Alignments von teilweise codierenden Regionen:
Das code2aln - Projekt und seine Anwendungen

Roman STOCSITS

Institut für Theoretische Chemie und Molekulare Strukturbiologie
Universität Wien

Defensio

TBI, Februar 2003

- Alignments im Prinzip
- Einführung in das Problem
- Das `code2aln` - Projekt
- Hepatitis B Virus und Levivirus
- Resultate und Diskussion

Ein Alignment ist einer der grundlegendsten Sequenzanalyse-Tasks.

Wir nutzen es, um festzustellen, ob zwei oder mehrere Sequenzen verwandt sind und wie nah die Verwandtschaft bezüglich der Sequenzähnlichkeit ist.

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
            ++ ++++H+ KV   + +A  ++          +L+ L+++H+ K
LGB2_LUPLU  NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG
```

Beispiel für ein Alignment zweier kurzer Proteinteile: identische und ähnliche Positionen sind gekennzeichnet.

```
ADI-MAL       CUCUAUACAACAGGGAUAGUAGGA---------------GAUAUAAGAAGAGCAUAUUGUACU
AE-90CF402    UUCCAUACAACAGGAAACAUAAAUGGU------------GAUAUAAGAAAAGCAUAUUGUGAA
AE-CM240      UUCUAUAGAACAGGAGAUAUAAUAGGA------------AAUAUAAGAAAAGCAUAUUGUGAG
B-896         UUUUAUGCAAGAAGAAACAUAAUAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-ACH320A     UUUUAUGCAACAGGACAAAUAAUAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-BCSG3       UAUUAUACAACAGGAGAAAUAGUAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-CAM1        GUUUAUGCAACAGACAGAAUAAUAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-D31         UUUUAUACAAAAGGAAAAAUAAUAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-HIV1AD8     UUUUAUACAACAGGAGACAUAAUAGGA------------GAUAUAAGACAAGCACAUUGCAAC
B-HXB2        UUUGUUACAAUAGGAAAA---AUAGGA------------AAUAUGAGACAAGCACAUUGUAAC
B-JRCSF       UUUUAUACAACAGGAGAAAUAAUAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-LAI         UUUGUUACAAUAGGAAAA---AUAGGA------------AAUAUGAGACAAGCACAUUGUAAC
B-MANC        UUUCAUGUAACAAGAGCCGUAACAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-OYI         UUUCAUACAACAAAACAAAUAAUAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-SF2         UUUCAUACAACAGGAAGAAUAAUAGGA------------GAUAUAAGAAAAGCACAUUGUAAC
B-WEAU        CUUUAUACAACAGGAGAAAUAAUAGGA------------GAUAUAAGACGAGCACAUUGUAAC
B-YU2         UUGUAUACAACAGGAGAAAUAAUAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-pNL43       UUUGUUACAAUAGGAAAA---AUAGGA------------AAUAUGAGACAAGCACAUUGUAAC
D-ELI         CUCUAUACUACAAGAUCAAGAUCA---------------AUAAUAGGACAAGCACAUUGUAAU
D-NDK         CUCUAUACAAUAACAGGAAAAAAGAAGAAAACAGGA---UACAUAGGACAAGCACAUUGUAAA
O-ANT70       UACAGCAUGGGAAUAGGGGGAACAGCAGGAAACAGC------UCAAGGGCAGCUUAUUGCAAG
O-MVP5180     CGCAGUAUGACACUUAAAAGAAGUAACAAUACAUCACCAAGAUCAAGGGUAGCUUAUUGUACA
SIVCPZGAB     UUUUAUAAUAUAGAAAAUGUAGUAGGA------------GAUACCAGAUCUGCCUACUGUAAG
```

Beispiel für ein multiples Nukleinsäure-Alignment: ein kurzer Ausschnitt des *env*-Gens von HIV-1.

Es gibt für 2 Sequenzen der Länge n:

1 mögliches Alignment ohne Gaps,

aber:

$$\frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{2\pi n}} \tag{1}$$

mögliche Alignments mit Gaps verschiedener Länge.

Deshalb: `dynamic programming`

Das optimale Alignment wird rekursiv berechnet.

Optimale Teilalignments kleinerer Subsequenzen führen zur Gesamtlösung.

Gap-Penalties

Lineare Gap-Penalties:

$$\gamma(g) = -gd \tag{2}$$

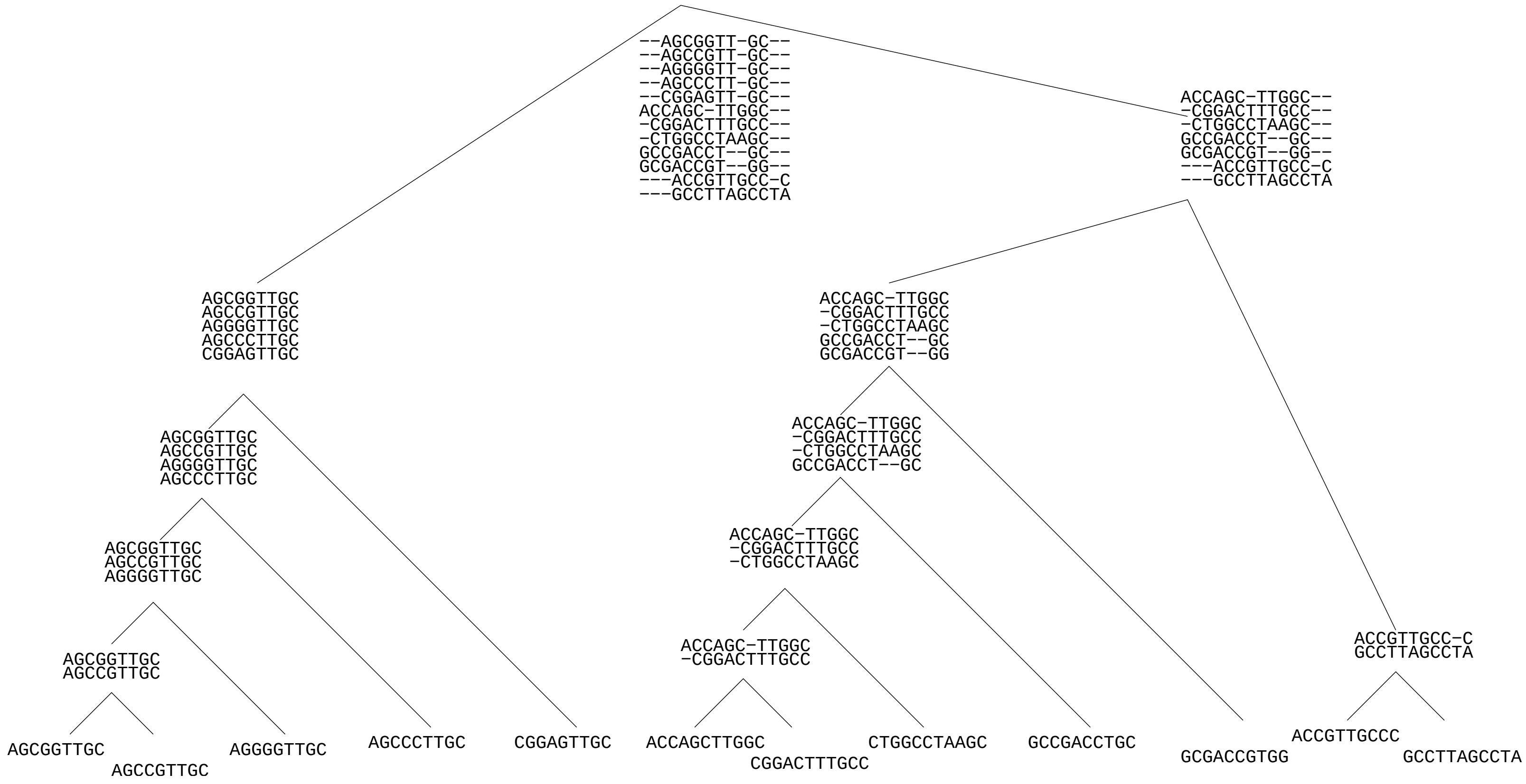
Affine Gap-Penalties:

$$\gamma(g) = -d - (g-1)e \tag{3}$$

d Gap Open-Penalty; e Gap Extension-Penalty; g Gap-Länge

Multiples Alignment

Ein Guide-Tree repräsentiert den Prozess des multiplen (Profil-) Alignments.



Die höhere Sequenzheterogenität von Nukleinsäuren im Vergleich zu Protein:

```
SARGLSSTVSLGQFEHWSPR
+AR+LS+TVSL+QF+H SPR
NARNLSDTVSLSQFDHPSPR
```

```
AGTGCAAGAGGATTAAGTAGTACAGTAAGTTTAGGACAATTTGAACATTGGAGTCCAAGA
   GC  G G   T       AC G      T    CA TT GA CA       CC  G
GACGCCCGCGACCTCTCCGACACCGCTTCCCTCTCCCAGTTCGACCACCCCTCCCCCCGC
```

Oben ein Protein-Alignment mit hohem Grad an Identität.

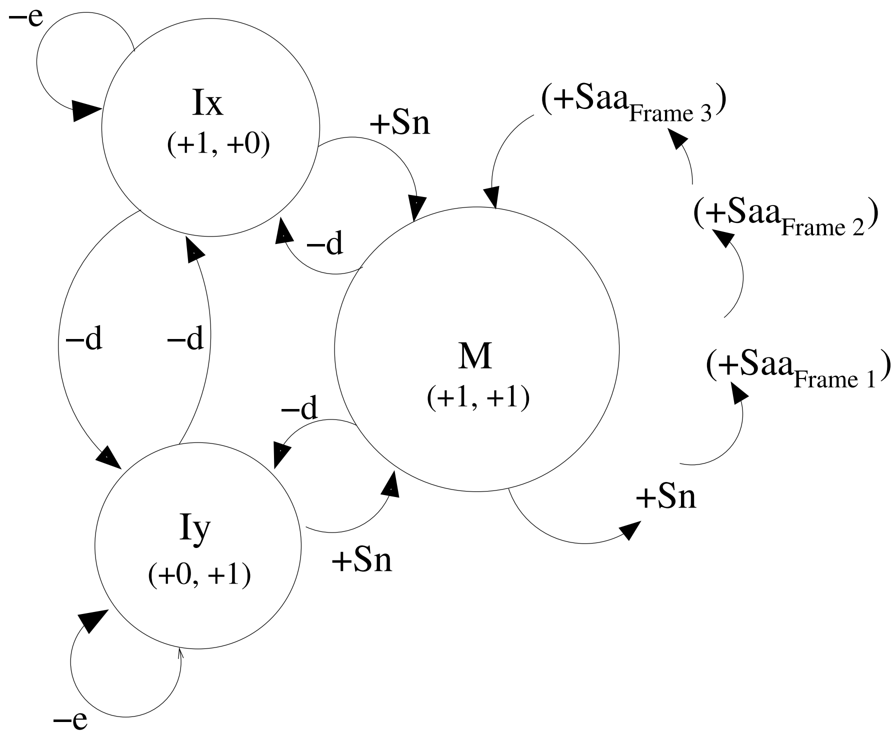
Darunter diesselbe Sequenz auf Ebene der Nukleinsäuren: viel heterogener.

Die paarweise Identität ist nur 33%.

Code2aln

- produziert multiple Nukleinsäure-Alignments
- verwendet affine Gap-Penalties
- findet Information über codierende und nicht-codierende Regionen
- nutzt diese als Teil der Scoring-Funktion
- und berücksichtigt überlappende codierende Regionen

Die Zustände und Scoring-Terme im `code2aln`-Algorithmus:



$\underline{M}$ = (Mis)match state

$\underline{Ix,\ Iy}$ = InDel states

$\underline{Sn}$ = Inkrement des Nukleinsäure-Scores

$\underline{d,\ e}$ = Gap-open- und -extension-Penalties

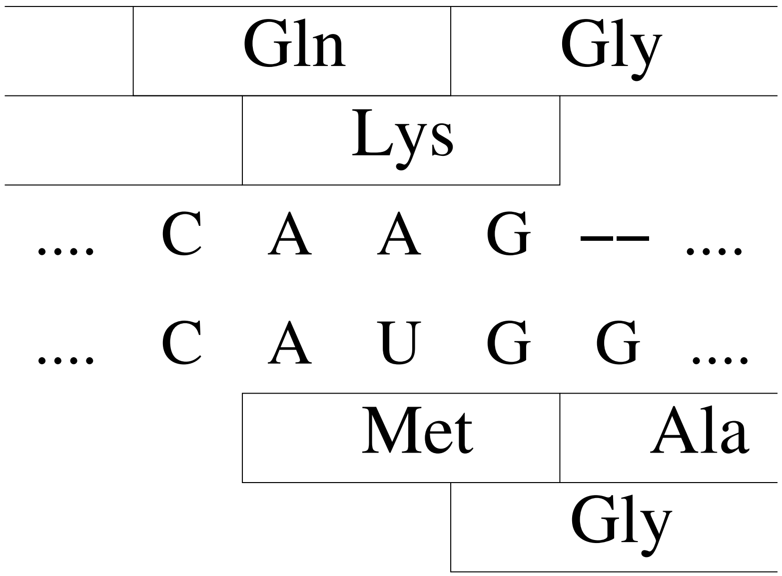
$\underline{Saa}$ = Aminosäure-Scores in drei möglichen Frames

Die Scoring - Funktion des `code2aln` - Algorithmus.

$$S = S_N + \sum_{i=1}^{3} S_{A_i} \tag{4}$$

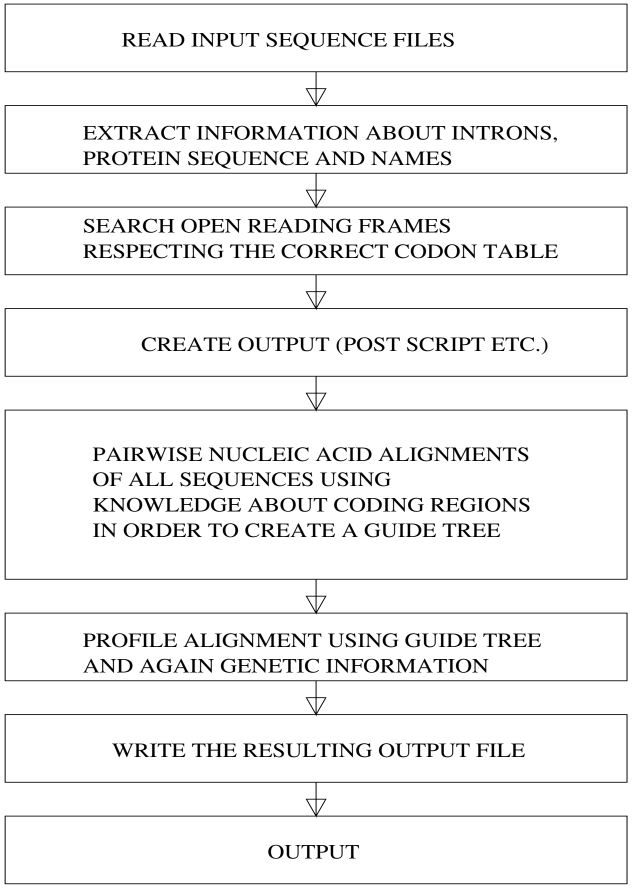
S Gesamt-Score des betrachteten Nukleotid-(Mis)matches;

S_N Nukleotid-Score; S_{A_i} Aminosäure-Score in den 3 Frames



Match number:	1	2	3	4	5
Nucleic acid score:	10	10	0	10	0
Amino acid score (frame 1):	0	4	0	12	0
Amino acid score (frame 2):	0	0	4	0	0
Amino acid score (frame 3):	0	0	0	4	0
Gap penalty:	0	0	0	0	–10
Total sum:	10	14	4	26	–10

Der `code2aln` - Algorithmus im Ablaufdiagramm.

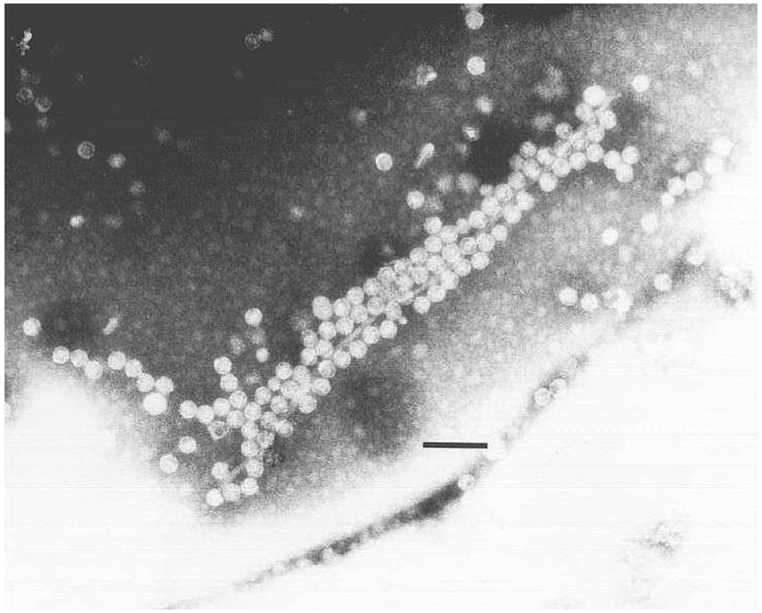


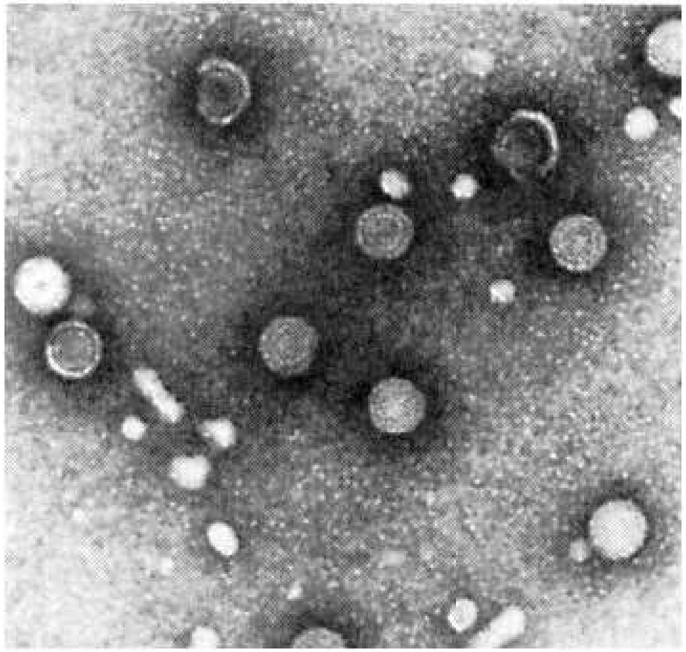
Anwendung: Leviviridae und Hepatitis B Virus

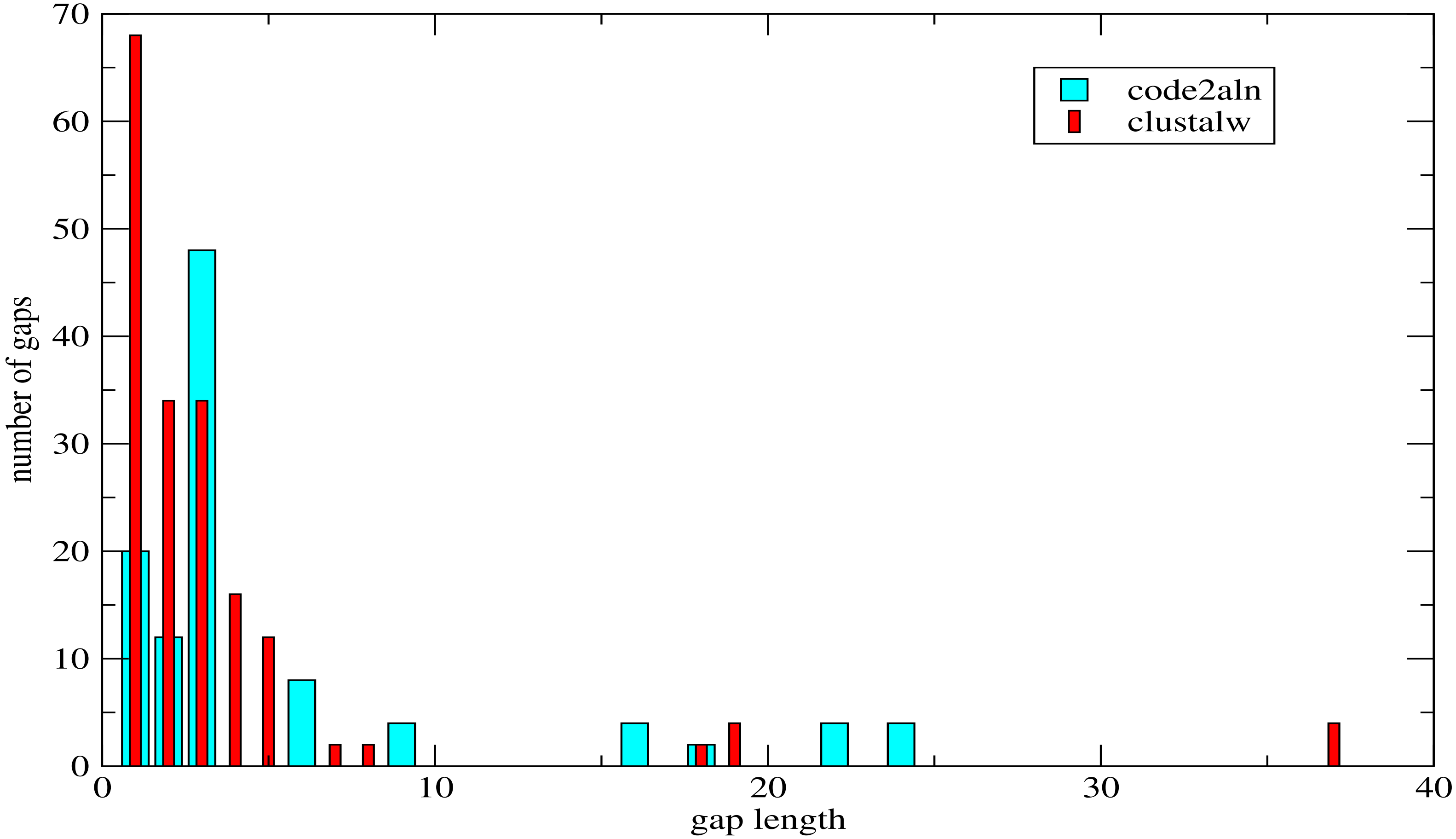
- Virus-Genome sind ein guter Testfall für `code2aln`, weil viele mehrere nicht-codierende und überlappende codierende Bereiche enthalten.

- Die Alignments wurden genutzt als Input für den `Alidot`-Prozess zur Vorhersage konservierter Sekundärstrukturelemente.

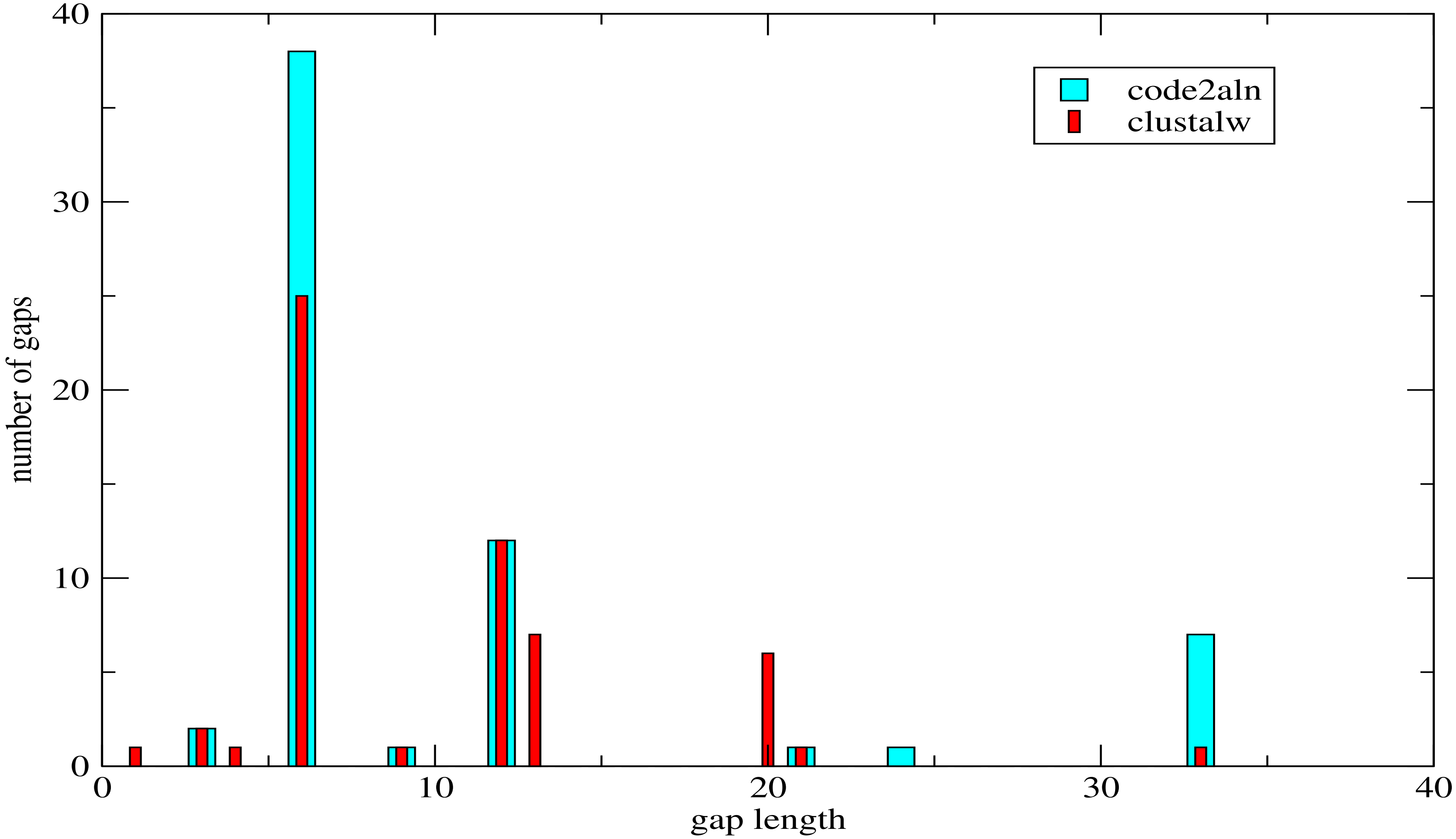
Elektronenmikroskopische Aufnahmen der Viruspartikel von Levivirus und Hepatitis B Virus.





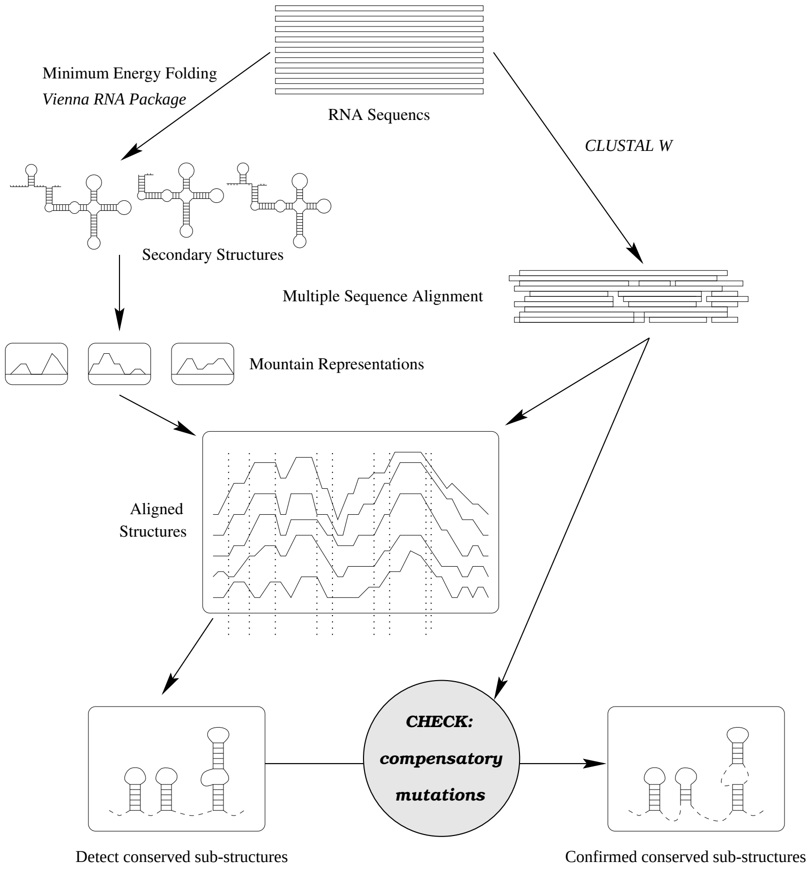


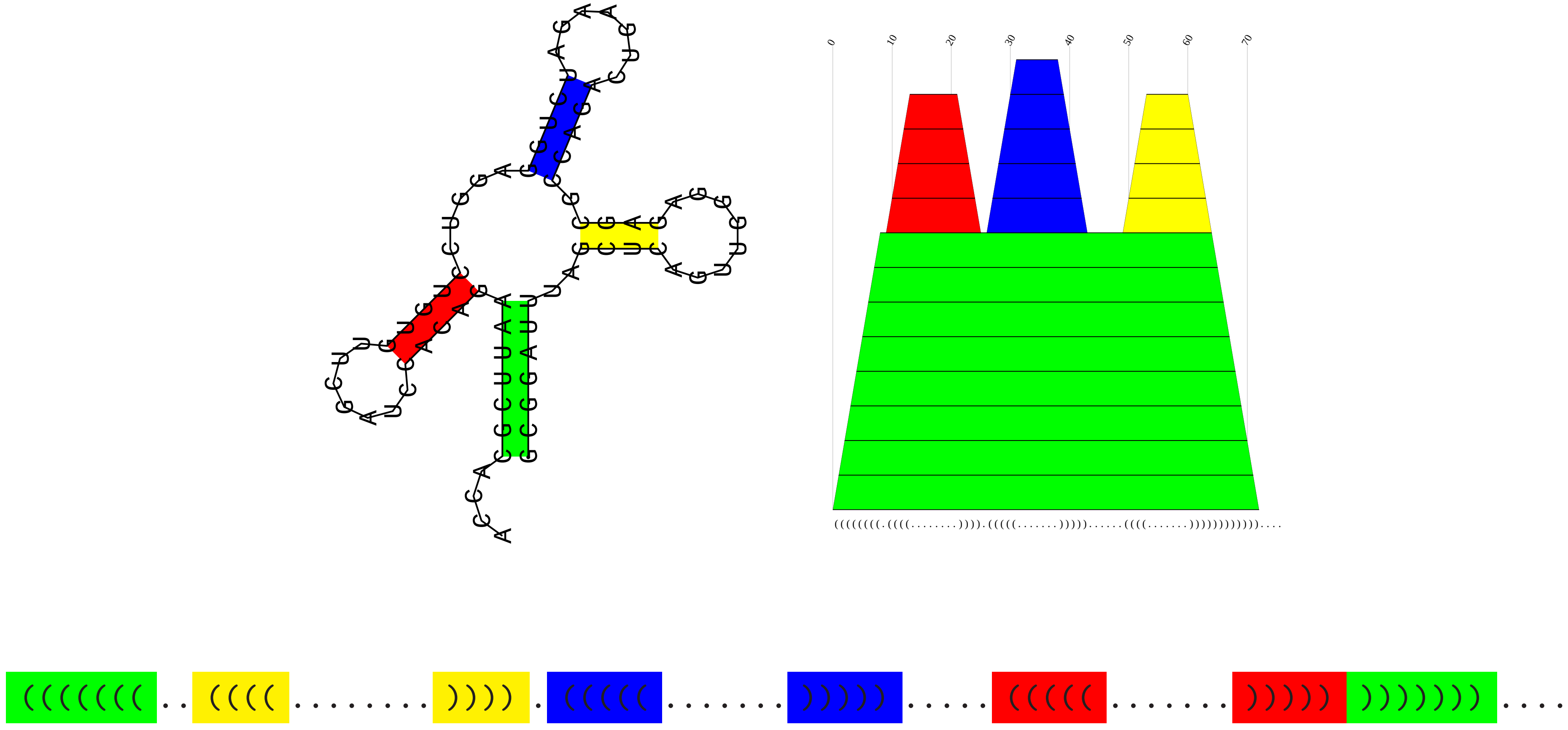
Leviviridae: `code2aln` hat die starke Tendenz, Codons in codierenden Bereichen nicht zu zerreissen. Dadurch ergibt sich ein viel größerer Anteil an Gaps, die durch 3 teilbar sind.



Hepatitis B Virus: `code2aln` tendiert abermals mehr dazu, Codons in codierenden Bereichen nicht zu zerreissen. Mehr Gaps sind durch 3 teilbar, ansonsten sehr ähnliche Verteilung der Gap-Längen.

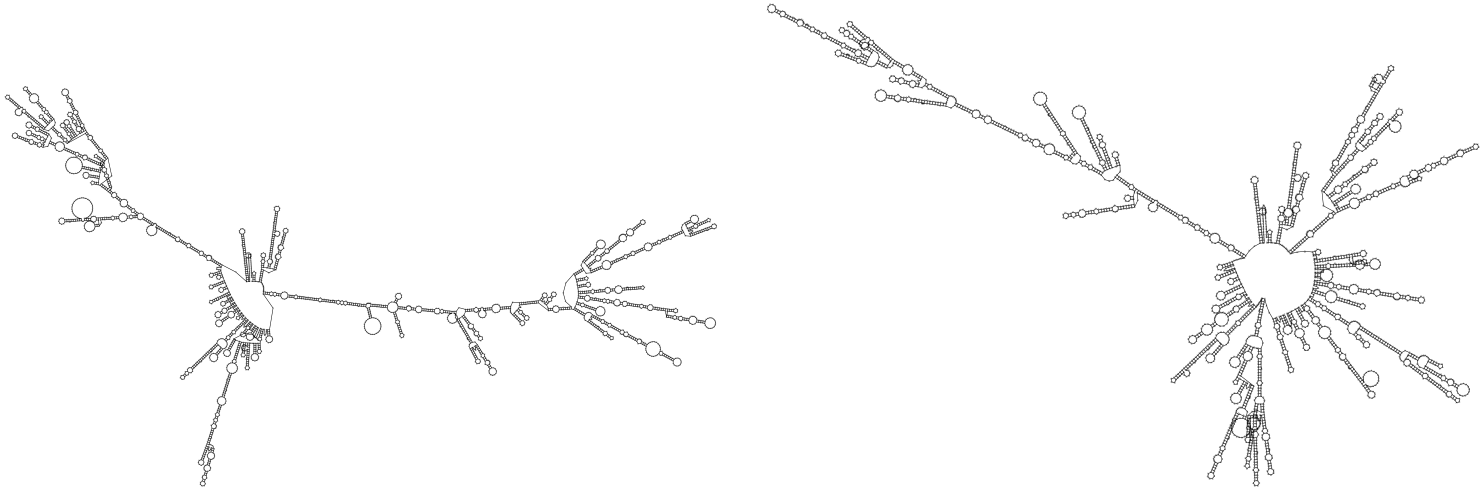
Ablaufdiagramm von `alidot`:



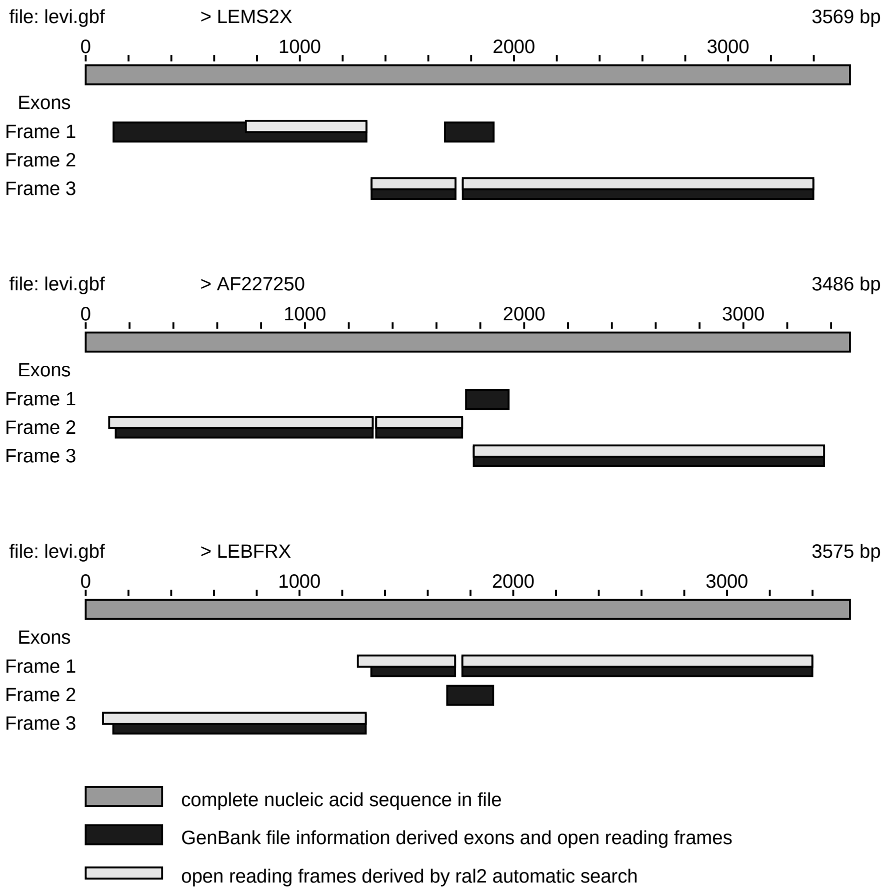


Arten der Darstellung von RNA-Sekundärstrukturen.

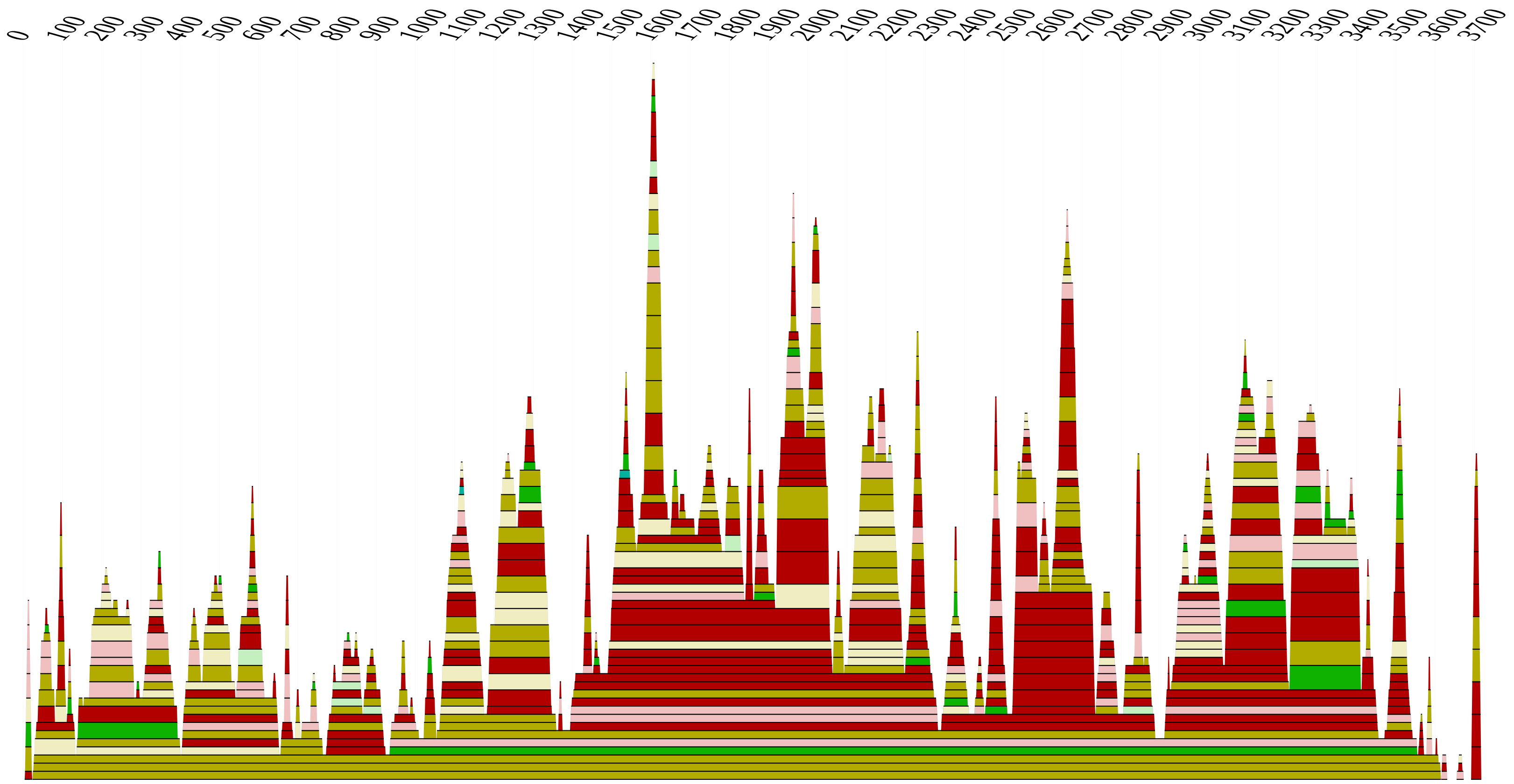
Zwei vorhergesagte Sekundärstrukturen gesamter Hepatitis B Virus-Genome:



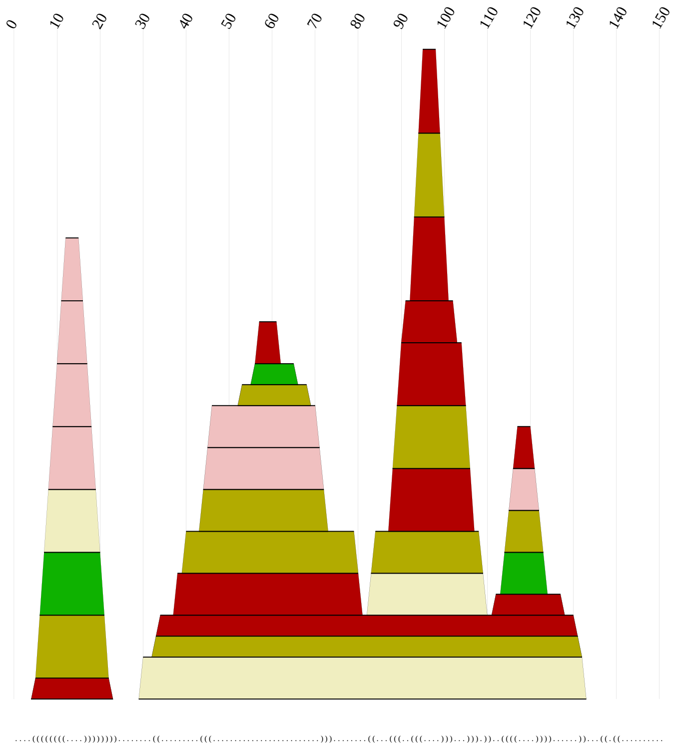
Kaum Ähnlichkeit: kleine Änderungen der Sequenz können enorme Änderungen der Strukturen bewirken.

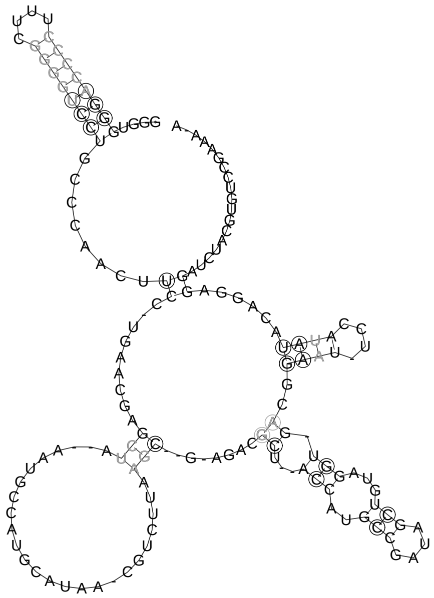


Die Genom-Organisation von 3 Levivirus-Sequenzen, wie `code2aln` sie detektiert.

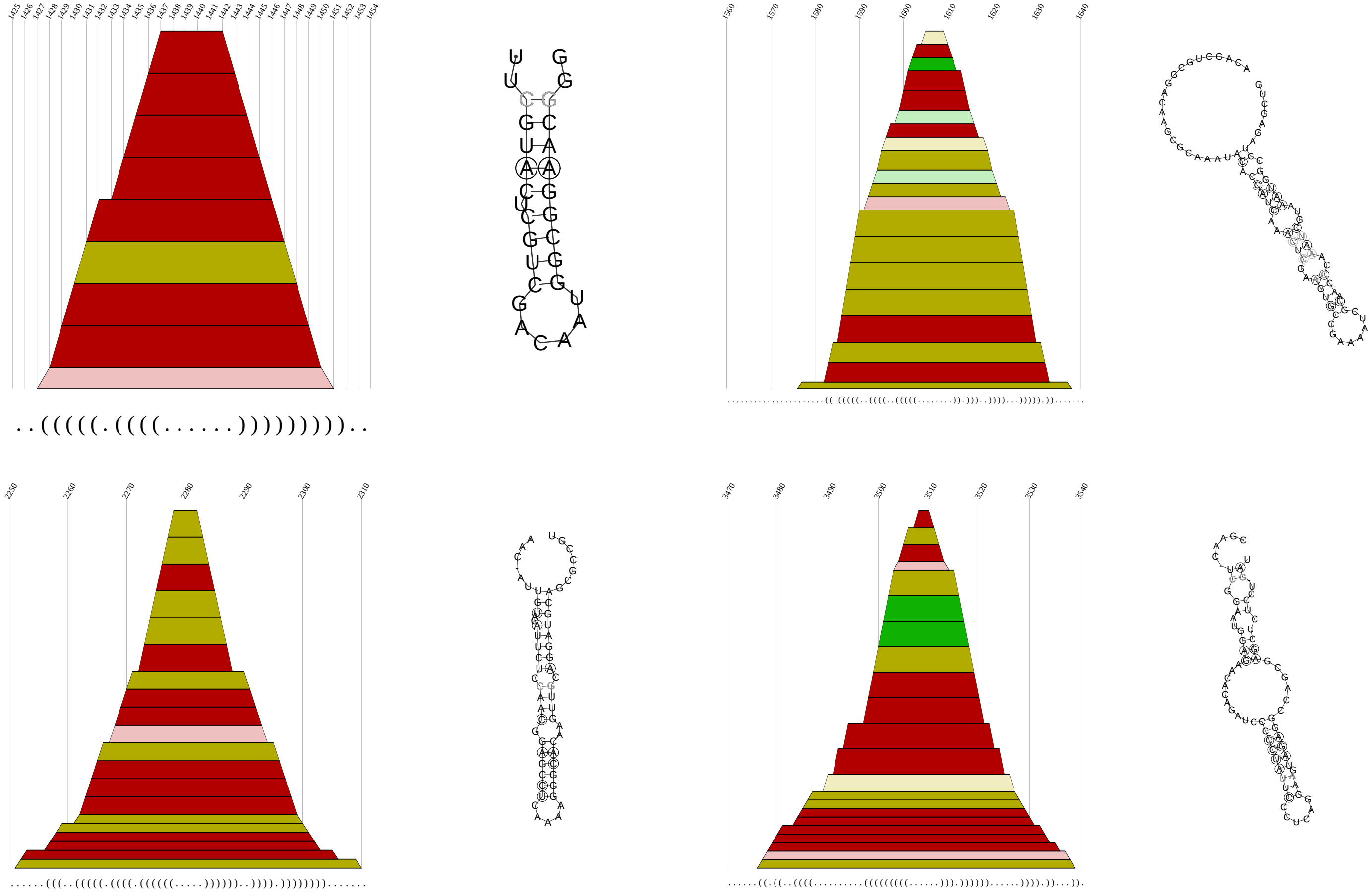


Die Strukturen des kompletten RNA-Genoms von Levivirus.

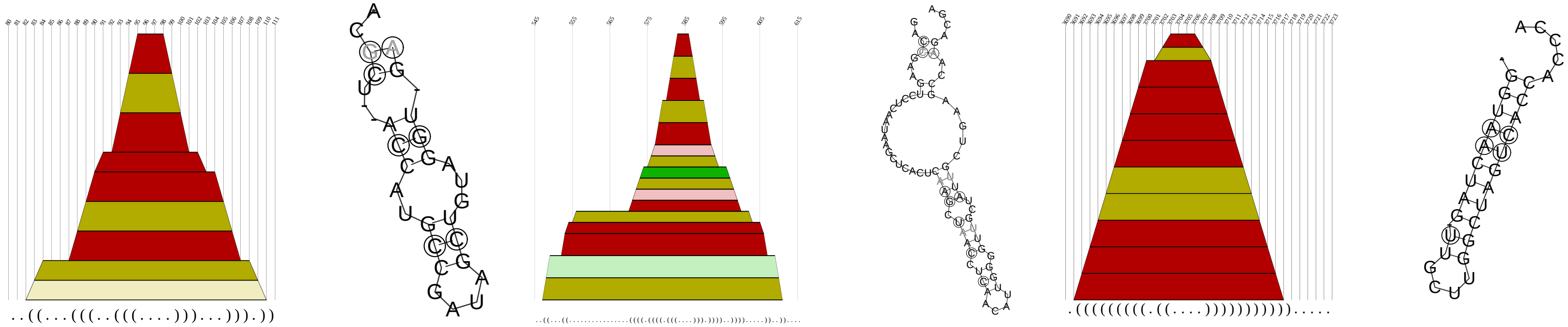




Position 1-150: Der 5'-terminale Hairpin ist wahrscheinlich das Analogon zur Recognition Signal Site für die RNA-Replikase in Allolevivirus.



Die Positionen 1426-1454, 1560-1640, 2250-2310 und 3470-3540.



Die Positionen 80-110, 545-615 und 3690-3723; drei Tetraloops.

S: Surface-Antigene
C: Core protein
P: Reverse Transcriptase, mit DNA-Polymerase- und RNase H - Funktion
X: Transaktivationsfunktion

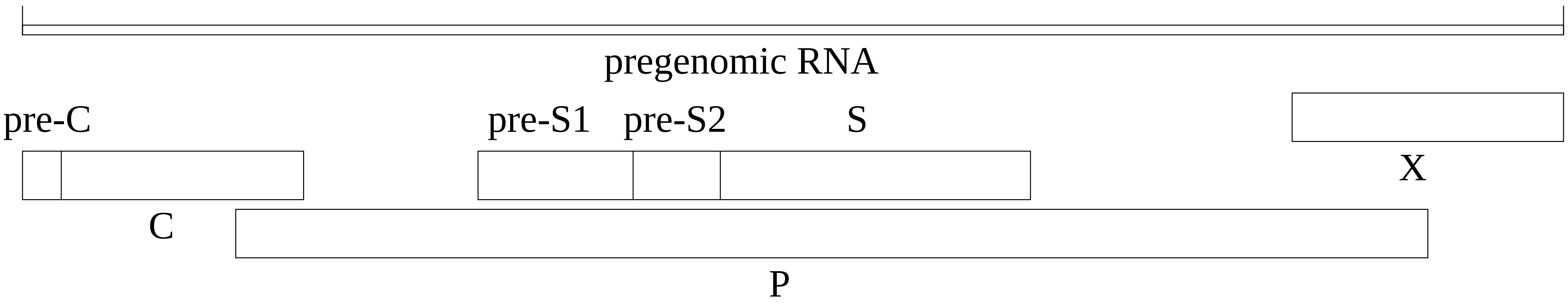
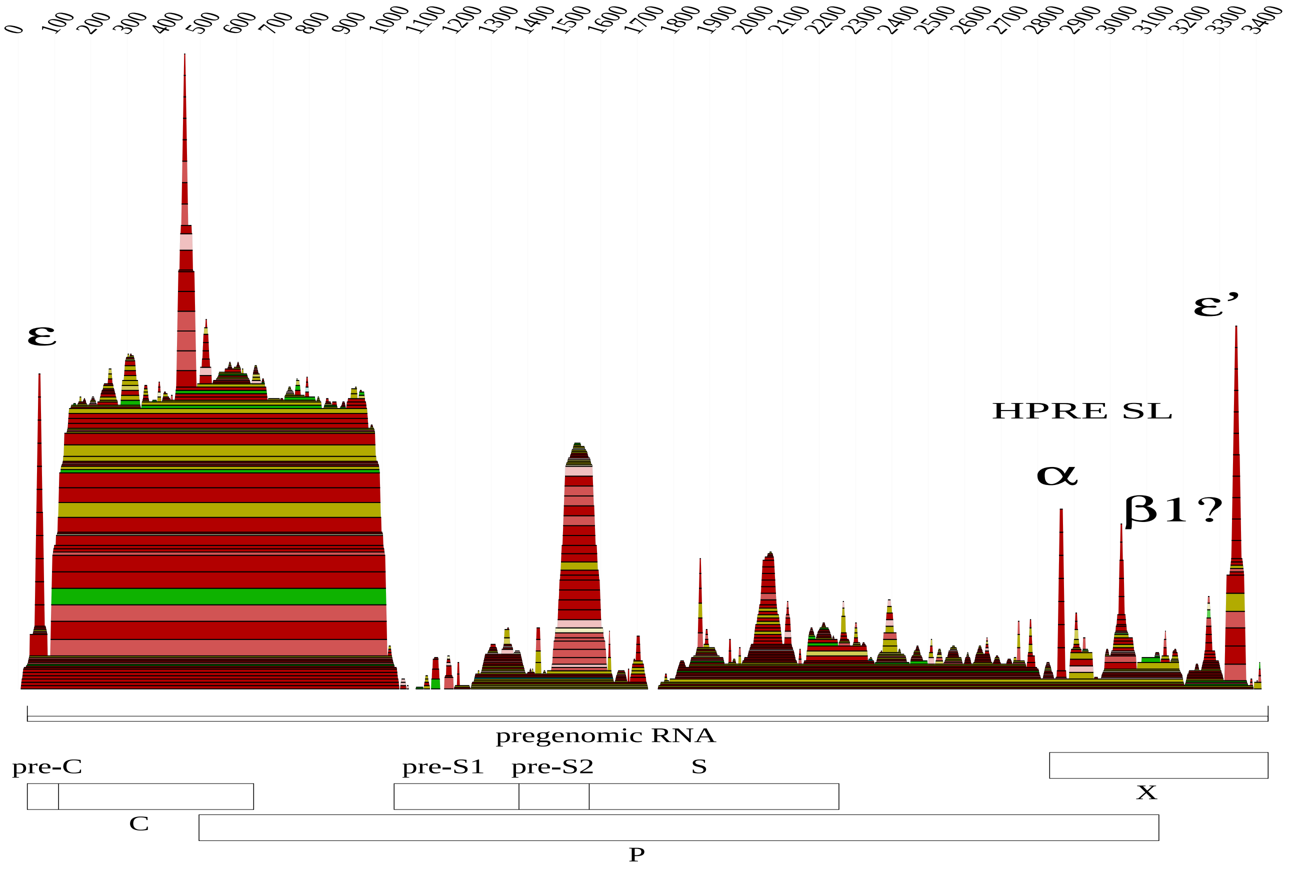
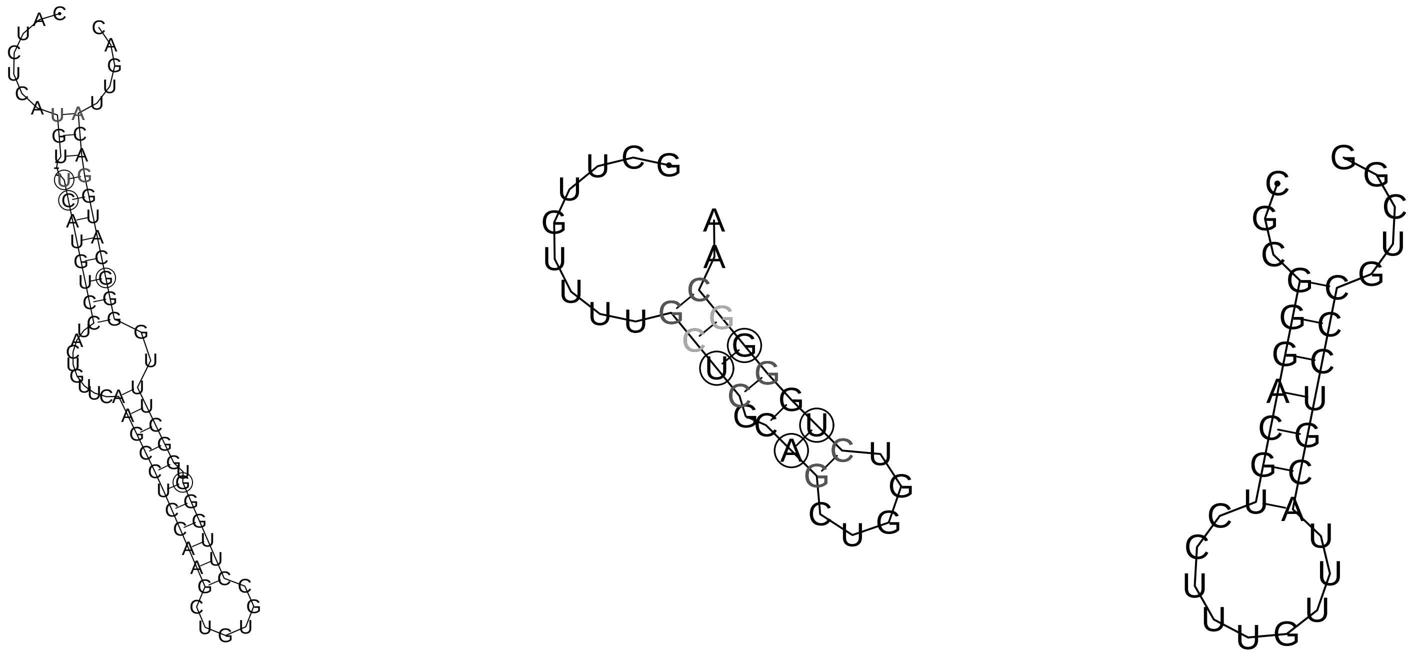


Diagramm der genetischen Organisation vom Hepatitis B Virus.



Die Strukturen des Hepatitis B Virus - Prägenoms.



Die Sekundärstrukturelemente ε, α und $\beta1$.

Conclusio:

- Durch die Integration genetischer Information in die Scoring-Funktion werden Alignments von codierenden Regionen signifikant verbessert.

- Der Anteil an durch 3 teilbaren Gaps wird stark erhöht.

- Das deutet auf die Tendenz hin, Codons nicht zu zerreissen.

- Die resultierenden Alignments sorgen als Input für `alidot` für die Detektion von Elementen, die man sonst nicht immer sieht.

- Das Wissen um die Lokalisation dieser Strukturelemente erleichtert weitere Studien (im Labor).

DANKE!

Peter Stadler

Ivo Hofacker

Peter Schuster

und alle meine Freunde vom TBI!