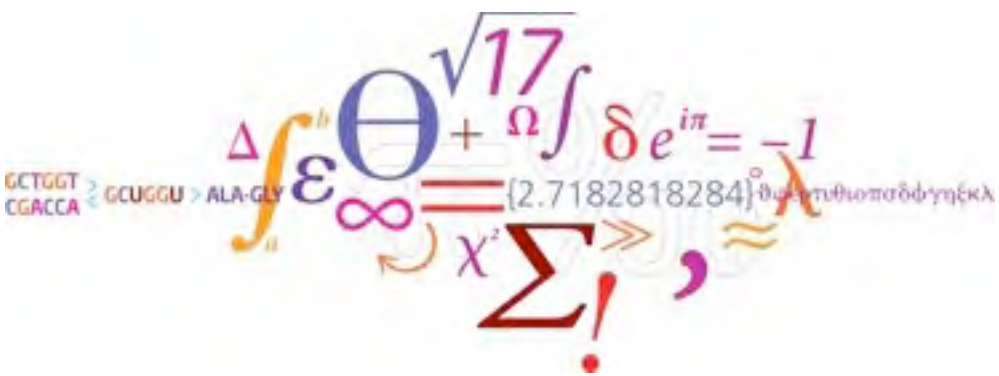
Comparative Genomics

Introduction to DNA atlases

(DNA structures)



Dave Ussery

Genomics of Prokaryotes Workshop
Universidad Miguel Hernández
Alicante, Spain

Monday, 12 December, 2011

Center for Biological Sequence Analysis
Department of Systems Biology

Monday, 12 December, 2011

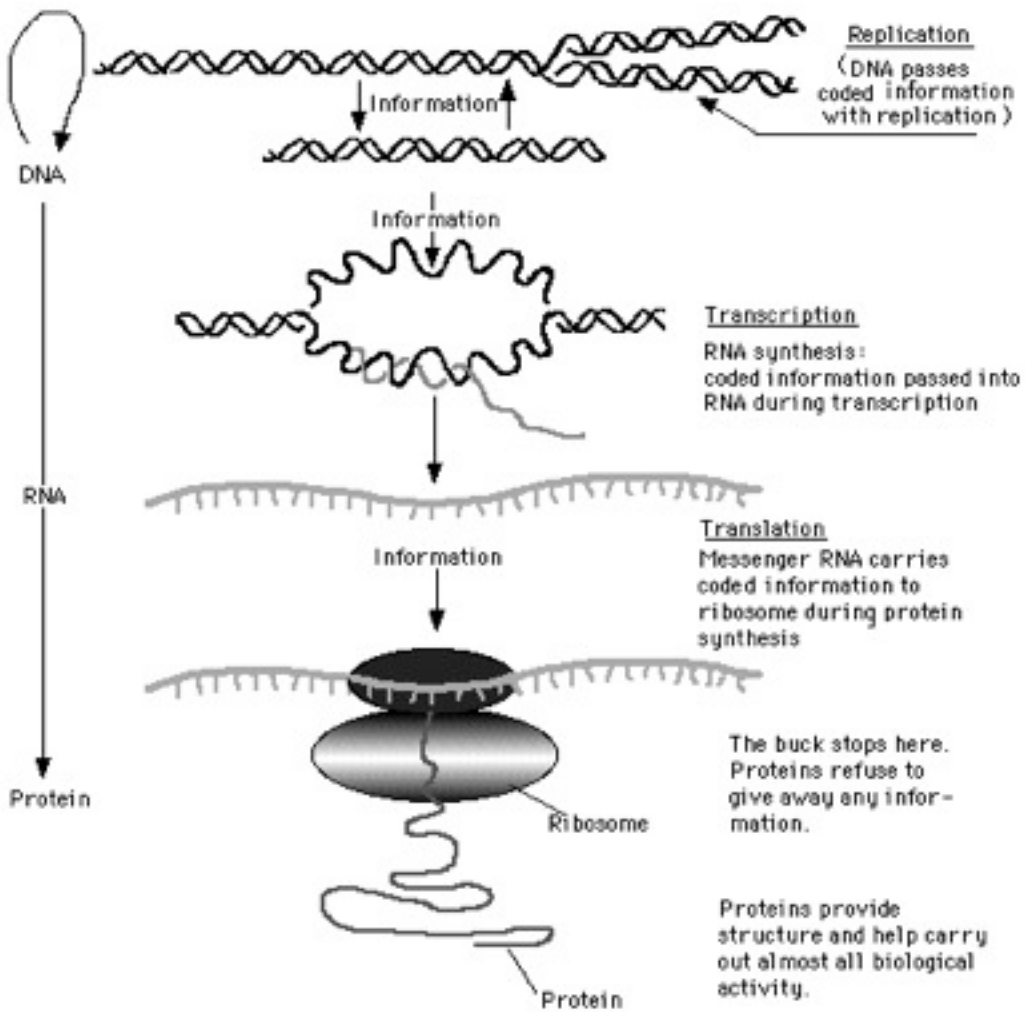
THE OXFORD ENGLISH DICTIONARY

bioinformatics, n.

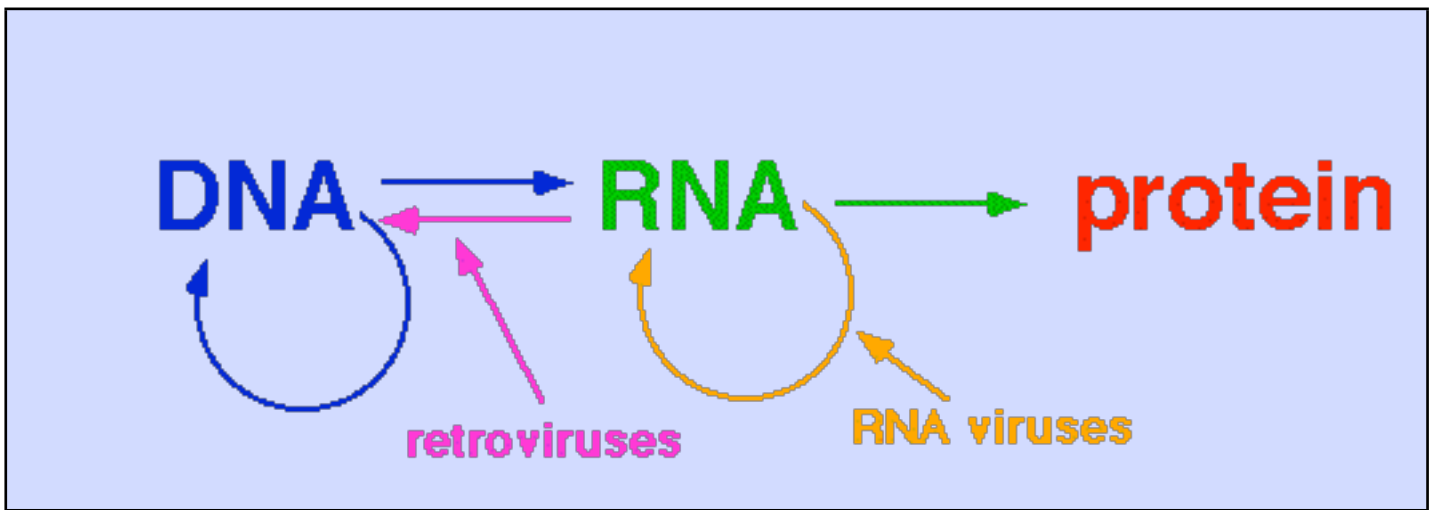
The science of information and information flow in biological systems, esp. the use of computational methods in genetics and genomics.

1978 P. HOGEWEG in *Simulation* **31** 90/1 Since 1970 she has been a staff member at the Subfaculty of Biology of the University of Utrecht, with her main field of research in bioinformatics. **1985** *Jrnl. Theoret. Biol.* **113** 719 (*heading*) Tumor escape from immune elimination... R. J. De Beer, Bioinformatics Group, University of Utrecht. **1986** *Philos. Trans. Royal Soc.* A. **317** 324 The area of modelling mutants from a known structure has been revolutionized by the latest tools of molecular graphics... This is a key element in the whole technology and has attracted much interest (for example, the recent E.E.C. 'Bioinformatics' programme). **1987** *Science* 4 Sept. 1108/3 One of the latest developments [at the European Molecular Biology Laboratory] has been the creation of a new research program in bioinformatics. This is intended to bring together research in computing science, structural biology, and molecular genetics. **1996** *Fast Company* Aug.-Sept. 32/3 A lot of breakthroughs in medicine will come out of the efforts of bio-informatics. **2001** *N.Y. Times* 4 Jan. B6/2 The hope..is to make New York a leader in cutting-edge fields like bioinformatics, in which computers are used to decipher genes and proteins.

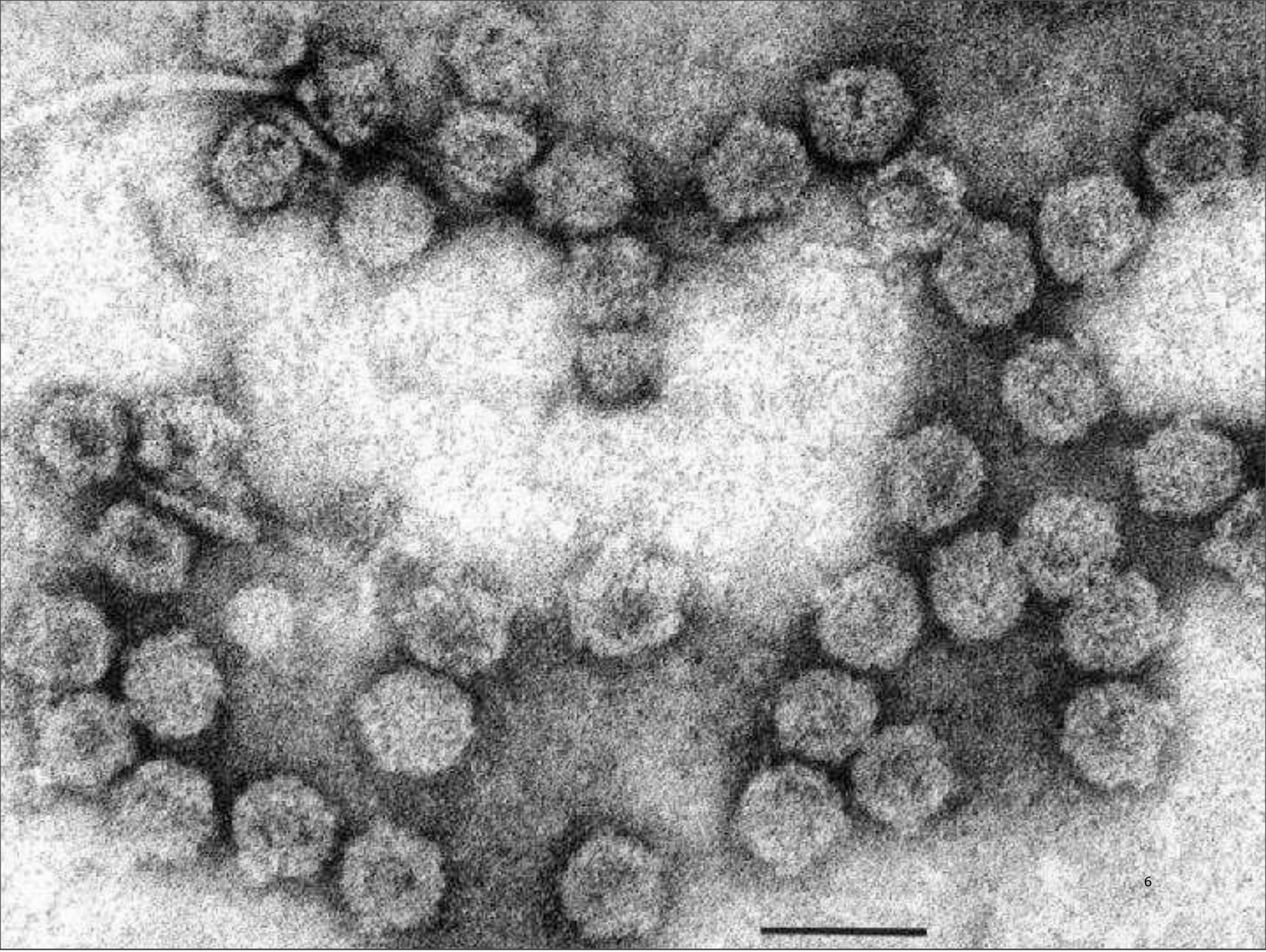
The Central Dogma of Molecular Biology



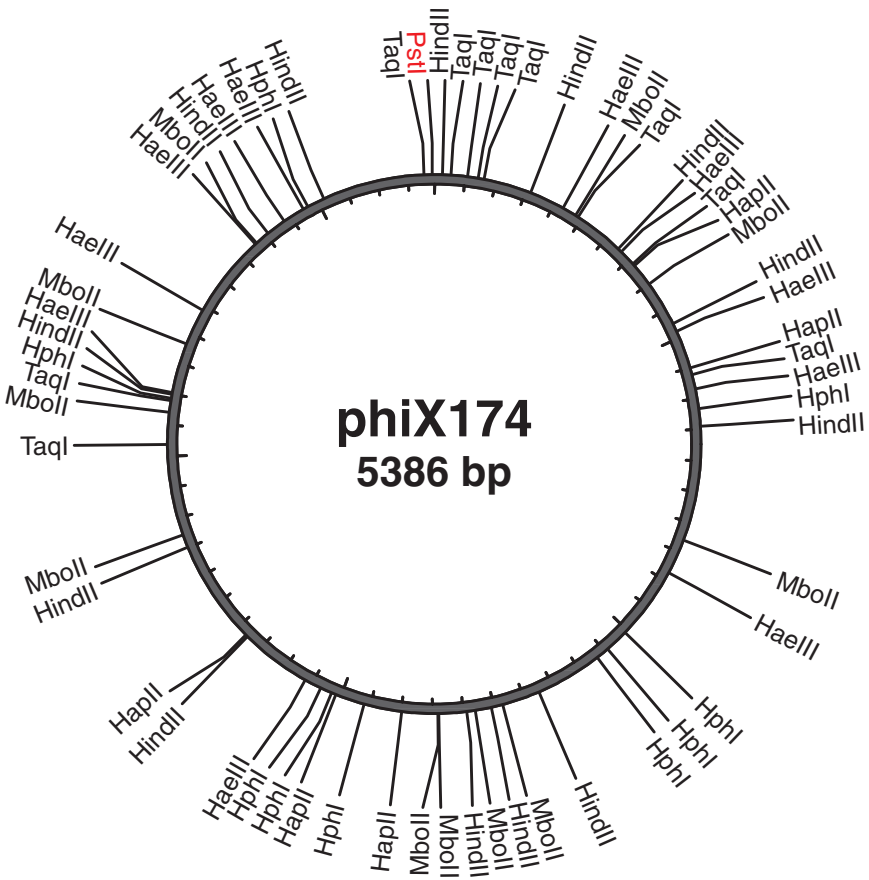
What is Biological Information?

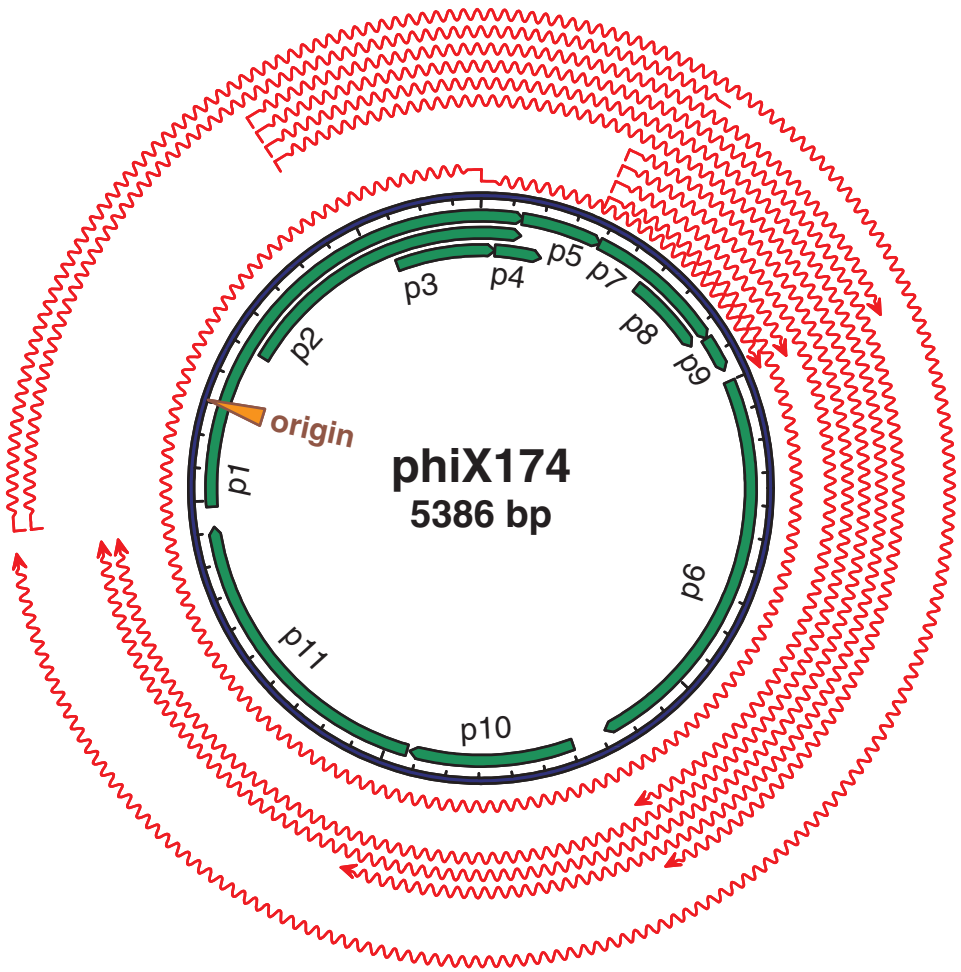


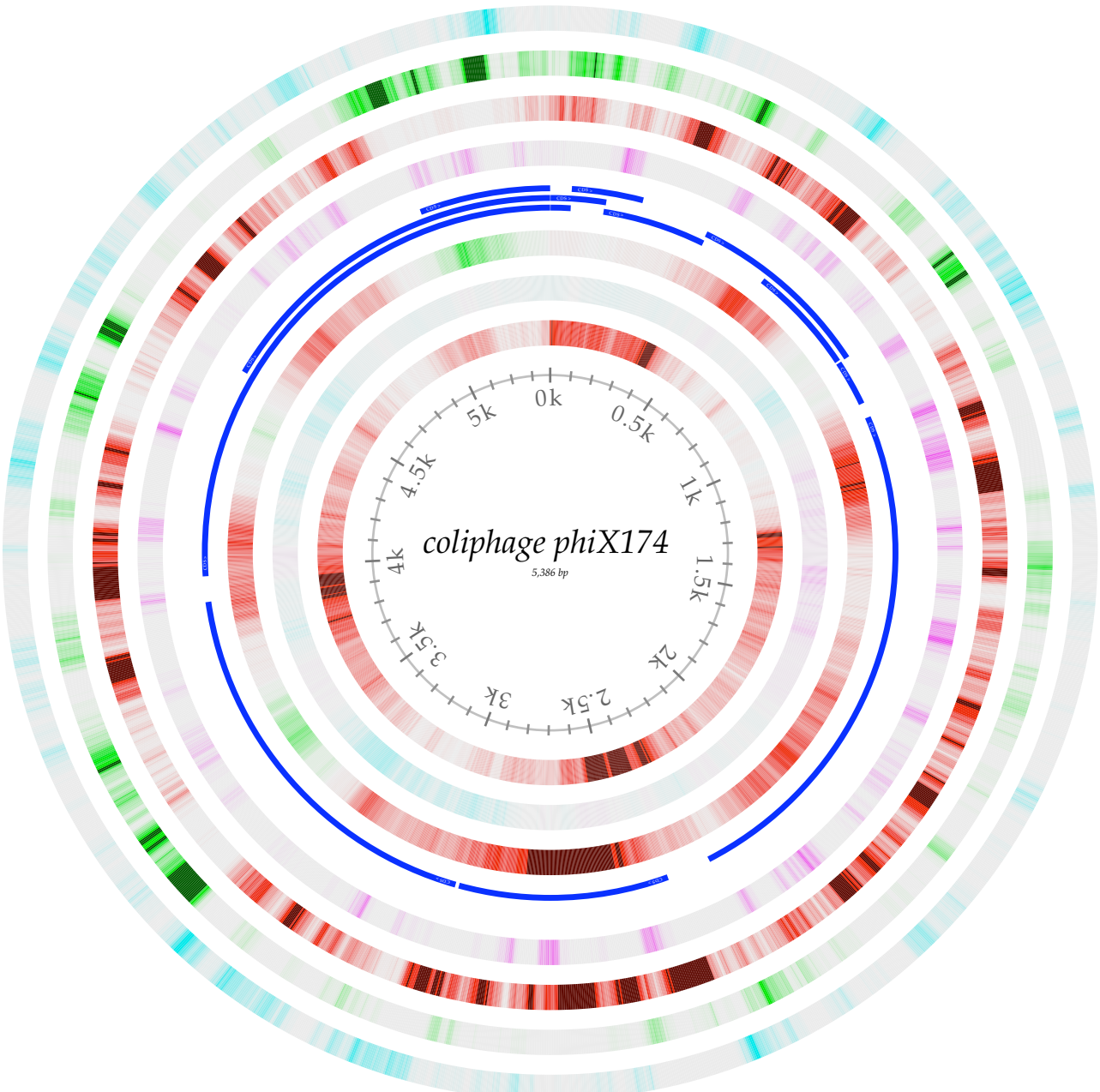
Genome -> Transcriptome ->Proteome



gagttttatc gcttccatga cgcagaagtt aacactttcg gatatttctg atgagtcgaa aaattatctt gataaagcag gaattactac tgcttgttta cgaattaaat cgaagtggac
tgctggcgga aaatgagaaa attcgaccta tccttgcgca gctcgagaag ctcttacttt gcgacctttc gccatcaact aacgattctg tcaaaaactg acgcgttgga tgaggagaag
tggcttaata tgcttggcac gttcgtcaag gactggttta gatatgagtc acattttgtt catggtagag attctcttgt tgacatttta aaagagcgtg gattactatc tgagtccgat
gctgttcaac cactaatagg taagaaatca tgagtcaagt tactgaacaa tccgtacgtt tccagaccgc tttggcctct attaagctca ttcaggcttc tgccgttttg gatttaaccg
aagatgattt cgattttctg acgagtaaca aagtttggat tgctactgac cgctctcgtg ctcgtcgctg cgttgaggct tgcgtttatg gtacgctgga ctttgtggga taccctcgct
ttcctgctcc tgttgagttt attgctgccg tcattgctta ttatgttcat cccgtcaaca ttcaaacggc ctgtctcatc atggaaggcg ctgaatttac ggaaaacatt attaatggcg
tcgagcgtcc ggttaaagcc gctgaattgt tcgcgtttac cttgcgtgta cgcgcaggaa acactgacgt tcttactgac gcagaagaaa acgtgcgtca aaaattacgt gcggaaggag
tgatgtaatg tctaaaggta aaaaacgttc tggcgctcgc cctggtcgtc cgcagccgtt gcgaggtact aaaggcaagc gtaaaggcgc tcgtctttgg tatgtaggtg gtcaacaatt
ttaattgcag gggcttcggc cccttacttg aggataaatt atgtctaata ttcaaactgg cgccgagcgt atgccgcatg acctttccca tcttggcttc cttgctggtc agattggtcg
tcttattacc atttcaacta ctccggttat cgctggcgac tccttcgaga tggacgccgt tggcgctctc cgtctttctc cattgcgtcg tggccttgct attgactcta ctgtagacat
ttttactttt tatgtccctc atcgtcacgt ttatggtgaa cagtggatta agttcatgaa ggatggtgtt aatgccactc ctctcccgac tgttaacact actggttata ttgaccatgc
cgcttttctt ggcacgatta accctgatac caataaaatc cctaagcatt tgtttcaggg ttatttgaat atctataaca actattttaa agcgccgtgg atgcctgacc gtaccgaggc
taaccctaat gagcttaatc aagatgatgc tcgttatggt ttccgttgct gccatctcaa aaacatttgg actgctccgc ttcctcctga gactgagctt tctcgccaaa tgacgacttc
taccacatct attgacatta tgggtctgca agctgcttat gctaatttgc atactgacca agaacgtgat tacttcatgc agcgttacca tgatgttatt tcttcatttg gaggtaaaac
ctcttatgac gctgacaacc gtcctttact tgtcatgcgc tctaatctct gggcatctgg ctatgatgtt gatggaactg accaaacgtc gttaggccag ttttctggtc gtgttcaaca
gacctataaa cattctgtgc cgcgtttctt tgttcctgag catggcacta tgtttactct tgcgcttgtt cgttttccgc ctactgcgac taaagagatt cagtacctta acgctaaagg
tgctttgact tataccgata ttgctggcga ccctgttttg tatggcaact tgccgccgcg tgaaatttct atgaaggatg ttttccgttc tggtgattcg tctaagaagt ttaagattgc
tgagggtcag tggtatcgtt atgcgccttc gtatgtttct cctgcttatc accttcttga aggcttccca ttcattcagg aaccgccttc tggtgatttg caagaacgcg tacttattcg
ccaccatgat tatgaccagt gtttccagtc cgttcagttg ttgcagtgga atagtcaggt taaatttaat gtgaccgttt atcgcaatct gccgaccact cgcgattcaa tcatgacttc
gtgataaaag attgagtgtg aggttataac gccgaagcgg taaaaatttt aatttttgcc gctgaggggt tgaccaagcg aagcgcggta ggttttctgc ttaggagttt aatcatgttt
cagacttta tttctcgcca taattcaaac tttttttctg ataagctggt tctcacttct gttactccag cttcttcggc acctgtttta cagacaccta aagctacatc gtcaacgtta
tattttgata gtttgacggt taatgctggt aatggtggtt ttcttcattg cattcagatg gatacatctg tcaacgccgc taatcaggtt gtttctgttg gtgctgatat tgcttttgat
gccgacccta aatttttgc ctgtttggtt cgctttgagt cttcttcggt tccgactacc ctcccgactg cctatgatgt ttatcctttg aatggtcgcc atgatggtgg ttattatacc
gtcaaggact gtgtgactat tgacgtcctt ccccgtacgc cgggcaataa cgtttatgtt ggtttcatgg tttggtctaa ctttaccgct actaaatgcc gcggattggt ttcgctgaat
aagagattat ttgtctccag ccacttaagt gaggtgattt atgtttggtg ctattgctgg cggtattgct tctgctcttg ctggtggcgc catgtctaaa ttgtttggag gcggtcaaaa
agccgcctcc ggtggcattc aaggtgatgt gcttgctacc gataacaata ctgtaggcat gggtgatgct ggtattaaat ctgccattca aggctctaat gttcctaacc ctgatgaggc
cgcccctagt tttgtttctg gtgctatggc taaagctggt aaaggacttc ttgaaggtac gttgcaggct ggcacttctg ccgtttctga taagttgctt gatttggttg gacttggtgg
caagtctgcc gctgataaag gaaaggatac tcgtgattat cttgctgctg catttcctga gcttaatgct tgggagcgtg ctggtgctga tgcttcctct gctggtatgg ttgacgccgg
atttgagaat caaaaagagc ttactaaaat gcaactggac aatcagaaag agattgccga gatgcaaaat gagactcaaa aagagattgc tggcattcag tcggcgactt cacgccagaa
tacgaaagac caggtatatg cacaaaatga gatgcttgct tatcaacaga aggagtctac tgctcgcgtt gcgtctatta tggaaaacac caatctttcc aagcaacagc aggtttccga
gattatgcgc caaatgctta ctcaagctca aacggctggt cagtatttta ccaatgacca aatcaaagaa atgactcgca aggttagtgc tgaggttgac ttagttcatc agcaaacgca
gaatcagcgg tatggctctt ctcatattgg cgctactgca aaggatattt ctaatgtcgt cactgatgct gcttctggtg tggttgatat ttttcatggt attgataaag ctgttgccga
tacttggaac aatttctgga aagacggtaa agctgatggt attggctcta atttgtctag gaaataaccg tcaggattga caccctccca attgtatgtt ttcatgcctc caaatcttgg
aggctttttt atggttcgtt cttattaccc ttctgaatgt cacgctgatt attttgactt tgagcgtatc gaggctctta aacctgctat tgaggcttgt ggcatttcta ctctttctca
atccccaatg cttggcttcc ataagcagat ggataaccgc atcaagctct tggaagagat tctgtctttt cgtatgcagg gcgttgagtt cgataatggt gatatgtatg ttgacggcca
taaggctgct tctgacgttc gtgatgagtt tgtatctgtt actgagaagt taatggatga attggcacaa tgctacaatg tgctccccca acttgatatt aataacacta tagaccaccg
ccccgaaggg gacgaaaaat ggtttttaga gaacgagaag acggttacgc agttttgccg caagctggct gctgaacgcc ctcttaagga tattcgcgat gagtataatt accccaaaaa
gaaaggtatt aaggatgagt gttcaagatt gctggaggcc tccactatga aatcgcgtag aggctttgct attcagcgtt tgatgaatgc aatgcgacag gctcatgctg atggttggtt
tatcgttttt gacactctca cgttggctga cgaccgatta gaggcgtttt atgataatcc caatgctttg cgtgactatt ttcgtgatat tggtcgtatg gttcttgctg ccgagggtcg
caaggctaat gattcacacg ccgactgcta tcagtatttt tgtgtgcctg agtatggtac agctaatggc cgtcttcatt tccatgcggt gcactttatg cggacacttc ctacaggtag
cgttgaccct aattttggtc gtcgggtacg caatcgccgc cagttaaata gcttgcaaaa tacgtggcct tatggttaca gtatgcccat cgcagttcgc tacacgcagg acgcttttc
acgttctggt tggttgtggc ctgttgatgc taaaggtgag ccgcttaaag ctaccagtta tatggctgtt ggtttctatg tggctaaata cgttaacaaa aagtcagata tggaccttgc
tgctaaaggt ctaggagcta aagaatggaa caactcacta aaaaccaagc tgtcgctact tcccaagaag ctgttcagaa tcagaatgag ccgcaacttc gggatgaaaa tgctcacaat
gacaaatctg tccacggagt gcttaatcca acttaccaag ctgggttacg acgcgacgcc gttcaaccag atattgaagc agaacgcaaa aagagagatg agattgaggc tgggaaaagt
tactgtagcc gacgttttgg cggcgcaacc tgtgacgaca aatctgctca aatttatgcg cgcttcgata aaaatgattg gcgtatccaa cctgca


phiX174
5386 bp
TaqI
PstI
HindII
TaqI
TaqI
TaqI
TaqI
HindII
HaeIII
MboII
TaqI
HindII
HaeIII
TaqI
HapII
MboII
HindII
HaeIII
HapII
TaqI
HaeIII
HphI
HindII
MboII
HindII
HaeIII
HphI
TaqI
MboII
TaqI
MboII
HindII
HapII
HindII
HaeIII
HphI
HphI
HapII
HphI
HapII
MboII
MboII
HindII
MboII
HindII
HphI
HphI
HphI
HindII


phiX174
5386 bp
origin
p1
p2
p3
p4
p5
p7
p8
p9
p6
p10
p11



G Content
fix avg
0.00 0.40

A Content
fix avg
0.00 0.40

T Content
fix avg
0.00 0.40

C Content
fix avg
0.00 0.40

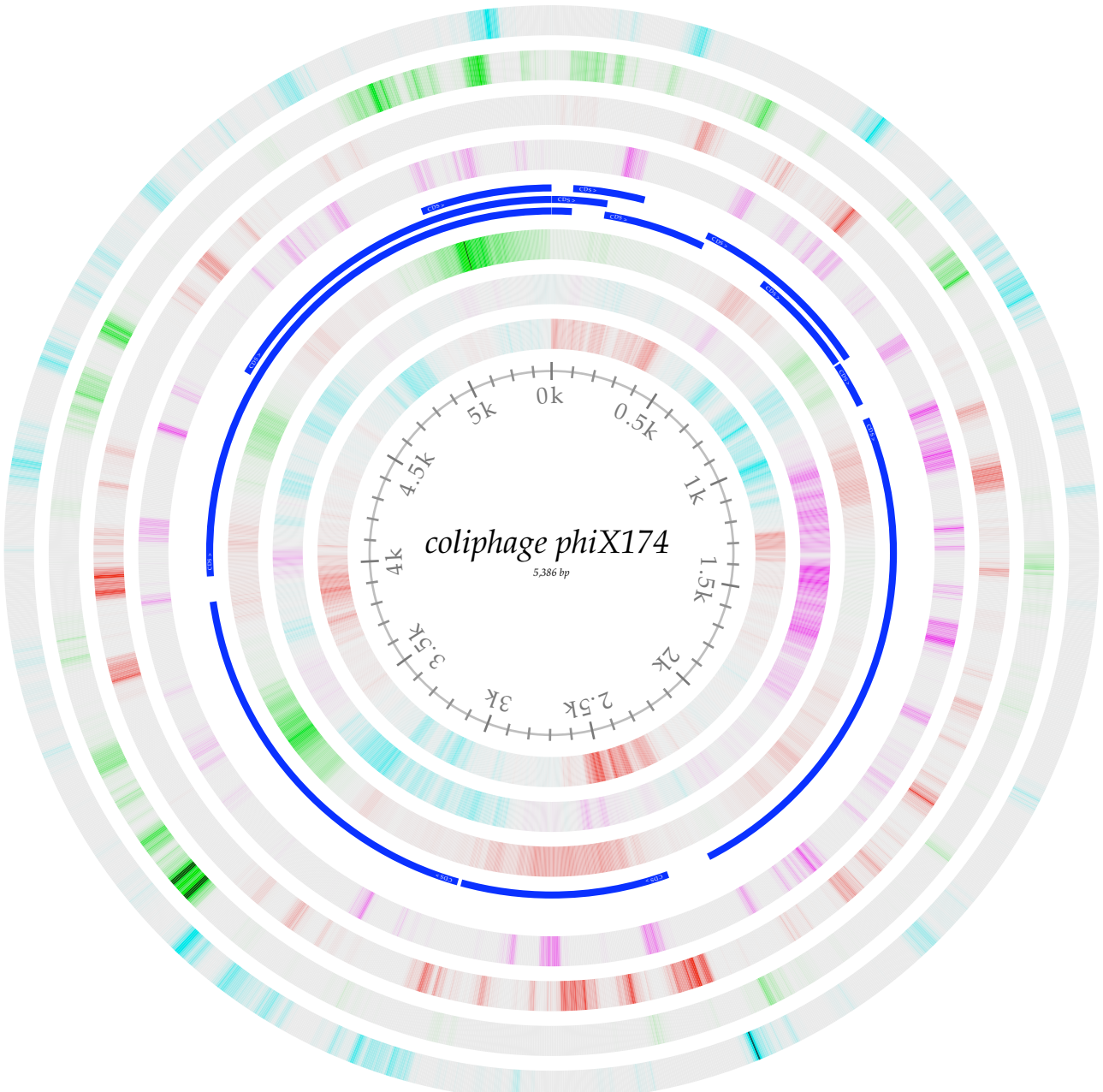
Annotations:
CDS +

AT Skew
fix avg
-0.20 0.20

GC Skew
fix avg
-0.20 0.20

Percent AT
fix avg
0.40 0.60

Resolution: 3



G Content
dev avg
0.07 0.39

A Content
dev avg
0.01 0.47

T Content
dev avg
0.10 0.53

C Content
dev avg
0.04 0.39

Annotations:
CDS +

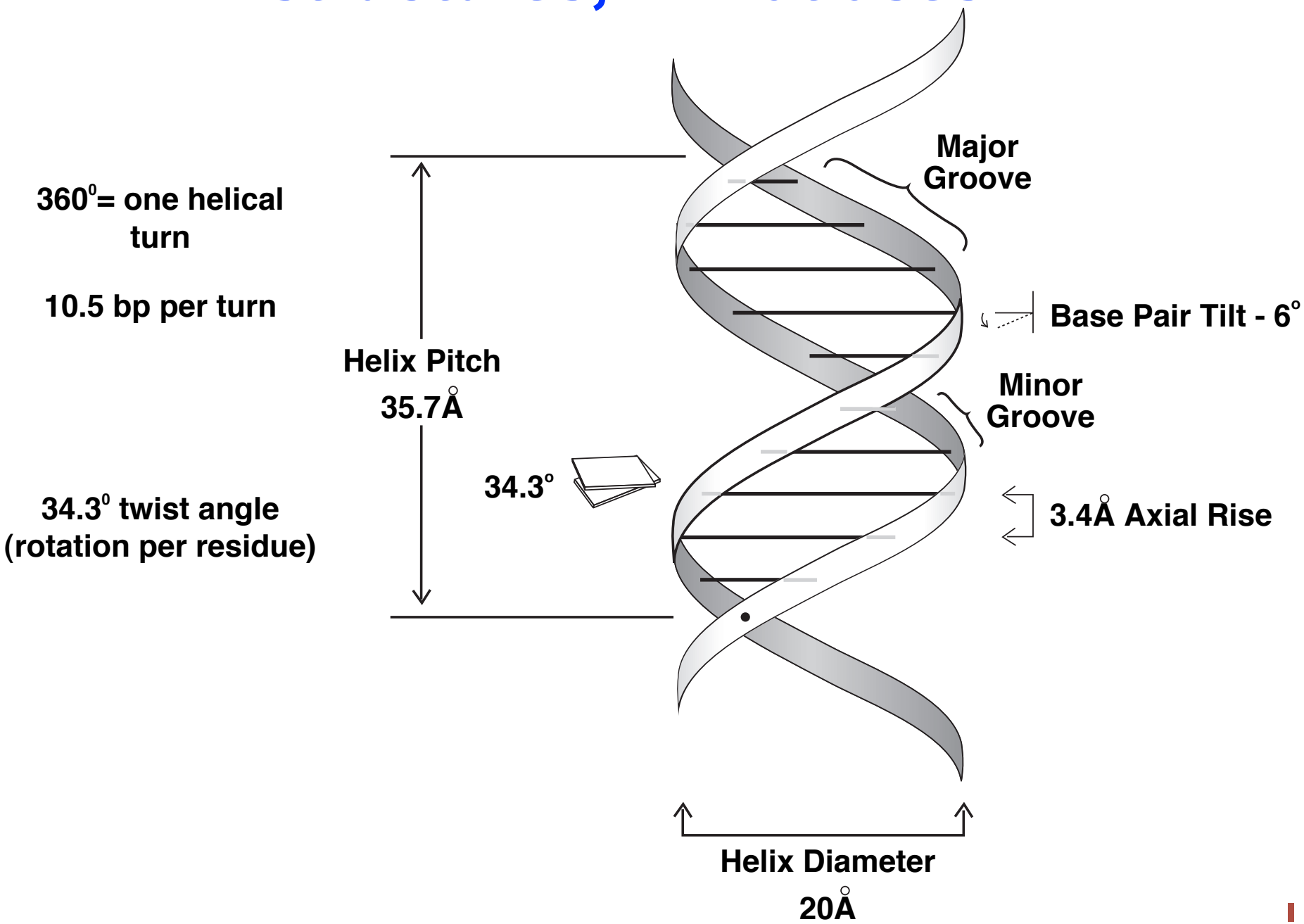
AT Skew
dev avg
-0.33 0.18

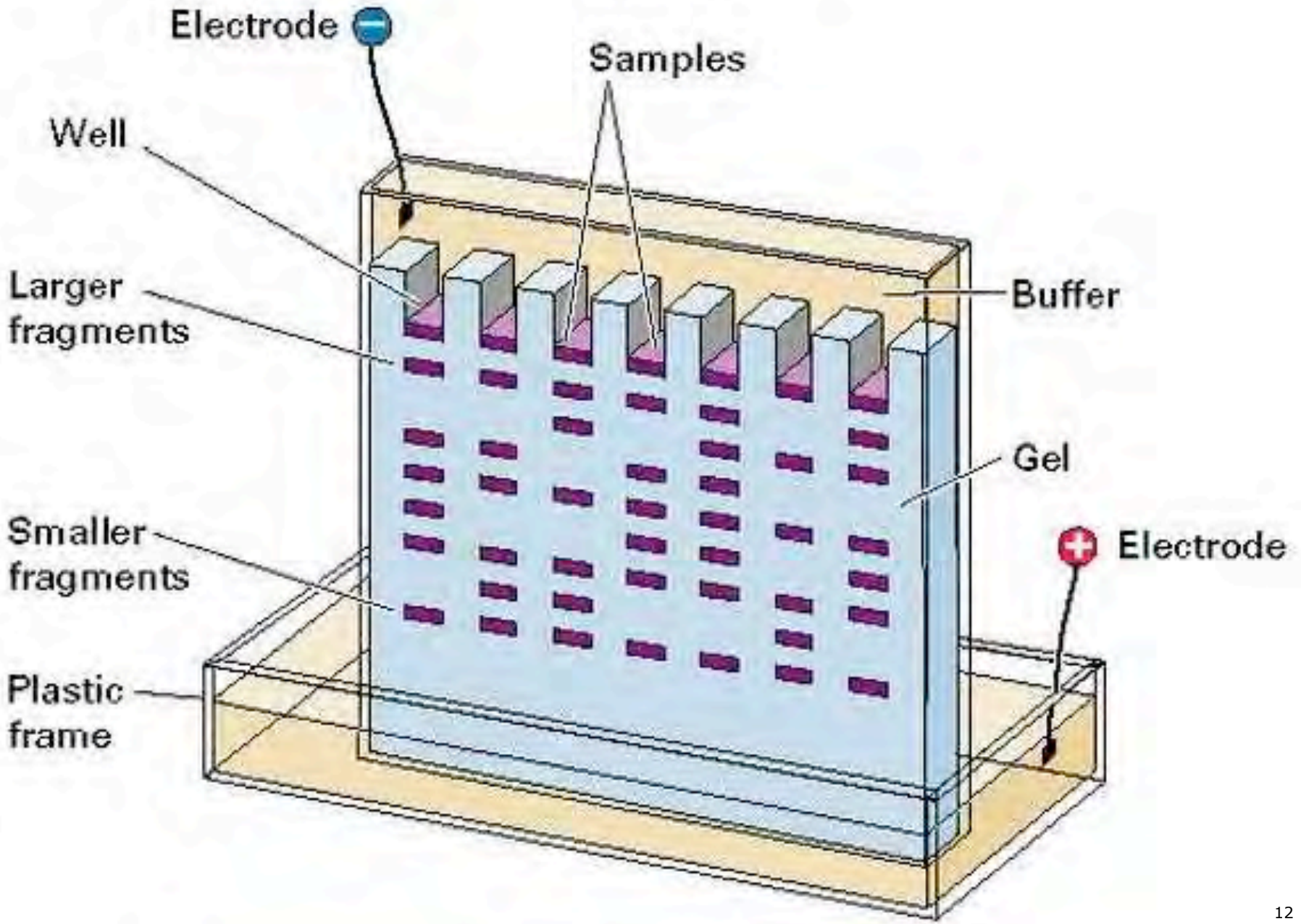
GC Skew
dev avg
-0.10 0.14

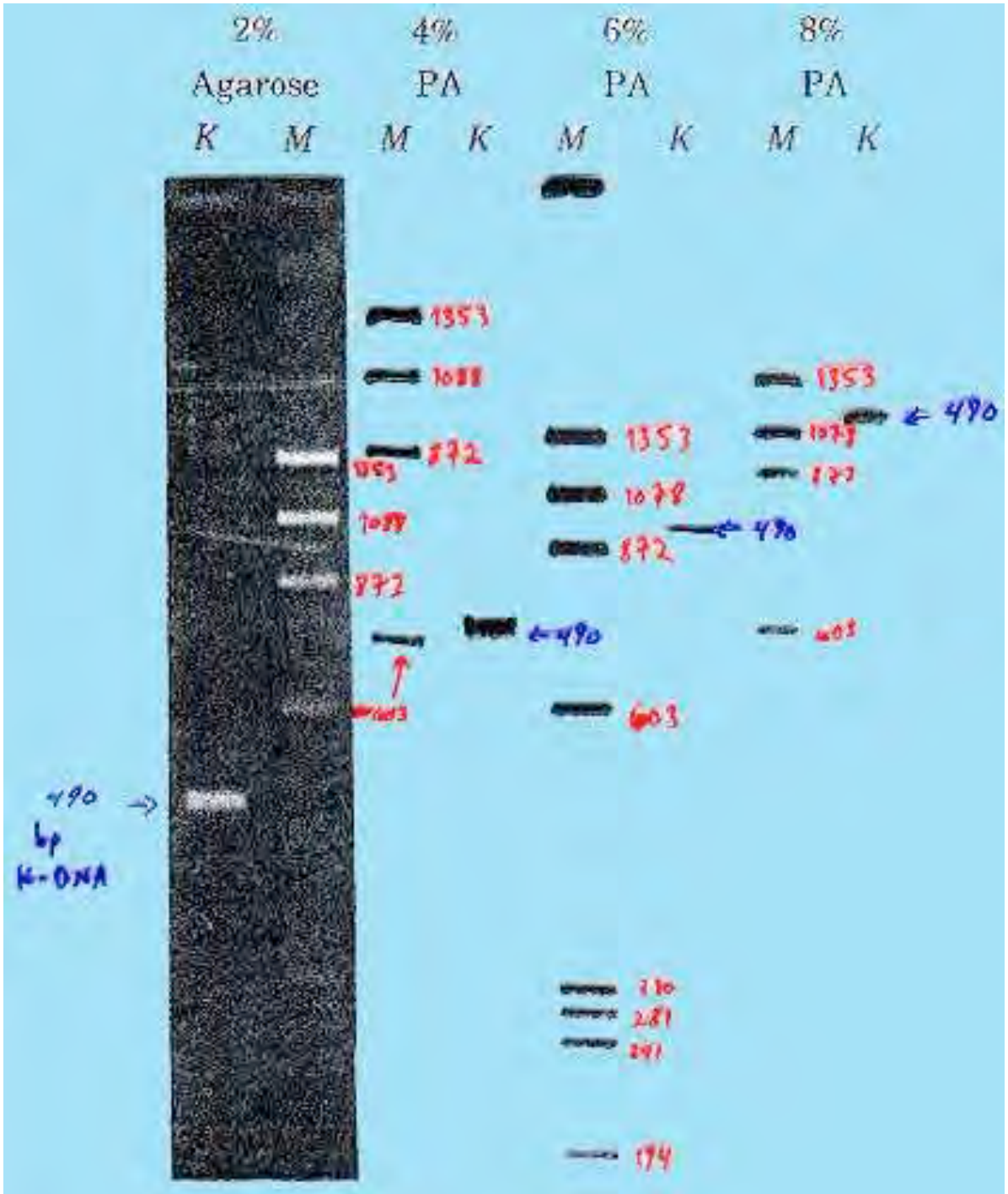
Percent AT
dev avg
0.47 0.63

Resolution: 3

a digression - a few words about DNA structures, DNAatlases....



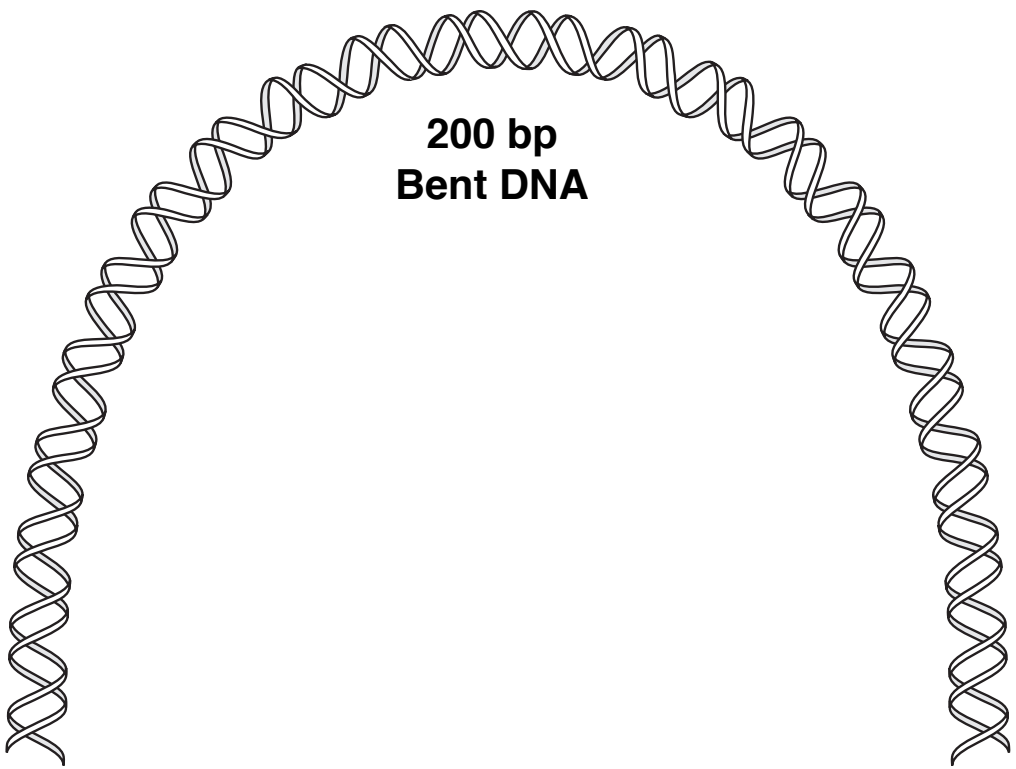

Electrode
Samples
Well
Larger fragments
Buffer
Gel
Smaller fragments
Electrode
Plastic frame


2%
Agarose
4%
PA
6%
PA
8%
PA
K
M
M
K
M
K
M
K
1353
1078
872
603
1078
872
603
490
1353
1078
872
603
490
490
bp
λ-DNA
310
281
271
194

A.

200 bp
Straight DNA

B.



200 bp
Bent DNA

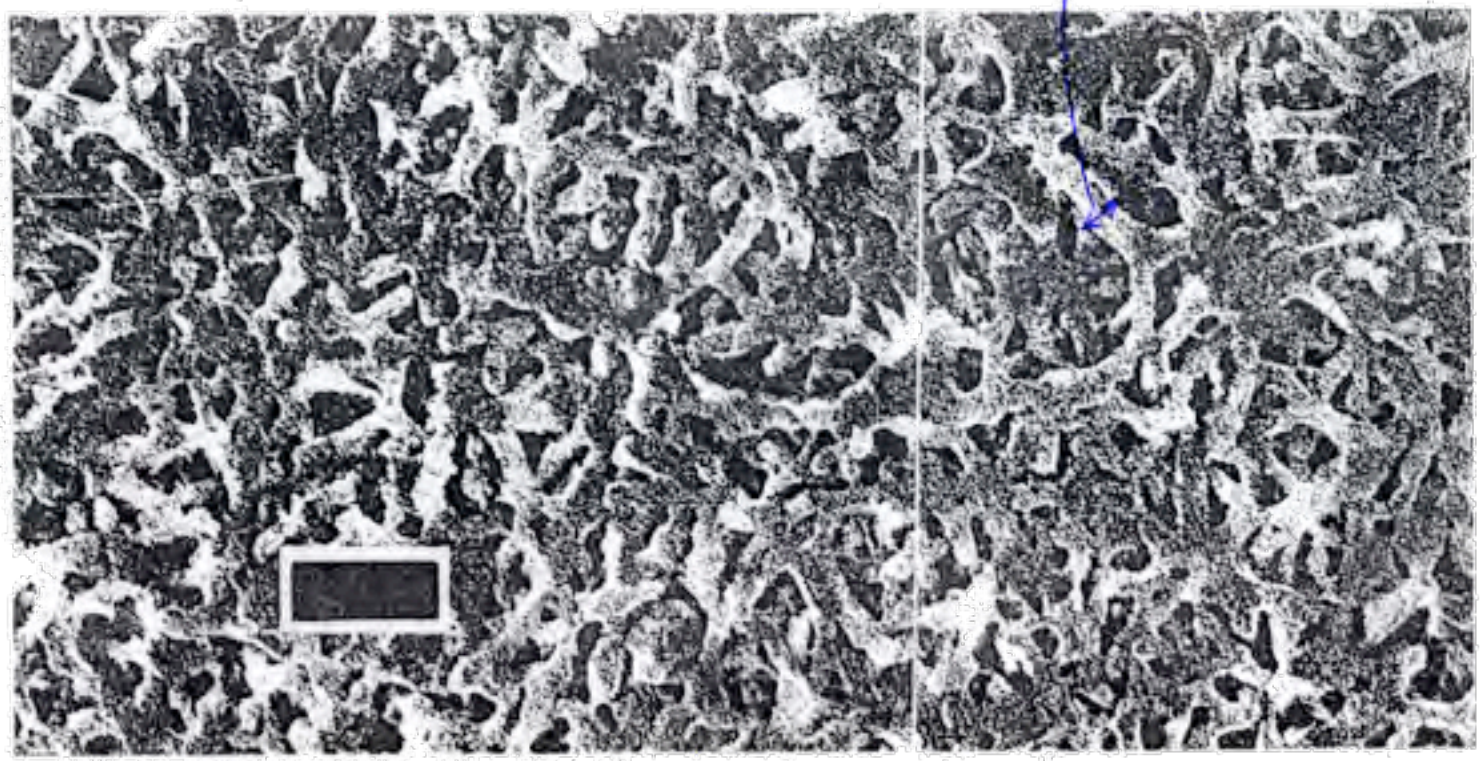
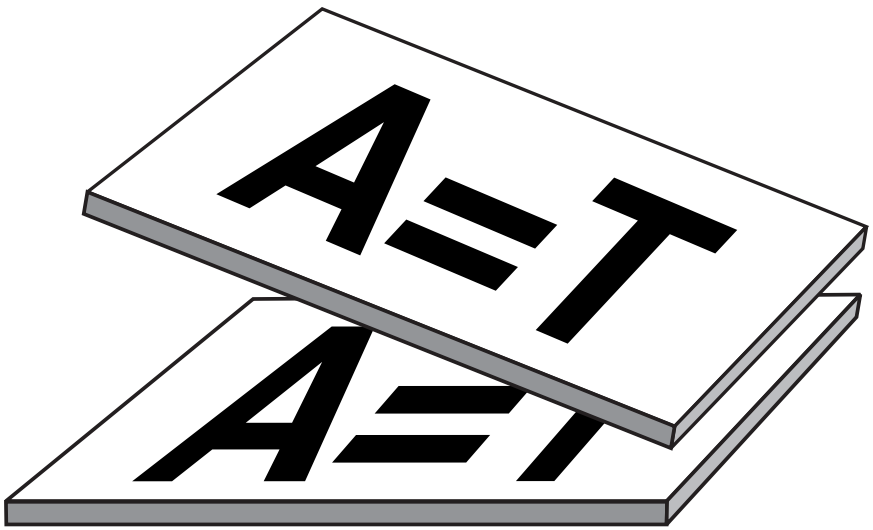
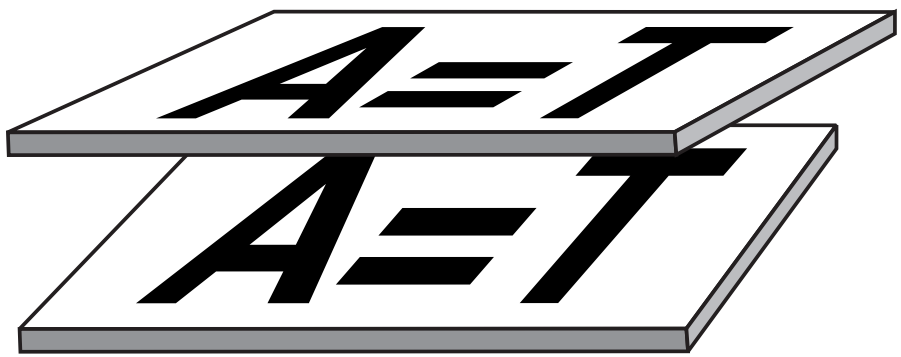


Figure 8.10 Electron micrograph of a portion of a 2% agarose gel, 1 μm × 0.5 μm overall: small black rectangle is 1000 Å × 500 Å. Individual gel fibers are about 100 Å wide. Courtesy of Sue Whytock and John Finch.

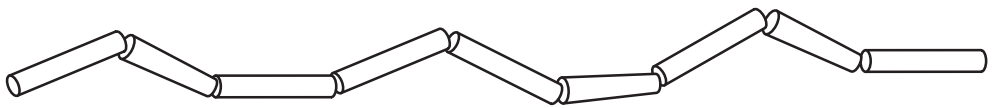


Tilt

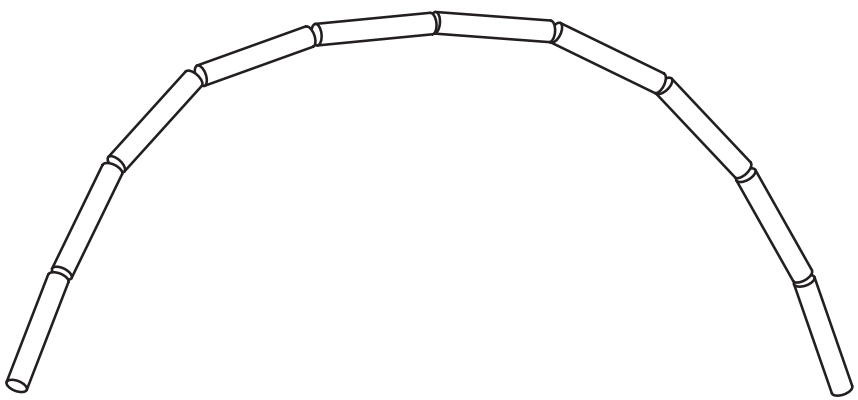


Roll

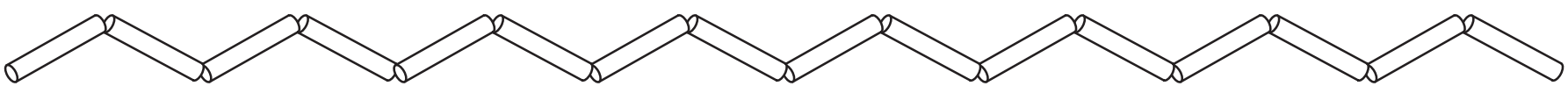
Random Bends



Bends Phased at 10.5 bp



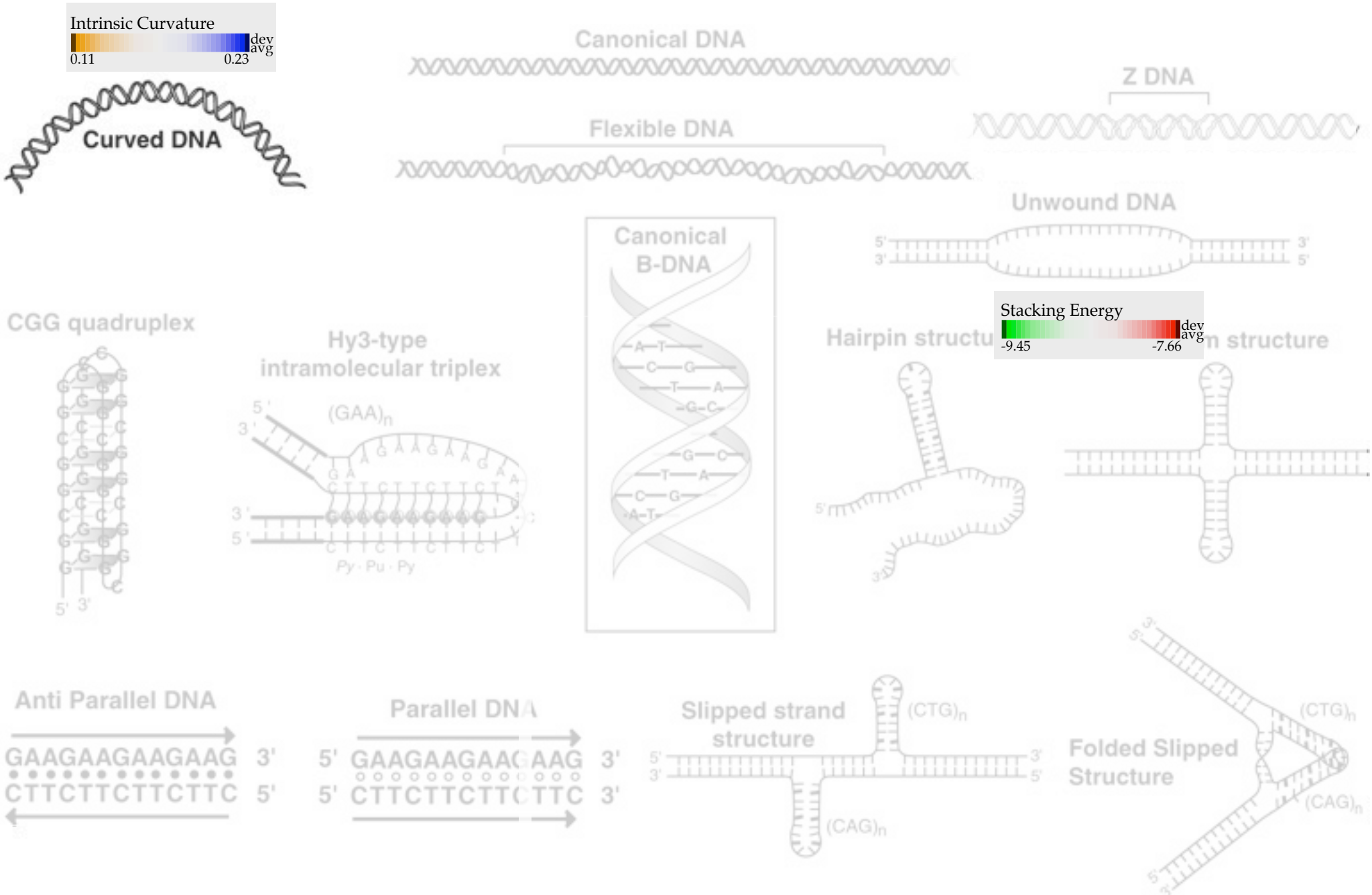
Bends Phased at 16 bp

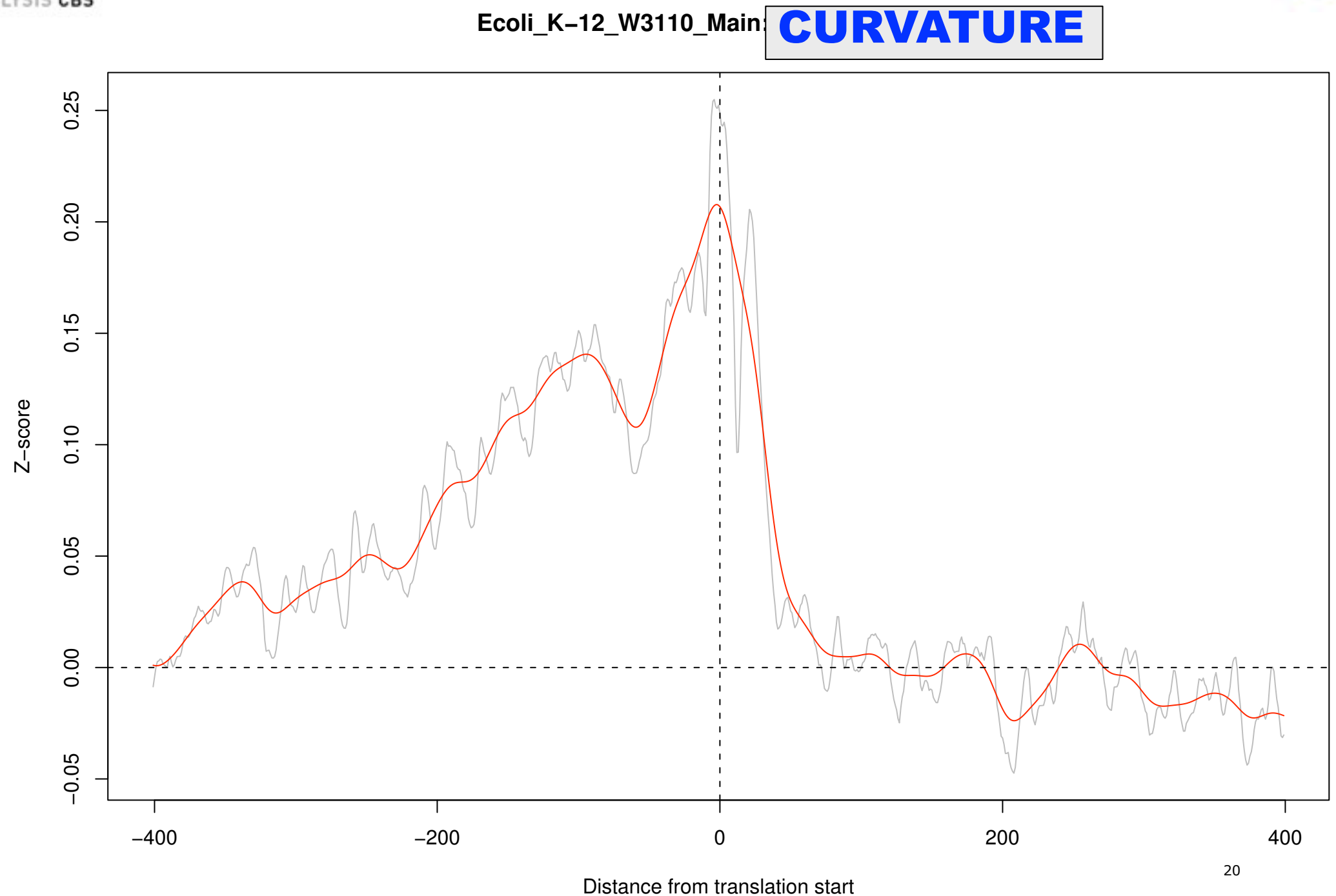


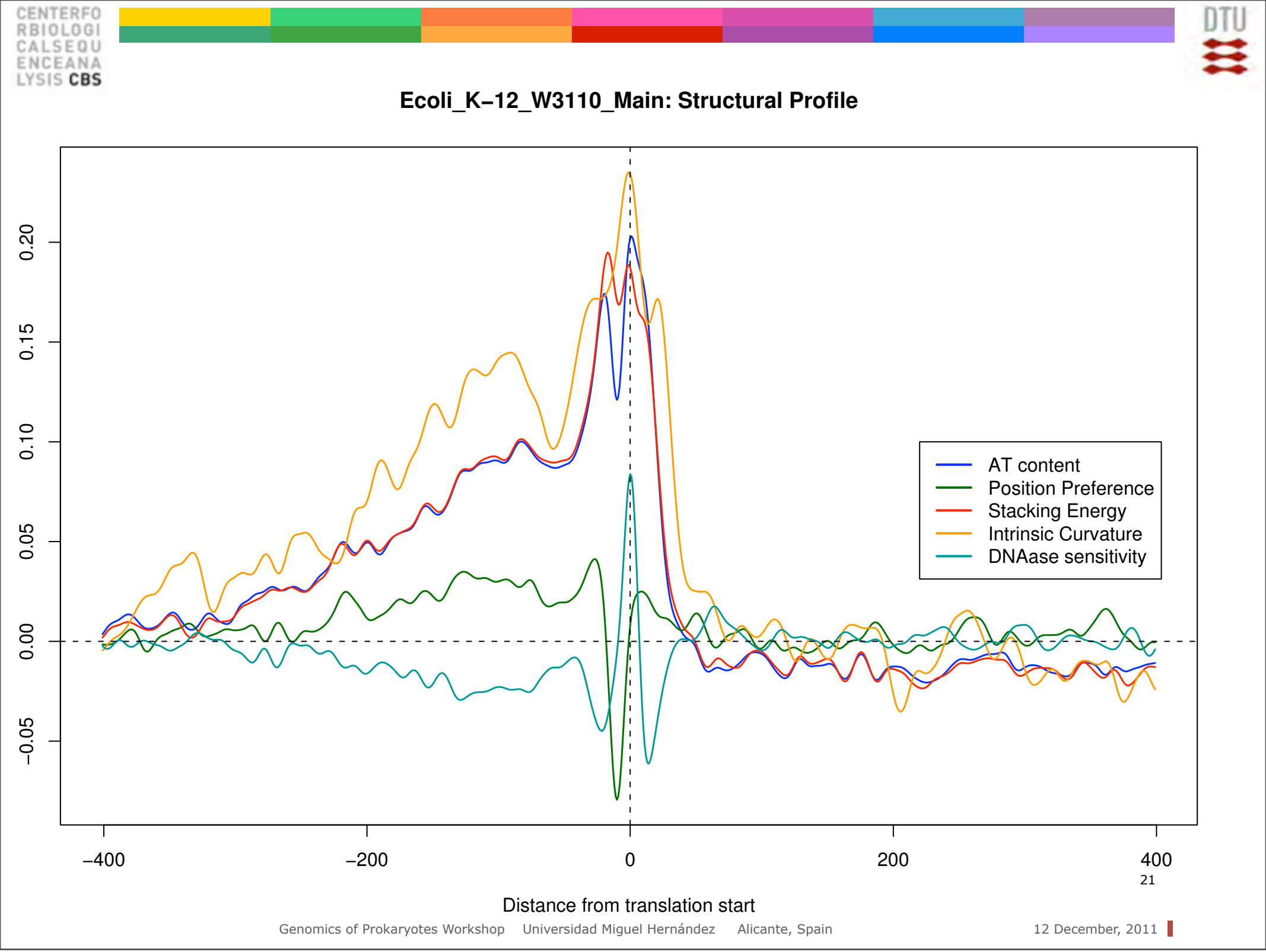
1 11 21 31 41

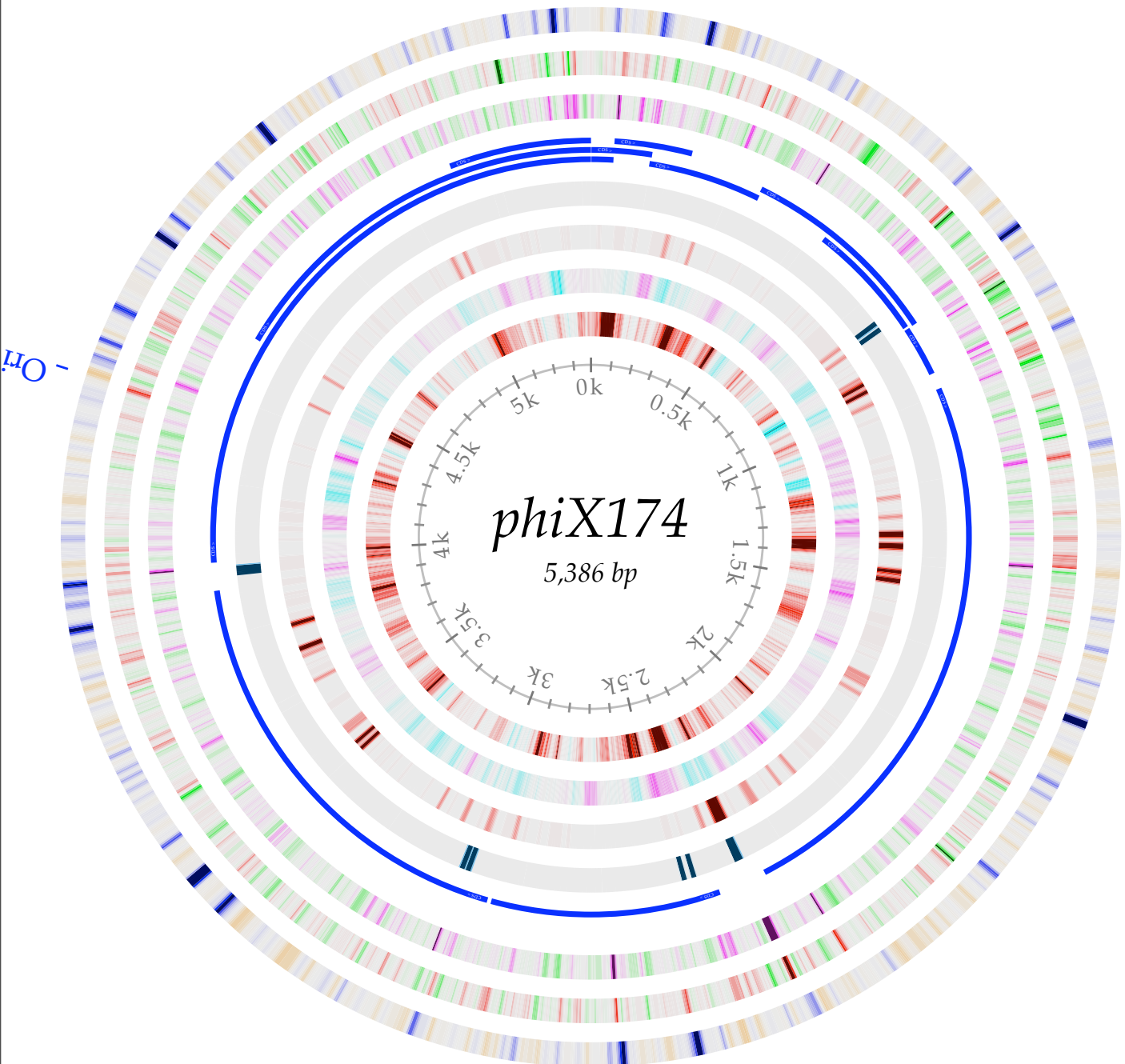
. . . . .

CCCAAAAATGTCAAAAAAATAGGCAAAAAATGCCAAAAATCCCAAC


Intrinsic Curvature
dev
avg
0.11
0.23
Curved DNA
Canonical DNA
Z DNA
Flexible DNA
Unwound DNA
Canonical
B-DNA
Stacking Energy
dev
avg
-9.45
-7.66
CGG quadruplex
Hy3-type
intramolecular triplex
Hairpin structu
n structure
Anti Parallel DNA
5' GAAGAAGAAGAAG 3'
3' CTTCTTCTTCTTC 5'
Parallel DNA
5' GAAGAAGAAGAAG 3'
5' CTTCTTCTTCTTC 3'
Slipped strand
structure
(CTG)n
(CAG)n
Folded Slipped
Structure
(CTG)n
(CAG)n


Ecoli_K−12_W3110_Main:
CURVATURE
Z−score
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
−400
−200
0
200
400
Distance from translation start


Ecoli_K−12_W3110_Main: Structural Profile
0.20
0.15
0.10
0.05
0.00
−0.05
AT content
Position Preference
Stacking Energy
Intrinsic Curvature
DNAase sensitivity
−400
−200
0
200
400
Distance from translation start



Intrinsic Curvature

-0.09 0.51

Stacking Energy

-10.31 -5.40

Position Preference

0.03 0.25

Annotations:

CDS +

CDS -

Perfect-palindromes

-0.25 0.28

Inverted Repeats

3.54 7.83

GC Skew

-0.25 0.29

Percent AT

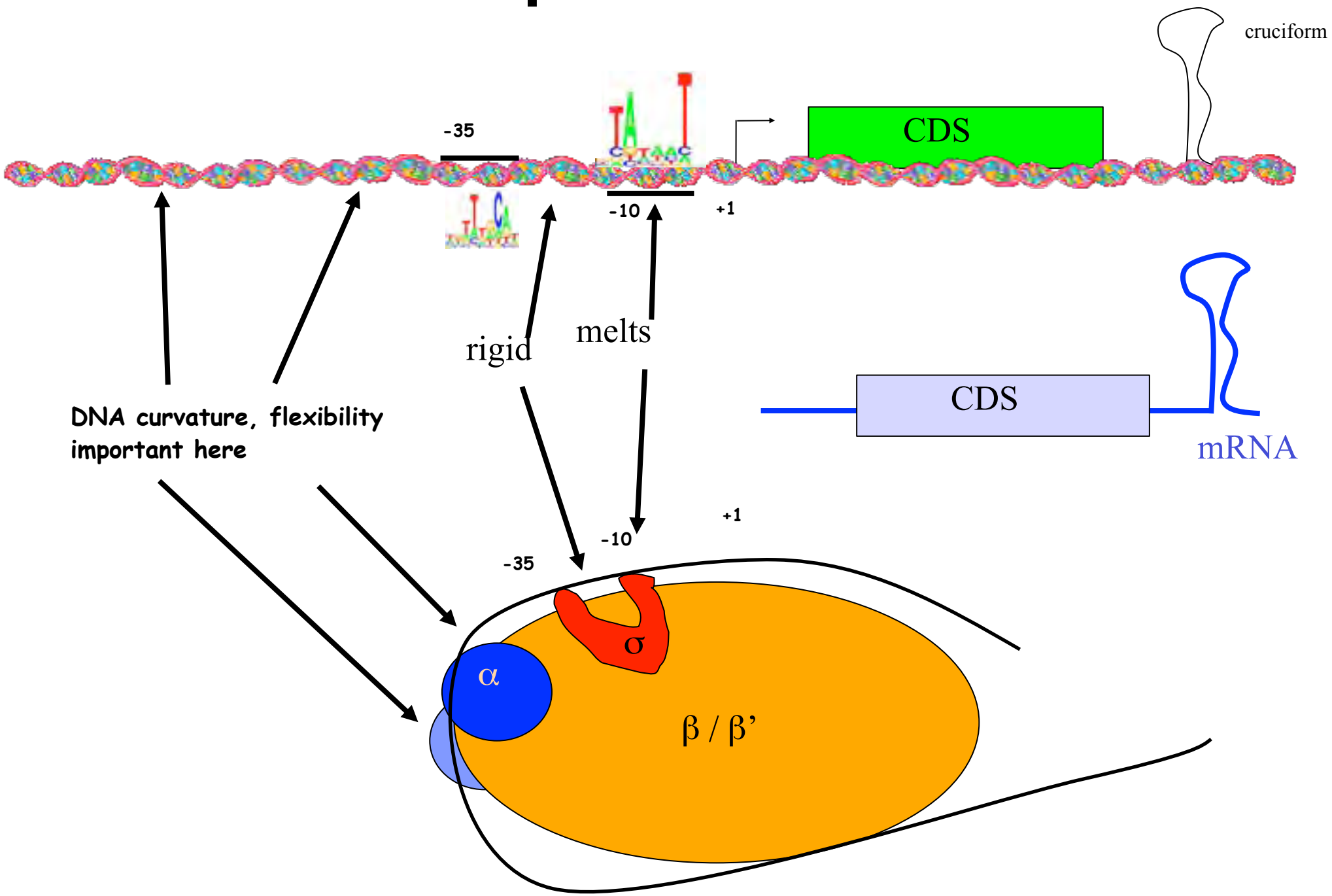
0.35 0.65

Resolution: 3

http://www.cbs.dtu.dk/
Center for Biological Sequence Ar

GENOME ATLAS

Promoter Structural profile



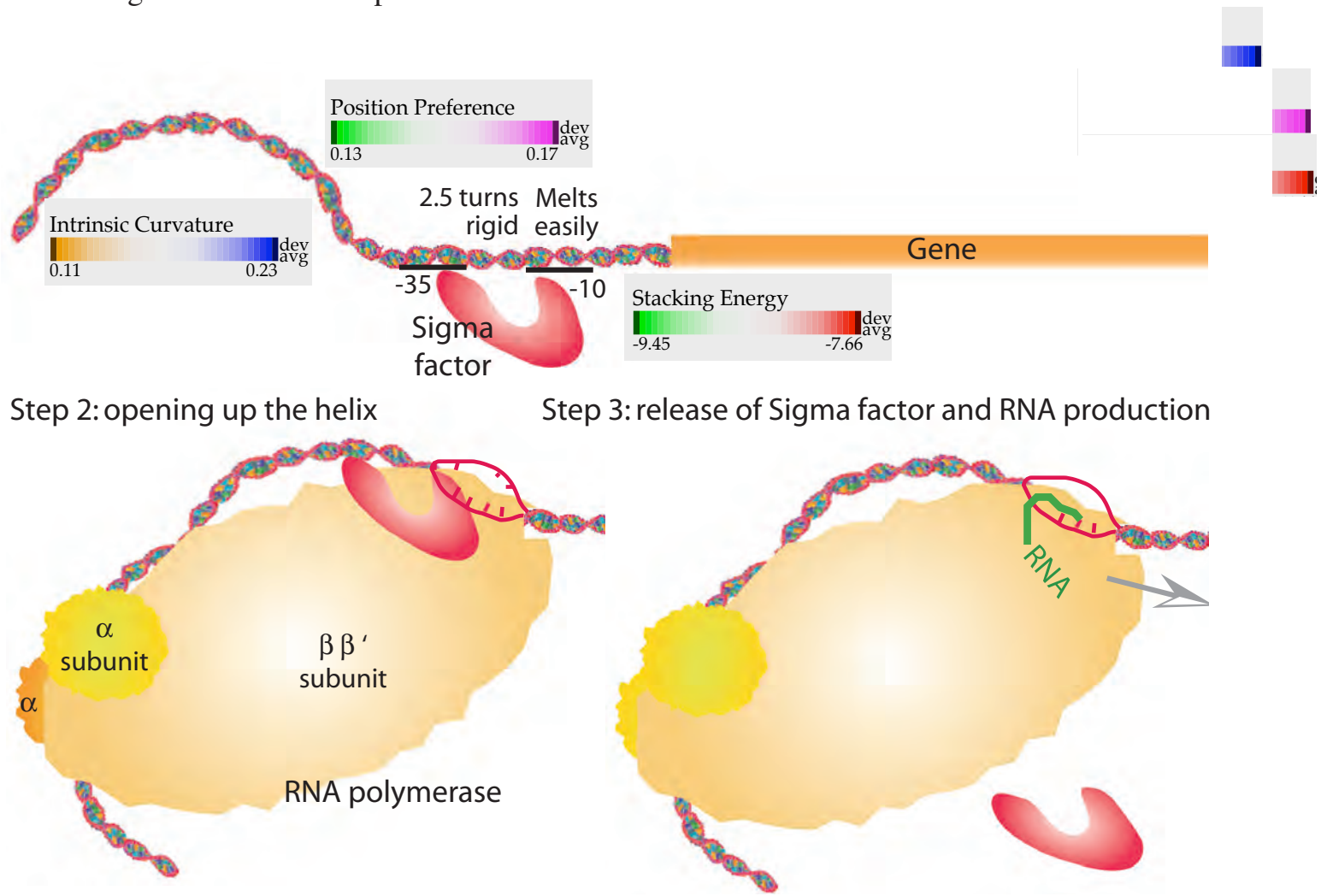
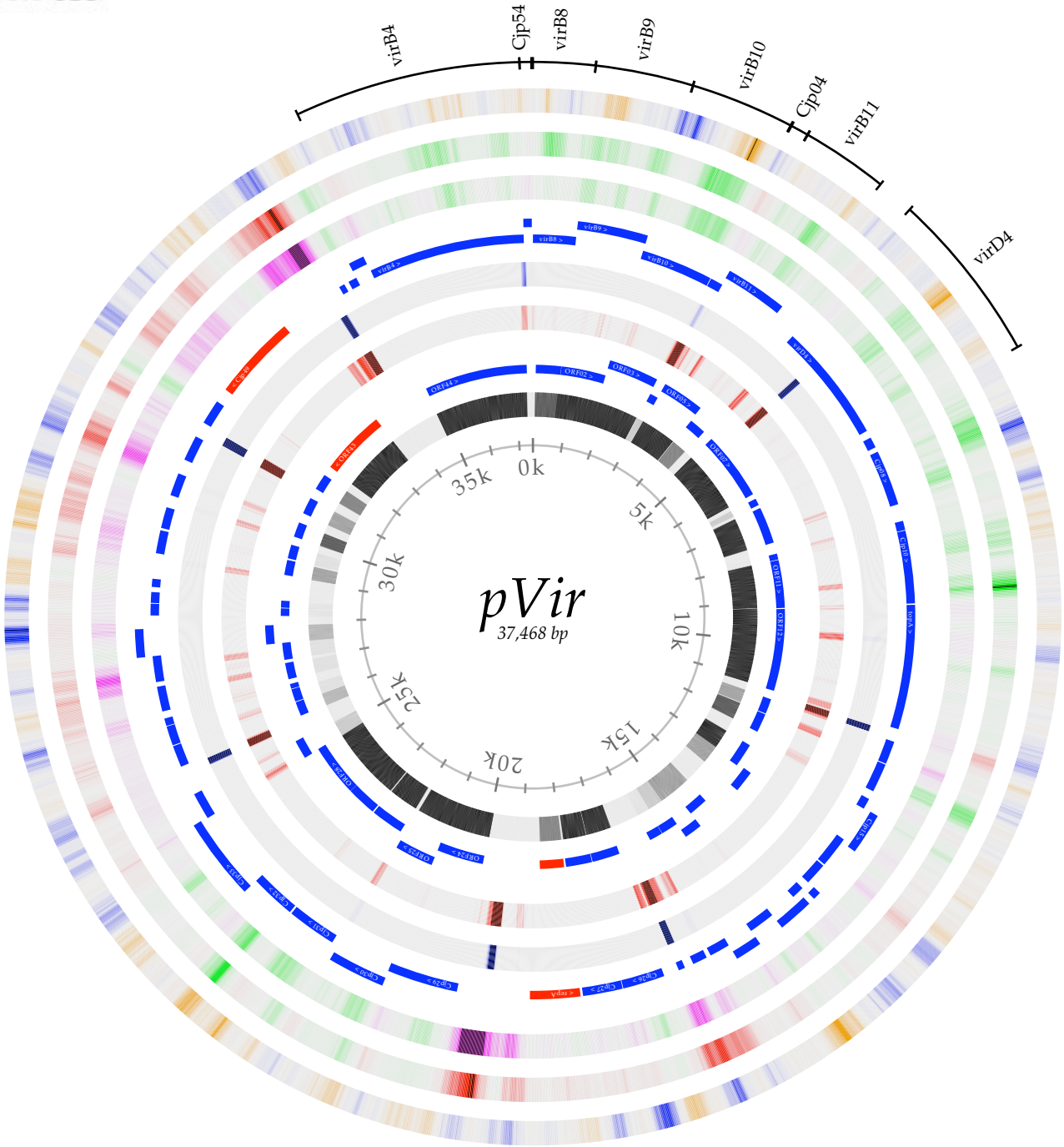


Fig. 10.5 Initiation of transcription in bacteria. In the first step, Sigma factor binds to the DNA on two locations (in the case of Sigma 70 the -35 and -10 sites). RNA polymerase (a complex of two α, one β and one β' subunit) binds next, after which the DNA wraps around the protein. Sigma induces local strand separation so that RNA polymerase starts producing RNA (*in green*). The Sigma factor is then released and RNA polymerase proceeds along the DNA, moving with a local bubble of melted DNA (*indicated by the arrows*)



Intrinsic Curvature
dev avg
0.18 0.34

Stacking Energy
dev avg
-7.55 -6.04

Position Preference
dev avg
0.14 0.22

Annotations:
CDS +
CDS -

Probability of Opening
dev avg
-0.34 0.38

Destabilisation energy
dev avg
1.88 14.98

Annotations:
eCDS +
eCDS -

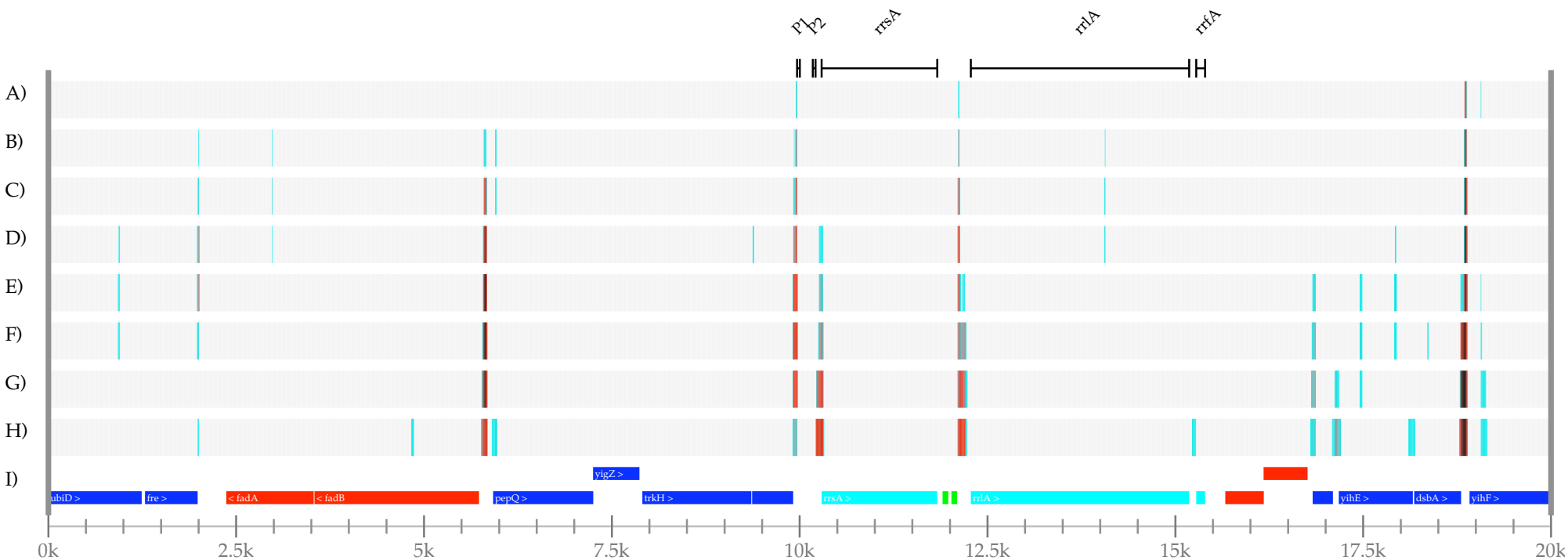
EasyGene -logP
fix avg
0.00 10.00

Resolution: 15

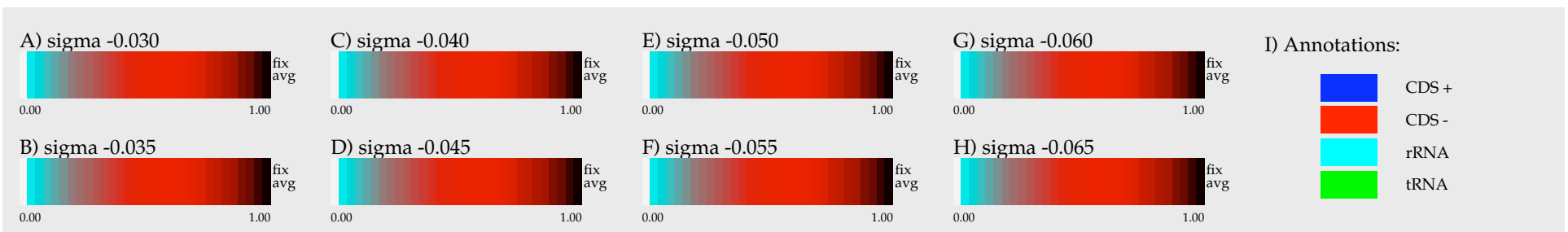
http://www.cbs.dtu.dk/
Center for Biological Sequence Analysis

Ecoli_K12_MG1655 rrsA

20,000 bp



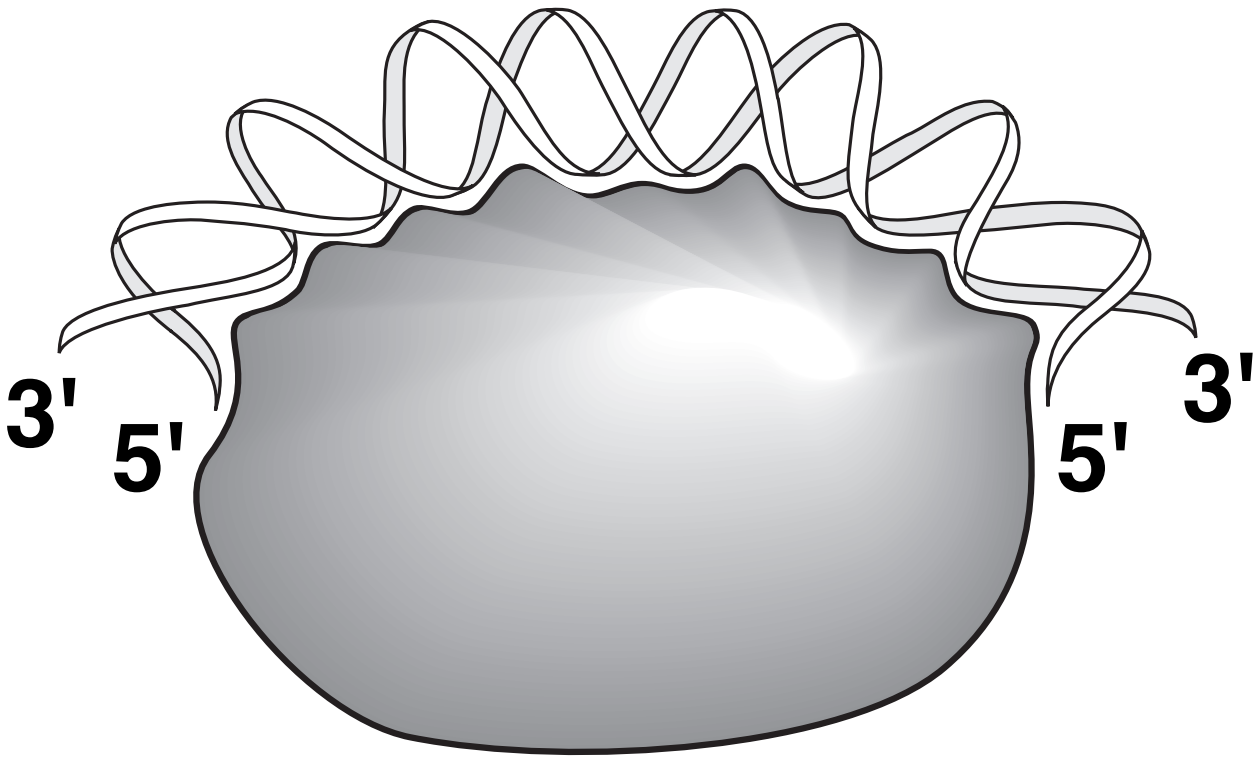
Resolution: 9

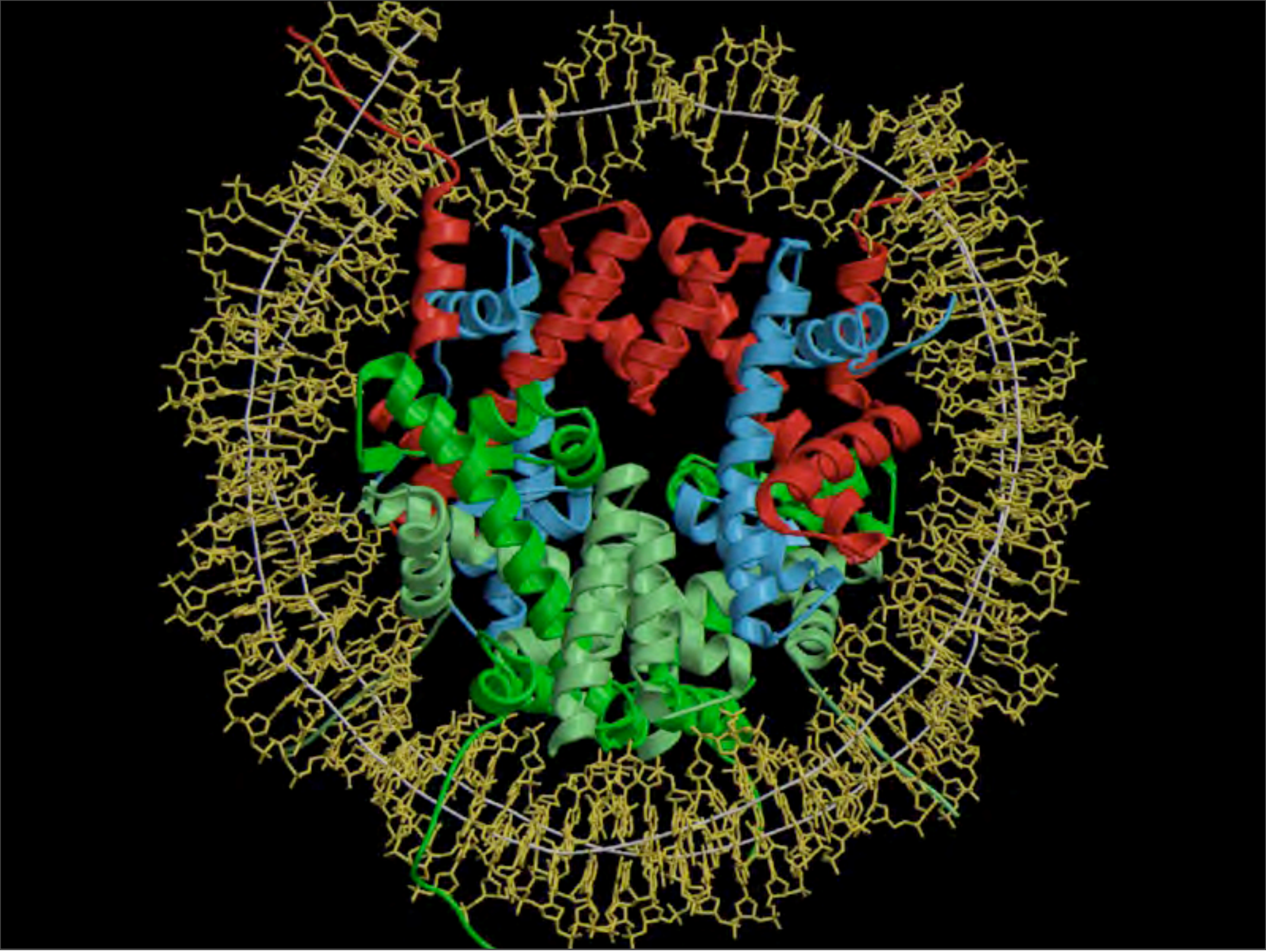


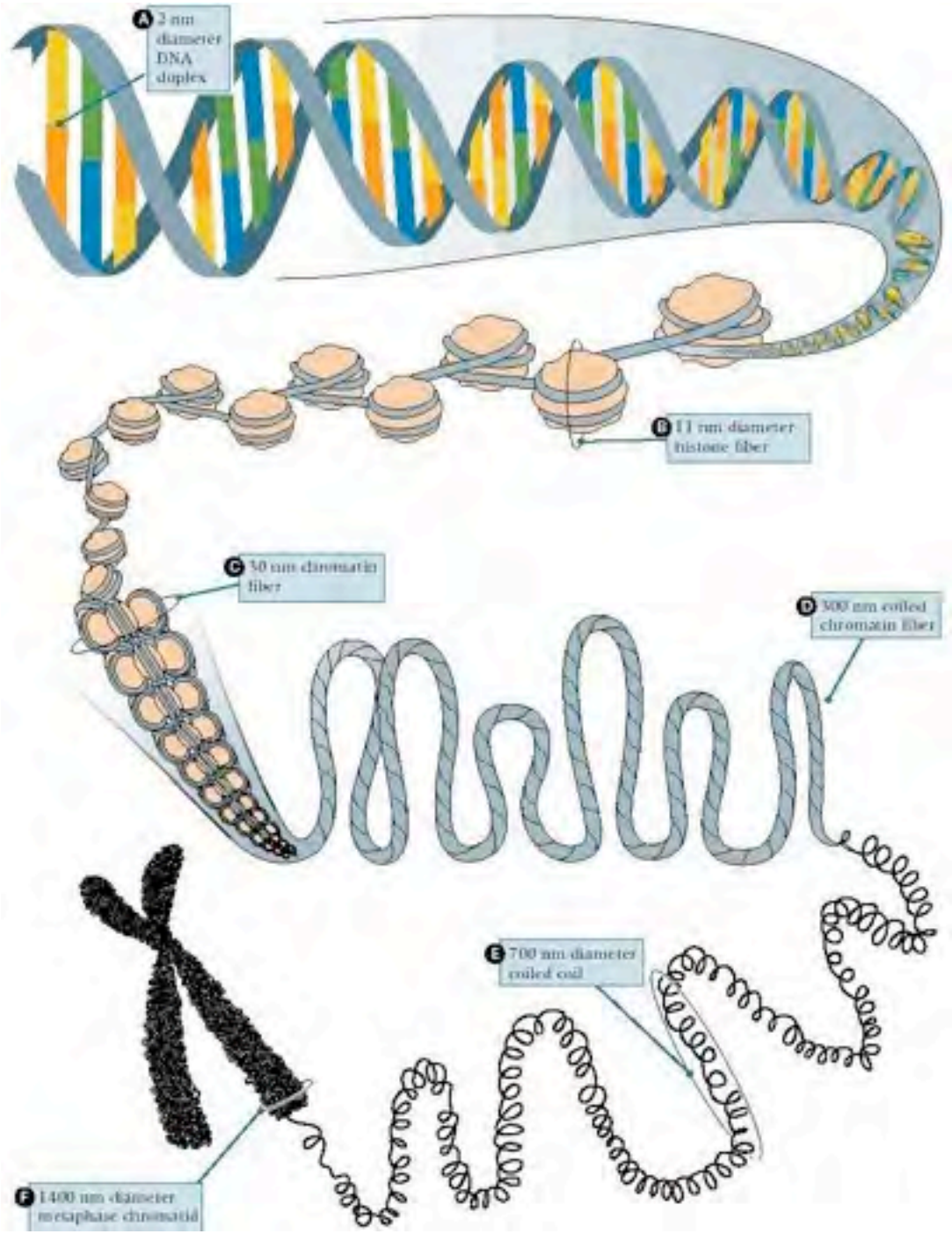
Center for Biological Sequence Analysis
http://www.cbs.dtu.dk/

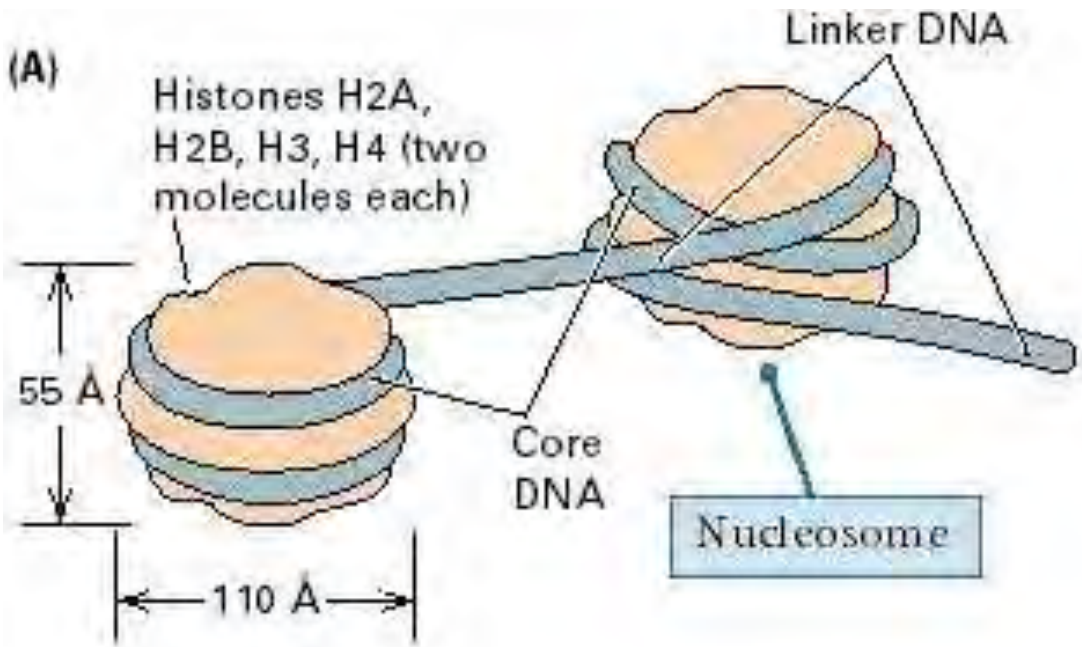
SIDD ATLAS

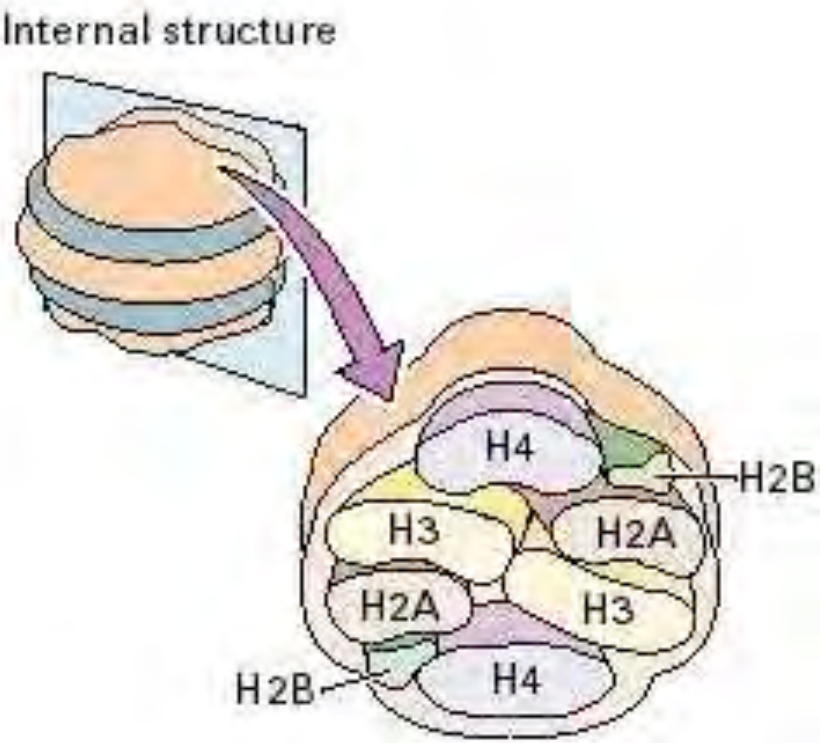
Pause

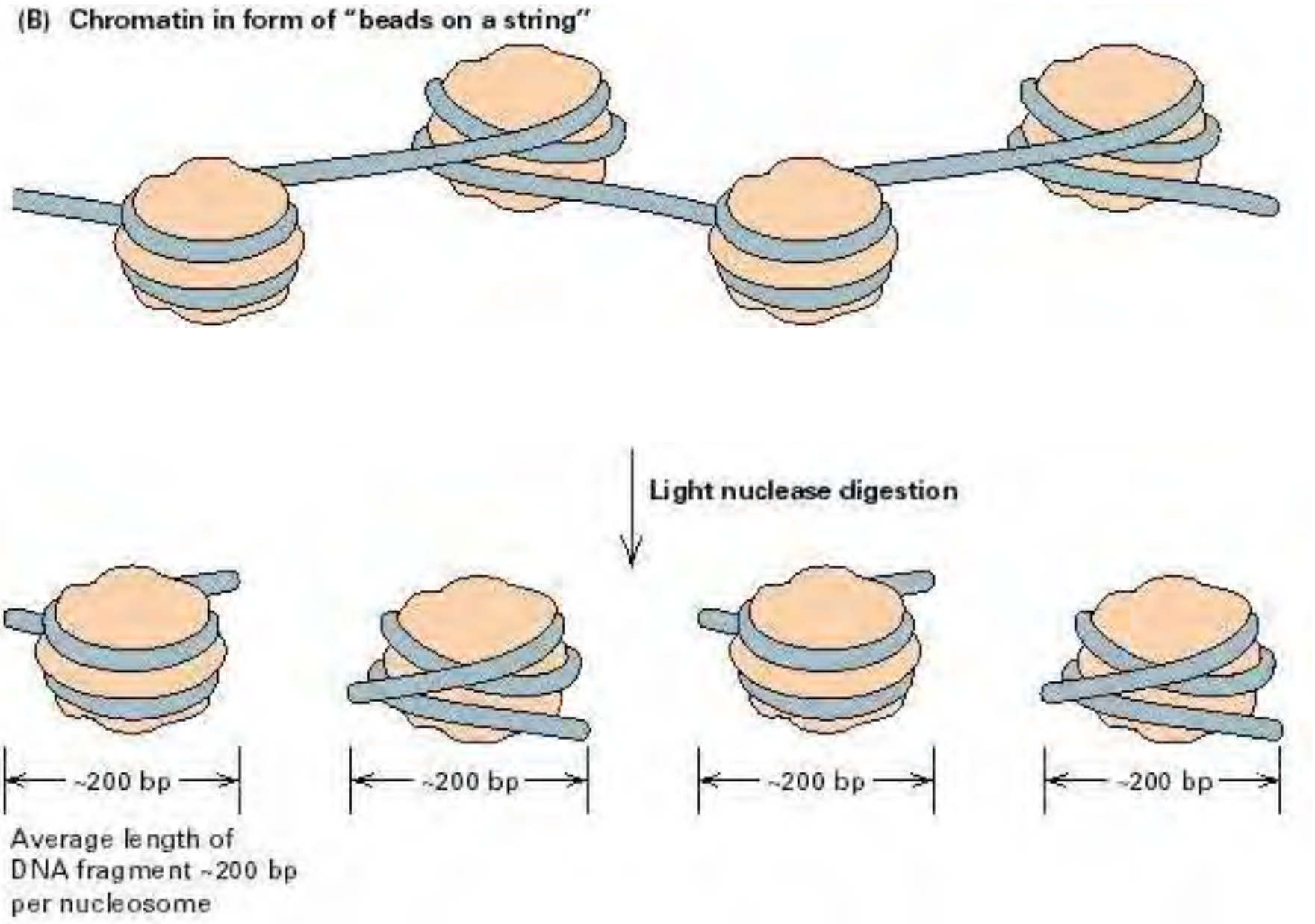

3'
5'
5'
3'

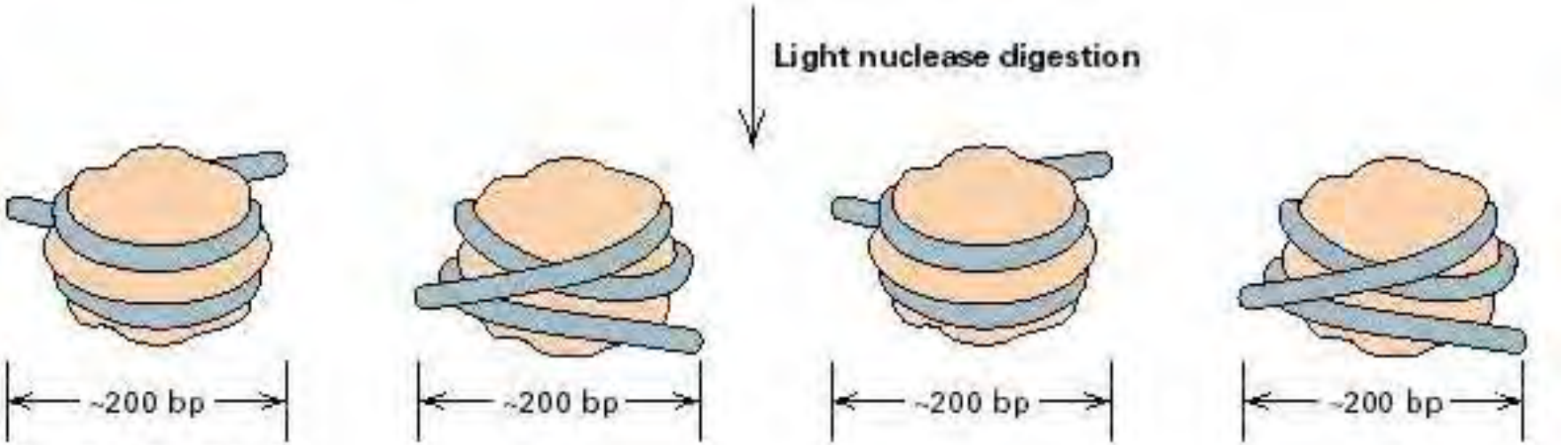



A 2 nm diameter DNA duplex
B 11 nm diameter histone fiber
C 30 nm chromatin fiber
D 300 nm coiled chromatin fiber
E 700 nm diameter coiled coil
F 1400 nm diameter metaphase chromatid

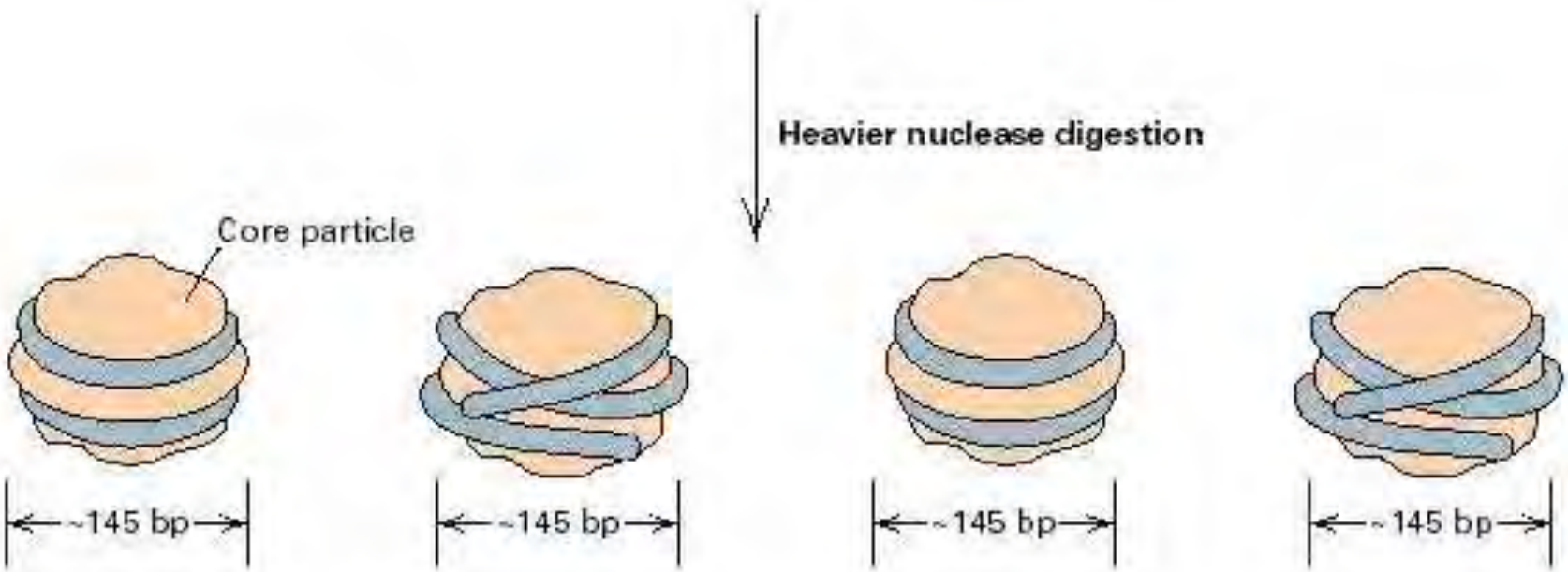

(A)
Histones H2A, H2B, H3, H4 (two molecules each)
Linker DNA
55 Å
110 Å
Core DNA
Nucleosome


Internal structure
H4
H2B
H3
H2A
H2A
H3
H2B
H4


(B) Chromatin in form of "beads on a string"
Light nuclease digestion
~200 bp
~200 bp
~200 bp
~200 bp
Average length of
DNA fragment ~200 bp
per nucleosome



Average length of DNA fragment ~200 bp per nucleosome



Average length of core DNA ~145 bp per core particle

"Travers" trinucleotide scale:

Satchwell,S.C., Drew,H.R., and Travers,A.A., "Sequence periodicities in chicken nucleosome core DNA", J. Mol. Biol., **191**:659-675, (1986).

NOTE: we use a (slight) modification, in which the absolute value (magnitude) of the values is used to reflect trinucleotides which tend to exclude nucleosomes.

Baldi,P., Brunak,S., Chauvin,Y., and Krogh,A., "Naturally occurring nucleosome positioning signals in human exons and introns",

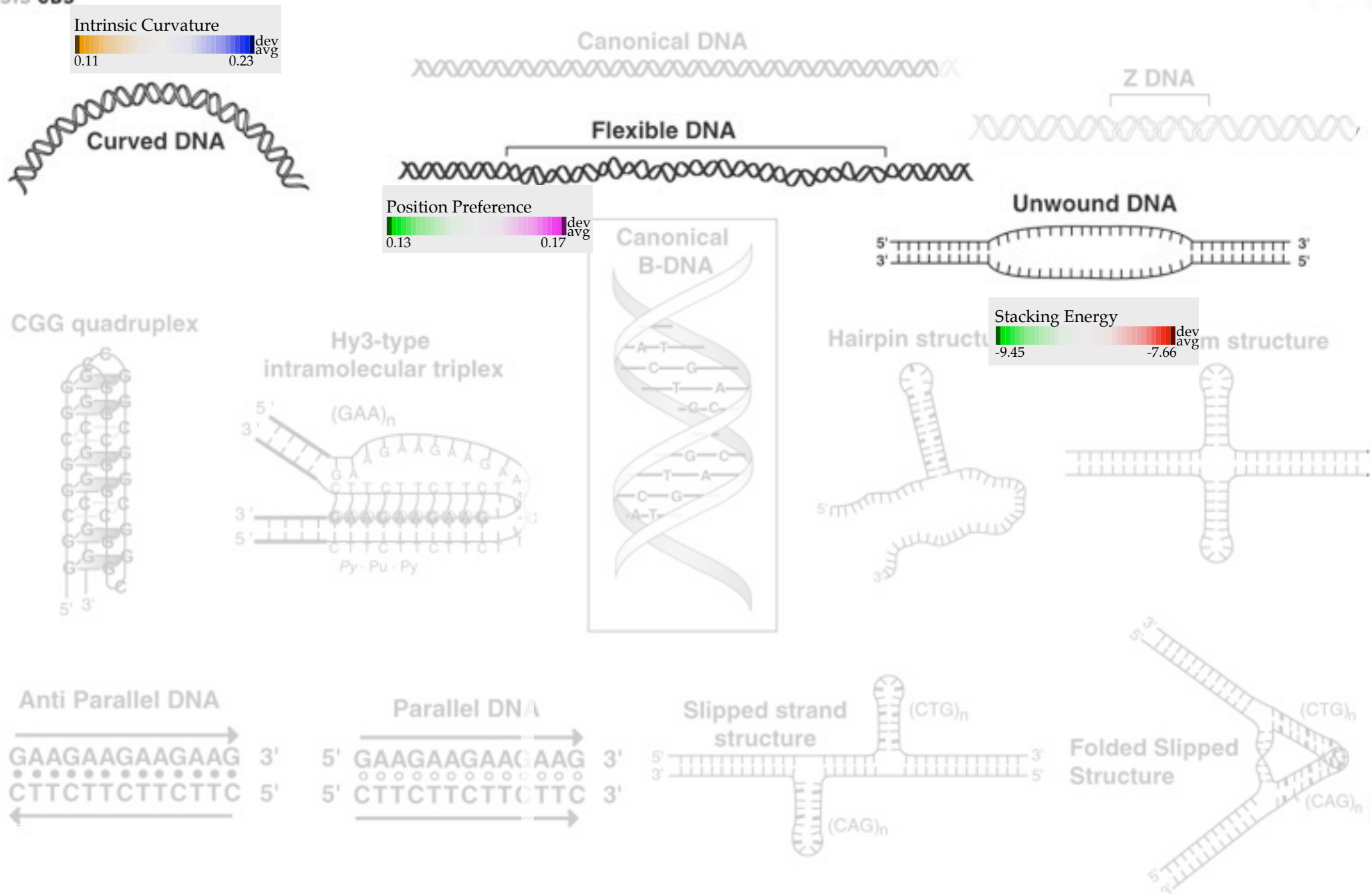
J. Mol. Biol., **263**:503-510, (1996).

trinuc.	% Out
AAT/ATT	-0.280
AAA/TTT	-0.274
CCA/TGG	-0.246
AAC/GTT	-0.205
ACT/AGT	-0.183
CCG/CGG	-0.136
ATC/GAT	-0.110
AAG/CTT	-0.081
CGC/GCG	-0.077
AGG/CCT	-0.057
GAA/TTC	-0.037
ACG/CGT	-0.033
ACC/GGT	-0.032
GAC/GTC	-0.013
CCC/GGG	-0.012
ACA/TGT	-0.006
CGA/TCG	-0.003
GGA/TCC	0.013
CAA/TTG	0.015
AGC/GCT	0.017
GTA/TAC	0.025
AGA/TCT	0.027
CTC/GAG	0.031
CAC/GTG	0.040
TAA/TTA	0.068
GCA/TGC	0.076
CTA/TAG	0.090
GCC/GGC	0.107
ATG/CAT	0.134
CAG/CTG	0.175
ATA/TAT	0.182
TCA/TGA	0.194

High position pref.

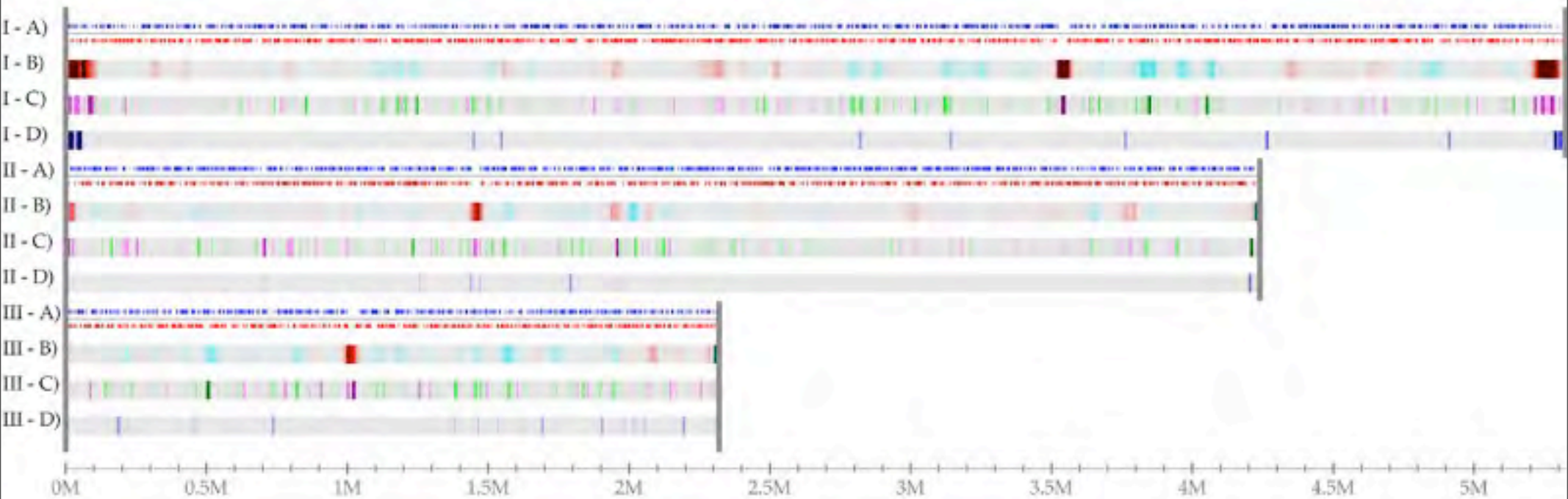
Low position pref.

High position pref.

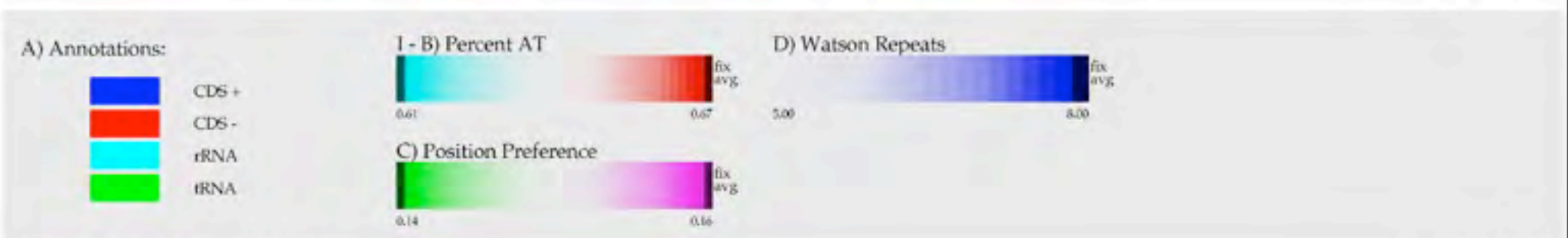

Intrinsic Curvature
dev
avg
0.11
0.23
Curved DNA
Canonical DNA
Z DNA
Flexible DNA
Position Preference
dev
avg
0.13
0.17
Unwound DNA
5'
3'
3'
5'
Canonical
B-DNA
Stacking Energy
dev
avg
-9.45
-7.66
CGG quadruplex
Hy3-type
intramolecular triplex
(GAA)n
Py · Pu · Py
Hairpin structu
m structure
Anti Parallel DNA
5' GAAGAAGAAGAAG 3'
3' CTTCTTCTTCTTC 5'
Parallel DNA
5' GAAGAAGAAGAAG 3'
5' CTTCTTCTTCTTC 3'
Slipped strand
structure
(CTG)n
(CAG)n
Folded Slipped
Structure
(CTG)n
(CAG)n

Schizosaccharomyces pombe

All Three Chromosomes *11,896,623 bp total*



Resolution: 1066



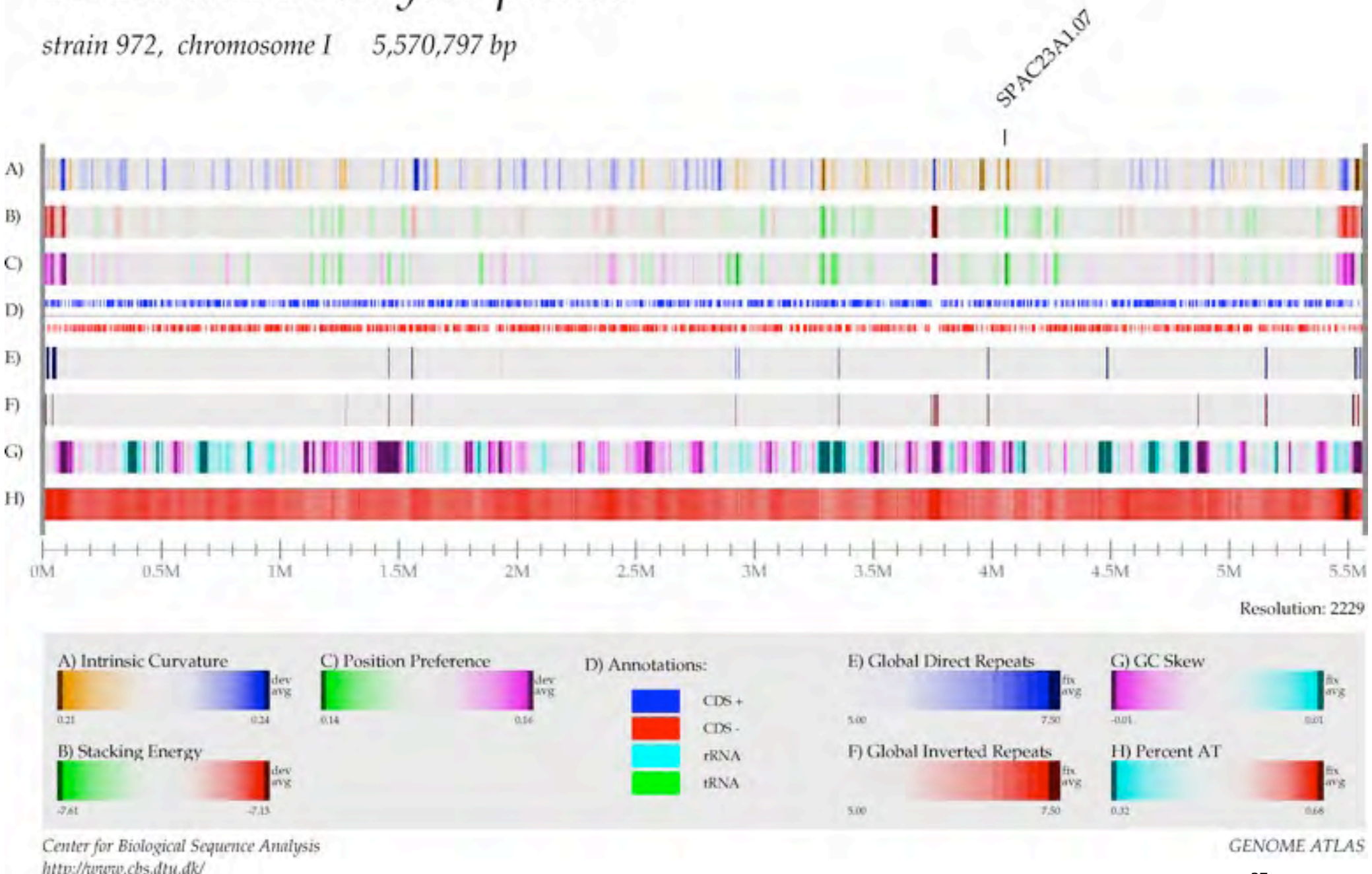
Center for Biological Sequence Analysis
http://www.cbs.dtu.dk/

GENOME ATLAS

Nature, **415**:871-880, (2002).

Schizosaccharomyces pombe

strain 972, chromosome I 5,570,797 bp



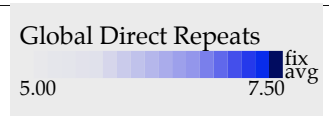
Center for Biological Sequence Analysis
http://www.cbs.dtu.dk/

GENOME ATLAS

DNA Conformation and Transcription, chapter 14, (edited by Takashi Ohyama, 2004).

DNA Conformation and Transcription, chapter 14, (edited by Takashi Ohyama, 2004).

Direct repeats



5'-AGAGAGAGATCTAGACTAGAGAGAGAGAGATACCCTAGAGAGAGAGAGAGAAAG-3'
3'-TCTCTCTCTAGATCTGATCTCCCTCTCTCTATGGGATCTCTCTCTCTCTCTTTC-5'

Local direct simple repeats

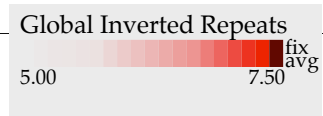
5'-AGGCAAAGTCGCCATGCTGCCATCTGGAGGCAAAGTCGCCATGCTGCCATCTGG-3'
3'-TCCGTTTCAGCGGTACGACGGTAGACGTCCGTTTCAGCGGTACGACGGTAGACC-5'

Local direct repeats

5'-AGGCAAAGTCGCCATGCTGCCATCTGGTGCCCAGGCAAAGTCGCCATGCTGCCATCTGG-3'
3'-TCCGTTTCAGCGGTACGACGGTAGACGACGGGTCCGTTTCAGCGGTACGACGGTAGACC-5'

Local direct repeats with spacer

Inverted repeats



5'-AGGCAAAGTCGCCATGCTGCCATCTGGCCAGATGGCAGCATGGCGACTTTGCCT-3'
3'-TCCGTTTCAGCGGTACGACGGTAGACCGGTCTACCGTCGTACCGCTGAAACGGA-5'

Local inverted repeat

5'-AGGCAAAGTCGCCATGCTGCCATCTGGTGCCCCCAGATGGCAGCATGGCGACTTTGCCT-3'
3'-TCCGTTTCAGCGGTACGACGGTAGACCACGCCGGTCTACCGTCGTACCGCTGAAACGGA-5'

Inverted repeat with spacer

5'-AGGCACGTGAGAATGAATTCCTGCTTTAGATCTAAAAGGCAAAGCTTTGCCT-3'
3'-TCCGTGCACTCTTACTTAAGGACGAAATCTAGATTTTCCGTTTCGAAACGGT-5'

Palindromic repeats

Mirror and Everted repeats

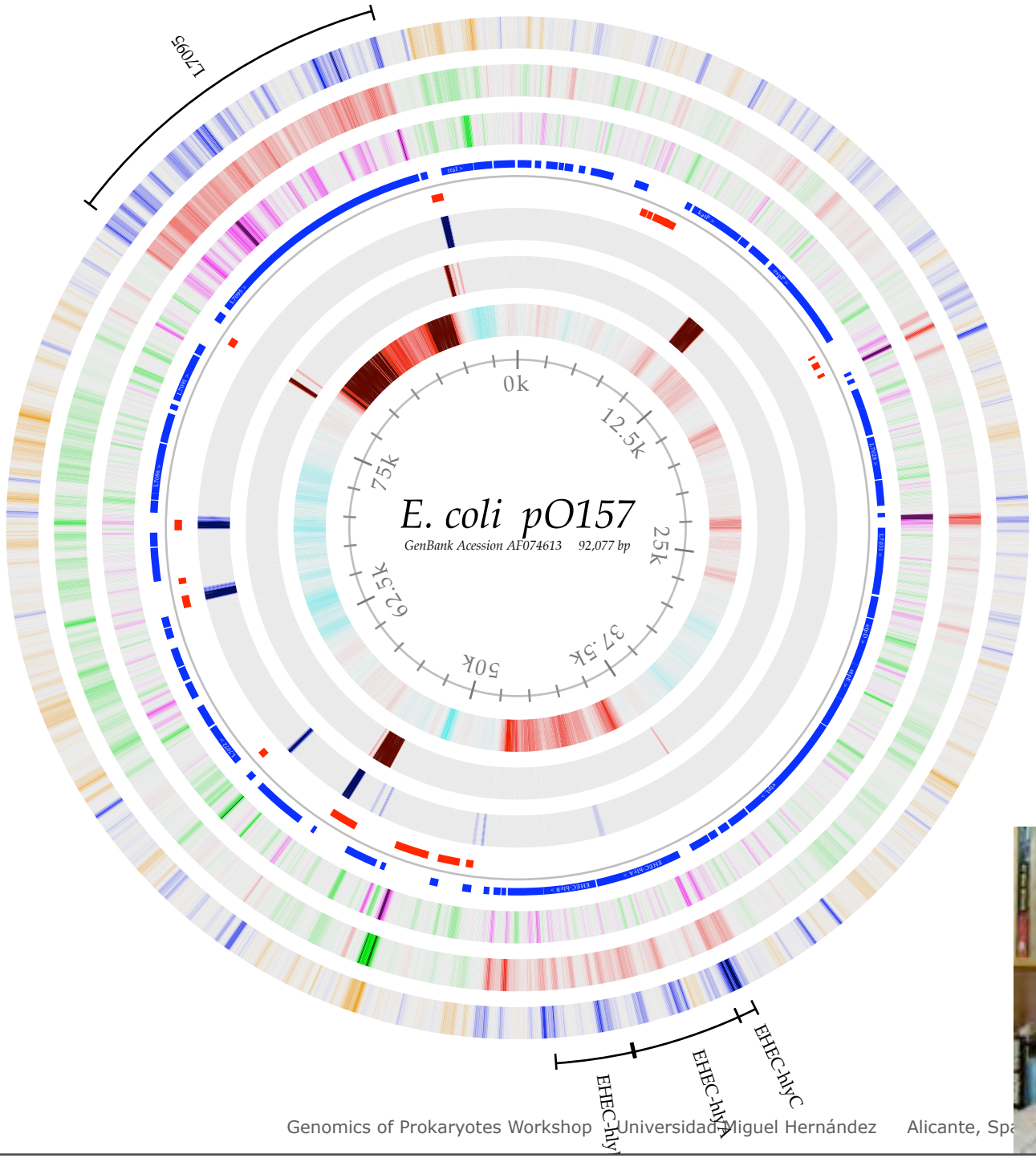
5'-AGGCAAAGTCGCCATGCTGCCATCTGGGGTCTACCGTCGTACCGCTGAAACGGA-3'
3'-TCCGTTTCAGCGGTACGACGGTAGACCCCAGATGGCAGCATGGCGACTTTGCCT-5'

Mirror repeat

5'-AGGCAAAGTCGCCATGCTGCCATCTGGTCCGTTTCAGCGGTACGACGGTAGACC-3'
3'-TCCGTTTCAGCGGTACGACGGTAGACCAGGCAAAGTCGCCATGCTGCCATCTGG-5'

Everted repeat

Introducing the "Genome Atlas"



Intrinsic Curvature

0.08 0.30

Stacking Energy

-9.52 -6.41

Position Preference

0.11 0.17

Annotations:

CDS +

CDS -

Watson Repeats

5.00 7.50

Crick Repeats

5.00 7.50

Percent AT

0.30 0.70

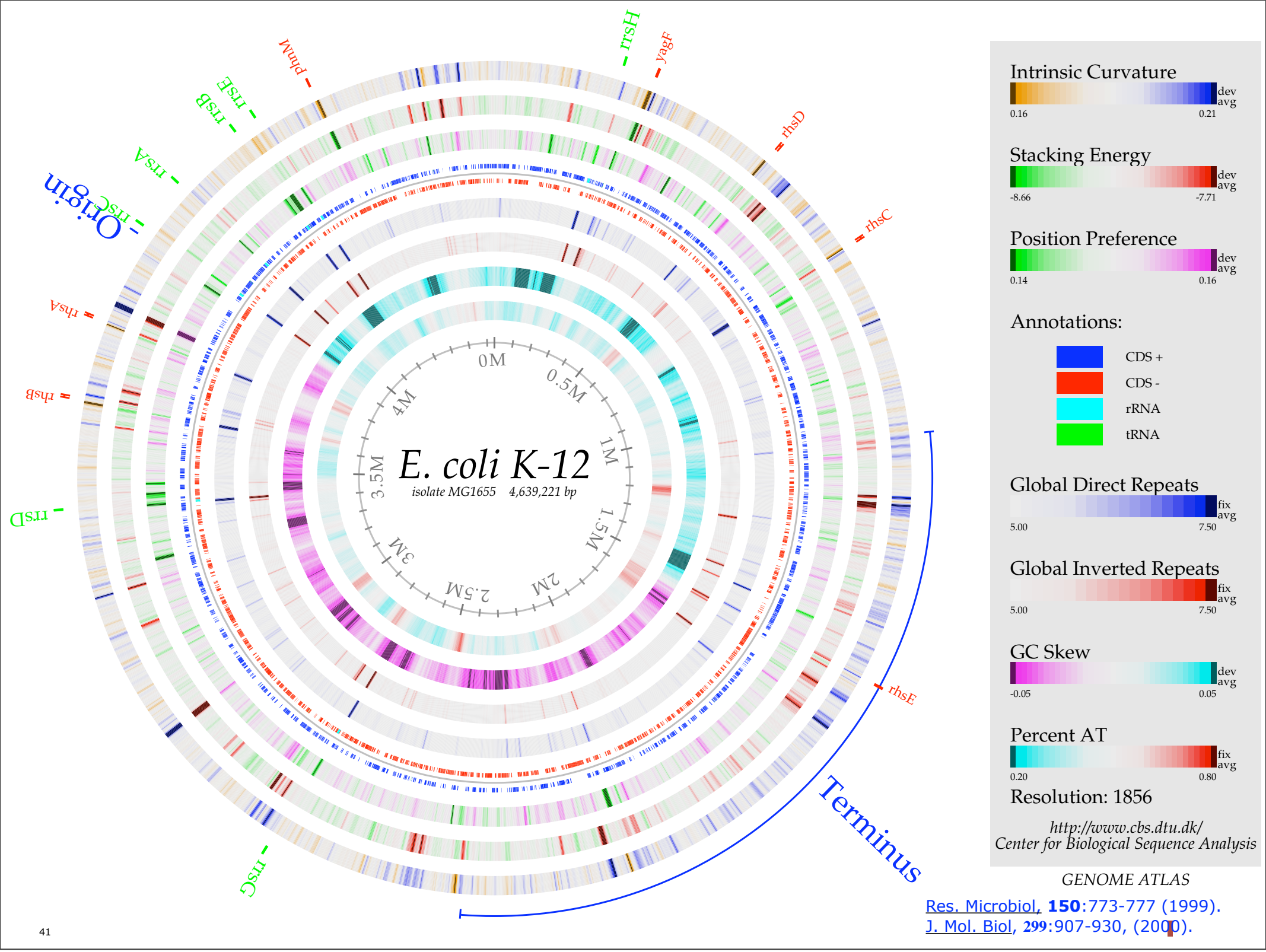
Resolution: 37

Friis, Jensen, and Ussery

Genetica, **108**:47-51, (2000).

http://www.cbs.dtu.dk/
Center for Biological Sequence Ar

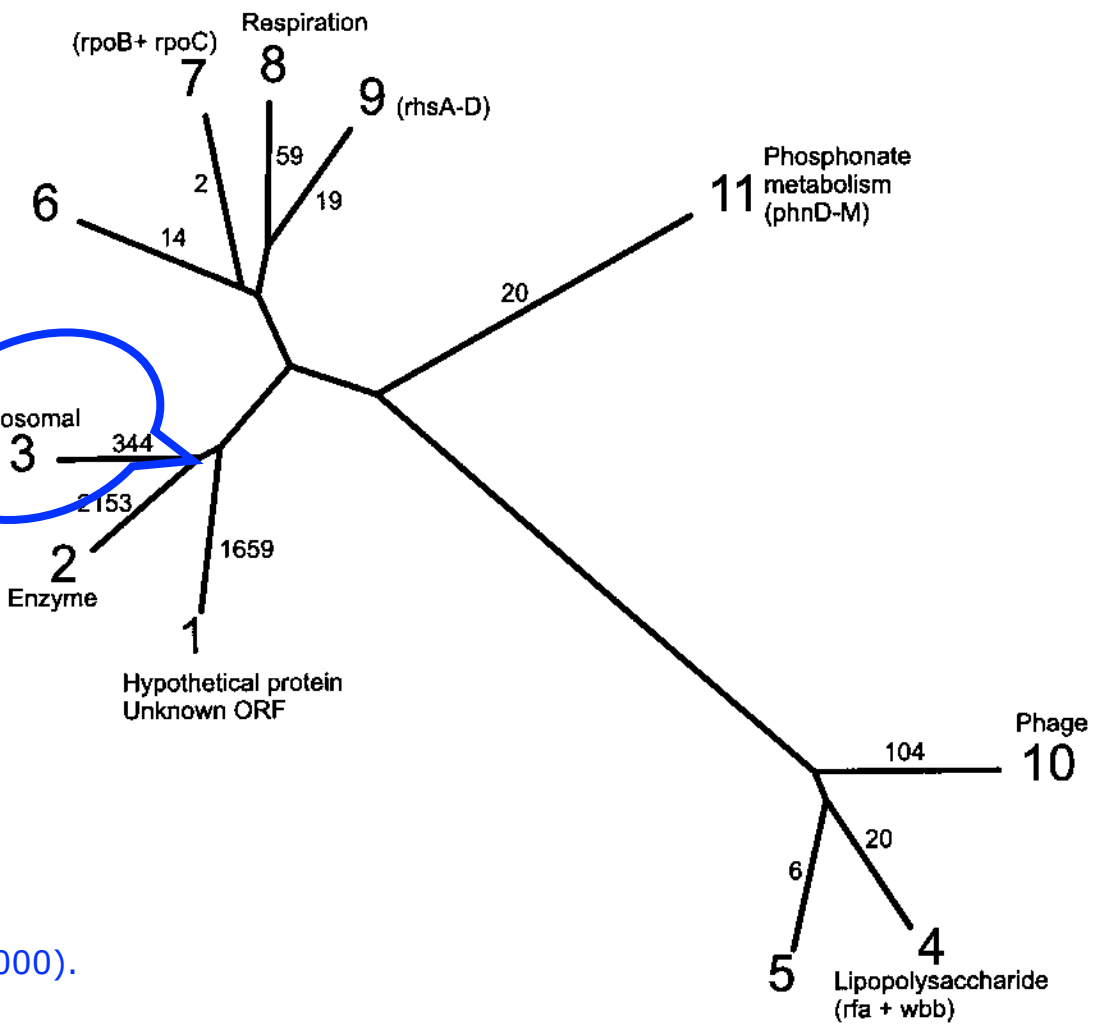
GENOME ATLAS



GENOME ATLAS

Res. Microbiol, **150**:773-777 (1999).
J. Mol. Biol, **299**:907-930, (2000).

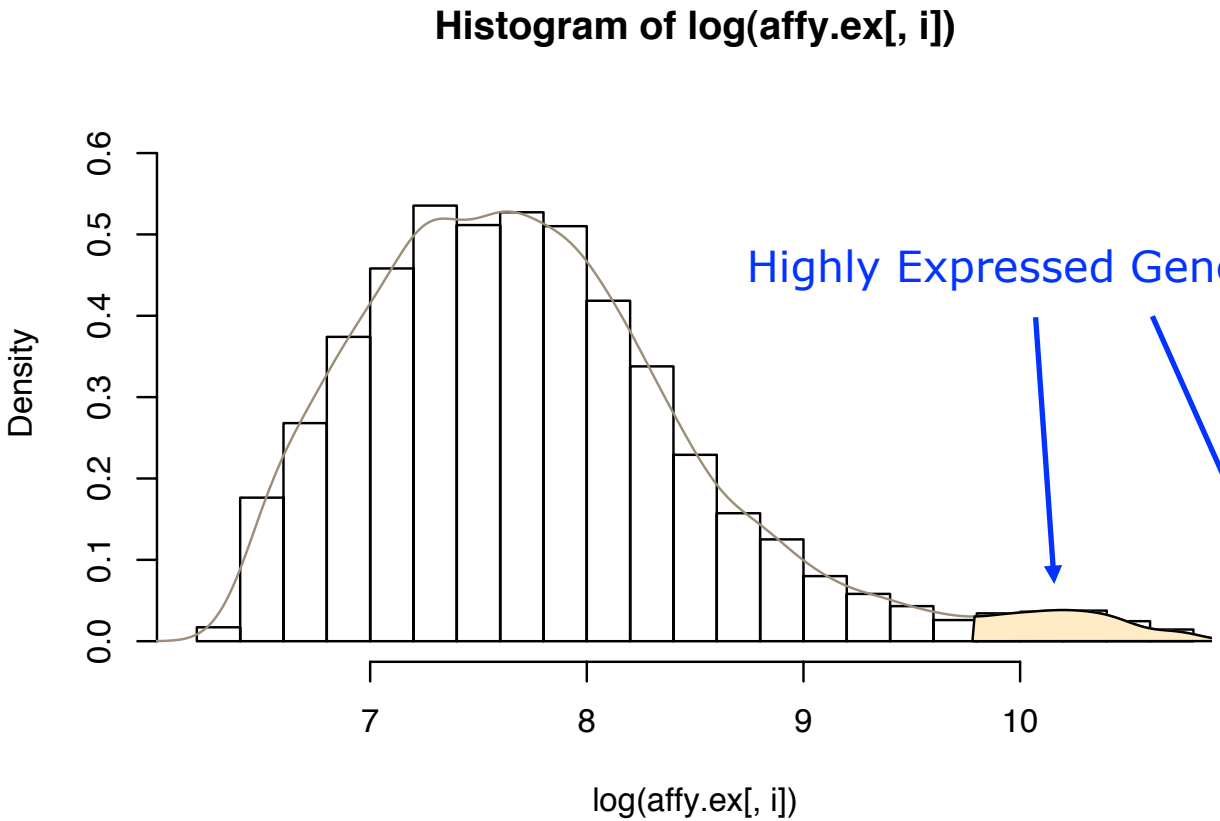
DNA Structural Atlas for Escherichia coli



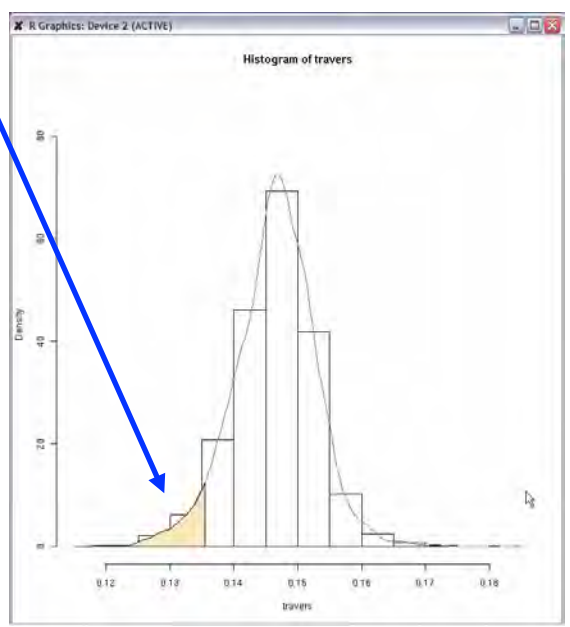
low position preference

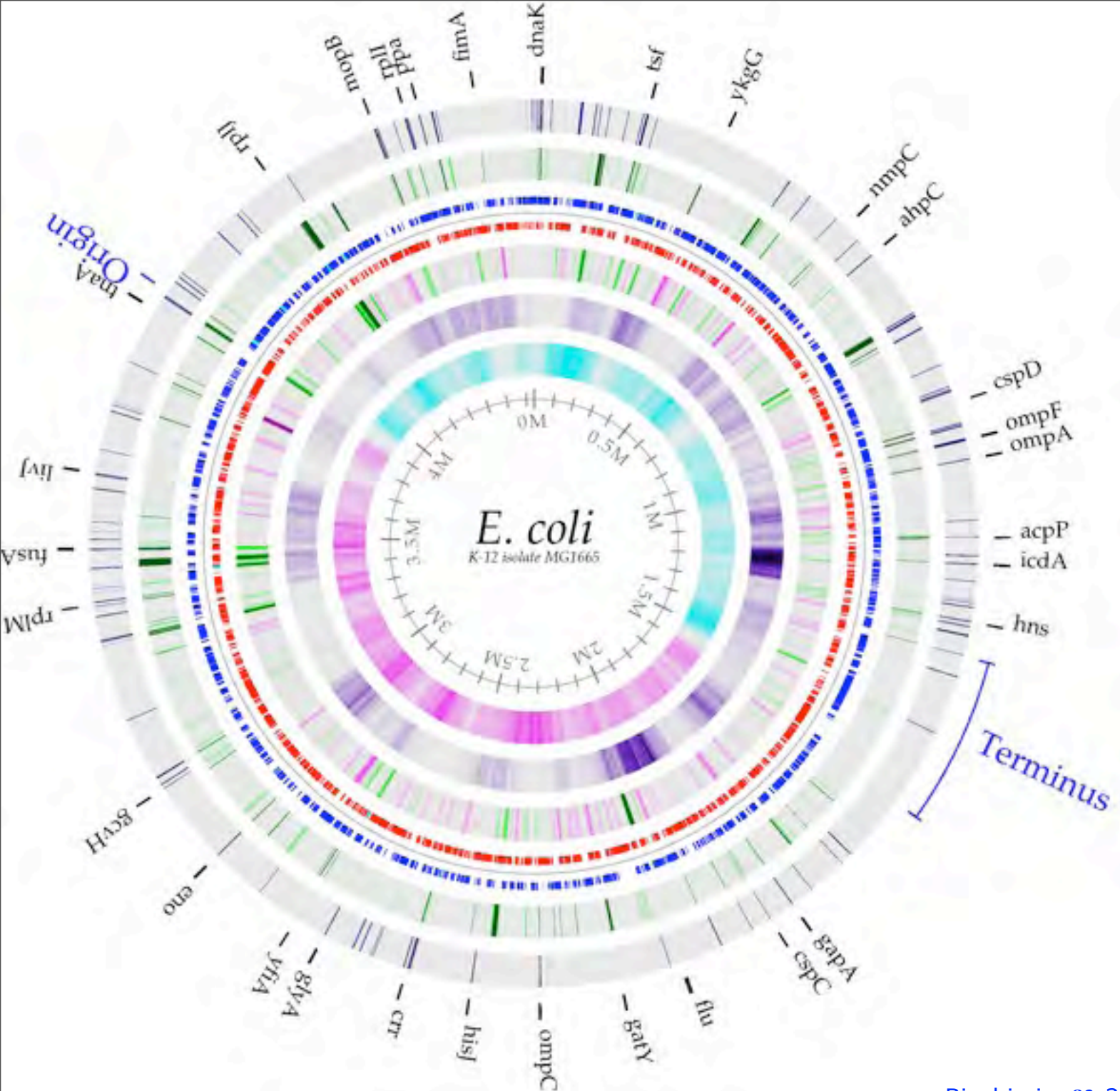
J. Mol. Biol, **299**:907-930, (2000).

Figure 7. Structural cluster analysis. Distance tree showing the relative location of 11 gene clusters based on average structural measures. The number of genes in



Highly Expressed Genes





Biochimie, 83:201-212, (2001).

Monday, 12 December, 2011