Instituto Tecnológico de Costa Rica
Escuela de Ingeniería en Diseño Industrial

Visualización de Datos Proteómicos:
Arquitectura y representación de información de proteínas relacionadas con el cáncer de mama.

Para optar por el título de Ingeniero en Diseño Industrial con el grado académico de Bachiller Universitario

Asesor Académico:
Ph. D. Franklin Hernández-Castro

Autores:
Alfaro Arias, Verónica
Solano Román, Antonio

2014

Agradecimientos

Gracias a todos aquellos que nos tendieron una mano para poder realizar este proyecto: A Allan Orozco de Indromics por darnos libertad para desarrollar nuestras ideas. A todos los especialistas que participaron de nuestras entrevistas, por mostrarnos caminos a seguir. Gracias a Luis Ruiz y a Berny Alvarado por su asesoría técnica.

Gracias especiales a Carlos Cruz por su ayuda incondicional en los aspectos técnicos del proyecto y a los desarrolladores de 3VOT.

Finalmente, gracias al profesor Franklin Hernández por su incondicional apoyo, rigurosa metodología y constante atención.

Verdaderamente los consideramos a todos ustedes parte de nuestro equipo y estamos infinitamente agradecidos.

Dedicatoria

Dedicamos este proyecto a nuestras familias, a su apoyo incondicional a través de los años de la carrera y especialmente durante estos últimos meses de arduo trabajo.

Le dedico este proyecto también a mi abuelita Alice, quien aunque me hubiera gustado que viera el proceso, ha unido a nuestra familia por casi quince años y sigue aquí motivándonos a ser mejores.

Índice de Contenidos

Índice de gráficos:

Índice de tablas:

16 Introducción

En el presente trabajo se desarrolla el diseño de una herramienta de visualización de datos proteómicos y genómicos de proteínas relacionadas con el cáncer de mama. Se realizó un trabajo conjunto con la empresa TecApro, específicamente con la división Indromics con el fin de optar por el título de Bachiller Ingeniero en Diseño Industrial en el Tecnológico de Costa Rica.

A continuación se presenta información sobre el desarrollo de la propuesta, desde su conceptualización hasta el proceso de implementación, con el fin de generar una propuesta de visualización que contribuya a la investigación de proteínas y la lucha contra el cáncer de mama.

Palabras Clave: cáncer, dataviz, proteómica, genómica, infográfica, visualización de información.

1 Introduction

The present work is focused on the design of a tool for visualizing genomic and proteomic data of proteins related to the incidence of breast cancer. A joint work was done in collaboration with the company TecApro, specifically with the Indromics division in order to obtain the title of Bachelor Engineer in Industrial Design at the *Tecnológico de Costa Rica*.

In the following pages the reader will find information about the development of the proposal in detail; from it's conceptualization to the implementation process, with the aim of generating a visualization proposal that contributes to the field of protein research and the fight against breast cancer.

Keywords: cancer, dataviz, proteomics, genomics, infographics, informiation visualization.

16.1 Descripción de la empresa

Indromics es una empresa fundada en 2012 a partir de la diversificación de la empresa nacional TecApro. Es una empresa líder en centroamérica en el área de la Bioinformática, Biocomputación Molecular, Biología de Sistemas y Sintética. Brinda los servicios de diagnóstico diferencial de pacientes (secuenciado de ADN), mejoramiento en selección de especies y detección de trazas de transgénicos. (Indromics, 2014).

La empresa busca mejorar la forma en que los investigadores acceden e interactúan con los datos proteómicos y genómicos (de proteínas y genes), puesto que las bases de datos actuales si bien son muy completas, cuentan con una arquitectura de información poco usable, una pobre visualización de los datos y además se encuentran aisladas unas de otras.

16.2 Planteamiento del Problema

16.2.1 Planteamiento del problema:

Los datos proteómicos y genómicos existentes ven reducido su valor y funcionalidad debido a que la información no está debidamente organizada y jerarquizada.

Debido al desarrollo de herramientas de procesamiento y extracción de datos proteómicos y genómicos, la cantidad y calidad de datos va en aumento. Esto sumado al aumento en los casos de cáncer de mama revela la importancia del análisis de estructuras que interfieren en el mismo.

Las herramientas de visualización de datos genómicos y proteómicos son muy difíciles de interpretar y causan que el usuario sea propenso a cometer errores en el procesamiento de datos.

Las relaciones entre un tipo de datos y otro son poco claras y en algunos casos inexistentes, causando que los investigadores tengan que generar dichas correlaciones manualmente. La falta de automatización genera pérdida de confiabilidad en los datos obtenidos con las herramientas existentes.

La organización jerárquica entre los elementos presentes en las herramientas existentes es ineficiente y la navegación por parte del usuario es poco intuitiva, lo cual genera un problema por el hecho de que los investigadores deben invertir gran cantidad de tiempo aprendiendo a usar correctamente la visualización para no cometer errores.

Se ha invetido una enorme cantidad de recursos para que los datos estén a la disposición de los investigadores, pero al no presentar una herramienta intuitiva de visualización, el valor de los datos es reducido y su funcionalidad es pobre, dificultando el trabajo de los investigadores.

16.2.2 Definición del usuario:

Se definió un tipo de usuario que va a utilizar la herramienta de visualización que se va a diseñar. Las características del usuario, sus intereses y su dinámica de trabajo determinó los requisitos de diseño y funcionalidad de la herramienta.

Se establecieron tres grupos generales de investigadores que actualmente consultan las bases de datos proteómicas y genómicas y que tendrían la posibilidad de utilizar la herramienta de visualización.

Los tres grupos de profesionales son: Investigadores Bioquímicos, Investigadores Farmacéuticos e Investigadores Bioinformáticos.

Si bien es cierto, los tres grupos de investigadores tienen intereses específicos, es posible generar una herramienta que brinde información valiosa a todos los involucrados.

A continuación se desarrolla un análisis de las partes involucradas en el proyecto, con el fin de especificar mejor las necesidades y requerimientos del usuario.

16.2.3 Análisis de Involucrados

Se analizaron las partes que se encuentran involucradas en el proyecto con el fin de aportar soluciones integrales que se adapten a las necesidades de todas las personas y entidades:

Grupos	Intereses	Problemas percibidos	Recursos	Interés en estrategias	Conflictos potenciales
Investigadores bioquímicos	Mejor organización de los datos protéicos y genómicos Adecuada centralización de datos pertinentes	Falta de adecuada arquitectura de la información Pérdida de tiempo y recursos en la búsqueda de los datos	Recursos tecnológicos para accesar y utilizar la plataforma	Plataforma que centralice las bases de datos protéicas y genómicas	Restricción de la organización y tipos de datos disponibles
Investigadores farmacéuticos	Evidenciar la relación entre datos protéicos y fármacos correspondientes	Pérdida de la relevancia de la información Necesidad de procesar los datos previamente		Facilitar la relación entre datos y fármacos mediante la manipulación de los mismos	Falta de valor agregado debido a la inexistencia de fármacos relacionados a ciertas proteínas
Invesigadores bioinformáticos	Mejor acceso a los datos fuente Mejor manipulación e intercambio de datos	Dificultad de acceso a datos protéicos y genómicos Limitada posibilidad de manipulación de datos		Plataforma que ofrezca fácil acceso e intercambio de los datos fuente	Restricción de la cantidad y tipo de formatos disponibles

Tabla 1. Análisis de involucrados

16.2.4 Árbol de problemas

En el árbol de problemas se tomaron en cuenta las causas y efectos que forman parte de la determinación del planteamiento del mismo. De esta forma se desglosa la situación para realizar un análisis más exhaustivo.

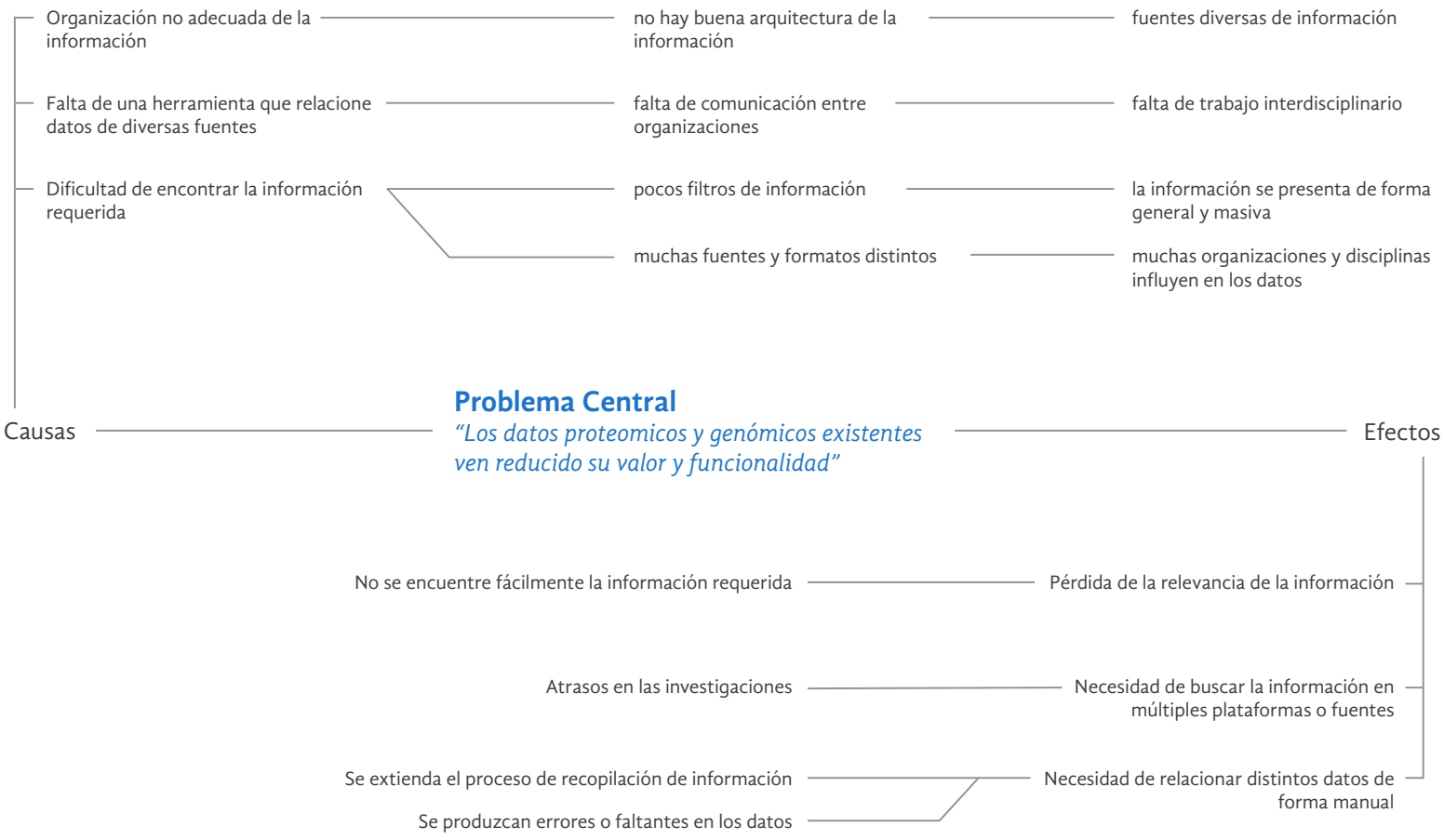


Tabla 2. Árbol de problemas.

16.2.5 Objetivo general

Diseñar una herramienta de visualización de datos proteómicos y genómicos existentes que brinde mayor valor y funcionalidad a los mismos a través de una buena arquitectura y análisis de la información.

16.2.6 Objetivos específicos

1. Investigar sobre las distintas tecnologías de visualización de información adecuadas a este tipo de base de datos e investigar la naturaleza de los datos proteicos y genómicos con el fin de mejorar el análisis de los mismos.

2. Desarrollar las alternativas de visualización así como las posibilidades de relación de los datos que respondan a las necesidades de análisis de los mismos.

3. Desarrollar la herramienta que facilite el análisis de los datos proteicos y genómicos.

16.2.7 Árbol de objetivos

Mediante el árbol de objetivos, se visualizan los medios y los fines del proyecto con el fin de determinar el objetivo general que va a dar el rumbo al mismo.

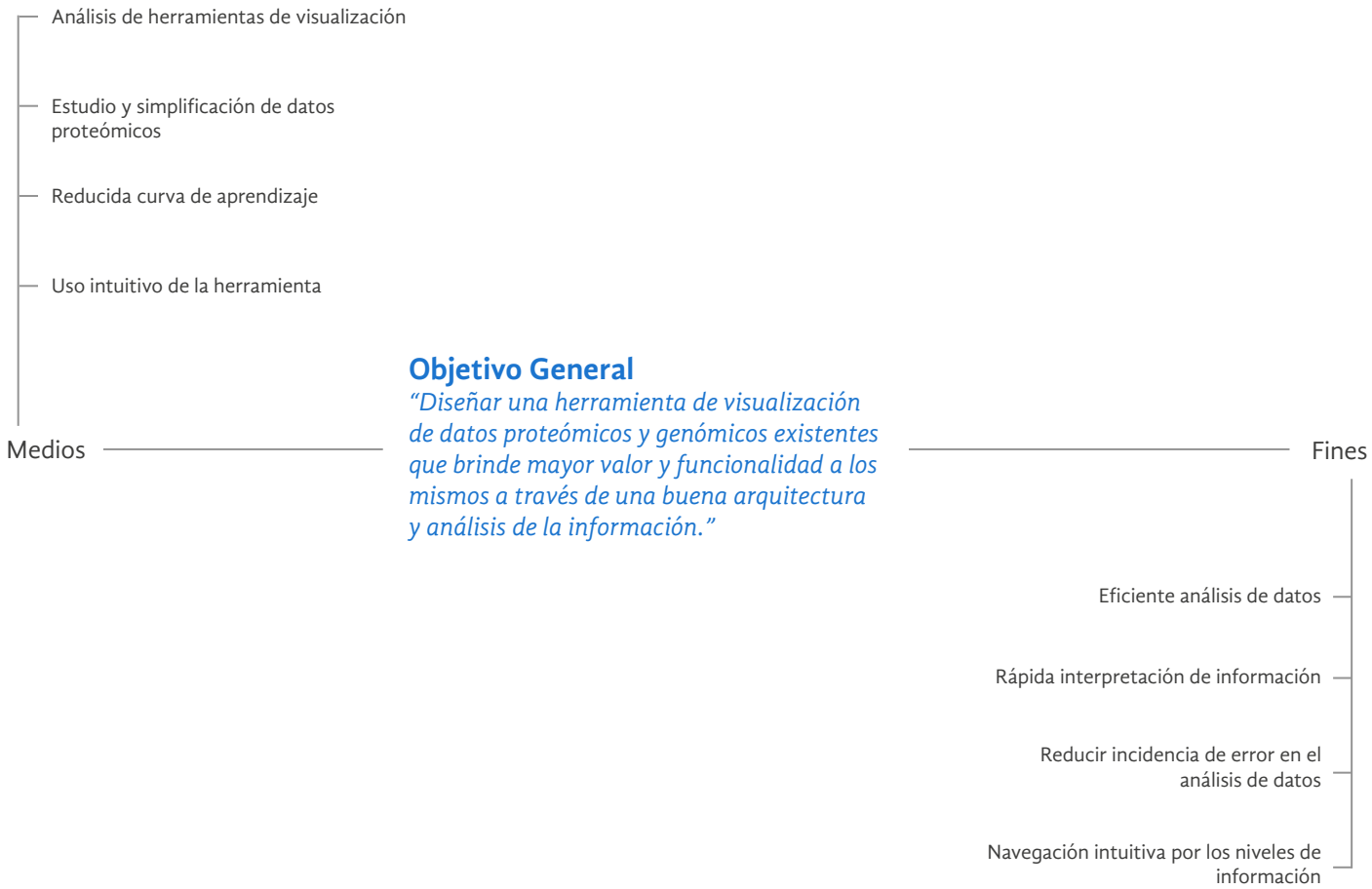


Tabla 3. Árbol de objetivos.

16.2.8 Soporte lógico

Se establece el alcance que va a tener el proyecto, las limitaciones que pueden afectar el correcto desempeño del mismo, así como los supuestos que se dan por cumplidos para el desarrollo del proyecto y los riesgos que se corren y que podrían afectar tanto el proyecto como la empresa.

Alcance	Limitaciones	Supuestos	Riesgos	Impacto
Se presenta una maqueta funcional que contenga datos relacionados, debidamente organizados de acuerdo con las necesidades detectadas.	Es difícil contactar a los expertos en el campo debido a que el área se encuentra en desarrollo en el país y cuentan con el tiempo muy reducido.	Los datos permanecerán disponibles, libres y debidamente manipulables.	Atraso en las fechas que se plantean en el marco metodológico como consecuencia de falta de coordinación entre las partes del proyecto.	Representación alternativa de los datos que va a permitir visualizar relaciones que anteriormente estaban ocultas.
	La falta de comprensión profunda de los datos en el área de la bioquímica dificulta el procesamiento y organización de la información.	Los software a utilizar estarán disponibles y actualizados conforme los requerimientos actuales.	Problemas externos que impidan el acceso a las bases de datos públicas.	Mayor eficiencia de interpretación de datos que se va a ver reflejado en menor tiempo destinado a la investigación y correlación y mayor tiempo al análisis de los mismos.
	La cantidad y calidad de la información dificulta la organización de los mismos.			

Tabla 4. Soporte lógico.

16.3 Justificación

El avance de la tecnología ha traído consigo nuevas formas de investigación, obtención y procesamiento de datos, los cuales día a día son más abundantes y de fuentes más confiables y con respaldo.

Las herramientas de visualización que existen en la actualidad no tienen la capacidad de procesar la gran cantidad de datos existentes por lo que su potencial uso en investigaciones científicas se ve reducido y comprometido, ya que los investigadores son más propensos a cometer errores al correlacionar datos de forma manual.

Mediante una visualización que sintetice diferentes bases de datos genómicas y proteómicas y las relaciones de forma sistemática, se pretende aumentar valor a los datos que se encuentran disponibles.

El correcto análisis y procesamiento de datos va a reducir el tiempo que los investigadores invierten en correlacionar datos, por ende las investigaciones tendrán resultados en un menor tiempo, trayendo beneficios a la comunidad médica y científica y va a garantizar que mediante una herramienta intuitiva, se obtengan datos con menor posibilidad de cometer errores de análisis.

17 Antecedentes

17.1 El cáncer de mama:

De acuerdo con información de Roche, "El cáncer de mama es el cáncer más común en mujeres. Cada año se diagnostican cerca de 1.4 millones de nuevos casos y cerca de 450,000 mujeres morirán a causa de esta enfermedad" (2013). Los datos de la IARC (Agencia Internacional para la Investigación del Cáncer por sus siglas en Inglés) respaldan esta información, que puede verse en el gráfico 1. Para ver información comparativa con respecto a otros tipos de cáncer, remitirse al Anexo 14.1

En Costa Rica los datos más actualizados son del 2013, donde se registraron 348 muertes de acuerdo con la información recopilada por el INEC (Instituto Nacional de Estadística y Censos). En general, se puede decir que cerca del 32% de los casos de cáncer de mama terminan en muertes. Es importante aclarar que no todos los casos de cáncer ocurren por causas genéticas, sin embargo, en el país y en el mundo actualmente se realizan investigaciones para averiguar la incidencia del cáncer por causas genéticas en la población femenina. En Costa Rica estas investigaciones las realiza, por ejemplo, la Facultad de Biología de la Universidad de Costa Rica en conjunto con al Hospital Calderón Guardia (Gutiérrez, Espeleta, Moreno, & García, 2014). Investigaciones estadísticas como esta son de suma importancia para la planificación, prevención y atención del cáncer de mama en el país.

La investigación en cáncer es abordada desde un enfoque altamente interdisciplinario: biólogos, bioquímicos, químicos, físicos, médicos y en los últimos años, debido a la gran cantidad de información recopilada, investigada e indexada, arquitectos de información, bioinformáticos, informáticos, estadistas, ingenieros en tecnologías de la información y diseñadores.

Incidencia del Cáncer de Mama (IARC, 2012)

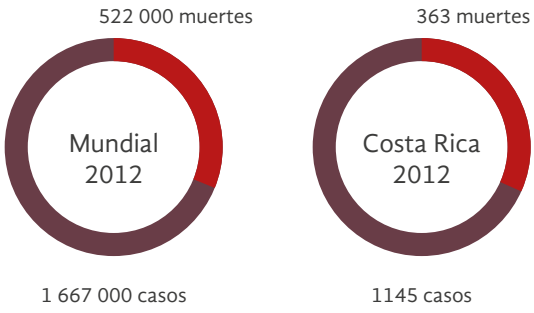


Gráfico 1. El cáncer de mama.

17.2 La bases de datos genómicas y proteómicas:

Como se mencionó previamente, la cantidad de información que se ha investigado en relación con el cáncer de mama es vasta y data desde 1994, cuando se descubrió el gen BRCA-1, uno de los genes conocidos por tener relación directa con la incidencia del cáncer de mama. (National Institutes of Health, 2012). Esta situación generó que los gobiernos y los centros de investigación internacionales buscaran soluciones para centralizar, almacenar, organizar, publicar, visualizar y distribuir los datos de genes y de proteínas. Como ejemplos se pueden mencionar la Protein Data Bank[1], perteneciente a The Research Collaboratory for Structural Bioinformatics (RCSB, s.f.) y también al navegador genómico Ensembl[2], (European Bioniformatics Institute, s.f.).

Estas plataformas recopilan información proveniente de fuentes de todos los rincones del planeta y buscan indexarla de forma que los investigadores puedan accesarla para diferentes fines, por el ejemplo el estudio de la forma y comportamiento de las proteínas ligadas al cáncer o también la búsqueda de fármacos con base en las estructuras proteómicas descubiertas.

17.3 Big Data:

La historia de la Big Data se remonta sorpresivamente a los años 1970, con la implementación en Chile de un sistema nacional de información llamado Cybersin (Jackson, 2014). Este sistema actualizado a mano y utilizando tecnologías análogas, permitía recibir información en tiempo real en un "centro de comando", captando la mayor cantidad de información y tomando decisiones estratégicas con base en los datos.

Hoy en día el análisis de Big Data se utiliza para identificar tendencias de comportamiento, precedir consumo el consumo energético en el hogary detectar emergencias antes de que ocurran, para mencionar algunas aplicaciones. (MongoDB, 2014).

En setiembre de este año, la empresa GBM en alianza con IBM introdujo procesadores optimizados para el manejo de Big Data en Costa Rica, usando la tecnología Power8. (Salas Víquez, 2014). Para el país, esto implica que las empresas transnacionales van a requerir de profesionales especializados tanto en el manejo de datos, como en la visualización de los mismos. De acuerdo con el artículo citado, "se estima que actualmente en el mundo se producen 2,3 trillones de gigabytes por día, [...] para el 2015 se producirá 40 veces esa cifra". El hecho de que en Costa Rica se comienzen a implementar estas tecnologías de servicios de análisis es un indicador de que la fuerza laboral nacional está preparada para afrontar estos nuevos retos tecnológicos.

[1] Protein Data Bank http://pdb.org/

[2] Ensembl Genome Browser http://www.ensembl.org/

18 Marco metodológico:

Primera fase:

1 Planteamiento del problema:
Analizar la problemática existente y plantear el problema del proyecto.

2 Planteamiento de objetivos:
Definición del objetivo general así como los objetivos específicos del proyecto.

3 Investigación de contenido:
Investigar el área de la bioquímica y la bioinformática con el fin de conocer suficiente del tema para poder trabajar con los datos proteicos y genómicos.

4 Análisis de lo existente
Investigar el área de la visualización de la información con el fin de establecer las diferentes alternativas y acercamientos que pueden existir para abordar el proyecto.

5 Definición del concepto
Conceptualizar diversas propuestas gráficas y de interacción con el fin de evaluar la funcionalidad y uso de la herramienta.

Segunda fase:

6 Generación de propuestas:
Prototipar distintas propuestas trabajando con una muestra pequeña de datos ejemplares.

7 Prueba y evaluación:
Evaluar las propuestas con el fin de tomar decisiones de mejora de los prototipos.

8 Definición de propuesta final:
Seleccionar una propuesta y comenzar la implementación con datos ejemplares.

9 Detallado de propuesta:
Refinar la propuesta funcional y estéticamente.

10 Implementación:
Desarrollar la propuesta final.

11 Informe y Resultados:
Generar el informe y documentación de la totalidad de las fases del proyecto.

19 Marco teórico:

19.1 Tecnologíasdelavisualizaciónde la información:

19.1.1 Métodos/paradigmas de visualización

En el área de la visualización de la información existen una serie de taxonomías distintas para clasificar los distintos métodos o paradigmas usados para visualizar información.

Esta área no es necesariamente una disciplina nueva: ejemplos del uso de visualización de datos para el análisis y la búsqueda de patrones remontan a 1854, cuando el médico John Snow trazó sobre un mapa del centro de Londres la ubicación de personas infectadas del Cólera, con el fin de dar con la "zona cero" o zona de infección y así detener la propagación de la enfermedad.(DashingD3.js, 2014); en la actualidad, este tipo de paradigma se conoce como "mapa de distribución de puntos". De igual forma, existe una serie de clasificaciones de acuerdo con distintos autores para nombrar a los diferentes tipos de visualizaciones existentes. Con el motivo de facilitar este estudio, se decidió seguir la taxonomía de Los Servicios de Datos y SIG de la Universida de Duke (Zoss, s.f.): ver siguiente página.

2D/Plano

Choropleth: Herramienta que funciona para visualizar datos territoriales con relaciones de otros factores asociados.

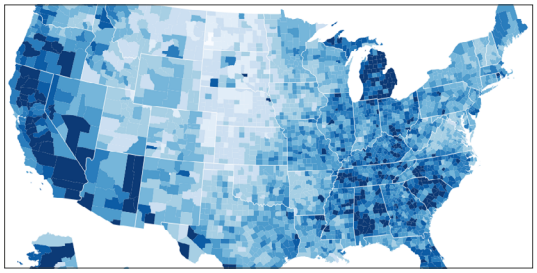


Gráfico 2. Choropleth.

Distribución de puntos: Permite la representación de aparición de un elemento mediante el uso de elementos gráficos como puntos. Es posible relacionar dos tipos de datos diferentes.

Gráfico 3. Distribución de puntos

Temporal:

Gráfico de flujo: Es utilizado para representación de datos continuos tales como series de tiempo, puede reemplazar barras apiladas y demuestra lo que sucede al interpolar datos temporales.

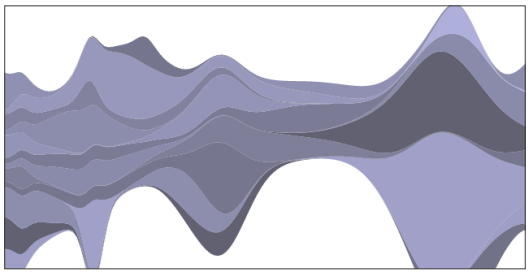


Gráfico 4. Gráfico de flujo

Diagrama de arco: Muestra relación entre diferentes elementos, mediantes arcos, a través del tiempo. El grosor de los arcos puede variar para mostrar que ese dato es más recurrente o más importante en la visualización.

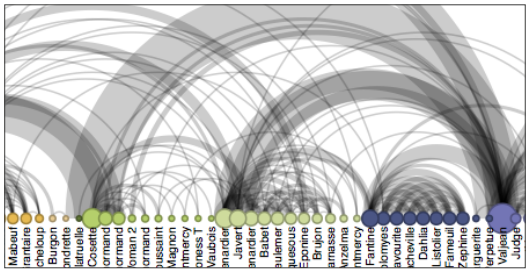


Gráfico 5. Diagrama de arco

Multidimensional

Nube de burbujas: Es un gráfico de burbujas desordenadas, que muestra diferentes datos mediante diferenciación de tamaño y color. La agrupación de los elementos no muestra ningún orden específico, por lo que los datos se interpretan por medio del cambio dimensional de los elementos.

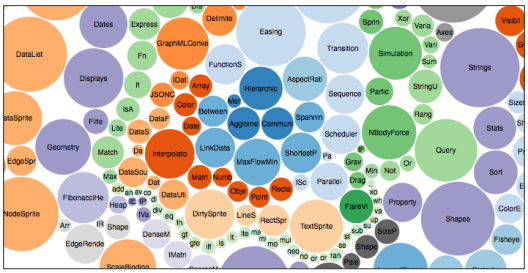


Gráfico 6. Nube de burbujas

Coordenadas paralelas: Muestra relaciones entre diferentes ejes, mediante líneas, el usuario tiene la posibilidad de filtrar cuál sección de una coordenadas desea visualizar y de esta forma no mostrar la totalidad de las relaciones para brindar mayor claridad.

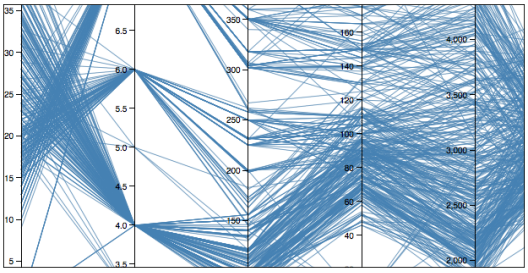


Gráfico 7. Coordenadas paralelas

Árbol/Jerárquico

Árbol hiperbólico: Muestra relaciones jerárquicas en una distribución radial de los elementos, es posible determinar cuál es la raíz de cada elemento presente en el árbol y cuáles son las ramas que comprenden el mismo.

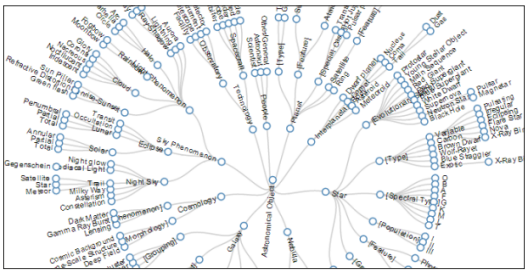


Gráfico 8. Árbol/Jerárquico

Mapa de árbol: Muestra relaciones entre diferentes elementos, brinda la posibilidad de "zoom" mediante animaciones que permiten una navegación más intuitiva en la herramienta e identificar visualmente cuál es la jerarquía entre los elementos presentes en el diagrama.

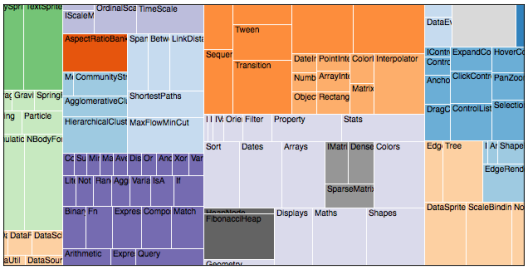


Gráfico 9. Mapa de árbol

Red

Matriz: El diagrama de matriz correlaciona datos de dos coordenadas diferentes, mostrando las coincidencias mediante cuadros, los cuales corresponden a diferentes categorías de acuerdo a los colores en que se dibujan.

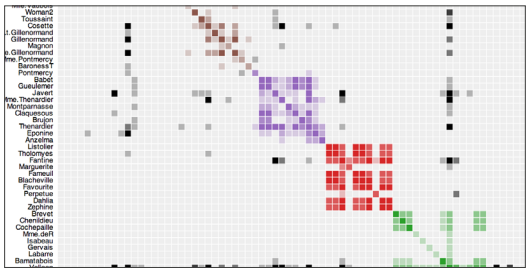


Gráfico 10. Matriz

Diagrama de cuerdas: Los datos se encuentran distribuidos en el perímetro de un círculo y de esta forma se dan relaciones del tipo “todos con todos”, mediante curvas que muestran el inicio y el fin de la correlación.

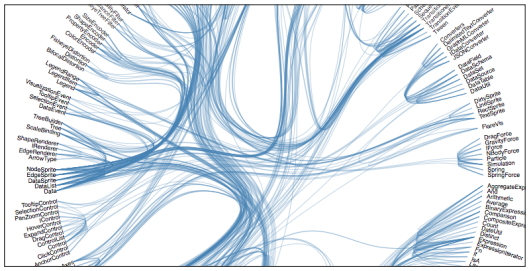


Gráfico 11. Diagrama de cuerdas

Comparación de paradigmas

Se definieron las ventajas, desventajas y el uso que se le da a cada uno de los paradigmas para analizar cuál es el más adecuado para visualizar los datos del proyecto.

Paradigma	Ventajas	Desventajas	Uso
2D Plano	Buena distribución espacial	Se necesita un diagrama espacial correcto antes de colocar los datos	Distribución geográfica y ubicación espacial
Temporal	Facilidad de correlacionar datos a través del tiempo	Si los períodos de tiempo son extensos es difícil visualizar todos los datos	Visualización de períodos de tiempo
Multidimensional	Posibilidad de correlacionar datos de diferente naturaleza	Conforme aumenta la cantidad de niveles es más dificil relacionarlos	Diagramas de agrupación de elementos
Árbol/Jerárquico	Facilidad de representación de pertenencia por categorías	No siempre es fácil comprender los niveles de jerarquía	Diagramas de visualización de niveles de jerarquía
Red	Posibilidad de correlacionar todos los datos sin necesidad de agrupación	Dificultad de comprensión de las correlaciones	Matrices de correlación de datos variados

Tabla 5. Tabla de comparación de paradigmas.

19.2 Area de la bioquímica y la bioinformática:

19.2.1 Introducción:

Como se mencionó anteriormente, las bases de datos como Ensembl y PDB albergan información referente a datos de genes y proteínas. Antes de hablar sobre la naturaleza de dichos datos, es importante mencionar algunos conceptos claves para la comprensión de la información.

En primera instancia, es clave comprender la relación entre el ADN y las proteínas. El ADN contiene las instrucciones necesarias para que el cuerpo fabrique nuevos seres vivos y nuevas estructuras biológicas, como células y proteínas. (Biology-Online, 2008). Las proteínas entonces, se fabrican a partir de porciones de ADN, estas porciones se conocen como ADN codificante o exones.

Las proteínas se componen de unidades básicas llamadas aminoácidos, estos a su vez se componen de unidades moleculares más pequeñas llamadas bases nitrogenadas. Una porción del ADN puede representarse de la siguiente manera:

Representación de la doble hélice

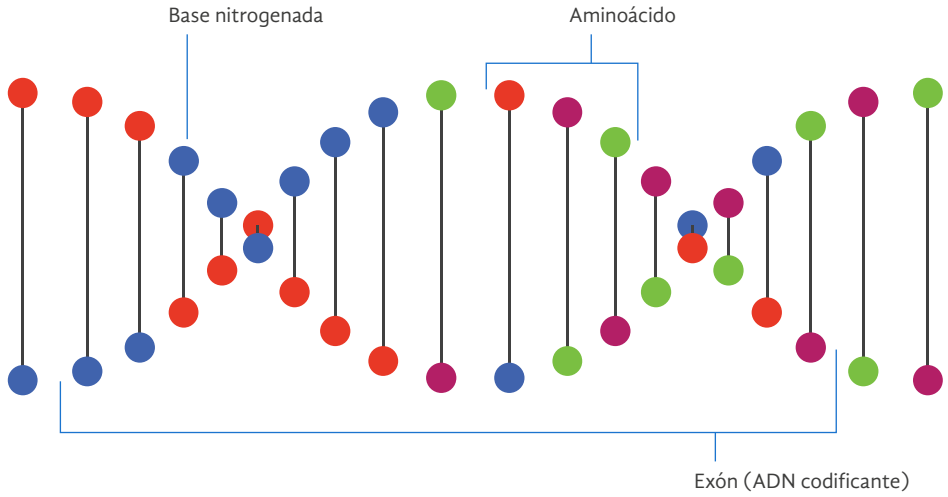


Gráfico 10. Representación de una cadena de ADN codificante

Las bases nitrogenadas, señaladas en azul, rojo, verde y púrpura, siempre se organizan en pares ordenados como muestra el gráfico de la página anterior. Gracias a esto, los investigadores solo necesitan estudiar un lado de la doble hélice y por lo tanto, la carga cognitiva de la información se simplifica.

Un lado de la hélice con las iniciales de las bases nitrogenadas

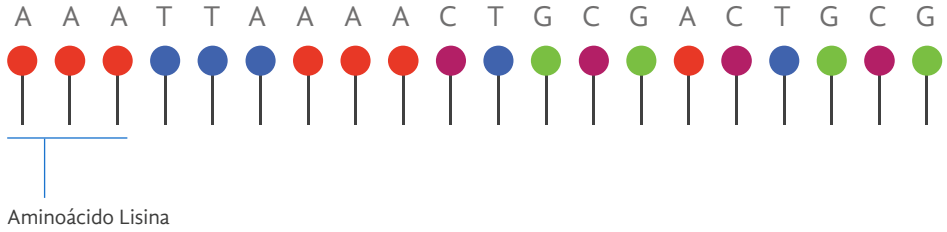


Gráfico 10. Representación de una porción de una cadena de aminoácidos

Una vez comprendida la noción básica de qué es el ADN y por qué está conformado, es importante explicar qué es una proteína; en términos sencillos las proteínas son "máquinas biológicas", compuestas por "cadenas de aminoácidos, las cuales determinan la función y estructura de la misma" (Biology-Online, 2008). Las proteínas cumplen una diversidad de funciones dentro del cuerpo.

En el siguiente gráfico se representa una base nitrogenada que ha cambiado con respecto a la secuencia natural; esto produce también un cambio en el aminoácido y finalmente una variante en la función normal de la proteína.

Un cambio anormal en la primera base nitrogenada

Gráfico 11. Representación de una porción de una cadena de aminoácidos y una posible mutación.

En resumen, una proteína está compuesta por aminoácidos, que a su vez se componen de unidades más pequeñas llamadas bases nitrogenadas. Cuando surge un cambio en una de estas bases, se altera la secuencia de aminoácidos de la proteína y por consiguiente, su función en el organismo.

Por ejemplo, la proteína conocida como BRCA1 (Breast Cancer Type 1 Susceptibility protein) (National Institutes of Health, 2012), es una proteína que en su estado sano se encarga, en terminos sencillos, de eliminar células malignas en el tejido mamario. Cuando BRCA1 está mutada, pierde la capacidad de neutralizar éstas células provocando que se repliquen de forma incontrolada.

19.2.2 Naturaleza de los datos proteómicos y genómicos:

Con base en la información anterior y en las bases de datos analizadas, se realiza un breve resumen de la naturaleza de los datos ahí contenidos con en fin de comprender qué tipo de información es, cómo está compuesta y posteriormente en la fase de diseño, cómo poder visualizarla de mejor manera:

Múltiples clases: Existen dos clasificaciones generales en los datos, por un lado están las mutaciones y por otro lado está la cadena de aminoácidos. Estos dos tipos de datos deben ser correlacionados.

Gráfico 12. Representación gráfica de múltiples clases de datos.

Múltiples correlaciones: Debido a la gran cantidad y variedad de datos que se presentan en la herramienta de visualización, se requiere hacer múltiples correlaciones entre los mismos, para mostrar cuáles mutaciones están presentes en cada aminoácido, así como información complementaria adjunta.

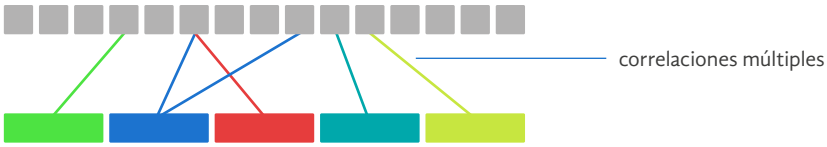


Gráfico 13. Representación gráfica de correlaciones entre datos.

Distintas fuentes: Los investigadores actualmente encuentran los datos a través de diversas fuentes, por lo que información que se van a visualizar mediante la herramienta no se encuentra compilada ni correlacionada en ninguna plataforma.

Secuenciales y no secuenciales: Los aminoácidos son datos secuenciales, esto quiere decir, que al ser una cadena, tienen un principio y un fin, y de esta forma deberán ser visualizados. Por otra parte las mutaciones no son secuenciales, así que no es relevante en el orden en que se presenten, siempre y cuando las relaciones con los aminoácidos sean correctas.

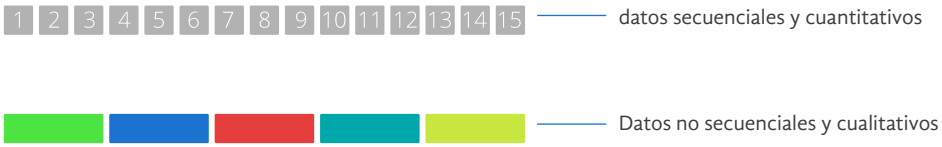


Gráfico 14. Representación gráfica de múltiples tipos de datos.

19.2.3 Formatos y clasificación de los datos:

En el área de la bioinformática entonces, no resulta extraño que exista una gran variedad de formatos especiales en los cuales los datos se han codificado. Si bien muchas veces los datos pueden descargarse en formato .txt (documento de texto sencillo), las bases de datos antes mencionadas también han creado sus propios formatos. A continuación se mencionan algunos importantes:

GFF (General Feature Format): Este formato es común en la base de datos de Ensembl y consiste en un archivo de texto delimitado por tabulaciones que describe características genómicas. (Broad Institute, s.f.). Consiste de una línea por característica, cada una con nueve columnas de datos. (Ensembl, s.f.).

FASTA: Un formato de líneas que contienen data de secuenciación de los aminoácidos. La primera línea contiene datos descriptivos y las siguientes contienen la secuencia de aminoácidos de una proteína en específico. (NCBI, s.f.).

DSSP: Más que un formato, DSSP es un sistema de a|signación de estructuras secundarias de la proteína a una secuencia de aminoácidos. Cuenta con siete letras para definir los tipos existentes de estructura, por ejemplo: H para Hélice Alfa y B para Puente Beta. (CMBI, s.f.).

19.2.4 Bases de datos y fuentes de información proteómica y genómica:

Existen diversas bases de datos disponibles al público en general y a los investigadores de distintas áreas que pueden ser consultadas gratuitamente. La información de estas bases es pública y el resultado de una inmensa cooperación internacional por centralizar y facilitar el acceso a la información.

Se analizó además la forma en que las mismas facilitan la visualización de datos para los usuarios. La forma en la que los investigadores interpretan los datos está directamente relacionado a la estrategia de visualización y diseño que se utiliza en las plataformas de bases de datos. A continuación se presentan tres bases de datos que fueron de enorme ayuda para la investigación del presente proyecto:

PDB: Del inglés Protein Data Bank, es un portal Estadounidense de información de estructuras macromoleculares biológicas. La herramienta PDB contiene compilaciones de estudios proteómicos que se han llevado a cabo a nivel mundial y han sido validados para garantizar la veracidad de la información que se brinda a los investigadores que consultan la página.

En la plataforma PDB se analizó específicamente la visualización general de la proteína BRCA-1. En este caso es evidente la falta de una simbología clara para que el investigador entienda la información: la visualización presenta gran cantidad de colores y formas diferentes que representan características específicas de la proteína, pero no brinda una diferenciación clara entre los mismos.

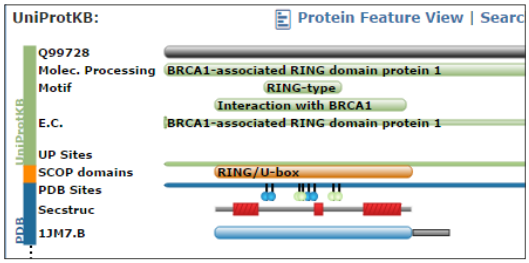


Gráfico 15. Vista general de los datos proteómicos de BRCA1 (en pdb).

EBI: Es una plataforma del Instituto Europeo de Bioinformática que brinda información sobre proteínas, genes y químicos. Ofrece información y datos en su plataforma, así como redireccionamiento a páginas complementarias y cuenta con secciones informativas como la de noticias y los apartados que incluyen formación profesional.

Ensemble: Es una sección de información complementaria que es parte de EBI fundada en 1999. Brinda información genómica y visualizaciones gráficas de los datos. Además da la posibilidad de obtener los datos genómicos en formatos que pueden ser procesados y analizados posteriormente.

Gráfico 16. Imagen de las variantes de BRCA-1 (Ensembl, s.f.).

19.3 Big Data:

Como se puede asumir , el hecho de constituir tantas y tan diversas bases de datos, sumado a las décadas de investigación en el área genómica y proteómica han generado una enorme cantidad de información que debe, de alguna manera, ser controlada y dosificada para su estudio y análisis. En este punto entonces entra en juego el concepto de Big Data que "aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales". (Barranco, 2012).

En el caso de los datos genómicos, los análisis estadísticos tradicionales arrojan hallazgos limitados y básicos, que no aprovechan realmente el torrente de información; es por eso que al recurrir a procesos más inteligentes de tratamiento y visualización de datos, es en teoría posible lograr nuevas conclusiones e inferencias a partir de los mismos datos. La Big Data en general está compuesta por tres componentes básicos:

Volumen: El genoma humano promedio se estima en aproximadamente 3200 Mb or 3.2×10^9 bp. (es decir, 3 200 000 pares de bases nitrogenadas). En términos de datos, esto genera a partir de una secuenciación de ADN unos 200 gigabytes de información. (Robinson, 2014).

Variedad: Los datos genómicos y proteómicos no sólo están compuestos por la cadena de bases nitrogenadas, sino que existen datos adicionales, secuenciales y no secuenciales, cualitativos y cuantitativos, etc. Tal y como se explicó en el apartado 4.2.1.

Velocidad: Estos datos deben ser procesados, visualizados y extraídos rápidamente. Con los mínimos tiempos de espera puesto que es información pertintente para realizar investigaciones de toda índole, por eso los sistemas informáticos deben estar estructurados para que eso sea una realidad.

19.4 Implementación:

Para implementar el diseño propuesto, se cuenta con una variedad de opciones dependiendo de los alcances internos del proyecto y de los recursos disponibles. En esta sección se presentan aquellas herramientas que se considera son las más importantes:

19.4.1 Adobe Edge Animate:

Adobe Edge Animate es una herramienta de animación en HTML que utiliza JavaScript como lenguaje de programación para interacciones avanzadas. Es una "herramienta nueva de diseño interactivo y movimiento" (Adobe, 2012). Edge permitiría generar una maqueta básica con funcionalidad limitada que se usaría para analizar parcialmente la usabiliad, interacción y ejemplos de uso de la herramienta de visualización.

19.4.2 Processing:

Processing es un lenguaje de programación que ha "evolucionado en una herramienta de desarrollo para profesionales". (Processing, s.f.). Este lenguaje es muy últil para realizar prototipos funcionales utilizando Java como base de lenguaje, pero con una estructura más sencilla. El equipo de diseño además tiene experiencia previa utilizando esta herramienta en proyectos de diseño de interacción y de diseño visual.

19.4.3 JavaScript (con HTML5 y CSS3):

JavaScript es el "lenguaje de programación más popular del mundo". (W3 Schools, s.f.). JS (abreviado) es el lenguaje para HTML, para la web y es el estándar de la industrial. Junto con la versión 5 de HTML y la versión 3 de CSS, conforma un poderoso y relevante ambiente de desarrollo que asegura que exista respaldo, actualizaciones y vigencia.

JS es un lenguaje más complejo que Processing, pero con la ventaja de que es más abierto y existe una gran cantidad de bibliotecas que funcionan en conjunto, además de que la distribución de la aplicación se vuelve mucho más sencilla debido a que los dispositivos como computadoras y tablets están diseñados para leer HTML nativamente.

19.4.4 D3.js:

D3.js (Data-Driven Documents) es una biblioteca para JS que fue diseñada específicamente para la visualización de datos. D3 ayuda a manipular documentos de datos usando HTML, SVG y CSS, (Bostock, 2013) y permite la creación de una diversa cantidad de gráficos y visualizaciones tanto estáticas como dinámicas en un ambiente estandarizado, escalable y abierto.

19.4.5 3VOT:

"3VOT es una Plataforma de Publicación Web, es una forma simple y sencilla para publicar contenido en internet" (Rodríguez, 2014). 3VOT está diseñado especialmente para trabajar con JavaScript, CSS, HTML y multimedia y también cuenta con un sistema de almacenamiento en la nube.

Además, 3VOT es un proyecto costarricense, por lo que el equipo de diseño consideró como buena opción para implementar el proyecto de manera más sólida, escalable y dinámica y también para dar a conocer otras innovaciones nacionales a través de su uso.

20 Análisis de referenciales:

20.1 Referenciales:

Se revisó una colección de diferentes tipos de visualización con el fin de identificar elementos positivos en las mismas y ser utilizados como punto de partida en la visualización que se pretende generar, así como identificar aspectos negativos, ésto para evitar caer en los mismos errores de visualizaciones existentes.

20.1.1 Map your Moves:

Map your moves es una exploración visual de big data sobre 4000 mudanzas en New York.

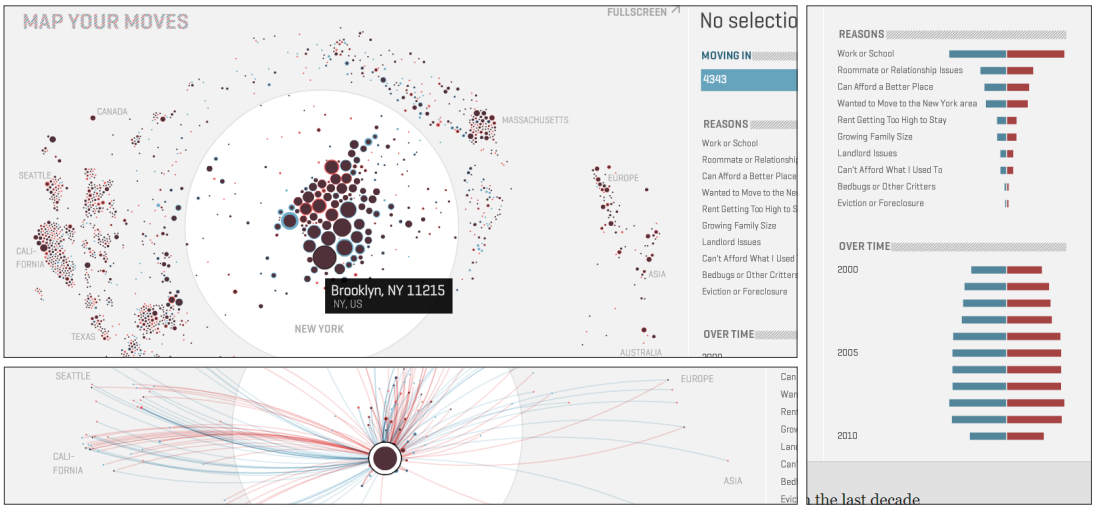


Gráfico 17. (Stefaner, 2010).

Aspectos Positivos

Interfaz limpia e intuitiva
Gran cantidad de datos en revelación progresiva
Capacidad de ver muchos o pocos datos fácilmente
Web responsive

Aspectos Negativos

Gráficos de la derecha estáticos
Colores no intuitivos
Distribución no representa posiciones geográficas reales

20.1.2 Flight Risk – Every Major Commercial Plane Crash of the Last 20 Years:

Flight Risk es una infografía interactiva que representa datos de accidentes aéreos de los últimos 20 años.

Gráfico 18. (BBC, s.f).

Aspectos Positivos

Diversos filtros de información
Revelación progresiva de datos
Transiciones/animaciones agradables
Distintas formas de accesar la información
Sobreposición estética de elementos gráficos

Aspectos Negativos

Sólo un eje de información (tamaño)
Sobresimplicidad en términos de big data
Los filtros podrían ser más claros

20.1.3 Expression data from healthy controls and early stage CRC patient's tumor:

Visualización de características de pacientes de cáncer colorrectal en diferentes fases de evolución de la enfermedad.

Gráfico 19. (Biogps, s.f).

Aspectos Positivos

Simplicidad de representación de información
Filtros que cambian jerarquía de visualización
Zoom intuitivo que dosifica el nivel de detalle
Barra de búsqueda

Aspectos Negativos

Sólo hay un eje de información.
La interacción de los filtros no es intuitiva
La barra de búsqueda no brinda resultados claramente

20.1.4 Beer Viz:

Herramienta de visualización que brinda correlaciones entre marcas y tipos de cervezas.

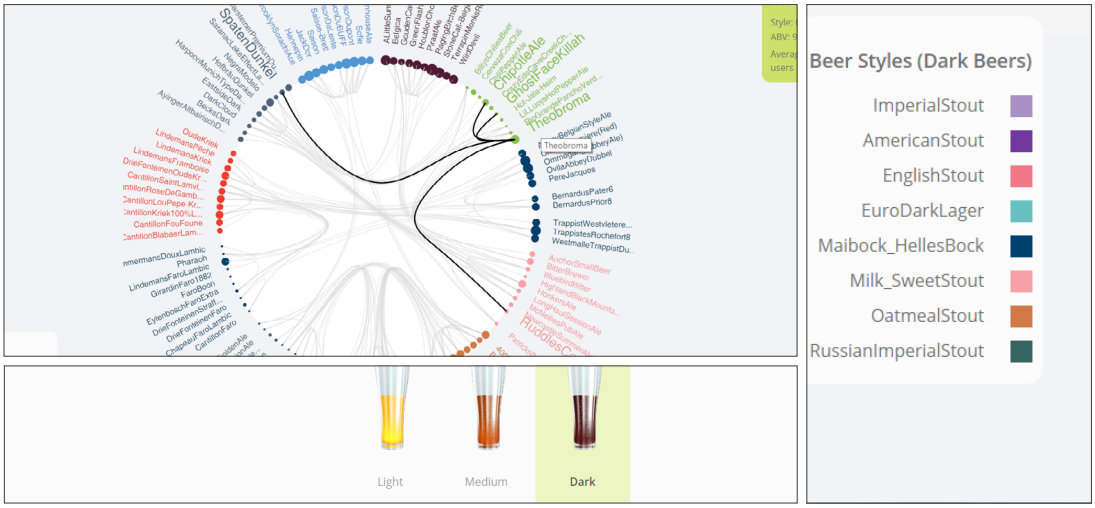


Gráfico 20. (Seekshreyas, s.f).

Aspectos Positivos

Información complementaria a la visualización principal
Filtros para revelar información
DIferentes niveles de información
Diferenciación de color por categorías

Aspectos Negativos

Poca claridad en el uso de filtros
La interacción entre los elementos no es clara
Poca correlación entre una visualización y otra

20.2 Conclusiones y mínimos comunes:

Se analizaron diferentes visualizaciones existentes con el fin de determinar cuáles elementos positivos se pueden rescatar para ser implementados en el diseño de la aplicación.

A partir de análisis que se definieron pautas a seguir con el fin de que la visualización sea una herramienta que realmente tenga un impacto positivo en la comunidad de investigadores.

Se determina entonces que la visualización deberá ser simple, además de contar con simbología y diferenciación de colores, que guíen al usuario y le indiquen de qué manera se interpreta la información que se presenta.

Deberá tener posibilidad de acercamiento y brindar información adicional si el usuario desea conocer más del tema, pero al mismo tiempo la herramienta debe brindar la posibilidad de ver la totalidad de los datos en una visualización principal y que los investigadores puedan ver, por ejemplo, la totalidad de mutaciones que se encuentran presente en una cadena de aminoácidos.

Además se determinaron elementos que han sido mal utilizados y por ello los resultados de visualización no han dado los resultados esperados, de esta forma se va a evitar cometer los mismos errores y desarrollar una herramienta intuitiva y con buenos elementos de diseño.

21 Concepto de diseño:

21.1 Síntesis de los análisis previos al proceso de diseño

Se procedió a a hacer una búsqueda de paradigmas y elementos gráficos como tipografía, color y estilo de línea, con el fin de determinar el componente estético de la visualización de proteínas a desarrollar.

Con base en la colección de imágenes se realizó un moodboard, que ayuda a determinar el estilo gráfico que se busca alcanzar en la herramienta.

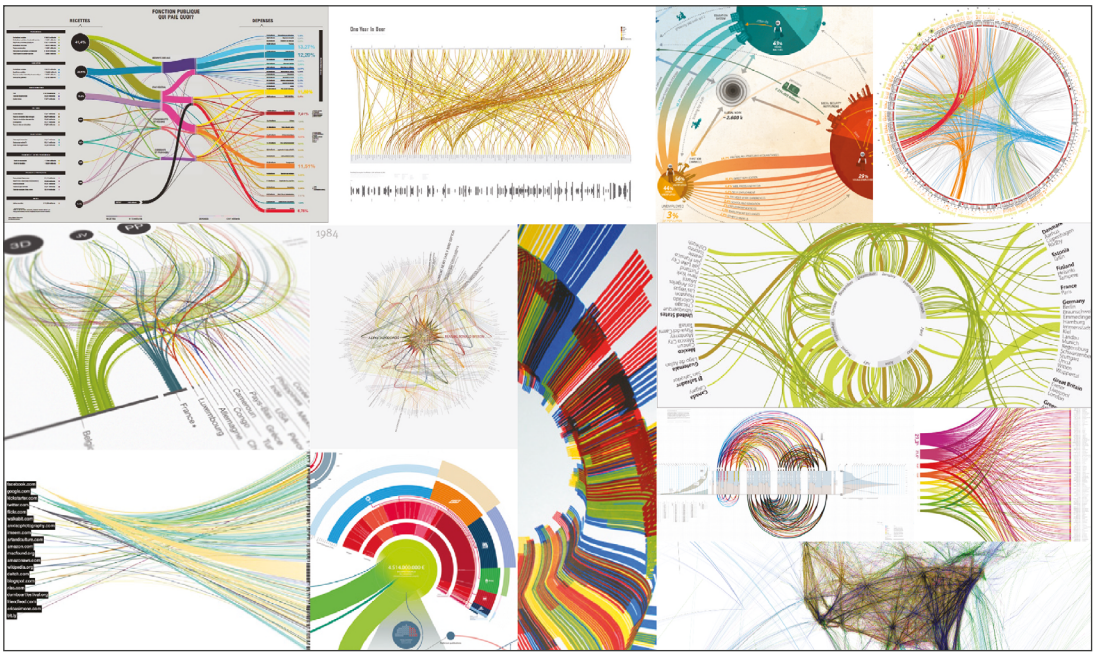


Gráfico 21. Moodboard.

21.2 Problema y criterios de investigación

21.2.1 Problema:

El problema que se planteó al inicio de la investigación define que los datos proteómicos y genómicos existentes ven reducido su valor y funcionalidad, debido a las herramientas de visualización que están siendo utilizadas para dicho fin.

Los investigadores bioquímicos, bioinformáticos y farmacéuticos deben buscar la información en múltiples plataformas o fuentes además de que se ven en la necesidad de correlacionar distintos datos de forma manual. Dicha situación propicia el error debido a la gran cantidad de información que se debe manejar. El hecho de que las correlaciones no estén automatizadas genera confusión así como pérdida de la relevancia de la información.

Como resultado el tiempo que se dedica a las investigaciones en el área de la genómica y proteómica es más extenso de lo que debería ser, si se tuvieran herramientas adecuadas de visualización e interpretación de datos. Debido a esto se infirió que el concepto de diseño que se va a desarrollar debe responder a dicha problemática que aqueja a los investigadores en las áreas anteriormente mencionadas. De esta forma la visualización de datos proteómicos y genómicos va a agilizar el trabajo de los investigadores, mediante una interfaz intuitiva y que garantice la veracidad de los datos obtenidos.

21.2.2 Criterios de investigación:

El desarrollo de la investigación requirió de una lista de criterios que debieron ser tomados en cuenta con el fin de resolver la problemática que se planteó. Se analizó una amplia cantidad de información y se realizaron entrevistas con expertos (Alpízar, 2014), (Gutiérrez, Espeleta, Moreno, & García, 2014), (Orozco, 2014) y (Hernández, 2014), con el fin de seleccionar los datos más pertinentes para los investigadores, que se convirtieron en los requisitos de información.

Con esto se determinaron dos variables de análisis distintas: por un lado los criterios que el equipo de diseñadores consideró importante analizar y por otro lado, la lista de criterios que la empresa facilitó tomando en cuenta las necesidades de los investigadores que van a tener relación directa con la herramienta de visualización.

Finalizando el proceso de selección se definión una lista de criterios de investigación reducida, tomando en cuenta las limitantes inherentes al proyecto y que se tomaron en cuenta al diseñar la herramienta de visualización.

A continuación se presenta la lista de criterios:

1. Secuencia de aminoácidos
2. Exones (ADN codificante)
3. Dominios y motivos asociados
4. Genes y mutaciones correlacionados

21.3 Requisitos de la interfaz:

La interfaz debe generar beneficios en tiempo de procesamiento y calidad de la información que se extrae de la misma. Para poder lograr dichos propósitos y además generar un acercamiento a la solución del problema planteado se debe seguir una serie de requisitos de visualización que se desarrollan a continuación.

Simple:

La herramienta debe ser simple, desde la forma en que se visualiza la información, hasta la manera en que se interactúa con la herramienta para obtener los datos. El objetivo principal va a ser que el investigador obtenga los datos más rápido y con mayor seguridad de no incurrir en errores, lo cual sólo puede ser alcanzado si la visualización es lo suficientemente sencilla para que usuarios multidisciplinarios puedan acceder a ella.

Intuitiva:

Debido a la cantidad de información que la herramienta va a desplegar, la misma debe ser fácil de usar por los investigadores, y ella misma debe indicar los pasos a seguir para acceder a información, sin necesidad de seguir un instructivo de uso.

Información clara:

La herramienta va a centralizar diferentes tipos de datos proteómicos, los cuales deben estar representados de forma clara, para que los investigadores no cometan el error de confundir un tipo de dato con otro.

Interactiva:

Se busca alcanzar una visualización con interacción del usuario, esto quiere decir que los investigadores tengan la posibilidad de visualizar mediante dosificación de información y filtros, para que la misma sea revelada de forma progresiva y no estática.

Multidimensional:

La herramienta va a permitir que diferentes dimensiones de información coincidan en una sola visualización, esto con el fin de unificar información que actualmente se encuentra distribuida en diferentes bases de datos y los investigadores deben acceder manualmente.

Escalable:

La herramienta que se va a plantear debe ser escalable a otras aplicaciones proteómicas, por lo tanto si se desea visualizar los datos de otra proteína, va a ser posible hacerlo por medio del diseño y visualización que se está planteando para la presente investigación.

Uso de tecnologías estándar:

Es preciso garantizar el acceso a los datos a todos los investigadores, sin importar el equipo o tecnología con la que cuente cada uno de ellos. Es por esto que los recursos tecnológicos que se utilizan deben ser el estándar que se utiliza en la industria.

Información correlacionable:

Por el hecho de que la herramienta va a unificar diferentes dimensiones de información, esta debe contar con la posibilidad de correlacionar estos datos automáticamente, con el fin de que los investigadores ya no procedan a hacerlo de forma manual y disminuir los errores humanos al momento de la obtención de datos.

21.4 Aspectos fundamentales para definir el concepto de diseño

Al tomar en cuenta el problema de diseño, los criterios y los requisitos de la interfaz, se determinaron los aspectos fundamentales que definen el concepto de diseño del proyecto:

- Diseñar una herramienta de visualización genómica y proteómica, intuitiva, simple y fácil de usar.
- Definir los niveles de información y plataformas de bases de datos que serán tomadas en cuenta para el diseño de la herramienta.
- Dar más valor a los datos que se encuentran disponibles a los investigadores, mediante una herramienta simple, de esta forma el usuario será capaz de obtener los datos que necesita para continuar con los siguientes estadíos en las investigaciones científicas.

21.5 Conclusiones sobre la conceptualización:

Se realizó el análisis de diferentes aspectos para ser tomados en cuenta en el desarrollo del proyecto.

El análisis de los referenciales dio la posibilidad de determinar elementos de diseño, interacción y funcionalidad que son positivos, y por ello podrían ser aplicados al desarrollo de la herramienta de visualización.

De la misma forma se determinaron elementos negativos que presentan las visualizaciones, con el fin de evitar cometer los mismos errores en las propuestas a desarrollar en la siguiente etapa del proyecto.

Con el desarrollo del moodboard se concluyó el clima gráfico que debe de tener la aplicación y en general, se logró el establecimiento del concepto de diseño que se va a seguir en las próximas etapas de desarrollo del proyecto de visualización de datos proteómicos.

Con respecto a las conclusiones que se extrajeron de los diferentes análisis, tomando en cuenta diseño, interacción y funcionalidad, se desarrollaron las alternativas que se muestran a continuación.

22 Desarrollo de alternativas y selección de la propuesta de diseño:

22.1 Alternativas de diseño:

Se desarrollaron 3 diferentes alternativas con el fin de generar exploración gráfica, de interacción y de funcionalidad, y posteriormente determinar cuáles son los elementos que se van a utilizar para la propuesta final de visualización de información.

22.1.1 “Double Chord”:

Se definió una estructura circular, la cual está compuesta de dos sub-estructuras, una la parte interna se encuentran las 14 mutaciones que se pueden dar, y en la parte exterior está la cadena de aminoácidos.

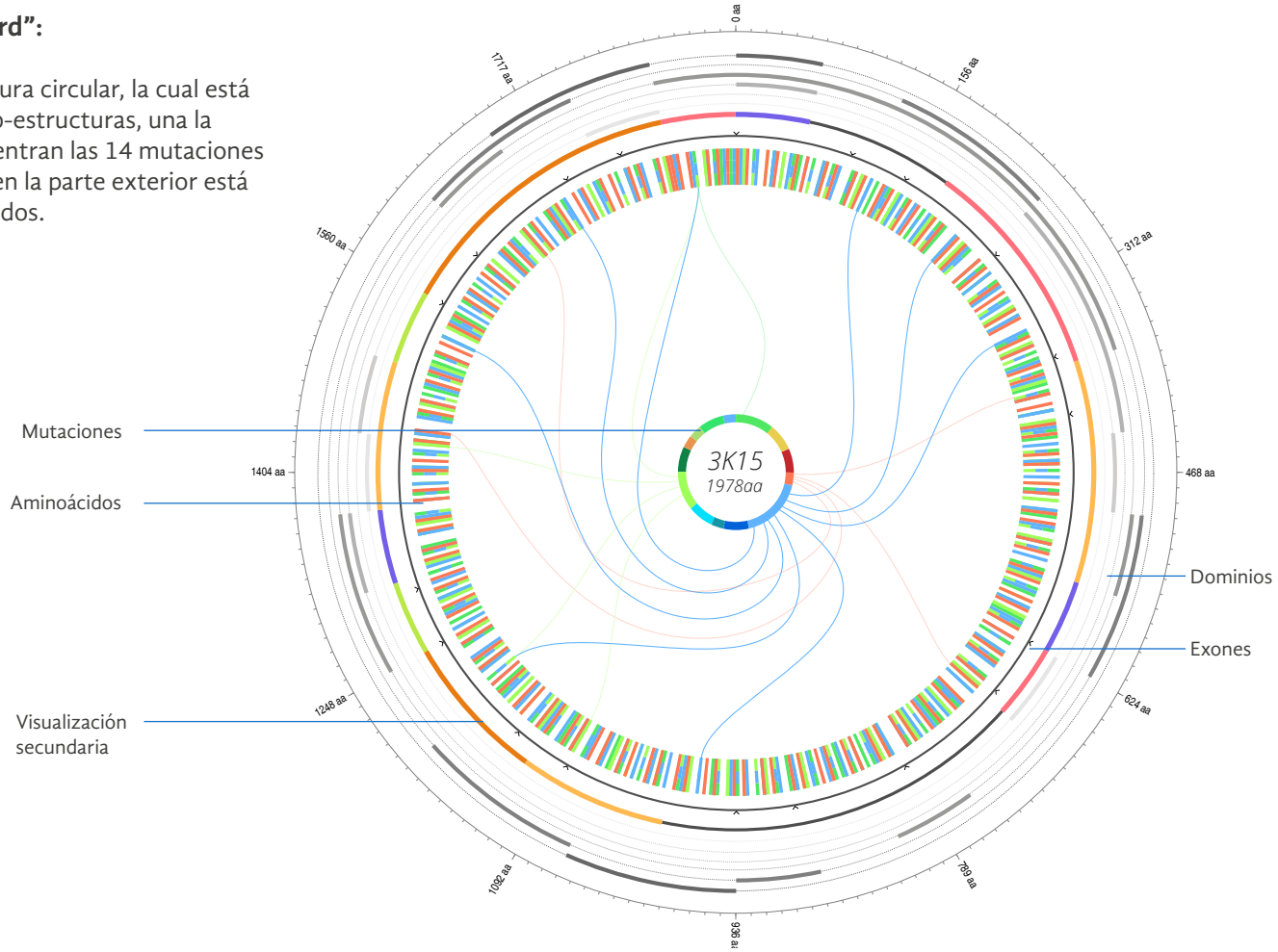


Gráfico 22. Propuesta conceptual 1: “Double Chord”

Las mutaciones se relacionan mediante líneas o cuerdas, a los aminoácidos en donde se encuentran las mismas. Además cuenta con una visualización secundaria en la cual se puede ver un acercamiento de los dominios de la proteína y sus estructuras secundarias relacionadas.

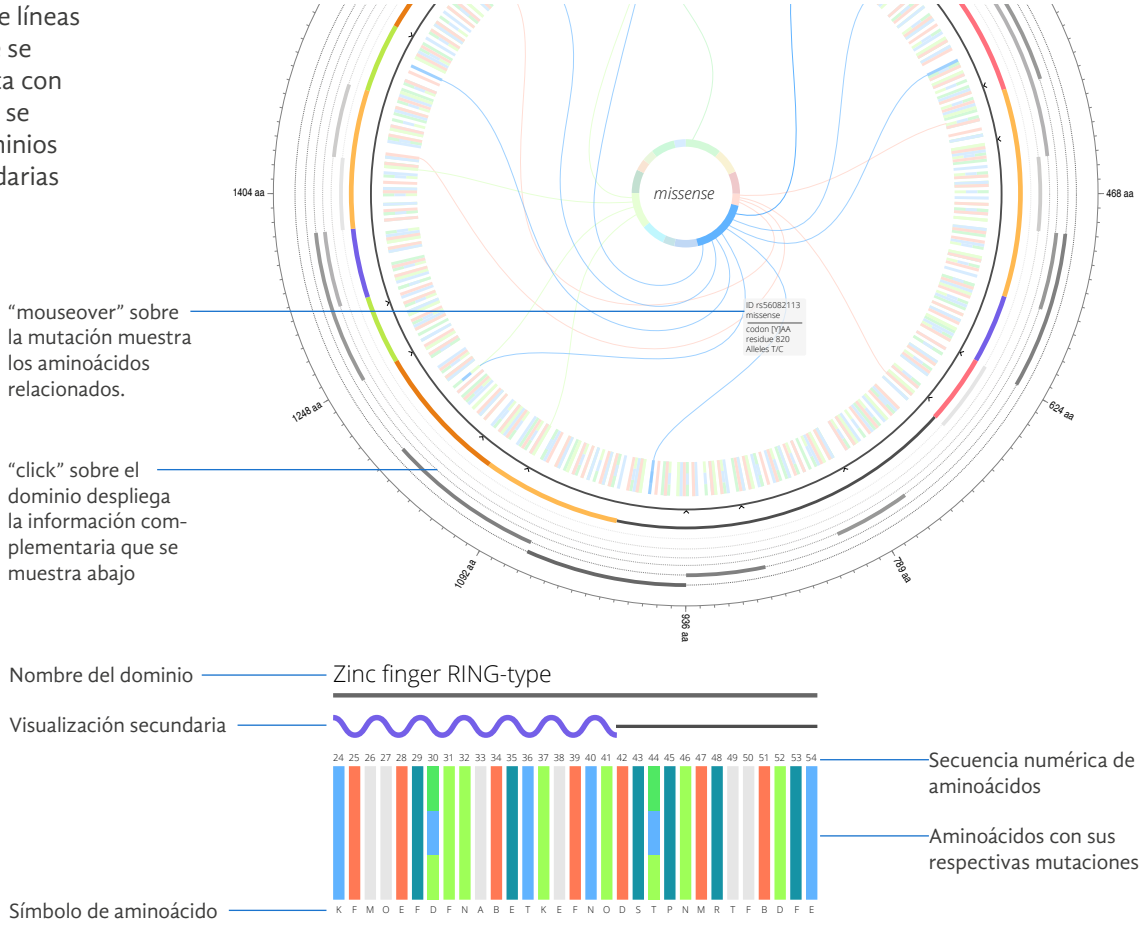


Gráfico 23. Detalles de propuesta conceptual 1: "Double Chord"

22.1.2 “Single Chord”:

Se utilizó la el paradigma del “Chord diagram” en el cual se distribuye la cadena de aminoácidos en una sección de la circunferencia y la porción restante es utilizada para distribuir los catorce tipos de mutaciones.

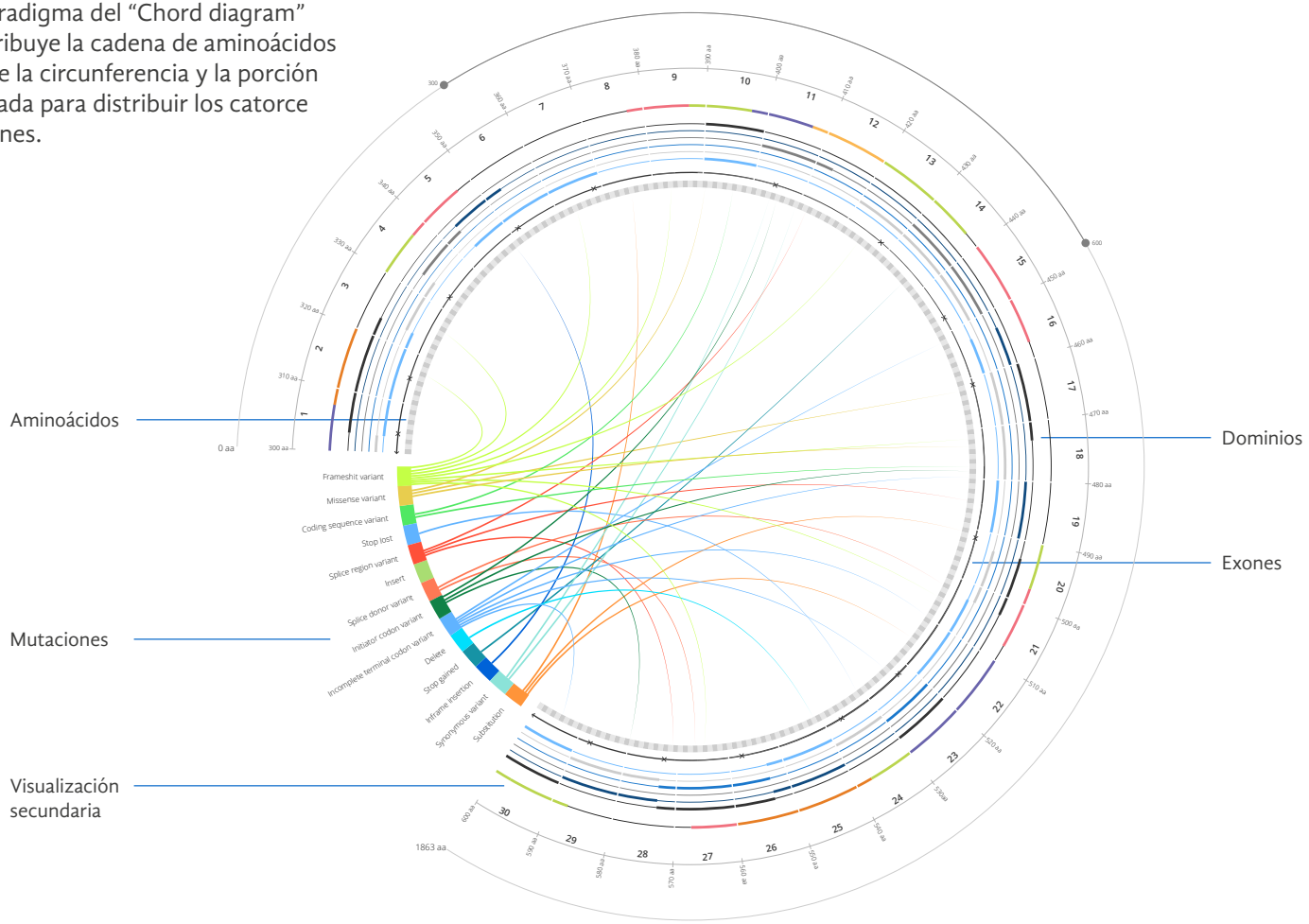


Gráfico 24. Popuesta conceptual 2: “Single Chord”

De esta forma en el interior del círculo se encuentran las correlaciones entre aminoácidos y mutaciones.

"click"sobre un tipo de mutación hace que el resto de las correlaciones cambien a un valor de opacidad menor.

En el exterior de los exones se colocan seis ejes con los dominios y por último en el eje externo se encuentra una visualización de la estructura secundaria de la proteína.

La visualización secundaria aumenta la definición de los datos con respecto al zoom. Mostrando formas pictográficas cuando el zoom es mayor.

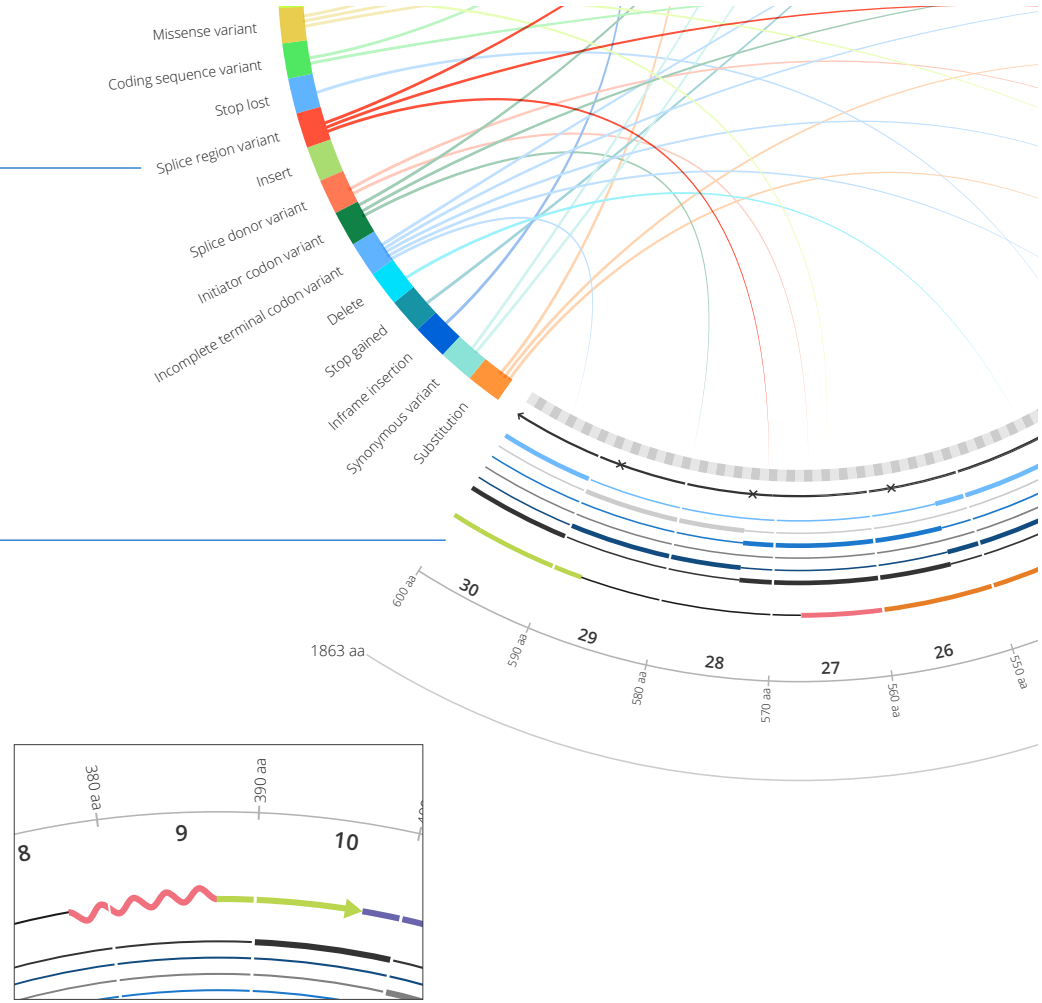


Gráfico 25. Detalles de propuesta conceptual 1: "Dobule Chord"

22.1.3 “Parallel Coordinates”:

Se diseñó una estructura lineal que cuenta con dos ejes principales, en el eje superior se encuentra la cadena de aminoácidos y en el eje inferior se colocaron las mutaciones.

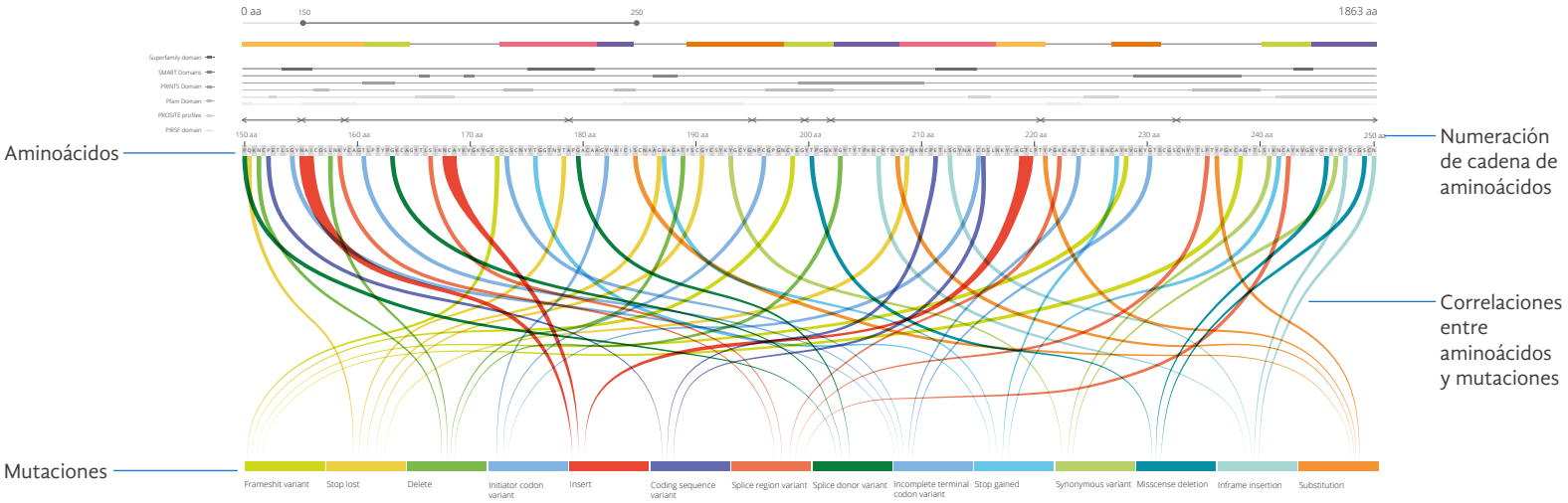


Gráfico 26. Propuesta conceptual 3: “Parallel Coordinates”

Los ejes se correlacionan mediante curvas del color correspondiente a la mutación. El usuario determina cuál mutación necesita visualizar y la herramienta muestra los aminoácidos que tienen presente dicha mutación.

Además esta visualización cuenta con la posibilidad de realizar “zoom” mediante una barra, controlada por dos puntos, el usuario decide el inicio y el fin de la porción de la cadena de aminoácidos que quiere visualizar.

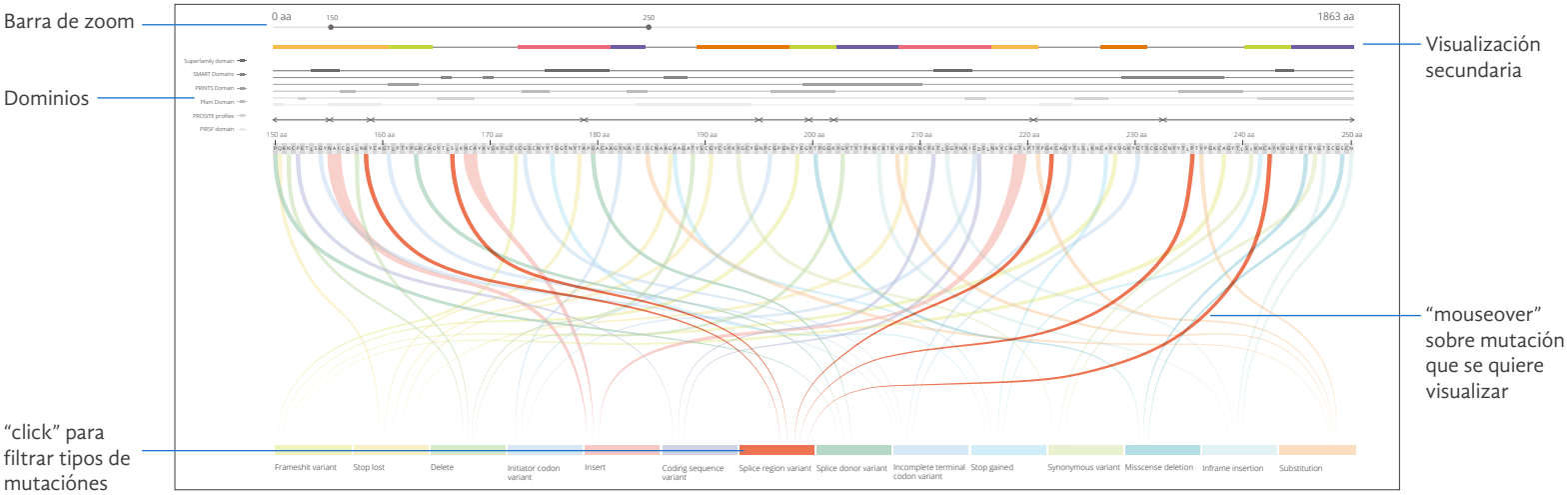


Gráfico 27. Detalles de propuesta conceptual 3: “Parallel Coordinates”

22.2 Casos de estudio:

A continuación se analizan situaciones de uso de las 3 propuestas de visualización, con el fin de determinar las acciones que el usuario debe realizar para obtener información, y de esta forma determinar cuál alternativa cumple con los requisitos de usabilidad.

22.2.1 Búsqueda de mutaciones

Se describe el caso en el que el investigador busca cuál mutación está relacionada a los aminoácidos de la cadena.

Double Chord

El usuario selecciona el tipo de mutación que necesita visualizar, posteriormente tiene la posibilidad de realizar "mouse-over" y se despliega un tooltip con información complementaria de la correlación.

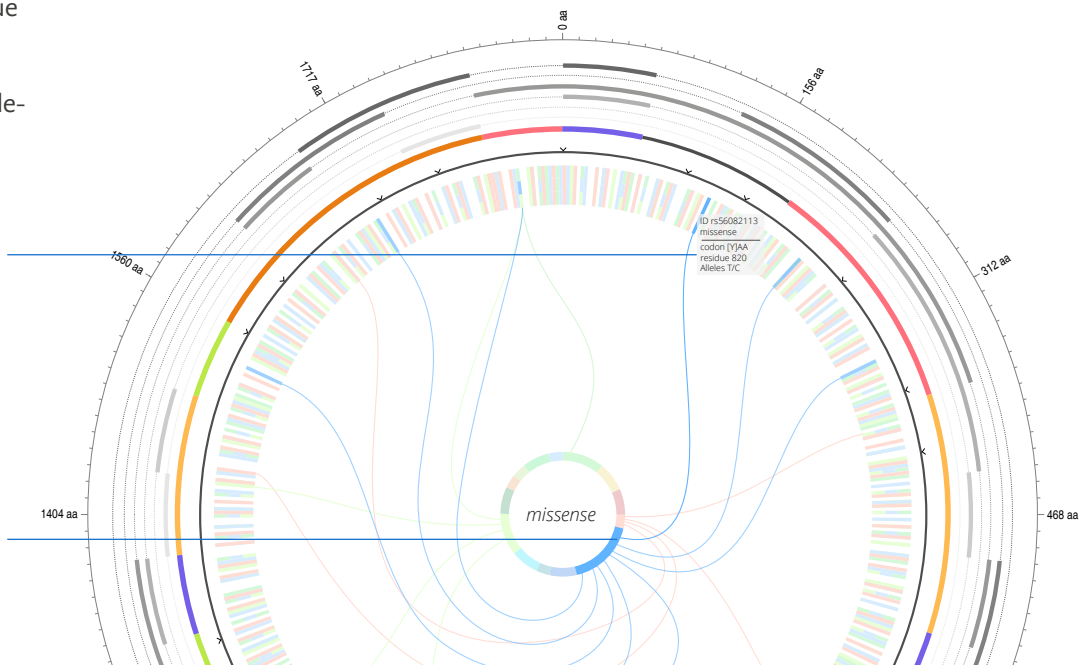


Gráfico 28. Caso 1, Double Chord.

Single Chord

Se selecciona el tipo de mutación, las relaciones correspondientes se hacen más visibles y las demás pierden opacidad para dar mayor enfoque a las que el usuario necesita ver.

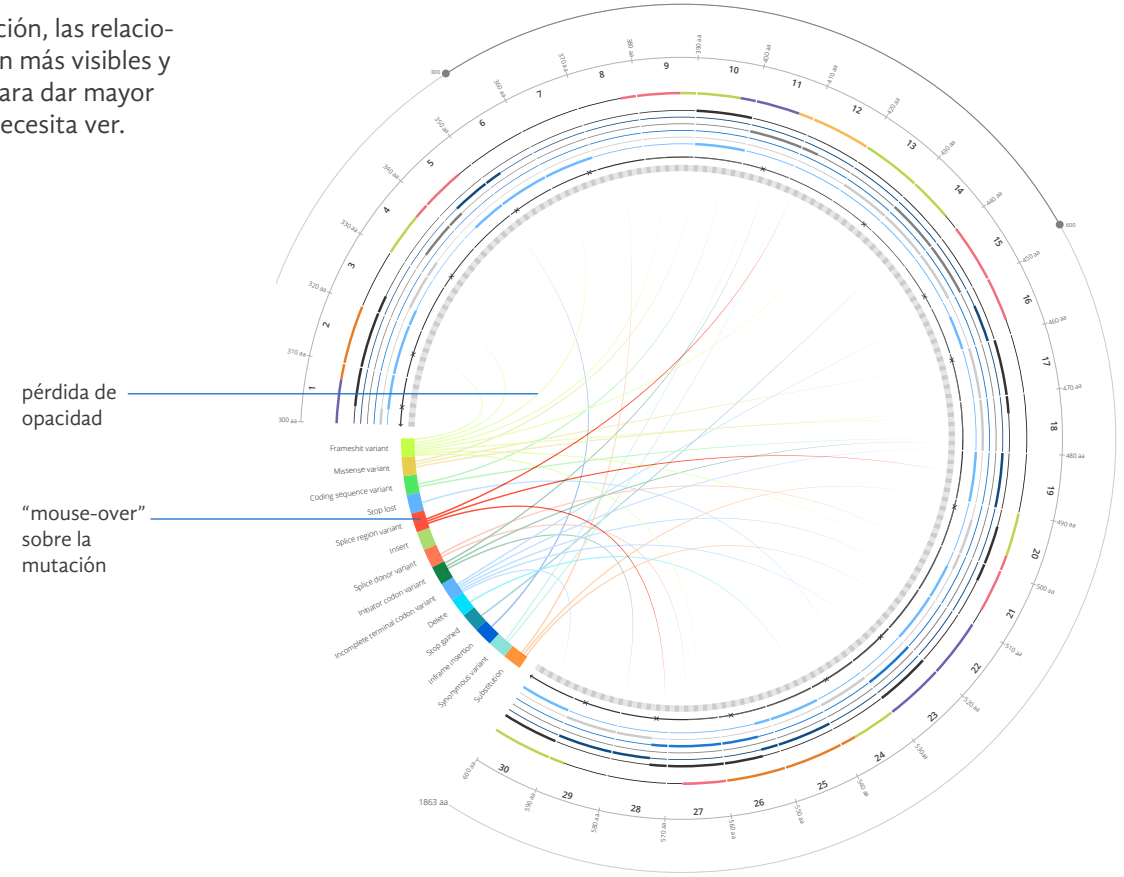
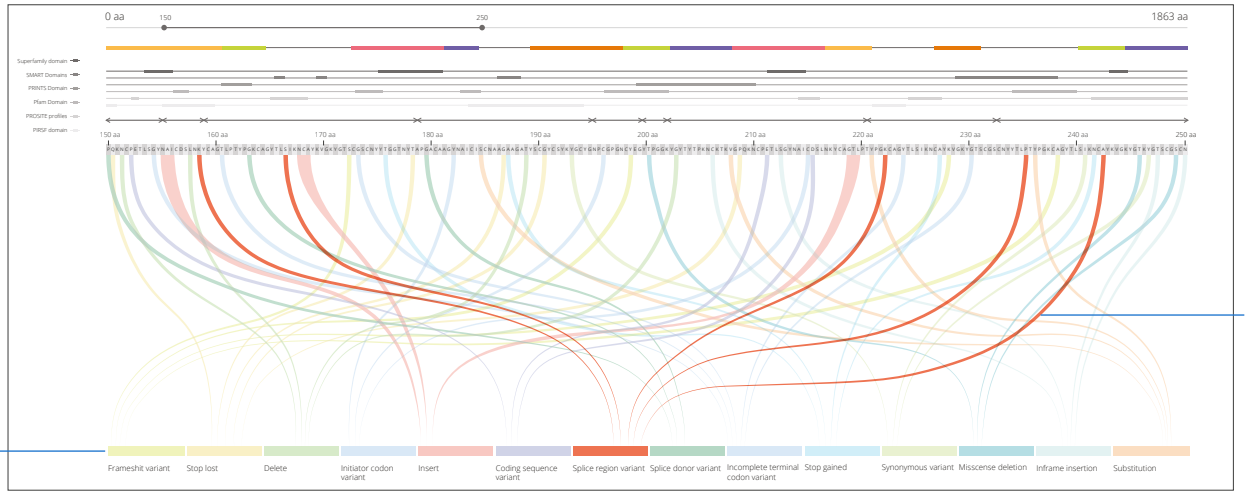


Gráfico 29. Caso 1, Single Chord.

Parallel Coordinates

En este caso tanto las correlaciones como el recuadro del tipo de mutación se hacen más visibles para facilitar la interpretación de los datos de la proteína.



"mouseover" sobre mutación que se quiere visualizar

menos opacidad en los tipos de mutaciones restantes

Gráfico 30. Caso 1, Parallel Coordinates.

22.2.2 Búsqueda de dominios

Se describen las acciones que el investigador realiza para poder obtener información sobre los dominios de la proteína.

Double Chord

En esta propuesta el usuario puede escoger entre botones/filtros cuál es la clasificación de dominios que desea visualizar, además se despliega una visualización secundaria al seleccionar un dominio específico.

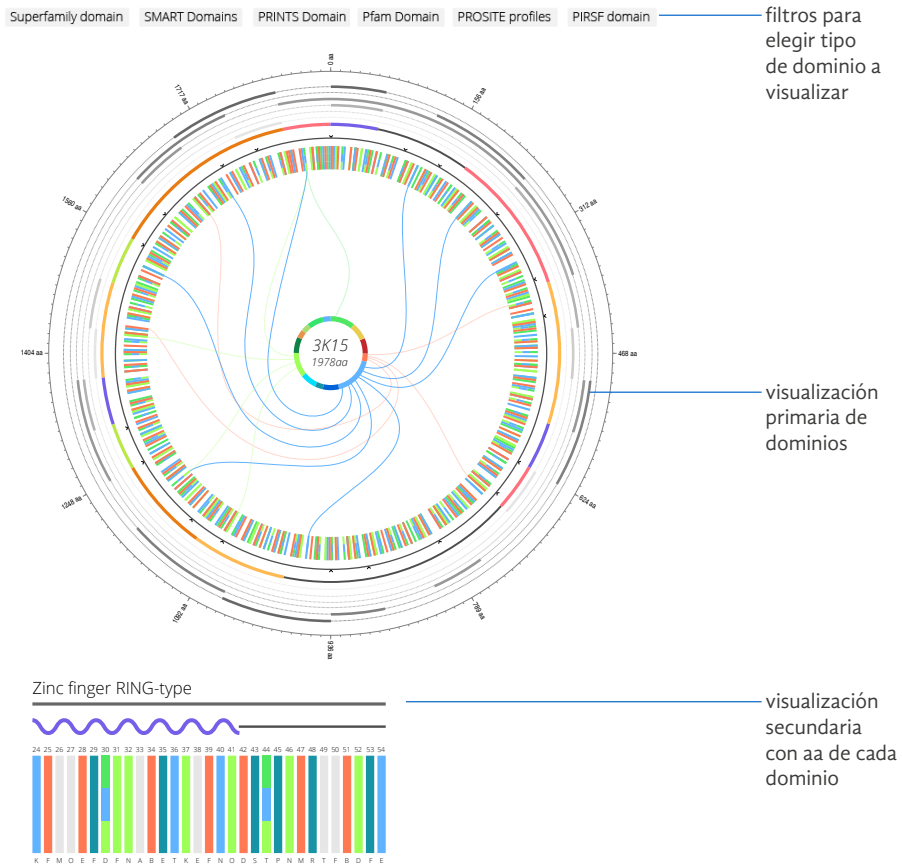


Gráfico 31. Caso 2, Double Chord

Single Chord

En esta propuesta se cuenta con una simbología por color, intercalando líneas grises y azules, con el fin de ayudar al usuario a seguir las líneas alrededor de la circunferencia.

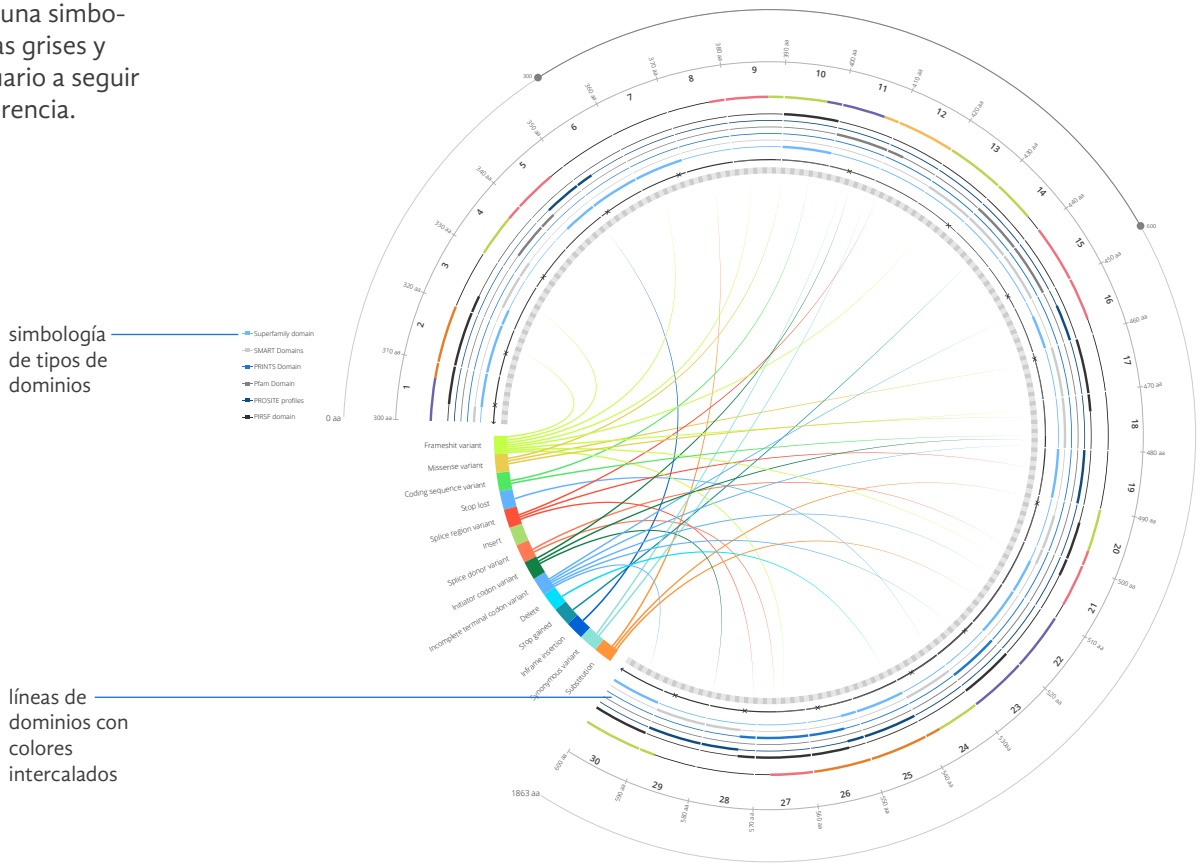


Gráfico 32. Caso 2, Single Chord

Parallele Coordinates

En esta propuesta, el dominio seleccionado queda visible junto con sus aminoácidos y correlaciones correspondientes, mientras que el resto de dominios pierden opacidad.

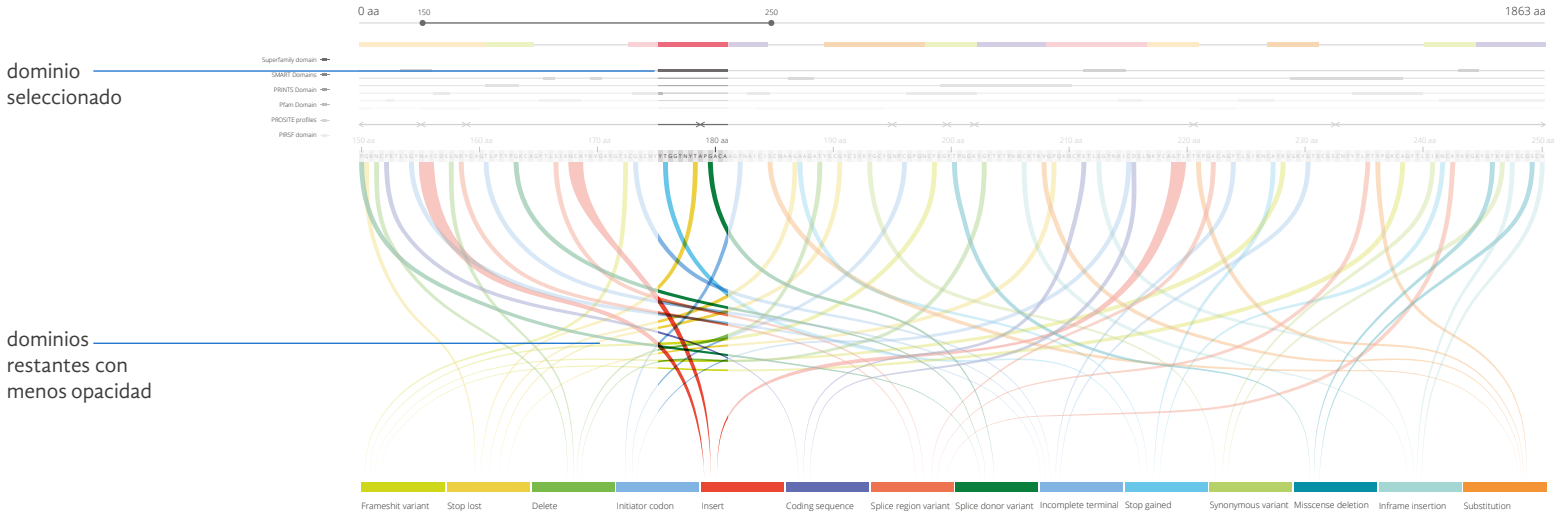


Gráfico 33. Caso 2, Parallel Coordinates.

22.2.3 Búsqueda de estructura secundaria

Se describen los pasos a seguir por parte de los usuarios para visualizar la estructura secundaria de las proteínas en las diferentes propuestas de diseño.

Double Chord

En esta propuesta la estructura secundaria se encuentra representada mediante simbología de color alrededor de la circunferencia. Además se cuenta con una visualización complementaria en la cual se representa de forma pictórica cada estructura.

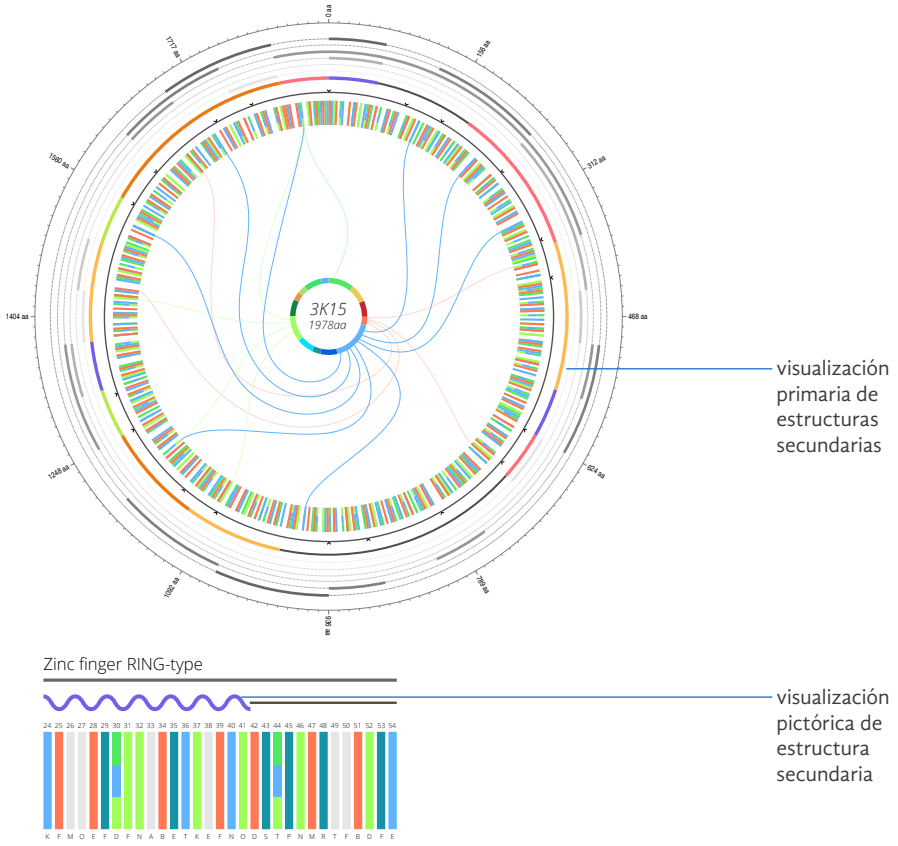


Gráfico 34. Caso 3, Double Chord.

Single Chord

Las estructuras secundarias se representan mediante simbología de color, conforme disminuye la cantidad de información mostrada (por medio el zoom), aumenta el nivel de detalle y con esto las mismas se representan de forma pictórica.

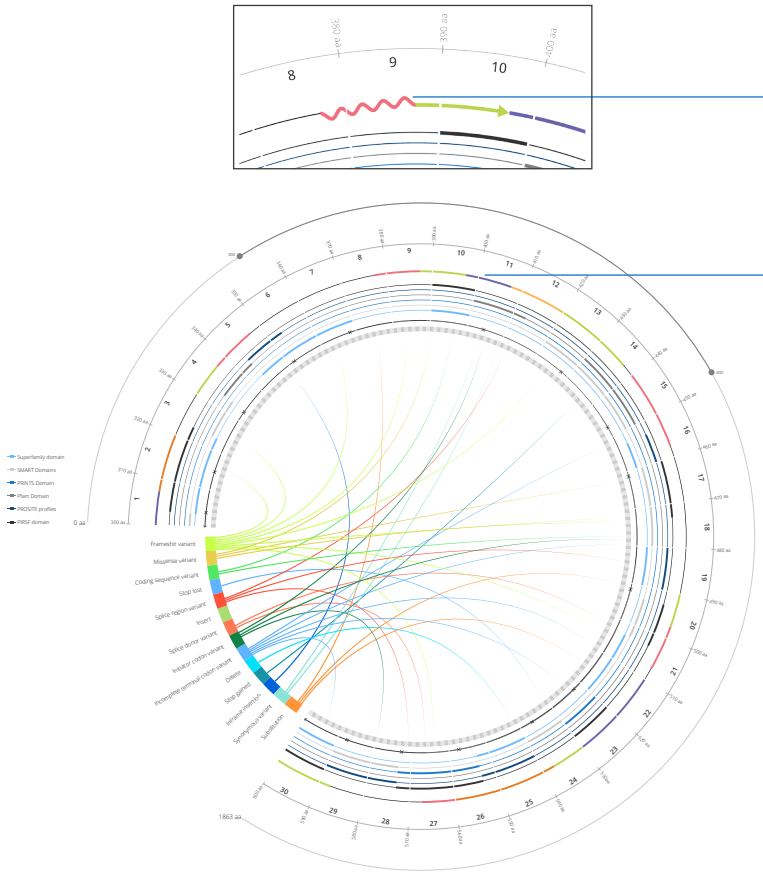


Gráfico 35. Caso 3, Single Chord.

Parallele Coordinates

Las estructuras secundarias se encuentran representadas de manera lineal siguiendo la simbología de color, además conforme aumenta el nivel de detalle se representan de forma pictórica.

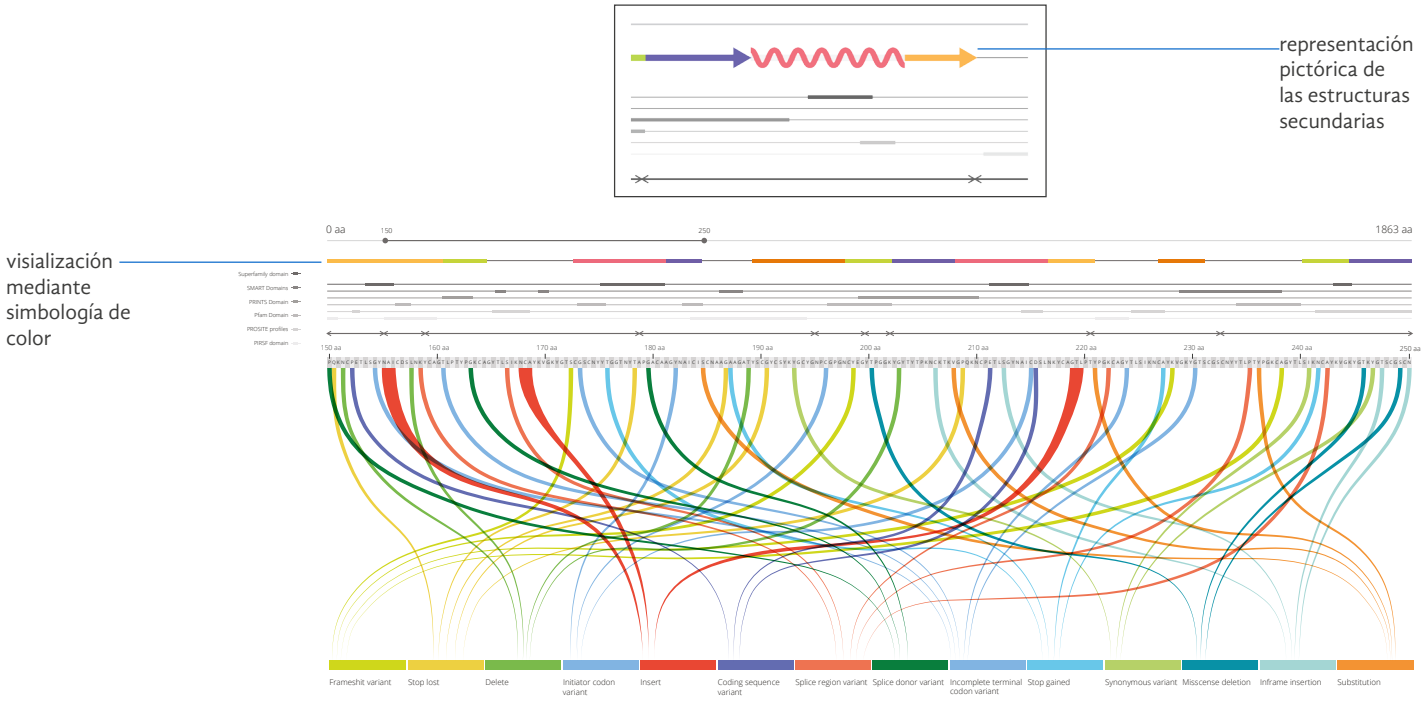


Gráfico 36. Caso 3, Parallel Coordinates.

22.3 Selección:

Con base en los casos, se evaluaron las propuestas junto con el asesor de empresa, con quien se discutieron las diversas ventajas y desventajas desde el punto de vista de usabilidad, diseño, implementación y uso. Adicionalmente, se realizó una matriz con el fin de seleccionar la propuesta que mejor cumpliera con los requisitos del proyecto.

En la matriz se inscriben dos ejes, uno con las propuestas y el otro con los criterios de selección de las mismas. Se colocó un valor del 1 al 5 (1 siendo "menos cumple con el criterio" y 5 "más cumple con el criterio"). Estos criterios se detallan a continuación:

Intuitivo: Evalúa si la herramienta es fácil de utilizar y si el usuario comprende los pasos que debe seguir para obtener los datos.

Legible: Evalúa la legibilidad de datos, textos y correlaciones. Busca determinar si las propuestas brindan la información que el usuario necesita en un tiempo reducido y con baja posibilidad de incurrir en errores.

Focus + Context: Se refiere a la posibilidad de visualizar la cadena de aminoácidos completa ("el todo") y sus correlaciones a las mutaciones, pero a la vez poder realizar acercamientos con dosificación del nivel de detalle, para brindar información más específica sin perder de vista el contexto global de la visualización.

Simple: Evalúa cual propuesta presenta mayor simplicidad en la visualización de datos, logrando con esto que el usuario interprete los datos con mayor facilidad. Dicho componente es determinante por el hecho de que el objetivo del proyecto es dar más valor a los datos y facilitar su interpretación.

Escalable: Se refiere al hecho de que la herramienta permita contar con cadenas de

aminoácidos de diferentes longitudes, y de igual forma conservar estabilidad y la legibilidad y buena comprensión de los datos. Si la herramienta perrmite diferentes escalas de datos, se va a poder adaptar a la visualización de gran número de proteínas.

A continuación se muestra la tabla con los criterios de selección que se tomaron en cuenta y las tres propuestas de diseño:

	Double Chord	Single Chord	Parallel Coordinates
Intuitivo	2	3	4
Legible	3	2	5
Focus + Context	4	4	5
Simple	2	5	4
Escalable	5	5	4
Totales	17	19	**22**

Tabla 5. Tabla de criterios de selección vs. Propuestas de diseño.

De esta forma se concluyó que la propuesta "Parallel Coordinates" es la que mejor cumple con los criterios de selección, pero se determina además que las otras herramientas cuentan con características que se pueden recatar. De esta forma la propuesta final va a contar también con elementos que se extraen del "Double Chord" y el "Single Chord".

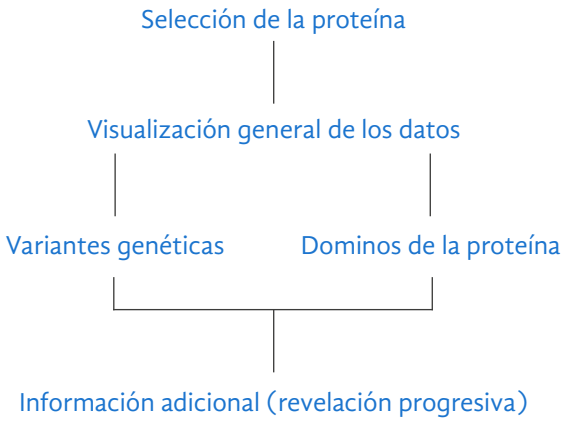
Mediante el uso de esta tabla es posible cuantificar los criterios que se cumplen en cada una de las propuestas, y elegir la que va a tener mejor desempeño en el área de la investigación de datos proteómicos.

23 Generación de la propuesta:

En este apartado se describen las decisiones que el equipo de diseño tomó en cuanto a aspectos básicos de diseño, como son el Look & Feel de la propuesta, la arquitectura de la información y el layout.

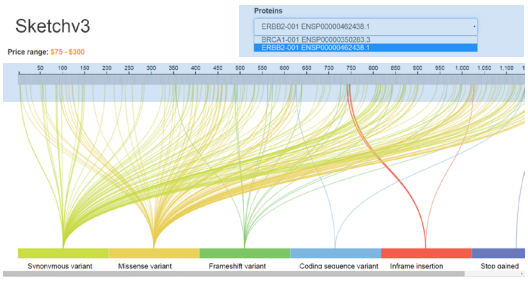
23.1 Arquitectura de la información:

La información debió organizarse y dosificarse de manera cuidadosa, con el fin de no sobrecargar al usuario cognitivamente y también para guiarlo de la forma más intuitiva posible. Como se describió en el apartado 4.2.2, existen múltiples clases de datos de distintas naturalezas que deben ser visualizados simultáneamente. A continuación se presentan los niveles de información que se diseñaron:



23.2 Interfaz e interacción:

El diseño de la interfaz de la propuesta debe responder a la arquitectura previamente diseñada, a través de un buen diseño de la distribución de los elementos, al igual que una adecuada interacción con los mismos.



Idealmente, se trabajaría con más de una proteína, por lo que se requiere de un menú desplegable o "flyout" que permita al usuario seleccionar una proteína en particular de entre la lista de disponibles.

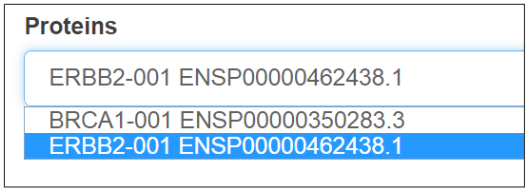


Gráfico 37. Menú desplegable.

A continuación, la vista general de los datos presentaría la cadena completa de aminoácidos. Esta vista no es particularmente funcional ya que la cantidad de información es masiva, pero el panorama completo permite al usuario poder decidir qué pasos siguientes tomar y también identificar peculiaridades en el comportamiento de los datos a través de la visualización.

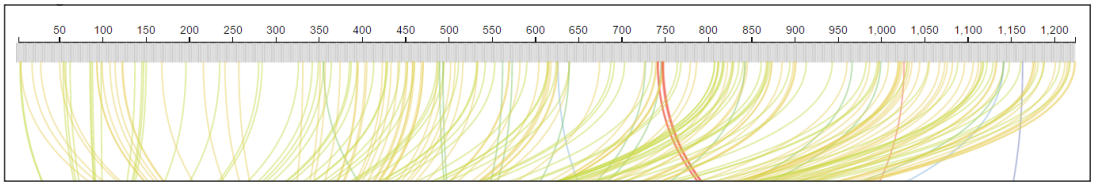
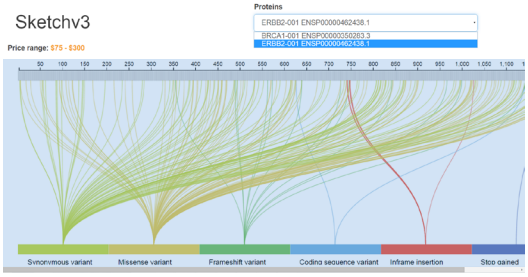
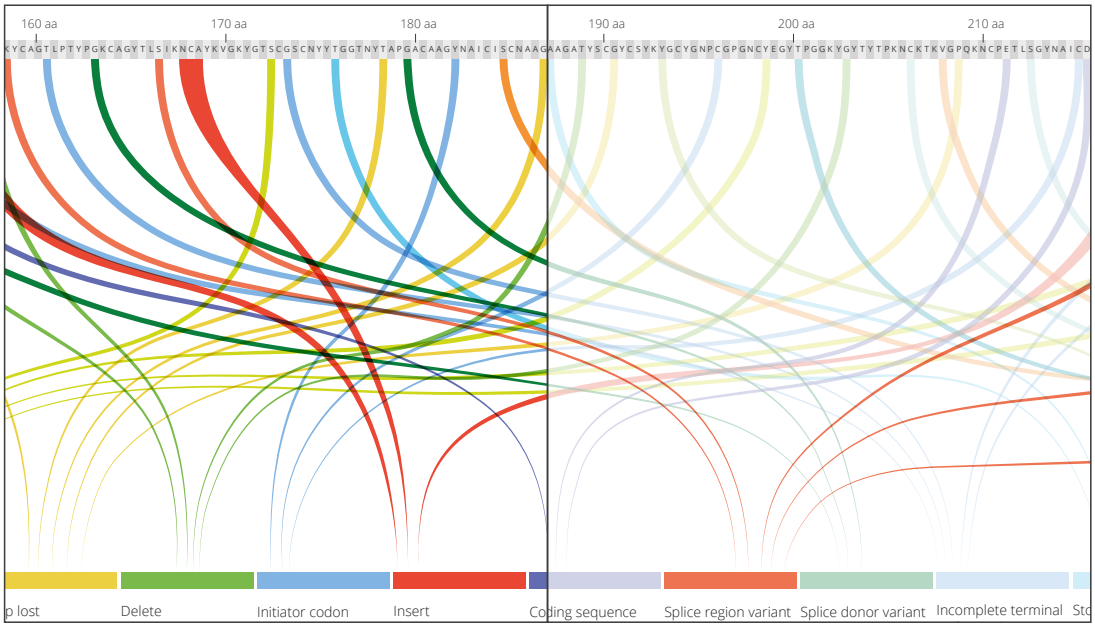


Gráfico 29. Vista general de los aminoácidos.

Las variantes genéticas de la proteína se visualizan mediante una correlación por medio de líneas entre los aminoácidos y el tipo de variante. Además, las variantes actúan como filtos que al hacer click sobre ellos, cambian la opacidad de otros elementos, para darle reelevancia a la información pertinente:





correlaciones visuales de las variantes genéticas

vista con el uso de variantes como filtros dosificadores

Gráfico 38. Correlaciones y filtros.

Además, se diseñó que la interacción con el "mouse over" sea tal que se despliege información adicional al colocar el puntero sobre alguna de las cuerdas o líneas conectoras:

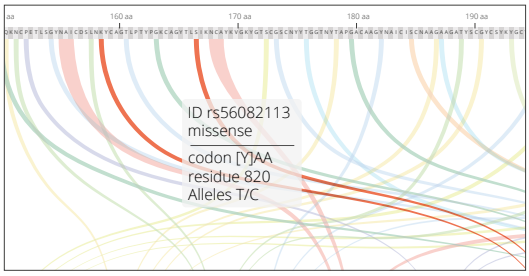
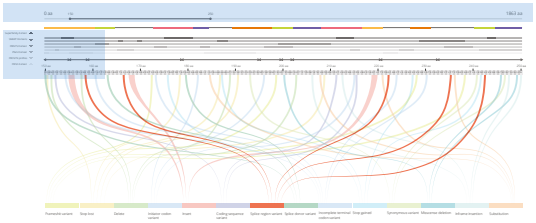


Gráfico 39. Tooltips.

Los dominos de la proteína se muestran en la parte superior, como información complementaria correlacionada, que aporta datos sobre las estructuras moleculares que forman la cadena de amioácidos:



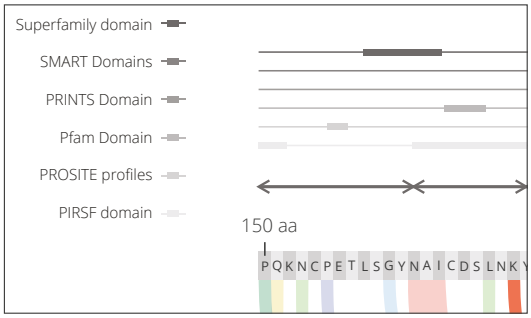


Gráfico 40. Simbología.

Finalmente, se propuso el diseño de un sistema de zoom, que permitiera dosificar los datos mediante una barra o con el uso del scroll del mouse y así mostrar sólo una porción de la cadena de amioácidos.

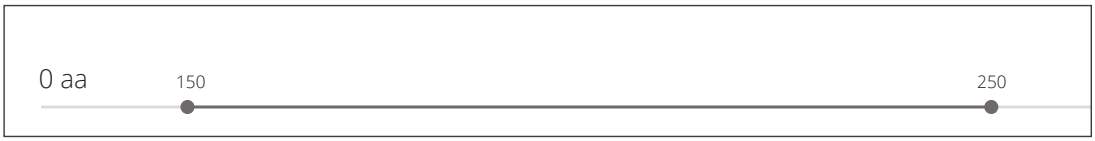


Gráfico 41. Barra de zoom.

El equipo de diseño consideró muy importante plantear la implementación de transiciones entre diversos datos; de acuerdo con estudios realizados en el VisualizationLab de la Universidad de California, Berkeley (Heer y Robertson, 2007), las transiciones animadas entre ciertos tipos de datos permiten al usuario comprender mejor el comportamiento de los mismos y “pueden mejorar significativamente la percepción gráfica”.

23.2.1 Layout:

Se estableció una diagramación que de la posibilidad observar la totalidad de los datos y correlaciones, y a la misma vez, realizar acercamientos para visualizar detalles específicos, todo esto sobre la misma retícula. La distribución de los elementos permite la navegación y obtención de información de forma sencilla e intuitiva.

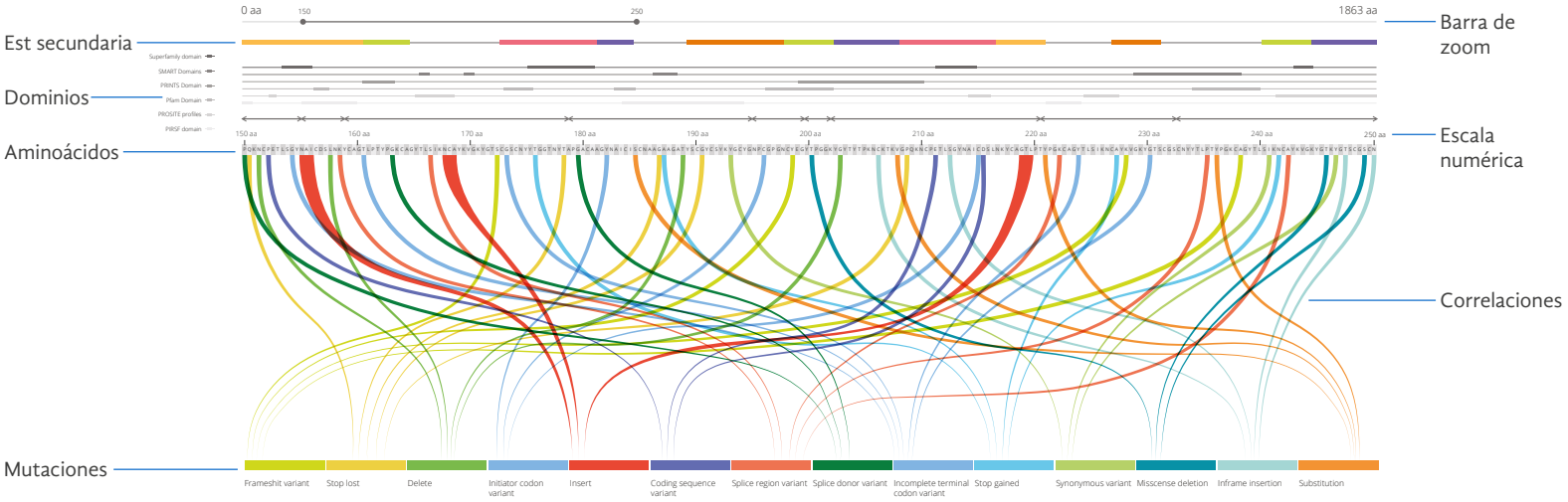


Gráfico 42. Layout de la propuesta final.

Se estableció una diagramación que de la posibilidad observar la totalidad de los datos y correlaciones, y a la misma vez, realizar acercamientos para visualizar detalles específicos, todo esto sobre la misma retícula. La distribución de los elementos permite la navegación y obtención de información de forma sencilla e intuitiva.

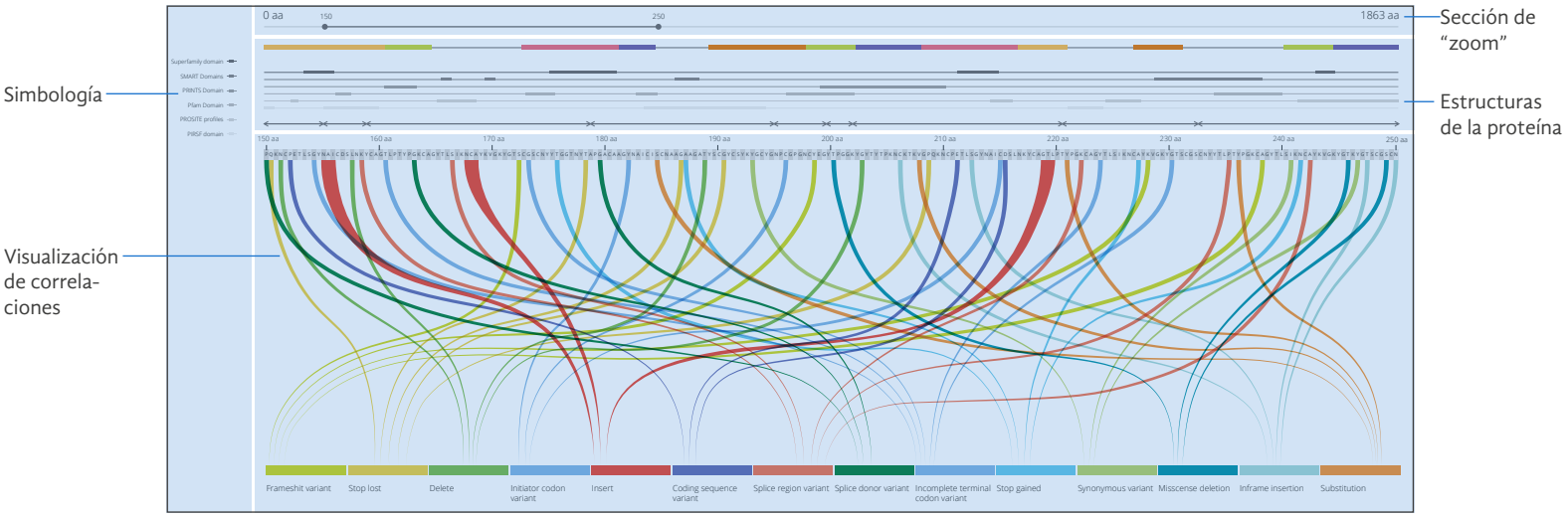


Gráfico 43. Secciones de la propuesta final.

23.2.2 Identidad:

Se diseñó una sencilla identidad corporativa para la aplicación que representara la limpieza, seriedad y estilo gráfico del proyecto con el fin de utilizarlo como encabezado en la visualización misma, al igual que en cualquier otra aplicación que requiera el uso de la marca.

Resultado final:

Construcción del isologotipo:

Construcción del isotipo:

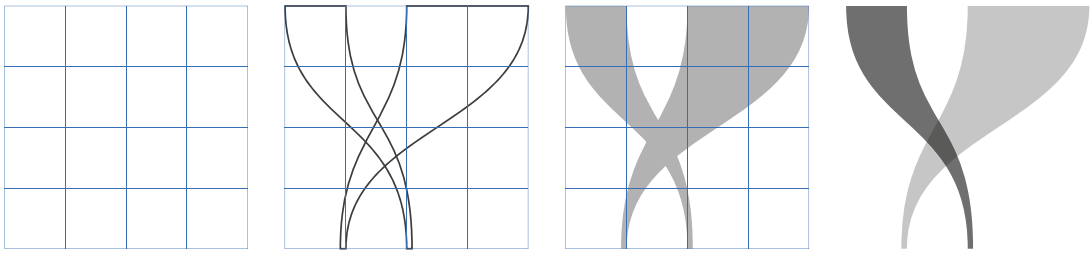


Gráfico 44. Desarrollo de la identidad para la visualización.

24 Implementación de la propuesta:

La naturaleza de los datos y la cantidad de información que se pretende procesar mediante la herramienta de visualización establecen características que deben cumplirse en la etapa de implementación, con el fin de que el usuario pueda dar un buen uso a la herramienta.

Con el fin de generar una propuesta real del diseño de una herramienta de investigación de datos genómicos y proteómicos y no únicamente una maqueta funcional que muestre un nivel más abstracto de interacción, se tomó la decisión de realizar la implementación por medio de JavaScript, utilizando la librería d3.js junto con el framework de 3VOT. Esto por cuanto si la implementación se realiza únicamente con datos falsos, es posible suponer que el diseño va a mejorar la obtención de datos por parte de los investigadores, pero no es posible comprobar si esa suposición se va a cumplir.

Inicialmente la implementación se trabajó con datos falsos y código estático, como se verá en los siguientes apartados, para luego pasar a una segunda etapa en la cual se tomó la decisión de utilizar JavaScript con el fin de trabajar con el estándar de la industria web, sumado a HTML5 y CSS3. 3VOT se utilizó para tener la posibilidad de procesar gran cantidad de datos reales en la nube, y de esta forma evaluar verdaderamente si la propuesta de diseño responde a las necesidades de los investigadores y mejora la situación actual de las bases de datos genómicas y proteómicas.

24.1 Código con datos Falsos:

Con el fin de empezar el proceso de implementación, el equipo de diseño decidió comenzar a escribir pequeños bloques de código con datasets “falsos” generados por el equipo mismo. Estos sets contaban con datos muy básicos y a una escala mucho menor, pero permitieron genera una estructura básica para después implementar funciones paramétricas más complejas. (Mayores explicaciones sobre el código que se utilizó para esta fase pueden encontrarse en la sección de anexos).

Con base en estos datos se crearon funciones que tomarían esos datos y los desplegarían en forma gráfica. Con eso se obtuvo la siguiente visualización:

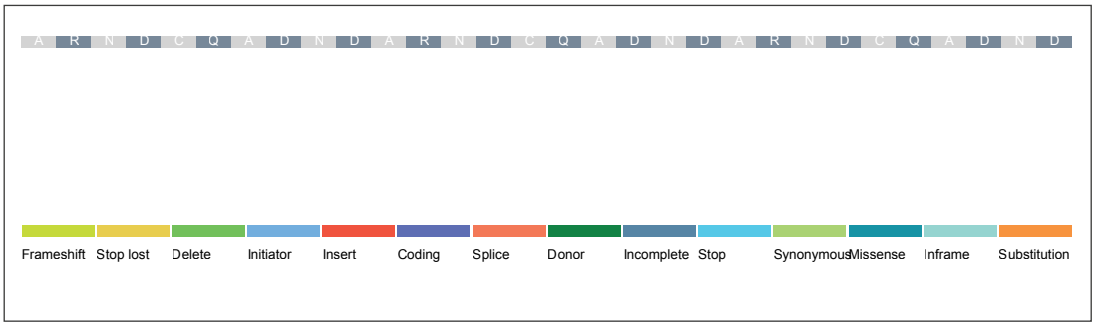


Gráfico 45. Código con datos falsos.

También se generó una tercera visualización, en la cual se mostraban bloques de color (que corresponden al tipo de mutación), junto con información adicional, como la letra del aminoácido. Esto se hizo con base en los mismos datasets:

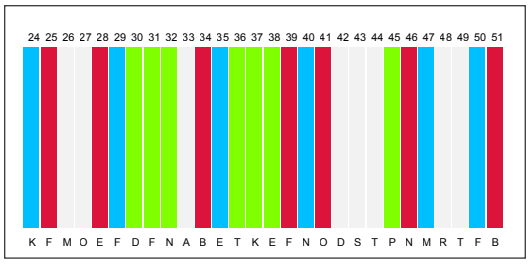


Gráfico 46. Visualización extra con datos falsos.

Finalmente los tres gráficos se implementaron en una sola vista, lo cual requería coordinar tres funciones distintas dentro del mismo código estático. Además se implementó un sistema de eje y escala horizontal para contabilizar los datos. Todo esto se realizó con ayuda de funciones de la biblioteca de d3.js.

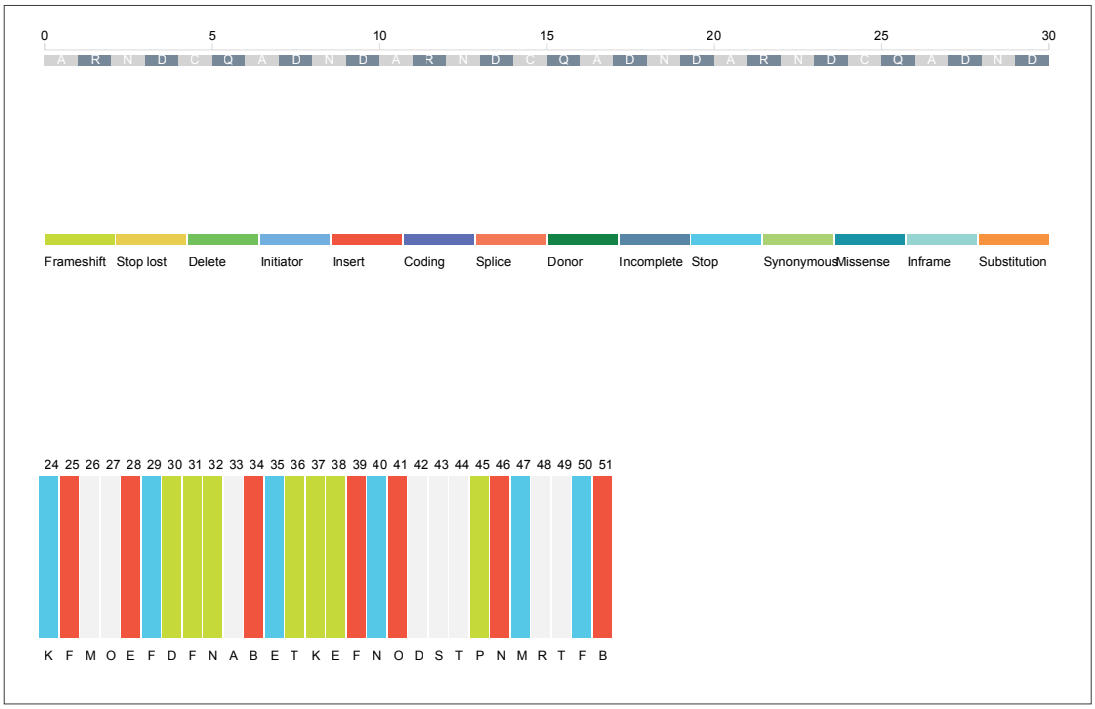


Gráfico 47. Progreso de visualización con datos falsos.

24.2 El tratamiento de los datos reales:

Una vez que hubo un primer acercamiento a la programación habiendo estudiado JavaScript y d3.js, se procedió a analizar los datos reales. Para esto se decidió trabajar primero con la proteína HER-2/ERBB2 ya que su dataset es relativamente pequeño (cerca de 1000 nodos/variantes).

Los datos se descargaron manualmente desde la base de datos de Ensembl, pero se encontraban en un formato que realmente no permitía manejar la información de manera adecuada.

El archivo fue procesado con Microsoft Excel y los datos se separaron en filas y columnas. Además se colocó una fila de títulos para cada columna.

Cuando se analizaron las tablas de datos, se concluyó que dentro del archivo existían tres grupos bien diferenciados de datos, por lo que al final se construyeron tres tablas distintas, estas tablas se describen a continuación:

Tabla "Variation":

source	datatype	startaa	endaa	variationname	allelles	class	type	alternativeresidues	codon
DbSNP	Variation	1162	1163	rs182815010	G/T	snp	stop_gained	E, *	[K]AG

Esta tabla es la que contiene más datos. Se compone por 10 columnas de datos referentes a las variaciones en la cadena de aminoácidos. Los datos de las columnas startaa y endaa son de suma importancia ya que indican la posición de la variante dentro de la cadena de aminoácidos (en este ejemplo, la variante existe entre los aminoácidos 1162 y 1163). Igualmente, el valor type indica el tipo de variante. Estos tres datos permiten generar las correlaciones entre las dos barras horizontales que ya se han comentado en las páginas anteriores de este documento.

El resto de los datos son considerados información adicional para el investigador.

Tabla 6. Variation.

Tabla “Domains”:

source	datatype	startaa	endaa	domainid	description
Superfamily	domain	674	998	SSF56112	Protein kinase-like domain

Esta tabla se conforma por seis columnas. Los datos startaa y endaa se repiten ya que se refieren a las mismas posiciones en la cadena de aminoácidos que en la tabla anterior. El dato source indica la fuente o base de datos de donde se obtuvo la información; por cada fuente se agrega una línea horizontal al gráfico, como se pudo apreciar en las propuestas de diseño. El resto de los datos son información adicional para los investigadores.

Tabla 7. Domains.

Tabla “Exons”:

source	datatype	startaa	endaa	exonid	startphase	endphase
Ensembl	Protein	1	45	ENSE00003693399	1	1

Esta tabla se conforma por siete columnas. Los exones representan las porciones de ADN codificante de la proteína (ver 4.2.1) Los datos startaa y endaa de nuevo son recurrentes pero en este caso son siempre consecutivos, por ejemplo la primera fila tiene un rango de 1-45, mientras que la segunda tendría uno de 46-n, y así sucesivamente.

Tabla 8. Exons.

25 Propuesta final:

En este apartado se describe la propuesta final que se desarrolló, profundizando en el layout, la interfaz, el diseño de experiencia y demás características de diseño que se implementaron.

25.1 Layout:

Se generó un layout con secciones claramente definidas,
menús, barras interactivas y diversos niveles de visualización.

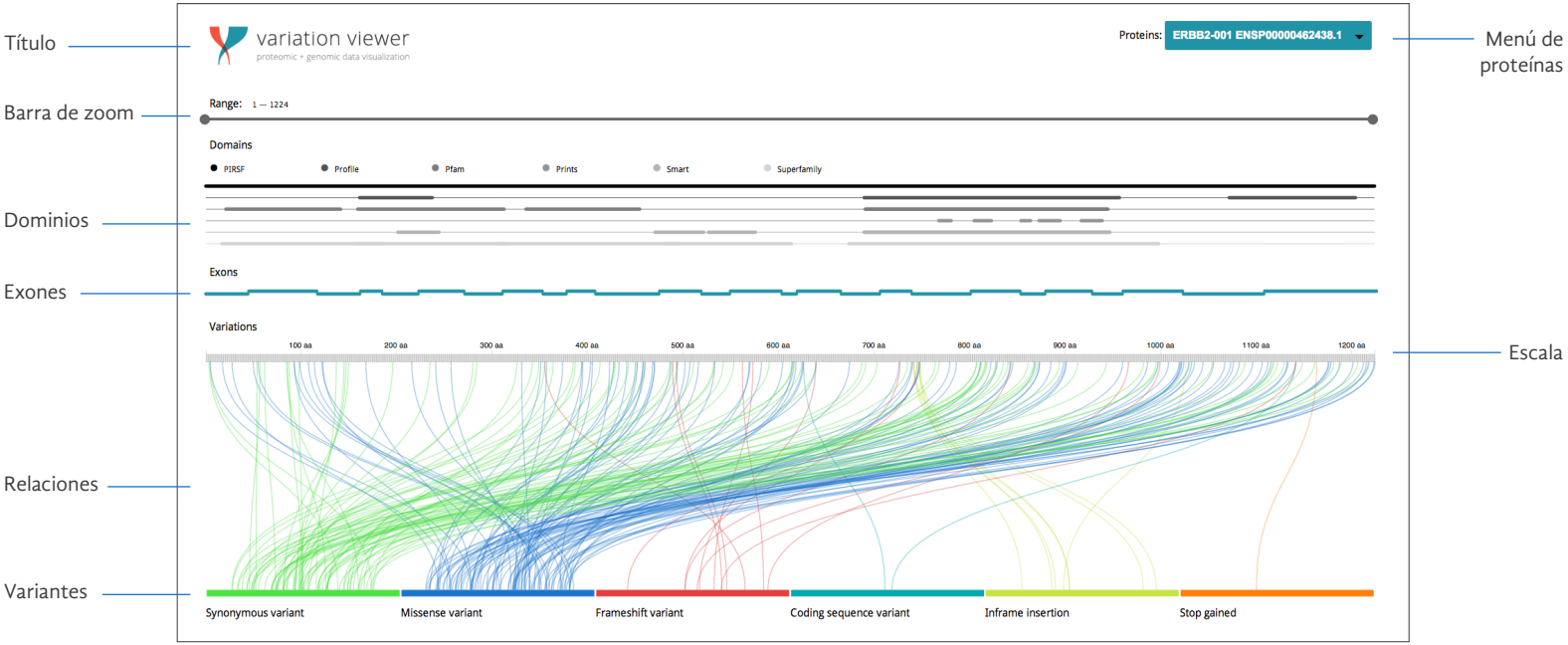


Gráfico 48. Layout final.

25.2 Secciones:

Se generaron tres secciones distintas en orden de jerarquía y uso. El alto total de la página se consideró de modo que sea visible en su totalidad en la mayoría de los display modernos.

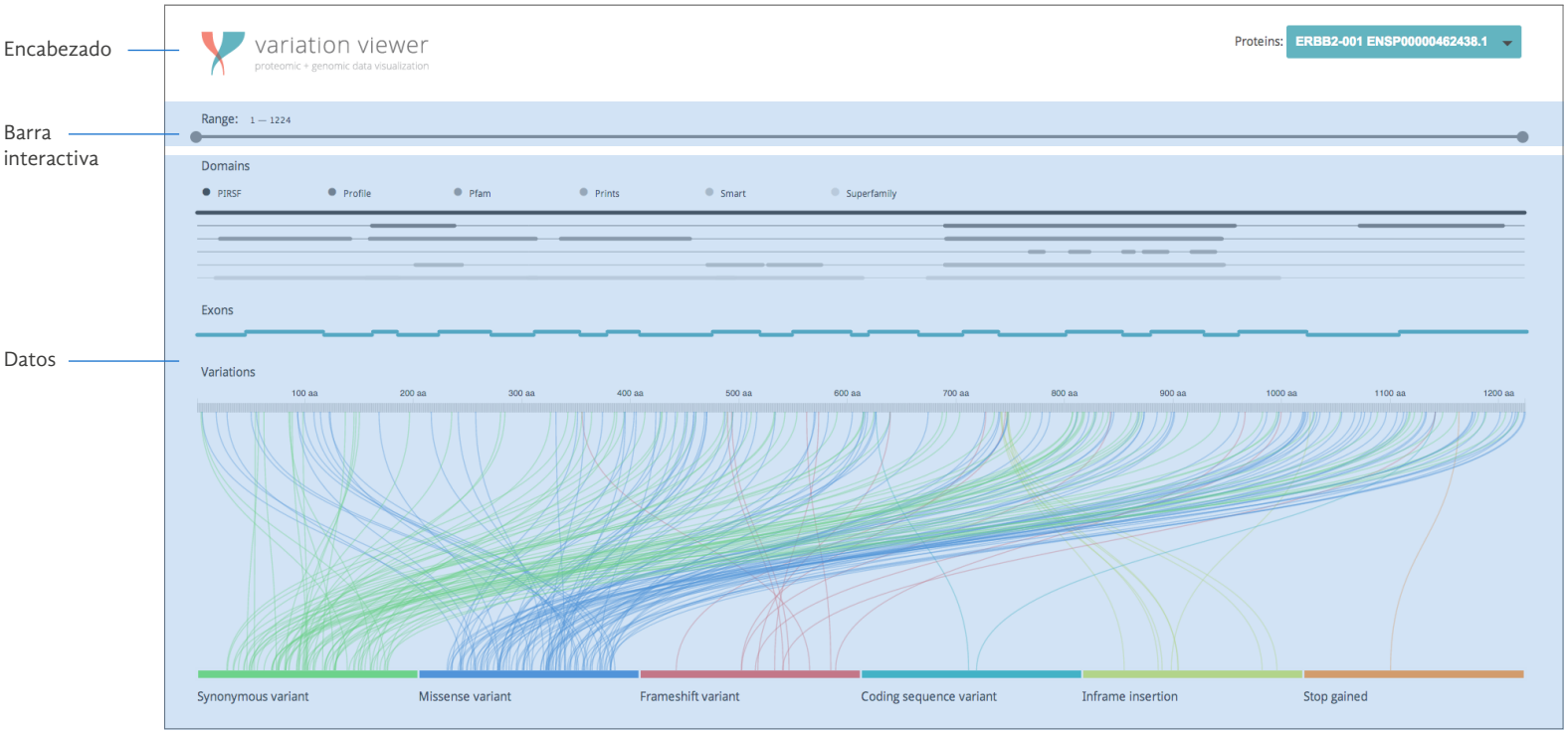
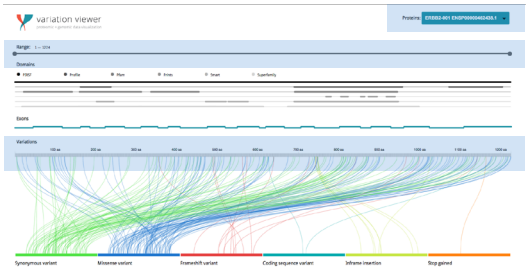


Gráfico 49. Seccones en layout final.

25.3 Interfaz de usuario:



La interfaz cuenta con un menú desplegable principal con el título “Proteinas” que permite al usuario seleccionar una de los set de datos disponibles.

El menú muestra el nombre de la proteína junto con la versión de los datos (de acuerdo con la base de datos de ensembl.org)

Gráfico 50. Menú desplegable.

El siguiente elemento de la interfaz es la barra de zoom que está compuesta por dos círculos que marcan el inicio y fin de un rango específico de datos.

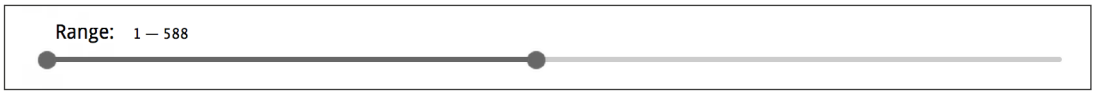
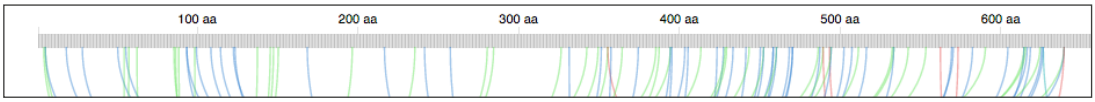


Gráfico 51. Barra de zoom final.

La escala se actualizará automáticamente para mostrar incrementos de acuerdo con el rango de datos:

Incremento de 100 en 100:



Incremento de 50 en 50:

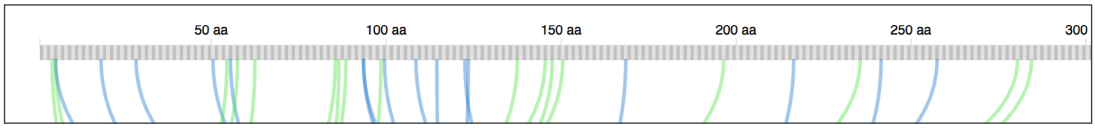
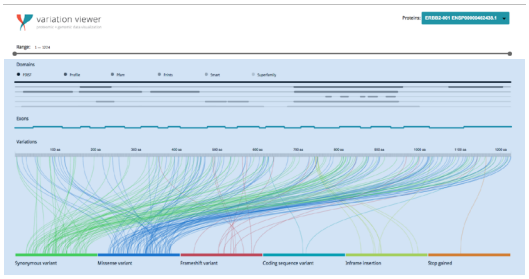


Gráfico 52. Escala dinámica final.

El área de dominios muestra las diversas fuentes de información de dominios (estructuras moleculares conocidas) de la proteína. Se fía de una simbología por color en la parte superior.



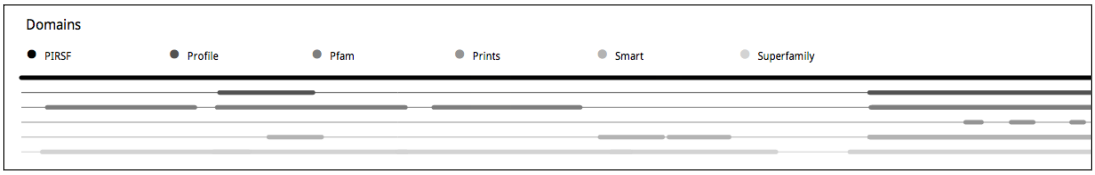


Gráfico 53. Sección de Dominios final.

Los exones, que representan el ADN codificante, se muestran en una línea de altura intercalada que representan el inicio y fin de las porciones del ADN.

Gráfico 54. Sección de Exones final.

Las líneas de colores muestran la correlación entre un aminoácido de la cadena (barra gris superior) y el tipo de variante que se ha descubierto (barra inferior)

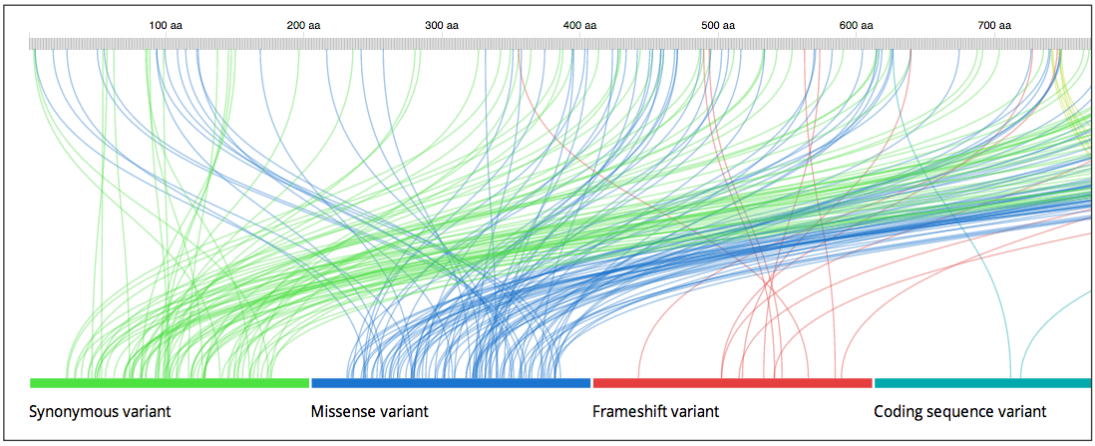


Gráfico 55. Sección de Variantes final.

La barra inferior funciona como una simbología/filtro que muestra las diferentes clases de variantes que hay en una proteína. La barra se actualiza dinámicamente para sólo mostrar aquellas variantes que aparecen y elminar datos superfluos de la visualización.

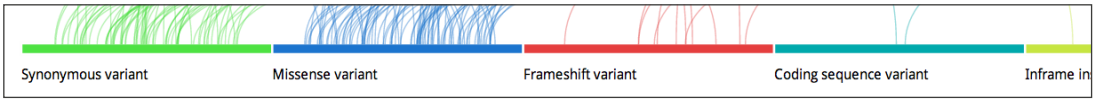


Gráfico 56. Barra inferior de simbología.

25.4 Interacción de usuario y experiencia de uso:

El menú desplegable se abre mediante un click que despliega la informacción y el usuario selecciona con otro click la proteína deseada.

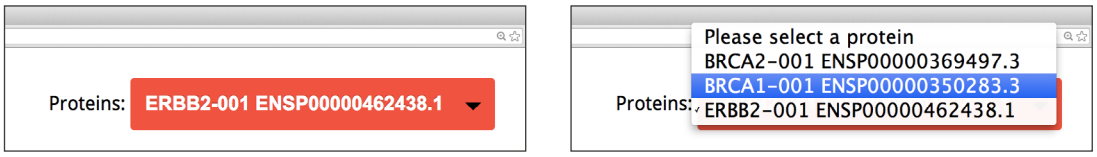


Gráfico 57. Interacción con menú desplegable.

Una vez que se selecciona una proteía el usuario tiene una vista general de todos los datos. A continuación el usuario puede utilizar la barra de zoom para seleccionar un rango menor de datos e inspeccionar con más detalle la información.

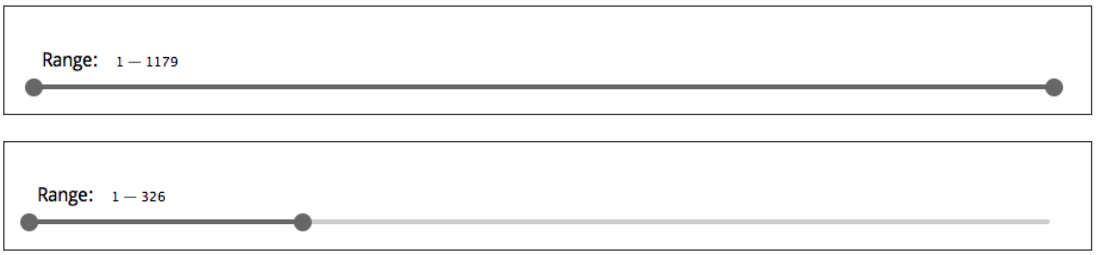


Gráfico 58. Interacción con rango dinámico.

Cuando el zoom se actualiza, se despliegan menos datos, pero con mayor definición.

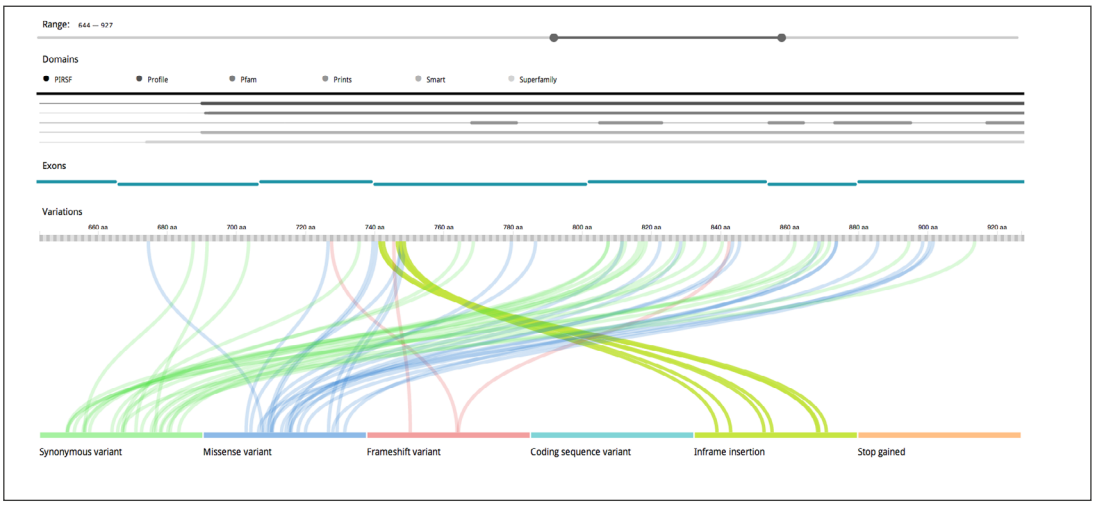


Gráfico 59. Update de datos interactivo.

Cuando el zoom se actualiza, se despliegan menos datos, pero con mayor definición. Las líneas curvas de los datos se generan de forma aleatora, lo que agrega un efecto orgánico y agradable a la vista.

Finalmente, un “mouse over” sobre una curva activa un tooltip con transiciones suaves, que permite al usuario accesar información específica de cada variante.

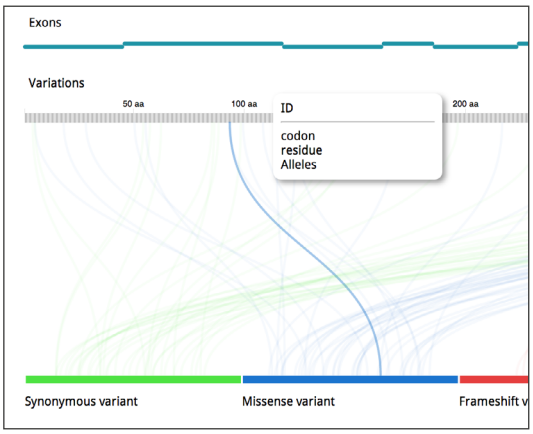


Gráfico 60. Tooltip con el hover del mouse.

25.5 Responsividad:

La aplicación final cuenta con responsividad limitada que permite ajustar el contenido al ancho de la pantalla. Esto es muy útil debido a que esto permite utilizar la aplicación en distintas plataformas y dispositivos, como computadoras, tablets, pantallas de televisor y diversos sistemas operativos.

Gráfico 61. Responsividad.

25.6 Cromática:

Se establecieron colores para ser utilizados en cada parte de la visualización, por las naturaleza de los datos, es necesario escoger una paleta cromática amplia, y categorizarla de acuerdo al tipo de datos que va a representar. En este caso se presenta la cromática utilizada para las mutaciones y para los dominios.

Mutaciones:

#4de242 #1b74ce #e53e3e #00a9ac #c6e541 #ff810

#2935cc #991010 #ffcf2d

Dominios:

#030303 #525252 #7d7d7d #949494 #b3b3b3 #d3d3d3

25.7 Cromática:

Se decidió Droid Sans, con el fin de brindar buena legibilidad al investigador y que este sea capaz de entender los datos y prevenir errores de interpretación.

"Droid Sans se ha optimizado para las interfaces de usuario y para ser cómodo para la lectura en un teléfono móvil, en los menús, en navegadores web y otros textos en pantalla". (Matteson, 2014)

Droid Sans Regular

1 2 3 4 5 6 7 8 9 0
a b c d e f g h i j k l m n ñ o p q r s t u v w x y z
A B C D E F G H I J K L M N Ñ O P Q R S T U V W X Y Z

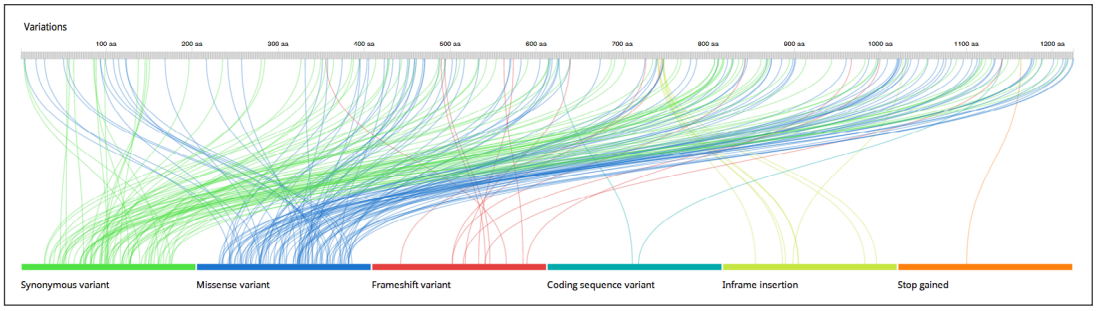
26 Validación de la propuesta:

26.1 Casos de estudio:

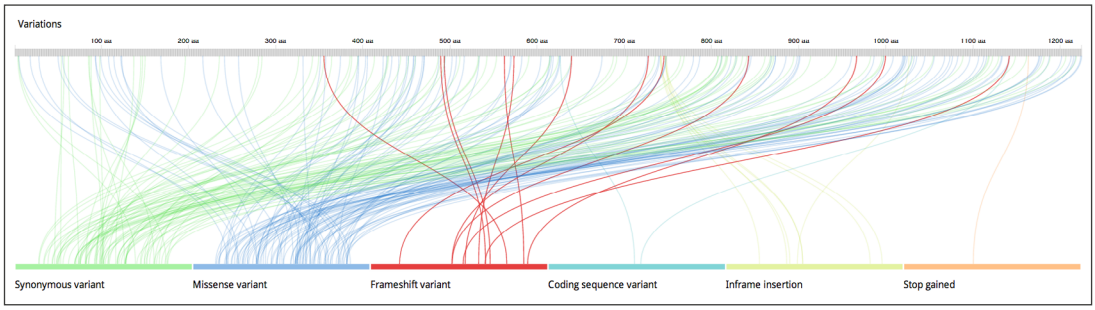
A continuación se analizan situaciones de uso de la herramienta de visualización, con el fin de determinar las acciones que el usuario debe realizar para obtener información, y de qué manera la herramienta brinda la misma.

26.1.1 Vista General

En este caso de estudio el investigador necesita visualizar la totalidad de la cadena de aminoácidos con sus variantes relacionadas.



Además de visualizar la totalidad de las correlaciones, es posible seleccionar un tipo de mutación con el fin de que éstas tengan mayor nivel de atención que las demás.

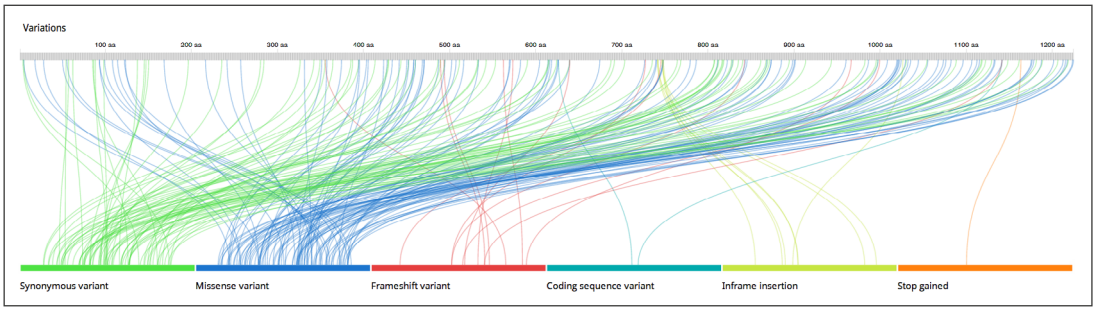


En este caso se selecciona el tipo de mutación que es representada con el color rojo, las correlaciones asociadas se hacen más visibles, mientras que las demás pierden importancia.

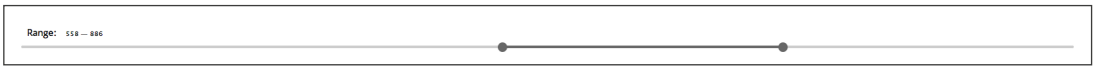
Gráfico 62. Caso de vista general.

26.1.2 Correlaciones + Tooltip

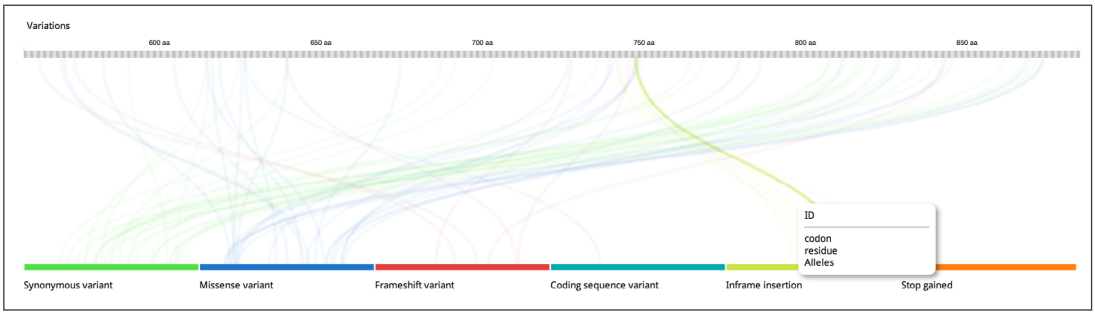
Inicialmente se presenta la totalidad de la cadena de aminoácidos con las correlaciones asociadas a las mutaciones.



Posteriormente el usuario realiza zoom con el fin de seleccionar una porción de la cadena que contenga la correlación que se necesita visualizar.



Debido a la herramienta de zoom, se acorta la cantidad de aminoácidos que se visualiza y se aumenta el grosor de la línea de correlaciones entre las mutaciones.

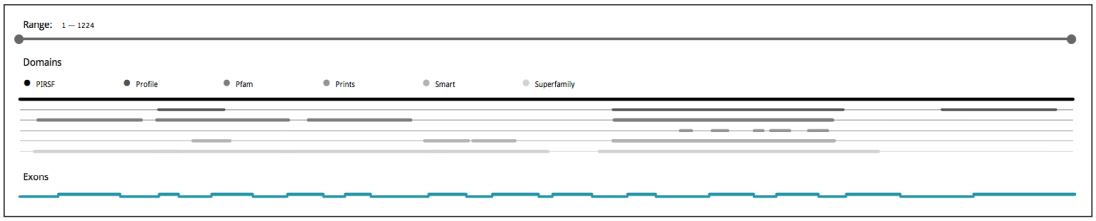


Al realizar "mouse over" sobre una correlación, ésta se hace más visible y se despliega un tooltip con información complementaria.

Gráfico 63. Caso de correlaciones + tooltip

26.1.3 Datos asociados

Además de los aminoácidos y las mutaciones, la visualización brinda información de datos asociados, como son los exones y los dominios de la proteína.



Al realizar zoom, la información de los dominios y los exones se actualiza con respecto a la cantidad de aminoácidos que se estén mostrando.

Gráfico 64. Caso de datos asociados

26.2 Resultados de la validación:

Con base en el uso de los casos previamente descritos para validar la propuesta, se encontró que existe una mejor comprensión de los datos, además de una interfaz más intuitiva y accesible para los mismo. Se encontró que la interfaz invita al usuario a interactuar con los datos por medio de clicks, arrastres y eventos de “hover” que despliegan más información.

También se detectó que la simbología estática de los Dominios es algo confusa, ya que el usuario puede pensar que es interactiva, por lo que sería deseable implementar eso en el futuro.

Adicionalmente, se sugiere implementar cambios de opacidad en estados de hover para los Exones los Dominos con el fin de tener sistemas de Focus+Context más efectivos.

26.3 Gradientes de mejora:

El proyecto brinda una visualización alternativa de datos proteómicos y genómicos con el fin de optimizar el proceso de investigación médica y científica en el área.

La manera en que se interactúa con la herramienta viene a mejorar el paradigma utilizado hasta ahora y aumenta la eficiencia, al reducir el tiempo que los investigadores dedican a la extracción e interpretación de los datos.

Más allá de eso, la herramienta permite visualizar nuevas relaciones entre los datos y revelar hallazgos que antes hubiera sido imposible de encontrar, los cuales posibilitan nuevos frentes de conocimiento para los investigadores.

26.3.1 Área social:

Incrementar la facilidad de acceso a los datos genera una mayor apropiación de la información y del conocimiento, por lo que a largo plazo puede sentar las bases para la mejor divulgación en el tema de la genómica, la proteómica y el cáncer.

26.3.2 Área económica:

Actualmente se dedican gran cantidad de recursos económicos a la investigación de proteínas y obtención de datos de las mismas.

Las investigaciones ven extendido su plazo, debido al tiempo que se requiere para extraer información de las bases de datos proteómicas existentes. Es por esto que la inversión de

entidades de salud e investigación, ya sean públicas o privadas, se podría ver reducida mediante una optimización del proceso de interpretación de datos.

26.3.3 Optimización:

La herramienta de visualización pretende lograr que los resultados de las investigaciones de cáncer de mama se alcancen con mayor rapidez y efectividad.

Además busca que los sistemas sean más rápidos y multiplataforma, que la información se distribuya más fácilmente y que sea más accesible a través del mundo.

27 Conclusiones:

Se investigó sobre las diferentes tecnologías de visualización de información, sobre los distintos paradigmas, el software, los métodos y metodologías que podían resultar útiles para aplicar a las bases de datos genómicas y proteómicas ya existentes.

Se desarrollaron tres alternativas de visualización distintas junto con las experiencias, interfazes y diversas formas de interrelacionar o correlacionar los datos. Esto con base en las necesidades detectadas a través de la investigación y la realización de entrevistas.

Se diseñó una herramienta con base en una de las propuestas de diseño que facilita la visualización de los datos y permite hacer más eficiente la generación de nuevas conclusiones por parte de los investigadores.

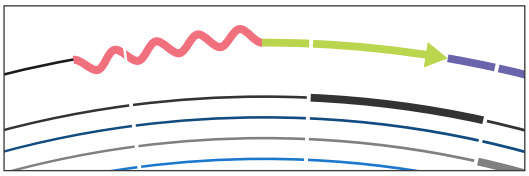
Más relevante aún, se logró un nivel de implementación lo suficientemente alto para que se pudieran comprarar "lado a lado" el estado del arte con la nueva propuesta, esto debido a que se pudieron cargar datos reales de la base de datos de ensembl.org.

Adicionalmente, se lograron visualizar los datos de una forma innovadora y diferente, que permitió de inmediato encontrar nuevas relaciones, comportamientos y patrones en los datos. Además se pudo implementar un sistema escalable que permitió visualizar distintas proteínas y afinar la visualización de modo que funcione para la mayor cantidad de casos posibles.

28 Recomendaciones:

28.1 Mostrar visualización secundaria y cambio de forma de acuerdo al zoom:

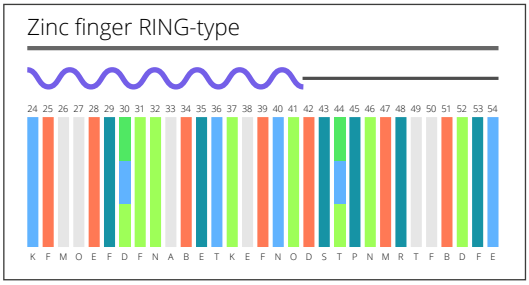
En la etapa de diseño se definió una sección específica para colocar las estructuras secundarias de la proteína y una función que permitía brindar mayor cantidad de detalle y mostrar la forma de las estructuras de forma pictórica, estos datos se pretendían importar de la base de datos PDB; esto puede considerarse como una primera oportunidad de mejora a futuro.



28.2 Correlacionar mayor cantidad de datos de otras bases de datos proteómicas y genómicas:

Durante la etapa de investigación de los datos, se encontraron características de las proteínas que sería importante visualizar, se escogieron varios datos tales como las mutaciones, los exones, los aminoácidos y los dominios de la proteína, pero otras quedaron por fuera y no fue posible visualizarlas en esta etapa del proyecto.

En una etapa futura se podría implementar mayor cantidad de datos con el fin de que la herramienta se pueda utilizar en una mayor cantidad de ámbitos de investigación médica y científica.



28.3 API para cargar proteínas automáticamente:

En este momento es posible visualizar varias proteínas en la misma plataforma, esto se logra mediante un archivo de excel administrativo que da la posibilidad al usuario de seleccionar la proteína que necesita visualizar.

En una etapa muy futura del proyecto, se podría implementar un API que se conecte a las bases de datos y que permita cargar los datos directamente desde ahí. Esta gradiente es posiblemente una de las más complejas de implementar debido a la forma en que las APIs actuales de Ensembl y PDB están configuradas.

28.4 Responsividad verdadera:

Implementar verdadera responsividad como recurso de diseño para el soporte multi display sería de gran beneficio para la democratización de la herramienta. Se sugiere el uso de un framework como Bootstrap o el uso de Media Queries de CSS

28.5 Reducir el “lagging”:

El lagging que a veces se aprecia al actualizar los datos puede ser molesto y reducir la productividad de los investigadores. Mejores técnicas programáticas y una profundización en el uso de D3 podrían reducir este problema y generar transiciones más suaves y contínuas.

29 Bibliografía:

Adobe. (2012). Adobe Edge Animate. Obtenido de Adobe TV: http://tv.adobe.com/es/product/edge-animate/

Alpízar, W. (8 de Agosto de 2014). Indagación de requerimientos con experto en genómica del cáncer. (A. Solano, & V. Alfaro, Entrevistadores)

Barranco, R. (18 de Junio de 2012). ¿Qué es Big Data? Obtenido de IBM DeveloperWorks: http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

BBC. (s.f.). FlightRisk. Obtenido de BBC Future: http://www.bbc.co.uk/bbc.com/future/bespoke/20140724-flight-risk/

Biogps. (s.f.). Expression data from healthy controls and early stage CRC patient's tumor:. Obtenido de Biogps.org: http://plugins.biogps.org/data_chart_2/#1017/1839

Biology-Online. (22 de Junio de 2008). DNA. Obtenido de Biology-Online: http://www.biology-online.org/dictionary/Dna

Biology-Online. (27 de Julio de 2008). Protein. Obtenido de Biology-Online: http://www.biology-online.org/dictionary/Protein

Bostock, M. (2013). D3 Data-Driven Documents. Obtenido de D3 Data-Driven Documents: http://d3js.org/

Broad Institute. (s.f.). GFF/GTF. Obtenido de Broad Institute: http://www.broadinstitute.org/igv/GFF

CMBI. (s.f.). http://www.cmbi.ru.nl/dssp.html. Obtenido de Centre for Molecular and Biomolecular Informatics: http://www.cmbi.ru.nl/dssp.html

DashingD3.js. (2014). Why Data Visualizations. Obtenido de DashingD3.js: https://www.dashingd3js.com/why-data-visualizations

Ensembl. (s.f.). GFF/GTF File Format - Definition and supported options. Obtenido de Ensembl Genome Browser: http://www.ensembl.org/info/website/upload/gff.html

Ensembl. (s.f.). Transcript: BRCA1-001. Obtenido de Ensembl: http://www.ensembl.org/Homo_sapiens/Transcript/ProteinSummary?db=core;g=ENSG00000012048 ;r=17:43044295-4312548 3;t=ENST00000357654

European Bioinformatics Institute. (s.f.). Ensembl Genome Browser. Obtenido de Ensembl Genome Browser: http://www.ensembl.org/

Gutiérrez, Espeleta, G., Moreno, K., & García, L. (22 de Agosto de 2014). Indagación de Requeri mientos en la Facultad de Biología de la UCR. (A. Solano, & V. Alfaro, Entrevistadores) Obtenido de Sitio Oficial del Tecnológico de Costa Rica: http://www.tec.ac.cr/sitios/Docencia/Paginas/Acreditacion.aspx

Heer, J., & Robertson, G. G. (2007). Animated Transitions in Statistical Data Graphics. Obtenido de Visualization Lab, University of California, Berkeley: http://vis.berkeley.edu/papers/animated_transitions/2007-AnimatedTransitions-InfoVis.pdf

Hernández, A. (17 de Agosto de 2014). Indagación de requerimientos con experto en bioinfor mática. (A. Solano, & V. Alfaro, Entrevistadores)

IARC. (28 de Agosto de 2014). Estimated Incidence, Mortality and Prevalence Worldwide in 2012. Obtenido de Cancer Fact Sheets: http://globocan.iarc.fr/Pages/fact_sheets_can cer.aspx

Indromics. (2014). Acerca de: Indromics. Obtenido de Indromics: http://www.indromics.com/servicios/servicios?es

Jackson, J. (10 de Octubre de 2014). Big data digest: Set your happiness gadget to bliss. Obte nido de PC World: http://www.pcworld.com/article/2825092/big-data-digest-set-your-happiness-gadget-to-bliss.html#tk.rss_all?utm_medium=referral&utm_sour ce=pulsenews

Matteson, S. (2014). Droid Sans. Obtenido de Google Fonts: http://www.google.com/fonts/specimen/Droid+Sans

Mbostock. (12 de Noviembre de 2012). Choropleth. Obtenido de Mbostock's block #4060606: http://bl.ocks.org/mbostock/4060606

MongoDB. (2014). Big Data Explain. Obtenido de MongoDB: http://www.mongodb.com/big-data-explained

National Institute of Health. (Mayo de 2012). Questions About the BRCA1 and BRCA2 Gene Study and Breast Cancer. Obtenido de National Human Genome Research Institute: http://www.genome.gov/10000940

NCBI. (s.f.). Query Input and database selection. Obtenido de NCBI: http://www.ncbi.nlm.nih.gov/blast/blastcgihelp.shtml

Orozco, A. (4 de Agosto de 2014). Indagación de requerimientos con experto en bioinformática. (A. Solano, & V. Alfaro, Entrevistadores)

Processing. (s.f.). Processing. Obtenido de Processing: http://www.processing.org/

Protein Data Bank. (s.f.). Solution structure of the BRCA1/BARD1 RING-domain heterodimer. Obtenido de Protein Data Bank: http://www.rcsb.org/pdb/explore/images.do?structureId=1JM7

RCSB. (s.f.). Protein Data Bank. Obtenido de Protein Data Bank: http://www.pdb.org/pdb/home/home.do

Robinson, R. (5 de Enero de 2014). How big is the human genome? Obtenido de Medium: https://medium.com/precision-medicine/how-big-is-the-human-genome-e90caa3409b0

Roche. (Mayo de 2013). HER2-positive breast cancer. An aggresive type of breast cancer. Obtenido de Sitio web de Roche: http://www.roche.com/med-her2-cancer.pdf

Rodríguez, R. (21 de Julio de 2014). 3VOT, Getting Started. Obtenido de GitHub: https://github.com/3vot/3vot-cli/wiki/Getting-Started

Salas Víquez, D. L. (3 de Setiembre de 2014). GBM introducirá en Costa Rica procesadores especializados en big data. Obtenido de El Financiero: http://www.elfinancierocr.com/tecnologia/GBM-big_data-procesadores_0_584941512.html

Seekshreyas. (s.f.). Beer Viz. Obtenido de seekshreyas: http://seekshreyas.com/beerviz/

Stefaner, M. (2010). Map Your Moves. Obtenido de MoritzStefaner: http://moritz.stefaner.eu/projects/map%20your%20moves/

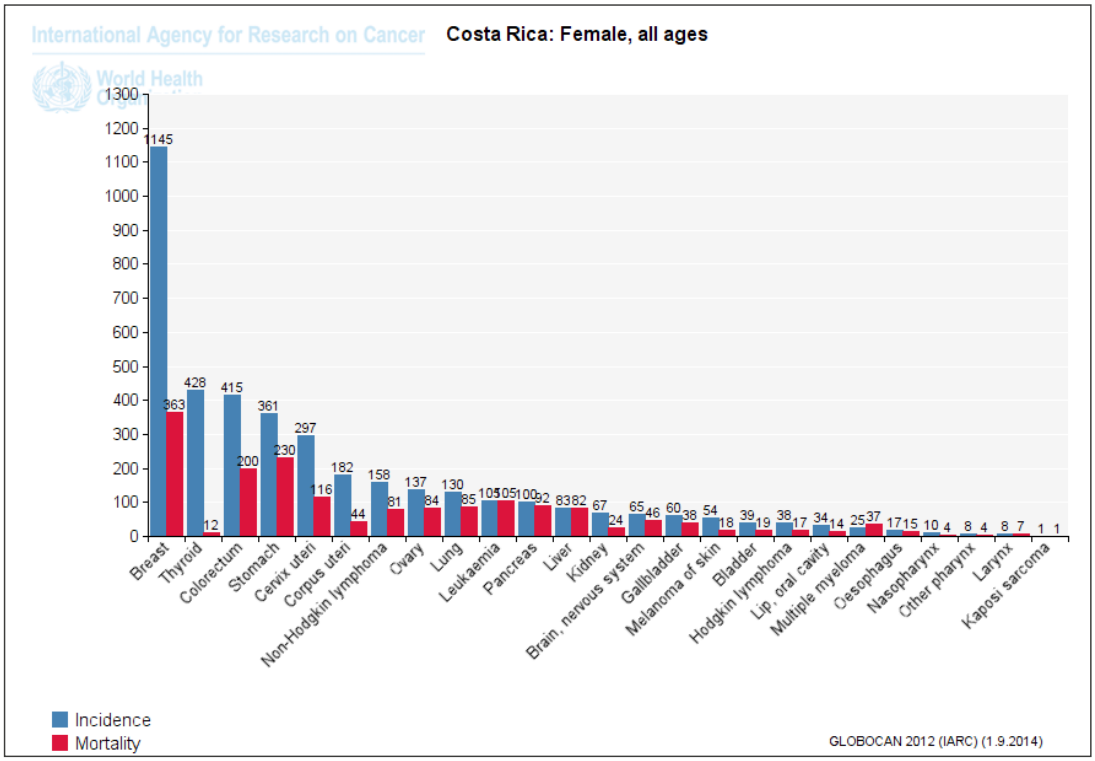
The Research Collaboratory for Structural Bioinformatics. (s.f.). Protein Data Bank. Obtenido de Protein Data Bank: http://www.rcsb.org/pdb/home/home.do

W3 Schools. (s.f.). JavaScript Introduction. Obtenido de W3 Schools: http://www.w3schools.com/js/js_intro.asp

Zoss, A. (s.f.). Common Static Visualization Types. Obtenido de Introduction to Data Visualization, Duke University: http://guides.library.duke.edu/vis_types

30 Anexos:

30.1 Incidencia del cáncer en Costa Rica:



30.2 Los datos falsos:

Como se describe en 9.1, se generaron tablas de datos falsos en formatos .csv (comma separated values) que se accesaron mediante funciones utilizando JavaScript y d3.js.

Los datos para generar una “barra de variantes” (la barra inferor) se escribieron de la forma:

```
“name”,”number”
“Frameshift”, 1
“Stop lost” , 2
“Delete”, 3
“Initiator”, 4 [...]”
```

Los datos para generar una “barra de aminoácidos” se escribieron de la forma:

```
“name”,”number”
“A”, 1
“R” , 2
“N”, 3
“D”, 4 [...]”
```

30.3 El tratamiento de los datos:

Los datos se descargaron manualmente desde la base de datos de ensembl en el formato GFF (ver 4.2.3) . Este formato en crudo realmente no permitía manejar la información de manera adecuada, como se puede ver en la siguiente imagen:

```
Ensembl Protein 854     879     exon_id=ENSE00003580476  start_phase=1   end_phase=2
Ensembl Protein 880     928     exon_id=ENSE00003650103  start_phase=2   end_phase=2
Ensembl Protein 929     960     exon_id=ENSE00003465438  start_phase=2   end_phase=1
Ensembl Protein 961     1023    exon_id=ENSE00003585058  start_phase=1   end_phase=1
Ensembl Protein 1024    1108    exon_id=ENSE00003528993  start_phase=1   end_phase=2
Ensembl Protein 1109    1226    exon_id=ENSE00002695559  start_phase=2   end_phase=1
Superfamily     domain  307     496     id=SSF52058
Superfamily     domain  18      186     id=SSF52058
Superfamily     domain  674     998     id=SSF56112      description=Protein kinase-like domai
Superfamily     domain  480     614     id=SSF57184      description=Insulin-like growth factor
Superfamily     domain  157     314     id=SSF57184      description=Insulin-like growth factor
Smart   domain  690     946     id=SM00219       description=Tyrosine-protein kinase, catalytic
Smart   domain  690     947     id=SM00220       description=Serine/threonine- /dual specificit
```

El archivo se analizó a través de Microsoft Excel y se determinó que se generaran columnas entre los espacios de los bloques de texto. Además se agregaron títulos a cada columna, llamados “fields”, los cuales facilitan poder llamar a un grupo de datos desde JavaScript.

	A	B	C	D	E	F	G
1	source	datatype	startaa	endaa	variationname	allelles	class
2	DbSNP	Variation	562	562	rs72478177	-/G	insert
3	DbSNP	Variation	842	842	rs66920285	-/T	insert
4	DbSNP	Variation	573	573	rs72125310	-/G	insert
5	DbSNP	Variation	639	639	rs67526367	-/G	insert
6	DbSNP	Variation	493	493	rs67881774	-/T	insert
7	ClinVar	Variation	746	746	rs397516979	-/TCT/TGT/TTT	insert
8	DbSNP	Variation	1026	1026	rs145292805	-/GGGGGG	insert
9	ClinVar	Variation	741	741	rs397516975	-/GCATACGTGATG	insert