Complex Motif Search: `fragrep`

part of “Genomik der Genregulation”

Sonja Prohaska

Computational EvoDevo
University Leipzig

Leipzig, WS 2011/12

Homology Search for structured RNAs

- “Easy”
  - relatively well-conserved sequence
  - very strong conservation of secondary structure
  - little variation in size
- “Hard”
  - short structured motifs in unconstrained context
  - very variable in sequence **and** length
  - few well-characterized example sequences/structures
  - high flexibility also in structure

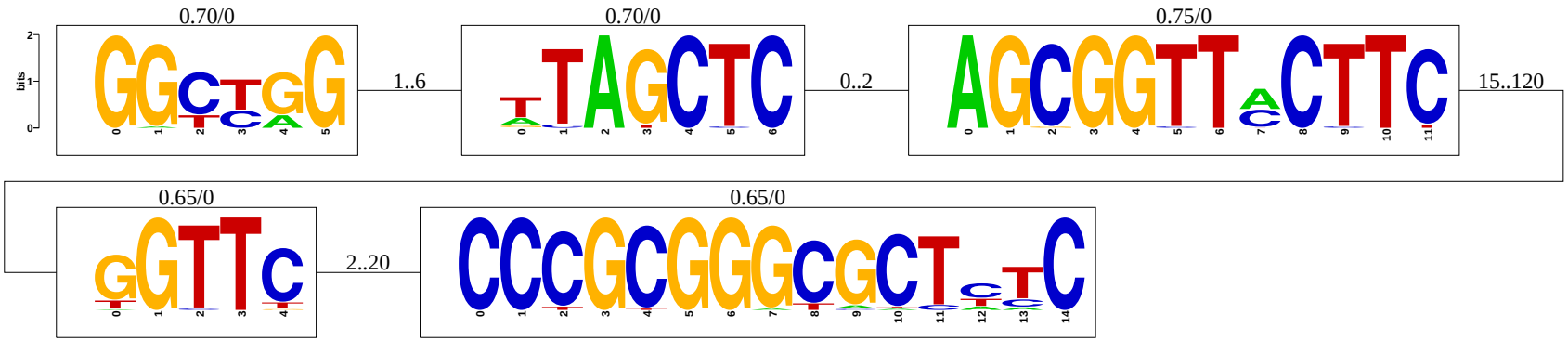
Complex Search Pattern

Given

- motifs as PFMs, for each motif W_i for $i \in 1..k$
  - a threshold match score θ_i
  - a threshold number of deletions d_i
- an order of the motifs
- lower l_i and upper u_i bounds for the sequence length between motifs, better, 5' of the motif W_i
- number δ of allowed motif deletions

Search for the best scoring subsequence of the complex pattern.

Complex Pattern: vault RNA



Implementation

- Match each PFM allowing for d_i deletions
  - this requires a Dynamic Programming version of the MATCH algorithm
  - fractional programming solution
- match the most informative PFMs first and search their surrounding for matches of others
  - calculate information content H_i for each motif W_i
- Match the complex pattern allowing for δ deletions of motifs
  - this requires Dynamic Programming

Matching PFMs *without* deletions

Given a PWM W and a nucleotide sequence x, both of length ℓ, compute the MATCH score:

$$\text{score} = \frac{\text{Current} - \text{Min}}{\text{Max} - \text{Min}} \tag{1}$$

where

$$\text{Current} = \sum_{i=1}^{\ell} H_i W_{i,x_i} \tag{2}$$

$$\text{Max} = \sum_{i=1}^{\ell} H_i f_i^{\max} \tag{3}$$

$$\text{Min} = \sum_{i=1}^{\ell} H_i f_i^{\min} \tag{4}$$

Matching PFMs *without* deletions

- $W_{i,\alpha}$: frequency of nucleotide α at position i
- H_i: information content of column i of the matrix W
  $H_i = \sum_{\alpha \in [A,C,G,T]} W_{i,\alpha} \ln W_{i,\alpha}$
- $f_i^{\max}$: frequency of most frequent nucleotide in column i
- $f_i^{\min}$: frequency of least frequent nucleotide in column i

$$\text{score}(W, x) := \max_A \frac{\sum_{i=1}^{\ell} H_i W_{i,x_i} - H_i f_i^{\min}}{\sum_{i=1}^{\ell} H_i f_i^{\max} - H_i f_i^{\min}} \tag{5}$$

Matching PFMs *with* deletions

- Given a PFM W of length $(\ell + d)$ and a sequence x of length ℓ (deletion of matrix columns allowed)
- What is the best possible match score between W and x after deleting d columns of W?
- Deleting columns achieved formally through an alignment A

$$\begin{aligned} E_{i,j} &:= H_i \left(W_{i,x_j} - f_i^{\min} \right) \quad \text{and} \\ D_{i,j} &:= H_i \left(f_i^{\max} - f_i^{\min} \right), \end{aligned}$$

$$\text{score}(W, x) := \max_A \frac{\sum_{(i,j)\in A} E_{i,j}}{\sum_{(i,j)\in A} D_{i,j}} \tag{6}$$

where A are the alignments

needed to calculate the maximum of a fraction:
find the optimal value for threshold $\theta \in [0, 1]$

$$S_{\theta,A} := \sum_{(i,j)\in A} E_{i,j} - \theta \sum_{(i,j)\in A} D_{i,j} \quad (\geq 0)$$

the best alignment is obtained by Dynamic Programming

$$S_\theta(i,j) = \max \begin{cases} S_\theta(i,j-1) \\ S_\theta(i-1,j-1) + E_{i,j} - \theta D_{i,j}. \end{cases}$$

do a "binary search" on values for θ to aproach the optimal θ.
After m steps we obtain precision 2^{-m}.

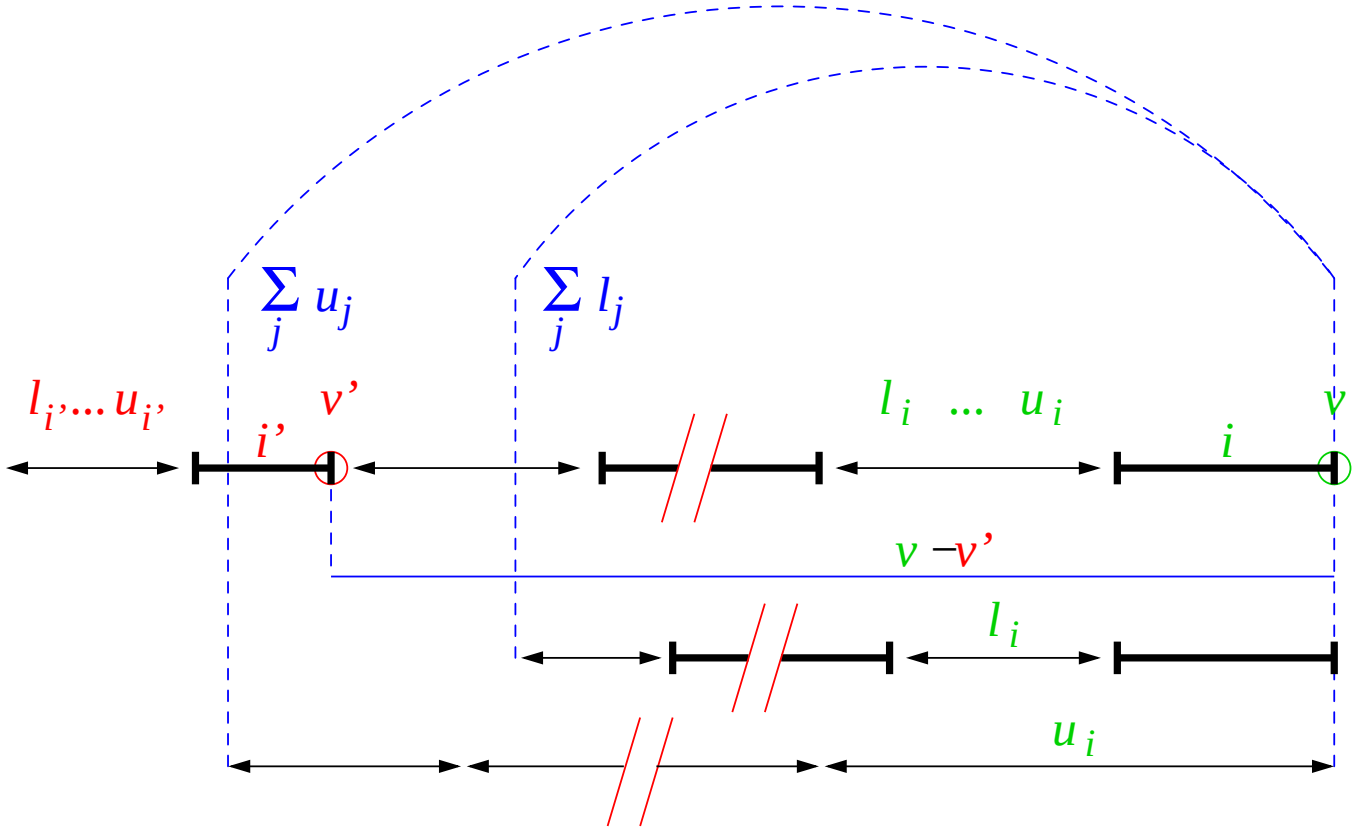
Obtain Partial Ordering of Motif Occurences

- denote by (i, ν) an occurrence of motif W_i at position ν
- a motif (i', ν') is a predecessor of (i, ν) if:

$$(i', \nu') \prec (i, \nu) \Leftrightarrow \sum_{j=i'}^{i} u_j \geq \nu - \nu' \geq \sum_{j=i'}^{i} (l_j + n_j + d_j)$$

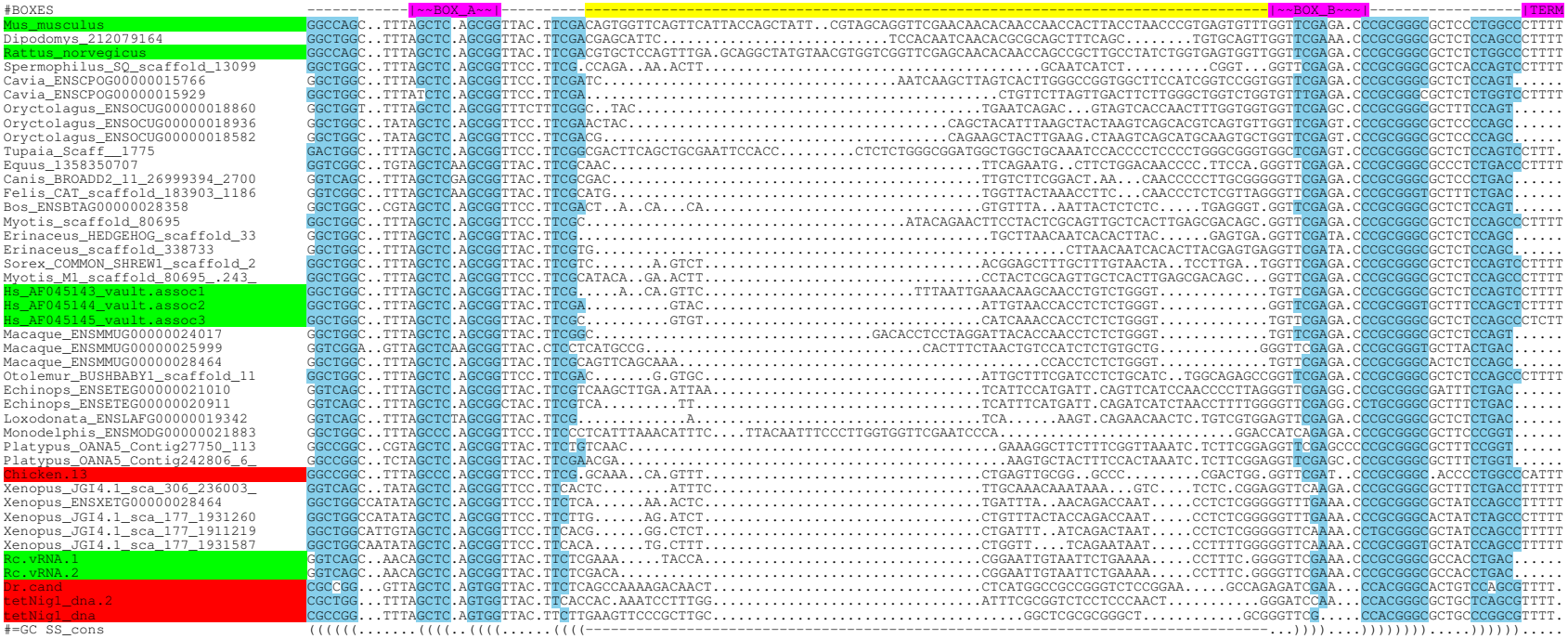
  where n_j is the length of motif j
- find chains of length $k - \delta$ in this partially ordered set
- derive the chain with the best score using Dynamic Programming

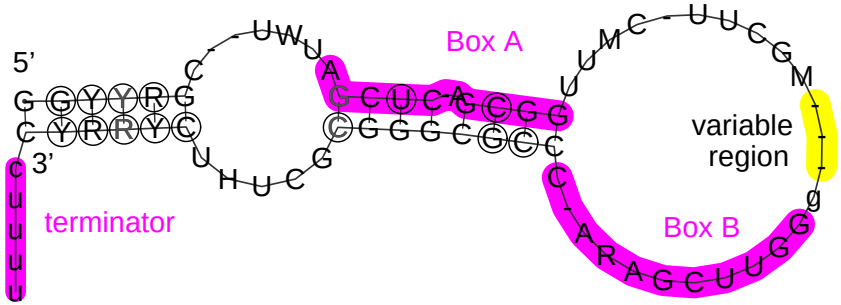
a motif (i', ν') is a predecessor of (i, ν) if...



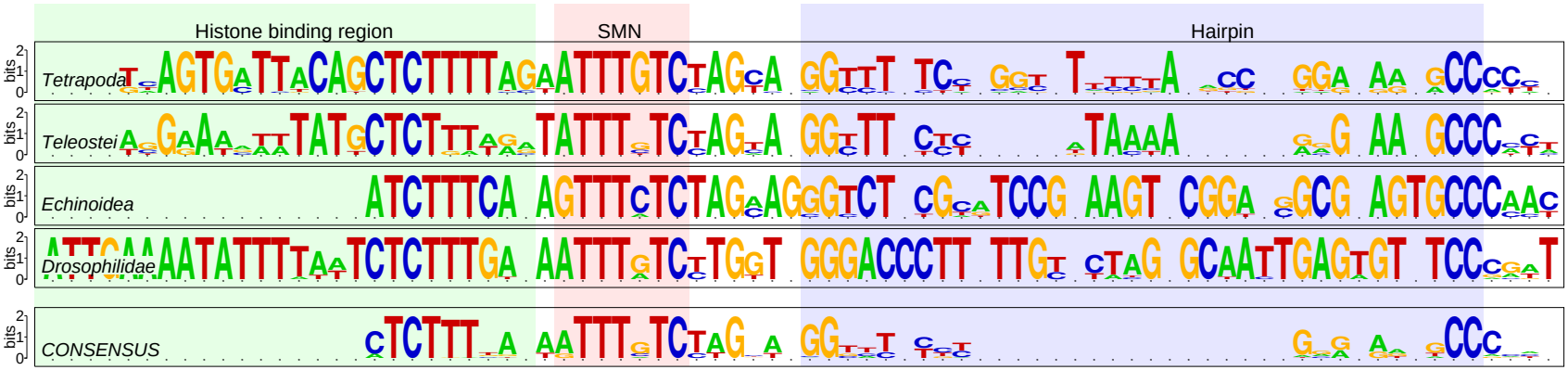
$\sum_{j=i'}^{i} u_j$ . . . is the sum over all upper limits that belong to a ν_j where $\nu' < \nu_j <= \nu$

A Survey for vault RNAs





U7 snRNAs




```
#                    |<Histone-binding-region>|.|<SMN>|.......<<<<<.<<<.<<<<.....>>>>..>>>.>>.>>>....
Homo                 .....CAGTG.TTACAGCTCTTTTAGAATTTGTCTAGTA..GGCTT.TCT.GGC.TTTTT..ACC..GGA.AA.GCCCCT.
Mus                  .....AAGTG.TTACAGCTCTTTTAGAATTTGTCTAGCA..GGTTT.TCT.GAC..TTCG..GTC..GGA.AA.ACCCCT.
Xenopus_l            .....AAGTG.TTACAGCTCTTTTACTATTTGTCTAGCC..GGTTT.TTA.C....TCT.....G..TTG.GA.GCCACA.
Takifugu             ....AGGAATGATT..GCTCTTTAGATATTTCTCTAGTA..GGCTT.TTC.....ATACA.......GAG.AA.GCCCCCT
Petromyzon-c1        ..........ATTGAGGATCTTTGAC.TTTTGTCTTTGTGTGGTGCACC.......GAAA........GGAGC.ACC....
Branchiostoma-c1     .....ACTGG.TAAC.GCTCTTTCAC.CTTTATCCGCG...GGGTA.A........CCT.........T.TA.TCCGTA.
Branchiostoma-c2     .....GAGTG.TAAC.GTTCTTTCAC.CTTTATCCGCG...GGGTA.........ACCTA..........TA.TCCGTT.
Psammechinus_1       ................ATCTTTCA.AGTTTCTCTAGAA.GGGTCT.CGCGTCCG.AAGT.CGGA.GGCG.AGTGCCCAAC
Bombyx_mori-c1       TCCATCAAT.ATGTTCTATCTTTTA..ATTTATCGAAAA.CGGTCA.AG.A....ACTAGTC....G.CT.TG.GCC....
Bombyx_mori-c2       AAGATTTTG.GTGTGTAATCTTTAACTGTTTATCTTTTG.CGGTAGG...T.AGCGGCTTGGCT.......CT.GCC....
Dr_melanogaster      ATTGAAAAT.TTTTATTCTCTTTGA.AATTTGTCTTGGT..GGGACCCTT..TGT.CTAG.GCA.TTGAGTGT.TCCCGTT
#                    |<Histone-binding-region>|.|<SMN>|.......<<<<<.<<<.<<<<.....>>>>..>>>.>>.>>>....
```

Literature

[Mosig, 2007] Axel Mosig, Julian L. Chen, Peter F. Stadler . *Homology Search with Fragmented Nucleic Acid Sequence Patterns*. WABI 2007 (R. Giancarlo & S. Hannehalli, eds.), Springer Verlag, Berlin, LNBI 4645, pp. 335-345.