ANÁLISIS DE UNA SECUENCIA USANDO
HERRAMIENTAS DISPONIBLES EN LA RED

Samuel Pintos González

1. SELECCIÓN DE LA SECUENCIA E IDENTIFICACIÓN DE ORFs

He utilizado “ORF finder” para el primer ejercicio, copiando la secuencia proporcionada en esa página web con los parámetros por defecto, salvo en la sección “ORF start codon to use”, ya que como no sé a qué organismo pertenece la secuencia, he seleccionado “ATG and alternative initiation codons”

Elijo ORF1 [Figura 1] ya que encapsula al resto de ORFs en la secuencia y es el más grande.

2. ANÁLISIS FUNCIONAL DE LA SECUENCIA

La secuencia de aminoácidos que viene dada en la propia herramienta de “ORF finder” al seleccionar el ORF1 es:

MNYSFTEKKRIRKSFAKRENVLEVPFLLATQIDSYAKFLQLENAFDKRTDDGLQAAFNSIFPIVSHNGYARLE
FVYYTLGEPLFDIPECQLRGITYAAPLRARIRLVILDKEASKPTVKEVRENEVYMGEIPLMTPSGSFVINGTER
VIVSQLHRSPGVFFEHDKGKTHSSGKLLFSARIIPYRGSWLDFEFDPKDLLYFRIDRRRKMPVTILLKALGYN
NEQILDIFYDKETFYLSSNGVQTDLVAGRLKGETAKVDILDKEGNVLVAKGKRITAKNIRDITNAGLTRLDVE
QESLLGKALAADLIDSETGEVLASANDEITEELLAKFDINGVKEITTLYINELDQGAYISNTLRTDETAGRQAA
RVAIYRMMRPGEPPTEEAVEQLFNRLFFSEDSYDLSRVGRMKFNTRTYEQKLSEAQQNSWYGRLLNETF
AGAADKGGYVLSVEDIVASIATLVELRNGHGEVDDIDHLGNRRVRSVGELTENQFRSGLARVERAVKERLN
QAESENLMPHDLINAKPVSAAIKEFFGSSQLSQFMDQTNPLSEVTHKRRVSALGPGGLTRERAGFEVRDV
HPTHYGRVCPIETPEGPNIGLINSLSVYARTNDYGFLETPYRRVIDGKVTEEIDYLSAIEEGRYVIAQANADL
DSDGNLIGDLVTCREKGETIMATPDRVQYMDVATGQVVSVAASLIPFLEHDDANRALMGANMQRQAVP
CLRPEKPMVGTGIERSVAVDSATAIVARRGGVVEYVDANRVVIRVHDDEATAGEVGVDIYNLVKFTRSNQS
TNINQRPAVKAGDVLQRGDLVADGASTDLGELALGQNMTIAFMPWNGYNYEDSILISEKVAADDRYTSI
HIEELNVVARDTKLGAEDITRDIPNLSERMQNRLDESGIVYIGAEVEAGDVLVGKVTPKGETQLTPEEKLLR

AIFGEKASDVKDTSLRMPTGMSGTVIDVQVFTREGIQRDKRAQSIIDSELKRYRLDLNDQLRIFDNDAFDR
IERMIVGQKANGGPMKLAKGSEITTEYLAGLPSRHDWFDIRLTDEDLAKQLELIKLSLQQKREEADELYEIK
KKKLTQGDELQPGVQKMVKVFIAIKRRLQAGDKMAGRHGNKGVVSRILPVEDMPYMADGRPVDIVLNP
LGVPSRMNIGQILEVHLGWAAKGIGERIDRMLKERRKAGELREFLNKLYNGSGKKEDLDSLTDEEIIELASN
LRKGASFASPVFDGAKESEIREMLNLAYPSEDPEVEKLGFNDSKTQITLYDGRSGEAFDRKVTVGVMHYLK
LHHLVDEKMHARSTGPYSLVTQQPLGGKAQFGGQRFGEMEVWALEAYGAAYTLQEMLTVKSDDVNGR
TKMYENIVKGEHKIDAGMPESFNVLVKEIRSLGLDIDLERY

Copio la secuencia de aminoácidos en protein BLAST seleccionando la base de datos "Reference proteins (refseq_protein)". A continuación me aparece que su cobertura es el 100% de la secuencia con un porcentaje de identidad del 100% y un E-valor de 0.0 [Figura 2]. Pertenecería a la especie *Neisseria gonorrhoeae*, donde la descripción de esta proteína sería "DNA-directed RNA polymerase subunit beta".

Con core nucleotide database (core nt) para el nucleotide blast, me aparece la misma cobertura, E-valor y porcentaje de identidad pero con descripciones distintas una de otra, a pesar de que el organismo coincide también [Figura 3].

3. COMPARACIÓN EVOLUTIVA

WP_003690105.1 Neisseria gonorrhoeae (el organismo al que pertenece la secuencia)
WP_119518745.1 Pseudomonas aeruginosa
WP_353524646.1 Escherichia coli
[Figura 4].

Haciendo el alineamiento a 3Bits para ser más estrictos que a 2Bits, las 2 regiones regiones que tienen una conservación más grande (en cuanto a longitud de la cadena de amino acidos sin variaciones) entre estos 3 organismos, siendo estas para *Neisseria gonorrhoeae* son:
aminoácido 527 al 603 y aminoácido 1260 al 1361. Hay varias regiones donde faltan datos (en gris, siendo la más grande la del aminoácido 420 al 444) y hay diferentes regiones bastante variables [Figura 5].

4. ANÁLISIS ESTRUCTURAL

Utilizando la herramienta PROSITE, introduzco la secuencia de aminoácidos que me dio ORFfinder.

Obtengo un dominio proteico entre los aminoácidos 1104 y 1116 (correspondiendo al código “PS01166” y con la serie de aminoácidos “GDKMAGRHGNKGV”), el cual coincide con una zona muy conservada entre el organismo de la secuencia y el resto de organismos del alineamiento (aunque no está dentro de las zonas conservadas más grandes que se han destacado previamente).

Según los datos aportados por la propia herramienta seleccionando la secuencia conservada, contiene dos lisinas muy bien conservadas que son parte del sitio activo del extremo C-terminal de todas las cadenas beta y este sería el motivo por el que varía tan poco entre especies.

ANEXO: IMÁGENES DE APOYO AL TRABAJO Y SU DESCRIPCIÓN

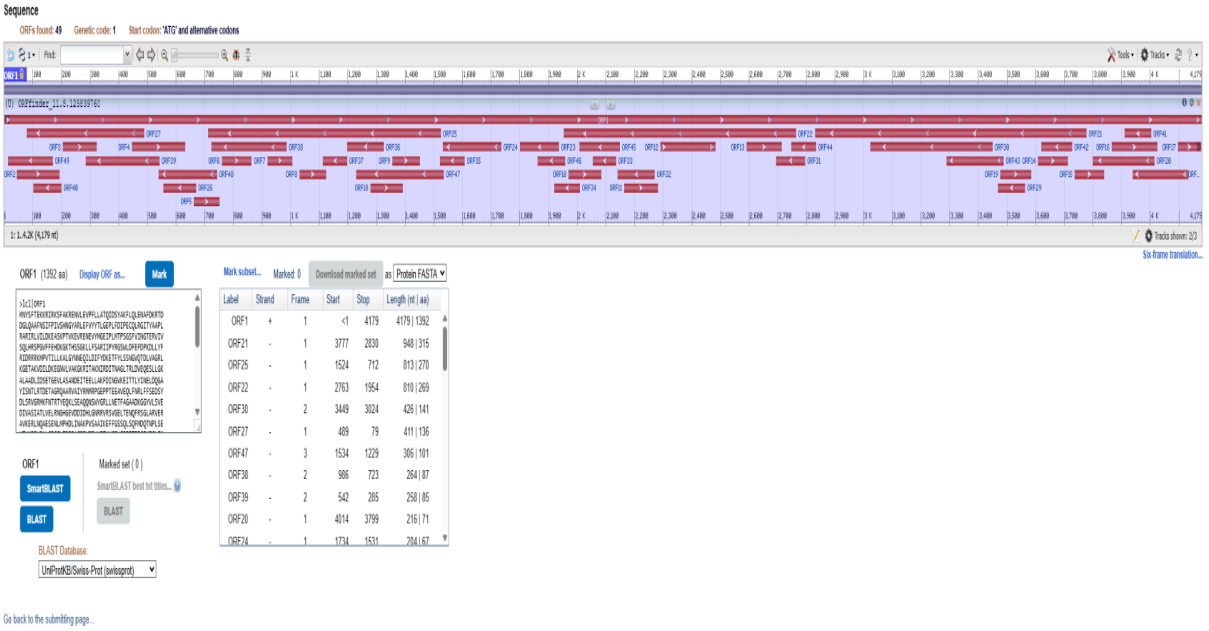

Figura 1. Representación de ORFfinder de todos los ORFs de la secuencia seleccionada en barras horizontales rojas, siendo ORF1 el primero y el más grande, abarcando al resto de ORFs

Query Cover	E value	Per. Ident	Acc. Len	Accession
100%	0.0	99.93%	1392	WP_404518383.1
100%	0.0	99.93%	1392	WP_218422894.1
100%	0.0	100.00%	1392	WP_003690105.1

Figura 2. Resultados del análisis por protein BLAST de la secuencia de aminoácidos que se ha obtenido previamente con ORFfinder

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Neisseria gonorrhoeae strain 9035 chromosome, complete genome	Neisseria gonorrhoeae	7718	7718	100%	0.0	100.00%	2223133	CP104546.2
Neisseria gonorrhoeae isolate G97687 genome assembly, chromosome: I	Neisseria gonorrhoeae	7718	7718	100%	0.0	100.00%	2174841	LS999565.1
Neisseria gonorrhoeae strain 10538 chromosome, complete genome	Neisseria gonorrhoeae	7718	7718	100%	0.0	100.00%	2223795	CP104548.2
Neisseria gonorrhoeae strain AT159 chromosome, complete genome	Neisseria gonorrhoeae	7718	7718	100%	0.0	100.00%	2232771	CP097846.1
Neisseria gonorrhoeae strain TH2288 chromosome, complete genome	Neisseria gonorrhoeae	7718	7718	100%	0.0	100.00%	2227834	CP138339.1
Neisseria gonorrhoeae strain CT530 chromosome, complete genome	Neisseria gonorrhoeae	7718	7718	100%	0.0	100.00%	2169323	CP048254.1
Neisseria gonorrhoeae strain JS-23-24 chromosome 1, complete sequence	Neisseria gonorrhoeae	7718	7718	100%	0.0	100.00%	2231116	CP195874.1
Neisseria gonorrhoeae strain WHO K genome assembly, chromosome: 1	Neisseria gonorrhoeae	7718	7718	100%	0.0	100.00%	2169846	LT591908.1

Figura 3. Nucleotide BLAST donde se analiza la secuencia de nucleótidos obtenida del enunciado del ejercicio donde aparece la especie a la que pertenece *(Neisseria gonorrhoeae)* y distintos valores como su cobertura y porcentaje de identidad.

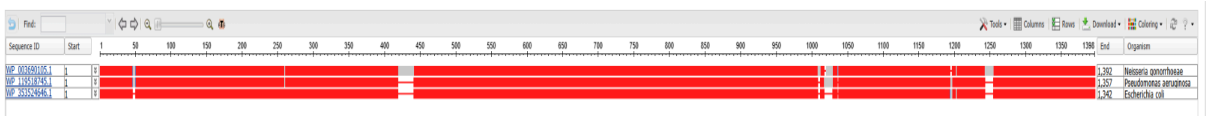

Figura 4. Selección de 3 organismos distintos para hacer un alineamiento. Las barras rojas horizontales indican zonas conservadas entre estos organismos, mientras que los espacios se deben a zonas variables o desconocidas

```
WP_003690105.1   1     MNYSFTEKKRIRKSFAKRENVLEVPFLLATQIDSYAKFLQLENAFDKRTDDGLQAAFNSIFPIVSHNGYARLEFVYYTLG  80
WP_119518745.1   1     MAYSYTEKKRIRKDFSKLPDVMDVPYLLAIQLDSYREFLQAGATKEQFRDIGLHAAFKSVFPIISYSGNAALEYVGYRLG  80
WP_353524646.1   1     MVYSYTEKKRIRKDFGKRPQVLDVPYLLSIQLDSFQKFIEQDPEGQ----YGLEAAFRSVFPIQSYSGNSELQYVSYRLG  76

WP_003690105.1   81    EPLFDIPECQLRGITYAAPLRARIRLVILDKEASKPTVKEVRENEVYMGEIPLMTPSGSFVINGTERVIVSQLHRSPGVF  160
WP_119518745.1   81    EPAFDVKECVLRGVTFAVPLRVKVRLIIFDRESSNKAIKDIKEQEVYMGEIPLMTENGTFIINGTERVIVSQLHRSPGVF  160
WP_353524646.1   77    EPVFDVQECQIRGVTYSAPLRVKLRLVIYEREAPEGTVKDIKEQEVYMGEIPLMTDNGTFVINGTERVIVSQLHRSPGVF  156

WP_003690105.1   161   FEHDKGKTHSSGKLLFSARIIPYRGSWLDFEFDPKDLLYFRIDRRRKMPVTILLKALGYNNEQILDIFYDKETFYLSSNG  240
WP_119518745.1   161   FDHDRGKTHSSGKLLYSARIIPYRGSWLDFEFDPKDCVFVRIDRRRKLPASVLLRALGYSTEEILNAFYATNVFHIKGET  240
WP_353524646.1   157   FDSDKGKTHSSGKVLYNARIIPYRGSWLDFEFDPKDNLFVRIDRRRKLPATIILRALNYTTEQILDLFFEKVIFEIRDNK  236

WP_003690105.1   241   VQTDLVAGRLKGETAKVDILDKEGNVLVAKGKRITAKNIRDITNAGLTRLDVEQESLLGKALAADLIDSETGEVLASAND  320
WP_119518745.1   241   LNLELVPQRLRGEVASIDIKDGSGKVIVEQGRRITARHINQLEKAGVSQLEVPFDYLIGRTIAKAIVHPATGEIIAECNT  320
WP_353524646.1   237   LQMELVPERLRGETASFDI-EANGKVYVEKGRRITARHIRQLEKDDVKLIEVPVEYIAGKVVAKDYIDESTGELICAANM  315

WP_003690105.1   321   EITEELLAKFDINGVKEITTLYINELDQGAYISNTLRTDETAGRQAARVAIYRMMRPGEPPTEEAVEQLFNRLFFSEDSY  400
WP_119518745.1   321   ELTLDLLAKVAKAQVVRIETLYTNDIDCGPFISDTLKIDNTSNQLEALVEIYRMMRPGEPPTKEAAETLFGNLFFSAERY  400
WP_353524646.1   316   ELSLDLLAKLSQSGHKRIETLFTNDLDHGPYISETLRVDPTNDRQSALVEIYRMMRPGEPPTREAAESLFENLFFSEDRY  395

WP_003690105.1   401   DLSRVGRMKFNTRTYEQKL[22]KGGYVLSVEDIVASIATLVELRNGHGEVDDIDHLGNRRVRSVGELTENQFRSGLARV  498
WP_119518745.1   401   DLSAVGRMKFNRRIGRTEI    EGPGVLSKEDIIDVLKTLVDIRNGKGIVDDIDHLGNRRVRCVGEMAENQFRVGLVRV  476
WP_353524646.1   396   DLSAVGRMKFNRSLLREEI    EGSGILSKDDIIDVMKKLIDIRNGKGEVDDIDHLGNRRIRSVGEMAENQFRVGLVRV  471

WP_003690105.1   499   ERAVKERLNQAESENLMPHDLINAKPVSAAIKEFFGSSQLSQFMDQTNPLSEVTHKRRVSALGPGGLTRERAGFEVRDVH  578
WP_119518745.1   477   ERAVKERLSMAESEGLMPQDLINAKPVAAAIKEFFGSSQLSQFMDQNNPLSEITHKRRVSALGPGGLTRERAGFEVRDVH  556
WP_353524646.1   472   ERAVKERLSLGDLDTLMPQDMINAKPISAAVKEFFGSSQLSQFMDQNNPLSEITHKRRISALGPGGLTRERAGFEVRDVH  551

WP_003690105.1   579   PTHYGRVCPIETPEGPNIGLINSLSVYARTNDYGFLETPYRRVIDGKVTEEIDYLSAIEEGRYVIAQANADLDSDGNLIG  658
WP_119518745.1   557   PTHYGRVCPIETPEGPNIGLINSLATYARTNKYGFLESPYRVVKDSLVTDEIVFLSAIEEADHVIAQASATLNEKGQLVD  636
WP_353524646.1   552   PTHYGRVCPIETPEGPNIGLINSLSVYAQTNEYGFLETPYRKVTDGVVTDEIHYLSAIEEGNYVIAQANSNLDEEGHFVE  631

WP_003690105.1   659   DLVTCREKGETIMATPDRVQYMDVATGQVVSVAASLIPFLEHDDANRALMGANMQRQAVPCLRPEKPMVGTGIERSVAVD  738
WP_119518745.1   637   ELVAVRHLNEFTVKAPEDVTLMDVSPKQVVSVAASLIPFLEHDDANRALMGSNMQRQAVPTLRADKPLVGTGMERNVARD  716
WP_353524646.1   632   DLVTCRSKGESSLFSRDQVDYMDVSTQQVVSVGASLIPFLEHDDANRALMGANMQRQAVPTLRADKPLVGTGMERAVAVD  711

WP_003690105.1   739   SATAIVARRGGVVEYVDANRVVIRVHDDEATAGEVGVDIYNLVKFTRSNQSTNINQRPAVKAGDVLQRGDLVADGASTDL  818
WP_119518745.1   717   SGVCVVARRGGVIDSVDASRVVVRVADDEVETGEAGVDIYNLTKYTRSNQNTCINQRPLVSKGDVVARGDILADGPSTDM  796
WP_353524646.1   712   SGVTAVAKRGGVVQYVDASRIVIKVNEDEMYPGEAGIDIYNLTKYTRSNQNTCINQMPCVSLGEPVERGDVLADGPSTDL  791

WP_003690105.1   819   GELALGQNMTIAFMPWNGYNYEDSILISEKVAADDRYTSIHIEELNVVARDTKLGAEDITRDIPNLSERMQNRLDESGIV  898
WP_119518745.1   797   GELALGQNMRVAFMPWNGFNFEDSICLSERVVQEDRFTTIHIQELTCVARDTKLGPEEITADIPNVGEAALNKLDEAGIV  876
WP_353524646.1   792   GELALGQNMRVAFMPWNGYNFEDSILVSERVVQEDRFTTIHIQELACVSRDTKLGPEEITADIPNVGEAALSKLDESGIV  871

WP_003690105.1   899   YIGAEVEAGDVLVGKVTPKGETQLTPEEKLLRAIFGEKASDVKDTSLRMPTGMSGTVIDVQVFTREGIQRDKRAQSIIDS  978
WP_119518745.1   877   YVGAEVQAGDILVGKVTPKGETQLTPEEKLLRAIFGEKASDVKDTSLRVPTGTKGTVIDVQVFTRDGVERDSRALSIEKM  956
WP_353524646.1   872   YIGAEVTGGDILVGKVTPKGETQLTPEEKLLRAIFGEKASDVKDSSLRVPNGVSGTVIDVQVFTRDGVEKDKRALEIEEM  951

WP_003690105.1   979   ELKRYRLDLNDQLRIFDNDAFDRIERMIVG[4]GGPMK---LAKGSEITTEYLAGLPsRHDWFDIRLTDEDLAKQLELIK  1056
WP_119518745.1   957   QLDQIRKDLNEEFRIVEGATFERLRAALVG   AKAEGgpaLKKGTEITDDYLDGLE-RGQWFKLRMADDALNEQLEKAQ  1032
WP_353524646.1   952   QLKQAKKDLSEELQILEAGLFSRIRAVLVA   GGVEA-----------EKLDKLP-RDRWLELGLTDEEKQNQLEQLA  1015

WP_003690105.1   1057  LSLQQKREEADELYEIKKKKLTQGDELQPGVQKMVKVFIAIKRRLQAGDKMAGRHGNKGVVSRILPVEDMPYMADGRPVD  1136
WP_119518745.1   1033  AYISDRRQLLDDKFEDKKRKLQQGDDLAPGVLKIVKVYLAIKRRIQPGDKMAGRHGNKGVVSVIMPVEDMPHDANGTPVD  1112
WP_353524646.1   1016  EQYDELKHEFEKKLEAKRRKITQGDDLAPGVLKIVKVYLAVKRRIQPGDKMAGRHGNKGVISKINPIEDMPYDENGTPVD  1095

WP_003690105.1   1137  IVLNPLGVPSRMNIGQILEVHLGWAAKGIGERIDRMLKERRKAGELREFLNKLYN---GSGKKEDLDSLTDEEIIELASN  1213
WP_119518745.1   1113  IVLNPLGVPSRMNVGQILETHLGLAAKGLGEKINRMLEEQRKVAELRKFLHEIYN-GIGGREE-NLDELGDNEILALAKN  1190
WP_353524646.1   1096  IVLNPLGVPSRMNIGQILETHLGMAAKGIGDKINAMLKQQQEVAKLREFIQRAYDlGADVRQKVDLSTFSDEEVMRLAEN  1175

WP_003690105.1   1214  LRKGASFASPVFDGAKESEIREML[12]KLGFNDSKTQITLYDGRSGEAFDRKVTVGVMHYLKLHHLVDEKMHARSTGPY  1301
WP_119518745.1   1191  LRGGVPMATPVFDGAKEREIKAML    KLADLPESGQMRLFDGRTGNQFERPTTVGYMYMLKLNHLVDDKMHARSTGSY  1266
WP_353524646.1   1176  LRKGMPIATPVFDGAKEAEIKELL    KLGDLPTSGQIRLYDGRTGEQFERPVTVGYMYMLKLNHLVDDKMHARSTGSY  1251

WP_003690105.1   1302  SLVTQQPLGGKAQFGGQRFGEMEVWALEAYGAAYTLQEMLTVKSDDVNGRTKMYENIVKGEHKIDAGMPESFNVLVKEIR  1381
WP_119518745.1   1267  SLVTQQPLGGKAQFGGQRFGEMEVWALEAYGAAYTLQEMLTVKSDDVNGRTKMYKNIVDGDHRMEAGMPESFNVLIKEIR  1346
WP_353524646.1   1252  SLVTQQPLGGKAQFGGQRFGEMEVWALEAYGAAYTLQEMLTVKSDDVNGRTKMYKNIVDGNHQMEPGMPESFNVLLKEIR  1331

WP_003690105.1   1382  SLGLDIDLERY  1392
WP_119518745.1   1347  SLGIDIELETE  1357
WP_353524646.1   1332  SLGINIELEDE  1342
```

Figura 5. Alineamiento de los 3 organismos seleccionados donde se representan los aminoácidos conservados (rojo), las zonas variables (azul) y las zonas desconocidas (gris).