|  |  |
| --- | --- |
| Logo Universidad de los Andes web | EIEI 2018 - Gestión, calidad y  desarrollo en las facultades de ingenieríaEIEI 2018 - Gestión, calidad y  desarrollo en las facultades de ingeniería | **BCOM4006 – Algoritmos en Biología Computacional**  **Semestre: 2023-20**  **Jhon Stewar Rayo Mosquera** |

**Tarea 1 – Bases de datos de ADN**

1. La función *BLAST* – *The Basic Local Alignment Search Tool -* es un algoritmo diseñado para comparar secuencias biológicas tales como secuencias de aminoácidos para proteínas o secuencias de nucleótidos para secuencias de ADN. [1]. En el sitio web de *National Center for Biotechnology Information - NCBI* existen 4 variedades de BLAST con diferentes entradas y salidas de datos. [2]

***Nucleotide BLAST - blastn****.* Compara una secuencia entrada de nucleótidos con secuencias de nucleótidos en bases de datos. Es útil en la medida que permite comparar familia de genes entre distintas especies, por lo que se utiliza en este ejercicio para llevar a cabo el análisis de variabilidad del gen GBSSI.

***Protein BLAST –*** ***blastp.*** Compara una secuencia entrada de proteínas con secuencias de proteínas en bases de datos.

***Blastx.*** Compara una secuencia entrada de nucleótidos traducidas en secuencias de proteínas con bases de datos de secuencias de proteínas.

***Blastn.*** Compara una secuencia de proteína con las traducciones de secuencias de nucleótidos en bases de datos.

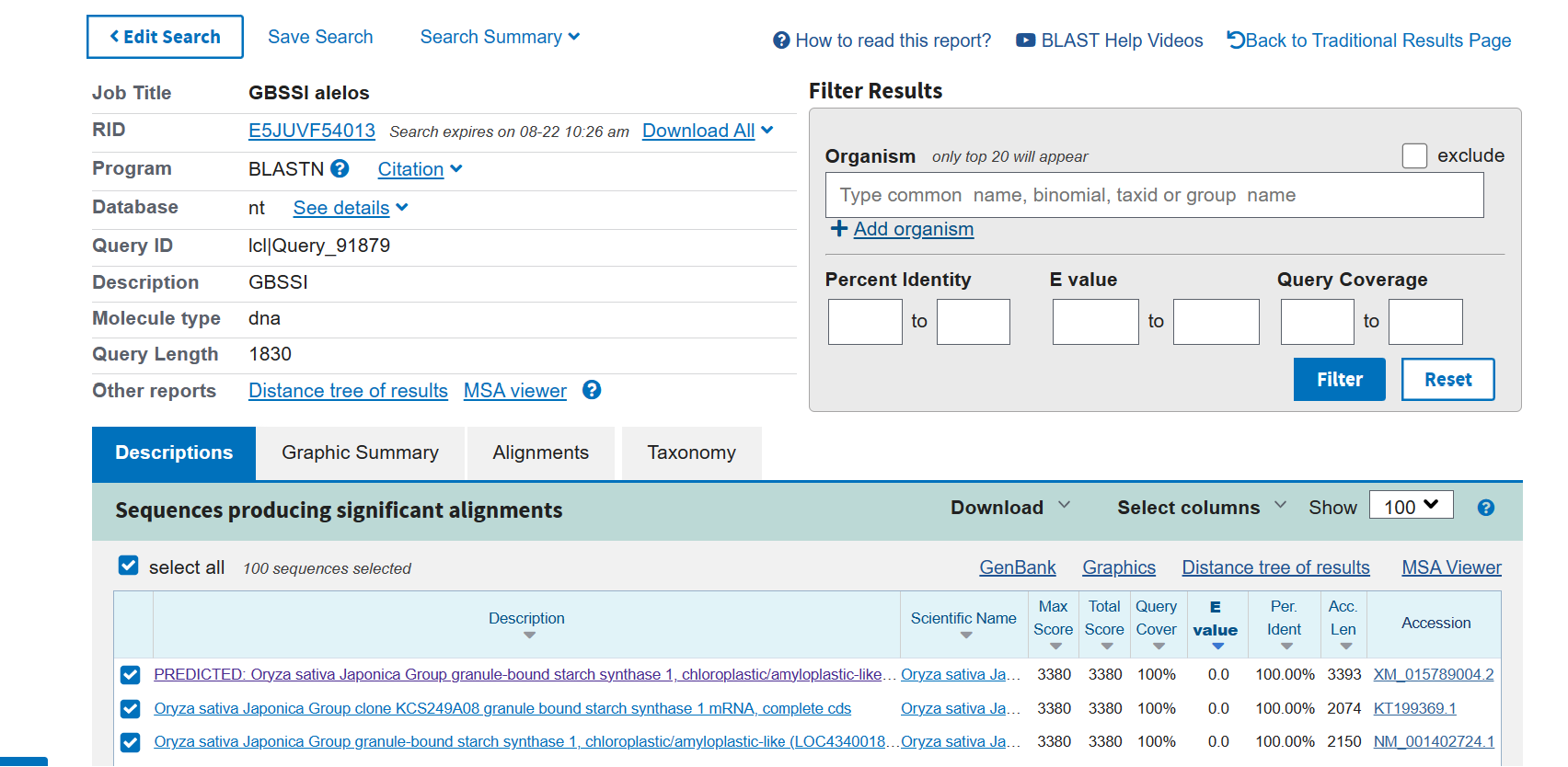
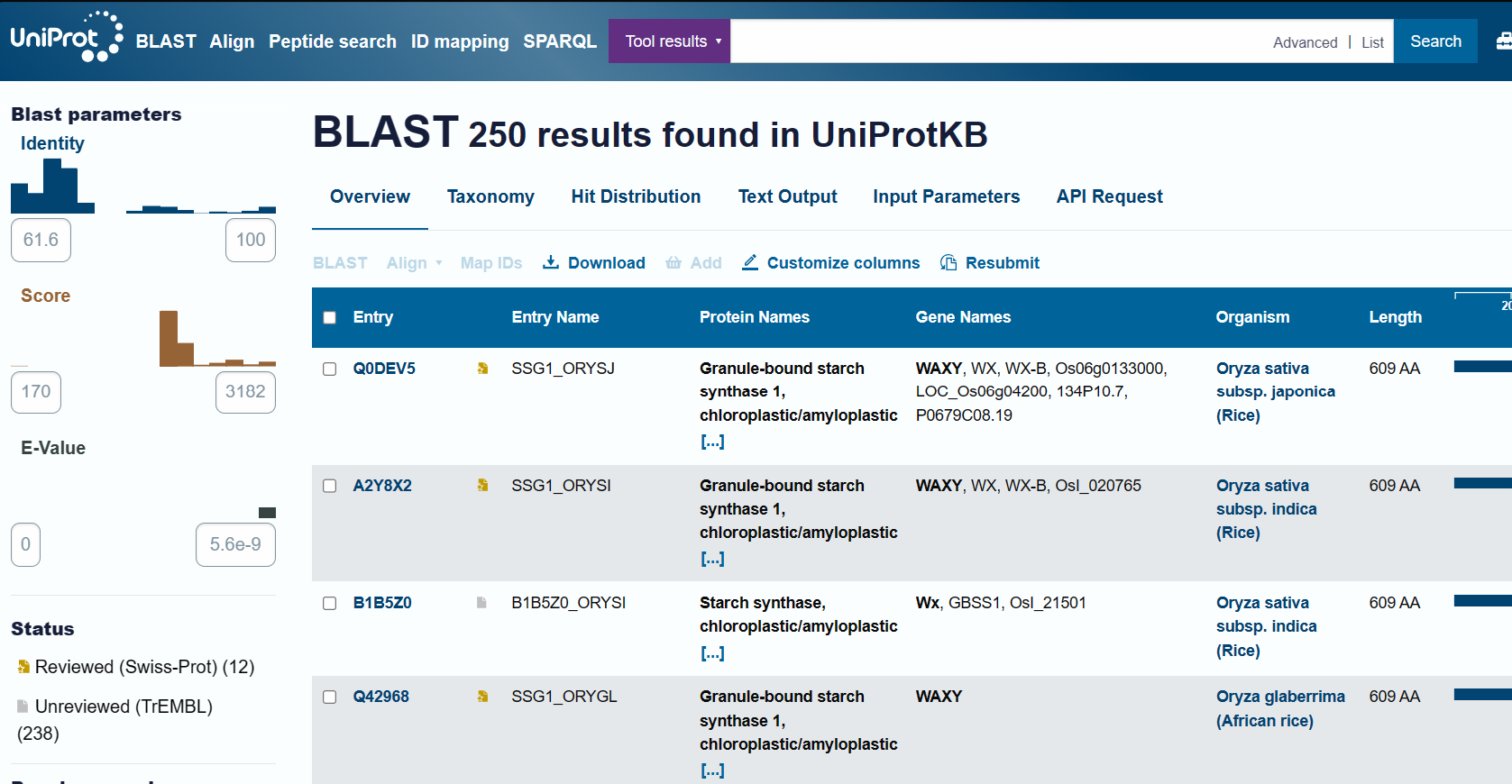


Figura 1. Resultado ejecución BLAST en *NCBI*.

Para garantizar que solo se retornaran secuencias de ARN maduro, se ejecutó el programa *BLAST* con la base de datos ***refseq\_rna*** que corresponde a una colección de secuencias de ADN comprensiva y poco redundante. [3], se obtuvo 40 secuencias con alta similitud.

Luego, para el caso de *Uniprot,* se ejecuta ***blastx*** con la base de datos ***uniprotkb\_refprotswissprot*** con el tipo ***dna***.



*Figura 2. Resultado ejecución BLAST en UniProt.*

Finalmente, en la plataforma de *Phytozome* se corrió el algoritmo BLAST para genomas de tipo ***blastn.***

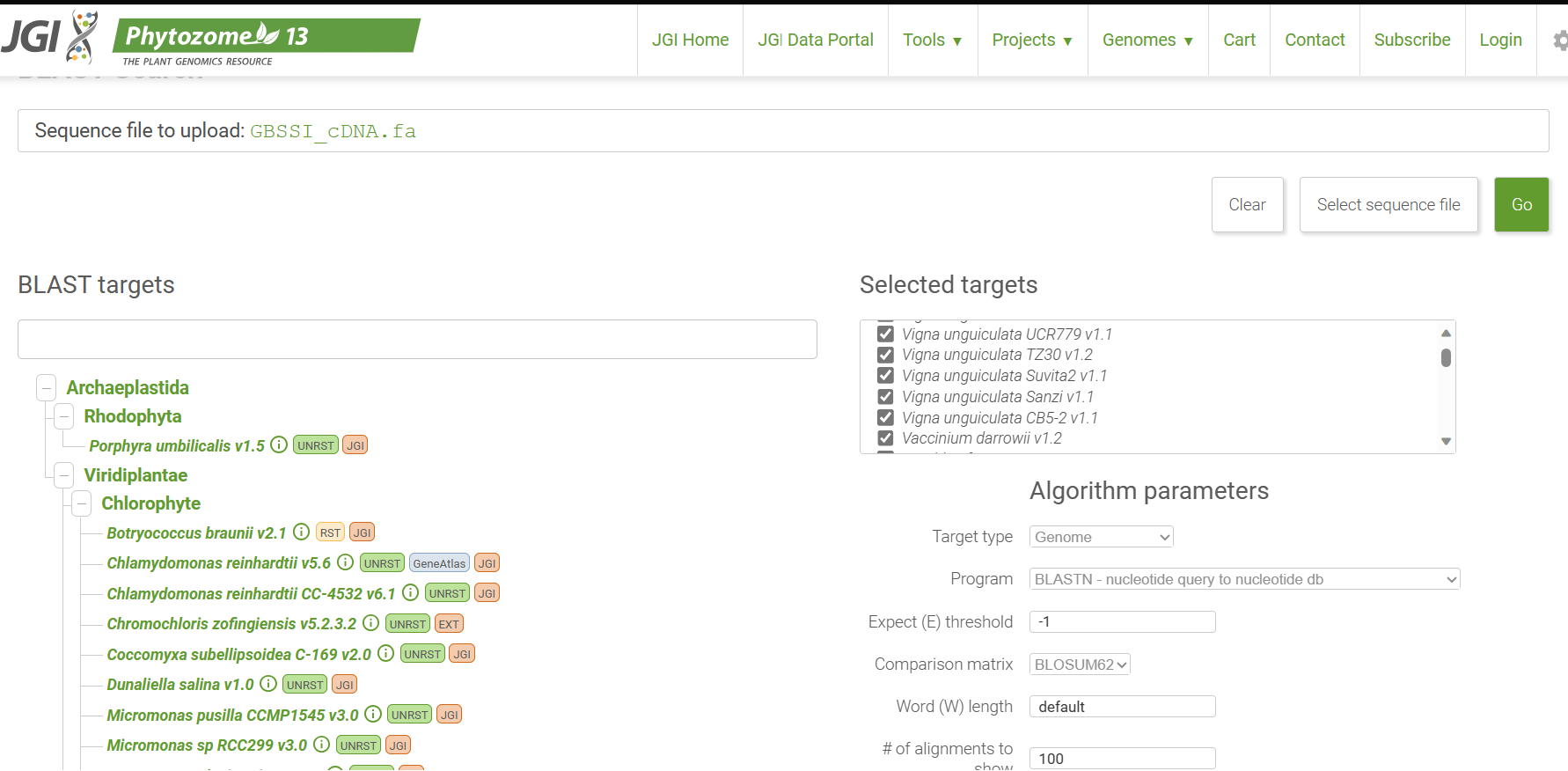


Figura 3. Ejecución BLAST en *Phytozome.*

Luego de analizar los resultados obtenidos, se agruparon las secuencias obtenidas de las bases de datos en un solo archivo de formato *fasta*, con múltiples cabeceras indicando la información de la secuencia. Se hizo uso de otros programas bioinformáticos como *mview* para convertir la salida BLAST en formato FASTA segun los formatos retornados por las distintas plataformas.

1. Se creo un programa ***remover-secuencias-repetidas.cpp*** que recibe un archivo *fasta* denominado ***GBSSI\_nucl.fa***con múltiples cabeceras siguiendo el siguiente formato.

*>secuencia1*

*ATGTCGGCTC…*

*>secuencia2*

*TCGGTACGAC…*

*>secuencia3*

*TGCTCCTTGA…*

El algoritmo recibe el archivo original y retorna un nuevo archivo con formato *fasta* sin secuencias duplicadas llamado ***GBSSI\_nucl-out.fa***. El algoritmo utiliza un mapa para eficientemente determinar si existen secuencias repetidas.

La complejidad del algoritmo es donde equivale a la cantidad de secuencias en el archivo original, y equivale al tamaño de la secuencia más larga.

El readme que acompaña el código fuente contiene instrucciones más detalladas de como ejecutar el programa.

1. Se creo un programa ***obtener-secuencias-amicds.cpp*** que recibe un archivo *fasta* denominado ***GBSSI\_nucl-out.fa,*** y retorna un archivo ***GBSSI\_amicds-out.fa***, donde cada cabecera describe la secuencia de aminoácidos que representan la traducción a proteína de las secuencias de nucleótidos correspondientes. Este programa contiene pasos tanto para transcribir la secuencia de ADN en una secuencia de ARN, para posteriormente hacer la traducción a las secuencias de proteínas teniendo en cuenta los codones de inicio y terminación. El algoritmo implementado tiene una complejidad de donde equivale a la cantidad de secuencias de nucleótidos y a la secuencia más larga.

Para este punto, es importante notar que, aunque no haya secuencias repetidas de nucleótidos en la entrada, es posible que haya secuencias de proteínas repetidas en la salida pues, aunque las secuencias no sean idénticas, pueden contener subsecuencias iguales que correspondan al inicio y terminación de un codón. De hecho, en las pruebas realizadas se encontraron proteínas repetidas.

# **Referencias**

|  |  |
| --- | --- |
| [1] | National Center for Biotechnology Information, "NIH BLAST," [Online]. Available: https://blast.ncbi.nlm.nih.gov/Blast.cgi. |
| [2] | National Center for Biotechnology Information, "NIH BLAST," [Online]. Available: https://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo\_BLASTGuide.pdf. |
| [3] | National Center for Biotechnology Information, "RefSeq," [Online]. Available: https://www.ncbi.nlm.nih.gov/refseq/about/. |