**BIOINF101 - Aufgabe 5**

*Human T-cell leukemia virus type I (NC\_001436)*

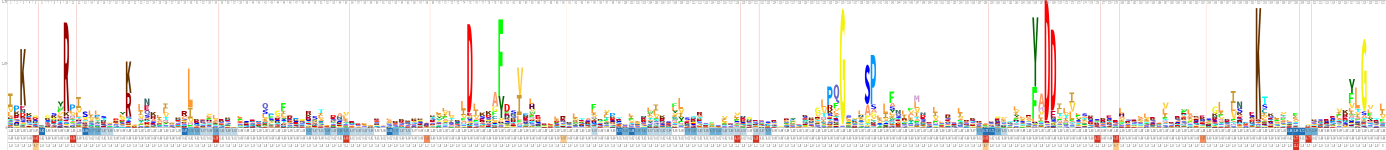
* Genom - die 100 ersten Nukleotide: GGCTCGCATCTCTCCTTCACGCGCCCGCCGCCTTACCTGAGGCCGCCATCCACGCCGGTTGAGTCGCGTTCTGCCGCCTCCCGCCTGTGGTGCCTCCTGA
* Peptidsequenz - die ersten 30 Aminosäuren des 1. 5'3' Frames

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)*

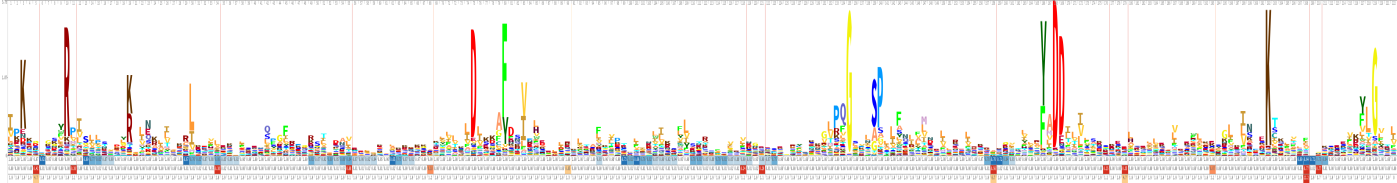
* Genom - die ersten 100 Nukleotide: GGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCCACTGCTTAAGCCTCAATAAAGCTTGCCTTGAGTGCTTCAAGT
* Peptidsequenz - die ersten 30 Aminosäuren des 2. 5'3' Frames

VSLVRPDLSLGALWLTREPTA-ASIKLALS

* Hidden Markov Model Logo

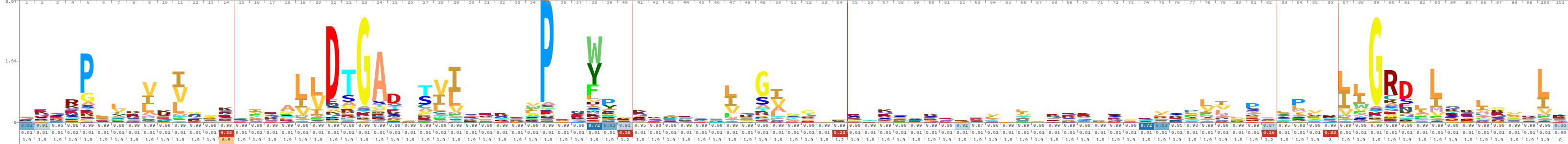
*Reverse transcriptase (RNA-dependent DNA polymerase)*

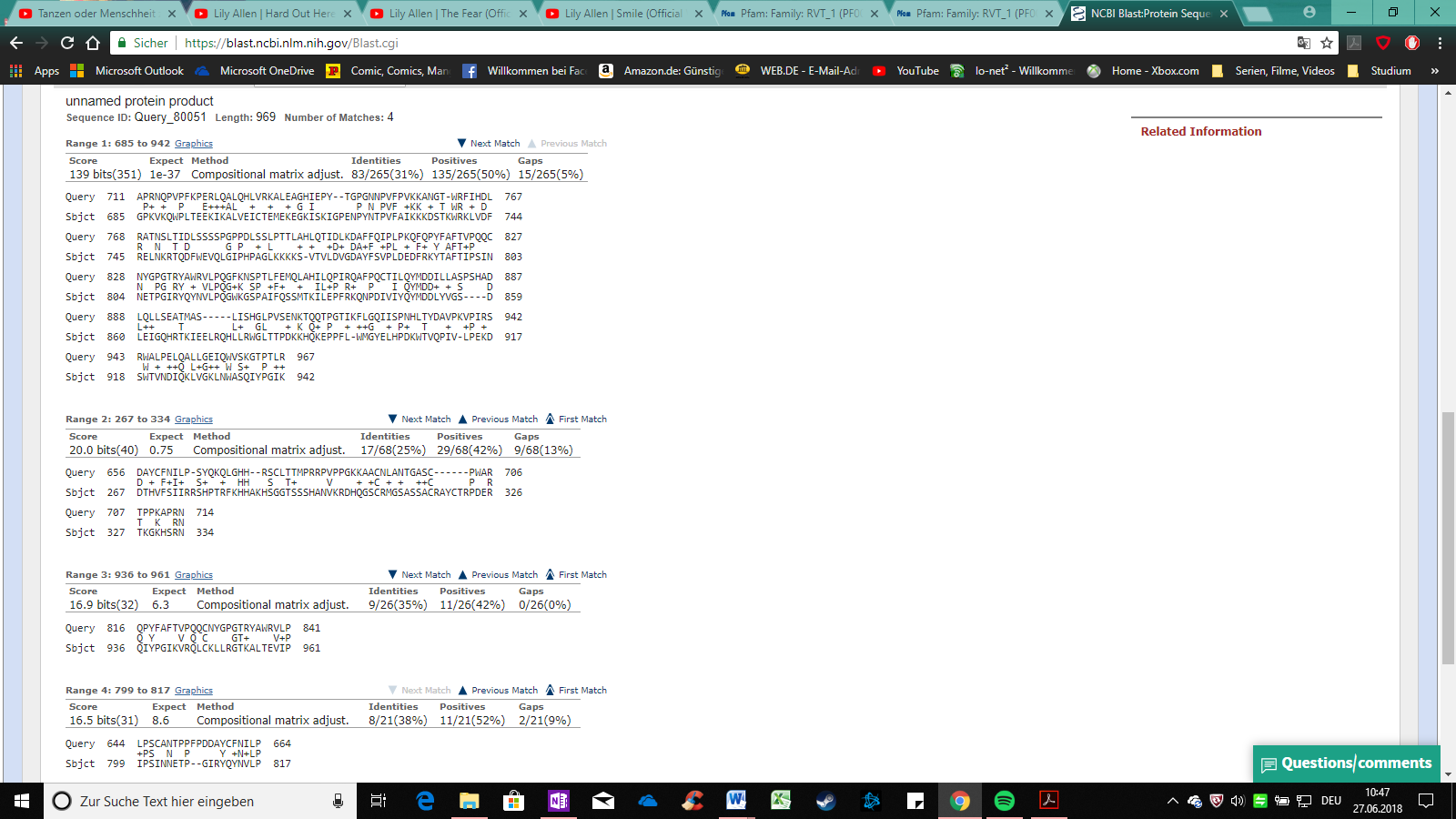
<http://pfam.xfam.org/family/PF00078.26#tabview=tab4>



*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.

|  |  |
| --- | --- |
| 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 |