

**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 code2a1n - 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 | NNPELQAHAGKVFKLVYEAAIQLQVTGVVV | TDLKLNLSVHVSKG       |

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

|              |                                                                 |
|--------------|-----------------------------------------------------------------|
| ADI - MAL    | CUCUAUACAACAGGGAUAGUAGGA-----GAUAUAAGAAGAGCAUAUUGUACU           |
| AE - 90CF402 | UUCCAUACAACAGGAAACAUAAUUGGU-----GAUAUAAGAAAAGCAUAUUGUGAA        |
| AE - CM240   | UUCUAUAGAACAGGAGAUAAUAGGA-----AAUAUAAGAAAAGCAUAUUGUGAG          |
| B - 896      | UUUUUAGCAAGAAGAAACAUAAUAGGA-----GAUAUAAGACAAGCACAUUGUAAC        |
| B - ACH320A  | UUUUUAGCAACAGGACAAAUAAUAGGA-----GAUAUAAGACAAGCACAUUGUAAC        |
| B - BCSG3    | UAUUUAUACAACAGGAGAAUAGUAGGA-----GAUAUAAGACAAGCACAUUGUAAC        |
| B - CAM1     | GUUUAUGCAACAGACAGAAUAAUAGGA-----GAUAUAAGACAAGCACAUUGUAAC        |
| B - D31      | UUUUUAUACAAAAGGAAAAAUAAUAGGA-----GAUAUAAGACAAGCACAUUGUAAC       |
| B - HIV1AD8  | UUUUUAUACAACAGGAGACAUAAUAGGA-----GAUAUAAGACAAGCACAUUGCAAC       |
| B - HXB2     | UUUGUUACA AUAGGAAAA - - - AUAGGA-----AAUAUGAGACAAGCACAUUGUAAC   |
| B - JRCSF    | UUUUUAUACAACAGGAGAAUAAUAGGA-----GAUAUAAGACAAGCACAUUGUAAC        |
| B - LAI      | UUUGUUACA AUAGGAAAA - - - AUAGGA-----AAUAUGAGACAAGCACAUUGUAAC   |
| B - MANC     | UUUCAUGUAACAAGAGCCGUAAACAGGA-----GAUAUAAGACAAGCACAUUGUAAC       |
| B - OYI      | UUUCAUACAACAAAACAAAUAAUAGGA-----GAUAUAAGACAAGCACAUUGUAAC        |
| B - SF2      | UUUCAUACAACAGGAAGAAUAAUAGGA-----GAUAUAAGAAAAGCACAUUGUAAC        |
| B - WEAU     | CUUUUAUACAACAGGAGAAUAAUAGGA-----GAUAUAAGACGAGCACAUUGUAAC        |
| B - YU2      | UUGUAUACAACAGGAGAAUAAUAGGA-----GAUAUAAGACAAGCACAUUGUAAC         |
| B - pNL43    | UUUGUUACA AUAGGAAAA - - - AUAGGA-----AAUAUGAGACAAGCACAUUGUAAC   |
| D - ELI      | CUCUAUACUACAAGAUCAAGAUCA-----AUAAUAGGACAAGCACAUUGUAAU           |
| D - NDK      | CUCUAUACA AUACAGGAAAAAAGAAGAAAACAGGA---UACAUAGGACAAGCACAUUGUAAA |
| O - ANT70    | UACAGCAUGGGAAUAGGGGGAACAGCAGGAAACAGC-----UCAAGGGCAGCUUAUUGCAAG  |
| O - MVP5180  | CGCAGUAUGACACUUAAGAAGUAACAAUACAUCACCAAGAUCAAGGGUAGCUUAUUGUACA   |
| SIVCPZGAB    | UUUUUAUAAUUAAGAAAAUGUAGUAGGA-----GAUACCAGAUUCUGCCUACUGUAAG      |

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:*

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

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:

$$\gamma(g) = -gd \quad (2)$$

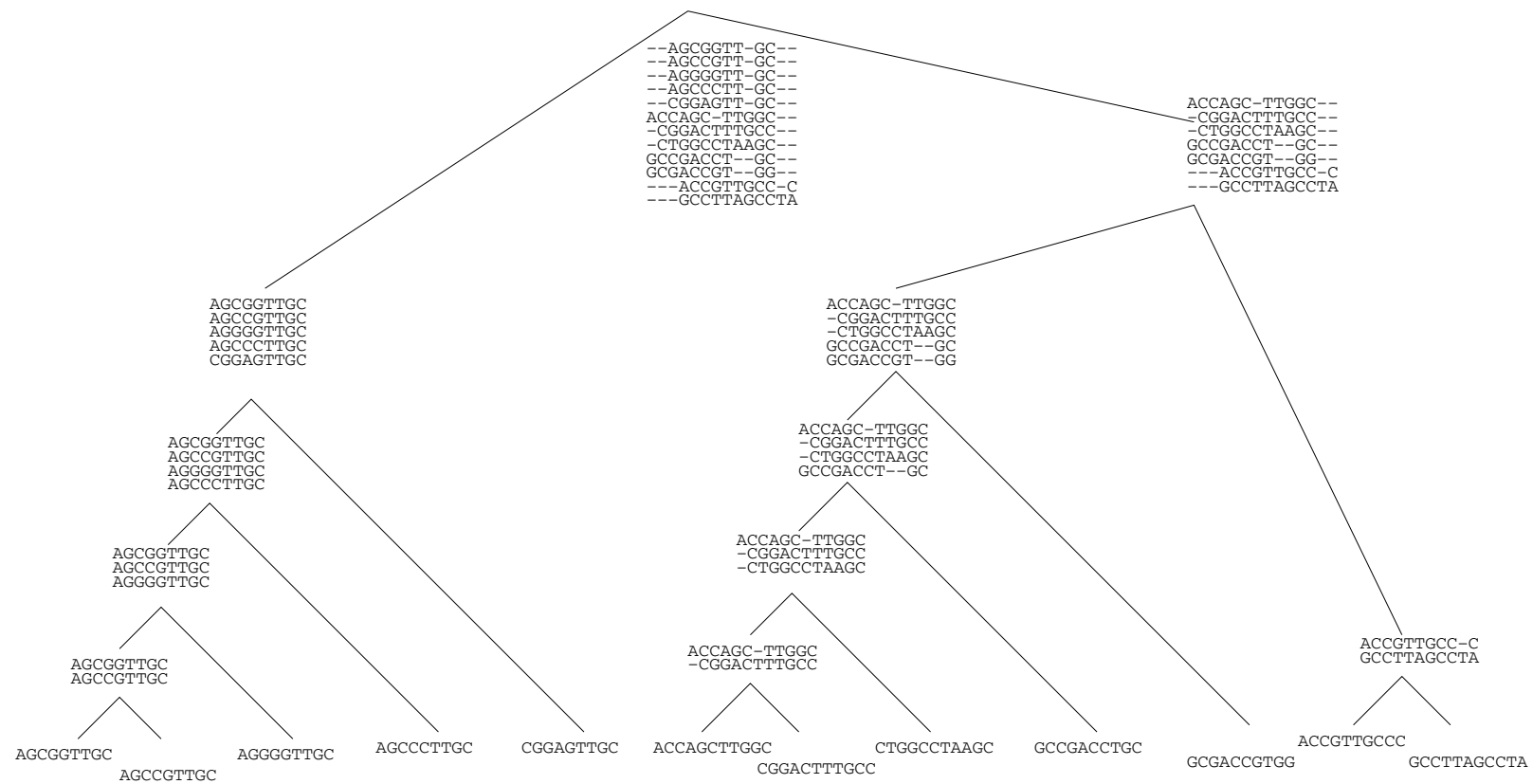
Affine Gap-Penalties:

$$\gamma(g) = -d - (g - 1)e \quad (3)$$

$d$  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

GACGCCCCGCGACCTCTCCGACACCGCTTCCCTCTCCAGTTCGACCACCCCTCCCCCGC

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:

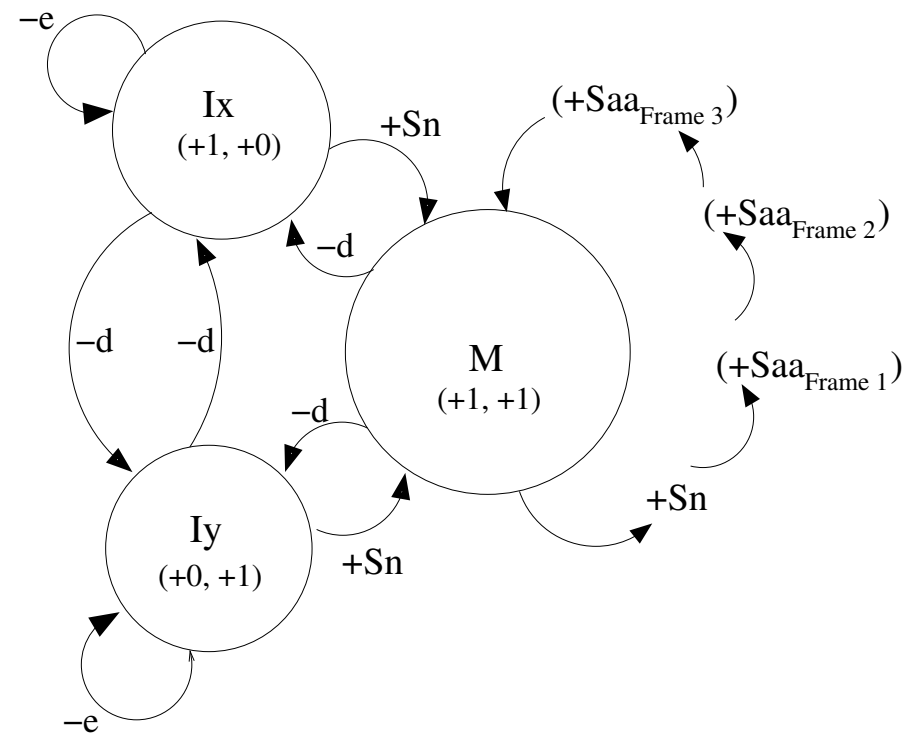
M = (Mis)match state

I<sub>x</sub>, I<sub>y</sub> = InDel states

S<sub>n</sub> = Inkrement des Nukleinsäure-Scores

d, e = Gap-open- und -extension-Penalties

S<sub>aa</sub> = Aminosäure-Scores in drei möglichen Frames



### Die Scoring - Funktion des code2aln - Algorithmus.

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

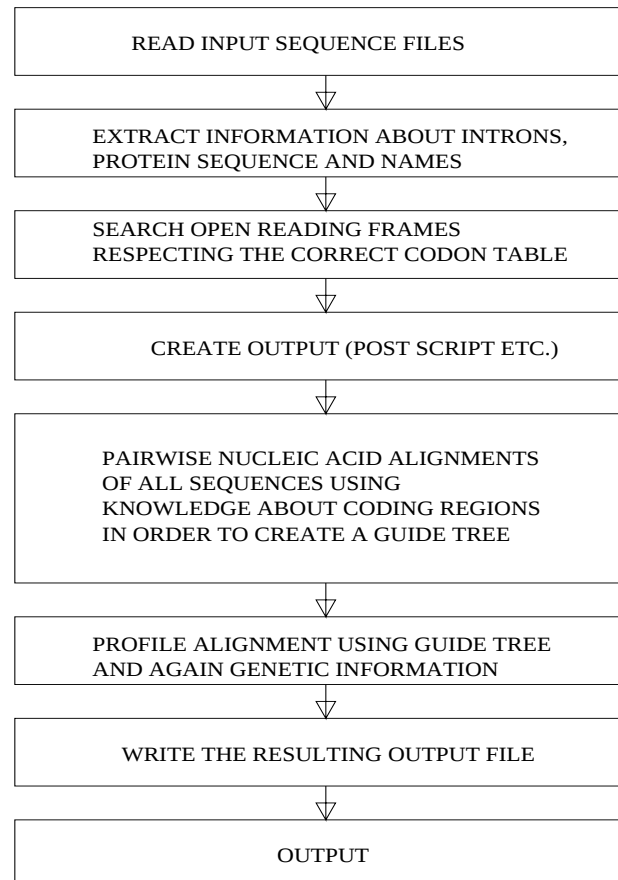
$S$  Gesamt-Score des betrachteten Nukleotid-(Mis)matches;

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

|      |     |     |   |     |     |      |
|------|-----|-----|---|-----|-----|------|
|      | Gln |     |   | Gly |     |      |
|      | Lys |     |   |     |     |      |
| .... | C   | A   | A | G   | --  | .... |
| .... | C   | A   | U | G   | G   | .... |
|      |     | Met |   |     | Ala |      |
|      |     |     |   | Gly |     |      |

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

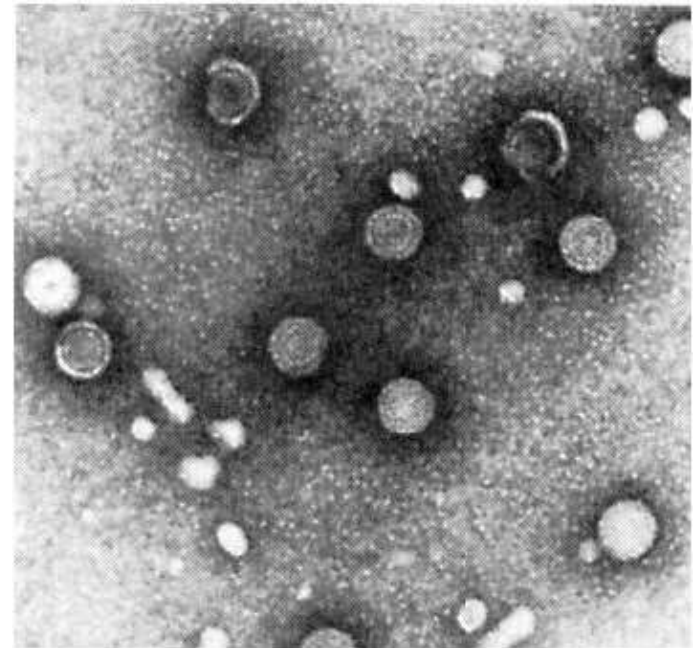
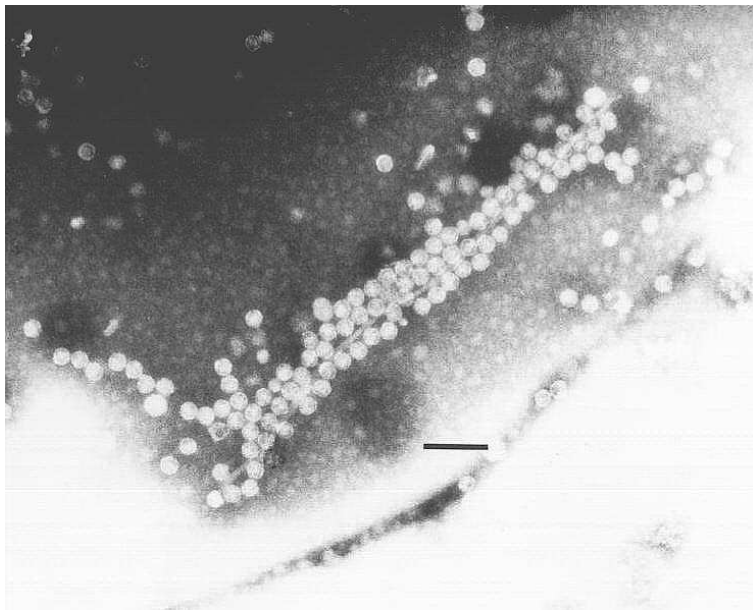
## Der code2aln - Algorithmus im Ablaufdiagramm.

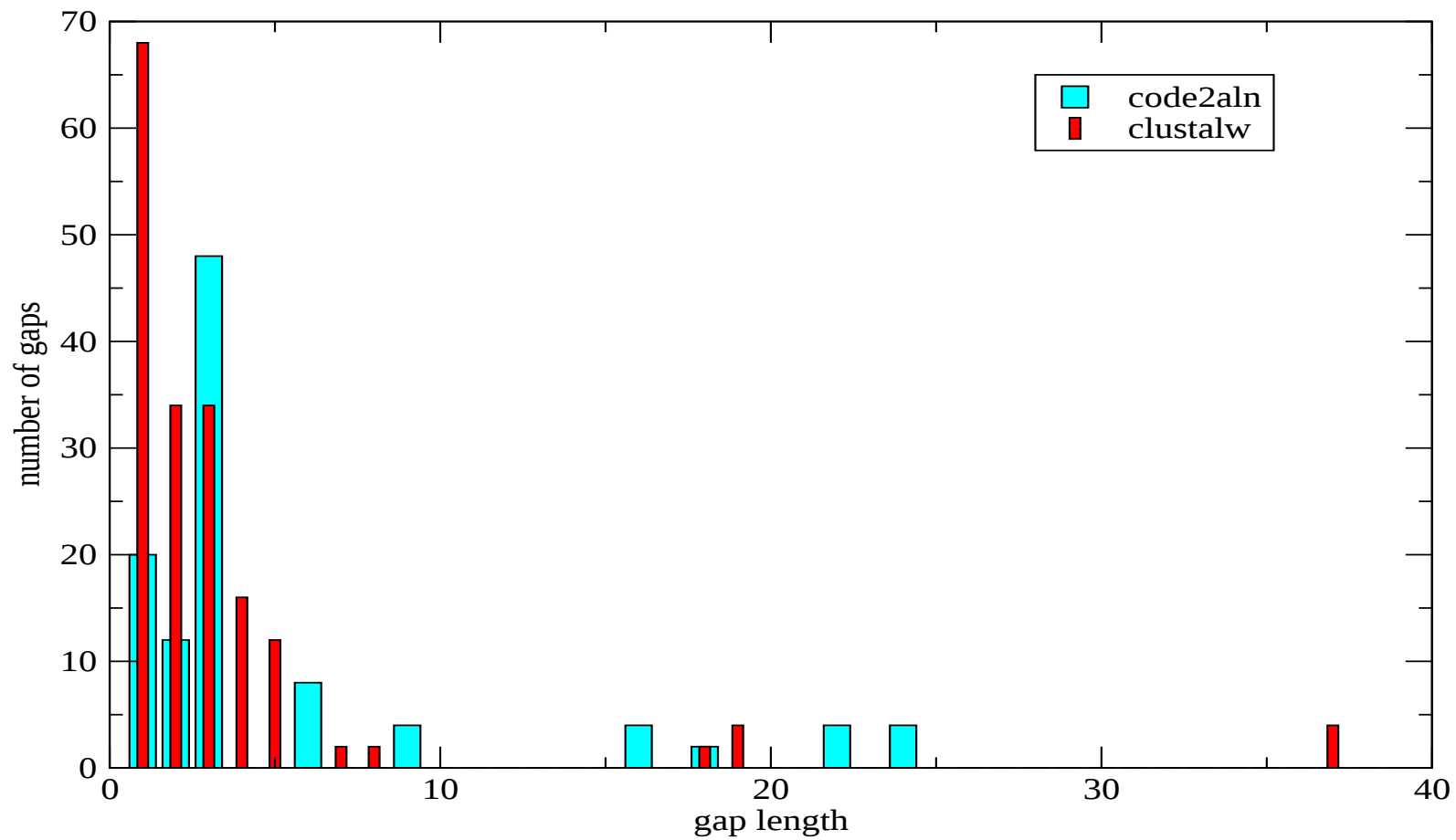


## Anwendung: Leviviridae und Hepatitis B Virus

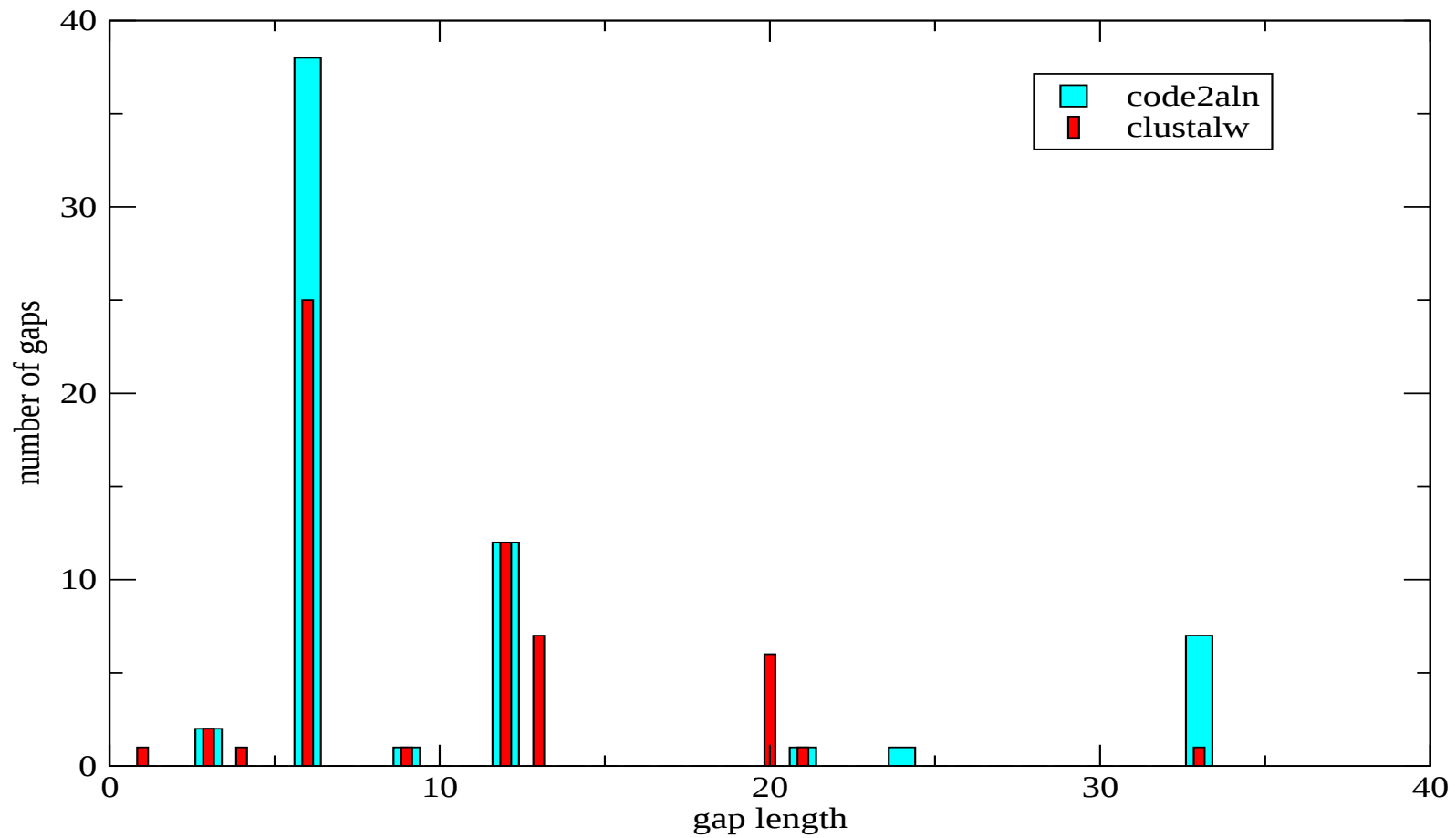
- Virus-Genome sind ein guter Testfall für `code2aln`, weil viele mehrere nicht-codierende und überlappende codierende Bereiche enthalten.
- Die Alignments wurden genutzt als Input für den `Alidot`-Prozess zur Vorhersage konservierter Sekundärstrukturelemente.

Elektronenmikroskopische Aufnahmen der Viruspartikel von Levivirus und Hepatitis B Virus.



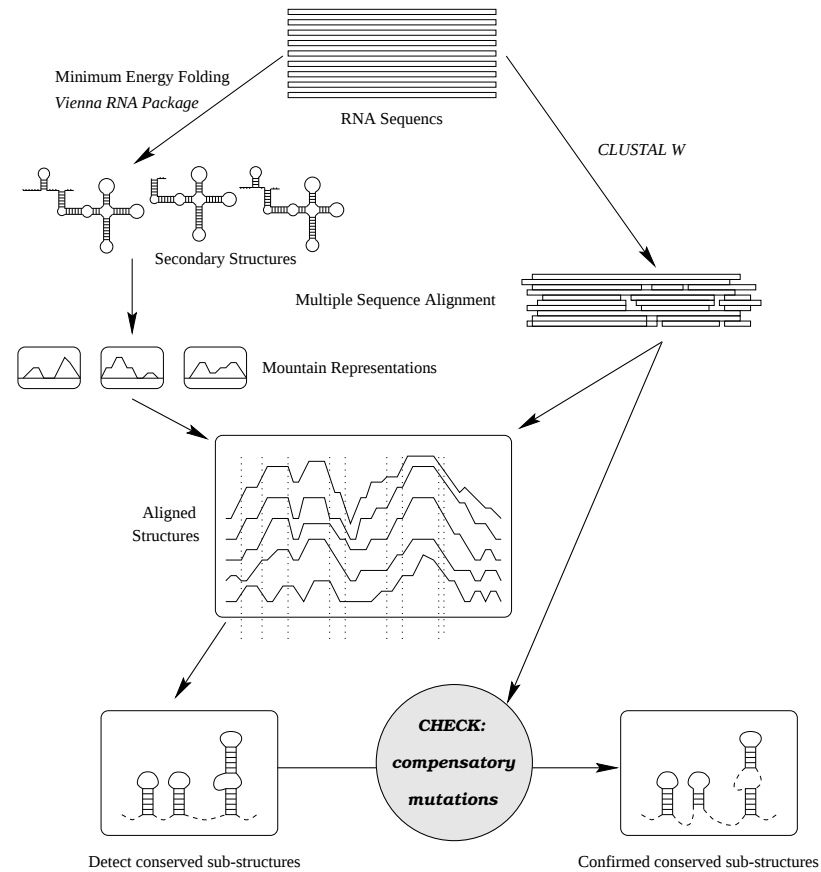


Leviviridae: code2aln hat die starke Tendenz, Codons in codierenden Bereichen nicht zu zerreißen. Dadurch ergibt sich ein viel größerer Anteil an Gaps, die durch 3 teilbar sind.



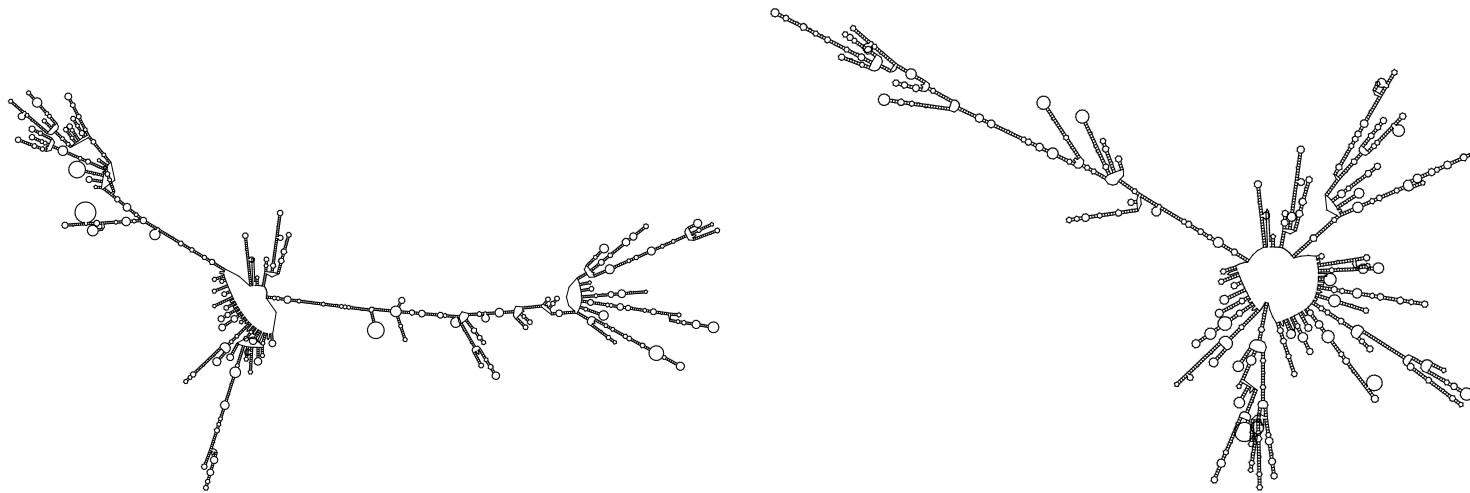
Hepatitis B Virus: code2aln tendiert abermals mehr dazu, Codons in codierenden Bereichen nicht zu zerreißen. Mehr Gaps sind durch 3 teilbar, ansonsten sehr ähnliche Verteilung der Gap-Längen.

## Ablaufdiagramm von alidot:

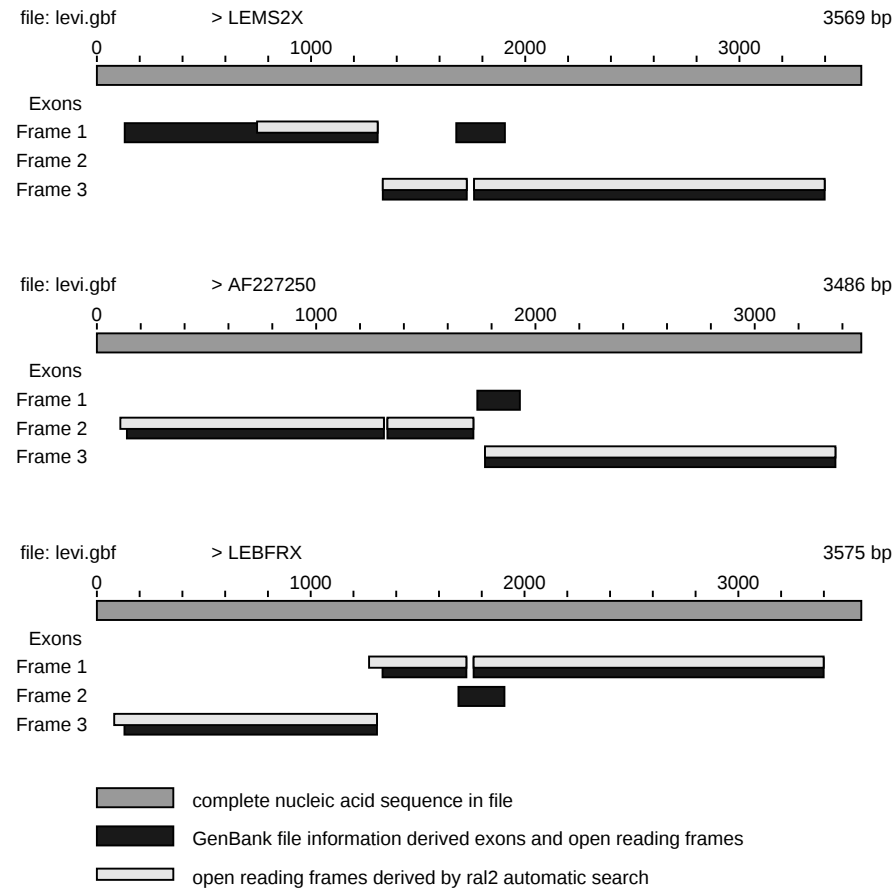




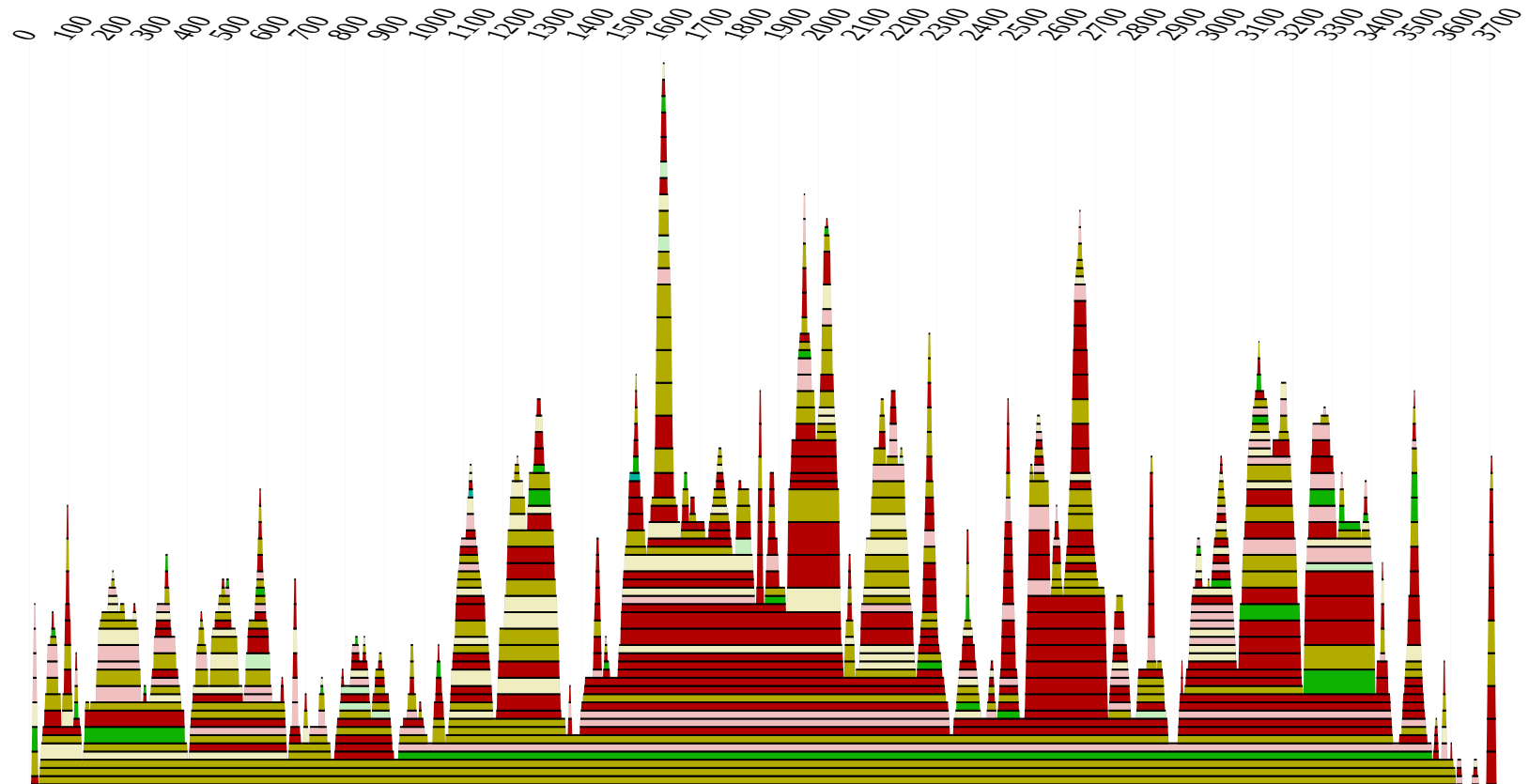
Zwei vorhergesagte Sekundärstrukturen gesamter Hepatitis B Virus-Genome:



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



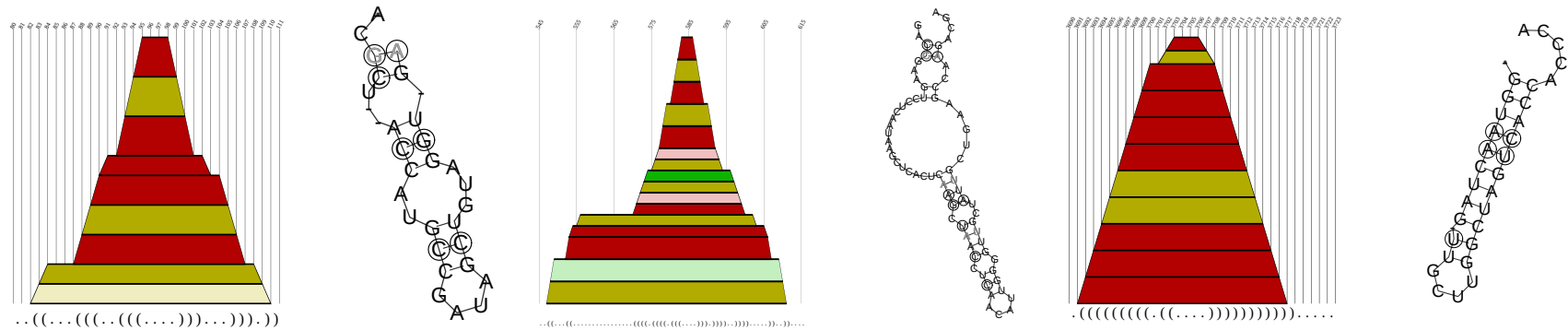
Die Genom-Organisation von 3 Levivirus-Sequenzen, wie code2a1n sie detektiert.



Die Strukturen des kompletten RNA-Genoms von Levivirus.







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

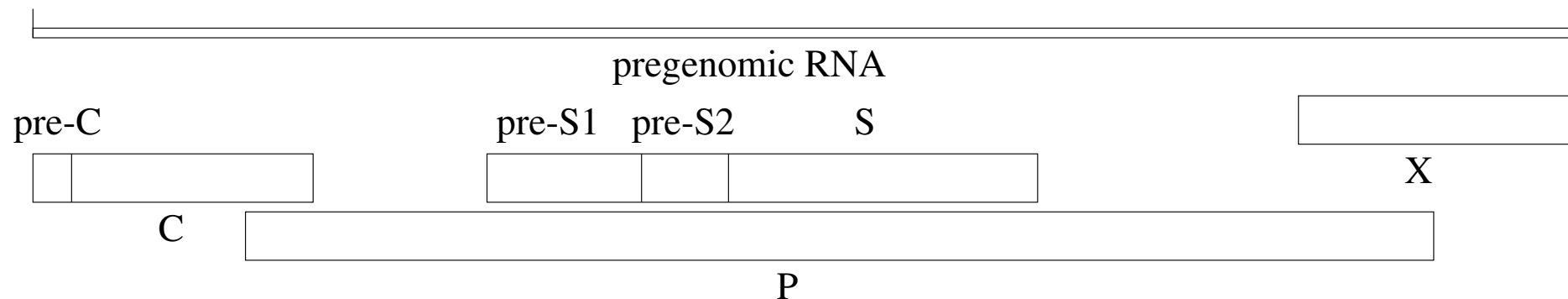
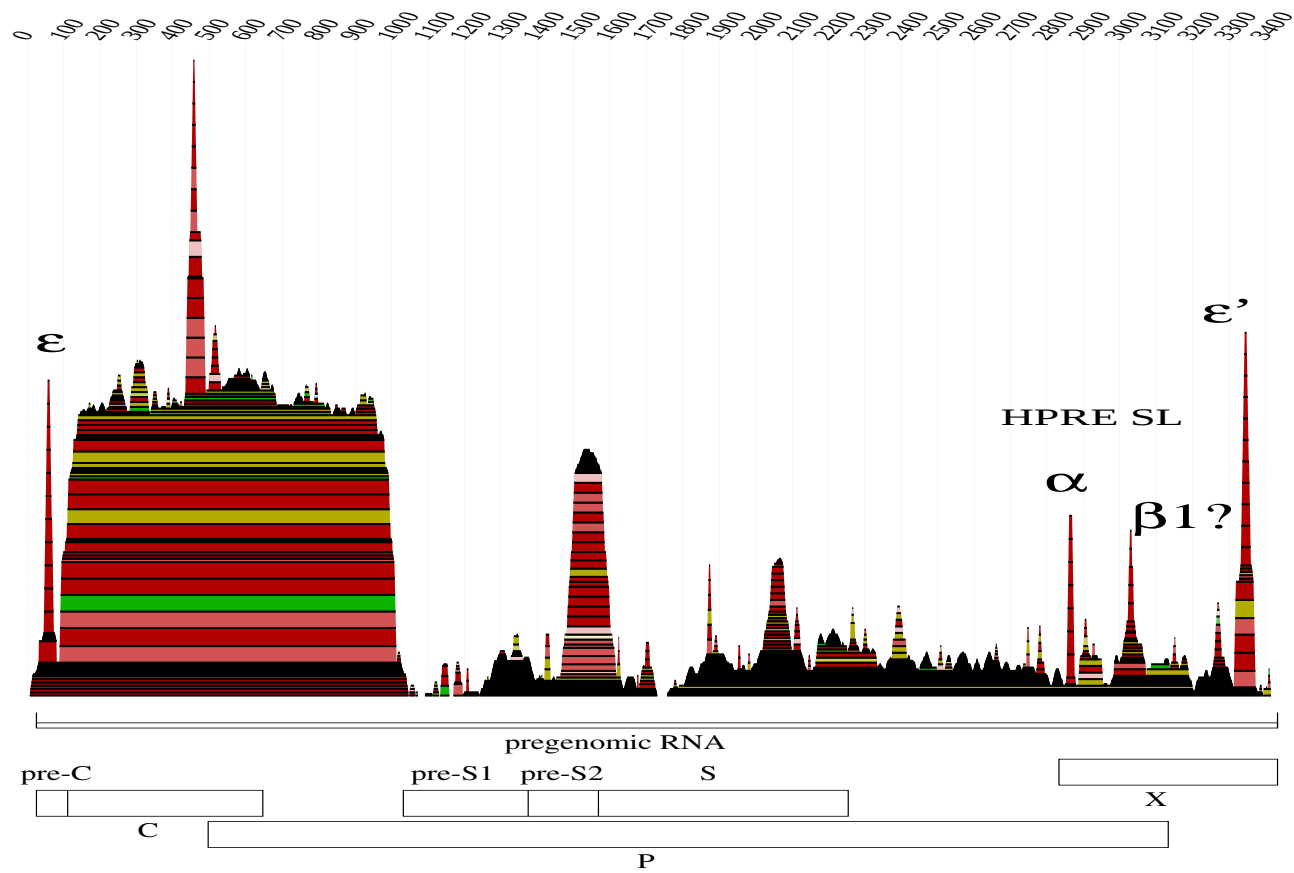


Diagramm der genetischen Organisation vom Hepatitis B Virus.



Die Strukturen des Hepatitis B Virus - Prägenoms.



## 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!