Part 2
Comparative Analysis of RNAs

Example

Given: set of related RNA sequences

```
>AF008220 GGAGGAUUAGCUCAGCUGGGAGAGCAUCUGCCUUACAAGCAGAGGGUCGGCGGUUCGAGCCCGUCAUCCUCCA
>M68929 GCGGAUAUAACUUAGGGGUUAAAGUUGCAGAUUGUGGCUCUGAAAACACGGGUUCGAAUCCCGUUAUUCGCC
>X02172 GCCUUUAUAGCUUAGUGGUAAAGCGAUAAACUGAAGAUUUAUUUACAUGUAGUUCGAUUCUCAUUAAGGGCA
>Z11880 GCCUUCCUAGCUCAGUGGUAGAGCGCACGGCUUUUAACCGUGUGGUCGUGGGUUCGAUCCCCACGGAAGGCG
>D10744 GGAAAAUUGAUCAUCGGCAAGAUAAGUUAUUUACUAAAUAAUAGGAUUUAAUAACCUGGUGAGUUCGAAUCUCACAUUUUCCG
```

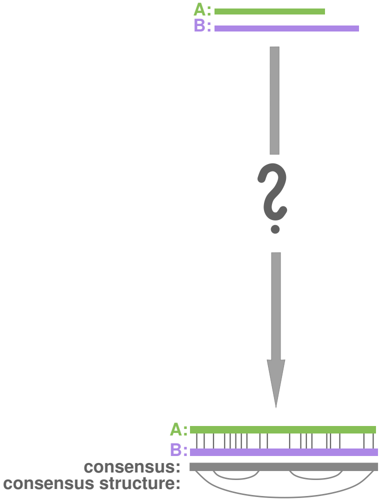
Wanted: learn about evolutionary relation

```
AF008220   GGAGGAUU-AGCUCAGCUGGGAGAGCAUCUGCCUUACAAGC---------AGAGGGUCGGCGGUUCGAGCCCGUCAUCCUCCA
M68929     GCGGAUAU-AACUUAGGGGUUAAAGUUGCAGAUUGUGGCUC---------UGAAAA-CACGGGUUCGAAUCCCGUUAUUCGCC
X02172     GCCUUUAU-AGCUUAG-UGGUAAAGCGAUAAACUGAAGAUU---------UAUUUACAUGUAGUUCGAUUCUCAUUAAGGGCA
Z11880     GCCUUCCU-AGCUCAG-UGGUAGAGCGCACGGCUUUUAACC---------GUGUGGUCGUGGGUUCGAUCCCCACGGAAGGCG
D10744     GGAAAAUUGAUCAUCGGCAAGAUAAGUUAUUUACUAAAUAAUAGGAUUUAAUAACCUGGUGAGUUCGAAUCUCACAUUUUCCG

consensus   (((((((...((((........))))((((((.......)).........))))....(((((.......)))))))))))).
```

Remarks

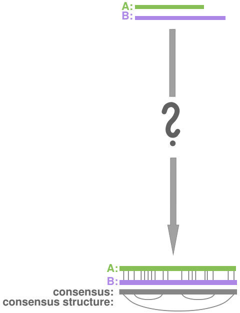
- Usually, we only know the sequences of RNAs. Why?
- Important for evolution: sequence AND structure. Why?

Comparative RNA Analysis



adopted from:
[Gardner & Giegerich BMC 2004]

Comparative RNA Analysis

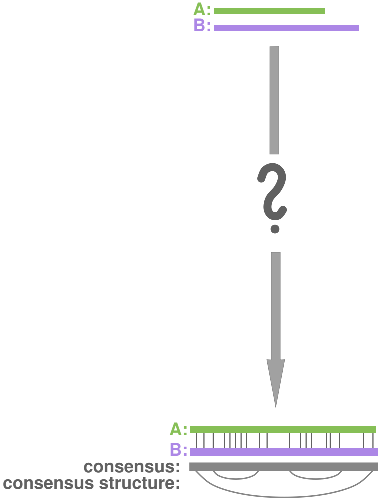


adopted from:
[Gardner & Giegerich BMC 2004]

Remarks

- Here, *Comparative RNA Analysis* refers to this problem: given a set of RNA sequences, how to match them (alignment) and what's their common structure (consensus structure).
- in general: multiple sequences, here: only pairwise

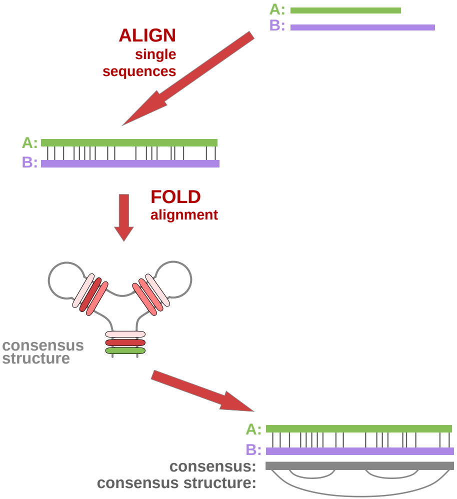
Comparative RNA Analysis



adopted from:
[Gardner & Giegerich BMC 2004]

Comparative RNA Analysis

Plan A



A:
B:

ALIGN
single sequences

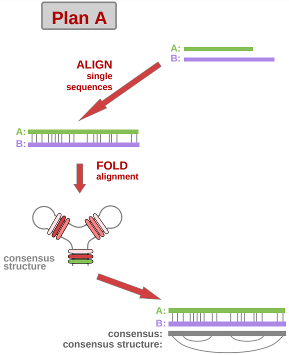
A:
B:

FOLD
alignment

consensus structure

A:
B:
consensus:
consensus structure:

Comparative RNA Analysis



Remarks

- first, simplest way. We will see two further plans.
- ALIGN: sequence alignment
- FOLD: we will generalize prediction for single sequences

Sequence Alignment, a slightly new definition

Example

```
In:  A=ACGTAA,  B=ACCCT
Out: AC-GTAA
     ACCCT--
```

“match/mismatch”, “insertion”, “deletion”

Definition (Alignment (as set of alignment edges))

An *alignment of two (RNA) sequences A and B*, $n = |A|$, $m = |B|$, is a set $\mathcal{A}$ of alignment edges, where

1. for $1 \leq i \leq n$ and $1 \leq j \leq m$, an *alignment edge* is either a matching edge (i, j) or a gap edge $(i, -)$ or $(-, j)$.
2. matching edges do not conflict
   $\forall (i, j), (i', j') \in \mathcal{A} : i < i' \implies j < j'$
3. “degree is 1”:
   - $\forall i : (i, -) \in \mathcal{A} \vee \exists! j : (i, j) \in \mathcal{A}$
   - $\forall j : (-, j) \in \mathcal{A} \vee \exists! i : (i, j) \in \mathcal{A}$

Sequence Alignment, a slightly new definition

Definition (Alignment (as set of alignment edges))

An *alignment of two (RNA) sequences A and B*, $n = |A|$, $m = |B|$, is a set $\mathcal{A}$ of alignment edges, where

1. for $1 \leq i \leq n$ and $1 \leq j \leq m$, an *alignment edge* is either a matching edge (i, j) or a gap edge $(i, -)$ or $(-, j)$.
2. matching edges do not conflict
   $\forall (i, j), (i', j') \in \mathcal{A} : i < i' \implies j < j'$
3. "degree is 1":
   - $\forall i : (i, -) \in \mathcal{A} \vee \exists! j : (i, j) \in \mathcal{A}$
   - $\forall j : (-, j) \in \mathcal{A} \vee \exists! i : (i, j) \in \mathcal{A}$

Remark

New definition equivalent to previous one via alignment strings

```
AC-GTAA
ACCCT--
```

$\equiv \quad \{(1,1), (2,2), (-,3), (3,4), (4,5), (5,-), (6,-)\}$

Recall: The Best Sequence Alignment

Idea: define best alignment as alignment with minimal edit distance

Definition (Sequence Alignment Problem)

Given two (RNA) sequences A and B, find the alignment $\mathcal{A}$ of A and B with minimal edit distance

$$\mathrm{dist}_{A,B}(\mathcal{A}) = \sum_{(i,j)\in\mathcal{A}} d(i,j),$$

where $d(i,j) = \begin{cases} \gamma & i = - \text{ or } j = - \\ w_m & A_i \neq B_j \\ 0 & A_i = B_j. \end{cases}$

- idea: how can we transform A into B? Find sequence of edit operations (match/mismatch, insertion, deletion) with minimal weight
- $d(i,j)$ weights the edit operation from positions i to j

Recall: Needleman-Wunsch Algorithm

Idea: Minimize edit distance by DP. Get best alignment by traceback.

Definition (Needleman-Wunsch Matrix)

Define the matrix $D = (D_{ij})_{0 \leq i \leq n, 0 \leq j \leq m}$ by

$$D_{ij} := \min\{\mathrm{dist}_{A,B}(\mathcal{A}) \mid \mathcal{A} \text{ alignment of } A_1, \dots, A_i \text{ and } B_1, \dots, B_j\}.$$

for $1 \leq i \leq n$, $1 \leq j \leq m$:
Init: $D_{00} = 0$, $D_{i0} = i\gamma$, $D_{0j} = j\gamma$,

Recurse: $D_{ij} = \begin{cases} D_{i-1j-1} + d(i,j) \\ D_{i-1j} + d(i,-) \\ D_{ij-1} + d(-,j) \end{cases}$

Remarks:
- recursively compute edit distances of prefix alignments
- obtain alignment by trace-back

Recall: From Pairwise to Multiple

Problem: Given set of k RNA sequences, find best multiple alignment

Definition (Multiple Alignment)

Define a *multiple alignment* $\mathcal{A}$ *of* K *(RNA) sequences* $S_1, \dots, S_K$ as a matrix of $a_{\ell i} \in \{A, C, G, U, -\}$ $(1 \leq \ell \leq K, 1 \leq i \leq m)$, s.t.

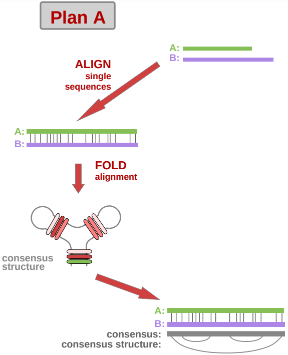
- for ℓ: deleting each occurrence of $-$ from $a_{\ell 1} \dots a_{\ell m}$ yields S_ℓ.
- for i: $a_{1i} \dots a_{Ki} \neq - \cdots -$.

Call m the *length of* $\mathcal{A}$.

Recall: *Progressive Alignment*

- pairwise alignments all-vs-all
- construct guide tree
- progressivly construct multiple alignment following guide tree

You are here



Example: S_1=CGAUACG, S_2=CGAAUACG, S_3=CCGAUUCGG

```
C-GA-UAC-G
C-GAAUAC-G
CCGA-UUCGG
```

Next: fold the alignment

How to fold an alignment

The Idea of RNAalifold

Given a K-way multiple alignment of length m.

Goal: predict the (non-crossing) consensus structure of the alignment. A consensus structure is a (non-crossing) RNA structure of length m. An optimal consensus structure minimizes a combination of

- sum of free energy over all K RNA sequences and
- a conservation score (= evidence for base pairing).

Remarks

- Think of the alignment as sequence of alignment columns. Folding of this sequence is analogous to folding of an RNA sequence. The consensus structure is a structure of the alignment.
- Thus, same decomposition as Zuker; except modified scoring: sum loop energies for all sequences & add conservation score
- Conservation score $\gamma(i,j)$ for each base pair (i,j), awards mutation — penalizes non-complementarity

RNAalifold — Example

```
AF008220  GGAGGAUU-AGCUCAGCUGGGAGAGCAUCUGCCUUACAAGC---------AGAGGGUCGGCGGUUCGAGCCCGUCAUCCUCCA
M68929    GCGGAUAU-AACUUAGGGGUUAAAGUUGCAGAUUGUGGCUC---------UGAAAA-CACGGGUUCGAAUCCCGUUAUUCGCC
X02172    GCCUUUAU-AGCUUAG-UGGUAAAGCGAUAAACUGAAGAUU---------UAUUUACAUGUAGUUCGAUUCUCAUUAAGGGCA
Z11880    GCCUUCCU-AGCUCAG-UGGUAGAGCGCACGGCUUUUAACC---------GUGUGGUCGUGGGUUCGAUCCCCACGGAAGGCG
D10744    GGAAAAUUGAUCAUCGGCAAGAUAAGUUAUUUACUAAAUAAUAGGAUUUAAUAACCUGGUGAGUUCGAAUCUCACAUUUUCCG

alifold   (((((((...((((........))))((((((.......)).........))))....(((((.......)))))))))))).
    (-49.58 = -17.46 + -32.12)
```

RNAalifold Recursions

$$W_{ij} = \min \begin{cases} W_{ij-1} \\ \min_{i \leq k < j-m} W_{ik-1} + V_{kj} \end{cases}$$

$$V_{ij} = \beta\gamma(i,j) + \min \begin{cases} \sum_{1 \leq \ell \leq K} \mathrm{eH}(i,j,S_\ell) \\ \sum_{1 \leq \ell \leq K} \min_{i<i'<j'<j} V_{i'j'} + \mathrm{eSBI}(i,j,i',j',S_\ell) \\ \min_{i<k<j} WM_{i+1k} + WM_{k+1j-1} + aK \end{cases}$$

$$WM_{ij} = \min \begin{cases} WM_{ij-1} + cK, WM_{i+1j} + cK, V_{ij} + bK \\ \min_{i<k<j} WM_{ik} + WM_{k+1j} \end{cases}$$

Remarks

- $\mathrm{eH}(i,j,S_\ell)$ and $\mathrm{eSBI}(i,j,i',j',S_\ell)$ yield energy contributions for the respective S_ℓ.

RNAalifold Recursions

$$W_{ij} = \min \begin{cases} W_{ij-1} \\ \min_{i \leq k < j-m} W_{ik-1} + V_{kj} \end{cases}$$

$$V_{ij} = \beta\gamma(i,j) + \min \begin{cases} \sum_{1 \leq \ell \leq K} \mathrm{eH}(i,j,S_\ell) \\ \sum_{1 \leq \ell \leq K} \min_{i<i'<j'<j} V_{i'j'} + \mathrm{eSBI}(i,j,i',j',S_\ell) \\ \min_{i<k<j} WM_{i+1k} + WM_{k+1j-1} + aK \end{cases}$$

$$WM_{ij} = \min \begin{cases} WM_{ij-1} + cK, WM_{i+1j} + cK, V_{ij} + bK \\ \min_{i<k<j} WM_{ik} + WM_{k+1j} \end{cases}$$

Remarks

- $\mathrm{eH}(i,j,S_\ell)$ and $\mathrm{eSBI}(i,j,i',j',S_\ell)$ yield energy contributions for the respective S_ℓ.
- RNAalifold implements an unambiguous variant of these recursions for computing partition function and base pair probabilities for the consensus structure.
- β weights conservation score vs. sum of free energy. For γ see next slide.

RNAalifold Conservation Score

conservation score = covariation + penalty

$$\gamma(i,j) =$$

$$-\frac{1}{2}\sum_{1\leq\ell<\ell'\leq K}\begin{cases} h(a_{\ell i}, a_{\ell' i}) + h(a_{\ell j}, a_{\ell' j}) & a_{\ell i} - a_{\ell j},\ a_{\ell' i} - a_{\ell' j} \text{ compl.} \\ 0 & \text{otherwise,} \end{cases}$$

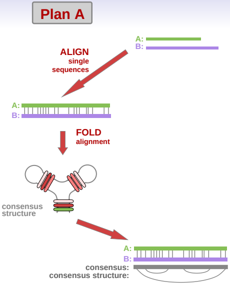
(*covariation*)

$$+\,\delta\sum_{1\leq\ell\leq K}\begin{cases} 0 & a_{\ell i} - a_{\ell j} \text{ complementary} \\ 0.25 & a_{\ell i}, a_{\ell j} \text{ are both gaps} \\ & \\ 1 & \text{otherwise,} \end{cases}$$

(*penalty*)

hamming distance $h(x,y) = \begin{cases} 1 & x \neq y \\ 0 & x = y \end{cases}$

Comparative RNA Analysis: Plan A — summary



- alignment doesn't look at structure

  → misalignment likely (when?)

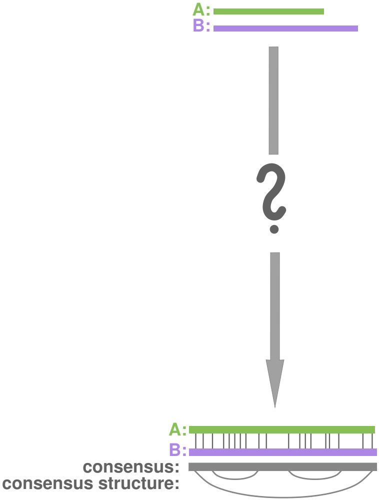
  folding step cannot revise alignment

  → misalignment cannot fold correctly
- very useful, when
  - sequence similarity high
  - alignment is already given/known

    → infer consensus structure

    → measure alignment quality

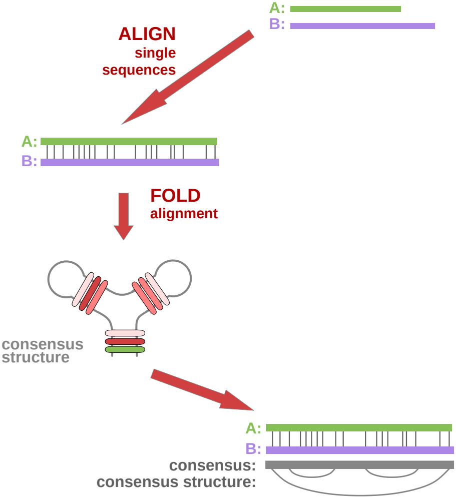
Revisit Comparative RNA Analysis



adopted from:
[Gardner & Giegerich BMC 2004]

Revisit Comparative RNA Analysis

Plan A



A:

B:

ALIGN
single sequences

A:

B:

FOLD
alignment

consensus structure

A:

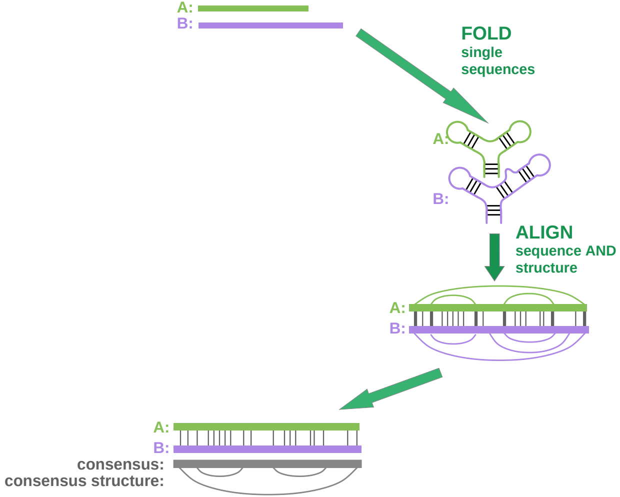
B:

consensus:

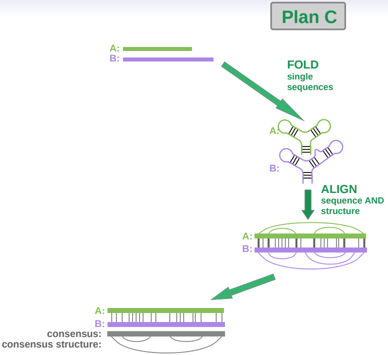
consensus structure:

Revisit Comparative RNA Analysis

Plan C



Comparative RNA Analysis: Plan C



Remarks

- we already know step one FOLD!
- remaining: ALIGN — given RNA (sequences and) structures, align using sequence and structure information!
- how will this differ from sequence alignment/edit distance
- what is better/worse than in plan A?

Aligning Sequence and Structure

General Sequence Structure Alignment Problem

Given two RNA sequences A and B with resp. RNA structures P_A and P_B. Find the best alignment of the two RNAs.

Aligning Sequence and Structure

General Sequence Structure Alignment Problem

Given two RNA sequences A and B with resp. RNA structures P_A and P_B. Find the best alignment of the two RNAs.

More questions than answers

- what means best? how to use structure information?
- are the structures restricted?
- what means alignment?

Aligning Sequence and Structure

General Sequence Structure Alignment Problem

Given two RNA sequences A and B with resp. RNA structures P_A and P_B. Find the best alignment of the two RNAs.

More questions than answers

- what means best? how to use structure information?
  penalize structural mismatch → edit distance
- are the structures restricted?
- what means alignment?

Aligning Sequence and Structure

General Sequence Structure Alignment Problem

Given two RNA sequences A and B with resp. RNA structures P_A and P_B. Find the best alignment of the two RNAs.

More questions than answers

- what means best? how to use structure information?
  penalize structural mismatch → edit distance
- are the structures restricted?
  distinguish crossing/non-crossing input
- what means alignment?
  necessarily the same as sequence alignment?

Non-Crossing Sequence Structure ≡ Tree

Idea: for non-crossing RNA, reduce RNA comparison to comparing trees (i.e. reduce to a more general problem in computer science).

Example:

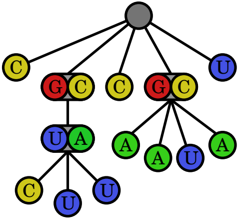
```
CGUCUUACCGAAUACU    AGUCUUCGAAAACU
.((...)).(....).    ((....)(....))
```

Non-Crossing Sequence Structure ≡ Tree

Idea: for non-crossing RNA, reduce RNA comparison to comparing trees (i.e. reduce to a more general problem in computer science).

Example:

```
CGUCUUACCGAAUACU    AGUCUUCGAAAACU
.((...)).(....).    ((....)(....))
```

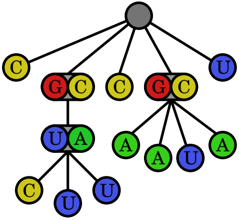


Non-Crossing Sequence Structure ≡ Tree

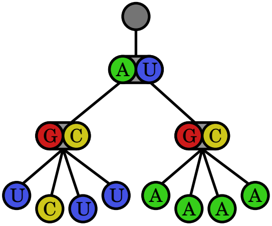
Idea: for non-crossing RNA, reduce RNA comparison to comparing trees (i.e. reduce to a more general problem in computer science).

Example:

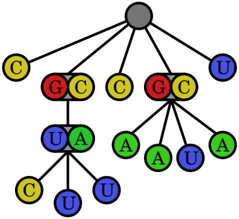
```
CGUCUUACCGAAUACU
.((...)).(....).
```



```
AGUCUUCGAAAACU
((....)(....))
```



RNA Tree

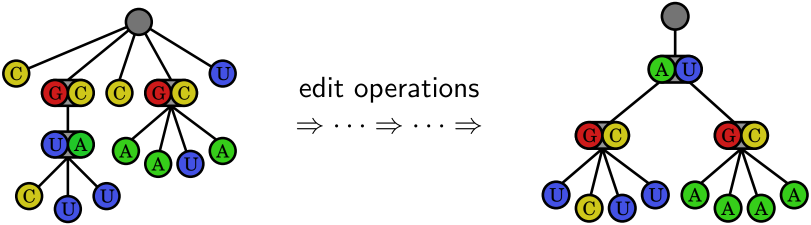


Definition (RNA tree)

An *RNA tree* is an ordered tree G. The nodes $v \in V_G$ are either base nodes or base pair nodes (or root). Nodes are labled. For base nodes, label$(v) \in \{A, C, G, U\}$ and for base pair nodes label$(v) \in \{AU, UA, CG, GC, GU, UG\}$.

How to Compare Trees I: Tree Editing

Idea: tranform the first tree into the second tree by edit operations



edit operations

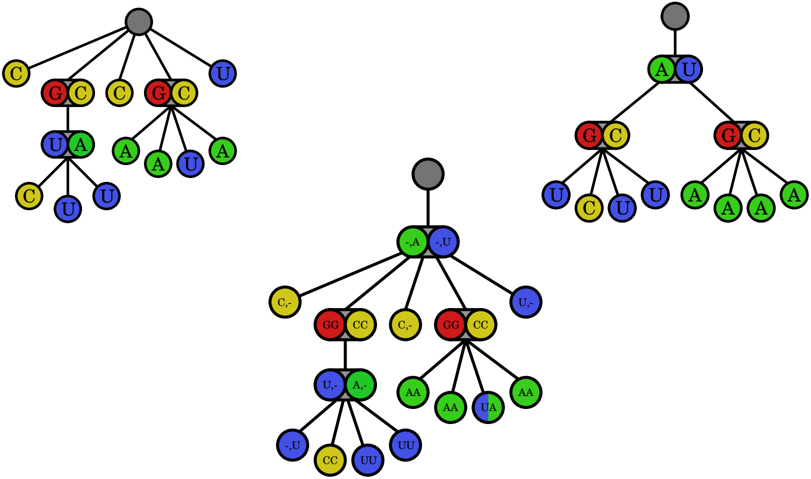
$\Rightarrow \cdots \Rightarrow \cdots \Rightarrow$

- rename base
- insert/delete base node
- rename base pair
- insert/delete base pair node

Remark: assign cost to edit ops and find best sequence of edit ops

How to Compare Trees II: Tree Alignment

Idea: common super-tree = *tree alignment*



Remark: assign cost to nodes of tree alignment and find best one

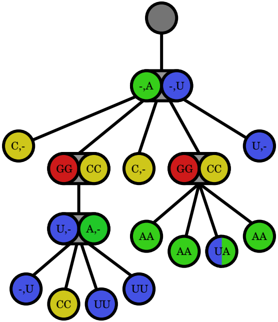
How to Compare Trees II: Tree Alignment

Alignment of two strings = string with tuples as characters.

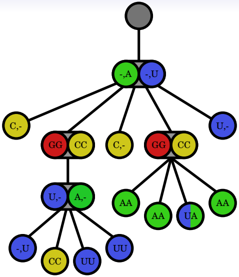
-CGU-CUUACCGAAUACU-

A-G-UCUU-C-GAAAAC-U

Alignment of two trees = tree with tuples as labels



Tree Alignment



Definition (RNA tree alignment)

An *RNA tree alignment* is an ordered tree T. The nodes $v \in V_T$ are either base nodes or base pair nodes (or root). Nodes have pairs of labels $(\mathrm{label}_1(v), \mathrm{label}_2(v))$. For base nodes, $\mathrm{label}_i(v) \in \{A, C, G, U, -\}$ and for base pair nodes $\mathrm{label}_i(v) \in \{AU, UA, CG, GC, GU, UG, --\}$ $(i = 1, 2)$.
An RNA tree alignment T is *RNA tree alignment of two RNA trees F and G* iff "projecting" T to the first or second labels is F or G respectively. (Projection deletes "gap nodes".)

Tree Alignment Problem

Definition (RNA tree alignment problem)

We define a cost w for each node of an RNA tree alignment depending on the node labels. Given two RNA trees $F = (V_F, E_F)$ and $G = (V_G, E_G)$, the *RNA tree alignment problem* is finding the minimal cost RNA tree alignment $T = (V_T, E_T)$ of F and G, where cost of T is

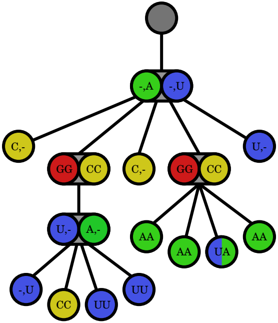
$$\mathrm{cost}(T) = \sum_{v \in V_T} \mathrm{w}(v).$$

Remark

RNAforester (Hoechsmann et al.) implements a solution of this kind of tree alignment problem.

Tree Alignment Yields Alignment of Arc Annotated Sequences

Tree alignment:



Alignment of arc annotated sequences:

```
-.((-...)).(....).-
-CGU-CUUACCGAAUACU-

A-G-UCUU-C-GAAAAC-U
(-(-....-)-(....)-)
```

Tree Alignment Limitations

Some alignments of arc annotated sequences cannot be obtained from tree alignments:

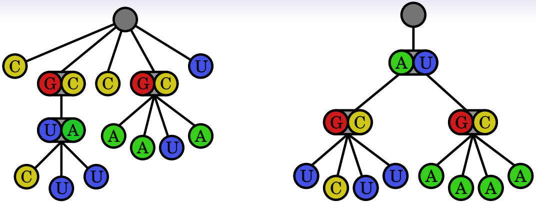
```
(.....)...
GCA-UGCAC-

...(.....)
-CACUG-ACG
```

Limitation:

Tree alignment does not allow alignments where the combination of the single structures forms a crossing structure.

structure 1	(.....)...
structure 2	...(.....)
combination	(..[..)..]

Edit Ops on Trees are Ops on Arc-annotated Sequences



```
CGUCUUACCGAAUACU      AGUCUUCGAAAACU
.((...)).(....).      ((....)(....))
```

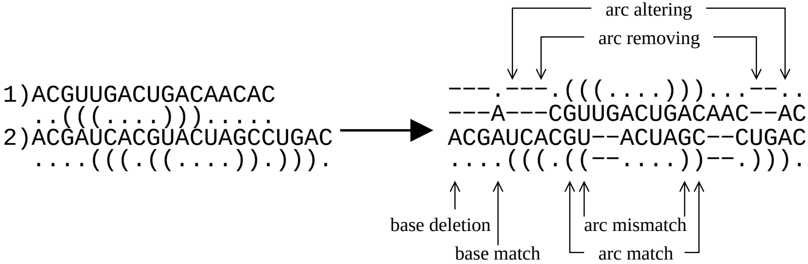
Remarks

- Therefore, tree editing is more flexible then tree alignment.
- Tree alignment limits possible alignment (must correspond to tree alignment).
- In tree editing insertions and deletions of arcs can “cross”.
- More flexible edit operations.

```
T.-Alignment: -.((-...)).(....).-   T.-Editing: .((...)).(....).
              -CGU-CUUACCGAAUACU-               CGUCUUACCGAAUACU

              A-G-UCUU-C-GAAAAC-U               AGUCUU-C-GAAAACU
              (-(-....-)-(....)-)               ((....-)-(....))
```

General Edit Operations

Arc annotated sequence view allows introducing more general edit operations



```
1)ACGUUGACUGACAACAC
  ..(((....)))..... 
2)ACGAUCACGUACUAGCCUGAC
  ....(((.((....)).))).

---.---.(((....)))...--..
---A---CGUUGACUGACAAC--AC
ACGAUCACGU--ACUAGC--CUGAC
....(((.((--....))--.))).
```