BPFIND

Authors: Jhuma Das, Shayantani Mukherjee, Sukanya Halder, Debasish Mukherjee and Dhananjay Bhattacharyya
Saha Institute of Nuclear Physics
Kolkata 700064, INDIA
E-mail: dhananjay.bhattacharyya@saha.ac.in bhattasinp@gmail.com

This program reads a mmCIF formatted file containing coordinates of nucleic acids and detects all base pairs stabilized by two or more hydrogen bonds between the pairs of bases involved. These hydrogen bonds can give rise to usual canonical Watson-Crick type base pairs or any non-canonical base pairs, all of which are detected. The default criteria for hydrogen bonds are not usual distance and angle cut-offs but in terms of few defined pseudo-bonds and pseudo-angles as hydrogen atoms forming the hydrogen bonds are generally not present in the mmCIF files. The default distance cut-off is set to 3.8A while the default pseudo-angle cut-off is set to 120^{o}. These values were chosen by looking at various structures and comparing detection by other equivalent programs. These can be changed during run-time by using appropriate run-time options. The program was initially described in J. Das, S. Mukherjee, A. Mitra and D. Bhattacharyya (2006) Non-Canonical Base Pairs and Higher Order Structures in Nucleic Acids: Crystal Structure Database Analysis *J. Biomol. Struct. Dynam.* **24**, 149-161 and few updates was presented in M. Chawla, P. Sharma, S. Halder, D. Bhattacharyya and A. Mitra (2011) Protomation of base pairs in RNA: Context Analysis and Quantum Chemical Investigations of their Geometries and Stabilities, J. Phys, Chem. B **115**: 1469-1484.

Few files (AdeVariants.name, GuaVariants.name, CytVariants.namd and UraVariants.name), supplied along with this package, needs to be placed in specific location and those need to be indicated in the code. The directory name needs to be passed to the code through environmental variable “NUCLEIC_ACID_DIR” using shell command.

Different run-time options are the following:

-HD [value] to set default hydrogen bond distance cutoff (default = 3.8)
-VA [value] to set default pseudo angle cutoff (default = 120.0)
-EN [value] to set default E-value cutoff (default = 1.8)
-ML [character] to select desired chain identifier in PDB File (default = all)
-HT to include HETATM entries in PDB
-CH to avoid identification of base pairs stabilized by C-H...O/N H-bonds
-SG to avoid identification of base pairs involvingsugar O2' atoms
-AB to avoid base pairing between residue no. i and i+1
-NMR to calculate base pairing information for the first model NMR derived structure
-MD [number] to calculate base pairing information fora particular NMR model [number]
-CIF to calculate base pairing information from mmCIF formatted file (PDB is the default). This considers only the first model of NMR derived structures

As PDBe is primarily focusing on working with mmCIF formatted files, main emphasis was given so that mmCIF files, such as PDB_ID_updated.cif, can be used for analysis for obtaining outputs in desired csv format. Hence, the default run script should be:
bpfind PDB_ID_updated.cif -cif

Two output files would be created along with few more for other purposes, such as PDB_ID_updated_pairing.csv and PDB_ID_updated_structure.csv. The first one contains detail information of each base pair, the pairs of bases involved, the type of base pairing, etc. The columns

have the following information:
1st column: PDB_ID of the file
2nd column: RNA residue serial number
3rd column: The information provided in the _atom_site.auth_sequ_id column of the mmCIF file
4th column: Single character residue name
5th column: This contains the information provided by the _atom_site.pdbx_PDB_ins_code column of mmCIF file. Generally this column contains "?" but often this contains some other character also. The program internally uses the information provided here to assign proper RNA residue serial number
6th column: Chain ID as available from the _atom_site.auth_asym_id of the mmCIF file (this column can be upto 4 character long)

In case of single stranded molecules or single stranded regions without any base pairing, the line may not have anything else. When a base is paired to some other base, the similar informations of the paired base are given in 7th column onwards. Hence, 7th column can have a residue serial number of the paired base, 8th column can have residue number given in the _atom_site.auth_sequ_id field, 9th column can have single character residue name, 10th column would have chain name of the paired base, 11th column would have the information _atom_site.pdbx_PDB_ins_code of the paired base, 12th column would have chain name of the paired base. The 13th column indicates type of base pairing between the two bases. It can be one of several type, such as W:W, W:H, H:W, H:H, W:S, S:W, H:S, S:H, or S:S along with smaller case types or types containing "+" or "z". The 14th column further gives type of base pairing in terms of *Cis* or *Trans.* The 15th column, if present, contains BP always. The 16th column gives a quantitative composite measure of quality of the base pair, indicating how strong the hydrogen bonds are, how planar the two bases are, etc.

There can, sometimes, be 17th and further columns, which contain informations identical to 7th to 13th column, indicating that the base (identified by the 2nd column) is paired simultaneously to bases indicated by 7th and 17th column residues. In this case the columns 17th through 22nd would be filled with values, the 23rd column would be indicated by "TP" or "BF" and 24th column would contain measure of the pairing. There can be values in the columns 25th to 32nd when the residue serial number (mentioned in 2nd column) is paired to three bases simultaneously using its three edges.

The second file created by BPFIND contains information about secondary structure of any of its residues. It contains residue serial number, residue number in PDB (from _atom_site.auth_sequ_id of mmCIF file), residue name, PDB_ins_code, chain name and single character secondary structure information. The secondary structure are mentioned by "C" for unpaired residue which may be in coil conformation; "H" indicating double helical region; "T" indicating when a base is paired to two bases simultaneously; "N" indicating the base is involved in non-canonical base pairing; "W" indicating a residue is paired to another by canonical Watson-Crick way but not having other double helical characters, such as anti-parallel arrangement, stacking with successive base pairs, etc.; "L" indicates bases in hairpin loop conformation in between two strands in double helical structures; "B" indicates single stranded asymmetric bulged out residue within helical region.

The file with ".dbn" extension contains secondary structure information of the selected RNA in Dot-Bracket format, which can be viewed by VARNA or other packages. The BPFIND code resolves Pseudoknot problem for upto two levels of pseudoknot at present by using "(", "[" and "{" opening braces, with appropriate closing ones. The triplets, however, are not considered in the ".dbn" file.