

Bericht zur Analyse des Genom Human T-Cell leukemia virus type 1 mittels HMM-Profilen

Von Gensequenz zur Proteinsequenz:

Bei dem Human T-Cell leukemia virus Typ 1 handelt es sich um ein Retrovirus, welches eine Art Leukämie in den T-Zellen des Immunsystems auslöst. Mittels der HMM-Profile ist es möglich, Domänen innerhalb der Gen- bzw. Proteinsequenz des Virus zu finden.

Über die Genom-Datenbank von NCBI erhält man die Basensequenz. Exemplarisch sind hier die ersten 100 Basen dargestellt:

```
GGCTCGCATCTCTCCTTCACGCGCCCGCCGCTTACCTGAGGCCGCCATCCACGCCGGTTGAGTCGCGTT  
CTGCCGCCTCCCGCCTGTGGTGCCTCCTGA
```

Über die Basenpaarung A – T und G – C erhält man auch die ersten 100 Basenpaare:

```
GGCTCGCATCTCTCCTTCACGCGCCCGCCGCTTACCTGAGGCCGCCATCCACGCCGGTTGAGTCGCGTT  
CCGAGCGTAGAGAGGAAGTGC GCGGGCGGCGGAATGGACTCCGGCGGTAGGTGCGGCCAACTCAGCGCAA
```

```
CTGCCGCCTCCCGCCTGTGGTGCCTCCTGA  
GACGGCGGAGGGCGGACACCACGGAGGACT
```

Mittels eines Übersetzungstools (ExpASY) wird die Basensequenz in eine Proteinsequenz übersetzt. Dabei wird sowohl die 3'5'-Richtung als auch die 5'3'-Richtung verwendet, da man nicht weiß in welche Richtung die Translation startet.

Exemplarisch sind hier die ersten 30 Aminosäuren im Einbuchstabencode des 1. 5'3' Frames dargestellt:

```
GSHLSFTRPPPYLRPPSTPVESRSAASRLW
```

Beantwortung der Fragen der Aufgabe 3 a und b:

3a) Wieso ist die Suche in der Proteinsequenz der in der Genomsequenz vorzuziehen?

Die Gensequenz besteht aus 4 Buchstaben/Basen (ACGT), jeweils 3 Basen werden zu einer Aminosäure übersetzt(translatiert). Dabei ist der Übersetzungscode hoch konserviert, sodass eine Abfolge von Basen immer zu der gleichen Aminosäure führt. Der erste Grund ist daher, dass die Proteinsequenz immer kürzer als die Genomsequenz ist. Vielbedeutender in der Beantwortung ist jedoch der Übersetzungsprozess selbst. Es gibt 20 natürliche Aminosäuren, welche durch die Gene codiert werden müssen. Bei 4 verschiedenen Basen braucht man hierzu mindestens eine Folge von 3 Basen pro Aminosäure. Würde man nur eine Base verwenden, hätte man nur 4 Möglichkeiten und somit auch nur 4 Aminosäuren, die codiert werden können. Bei einer Folge von 2 Basen pro Aminosäure gäbe es insgesamt 16 Möglichkeiten. Auch dies ist zu wenig. Verwendet man jedoch jeweils eine Sequenz der Länge 3, erhält man $4^3=64$ Möglichkeiten. Diese werden ausgenutzt, indem eine Aminosäure durch mehrere Kombinationen der 4 Basen codiert wird. So können Punktmutationen in der Gensequenz auf die Proteinsequenz keine Auswirkung haben. Verwendet man jedoch die Gensequenz zur Analyse, würden bestimmte Motive unter Umständen nicht erkannt werden, obwohl die Proteinsequenz noch immer übereinstimmt. Die Analyse nach bekannten Motiven eignet sich daher mehr in der Proteinsequenz. Denn diese muss zum Ausbilden ein und derselben Funktion stark mit dem bekannten Motiv übereinstimmen. Durch die Verwendung der Gensequenz müsste man mit viel mehr Fehlern rechnen.

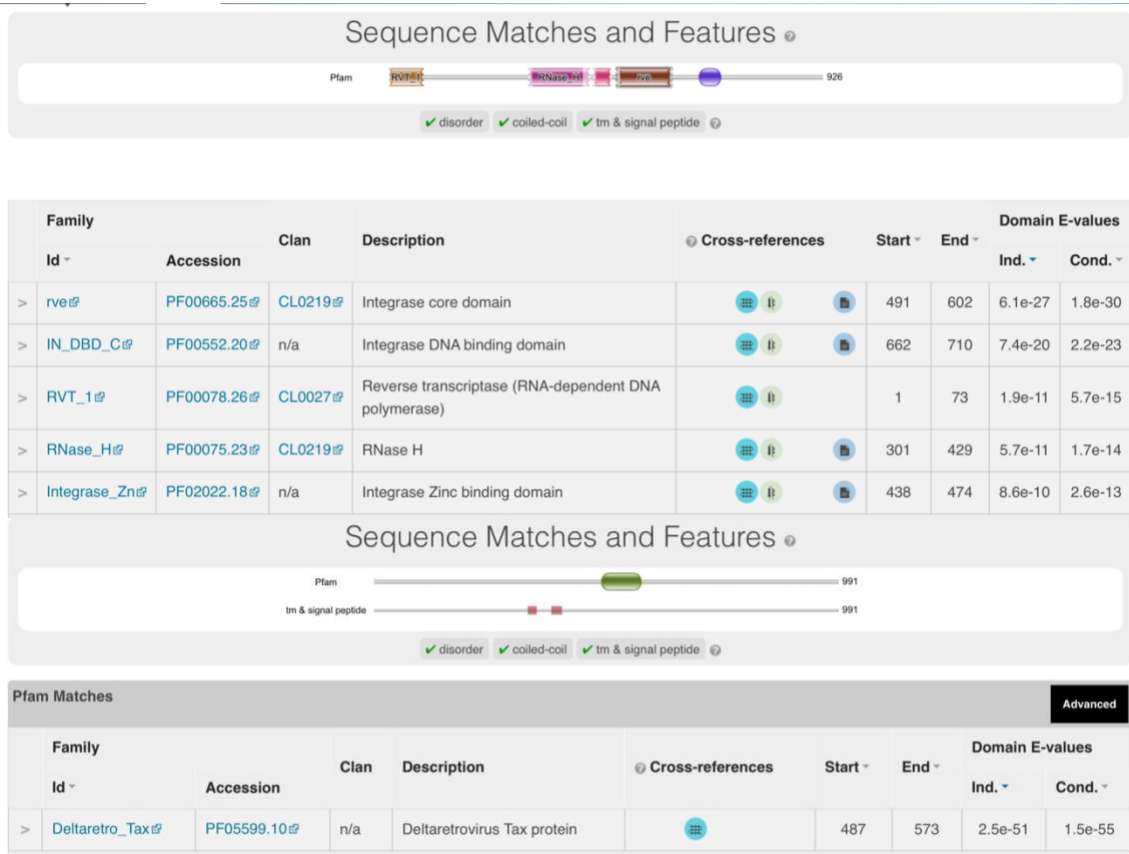
3b) Wieso ist es sinnvoll alle 6 möglichen Übersetzungsframes zu untersuchen?

Die Translation startet immer bei einer bestimmten Startsequenz. Bei dieser Suche wird jedoch einfach im ersten Frame einfach die erste Base als Startpunkt verwendet. Je nachdem, ob man den Startpunkt bei der ersten, zweiten oder dritten Base festlegt, ändert sich der „reading frame“ und somit auch die Proteinsequenz, die man erhält. Da nicht bekannt ist, welche dieser drei Möglichkeiten der im Organismus entspricht, ist es sinnvoll alle mit den Profilen zu vergleichen, um so möglichst genaue Ergebnisse zu erzielen. Zusätzlich ist es möglich die Übersetzung am 3' bzw. 5' Ende zu starten. Auch dies führt zu unterschiedlichen Proteinsequenzen. Somit ergeben sich insgesamt 6 Möglichkeiten die Gensequenz in eine Proteinsequenz zu übersetzen. Jeweils drei Frames in 5'3'- und in 3'5'-Richtung.

Durchsuchung des 1. Frame 5'3' mit den HMM-Profilen:

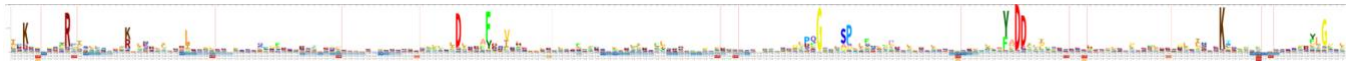
Der erste Frame in 5'3'-Richtung wird nun in die Datenbank mit den HMM-Profilen eingespeist und untersucht.

Gefunden wurden im ersten 5'3' Frame die Profile: reverse Transkriptase RVT_1, Rnase H, eine Integrase Zink-binde-Domäne, eine Integrase core domain, eine Integrase DNA Bindedomäne und das Deltaretrovirus Tax Protein.



Das HMM Logo der Reversen Transkriptase RVT_1 passt in weiten Teilen zu der Proteinsequenz. Vor allem hoch konservierte Aminosäuren (in HMM Logo größer als 1) stimmen mit der Sequenz überein. In der folgenden Abbildung wird dies durch die großen Buchstaben zwischen Model und Query angezeigt. Da das HMM Profil sehr lang ist und daher in diesem Bericht nur sehr klein dargestellt ist, zeige ich die Ähnlichkeit an dem Alignment von Model und eingespeister Proteinsequenz. Auffällig ist, dass es viele größere

Id		Accession	Clan	Description	Cross-references	Start	End	Ind.	Cond.
	RVT_1	PF00078.26		<i>polymerase</i>				6.7e-38	4.0e-42
Model	1	<p>.....*.....*.....*.....*.....*.....*.....*.....*.....*</p> <p>1 ipkkgpsyRpsllsvdykalkliakrLkdvlekllengpggfrgrstldaveellkalkkkkkakllkldlkkaF 80</p> <p>++K + g+tR i d+a n+l+++ ls ++pg ++ s+ ++ +l ++Dlk+aF</p>							
Query	17	<p>VKAN--GTWRFI--HDLRATNSLTID-----LSSSSGPPDLSSL-----PTTLAHLQTLDKDAF 69</p>							
PP		<p>57888.*****8...8*****777.....56789*****</p>							
Model	81	<p>.....*.....*.....*.....*.....*.....*.....*.....*.....*</p> <p>81 dsvpleellrlkltafkvptkllinliksflstrsfsvrvnge.esegryekkglpqGsvlSPlllnlrmellirelkrak 159</p> <p>+++pl+++++++af+vp++ + ++ry ++lpgG+++SP+l f + + +l+++++++</p>							
Query	70	<p>FOIPLPKQFOQPYFAFTVQQ-----CNYGPGTRYAWRVLPQGFKNSTPLFEMOLAHILQPIROAF 130</p>							
PP		<p>*****77.....33489*****</p>							
Model	160	<p>.....*.....*.....*.....*.....*.....*.....*.....*.....*</p> <p>160 gvtlirYaDDililskkeelqellaveewlkesgiknpeKtklvlfsgkseevkylGvt 221</p> <p>++t+l++DDil++s s+ +lq l ea+ l ++gl +++Kt+ + + ++kLG +</p>							
Query	131	<p>QCTILQXMDILLASPSHADLQLLSEATMASLISHGLPVSENKTOQT-----PGTIKFLQOI 187</p>							
PP		<p>*****.....9*****86</p>							



	Id ▾	Accession	Clan	Description	🔗 Cross-references	Start ▾	End ▾	Ind. ▾	Cond. ▾
	RVT_1	PF00078.26						1.9e-11	5.7e-15
V	RNase_H	PF00075.23	CL0219	RNase H		301	429	5.7e-11	1.7e-14

.....*.....*.....*.....*.....*

Model 5 vvtDGscignseggagavlykagar...nisaple.caqtnnraELsAvicaalkaksdekviyvtDSkYvvkgitgw 80

+++DGs +++a+++++ + + ++s+pl+ +++++raEL + + L++++s+ ++ni+ DSkY+++ +

Query 305 CLFSGDS-----TSQAAYILWDK--HilsQRSFPLPpPHKSAQRAELLGLLHGLSSARSWRCLNIFLDSKLYLH----Y 372

PP 5678888.....669***97766..4445999***99*****88*****...4

.....*.....*.....*.....*.....*




Model 81 vhwkknawkttsqkpvknkelaellkelakkkkvqlkhvkgHaGd 128

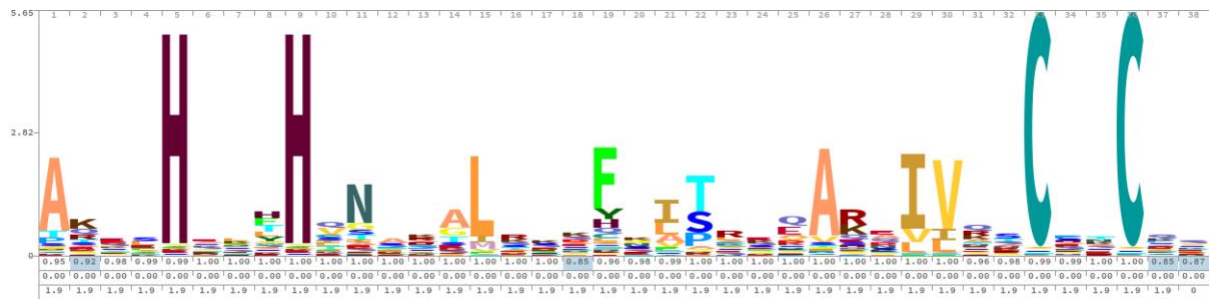
+ + + + s+q p+ + l +l ++k v+l+hv+ H++ p

Query 373 LRTLALGTFGGRSSQAPFOA--LLPRL---LSRKVVYLHHVRSHTNLP 415

PP 445 5 666767777877777..44444...68899*****9984

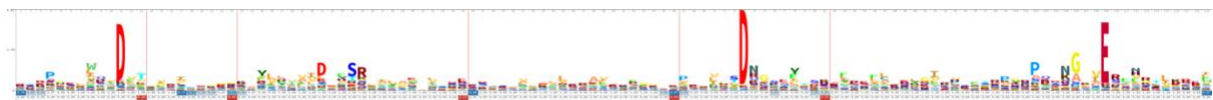


V	Integrase_Zn	PF02022.18	n/a	Integrase Zinc binding domain	 		438	474	8.6e-10	2.6e-13
Model	3	<pre>*.....*.....*..... esHelhHqNakALrkKfKitreqAreIVqsCptCg ++H+++H++++aL++ ++t+++A +I++sC++C+ </pre>		37						
Query	440	<pre> DLHSFTHCGOTALTI-QGATTTEASNILRSCHACR </pre>		473						
PP	89	<pre> *****.******7 </pre>								




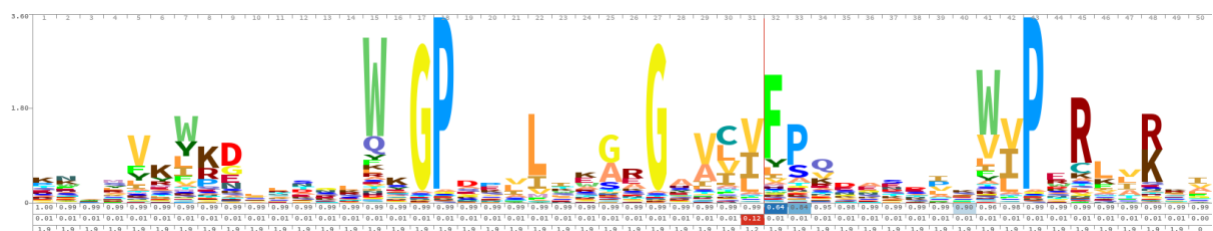
Der Vergleich der Integrase core domain zeigt, dass die konservierten Aminosäuren sich auch in dieser Sequenz wiederfinden. Dies sieht man z.B. an der Aminosäure Asparaginsäure (D) an den Positionen 12 und 74. Es gibt jedoch auch größere Lücken. Das HMM Profil zeigt jedoch deutlich, dass es 3 Stellen gibt, an denen eine ganz bestimmte Aminosäure sein muss.

Family	Id	Accession	Clan	Description	Cross-references	Start	End	Domain E-values	
								Ind.	Cond.
V	rve	PF00665.25	CL0219	Integrase core domain		491	602	6.1e-27	1.8e-30
Model	4	<p>.....*.....*.....*.....*.....*.....*.....*</p> <p>ggelwqvDvfvripdgggkayllviddferlllealsdmdastvllaleravrfggvepervlsDngseytskaf 83</p> <p>p+++wq+D+t+++++ + l+v++d+fs+ i a + t+e ++s ++ l +a+a+ g+ p +++Dng++y+s++f</p>							
Query	492	<p>PNHIWQGDITHFKYKNTL--YRLHVWVDTFSGAISATQKRKE--TSSEAISLLQAIAYLGR--PSYINTDNGPAYISQDF 566</p>							
PP		<p>99*****77666..*****.....,.....9*****</p>							
Model	84	<p>.....*.....*.....*.....*</p> <p>reflaelglvsvftrpgrpqdnGkvErfrtlkd 117</p> <p>+++ +l ir++++ p++p++ G+vEr n++lk+</p>							
Query	567	<p>LNMCTSLAIRHTTHVPYNPTSSGLVERSNGILKT 600</p>							
PP		<p>*****96</p>							



Auch beim Vergleich der DNA bindenden Domäne zeigt sich, dass hoch konservierte Aminosäuren übereinstimmen. Insbesondere die Abfolge W_GP an der Stelle 16 und W_P_R an der Stelle 41 finden sich in der Proteinsequenz.

Id ▾		Accession	Clan	Description	⦿ Cross-references	Start ▾	End ▾	Ind. ▾	Cond. ▾
	rve	PF00665.25						6.1e-27	1.8e-30
V	IN_DBD_C	PF00552.20	n/a	Integrase DNA binding domain		662	710	7.4e-20	2.2e-23
Model	1	<div>.....*.....*.....*.....*</div> <div>knamvkwkdlllgllwkgpdpvlikgrGavcvFpqdasdikvPeRlvrki 50</div> <div>+++++k+++l+s++WkGP+++L++++Ga+++ P++as+++w+P+Rl++++</div>							
Query	662	<div>HWYFPLPGLNSRQWKGQEAALQEAAGAALI--PVSASSAQWIPWRLIKRA 710</div>							
PP		<div>7*****85</div>							

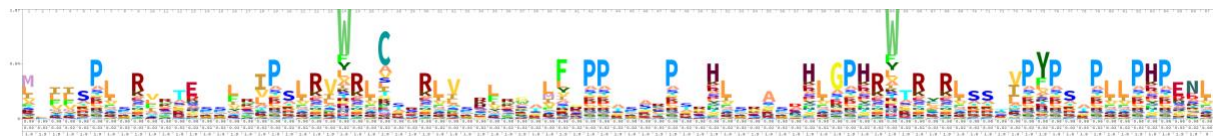


Bei dem Deltaretrovirus Tax Protein erkennt man eine sehr hohe Übereinstimmung. Nur 14 Aminosäuren unterscheiden sich zwischen Modell und Proteinsequenz. Diese hohe Übereinstimmung deutet auf eine sehr spezialisierte Funktion des Proteins hin, welche genau diese Proteinsequenz für die Struktur bzw. die Funktion benötigt. Die genaue Funktion der Tax Proteine ist noch nicht geklärt. Man weiß jedoch, dass sie für die Virusexpression von Bedeutung sind. Dies würde die hohe Ähnlichkeit aller Tax Proteine erklären. Das HMM Profil zeigt jedoch auch, dass an den meisten Positionen mehrere Aminosäuren gleich wahrscheinlich anzutreffen sind. Oftmals handelt es sich dabei um Aminosäuren, die ähnliche Funktionen/chemische Umgebungen bieten.

Family				Cross-references	Start	End	Domain E-values	
Id	Accession	Clan	Description				Ind.	Cond.
v	Deltaretro_Tax	PF05599.10	n/a	Deltaretrovirus Tax protein	487	573	2.5e-51	1.5e-55

Model	1*.....*.....*.....*.....*	80
Query	487	mliisplrvrtessripslrwrlcrrrlv+lg+fgpp+s+rp++hlrsasdhlgphwtryrlsstvpypsapl	566
PP	99	*****	

Model	81	lphpenl	87
Query	567	LPHPENL	573
PP		*****8	



Die gefundenen Profile zeigen, dass oftmals nur einige wenige hochkonservierte Aminosäuren übereinstimmen müssen. Auch passen alle gefundenen Profile zu einem Retrovirus.

Wiederholung mit dem Human Immunodeficiency Virus 1

Bei dem HI-Virus handelt es sich um eines der bekanntesten Geschlechtskrankheiten, welche bis heute nicht geheilt werden kann.

Im Folgenden sind die ersten 100 Basen des Genoms:

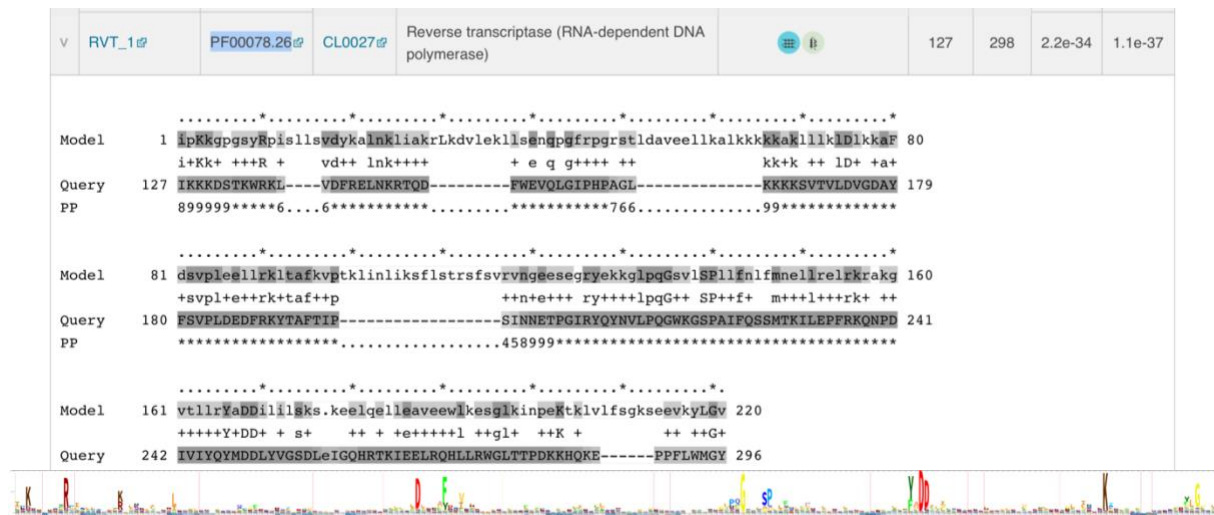
```
GGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCCACTGCTTAAGCC
TCAATAAAGCTTGCCTTGAGTGCTTCAAGT
```

Mittels Übersetzung in eine Proteinsequenz erhält man:

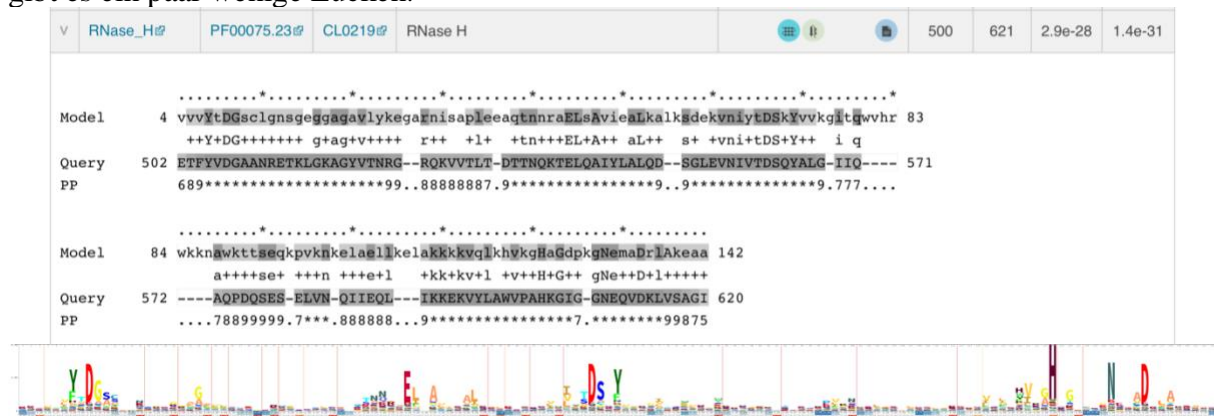
GLSG-TRSEPGSSLAN-GTHCLSLNKACLEC

Die Untersuchung des 2. Frames 5`3` findet folgende Motive:
reverse Transkriptase, RNaseH und retrovirale aspartyl Protease.

Der Vergleich des Modells der Reversen Transkriptase zeigt, dass die hochkonservierten Aminosäuren auch in der Proteinsequenz vorhanden sind. Allerdings gibt es auch einige Lücken in der Sequenz.



Auch bei der RNase H stimmen hoch konservierte Aminosäuren überein. Zwischen diesen gibt es ein paar wenige Lücken.



Der Vergleich des Modells der Retroviralen Aspartyl Protease und der Proteinsequenz zeigt, obwohl einige Aminosäuren übereinstimmen, viele in der Proteinsequenz nicht wieder zu finden sind. Das liegt möglicherweise daran, dass nur ein Teil der Sequenz verwendet wird.

