A Common Framework for Linear and Cyclic Multiple Sequence Alignment

Sebastian Will and Peter Stadler

Bioinformatics, University Leipzig

WABI 2014

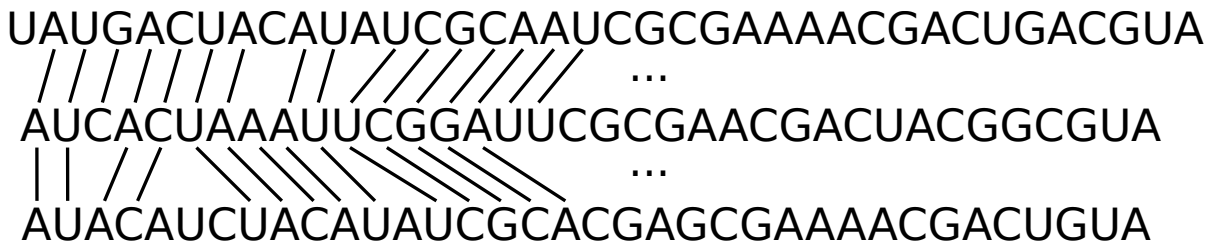
Comparing Linear and Circular RNAs

UAUGACUACAUAUCGCAAUCGCGAAAACGACUGACGUA

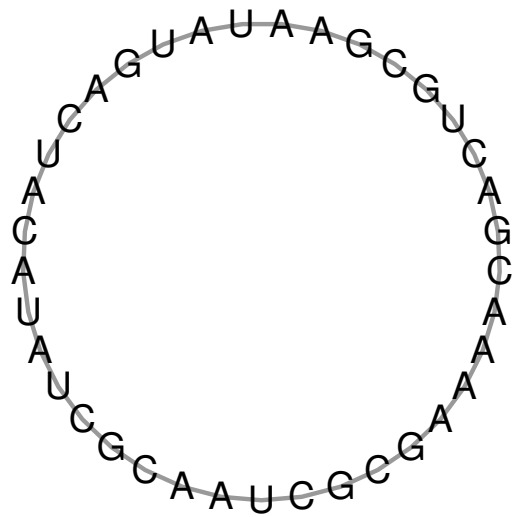
AUCACUAAAUUCGGAUUCGCGAACGACUACGGCGUA

AUACAUCUACAUAUCGCACGAGCGAAAACGACUGUA

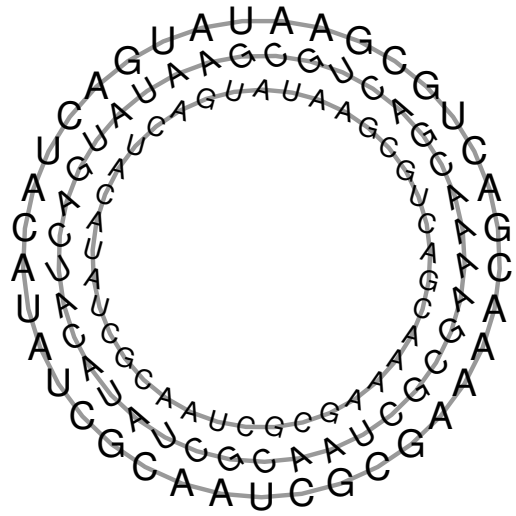
Comparing Linear and Circular RNAs



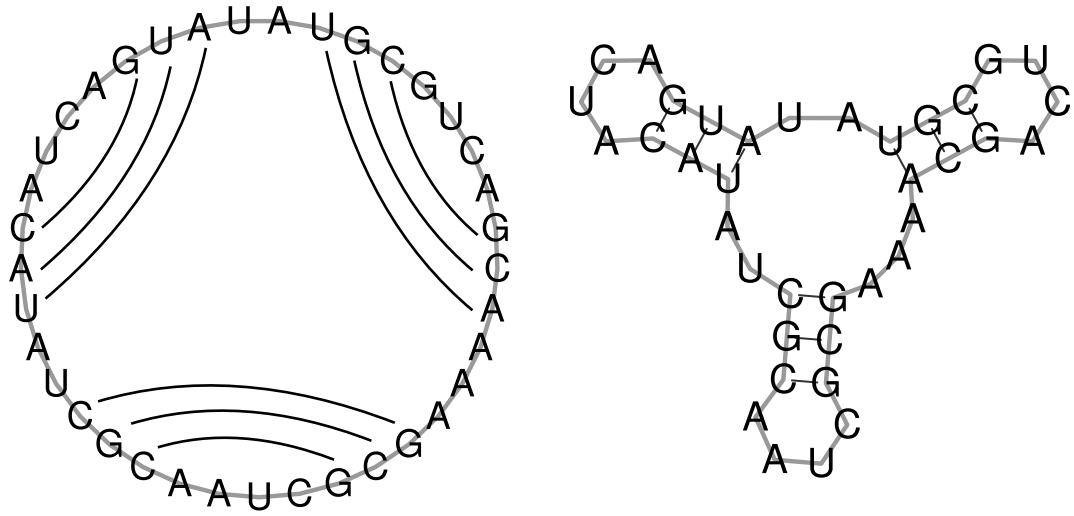
Comparing Linear and Circular RNAs



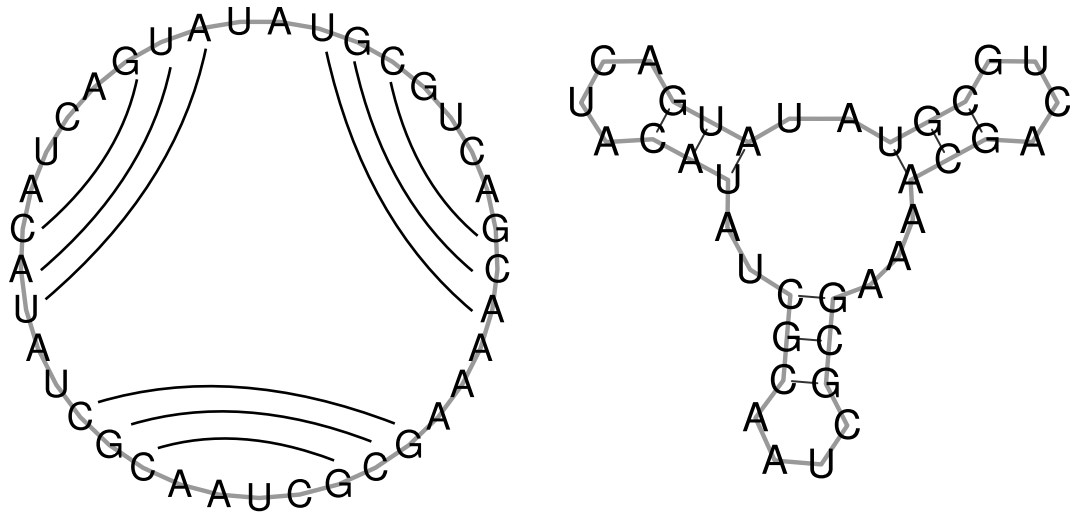
Comparing Linear and Circular RNAs



Comparing Linear and Circular RNAs

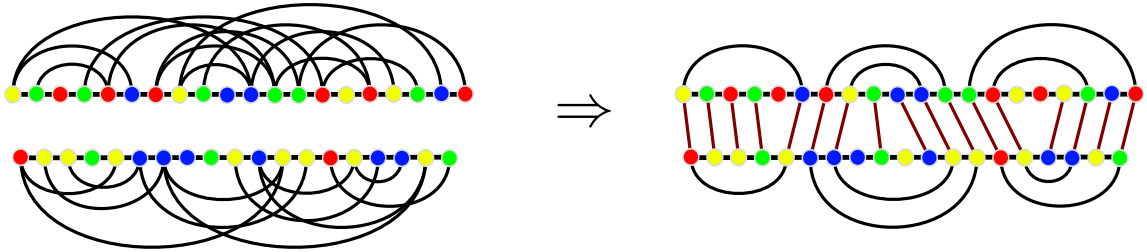


Comparing Linear and Circular RNAs



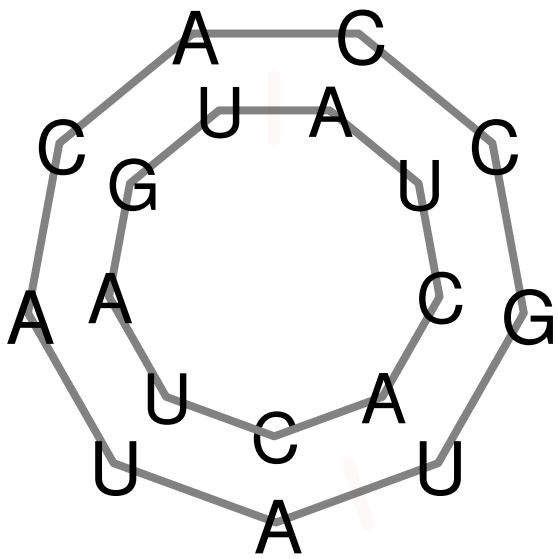
Structure-based alignment of RNAs:

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



Here: **multiple circular RNAs**

(Pairwise) Cyclic Alignment (vs. Linear)



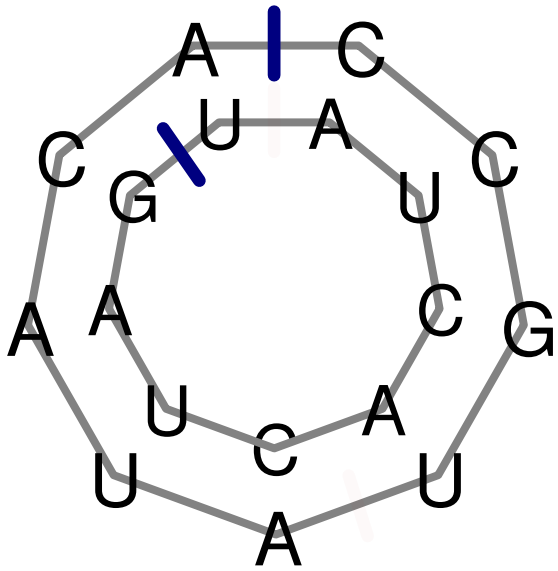
UAUCACUAG

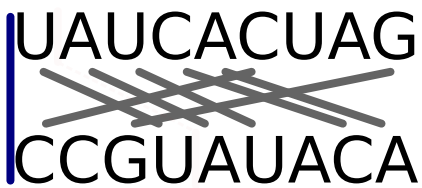
CCGUAUACA

Pairwise: Dynamic Programming + find best rotation/cut.

Pairwise cyclic sequence alignment [Mosig et al., 2006].

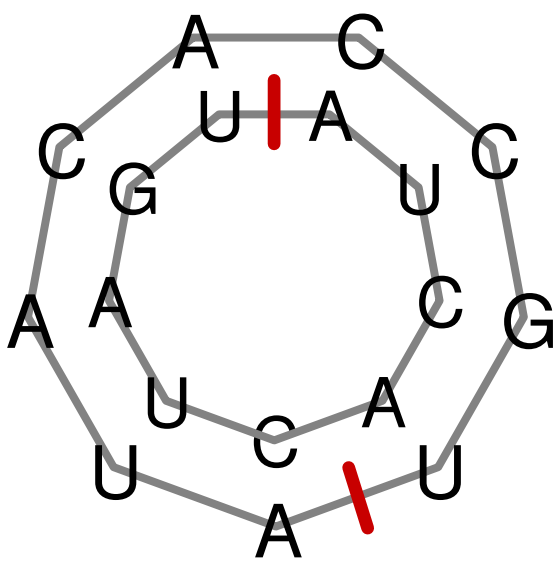
(Pairwise) Cyclic Alignment (vs. Linear)

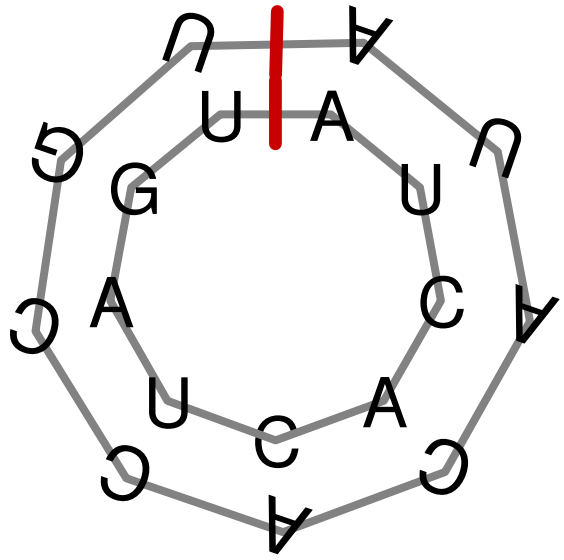


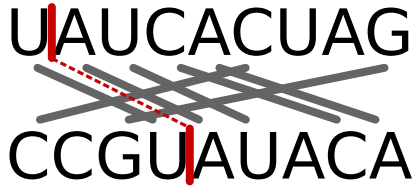


Pairwise: Dynamic Programming + find best rotation/cut.
Pairwise cyclic sequence alignment [Mosig et al., 2006].

(Pairwise) Cyclic Alignment (vs. Linear)

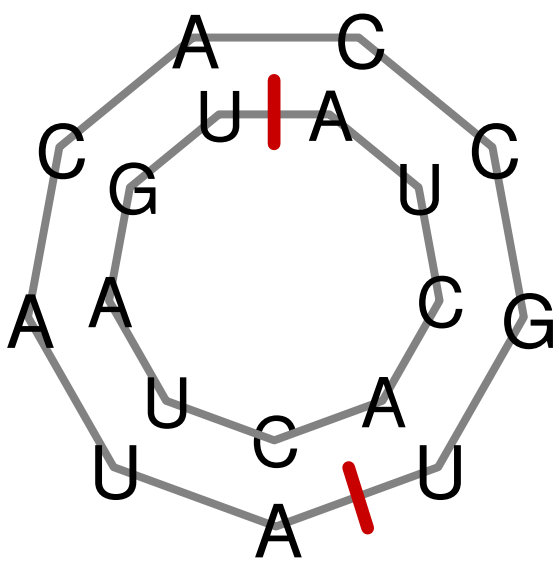


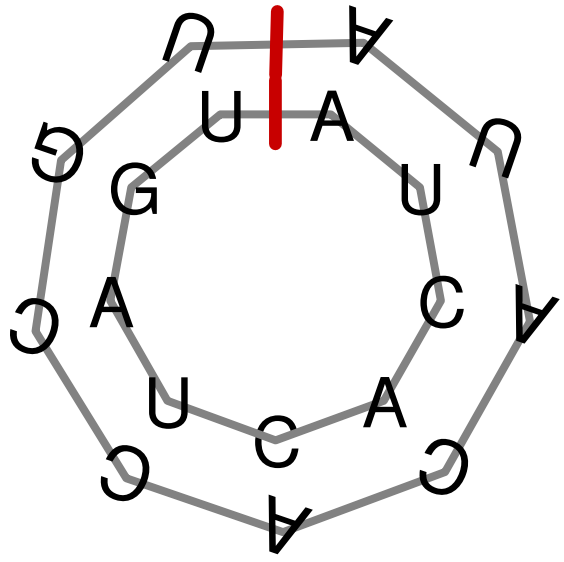


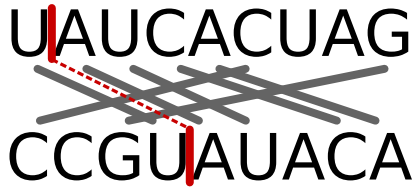


Pairwise: Dynamic Programming + find best rotation/cut.
Pairwise cyclic sequence alignment [Mosig et al., 2006].

(Pairwise) Cyclic Alignment (vs. Linear)

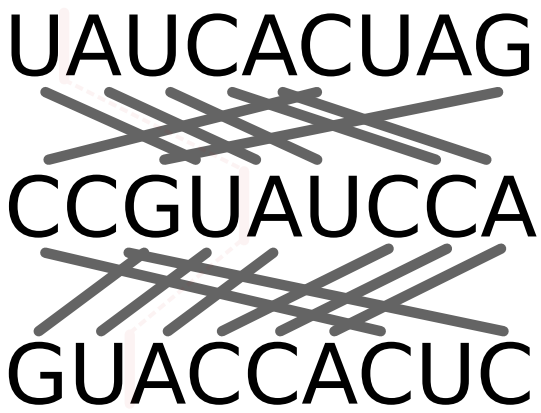






Pairwise: Dynamic Programming + find best rotation/cut.
Pairwise cyclic sequence alignment [Mosig et al., 2006].

Multiple Cyclic Sequence Alignment

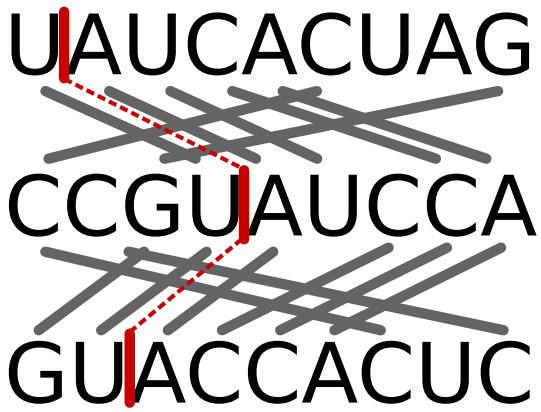


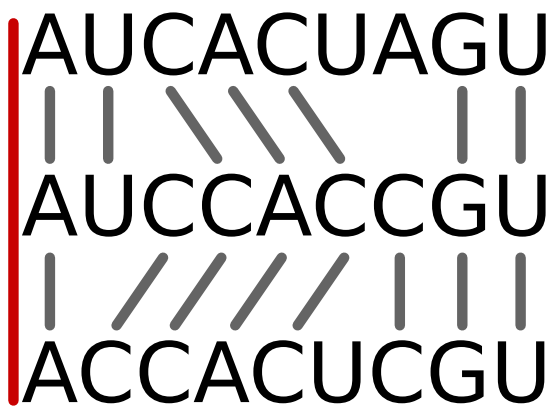
Two major problems at once

1) (Linear) multiple sequence alignment is NP-hard.

2) Search over all cuts is exponential!

Multiple Cyclic Sequence Alignment



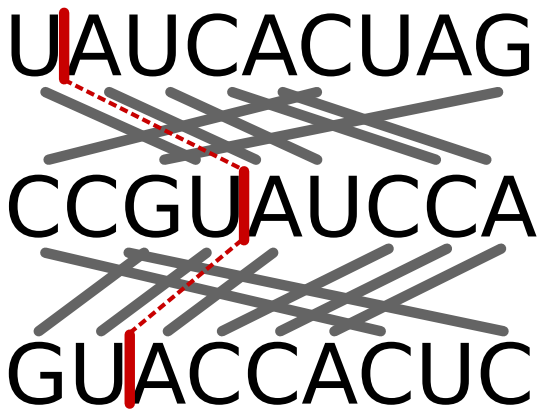


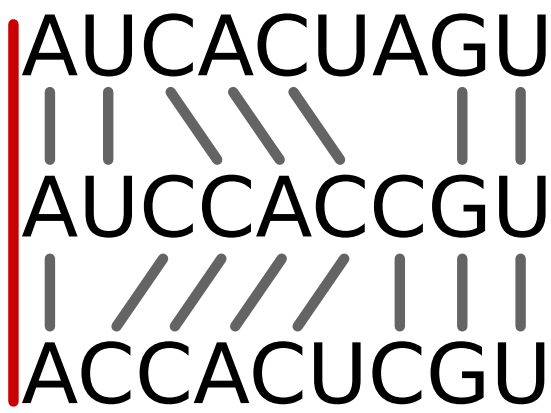
Two major problems at once

1) (Linear) multiple sequence alignment is NP-hard.

2) Search over all cuts is exponential!

Multiple Cyclic Sequence Alignment





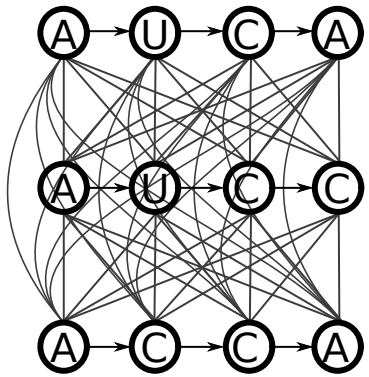
Two major problems at once

1) (Linear) multiple sequence alignment is NP-hard.

2) Search over all cuts is exponential!

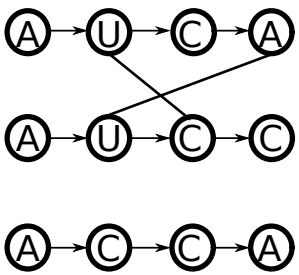
Graph-based Model: Maximum Weight Trace
(for linear alignment)

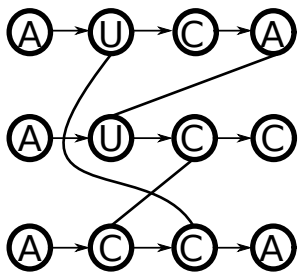
Alignment graph for input sequences (a =AUCA, b =AUCC, and c =ACCA)



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

Mixed cycle constraints



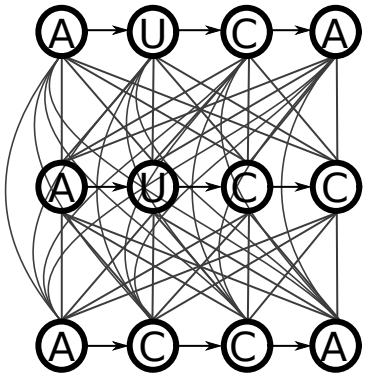


There are **exponentially many** mixed cycles!

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

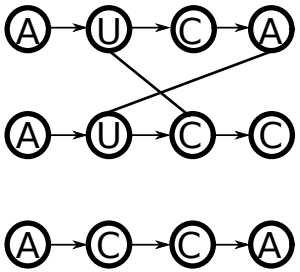
Graph-based Model: Maximum Weight Trace
(for linear alignment)

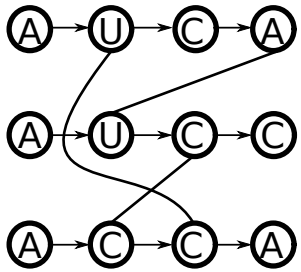
Alignment graph for input sequences (a =AUCA, b =AUCC, and c =ACCA)



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

Mixed cycle constraints



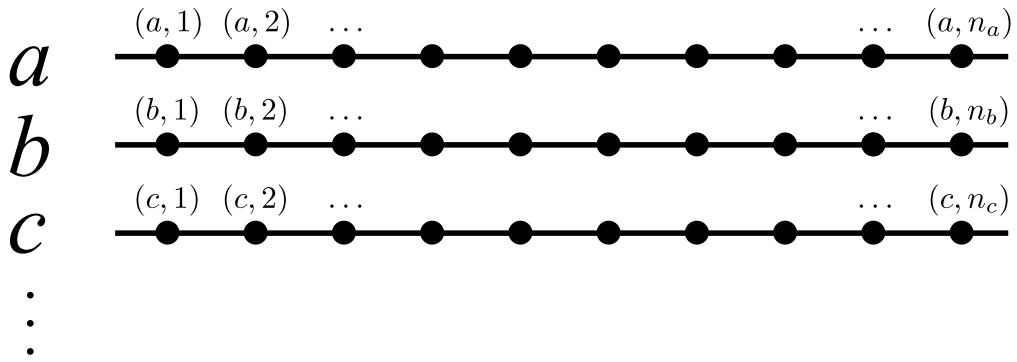


Transfer to cyclic MSA!?

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

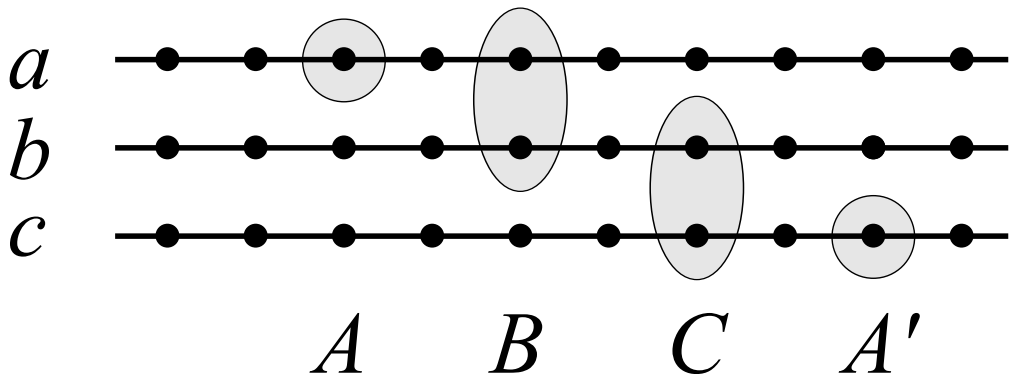
Define linear order relation on classes $A, B \subseteq X$: $A \prec B$ iff

- $A \neq B$ (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

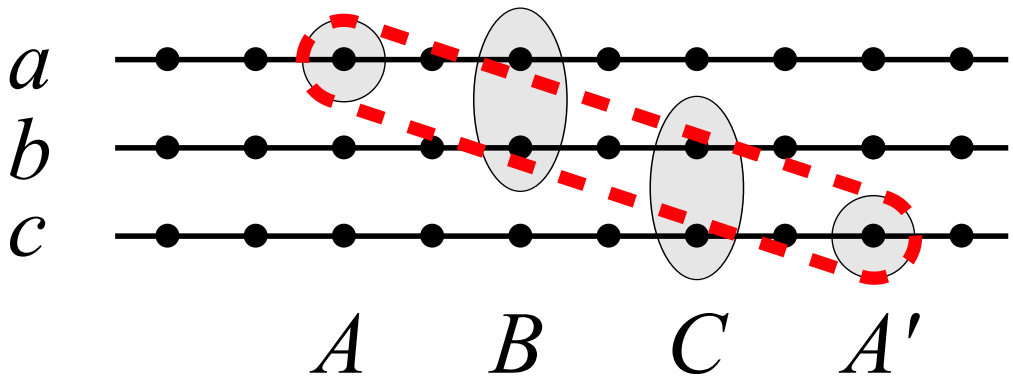
Define linear order relation on classes $A, B \subseteq X$: $A \prec B$ iff

- $A \neq B$ (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



$A^+ = A \cup A'$

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

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

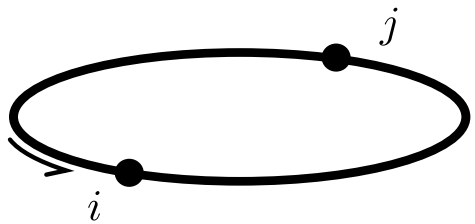
- $A \neq B$ (irreflexive)
- $\exists\, (a, i) \in A, (a, j) \in B$: $i < j$ (ordered for at least one sequence)
- $\nexists\, (a, i) \in A, (a, j) \in B$: $i > j$ (no conflicts)

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)

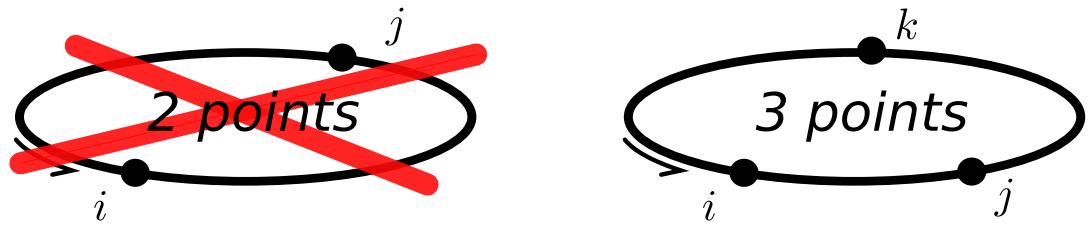


The ternary relation $\lhd$ is a *cyclic order* iff

- $\lhd i\ j\ k$ implies i, j, k pairwise distinct (irreflexive)
- $\lhd i\ j\ k$ implies $\lhd k\ i\ j$ (cyclic)
- $\lhd i\ j\ k$ implies $\neg\lhd k\ j\ i$ (antisymmetric)
- $\lhd i\ j\ k$ and $\lhd i\ k\ l$ implies $\lhd i\ j\ l$ (transitive)
- If i, j, k are pairwise distinct then $\lhd i\ j\ k$ or $\lhd k\ j\ i$ (total)

Towards a Set-theoretic Model for Cyclic MSA

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

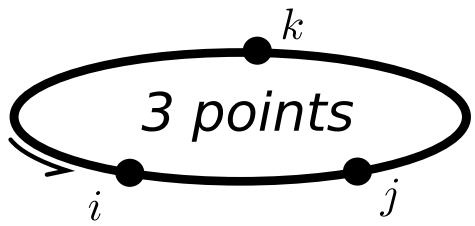


The ternary relation $\lhd$ is a *cyclic order* iff

- $\lhd i\ j\ k$ implies i, j, k pairwise distinct (irreflexive)
- $\lhd i\ j\ k$ implies $\lhd k\ i\ j$ (cyclic)
- $\lhd i\ j\ k$ implies $\neg\lhd k\ j\ i$ (antisymmetric)
- $\lhd i\ j\ k$ and $\lhd i\ k\ l$ implies $\lhd i\ j\ l$ (transitive)
- If i, j, k are pairwise distinct then $\lhd i\ j\ k$ or $\lhd k\ j\ i$ (total)

Towards a Set-theoretic Model for Cyclic MSA

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

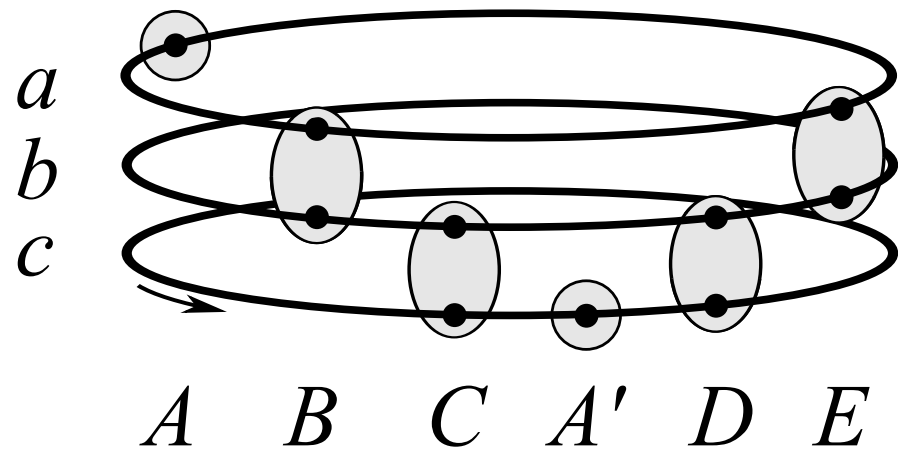


$\triangleleft\, i\, j\, k$

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

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

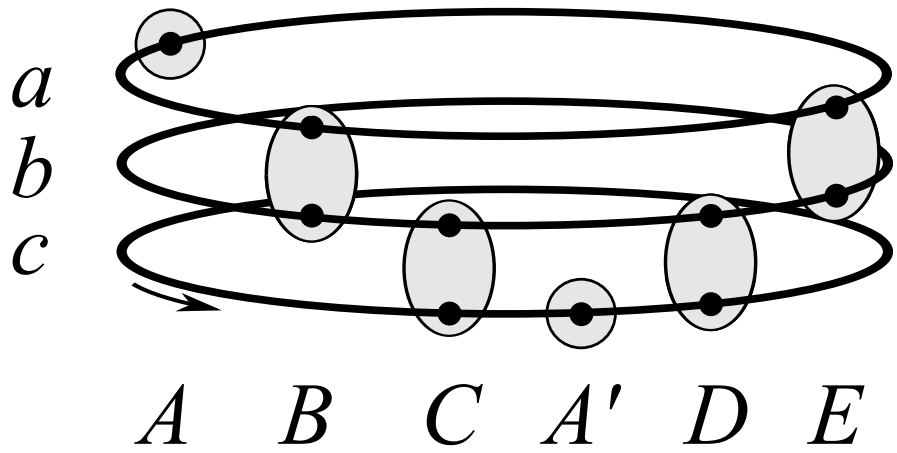
Towards a Set-theoretic Model for Cyclic MSA



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

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

Set-theoretic Cyclic MSA Model

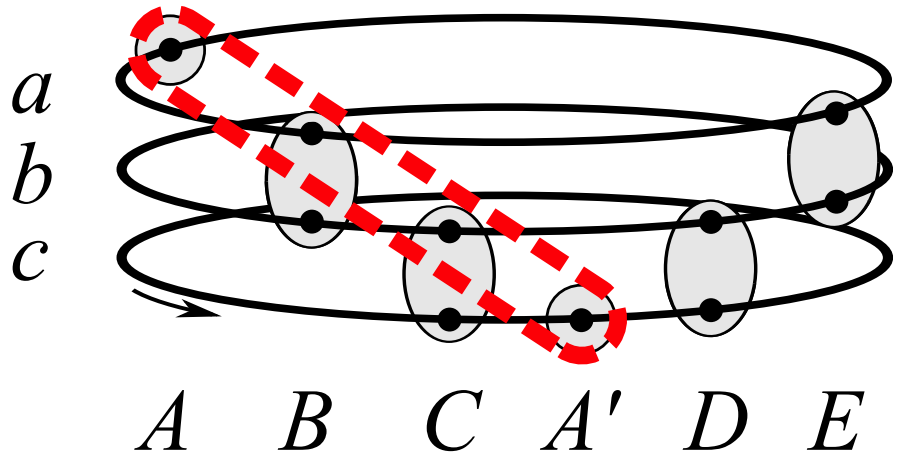


Definition (Cyclic MSA)

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 $\overline{\blacktriangleleft}$ of $\blacktriangleleft$ is a partial cyclic order of $\mathcal{A}$.

Set-theoretic Cyclic MSA Model



$A^+ = A \cup A'$

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 $\overline{\blacktriangleleft}$ of $\blacktriangleleft$ 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 $\alpha \overline{\prec} \beta$
- ... (further variables for objective function: base matches: **E**, affine gap cost: **G**, **GO**, RNA structure matches **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 $\overline{\prec}$
- $\mathbf{O}\alpha\beta$ 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 $\alpha \overline{\prec} \beta$~~ **$\mathbf{O}\alpha\beta\gamma = 1$ iff $\overline{\blacktriangleleft}\,\alpha\beta\gamma$**
- ... (further variables for objective function: base matches: **E**, affine gap cost: **G**, **GO**, RNA structure matches **B**)

Integer Linear Program (ILP):

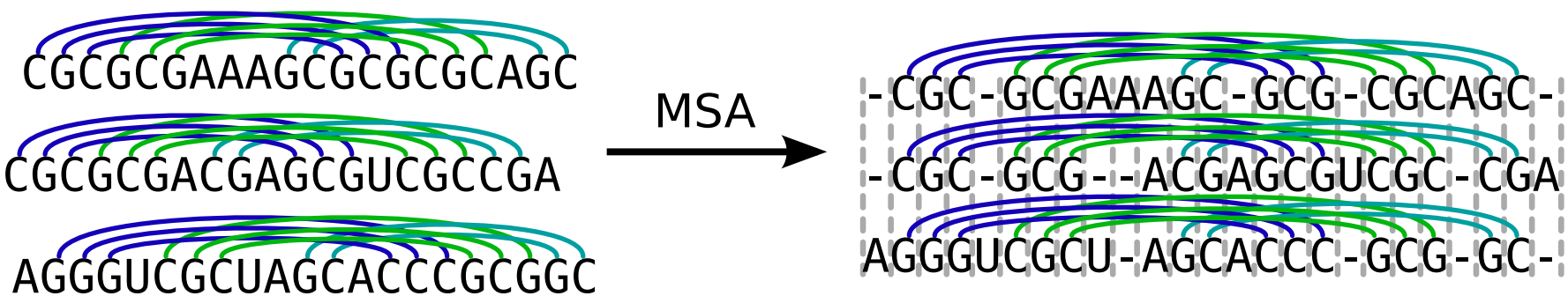
max alignment-score($\mathbf{E}, \mathbf{G}, \mathbf{GO}, \mathbf{B}$)

s.t.

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

Preliminary Results with CPLEX Solver

Instance			Model	Solving Time (s)	
#seqs	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)

  $\Rightarrow$ 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)

  $\Rightarrow$ branch-and-cut or Lagrange relaxation for cyclic MSA

Thank you!