

Thomas Rinner Bioinformatik Übungszettel 5:

Aufgabe 1:

<https://github.com/rinnerthomas/bioinformatics-BC/tree/master/assignment5>

Aufgabe 2:

Der Datenbank <https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239> wurde die Genomsequenz des Menschlichen T-Zellen Leukämie Virus Typ 1 entnommen (NC_001436), die 8507 Nukleotide umfasst und für 6 Proteine kodiert. Die ersten 100 Basen sind exemplarisch aufgeführt:

ORIGIN¹

```
1  tgacaatgac  catgagcccc  aaatatcccc  cgggggctta  gagcctccca  gtgaaaaaca
61  tttccgcgaa  acagaagtct  gaaaagggtc  gggcccagac  taaggctctg  acgtctcccc
121  cgggagggac  agctcagcac  cggctcaggc  taggcctga  cgtgtccccc  tgaagacaaa
```

Aufgabe 3:

- Die Suche in der Aminosäuresequenz ist der Suche in der Genomsequenz aus verschiedenen Gründen vorzuziehen. Einerseits besitzt die Aminosäuresequenz eine reduzierte Datenmenge, da die Übersetzung des Triplet-Codes die Länge der Sequenz bereits um einen Faktor 3 reduziert. Weiterhin enthält die Aminosäuresequenz bereits die Information für welches Leseraster man sich entschieden hat. Weitere Argumente wären, dass Programme auf Basis der Aminosäuresequenz die degenerierte Eigenschaft des genetischen Codes vernachlässigen können und nicht das Risiko beinhalten, dass der genetische Code, der eigentlich ubiquitär ist, bei wenigen Organismus leicht abweichen kann.
- Je nach den gewählten Leserastern können unterschiedliche Aminosäuresequenzen entstehen, daher wäre es unzureichend nur ein Leseraster zu betrachten. Auf der Suche nach Motiven innerhalb der Aminosäuresequenz sollte man daher auch alle möglichen Aminosäuresequenz betrachten.

Übersetzung der Genomsequenz mit ExPasy²

Bereich 1-756 „5' LTR“ liefert als Aminosäuresequenz im 5'-3' Reading Frame 1:

5'3' Frame 1

-Q-P-APNIPRGLRASQ-KTFPRNRS�KRSQPRRLRL-RLPPEGQLSTGSG-ALTCPPEDKS-AQTSGKPPEPPISSPCLSSRPQALTTTPHLKKLF-MARIWLNKLT
GVYKSVETVQEGARISPSRARRPT-GRHPRRLSRVLPPPACGAS-TASAV-VSLELRSRPGLCALPWSLPRLSRFSTLCLTLAQLCVFVSFVLRRYRSKVPPL
SLSFTTDCRLGPRPSTGDSVGSEPATASIALSRRET-YT

¹ <https://www.ncbi.nlm.nih.gov/nuccore/LC183873.1>

² <https://web.expasy.org/cgi-bin/translate/dna2aa.cgi>

Bereich 802-2092 des Gens ‚gag‘ liefer als Aminosäuresequenz im 5´-3´ Reading Frame1:

```
5'3' Frame 1
MGQIFSRASASPIRPPRGLAAHHLNFLQAAYRLEPGPSSYDFHQLKKFLKIALETVPVWICPINYSLLASLLPKGYPGRVNEILHILITQQAQIPSRPAPPPSSP
THDPPDSDPQIPPPYVEPTAPQVLPVMHPHGAPPNHRPWQMKDLQAIKQEVSAAPGSPQFMQTIRLAVQQFDPTAKDLQDLLQYLCSSIVASLHHQQLDLSISEA
ETRGITGYNPLAGPLRVQANNPQQGLRREYQQLWLAFAALPGSAKDPSPWASILQGLEEPPYHAFVERLNIALDNLPEGTPKDPILRSLAYSANKECCQKLLQAR
GHTNSPLGDMLRACQWTTPKDKTKVLVVQPKKPPNPQPCFRCKAGHWSRDCTQPRPPPGPCPLCQDPHWRKDCPRLKPTIPEPEPEEDALLDLDPADIPHPKNS
IGGEV-
```

Aufgabe 4:

Auf die translatierte Aminosäuresequenz des gag-Gens wurde mit HMM-Profilen aus der Pfam-Datenbank gesucht um die HMM-Logos der Profile zu finden. Die verwendete Aminosäuresequenz lieferte 3 Treffer:

Sequence Matches and Features

Pfam

gag_p24

429

disorder

coiled-coil

tm & signal peptide

Pfam Matches

Advanced

Family	Accession	Clan	Description	Cross-references	Start	End	Domain E-values	
Id							Ind.	Cond.
> Gag_p24	PF00607.19	CL0148	gag gene protein p24 (core nucleocapsid protein)		147	344	7.5e-65	1.3e-68
> Gag_p19	PF02228.15	CL0074	Major core protein p19		1	92	2.1e-57	3.8e-61
> zf-CCHC	PF00098.22	CL0511	Zinc knuckle		355	372	3.3e-07	5.9e-11

Abbildung 1: Treffer der Suche nach HMM-Profilen aus der Pfam-Datenbank in der gag-Genomsequenz des HTLV1

Für jeden der Pfam Matches wurde daraufhin das HMM-Logo ermittelt.

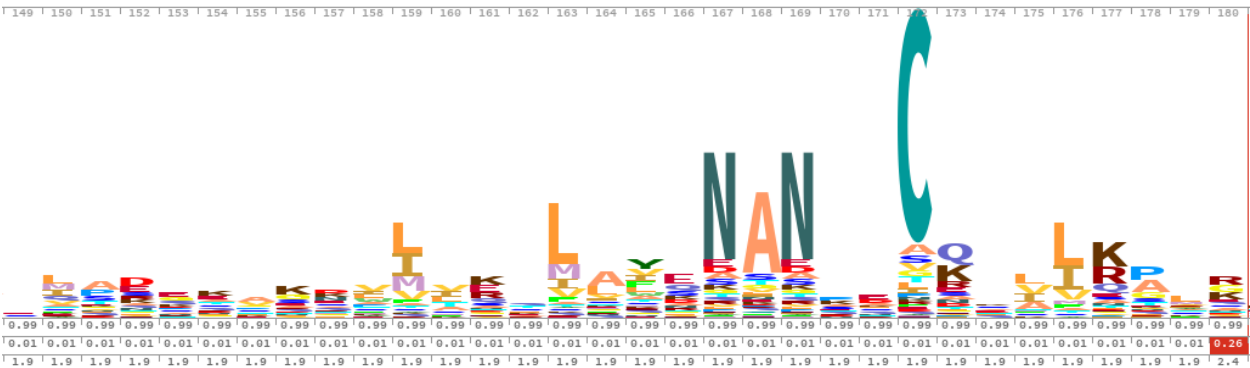


Abbildung 2: Ausschnitt aus dem Gag_p24 HMM-Logo



Suchsequenz Position 355-372: QP**C**FR**C**GKA**G**HWS**R**D**C**T**Q**

Wie zu erwarten treten die zu 100% konservierten Cysteine an Stelle 3, 6 und 16 auf ebenso wie das Histidin an Stelle 11. Ebenso sind die Glycine an Stelle 2 und 10 vorhanden. Die ersten beiden Stellen im Motiv so wie die letzten weisen eine relativ hohe Variabilität auf. In 9 von 18 Stellen trat der Buchstabe der most frequent letter an der erwarteten Stelle auf, wobei in 3 weiteren Fällen (Pos 4, 13 und 14) der zweithäufigste Buchstabe auftrat.

Aufgabe 5: 2. Bericht für Enterobacteria Phage Lambda:

Genom: https://www.ncbi.nlm.nih.gov/nuccore/NC_001416.1

Die ersten 100 Basenpaare:

ORIGIN

```
1  gggcgggcgac  ctgcgggggtt  ttgcgtatatt  atgaaaattt  tccggtttaa  ggcgtttccg
61  ttctttcttcg  tcataactta  atgtttttat  ttaaaatacc  ctctgaaaag  aaaggaaacg
121 acaggtgctg  aaagcgaggc  tttttggcct  ctgtcgtttc  ctttctctgt  ttttgccgt
181 ggaatgaaca  atggaagtca  acaaaaagca  gctggctgac  attttcggtg  cgagtatccg
241 taccattcag  aactggcagg  aacagggaat  gcccgttctg  cgaggcggtg  gcaagggtaa
```

Betrachteter Bereich für die HMM-Profil-Analyse:

CDS

```
complement(28860..29078)
/gene="xis"
/locus_tag="lambdap34"
/codon_start=1
/transl_table=11
/product="Excisionase"
/protein_id="NP_040610.1"
/db_xref="GeneID:2703504"
/translation="MYLTLQEWNARQRRPRSLETVRRWVRECRIFFPPVKDGREYLFH
ESAVKVDLNRPVTGLLKRIRNGKKA"
```

Reading Frame 1:

MYLTLQEWNARQRRPRSLETVRRWVRECRIFFPPVKDGREYLFHESAVKVDLNRPVTGLLKRIRNGKKA

HMM-Profil:




Sequence Matches and Features

Pfam **Exc** 72

☐ disorder ☒ coiled-coil ☒ tm & signal peptide

Pfam Matches

Advanced

Family		Clan	Description	Cross-references	Start	End	Domain E-values		
Id	Accession						Ind.	Cond.	
>	Exc	PF07825.10	CL0123	Excisionase-like protein	  	1	72	2.4e-41	1.5e-45

Your search took: 0.02 secs

Abbildung 5: Treffer der Suche nach HMM-Profilen aus der Pfam-Datenbank in der Genomsequenz (Pos. 28860-29078) der Enterobacteria Phage Lambda

Logo der ersten 34 Positionen³:

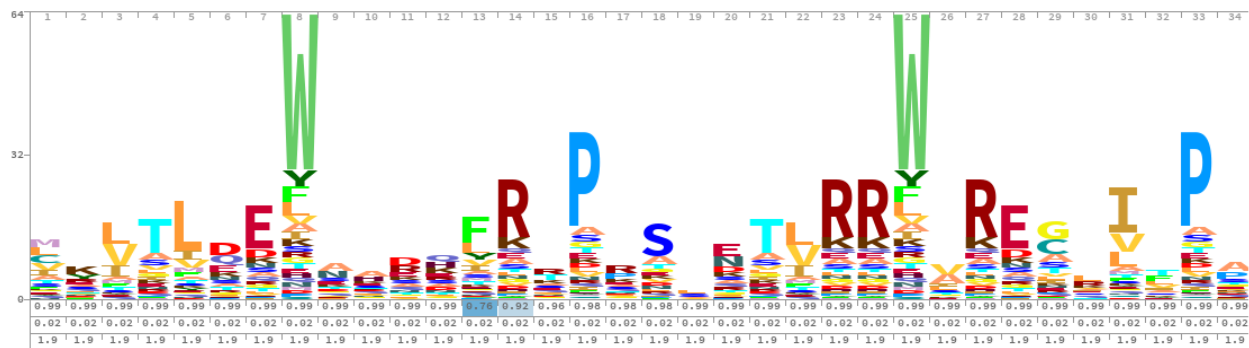


Abbildung 6: Ausschnitt der ersten 34 Positionen des HMM-Logo des Excisionase-Proteins

Sequenzvergleich: MYLTLQEWNARQ - RR - PRSLETVRRWVRECRIFP

Enthielt die Suchsequenz an der betrachteten Position die höchstwahrscheinliche Aminosäure wurde diese in der entsprechenden Farbe gefärbt, handelte es sich um die zweit wahrscheinlichste Aminosäure wurde der Buchstabe der Sequenz unterstrichen. Der durch – eingerahmte Bereich könnte eine Deletion zwischen Position 9-15 beinhalten, wenn man sich an den Bereichen hoher Übereinstimmung orientiert. Denn die Sequenzen stimmen in den hochkonservierten Bereichen 3-8 und 23-29 gut überein, aber nur unter der Annahme, dass in diesem Bereich eine Deletion stattgefunden hat.

³ <http://pfam.xfam.org/family/PF07825.10#tabview=tab4>