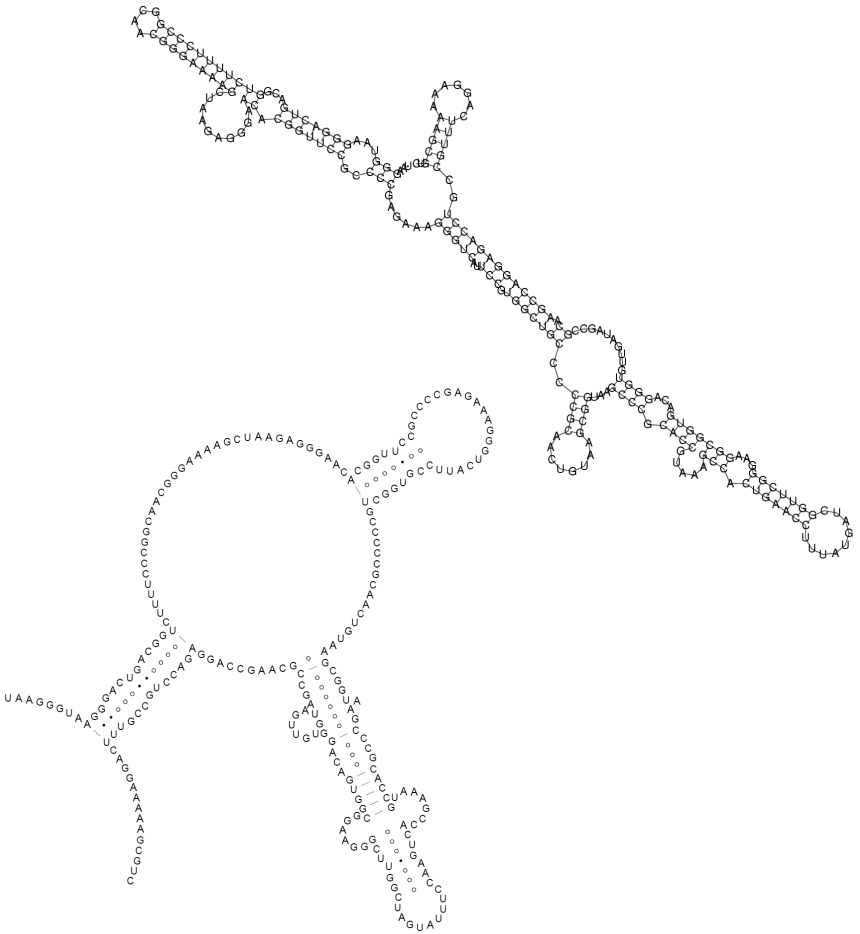
Reformulate the folding problem

RNA secondary structure prediction problem

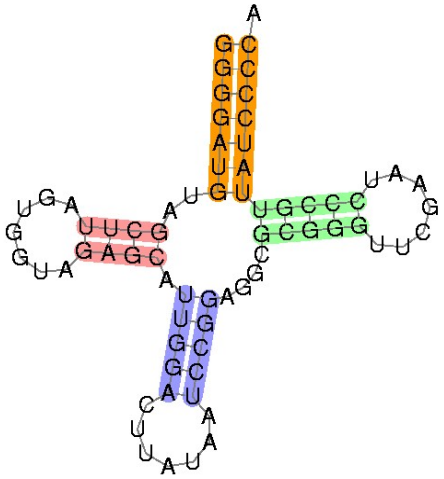
- Algorithms/programs to compute the minimum energy:
  - Nussinov et al (1978), Waterman (1978), Smith and Waterman (1978), and Zuker and Sankoff (1984).
  - **Mfold** (Zuker 2003) and **RNAfold (ViennaRNA)** (Hofacker 2003).

- RNA folding via energy minimization has its shortcomings:
  - Prediction depends on correct energy parameters.
  - Sometimes, the true structure does not have the minimum energy.



How to find ncRNA genes from a multiple alignment

- RNAs with similar functions often have similar structures.
- Sequence changes can be tolerated by covarying mutations.



RNAalifold

- If we have correct multiple alignments, looking for covarying mutations and finding consensus structure is a good way to do structure prediction.
  - **RNAalignfold** (Hofacker et al. 2002)
  - The consensus structure prediction is more accurate.
  - To find energetically stable consensus structure is more statistically significant.
  - Still compute the MFE.
  - Covariance information is incorporated into the energy model by rewarding compensatory and consistent mutations.

$$E_{i,j} = \min\left\{ E_{i,j-1};\ \min_{\substack{k:\ i+m<k\le j \\ \Pi_{ik}=1}} E_{i+1,k-1} + E_{k+1,j} + \beta_{ik} \right\}$$

Problem:

- Correctly aligning multiple and divergent RNA sequences without taking into account the structural information is difficult.
- To get covarying mutation, we need an alignment.
- Do the multiple alignment and structure prediction at the same time.

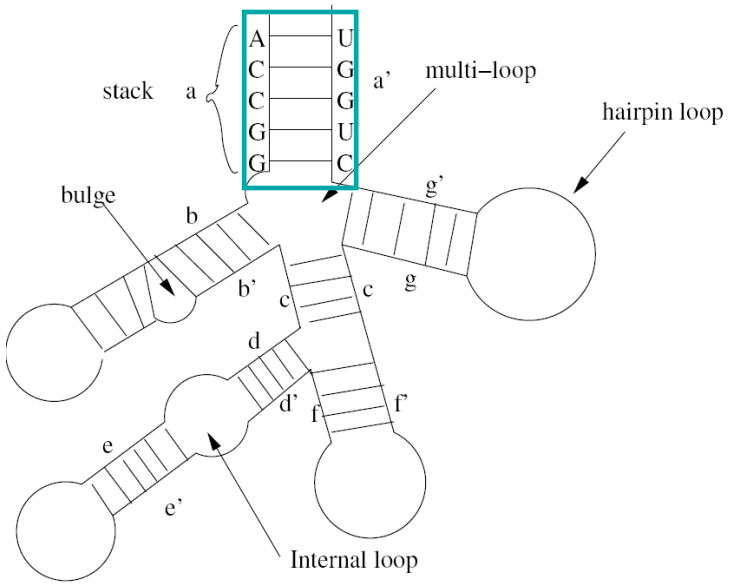
RNA consensus folding problem

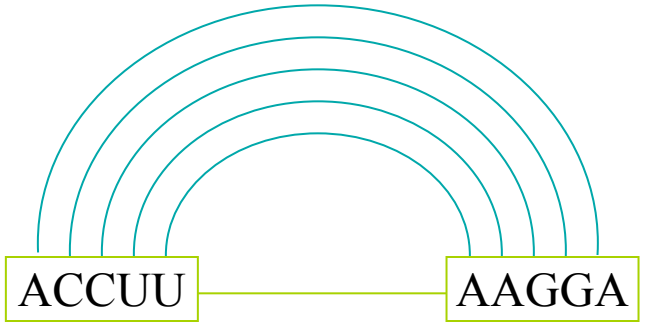
RNA consensus folding problem: computing the common secondary structure for a set of unaligned RNA sequences

- Sankoff (1985) first proposed an algorithm simultaneously align RNA sequences and find the optimal common fold.
  - Time complexity is $O(n^6)$ for two sequences with length n.
  - Implemented as **Dynalign** (Mathews and Turner 2002)
  - It's not practical for multiple sequences.

- Eddy and Durbin (1994) and other groups used stochastic context-free grammars to predict the consensus structure.
  - Start from a seed alignment.
  - Stochastic iteration to improve the alignment and predict structure.
  - Need a good seed alignment.

Motivation to a new approach

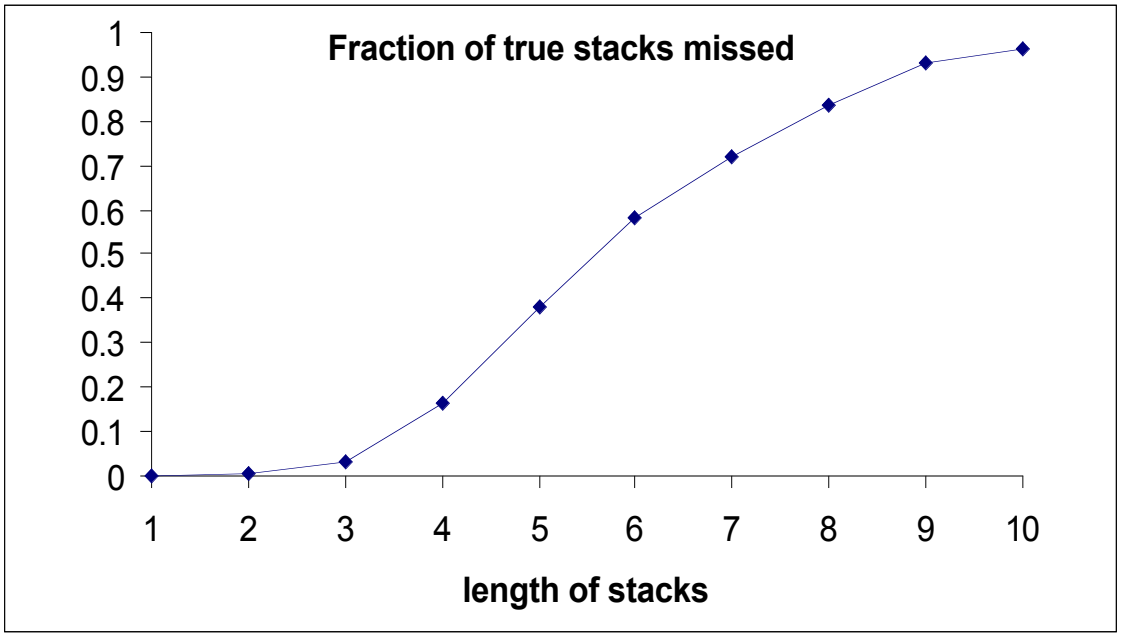




$p = (1/4)^5 < 0.001$.

- Base-pairs appear in ‘clusters’: we call them stacks, which is energetically favorable.
- Most of the stability of the RNA secondary structure is determined by stacks.
- Stacks are much less likely to occur by chance.

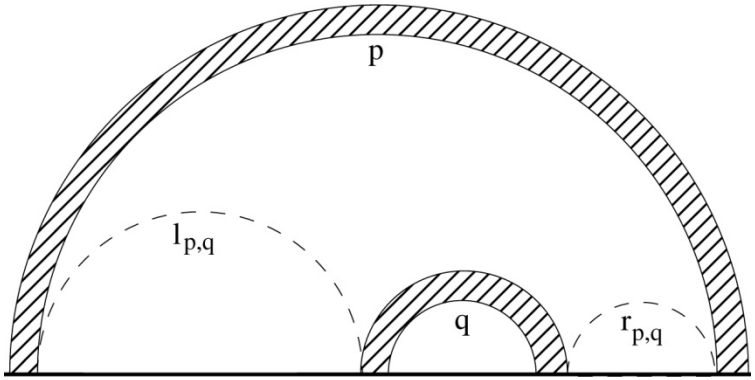
Statistics of the stacks in Rfam database



The *Stack-based Folding* algorithm based on Nussinov model

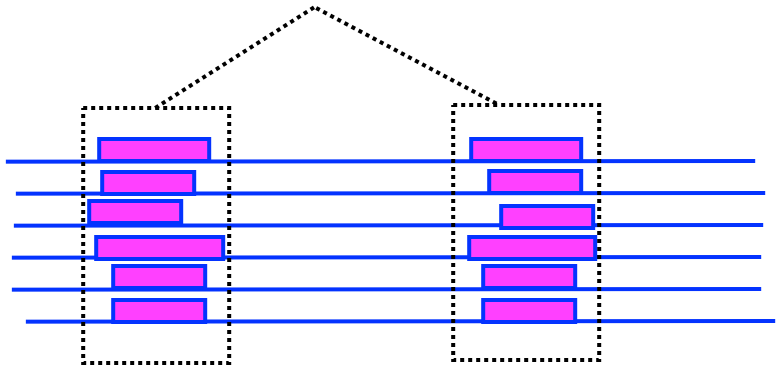
- The Nussinov model (simplified model)
  - Basically counts the number of base pairs
- The *Stack-based Folding* algorithm
  - N(p) : the maximal number of base pairs of all stack configurations within *p* (including *p* itself)

$$N(p) = p_l + \max_{\forall q \in \mathcal{F}_I(p)} \{N(q) + N(l_{p,q})\}$$



Using stacks as anchors for predictions

- The idea of anchors as constraints has been used in multiple genomic sequence alignment.
  - MAVID (Bray and Pachter, 2004)
  - TBA (Blanchette et al., 2004)



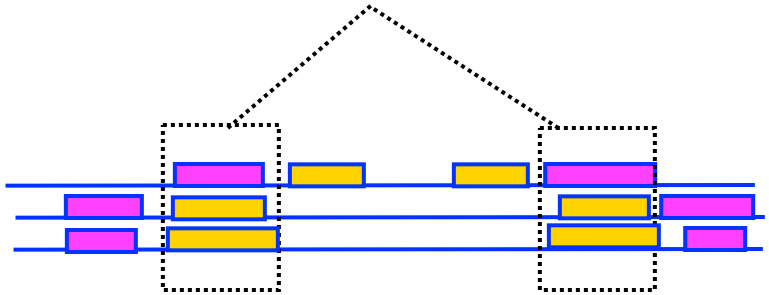
Several heuristic methods have been developed by finding anchored stacks:

- Waterman (1989) used a statistical approach to choose conserved stacks within fixed-size windows.
- Ji and Stormo (2004) and Perriquet et al. (2003) use primary sequence conservation of the stacks and the length of loop regions to reduce the searching space.

- stack anchor has low sequence similarity.
- It's hard to find correct anchors

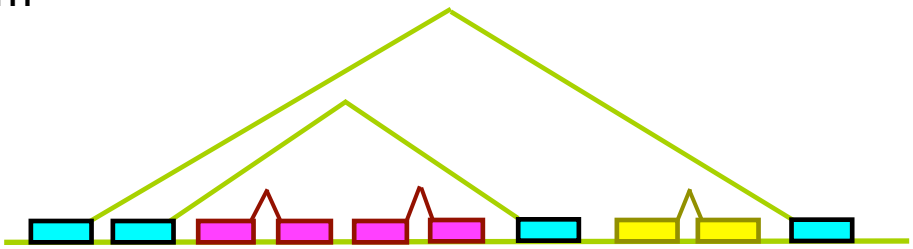
Problem:

- Selecting one stack at a time may cause wrong matching stacks.



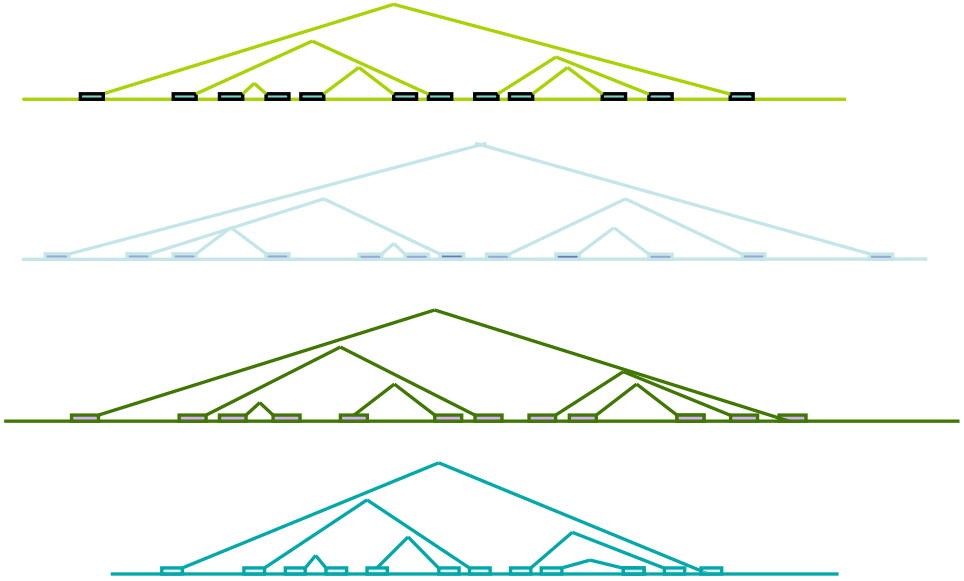
A global approach: configuration of stacks

- RNA secondary structure can be viewed as stacks plus unpaired loops. (no individual base-pairs)
- The energy of the structure is the sum of the energies of stacks and loops.
- Stack configuration:
  - Nested stacks
  - Parallel stacks
  - Crossing stacks (pseudo knots)
- More generalized stacks can include mismatches in the stacks.

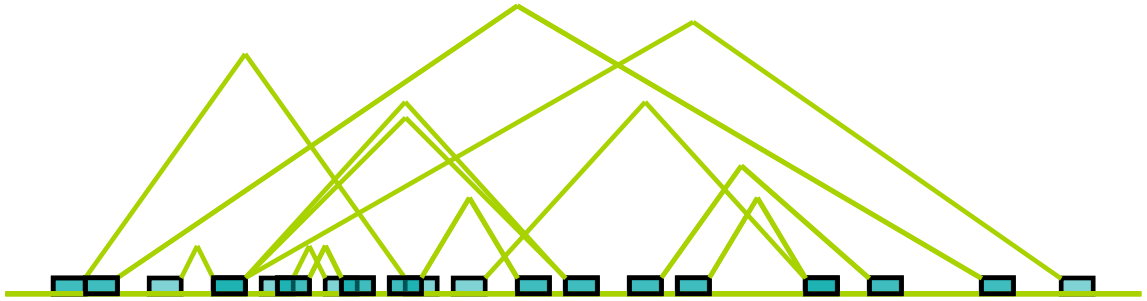


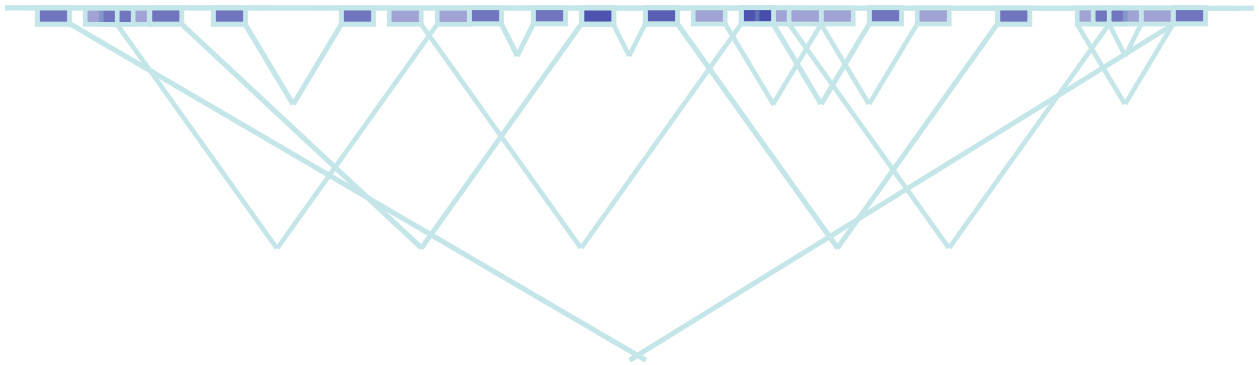
RNA Stack-based Consensus Folding (RNAscf) problem

- Find conserved stack configurations for a set of unaligned RNA sequence.
- Optimize both stability (free energy) of the structure and sequence similarity computed based on these common stacks as anchors.

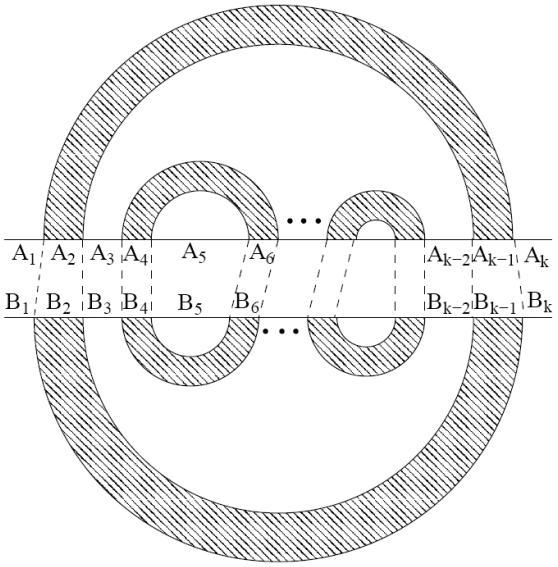


RNA stack-based consensus folding for pairwise sequences





A matching stack-configurations on two sequences



$$M(\mathcal{P}(A,B)) = w_1\Phi(\mathcal{P}(A,B)) + w_2 \sum_{i\in\mathcal{P}(A,B)} \mathcal{S}(A_i,B_i) + w_3 \sum_{i\notin\mathcal{P}(A,B)} \mathcal{S}(A_i,B_i)$$

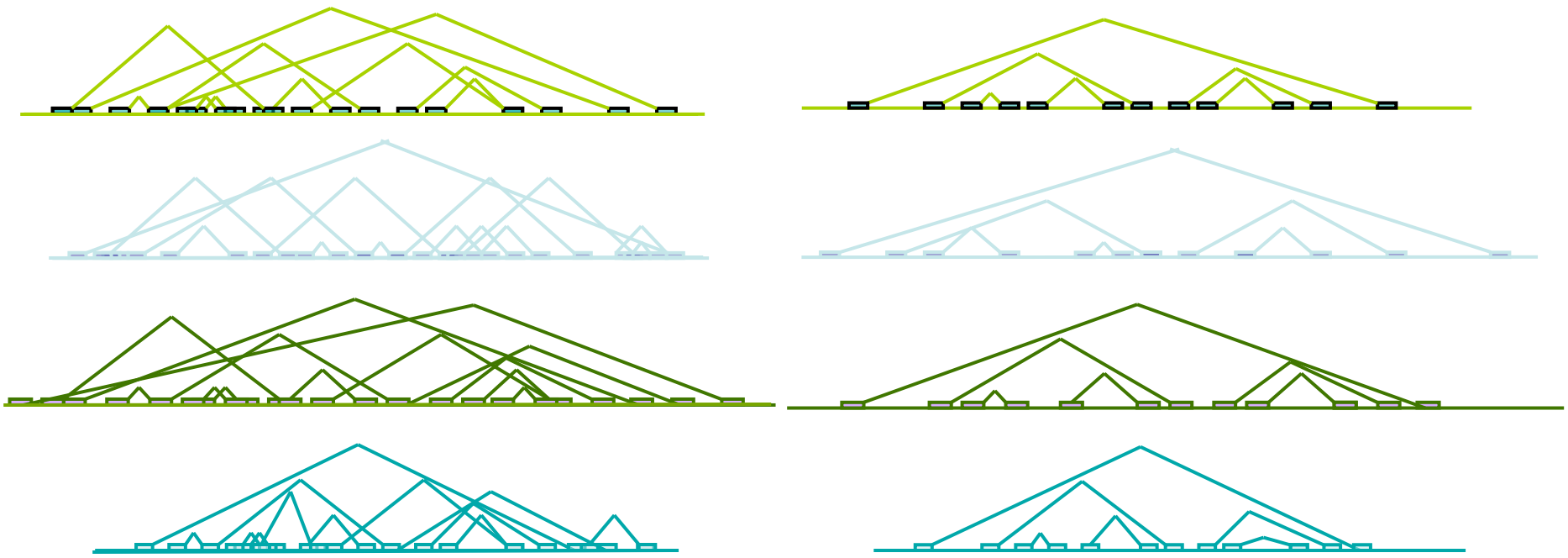
Weights of different costs

Energy of the consensus structure

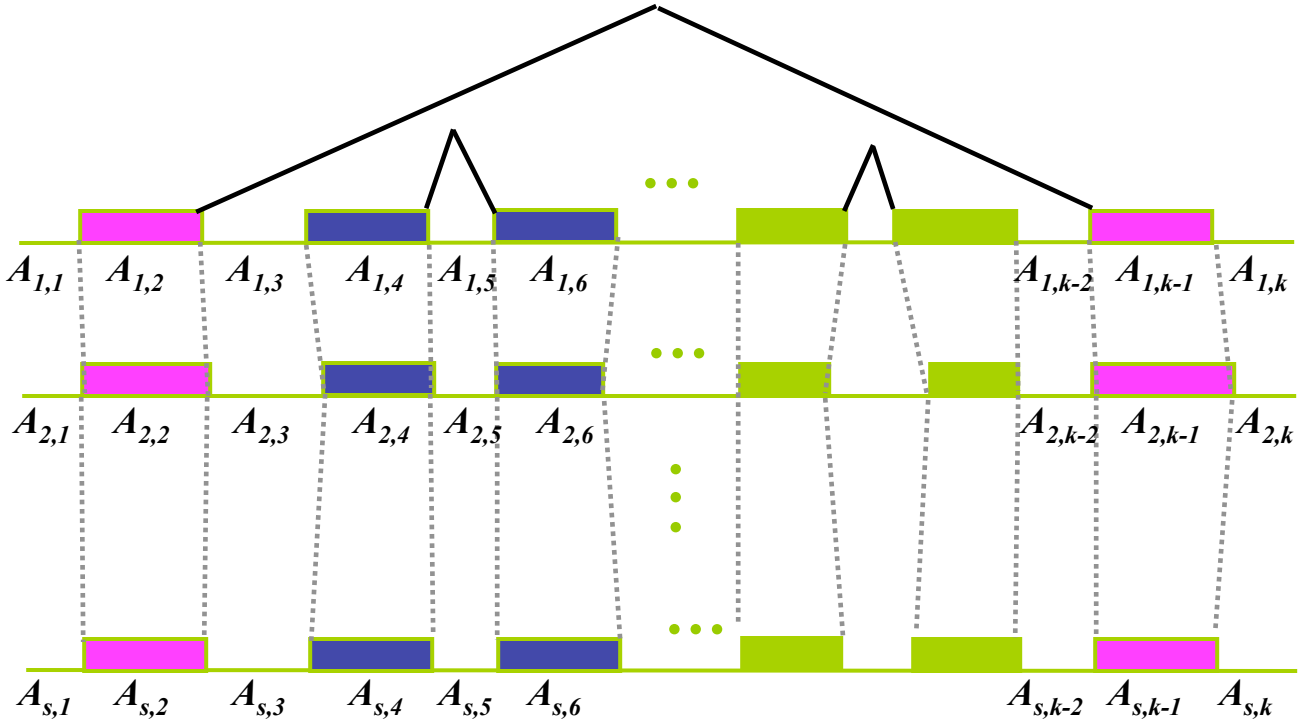
Sequence similarity of unpaired regions

Sequence similarity of stacks

RNA Stack-based Consensus Folding for multiple sequences



Cost function for multiple sequences

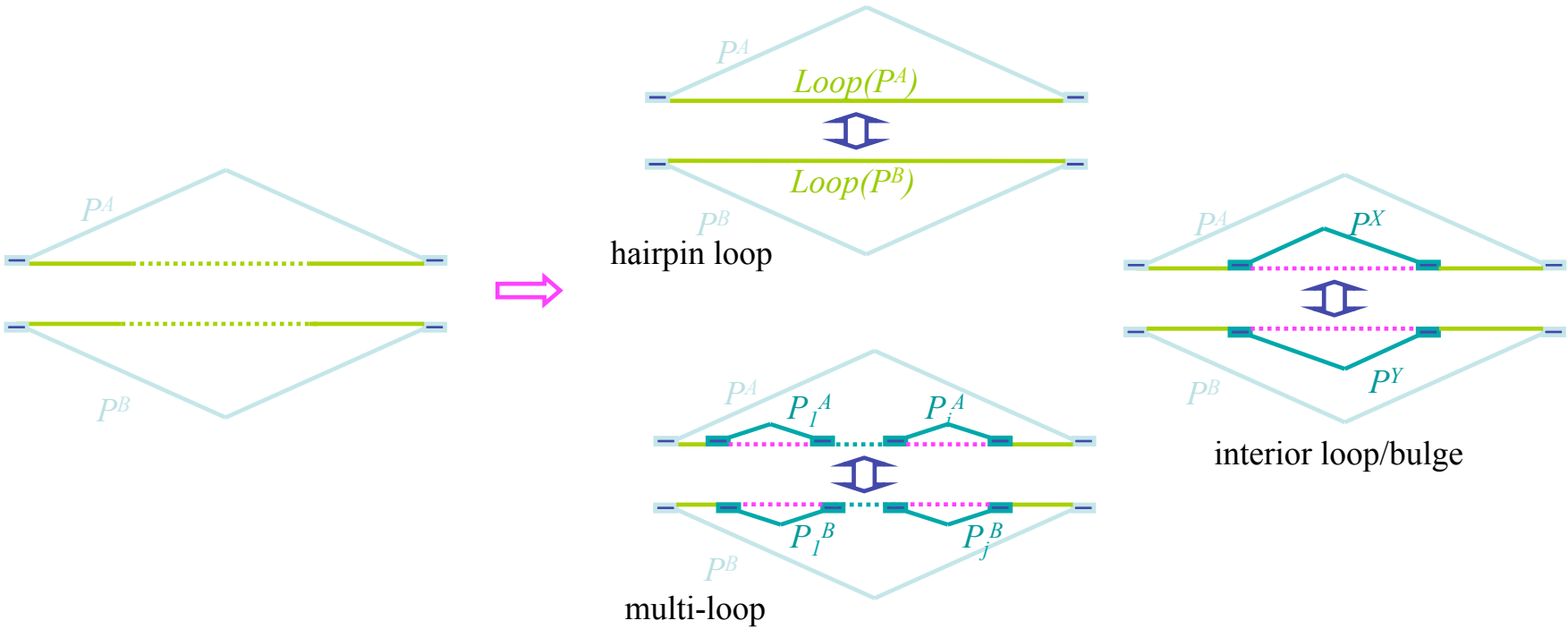


$$M(\mathcal{P}(A_1,\ldots,A_s)) = w_1\Phi(\mathcal{P}(A_1,\ldots,A_s)) + w_2\sum_{j\in\mathcal{P}}\mathcal{S}\begin{pmatrix}A_{1,j},\\A_{2,j},\\\ldots,\\A_{s,j}\end{pmatrix} + w_3\sum_{j\notin\mathcal{P}}\mathcal{S}\begin{pmatrix}A_{1,j},\\A_{2,j},\\\ldots,\\A_{s,j}\end{pmatrix}$$

Compute an optimal stack configuration for two sequences

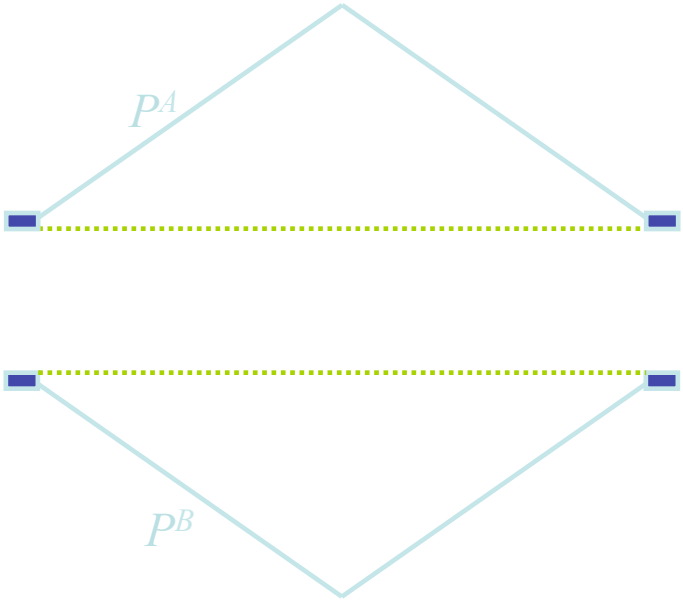
- Dynamic programming algorithm is used to align RNA sequences and find an optimal configuration at the same time.
  - The algorithm is similar to prior work (Sankoff 1985, Bafna et al. 1995)
  - Differences:
    - We use stacks as the basic structural elements.
    - Prior work used individual base pairs.
  - The computational time is $O(n^4)$ (n is the number of stacks).
    - Sankoff's algorithm is $O(m^6)$, (m is the length of the sequences).
    - The number of possible stacks (size >= 4) is much smaller than the length of the sequence.
    - It's much faster.

For any pair of stacks, there are three choices:



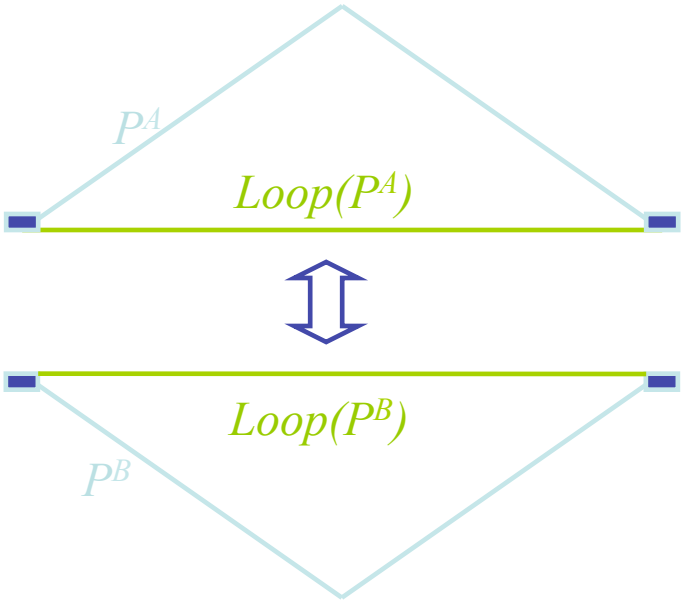
$$M[P^A, P^B] = M_s[P^A, P^B] + \min \left\{ \begin{array}{ll} M_h[P^A, P^B], & \text{(* hairpin loop *)} \\ M_b[P^A, P^B], & \text{(* interior loop/bulge *)} \\ M_m[P^A, P^B] & \text{(* multi-loop *)} \end{array} \right\}$$

The score of matching stacks:



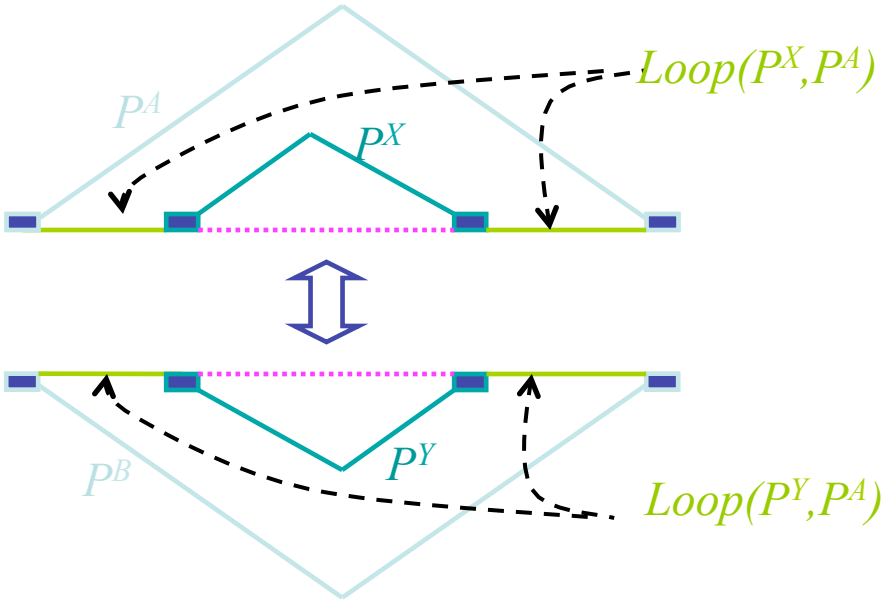
$$M_s[P^A, P^B] = w_1 \max \left\{ \begin{array}{l} \mathcal{E}_s(P^A), \\ \mathcal{E}_s(P^B) \end{array} \right\} + w_2 \mathcal{S} \left(\begin{array}{l} \mathrm{Seq}(P^A) \\ \mathrm{Seq}(P^B) \end{array} \right)$$

The score of matching hairpin loops:



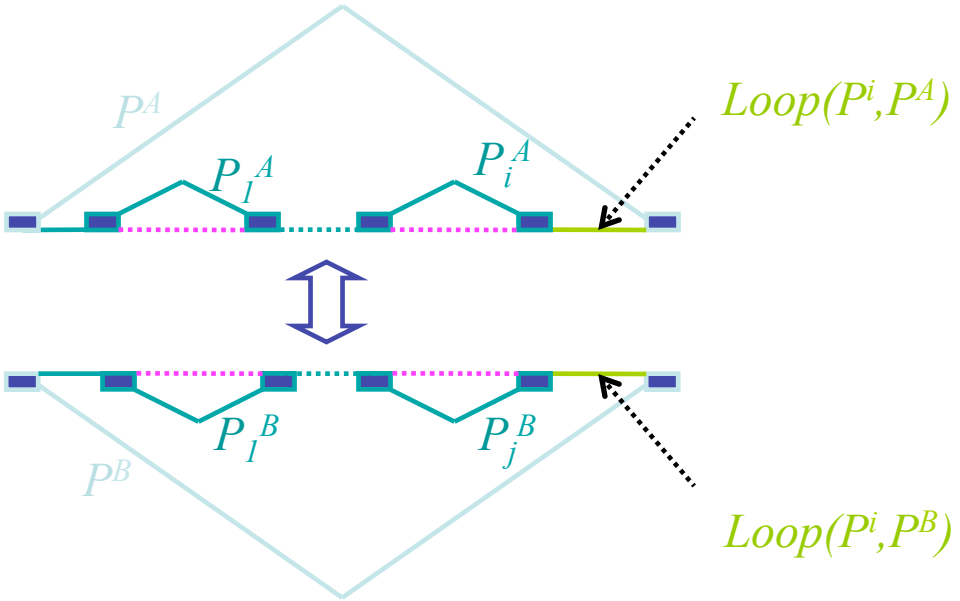
$$M_h[P^A, P^B] = w_1 \max \left\{ \begin{array}{l} \mathcal{E}_h(|\text{Loop}(P^A)|), \\ \mathcal{E}_h(|\text{Loop}(P^B)|) \end{array} \right\} + w_3 \mathcal{S} \left(\begin{array}{l} \text{Loop}(P^A), \\ \text{Loop}(P^B) \end{array} \right)$$

The score of matching interior loops or bulges:



$$M_b[P^A, P^B] = \min_{\substack{P^x <_I P^A, \\ P^y <_I P^B}} \left\{ w_1 \max \left\{ \begin{array}{l} \mathcal{E}_b(|\mathrm{Loop}(P^x, P^A)|), \\ \mathcal{E}_b(|\mathrm{Loop}(P^y, P^B)|) \end{array} \right\} + w_3 \mathcal{S} \left(\begin{array}{l} \mathrm{Loop}(P^x, P^A), \\ \mathrm{Loop}(P^y, P^B) \end{array} \right) + M[P^x, P^y] \right\}$$

The score of matching two multi-loops:



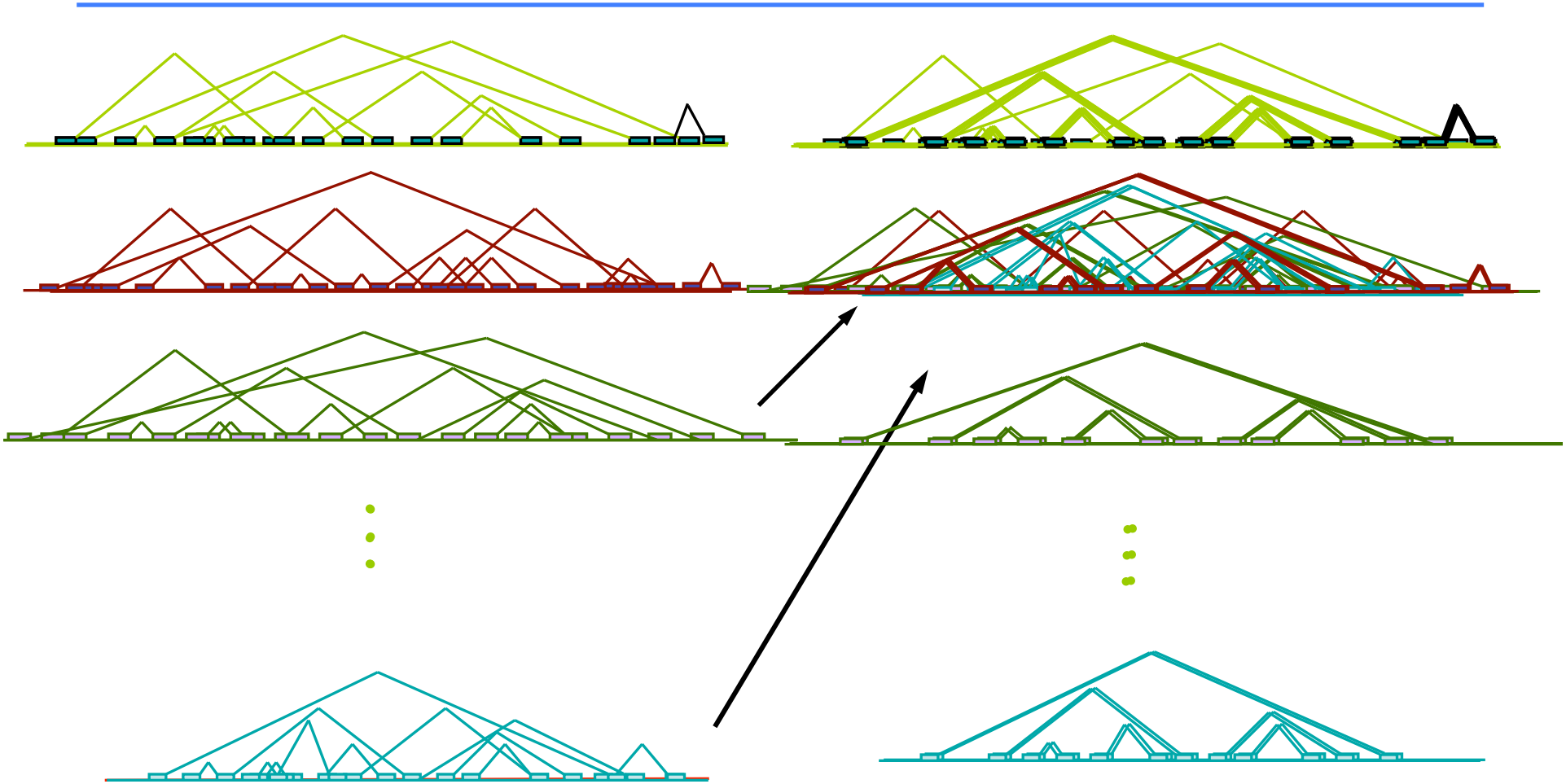
$$M_m[P^A, P^B] = \min_{i,j} \left\{ M_c[P_i^A, P_j^A] + w_1 \max \left\{ \begin{array}{l} \mathcal{E}_m(|\text{Loop}(P^i, P^A)|), \\ \mathcal{E}_m(|\text{Loop}(P^j, P^B)|) \end{array} \right\} + w_3 \mathcal{S} \left(\begin{array}{l} \text{Loop}(P^i, P^A), \\ \text{Loop}(P^j, P^B) \end{array} \right) \right\}$$

Consensus folding for multiple sequences

$$M(\mathcal{P}(A_1,\ldots,A_s)) = w_1\Phi(\mathcal{P}(A_1,\ldots,A_s)) + w_2\sum_{j\in\mathcal{P}}\mathcal{S}\begin{pmatrix}A_{1,j},\\A_{2,j},\\\ldots,\\A_{s,j}\end{pmatrix} + w_3\sum_{j\notin\mathcal{P}}\mathcal{S}\begin{pmatrix}A_{1,j},\\A_{2,j},\\\ldots,\\A_{s,j}\end{pmatrix}$$

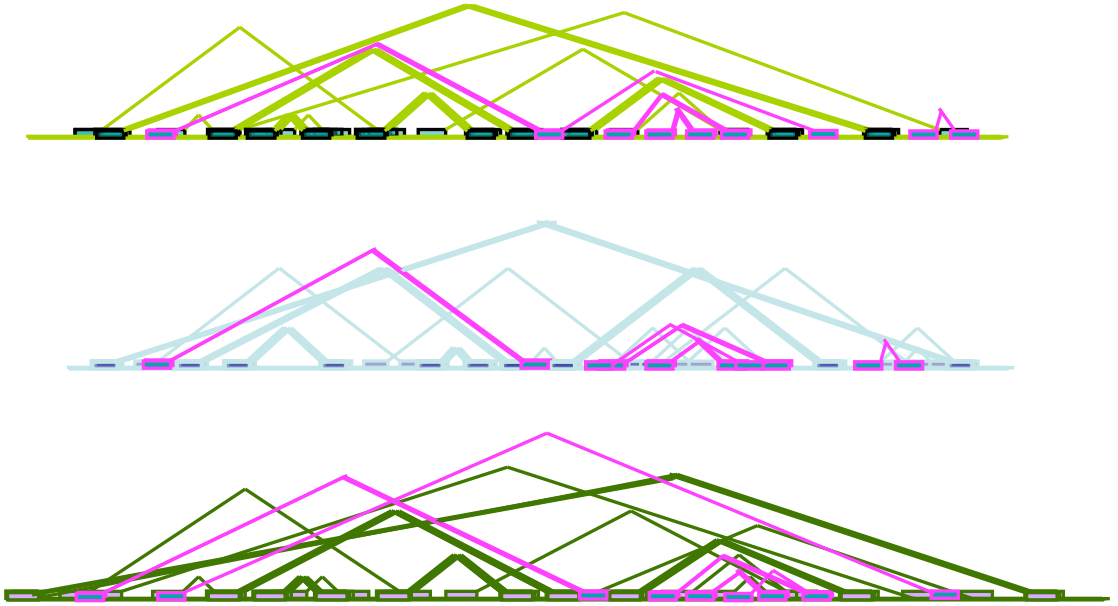
- We use a heuristic method based on the notion of star-alignment.
  - Compute an optimal configuration from a random seed pair.
  - Align all individual sequences to this configuration.
  - Choose the conserved stack configuration in all sequences.
  - Allow some stacks to be partially conserved (at least appear in a certain fraction of the sequences).

Compute the stack configuration for multiple sequences: **RNAscf(k,h,f)**



Iterative procedure for **RNAscf**

1. P = RNAscf(k, h, f).
2. In each sequence, extract the unpaired regions according to the loop regions in P.
3. Predict additional putative stacks that are not crossing with P using smaller k’ and h’.
4. Recompute the alignment for with additional putative stacks using RNAscf(k’,h’,f).



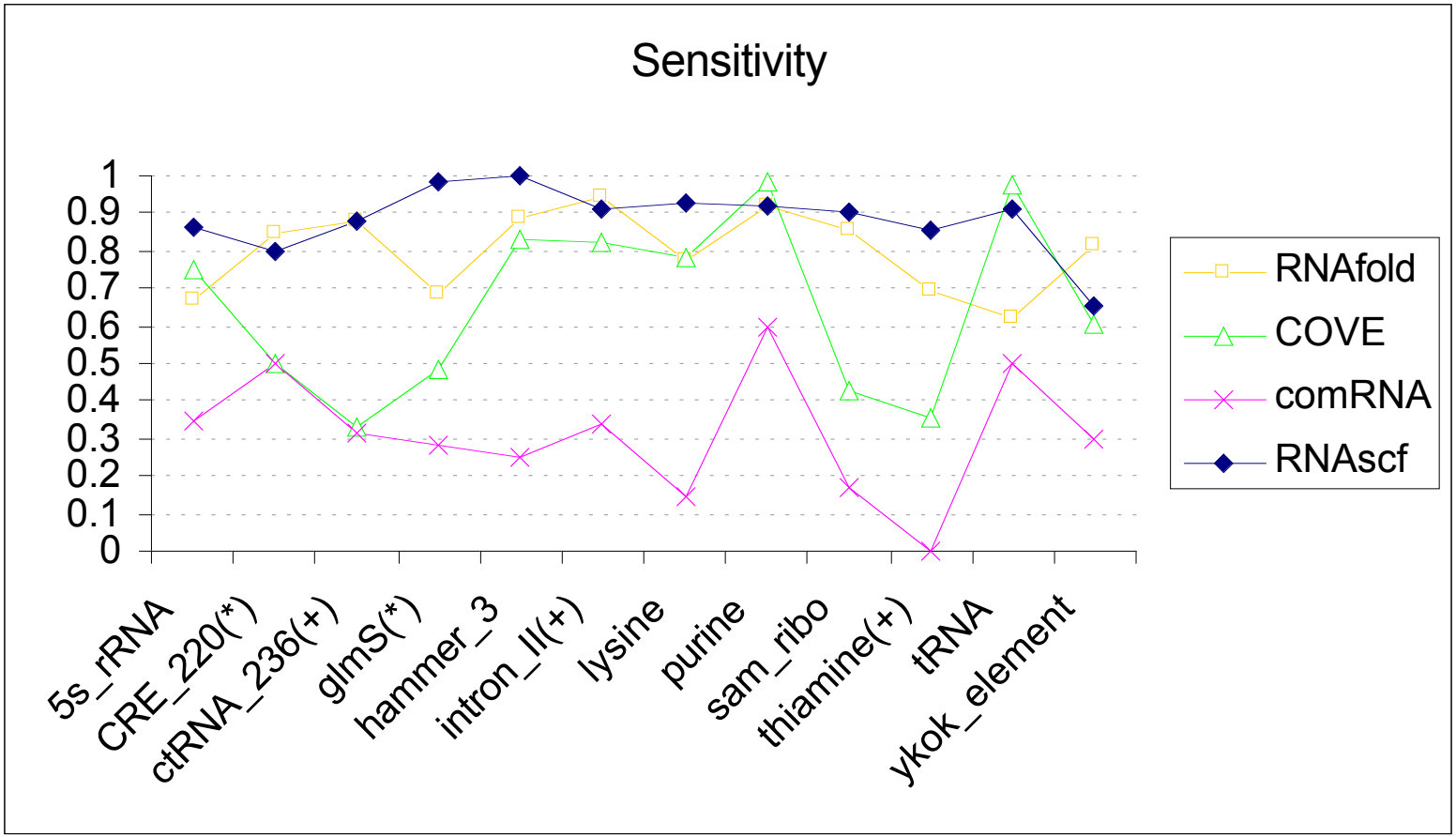
Test dataset

- We choose a set of 12 RNA families from Rfam database:
  - 20 sequences chosen from the families. (except for CRE and glms, we choose 10 sequences) with annotated structures.
  - There are 953 stacks.
  - We compare **RNAscf** with 3 other programs that are available online for RNA folding:
    - **RNAfold** (energy based minimization) (Hofacker 2003)
    - **COVE** (covariance model) (Eddy and Durbin 1994)
      - Cove need a staring seed alignment which is produced by **ClustalW**.
    - **comRNA** (computing anchors in multiple sequences) (Ji, Xu and Stormo 2004).
  - *Sensitivity*: the fraction of true stacks that overlapped with predicted stacks.
  - *Accuracy*: the fraction of predicted stacks that overlapped with true stacks

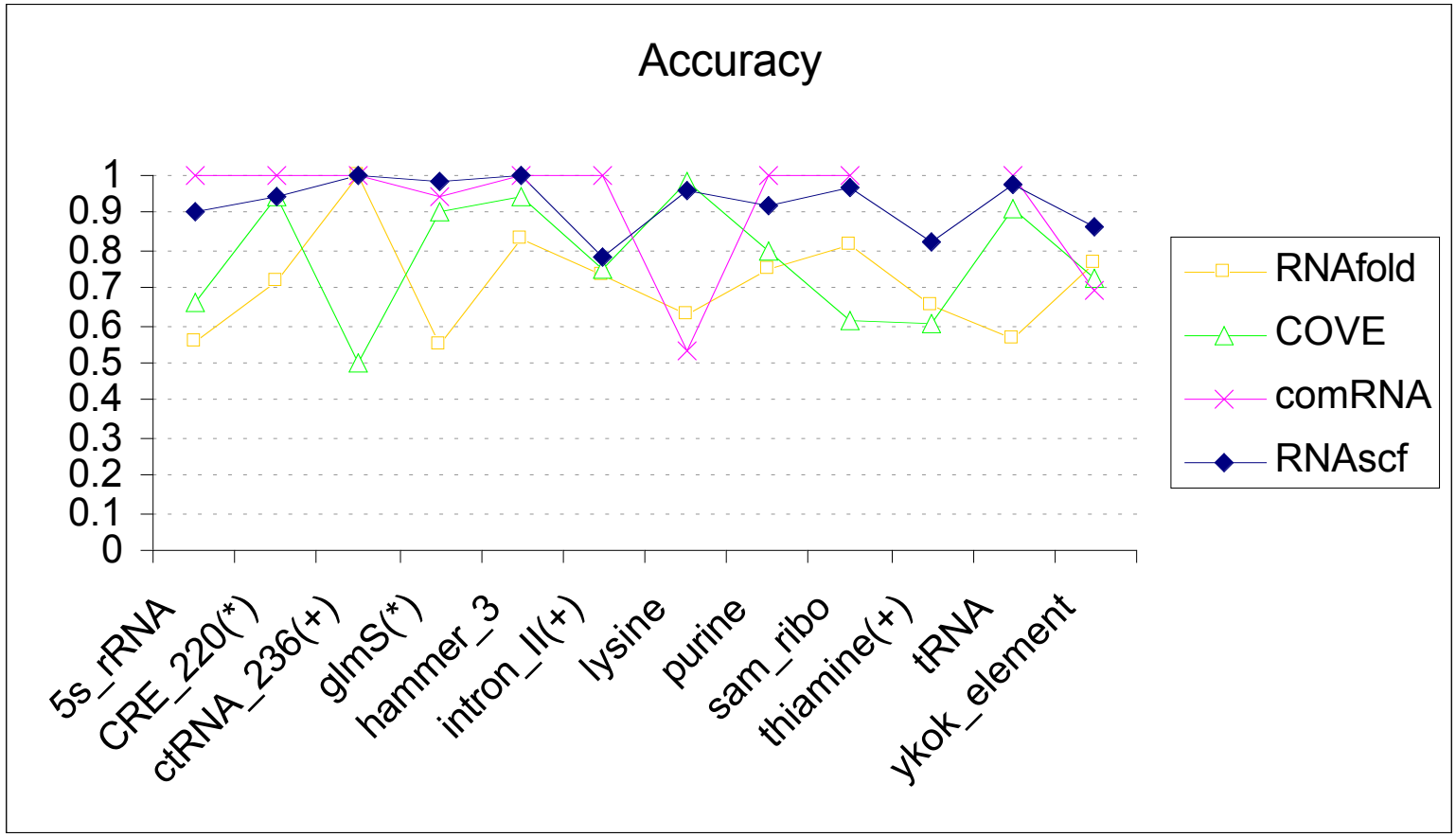
Test results

Name (Rfam_id)	Stacks	Sensitivity				Accuracy			
		RNAscf	RNAfold	Cove	comRNA	RNAscf	RNAfold	Cove	comRNA
5s_rRNA (RF00001)	100	0.86	0.67	0.75	0.35	0.9	0.558	0.658	1
Rhino_CRE (RF00220)	20	0.8	0.85	0.5	0.5	0.941	0.72	0.941	1
ctRNA_pGA1 (RF00236)	51	0.882	0.882	0.333	0.313	1	1	0.5	1
glmS(RF00234)	60	0.983	0.683	0.483	0.283	0.983	0.552	0.906	0.944
Hammerhead_3 (RF00008)	60	1	0.883	0.833	0.25	1	0.828	0.943	1
Intron_gpII (RF00029)	56	0.91	0.946	0.821	0.339	0.782	0.731	0.754	1
Lysine (RF00168)	120	0.925	0.775	0.783	0.142	0.958	0.633	0.984	0.531
Purine (RF00167)	60	0.917	0.917	0.983	0.6	0.917	0.753	0.797	1
Sam_riboswitch (RF00162)	113	0.903	0.858	0.425	0.168	0.966	0.813	0.613	1
Thiamine (RF00059)	80	0.858	0.690	0.354	0	0.824	0.654	0.606	-
tRNA (RF00005)	113	0.912	0.625	0.975	0.5	0.973	0.567	0.910	1
ykok (RF00380)	180	0.656	0.817	0.606	0.3	0.863	0.762	0.727	0.692
Average		**0.884**	0.8	0.654	0.312	**0.926**	0.714	0.778	0.924

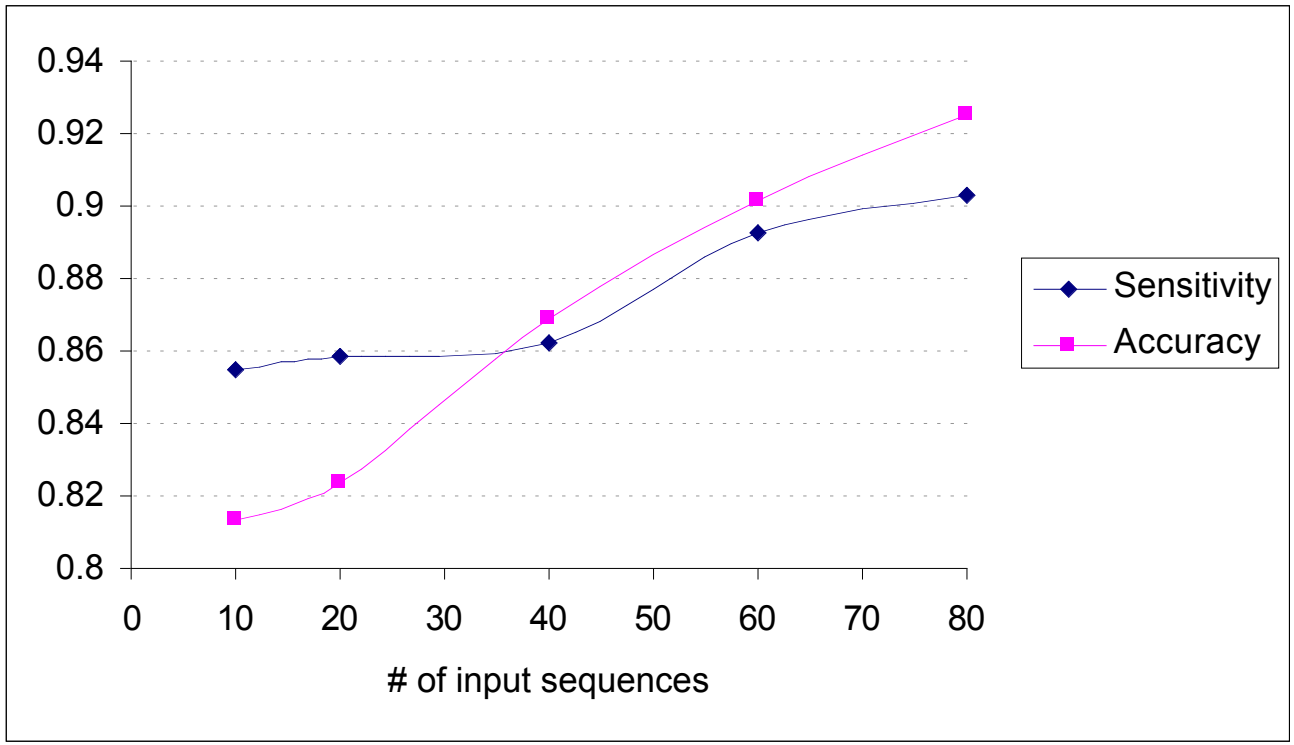
Test results



Test results

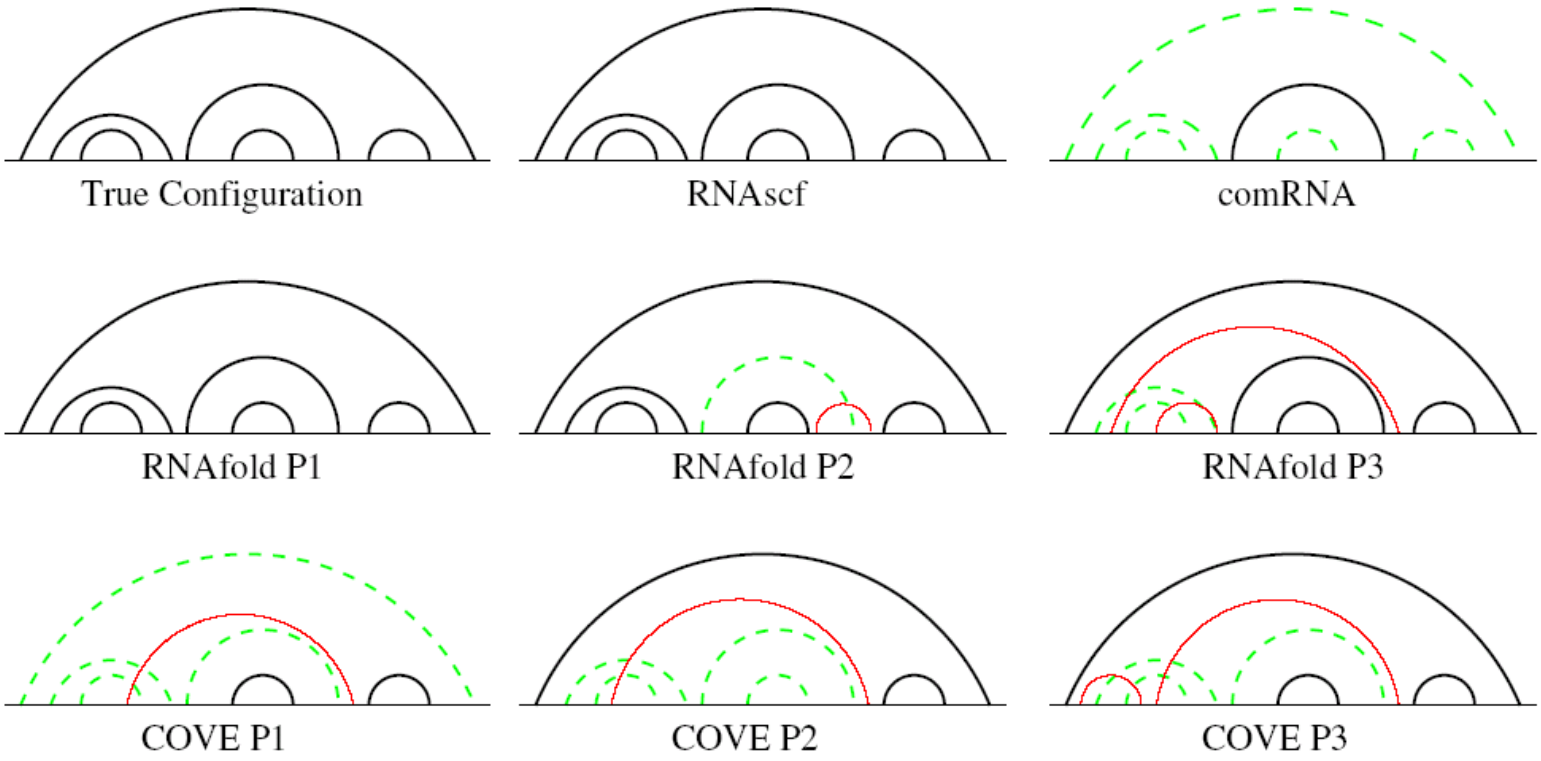


Performance improves when the number of sequences increases



(Using Thiamine riboswitch subfamily (RF00059))

RNAscf always finds the right consensus stack configuration.



(Sam riboswitch (RF00162))

Conclusion and future work

- **RNAscf** is a valid approach to RNA consensus structure prediction.
  - Use stack configuration to represent RNA secondary structure.
  - Propose a dynamic programming algorithm to find optimal stack configuration for pairwise sequences.
  - Use both primary sequence information and energy information.
  - Use a star-alignment-like heuristic method to get the consensus structure for multiple sequences.
- Future work:
  - Correcting errors by using a stochastic iterative scheme (such as Gibbs sampling).
  - Provide P-value for each prediction.
  - Use RNAscf to find new families of ncRNAs.
  - Perform constraint folding to refine the predicted structure by adding some minor basepairs.