Übungszettel 5

Suche nach der Genomsequenz des Human T-cell leukemia virus type I mit Hilfe des Accession Code

https://www.ncbi.nlm.nih.gov/nuccore/9626453

In der Genomsequenz wird nach der kodierenden Sequenz, CDS (coding sequence) gesucht.

```
LOCUS       LC378575                8980 bp    DNA     linear   VRL 15-MAY-
2018
DEFINITION  Human T-cell leukemia virus type I OATL9B proviral DNA,
complete
            genome.
ACCESSION   LC378575
VERSION     LC378575.1
SOURCE      Human T-cell leukemia virus type I
ORGANISM  Human T-cell leukemia virus type I
            Viruses; Ortervirales; Retroviridae; Orthoretrovirinae;
            Deltaretrovirus.

CODING SEQUENCE
                                    atgggcca 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 ggtgccccctc 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 aa
```

https://www.ncbi.nlm.nih.gov/nuccore/LC378575.1?&feature=CDS

Aus der so erhaltenen CDS wird mittels expasy die Sequenz der codierenden Tripletts in die jeweiligen Aminosäuren umgewandelt.

5' 3' Frame 1

MGQIFSRSASPIPRPPRGLAAHHWLNFLQAAYRLEPGPSSYDFHQLKKFLKIALETPVWICP
INYSLLASLLPKGYPGRVNEILHILIQTQAQIPSRPAPPPPSSPTHDPPDSDPQIPPPYVEP
TAPQVLPVMHPHGAPPNHRPWQMKDLQAIKQEVSQAAPGSPQFMQTIRLAVQQFDPTAKDLQ
DLLQYLCSSLVASLHHQQLDSLISEAETRGITGYNPLAGPLRVQANNPQQQGLRREYQQLWL
AAFAALPGSAKDPSWASILQGLEEPYHAFVERLNIALDNGLPEGTPKDPILRSLAYSNANKE
CQKLLQARGHTNSPLGDMLRACQTWTPKDKTKVLVVQPKKPPPNQPCFRCGKAGHWSRDCTQ
PRPPPGPCPLCQDPTHWKRDCPRLKPTIPEPEPEEDALLLDLPADIPHPKNSIGGEV

Die Suche in AS-Sequenzen verringert die zu verarbeitende Datenmenge um das Dreifache, da in der DNA die jeweilige AS über drei Basen bzw. ein Basentriplett codiert ist.
Außerdem umgeht man einen zusätzlichen Rechenaufwand der für mehrfach codierende Tripletts für die gleiche AS anfällt.

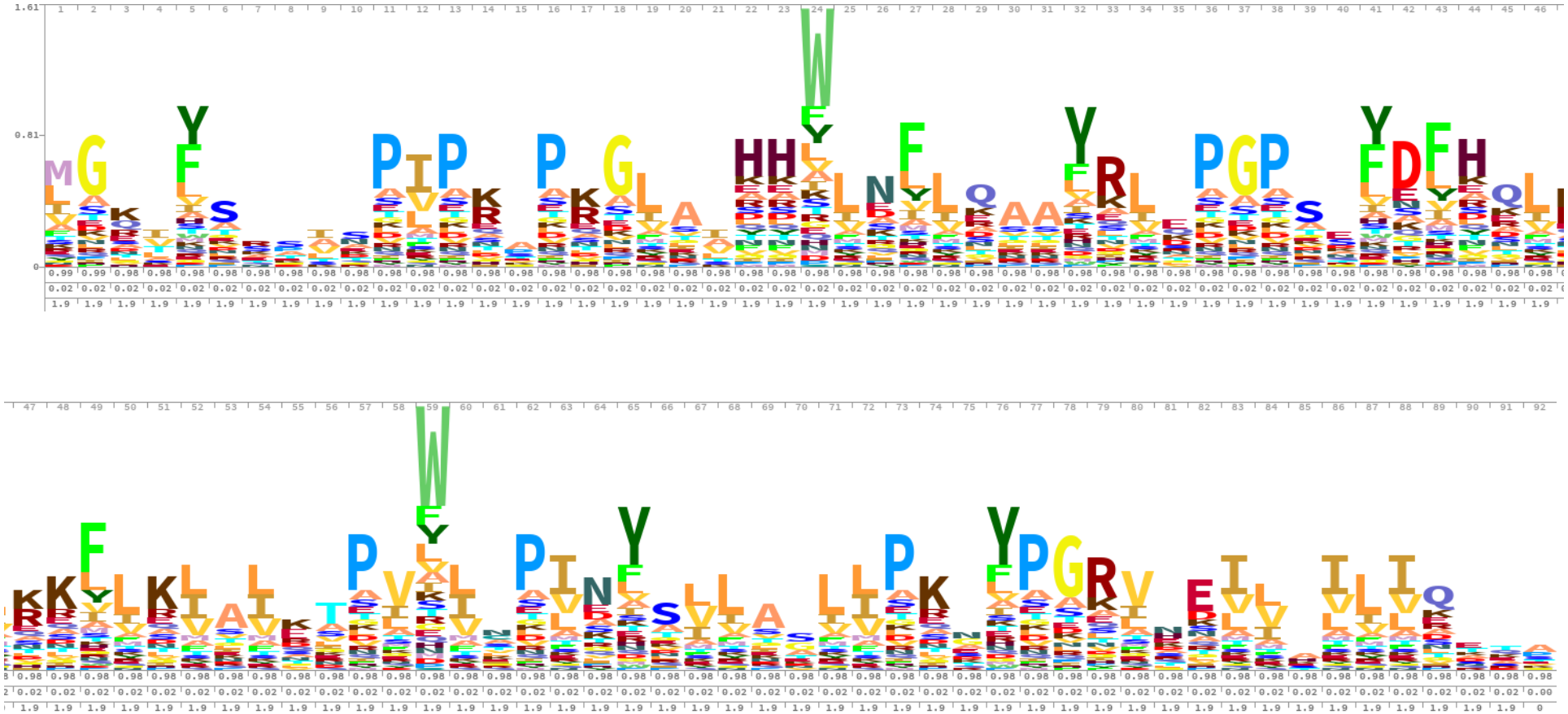
Beispiel Triplett T C X

- X = A
- X = T
- X = C
- X = G

Alle damit möglichen Tripletts codieren für die gleiche AS Serin.

Betrachtet man jetzt das HMM in Bezug auf die DNA-Basen-Wahrscheinlichkeiten, gilt für die statistische Voraussage der einzelnen Basen jeweils ein anderer Wert, obwohl für die gleiche AS codiert wird. An dieser Stelle wird also zusätzlich an Rechenleistung gespart, wenn direkt mit der AS-Sequenz gearbeitet wird.

Die so erhaltene Aminosäuresequenz, der sogenannte open reading frame, wird auf HMM verarbeitet.



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

Vergleicht man die HMM und die expasy-Aminosäure(AS)-Sequenz fällt bereits am Anfang eine hohe Übereinstimmung der beiden Sequenzen auf.
Die gelb unterlegten AS sind in beiden Sequenzoutputs gleich, bzw. bei der HMM an oberster, wahrscheinlichster Stelle.

MGQIFSRSASPIPRPPRGLAAHHWLNFLQAAYRLEPGPSSYDFHQLKKFLKIALETPVWICP

Gal4
Gal4 coding-sequence vom Vector pRmHa3-GAL4
https://dgrc.bio.indiana.edu/product/View?product=1042

atgaagctactgtcttctatcgaacaagcatgcgatatttgccgacttaaaaagctcaagtgctccaaaga
aaaaccgaagtgcgccaagtgtctgaagaacaactgggagtgtcgctactctcccaaaaccaaaaggtctc
cgctgactagggcacatctgacagaagtggaatcaaggctagaaagactggaacagctatttctactgatt
tttcctcgagaagaccttgacatgattttgaaaatggattctttacaggatataaaagcattgttaacagg
attatttgtacaagataatgtgaataaagatgccgtcacagatagattggcttcagtggagactgatatgc
ctctaacattgagacagcatagaataagtgcgacatcatcatcggaagagagtagtaacaaaggtcaaaga
cagttgactgtatcgattgactcggcagctcatcatgataactccacaattccgttggattttatgcccag
ggatgctcttcatggatttgattggtctgaagaggatgacatgtcggatggcttgcccttcctgaaaacgg
accccaacaataatgggttctttggcgacggttctctcttatgtattcttcgatctattggctttaaaccg
gaaaattacacgaactctaacgttaacaggctcccgaccatgattacggatagatacacgttggcttctag
atccacaacatcccgtttacttcaaagttatctcaataattttcacccctactgccctatcgtgcactcac
cgacgctaatgatgttgtataataaccagattgaaatcgcgtcgaaggatcaatggcaaatccttttaac
tgcatattagccattggagcctggtgtatagaggggaatctactgatatagatgtttttttactatcaaaa
tgctaaatctcatttgacgagcaaggtcttcgagtcaggttccataattttggtgacagccctacatcttc
tgtcgcgatatacacagtggaggcagaaaacaaatactagctataatttcacagcttttccataagaatg
gccatatcattgggcttgaatagggacctcccctcgtccttcagtgatagcagcattctggaacaaagacg
ccgaatttggtggtctgtctactcttgggagatccaattgtccctgctttatggtcgatccatccagcttt
ctcagaatacaatctccttcccttcttctgtcgacgatgtgcagcgtaccacaacaggtcccaccatatat
catggcatcattgaaacagcaaggctcttacaagttttcacaaaaatctatgaactagacaaaacagtaac
tgcagaaaaaagtcctatatgtgcaaaaaaatgcttgatgatttgtaatgagattgaggaggtttcgagac
aggcaccaaagtttttacaaatggatatttccaccaccgctctaaccaatttgttgaaggaacacccttgg
ctatcctttacaagattcgaactgaagtggaaacagttgtctcttatcatttatgtattaagagatttttt
cactaatttacccagaaaaagtcacaactagaacaggatcaaaatgatcatcaaagttatgaagttaaac
gatgctccatcatgttaagcgatgcagcacaaagaactgttatgtctgtaagtagctatatggacaatcat
aatgtcaccccatattttgcctggaattgttcttattacttgttcaatgcagtcctagtacccataaagac
tctactctcaaactcaaaatcgaatgctgagaataacgagaccgcacaattattacaacaaattaacactg
ttctgatgctattaaaaaaactggccacttttaaaatccagacttgtgaaaaatacattcaagtactggaa
gaggtatgtgcgccgtttctgttatcacagtgtgcaatcccattaccgcatatcagttataacaatagtaa
tggtagcgccattaaaaatattgtcggttctgcaactatcgcccaataccctactcttccggaggaaaatg
tcaacaatatcagtgttaaatatgtttctcctggctcagtagggccttcacctgtgccattgaaatcagga
gcaagtttcagtgatctagtcaagctgttatctaaccgtccaccctctcgtaactctccagtgacaatacc
aagaagcacaccttcgcatcgctcagtcacgccttttctagggcaacagcaacagctgcaatcattagtgc
cactgaccccgtctgctttgtttggtggcgccaattttaatcaaagtgggaatattgctgatagctcattg
tccttcactttcactaacagtagcaacggtccgaacctcataacaactcaaacaaattctcaa

Die daraus erhaltene AS-Sequenz (mittels expasy)

5' 3' Frame 1

MKLLSSIEQACDICRLKKLKCSKEKPKCAKCLKNNWECRYSPKTKRSPLTRAHLTEVESRLE
RLEQLFLLIFPREDLDMILKMDSLQDIKALLTGLFVQDNVNKDAVTDRLASVETDMPLTLRQ
HRISATSSSEESSNKGQRQLTVSIDSAAHHDNSTIPLDFMPRDALHGFDWSEEDDMSDGLPF
LKTDPNNNGFFGDGSLLCILRSIGFKPENYTNSNVNRLPTMITDRYTLASRSTTSRLLQSYL
NNFHPYCPIVHSPTLMMLYNNQIEIASKDQWQILFNCILAIGAWCIEGESTDIDVFYYQNAK
SHLTSKVFESGSIILVTALHLLSRYTQWRQKTNTSYNFHSFSIRMAISLGLNRDLPSSFSDS
SILEQRRRIWWSVYSWEIQLSLLYGRSIQLSQNTISFPSSVDDVQRTTTGPTIYHGIIETAR
LLQVFTKIYELDKTVTAEKSPICAKKCLMICNEIEEVSRQAPKFLQMDISTTALTNLLKEHP
WLSFTRFELKWKQLSLIIYVLRDFFTNFTQKKSQLEQDQNDHQSYEVKRCSIMLSDAAQRTV
MSVSSYMDNHNVTPYFAWNCSYYLFNAVLVPIKTLLSNSKSNAENNETAQLLQQINTVLMLL
KKLATFKIQTCEKYIQVLEEVCAPFLLSQCAIPLPHISYNNSNGSAIKNIVGSATIAQYPTL
PEENVNNISVKYVSPGSVGPSPVPLKSGASFSDLVKLLSNRPPSRNSPVTIPRSTPSHRSVT
PFLGQQQQLQSLVPLTPSALFGGANFNQSGNIADSSLSFTFTNSSNGPNLITTQTNSQ

HMM Logo des Gal4-Dimers