Annotation of Nucleic Acid Structures at *dnatco.datmos.org*

Bohdan Schneider

bohdan.schneider@gmail.com
bs.structbio.org

Jiří Černý,

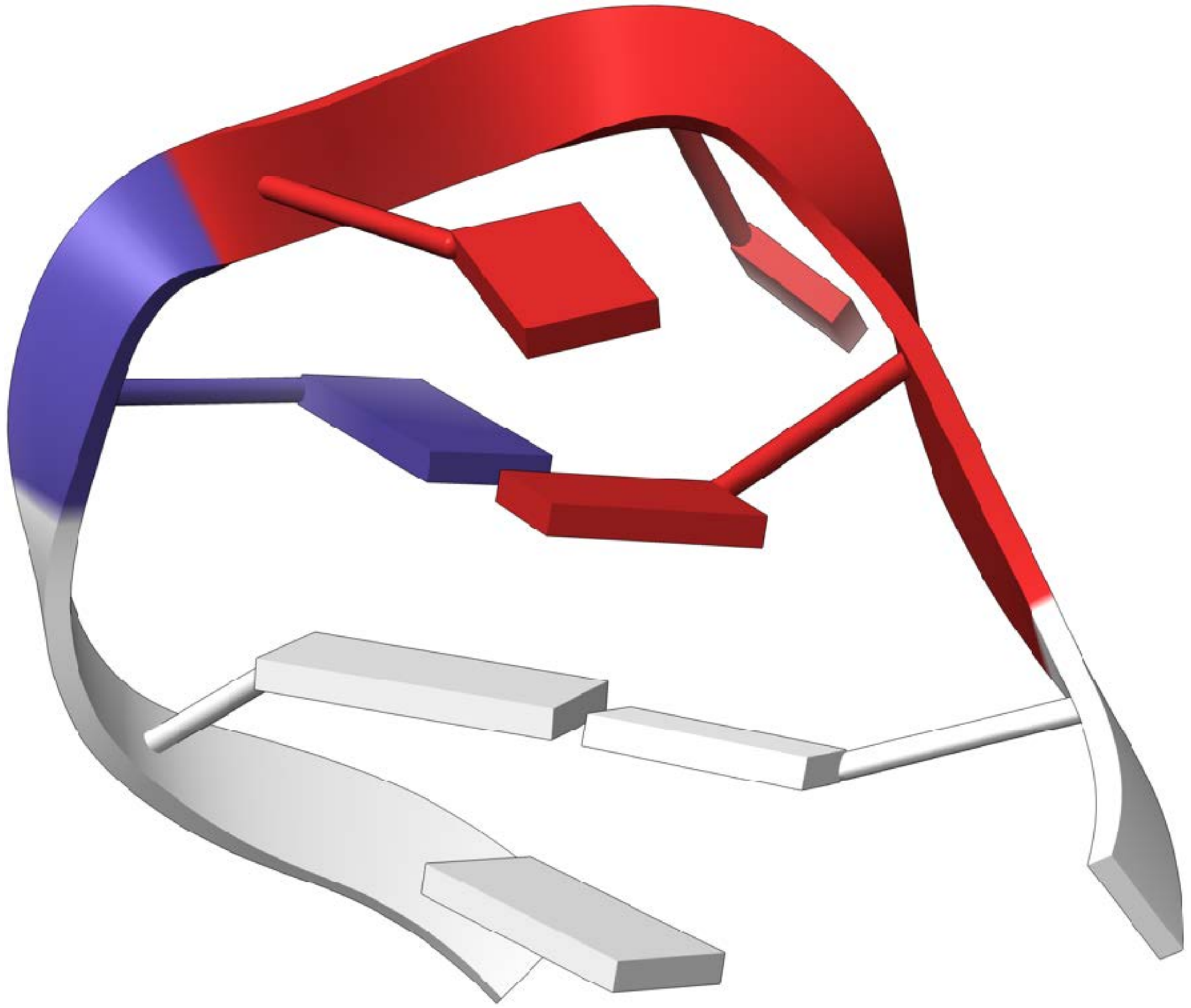
Paulína Božíková, Lada Biedermannová,
Michal Malý, Jakub Svoboda

Institute of Biotechnology of the Czech Academy of Sciences

Computational Approaches to RNA Structure nd Function
2022-08-10

Motivation

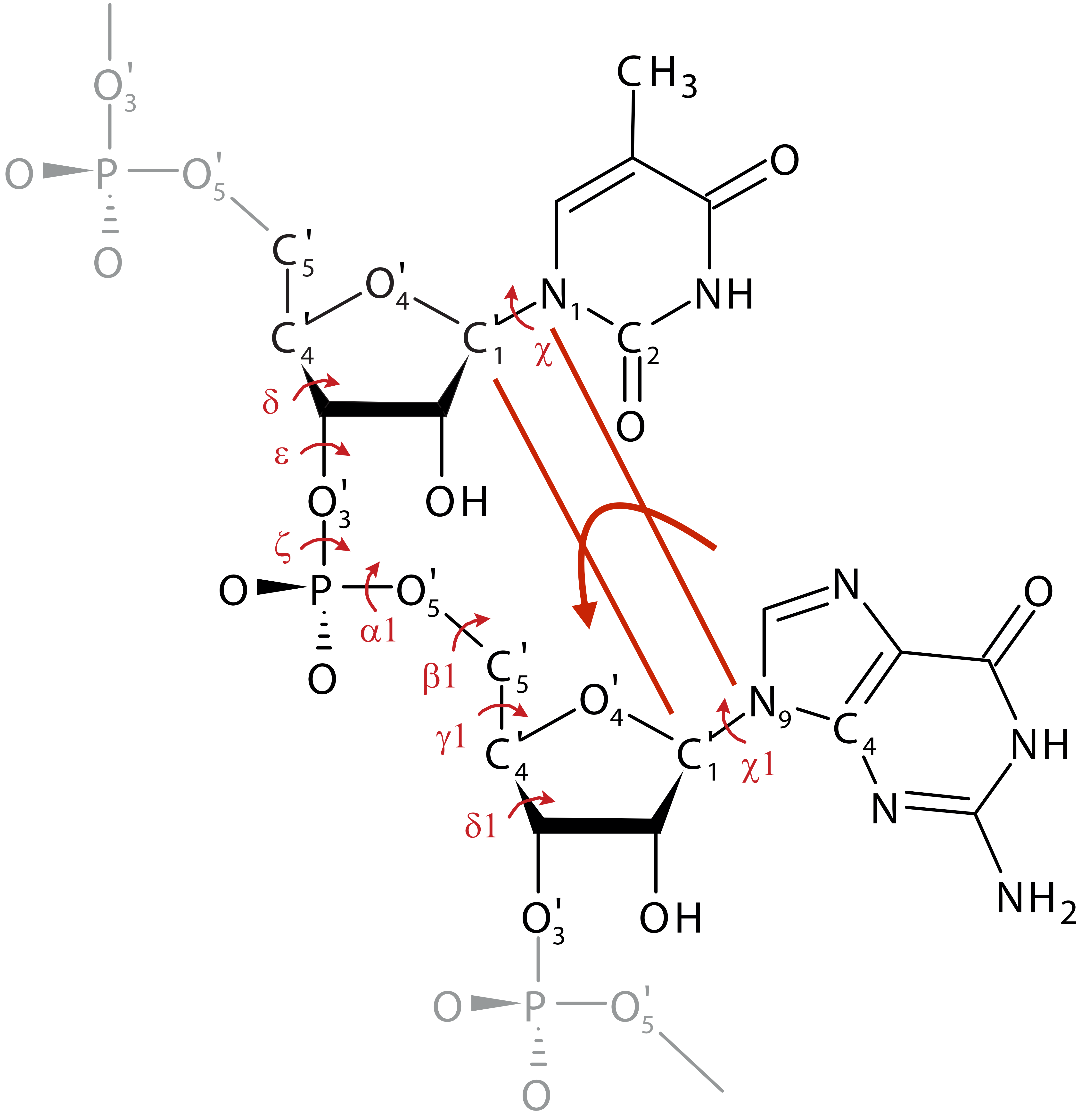
- Provide tools to make annotation, refinement, validation, and modeling of nucleic acids more reliable



Nucleotide Conformers, ***NtC***

- Both DNA & RNA analyzed:
  - sequentially non-redundant set of crystal structures
  - containing ~120 thousand dinucleotide steps
- Analysis in 12D torsion space
  - fragment is not suite
- Cluster analysis + empirical rules
- 96 ***NtC*** classes

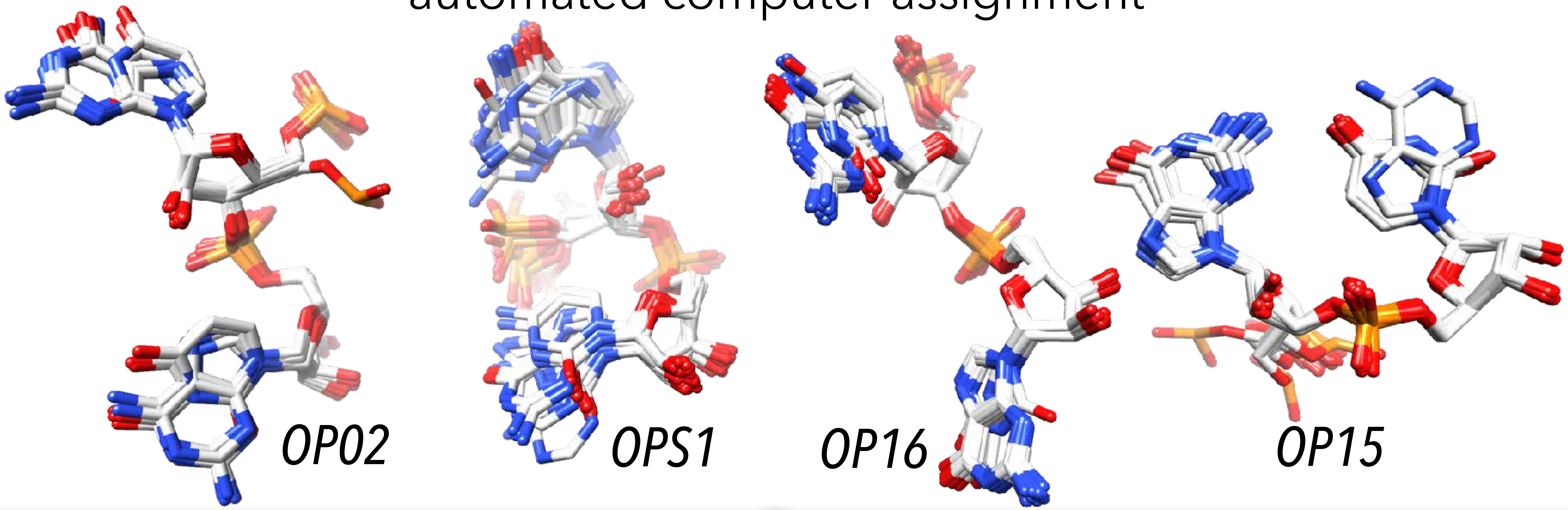
Černý et al.: Nucleic Acids Research **48**: 6367 (2020).



Two levels of analysis:
NtC = Geometry assignment; CANA = symbolic annotation

96 ***NtC*** *classes: geometry of ensembles*

automated computer assignment



NtCs *OPxx grouped to* ***CANA*** *"letter" OPN*

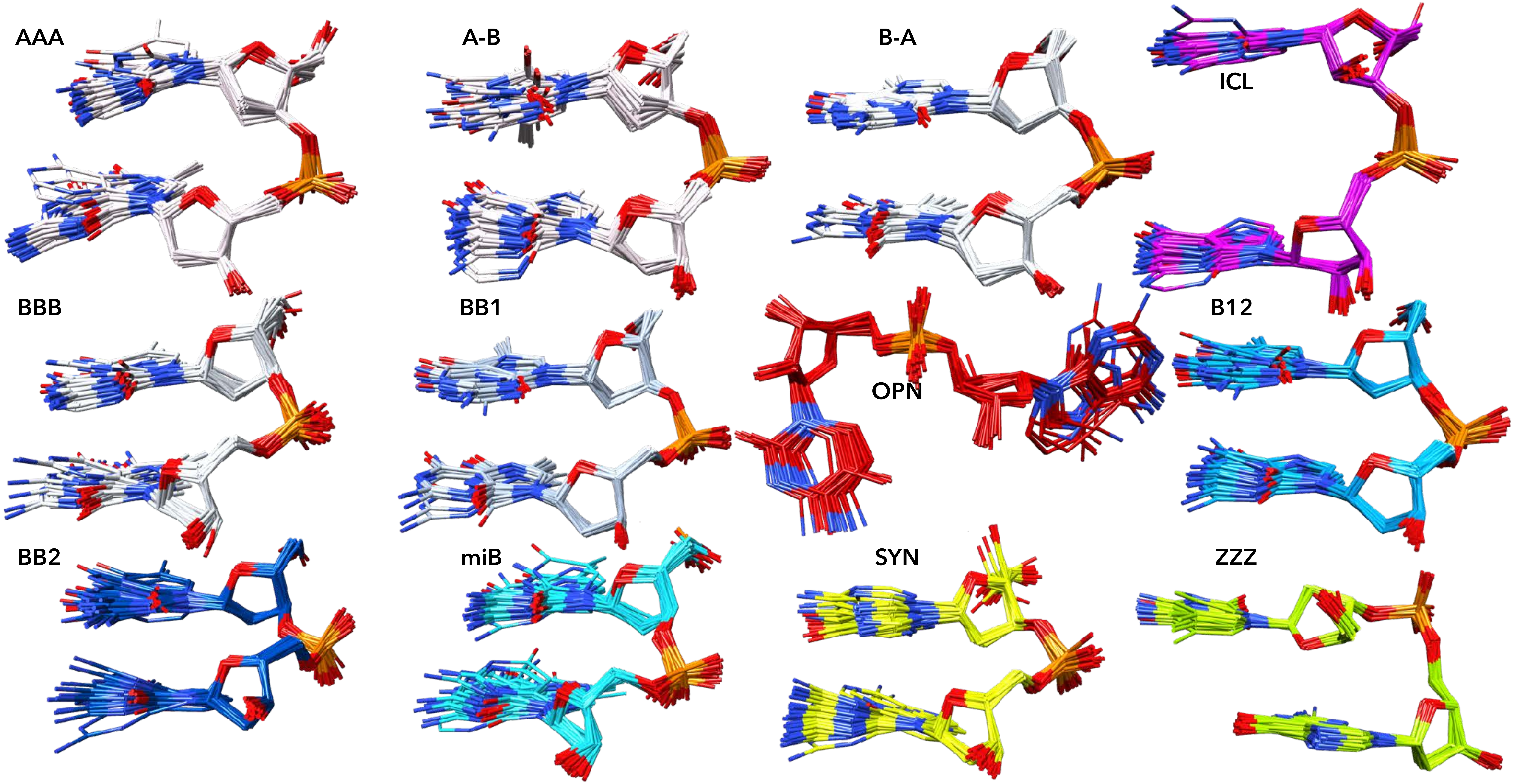
NtC & *CANA* summary, more at *dnatco.datmos.org*

Annotation	***CANA***	***NtC***	Numbers of Steps in			
			DNA		RNA	
			#	%	#	%
the most frequent, canonical A form	*AAA*	*AA00*	1 805	3.1	21 177	37.1
A-DNA with B like χ	*AAA*	*AA02*	595	1.0	80	0.1
A like, similar to AA00, high α	*AAA*	*AA03*	32	0.1	806	1.4
A like, similar to AA00, low α	*AAA*	*AA04*	148	0.3	1 830	3.2
A form variant	*AAA*	*AA08*	150	0.3	8 527	15.0
A form variant	*AAA*	*AA09*	22	0.0	230	0.4
A like, α/γ switch (150/180)	*AAw*	*AA01*	284	0.5	1 859	3.3
A like, α/γ switch (130/180)	*AAw*	*AA05*	7	0.0	385	0.7
A like, α/γ switch (130/180)	*AAw*	*AA06*	18	0.0	513	0.9
A like, α/γ switch (210/140)	*AAw*	*AA10*	15	0.0	488	0.9
A like, α/γ switch (90/190)	*AAw*	*AA11*	6	0.0	240	0.4
A form with distant bases, low ζ	*AAu*	*AA07*	1	0.0	95	0.2
A like backbone, bases can be intercalated	*AAu*	*AA12*	7	0.0	616	1.1
A form with distant bases, low ζ	*AAu*	*AA13*	0	0.0	105	0.2

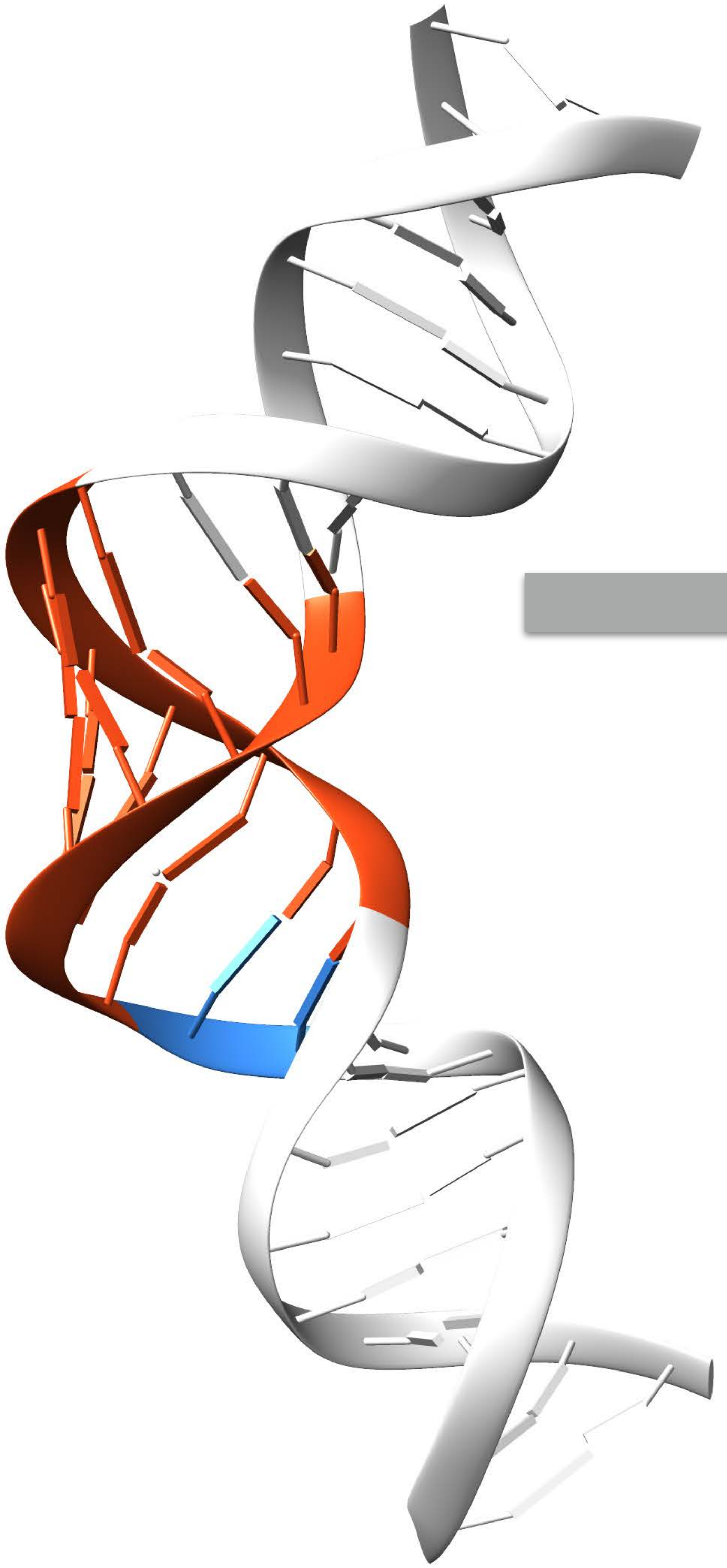
NtC & *CANA* summary, more at *dnatco.datmos.org*

The main features of CANA letters	*CANA* letter	Number of *NtC*	DNA %	RNA %
A-form	AAA	6	4.7	57.0
A-like, switched α/γ values	AAw	5	0.5	5.8
A-like, unstacked bases	AAu	3	0.0	1.3
conformers bridging A- to B-form	A-B	5	4.9	2.1
conformers bridging B- to A-form	B-A	8	6.7	0.2
canonical B-form	BBB	2	39.5	0.0
less populated BI conformers, switched α/γ	BBw	5	6.0	0.0
conformers bridging BI- to BII-form	B12	2	6.3	0.0
BII form	BB2	2	4.8	0.0
minor B-like, untypical torsion combinations	miB	6	3.6	0.0
parallel bases, can be intercalated	ICL	7	0.2	0.6
unstacked, often distant bases	OPN	33	0.3	4.1
parallel bases, one in syn orientation	SYN	6	0.7	0.3
Z-forms	ZZZ	6	0.6	0.6
All Assigned Steps	AlA	96	79.0	72.2
non-Assigned Steps	NAN	1	21.1	27.8
All Steps			100.0	100.0

NA structural alphabet ***CANA***



Analysis of NA structures by their “translation” into letters of the Structural Alphabet

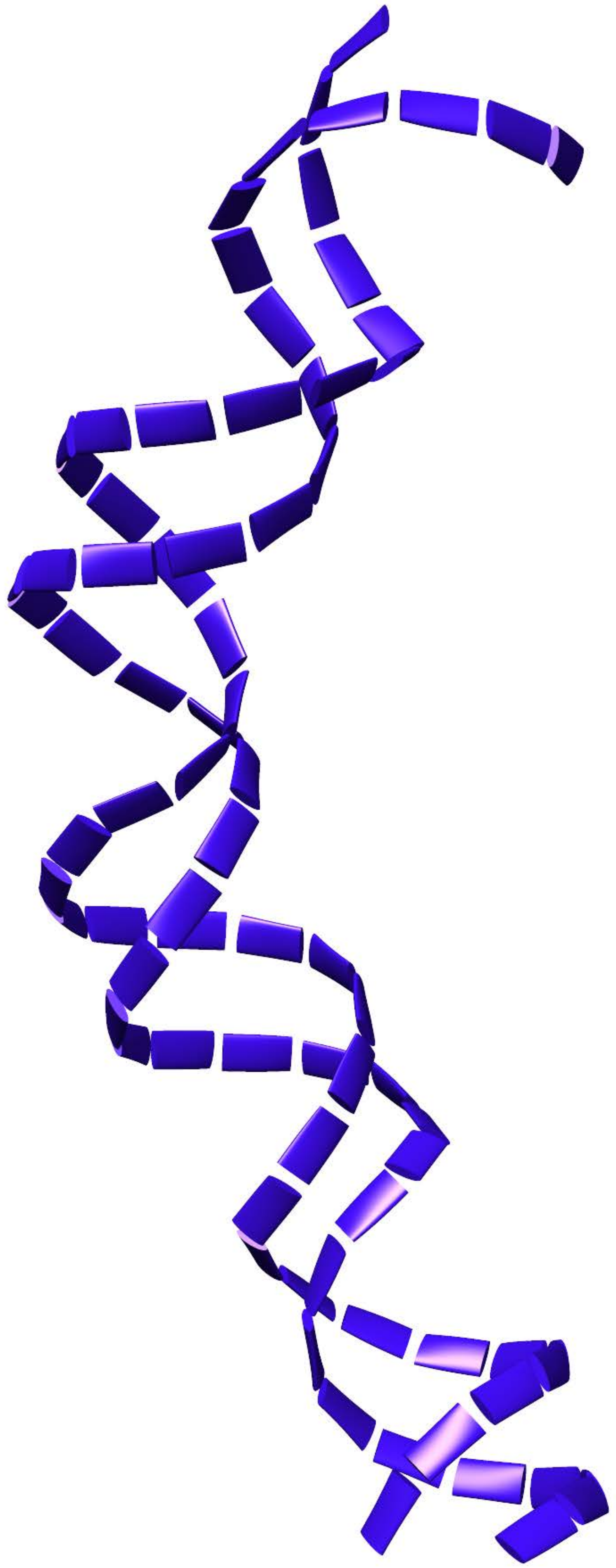


chain A:

```
miB B-A NAN miB B12 BBB 2B1 NAN
AAA AAA AAA AAA AAA AAA AAA A-B
BB2 2B1 BBB B12 B-A 2B1 B12 B-A
A-B B-A
```

chain B:

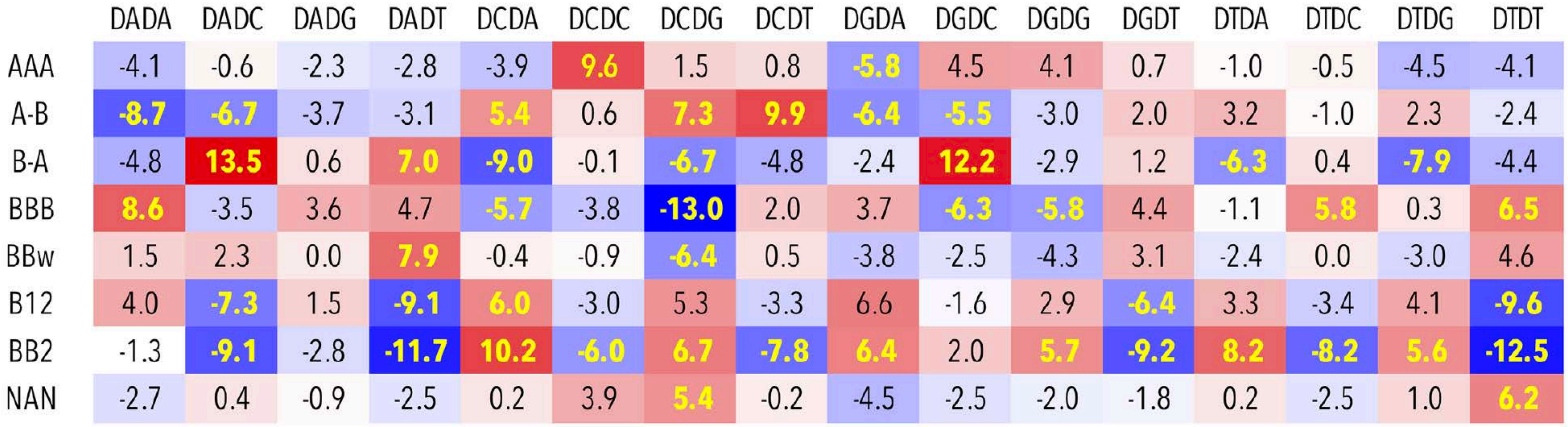
```
3B1 2B1 2B1 2B1 BBB B-A A-B BBB
B-A AAA AAA AAA AAA AAA AAA AAA
A-B 2B1 2B1 BBB NAN A-B B-A 2B1
B12 miB
```



Results

CANA/sequence associations in the nucleosome core particle and in complexes with transcription factors

Populated CANA codes in DNA structures



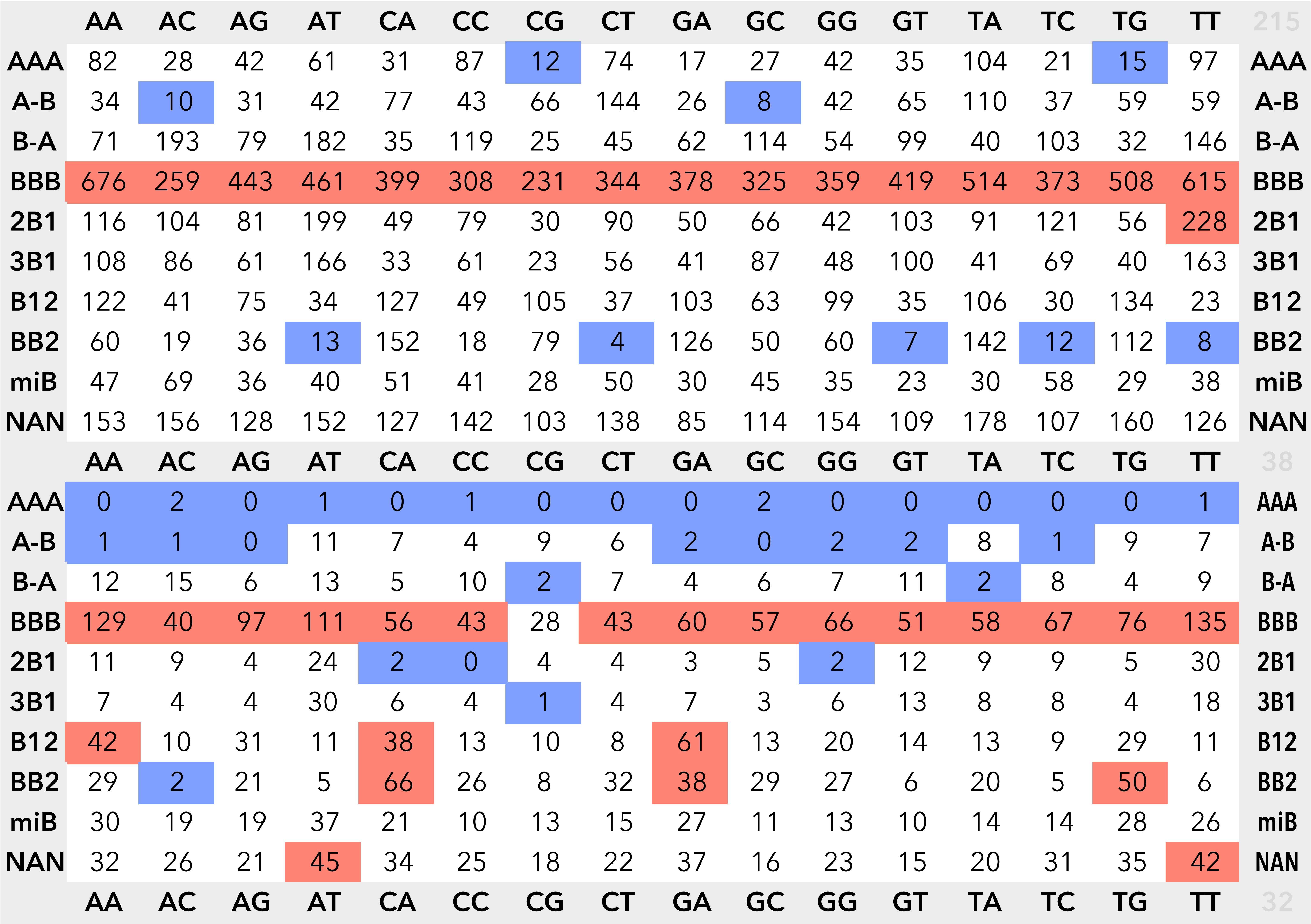
	DADA	DADC	DADG	DADT	DCDA	DCDC	DCDG	DCDT	DGDA	DGDC	DGDG	DGDT	DTDA	DTDC	DTDG	DTDT
AAA	-4.1	-0.6	-2.3	-2.8	-3.9	**9.6**	1.5	0.8	**-5.8**	4.5	4.1	0.7	-1.0	-0.5	-4.5	-4.1
A-B	**-8.7**	**-6.7**	-3.7	-3.1	**5.4**	0.6	**7.3**	**9.9**	**-6.4**	**-5.5**	-3.0	2.0	3.2	-1.0	2.3	-2.4
B-A	-4.8	**13.5**	0.6	**7.0**	**-9.0**	-0.1	**-6.7**	-4.8	-2.4	**12.2**	-2.9	1.2	**-6.3**	0.4	**-7.9**	-4.4
BBB	**8.6**	-3.5	3.6	4.7	**-5.7**	-3.8	**-13.0**	2.0	3.7	**-6.3**	**-5.8**	4.4	-1.1	**5.8**	0.3	**6.5**
BBw	1.5	2.3	0.0	**7.9**	-0.4	-0.9	**-6.4**	0.5	-3.8	-2.5	-4.3	3.1	-2.4	0.0	-3.0	4.6
B12	4.0	**-7.3**	1.5	**-9.1**	**6.0**	-3.0	5.3	-3.3	6.6	-1.6	2.9	**-6.4**	3.3	-3.4	4.1	**-9.6**
BB2	-1.3	**-9.1**	-2.8	**-11.7**	**10.2**	**-6.0**	**6.7**	**-7.8**	**6.4**	2.0	**5.7**	**-9.2**	**8.2**	**-8.2**	**5.6**	**-12.5**
NAN	-2.7	0.4	-0.9	-2.5	0.2	3.9	**5.4**	-0.2	-4.5	-2.5	-2.0	-1.8	0.2	-2.5	1.0	**6.2**

Populated CANA codes in RNA structures

	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
AAA	-4.3	0.5	-4.1	-4.2	-2.8	**9.5**	2.6	4.3	**-11.0**	**7.9**	**5.9**	1.5	**-8.5**	0.2	-4.5	-3.9
AAw	-2.6	-2.2	4.7	-3.1	0.7	-1.7	**8.7**	-1.6	**-5.4**	-2.2	**8.1**	**-6.9**	**-7.8**	-1.0	3.4	**-5.3**
A-B	-2.7	-3.3	-3.2	1.7	0.9	-2.3	1.2	**5.2**	-3.8	-3.9	0.5	1.3	0.4	-2.0	**5.1**	1.5
OPN	-1.3	-1.8	3.2	-0.7	-3.4	**-8.7**	**-6.7**	**-5.5**	**11.2**	-2.1	**-10.7**	**5.2**	**6.7**	3.7	0.5	0.4
NAN	**10.7**	3.0	**6.4**	**9.9**	5.0	**-18.5**	**-8.8**	**-9.3**	**12.9**	**-14.9**	**-13.6**	-3.5	**11.3**	-1.4	2.9	**7.9**

Černý et al.: **NAR 48**: 6367 (2020).

CANA/sequence associations in the nucleosome core particle and in complexes with transcription factors



DNA in contact with proteins

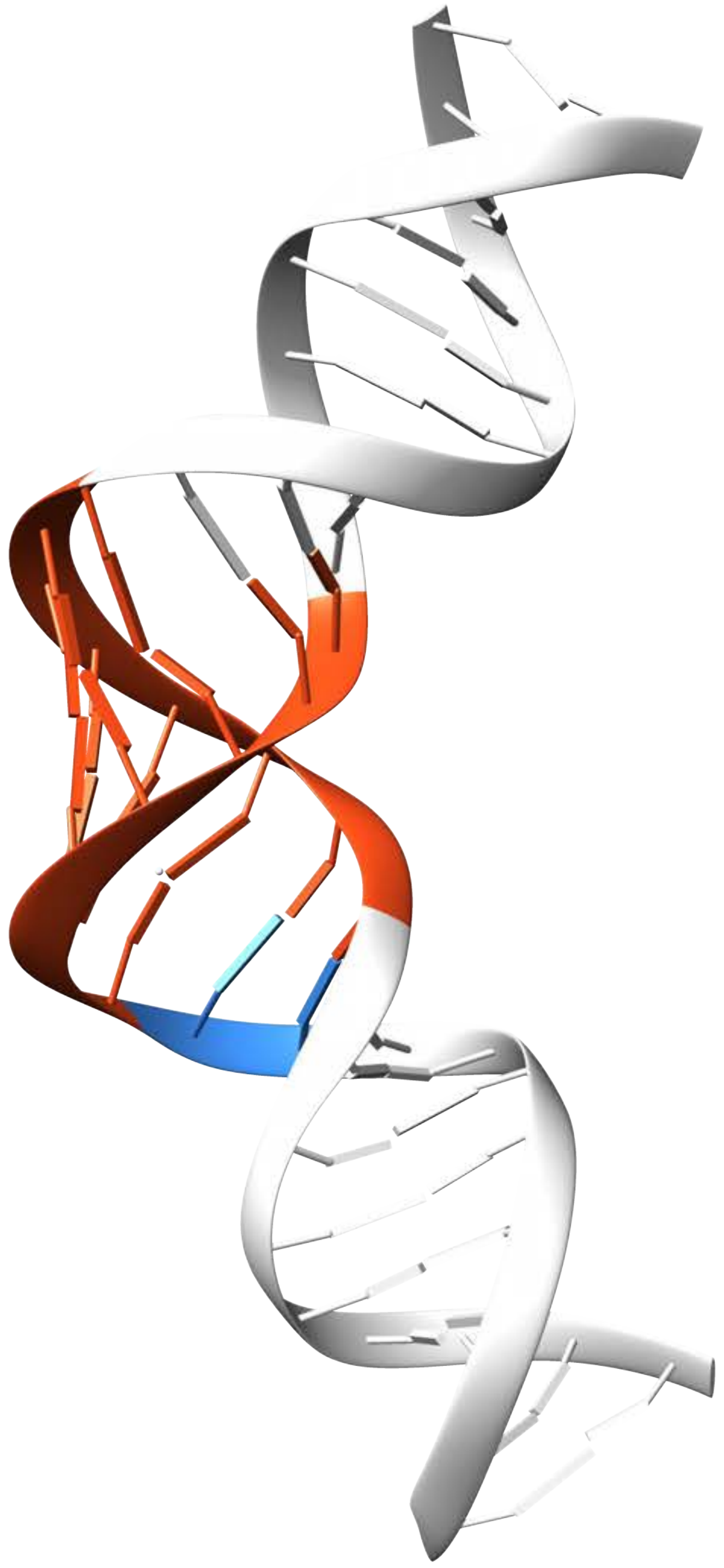
DNA Complexed to regulatory proteins

DNA in nucleosome core particles

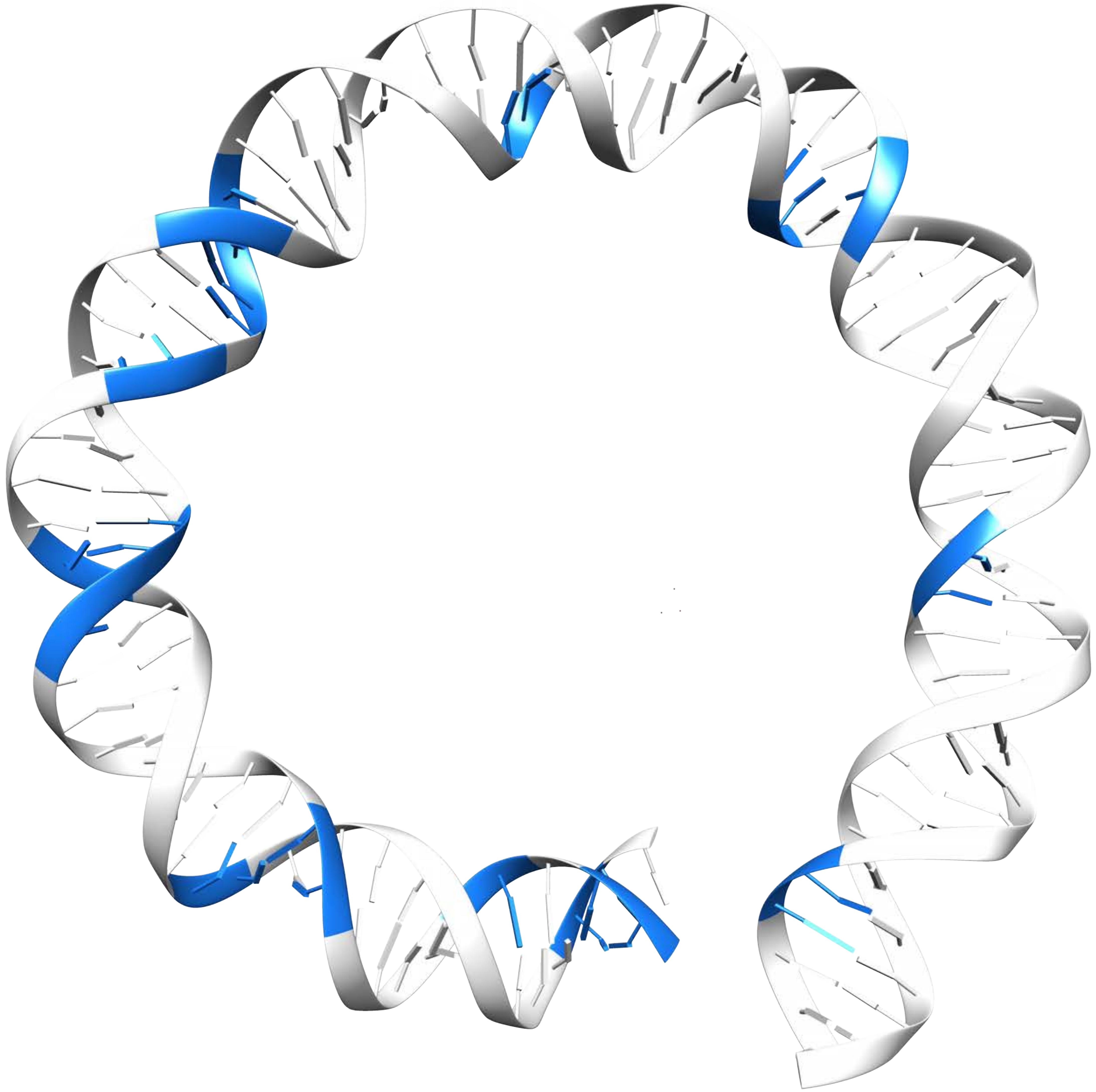
	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	215
AAA	82	28	42	61	31	87	12	74	17	27	42	35	104	21	15	97	**AAA**
A-B	34	10	31	42	77	43	66	144	26	8	42	65	110	37	59	59	**A-B**
B-A	71	193	79	182	35	119	25	45	62	114	54	99	40	103	32	146	**B-A**
BBB	676	259	443	461	399	308	231	344	378	325	359	419	514	373	508	615	**BBB**
2B1	116	104	81	199	49	79	30	90	50	66	42	103	91	121	56	228	**2B1**
3B1	108	86	61	166	33	61	23	56	41	87	48	100	41	69	40	163	**3B1**
B12	122	41	75	34	127	49	105	37	103	63	99	35	106	30	134	23	**B12**
BB2	60	19	36	13	152	18	79	4	126	50	60	7	142	12	112	8	**BB2**
miB	47	69	36	40	51	41	28	50	30	45	35	23	30	58	29	38	**miB**
NAN	153	156	128	152	127	142	103	138	85	114	154	109	178	107	160	126	**NAN**
	AA	**AC**	**AG**	**AT**	**CA**	**CC**	**CG**	**CT**	**GA**	**GC**	**GG**	**GT**	**TA**	**TC**	**TG**	**TT**	38
AAA	0	2	0	1	0	1	0	0	0	2	0	0	0	0	0	1	**AAA**
A-B	1	1	0	11	7	4	9	6	2	0	2	2	8	1	9	7	**A-B**
B-A	12	15	6	13	5	10	2	7	4	6	7	11	2	8	4	9	**B-A**
BBB	129	40	97	111	56	43	28	43	60	57	66	51	58	67	76	135	**BBB**
2B1	11	9	4	24	2	0	4	4	3	5	2	12	9	9	5	30	**2B1**
3B1	7	4	4	30	6	4	1	4	7	3	6	13	8	8	4	18	**3B1**
B12	42	10	31	11	38	13	10	8	61	13	20	14	13	9	29	11	**B12**
BB2	29	2	21	5	66	26	8	32	38	29	27	6	20	5	50	6	**BB2**
miB	30	19	19	37	21	10	13	15	27	11	13	10	14	14	28	26	**miB**
NAN	32	26	21	45	34	25	18	22	37	16	23	15	20	31	35	42	**NAN**
	AA	**AC**	**AG**	**AT**	**CA**	**CC**	**CG**	**CT**	**GA**	**GC**	**GG**	**GT**	**TA**	**TC**	**TG**	**TT**	32

Schneider et al.: *Genes* (2017)

Transcription factors: specific recognition, DNA may be straight or bent
Nucleosome core particle: “nonspecific” binding, bent DNA

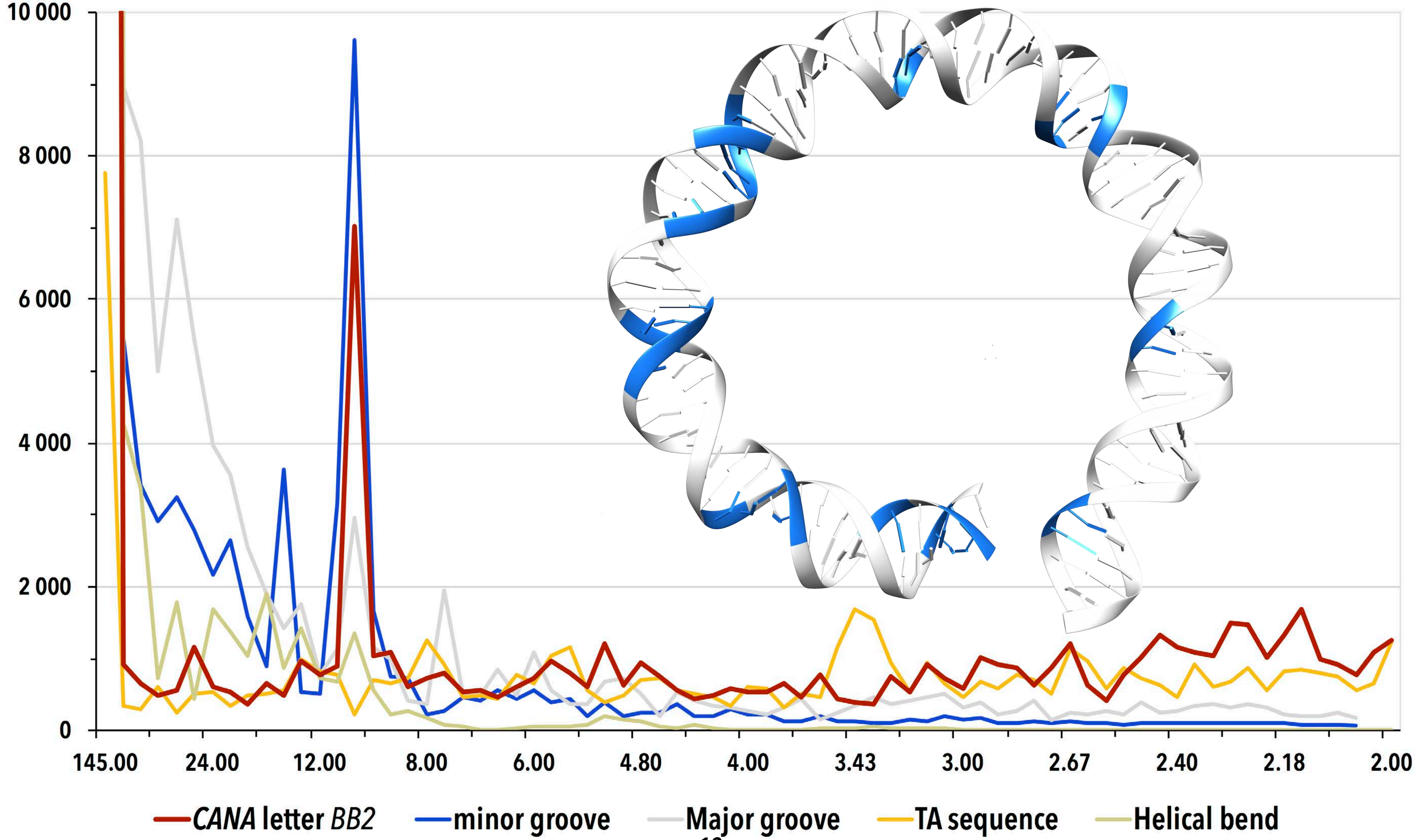


DNA bound by transcription factor in 4roc



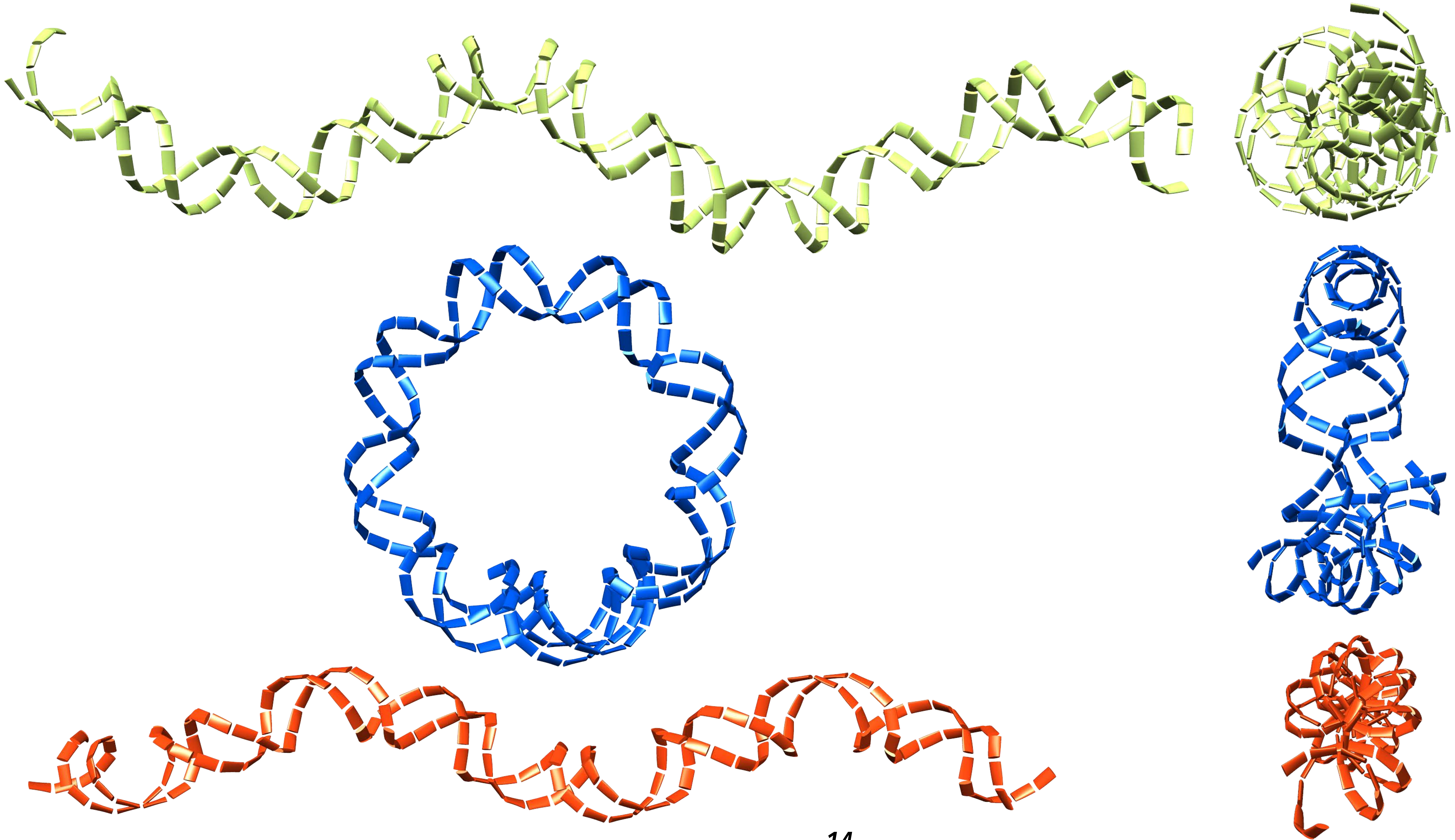
DNA in nucleosome core particle 5f99

Fourier transform of several DNA properties along the strand: BII every 10th nucleotide bends DNA duplex



Schneider et al.: *Genes* (2017)

BI-BII-BI models with periodicity of BII 8, 10, and 12



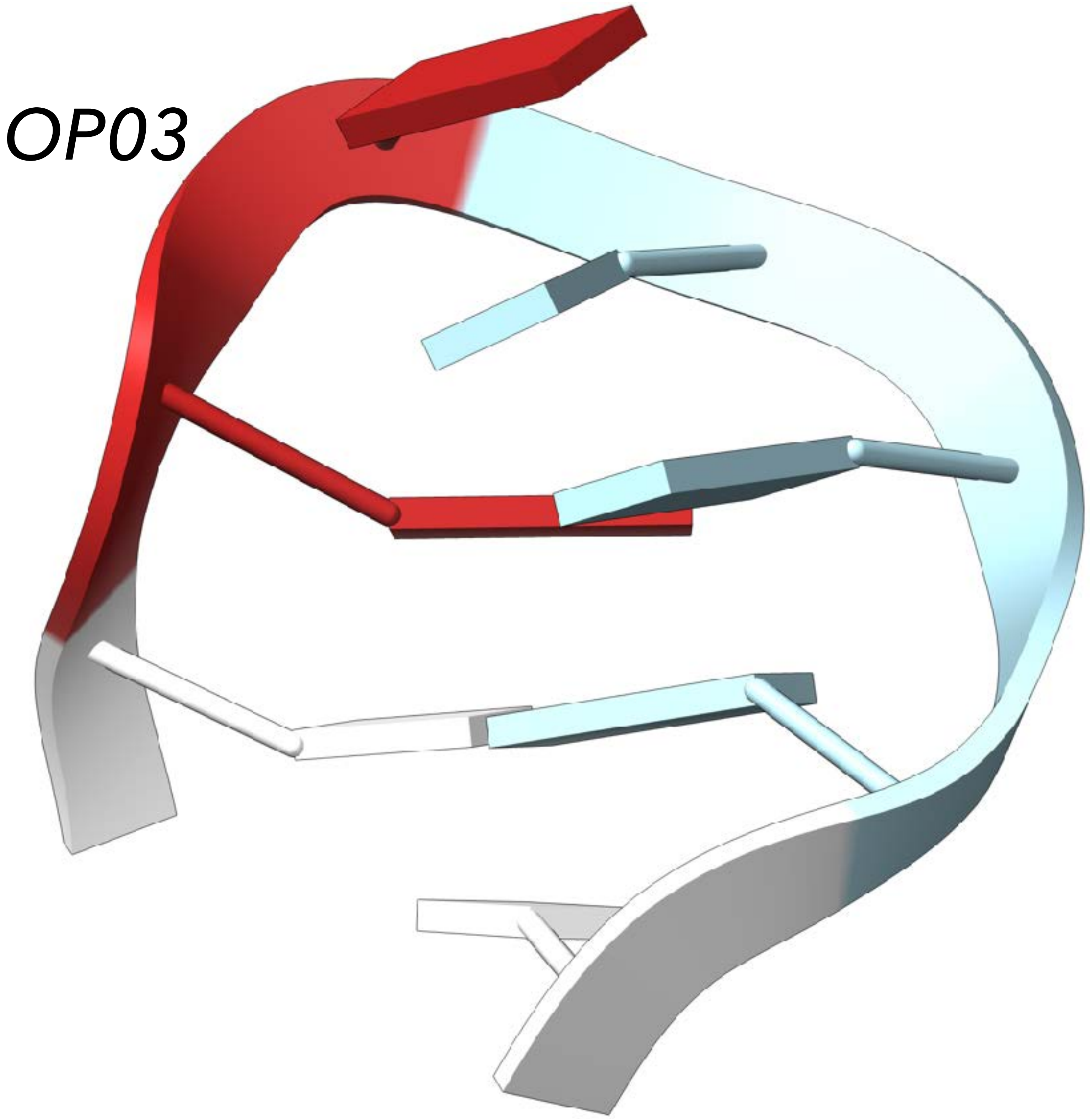
8 steps

10 steps

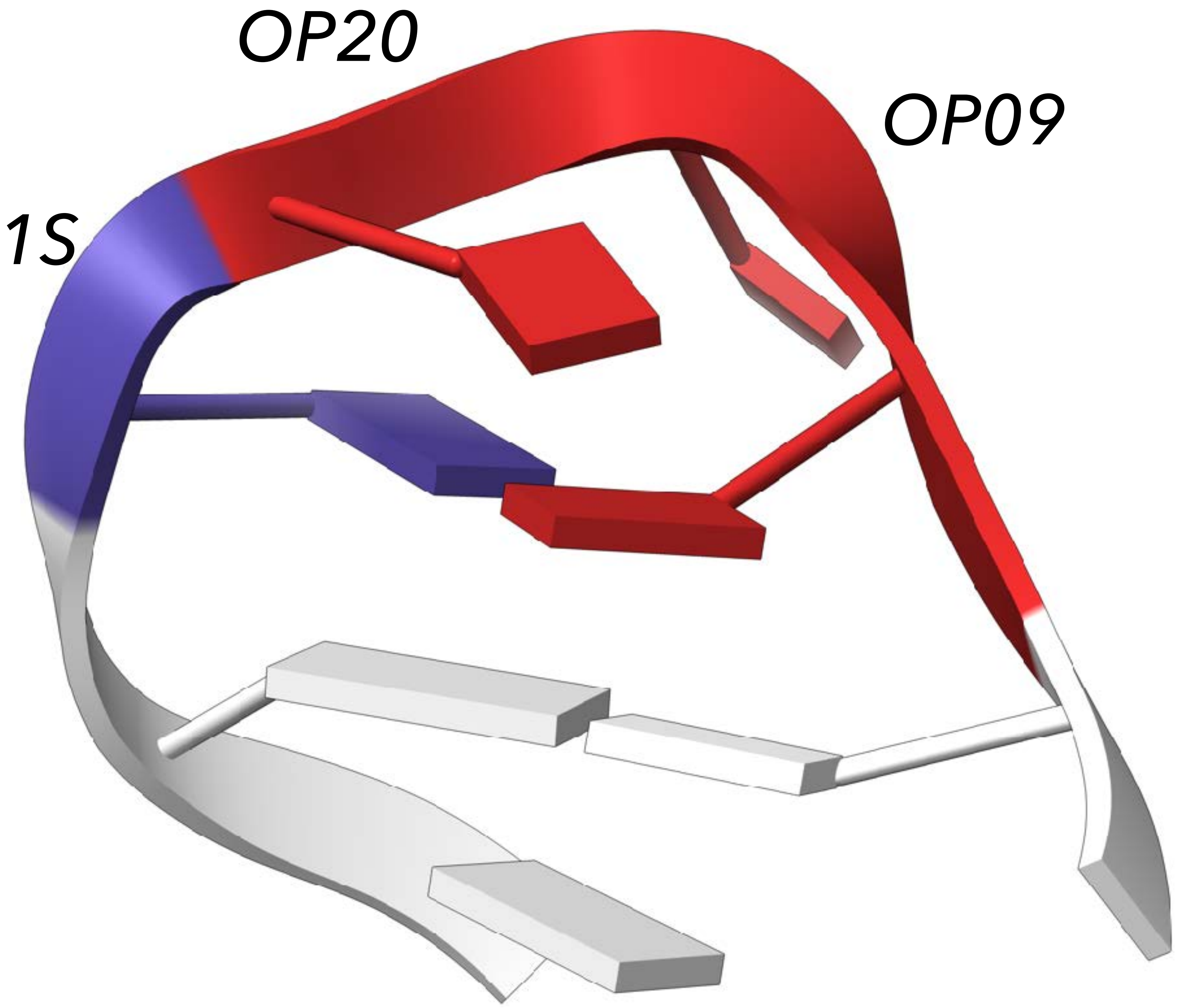
12 steps

Examples of tetraloop motifs:
sequential search → structural motifs

Deeper analysis



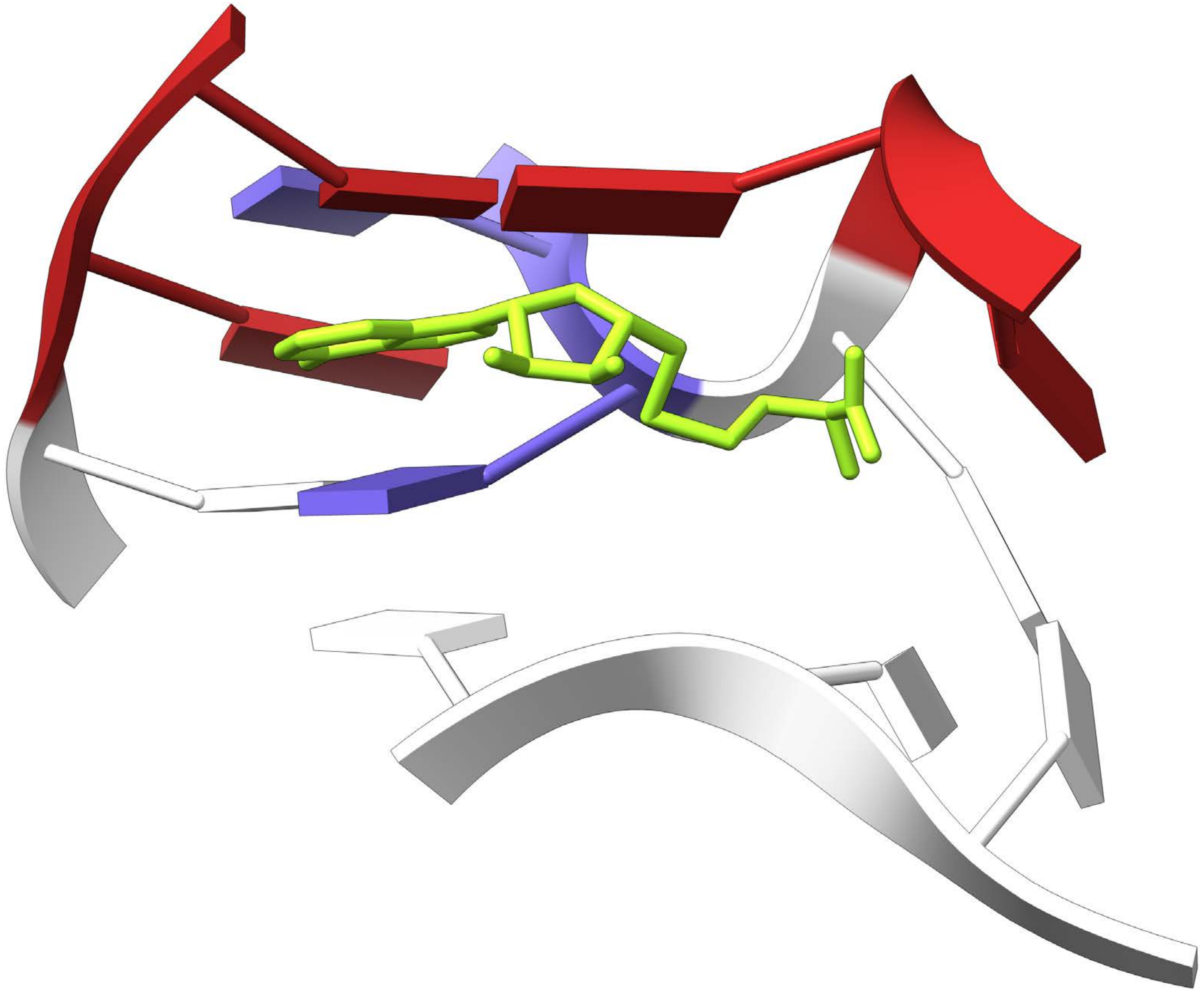
4lvz: relatively simple motif



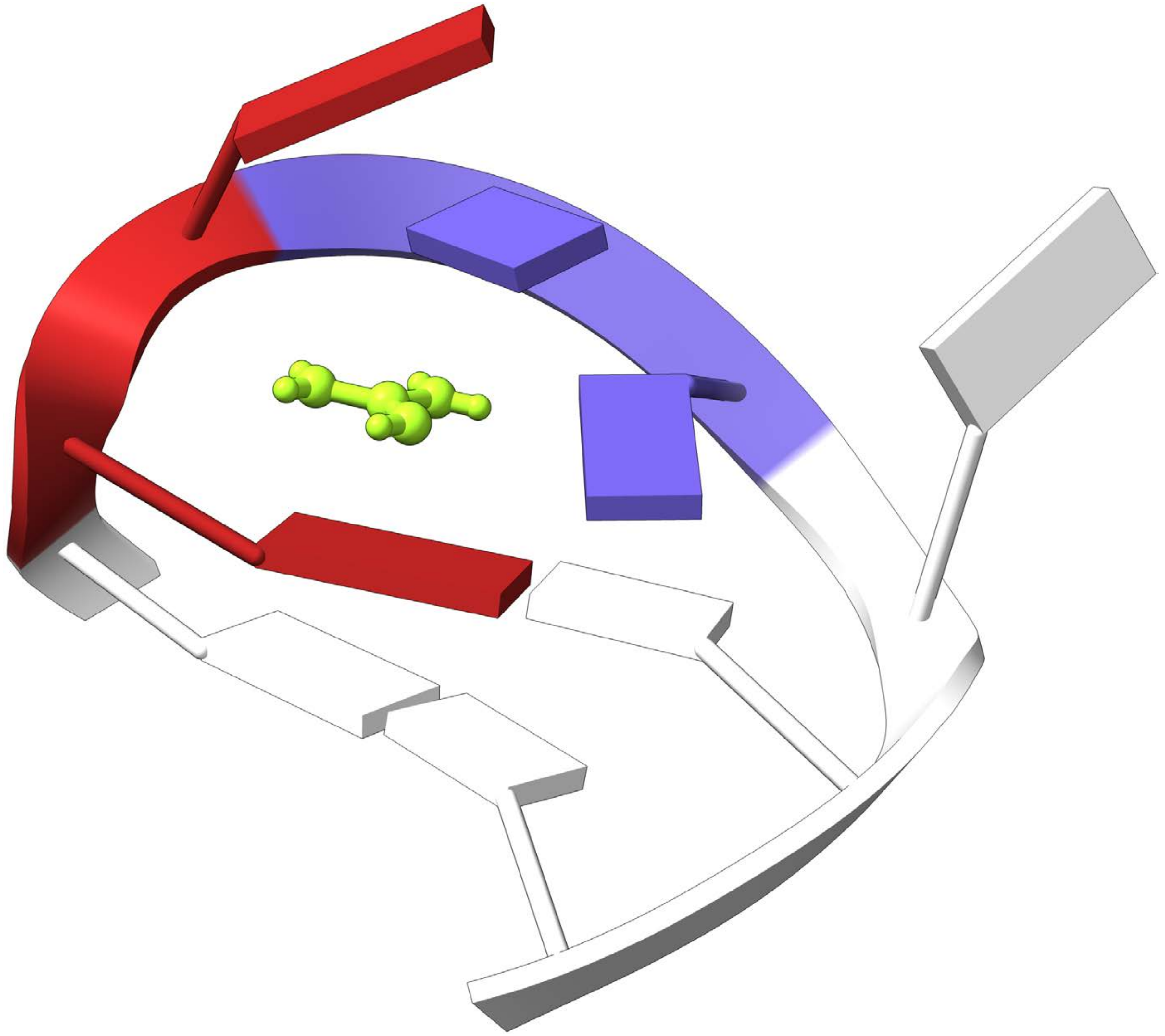
4qvi: made of rare NtCs

Černý et al.: NAR **48**: 6367 (2020).

Annotation of riboswitch binding sites: each is unique, contain many unassigned dint



4npq: S-adenosyl homocysteine riboswitch



5ndh: Guanidine II riboswitch

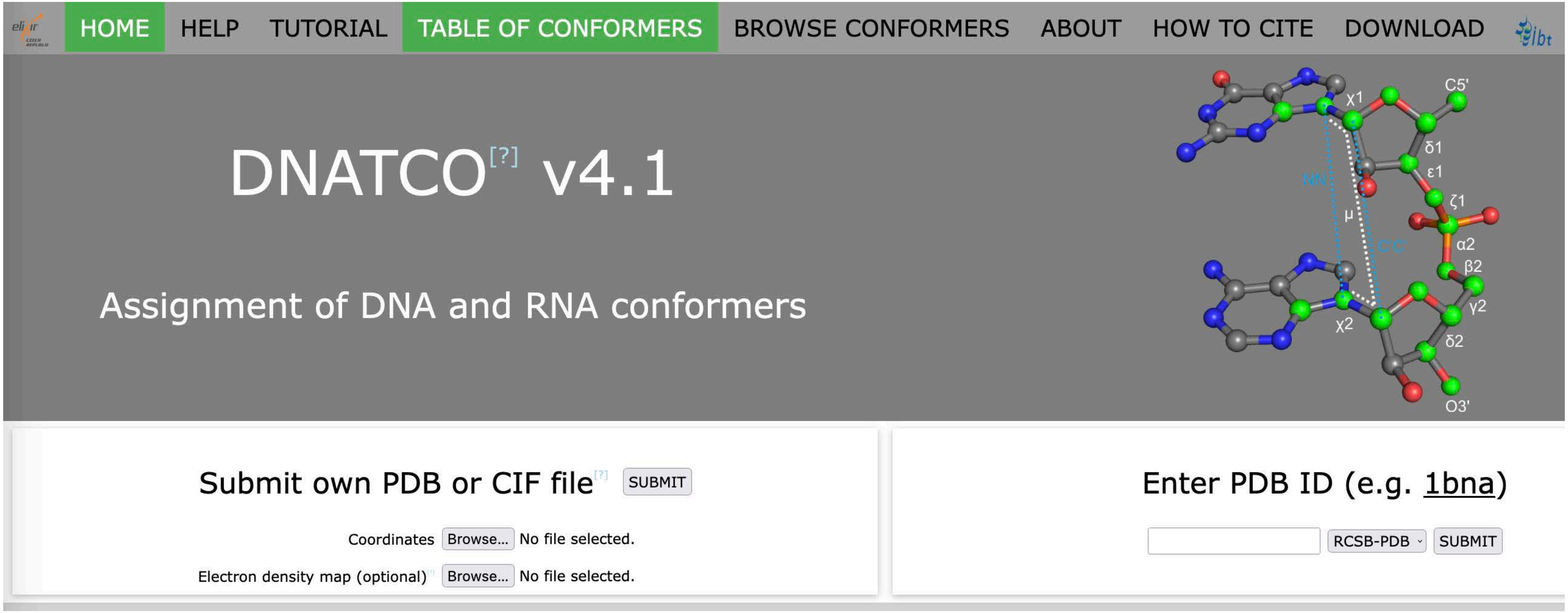
We assigned dinucleotide geometries across the database …

Bad news

- Of ~7 million dinucleotides assigned, *about 30% remain unassigned*
- We know that
  - ☹️no major *NtC* class is missing on our list
  - 😊geometries of up to 50% of unassigned close to the *NtC* geometries

- New refinement protocols possible & needed

Web services



DNATCO[?] v4.1

Assignment of DNA and RNA conformers

Submit own PDB or CIF file[?]

Coordinates No file selected.

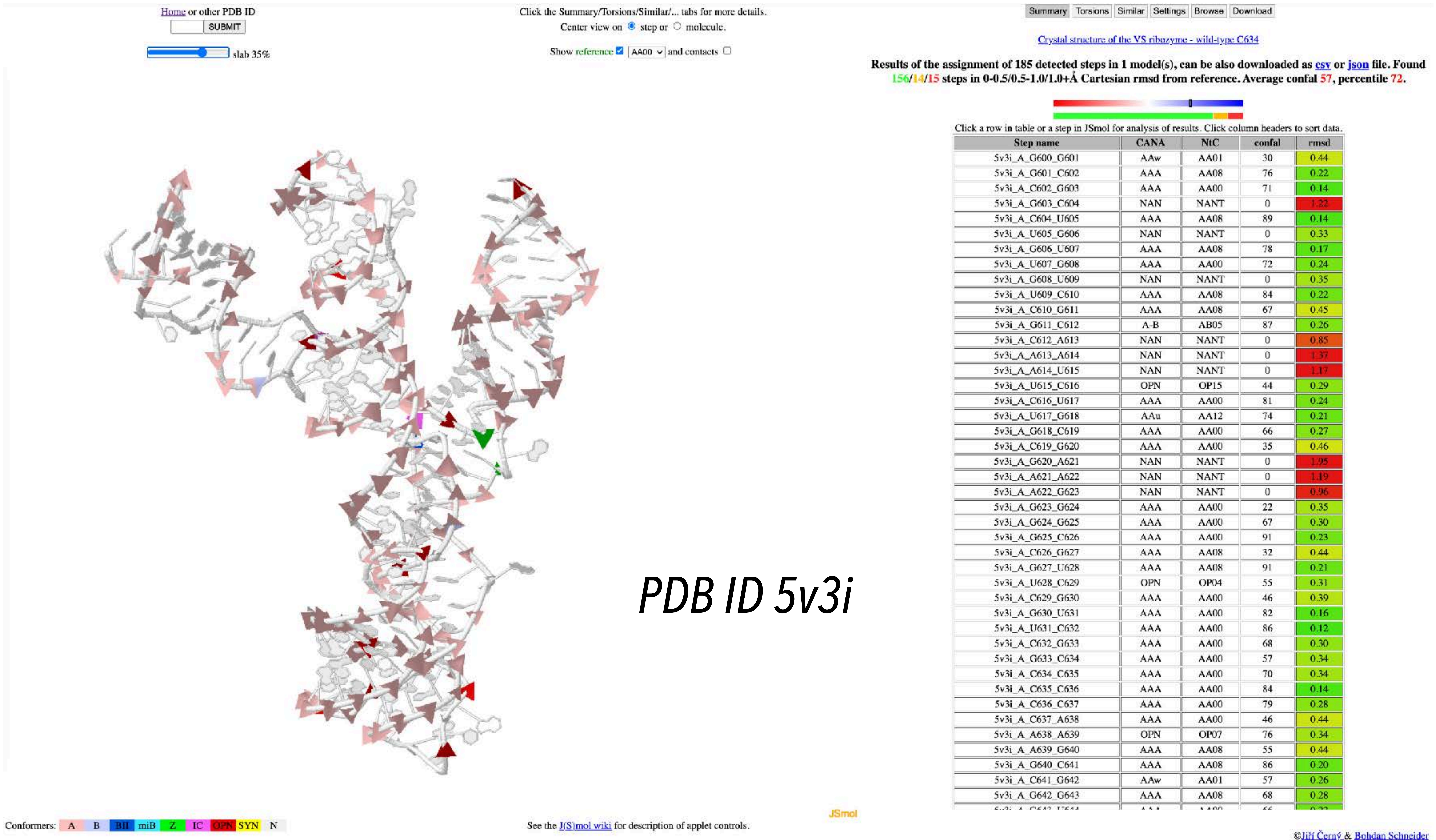
Electron density map (optional)[?] No file selected.

Enter PDB ID (e.g. 1bna)

dnatco.datmos.org: annotation tools from intuitive to expert

Černý, Božíková & Schneider: Acta Cryst **D76**: 806 (2020)

Annotation



Crystal structure of the VS ribozyme - wild-type C634

Results of the assignment of 185 detected steps in 1 model(s), can be also downloaded as csv or json file. Found 156/14/15 steps in 0-0.5/0.5-1.0/1.0+Å Cartesian rmsd from reference. Average confal 57, percentile 72.

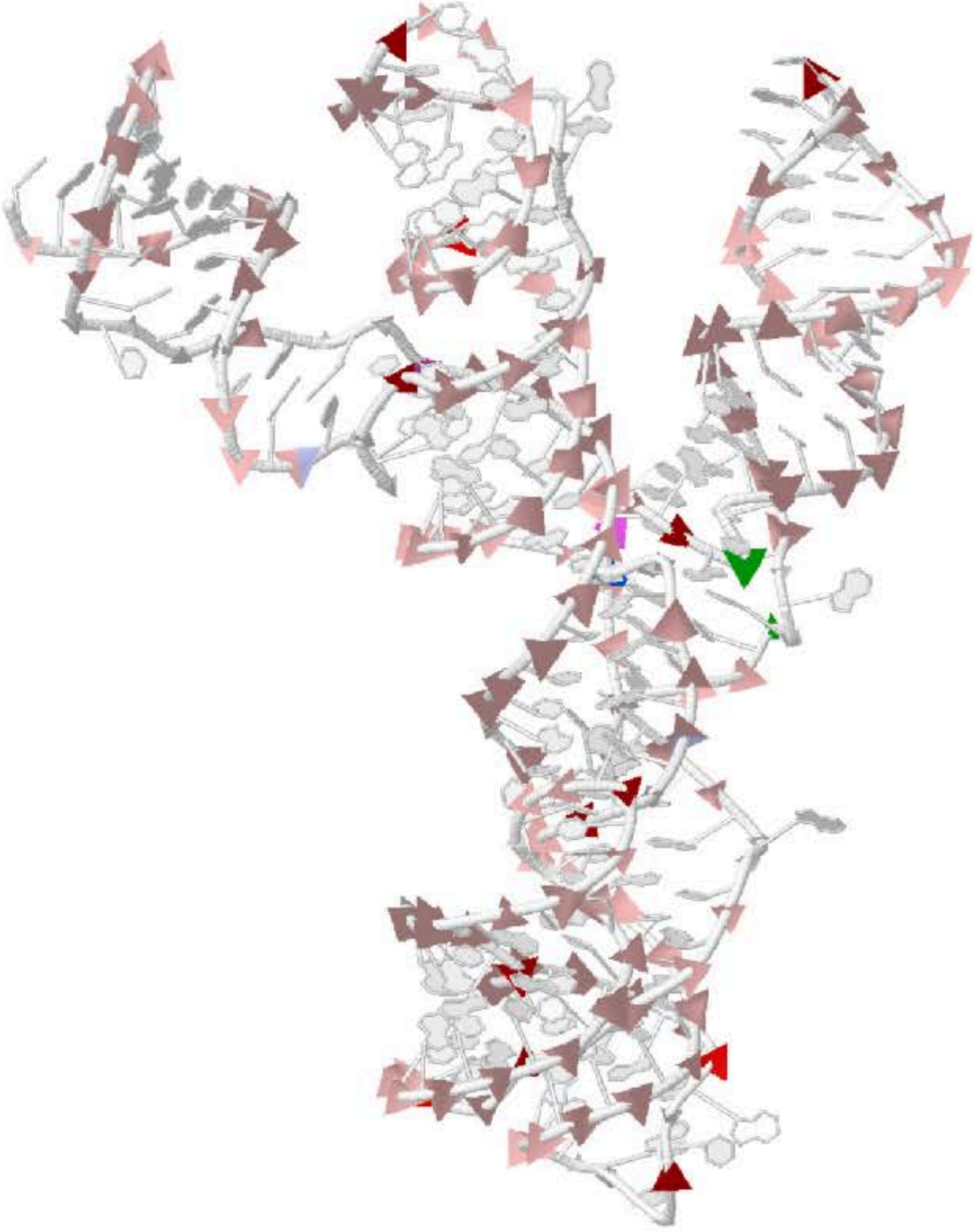
Click a row in table or a step in JSmol for analysis of results. Click column headers to sort data.

Step name	CANA	NtC	confal	rmsd
5v3i_A_G600_G601	AAw	AA01	30	0.44
5v3i_A_G601_C602	AAA	AA08	76	0.22
5v3i_A_C602_G603	AAA	AA00	71	0.14
5v3i_A_G603_C604	NAN	NANT	0	1.22
5v3i_A_C604_U605	AAA	AA08	89	0.14
5v3i_A_U605_G606	NAN	NANT	0	0.33
5v3i_A_G606_U607	AAA	AA08	78	0.17
5v3i_A_U607_G608	AAA	AA00	72	0.24
5v3i_A_G608_U609	NAN	NANT	0	0.35
5v3i_A_U609_C610	AAA	AA08	84	0.22
5v3i_A_C610_G611	AAA	AA08	67	0.45
5v3i_A_G611_C612	A-B	AB05	87	0.26
5v3i_A_C612_A613	NAN	NANT	0	0.85
5v3i_A_A613_A614	NAN	NANT	0	1.37
5v3i_A_A614_U615	NAN	NANT	0	1.17
5v3i_A_U615_C616	OPN	OP15	44	0.29
5v3i_A_C616_U617	AAA	AA00	81	0.24
5v3i_A_U617_G618	AAu	AA12	74	0.21
5v3i_A_G618_C619	AAA	AA00	66	0.27
5v3i_A_C619_G620	AAA	AA00	35	0.46
5v3i_A_G620_A621	NAN	NANT	0	1.95
5v3i_A_A621_A622	NAN	NANT	0	1.19
5v3i_A_A622_G623	NAN	NANT	0	0.96
5v3i_A_G623_G624	AAA	AA00	22	0.35
5v3i_A_G624_G625	AAA	AA00	67	0.30
5v3i_A_G625_C626	AAA	AA00	91	0.23
5v3i_A_C626_G627	AAA	AA08	32	0.44
5v3i_A_G627_U628	AAA	AA08	91	0.21
5v3i_A_U628_C629	OPN	OP04	55	0.31
5v3i_A_C629_G630	AAA	AA00	46	0.39
5v3i_A_G630_U631	AAA	AA00	82	0.16
5v3i_A_U631_C632	AAA	AA00	86	0.12
5v3i_A_C632_G633	AAA	AA00	68	0.30
5v3i_A_G633_C634	AAA	AA00	57	0.34
5v3i_A_C634_C635	AAA	AA00	70	0.34
5v3i_A_C635_C636	AAA	AA00	84	0.14
5v3i_A_C636_C637	AAA	AA00	79	0.28
5v3i_A_C637_A638	AAA	AA00	46	0.44
5v3i_A_A638_A639	OPN	OP07	76	0.34
5v3i_A_A639_G640	AAA	AA08	55	0.44
5v3i_A_G640_C641	AAA	AA08	86	0.20
5v3i_A_C641_G642	AAw	AA01	57	0.26
5v3i_A_G642_G643	AAA	AA08	68	0.28

PDB ID 5v3i

See the J(S)mol wiki for description of applet controls.

©Jiří Černý & Bohdan Schneider



Prevailing A form

→

Highlight of "hot" regions

Future development: open our tools for the community

- Reorganize *dnatco.datmos.org*
  - annotation, validation, refinement, modeling tabs
- Help to improve refinement and modeling pipelines
  - Phenix
  - Refmac
  - Macromolecular Model Builder (MMB) by Sam Flores uses ***NtC*** to build from 2D or fit to electron density
- Collaborate on development of the NAKB

The Czech Academy of Sciences

BIOCEV

Institute of Biotechnology of the Czech Academy of Sciences

Thank you for your time!

dnatco.datmos.org

bohdan.schneider@gmail.com

Building the capacity:
CZ.02.1.01/0.0/0.0/16_013/0001777

instruct Integrating Biology

CIISB project:
LM2015043

MINISTRY OF EDUCATION, YOUTH AND SPORTS

LTAUSA18197