MIRANE FLORENCIO ZABALETA

Biologia

2019/05/10

AN7863 GENEAREN ANALISI PROTEIKOA

## Aurkibidea

[**Helburua** 2](#_Toc8300187)

[**Azterketaren oinarria** 2](#_Toc8300188)

[**Prozedura** 2](#_Toc8300189)

[Prozeduraren laburpena 11](#_Toc8300190)

[**Emaitzen azterketa** 11](#_Toc8300191)

[**Bibliografia** 11](#_Toc8300192)

## Helburua

**Aspergillus nidulans** onddoaren **AN7863** genearen analisi proteikoa gauzatuko dugu.

Geneak, DNAren parte dira, informazioa gordetzen duten unitateak dira. Hauek, proteinak sintetizatzeko behar den informazio genetikoa dute eta sintetizatuko luketeen proteinan oinarrituz, hauei buruzko informazioa eskuratuko dugu.

## Azterketaren oinarria

Analisia gauzatzeko datu base batzuek eskaintzen diguten informazioan oinarrituko gara.

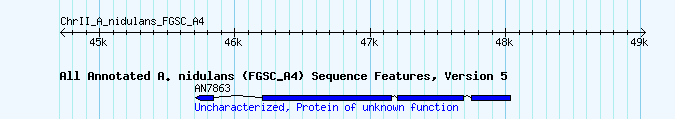
Programa hauekin geneak sintetizatzen duen proteinaren sekuentzia eskuratu eta antzekotasuna duten beste proteina batzuekin alderatuko ditugu, azkenik zuhaitz filogenetikoa eraikiz. Azken honekin proteinen arteko erlazio ebolutiboak aztertuko ditugu, proteina horren funtzioa determinatzen laguntzeko.

## Prozedura

Gene zenbakia hausaz aukeraturikoa da eta honi buruzko lehen informazioa hainbat datu baseetan eskura daitekeen arren, guk *AspGD* erabiliko dugu.

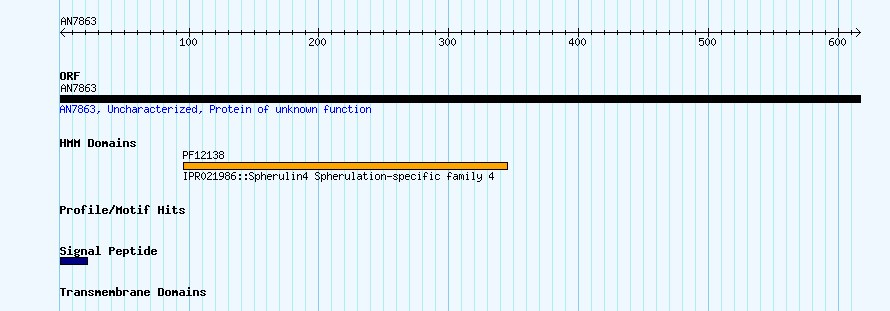
Orrialde horretan, *Aspergillus Genome Database*[1]-en zehazki, gure genea eta honek sintetizatzen dituen proteinei buruzko informazioa jasoko dugu.

Hasiera batean, informazio basikoa eskuratuko dugu “summary” deituriko leihatilan eta bertan lehen grafiko hau aurkituko dugu, non aukeratutako genearen kokapen kromosomikoa adierazten den, honek dituen exoi eta introi kopurua adieraziz.



Gure geneari erreparatuz, transkripzio prozesu bakarra ematen da, hau da, ARN eraketa prozesu bakarra. Horretaz gain, zati kodifikantea (exoiak) eta ez-kodifikantea (introiak) bereizi dezakegu, lehenetik 4 eta bigarrenetik 3 dituela hausnartuz.

Jarraian, orrialde berean geneak sintetizatuko dituen proteinen funtzioari buruzko informazioa jasoko dugu, “protein” izeneko leihatilan. Bertan hurrengo argazkia behatu dezakegu, honek proteinaren eremuak azaltzen ditu, zati bakoitza dagokion funtzio jakinarekin lotuz.



Kasu honetan, ez dago funtziorik definituta sintetizatzen duen proteinarako, gene zehatz hau aurretik ikertua izan ez delako izan daiteke. Hala ere, peptido seinale bat dugunez, bertan oinarrituko gara proteinari buruzko informazio gehiago eskuratzeko.

**Peptido seinalea**, peptido motz bat da, orokorrean 16-30 aminoazido luzerakoa. Hau N-terminalean presente dago sintetizatu berriak diren proteinetan eta hauen helburua erabaki edo ezarriko du, zelulei proteinak translokatzeko abisatuz.

Azken hauek sekrezio rutara bideratuak daude, hau da, proteinak ingurugiro extrazelularrera isurtzen dituen bidera. Hona zuzenduriko proteina gehienak exozitosia jasaten duten arren, pasabide honetan proteinak prozesatze dira ere, batik bat, zelula mintzaren parte izango direnak edo bertan biziko direnak.

Gure geneari dagokionez, ziurrenik jariatua izango da, peptido seinaleaz gain ez dugulako beste inongo asoziazio edo funtzio zehatzik, ez beste inolako informaziorik.

Behin informazio hau eskuraturik, geneak sintetizatzen duen proteinaren sekuentzia kopiatu eta Blast-a egingo dugu. Honela, gure geneak sintetizatzen dituen proteinaren ortologoak direnak bilatuko ditugu, gure proteinaren jatorria eta honela bere funtzioari buruzko informazioa eskuratzeko.

Hau da geneak sintetizatzen duen proteinaren sekuentzia:

**>AN7836 A.nidulans**

MKSISLSLSLATTILLSTAPLVTAKTYTTNIPVKEIQGAWSIHGNSISWTEDGFKTSIDC

DDQDGNKKLSLSNNKKFAGCCLSGQRLVGSPDTAFDCCANAHDLAGSKETGYRCCPEGET

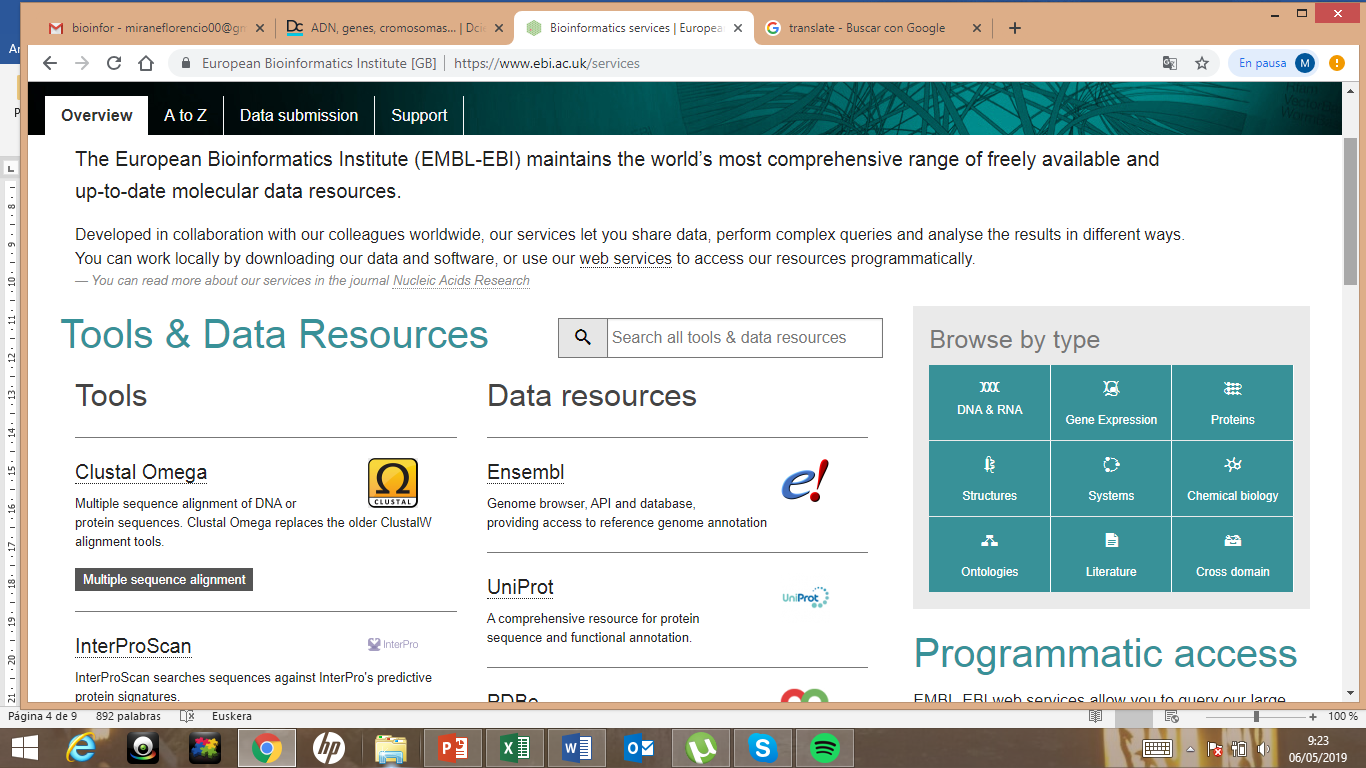
YDGVTCKADDPVCQNGKLLKNGECACPKGTKEDENGICAPAKCSSGLETGKCYTFTGENG

HRLGFGGSWFITAPESMSLKSGRFKLCKDEECKAGETINPADQIYIKDIHGNPGNGALPN

RWLNSAMNGNHVGKTDNFAQAGKFSMTKWPCGKYCLGGFDYGLGPACPSNTPALTFFQND

KQACVPFDFTEVPCDVKAEANNCIWKTNEDQCCGGAVDCEGN\*

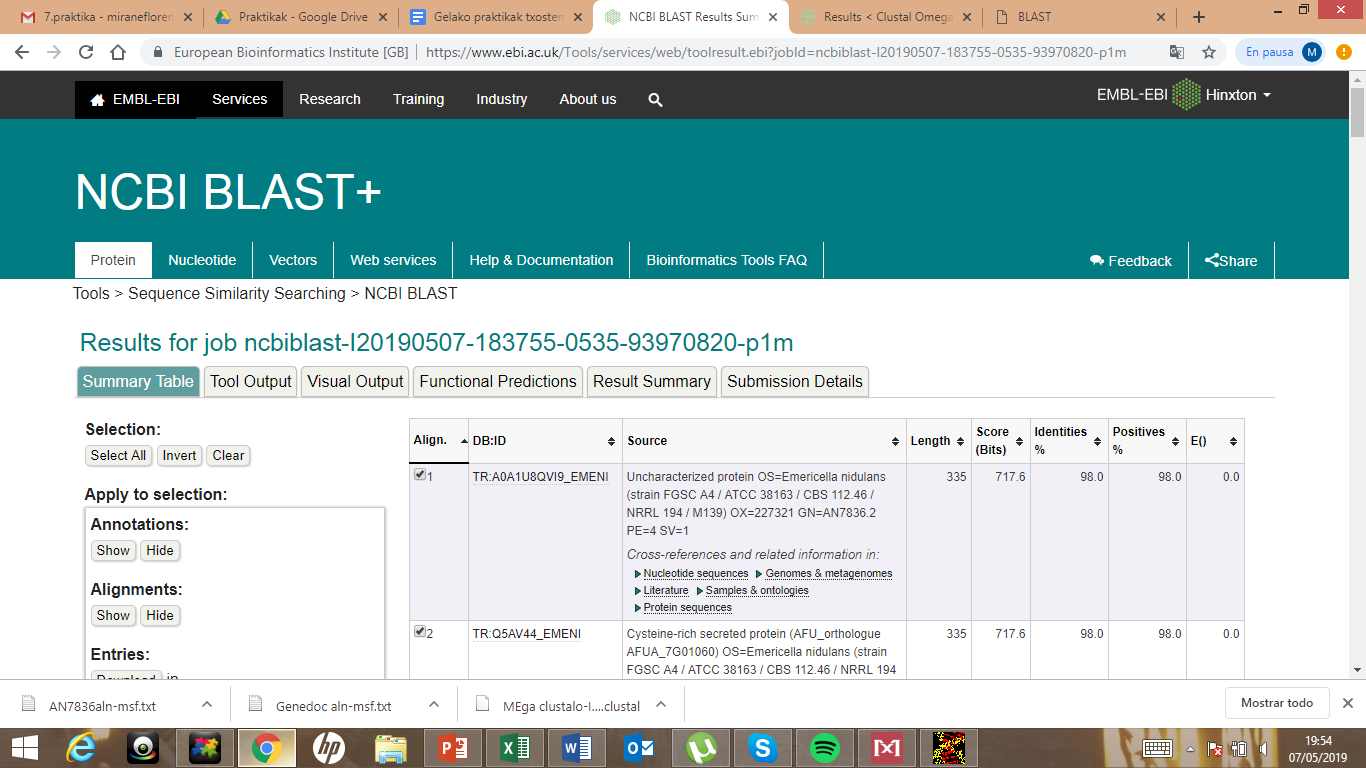
***“EBI search”***[2]erabiliko dugu, honek informazio biologikoa era uniforme eta erraz batean eskuratzea ahalbidetzen du.

Hainbat datu desberdin jaso ditzazkegu datu base honetan:

Argazkian ikus daitekeen bezala, denetariko informazioa eskura dezakegu, baina guk proteinaren sekuentziaren *blast*-a eskuratuko dugu, *NCBI BLAST+*[3] portal elektronikoan.

Blast-ak Antzekotasuna duten edo ortologoak diren proteinak bilatzen ditu, hauen sekuentzien alineazioaren bidez. Kasu honetan gure proteinak dato basean dituen proteinekin konparatuko du. Esan beharra dago, honek logaritmo heurestiko bat erabiltzen duela, ondorioz, ezin izango da ziurtatu bilatzen duen informazioa guztiz zehatza denik.

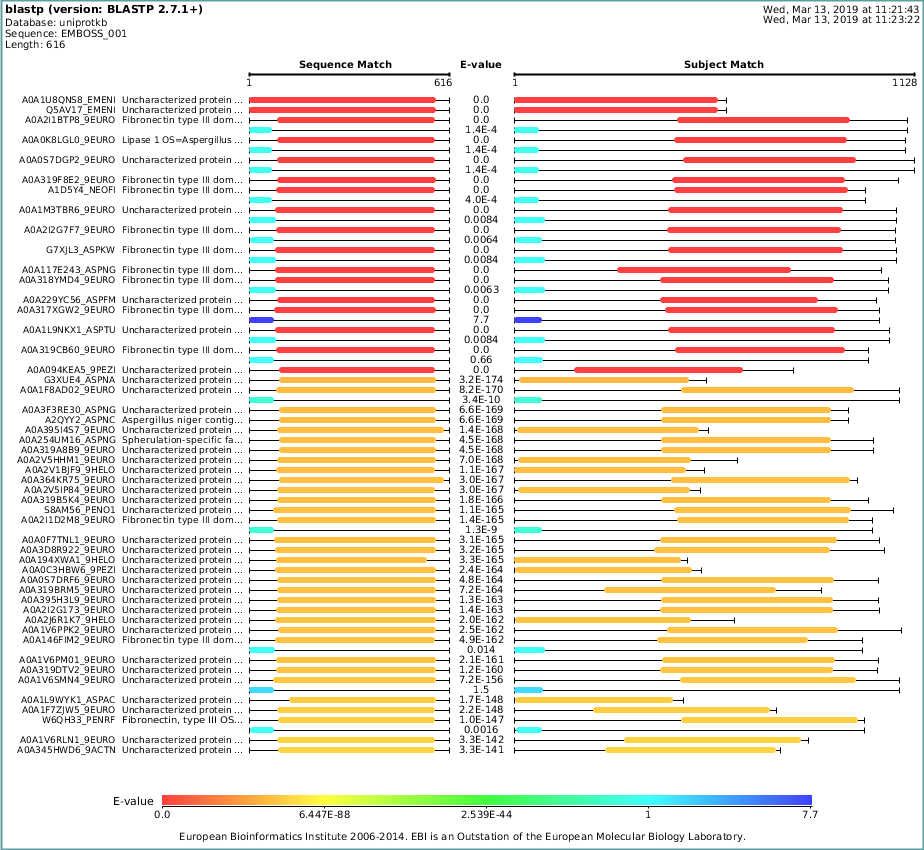
Portal honetan, informazioa modu desberdinean eskura dezakegu, leihatila bakoitzak informazioa modu batetara ematen digu.



Lehenengoan, bilaturiko proteina ortologo desberdinak ditugu, beraien izen funtzio eta antzekotasunari buruzko informazioa. Bertan, “identity” izeneko zutabean, proteinaren aminoazido guztien artean kontserbazio portzeintaia ikusi dezakegu. Baita *score* eta *E-value* zutabeak ere, hauek kontsertbatutako zatiei buruzko informazioa ematen digutelarik. Gure kasuan, *E-value* erabiliko dugu.

“Visual output”-en grafiko batean irudikatua kontserbazioari buruzko informazioa jaso dezakegu *E-value* eskalarekin baloratua eta “Functional prediction”-en, aurkitutako *hit* desberdinak.

“Visual output”-en honako grafiko hau eskuratu dugu:



Bertan gure proteina eta bere ortologoak dituzten antzekotasunak ikusi ahal izango ditugu, azken hauen eta gurearen artean dauden kontserbazio guneak aztertuz.

Programa informatiko honek, *Query* bilaketa terminoaren bidez aurkitzen ditu ortologoak eta bilaturiko termino antzekoak, *hit* bezala adierazten dira. Kontserbatutako guneen arteko antzekotasuna adierazteko, *E-value* izeneko eskala erabiltzen da eta honek ausaz sekuentzia aurkitzeko aukera definitzen du, zenbat eta hurbilago zerora, zenbat eta esanguratsuagoa izango da sekuentziak bat etortzea.

Orduan, esan dezakegu gorriz azaltzen direnak, guztiz kontserbaturiko zatiak direla, laranjak guztiz kontserbaturik ez daudenak eta urdinak gutxien, koloreak iluntasuna erakusten duen heinean.

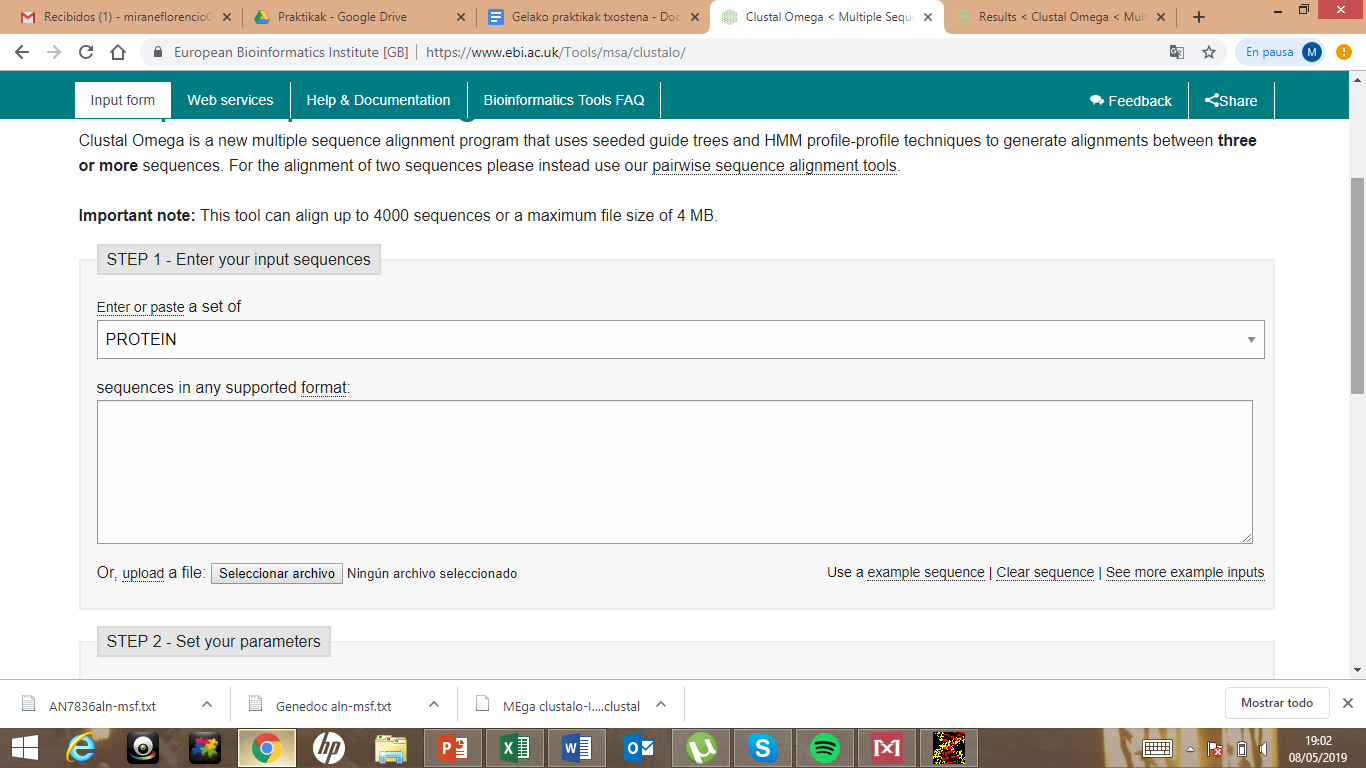
Hala ere, *E-value*-k ez du proteina osoa ebaluatzen, kontserbatutako zatia besterik ez. Ondorioz, bakarrik kontserbaturiko zatietan duten antzekotasuna jakin dezakegu, ez proteina osoaren baitan dagoena.

Behin proteina ortologoak aurkituta, proteina desberdinen *hit*-ak hartu eta elkarren artean alineatu egingo ditugu. Honela, beste programa batzuen bidez informazio gehiago eskuratu ahal izango dugu; Batik bat, proteinaren eboluzioari buruzkoa.

Aurretik aipatutakoa burutzeko, antzekotasuna duten proteinen *fast-*ak biltzen dituen dokumentu bat sortu behar dugu, bertan proteina ortologo horien sekuentziak eta identifikatu ahal izateko izen bat. Behin sekuentzia guztiak dokumentu berean bilduta ditugula, *Clustal Omega[4]* atari elektronikoan sartuko gara.

Datu-base honekin, sekuentzien linealizazioa egingo da hainbat proteinen artean, gure kasuan 16 izango dira, sekuentzia guztiak bata bestearen azpian jarri eta antolatuz, kontserbazio maila altuena duen alineamendua emateko. Eremu kontserbakorrak adieraziz, aminoazidoak senkuentzien zein tokitan mantentzen diren jakingo dugu.

Behin orrialdean, hainbat pauso jarraitu behar dira alineamendua lortzeko. Lehenengo *STEP 1* -en, aurretik aipatutako dokumentuan bateratutako 16 sekuentziak itsatsi behar ditugu.



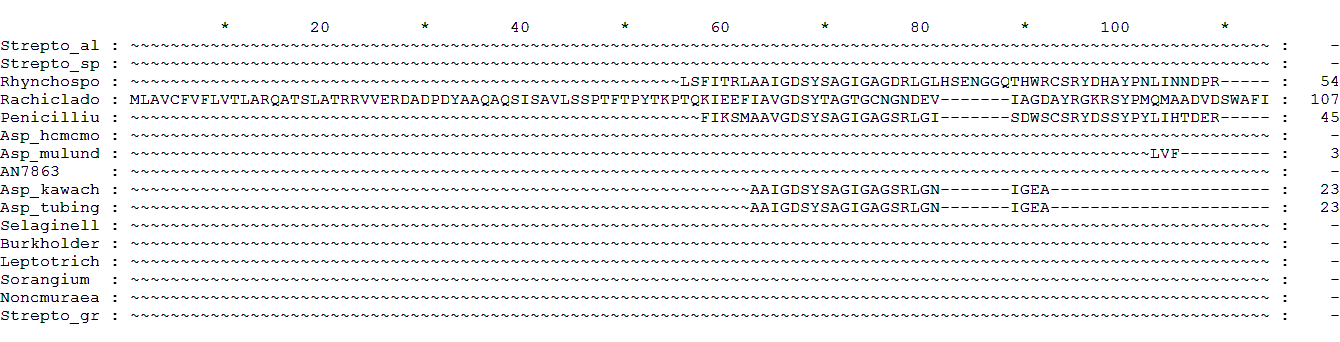
Ondoren, *STEP 2* -an emaitzaren formatua hautatuko dugu. Gure kasuan bi formatutan jaso nahi dugu informazioa, dagokien aplikazioetan alinamendu horretatik informazioa eskuratu ahal izateko.

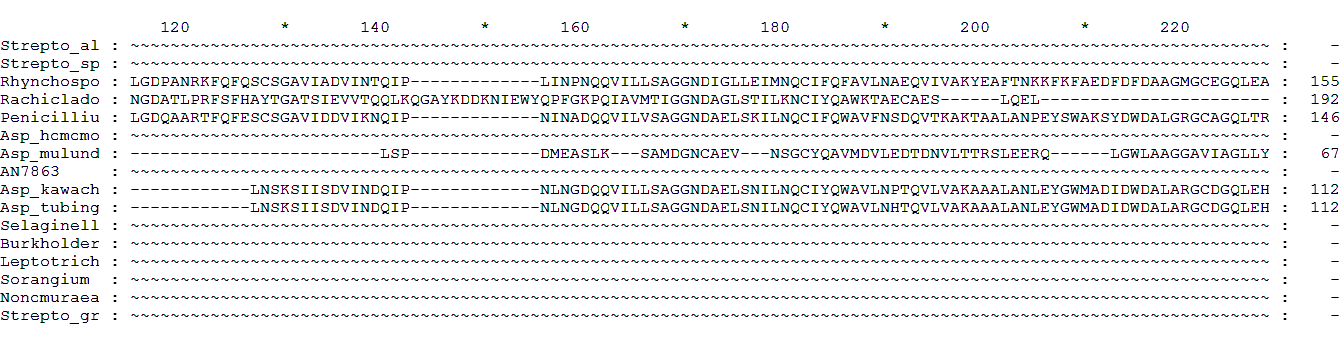
Emaitza lortzean, kontserbazioaren lehen irudikapen bat lortzen dugu. Artxibo hau gorde behar dugu bi formatuetan, estentsioa mantenduz.

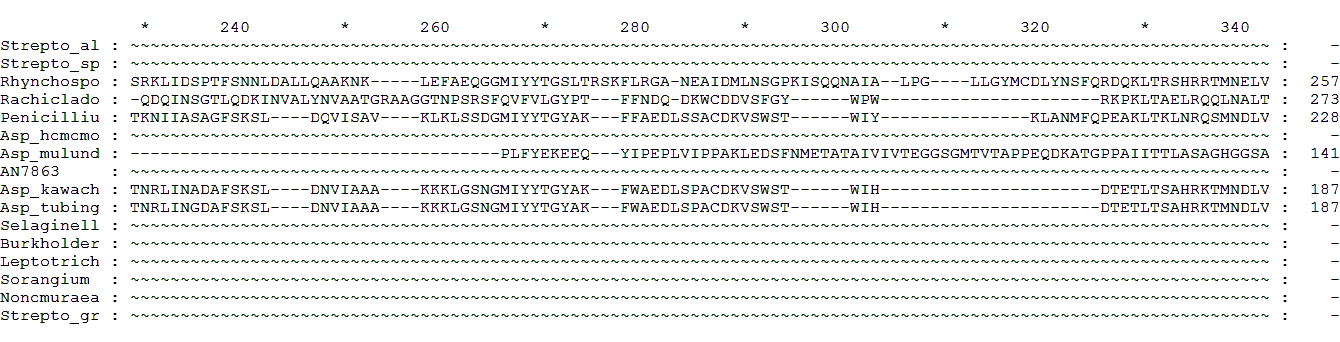
* *Msf* formatua, ondoren *GeneDoc[5]*-en zabaldu ahal izateko.
* *Clustalw* formatua, *MEGAX[6]*-en zabaldu ahal izateko

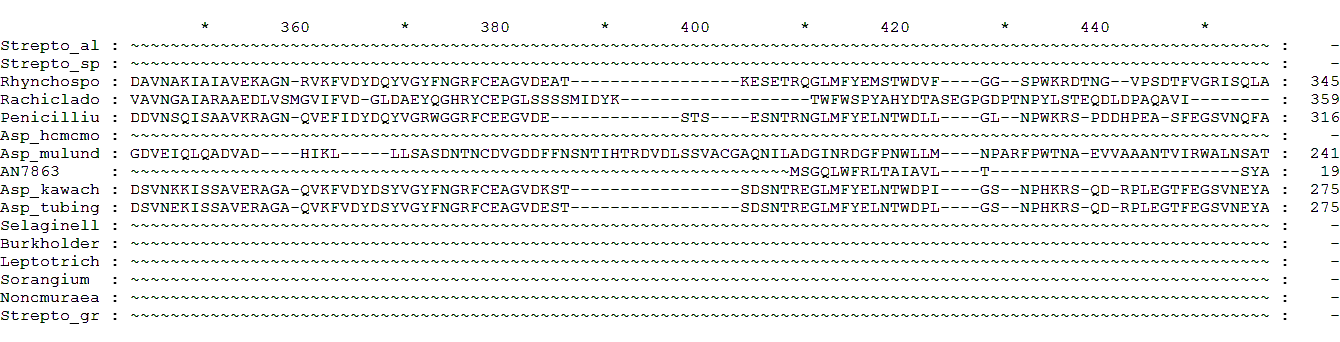
Dagokien dokumentuak izanda, aplikazioetan zabalduko ditugu. Lehenengo *GeneDoc*-eko emaitzak aztertuko ditgu. Programa honetan, gure proteina eta bere ortologoen arteko sekuentzien arteko kontserbazioak garbi ikusi ahal izango ditugu, oso era sinplean adierazten du eta

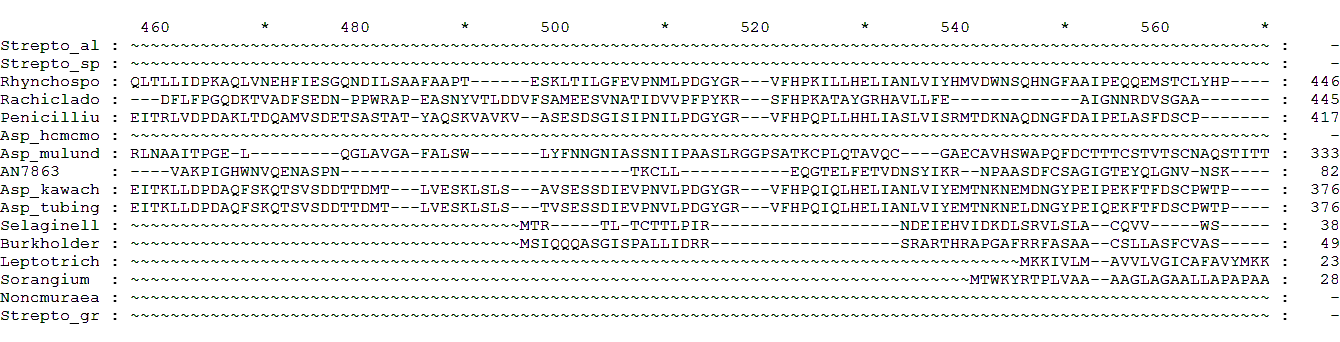
Honakoa eskuratu dut *msf* artxiboa bertan irakitzean:

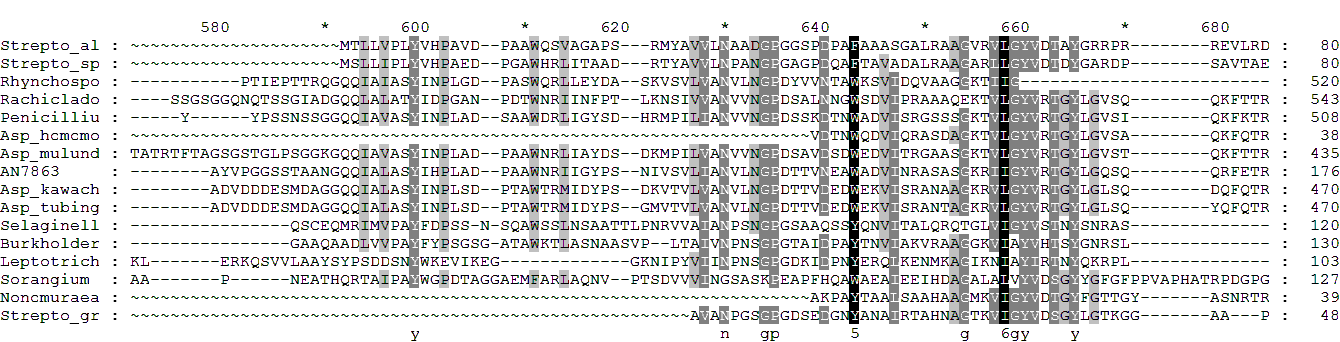


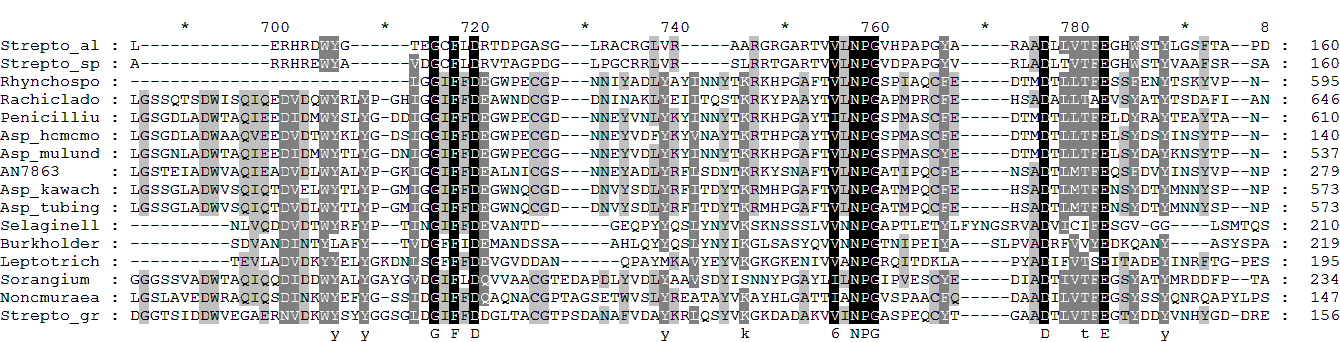


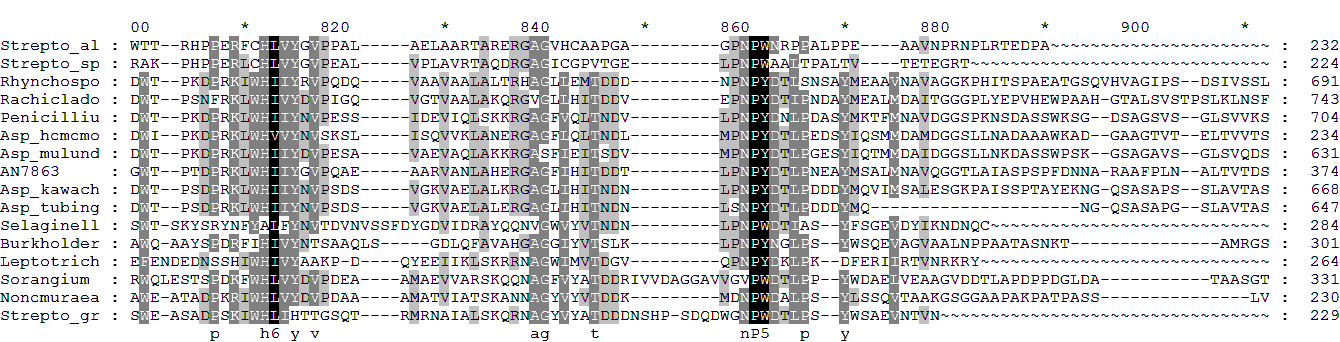


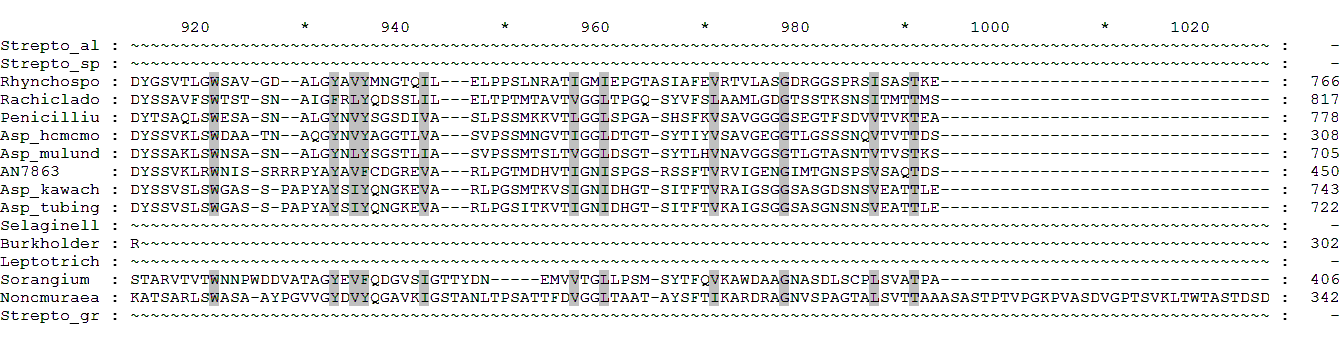


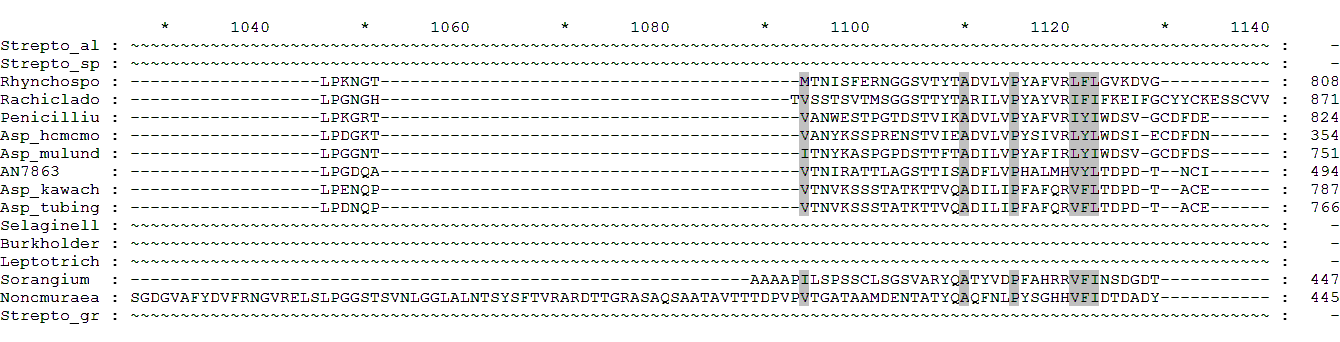


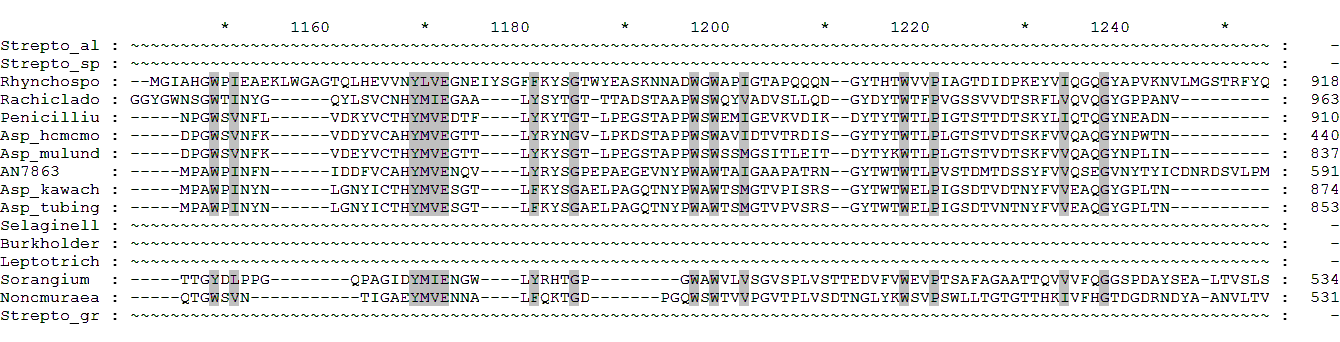


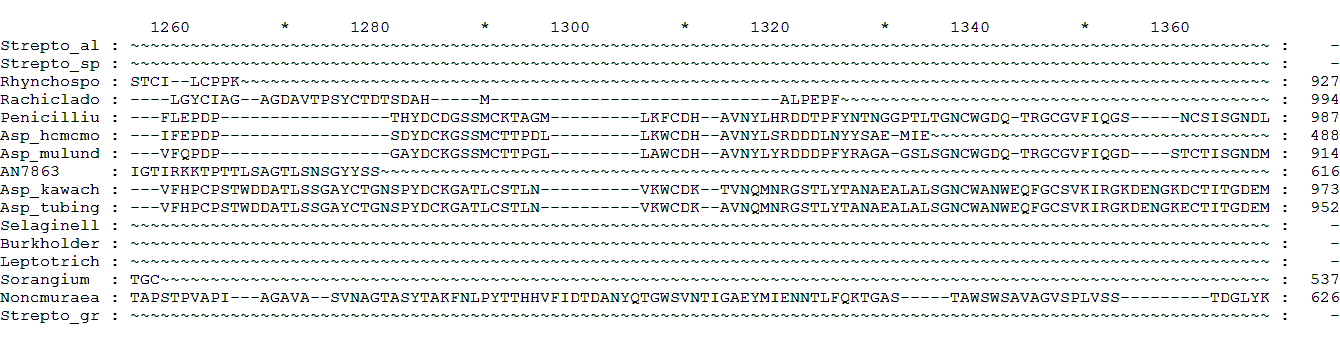


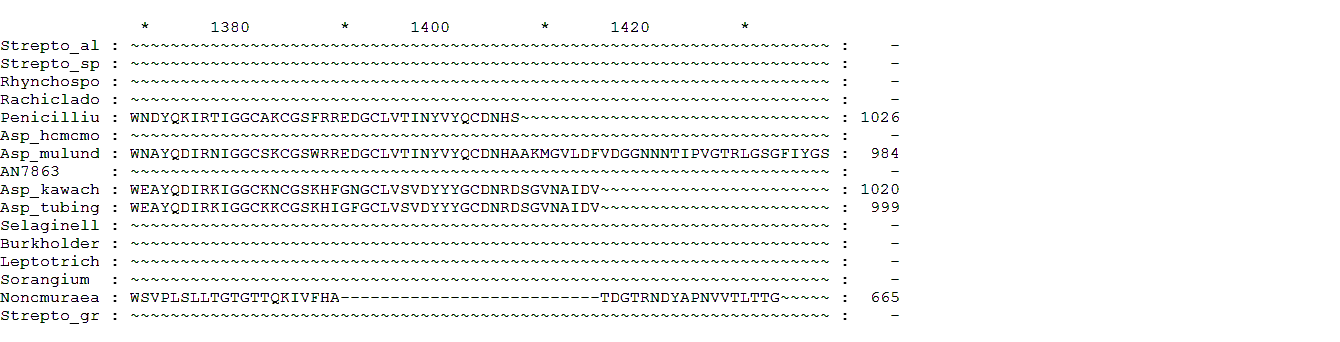












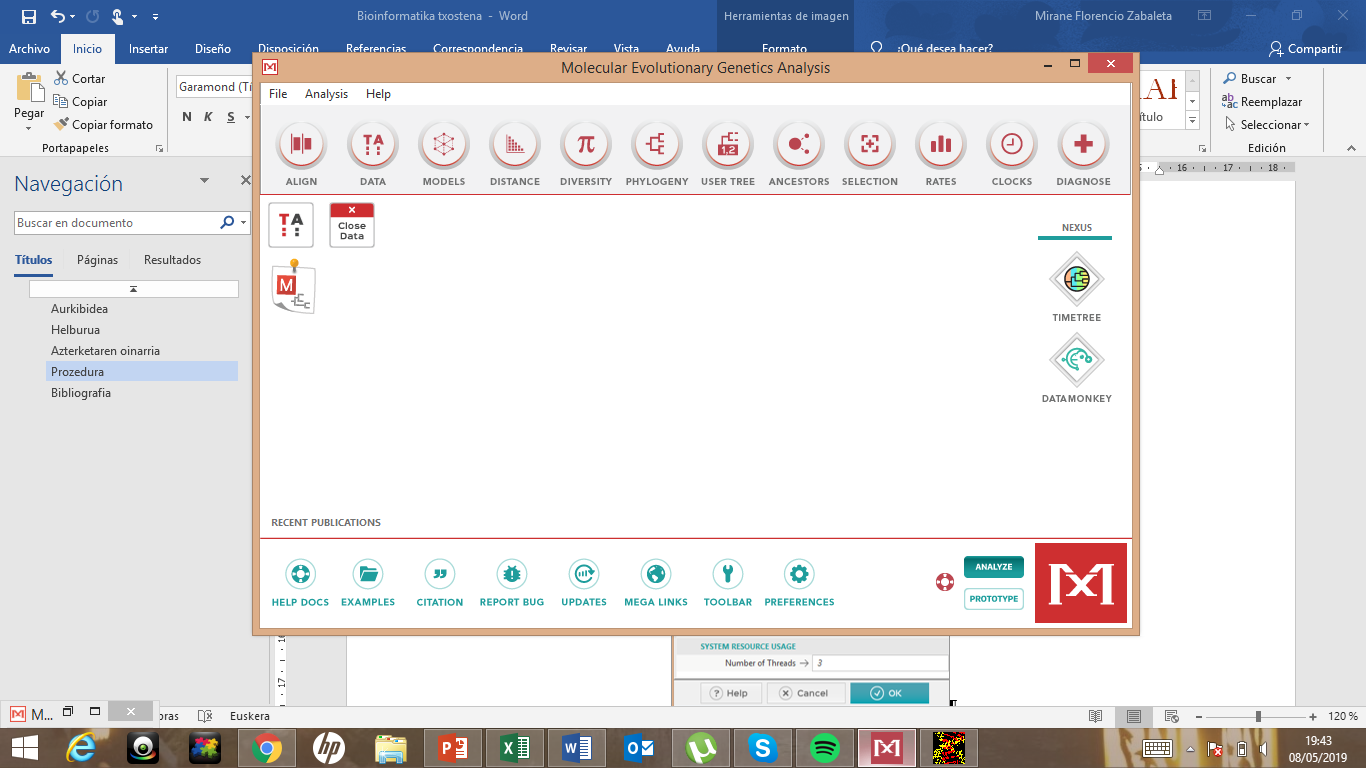
Grafiko honetan kontserbazioak behatu ditzakegu kolore eskala baten bidez. Beltzak guztiz kontserbatzen direnak, grisak partzialki eta txurian ez dago inongo kontserbaziorik. Kontserbazioaren faktorea aminoazidoen araberakoa da, aminoazidoak berdinak ez izan arren, izaera berdina badute, partzialki kontserbatzen direla onar daiteke. Eskuin aldean ere, zenbaki batzuk aurkitzen ditugu. Hauek aminoazidoak proteinan duten kokapena adierazten du.

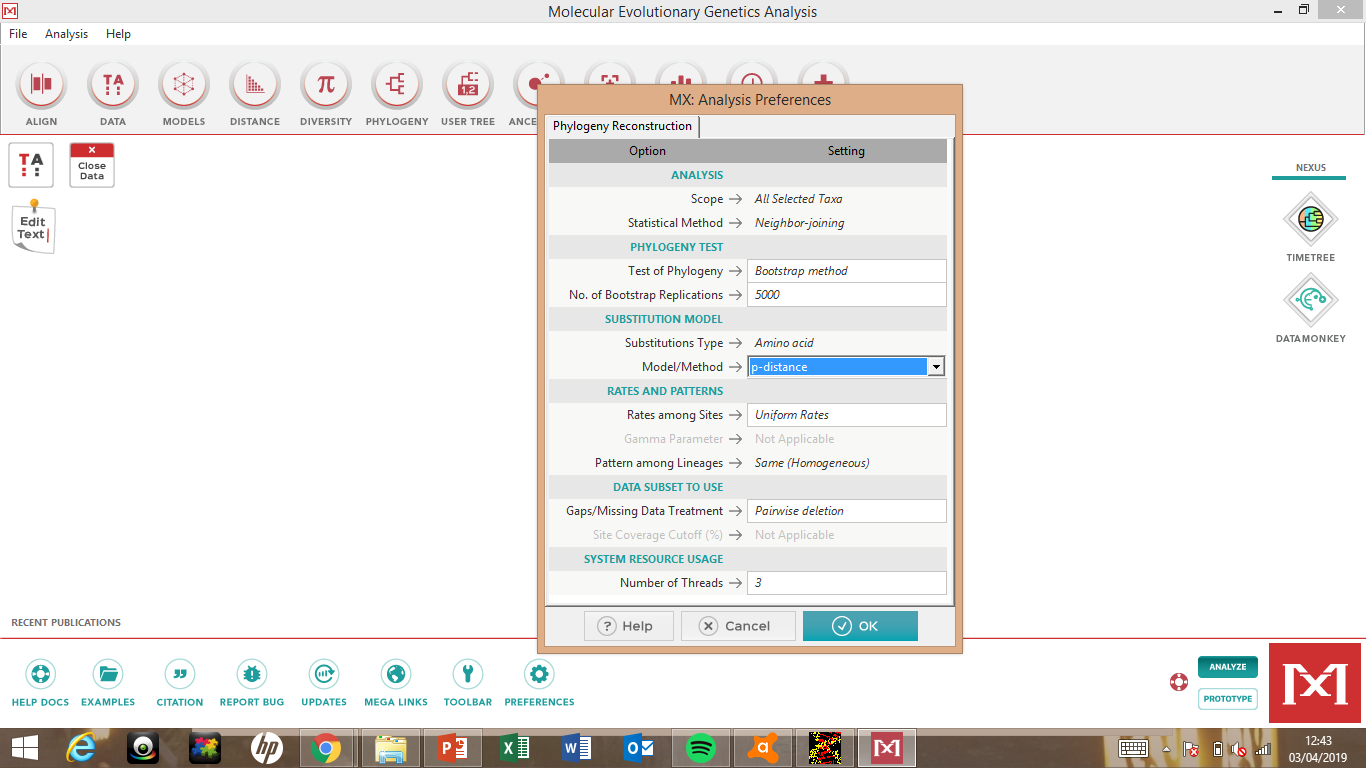
Jarraian *MEGAX-*en zuhaitz filogenetikoa eraikiko dugu, honela jatorrizko proteinari buruz informazioa eskuratu ahal izango dugu. Hau, *Clustal*-en egindako linealizaziotik abiatzen da, baina programak ireki ahal izateko honek berak artxibo hori eraldatu behar du, *Mega* artxibo bihurtuz.

Ondoren, eraldatutako dokumentua ireki eta behar dugun informazioa bilatuko dugu.

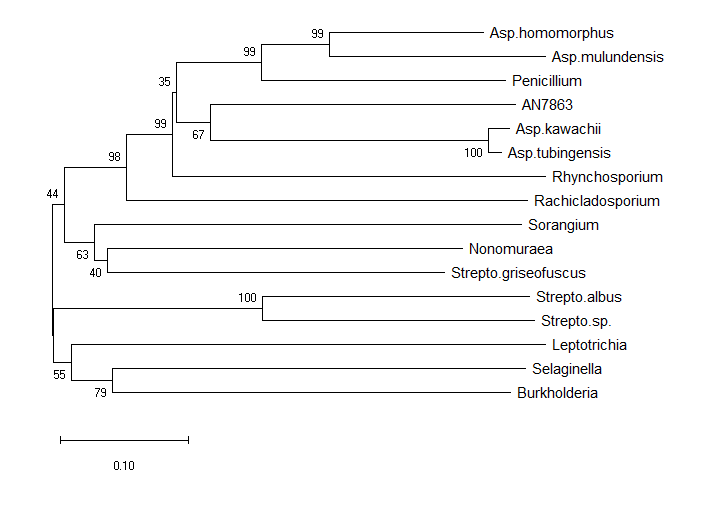
Aplikazioak hainbat aukera desberdin ematen dizkigu proteinari buruzko informazioa eskuratzeko.

Zuhaitz filogenetikoa sortzeko *phylogeny* aukerari eman eta *neighbour-joining* sakatuko dugu, jarraian datorren argazkian azaltzen diren gelaxkak aukeratuz edo betez.





Honako zuhaitz filogenetikoa eskuratuko da:



Zuhaitz filogenetikoak[7] organismoen arteko harreman ebolutiboak erakusten dituen diagrama da. Eskuinaldean gaur egungo espezieak ditugu eta ezkerraldean aitzindariak, zenbat eta denboran atzerago egon komunean duten aitzindaria, gauza gutxiago izango dituzte amankomunean.

Zenbakiek ziurgabetasuna adierazten dute adarraren kokapen luzerari buruz, ehunekoetan. Lerroen azpiko zenbakia zenbat eta handiagoa, programaren ziurtasuna proteinaren kokapen ebolutiboarekiko handiagoa izango da (99 jartzen badu, honen ziurgabetasuna %1-a izango da).

Azpiko distantziak *aminoacides per site* adierazten du. Programaren algoritmoak sekuentzien alineamentuetan hauen artean posizio bakoitzean dauden kontserbazio desberdintasunak neurtu eta horren arabera kokatzen ditu zuhaitzaren adarrak. Hauek denborarekin asoziatuz.

Hau kontuan izanik, ikus dezakegu *Asp. Kawachii-*rekin edo Asp. *Tubingensis-*ekin antzekotasun handiagoa duela *Penicillium*-ekin baino, arbaso gertuagoak dituztelako.

### Prozeduraren laburpena

1. *AspGD*-en gure geneari buruzko informazio bilatu, honek sintetizatzen duen proteina, bere funtzioa eta sekuentzia.
2. Proteinaren sekuentzia izanik, honen ortologoak bilatuko ditugu *Blast*-a eginez.
3. Proteina eta bere ortologoen *fast*-ekin dokumentu bat osatu eta sekuentzien alineazioa gauzatu *Clustal Omega* portal elektronikoaren bidez.
4. Alineazioa bi formato desberdinetan deskargatu, *msf eta clustalw*, hauek *GeneDoc* eta *MegaX-*en zabaldu ahal izateko.
5. *Genedoc-*en sekuentzien arteko kontserbazioak behatu proteina eta haren ortologoen artean.
6. Azkenik, *MegaX*-en zuhaitz filogenetikoa eraiki proteinaren jatorria eta eboluzioari buruz gehiago jakiteko.

## Emaitzen azterketa

Aurretik prozesuan zehar aipatu bezala, gure geneak sintetizatzen duen proteinak ez du funtzio definitu bat, peptido seinalean oinarritu arren, hainbat aukera ditugu gure proteinaren funtzio izan daitezkeenak baina ez dugu ziurtasunik honi buruz. Lehen orrialdean aurkitutako informazioari dagokionez, ziurrenik proteina jariatua izango da sekrezio rutan.

Ortologoak bilatu eta hauen arteko sekuentziak alderatu ditugu ere *GeneDoc-*en, bertan ikus dezakegu proteinaren zonalde askotan ez dagoela aminoazidoen kontserbaziorik baina baina ba dago antzekotasun minimo bat denen artean, gehienbat 160. aminoazidoan.

Azkenik zuhaitz filogenetikoa eraiki dugu. Honetan *Asp. Kawachii-*rekin edo Asp. *Tubingensis-*ekin antzekotasun handia dutela ikusi dugu. Hala ere, hauen arbasoa identifikatzerakoan ziurgabetasuna %33-koa da, hortaz ezin dugu zehazki jakin hauen funtzio bera edo antzekoa beteko duen.

Ondorioz, azken bi hauekin arbaso komuna izan arren, ziurrenik sekrezio rutara iristean jariatua izango da ingurune extrazelularrera golgi aparatua edo jariaketa sisteman parte hartzen dutenen bidez.

## Bibliografia

[1]**AspGD.**

The Aspergillus Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. *Nucleic Acids Res* 42 (1);

D705-10.Cerqueira GC, Arnaud MB, Inglis DO, Skrzypek MS, Binkley G, Simison M, Miyasato SR, Binkley J, Orvis J, Shah P, Wymore F, Sherlock G, Wortman JR (2014).

[2]**EBI Search.**

*The EMBL-EBI search and sequence analysis tools APIs in 2019.*.   
Nucleic Acids Research published online on April 12, 2019.

Madeira F., Park Y.M., Lee J., Buso N., Gur T., Madhusoodanan N., Basutkar P., Tivey ARN., Potter SC., Finn RD., Lopez R. (2019)

[3] **NCBI BLAST+.**

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17); 3389-3402.

Altschul, Stephen F., Gish, Warren, Miller, Webb, Myers, Eugene W., and Lipman, David J. (1990). Basic local alignment search tool. J. Mol. Biol. 215; 403-410. Gapped BLAST is described in Altschul, Stephen F., Madden, Thomas L., Schaffer, Alejandro A., Zhang, Jinghui, Zhang, Zheng, Miller, Webb, and Lipman, David J. (1997).

[4]**Clustal Omega**.

**Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.** Sievers F., Wilm A., Dineen D., Gibson T.J., Karplus K., Li W., Lopez R., McWilliam H., Remmert M., Söding J., Thompson J.D. and Higgins D.G. (2011)

[5] **GeneDoc.**

GeneDoc: A tool for editing and annotating multiple sequence.

[KB Nicholas](https://www.scienceopen.com/search#author/3d743c02-aca9-47b1-98d2-1cc81b78a323), [HB Nicholas](https://www.scienceopen.com/search#author/3d35531b-7d14-4d97-abc9-a880f0e1f580), [DW Deerfield](https://www.scienceopen.com/search#author/e2660b87-4776-4b08-a2ce-99e0b669eb79), [HBJ Nicholas](https://www.scienceopen.com/search#author/1c95194f-4b1d-465c-9c42-f9bb6f8c6cd8), [K Nicholas](https://www.scienceopen.com/search#author/cd3032c2-8d79-4799-935a-1e97b48e133b), [HJ Nicholas](https://www.scienceopen.com/search#author/b9442f2a-cd97-4462-b82e-94ccc84d87f0), [K. R. Nicholas](https://www.scienceopen.com/search#author/f525d8fe-5a79-47d0-af20-e45c74142844), [H. B. J. Nicholas](https://www.scienceopen.com/search#author/e36bbd08-d5c8-47bf-8308-1311053cfc73), [A. Nicholas](https://www.scienceopen.com/search#author/efc0529f-7c48-4f0d-8049-75069f60588c), [DW Deerﬁeld](https://www.scienceopen.com/search#author/792def6a-5dec-465e-b837-00f6bc47dd2e), [H Nicholas](https://www.scienceopen.com/search#author/02e61d95-83f5-4eb0-af0d-1e80301b01ab), [H. Gauch](https://www.scienceopen.com/search#author/6ec35a8d-8090-4fca-8796-28efe0f710fd) (1997).

[6]**MEGAX.**

MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Molecular Biology and Evolution 35:1547-1549Sudhir Kumar, Glen Stecher, Michael Li, Christina Knyaz, and Koichiro Tamura (2018)

[7] **Zuhaitz filogenetikoa.**

"Taxonomy and phylogeny (Taxonomía y filogenia),"

Robert Bear, David Rintoul, Bruce Snyder, Martha Smith-Caldas, Christopher Herren y Eva Horne.