Lucia Lopez Clavain

Student number:2001133

Word count: 2707 words

Title: Protein sequence analysis of melanocyte-stimulating hormone receptor

Abstract

The use of several bioinformatics tools makes possible the analysis of DNA and protein sequences. The protein given is homologous with the melanocyte-stimulating hormone receptor protein. It is part of the G-protein coupled receptors 1 family. The reference protein and the unknown protein had very similar sequences, so HMMER was the last tool used to confirm the identification of the protein. Secondly, PCR primers are a short sequence of oligonucleotides that provide a start and end reference point in the DNA sequence when applying the PCR technique. Two primers with appropriate characteristics have been designed for melanocyte-stimulating hormone receptors. Further research can be done using other websites and tools with more in-depth information.

Introduction

Bioinformatics is an important modern advance in the biological and informatics field. It uses informatics tools and software to understand, analyse and interpret large amounts of biological data. In this report, a protein has to be identified from a cDNA sequence given. Many websites that contribute to protein and DNA sequences analysis, but only specific tools and websites will be used to achieve this task in this report.

Sequence manipulation suite (SMS) is a group of programs for generating and analysing DNA and protein sequences. It will be used in this report to find the reading frames of the DNA sequence, to obtain the cDNA-translated protein sequence, and to calculate the characteristics of the protein such as molecular weight and the isoelectric point. BLAST is an excellent tool to do homology searchers. To find the closest homologs to the melanocyte-stimulating hormone receptor protein the ncbiblast tool was used. The database UniProtKB/Swiss-Pro includes annotated proteins and further research can be done by comparison. Therefore, the comparison of the unknown protein and the reference protein sequences will be done using this program. The 3D protein structure will be determined by BLAST and two other websites. The significance of the similarity of both proteins sequence was confirmed using the program HMMER.

The development of the PCR technique during the years was key to introducing new improvements that transformed and simplified genetic profiling. An appropriate primer design is key to creating a successful PCR, as a bad primer can lead to miss-amplification of the target sequence or no amplification at all. Two primers were created for the cDNA sequence given using the website US National library of medicine.

Results and discussion

The cDNA sequence given is the following:

```
ATGGCGGTGCAGGGCAACCAGCGCCGCCTGCTGGGCAGCCTGAACAGCACCCCGACCGCG
ATTCCGCAGCTGGGCCTGGCGGCGAACCAGACCTGCGCGCGCTGCCTGGAAGTGAGCATT
AGCGATGGCCTGTTTCTGAGCCTGGGCCTGGTGAGCCTGGTGGAAAACGCGCTGGTGGTG
GCGACCATTGCGAAAAACCGCAACCTGCATAGCCCGATGTATTGCTTTATTTGCTGCCTG
GCGCTGAGCGATCTGCTGGTGAGCGGCAGCAACGTGCTGGAAACCGCGGTGATTCTGCTG
CTGGAAGCGGGCGCGCTGGTGGCGCGCGCGGCGGTGCTGCAGCAGCTGGATAACGTGATT
GATGTGATTACCTGCAGCAGCATGCTGAGCAGCCTGTGCTTTCTGGGCGCGATTGCGGTG
GATCGCTATATTAGCATTTTTTATGCGCTGCGCTATCATAGCATTGTGACCCTGCCGCGC
GCGCGCCGCGCGGTGGCGGCGATTTGGGTGGCGAGCGTGGTGTTTAGCACCCTGTTTATT
GCGTATTATGATCATGTGGCGGTGCTGCTGTGCCTGGTGGTGTTTTTTCTGGCGATGCTG
GTGCTGATGGCGGTGCTGTATGTGCATATGCTGGCGCGCGCGTGCCAGCATGCGCAGGGC
ATTGCGCGCCTGCATAAACGCCAGCGCCCGGTGCATCAGGGCTTTGGCCTGAAAGGCGCG
GTGACCCTGACCATTCTGCTGGGCATTTTTTTTCTGTGCTGGGGCCCGTTTTTTCTGCAT
CTGACCCTGATTGTGCTGTGCCCGGAACATCCGACCTGCGGCTGCATTTTTAAAAACTTT
AACCTGTTTCTGGCGCTGATTATTTGCAACGCGATTATTGATCCGCTGATTTATGCGTTT
CATAGCCAGGAACTGCGCCGCACCCTGAAAGAAGTGCTGACCTGCAGCTGGTGA
```

Figure 1: cDNA sequence to study

Sequence manipulation Suite

DNA sequences find out what proteins they encode. They have six reading frames in which the protein could be encoded by the double-stranded DNA. Each strand has 3 frames, but they have different directions. In addition, every amino acid of the protein is determined by a codon of three bases. The website used for the development of the reading frames is called Sequence Manipulation Suite. The translation Map gets a DNA sequence and gives a textual map showing protein translations. The DNA sequence used for this report was converted by the translation map to obtain the following result.

```
  1   G  G  A  G  Q  P  A  P  P  A  G  Q  P  E  Q  H  P  D  R  D
      1  W  R  C  R  A  T  S  A  A  C  W  A  A  *  T  A  P  R  P  R
      1 M  A  V  Q  G  N  Q  R  R  L  L  G  S  L  N  S  T  P  T  A
      1ATGGCGGTGCAGGGCAACCAGCGCCGCCTGCTGGGCAGCCTGAACAGCACCCCGACCGCG
      1          10        20        30        40        50
      1TACCGCCACGTCCCGTTGGTCGCGGCGGACGACCCGTCGGACTTGTCGTGGGGCTGGCGC
     21   S  A  A  G  P  G  G  E  P  D  L  R  A  L  P  G  S  E  H  *
     21  F  R  S  W  A  W  R  R  T  R  P  A  R  A  A  W  K  *  A  L
     21 I  P  Q  L  G  L  A  A  N  Q  T  C  A  R  C  L  E  V  S  I
     61 ATTCCGCAGCTGGGCCTGGCGGCGAACCAGACCTGCGCGCGCTGCCTGGAAGTGAGCATT
     61          70        80        90        100       110
     61 TAAGGCGTCGACCCGGACCGCCGCTTGGTCTGGACGCGCGCGACGGACCTTCACTCGTAA
     41   R  W  P  V  S  E  P  G  P  G  E  P  G  G  K  R  A  G  G  G
     41  A  M  A  C  F  *  A  W  A  W  *  A  W  W  K  T  R  W  W  W
     41 S  D  G  L  F  L  S  L  G  L  V  S  L  V  E  N  A  L  V  V
    121 AGCGATGGCCTGTTTCTGAGCCTGGGCCTGGTGAGCCTGGTGGAAAACGCGCTGGTGGTG
    121          130       140       150       160       170
    121 TCGCTACCGGACAAAGACTCGGACCCGGACCACTCGGACCACCTTTTGCGCGACCACCAC
     61   D  H  C  E  K  P  Q  P  A  *  P  D  V  L  L  Y  L  L  P  G
     61  R  P  L  R  K  T  A  T  C  I  A  R  C  I  A  L  F  A  A  W
     61 A  T  I  A  K  N  R  N  L  H  S  P  M  Y  C  F  I  C  C  L
    181 GCGACCATTGCGAAAAACCGCAACCTGCATAGCCCGATGTATTGCTTTATTTGCTGCCTG
    181          190       200       210       220       230
```

```
181 CGCTGGTAACGCTTTTTGGCGTTGGACGTATCGGGCTACATAACGAAATAAACGACGGAC
 81   A  E  R  S  A  G  E  R  Q  Q  R  A  G  N  R  G  D  S  A  A
 81   R  *  A  I  C  W  *  A  A  A  T  C  W  K  P  R  *  F  C  C
 81 A  L  S  D  L  L  V  S  G  S  N  V  L  E  T  A  V  I  L  L
241 GCGCTGAGCGATCTGCTGGTGAGCGGCAGCAACGTGCTGGAAACCGCGGTGATTCTGCTG
241          250       260       270       280       290
241 CGCGACTCGCTAGACGACCACTCGCCGTCGTTGCACGACCTTTGGCGCCACTAAGACGAC
101   G  S  G  R  A  G  G  A  R  G  G  A  A  A  A  G  *  R  D  *
101   W  K  R  A  R  W  W  R  A  R  R  C  C  S  S  W  I  T  *  L
101 L  E  A  G  A  L  V  A  R  A  A  V  L  Q  Q  L  D  N  V  I
301 CTGGAAGCGGGCGCGCTGGTGGCGCGCGCGGCGGTGCTGCAGCAGCTGGATAACGTGATT
301          310       320       330       340       350
301 GACCTTCGCCCGCGCGACCACCGCGCGCGCCGCCACGACGTCGTCGACCTATTGCACTAA
121   C  D  Y  L  Q  Q  H  A  E  Q  P  V  L  S  G  R  D  C  G  G
121   M  *  L  P  A  A  A  C  *  A  A  C  A  F  W  A  R  L  R  W
121 D  V  I  T  C  S  S  M  L  S  S  L  C  F  L  G  A  I  A  V
361 GATGTGATTACCTGCAGCAGCATGCTGAGCAGCCTGTGCTTTCTGGGCGCGATTGCGGTG
361          370       380       390       400       410
361 CTACACTAATGGACGTCGTCGTACGACTCGTCGGACACGAAAGACCCGCGCTAACGCCAC
141   S  L  Y  *  H  F  L  C  A  A  L  S  *  H  C  D  P  A  A  R
141   I  A  I  L  A  F  F  M  R  C  A  I  I  A  L  *  P  C  R  A
141 D  R  Y  I  S  I  F  Y  A  L  R  Y  H  S  I  V  T  L  P  R
421 GATCGCTATATTAGCATTTTTTATGCGCTGCGCTATCATAGCATTGTGACCCTGCCGCGC
421          430       440       450       460       470
421 CTAGCGATATAATCGTAAAAAATACGCGACGCGATAGTATCGTAACACTGGGACGGCGCG
161   A  P  R  G  G  G  D  L  G  G  E  R  G  V  *  H  P  V  Y  C
161   R  A  A  R  W  R  R  F  G  W  R  A  W  C  L  A  P  C  L  L
161 A  R  R  A  V  A  A  I  W  V  A  S  V  V  F  S  T  L  F  I
481 GCGCGCCGCGCGGTGGCGGCGATTTGGGTGGCGAGCGTGGTGTTTAGCACCCTGTTTATT
481          490       500       510       520       530
481 CGCGCGGCGCGCCACCGCCGCTAAACCCACCGCTCGCACCACAAATCGTGGGACAAATAA
181   V  L  *  S  C  G  G  A  A  V  P  G  G  V  F  S  G  D  A  G
181   R  I  M  I  M  W  R  C  C  C  A  W  W  C  F  F  W  R  C  W
181 A  Y  Y  D  H  V  A  V  L  L  C  L  V  V  F  F  L  A  M  L
541 GCGTATTATGATCATGTGGCGGTGCTGCTGTGCCTGGTGGTGTTTTTTCTGGCGATGCTG
541          550       560       570       580       590
541 CGCATAATACTAGTACACCGCCACGACGACACGGACCACCACAAAAAAGACCGCTACGAC
201   A  D  G  G  A  V  C  A  Y  A  G  A  R  V  P  A  C  A  G  H
201   C  *  W  R  C  C  M  C  I  C  W  R  A  R  A  S  M  R  R  A
201 V  L  M  A  V  L  Y  V  H  M  L  A  R  A  C  Q  H  A  Q  G
601 GTGCTGATGGCGGTGCTGTATGTGCATATGCTGGCGCGCGCGTGCCAGCATGCGCAGGGC
601          610       620       630       640       650
601 CACGACTACCGCCACGACATACACGTATACGACCGCGCGCGCACGGTCGTACGCGTCCCG
221   C  A  P  A  *  T  P  A  P  G  A  S  G  L  W  P  E  R  R  G
221   L  R  A  C  I  N  A  S  A  R  C  I  R  A  L  A  *  K  A  R
221 I  A  R  L  H  K  R  Q  R  P  V  H  Q  G  F  G  L  K  G  A
661 ATTGCGCGCCTGCATAAACGCCAGCGCCCGGTGCATCAGGGCTTTGGCCTGAAAGGCGCG
661          670       680       690       700       710
661 TAACGCGCGGACGTATTTGCGGTCGCGGGCCACGTAGTCCCGAAACCGGACTTTCCGCGC
241   D  P  D  H  S  A  G  H  F  F  S  V  L  G  P  V  F  S  A  S
241   *  P  *  P  F  C  W  A  F  F  F  C  A  G  A  R  F  F  C  I
241 V  T  L  T  I  L  L  G  I  F  F  L  C  W  G  P  F  F  L  H
721 GTGACCCTGACCATTCTGCTGGGCATTTTTTTTCTGTGCTGGGGCCCGTTTTTTCTGCAT
721          730       740       750       760       770
721 CACTGGGACTGGTAAGACGACCCGTAAAAAAAAGACACGACCCCGGGCAAAAAAGACGTA
261   D  P  D  C  A  V  P  G  T  S  D  L  R  L  H  F  *  K  L  *
261   *  P  *  L  C  C  A  R  N  I  R  P  A  A  A  F  L  K  T  L
261 L  T  L  I  V  L  C  P  E  H  P  T  C  G  C  I  F  K  N  F
781 CTGACCCTGATTGTGCTGTGCCCGGAACATCCGACCTGCGGCTGCATTTTTAAAAACTTT
781          790       800       810       820       830
781 GACTGGGACTAACACGACACGGGCCTTGTAGGCTGGACGCCGACGTAAAAATTTTTGAAA
281   P  V  S  G  A  D  Y  L  Q  R  D  Y  *  S  A  D  L  C  V  S
281   T  C  F  W  R  *  L  F  A  T  R  L  L  I  R  *  F  M  R  F
281 N  L  F  L  A  L  I  I  C  N  A  I  I  D  P  L  I  Y  A  F
841 AACCTGTTTCTGGCGCTGATTATTTGCAACGCGATTATTGATCCGCTGATTTATGCGTTT
841          850       860       870       880       890
841 TTGGACAAAGACCGCGACTAATAAACGTTGCGCTAATAACTAGGCGACTAAATACGCAAA
301   *  P  G  T  A  P  H  P  E  R  S  A  D  L  Q  L  V
```

```
301  I  A  R  N  C  A  A  P  *  K  K  C  *  P  A  A  G
301 H  S  Q  E  L  R  R  T  L  K  E  V  L  T  C  S  W  *
901 CATAGCCAGGAACTGCGCCGCACCCTGAAAGAAGTGCTGACCTGCAGCTGGTGA
901          910       920       930       940       950
901 GTATCGGTCCTTGACGCGGCGTGGGACTTTCTTCACGACTGGACGTCGACCACT
```

Figure 2: Results for 954 residue sequence "Untitled" starting "ATGGCGGTGC".

A possible protein-coding sequence starts with a specific codon (generally ATG which encodes methionine, met) followed by some amino acids and it is terminated by a stop codon. The longest coding sequence (CDS) the DNA sequence encodes is shown below. This is the cDNA-translated protein sequence.

```
M  A  V  Q  G  N  Q  R  R  L  L  G  S  L  N  S  T  P  T  A  I  P  Q  L  G  L  A  A
N  Q  T  C  A  R  C  L  E  V  S  I  S  D  G  L  F  L  S  L  G  L  V  S  L  V  E  N
A  L  V  V  A  T  I  A  K  N  R  N  L  H  S  P  M  Y  C  F  I  C  C  L  A  L  S  D
L  L  V  S  G  S  N  V  L  E  T  A  V  I  L  L  L  E  A  G  A  L  V  A  R  A  A  V
L  Q  Q  L  D  N  V  I  D  V  I  T  C  S  S  M  L  S  S  L  C  F  L  G  A  I  A  V
D  R  Y  I  S  I  F  Y  A  L  R  Y  H  S  I  V  T  L  P  R  A  R  R  A  V  A  A  I
W  V  A  S  V  V  F  S  T  L  F  I  A  Y  Y  D  H  V  A  V  L  L  C  L  V  V  F  F
L  A  M  L  V  L  M  A  V  L  Y  V  H  M  L  A  R  A  C  Q  H  A  Q  G  I  A  R  L
H  K  R  Q  R  P  V  H  Q  G  F  G  L  K  G  A  V  T  L  T  I  L  L  G  I  F  F  L
C  W  G  P  F  F  L  H  L  T  L  I  V  L  C  P  E  H  P  T  C  G  C  I  F  K  N  F
N  L  F  L  A  L  I  I  C  N  A  I  I  D  P  L  I  Y  A  F  H  S  Q  E  L  R  R  T
L  K  E  V  L  T  C  S  W  *
```

Figure 3: Longest coding sequence (CDS) of the cDNA-translated protein sequence.

The tool Filter Protein of the website Sequence Manipulation Suite was used to remove unnecessary non-protein characters such as blank spaces or digits from a sequence. When the protein is reformatted, it is easier to work with.

```
MAVQGNQRRLLGSLNSTPTAIPQLGLAANQTCARCLEVSISDGLFLSLGLVSLVENALVVATIAKNRN
LHSPMYCFICCLALSDLLVSGSNVLETAVILLLEAGALVARAAVLQQLDNVIDVITCSSMLSSLCFLI
AVDRYISIFYALRYHSIVTLPRARRAVAAIWVASVVFSTLFIAYYDHVAVLLCLVVFFLAMLVLMAVL
YVHMLARACQHAQGIARLHKRQRPVHQGFGLKGAVTLTILLGIFFLCWGPFFLHLTLIVLCPEHPTCC
IFKNFNLFLALIICNAIIDPLIYAFHSQELRRTLKEVLTCSW
```

Protein Stats of the website Sequence Manipulation Suite was used to find the number of amino acids of the protein in the correct format. This tool displays the number of occurrences of each amino acid in the sequence. The residues are also given with their percentages and in addition, groups of residues are also shown with their percentages. This allows a quickly comparation of the results obtained.

Table 1: Results for 314 residue sequence "sample sequence" starting "MAVQGNQRRL"

Pattern:	Times found:	Percentage:
A	34	10.83
B	0	0.00
C	15	4.78
D	7	2.23
E	7	2.23
F	17	5.41
G	13	4.14
H	10	3.18
I	23	7.32

K	5	1.59
L	55	17.52
M	6	1.91
N	11	3.50
P	9	2.87
Q	11	3.50
R	16	5.10
S	20	6.37
T	14	4.46
V	30	9.55
W	3	0.96
X	0	0.00
Y	8	2.55
Z	0	0.00
Aliphatic G,A,V,L,I	155	49.36
Aromatic F,W,Y	28	8.92
Sulphur C,M	21	6.69
Basic K,R,H	31	9.87
Acidic B,D,E,N,Q,Z	36	11.46
Aliphatic hydroxyl S,T	34	10.83
tRNA synthetase class I Z,E,Q,R,C,M,V,I,L,Y,W	174	55.41
tRNA synthetase class II B,G,A,P,S,T,H,D,N,K,F	140	44.59

The highest number of amino acids found in the sequence is leucine (L) with 55 times counted, 17.52% of the protein. The aliphatic group has the highest number of amino acids in the protein (155 times found, 49.36%). An aliphatic compound represents any chemical organic element with atoms connected by triple, double, or single bonds to make nonaromatic structures. This includes the alkenes, alkanes, alkynes, and any compound derived from them (Britannica, 2022).

For finding the molecular weight of the protein, the tool protein molecular weight of the website Sequence Manipulation Suite was used. This can be used when a prediction of the location of a target protein is made. The protein molecular weight result was 34.60 kDa.

The tool Protein Isoelectric Point of the website Sequence Manipulation Suite was used to estimate the protein isoelectric point (pl). The website determines the theoretical isoelectric point for the protein sequence. This is useful to calculate where on a 2-D gel a certain protein will be found. The result was pH 8.39.

BLAST

After the unknown protein has been found, resources with databases containing annotated proteins are a good place to look for more information about the protein. This is because if the unknown protein shares a sufficient degree of homology with a known protein, it is assumed that those two proteins have the same function. For this task, the tool used is BLAST (Basic Local Alignment Search Tool), BLASTP when working with proteins. This is an American website specialising in finding parts between sequences with local similarity. It works with protein or nucleotide sequences and compares them to sequence databases, considering the

statistical significance. It can also be used to determine gene families, design PCR primers, and search T cell and immunoglobulins receptor sequences. Homology searches are normally carried out by BLAST tools.

The best homolog found for the target protein (the highest % identity / lowest E value) is UniProtKB - Q01726 (MSHR_HUMAN). The protein the cDNA-translated protein sequence encodes is called melanocyte-stimulating hormone receptor, and the gene name is MC1R. The target protein is found in Homo sapiens (humans).

The protein family is the G-protein coupled receptors 1 family and their main function is to transduce extracellular signals by interaction with guanine nucleotide-binding (G) proteins. G-coupled receptors (GPCRs) represent a large protein family that has a broad range of autocrine, endocrine and paracrine functions. Overall, more than 400 receptors are part of the GPCRs family. They are usually drug targets which gives them a high analytical interest. The subgroup of GPCRA, rhodopsin-like GPCRs, illustrates a big protein family including neurotransmitter, hormone and light receptors that convert extracellular signals by interacting with nucleotide-binding (G) protein and guanine. They have high similarity in their amino acid sequences, even though their activating ligands have different characters and structures. It is believed that they assume a general structure forming 7 transmembrane helices. The MHS receptor is found abundantly in melanomas, melanocytes and their similar cell lines.

The homolog protein (melanocyte-stimulating hormone receptor) sequence is the following:
MAVQGSQRRLLGSLNSTPTAIPQLGLAANQTGARCLEVSISDGLFLSLGLVSLVENALVVATIAKNRN
LHSPMYCFICCLALSDLLVSGSNVLETAVILLLEAGALVARAAVLQQLDNVIDVITCSSMLSSLCFLG
AIAVDRYISIFYALRYHSIVTLPRARRAVAAIWVASVVFSTLFIAYYDHVAVLLCLVVFFLAMLVLMA
VLYVHMLARACQHAQGIARLHKRQRPVHQGFGLKGAVTLTILLGIFFLCWGPFFLHLTLIVLCPEHPT
CGCIFKNFNLFLALIICNAIIDPLIYAFHSQELRRTLKEVLTCSW
Figure 4: homolog protein sequence

BLAST summary table supplies much information about the homology proteins and the reference proteins. The sum of substitution and gap scores calculates the score of an alignment. To obtain this result, substitution values are given by a table done by sites like BLOSUM or PAM, and the gap scores are the result of the sum of the gap opening penalty, G and L. The score of the reference protein is 603 bits. The identity represents the amount of similarity between two amino acid or nucleotide sequences regarding the residues at the same location in an alignment. In this case, the identities are 312/317, so 98%. Similarities show how much two sequences are related and the similarity obtained for the protein melanocyte-stimulating hormone receptor is 99%. The E value illustrates the number of alignments with scores better than or equivalent to S that are expected to happen in a database by chance. The highest the E value, the less significant the alignment and the score. The E value for the reference protein is 0.0.

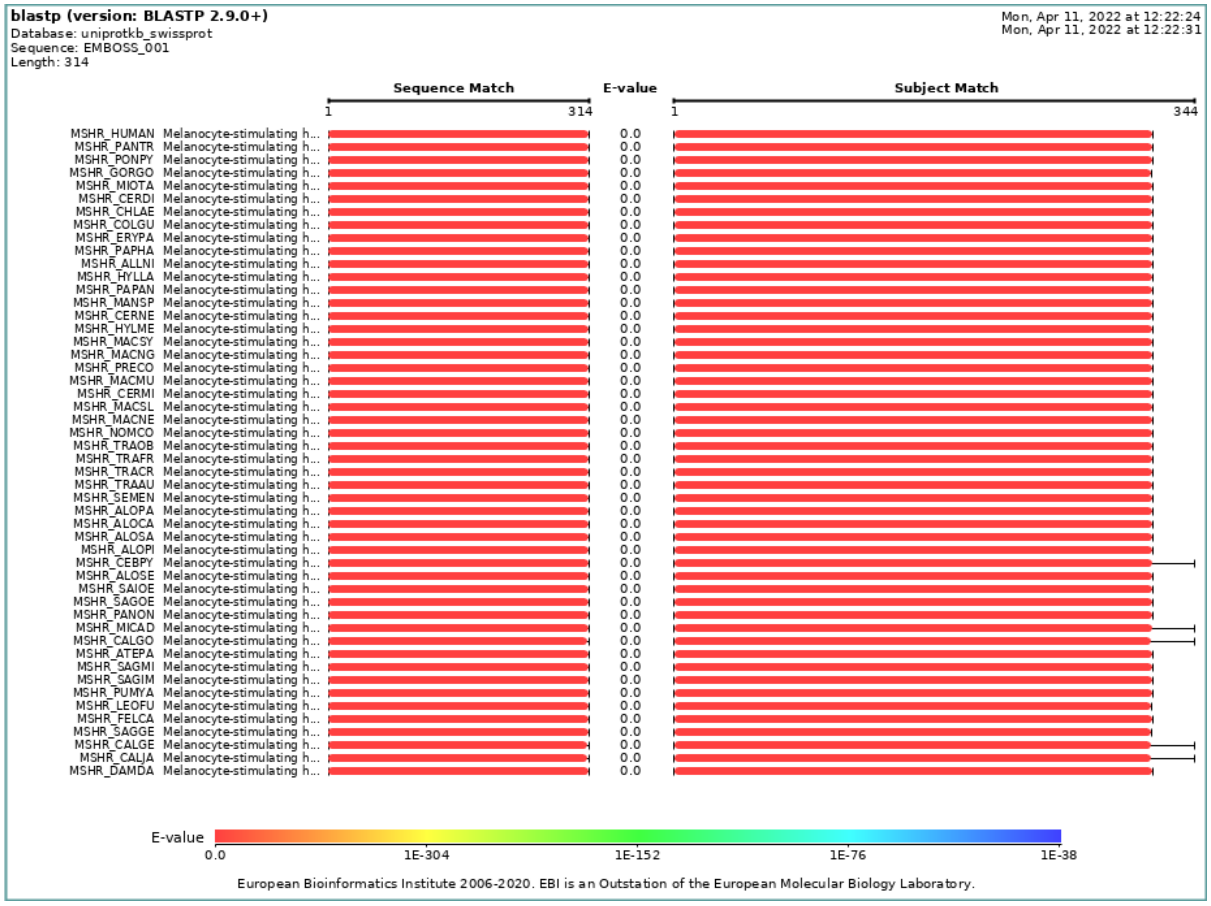

Figure 5: a picture showing the visual output of BLAST for the analysis of the protein melanocyte-stimulating hormone receptor.

Sequence comparation

Often analysing a novel protein requires the comparison of two protein sequences, the disease-associated sequence, and the known protein sequence. Pairwise BLAST is usually used to determine the differences between two sequences. Consequently, any changes in the location of the amino acids can be observed in the result. EMBOSS Water works with the Smith-Waterman algorithm to analyse the local alignment of two sequences. In this case, the reference protein is being compared with the homolog one. The result can be a guide to designing PCR primers, allowing further analysis of the mutated section in the protein. Investigators amplify and analyse the specific part of the gene and study the frequency these mutations show with other people.

```
1 MAVQGNQRRLLGSLNSTPTAIPQLGLAANQTCARCLEVSISDGLFLSLGL     50
  |||||:|||||||||||||||||||||||.||||||||||||||||||||
1 MAVQGSQRRLLGSLNSTPTAIPQLGLAANQTGARCLEVSISDGLFLSLGL     50

51 VSLVENALVVATIAKNRNLHSPMYCFICCLALSDLLVSGSNVLETAVILL    100
   ||||||||||||||||||||||||||||||||||||||||||||||||||
51 VSLVENALVVATIAKNRNLHSPMYCFICCLALSDLLVSGSNVLETAVILL    100

101 LEAGALVARAAVLQQLDNVIDVITCSSMLSSLCFL--IAVDRYISIFYAL    148
    |||||||||||||||||||||||||||||||||||  ||||||||||||
101 LEAGALVARAAVLQQLDNVIDVITCSSMLSSLCFLGAIAVDRYISIFYAL    150

149 RYHSIVTLPRARRAVAAIWVASVVFSTLFIAYYDHVAVLLCLVVFFLAML    198
    ||||||||||||||||||||||||||||||||||||||||||||||||||
151 RYHSIVTLPRARRAVAAIWVASVVFSTLFIAYYDHVAVLLCLVVFFLAML    200
```

```
199 VLMAVLYVHMLARACQHAQGIARLHKRQRPVHQGFGLKGAVTLTILLGIF    248
    ||||||||||||||||||||||||||||||||||||||||||||||||||||
201 VLMAVLYVHMLARACQHAQGIARLHKRQRPVHQGFGLKGAVTLTILLGIF    250

249 FLCWGPFFLHLTLIVLCPEHPTC-CIFKNFNLFLALIICNAIIDPLIYAF    297
    |||||||||||||||||||||| ||||||||||||||||||||||||||
251 FLCWGPFFLHLTLIVLCPEHPTCGCIFKNFNLFLALIICNAIIDPLIYAF    300

298 HSQELRRTLKEVLTCSW    314
    |||||||||||||||||
301 HSQELRRTLKEVLTCSW    317
```

Figure 6: Comparison of the homolog protein sequence and the cDNA-translated protein sequence. The figures highlighted in yellow are the figures that differ from the original sequence.

It can be observed the high resemblance between the homolog protein sequence and the reference protein. Although, the sequences differ in two amino acids (N instead of S, and C instead of G) and three missing ones (they are highlighted in yellow on the result). For this reason, it can be concluded that the ADN sequence section in those specific parts is different. Some codons can express the same protein in the Universal Genetic Code.

3D structure

Three websites were used to find the 3D model structure for the reference protein. The first one is Swiss-model from Biozentrum of University of Basel.

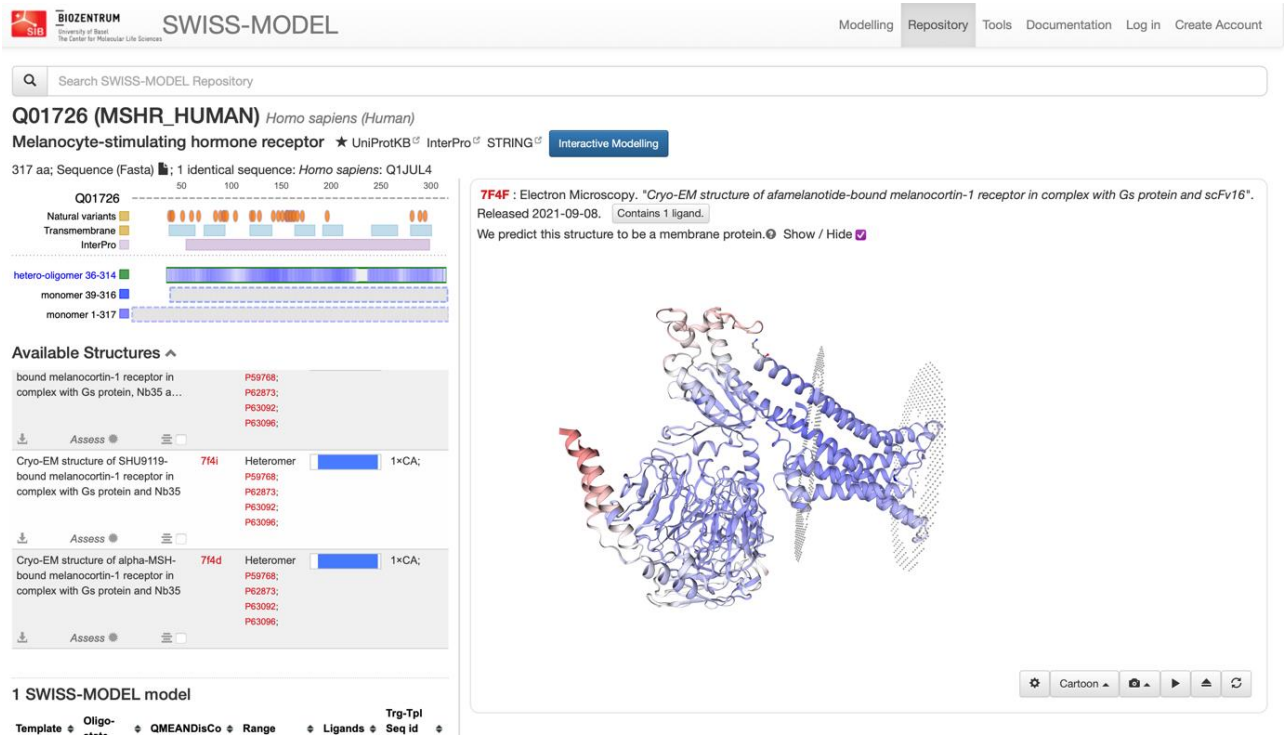

Figure 7: A picture of the 3D structure and analysis of Q01726, Melanocyte-stimulating hormone receptor, in the website Swiss-Model.

The Swiss model also shows the hetero-oligomers found in the whole protein, such as P59768 or P62873. The main function of these proteins is to act as a transducer or modulator in several transmembrane signalling pathways. P59768 is also called guanine nucleotide-binding protein

G subunit gamma-2 a, and the protein P62873, guanine nucleotide-binding protein G subunit beta-1.

The second website was UniProt. It is a group of databases that provides a comprehensive resource for annotation data and protein sequence. It is a cooperation project between the SIB Swiss Institute of bioinformatics, the European bioinformatics institute (EMBL-EBI), and the Protein Information Resource (PIR). When using BLAST, a lot of information about the protein can be generated, such as function, structure, similar proteins, and publications about the protein. The structure of our protein, the melanocyte-stimulating hormone receptor was found and AlphaFold produces a confidence score between 0 and 100, which can be seen on the left side of Figure 8.

Figure 8: A picture showing the 3D structure of the protein melanocyte-stimulating hormone receptor. It was generated by BLAST from the website UniProt.

The third website was RCSB Protein Data Bank. The website Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi was used to obtain the pbd archive needed to create the 3D view in RCSB. This tool makes a double-helical secondary structure of DNA using previous research.

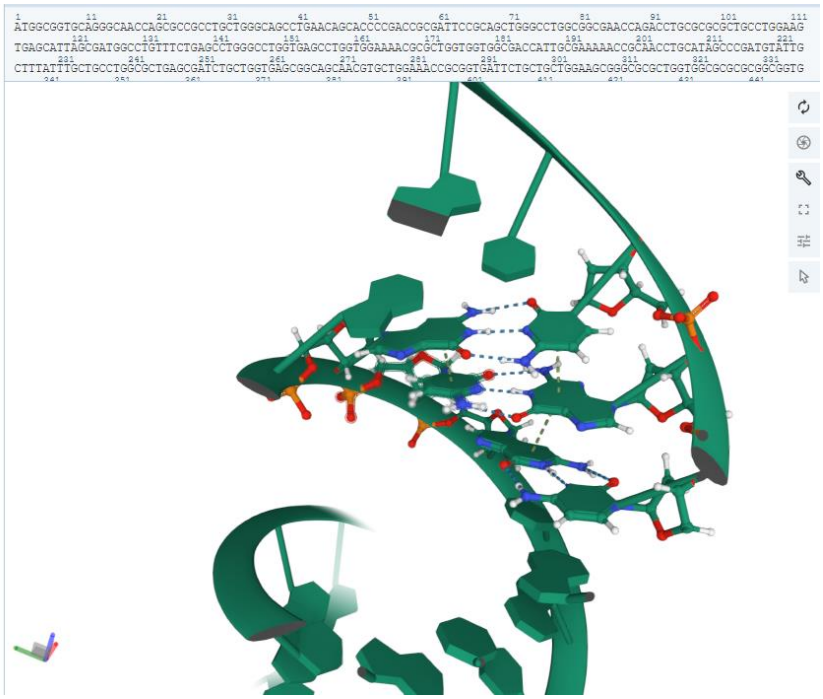

Figure 9: A picture showing the 3D structure of the protein melanocyte-stimulating hormone receptor. It was generated by RCSB Protein Data Bank.

Hhmer

Hhmer is a tool for bio sequence analysis that implements processes using profile hidden Markov models to compare domains or conserved sequences in proteins with the same function. It is used to search databases for sequence homologs (between proteins with functional annotations in databases and unknown proteins) and to create sequence alignments. The downside of Hhmer is that it works well with very known proteins, but it is not strong and reliable enough with unusual species. Hhmer is commonly used along with other profile databases such as Pfam.

Looking at the domain of the output given by Hmmer, the same results as the BLAST output were obtained. The protein is a 7 transmembrane receptor (rhodopsin family), also known as G protein-coupled receptors. Since the results were the same, it is enough evidence to suggest that the protein is a functional homolog of the G-protein coupled receptors 1 family.

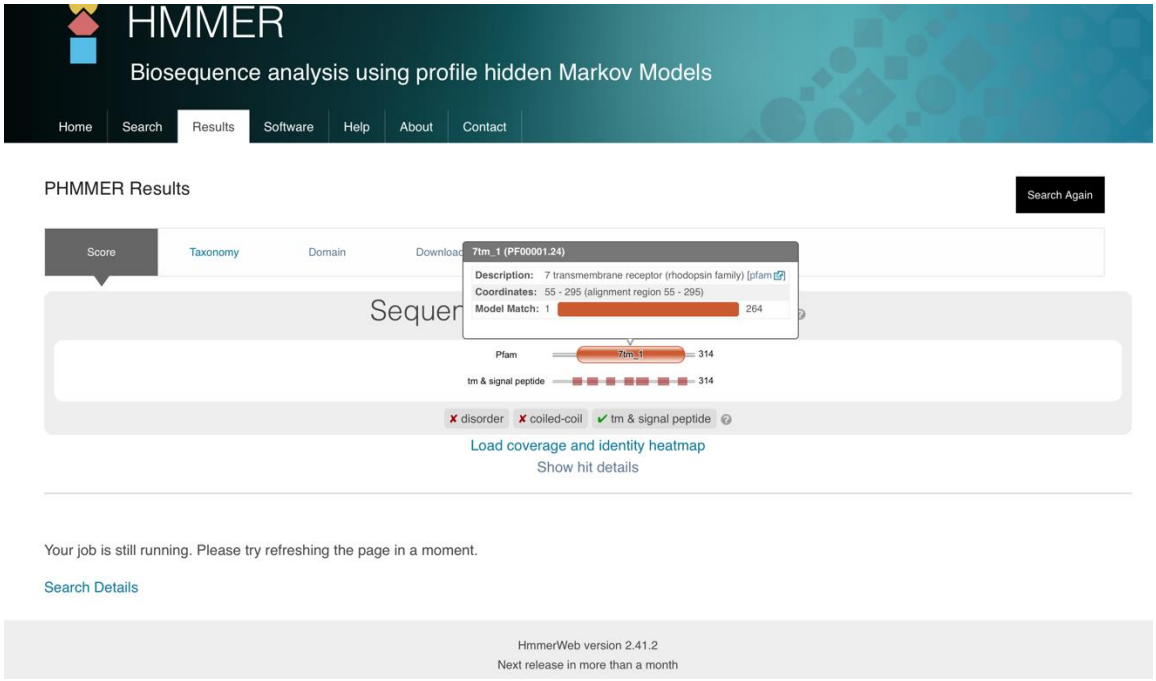

Figure 10: Hmmer analysis output for 7 transmembrane receptor protein (rhodopsin family).

PCR primers

The polymerase chain reaction (PCR) was developed by Karis Mullis (1983) to amplify DNA. It rapidly became a popular technique, which lead to further development for isolating specific genes in cloning. Nowadays, it is the main technique used for mutagenesis, cloning, and amplification of a target sequence for Sanger sequencing (combined or separate reactions). Having an amplified ADN sequence can be useful to make a further analysis of the targeted sequence and to apply different forensic and genetic techniques. PCR can also detect viral and bacterial pathogens, as well as mutations linked with disease and multiple recurrent loci. In forensic science, identification of DNA and genetic testing is the best proof to incriminate someone in a case, as it has a great sensibility. This is because it has the highest possibility of estimating the level of uncertainty of the results due to their sensitiveness and power.

PCR replicates the DNA sequence of interest by semiconservative replication, which is repeated many times to create a very large amount of DNA. This technique needs some components to work. A DNA template containing the target sequence is an essential component; PCR also requires BSA to avoid unwanted protein-DNA interactions, DNA polymerase, and PCR buffer to add Mg2+ and to change pH, nucleotides, and two primers in the two ends of the target (Spiers, 2022).

Each cycle has three different steps repeated from 25 to 35 times in the whole process. If the reaction is successful, it can produce a milliard of copies from one template. This is because the template used in the next cycle is the DNA strand previously synthesized, it has a logarithmic growth. The first basic step is denaturation. It works at 96 °C due to the use of heat to separate the DNA strand. The next step is called annealing and it should have a temperature of 55 - 65° C° to cool down the reaction and let the primers attach to their complementary sequences in the DNA strand. Although the annealing temperature depends on the characteristics of the primer and how much heat is required to dematerialize 50% of the primer-

DNA complex. Finally, the DNA polymerase is activated, and the new ADN strands are synthesised. This last step is called extension (Spiers, 2022).

Primers aim to amplify the target sequence without the difficulties of non-specific amplicons. They define the two ends of the DNA sequence, and they are commonly 6-30 nucleotides long. Short primers are not very specific due to their ability to bind many times to a different part of the DNA sequence; longer ones are more likely to be distinctive and specific. Primers help to start synthesizing the new DNA by providing a 3'-OH for the DNA polymerase and binding to their corresponding base pair. ADN polymerases can only work if there is a primer present (Spiers, 2022).

The primer design website tool of the US National Library of Medicine was used to create two primers that will amplify all the protein-coding sequences for the ADN sequence given. It finds specific primers to the PCR target template by using BLAST and Primer3. Firstly, the correct reading frame was determined for the protein given in the DNA sequence. For this purpose, the website used was emboss sixpack. Six-Pack output with Melanocyte-stimulating hormone receptor protein identified is shown below. The correct reading frame was F1 so all the other reading frames were eliminated.

```
M  A  V  Q  G  N  Q  R  R  L  L  G  S  L  N  S  T  P  T  A     F1

  1 ATGGCGGTGCAGGGCAACCAGCGCCGCCTGCTGGGCAGCCTGAACAGCACCCCGACCGCG 60

    ----:----|----:----|----:----|----:----|----:----|----:----|

  1 TACCGCCACGTCCCGTTGGTCGCGGCGGACGACCCGTCGGACTTGTCGTGGGGCTGGCGC 60

                                                        5'-GAGCATT

  I  P  Q  L  G  L  A  A  N  Q  T  C  A  R  C  L  E  V  S  I     F1

 61 ATTCCGCAGCTGGGCCTGGCGGCGAACCAGACCTGCGCGCGCTGCCTGGAAGTGAGCATT 120

    ----:----|----:----|----:----|----:----|----:----|----:----|

 61 TAAGGCGTCGACCCGGACCGCCGCTTGGTCTGGACGCGCGCGACGGACCTTCACTCGTAA 120


   AGCGATGGCCTGT- 3'

    S  D  G  L  F  L  S  L  G  L  V  S  L  V  E  N  A  L  V  V     F1

121 AGCGATGGCCTGTTTCTGAGCCTGGGCCTGGTGAGCCTGGTGGAAAACGCGCTGGTGGTG 180

    ----:----|----:----|----:----|----:----|----:----|----:----|

121 TCGCTACCGGACAAAGACTCGGACCCGGACCACTCGGACCACCTTTTGCGCGACCACCAC 180



    A  T  I  A  K  N  R  N  L  H  S  P  M  Y  C  F  I  C  C  L     F1

181 GCGACCATTGCGAAAAACCGCAACCTGCATAGCCCGATGTATTGCTTTATTTGCTGCCTG 240

    ----:----|----:----|----:----|----:----|----:----|----:----|

181 CGCTGGTAACGCTTTTTGGCGTTGGACGTATCGGGCTACATAACGAAATAAACGACGGAC 240

    3'-CTGGTAACGCTTTTTGGCGT-5'
```

```
    A  L  S  D  L  L  V  S  G  S  N  V  L  E  T  A  V  I  L  L     F1
241 GCGCTGAGCGATCTGCTGGTGAGCGGCAGCAACGTGCTGGAAACCGCGGTGATTCTGCTG 300
    ----:----|----:----|----:----|----:----|----:----|----:----|
241 CGCGACTCGCTAGACGACCACTCGCCGTCGTTGCACGACCTTTGGCGCCACTAAGACGAC 300

    L  E  A  G  A  L  V  A  R  A  A  V  L  Q  Q  L  D  N  V  I     F1
301 CTGGAAGCGGGCGCGCTGGTGGCGCGCGCGGCGGTGCTGCAGCAGCTGGATAACGTGATT 360
    ----:----|----:----|----:----|----:----|----:----|----:----|
301 GACCTTCGCCCGCGCGACCACCGCGCGCGCCGCCACGACGTCGTCGACCTATTGCACTAA 360

    D  V  I  T  C  S  S  M  L  S  S  L  C  F  L  G  A  I  A  V     F1
361 GATGTGATTACCTGCAGCAGCATGCTGAGCAGCCTGTGCTTTCTGGGCGCGATTGCGGTG 420
    ----:----|----:----|----:----|----:----|----:----|----:----|
361 CTACACTAATGGACGTCGTCGTACGACTCGTCGGACACGAAAGACCCGCGCTAACGCCAC 420

    D  R  Y  I  S  I  F  Y  A  L  R  Y  H  S  I  V  T  L  P  R     F1
421 GATCGCTATATTAGCATTTTTTATGCGCTGCGCTATCATAGCATTGTGACCCTGCCGCGC 480
    ----:----|----:----|----:----|----:----|----:----|----:----|
421 CTAGCGATATAATCGTAAAAAATACGCGACGCGATAGTATCGTAACACTGGGACGGCGCG 480

    A  R  R  A  V  A  A  I  W  V  A  S  V  V  F  S  T  L  F  I     F1
481 GCGCGCCGCGCGGTGGCGGCGATTTGGGTGGCGAGCGTGGTGTTTAGCACCCTGTTTATT 540
    ----:----|----:----|----:----|----:----|----:----|----:----|
481 CGCGCGGCGCGCCACCGCCGCTAAACCCACCGCTCGCACCACAAATCGTGGGACAAATAA 540

       5’-TATGATCATGTGGCGGTGCT-3'
    A  Y  Y  D  H  V  A  V  L  L  C  L  V  V  F  F  L  A  M  L     F1
541 GCGTATTATGATCATGTGGCGGTGCTGCTGTGCCTGGTGGTGTTTTTTCTGGCGATGCTG 600
    ----:----|----:----|----:----|----:----|----:----|----:----|
541 CGCATAATACTAGTACACCGCCACGACGACACGGACCACCACAAAAAAGACCGCTACGAC 600
```

```
    V  L  M  A  V  L  Y  V  H  M  L  A  R  A  C  Q  H  A  Q  G     F1
601 GTGCTGATGGCGGTGCTGTATGTGCATATGCTGGCGCGCGCGTGCCAGCATGCGCAGGGC 660
    ----:----|----:----|----:----|----:----|----:----|----:----|
601 CACGACTACCGCCACGACATACACGTATACGACCGCGCGCGCACGGTCGTACGCGTCCCG 660

    I  A  R  L  H  K  R  Q  R  P  V  H  Q  G  F  G  L  K  G  A     F1
661 ATTGCGCGCCTGCATAAACGCCAGCGCCCGGTGCATCAGGGCTTTGGCCTGAAAGGCGCG 720
    ----:----|----:----|----:----|----:----|----:----|----:----|
661 TAACGCGCGGACGTATTTGCGGTCGCGGGCCACGTAGTCCCGAAACCGGACTTTCCGCGC 720

    V  T  L  T  I  L  L  G  I  F  F  L  C  W  G  P  F  F  L  H     F1
721 GTGACCCTGACCATTCTGCTGGGCATTTTTTTTCTGTGCTGGGGCCCGTTTTTTCTGCAT 780
    ----:----|----:----|----:----|----:----|----:----|----:----|
721 CACTGGGACTGGTAAGACGACCCGTAAAAAAAAGACACGACCCCGGGCAAAAAAGACGTA 780

    L  T  L  I  V  L  C  P  E  H  P  T  C  G  C  I  F  K  N  F     F1
781 CTGACCCTGATTGTGCTGTGCCCGGAACATCCGACCTGCGGCTGCATTTTTAAAAACTTT 840
    ----:----|----:----|----:----|----:----|----:----|----:----|
781 GACTGGGACTAACACGACACGGGCCTTGTAGGCTGGACGCCGACGTAAAAATTTTTGAAA 840

    N  L  F  L  A  L  I  I  C  N  A  I  I  D  P  L  I  Y  A  F     F1
841 AACCTGTTTCTGGCGCTGATTATTTGCAACGCGATTATTGATCCGCTGATTTATGCGTTT 900
    ----:----|----:----|----:----|----:----|----:----|----:----|
841 TTGGACAAAGACCGCGACTAATAAACGTTGCGCTAATAACTAGGCGACTAAATACGCAAA 900
              3'-GCGACTAATAAACGTTGCGC-5'

    H  S  Q  E  L  R  R  T  L  K  E  V  L  T  C  S  W  *           F1
901 CATAGCCAGGAACTGCGCCGCACCCTGAAAGAAGTGCTGACCTGCAGCTGGTGA 954
    ----:----|----:----|----:----|----:----|----:----|----
901 GTATCGGTCCTTGACGCGGCGTGGGACTTTCTTCACGACTGGACGTCGACCACT 954
```

Figure 11: Reading frame F1 of the protein sequence with the primers designed to do the PCR. The amino acid and nucleotides highlighted in blue are the regions which will be amplified.

Two primers were designed. Primer 1 has a length of 20 bases both the forward and reverse primer. The forward primer goes from position 114 to 202. It has a 55% GC content, a melting

temperature (Tm) of 60.53°C, and the following sequence (5'->3') GAGCATTAGCGATGGCCTGT. The reverse primer starts in the 202nd position, and it finishes in the 183rd position. It has a 50% GC content, a melting temperature (Tm) of 60.32 °C, and the following sequence (5'->3') TGCGGTTTTTCGCAATGGTC.

Primer 2 has a length of 20 bases both the forward and reverse primer. The forward primer goes from position 547 to 566. It has a 50% GC content, a melting temperature (Tm) of 59.53°C, and the following sequence (5'->3') TATGATCATGTGGCGGTGCT. The reverse primer starts in the 854th position and it finishes in the 873rd position. It has a 50% GC content and a melting temperature (Tm) of 59.17 °C, and the following sequence (5'->3') CGCGTTGCAAATAATCAGCG.

Summary

To summarize, bioinformatic tools can be used to handle different protein and DNA sequences and make further analysis on them. The homologous protein of the unknown was melanocyte-stimulating hormone receptor protein and further examination was done for characteristics like molecular weight, isoelectronic point, and 3D structure. Regarding the PCR primers, if the experiment were to be carried out, it would be necessary to check whether the primers are suitable to make a good PCR. Further research can be done by checking more PCR primers and doing more analysis on the protein functions and their intervention in different signalling pathways.

References

Alejandro A. Schaffer, L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements", Nucleic Acid. Res. 29:2994-3005.

European Bioinformatics Institute (2022) *EMBOSS SixPack.* Available at: https://www.ebi.ac.uk/Tools/st/emboss_sixpack/

European Bioinformatics Institute (2022) *EMBOSS Water*, *Pairwise Sequence Alignment.* Available at: www.ebi.ac.uk/Tools/psa/emboss_water/

European Bioinformatics Institute (2022) HMMER, P*rotein sequence vs protein sequence database.* Available at: www.ebi.ac.uk/Tools/hmmer/search/phmmer

European Bioinformatics Institute (2022) *Pfam*. Available at: http://pfam.xfam.org/family/PF00001.24#tabview=tab0

European Bioinformatics Institute (2022) *USA NCBI's Blast tool.* Available at: https://www.ebi.ac.uk/Tools/sss/ncbiblast/

European Bioinformatics Institute (2022) *USA NCBI's Blast tool.* Available at: ncbiblast-I20220502-213028-0866-4635767-p1m
https://swissmodel.expasy.org/repository/uniprot/Q01726?csm=CB67405A562C29B2
https://www.uniprot.org/uniprot/Q01726

PDB ID, D. Sehnal, S. Bittrich, M. Deshpande, R. Svobodová, K. Berka, V. Bazgier, S. Velankar, S.K. Burley, J. Koča, A.S. Rose (2021) doi: 10.1093/nar/gkab314, RCSB PDB. S. Arnott, P.J. Campbell-Smith & R. Chandrasekaran. In Handbook of Biochemistry and Molecular Biology, 3rd ed. Nucleic Acids--Volume II, G.P. Fasman, Ed. Cleveland: CRC Press, (1976). pp. 411-422.

Sequence Manipulation Suit (2020) *Filter Protein.* Available at: https://www.bioinformatics.org/sms2/filter_protein.html

Sequence Manipulation Suit (2020) *Protein Isoelectric Point.* Available at:https://www.bioinformatics.org/sms2/protein_iep.html

Sequence Manipulation Suit (2020) *Protein Molecular Weight.* Available at:https://www.bioinformatics.org/sms2/protein_mw.html

Sequence Manipulation Suit (2020) *Translation Map.* Available at: https://www.bioinformatics.org/sms2/trans_map.html

Swiss model (2021) *Swiss model repository.* Available at: https://swissmodel.expasy.org/repository/uniprot/Q01726?csm=CB67405A562C29B2

U.S National Library of Medicine (no date) *Primer-Blast.* Available at: https://www.ncbi.nlm.nih.gov/tools/primer-blast/primertool.cgi?ctg_time=1651346610&job_key=lJ5L6DJ-P9YY6CXtKI0B31KWEO1_hQvwfg

UniProt (2000) *UniProtKB.* Available at: https://www.uniprot.org/uniprot/Q01726