

A Common Framework for Linear and Cyclic Multiple Sequence Alignment

Sebastian Will and Peter Stadler

Bioinformatics, University Leipzig

WABI 2014

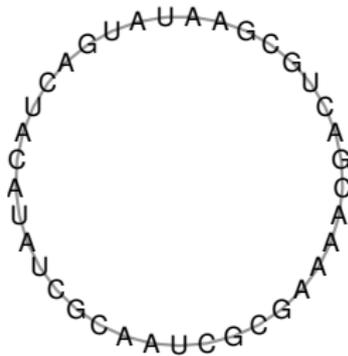
Comparing Linear and Circular RNAs

UAUGACUACAUAUCGCAAUCGCGAAAACGACUGACGUA

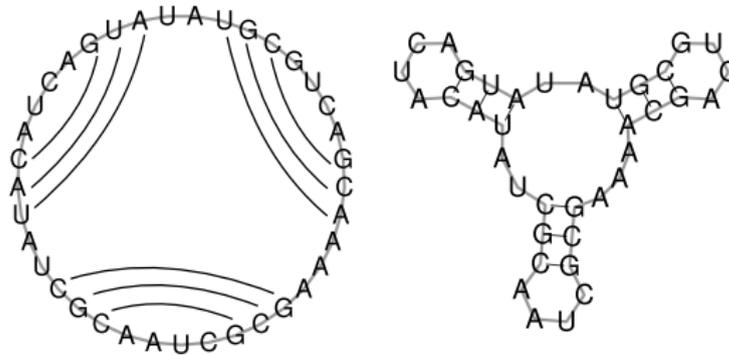
AUCACUAAAUUCGGAUUCGCGAACGACUACGGCGUA

AUACAUCUACAUAUCGCACGAGCGAAAACGACUGUA

Comparing Linear and Circular RNAs

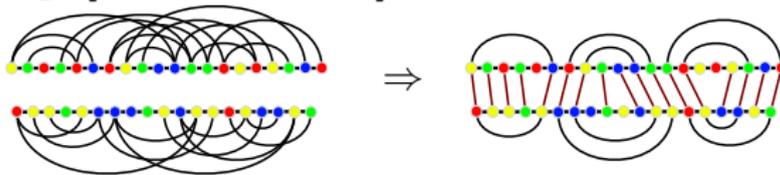


Comparing Linear and Circular RNAs



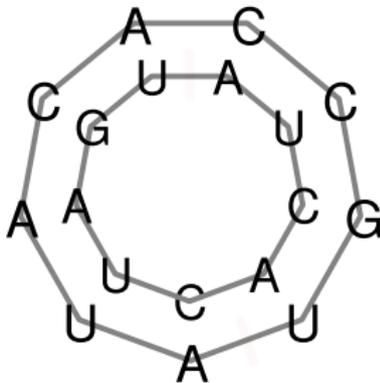
Structure-based alignment of RNAs:

e.g. [Will et al., 2007]: **two linear RNAs**



Here: **multiple circular RNAs**

(Pairwise) Cyclic Alignment (vs. Linear)



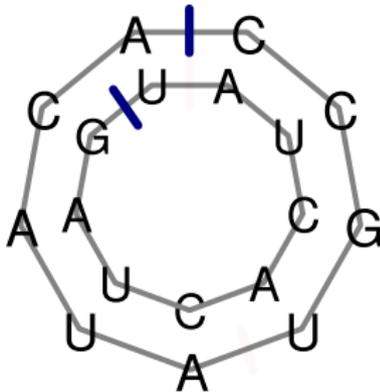
UAUCACUAG
CCGUAUACA



UAUCACUAGU
UAUCACCGU

Pairwise: Dynamic Programming + find best rotation/cut.
Pairwise cyclic sequence alignment [Mosig et al., 2006].

(Pairwise) Cyclic Alignment (vs. Linear)



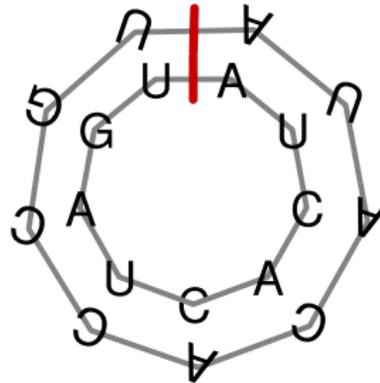
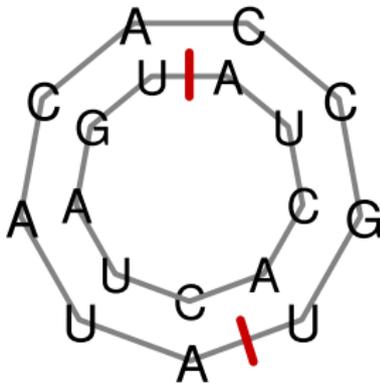
UAUCACUAG
CCGUAUACA



UAUCACUAGU
UAUCACCGU

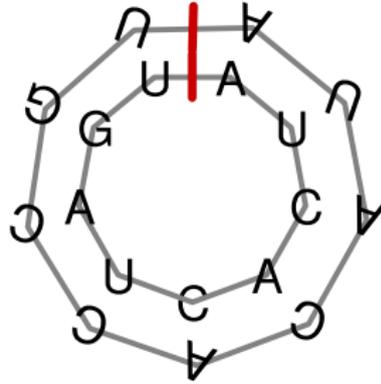
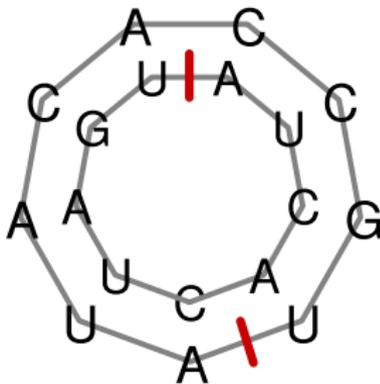
Pairwise: Dynamic Programming + find best rotation/cut.
Pairwise cyclic sequence alignment [Mosig et al., 2006].

(Pairwise) Cyclic Alignment (vs. Linear)



Pairwise: Dynamic Programming + find best rotation/cut.
 Pairwise cyclic sequence alignment [Mosig et al., 2006].

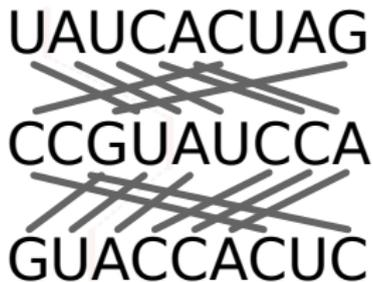
(Pairwise) Cyclic Alignment (vs. Linear)



Pairwise: Dynamic Programming + find best rotation/cut.
 Pairwise cyclic sequence alignment [Mosig et al., 2006].

Multiple Cyclic Sequence Alignment

UAUCACUAG
CCGUAUCCA
GUACCACUC



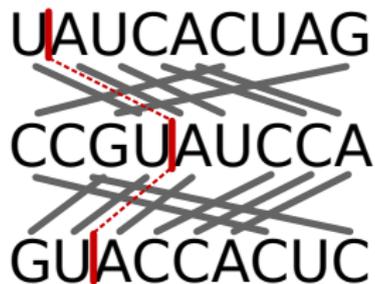
AUCACUAGU
AUCACCGU
ACCACUCGU



Two major problems at once

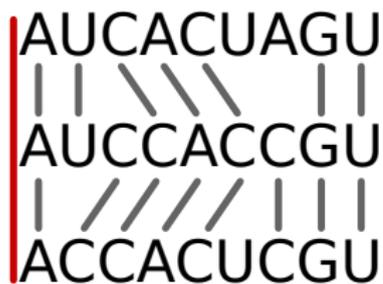
- 1) (Linear) multiple sequence alignment is NP-hard.
- 2) Search over all cuts is exponential!

Multiple Cyclic Sequence Alignment



U|AUCACUAG
CCGU|AUCCA
GU|ACCACUC

The diagram shows three RNA sequences: U|AUCACUAG, CCGU|AUCCA, and GU|ACCACUC. A red path highlights a specific alignment: U (row 1, col 1) to C (row 2, col 1) to G (row 3, col 1) to U (row 3, col 2) to A (row 2, col 2) to U (row 1, col 2) to C (row 1, col 3) to A (row 1, col 4) to U (row 1, col 5) to A (row 1, col 6) to G (row 1, col 7). Grey lines represent other possible alignments between the sequences.



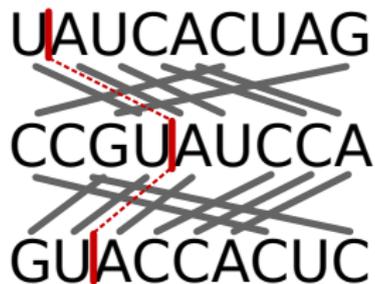
AUCACUAGU
| | \ \ \ | |
AUCCACCGU
| / / / / | | |
ACCACUCGU

The diagram shows three RNA sequences: AUCACUAGU, AUCCACCGU, and ACCACUCGU. A red vertical line is on the left. Vertical lines connect corresponding positions: A to A, U to C, C to C, A to A, C to C, U to U, A to G, and G to U. Diagonal lines connect A to C, C to U, U to C, and C to G.

Two major problems at once

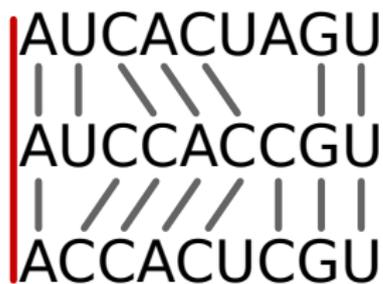
- 1) (Linear) multiple sequence alignment is NP-hard.
- 2) Search over all cuts is exponential!

Multiple Cyclic Sequence Alignment



U|AUCACUAG
CCGU|AUCCA
GU|ACCACUC

The diagram shows three sequences: U|AUCACUAG, CCGU|AUCCA, and GU|ACCACUC. A red path starts at the first 'U' in the first sequence, goes down to the 'U' in the second sequence, then down to the 'U' in the third sequence, and finally back up to the 'U' in the first sequence. Grey lines represent other possible alignments between the sequences.



AUCACUAGU
| | \ \ \ | |
AUCCACCGU
| / / / | | |
ACCACUCGU

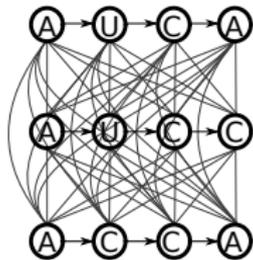
The diagram shows three sequences: AUCACUAGU, AUCCACCGU, and ACCACUCGU. A red vertical line is on the left of the first sequence. Vertical lines connect the 'A' in the first sequence to the 'A' in the second, and the 'U' in the first to the 'U' in the second. Diagonal lines connect the 'C' in the first to the 'C' in the second, and the 'A' in the first to the 'A' in the second. Vertical lines connect the 'U' in the first to the 'U' in the second, and the 'G' in the first to the 'G' in the second. Diagonal lines connect the 'U' in the first to the 'U' in the second, and the 'A' in the first to the 'A' in the second.

Two major problems at once

- 1) (Linear) multiple sequence alignment is NP-hard.
- 2) Search over all cuts is exponential!

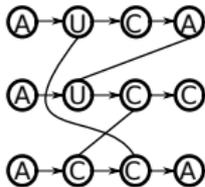
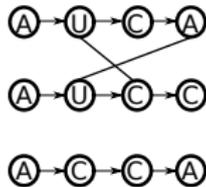
Graph-based Model: Maximum Weight Trace (for linear alignment)

Alignment graph for
input sequences
($a = \text{AUCA}$, $b = \text{AUCC}$,
and $c = \text{ACCA}$)



MWT-problem: find trace of maximum weight, where
trace := set of edges that corresponds to a valid MSA

Mixed cycle constraints

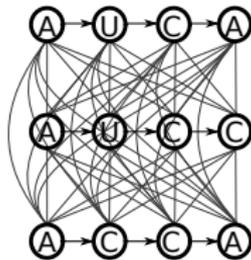


There are **exponentially many** mixed cycles!

Branch-and-cut: mixed cycles as cuts [Reinert et al., 1997]

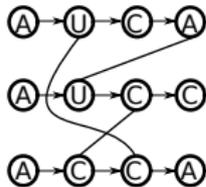
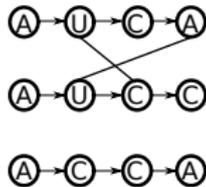
Graph-based Model: Maximum Weight Trace (for linear alignment)

Alignment graph for
input sequences
($a = \text{AUCA}$, $b = \text{AUCC}$,
and $c = \text{ACCA}$)



MWT-problem: find trace of maximum weight, where
trace := set of edges that corresponds to a valid MSA

Mixed cycle constraints

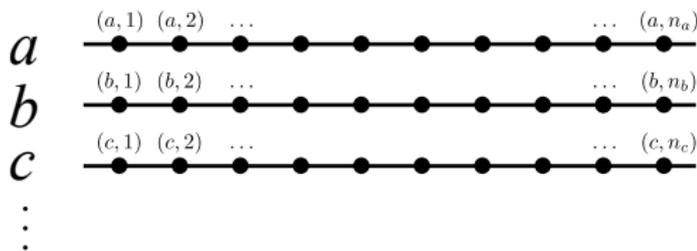


Transfer to cyclic MSA!?

There are **exponentially many** mixed cycles!

Branch-and-cut: mixed cycles as cuts [Reinert et al., 1997]

Set-theoretic Model of Linear MSA



$X :=$ set of all positions (a, k) of all sequences

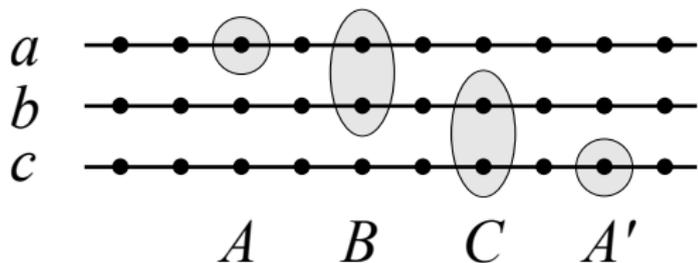
Define linear order relation on classes $A, B \subseteq X$: $A \prec B$ iff

- $A \neq B$ (irreflexive)
- $\exists (a, i) \in A, (a, j) \in B: i < j$ (ordered for at least one sequence)
- $\nexists (a, i) \in A, (a, j) \in B: i > j$ (no conflicts)

Definition [Morgenstern et al., 1999]: A partition \mathcal{A} is called *multiple sequence alignment* iff

- $\forall A \in \mathcal{A}$: at most one position per sequence,
- $\forall A \neq B \in \mathcal{A}$: $A \prec B$ or $B \prec A$ (non-crossing)
- transitive closure $\overline{\prec}$ of \prec : partial order on \mathcal{A} .

Set-theoretic Model of Linear MSA



$X :=$ set of all positions (a, k) of all sequences

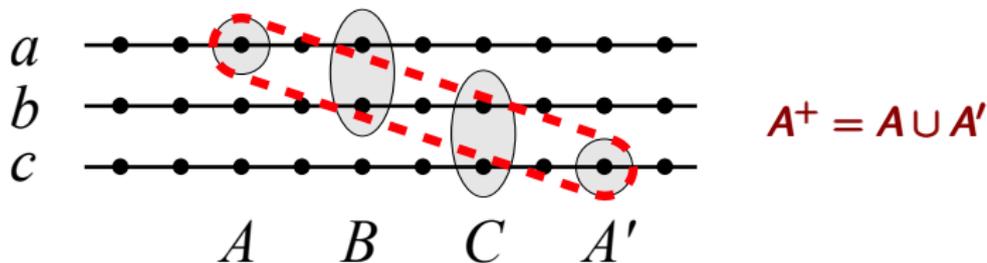
Define linear order relation on classes $A, B \subseteq X$: $A \prec B$ iff

- $A \neq B$ (irreflexive)
- $\exists (a, i) \in A, (a, j) \in B: i < j$ (ordered for at least one sequence)
- $\nexists (a, i) \in A, (a, j) \in B: i > j$ (no conflicts)

Definition [Morgenstern et al., 1999]: A partition \mathcal{A} is called *multiple sequence alignment* iff

- $\forall A \in \mathcal{A}$: at most one position per sequence,
- $\forall A \neq B \in \mathcal{A}$: $A \prec B$ or $B \prec A$ (non-crossing)
- transitive closure $\overline{\prec}$ of \prec : partial order on \mathcal{A} .

Set-theoretic Model of Linear MSA



$X :=$ set of all positions (a, k) of all sequences

Define linear order relation on classes $A, B \subseteq X$: $A \prec B$ iff

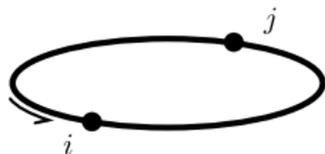
- $A \neq B$ (irreflexive)
- $\exists (a, i) \in A, (a, j) \in B: i < j$ (ordered for at least one sequence)
- $\nexists (a, i) \in A, (a, j) \in B: i > j$ (no conflicts)

Definition [Morgenstern et al., 1999]: A partition \mathcal{A} is called *multiple sequence alignment* iff

- $\forall A \in \mathcal{A}$: at most one position per sequence,
- $\forall A \neq B \in \mathcal{A}$: $A \prec B$ or $B \prec A$ (non-crossing)
- transitive closure $\overline{\prec}$ of \prec : partial order on \mathcal{A} .

Towards a Set-theoretic Model for Cyclic MSA

Key idea: use cyclic order (in place of linear order)

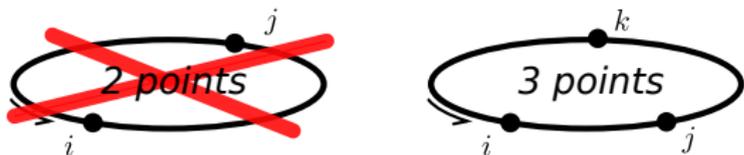


The ternary relation \triangleleft is a *cyclic order* iff

- $\triangleleft ijk$ implies i, j, k pairwise distinct (irreflexive)
- $\triangleleft ijk$ implies $\triangleleft kji$ (cyclic)
- $\triangleleft ijk$ implies $\neg \triangleleft kji$ (antisymmetric)
- $\triangleleft ijk$ and $\triangleleft ikl$ implies $\triangleleft ijkl$ (transitive)
- If i, j, k are pairwise distinct then $\triangleleft ijk$ or $\triangleleft kji$ (total)

Towards a Set-theoretic Model for Cyclic MSA

Key idea: use cyclic order (in place of linear order)



The ternary relation \triangleleft is a *cyclic order* iff

- $\triangleleft ijk$ implies i, j, k pairwise distinct (irreflexive)
- $\triangleleft ijk$ implies $\triangleleft kij$ (cyclic)
- $\triangleleft ijk$ implies $\neg \triangleleft kji$ (antisymmetric)
- $\triangleleft ijk$ and $\triangleleft ikj$ implies $\triangleleft ijkl$ (transitive)
- If i, j, k are pairwise distinct then $\triangleleft ijk$ or $\triangleleft kji$ (total)

Towards a Set-theoretic Model for Cyclic MSA

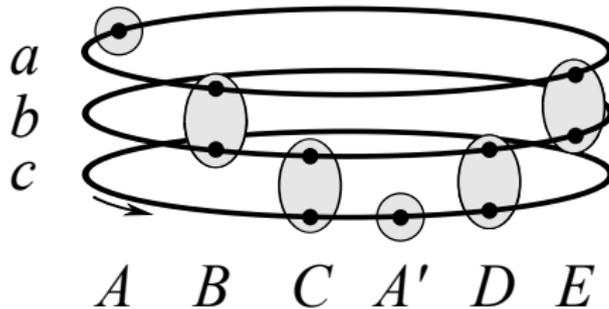
Key idea: use cyclic order (in place of linear order)



The ternary relation \triangleleft is a *cyclic order* iff

- $\triangleleft ijk$ implies i, j, k pairwise distinct (**irreflexive**)
- $\triangleleft ijk$ implies $\triangleleft kji$ (**cyclic**)
- $\triangleleft ijk$ implies $\neg \triangleleft kji$ (**antisymmetric**)
- $\triangleleft ijk$ and $\triangleleft ikl$ implies $\triangleleft ij l$ (**transitive**)
- If i, j, k are pairwise distinct then $\triangleleft ijk$ or $\triangleleft kji$ (**total**)

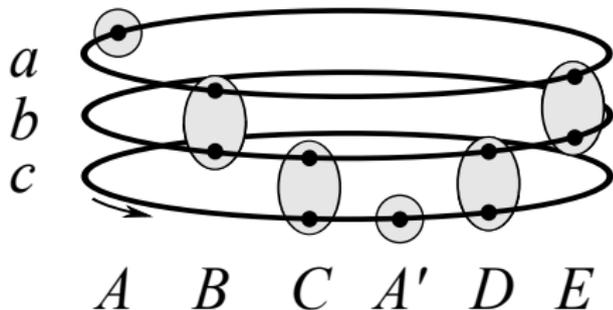
Towards a Set-theoretic Model for Cyclic MSA



Define relation \triangleleft on classes (of X); $\triangleleft ABC$ iff

- A , B , and C are pairwise distinct (**irreflexive**)
- $\exists (a, i) \in A, (a, j) \in B, (a, k) \in C: \triangleleft ijk$ (**comparable**)
- $\nexists (a, i) \in A, (a, j) \in B, (a, k) \in C: \neg \triangleleft ijk$ (**no conflicts**)

Set-theoretic Cyclic MSA Model

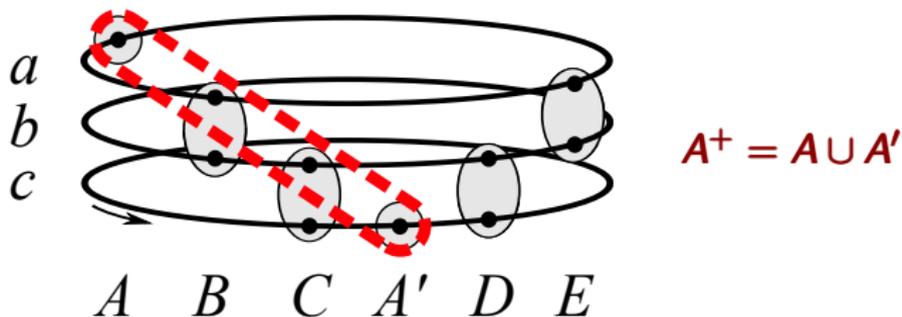


Definition (Cyclic MSA)

A cyclic MSA is a partition \mathcal{A} of X iff

- $\forall A \in \mathcal{A}$: at most one position per sequence,
- for all $A, B, C \in \mathcal{A}$: A, B, C are cyclically non-crossing
- The transitive closure $\overleftarrow{\triangleleft}$ of \triangleleft is a partial cyclic order of \mathcal{A} .

Set-theoretic Cyclic MSA Model



Definition (Cyclic MSA)

A cyclic MSA is a partition \mathcal{A} of X iff

- $\forall A \in \mathcal{A}$: at most one position per sequence,
- for all $A, B, C \in \mathcal{A}$: A, B, C are cyclically non-crossing
- The transitive closure $\overline{\triangleleft}$ of \triangleleft is a partial cyclic order of \mathcal{A} .

Formal Linear MSA Model \Rightarrow ILP Model

Model the linear MSA \mathcal{A} by Boolean variables and linear inequations and maximize alignment score.

Variables:

- $\mathbf{P}_{x\alpha} = 1$ iff $x = (a, i)$ is in class α of the partition \mathcal{A}
- $\mathbf{O}_{\alpha\beta} = 1$ iff $\alpha \overline{\succ} \beta$
- ... (further variables for objective function: base matches: \mathbf{E} , affine gap cost: \mathbf{G} , \mathbf{GO} , RNA structure matches \mathbf{B})

Integer Linear Program (ILP):

max alignment-score(\mathbf{E} , \mathbf{G} , \mathbf{GO} , \mathbf{B})

s.t.

- Variables $\mathbf{P}_{x\alpha}$ represent a partition of \mathcal{A}
- Variables $\mathbf{O}_{\alpha\beta}$ represent $\overline{\succ}$
- $\mathbf{O}_{\alpha\beta}$ describe a partial order
- ... (constrain the variables of the objective function)

Formal **Cyclic MSA Model** \Rightarrow **ILP Model**

Model the cyclic MSA \mathcal{A} by Boolean variables and linear inequations and maximize alignment score.

Variables:

- $\mathbf{P}_{x\alpha} = 1$ iff $x = (a, i)$ is in class α of the partition \mathcal{A}
- ~~$\mathbf{O}_{\alpha\beta} = 1$ iff $\alpha \succ \beta$~~ **$\mathbf{O}_{\alpha\beta\gamma} = 1$ iff $\triangleleft \alpha\beta\gamma$**
- ... (further variables for objective function: base matches: **E**, affine gap cost: **G**, **GO**, RNA structure matches **B**)

Integer Linear Program (ILP):

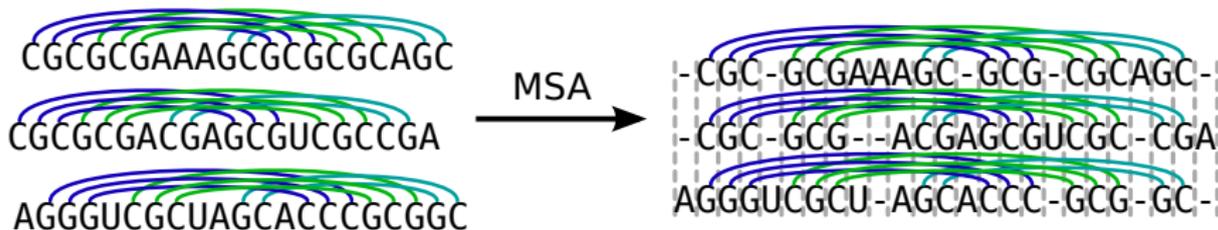
max alignment-score(**E**, **G**, **GO**, **B**)

s.t.

- Variables $\mathbf{P}_{x\alpha}$ represent a partition of \mathcal{A}
- Variables ~~$\mathbf{O}_{\alpha\beta}$~~ represent ~~\succ~~ **$\mathbf{O}_{\alpha\beta\gamma}$ represent \triangleleft**
- ~~$\mathbf{O}_{\alpha\beta}$~~ **$\mathbf{O}_{\alpha\beta\gamma}$** describe a partial **cyclic** order
- ... (constrain the variables of the objective function)

Preliminary Results with CPLEX Solver

#seqs	Instance		Model	Solving Time (s)	
	length	structure		95% opt.	optimal
3	10	2-knot	linear	1.4	1.4
3	10	2-knot	cyclic	170	176
3	10	2-knot	cyclic	229	273
3	15	3-knot	linear	129	143
3	20	3-knot	linear	287	> 300
3	20	3-knot	linear $\Delta 3$	8.4	8.4
4	10	2-knot	linear	10	28
4	10	2-knot	linear $\Delta 3$	4.8	6.4



Conclusions

- Systematic analysis of cyclic MSA
 - Framework for linear and cyclic MSA; single difference: **order**
 - Only polynomially many constraints (graph-based: exponential)
 - Model is flexibly extensible:
structure-based alignment of circular RNAs
 - Current ILP model: CPLEX solves only small instances
 - Future work:
 - are other solvers more suitable?
 - “tricks” like variable reduction
 - transfer “critical mixed cycles” from linear to cyclic MSA (some reassuring theoretical results given in the paper)
- ⇒ branch-and-cut or Lagrange relaxation for cyclic MSA

Conclusions

- Systematic analysis of cyclic MSA
 - Framework for linear and cyclic MSA; single difference: **order**
 - Only polynomially many constraints (graph-based: exponential)
 - Model is flexibly extensible:
structure-based alignment of circular RNAs
 - Current ILP model: CPLEX solves only small instances
 - Future work:
 - are other solvers more suitable?
 - “tricks” like variable reduction
 - transfer “critical mixed cycles” from linear to cyclic MSA (some reassuring theoretical results given in the paper)
- ⇒ branch-and-cut or Lagrange relaxation for cyclic MSA

Thank you!