BIOINF101 - Aufgabe 5

Human T-cell leukemia virus type I (NC_001436)

- Genom - die 100 ersten Nukleotide:

  GGCTCGCATCTCTCCTTCACGCGCCCGCCGCCTTACCTGAGGCCGCCATCCACGCCGGTTGAGTCGCGTTCTGCC
  GCCTCCCGCCTGTGGTGCCTCCTGA

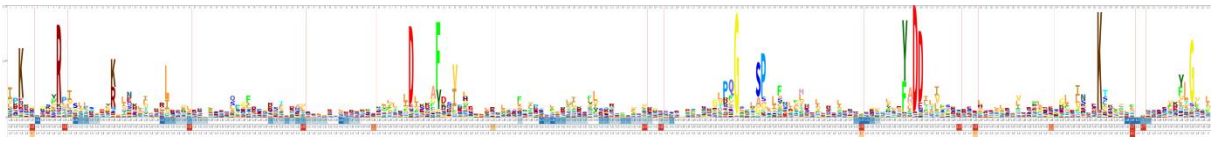
- Peptidsequenz - die ersten 30 Aminosäuren des 1. 5'3' Frames

  GSHLSFTRPPPYLRPPSTPVESRSAASRLW

- Hidden Markov Model Logo

  Reverse transcriptase (RNA-dependent DNA polymerase) (754 - 925)

  http://pfam.xfam.org/family/PF00078.26#tabview=tab4



Einschätzung: die hochkonservierten Bereiche, an denen bestimmte Aminosäuren mit sehr hoher Wahrscheinlichkeit auftreten, sind in der Sequenz vorhanden und große Bereiche des Profils stimmen weitestgehend mit der Sequenz überein, da an vielen Stellen eine Vielzahl verschiedener Aminosäuren vom Profil zugelassen wird.

Human immunodeficiency virus 1 (NC_001802.1)

- Genom - die ersten 100 Nukleotide:

  GGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCCACTGCTTAAGCCTCAAT
  AAAGCTTGCCTTGAGTGCTTCAAGT

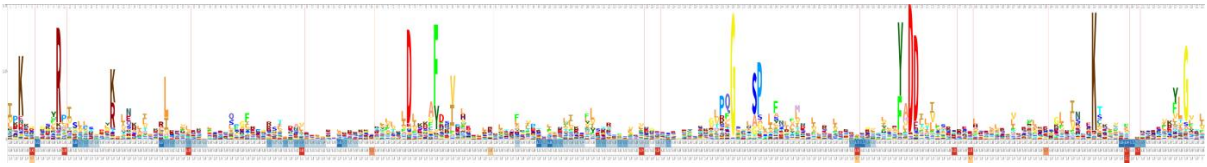
- Peptidsequenz - die ersten 30 Aminosäuren des 2. 5'3' Frames

  VSLVRPDLSLGALWLTREPTA-ASIKLALS

- Hidden Markov Model Logo

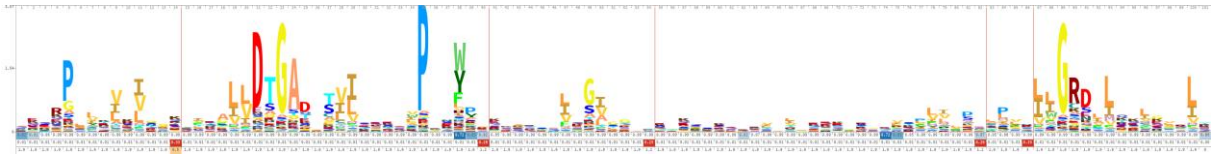
  Reverse transcriptase (RNA-dependent DNA polymerase)

  http://pfam.xfam.org/family/PF00078.26#tabview=tab4



Retroviral aspartyl protease

http://pfam.xfam.org/family/PF00077.19#tabview=tab4



Einschätzung: Aminosäuresequenz des *immunodefiency virus* stimmt mit der Sequenz des *T-cell leukemia virus type I* nur in wenigen, hochkonservierten Sequenzen überein (darunter VLPQG im *RVT_1 profile*), die auch im HMM als konserviert zu erkennen sind. Das HMM lässt an vielen Stellen Variabilität in der As-Sequenz zu. Deshalb passt es zu beiden Sequenzen sehr gut, auch wenn sie sich stark unterscheiden. (siehe *BLAST Results*, oben *T-cell virus* unten *immunodefiency virus*). Mit der *Retroviral aspartyl protease* verhält es sich ähnlich.

Range 1: 685 to 942 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
139 bits(351)	1e-37	Compositional matrix adjust.	83/265(31%)	135/265(50%)	15/265(5%)

```
Query  711  APRNQPVPFKPERLQALQHLVRKALEAGHIEPY--TGPGNNPVFPVKKANGT-WRFIHDL  767
            P+ +  P   E+++AL  +  +  + G I       P N PVF +KK + T WR + D
Sbjct  685  GPKVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTPVFAIKKKDSTKWRKLVDF  744

Query  768  RATNSLTIDLSSSSPGPPDLSSLPTTLAHLQTIDLKDAFFQIPLPKQFQPYFAFTVPQQC  827
            R  N  T D     G P  + L    + +  +D+ DA+F +PL + F+ Y AFT+P
Sbjct  745  RELNKRTQDFWEVQLGIPHPAGLKKKKS-VTVLDVGDAYFSVPLDEDFRKYTAFTIPSIN  803

Query  828  NYGPGTRYAWRVLPQGFKNSPTLFEMQLAHILQPIRQAFPQCTILQYMDDILLASPSHAD  887
            N  PG RY + VLPQG+K SP +F+  +  IL+P R+  P   I QYMDD+ + S    D
Sbjct  804  NETPGIRYQYNVLPQGWKGSPAIFQSSMTKILEPFRKQNPDIVIYQYMDDLYVGS----D  859

Query  888  LQLLSEATMAS-----LISHGLPVSENKTQQTPGTIKFLGQIISPNHLTYDAVPKVPIRS  942
            L++    T        L+  GL   + K Q+ P  + ++G  + P+  T   +P +
Sbjct  860  LEIGQHRTKIEELRQHLLRWGLTTPDKKHQKEPPFL-WMGYELHPDKWTVQPIV-LPEKD  917

Query  943  RWALPELQALLGEIQWVSKGTPTLR  967
             W + ++Q L+G++ W S+  P ++
Sbjct  918  SWTVNDIQKLVGKLNWASQIYPGIK  942
```

Range 2: 267 to 334 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
20.0 bits(40)	0.75	Compositional matrix adjust.	17/68(25%)	29/68(42%)	9/68(13%)

```
Query  656  DAYCFNILP-SYQKQLGHH--RSCLTTMPRRPVPPGKKAACNLANTGASC------PWAR  706
            D + F+I+  S+  +  HH   S  T+     V    + +C + +  ++C      P  R
Sbjct  267  DTHVFSIIRRSHPTRFKHHAKHSGGTSSSHANVKRDHQGSCRMGSASSACRAYCTRPDER  326

Query  707  TPPKAPRN  714
            T  K  RN
Sbjct  327  TKGKHSRN  334
```

Range 3: 936 to 961 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
16.9 bits(32)	6.3	Compositional matrix adjust.	9/26(35%)	11/26(42%)	0/26(0%)

```
Query  816  QPYFAFTVPQQCNYGPGTRYAWRVLP  841
            Q Y    V Q C    GT+     V+P
Sbjct  936  QIYPGIKVRQLCKLLRGTKALTEVIP  961
```

Range 4: 799 to 817 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
16.5 bits(31)	8.6	Compositional matrix adjust.	8/21(38%)	11/21(52%)	2/21(9%)

```
Query  644  LPSCANTPPFPDDAYCFNILP  664
            +PS  N  P      Y +N+LP
Sbjct  799  IPSINNETP--GIRYQYNVLP  817
```

3 a)	3 b)
• Genomsequenzen sind größere Datensätze als Aminosäuresequenzen • es würde viele falschpositive Ergebnisse geben, weil die Suche mit 4 Basen unspezifischer ist als mit 20 Aminosäuren • außerdem würden viele unzutreffende Ergebnisse geliefert werden, da der *frame* nicht bekannt ist (kontextabhängig) • Genomsequenzen weisen nochmal mehr Variabilität auf als Aminosäuresequenzen	• jeder Frame ergibt eine völlig andere Aminosäuresequenz • jeder Frame wird andere HMM-Profile liefern