



# Εργασία με θέμα το σεμινάριο Elixir που πραγματοποιήθηκε στη Λαμία στις 13-15/12/2019 για το μάθημα Υπολογιστική Ανάλυση Βιολογικών Αλληλουχιών

Βλασιάδης Λοΐζος Α.Μ: 00685

## Πίνακας περιεχομένων

Επιλογή Πρωτεΐνης.....	2
Στοίχιση δυο ακολουθιών - BLAST.....	3
Multiple Alignment T-coffee.....	7
CLUSTAL Omega.....	9
PSI-Blast.....	11
HMMER.....	13
Pfam.....	16
Πρόγνωση Τοπολογίας Διαμεμβρανικών Πρωτεϊνών.....	17
A-helix.....	18
β-Barrels.....	20

# Επιλογή Πρωτεΐνης

Για να επιλέξουμε την πρωτεΐνη με την οποία θα ασχοληθούμε θα μεταβούμε στη βάση δεδομένων [Uniprot](#). Η τελευταία τροποποίηση της βάσης πραγματοποιήθηκε στις 15 Οκτωβρίου 2019.

Θα ψάξουμε στη βάση Uniprot για διαμεμβρανικές πρωτεΐνες και πιο συγκεκριμένα για υποδοχείς GPCRs (G protein-coupled receptors).

The screenshot shows the UniProt search interface. On the left, there are filters for 'Filter by' (Reviewed (843), Unreviewed (344)), 'Popular organisms' (Human (1,187)), 'Proteomes' (UP000005640 (1,187)), and 'Search terms' (Filter "gpcr" as:). The main search area has a search bar with 'gpcr' and a dropdown menu for 'Subcellular location > Transmembrane'. The search results table shows the following data:

Q8NFJ5	RAI3_HUMAN	Retinoic acid-induced protein 3 (G-protein coupled receptor family C group 5 member A) (Phorbol ester induced gene 1, PEIG-1) (Retinoic acid-induced gene 1 protein, RAIG-1)	GPCR5A GPCR5A, RAI3, RAIG1	Homo sapiens (Human)	357
Q9H3N8	HRH4_HUMAN	Histamine H4 receptor, H4R, HH4R (AXOR35) (G-protein coupled receptor 105) (GPRV53) (Pli-013) (SP9144)	HRH4 GPCR105	Homo sapiens (Human)	390
Q9Y5N1	HRH3_HUMAN	Histamine H3 receptor, H3R, HH3R (G-protein coupled receptor 97)	HRH3 GPCR97	Homo sapiens (Human)	445

Εικόνα 1

Στην παραπάνω εικόνα, στιγμιότυπο (screenshot), φαίνεται το ερώτημα και οι παράμετροι που θέσαμε στη βάση δεδομένων. Πιο συγκεκριμένα, στο πεδίο του οργανισμού (Organism OS) επιλέξαμε για είδος τον Homo Sapiens, δηλ τον άνθρωπο μιας και το πρωτέωμά του είναι από τα καλύτερα μελετημένα. Συνεπώς στα πεδία αναζήτησης περάσαμε και το πρωτέωμα του Homo Sapiens. Στη συνέχεια επιλέξαμε στο πεδίο Subcellular location την επιλογή Transmembrane, επειδή επιθυμούμε να βρούμε διαμεμβρανικά τμήματα. Τέλος, επιλέξαμε τον όρο GPCR επειδή θέλουμε να ασχοληθούμε με receptors.

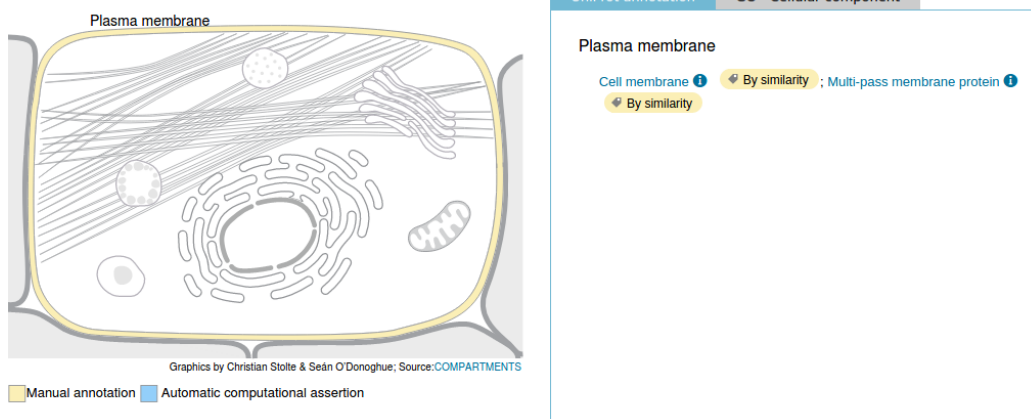
**Protein** | **G-protein coupled receptor family C group 6 member A**  
**Gene** | **GPCR6A**  
**Organism** | *Homo sapiens (Human)*  
**Status** | **Reviewed** - Annotation score: - Experimental evidence at protein level<sup>1</sup>

Εικόνα 2

Από τα 1187 αποτελέσματα που μας επιστρέφει, θα επιλέξουμε να ασχοληθούμε με το γονίδιο **GPCR6A** με κωδικό entry στην Uniprot [Q5T6X5](#) και μέγεθος 926. Το συγκεκριμένο γονίδιο είναι καλά μελετημένο και μάλιστα αυτό που λέμε ότι έχει καταχωρηθεί “από χέρι”. Δηλαδή όπως

παρατηρούμε στην εικόνα 2, προέρχεται από τη Swiss-Prot, κομμάτι της Uniprot, για αυτό και είναι με χρυσό. Επίσης το annotation score, ένας δείκτης της ίδιας της Uniprot, το βαθμολογεί με 5 κύκλους, το καλύτερο σκορ, που σημαίνει ότι η πρωτεΐνη υπάρχει και έχει αποδειχθεί πειραματικά η ύπαρξή της.

### Subcellular location<sup>i</sup>



Εικόνα 3

Στις εικόνες 3 και 4 βλέπουμε στο πεδίο Subcellular Location πληροφορίες για την διαμεμβρανική πρωτεΐνη καθώς και τα σημεία, επτά (7), καθώς και μέγεθος του καθενός από αυτά τα σημεία, στα οποία διαπερνάει την κυτταρική μεμβράνη.

Topology					
Feature key	Position(s)	Description	Actions	Graphical view	Length
Topological domain <sup>i</sup>	19 – 594	Extracellular <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		576
Transmembrane <sup>i</sup>	595 – 615	Helical; Name=1 <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		21
Topological domain <sup>i</sup>	616 – 631	Cytoplasmic <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		16
Transmembrane <sup>i</sup>	632 – 652	Helical; Name=2 <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		21
Topological domain <sup>i</sup>	653 – 669	Extracellular <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		17
Transmembrane <sup>i</sup>	670 – 690	Helical; Name=3 <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		21
Topological domain <sup>i</sup>	691 – 704	Cytoplasmic <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		14
Transmembrane <sup>i</sup>	705 – 725	Helical; Name=4 <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		21
Topological domain <sup>i</sup>	726 – 748	Extracellular <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		23
Transmembrane <sup>i</sup>	749 – 769	Helical; Name=5 <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		21
Topological domain <sup>i</sup>	770 – 782	Cytoplasmic <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		13
Transmembrane <sup>i</sup>	783 – 803	Helical; Name=6 <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		21
Topological domain <sup>i</sup>	804 – 810	Extracellular <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		7
Transmembrane <sup>i</sup>	811 – 831	Helical; Name=7 <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		21
Topological domain <sup>i</sup>	832 – 926	Cytoplasmic <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		95

Εικόνα 4

## Στοίχιση δυο ακολουθιών - BLAST

Το BLAST είναι ένα εργαλείο για να κάνουμε τοπικές στοιχίσεις ψάχνοντας σε διάφορες βάσεις δεδομένων.

Χρησιμοποιεί την τεχνική των λέξεων (words), επιλέγει μια λέξη “γραμμάτων” αμινοξέων ή νουκλεοτιδίων ότι μεγέθους θέλουμε, συνήθως όμως 3 . Με βάση αυτή τη λέξη και με ένα κατώφλι (threshold) που μπορούμε να το ορίσουμε και αυτό εμείς, βρίσκει λέξεις που είναι πάνω από το

κατώφλι. Στη συνέχεια με βάση τη λέξη που θα βρει θα πραγματοποιήσει αναζήτηση δεξιά και αριστερά κατά μήκος της ακολουθίας.

Εμείς στην αρχική σελίδα του [BLAST](#), επιλέξαμε να κάνουμε αναζήτηση για πρωτεΐνες μόνο.

Αφού εισάγουμε σε FASTA format την ακολουθία μας στη συνέχεια επιλέγουμε σε ποιες βάσεις θέλουμε να γίνει η αναζήτηση. Αν δεν το ξεκαθαρίσουμε επιλέγοντας μια βάση συγκεκριμένα, θα αποτελέσματα που θα μας επιστρέψει θα βασίζονται σε μια τυχαία επιλογή. Μπορεί να είναι είτε της PDB είτε της TREMBL είτε οποιασδήποτε άλλης βάσης από τις προτεινόμενες.

Εμείς θα επιλέξουμε την PDB επειδή είναι πιο μικρή και επομένως θα μας φέρει πιο γρήγορα αποτελέσματα.

Δεν αλλάξαμε κάτι από τις παραμέτρους του προγράμματος και θα δούμε παρακάτω ποιες είναι αυτές ακριβώς.

Αφού το τρέξουμε, τα αποτελέσματα που μας επιστρέφει είναι ένας πίνακας της παρακάτω μορφής, όπως βλέπουμε και στην εικόνα 5.

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) [Query subrange](#)

From

To

Or, upload file  Δεν επιλέχθηκε αρχείο. [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Choose Search Set

Database  [?](#)

Organism [Optional](#)  ☐ exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Exclude [Optional](#) ☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WP) ☐ Uncultured/environmental sample sequence

Program Selection

Algorithm ☒ blastp (protein-protein BLAST)

Εικόνα 5

Για κάθε πρωτεΐνη που κατάφερε να στοιχίσει, μας δίνει το score της στοίχισης, το ποσοστό καταλοίπων, το ποσοστό της πρωτεΐνης που κάλυψε, το E-value καθώς και το accession της βάσης, εδώ της PDB.

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
✓	<a href="#">Chain A, Metabotropic glutamate receptor 5, Endolysin [Homo sapiens]</a>	62.4	120	26%	1e-09	32.19%	<a href="#">6FFH_A</a>
✓	<a href="#">Chain A, Metabotropic glutamate receptor 5, Lysozyme, Metabotropic glutamate receptor 5 chimera [Homo sapiens]</a>	62.4	120	26%	1e-09	32.19%	<a href="#">4OO9_A</a>
✓	<a href="#">Chain A, Metabotropic glutamate receptor 5, Endolysin, Metabotropic glutamate receptor 5 [Homo sapiens]</a>	62.4	119	26%	2e-09	32.19%	<a href="#">5CGC_A</a>
✓	<a href="#">Chain A, Soluble cytochrome b562, Metabotropic glutamate receptor 1 [Homo sapiens]</a>	112	112	28%	3e-26	31.75%	<a href="#">4OR2_A</a>
✓	<a href="#">Chain A, Extracellular Calcium-sensing Receptor [Homo sapiens]</a>	229	229	48%	9e-65	30.88%	<a href="#">5FBH_A</a>
✓	<a href="#">Chain B, Taste Receptor, Type 1, Member 3 [Oryzias latipes]</a>	208	208	48%	4e-58	30.35%	<a href="#">5X2M_B</a>
✓	<a href="#">Chain A, Metabotropic glutamate receptor 3 [Homo sapiens]</a>	157	157	34%	1e-40	30.06%	<a href="#">3SM9_A</a>
✓	<a href="#">Chain A, Metabotropic glutamate receptor 3 [Homo sapiens]</a>	158	158	34%	2e-40	30.06%	<a href="#">4XAR_A</a>
✓	<a href="#">Chain A, Extracellular calcium-sensing receptor [Homo sapiens]</a>	275	275	62%	7e-81	29.62%	<a href="#">5K5T_A</a>
✓	<a href="#">Chain A, Extracellular calcium-sensing receptor [Homo sapiens]</a>	274	274	62%	1e-80	29.62%	<a href="#">5K5S_A</a>
✓	<a href="#">Chain A, Metabotropic glutamate receptor 2 [Homo sapiens]</a>	150	150	35%	5e-38	28.36%	<a href="#">4XAQ_A</a>
✓	<a href="#">Chain A, Metabotropic glutamate receptor 2 [Homo sapiens]</a>	150	150	35%	7e-38	28.36%	<a href="#">5CNJ_A</a>
✓	<a href="#">Chain A, Extracellular Ligand-binding Receptor [Desulfitobacterium hafniense DCB-2]</a>	32.3	32.3	10%	2.7	28.00%	<a href="#">4MLC_A</a>
✓	<a href="#">Chain A, Metabotropic glutamate receptor 5 [Homo sapiens]</a>	272	272	89%	1e-77	27.95%	<a href="#">6NS2_A</a>
✓	<a href="#">Chain A, Metabotropic glutamate receptor 5 [Homo sapiens]</a>	272	272	89%	1e-77	27.95%	<a href="#">6N4X_A</a>
✓	<a href="#">Chain A, Metabotropic glutamate receptor 5 [Homo sapiens]</a>	259	259	87%	1e-73	27.68%	<a href="#">6NS1_A</a>
✓	<a href="#">Chain A, Gamma-aminobutyric Acid Type B Receptor Subunit 1 [Homo sapiens]</a>	45.4	45.4	13%	3e-04	27.64%	<a href="#">4MQE_A</a>
✓	<a href="#">Chain A, Gamma-aminobutyric Acid Type B Receptor Subunit 2 [Homo sapiens]</a>	49.7	49.7	14%	1e-05	26.81%	<a href="#">4F11_A</a>
✓	<a href="#">Chain A, Actin-binding LIM protein 3 [Homo sapiens]</a>	29.6	29.6	4%	4.0	26.67%	<a href="#">2DJ7_A</a>

Εικόνα 6

Πριν πάμε να δούμε τα αποτελέσματα του BLAST, να παραθέσουμε μια εικόνα με τις default παραμέτρους που έγινε η στοίχιση (εικόνα 7).

Search Parameters	
Program	blastp
Word size	6
Expect value	10
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	21
Composition-based stats	2

Εικόνα 7

Παρατηρούμε ότι το εργαλείο που χρησιμοποιήσαμε είναι το blastp, το μέγεθος της λέξης ήταν 6 και το κατώφλι 21, η ποινή για τα κενά 11 και 1 για κάθε επόμενο ενώ χρησιμοποιήθηκε ο πίνακας BLOSUM62.

Επιλέγοντας τώρα το πρώτο αποτέλεσμα, βλέπουμε πληροφορίες για τη στοίχιση. Η ακολουθία μας αντιστοιχεί στο Query ενώ στο Subject η ακολουθία που κατάφερε να στοιχιστεί από την PDB. Όπως παρατηρούμε, η ακολουθία ξεκίνησε να στοιχίζεται από το κατάλοιπο 10 κάτι που σημαίνει για τα προηγούμενα γονίδια δεν έχουμε συγκεκριμένη προσδιορισμένη δομή. Ίσως να αφορά το εξωκυττάριο μέρος όπου προσδένεται ο υποδοχέας. Ενώ η στοίχιση φτάνει μέχρι το κατάλοιπο 588.

Score	Expect	Method	Identities	Positives	Gaps
275 bits(702)	7e-81	Compositional matrix adjust.	181/611(30%)	306/611(50%)	42/611(6%)
Query 10	CFVILAT-SQPCQTPDDFVAATSPGHIIIGGLFAIHEKMLS-SEDSRRPQIQECVGFE	67			
Sbjct 7	C+V++ T PD A G II+GGLF IH + + +D RP+ EC+ +	64			
Query 68	ISVFLQTLAMIHSIEMINNS-TLLPGVKLGYEIYDTCTEVTVAMAATLRFLSKFNCSRET	126			
Sbjct 65	F AMI +IE IN+S LLP + LGY I+DTC V+ A+ ATL F+++	124			
Query 127	VEFKCDYSSYMPRVKAVIGSGYSEITMAVSRMLNLQMPQVGYESTAEILSDKIRFPSFL	186			
Sbjct 125	++ C+ S ++P AV+G+ S ++ AV+ +L L +POV Y S++ +LS+K +F SFL	184			
Query 187	LDEFNCNSEHIPSTIAVVGATGSGVSTAVANLLGLFYIPQVSYASSSRLLSNKNQKFSFL	246			
Sbjct 185	RTVPSDFHQIKAMAHLIQSGWNWIGIITTTDDDYGRALNTFIIQAEANNVCIAFKEVLP	244			
Query 247	RT+P+D HQ AMA +I+ WNW+G I DDDYGR + F +AE ++CI F E++	305			
Sbjct 245	RTIPNDEHQATAMADIIEYFRWNWVGITIAADDDYGRPGIEKFREEAEERDIDFSELIS	299			
Query 306	AFLSDNTIEVRINRTLKKIILEAQVNVIVVFLRQFHVFDLFNKAITEMNI-NKMWIASDNW	360			
Sbjct 300	+ + I+ + ++I + VIVVF + L + + NI K+W+AS+ W	359			
Query 361	QYSDEEEIQHVV-----EVIQNSTAKVIVVFSSGPDLEPLIKEIVRRNITGKIWLASEAW	412			
Sbjct 360	STATKITTIPNVKKIGKVVGFARRGNISFHSFLQNLHLLPSDSHKLLHE-----YAMH	417			
Query 413	++++ I +G +GFA + G I F FL+ +H S + E + H	462			
Sbjct 418	ASSSLIAMPQYFHVVGTTIGFALKAGQIPGFREFLKKVHPRKSVHNGFAKEFWETFNCH	477			
Query 463	LSACA-----YVKDTLSQCIFNHSQRTLAYKANKAIERNFVMRNDFLWDYAEPGLI	517			
Sbjct 478	L A +++ + S F SQ + A++ + N DY +	537			
Query 518	LQEGAKGPLPVDTFLRGHEESGDRF--SQSSTAFRPLCTGDENINSVETPYIDYTHLRIS	577			
Sbjct 538	HSIQLAVFALGYAIRD-----LCQARDCQNPNAFQPWELLGVLKNVTFTDGNWS-F	596			
Query 578	+++ LAV+++ +A++D L C + + W++L L+++ FT+				
Sbjct 597	YNVYLAVYSIAHALQDIYTCLPGRGLFTNGSCADIKKVEAWQVLKHLRHLNFTNNMGEQV				
Query 588	HFDAHGDLNTGYDVVLW--KEINGHMTVTKMAEYDL---QNDVFIIPDQETKNEFRNLKQ				
Sbjct 607	FD GDL Y ++ W +G + ++ Y++ + + I +++ + +				
Query 588	TFDECGDLVGNYSIINWHLSPEDGSIVFKEVGYNVYAKKGERLFINEEKILWSGFSREV				
Query 577	IQSKCSKECSPGQMKKTTRSQHICCYEQNCPENHYTNQTDMPHCLLCNNKTHWAPVRST				
Sbjct 596	S CS++C G K + CC+EC CP+ Y+++TD C C + W+ T				
Query 588	PFSNCSRDCLAGTRKGIIEGEPTCCFECVECPDGEYSDETDASACNKCPDD-FWSNENHT				
Query 588	MCFEKEVEYLN 588				
Sbjct 597	C KE+E+L+ 607				
Query 588	SCIAKEIEFLS 607				

Εικόνα 8

Έχουμε ένα καλό E-value= 7e-81 που μας δείχνει τη στατιστική σημαντικότητα. Το E-value είναι ο αριθμός στοιχίσεων που περιμένουμε να έχουν score ίσο ή μεγαλύτερο με το score που βλέπουμε και να έχει προκύψει αυτό το score τυχαία. Στον καθορισμό του παίζουν ρόλο τα εξής: το μέγεθος της βάσης, το μέγεθος της ακολουθίας που ψάχνουμε και το σύστημα βαθμονόμησης.

Οι τιμές που μπορεί να πάρει κυμαίνονται από το 0 (αυτό θέλουμε ιδανικά δλδ δεν είναι τυχαίο) μέχρι +άπειρο.

Στο συγκεκριμένο παράδειγμα πρέπει να αναλογιστούμε ότι η PDB είναι μια μικρή βάση δεδομένων.



## Multiple Alignment T-coffee

Για να προχωρήσουμε στην πολλαπλή στοίχιση ακολουθιών, μεταβαίνουμε στην Uniprot ξανά και αφού επαναλάβουμε το ερώτημα της εικόνας 1 (δλδ στον οργανισμό Homo Sapiens, πρωτέωμα του Homo Sapiens, στην Subcellular location για transmembrane και γενικά για GPCR) επιλέγουμε 7 ακολουθίες. Τα assertion numbers τους είναι τα εξής: [Q5T6X5](#), [Q14833](#), [O15303](#), [Q14416](#), [Q14831](#), [Q02643](#) και [Q14832](#). Τις “κατεβάζουμε” και τις αποθηκεύουμε σε αρχείο σε FASTA format. Στη συνέχεια , μεταβαίνουμε στο site του EBI (Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής) και επιλέγουμε το εργαλείο [T-Coffee](#).

Το συγκεκριμένο εργαλείο, μας παρέχει τη δυνατότητα να κάνουμε πολλαπλές στοιχίσεις. Γενικά το T-Coffee δεν μας δίνει την δυνατότητα για πολλές παραμετροποιήσεις. Ανάλογα με το τι δεδομένα επιλέγουμε να εισάγουμε, πρωτεΐνες, DNA ή RNA, κάνει τις ανάλογες ρυθμίσεις για να πραγματοποιηθεί η ανά δύο στοιχίση των επιλεγμένων ακολουθιών μας.

Αφού εισάγουμε τα δεδομένα μας σε fasta format επιλέγουμε να τα στοιχίσουμε χωρίς να αλλάξουμε κάτι στις default επιλογές. Τέλος επιλέγουμε τα αποτελέσματα να μας παρουσιαστούν σε CLUSTALW μορφή όπως φαίνεται και στην εικόνα 9.

CLUSTAL W (1.83) multiple sequence alignment

```

sp|Q15303|GRM6_HUMAN    MARPPRA-----REPLLV--ALLPLAWLAQAG--LARAAGSVRLAGG
sp|Q02643|GHRHR_HUMAN   MDRRMWG-----AHVFCVLSPL-----P
sp|Q14416|GRM2_HUMAN    MGSLLAL-----LALLL-L-----WGAVAE--GPAKKVLTLEGD
sp|Q14831|GRM7_HUMAN    MVQLRKLRLVLTLMKFPCCVLEVLLCALAAAARGQ--EYAPHSIIRIEGD
sp|Q14832|GRM3_HUMAN    MKMLTRL-----QVLTL--ALFSKGFLSLGD--HNFRLRRIKIEGD
sp|Q14833|GRM4_HUMAN    MPGKRGGLG--WwWwARLPLCLLLSLYGPMWPSGLGPKGHPHMNSIRIDGD
sp|Q5T6X5|GPC6A_HUMAN   MAFLIIL-----ITCFV--IILA--TSQPCQ--TPDDFVAATSPGH

```

```

sp|Q15303|GRM6_HUMAN      LTLGGLFPVHARGAAGRACGQ---LKKEQG----VHRL EAMLYALDRVN
sp|Q02643|GHRHR_HUMAN     TVLGHMHP-----ECDF---ITQLR-----EDESACLQAAEEMP
sp|Q14416|GRM2_HUMAN       LVLGGLFPVHKGGAEDCGP---VNEHRG-----IQRL EAMLFALDRIN
sp|Q14831|GRM7_HUMAN       VTLGGLFPVHAKGSGVPCGD---IKRENG-----IHRLEAMLYALDQIN
sp|Q14832|GRM3_HUMAN       LVLGGLFPVINEKGTGTEECGR---INEDRG-----IQRL EAMLYAIDEIN
sp|Q14833|GRM4_HUMAN       ITLGGLFPVHGRGSEKPCGE---LKKEKG-----IHRLEAMLFALDRIN
sp|Q5T6X5|GPC6A_HUMAN     IIIIGLFAIHEKMLSSIEDSPRRPQIQECVGF EISVFLQTLAMHSIEMIN

```

sp|015303|GRM6\_HUMAN ADPELLPGVRLGARLLDTCSCRDYALEQALSFVQALIRGRGDGDEVGVRC  
sp|Q02643|GHRHR\_HUMAN NTTLGCPATWDGLLCWPTAGSGEWTLPDPFFSHFSE---SGAVKRDG  
sp|Q14416|GRM2\_HUMAN RDPHLLPGVRLGAHLDTSCSDTHALEQALDFVVRASLRG--ADGSRHC  
sp|Q14831|GRM7\_HUMAN SDPNLLPNVTLGARILDTCSDRDYALEQSLTFVQALIK---DTSDVRC  
sp|Q14832|GRM3\_HUMAN KDDYLLPGVKLVGHLLDTCSDRYALEQSLFVRASLTK---VDEAEYMC  
sp|Q14833|GRM4\_HUMAN NDPDLLPNITLGARILDTCSDTHALEQSLTFVQALIEK---DGTEVRC  
sp|Q5T6X5|GPC6A\_HUMAN NS-TLLPGVKLVGYEYDTCETVTVMAAATLRLFSKFNC---RETVEFKC

```

sp|015303|GRM6_HUMAN      P--GGV--PPLR-----PAPPERVVAVVGASASVSMV
sp|Q02643|GHRHR_HUMAN    TITGWSEFPFPYPVACPVPLELLAEESYFSTVKIITYVGHISISVALFV
sp|Q14416|GRM2_HUMAN      P--DGS--YATH-----GDAPTITGVIGGSYSVDSIQV
sp|Q14831|GRM7_HUMAN      T--NGE--PPVF-----VKPEKVVGVIGASGGSVSIMV
sp|Q14832|GRM3_HUMAN      P--DGS--YAIQ-----ENIPLLTAGVIGGSYSVDSIQV
sp|Q14833|GRM4_HUMAN      G--SGG--PPII-----TKPERVVGVIGASGGSVSIMV
sp|Q5T6X5|GPC6A_HUMAN    D--YSS--Y-----MPRVKAVIGGSYIEITMAV

```

```

sp|015303|GRM6_HUMAN  AN---VLRIFAIPQISYASTAPELSDSTRDYFFSRVVPVDSYQAMVD
sp|Q02643|GHRHR_HUMAN AITILVALRRLHCPR-----
sp|Q14416|GRM2_HUMAN  AN---LLRLFQIPQISYASTASAKLSDSKSRDYDFARTVPPDFFQAKAMAE
sp|Q14831|GRM7_HUMAN  AN---ILRLFQIPQISYASTAPELSDDRDYFFSRVVPVDSFQAMVD
sp|Q14832|GRM3_HUMAN  AN---LLRLFQIPQISYASTASAKLSDSKSRDYDFARTVPPDFFQAKAMAE
sp|Q14833|GRM4_HUMAN  AN---ILRLFQIPQISYASTAPDLSNDRDYFFSRVVPVSDTYQAMVD
sp|Q5T6X5|GPC6A_HUMAN SRM---LNLQLMPQVGVESTAEILSDKIRFPFLRTVPVSDFHQIKAMAH

```

sp|Q15303|GRM6\_HUMAN IVRALGWNYVSTLASEGNYGESGV EAFVQISR EAGGVCTAQSIKIPREP K  
sp|Q02643|GHRHR\_HUMAN ----- NYVHT ----- QLF TTFILKAGAVFLKDAALFHSDD T  
sp|Q14416|GRM2\_HUMAN ILRFFNW TYVSTVASEG DYGETG IEAFEEAR -ARNICVATSEKVG RAMS  
sp|Q14831|GRM7\_HUMAN IKVALGWN YVSTLASEGSYGETGVESFTQISK EAGGLCIAQSVRI PQRK  
sp|Q14832|GRM3\_HUMAN ILRFFNW TYVSTVASEG DYGETG IEAFEEAR -LRNICATAEAKVGSNI  
sp|Q14833|GRM4\_HUMAN IVRALKWN YVSTVASEGSYGESGV EAFIQKSR EDGGVCTAQSVKIPREP K  
sp|Q5T6X5|GPC6A\_HUMAN LTIKSGWN WIGITTD DGYRLALNTFI IQAE -ANNVCTAFKEVLP AFLS

Λόγω μεγέθους δεν βάλουμε ολόκληρη τη στοίχιση των ακολουθιών. Ολόκληρη θα την παραθέσουμε στο τέλος της εργασίας.

Επιλέξαμε να μας εμφανιστούν με χρώμα ώστε να είναι πιο ευπαρουσίαστα στο ανθρώπινο μάτι. Για την περαιτέρω βοήθεια μας, το T-Coffee κάτω από τις στοιχίσεις εμφανίζει κάποια σύμβολα. Τα σύμβολα αυτά είναι τα \* (αστερίσκος), : (άνω κάτω τελεία), . (τελεία) καθώς και το κενό ( ).

- Ο αστερίσκος \* αντιστοιχεί σε θέσεις όπου υπάρχει απόλυτη ταύτιση στη στοίχιση.
- Η άνω κάτω τελεία : μας δείχνει τις θέσεις όπου υπάρχει μια σχετική διατήρηση και συνήθως είναι και ίδιου χρώματος.
- Η τελεία . δείχνει τις θέσεις όπου υπάρχει μια χαλαρή διατήρηση μεταξύ των στοιχίσεων. Είναι μια επιτρεπτή αντικατάσταση αλλά όχι τόσο κοντινές.
- Τέλος το κενό ( ) φανερώνει τις θέσεις όπου δεν υπάρχει καμία σύνδεση.

Προφανώς και τα χρώματα δεν είναι τυχαία αλλά βασίζονται στον παρακάτω πίνακα.

Residue	Colour	Property
AVFPMILW	RED	Small (small+ hydrophobic (incl.aromatic -Y))
DE	BLUE	Acidic
RK	MAGENTA	Basic - H
STYHCNGQ	GREEN	Hydroxyl + sulfhydryl + amine + G
Others	Grey	Unusual amino/imino acids etc

Εικόνα 10

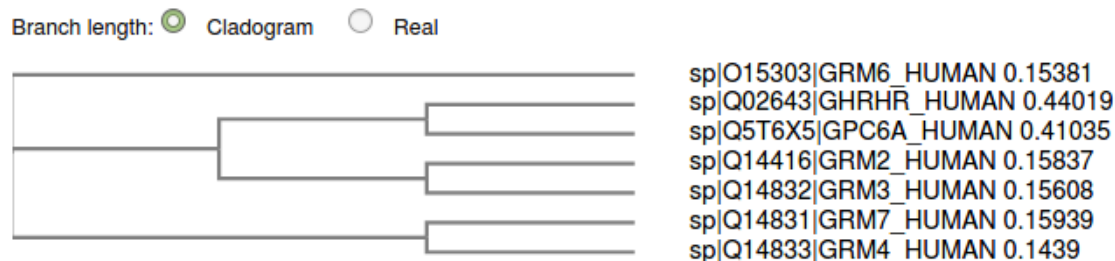
Χρειαζόμαστε αστεράκια (\*) στα αποτελέσματα μας αλλά θέλουμε να υπάρχει διαφοροποίηση αλλιώς δεν υπάρχει ενδιαφέρον και ενδεχομένως να έχουμε πάρει την ίδια ακολουθία αν εμφανίζονται μόνο αστερίσκοι.

Το T-coffee μας δίνει την δυνατότητα να έχουμε και το φυλογενετικό δέντρο των ακολουθιών που επιλέξαμε για πολλαπλή στοίχιση όπως φαίνεται και στην παρακάτω εικόνα 11.



# Phylogenetic Tree

*This is a Neighbour-joining tree without distance corrections.*



Εικόνα 11

Όπως και τον Percent Identity Matrix που μας δείχνει πόσο τοις εκατό (%) μοιάζουν μεταξύ τους οι ακολουθίες.

```
# Percent Identity Matrix - created by Clustal2.1
#
#
```

1:	sp O15303 GRM6_HUMAN	100.00	23.31	46.65	67.13	46.56	69.85	26.34
2:	sp Q02643 GHRHR_HUMAN	23.31	100.00	21.33	24.08	23.47	22.57	14.95
3:	sp Q14416 GRM2_HUMAN	46.65	21.33	100.00	45.12	68.55	46.86	27.52
4:	sp Q14831 GRM7_HUMAN	67.13	24.08	45.12	100.00	44.65	69.67	25.26
5:	sp Q14832 GRM3_HUMAN	46.56	23.47	68.55	44.65	100.00	46.14	27.80
6:	sp Q14833 GRM4_HUMAN	69.85	22.57	46.86	69.67	46.14	100.00	25.03
7:	sp Q5T6X5 GPC6A_HUMAN	26.34	14.95	27.52	25.26	27.80	25.03	100.00

Εικόνα 12

## CLUSTAL Omega

Πέρα από το T-coffee, το EBI μας παρέχει και άλλα εργαλεία, όπως πχ το CLUSTAL Omega. Το συγκεκριμένο εργαλείο είναι η βελτιωμένη έκδοση του ClustalW. Το CLUSTAL Omega χρησιμοποιεί HMM profile και profile τεχνικές γενικότερα για να μας δώσει καλύτερα αποτελέσματα για στοιχίσεις τριών (3) ακολουθιών και πάνω.

Επίσης εδώ μπορούμε να έχουμε περισσότερες επιλογές στην παραμετροποίηση.

Στην εμφάνιση των αποτελεσμάτων μας παρέχει την δυνατότητα για χρωματισμό των στοιχίσεων όπως επίσης και το σχολιασμό με τα ίδια σύμβολα, αστερίσκο \*, τελεία ., άνω κάτω τελεία : καθώς και το κενό ( ).

```

CLUSTAL O(1.2.4) multiple sequence alignment

sp|Q02643|GHRHR_HUMAN      ----- 0
sp|Q5T6X5|GPC6A_HUMAN      ----- 44
sp|Q15303|GRM6_HUMAN      ----- 47
sp|Q14833|GRM4_HUMAN      ----- 57
sp|Q14831|GRM7_HUMAN      ----- 57
sp|Q14416|GRM2_HUMAN      ----- 40
sp|Q14832|GRM3_HUMAN      ----- 47

sp|Q02643|GHRHR_HUMAN      ----- 0
sp|Q5T6X5|GPC6A_HUMAN      ----- 103
sp|Q15303|GRM6_HUMAN      ----- 99
sp|Q14833|GRM4_HUMAN      ----- 109
sp|Q14831|GRM7_HUMAN      ----- 109
sp|Q14416|GRM2_HUMAN      ----- 92
sp|Q14832|GRM3_HUMAN      ----- 99

sp|Q02643|GHRHR_HUMAN      ----- 0
sp|Q5T6X5|GPC6A_HUMAN      ----- 154
sp|Q15303|GRM6_HUMAN      ----- 159
sp|Q14833|GRM4_HUMAN      ----- 164
sp|Q14831|GRM7_HUMAN      ----- 164
sp|Q14416|GRM2_HUMAN      ----- 150
sp|Q14832|GRM3_HUMAN      ----- 156

sp|Q02643|GHRHR_HUMAN      ----- 0
sp|Q5T6X5|GPC6A_HUMAN      ----- 214
sp|Q15303|GRM6_HUMAN      ----- 219
sp|Q14833|GRM4_HUMAN      ----- 224
sp|Q14831|GRM7_HUMAN      ----- 224
sp|Q14416|GRM2_HUMAN      ----- 210
sp|Q14832|GRM3_HUMAN      ----- 216

sp|Q02643|GHRHR_HUMAN      ----- 0
sp|Q5T6X5|GPC6A_HUMAN      ----- 273
sp|Q15303|GRM6_HUMAN      ----- 275
sp|Q14833|GRM4_HUMAN      ----- 280
sp|Q14831|GRM7_HUMAN      ----- 282
sp|Q14416|GRM2_HUMAN      ----- 265
sp|Q14832|GRM3_HUMAN      ----- 271

```

Εικόνα 13

Αυτό που παρατηρούμε είναι ενώ το T-coffee από την αρχή των ακολουθιών προβαίνει σε στοιχίσεις μεταξύ τους, αντίθετα το Clustal Omega στην ίδια περιοχή μας εμφανίζει ότι δεν υπάρχει συσχέτιση μεταξύ των αλληλουχιών.

Παρ' όλες τις διαφορές που εμφανίζουν στη στοιχίση, τα φυλογενετικά δέντρα είναι ίδια.

# Phylogenetic Tree

*This is a Neighbour-joining tree without distance corrections.*

Branch length: ☒ Cladogram ☐ Real



Εικόνα 14

Ενώ αν δούμε και το Identity Percent Matrix παρατηρούμε και εκεί διαφορές όπως ήταν αναμενόμενο.

```
# Percent Identity Matrix - created by Clustal2.1  
#  
#
```

1:	sp Q02643 GHRHR_HUMAN	100.00	14.08	15.74	18.62	16.82	18.27	18.81
2:	sp Q5T6X5 GPC6A_HUMAN	14.08	100.00	25.58	25.38	25.35	27.38	27.95
3:	sp O15303 GRM6_HUMAN	15.74	25.58	100.00	69.51	67.28	46.64	46.19
4:	sp Q14833 GRM4_HUMAN	18.62	25.38	69.51	100.00	69.67	46.67	45.83
5:	sp Q14831 GRM7_HUMAN	16.82	25.35	67.28	69.67	100.00	44.86	44.51
6:	sp Q14416 GRM2_HUMAN	18.27	27.38	46.64	46.67	44.86	100.00	68.21
7:	sp Q14832 GRM3_HUMAN	18.81	27.95	46.19	45.83	44.51	68.21	100.00

Εικόνα 15

## PSI-Blast

Μέθοδος εύρεσης απομακρυσμένων ομοιοτήτων πιο ευαίσθητη και κυρίως για τοπική στοίχιση.

Το PSI-Blast είναι ένα εργαλείο που παίρνει την ακολουθία μας και κάνει αναζήτηση στη βάση δεδομένων που έχουμε επιλέξει. Μπορούμε να ορίσουμε εμείς την επιθυμητή τιμή του E-value που θέλουμε. Αφού βρίσκει τα διάφορα αποτελέσματα, τα οποία μπορεί να μην είναι και τα καλύτερα, προχωράει σε πολλαπλή στοίχιση μεταξύ τους, όπου μπορούν να στοιχιστούν, και με βάση αυτά φτιάχνει ένα πίνακα υποκατάστασης (PSSM). Στη συνέχεια χρησιμοποιεί αυτό τον πίνακα και κάνει αναζήτηση έναντι της βάσης. Άρα έχουμε ένα προσαρμοσμένο πινάκα αντικατάστασης των αμινοξέων με βάση τις ομοιότητες που έχει ήδη βρει και επαναλαμβάνει την διαδικασία. Στο τέλος όλες οι πρωτεΐνες έχουν μια ομοιότητα με την πρωτεΐνη μας, έστω και απομακρυσμένη.

# PSI-BLAST

[Input form](#)[Web services](#)[Help & Documentation](#)[Bioinformatics Tools FAQ](#)[Feedback](#)[Share](#)

Tools > Sequence Similarity Searching > PSI-BLAST

## Protein Similarity Search

PSI-BLAST allows users to construct and perform a NCBI BLAST search with a custom, position-specific, scoring matrix which can help find distant evolutionary relationships. Users can specify pattern files to restrict search results using the PHI-BLAST functionality under 'more options'.

### STEP 1 - Select your database

#### PROTEIN DATABASES

UniProtKB/Swiss-Prot isoforms (The manually annotated isoforms of UniProtKB/Swiss-Prot)

### STEP 2 - Enter your input sequence

Enter or paste a PROTEIN sequence in any supported format:

Upload a file:

Αναζήτηση...

Δεν επιλέχθηκε αρχείο.

[Use a example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

### STEP 3 - Set your parameters

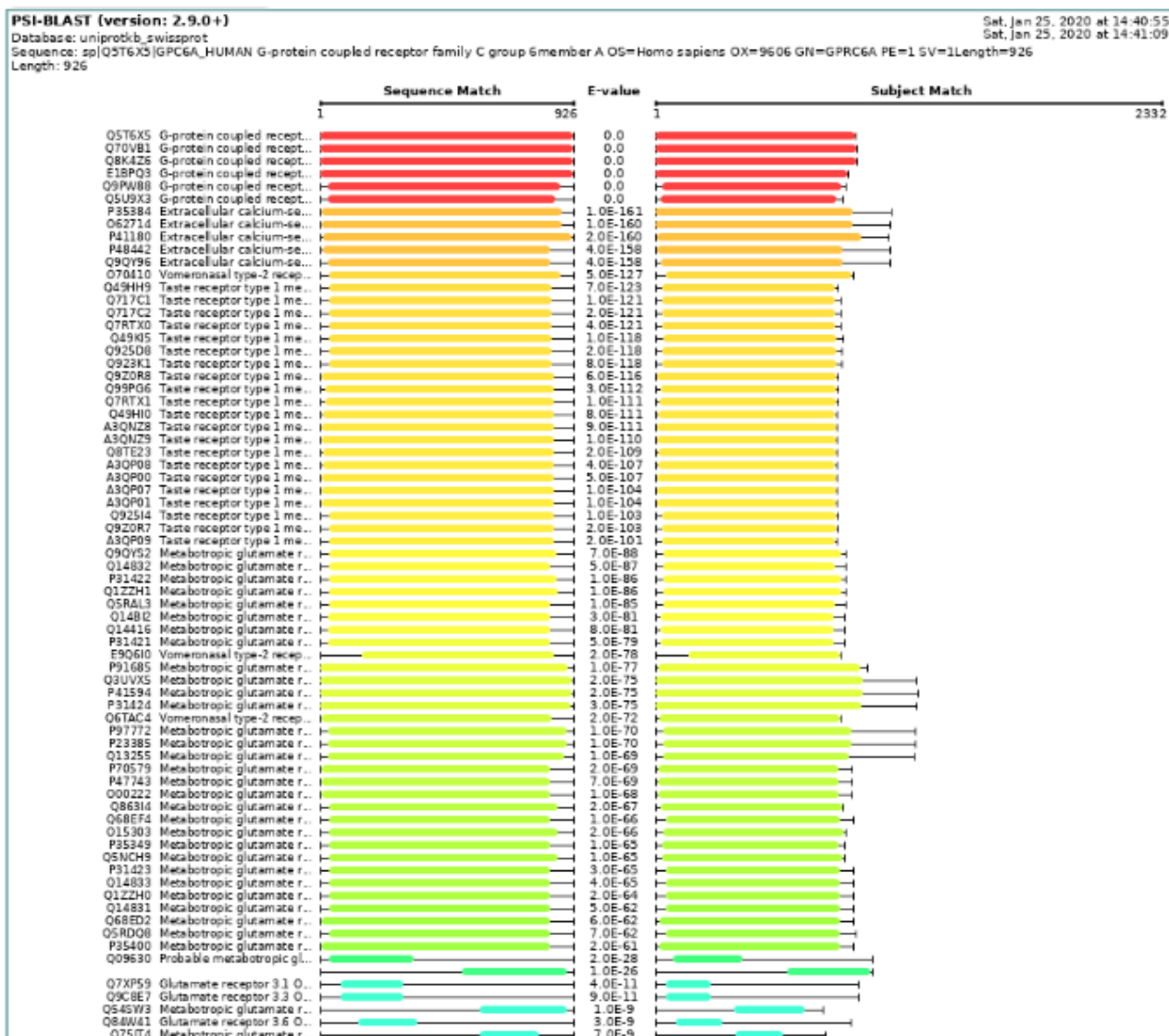
#### PSSM E-VALUE CUT-OFF

1.0e-3

Εικόνα 16

Και μπορούμε να πάρουμε τα αποτελέσματα σε διάφορες μορφές. Στην παρακάτω μας παρουσιάζονται με χρώμα και με βάση το E-value. Οι πρώτες έχουν μηδενικό οπότε μιλάμε ουσιαστικά για την ίδια ακολουθία.

Επίσης μας δείχνει ποια βάση δεδομένων χρησιμοποιήσαμε, το assertion number της κάθε πρωτεΐνης, το όνομα της και σε ποια σημεία έχουμε ομοιότητα.



Εικόνα 17

## HMMER

Το εργαλείο αυτό, χρησιμοποιεί τα profile Hidden Markov Models (HMM) προκειμένου να εντοπίσει ομόλογες ακολουθίες και να κάνει στοιχίσεις ακολουθιών. Για να το επιτύχει αυτό συγκρίνει ένα profile HMM είτε με μια βάση δεδομένων που επιλέγουμε εμείς είτε με μια ακολουθία.

Μπορούμε να το “κατεβάσουμε” τοπικά στον υπολογιστή μας είτε να το χρησιμοποιήσουμε online. Μεταβαίνουμε στη σελίδα του Ευρωπαϊκού Ινστιτούτου Βιοπληροφορικής (EBI) και επιλέγουμε το hmmer.



**HMMER**  
Biosequence analysis using profile hidden Markov Models

Home Search Results Software Help About Contact

phmmer hmmscan hmmsearch jackhmmer

## protein sequence vs protein sequence database

Paste a Sequence | Upload a File | Accession Search

Paste in your sequence or use the example

Submit Reset

▼ Sequence Database

Frequently used databases: Reference Proteomes UniProtKB SwissProt PDB Ensembl

Current database selection:  
Reference Proteomes

Εικόνα 18

Όπως βλέπουμε και στην παραπάνω εικόνα, μπορούμε να επιλέξουμε διάφορους τύπους αναζήτησης βασιζόμενοι σε πιο εργαλείο θα επιλέξουμε (phmmer, jackhmmer κτλ.)

Στη συνέχεια κάνουμε επιλογή και της βάσης έναντι της οποίας θα γίνει η αναζήτηση της πρωτεϊνικής ακολουθίας (πχ PDB)

Στα αποτελέσματα αυτό που βλέπουμε είναι ότι τα πρώτα αποτελέσματα αφορούν ουσιαστικά τον ίδιο τον υποδοχέα – ακολουθία. Ανοίγοντας ένα αποτέλεσμα έχουμε ένα e-value το οποίο δεν είναι ίδιο με αυτό του BLAST αλλά έχει ως στόχο να μας δώσει ένα score του πόσες ακολουθίες μπορούν να έχουν ίσο σκορ με αυτό που παίρνουμε στη στοίχιση τυχαία.

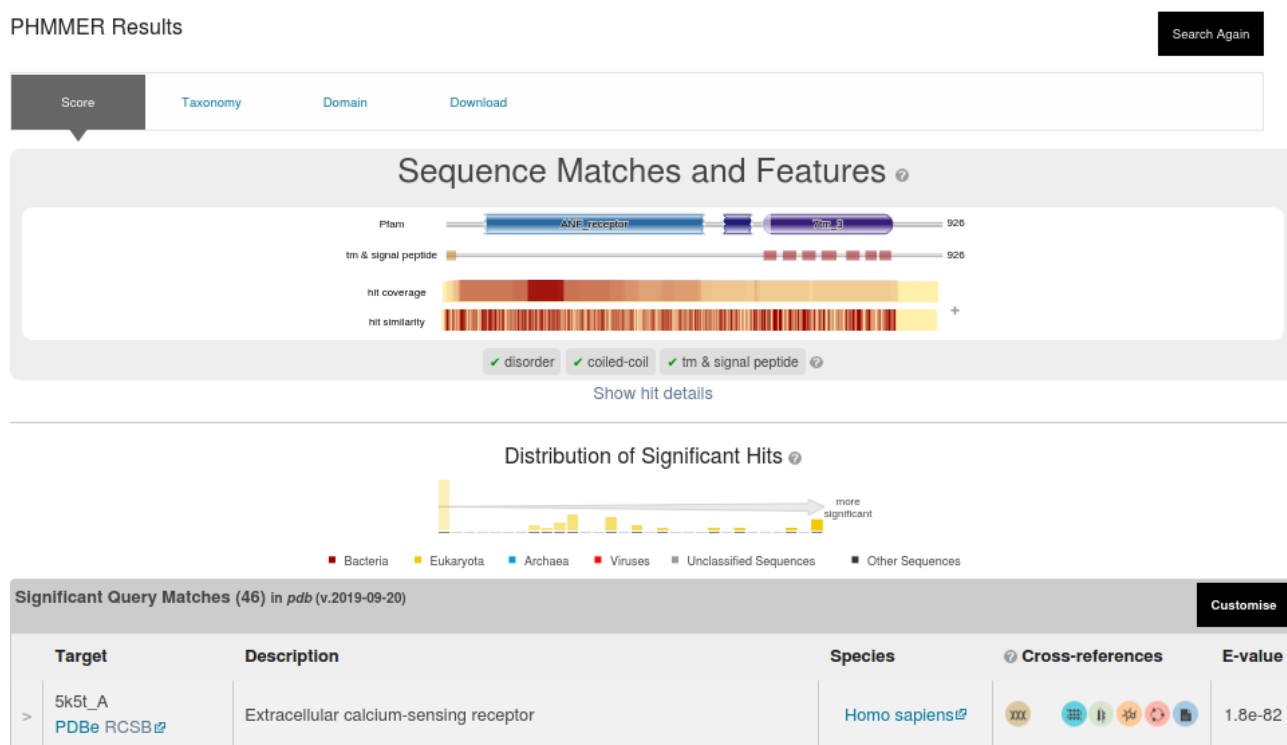
Τα profile HMM είναι εξειδικευμένα στο να πραγματοποιούν πολλαπλές στοίχισεις. Στατιστικός τρόπος για να μοντελοποιηθεί μια πολλαπλή στοίχιση και ειδικά εκεί που μπορεί να έχουμε κενά σε κάποιες από τις ακολουθίες.

Τα profile HMM έχουν αρχιτεκτονική από τα αριστερά προς τα δεξιά, πάντα θα πάμε στην επόμενη θέση στο επόμενο κατάλοιπο, δεν μένουμε σταθεροί. Σε κάθε κατάλοιπο έχουμε 3 εν δυνάμει καταστάσεις όπου μπορεί η ακολουθία μας να εκπέμπει. Match δηλ. να έχουμε το ίδιο κατάλοιπο με την στοίχιση, Insert έχουμε παραπάνω αμινοξύ και Delete που σημαίνει ότι έχουμε κενό στη στοίχιση. Με βάση αυτές τις καταστάσεις και την πολλαπλή στοίχιση υπολογίζονται τα αντίστοιχα, emission probabilities δηλαδή οι πιθανότητες στη συγκεκριμένη θέση να βρίσκεται κάποιο από τα 20 αμινοξέα.

Όπως και στα απλά HMM έτσι και εδώ κάθε θέση δεν θεωρείται ανεξάρτητη αλλά συνδέεται με την προηγούμενη θέση.

Στο HMMER μπορούμε να χρησιμοποιήσουμε είτε μια μόνο ακολουθία είτε μια πολλαπλή στοίχιση.

Τα αποτελέσματα χρησιμοποιώντας τον υποδοχέα από την Uniprot με assertion number [Q5T6X5](#) είναι τα παρακάτω:



Εικόνα 19

Μας επιστρέφει όπως παρατηρούμε και στην εικόνα, το πετίδιο οδηγητή και τα 7 διαμεμβρανικά τμήματα. Επίσης μας επιστρέφει και παρόμοιες ακολουθίες με ένα E-value που όμως είναι διαφορετικό από του BLAST και συνεπώς δεν μπορεί να υπάρξει σύγκριση.

Στο hmmer όμως μπορούμε αντί για μια μόνο ακολουθία να εισάγουμε μια πολλαπλή στοίχιση. Εμείς θα εισάγουμε την πολλαπλή στοίχιση που κάναμε με το T-Coffee. Τα assertion numbers από την Uniprot είναι τα εξής: [Q5T6X5](#), [Q14833](#), [O15303](#), [Q14416](#), [Q14831](#) και [Q14832](#).

Στο πεδίο search επιλέγουμε αυτή τη φορά το hmsearch. Εδώ μας δίνονται η δυνατότητα να εισάγουμε είτε ένα profile HMM είτε μια πολλαπλή στοίχιση. Αφού εισάγουμε το αρχείο με τα αποτελέσματα που πήραμε από το T-Coffee και επιλέξουμε τη βάση που επιθυμούμε, εδώ διαλέξαμε την PDB, πατάμε Submit.

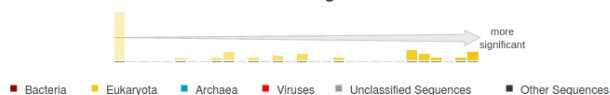
Αυτό που μας ενδιαφέρει δεν είναι να βρούμε με ποιες μοιάζει μια ακολουθία, αυτό το κάνει το BLAST αλλά να βρούμε τις θέσεις πάνω στην ακολουθία που είναι πιο σημαντικές για τη λειτουργία της πρωτεΐνης. Τις πιο συντηρημένες θέσεις δλδ.

## HMMSEARCH Results

Search Again

Score	Taxonomy	Domain	Download
-------	----------	--------	----------

### Distribution of Significant Hits



« First « Previous Page 1 of 2 Next » Last »

### Significant Query Matches (52) in pdb (v.2019-09-20)

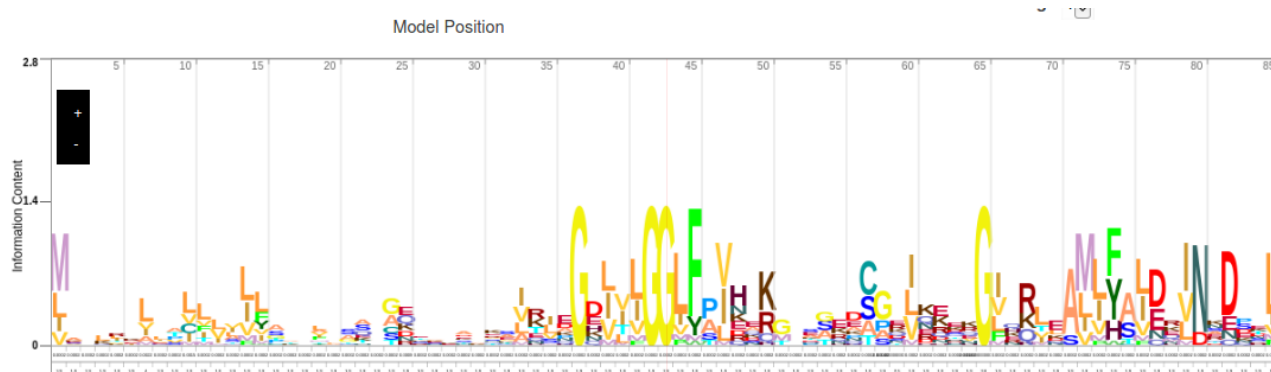
Customise

Target	Description	Species	Cross-references	E-value
> 6n4x_A PDBe RCSB	Metabotropic glutamate receptor 5	<a href="#">Homo sapiens</a>		3.2e-279
> 6n52_A PDBe RCSB	Metabotropic glutamate receptor 5	<a href="#">Homo sapiens</a>		4.1e-279
> 2e4u_A PDBe RCSB	Metabotropic glutamate receptor 3	<a href="#">Rattus norvegicus</a>		3.7e-277

Εικόνα 20

Το Hmmer την πολλαπλή στοίχιση που δώσαμε την μετέτρεψε αυτόματα σε profile HMM χωρίς περαιτέρω επεξεργασία εκ μέρους μας.

Επίσης μας παρέχει και το παρακάτω διάγραμμα όπου φαίνεται πιο αμινοξύ επικρατεί στις διάφορες θέσεις.



Column number:

Εικόνα 21

## Pfam

Η Pfam είναι πρωτεϊνική βάση δεδομένων που χρησιμοποιεί profile HMM τα οποία φτιάχνει μέσω του Hmmer.

Η Pfam μας δίνει πληροφορίες σχετικά με μια πρωτεϊνική οικογένεια και πιο συγκεκριμένα πληροφορίες για τις στοιχίσεις από τις οποίες φτιάχτηκε η εν λόγω οικογένεια και αν υπάρχει πληροφορία για τη δομή μας δίνει και την αντίστοιχη δομή ή λίστα δομών.

**Protein: GPC6A\_HUMAN (Q5T6X5)**

1 architecture 1 sequence 0 interactions 1 species 0 structures

**Summary**

This is the summary of UniProt entry [GPC6A\\_HUMAN](#) (Q5T6X5).

**Description:** G-protein coupled receptor family C group 6 member A


**Source organism:** [Homo sapiens \(Human\)](#) (NCBI taxonomy ID [9606](#))

**Length:** 926 amino acids

**Reference Proteome:** ✓

**Pfam domains**

This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will take you to the page describing that Pfam entry. The table below gives the domain boundaries for each of the domains. [More...](#)



[Download](#) the data used to generate the domain graphic in JSON format.

Source	Domain	Start	End
sig_p	n/a	1	18
Pfam	<a href="#">ANF_receptor</a>	72	483
Pfam	<a href="#">NCD3G</a>	518	572
low_complexity	n/a	591	614
Pfam	<a href="#">7tm_3</a>	593	836
transmembrane	n/a	593	616
transmembrane	n/a	628	652
transmembrane	n/a	664	688
transmembrane	n/a	700	727
transmembrane	n/a	747	771
low_complexity	n/a	758	770

Εικόνα 22

Όπως παρατηρούμε και στην εικόνα 22, πατώντας πάνω σε κάποιο domain μπορούμε να βρούμε πληροφορίες σχετικά με αυτό το domain της πρωτεΐνης μας. Με βάση τα domain μπορούμε να ξεχωρίσουμε τις διαφορετικές οικογένειες που ανήκουν οι πρωτεΐνες. Στο συγκεκριμένο παράδειγμα δεν μας εμφανίζει κάποια δομή (0 structures). Μας δίνει πληροφορίες για το πεπτίδιο οδηγητή, που ξεκινάνε και τελειώνουν τα διαμεμβρανικά τμήματα. Καθώς και πληροφορίες σχετικά με την ακολουθία αλλά και το είδος που ανήκει η πρωτεΐνη.

## Πρόγνωση Τοπολογίας Διαμεμβρανικών Πρωτεϊνών

Σκοπός της πρόγνωσης τοπολογίας στις διαμεμβρανικές πρωτεΐνες είναι να βρούμε που είναι το enter ή αλλιώς N-terminal, η είσοδος δλδ, τα διαμεμβρανικά τμήματα που διαπερνούν την μεμβράνη και τέλος το c-terminal.

Επειδή έχουν διαφορετικά χαρακτηριστικά, τα α-ελικοειδή (α-helix) και τα β-βαρέλια (β-barrels), υπάρχουν και διαφορετικές μέθοδοι πρόγνωσης για την κάθε υποκατηγορία ξεχωριστά.

# A-helix

## TOPCONS

Το TOPCONS είναι ένα εργαλείο που χρησιμοποιεί HMM καθώς και consensus για την πρόβλεψη της τοπολογίας. Πιο συγκεκριμένα μπορεί να χρησιμοποιήσει μόνο την ακολουθία, να φτιάξει profile από πολλαπλή στοίχιση καθώς και να εντοπίσει το πεπτίδιο οδηγητή.

Χρησιμοποιεί πέντε (5) μεθόδους οι οποίες είναι εκπαιδευμένες σε διαφορετικά σύνολα εκπαίδευσης η καθεμία. Δεν έχουν όλες τις ίδιες ιδιότητες ή δεν προβλέπουν με την ίδια ακρίβεια. Η Philius δεν χρησιμοποιεί profile παρά μόνο την ακολουθία σε μορφή FASTA. Αφού πάρουμε τις προβλέψεις φτιάχνουμε ένα topology profile και στη συνέχεια εφαρμόζουμε consensus. Στη συνέχεια μέσω του αλγόριθμου Vitterbi πραγματοποιούμε ένα φιλτράρισμα και καταλήγουμε στο most optimal path βασισμένο στο μοντέλο που επιλέξαμε.

Ένα μη επιτρεπτό μονοπάτι δεν θα εμφανιζόταν επειδή δεν θα είχε υψηλό score.

Μεταβαίνουμε στη σελίδα του [TOPCONS](#) και εισάγουμε τις ακολουθίες που θέλουμε. Στη συγκεκριμένη περίπτωση μέσω της Uniprot επιλέξαμε να εισάγουμε ion channels.

### Results

- Submitted: 2020-01-31 16:43:34 UTC
- Status: **Finished**
- Waiting time: 5 secs
- Running Time: 0 sec

Results of your prediction with jobid: **rst\_0reOgS**

Zipped folder of your result can be found in [rst\\_0reOgS.zip](#)

Dumped prediction in one text file can be found in [query.result.txt](#)

The sequence(s) you submitted can be found in [query.raw.fa](#)

Total number of protein sequences:	5
TM proteins predicted by the consensus:	4 / 5 (80.0 %)
TM proteins predicted by any of the sub-predictors:	5 / 5 (100.0 %)
Non-TM proteins predicted by the consensus:	1 / 5 (20.0 %)
Non-TM proteins not predicted as TM by any of the sub-predictors:	0 / 5 (0.0 %)
Proteins with signal peptide predicted by the consensus:	0 / 5 (0.0 %)
Proteins with signal peptide predicted by any of the sub-predictors:	0 / 5 (0.0 %)

Please click the link of each sequence below to see the result of individual sequence  
Note that the table below is sortable by clicking the table header

Show  entries

Search:

No.	Length	numTM	SignalPeptide	RunTime(s)	SequenceName	Source	FinishDate
1	241	0	No	0.0	<a href="#">sp O00299 CLIC1_HUMAN Chloride</a>	cached	2020-01-31 17:43:39 CET
2	123	1	No	0.0	<a href="#">sp Q9Y6J6 KCNE2_HUMAN Potassiu</a>	cached	2020-01-31 17:43:39 CET
3	528	2	No	0.0	<a href="#">sp P78348 ASIC1_HUMAN Acid-sen</a>	cached	2020-01-31 17:43:39 CET
4	531	1	No	0.0	<a href="#">sp Q9UHC3 ASIC3_HUMAN Acid-sen</a>	cached	2020-01-31 17:43:39 CET
5	512	1	No	0.0	<a href="#">sp Q16515 ASIC2_HUMAN Acid-sen</a>	cached	2020-01-31 17:43:39 CET

Showing 1 to 5 of 5 entries

Previous

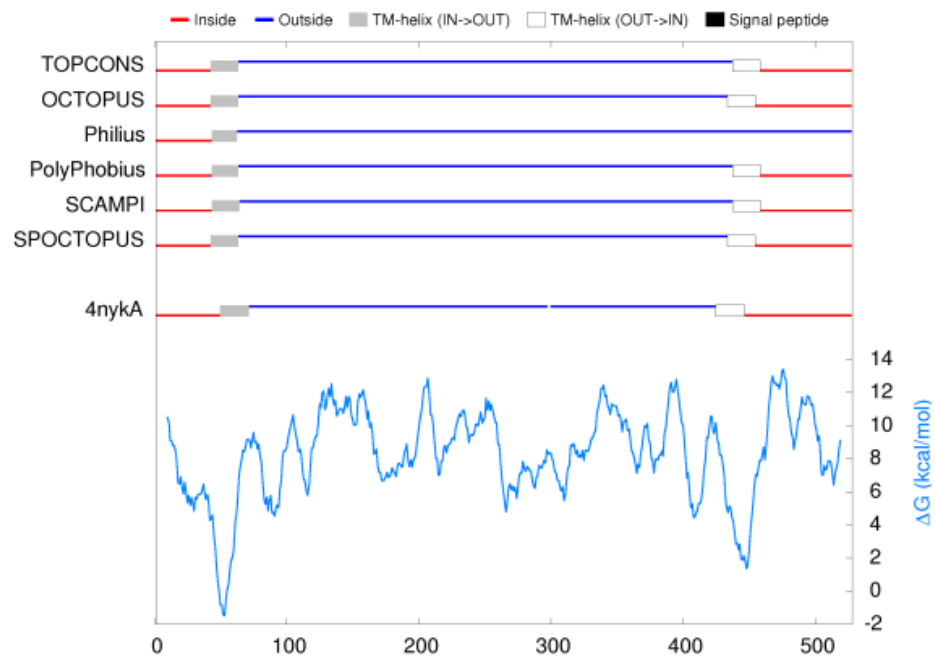
1

Next

### Εικόνα 23

Μπορούμε να δούμε περισσότερες και λεπτομερέστερες πληροφορίες για κάθε μια πρωτεΐνη επιλέγοντας την καθώς και τα αποτελέσματα της κάθε μεθόδου.





Εικόνα 24

Βλέπουμε αν ξεκινάει από Outside ή από Inside καθώς και σε πιο κατάλοιπο ξεκινάει το κάθε τμήμα.

Στη συνέχεια θα χρησιμοποιήσουμε και το εργαλείο HMM-TM και θα συγκρίνουμε τα αποτελέσματα από τα δυο εργαλεία.

You have submitted 5 sequence(s)  
The sequence(s) you submitted can be found [here](#)

The status of your job is: **Finished**  
Time running: 3.0 seconds

HMM-TM results				
Entry	Length	#TM	Reliability	Image
sp_O00299_CLIC1_HUMA	246	--	0.995	--
sp_P78348_ASIC1_HUMA	537	2	0.978	<a href="#">Download</a>
sp_Q16515_ASIC2_HUMA	520	2	0.927	<a href="#">Download</a>
sp_Q9UHC3_ASIC3_HUMA	540	2	0.821	<a href="#">Download</a>
sp_Q9Y6J6_KCNE2_HUMA	126	1	0.743	<a href="#">Download</a>

[Download](#) results in plain text format  
[Download](#) topologies only in plain text format  
[Download](#) all files related to this submission as a tarball

[Run again](#)

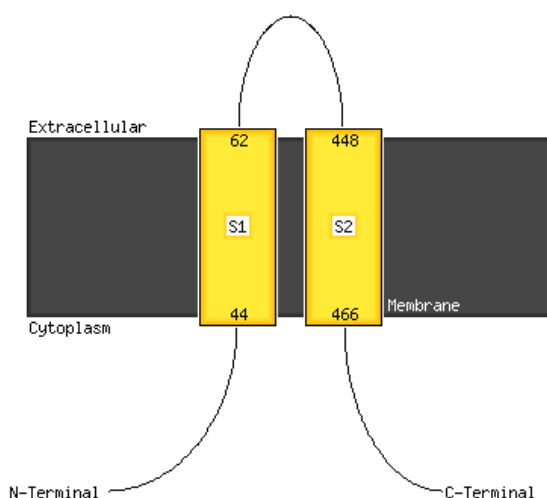
Εικόνα 25

Παρατηρούμε ότι ενώ και το HMM-TM βρήκε στις 4 από τις 5 ακολουθίες μας διαμεμβρανικά τμήματα όπως ακριβώς και το TOPCONS, στον αριθμό των διαμεμβρανικών τμημάτων στις

ακολουθίες μας και πιο συγκεκριμένα στις Q9UHC3 Q16515 βρήκε 2 τμήματα αντί ενός που βρήκε το TOPCONS.

Το HMM-TM μας δίνει και μια οπτικοποίηση των διαμεμβρανικών τμημάτων της ακολουθίας. Πόσα τμήματα, σε πιο κατάλοιπο ξεκινάνε και τελειώνουν, αν ξεκινάει από Inside ή από την εξωκυττάρια περιοχή.

HMM-TM topology prediction image for sp\_P78348\_ASIC1\_HUMA



Εικόνα 26

Οι διαφορές που εμφανίζουν οι διάφορες μέθοδοι οφείλονται στον τρόπο που έχουν σχεδιαστεί τα μοντέλα και τα σύνολα εκπαίδευσης που χρησιμοποιεί η καθεμία.

## β-Barrels

Τα β-βαρέλια είναι μια πιο μικρή κατηγορία των διαμεμβρανικών πρωτεϊνών. Εντοπίζονται αποκλειστικά και μόνο σε αρνητικά κατά Gram βακτήρια, στα μιτοχόνδρια και στους χλωροπλάστες. Επειδή είναι μικρός αριθμός, γύρω στις 50, δυσκολευόμαστε να τις προβλέψουμε μιας και δεν έχουμε ένα ικανοποιητικό σύνολο εκπαίδευσης.

### PRED-TMBB2

Εργαλείο για την πρόβλεψη της τοπολογίας αλλά και για ταξινόμηση. Δουλεύει με εισαγωγή ακολουθίας και όχι με profile. Επειδή όπως είπαμε ο αριθμός των πρωτεϊνών για τις οποίες ξέρουμε τη δομή και είναι διαφορετικές μεταξύ τους είναι μικρός χρησιμοποιούμε cross validation για την εκπαίδευση του.

Πάλι από την Uniprot παίρνουμε την πρωτεΐνη με assertion number [Q9Y277](#) και εισάγουμε στο PRED-TMBB2 την ακολουθία σε FASTA μορφή.

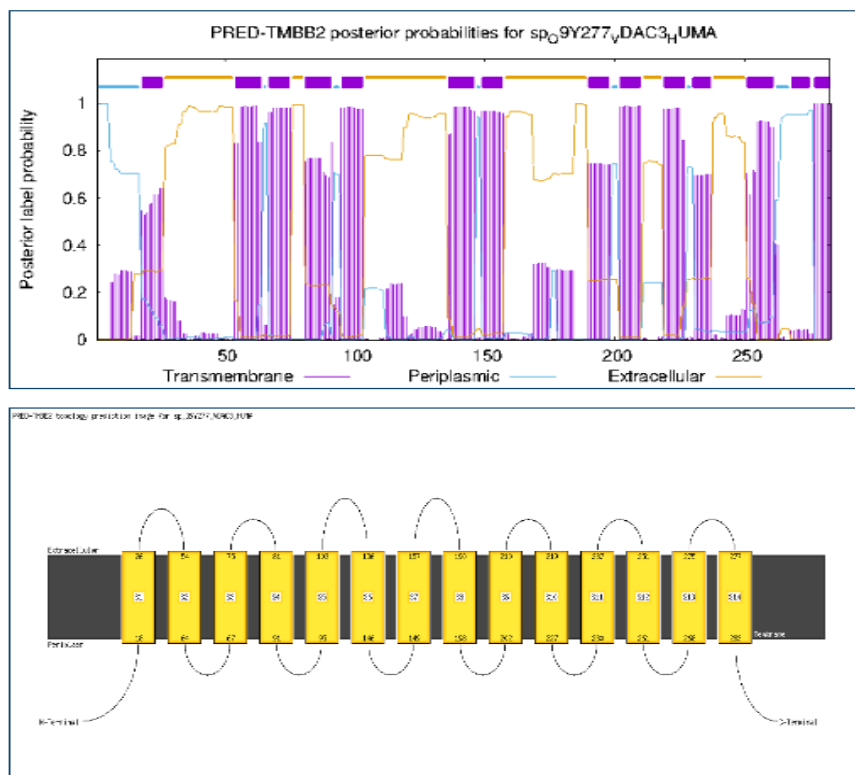
## PRED-TMBB2 results

[illegible]

Εικόνα 27

Όπως βλέπουμε και από την εικόνα 27 παίρνουμε πληροφορίες σχετικά με το μήκος της ακολουθίας, με το αν έχει πεπτίδιο-οδηγητή, τον αριθμό των β-κλώνων ( $\beta$ -strands), την αξιοπιστία της πρόβλεψης. Ακόμη μας παρουσιάζει και την προβλεπόμενη τοπολογία σε FASTA format.

Τέλος υπάρχει και γραφική απεικόνιση των αποτελεσμάτων αυτών.



Εικόνα 28

