The database of RNA secondary structure elements

E.F. Baulin, M.A. Roytberg

Higher School of Economics, Moscow, Russia; IMPB RAS, Pushino, Russia

baulin@rambler.ru, *mroytberg@lpm.org.ru*

Introduction

We propose a new definition of the loop, which on the one hand is a generalization of Nearest Neighbour Model (NNM), developed by Zucker-Matthews-Turner [1], and on the other hand allows to divide into loops an arbitrary secondary structure, not only the pseudoknot-free structure. Based on the description we have created the database of elements of secondary structures of experimentally determined structures of RNA.

Generalized loops

1. Basic Definitions

A *helix* is a non-extendable sequence of base pairs of the form *(i, j), (i+1, j-1),…, (i+k, j-k), k ≥ 2*. The chain fragment *[i, i + k]* is a *left wing* of the helix, the fragment *[j - k, j]* is a *right wing*.

Pair *(i, j)* is an *external pair* of the helix or a *face*. Pair *(i + k, j – k)* will be called an *internal pair* of the helix.

The position *t* of the chain *belongs to* a helix *H*, if it lies between the nucleotides forming its internal pair and there is no helix *H1*, lying inside *H*, such that $x < t < y$, where *(x, y)* is the face of H1.

Loop of a helix *H* is the set of all positions that belong to the helix *H*. ***There is one-to-one correspondence between loops and helices***

2. Closed Regions and Planes

Let *H* be a helix and *(u, v)* be its internal pair.

Region *[i, j]* is *closed under H* if: *[i, j]* lies inside *H*; there is no such pair *(k, t)* that $(i \le k \le j < t < v)$ or $(u < k < i \le t \le j)$; there are pairs *(i, k)* and *(t, j)*, where $k \le j$; $i \le t$; there is no other than *[i, j]* region *[i', j']* such that $i' \le i < j \le j'$ and the region *[i', j']* satisfies three previous conditions.

Closed region is *simple* if its face is the face of a helix and *complex* otherwise.

A plane is a fragment between two faces. Namely, let (s_1, t_1), …, (s_n, t_n) – faces of all closed under *H* regions; $s_1 < t_1 < \ldots < s_n < t_n$. $t_0 = u$; $s_{n+1} = v$. Then the *k*-th *plane* of the loop of *H* is a fragment $[t_{k-1}+1, s_k - 1]$, here $1 \le k \le n+1$.

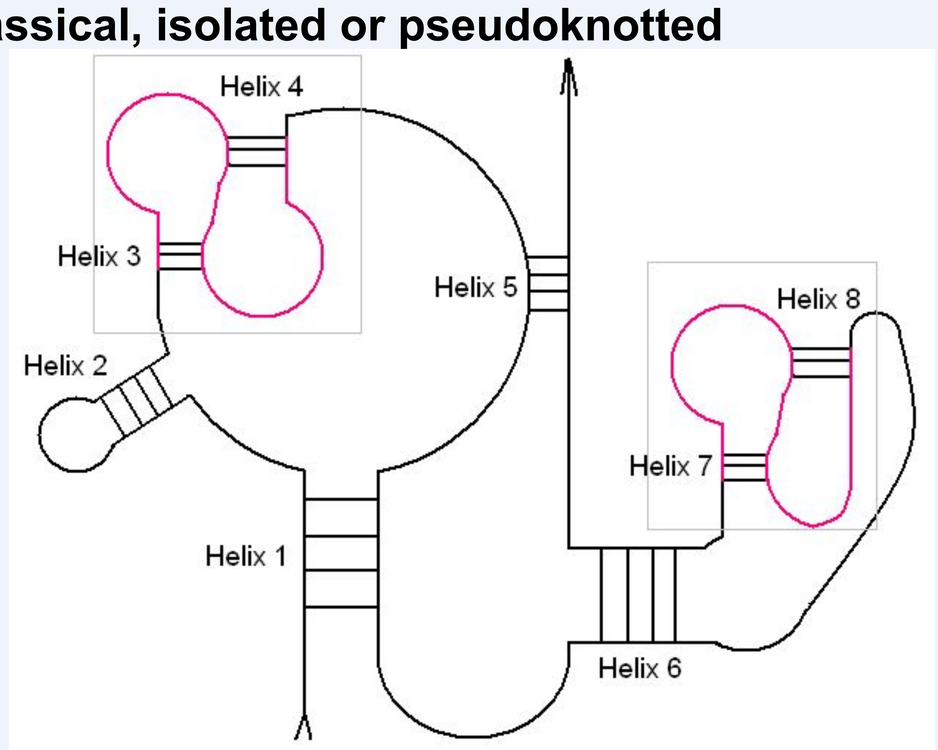
3. Loop Classification

A loop is called *classical* if it does not contain wings and faces of complex closed regions.

A non-classical loop is *isolated* if it does not contain wings and *pseudoknotted* otherwise.

Hairpins, bulges, internal **and** ***multiple*** **loops** are defined as usual depending of number of planes within a loop. They **can be classical, isolated or pseudoknotted**

.



Loops: 1 - pseudoknotted multiple; 2 - classical hairpin; 3,4,7,8 - pseudoknotted hairpin; 5 - pseudoknotted internal; 6 - isolated internal

*Red boxes mark two complex closed regions: 1) from the beginning of Helix1's left wing to the end of the Helix2's right wing, and : 2) from the beginning of Helix7's left wing to the end of the Helix8's right wing.

Database

Input data: all PDB v.3.3 documents containing RNA chains. We have analyzed 6716 models from 1674 documents, 3169 RNA chains were handled (excluding the representation of the same chain in several models). The database contains tables of helices, loops, faces, RNA-Protein contacts, etc. Currently DB contains information about more than 78000 loops, around 149000 non-paired regions (threads), and more than 1 million residue contacts. The beta-version of the database: http://server2.lpm.org.ru/~baulin/home.html.

The web-interface allows one search by documents, loops and different types of contacts between residues or atoms using various arguments.

Table of Loops

	Type/Class	*Hairpin*	*Internal*	*Buldge*	*Multiple*	*Total*
Number	*Classical*	26219	16973	16666	9003	68861
	Isolated	0	122	1175	1217	2514
	Pseudoknotted	4416	674	35	1517	6642
	Total	30635	17769	17876	11737	78017
Total Length	*Classical*	165261	117638	19578	133215	435692
	Isolated	0	788	1362	20028	22178
	Pseudoknotted	42742	18847	320	97278	159187

Example of the Web-interface

Monomer contacts | Atom contacts

PDB file like

Pair type RNA-Protein

Contact type All

Distance (in A.) >

Atom name 1 like

Atom name 2 like

One model per file

Submit

Example of the Web-interface

PDB file like Example: "1.D." (help)

Title like

Date > Format: YYYY-MM-DD

Resolution > Exclude not stated

Rna chains >

Max rna chain length >

Protein chains >

Ligand chains >

Type all

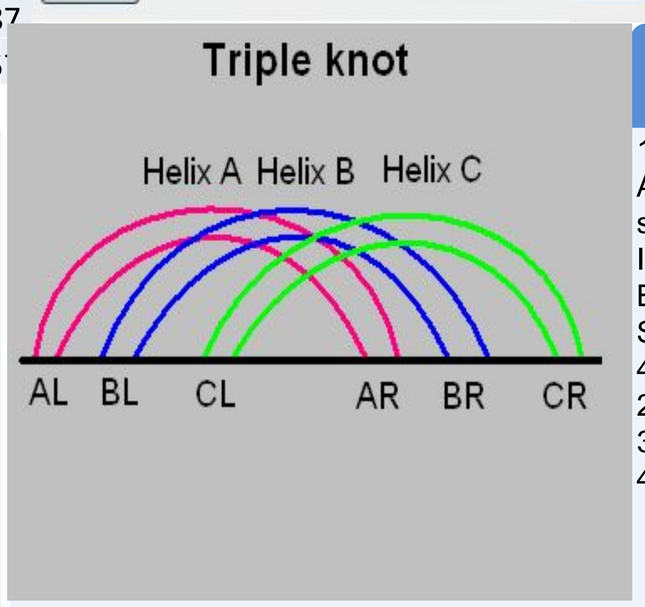
One model per file

Submit

Triple knots

Question: whether there are pseudoknotted structures that cannot be represented as a planar graph?. In other words: can a position belong to three loops? In this case a structure should contain a *triple knot.* A structure of RNA is called *triple knot* if it contains three helices *A, B, C,* their wings are in following order: AL, BL, CL, AR, BR. CR.

A search detected 233 structures that contain a triple knot. All such knots are in homologous to each other sections of 23S RNA; in two of the three helices wings are complementary, and in one - not (in all cases except one, the "bad" helix is A); all helices are usually short (2-3 pairs) in a small number of cases helix C contains 4 pairs



References

1. Zuker M., Mathews D.H., Turner D.H., et al. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: RNA Biochemistry and Biotechnology, J. Barciszewski & B.F.C. Clark, eds., NATO ASI Series, Kluwer Academic Publishers, 1999. P. 42-57.

2. http://x3dna.org/

3.http://www.matbio.org/2012/Baulin_7_567.pdf

4. http://pdb.org/pdb/home/home.doA

We thank S.A.Spirin, D.N. Ivankov and D.V.Khachko for useful discussions.