Diseño de herramienta bioinformática para tabulación de secuencias de ácidos nucleicos y proteínas de bases de datos públicas para usos biomédicos y de ciencia básica.

Ruben Oswaldo Duarte Bernal

Trabajo de Grado para optar al título de Ingeniero de Sistemas

Director

PhD. Lola Xiomara Bautista Rozo

Doctor en Ciencias y Tecnologías de la Comunicación y la Información

Codirector

PhD. Francisco José Martinez Pérez

Doctor en Ciencias Biológicas

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingeniería de Sistemas e Informática

Bucaramanga

2020

Dedicatoria

Este trabajo viene dedicado para todas aquellas personas que apoyaron el desarrollo y ejecución de este trabajo de grado.

En especial reconozco la permanente presencia de Dios en mi camino de vida, en segundo lugar a mi familia por todo su apoyo a lo largo de mi carrera universitaria y a mis profesores en especial a Lola Xioamara Bautista Rozo, Francisco José Martínez Pérez y Fabio Martinez Carrillo por su guía incansable y ejemplo a seguir brindado durante este proceso.

Agradecimientos

Agradezco a mi familia por el apoyo ecónomico y moral que tuvieron para conmigo durante el desarrollo de mi carrera. También agradezco a mis amigos y compañeros por las vivencias de estos inolvidables años de universidad.

Un reconocimiento y agradecimiento importante lo realizo a mi director de trabajo de grado, por dedicar su tiempo, experiencia y conocimiento en la guía de mi proyecto a mi coodirector por toda su contribución a nivel personal en este proyecto, al laboratorio Genómica de Celomados GCL por mostrarme lo apasionante de la ciencia desde mis inicios en la carrera y al grupo BivL2ab por su guía en la iniciativa y dirección de este proyecto.

Un agradecimiento final a Anderson Cepeda por su colaboración en mis estudios.

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Glosario

Ácidos nucléicos: Son moléculas orgánicas compuestas de bases púricas y pirimídicas, en las que se almacena toda la información de los genes que codifican los ARN necesarios para la síntesis de proteínas y regulación de procesos para su metabolismo, crecimiento y muerte programada. En células eucariontes se encuentra en el núcleo, la mitocondria o los plástidos .

Codón: Conjunto de tres nucleótidos que indican la posición de un aminoaciltRNA requeridos en la síntesis de proteínas.

Oligo nucleotidos: Cadena de 2 a 100 nucleótidos.

Registro GenBank: Documento en formato de texto plano, donde los laboratorios reportan las secuencias de ácidos nucleicos obtenidas por procesos de clonación o secuenciación de nueva generación.

Secuencia: Cadena de caracteres codificados que representa la posición de ácidos nucleicos o aminoácidos.

Resumen

Título: Diseño de herramienta bioinformática para tabulación de secuencias de ácidos nucleicos y proteínas de bases de datos públicas para usos biomédicos y de ciencia básica. [*]

Autor: Ruben Oswaldo Duarte Bernal [**]

Palabras Clave: Bioinformática, Secuencias, Biología, Clasificación.

Descripción: Este proyecto presenta el desarrollo de una herramienta bioinformática para la tabulación de secuencias de ácidos nucleicos. Esta herramienta nace de la investigación desarrollada en el Laboratorio de Genómica de Celomados (GCL) en el diagnóstico del virus de influenza AH1N1 y en el estudio del uso preferencial de codones en Dopamina para posibles aplicaciones en psicosis, esquizofrenia y enfermedad de Parkinson. En el desarrollo de la investigación realizada, se abordó la etapa denominada preparación de datos, en la cual se tabulaban cada uno de los registros descargados del Genbank del National Center for Biotechnology Information (NCBI) con el fin de depurar la base de datos de registros incompletos o no congruentes. Este proceso demoraba el desarrollo de los proyectos de investigación elevando los tiempos de ejecución y los costos. Por lo mencionado anteriormente, se desarrolla una herramienta informática que permite agilizar esta etapa, enfocada en particular en la tabulación de las secuencias, usando un script de Python que le permitiría al investigador cargar los conjuntos de registros descargados sin tabular (en formato de texto plano), tomando estos registros y extrayendo las características del primer registro como criterios de búsqueda en el conjunto de datos, posteriormente extrayendo estas características en un tiempo menor de cada uno de los registros, construyendo una tabla de excel amigable con el investigador.

---

[*] Trabajo de grado

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería de Sistemas e Informática Director: Lola Xiomara Bautista Rozo. Codirector: Francisco José Martínez Pérez

Abstract

Title: Design of a bioinformatic tool for the clasification of nucelic acid and protein sequences from public data bases for biomedical and basic cience uses. *

Author: Ruben Oswaldo Duarte Bernal **

Keywords: Bioinformatics, Sequences, Biology and Classification.

Description: This work presents the development of a bioinformatic tool for the clasification of nucleic acid and protein sequences. This tool was borned from the research at the Genomic of Celomathe Laboratory (GCL) in the diagnose of the Influenza AH1N1 virus and the study of the preferential codon usage in Dopamin for possibles aplications on Psicosis, Schizophrenia and Parkinson disease. The first phase was the data preparation phase, in which the researchers clasified each downloaded record from the GenBank of the National Center for Biotechnology Information (NCBI) manually, to clean the database from uncompleted, or non congruent records. This process is time consuming and expensive. As mentioned above, we developed an informatic tool that agilizes this phase during the research, focus in the sequence tabulation using a Python script that allows the researcher to load the plain text records sets, the script takes this records and pulls out the characteristics of the first record as search criteria in the data set, then uses this criteria as search parameters for each one of the records and build an excel spreadsheet in less time friendly to the researcher.

---

* Bachelor Thesis

** Faculty of Physical-Mechanical Engineering. Systems Engineering and Informatics Department Advisor: Lola Xiomara Bautista Rozo. Co-Advisor: Francisco José Martínez Pérez.

Introducción

Una de las aplicaciones de la bioinformática es en las ciencias biológicas y de salud (Jimenez-Gutierrez et al 2016), ella contribuye a realizar estudios en diversos campos de estas áreas de investigación, por medio de herramientas informáticas que permiten el análisis de una grán cantidad de datos, dentro de los que se cuentan las secuencias de ácidos nucleícos y de proteínas. Este proyecto, se focalizará en el campo de la genómica, la cual usa secuencias de ácidos nucleicos sea bien ácido desoxirribonucleico (ADN) y/o ácido ribonucleico (ARN) como fuente de datos. Ellas son reportadas por los investigadores en bases de datos públicas, como es el GenBank del Nacional Center for Biotechnology Information (NCBI) y con ello se comparten los hallazgos científicos Jimenez-Gutierrez et al. (2016).

En el Laboratorio de Genómica de Celomados (GCL) adscrito a los Grupos de Investigación de Microbiología y Genética y al grupo de Grupo de investigación de Cómputo Avanzado y a gran escala (CAGE) junto con el Grupo de Investigación Biomedical Imaging vision and Learning Laboratory (BIVL2ab) de la Universidad Industrial de Santander (UIS) se han realizado proyectos de investigación que emplean la bioinformática como herramienta para obtener información relevante de genes, ARNs y proteínas para el desarrollo de la investigación básica y aplicada.

Ejemplo de lo anterior son un par de proyectos que hicieron uso de la bioinformática como herramienta fundamental. El primero estudió la correlación entre el uso preferencial de codones y

la estructura secundaria de receptores de dopamina en invertebrados y vertebrados para establecer su evolución y posible aplicación en psicosis, esquizofrenia y la enfermedad de Parkinson. Además contribuyó a un nuevo sistema de secuenciación genómica de nueva generación cuya patente esta en proceso de validación Barrios Hernández et al. (2018).

El segundo correspondió al diagnóstico del virus de Influenza AH1N1 para establecer las causas de cepas atípicas que causaron la muerte a algunos pacientes durante la pandemia del año 2009. La aplicación de los procedimientos bioinformáticos permitió establecer el patrón de mutación de los genes empleados en el diagnóstico y con ello se diseñaron nuevas moléculas; un juego de ellas le salvó la vida a 150 pacientes y generó una propuesta de patente de la UIS y otras dos instituciones participantes. Gonzáles Barrios et al. (2017).

Una parte fundamental de ambos proyectos fue la obtención de la información de los respectivos genes a estudiar a partir de las secuencias reportadas en la base de datos pública GenBank del NCBI. De ella se descargó las secuencias, que fueron preseleccionadas para su posterior clasificación y verificación, esta etapa se definió como “Preparación de Datos”. Es de resaltar que es la más importante, relevante y fundamental para ambos proyectos y cualquiera que se quiera realiza por los grupos de investigación u otros; debido a que una incorrecta selección de una o más secuencias genera la invalidez del modelo de estudio, lo que se traduce en: pérdida de tiempo, insumos, esfuerzo de los depuradores y sobretodo retraso en la entrega de los resultados que generarían el producto final, sea bien: un diagnóstico médico para pacientes, publicación científica, una patente o informes a los entes financiadores de los proyectos.

La verificación de las secuencias en estos estudios, fue realizada de manera manual, lo cual fue extensa en el caso del estudio del virus de Influenza A H1N1 dada la cantidad de secuencias, retrasando el desarrollo del proyecto, elevando los costos de ejecución y los tiempos de entrega.

Por lo anterior en este proyecto se plantea el desarrollo de una herramienta computacional por medio de la bioinformática como solución, para optimizar el procesado de la información relevante de las secuencias reportadas en la base de datos del GenBank, que sustituirá la manera convencional, el tiempo de selección y depuración para la obtención de bases de datos de ácidos nucleicos.

El contenido de este proyecto de grado, esta dado de la siguiente forma: Inicia con el capítulo de objetivos, donde encontramos presentación de los objetivos establecidos para el proyecto entre los dos grupos de trabajo. En el segundo capítulo, estado del arte, se muestra el contexto del entorno de trabajo del investigador y las herramientas que se usan en la actualidad. Luego se aborda la metodología, donde se describe el modelo usado para desarrollar la herramienta y las etapas del desarrollo. Para el capítulo de resultados se exponen los resultados obtenidos de la herramienta y su validación. En el capítulo recomendaciones se muestran las recomendaciones que surgieron con base al trabajo realizado. A continuación, se presenta el trabajo futuro, capítulo que sugiere una serie de desarrollos futuros como trabajo complementario a la herramienta y finalmente se presentan en el capítulo conclusiones, las conclusiones obtenidas con base en el proyecto realizado y sus resultados.

1. Objetivos

1.1. Objetivo general

Crear una herramienta para agilizar la clasificación de secuencias de ácidos nucleicos de la base de datos pública GenBank.

1.2. Objetivos específicos

Establecer los parámetros del formato de descarga de los registros del GenBank, para que sean compatibles con la herramienta propuesta;

Crear una interfaz amigable con el usuario, que permita la fácil interpretación de los datos generados por la herramienta;

Validar los resultados, comparando los tiempos de cómputo de la herramienta propuesta con los desarrollados por investigadores del laboratorio GCL.

2. Estado del arte

En la actualidad, existen diferentes tipos de herramientas bioinformáticas, de acuerdo a las necesidades que se han planteando desde el campo de la investigación Chou (2009) Priyadarshi (2014), en estas, se pueden mencionar las siguientes áreas como focos principales de aplicación de estas herramientas entre otras: estudios filogenéticos, ensamblaje de genomas, alineamiento de secuencias, generación de oligo nucleótidos, visores de secuencias, generadores de estructura secundaria de proteínas, visualización y edición de estructuras secundarias, control de poblaciones. En este proyecto nos enfocaremos en las herramientas relacionadas con la obtención de información a partir de los registros de las secuencias obtenidos del GenBank. Estas herramientas, usan como entrada, secuencias de ADN y/o ARN para el procesamiento y generación de análisis bioinformáticos de las mismas. A continuación se presentan las herramientas que se usan para realizar la lectura y edición de secuencias.

2.1. Herramientas bioinformáticas

2.1.1. Bioedit. BioEdit Hall (1999), esta herramienta permite al usuario, cargar secuencias en distintos formatos, permite realizar alineamientos, árboles filogenéticos incluyendo varios modelos, consultas a bases de datos externas, visualización de secuencias en bloque y edición de alineamientos. Al momento de realizar análisis de las secuencias, la herramienta permite usar distintos códigos de color en la visualización de bloque para caracterizar, cada uno de los caracteres en las secuencias y determinar zonas cambiantes o regiones con cambios puntuales. Por ejemplo

en la figura 1 se cargó el conjunto de datos de secuencias del virus de Influenza A H1N1 correspondientes al año 2000 el cual contiene 122 secuencias, la herramienta, colorea por defecto los distintos nucleotidos con el fin de facilitar la lectura de las secuencias.

Aunque este software acepta los archivos en formato GenBank con extensión ".gb", no permite extraer las características de la secuencia en el registro, solo la secuencia en formato Fasta y los análisis que pueda generar a partir de esta.

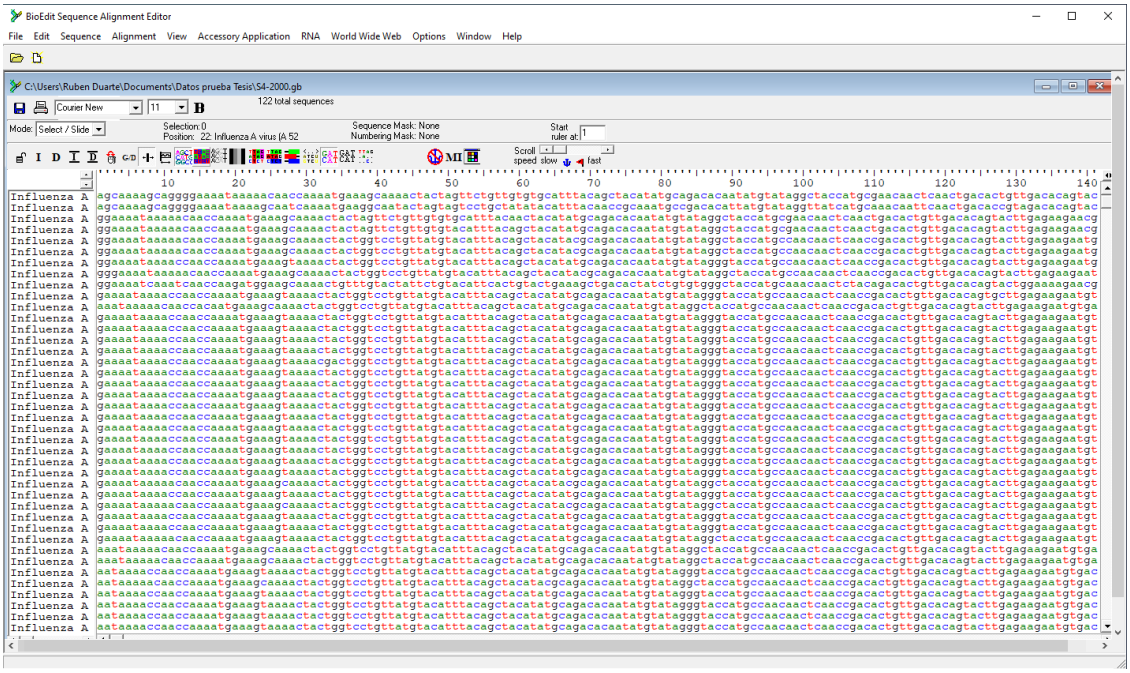

Figura 1. Captura del software BioEdit corriendo en Windows 10

2.1.2. Ugene. UGENE Okonechnikov et al. (2012) es un paquete de herramientas de distintas aplicaciones. En estas, podemos realizar búsquedas de patrones, búsquedas de cuadros abiertos de lectura, búsquedas de sitios de restricción, búsquedas de secuencias similares en el GenBank,

aplicación de modelos estadísticos para la generación de análisis filogenéticos, realizar alineamientos de secuencias, generar modelos de trabajo para análisis, diseñar oligo nucleótidos, generación de estructura secundaria y previsualización de la misma. Al igual que la anterior, esta herramienta permite al usuario usar códigos de color para caracterizar cada carácter de la secuencia. Por ejemplo en la figura 2 se cargó el mismo archivo correspondiente a las secuencia del virus de Influenza A H1N1 conformado por 122 secuencias. Como podemos observar, el programa indentifica regiones de la secuencia y las resalta usando códigos de color, al igual que la herramienta anterior, acepta archivos en formato GenBank con extension ".gb"pero no permite extraer las características de la secuencia en el registro, solo la secuencia en formato Fasta y los análisis que puede generar a partir de esta.

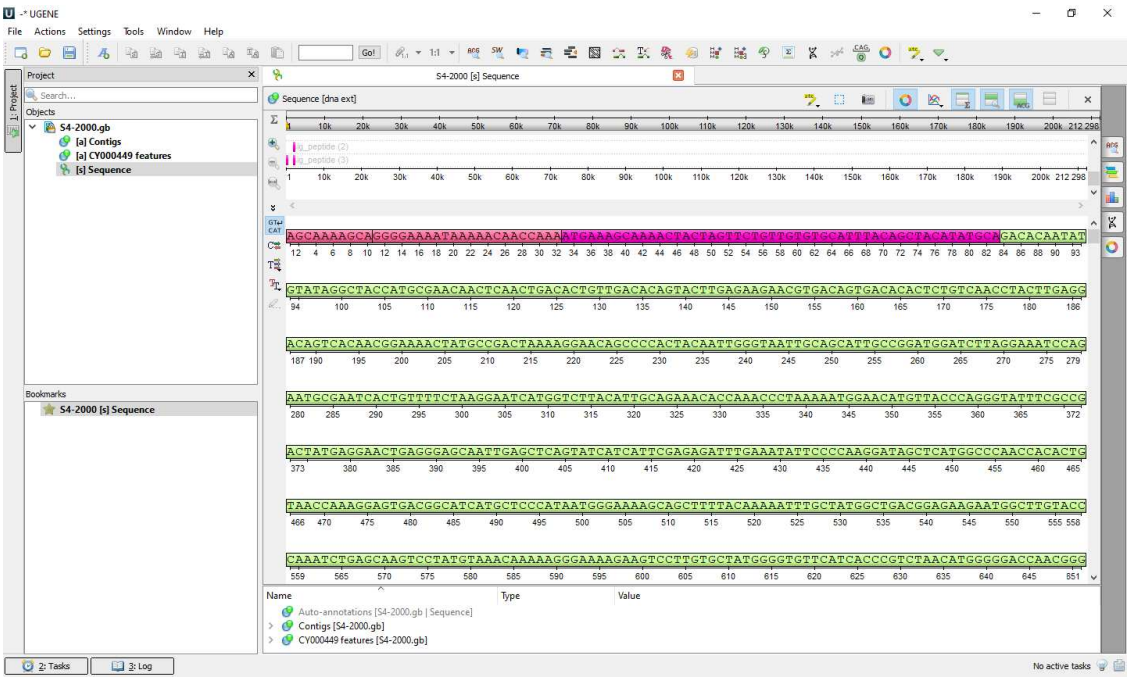

Figura 2. Captura del software UGENE en MacOSX

2.1.3. BioPython. BioPython Cock et al. (2009) es una recopilación de rutinas, conocida como librería, que permiten a un desarrollador interactuar con datos provenientes del GenBank, este paquete está escrito en el lenguaje de programación Python PythonCoreTeam (2015). La librería BioPython no constituye una aplicación para usuario final, por lo tanto, es necesaria la implementación en un desarrollo complementario, para permitir la interacción del usuario final con el paquete. A diferencia de las dos aplicaciones mostradas previamente, BioPython, permite obtener toda la información reportada en la secuencia como el tipo de secuencia, el lugar de obtención, la fecha, la traducción, entre otros y no solo la secuencia de ácidos nucleicos.

2.2. Contexto de trabajo

Durante el desarrollo de proyectos de investigación en el GCL, se realiza una preparación de datos previa al desarrollo y/o ejecución de modelos para obtener información relevante de lo datos. Durante esta fase, las secuencias nucleotídicas fueron seleccionadas de los resultados arrojados por las consultas al GenBank como se puede observar en la figura 3, cada uno de estos registros se tenía que consultar de manera individual para verificar su relevancia biológica, lo que conllevó a la prolongación de los proyectos hasta que la etapa fuese concluida.

Ejemplo de ello, es la base de datos que se generó del Virus de Influenza A H1N1 para su diagnóstico. En un principio se emplearon todas las secuencias reportadas en el GenBank hasta el año 2010 es decir se seleccionaron manualmente más de 45.000 secuencias, este volumen de

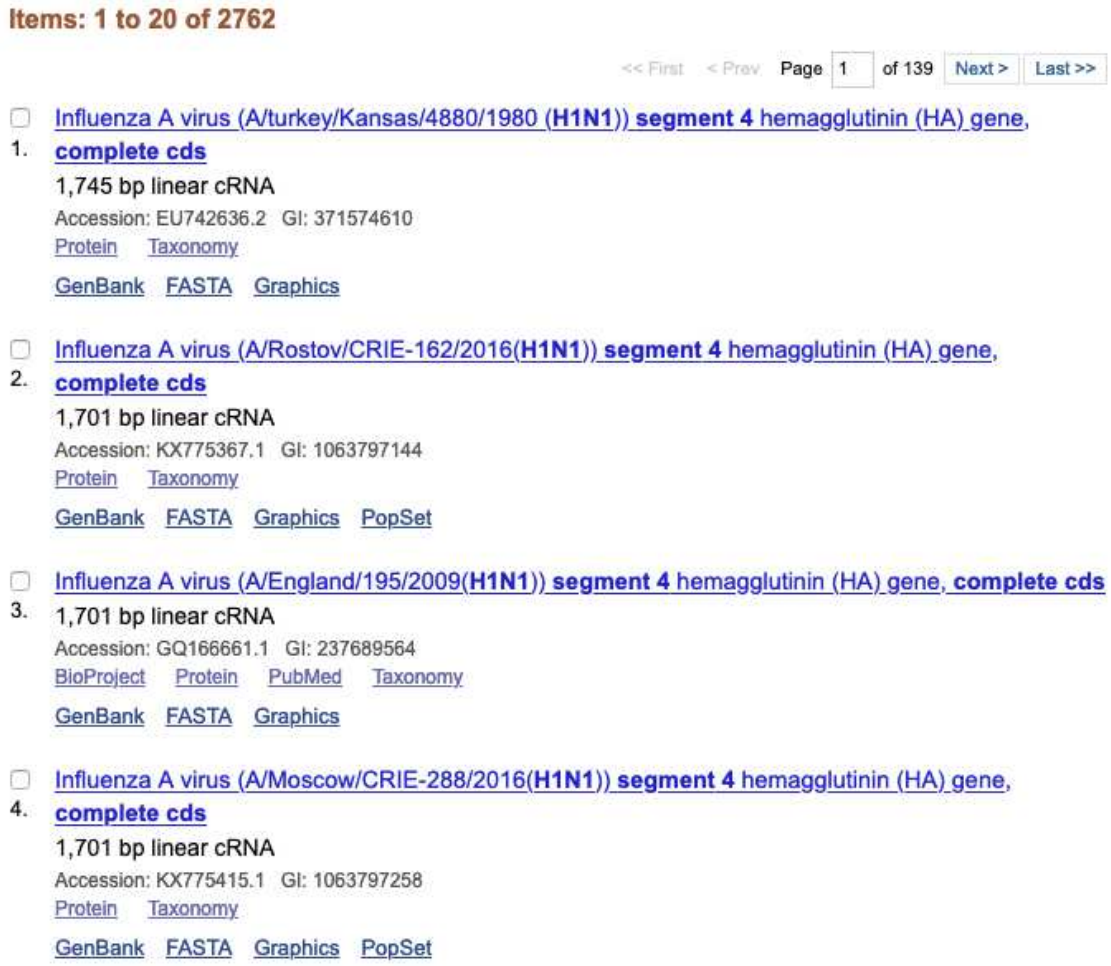

Figura 3. Consulta al GenBank

datos conllevó a que la preparación de datos, requiriese más de 12 personas, 3 de ellas contratadas específicamente para esto, durante un periodo de 15 meses para obtener 37708 secuencias de los ocho segmentos del genoma del Virus de Influenza A H1N1. Posteriormente, por la aplicación de modelos evolutivos se obtuvo la base de datos final de 31835 la cual llevó un año y medio de su depuración por medio de 14 personas. Después fue necesaria la edición de cada una de las secuencias para poder ser empleadas en paquetes bioinformáticos que solicitan que ellas tengan el formato FASTA con una nomenclatura específica en el caso de la súper computadora GUANE.

Este tipo de depuración continúa en el grupo de investigación debido a que no existe una herramienta computacional que permita agilizar el proceso. Por lo tanto, se evidenció la necesidad de la generación de una herramienta computacional que optimizara la etapa de la preparación de los datos, para disminuir los tiempos y hacer operante la obtención, selección y procesamiento de datos, para con ello prevenir la inviabilidad de cualquiera de los proyectos, además de la disminución de: el número de personas, equipos de computación, tiempos de ejecución de la etapa y costos económicos.

En algunos campos de la biología como la genética y el estudio de poblaciones, entre otros, se usa la aplicación de herramientas informáticas para estudiar comportamientos, patrones evolutivos y realizar distintos estudios. Estas herramientas se definen dentro del campo de la bioinformática, las cuales usan como datos de entrada, las secuencias de ácidos nucleicos y/o de proteínas que se recopilan de las bases de datos públicas como el GenBank del NCBI.

En estos estudios se descargan los registros de cada una de las secuencias reportadas con respecto al tema de estudio; estos registros se pueden descargar en distintos formatos, donde los más usados son los formatos FASTA (figura 4) y GenBank (figura 5), ambos son archivos de texto plano, formato FASTA (figura 4), consiste en la representación de la secuencia de ácidos nucleicos indicada por el autor, usualmente se usa la secuencia de codones, donde cada codón está conformado por 3 bases, la secuencia de codones es la parte del gen que codifica a una cadena de aminoácidos y posteriormente a una proteína, considerando que las secuencias tienen secciones que codifican y otras que no, estas llevan un signo > como indicador del inicio del archivo seguido por el nombre designado para la secuencia y la secuencia de ácidos en las siguientes líneas.

```
>KX802662.1:17-256 Synthetic construct clone BS22073 Akh-RA (Akh) gene, complete
cds
ATGAATCCCAAGAGCGAAGTCCTCATTGCAGCCGTGCTCTTCATGCTGCTGGCCTGCGTCCAGTGTCAAT
TGACCTTCTCGCCGGATTGGGGCAAGCGTTCGGTGGGCGGAGCTGGTCCTGGAACCTTTTTCGAGACACA
GCAGGGCAACTGCAAGACCTCCAACGAAATGCTGCTCGAGATCTTCCGCTTCGTGCAATCTCAGGCACAG
CTCTTTCTGGACTGCAAGCACCGCGAGTAG
```

Figura 4. Formato Fasta.

El formato GenBank full (figura 5) , contiene toda la información relacionada al registro como por ejemplo: Nombre, Localizador, Identificador de Gen, Palabras clave, Revista (en caso de ser publicado como artículo), Especie, Autores, Taxonomía, Origen, Fechas, Laboratorio, Tamaño, Secuencia de codones, Traducción, entre otras dependiendo de la información reportada por el autor.

Cuando se realiza la consulta en la base de datos, esta devuelve una lista de registros (figura 3), que coinciden con las palabras usadas como criterio de búsqueda, en la que cada resultado muestra una versión resumida del registro. Cada uno de estos resultados se selecciona uno a uno conforme al objeto de estudio para la posterior construcción de la base de datos que se usa para el estudio.

Este resultado, muestra el registro en su formato inicial GenBank. Para obtener el segundo tipo de archivo, en formato Fasta, desde la vista del archivo en formato GenBank se puede escoger la secuencia de codones, o la secuencias completa en formato Fasta y se procede, a descargar el archivo. En los trabajos realizados y en curso por el GCL, se usan bases de datos que varían su tamaño en función del objeto de estudio, las secuencias reportadas, y el alcance del proyecto, variando desde algunos registros, hasta miles de registros.

```
LOCUS       KX802662                272 bp    DNA     linear   SYN 21-SEP-2016
DEFINITION  Synthetic construct clone BS22073 Akh-RA (Akh) gene, complete cds.
ACCESSION   KX802662
VERSION     KX802662.1
KEYWORDS    .
SOURCE      synthetic construct
  ORGANISM  synthetic construct
            other sequences; artificial sequences.
REFERENCE   1  (bases 1 to 272)
  AUTHORS   Carlson,J., Booth,B., Frise,E., Park,S., Wan,K., Yu,C. and
            Celniker,S.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-AUG-2016) Berkeley Drosophila Genome Project,
            Lawrence Berkeley National Laboratory, One Cyclotron Road,
            Berkeley, CA 94720, USA
COMMENT     Sequence submitted by:
            Berkeley Drosophila Genome Project
            Lawrence Berkeley National Laboratory
            Berkeley, CA 94720
            For further information about this clone please visit our Web
            site (http://www.fruitfly.org) or send email to cdna@fruitfly.org.
            This clone is distributed from the Drosophila Genomic Resource
            Center (http://dgrc.cgb.indiana.edu).
FEATURES             Location/Qualifiers
     source          1..272
                     /organism="synthetic construct"
                     /mol_type="other DNA"
                     /db_xref="taxon:32630"
                     /clone="BS22073"
                     /focus
     source          17..256
                     /organism="Drosophila melanogaster"
                     /mol_type="other DNA"
                     /strain="y; cn bw sp"
                     /db_xref="taxon:7227"
                     /note="construct generated from ORF of Drosophila
                     melanogaster clone IP01764 with stop codon retained"
     misc_feature    <1..8
                     /note="loxP site"
     misc_feature    11..16
                     /note="SalI site"
     gene            17..256
                     /gene="Akh"
                     /note="ORF for Akh-RA"
                     /db_xref="FLYBASE:FBgn0004552"
     CDS             17..256
                     /gene="Akh"
                     /codon_start=1
                     /transl_table=11
                     /product="Akh-RA"
                     /protein_id="AOQ11599.1"
                     /db_xref="FLYBASE:FBgn0004552"
                     /translation="MNPKSEVLIAAVLFMLLACVQCQLTFSPDWGKRSVGGAGPGTFF
                     ETQQGNCKTSNEMLLEIFRFVQSQAQLFLDCKHRE"
     misc_feature    257..262
                     /note="HindIII site"
ORIGIN
        1 gaagttatca gtcgacatga atcccaagag cgaagtcctc attgcagccg tgctcttcat
       61 gctgctggcc tgcgtccagt gtcaattgac cttctcgccg gattggggca agcgttcggt
      121 gggcggagct ggtcctggaa ccttttcga gacacagcag ggcaactgca agacctccaa
      181 cgaaatgctg ctcgagatct tccgcttcgt gcaatctcag gcacagctct ttctggactg
      241 caagcaccgc gagtagaagc tttctagacc at
//
```

Figura 5. Formato GenBank full.

3. Metodología

En el presente capítulo se describe la metodología usada en el desarrollo de la herramienta, desde la selección del modelo de desarrollo, hasta la herramienta en ejecución.

3.1. Planeación

Durante la ejecución de los proyectos Estudio de correlación de codones y estructura secundaria en receptores de dopamina y el diagnóstico del virus de Influenza A H1N1 por parte del GCL, se observó que la fase de preparación de datos, era una tarea fundamental para la ejecución de los mismos, dado que la mala práctica en esta fase podría conllevar a la invalidez del modelo.

Además se observó que esta tarea se realizaba de forma manual, consultando en la base de datos del NCBI, seleccionando, validando y descargando cada secuencia y consolidando las secuencias descargadas en archivos de formato Fasta con las secuencias descargadas, como se puede observar el proceso en la figura 6.

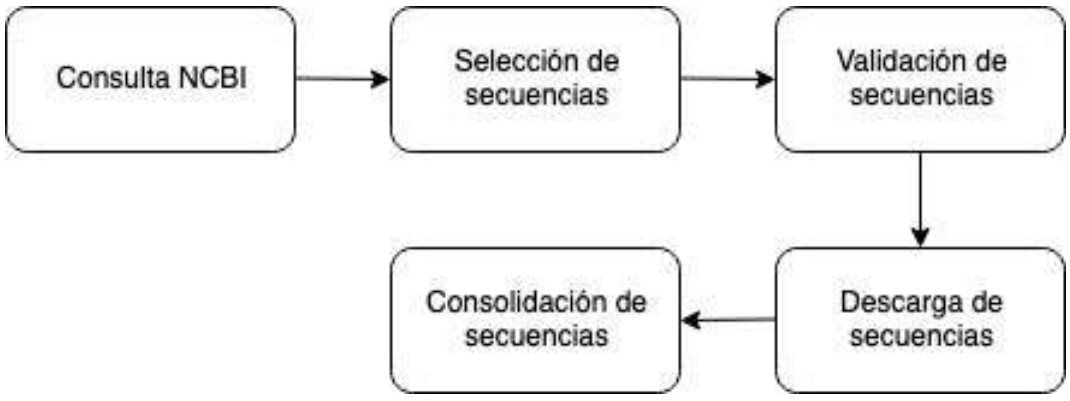

Figura 6. Diagrama de preparación de datos

Las etapas de la preparación de datos, se describen a continuación:

Consulta NCBI: Consiste en consultar la base de datos del NBI disponible en www.ncbi.nlm.nih.gov, introduciendo palabras clave como criterios de búsqueda para obtener los registros que coinciden con estas.

Selección de secuencias: Dadas las diferencias entre los equipos de trabajo y metodologías usadas para la obtención de las secuencias, los registros consignados en la base de datos, suelen diferir en sus datos y contienen errores de caracterización. Dadas esta situación, a cada registro se le verifica que efectivamente corresponda al objeto de estudio con el título y el resumen que muestra la consulta.

Validación de secuencias: Posterior a la selección de la secuencia, el registro se consulta individualmente en formato Genbank, se verifica que la secuencia de codones esté completa y con base en las características, se realiza una verificación más profunda y detallada para saber si el registro coincide con el criterio de búsqueda.

Descarga de secuencias: Se selecciona la secuencia o secuencias de codones correspondiente, según el tipo de secuencia y el objeto de estudio, se selecciona el formato Fasta y se copia directamente de la página o se descarga el archivo para consolidarlo en un archivo de texto local.

Consolidación de secuencias: Con las secuencias descargadas, se ensamblan archivos de texto con las secuencias clasificadas por distintos criterios como lo son año, segmento, país,

entre otros. Copiando canda una de las secuencias en un archivo de texto.

Dadas las múltiples disciplinas que conformaban el equipo de trabajo, se necesitaban realizar pruebas con un modelo que permitiera realizar pruebas tempranas y correcciones durante el desarrollo de la herramienta. Esto conllevó a escoger el modelo de desarrollo de prototipo evolutivo Sherrell (2013) Pressman (2010), el cual está comprendido por las siguientes cinco etapas:

Requisitos del sistema: Se establece un canal de comunicación con el cliente o usuario final interesado en el software, se establecen los objetivos y alcances para identificar los requerimientos del software.

Especificaciones de requisitos del prototipo: Con base en la comunicación, los objetivos y alcance establecidos en la etapa anterior, se diseña un plan de para desarrollar la herramienta de acuerdo con el ciclo de vida del prototipado. Este plan se es entregado al usuario para ser evaluado y recopilar apreciaciones con el fin de tener la retroalimentación necesaria, para validar requerimientos y dar paso a la siguiente etapa.

Diseño del prototipo: Se diseñan las vías de interacción entre el usuario y la herramienta, para este proyecto, la interfaz gráfica de usuario.

Implementación del prototipo y prueba del prototipo: Se ejecuta el plan diseñado para desarrollar la herramienta computacional.

Evaluación y comunicación para refinamiento: Durante esta etapa, se entrega la versión prototipo de la herramienta al usuario con el objetivo de realizar validaciones de los requeri-

mientos, objetivos y alcance. Los resultados de esta validación, son usados como retroalimentación para dar lugar a la siguiente versión de prototipo de la herramienta, hasta lograr los objetivos.

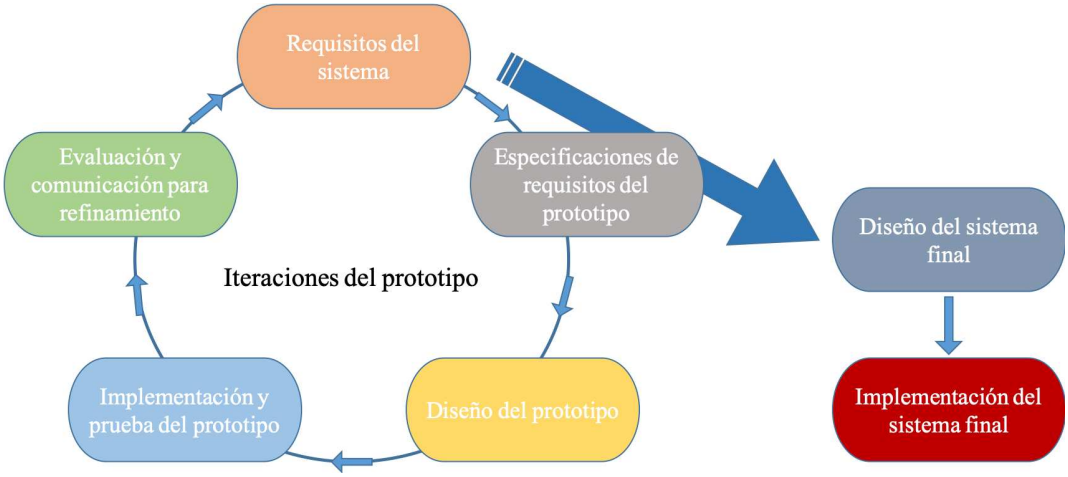

Figura 7. Ciclo de vida del prototipado evolutivo. Adaptado de "Modelo de Ciclo de Vida de Prototipado Evolutivo", por I Cornejo, 2017, basado en Pressman (2010), disponible en: https://www.slideshare.net/IvnCornejo/modelo-de-ciclo-de-vida-de-prototipado-evolutivo-79872674

Se plantearon 2 actores principales, el primer grupo conformado por los investigadores del GCL y su director en calidad de codirector de este proyecto, como usuarios finales de la herramienta y el segundo conformado por el autor de este proyecto y la directora del mismo como equipo desarrollador. Este modelo se adaptaba a las necesidades del desarrollo puesto que la herramienta debía ser probada por usuarios finales en su etapa de desarrollo, por lo cual todas las versiones inicialmente, eran orientadas a usuario final, en cada versión se realizaba la respectiva verificación y esta se usaba como retroalimentación para el desarrollo de la siguiente versión.

Durante la etapa de comunicación, se determinó que filtrar los resultados de la consulta era una de las tareas que más ocupaba tiempo en la clasificación de secuencias, junto con la extracción y tabulación de las mismas en formato Fasta. Por lo cual se decidió tomar todos los resultados de la consulta como bloque de datos inicial para la herramienta, así evitando realizar la selección de secuencias una a una. Estos bloques se descargaban en formato GenBank (Full), el cual es un archivo de texto plano con todos los registros de la consulta con extensión ".gb".

Ya que los usuarios debían interactuar desde la primera versión de la herramienta, se definió desarrollar la herramienta en el lenguaje de programación Python PythonCoreTeam (2015), usando el framework Jupyter Pérez and Granger (2007) para permitir a los usuarios interactuar con la herramienta sin tener que diseñar una interfaz gráfica en una etapa temprana. Este modelo de trabajo, permitió desarrollar todas las versiones de BioDataToolkit hasta la final, cada una de estas versiones fueron validadas por el equipo del GCL.

Dado que se necesitaba poder interactuar de forma dinámica con la información obtenida de las secuencias, se determinó que la salida óptima para los usuarios finales se entregaría en una tabla de Excel, permitiendo a los usuarios tener una mejor visualización de los datos y realizar filtrados de datos, con el fin de identificar las secuencias que serían de relevancia para el estudio. Para usar estas secuencias preseleccionadas en los diferentes paquetes bioinformáticos, es necesario que estén consolidadas en un archivo en formato Fasta, por lo tanto la hoja de Excel permitiría listar las secuencias en formato Fasta, facilitando la extracción de las mismas.

El trabajo conjunto entre los equipos durante la fase inicial, conllevó a establecer los siguientes

requerimientos para el desarrollo de la herramienta.

Requerimiento
La herramienta debe aceptar conjuntos de secuencias de gran tamaño, del orden de las 10.000 secuencias.
La herramienta debe diferenciar las características de cada secuencia.
La herramienta debe obtener la secuencia de codones del registro descargado.
La herramienta debe entregar los resultados en una tabla que permita filtrar los mismos.
La herramienta debe codificar las secuencias obtenidas de acuerdo a la nomenclatura diseñada por el GCL.
La herramienta debe ser compatible con los principales sistemas operativos.

Tabla 1
Tabla de requerimientos

3.2. Desarrollo

Con base en lo establecido en la sección anterior, se desarrolló la primera versión de la herramienta siguiento las fases establecidas para cada versión. Se usó el lenguaje de programación Python con el framework de notebooks de Jupyter, para facilitar la interacción entre el usuario y la herramienta. Usando el archivo descargado resultado de la consulta a la base de datos GenBank como entrada de la herramienta, se planteó en la primera versión, una rutina que recorría este archivo línea por línea extrayendo las características especificadas previamente en el código, usando una estructura de datos de un diccionario de listas, donde cada una de las llaves del diccionario, correspondía a una característica de la secuencia y su valor la lista que almacenaba cada uno de los valores recopilados de la lectura del archivo, posteriormente este diccionario se escribía en un archivo de salida de Excel usando la librería xlsxwriter.

El usuario debía cargar el archivo descargado del GenBank al directorio de ejecución del notebook, en el código, debía escribir el nombre del archivo de entrada y el nombre del archivo de salida y ejecutar la rutina.

3.2.1. Version 1.0. En la versión 1.0 se mostró que la herramienta escribía correctamente el archivo de Excel, ubicando la información en columnas, pero incluía la primera palabra de cada campo, por lo tanto se perdía información valiosa, la cual se pretendía obtener. Usando esta información, se dio lugar a la segunda version de BioDataToolkit.

3.2.2. Version 2.0. Con la retroalimentación dada por el GCL, se decide implementar el uso de la librería BioPython Cock et al. (2009) en la segunda versión, dada que esta permitía obtener los valores de cada campo completos y optimizaba la herramienta, al tratar la secuencia como un objeto donde los atributos se representaban como diccionarios siendo la característica la llave y el atributo el valor,y no como un texto plano.

La versión 2.0 escribía en el archivo de Excel de salida, los valores de cada atributo completos. Tras la validación realizada por el GCL, se encontró que habían secuencias con más de una secuencia de codones puesto que esta característica presentaba información encapsulada, lo cual era un problema para la herramienta dado que esta solo tomaba el primero que encontraba.

3.2.3. Version 3.0. Con base en esto se da lugar a la tercera versión (3.0) de BioDataToolkit, en la cual se usa una estructura de datos específica para las secuencias de codones, definiendo una clase para dar caracter de objeto a la secuencia de codones y almacenar las características de esta como atributos, dado que se necesitaba caracterizar estas secuencias de codones sin im-

portar la variablidad de la encapsulación en las secuencias. Se decide incluir en la recolección de información de la secuencia de codones: el gen, la posición del codón de inicio, el producto, el identificador de la proteína y la traducción.

La inclusión de esta información llevó a que la herramienta proporcionara una información más completa sobre la secuencia, caracterizando cada uno de sus productos. Los criterios de búsqueda fueron automatizados, puesto que inicialmente el usuario debía digitar una lista de características de interés de la secuencia a obtener, se procede a parametrizar estas características, excluyendo las característica de literatura relacionada con la secuencia, usando como parámetro de búsqueda, una lista creada con todas las características del primer registro dentro del archivo de entrada.

En la versión 3.0 se encontraron los siguientes problemas: los registros obtenidos de la base de datos, no eran congruentes. Al momento de realizar la extracción de características previamente definidas, encontramos que muchas de las secuencias no cumplían con los criterios de búsqueda, lo cual derivaba en incongruencias en la tabla generada, dado que el tamaño de las columnas variaba al no encontrarse la característica buscada, no se anexaba valor correspondiente a la lista. Esto derivó en que los datos se traslapaban, haciéndolos no usables, además, algunas características tienen valores complejos en sus campos, lo cual hacía que la herramienta tomara información parcial de los campos y no el valor del campo completo.

3.2.4. Version 4.0. La versión 4.0 se enfocó en normalizar las columnas para no tener problemas dada la heterogeneidad de los registros, en esta, la herramienta deja espacios en blanco en los campos en los que no encuentra el valor.

Pero esta versión aunque hacía evidente la falta de información de algunas secuencias, seguía ge-

nerando incongruencias en la tabla dado que al dejar el espacio vacío, el proceso de exportación de los datos a la taba de excel, generaba columnas de distintos tamaños, además, el equipo del GCL solicitó que se implementara una columna en la que aparecieran las secuencias en formato Fasta pero con el nombre codificado, según la nomenclatura diseñada por ellos, para brindar la posibildad de ejecutar otros análisis en paquetes con la limitante de caracteres en el encabezado y omitir la tarea de renombrar cada secuencia manualmente (como se hacía en los proyectos), lo cual da lugar a la siguiente versión.

Tomando la solicitud y la retroalimentación presentada por el equipo GCL. Se garantiza la congruencia de las columnas, haciendo que la herramienta escriba la palabra “empty” en los campos en donde la secuencia no tenía valor correspondiente al criterio de búsqueda, además, se adiciona a la herramienta una rutina que en base a la información obtenida por la herramienta, ensambla el nombre de la secuencia codificado con la nomenclatura diseñada por el GCL.

Para renombrar cada una de las secuencias con la nomenclatura diseñada por el equipo del GCL, se usa el país de origen, el estado (solo para Estados Unidos) y el segmento (En caso de ser referenciado) para generar un nombre codificado. Por ejemplo: Una secuencia reportada del virus de Influenza A H1N1 del país Estados Unidos, estado de Iowa del segmento 4 del virus; es renombrada así. El primer caracter de la nueva sintaxis corresponde al segmento en este caso "4", el segundo caracter corresponde al continente norte América "N", los siguientes dos caracteres, corresponden a la abreviatura del país “US”, para las secuencias de Estados Unidos, los siguientes dos caracteres, corresponden a la abreviación del estado, en este caso Iowa “IA” y por último un consecutivo, para identificar la secuencia. Ensamblando las partes mencionadas, el nuevo nombre de la secuencia

sería de la siguiente manera; ">4NUSIA0", el cual anteriormente era ">MG978627.1 Influenza A virus (A/Iowa/43/2017(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds".

A continuación en la figura 8 mostramos el diagrama de secuencias de la última versión de BioDatatollkit, este describe la interacción entre usuario, métodos y objetos.

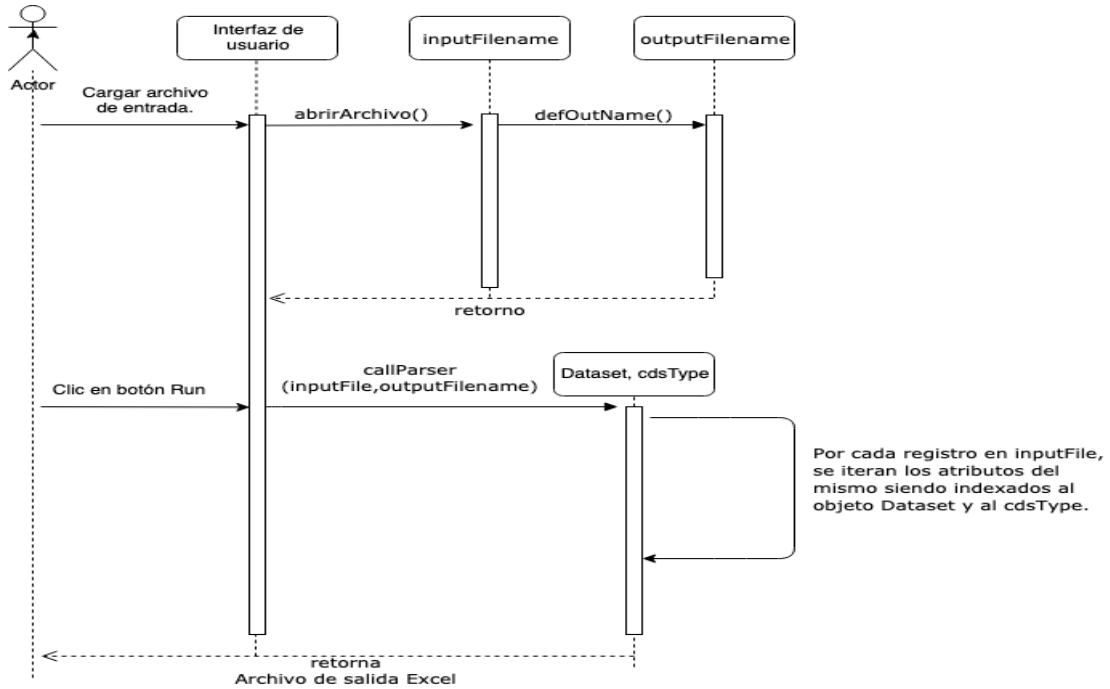

Figura 8. Diagrama de secuencias de BioDataToolkit, interacción de usuario-software.

Con el fin de facilitar el uso por parte de usuarios no familiarizados con aplicaciones de consola, se decidió implementar una interfaz gráfica GUI, la cual se muestra a continuación en la figura 9.

La interfaz gráfica es compatible con los principales sistemas operativos, permitiendo que la interacción entre la herramienta y el usuario sea intuitiva y natural.

Figura 9. BioDataToolkit GUI

4. Resultados

La verificación de esta herramienta fue realizada por el equipo investigador del GCL, obteniendo como resultado el trabajo de grado “EVALUACION DE UN PROGRAMA PARA LA GENERACION DE BASES DE DATOS CON SECUENCIAS DEL AÑO 2017 DEL GEN DE LA HEMAGLUTININA DEL VIRUS DE INFLUENZA A H1N1” realizado por Cristina Acuña estudiante de Biología, investigadora del GCL de la UIS. Esta colaboración brindó la fundamentación y retroalimentación para el desarrollo de la herramienta en conjunto con el grupo de trabajo por parte del GCL Acuña Carvajal (2019).

Estas validaciones fueron realizadas por un grupo de 3 investigadores del GCL, verificando la usabilidad de la herramienta con el virus de Influenza AH1N1, neuropéptidos y el reciente Coronavirus nCov-2019.

La herramienta evidenció una notoria reducción de tiempo al disminuir la duración de la fase de preparación de datos en el proyecto del diagnóstico del virus de Influenza A H1N1, dado que en el proyecto, esta fase en su periodo inicial tomó 15 meses. Con BioDataToolkit, la construcción de la base de datos de cada uno de los segmentos se redujo a tiempos inferiores a 1 minuto dado que usando la metodología propuesta, se descargaron todos los resultados obtenidos de las consultas realizadas en la base de datos GenBank del NCBI, los cuales fueron procesados por la herramienta BioDataToolkit, entregando tablas de Excel de salida que permitieron a los investigadores realizar el proceso planteado en la figura 6 para realizar posteriores análisis con la información obtenida, como se muestra a continuación en la figura 10.

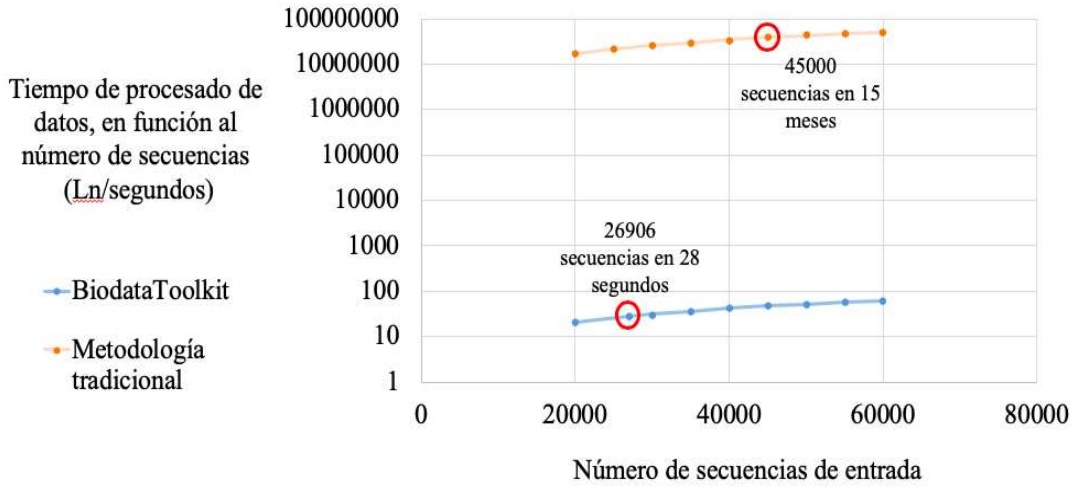

Figura 10. Comparación de tiempos de procesado de datos entre la metodología tradicional y la propuesta por este proyecto.

Dada la significativa diferencia en los tiempos de procesamiento, se hizo necesario graficar el tiempo en escala logarítmica para poder apreciar los resultados, puesto que la elaboración de la base de datos realizada por el grupo de investigadores iniciales, no llevaba una trazabilidad del proceso con las métricas de cuantas secuencias se procesaban por unidad de tiempo, se usó el tiempo final como punto de referencia para generar la línea superior, extrapolando los demás puntos.

En el caso de la representación del proceso usando BiodataToolkit (línea azul), se resaltó el punto con 26906 secuencias dado que es el que se toma como ejemplo en el video que se referencia en el documento, como ejemplo del uso de la herramienta.

Se recomienda consultar el trabajo de grado Acuña Carvajal (2019) en las páginas 28-36 para econtrar al detalle la verificación realizada por parte del GCL, disponible en: Link biblioteca UIS.

El proceso de construcción del archivo Fasta con todas las secuencias, se simplificó a un par de

clics, dado que al obtener todas las secuencias en la tabla de Excel, se pueden copiar de la columna correspondiente al Fasta todas al tiempo.

El proceso de tabulación se optimizó, puesto que la herramienta BioDataToolkit, realiza la extracción de las características de la secuencia y las convierte en las columnas de la tabla de Excel de salida. Al ser una tabla de Excel el archivo de salida, se pudo implementar el uso de la herramienta de Excel filtro, que permite al investigador tener las secuencias clasificadas, según una o varias características de criterio de selección de acuerdo con Acuña Carvajal (2019). También se sugiere consultar el video a continuación Link, el cual muestra la comparación directa entre la metodología tradicional y la planteada usando la herramienta propuesta en este proyecto.

Este proyecto se encuentra publicado en el siguiente repositorio de GitHub disponible para su distribuición y validación de libre acceso Link.

5. Recomendaciones

Durante el desarrollo de la herramienta y la conceptualización de la misma por parte del equipo multidisciplinario, se evidenció la necesidad de desarrollar herramientas que faciliten el estudio en el campo de la bioinformática, ya que esta es una de las áreas con mayor potencial de generación de conocimiento científico y evidente importancia para el país Benitez-Paez and Cárdenas-Brito (2010).

La mayoría de tareas en cuanto a clasificación y análisis de datos, se realizan de manera manual en la actualidad, por lo tanto se recomienda que se implementen soluciones como esta herramienta en el grupo de herramientas usadas por los investigadores en esta área, y se extiende la invitación a profundizar en la alianza multidisciplinaria entre las distintas escuelas involucradas en el desarrollo de este proyecto.

6. Trabajo futuro

Se propone como trabajo futuro, la profundización del proyecto, dado que pretende ser un conjunto de herramientas unificadas. Durante el desarrollo de la herramienta, se observó que es necesario implementar módulos que permitan: la generación de las secuencias complementarias, las traducciones correspondientes, la generación secuencias consenso, obtener información sobre la composición de las secuencias y codificar el nombre de un bloque de secuencias de acuerdo a la nomenclatura propuesta por GCL u otra.

Cada una de estas actividades constituye el desarrollo futuro de un módulo a acoplar en la herramienta de este trabajo, dando solución a las necesidades por parte de los investigadores en el GCL. También se recomienda realizar una validación con una muestra significativa entre proyectos que tengan registros de la duración de la preparación de sus proyectos, y bloques de secuencias correspondientes para caracterizar el comportamiento del rendimiento de la herramienta en función del tamaño del bloque de datos de entrada.

7. Conclusiones

A partir de los desarrollos presentados y los resultados obtenidos en el presente trabajo de grado, es posible enunciar la siguiente conclusión general:

Los resultados obtenidos validan la herramienta, mostrando una significativa reducción en los tiempos de ejecución de la fase de preparación de datos durante el desarrollo de proyectos de investigación en ciencia básica y aplicada, reduciendo los costos de ejecución de los proyectos y mejorando el aprovechamiento de recursos.

De manera más puntual:

Se recopilaron datos sobre la etapa de clasificación de secuencias para la ejecución de proyectos de investigación por parte del GCL, mediante los cuales se concluyó que la metodología tradicional para realizar este proceso, es obsoleta y aumenta notoriamente el tiempo de ejecución de estos proyectos.

La herramienta desarrollada permite incluir todos los resultados obtenidos de la base de datos, para una posterior clasificación usando la misma, logrando obtener un análisis más detallado. Se concluye que se logró ampliar la capacidad de análisis de los datos en los proyectos de investigación.

Se concluye que el uso del modelo de desarrollo de software prototipado evolutivo, permitió satisfactoriamente ejecutar el desarrollo de este proyecto, conforme a las necesidades plan-

teadas por el mismo. Facilitando la retroalimentación de la herramienta por parte del usuario en cada una de las versiones desarrolladas, permitiendo solucionar necesidades específicas.

Referencias Bibliográficas

Acuña Carvajal, C. I. (2019). Evaluación de un programa para la generación de bases de datos con secuencias del año 2017 del gen de la hemaglutinina del virus de influenza a h1n1. http://tangara.uis.edu.co/biblioweb/tesis/2019/176738.pdf. Accessed: 2020-01-26.

Barrios Hernández, C. J., Bautista Rozo, L. X., Gonzales Barrios, J. A., Martínez Fong, D., Martínez Pérez, F. J., Mejía Ospino, E., Munive Argüelles, N. M., Pedraza Ferreira, G. R., Ramírez Ardila, S. D., Rodríguez Vásquez, R., and et al. (2018). Nucleotide mixtures for the amplification and sequencing of nucleic acid polymers. WO2018203280 World Intellectual Property Organization.

Benitez-Paez, A. and Cárdenas-Brito, S. (2010). Bioinformática en colombia: presente y futuro de la investigación biocomputacional. *Biomédica*, 30:170.

Chou, K.-C. (2009). Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Current Proteomics - CURR PROTEOMICS*, 6.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

Gonzáles Barrios, J. A., Thompson Bonilla, M. d. R., Martínez Pérez, F. J., Barrios Hernández,

C. J., Bautista Rozo, L. X., Rodríguez Vázquez, R., and Martínez Fong, D. (2017). Oligonucleotides and process for detecting h1n1 influenza a virus. WO2017195063 World Intellectual Property Organization.

Hall, T. (1999). Bioedit: A user-friendly biologycal sequece alignment editor and analysis program for windows 95/98/nt. *Nucleic Acids Symposium Series*, 41:95–98.

Jimenez-Gutierrez, L. R., Barrios-Hernández, C. J., Pedraza-Ferreira, G. R., Vera-Cala, L., and Martinez-Perez, F. (2016). Importance of databases of nucleic acids for bioinformatic analysis focused to genomics. *Journal of Physics: Conference Series*, 743:012009.

Okonechnikov, K., Golosova, O., Fursov, M., and the UGENE team (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*, 28(8):1166–1167.

Pérez, F. and Granger, B. E. (2007). IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29.

Pressman, R. S. (2010). *Ingenieria del software: un enfoque practico*. McGraw-Hill, 7th edition.

Priyadarshi, M. B. (2014). Applications of bioinformatics. https://www.biotecharticles.com/Bioinformatics-Article/Applications-of-Bioinformatics-3270.html. Accessed: 2020-01-26.

PythonCoreTeam (2015). Python: A dynamic, open source programming language. http://www.python.org/. Accessed: 2020-01-26.

Sherrell, L. (2013). *Evolutionary Prototyping*, pages 803–803. Springer Netherlands, Dordrecht.