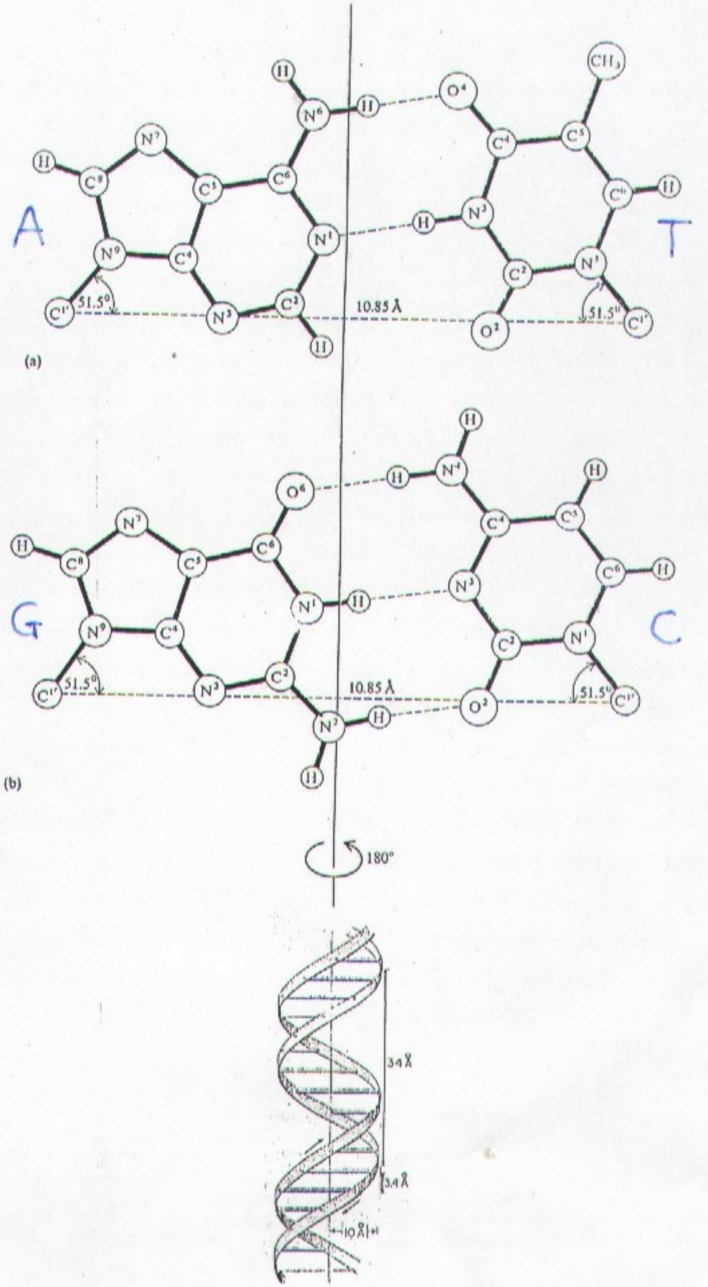
GENETIC CODES



A
T
G
C
(a)
(b)
H
N⁶
O⁴
CH₃
N⁷
C⁸
C⁵
C⁶
C⁴
N⁹
N¹
N³
C²
C¹'
O²
N⁴
N²
51.5°
10.85 Å
180°
34 Å
3.4 Å
10 Å

The idea on

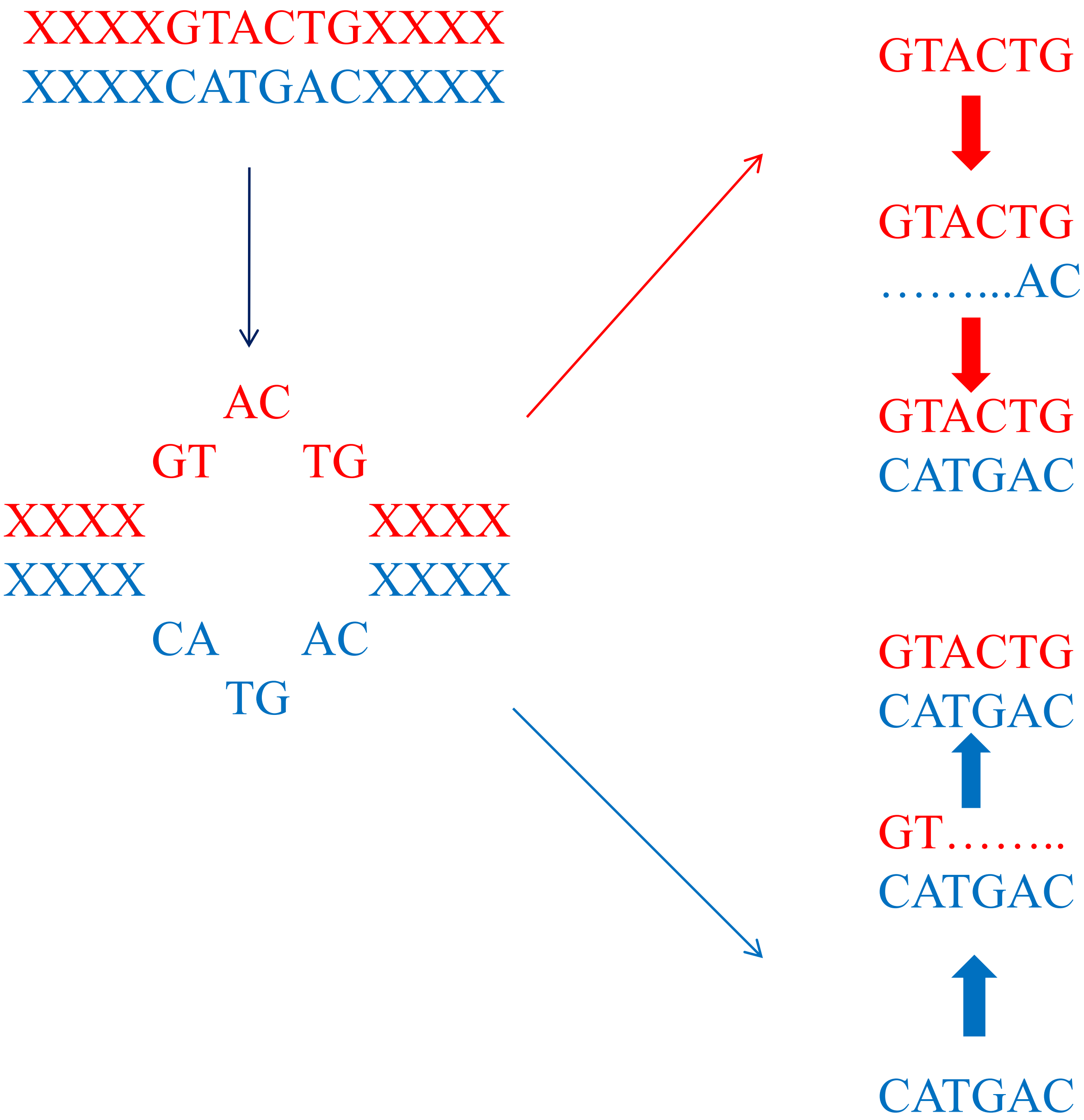
molecular complementarity
in macromolecular interactions

was outlined by
Linus Pauling and Max Delbruck
in 1940

Nature 371, 285, 1994

The paper of
Rosalind Franklin and Wilkins
with x-ray diffraction of A-DNA

appeared in the same issue of Nature
as the paper by Watson and Crick



XXXXGTACTGXXXX
XXXXCATGACXXXX

AC
GT TG
XXXX XXXX
XXXX XXXX
CA AC
TG

GTACTG

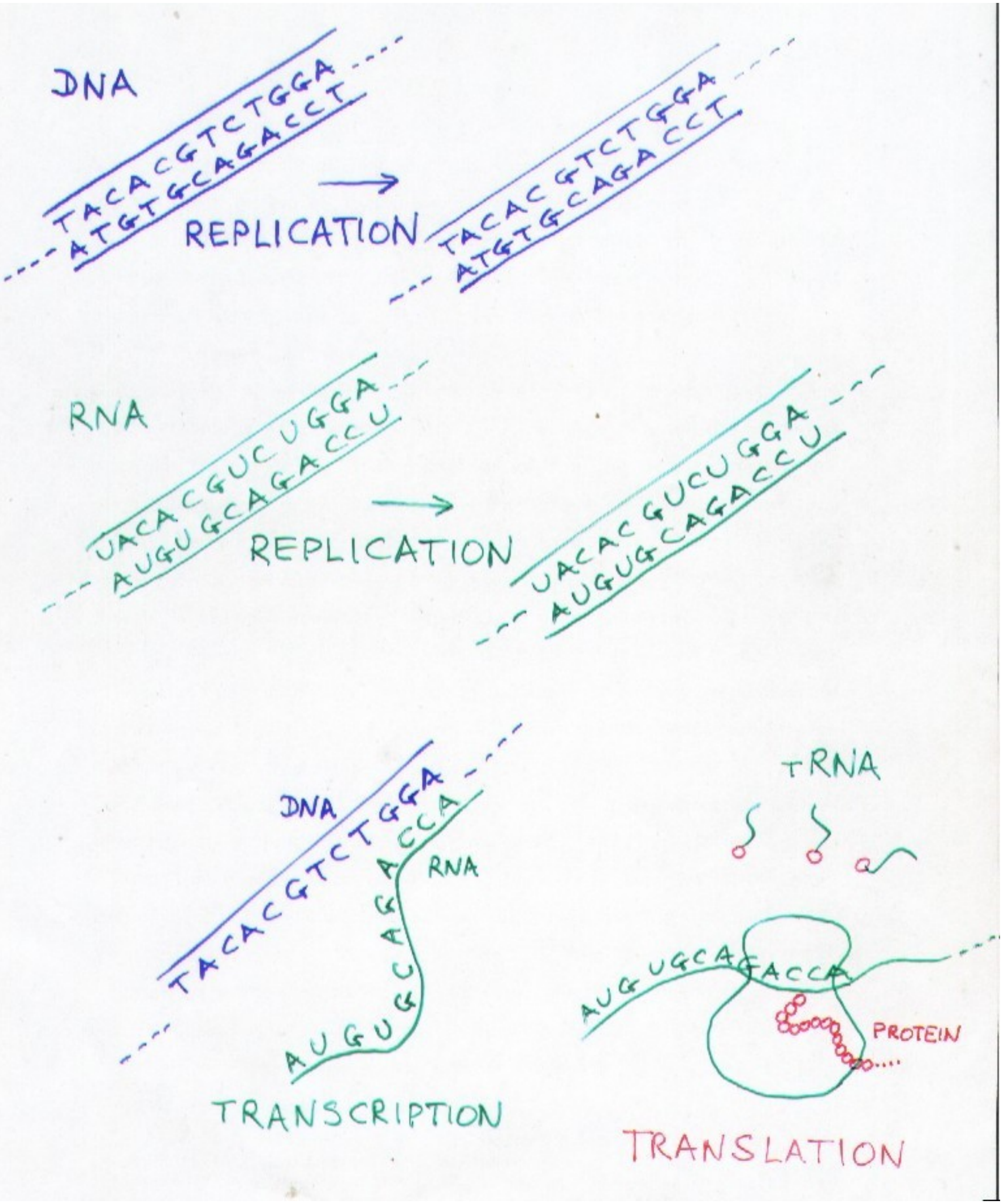
GTACTG
.........AC

GTACTG
CATGAC

GTACTG
CATGAC

GT........
CATGAC

CATGAC


DNA
TACACGTCTGGA
ATGTGCAGACCT
REPLICATION
TACACGTCTGGA
ATGTGCAGACCT
RNA
UACACGUCUGGA
AUGUGCAGACCU
REPLICATION
UACACGUCUGGA
AUGUGCAGACCU
DNA
TACACGTCTGGA
AUGUGCAGACCA
RNA
TRANSCRIPTION
tRNA
AUGUGCAGACCA
PROTEIN
TRANSLATION

“And now the announcement of Watson and Crick about DNA. This is for me the real proof of the existence of God”

Salvador Dali

Friedrich Miescher looked for hereditary material in sperm

and discovered DNA (1869).

He thought (1882) that the genetic information
may exist in the form of a molecular text,
a linear sequence of chemical symbols,

"just as the words and concepts of all languages
can find expression in twenty-four to thirty
letters of the alphabet"

Astbury and Bell (1938)
discovered
3.3 A periodicity in the fiber
x-ray diffraction of DNA –

-stacking of flat DNA bases

They also hypothesized that the
bases
"form the long scroll on which
is written the pattern of life".

Transforming activity of DNA

was first demonstrated by
O. Avery, S. MacLeod and M. McCarty
in 1944

For a long time (1906-1948)
DNA was viewed
as monotonous repetition of

identical tetranucleotide units

(Steudel, 1906; Levene and Simms, 1925)

Erwin Chargaff established the “Chargaff’s rule” in 1948:

A = T, and G = C

He was at the very doors of the discovery of DNA duplex structure.
Ruining the tetranucleotide theory, he was cautious with the obvious speculation, fearing to get in the shoes of Steudel and Levene,

…and missed the great discovery.

To the end of his days he was openly very bitter about that.

```
tgccattgcg ctccaaaaaa aaaaaaaaaa aagacattaa cataaattta aatattttat      2580
aatgacaatc cacattaact acttaaagca taagctattt tccaggagag gcagcaagtg      2640
cattctactc ccatgcccaa gaagaaagga gcgtgacttt ggtgggagta ctaggagttt      2700
ctactggagc acttgcccgc agagtgagaa acgttcctag agaggaagtt atacctgctg      2760
tggaatttaa gagaatcttg tcatattttg acaagttttt tgagatggaa gtctcactct      2820
gtcgcccagg ctggagtgca gtggcgcaat ctcagctcac tgcagcctgc acctcctcgg      2880
ctccagctat tctcttgtct cagcctcctg agtaactggg attacaggcg cccgccacta      2940
cgcctggcta atttttgtat ttttagtaga aatggggttt taccatgttg gccagactgg      3000
tctcaaactc ccgacctcag gtgatctgcc tgcctcagcc tcccaaagtg ctggaattac      3060
aggcgtgtgc cactgcgcct ggctaatttt tttttttttt tttttttagt agagacggtg      3120
gtttcaccat gtcatccagg ctggtctcaa actcctgacc tcaggtgatc cacccacctt      3180
ggtctaccaa agtgctcgga ttacaggcat gagccaccag gcccagtcaa cgtgatgtgt      3240
tttggaaccc tgaattcctt ggcttgcccg gagggttttc tttttgttaa tatctttgct      3300
tgctttctag tatttaaaaa attgtgtttt gctctaacta tgcaatggct ttaagtctta      3360
```

Sequence fragment from rDNA spacer of *Arabidopsis thaliana*

```
MSVNYMRLLCLMACCFSVCLAYRPSGNSYRSGGYGEYIKPVETAEAQAAALTNAAGAAASS
AKLDGADWYALNRYGWEQGKPLLVKPYGPLDNLYAAALPPRAFVAEIDPVFKRNSYGGAYG
ERTVTLNTGSKLAVSAAIGREAIVGAGLQGPFGGPWPYDALSPFDMPYGPALPAMSCGAGS
FGPSSGFAPAAAYGGGLAVTSSSPISPTGLSVTSENTIEGVVAVTGQLPFLGAVVTDGIFP
TVGAGDVWYGCGDGAVGIVAETPFASTSVNPAMSKSGVPRLLTASERERLEPIDQIHYSPR
ADDEYEYRHMLPKAMLKAIPTDYFNPETGTLRILQEEEWRGLGITQSGWEMYEVHVPEPHI
LLFKREKDYQMKFSQQRGGMLLNRTSFVTLFAAGMLVSALAQAHPKLVSSTPAEGSEGAAP
AKIELHFSENLVTQFSGAKLVMTAMPGMEHSPMAVKAAVSGGGDPKTMVITPASPLTAGTY
KVDWRAVSSDTHPITGSVTFKVKMSSQQQKQPCTLPPQLQQHQVKQPCQPPPQEPCVPKTK
EPCQPKVPEPCQPKVPEPCQPKVPEPCQPKVPQPCQPKVPEPCQPKVPEPCQPKVPEPCQP
KVPEPCQSKVPQPCQPKVPEPCQTKQKMADNLSQSFDKSAMTEEERRHIKKEIRKQIVAFA
LMIFLTLMSFMAVATDVIPRSFAIPFIFILAVIQFALQLFFFMHMKDKDHGWANAFMISGI
FITVPIAALMLLLGVNKISKIVKFLKELATPSHSMEFFHKPASNSLLASELNFVRRNIKRE
DFGHEVLTGAFGTLKSPVIVSIFHSRIVACEGGDGEEHDILFHTVAEKKPTICLDGQVFKL
KHISSEGEVMYYMFRQCAKRYASSLPPNALKPAFGPPDKVAAQKFKESLMATEKHAKDTSN
MWVKISVWVALPAIALTAVNTYFVEKEHAEHREHLKHVPDSEWPRDYEFMNIRSKPFFWGD
GDKTLFWNPVVNRHIEHDDQSTVHIVGDNTGWSVPSSPNFYSQWAAGKTFRVGDSLQFNFP
ANAHNVHEMETKQSFDACNFVNSDNDVERTSPVIERLDELGMHYFVCTVGTHCSNGQKLSI
NVVAANATVSMPPPSSSPPSSVMPPPVMPPPSPS
```

Aus der Harzreise, 1824,
Heinrich Heine.

Auf die Berge
Will ich steigen,

Wo die dunkeln
Tannen ragen,

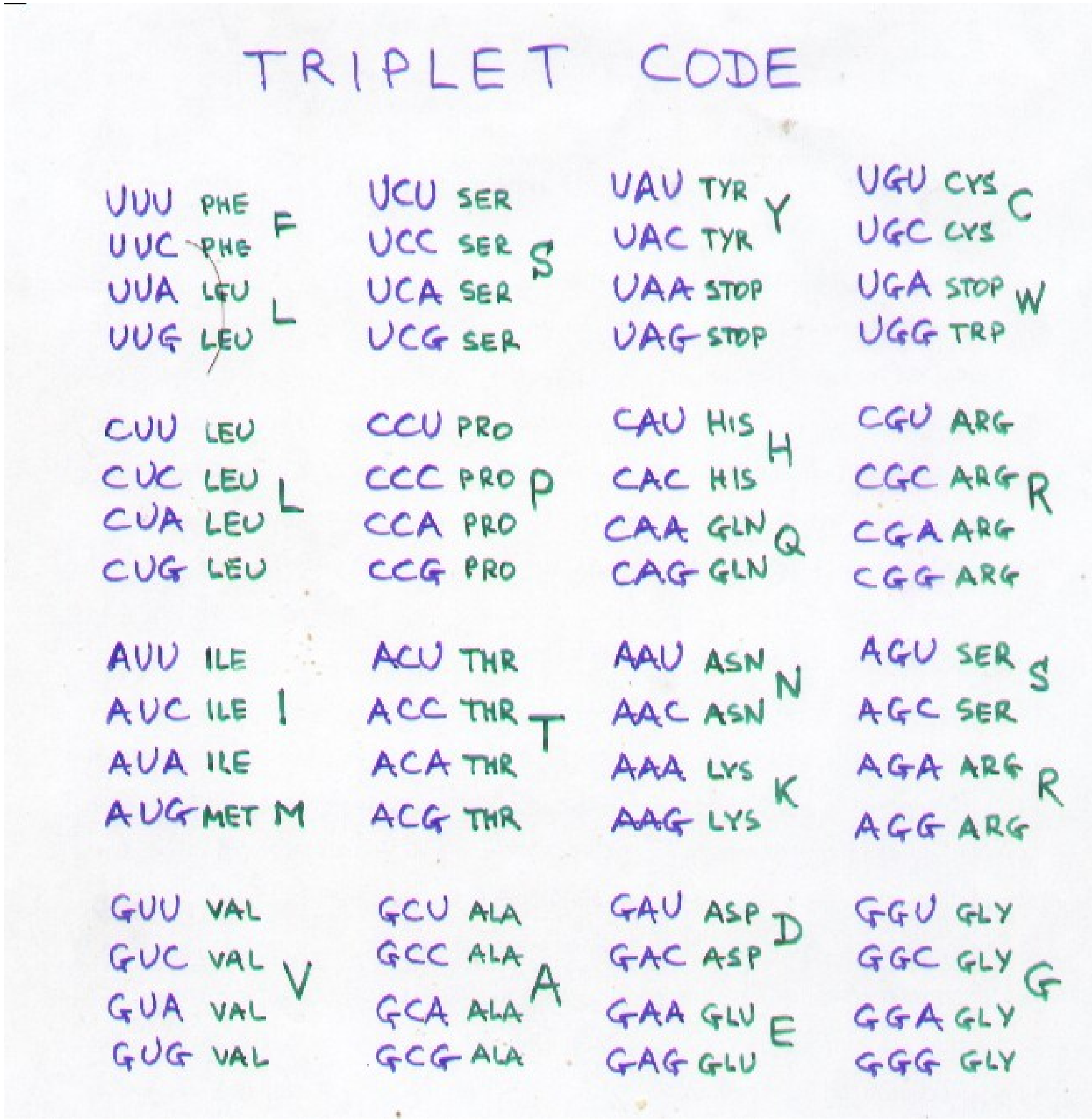
Bäche rauschen,
Vögel singen,

Und die stolzen
Wolken jagen.

Acrostic of Guido d'Arezzo (1025)

(on the hymn to St. John the Baptist)

Do (**Ut** in France)	***Ut*** *queant laxis*
Re	***Re**sonare fibris* (vocal chords)
Mi	***Mi**ra gestorum*
Fa	***Fa**muli tuorum*
Sol	***So**lve polluti*
La	***La**bii reatum* (tight lips)



TRIPLET CODE

UUU	PHE	F	UCU	SER	S	UAU	TYR	Y	UGU	CYS	C
UUC	PHE		UCC	SER		UAC	TYR		UGC	CYS	
UUA	LEU	L	UCA	SER		UAA	STOP		UGA	STOP	
UUG	LEU		UCG	SER		UAG	STOP		UGG	TRP	W
CUU	LEU	L	CCU	PRO	P	CAU	HIS	H	CGU	ARG	R
CUC	LEU		CCC	PRO		CAC	HIS		CGC	ARG	
CUA	LEU		CCA	PRO		CAA	GLN	Q	CGA	ARG	
CUG	LEU		CCG	PRO		CAG	GLN		CGG	ARG	
AUU	ILE	I	ACU	THR	T	AAU	ASN	N	AGU	SER	S
AUC	ILE		ACC	THR		AAC	ASN		AGC	SER	
AUA	ILE		ACA	THR		AAA	LYS	K	AGA	ARG	R
AUG	MET	M	ACG	THR		AAG	LYS		AGG	ARG	
GUU	VAL	V	GCU	ALA	A	GAU	ASP	D	GGU	GLY	G
GUC	VAL		GCC	ALA		GAC	ASP		GGC	GLY	
GUA	VAL		GCA	ALA		GAA	GLU	E	GGA	GLY	
GUG	VAL		GCG	ALA		GAG	GLU		GGG	GLY	

Experiment of Nirenberg and Matthaei (1961):

UUU	UUU	UUU	UUU	UUU	UUU	UUU	UUU	UUU	UUU
F	F	F	F	F	F	F	F	F	F

After random "mutations", incorporation of C instead of U, expected NEW triplets: CUU, UCU, UUC.
Three or less NEW aminoacids expected in the product

Only two new aminoacids detected:
serine (S) and leucine (L)

UUU	UCU	UUU	CUU	UUU	UUU	UCU	UUU	UUC	UUU
F	F	F	F	F	F	F	F	F	F
	or		or			or		or	
	S		S			S		S	
	or		or			or		or	
	L		L			L		L	
	or		or			or		or	
	none		none			none		none	

Final answer: CUU L
UCU S
UUC F

Multiple overlapping codes in the biological sequences

```
MnnnnnMnnnMMnnnnMnnMMMnnnMMnnnnnMnnMnnnnn  No.1
      |         |   ||    |
MnnnMnMnnnMMnMnnMnnMMMnMnMMnnnMnMnMMnnMnn  No.1 and No.2
      |         |   ||    |                  superimposed
nnnnMnMnnnnnnMnnMnnnMMnMnnMnnnMnnnMnnnMnn  No.2
```

Sidney Brenner:

The non-coding sequences
could not have been called "garbage“
instead of "junk", since
the garbage is to throw away
while the junk is to carry with.

Definition of the sequence code:

Any sequence pattern or bias responsible for specific biological or biomolecular function

(ENT, 1989)

There are, thus, many codes

Second Genetic Code Deciphered

The New York Times May 13, 1988

reported in today's issue of nature, by Ya-Ming Hou and Paul Schimmel

1988

1

1

work is important, but hardly most of the answer to the puzzle

that some call "the second genetic code“
and others call "the protein recognition problem."

C. Vaughan, Science News, May 28, 1988

DNA methylation, DNA's *[second !]*Second Code,

has been first announced under this name by Orion Genomics Company in 2001,
after publication: Martindale, Diane; "Genes Are Not Enough,"
Scientific American, 285:22, October 2001; and is broadly accepted since then.

See, e. g.:

Crack the Second Code: Methylated DNA Sequencing for Epigenetic Analysis
ETON Bioscience Inc 2003;

Imprinted Genes Offer Key to Some Diseases and to Possible Cures. By Sharon Begley,
Wall Street Journal. 24 June 2005.

2001

Packaging proteins may be

[third !] **second genetic code**

NewScientist **09 August 2001 by Emma Young**

Science (vol 293, from p 1068)

2001

I′ m done with seconds, can I have a third?

As an aside, the authors of the editorial summary coined the work as the second genetic code. I find this amusing, because this would be the third second genetic code.

The aminoacyl tRNA code was also coined the second genetic code, but people must have forgotten that, because another second genetic code was proposed in 2001. This genetic code describes how methylated DNA sequences regulate chromatin structure and gene regulation.

(*Todd Smith* , FINCHTALK Journal Club, May 11, 2010)

A genomic code for nucleosome positioning

Eran Segal, Yvonne Fondufe-Mittendorf, Lingyi Chen, AnnChristine Thastrom, Yair Field, Irene K. Moore, Ji-Ping Z. Wang & Jonathan Widom

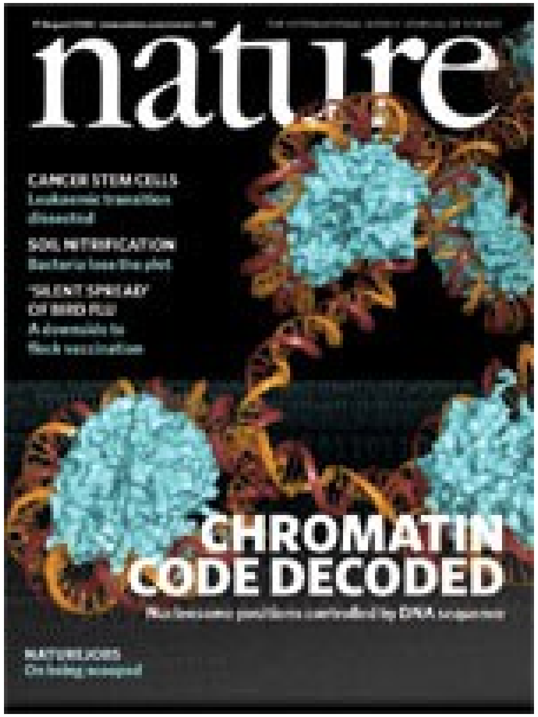
nature 442, 772-778, 2006

“a *[fourth !]*second code in DNA

in addition to the genetic code”

The New York Times July 25, 2006

2006


nature
CANCER STEM CELLS
Leukaemic transition dissected
SOIL NITRIFICATION
Bacteria lose the plot
'SILENT SPREAD' OF BIRD FLU
A downside to flock vaccination
CHROMATIN CODE DECODED
Nucleosome positions controlled by DNA sequence
NATUREJOBS
On being scooped

2006

The tendency of the dinucleotides to fit to … 10.5 or so base frame … can be considered as another message… two codes …

Trifonov, Nucl. Acids Res. 1980

“Chromatin code” – **chapter by Trifonov in "International Cell Biology 1980-1981"**

```
          minor
         groove
           out
            |
            |
n n n A A n n n T T n n n     team of Trifonov
            |                        1980-1996
            |
A A A n n G G C n n A A A     Satchwell et al.
T T T     G C C     T T T            1986
A A T     A G C     A A T
A T T     G C T     A T T
            |
            |
 A A n n n G C n n n A A      Segal et al.
 T T         |         T T           2006
 T A         |         T A
            |
            |
 C G R A A A T T T Y C G      team of Trifonov
                                     2009, 2010
```

Cracking the *[fifth !]* Second Genetic Code

Tim Hughes, *The FASEB Journal*. 2008;22:262.2

The interaction specificities between proteins and DNA has been termed the "second genetic code".

2008

Deciphering the splicing code

Yoseph Barash, John A. Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J. Blencowe & Brendan J. Frey

Breaking the *[sixth !]* second genetic code

J. Ramón Tejedor and Juan Valcárcel

nature, May 6, 2010

2010

SIX SECOND CODES:

three in nature,
one in Scientific American,
one in Science,
one in The FASEB Journal
one in common use

Many scientists have become "zombies":
they do not need to think
about important biological problems anymore,
instead, they simply go to the laboratory
and use the technical facilities available
to collect large quantities of data.

(Sidney Brenner)

The truth is that there are MANY codes in the sequences:

	discovered	cracked
1. RNA-protein translation (triplet) code	(1961)	(1961)
2. Genomic code (isochores)	(1973)	(1973-1990)
3. Chromatin (nucleosome positioning) code	(1980,1981)	(1980-2009)
4. DNA shape code (curved DNA)	(1980,1981)	(1980-1996)
5. Gene splicing code (Chambon rules)	(1981)	not yet
6. N-end rule (protein lifetime)	(1986)	(1986-1996)
7. Translation framing code	(1987)	(1987)
8. Fast adaptation (modulation) code	(1989)	(1989)
9. Genome segmentation code	(1994)	not yet
10. Codes of small RNAs	(1998)	(1998)
11. Translation pausing code	(2002)	(2002)
12. Proteomic code (proteins)	(2003)	(2003-2008)
13. Genome inflation code	(2010)	(2010)

....

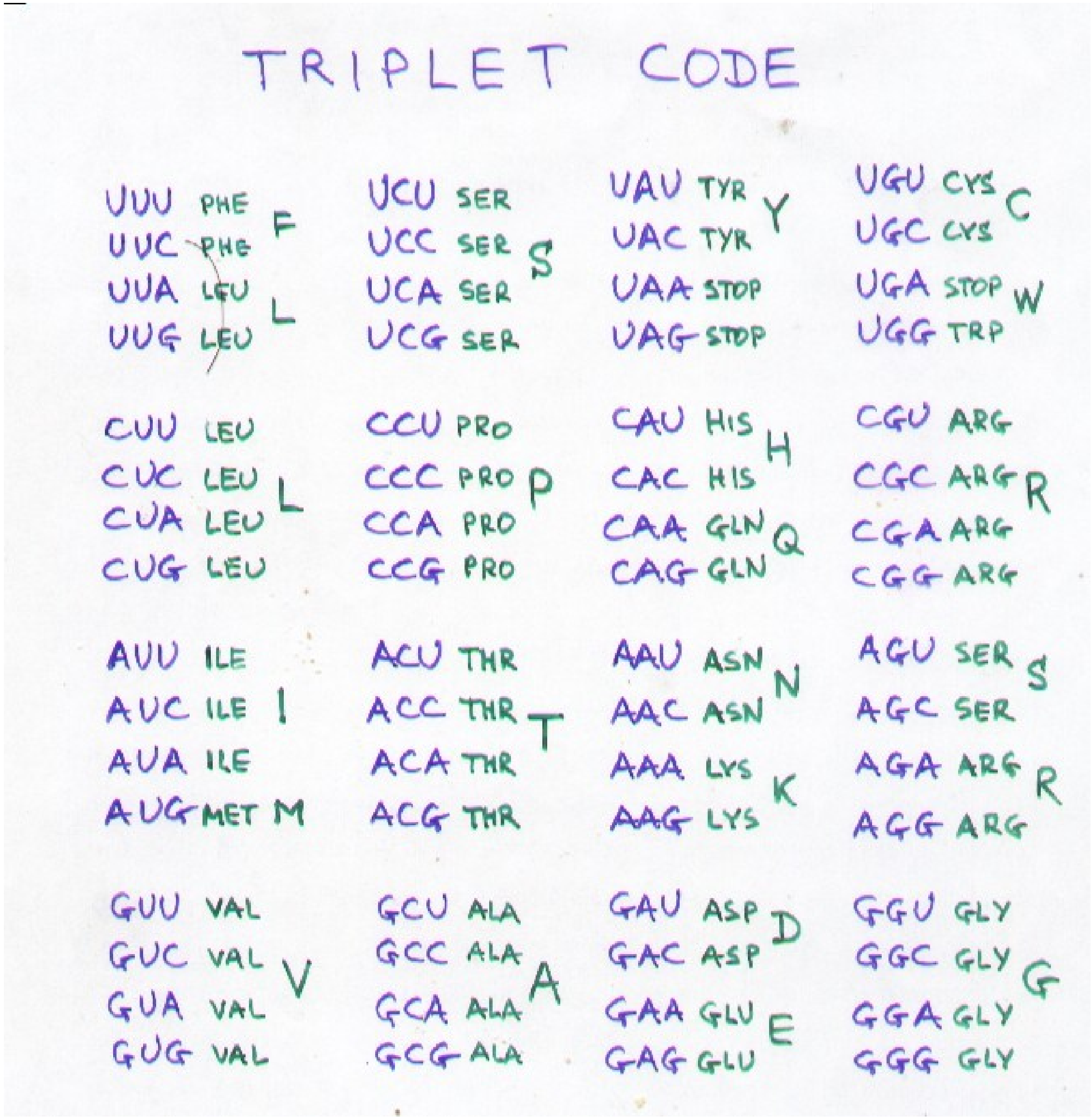
Several more sequence patterns are known, that qualify as general codes:

Transcription initiation code (promoters)
Transcription termination code (terminators)
Poly-adenylation code

And this is common knowledge, essentially, since 1989:

Trifonov, E. N., Bull. Math. Biol. 51, 417-432 (1989)
Trifonov, E. N., Sequence codes. In: "Encyclopedia of Molecular Biology", 1999

Triplet code
(RNA-protein translation code)



TRIPLET CODE

UUU	PHE	F	UCU	SER	S	UAU	TYR	Y	UGU	CYS	C
UUC	PHE		UCC	SER		UAC	TYR		UGC	CYS	
UUA	LEU	L	UCA	SER		UAA	STOP		UGA	STOP	
UUG	LEU		UCG	SER		UAG	STOP		UGG	TRP	W
CUU	LEU	L	CCU	PRO	P	CAU	HIS	H	CGU	ARG	R
CUC	LEU		CCC	PRO		CAC	HIS		CGC	ARG	
CUA	LEU		CCA	PRO		CAA	GLN	Q	CGA	ARG	
CUG	LEU		CCG	PRO		CAG	GLN		CGG	ARG	
AUU	ILE	I	ACU	THR	T	AAU	ASN	N	AGU	SER	S
AUC	ILE		ACC	THR		AAC	ASN		AGC	SER	
AUA	ILE		ACA	THR		AAA	LYS	K	AGA	ARG	R
AUG	MET	M	ACG	THR		AAG	LYS		AGG	ARG	
GUU	VAL	V	GCU	ALA	A	GAU	ASP	D	GGU	GLY	G
GUC	VAL		GCC	ALA		GAC	ASP		GGC	GLY	
GUA	VAL		GCA	ALA		GAA	GLU	E	GGA	GLY	
GUG	VAL		GCG	ALA		GAG	GLU		GGG	GLY	

Note to degeneracy of triplet code

```
Original sequence:   TACTCGCTAACCGTAGGGGCCCGG
       Sequence I:   T  T  C  A  G  G  G  C
      Sequence II:    A  C  T  C  T  G  C  G
     Sequence III:     C  G  A  C  A  G  C  G
```

It turned out that
the third position sequence
is the most deviant from random)

(Sasha Rapoport, 2008)

OUT-OF-CONTEXT SEQUENCES I, II and III

```
original seq.  ACC GCU AUA CAG AUG UGU CAU ACC GCC CAU GAC GGC ACU UGC AAU GCA CGU UUA
     I         A   G   A   C   A   U   C   A   G   C   G   G   A   U   A   G   C   U
     II         C   C   U   A   U   G   A   C   C   A   A   G   C   G   A   C   G   U
     III         C   U   A   G   G   U   U   C   C   U   C   C   U   C   U   A   U   A
```

original seq.   ACCGCUAUACAGAUGUGUCAUACCG**CCC**AUGACGGCA**CUU**GCAAUGCACG**UUU**A

I   AGACAUCAGCGGAUAGCU

II   **CCU**AUGACCAAGCGACGU

III   CUAGG**UUCCUCCUCU**AUA

A. Rapoport, 2008

The end of the first lecture
(Brno 2011)

```
(a)
. . . G A G T C C T G G C A A G A A T A C C A A G A G T T C C T C G G T T T G C C A G T T . . .

      G A   T C   T G   C A   G A   T A   C A   G A   T T   C T   G G   T T   C C   G T     1) Gene TRP1
      glu   ser   trp   gln   glu   tyr   gln   glu   phe   leu   gly   leu   pro   val

      G                       G                 G                 G                 G       2) framing of TRP1

      G A G               A A G A       C C   A G A G     C C T C             C C           3) nucleosome

(b)
. . . A C A G T T G T C A C G C T G A T T G G T G T C G T T A C A A T C T A A C G C . . .

      A C   G T   G T   A C   C T   A T   G G   G T   G T   A C   A T   T A A              1) end of frdD gene
      thr   val   val   thr   leu   ile   gly   val   val   thr   ile   term

            G     G                       G     G     G                                     2) framing of frdD

              T T G   C A                                 T A   A A T                       3) promoter P1
                                                                                               of ampC gene

(c)
. . . T C G A A G T G G A C T G C T G G T G G A A A A T G A G G A A A T T C A A . . .

      T C   A A   T G   A C   G C   G G   G G   A A   T G A                                1) Gene A,A*
      ser   lys   trp   thr   ala   gly   gly   lys   term

                              G     G     G                                                 2) framing of A,A*

        C G   A G   G G   C T   C T   G T   G A   A A   G A   G A   A T   C A              3) Gene K
        arg   ser   gly   leu   leu   val   glu   asn   glu   glu   ile   gln

                    G                 G     G           G     G                             4) framing of K

                                                    A T G A G   A A   T T   A A             5) Gene C
                                                    fmet  arg   lys   phe   asn
```

Translation framing code

Distribution of bases in three codon positions

	I	II	III	I	II	III	I	II	III
A	2442	2756	1290	1212	1243	766	557	488	481
C	2005	1900	2999	859	1032	1316	194	486	475
G	2723	1618	2688	1257	780	1036	561	344	180
U	1612	2508	1805	772	1045	982	395	389	571
		Human			Mouse			Ciliates	
A	538	495	478	1496	1573	1044	660	830	606
C	263	470	317	561	1271	1229	503	517	666
G	575	290	98	1690	652	848	798	373	490
U	383	504	866	1063	1314	1689	396	637	595
		Dictyostelium			Yeast			Plants	
A	4933	6064	3608	662	824	603	463	569	323
C	4723	4479	5586	401	535	450	480	479	600
G	7314	3497	5311	773	359	550	729	340	595
U	2767	5697	5232	449	567	682	312	596	466
		E. coli			Bacilli			*S. typhimurium*	
A	387	455	242	4701	3025	6212	1273	1355	1555
C	382	385	575	3121	3620	3917	985	1339	951
G	599	312	459	3173	1808	749	1990	1100	681
U	241	457	333	3597	6139	3714	1290	1744	2351
		Rhizobiaceae			Mitochondria			Chloroplasts	
A	551	596	495	682	705	556	861	916	793
C	292	380	238	657	738	721	410	462	546
G	547	316	353	912	569	849	641	311	390
U	354	452	658	474	713	599	391	614	574
		SV40			RSV			CMV	
A	1048	1119	958	945	1162	653	641	688	499
C	490	712	419	662	691	924	557	586	625
G	1107	547	380	1164	594	828	880	494	736
U	620	887	1508	554	878	920	461	771	679
		T4			T7			Transposons	
A	883	948	906	660	685	571	25595	26496	22639
C	209	418	157	551	617	674	18305	21117	23385
G	684	348	185	841	459	584	28958	15111	17990
U	614	676	1142	464	755	687	17209	27343	26053
		Plasmid K1			Plasmid Ti			Total	

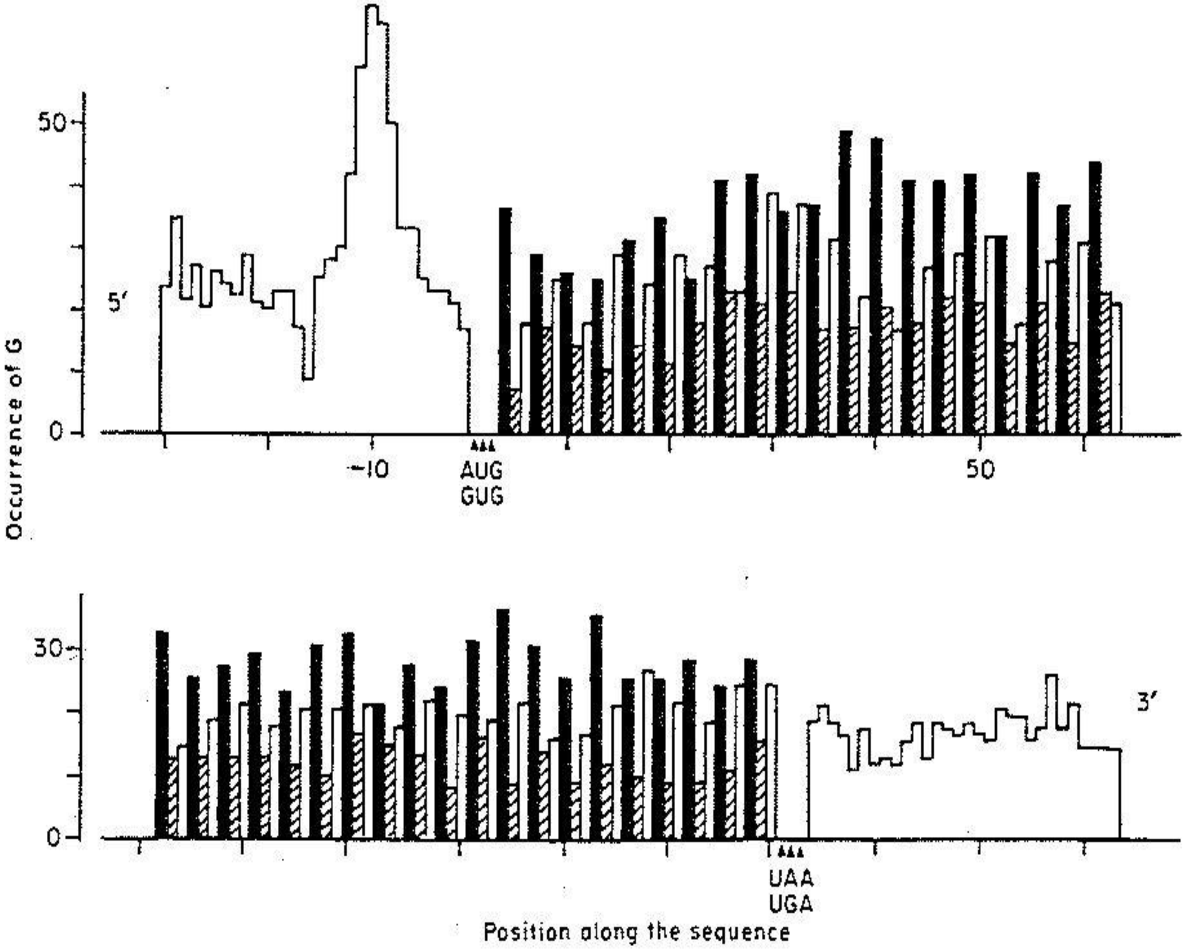


Figure 1. Distribution of guanines along *E. coli* mRNA. Filled bars, first positions of the codons; hatched bars, second positions. Only the first and last 60 bases of the coding regions are presented.

The three-base periodicity suggests that the ribosome may recognize correct reading frame far away from initiation triplet AUG.

Why that would be needed?

Does ribosome always move by exactly three steps?

It does not!

Occasionally, ribosome makes mistakenly two base steps instead, or 4 base steps.

That is, the ribosome may spoil the reading frame,
and synthesize protein with wrong sequence,
starting from the site of the mistake.

I n 1972 John Atkins (Ireland) discovered that
a mutant bacterial strain with frameshift mutation
is still able to produce normal gene product
in small amount.

Despite various measures to exclude contamination
by wild type strain the effect persisted.

In discussion Atkins suggested several possible
reasons why the apparently mutated gene was still able
to direct synthesis of normal protein, and concluded:

But, of course, the ribosome can not possibly
jump forward or backwards.

And that, actually, was exactly what was happening.

Frameshift mutation,
and translational frameshifting
are **different phenomena.**

First is a mishap caused by insertion/deletion
(gene sequence changed)

Second is a mishap (or happy accident)
caused by failure of the ribosome
to correctly count triplets
(no change in the gene sequence)

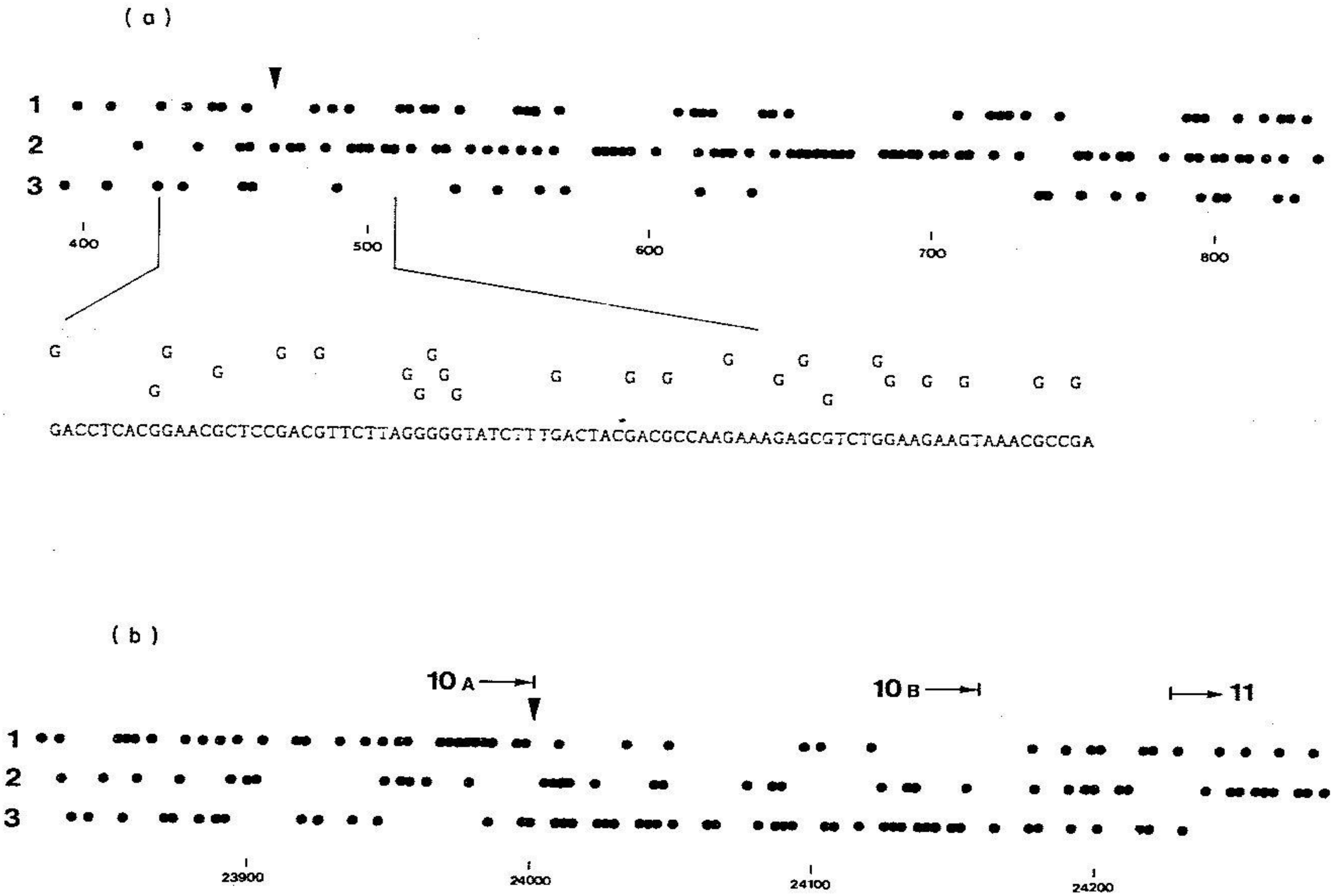


Figure 3. Actual distribution of guanines in 3 frames of the *RF-2* gene of *E. coli* (a) and the *10A,B* gene of bacteriophage T7 (b). The sequence around the ribosome slippage site is also shown (a). Every occurrence of G is indicated by a dot. Arrowheads indicate positions of ribosome frameshifting. Sequence co-ordinates correspond to those in original papers (Craigen *et al.*, 1985; Dunn & Studier, 1983).

Potential mRNA binding sites in 16 S rRNA

$(NNC)_n$ sites	Stickiness to *E. coli* $(GNN)_n$ mRNA	Exposed loops
(1395)caCacCucC	1·19	+
(517)gcCagCagCcgC	1·17	+
(629)aaCugCauC	1·15	
(499)agCacCggC	1·13	
(1061)guCguCagC	1·13	
(803)guCcaCgcC	1·11	
(306)acCagCcaC	1·11	
(1312)guCugCaaC	1·10	
(874)guCgaCcgC	0·97	
(1531)auCacCucC	0·96	+
(891)uaCggCcgC	0·92	
(993)gaCauCcaC	0·89	
(1095)ucCcgCaaC	0·88	
(1257)agCgaCcuC	0·80	
(730)ggCggCccC	0·73	
(1320)cuCgaCucC	0·52	
(337)gaCucCuaC	0·44	

mRNA binding sites in 16 S rRNA

(517)G C $\overset{\bullet}{C}$ A G $\overset{\bullet}{C}$ A G $\overset{\bullet}{C}$ C G $\overset{\bullet}{C}$ G G U A A U(534)

(1392)G U A C A $\overset{\bullet}{C}$ A C $\overset{\bullet}{C}$ G C $\overset{\bullet}{C}$ C G U C A(1408)

(1530)G A U $\overset{\bullet}{C}$ A C $\overset{\bullet}{C}$ U C $\overset{\bullet}{C}$ U U A(1542)

mRNA consensus (J. Lagunez-Otero, 1992)

$(\mathrm{GHN})_n$ - obvious pattern (1987)

$(\mathrm{GHU})_n$ - normalized base distributions

$(\mathrm{GCU})_n$ - dinucleotide preferences

$(\mathrm{GCU})_n$ - avoidance of bad mismatches

$(\mathrm{GCU})_n$

```
5’-U GCU GCU GCU GCU G   mRNA consensus
   • ••• ••• ••• ••• •
3’-A UGG CGC CGA CGA C   525 site of 16S rRNA
                          (proof-reading site)
```

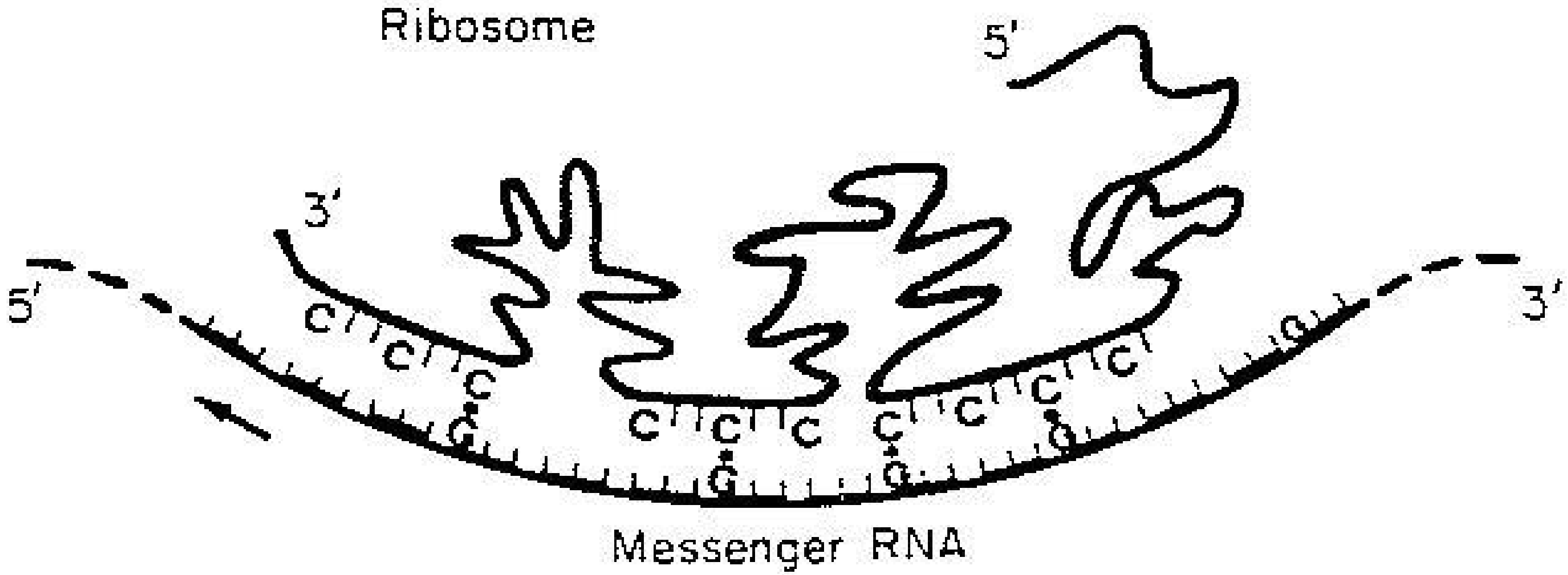
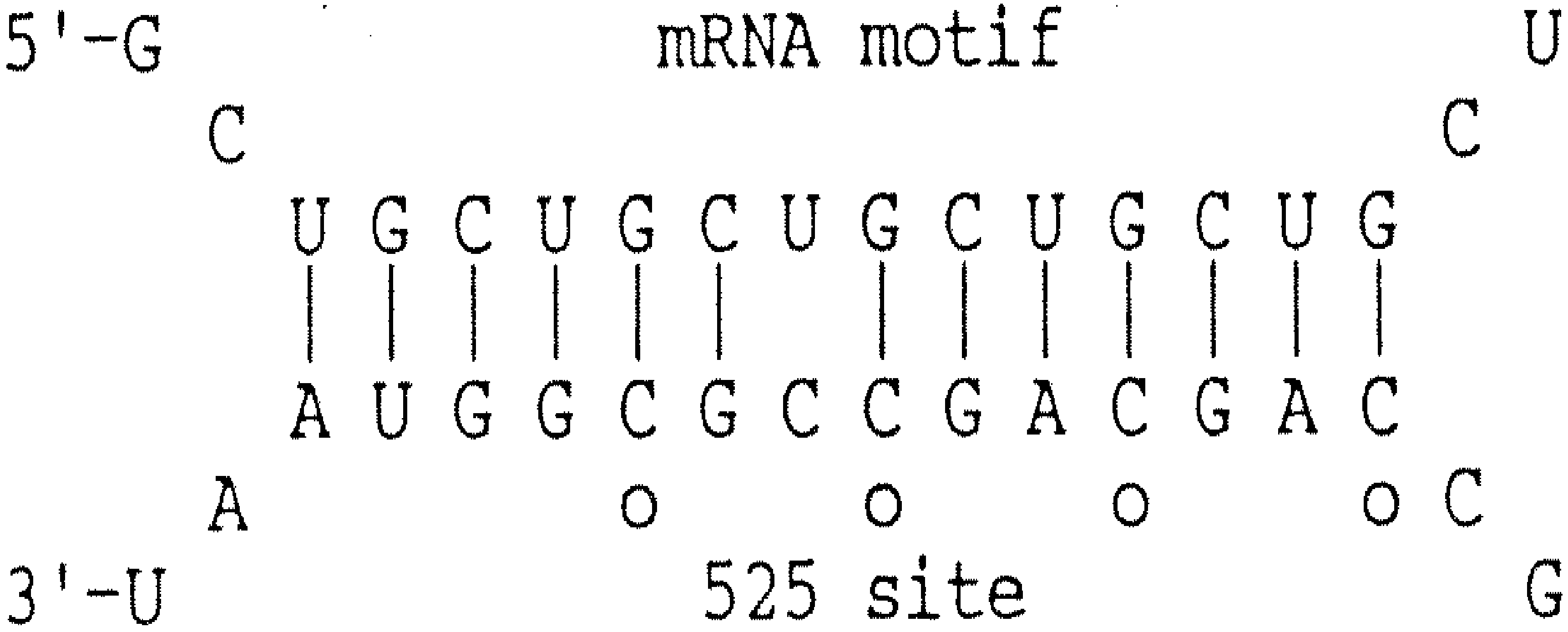


Figure 4. Scheme of the translation frame-monitoring mechanism.

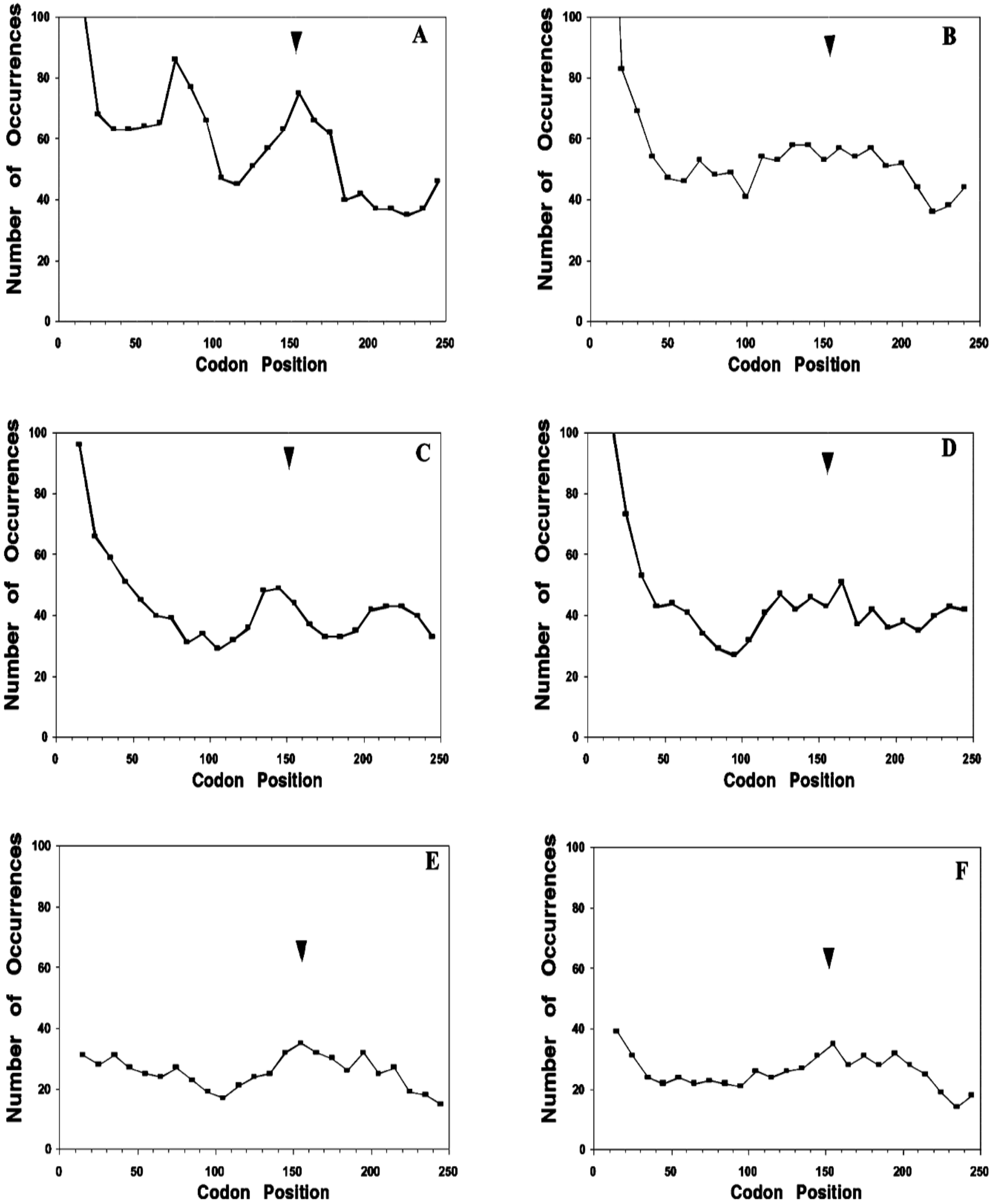
ENT, 1987

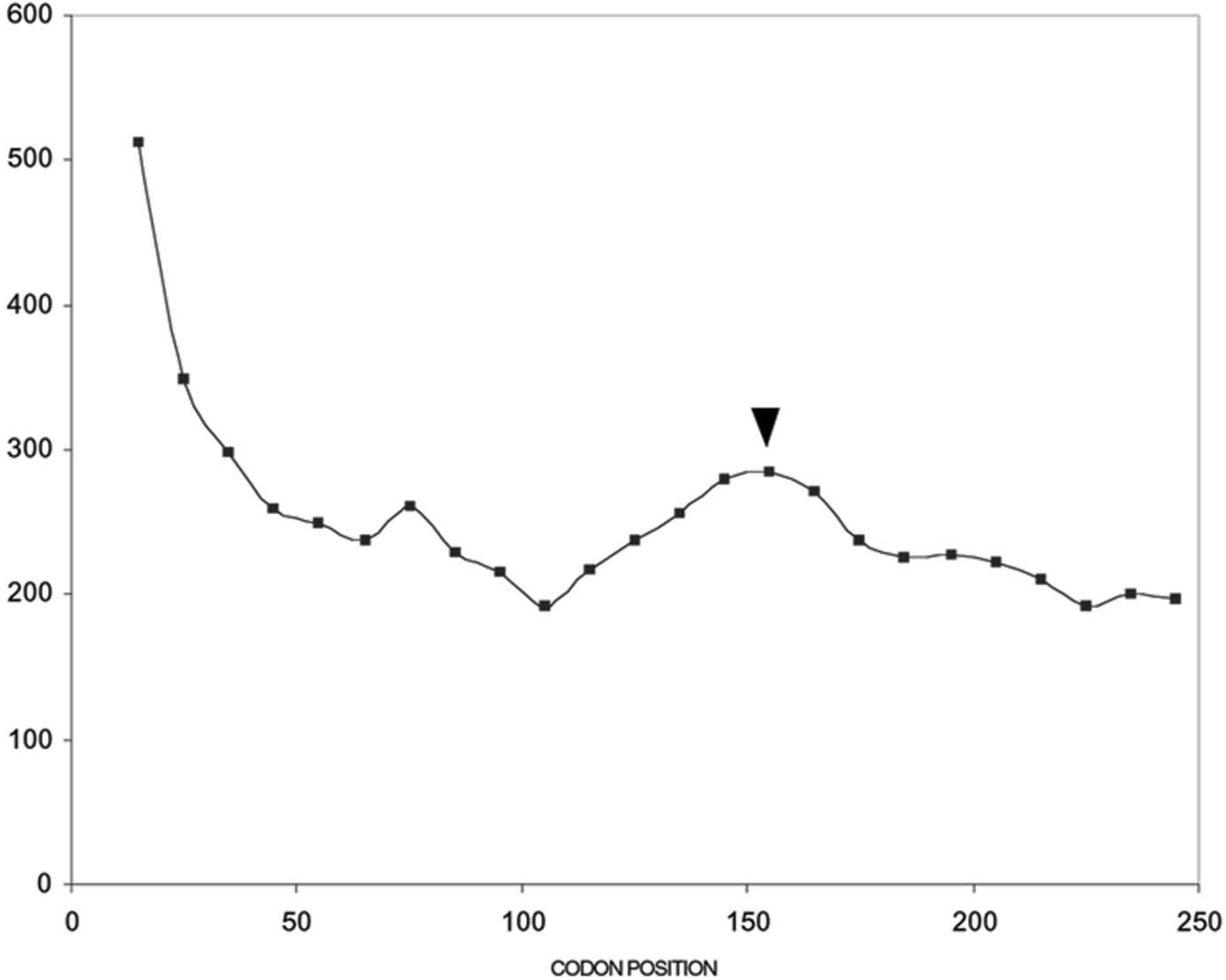


```
5'-G          mRNA motif            U
    C                             C
      U G C U G C U G C U G C U G
      | | | | | |   | | | | | | |
      A U G G C G C C G A C G A C
    A         o     o     o     o C
3'-U            525 site            G
```

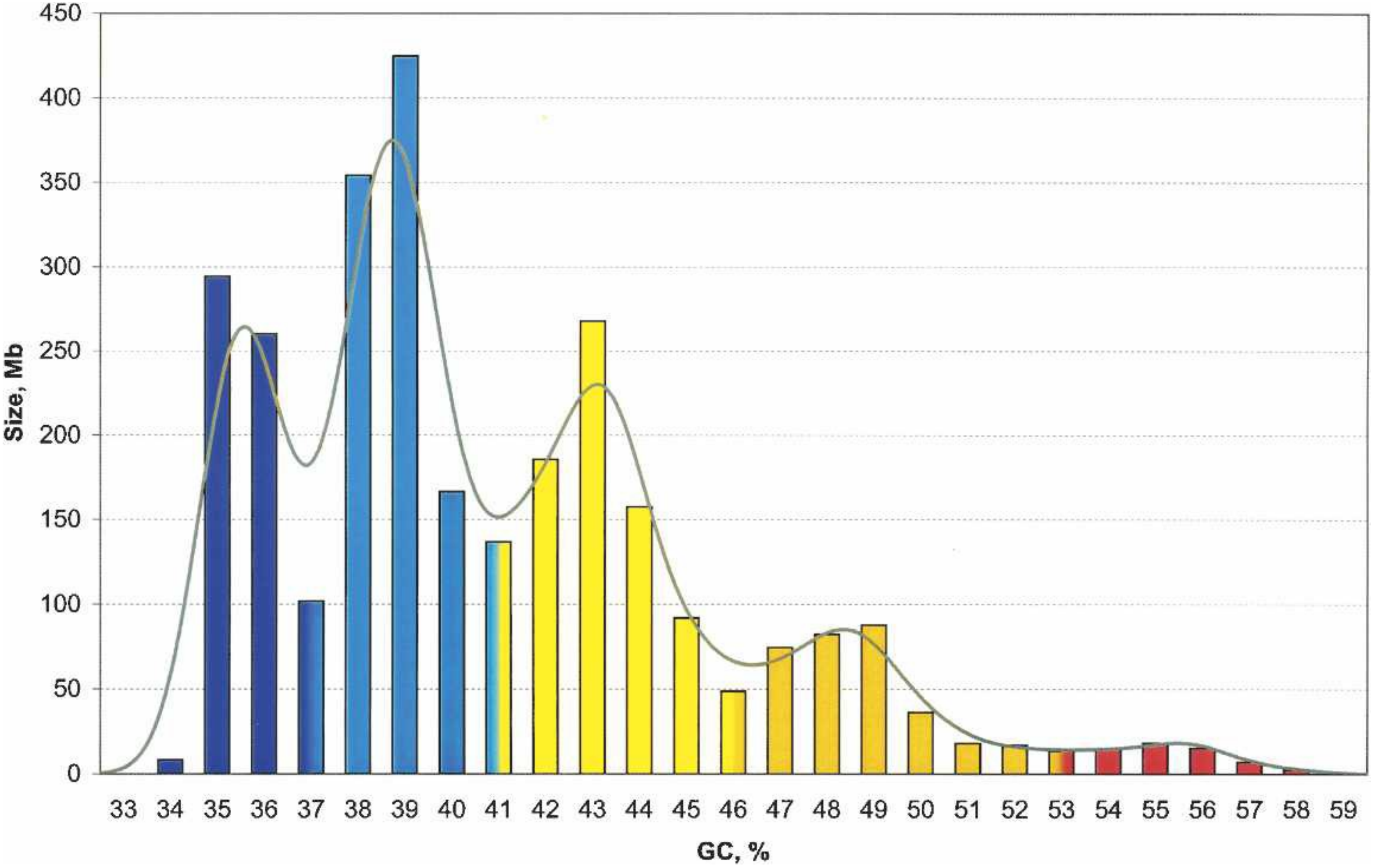
Which one is more ancient?

Translation pausing code


A
Number of Occurrences
100
80
60
40
20
0
0
50
100
150
200
250
Codon Position
B
Number of Occurrences
100
80
60
40
20
0
0
50
100
150
200
250
Codon Position
C
Number of Occurrences
100
80
60
40
20
0
0
50
100
150
200
250
Codon Position
D
Number of Occurrences
100
80
60
40
20
0
0
50
100
150
200
250
Codon Position
E
Number of Occurrences
100
80
60
40
20
0
0
50
100
150
200
250
Codon Position
F
Number of Occurrences
100
80
60
40
20
0
0
50
100
150
200
250
Codon Position

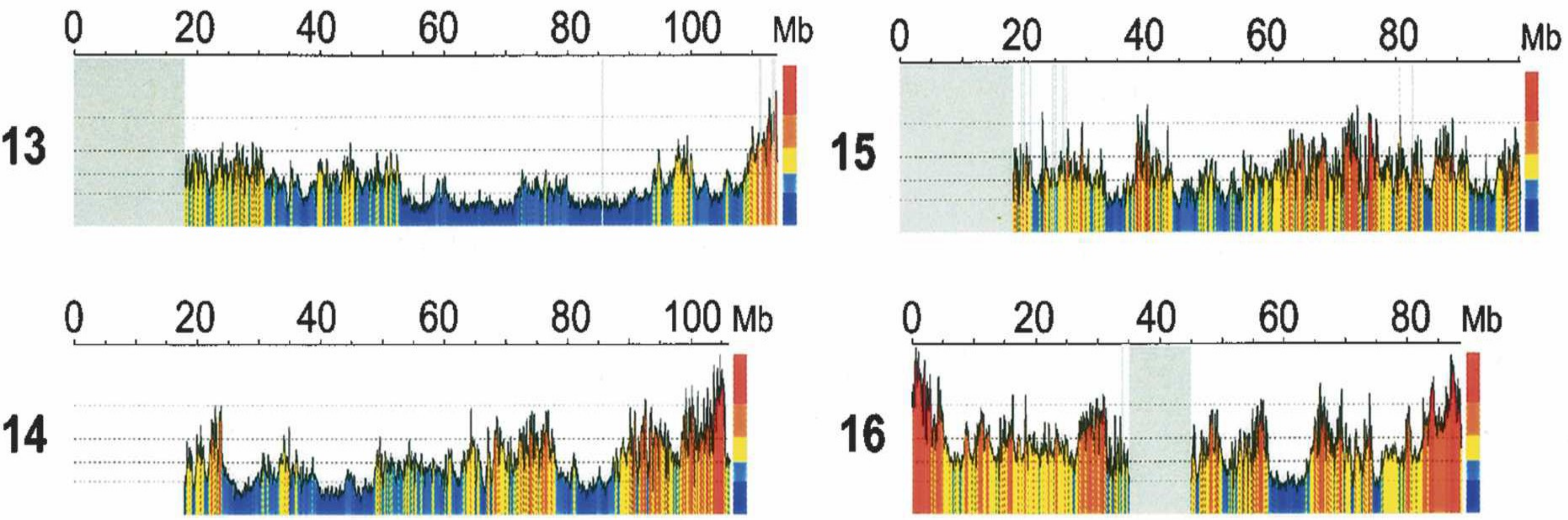

600
500
400
300
200
100
0
0
50
100
150
200
250
CODON POSITION

Genomic code (isochores)



450
400
350
300
250
200
150
100
50
0
Size, Mb
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
GC, %

H3	🟥	>53
H2	🟧	46-53
H1	🟨	41-46
L2	🟦	37-41
L1	🟦	<37



Isochores

Lab of G. Bernardi, 2006

Transcription factor binding sites
in G+C rich isochores are G+C rich as well

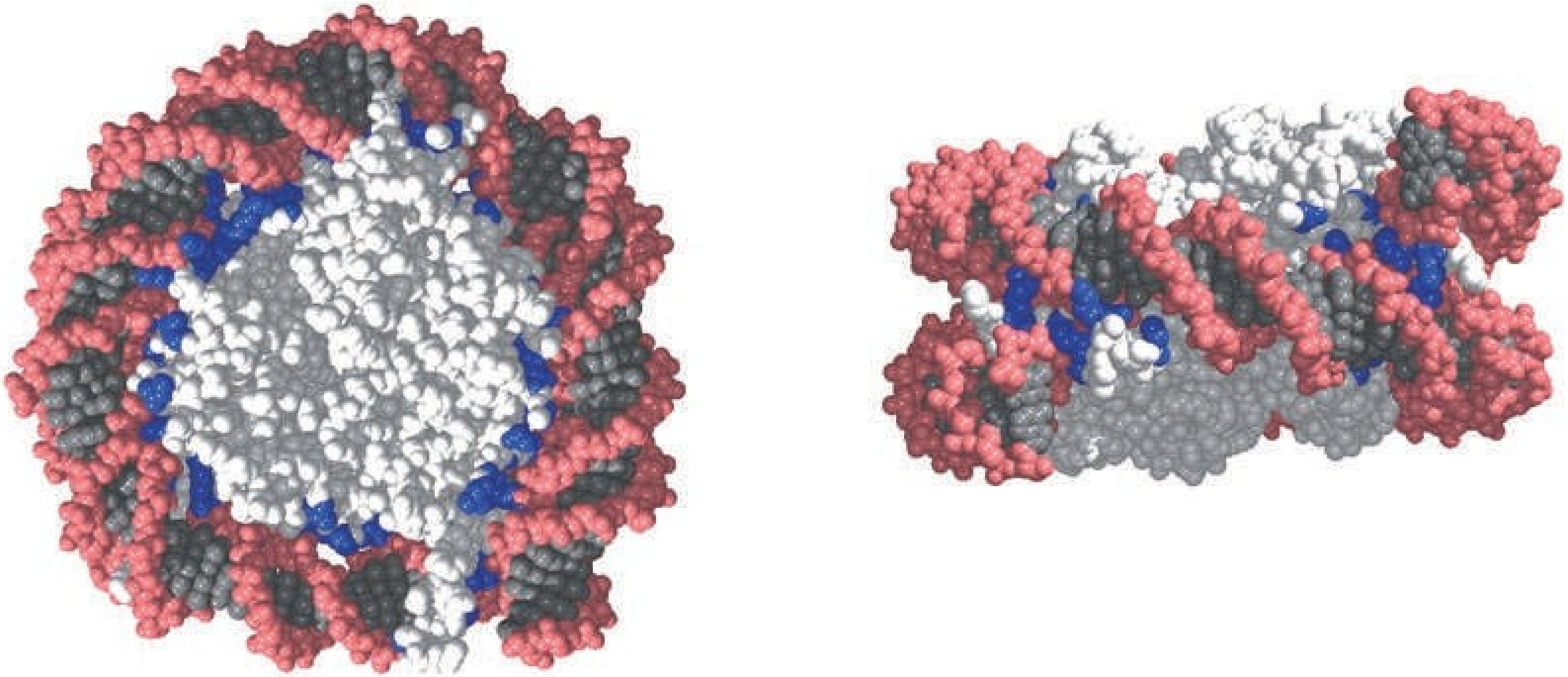
This results in different usage of transcription factors
in different isochores

In other words, each isochore type in the genome
is under isochore-specific separate regulatory system

In that sense isochores appear as individual mini-genomes
within the genomes

Apparently, modern eukaryotic genomes are mosaics of
many fused small ancestral genomes

DNA SHAPE CODE (CURVED DNA)



S. Tan, Pennsylvania State University, USA.

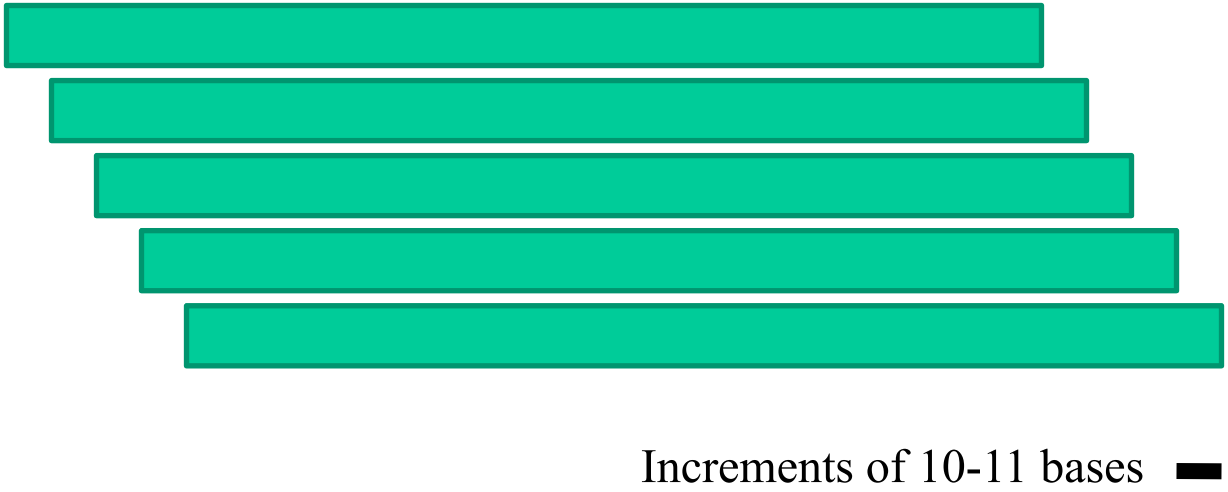
Since 1974 the experimental evidence started to accumulate suggesting that

1. Nucleosomes prefer some specific sequences

2. Comparisons of the sequences do not show anything in common

3. Often there are several alternative nucleosome positions on the same sequence

4. The alternative positions are separated by 10-11 bases



Increments of 10-11 bases

Separation of the nucleosome positions by 10-11 bases
(one structural period of DNA helix)
means that

The DNA molecule binds to histone octamers by one side

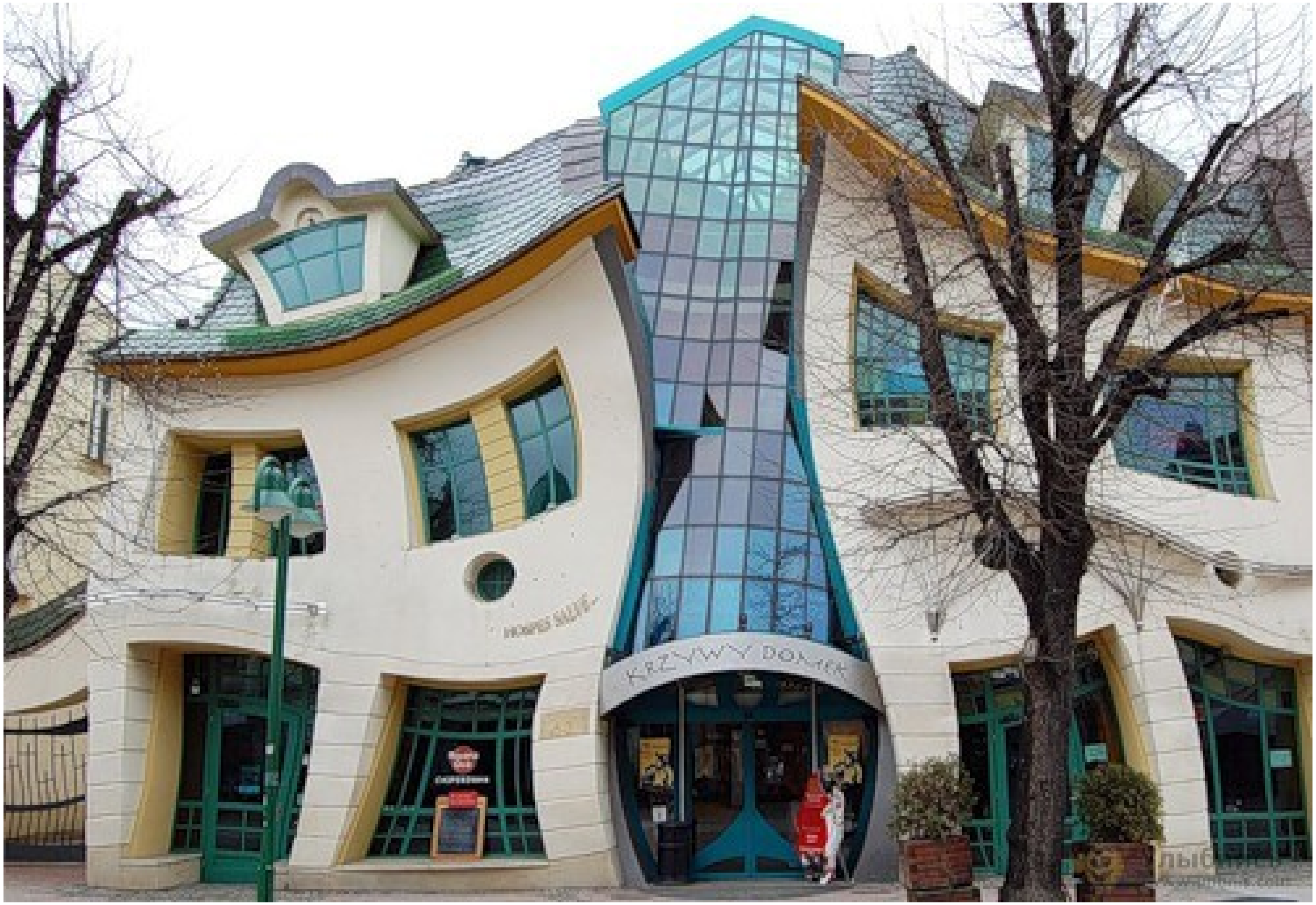
Physically, there are two ways to make DNA sided:

1. DNA may have the curvilinear shape, with arc-like axis – **Curved DNA**

2. DNA (straight DNA) could be easier bent in certain direction – **Bent DNA**

One is arc-like because it has that shape (like banana) – no force applied (curved DNA)

Another one is arc-like because the bending force is applied to it (bent DNA)



Krzywy domek (Curved house), Sopot, Poland

Object of curvilinear shape is called

Кривой	Согнутый	(Russian)
Křivý	Ohnutý	(Cžech)
Krzywy	?	(Polish)
Krumm	?	(German)
Curved	Bent, (but also **Curved**)	(English)
↑	↑	
no force applied	**actively deformed**	

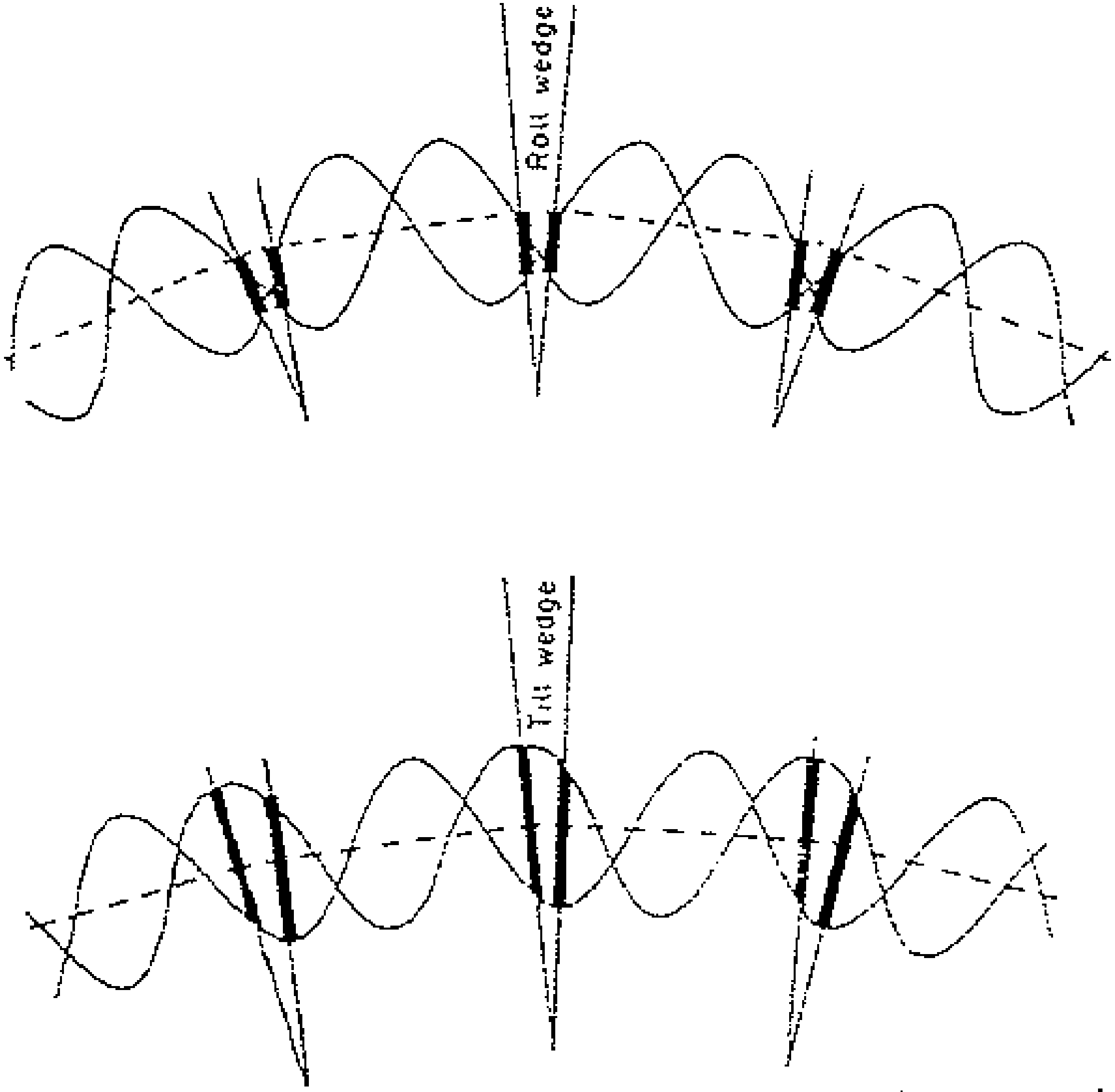


Figure 2. Wedge components of curved DNA (scheme). two interwound strands of double helical DNA molecule are presented by their sinusoidal projections. Only those base-pairs are shown which are non-parallel making the coresponding angles in their in-plane projections (From Ulanovsky and Trifonov, 1987, with permission).

```
aacaagctaagtaccgtactgaagcgcattttaattacgataaggcttatcttaatttcgccgatggcaatgaatgacgtaagcttac
.  .    .             .          .        .           .              .   .       .
0  3    8            21         32        41          53             68  72      80
   0    5            18         29        38          50             65  69      77
        0            13         24        33          45             60  64      72
                      0         11        20          32             47  51      59
                                 0        9          21             36  40      48


                      *          *
    * *  ** * *    * **  *    *  **  * * **   * ** ** *
.........................................................
0         10        20        30        40        50
```

```
aacgaacgatccgcaattaagtcgcgtctggtgcaagggtacttaacagattggaagtaaccgtaactgtcaggaacgtaaggtccat
.    .         .   .               .         .         .   .     .         .    .
0    4        14  18              34        44        54  58    64        74   79
     0        10  14              30        40        50  54    60        70   75
               0   4              20        30        40  44    50        60   65
                   0              16        26        36  40    46        56   61
                                   0        10        20  24    30        40   45

                                        *
                              *         *
    *     *   *     *         *         *   *     *     *
    *     *   * * * *   * *   *   * *   *   ***   *     *
................................................... ......
0         10        20        30        40        50
```

ANGLES DESCRIBING SHAPE OF DNA
(DNA SHAPE CODE)

	Roll	Tilt	Twist
AA	-6.5	3	35.6
AC	(-1)	(-1)	34
AG	8	(0)	28
AT	3		31.5
CA	2	3	34.5
CC	1	2	33.7
CG	7		30
GA	-3	-5	37
GC	-5		40
TA	1		36

Positive Roll opens towards minor groove
Positive Tilt opens towards phosphates

Bolshoy et al., 1991
Kabsch et al., 1982

One of the curviest known DNA is

(GAAAATTTTC)n

P. Hagerman, 1986

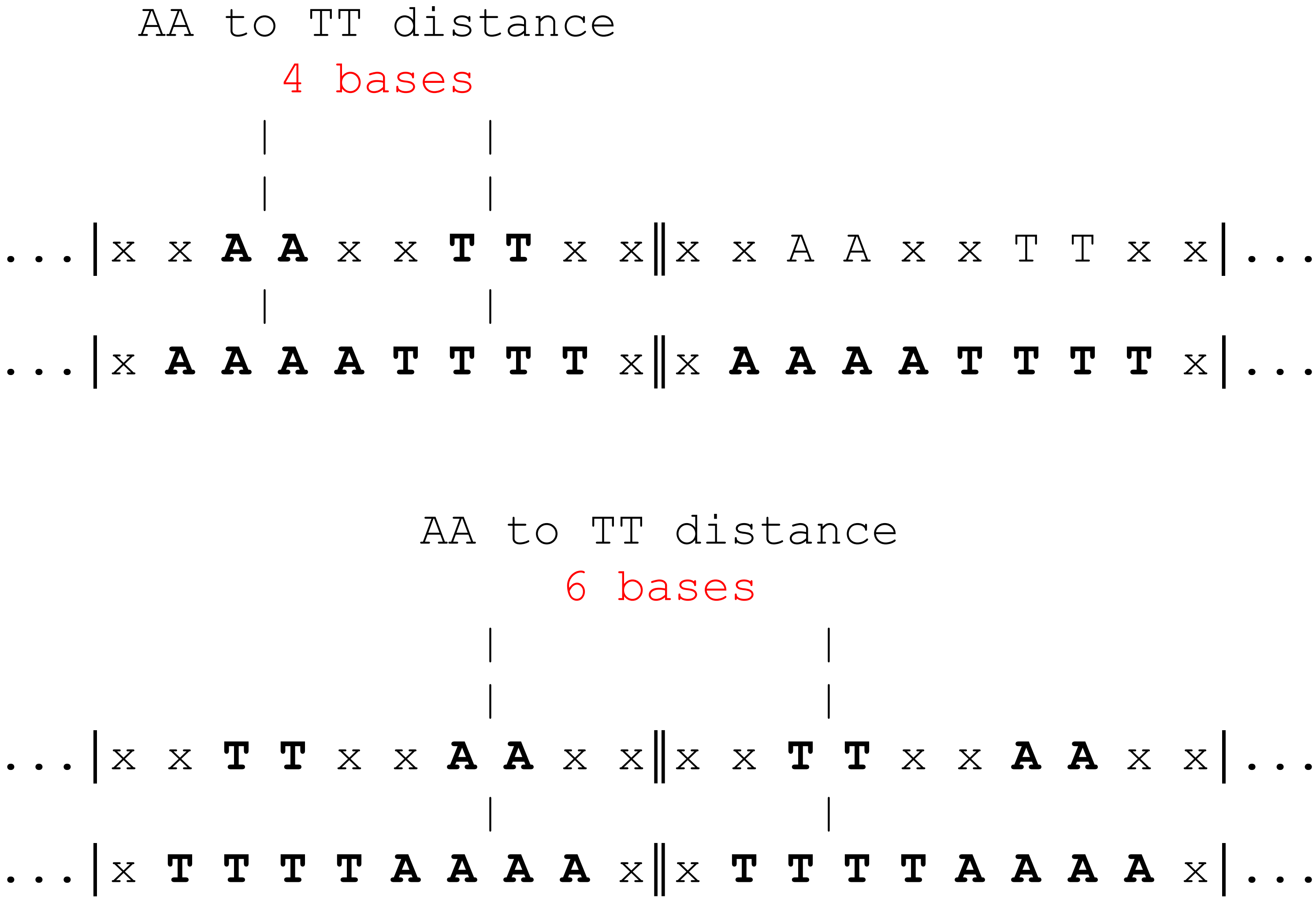
One way to experimentally observe DNA curvature is to watch DNA moving in gel electrophoresis

DNA moves head-on through the narrow pores of the polyacrylamide gel – reptation

The curvature is an obstacle, since the curved molecule keeps deflecting from the along field direction,
and it has to be made straight (force applied) to get through

In the experiments of Hagerman he discovered that repeating GAAAATTTTC behaves in the gel like curved DNA (slow migration)

While repeating GTTTTAAAAC behaves like straight DNA



AA to TT distance

4 bases

...|x x A A x x T T x x‖x x A A x x T T x x|...

...|x A A A A T T T T x‖x A A A A T T T T x|...

AA to TT distance

6 bases

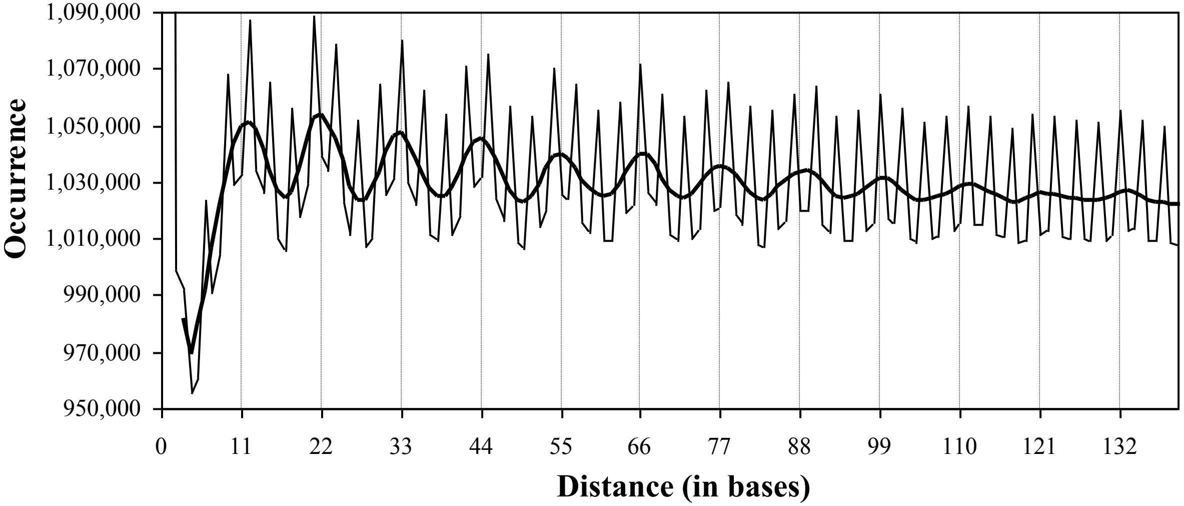
...|x x T T x x A A x x‖x x T T x x A A x x|...

...|x T T T T A A A A x‖x T T T T A A A A x|...

Original calculations on a small sequence ensemble (30 000 bases only)
indicated that the sequence periodicity of 10-11 bases is characteristic
of only eukaryotic sequences

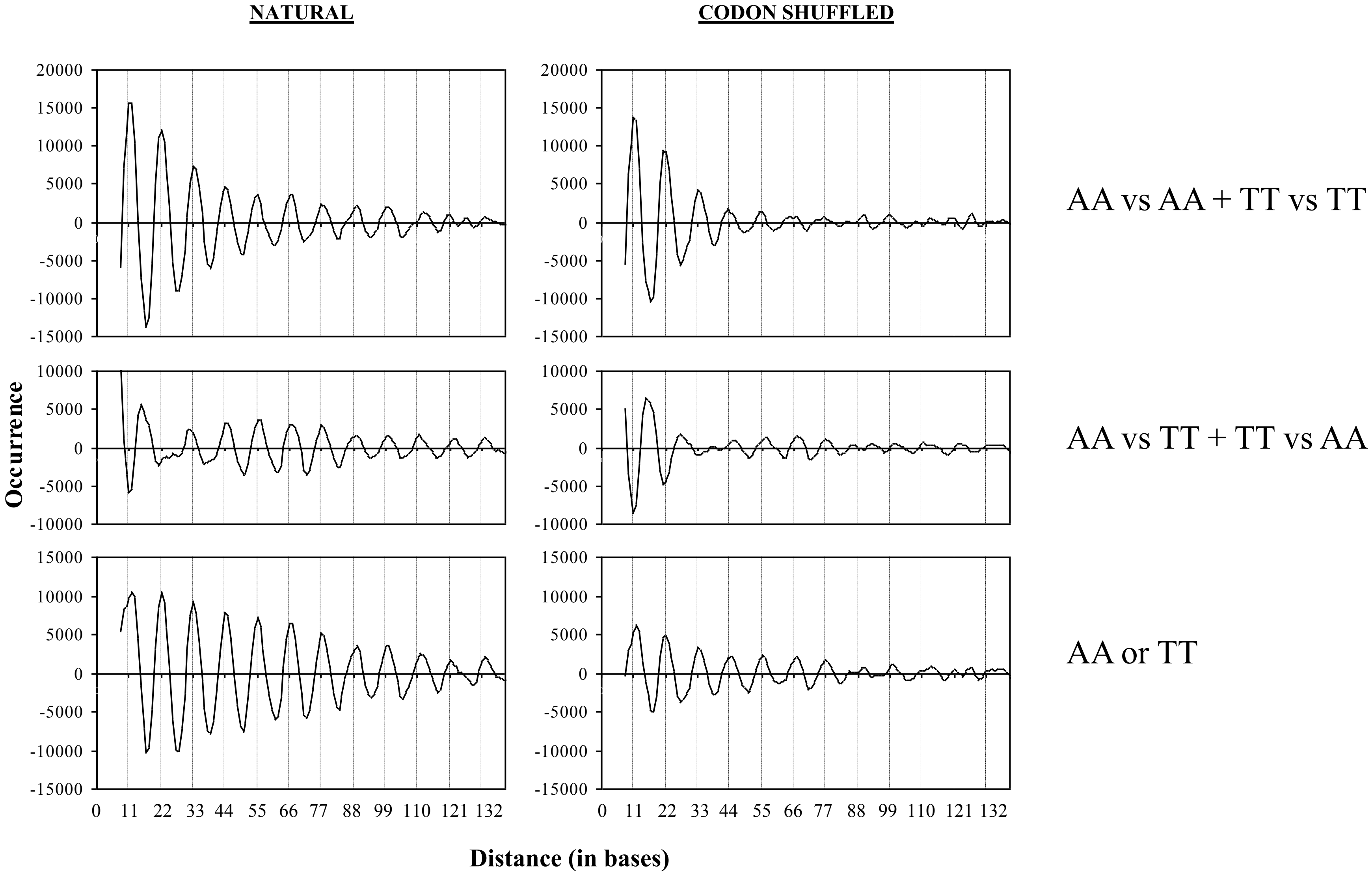
Later on it turned out that prokaryotic genomes are periodical as well,
apparently to maintain DNA superhelicity

In prokaryotes where 85% of genome are protein-coding
the DNA curvature signal (10-11 base period) massively overlaps
with the protein-coding signal (3 base period)

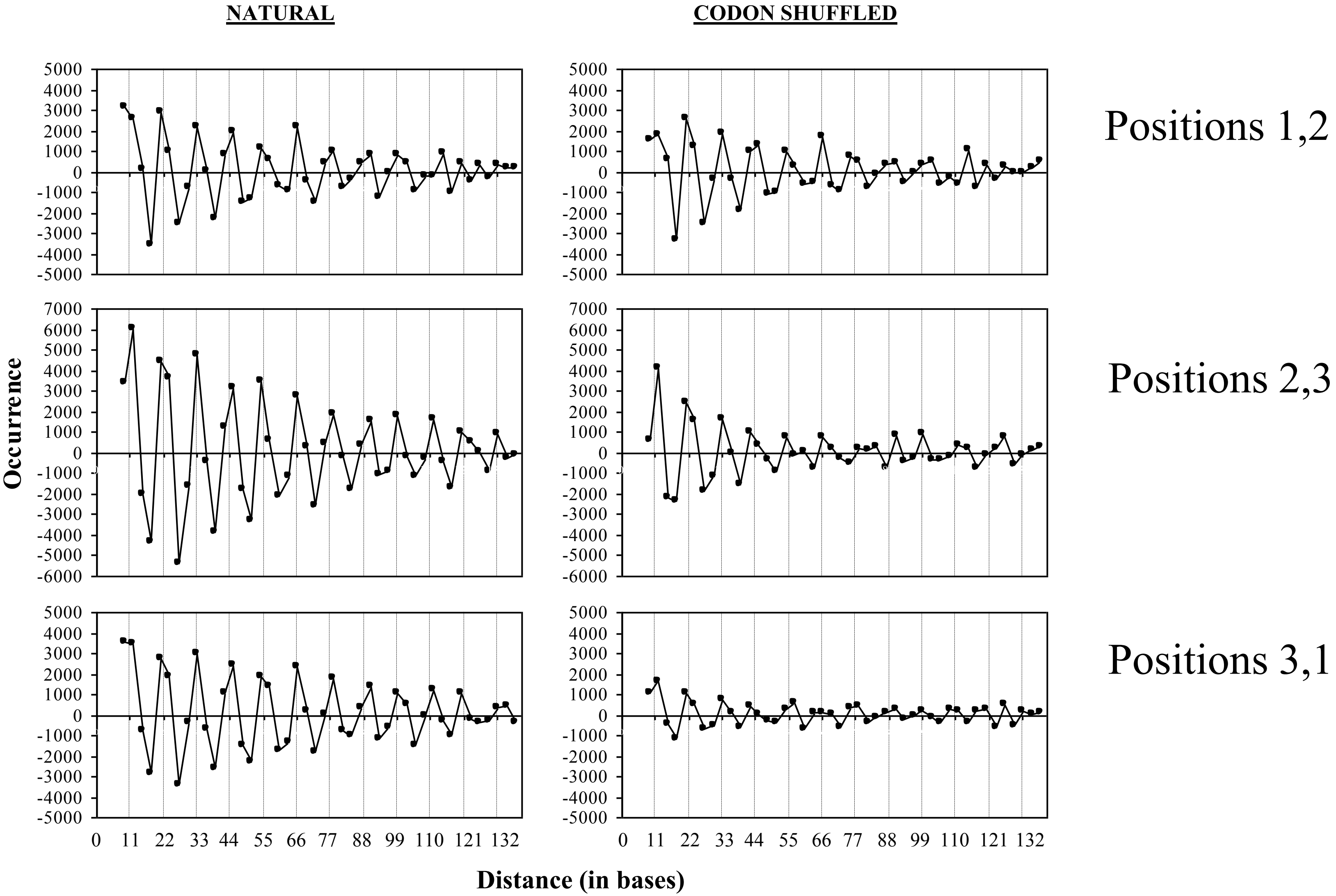


Cohanim, 2006
Eubacteria

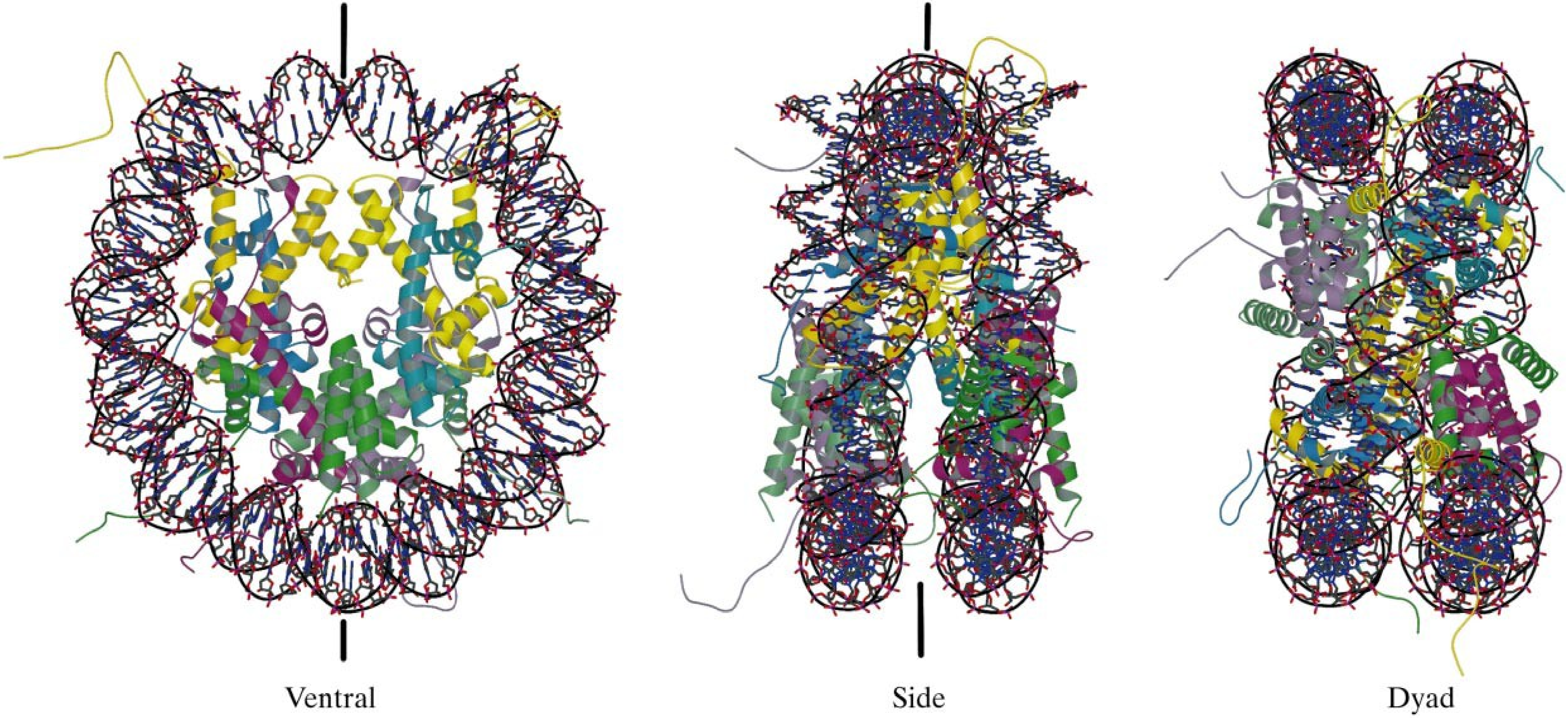
Randomizing third positions brings the oscillations down



The end of the second lecture
(Brno 2011)


NATURAL
CODON SHUFFLED
Positions 1,2
Positions 2,3
Positions 3,1
Occurrence
5000
4000
3000
2000
1000
0
-1000
-2000
-3000
-4000
-5000
7000
6000
-6000
0
11
22
33
44
55
66
77
88
99
110
121
132
Distance (in bases)

CHROMATIN CODE

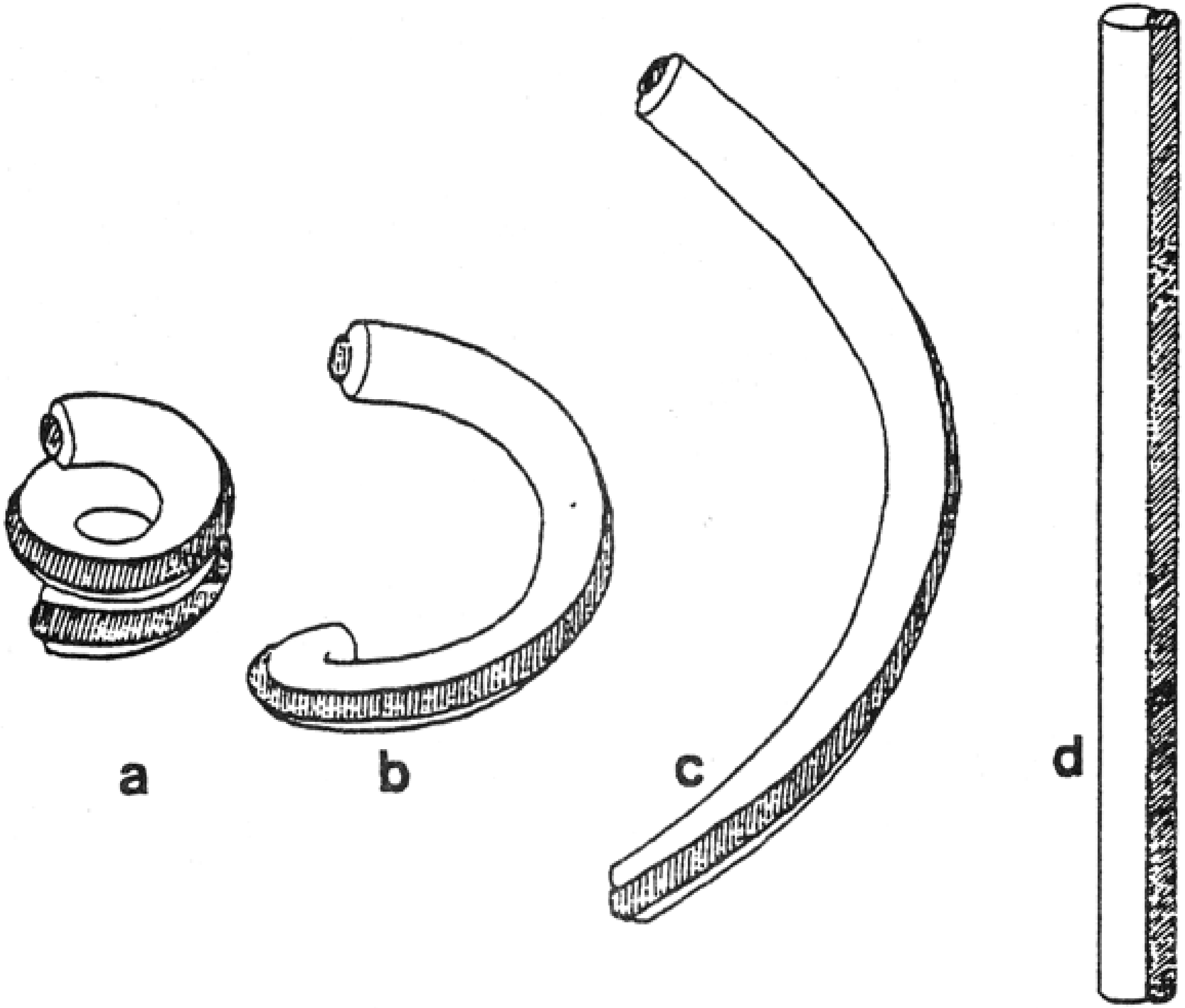


Ventral

Side

Dyad

Lab of G. Bunick, 2000

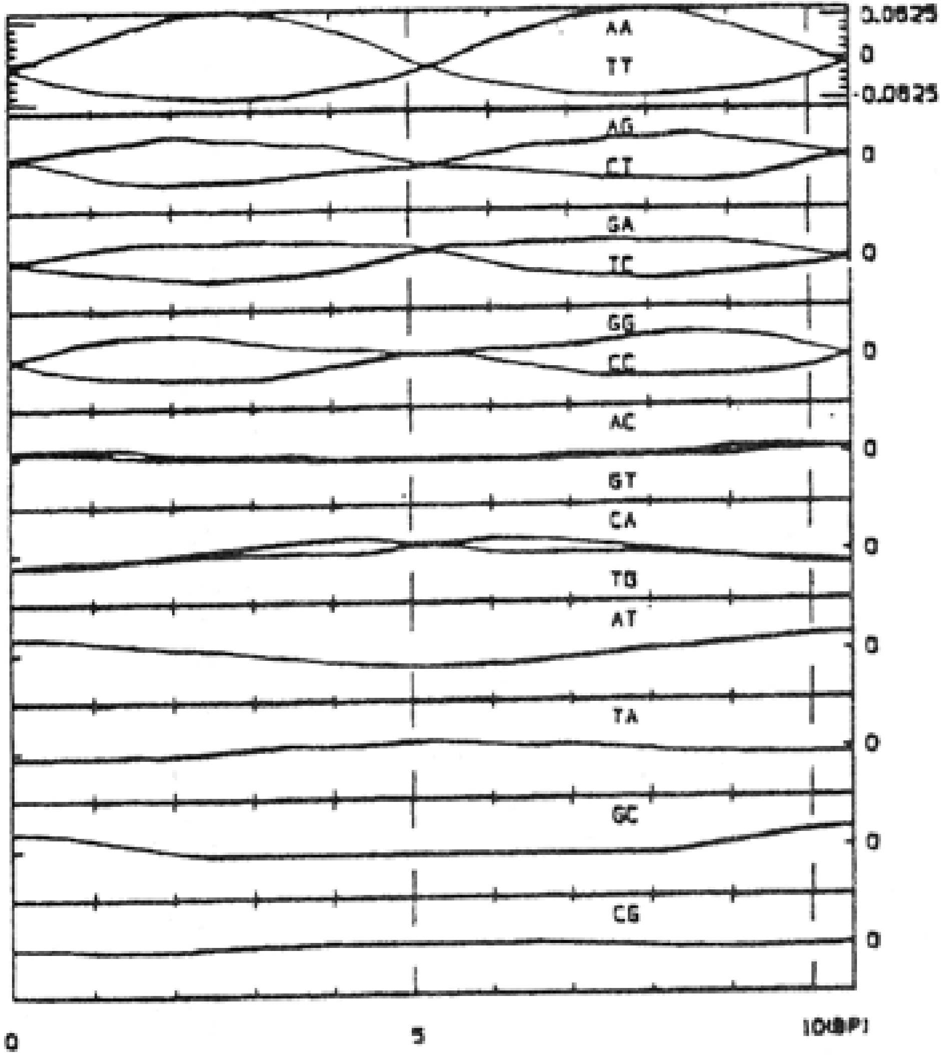

a
b
c
d

Nucleosome core -
particle built
of two side-by-side superhelices
(histones and DNA),
1.5 turns each

It contains ~125 bp of DNA
with structural period 10.4 bp

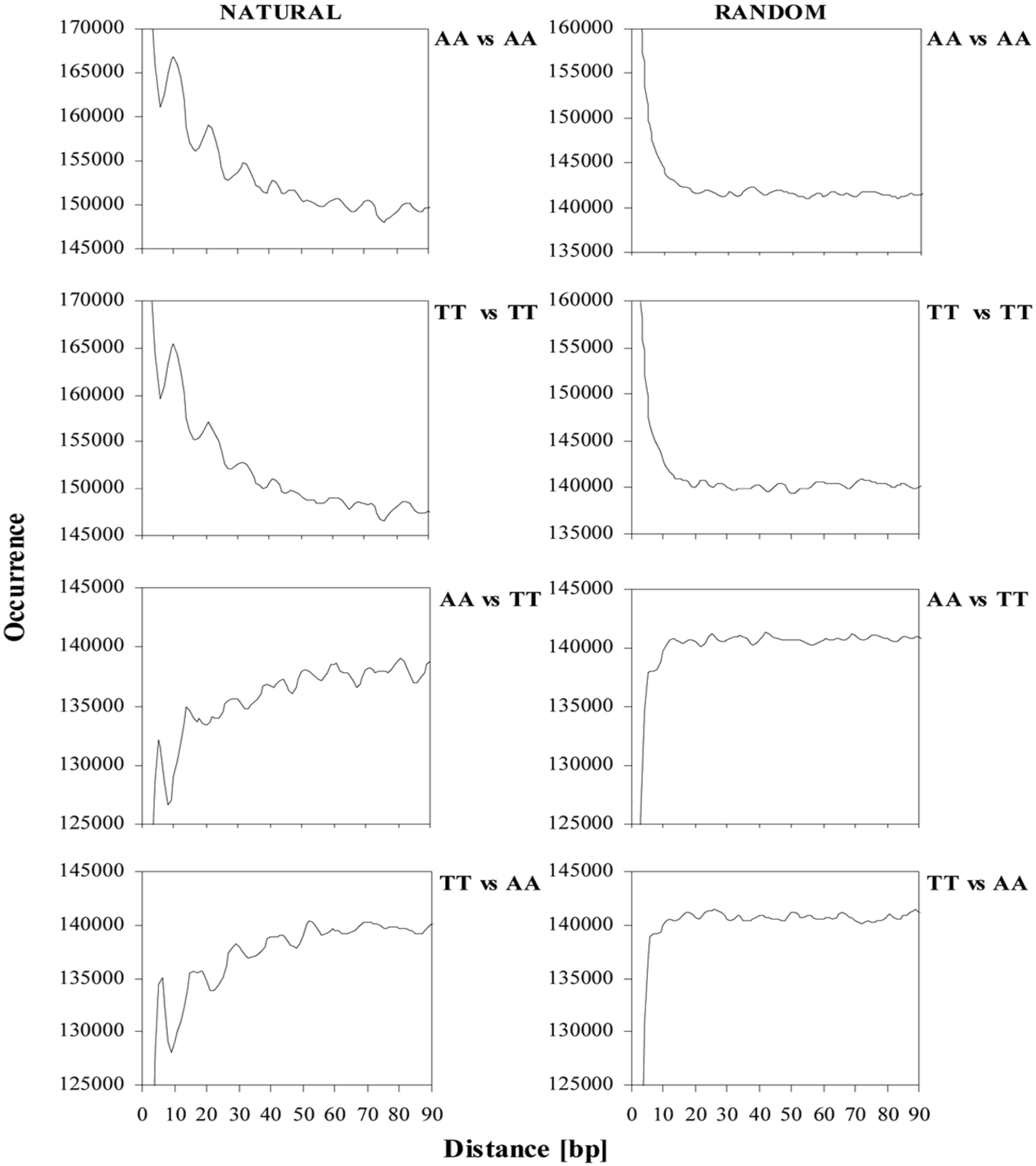
The topologically linear structure
suggests a simple mode
of nucleosome unfolding
during template processes

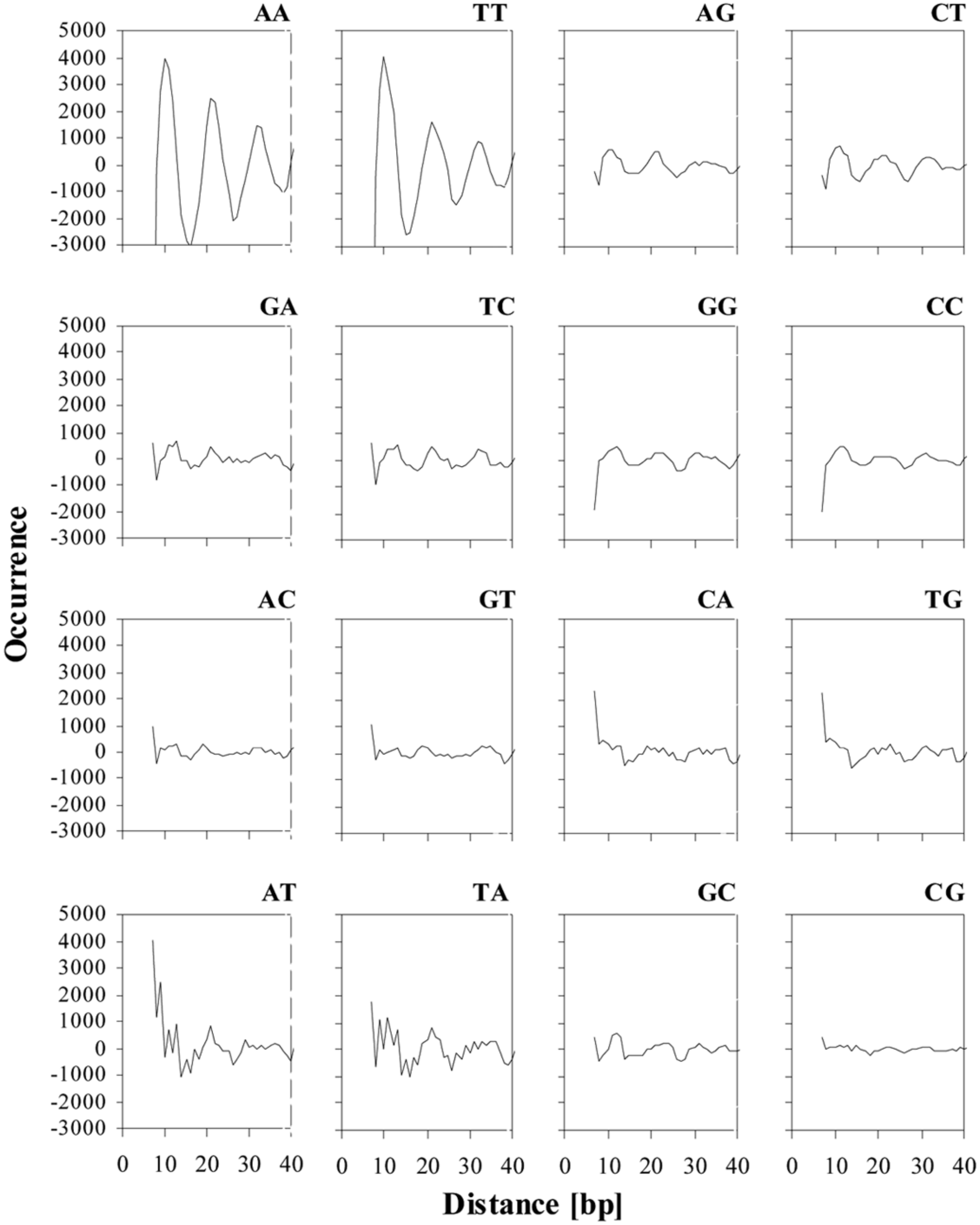
First matrix of nucleosome DNA bendability

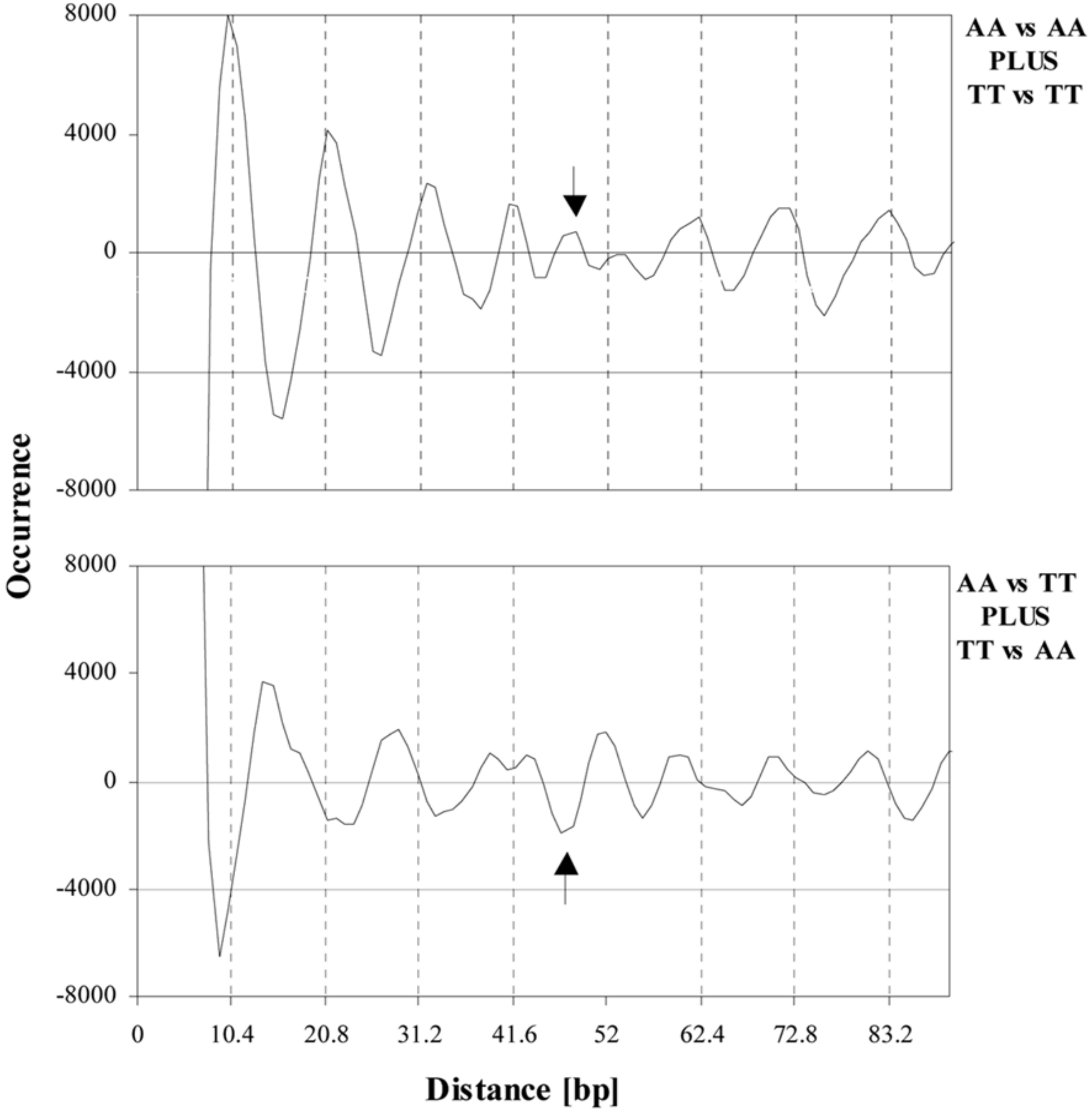


Mengeritsky and ENT, 1983

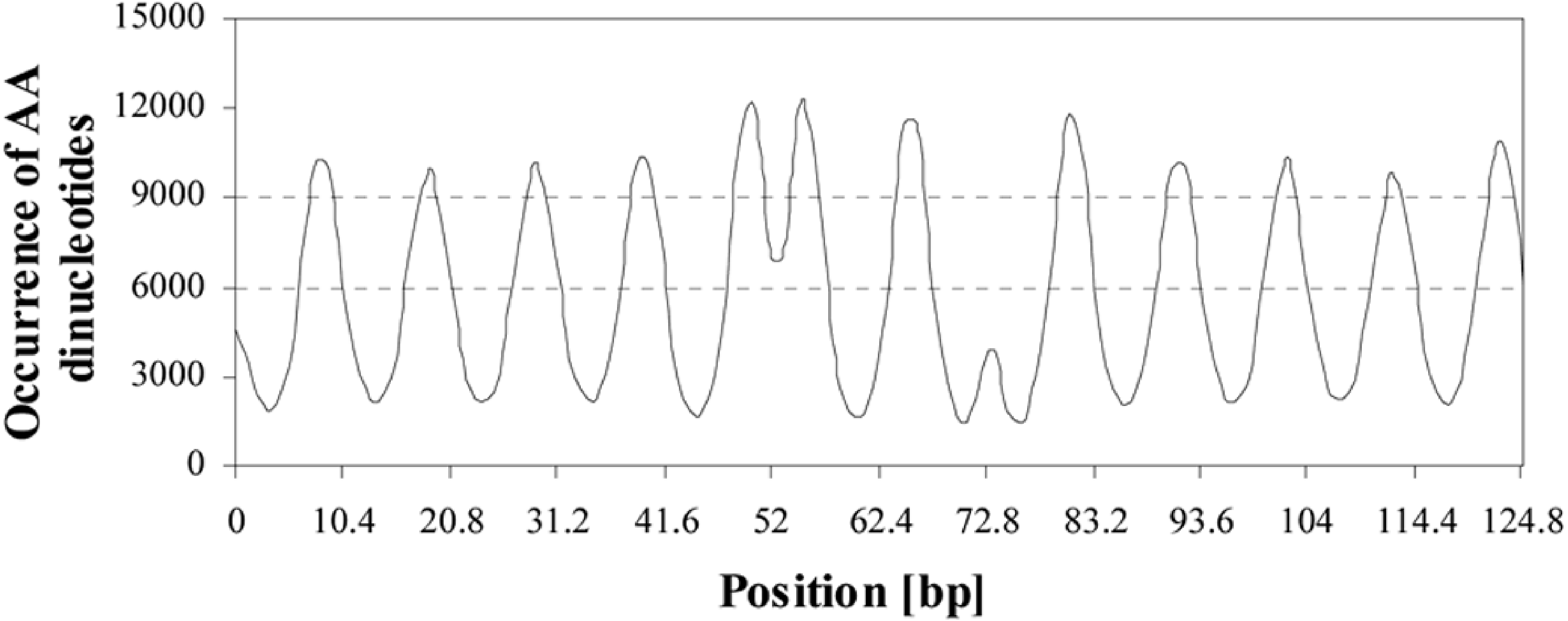
Yeast
Cohanim, 2005




AA
TT
AG
CT
GA
TC
GG
CC
AC
GT
CA
TG
AT
TA
GC
CG
5000
4000
3000
2000
1000
0
-1000
-2000
-3000
Occurrence
0
10
20
30
40
Distance [bp]


8000
4000
0
-4000
-8000
AA vs AA
PLUS
TT vs TT
Occurrence
8000
4000
0
-4000
-8000
AA vs TT
PLUS
TT vs AA
0
10.4
20.8
31.2
41.6
52
62.4
72.8
83.2
Distance [bp]

Calculated nucleosome positioning pattern for yeast genome (Cohanim, 2005)



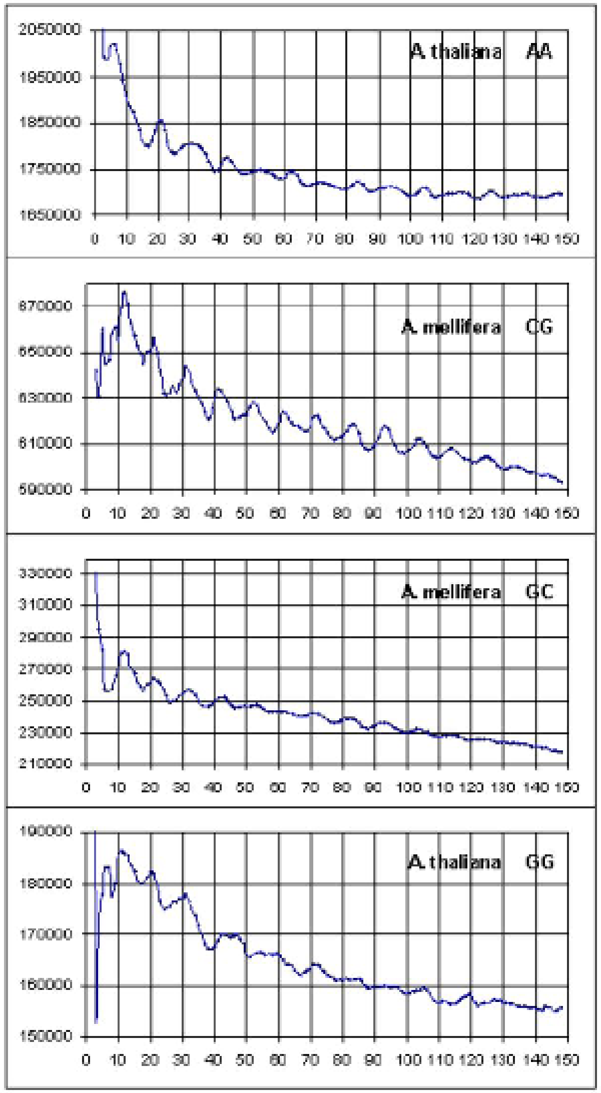

A. thaliana AA
2050000
1950000
1850000
1750000
1650000
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
A. mellifera CG
670000
650000
630000
610000
590000
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
A. mellifera GC
330000
310000
290000
270000
250000
230000
210000
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
A. thaliana GG
190000
180000
170000
160000
150000
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150

Figure 1

History of the chromatin code

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

Pre-genomic studies

...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)
...S S S S x W W W W x S S S S x W W W W...	Satchwell *et al.* (1986)
...x S S S x x W W W x x S S S x x W W W...	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)
...A C G C C T A T A A A C G C C T A T A...	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)

Genome-scale analyses

...T T A A A A A T T T T T A A A A A T T...	Cohanim et al. (2006)
...Y T A R A A A T T T Y T A R A A A T Y...	Salih et al. (2008)
...Y Y R R R R R Y Y Y Y Y R R R R R Y Y...	Salih et al. (2008)
...S S S S x W W W W x S S S S x W W W W...	Chung, Vingron (2009)

Whole-genome nucleosome databases

...C C G G A A A T T T C C G G A A A T T...	Gabdank et al. (2009)

Physics

...C C G G A A A T T T C C G G A A A T T...	Trifonov (2010)

Methods of sequence analysis used for detection of nucleosome pattern(s)

1. Distance analysis (positional correlation)
2. Iteration with random start
3. Multiple alignment
4. Regeneration of the signal from its parts
5. Shannon N-gram extension

Methods that failed:
Fourier transform
Hidden Markov model
Many more failures not publicized

Nucleosome positioning sequence pattern is very weak

(as the nucleosomes should be easy to unfold)

That is why it took so long to crack the code.

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 the signal and describe it in detail

It is easy, however, to detect the signal

Only few properly positioned dinucleotides per nucleosome
are sufficient to claim unique position for the nucleosome

Two good nucleosomes may have completely different sequence.

cacgaaagccacgccggaatc
gcgcggcttgtgtgaatccag

These two sequences
have not a single common base.
But both are very good for nucleosome

ccggaaatttccggaaatttc

The ideal sequence
to which they both match

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	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-
A. gambiae	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C. reinhardtii	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
G. gallus	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-
D. discoideum	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
H. sapiens	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
M. musculus	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

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

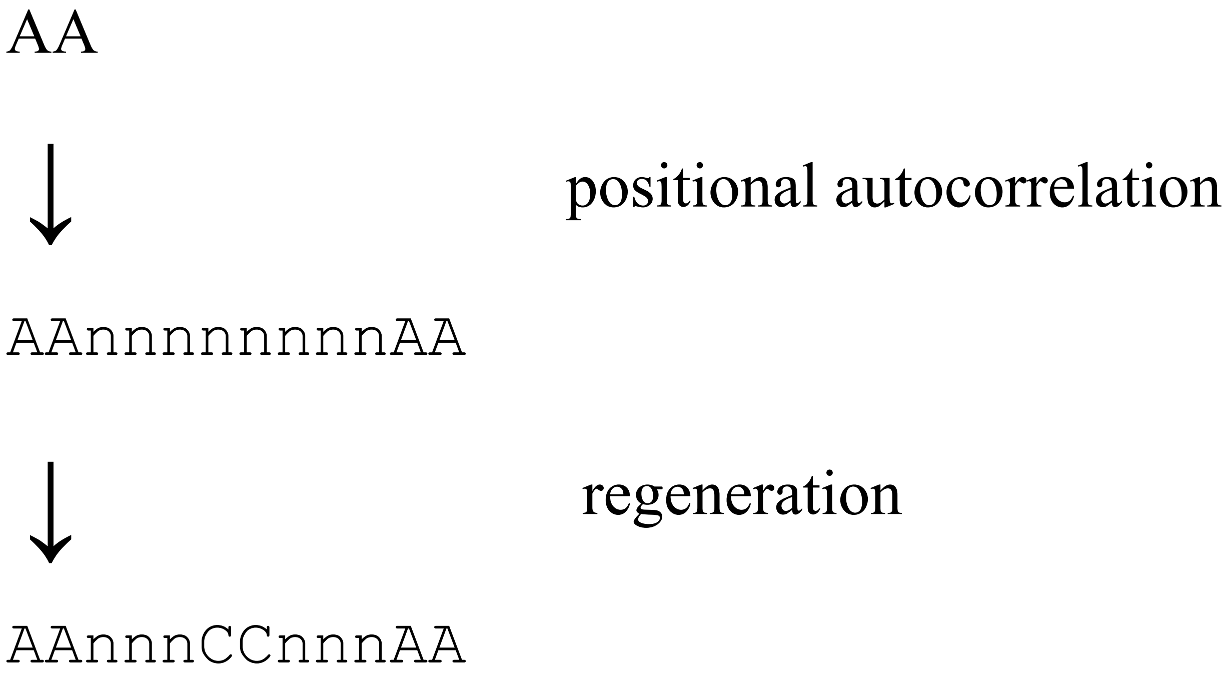
Available databases

of natural nucleosome DNA sequences :

S. Satchwell et al., 1986	115 sequences (chicken)
I. Ioshikhes et al., 1996	~200 sequences (mixture)
M. Kato et al., 2003	~1,300 sequences (human)
S. Johnson et al., 2006	163,651 sequences (*C. elegans*)

Mavrich et al., 2008	$\sim 10^5$ sequences (yeast
Schones et al., 2008	$\sim 10^6$ sequences (H. sapiens)
Mavrich et al., 2008	$\sim\ 10^6$ sequences (fruit fly)

Regeneration of signal from its incomplete versions:



AA

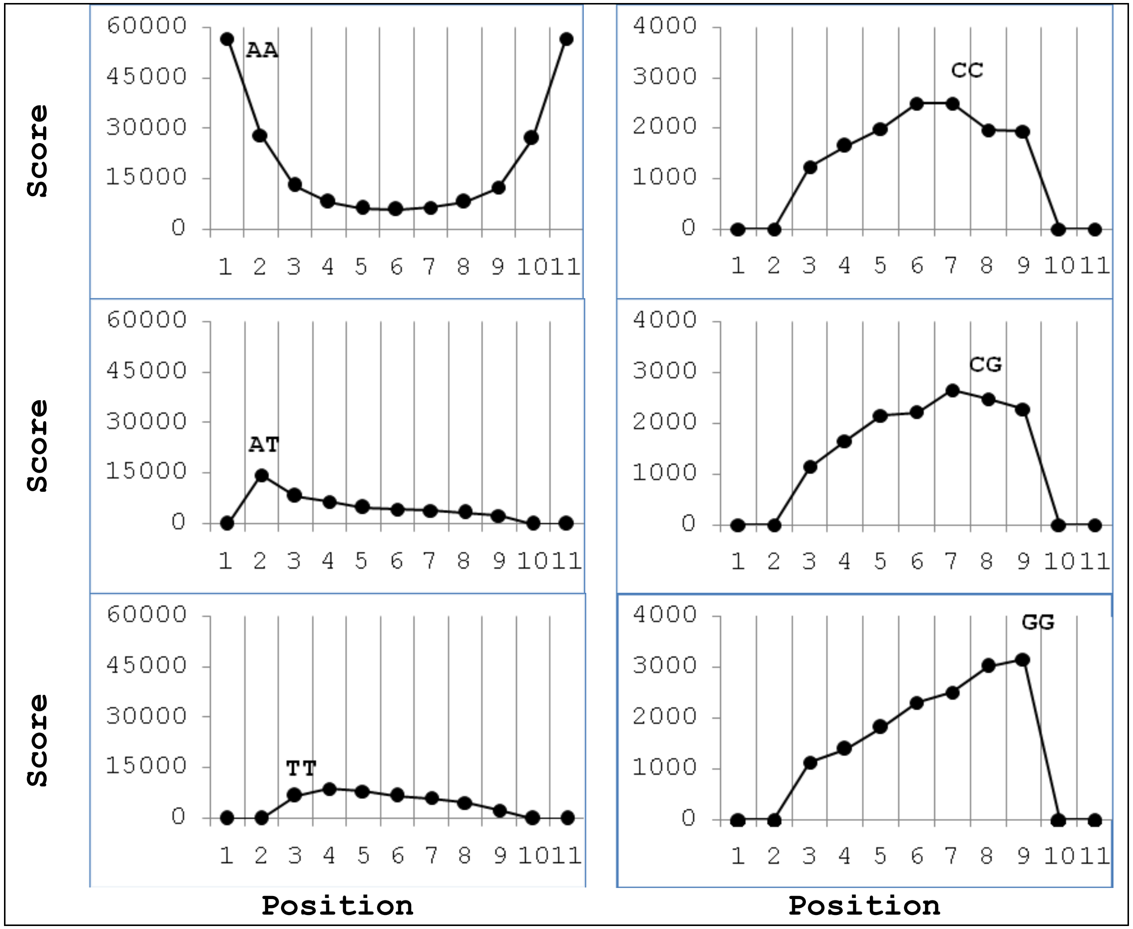
↓ positional autocorrelation

AAnnnnnnnnAA

↓ regeneration

AAnnnCCnnnAA

AAnnnnnnnnAA repeat structure (*C. elegans*)



Regenerated pattern (AAATTTCCGG)(AAAT…

Several reasons for a given dinucleotide to occupy specific position within the repeat:

1. Physical (deformational) preference.

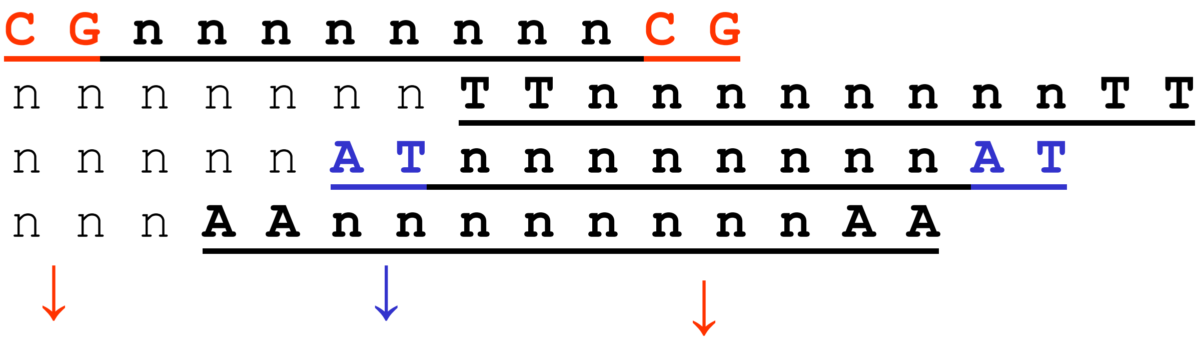
2. Sequence linkage (inclusion effect). Dinucleotide AB has to have neighbors NA and BN.

3. Exclusion effect. Less committed elements are pushed away from strong positions.

4. Compositional bias. Frequent dinucleotides contribute more to the periodicity.

5. Existence of many different codes overlapping on the same sequence (e. g. triplet code, framing code, splicing code, amphipatic helices)

Combination of four matrices:



C G n n n n n n n n C G

n n n n n n n T T n n n n n n n n T T

n n n n n A T n n n n n n n n A T

n n n A A n n n n n n n n A A

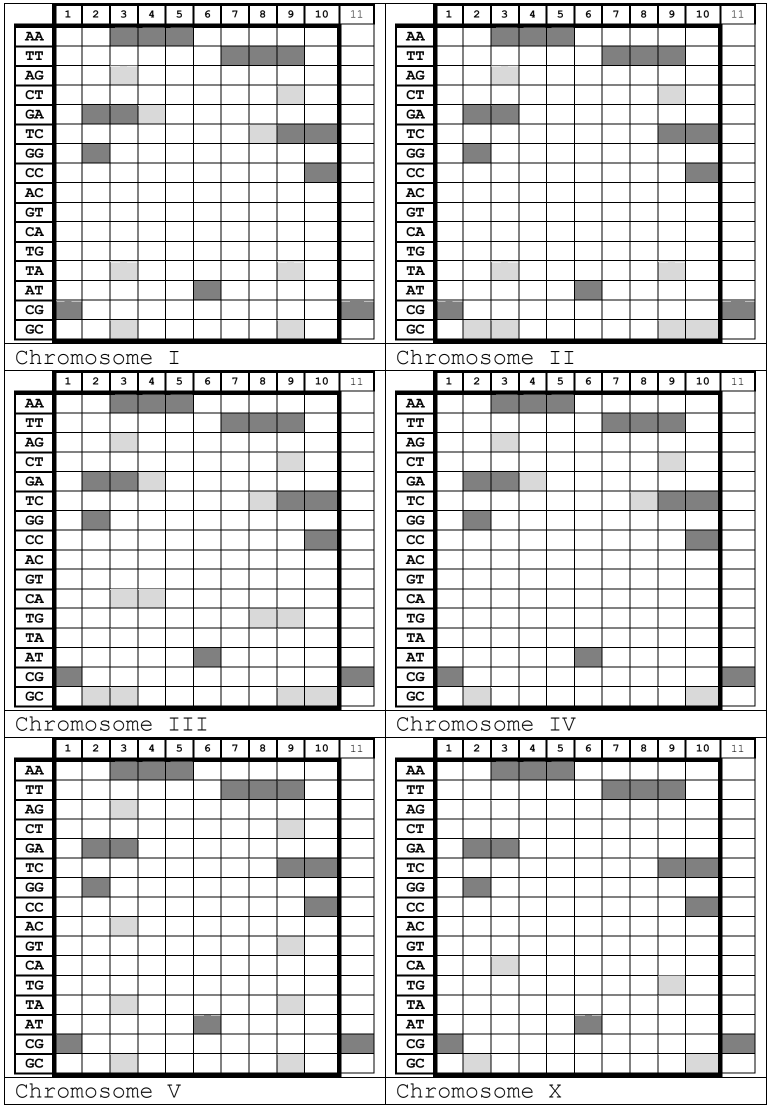
	1	2	3	4	5	6	7	8	9	10	11
AA	2	2	**4**	**4**	**4**	2	2	0	0	0	2
TT	2	0	0	0	2	2	**4**	**4**	**4**	2	2
AG	3	3	**4**	3	2	2	1	1	0	1	3
CT	3	1	0	1	1	2	2	3	**4**	3	3
GA	3	**4**	**4**	**4**	2	0	0	0	1	2	3
TC	3	2	0	0	1	0	2	**4**	**4**	**4**	3
GG	3	**4**	3	1	1	1	0	1	**4**	2	3
CC	3	2	3	1	0	1	1	2	**4**	**4**	3
AC	2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

The matrix turns out to be complementarily symmetrical.

Indeed, symmetrically positioned complementary base-pair stacks should have the same deformations.

Matrices of positional preferences
for six chromosomes
of *C. elegans*

Common symmetrical elements:
AA/TT, GA/TC, GG/CC, AT and CG



	1	2	3	4	5	6	7	8	9	10	11
AA			▓	▓	▓						
TT							▓	▓	▓		
AG			░								
CT									░		
GA		▓	▓	░							
TC								░	▓	▓	
GG		▓									
CC										▓	
AC											
GT											
CA											
TG											
TA			░						░		
AT						▓					
CG	▓										▓
GC			░						░		

Chromosome I

	1	2	3	4	5	6	7	8	9	10	11
AA			▓	▓	▓						
TT							▓	▓	▓		
AG			░								
CT									░		
GA		▓	▓								
TC									▓	▓	
GG		▓									
CC										▓	
AC											
GT											
CA											
TG											
TA			░						░		
AT						▓					
CG	▓										▓
GC		░	░						░	░	

Chromosome II

	1	2	3	4	5	6	7	8	9	10	11
AA			▓	▓	▓						
TT							▓	▓	▓		
AG			░								
CT									░		
GA		▓	▓	░							
TC								░	▓	▓	
GG		▓									
CC										▓	
AC											
GT											
CA			░	░							
TG								░	░		
TA											
AT						▓					
CG	▓										▓
GC		░	░						░	░	

Chromosome III

	1	2	3	4	5	6	7	8	9	10	11
AA			▓	▓	▓						
TT							▓	▓	▓		
AG			░								
CT									░		
GA		▓	▓	░							
TC								░	▓	▓	
GG		▓									
CC										▓	
AC											
GT											
CA											
TG											
TA											
AT						▓					
CG	▓										▓
GC		░								░	

Chromosome IV

	1	2	3	4	5	6	7	8	9	10	11
AA			▓	▓	▓						
TT							▓	▓	▓		
AG			░								
CT									░		
GA		▓	▓								
TC									▓	▓	
GG		▓									
CC										▓	
AC			░								
GT									░		
CA											
TG											
TA			░						░		
AT						▓					
CG	▓										▓
GC			░						░		

Chromosome V

	1	2	3	4	5	6	7	8	9	10	11
AA			▓	▓	▓						
TT							▓	▓	▓		
AG											
CT											
GA		▓	▓								
TC									▓	▓	
GG		▓									
CC										▓	
AC											
GT											
CA			░								
TG									░		
TA											
AT						▓					
CG	▓										▓
GC		░								░	

Chromosome X

Positional matrix of bendability

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

Same in simplified forms:

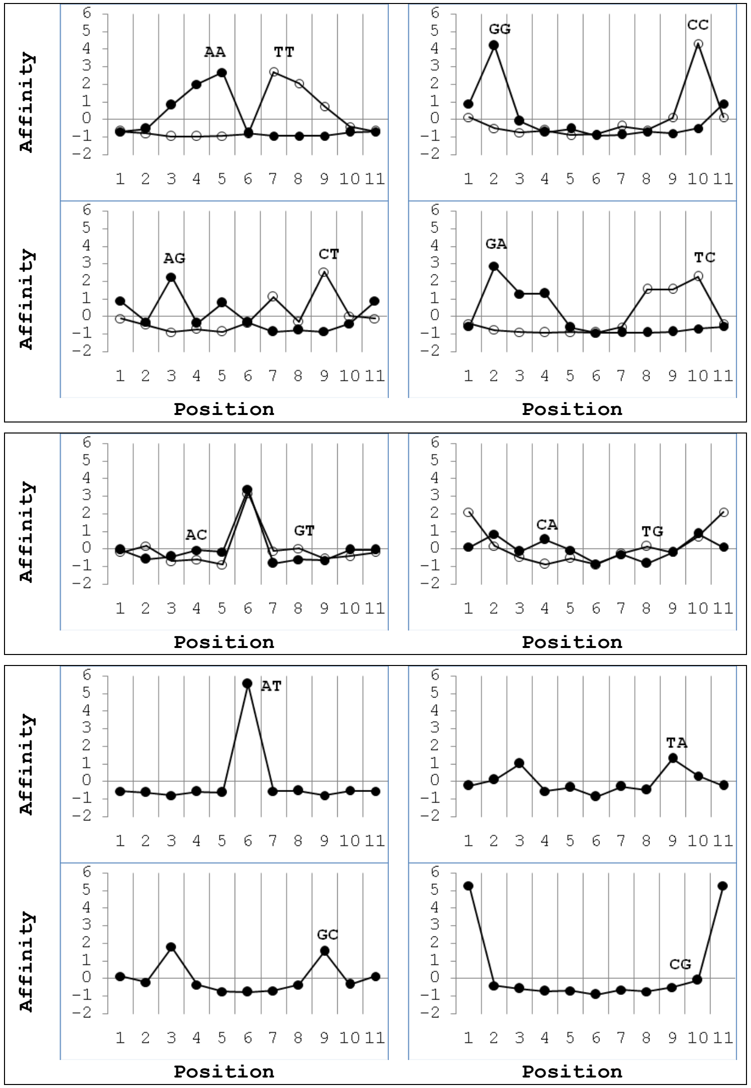
```
  ▼                   ▼                   ▼
  1   2   3   4   5   6   7   8   9  10  11

 CG  GG                              CC  CG
     GA  GA                      TC  TC
         AA  AA  AA  AT  TT  TT  TT
---------------------------------------------
  ▼                   ▼                   ▼
C   G   G   A   A   A   T   T   T   C   C   G    -  one-line form
  ▼                   ▼                   ▼
Y   R   R   R   R   R   Y   Y   Y   Y   Y   R    -  [R,Y] form

---------------------------------------------

x   x   R   R   R   x   x   Y   Y   Y   x   x    -  matrix of bendability,
---------------------------------------------        Mengeritsky, 1983

Y   R   x   x   x   R   Y   x   x   x   Y   R    -  YR/RY form,
                                                     Zhurkin, 1983
```

LINEAR FORM OF
THE POSITIONAL MATRIX OF BENDABILITY:

CGRAAATTTYCG

Matrix of bendability

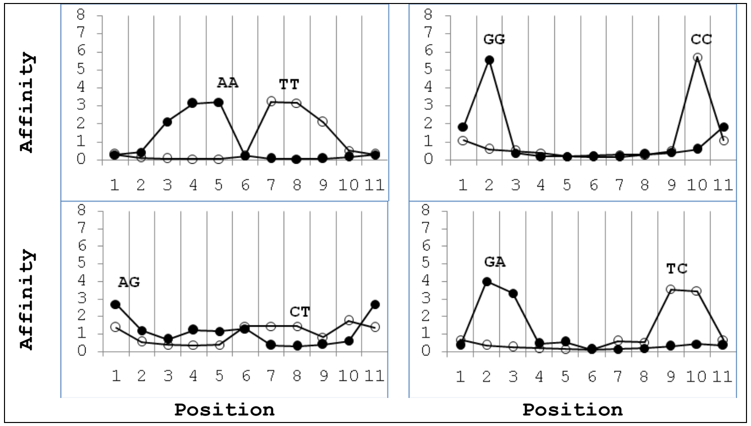
for Chromosome I

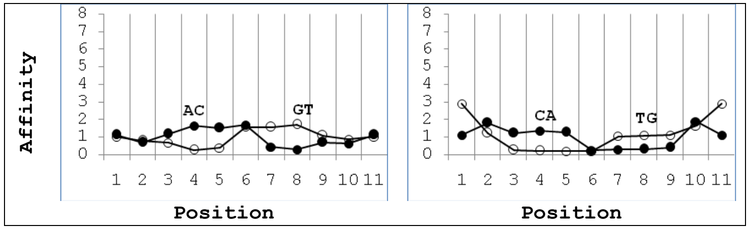


(no symmetrization applied)

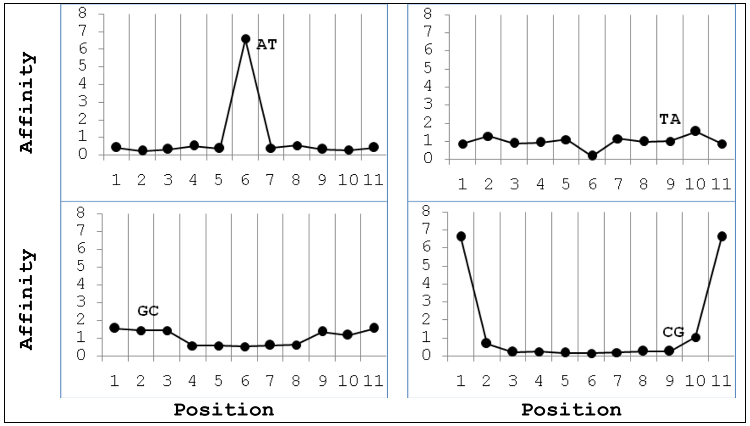
Matrix of bendability

for all 6 chromosomes of *C. elegans*





Self-complementary elements AT and CG are separated by 5 bases (half-period) and positioned at the axes of complementary symmetry



NUCLEOSOME DNA PATTERNS IN 2-LETTER ALPHABETS

R = A, G Y = C, T

```
           |         |         |
. . . Y Y Y R R R R R Y Y Y Y Y R R R . . .
```

E. Trifonov, J. Sussman, 1980
G. Mengeritsky, E. Trifonov, 1983

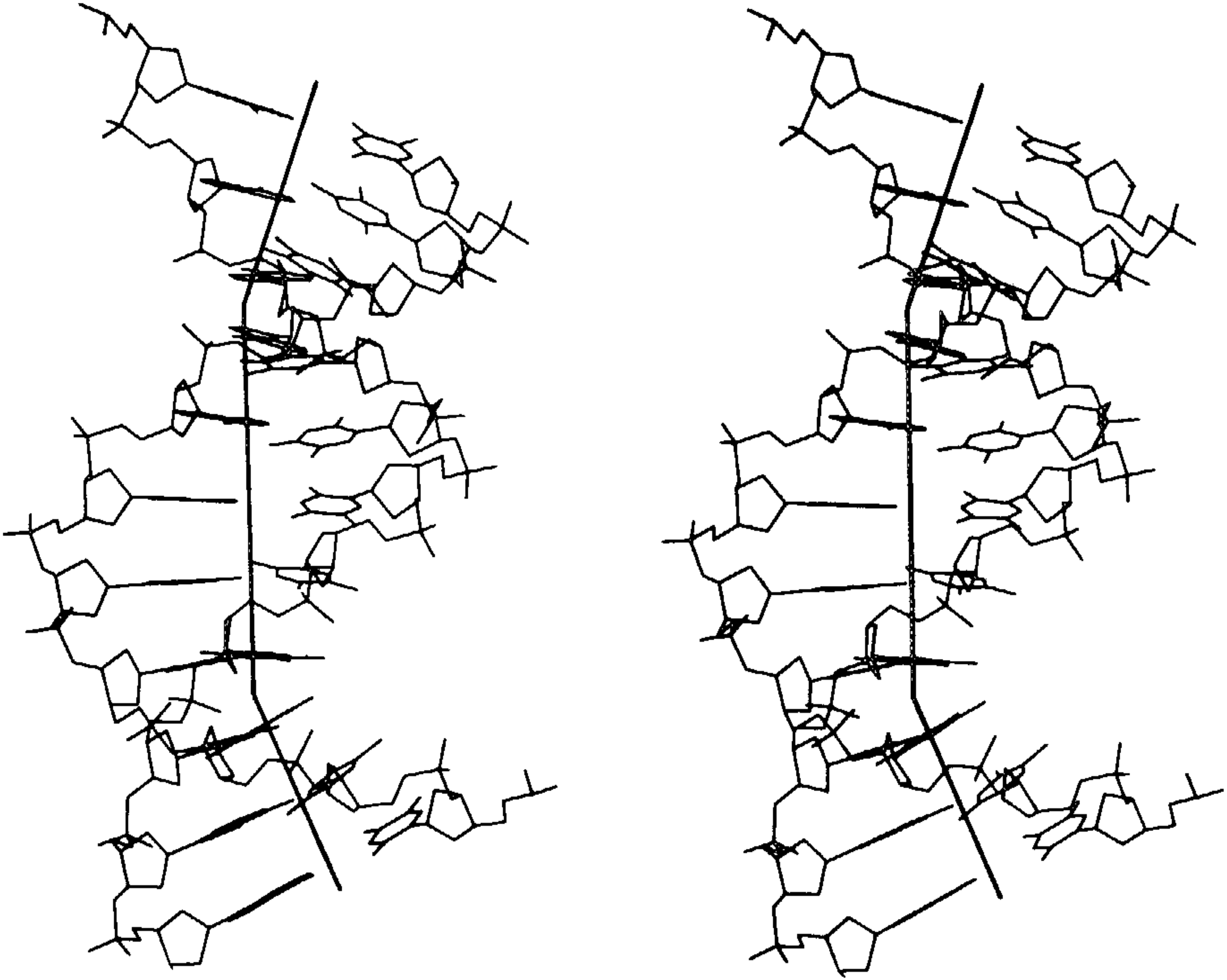
V. Zhurkin, 1983

F. Salih et al, 2007, 2008

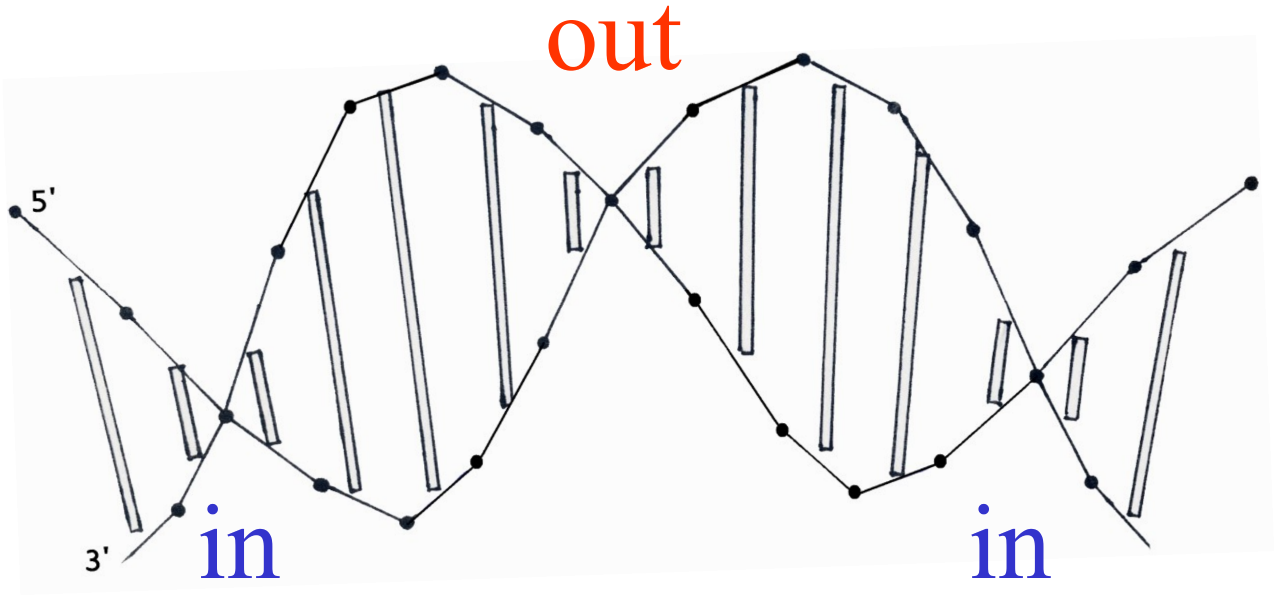
S = G, C W = A, T

```
           |         |         |
. . . S S S W W W W W S S S S S W W W . . .
```

S. Satchwell et al, 1986
H. Chung, M. Vingron, 2009



Ulyanov and Zhurkin, JBSD, 1984



Mere physics

SSSS WWWW SSSS ←

weak base pair stacks
should be OUT,
as they are easier
to deform (unstack).

YR RY YR ←

R stacks are on the surface,
i. e. IN (Zhurkin, 2010)

Y RRR YYY R ←

rines, with stronger stacking
between them,
should be on the surface

CCGGRAATTYCCGG ←

a unique merger
of the binary patterns

CCGGAAATTTCCGG ←

A+T rich genomes

Sequence analysis:	CGRAAATTTYCG
Physics:	CGGAAATTTCCG

```
R Y Y Y Y Y R R R R R Y Y Y Y Y R R R R R Y
 |         |         |         |         |
A|T T T T T|A A A A A|T T T T T|A A A A A|T
 |         |         |         |         |
 |        T|G        |        T|G        |      isochores L1
A|T T T T  |  A A A A|T T T T  |  A A A A|T
 |        C|A        |        C|A        |
 |         |         |         |         |
A|T T T T C|G A A A A|T T T T C|G A A A A|T
A|T T T C C|G G A A A|T T T C C|G G A A A|T     most
A|T T C C C|G G G A A|T T C C C|G G G A A|T     frequent
A|T C C C C|G G G G A|T C C C C|G G G G A|T     patterns
 |         |         |         |         |
A|C        |        A|C        |        A|C
 |  C C C C|G G G G  |  C C C C|G G G G  |
G|T        |        G|T        |        G|T     isochores H3
 |         |         |         |         |
G|C C C C C|G G G G G|C C C C C|G G G G G|C
```

10.4 base periodical contributions of SS and WW dinucleotides in various species

	Human	Mouse	Arabidopsis	C. elegans
SS	0.312	0.286	0.099	~0
WW	~0	0.050	0.092	0.185

S. Kogan, 2005

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
....		....

Shannon N-gram extension

```
                     AAA
                    AAA        A. Rapoport,
                      AAT      Z. Frenkel,
                   GAA ATT     E.N.T., 2010
                  TGA   TTT
                 TTG     TTT
                TTT       TTC
               TTT         TCA
              ATT           CAA
             AAT             AAA
            AAA               AAA
           AAA                 AAT
          GAA                   ATT
         TGA                     TTT
        TTG                       TTT
       TTT                         TTC
      TTT                           TCA
   ...TTTTGAAAATTTTGAAAATTTTCAAAATTTTCA...

   ...AAA... : TTTtgAAAATTTTcaAAA
   ...CGA... : TTTcgAAAATTTTcgAAA
regeneration : TTYCGRAAATTTYCGRAA
```

TOPMOST TRINUCLEOTIDES MAKE TOGETHER THE DOMINANT PATTERN

GAAAATTTTC:

GAAAATTTTC
GAAAATTTTC
GAAAATTTTC
GAAAATTTTC
GAAAATTTTC
GAAAATTTTC
GAAAATTTTC
GAAAATTTTC

extention motifs	species	starting triplets
C AAAAA TTTTT G	A.gamb	TTT
T AAAAA TTTTT A	A.mell	TTT
AAAAA TTTTT	A.thali	AAA
TTTTC AAAAA TTTTT GAAAA	C.albic	AAA
GAAAA TTTTC	C.eleg	AAA
GG CC	C.reinh	GGC
AAAAA TTTTT	D.disc	AAA
C AAAAA TTTTT G	D.melan	AAA
AAAAA TTTTT	D.rerio	AAA
C AGAAA TTTCT G	G.gall	TTT
AAAAA TTTTT	H.sapi	TTT
GAAAA TTTTC	M.musc	TTT
GAAAA TTTTC	S.cerev	AAA

Fig. 3. N-gram Shannon extensions of the most frequent trinucleotides of various genomes, as indicated. Only the central parts of the extensions (underlined) are shown.

extention motifs	species	starting triplets
C AAAAA TTTTC GAAAA TTTTT G	A.gamb	TCG
AAAAA TTTTC GAAAA TTTTT	A.mell	CGA
AAAAA TTTTC GAAAA TTTTT	A.thali	TCG
AAAAA TTTTC GAAAA TTTTT	C.albic	TCG
GAAAA TTTTC GAAAA TTTTC	C.eleg	CGA
AAAAA TTTTC GAAAA TTTTT	D.disc	TCG
GC AAAAA TTTTC GAAAA TTTTT GC	D.melan	TCG
AAAAA TTTCC GGAAA TTTTT	H.sapi	CGG
GAAAA TTTTC GAAAA TTTTC	S.cerev	CGA
GGC GCC	C.reinh	CGC
TTTT AAAAC GTTTT AAAA	D.rerio	ACG
A GAAAC GTTTC T	G.gall	CGT
AC GT	M.musc	CGT

Fig. 4. Extensions of the topmost CG-containing trinucleotides of various genomes, as indicated. Only the central parts of the extensions (underlined) are shown. Four genomes with extensions that do not conform to others, are separated.

Rapoport et al., 2010

The end of the third lecture (Brno 2011)

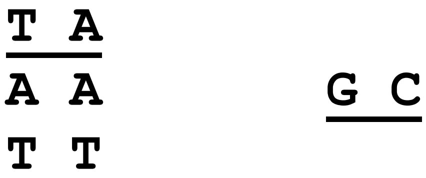
CHROMATIN CODE:

CGRAAATTTYCG

It is derived by 3 independent methods:

1. From physics of DNA deformation
2. From nucleosome database of C. elegans
3. By Shannon N-gram extension

TA/GC pattern (Segal/Widom)



T A
A A
T T

G C

at 5 bases distance

The pattern **TA/GC** is derived from SELEX experiments (*artificial sequences*)

CG/AT pattern is derived from *natural ones* (nematode, confirmed in other eukaryotes)

TA*TA stack is of the lowest stacking energy.
In symmetrical groove positions it would readily kink.
That would create mutational hot spot.

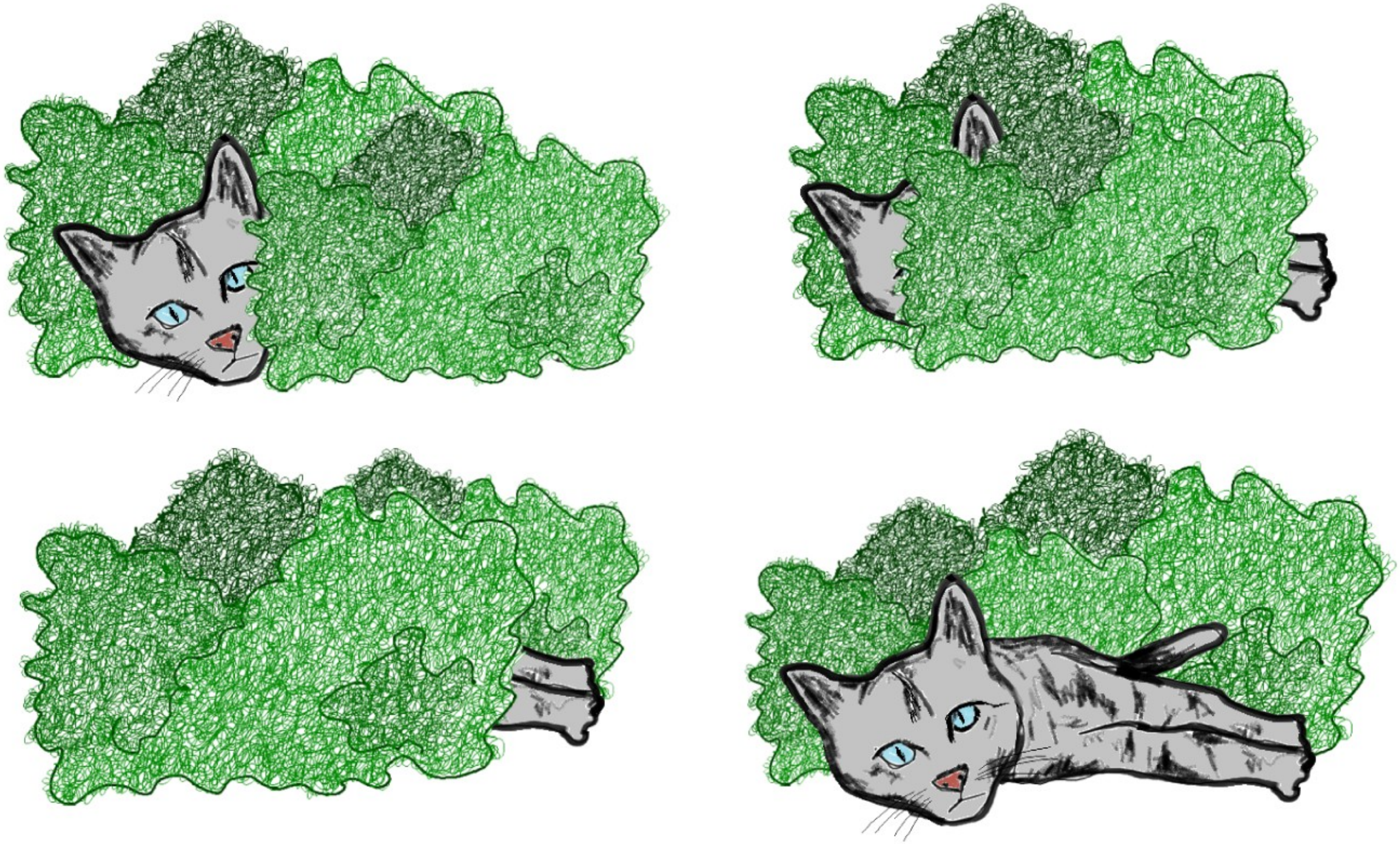
The hidden chromatin code is described by the motif:

CGRAAATTTYCG

○ ○ ○

An ideal nucleosome DNA in simple sequence form
is periodical repetition of this motif:

CGRAAATTTYCGRAAATTTYCGRAAATTTYCGRAAATTTYCGRAAATTTYCGRAAATTTYCGRAAATTTYCGRAAATTTYCGRAAATTTYCGRAAATTTYCGRAAATTTYCG



Cat in bushes. Courtesy of I. Gabdank

...TTTCCGGAAATTTCCGGAAA...

...ATTCGTTCCATTGAAGGCCG...

...CGAACGCTTGGTTAGCGATT...

...CCAGAATAAATACAGTCCAA...

...AATCGCCTTTAAAGGGGTTT...

...GAGTTCGACTCCAATCAGGG...

...CGGTACCCTCAGACCCATTC...

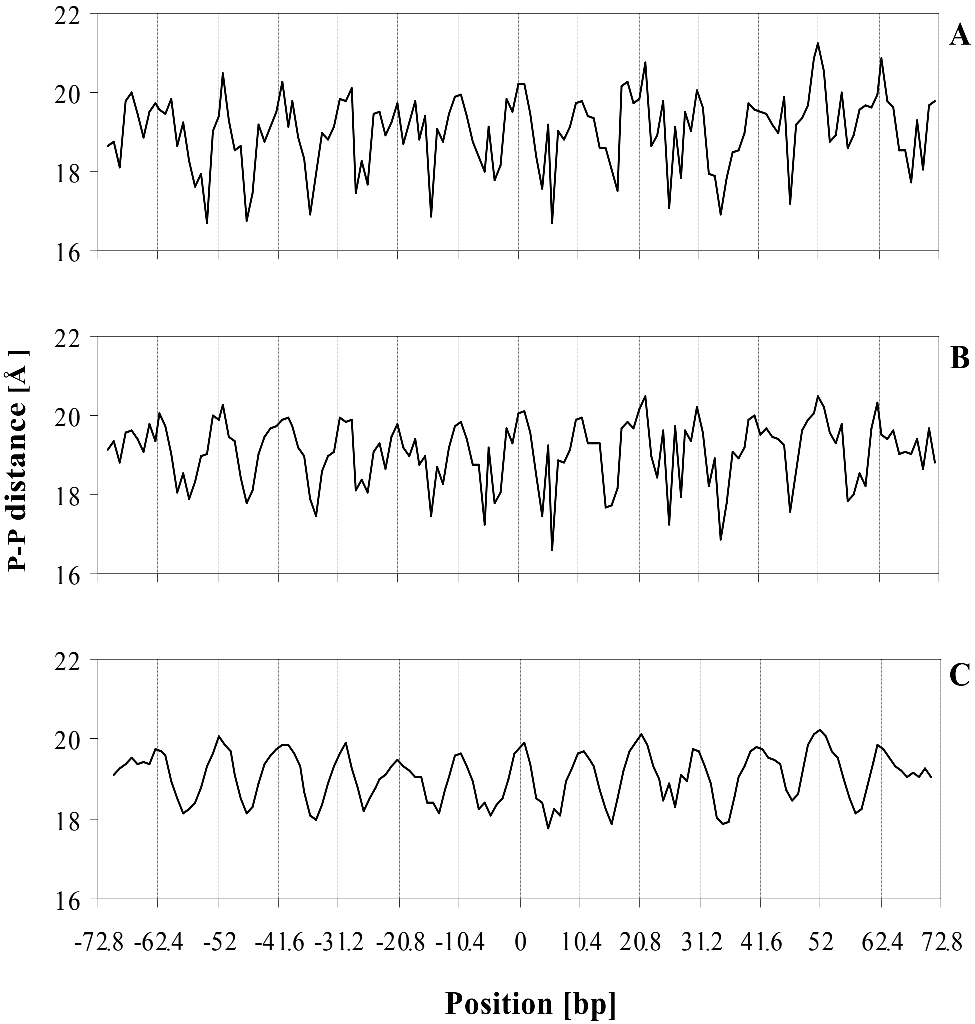
...CATCTATTCCAAATTTTCGC...

The nucleosome DNA structural period is between 10.333 and 10.400

pitch of DNA						local dyads							
(base pairs)	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII
10.000-10.100	+	+										+	+
10.100-10.125		+	+								+	+	
10.125-10.167			+	+						+	+		
10.167-10.222				+	+				+	+			
10.222-10.273	+				+				+				+
10.273-10.333		+			+				+			+	
10.333-10.400													
10.400-10.444	+					+		+					+
10.444-10.556				+		+		+		+			
10.556-10.600	+					+		+					+
10.600-10.667													
10.667-10.727		+			+				+			+	
10.727-10.778	+				+				+				+
10.778-10.833				+	+				+	+			
10.833-10.875			+	+						+	+		
10.875-10.900		+	+								+	+	
10.900-11.000	+	+										+	+

Noninteger Pitch and Nuclease Sensitivity of Chromatin DNA
Edward N. Trifonov and Thomas Bettecken, Biochemistry, 1979

Nucleosome crystal data reveal the 10.4-base structural period of the nucleosome DNA (A. Cohanim et al., 2006)



1KX5
(C. Davey et al., 2002)

1AOI+1KX4
(K. Luger et al. 1997)
+1KX5

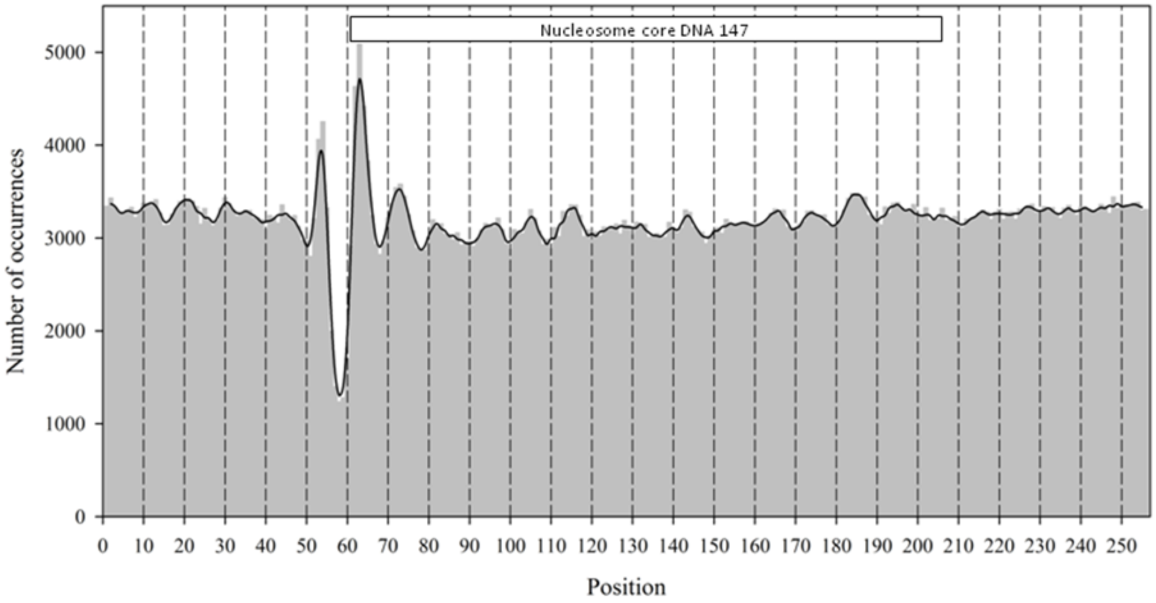
Same,
smoothed

There are 12 contact sites of the minor grooves with the histones – 12 positions for CG.

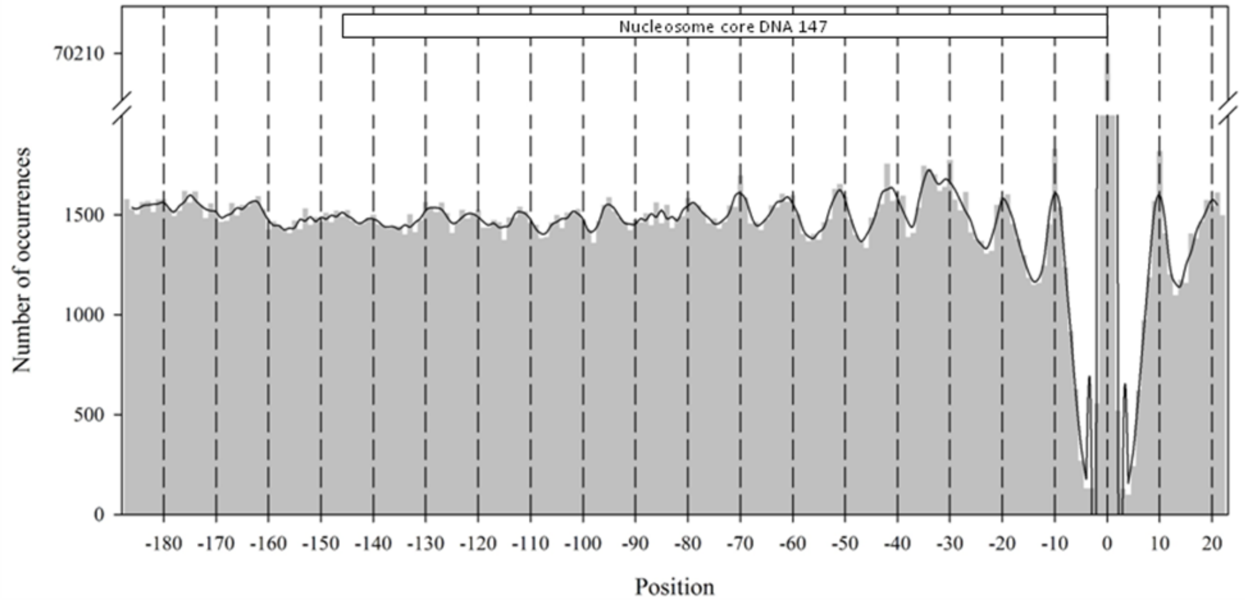
Total length of the DNA in contact with histone octamers is 10.4x11+1 = 115 bp

Micrococcal nuclease (MNase)
is popular nuclease for digestion of chromatin.
It cuts preferentially at ↓WWWW (↓AATT)
sites
at the ends of the nucleosome DNA

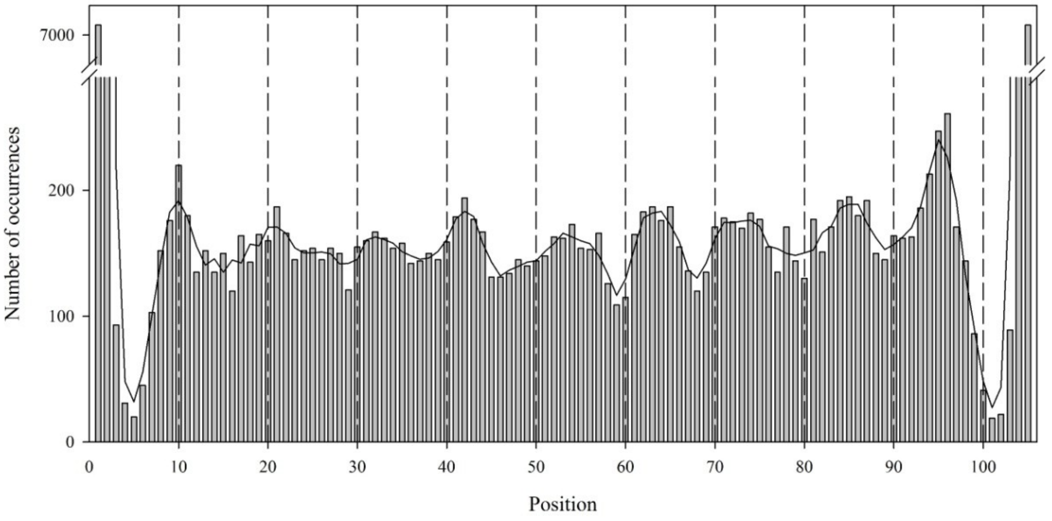
Alignment of nucleosome DNA sequences (C.elegans) by left ends

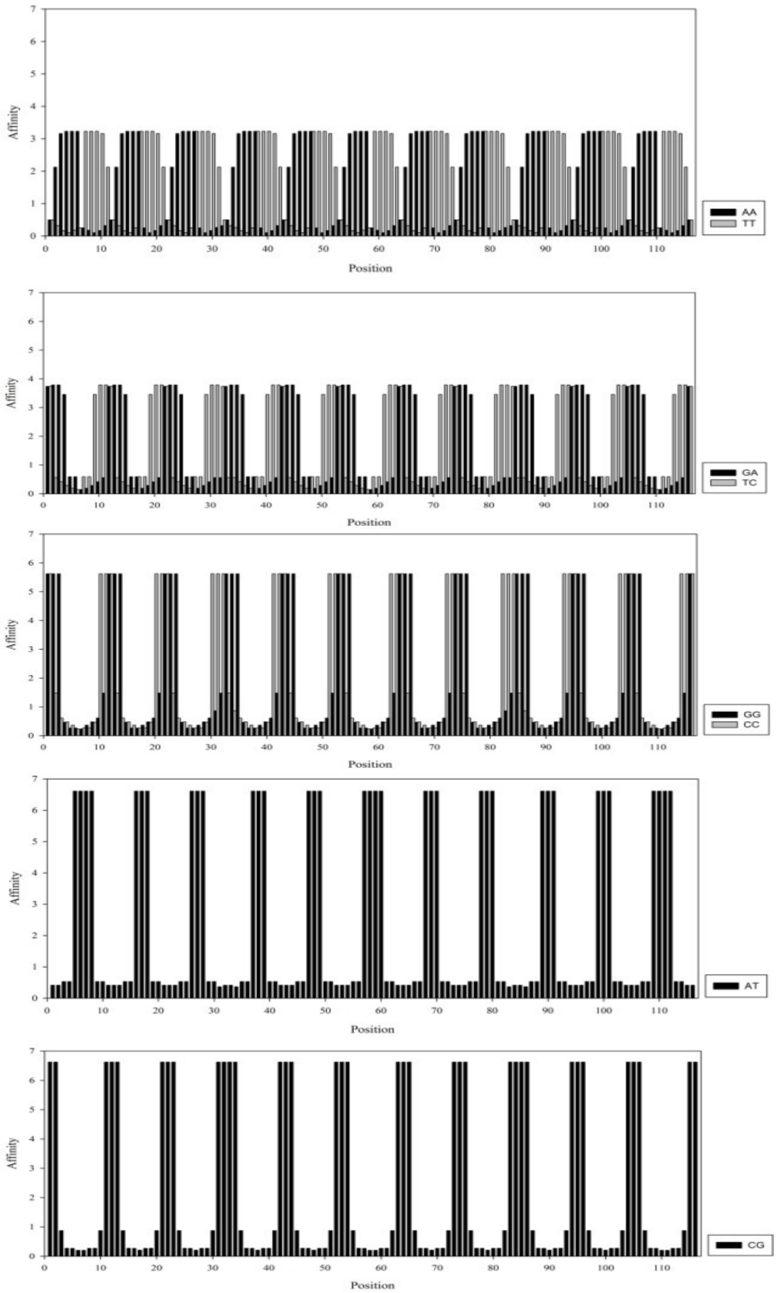


Alignment by right ends

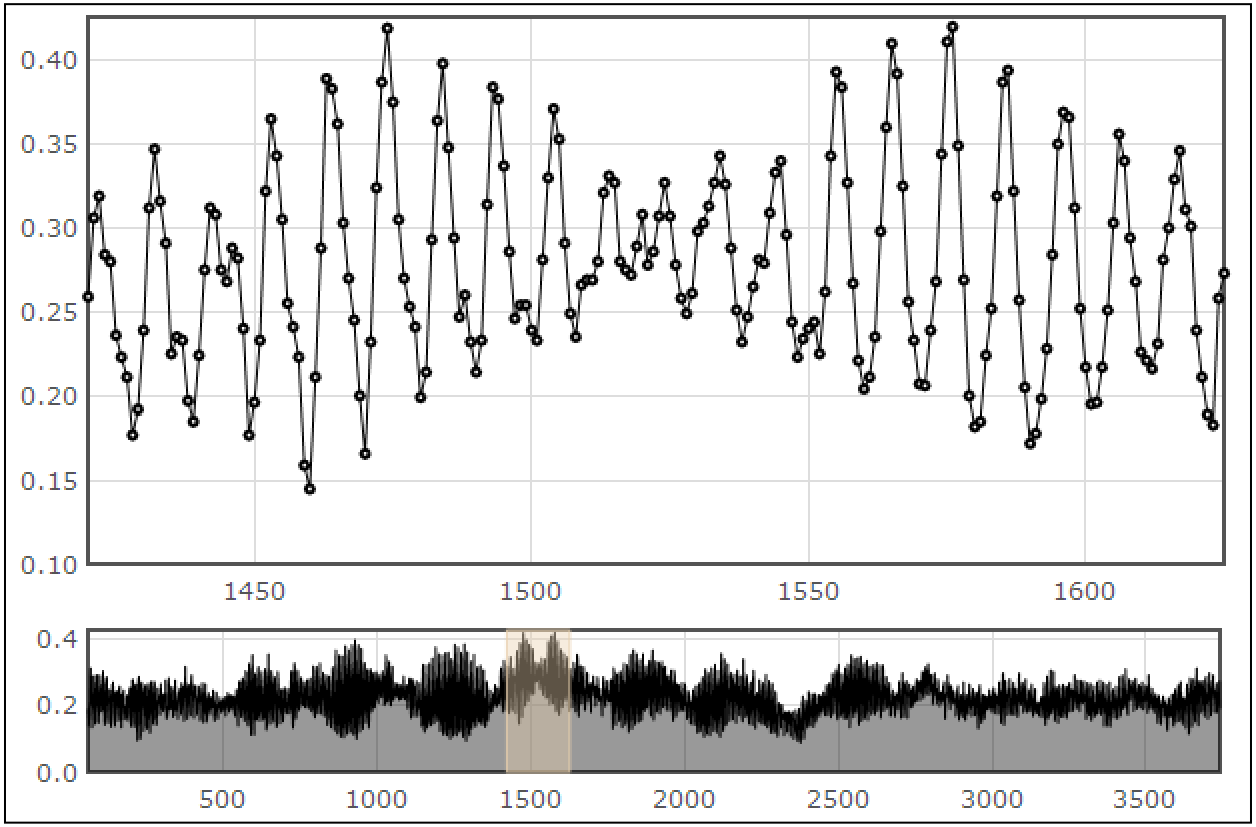


Periodicity all along



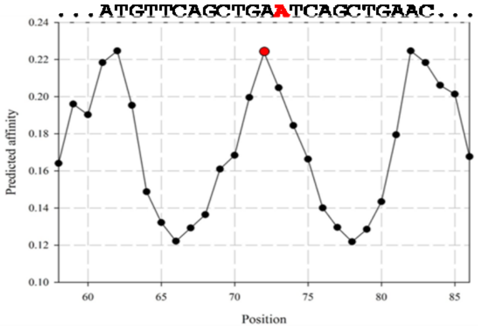


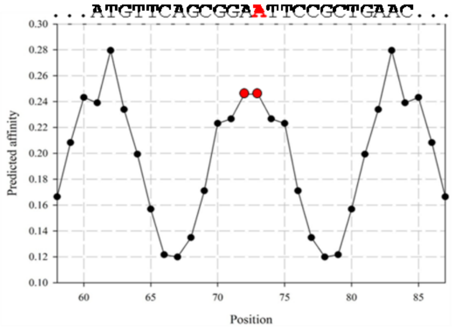
Full length (11 periods) matrix of bendability – nucleosome probe

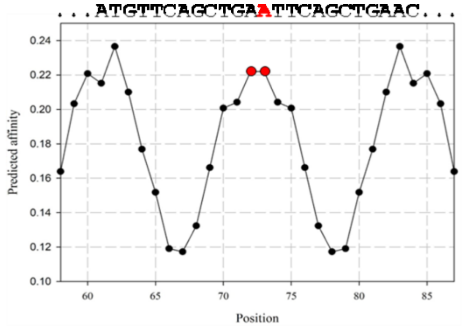


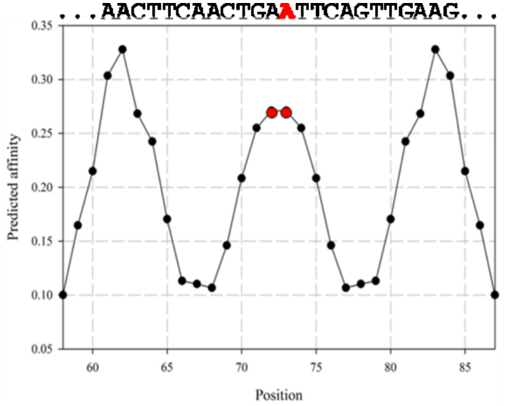
Example of the output from the nucleosome mapping server
http://www.cs.bgu.ac.il/~nucleom

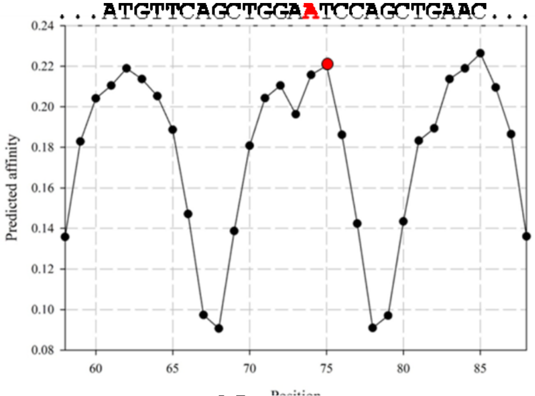
Examples of mapping of sharply positioned nucleosomes

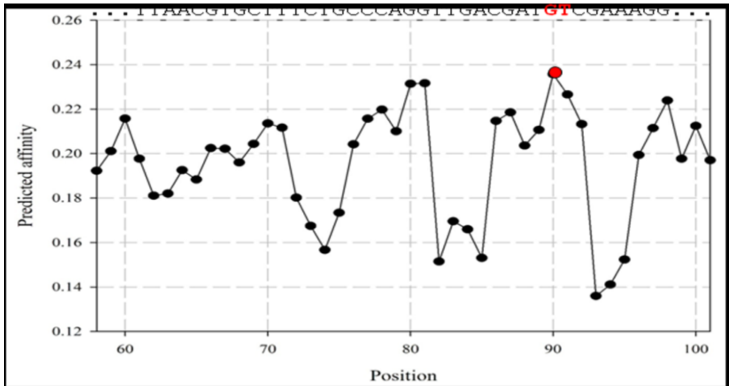


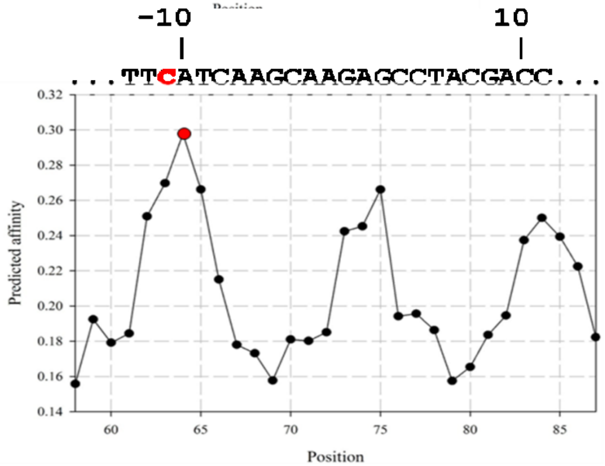


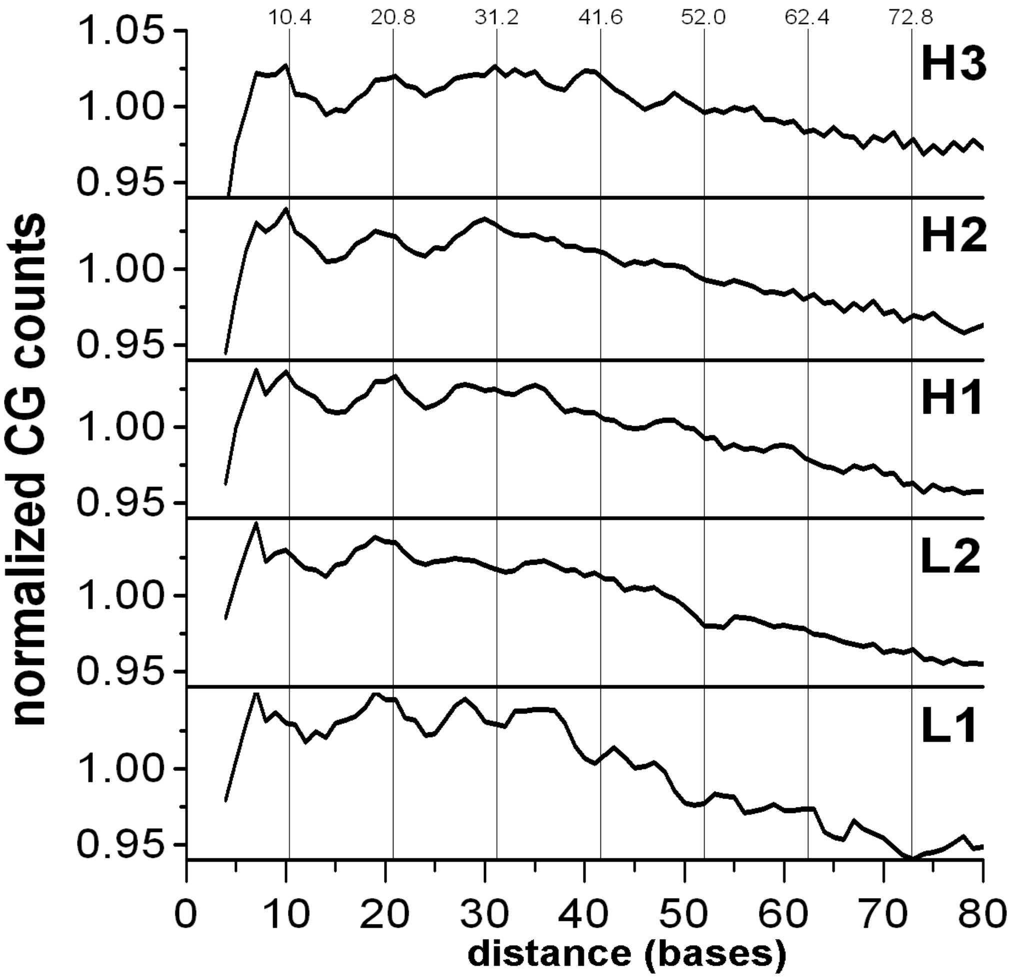




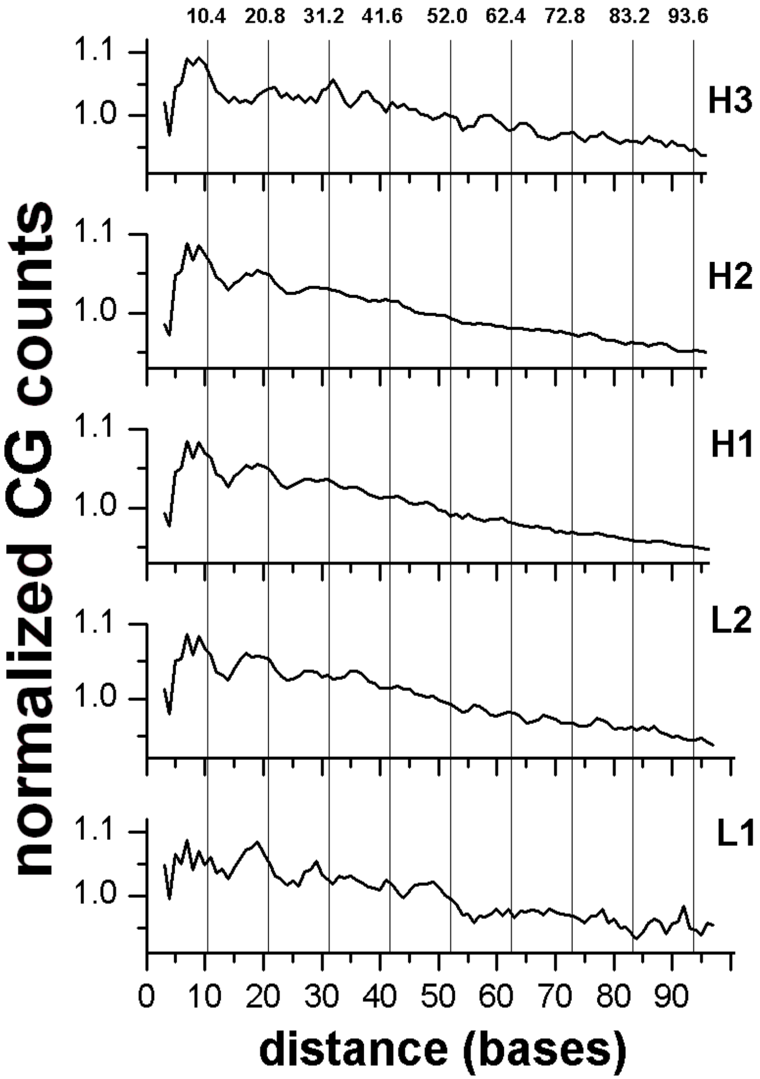




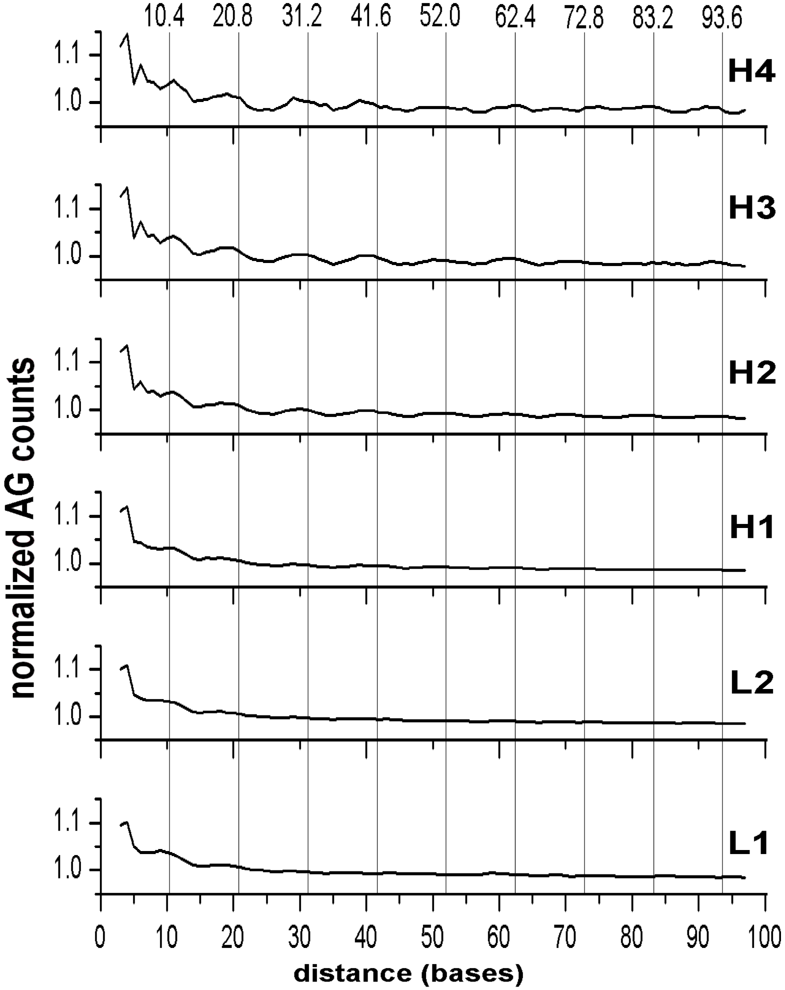



10.4
20.8
31.2
41.6
52.0
62.4
72.8
H3
H2
H1
L2
L1
1.05
1.00
0.95
normalized CG counts
0
10
20
30
40
50
60
70
80
distance (bases)

human


10.4
20.8
31.2
41.6
52.0
62.4
72.8
83.2
93.6
H3
H2
H1
L2
L1
1.1
1.0
normalized CG counts
0
10
20
30
40
50
60
70
80
90
distance (bases)

mouse


10.4
20.8
31.2
41.6
52.0
62.4
72.8
83.2
93.6
H4
H3
H2
H1
L2
L1
1.1
1.0
normalized AG counts
0
10
20
30
40
50
60
70
80
90
100
distance (bases)

chicken

extention motifs	isochores	starting triplets
AAAAA TTTTT	L1	TTT (top)
AAAAA TTTTT	L2	TTT (top)
C AGAAA TTTCT G	H1	TTT (top)
C AGAAA TTTCC GGAAA TTTCT G	H1	CGG_
TCCCC AGGGG	H2	CAG (top)
CCCCT GGGGA	H2	CTG (top)
TCCCC GGGGA	H2	CCG
AGGGG CCCCT	H3	GGG (top)
AGGGG CCCCC GGGGG CCCCT	H3	CGG
Y RRRRR YYYYY RRRRR YYYYY R		human

extention motifs				isochores	starting triplets (top)
AAAAA	TTTTT			L1	TTT
AAAAA	TTTTT			L2	AAA
	TTTCT	G		H1	TTT
	C	AGAAA		H1	AAA
	TCCCC	AGGGG		H2	CAG
	CCCCT	GGGGA		H2	CTG
AGGGG	CCCCT	GGGGG	CCCCC	H3	CTG
GGGGG	CCCCC	AGGGG	CCCCT	H3	CAG

RRRRR YYYYY RRRRR YYYYY

mouse

extention motifs	isochores	starting triplets
AAAAA TTTTT	L1	AAA (top)
GAAAA TTTTC	L2	TTT (top)
TTTCT G	H1	TTT (top)
C AGAAA	H1	AAA (top)
G CTCCC GGGAG C	H2	CCG
G CTCCC GGGAG C	H3	CCG
TG CCCCC GGGGG CA	H4	CCG
Y RRRRR YYYYY RRRRR Y		chicken

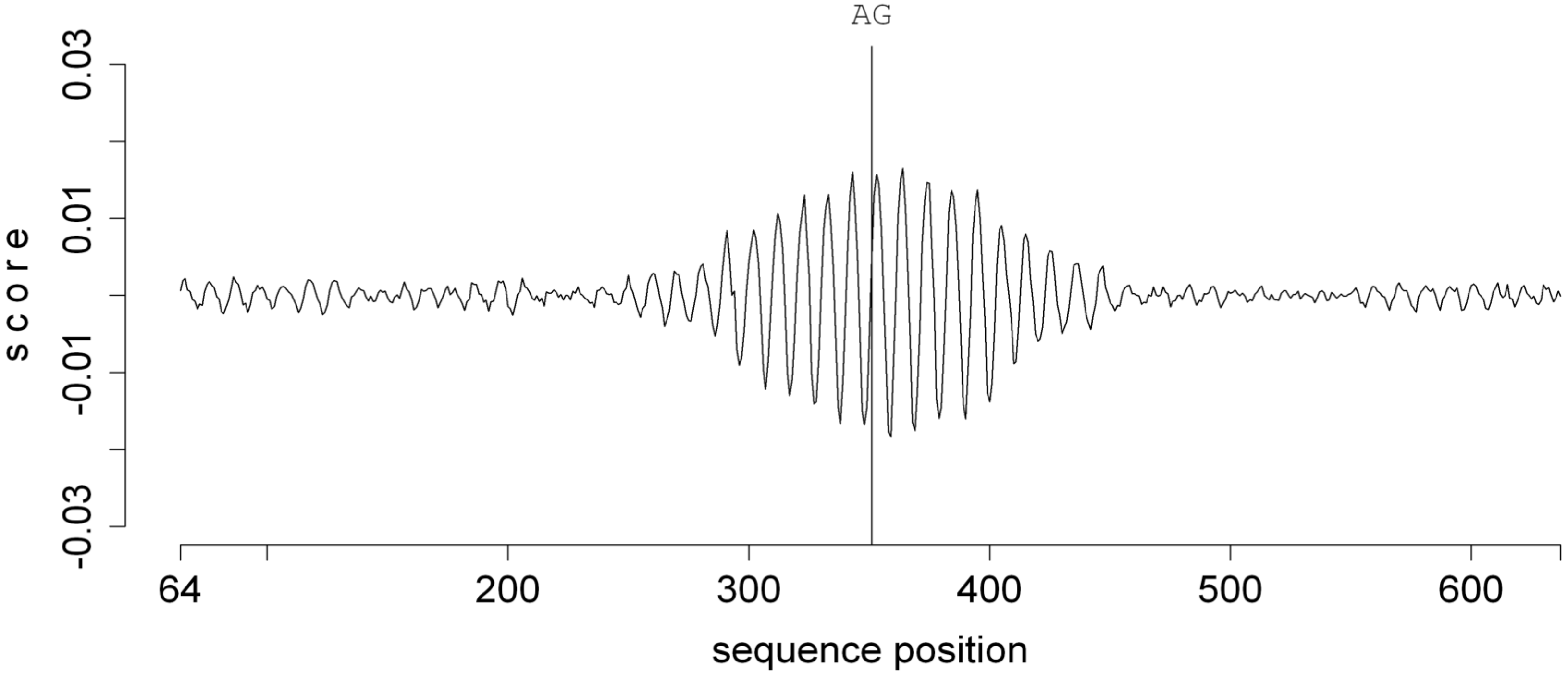
```
human     AAAAA TTTTT
mouse     AAAAA TTTTT               L1
chicken   AAAAA TTTTT

human     AAAAA TTTTT
mouse     AAAAA TTTTT               L2
chicken   GAAAA TTTTC

human   C AGAAA TTTCT G             H1
mouse           TTTCT G
        C AGAAA
chicken         TTTCT G
        C AGAAA

human           TCCCC AGGGG
                CCCCT GGGGA
mouse           TCCCC AGGGG
                CCCCT GGGGA
chicken       G CTCCC GGGAG C
Consensus       YCCCY RGGGR         H2

human     AGGGG CCCCT
mouse     AGGGG CCCCT GGGGG CCCCC
          GGGGG CCCCC AGGGG CCCCT
chicken       G CTCCC GGGAG C
Consensus RGGGG CCCCY RGGGG CCCCY   H3

chicken      TG CCCCC GGGGG CA      H4


        Y RRRRR YYYYY RRRRR YYYYY
```

```
R Y Y Y Y Y R R R R R Y Y Y Y Y R R R R R Y
 |         |         |         |         |
A|T T T T T|A A A A A|T T T T T|A A A A A|T
 |         |         |         |         |
 |        T|G        |        T|G        |
A|T T T T  |  A A A A|T T T T  |  A A A A|T
 |        C|A        |        C|A        |
 |         |         |         |         |
A|T T T T C|G A A A A|T T T T C|G A A A A|T
A|T T T C C|G G A A A|T T T C C|G G A A A|T
A|T T C C C|G G G A A|T T C C C|G G G A A|T
A|T C C C C|G G G G A|T C C C C|G G G G A|T
 |         |         |         |         |
A|C        |        A|C        |        A|C
 |  C C C C|G G G G  |  C C C C|G G G G  |
G|T        |        G|T        |        G|T
 |         |         |         |         |
G|C C C C C|G G G G G|C C C C C|G G G G G|C
```

isochores L1

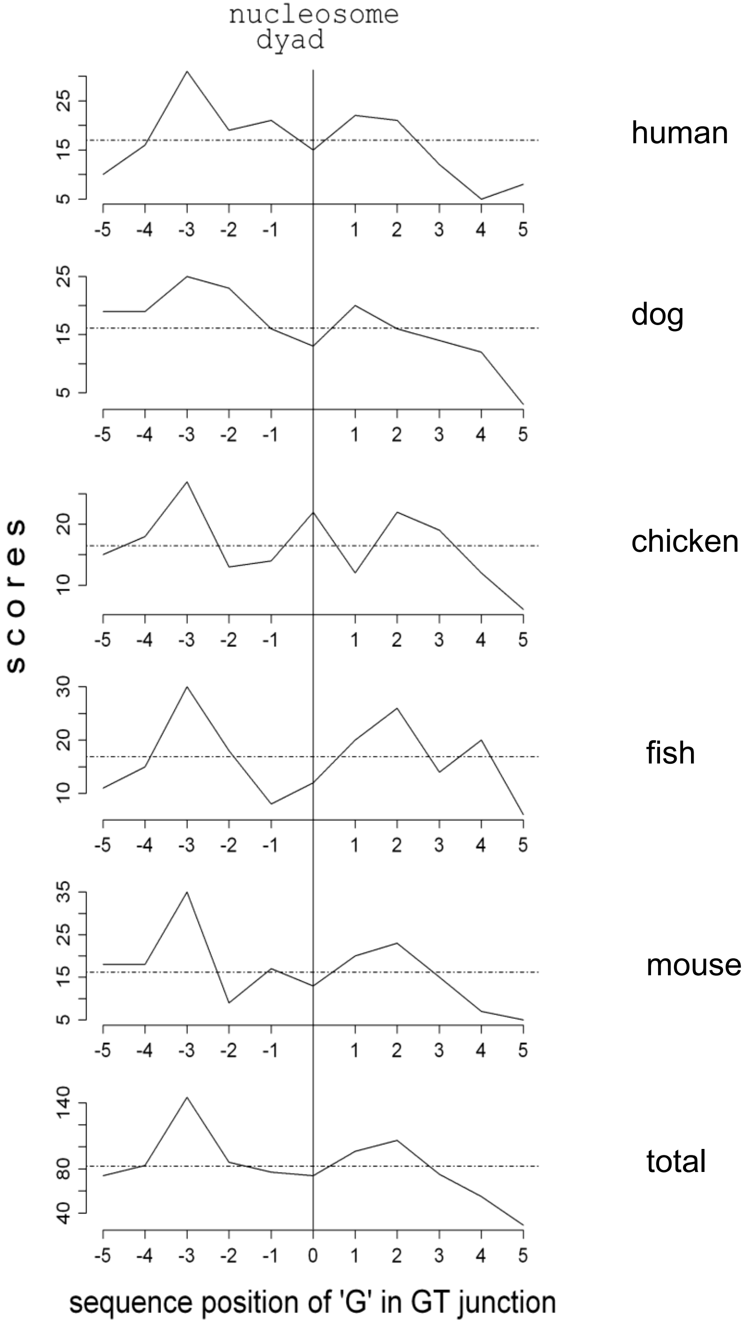
most frequent patterns

isochores H3

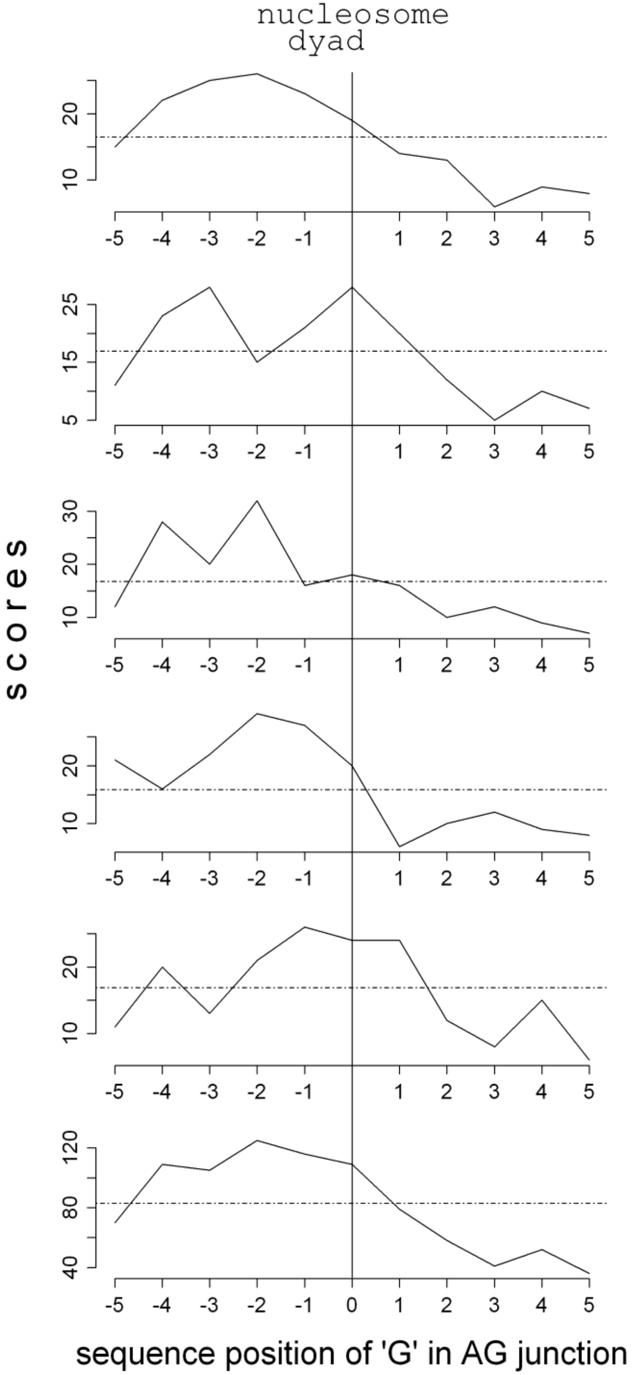


Splice junctions preferably reside in the nucleosomes, preferably at certain distance from the nearest nucleosome center

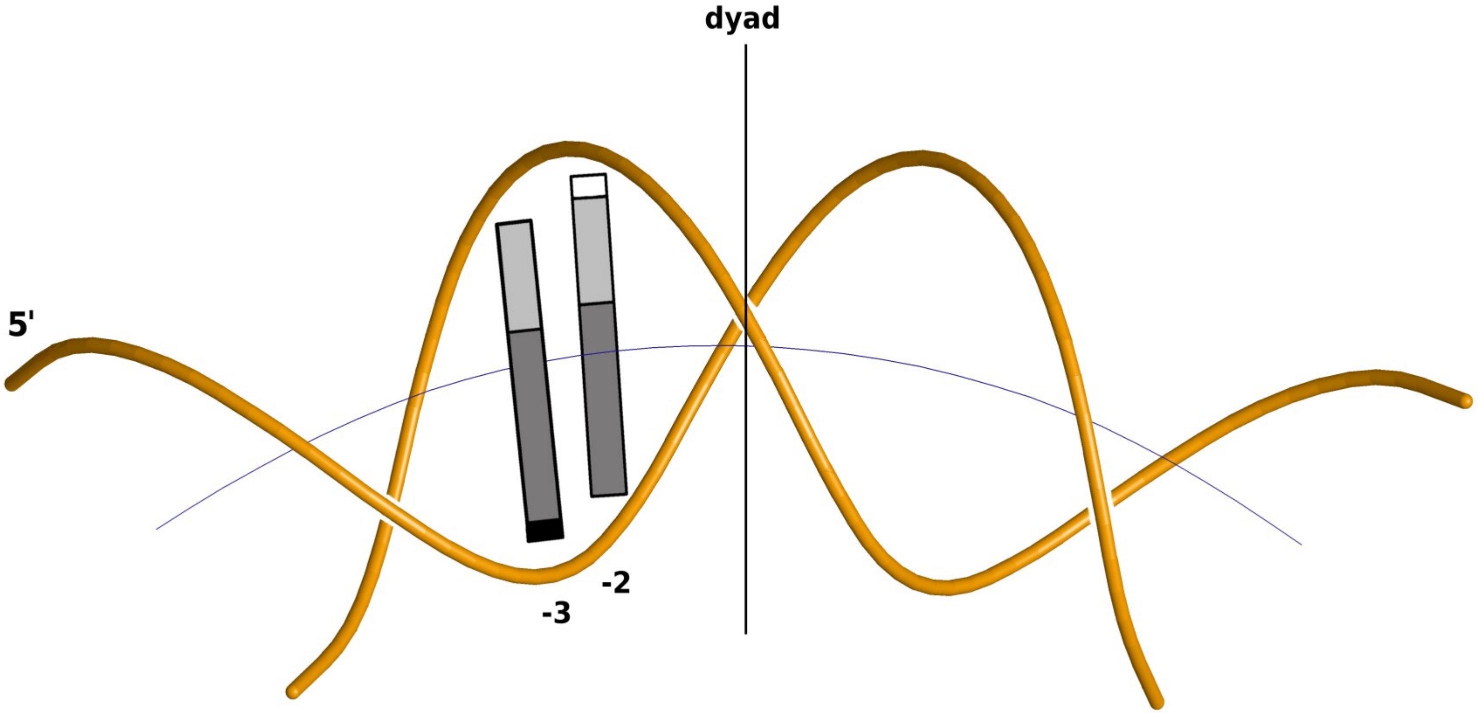
Jan Hapala 2010



Position -3
preferred



Position -2
preferred


dyad
5'
-2
-3

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

Such orientation protects guanines from spontaneous depurination and oxidation

The most frequent spontaneous damages to DNA bases:

depurination of G

oxidation of G

deamination of C

Origin of the chromatin code
is to be looked for in

prokaryotes

Triplet extension (Shannon) patterns for A+T rich prokaryotic genomes

species	G+C content %	extension motif
F. nucleatum	27.2	[(a)t]**(A)(T)**[(a)t]
N. equitans	31.6	(ta)t**(A) t**(at)
- " -		(at)**a (T)**a(ta)
S. solfataricus	35.8	[(t)a]ttt**(A)(T)**[(a)(t)]
T. denicola	37.9	[(a)t]**(A)(T)**[a(t)]
C. pneumoniae	40.0	[g(a)]**G(A)**[g(a)
- " -		[(t)c]**(T)C**[(t)c]
M. acetivorans	42.7	[g(a)]**G(A)(T)C**[(t)c]
A. aeolicus	43.3	[gg(a)]**gG(A)**[gg(a)]
- " -		[(t)cc]**(T)C**c[(t)cc]
B. subtilis	43.5	[g(a)(t)]**G(A)(T)C**[(a)(t)c]
T. maritima	46.2	(gaa)**G(A)**[g(a)]
- " -		[(t)c]**(T)C**(ttc)
D. ethenogenes	48.9	(cggc)cggc**(T)C**agccg(gccg)
consensus		**G(A)(T)C**

CGAAAATTTTCG

same as in eukaryotes!:

CGRAAATTTYCG

α-helices

10-15 aa long
(30-45 bases in DNA)

often **amphipatic**
(alternating hydrophobic/hydrophilic aa)

Period ~3.5 residues
(~10.5 bases in DNA)

Leu (L) - TTx in DNA
Lys (K) - AAx in DNA

What this periodical motif codes for in prokaryotes?

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

GAA AAT TTT CGA AAA TTT TCG AAA ATT TTC
glu asn phe arg lys phe ser lys ile phe
```

non-polar amino acids

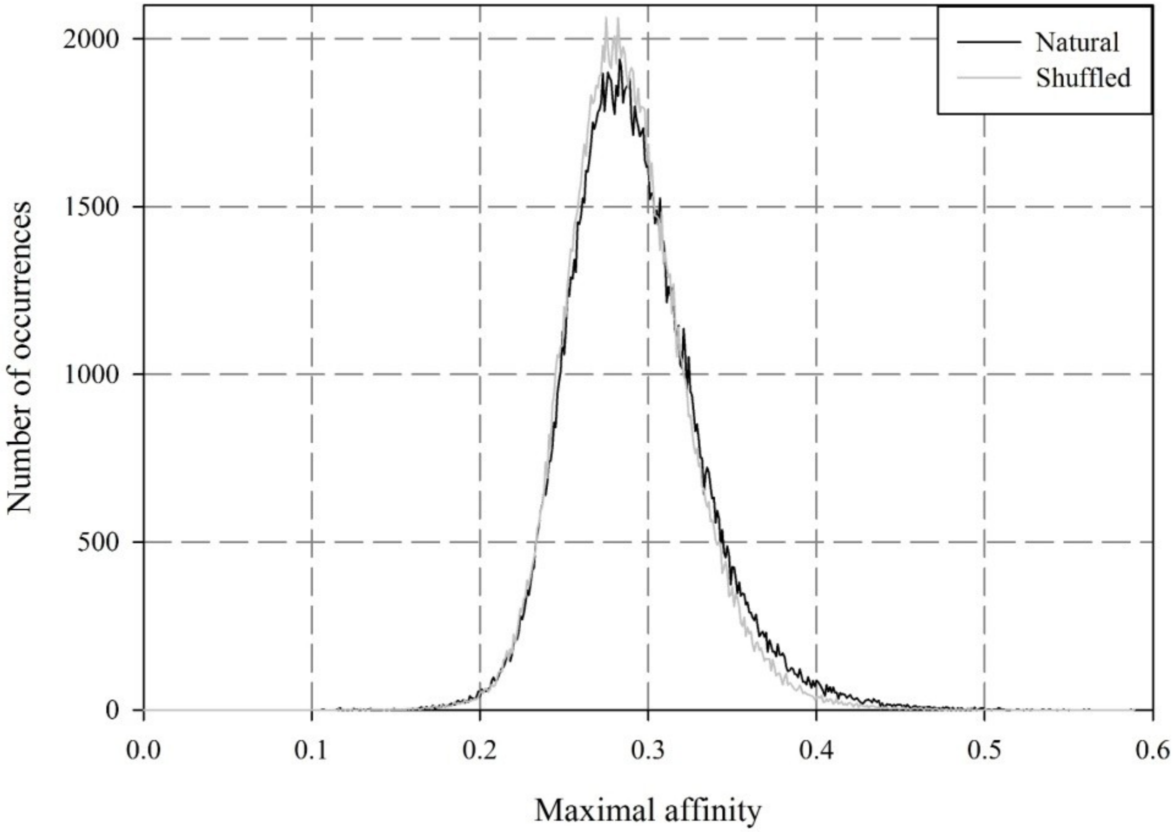
ala
gly
ile
leu
met
phe
pro
val

polar amino acids

arg
asn
asp
cys
glu
gln
his
lys
ser
thr
trp
tyr

Natural nucleosome sequence periodicity is only slightly higher than in random sequences.

Match to simple periodical probe:



Deciphering of the chromatin code opens a new era of high resolution chromatin studies

One can now obtain accurate information on translational and rotational positioning of DNA in the nucleosomes,

for any sequence,
in no time

Nucleosome mapping in no time, with 1 base resolution:

http://www.cs.bgu.ac.il/~nucleom/

Gabdank et al., 2010

THE COLLEAGUES WITH WHOM WE AGONIZED TOGETHER ALL THESE YEARS (1978-2010) TO FINALLY REACH THE GOAL:

Joel Sussman (1978)
Thomas Bettecken (1979)
Galina Mengeritsky (1983)
Levy Ulanovsky (1983)
Roni Wartenfeld (1984)
Jacqui Beckmann (1991)
Ilya Ioshikhes (1992)
Alex Bolshoy (1992)
Konstantin Derenshtein (1996)
Mark Borodovsky (1996)
Dmitry Denisov (1997)
Edward Shpigelman (1997)
Kevin Shapiro (1997)
Hanspeter Herzel (1998)
Ivo Grosse (1998)
Olaf Weiss (1998)
Yuko Wada-Kiyama (1999)
Kentaro Kuwabara (1999)
Yasuo Sakuma (1999)
Ryoiti Kiyama (1999)
Yoshiaki Ohnishi (1999)
Michael Zhang (1999)
Jiri Fajkus (2001)
Toshimichi Ikemura (2003)
Takashi Abe (2003)
Simon Kogan (2003)
M.Kato (2003)
Amir Cohanim (2005)
Yehezkiel Kashi (2005)
Fadil Salih (2007)
Bilal Salih (2007)
Idan Gabdank (2009)
Danny Barash (2009)
Zakharia Frenkel (2009)
Alexandra Rapoport (2010)
Jan Hapala (2010)

Alu NUCLEOSOMES

Alu sequence (consensus)

```
                 ggccgggcgcggtgg  15
ctcacgcctgtaatcccagcactttgggaggc  47
CGaggcgggCGgatcacctgaggtcaggagtt  79
CGagaccagcctggc-caacatggtgaaaccc 110
CGtctctactaaaaatacaaaaattagccggg 142
CGtggtggcgCGcgcctgtaatcccagctact 174
CGggaggctgaggcaggagaatCGcttgaacc 206
CGggaggcggaggttgcagtgagccgagatcg 238
CGccactgcactccagcctgggCGacagagcg 270
agactccgtctcaaaaaaaa
```

Alu, hidden 8-base repeat

```
                   ggccggg cgcggtgg  15
ctcacgcc tgtaatcc cagcactt tgggaggc  47
CGaggcgg gcggatca cctgaggt caggagtt  79
CGagacca gcctggc- caacatgg tgaaaccc 110
CGtctcta ctaaaaat acaaaaat tagccggg 142
CGtggtgg cgcgcgcc tgtaatcc cagctact 174
CGggaggc tgaggcag gagaatcg cttgaacc 206
CGggaggc ggaggttg cagtgagc cgagatcg 238
CGccactg cact-cca -gcctggg cgacagag 268
CGagactc cgtctcaa aaaaaa
Yrrrrxxx Yrrrrxxx Yrrrrxxx Yrrrrxxx
```

that is, the Alu repeat is itself a degenerate simple tandem repeat

Two halves of Alu

```
                   ggccggg cgcggtgg  15
ctcacgcc tgtaatcc cagcactt tgggaggc  47
CGaggcgg gcggatca cctgaggt caggagtt  79
CGagacca -gcctggc caacatgg tgaaaccc 110
CGtctcta ctaaaat  acaaaaa           133
                t tagccggg CGtggtgg 150  (15)
cgcgcgcc tgtaatcc cagctact CGggaggc 182  (47)
tgaggcag gagaatcg cttgaacc CGggaggc 214  (79)
ggagg
     ttg cagtgagc cgagatcg CGccactg 246  31 base
cact                                      insert
    -cca -gcctggg cgacagag CGagactc 276 (110)
cgtctcaa aaaaaa                     290 (133)
```

The insert is of very proper size, apparently, to maintain/improve the $(31\text{-}32)_n$ pattern

Alu is made of two repeating pieces of 7S RNA

```
                 ggccgggcgcggtgg  15
                  ==============
ctcacgcctgtaatcccagcactttgggaggc  47
=G=GT=======G=======TAC=C========      7S RNA
CGaggcgggcggatcacctgaggtcaggagtt  79
T====T===A=====G=T====TC========
CGagaccagcctggc-caacatggtgaaaccc 110
=TG=G=TGTAG==CG-=T=T
CGtctctactaaaaatacaaaaattagccggg 142
                          ======
CGtggtggcgcgcgcctgtaatcccagctact 174
==C=========T=======G===========      7S RNA
CGggaggctgaggcaggagaatcgcttgaacc 206
==============T====G=========GT=
CGggaggcggaggttgcagtgagccgagatcg 238
=A====TTCTG==C==T====C==TAT
CGccactgcact-cca-gcctgggcgacagag 268
CGagactccgtctcaaaaaaaa
```

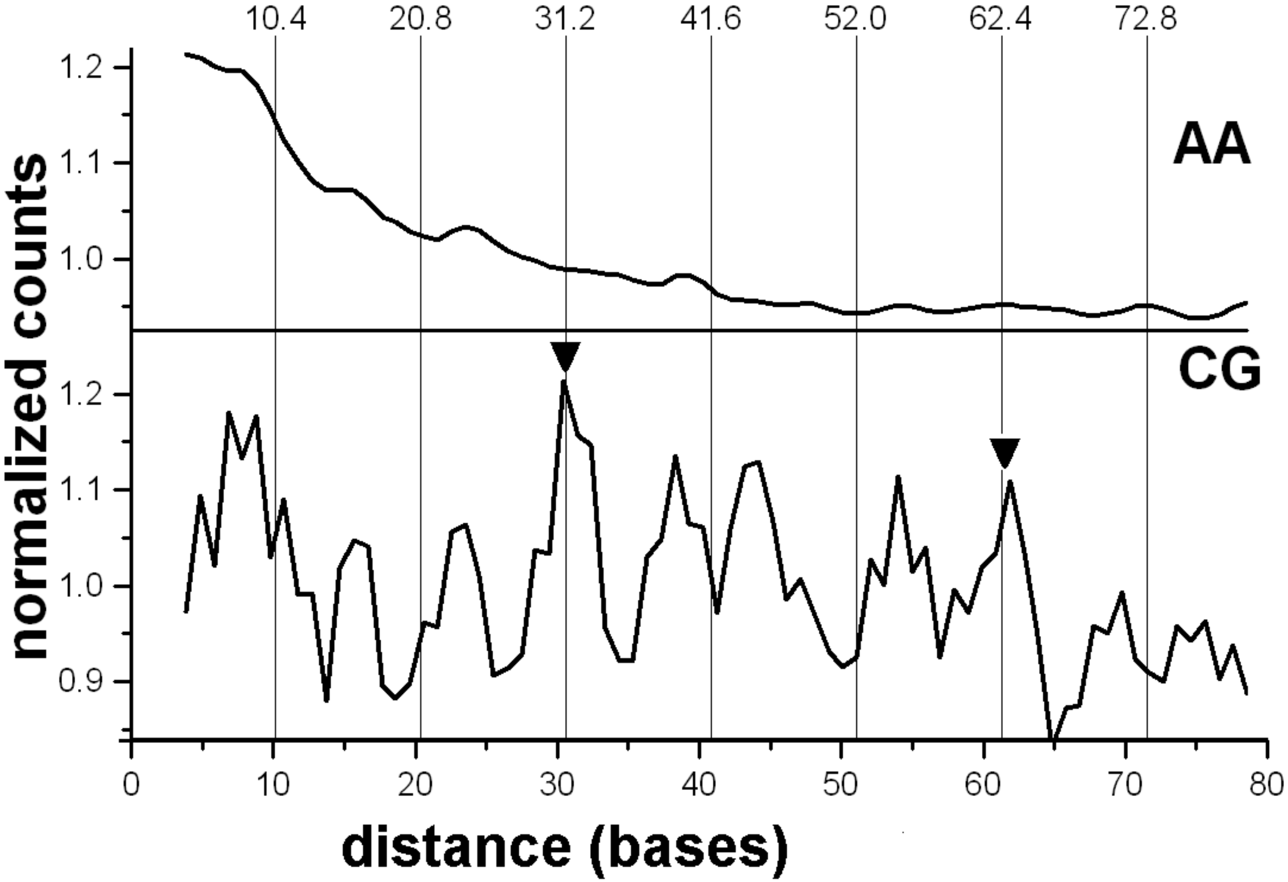
All major types of the Alu repeats have regularly positioned CG

97
nucleosome 1 bends: ↓
AluJ agcactttgggaggc**CG**aggcgggaggatcacttgagcccaggagtt**CG**agaccagcctgggcaacatagtgaaaccc**CG**tctctacaaaaaatacaaaaattagccggg**CG**tggtggcgcgcgcct
AluSx agcactttgggaggc**CG**aggcgggcggatcacctgaggtcaggagtt**CG**agaccagcctggccaacatggtgaaaccc**CG**tctctactaaaaatacaaaaattagccggg**CG**tggtggcgcgcgcct
AluSq agcactttgggaggc**CG**aggcgggtggatcacctgaggtcaggagtt**CG**agaccagcctggccaacatggtgaaaccc**CG**tctctactaaaaatacaaaaattagccggg**CG**tggtggcgggcgcct
AluSp agcactttgggaggc**CG**aggcgggcggatcacctgaggtcgggagtt**CG**agaccagcctgaccaacatggagaaaccc**CG**tctctactaaaaatacaaaaattagccggg**CG**tggtggcgcatgcct
AluSc ccagcactttgggaggc**CG**aggcgggcggatcacgaggtcaagagat**CG**agaccatcctggccaacatggtgaaaccc**CG**tctctactaaaaatacaaaaattagctggg**CG**tggtggcgcgcgcct
AluY cagcactttgggaggc**CG**aggcgggcggatcacgaggtcaggagat**CG**agaccatcctggctaacacggtgaaaccc**CG**tctctactaaaaatacaaaaaattagccggg**CG**tggtggcgggcgcct
AluYa5 cagcactttgggaggc**CG**aggcgggcggatcacgaggtcaggagat**CG**agaccatcccggctaaaacggtgaaaccc**CG**tctctactaaaaatacaaaaaattagccggg**CG**tagtggcgggcgcct
AluYa8 ccagcactttgggaggc**CG**aggcgggcggatcacgaggtcaggagat**CG**agaccatcccggctaaaacggtgaaaccc**CG**tctctactaaaactacaaaaaatagccggg**CG**tagtggcgggcgcct
AluYb8 cagcactttgggaggc**CG**aggcgggtggatcatgaggtcaggagat**CG**agaccatcctggctaacaaggtgaaaccc**CG**tctctactaaaaatacaaaaaattagccggg**CG**cggtggcgggcgcct

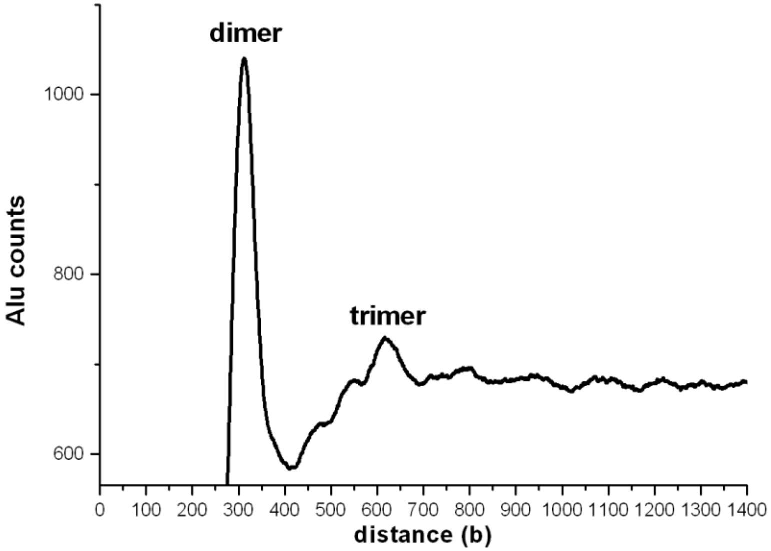
223
nucleosome 2 bends: ↓
AluJ gtagtcccagctact**CG**ggaggctgaggcaggagaatcgcttgaacc**CG**ggaggcggaggttgcagtgagccgtgat**CGCG**ccactgcactccagcctgggcgacagag**CG**agaccctgtctcaaa
AluSx gtaatcccagctact**CG**ggaggctgaggcaggagaatcgcttgaacc**CG**ggaggcggaggttgcagtgagccgagat**CGCG**ccactgcactccagcctgggcgacagag**CG**agactccgtctcaaa
AluSq gtaatcccagctact**CG**ggaggctgaggcaggagaatcgcttgaacc**CG**ggaggcggaggttgcagtgagccgagat**CGCG**ccactgcactccagcctgggcaacaagag**CG**aaactccgtctcaa
AluSp gtaatcccagctact**CG**ggaggctgaggcaggagaatcgcttgaacc**CG**ggaggcggaggttgcggtgagccgagat**CGCG**ccattgcactccagcctgggcaacaagag**CG**aaactccgtctcaa
AluSc tgtagtcccagctact**CG**ggaggctgaggcaggagaatcgcttgaacc**CG**ggaggcggaggttgcagtgagccgagat**CG**cgccactgcactccagcctggcgacagag**CG**agactccgtctcaaa
AluY tgtagtcccagctact**CG**ggaggctgaggcaggagaatggcgtgaacc**CG**ggaggcgcaggttgcagtgagccgagat**CG**cgccactgcactccagcctgggcgacagag**CG**agactccgtctcaa
AluYa5 gtagtcccagctacttgggaggctgaggcaggagaatggcgtgaacc**CG**ggaggcgcaggttgcagtgagccgagatcc**CG**ccactgcactccagcctgggcgacagag**CG**agactccgtctcaaa
AluYa8 gtagtcctagctacttgggaggctgaggcaggagaatggcgtgaacc**CG**ggaggcgcaggttgcagtgagccgagatcc**CG**ccactgcactccagcctgggcgacagag**CG**agactccgtctcaaa
AluYb8 gtagtcccagctact**CG**ggaggctgaggcaggagaatggcgtgaacc**CG**ggaagcgcaggttgcagtgagccgagattg**CG**ccactgcagtccagcagtccggcctggg**CG**acagagcgagactcc

Methylation/demethylation of properly positioned CG
in the nucleosome DNA
leads to weakening/strengthening
of the nucleosome,
which is, thus, an epigenetic nucleosome

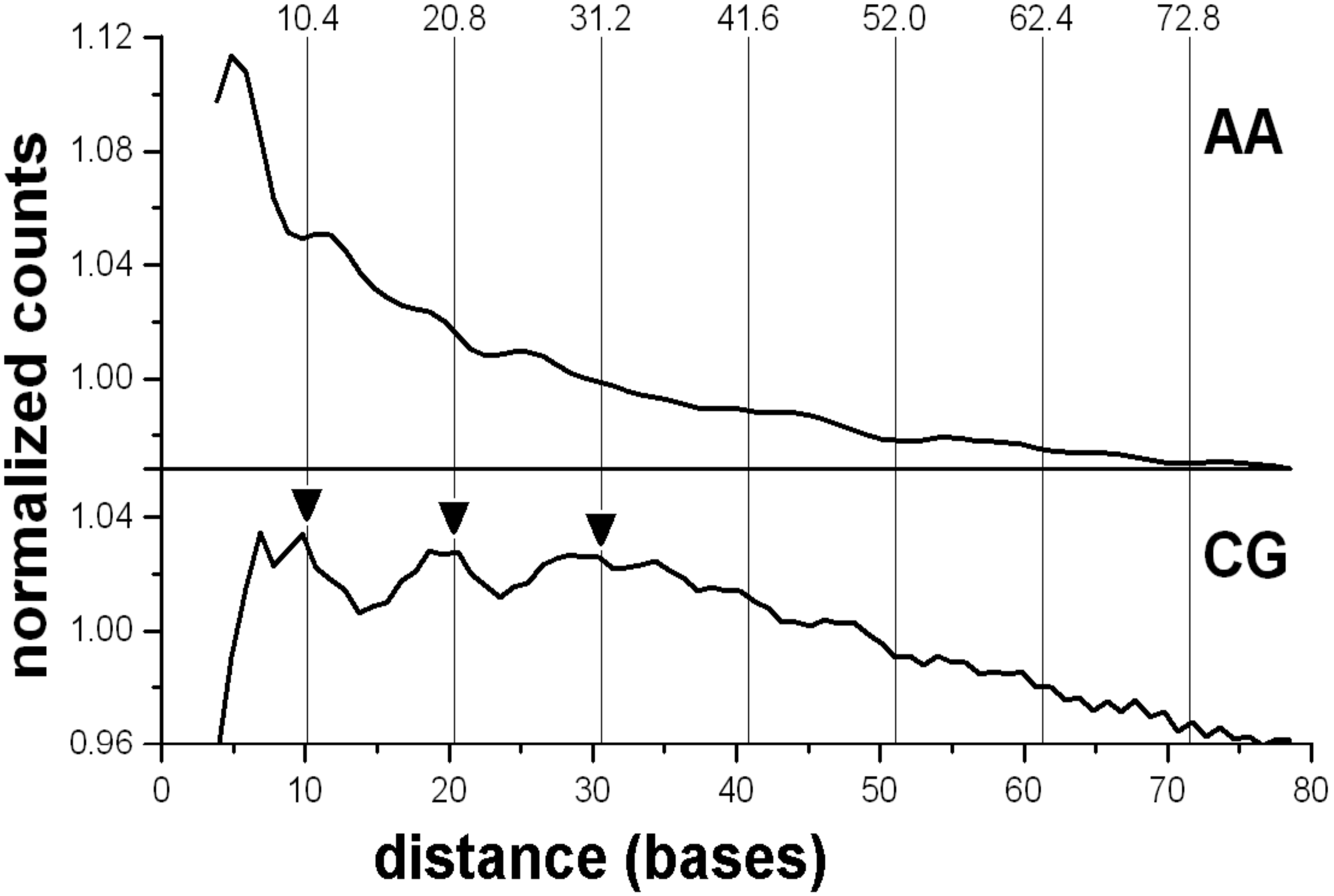
Whole genome (human) shows only 31n periodicity



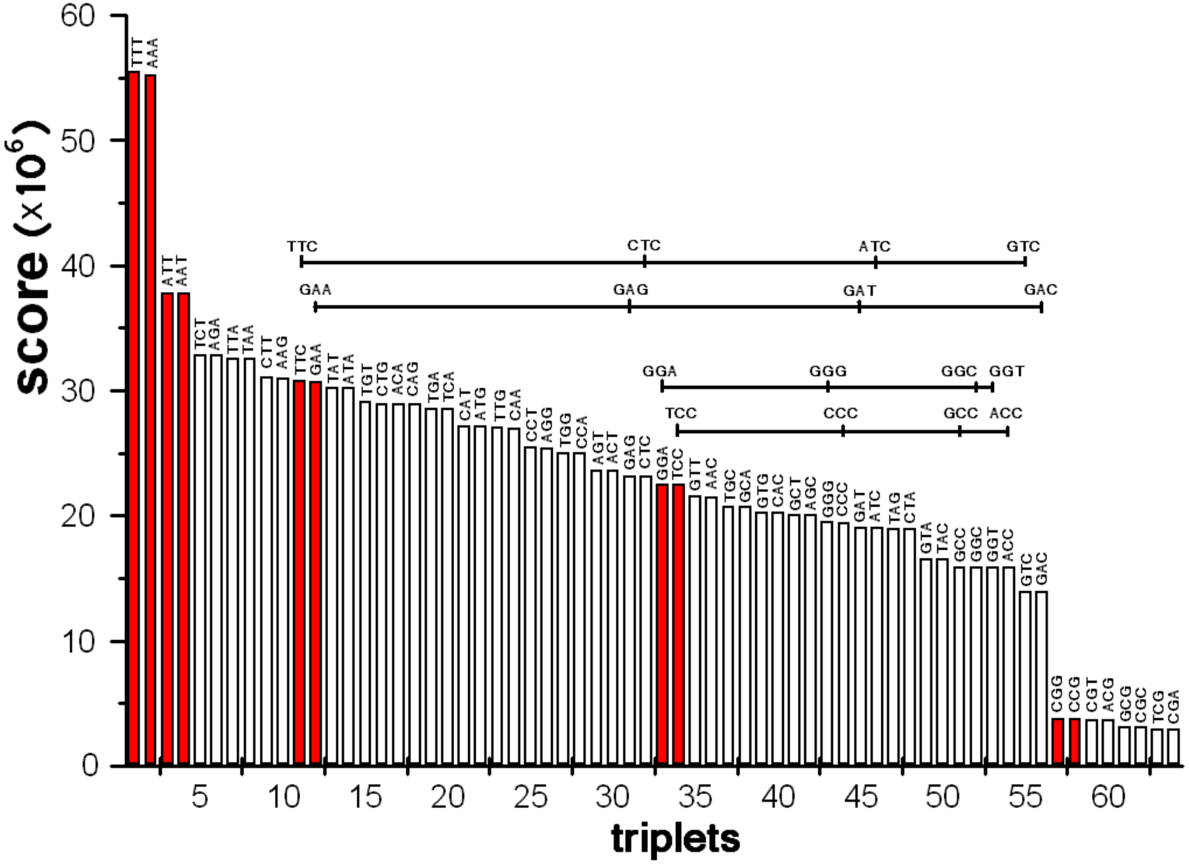
Alu sequences often make tandem clusters



After removal of Alu sequences CG periodicity is seen

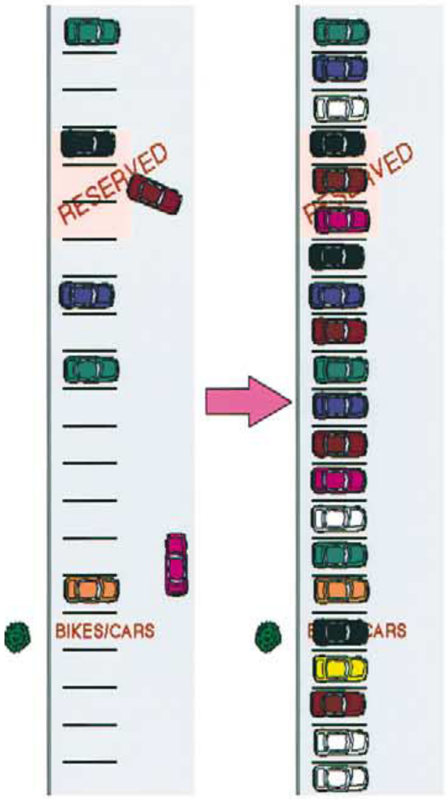


Trinucleotides of human genome fuse in the sequence CC GGAAA TTTCC GG

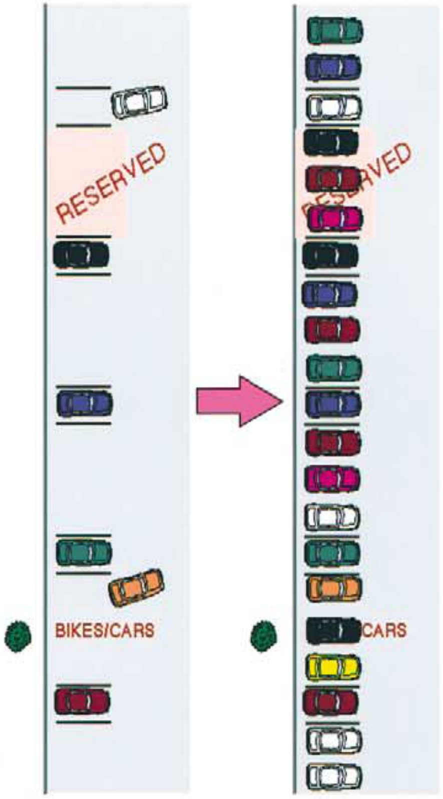


Parking Lot

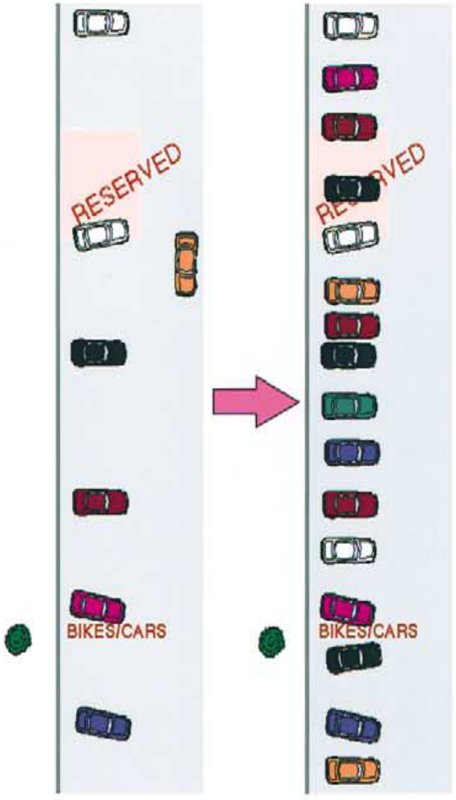
1. Perfect Positioning



2. Partial Positioning

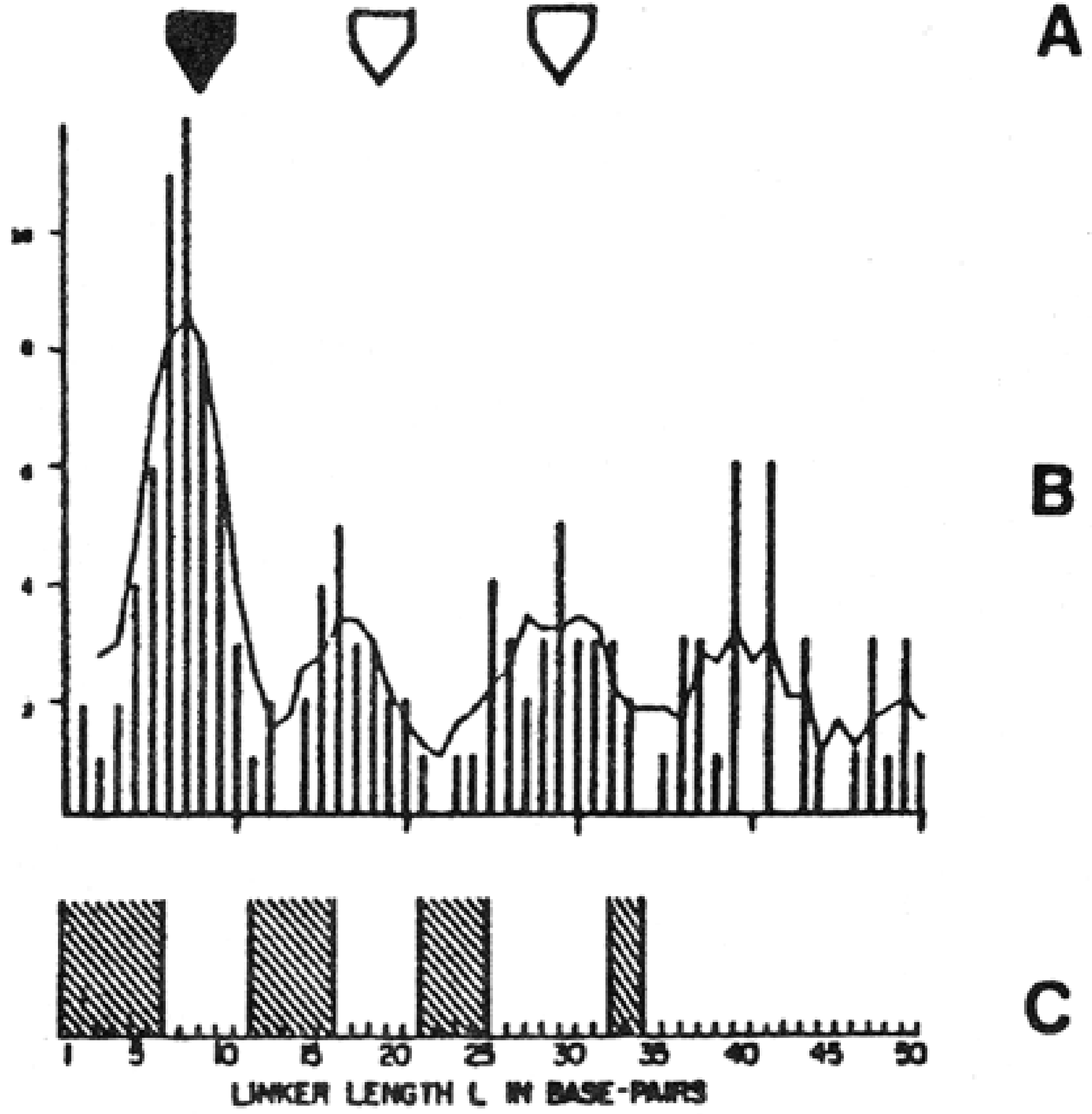


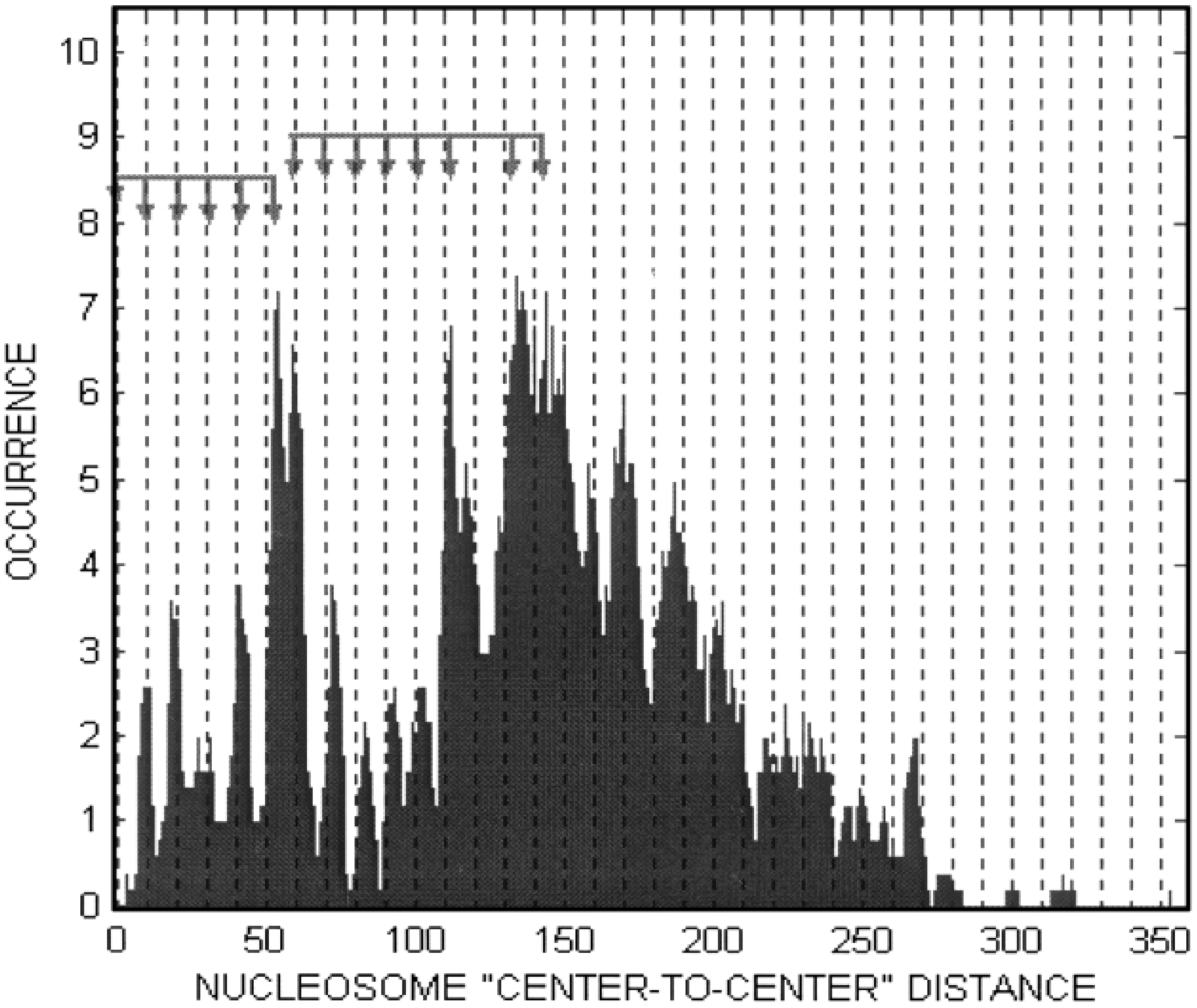
3. Random Placement

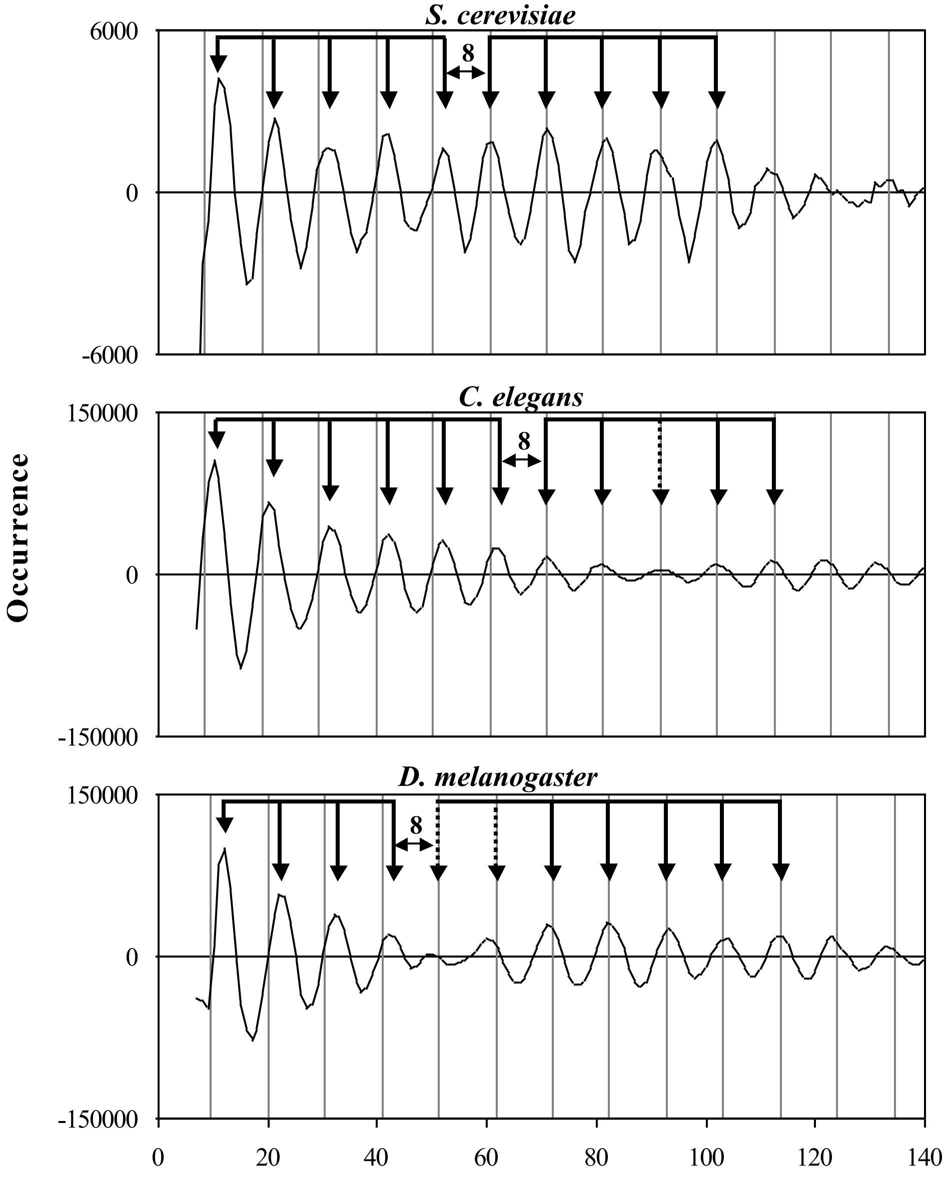


The deformational properties of DNA
is not the only sequence-dependent
factor of nucleosome positioning.

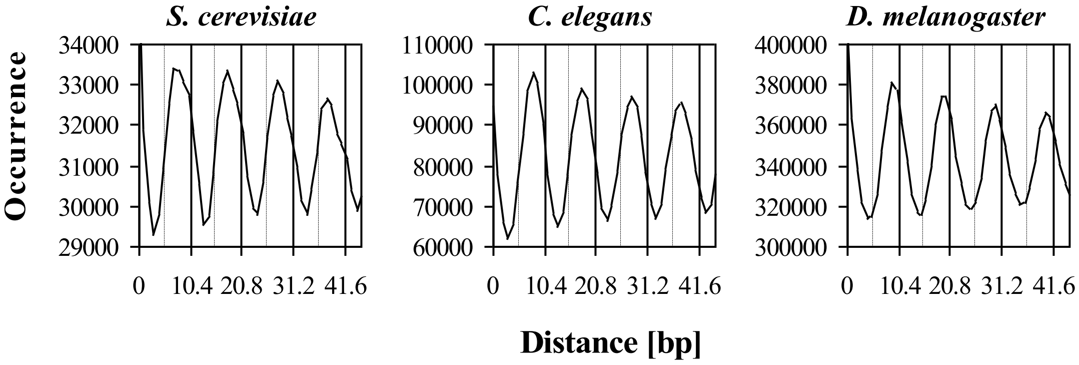
The second factor is the steric exclusion rules,
imposing limitations to the linker lengths.


A
B
C
10
8
6
4
2
1
5
10
15
20
25
30
35
40
45
50
LINKER LENGTH L IN BASE-PAIRS

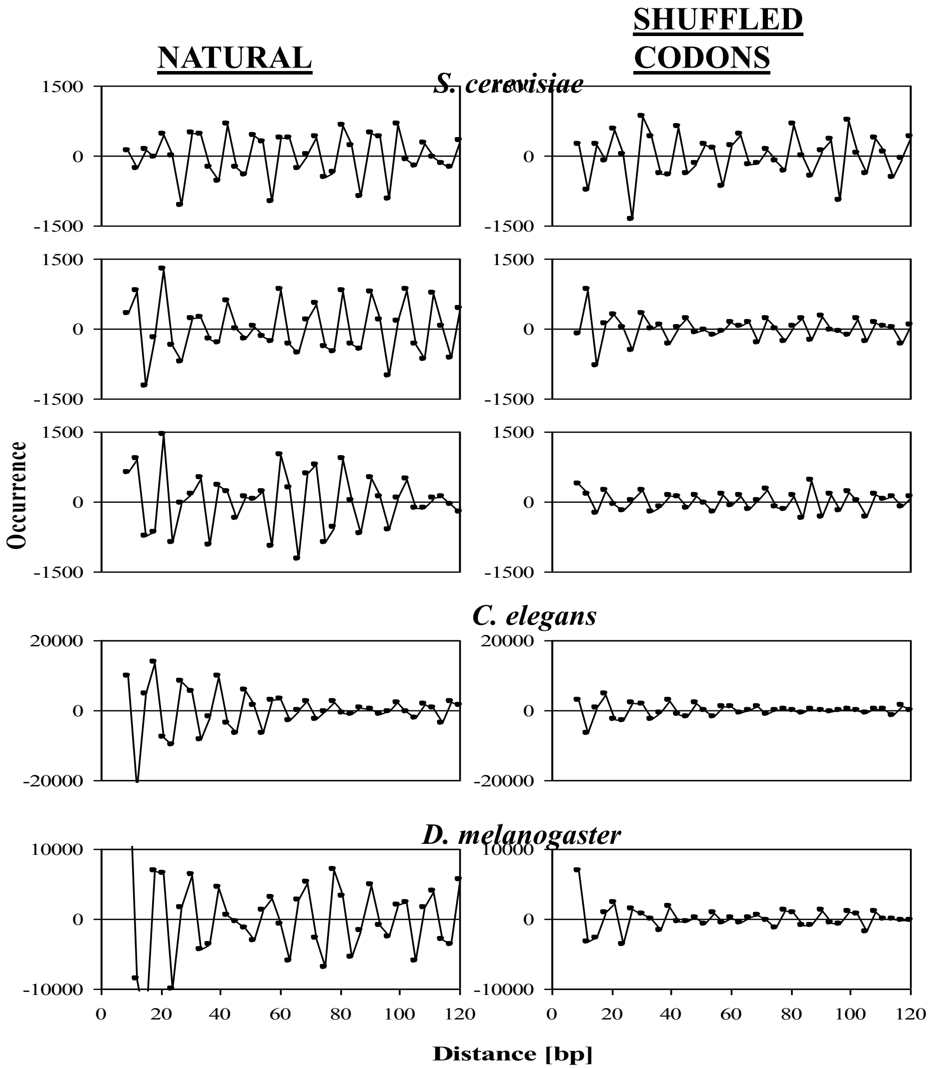

OCCURRENCE
0
1
2
3
4
5
6
7
8
9
10
0
50
100
150
200
250
300
350
NUCLEOSOME "CENTER-TO-CENTER" DISTANCE


S. cerevisiae
6000
0
-6000
8
C. elegans
150000
0
-150000
8
D. melanogaster
150000
0
-150000
8
Occurrence
0
20
40
60
80
100
120
140
Distance [bp]

Linker lengths are 7-8   10.4•n bp



AA-PERIODICITY DISAPPEARS WHEN THE THIRD POSITIONS ARE RANDOMIZED



CODON POSITIONS

1,2

2,3

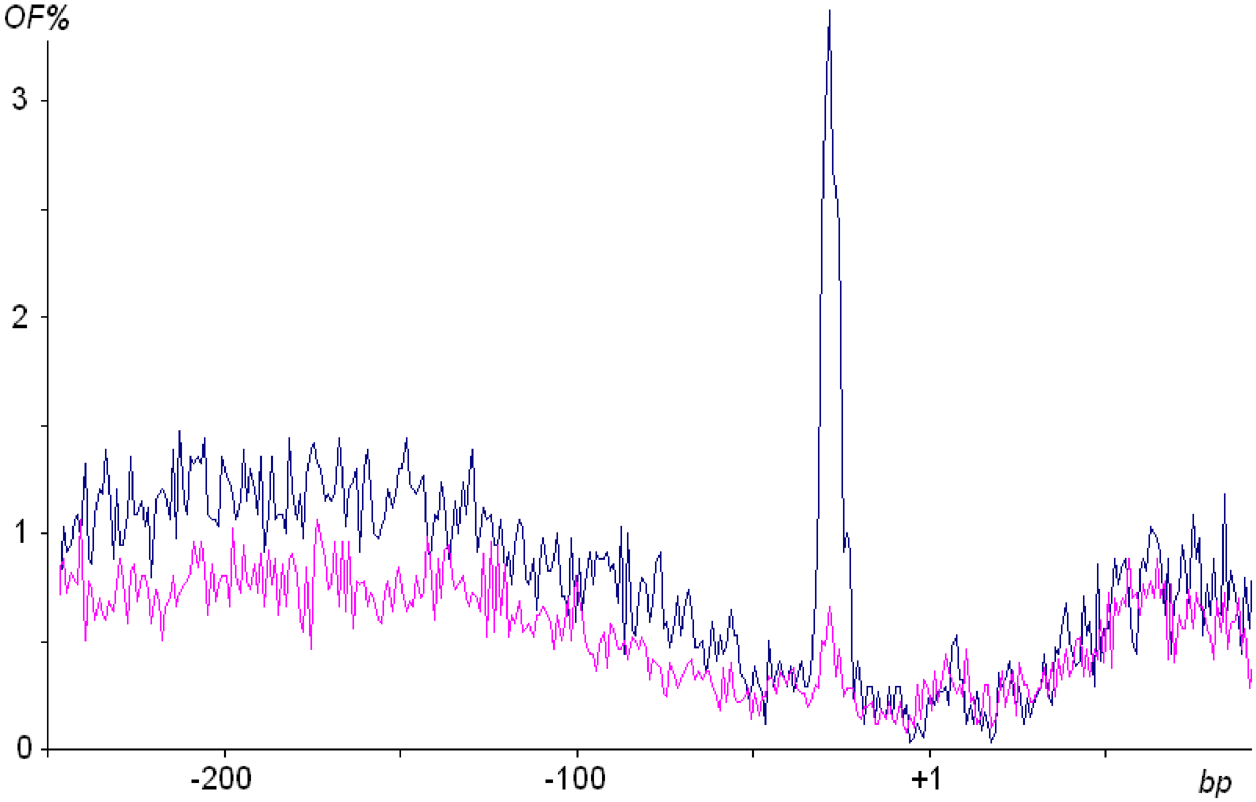
3,1

3,1

3,1

Cohanim 2006

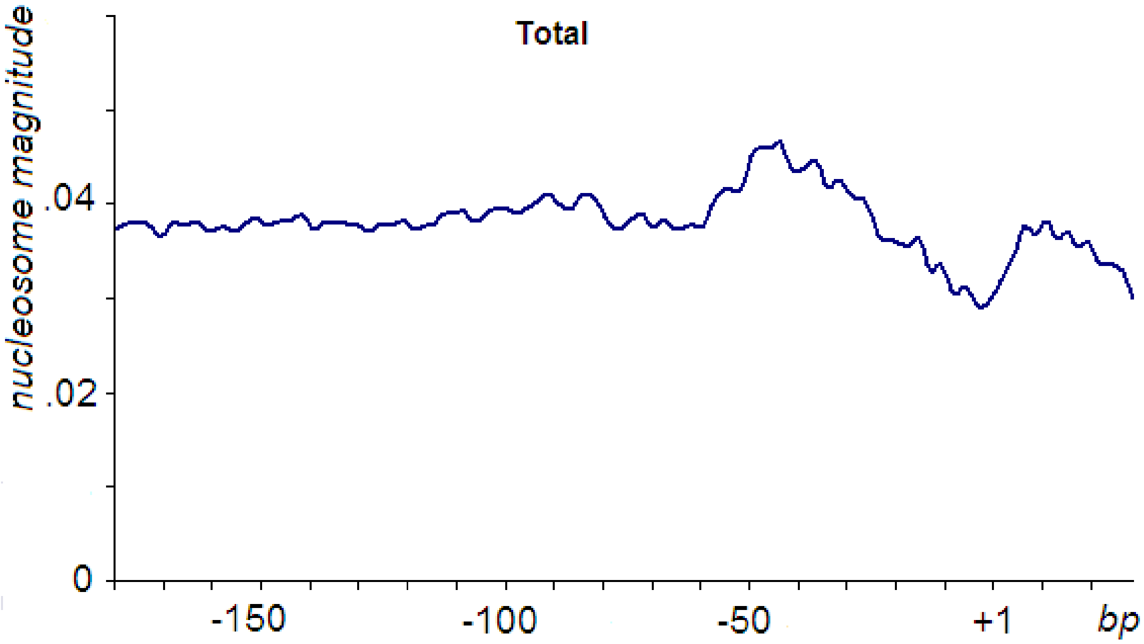
TATA-box



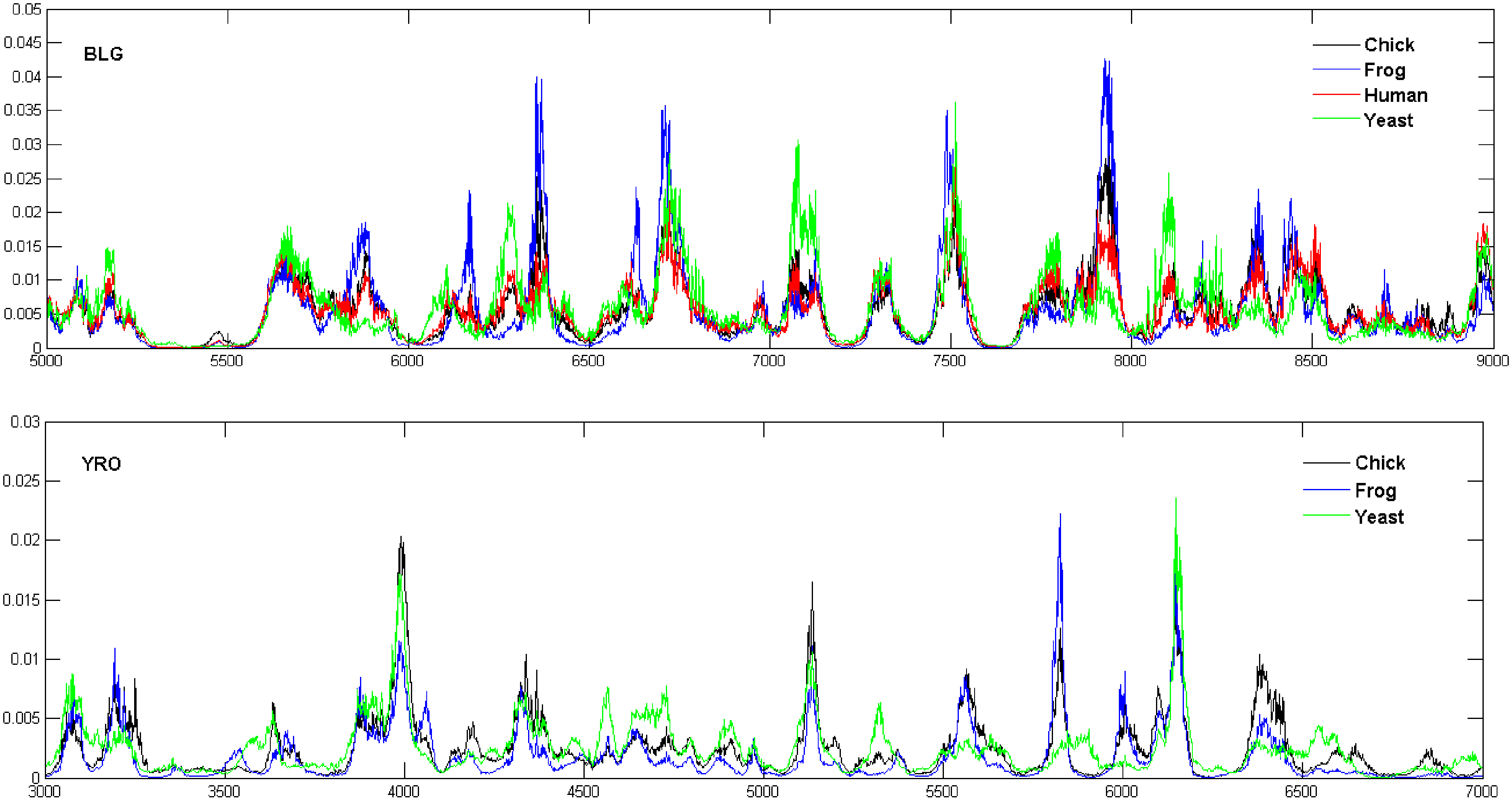
TSS

Gershenzon, Drosophila, 2006

Nucleosomes around transcription start sites (Drosophila)

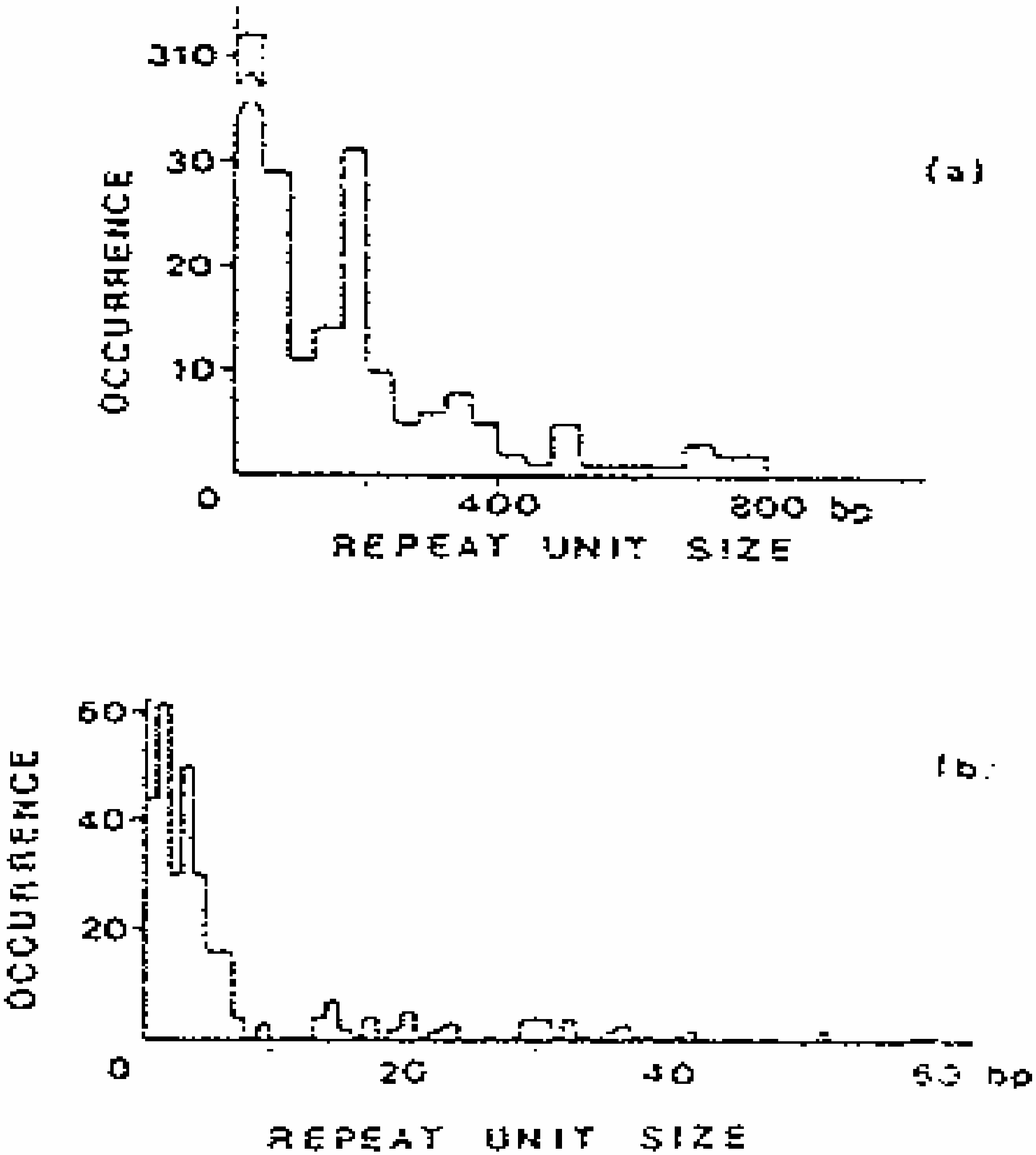


TSS



Species-specificity of nucleosome positioning
Allan et al. JMB, 2010

Modulation
(fast adaptation)
code


(a)
310
30
20
10
0
400
800 bp
OCCURRENCE
REPEAT UNIT SIZE
(b)
60
40
20
0
20
40
60 bp
OCCURRENCE
REPEAT UNIT SIZE

MODULATION OF TRANSCRIPTION

Unit / No. of repeats / location / reference

A 20-55 upstream of *ADR2* gene of *S. cerevisiae* Nature 304, 652, 1983

T 11-45 upstream of *Dictyostellium* actin genes NAR 22, 5099, 1994

T 9-42 Gcn4-activated transcription, *his3* gene, yeast EMBO J 14, 2570, 1995

T 10-80 upstream, vaccinia virus late promoters JMB 210, 771, 1989

GT 30-130 *CAT* constructs, monkey, human cells MCB 4, 2622, 1984

RY 94,144 mouse *ADH1* gene, first intron Gene 57, 27, 1987

ACCGA 5-12 UAS1 site of yeast *CYC1* gene MCB 6, 4690, 1986

CTTCC 2,3 upstream activator of yeast *PGK* gene NAR 16, 8245, 1988

AARKGA 2-8 human IFN beta gene, PRDI element Science 236, 1237, 1987; EMBO J 8, 101, 1989

ATCTTTC 15-28 Between promoters P2 and P1 of adhesin genes of *H. influenzae,* PNAS 96, 1077, 1999

AGGGCAGAGC 1-3 mouse •DRE element, •-globin promoter MCB 10, 972, 1990

GGGGCGGGGC 1,2 Sp1 sites, adenovirus early promoter JBC 266, 20406, 1991

CAAAAATGCC 9-35 transient expression of galactokinase BBRC 180, 1273, 1991

11 bp 1-4 mouse metallothionein I gene, MREa element, MCB 5, 1480, 1985

12 bp 1,3 bovine papilloma virus, E2 site EMBO J 7, 525, 1988

12 bp 1-4 human IFN beta gene, PRDII element EMBO J 8, 101, 1989

12 bp 1-6 MRE element of mouse metallothionein-I promoter, Nature 317, 828, 1985

14 bp 1-4 soybean heat shock promoter element JMB 199, 549, 1988

14 bp 1-4 C. elegans HS element in mouse cells MCB 6, 3134, 1986

14 bp 1-4 Drosophila HS element in yeast cells NAR 14, 8183, 1986

14 bp 1-5 cell-cycle dependent transcription of the yeast *HO* gene, Cell 42, 225, 1985

16 bp 1,5 human oligoA synthetase gene EMBO J 7, 411, 1988

17 bp 1,3 yeast allantoate permease gene, GATAA containing element, MCB 9, 602, 1989

17 bp 1-8 SV40-rat construct, preproinsulin gene MCB 8, 2737, 1988

17 bp 1,5 yeast allantoate permease gene MCB 9, 602, 1989

18 bp 1-5 immediately early genes, human cytomegalovirus, JV 63, 1435, 1989

31 bp 1-8 NF-•B factor binding site upstream of mouse beta-globin gene, JMB 214, 373, 1990

32 bp 1,2 yeast allantoate permease gene MCB 9, 602, 1989

32 bp 1,2 immediately early genes, human cytomegalovirus, JV 63, 1435, 1989

32 bp 1-4 upstream of the *SUC2* gene of *S. cerevisiae,* MCB 6, 2324, 1986

39 bp 1,2 copper-induced transcription of yeast copper-metallothionein gene, MCB 6, 1158, 1986

57 bp 1-4 H element, Ty1 transposon, yeast *CYC7* MCB 8, 5299, 1988

60 bp 1-3 cauliflower mosaic virus activator EMBO J 7, 1589, 1988

113 bp n expression of a reporter gene Gene 189, 13, 1997

122 bp 1-4 maize streak virus activator element EMBO J 7, 1589, 1988

240 bp n rDNA spacer in Drosophila NAR 10, 7017, 1982; PNAS 85, 5508, 1988; MCB 10, 4667, 1990

ENHANCERS

Unit / No. of repeats / location / reference

12 bp 1-3 SV40 constructs expressing E2 peptide of bovine papilloma virus, EMBO J 7, 525, 1988

12 bp 2-6 ftz-dependent enhancer, Drosophila Nature 336, 744, 1988

14 bp 1,2 phorbol ester induction, HIV, R region MCB 7, 3994, 1987

16 bp 1,5 interferon-responsive, *tk* gene constructs, transfected monkey cells, EMBO J 7, 1411, 1988

17 bp 1,2 yeast upstream activator sequence, in HeLa cells, Cell 52, 169, 1988

17 bp 1,4 CRE enhancer of human vasoactive intestinal peptide gene, PNAS 85, 6662, 1988

18 bp 1,2 cAMP responsive, human glycoprotein hormone, MCB 7, 3759, 1987

20 bp 4,8 core of SV40 enhancer, constructs JMB 201, 81, 1988

30 bp 11-21 EBV transcription and replication MCB 6, 3838, 1986

50 bp 1-6 herpes virus saimiri JMB 201, 81, 1988

57 bp 1-4 H element of Ty1 transposon, *CYC7* gene MCB 8, 5299, 1988

60 bp n rDNA spacer, *X. laevis* Cell 35, 449, 1983

68 bp 1-3 BKV transcription Science 222, 749, 1983

72 bp 1-3 SV40, constructs JV 55, 823, 1981

81 bp n rDNA spacer, *X. laevis* Cell 35, 449, 1983

99 bp 1,2 murine Akv retrovirus JV 64, 3185, 1990

109 bp 1,2 MCF virus, oncogenicity JV 63, 1284, 1989

140 bp 1-13 mouse rRNA gene spacer PNAS 87, 7527, 1990

OTHER ACTIVITIES

Unit / No. of repeats / location / reference

A 17-20 promoter region, *Mycoplasma* surface antigen variation, EMBO J 10, 4069, 1991

C 8-44 5'-UTR, virulence of mengovirus JV 70, 2027, 1996

GT n recombination, mouse somatic cells MCB 6, 3948, 1986

GT n recombination, Rec A binding JMB 273, 105, 1997

GT n meiosis, yeast MCB 6, 3934, 1986

CG n recombination, mouse somatic cells MCB 6, 3948, 1986

AAG 2-8 exon M2 of mouse IG• gene, enhancement of splicing, MCB 14, 1347, 1994

GACA 22-35 phenotypic switching of a lypopolysaccharide epitope, PNAS 93, 11121, 1996

AAGTGA 4-8 upstream inducible element, human beta interferon gene, JV 64, 3063, 1990

GAAAGT 2,4 mediates virus-inducible transcription of human interferon genes, PNAS 88, 1369, 1991

ATAGTAAA 13,17 iteron in plasmid pAD1 of *E. faecalis*, mating response to sex pheromone, J Bact 177, 5453, 1995

CTGAGGTCAA 1-5 F2 half-element of chicken lysozyme silencer S-2.4 kb, Cell 61, 505, 1990

14 bp 1-5 3'-terminal UTR, tobacco vein mottling virus, disease symptom severity, PNAS 88, 9863, 1991

17 bp 1-8 modulation of translation, rat preproinsulin, MCB 8, 2737, 1988

31 bp 1-6 packaging of Adenovirus Type 5 DNA JV 64, 2047, 1990

40 bp 1,2 polyoma virus expression JV 62, 3896, 1988

46 bp 1-4 virus-responsive element of IFN•1 promoter, induced expression, Cell 50, 1057, 1987

48 bp 2,5 transforming activity of a retrovirus NAR 26, 4868, 1998

68 bp 1-3 BK virus, transforming activity JV 55, 867 & 823, 1985

240 bp 13-350 modulation of meiotic drive, Rsp of SD system of *Drosophila* Nature 332, 394, 1988; Cell 54, 179, 1988

TG 20-30 regulation of period in circadian rhythm Science 278, 2117, 1997

SKQPFRK 2-7 chloroplast ribosomal protein S18 FEBS Let 279, 190, 1991

YSPTSPS 9-26 yeast RNApolII, modulation, response to enhancer signals Nature 347, 491, 1990; MCB 8, 321, 1988

YSPTSPS 3-78 mouse RNApolII, modulation MCB 8, 330, 1988

12 aa 7-11 Mycoplasma surface antigen variation EMBO J 10, 4069, 1991

31 aa 3,4 stage- and tissue specificity of human microtubule-associated protein tau, EMBO J 8, 393, 1989

34 aa 0-17 plant resistance to bacterial spot disease, Nature 356, 172, 1992

42 aa 3-13 segment polarity armadillo gene, *Drosophila*, phenotypic series, Cell 63, 1167, 1990

53 aa 11-50 kringle IV, processing and secretion of apolipoprotein (a), JBC 271, 32403, 1996

82 aa 1-9 alpha C protein, *Streptococci*, modulation of host immunity, PNAS 93, 4131, 1996

Diseases with repeats in non-coding regions

	Triplet	**n** in norm/pathology
FRAXA (fragile X syndrome)	CGG	6-53/230+
FXTAS (FRAXA associated tremor/ataxia syndrome)	CGG	6-53/55-200
FRAXE (fragile XE mental retardation)	GCC	6-35/200+
FRDA (Friedreich’s ataxia)	GAA	7-34/100+
DM (myotonic dystrophy)	CTG	5-37/50+
SCA8 (spinocerebellar ataxia Type 8)	CTG	16-37/110-250

from Wikipedia

…GCUGCUGCUGCUGCU…

…AGCAGCAGCAGCAGC…

this is
GCU repeat,
but also CUG repeat,
UGC repeat,
AGC repeat,
GCA repeat,
and CAG repeat

Diseases with repeats in non-coding regions

	Triplet	n in norm/pathology
FRAXA (fragile X syndrome)	CGG GCC	6-53/230+
FXTAS (FRAXA associated tremor/ataxia syndrome)	CGG GCC	6-53/55-200
FRAXE (fragile XE mental retardation)	GCC GCC	6-35/200+
FRDA (Friedreich’s ataxia)	GAA GAA	7-34/100+
DM (myotonic dystrophy)	CTG GCU	5-37/50+
SCA8 (spinocerebellar ataxia Type 8)	CTG GCU	16-37/110-250

Polyglutamine diseases (polyCAG = polyGCU)

		n in norm/pathology
DRPLA	(dentatorubropallidoluysian atrophy)	6-35/49-88
HD	(Huntington’s disease	10-35/35+
SBMA	(spinobulbar muscular atrophy)	9-36/38-62
SCA1	(spinocerebellar ataxia Type 1)	6-35/49-88
SCA2		14-32/33-77
SCA3		12-40/55-86
SCA6		4-18/21-30
SCA7		7-17/38-120
SCA17		25-42/47-63

from Wikipedia

Tandem repeat expansion diseases and disorders

Repeat/Copy number ***n*** range/Location/Disease or disorder/References

(3 bp/1 aa) *n* 5 to over 200 5′-, 3′- and over coding regions
15 different neurodegenerative and other diseases Usdin and Grabczyk, 2000 Brais et al., 1998 Delot et al., 1999

(4 bp) ***n*** 75 to 11.000 intron 1 of *ZNF9* myotonic dystrophy gene type 2 Liquori et al., 2001

(5 bp) ***n*** 10 to 4.500 intron 9 of *SCA10* gene type 10 spinocerebellar ataxia Matsuura et al., 2000

(12 bp) ***n*** 2 to over 60 5′ from cystatin B gene progressive myoclonus epilepsy Lalioti et al., 1997

(14 bp) ***n*** 40 to 150 5′ from insulin gene type 1 susceptibility to diabetes Bennett et al., 1995, Kennedy et al., 1995

(15 bp) and (18 bp) ***n*** few to 90 5′ from cystatin B gene progressive myoclonus epilepsy Virtaneva et al., 1997

(24 bp/8 aa) ***n*** 5 to 34 coding region of the prion protein gene Creutzfeldt-Jakob disease Cochran et al., 1996

(28 bp) ***n*** 30 to 100 3′ from *HRAS1* proto-oncogene ovarian cancer risk Phelan et al., 1996

(342 bp/114 aa) ***n*** 15 to 37 apo(a) coding region Lp(a) level, susceptibility to atherosclerosis and thrombosis, Lindahl et al., 1990, Koschinsky et al., 1990

(3200 bp)***n*** 2 to 100 *FSHD* gene region FSHD muscular dystrophy van Deutekom et al., 1993

There is only few percent difference between genomes of human and chimpanzee.
Mostly in copy numbers of simple repeats.

PROTEOMIC CODE

(PROTEIN SEQUENCE MODULES)

Two related sequences, aligned

33% match

```
Q816J5
DVNLPKFDGFYWCRQIRHESTCPIIFISARAGEMEQIMAIESGADDYITKPFHYDVVMAKIKGQLRR
|||||-|||----|--|--|----------------------||||---|||------|-----|||
DVNLPGIDGWDLLRRLRERSSARVMMLTGHGRLTDKVRGLDLGADDFMVKPFQFPELLARVRSLLRR
Q7DCC5
```

```
CPIIFISARAGEMEQIMAIE Q816J5 Two-component response regulator B. cereus
 ||||||||   | | ||||
VPIIFISARDSDMDQVMAIE Q97IX4 Response regulator                C. acetobutylicum
|| ||||||| | | |   |
VPVIFISARDADIDRVLGLE O32192 Transcr. regulatory protein cssR B. subtilis
||  | ||||  ||||||||
VPILFLSARDEEIDRVLGLE Q89D26 Two-component response regulator B. japonicum
 ||  | || || | |||||
IPIIMLTARSEEFDKVLGLE Q8R9H7 Response regulators               Th. tengcongensis
  | ||||||   ||| |||
SRIMMLTARSRLADKVRGLE Q88RT2 heavy metal response regulator   Ps. Putida
 | ||||   || ||||||
ARVMMLTGHGRLTDKVRGLD Q7DCC5 Two-component response regulator Ps. Aeruginosa
```

```
Q816J5 Two-component response regulator
DVNLPKFDGFYWCRQIRHESTCPIIFISARAGEMEQIMAIESGADDYITKPFHYDVVMAKIKGQLRR
|||||-|||----|--|--|----------------------||||---|||------|-----|||
DVNLPGIDGWDLLRRLRERSSARVMMLTGHGRLTDKVRGLDLGADDFMVKPFQFPELLARVRSLLRR
Q7DCC5 Probable two-component response regulator
```

No-match relatives

```
LEVALALSQADIIVRDALVS Q8UBQ7 Uroporphyrin-III C-methyltransferase          A. tumefaciens
|  | || ||| || ||||
LHAANALRQADVIVHDALVN Q92P47 probable Uroporphyrin-III C-methyltransferase   Rh. meliloti
| |   |  ||||||||||
LRAQRVLMEADVIVHDALVP Q8YEV9 Uroporphyrin-III C-methyltransferase          B. melitensis
||| | ||||||||||||||
LRAHRLLMEADVIVHDALVP Q98GP6 Siroheme synthase (precorrin methyltransferase) Rh. loti
|   ||| |||||
LKGQRLLQEADVILYADSLV Q8DLD2 Precorrin-4 C11-methyltransferase             S. elongatus
 ||||   ||||| || |||
IKGQRIVKEADVIIYAGSLV Q8REX7 Precorrin-4 C11-methyltransferase             F. nucleatum
 ||||      |||||||||
VKGQRLIRQCPVIIYAGSLV Q88HF0 Precorrin-4 C11-methyltransferase             Ps. putida
| |  ||  |||  ||||||
VRGRDLIAACPVCLYAGSLV Q8UBQ5 Precorrin-4 C11-methyltransferase             A. tumefaciens
```

```
Q8UBQ7 methyltransferase
HVWLAGAGPGDVRYLTLEVALALSQADIIVRDALVS
-|---|||||-----|--------------------
TVHFIGAGPGAADLITVRGRDLIAACPVCLYAGSLV
Q8UBQ5 methyltransferase
```

No-match relatives

Methyltransferases

```
LEVALALSQADIIVRDALVS Q8UBQ7
|  | || ||| || ||||
LHAANALRQADVIVHDALVN Q92P47
| |   |  ||||||||||
LRAQRVLMEADVIVHDALVP Q8YEV9
||| | ||||||||||||||
LRAHRLLMEADVIVHDALVP Q98GP6
|   ||| |||||
LKGQRLLQEADVILYADSLV Q8DLD2
 ||||   ||||| || |||
IKGQRIVKEADVIIYAGSLV Q8REX7
 ||||      |||||||||
VKGQRLIRQCPVIIYAGSLV Q88HF0
| |  ||  |||  ||||||
VRGRDLIAACPVCLYAGSLV Q8UBQ5
```

No-match relatives

LEVALALSQADIIVRDALVS	Q8UBQ7
VRGRDLIAACPVCLYAGSLV	Q8UBQ5

To be related

the sequences

do not have to be similar

(upto even complete mismatch)

Existing most advanced
sequence alignment techniques
(e. g. BLAST)
would not be able to qualify
such fully dissimilar sequences
as relatives

unless many intermediate sequences
are analyzed
(that amounts to a whole research project)

One can make long

walks

from fragment to fragment in the

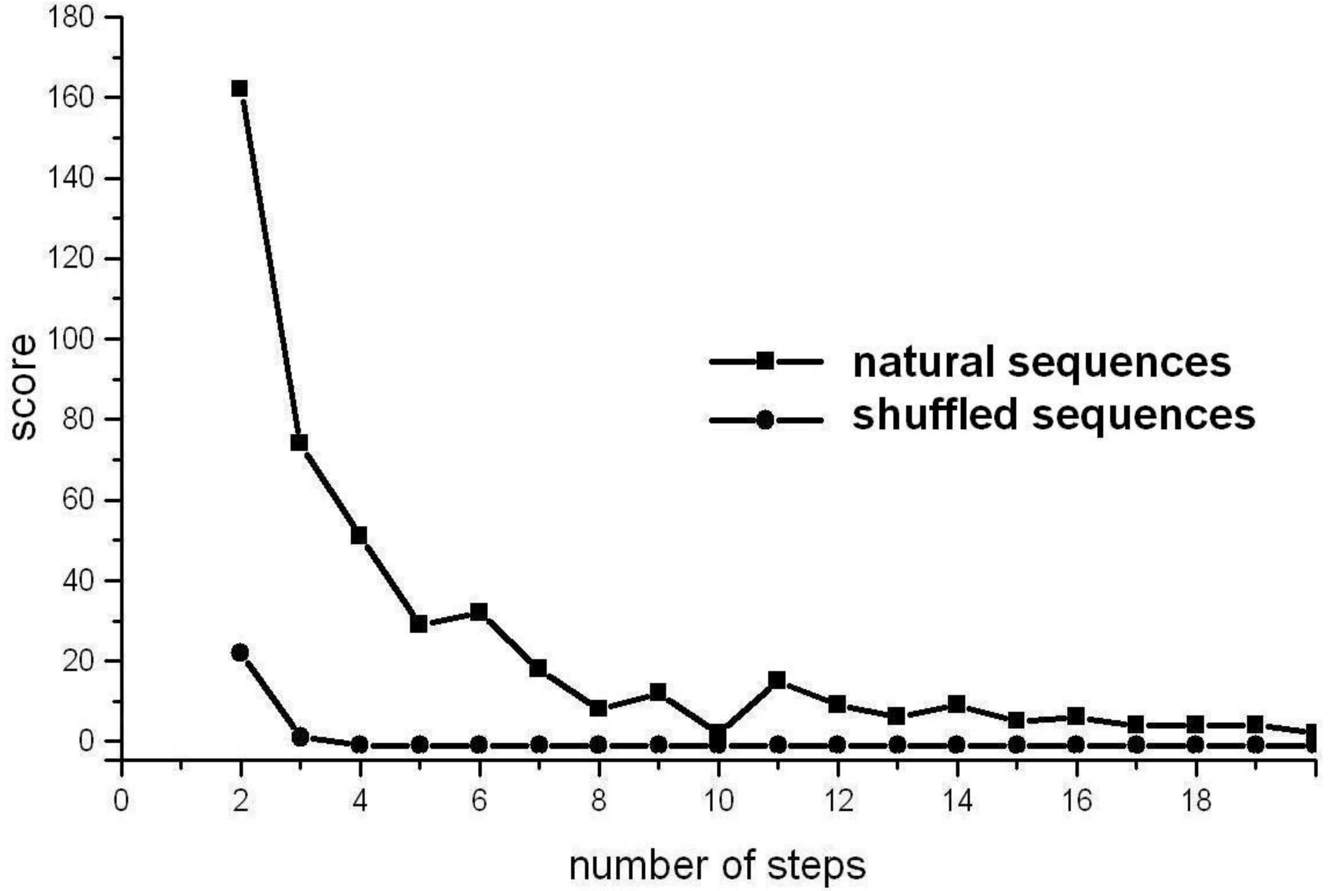
formatted protein sequence space

(sequence fragments of the same length, 20 residues, gathered from all or many proteomes)

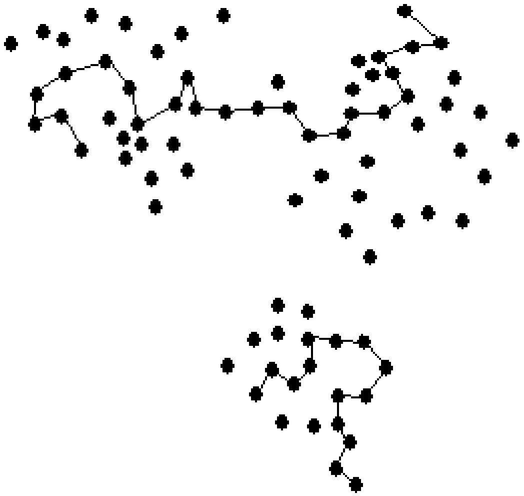
Pair-wise connected matching fragments make also

networks

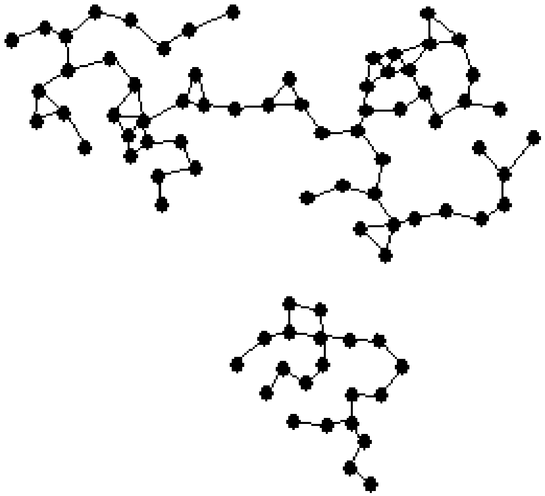
Natural sequence space has longer walks
than random sequence space of the same size



WALK



NETWORK



Frenkel, 2006

60% match threshold networks:

320,000 proteins from 120 prokaryotes, ~100,000,000 fragments

The largest (monster) network 9,368,905 sequence fragments (~10% of all)

Next largest 2,535 fragments

Networks of sizes 120 to 2,535 fragments (several thousand, 3.8% of all fragments)

Small networks cover 86% of the space

35% of fragments are single, no relatives

Number of different fragments in complete (random) space:

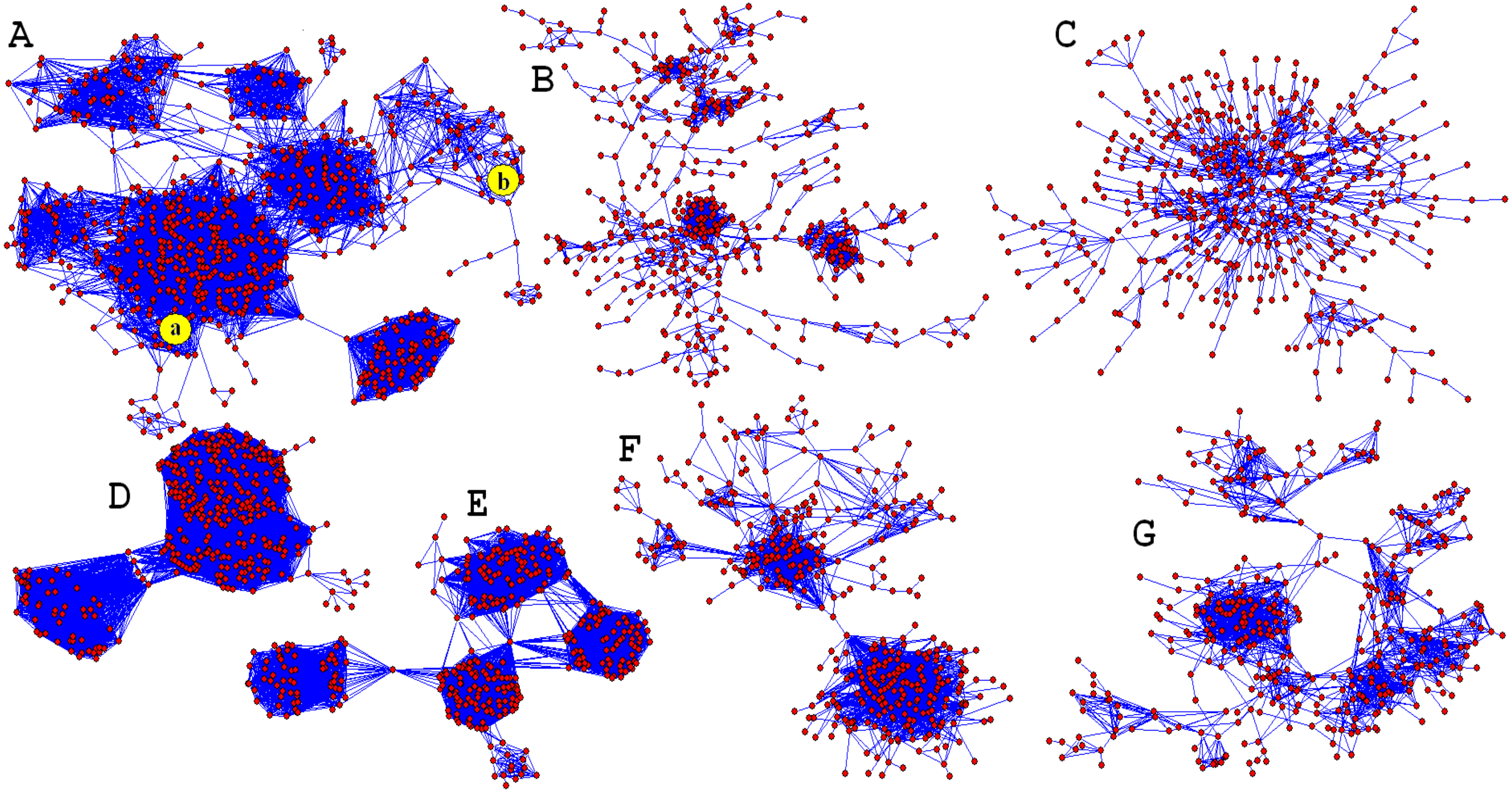
$20^{20} \sim 10^{26}$

Number of fragments in complete natural space:

$10^{7} \bullet 3 \bullet 10^{4} \bullet 300 \sim 10^{14}$

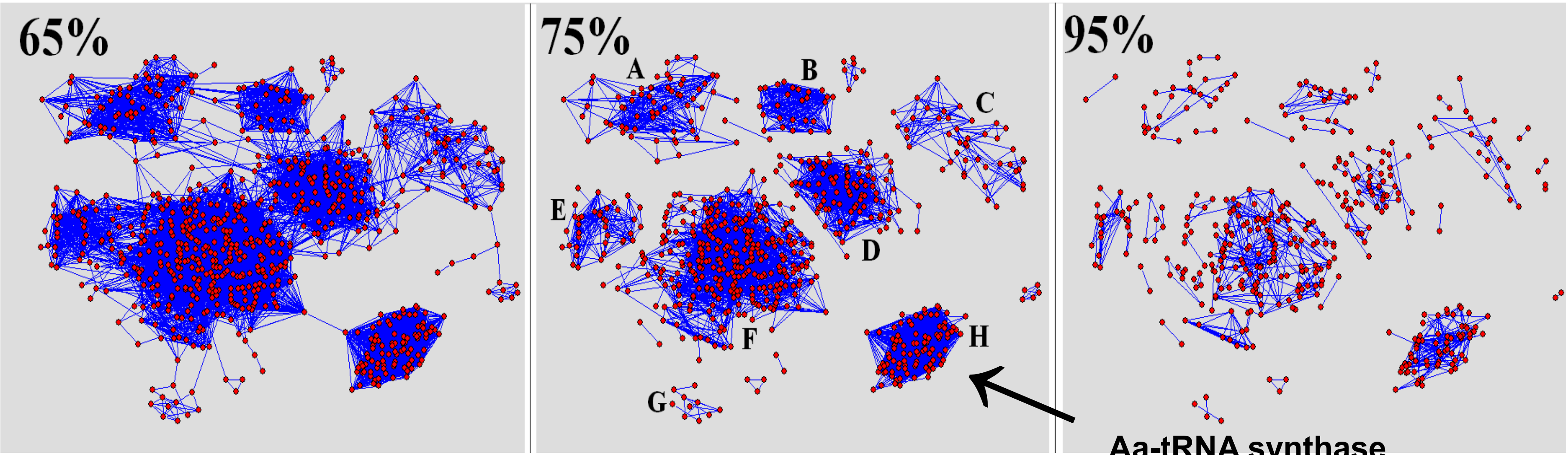
Probability that a given fragment in natural space

is randomly generated is 10^{-12}


A
B
C
a
b
D
E
F
G

Networks of fragments of aa-tRNA synthetases

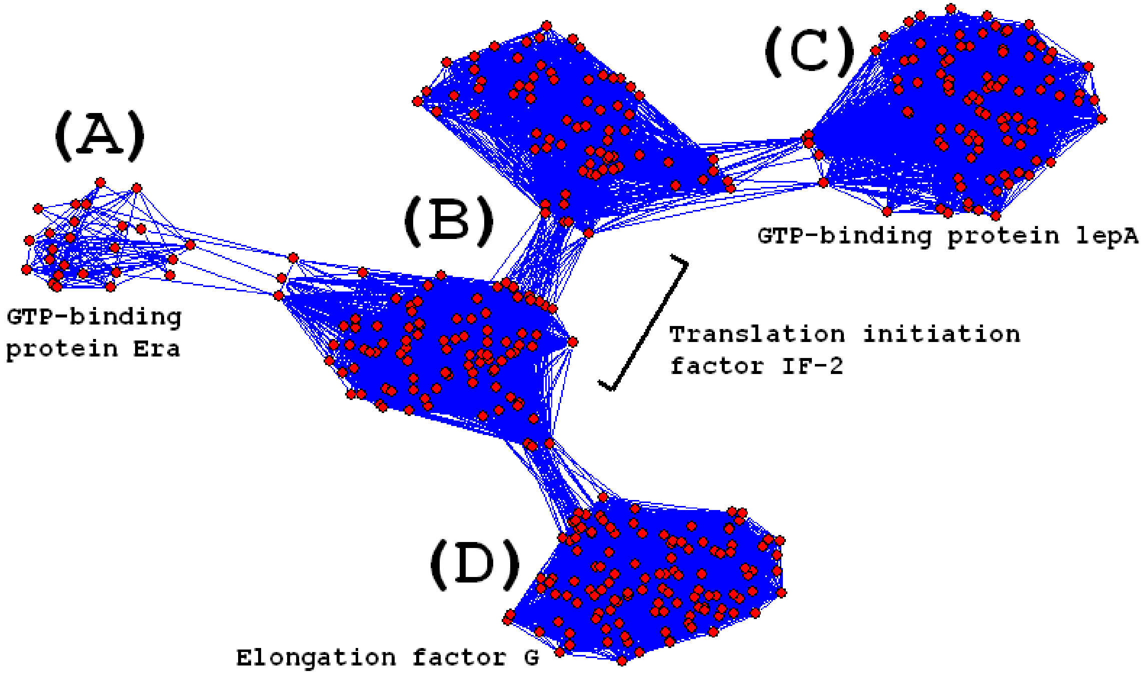
at various thresholds of sequence match



Aa-tRNA synthase module of lepA

A tyr trp B met C arg trp D cys
E leu F met leu ile val G ile H lepA

Network of GTP binding proteins



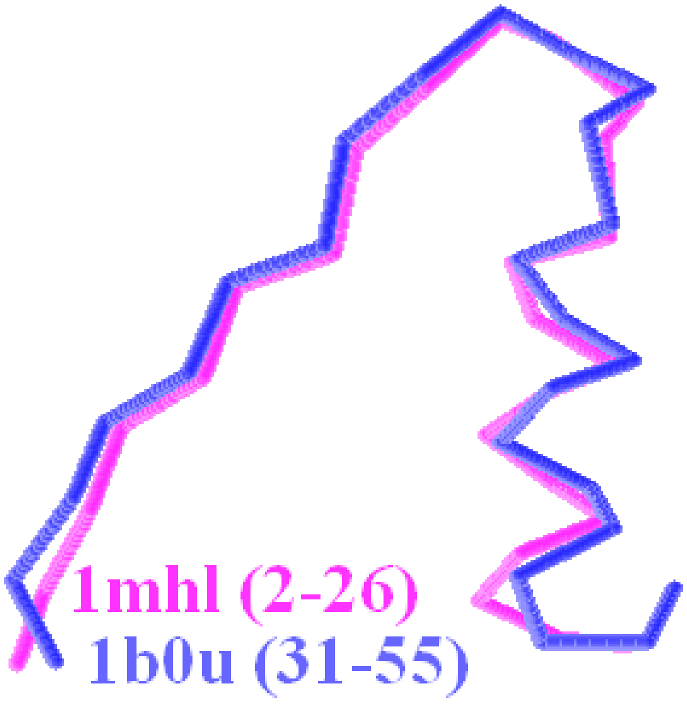
Sequence fragments with the same function
are found in the same network

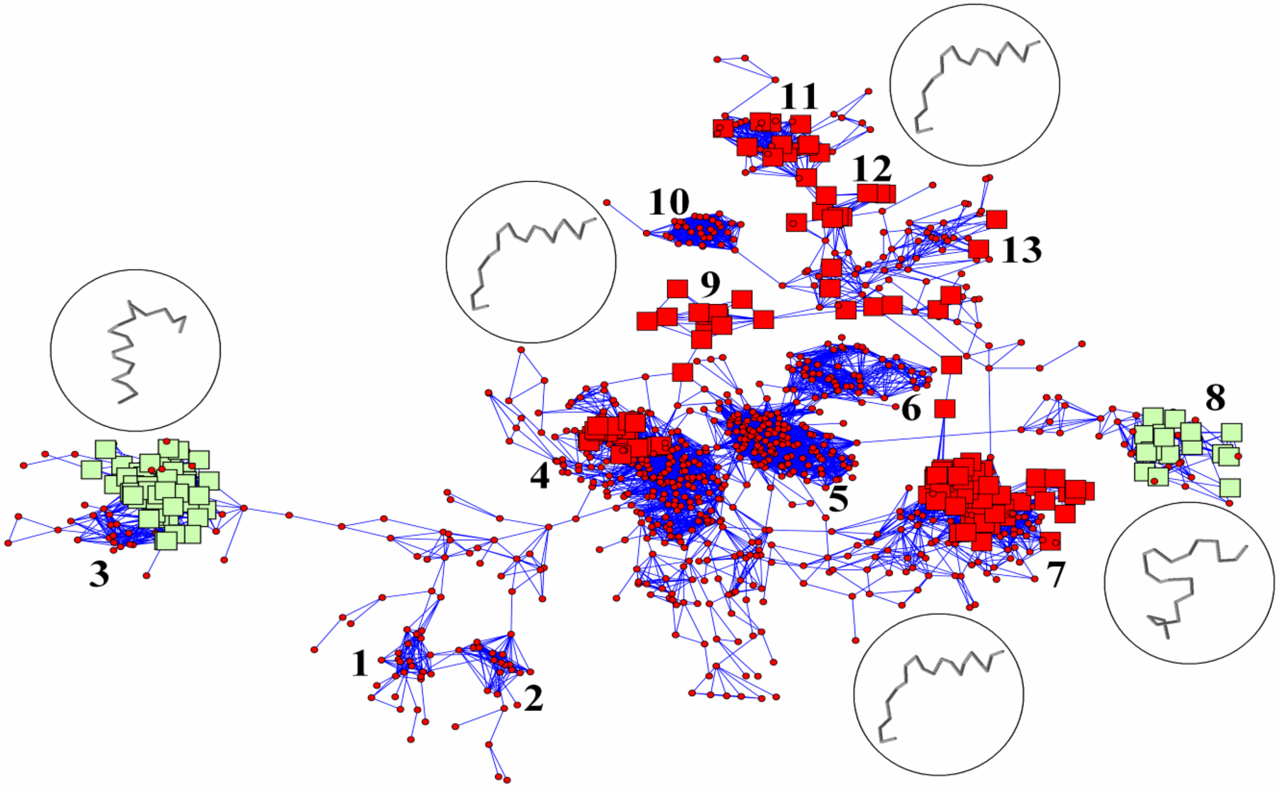
1mh1_ c.37.1.8 Rac (GTP-binding)
{Human (Homo sapiens)}

```
2                       26
QAIKCVVVGDGAVGKTCLLISYTTN
        |    ||   |
AGDVISIIGSSGSGKSTFLRCINFL
31                      55
```

1b0ua_ c.37.1.12 (A:) ATP-binding subunit
of the histidine permease
{Salmonella typhimurium}

Fig. 2



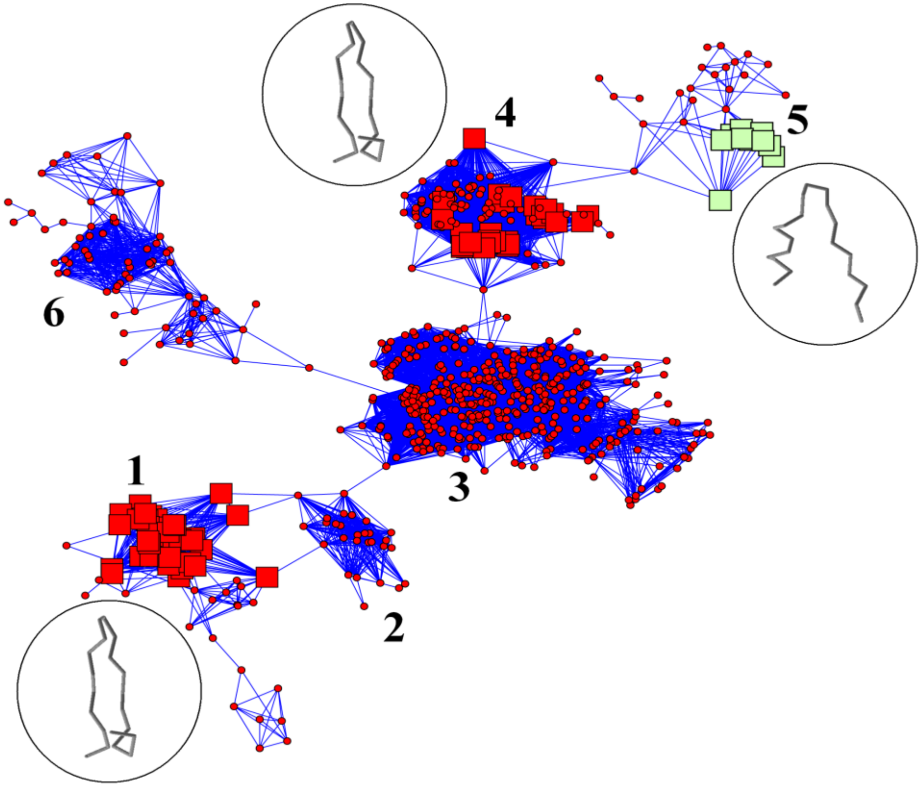


1 Putative peptidoglycan bound protein
2 Collagen adhesion protein
3 Ribosomal protein L11
4 Penicillin-binding protein 2x
5 Penicillin-binding protein 1
6 Penicillin binding protein 2A
7 D-alanyl-D-alanine carboxypeptidase

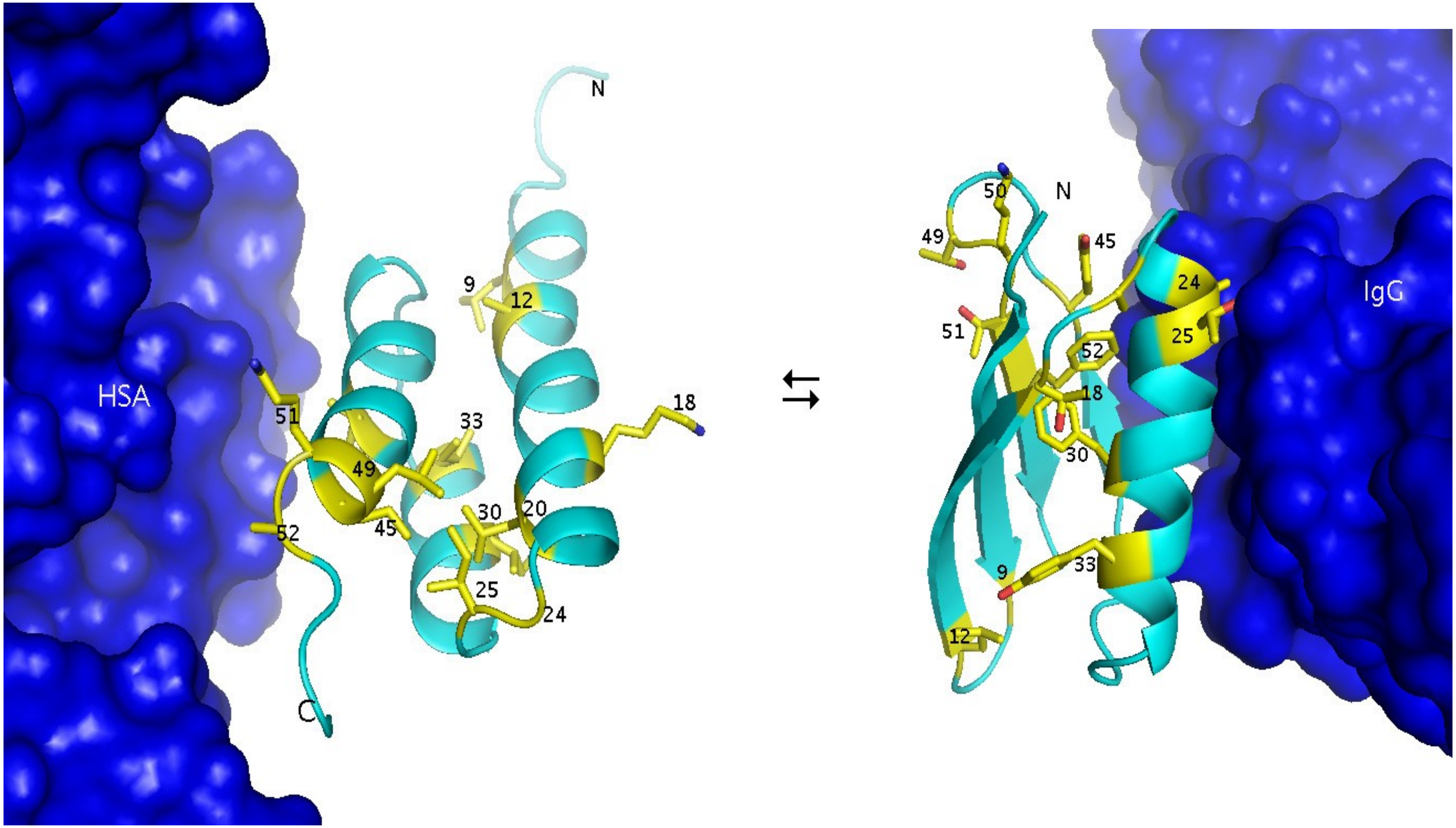
8 cytochrome

9 Beta-Lactamase
10 Mannitol-1-phosphate 5-dehydrogenase
11 glutaminase
12 Beta-lactamase
13 Esterase EstB

Fragments of the same network
have, essentially, the same structure.
Periferal fragments may be different


4
5
6
3
1
2

Two alternative structures with the same sequence



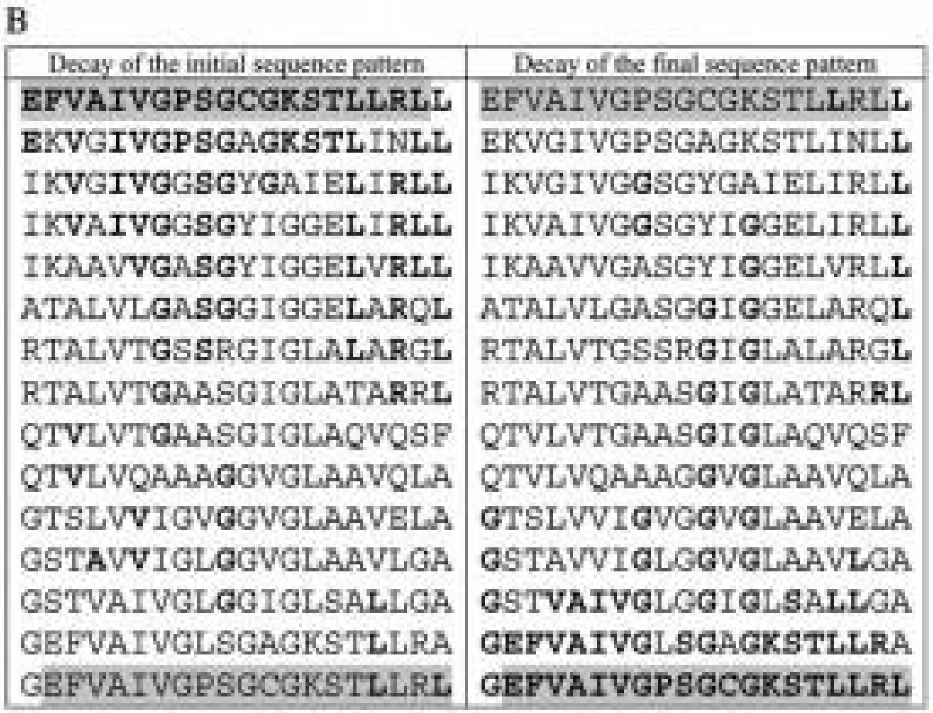
Lab of P. N. Bryan, 2009

New definition of sequence relatedness:

fragments of the same network
are relatives

	Decay of the initial sequence pattern (bottom up)	Decay of the final sequence pattern (bottom up)	Every two nearest neighbors share at least 60% identity
1	LED**A**IKA**A**KAGA**D**IIMLDNM	**LEDAIKAAKAGADIIMLDNM**	LEDAIKAAKAGADIIMLDNM
2	PED**A**PRA**A**DAGA**D**IV**L**LDNM	P**EDA**PR**AA**D**AGADI**VL**LDNM**	PEDAPRAADAGADIVLLDNM
3	PEA**A**ERA**A**ATGA**DG**VGLLRM	P**E**A**A**ER**AA**AT**GAD**GVG**L**LR**M**	PEAAERAAATGADGVGLLRM
4	PEA**A**RKA**A**ATGA**DG**VGLLRT	P**E**A**A**R**KAA**AT**GAD**GVG**L**LRT	PEAARKAAATGADGVGLLRT
5	PAD**A**RAARAFGAE**G**IGLCRT	PA**DA**RA**A**RAF**GA**EG**I**G**L**CRT	PADARAARAFGAEGIGLCRT
6	PTDFKKALLFGAE**G**VGLCRT	PT**D**FK**KA**LLF**GA**EGVG**L**CRT	PTDFKKALLFGAEGVGLCRT
7	PLDIIKALVLGAKAVGLSRT	PL**D**I**IKA**LVL**GA**KAVG**L**SRT	PLDIIKALVLGAKAVGLSRT
8	GTDIIKALAIGANLVGL**G**RM	GT**D**I**IKA**LAI**GA**NLVG**L**GR**M**	GTDIIKALAIGANLVGLGRM
9	GTDIVKAIAAGA**D**LVGI**G**RL	GT**D**IV**KA**IA**AGAD**LVGIGRL	GTDIVKAIAAGADLVGIGRL
10	**S**GDIAKAIAAGA**D**AVML**G**SL	SG**D**IA**KA**IA**AGAD**AV**ML**GSL	SGDIAKAIAAGADAVMLGSL
11	IGLIEKAKAEGA**D**AVIL**G**CT	IGLIE**KA**KAE**GAD**AVI**L**GCT	IGLIEKAKAEGADAVILGCT
12	KRLVEIAKLEGA**D**AICH**G**CT	KRLVEI**A**KLE**GAD**A**I**CHGCT	KRLVEIAKLEGADAICHGCT
13	ARIVEIAKACGA**D**AIHP**G**YG	ARIVEI**A**KAC**GAD**A**I**HPGYG	ARIVEIAKACGADAIHPGYG
14	EKIIAAAKASGAEAIHP**G**YG	EKIIAA**A**KAS**GA**EA**I**HPGYG	EKIIAAAKASGAEAIHPGYG
15	EKLLAVAKRSGA**D**AVHP**G**YG	EKLLAV**A**KRS**GAD**AVHPGYG	EKLLAVAKRSGADAVHPGYG
16	EK**A**LAALESSGA**D**AVMI**G**RG	EKALAALESS**GAD**AV**M**IGRG	EKALAALESSGADAVMIGRG
17	LK**A**RAVLDYTGA**D**ALMI**G**R**A**	LKARAVLDYT**GAD**AL**M**IGRA	LKARAVLDYTGADALMIGRA
18	KK**A**F**E**VLQITQA**DG**LMI**G**R**A**	KKAFEVLQITQ**AD**GL**M**IGRA	KKAFEVLQITQADGLMIGRA
19	Q**NA**K**E**VYKITKC**DG**LMI**G**R**A**	QNAKEVYKITKC**D**GL**M**IGRA	QNAKEVYKITKCDGLMIGRA
20	Q**NA**K**E**ILGIDSV**DG**L**L**I**GSA**	QNAKEILGIDSV**D**GLLIGSA	QNAKEILGIDSVDGLLIGSA
21	**SNA**K**EL**MGVANV**DGAL**I**GGA**	SNAKELMGV**A**NV**D**GALIGGA	SNAKELMGVANVDGALIGGA
	SNAAELFAQPDIDGALVGGA	SNA**A**ELF**A**QPDI**D**GALVGGA	SNAAELFAQPDIDGALVGGA

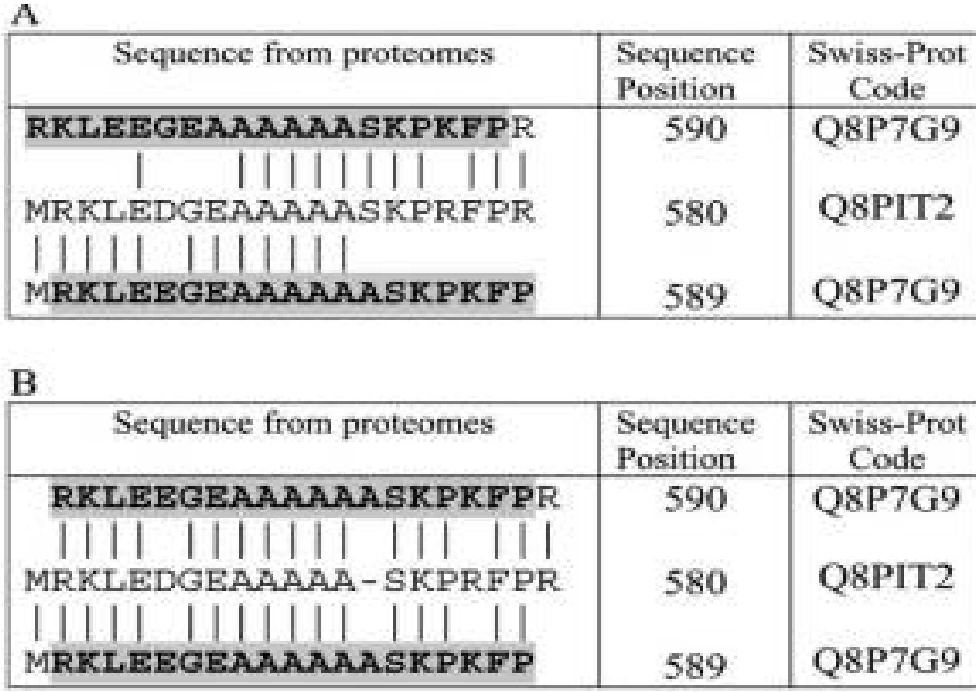
Sequences shifted by one residue may belong to the same network



B

Decay of the initial sequence pattern	Decay of the final sequence pattern
EFVAIVGPSGCGKSTLLRLL	EFVAIVGPSGCGKSTLLRLL
EKVGIVGPSGAGKSTLINLL	EKVGIVGPSGAGKSTLINLL
IKVGIVGGSGYGAIELIRLL	IKVGIVGGSGYGAIELIRLL
IKVAIVGGSGYIGGELIRLL	IKVAIVGGSGYIGGELIRLL
IKAAVVGASGYIGGELVRLL	IKAAVVGASGYIGGELVRLL
ATALVLGASGGIGGELARQL	ATALVLGASGGIGGELARQL
RTALVTGSSRGIGLALARGL	RTALVTGSSRGIGLALARGL
RTALVTGAASGIGLATARRL	RTALVTGAASGIGLATARRL
QTVLVTGAASGIGLAQVQSF	QTVLVTGAASGIGLAQVQSF
QTVLVQAAAGGVGLAAVQLA	QTVLVQAAAGGVGLAAVQLA
GTSLVVIGVGGVGLAAVELA	GTSLVVIGVGGVGLAAVELA
GSTAVVIGLGGVGLAAVLGA	GSTAVVIGLGGVGLAAVLGA
GSTVAIVGLGGIGLSALLGA	GSTVAIVGLGGIGLSALLGA
GEFVAIVGLSGAGKSTLLRA	GEFVAIVGLSGAGKSTLLRA
GEFVAIVGPSGCGKSTLLRL	GEFVAIVGPSGCGKSTLLRL

Formation of shifted self by deletion of repeating residue



A

Sequence from proteomes	Sequence Position	Swiss-Prot Code
RKLEEGEAAAAAASKPKFPR	590	Q8P7G9
ǀ ǀǀǀǀǀǀǀǀ ǀǀǀ		
MRKLEDGEAAAAASKPRFPR	580	Q8PIT2
ǀǀǀǀǀ ǀǀǀǀǀǀǀ		
M**RKLEEGEAAAAAASKPKFP**	589	Q8P7G9

B

Sequence from proteomes	Sequence Position	Swiss-Prot Code
RKLEEGEAAAAAASKPKFPR	590	Q8P7G9
ǀǀǀǀ ǀǀǀǀǀǀǀǀ ǀǀǀ ǀǀǀ		
MRKLEDGEAAAAA-SKPRFPR	580	Q8PIT2
ǀǀǀǀǀ ǀǀǀǀǀǀǀ ǀǀǀ ǀǀ		
M**RKLEEGEAAAAAASKPKFP**	589	Q8P7G9

Careful with consensus!

The words

COOKY

MANGO

MELON

HONEY

SWEET

all suggest something sweet or sweet-sour
and could be considered, thus, as recognition sequences for the 'sweet' quality. Their consensus sequence, however, conveys a rather different message:

MONEY

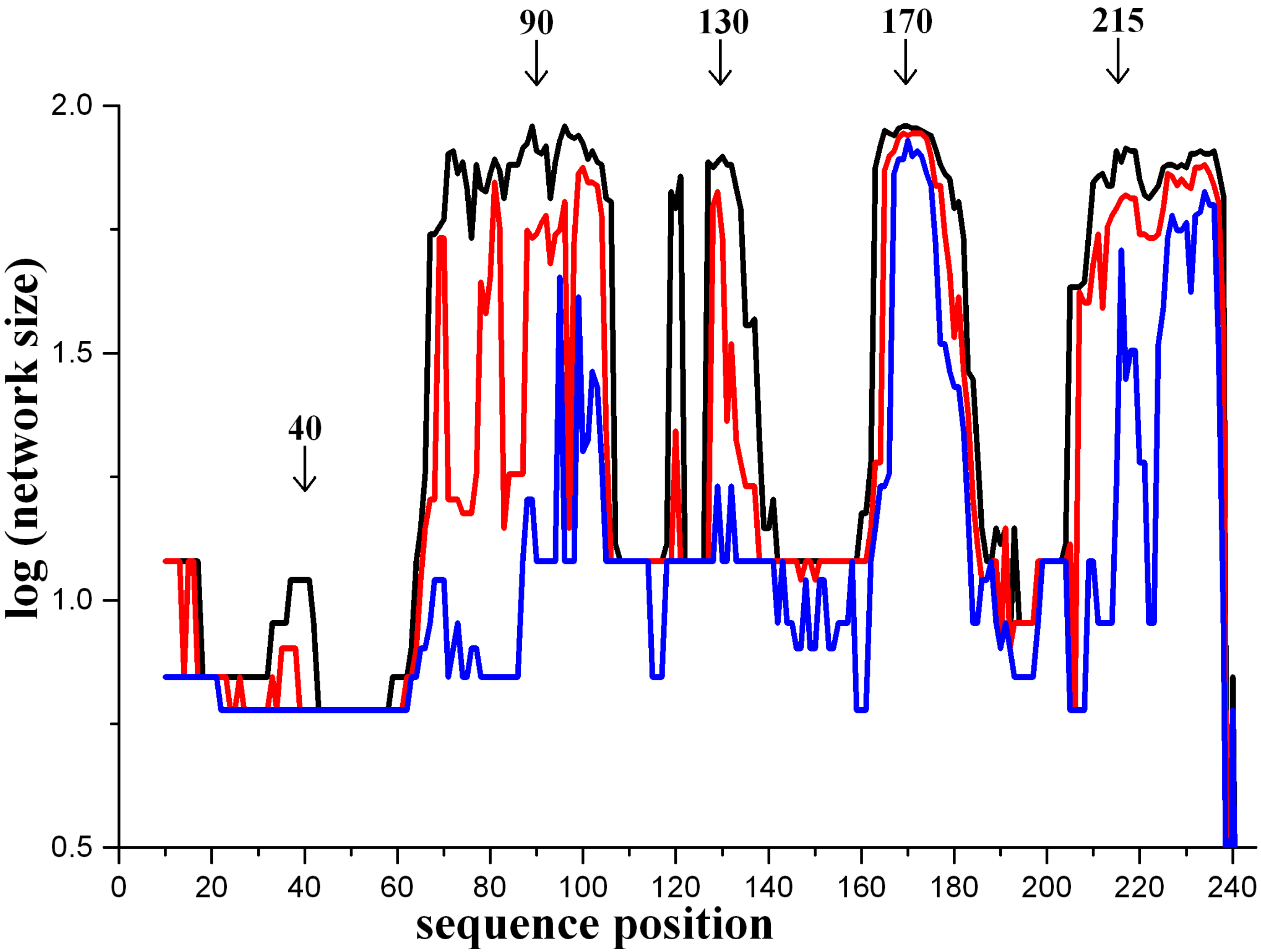
Every fragment
of the precalculated space
is tagged (protein, species)

It is also uniquely located in it s family network.

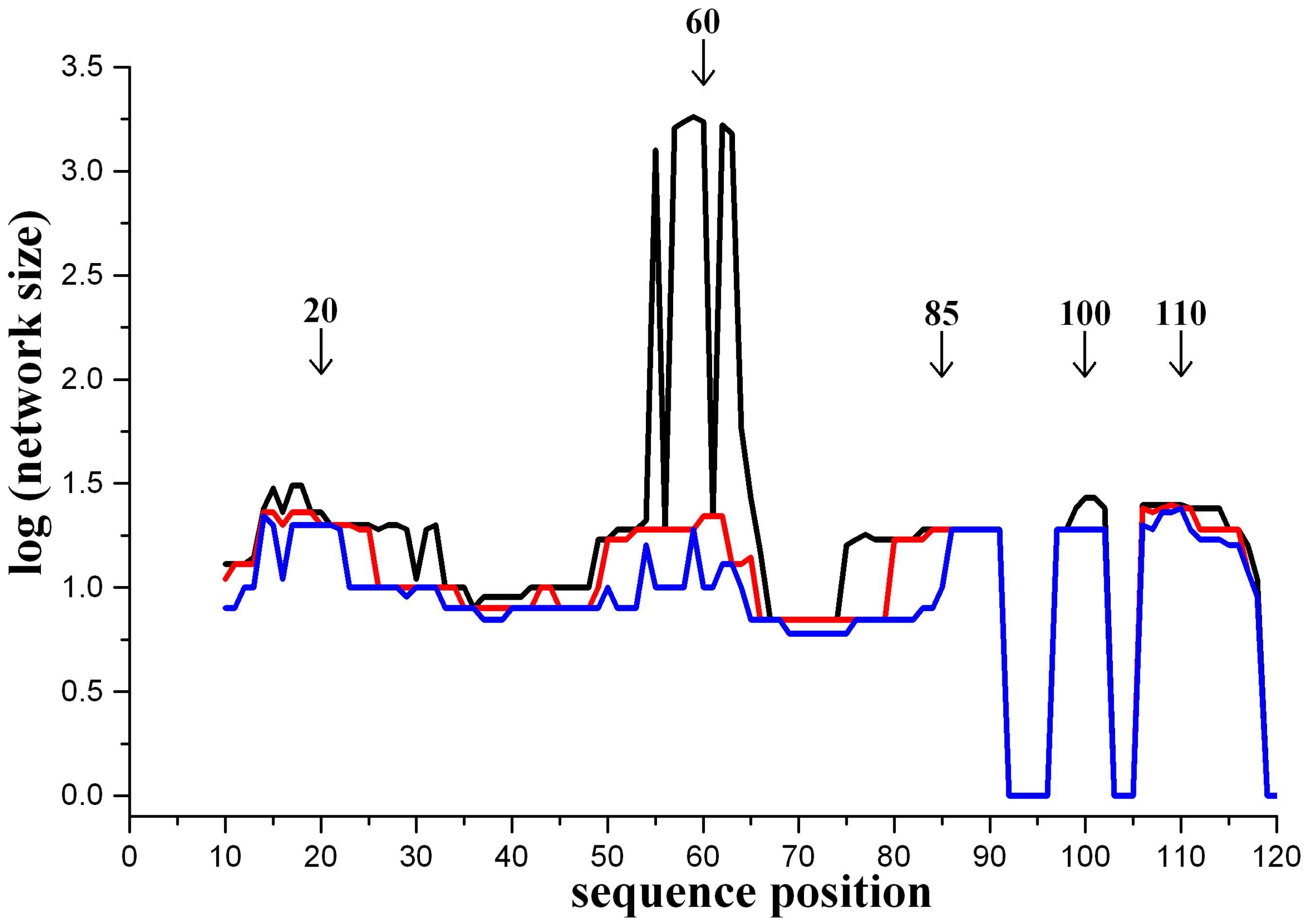
The size of the network says
how many relatives the fragment has

Thus, one can take a sequence
and for all fragments of it
find their networks and plot the sizes

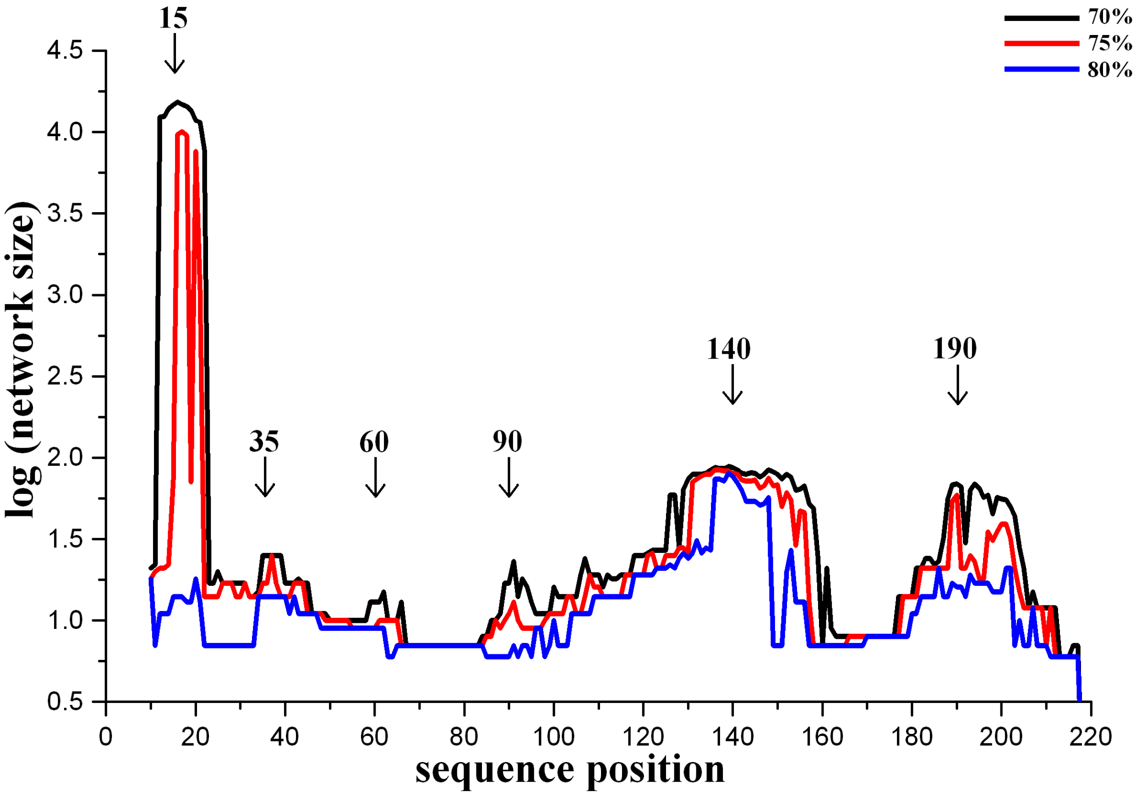
Modules of TIM-barrell protein



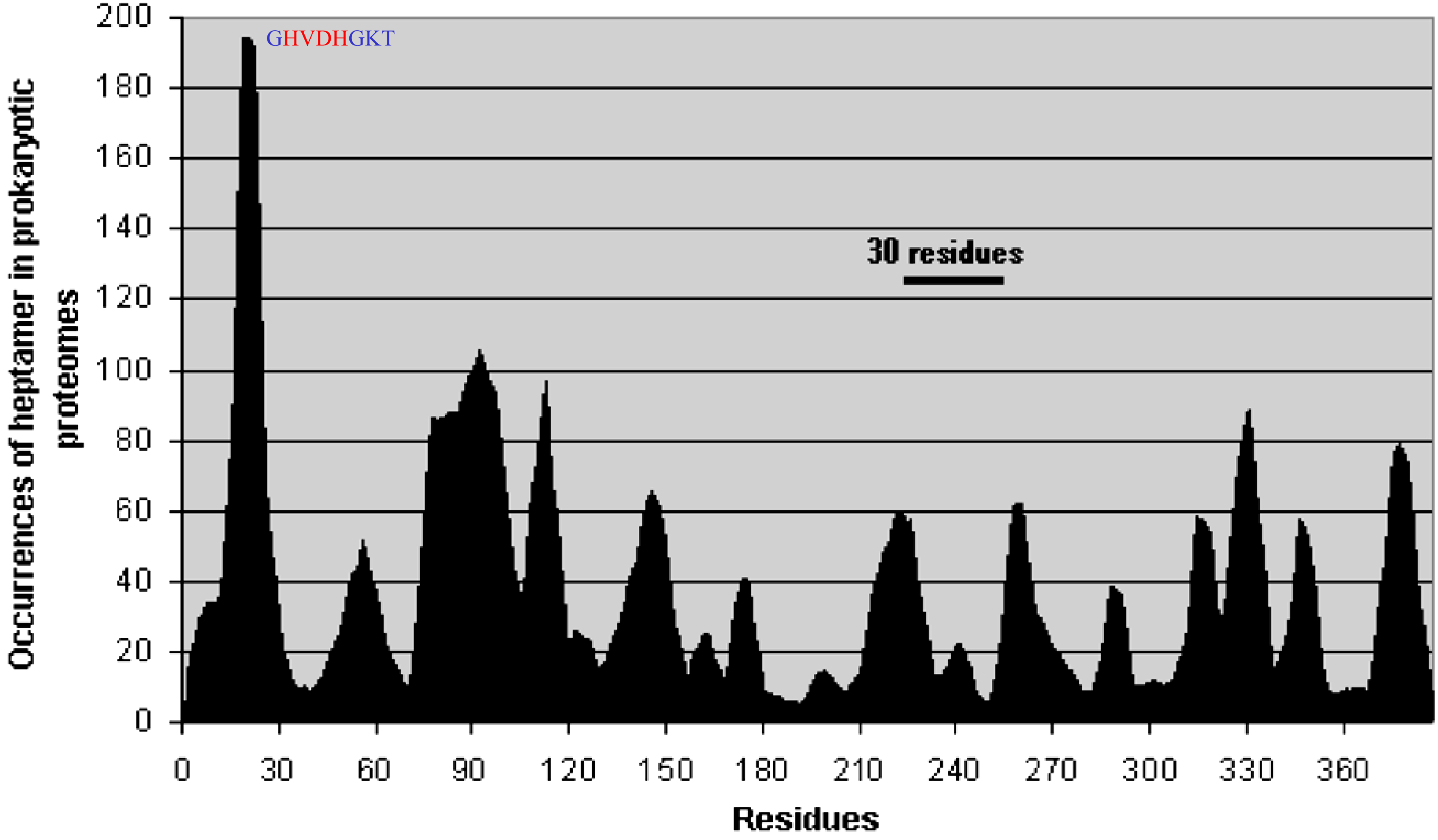
Modules of chemotaxis protein cheY



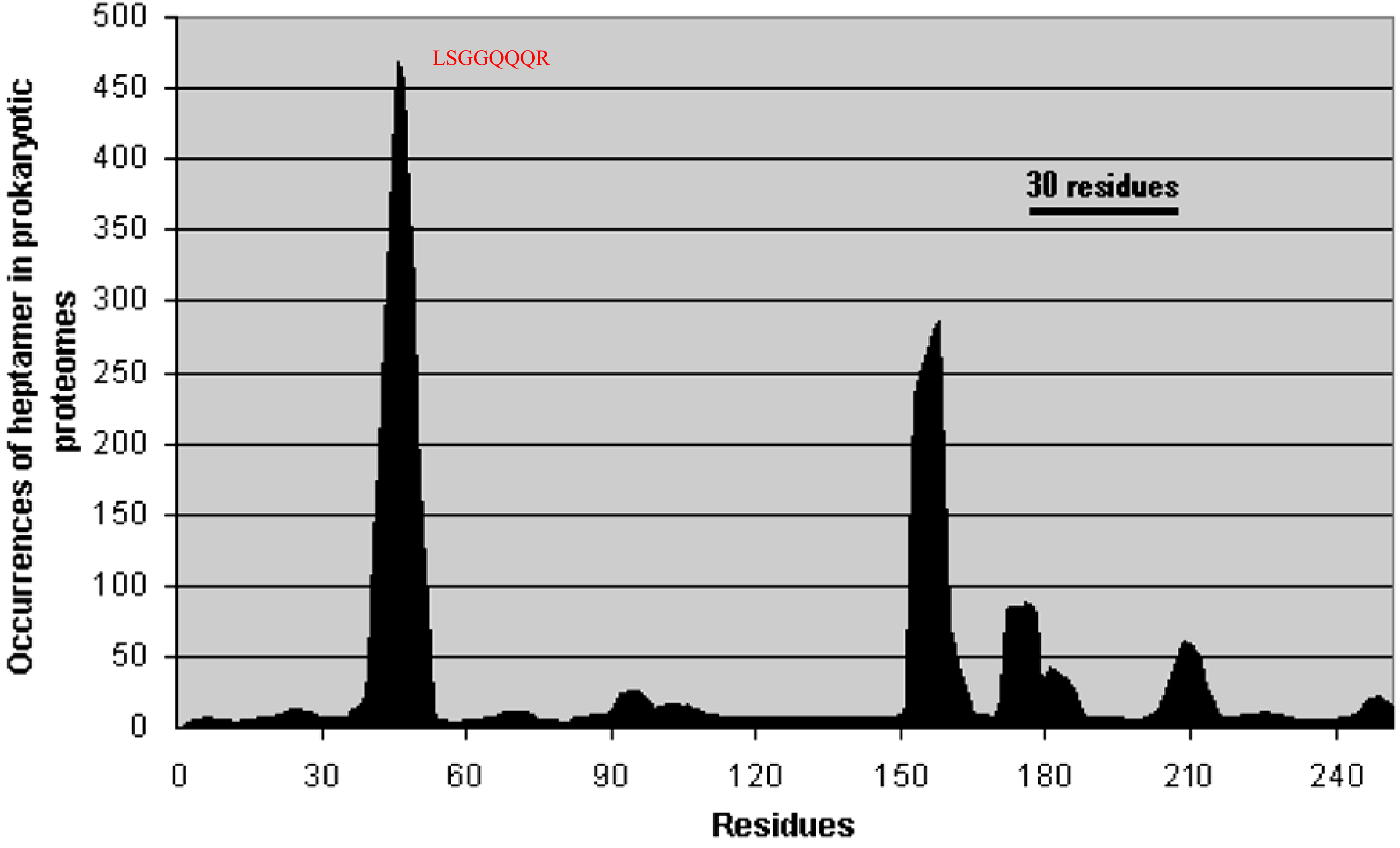
Modules of cytidylate kinase



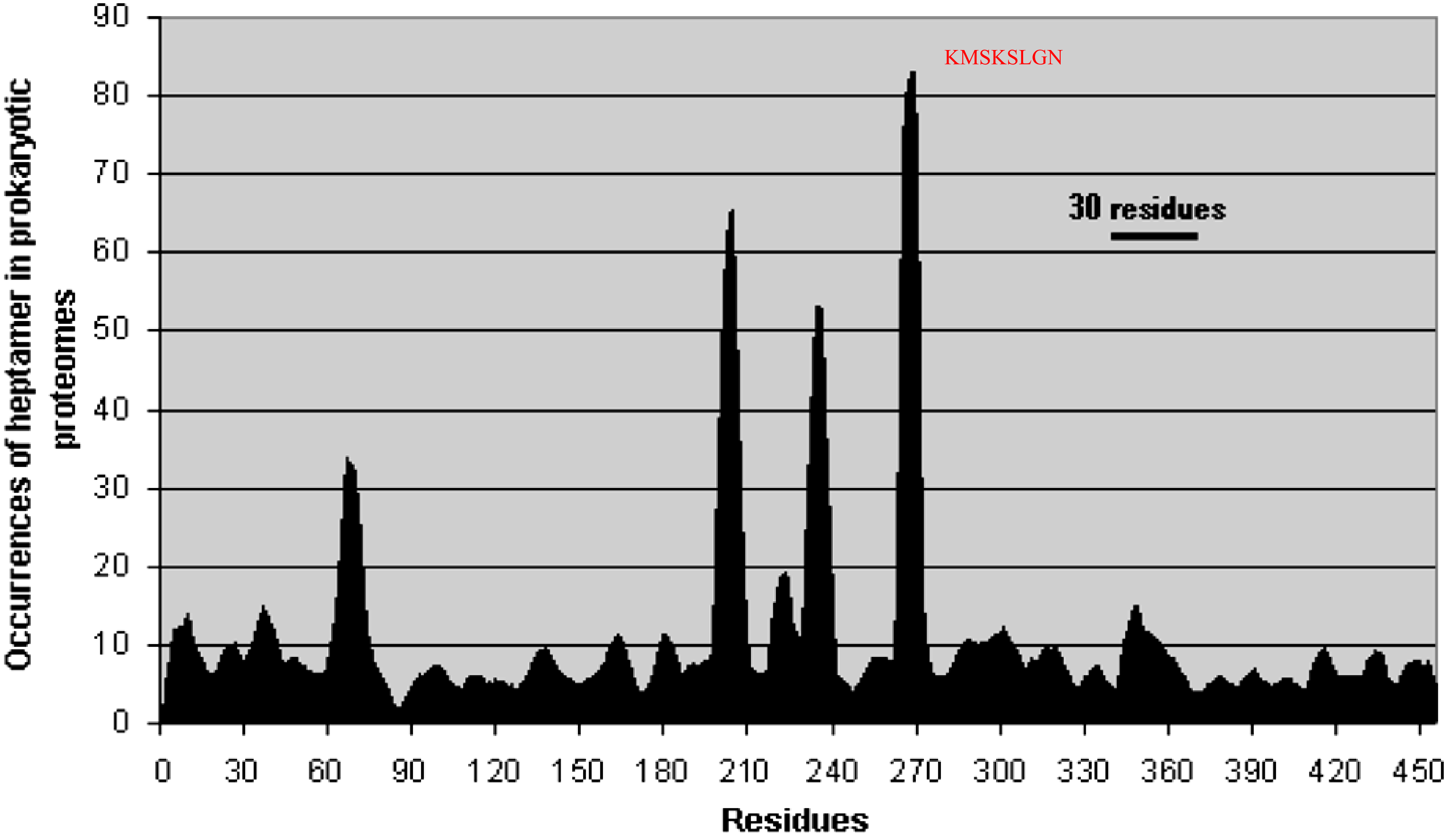
Intact elongation factor, Chain A, *E. Coli*



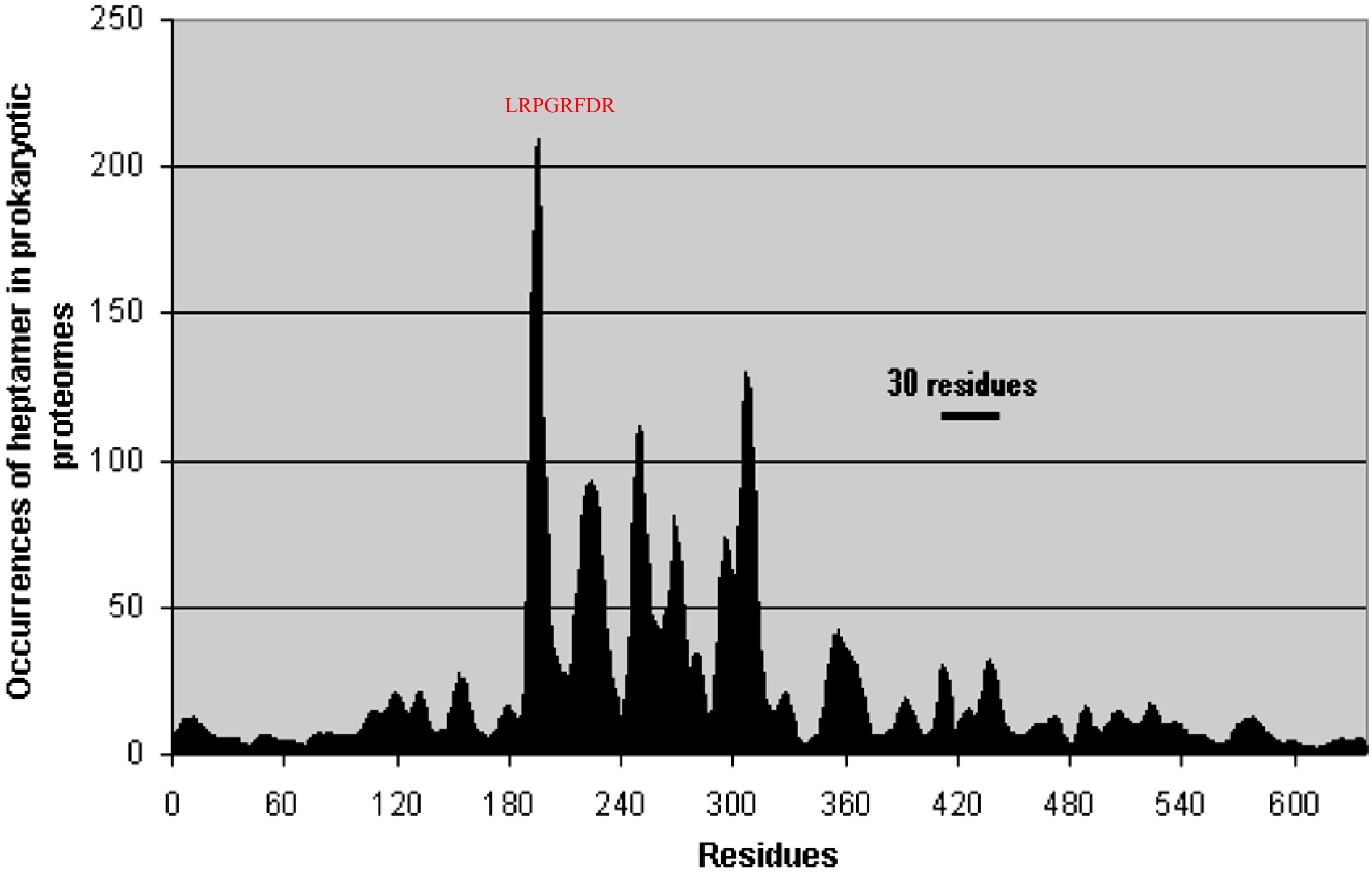
ATP-binding component of high-affinity phosphate-specific transport system, *E. Coli*


LSGGQQQR

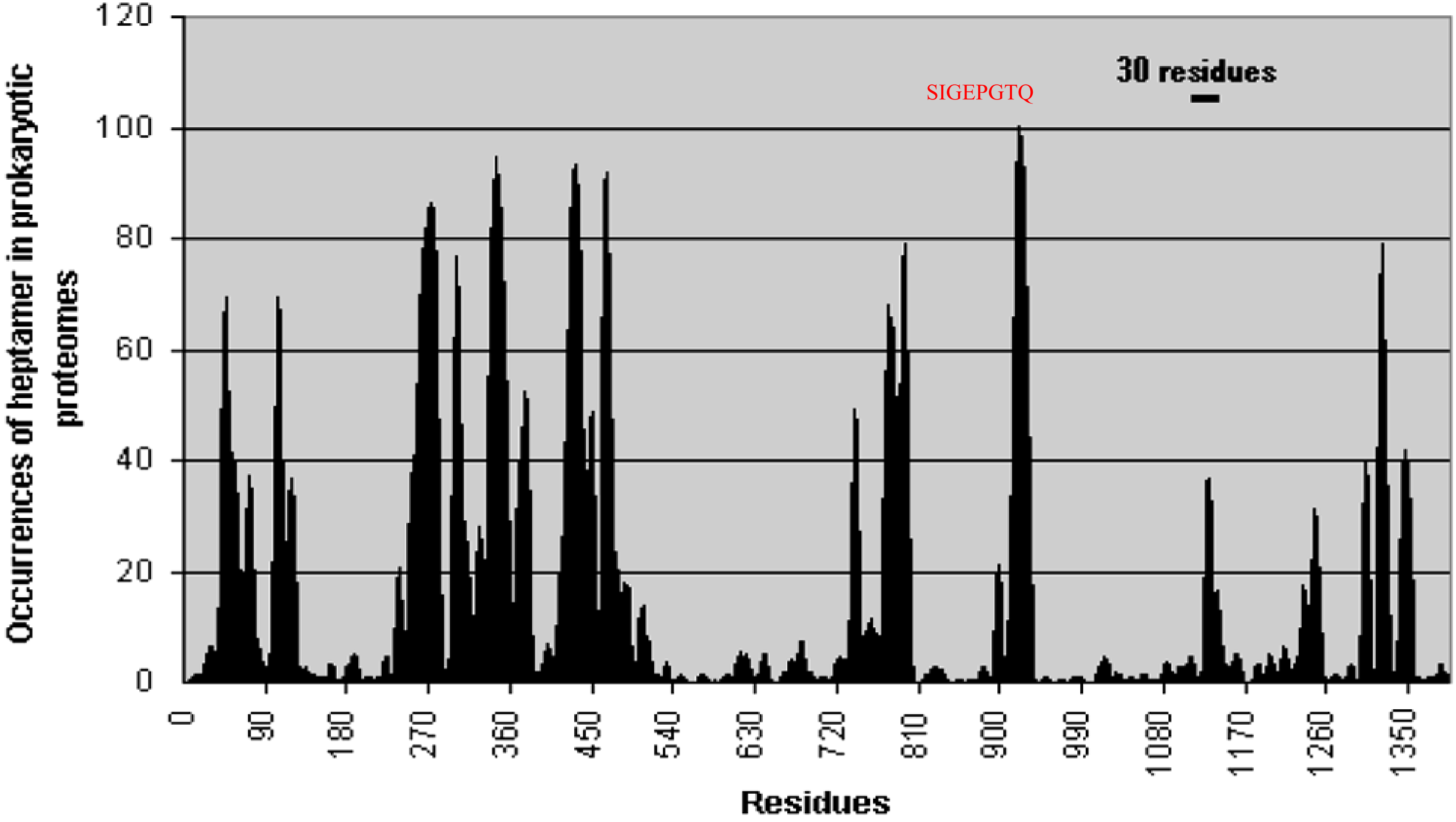
cysteine tRNA synthetase, *E. Coli* K12



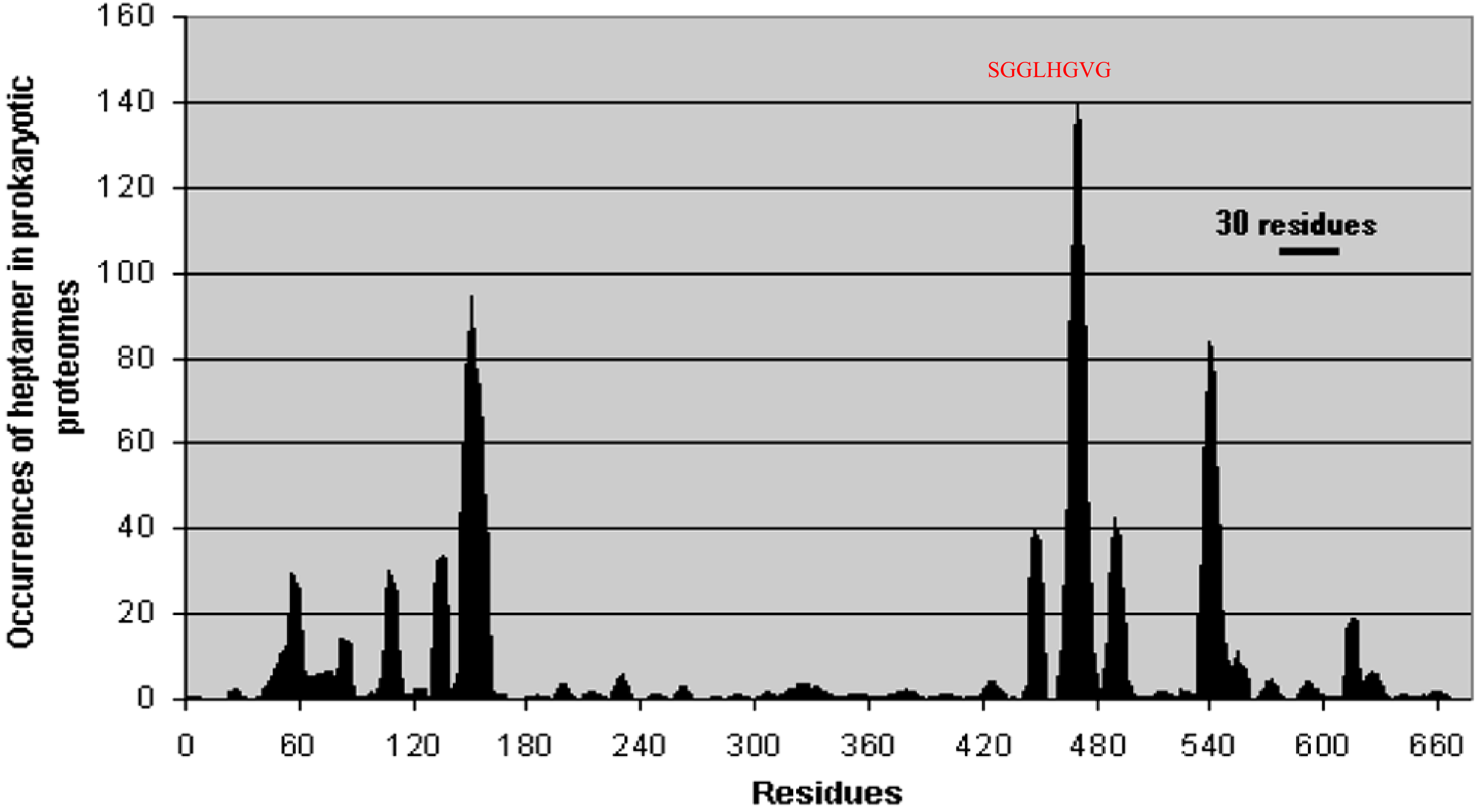
Cell division protein ftsH, *E. Coli*



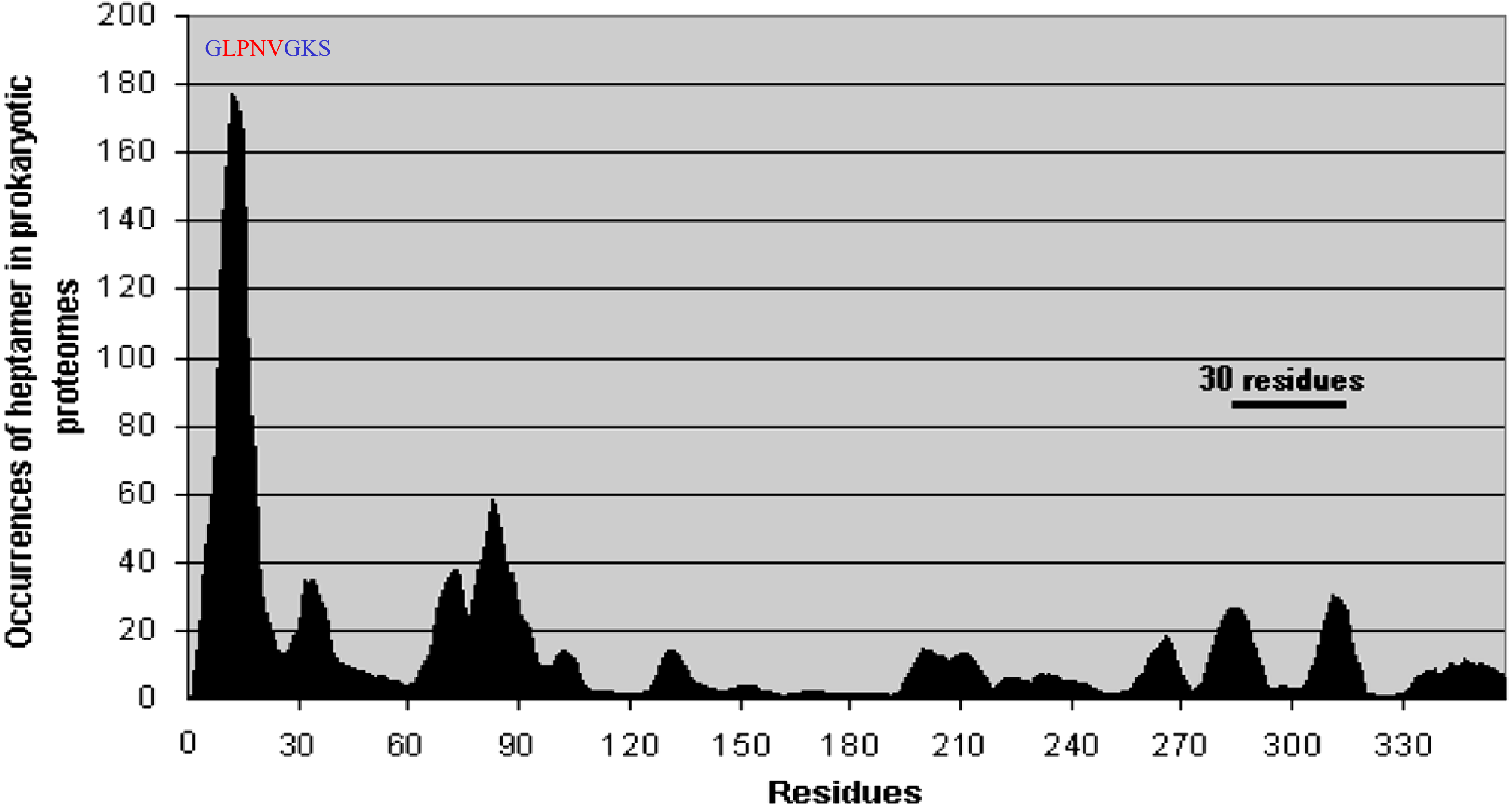
RNA polymerase beta subunit, *Rhodopseudomonas palustris CGA009*



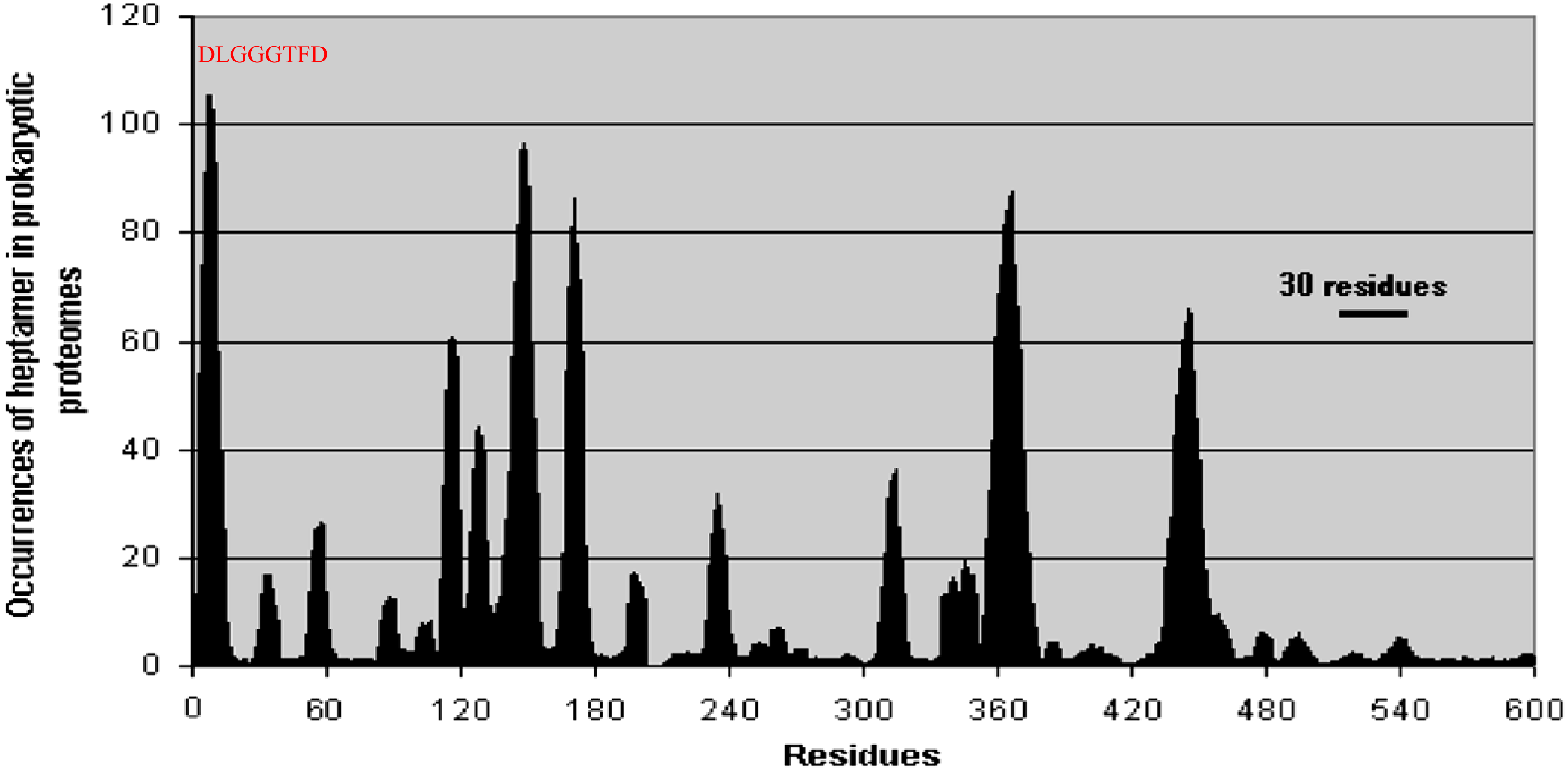
DNA topoisomerase,
Rhodopseudomonas palustris CGA009



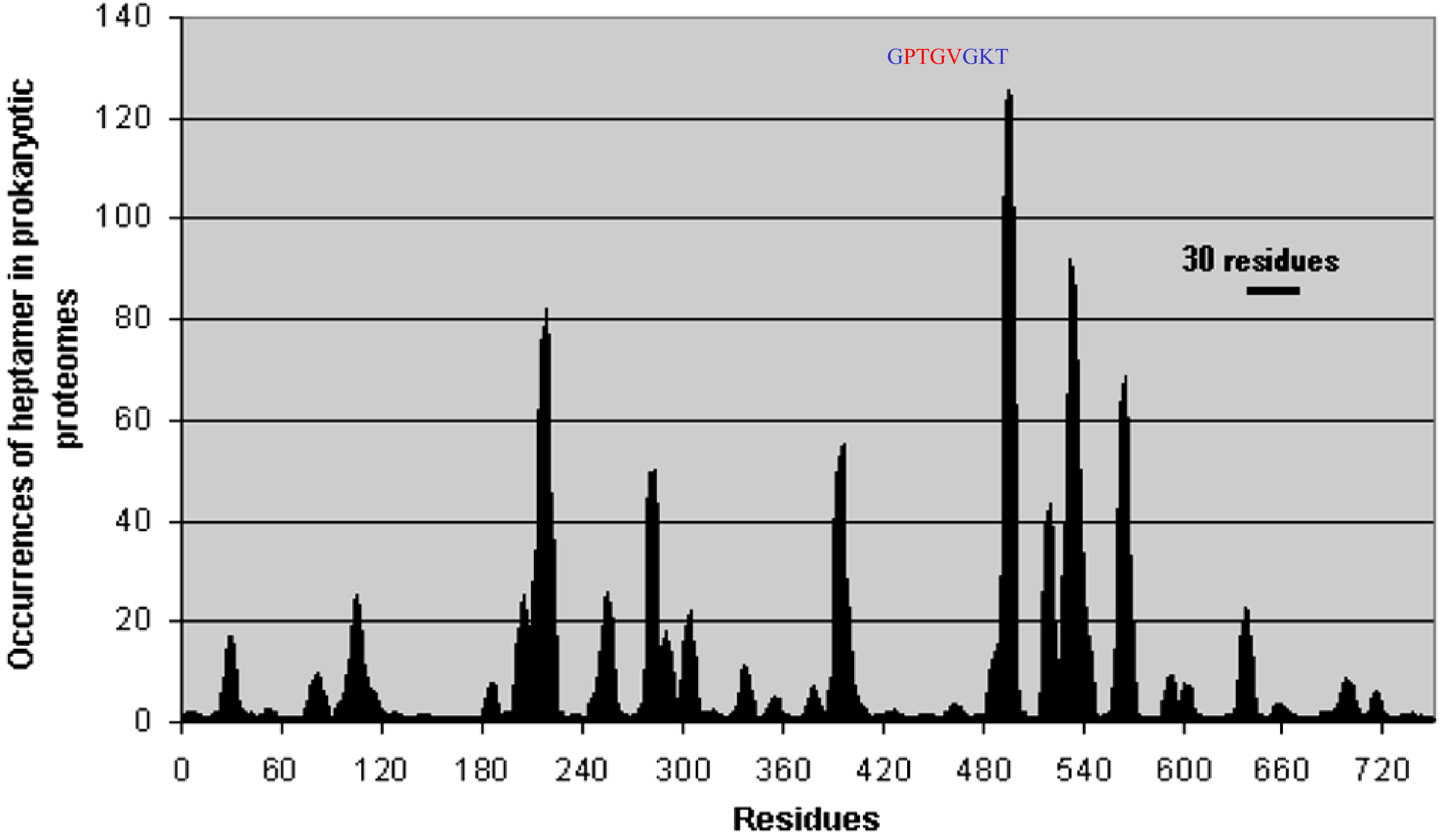
GTP-binding protein,
Hæmophilus influenzæ Rd KW20



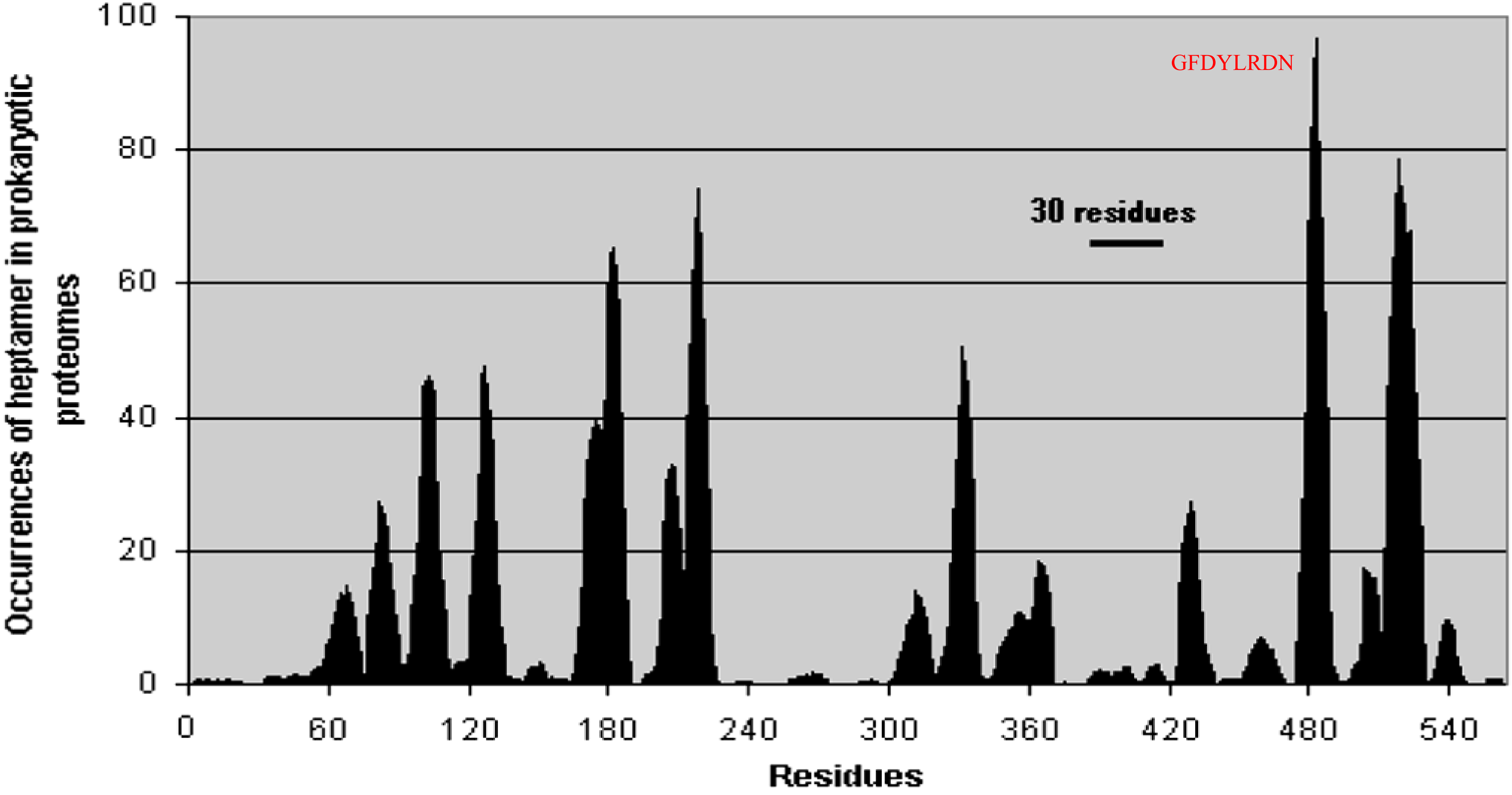
Heat shock protein DnaK
Fusobacterium nucleatum subsp. polymorphum

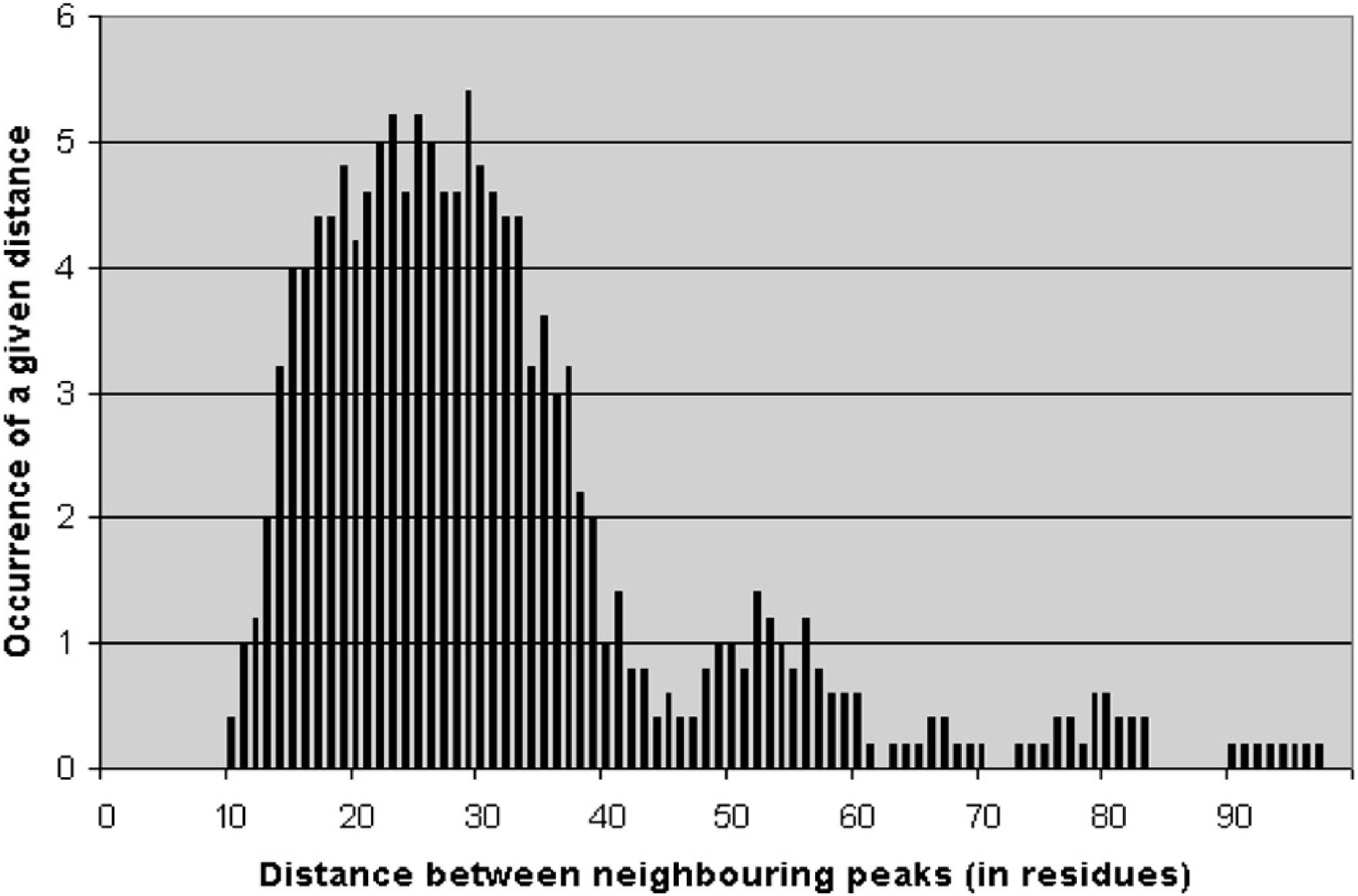


ClpA, ATP dependent protease, chaperonin
Nitrosomonas europæa ATCC 19718



protein translocase subunit SecA
Heliobacillus mobilis




Occurrence of a given distance
0
1
2
3
4
5
6
0
10
20
30
40
50
60
70
80
90
Distance between neighbouring peaks (in residues)

ABC transporters

(… GPS S LTA S LSG S IYV …)

GPS (Aleph) **LTA (Dalet)** **LSG, LAD (Beth)** **IYV (Zayin)**

(36) GPSGSGKsTmL (38) fVFQqfnLiPlLTALENV (40) QLSGGQQQRVAIARAL(6)iLADEPTgALD (22) vvVTHDi (30) 1F3O

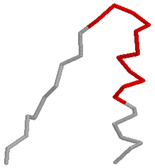
(32-72)**GPSGSGKTTLL**(29-41)MVFQNYALFPHLTALENV(31-42)Q**LSGGQQQRVAIARAL**(6 L**LADEPT**SALD(21-22)IYVTHDQ(28-263) **consensus**

The consensus sequences of the modules are built from overlapping motifs that appear in at least half of the 15 representative species.
There are representatives of the above cassette in every species.
Thus the ABC cassette as outlined above is OMNIPRESENT

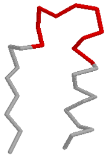
Proteases (cell division proteins FtsH)

(… GPP FVE FID DER RPG …)

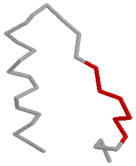
GPP (Aleph)



FVE



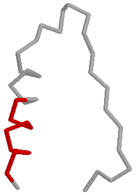
FID



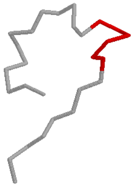
(197) LLVGPPGTGKTLLARAVAGEA(7)SGSDFVELFVGVGAARVRD(9)PCIVFIDEIDAVGR (10) 2CEA

(146-463)LLV**GPPGTGKTLLA**RAVAGEA(7)SGSDFVEMFVGVGASRVRD(9)PCII**FIDEID**AVGR(7-11) consensus

DER



RPG



DEREQTLNQLLVEMDGF(8)MAATNRPDILDPALLRPGRFDKK (297) 2CEA

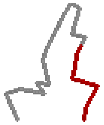
DEREQTLNQLLVEMDGF(8)IAATNRPDxLDPALLRPGRFDRQ (95-415) consensus

- another example of the omnipresent cassette

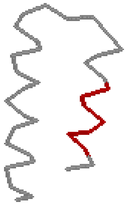
Omnipresent cassette of RNA polymerases

(... FAT NEK S NLL S S VLL NAD ...)

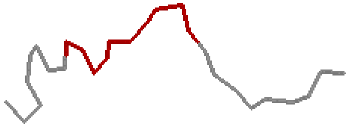
FAT



NEK

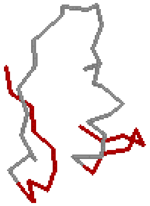


NLL



```
    (529)  VDGGRFATSDLNDLYRRLINRNNNRLK (12) RNEKRMLQEAVDAL  (27) GKQGRFRQNLLGKRVDYSGRSVIVVGP 2A6E

(224-518)LDGGRFATSDLNDLYRRVINRNNNRLK (12)  RNEKRMLQEAVDAL(25-27)GKQGRFRQNLLGKRVDYSGRSVIVVGP consensus
```

VLL **NAD**

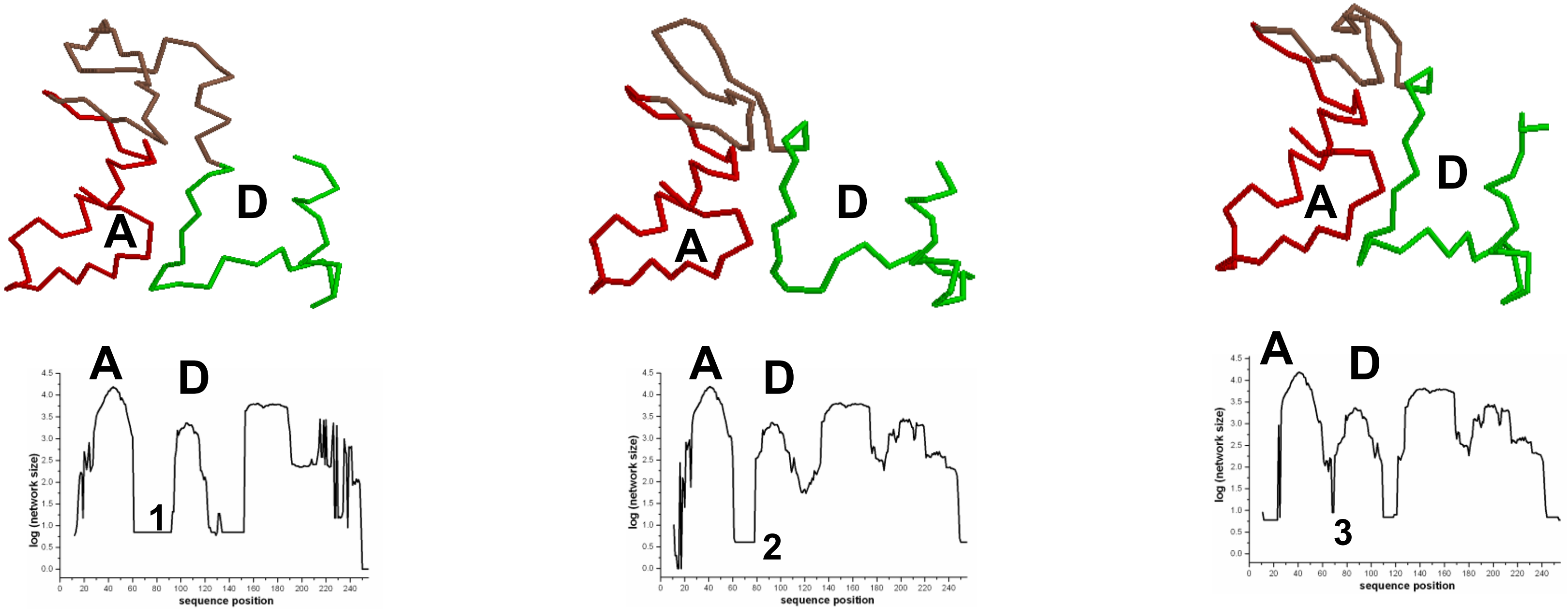


```
     (62) KVVLLNRAPTLHRLGIQAF (18) AFNADFDGDQMAVH   (776)   2A6E

  (59-84)HPVLLNRAPTLHRLGIQAF (18) AFNADFDGDQMAVH (131-961) consensus
```

The maps of the modules show as well
the “silent” regions
– least conserved, least related to anything
and, perhaps, not very much loaded functionally.

These would be of not much interest
for the sequence alignment community

silent module 1 silent module 2 silent module 3



```
             A              silent modules 1-3              D
IVLLVGPSGSGKTTLLRALAGLLGPDGG                              RRGIGMVFQEYALFPHLTVLENVALGL
     | ||||| | ||    |  |  |                              |    ||||   |  | ||||||
VISIIGSSGSGKSTFLRCINFLEKPSEGSIVVNGQTINLVRDKDGQLKVADKNQLRLLRTRLTMVFQHFNLWSHMTVLENVMEAP 1
     | ||||| | || |  || || |          || |      |  |           ||||   |  ||||  |
FMILLGPSGCGKTTTLRMIAGLEEPSRG---QIYIGDRLVADPEKGIFVPPK------DRDIAMVFQSYALYPHMTVYDNIAFPL 2
|    ||||||| | ||||||||    |          |          ||        |   |||||||||||| |  |  | |
FVVFVGPSGCGKSTLLRMIAGLETITSG---------DLFIGEKRMNDTPPA------ERGVGMVFQSYALYPHLSVAENMSFGL 3
```

The silent modules appear to maintain
3D structural relationships between functionall modules

When long sequences are compared it is worth first to identify which segments are more informative.

This is done by mapping of the modules.

The list of modules revealed in the map
for a given protein sequence,
with reference to corresponding
(characterized) networks
of the precalculated sequence space

provides full annotation of the protein

V. Alva et al., PROTEIN SCIENCE 19 , 124-130, 2010

"…modular peptide fragments of between 20 and 40 residues
that co-occur in the connected folds
in disparate structural contexts.
These may be
descendants of an ancestral pool of peptide modules…"

V. Alva et al., PROTEIN SCIENCE 19 , 124-130, 2010

What are the protein modules:

Their sequences are represented by networks
in the protein sequence space -
separate network (or group of related networks) for each module.

Each module has its own unique structure.
Typically, these are closed loops of the contour length 25-30 residues.

Apart from general activity ascribed to the protein that harbors given module,
each module type has its own specific function.

Individual modules even of the same type are sequence-wise often different.
Their evolution from ancestral prototypes
may be traced along walks and networks in the sequence space.

Proteins are made
from standard size modules
of many types.

Each type has its unique structure and function,
but highly variable sequence

All current protein science turns inside out:

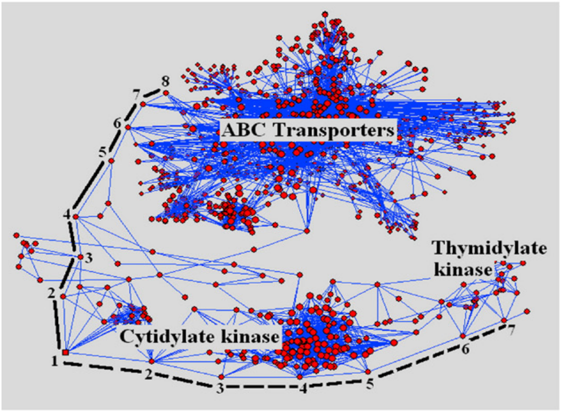
Protein world is world of modules

Every breakthrough that opens new vistas
also removes the ground
from under the feet of other scientists.

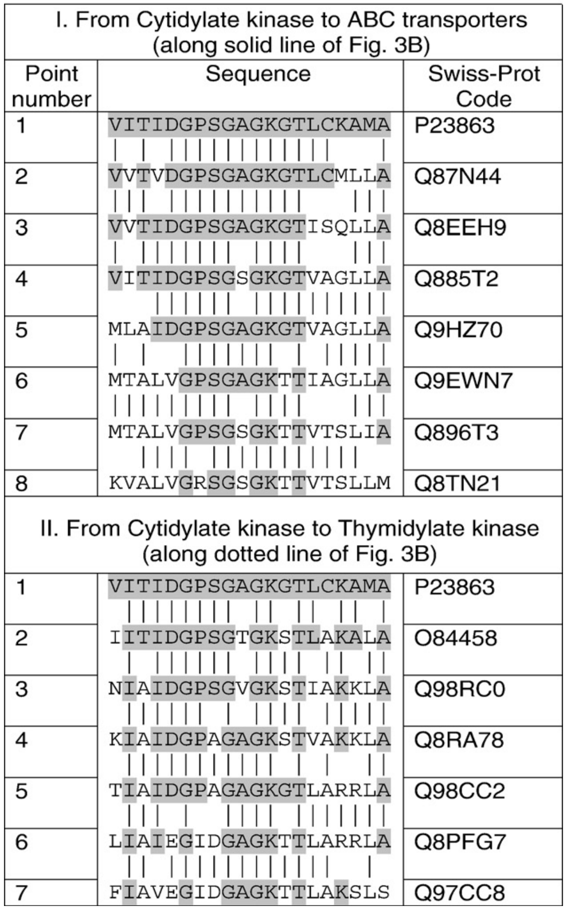
The scientific joy of those who have seen the new light
is accompanied by the dismay
of those whose way of life has been changed for ever.

Fersht A, Nature Rev Mol Cell Biol, 2008

B



C



I. From Cytidylate kinase to ABC transporters (along solid line of Fig. 3B)

Point number	Sequence	Swiss-Prot Code
1	VITIDGPSGAGKGTLCKAMA	P23863
2	VVTVDGPSGAGKGTLCMLLA	Q87N44
3	VVTIDGPSGAGKGTISQLLA	Q8EEH9
4	VITIDGPSGSGKGTVAGLLA	Q885T2
5	MLAIDGPSGAGKGTVAGLLA	Q9HZ70
6	MTALVGPSGAGKTTIAGLLA	Q9EWN7
7	MTALVGPSGSGKTTVTSLIA	Q896T3
8	KVALVGRSGSGKTTVTSLLM	Q8TN21

II. From Cytidylate kinase to Thymidylate kinase (along dotted line of Fig. 3B)

Point number	Sequence	Swiss-Prot Code
1	VITIDGPSGAGKGTLCKAMA	P23863
2	IITIDGPSGTGKSTLAKALA	O84458
3	NIAIDGPSGVGKSTIAKKLA	Q98RC0
4	KIAIDGPAGAGKSTVAKKLA	Q8RA78
5	TIAIDGPAGAGKGTLARRLA	Q98CC2
6	LIAIEGIDGAGKTTLARRLA	Q8PFG7
7	FIAVEGIDGAGKTTLAKSLS	Q97CC8

Examples of evolutionary paths

MOST COMMON
PROTEIN SEQUENCE MODULES (PROTOTYPES)

Aleph	**GEIVLLVGPSGSGKTTLLRALAGLLGPDGG**
Beth	**LSGGQRQRVAIARALALEPKLLLLDEPTSALD**
Gimel	DVVVIGAGGAGLAAALALARAGAKVVVVE
Dalet	RRGIGMVFQEYALFPHLTVLENVALGL
Heh	PVIMLTARGDEEDRVEALLEAGADDYLTKPF
Vav	LLGLSKKEARERALELLELVGLEEKADRYP
Zayin	LLLKLLKELGLTVLLVTHDLEEA

Berezovsky et al. 2000-2003

The underlined motifs are omnipresent

```
KVALVGRSGSGKTTVTSLLM
FIAVEGIDGAGKTTLAKSLS
     GxxxxGKT  -  Walker A motif
                  (NTP binding)
```

Omnipresent 6-9 mers of 15 prokaryotes from different phyla

ALEPH ATP/GTP binding

1	HVDHGKTTL
2	GPPGTGKT
3	GHVDHGKT
4	GSGKTTLL
5	IDTPGHV
6	GPSGSGK
7	PTGSGKT
8	NGSGKTT
9	GKSTLLN
10	SGSGKT
11	TGSGKS
12	PGVGKT
13	PNVGKS
14	GVGKTT
15	GTGKTT
16	DHGKST
17	GKTTLA
18	GKTTLV
19	KSTLLK

BETH ATPases of ABC transporters

20	QRVAIARAL
21	LSGGQQQRV
22	LADEPT
23	TLSGGE

Other omni:

24	FIDEID
25	KMSKSL
26	WTTTPWT
27	NADFDGD

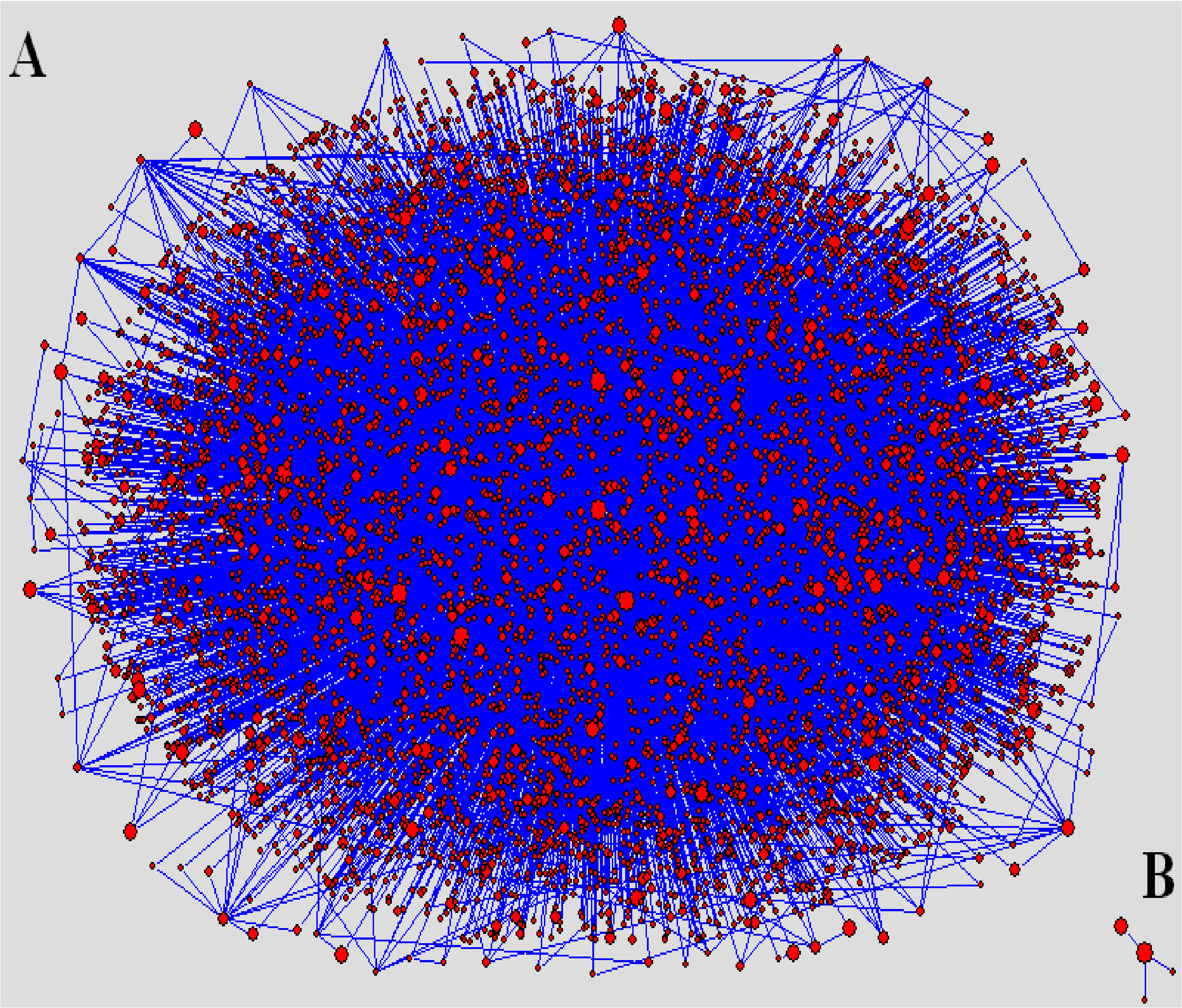
**Omnipresence is a new measure of sequence conservation.
These elements are the most conserved ones,
coming, presumably from last common ancestor**

ALEPH and BETH reconstructed from overlapping omnipresent motifs turn out to be relatives, though they do not match:

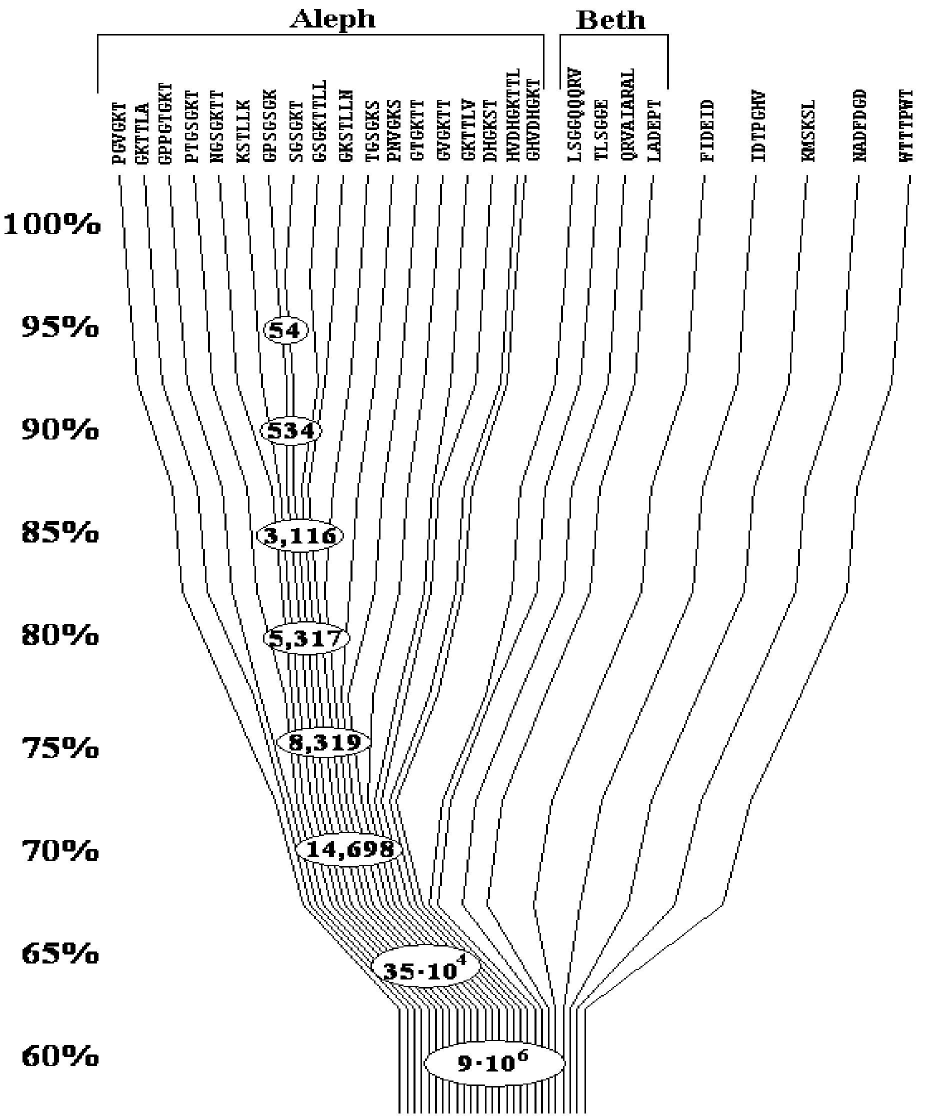
```
IDTPGHVDHGKTTLLN      ALEPH
  |
  TLSGGQQQRVAIARAL    BETH
```

They both belong to 10% monster network.

All 27 omnipresent elements belong to the same network



10% MONSTER network (10^7 fragments)



Sequence space based
evolutionary tree of omnipresent elements

TO CONCLUDE THE CHAPTER ON NETWORKS:

I. Protein sequence characterization via networks in the sequence space does not require

gap penalties,
nor substitution matrices,
nor statistics of alignment

II. The networks in the sequence space represent protein modules. Each sequence fragment belongs to only one specific network, and, thus, is given an unequivocal annotation.

III. Each protein can be described as linear combination of several different modules, and presented as word in the alphabet of the modules – the proteomic code

Paths from Aleph to Beth and back

	A		**B**	
1	GEFVAIVGPSGCGKSTLLRL	Q825G5	GEFVAIVGPSGCGKSTLLRL	Q825G5
2	**GESLALTGESGSGKSTLLHL**	**Q7CP38**	**GEVVVIIGPSGSGKSTLLRS**	**Q97RJ0**
3	AQTI**ALIGESGSGKSTLLGI**	**Q8ZCB4**	**QVVVVGAGPSGSTVSALLKS**	**Q87R97**
4	**ATLAALIGAGGLGKLILLGI**	**Q813M6**	**DVVVVGAGPSGSSAARYLSE**	**O66509**
5	**AVIAALIGAGGFGALVFQGL**	**Q8X670**	**DVVVIGAGPGGYVAAIRASQ**	**Q9A7J2**
6	V**VLAGLVGAGGLGAEVTRGL**	**Q8U8Y4**	**DAVIIGGGPGGYVCAIKLAQ**	**Q9WYL2**
7	**VVGGGVVGAGTALDAVTRGL**	**Q82DH4**	**FAVITGGGPGAMEAANKGAQ**	**Q8KC62**
8	**VVGGGSTGAGVARDLAMRGL**	**Q9HNS4**	**LTVATGGGPGAMEAANLGAY**	**O86748**
9	**VVGGGFTGQSAALHLAEGGL**	**Q8UCD8**	**LDVGTGSGVLAMAAAKLGAA**	**Q9RU72**
10	LC**GGGFTGQSQALRLAIARA**	**Q8A0Z5**	**LDLGTGSGALAVHAARLGAR**	**Q826J9**
11	**LSGGERIALSIALRLAIAKA**	**Q97WH0**	**LDTGIMSGADIVAAIALGAR**	**Q9CBF2**
12	**LSGGQRRALGIALALASNPE**	**Q9YBQ1**	**MDGGIRSGQDVLKAVALGAR**	**Q8UD10**
13	**LSGGQRQRVAIARALALDPD**	**Q82BU6**	**VSGGIRSGADVAKALALGAD**	**Q8U870**
14	A**SGGMRDGVMMAKALAMGAS**	**O58893**		
15	L**SGGMRQRVMIAIALACGPD**	**Q89KL2**		
16	**LSGGQRQRVAIARALALDPD**	**Q82BU6**		

	C		**D**	
1	GEFVAIVGPSGCGKSTLLRL	Q825G5	GEFVAIVGPSGCGKSTLLRL	Q825G5
2	**GQVVVVLGPSGSGKSTLCRT**	**Q8RQL7**	**GKLVALLGPSGSGKSTLLRL**	**Q8Z0H0**
3	**GQVVMVTGAGGSIGSELCRQ**	**Q9HZ86**	**NKLVLLTGPSGSGKSTLALD**	**Q9KEY5**
4	RK**VAFVTGGAGGIGSETCRQ**	**Q9KCM1**	**IHLVNLSGPAGSGKTILALA**	**Q887P5**
5	GR**VAFVTGGAGGIGRATAER**	**Q8UA89**	**GHLQSASGPLGLMKTILALR**	**O50436**
6	**GKTAFITGGGQGIGLACAEA**	**Q89QA5**	**GHMDAAAGIGGLIKTVLALR**	**Q8U9Q4**
7	LV**TGANTGLGQGIALALAEA**	**Q8PE31**	**GHTGGAAGIAGLLKAVLAIE**	**O06586**
8	**LVTGANKGIGLAIARQLGAA**	**Q7CP30**	**GRTGGWAAIAGLLAAIGATV**	**Q98BE5**
9	**LVTGSSQGIGAAIAAGLARA**	**Q9RK29**	**GSRGIGAAIARRLAADGAHV**	**Q8XT12**
10	SAC**GSSSGSGAAVAAGLAPL**	**Q9A5H4**	**ASRGIGKAIAEVAARDGAPV**	**Q92PY2**
11	LPG**GSSSGAGVVVAAGLVPV**	**Q8UAX4**	**SSGKMGYAIAEVAANLGADV**	**Q819T8**
12	IS**GGSSGGSAVAVALGLVDV**	**Q975D0**	**SSGKMGYAVAQVARELGATV**	**Q88WL5**
13	**LSGGESFMAALALALGLSDV**	**Q87HE3**	**SSGNHAQAVALAARELGTTA**	**Q9XAA4**
14	**LSGGESFIAALALALSLAEV**	**Q830T3**	**SSGNHAQGVALAARLHGIPA**	**Q8UBW5**
15	**LSGGMIKRAALARALSLDPD**	**Q8UEV8**	**VSGGQAQRVALALALAGTPA**	**Q9EWP7**
16	**LSGGQRQRVAIARALALDPD**	**Q82BU6**	**LSGGQRQRVAIARALALDPD**	**Q82BU6**

GENOME SEGMENTATION CODE

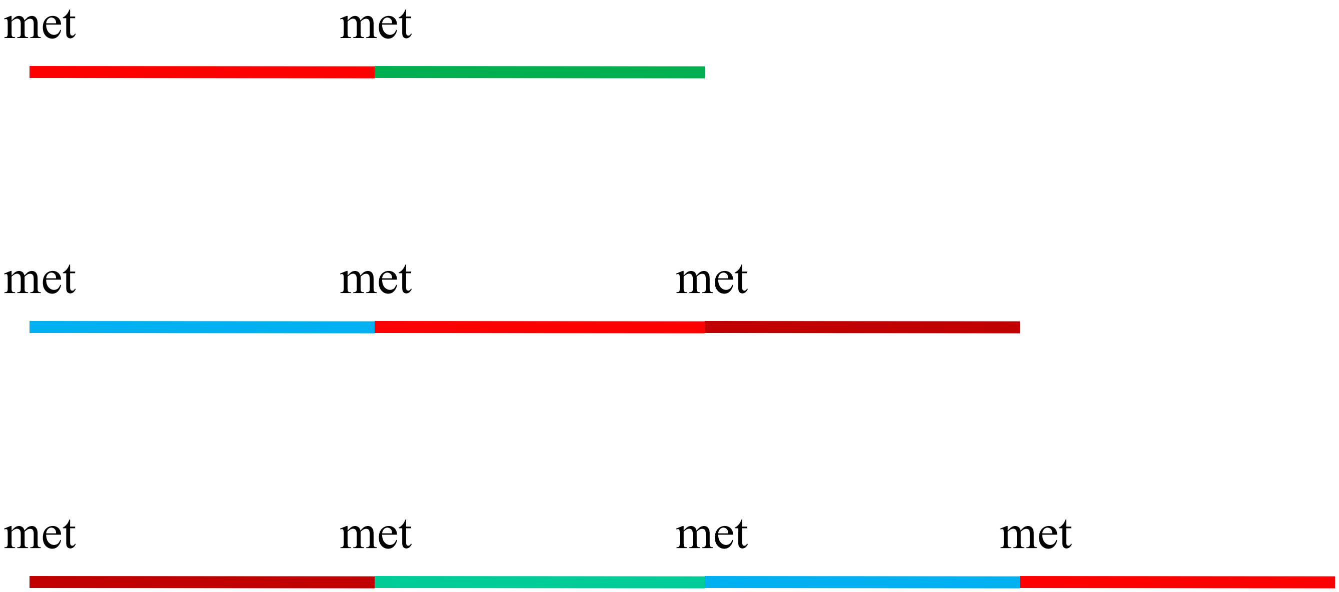
“The proteins… can, with regard to molecular weight, be divided into four subgroups… The molecular masses characteristic of the three higher subgroups are – as a first approximation – derived from the molecular mass of the first subgroup by multiplying by the integers…”

The Svedberg
Mass and size of protein molecules
Nature 123, 871 (1929)

~ 160 aa unit (Svedberg, 1937)

"…proteins of molecular weight greater than about 20 000 are often built up not as a single unit but by a combination of two or three large substructures. This finding suggests that a 3D structure based on the principle of a polar exterior surrounding a hydrophobic core can be conveniently achieved with a polypeptide molecular weight of about 10 000 – 16 000."

B. W. Matthews et al. (P. Sigler)
Nature New Biology
238, 37, 1972

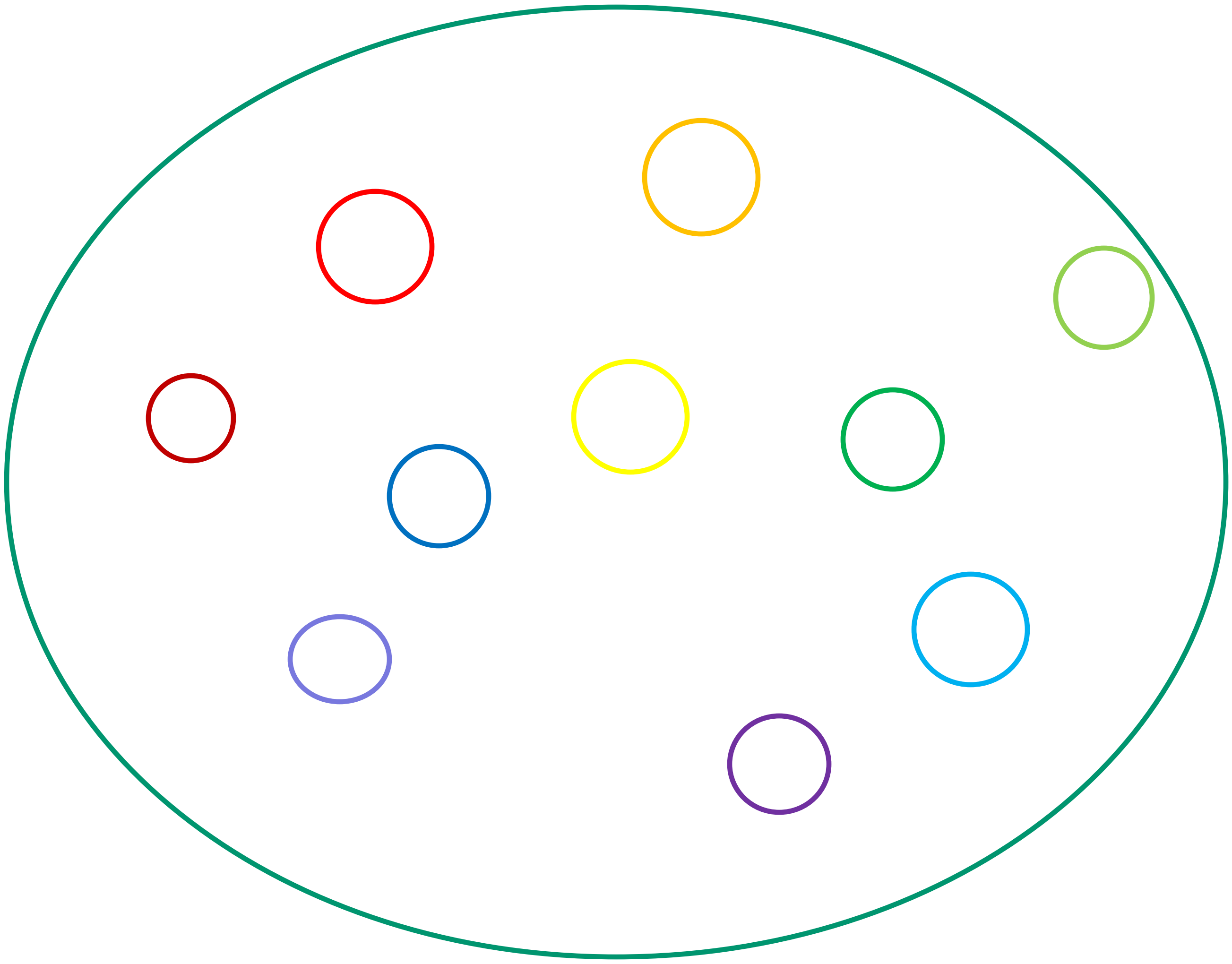

met
met
met
met
met
met
met
met
met

The Lord Of The Rings

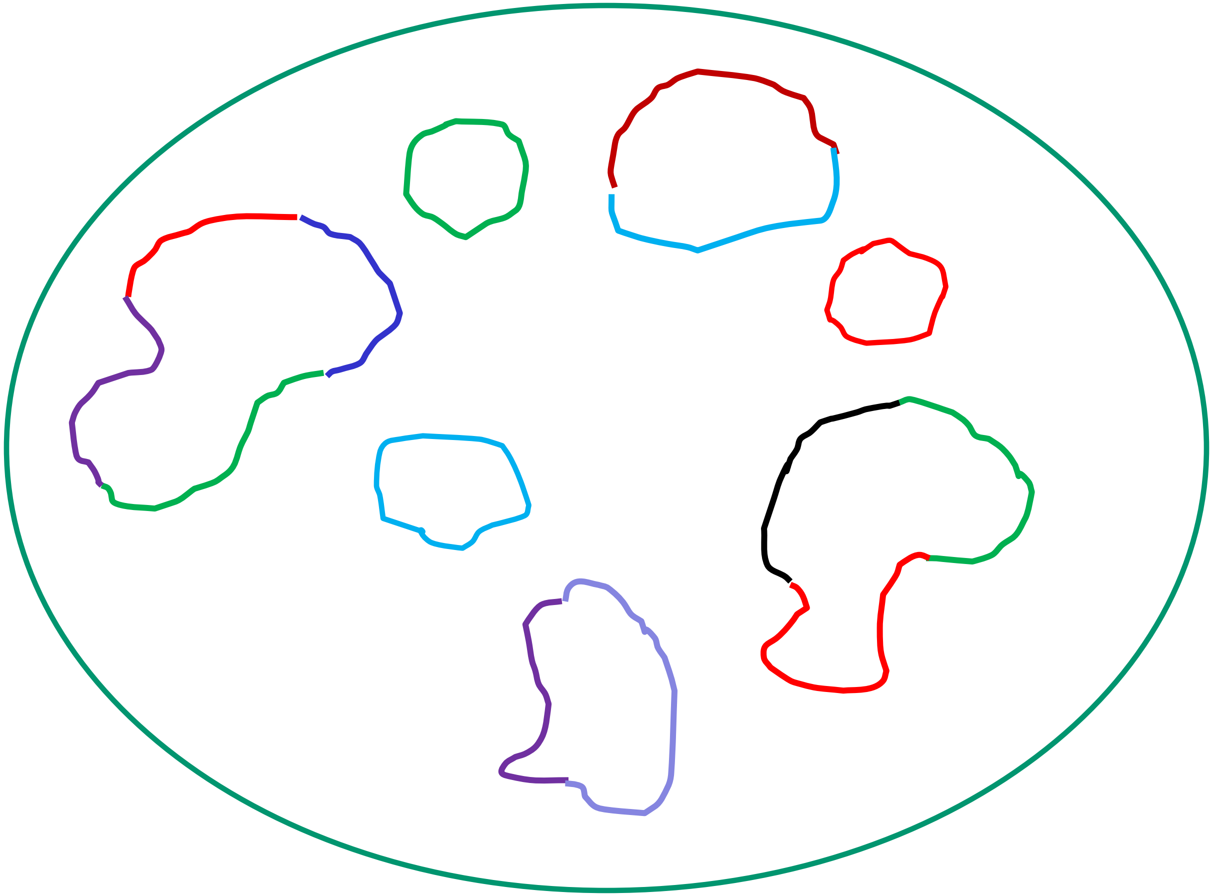
Three rings for the Elven-kings under the sky,
Seven for the Dwarf-lords in their halls of stone,
Nine for Mortal Men doomed to die,
One for the Dark Lord on his dark throne.

J. R. R. Tolkien

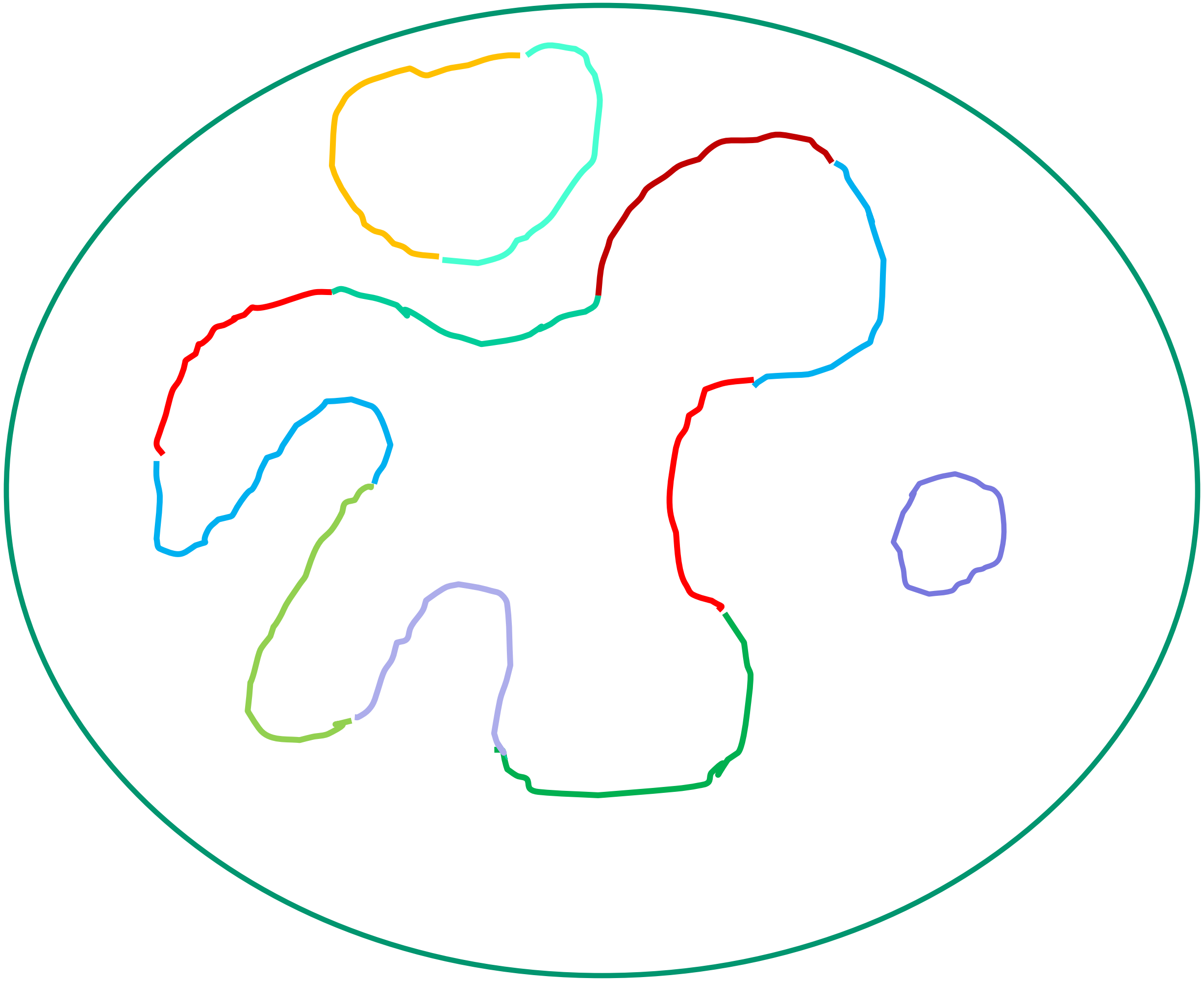
Pre-genomic, pre-recombination stage



Pre-genomic, recombination stage

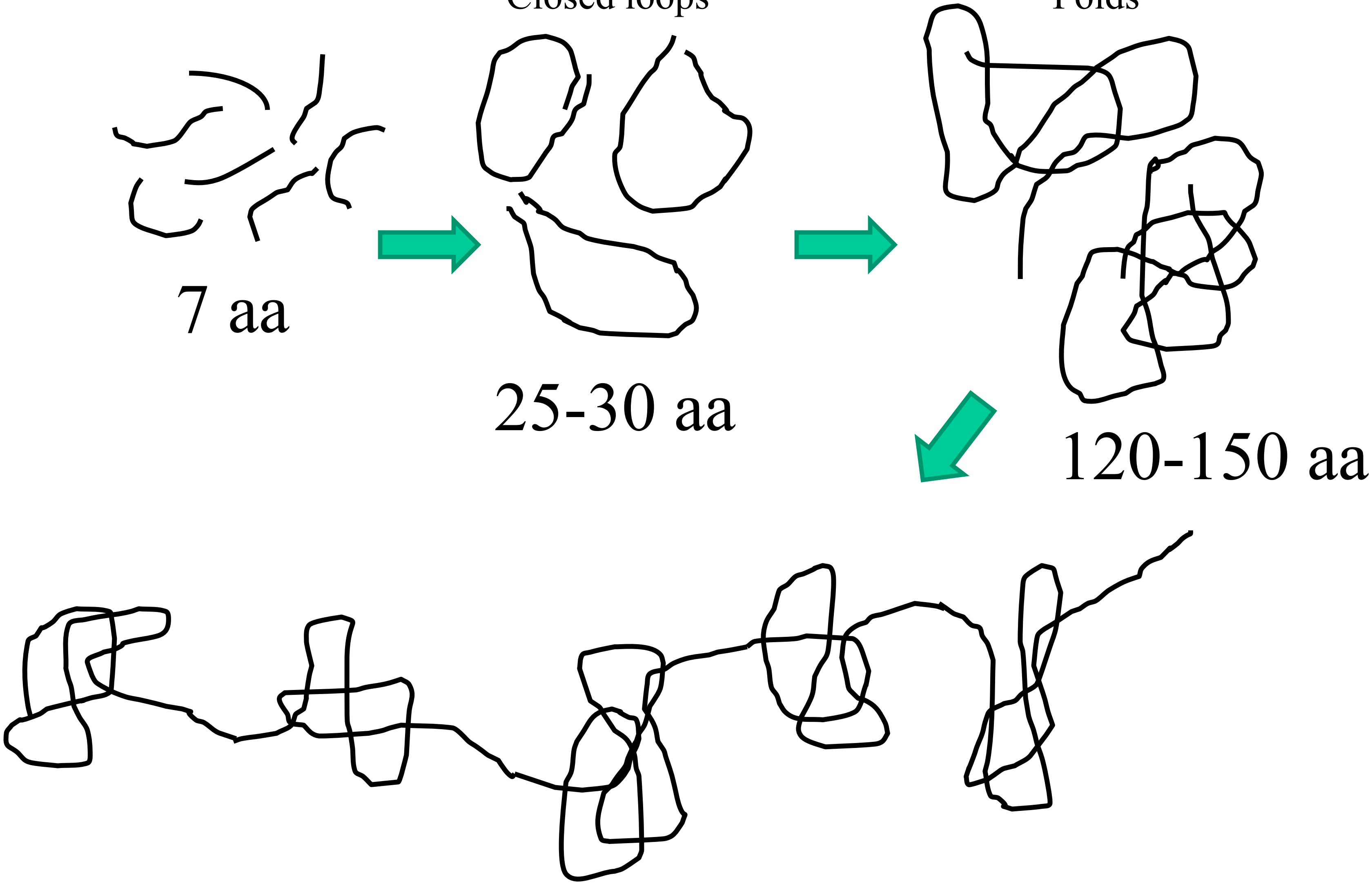


Early genomic stage



“Evolution may have proceeded largely, rather than periferally, through extrachromosomal elements”

D. Reanney
Bact. Rev. 40, 552, 1976



Closed loops

Folds

7 aa

25-30 aa

120-150 aa

Multifold proteins

One striking case of overlapping codes

Triplet extension patterns for A+T rich prokaryotic genomes

species	G+C content %	extension motif
F. nucleatum	27.2	[(a)t]**(A)(T)**[(a)t]
N. equitans	31.6	(ta)t**(A) t**(at)
- " -		(at)**a (T)**a(ta)
S. solfataricus	35.8	[(t)a]ttt**(A)(T)**[(a)(t)]
T. denicola	37.9	[(a)t]**(A)(T)**[a(t)]
C. pneumoniae	40.0	[g(a)]**G(A)**[g(a)
- " -		[(t)c]**(T)C**[(t)c]
M. acetivorans	42.7	[g(a)]**G(A)(T)C**[(t)c]
A. aeolicus	43.3	[gg(a)]**gG(A)**[gg(a)]
- " -		[(t)cc]**(T)C**c[(t)cc]
B. subtilis	43.5	[g(a)(t)]**G(A)(T)C**[(a)(t)c]
T. maritima	46.2	(gaa)**G(A)**[g(a)]
- " -		[(t)c]**(T)C**(ttc)
D. ethenogenes	48.9	(cggc)cggc**(T)C**agccg(gccg)
consensus		**G(A)(T)C**

CGAAAATTTTCG

same as in eukaryotes!:

CGRAAATTTYCG

What this periodical motif codes for in prokaryotes?

```
(GAAAATTTTC)(GAAAATTTTC)....
   AAAATTTTC)(GAAAATTTTC)(G....
    AAATTTTC)(GAAAATTTTC)(GA....
```

```
GAA AAT TTT CGA AAA TTT TCG AAA ATT TTC
glu asn phe arg lys phe ser lys ile phe

AAA ATT TTC GAA AAT TTT CGA AAA TTT TCG
lys ile phe glu asn phe arg lys phe ser

AAA TTT TCG AAA ATT TTC GAA AAT TTT CGA
lys phe ser lys ile phe glu asn phe arg
```

non-polar
amino acids

ala
gly
ile
leu
met
phe
pro
val

polar
amino acids

arg
asn
asp
cys
glu
gln
his
lys
ser
thr
trp
tyr

Our pattern shows alternation of polar and non-polar residues, with the period 3.5 residues

(glu asn phe arg lys phe ser lys ile phe)glu asn phe

period 3.5

period 3.5

α-helices

10-15 aa long

(30-45 bases in DNA)

are often amphipathic

(alternating polar/non-polar aa)

with period ~3.5 residues

(~10.5 bases in DNA)

That keeps polar and non-polar residues on opposite sides of the helix

NF kappaB recognition sequences

(NF kappaB is the heaviest duty transcription factor)

IL-1β-κB	GGGAAAA	TCC	T	
TNFα	GGGAAAG	CCC		C
Urokinase	GGGAAAG	TAC		C
E-selectin (PD3)	GGGAAAG	TTT		C
Ifn-B	GGGAAA	TTCC		C
Lymphotoxin	GGGAAG	CCCC		C
TCR-β	GGGAGA	TTCC		C
PRDII	GGGAAA	TTCCT	T	
GCR	GGGGGG	CACC	T	
ICAM1	TGGAAA	TTCC	H	
κB-33	TGGAAA	TTTC	H	
IL-2	AAGAA	TTTCC	H	
GM-CSF CK1	AGAAA	TTCC		C
G-CSF CK1	AGAAA	TTCC		C
IL-2 CD28RE	AGAAA	TTCC		C
IL-8 CD28RE	GGAAA	TTCC		C
GM-CSF	GGGAA	CTACC		C
TNFα (-655)	GGGAA	TTCAC		C
IL-2R	GGGAA	TTCCC		C
H2	GGGGA	TTCCC		C
E-selectin	GGGGA	TTTCC		C
LCAM	GGGGA	TTTCC		C
Lymphotoxin	GGGGG	CTTCC		C
GMCSF	TAGAA	TCTCC		C
IL-3 CD28RE	TGAGA	TTCC		C
IL-8	TGGAA	TTCCC	H	
Human P sequence	AAAA	TTTCC		C
TF	GGAG	TTTCC		C
Igκ	GGGA	CTTTCC		C
IL-2	GGGA	TTTCAC		C
IL-6	GGGA	TTTCC		C
Angiotensinogen	GGGA	TTTCCC		C
TNFα	GGGG	CTTTCC		C
VCAM	GGGG	TTTCCC		C
Mouse P sequence	AAA	TTTTCC		C
IFNγ	GAA	TTTTCC		C
6-16 ISRE	TCA	TTTTCC		C

GGRAA TTYCC

DNA curvature	**GAAAATTTTC**
Chromatin code	**GRAAATTTYC**
Amphipathic helices	**GAAAATTTTC**
NF kappaB	**GGRAATTYCC**
They all	**GRRAATTYYC**

Reading only one message, one gets three more, practically GRATIS !

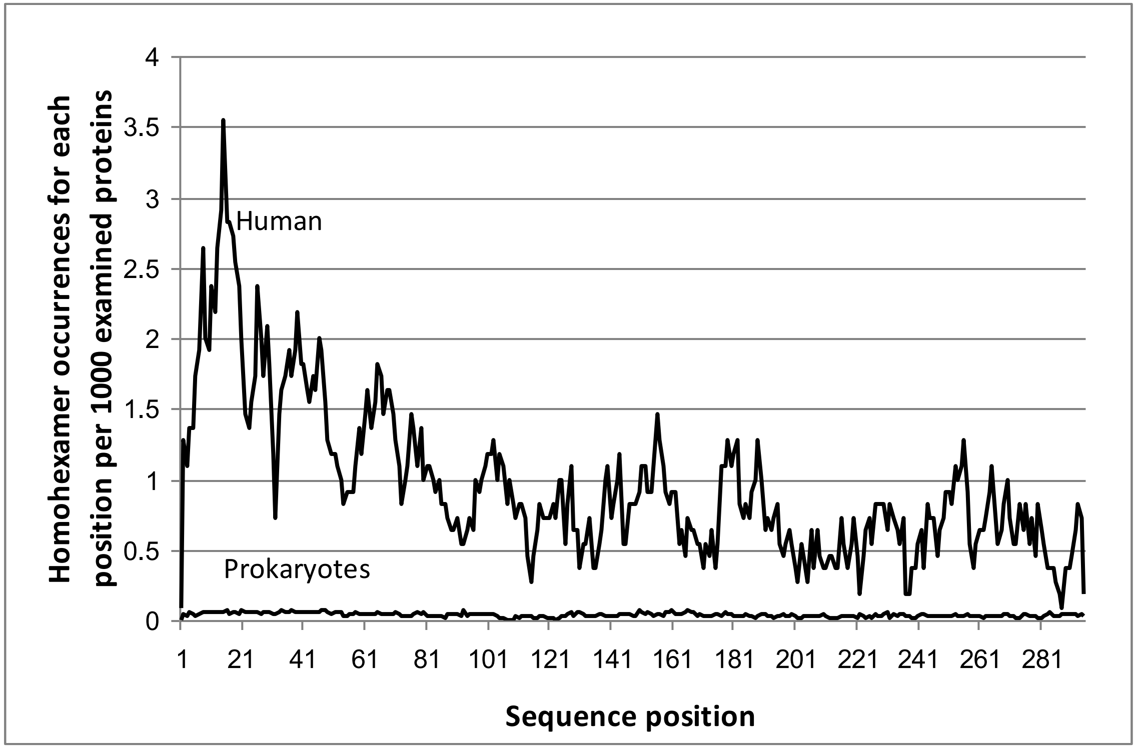
Not only there are many different codes
in the sequences,

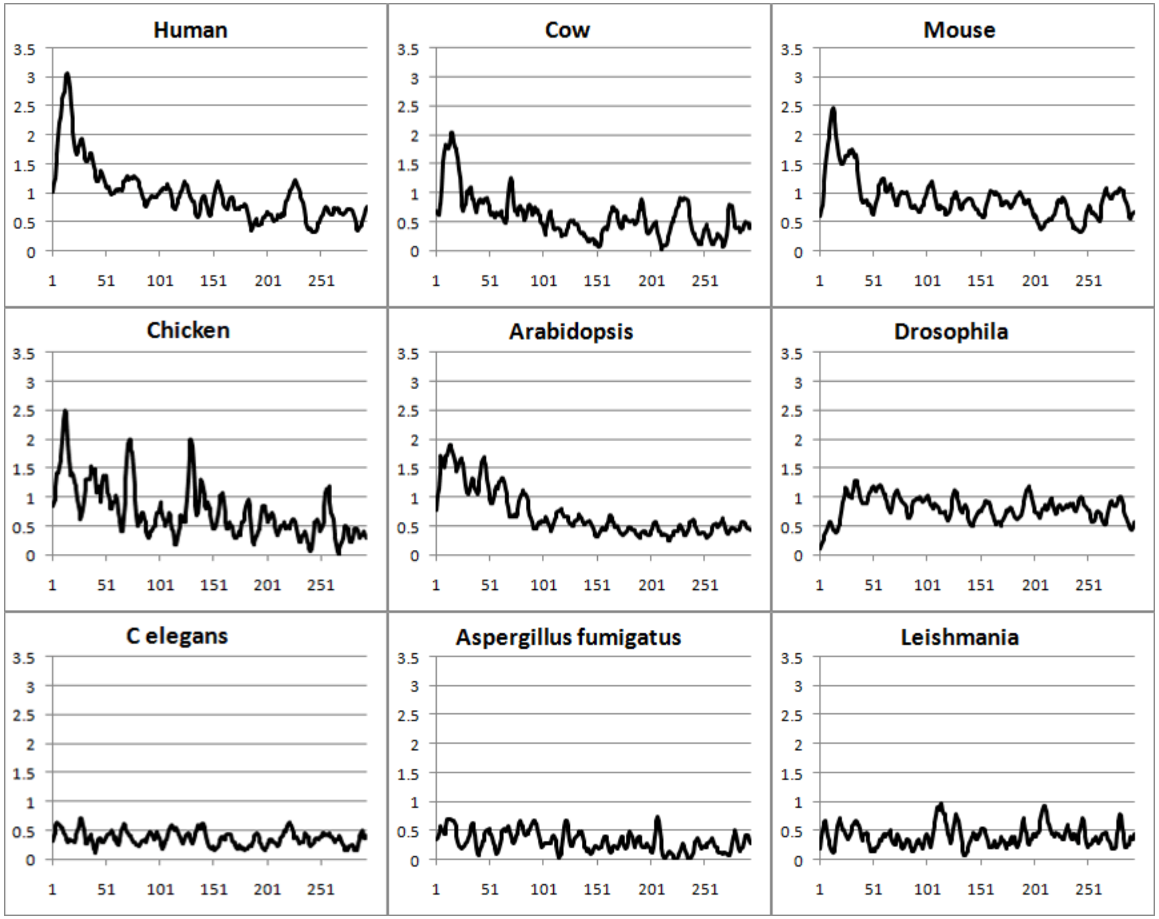
but also they overlap,

so that the same letters in a sequence
may take part simultaneously
in several different messages

Genome inflation code

Occurrence of homopeptides in protein sequences




Human
Cow
Mouse
Chicken
Arabidopsis
Drosophila
C elegans
Aspergillus fumigatus
Leishmania
3.5
3
2.5
2
1.5
1
0.5
0
1
51
101
151
201
251

Three known pathologically expanding ("aggressive") classes of triplets

GCU (GCU, CUG, UGC, AGC, GCA, CAG) ,

GCC (GCC, CCG, CGC, GGC, GCG, CGG) and

AAG (AAG, AGA, GAA, CTT, TTC, TCT).

Aggressive amino acids encoded by expanding triplets

L is encoded by **CTG** (GCT group) and **CTT** (AAG group),
A – by **GCT, GCA** (both GCT group), **GCC and GCG** (GCC group),
G – by **GGC** (GCC group),
P – by **CCG** (GCC group),
S – by **AGC** (GCT group) and **TCT** (AAG group),
E – by **GAA** (AAG group),
R – by **CGG, CGC** (both GCC group) and **AGA** (AAG group),
Q – by **CAG** (GCT group), and
K – by **AAG** (AAG group),
F – by UUC (AAG group),
C – by UGC (GCU group).

Majority of homopeptides are built from aggressive amino acids

human		eukar.	prokar.
tripeptides	Score	(Faux	(Faux
1st exons	(tripept.)	et al.)	et al.)
1. L3	**4552**	**1446**	**70(5)**
2. A3	**4046**	**5465(3)**	**251(3)**
3. G3	**2972**	**5002(5)**	**310(2)**
4. P3	**2258**	**4157(7)**	**217(4)**
5. S3	**1981**	**5424(4)**	**378(1)**
6. E3	**1630**	**4334(6)**	**67(6)**
7. R3	**1145**	**462**	**60(8)**
8. Q3	**802**	**8022(1)**	**52(9)**
9. K3	**535**	**1920(9)**	**25**
10. V3	414	94	9
11. H3	273	1049	32
12. D3	269	1554	34
13. T3	267	2492(8)	63(7)
14. I3	109	34	3
15. F3	**103**	**175**	**1**
16. C3	**92**	**38**	**0**
17. N3	79	6962(2)	31
18. M3	34	19	0
19. Y3	32	39	4
20. W3	14	3	0
	92%	**75%**	**89%**

Codons, preferentially used for repeating amino acids in various eukaryotes

	G+C%	**E**	**G**	**K**	**L**	**P**	**Q**	**R**	**S**
A.gambiae	55.8	GAG/**GAA**	GGU	AAA	-	CCA	**CAG**	-	**AGC**
D.melan.	53.9	GAG	GGA	AAA/**AAG**	-	CCA	**CAG**	AGG	**AGC**
T.rubrip.	53.5	GAG	-	-	-	-	**CAG**	-	-
R.norveg.	52.6	GAG	**GGC**	AAA/**AAG**	**CUG**	**CCG**	**CAG**	**AGA**	**AGC**
H.sapiens	52.3	GAG	**GGC**	AAA/**AAG**	**CUG**	CCA/**CCG**/CCU	**CAG**	**CGG**	**AGC**
M.musc.	52.0	GAG	**GGC**	AAA/**AAG**	**CUG**	CCA/CCU	**CAG**	**CGG**	**AGC**
G.gallus	51.4	GAG	**GGC**	**AAG**	**CUG**	-	**CAG**	**CGC**	**AGC**
D.rerio	50.2	GAG	-	**AAG**	**CUG**	CCU	**CAG**	**AGA**	UCC
A.thal.	44.6	**GAA**	GGU	**AAG**	**CUU**	CCU	CAA	-	**UCU**
A.mellif.	43.5	-	GGA	AAA/**AAG**	-	-	CAA	AGG	**AGC**
C.elegans	42.9	**GAA**	GGA	**AAG**	**CUU**	CCA	CAA	CGA	UCA
S.cerev.	39.8	**GAA**	-	**AAG**	-	CCA	CAA/**CAG**	-	**AGC**
P.falcip.	23.8	**GAA**	GGA/GGU	AAA	UUA	CCA	CAA	**AGA**	AGU
Dominant codons:		GAG	**GGC**	**AAG**	**CUG**	CCA	**CAG**	**AGA**	**AGC**

Codons most frequently used by aggressive amino acids

	G+C%	**F**	**L**	**S**	**P**	**Q**	**K**	**E**	**C**	**R**	
A.gambiae	55.8	**UUC**	**CUG**	**AGC**	CCC	**CAG**	**AAG**	GAG	**UGC**	**CGG**	**GGC**
D. melan	53.9	**UUC**	**CUG**	**AGC**	CCC	**CAG**	**AAG**	GAG	**UGC**	**CGC**	**GGC**
T. rubrip	53.5	**UUC**	**CUG**	**AGC**	CCC	**CAG**	**AAG**	GAG	**UGC**	AGG	**GGC**
R. norveg	52.6	**UUC**	**CUG**	**AGC**	CCC	**CAG**	**AAG**	**GAA**	**UGC**	AGG	**GGC**
H. sapiens	52.3	**UUC**	**CUG**	**AGC**	CCC	**CAG**	**AAG**	GAG	**UGC**	**CGG**	**GGC**
M. muscul	52.0	**UUC**	**CUG**	**AGC**	CCU	**CAG**	**AAG**	GAG	**UGC**	AGG	**GGC**
G. gallus	51.4	**UUC**	**CUG**	**AGC**	CCC	**CAG**	**AAG**	GAG	**UGC**	**AGA**	**GGC**
D. rerio	50.2	**UUC**	**CUG**	**AGC**	CCU	**CAG**	**AAG**	GAG	UGU	**AGA**	GGA
A. thal	44.6	UUU	**CUU**	**UCU**	CCU	CAA	**AAG**	**GAA**	UGU	**AGA**	GGA
A. mellif	43.5	**UUC**	UUG	**UCU**	CCA	CAA	AAA	**GAA**	**UGC**	**AGA**	GGA
C. eleg	42.9	**UUC**	**CUU**	UCA	CCA	CAA	AAA	**GAA**	UGU	**AGA**	GGA
S. cerev	39.8	UUU	UUG	**UCU**	CCA	CAA	AAA	**GAA**	UGU	**AGA**	GGU
P. falcip	23.8	UUU	UUA	AGU	CCA	CAA	AAA	**GAA**	UGU	AGU	GGA
dominant codon:		**UUC**	**CUG**	**AGC**	CCC	**CAG**	**AAG**	GAG	**UGC**	**AGA**	**GGC**

Protein sequences evolve as a mosaic of expanding amino acids,
homopeptides at the moment of expansion event,
gradually mutating to their modern sequence appearance
not recognizable as repeats anymore

Edward N. Trifonov

(kakhol ve lavan)
(blue and white)