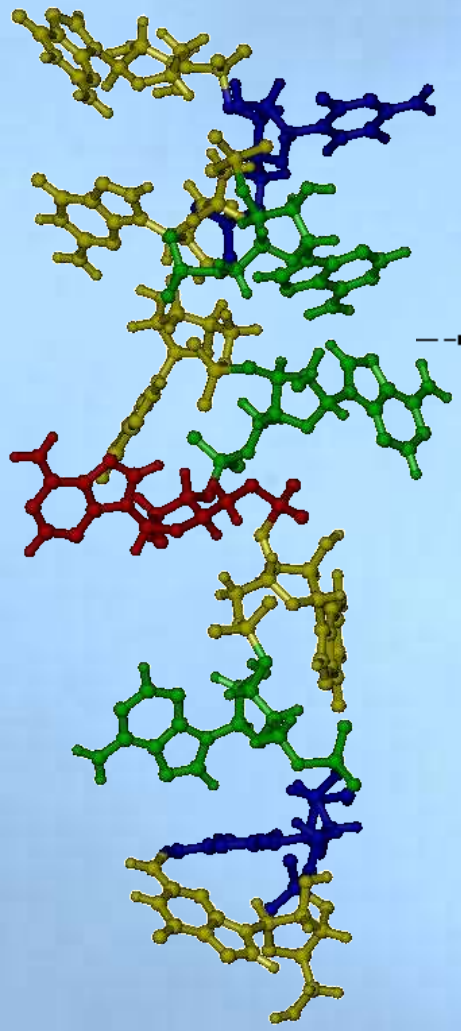
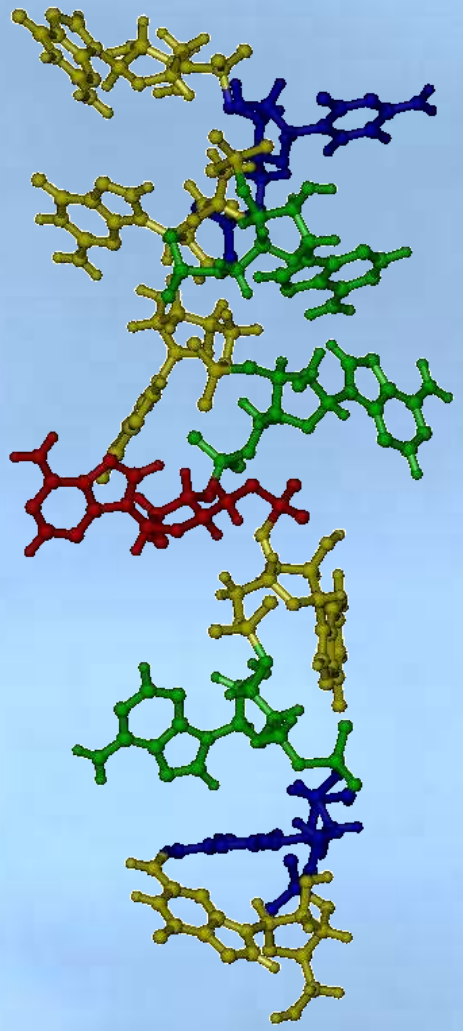


RNA fragments database: project & implementation

Agenda

1. Introduction
   - trends in structural bioinformatics
   - about RNA...
2. Project of RNA fragments database
   - main concept
   - database scheme & contents
3. RNA FRABASE
   - implementation
   - search examples
   - performance



Structural bioinformatics

Basic streams in structural bioinformatics:

--> • Structure determination

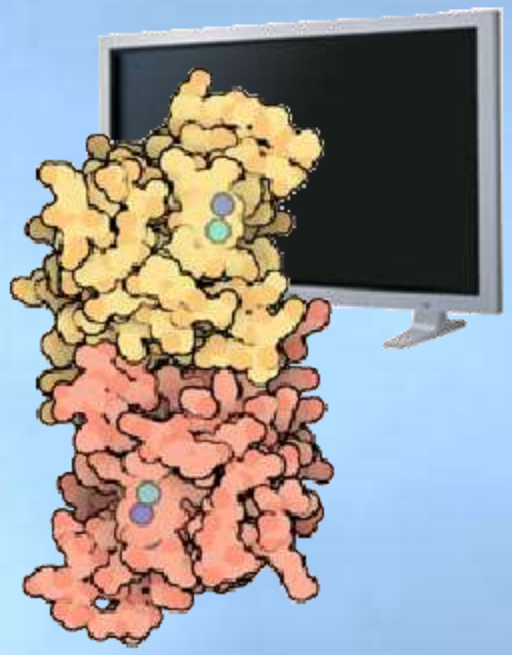
• Discovering dependencies: molecular structure - functions

--> • Structure prediction

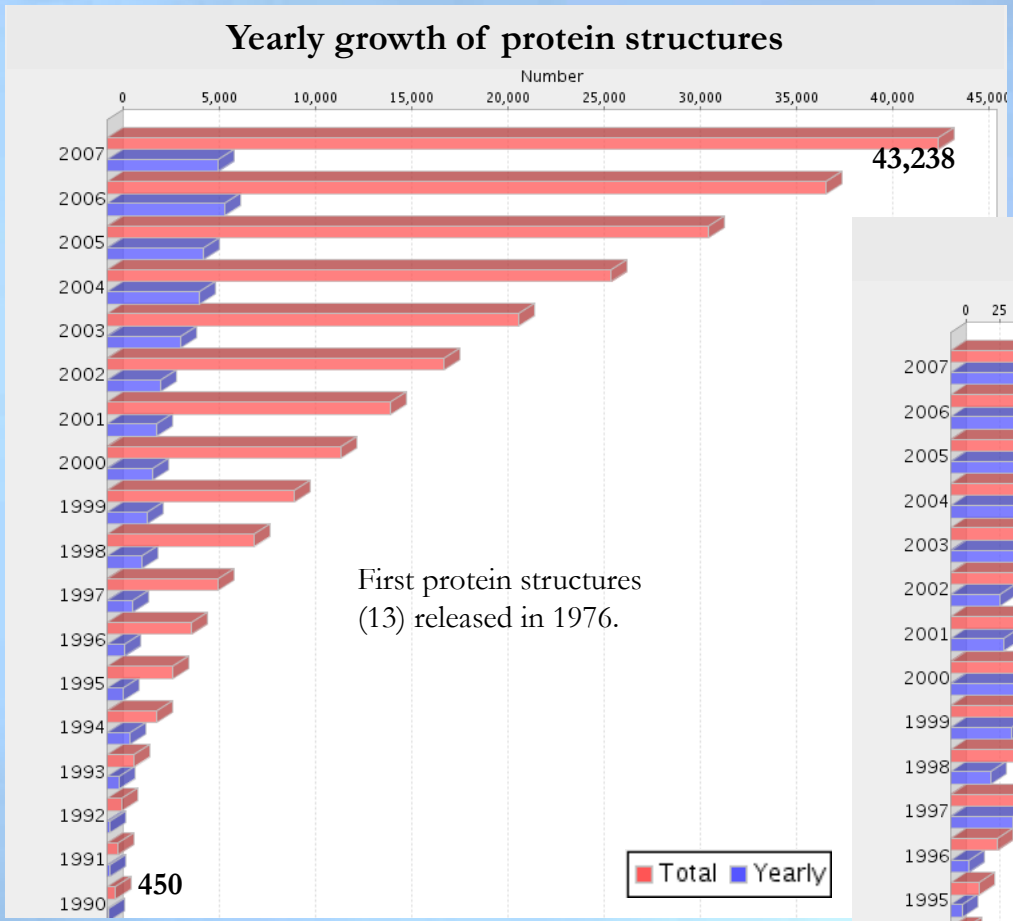
--> • Structure comparison

⇨ • Structural databases

• Visualization of molecular models

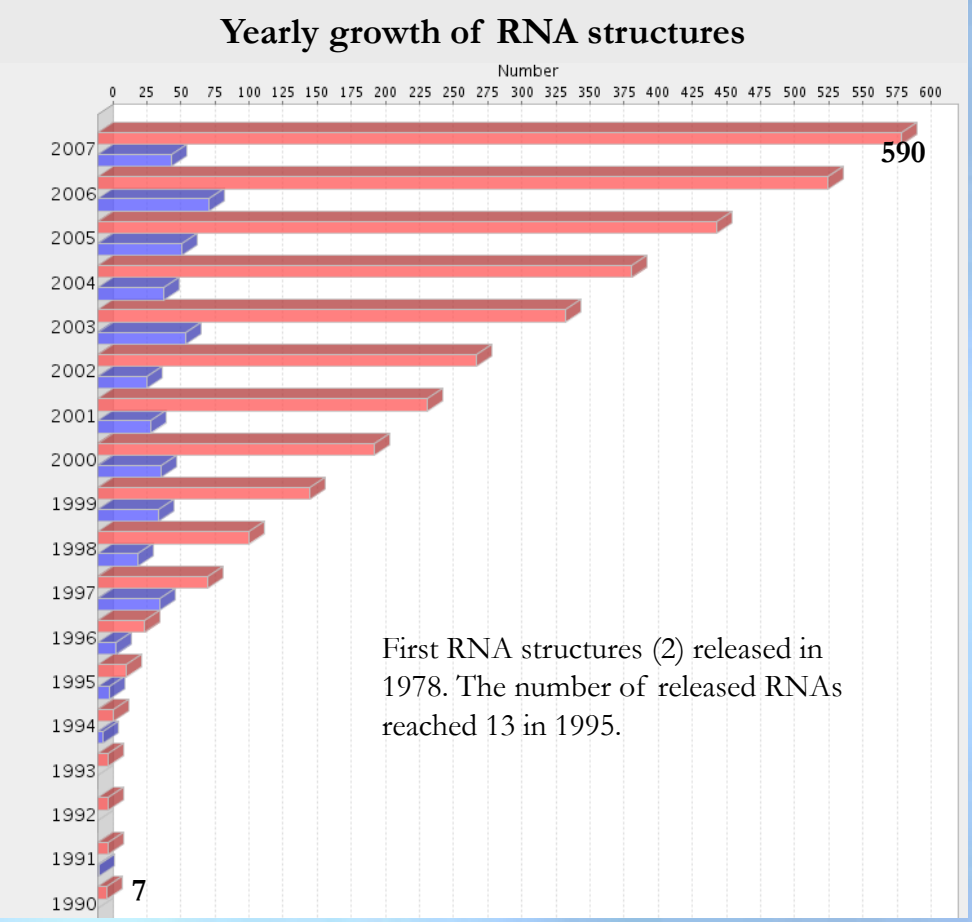


Trends in structural research



First protein structures (13) released in 1976.

Growth of released structures per year: proteins vs RNA



First RNA structures (2) released in 1978. The number of released RNAs reached 13 in 1995.

Statistics updated: Tuesday, 30.10.2007

About RNA...

Proteins + DNA + RNA = molecular model of life

C G U A

- transmits genetic information from DNA into proteins,
- catalyzes chemical reactions,
- controls certain chemical processes in the cell,
- initiates the process of gene silencing (RNAi),
- forms the genetic material of some viruses,
- is a component of the ribosomes.

Nobel 2006

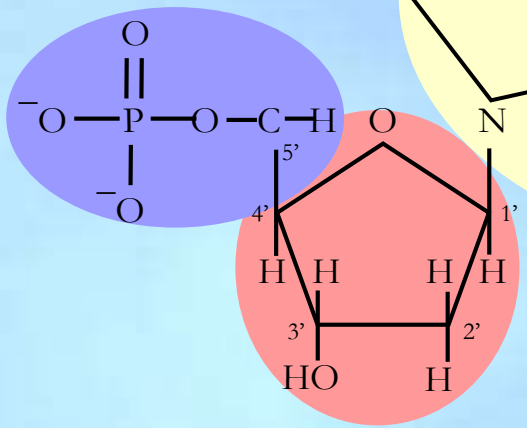
Andrew Z. Fire i Craig C. Mello - laureaci Nagrody Nobla z medycyny i fizjologii w 2006 roku.

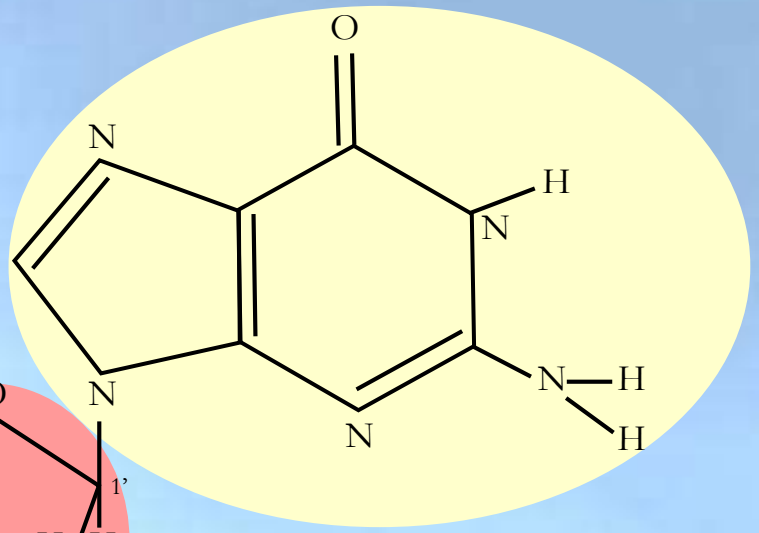
Fot. MICHAEL PROBST AP

RNA structure

basic component of RNA chain: nucleotide

phosphate group





sugar: ribose

organic base:

Adenine

Guanine

Cytosine

Uracil

I- and II-dimensional structure of RNA

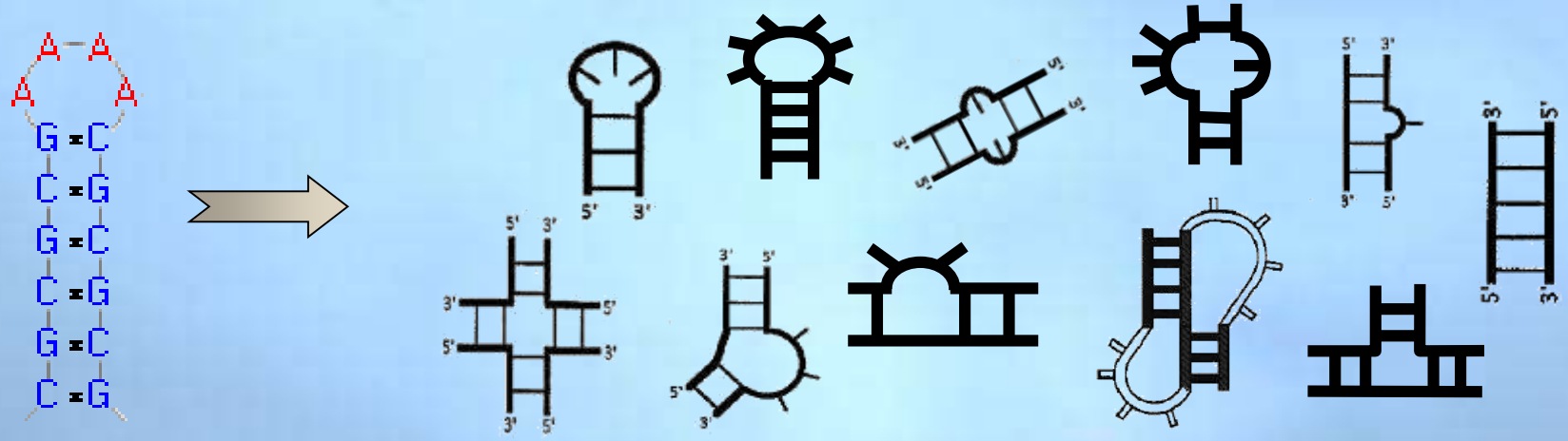
Primary structure

describes nucleotide sequence in the chain

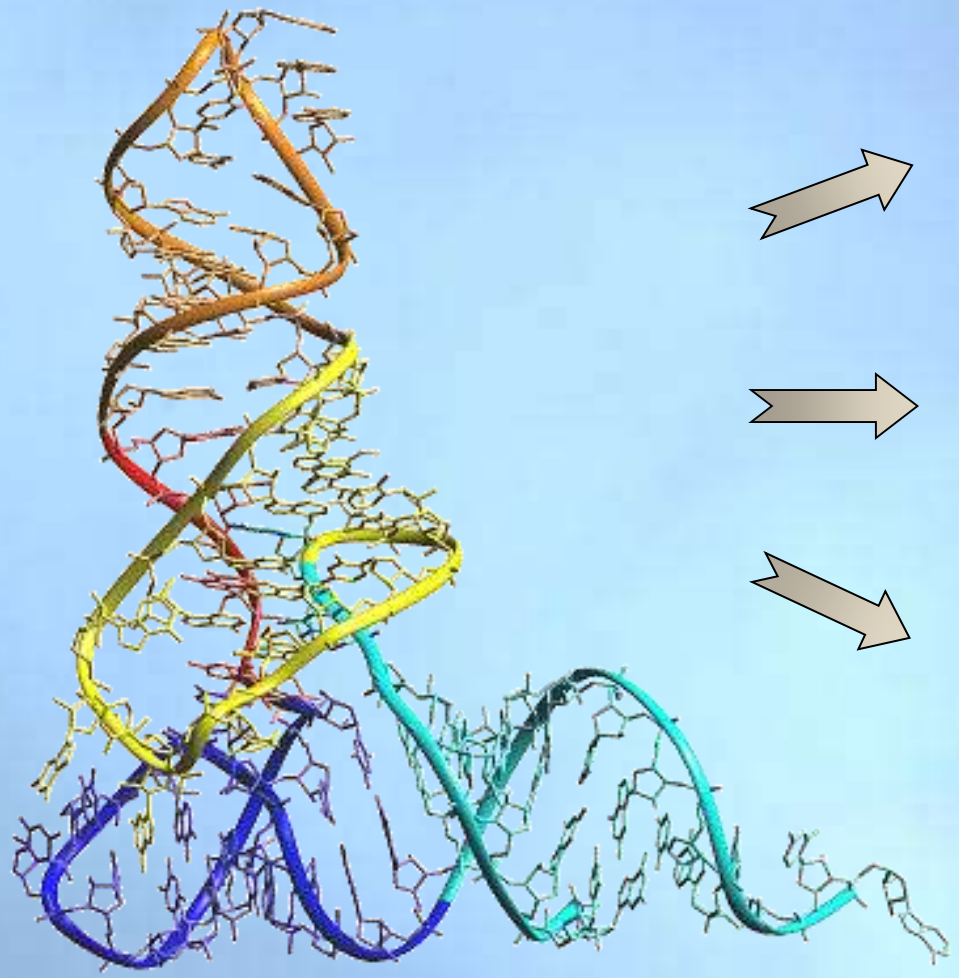
$$5' - C_1\ G_2\ C_3\ G_4\ A_5\ U_6\ C_7\ U_8\ G_9 - 3'$$

Secondary structure

describes one- and two-dimensional fragments (non-/canonical base-pairs)



III-dimensional structure of RNA



Tertiary structure is a 3D shape an entire chain assumes.

Determines molecule physiological roles inside a cell.

Describes ribose ring folding, torsion angles, base-pairing (canonical, Watson-Crick A:U, C:G, non-canonical>100 types).

Goals of structure determination

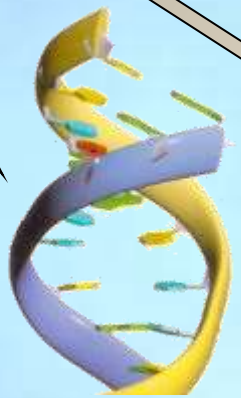
molecular structure

relation

biological roles

designing new medicaments

growing new specimens of plants and animals

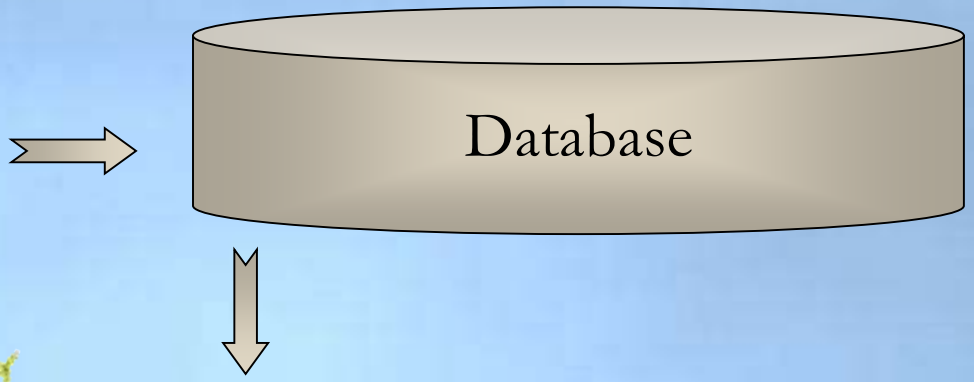


explaining biological functions of molecules

disease identification

What follows structure determination...

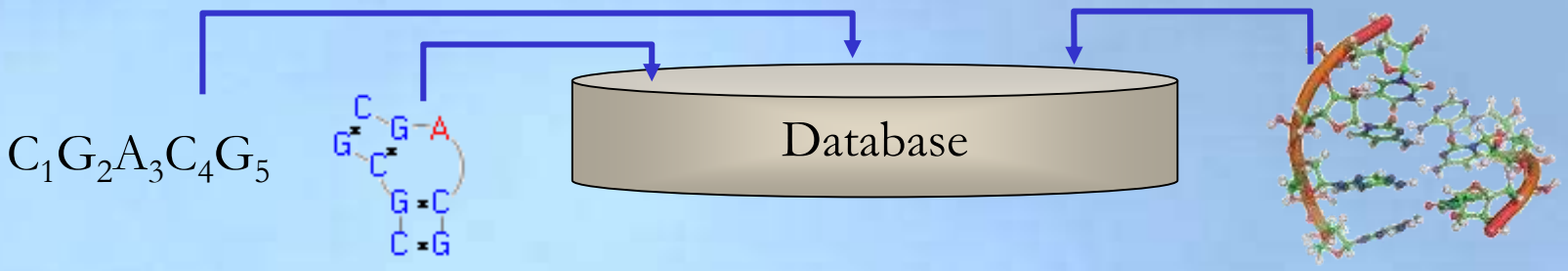
1DRZ structure



- Protein Data Bank (PDB)
  http://www.rcsb.org/pdb/
- Nucleic Acid Database (NDB)
  http://ndbserver.rutgers.edu/
- The RNA Structure Database (RNABase)
  http://www.rnabase.org/
- The European Ribosomal RNA Database (rRNA)
  http://www.psb.ugent.be/rRNA/
- A Database of Metal Ion Binding Sites in RNA (MERNA)
  http://merna.lbl.gov/
- The Structural Classification of RNA (SCOR)
  http://scor.lbl.gov/scor.html

RNA fragments database: main concept

Construct a relational database storing the information about primary, secondary and tertiary structures of RNA molecules and / or their fragments.



Find an answer to the following problems / questions…

⇨ look for 3D structures of RNA with the given sequence …

C G C G U A C …

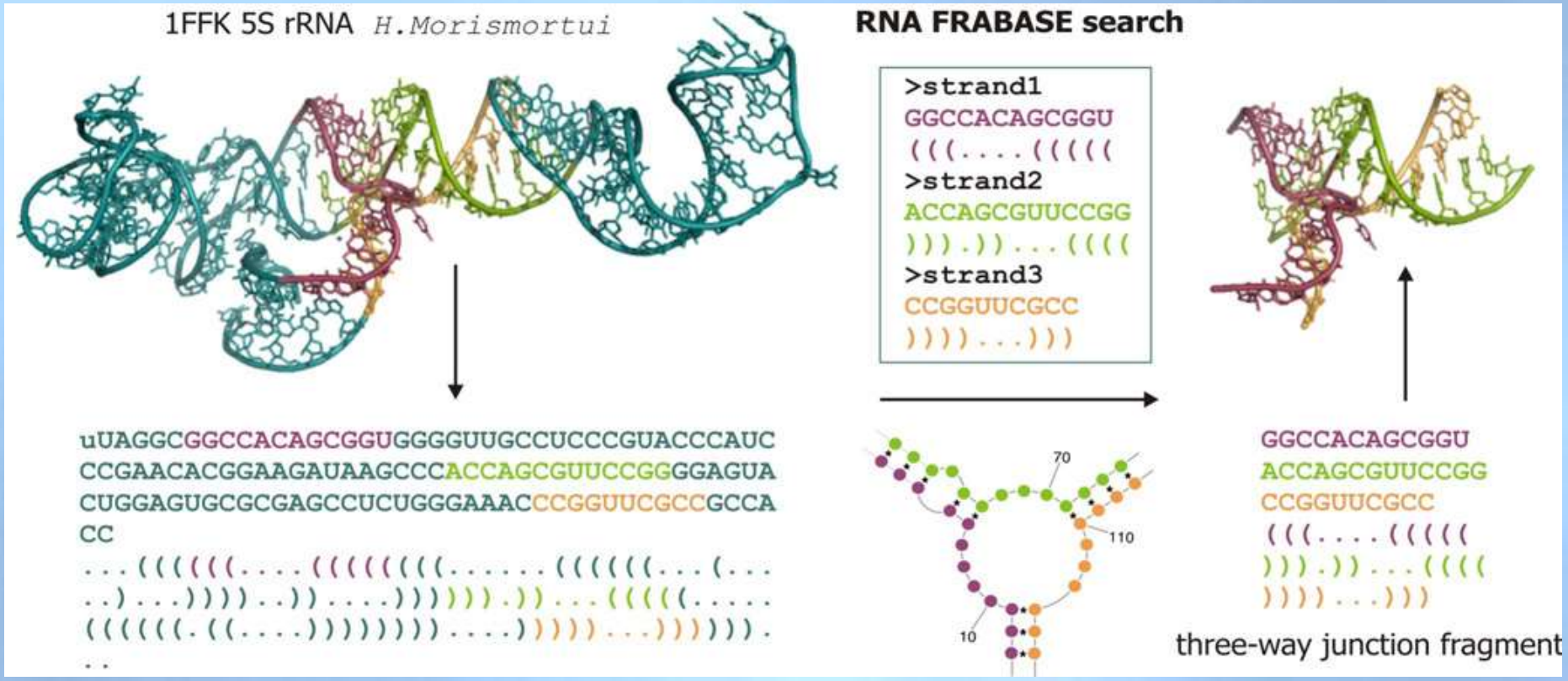
⇨ look for 3D structures of RNA with the given secondary structure motif …

RNA fragments database: main concept

Look for 3D structure of the RNA fragment defined by:

(a) Sequence and / or

(b) Secondary structure

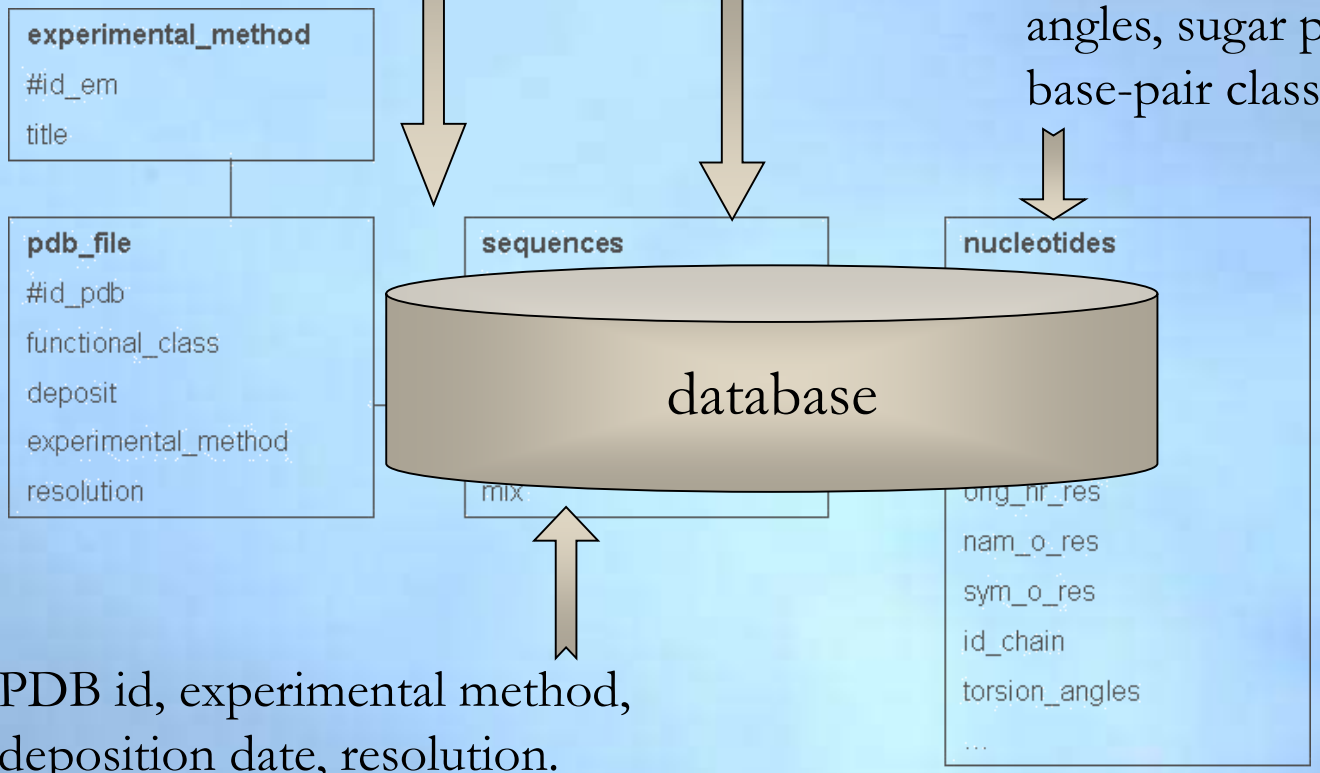


RNA fragments database: contents

Primary structure (sequence) given in IUPAC codes.

Secondary structure given in dot-bracket notation.

Atom coordinates, torsion angles, sugar pucker parameters, base-pair classification



PDB id, experimental method, deposition date, resolution.

Primary structure encoding

IUPAC-IUB codes for RNA

Code	Description
A	Adenine
C	Cytosine
G	Guanine
U	Uracil
R	Purine {A, G}
Y	Pyrimidine {C, U}
M	{C, A}
K	{U, G}
W	{U, A}
S	{C, G}
B	{C, U, G} (not A)
D	{A, U, G} (not C)
H	{A, U, C} (not G)
V	{A, C, G} (not U)
N	any base {A, C, G, U}

The sequence in the database is defined with the code over 4-letter alphabet **{A, C, G, U}**.

Upper-case letter codes for unmodified residue, lower-case letter codes for modified residue.

Pattern sequence can be defined with **IUPAC-IUB** code for RNA.

Search engine uses regular expression matching procedures to find the pattern within the deposited RNA structures.

Secondary structure encoding

Dot-bracket notation: . = unpaired residue ( ) = base-pair

bulge

```
( ( ( (
) ) . ) )
```

hairpin loop

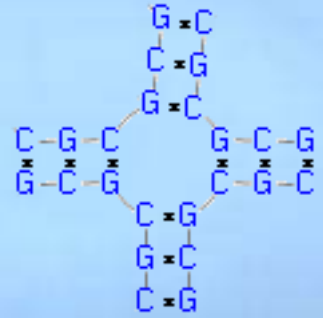
```
( ( ( . . . ) ) )
```

internal loop

```
( ( ( ( . . . ( ( ( (
) ) ) ) . . ) ) ) )
```

4-way junction

```
( ( ( ( ( (
) ) ) ( ( (
) ) ) ( ( (
) ) ) ) ) )
```



kissing loops

```
( ( ( ( . . [ . . ) ) ) )
( ( ( ( . . ] . . ) ) ) )
```

Filtering structures

Look for structures that suit the pattern and satisfy the given requirements:

(a) have been determined by the given method (e.g. X-ray, NMR),

(b) having torsion angle value(s) within the given range,

(c) with the given sugar pucker parameters, etc.

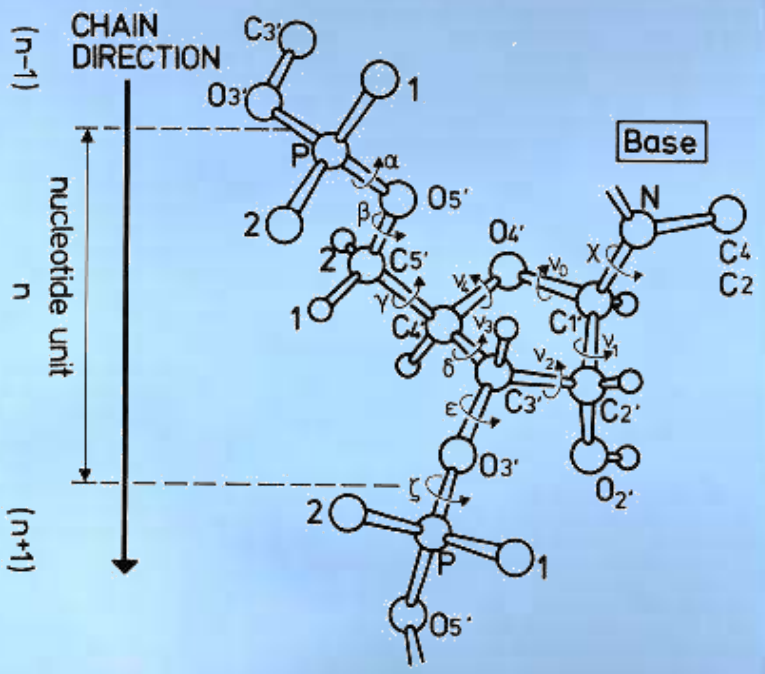
<u>example</u>

Res. 2 contains torsion angles similar to those in B-DNA structure

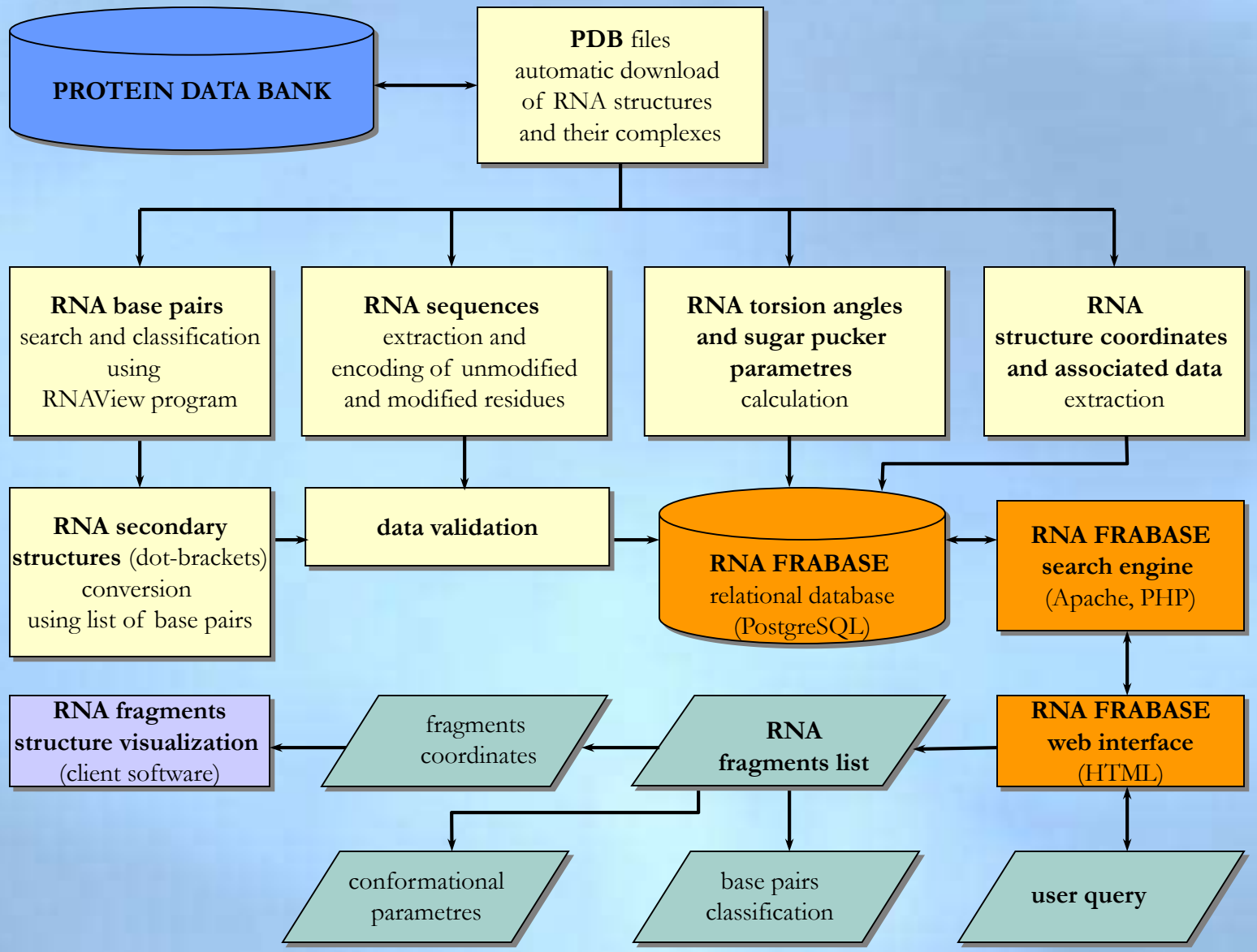
```
2 N ALPHA    -46  -36
2 N BETA     131  141
2 N GAMMA     33   43
2 N DELTA    134  144
2 N CHI     -137  -67
2 N EPSILON -138 -128
2 N ZETA    -163 -157
```

Residue 3 is in anti & C3'-endo conformation. Res. 4 is in syn & C2'-endo conformation (conformations characteristic for left-handed structures).

```
3 C CHI  -180 -110
3 C P       0   36
4 G CHI    50   90
4 G P     144  180
```



RNA fragments database: scheme



RNA FRABASE - RNA FRAgments search engine & dataBASE - Mozilla Firefox

Plik Edycja Widok Historia Zakładki ScrapBook Narzędzia Pomoc

Wstecz Do przodu Odśwież Zatrzymaj Strona startowa http://rnafrabase.ibch.poznan.pl/index2.php?act=examples&act2=6 Przejdź Google Szukaj

RNA FRABASE version 1.0

RNA FRAgments search engine & dataBASE

Search
About
Help
PDB list
References
Links
Contact us

You are **857.** visitor.

Currently **1** user(s) on-line.

Supported by
Laboratory of Structural Chemistry of Nucleic Acids,
Institute of Bioorganic Chemistry,
Polish Academy of Sciences

RNA FRABASE: an engine with database to search the three-dimensional fragments within 3D RNA structures using as an input the sequence(s) and / or secondary structure(s) given in the dot-bracket notation.

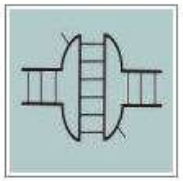
Database contains (updated **2007-07-30**):

- RNA sequences and secondary structures, described in dot-bracket notation, derived from the **1065** PDB-deposited RNA structures and their complexes,
- atom coordinates of the unmodified and mod
PDB-deposited RNA structures,
- torsion angle values, sugar pucker paramete

Sequence(s) and / or secondary structure(s):

Load example: 1 2 3 4 5 6

```
#Kissing loops
>strand1
(((.[[[[[[)))
>strand2
(((.]]]]]]]))
```



Search Advanced Search

Check for modified residues sensitive search: ☐

Experimental method: X-Ray Difraction ☑ NMR ☑ Electron Microscopy ☑ Other ☑

Resolution ≤ Å

Limits for pseudorotation parameters, sugar pucker amplitude and torsion angles:

#all	alpha	-180	180
#all	beta	-180	180
#all	gamma	-180	180
#all	delta	-180	180
#all	epsilon	-180	180
#all	zeta	-180	180
#all	chi	-180	180
#all	p	-180	180
#all	vmax	-180	180
#all	v0	-180	180
#all	v1	-180	180

Search Reset

http://rnafrabase.ibch.poznan.pl/index2.php?act=examples&act2=6

RNA FRABASE: *Search* and *Advanced Search*

Submission results

Number of matching fragments: **33**

Select all | Clear selection | Show coordinates | Show torsion angles | Show base pairs | Save table

	No.	PDB id	sequence	secondary structure	chain	start	end	method	class	date	Å
☐	1	1FFK	CAGGCAUCGACUG CUGCUUGAUGCG	(((.[[[[[(((.]]]]]							
☑	2	1JJ2	CAGGCAUCGACUG CUGCUUGAUGCG	(((.[[[[[(((.]]]]]							
☐	3	1K73	CAGGCAUCGACUG CUGCUUGAUGCG	(((.[[[[[(((.]]]]]							

Base pairs classification for selected fragment(s) → Save table

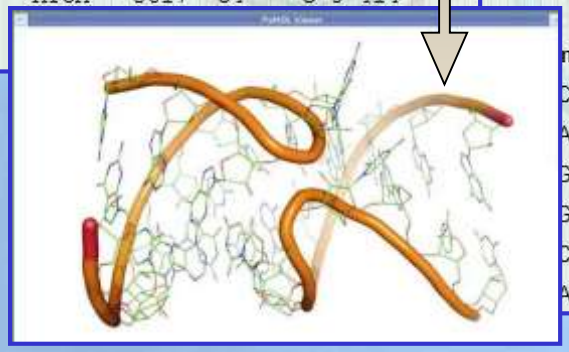
residue I					residue II				classification	
PDB	name	symbol	chain	res no.	PDB name	symbol	chain	res no.	Westhof's notation	Saenger's notation
	C	C	0	414	G	G	0	426	+/+ cis	XIX
								25	-/- cis	XX
								24	+/+ cis	XIX
								49	+/+ cis	XIX
								48	-/- cis	XX
								47	-/- cis	XX

Atom coordinates for selected fragment(s) → Save table

« Back

No.	PDB id	sequence	secondary structure	chain	start	end	method	class	date	Å
2	1JJ2	CAGGCAUCGACUG CUGCUUGAUGCGG	(((.[[[[[[]))) (	0	414	426	X-Ray	RIBOSOME	2001-07-03	2.4

```
ATOM   8611  P    C 0 414
ATOM   8612  O1P  C 0 414
ATOM   8613  O2P  C 0 414
ATOM   8614  O5*  C 0 414
ATOM   8615  C5*  C 0 414
ATOM   8616  C4*  C 0 414
ATOM   8617  O4*  C 0 414
```



Torsion angles for selected fragment(s) → Save table

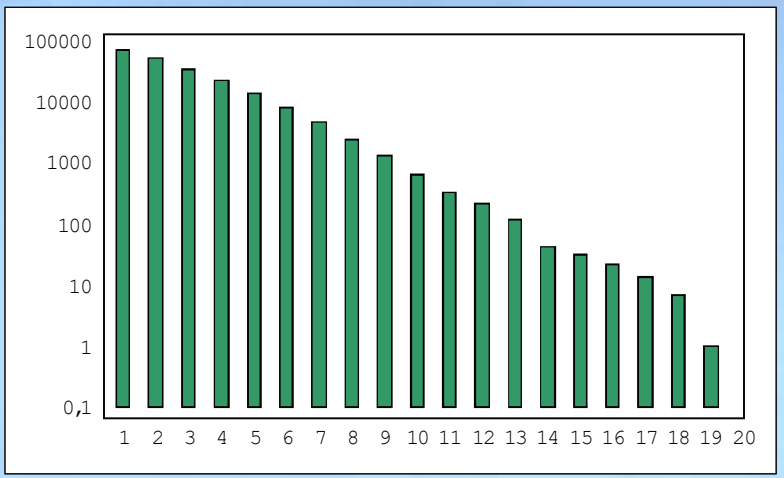
« Back

No.	PDB id	sequence	secondary structure	chain	start	end	method	class	date	Å
2	1JJ2	CAGGCAUCGACUG CUGCUUGAUGCGG	(((.[[[[[[]))) (((.]]]]]]]))))	0 0	414 2440	426 2452	X-Ray Difraction	RIBOSOME	2001-07-03	2.4

name	symbol	str.	chain	res no.	α	β	γ	δ	ε	ζ	χ	P	vmax	v0	v1	v2	v3	v4
C	C	(	0	414	-66.7	166.4	58.2	81.0	-152.0	-72.3	-162.3	11.7	42.6	5.2	-29.8	41.7	-39.7	21.7
A	A	(	0	415	-65.6	176.7	60.2	87.2	-157.2	-68.2	-166.2	14.5	36.5	2.7	-24.3	35.4	-34.9	20.3
G	G	(	0	416	-68.2	-177.4	50.5	85.6	-126.4	72.8	-148.1	15.7	35.2	2.0	-22.9	33.8	-34.0	20.2
G	G	.	0	417	-127.4	-100.0	76.8	130.9	-93.1	176.0	-93.9	151.9	31.0	-23.2	31.3	-27.4	14.7	5.2
C	C	[	0	418	-54.3	-142.6	57.6	82.8	-159.8	-70.2	-165.7	12.0	38.9	4.7	-27.2	38.0	-36.6	20.1
A	A	[	0	419	-71.5	-172.2	50.9	85.4	-149.0	-74.7	-154.8	9.7	38.5	6.1	-28.0	38.0	-35.6	18.5

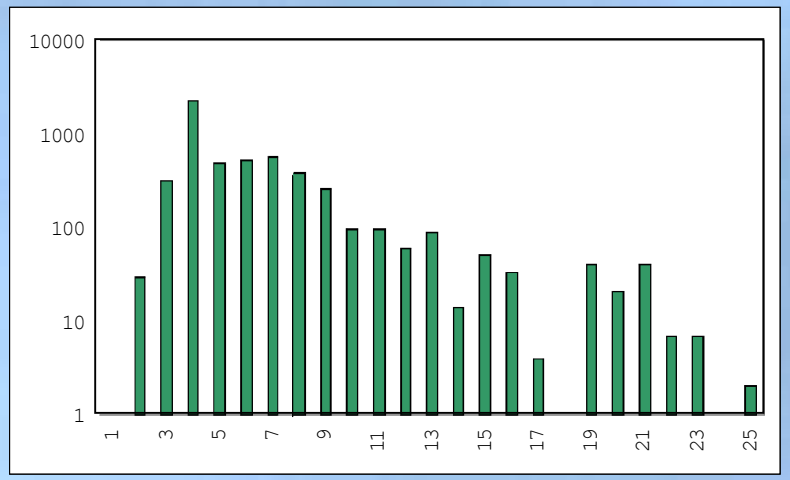
Selected query statistics

A



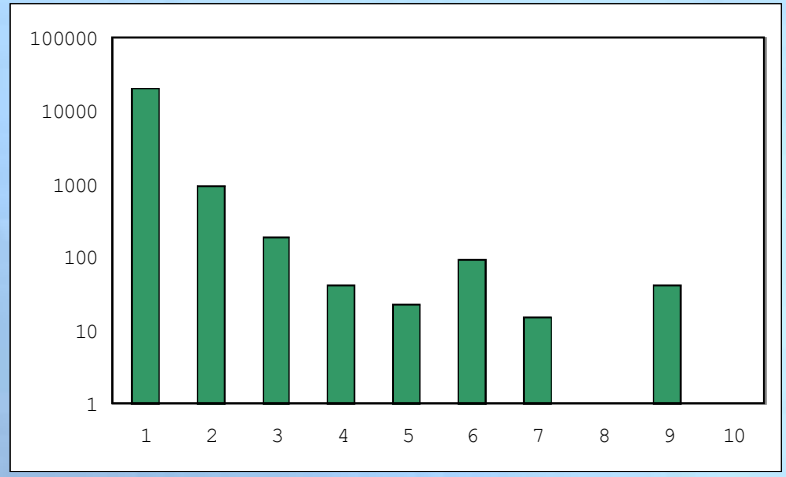
Number of RNA duplexes for different number of base-pairs.

B



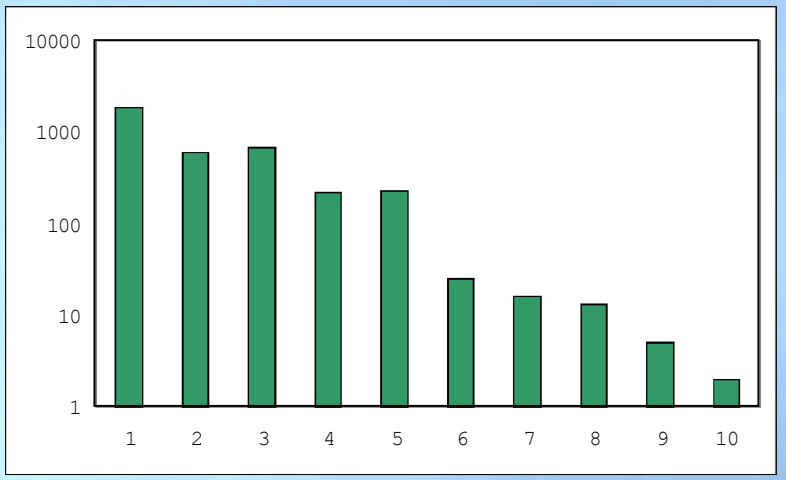
Number of internal loops for different number of unpaired residues.

C



Number of bulges for different number of unpaired residues.

D

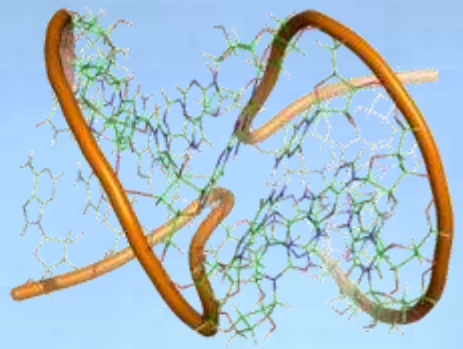


Number of symmetric internal loops for different number of unpaired residues.

Kissing loop search: selected results

```
>Strand1
(((((((((())))
>Strand2
((((]]]]]]]]))))

PDB ID : 1BJ2
```

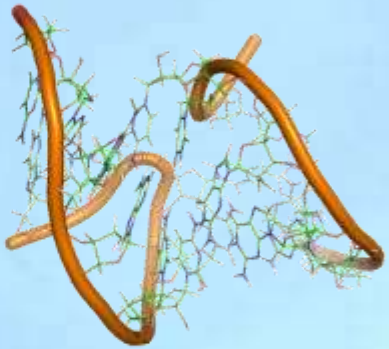


```
>Strand1
(((..[[)))
>Strand2
(((..]])))

PDB ID: 1F5U
```

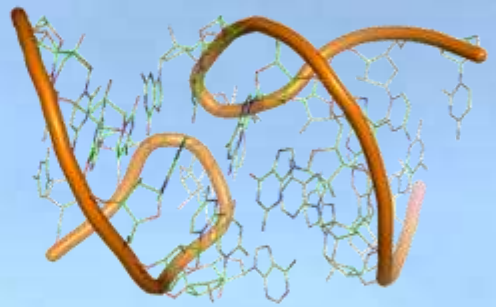
```
>Strand1
(.....[)
>Strand2
(.]......)

PDB ID: 1KIS
```



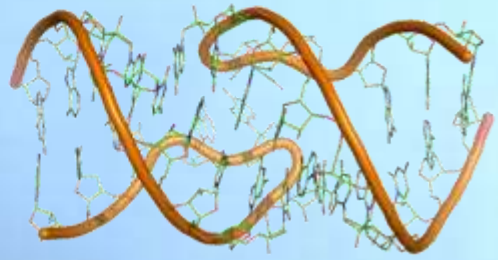
```
>Strand1
((..[[[[))
>Strand2
((..]]]]..))

PDB ID: 1JZX
```



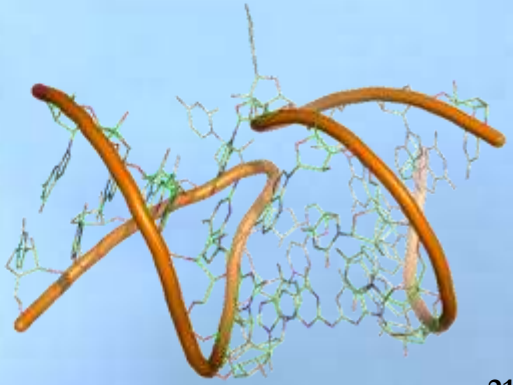
```
>Strand1
(((..[[[[[[.)))
>Strand2
(((..]]]]]].)))

PDB ID: 1ZCI
```

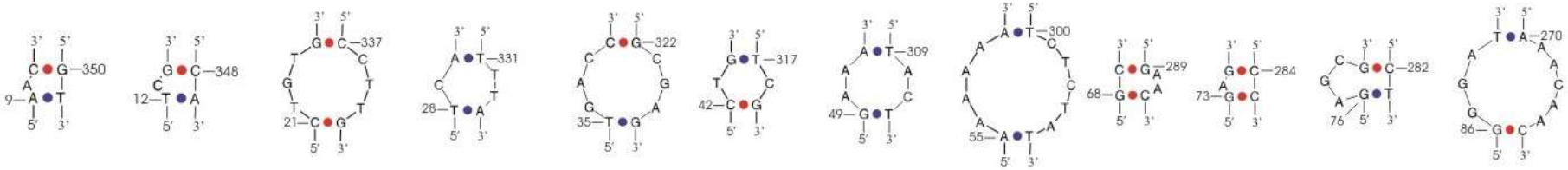


```
>Strand1
(((.[[[[[[)))
>Strand2
(((.]]]]]]]))))

PDB ID: 1VQO
```



Selected search results



I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
A 5'-A C-3' 3'-U-G-5' 8	C 5'-U G-3' 3'-A-C-5' 2	UGU 5'-C G-3' 3'-G C-5' 0 UUC	UU 5'-U A-3' 3'-A U-5' 0 C-	GAC 5'-U C-3' 3'-G G-5' 0 AGC	U 5'-C G-3' 3'-G U-5' 0 C	AA 5'-G A-3' 3'-U U-5' 0 CA	AA-AA 5'-A A-3' 3'-U U-5' 0 AUCUC	5'-G--C-3' 3'-C G-5' AA 0	A 5'-G G-3' 3'-C-C-5' 84	AGC 5'-G G-3' 3'-U---C-5' 0	G-G-A 5'-G U-3' 3'-C A-5' 0 AACAA
A 5'-R Y-3' 3'-Y-R-5' 192	C 5'-Y R-3' 3'-R-Y-5' 56	UGU 5'-Y R-3' 3'-R R-5' 0 UUC	UU 5'-Y R-3' 3'-Y Y-5' 0 C-	GAC 5'-Y Y-3' 3'-R R-5' 0 AGC	U 5'-Y R-3' 3'-R Y-5' 67 C	AA 5'-R R-3' 3'-Y Y-5' 0 CA	AA-AA 5'-R R-3' 3'-Y Y-5' 0 AUCUC	5'-R--Y-3' 3'-Y R-5' AA 1	A 5'-R R-3' 3'-Y-Y-5' 144	AGC 5'-R R-3' 3'-Y---Y-5' 0	G-G-A 5'-R Y-3' 3'-Y R-5' 0 AACAA
A 5'-N N-3' 3'-N-N-5' 848	C 5'-N N-3' 3'-N-N-5' 323	UGU 5'-N N-3' 3'-N N-5' 0 UUC	UU 5'-N N-3' 3'-N N-5' 0 C-	GAC 5'-N N-3' 3'-N N-5' 0 AGC	U 5'-N N-3' 3'-N N-5' 148 C	AA 5'-N N-3' 3'-N N-5' 39 CA	AA-AA 5'-N N-3' 3'-N N-5' 0 AUCUC	5'-N--N-3' 3'-N N-5' AA 150	A 5'-N N-3' 3'-N-N-5' 848	AGC 5'-N N-3' 3'-N---N-5' 0	G-G-A 5'-N N-3' 3'-N N-5' 0 AACAA
R 5'-A C-3' 3'-U-G-5' 8	Y 5'-U G-3' 3'-A-C-5' 4	YRY 5'-C G-3' 3'-G C-5' 0 YYY	YY 5'-U A-3' 3'-A U-5' 0 Y-	RRY 5'-U C-3' 3'-G G-5' 0 RRY	Y 5'-C G-3' 3'-G U-5' 31 Y	RR 5'-G A-3' 3'-U U-5' 0 YR	RR-RR 5'-A A-3' 3'-U U-5' 0 RYYYY	5'-G--C-3' 3'-C G-5' RR 2	R 5'-G G-3' 3'-C-C-5' 84	RRY 5'-G G-3' 3'-U---C-5' 0	R-R-R 5'-G U-3' 3'-C A-5' 0 RRYRR
R 5'-R Y-3' 3'-Y-R-5' 258	Y 5'-Y R-3' 3'-R-Y-5' 82	YRY 5'-Y R-3' 3'-R Y-5' 0 YYY	YY 5'-Y R-3' 3'-R Y-5' 0 Y-	RRY 5'-Y Y-3' 3'-R R-5' 43 RRY	Y 5'-Y R-3' 3'-R Y-5' 140 Y	RR 5'-R R-3' 3'-Y Y-5' 0 YR	RR-RR 5'-R R-3' 3'-Y Y-5' 0 RYYYY	5'-R--Y-3' 3'-Y R-5' RR 44	R 5'-R R-3' 3'-Y-Y-5' 203	RRY 5'-R R-3' 3'-Y---Y-5' 0	R-R-R 5'-R R-3' 3'-R R-5' 0 RRYRR
R 5'-N N-3' 3'-N-N-5' 1206	Y 5'-N N-3' 3'-N-N-5' 686	YRY 5'-N N-3' 3'-N N-5' 0 YYY	YY 5'-N N-3' 3'-N N-5' 39 Y-	RRY 5'-N N-3' 3'-N N-5' 73 RRY	Y 5'-N N-3' 3'-N N-5' 376 Y	RR 5'-N N-3' 3'-N N-5' 73 YR	RR-RR 5'-N N-3' 3'-N N-5' 0 RYYYY	5'-N--N-3' 3'-N N-5' RR 370	R 5'-N N-3' 3'-N-N-5' 1206	RRY 5'-N N-3' 3'-N---N-5' 19	R-R-R 5'-N N-3' 3'-N N-5' 0 RRYRR
N 5'-A C-3' 3'-U-G-5' 8	N 5'-U G-3' 3'-A-C-5' 44	NNN 5'-C G-3' 3'-G C-5' 108 NNN	NN 5'-U A-3' 3'-A U-5' 9 N-	NNN 5'-U C-3' 3'-G G-5' 104 NNN	N 5'-C G-3' 3'-G U-5' 63 N	NN 5'-G A-3' 3'-U U-5' 5 NN	NN-NN 5'-U U-3' 3'-A A-5' 1 NNNNN	5'-G--C-3' 3'-C G-5' NN 14	N 5'-G G-3' 3'-C-C-5' 274	NNN 5'-G G-3' 3'-U---C-5' 0	N-N-N 5'-G U-3' 3'-C A-5' 0 NNNNN
N 5'-R Y-3' 3'-Y-R-5' 357	N 5'-Y R-3' 3'-R-Y-5' 529	NNN 5'-Y R-3' 3'-Y Y-5' 279 NNN	NN 5'-Y R-3' 3'-R Y-5' 157 N-	NNN 5'-Y Y-3' 3'-R R-5' 350 NNN	N 5'-Y R-3' 3'-R Y-5' 579 N	NN 5'-R R-3' 3'-Y Y-5' 310 NN	NN-NN 5'-Y Y-3' 3'-R R-5' 34 NNNNN	5'-R--Y-3' 3'-Y R-5' NN 265	N 5'-R R-3' 3'-Y-Y-5' 464	NNN 5'-R R-3' 3'-Y---Y-5' 98	N-N-N 5'-R Y-3' 3'-Y R-5' 12 NNNNN
N 5'-N N-3' 3'-N-N-5' 1883	N 5'-N N-3' 3'-N-N-5' 1883	NNN 5'-N N-3' 3'-N N-5' 667 NNN	NN 5'-N N-3' 3'-N N-5' 753 N-	NNN 5'-N N-3' 3'-N N-5' 667 NNN	N 5'-N N-3' 3'-N N-5' 1819 N	NN 5'-N N-3' 3'-N N-5' 597 NN	NN-NN 5'-N N-3' 3'-N N-5' 517 NNNNN	5'-N--N-3' 3'-N N-5' NN 947	N 5'-N N-3' 3'-N-N-5' 1883	NNN 5'-N N-3' 3'-N---N-5' 188	N-N-N 5'-N N-3' 3'-N N-5' 111 NNNNN

References

M. Popenda, M. Błażewicz, M. Szachniuk, R.W. Adamiak (2007) RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures, *Nucleic Acids Research*, doi: 10.1093/nar/gkm786.

M. Błażewicz, RNA fragments search engine – project and implementation, *M.Sc. thesis* to be completed in 2008

The database is freely accessible at:

http://rnafrabase.ibch.poznan.pl