

Übungszettel 5, Ponimaskine

2) Genom-Sequenz von „Human T-cell leukemia virus type I“ (ersten 100 Basen der coding sequence, CDS):

```
atgggccaaatcttttcccgtagcgctagccctattccgcggccgccccgggggctggccgctcatcactggcttaacttctccaggc  
ggcatatcgcc
```

Die Basenpaarung erfolgt mit A-T und G-C.

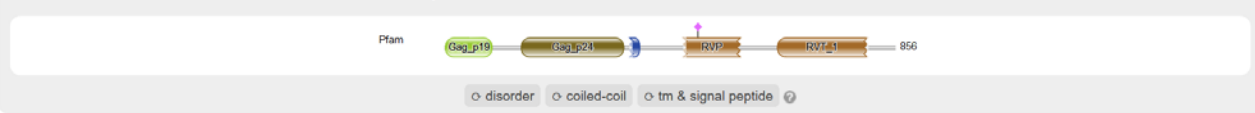
3) Die Gensequenz wurde mittels eines Übersetzungstools in die möglichen Aminosäuresequenzen übersetzt, dabei erhält man sechs mögliche Frames, da sowohl in 5'-3' Richtung, als auch in 3'5' Richtung übersetzt und der Reading-Frame etwas verschoben wird (3 Basenpaare dienen der Übersetzung einer Aminosäure, das sogenannte Codon). Hier beispielhaft die ersten 30 AS der CDS des ersten 5'3' Frames:














MGQIFSRASPIPRPPRGLAAHHWLNFLQA

a) Die Übersetzung der Gensequenz in eine AS-Sequenz ist hoch konserviert, dabei codieren 3 Basen jeweils eine AS, allerdings können auch unterschiedliche Kombinationen der Basen-Triplets dieselbe AS codieren, wie z.B. UCA und UCC Serin entsprechen. Somit muss eine Mutation in der Gensequenz nicht zwingend zu einer Veränderung AS-Sequenz/des Proteins führen und damit liefert auch die Untersuchung der AS-Sequenz möglicherweise eine höhere mögliche Übereinstimmung und ermöglicht es selbst bei Mutationen noch bestimmte Motive zu erkennen. Zusätzlich ist die AS-Sequenz kürzer als die Gensequenz und für die Untersuchung könnten außerdem die 3D Struktur und die Konformation des Proteins beachtet werden.

b) Es ist sinnvoll alle sechs Frames zu überprüfen, da man nicht genau weiß, welcher dieser sechs Frames tatsächlich im Organismus vorkommt.

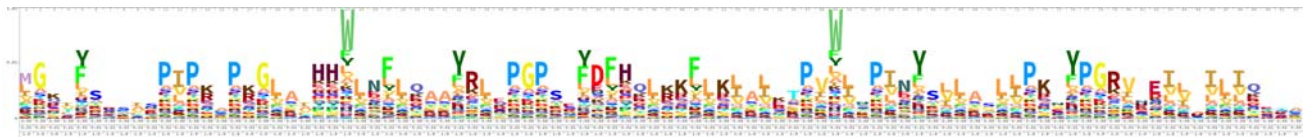
4) Die Suche mittels HMM der ersten 856 AS der CDS ergab folgendes:



Pfam Matches									
								Advanced	
Family								Domain E-values	
Id	Accession	Clan	Description	Cross-references	Start	End	Ind.	Cond.	
> Gag_p24	PF00607.19	CL0148	gag gene protein p24 (core nucleocapsid protein)	  	147	344	9.6e-65	2.9e-68	
> Gag_p19	PF02228.15	CL0074	Major core protein p19	  	1	92	6.0e-57	1.8e-60	
> RVT_1	PF00078.26	CL0027	Reverse transcriptase (RNA-dependent DNA polymerase)	 	633	804	9.2e-37	2.7e-40	
> RVP	PF00077.19	CL0129	Retroviral aspartyl protease	 	457	563	3.4e-15	1.0e-18	
> zf-CCHC	PF00098.22	CL0511	Zinc knuckle	  	355	372	7.2e-07	2.2e-10	

Dabei passen alle gefundenen Proteine anhand ihrer Eigenschaften und Funktionen zu denen eines (Retro) Virus.

Beispielhaft wird das HMM Logo des **Gag_p19** dargestellt:



An fast allen Stellen des Modells sind relativ viele AS wahrscheinlich. Nur einige sind konserviert (z.B. Tryptophan an 24. Und 59. Stelle). Das entsprechende Logo ist recht ähnlich der Suchsequenz (von Stelle 1 bis 92) mit einer Übereinstimmung von 76 %:

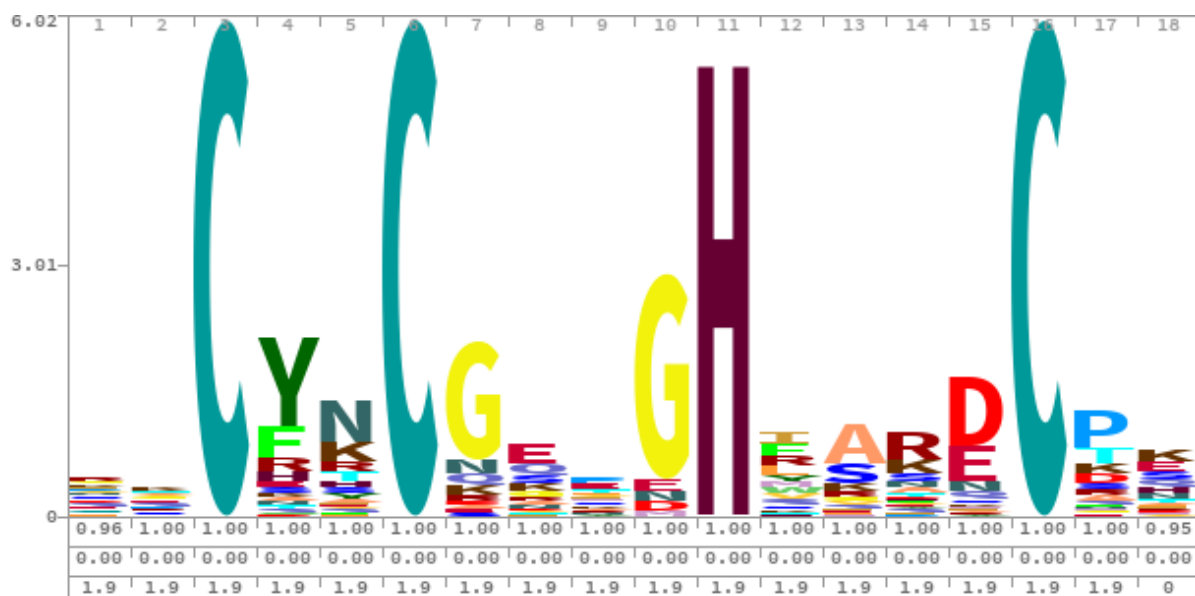
```

.....*.....*.....*.....*.....*.....*.....*.....*.....*
Model      1 mgklysrssispipkapkglaiahwlnflqaayrlepdpseydfhqlkkflklalktpvwnpinysllasllpknyprv 80
mg+i+srs+spip++p+gla+hhwlnflqaayrlepdp+yd fhqlkkflk+al+tpvw++pinysllasllpk+ypgrv
Query      1 MGQIFSRASPIPRPPRGLAAHHWLNFLQAAYRLEPGSSYDFHQLKKFLKIALETVPWICPINYSLASLLPKGYPRV 80
PP
9*****

.....*..
Model      81 neilailliqtg 92
neil+iliq++a
Query      81 NEILHILIQTQA 92
PP
*****996

```

Im Vergleich dazu hat das **Zinkfinger-Motiv** ein deutlich kürzeres Logo, wo allerdings an einigen Stellen bestimmte AS deutlich konserviert sind (z.B. an Stellen 3, 6 und 16 Cystein):



Hier stimmen die Sequenzen des Modells und der Suchsequenz (von Stelle 355 bis 372) weniger überein (nur 43%), allerdings sind die AS an den „entscheidenden“ Stellen (Cystein an 3., 6. und 16. Stelle, Histidin an 11. Stelle, ...) identisch und somit wird das Motiv trotzdem erkannt.

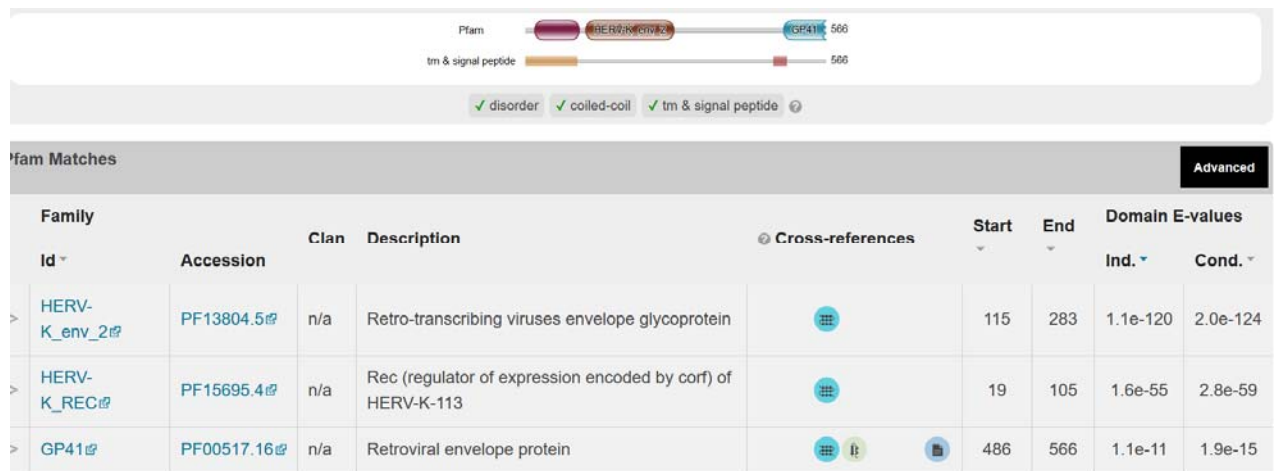
5) Zusätzlich wird die AS-Sequenz des „Human endogenous retrovirus K113“ auf HMM untersucht. Die ersten 100 Basen(paare) der CDS sind dabei:

Atgaacccatcggagatgcaaagaaaagcacctccgcgagacggagacaccgcaatcgagcaccgttgactcacaagatgaac
aaaatggtgacgtcag

Somit sind die ersten 30 AS des 1. 5'3'-Frames:

MNPSEMQRKAPPRRRRHRNRAPLTHKMNMK

Die Suche mittels HMM lieferte folgendes Ergebnis:



The image shows a bioinformatics search result. At the top, there is a Pfam logo for HERV-K env_2 and GP41. Below it, a table titled 'Pfam Matches' lists three protein families: HERV-K env_2, HERV-K_REC, and GP41. The table includes columns for Family, Accession, Clan, Description, Cross-references, Start, End, Domain Ind., and E-values. The HERV-K env_2 family is highlighted in green.

Family	Accession	Clan	Description	Cross-references	Start	End	Domain Ind.	E-values
HERV-K env_2	PF13804.5	n/a	Retro-transcribing viruses envelope glycoprotein		115	283	1.1e-120	2.0e-124
HERV-K_REC	PF15695.4	n/a	Rec (regulator of expression encoded by corf) of HERV-K-113		19	105	1.6e-55	2.8e-59
GP41	PF00517.16	n/a	Retroviral envelope protein		486	566	1.1e-11	1.9e-15

Auch hier stimmen die gefundenen Proteine mit denen eines (Retro)Virus überein.

Folglich beispielhaft das HMM-Logo des **HERV-K env_2**, auch hier sind im Modell relativ viele AS an einer Stelle möglich/wahrscheinlich:



Die Sequenzen stimmen zu 61% (**HERV-K_REC**), zu 41% (**HERV-K env_2**) und zu 28% (**GP41**) überein. Es wird deutlich dass für die Modelle vor allem entscheidend ist, ob stark konservierte AS übereinstimmen.