Tarea 1 - Bases de Datos de ADN

23 de agosto de 2018

Stephannie Jiménez **Código:** 201423727

El gen GBSSI es uno de los genes más importantes a estudiar en relación con el contenido de almidón en varias especies de plantas. En este punto se busca realizar un análisis de variabilidad de este gen en diferentes especies utilizando la información disponible en diferentes bases de datos, tanto de genomas completos como de secuenciación dirigida.

1. [40 %] Utilizar la función BLAST disponible en diferentes bases de datos de ADN para buscar alelos secuenciados del gen GBSSI.

1.1. NCBI

Esta base de datos, permite realizar cuatro tipos de BLAST en línea. Estos son *nucleotide* BLAST, blastx, tblastn y *protein* BLAST. La principal diferencia entre estas técnicas es sobre que tipo de dato se hace la comparación para encontrar similaridad entre secuencias de proteínas o nucleótidos según sea el caso. Es importante resaltar, que esta base de datos permite realizar otro tipo de búsquedas más especializadas con otro tipo de fines.

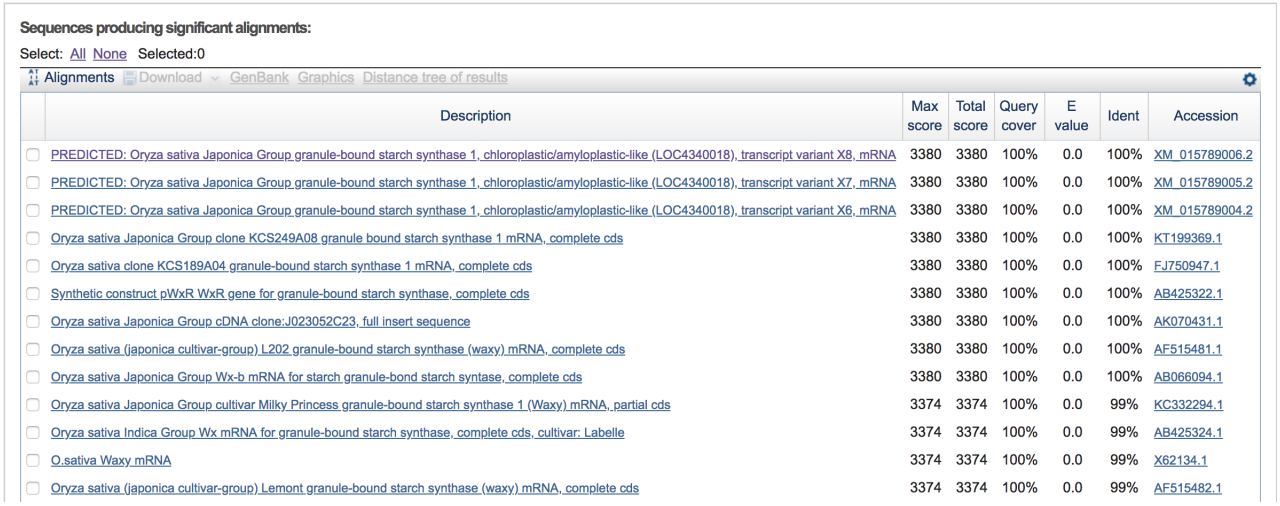
Cabe resaltar que los análisis de alineamiento de secuencias, tienden a retornar en un mismo formato. Los datos más importantes son la información general de la búsqueda, una gráfica que representa los alineamientos más importantes, junto con parámetros y estadísticos. Entre los estadísticos, se encuentra el valor esperado, *E-value*, que determina la probabilidad que el alineamiento se deba al azar y el porcentaje de identidad. También se muestran los alineamientos de las secuencias encontradas contra la ingresada, de esta forma se puede conocer en donde las secuencias son idénticas.

1.1.1. Nucleotide BLAST

En primer lugar, el *nucleotide* BLAST, *blastn*, como su nombre lo indica realiza un alineamiento de secuencias teniendo como entrada la secuencia FASTA de nucleótidos para realizar la comparación en la base de datos de nucleótidos. Su salida, depende del algoritmo que se elija, por defecto utiliza *megablast* por lo que muestra la identificación de la secuencia y una comparación intra-especies. Otros tipos de algoritmos que se pueden utilizar son *discontiguous megablast* y *blastn*.

Utilizando la secuencia del gen GBSSI dado, se logró realizar este tipo de BLAST obteniendo los resultados mostrados en la figura 1. Donde se puede ver que la secuencia tiene un alineamiento del 100 % con una gran cantidad de variantes del transcrito que codifica la misma proteína. Adicionalmente, esta prueba muestra la región donde el alineamiento de las variantes encontradas en la base de datos es igual a la consulta, en caso de que haya uno o más nucleótidos que no se alineen correctamente, se muestra una x. La figura 2 muestra el alineamiento de la primer secuencia encontrada.

Es importante mencionar, que con los resultados obtenidos se puede determinar que hay una alta variabilidad de este gen puesto que hay un 100 % de alineamiento con al menos nueve variedades de *Oryza Sativa*. Cabe mencionar que el valor esperado de las nueve primeras secuencias encontradas son cercanas a cero, por lo que la probabilidad de que el alineamiento sea al azar es nulo.



Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
PREDICTED: Oryza sativa Japonica Group granule-bound starch synthase 1, chloroplastic/amyloplastic-like (LOC4340018), transcript variant X8, mRNA	3380	3380	100%	0.0	100%	XM_015789006.2
PREDICTED: Oryza sativa Japonica Group granule-bound starch synthase 1, chloroplastic/amyloplastic-like (LOC4340018), transcript variant X7, mRNA	3380	3380	100%	0.0	100%	XM_015789005.2
PREDICTED: Oryza sativa Japonica Group granule-bound starch synthase 1, chloroplastic/amyloplastic-like (LOC4340018), transcript variant X6, mRNA	3380	3380	100%	0.0	100%	XM_015789004.2
Oryza sativa Japonica Group clone KCS249A08 granule bound starch synthase 1 mRNA, complete cds	3380	3380	100%	0.0	100%	KT199369.1
Oryza sativa clone KCS189A04 granule-bound starch synthase 1 mRNA, complete cds	3380	3380	100%	0.0	100%	FJ750947.1
Synthetic construct pWxR WxR gene for granule-bound starch synthase, complete cds	3380	3380	100%	0.0	100%	AB425322.1
Oryza sativa Japonica Group cDNA clone:J023052C23, full insert sequence	3380	3380	100%	0.0	100%	AK070431.1
Oryza sativa (japonica cultivar-group) L202 granule-bound starch synthase (waxy) mRNA, complete cds	3380	3380	100%	0.0	100%	AF515481.1
Oryza sativa Japonica Group Wx-b mRNA for starch granule-bond starch syntase, complete cds	3380	3380	100%	0.0	100%	AB066094.1
Oryza sativa Japonica Group cultivar Milky Princess granule-bound starch synthase 1 (Waxy) mRNA, partial cds	3374	3374	100%	0.0	99%	KC332294.1
Oryza sativa Indica Group Wx mRNA for granule-bound starch synthase, complete cds, cultivar: Labelle	3374	3374	100%	0.0	99%	AB425324.1
O.sativa Waxy mRNA	3374	3374	100%	0.0	99%	X62134.1
Oryza sativa (japonica cultivar-group) Lemont granule-bound starch synthase (waxy) mRNA, complete cds	3374	3374	100%	0.0	99%	AF515482.1

Figura 1: Resultados de *nucleotide* BLAST.

Download GenBank Graphics ▼ Next ▲ Previous Descriptions

PREDICTED: Oryza sativa Japonica Group granule-bound starch synthase 1, chloroplastic/amyloplastic-like (LOC4340018), transcript variant X8, mRNA

Sequence ID: XM_015789006.2 **Length:** 2254 **Number of Matches:** 1

Range 1: 181 to 2010 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
3380 bits(1830)	0.0	1830/1830(100%)	0/1830(0%)	Plus/Plus

Related Information

New Genome Data Viewer - aligned genomic context

```
Query  1     ATGTCGGCTCTCACCACGTCCCAGCTCGCCACCTCGGCCACCGGCTTCGGCATCGCCGAC  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181   ATGTCGGCTCTCACCACGTCCCAGCTCGCCACCTCGGCCACCGGCTTCGGCATCGCCGAC  240

Query  61    AGGTCGGCGCCGTCGTCGCTGCTCCGCCACGGGTTCCAGGGCCTCAAGCCCCGCAGCCCC  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241   AGGTCGGCGCCGTCGTCGCTGCTCCGCCACGGGTTCCAGGGCCTCAAGCCCCGCAGCCCC  300

Query  121   GCCGGCGGCGACGCGACGTCGCTCAGCGTGACGACCAGCGCGCGCGCGACGCCCAAGCAG  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301   GCCGGCGGCGACGCGACGTCGCTCAGCGTGACGACCAGCGCGCGCGCGACGCCCAAGCAG  360

Query  181   CAGCGGTCGGTGCAGCGTGGCAGCCGGAGGTTCCCCTCCGTCGTCGTGTACGCCACCGGC  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361   CAGCGGTCGGTGCAGCGTGGCAGCCGGAGGTTCCCCTCCGTCGTCGTGTACGCCACCGGC  420

Query  241   GCCGGCATGAACGTCGTGTTCGTCGGCGCCCGAGATGGCCCCCTGGAGCAAGACCGGCGGC  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421   GCCGGCATGAACGTCGTGTTCGTCGGCGCCCGAGATGGCCCCCTGGAGCAAGACCGGCGGC  480

Query  301   CTCGGTGACGTCCTCGGTGGCCTCCCCCCTGCCATGGCTGCGAATGGCCACAGGGTCATG  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481   CTCGGTGACGTCCTCGGTGGCCTCCCCCCTGCCATGGCTGCGAATGGCCACAGGGTCATG  540

Query  361   GTGATCTCTCCTCGGTACGACCAGTACAAGGACGCTTGGGATACCAGCGTTGTGGCTGAG  420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541   GTGATCTCTCCTCGGTACGACCAGTACAAGGACGCTTGGGATACCAGCGTTGTGGCTGAG  600

Query  421   ATCAAGGTTGCAGACAGGTACGAGAGGGTGAGGTTTTTCCATTGCTACAAGCGTGGAGTC  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  601   ATCAAGGTTGCAGACAGGTACGAGAGGGTGAGGTTTTTCCATTGCTACAAGCGTGGAGTC  660

Query  481   GACCGTGTGTTCATCGACCATCCGTCATTCCTGGAGAAGGTTTGGGGAAAGACCGGTGAG  540
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661   GACCGTGTGTTCATCGACCATCCGTCATTCCTGGAGAAGGTTTGGGGAAAGACCGGTGAG  720

Query  541   AAGATCTACGGACCTGACACTGGAGTTGATTACAAAGACAACCAGATGCGTTTCAGCCTT  600
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  721   AAGATCTACGGACCTGACACTGGAGTTGATTACAAAGACAACCAGATGCGTTTCAGCCTT  780

Query  601   CTTTGCCAGGCAGCACTCGAGGCTCCTAGGATCCTAAACCTCAACAACAACCCATACTTC  660
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  781   CTTTGCCAGGCAGCACTCGAGGCTCCTAGGATCCTAAACCTCAACAACAACCCATACTTC  840
```

Figura 2: Alineamiento de la primer secuencia encontrada utilizando *nucleotide* BLAST.

1.1.2. Protein BLAST

En cambio, el *protein* BLAST, *blastp*, permite comparar una proteína con la base de datos de proteínas. Para poder utilizar este tipo de análisis se necesita la secuencia FASTA de aminoácidos como entrada. De esta forma, se está comparando una proteína con la base de datos de proteínas. Al igual que en el caso anterior, este tipo de BLAST tiene una gran cantidad de algoritmos que se pueden utilizar. Por defecto, se encuentra *blastp* que se encarga de hacer una identificación general de la secuencia y de similitud. Otro tipo de algoritmos que se pueden utilizar son *Delta BLAST*, *PSI-BLAST* y *PHI-BLAST*.

Se realizó este tipo de BLAST utilizando la secuencia de aminoácidos para el gen GBSSI. Los resultados de esta prueba se muestran en la figura 3, en este se muestran los dominios putativos conservados que se encontraron. En este se ve que se clasifico la proteína como una glicógeno sintasa.

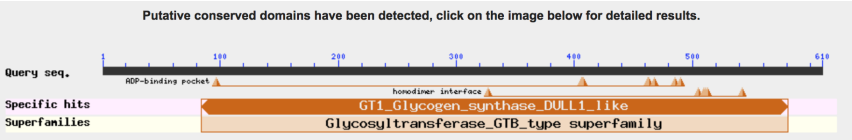


Figura 3: Dominios putativos conservados en la proteina del gen GBSSI.

Al revisar los resultados del alineamiento, mostrados en la figura 4 se puede determinar que la proteína pertenece a la sintasa del almidón en *Oryza Sativa Japonica Group* puesto a que su alineamiento es del 100 %. En los demás casos, el alineamiento es del 99 % con otras variedades de la misma especie.

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
granule-bound starch synthase 1, chloroplastic/amyloplastic isoform X2 [Oryza sativa Japonica Group]	1262	1262	99%	0.0	100%	XP_015644490.1
GBSS1 [Oryza sativa Japonica Group]	1261	1261	99%	0.0	99%	CCW36717.1
granule-bound starch synthase I [Oryza sativa Indica Group]	1261	1261	99%	0.0	99%	ACJ86344.1
granule-bound starch synthase I [Oryza sativa]	1261	1261	99%	0.0	99%	ACJ86356.1
RecName: Full=Granule-bound starch synthase 1, chloroplastic/amyloplastic; AltName: Full=Granule-bound starch synthase I; Short=GBSS-I; Flags: Precurs	1260	1260	99%	0.0	99%	Q42968.1
OsGBSS1 [Oryza sativa Indica Group]	1260	1260	99%	0.0	99%	CCW36719.1
granule-bound starch synthase I [Oryza sativa Indica Group]	1260	1260	99%	0.0	99%	ACJ86363.1
granule-bound starch synthase [Oryza sativa Japonica Group]	1260	1260	99%	0.0	99%	ACU82451.1
granule-bound starch synthase I [Oryza sativa]	1259	1259	99%	0.0	99%	ACJ86360.1
granule-bound starch synthase 1 [Oryza sativa Japonica Group]	1259	1259	99%	0.0	99%	AGK90263.1
granule-bound starch synthase I [Oryza sativa Indica Group]	1259	1259	99%	0.0	99%	ACJ86351.1
granule-bound starch synthase [Oryza sativa Japonica Group]	1259	1259	99%	0.0	99%	AAF72561.1
granule-bound starch synthase [Oryza sativa Japonica Group]	1259	1259	99%	0.0	99%	AAC61675.2
OsGBSS1 [Oryza sativa Japonica Group]	1259	1259	99%	0.0	99%	CCW36718.1
waxy [Oryza sativa Indica Group]	1259	1259	99%	0.0	99%	ABD77490.1
starch granule-bound starch synthase [Oryza sativa Japonica Group]	1258	1258	99%	0.0	99%	BAB88209.1
granule-bound starch synthase [Oryza sativa Japonica Group]	1258	1258	99%	0.0	99%	ACT52669.1
granule-bound starch synthase I [Oryza sativa Indica Group]	1257	1257	99%	0.0	99%	ACJ86357.1

Figura 4: Resultados de *protein* BLAST.

1.1.3. BLASTx

Este tipo de BLAST se fundamenta en traducir la secuencia de nucleótidos en seis secuencias de proteínas. Después, estas secuencias de proteínas se comparan contra la base de datos de proteínas. Al utilizar la secuencia del gen GBSSI se encuentran los siguientes resultados. En primer lugar, están los dominios putativos conservados en la proteína mostrados en la figura 5. Se puede observar que pertenece a la super-familia GlgA.

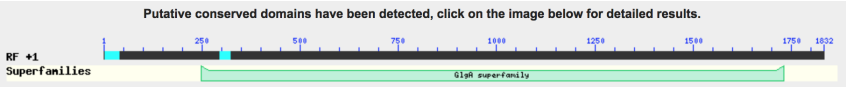


Figura 5: Resultados de *protein* BLAST.

Por otra parte, los resultados del alineamiento con la base de datos de proteínas se ven en la figura 6. Se puede observar que se encontró el mismo registro que en la prueba de *protein* BLAST. Así se determina que la proteína de interés tiene un alineamiento completo con la sintasa del almidón en *oryza Sativa Japonica Group*.

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics

Description	Max score	Total score	Query cover	E value	Ident	Accession
granule-bound starch synthase 1, chloroplastic/amyloplastic isoform X2 [Oryza sativa Japonica Group]	1211	1211	97%	0.0	100%	XP_015644490.1
GBSS1 [Oryza sativa Japonica Group]	1211	1211	97%	0.0	99%	CCW36717.1
granule-bound starch synthase I [Oryza sativa]	1210	1210	97%	0.0	99%	ACJ86356.1
RecName: Full=Granule-bound starch synthase 1, chloroplastic/amyloplastic; AltName: Full=Granule-bound starch synthase I; Short=GBSS-I; Flags: Precurs	1209	1209	97%	0.0	99%	Q42968.1
OsGBSS1 [Oryza sativa Indica Group]	1209	1209	97%	0.0	99%	CCW36719.1
granule-bound starch synthase I [Oryza sativa Indica Group]	1209	1209	97%	0.0	99%	ACJ86344.1
granule-bound starch synthase I [Oryza sativa Indica Group]	1209	1209	97%	0.0	99%	ACJ86363.1
granule-bound starch synthase I [Oryza sativa Indica Group]	1209	1209	97%	0.0	99%	ACJ86351.1
granule-bound starch synthase [Oryza sativa Japonica Group]	1209	1209	97%	0.0	99%	AAF72561.1
granule-bound starch synthase [Oryza sativa Japonica Group]	1209	1209	97%	0.0	99%	ACU82451.1
OsGBSS1 [Oryza sativa Japonica Group]	1209	1209	97%	0.0	99%	CCW36718.1

Figura 6: Resultados de BLASTx.

1.1.4. tBLASTn

Este algoritmo recibe como entrada la secuencia de aminoácidos y traduce la base de datos de ADN a 6 posibles proteínas para realizar la comparación. De esta forma, se esta realizando un alineamiento de proteínas. Cuando se realiza la búsqueda del gen GBSSI se obtienen los resultados mostrados en la figura 7. Se puede observar que al igual que en las otras búsquedas en la base de datos se obtuvo una alto alineamiento de la secuencia con una variedad de la *Oryza Sativa Japonica Group*.

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics

Description	Max score	Total score	Query cover	E value	Ident	Accession
Oryza sativa Japonica Group Wx gene for OsGBSS1, cultivar Meifeng6, allele A114S	1211	1211	97%	0.0	98%	HF951690.1
Synthetic construct pWxR WxR gene for granule-bound starch synthase, complete cds	1211	1211	97%	0.0	98%	AB425322.1
Oryza sativa (japonica cultivar-group) L202 granule-bound starch synthase (waxy) mRNA, complete cds	1211	1211	97%	0.0	98%	AF515481.1
Oryza sativa Indica Group Wx gene for OsGBSS1, cultivar 32657, allele T543A	1210	1210	97%	0.0	98%	HF951692.1
Oryza sativa (indica cultivar-group) Rexmont granule-bound starch synthase (waxy) mRNA, complete cds	1209	1209	97%	0.0	98%	AF515480.1
Oryza sativa Japonica Group Wx gene for OsGBSS1, cultivar 38138, allele EA314-315QT	1209	1209	97%	0.0	98%	HF951691.1
Oryza sativa Japonica Group cultivar Milky Princess granule-bound starch synthase 1 (Waxy) mRNA, partial cds	1209	1209	97%	0.0	98%	KC332294.1
Oryza sativa Indica Group Wx mRNA for granule-bound starch synthase, complete cds, cultivar: Labelle	1209	1209	97%	0.0	98%	AB425324.1
Oryza sativa (japonica cultivar-group) Lemont granule-bound starch synthase (waxy) mRNA, complete cds	1209	1209	97%	0.0	98%	AF515482.1

Figura 7: Resultados de tBLASTn.

1.2. Uniprot

Esta base de datos permite realizar un alineamiento utilizando como entrada una cadena de aminoácidos o de nucleótidos y la compara contra la base de datos de proteínas. La figura 8 muestra el resultado general que se obtiene de la búsqueda en esta base de datos con la secuencia de nucleótidos proporcionada.

Figura 8: Resultados del BLAST realizado en la base de datos Uniprot.

Al revisar el detalle de la primer muestra, se puede observar que se encontró la misma proteína que se encontró en la base de datos de NCBI. Pues la secuencia que se ingresó tiene un 100 % de similitud con la sintasa del almidón en *Oryza Sativa Japonica Group*.

1.3. Phytozome

Por último, se realizo el BLAST en esta base de datos. Esta tiene una gran variedad de plantas, al igual que tres tipos de BLAST que se previamente se explicaron. Se realizó el BLASTn utilizando la secuencia de nucleótidos dada y se obtuvieron una gran variedad de resultados. Se decidieron tomar los 10 primeros resultados obteniendo tanto la secuencia de nucleótidos como la de aminoácidos para cada gen. Un ejemplo de los resultados encontrados en esta base de datos se muestra en la figura 9.

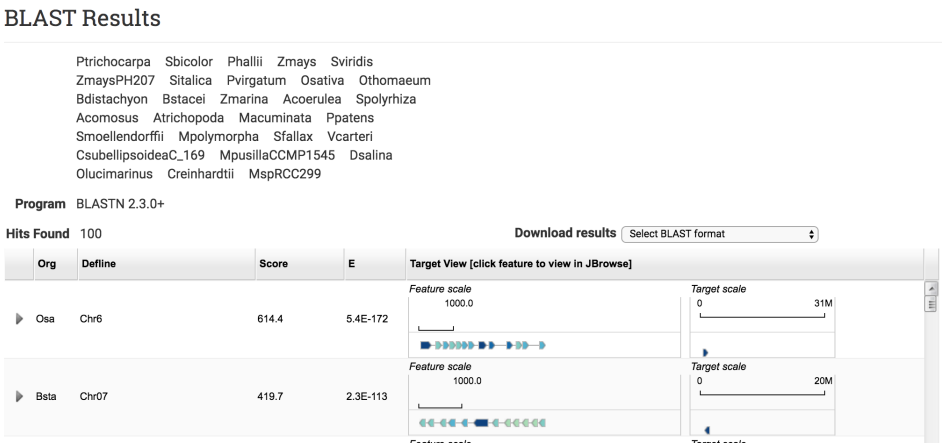


Figura 9: Resultados del BLAST realizado en la base de datos Phytozome.

Nota: Se anexan todas las secuencias de nucleótidos y proteínas usando el formato FASTA en varios archivos.

2. [30 %] Realizar un alineamiento múltiple de secuencias con las secuencias de nucleótidos y otro con la secuencia de proteinas recolectadas en el punto anterior.

Se utilizó el algoritmo *Multiple Sequence Comparison by Log-Expectation* (MUSCLE) para realizar la comparación entre las secuencias que se encontraron en el anterior. Para llevar a cabo este proceso, se necesitan dos secuencias de nucleótidos o de aminoácidos. Adicionalmente, este tipo de prueba tiene como salida una gran variedad de formatos donde se incluye el *clustalW*, FASTA, HTML, entre otros. El formato *clustalW* permite generar un archivo donde se muestran los resultados de la comparación de dos o más secuencias. Este genera bloques por cada par de secuencias comparadas, donde se muestra el nombre de las secuencias, hasta 60 símbolos de las secuencias y una línea que muestra el grado de conservación.

Se realizó el alineamiento múltiple por nucleótidos, y los resultados se muestran en la imagen 10, mientras que el resultado del alineamiento múltiple por proteínas se muestra en la figura 11. Al igual que en el numeral anterior, se anexan todos los resultados en formato *clustalW*.

```
Brast07G026000.1        ATGGCGGCTCTGGTCACGTCCCAGCTCGCCCCCACGTGTGCCGGCAT-GCCGCCGCCGCC
Bradi1g50090.1          ATGGCGGCTCTGGTCACGTCCCAGCTCGCCGCCACGTGTGCCGGCAT-GCCGCCGCCGCC
HF951690.1              ATGTCGGCTCTCACCACGTCCCAGCTCGCCACCTCGGCCACCGGCTTCGGCATCGCCGAC
LOC_Os06g04200.3        ATGTCGGCTCTCACCACGTCCCAGCTCGCCACCTCGGCCACCGGCTTCGGCATCGCCGAC
XM_015789006.2          ATGTCGGCTCTCACCACGTCCCAGCTCGCCACCTCGGCCACCGGCTTCGGCATCGCCGAC
Seita.4G022400.1        ATGGCGGCTCTGGCCACTTCCCAGCTCGTCACCACCCGCGCCGGCTTCGGCCTCGGCGAC
Sevir.4G021900.1        ATGGCGGCTCTGGCCACTTCCCAGCTCGCCACCACCCGCGCCGGCTTCGGCCTCGGCGAC
Pahal.J00308.1          ATGGCGGCTCTGGCCACGTCCCAGCTCGCCACCACCCACGCCGGCTTCGGCCTCGGCGGC
Pavir.J25731.1          ATGGCGGCTCTGGCAACGTCCCAGCTCGCCACCACTCACGCCGGCTTCGGCCTCGGCGGC
Pavir.Db02291.1         ATGGCGGCTCTGGCAACGTCCCAGCTCGCCACCACCCACGCCGGCTTCGGCCTCGGCGGC
Sobic.010G022600.1      ATGTCGACTCTAGCCACGTCGCAGCTCGTCGCCACGCACGCCGGCCTAGGCGTCCCGGAC
Zm00008a033823_T01      ATGGCGGCTCTGGCCACGTCGCAGCTCGTCGCAACGCCCGCCGGCCTGGGCGTCCCGGAC
                        *** ** ****    ** ** ******* * *  *     ***** * * *  *   * *

Brast07G026000.1        --------------GTCCATGCTCCGGCGCGGCCAC---CATGGC---AAGATCG-----
Bradi1g50090.1          -----------GTCGTCCGTGCTCCGGCGCGGCCAC---CATGGC---AAGATCG-----
HF951690.1              AGGTCGGCGCCGTCGTCGCTGCTCCGCCACGGGTTC---CAGGGCCTCAAGCCCCGCAGC
LOC_Os06g04200.3        AGGTCGGCGCCGTCGTCGCTGCTCCGCCACGGGTTC---CAGGGCCTCAAGCCCCGCAGC
XM_015789006.2          AGGTCGGCGCCGTCGTCGCTGCTCCGCCACGGGTTC---CAGGGCCTCAAGCCCCGCAGC
Seita.4G022400.1        ------GCCTCCTCCTCCATGTTCCGCCCCGGCGTC---CAGGGCCT-CAG---GGGCTC
Sevir.4G021900.1        ------GCCTCCTCCTCCATGTTCCGCCCCGGCGTC---CAGGGCCT-CAGCAGGGGCTC
Pahal.J00308.1          ---------GACACCTCCATGTTCCGCCGCGGCGTC---CAGGGCCT-GAG---GGGGCC
Pavir.J25731.1          ---------GACACCTCCATGTTCCGCCGCGGCGTC---CAGGGCCT-GAG---GGCCCC
Pavir.Db02291.1         ---------GACACCTCCATGTTCCGCCGCGGCGTC---CAGGGCCT-GAG---GGGGCC
Sobic.010G022600.1      ------------GCGTCCATGTTCCGCCGCGGCGGCGTGCAGGGCCT-GAGGGCGGCGGC
Zm00008a033823_T01      ------------GCGTCCACGTTCCGCCGCGGCGCCGCGCAGGGCCT-GAG---GGGGGC
                                   **   * **** * ***   *   ** ***    **
```

Figura 10: Resultados del alineamiento múltiple de las secuencias de nucleótidos obtenidas en el numeral anterior.

A partir de los resultados, se puede determinar que hubo un mejor alineamiento entre las secuencias de proteínas y no las de nucleótidos. Esto se debe a que ciertos aminoácidos de las proteínas pueden ser generados por una combinación de nucleótidos, por lo que pueden existir pequeñas diferencias entre las cadenas que no generan una diferencia en el peptido.

Figura 11: Resultados del alineamiento múltiple de las secuencias de aminoácidos obtenidas en el numeral anterior.

3. [30 %] Investigar herramientas que permitan construir dendogramas con el algoritmo UPGMA y con el algoritmo Neighbor Joining a partir de cada uno de los alineamientos construidos en el punto anterior.

Se realizó una búsqueda sobre algunas herramientas ya desarrolladas que permiten generar dendogramas tanto con el algoritmo UPGMA y Neighbour Joining. En esta sección se describiran algunas de las herramientas encontradas que son de software libre y se pueden ejecutar teniendo como entrada el formato *clustalW* o similar.

3.1. Algoritmo UPGMA

En primer lugar, se encontraron dos herramientas que permiten generar dendogramas utilizando este tipo de algoritmo. Uno de ellos es un servicio de tipo web, Simple Phylogeny[1], que permite generar los dendogramas a partir de los alineamientos que se encontraron en el numeral anterior. Para poder utilizarlo de forma efectiva, es necesario tener el alineamiento en formato *clustalW*. Al utilizar este programa, se lograron generar los dendogramas mostrados en las figuras 12 y 13. Para utilizar esta herramienta, es importante mencionar que no puede existir un id duplicado porque sino el programa arroja un error.

Por otra parte, se encontró otra herramienta web desarrollada por un grupo francés. [2]. Se utiliza para generar un análisis para datos científicos, como para manipulación de estos, visualización, estadística, análisis espacial, entre otros. Una de sus funciones permite generar dendogramas a partir de una función de clustering jerárquico que se basa en el algoritmo UPGMA. Los árboles generados utilizando este programa se pueden observar en las figuras 14 y 15.

[1] Disponible en https://www.ebi.ac.uk/Tools/phylogeny/simple_phylogeny/
[2] http://phylogeny.lirmm.fr/

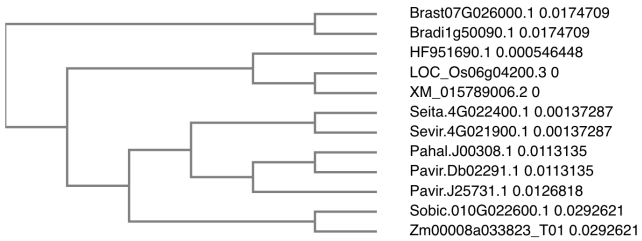


Figura 12: Resultados del dendograma utilizando el archivo de nucleótidos obtenidos en el numeral anterior con el algoritmo UPGMA.

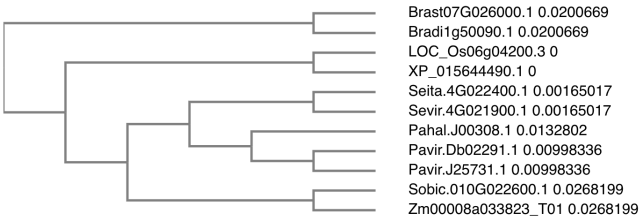


Figura 13: Resultados del dendograma utilizando el archivo de aminoácidos obtenidos en el numeral anterior con el algoritmo UPGMA.

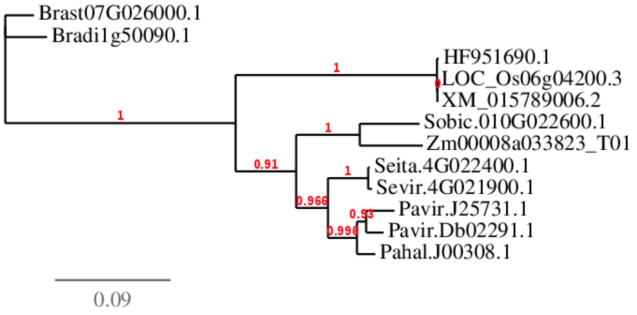


Figura 14: Resultados del dendograma utilizando el archivo de nucleótidos obtenidos en el numeral anterior con el algoritmo UPGMA.

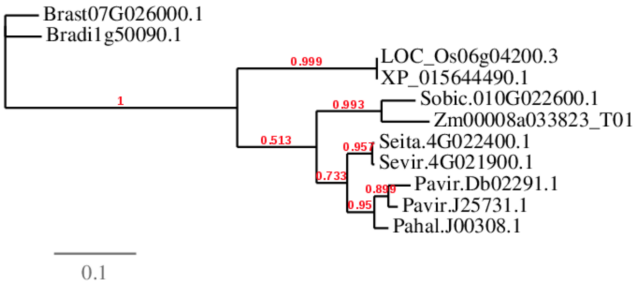


Figura 15: Resultados del dendograma utilizando el archivo de nucleótidos obtenidos en el numeral anterior con el algoritmo UPGMA.

Se puede observar que los árboles entre las dos herramientas utilizadas difieren entre sí. Pero el algoritmo de clusterización intenta generar una jerarquía de forma aglomerativa. Es decir, genera el árbol usando un *bottom-up* utilizando la matriz de semejanza. De esta forma, calcula la distancia entre los genes ingresados y empieza a clusterizar los que se encuentren más cerca de cierto umbral, hasta que solo quede un grupo grande que contenga todos los genes. Es de esperar, que el algoritmo realice diferentes árboles según la noción de distancia o el umbral elegido de forma automática. Cabe resaltar que la diferencia entre los árboles generados con la misma herramienta pero con la secuencia de nucleótidos o aminoácidos tienen una forma similar, lo cuál es lo esperado.

3.2. Algoritmo Neighbour Joining

Al igual que en el caso anterior, se encontraron dos herramientas que permiten generar dendogramas utilizando este tipo de algoritmo. El primero, es el mismo servicio web Simple Phylogeny[3], que permite generar los dendogramas a partir de los alineamientos que se encontraron. Para poder utilizarlo de forma efectiva, es necesario tener el alineamiento en formato *clustalW*. Al utilizar este programa, se lograron generar los dendogramas mostrados en las figuras 16 y 17. De igual forma, se utilizó la misma página web de *phylogeny* para generar los dendogramas utilizando el algoritmo de Neighbour Joining. Las figuras 18 y 19 muestra los árboles generados.

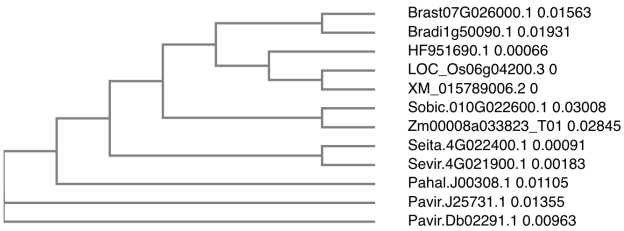


Figura 16: Resultados del dendograma utilizando el archivo de nucleótidos obtenidos en el numeral anterior con el algoritmo Neighbour Joining.

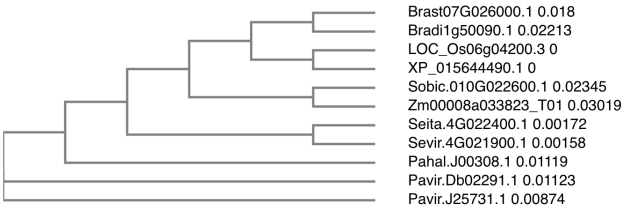


Figura 17: Resultados del dendograma utilizando el archivo de aminoácidos obtenidos en el numeral anterior con el algoritmo Neighbour Joining.

Entre las dos herramientas que se utilizaron, se puede mostrar que tienen una forma similar, y que se agruparon los mismos genes entre los árboles, pero, no son idénticos. El algoritmo de Neighbour Joining se fundamenta en un método de clusterización *bottom-up*, por lo cual, funciona de forma aglomerativa. Es decir, se empiezan a juntar los genes dependiendo de la noción de distancia que se determine y utiliza la matriz de distancia utilizando la medida de distancia euclidiana.

Cabe mencionar, que pese a que ambos métodos utilizados para generar los árboles son de tipo aglomerativo y *bottom-up*, los resultados son diferentes. Esto se debe a que el espacio de representación

[3]Disponible en https://www.ebi.ac.uk/Tools/phylogeny/simple_phylogeny/

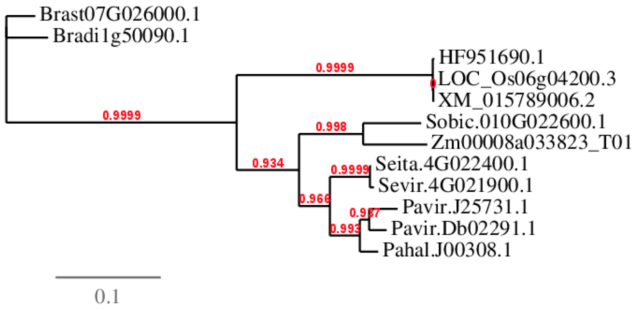


Figura 18: Resultados del dendograma utilizando el archivo de nucleótidos obtenidos en el numeral anterior con el algoritmo Neighbour Joining.

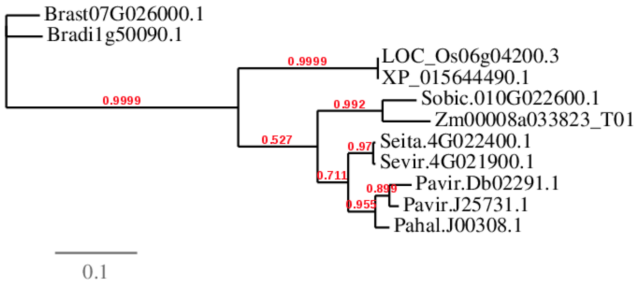


Figura 19: Resultados del dendograma utilizando el archivo de aminoácidos obtenidos en el numeral anterior con el algoritmo Neighbour Joining.

utilizado en conjunto con la noción de distancia elegida forma diferentes árboles. Por lo cual, los umbrales con los cuales se determinan los clusters de genes difieren incluso entre las cadenas de aminoacidos y las de nucleótidos. Aún así, se encuentra que los árboles encontrados tienen una similitud alta entre sí.

4. (Bono [10 %]) Para verificar la estabilidad de las ramas se realiza un procedimiento llamado *"bootstrapping"*.

Generalmente, se realiza un análisis de *Bootstrap* para encontrar los intervalos de confianza entre las filogenias. Este es un análisis estadístico que permite determinar cuan es la estabilidad de las ramas que se encontraron en el árbol filogenético. Se fundamenta en realizar un muestreo sobre los datos originales. Por lo general, se realizan varios árboles con los mismos datos para generar alternativas y medir el número de veces que cierto gen se agrupo por cluster.

Una de las herramientas utilizada en el numeral anterior, permite generar este tipo de análisis y requiere como parámetro el número de *bootstraps* que se quieren realizar. Esto significa, que se generan este n numero de árboles y se empieza a registrar cuantas veces aparecen los genes en ciertos clusters. Para este tipo de análisis, se requiere un gasto computacional alto, en especial si hay un gran número de secuencias, que tienen una longitud considerable y se escoge un n alto. Las figuras 20 y 21 muestran los árboles que se generaron con las secuencias alineadas de nucleotidos y aminoácidos respectivamente.

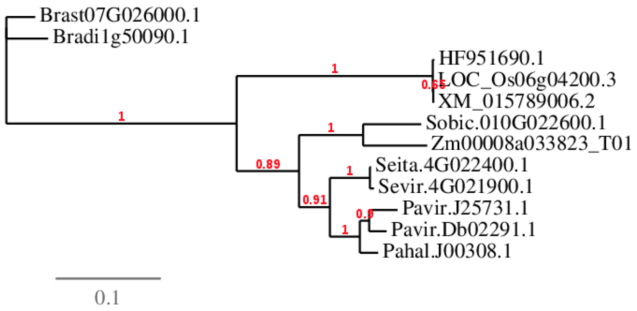


Figura 20: Resultados del dendograma utilizando el archivo de nucleótidos usando *bootstrapping*.

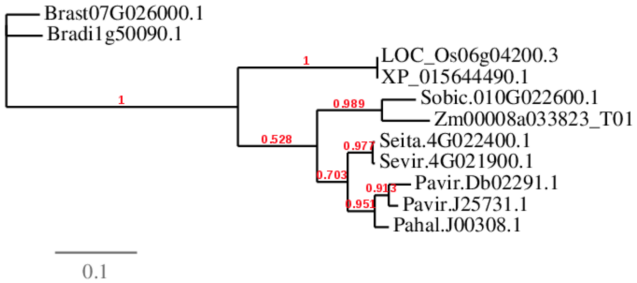


Figura 21: Resultados del dendograma utilizando el archivo de aminoácidos usando *bootstrapping*.

Referencias

[1] Moura, "Introduction Distance-Based Methods Character-Based Methods Conclusion Algorithms in Bioinformatics: Lecture 15-16: Phylogeny Reconstruction," 2010.

[2] "UPGMA Method." [Online]. Available: https://www.sequentix.de/gelquest/help/upgma_method.htm. [Accessed: 23-Aug-2018].

[3] "Phylogeny.fr: Home." [Online]. Available: http://phylogeny.lirmm.fr/phylo_cgi/index.cgi. [Accessed: 23-Aug-2018].

[4] "Simple Phylogenetic Tree < Phylogeny < EMBL-EBI." [Online]. Available: https://www.ebi.ac.uk/Tools/phylogeny/simple_phylogeny/. [Accessed: 23-Aug-2018].

[5] "ClustalW2 < Multiple Sequence Alignment < EMBL-EBI." [Online]. Available: https://www.ebi.ac.uk/Tools/msa/clustalw2/. [Accessed: 23-Aug-2018].

[6] "MUSCLE < Multiple Sequence Alignment < EMBL-EBI." [Online]. Available: https://www.ebi.ac.uk/Tools/msa/muscle/. [Accessed: 23-Aug-2018].

[7] "BLAST TOPICS." [Online]. Available: https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=B [Accessed: 23-Aug-2018].

[8] "Phytozome v12.1: Home." [Online]. Available: https://phytozome.jgi.doe.gov/pz/portal.html#. [Accessed: 23-Aug-2018].

[9] "UniProt." [Online]. Available: https://www.uniprot.org/. [Accessed: 23-Aug-2018].

[10] "BLAST: Basic Local Alignment Search Tool." [Online]. Available: https://blast.ncbi.nlm.nih.gov/Blast.cgi. [Accessed: 23-Aug-2018].

[11] "Weblet Importer." [Online]. Available: http://bioinformatics.psb.ugent.be/downloads/psb/Userman/treecon_bo [Accessed: 23-Aug-2018].