The database of RNA secondary structure elements

E.F. Baulin,

Institute of Mathematical Problems in Biology, Pushchino, Russia, Higher School of Economics, Moscow, Russia, baulin@rambler.ru

M.A. Roytberg,

Institute of Mathematical Problems in Biology, Pushchino, Russia, Higher School of Economics, Moscow, Russia, mroytberg@lpm.org.ru

We propose a new definition of the loop, which on the one hand is a generalization of the definition of Mathews-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.

We use the following terminology: Helix is a non-extendable sequence of hydrogen bonds of form: *(x, y), (x+1, y-1),....,(x+s, y-s).* Here *(i, j)* is the bond between *i*-th and *j*-th nucleotides of the chain determined according the X3DNA program [2]. The fragment *[x, x + s]* is called a left wing of the helix, the fragment *[y – s, y]* is called a right wing. Pair *(x, y)* is called an external pair of the helix or a face of the helix. The pair *(x + s, y – s)* is an internal pair of the helix. The position of the chain *t* belongs to the 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 the helix *H* is the set of all positions that belong to helix *H*. Thread (or unpaired region) is a chain fragment *[i, j]*, such that each nucleotide from *i* to *j* is unpaired. See [3] for details.

A database of 3D-structures of RNAs and RNA-protein complexes is developed using the proposed classification. The database contains tables of helices, loops, faces, RNA-Protein contacts, etc. Currently we have detailed information about more than 78000 loops and around 149000 threads. Also more than 1 million residue contacts are represented in the database.

The input data were selected from documents of a database of spatial structures PDB (Protein Data Bank, version 3.3) [4]. All selected documents were divided into two groups – containing only RNA chains and containing RNA-protein complexes. Since the certain structures were presented in the same document in several variations, all documents have

been divided according to the principle “one file - one variation of a structure”. One variation of a structure is called model. Currently we have analyzed 6716 models from 1674 documents, 3169 RNA chains were handled (excluding the representation of the same chain in several models). To mark up the hydrogen bonds forming RNA secondary structure function find_pair from toolkit X3DNA (version 1.5) [2] was used.

The beta-version of the database is available at 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.

Creating the database we have studied a question whether there is a pseudoknotted secondary structures that cannot be represented in the form of a planar graph. It was discovered that non-planar graph of structure can take place, if it contains a triple knot. A structure of RNA will be called triple knot if it contains three helices which are in conflict each to each. More precisely, let A, A'; B, B' and C, C' be complementary parts (wings) of helices A, B and C. We say that helices A, B and C form a triple knot if their wings are located on the RNA chain in the following order: A B C A' B' C'. The interest in triple knots is determined by various factors. First, such a structure should be very compact, and it is interesting whether it takes a place in real structures. Second, the question about the presence of RNA triple knots is interesting from the theoretical point of view.

A search detected 233 structures that contain a triple knot. What is more: 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 "wrong" helix is *A*); all helices are usually short (2-3 pairs) in a small number of cases helix *C* contains 4 pairs.

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

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.do