Final Project Bioinformatics Course

Ben Kapner
Lihi Bik

We were given to do the final project on the chimera **ERG - TMPRSS2**

Question 1:

Find FASTA sequences of chimeric RNAs of the parental genes: ERG and TMPRSS2.
Collect all the found sequences in the project (at least 6 sequences).

Answer 1:

We want to find info and data in the Bioinformatics Databases: NCBI, GenBank, GO and others and then find FASTA sequences.
First thing that we did, we went to NCBI and found all the results for ERG - TMPRSS2, we needed to search it as ERG:TMPRSS2 under the Nucleotide DB and Homo sapiens and we found exactly 6 sequences.

This is our results link:
https://www.ncbi.nlm.nih.gov/nuccore?term=(ERG+%3A+TMPRSS2)+AND+%22Homo+sapiens%22%5Bporgn%5D&cmd=DetailsSearch&log$=activity

We copied all the sequences and created a FASTA file with all 6:

First sequence:
>DQ831522.1 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
GGAGGCGGAGGCGGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAGGTTATTCC
AGGATCTTTGGAGACCCGAGGAAAGCCGTGTTGACCAAAAGCAAGACAAATGACTCACAGAGAAAAAA
GATGGCAGAACCAAGGGCAACTAAAGCCGTCAGGTTCTGAACAGCTGGTAGATGGGCTGGCTTACTGA
AGGACATGATTCAGACTGTCCCGGACCCAGCAGCTCATATCAAGGAAGCCTTATCAGTTGTGAGTGAGG
ACCAGTCGTTGTTTGAGTGTGCCTACA

Second sequence:
>DQ831521.1 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
GCGGAGTGCGAGGGGCGGGGACGCCGCCTGGACGCGGCAGCCGTCAGGTTCTGAACAGCTGGTAGAT
GGGCTGGCTTACTGAAGGACATGATTCAGACTGTCCCGGACCCAGCAGCTCATATCAAGGAAGCCTTAT
CAGTTGTGAGTGAGGACCAGTCGTTGTTTGAGTGTGCCTACGGAACGCCACACCTGGCTAAGACAGAG
ATGACCGCGTCCTCCTCCAGCGACTATGGACAGACTTCCAAGATGAGCCCACGCGTCCCTCAGCAGGAT
TGGCTGTCTCAACCCCCAGCCAGGGTCACCATCAAAATGGAATGTAACCCTAGCCAGGTGAATGGCTCA
AGGAACTCTCCTGATGAATGCAGTGTGGCCAAAGGCGGGAAGATGGTGGGCAGCCCAGACACCGTTGG
GATGAACTACGGCAGCTACATGGAGGAGAAGCACATGCCACCCCCAAACATGACCACGAACGAGCGCA
GAGTTATCGTGCCAGCAGATCCTACGCTATGGAGTACAGACCATGTGCGGCAGTGGCTGGAGTGGGCG
GTGAAAGAATATGGCCTTCCAGACGTCAACATCTTGTTATTCCAGAACATCGATGGGAAGGAACTGTGC
AAGATGACCAAGGACGACTTCCAGAGGCTCACCCCCAGCTACAACGCC

Third sequence:
>EF194202.1 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
TTCGCTTGCTGTTTCTGTGACTTTACGCTCTGCTGACCTAGAGGTCTTAGTTCCGGAGGGAGGAATGCTG
CCACCAGGAGACACAACAATGATTCAATTAAACTAGAATTTACGACTGCCACCTGGCCACGCTGAGCTCC
ACATGCCTCTGAATCAAAAGGCAAAGAGAGAGTTATGCATTGGCTGGGGAGACCCATCTGGACTACCAA
GGAGAAGCTATAGACTACTTCTACTCCACCAGGAAGGAAGCCTTATCAGTTGTGAGTGAGGACCAGTCG
TTGTTTGAGTGTGCCTACGGAACGCCACACCTGGCTAAGACAGAGATGACCGCGTCCTCCTCCAGCGAC
TATGGACAGACTTCCAAGATGAGCCCACGCGTCCCTCAGCAGGATTGGCTGTCTCAACCCCCAGCCAGG
GTCACCATCAAAATGGAATGTAACCCTAGCCAGGTGAATGGCTCAAGGAACTCTCCTGATGAATGCAGT
GTGGCCAAAGGCGGGAAGATGGTGGGCAGCCCAGACACCGTTGGGATGAACTACGGCAGCTACATGG
AGGAGAAGCACATGCCACCCCCAAACATGACCACGAACGAGCGCAGAGTTATCGTGCCAGCAGATCCT
ACGCTATGGAGTACAGACCATGTGCGGCAGTGGCTGGAGTGGGCGGTGAAA

Forth sequence:
>FJ423744.1 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
GGAGTAGGCGCGAGCTAAGCAGGAGGCGGAGGCGGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGC
CTGGAGCGCGGCAGGAAGCCTTATCAGTTGTGAGTGAGGACCAGTCGTTGTTTGAGTGTGCCTACGGA
ACGCCACACCTGGCTAAGACAGAGATGACCGCGTCCTCCTCCAGCGACTATGGACAGACTTCCAAGATG
AGCCCACGCGTCCCTCAGCAGGATTGGCTGTCT

Fifth sequence:
>EU432099.1 Homo sapiens TMPRSS2-ERG prostate cancer specific isoform 1 (ERG) mRNA, complete cds, alternatively spliced
GCAGGAGGCGGAGGCGGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAGGAAG
CCTTATCAGTTGTGAGTGAGGACCAGTCGTTGTTTGAGTGTGCCTACGGAACGCCACACCTGGCTAAGA
CAGAGATGACCGCGTCCTCCTCCAGCGACTATGGACAGACTTCCAAGATGAGCCCACGCGTCCCTCAGC
AGGATTGGCTGTCTCAACCCCCAGCCAGGGTCACCATCAAAATGGAATGTAACCCTAGCCAGGTGAATG
GCTCAAGGAACTCTCCTGATGAATGCAGTGTGGCCAAAGGCGGGAAGATGGTGGGCAGCCCAGACACC
GTTGGGATGAACTACGGCAGCTACATGGAGGAGAAGCACATGCCACCCCCAAACATGACCACGAACGA
GCGCAGAGTTATCGTGCCAGCAGATCCTACGCTATGGAGTACAGACCATGTGCGGCAGTGGCTGGAGT
GGGCGGTGAAAGAATATGGCCTTCCAGACGTCAACATCTTGTTATTCCAGAACATCGATGGGAAGGAAC
TGTGCAAGATGACCAAGGACGACTTCCAGAGGCTCACCCCCAGCTACAACGCCGACATCCTTCTCTCACA
TCTCCACTACCTCAGAGAGACTCCTCTTCCACATTTGACTTCAGATGATGTTGATAAAGCCTTACAAAACT
CTCCACGGTTAATGCATGCTAGAAACACAGGGGGTGCAGCTTTTATTTTCCCAAATACTTCAGTATATCCT
GAAGCTACGCAAAGAATTACAACTAGGCCAGTCTCTTACAGATAAAACAACAGAACCAGTGCCAGAAAG
CAGCCTTCCCTTACATGGGCACTTCTGCCAAGCATATGAGTTCATTGCCTTGAAGATCAAAGTCAAAGAG
AAATGGAGAGGGTGTTGAAATGATCAGCGAAAATTAAATGTAAAATATATTCTTATTGGAAGTCTGATG
CTCTATTATCAATAAAGGACACATAGCAAAGATAAAAAAAAAAAAAAAAAAAAAA

Sixth sequence:
>EU090248.1 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
CGCGAGCTAAGCAGGAGGCGGAGGCGGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCG
CGGCAGGTTATTCCAGGATCTTTGGAGACCCGAGGAAAGCCGTGTTGACCAAAAGCAAGACAAATGAC
TCACAGAGAAAAAAGATGGCAGAACCAAGGGCAACTAAAGCCGTCAGGTTCTGAACAGCTGGTAGATG
GGCTGGCTTACTGAAGGACATGATTCAGACTGTCCCGGACCCAGCAGCTCATATCAAGGAACTCTCCTG
ATGAATGCAGTGTGGCCAAAA

Question 2+3:

Question 2:

Translate the chimeric RNAs in 6 frames and find the correct frame of chimeric proteins. Explain what 6 frames are.

Question 3:

Find a correct protein sequence in FASTA format (using transeq or any other software). Give the longest protein sequence found for the ERG-TMPRSS2 chimeric protein in NCBI and/or other databases. Does this sequence have a correct "start codon"?

Answer 2+3:

In order to translate the chimeric RNAs in 6 frames we used transeq
https://www.ebi.ac.uk/jdispatcher/st/emboss_transeq
We copied each of the chimeric RNAs sequences to the input box, checked that the parameters where 6 frames.
We received the following 6 results for the sequences:

First sequence:

```
>DQ831522.1_1 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
GGGGGGGGRGAGSAAWSAAGYSRIFGDPRKAVLTKSKTNDSQRKKMAEPRATKAVRF*TA
GRWAGLLKDMIQTVPDPAAHIKEALSVVSEDQSLFECAYX
>DQ831522.1_2 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
EAEAEAEGEGRGAPPGARQVIPGSLETRGKPC*PKARQMTHREKRWQNQGQLKPSGSEQL
VDGLAY*RT*FRLSRTQQLISRKPYQL*VRTSRCLSVPT
>DQ831522.1_3 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
RRRRRRRARGGERRLERGRLFQDLWRPEESRVDQKQDK*LTEKKDGRTKGN*SRQVLNSW
*MGWLTEGHDSDCPGPSSSYQGSLISCE*GPVVV*VCLX
>DQ831522.1_4 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
VGTLKQRLVLTHN**GFLDMSCWVRDSLNHVLQ*ASPSTSCSEPDGFSCPWFCHLFSL*V
ICLAFGQHGFPRVSKDPGITCRAPGGAPRPSPSASASAS
>DQ831522.1_5 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
CRHTQTTTGPHSQLIRLP*YELLGPGQSESCPSVSQPIYQLFRT*RL*LPLVLPSFFSVS
HLSCFWSTRLSSGLQRSWNNLPRSRRRSPPLALRLRLRLX
>DQ831522.1_6 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
*AHSNNDWSSLTTDKASLI*AAGSGTV*IMSFSKPAHLPAVQNLTALVALGSAIFFLCES
FVLLLVNTAFLGSPKILE*PAALQAALPAPRPPPPPPPP
```

Second sequence:

```
>DQ831521.1_1 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
AECEGRGRRLDAAAVRF*TAGRWAGLLKDMIQTVPDPAAHIKEALSVVSEDQSLFECAYG
TPHLAKTEMTASSSSDYGQTSKMSPRVPQQDWLSQPPARVTIKMECNPSQVNGSRNSPDE
CSVAKGGKMVGSPDTVGMNYGSYMEEKHMPPPNMTTNERRVIVPADPTLWSTDHVRQWLE
WAVKEYGLPDVNILLFQNIDGKELCKMTKDDFQRLTPSYNA
>DQ831521.1_2 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
RSARGGDAAWTRQPSGSEQLVDGLAY*RT*FRLSRTQQLISRKPYQL*VRTSRCLSVPTE
RHTWLRQR*PRPPPATMDRLPR*AHASLSRIGCLNPQPGSPSKWNVTLAR*MAQGTLLMN
AVWPKAGRWWAAQTPLG*TTAATWRRSTCHPQT*PRTSAELSCQQILRYGVQTMCGSGWS
GR*KNMAFQTSTSCYSRTSMGRNCAR*PRTTSRGSPPATTP
>DQ831521.1_3 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
GVRGAGTPPGRGSRQVLNSW*MGWLTEGHDSDCPGPSSSYQGSLISCE*GPVVV*VCLRN
ATPG*DRDDRVLLQRLWTDFQDEPTRPSAGLAVSTPSQGHHQNGM*P*PGEWLKELS**M
QCGQRREDGGQPRHRWDELRQLHGGEAHATPKHDHERAQSYRASRSYAMEYRPCAAVAGV
GGERIWPSRRQHLVIPEHRWEGTVQDDQGRLPEAHPQLQRX
>DQ831521.1_4 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
GVVAGGEPLEVVLGHLAQFLPIDVLE*QDVDVWKAIFFHRPLQPLPHMVCTP*RRICWHD
NSALVRGHVWGWHVLLLHVAAVVHPNGVWAAHHLPAFGHTAFIRRVP*AIHLARVTFHFD
GDPGWGLRQPILLRDAWAHLGSLSIVAGGGRGHLCLSQVWRSVGTLKQRLVLTHN**GFL
DMSCWVRDSLNHVLQ*ASPSTSCSEPDGCRVQAASPPLALR
>DQ831521.1_5 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
RCSWG*ASGSRPWSSCTVPSHRCSGITRC*RLEGHILSPPTPATAAHGLYSIA*DLLAR*
LCARSWSCLGVACASPPCSCRSSSQRCLGCPPSSRLWPHCIHQESSLSHSPG*GYIPF*W
*PWLGVETANPAEGRVGSSWKSVHSRWRRTRSSLS*PGVAFRRHTQTTTGPHSQLIRLP*
YELLGPGQSESCPSVSQPIYQLFRT*RLPRPGGVPAPRTPX
>DQ831521.1_6 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
AL*LGVSLWKSSLVILHSSFPSMFWNNKMLTSGRPYSFTAHSSHCRTWSVLHSVGSAGTI
TLRSFVVMFGGGMCFSSM*LP*FIPTVSGLPTIFPPLATLHSSGEFLEPFTWLGLHSILM
VTLAGG*DSQSC*GTRGLILEVCP*SLEEDAVISVLARCGVP*AHSNNDWSSLTTDKASL
I*AAGSGTV*IMSFSKPAHLPAVQNLTAAASRRRPRPSHSA
```

Third sequence:

```
>EF194202.1_1 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
FACCFCDFTLC*PRGLSSGGRNAATRRHNNDSIKLEFTTATWPR*APHASESKGKERVMH
WLGRPIWTTKEKL*TTSTPPGRKPYQL*VRTSRCLSVPTERHTWLRQR*PRPPPATMDRL
PR*AHASLSRIGCLNPQPGSPSKWNVTLAR*MAQGTLLMNAVWPKAGRWWAAQTPLG*TT
AATWRRSTCHPQT*PRTSAELSCQQILRYGVQTMCGSGWSGR*X
>EF194202.1_2 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
SLAVSVTLRSADLEVLVPEGGMLPPGDTTMIQLN*NLRLPPGHAELHMPLNQKAKRELCI
GWGDPSGLPRRSYRLLLLHQEGSLISCE*GPVVV*VCLRNATPG*DRDDRVLLQRLWTDF
QDEPTRPSAGLAVSTPSQGHHQNGM*P*PGEWLKELS**MQCGQRREDGGQPRHRWDELR
QLHGGEAHATPKHDHERAQSYRASRSYAMEYRPCAAVAGVGGEX
>EF194202.1_3 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
RLLFL*LYALLT*RS*FRREECCHQETQQ*FN*TRIYDCHLATLSSTCL*IKRQRESYAL
AGETHLDYQGEAIDYFYSTRKEALSVVSEDQSLFECAYGTPHLAKTEMTASSSSDYGQTS
KMSPRVPQQDWLSQPPARVTIKMECNPSQVNGSRNSPDECSVAKGGKMVGSPDTVGMNYG
SYMEEKHMPPPNMTTNERRVIVPADPTLWSTDHVRQWLEWAVK
>EF194202.1_4 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
SPPTPATAAHGLYSIA*DLLAR*LCARSWSCLGVACASPPCSCRSSSQRCLGCPPSSRLW
PHCIHQESSLSHSPG*GYIPF*W*PWLGVETANPAEGRVGSSWKSVHSRWRRTRSSLS*P
GVAFRRHTQTTTGPHSQLIRLPSWWSRSSL*LLLGSPDGSPQPMHNSLFAF*FRGMWSSA
WPGGSRKF*FN*IIVVSPGGSIPPSGTKTSRSAERKVTETASE
>EF194202.1_5 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
FTAHSSHCRTWSVLHSVGSAGTITLRSFVVMFGGGMCFSSM*LP*FIPTVSGLPTIFPPL
ATLHSSGEFLEPFTWLGLHSILMVTLAGG*DSQSC*GTRGLILEVCP*SLEEDAVISVLA
RCGVP*AHSNNDWSSLTTDKASFLVE*K*SIASPW*SRWVSPANA*LSLCLLIQRHVELS
VARWQS*ILV*LNHCCVSWWQHSSLRN*DL*VSRA*SHRNSKRX
>EF194202.1_6 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
FHRPLQPLPHMVCTP*RRICWHDNSALVRGHVWGWHVLLLHVAAVVHPNGVWAAHHLPAF
GHTAFIRRVP*AIHLARVTFHFDGDPGWGLRQPILLRDAWAHLGSLSIVAGGGRGHLCLS
QVWRSVGTLKQRLVLTHN**GFLPGGVEVVYSFSLVVQMGLPSQCITLSLPFDSEACGAQ
RGQVAVVNSSLIESLLCLLVAAFLPPELRPLGQQSVKSQKQQAX
```

Forth sequence:

```
>FJ423744.1_1 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
GVGAS*AGGGGGGGGRGAGSAAWSAAGSLISCE*GPVVV*VCLRNATPG*DRDDRVLLQR
LWTDFQDEPTRPSAGLAVX
>FJ423744.1_2 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
E*ARAKQEAEAEAEGEGRGAPPGARQEALSVVSEDQSLFECAYGTPHLAKTEMTASSSSD
YGQTSKMSPRVPQQDWLS
>FJ423744.1_3 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
SRRELSRRRRRRRRARGGERRLERGRKPYQL*VRTSRCLSVPTERHTWLRQR*PRPPPAT
MDRLPR*AHASLSRIGCL
>FJ423744.1_4 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
DSQSC*GTRGLILEVCP*SLEEDAVISVLARCGVP*AHSNNDWSSLTTDKASCRAPGGAP
RPSPSASASASCLARAYS
>FJ423744.1_5 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
RQPILLRDAWAHLGSLSIVAGGGRGHLCLSQVWRSVGTLKQRLVLTHN**GFLPRSRRRS
PPLALRLRLRLLLSSRLLX
>FJ423744.1_6 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
TANPAEGRVGSSWKSVHSRWRRTRSSLS*PGVAFRRHTQTTTGPHSQLIRLPAALQAALP
APRPPPPPPPPA*LAPTP
```

Fifth sequence:

```
>EU432099.1_1 Homo sapiens TMPRSS2-ERG prostate cancer specific isoform 1 (ERG) mRNA, complete cds, alternatively spliced
AGGGGGGGGRGAGSAAWSAAGSLISCE*GPVVV*VCLRNATPG*DRDDRVLLQRLWTDFQ
DEPTRPSAGLAVSTPSQGHHQNGM*P*PGEWLKELS**MQCGQRREDGGQPRHRWDELRQ
LHGGEAHATPKHDHERAQSYRASRSYAMEYRPCAAVAGVGGERIWPSRRQHLVIPEHRWE
GTVQDDQGRLPEAHPQLQRRHPSLTSPLPQRDSSSTFDFR*C**SLTKLSTVNAC*KHRG
CSFYFPKYFSIS*SYAKNYN*ASLLQIKQQNQCQKAAFPYMGTSAKHMSSLP*RSKSKRN
GEGVEMISEN*M*NIFLLEV*CSIINKGHIAKIKKKKKKX
>EU432099.1_2 Homo sapiens TMPRSS2-ERG prostate cancer specific isoform 1 (ERG) mRNA, complete cds, alternatively spliced
QEAEAEAEGEGRGAPPGARQEALSVVSEDQSLFECAYGTPHLAKTEMTASSSSDYGQTSK
MSPRVPQQDWLSQPPARVTIKMECNPSQVNGSRNSPDECSVAKGGKMVGSPDTVGMNYGS
YMEEKHMPPPNMTTNERRVIVPADPTLWSTDHVRQWLEWAVKEYGLPDVNILLFQNIDGK
ELCKMTKDDFQRLTPSYNADILLSHLHYLRETPLPHLTSDDVDKALQNSPRLMHARNTGG
AAFIFPNTSVYPEATQRITTRPVSYR*NNRTSARKQPSLTWALLPSI*VHCLEDQSQREM
ERVLK*SAKIKCKIYSYWKSDALLSIKDT*QR*KKKKKKX
>EU432099.1_3 Homo sapiens TMPRSS2-ERG prostate cancer specific isoform 1 (ERG) mRNA, complete cds, alternatively spliced
RRRRRRRRARGGERRLERGRKPYQL*VRTSRCLSVPTERHTWLRQR*PRPPPATMDRLPR
*AHASLSRIGCLNPQPGSPSKWNVTLAR*MAQGTLLMNAVWPKAGRWWAAQTPLG*TTAA
TWRRSTCHPQT*PRTSAELSCQQILRYGVQTMCGSGWSGR*KNMAFQTSTSCYSRTSMGR
NCAR*PRTTSRGSPPATTPTSFSHISTTSERLLFHI*LQMMLIKPYKTLHG*CMLETQGV
QLLFSQILQYILKLRKELQLGQSLTDKTTEPVPESSLPLHGHFCQAYEFIALKIKVKEKW
RGC*NDQRKLNVKYILIGSLMLYYQ*RTHSKDKKKKKKK
>EU432099.1_4 Homo sapiens TMPRSS2-ERG prostate cancer specific isoform 1 (ERG) mRNA, complete cds, alternatively spliced
FFFFFFYLCYVSFIDNRASDFQ*EYILHLIFADHFNTLSISL*L*SSRQ*THMLGRSAHV
REGCFLALVLLFYL*ETGLVVILCVASGYTEVFGKIKAAPPVFLACINRGEFCKALSTSS
EVKCGRGVSLR*WRCERRMSAL*LGVSLWKSSLVILHSSFPSMFWNNKMLTSGRPYSFTA
HSSHCRTWSVLHSVGSAGTITLRSFVVMFGGGMCFSSM*LP*FIPTVSGLPTIFPPLATL
HSSGEFLEPFTWLGLHSILMVTLAGG*DSQSC*GTRGLILEVCP*SLEEDAVISVLARCG
VP*AHSNNDWSSLTTDKASCRAPGGAPRPSPSASASASC
>EU432099.1_5 Homo sapiens TMPRSS2-ERG prostate cancer specific isoform 1 (ERG) mRNA, complete cds, alternatively spliced
FFFFFFLSLLCVLY***SIRLPIRIYFTFNFR*SFQHPLHFSLTLIFKAMNSYAWQKCPC
KGRLLSGTGSVVLSVRDWPSCNSLRSFRIY*SIWENKSCTPCVSSMH*PWRVL*GFINII
*SQMWKRSLSEVVEM*EKDVGVVAGGEPLEVVLGHLAQFLPIDVLE*QDVDVWKAIFFHR
PLQPLPHMVCTP*RRICWHDNSALVRGHVWGWHVLLLHVAAVVHPNGVWAAHHLPAFGHT
AFIRRVP*AIHLARVTFHFDGDPGWGLRQPILLRDAWAHLGSLSIVAGGGRGHLCLSQVW
RSVGTLKQRLVLTHN**GFLPRSRRRSPPLALRLRLRLLX
>EU432099.1_6 Homo sapiens TMPRSS2-ERG prostate cancer specific isoform 1 (ERG) mRNA, complete cds, alternatively spliced
FFFFFFFIFAMCPLLIIEHQTSNKNIFYI*FSLIISTPSPFLFDFDLQGNELICLAEVPM
*GKAAFWHWFCCFICKRLA*L*FFA*LQDILKYLGK*KLHPLCF*HALTVESFVRLYQHH
LKSNVEEESL*GSGDVREGCRRCSWG*ASGSRPWSSCTVPSHRCSGITRC*RLEGHILSP
PTPATAAHGLYSIA*DLLAR*LCARSWSCLGVACASPPCSCRSSSQRCLGCPPSSRLWPH
CIHQESSLSHSPG*GYIPF*W*PWLGVETANPAEGRVGSSWKSVHSRWRRTRSSLS*PGV
AFRRHTQTTTGPHSQLIRLPAALQAALPAPRPPPPPPPPA
```

Sixth sequence:

```
>EU090248.1_1 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
RELSRRRRRRRRARGGERRLERGRLFQDLWRPEESRVDQKQDK*LTEKKDGRTKGN*SRQ
VLNSW*MGWLTEGHDSDCPGPSSSYQGTLLMNAVWPK
>EU090248.1_2 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
AS*AGGGGGGGGRGAGSAAWSAAGYSRIFGDPRKAVLTKSKTNDSQRKKMAEPRATKAVR
F*TAGRWAGLLKDMIQTVPDPAAHIKELS**MQCGQX
>EU090248.1_3 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
RAKQEAEAEAEGEGRGAPPGARQVIPGSLETRGKPC*PKARQMTHREKRWQNQGQLKPSG
SEQLVDGLAY*RT*FRLSRTQQLISRNSPDECSVAKX
>EU090248.1_4 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
FWPHCIHQESSLI*AAGSGTV*IMSFSKPAHLPAVQNLTALVALGSAIFFLCESFVLLLV
NTAFLGSPKILE*PAALQAALPAPRPPPPPPPPA*LA
>EU090248.1_5 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
LATLHSSGEFLDMSCWVRDSLNHVLQ*ASPSTSCSEPDGFSCPWFCHLFSL*VICLAFGQ
HGFPRVSKDPGITCRAPGGAPRPSPSASASASCLARX
>EU090248.1_6 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
FGHTAFIRRVP*YELLGPGQSESCPSVSQPIYQLFRT*RL*LPLVLPSFFSVSHLSCFWS
TRLSSGLQRSWNNLPRSRRRSPPLALRLRLRLLLSSR
```

What are 6 Frames?

Reading Frames:

- A reading frame is a way of dividing a nucleotide sequence into a set of consecutive, non-overlapping codons.
- There are three possible reading frames in the forward direction and three in the reverse direction.

Forward Reading Frames:

- **Frame 1**: Starts from the first nucleotide.
- **Frame 2**: Starts from the second nucleotide.
- **Frame 3**: Starts from the third nucleotide.

Reverse Reading Frames:

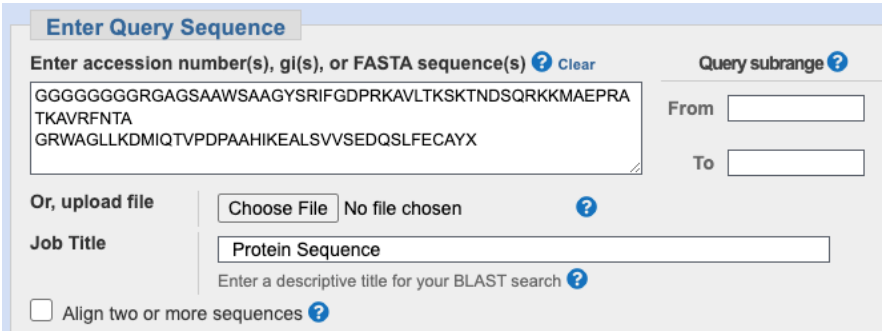
- **Frame -1**: Starts from the first nucleotide of the reverse complement.
- **Frame -2**: Starts from the second nucleotide of the reverse complement.
- **Frame -3**: Starts from the third nucleotide of the reverse complement.

In order to find the correct frame of chimeric proteins we needed to do some changes in our protein sequence.
One change that we did was because we had a lot of end codons in our sequences, we first choose the sequences with the least amount of end codons.
Then we changed the end codons * in the sequences into amino acid N which stands for Asparagine. We did it because changing the stop codon to an amino acid like asparagine can restore the full-length, functional protein.

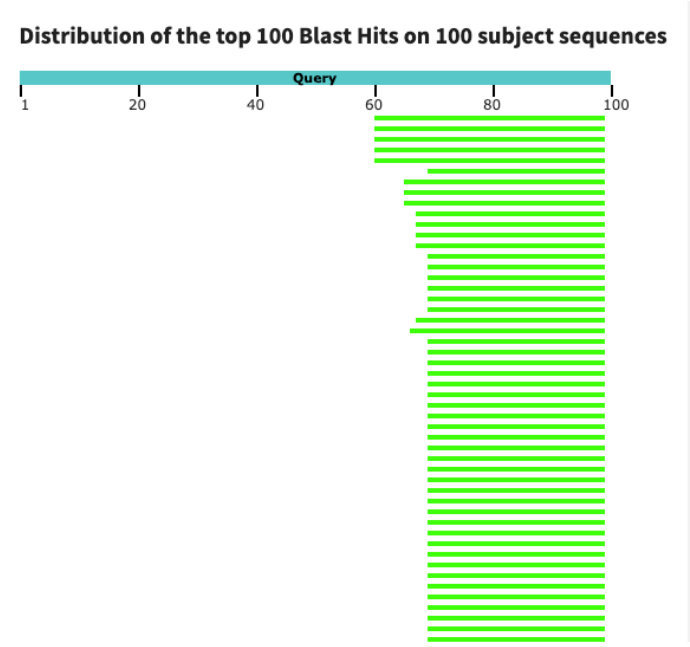
To find the correct protein sequence we used BLAST.
We had 3 different frames (from 3 different sequences) that had only 1 stop codon, so we checked all 3 but only 1 fitted what we were looking for. This frame is:

```
>DQ831522.1_1 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
GGGGGGGGRGAGSAAWSAAGYSRIFGDPRKAVLTKSKTNDSQRKKMAEPRATKAVRFNTA
GRWAGLLKDMIQTVPDPAAHIKEALSVVSEDQSLFECAYX
```

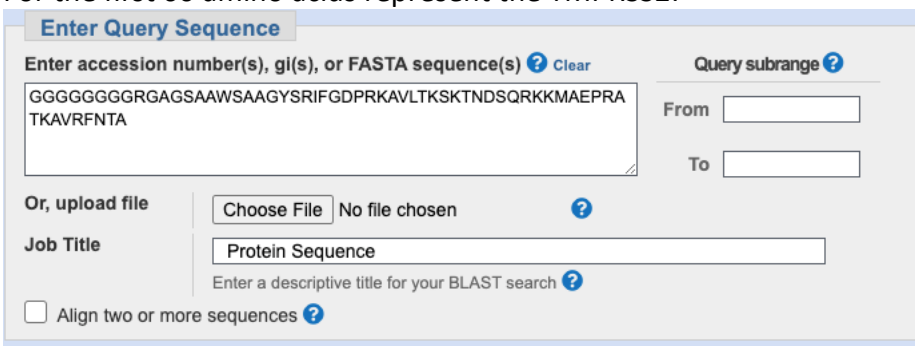
We passed the “fixed” protein sequence into BLAST:



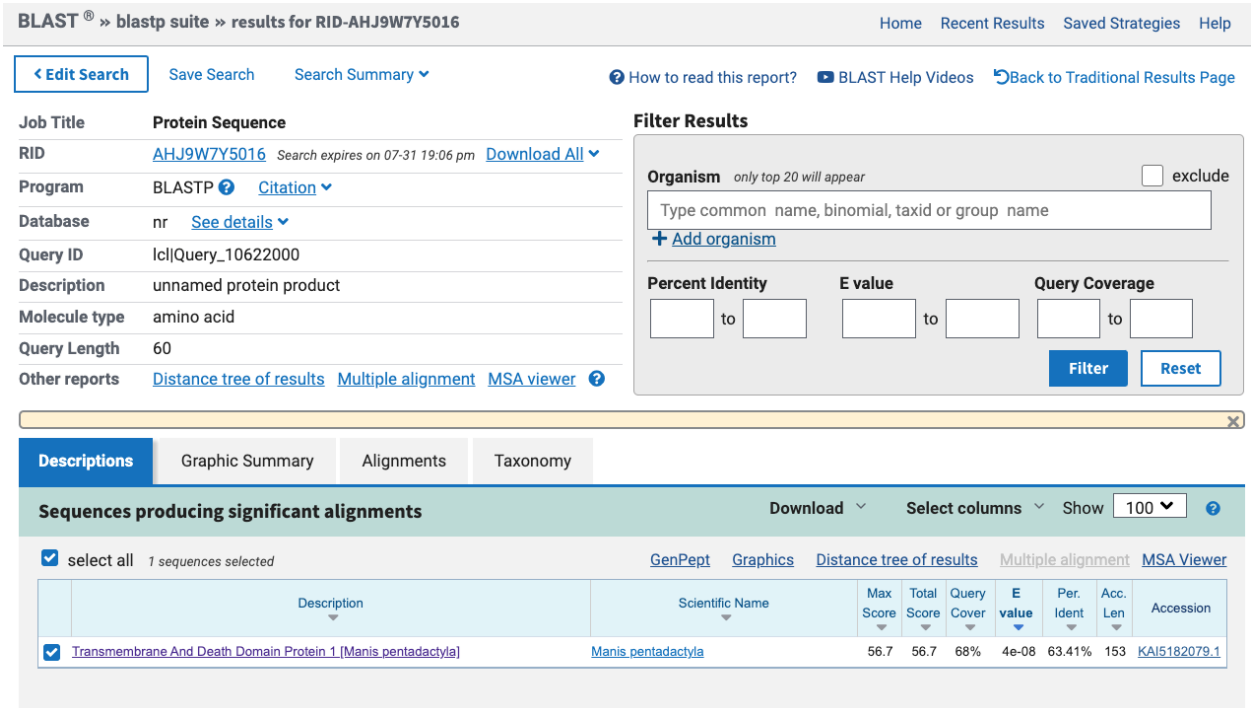
And received the following graphic summary:



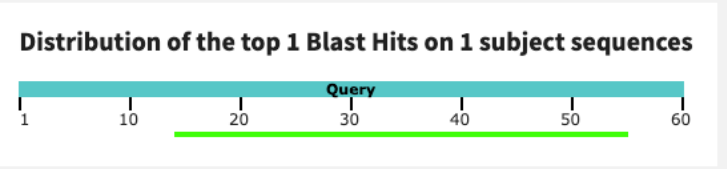
We understand from here that a part of the sequence is the ERG and part of it is the TMPRSS2
From the above image we understand that the first 60 amino acids represent the TMPRSS2 and the last represent ERG.
So, we checked them separately.
For the first 60 amino acids represent the TMPRSS2:



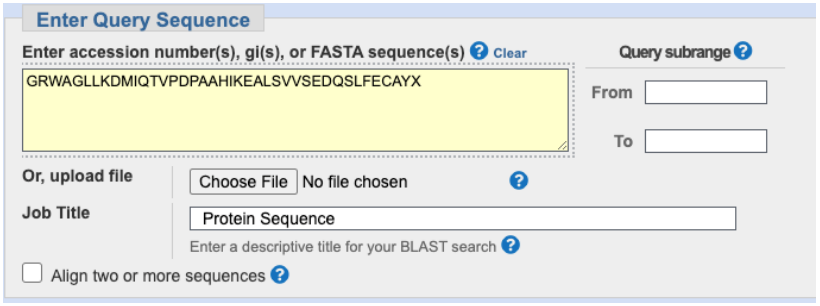
This is the BLAST results:



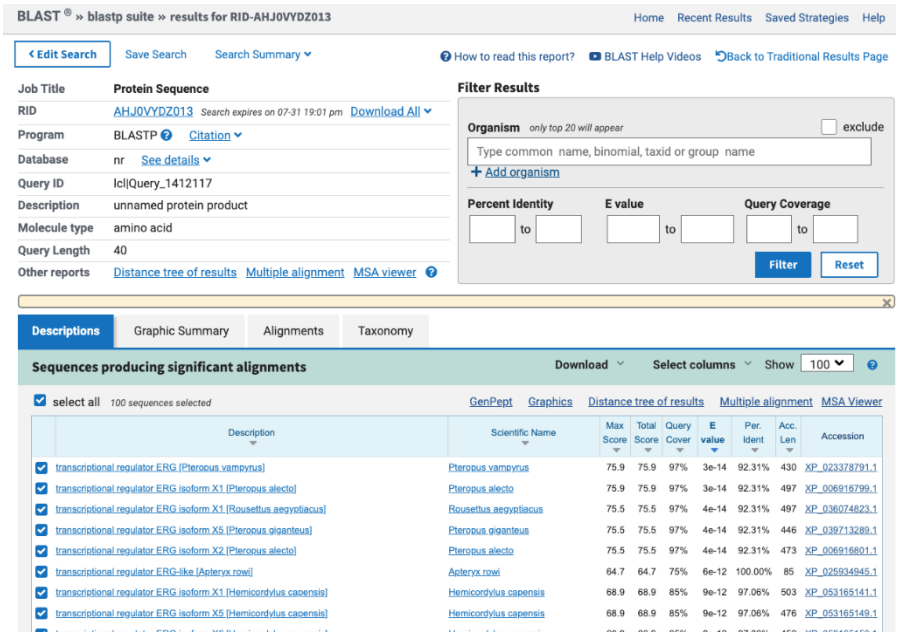
And this is the graphic summary:



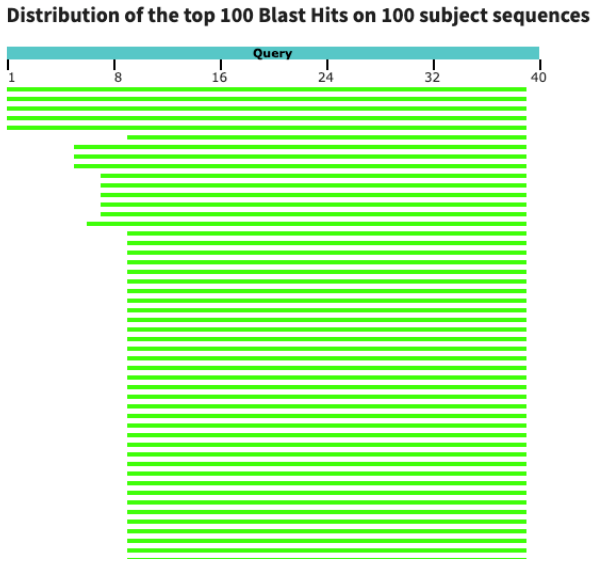
For the ERG part:



This is the BLAST results:



And this is the graphic summary:



Regarding the “start codon”:

First sequence:
>DQ831522.1 Homo sapiens TMPRSS2/ERG fusion mRNA, partial sequence
GGAGGCGGAGGCGGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAGGTTATTCC
AGGATCTTTGGAGACCCGAGGAAAGCCGTGTTGACCAAAAGCAAGACAAATGACTCACAGAGAAAAAA
GATGGCAGAACCAAGGGCAACTAAAGCCGTCAGGTTCTGAACAGCTGGTAGATGGGCTGGCTTACTGA
AGGACATGATTCAGACTGTCCCGGACCCAGCAGCTCATATCAAGGAAGCCTTATCAGTTGTGAGTGAGG
ACCAGTCGTTGTTTGAGTGTGCCTACA

The provided sequence includes the correct start codon "ATG" (which corresponds to "AUG" in RNA) at several positions, one of them is position 88-90. This indicates that translation could potentially initiate at this site, leading to the production of a protein.

Question 4:
Optional (bonus – 5 points): write a program in Python for transeq.

Answer 4:

```
# Auxiliary function
# This function takes a DNA sequence and returns its reverse complement.
def reverse_complement(seq):
    complement = {'A': 'T', 'C': 'G', 'G': 'C', 'T': 'A'} # DNA dictionary
    return ''.join(complement.get(base, base) for base in reversed(seq))

# This function translates the DNA sequence into a protein sequence
def translate(seq):
    codon_table = {
        'ATA':'I', 'ATC':'I', 'ATT':'I', 'ATG':'M',
        'ACA':'T', 'ACC':'T', 'ACG':'T', 'ACT':'T',
        'AAC':'N', 'AAT':'N', 'AAA':'K', 'AAG':'K',
        'AGC':'S', 'AGT':'S', 'AGA':'R', 'AGG':'R',
        'CTA':'L', 'CTC':'L', 'CTG':'L', 'CTT':'L',
        'CCA':'P', 'CCC':'P', 'CCG':'P', 'CCT':'P',
        'CAC':'H', 'CAT':'H', 'CAA':'Q', 'CAG':'Q',
        'CGA':'R', 'CGC':'R', 'CGG':'R', 'CGT':'R',
        'GTA':'V', 'GTC':'V', 'GTG':'V', 'GTT':'V',
        'GCA':'A', 'GCC':'A', 'GCG':'A', 'GCT':'A',
        'GAC':'D', 'GAT':'D', 'GAA':'E', 'GAG':'E',
        'GGA':'G', 'GGC':'G', 'GGG':'G', 'GGT':'G',
        'TCA':'S', 'TCC':'S', 'TCG':'S', 'TCT':'S',
        'TTC':'F', 'TTT':'F', 'TTA':'L', 'TTG':'L',
        'TAC':'Y', 'TAT':'Y', 'TAA':'*', 'TAG':'*',
        'TGC':'C', 'TGT':'C', 'TGA':'*', 'TGG':'W', }
    protein = ""
    for i in range(0, len(seq) - 2, 3):
        codon = seq[i:i+3]
        if len(codon) == 3:
            amino_acid = codon_table.get(codon, 'X')  # 'X' for unknown codons
            protein += amino_acid
    return protein

# This function translates the input sequence in all six reading frames.
def transeq(seq):
    seq = seq.upper()
    rev_seq = reverse_complement(seq)
    translations = []
    for i in range(3):
        translations.append(translate(seq[i:]))
        translations.append(translate(rev_seq[i:]))

    return translations
```

EXAMPLE USAGE IN JUPYTER NOTEBOOK:

```python
dna_sequence = "GGAGGCGGAGGCGGAGGCGGAGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAGGTTATTCCAGGATCTTTGGAGACCCGAGGAAAGCCGTGTTGACCAA

translations = transeq(dna_sequence)

for i, translation in enumerate(translations):
    frame = f"{'Forward' if i < 3 else 'Reverse'} Frame {i % 3 + 1}"
    print(f"{frame}:")
    print(translation)
    print()
```

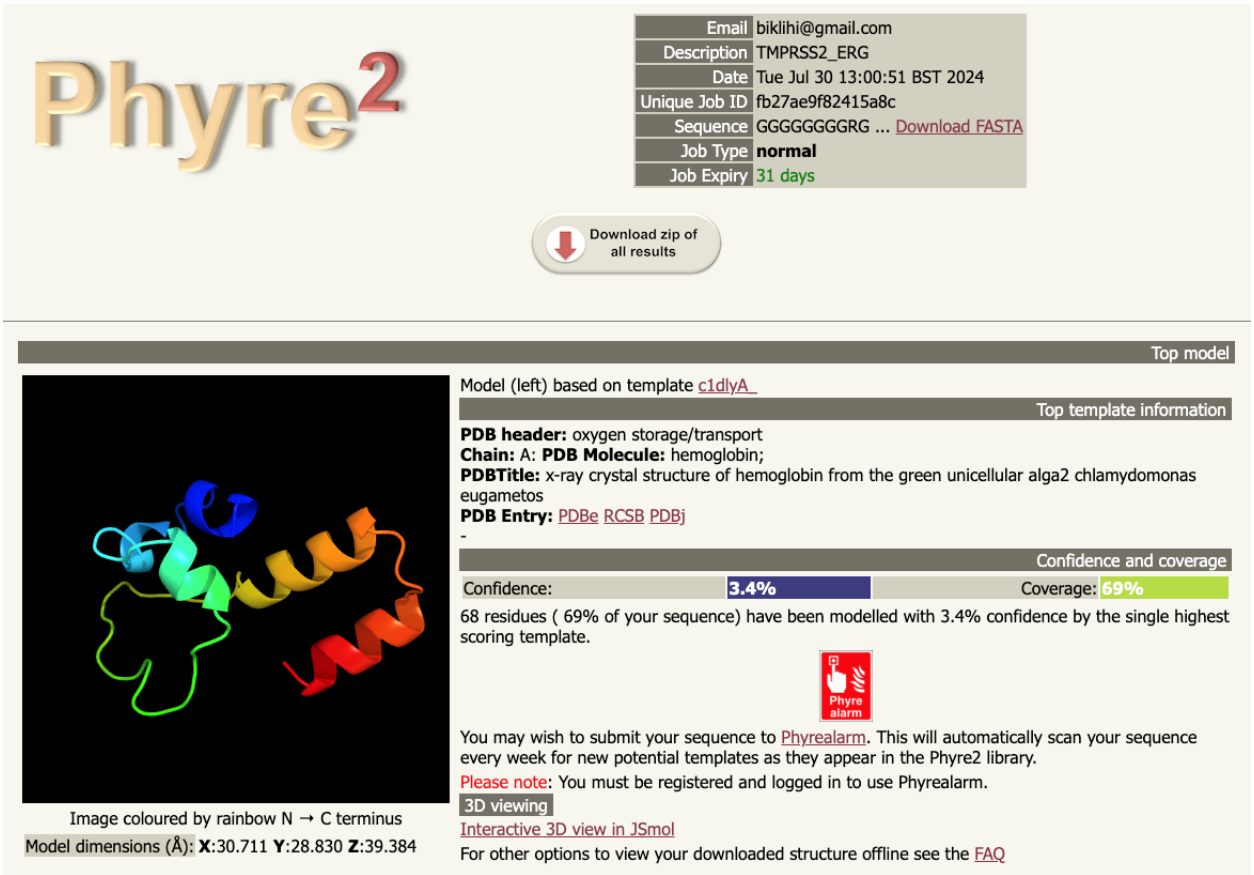
```
Forward Frame 1:
GGGGGGGGRGAGSAAWSAAGYSRIFGDPRKAVLTKSKTNDSQRKKMAEPRATKAVRF*TAGRWAGLLKDMIQTVPDPAAHIKEALSVVSEDQSLFECAY

Forward Frame 2:
CRHTQTTTGPHSQLIRLP*YELLGPGQSESCPSVSQPIYQLFRT*RL*LPLVLPSFFSVSHLSCFWSTRLSSGLQRSWNNLPRSRRRSPPLALRLRLRL

Forward Frame 3:
EAEAEAEGEGRGAPPGARQVIPGSLETRGKPC*PKARQMTHREKRWQNQGQLKPSGSEQLVDGLAY*RT*FRLSRTQQLISRKPYQL*VRTSRCLSVPT

Reverse Frame 1:
VGTLKQRLVLTHN**GFLDMSCWVRDSLNHVLQ*ASPSTSCSEPDGFSCPWFCHLFSL*VICLAFGQHGFPRVSKDPGITCRAPGGAPRPSPSASASAS

Reverse Frame 2:
RRRRRRRARGGERRLERGRLFQDLWRPEESRVDQKQDK*LTEKKDGRTKGN*SRQVLNSW*MGWLTEGHDSDCPGPSSSYQGSLISCE*GPVVV*VCL

Reverse Frame 3:
*AHSNNDWSSLTTDKASLI*AAGSGTV*IMSFSKPAHLPAVQNLTALVALGSAIFFLCESFVLLLVNTAFLGSPKILE*PAALQAALPAPRPPPPPPP
```

Question 5:

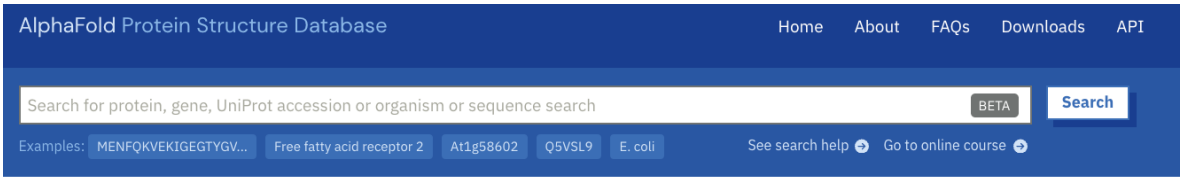
Predict the 3D protein structure of the ERG-TMPRSS2 chimeric protein. Make “print screen” of the results. Explain the obtained results, homologous proteins, and their function.

Answer 5:

In order to predict the 3D protein structure of the ERG-TMPRSS2 chimeric protein via PYMOL we needed first to download the PDB file. We tried to do it using Phyre2 as we studied in class, but I didn’t receive the exact chimera:

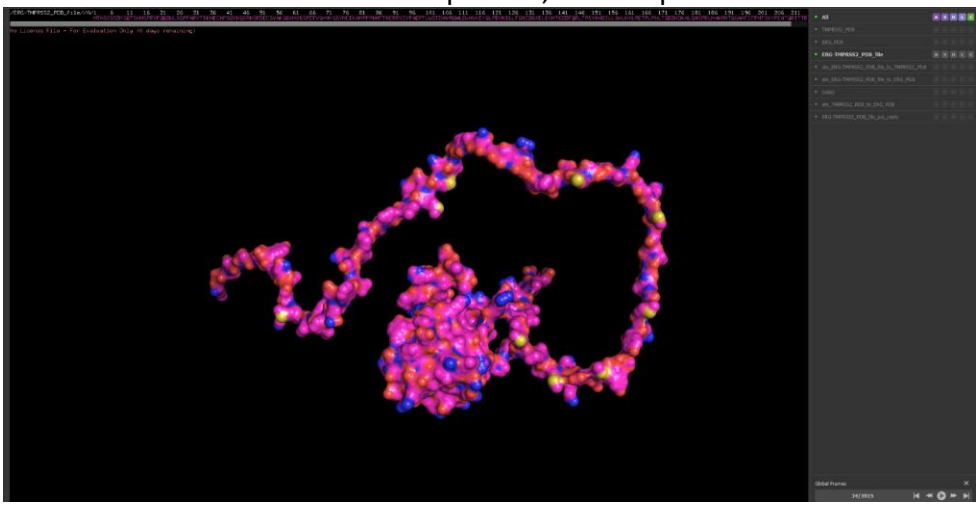


So instead, we used the AlphaFold Protein Structure database which gave use the exact results that we needed:

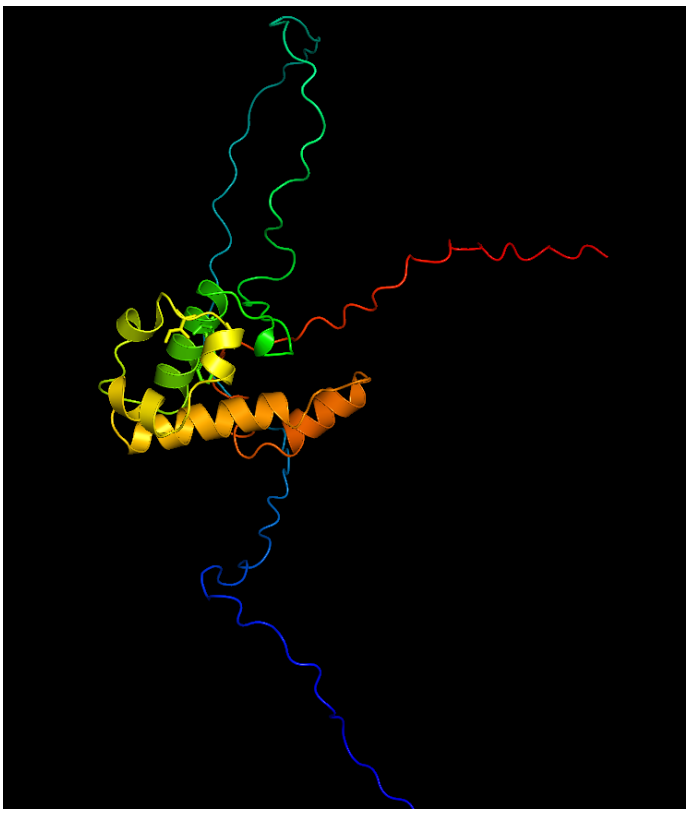


After we received the PDB file we uploaded it to PYMPL and received the following 3D structure:

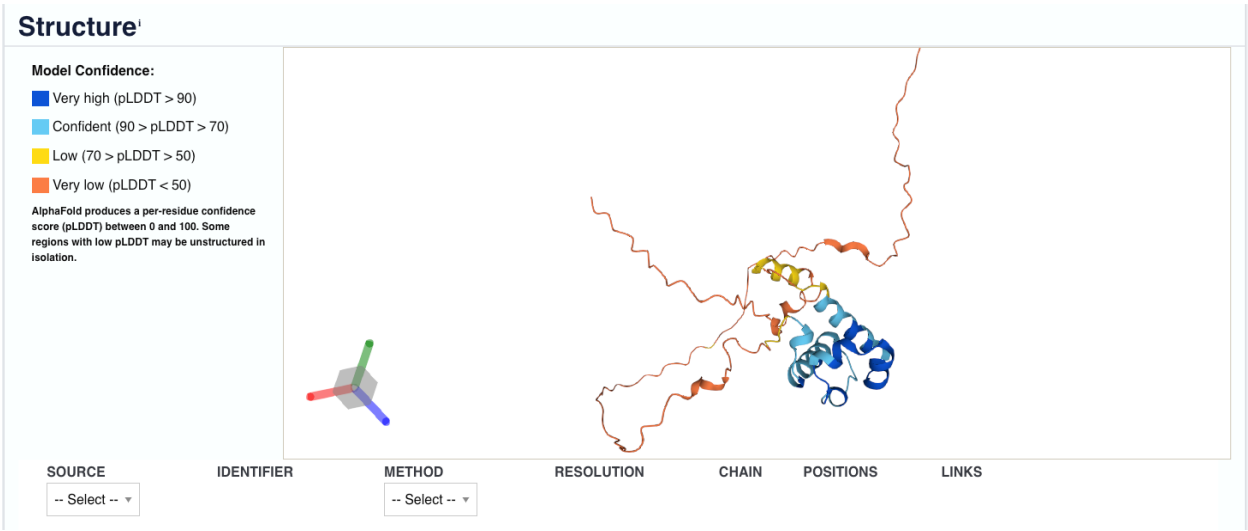
we took the PDB file of the fusion protein, and we present it in PYMOL in dots+ surface mode.



We also added the cartoon mode:



In order to check the obtained results, we compared the results that we got to results from UniProt for the same **TMPRSS2-ERG** chimera



https://www.uniprot.org/uniprotkb/B2Y833/entry#structure

we can see that the results are basically the same because they also used AlphaFolds.
Regarding homologous proteins and their function:

Homologous proteins are proteins that share a common evolutionary origin, often reflected in their sequence or structural similarity. When discussing the ERG-TMPRSS2 chimeric protein, homologous proteins would refer to proteins that are evolutionarily related to either ERG, TMPRSS2, or both.

There are 2 types of Homologous Proteins:

- **Orthologs**: Homologous proteins in different species that originated from a common ancestral gene through speciation. They often retain similar functions.
- **Paralogs**: Homologous proteins within the same species that originated from gene duplication. They may evolve new functions.

Homologous Proteins in the Context of ERG-TMPRSS2 Chimeric Protein:

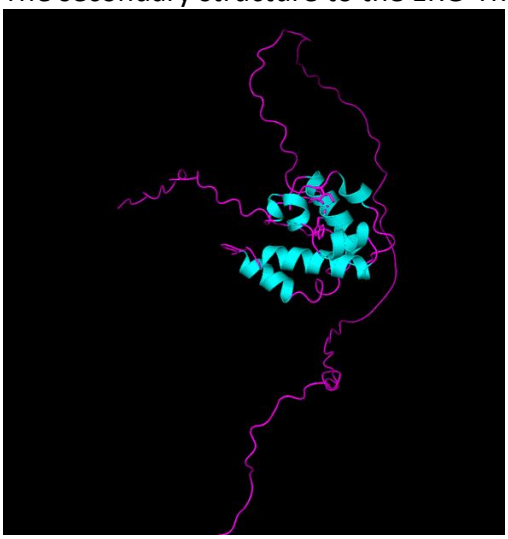
1. **ERG (ETS-related gene)**:
    - ERG is a member of the ETS (E26 transformation-specific) family of transcription factors.
    - Homologous proteins to ERG would include other ETS family members that share sequence and functional similarities, such as ETV1, ETV4, and ETV5.
2. **TMPRSS2 (Transmembrane Protease, Serine 2)**:
    - TMPRSS2 is a serine protease.
    - Homologous proteins to TMPRSS2 would include other serine proteases that share sequence motifs and structural features essential for protease activity.
3. **ERG-TMPRSS2 Chimeric Protein**:
    - The ERG-TMPRSS2 chimeric protein is a fusion protein resulting from a gene fusion event, often observed in prostate cancer.
    - Homologous proteins to the ERG-TMPRSS2 chimeric protein could be other fusion proteins involving ERG or TMPRSS2 or proteins with significant sequence or structural similarity to the domains contributed by ERG or TMPRSS2.

Question 6:

Explain the secondary structure of the ERG-TMPRSS2 chimeric protein according to the 3D structure prediction and classes in SCOP.

Answer 6:

Secondary protein structure describes localized conformation of the chain.
The secondary structure to the ERG-TMPRSS2 is:



Alpha Helices:

- **Description**: Alpha helices are right-handed coiled structures stabilized by hydrogen bonds between the carbonyl oxygen of one amino acid and the amide hydrogen of another amino acid four residues away.
- **Observation**: In the image, alpha helices are typically represented by the spiral or helical structures colored in light blue, we see 7 of them.

Question 7:

Characterize the potential function of ERG-TMPRSS2 chimeric protein. Explain your results based on the parental proteins.

Answer 7:

The ERG-TMPRSS2 chimeric protein is a fusion of two proteins found in prostate cancer. It combines parts of ERG, a gene-controlling protein, with TMPRSS2, a protein that responds to male hormones.

This fusion protein likely works as follows:

1. It can turn genes on or off, like ERG normally does. These genes control how cells grow, specialize, and form blood vessels.
2. It becomes active when male hormones are present, due to the TMPRSS2 part.
3. It makes cells grow and divide more than they should.
4. It helps cancer cells move and spread to other parts of the body.
5. It stops prostate cells from developing normally.

The ERG part of the fusion protein keeps its ability to control genes but loses its usual controls. The TMPRSS2 part mainly contributes its response to male hormones.

This combination leads to too much ERG protein being made in prostate cells when male hormones are present. This excess ERG then affects many genes and cell processes. The result is that cells grow out of control and act abnormally, which can lead to cancer.

In simple terms, this fusion protein takes a powerful gene controller (ERG) and makes it overactive in prostate cells, driving the development and worsening of prostate cancer.

Question 8:

Find function of the ERG-TMPRSS2 chimeric protein in cancers. Explain your results and give 10 references from PubMed supporting your results.

Answer 8:

The ERG-TMPRSS2 chimeric protein plays a crucial role in cancer, particularly in prostate cancer. This fusion protein primarily functions as an oncogenic driver by promoting uncontrolled cell proliferation and altering gene expression patterns. It enhances cancer cell invasion and migration, facilitating metastasis. The chimeric protein disrupts normal cellular differentiation, keeping cells in an immature, rapidly dividing state. It also promotes angiogenesis, supporting tumor growth through increased blood supply. The fusion confers androgen sensitivity to cancer cells, making them more responsive to male hormones. Additionally, it contributes to genomic instability, alters cellular metabolism to support rapid growth, and helps cancer cells evade normal cell death mechanisms. The ERG-TMPRSS2 protein also modulates the tumor microenvironment, creating conditions favorable for cancer progression. These combined effects make the ERG-TMPRSS2 fusion a significant factor in the initiation, progression, and metastasis of prostate cancer, occurring in approximately 50% of cases.

10 supporting references from PubMed:

1. Tomlins SA, et al. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science, 310(5748), 644-648.
2. Carver BS, et al. (2009). ETS rearrangements and prostate cancer initiation. Nature, 457(7231), E1.
3. Yu J, et al. (2010). An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. Cancer Cell, 17(5), 443-454.
4. Kron KJ, et al. (2017). TMPRSS2-ERG fusion co-opts master transcription factors and activates NOTCH signaling in primary prostate cancer. Nature Genetics, 49(9), 1336-1345.
5. Adamo P, Ladomery MR. (2016). The oncogene ERG: a key factor in prostate cancer. Oncogene, 35(4), 403-414.
6. Bose R, et al. (2017). ERF mutations reveal a balance of ETS factors controlling prostate oncogenesis. Nature, 546(7660), 671-675.

7. Nhili R, et al. (2016). Targeting the DNA-binding activity of the human ERG transcription factor using new heterocyclic dithiophene diamidines. Nucleic Acids Research, 44(8), 3578-3590.
8. Mani RS, et al. (2016). TMPRSS2-ERG-mediated feed-forward regulation of wild-type ERG in human prostate cancers. Cancer Research, 76(21), 6227-6236.
9. Deplus R, et al. (2017). TET2 and TET3 regulate GlcNAcylation and H3K4 methylation through OGT and SET1/COMPASS. EMBO Journal, 36(15), 2233-2250.
10. Humphrey PA. (2012). Histological variants of prostatic carcinoma and their significance. Histopathology, 60(1), 59-74.

Question 9:

Predict the RNA fold of at least 6 RNA sequences of ERG-TMPRSS2 chimeric protein using the dedicated software. What is the minimal energy function? What is the optimal RNA fold? Explain your results and print the screen.

Answer 9:

Here we took the 6 sequences we have from the FASTA file and we used RNAfold WebServer: (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi)

Results for minimum free energy prediction

The optimal secondary structure in dot-bracket notation with a minimum free energy of **-91.00** kcal/mol is given below.
[color by base-pairing probability | color by positional entropy | no coloring]

```
1      CGCGAGCUAAGCAGGAGGCGGAGGCGGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCUGGAGCGCGGCAGGUUAUUCCAGGAUCUUUGGAGACCCGAGGAAAGCCGUGUUGACCAAAAGCAAGACAAAUGACUCACAGAGAAAAAAGAUGGCAGAACCA
160    AGGGCAACUAAAGCCGUCAGGUUCUGAACAGCUGGUAGAUGGGCUGGCUUACUGAAGGACAUGAUUCAGACUGUCCCGGACCCAGCAGCUCAUAUCAAGGAACUCUCCUGAUGAAUGCAGUGUGGCCAAAA
```

Results for minimum free energy prediction

The optimal secondary structure in dot-bracket notation with a minimum free energy of **-289.30** kcal/mol is given below.
[color by base-pairing probability | color by positional entropy | no coloring]

```
1      GCAGGAGGCGGAGGCGGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCUGGAGCGCGGCAGGAAGCCUUAUCAGUUGUGAGUGAGGACCAGUCGUUGUUUGAGUGUGCCUACGGAACGCCACACCUGGCUAAGACAGAGAUGACCGCGUCCUCCUCCAGC
160    GACUAUGGACAGACUUCCAAGAUGAGCCCACGCGUCCCUCAGCAGGAUUGGCUGUCUCAACCCCCAGCCAGGGUCACCAUCAAAAUGGAAUGUAACCCUAGCCAGGUGAAUGGCUCAAGGAACUCUCCUGAUGAAUGCAGUGUGGCCAAAGGCGGGAAGA
320    UGGUGGGCAGCCCAGACACCGUUGGGAUGAACUACGGCAGCUACAUGGAGGAGAAGCACAUGCCACCCCCAAACAUGACCACGAACGAGCGCAGAGUUAUCGUGCCAGCAGAUCCUACGCUAUGGAGUACAGACCAUGUGCGGCAGUGGCUGGAGUGGGC
480    GGUGAAAGAAUAUGGCCUUCCAGACGUCAACAUCUUGUUAUUCCAGAACAUCGAUGGGAAGGAACUGUGCAAGAUGACCAAGGACGACUUCCAGAGGCUCACCCCCAGCUACAACGCCGACAUCCUUCUCUCACAUCUCCACUACCUCAGAGAGACUCCU
640    CUUCCACAUUUGACUUCAGAUGAUGUUGAUAAAGCCUUACAAAACUCUCCACGGUUAAUGCAUGCUAGAAACACAGGGGGUGCAGCUUUUAUUUUCCCAAAUACUUCAGUAUAUCCUGAAGCUACGCAAAGAAUUACAACUAGGCCAGUCUCUUACAGAU
800    AAAACAACAGAACCAGUGCCAGAAAGCAGCCUUCCCUUACAUGGGCACUUCUGCCAAGCAUAUGAGUUCAUUGCCUUGAAGAUCAAAGUCAAAGAGAAAUGGAGAGGGUGUUGAAAUGAUCAGCGAAAAUUAAAUGUAAAAUAUAUUCUUAUUGGAAGUC
960    UGAUGCUCUAUUAUCAAUAAAGGACACAUAGCAAAGAUAAAAAAAAAAAAAAAAAAAAAAA
```

Results for minimum free energy prediction

The optimal secondary structure in dot-bracket notation with a minimum free energy of **-85.90** kcal/mol is given below.
[color by base-pairing probability | color by positional entropy | no coloring]

```
1      GGAGUAGGCGCGAGCUAAGCAGGAGGCGGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCUGGAGCGCGGCAGGAAGCCUUAUCAGUUGUGAGUGAGGACCAGUCGUUGUUUGAGUGUGCCUACGGAACGCCACACCUGGCUAAGACAGAGAUG
160    ACCGCGUCCUCCUCCAGCGACUAUGGACAGACUUCCAAGAUGAGCCCACGCGUCCCUCAGCAGGAUUGGCUGUCU
```

Results for minimum free energy prediction

The optimal secondary structure in dot-bracket notation with a minimum free energy of **-213.70** kcal/mol is given below.
[color by base-pairing probability | color by positional entropy | no coloring]

```
1      UUCGCUUGCUGUUUCUGUGACUUUACGCUCUGCUGACCUAGAGGUCUUAGUUCCGGAGGGAGGAAUGCUGCCACCAGGAGACACAACAAUGAUUCAAUUAAACUAGAAUUUACGACUGCCACCUGGCCACGCUGAGCUCCACAUGCCUCUGAAUCAAAAG
160    GCAAAGAGAGAGUUAUGCAUUGGCUGGGGAGACCCAUCUGGACUACCAAGGAGAAGCUAUAGACUACUUCUACUCCACCAGGAAGGAAGCCUUAUCAGUUGUGAGUGAGGACCAGUCGUUGUUUGAGUGUGCCUACGGAACGCCACACCUGGCUAAGACA
320    GAGAUGACCGCGUCCUCCUCCAGCGACUAUGGACAGACUUCCAAGAUGAGCCCACGCGUCCCUCAGCAGGAUUGGCUGUCUCAACCCCCAGCCAGGGUCACCAUCAAAAUGGAAUGUAACCCUAGCCAGGUGAAUGGCUCAAGGAACUCUCCUGAUGAAU
480    GCAGUGUGGCCAAAGGCGGGAAGAUGGUGGGCAGCCCAGACACCGUUGGGAUGAACUACGGCAGCUACAUGGAGGAGAAGCACAUGCCACCCCCAAACAUGACCACGAACGAGCGCAGAGUUAUCGUGCCAGCAGAUCCUACGCUAUGGAGUACAGACCA
640    UGUGCGGCAGUGGCUGGAGUGGGCGGUGAAA
```

Results for minimum free energy prediction

The optimal secondary structure in dot-bracket notation with a minimum free energy of **-223.10** kcal/mol is given below.

[color by base-pairing probability | color by positional entropy | no coloring]

```
1      GCGGAGUGCGAGGGGCGGGGACGCCGCCUGGACGCGGCAGCCGUCAGGUUCUGAACAGCUGGUAGAUGGGCUGGCUUACUGAAGGACAUGAUUCAGACUGUCCCGGACCCAGCAGCUCAUAUCAAGGAAGCCUUAUCAGUUGUGAGUGAGGACCAGUCGU
160    UGUUUGAGUGUGCCUACGGAACGCCACACCUGGCUAAGACAGAGAUGACCGCGUCCUCCUCCAGCGACUAUGGACAGACUUCCAAGAUGAGCCCACGCGUCCCUCAGCAGGAUUGGCUGUCUCAACCCCCAGCCAGGGUCACCAUCAAAAUGGAAUGUAA
320    CCCUAGCCAGGUGAAUGGCUCAAGGAACUCUCCUGAUGAAUGCAGUGUGGCCAAAGGCGGGAAGAUGGUGGGCAGCCCAGACACCGUUGGGAUGAACUACGGCAGCUACAUGGAGGAGAAGCACAUGCCACCCCCAAACAUGACCACGAACGAGCGCAGA
480    GUUAUCGUGCCAGCAGAUCCUACGCUAUGGAGUACAGACCAUGUGCGGCAGUGGCUGGAGUGGGCGGUGAAAGAAUAUGGCCUUCCAGACGUCAACAUCUUGUUAUUCCAGAACAUCGAUGGGAAGGAACUGUGCAAGAUGACCAAGGACGACUUCCAGA
640    GGCUCACCCCCAGCUACAACGCC
```

Results for minimum free energy prediction

The optimal secondary structure in dot-bracket notation with a minimum free energy of **-97.90** kcal/mol is given below.

[color by base-pairing probability | color by positional entropy | no coloring]

```
1      GGAGGCGGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCUGGAGCGCGGCAGGUUAUUCCAGGAUCUUUGGAGACCCGAGGAAAGCCGUGUUGACCAAAAGCAAGACAAAUGACUCACAGAGAAAAAAGAUGGCAGAACCAAGGGCAACUAAAG
160    CCGUCAGGUUCUGAACAGCUGGUAGAUGGGCUGGCUUACUGAAGGACAUGAUUCAGACUGUCCCGGACCCAGCAGCUCAUAUCAAGGAAGCCUUAUCAGUUGUGAGUGAGGACCAGUCGUUGUUUGAGUGUGCCUACA
```

To determine the optimal RNA fold among these 6 sequences, we need to consider that the more negative the free energy, the more stable and likely the RNA structure is. Based on the minimum free energy values we got, the optimal RNA fold would be the one with the lowest (most negative) free energy, which in our case is -289.30 kcal/mol:

>EU432099.1 Homo sapiens TMPRSS2-ERG prostate cancer specific isoform 1 (ERG) mRNA.

GCAGGAGGCGGAGGCGGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAGGAAG
CCTTATCAGTTGTGAGTGAGGACCAGTCGTTGTTTGAGTGTGCCTACGGAACGCCACACCTGGCTAAGA
CAGAGATGACCGCGTCCTCCTCCAGCGACTATGGACAGACTTCCAAGATGAGCCCACGCGTCCCTCAGC
AGGATTGGCTGTCTCAACCCCCAGCCAGGGTCACCATCAAAATGGAATGTAACCCTAGCCAGGTGAATG
GCTCAAGGAACTCTCCTGATGAATGCAGTGTGGCCAAAGGCGGGAAGATGGTGGGCAGCCCAGACACC
GTTGGGATGAACTACGGCAGCTACATGGAGGAGAAGCACATGCCACCCCCAAACATGACCACGAACGA
GCGCAGAGTTATCGTGCCAGCAGATCCTACGCTATGGAGTACAGACCATGTGCGGCAGTGGCTGGAGT
GGGCGGTGAAAGAATATGGCCTTCCAGACGTCAACATCTTGTTATTCCAGAACATCGATGGGAAGGAAC
TGTGCAAGATGACCAAGGACGACTTCCAGAGGCTCACCCCCAGCTACAACGCCGACATCCTTCTCTCACA
TCTCCACTACCTCAGAGAGACTCCTCTTCCACATTTGACTTCAGATGATGTTGATAAAGCCTTACAAAACT
CTCCACGGTTAATGCATGCTAGAAACACAGGGGGTGCAGCTTTTATTTTCCCAAATACTTCAGTATATCCT
GAAGCTACGCAAAGAATTACAACTAGGCCAGTCTCTTACAGATAAAACAACAGAACCAGTGCCAGAAAG
CAGCCTTCCCTTACATGGGCACTTCTGCCAAGCATATGAGTTCATTGCCTTGAAGATCAAAGTCAAAGAG
AAATGGAGAGGGTGTTGAAATGATCAGCGAAAATTAAATGTAAAATATATTCTTATTGGAAGTCTGATG
CTCTATTATCAATAAAGGACACATAGCAAAGATAAAAAAAAAAAAAAAAAAAAAA

Question 10:

Find folds, families, and superfamilies of parental proteins ERG and TMPRSS2 in the SCOP database, find corresponding reactions in the KEGG database. Explain the results – half a page

Answer 10:

Regarding ERG the results from SCOP:

Folds:

Search results for *ERG*

Folds [1] Superfamilies [9] Families [16]

- 2000375 **Amb V allergen**

FOLD

Amb V allergen

disulfide-rich, alpha+beta: 3 antiparallel strands followed by a short alpha helix

Superfamilies:

Search results for *ERG*

Folds [1] Superfamilies [9] Families [16]

- 3002819 **STAT1-TAD**
- 3002820 **STAT2-TAD**
- 3002482 **AZ1 domain-like**
- 3002185 **Blo t 5 dust mite allergen-like**
- 3000177 **Pollen allergen ole e 6**
- 3000548 **Amb V allergen**
- 3001008 **PHL pollen allergen**
- 3001033 **DmpA/ArgJ-like**
- 3001079 **Group V grass pollen allergen**

Families:

Search results for *ERG*

Folds [1] Superfamilies [9] **Families [16]**

- 4000576 **Early switch protein XOL-1, N-terminal domain**
- 4000807 **Apolipoprotein**
- 4001283 **HEAT repeat**
- 4002255 **phl pollen allergen**
- 4002306 **Group V grass pollen allergen**
- 4002520 **Divergent polysaccharide deacetylase**
- 4002702 **Pollen allergen PHL P 1 N-terminal domain**
- 4003023 **Pollen allergen ole e 6**
- 4003283 **Amb V allergen**
- 4003372 **Type III secretory system chaperone**
- 4003742 **Polcalcin**
- 4004163 **Jas motif-like**
- 4004702 **MJ0951-like (UPF0348)**
- 4005078 **Group 7 allergen-like**
- 4005314 **Blo t 5 dust mite allergen-like**
- 4007748 **Divergent PilZ domain**

Regarding TMPRSS2 the results from SCOP:
We tried to search it in SCOP but couldn't find this gene.

Search results for *TMPRSS2*

What did instead is for some proteins the structure is conserved therefore I added the print screen above because I didn't find TMPRSS2 in SCOP.
In the homology analysis we understand that TMPRSS2 is a serine protease, so we looked for other serine proteases in SCOP

ebi.ac.uk/pdbe/scop/term/4000286

SUPERFAMILY Trypsin-like serine proteases

FAMILY

Eukaryotic proteases

SCOP ID: 4000286

Function functionally relevant complex structure(s) determined

Show ancestry ☐

Domains [65 entries]	**ID**	**Region**	**Links**
Protein **Prothrombin** Species *Homo sapiens* **Representative domain 8029775** Represented structures [971]	**P00734** **1HAG**	328-622 E:1-247	UniProt PDBe RCSB PDB
Protein **Serine protease 1** Species *Bos taurus* **Representative domain 8025117** Represented structures [678]	**P00760** **2AYW**	24-246 A:16-245	UniProt PDBe RCSB PDB
Protein **Chymotrypsinogen A** Species *Bos taurus* **Representative domain 8029749** Screenshot structures [196]	**P00766** **2CGA**	1-245 A:1-245	UniProt PDBe RCSB PDB

We went to KEGG database and searched for the TMPRSS2 and ERG:

<u>TMPRSS2:</u>

Search KEGG **for** TMPRSS2 Go Clear

Database: KEGG - Search term: TMPRSS2

KEGG ORTHOLOGY

K09633
TMPRSS2; transmembrane protease serine 2 [EC:3.4.21.122]

KEGG GENES

hsa:7113
K09633 transmembrane protease serine 2 [EC:3.4.21.122] | (RefSeq) TMPRSS2, PRSS10; transmembrane serine protease 2
ptr:474007
K09633 transmembrane protease serine 2 [EC:3.4.21.122] | (RefSeq) TMPRSS2; transmembrane protease serine 2 isoform X1
pps:100988626
K09633 transmembrane protease serine 2 [EC:3.4.21.122] | (RefSeq) TMPRSS2; transmembrane protease serine 2 isoform X1
ggo:101134135
K09633 transmembrane protease serine 2 [EC:3.4.21.122] | (RefSeq) TMPRSS2; transmembrane protease serine 2 isoform X4
pon:100446360
K09633 transmembrane protease serine 2 [EC:3.4.21.122] | (RefSeq) TMPRSS2; transmembrane protease serine 2 isoform X1
• • • » display all

TMPRSS2 In KEGG is associated with pathways related to protein activation and cell surface remodeling. TMPRSS2 is known to be regulated by androgens, which is significant in the context of prostate cancer so we got a lot of result that are connected to prostate cancer:

ORTHOLOGY: K09633

Help

Entry	K09633 KO
Symbol	TMPRSS2
Name	transmembrane protease serine 2 [EC:3.4.21.122]
Pathway	map05164 Influenza A map05171 Coronavirus disease - COVID-19 map05202 Transcriptional misregulation in cancer map05215 Prostate cancer
Brite	KEGG Orthology (KO) [BR:ko00001] 09160 Human Diseases 09161 Cancer: overview 05202 Transcriptional misregulation in cancer K09633 TMPRSS2; transmembrane protease serine 2 09162 Cancer: specific types 05215 Prostate cancer

ERG:

Search KEGG for ERG Go Clear

Database: KEGG - Search term: ERG

KEGG PATHWAY

map04261
Adrenergic signaling in cardiomyocytes
map04724
Glutamatergic synapse
map04725
Cholinergic synapse
map04726
Serotonergic synapse
map04727
GABAergic synapse
• • • » display all

KEGG MODULE

M00102
Ergocalciferol biosynthesis, FPP => ergosterol/ergocalciferol

KEGG ORTHOLOGY

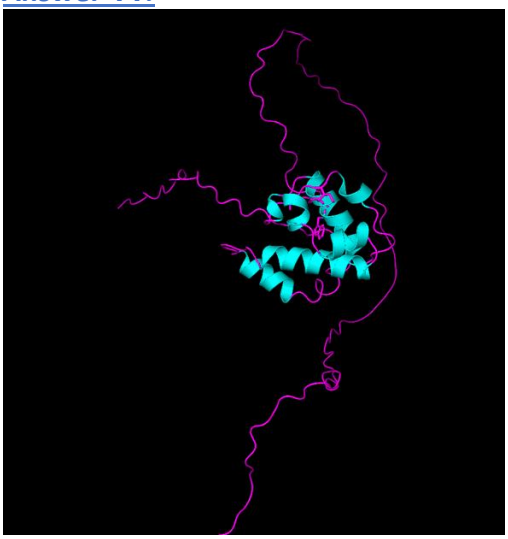
K00222
TM7SF2, ERG24; Delta14-sterol reductase [EC:1.3.1.70]
K00223
ERG4; Delta24(24(1))-sterol reductase [EC:1.3.1.71]
K00227
SC5DL, ERG3; Delta7-sterol 5-desaturase [EC:1.14.19.20]
K00511
SQLE, ERG1; squalene monooxygenase [EC:1.14.14.17]
K00559
SMT1, ERG6; sterol 24-C-methyltransferase [EC:2.1.1.41]
• • • » display all

ERG is a transcription factor belonging to the ETS family (which is the DNA-binding domain of ERG). In KEGG, it is found in pathways related to cancer development and progression. ERG has a relation to prostate cancer development when is fused with TMPRSS2. In KEGG pathways, this fusion would likely be represented in prostate cancer-specific signaling cascades.

Question 11:

Present the predicted structure in PyMol using the cartoon presentation. Print screen.

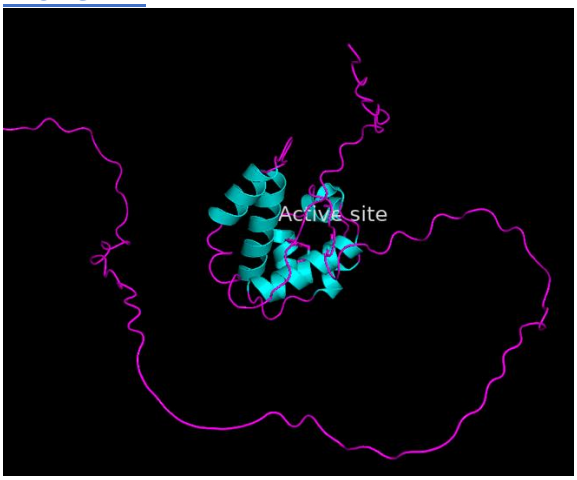
Answer 11:

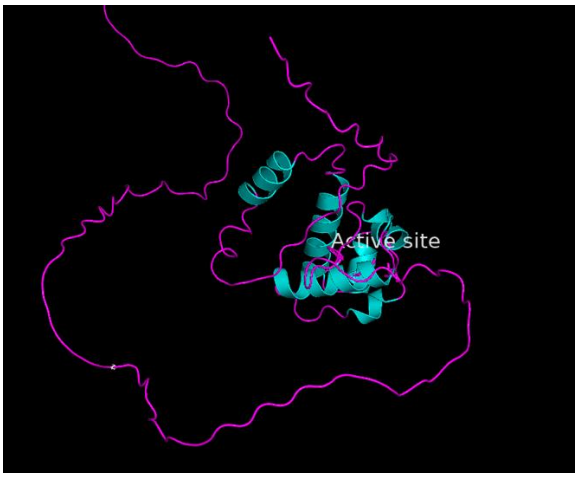


Question 12:

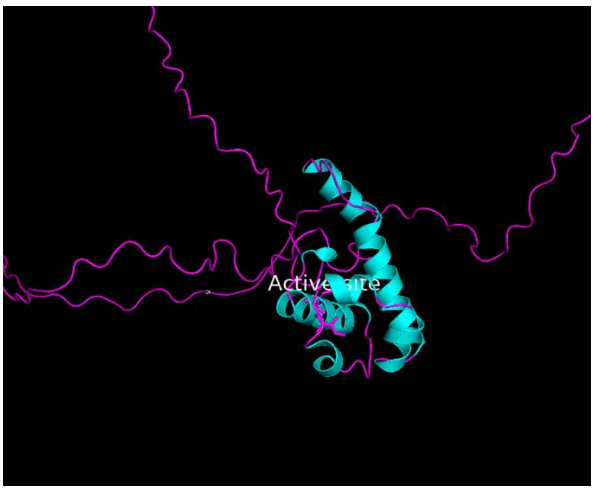
Find the active site in the 3D protein molecule, present it by the surface presentation and label all the atoms in the active site. Make 10 images with different presentations, rotations, and views

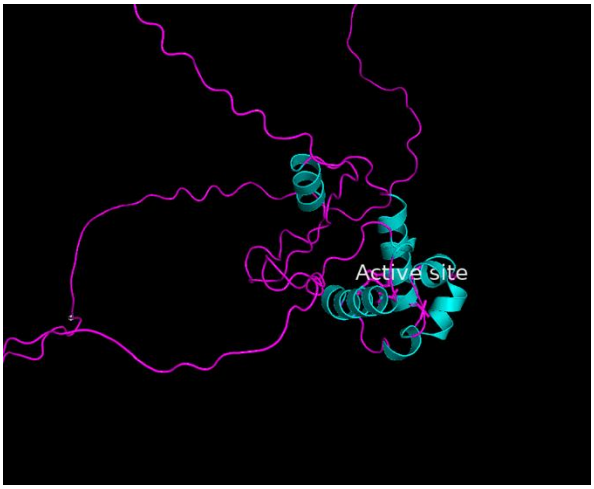
Answer 12

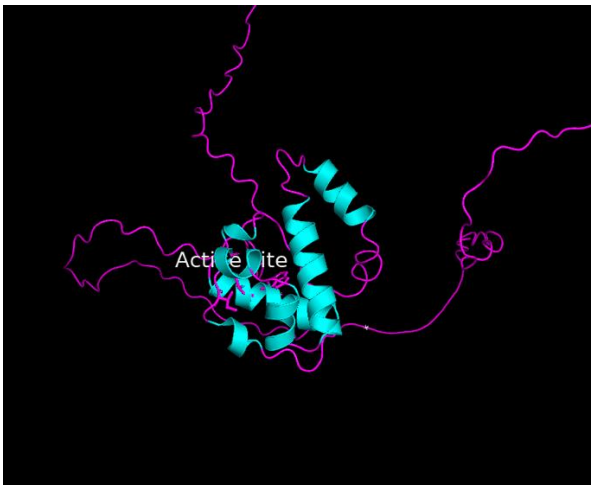


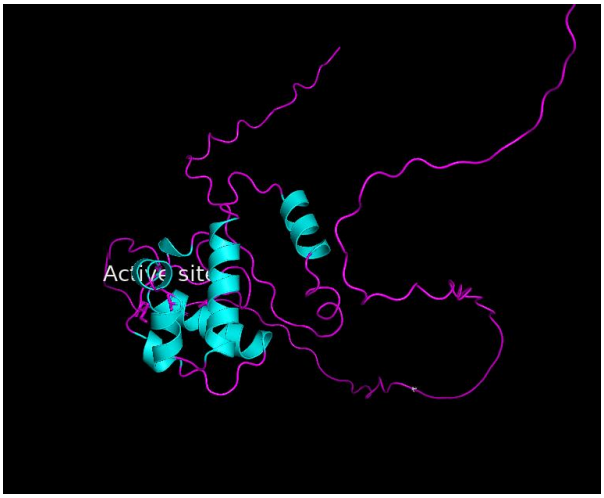



Active site

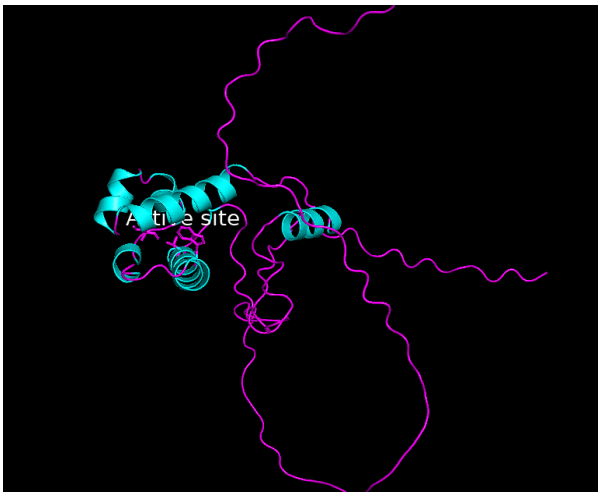

Active site

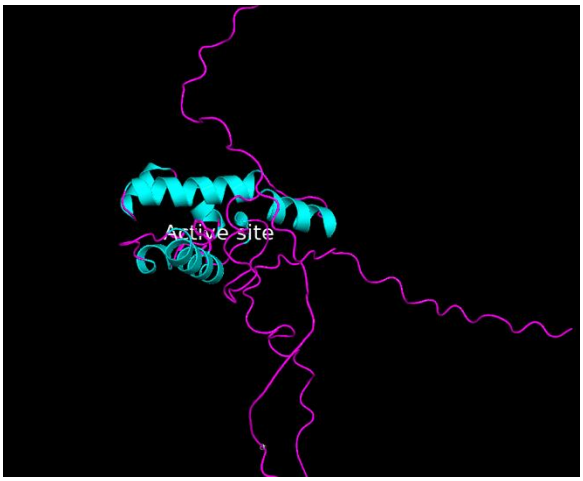

Active site


Active site


Active site


Active site





Question 13+14:

Question 13:

Make a movie of the 3D structure of the chimeric protein in different presentations from cartoon to the surface presentation, zoom in to the active site for the movie #1 of 20 sec.

Question 14:

Use different selections to select the active site atoms.

Answer 14:

The movie of the chimera all was sent in an added video named CHIMERA_ERG_TMPRSS2

Question 15:

Find pdb files of the parental proteins: ERG and TMPRSS2 or predict their protein structure. Produce structural alignment with the chimera and the pdb files of parental proteins ERG and TMPRSS2, or their predicted 3D protein structure. Make a short movie #2 to present the structural alignment for 10 seconds in different presentations and views.

Answer 15:

The movie 0f the parental proteins and the chimera all together was sent in an added video named ERG&TMPRSS2&CHIMERA_ERG_TMPRSS2

Transeq links:

1. https://www.ebi.ac.uk/jdispatcher/st/emboss_transeq/summary?jobId=emboss_transeq-I20240727-152933-0881-18103251-p1m&js=pass
2. https://www.ebi.ac.uk/jdispatcher/st/emboss_transeq/summary?jobId=emboss_transeq-I20240727-153058-0474-10559950-p1m&js=pass
3. https://www.ebi.ac.uk/jdispatcher/st/emboss_transeq/summary?jobId=emboss_transeq-I20240727-153137-0211-67548566-p1m&js=pass
4. https://www.ebi.ac.uk/jdispatcher/st/emboss_transeq/summary?jobId=emboss_transeq-I20240727-153158-0046-7829944-p1m&js=pass
5. https://www.ebi.ac.uk/jdispatcher/st/emboss_transeq/summary?jobId=emboss_transeq-I20240727-153224-0025-10523302-p1m&js=pass
6. https://www.ebi.ac.uk/jdispatcher/st/emboss_transeq/summary?jobId=emboss_transeq-I20240727-153244-0664-91866022-p1m&js=pass