

# Übungszettel 5

## Aufgabe 2

Das Genom des Human T-Cell leukemia virus Type I ist unter dem Eintrag NC\_001436.1 zu finden.

### Human T-lymphotropic virus 1, complete genome

```
>NC_001436.1 Human T-lymphotropic virus 1, complete genome
GGCTCGCATCTCTCCTTCACGCGCCCGCCGCTTACCTGAGGCCGCCATCCACGCCGGTTGAGTCGCGTT
CTGCCGCTCCCGCCTGTGGTGCCTCCTGAACCTACGTCCGCCGTCTAGGTAAGTTTAGAGCTCAGGTCGA
GACCGGGCCTTTGTCCGGCGCTCCCTTGGAGCCTACCTAGACTCAGCCGGCTCTCCACGCTTTGCCTGAC
CCTGCTTGCTCAACTCTACGTCTTTGTTTCGTTTTCTGTTCTGCGCCGTTACAGATCGAAAGTTCCACCC
```



## Aufgabe 3

- Die HMMs werden bevorzugt bei Aminosäuresequenzen angewendet, da sich die Basensequenzen der unterschiedlichen Organismen stark unterscheiden können, obwohl die Aminosäuresequenzen sehr ähnlich sind. Da teilweise mehrere Codons (vor allem durch Variation an der dritten Codonstelle) für dieselbe Aminosäure codieren, können auch unterschiedliche Sequenzen das gleiche oder sehr ähnliche Proteine erzeugen. Dementsprechend variieren die Muster in den Nukleotidsequenzen deutlich stärker als die Muster in den Proteinen. Die Wahrscheinlichkeit, eine Ähnlichkeit festzustellen ist demnach bei Aminosäuresequenzen höher. Ein weiteres Argument ist, dass - zwar nur bei einigen wenigen Organismen – der genetische Code einige wenige Unterschiede aufweist. So codieren nicht alle Tripletts in allen Organismen die gleichen Aminosäuren.
- Die DNA-Sequenz wird in Triplets von Nukleotiden in Aminosäuresequenz translatiert, beginnend bei einem Startcodon (ATG/AUG), das für Methionin codiert. Dadurch ergibt sich zwangsweise für das Protein ein *reading frame*. Es kann also ab jedem Auftreten dieser drei Basen hintereinander, ein Protein codiert sein, unabhängig davon, an welcher Stelle (1., 2., 3.) sich der Beginn ‚A‘ des Codons befindet. Die RNA-Polymerase bindet ebenfalls unabhängig von der Position im Verhältnis zum ‚Start‘ des Genoms. Dementsprechend würden viele Proteine nicht untersucht werden, wenn man nur einen der 6 *reading frames* untersucht. Es ergeben sich drei unterschiedliche *reading frames* auf dem *leading strand* und genauso viele auf dem *lagging strand*. Auf beiden DNA-Strängen kann die Information für ein Protein codiert sein.

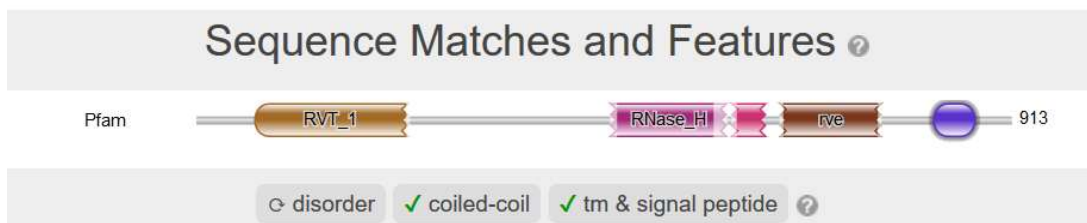
Die gesamte Sequenz wurde im ersten *reading frame* translatiert. Die „–“ stehen dafür für Stoppcodons, bei denen die Translation abbrechen würde. Es wurde für die weitere Untersuchung ein Teil der Sequenz gewählt, bei dem besonders lange kein Stoppcodon auftritt (Aminosäure 716 bis 1611).















Translatierte Sequenz im ersten 5' → 3' reading frame (<https://web.expasy.org/cgi-bin/translate/dna2aa.cgi>):





#### Aufgabe 4:

Für die eingegebene Sequenz konnten folgende Matches mittels HMM gefunden werden:









Pfam Matches														Standard	
Family			Clan	Description	Cross-references	Start	End	Alignment		Model		Bit Score	Domain E-values		
Id	Accession	Start						End	Start	End	Length		Ind.	Cond.	
>	<a href="#">RVT_1</a>	<a href="#">PF00078.26</a>	<a href="#">CL0027</a>	Reverse transcriptase (RNA-dependent DNA polymerase)	 	67	238	67	237	1	221	222	126.50	9.9e-37	3.0e-40
>	<a href="#">rve</a>	<a href="#">PF00665.25</a>	<a href="#">CL0219</a>	Integrase core domain	  	656	767	657	765	4	117	119	94.14	5.8e-27	1.7e-30
>	<a href="#">IN_DBD_C</a>	<a href="#">PF00552.20</a>	n/a	Integrase DNA binding domain	  	827	875	827	875	1	50	50	69.94	7.1e-20	2.1e-23
>	<a href="#">RNase_H</a>	<a href="#">PF00075.23</a>	<a href="#">CL0219</a>	RNase H	  	466	594	470	580	5	128	143	42.75	5.5e-11	1.6e-14
>	<a href="#">Integrase_Zn</a>	<a href="#">PF02022.18</a>	n/a	Integrase Zinc binding domain	  	603	639	605	638	3	37	38	38.35	8.3e-10	2.5e-13
Your search took: 0.1 secs															

Your search took: 0.1 secs

v	RVT_1	PF00078.26	CL0027	Reverse transcriptase (RNA-dependent DNA polymerase)	 	67	238	9.9e-37	3.0e-40
Model	1	ipkkgpsvypslslsvdykalnkliakrLkdvleklisengpgrgrgrstldaveellkalkkkkkaklllklDlkkaF			80				
Query	67	VKKAN-GTWRFI---HDLRATNSLTID-----LSSSPGPPDLSSI-----PTTLAHLQITIDLKDAF			119				
PP		57888.*****8.....8*****777.....56789*****							
Model	81	dsvpleellrkltafkvtkliniksflstrsfsvrvnge.esegryekkglpqGsvlSPllfnlfnmellrelrkra			159				
Query	120	FQIPLPKQFQPYFAFTVPQ-----CNYGPGTRYAWRVLPQGFKNSTLFFEMQLAHLQIPIRQAF			180				
PP		*****77.....33489*****							
Model	160	gvlllrYaDDililiskkeelldelleaveewlkesglikinpeKtklVlfsgkseevkylGvt			221				
Query	181	QCTILQYMDLILASPSHADLQLLSEATMASLISHGLPVSENKTOOT-----PGTIKFLGQI			237				
PP		*****.....9*****86							

This row shows the alignment between your sequence and the matching HMM.

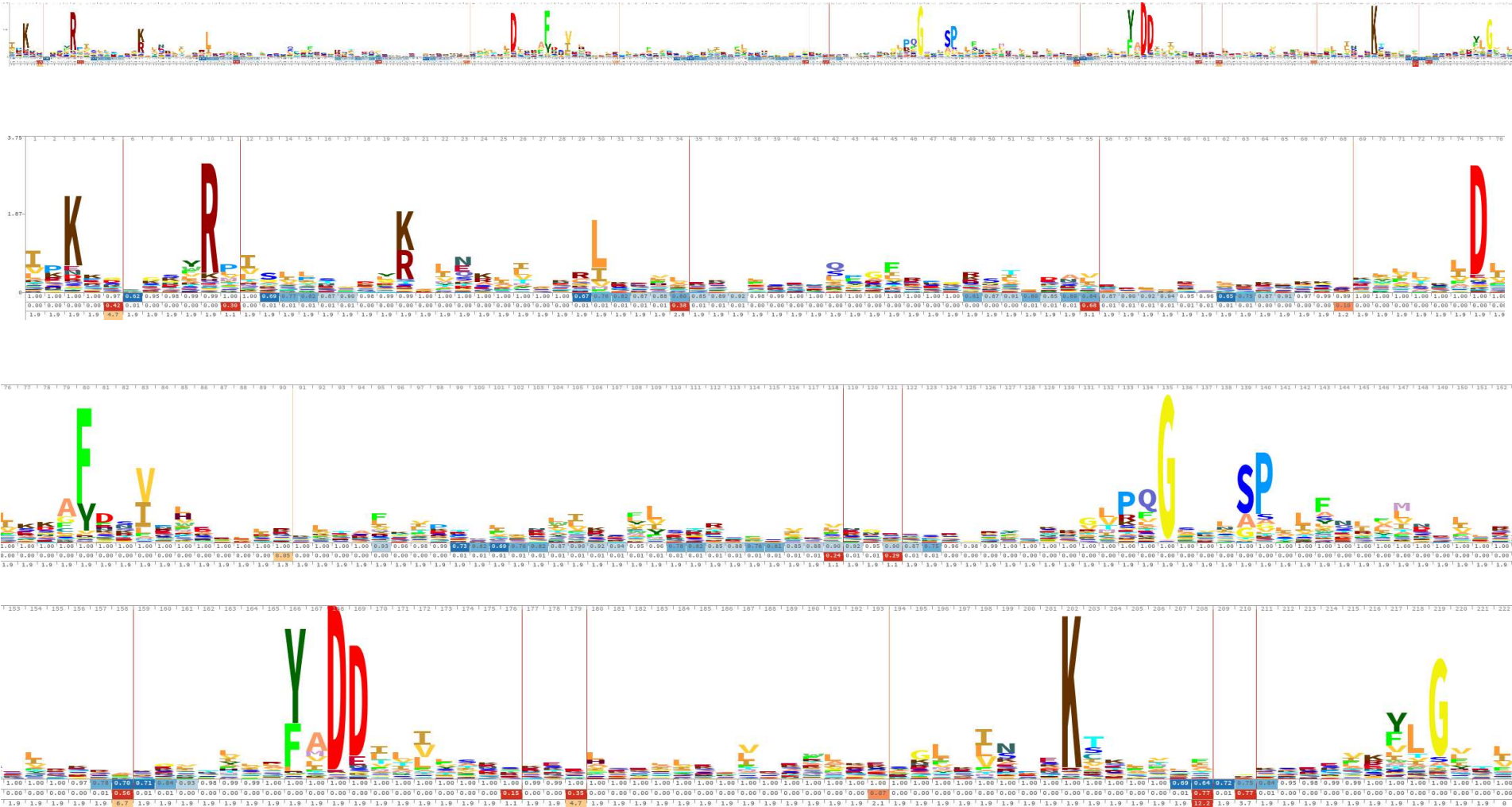
Model:	consensus sequence of the HMM, coloured according to the match: identical residues  Similar residues 
Match line:	the match between the query sequence and the HMM
Query	query sequence, coloured according to the posterior probability: 0%  100%
PP:	posterior probability, or the degree of confidence in each individual aligned residue

> rve	PF00665.25	CL0219	Integrase core domain		656	767	5.8e-27	1.7e-30
> IN_DBD_C	PF00552.20	n/a	Integrase DNA binding domain		827	875	7.1e-20	2.1e-23
> RNase_H	PF00075.23	CL0219	RNase H		466	594	5.5e-11	1.6e-14
Your search took: 0.09 secs	PF02022.18	n/a	Integrase Zinc binding domain		603	639	8.3e-10	2.5e-13

Der erste Treffer ist die Reverse transkriptase, sie hat außerdem den höchsten Bit Score. Die Sequenz stimmt nicht exakt überein, im Alignment sind viele abweichende Aminosäuren (allerdings auch häufig mit ähnlichen Eigenschaften wie die Aminosäuren im Model) zu sehen. An vielen Stellen gibt es große Bereiche, die nicht übereinstimmen, zu Ende der Sequenz hin, stimmt diese häufiger überein. An den hoch konservierten Stellen stimmt die Sequenz sehr gut mit dem Model überein. Eine gleiche oder mindestens sehr ähnliche Funktionsweise der Enzyme ist wahrscheinlich.



Das HMM Logo des ersten Treffers, der Reversen Transkriptase (PF00078.26, RVT\_1 ):



## Aufgabe 5

### Cauliflower mosaic virus, complete genome

NCBI Reference Sequence: NC\_001497.2

```
>NC_001497.2 Cauliflower mosaic virus, complete genome
TGGTATCAGAGCCATGAATCGGTTTAAAGACCAAACTCAAGAGGGTAAAACCTCACCAAAATACGAAAGA
GTTCTTAACTCTAAAAATAAAAGATCTTTCAAGATCAAACATAGTTCCCTCACACCGGTGACCGACAGGA
TTACCACCGTAAGGTTTCAGAACAACATCGAAAGCGTTTACGCCAACTT
```

NC\_001497.2 Cauliflower mosaic virus, complete genome [dna]

1 500 1k 1.5k 2k 2.5k 3k 3.5k 4k 4.5k 5k 5.5k 6k 6.5k 7k 7.5k 8 024

1 500 1k 1.5k 2k 2.5k 3k 3.5k 4k 4.5k 5k 5.5k 6k 6.5k 7k 7.5k 8 024

TGGTATCAGAGCCATGAATCGGTTTAAAGACCAAACTCAAGAGGGTAAAACCTCACCAAAATACGAAAGAGTTCTTAACTCTAAAAATAAAAGATC  
ACCATAGTCTCGGTACTTAGCCAAATCTGGTTTTGAGTTCTCCCATTTTGGAGTGGTTTTATGCTTCTCAAGAATTGAGATTTTTATTTCTAG  
TTTCAAGATCAAACATAGTTCCCTCACACCGGTGACCGACAGGATTACCACCGTAAGGTTTCAGAACAACATCGAAAGCGTTTACGCCAACTTCTGA  
AAAGTTTCTAGTTTGTATCAAGGGAGTGTGGCCACTGGCTGCTCTAATGGTGGCATTCCAAAGTCTTGTGTAGCTTTCGCAAAATGCGGTTGAAGCT  
CTCTCAACTCAAGTCGTACGATGGTAGATCTAAAAGATCAAGACTCTAAGCCTTAAAAATCTTAGATGTTACGAAGCCTTCCTCAGGAAGTA  
GAGAGTTGAGTTCTAGCAGCATGCTACCATCTAGATTTTCTAGTTCTGAGATTCTCGGAATTTTAGAATCTACAATGCTTCGGAAGGAGTCCCTCAT  
CCTTCTGGAACAATAAATCTCTCTGAGAATAGTACTCTATTGAGTATCCACAGGAAAAATAACCTTCTGTGTTGAGATGGATTGTATCCAGAAGA  
GGAAGACCTTGTATTTAGAGAGACTCTTATCATGAGATAACTCATAGGTGTCCTTTTATTGGAAGACACAACCTCTACCTAAACATAGGTCTTCT  
AAATACCCAAAGCGAGCAATTCGAGAATCTGAAAATAATATGCAATATTTAAATCAGAAAATTCGGATGGATTCTCTCCGATCTAATGATCTC  
TTTATGGGTTTCGCTCGTTAGCGTCTTAAGACTTTTATTATACGTTTATAAATTTAGTCTTTTAAGCCTACCTAAGAGGAGGCTAGATTACTAGAG  
AAACGATCAATTAATAAATATCTCTAAAACCAATTAACCTTGGAGAAAGAAAAGATATTTAAATGCCTAACGTTTTATCTCAAGTTATGAAAAA  
TTTGCTAGTTAATTTTTTATAGAGATTTTGGGTTAATTGGAACCTCTTTCTTTCTATAAATTTTACGGATTGCAAAATAGAGTTCAATACTTTTT

Die Sequenz wurde translatiert:

NC\_001436AA [amino]

1 100 200 300 400 500 600 700 800 900 1k 1.1k 1.2k 1.3k 1.4k 1.5k 1.6k 1.7k 1.8k 1.9k 2k 2.1k 2.2k 2.3k 2.4k 2.5k 2.6k 2.7k 2 835

Untersuchte Sequenz (1) Untersuchte Sequenz

1 100 200 300 400 500 600 700 800 900 1k 1.1k 1.2k 1.3k 1.4k 1.5k 1.6k 1.7k 1.8k 1.9k 2k 2.1k 2.2k 2.3k 2.4k 2.5k 2.6k 2.7k 2 835

GRFNLPPHITASFS-PRPNIYSASYTVRSRPSARN-SPD-HPDQPKDYRSSTRYSRHDSPDSLVLK-YSPQKHIRVRGRGPNPRSL-AHLPS  
577 580 585 590 595 600 605 610 615 620 625 630 635 640 645 650 655 660 665 672  
ANTPPFPDDAYCFNILFS-YQQLGHRS-CLTTPRRPVPFP-EKKAACNLANTGASCPWARTPPKAPRNQVPEKPERLOALQHLVRKALEAGH  
673 680 685 690 695 700 705 710 715 720 725 730 735 740 745 750 755 760 765 768  
EEXTGPGNNPVFPVKANGTWREIHDLRATNSLTIDLSSSSPGPPDISLPTTLAHLQITIDLKDAFFOIEPLPKQFOFYFAFTVPOOCNYVGRTRYA  
769 775 780 785 790 795 800 805 810 815 820 825 830 835 840 845 850 855 860 864  
MRVLPQGFKNSTPLFEMOLAHILOPIROAFFOCTILOYMDILLASPSHADLOLLSEATMASLISHGLPVSENKTOOTPGTIKELGQIISNHLTY  
865 870 875 880 885 890 895 900 905 910 915 920 925 930 935 940 945 950 955 960  
DAVPKVPISRWRALPELOALIGEIQWVSKGTPTLRQPLHSLYCALORHTDPRDQIYLNPSQVQSLVQLRQALSQNCPSRIVOTIPLIGAIMLTLTG  
961 965 970 975 980 985 990 995 1k 1005 1010 1015 1020 1025 1030 1035 1040 1045 1050 1056  
TTTVYFQSKQOQWELVWLHAPLEHTSOCPWCOLLASAVILLDRYTLQSYGLICQTIHHNISTQTFNQFIQTSDDHESVPIILHSHREKNLGAQTGEI  
1 057 1065 1070 1075 1080 1085 1090 1095 1.1k 1105 1110 1115 1120 1125 1130 1135 1140 1145 1152  
WNTFLKTTAPLAPVKALMPVETLSPVIINTAPCLFSDGSTSOAAAYILLWDKHILSORSEPLPPPHKSAQRAELGLLHGLSSARSWRCLNIFLPSKY  
1 153 1160 1165 1170 1175 1180 1185 1190 1195 1.2k 1205 1210 1215 1220 1225 1230 1235 1240 1248

Sequence Matches and Features





Plan

Peptidase\_A3 RVT\_1 679

✓ disorder ✓ coiled-coil ✓ tm & signal peptide

Pfam Matches

Standard



Family		Accession	Clan	Description	Cross-references	Start	End	Alignment		Model		Length	Bit Score	Domain E-values	
Id								Start	End	Start	End			Ind.	Cond.
>	Peptidase_A3	PF02160.14	CL0129	Cauliflower mosaic virus peptidase (A3)	 	20	224	20	224	1	205	205	388.89	4.0e-117	4.7e-121
>	RVT_1	PF00078.26	CL0027	Reverse transcriptase (RNA-dependent DNA polymerase)	 	293	452	295	452	3	222	222	124.66	3.6e-36	4.3e-40

Your search took: 0.09 secs

Die HMM-Suche ergab zwei Treffer, die Peptidase hat einen deutlich höheren Bit Score als die RVT (siehe oben).

Pfam Matches

Standard

Family		Accession	Clan	Description	Cross-references	Start	End	Alignment		Model		Length	Bit Score	Domain E-values	
Id								Start	End	Start	End			Ind.	Cond.
>	Peptidase_A3	PF02160.14	CL0129	Cauliflower mosaic virus peptidase (A3)	 	20	224	20	224	1	205	205	388.89	4.0e-117	4.7e-121
<div> <div>Model</div> <div>1</div> <div>nPnsiyykvrlykknnykkelhcyvdtGaalciakkkvipeenWvllkrPkeWkiadkkitiskvkdvdllkhenkifi</div> <div>80</div> </div> <div> <div>Query</div> <div>20</div> <div>nPnsiyykvrlykknnykkelhcyvdtGaalciakkkvipeenWvllkrPkeWkiadkkitiskvkdvdllkhenkifi</div> <div>99</div> </div> <div> <div>PP</div> <div>g*****</div> </div>															
<div> <div>Model</div> <div>81</div> <div>iPiwyigesGldliiGnnflklyePfiqflerwlrkkklyPkeiaklskklvlyktevikeslkkrsktlekalwfkzie</div> <div>160</div> </div> <div> <div>Query</div> <div>100</div> <div>iP+vy+gesG+d+iiGnnfl+lyePfiqf+rvvi+k+k+yP+tiakl+r+v+v+te++tes+kkrskt++e+v++s++</div> <div>179</div> </div> <div> <div>PP</div> <div>*****</div> </div>															
<div> <div>Model</div> <div>161</div> <div>kienpleevailseidrlseeklkiellirKkeellekvtsaiP</div> <div>205</div> </div> <div> <div>Query</div> <div>180</div> <div>KIENPLEETAILSSGRRLSEELFITQQRMQKIEELLEKVCSNP</div> <div>224</div> </div> <div> <div>PP</div> <div>*****g</div> </div>															
>	RVT_1	PF00078.26	CL0027	Reverse transcriptase (RNA-dependent DNA polymerase)	 	293	452	295	452	3	222	222	124.66	3.6e-36	4.3e-40
Your search took: 0.09 secs															
Download your results in various formats															

Die untersuchte Sequenz stimmt in hohem Maße mit dem Model überein (deutlich besser als die RVT in Aufgabe 4). Die Zusätzlich zu den hoch konservierten Stellen, stimmen auch viele Bereiche dazwischen überein oder haben Aminosäuren mit ähnlichen Resten. Es gibt keine Bereiche größerer Abweichungen. Die Sequenz ist dem Model sehr ähnlich.



HMM Logo der Peptidase A 3 (Accession: PF02160.14):

