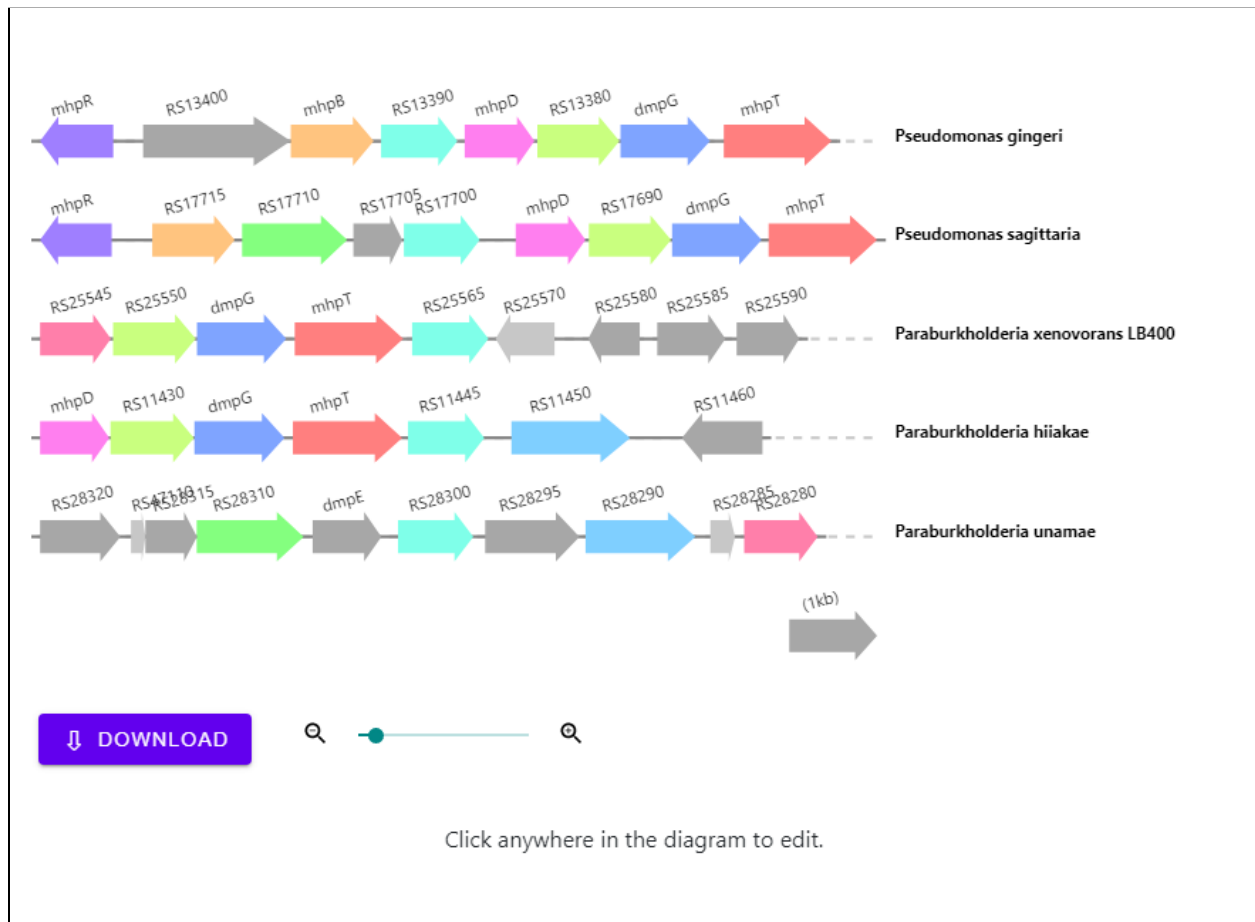


**Figura 35:** Interfaz inicial de visualización y edición del diagrama al graficar el contexto de la alfa/beta hidrolasa en un conjunto de bacterias elegidas arbitrariamente.

**Fuente:** Captura de pantalla de <http://200.1.19.244:3030/>

El diagrama es un lienzo SVG que contiene cada uno de los contextos uno arriba del otro, con un indicador de escala en la parte inferior derecha.

Cada contexto tiene una línea central continua en el segmento graficado y discontinua en las partes del diagrama que no hayan sido consideradas para graficar dicho contexto. En otras palabras, si un contexto considera una región genómica de 200 kb y otro considera sólo 60, el segundo tendrá 140 kb de línea discontinua.

Sobre esta línea central se presentan los rasgos (*features*) presentes en la anotación original del genoma en forma de flecha, desde el menor par de bases hasta el mayor par de bases incluido en dicho rasgo. La dirección de la flecha indica si el rasgo está presente en una hebra del ADN o en la complementaria. Sobre estos, un texto correspondiente al nombre del gen u otra información que permita identificarlo, en caso de estar presente, o al *locus tag* en caso contrario. Esta etiqueta se presenta rotada en quince grados contrarreloj, para evitar colisión entre ellas.

El color por defecto de cada flecha es #BD3B32 (rojo) en caso de que el gen sea uno de los resultados de la búsqueda BLAST o la secuencia *query*. Este color fue elegido para destacar con respecto al resto, incluso para los usuarios con algún grado de dificultad para distinguir colores, al ser más oscuro. Para todos los demás genes, el color es asignado según el producto del gen de manera arbitraria. Es decir, todos los genes que tienen el mismo producto tendrán el mismo color, pero cada producto puede corresponder a un color diferente al graficarlo en distintos contextos.

A la derecha de cada contexto se presenta un texto en negrita para identificar la especie, cepa, operón u otro al que corresponda el contexto. Por defecto, este texto corresponde al nombre de la especie, en caso de estar disponible dicha información, o al nombre del archivo de genoma anotado, en otro caso.

Tanto las etiquetas de contexto como las de gen permiten ser arrastradas con el cursor, lo cual es demostrado con el cambio de imagen del cursor al posar este sobre algún texto, como se muestra en la figura 36. Esto está pensado especialmente para los casos de genes muy pequeños en comparación al resto y que estén muy cercanos, lo cual podría generar que dos etiquetas se solapen ocupando el mismo espacio, resultando ilegibles.

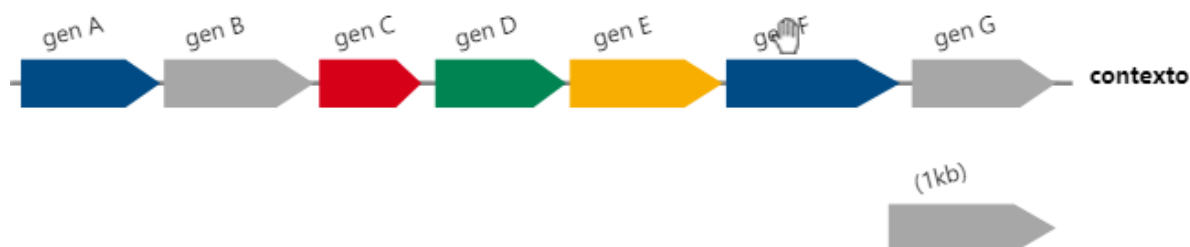


**Figura 36:** Cambio de la apariencia del cursor al ser posado sobre un texto arrastrable.

**Fuente:** Captura de pantalla de <http://200.1.19.244:3030/>

Alineada con la parte derecha de los contextos, en la parte inferior se presenta la escala de mil bases nitrogenadas. Esta comparte todas las capacidades de edición de los genes, por lo que se le puede cambiar el color, el texto, reposicionar la etiqueta, etcétera.

Mientras el usuario se mantenga en esta página, el diagrama se mostrará visualmente fiel a la versión final descargable. Además, mientras se tenga alguna porción del diagrama en pantalla, será visible la barra de herramientas generales, que incluye un botón para descargar el gráfico en formato SVG (*Scalable Vector Graphics*) y una barra deslizable para magnificar el diagrama, acompañada de íconos clicables, uno a la derecha para ampliar y uno a la izquierda para reducir la imagen. En caso de que el tamaño del diagrama sobrepase el de la ventana, el usuario puede navegar a través de él usando el sistema de desplazamiento por defecto de su navegador.

Al hacer clic en cualquier elemento del diagrama, quedará seleccionado, mostrando una barra contextual de edición y una tabla de información adicional de este elemento. Lo que contengan estas secciones depende del elemento seleccionado, que puede ser la representación de un gen, la etiqueta de un gen o un contexto completo, como se muestra en la figura 37. Además, al seleccionar un elemento se mostrará brevemente en la barra de herramientas generales un indicador de cambio de las secciones contextuales, como se muestra en la figura 38 y, de ser la primera vez que se selecciona un elemento, la ventana se desplazará hasta que sea visible el cambio.



**Figura 37:** Los tres distintos tipos de selección: En azul, etiqueta de gen, en verde, representación de gen y en rojo, contexto genómico.

Fuente: Elaboración propia

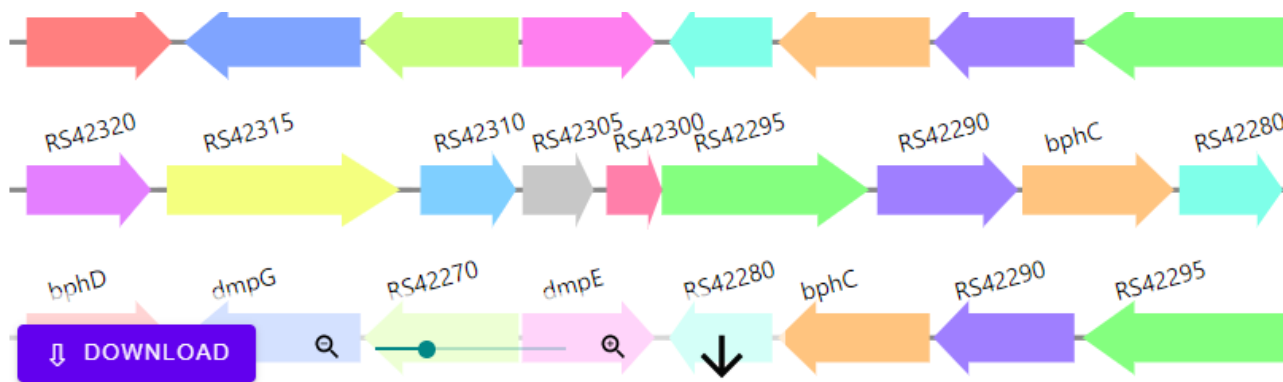


**Figura 38:** Indicador de reacción al seleccionar un elemento.

Fuente: Captura de pantalla de <http://200.1.19.244:3030/>

La barra de herramientas contextuales correspondiente a la representación de un gen incluye un campo de texto que permite cambiar el color de esta por medio de código hexadecimal de color (RGB). Para facilitar su uso, este campo de texto tiene por fondo el color seleccionado y al enfocarlo se abre una pequeña ventana de selección gráfica de colores, como se muestra en la figura 39. Cambiar el color de esta ventana modificará instantáneamente el valor escrito en el campo de texto, el fondo del campo de texto y el componente que corresponda en el diagrama.

La selección de color a través de un mapa como este se ha elegido porque tiende a ser la preferida por los usuarios con ocupaciones fuera del ámbito del diseño [Brathovde, K. et al.

(2019)] Además, La fácil selección de colores ha sido muy valorada por los usuarios entrevistados, y la presencia del código de color en un campo de texto permite que el usuario pueda copiar y pegar dicho texto para asignarle el mismo color a varios genes dentro del diagrama, como se suele hacer para relacionarlos por producto o grupo funcional.

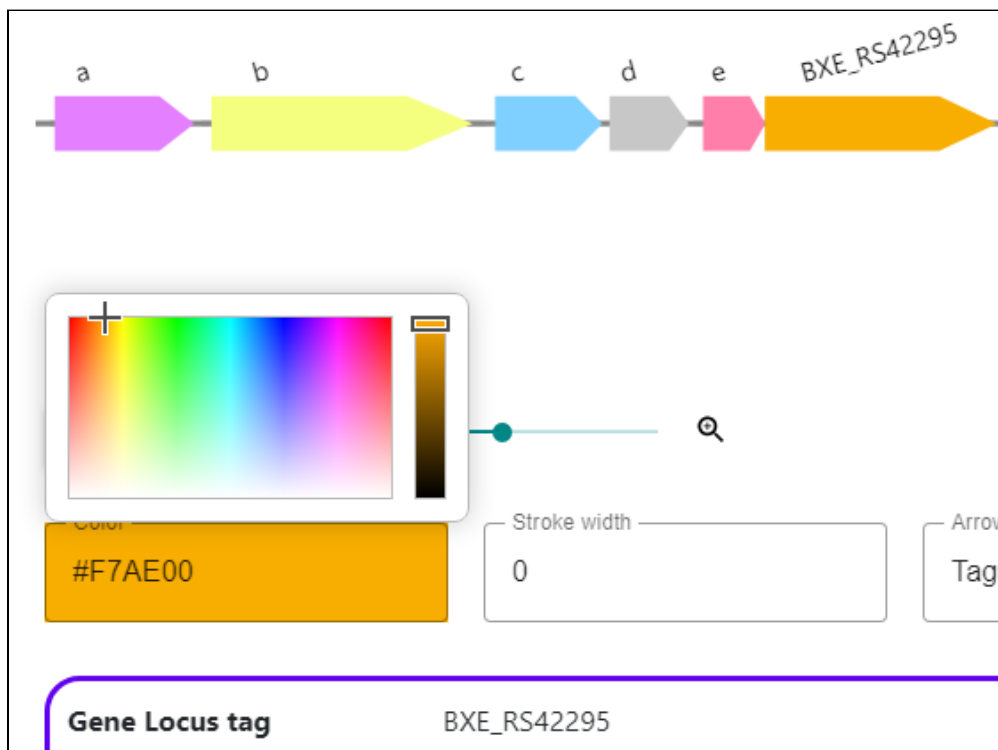


**Figura 39:** Paleta universal de colores y el efecto de seleccionar un color.

**Fuente:** Captura de pantalla de <http://200.1.19.244:3030/>

El segundo elemento de esta barra es un campo numérico que permite modificar el grosor del borde de las flechas, como se muestra en la figura 40. Se ha definido cero como el valor por defecto para darle a los diagramas un aspecto más moderno y que tenga consistencia con la estética de general de la aplicación, definida principalmente por los componentes del sistema de diseño Material. Además, se ha determinado que este número afecte a todos los genes por igual, porque corresponde a una opción estética del diagrama en general y no un rasgo de distinción entre un gen y otro, como sí lo es el color o la etiqueta que acompaña al gen.

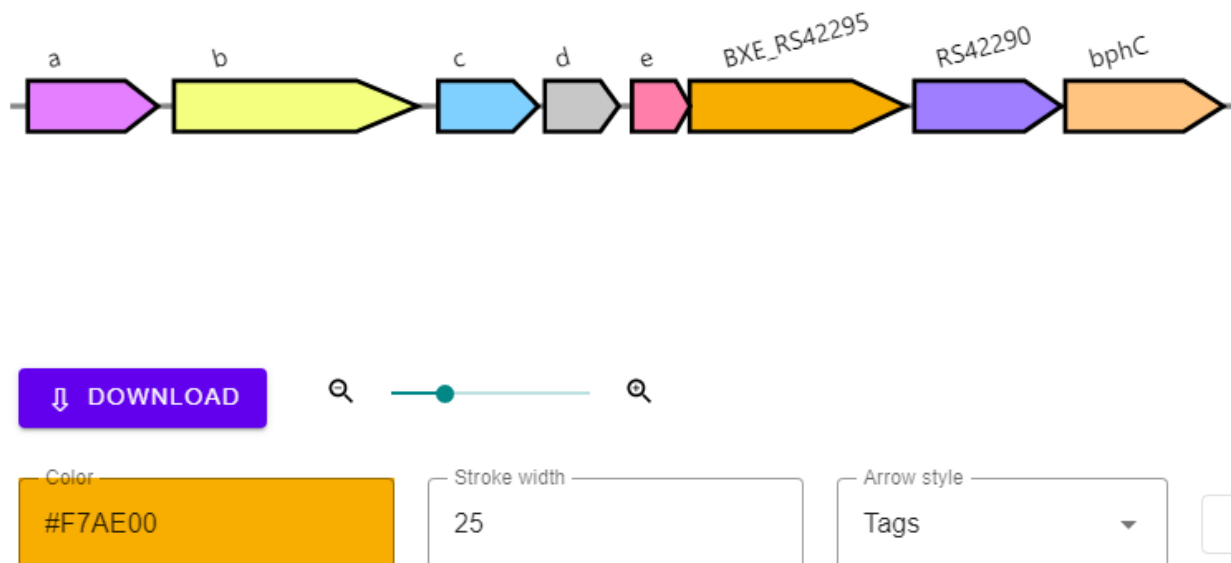


**Figura 40:** Grosor del trazo de las flechas.

**Fuente:** Captura de pantalla de <http://200.1.19.244:3030/>

El tercer campo de edición corresponde a un menú desplegable que permite editar la forma de las flechas. Las alternativas corresponden a las más tradicionales dentro de la literatura existente para diagramas de contexto genómico, que corresponde a las formas de flecha, triángulo y etiqueta, como se puede ver en la figura 41 [Garcia, P. et al. (2019)] [Saha, C. K. et al. (2020)].

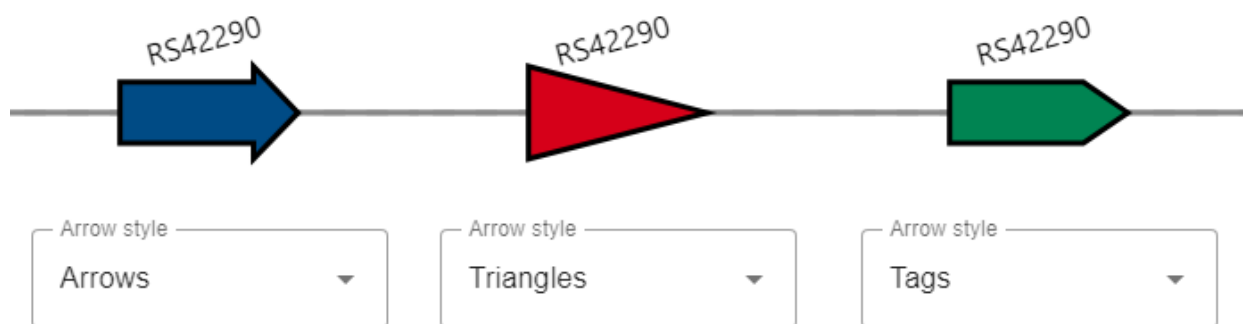


**Figura 41:** Alternativas de forma para representar gráficamente cada gen.

**Fuente:** Elaboración propia.

En la barra se presenta también un botón que permite dejar de mostrar un gen en particular. Esto se ha presentado como una funcionalidad necesaria, debido a que mostrar todos los genes le brinda al usuario toda la información que podría interesarle, sin embargo en muchos casos esto puede incluir secuencias de tRNA, rRNA o genes que sintetizan proteínas hipotéticas (de las cuales aún no se demuestra su existencia), que a criterio del usuario podrían dificultar la visualización de la información de interés más que aportar información útil.

Al elegir esconder un gen, tanto su representación como su etiqueta dejan de ser visibles, y el botón de esconder es reemplazado por uno de mostrar. Después de seleccionar otro elemento es posible volver a seleccionar el gen oculto solamente clicando el espacio que antes ocupaba. Se ha decidido no dejar ningún indicador visual para el gen oculto con el propósito de mantener la fidelidad de lo que se muestra en pantalla con el diagrama descargable.



**Figura 42:** Comportamiento al esconder el gen *dmpG* en *Paraburkholderia Xenovorans* LB400. Nótese que aún mientras el gen está escondido, se mantiene el área clicable para volver a mostrarlo.

**Fuente:** Elaboración propia

El último componente de esta barra contextual permite la alineación de los contextos a través de la selección de un gen de interés. En diagramas generados a través de la búsqueda de homólogos, dichos homólogos quedan automáticamente alineados y centrados, pero al seleccionar los contextos directamente no ocurre esto, por lo que queda para el usuario decidir si le interesa alinear los contextos de alguna manera en particular, y de qué manera.

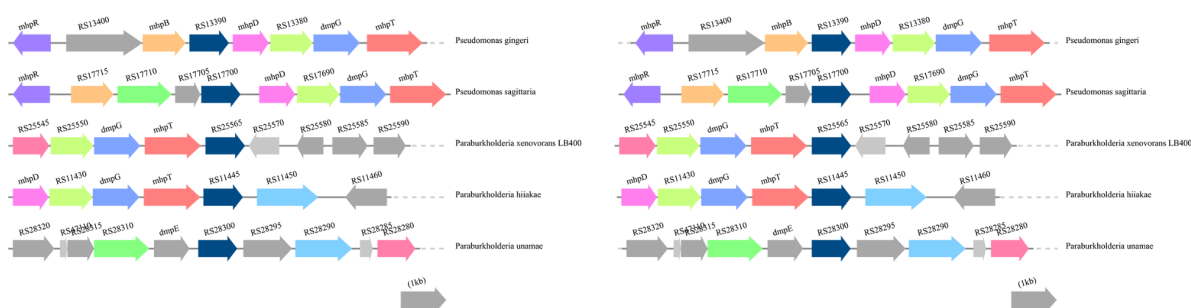


**Figura 43:** Diagrama de sintenia de una serie de bacterias que comparten el gen de la alfa/beta hidrolasa (en azul), con su alineación por defecto (izquierda) y después de seleccionar cada una de las hidrolasas y presionar el botón de alineamiento (derecha).

**Fuente:** Descargado desde <http://200.1.19.244:3030/>

En caso de que el elemento seleccionado corresponda a la etiqueta de un gen, las herramientas contextuales permiten modificar el nombre del gen y el tamaño de letra, como se muestra en la figura 44. El tamaño de letra es considerado un atributo de estilo compartido entre todas las etiquetas de gen. El texto, en cambio, es editable de manera independiente para cada etiqueta.

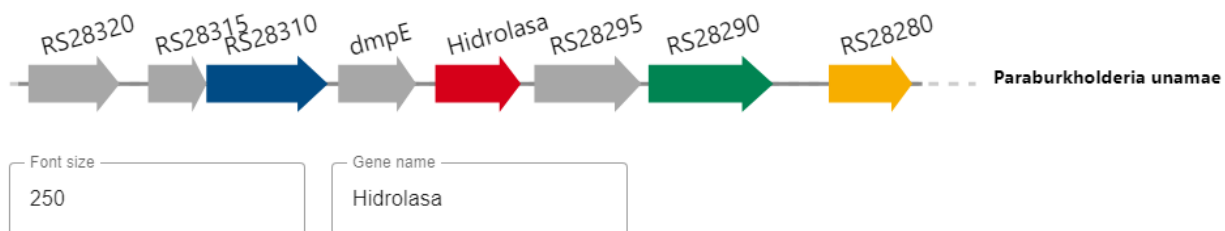


**Figura 44:** Efecto de ingresar un tamaño de fuente superior al por defecto y cambiar el texto de la etiqueta de un gen de hidrolasa en una sección de un genoma de *Paraburkholderia unamae*.  
**Fuente:** Elaboración propia.

Debajo de la barra de herramientas de edición, se muestra una tabla con los datos más relevantes del gen seleccionado, en caso de estar estos presentes en el archivo de origen, como se muestra en la figura 45.

|                |                                                                                                                                                                                                  |
|----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Gene Locus tag | BM297_RS17705                                                                                                                                                                                    |
| Gene inference | COORDINATES: similar to AA sequence:RefSeq:WP_012112954.1                                                                                                                                        |
| Product        | flavin reductase family protein                                                                                                                                                                  |
| Note           | Derived by automated computational analysis using gene prediction method: Protein Homology.                                                                                                      |
| Size           | 557                                                                                                                                                                                              |
| Translation    | MSKEINQLDWIPNADLAALPVVSADHRVGMRLAAAVTIITA<br>RDGDTHLGLTATAVCSVTAEPPRLVVFVNKKVAASEAILNSGALCVNVLSGEQEHEAK<br>VFAGMVEGVHGGEARFQHGQWRDLASGVPLDGAVANFDCRVVKVFDESTHHAFLCEVL ATRGQPGGEPLLYLNGAFRRRLAPQA |

**Figura 45:** Tabla de información adicional al seleccionar el locus BM297\_RS17705 de un genoma de *Pseudomonas sagittaria*.  
**Fuente:** Captura de pantalla de <http://200.1.19.244:3030/>

En caso de seleccionar un contexto genómico, se presenta una barra de herramientas contextuales que permite modificar el tamaño de fuente de las etiquetas, el texto de la etiqueta correspondiente y el grosor de la línea central del genoma, junto con un botón para eliminar el contexto seleccionado, como si este nunca hubiera existido. Tanto el tamaño de letra como el grosor de la línea media corresponden a atributos de estilización para todo el diagrama, mientras que las otras dos opciones afectan sólo al contexto seleccionado.



**Figura 46:** Herramientas de contexto genómico con valores distintos a los asignados por defecto en un fragmento de genoma de *Pseudomonas gingeri*.

**Fuente:** Elaboración propia

Por debajo de esta barra de herramientas de edición, se presenta una tabla con información del genoma utilizado y el origen de los datos, como se puede ver en la figura 47.

|                            |                                                                                                 |
|----------------------------|-------------------------------------------------------------------------------------------------|
| <b>Assembly Definition</b> | Pseudomonas gingeri strain G9001 run0228_G9001_S4_L004_contig_7, whole genome shotgun sequence. |
| <b>Assembly Accession</b>  | NZ_JACAQA010000007 NZ_JACAQA010000000                                                           |
| <b>Source link</b>         | ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/013/385/425/GCF_013385425.1_ASM1338542v1             |
| <b>Taxonomic ID</b>        | 117681                                                                                          |
| <b>Entry submitter</b>     | Wageningen University & Research                                                                |

**Figura 47:** Datos de genoma de *Pseudomonas gingeri*.

**Fuente:** Captura de pantalla de <http://200.1.19.244:3030/>

### 4.3 Flujo de trabajo

En esta sección se describe el proceso y flujo de los datos a lo largo de un uso típico de la aplicación.

Al solicitar el cliente la página principal, según las *cookies* presentes, el servidor corriendo en *Node.js* y *Express* retorna un documento predefinido. Todas las interacciones que se realizan durante el ingreso de los datos, hasta pulsar el botón “Draw it” se trabajan de manera local utilizando HTML, CSS y *Javascript* vainilla.

#### 4.3.1 Búsqueda de homólogos

En caso de realizarse un proceso de búsqueda de homólogos, se procede a recuperar la secuencia *query*. Si se especificó un identificador de genoma, se procede a buscar la ruta de dicho archivo en el archivo resumen de genomas anotados dispuesto por el NCBI “*assembly summary*”. Luego de identificar la ruta del archivo, se procede a revisar si está disponible en el servidor local. De no estarlo, se recupera desde el NCBI. Una vez que el archivo está disponible para su lectura, se recorre buscando el *locus tag* de interés, y, a través del uso de expresiones regulares, se obtiene la secuencia FASTA asociada.

Utilizando BLAST en el servidor local se obtienen los identificadores de los genes resultado, cobertura e identidad con respecto a la *query* y el identificador taxonómico correspondiente. Estos resultados se filtran según la cobertura e identidad especificadas por el usuario. En caso de corresponder, se utiliza *taxonkit* para obtener todos los grupos taxonómicos a los que pertenece cada resultado y filtrar según los criterios especificados por el usuario.

Con los datos filtrados se procede a buscar la ruta de los genomas anotados disponibles que correspondan a dicho taxón. En caso de no estar presentes en el servidor local, se recuperan desde NCBI. Se recorren estos hasta encontrar el gen de interés e identificar su contexto.

Con estos datos *parseados* se procede a la asignación de colores siguiendo las siguientes reglas: La *query* y sus homólogos van de color **#BD3B32** (rojo), los genes que sinteticen proteínas hipotéticas (es decir, cuyo producto haya sido anotado como “*hypotetical protein*”) tienen el color **#C7C7C7** (gris claro) y el resto de los genes van de color **#A7A7A7** (gris) a menos que su nombre o producto se encuentre repetido en otro gen presente en el diagrama.

A cada grupo de genes donde cada gen comparta nombre o producto con al menos uno de los genes del grupo (excluyéndose) se le asigna un color de igual saturación y luminosidad, y el valor del matiz se calcula utilizando el método de *hashing* multiplicativo de Fibonacci, que genera números bien espaciados entre ellos, convergiendo a la equidistancia [Preiss, B. (2001)]. Se ha optado por este método para maximizar la distinguibilidad de los distintos grupos de genes, obteniendo mayor claridad en la lectura de la información subyacente.

Estos datos se envían al cliente junto con los documentos web necesarios para generar el diagrama interactivo. En el navegador, se genera el gráfico por medio del uso de D3.js. Toda la edición y visualización posterior se produce de forma local en el cliente.

#### **4.3.2 Contextos especificados de forma directa**

Respecto a la especificación directa de contextos a graficar, en el caso de los contextos para los cuales el usuario haya ingresado un identificador de acceso, se procede a buscar la ruta de dichos genomas anotados en el resumen “*assembly summary*”. En caso de no estar presentes en el servidor local, se recuperan desde NCBI. Se recorren estos hasta encontrar la región de interés, *parsear* sus datos y asignar colores como fue descrito más arriba para enviarlos al cliente.

#### **4.3.3 Generación del gráfico**

En el cliente se utilizan las capacidades de la biblioteca D3.js para mapear los datos al diagrama, correspondiente a un lienzo SVG con un grupo por cada contexto dibujado y la escala. Cada grupo contiene su etiqueta a la derecha, una línea central continua que tiene largo igual a la región graficada del genoma, y una línea discontinua detrás de esta, con longitud

correspondiente a la región graficada en todos los genomas. Sobre estas, se genera un grupo con una flecha y una etiqueta de texto por cada gen.

Las coordenadas horizontales de inicio y término de los genes se calculan de la siguiente manera:

$$Inicio_{Diagrama}(gen A) = Inicio_{Genoma}(gen A) - Inicio_{Genoma}(gen de interés)$$

$$Fin_{Diagrama}(gen A) = Fin_{Genoma}(gen A) - Inicio_{Genoma}(gen de interés)$$

Donde  $Inicio_{Diagrama}$  y  $Fin_{Diagrama}$  corresponden respectivamente a las coordenada de inicio y fin de la representación del gen en el lienzo e  $Inicio_{Genoma}$  y  $Fin_{Genoma}$  corresponden al índice del primer y último par de bases nitrogenadas de un gen dentro del genoma anotado, respectivamente

Es decir, se establece el inicio del gen de interés como el origen del lienzo y se traza al resto con respecto a este, según sus posiciones relativas en la anotación original. Esto permite alinear fácilmente los distintos contextos según el gen de interés. En caso de diagramar los resultados de una búsqueda tipo BLAST, los homólogos son marcados como gen de interés y determinan la abscisa cero del gráfico, como se aprecia en la figura 48. En caso contrario, se utiliza el primer gen de cada contexto como gen de referencia, causando que estos queden alineados a la izquierda, como se muestra en la figura 49.

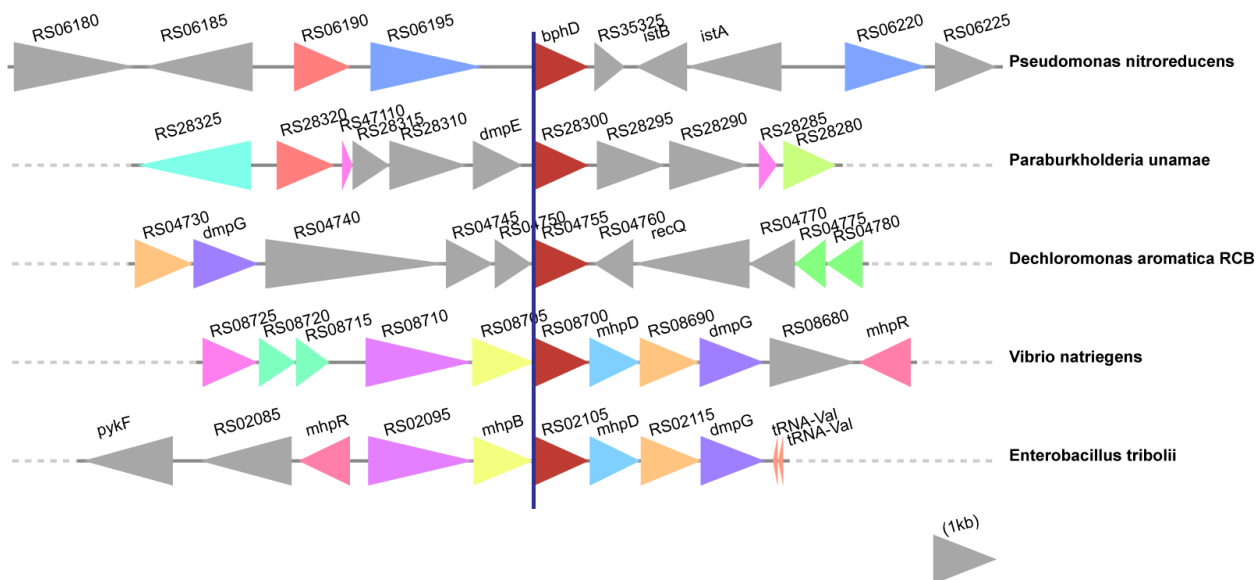


**Figura 48:** Alineamiento automático de contextos con respecto a los homólogos. En azul, la coordenada cero del lienzo. Nótese que el contexto de *Enterobacillus tribolii* tiene su inicio y término a la izquierda que los de *Vibrio natriegens*, como consecuencia de mantener el alineamiento de los genes de interés.

**Fuente:** Elaboración propia

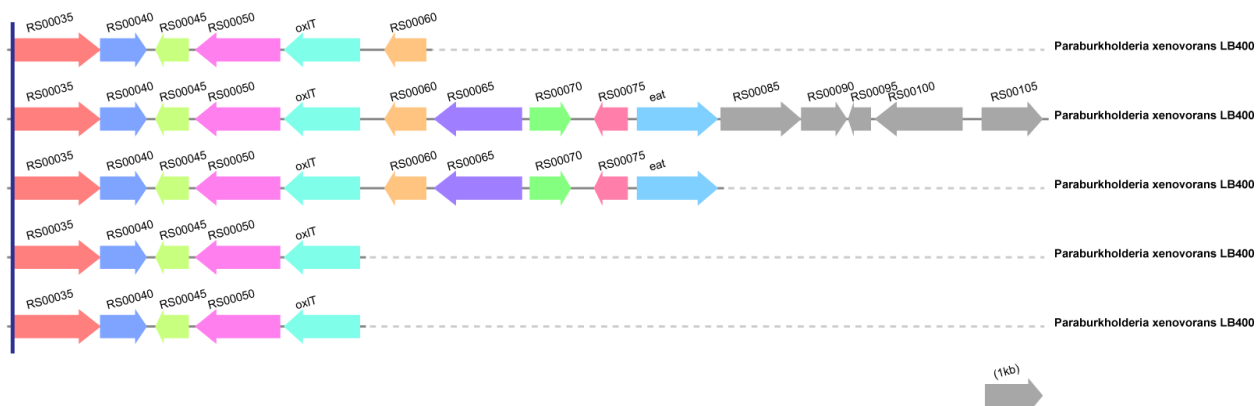


**Figura 49:** Alineamiento de contextos con respecto a su gen de inicio al graficar distintas regiones del genoma de *Paraburkholderia Xenovorans*. En azul, la coordenada cero del lienzo. Nótese que las regiones excluidas por el usuario de la región a graficar se representan con línea discontinua.

**Fuente:** Elaboración propia

El alto de cada flecha se calcula en base a un largo esperado de los genes, equivalente al largo total de la región graficada partido en la cantidad de genes a representar. La parte más gruesa de cada flecha tiene un alto igual a la mitad del largo esperado, y es equivalente al espacio entre un contexto y otro, como se muestra en la figura 50. El largo esperado del gen se calcula en base al primer contexto y sin considerar al resto, para mayor eficiencia computacional.

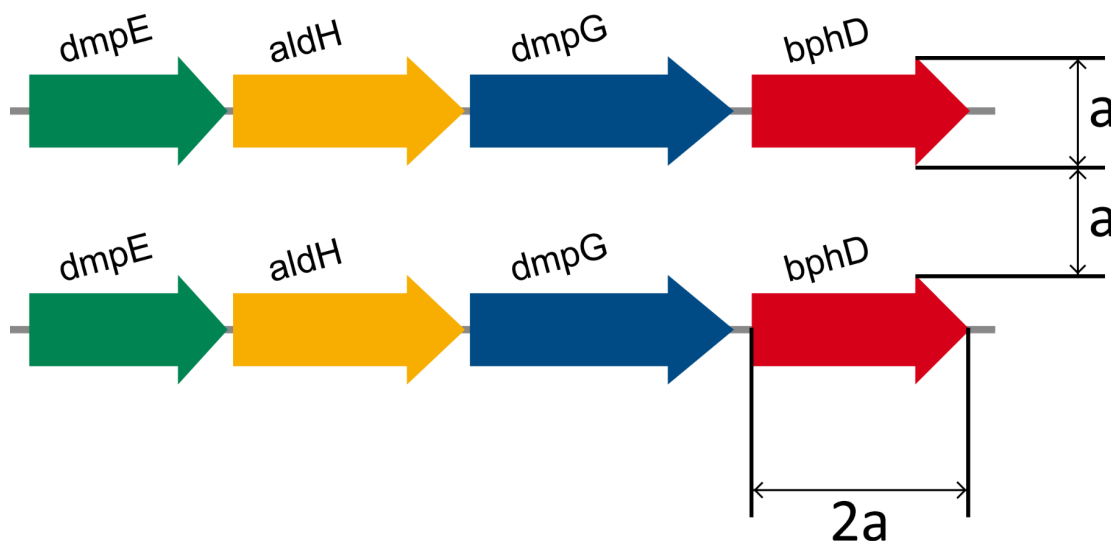


**Figura 50:** Espaciado entre los contextos graficados.

**Fuente:** Elaboración propia

La parte angosta de las flechas se relaciona con la parte más gruesa según la proporción áurea, como se muestra en la figura 51. Se ha tomado esta decisión por la creencia sostenida desde la antigüedad de que esta proporción es más atractiva a la vista, por lo que ha sido usada en el

diseño de diversas obras del ámbito de la pintura, escultura, arquitectura, música y diseño gráfico, entre otros [Llasera, J. (2021)].

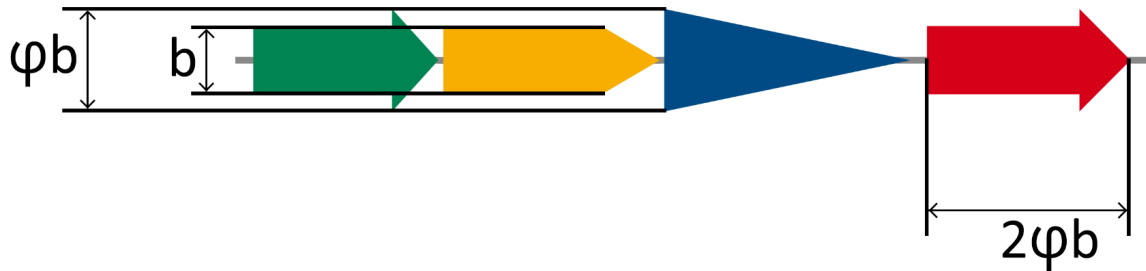


**Figura 51:** Proporciones internas de las representaciones de los genes, en sus formas alternativas.

**Fuente:** Elaboración propia

El texto de las etiquetas de cada gen se asigna, en primer lugar, según la naturaleza del gen. En caso de que sintetice ARN de transferencia, se le asigna el nombre de su producto, por ejemplo "tRNA-Ala". En el caso de sintetizar ARN ribosómico, se le asigna un nombre según su clasificación, por ejemplo "5S RNA". En otro caso, se le asigna el nombre del gen, como puede ser "*bphD*". En caso de no hallarse cualquiera de estos datos, se utiliza el *locus tag*.

## 5. Validación de la solución

Para validar que la aplicación aquí presentada resuelve el problema inicialmente planteado de manera satisfactoria, se han utilizado dos enfoques: En primer lugar, una evaluación estándar realizada por el mismo equipo de desarrollo, y en segundo lugar una medición desde la perspectiva de los usuarios. Ambos han entregado resultados satisfactorios y han permitido identificar oportunidades de mejora.

### 5.1. Validación experta

La interfaz se ha evaluado de forma preliminar con los 247 lineamientos de usabilidad web propuestos por el doctor David Travis [Travis, D. (2016)], diseñados para promover entre los desarrolladores web en general la consistencia y buenas prácticas en el diseño de interfaces. Esto consiste en una serie de preguntas objetivas acerca de la interfaz a las que se les asigna un puntaje del -1 al 1, donde -1 indica que no se cumple con el lineamiento, 1 indica que sí se cumple y 0 indica que se cumple parcialmente.

De los 247 lineamientos, se ha considerado que solo 200 aplican a esta interfaz, debido a que muchos están enfocados a páginas de contenido (como YouTube o Wikipedia) y a páginas de comercio electrónico (como Ripley o Amazon). De los que aplican, se ha determinado que la interfaz desarrollada cumple con el 81% en general.

| Summary of results               |           |             |           |       |
|----------------------------------|-----------|-------------|-----------|-------|
|                                  | Raw score | # Questions | # Answers | Score |
| Home Page                        | 13        | 20          | 17        | 88%   |
| Task Orientation                 | 20        | 44          | 39        | 76%   |
| Navigation & IA                  | 10        | 29          | 21        | 74%   |
| Forms & Data Entry               | 15        | 23          | 19        | 89%   |
| Trust & Credibility              | 3         | 13          | 11        | 64%   |
| Writing & Content Quality        | 16        | 23          | 22        | 86%   |
| Page Layout & Visual Design      | 33        | 38          | 37        | 95%   |
| Search                           | 0         | 20          | 0         |       |
| Help, Feedback & Error Tolerance | 20        | 37          | 34        | 79%   |
| Overall score                    |           | 247         | 200       | 81%   |

**Figura 52:** Resultados de la evaluación según lineamientos de usabilidad web.

**Fuente:** Elaboración propia

Si bien algunos de los lineamientos incumplidos no son posibles de cumplir debido a la naturaleza de la aplicación, este ejercicio sí ha permitido poner la atención sobre potencialidades de mejora en el sistema.

Las principales falencias detectadas con este ejercicio se incluyen en el apartado de “confianza y credibilidad”, y se deben a la carencia de un apoyo institucional por alguna entidad reconocida, y la inexistencia de una imagen de marca que permita identificar a la aplicación.

Una parte importante del puntaje restante se debe a la linealidad de las tareas, sin mostrar resultados preliminares y forzando al usuario a comprometerse con la información ingresada. Esto no es posible de mejorar en el flujo de datos que incluye una búsqueda de homólogos debido a la naturaleza del trabajo realizado, donde resultados intermedios no tienen sentido, pero sí en el flujo alternativo, por ejemplo, cargando el diagrama a medida que el usuario ingresa los genomas a graficar. Esto no se ha hecho porque el esfuerzo requerido para realizar esta modificación supera el disponible para este trabajo, aunque no se descarta para futuras intervenciones a la aplicación.

En general, podemos decir que los resultados de esta evaluación son muy positivos, aunque hay mucho espacio para mejorar.

## 5.2. Validación con usuarios

Para validar la aplicación desarrollada se ha utilizado el cuestionario AttrakDiff propuesto por Hassenzahl et al., desarrollado para medir de manera independiente la calidad pragmática y hedónica de un producto, donde la primera se relaciona con la facilidad y eficiencia con que el usuario es capaz de completar la tarea por la cual utiliza el *software* y la segunda con “dimensiones de calidad que no tienen una obvia relación con la tarea que el usuario desea completar con el sistema, como por ejemplo originalidad, innovación, belleza, etc. Aunque no estén directamente vinculados a la tarea, el usuario los considera aspectos importantes de calidad, por su propio bienestar.” [Hassenzahl et al (2000)].

Este cuestionario fue contestado por doce investigadores, profesionales y estudiantes de doctorado de áreas biológicas que necesitan realizar análisis de contextos genómicos como parte de su trabajo, después de utilizar la herramienta para completar una serie de tareas.

Todos los participantes lograron completar las tareas asignadas, y los resultados del cuestionario fueron muy positivos, como se puede apreciar en la figura 53. Aquí se muestra un punto central correspondiente al valor esperado calculado a partir de las respuestas de los distintos usuarios y un rectángulo más amplio mostrando el intervalo de confianza. Haber obtenido un rectángulo de confianza relativamente pequeño es considerado una ventaja porque implica mayor coincidencia entre los distintos usuarios, lo que nos muestra que estos resultados son más bien confiables e independientes del usuario y sus circunstancias.

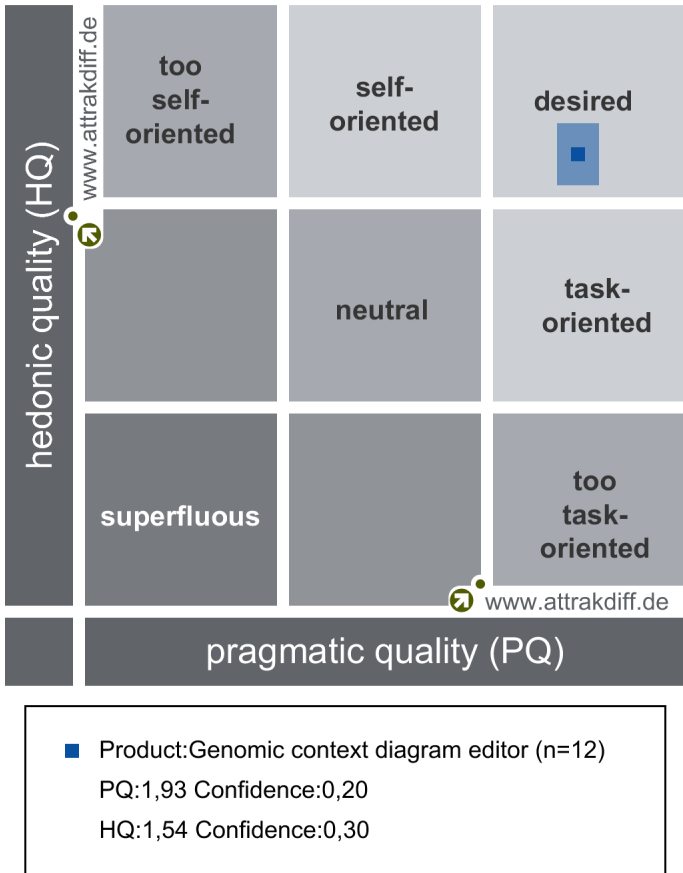


**Figura 53:** Resumen de los resultados del estudio realizado. El eje vertical indica la calidad hedónica, mientras que el horizontal muestra la calidad pragmática, ambos en escala del -3 al 3.

**Fuente:** AttrakDiff Result overview

Este cuestionario entrega un resultado en escala del -3 al 3 para cada dimensión, donde los valores mayores a 1 se consideran “deseables”. La aplicación ha obtenido un puntaje en el intervalo de confianza [1,73; 2,13] en su calidad pragmática y [1,24; 1,84] para su calidad hedónica. Esto nos permite afirmar que los usuarios encuestados **ciertamente perciben una calidad deseable en la presente interfaz.**

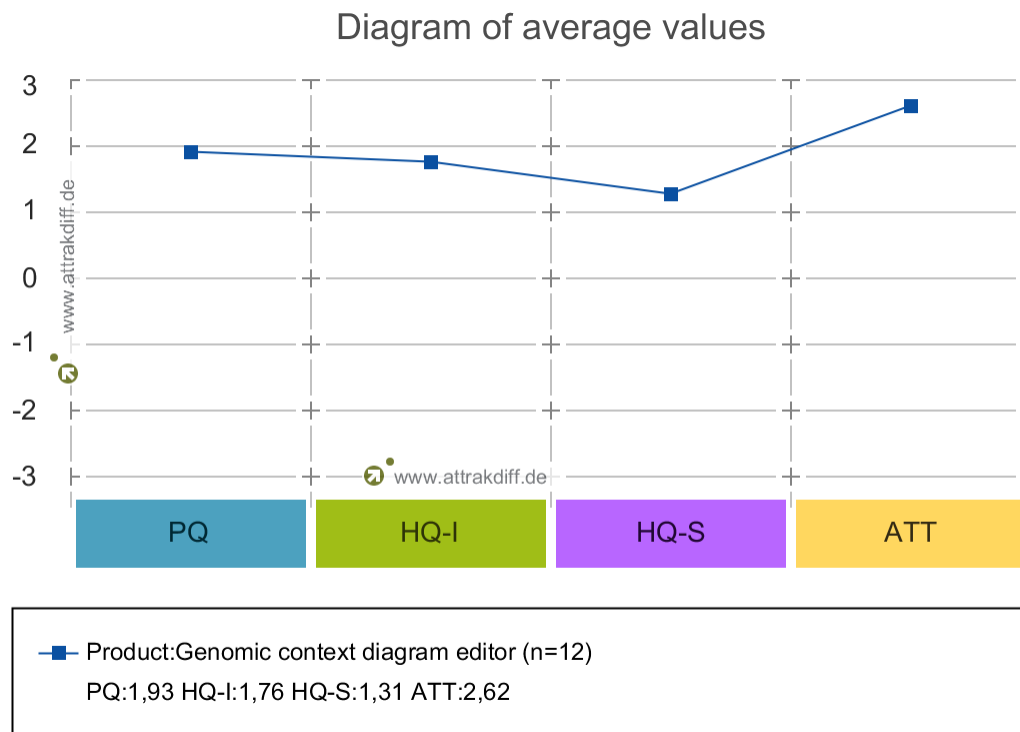


**Figura 54:** Resumen de los resultados del estudio realizado, por dimensión. PQ: *Pragmatic Quality* (calidad pragmática), HQ-I: *Hedonic Quality - Identity* (calidad hedónica - identidad), HQ-S: *Hedonic Quality - Stimulation* (calidad hedónica - estimulación), ATT: *attractiveness* (atractivo).

**Fuente:** AttrakDiff Result overview

Al ver los resultados en un grado algo mayor de detalle (figura 54), podemos notar que el puntaje más bajo se ha obtenido en el área de estimulación, lo cual es predecible al tratarse de una herramienta de trabajo profesional.

En muchos casos la estimulación se considera contraria a la calidad pragmática, pues llamar la atención del usuario requiere mostrarle elementos novedosos y en cierto grado misteriosos. Despertar la curiosidad del usuario implica generar elementos impredecibles en su comportamiento, lo cual reduce la facilidad de uso de la aplicación, lo cual no sería tan deseable en este caso, dado que el fin de la interfaz es facilitar un flujo de trabajo más que generar un periodo de esparcimiento para el usuario.

Esto se confirma al observar en la figura 55, que muestra los puntajes individuales por pregunta, que el puntaje más bajo de esta área corresponde al eje “sencillo - desafiante” (“undemanding - challenging”), lo que ratifica que se ha priorizado la facilidad de uso antes que la “dicha de uso”, por considerarse la primera una parte esencial de una herramienta de trabajo.

El segundo puntaje más bajo es el de la identidad, lo que reafirma las conclusiones de la revisión experta llevada a cabo con anterioridad, de que la carencia de una imagen de marca definida y un apoyo institucional afecta la usabilidad del producto.

De la figura 55, que detalla los puntajes individuales en cada pregunta del cuestionario, podemos ver que el puntaje más bajo, que indica que la aplicación es “muy técnica y poco humana” está íntimamente relacionado con la naturaleza de la aplicación como herramienta de trabajo en un área específica de la ciencia y que, como tal, debe contener términos técnicos propios del área.

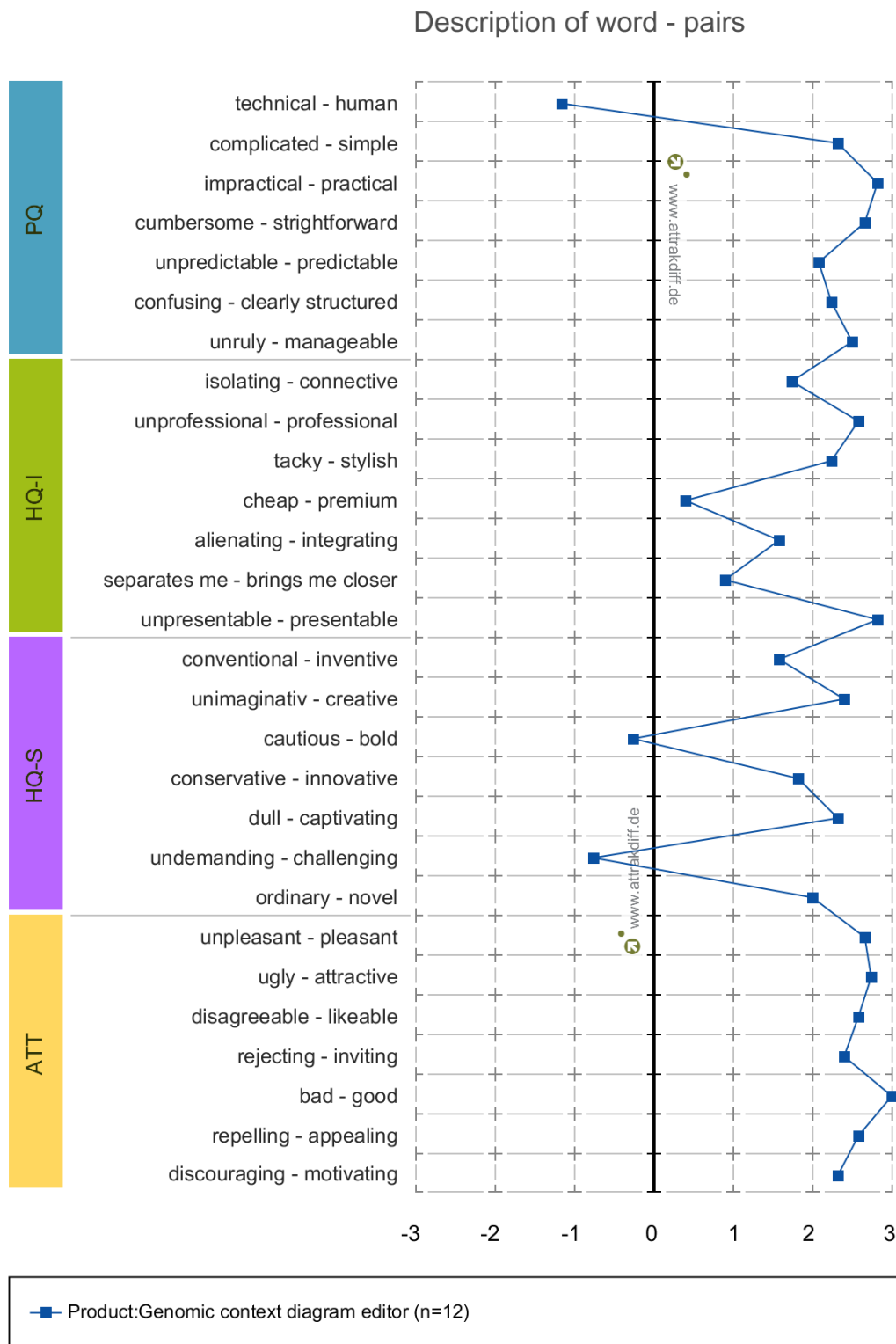


**Figura 55:** Resultados del estudio realizado, por pregunta.  
**Fuente:** AttrakDiff Result overview

Los resultados de este estudio no son estadísticamente representativos del público objetivo por la cantidad de usuarios encuestados y el sesgo en su selección, además de que todos los sujetos tenían una experiencia muy limitada utilizando la aplicación. Sin embargo, una valoración tan positiva valida los esfuerzos realizados al confirmar la pertinencia de este sistema para los fines a los que fue diseñado.

El detalle de los resultados se halla en el anexo.

## Conclusiones

La visualización y comparación de contextos genómicos es un paso esencial en la genómica comparativa y, a pesar de las herramientas que se han desarrollado para facilitar este proceso, sigue siendo un paso costoso y tedioso para los estudiosos del área, significando numerosas horas de trabajo calificado para un proceso fácilmente automatizable.

Se ha construido una interfaz de usuario para la visualización de bases de datos genómicos y otros datos provistos por el usuario para facilitar este tipo de análisis, específicamente a través del dibujo de diagramas de contexto genómico, instrumento muy utilizado en estos estudios.

Para ello, se han determinado las principales tareas de usuario que se requieren para la visualización de datos genómicos necesarios para la identificación de contextos genómicos de genomas bacterianos, a través de entrevistas con investigadores del área y otras técnicas de identificación de problemas.

Se ha realizado un estudio del estado del arte en cuanto a las herramientas bioinformáticas que sirven al análisis de contexto genómico, identificando fortalezas y falencias, que ha confirmado la necesidad de desarrollar un sistema fácil de usar, con productos altamente personalizables e integración con otras herramientas preexistentes.

Se ha desarrollado una interfaz funcional que permite a los usuarios realizar las tareas previamente identificadas, a través de un desarrollo iterativo y con múltiples instancias de retroalimentación por parte de los potenciales usuarios.

Finalmente, se ha validado el sistema con los usuarios, a través de entrevistas y un cuestionario estándar de facilidad de uso, afirmando su calidad tanto pragmática como hedónica, lo que significa que dicho sistema no sólo permite realizar las tareas esperadas, sino que además es agradable de usar. La aplicación está disponible para su uso en la dirección web <http://200.1.19.244:3030>.

Esta experiencia saca a relucir la necesidad de una mayor integración multidisciplinaria para enfrentar los distintos desafíos que enfrenta la biología contemporánea, que se refleja no sólo en los análisis comparativos de contexto genómico, sino también en otros tipos de análisis, que históricamente han sido atacados principalmente desde el área biológica y bioinformática, cuando una mayor integración de expertos de otras áreas como el diseño y las ciencias de la computación tiene el potencial de brindar grandes beneficios a bajo coste. Incluso en este proyecto en particular, si bien el trabajo finalmente fue largo, los primeros prototipos ya servían para ahorrar semanas de trabajo de un microbiólogo. Como dice el principio de Pareto, el 20% del trabajo resuelve el 80% de los problemas [Kiremire, A. (2011)].

Dado que el desarrollo de este tipo de sistemas requiere de un alto manejo de conocimiento biológico, es impensable su realización sin la participación de los usuarios, lo cual confirma la necesidad de un enfoque de diseño centrado en usuarios como el utilizado en este trabajo.

### **Trabajo futuro**

Si bien la aplicación está en un estado de desarrollo que permite su uso, es posible seguir desarrollándola ampliamente, para su mayor perfeccionamiento en varios aspectos:

### **Eficiencia y confiabilidad**

La actual implementación del sistema depende de una conexión permanente con los datos del NCBI, y por lo tanto no puede funcionar en caso de que este deje de hacerlo. Además, cada solicitud de un usuario puede implicar la descarga de varios Megabytes de datos, lo cual genera un retraso no despreciable en la respuesta y está determinado por el ancho de banda de la red de acceso del servidor.

Esta dificultad es posible sortearla simplemente almacenando todos los genomas secuenciados y anotados hasta la fecha y actualizándolos periódicamente, lo cual no es posible en este momento debido a que el disco utilizado no tiene el suficiente espacio para almacenar TeraBytes de datos.

### **Escalabilidad**

Hay mucho trabajo por realizar en esta área, para enfrentar las múltiples solicitudes a recibir, empezando por estimar el volumen de estas. Cuantificar el público objetivo de esta aplicación y por lo tanto la capacidad de respuesta que requiere el servidor se sale del alcance de este trabajo, pero es necesario para su adecuado provecho por parte de la comunidad científica global.

Según los resultados de dicho estudio, y considerando que el trabajo realizado por el servidor tiene un alto uso de procesador, para servir múltiples solicitudes en paralelo de forma eficiente se haría necesario utilizar una máquina con más núcleos de procesamiento o incluso múltiples servidores distribuidos físicamente, con el apoyo de algún sistema de balanceo de carga. Dado que las opciones son múltiples, un estudio apropiado debe determinar la conveniencia de alguna de ellas para su implementación.

### **Usabilidad**

En este trabajo se ha presentado una validación de la facilidad de uso de la interfaz desarrollado, pero también se han identificado importantes oportunidades de mejora, incluyendo:

- El desarrollo de una imagen de marca que les brinde a los usuarios una mayor seguridad y exprese confiabilidad y profesionalismo, haciéndoles sentir un respaldo institucional serio y robusto.
- Una mayor retroalimentación al usuario de los procesos que son llevados a cabo en el servidor, lo cual requiere de un mayor intercambio de datos entre el servidor y el cliente, para que este último pueda sentir mayor seguridad al usar la aplicación, especialmente en los periodos de espera, mientras se realiza la búsqueda de homólogos, filtro de resultados y análisis de genomas anotados.
- Una mayor flexibilidad en las formas de introducir datos, como la posibilidad de subir varios archivos simultáneamente, la selección de varias regiones a graficar dentro del mismo genoma, o la posibilidad de modificar las regiones a graficar, ya sea agregando o quitando o ampliando o reduciendo las ya presentes, mientras se visualizan las regiones ya especificadas. Cualquiera de estas modificaciones requiere un importante proceso de reingeniería de la interfaz presentada.
- Un análisis exhaustivo de las distintas herramientas de anotación de genomas, para asegurar la máxima compatibilidad con cada una de ellas, dentro de lo posible, considerando que esta aplicación ha sido desarrollada con enfoque en los instrumentos utilizados entre los integrantes del LabMMBA, pero existen otros métodos de anotación que no han sido considerados para el presente trabajo.
- Una mayor flexibilidad y completitud en las formas de visualizar datos, para presentar el máximo de información posible a los usuarios facilitando aún más el análisis de contexto genómico.
- La realización de un estudio de alternativas en la forma de mostrar las propiedades editables del diagrama, considerando que la forma de presentación puede no ser tan obvia para algunos usuarios, como se observó en las pruebas de *think-aloud testing*.
- Analizar los formatos de imagen que podrían ofrecerse para descargar el diagrama resultante, y la mejor manera de ofrecerlos en la interfaz, para facilitar su manejo en distintos casos de uso.

## Funcionalidad

Considerando la especificidad de las tareas de usuario consideradas para el desarrollo de este sistema en contraste a la diversidad de análisis que es posible realizar en el ámbito de la genómica comparativa, una natural evolución de esta aplicación pasa por estudiar las posibilidades de ampliar los casos de uso, desde cosas tan sencillas como, para graficar en una sola fila secuencias de una misma especie que estén en distintos cromosomas o *contigs*, hasta la integración con otro tipo de visualizaciones, como por ejemplo, la integración de árboles filogenéticos generados por iTOL (o la herramienta que corresponda) en la misma figura, o articular con visualizaciones de genoma completo con diseño circular.

## Referencias bibliográficas

1. Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D. (1990). "Basic Local Alignment Search Tool" *Journal of molecular biology*. 215. 403-10.  
doi: 10.1016/S0022-2836(05)80360-2
2. Aziz, R., Bartels, D., Best, A., DeJongh, M., Disz, T., Edwards, R., Formsma, K., Gerdes, S., Glass, E., Kubal, M., Meyer, F., Olsen, G., Olson, R., Osterman, A., Overbeek, R., McNeil, L., Paarmann, D., Paczian, T., Parrello, B., Pusch, G., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., Zagnitko, O. (2008) "The RAST Server: Rapid Annotations using Subsystems Technology" *BMC Genomics*. 2008; 9: 75. Publicado el 8 de febrero de 2008.  
doi: 10.1186/1471-2164-9-75
3. Bilotta, M., Tradigo, G., Veltri, P. (2019) "Bioinformatics Data Models, Representation and Storage" *Encyclopedia of Bioinformatics and Computational Biology* pp. 110-116.  
ISBN: 9780128114322 doi: 10.1016/B978-0-12-809633-8.20410-X
4. Brathovde, K., Farner, M., Brun, F., Sandnes, F. (2019) "Effectiveness of Color-Picking Interfaces Among Non-designers" *Cooperative Design, Visualization, and Engineering* pp. 181-189, Springer International Publishing. doi: 10.1007/978-3-030-30949-7\_21
5. Buchfink, B., Reuter, K., Drost, H. G., (2021) "Sensitive protein alignments at tree-of-life scale using DIAMOND", *Nature Methods* 18, 366–368. doi:10.1038/s41592-021-01101-x
6. Buchfink, B., Xie, C., Huson, D. H., (2015) "Fast and sensitive protein alignment using DIAMOND", *Nature Methods* 12, 59-60. doi:10.1038/nmeth.3176
7. Buchfink, B. (2020) "diamond slow, compared to blastp?" GitHub issue thread #413.  
Extraído de: <https://github.com/bbuchfink/diamond/issues/413>
8. Cantalapiedra, C., Hernandez-Plaza, A., Letunic, I., Bork, P., Huerta-Cepas, J. (2021) "eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale" *bioRxiv* 2021.06.03.446934  
doi: 10.1101/2021.06.03.446934
9. CDC: Centers for Disease Control and Prevention (2016) "Pulsed-field Gel Electrophoresis (PFGE)", extraído de: <https://www.cdc.gov/pulsenet/pathogens/pfge.html>
10. CDC: Centers for Disease Control and Prevention (2016) "Whole Genome Sequencing (WGS)", extraído de: <https://www.cdc.gov/pulsenet/pathogens/wgs.html>
11. Chaniotis, I., Kyriakou, K., Tselikas, N. (2015) "Is Node.js a viable option for building modern web applications? A performance evaluation study" *Computing* 97, 1023–1044, Springer-Verlag Wien. doi: 10.1007/s00607-014-0394-9

12. Dahm, R. (2005) "Friedrich Miescher and the discovery of DNA" *Developmental Biology*, Volume 278, Issue 2, pp. 274-288, ISSN 0012-1606 doi: 10.1016/j.ydbio.2004.11.028
13. Dale, J., von Schantz, M. (2002) "From genes to genomes: Concepts and applications of DNA technology", consultado de:  
[http://www.aun.edu.eg/molecular\\_biology/Procedure1/From%20Genes%20to%20Genomes%20-%20Concepts%20and%20Applications%20of%20DNA%20Tec.pdf](http://www.aun.edu.eg/molecular_biology/Procedure1/From%20Genes%20to%20Genomes%20-%20Concepts%20and%20Applications%20of%20DNA%20Tec.pdf)
14. Darling, A.C.E., Mau, B., Blattner, F.R., Perna, N.T. (2004) "Mauve: multiple alignment of conserved genomic sequence with rearrangements" *Genome research*  
<https://genome.cshlp.org/content/14/7/1394.full.pdf>
15. Durán, R.E., Alviz-Gazitua, P., Saona-Urmeneta, V., Buil, C., Seeger, M. (2021). "Genomic and functional assessment of the aromatic catabolic potential of the model heavy metal resistant bacterium *Cupriavidus metallidurans* CH34" Presentación de Panel. XXV Congreso Latinoamericano de Microbiología (ALAM 2021). 25 al 28 de Agosto del 2021. Asunción - Paraguay.
16. Garcia, P., Jauffrit, F., Grangeasse, C., Brochier-Armanet, C. (2019) "GeneSpy, a user-friendly and flexible genomic context visualizer" *Bioinformatics*, Volume 35, Issue 2, 15 January 2019, Pages 329–331 doi: 10.1093/bioinformatics/bty459
17. González-Rivera, E., Durán, R.E., Seeger, M. (2021) "Genetic determinants for Bacterial Heavy Metal resistance (HMR): An open-source and manually curated dataset for genomic and metagenomic analyses" Presentación de Panel. XXV Congreso Latinoamericano de Microbiología (ALAM 2021). 25 al 28 de Agosto del 2021. Asunción - Paraguay.
18. Google (2021) "Introduction to Material Design", consultado de:  
<https://material.io/design/introduction>
19. Guy, L., Kultima J., Andersson, S. (2010) "genoPlotR: comparative gene and genome visualization in R", *Bioinformatics*, Volume 26, Issue 18, 15 September 2010, Pages 2334–2335, doi: 10.1093/bioinformatics/btq413
20. Harrison, K., De Crécy-Lagard, V., Zallot, R. (2018) "Gene Graphics: a genomic neighborhood data visualization web application" *Bioinformatics*, Volume 34, Issue 8, 15 April 2018, pp. 1406-1408, doi: 10.1093/bioinformatics/btx793
21. Hassenzahl, M., Platz, A., Burmester, M., Lehner, K. (2000) "Hedonic and ergonomic quality aspect determine a software's appeal" Conference: Proceedings of the CHI 2000 Conference on Human factors in computing systems, The Hague, The Netherlands  
doi: 10.1145/332040.332432

22. Instituto Universitario de Conservación y Mejora de la Agrodiversidad Valenciana, Universitat Politècnica de València (2021) “Secuenciación de Sanger”, extraído de: [https://bioinf.comav.upv.es/courses/intro\\_bioinf/sanger.html](https://bioinf.comav.upv.es/courses/intro_bioinf/sanger.html)
23. Kiremire, A (2011) “the application of the Pareto principle in software engineering”, extraído de: [http://www2.latech.edu/~box/ase/papers2011/Ankunda\\_termpaper.pdf](http://www2.latech.edu/~box/ase/papers2011/Ankunda_termpaper.pdf)
24. Letunic I., Bork P. (2016) “Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees”. *Nucleic Acids Res.*, 44, W242–W245. doi: 10.1093/nar/gkw290
25. Llasera, J. (2021) “La proporción áurea: Qué es y cómo se aplica en diseño gráfico” Blog, 8 de marzo de 2021, Imborrable. Extraído de: <https://imborrable.com/blog/proporcion-aurea/>
26. LUMA Institute (2021) “Innovating for People: Handbook of Human-Centered Design Methods” LUMA Institute ISBN: 978-0-9857509-0-9
27. Madden, T. (2002). “The BLAST sequence analysis tool” *The NCBI Handbook*.
28. Martinez-Guerrero, C.E., Ciria, R., Abreu-Goodger, C., Moreno-Hagelsieb, G., Merino, E. (2008) “GeConT 2: gene context analysis for orthologous proteins, conserved domains and metabolic pathways” *Nucleic Acids Res.* 2008 Jul 1;36(Web Server issue):W176-80. doi: 10.1093/nar/gkn330
29. Muffato, M., Louis, A., Poinsel, C.E., Roest Crolius, H. (2010) “Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes” *Bioinformatics*, Volume 26, Issue 8, 15 April 2010, Pages 1119-1121. doi: 10.1093/bioinformatics/btq079
30. McClintock, B. (1930) “A cytological demonstration of the location of an interchange between two non-homologous chromosomes of *zea mays*” *Department of Botany, New York State College of Agriculture, Genetics*, vol. 16, pp. 791-196, extraído de: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC544331/pdf/pnas01013-0013.pdf>
31. McKilligan, S., Creeger, S. (2018) “Strategies to redefine the problem exploration space for design innovation” *Creativity and Innovation in Design and Engineering Education*, pp. 318-325. ISBN: 978-1-912254-02-6, Recuperado de: <https://www.designsociety.org/publication/40782/STRATEGIES+TO+REDEFINE+THE+PROBLEM+EXPLORATION+SPACE+FOR+DESIGN+INNOVATION>
32. NCBI (2021) “About RefSeq”, extraído de <https://www.ncbi.nlm.nih.gov/refseq/about/>
33. NCBI (2020) “NCBI RefSeq Select”, extraído de [https://www.ncbi.nlm.nih.gov/refseq/refseq\\_select/](https://www.ncbi.nlm.nih.gov/refseq/refseq_select/)

34. NCBI (2021) "GenBank and WGS Statistics", extraído de <https://www.ncbi.nlm.nih.gov/genbank/statistics/>
35. Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E., Zaslavsky, L., Lomsadze, A., Pruitt, K., Borodovsky, M., Ostell, J. (2016) "NCBI prokaryotic genome annotation pipeline" *Nucleic Acids Res.* 2016 Aug 19;44(14):6614-24. doi: 10.1093/nar/gkw569
36. Travis, D. (2016) "247 usability guidelines" Userfocus. Extraído de: <https://www.userfocus.co.uk/resources/guidelines.html>
37. Oberto, J. (2013) "SynTax: a web server linking synteny to prokaryotic taxonomy" *BMC Bioinformatics* 14, 4. doi: 10.1186/1471-2105-14-4
38. Odvárko, J. (2021) "jscolor: JavaScript Color picker with Opacity (alpha channel)", extraído de: <https://jscolor.com/>
39. Overmars, L., Kerkhoven, R., Siezen, R.J., Francke, C. (2013) "MGcV: the microbial genomic context viewer for comparative genome analysis" *BMC Genomics* 14, 209. doi: 10.1186/1471-2164-14-209.
40. Overmars, L., Van Hijum, S., Siezen, R., Francke, C. (2015) "CiVi: circular genome visualization with unique features to analyze sequence elements" *Bioinformatics*, Volume 31, Issue 17, 1 September 2015, Pages 2867–2869, doi: 10.1093/bioinformatics/btv249
41. Pearson, T. (2020) "Abstraction Laddering – An Easy, Effective Problem Framing Exercise", extraído de: <https://spin.atomicobject.com/2020/07/05/problem-framing-abstraction-ladder/>
42. Pearson, W. (2013) "An Introduction to Sequence Similarity ('Homology') Searching" *Current protocols in bioinformatics* vol. Chapter 3 (2013): Unit3.1. doi: 10.1002/0471250953.bi0301s42
43. Piérola, M. (2020) "Identificación de usuarios para software de visualización bioinformático" Trabajo realizado para la asignatura de Interfaces de Usuario con el profesor Carlos Buil, extraído de: <https://drive.google.com/file/d/136VOWH3GZoUatgrW7lrWa4h-56xJCwiw/view?usp=sharing>
44. Preiss, B. (2001) "Data Structures and Algorithms with Object-Oriented Design Patterns in C++" ed. John Wiley & Sons Inc. ISBN-13: 978-0471241348

45. Pertsemlidis, A., Fondon, J. (2001). "Having a BLAST with bioinformatics (and avoiding BLASTphemy)". *Genome biology*. 2. REVIEWS2002. doi: 10.1186/gb-2001-2-10-reviews2002
46. Saha, C. K., Pires, R. S., Brolin, H., Delannoy, M., Atkinson, G. C. (2020) "FlaGs and webFlaGs: discovering novel biology through the analysis of gene neighbourhood conservation", *Bioinformatics*, btaa788. doi: 10.1093/bioinformatics/btaa788
47. Seemann, T. (2014) "Prokka: rapid prokaryotic genome annotation" *Bioinformatics* 2014 Jul 15;30(14):2068-9 Oxford University Press. doi: 10.1093/bioinformatics/btu153
48. Segerman, B. (2020) "The Most Frequently Used Sequencing Technologies and Assembly Methods in Different Time Segments of the Bacterial Surveillance and RefSeq Genome Databases", *Frontiers in cellular and infection microbiology* vol. 10, 527102. 19 Oct. 2020, doi:10.3389/fcimb.2020.527102
49. Shen, W., Ren, H. (2021) "TaxonKit: A practical and efficient NCBI taxonomy toolkit" *Journal of Genetics and Genomics*, ISSN 1673-8527, doi: 10.1016/j.jgg.2021.03.006
50. Stothard, P., Wishart, D. (2005) "Circular genome visualization and exploration using CGView" *Bioinformatics*, Volume 21, Issue 4, 15 February 2005, Pages 537–539, doi: 10.1093/bioinformatics/bti054
51. UniProt Consortium (2021) "UniProtKB", extraído de: <https://www.uniprot.org/help/uniprotkb>
52. Van Dierendonck, R., van Tienhoven, S., Elid, T. (2015) "D3. JS: Data-Driven Documents" Extraído de: [https://staas.home.xs4all.nl/t/swtr/documents/wt2015\\_d3\\_js.pdf](https://staas.home.xs4all.nl/t/swtr/documents/wt2015_d3_js.pdf)
53. Villee, C. (1996) "Biología" octava edición McGraw-Hill Interamericana Editores ISBN 0-7216-9023-8
54. Wolf, Y., Rogozin, I., Kondrashov, A., Koonin, E. (2001) "Genome Alignment, Evolution of Prokaryotic Genome Organization, and Prediction of Gene Function Using Genomic Context" *Genome Res.* 2001. 11: 356-372 Cold Spring Harbor Laboratory Press doi: 10.1101/gr.161901
55. Zulko, Vegh, P., Bloom, J., Roy, L. (2020) "DNA Features Viewer" Repositorio en GitHub. Consultado de: <https://github.com/Edinburgh-Genome-Foundry/DnaFeaturesViewer#dna-features-viewer>

### Otra bibliografía relacionada

56. Kulmanov M., Hoehndorf R. (2020) "DeepGOPlus: improved protein function prediction from sequence" *Bioinformatics*, 36 (2): 422–429. doi: 10.1093/bioinformatics/btz595.
57. Nakano, F. K., Lietaert, M., Vens, C. (2019) "Machine learning for discovering missing or wrong protein function annotations : A comparison using updated benchmark datasets" *BMC bioinformatics*, 20(1), 485. doi: 10.1186/s12859-019-3060-6.
58. Sureyya Rifaioğlu, A., Doğan, T., Jesus Martin, M., Cetin-Atalay, R., Atalay, V. (2019). "DEEPred: Automated Protein Function Prediction with Multi-task Feed-forward Deep Neural Networks" *Scientific reports*, 9(1), 7344. doi: 10.1038/s41598-019-43708-3
59. Rohm, U., Diep, T. (2006) "How to BLAST your database - A Study of Stored Procedures for BLAST Searches", consultado desde:  
<https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.1474&rep=rep1&type=pdf>

## Anexo: Resultados de la encuesta a usuarios

Para validar la aplicación se le solicitó a 12 potenciales usuarios probar y evaluar la aplicación con el cuestionario AttrakDiff. Todos los sujetos están en el rango de edad entre 20 y 40 años, a excepción de uno que está en el rango entre 40 y 60. Nueve son mujeres, tres son hombres y todos son profesionales del área de la biología con estudios de grado y/o postgrado, que declararon como profesión alguna de las siguientes: Bioquímico, Ingeniero Agrónomo, Ingeniero en Biotecnología, Estudiante, Biólogo, Ingeniero Civil Ambiental, Doctor en Biotecnología, Bioingeniero e Ingeniero Químico.

Los resultados de dicha encuesta se presentan a continuación, tal como han sido obtenidos de la plataforma esurvey.uid.com, en escala del 1 al 7

|                               |     |      |     |      |     |     |     |     |     |     |     |     |
|-------------------------------|-----|------|-----|------|-----|-----|-----|-----|-----|-----|-----|-----|
| Excluded from evaluation      | N   | N    | N   | N    | N   | N   | N   | N   | N   | N   | N   | N   |
| Filled out completely         | Y   | Y    | Y   | Y    | Y   | Y   | Y   | Y   | Y   | Y   | Y   | Y   |
| Editing time (sec.)           | 243 | 1130 | 469 | 7928 | 649 | 302 | 957 | 919 | 361 | 742 | 646 | 301 |
| human - technical             | 1   | 2    | 2   | 3    | 1   | 3   | 5   | 6   | 2   | 4   | 3   | 2   |
| isolating - connective        | 4   | 7    | 6   | 6    | 6   | 7   | 5   | 6   | 4   | 5   | 6   | 7   |
| pleasant - unpleasant         | 6   | 7    | 7   | 7    | 7   | 6   | 5   | 7   | 7   | 7   | 7   | 7   |
| inventive - conventional      | 6   | 7    | 7   | 2    | 6   | 7   | 4   | 6   | 7   | 4   | 6   | 5   |
| simple - complicated          | 6   | 6    | 7   | 6    | 7   | 6   | 6   | 6   | 7   | 6   | 7   | 6   |
| professional - unprofessional | 7   | 7    | 7   | 7    | 7   | 6   | 4   | 7   | 7   | 6   | 7   | 7   |
| ugly - attractive             | 7   | 7    | 7   | 7    | 6   | 7   | 6   | 7   | 7   | 6   | 7   | 7   |
| practical - impractical       | 6   | 7    | 7   | 6    | 7   | 7   | 7   | 7   | 7   | 7   | 7   | 7   |
| likeable - disagreeable       | 6   | 6    | 7   | 6    | 6   | 7   | 6   | 7   | 7   | 7   | 7   | 7   |
| cumbersome - straightforward  | 6   | 7    | 6   | 7    | 7   | 6   | 7   | 7   | 7   | 6   | 7   | 7   |
| stylish - tacky               | 6   | 6    | 7   | 6    | 6   | 7   | 6   | 6   | 7   | 6   | 6   | 6   |
| predictable - unpredictable   | 5   | 7    | 7   | 6    | 7   | 6   | 6   | 5   | 7   | 6   | 6   | 5   |

|                                                       |   |   |   |   |   |   |   |   |   |   |   |   |
|-------------------------------------------------------|---|---|---|---|---|---|---|---|---|---|---|---|
| cheap - premium                                       | 4 | 4 | 5 | 5 | 5 | 7 | 4 | 5 | 4 | 1 | 5 | 4 |
| alienating - integrating                              | 4 | 5 | 6 | 5 | 6 | 7 | 6 | 6 | 4 | 7 | 4 | 7 |
| brings me closer to people - separates me from people | 4 | 4 | 4 | 4 | 4 | 7 | 5 | 4 | 4 | 6 | 6 | 7 |
| unpresentable - presentable                           | 6 | 7 | 7 | 7 | 7 | 7 | 6 | 7 | 7 | 7 | 7 | 7 |
| rejecting - inviting                                  | 4 | 7 | 7 | 7 | 7 | 7 | 6 | 5 | 7 | 7 | 6 | 7 |
| unimaginative - creative                              | 5 | 7 | 7 | 7 | 6 | 7 | 7 | 6 | 7 | 5 | 6 | 7 |
| good - bad                                            | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 |
| confusing - clearly structured                        | 5 | 7 | 7 | 7 | 7 | 7 | 6 | 6 | 7 | 5 | 6 | 5 |
| repelling - appealing                                 | 5 | 7 | 7 | 7 | 7 | 7 | 6 | 7 | 7 | 7 | 6 | 6 |
| bold - cautious                                       | 4 | 4 | 4 | 3 | 5 | 4 | 3 | 6 | 4 | 4 | 2 | 2 |
| innovative - conservative                             | 5 | 7 | 7 | 7 | 6 | 7 | 4 | 6 | 7 | 4 | 6 | 4 |
| dull - captivating                                    | 5 | 7 | 7 | 7 | 6 | 7 | 6 | 7 | 7 | 5 | 6 | 6 |
| undemanding - challenging                             | 2 | 4 | 7 | 2 | 5 | 2 | 2 | 3 | 1 | 4 | 5 | 2 |
| motivating - discouraging                             | 5 | 5 | 7 | 7 | 7 | 7 | 6 | 6 | 7 | 6 | 6 | 7 |
| novel - ordinary                                      | 6 | 7 | 7 | 7 | 6 | 7 | 4 | 6 | 7 | 5 | 6 | 4 |
| unruly - manageable                                   | 6 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 7 | 6 | 6 | 7 |