Conservation in Long Non-coding RNAs and other Updates

Peter F. Stadler

Bioinformatics Group, Dept. of Computer Science &
Interdisciplinary Center for Bioinformatics,
University of Leipzig
Max Planck Institute for Mathematics in the Sciences
RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology
Institute for Theoretical Chemistry, Univ. of Vienna (external faculty)
Center for non-coding RNA in Technology and Health, U. Copenhagen
The Santa Fe Institute (external faculty)

Universidad Nacional de Colombia (prof. hon.)

Bled, Feb 2020

Incongruent Evolution

```
CGUGGAAACCCACAG
.((((....))))..

CGUGGAAACC-CACAG
.((((....)-))).
.((-((....))))..
CGU-GAAACCUCACAG
```

```
CGUGAAACCUCACAG
.((((....))))..

.((((....))))..
CGUGGAAACCCACAG
CGUGAAACCUCACAG
.((((....))))..
```

exact conservation of the structure

Incongruent Evolution

Similar sequence, shifted structure

```
..((((.((((....)))).))))....
AAGGCUCUAUUAACUGGUAUCGGCUAUAG
**  *  ***** ***** **** *  *  ***
AAUGAUCUAUGAACUGUUAUCUGAUUUAG
...((((.((((....)))).))))...
```

```
..-((((.((((....)))).))))....
AA-GGCUCUAUUAACUGGUAUCGGCUAUAG
**  *          *           *   *  ***
AAUGAUCUAUGAACUGUUAUCUGAUU-UAG
...((((.((((....)))).))))-...
```

Is there really incongruent evolution?

Sequence and structure alignments mir-30a and mir-30b

```
consensus    ..U.AG....UGUAAACAUCCU..ACU...AGCUGU.A...CA....U.GGCU...A-GU.GGAUGUUUGC..C.GC...CU
mmu-mir-30a AGUGAGCGACUGUAAACAUCCUCGACUGGAAGCUGUGAAGCCACAAAUGGGCUUUCA-GUCGGAUGUUUGCAGCUGCCUACU
SS          ((((.(((.(((((((((((..(((((((((((.....))).....))))))))-)))))))))))))).))).))))
mmu-mir-30b -UUCAGUUCAUGUAAACAUCCUACACU--CAGCUGUCAU--CAUGCGUUGGCUGGGAUGU-GGAUGUUUACGUCAGCUGUCU
SS          -..(((((.((((((((((((.(((((--(((((..(..--.....)..)))))))).)))-)))))))))))).)))))...
consensus   ...[.[[[.[[[[[[[[[[[[.......[[[[[.............]]]]].......]]]]]]]]]]]].]]].]...

consensus    .....G....UGUAAACAUCCU.......A...........C........G..U.......GGAUGUUU.C....GC.U....
mmu-mir-30a AGUGAGCGACUGUAAACAUCCUCGACUGGAAGCUGUGAA-GCCACAAAUGGGCUUUCA-GUCGGAUGUUUGCAGCUGCCUACU-
SSi         ((((.(((.(((((((((((..(((((((((((..-..))).....)))))))-)))))))))))))).))).))))-
mmu-mir-30b UUCA-GUUCAUGUAAACAUCCU-ACACUCAGCUGUCAUCAUGCGUU-----GGCUGGGAUGUGGAUGUUUACGUCAGC-UGUCU
SS          ..((-(((.((((((((((((.-(((((((((..(.......)..-----))))))).)))))))))))))).)))-))...
consensus   ..[[.[[[.[[[[[[[[[[[[..[[[[[[[[[..[.......].......]]]]]]]].]]]]]]]]]]]]].]]].]]...
```

Incongruent Alignments

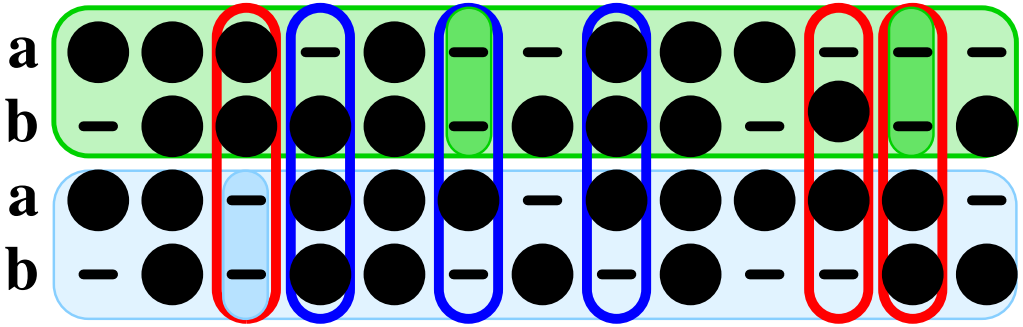
Basic idea: consider two or more alignments of the same objects (strings) simultaneously:

- implicitly defines alignments between the different copies of the same objects that do not allow mismatches
- Insertions and deletions in these same-object alignments correspond to *shifts* between the incongruent alignments
- scoring function:
  weighted scores of the consitutent alignments + scores for the "shifts"

... what exactly are "shifts"?

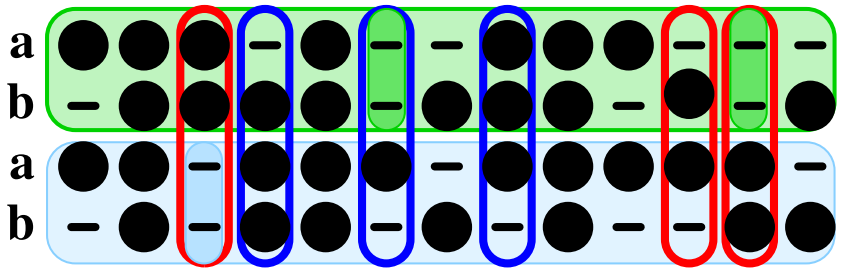
Formalization: Bi-Alignments

- two distinct alignments $\mathbb{U}$ and $\mathbb{V}$ of the same objects **a** and **b**
- an alignment $\mathbb{W}$ of the columns of $\mathbb{U}$ and $\mathbb{V}$
- score $= u(\mathbb{U}) + v(\mathbb{V}) + w(\mathbb{W})$
- Bi-alignment Problem: simultaneously optimize $\mathbb{U}$, $\mathbb{V}$, and $\mathbb{W}$.



Formalization: Shifts

- Gap patterns $c, d \in \binom{0}{1}$ in $\mathbb{U}$ and $\mathbb{V}$, respectively
- Congruent columns: $c_1 = d_1$ and $c_2 = d_2$.
- Incongruence $\|c - d\| = |c_1 - d_1| + |c_2 - d_2| \in \{0, 1, 2\}$
- score $w(\mathbb{W})$: proportional to the sum of the incongruences of the alignment columns.
- An alignment of alignments is again an alignment:
  $\mathbb{A} \simeq (\mathbb{U}, \mathbb{V}, \mathbb{W})$



- Number of in/dels between the two copies of **a** and **b**:
  $d(\mathbb{A}_{13}) = \sum_i |c_1(i) - d_1(i)|$ $\quad$ $d(\mathbb{A}_{24}) = \sum_i |c_2(i) - d_2(i)|$
- Columnwise scoring of $\mathbb{A}$:
  score of the projected alignments $\mathbb{U} \simeq \mathbb{A}_{12}$ and $\mathbb{V} \simeq \mathbb{A}_{34}$ plus the in/del-only scores $d(\mathbb{A}_{13})$ and $d(\mathbb{A}_{24})$.

Scoring Shifts

$$A \to A\binom{\bullet}{\bullet} \mid A\binom{\bullet}{-} \mid A\binom{-}{\bullet} \mid \cdots \mid A\binom{\bullet}{-} \mid A\binom{-}{\bullet} \mid A\binom{-}{-} \mid \varepsilon\,.$$

	$\binom{\bullet}{\bullet}$	$\binom{\bullet}{-}$	$\binom{-}{\bullet}$	$\binom{-}{-}$
$\binom{\bullet}{\bullet}$	0	Δ	Δ	2Δ
$\binom{\bullet}{-}$	Δ	0	2Δ	Δ
$\binom{-}{\bullet}$	Δ	2Δ	0	Δ
$\binom{-}{-}$	2Δ	Δ	Δ	-

$$M(0) = 0$$

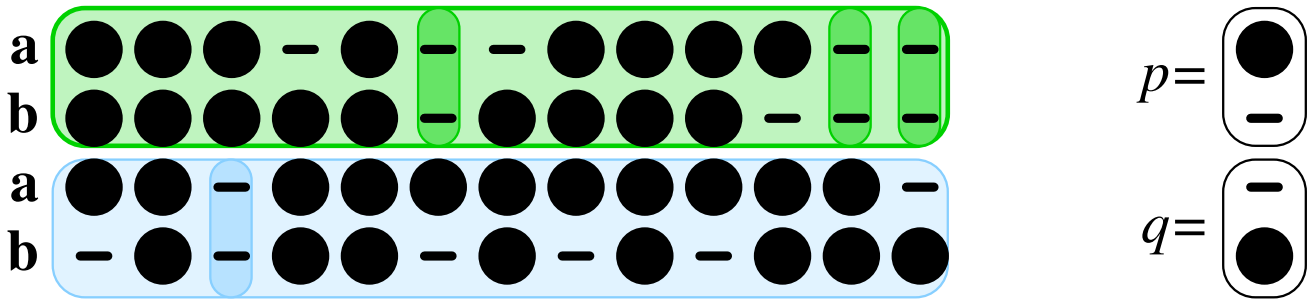
$$M(x) = \max_{c \in \mathcal{C}} M(x - c) + s(x, c)$$

A very preliminary scan survey

- small and medium-width `Rfam` seed alignments
  ($\leq$ 10 sequences, $\leq$ 120 columns)
  1181 `Rfam` families
- check if `Rfam` consensus-structure "oriented" Rfam alignment is significantly different from a `mafft` re-alignment of the sequences
  709 candidate families
- 10137 pairs of RNA sequences yield 143 cases in 72 families with prediced shifts in a sequence-based shift alignment

Affine Gap Costs in $\mathbb{U}$ and $\mathbb{V}$

- Gotoh's algorithm for each of $\mathbb{U}$ and $\mathbb{V}$
  scoring depending of the gap pattern of the penultimate column
- insufficient here: the penultimate column could be double-gap,
  i.e., an in/del of $\mathbb{W}$.
- remedy: keep *end gap pattern* defined for the last column that is **not** a double-gap:



Including secondary structure

$$A \rightarrow Ac \mid A\bar{c}Ac \mid \varepsilon$$

$$u(\mathbb{U}, \varphi_{\mathbb{U}}) + v(\mathbb{V}, \varphi_{\mathbb{V}}) + w(\mathbb{W})$$

$$M(x,y) = \max \begin{cases} \max\limits_{c \in \mathcal{C}} M(x, y-c) + s(y,c) \\ \max\limits_{\substack{(z,y)\in\mathcal{B}^* \\ (c,d)\in\mathcal{C}^*}} M(x,z-c) + M(z,y-d) + \tilde{s}(z,c;y,d) \end{cases}$$

$\mathcal{B}^*$ ... allowed index combinations, enforce base pairs

$$(c,d) \in \mathcal{C}' := \left\{ \begin{pmatrix} - \\ \bullet \end{pmatrix}, \begin{pmatrix} \bullet \\ \bullet \end{pmatrix}, \begin{pmatrix} - \\ \bullet \\ \bullet \end{pmatrix}, \begin{pmatrix} \bullet \\ \bullet \\ \bullet \end{pmatrix} \right\}^2$$

... Sankoff-style Bi-Alignments

Sankoff-style Bi-Alignments

- Complexity?
  $O(n^8)$ entries times $O(n^4)$ operations
- BUT: number of shifts is limited:

$$k\Delta \leq \delta^*(\mathbf{a},\mathbf{b}) := \max_{\mathbb{U}} u(\mathbb{U}) + \max_{\mathbb{U}} v(\mathbb{U}) - \max_{\mathbb{U}}[u(\mathbb{U}) + v(\mathbb{U})]\,.$$

  reduction to $O(n^4k^4)$ entries with $O(n^2k^2)$ operations, i.e., $O(n^6)$ like the Sankoff algorithm
- `locarna` approximation: only include $O(n)$ most frequent base pairs for each structure
- reduction to $O(n^2)$ space and time.
- Implementation: on the way

Matrices $M_{p,q}$ indexed by end gap patterns p and q for $\mathbb{U}$ and $\mathbb{V}$

$$M_{(p,q)}(x,y) = \max \begin{cases} \max\limits_{\substack{p'\neq 0\\ q'\neq 0}} M_{(p',q')}(x-q,y-q) + s(\binom{x}{y},\binom{p'}{q'},\binom{p}{q}) \\ \max\limits_{p'\neq 0} M_{(p',q)}(x-p,y) + s(\binom{x}{y},\binom{p'}{q},\binom{p}{0}) \\ \max\limits_{q'\neq 0} M_{p,q'}(x,y-q) + s(\binom{x}{y},\binom{p}{q'},\binom{0}{q}) \end{cases}$$

Collaborators

- Maria Waldl
- Sebastian Will
- Michael T. Wolfinger
- Christoph Flamm
- Christian Höner zu Siederdissen
- Ivo L. Hofacker