RNA Folding with Pseudoknots

A Topological Approach

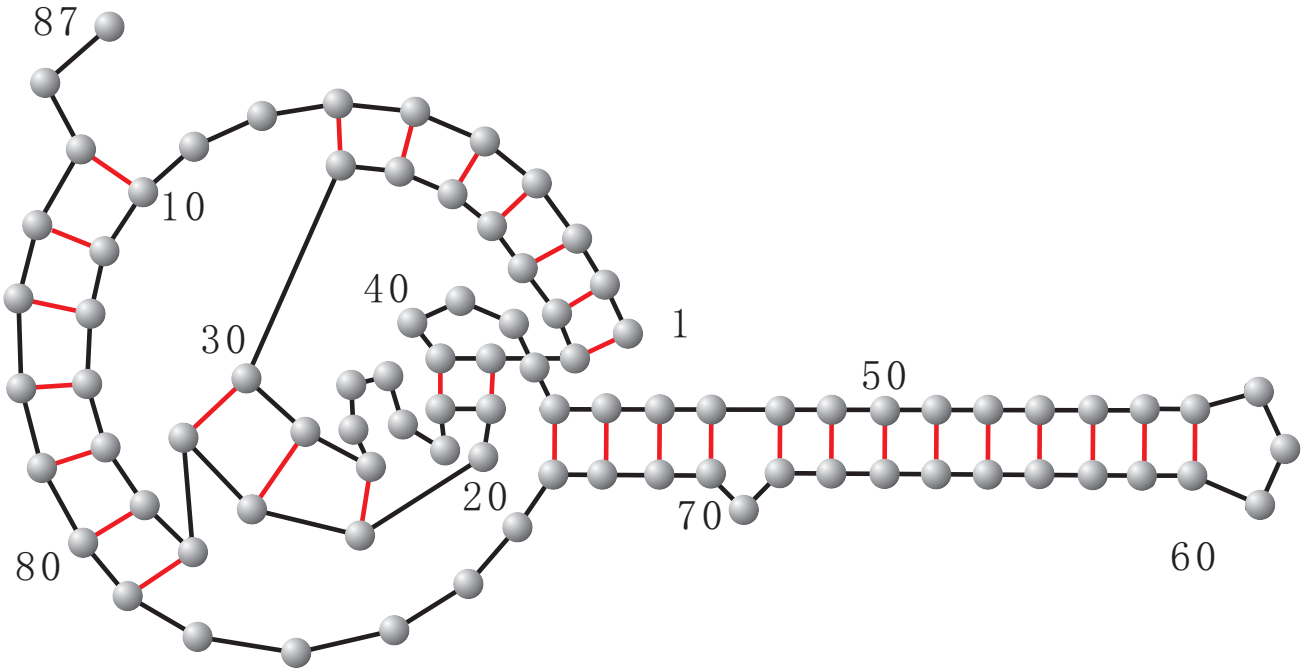
Peter F. Stadler

joint work with
Christian M. Reidys (Tianjin & Odense), Fenix D. Huang (Tianjin),
Jørgen E. Andersen, Robert C. Penner (Århus)
Markus E. Nebel (Kaiserslautern)

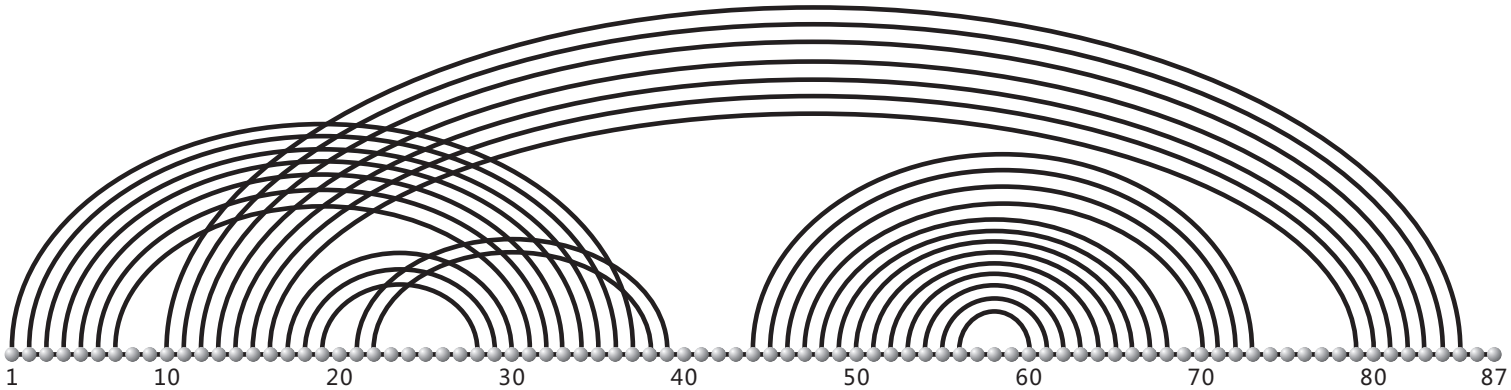
Bioinformatics Group, Dept. of Computer Science &
Interdisciplinary Center for Bioinformatics,
University of Leipzig
Max Planck Institute for Mathematics in the Sciences
RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology
Institute for Theoretical Chemistry, Univ. of Vienna (external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

Benasque, Aug 01 2012

Pseudoknotted RNA Structure



(a)



Folding of Pseudoknotted RNA Structures

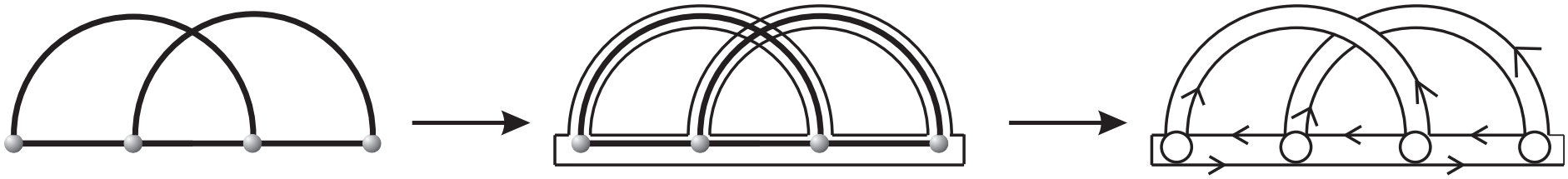
- The general problem can be see as a Maximum Matching Problem.
  not very useful since
  1. incompatible with 3D structures (too many crossing contacts)
  2. energy should depend on base pairing and loops
- The Stacking-based problem is NP hard
- Dynamic Programming algorithms have been devised for a large number of special classes of structures that have been chosen because of computational simplicity rather only
  1. Lygsø-Pedersen = Dirks-Pierce
  2. Akutsu-Uemura
  3. Uemura *et al.*
  4. Rivas-Eddy
  5. Cao-Chen
  6. Chen-Condon-Jabbari

  Mutual relations studied by Condon (2004), Nebel (2011).

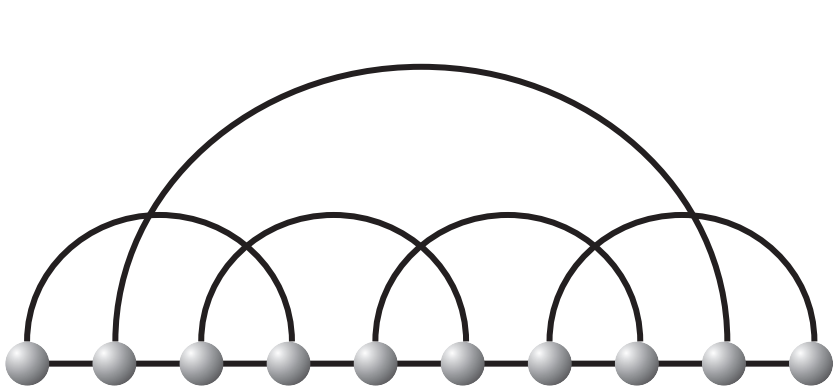
Classification of Pseudoknotted Structures

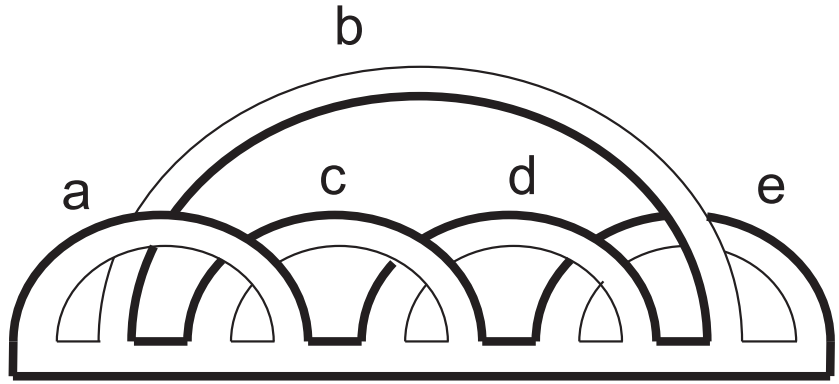
1. k-book-embeddable structure (Haslinger & Stadler)
   The structure is a superposition of at most k secondary structures.
   Non-recursive, no algorithms known.
2. k-non-crossing structures (Reidys)
   There is no subset of k base pairs in which each pair of pairs is crossing.
   Folding via enumeration of prototype structures, exponential in time and space
3. k-structures composed of irreducible components of topological genus at most k.
   Vernizzi, Orland, Bon and collaborators
   Nebel, Reidys, PFS, and collaborators

Secondary Structure $\longrightarrow$ Fat Graphs



Inflation of edges and vertices to ribbons and disks





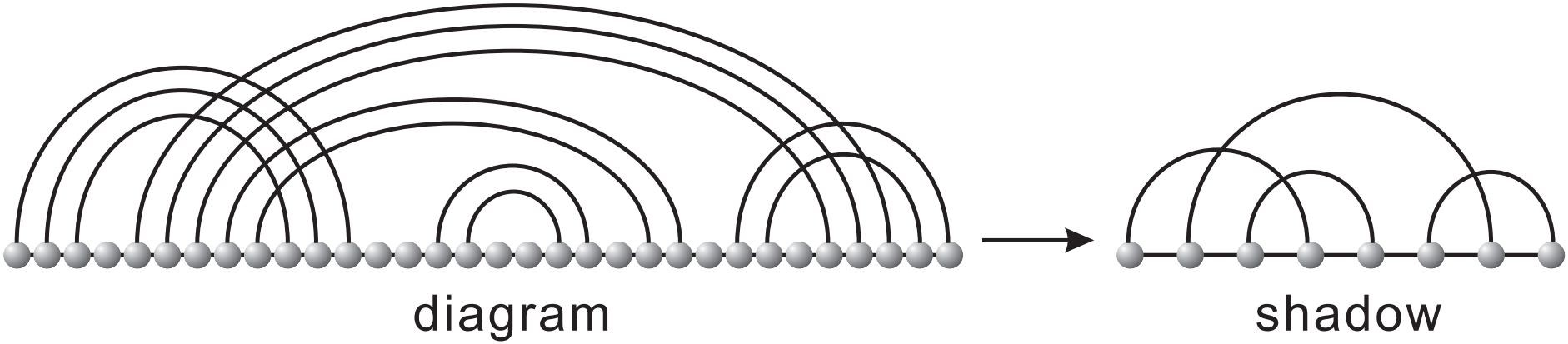
Computing the number of boundary components: $v = 10$ vertices, $e = 5 + 9$ edges; paths alternating between arc and backbone: $r = 2$ "boundary components".

Topological genus

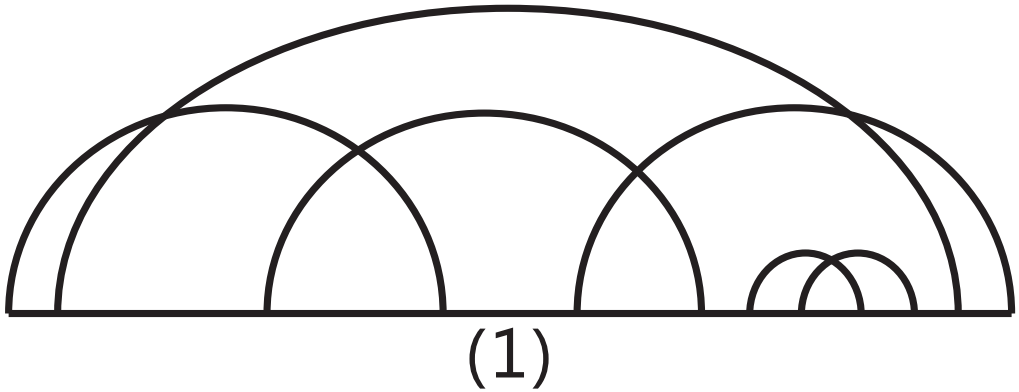
$$g = 1 - \frac{1}{2}(v - e + r) = 1 - (10 - 14 + 2)/2 = 2$$

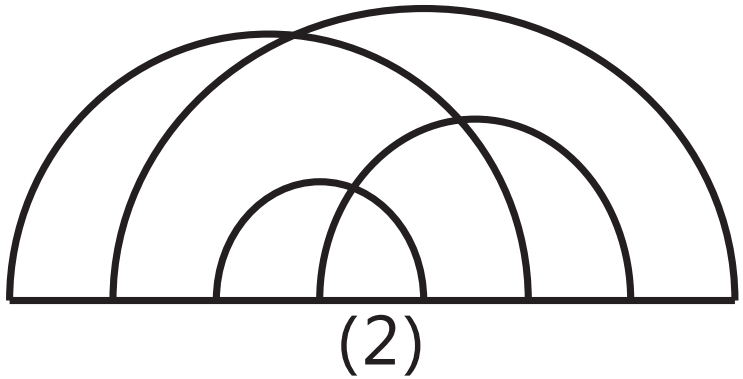
Orland *et al*: energy penalty proportional to g.

Shadows and γ-structures



shadow: collaps nested arcs $\Leftrightarrow$ Robert Giegerich's shapes





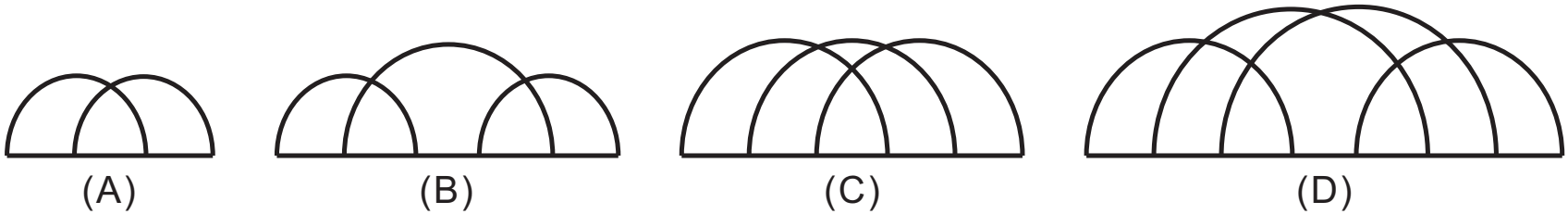
$\gamma = 1, g = 2$

$\gamma = 2, g = 2$

Classification Theorem

(the results for $\gamma = 1$ have been obtained by Orland and co-workers using a very different approach)

- A structure is a 0-structure if and only if it is (simple) secondary structure
- A structure is a 1-structure if and only if its shadow can be decomposed by iteratively removing one of the four shadows



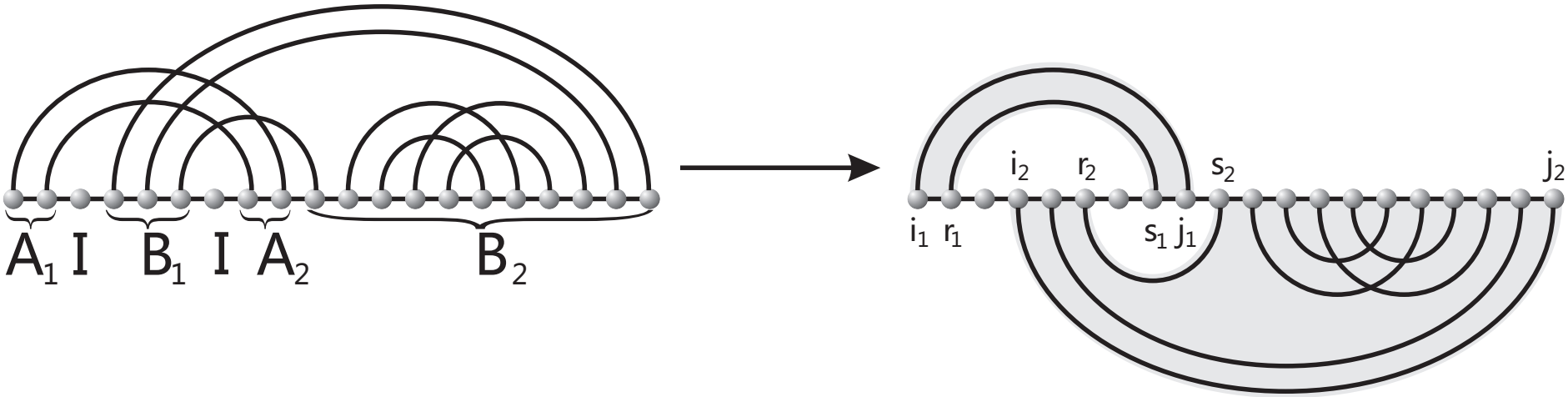
- A structure is a γ structure if and only if its shadow can be decomposed by iteratively removing shadows of genus at most γ.
- This set of distinct shadows is always finite for given γ.

Known Pseudoknots

With a single exception (HDV genome) all known RNA structures are 1-structures.
The genus is the sum of the genera of the irreducible components.
Biological sequences my have large genus, e.g., when they contain multiple pseudoknots.
The classification theorem suggests an dynamic programming approach.

Fragment Pairs

... are in essence the “gap matrices” of Rivas&Eddy

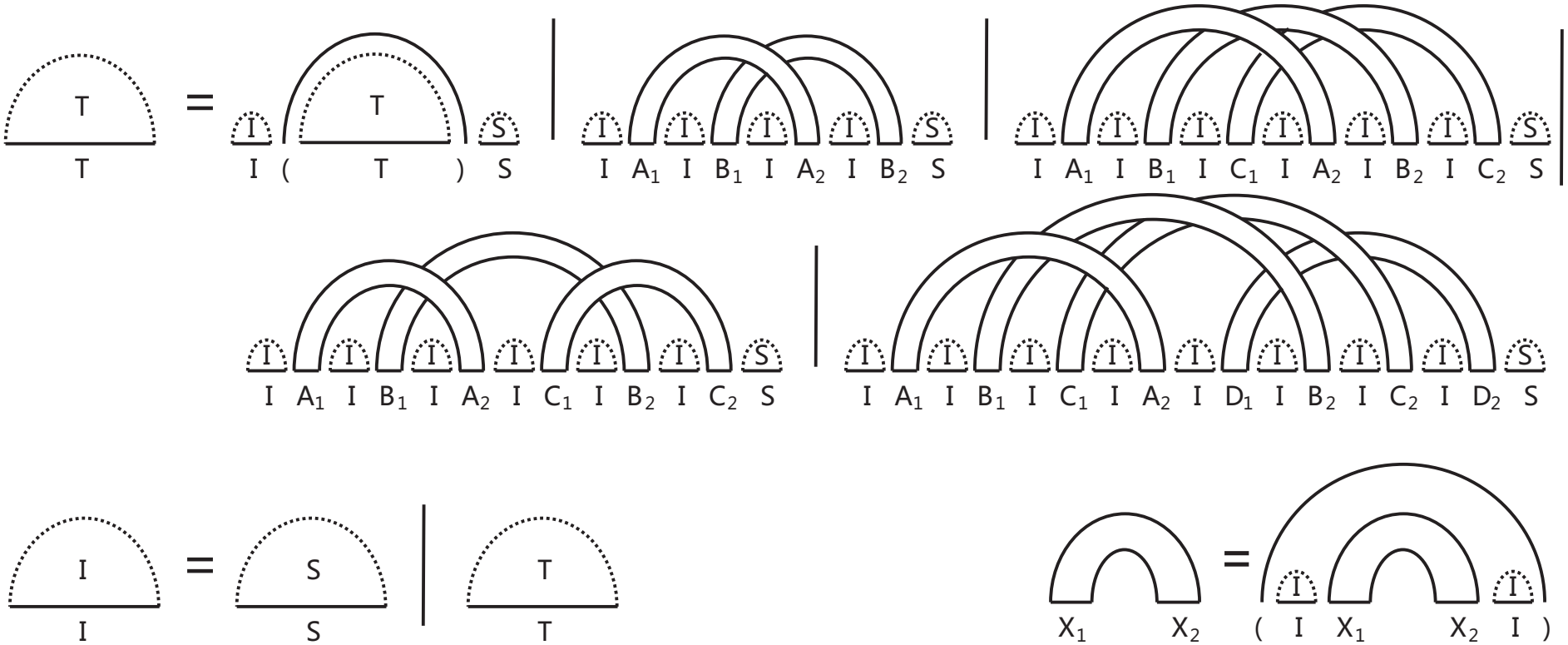


Rule $I \rightarrow IA_1IB_1IA_2IB_2S$ induces the fragment-pairs $[i_1, r_1]$, $[s_1, j_1]$ and $[i_2, r_2]$, $[s_2, j_2]$.
Arcs connecting the two fragments of a pair are non-crossing, while arcs with both endpoints within the same fragment may be crossing

... such as those within $[s_2, j_2]$.

Naïve Algorithm

Multi-Context-Free Grammar



useless in this form: $O(n^{18})$ time and $O(n^4)$ space

Naïve Algorithm: MCFG Form

$$
\begin{aligned}
I &\rightarrow S \mid T \\
S &\rightarrow (S)S \mid \ :S \mid \epsilon \\
T &\rightarrow I(T)S \\
T &\rightarrow IA_1IB_1IA_2IB_2S \\
T &\rightarrow IA_1IB_1IA_2IC_1IB_2IC_2S \\
T &\rightarrow IA_1IB_1IC_1IA_2IB_2IC_2S \\
T &\rightarrow IA_1IB_1IC_1IA_2ID_1IB_2IC_2ID_2S \\
\vec{X} &\rightarrow [(_X IX_1, X_2 I)_X] \mid [(_X,)_X],
\end{aligned}
$$

where $X \in \{A, B, C, D\}$ distinguishes the four types of pseudoknots.

More efficiency

An $O(n^6)$ and $O(n^4)$ space algorithm is obtainable by tabulating intermediate results, i.e., introducing additional non-terminals

$$\begin{aligned}
\vec{U} &\rightarrow [IX_1, IX_2] \\
\vec{V} &\rightarrow [U_1U_1', U_2U_2'] \\
\vec{W} &\rightarrow [U_1, U_1'U_2U_2'] \mid [V_1, U_1V_2U_2]
\end{aligned}$$

where (U_1', U_2') is a marked copy of (U_1, U_2) used to identify the components which must later be expanded in a coupled way.

$$\begin{aligned}
T &\rightarrow I(T)S \mid I'S \\
I' &\rightarrow V_1V_2 \mid U_1V_1U_2V_2 \mid U_1W_1U_2W_2
\end{aligned}$$

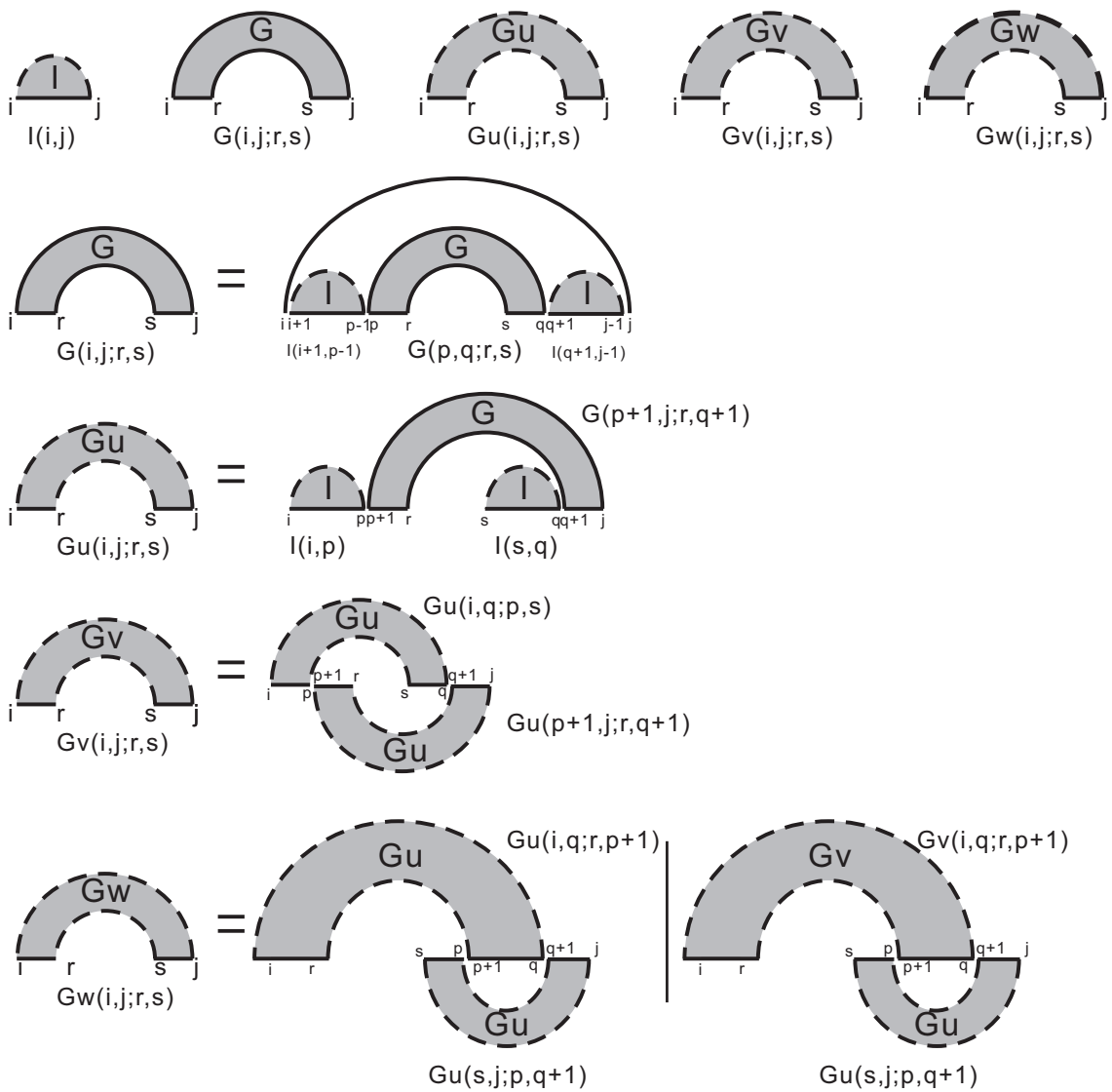
Energy Model?

- Multi-loop-like approach depending on the numbers $\#B$ of base pairs and $\#U$ of unpaired bases forming the pseudoknot.

$$G_{i,j}^{\text{pseudo}} = \beta_X + (\#B + 1) \cdot \beta_2 + \#U \cdot \beta_3,$$

- Pseudoknots in multiloop components or nested within other pseudoknots can get different energy parameters values
- Nice side effect: we can make β_X dependent on the type of pseudoknot.

More efficiency ... in diagrams



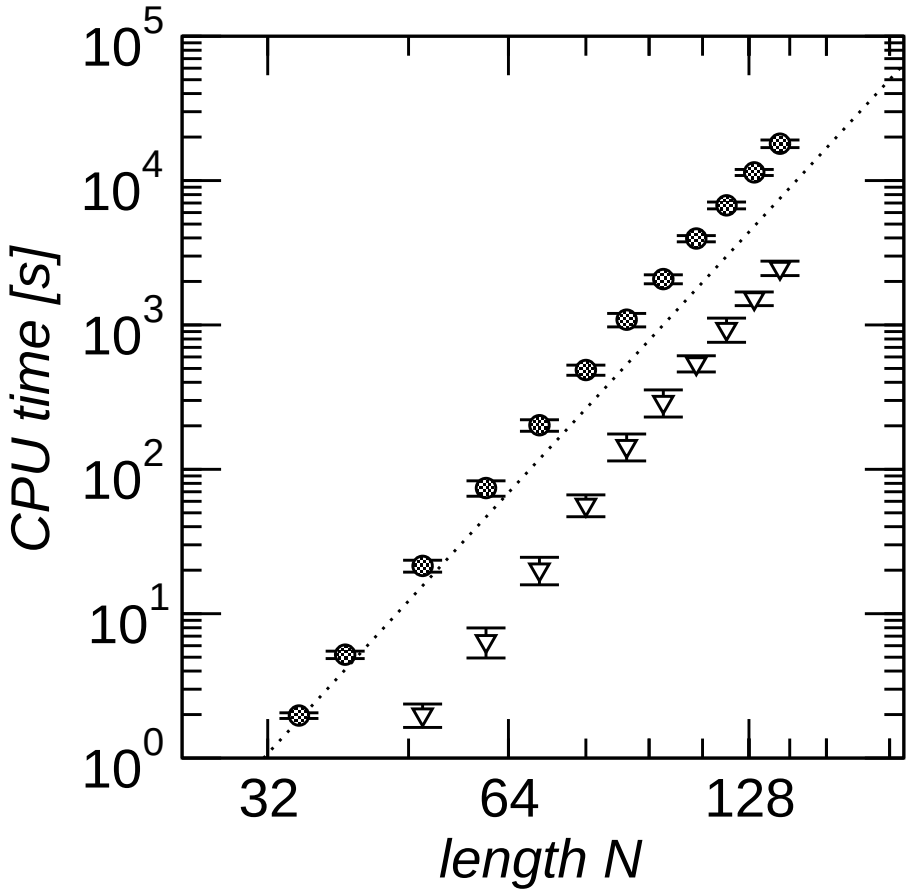
Decomposition for 4-dimensional matrices *G*, *Gu*, *Gv*, and *Gw*.

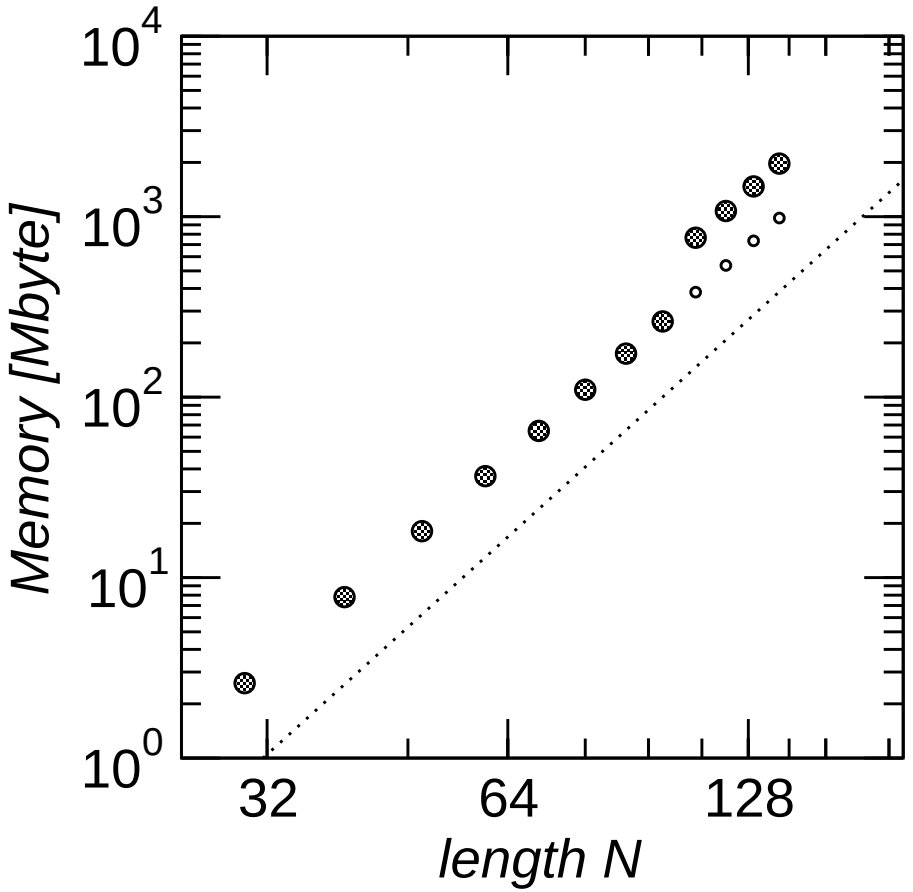
Implementation

- Variations:
  - MFE folding
  - partition function
  - stochastic backtracing
- Energy model: Different penalties for the four topological types of pseudoknots, optimized from known pseudoknots
- Available:
  `http://www.combinatorics.cn/cbpc/gfold.tar.gz`

Important: Pseudoknot penalty dependent on the irreducible diagrams rather than linear dependence on the genus g.

Performance

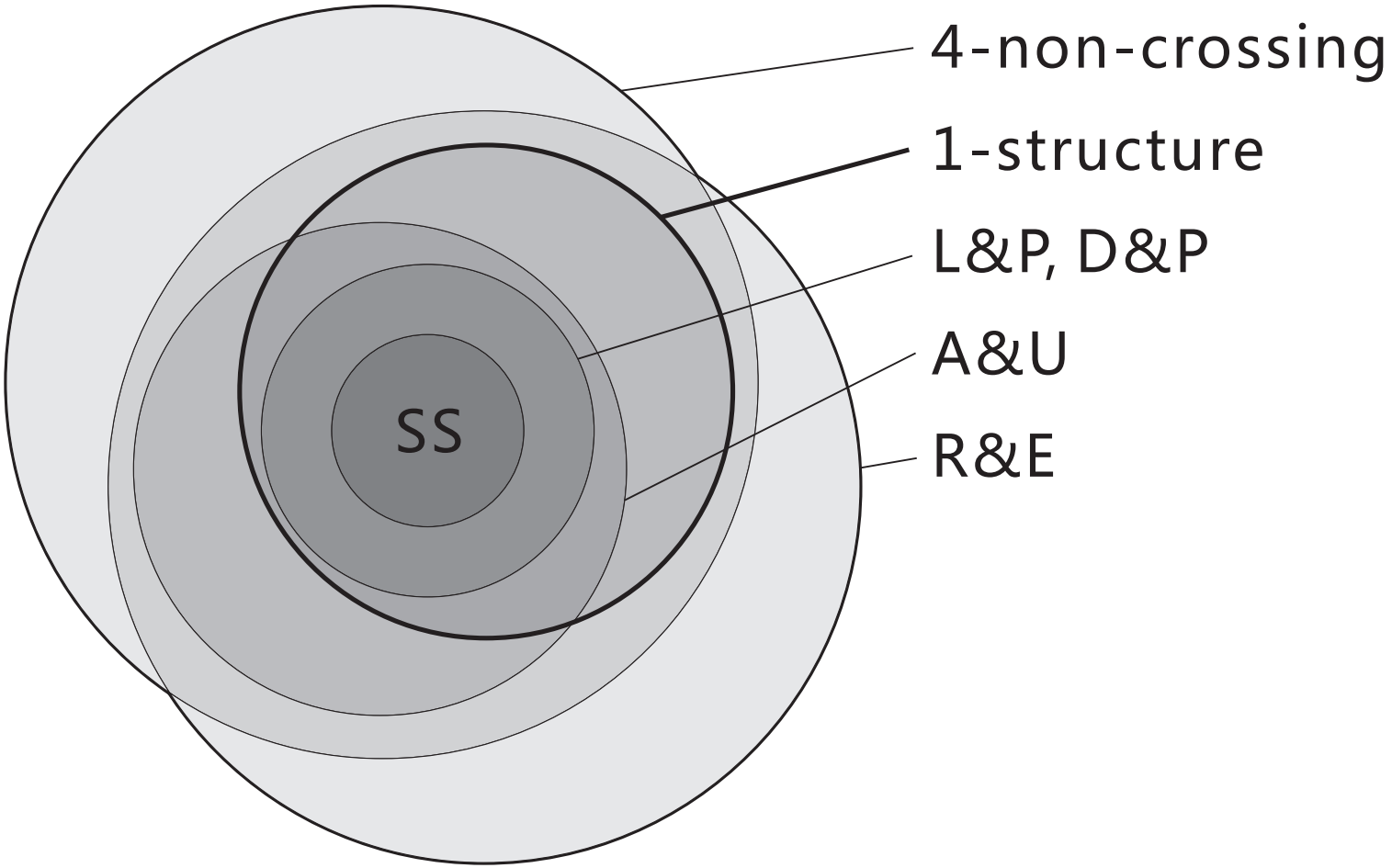




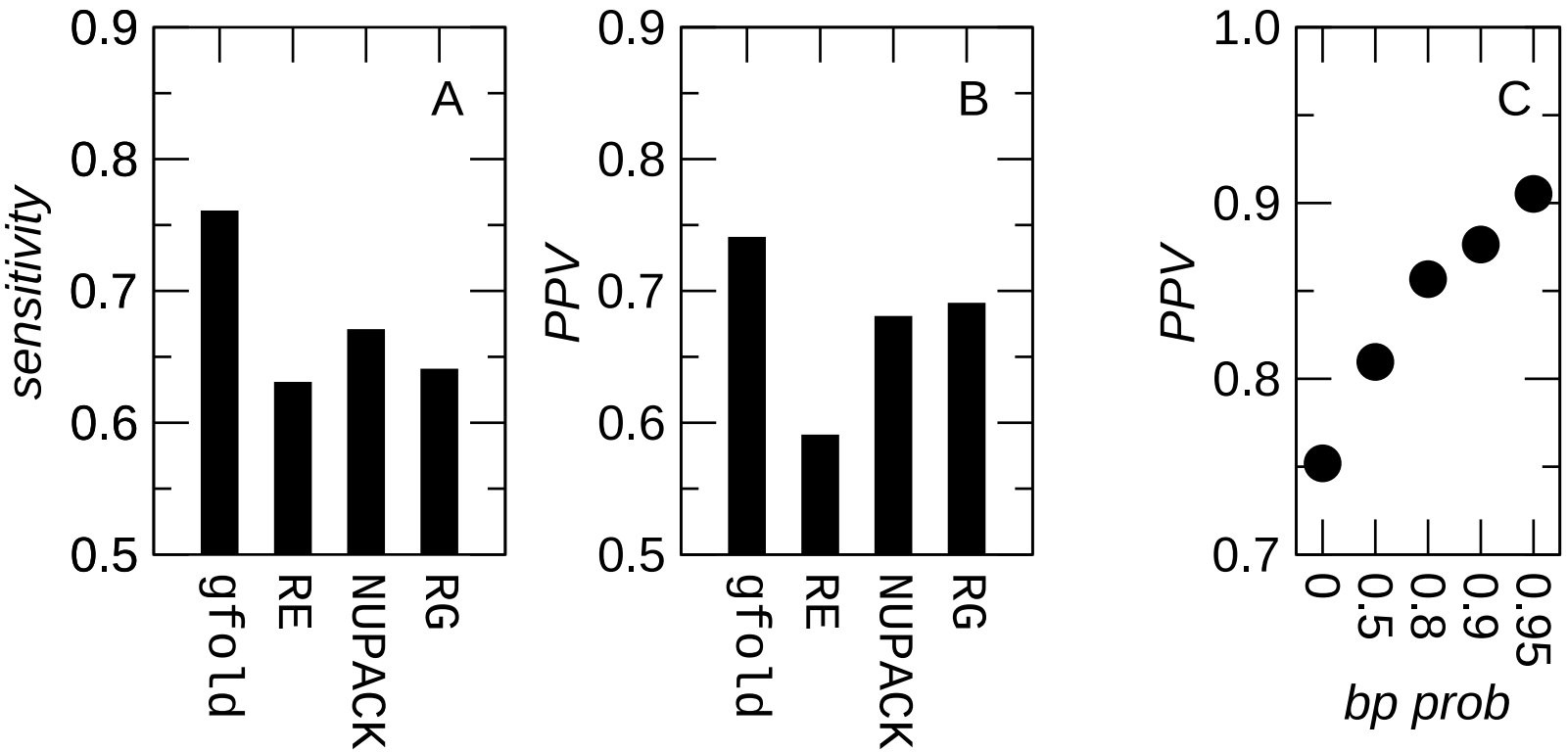
▽ MFE ○ paritition function

Feasible for most RFam families.

Comparison with other pseudoknot classes



Performance

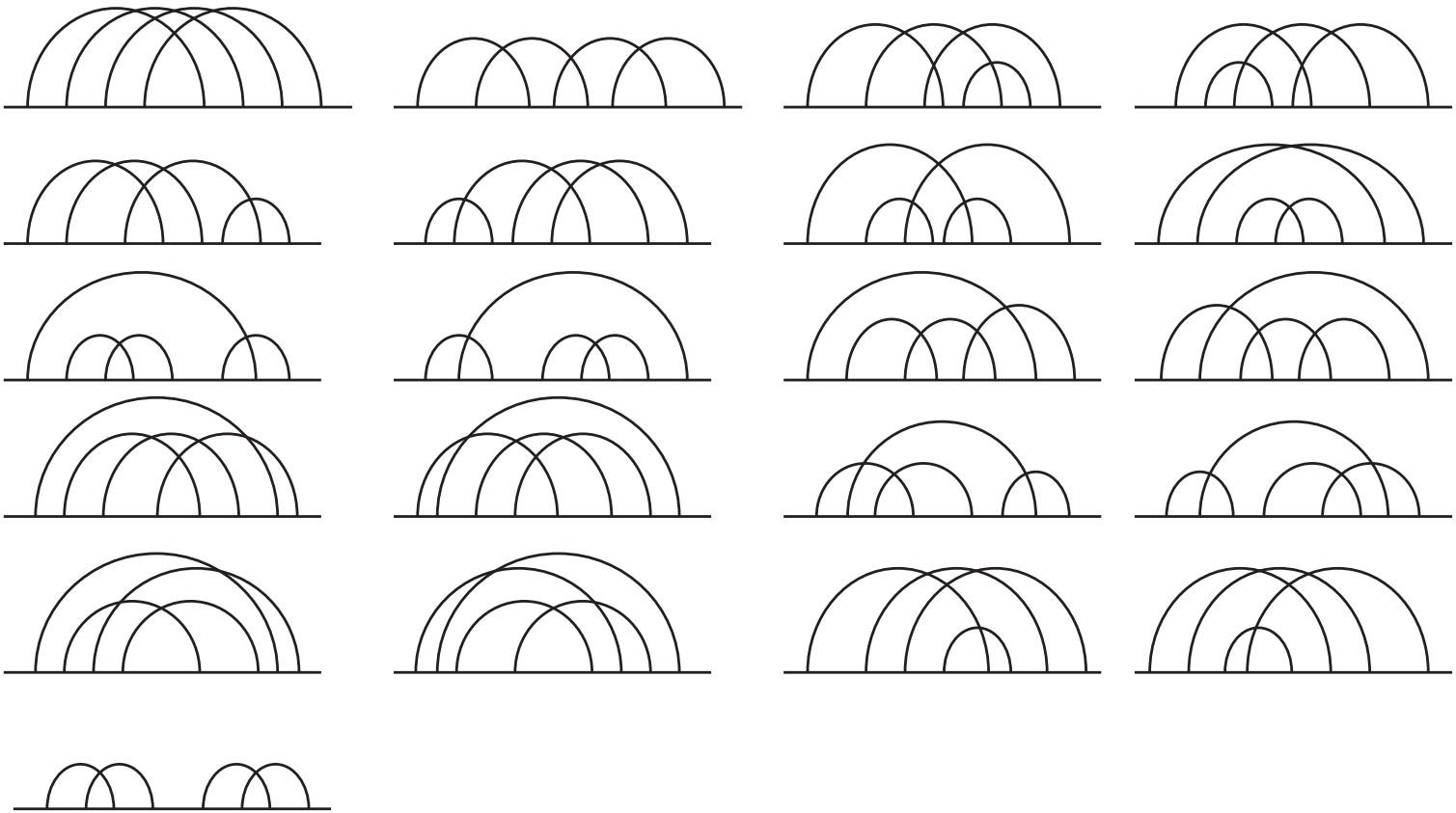


Comparison of the average sensitivity and PPV of different prediction algorithms on a sample of 32 structures from `Pseudobase`.
The PPV increases signficantly if only base pairs with larger pairing probabilities as predicted by the partition function version of `gfold` are included in the predicted structure.

Fewer false positive pseudoknots for pseudoknot-free benchmark structures.

Beyond 1-structures

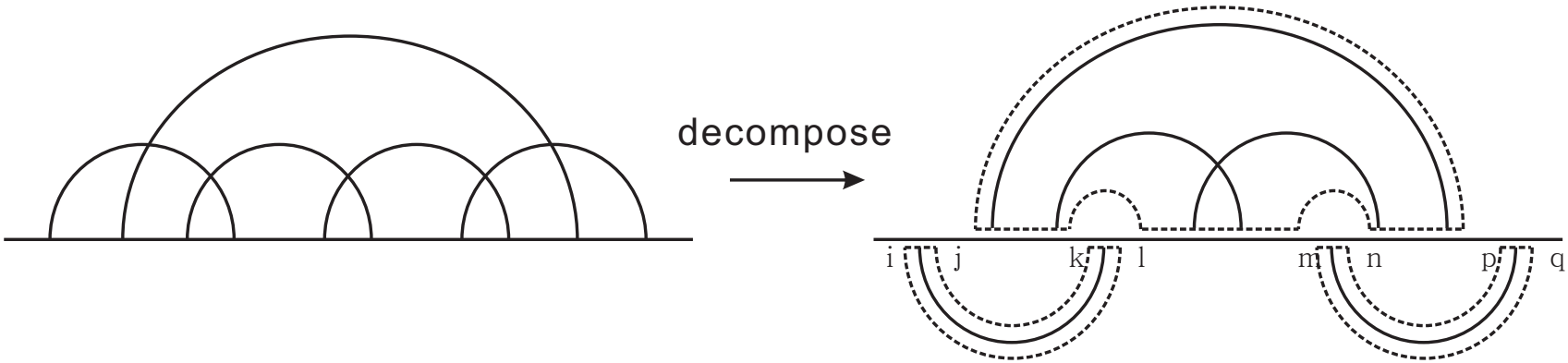
3472 shadows with $\gamma = 2$.



All shapes with 4 arcs, including the reducible ones with $\gamma = 1$.

Algorithms for 2-structures?

The HDV structure is in Rivas&Eddy and hence computable in $\mathcal{O}(n^6)$ time and $\mathcal{O}(n^4)$ space.
This following 2-structure, however, cannot be dealt with in terms of gap matrices.



$\mathcal{O}(n^8)$ time and $\mathcal{O}(n^6)$ space
Unknow if this suffices for all 2-structures