

CANAL USER GUIDE

(V1.5, updated October 2016)

Canal uses data from Curves+ (in a *file.cda*) to analyze structural information (helical, backbone and groove parameters). It is generally used for the analysis of molecular dynamics trajectory data (one or many trajectories simultaneously), but it can also be used for single conformations. Canal calculates averages and standard deviations and can optionally look for correlations between different parameters. Appropriate algorithms for dealing with both linear and circular parameters (such as backbone torsions) are used and all algorithms are single pass, so that the program can easily be applied to analyzing the data from large trajectories. The latest version of the code includes correct algorithms for obtaining correlation coefficients for pairs of circular variables as well as for combinations of circular and linear variables (see methods section of reference 3 below).

References:

- 1) Conformational analysis of nucleic acids revisited: Curves+. R. Lavery, M. Moakher, J.H. Maddocks, D. Petkeviciute, K. Zakrzewka *Nucleic Acids Res.* 37 (2009) 5917
- 2) A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. R. Lavery et al. *Nucleic Acids Res.* 38 (2010) 299
- 3) μ ABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. M. Pasi et al. *Nucleic Acids Res.* 42 (2014) 12272

Input data

The example below shows a simple input file. Here the compiled program is in the directory /Users/RL/Code/util and has the name canal. The compiled version is obtained by typing for example "gfortran -o canal canal.f" in the directory containing the canal.f (the Fortran 77 source code). Check that you have an appropriate Fortran compiler available.

The notation <<! allows the input data to be placed immediately after the line which calls the program. The input is then ended by the explanation mark.

```
rm test_ga.* test_ga_*.  
/Users/RL/Code/util/canal <<!  
&inp lis=test_ga,seq=GA,  
lev1=3,lev2=16,histo=t., &end  
GGGA_spc GCGAGGGAGGGAGGGAGC  
GAAA_spc GCAAGAAAGAAAGAAAGC  
!
```

The main options of Canal are controlled with a namelist input. All namelist parameters have default parameters (see below) that will be used unless new values are given in the namelist section of the input (i.e. between "&inp" and "&end").

The example above analyzes two MD trajectories for 18-mers with the sequences shown (contained in the files GGGA_spc.cda and GAAA_spc.cda). The analysis selects all GpA steps (see **seq**), avoiding the first and last two base pairs (see **lev1** and **lev2**). It will calculate the average, standard deviation,

range, minimum, maximum, and number of data points of all variables belonging to the appropriate steps in the two oligomers, avoiding the first and last two levels (i.e. base pairs). It will also generate histograms (e.g. test_twist.his) summarizing the fluctuations of each variable, accumulated for the selected steps.

Following the namelist (&inp...&end) input, there is an arbitrary number of lines, each specifying the names of the .cda file containing the data coming from Cur+, followed by a single blank space and then the sequence of the first strand of the corresponding DNA fragment. Note that you MUST give the correct sequence of the oligomer in each case. This data is not contained in the .cda files and consequently cannot be checked by Canal. Character strings should not be enclosed in quotes. All characters should be upper case.

For single structures (rather than MD trajectory input) use seq=* to summarize the variations of all parameters along the oligomer and series=.t. to generate files for plotting these parameters.

NAMELIST VARIABLES (the default values of each variable is given in brackets)

CHARACTER (strings without quotes, maximum length 128 characters):

lis (): root file name for output. The extensions to this file name are .lis for the printed output, *_variable.ser* for time series data and *_variable.his* for histograms.

seq (): sequence of bases to be searched for in the I/P data (default is blank, meaning no specified sequence). If this sequence is found and is odd in length, the central base will be analyzed. If seq is even in length, the central base pair step will be analyzed. "seq" can contain the upper case letters A, C, G, T (for the usual bases), R (purine = A or G), Y (pyrimidine = C or T) or * (any base A, C, G or T)

prop (): an optional file containing one value per snapshot in the .cda file. Used with **pmin** and **pmax** to filter input data. Any snapshot for which the corresponding value is less than pmin or greater than pmax is ignored.

nastr ('NA'): character string used to indicate missing data in .ser files see notes on .ser O/P)

REAL DATA

clim (0.4): value of curvature below which the direction of curvature is ignored

cormin (0.6): minimal absolute value for printing linear correlation coefficients between pairs of analyzed variables

pmin (-500.0): lower limit of value read in **prop** file for accepting data from the corresponding snapshot

pmax (500.0): upper limit of value read in **prop** file for accepting data from the corresponding snapshot

INTEGER DATA (default value of each variable is given):

lev1 (0): lower base level limit (i.e. base pairs) used for analysis

lev2 (0): upper base level limit used for analysis. If lev1 > 0 and lev2 = 0, lev2 is set to lev1 (i.e. analyze lev1 only). If lev1=lev2=0, lev1 is set to 1 and lev2 is set to the length of the oligomer (i.e. analyze all levels). Note that if many levels are analyzed, series or histogram files will contain many columns, one for each level. The 1st column gives the snapshot number.

itst (0): first snapshot of trajectory to analyze

itnd(0): last snapshot of trajectory to analyze. If itst>0 and itnd=0, itnd is set to itst (i.e. analyze snapshot number itst). If itst=0 and itnd=0, itst is set to 0 and itnd is not limited.

itdel (1): If itdel>1 only analyze every (itdel)th snapshot

iwin (1000): width of sliding window used to smooth magnitude and direction of curvature. Note that if you want to see curvature and register values for a single structure, set **iwin** to 1.

LOGICAL SWITCHES (enter as .t. or .f., the default value of each variable is given):

corr (.f.): if .t. then calculate linear correlation coefficients between all variables.

series (.f.): if .t. then O/P a spatial or time series. Only possible for the analysis of single structures or single trajectories

histo (.f.): if .t. then O/P histograms

LIST OUTPUT (.lis)

For each level (i.e. base pair) selected (i.e. within the limits imposed by **lev1** and **lev2**) and, optionally, corresponding to a chosen sequence (set with **seq**), Canal analyzes all the parameters coming from Cur+ and outputs the average, standard deviation, range (for angular variables, the number of 1° bins visited in the interval 0°-360°), minimum and maximum values (except for backbone dihedrals - see Curves+ article, reference 1 above) and the number of occurrences of the corresponding variable in the I/P data. Output referring to specific strands is indicated by "W" (Watson) for the first strand and by "C" (Crick) for the second strand.

When the analysis involves a sequence (**seq**), the output refers to the two central base pairs of **seq**, when it contains an even number of characters, and to the single central base pair, if it contains an odd number of characters. Note that the base pairs analyzed must lie within the range **lev1:lev2**, but the full string **seq** may go beyond these levels (as long as it fits within the corresponding oligomer).

Note that if a **seq** corresponds to a sequence fragment in the second strand of an oligomer, the backbone labels "W" and "C" (Watson and Crick) are automatically inverted so that "W" always refers to the strand containing seq. If **seq** has inversion symmetry (e.g. **seq**=AATT), it will be found in both strands at the same position, and it is treated as belonging to the first strand of the oligomer.

Extra output summarizes sugar puckers, epsilon/zeta and alpha/gamma backbones dihedrals (see below).

The detailed output of the variables analyzed is followed by a summary of the various backbone conformations.

- 1) Distribution of sugar puckers (phase angle and amplitude).
- 2) Distribution of BI-BII sub-states i.e. ϵ (epsilon) and ζ (zeta), where BI implies $\epsilon - \zeta < 0^\circ$ and BII implies $\epsilon - \zeta > 0^\circ$.
- 3) Distributions of α (alpha) / γ (gamma) grouped in terms of g+/t/g- (0°-120°, 120°-240°, 240°-360°) classes.

SERIES OUTPUT (.ser)

When analyzing MD trajectories, these files contain time series of the variables analyzed. The first column is simply a consecutive number from 1 to N (number of snapshots analyzed). The following columns give the value of the variables in each snapshot. The number of data elements per line depends on the number of levels separately analyzed (i.e. appearing in the .lis output). On the other hand, when single structures are analyzed, each line corresponds to a base pair level, and only one further data element per line is present.

A simple plot in gnuplot can be made with a command such as:

```
plot 'file_var.ser' using 1:3 with lines
```

- where *file_var.ser* is the appropriate file name. Note that:

- 1) backbone variables in each strand yield two separate files: *file_varW.ser* and *file_varC.ser*
- 2) missing data in .ser files (for example, no groove widths or depths towards the ends of an oligomer) is indicated by the standard symbol 'NA'. This symbol can be changed using the namelist variable **nastr**. If you use gnuplot to plot data from .ser files, use the notation '(\$X)' to plot the Xth column. This will correctly introduce line breaks where no data exists. For R users: The "read.table" command will accept "NA" by default as missing data. For Matlab users: The "textscan" command allows the definition of strings to be treated as missing data using the 'TreatAsEmpty' option.

HISTOGRAM OUTPUT (.his)

These files contain normalized histograms of the variables analyzed. The first column gives the value of the variable in the corresponding bin and the following columns specify the relative occurrence of this value. The number of data per line depends on the number of levels analyzed (i.e. appearing in the .lis output).

n.b. backbone variables in each strand yield two separate files: *file_varW.his* and *file_varC.his*

CORRELATION OUTPUT (.cor)

If corr=.t. linear, circular-circular and linear-circular correlations are calculated between all analyzed variables. The output gives the number of each variable (as it is listed above in the .lis file), the base pair level they belong to, their names and the correlation coefficient.

If series=.t. or if .ser output files have already been generated for the same analysis, then files .cor are generated containing the correlated variables as x,y pairs. The names of these files are constructed using the numbers defining each variable in the .lis output. For example, a stretch/opening correlation involving the first base pair level analyzed would generate file_2_6.cor, where the stretch is variable number 2 in the .lis output and opening is variable number 6.

NUMBERING VARIABLES

The standard variable numbering for the first base pair level analyzed is:

1 -6: shear, stretch, stagger, buckle, propel , opening
7-11: xdisp , ydisp, inclin, tip, ax-bend,
12-19: shift, slide, rise, tilt, roll, twist, h-ris, h-twi,
20-21: curvature, register
22-30: alpha, beta, gamma, delta, epsil, zeta, chi, phase, amp,
31-39: alpha, beta, gamma, delta, epsil, zeta, chi, phase, amp,
40-43: minw, mind, majw 'majd

Note that 22-30 and 31-39 correspond to the Watson and Crick backbones respectively and minw, mind, majw, majd correspond to the minor groove width and depth and the major groove width and depth respectively.

Sample output

The following is the analysis of the Dickerson dodecamer (from 1bna.pdb, see Curves+ guide for creating the .cda file used here). Parameters are averaged over the oligomer using the option seq=* (sequence averages are indicated by the letter "M". Note that the range of M is indicated in the output file).

```

*****
**** CANAL Version 1.5 10/2016 ****
*****

LIS : test                      SEQ : *
nastr: NA

cormin: 0.60

lev1 : 0 lev2 : 0 itst : 0 itnd : 0 itdel : 1
IWIN : 1

corr : F series: F histo : F

1 12 12 test CGCGAATTCGCG

Data from 1 trajectory and 12 levels ...

M) 1 2 3 4 5 6 7 8 9 10 11 12

No. Lev Var Aver Sdev Range Min Max Ndat
 0) - t bend 0.7 0.0 1.0 0.7 0.7 1
=====
1) M shear -0.04 0.27 0.9 -0.6 0.3 12
2) M stretch -0.17 0.09 0.3 -0.3 -0.0 12
3) M stagger 0.21 0.19 0.8 -0.1 0.7 12
4) M buckle 0.3 5.9 11.0 -10.8 10.1 12
5) M propel -13.7 6.9 9.0 -20.2 0.5 12
6) M opening 1.0 4.3 8.0 -4.7 10.2 12
-----
7) M xdisp 0.27 0.55 1.8 -0.4 1.5 12
8) M ydisp 0.11 0.23 1.0 -0.4 0.6 12
9) M inclin -0.1 4.7 9.0 -8.3 6.4 12
10) M tip -1.0 3.1 8.0 -5.9 5.6 12
11) M ax-bend 0.7 0.3 1.0 0.3 1.0 11
-----
12) M shift -0.02 0.55 2.1 -1.4 0.8 11
13) M slide 0.14 0.43 1.5 -0.6 0.9 11
14) M rise 3.36 0.18 0.7 3.0 3.7 11
15) M tilt -0.2 2.9 7.0 -5.0 3.2 11
16) M roll -0.3 5.9 10.0 -13.5 9.1 11
17) M twist 35.8 4.1 8.0 26.7 42.5 11
18) M h-ris 3.35 0.18 0.6 3.0 3.7 11
19) M h-twi 36.0 4.1 7.0 27.2 42.9 11
-----
20) M curv 0.37 0.22 0.7 0.1 0.9 9
21) M reg -163.4 62.9 9.0 9.0 9.0 9
-----
22) M alphaW -64.0 9.8 9.0 9.0 9.0 11
23) M betaW 169.6 14.0 9.0 9.0 9.0 11
24) M gammaW 60.5 33.8 10.0 10.0 10.0 12
25) M deltaW 125.4 19.0 12.0 12.0 12.0 12
26) M epsilonW -166.8 25.0 9.0 9.0 9.0 11
27) M zetaW -107.2 38.4 10.0 10.0 10.0 11
28) M chiW -116.1 13.5 11.0 11.0 11.0 12
29) M phaseW 132.0 21.9 12.0 12.0 12.0 12
30) M ampW 48.6 4.8 8.0 8.0 8.0 12
-----
31) M alphaC -61.6 5.9 9.0 9.0 9.0 11
32) M betaC 173.5 13.5 9.0 9.0 9.0 11
33) M gammaC 53.5 7.9 10.0 10.0 10.0 12
34) M deltaC 120.3 21.2 11.0 11.0 11.0 12
35) M epsilonC -173.9 24.1 9.0 9.0 9.0 11
36) M zetaC -102.4 28.4 9.0 9.0 9.0 11
37) M chiC -118.6 12.9 12.0 12.0 12.0 12
38) M phaseC 127.0 37.5 11.0 11.0 11.0 12
39) M ampC 45.5 3.1 8.0 8.0 8.0 12
-----
40) M minw 4.79 1.28 4.1 3.1 7.2 6
41) M mind 5.27 0.30 0.8 4.9 5.7 6
42) M majw 11.44 0.71 1.9 10.5 12.4 6

```

43) M majd 4.48 0.95 2.6 2.9 5.4 6

Sugar pucker distribution ...

		C3'n	C4'x	O1'n	C1'x	C2'n	C3'x	C4'n	O1'x	C1'n	C2'x
29)	M W	0	0	17	50	33	0	0	0	0	0
38)	M C	8	8	0	42	42	0	0	0	0	0
Tot		4	4	8	46	38	0	0	0	0	0

Epsilon/Zeta distribution ...

		BI	BII
26)	M W	82	18
35)	M C	91	9
Tot		86	14

Alpha/Gamma distribution ...

		g-/g-	g-/ t	g-/g+	t /g-	t /t	t /g+	g+/g-	g+/t	g+/g+
22)	M W	0	0	100	0	0	0	0	0	0
31)	M C	0	0	100	0	0	0	0	0	0
Tot		0	0	100	0	0	0	0	0	0