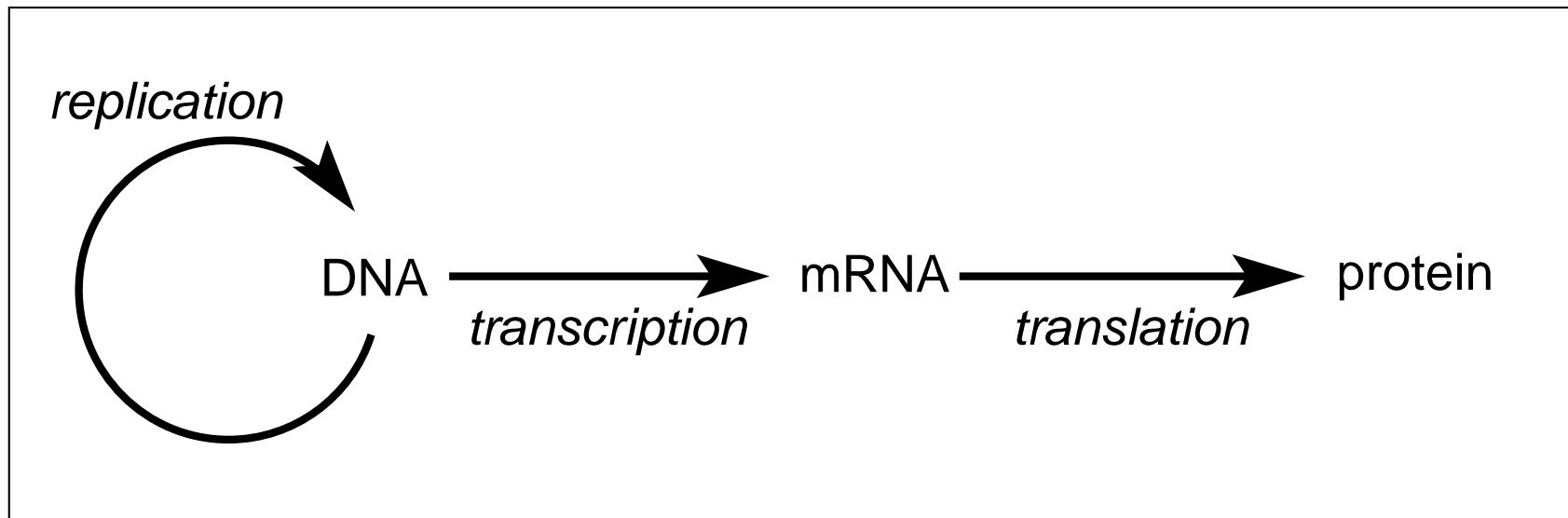


RNA SECONDARY STRUCTURE PREDICTION WITH A FEW HOMOLOGS

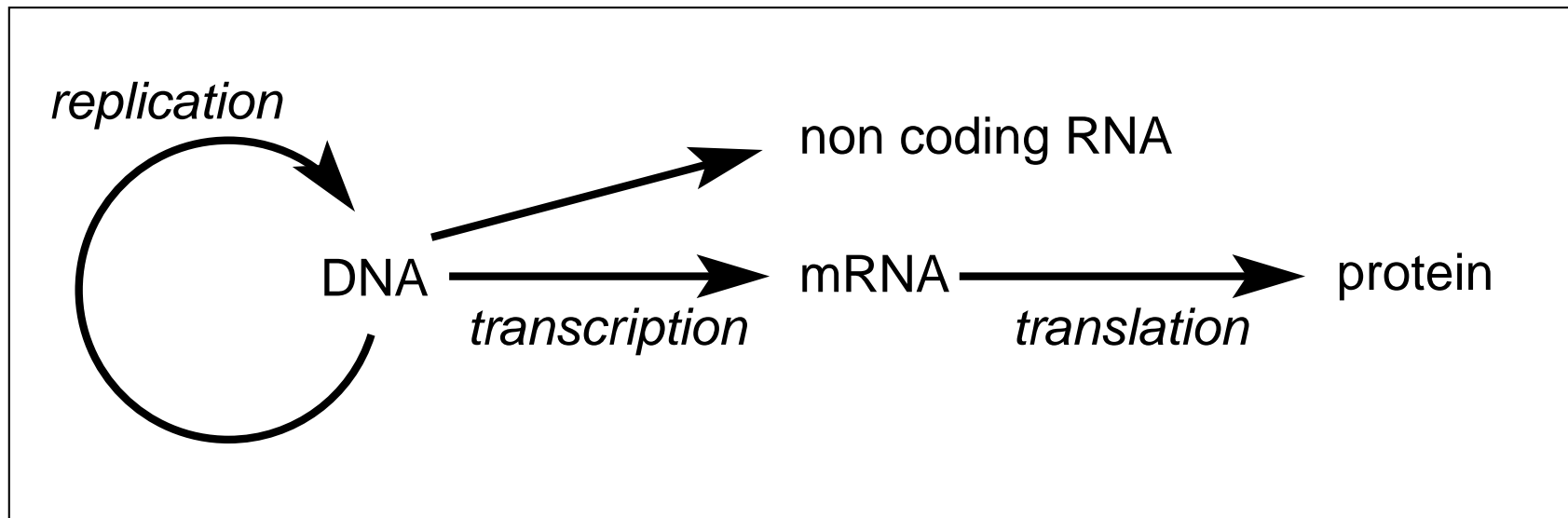
Centre For Artificial Intelligence
New University of Lisbon
<http://centria.di.fct.unl.pt>

OLIVIER PERRIQUET

The central dogma [Crick 1957]



The central dogma [Crick 1957]



RNA seen by an alborist

- The oriented sequence of nucleic acids (or bases) *Adenin*, *Cytosin*, *Guanin*, *Uracyl* is seen as a word on the alphabet { A, U, C, G }.

. . . gcuaauagcucagcugggagagcgccugcuuugcacgccacguaguacauacggggug
cccaucauagucgucagagaaggaggucugcgguucgaucccgcauagcguccacca . . .

- The bases can pair and while doing so, they drive the folding of the molecule in the 3D space.
- The base pairings obey some rules
 - The standard pairings are **A=U** and **C≡G** [Watson-Crick] and less often **G=U** [wooble]
 - They have a tendency to stack to form stems.

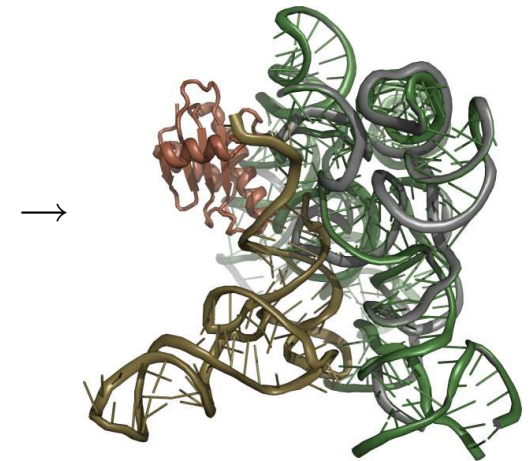
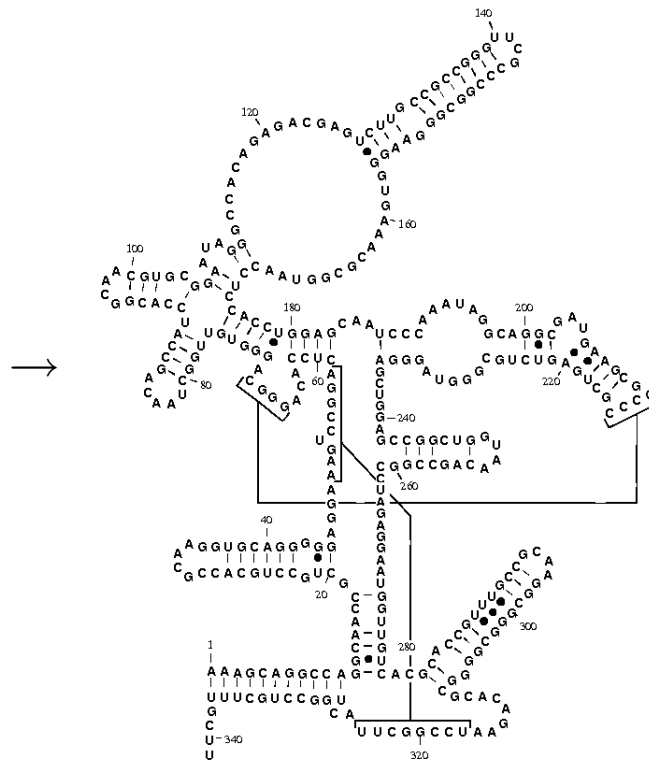
RNA Structure :: levels of description

Primary structure: the nucleic acids sequence

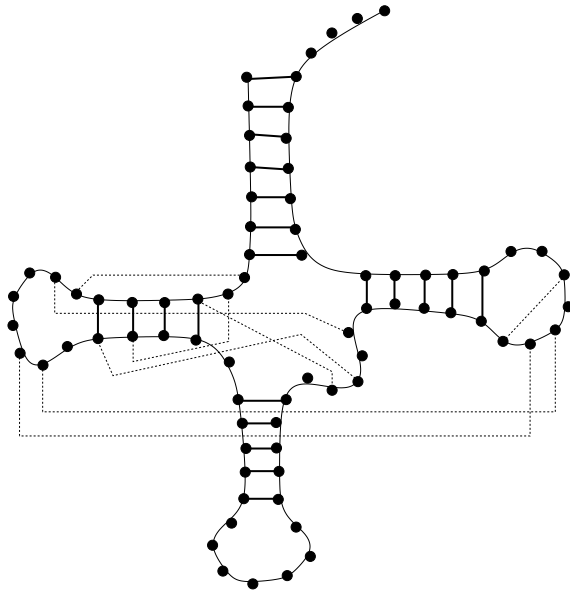
Topological structure: the graph of intra-molecular base-pairing interactions

Three-dimensional structure: the exact atomic positions

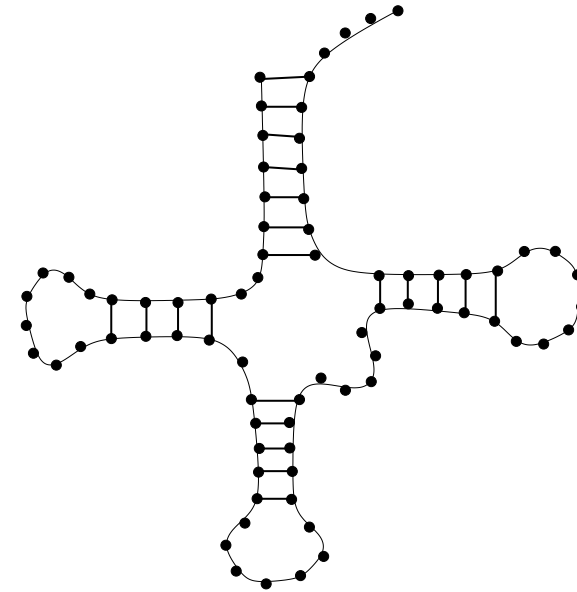
...gcuauagcucagcuggg
 agagcgcugcuuugcacgc
 cacguaguacauacggggug
 cccauc auagucgucagaga
 aggaggucugcgguucgauc
 ccgcauagcguccacca...



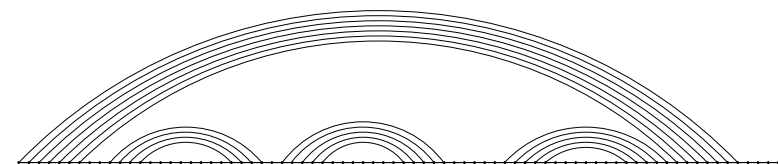
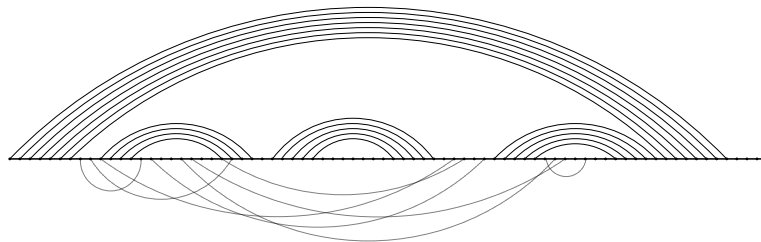
RNA topological structure :: tertiary and secondary



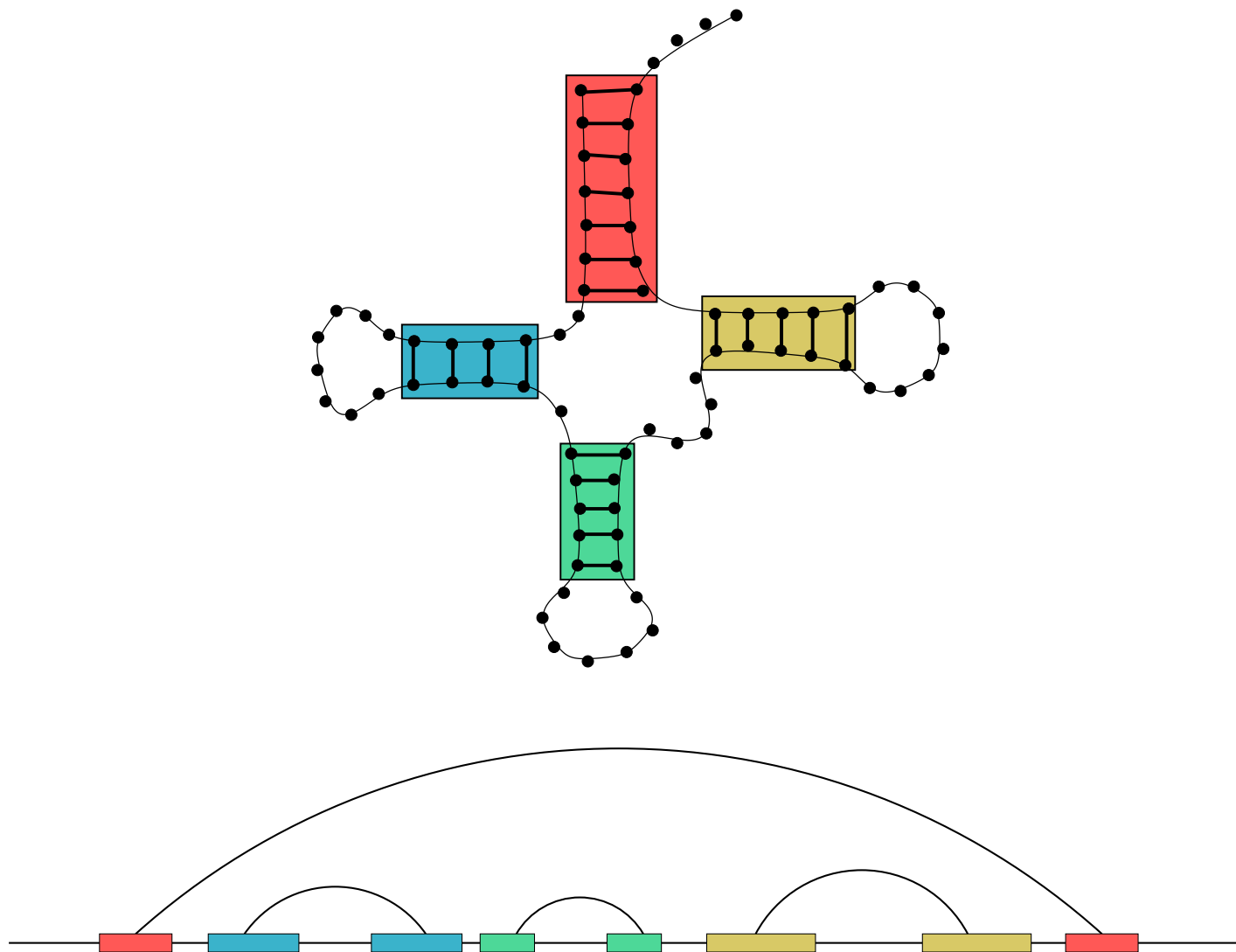
tertiary structure



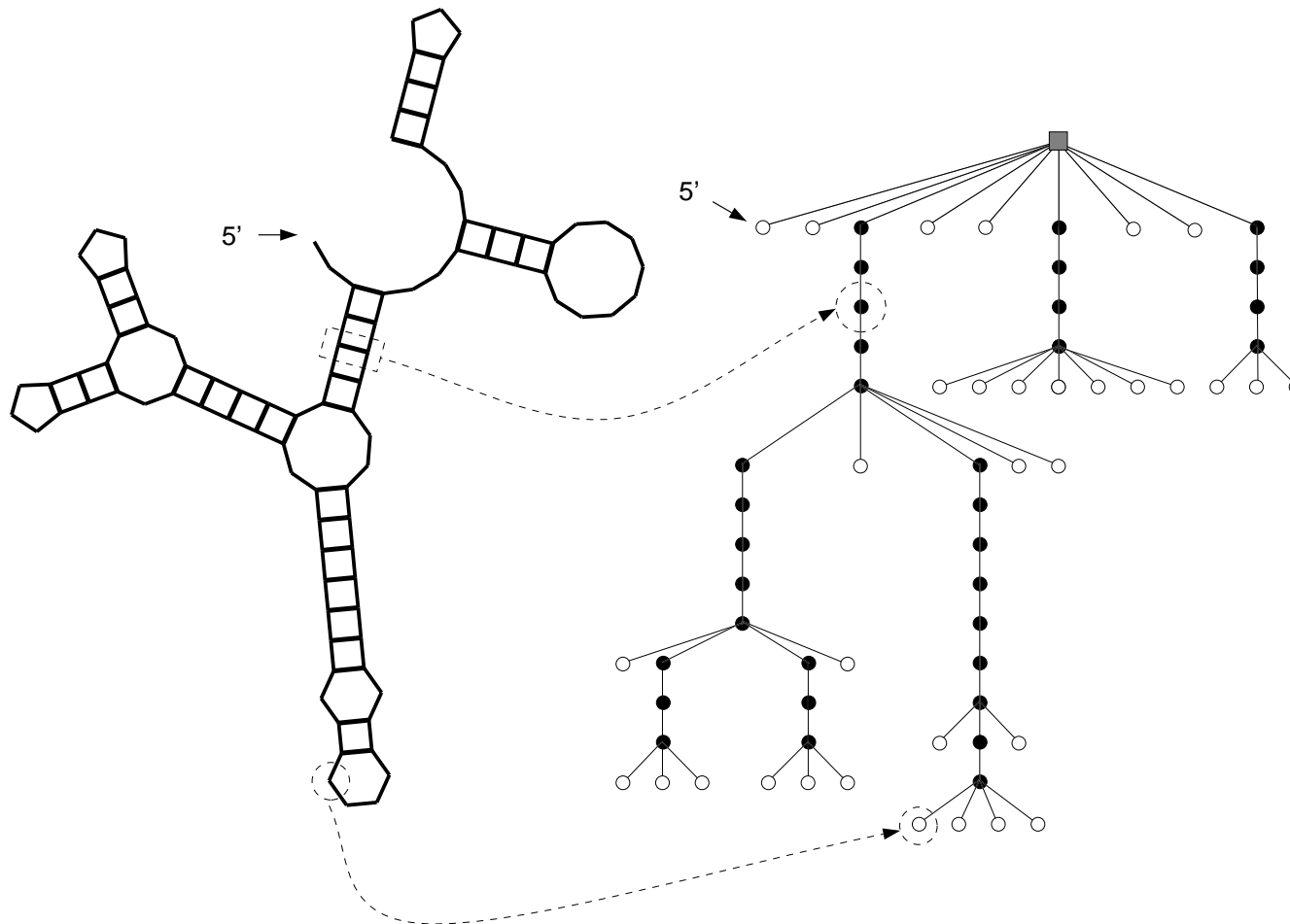
secondary structure



Secondary structure at stem level



One to one correspondance between secondary structures and rooted oriented Trees



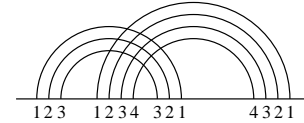
Secondary structure :: generative elements



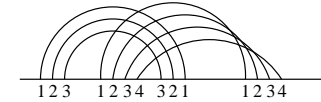
nested stems
(father and son)



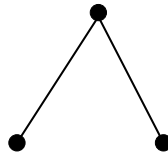
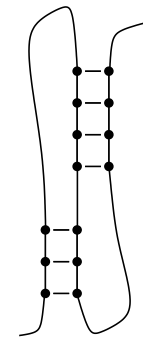
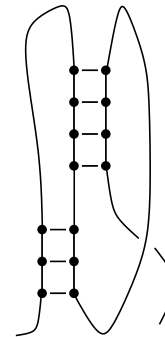
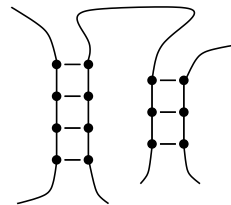
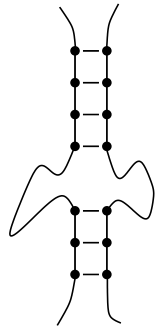
branching stems
(brothers)



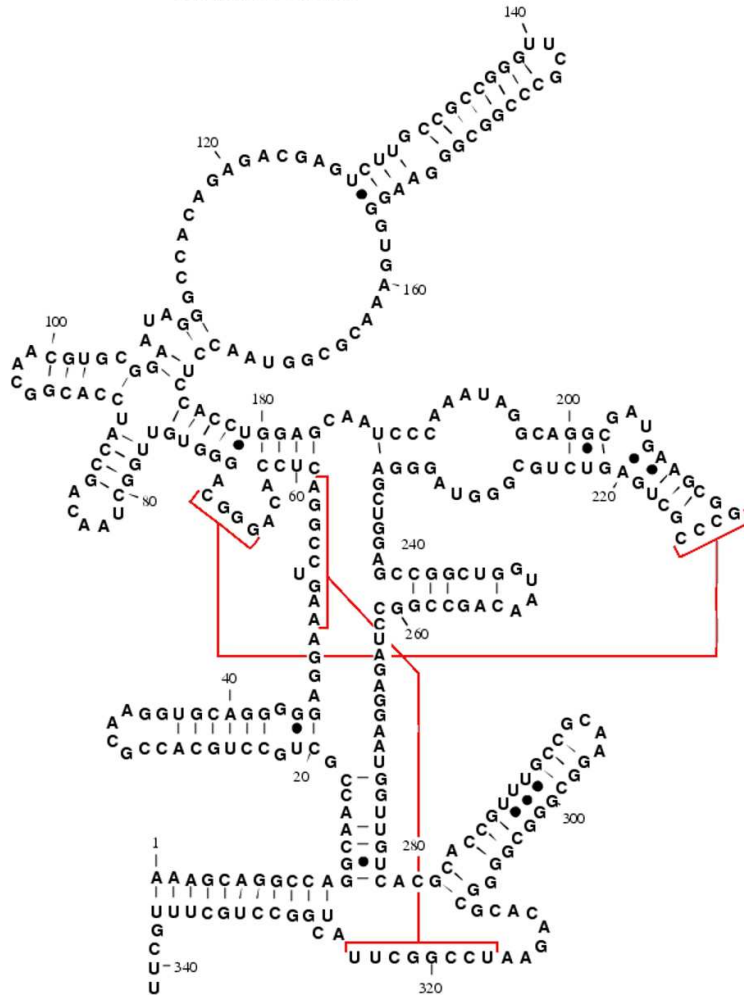
pseudoknot
EXCLUDED



pseudoknot
EXCLUDED



Alcaligenes eutrophus
RNase P RNA



Does it make sense to consider Secondary Structure ?

From a biological point of view
RNA folding is believed
to be partially hierarchical

From an algorithmic viewpoint,
secondary structure is far
more easy to compute
whereas it usually covers
80% to 90% of all the pairings

Secondary structure prediction

Computational prediction of the secondary structure given just the nucleic acid sequence.

depends on context

- 1. a single sequence (eg. UTR or unknown new sequence)
- 2. a large family of homologs (eg. tRNA, rRNA)
- 3. only a few (sometimes putative) homologs (eg. RNase)

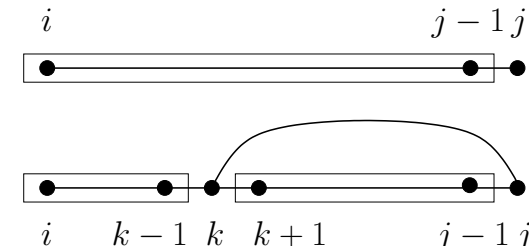
context 1 – a single sequence

combinatorial optimization
in a simplified thermodynamic model

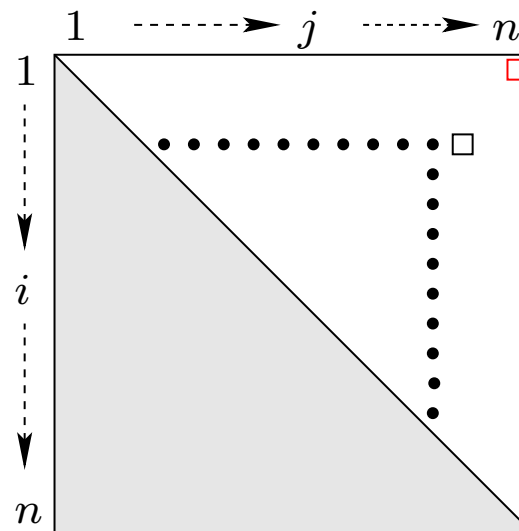
The underlying model :: Nussinov recursions – $O(n^2)/O(n^3)$

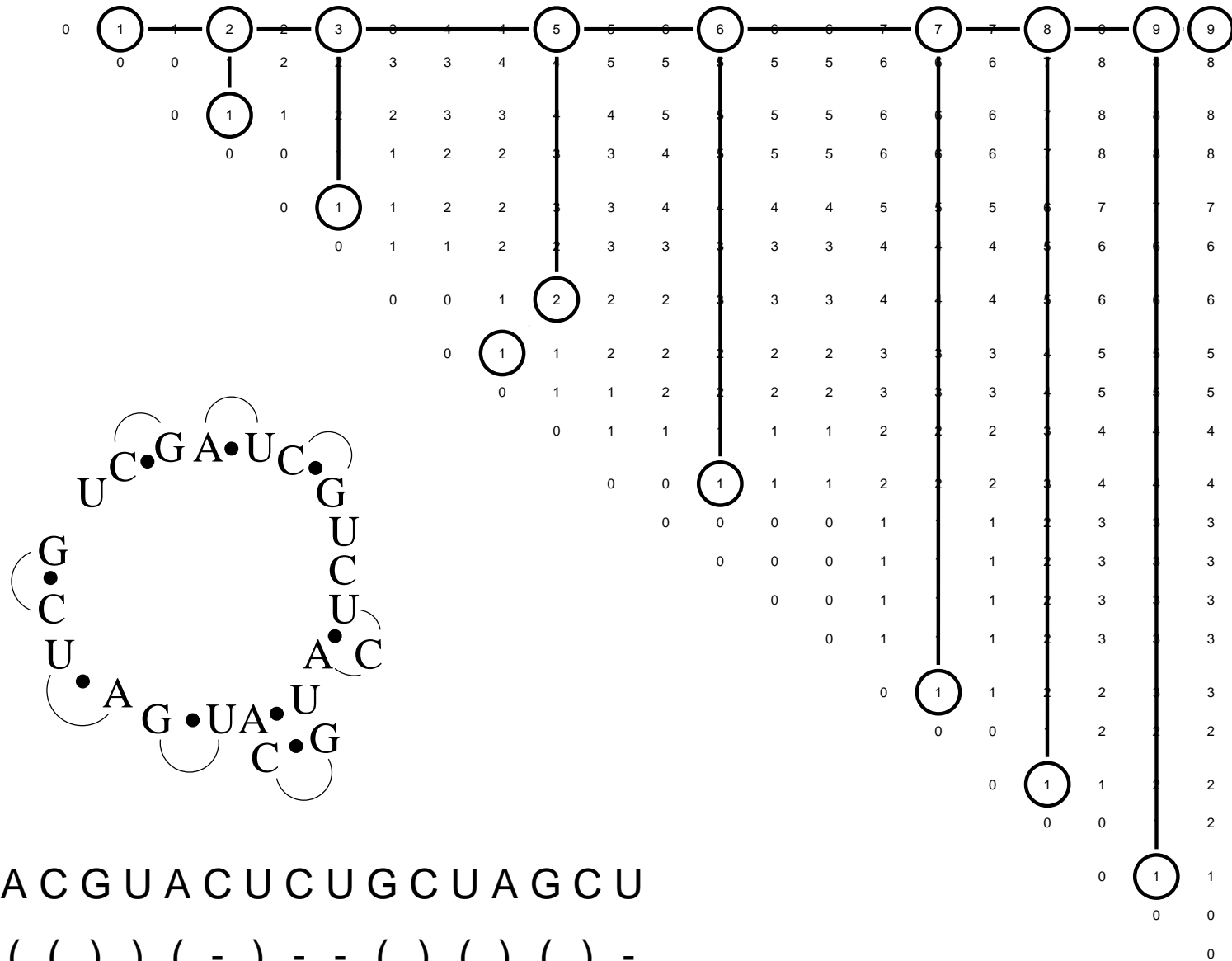
maximizing the number of base pairings A=U, C=G, G=U

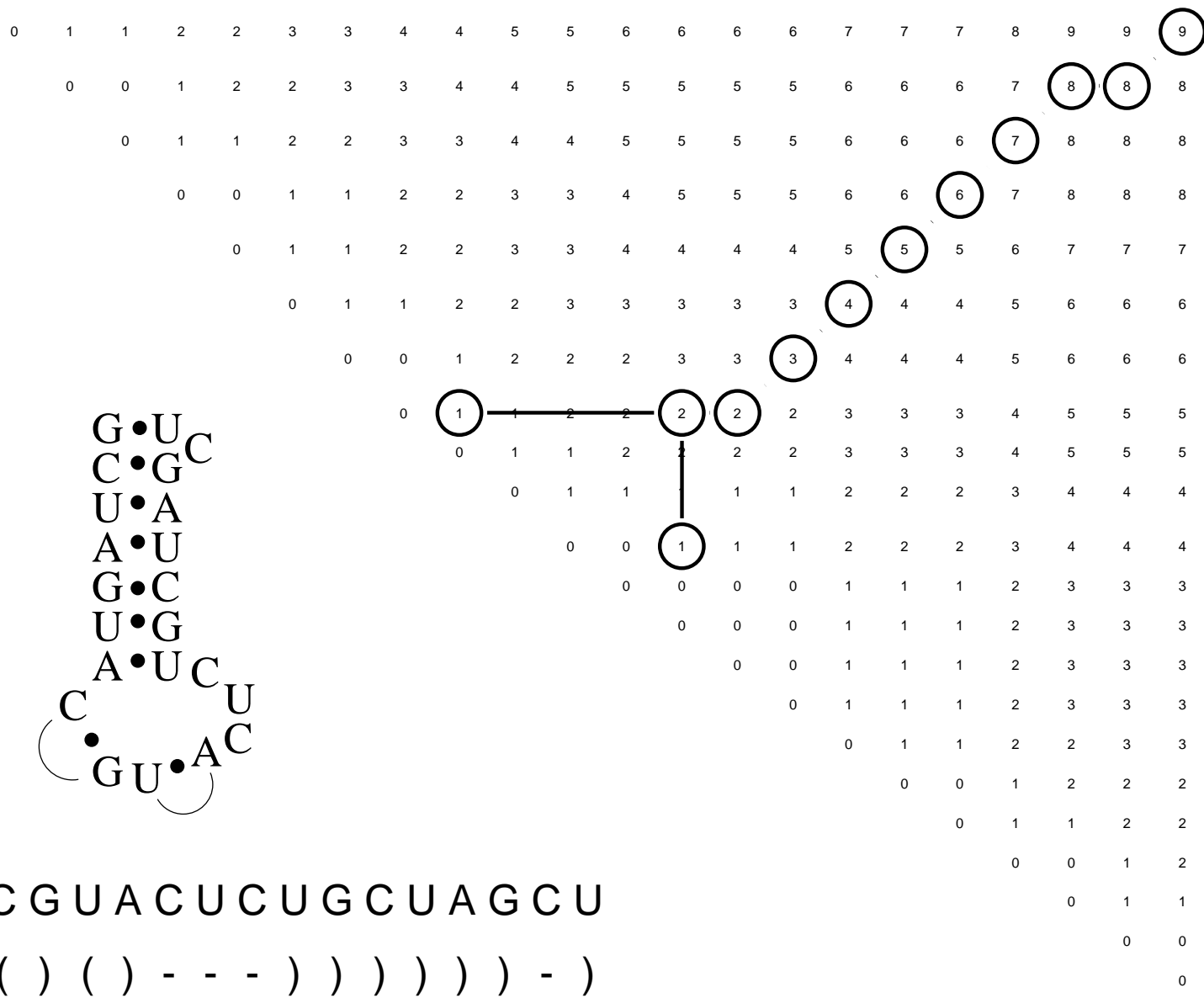
$$s[i..j] = \max \begin{cases} s[i..j-1] \\ s[i..k-1] + s[k+1..j-1] + 1 \end{cases}$$



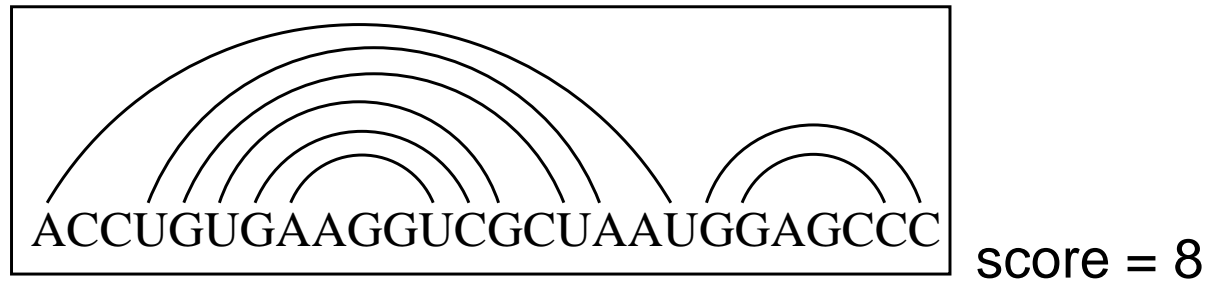
$(i \leq k < j)$



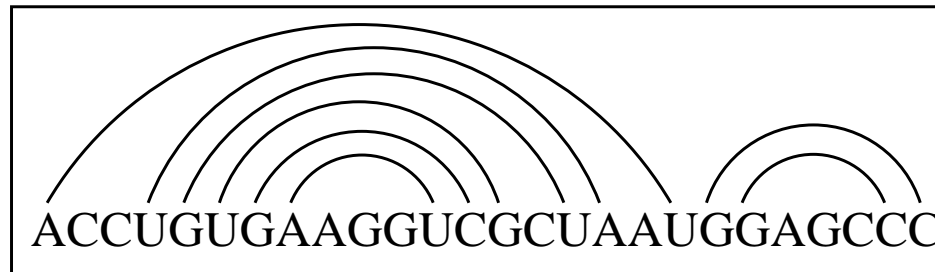




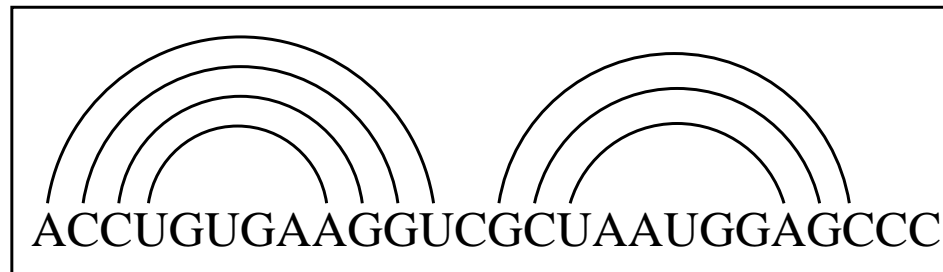
The model of Nussinov is too simple



The model of Nussinov is too simple



score = 8



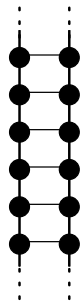
score = 7

Thermodynamic model

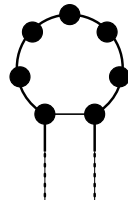
- takes into account stabilizing / destabilizing elements
- MFOLD [*Zuker 1986*]+ Vienna package [*Hofacker 1994*]
- Turner energetic parameters (Stacking energy model)
[*Salser 1977*]+ [*Turner 1986*]+ [*Mathews 1999*]

Thermodynamic model

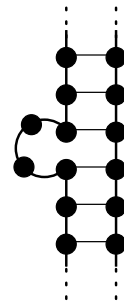
- takes into account stabilizing / destabilizing elements
- MFOLD [*Zuker 1986*]+ Vienna package [*Hofacker 1994*]
- Turner energetic parameters (Stacking energy model)
[*Salser 1977*]+ [*Turner 1986*]+ [*Mathews 1999*]



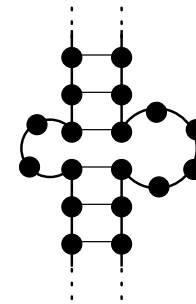
stem



hairpin



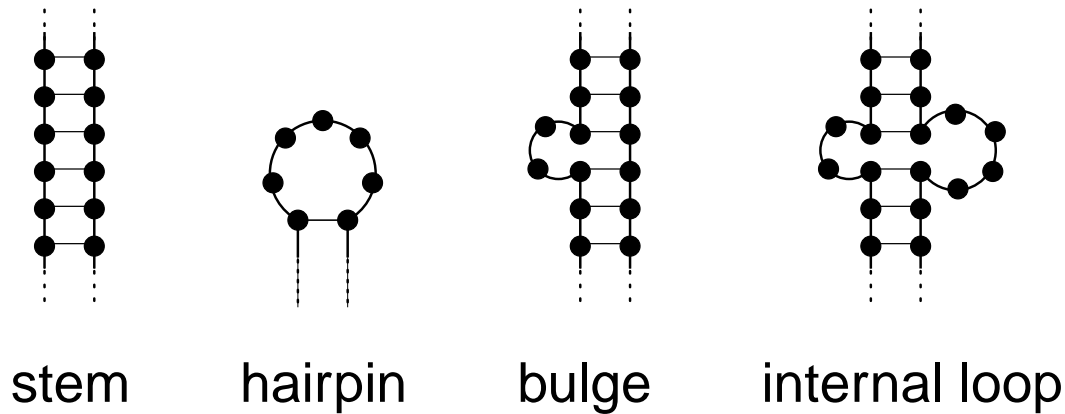
bulge



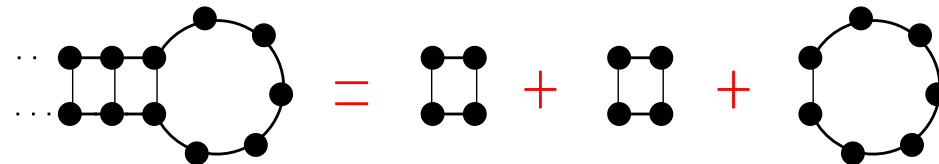
internal loop

Thermodynamic model

- takes into account stabilizing / destabilizing elements
- MFOLD [*Zuker 1986*]+ Vienna package [*Hofacker 1994*]
- Turner energetic parameters (Stacking energy model)
 [*Salser 1977*]+ [*Turner 1986*]+ [*Mathews 1999*]



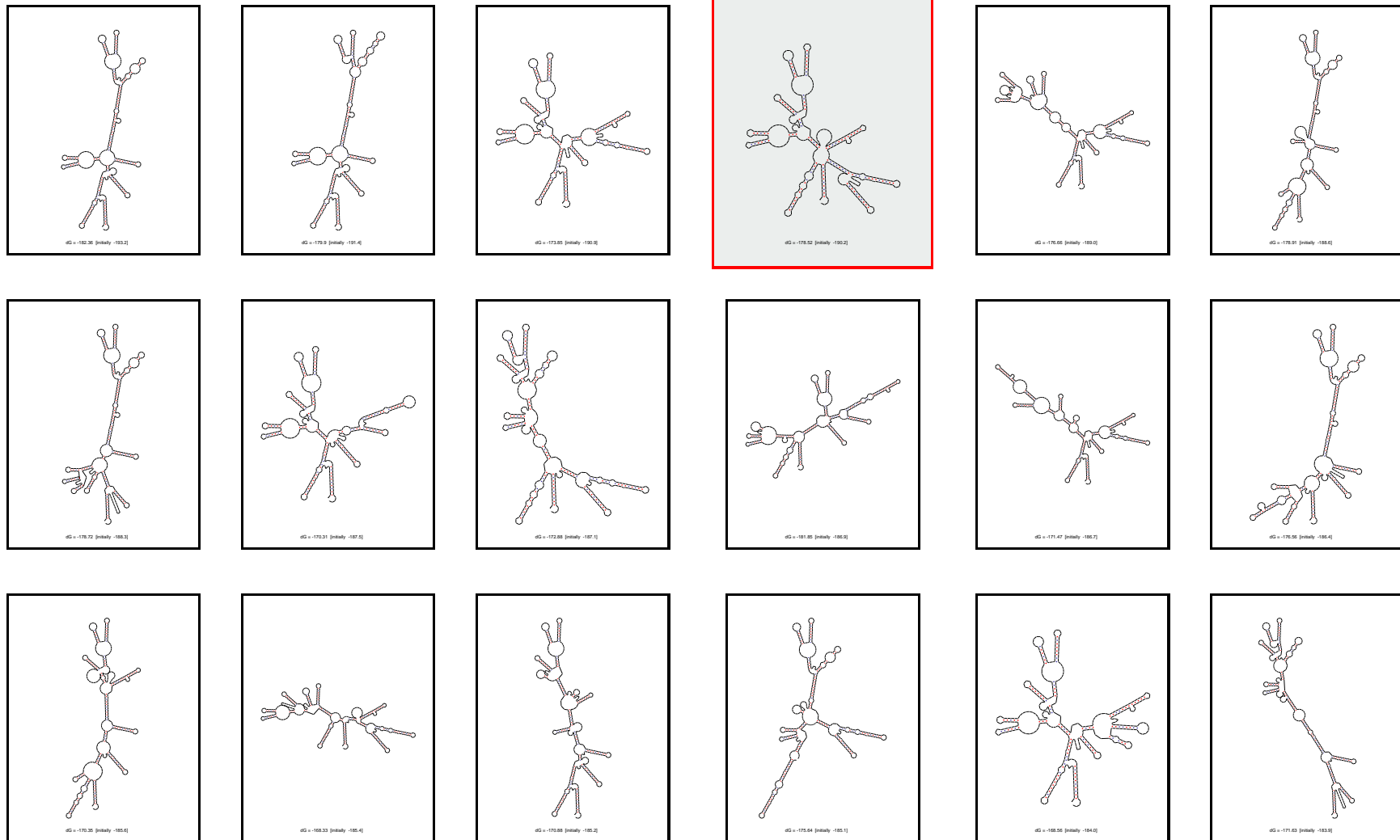
additive
model



Structure prediction with a single sequence

- Secondary structure
 - [*Nussinov 1978*]– dynamic programming recursions – $O(n^2)/O(n^3)$
maximizing the (weighed) number of base pairings A=U, C=G, G=U
 - [*Zuker 1986*]– **MFOLD** – dynamic programming – $O(n^2)/O(n^3)$
minimizing the free energy (stake vs. interval) + suboptimal structure
 - [*Mathews 2004*]– most up-to-date energy parameters
- Secondary structure with pseudoknots
 - [*Lyngsø 2000*]– NP complete in general
 - [*Rivas 1999*]– dynamic prog. (restricted class of pkn) – $O(n^4)/O(n^6)$

MFOLD: Suboptimal structures of *D.desulfuricans* RNase P RNA (5%)



context 2 – a large family of homologs

(from several tens to several thousands)

1. align the sequences
2. find the common structure by covariation analysis

Alignment of tRNA + secondary structure

```

<<<<<<  [[[[          ]]]] [[[[[          ]]]]]  [[[[[          ]]]]]>>>>>>
GUCCGAAuaGCUCagcuggauaGAGCaAUAGCcuucuaaGCUAUcggucGGGGGuucgaaUCCUCUUCGGACgcca
GCGCCCAuaGAUCaauuggauaGAUCgUUUGAcuacggaUCAAaagguuGAGGGuucgauuCCUUCUGGGCGCgcca
GCACUCGuaGCUUaac-ggauaAAGCaUCUGAcuacggaUCAGAagguuGCAGGuucgaaUCCUGCCGAGUGCa---
GUCCACGuaGCUCagcaggauaGAGCaCAGGAuuccuaaUCCUGggguuGGAGGuucgaaUCCUCUCGUGGACacca
GCGCCCGuaGCUCaauuggauaGAGCgUUUGAcuacggaUCAAGagguuAUGGGuucgacuCCUAUCGGGCGCg---
GCACCCAuaGCGCaacuggauaGAGUgUCUGAcuacgaaUCAGAagguuGUAGGuucgaguCCUACUGGGUGCa---
GCGCCCGuaGCUCaauuggauaGAGCgUUUGAcuacggaUCAAaagguuAGGGGuucgacuCCUCUCGGGCGCgcca
GCGCCCGuaGCUCagcuggauaGAGCgCUGCCcuccggaGGCAGaggucUCAGGuucgaaUCCUGUCGGGCGCgcca
GCGCCCUuaGCUCaguuggauaGAGCaACGACcuucuaaGUCGUgggccGCAGGuucgaaUCCUGCAGGGCGCgcca
    
```


Large families of homologs: Align then Fold

- [*Gutell 1992*]- Mutual Information (M.I.) + greedy incorporation
- [*Eddy + Durbin 1994*]- **COVE** – evolutionary SCFG – $O(pn^2)/O(p^3n^3)$
- [*Knudsen 1999*]- **PFOLD** – SCFG + phylogenetic tree – $O(pn^2)/O(p^3n^3)$
- [*Lindgreen 2006*]- deficiencies of the M.I. model

$$S \longrightarrow Sa \mid Su \mid Sc \mid Sg \qquad \text{(free base)}$$

$$S \longrightarrow aSu \mid uSa \mid cSg \mid gSc \mid uSg \mid gSu \qquad \text{(pairing)}$$

$$S \longrightarrow SS \qquad \text{(branching)}$$

$$S \longrightarrow \varepsilon$$

Align then Fold: SCFG

context 3 – only a few homologs

(a few means at least 2 ...)

With two homologs: fold then align

> *D.desulfuricans* RNase P RNA

```

1  GGAGUCGGAC  GGAUCGUCGC  CGCGGGGGCA  ACUCCGGGGA  GGAAAGUCCG
51  GGCUCCAAAG  GGCAGAACGC  UGGAUAAACAU  CCAGGGAGGG  CAACCUCGGG
101 ACAGCGCCAC  AGAAAGCAA  CCGCCCGGCC  UCGGCCGGGU  AAGGGUGAAA
151 CGGUGGUGUA  AGAGACCACC  AGAUGCCGUG  GUGACACGGC  AUGCUCGGCA
201 UACCCCGUUC  GGAGCAAGAC  CAAAUAGGGA  AGGCGGCCGG  CCCGGCCGAA
251 GCCUUCGGG  UAGGUUGCUU  GAGGGUGUGG  GCAACCGCAC  UCCUAGAGGA
301 AUGACGGUCA  CACGCGGGCA  ACCGUGUGGA  CAGAACCCGG  CUUACAGUCC
351 GACUCCCGCA

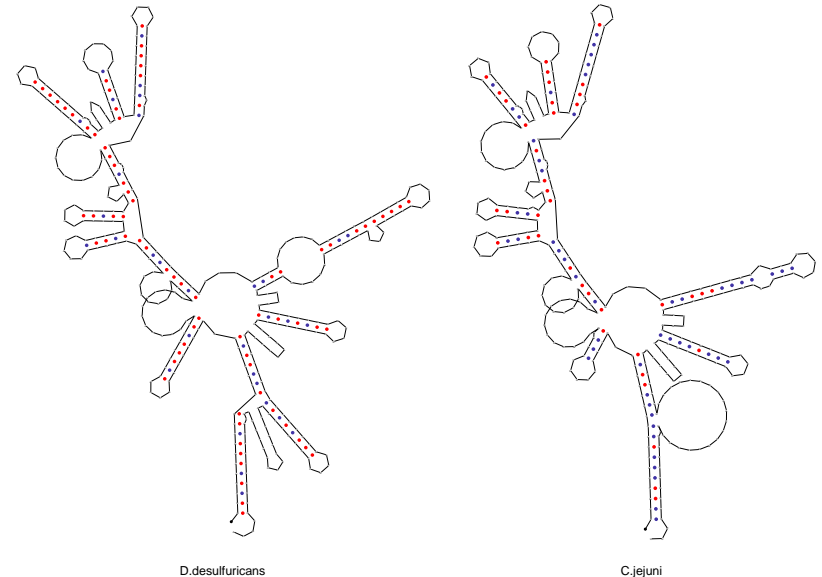
```

> *C.jejuni* RNase P RNA

```

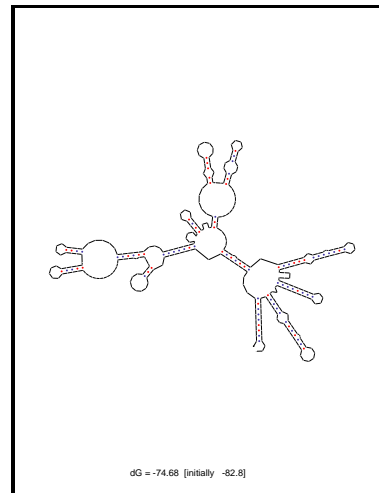
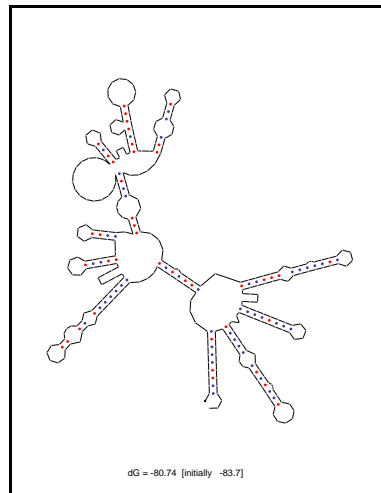
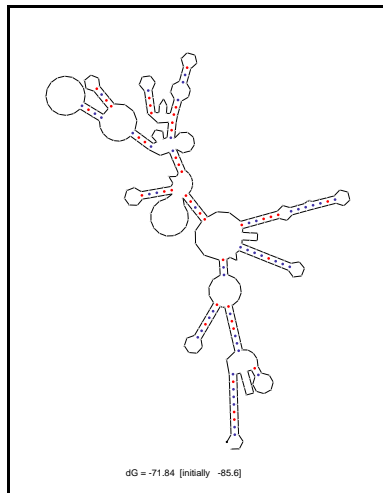
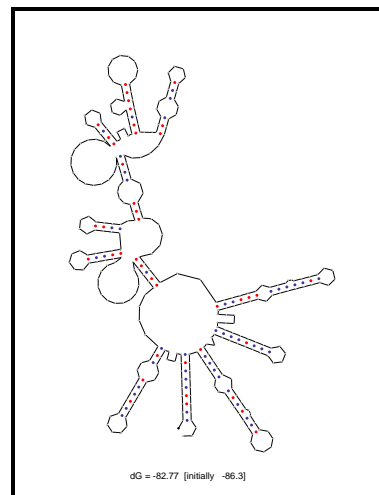
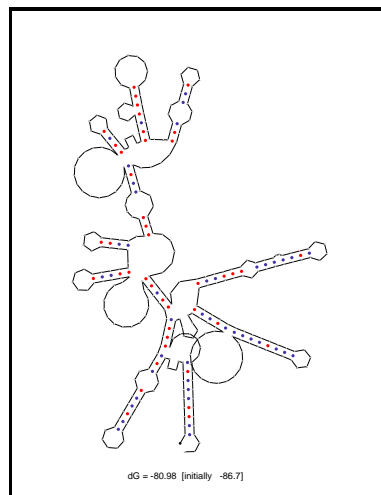
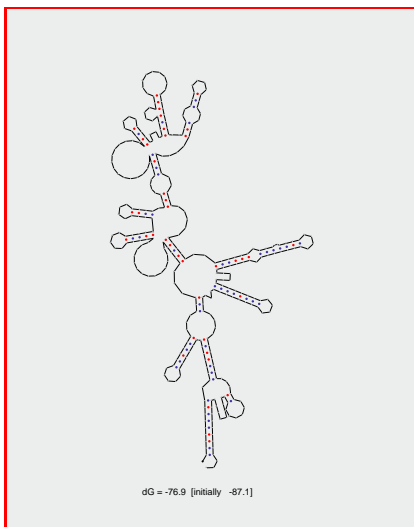
1  AAGCAUAGUA  AAUGCUCGCU  UCUUUUUAGG  AGAGGAAAGU  CCGAGCUGCU
51  AAAGACAAAC  AUUCCAUCUA  ACAGAUGGCU  AGGGUAACCU  AAGGGAUAGU
101 GCAACAGAAA  GAAAACUACC  ACGCAAAGUG  AAAAGGUGAA  ACGGCGGGGU
151 AAAAGCCCAC  CAGCGAUUUU  GGUAACAAU  UCGGCUAUGU  AAACCCAAUG
201 UGCAGCAAGA  AGGGAUGGUU  AGCGUCUUUG  UUUUAACCCU  UCGCUUGAUU
251 UUGUUUGCAA  AAACAAAACU  AGAUAAAUGA  GCAUUCAAGA  CAGAACUCGG
301 CUUAUCGCUA  UGCUUUUU

```

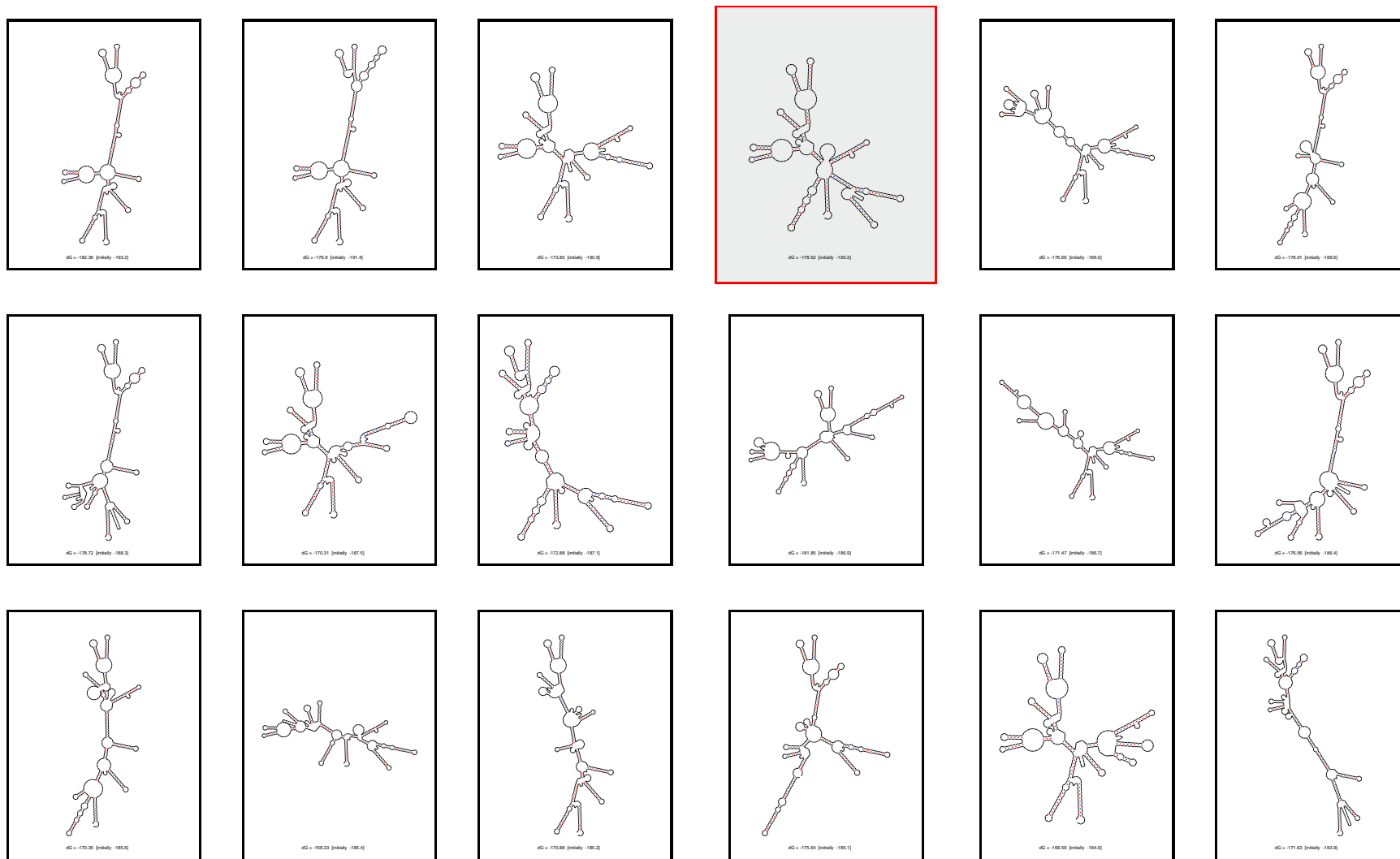


D.desulfuricans vs *C.jejuni* RNases P

C.jejuni : suboptimal structures given by MFOLD

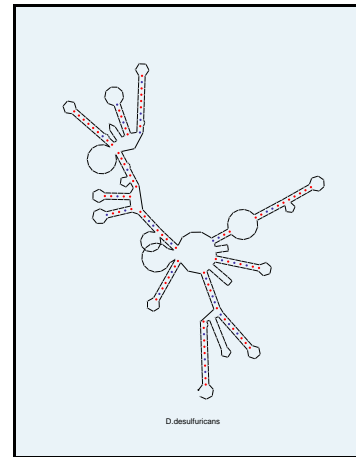


D.desulfuricans : suboptimal structures given by MFOLD

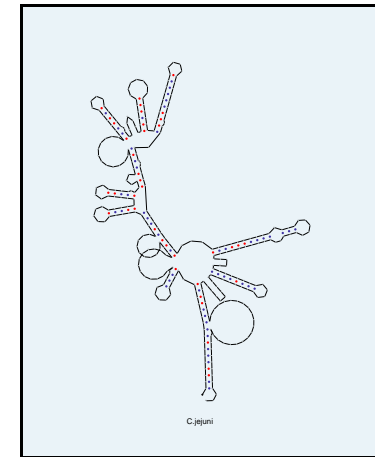


The known structures
D.desulfuricans & *C.jejuni*

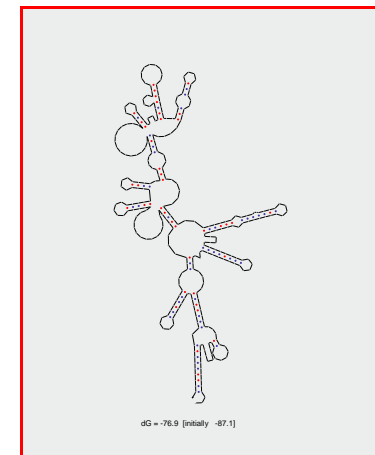
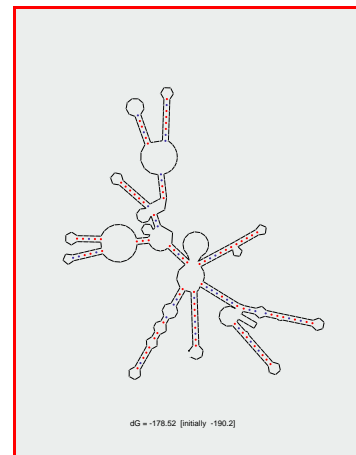
D.desulfuricans



C.jejuni



The best suboptimal structures
given by MFOLD have
69% et 65% base pairs in common
with the true structure



- the two optimal structures given by MFOLD are not similar to the known one
- these two optimal structures are not similar to each other

With two homologs: align then fold

global alignment (classic Needleman & Wunsch)

GGAGUCGGACGGAUCGUCGCCGCGGGGGCAACUCC.....GGGGAGGAAAGUCCGGGCUCCAAAGGGCAGAACGCUGGAUAA
AAGCAUAGU.....AAAUGCUCGCUUCUUUUUAGGAGAGGAAAGUCCGAGCUGCUGAAAGACA.AAC.....AUUC

CAUC...CAG..GG..AGGGCAACCU.CCGGACAGCGCCACAGAAAGCAAAC.....CGCCCGGCCUCGGCCGGGUAAGGGUGAAACGGU
CAUCUAACAGAUGGCUAGGGUAACCUAAGGGAUAGUGCAACAGAAAGAAAACUACCACGCAAG.....UGGAAAAGGUGAAACGGC

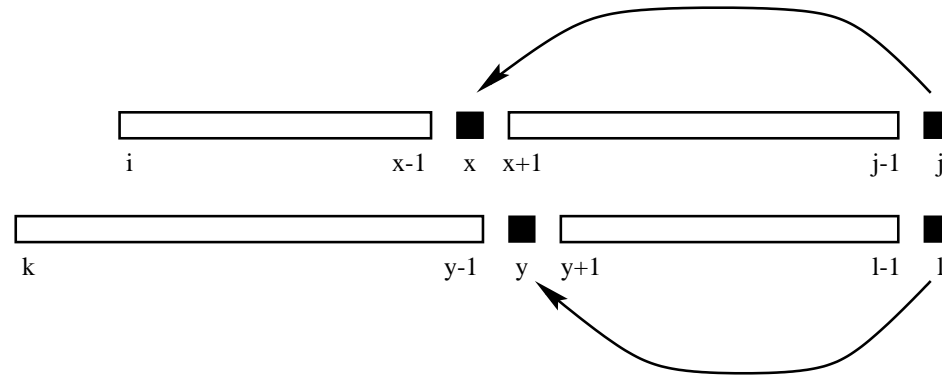
GGUGUAAGAGACCACCAGAUGCCG....UGGUGACA....CGGC.AUGCUCGGCAUACCCCGUUCGGAGCAAGACCAAUAGGGAAGG..
GGGGUAAAAGCCCACCAG....CGAUUUUGGUAACAAUUUCGGCUAU....GUAAACCCAUGUGCAGCAAGA.....AGGGAUGGUU

CGGCCGGCCCGGCCG.....AAGCCUUCGGGUAGGUUGCUUGAGGGUGUGGGCAACCGCACUCCUAGAGGAAUGACGGUCACAC
AG.....CGUCUUUGUUUAACCCUUC.....GCUUGAUUUUGUUUGCAAAAACAAAACUAGAUAAAUGA.....

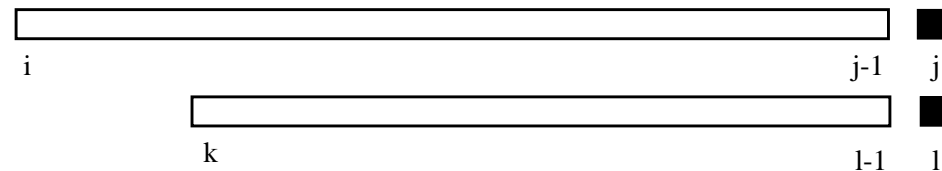
GCGGGCAACCGUGUGGACAGAACCCGGCUUACAGUCCGACUCCC...GCA
GCAUUCAA.....GACAGAACUCGGCUUA.....UCGCUAUGCUUUUU

Align and fold at the same time – $O(n^{2p})/O(n^{3p})$

common pairing



single base



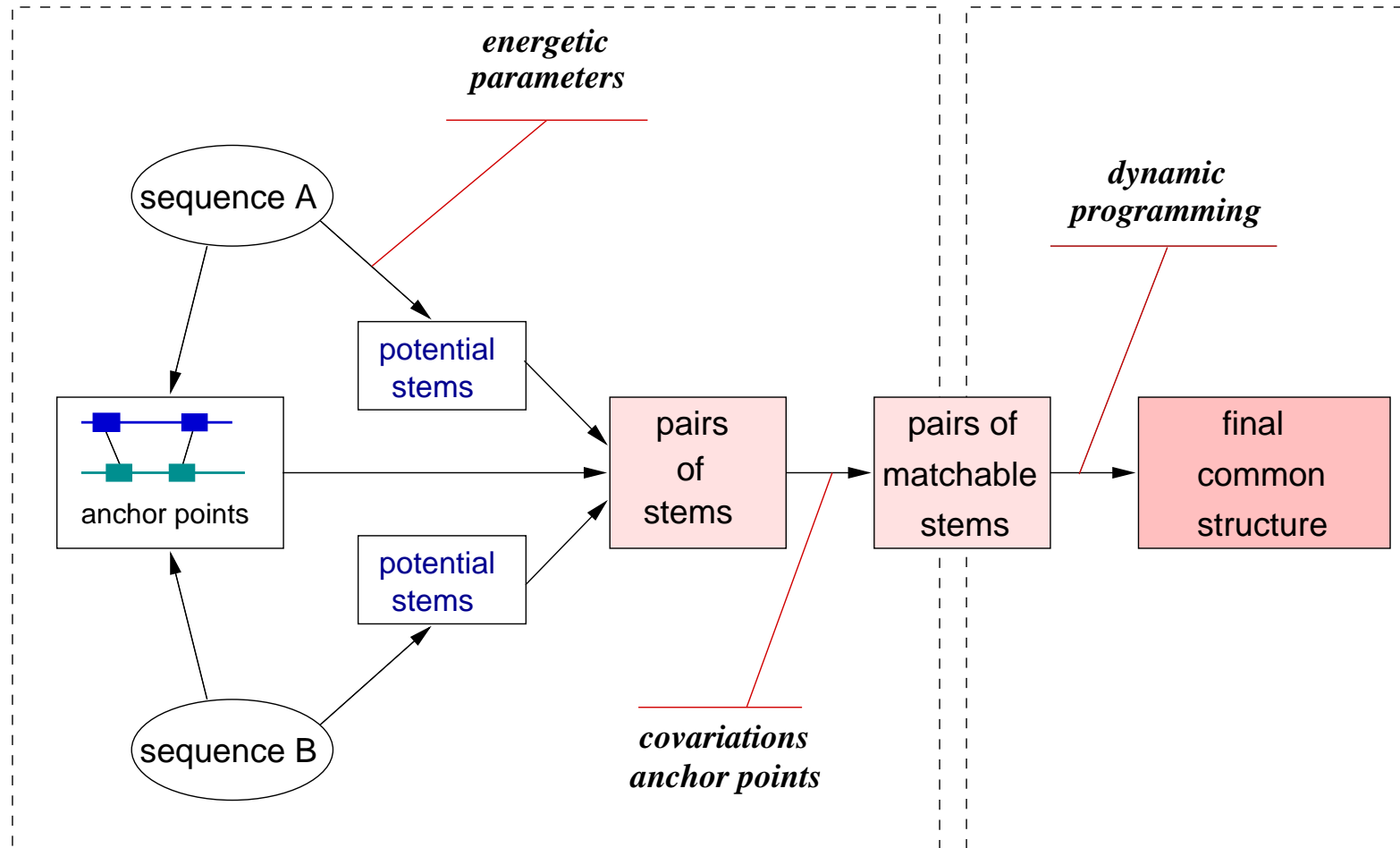
		space	time
1 sequence	[<i>Nussinov</i> 1978]	$O(n^2)$	$O(n^3)$
2 sequences	[<i>Sankoff</i> 1985]	$O(n^4)$	$O(n^6)$
p sequences	[<i>Sankoff</i> 1985]	$O(n^{2p})$	$O(n^{3p})$

Align and Fold at the same time :: 2 sequences

- [*Sankoff 1985*]— dynamic programming recursions – $O(n^4)/O(n^6)$
maximizing the number of common base pairings for 2 sequences
- [*Gorodkin 1997*]— **FOLDALIGN 1** – $O(n^4)/O(n^4)$
LOCAL HEURISTIC: output only the best common stem
- [*Mathews 2002*]— **DYNALIGN** – $O(M^2n^2)/O(M^3n^3)$
BANDING HEURISTIC: max allowed shift M between the sequences
- [*Perriquet 2004*]— **CARNAC** – $O(k^2N^2)/O(k^3N^3)$
HEURISTIC: anchor points on primary sequence (bound k)
HEURISTIC: stem level (N stems instead of n bases)
- [*Havgaard 2007*]— **FOLDALIGN 2** – $O(\lambda\delta n^2)/O(\lambda^2\delta^2n^2)$
SEQUENCE BANDING HEURISTIC: max allowed shift δ between the sequences
STRUCTURAL BANDING HEURISTIC: max size for a common motif bounded by λ
- [*Perriquet 2009*]— **ARNICA** – $O(\alpha^2)/O(\alpha^3)$ with $\alpha \sim \delta n$
LOCAL BANDING HEURISTIC: variable shift along the sequences

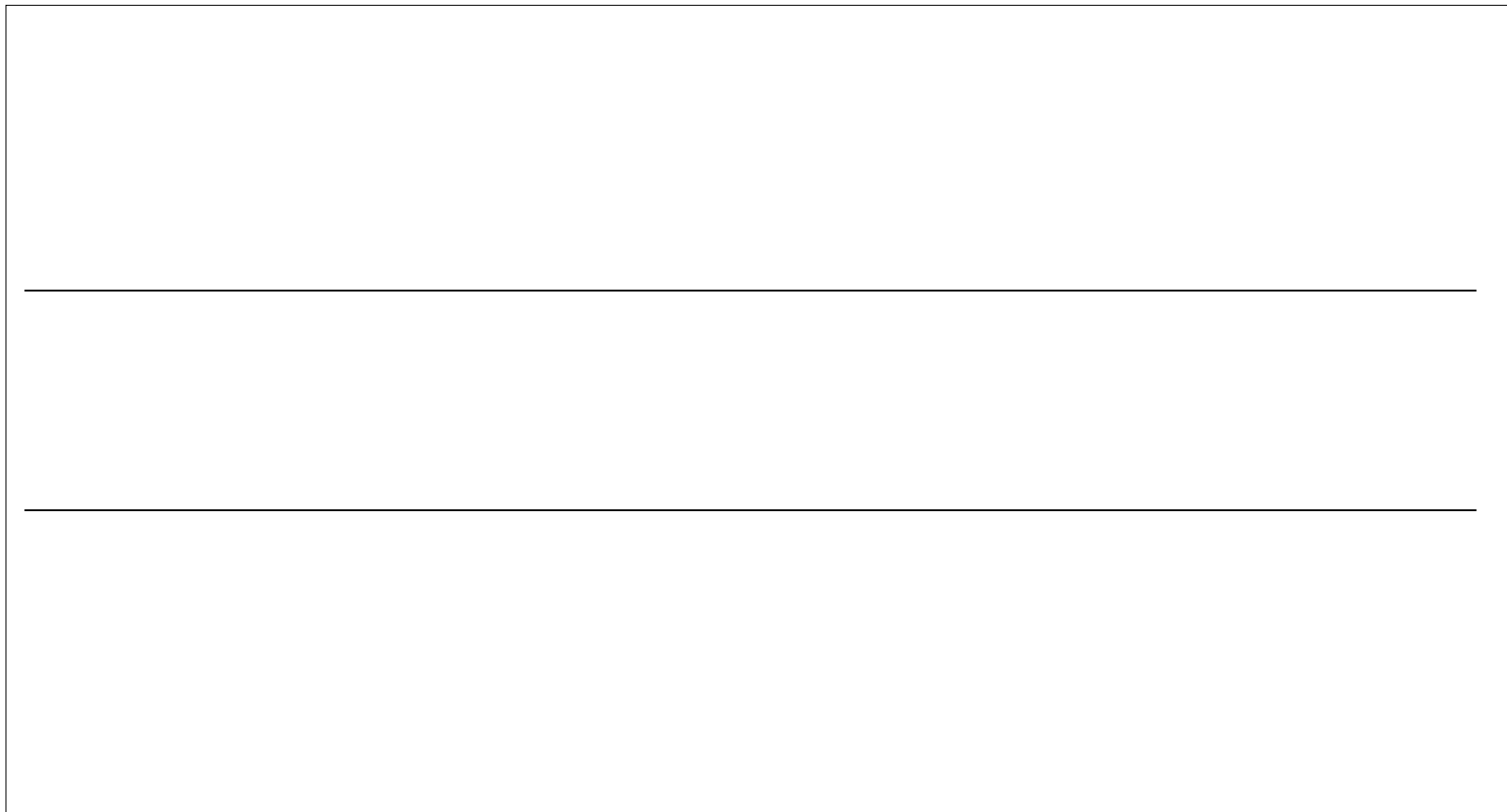
CARNAC – [*Perriquet 2004*]

CARNAC – [*Perriquet 2004*]



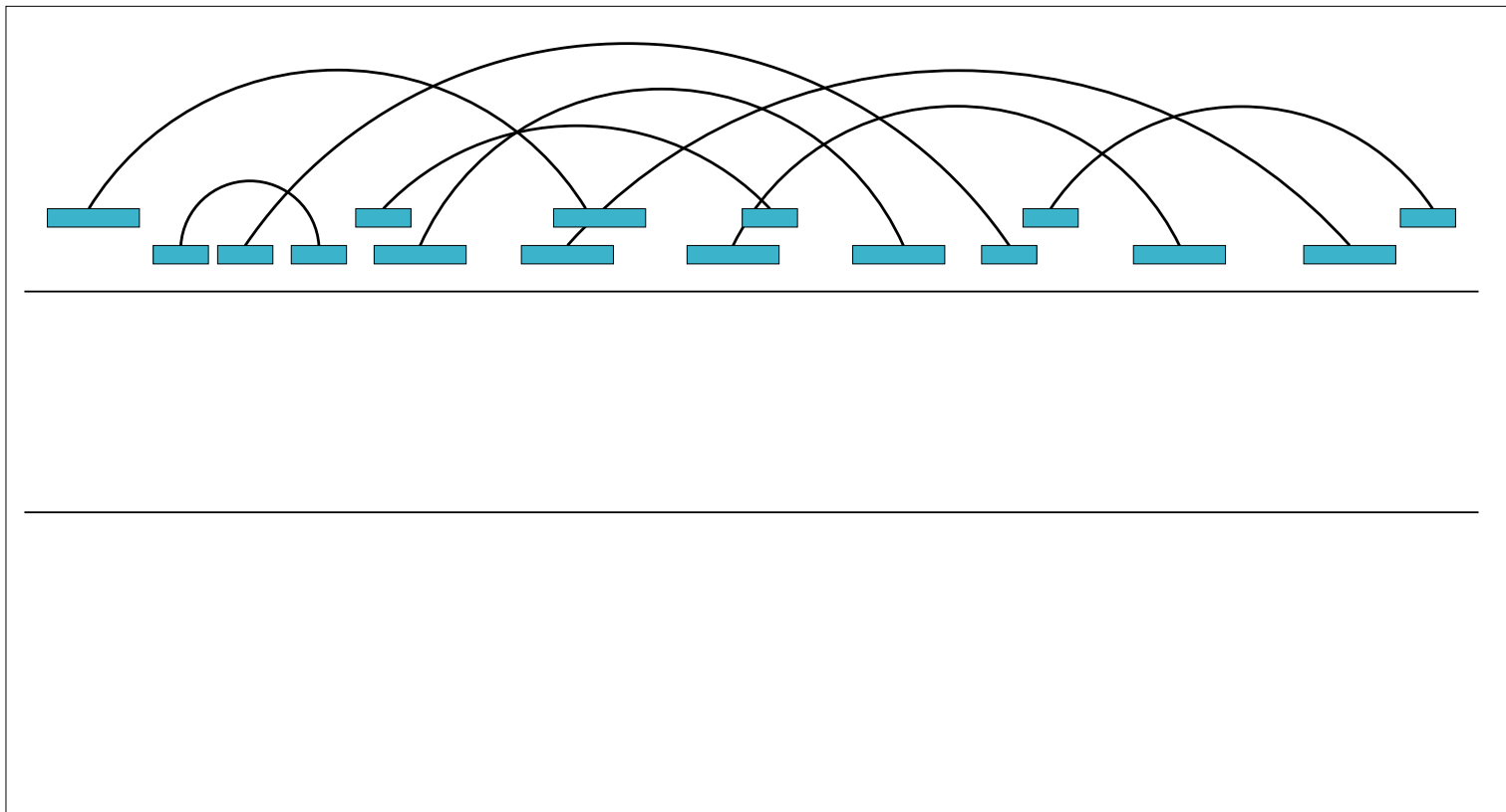
Stage 1. Searching for potential stems

- maximal stems with energy greater than a given threshold
- piecewise affine threshold



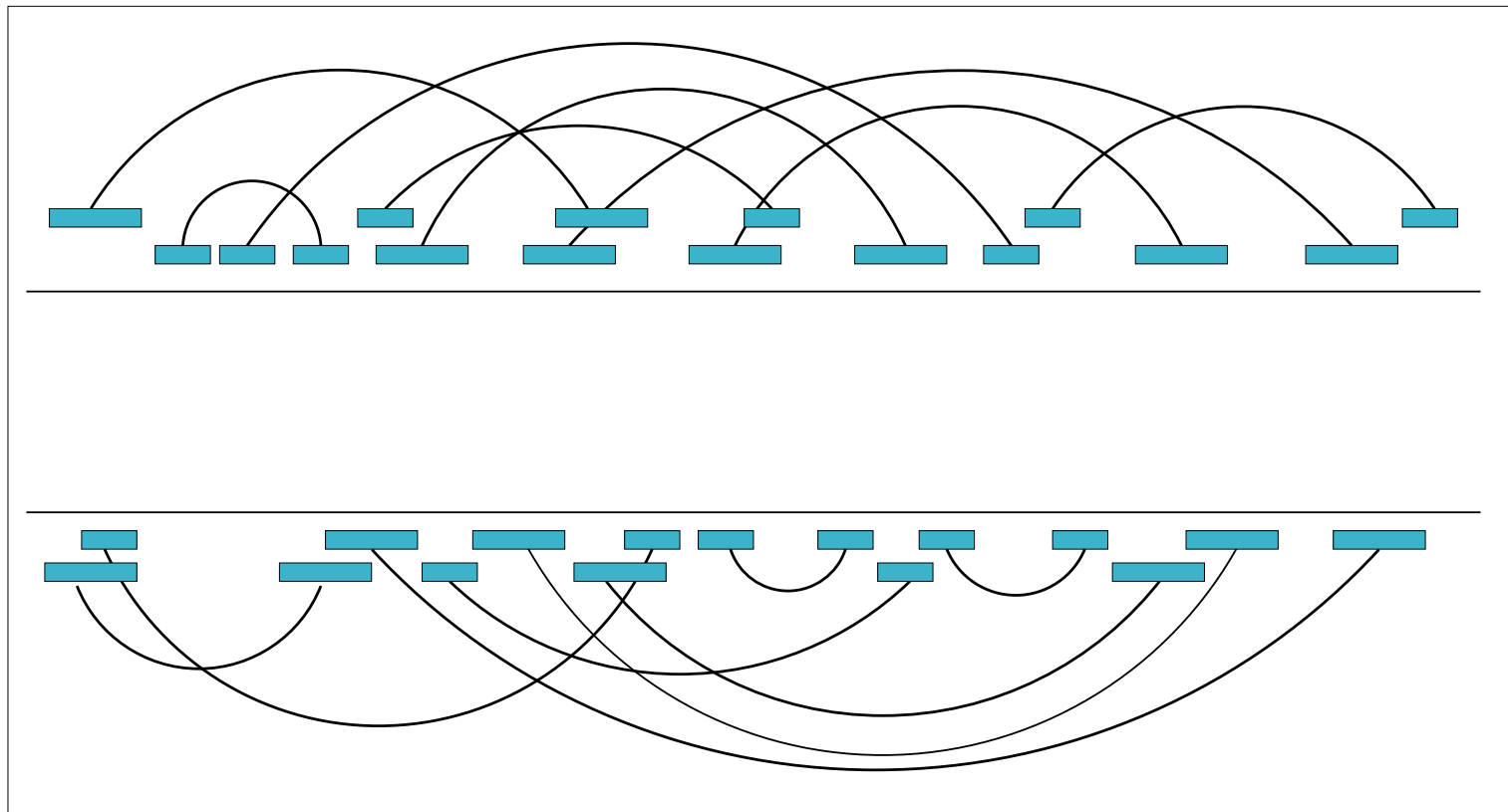
Stage 1. Searching for potential stems

- maximal stems with energy greater than a given threshold
- piecewise affine threshold



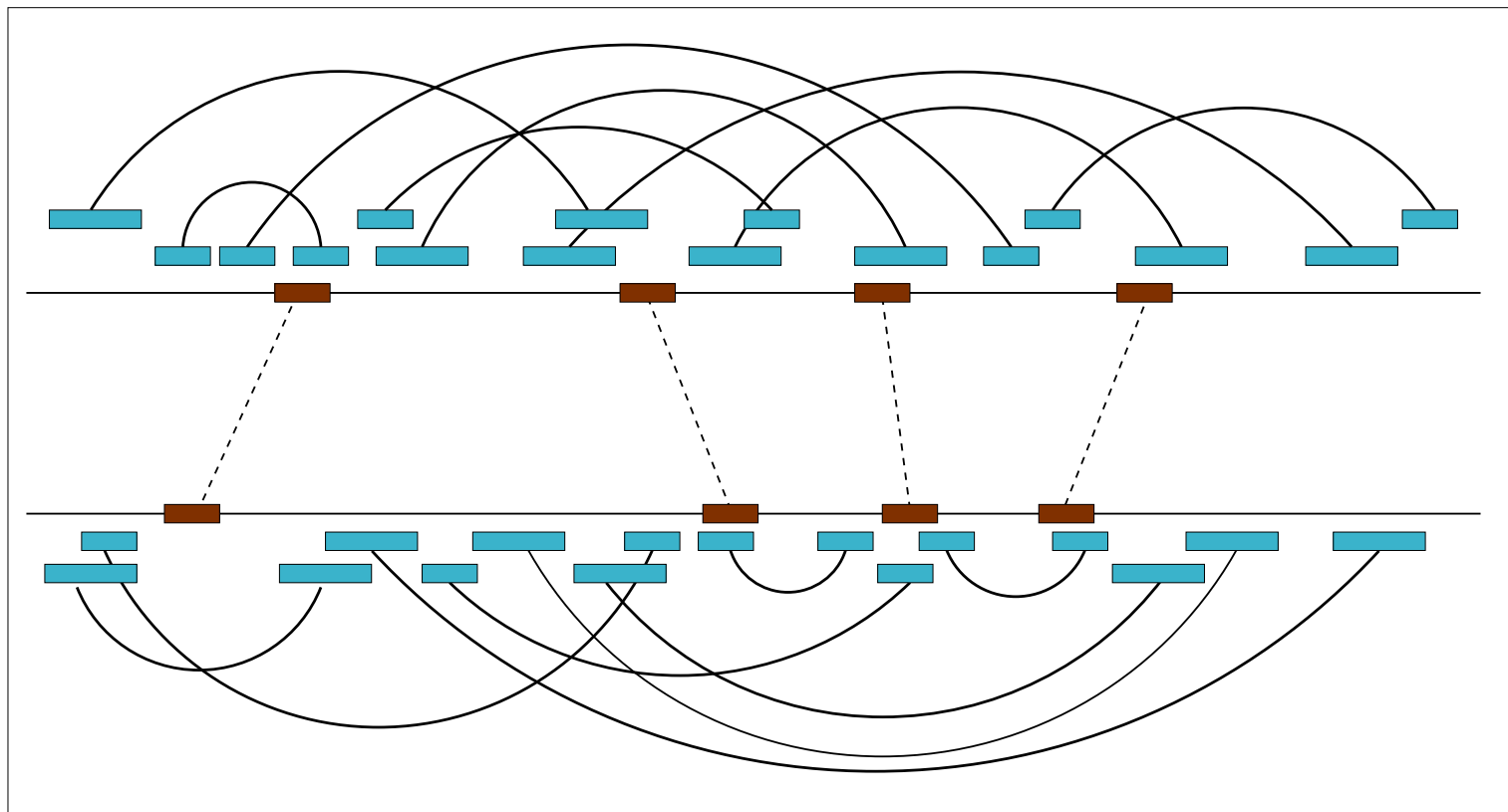
Stage 1. Searching for potential stems

- maximal stems with energy greater than a given threshold
- piecewise affine threshold



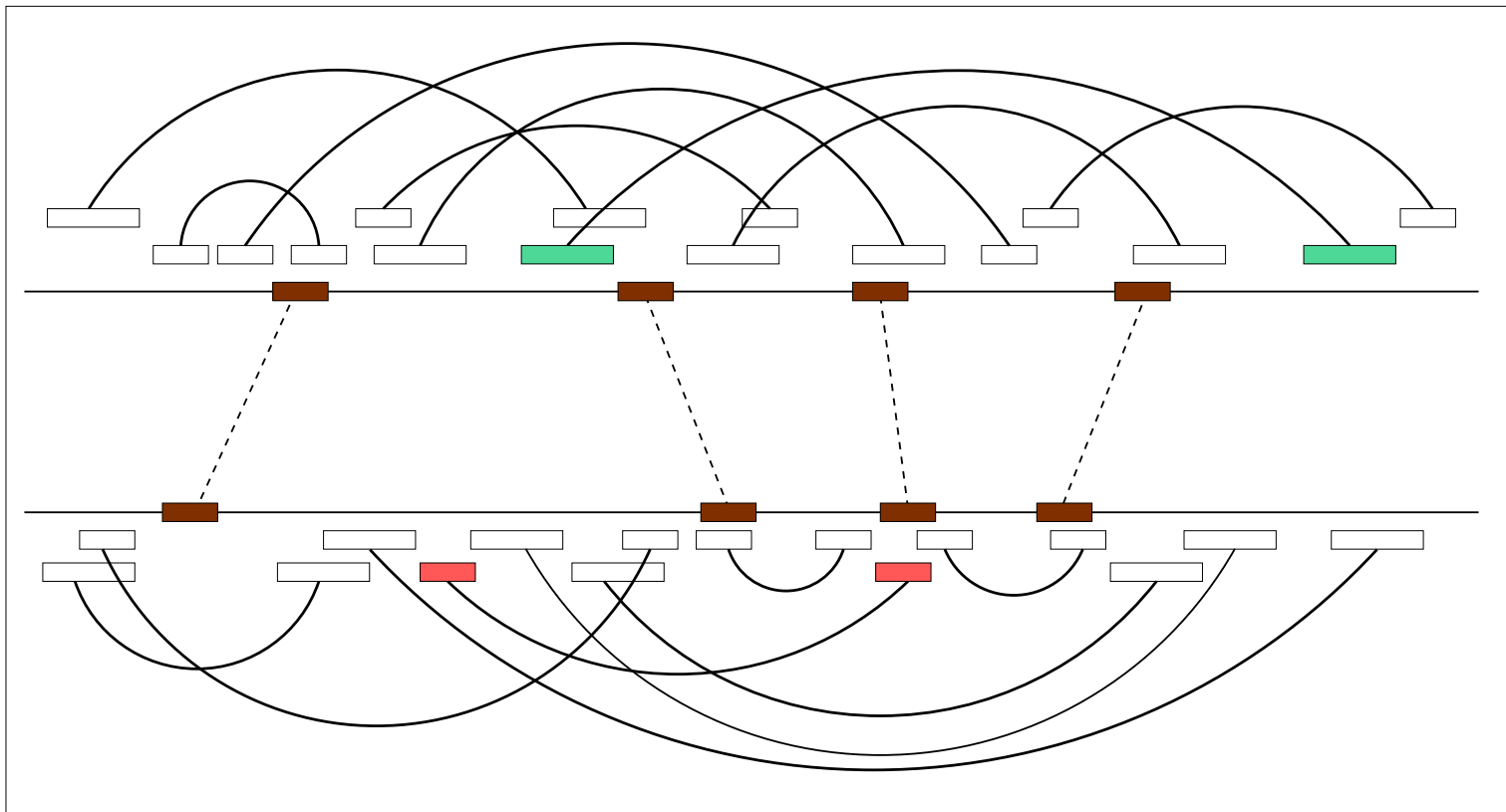
Stage 1. Anchor point detection

- highly conserved common patterns
- no indels



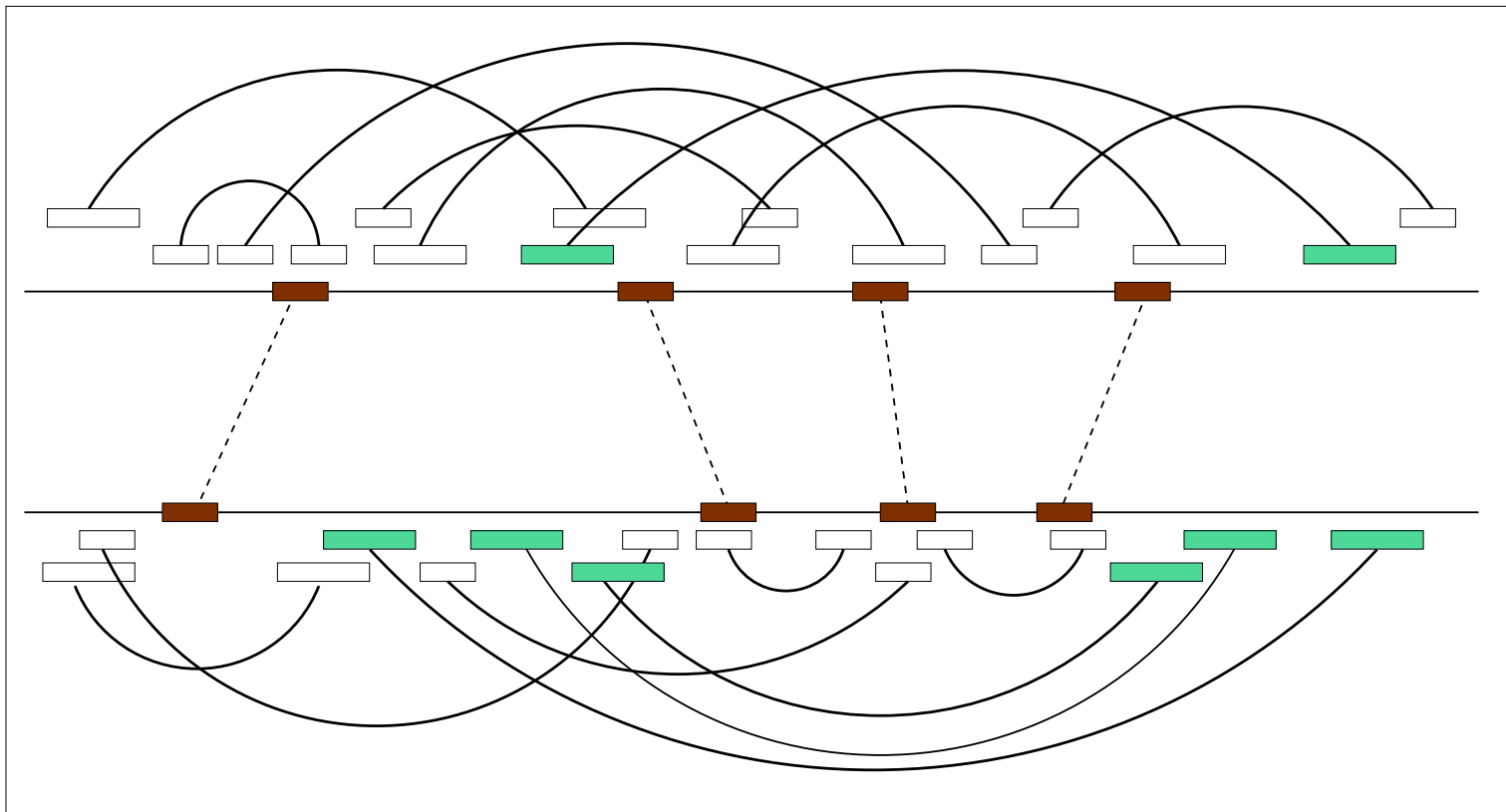
Stage 1. Filtering: matchable stems

- at least one covariation
- compatibility with anchor points



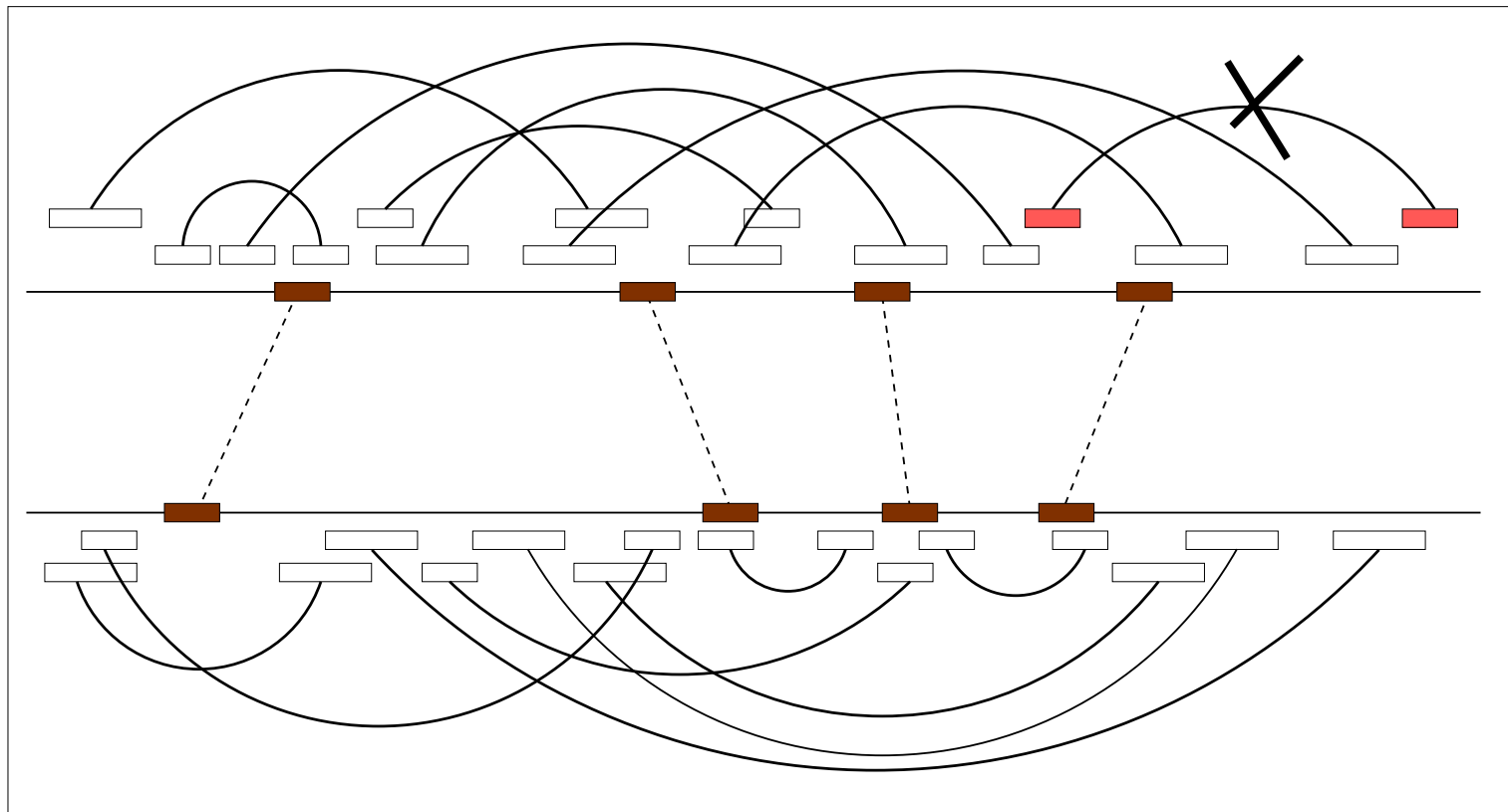
Stage 1. Filtering: matchable stems

- at least one covariation
- compatibility with anchor points



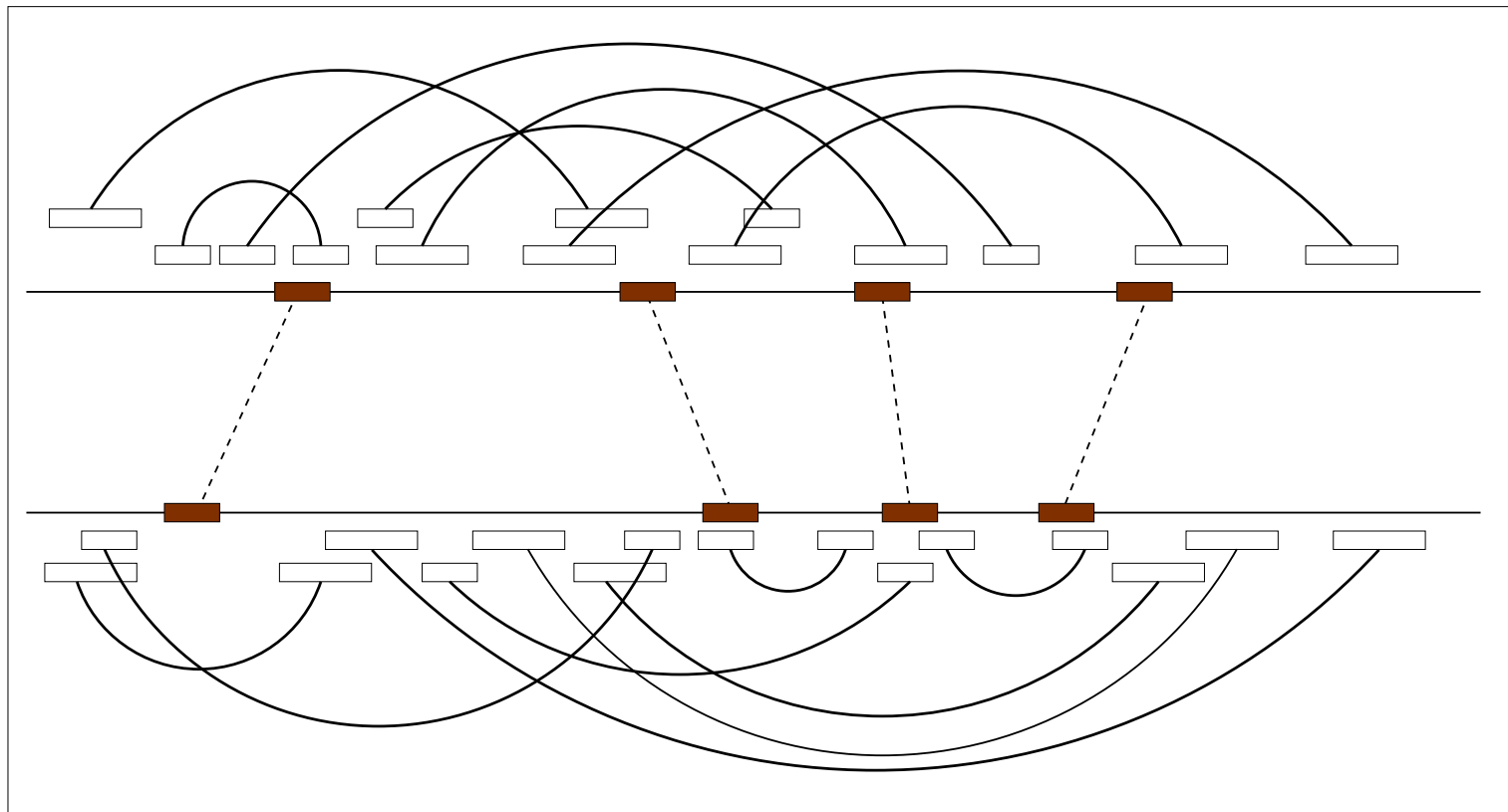
Stage 1. Elimination of single stems

- single: with no potential partner

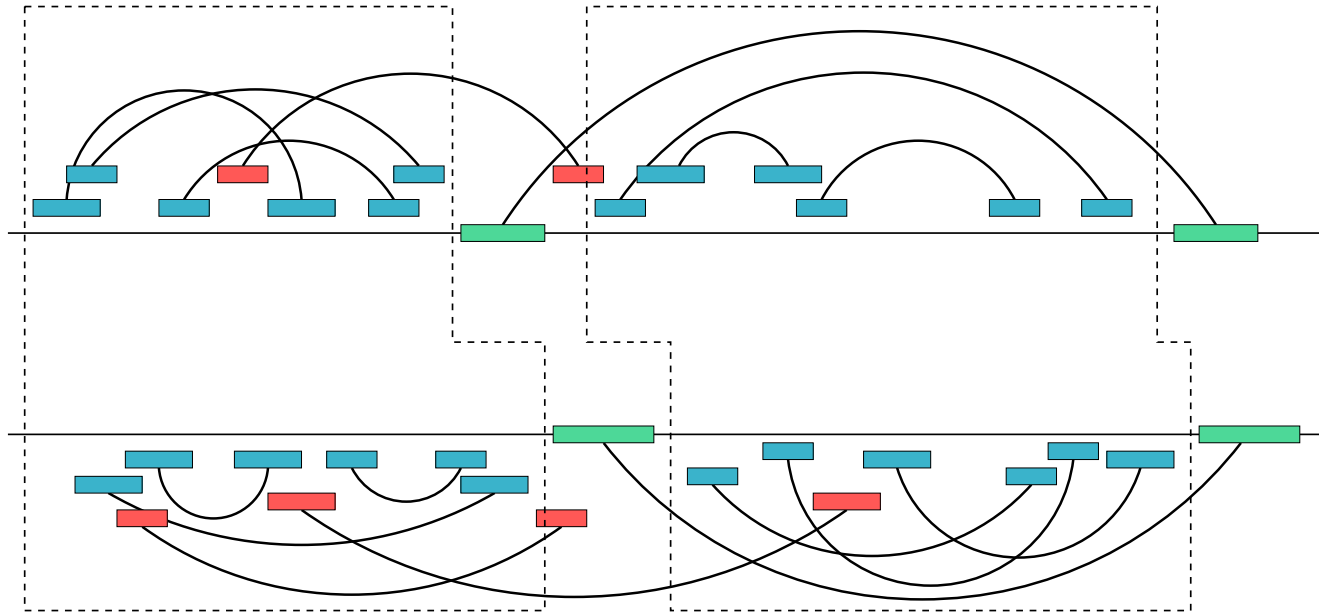


Stage 1. Elimination of single stems

- single: with no potential partner



Stage 2. Finding the common structure adaptation of Sankoff recursions

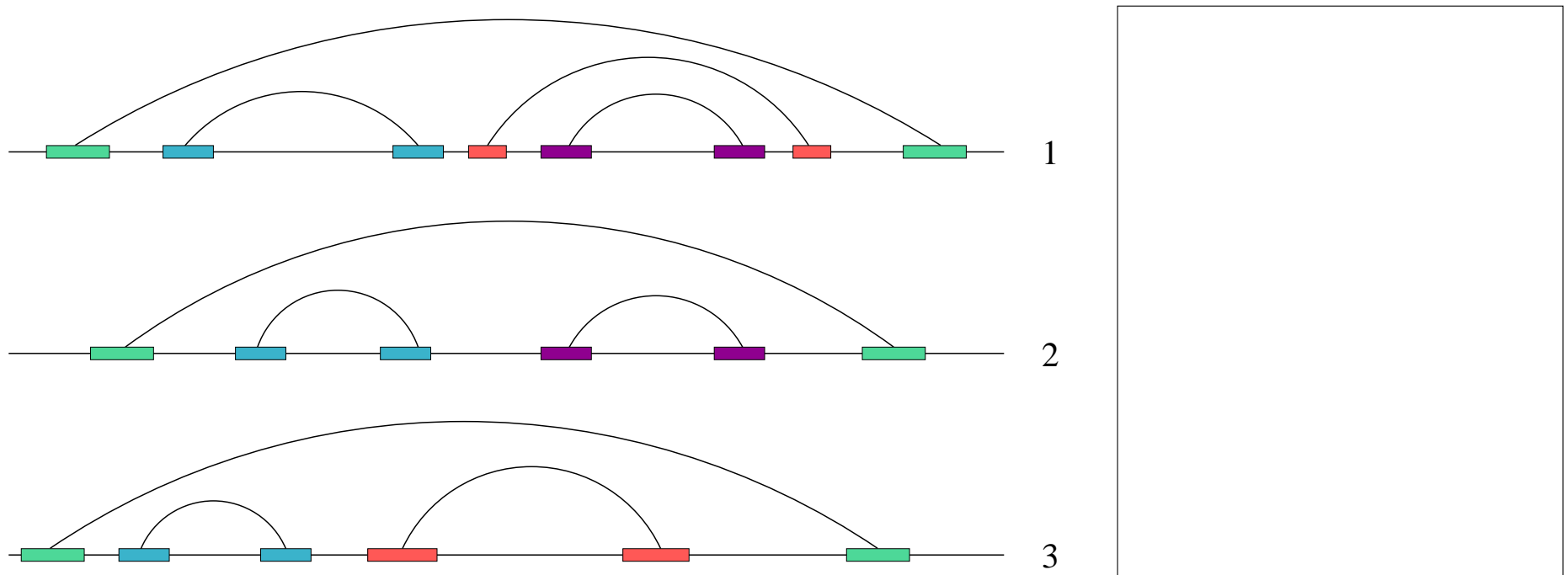


- dynamic programming
- deal with overlapping stems
- memory space optimisation: graph in $O(k^2 N^2)$
(N = number of stems after filtering, k = average nb of matchable stems)

From 2 to p sequences : CARNAC

Stage 0 – Computing of all pairwise foldings with CARNAC₂

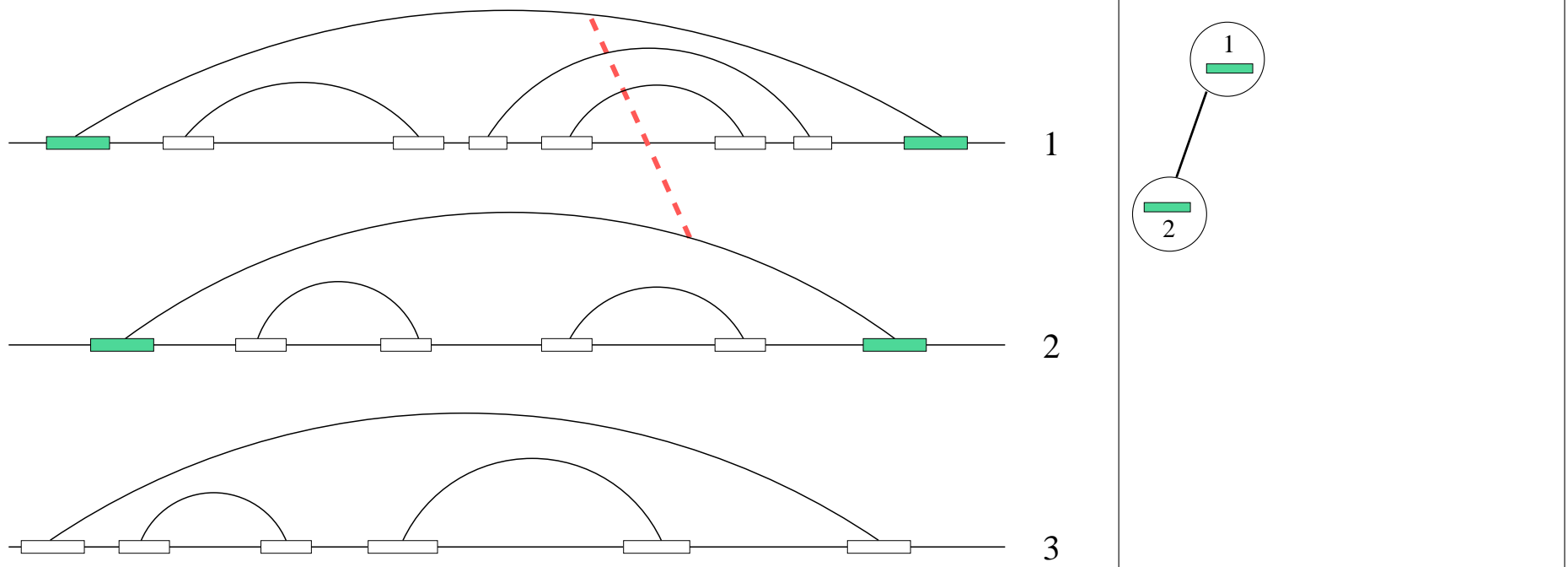
Stage 1 – Building the *graph of stems*



From 2 to p sequences : CARNAC

Stage 0 – Computing of all pairwise foldings with CARNAC₂

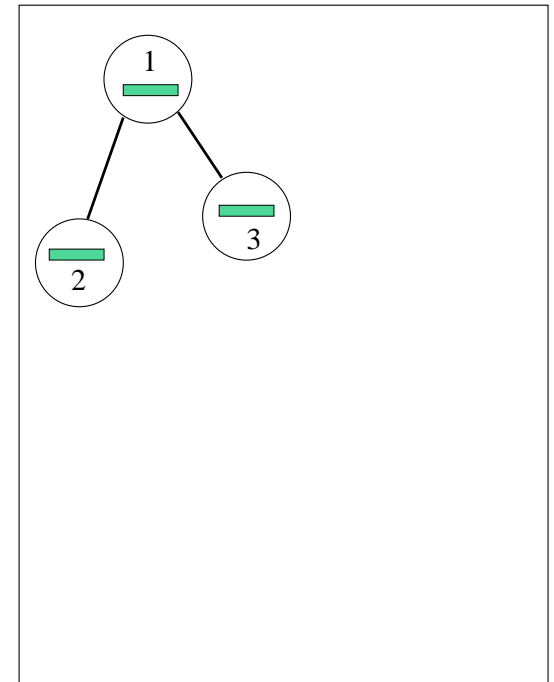
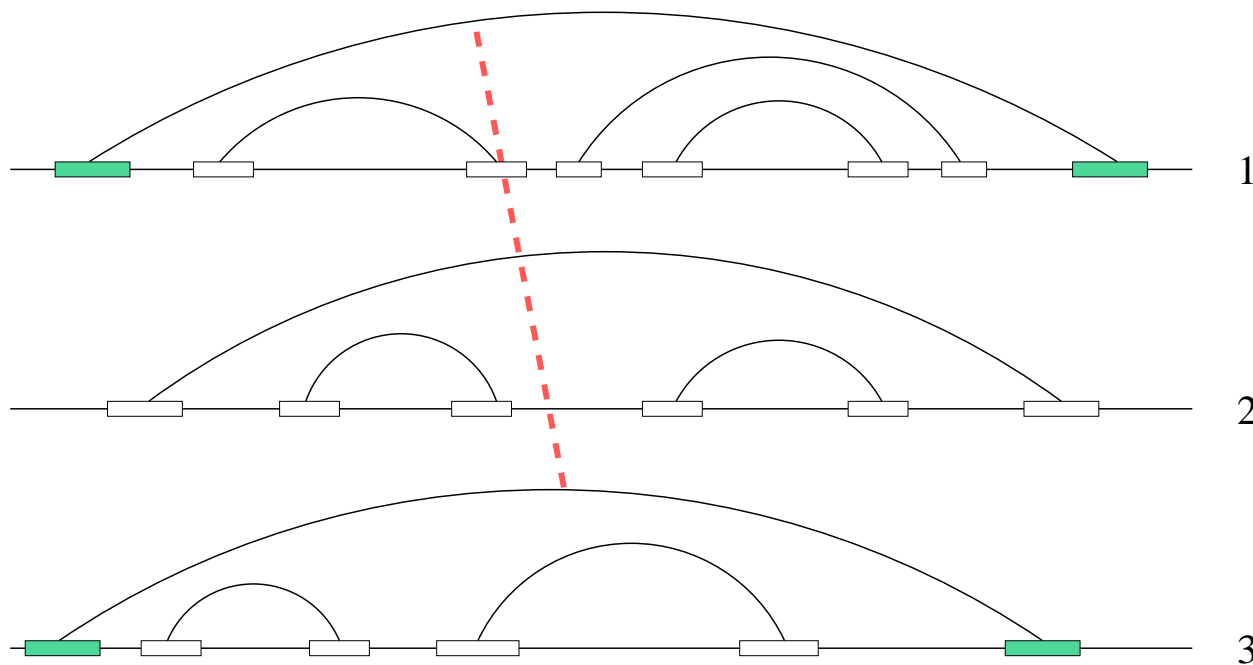
Stage 1 – Building the *graph of stems*



From 2 to p sequences : CARNAC

Stage 0 – Computing of all pairwise foldings with CARNAC₂

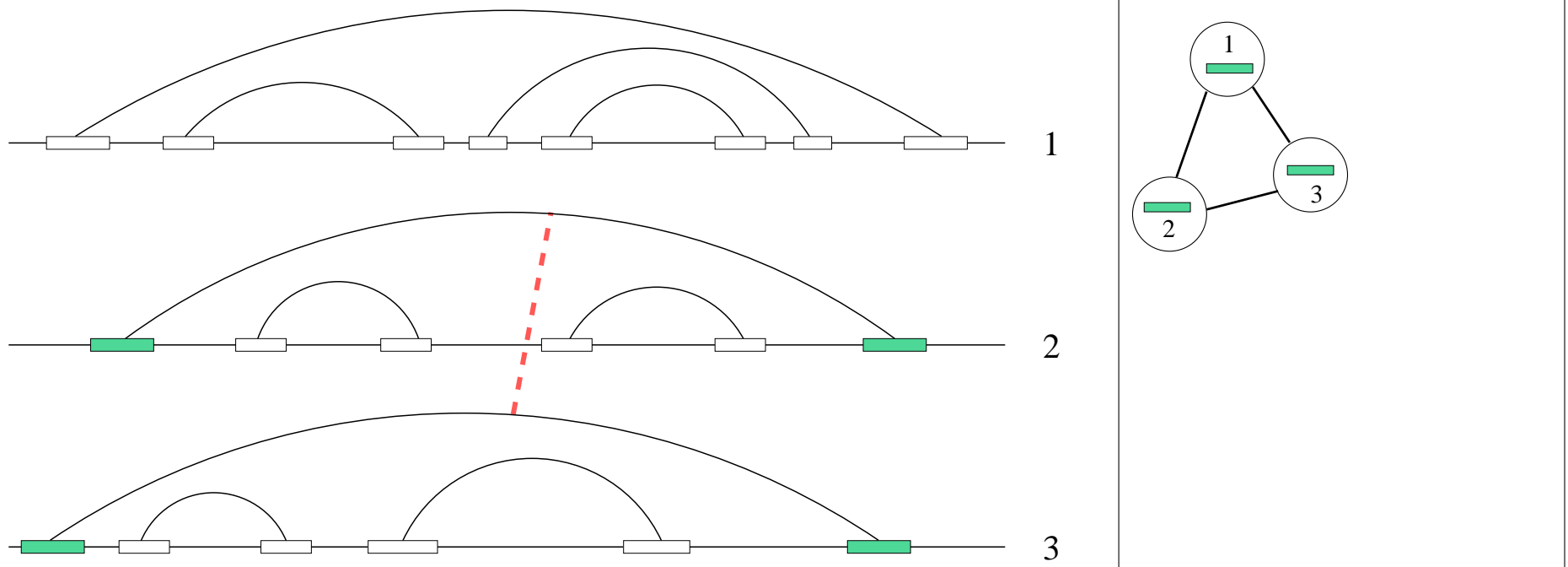
Stage 1 – Building the *graph of stems*



From 2 to p sequences : CARNAC

Stage 0 – Computing of all pairwise foldings with CARNAC₂

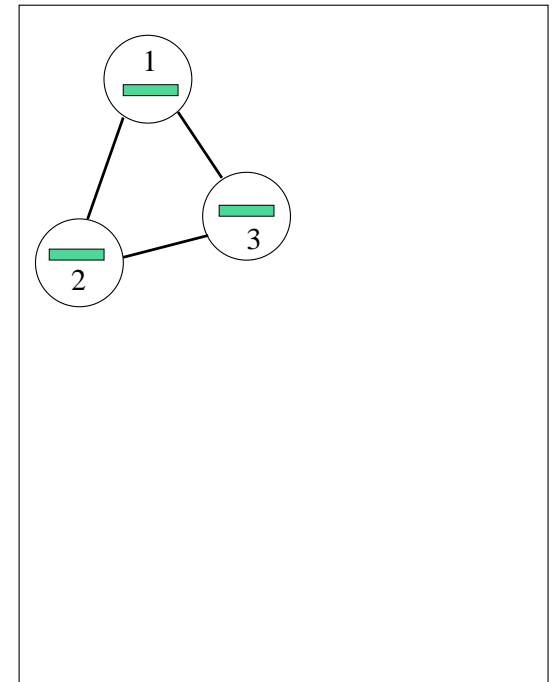
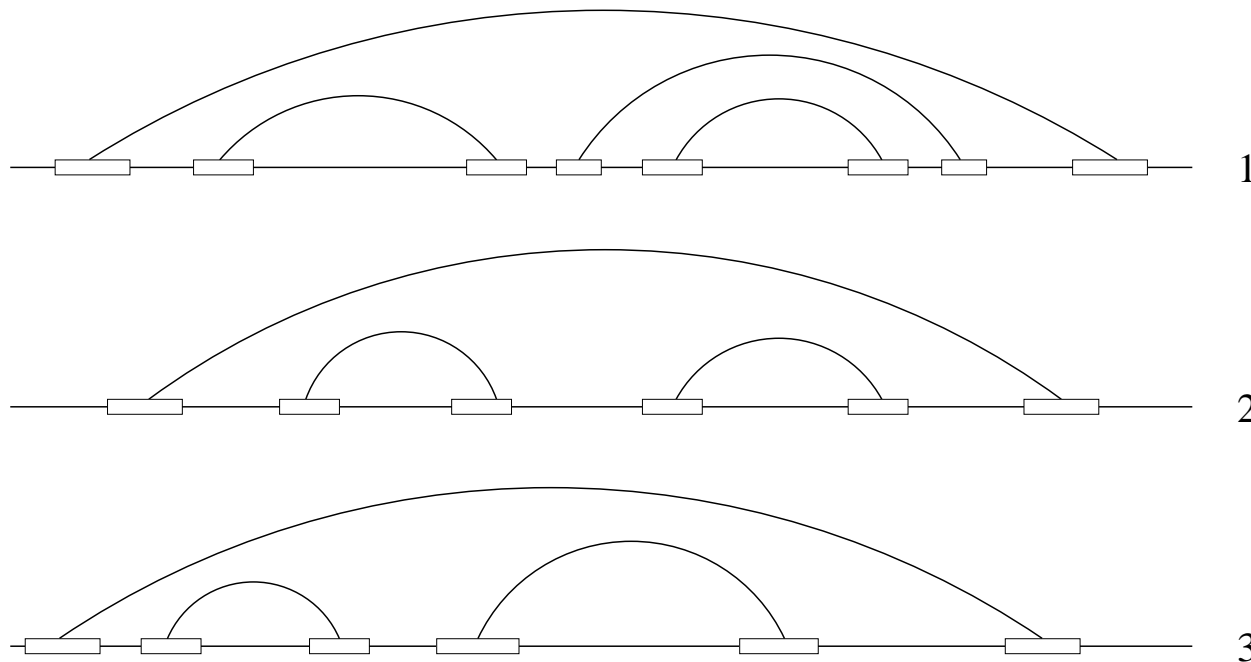
Stage 1 – Building the *graph of stems*



From 2 to p sequences : CARNAC

Stage 0 – Computing of all pairwise foldings with CARNAC₂

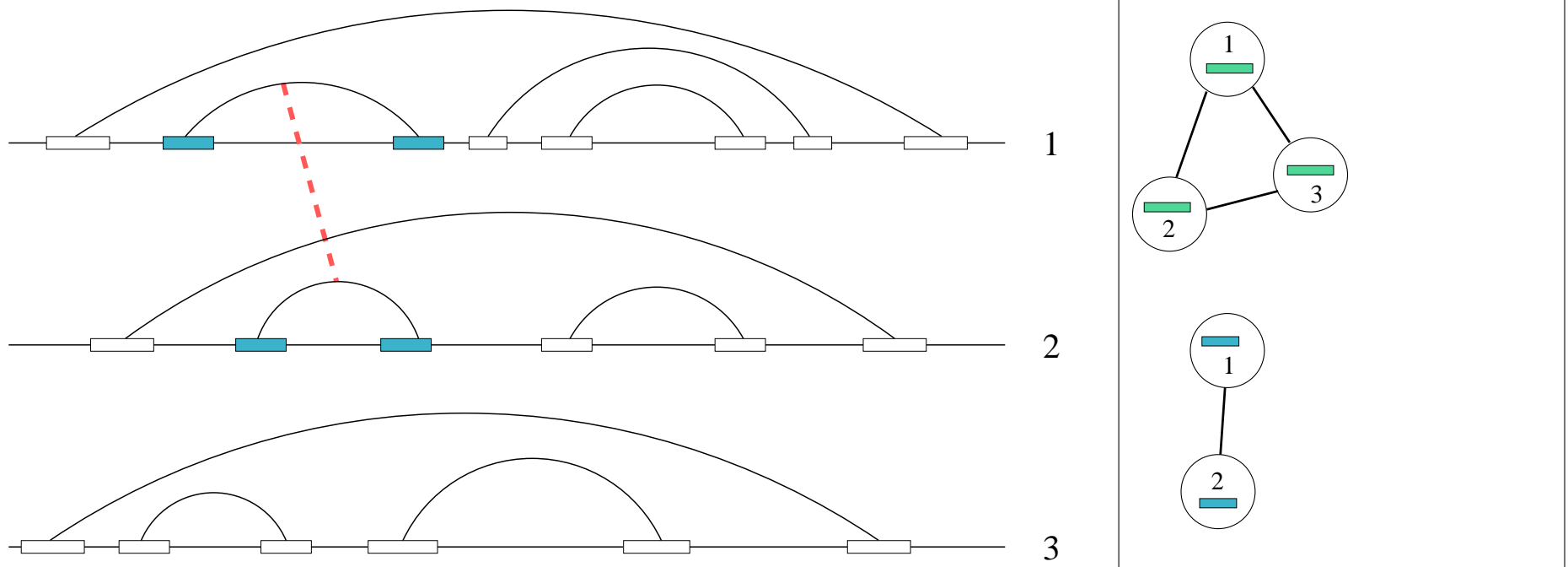
Stage 1 – Building the *graph of stems*



From 2 to p sequences : CARNAC

Stage 0 – Computing of all pairwise foldings with CARNAC₂

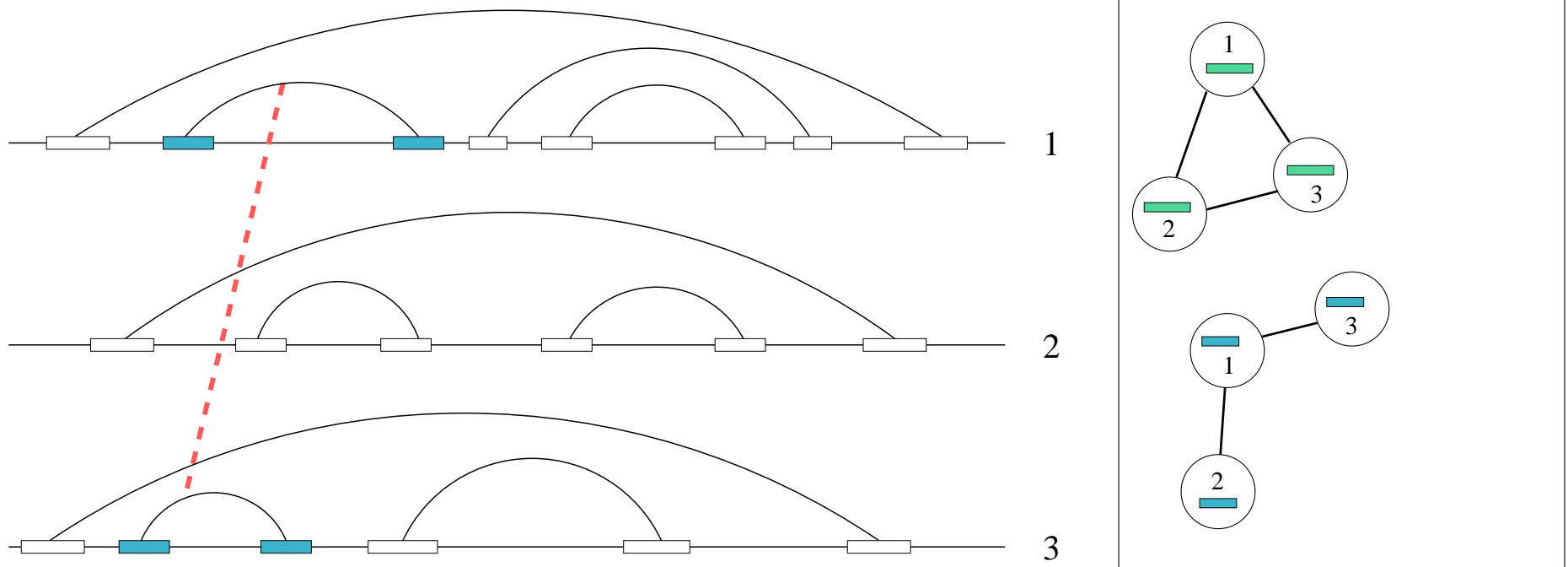
Stage 1 – Building the *graph of stems*



From 2 to p sequences : CARNAC

Stage 0 – Computing of all pairwise foldings with CARNAC₂

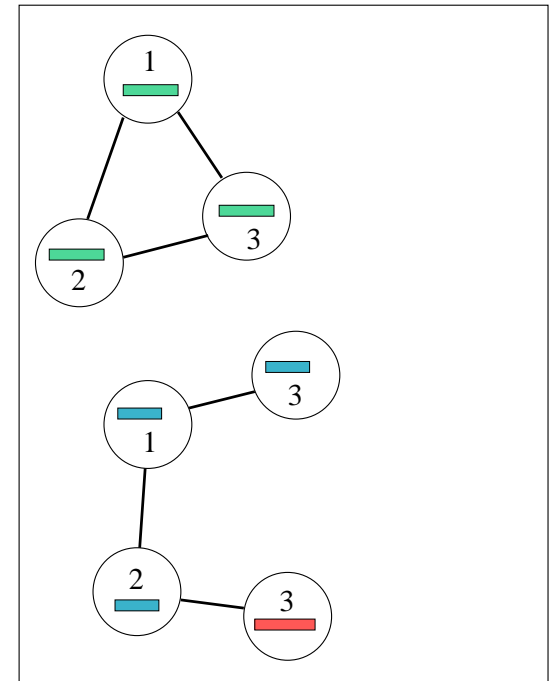
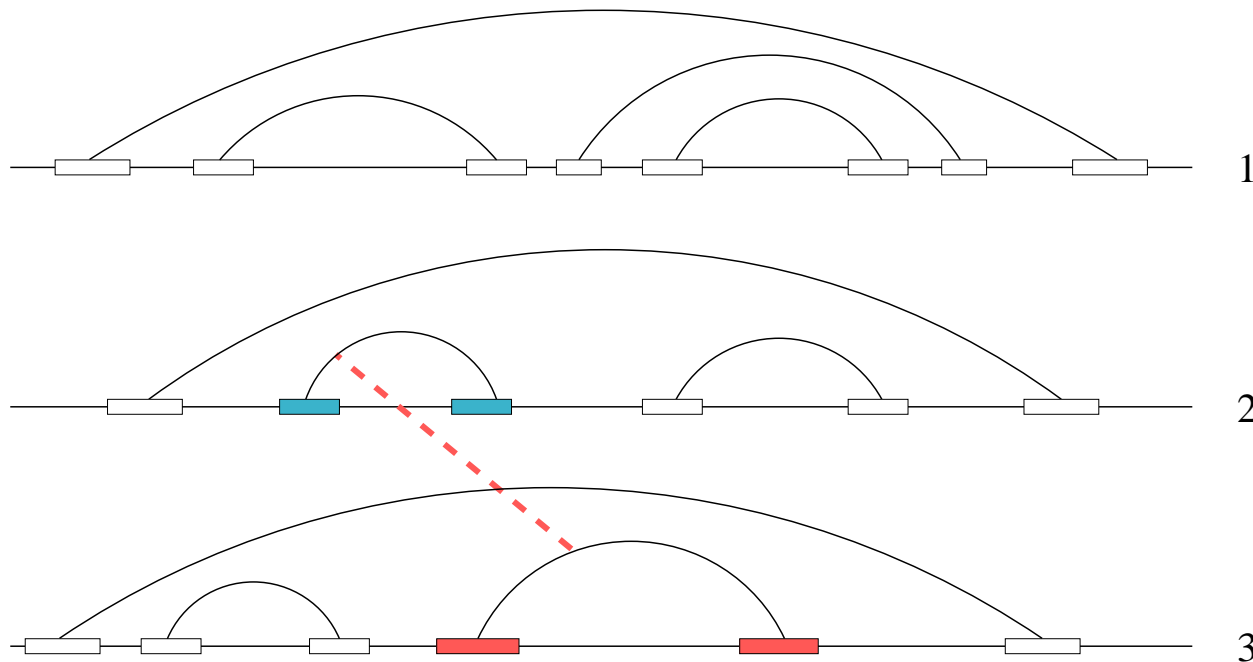
Stage 1 – Building the *graph of stems*



From 2 to p sequences : CARNAC

Stage 0 – Computing of all pairwise foldings with CARNAC₂

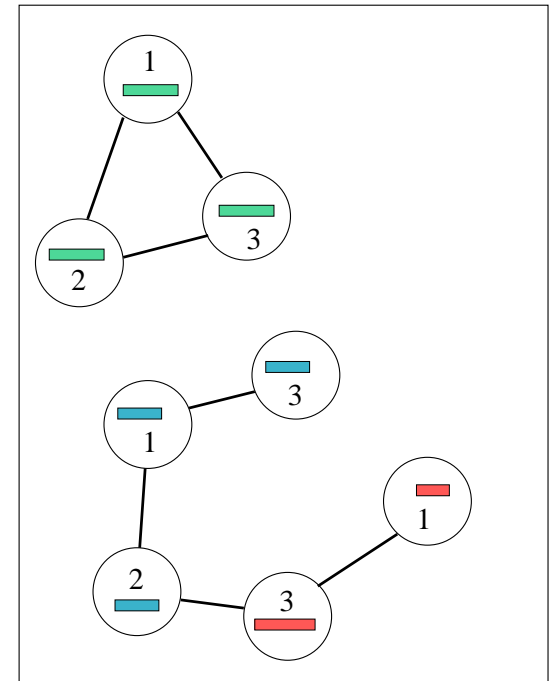
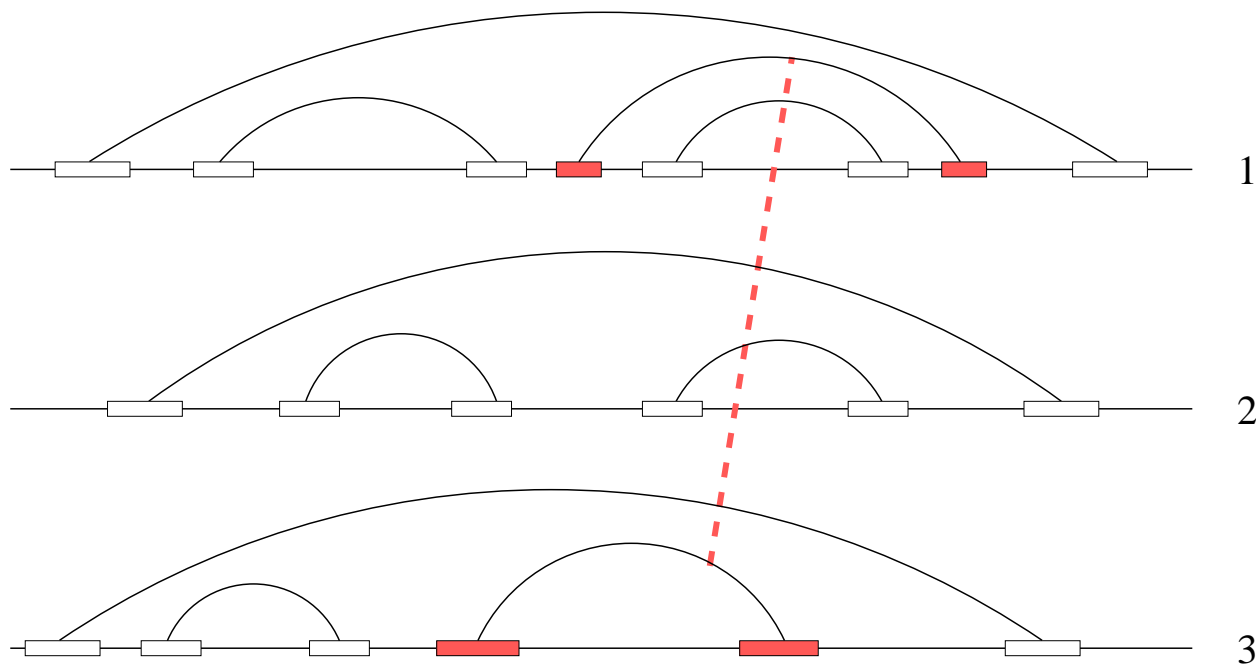
Stage 1 – Building the *graph of stems*



From 2 to p sequences : CARNAC

Stage 0 – Computing of all pairwise foldings with CARNAC₂

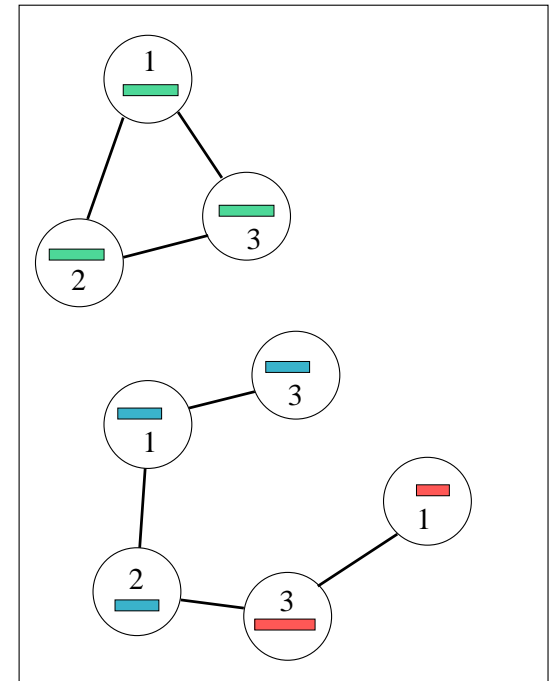
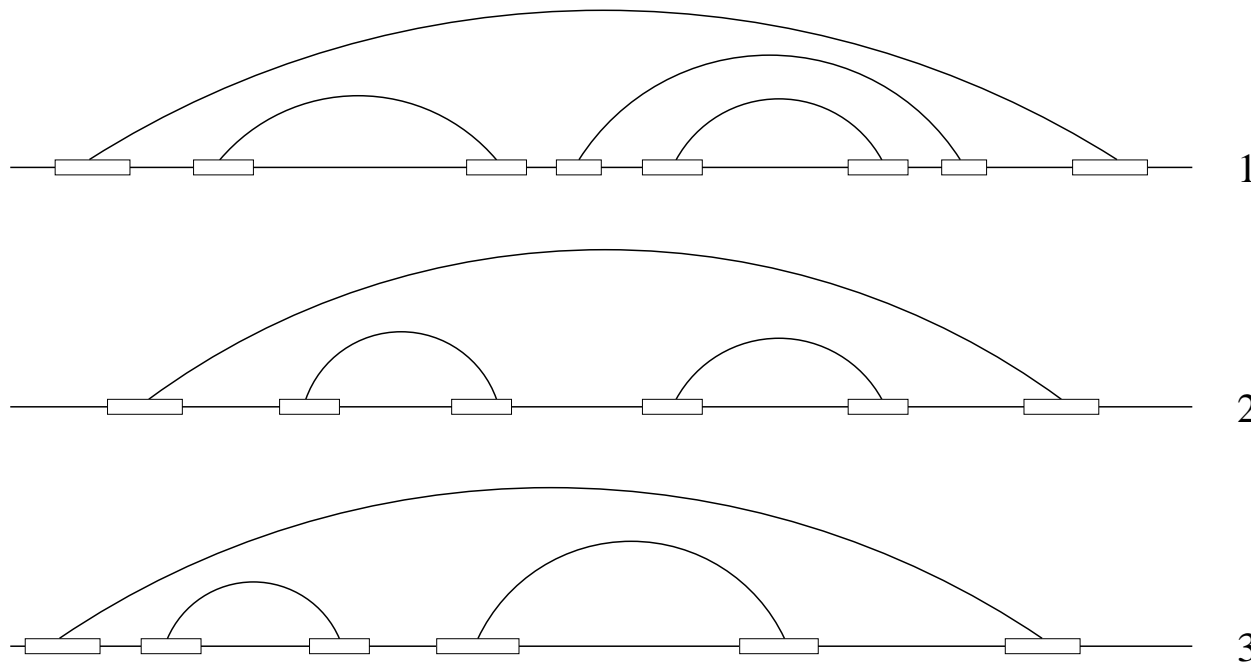
Stage 1 – Building the *graph of stems*



From 2 to p sequences : CARNAC

Stage 0 – Computing of all pairwise foldings with CARNAC₂

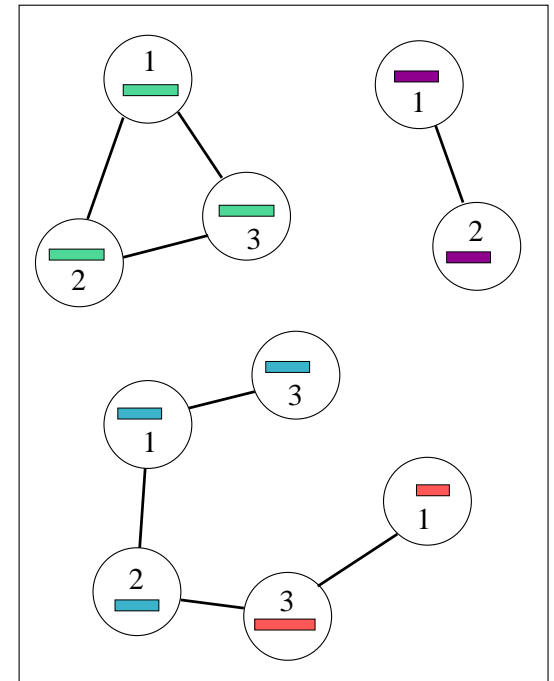
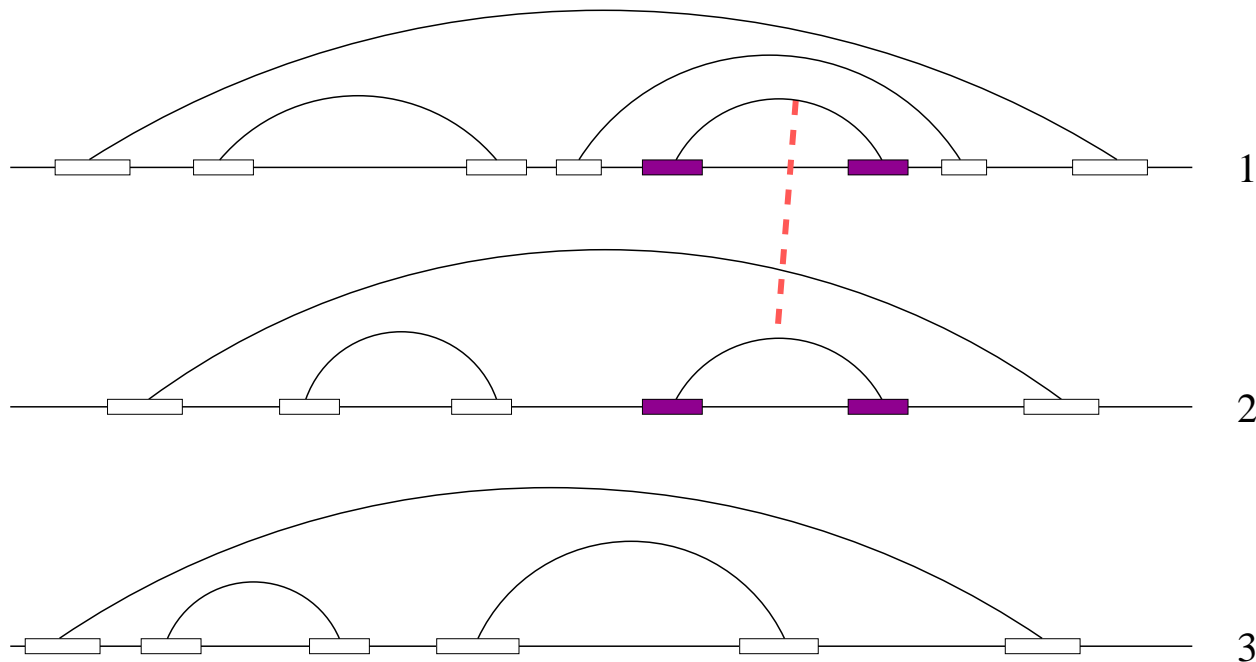
Stage 1 – Building the *graph of stems*



From 2 to p sequences : CARNAC

Stage 0 – Computing of all pairwise foldings with CARNAC₂

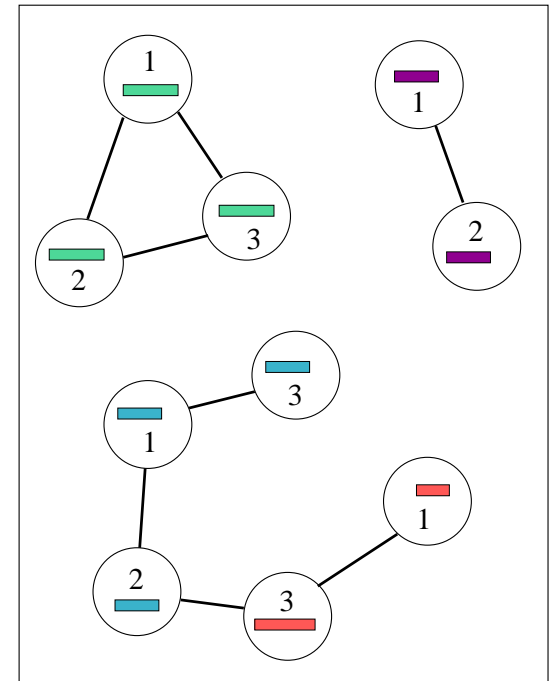
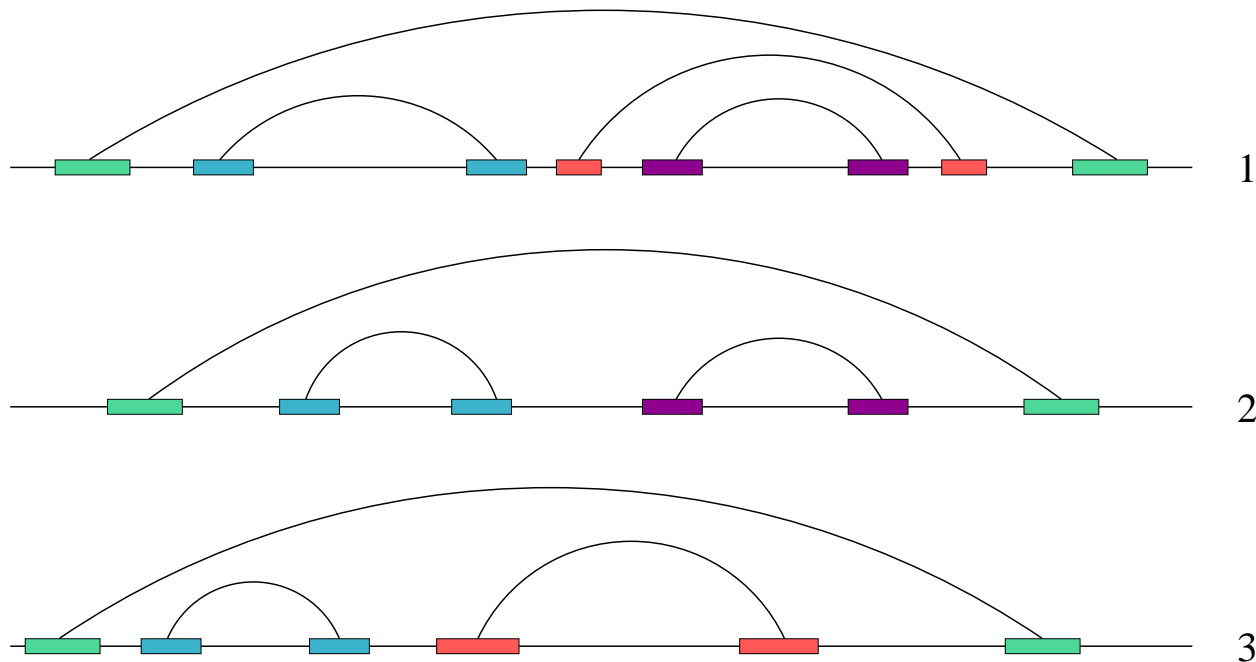
Stage 1 – Building the *graph of stems*



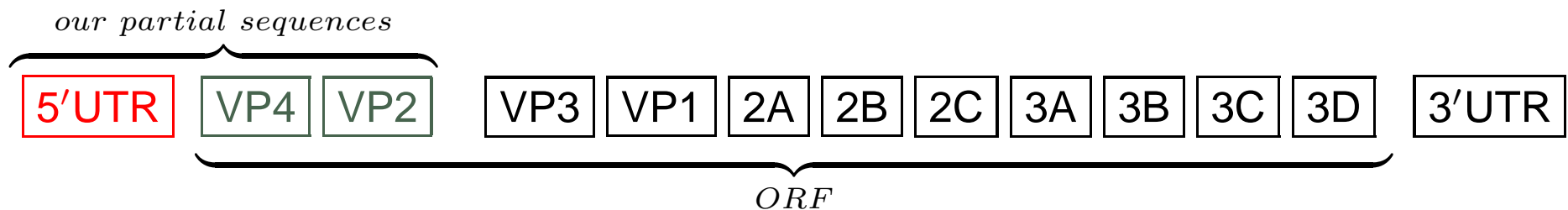
From 2 to p sequences : CARNAC

Stage 0 – Computing of all pairwise foldings with CARNAC₂

Stage 1 – Building the *graph of stems*



Enterovirus mRNA

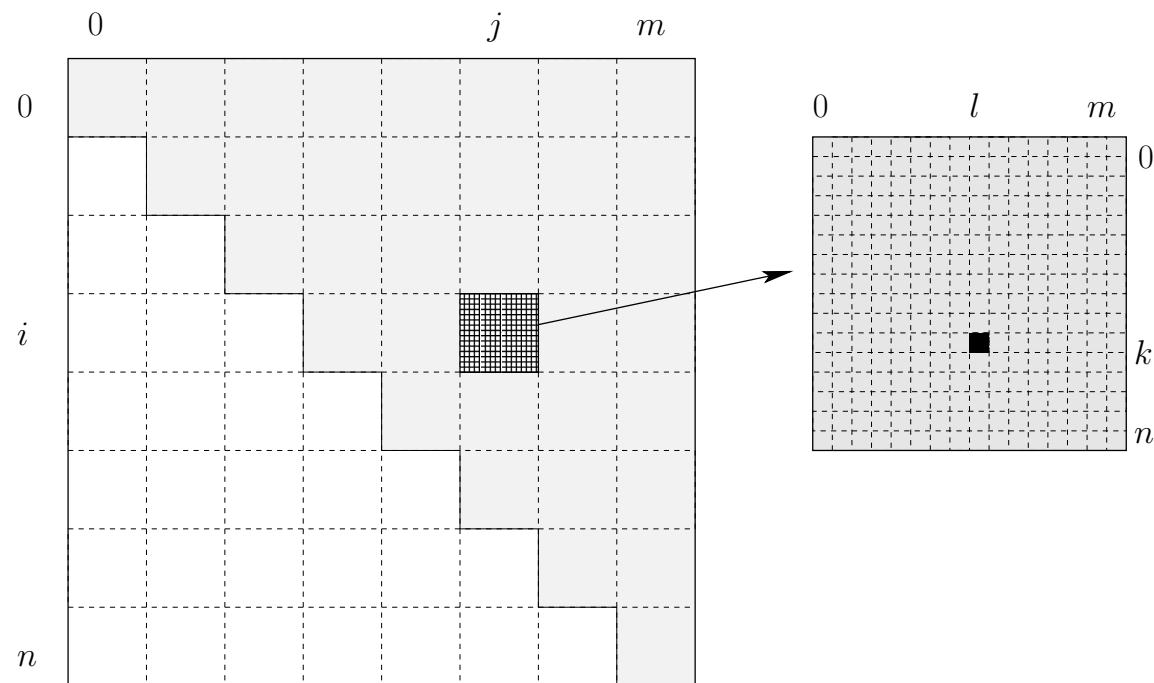


- length 1800 nt. (START codon at position 700)
- conserved structure within the 5'UTR [*Le 1990*]

more accuracy: back to the base level

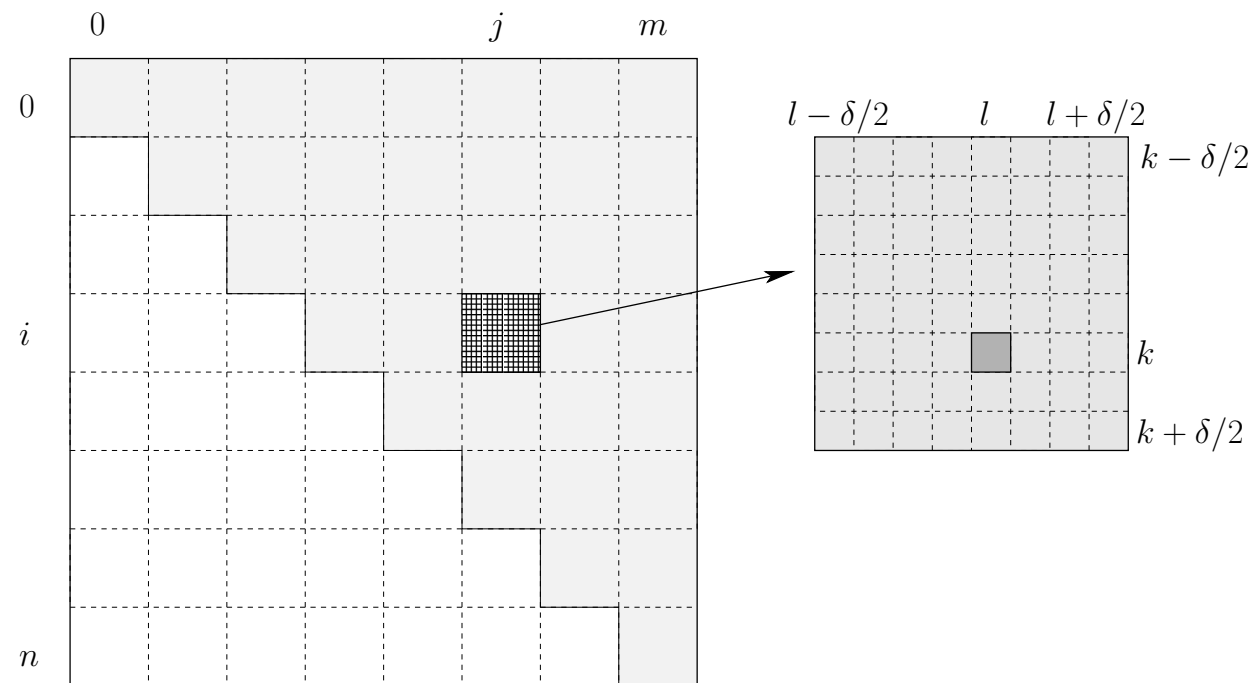
ARNICA – [*Perriquet 2009*]

4D matrix (Sankoff recursions)



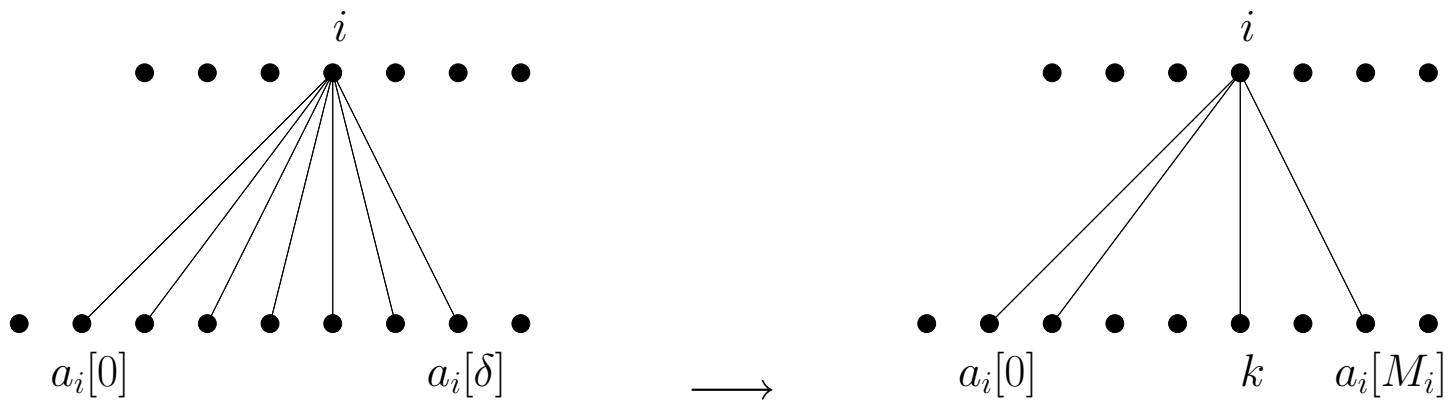
The value in the cell $[i, j, k, l]$ stores the score for the best structural alignment between the segments $seq_0[i..j]$ and $seq_1[k..l]$

4D matrix and banding heuristic

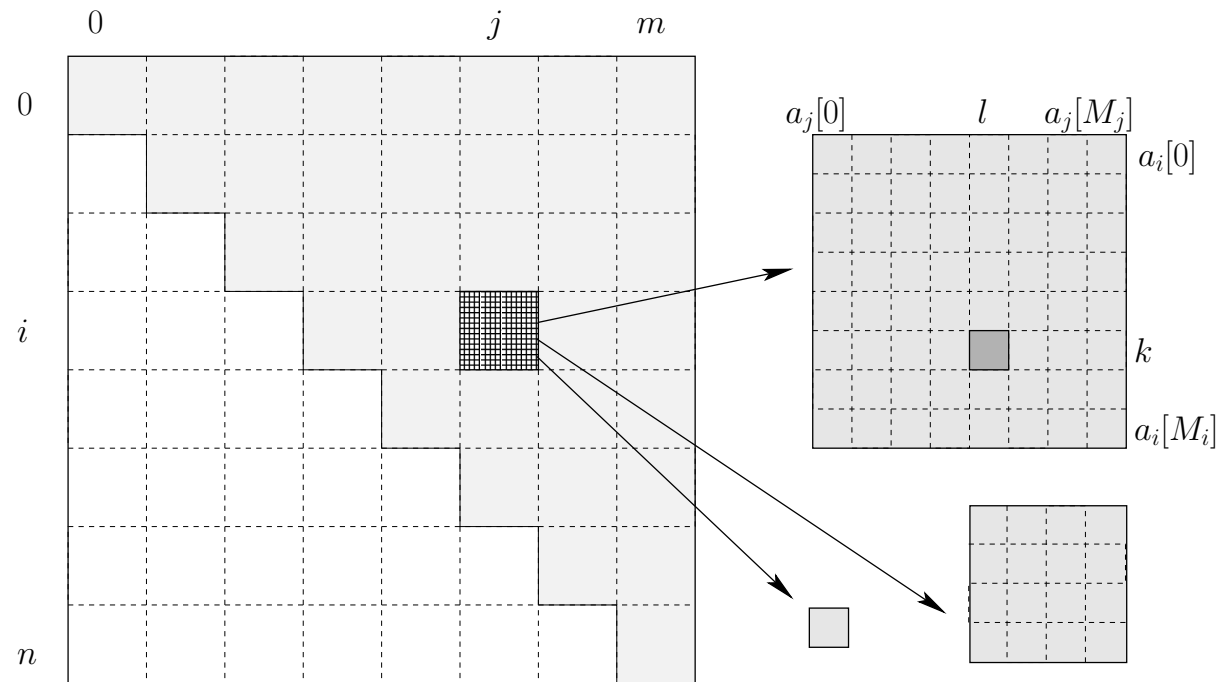


The value in the cell $[i, j, k, l]$ stores the score for the best structural alignment between the segments $seq_0[i..j]$ and $seq_1[k..l]$

local banding heuristic

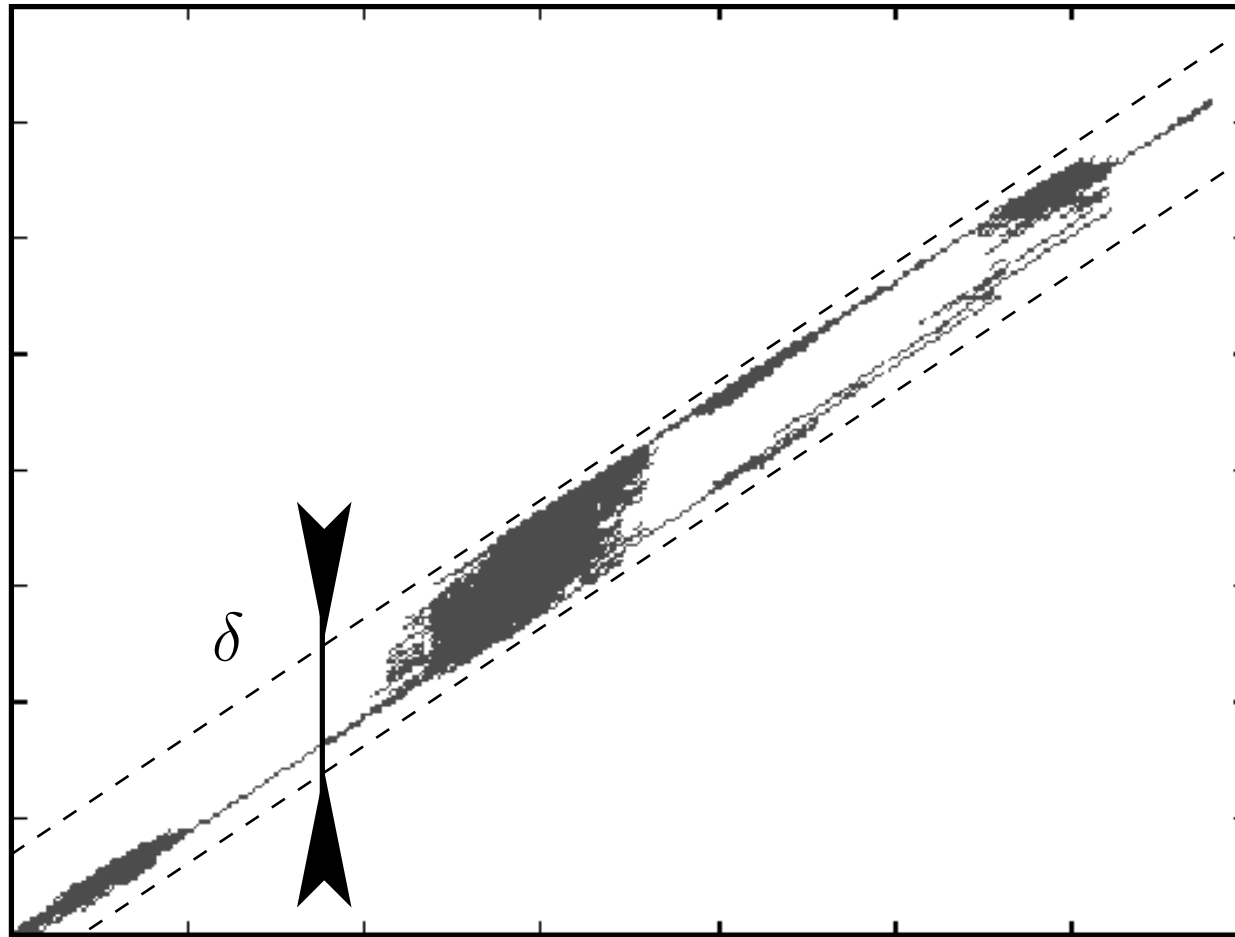


4D matrix and local banding heuristic



Each cell can be a 2D matrix of any size

ARNICA – [*Perriquet 2009*]



Suboptimal alignments of two RNase P RNA (*D. desulfuricans* vs *A. eutrophus*)

ARNICA vs. FOLDALIGN 2

	option	specificity	sensitivity	time (sec.)	space (Mb)
FOLDALIGN 2	local	56.2%	40.5%	1107	142.1
ARNICA	thd 0	73.6%	43.3%	6	16.5
	thd 10	74.1%	45.9%	7	16.7
	thd 30	75.7%	51.0%	10	17.5
	thd 50	76.2%	55.0%	20	19.1
	thd 80	75.7%	58.3%	77	23.7
	thd 100	74.7%	58.7%	148	29.0
	thd 150	73.9%	60.5%	536	46.5
	thd 200	73.2%	61.0%	1079	64.4

Average performance of ARNICA and FOLDALIGN 2
 (RNase P RNA, 11 seq. from <http://www.mbio.ncsu.edu/RNaseP/>)

ARNICA – [*Perriquet 2009*]

next steps

- . p sequences
- . integrate the full thermodynamic model
- . allow dynamic restraints on the fly

CARNAC – [*Perriquet 2004*]



?

ARNICA – [*Perriquet 2009*]

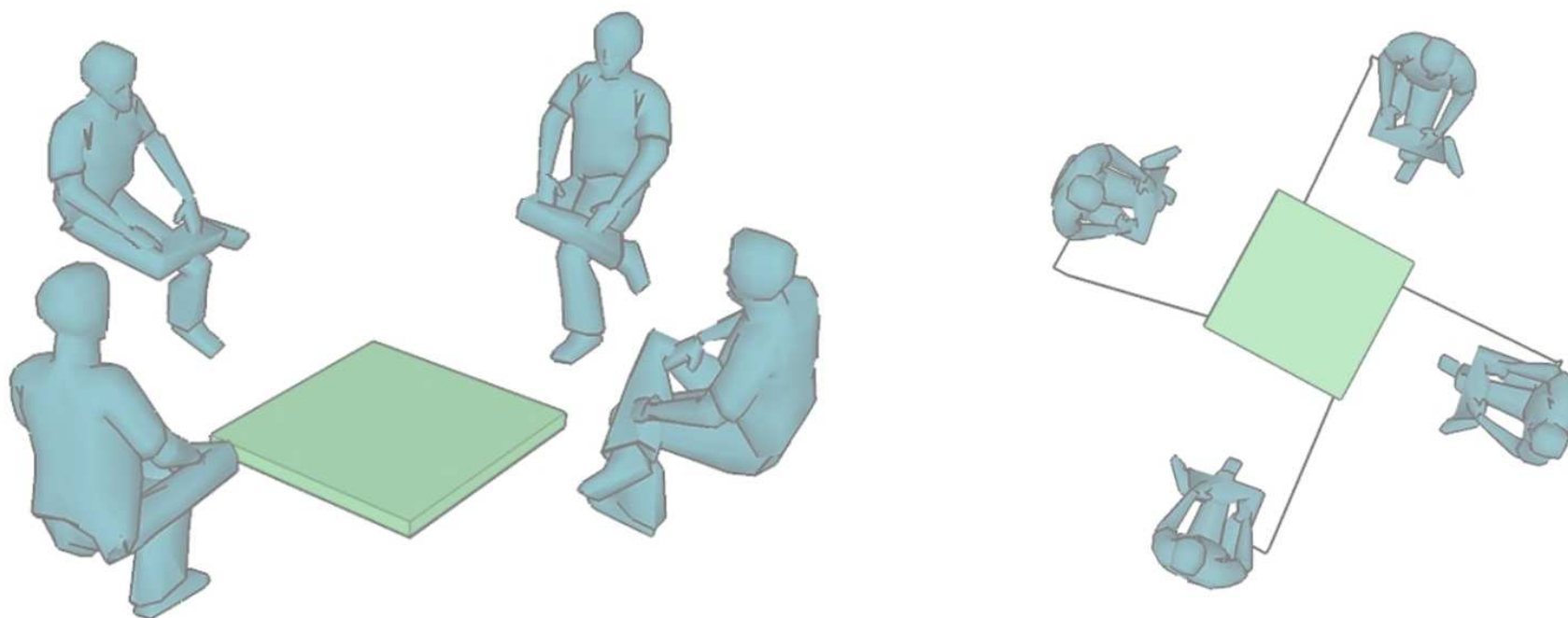
Prélude à Transformation Naturelle

$$\begin{array}{ccc} F(X) & \xrightarrow{F(f)} & F(Y) \\ \eta_X \downarrow & & \downarrow \eta_Y \\ G(X) & \xrightarrow{G(f)} & G(Y) \end{array}$$

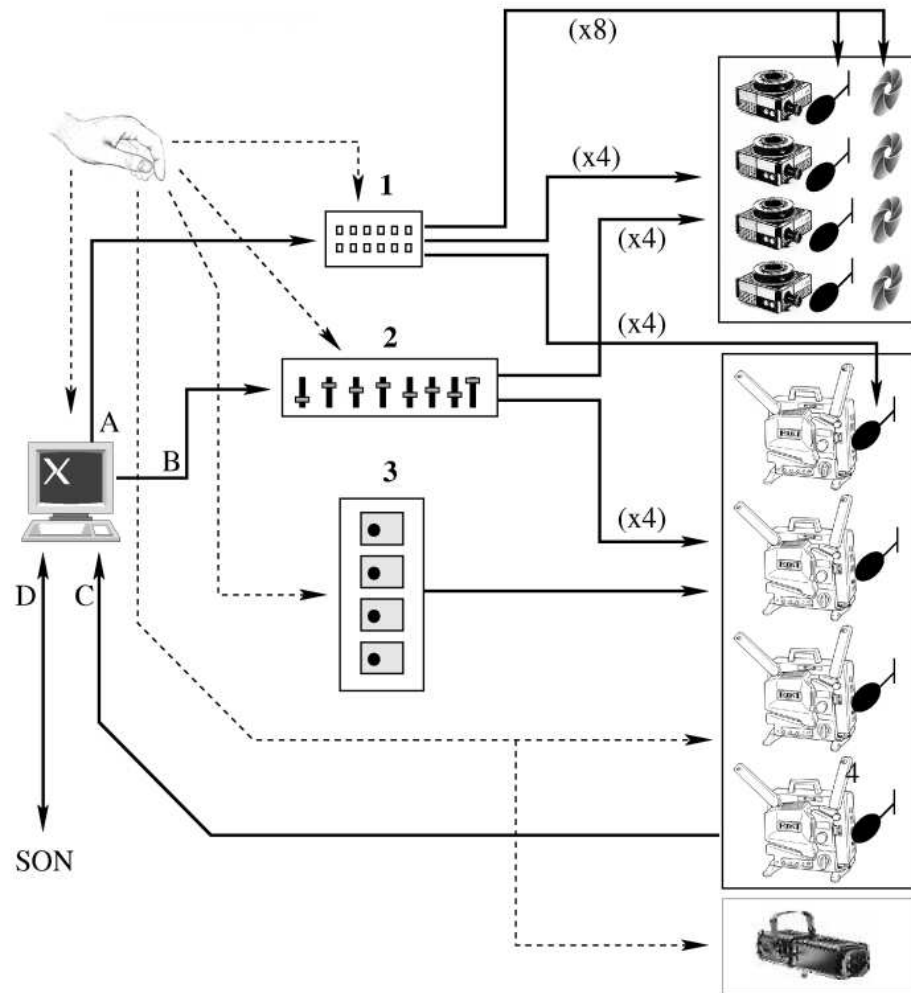


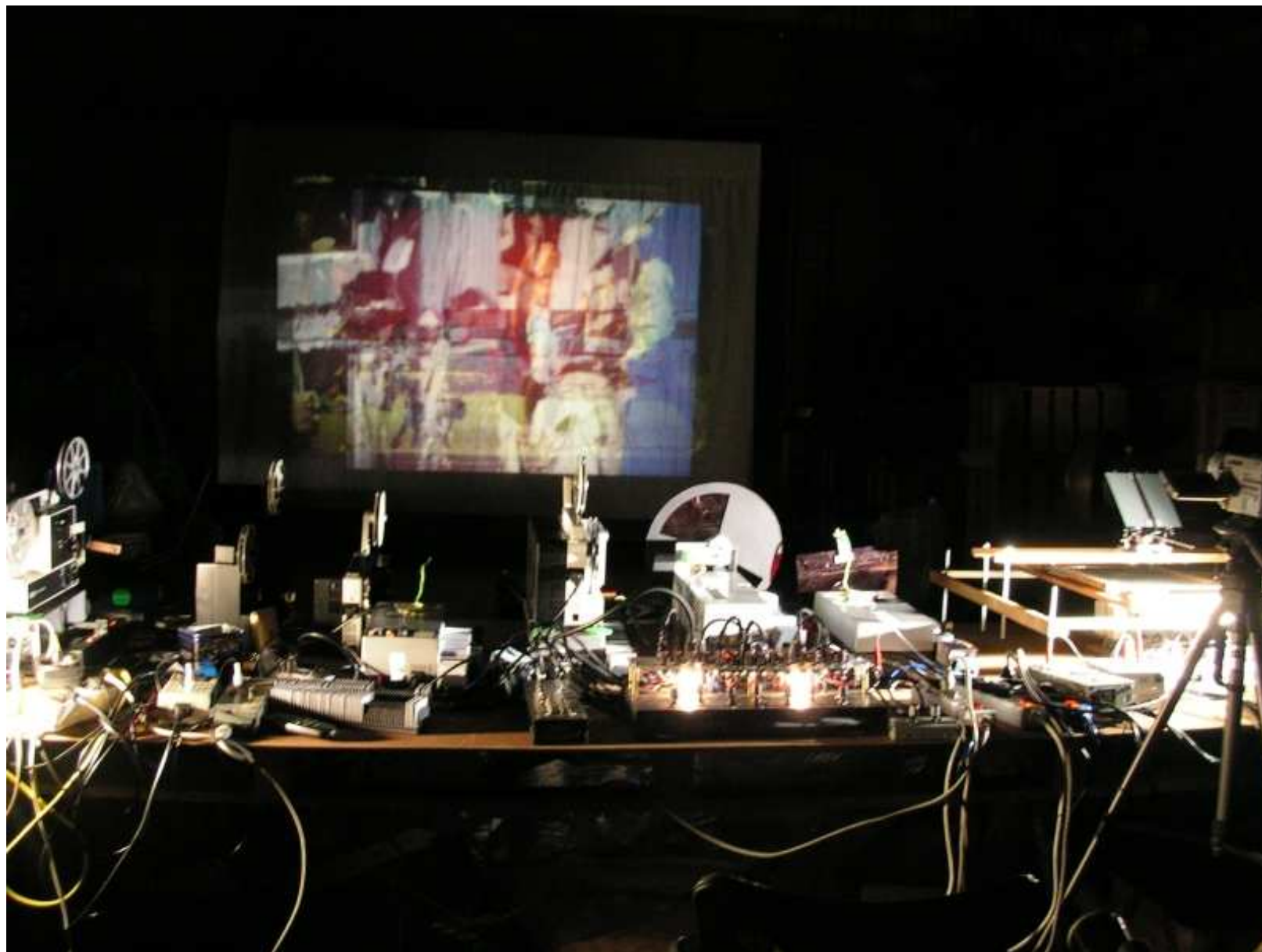
ENACTED EMERGENCE

> Hybrid Games <



ENACTED EMERGENCE





<http://cesium-133.net>