BIOINF101 – Aufgabe 5:

Nichanok Auevechanichkul (4781404)

Aufgabe 2:

Anhand der Webseite (https://www.ncbi.nlm.nih.gov/nuccore/LC378575.1) lässt sich eine Genomeseqeunz (**Human T-cell leukemia virus type I OATL9B proviral DNA, complete genome)** darunter darstellen.

```
   1 tgacaatgac catgagcccc aaatatcccc cgggggctta gagcctccca gtgaaaaaca
  61 tttccgcgaa acagaagtct gaaaaggtca gggcccagac taaggctctg acgtctcccc
 121 ccggagggac agctcagcac cggctcaggc taggccctga cgtgtccccc tgaagacaaa
 181 tcataagctc agacctccgg gaagccaccg gaaccaccca tttcctcccc atgtttgtca
 241 agccgccctc aggcgttgac gacaacccct cacctcaaaa aacttttcat ggcacgcata
 301 tggctgaata aactaacagg agtctataaa agcgtggaga cagttcagga gggggctcgc
 361 atctctcctt cacgcgcccg ccgccctacc tgaggccgcc atccacgccg gttgagtcgc
 421 gttctgccgc ctcccgcctg tggtgcctcc tgaactgcgt ccgccgtcta ggtaagttta
 481 gagctcaggt cgagaccggg cctttgtccg gcgctccctt ggagcctacc tagactcagc
 541 cggttctcca cgctttgcct gaccctgctt gctcaactct gcgtctttgt ttcgttttct
 601 gttctgcgcc gctacagatc gaaagttccg cccctttccc tttcattcac gactgactgc
 661 cggcttggcc cacggccaag taccggcgac tccgttggct cggagccagc gacagcccat
 721 tctatagcac tctccaggag agaaacttag tacacagttg ggggctcgtc cgggatacga
 781 gcgccccttt attccctagg caatgggcca aatcttttcc cgtagcgcta gccctattcc
 841 gcggccgccc cgggggctgg ccgctcatca ctggcttaac ttcctccaag cggcatatcg
 901 cctagaaccc ggtccctcca gttacgattt ccaccagttg aaaaaatttc ttaaaatagc
 961 tttagaaaca ccggtctgga tctgtcccat taactactcc ctcctagcca gcctactccc
1021 aaaaggatac cccggccggg tgaatgaaat tttacacata ctcatccaaa cccaagccca
1081 gatcccgtcc cgtcccgcgc caccgccgcc gtcatccccc acccacgacc ccccggattc
1141 tgatccacaa atccccccctc cctatgttga gcctacggcc ccccaagtcc ttccagtcat
1201 gcacccacat ggtgcccctc ccaaccatcg cccatggcaa atgaaggacc tacaggccat
1261 taagcaagaa gtctcccaag cagcccctgg gagcccccag tttatgcaga ccatccggct
1321 tgcggtgcag cagtttgacc ccactgccaa agacctccaa gacctcctgc agtacctttg
1381 ctcctccctc gtggcttccc tccatcacca gcagctagat agccttatat cagaggccga
1441 aacccgaggt attacaggtt ataacccctt agccggtccc ctccgtgtcc aagccaacaa
1501 tccacaacaa caaggattaa ggcgagaata ccagcaactc tggctcgccg ccttcgccgc
1561 cctgccaggg agtgccaaag acccttcctg ggcctctatc ctccaaggcc tggaggagcc
1621 ttaccacgcc ttcgtagaac gcctcaacat agctcttgac aatgggctgc cagaaggcac
1681 gcccaaagac cccatcttac gttccttagc ctactccaat gcaaacaaag aatgccaaaa
1741 attactacag gcccgaggac acactaatag ccctctagga gatatgttgc gggcttgtca
1801 gacctggacc cccaaagaca aaaccaaagt gttagttgtc cagcctaaaa aacccccccc
1861 aaatcagccg tgcttccggt gcgggaaagc aggccactgg agtcgggact gcactcagcc
1921 tcgtcccccc cccgggccat gccccctatg tcaagaccca actcactgga agcgagactg
1981 cccccgccta aagcccacta tcccagaacc agagccagag gaagatgccc tcctattaga
2041 cctccccgct gacatcccac acccaaaaaa ctccataggg ggggaggttt aacctccccc
2101 cccacattac agcaagtcct tcttaaccaa gacccagcat ctattctgcc agttataccg
2161 ttagatcccg cccgtcggcc cgtaattaaa gcccaggttg acacccagac cagccaccca
2221 aagactatcg aagctttact agatacagga gcagacatga cagtccttcc gatagccttg
2281 ttctcaagta atactccccc caaaaataca tccgtattag gggcaggggg ccaaacccaa
2341 gatcacttta agctcacctc ccttcctgtg ctaatacgcc tccctttccg gacaacgcct
2401 attgttttaa catcttgcct agttgatatc aaaaacaact gggccatcat aggtcgcgat
2461 gccttacaac aatgccaggg cgtcctgtac ctccctgagg caaaaaggcc gcctgtaatc
2521 ttgccaatac aggcgccagc cgtccttggg ctagaacacc tcccaaggcc ccccgaaatc
2581 agccagttcc ctttaaacca gaacgcctcc aggccttgca acacttggtc cggaaggccc
2641 tggaggcagg ccatatcgaa ccctacaccg gaccaggaaa taacccagta ttcccagtta
```

2701 aaaaggccaa tggaacctgg cgattcatcc acgacctgcg ggccactaac tctctaacca
2761 tagatctctc atcatcttcc cccgggcccc ctgacttgtc cagcctgcca actacactag
2821 cccacttgca aactatagac cttaaagacg ccttttttcca aatcccctta cctaaacagt
2881 tccagcccta ctttgctttc actgtcccac agcagtgtaa ctacggcccc ggcactagat
2941 acgcctggaa ggtactaccc caagggttta aaaatagtcc caccctgttc gaaatgcagc
3001 tggcccatat cctgcagccc attcggcaag ctttccccca atgcactatt cttcagtaca
3061 tggatgacat tctcctggca agcccctccc atgaggacct actactactc tcagaggcca
3121 caatggcttc cctaatctcc catgggttgc ctgtgtccga aaacaaaacc cagcaaaccc
3181 ctggaacaat taagttccta gggcaaataa tttcacccaa ccacctcact tatgatgcag
3241 tccccacggt acctatacgg tcccgctggg cgctacctga acttcaagcc ctacttggcg
3301 agattcagtg ggtctccaag ggaactccta ccttacgcca gccccttcac agtctctact
3361 gtgccttaca aaggcatact gatccccgag accaaatata tttaaatcct tctcaagttc
3421 aatcattagt gcagctgcgg caggccctgt cacagaactg ccgcagtaga ctagtccaaa
3481 ccctgccccct cctaggggct attatgctga ccctcactgg caccactact gtagtgttcc
3541 agtccaagca gcagtggcca cttgtctggc tacatgcccc cctaccccac actagccagt
3601 gcccctgggg gcagctactt gcctcagctg tgttattact cgacaaatac accttgcaat
3661 cctatgggct actctgccaa accatacatc ataacatctc cacccaaacc ttcaaccaat
3721 tcattcaaac atctgaccac cccagtgttc ctatcttact ccaccacagt caccgattca
3781 aaaatttagg tgcccaaact ggagaacttt ggaacacttt tcttaaaaca gctgccccat
3841 tggctcctgt aaaagccctc atgccagtgt ttactctttc cccggtgatc ataaacaccg
3901 ccccctgcct gttttcagac ggatctacct cccgggcagc ctatattctc tgggacaagc
3961 atacattgtc acaaagatca ttccccccttc cgccaccgca caagtcggcc caacgggccg
4021 aacttctcgg acttttgcat ggcctttcca gcgcccgttc gtggcgctgt ctcaacatat
4081 ttctagactc caagtatctt tatcattacc ttcggaccct tgccctgggc accttccaag
4141 gcaggtcctc tcaggccccc tttcaggccc tcctgccccg cttactatcg cgtaaggtcg
4201 tctatttgca ccacgttcgc agccatacca atctacctga tcccatctcc aggctcaacg
4261 ctctcacaga tgccctacta atcacccctg tcctgcagct ctctcctgca gaactacaca
4321 gtttcaccca ttgcggacag acggccctca cattgcaagg ggcaaccaca actgaggctt
4381 ccaatatcct gcgctcttgc cacgcctgcc gcaaaaataa cccacaacat cagatgcctc
4441 ggggacacat ccgccgtggc ctacttccta accacatctg gcaaggcgac attacccatt
4501 tcaaatataa aaatacgctg taccgccttc atgtatgggt agacaccttt tcaggagcca
4561 tctcagctac ccaaaagaga aaagaaacaa gctcagaagc tatttcctct ttgcttcagg
4621 ccattgccta tctaggcaag cctagctaca taaacacaga caacggccct gcctatattt
4681 cccaagactt cctcaatatg tgtacctccc ttgctattcg acatactacc catgtcccct
4741 acaatccaac cagctcagga cttgtagaac gctctaatgg cattcttaaa accctattat
4801 ataagtactt tactgacaaa cccgacctac ccatggataa tgctctatcc atagccctat
4861 ggacaatcaa ccacctgaat gtgttaacca actgccacaa aacccgatgg cagcttcacc
4921 actccccccg actccagccg atcccagaga cacattccct cagcaataaa caaacccatt
4981 ggtattattt caagcttcct ggtcttaata gccgccagtg gaaaggacca caggaggctc
5041 tccaagaagc tgccggcgct gctctcatcc cggtaagcgc tagttctgcc cagtggatcc
5101 cgtggagact cctcaagcga gctgcatgcc caagacccgt cggaggcccc gccgatccca
5161 aagaaaaaga ccaccaacac catgggtaag tttctcgcca ctttgatttt attcttccag
5221 ttctgccccc tcatcctcgg tgattacagc cccagctgct gtactctcac aattggagtc
5281 tcctcatacc actctaaacc ctgcaatcct gcccagccag tttgttcgtg gaccctcgac
5341 ctgccggccc tttcagcaga tcaggcccta cagcccccct gccctaatct agtaagttac
5401 tccagctacc atgccaccta ttccctatat ctattccctc attggattaa aaagccaaac
5461 cgaaatggcg gaggctatta ttcagcctct tattcagacc cttgttcctt aaagtgccca
5521 tacctggggt gccaatcatg gacctgcccc tatacaggag ccgtctccag cccctactgg
5581 aagtttcagc aagatgtcaa ttttactcaa gaagtttcac gcctcaatat taatctccat
5641 ttttcgaaat gcggttttcc cttctccctt ctagtcgacg ctccaggata tgaccccatc
5701 tggttcctta ataccgaacc cagccaactg cctcccaccg cccctcctct actccccccac
5761 tctaacctag accacatcct cgagccctct ataccatgga aatcaaaact cctgaccctt
5821 gtccagttaa ccctacaaag cactaattat acttgcattg tctgtatcga tcgtgccagc
5881 ctatccactt ggcacgtcct atactctccc aacgtctctg ttccatcctc ttcttctacc
5941 cccctccttt acccatcgtt agcgcttcca gccccccacc tgacgttacc atttaactgg
6001 acccactgct ttgaccccca gattcaagct atagtctcct ccccctgtca taactccctc
6061 atcctgcccc ccttttcctt gtcacctgtt cccaccctag gatcccgctc ccgccgagcg
6121 gtaccggtgg cggtctggct tgtctccgcc ctggccatgg gagccggggt ggctggcggg

6181 attaccggct ccatgtccct cgcctcagga aagagcctct tacatgaggt ggacaaagat
6241 atttcccaat taactcaagc aatagtcaaa aaccacaaaa atctactcaa aattgcgcag
6301 tatgctgccc agaacagacg aggccttgat ctcctgttct gggagcaagg aggattatgc
6361 aaagcattac aagaacagtg ctgttttctg aatattacta attcccatgt ctcaatacta
6421 caagagagac ccccccctgga gaatcgagtc ctgactggct ggggccttaa ctgggacctt
6481 ggcctctcac agtgggctcg agaggcctta caaactggaa tcacccttgt cgcgctactc
6541 cttcttgtta tccttgcagg accatgcatc ctccgtcagc tacgacacct cccctcgcgc
6601 gtcagatacc cccattactc tcttataaac cctgagtcat ccctgtaaac caagcacata
6661 attattgcaa ccacatcgcc tccagcctcc cctgccaata attaacctct cccattaaat
6721 cctccttctc ctgcagcaac ttcctccgtt cagcctccaa ggactccacc tcgccttcca
6781 actgtctagt atagccatca atccccaact cctgcatttt ttctttccta gcactatgct
6841 gtttcgcctt ctcagcccct tgtctccact tgcgctcacg gcgctcctgc tcttcctgct
6901 ttctccgggc gacgtcagcg gccttcttct ccgcccgcct cctgcgccgt gccttctcct
6961 cttccttcct tttcaaatac tcagcaatct gcttttcctc ctctttctcc cgctcttttt
7021 ttcgcttcct cttctcctca gcccgtcgct gccgatcacg atgcgtttcc ccgcgaggtg
7081 gcgctttctc ccctggaggg ccccgtcgca gccggccgcg gctttcctct tctaaggata
7141 gcaaaccgtc aagcacagct tcctcctcct ccttgtcctt taactcttcc tccaaggata
7201 atagcccgtc caccaattcc tccaccagca ggtcctccgg gcatggcaca ggcaagcatc
7261 gaaacagccc tacagataca aagttaacca tgcttattat cagcccactt cccagggttt
7321 ggacagagcc ttcttttcgg atacccagtc tacgtgtttg gagactgtgt acaaggcgac
7381 tggtgcccca tctctggggg actatgttcg gcccgcctac atcgtcacgc cctactggcc
7441 acctgtccag agcatcagat cacctgggac cccatcgatg gacgcgttat cggctcagct
7501 ctacagttcc ttatccctcg actcccctcc ttccccaccc agagaacctc taagaccctc
7561 aaggtcctta ccccgccaat cactcataca acccccaaca ttccaccctc cttcctccag
7621 gccatgcgca aatactcccc cttccgaaat ggatacatgg aacccaccct tgggcagcac
7681 ctcccaaccc tgtcttttcc agaccccgga ctccggcccc aaaacctgta caccctctgg
7741 ggaggctccg ttgtctgcat gtacctctac cagctttccc cccccatcac ctggcccctc
7801 ctgccccacg tgatttttg ccaccccggc cagctcgggg ccttcctcac caatgttccc
7861 tacaagcgaa tagaagaact cctctataaa atttccctta ccacaggggc cctaataatt
7921 ctacccgaag actgtttgcc caccaccctt ttccagcctg ttagggcacc cgtcacgcta
7981 acagcctggc aaaacggcct ccttccgttc cactcaaccc tcaccactcc aggccttatt
8041 tggacattta ccgatggcac gcctatgatt tccgggccct gccctaaaga tggccagcca
8101 tctttagtac tacagtcctc ctcctttata tttcacaaat ttcaaaccaa ggcctaccac
8161 ccctcattcc tactctcaca cggcctcata cagtactctt cctttcataa tttacatctc
8221 ctgtttgaag aatacaccaa catccccatt tctctacttt ttaacgaaaa agaggcagat
8281 gacaatgacc atgagcccca aatatccccc gggggcttag agcctcccag tgaaaaacat
8341 ttccgcgaaa cagaagtctg aaaaggtcag ggcccagact aaggctctga cgtctccccc
8401 cggagggaca gctcagcacc ggctcaggct aggccctgac gtgtcccccc gaagacaaat
8461 cataagctca gacctccggg aagccaccgg aaccacccat ttcctcccca tgtttgtcaa
8521 gccgccctca ggcgttgacg acaacccctc acctcaaaaa acttttcatg gcacgcatat
8581 ggctgaataa actaacagga gtctataaaa gcgtggagac agttcaggag ggggctcgca
8641 tctctccttc acgcgcccgc cgccctacct gaggccgcca tccacgccgg ttgagtcgcg
8701 ttctgccgcc tcccgcctgt ggtgcctcct gaactgcgtc cgccgtctag gtaagtttag
8761 agctcaggtc gagaccgggc ctttgtccgg cgctcccttg gagcctacct agactcagcc
8821 ggttctccac gctttgcctg accctgcttg ctcaactctg cgtctttgtt tcgttttctg
8881 ttctgcgccg ctacagatcg aaagttccgc ccctttccct ttcattcacg actgactgcc
8941 ggcttggccc acggccaagt accggcgact ccgttggctc

Aufgabe 3:

Durch das „Such-Tool: (https://web.expasy.org/cgi-bin/translate/dna2aa.cgi)“ lässt sich die codierten Aminosäuresequenzen aus der ersten 1000 Nukleinsäure alle 6-möglichen Übersetzungs-Frames durchsuchen. Die ersten und zweiten des 5'3' Frames sind darunter zu sehen.

5'3' Frame 1

Met A R I W L N K L T G V Y K S V E T V Q E G A R I S P S R A R R P T **Stop**

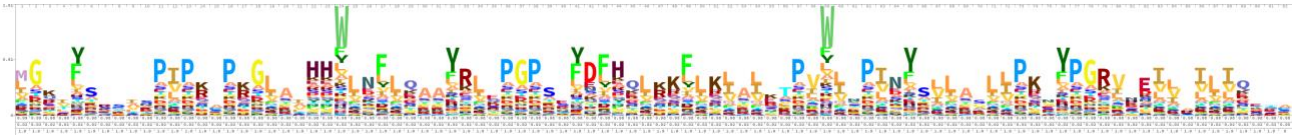
5'3' Frame 2

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

- Nur anhand Genomsequenzen weiß man nicht direkt welche Protein synthetisiert wird. Man muss die exons und introns Bereiche von dem genome wissen und wie die gene (alternativ) gespleißt wird.
- Außerdem, a Genomen pro 3 nucleotide (Codon) zu einem aminosäure entcodiert werden. ist es sinnvoll alle 6 möglichen Übersetzungs-Frames zu durchsuchen.
- Durch Aminosäuresequenzen weiß man direkt, welche Protein daraus synthetisiert wird. Daher wird Aminosäuresequenzen bevorzugt.

Aufgabe 4:

Die Aminosäuresequenzen von dem ersten und zweiten des 5'3' Frames werden in dem Such-Tool nach HMM-Profilen aus den Pfam Datenbank eingegeben. Leider gibt es bei der Sequenz des ersten 5’3’ Frames kein Match (hits = 34 score). Das HMM Logo zu dem Profil des zweiten 5’3’ Frames ist darunter gezeigt (Ref: http://pfam.xfam.org/family/PF02228.15).



Ref: http://pfam.xfam.org/family/PF02228.15#tabview=tab4

Im Vergleich zu der Suchsequenz in der Aufgabe 3, ist zu erkennen, dass die meiste Sequenz identisch zu der höchst vorkommenden Aminosäure (größte Schrift) von der jeweiligen Position ist.

Aufgabe 5:

Gene: **Infectious spleen and kidney necrosis virus strain MV257 major protein capsid (MCP) gene, partial cds**

```
        1 tcacacaagg tgaatctgcc attgatggcc accaatcccc tgtccgaggt gtcactcatt
       61 tacgagaaca ccctcggct ccaccagatg ggagtagact acttcacatc tgtcgacccc
      121 tactactttg cgcccagcat gcctgagatg gatggtgtta tgacctactg ctatacgttg
      181 gacatgggca atatcaaccc catgggttca accaactacg gccgcctgtc caacgtcacc
      241 ctgtcatgta aggtgtcgga caatgcaaag accaccgcgg cgggcggtgg cggcaacggc
      301 tccggctaca cggtggccca aaagtttgaa ctggtcgtt
```

Ref: https://www.ncbi.nlm.nih.gov/nuccore/KY354078.1

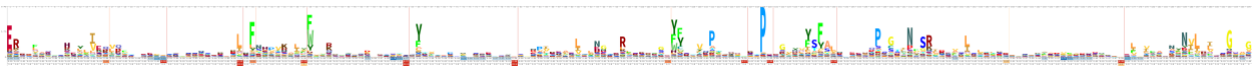
Aminosäuresequenz:

5'3' Frame 1

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

Ref: https://web.expasy.org/cgi-bin/translate/dna2aa.cgi

HMM Profile von der Pfam Databank:



Ref: http://pfam.xfam.org/family/PF04451.11#tabview=tab4

Im Vergleich zu der Suchsequenz aus der ExPASy sind die Sequenz aus beiden Tools ziemlich verschieden. N zum Teil der höchst conserved Aminosäure von beiden Tools sind identisch.