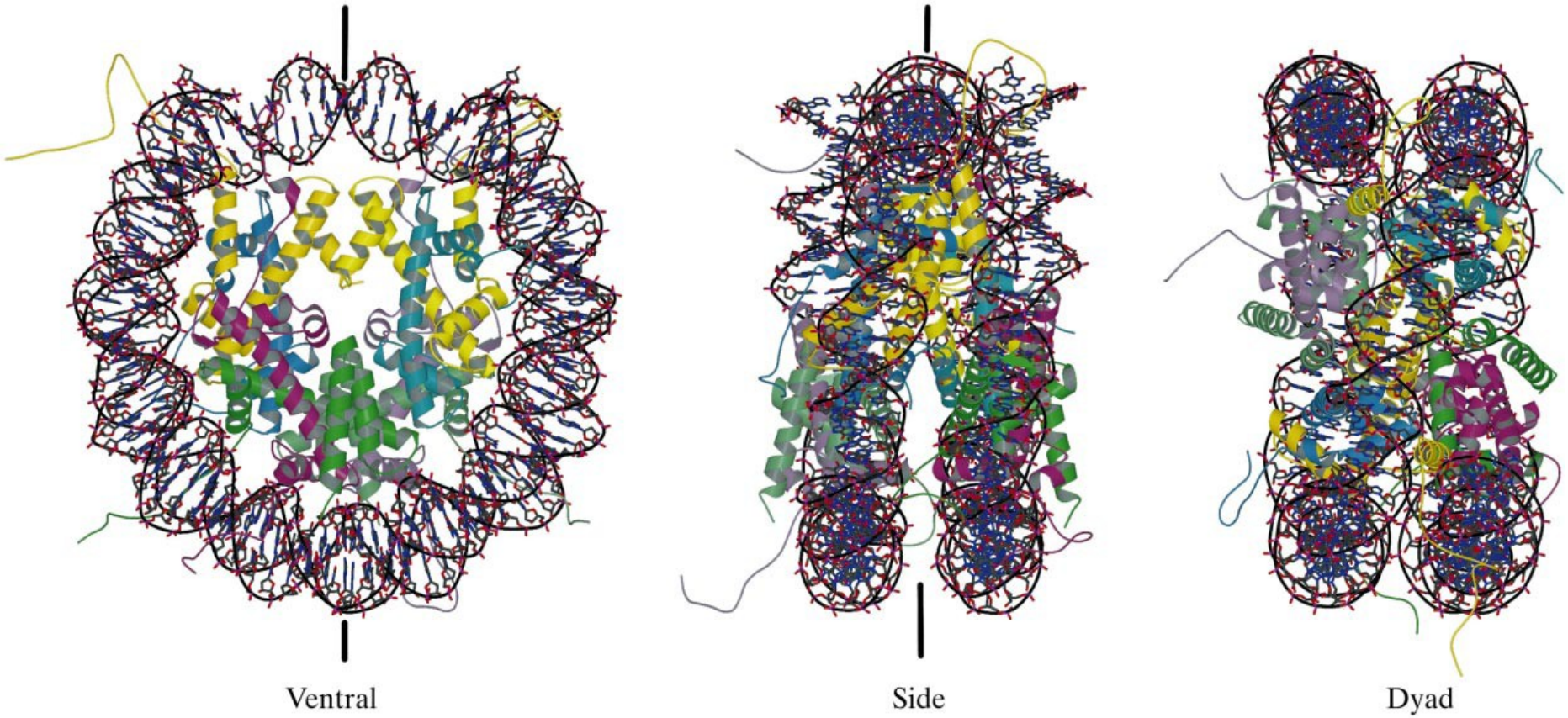
Nucleosome positioning sequence code: 33 years of agony and final picture

Edward N. Trifonov
University of Haifa, Israel

Prague, Brno
2013



Lab of G. Bunick, 2000

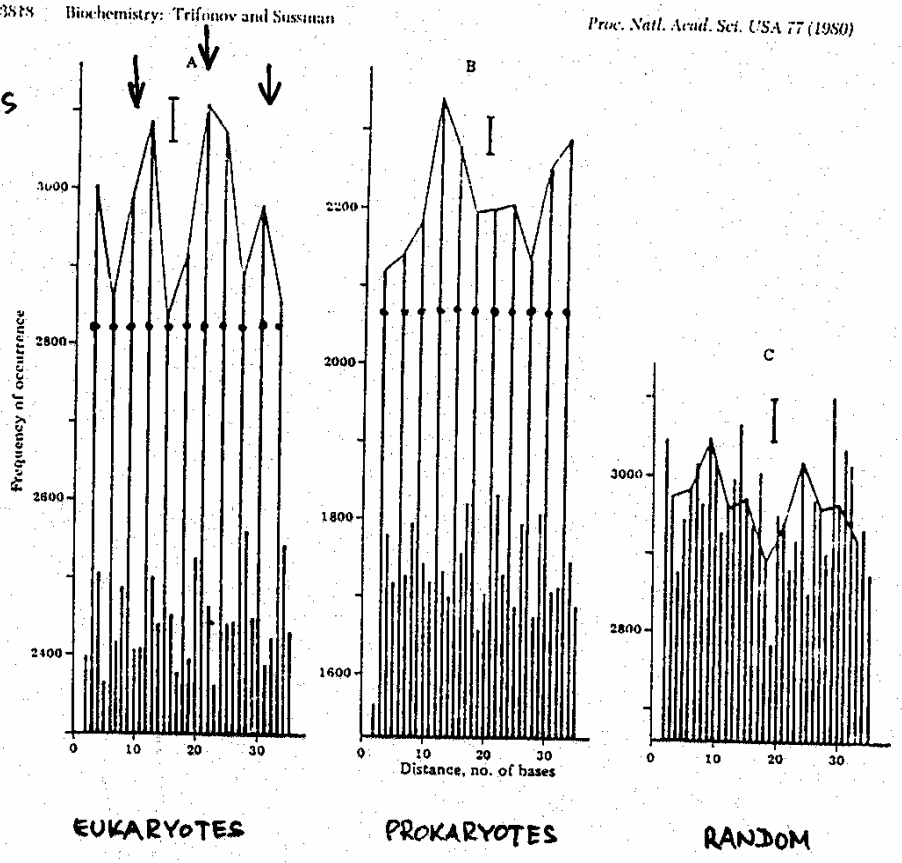
DNA in the nucleosome is severely deformed.

Neighboring base pairs become partially unstacked.

Some of the dinucleotide stacks
may be more deformable than others.
This also depends on their rotational orientations.

TRIFONOV, SUSSMAN, 1980

DISTANCE ANALYSIS (Autocorrelation)

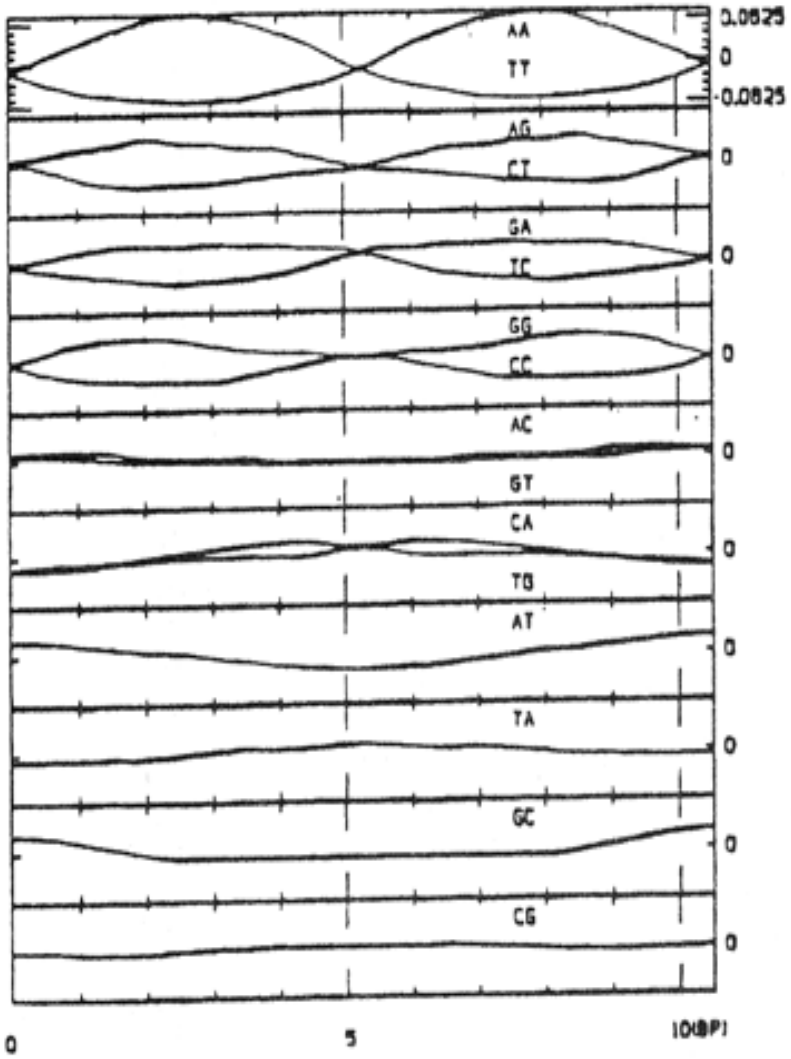


~ 10.5 BASES

3 BASES

~ 30 000 BASES

5’...RRRYYYYYYRRRRRRYYY...



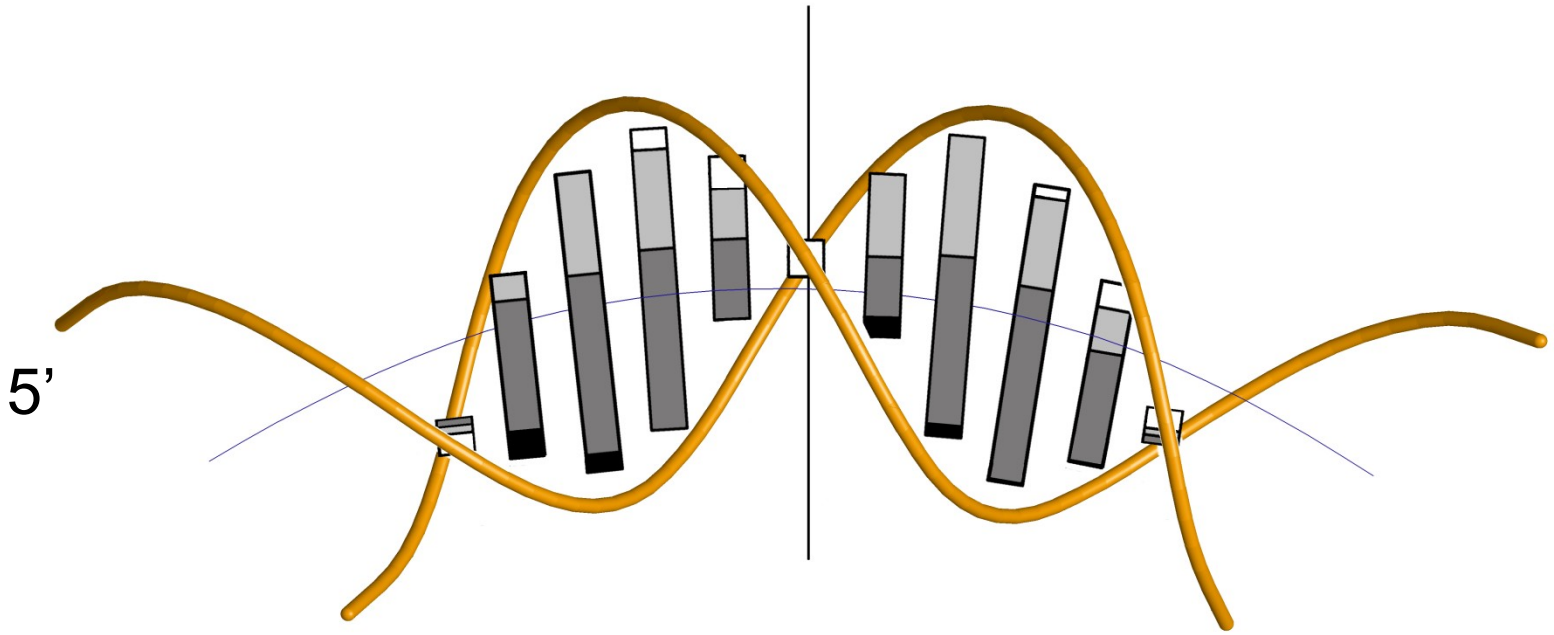
First matrix of nucleosome DNA bendability

Mengeritsky and ENT, 1983

Pattern of **1980-1983**

yrRRRryYYYyr
xxAAAxxTTTxx

Trifonov, Sussman , 1980
Trifonov, 1980
Mengeritsky, Trifonov, 1983



5’…YYYRRRRRYYYYYYRRR…

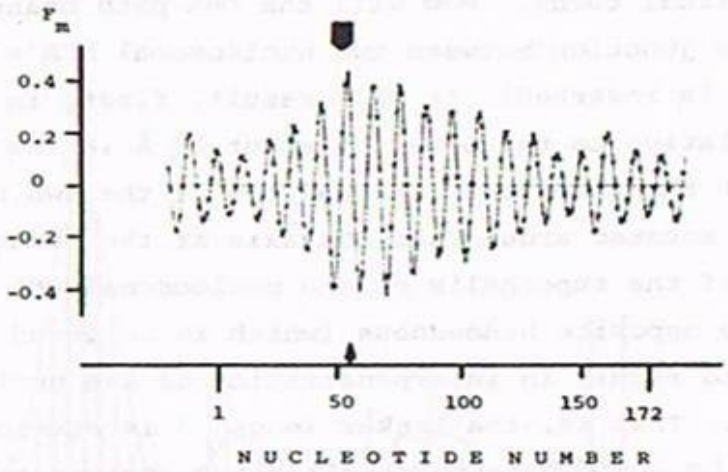


Fig. 5: The mapping function calculated for the nucleotide sequence of green monkey α-satellite. The numbering of the nucleotides is the same as used by Rosenberg *et al.*[13] The small arrow indicates position of the major maximum of the mapping function. The bigger arrow on the top points to the middle of the nucleosome found experimentally.[8,9] The width of the arrow corresponds to the error of the experimental mapping.

This achievement in the single-base accuracy mapping of the nucleosomes has not been accepted by chromatin research community.

The reasons:

1. Mistrust. The physics of the phenomenon and multiple alternative positions of the nucleosome centers are hard to grasp for non-physicists, and the sequences did not show any obvious periodicity

2. The chromatin research community was not ready yet to conduct high resolution experimental studies

History of the chromatin code. Pre-genomic studies 1980-2006

~10.5 base periodicity of some dinucleotides Trifonov, Sussman (1980)

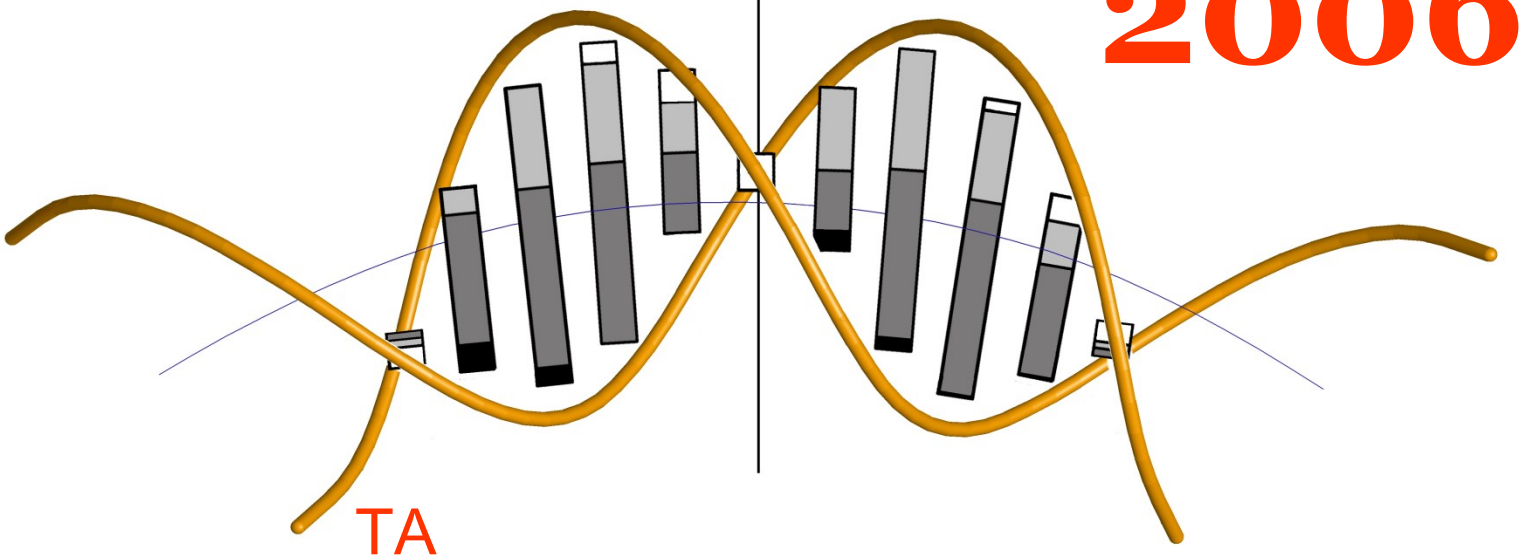
...T T A A A A A T T T T T A A A A A T T...	Mengeritsky, Trifonov (1983)
...Y Y R R R R R Y Y Y Y Y R R R R R Y Y...	Mengeritsky, Trifonov (1983)
...x Y R x x x R Y x x x Y R x x x R Y x...	Zhurkin (1983)
...W W W W x S S S S x W W W W x S S S S...	Satchwell *et al*. (1986)
...x W W W x x S S S x x W W W x x S S S...	Shrader, Crothers(1989), Tanaka et al.,(1992)
...C C x x x x x C C C C C x x x x x C C...	Bolshoy (1995)
...V W G x x x x x x x V W G x x x x x x...	Baldi *et al*. (1996)
...x x G G R x x x x x x x G G R x x x x...	Travers, Muyldermans (1996)
...C T A T A A A C G C C T A T A A A C G...	Widlund *et al*. (1997)
...C T A G x x x x x x C T A G x x x x x...	Lowary, Widom (1998)
...S S A A A A A S S S S S A A A A A S S...	Fitzgerald, Anderson (1998)
...C C G G G G G C C C C C G G G G G C C...	Kogan *et al*. (2006)

GG
CC
GC
CG

Suggestion of an approximate pattern by Segal,…,Widom, Nature 442, 772

2006

5’



TA
AA
TT

The work of Segal et al., 2006, was
the first high throughput whole-genome analysis.

It drew a lot of attention, and the approach
became very fashionable in the chromatin community.

But the emphasis was still on low resolution studies,
maps of “occupancy”, where the alternative positions
of the nucleosomes and rotational setting of DNA
are not seen.

No attempts were made to derive an exact
nucleosome positioning sequence pattern
from the whole genome sequences.

When we joined the high througput efforts our primary task was to derive the detailed nucleosome positioning sequence pattern

This involved three original techniques

A. Signal regeneration from its parts
B. Shannon N-gram extension
C. Extraction and analysis of strong nucleosomes

Nucleosome positioning patterns, species:

	species	authors	method
C GRAAA TTTYC G	*C. elegans*	Gabdank, 2009	A
C AAAAA TTTTT G	*C. elegans*	Rapoport, 2011	B
C AAAAA TTTTT G	*A. gambiae*	same	B
C AAAAA TTTTT G	*C. albicans*	same	B
C AAAAA TTTTT G	*D. melanogaster*	same	B
C AAAAA TTTTT G	*S. cerevisiae*	same	B
T AAAAA TTTTT A	*A. mellifera*	same	B
T AAAAA TTTTT A	*A. thaliana*	same	B
T AAAAA TTTTT A	*D. discoideum*	same	B
T AAAAA TTTTT A	*D. rerio*	same	B
T AAAAA TTTTT A	*G. gallus*	same	B
T AAAAA TTTTT A	*H. sapiens*	same	B
T AAAAA TTTTT A	*M. musculus*	same	B
c GGGGG CCccc G	*C. reinhardtii*	same	B
Y RRRRR YYYYY R	consensus		

A – signal regeneration, nucleosomes
B – Shannon N-gram extension, whole genome

Structural and sequence periodicity of nucleosome DNA

DNase I digestion of chromatin	10.30-10.40 bp	Prunell, Kornberg, Lutter, Klug, Levitt, Crick, **1979**
Beat effect, DNase I	10.33-10.40 bp	Bettecken, **1979**
Analytical geometry of nucl. DNA	10.30-10.50 bp	Ulanovsky, **1983**
DNA path in nucleosome crystals	10.36-10.44 bp	Cohanim, **2006**
CG periodicity, honey bee	10.36-10.44 bp	Bettecken, **2009**
DNase I digestion of chromatin	10.36-10.44 bp	Duke University, **2013**

Common range 10.36-10.40 bp

Magic distances, 10.4•n bases

	nearest integers
10.4	10
20.8	21
31.2	31
41.6	42
52.0	52
62.4	62
72.8	73
83.2	83
93.6	94
104.0	104
114.4	114

The ideal nucleosome positioning sequence
would contain some periodically repeating motif,
and all the distances between the same dinucleotides
would be magic distances.
Strong nucleosome DNA would show many magic distances.

Lowary and Widom (1998) took large ensemble of synthetic DNA fragments with random sequences, and selected those of them which formed

strong nucleosomes

The sequences demonstrated very strong periodicity of TA dinucleotides

Clone 601,

from collection of Lowary and Widom (1998)

...CAGCGCG**TA**CGTGCGTT**TA**AGCGGTGC**TA**GAGCTGTC**TA**C...

TACGTGCGTT**TA**
TAAGCGGTGC**TA**
TAGAGCTGTC**TA**

We took all **TA**nnnnnnnn**TA** segments
from the collection of Lowary/Widom,
and analysed which dinucleotides
are most frequently located in the
interval between **TA**, and in which positions

Regeneration of signal from its incomplete versions:

AA

positional autocorrelation

AAnnnnnnnnAA

regeneration
(all occurrences of AAnnnnnnnnAA are aligned, and other dinucleotides counted within the period)

AAnnnnCCnnAA

Gabdank, 2009

Bendability matrix for strong nucleosome DNAs of Lowary and Widom collection

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**0**
AA	0	16	3	0	0	1	0	0	0	0	0
AC	0	5	2	5	2	3	5	3	1	0	0
AG	0	**25**	11	**9**	2	4	1	1	1	0	0
AT	0	2	0	3	1	1	3	1	2	0	0
CA	0	0	1	0	2	4	3	1	0	0	0
CC	0	0	0	0	5	4	7	3	6	0	0
CG	0	0	4	4	4	4	4	5	3	0	0
CT	0	0	0	2	1	2	1	9	11	**22**	0
GA	0	0	**12**	4	3	3	0	0	0	0	0
GC	0	0	4	7	6	7	5	10	5	0	0
GG	0	0	7	4	3	3	7	0	1	0	0
GT	0	0	2	7	6	4	5	6	2	6	0
TA	**48**	0	1	1	4	1	2	3	0	0	**48**
TC	0	0	0	0	1	1	1	4	10	0	0
TG	0	0	0	1	8	6	4	2	1	0	0
TT	0	0	1	1	0	0	0	0	5	20	0

22.5 min

T A G A G x x x x **C T A** – manually

T A G A G G C C T C T A – by dynamic programming

Y R R R R R Y Y Y Y Y R

T A G A G G C C T C T A

The periodical pattern hidden in the sequences of Lowary and Widom is selfcomplementary, and manifests alternation of RRRRR and YYYYY

The strongest nucleosomes of *A. thaliana* display very clear though still imperfect periodicity

The ideal pattern for *A.thaliana* is repetition of TAAAAATTTTTA, again, alternation of RRRRR and YYYYY, and complementary symmetry

Before this picture was generated

(Dec. last year) nobody ever had seen

that the nucleosome sequences

look, indeed, periodical

From the bendability matrices

for the strong nucleosomes:

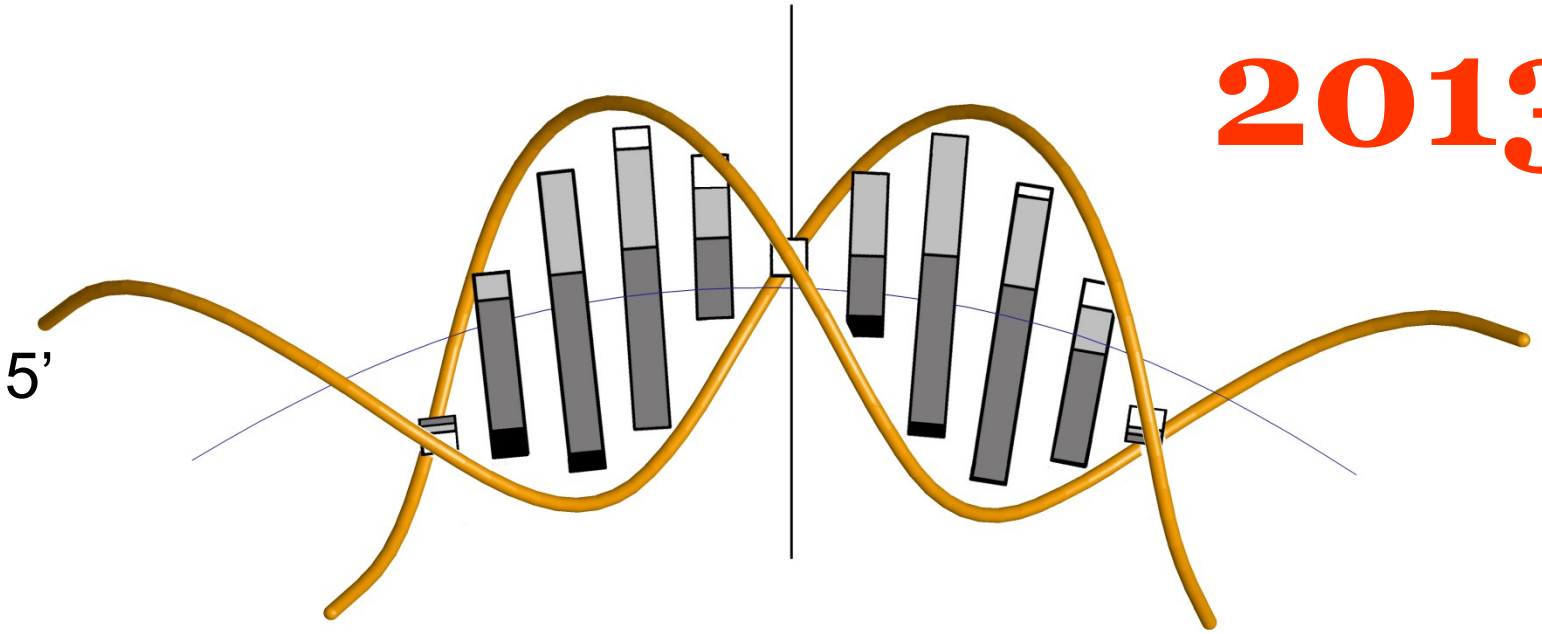
T AGAGG CCTCT A	Lowary and Widom
T AAAAA TTTTT A	A.thaliana
T AAAAA TTTTT A	C.elegans
T AAAAA TTTTT A	H.sapiens
T AAAAA TTTTT A	isochores L1, L2, H1 and H2
C GGGGG CCCCC G	isochores H3
Y RRRRR YYYYY R	common for all

A. thaliana	T AAAAA TTTTT A	strong nucleosomes
	T AAAAA TTTTT A	Shannon extension
C. elegans	T AAAAA TTTTT A	strong nucleosomes
	c grAAA TTT**yc g**	signal regeneration
isochores L1, L2	T AAAAA TTTTT A	strong nucleosomes
	T AAAAA TTTTT A	Shannon extension
isochores H1	T AAAAA TTTTT A	strong nucleosomes
	c gAAA TTT**c**T **g**	Shannon extension
isochores H2	T AAAAA TTTTT A	strong nucleosomes
	c ggggA T**cccc g**	Shannon extension
isochores H3	C GGGGG CCCCC G	strong nucleosomes
	C **a**GGGG CCCC**t** G	Shannon extension

Y RRRRR YYYYY R - all,
and all with complementary symmetry

Nucleosome positioning pattern

2013



5’…YYYRRRRRYYYYYRRR…

TA	**AT**
CG	**GC**
TG	AC
CA	GT
Contact with arginines	Exposed

The rest of the period is occupied by RR (AA,AG,GA,GG) and YY (TT, TC, CT, CC) dinucleotides, in their optimal partial unstacking positions

The dinucleotide stacks are placed in such positions within the nucleosome DNA period to ensure best possible bending.

The better the bending – the stronger the nucleosome.

But the bulk of the nucleosomes are only marginally stable.

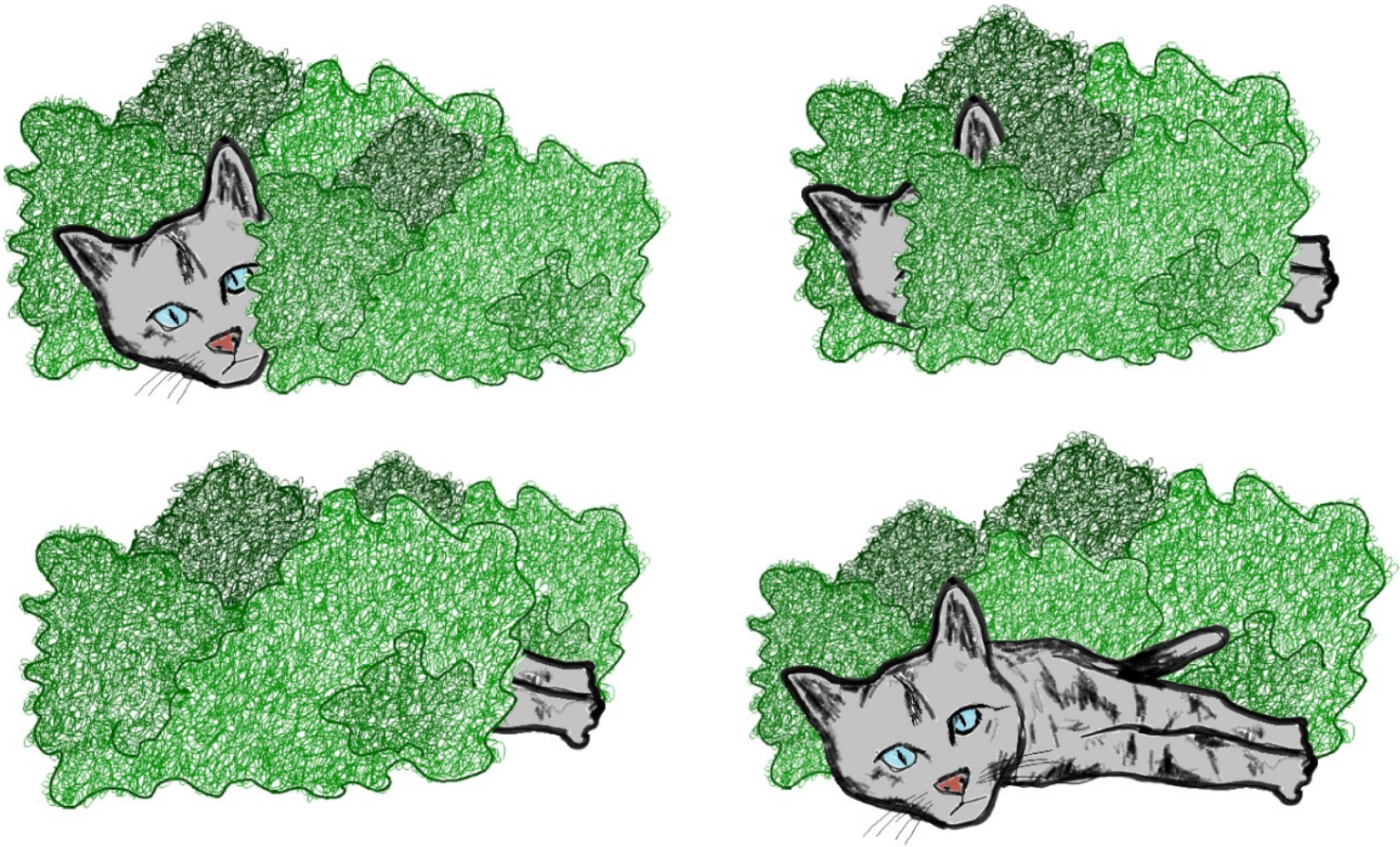
Only a fraction of properly positioned dinucleotides
is present in any given nucleosome DNA sequence.

Match of the BamHI nucleosome (typical semistable nucleosome) to the standard nucleosome probe $(GAAAATTTTC)_n$

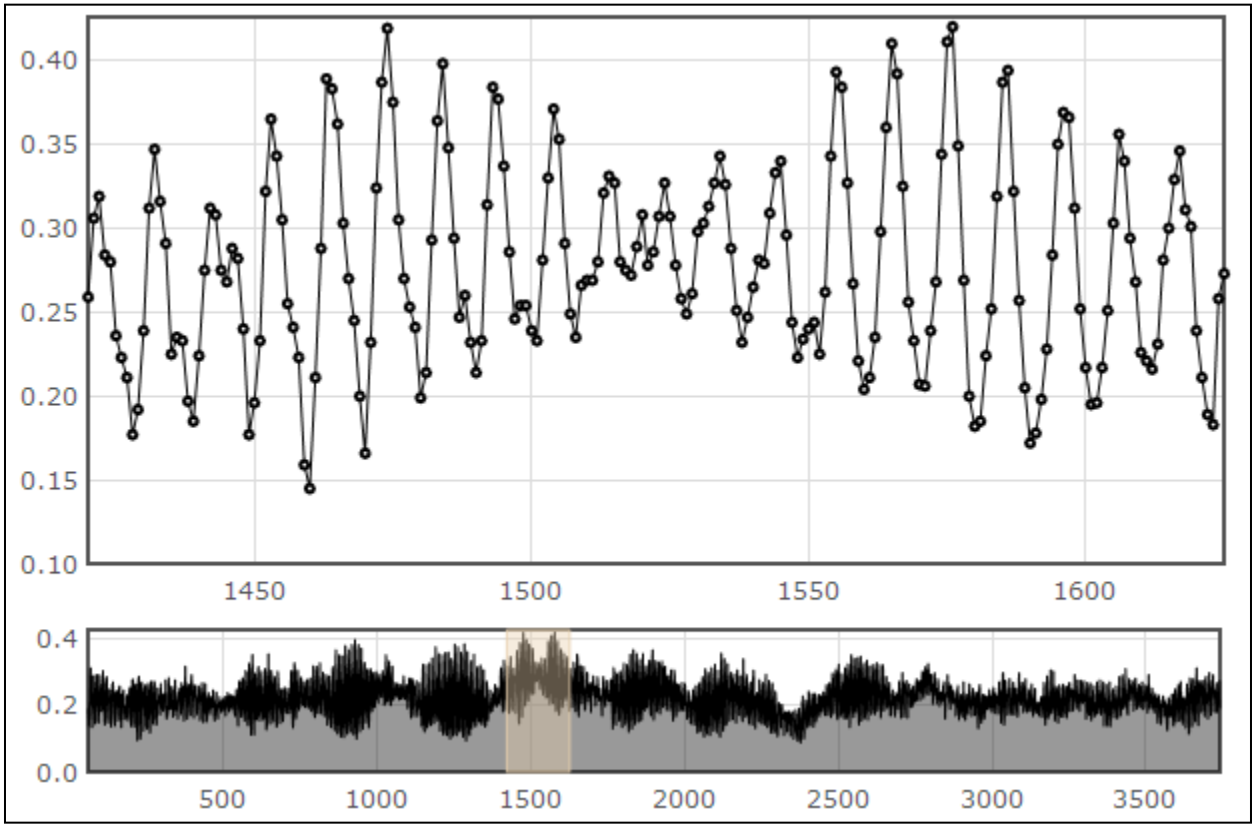
CGGAAATTTTCCGGAAATTTCCGGAAATTTCCGGGAAATTTCCGGAAATTTCCGGAAATTTTCCGGAAATTTCCGGAAATTTCCGGGAAATTTCCGGAAATTTTCC
Cagaggagcttcctgggga**TCC**a**GA**c**AT**gataagataca**TT**gat**GA**gt**TT**ggaca**AA**ccacaactag**AAT**gcagt**GAAA**aaaatgcttt**ATTT**gtga**AA**t**TT**gtgatgcta**TT**gct
YRRRRRag**YYYY**ct**RRR**ga**YYYRRR**c**RY**gata**RRR**taca**YY**gat**RRR**t**YY**ggac**RRR**ccacaact**RRRRY**gcagt**RRRR**aaaa**YR**ctttt**RYYY**gt**RRRR**t**YY**gtgatgcta**YY**g**YY**

The strongest nucleosomes of *A. thaliana* display very clear though still imperfect periodicity

The ideal pattern for *A. thaliana* is repetition of TAAAAATTTTTA, again, alternation of RRRRR and YYYYY, and complementary symmetry

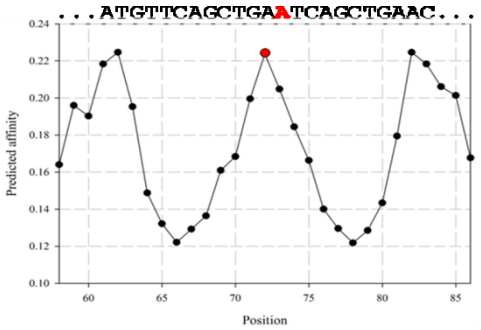


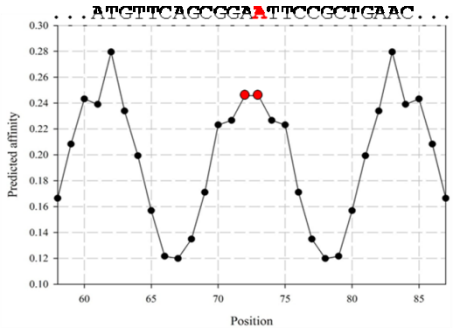
Cat in bushes. Courtesy of I. Gabdank

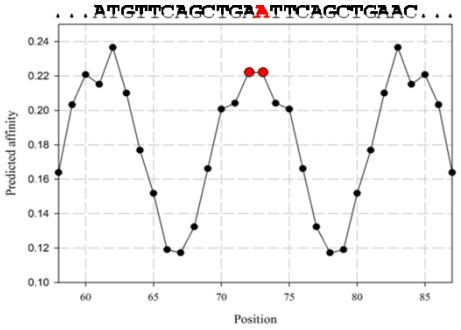


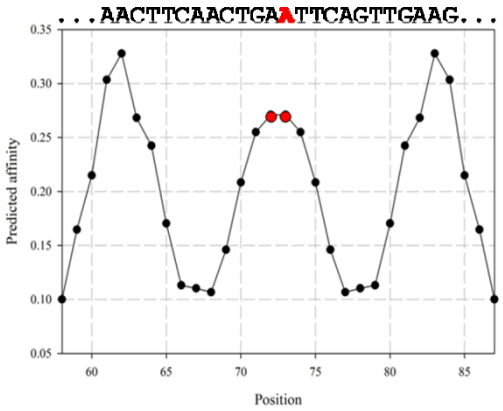
Example of the output from the nucleosome mapping server
http://www.cs.bgu.ac.il/~nucleom

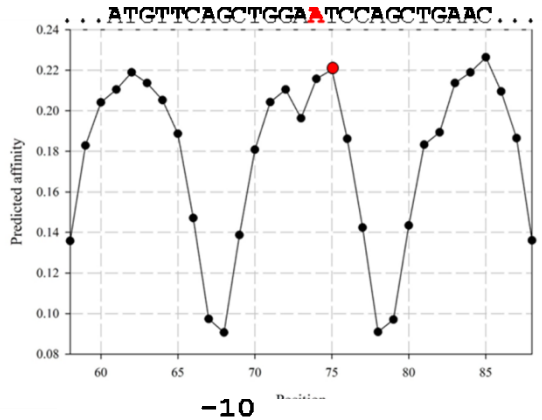
Mapping of sharply positioned nucleosomes

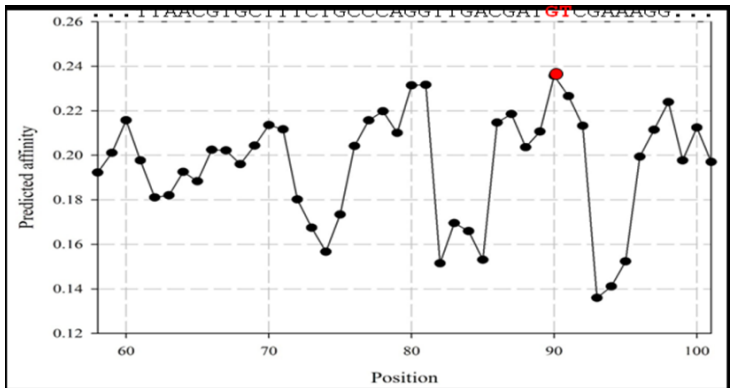


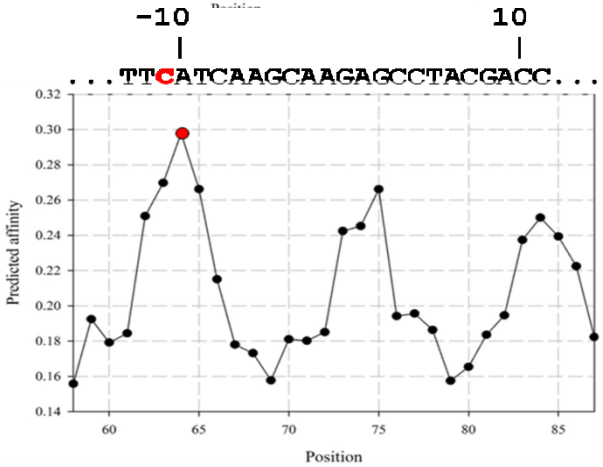


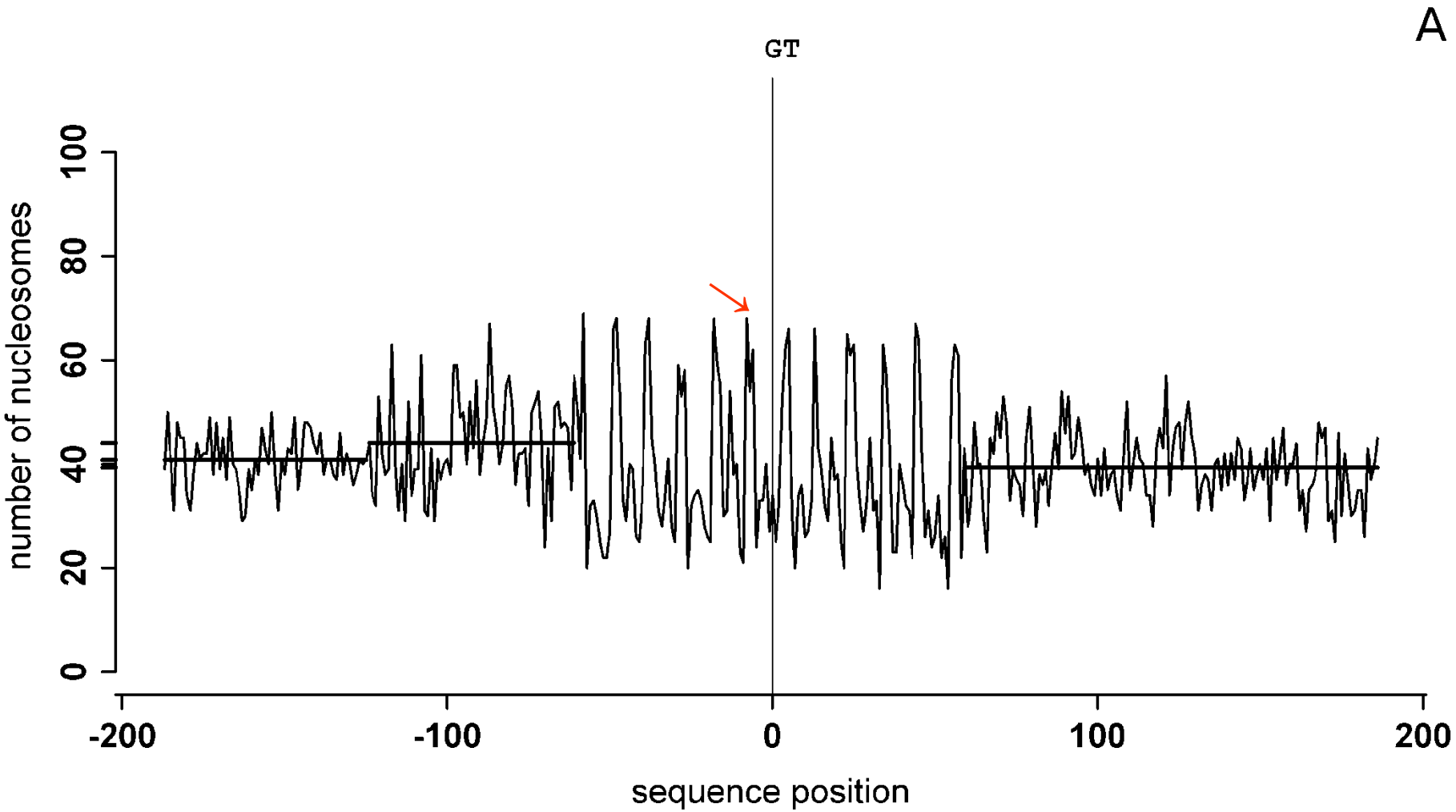




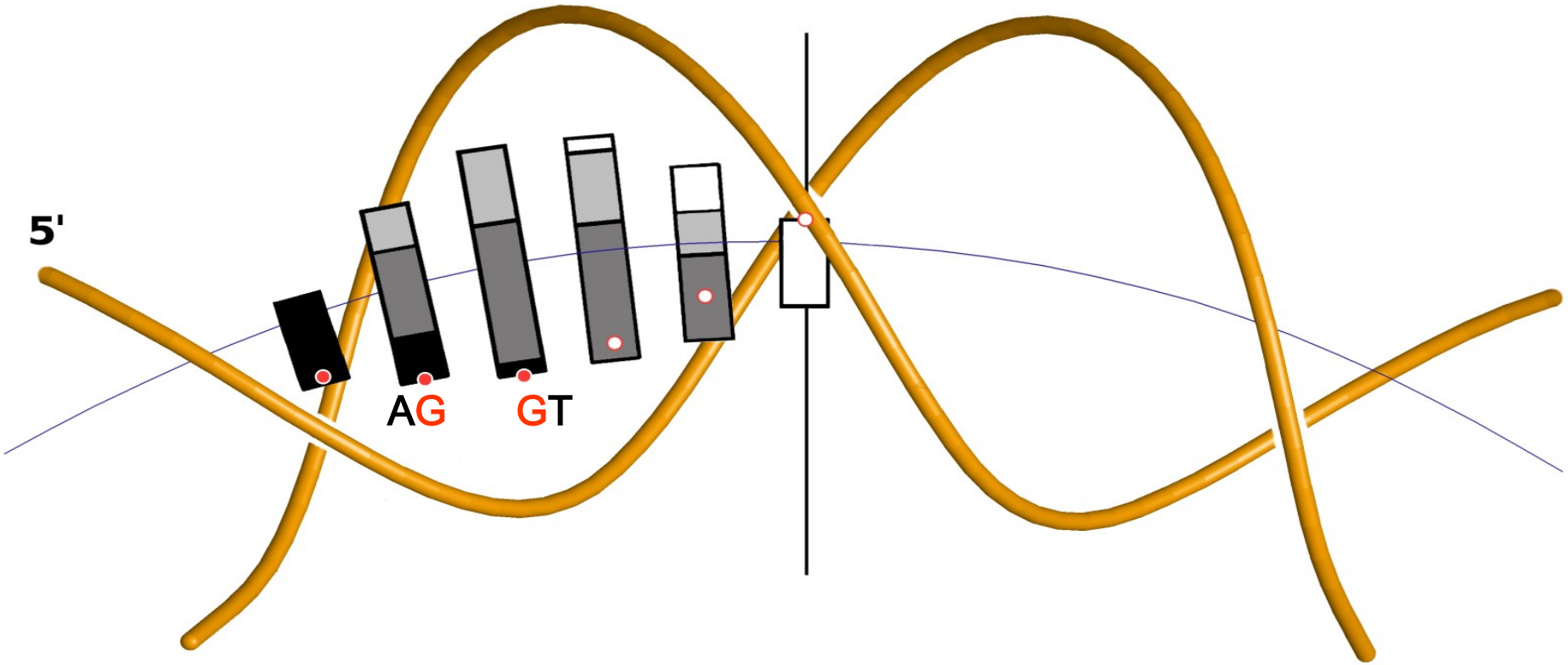








Nucleosomes around the GT splice junctions



Dots • - N9 atoms of guanines

Guanines of GT- and AG-ends of introns are oriented towards the surface of the histone octamer, away from exterior.

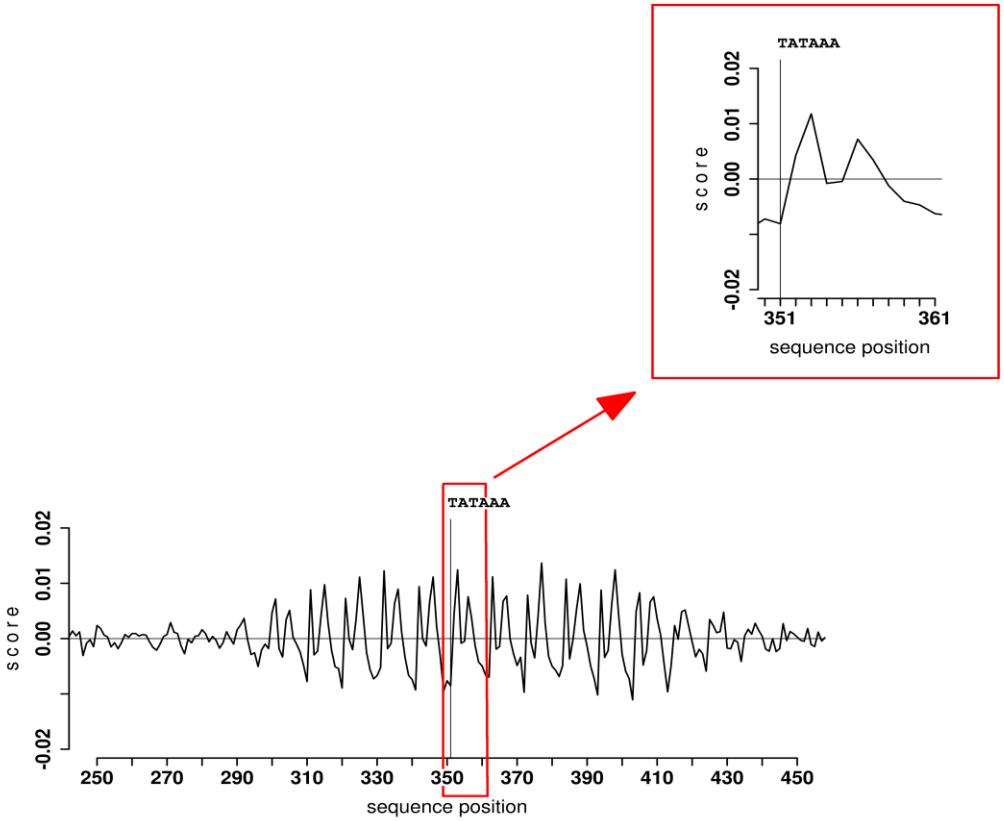
Such orientation is the best for guanines to minimize spontaneous depurination and oxidation

The most frequent spontaneous damages to DNA bases:

depurination of G (N9 atoms)

oxidation of G

deamination of C



Nucleosome DNA which carries promoter TATAAA box has two rotational settings encoded in the sequence (two peaks within one period)

TATA-switch

Two alternative positions of TATAAA box
in the promoter nucleosomes
are separated by 140 (220) degrees,
which closely correspond to exposed and inaccessible
orientations of the box.

By shifting the DNA along its path by 4(6) bases,
the promoter is switched ON or OFF.

The switch (shift) may be triggered by remodelers or transcription factors.

Today the single-base resolution nucleosome mapping
is the only practical tool to study fine structure
of chromatin and its role in

factor binding,
transcription,
replication,
DNA repair,
transposition,
recombination,
apoptosis,
chromatin domains,
and more

Immediate questions:

Where in genomes the strong nucleosomes are located?

What they are doing there?

Tentative answer:

Strong nucleosomes are chromatin organizers.

ACKNOWLEDGEMENTS

Recent contributions (2009-2013):

Idan **Gabdank** (Beer Sheva, Israel)
Zakharia **Frenkel** (Haifa, Israel)
Alexandra **Rapoport** (Haifa, Israel)
Thomas **Bettecken** (München, Germany)
Jan **Hapala** (Brno, Czech Republic)
Bilal **Salih** (Haifa, Israel)
Vijay **Tripathi** (Haifa, Israel)

Earlier contributions (1980-2008)

Thomas **Bettecken**
Joel **Sussman**
Galina **Mengeritsky**
Levy **Ulanovsky**
Alex **Bolshoy**
Ilya **Ioshikhes**
Amir **Cohanim**
Fadil **Salih**
Simon **Kogan**

Funding (2009-2012)
Israel Science Foundation,
and South Moravian Program

Why DNA binds to histone octamers
by one side?

It could be either intrinsic DNA curvature

or better bending in one specific direction
(deformational anisotropy of DNA)

Both should be sequence-dependent

The purine-purine • pyrimidyne-pyrimidyne stacks (RR • YY) are very asymmetric

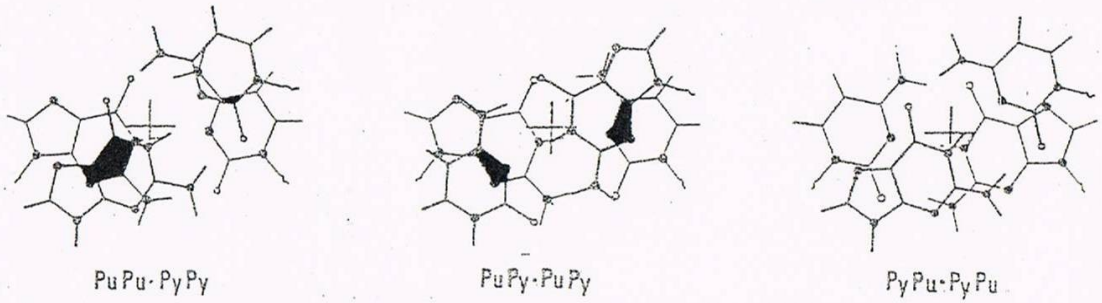


FIGURE 1. Projected views of two successive base-pairs of B DNA. Three possible cases of purine and pyrimidine base overlap are shown. Helix axes (perpendicular to the base-pairs) are indicated by crosses. Overlapping of the heterocyclic rings is shown in black. (From Arnott, S., Dover, S. D., and Wonacott, A. J., *Acta Crystallogr.*, B25, 2192, 1969. With permission.)

E.T.
CRC Crit. Rev. Bioch.
v. 19, 1985

Nucleosome positioning sequence pattern is very weak

(as the nucleosomes should be easy to unfold)

The weak pattern overlaps with other messages ("noise").

That makes the signal/noise ratio very low.

VERY large

database of the nucleosome DNA sequences is needed,

to extract and fully describe the signal

It is easy, however, to detect the signal

DISTANCE ANALYSIS (Autocorrelation)

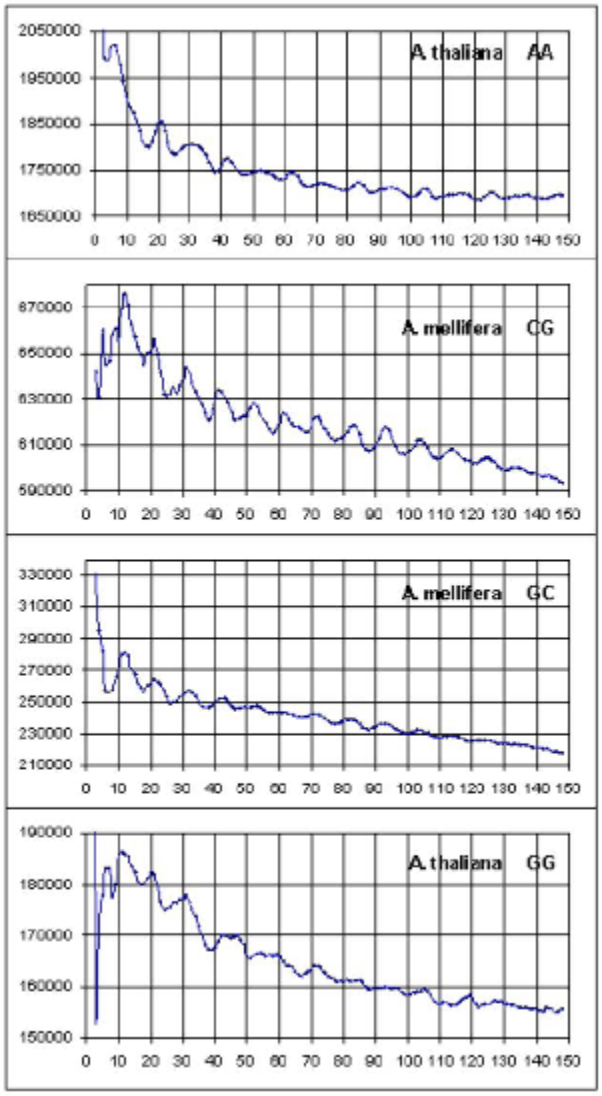


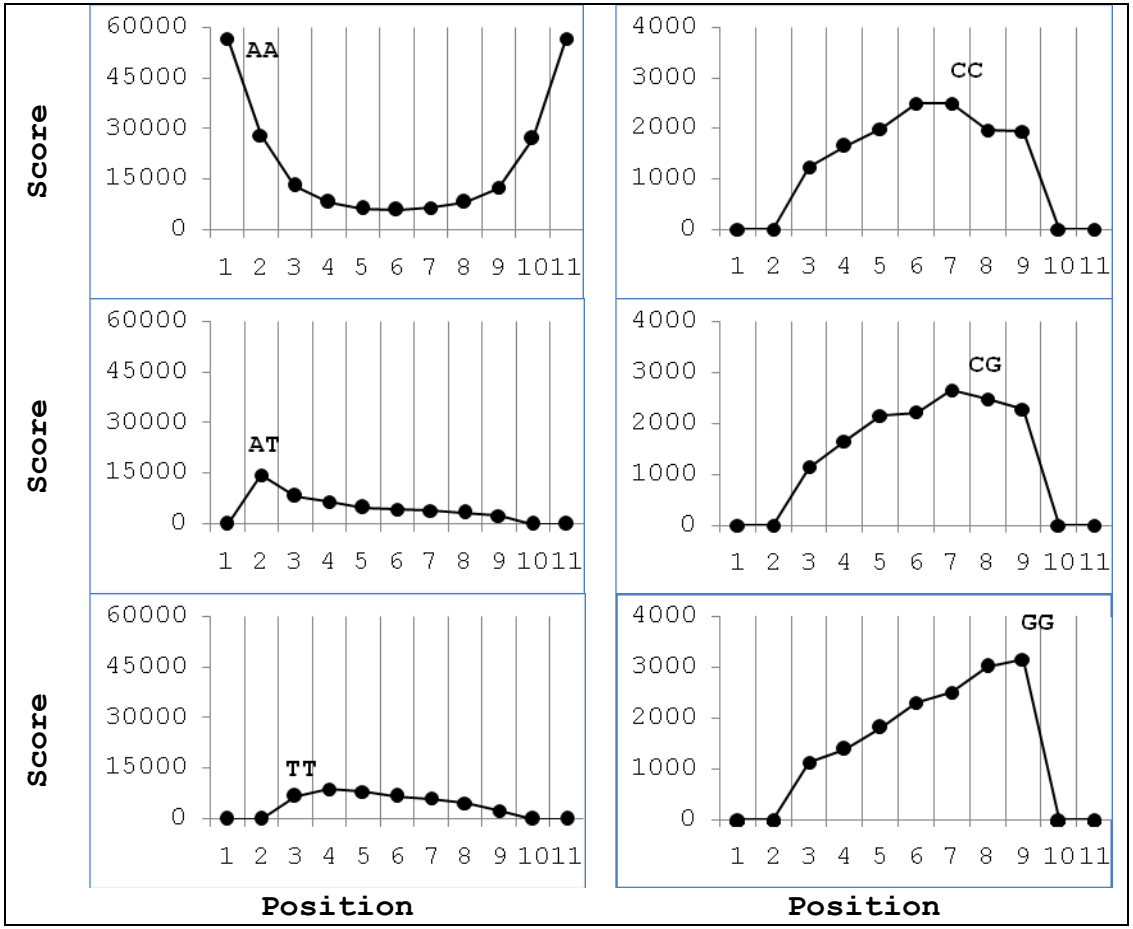
Figure 1

Whole-genome periodicities (distance analysis)

	AA	TT	CG	GC	CA	TG	AG	CT	AT	GG	CC	GA	TC	AC	GT	TA
S. cerevisiae														-	-	
C. elegans										-	-					-
A. thaliana			-				-	-			-	-	-	-	-	-
D. rerio			-		-	-	-	-	-			-	-	-	-	-
C. albicans			-	-			-	-	-	-	-	-	-	-	-	-
A. mellifera					-	-	-	-	-	-	-	-	-	-	-	-
D. melanogaster					-	-	-	-	-	-	-	-	-	-	-	-
G. gallus	-	-	-	-	-	-			-	-	-	-	-	-	-	-
A. gambiae			-	-	-	-	-	-	-	-	-	-	-	-	-	-
C. reinhardtii			-	-	-	-	-	-	-	-	-	-	-	-	-	-
D. discoideum	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-
H. sapiens	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-
M. musculus	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

T.Bettecken, E.N.T., 2009

AAnnnnnnnnAA repeat structure (*C. elegans*)



Regenerated pattern (AAATTTCCGG)(AAAT…

Positional matrix of bendability(C.elegans)

1	2	3	4	5	6	7	8	9	0	1	2
C	G									C	G
	G	G									
	G	A									
		G	A								
		A	A								
			A	A	A						
					A	T					
						T	T	T			
								T	T		
								T	C		
									T	C	
									C	C	
										C	G

LINEAR FORM OF THE POSITIONAL MATRIX OF BENDABILITY (*C.elegans*):

CGRAAATTTYCG

(YRRRRRYYYYYR)

Trinucleotides of *C. elegans* genome

		counts
1	**AAA**	4162266
2	**TTT**	4160750
3	**ATT**	2488998
4	**AAT**	2486813
5	**GAA**	1873844
6	**TTC**	1871673
7	CAA	1667120
8	TTG	1663842
9	TCA	1498069
10	TGA	1496493
.....		.....

TOPMOST TRINUCLEOTIDES MAKE TOGETHER THE DOMINANT PATTERN

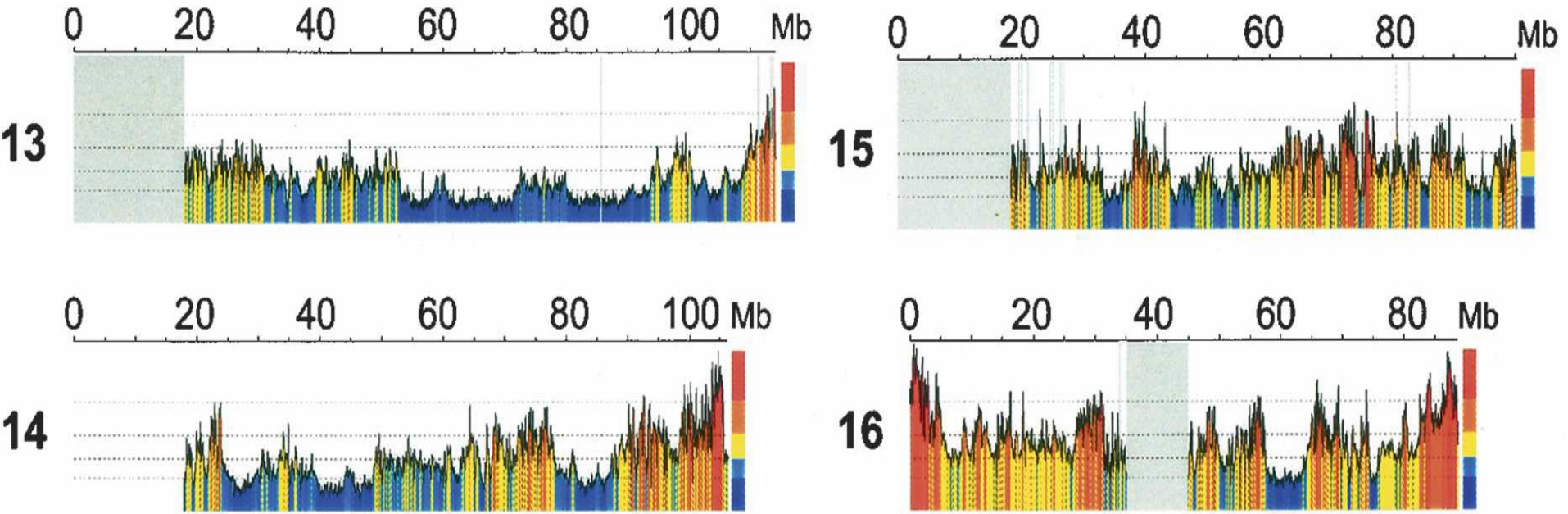
GAAAATTTTC:

GAAAATTTTC
GAAAATTTTC
GAAAATTTTC
GAAAATTTTC
GAAAATTTTC
GAAAATTTTC
GAAAATTTTC
GAAAATTTTC

This technique is known since 1948 –

Shannon N-gram extension

It has been very helpful
in further studies
of the nucleosome positioning patterns



Human isochores

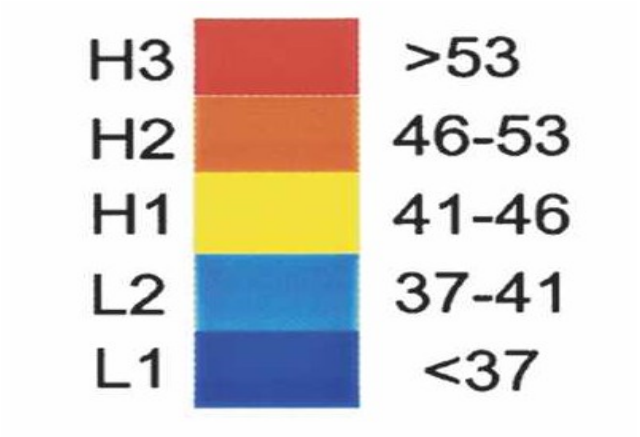
Lab of G. Bernardi, 2006

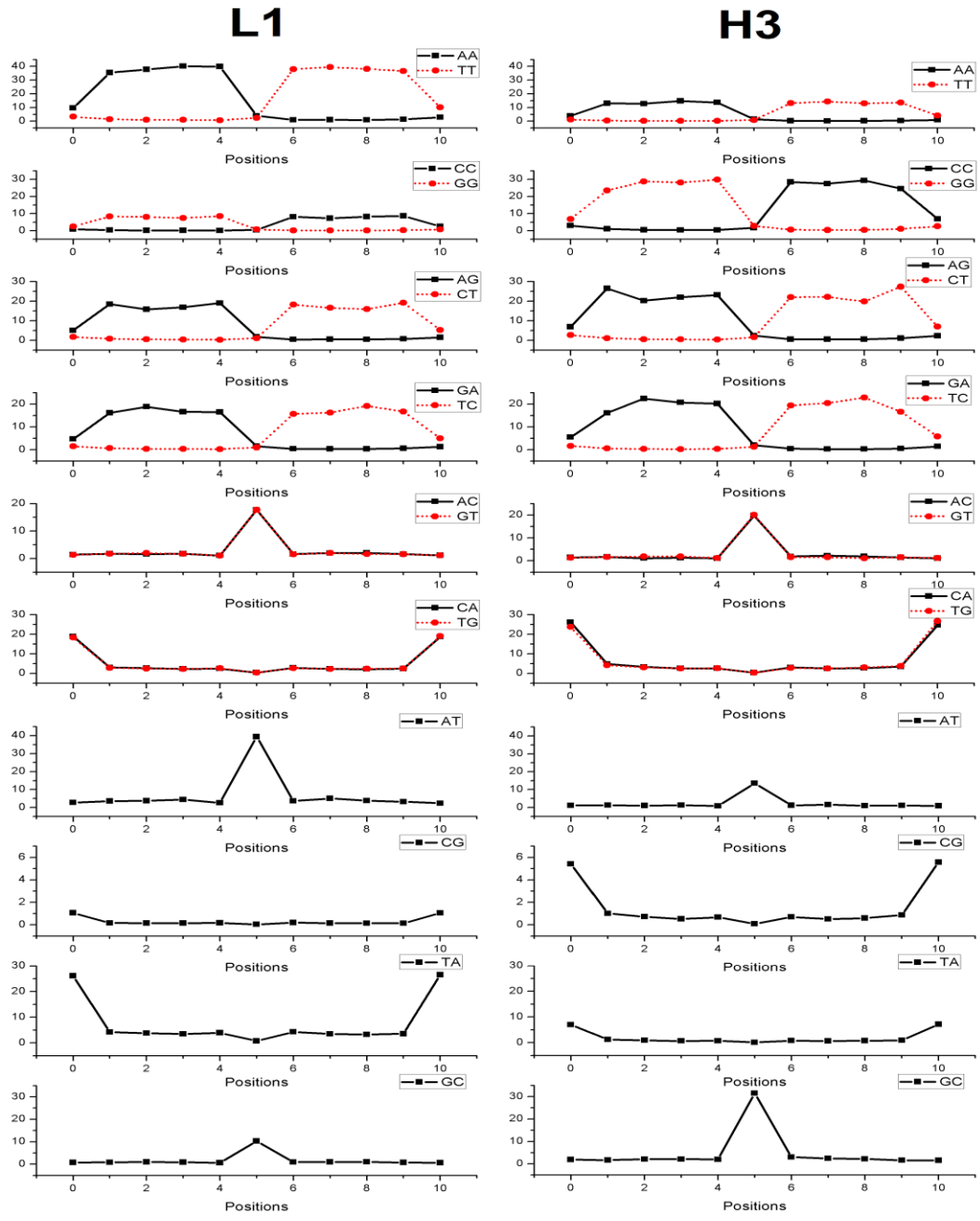
Nucleosome positioning patterns
of various isochores (Frenkel et al., 2011)
by N-gram extension

isochores G+C %

```
C AGGGG CCCCT G
C GGGGA TCCCC G
C AGAAA TTTCT G
T AAAAA TTTTT A
T AAAAA TTTTT A

Y RRRRR YYYYY R
```





Nucleosome positioning patterns for human isochores L1 and H3 derived by signal regeneration from apoptotic nucleosomes:

```
L1:   T AAAAA TTTTT A
H3:   C AGGGG CCCCT G
```

Frenkel et al., 2011

Nucleosome positioning patterns, isochores

(Frenkel, 2011, 2012)

		isochore	method
T AAAAA TTTTT A	L1	(<37% G+C)	B
T AAAAA TTTTT A		same	A
T AAAAA TTTTT A	L2	(37-41% G+C)	B
C AGAAA TTTCT G	H1	(41-46% G+C)	B
C GGGGA TCCCC G	H2	(46-53% G+C)	B
C AGGGG CCCCT G	H3	(>53% G+C)	B
C AGGGG CCCCT G		same	A
Y RRRRR YYYYY R		consensus	

A signal regeneration, nucleosomes

B Shannon N-gram extension, whole genome

Shannon N-gram reconstruction of linkers

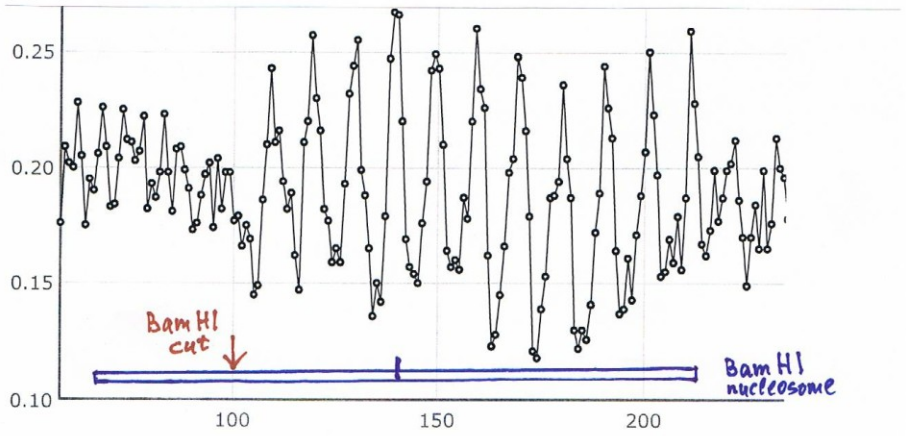
TTT**TA**TTT**TA**AAA**TA**AAA	human linkers
AAAA**TA**AAA**TA**TTT**TA**TTTT	yeast linkers
TAAAg**TA**cTT**TA**	human, apoptotic cuts

consensus:

TAxxx**TA**xxx**TA**xxx

(B. Salih, T. Bettecken, Z. Frenkel)

T**TA**AAAATTTT**TA**AAAATTTT**TA**A human L1 isochores, nucleosomes



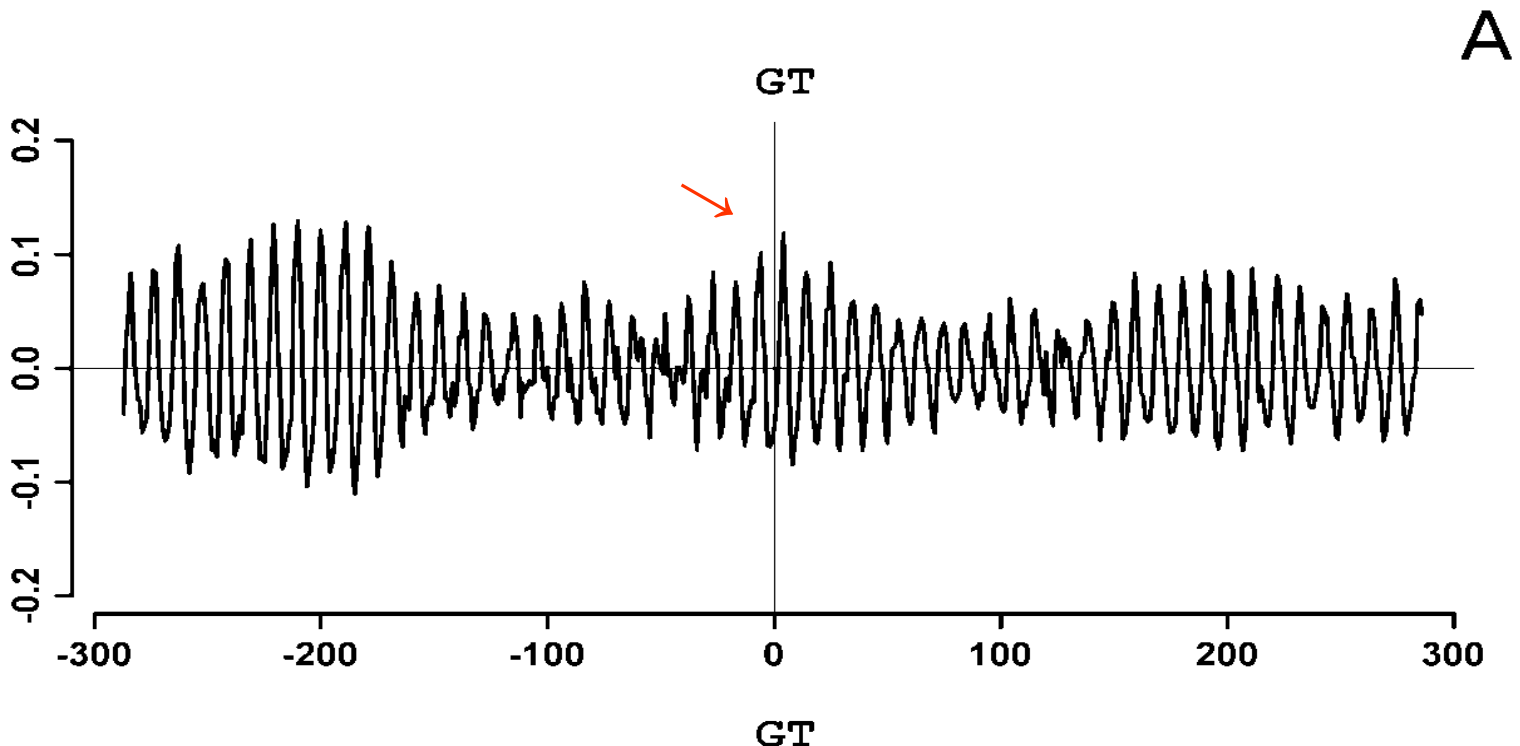
BamHI nucleosome of Ponder and Crawford, 1977

BamHI fragments of BamHI nucleosome DNA

Calculated	Observed in the gel	
24		
34		
43		
54	~53	
64	~63	misfit
(73)	(~73)	1 base
82	~83	
92	~93	
103		
112		
122		

Example of the nucleosomes at and around GT splice junction

Hapala, 2011



Plenty of various other nucleosome positioning patterns have been suggested during 30 years since the first observation of sequence periodicity.
At the best they provide occupancy maps (resolution of ~15 bases).

The (GRAAATTTYC)n and (RRRRRYYYYY)n are the only patterns that generate maps with single-base resolution, verified by crystal data.

The future of the chromatin structure/function is with the high resolution studies.