

## 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

```

                                atggggcca aatcttttcc cgtagcgcta gccctattcc
841 gcgggcgcgcc cggggggtgg ccgctcatca ctggcttaac ttcctccaag cggcataatcg
901 cctagaaccc ggtccctcca gttacgattt ccaccagttg aaaaaatttc ttaaaatagc
961 tttagaaaca ccggtctgga tctgtcccat taactactcc ctcctagcca gcctactccc
1021 aaaaggatac ccgggcccgg tgaatgaaat ttacacata ctcattccaa cccaagccca
1081 gatcccgctc cgtcccgcgc caccgcccgc gtcattcccc acccacgacc ccccggattc
1141 tgatccacaa atccccctc cctatgttga gcctacggcc cccaagtcc ttccagtcac
1201 gcaccacacat ggtgcccctc ccaaccatcg cccatggcaa atgaaggacc tacaggccat
1261 taagcaagaa gtctcccaag cagcccctgg gagccccag tttatgcaga ccatccggct
1321 tgcggtgcag cagtttgacc ccactgcaa agacctcaa gacctcctgc agtacctttg
1381 ctctccctc gtggcttccc tccatcacca gcagctagat agccttatat cagaggccga
1441 aacccgaggt attacaggtt ataaccctt agccggtccc ctccgtgtcc aagccaacaa
1501 tccacaacaa caaggattaa ggcgagaata ccagcaactc tggctcgccg ccttcgcccgc
1561 cctgccaggg agtgccaaag acccttcctg ggccctctat ctccaaggcc tggaggagcc
1621 ttaccacgcc ttcgtagaac gcctcaacat agctcttgac aatgggctgc cagaaggcac
1681 gcccaaagac cccattctac gttccttagc ctactccaat gcaaacaaag aatgccaaaa
1741 attactacag gcccgaggac acactaatag ccctctagga gatatgttgc gggcttgtca
1801 gacctggacc cccaaagaca aaaccaaagt gttagtgtgc cagcctaaaa aacccccccc
1861 aaatcagccg tgcttcgggt gcgggaaagc aggccactgg agtcgggact gcactcagcc
1921 tcgtcccccc ccggggccat gccccctatg tcaagaccca actcactgga agcgagactg
1981 ccccgcccta aagcccacta tcccagaacc agagccagag gaagatgccc tcctattaga
2041 cctccccgct gacatccac 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 Triplets in die jeweiligen Aminosäuren umgewandelt.

### 5' 3' Frame 1

```
MGQIFSRASPIPRPPRGLAAHHWLNFLQAAYRLEPGPSSYDFHQLKKFLKIALETPVWICP
INYSLLASLLPKGYPGRVNEILHILIQTAQIPSRPAPPPSSPTHDPDSDPQIPPPYVEP
TAPQVLPVMHPHGAPPNHRPWQMKDLQAIKQEVSAAPGSPQFMQTIRLAVQQFDPTAKDLQ
DLLQYLCSSLVASLHHQQLDSLISEAETRGITGYNPLAGPLRVQANNPQQQGLRREYQQLWL
AAFAALPGSAKDPSWASILQGLEEYPYHAFVERLNIALDNGLPEGTPKDPILRSLAYSNANKE
CQKLLQARGHTNSPLGDMLRACQWTWTPKDKTKVLVVQPKKPPPNQPCFRCGKAGHWSRDCTQ
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 Triplets für die gleiche AS anfällt.

---

#### Beispiel Triplett T C X

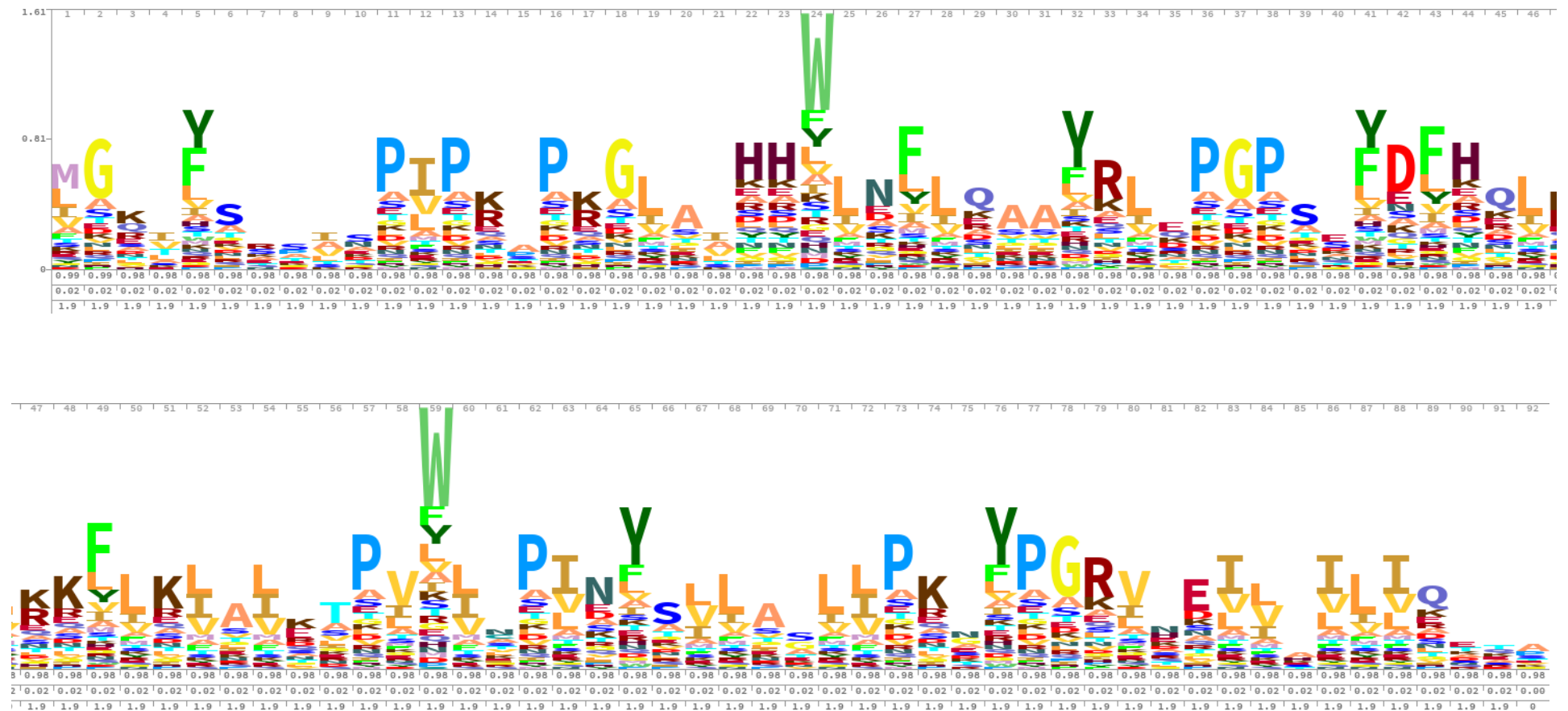
- X = A
- X = T
- X = C
- X = G

---

Alle damit möglichen Triplets 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.

MGQIFSRSA<sup>P</sup>IPRP<sup>P</sup>RGLAA<sup>H</sup>HWLNFLQAAYRLE<sup>P</sup>GPS<sup>S</sup>YDFHQLKKFLKIALE<sup>T</sup>TPVWIC<sup>P</sup>

## Gal4

Gal4 coding-sequence vom Vector pRmHa3-GAL4

<https://dgrc.bio.indiana.edu/product/View?product=1042>

```
atgaagctactgtcttctatcgaacaagcatgcgatatttgccgacttaaaaagctcaagtgtccaaaga
aaaaccgaagtgcgccaagtgtctgaagaacaactgggagtgctcgctactctcccaaaaccaaaggctctc
cgctgactagggcacatctgacagaagtggaatcaaggctagaaagactggaacagctatttctactgatt
tttctcgcagaagaccttgacatgattttgaaaatggattctttacaggatataaaagcattgttaacagg
attatttgtacaagataatgtgaataaagatgccgtcacagatagattggcttcagtggagactgatatgc
ctctaacattgagacagcatagaataagtgcgacatcatcatcggaagagagtagtaacaaagggtcaaaga
cagttgactgtatcgattgactcggcagctcatcatgataactccacaattccggttggaattttatgccag
ggatgctcttcatggatttgattggtctgaagaggatgacatgtcggatggcttgcccttctgaaaacgg
acccaacaataatgggttctttggcgacgggttctctcttatgtattcttcgatctattggctttaaacgg
gaaaattacacgaactctaacgttaacagggtcccgaccatgattacggatagatacacgttggttcttag
atccacaacatccggtttacttcaaagttatctcaataattttcacccctactgccctatcgtgcactcac
cgacgctaataatgatgttgataataaccagattgaaatcgcgctcgaaggatcaatggcaaatccttttaac
tgcatattagccattggagcctggtgtatagagggggaatctactgatataagatgttttttactatcaaaa
tgctaaatctcatttgacgagcaagggtcttcgagtcaggttcataaattttggtgacagccctacatcttc
tgtcgcgatatacacagtggaggcgaaaaacaaataactagctataaattttcacagcttttccataagaatg
gccatatcattgggcttgaatagggacctccctcgtccttcagtgtatagcagcattctggaacaaagacg
ccgaatttggtggtctgtctactcttgggagatccaattgtccctgctttatggctgatccatccagcttt
ctcagaatacaatctccttcccttcttctgtcgcagatgtgcagcgtaccacaacagggtcccaccatata
catggcatcattgaaacagcaaggctcttacaagttttcacaaaaatctatgaactagacaaaacagtaac
tgcagaaaaaagtcctatatgtgcaaaaaaatgcttgatgatttgaatgagattgaggaggtttcgagac
aggcaccaaagtttttacaaatggatatttccaccaccgctctaaccaatttgttgaaggaacacccttgg
ctatcctttacaagattcgaactgaagtggaaacagttgtctcttatcattttatgtattaagagatttttt
cactaattttacccagaaaaagtcacaactagaacaggatcaaatgatcatcaaagttatgaagttaaac
gatgtccatcatgttaagcgatgcagcacaaagaactgttatgtctgtaagtagctatatggacaatcat
aatgtcaccccatattttgcctggaattgttcttattacttgttcaatgcagtccttagtaccataaagac
tctactctcaaactcaaaatcgaatgctgagaataacgagaccgcacaattattacaacaaattaacactg
ttctgatgtatttaaaaaaactggccacttttaaaatccagacttgtgaaaaatacattcaagtactggaa
gaggtatgtgcgcgctttctgttatcacagtgtgcaatccattaccgcataatcagttataacaatagtaa
tggtagcgccattaaaaatattgtcggttctgcaactatcgcccaataccctactcttccggaggaaaatg
tcaacaatatcagtggttaaatatgtttctcctggctcagtagggccttcacctgtgccattgaaatcagga
gcaagtttcagtgtatctagtcaagctgttatctaaccgtccaccctctcgttaactctccagtgaataacc
aagaagcacaccttcgcacgcctcagtcacgccttttctagggcaacagcaacagctgcaatcattagtgc
cactgaccccgctctgctttgtttggtggcgccaatttttaataaagtgggaatattgctgatagctcattg
tccttcactttcactaacagtagcaacggtccgaacctcataacaactcaaacaattctcaa
```

Die daraus erhaltene AS-Sequenz (mittels expasy)

### 5' 3' Frame 1

```
MKLLSSIEQACDICRLKCLKSKEKPKCAKCLKNNWECRYSPKTKRSPLTRAHLTEVESRLE
RLEQLFLLIFPREDLDMILKMDSLQDIKALLTGLFVQDNVNKDAVTDRLASVETDMPLTLRQ
HRISATSSSEESSNKGQRQLTVSIDSAAHHDNSTIPLDFMPRDALHGFWDSEEDDMSDGLPF
LKTDPNNGFFGDGSLLCILRSIGFKPENYTNSNVNRLPTMITDRYTLASRSTTSRLLQSYL
NNFHPYCPIVHSPTLMMLYNNQIEIASKDQWQILFNCILAIGAWCIEGESTDIDVFYYQNAK
SHLTSKVFESGSIILVTALHLLSRYTQWRQKTNTSYNFHSFSIRMAISLGLNRDLPSSFSDS
SILEQRRRIWWSVYSWEIQLSLLYGRSIQLSQNTISFPSSVDDVQRTTTTGPTIYHGIIETAR
LLQVFTKIYELDKTVTAEKSPICAKKCLMICNEIEEVSROAPKFLQMDISTTALTNLLKEHP
WLSFTRFELKWKQLSLIIYVLRDFFTFNFTQKKSQLEQDQNDHQSYEVKRCSIMLSDAAQRTV
MSVSSYMDNHNVTPIFAWNCSYYLFNAVLVPIKTLLSNSKSNNAENNETAQLLQQINTVLMML
KKLATFKIQTCEKYIQVLEEVCAFPLLSQCAIPLPHISYNNNSGSAIKNIVGSATIAQYPTL
PEENVNNISVKYVSPGSPVPLKSGASFSDLVKLLSNRPPSRNSPVTIPRSTPSHRSVT
PFLGQQQQQLQSLVPLTPSALFGGANFNQSGNIADSSLSFTFTNSSNGPNLITTQTNQ
```

