TBA@TBI WinterSeminar

...as advertised on the event website

Yann Ponty

+ Bertrand Marchand*, Sebastian Will, Sarah Berkemer and Laurent Bulteau

AMIBio Team
CNRS & LIX, Institut Polytechnique de Paris

Automated Design of Dynamic Programming Algorithms

Another step towards the elusive perfect algorithm for folding RNAs with tertiary motifs

Yann Ponty

+ Bertrand Marchand*, Sebastian Will, Sarah Berkemer and Laurent Bulteau

AMIBio Team
CNRS & LIX, Institut Polytechnique de Paris

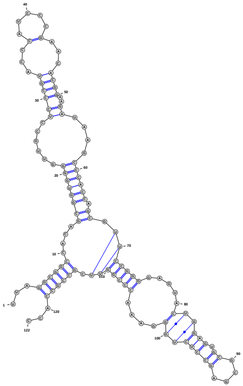
RNA structure(s)

RNA = Linear Polymer = Nucleotides sequence $w \in \{A, C, G, U\}^\star$

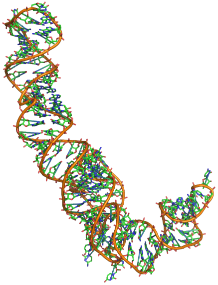
```
UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC
```

Primary struct.

aka sequence



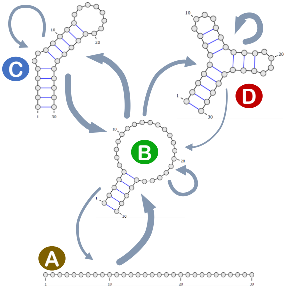
Secondary struct.



Tertiary (≈ 3D) struct.

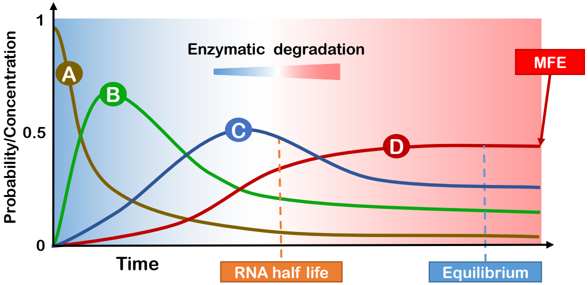
5s rRNA (PDB ID: 1K73:B)

Physics/chemistry paradigms in RNA structural bioinformatics



A – Kinetic Landscape

Continuous-time Markov chain

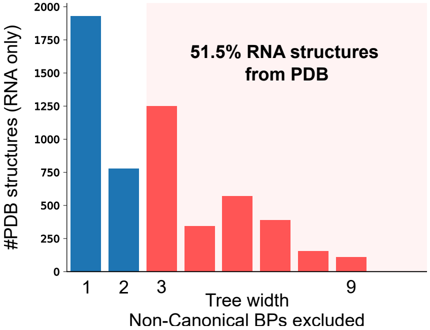


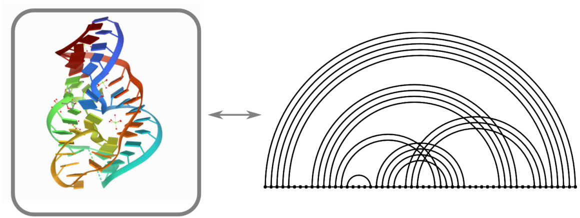
B – Evolution of concentrations

Given free-energy E at Boltzmann equilibrium $\mathbb{P}(S \mid w) \propto e^{-\Delta G(w,S)/RT}$

- Minimum Free-Energy (MFE): Functional structure = Most stable/probable
- Partition function: Equilibrium properties of Boltzmann ensemble
- Kinetics: Finite-time evolution of concentrations/probabilities

Why pseudoknots?





Twister ribozyme (PDB ID: 3OJI)

Pseudoknots are highly prevalent among non-coding RNAs:

- 48.5% RNA structures from PDB feature pseudoknots
- 565 RFAM families have pseudoknotted consensus structures

The CS/nerdy side of MFE prediction (with pseudoknots)

- Pseudoknot-free MFE solved in $\mathcal{O}(n^3)$ time for finite energy models

  Current best: $\mathcal{O}(n^{2.87})$ for BP-based models [Bringmann, Grandoni, Saha, & Williams 2016]

  Lower bounds: SETH $\rightarrow \mathcal{O}(n^2)$, k-clique $\rightarrow \mathcal{O}(n^{\omega})$, $\omega < 2.374$ [Abboud, Backurs & Williams 2015]

The CS/nerdy side of MFE prediction (with pseudoknots)

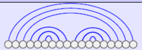
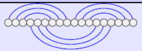
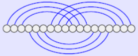
- Pseudoknot-free MFE solved in $\mathcal{O}(n^3)$ time for finite energy models

  Current best: $\mathcal{O}(n^{2.87})$ for BP-based models [Bringmann, Grandoni, Saha, & Williams 2016]

  Lower bounds: SETH $\to \mathcal{O}(n^2)$, k-clique $\to \mathcal{O}(n^\omega)$, $\omega < 2.374$ [Abboud, Backurs & Williams 2015]

- With pseudoknots $\to$ NP-hard for any realistic energy model

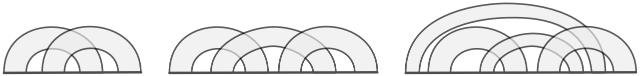
  [Lyngsø & Pedersen 2000; Akutsu & Uemura 2000; Sheikh, Backofen & Ponty 2012]

		Base-pairs	Stacking-Pairs	Nearest-Neighbor
	Comp.	P [Nussinov 80]	P [Ieong 03]	P [Zuker 81]
Non-crossing	Approx.	–	–	–
	Comp.	???	NP-Hard [Ieong 03]	NP-Hard [Ieong 03]
Planar	Approx.	2-approx. ≈[Ieong 03]	2-approx. [Ieong 03]	???
	Comp.	P [Tabaska 98]	NP-Hard [Lyngsø 04] **(any* Δ model)**	NP-Hard [Lyngsø 00, Akutsu 00]
General	Approx.	–	ε-approx. $\in \mathcal{O}(n^{4^{1/\varepsilon}})$ [Lyngsø 04] **1/5 (any Δ model)**	**APX-Hard**

History of RNA folding with PK (+ Interactions)

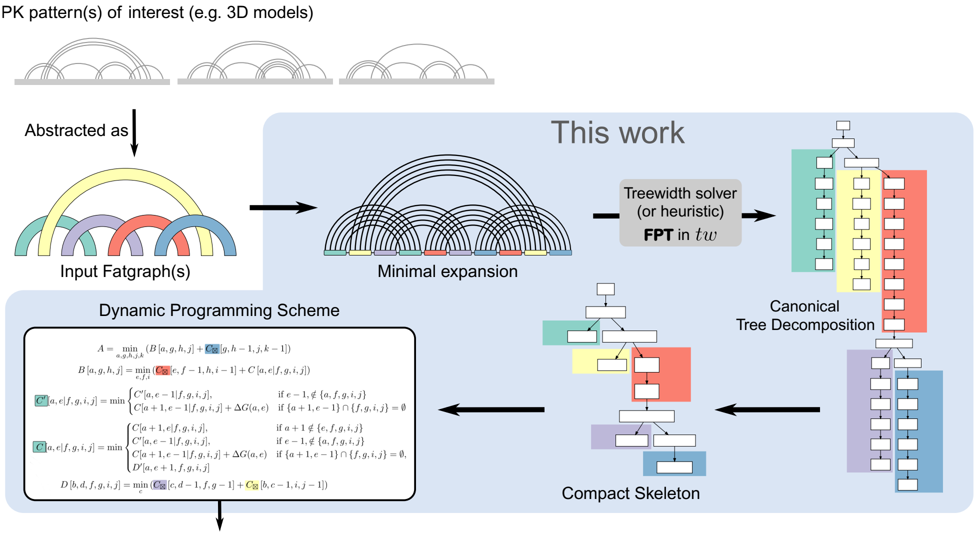
Tool	Reference	space	time	restriction
Pknots-RE	[Rivas and Eddy, 1999]	$O(n^4)$	$O(n^6)$	"one-hole structures"
NUPACK	[Dirks and Pierce, 2003]	$O(n^4)$	$O(n^5)$	"2 interleaved helices"
gfold	[Reidys et al., 2011]	$O(n^4)$	$O(n^6)$	genus ≤ 1
CCJ	[Chen et al., 2009]	$O(n^4)$	$O(n^5)$	"3 groups of bands"
Knotty	[Jabbari et al., 2018]	$O(n^3 + Z)$	$O(n^5)$	"CCJ-type + optims"
Pknots-RG	[Reeder and Giegerich, 2004]	$O(n^2)$	$O(n^4)$	"simple recursive PK"

- Manual design using Dynamic Programming (DP)
- Grammar-based frameworks [Giegerich & friends, Hoehner zu Siederdissen& Stadler]
- Topology-inspired (Genus → Fatgraphs) [Bon/Orland/Vernizzi, Reidys/Nebel/Stadler]



Our goal: To rationalize/automate the design of *conjunctive* DP schemes

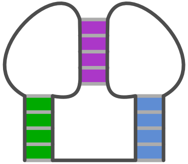
Workflow

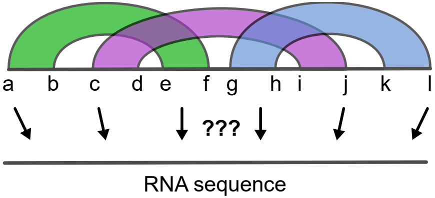


General algorithm for Fatgraph MFE folding problem
Correct for any input sequence (+ `C/C++` Code generation)

Complexity: $O(n^{tw})$ or $O(n^{tw+1})$ for simple energy models
$O(n^{tw+1})$ for full Turner model

Kissing hairpins: an example





MFE folding w.r.t. shadow = Find embedding of motifs anchors into RNA:

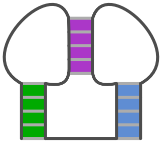
- Consistency: Consecutive anchors should be assigned increasing positions
- Helices scoring: 4 extremities jointly seen, or (inner | outer)ward progression

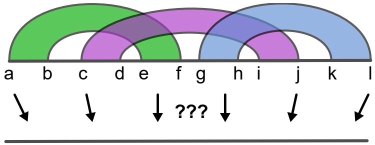
Option 1: Brute force 

$$MFE := \min_{\mathbf{a}<\mathbf{b}<\mathbf{c}<\mathbf{d}<\ldots<\mathbf{l}} \Delta G(a,b,e,f) + \Delta G(c,d,i,j) + \Delta G(g,h,k,l)$$

$\rightarrow \mathcal{O}(n^{12})$ time complexity 

Kissing hairpins: an example



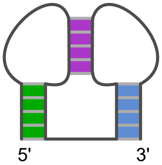


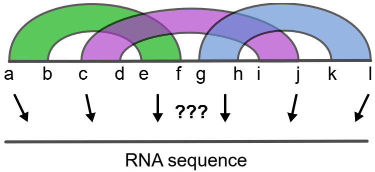
Option 2: Left to right strategy, starting from $(a, l) \longrightarrow MFE := \min_{\mathbf{a}<\mathbf{l}} A(a, l)$

Add anchors from left to right, forget unnecessary anchors:

$$A(a, l) = \min_{a<\mathbf{b}<l} B(a, b, l)$$

Kissing hairpins: an example



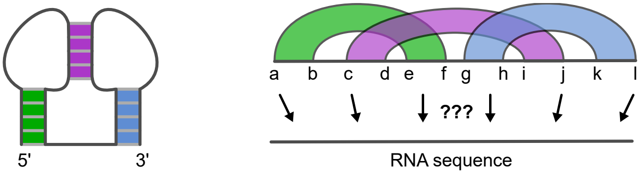


Option 2: Left to right strategy, starting from $(a, l) \longrightarrow MFE := \min_{\mathbf{a}<\mathbf{l}} A(a, l)$

Add anchors from left to right, forget unnecessary anchors:

$$A(a, l) = \min_{a<\mathbf{b}<l} B(a, b, l) \quad B(a, b, l) = \min_{b<\mathbf{c}<l} C(a, b, c, l)$$

Kissing hairpins: an example



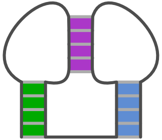
Option 2: Left to right strategy, starting from $(a, l) \longrightarrow MFE := \min_{\mathbf{a}<\mathbf{l}} A(a, l)$

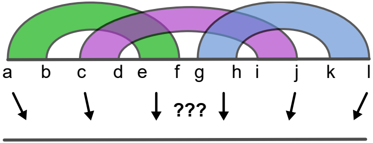
Add anchors from left to right, forget unnecessary anchors:

$$A(a,l) = \min_{a<\mathbf{b}<l} B(a,b,l) \quad B(a,b,l) = \min_{b<\mathbf{c}<l} C(a,b,c,l) \quad C(a,b,c,l) = \min_{c<\mathbf{d}<l} D(a,b,c,d,l)$$

$$D(a,b,c,d,l) = \min_{d<\mathbf{e}<l} E(a,b,c,d,e,l)$$

Kissing hairpins: an example





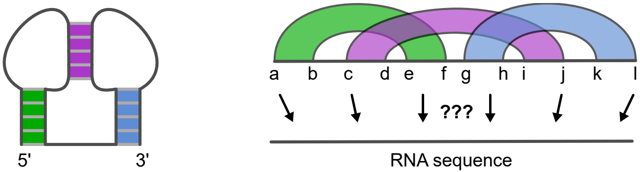
Option 2: Left to right strategy, starting from $(a, l) \longrightarrow MFE := \min_{\mathbf{a}<\mathbf{l}} A(a, l)$

Add anchors from left to right, forget unnecessary anchors:

$$A(a, l) = \min_{a<\mathbf{b}<l} B(a, b, l) \quad B(a, b, l) = \min_{b<\mathbf{c}<l} C(a, b, c, l) \quad C(a, b, c, l) = \min_{c<\mathbf{d}<l} D(a, b, c, d, l)$$

$$D(a, b, c, d, l) = \min_{d<\mathbf{e}<l} E(a, b, c, d, e, l) \quad E(a, b, c, d, e, l) = \min_{e<\mathbf{f}<l} F(\cancel{a}, \cancel{b}, c, d, \cancel{e}, f, l) + \Delta G(a, b, e, f)$$

Kissing hairpins: an example



Option 2: Left to right strategy, starting from $(a, l) \longrightarrow MFE := \min_{\mathbf{a}<\mathbf{l}} A(a, l)$

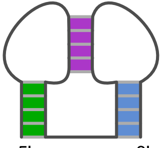
Add anchors from left to right, forget unnecessary anchors:

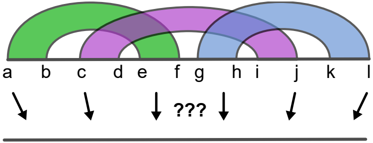
$$A(a,l) = \min_{a<\mathbf{b}<l} B(a,b,l) \quad B(a,b,l) = \min_{b<\mathbf{c}<l} C(a,b,c,l) \quad C(a,b,c,l) = \min_{c<\mathbf{d}<l} D(a,b,c,d,l)$$

$$D(a,b,c,d,l) = \min_{d<\mathbf{e}<l} E(a,b,c,d,e,l) \quad E(a,b,c,d,e,l) = \min_{e<\mathbf{f}<l} F(\cancel{a},\cancel{b},c,d,\cancel{e},f,l) + \Delta G(a,b,e,f)$$

$$F(c,d,f,l) = \min_{f<\mathbf{g}<l} G(c,d,f,g,l)$$

Kissing hairpins: an example





Option 2: Left to right strategy, starting from $(a, l) \longrightarrow MFE := \min_{\mathbf{a}<\mathbf{l}} A(a, l)$

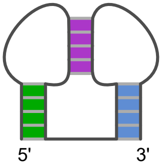
Add anchors from left to right, forget unnecessary anchors:

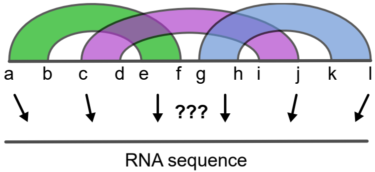
$$A(a,l) = \min_{a<\mathbf{b}<l} B(a,b,l) \quad B(a,b,l) = \min_{b<\mathbf{c}<l} C(a,b,c,l) \quad C(a,b,c,l) = \min_{c<\mathbf{d}<l} D(a,b,c,d,l)$$

$$D(a,b,c,d,l) = \min_{d<\mathbf{e}<l} E(a,b,c,d,e,l) \quad E(a,b,c,d,e,l) = \min_{e<\mathbf{f}<l} F(\cancel{a},\cancel{b},c,d,\cancel{e},f,l) + \Delta G(a,b,e,f)$$

$$F(c,d,f,l) = \min_{f<\mathbf{g}<l} G(c,d,f,g,l) \quad G(c,d,f,g,l) = \min_{g<\mathbf{h}<l} H(c,d,f,g,h,l)$$

Kissing hairpins: an example





Option 2: Left to right strategy, starting from $(a, l) \longrightarrow MFE := \min_{\mathbf{a}<\mathbf{l}} A(a, l)$

Add anchors from left to right, forget unnecessary anchors:

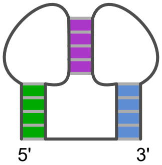
$$A(a,l) = \min_{a<\mathbf{b}<l} B(a,b,l) \quad B(a,b,l) = \min_{b<\mathbf{c}<l} C(a,b,c,l) \quad C(a,b,c,l) = \min_{c<\mathbf{d}<l} D(a,b,c,d,l)$$

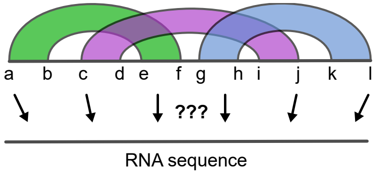
$$D(a,b,c,d,l) = \min_{d<\mathbf{e}<l} E(a,b,c,d,e,l) \quad E(a,b,c,d,e,l) = \min_{e<\mathbf{f}<l} F(\cancel{a},\cancel{b},c,d,\cancel{e},f,l) + \Delta G(a,b,e,f)$$

$$F(c,d,f,l) = \min_{f<\mathbf{g}<l} G(c,d,f,g,l) \quad G(c,d,f,g,l) = \min_{g<\mathbf{h}<l} H(c,d,f,g,h,l)$$

$$H(c,d,f,g,h,l) = \min_{h<\mathbf{i}<l} I(c,d,f,g,h,i,l)$$

Kissing hairpins: an example





Option 2: Left to right strategy, starting from $(a, l) \longrightarrow MFE := \min_{\mathbf{a}<\mathbf{l}} A(a, l)$

Add anchors from left to right, forget unnecessary anchors:

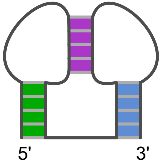
$$A(a,l) = \min_{a<\mathbf{b}<l} B(a,b,l) \quad B(a,b,l) = \min_{b<\mathbf{c}<l} C(a,b,c,l) \quad C(a,b,c,l) = \min_{c<\mathbf{d}<l} D(a,b,c,d,l)$$

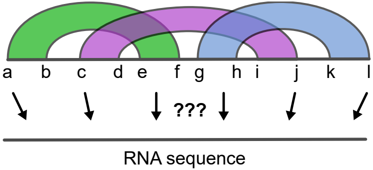
$$D(a,b,c,d,l) = \min_{d<\mathbf{e}<l} E(a,b,c,d,e,l) \quad E(a,b,c,d,e,l) = \min_{e<\mathbf{f}<l} F(\cancel{a},\cancel{b},c,d,\cancel{e},f,l) + \Delta G(a,b,e,f)$$

$$F(c,d,f,l) = \min_{f<\mathbf{g}<l} G(c,d,f,g,l) \quad G(c,d,f,g,l) = \min_{g<\mathbf{h}<l} H(c,d,f,g,h,l)$$

$$H(c,d,f,g,h,l) = \min_{h<\mathbf{i}<l} I(c,d,f,g,h,i,l) \quad I(c,d,f,g,h,i,l) = \min_{i<\mathbf{j}<l} J(\cancel{c},\cancel{d},f,g,h,\cancel{i},j,l) + \Delta G(c,d,i,j)$$

Kissing hairpins: an example





Option 2: Left to right strategy, starting from $(a, l) \longrightarrow MFE := \min_{\mathbf{a}<\mathbf{l}} A(a, l)$

Add anchors from left to right, forget unnecessary anchors:

$$A(a,l) = \min_{a<\mathbf{b}<l} B(a,b,l) \quad B(a,b,l) = \min_{b<\mathbf{c}<l} C(a,b,c,l) \quad C(a,b,c,l) = \min_{c<\mathbf{d}<l} D(a,b,c,d,l)$$

$$D(a,b,c,d,l) = \min_{d<\mathbf{e}<l} E(a,b,c,d,e,l) \quad E(a,b,c,d,e,l) = \min_{e<\mathbf{f}<l} F(\cancel{a},\cancel{b},c,d,\cancel{e},f,l) + \Delta G(a,b,e,f)$$

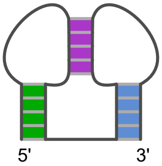
$$F(c,d,f,l) = \min_{f<\mathbf{g}<l} G(c,d,f,g,l) \quad G(c,d,f,g,l) = \min_{g<\mathbf{h}<l} H(c,d,f,g,h,l)$$

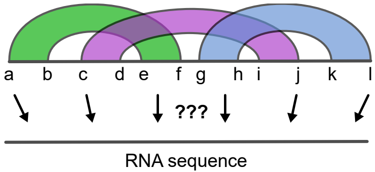
$$H(c,d,f,g,h,l) = \min_{h<\mathbf{i}<l} I(c,d,f,g,h,i,l) \quad I(c,d,f,g,h,i,l) = \min_{i<\mathbf{j}<l} J(\cancel{c},\cancel{d},f,g,h,\cancel{i},j,l) + \Delta G(c,d,i,j)$$

$$J(f,g,h,j,l) = \min_{j<\mathbf{k}<l} \Delta G(g,h,k,l)$$

$\rightarrow \mathcal{O}(n^8)$ time complexity (meh!), can we do better?

Treewidth and tree decomposition(s)





Option 1: Brute force $MFE := \min_{\mathbf{a}<\mathbf{b}<\ldots<\mathbf{l}} \Delta G(a,b,e,f) + \ldots \longrightarrow$ **a b c d e f g h i j k l**

Option 2: Left to right

$$MFE := \min_{\mathbf{a}<\mathbf{l}} A(a,l)$$

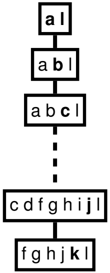
$$A(a,l) = \min_{a<\mathbf{b}<l} B(a,b,l)$$

$$B(a,b,l) = \min_{b<\mathbf{c}<l} C(a,b,c,l)$$

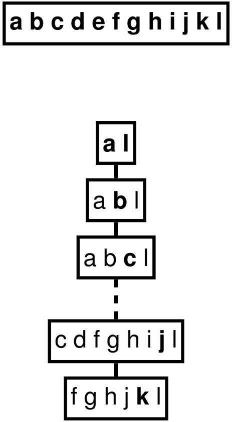
$$\vdots$$

$$I(c,d,f,g,h,i,l) = \min_{i<\mathbf{j}<l} J(f,g,h,j,l) + \Delta G(c,d,i,j)$$

$$J(f,g,h,j,l) = \min_{j<\mathbf{k}<l} \Delta G(g,h,k,l)$$



Treewidth and tree decomposition



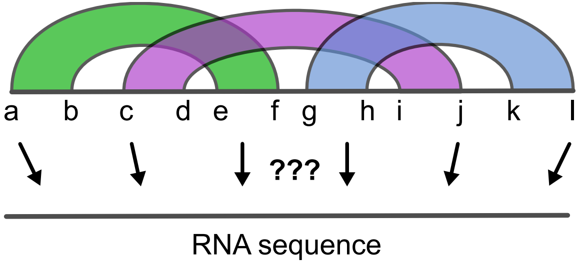
Definition: A tree decomposition for $G = (V, E)$ is a tree T over bags $\{X_i\}_i$, $X_i \subseteq V$, such that

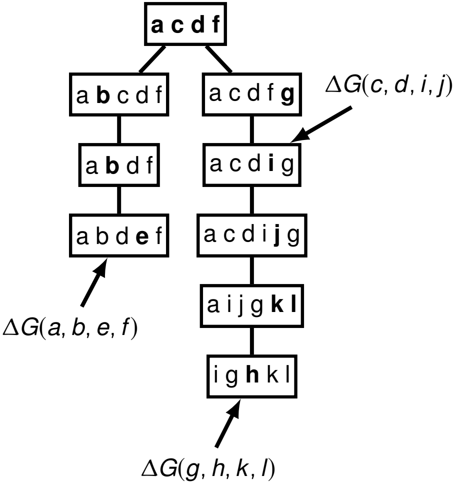
- Any $v \in V$ appears in some bag X_i
- $\forall$ edge $(u, v) \in E$, $\exists$ bag X_i such that $\{u, v\} \subseteq X_i$
- For any $v \in V$, the restriction of T to bags that contain v is a (connected) subtree

Treewidth tw = Size of largest bag -1

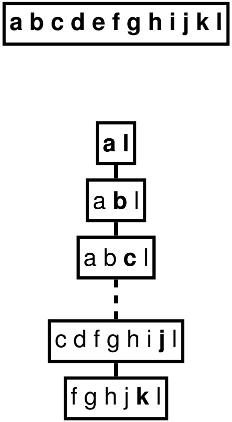
DP scheme for MFE anchors mapping in $\mathcal{O}((n + |E|).n^{tw+1})$
(XP algorithm, NP-hard from k-clique, probably not FPT)

Kissing hairpins: Alternative decomposition





Treewidth and tree decomposition



Definition: A tree decomposition for $G = (V, E)$ is a tree T over bags $\{X_i\}_i$, $X_i \subseteq V$, such that

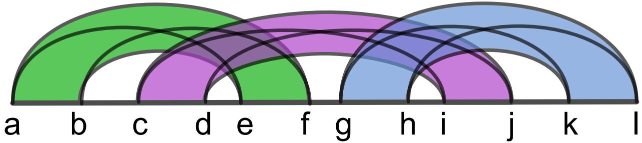
- Any $v \in V$ appears in some bag X_i
- $\forall$ edge $(u, v) \in E$, $\exists$ bag X_i such that $\{u, v\} \subseteq X_i$
- For any $v \in V$, the restriction of T to bags that contain v is a (connected) subtree

Treewidth *tw* = Size of largest bag −1

Theorem (Bodlaender *et al*): OPT (min treewidth) decomposition can be computed in time $O(n.2^{O(tw^3)})$

$\Rightarrow$ Optimal* DP schemes for MFE from OPT tree dec.??

Towards general (unbounded) structures



Property: Presence of bags for (4 anchors) helices could be ensured by cliques...

Helices can be generated/scored inwards/outwards, saving $\mathcal{O}(n)$ time/space

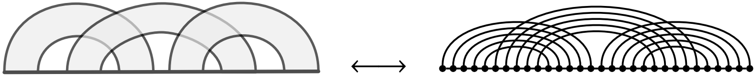
$\Rightarrow$ Instantiating shadow enables tree dec. to benefit from speedup

Problem: How to find generalizable tree decomposition/prevent shortcut?

Five base pairs (BPs) is all you need

Definitions:

- Structure space $\mathcal{S}_{\mathcal{F}}$ of fatgraph $\mathcal{F}$: (infinite) set of structures factoring into $\mathcal{F}$;
- Minimal expansion $\mathcal{C}_{\mathcal{F}}$ of $\mathcal{F}$: Structure with each helix expanded into 5 BPs.



Theorem: Contracting 5+ ladders to 5 base pairs preserves treewidth

Bound is tight as reduction to 4 may reduce treewidth

Corollary: $\mathcal{C}_{\mathcal{F}}$ has treewidth larger or equal to any structure in $\mathcal{S}_{\mathcal{F}}$

Intuition: Exploit tree dec. of $\mathcal{C}_{\mathcal{F}}$ to make general statements about folding over $\mathcal{S}_{\mathcal{F}}$

From canonical structures to DP algorithms

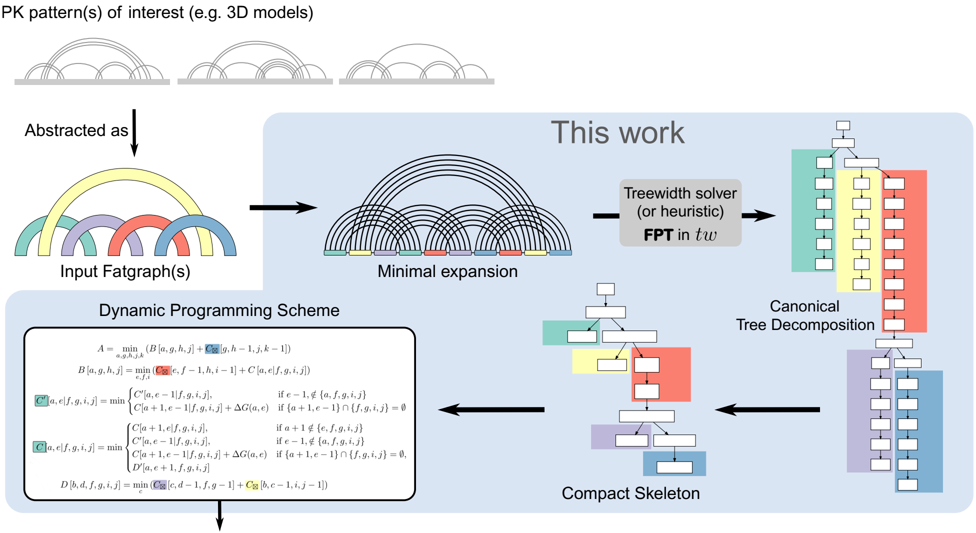
Theorem: Tree Decomposition of minimal expansion can be transformed into a TD (same width) partitioning helices into diagonals and cliques

- Cliques = bags featuring 4 helix anchors
- Diagonals = *triangulated* helices, amenable to inward/outward helix extension
- Transitionals = other bags, faithfully/boringly propagating anchor values

+ Generic DP rules for diagonals and cliques helices contribute energies under BP, Stacking or Turner models

Corollary: MFE folding over the structure space factoring into a given fatgraph $\mathcal{F}$ can be solved in complexity $\mathcal{O}(n^{tw+1})$

Putting everything together



$$A = \min_{a,g,h,j,k} \left(B\,[a,g,h,j] + C_{\boxtimes}[g,h-1,j,k-1]\right)$$

$$B\,[a,g,h,j] = \min_{e,f,i} \left(C_{\boxtimes}[e,f-1,h,i-1] + C\,[a,e|f,g,i,j]\right)$$

$$C'[a,e|f,g,i,j] = \min \begin{cases} C'[a,e-1|f,g,i,j], & \text{if } e-1, \notin \{a,f,g,i,j\} \\ C[a+1,e-1|f,g,i,j] + \Delta G(a,e) & \text{if } \{a+1,e-1\} \cap \{f,g,i,j\} = \emptyset \end{cases}$$

$$C[a,e|f,g,i,j] = \min \begin{cases} C[a+1,e|f,g,i,j], & \text{if } a+1 \notin \{e,f,g,i,j\} \\ C'[a,e-1|f,g,i,j], & \text{if } e-1, \notin \{a,f,g,i,j\} \\ C[a+1,e-1|f,g,i,j] + \Delta G(a,e) & \text{if } \{a+1,e-1\} \cap \{f,g,i,j\} = \emptyset, \\ D'[a,e+1,f,g,i,j] \end{cases}$$

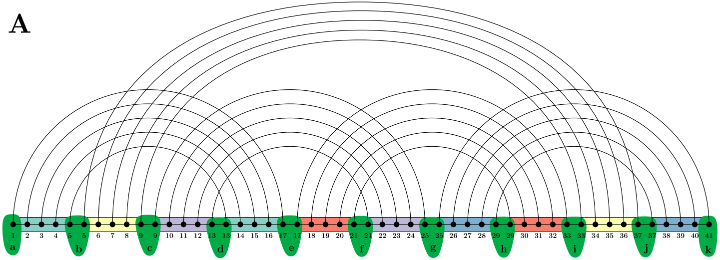
$$D\,[b,d,f,g,i,j] = \min_{c} \left(C_{\boxtimes}[c,d-1,f,g-1] + C_{\boxtimes}[b,c-1,i,j-1]\right)$$

General algorithm for Fatgraph MFE folding problem
Correct for any input sequence (+ `C/C++` Code generation)

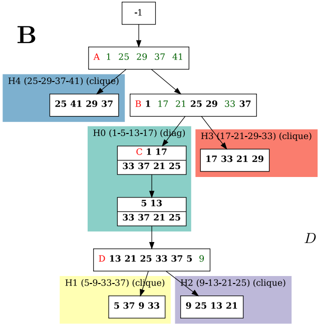
Complexity: $O(n^{tw})$ or $O(n^{tw+1})$ for simple energy models
$O(n^{tw+1})$ for full Turner model

Example

A



B



C

$$A = \min_{a,g,h,j,k} \left(B\,[a,g,h,j] + C_{\boxtimes}[g,h-1,j,k-1]\right)$$

$$B\,[a,g,h,j] = \min_{e,f,i} \left(C_{\boxtimes}[e,f-1,h,i-1] + C\,[a,e|f,g,i,j]\right)$$

$$C\,[a,e|f,g,i,j] = \min \begin{cases} C[a+1,e|f,g,i,j], \\ C[a,e-1|f,g,i,j], \\ C[a+1,e-1|f,g,i,j] + \Delta G(a,e), \\ D[a,e+1,f,g,i,j] \end{cases}$$

$$D\,[b,d,f,g,i,j] = \min_{c} \left(C_{\boxtimes}[c,d-1,f,g-1] + C_{\boxtimes}[b,c-1,i,j-1]\right)$$

Some results

			Complexities	
Type	Fatgraph	Treewidth	Full Turner	All others
H-type	([)]	4	$O\left(n^5\right)$	$O\left(n^4\right)$
Kissing hairpins	([)(])	4	$O\left(n^5\right)$	$O\left(n^4\right)$
"L" [Reidys et al, 2011]	([{)]}	5	$O\left(n^6\right)$	$O\left(n^6\right)$
"M" [Reidys et al, 2011]	([{)(]})	5	$O\left(n^6\right)$	$O\left(n^6\right)$
4-clique	([{<)]}>	5	$O\left(n^6\right)$	$O\left(n^6\right)$
5-clique	([{<A)]}>a	5	$O\left(n^6\right)$	$O\left(n^6\right)$
5-chain	({[)(][)}]	6	$O\left(n^7\right)$	$O\left(n^7\right)$

Extensions

- Embedding PKs/fatgraphs occurrences in 2D centric search spaces

  $\rightarrow$ Simply add a ghost edge, and *voila*: DP scheme now ready for inclusion within `RNAfold` closed structure table

- Computing partition function/thermodynamic quantities

  Generated recursions already unambiguous so $(\min, +) \rightarrow (\sum, \times)$ gives $\mathcal{Z}$
  Multiple fatgraphs $\rightarrow$ Disjoint search spaces
  Inside/outside (BP probabilities) obtainable with (a little) more sweat

- Recursive pseudoknots: Consecutive anchors *seen* together

  $\rightarrow$ Possibility to insert other 2D/PK structures in *unpaired* regions
  If non-recursive & affine energy model for PK unpaired regions, merge consecutive anchors (Gotoh trick) for reduced treewidth

- RNA-RNA interactions have fatgraphs too!

Musings

- Optimality of complexities?

  For which search space? Lower bounds are hard. . .

- Polluting conformation spaces! Konformationsraumverschmutzung???

  Monkeys & Typewriters $\longrightarrow$ Adding PK motif implies exponential growth of $\mathcal{Z}$
  Geometry: Terminal PKs are OK, recursive PKs are dangerous

- Tree decomposition vs Parsing (fight!)

Acknowledgments

Sebastian Will

Sarah Berkemer

Bertrand Marchand

AMIBio

Laurent Bulteau

+ Ronny (!!) and all members of PaRNAssus ANR/FWF project