Protein and DNA tools

at I*CGEB Trieste*

Sándor Pongor

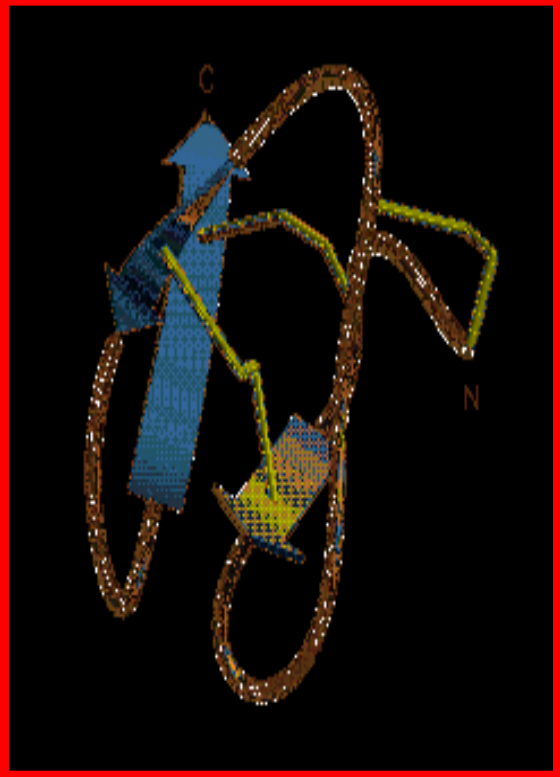
Protein Structure and Bioinformatics, ICGEB, Trieste

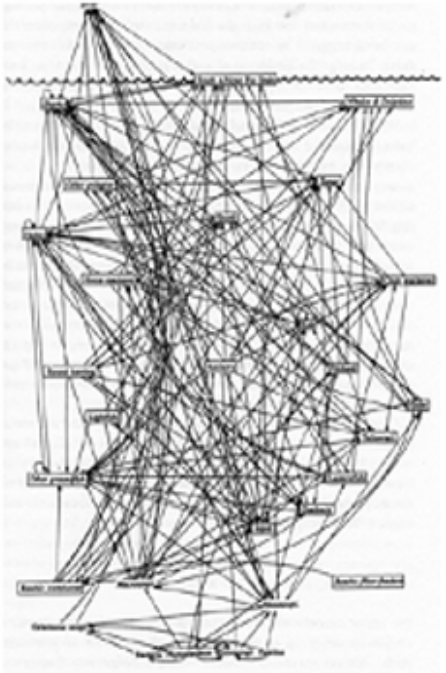
Lecture overview

- Theory: Protein domains and the “similarity space”
- SBASE dbase/search tool: domain sequence identification
- PRIDE : identification of folds in 3D structures
- CX, DPX: visualizing protein surface and interior
- Theory: DNA structure, curvature/bendability plots
- DNA sequence plots
- Excercises

Models

tassfvvswvsasdtvsgfrvey
elseegdepqyldlpstatsvni
pdllpgrkytvnvyeiseegeqn
lilstsqttapdappdptvdqvd
dtsivvrwsrprapitgyrivys
psvegsstelnlpetansvtlsd
lqpgvqynitiyaveenqestpv
fiqqettgvprsdkvppprdlqf
vevtdvkitimwtppespvtgyr
vdvipvnlpgehgqrlpvsrntf
aevtglspgvtyhfkvfavnqgr
eskpltaqqatkldaptnlqfin
etdttvivtwtpprarivgyrlt
vgltrggqpkqynvgpaasqypl
rnlqpgseyavslvavkgnqqsp
rvtgvfttlqplgsiphyntevt
ettivitwtpaprigfklgvrps
qggeaprevtsesgsivvsgltp
gveyvytisvlrdgqerdapivk

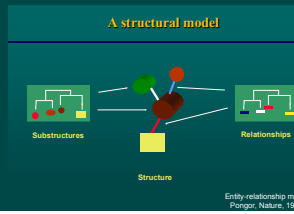




SEQUENCES

3D STRUCTURES

NETWORKS



Protein domain

3D

- Domains are the structural and functional building blocks of proteins (Pfam definition)
- Each organized into a characteristic three dimensional (3D) structure or fold type

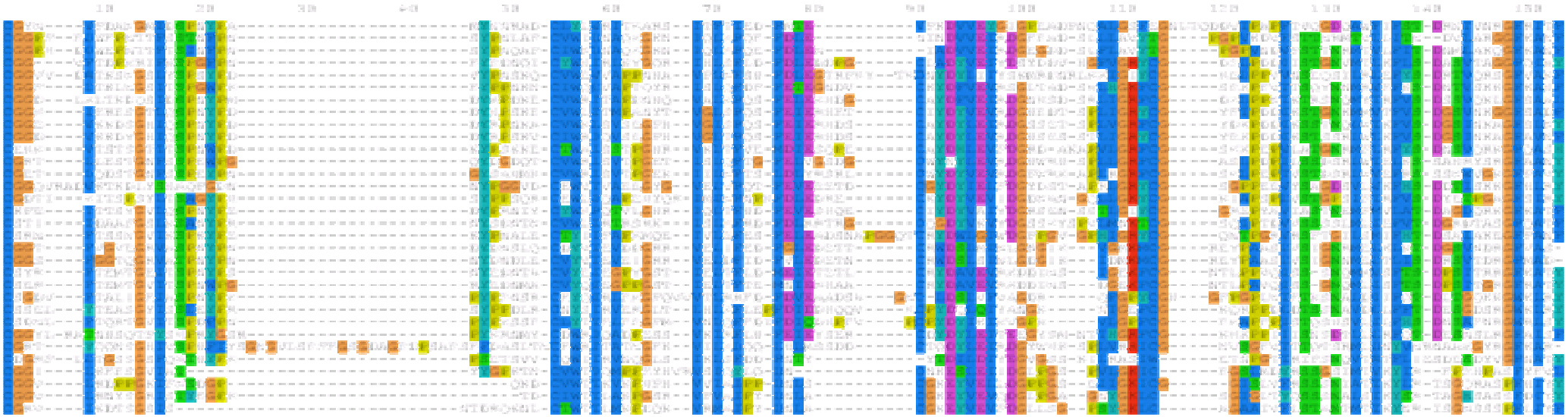
Major seminal plasma glycoprotein PSP-I/II

CUB domain

Acidic seminal fluid protein

Protein domain

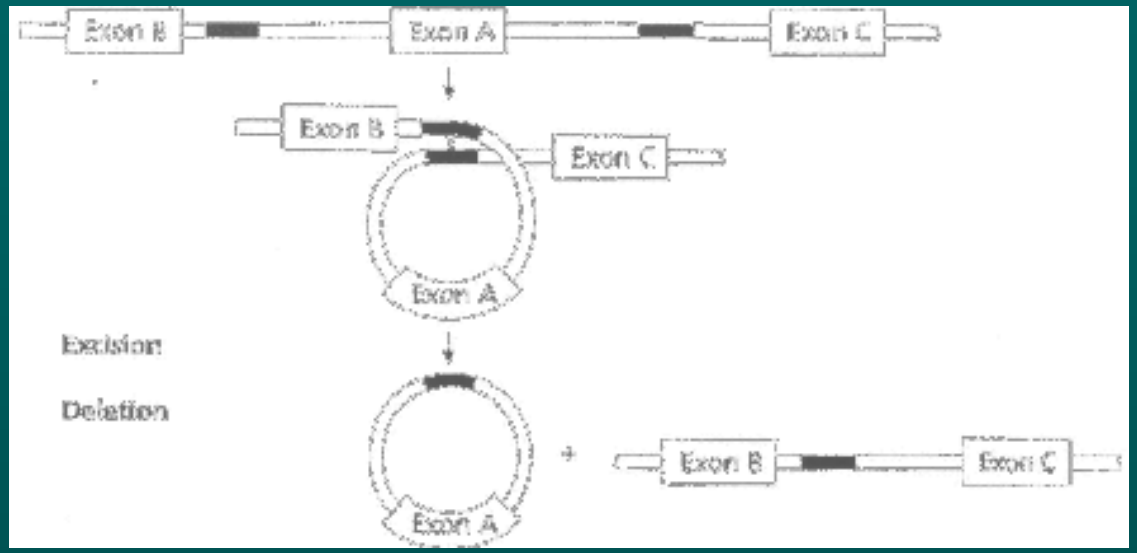
- More commonly domains correspond to a region of **sequence homology** identified in otherwise apparently unrelated proteins (PROSITE definition)



Pfam seed alignment for the CUB domain

Protein domains

Network, systems biology



Autonomous units of evolution, domain shuffling

PRESENT IN MORE THEN ONE FAMILY…

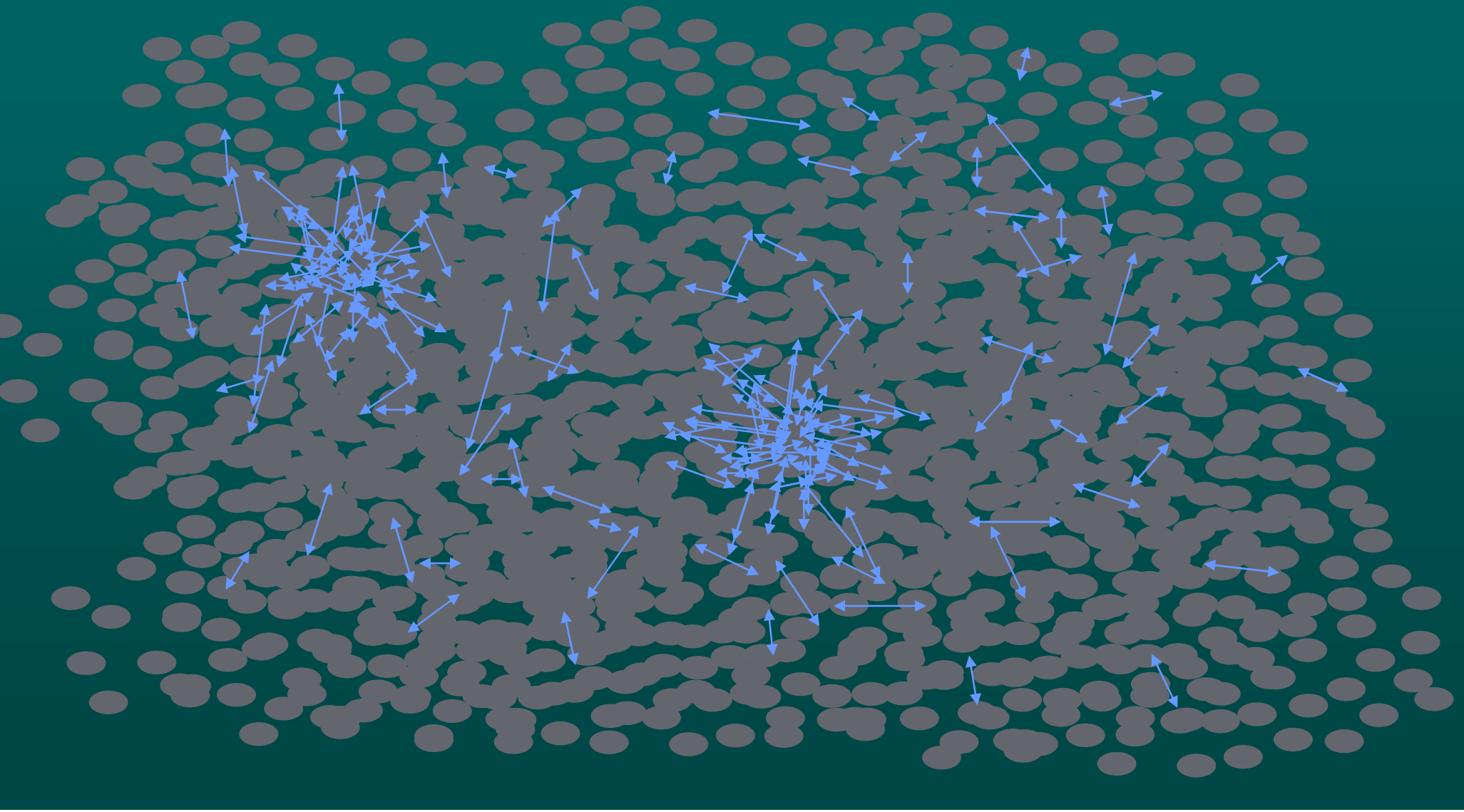
Frequent domains are evolutionary success stories

SBASE

Identification of protein domains in sequences via similarity analysis

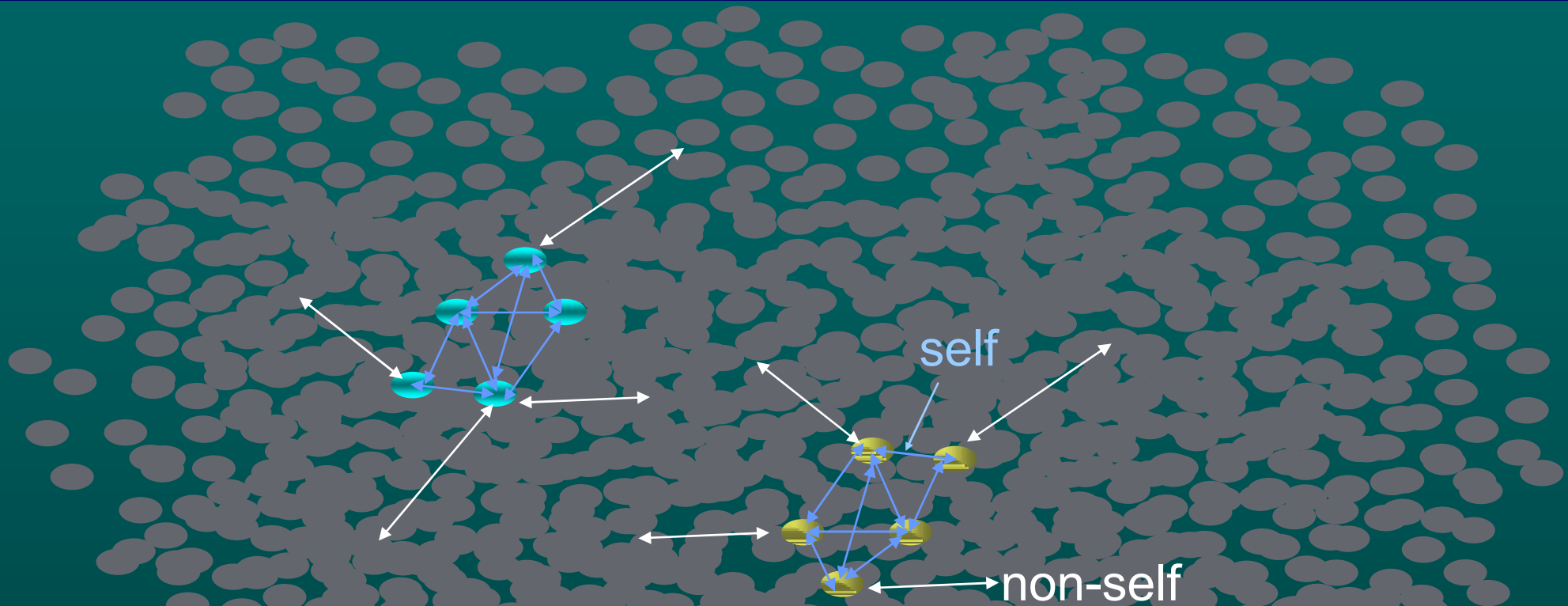
The database as network of similarities: A memory network model

1-D



Das Wohltemperierte Database: Similarity network as knowledge representation

1-D

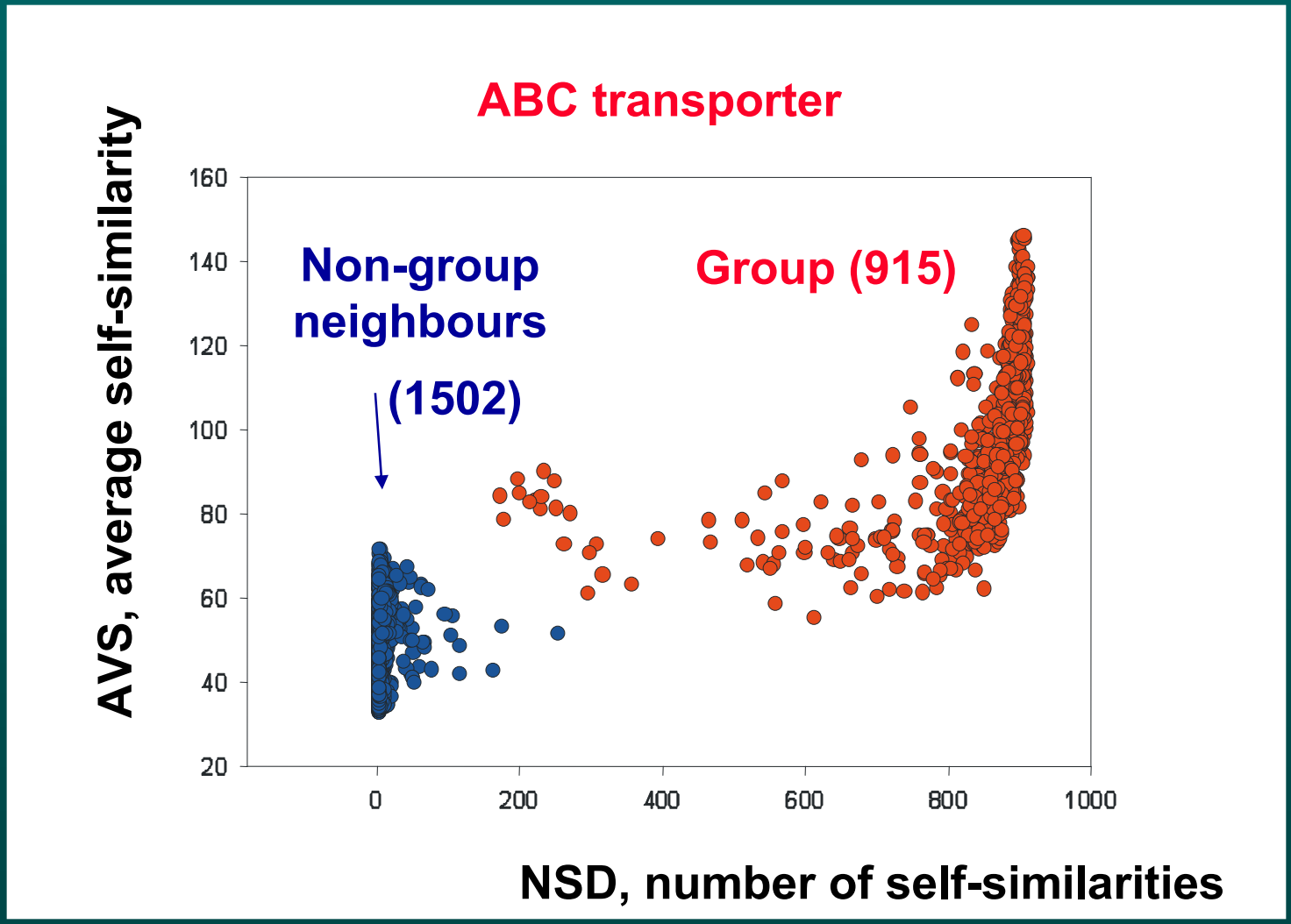


Knowledge-base of meaningful similarities
Signal and noise distinguished

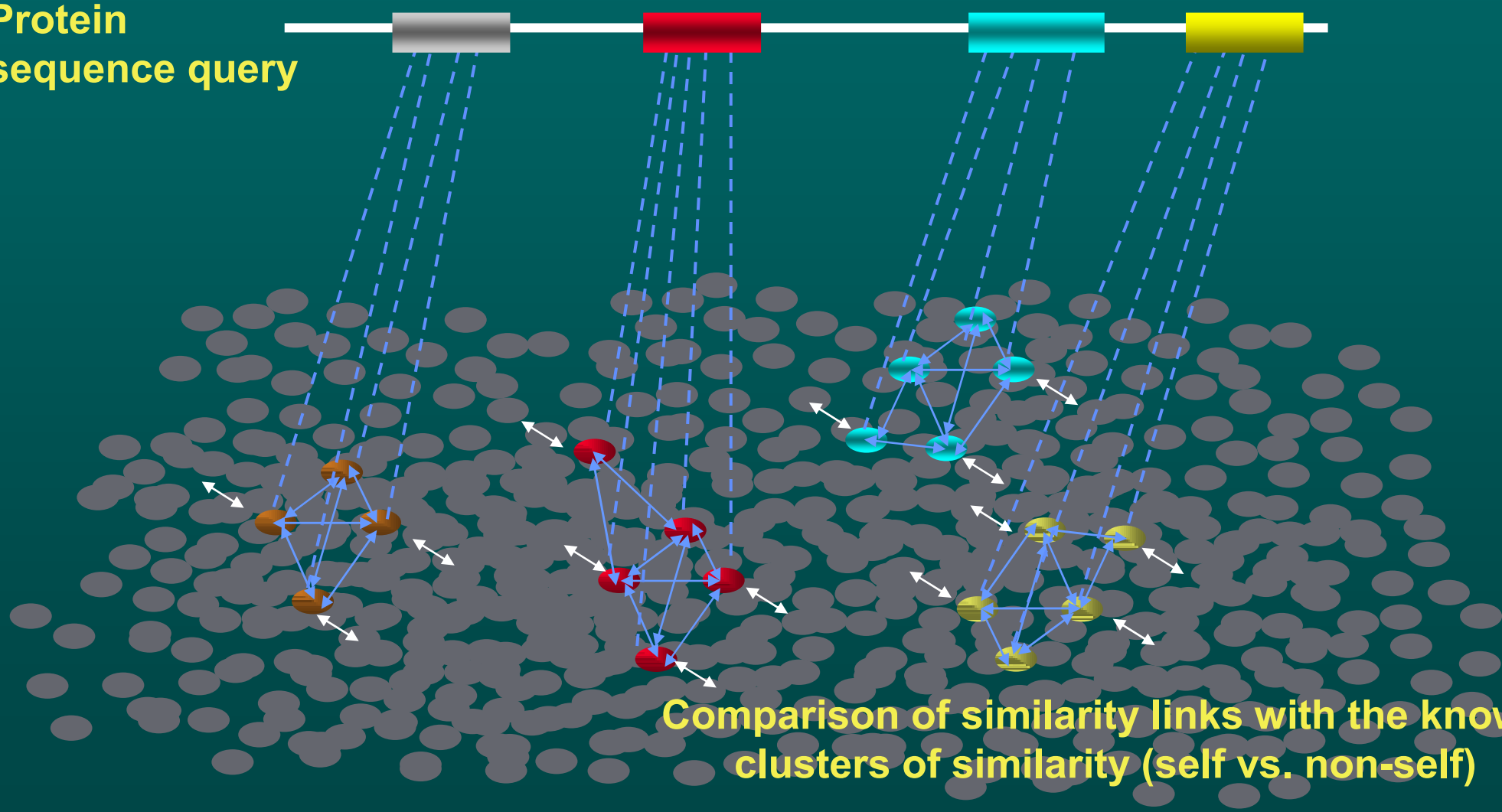
Representation of sequence group

(exemplars and a network of similarities)

1-D



Prediction based on the similarity network approach



Similarities recognized by machine learning techniques

1-D

- Machine learning: techniques for automated classification based on examples
- Best known: Neural networks, Support Vector Machines
- Overlapping positive/negative example regions problem persists, and must be resolved in the future!

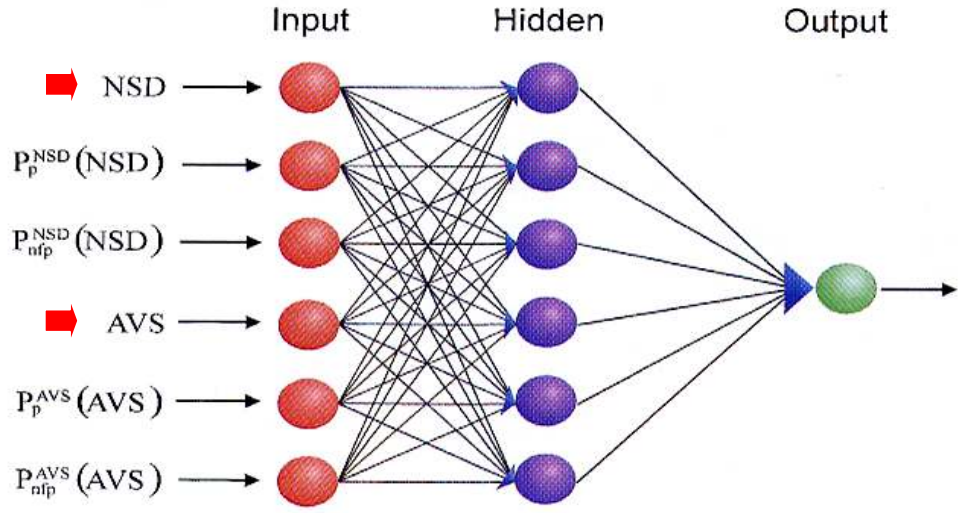
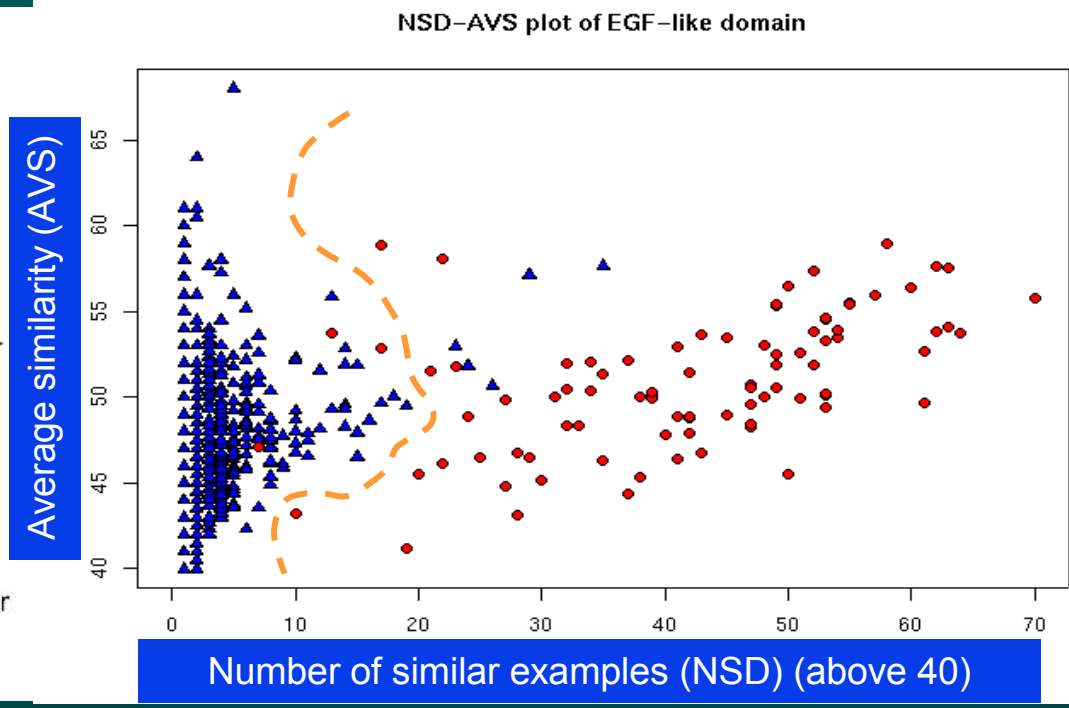
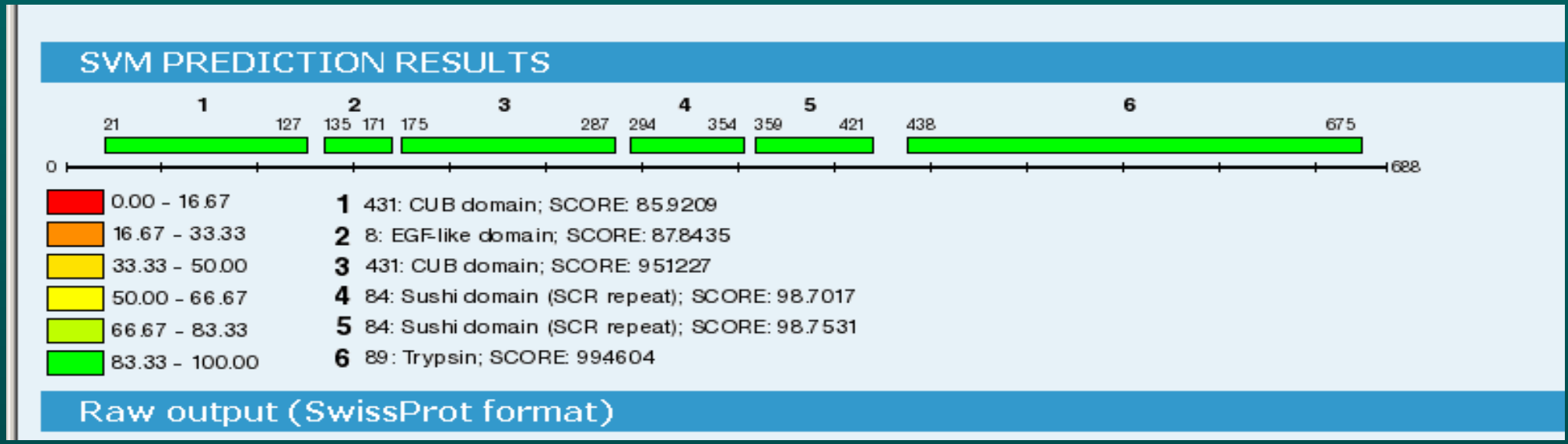


Figure 3 The backpropagation neural network architecture used for domain recognition.



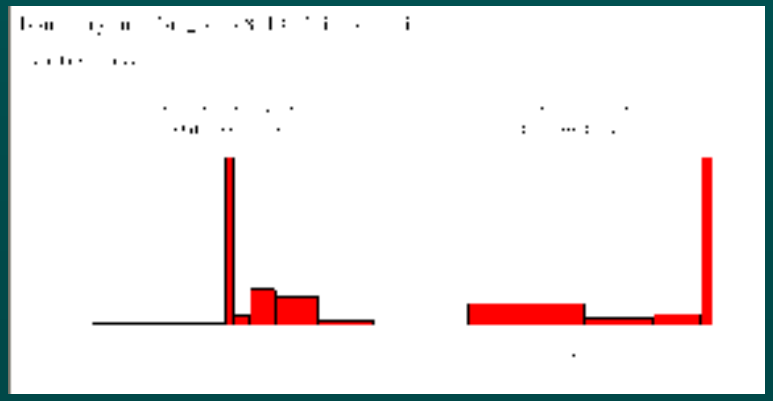
Domain prediction @SBASE



Domain prediction @SBASE

Raw output (SwissProt format)

```
ID   SBASE_PRED     STANDARD;      PRT;   688 AA.
AC   unknown;
DE   DOMAIN ARCHITECTURE PREDICTED BY SBASE SVM
KW
FT   DOMAIN       21    127       431: CUB domain; SCORE: 85.9209..
FT   DOMAIN      135    171       8: EGF-like domain; SCORE: 87.8435..
FT   DOMAIN      175    287       431: CUB domain; SCORE: 95.1227..
FT   DOMAIN      294    354       84: Sushi domain (SCR repeat); SCORE:
FT                                98.7017..
FT   DOMAIN      359    421       84: Sushi domain (SCR repeat); SCORE:
FT                                98.7531..
FT   DOMAIN      438    675       89: Trypsin; SCORE: 99.4604..
SQ   SEQUENCE   688 AA;  76685 MW;  85522647A4C47205 CRC64;
     MWCIVLFSLL AWVYAEPTMY GEILSPNYPQ AYPSEVEKSW DIEVPEGYGI HLYFTHLDIE
     LSENCAYDSV QIISGDTEEG RLCGQRSSNN PHSPIVEEFQ VPYNKLQVIF KSDFSNEERF
     TGFAAYYVAT DINECTDFVD VPCSHFCNNF IGGYFCSCPP EYFLHDDMKN CGVNCSGDVF
```

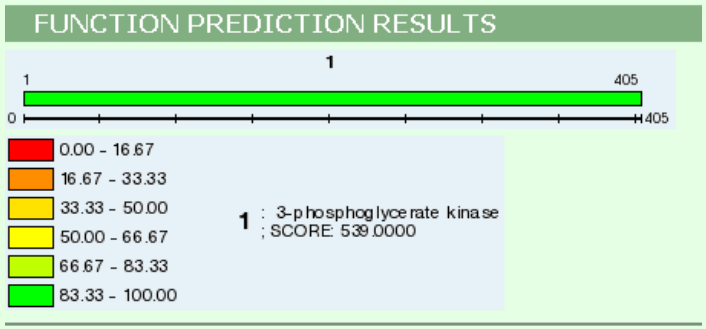


Boundary statistics

			1-5	6-10	11-20	21-50	50-
Total no. of predictions	278	100.00%					
With exact boundaries	156	56.12%					
With different boundaries	122	43.88%	20.50%	20.14%	3.24%	0.00%	0.00%
Differences in detail:							
Shorter	113	40.65%	18.35%	15.11%	6.83%	0.36%	0.00%
Longer	9	3.24%	1.80%	0.72%	0.72%	0.00%	0.00%
Left shifted*	0	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Right shifted*	0	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Left displaced	2	0.72%					
Right displaced	0	0.00%					
No fragments.							

*Equal start-end displacement

Function prediction @SBASE



SBASE
FUNCTION PREDICTION SYSTEM

Information Server status

Current release summary
Underlying theory
Related publications

Browse

Browse database by groups
Browse genomes

Analyze

new Quick predict!
Paste your sequence below (RAW)

Predict!

Predict functions of your sequence

Search

Quick search

Type either an accession, group id or keyword
(e.g. SBA12534, A367 or spectrin)

Advanced Search
BLAST search

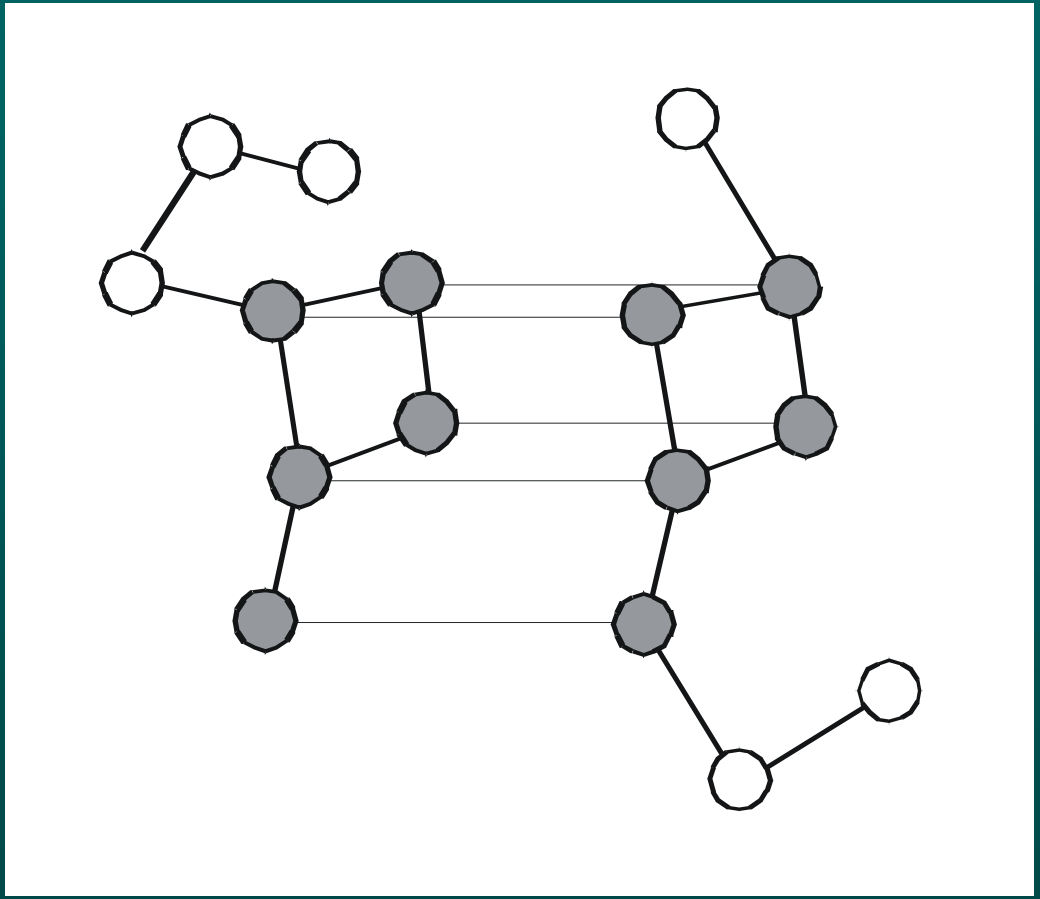
Retrieve

SBASE by FTP

PRIDE

Identification of folds (domain type) by 3-D similarity searching

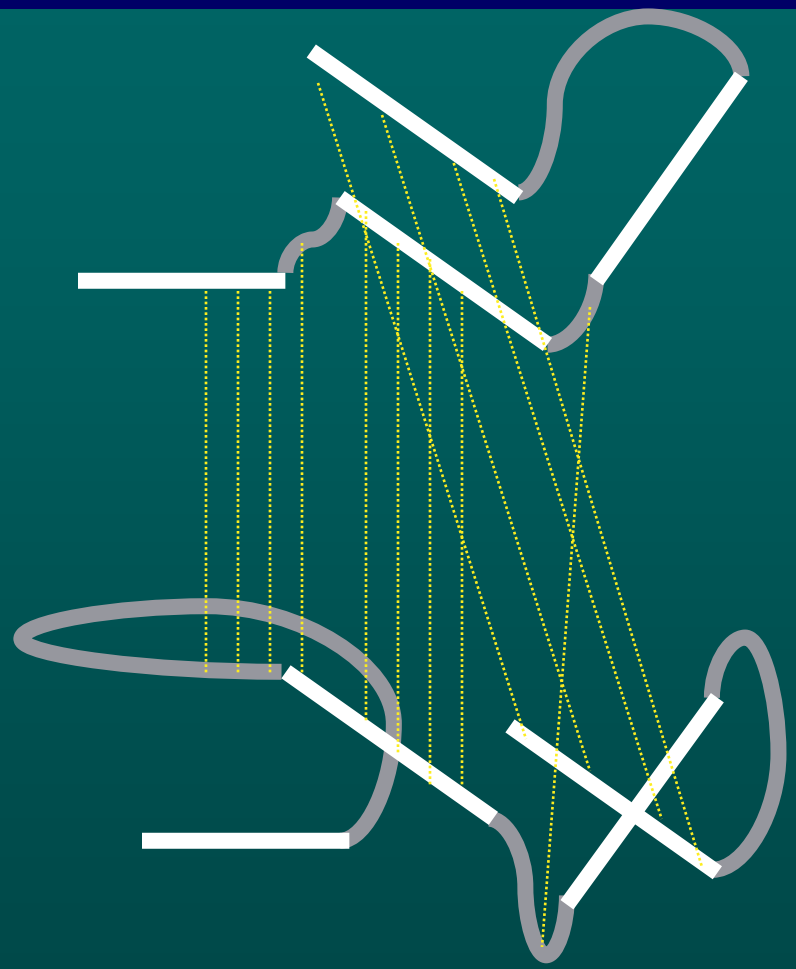
Substructure identity ~ similarity



”The similarity of objects can be best described as partial identities of components and relationships

Erich Goldmeier, The similarity of perceived forms, 1936

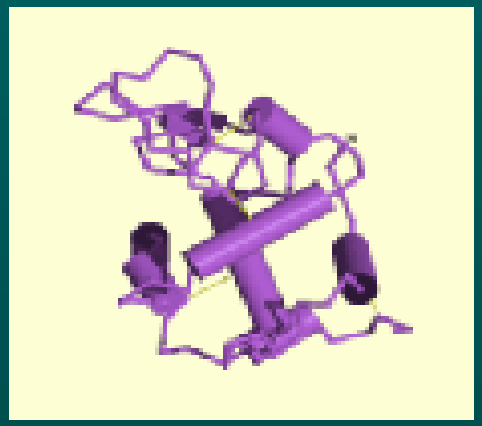
3-D comparison

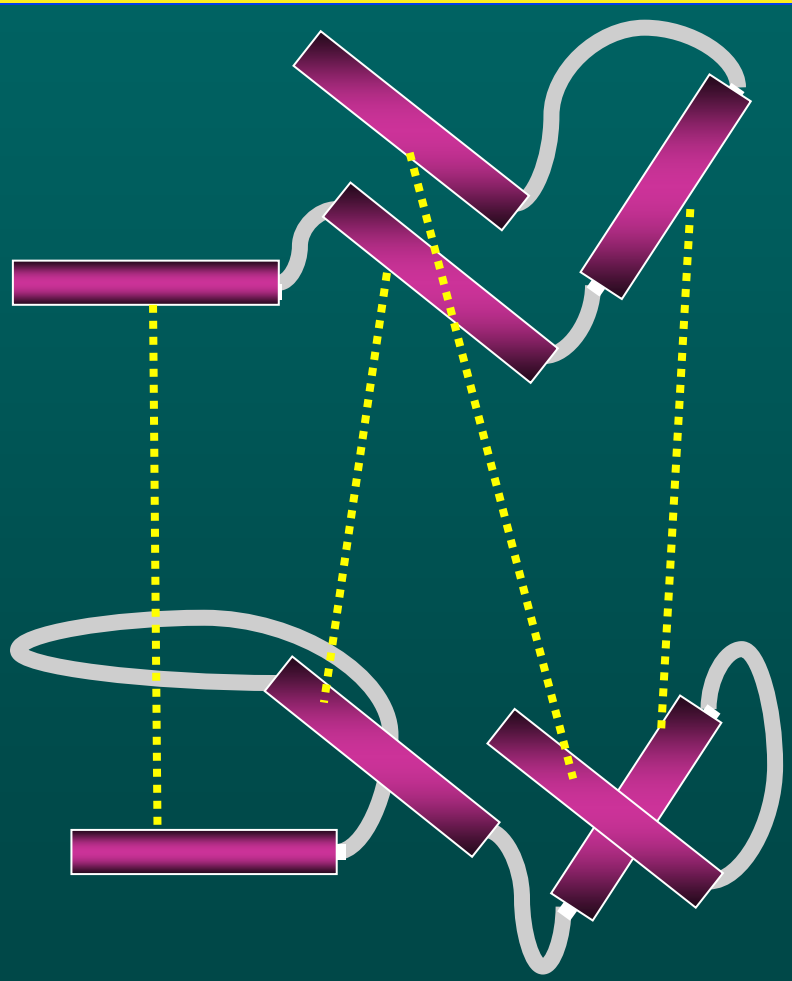


1. Find region of alignment (slow)
2. Calculate similarity score (fast)

Fold similarity from simplified representations 3-D

The secondary structure element approach





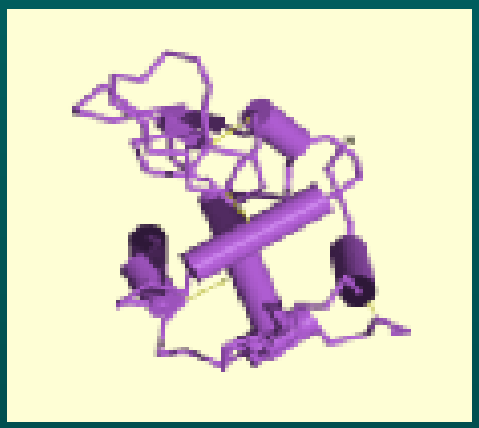
1. Find the SSE's (subjective)

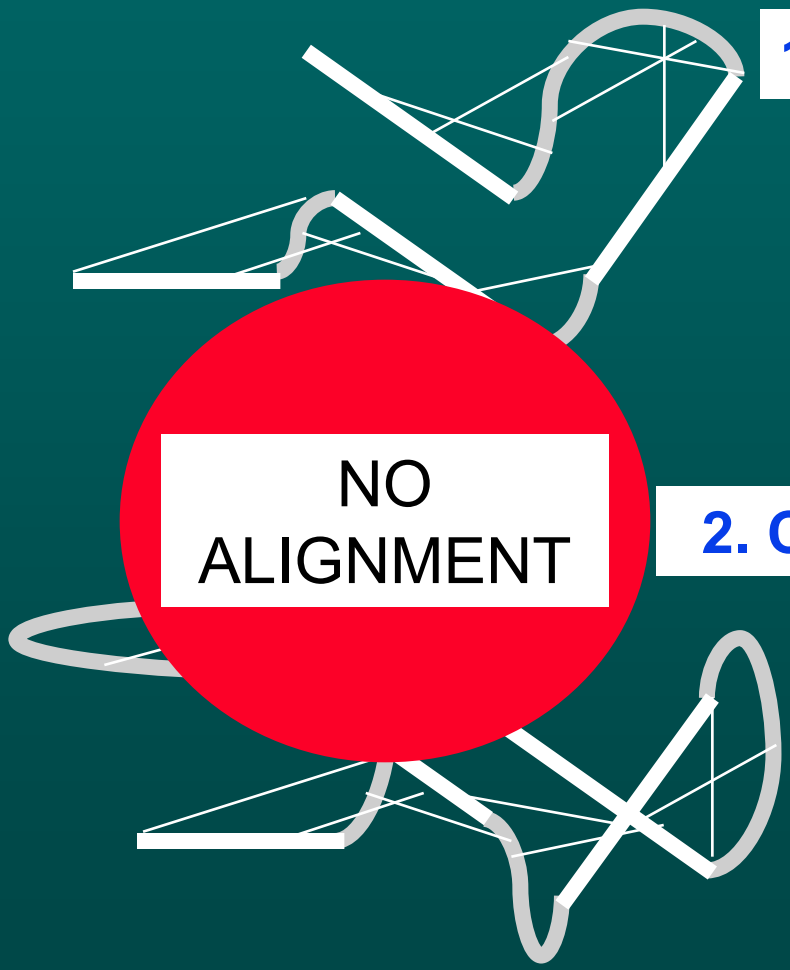
2. Align them (fast)

3. Compute "score" (topology, orientation etc.) (fast)

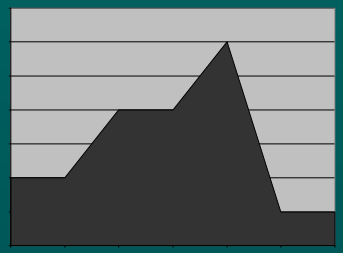
Fold similarity from C_α distance distribution (PRIDE)

3-D

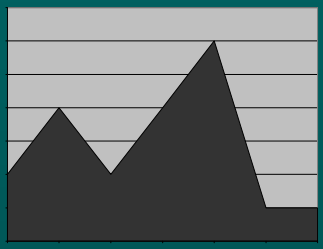




1. Take Cα -Cα distances (fast)

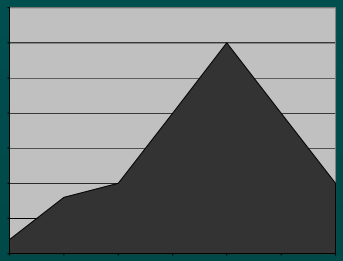


i-(i+6)

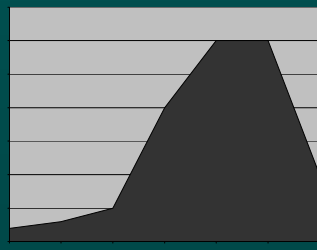


i-(i+10)

2. Calculate Cα -Cα distributions



i-(i+6)



i-(i+10)

3. Calculate Probability of Identity

Oliviero Carugo, 2001

Fold classification based on PRIDE

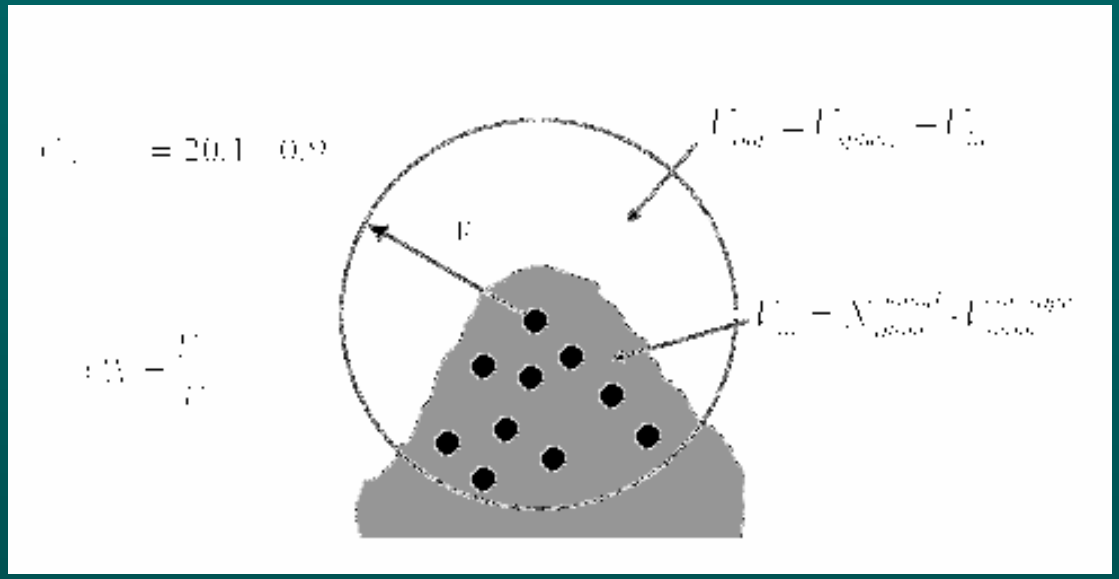
CATH category	PRIDE		Nearest neighbour within the same group
	self (within group)	non-self (outside the group)	
I Identical representatives	0.97 ± 0.06 (72,517)	0.17 ± 0.15 (106,909,361)	69.3 %
N Nearly-identical representatives	0.93 ± 0.12 (75,429)	0.17 ± 0.15 (106,833,932)	87.3 %
S Sequence family	0.56 ± 0.22 (800,060)	0.17 ± 0.15 (106,033,872)	98.5 %
H Homologous superfamily	0.44 ± 0.19 (1,071,841)	0.16 ± 0.15 (104,962,031)	98.8 %
T Topology	0.35 ± 0.17 (2,568,350)	0.16 ± 0.15 (102,393,681)	99.0 %
A Architecture	0.26 ± 0.19 (9,028,969)	0.15 ± 0.15 (93,364,712)	98.9 %
C Class (α, β, α/β or few sec. str.)	0.26 ± 0.20 (24,547,471)	0.11 ± 0.12 (68,817,241)	99.5 %

Up to 1000 comparisons per second

CX, DPX – protein 3-D visualization

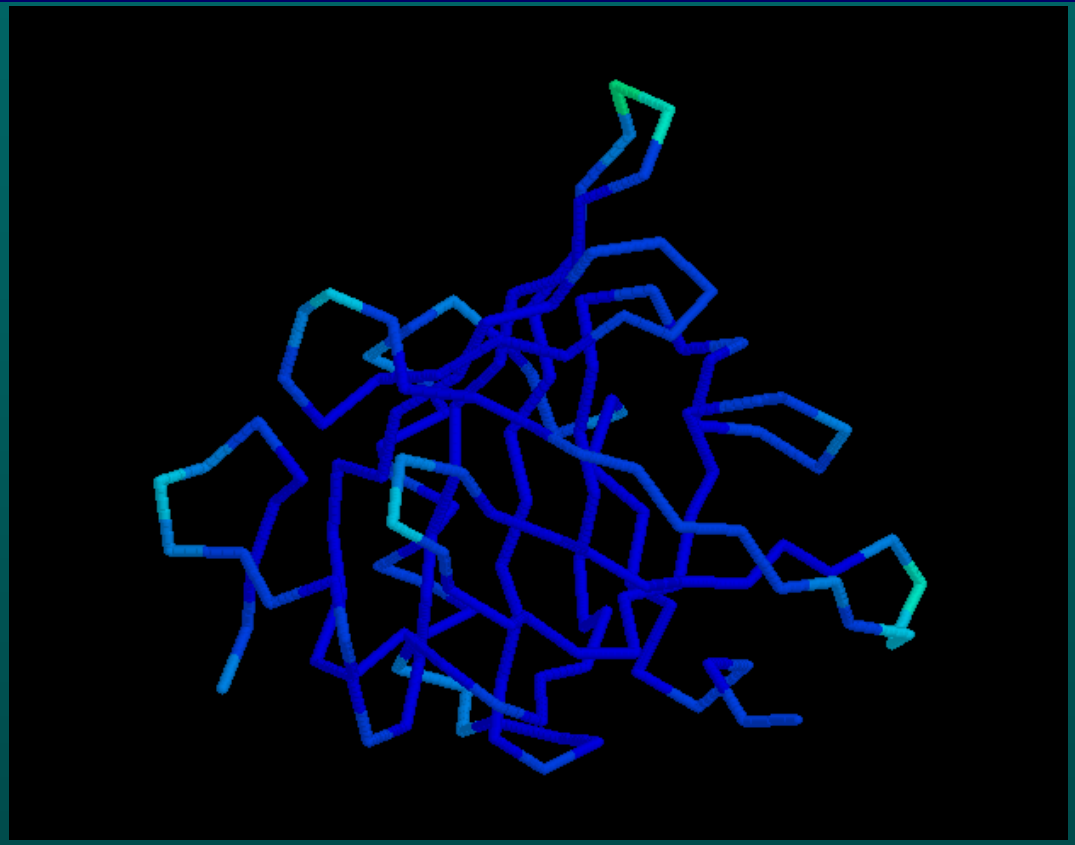
- Protein structures are complicated.
- Molecular graphics programs such as Rasmol or Swiss PDB viewer can color the structures according to various properties.
- How can we visualize new properties? We can pre-compute a property, write it back into an input file and use Rasmol or Swiss PDB viewer
- CX and DPX calculate geometric properties for visualization.

CX – identifying protruding atoms in proteins



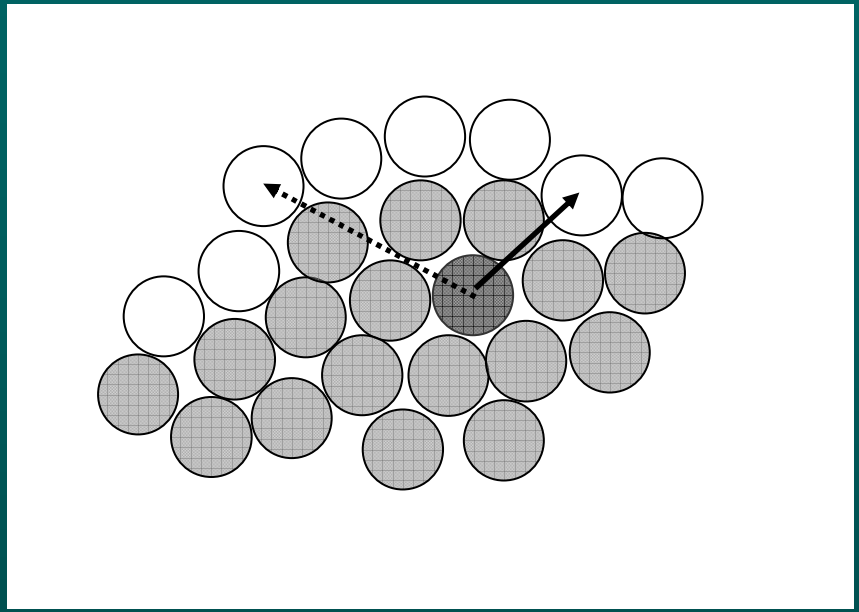
- An atom is “protruding” if it has very few neighbors in its vicinity. This can be calculated, and the result is put back into a PDB file (in place of the so-called Temperature factor data)

Protruding atoms in *1mup* (pheromon binding protein)



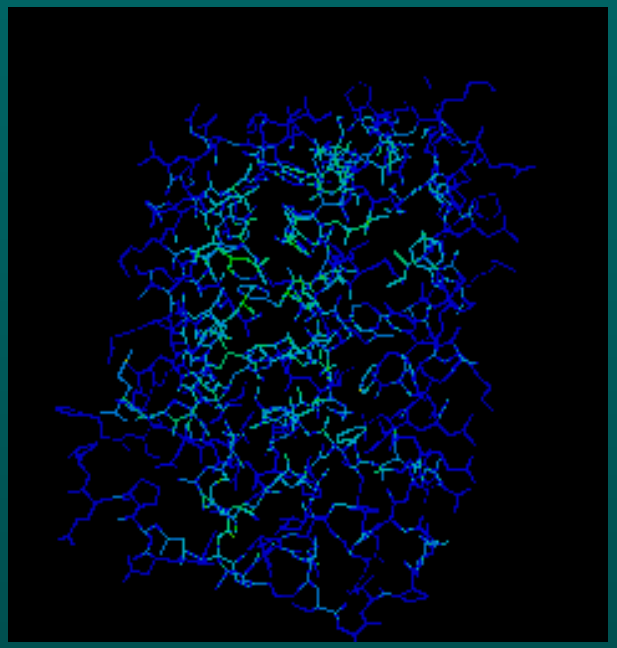
- 1mup.pdb submitted to the CX server. File visualized with Rasmol, Colored by “Temperature” and Backbone displayed

DPX visualization of atom depth



- Atom depth is the distance from the nearest surface atom. This is calculated and filled into the Temperature factor field of the PDB file

Atom depth in *1itf* (interferone)

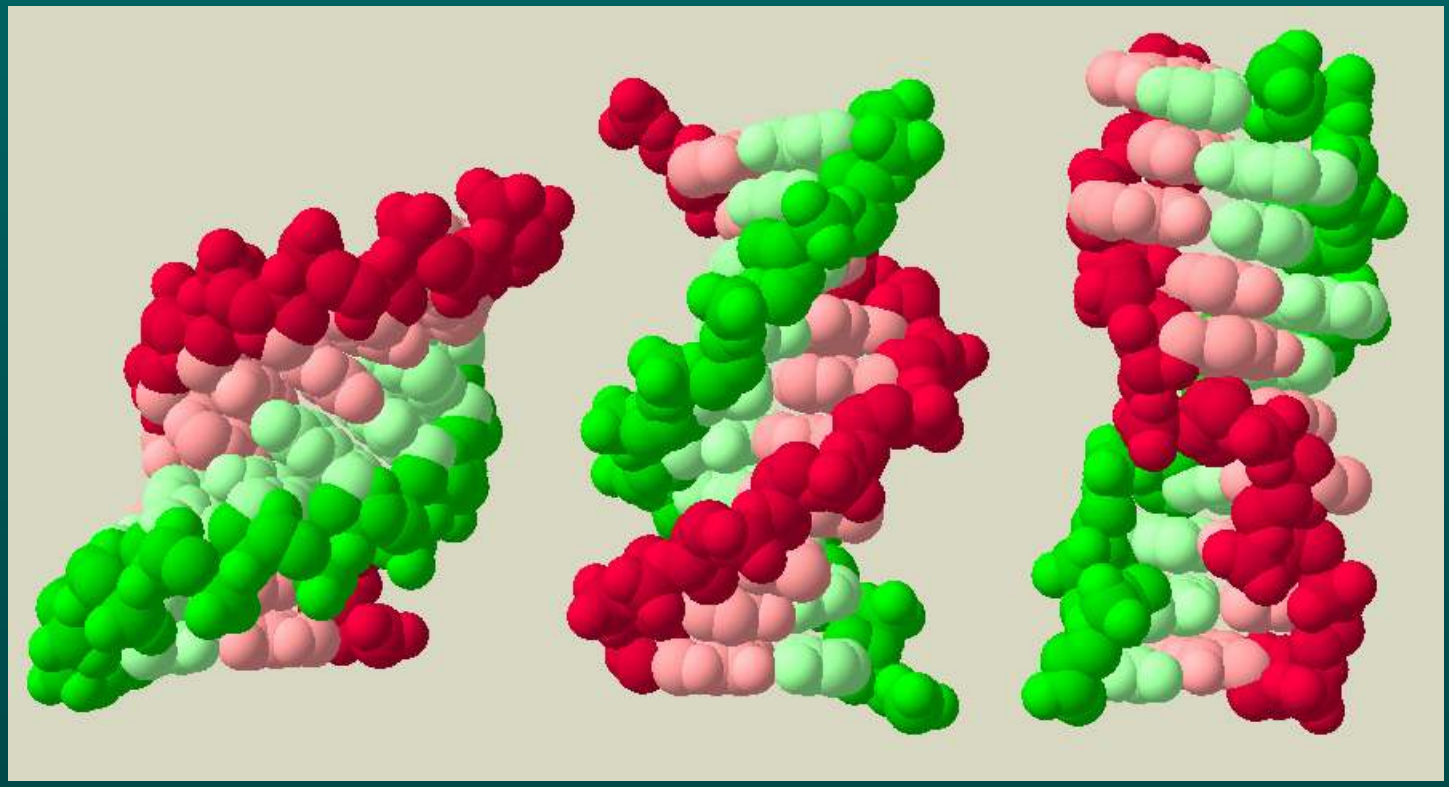


- 1itf.pdb submitted to the DPX server. File visualized with Rasmol, Colored by “Temperature” and Wireframe displayed

Prediction of structural DNA features from sequence

Visualization of DNA properties

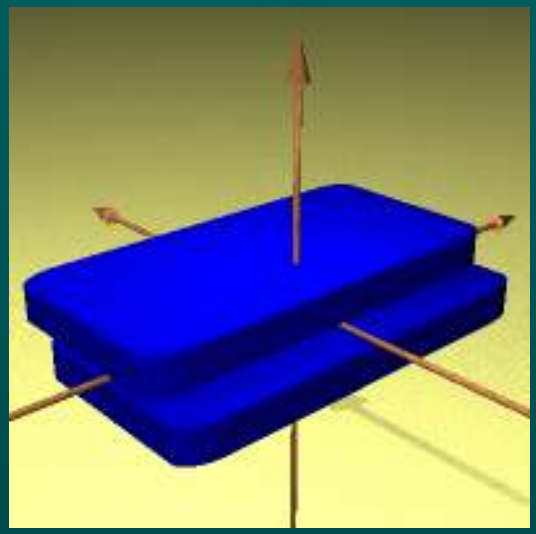
DNA structure polymorphism



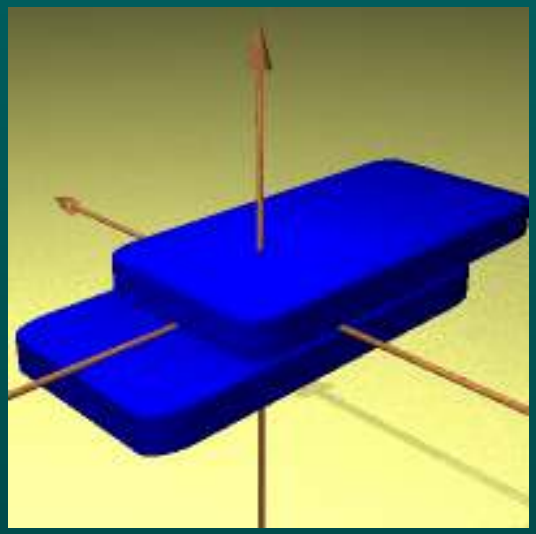
A B Z

Basepair geometry parameters

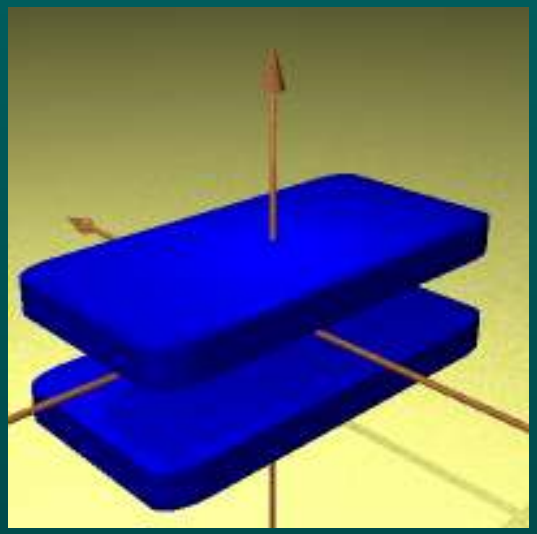
- Translational



Shift



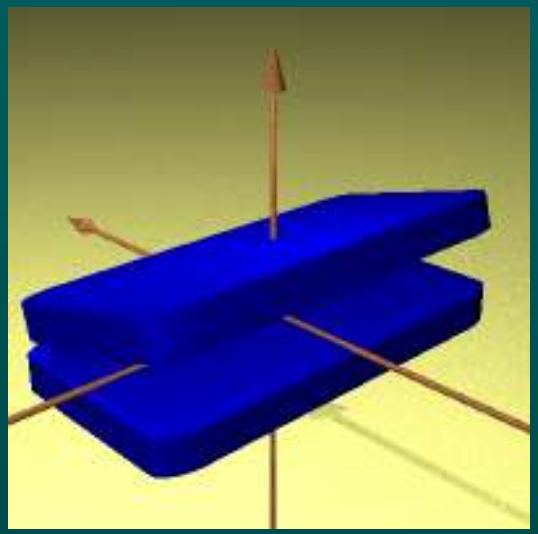
Slide



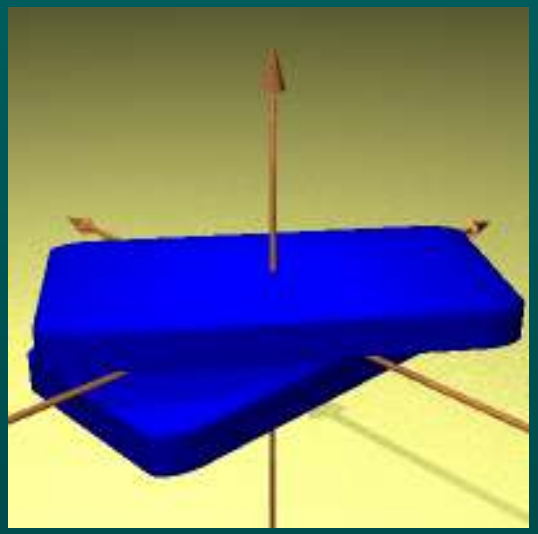
Rise

Basepair geometry parameters

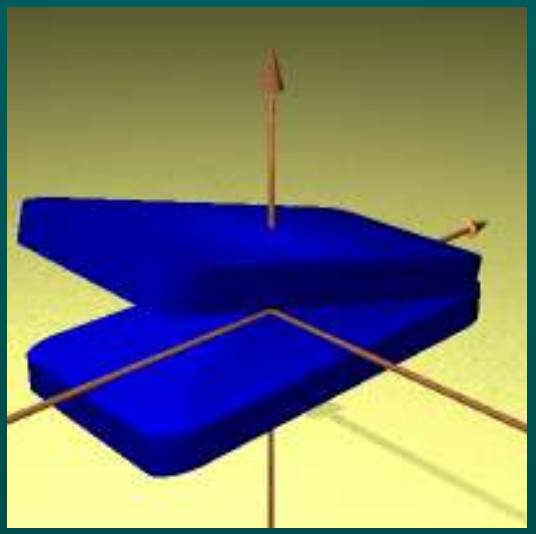
- Rotational



Roll

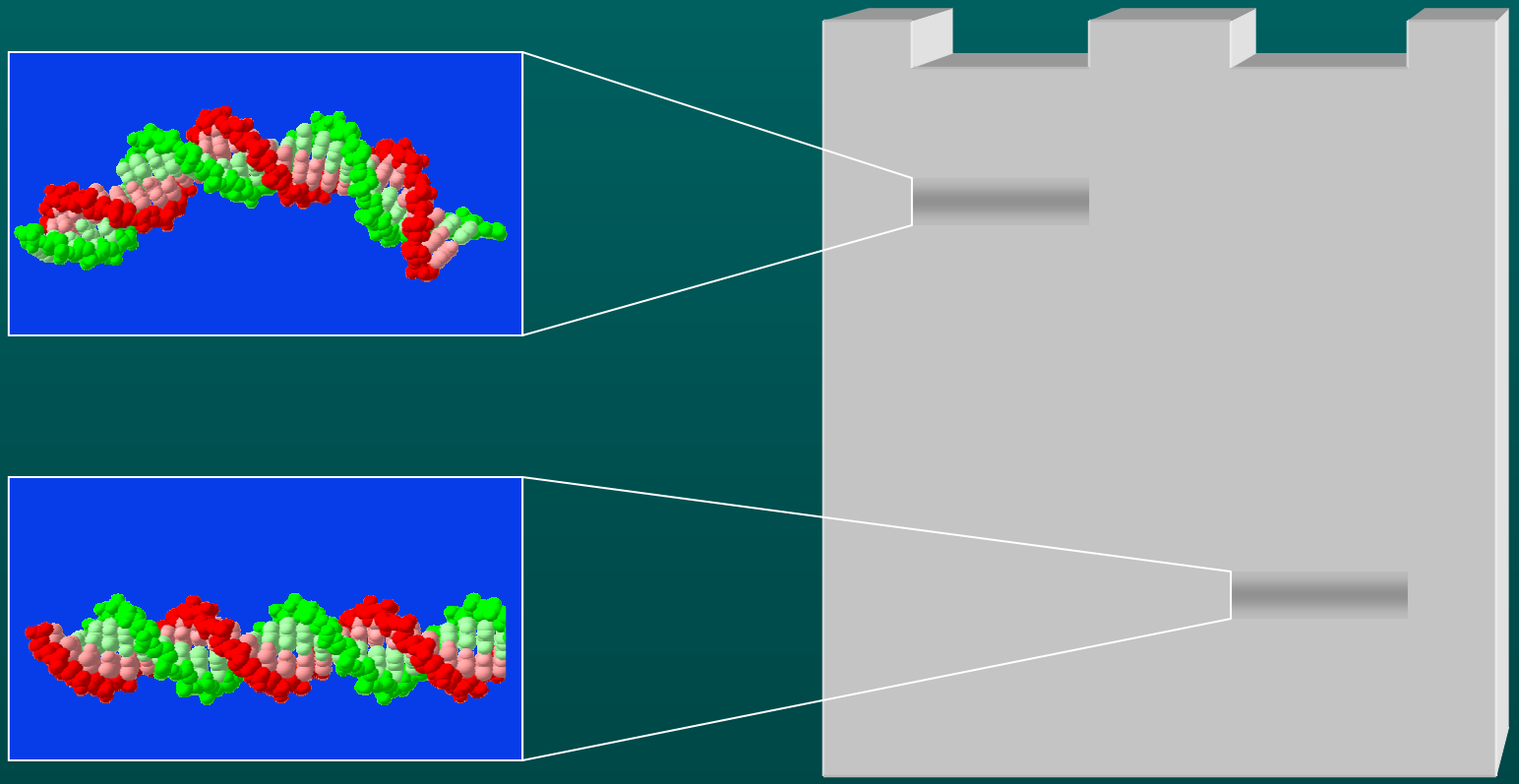


Twist



Tilt

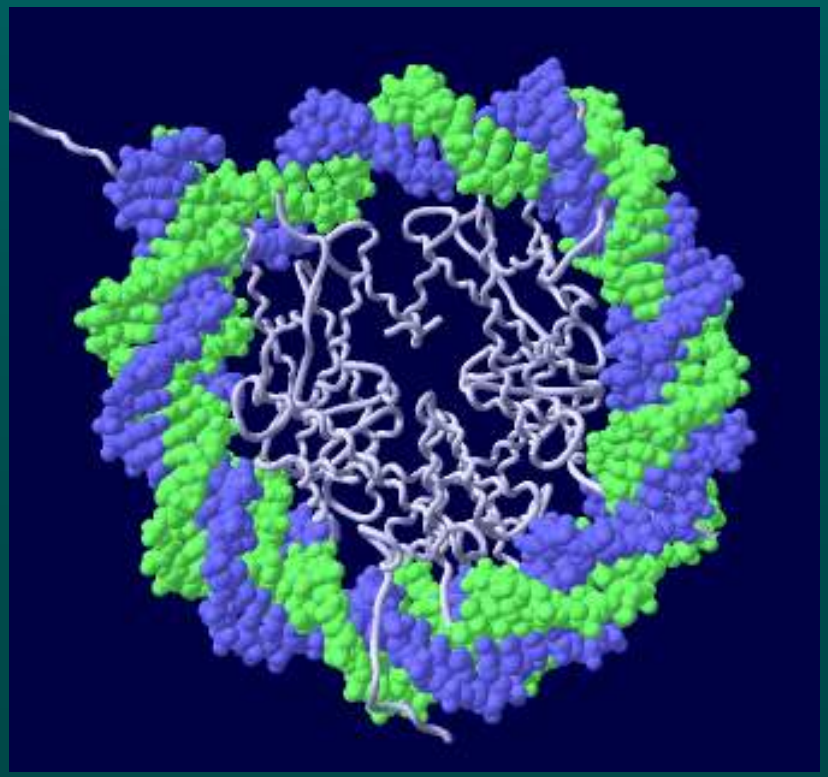
DNA is Curved

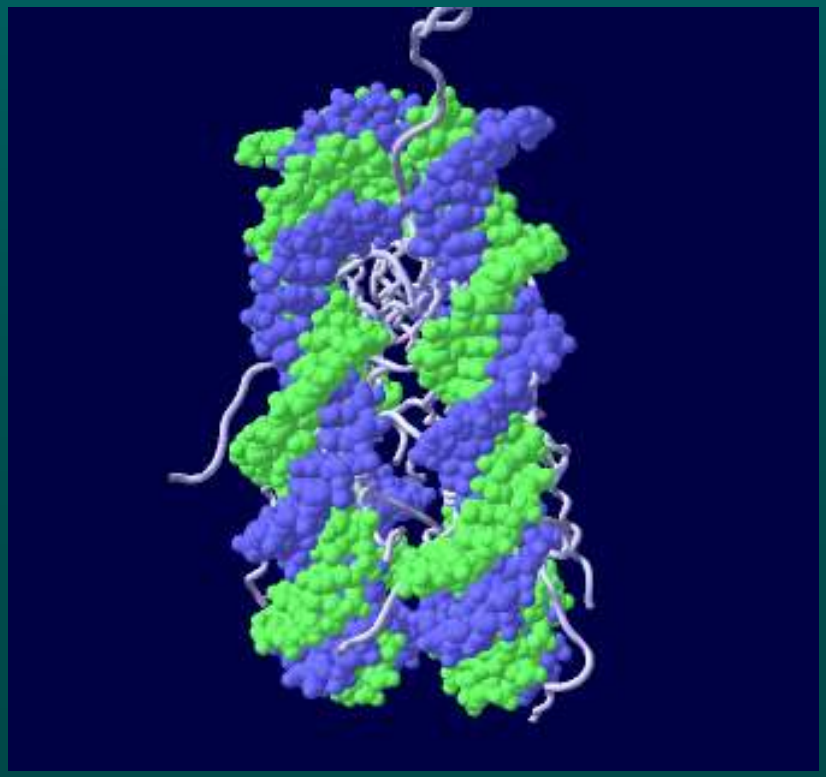


Wu & Crothers, Nature (1984)

Crystallographic example

- Nucleosome



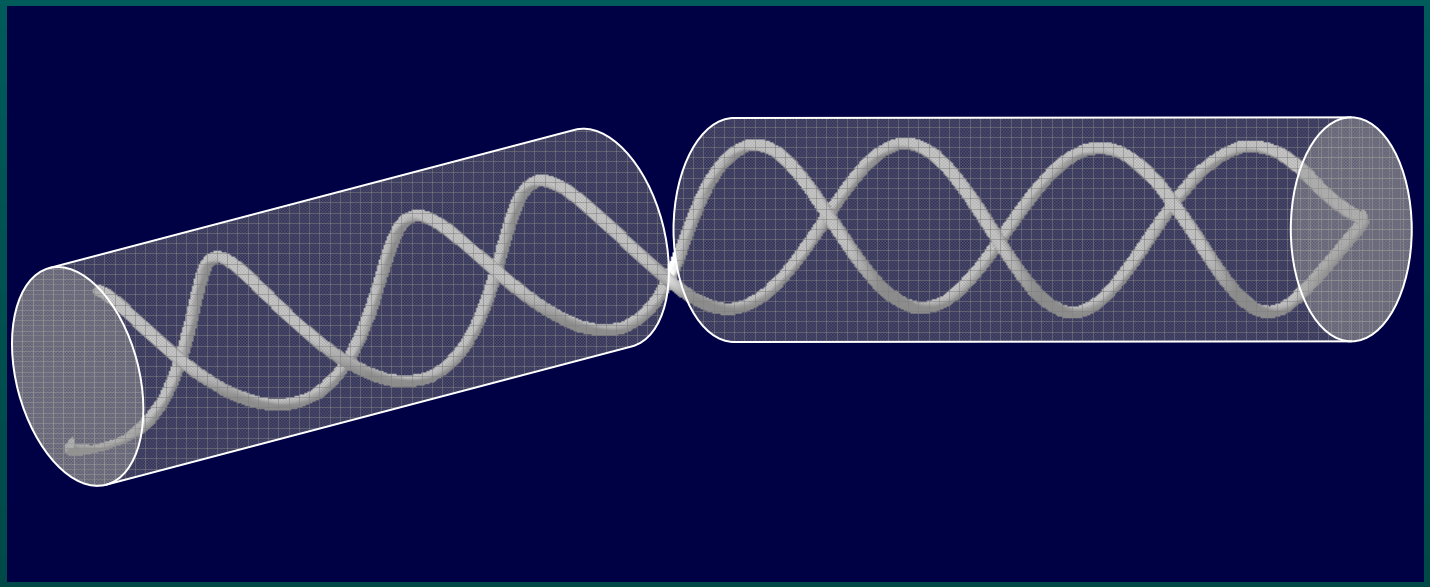


Curvature and Bendability

- Transcription signal
- Replication signal
- Chromatin condensation
- ?

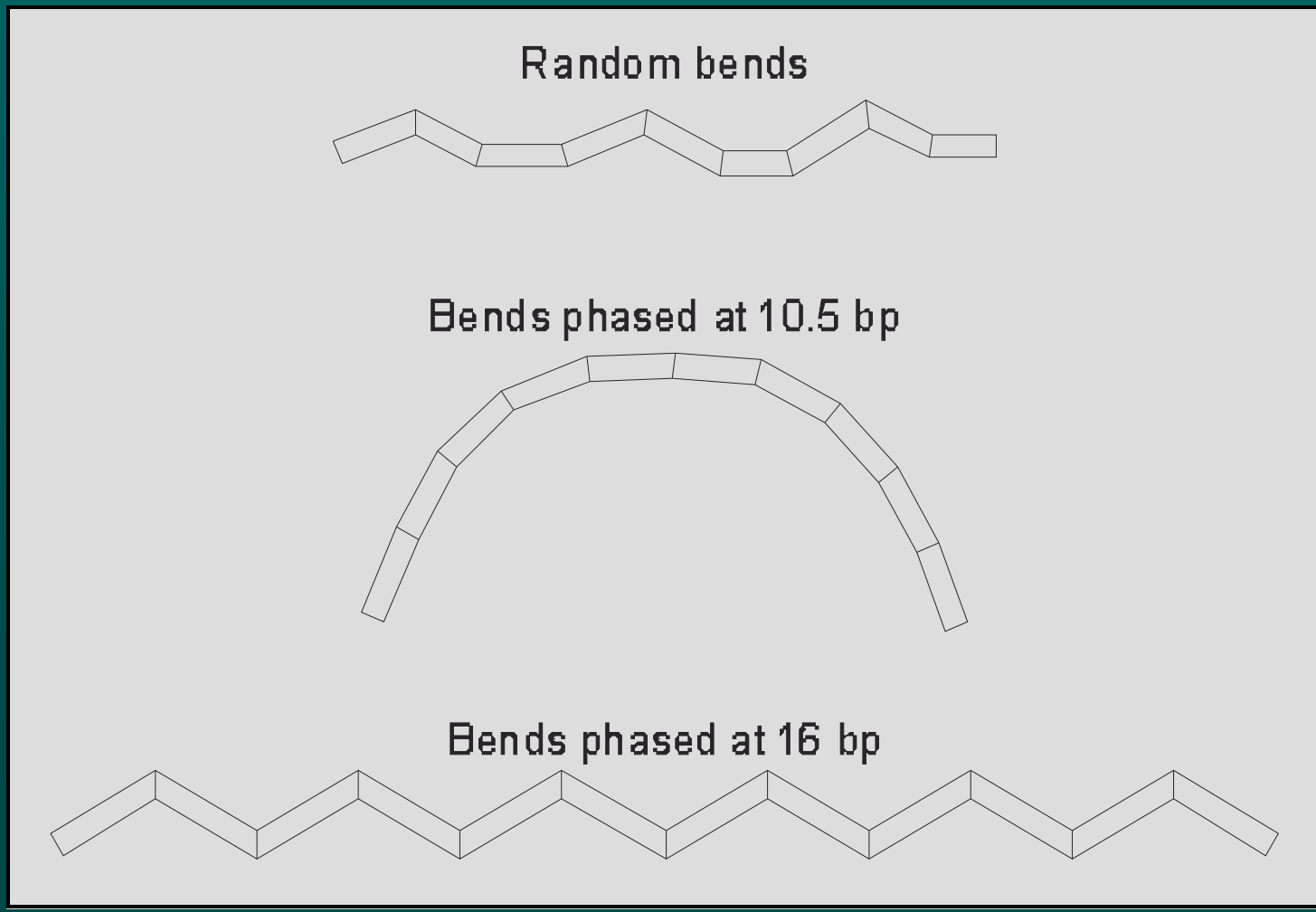
Curvature models

- Junctions between different DNA forms (A- & B-)



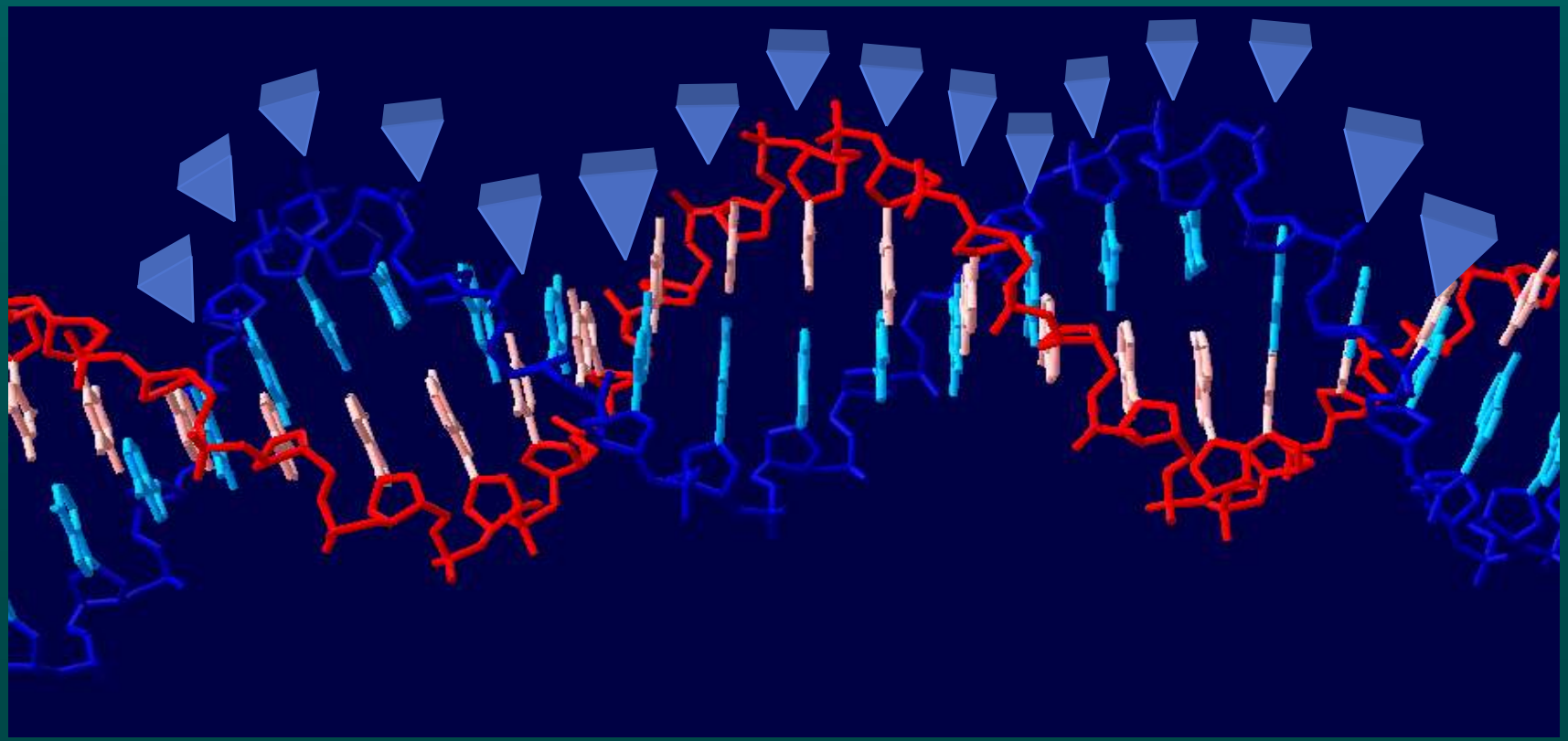
Curvature models

- Junction phasing



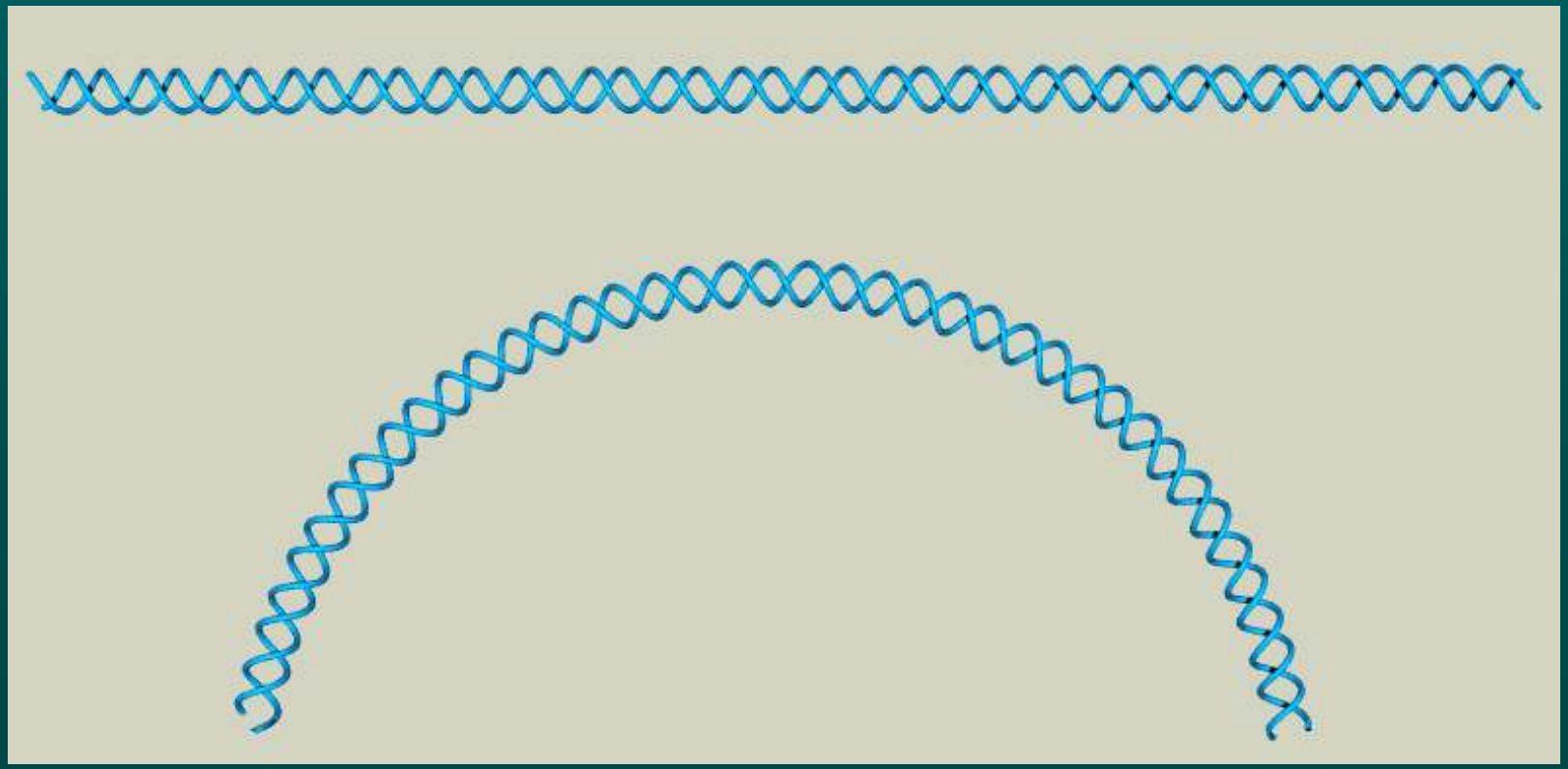
Curvature models

- ‘Wedge’ model



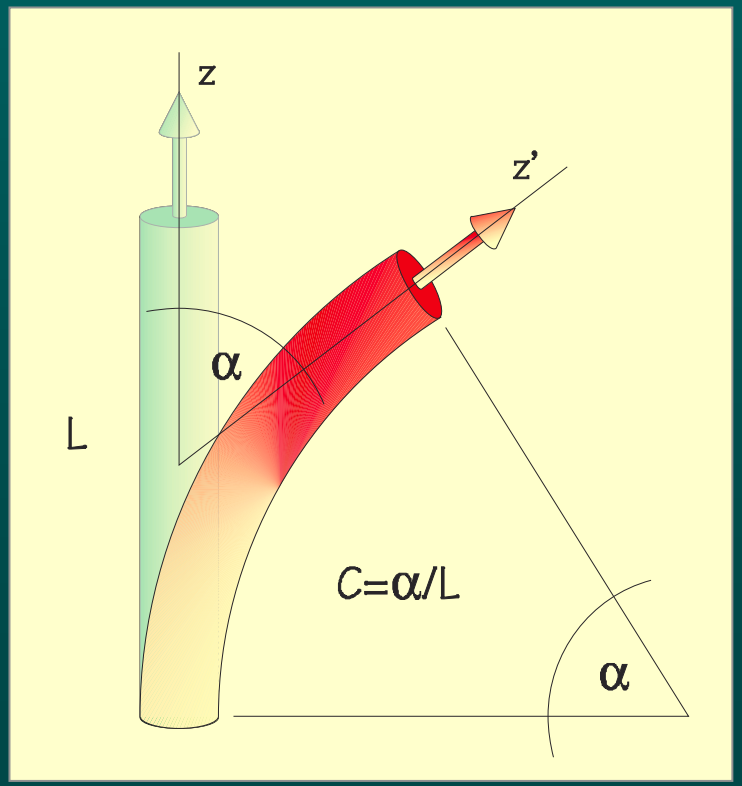
Curvature models

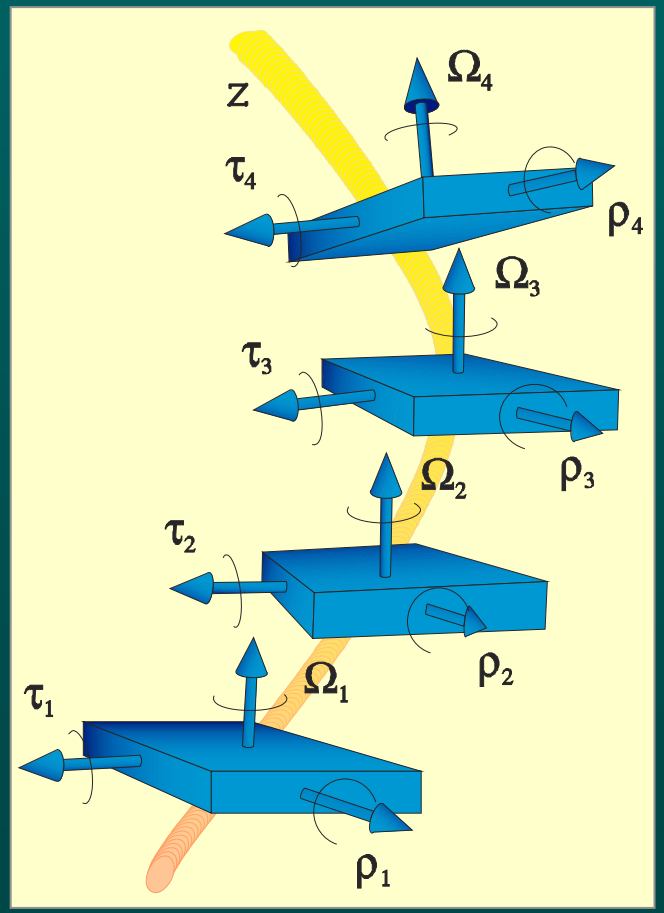
- 'Wedge' model



‘Measuring’ DNA curvature

- Curvature of an elastic rod





Curvature prediction

- **How many motifs?**

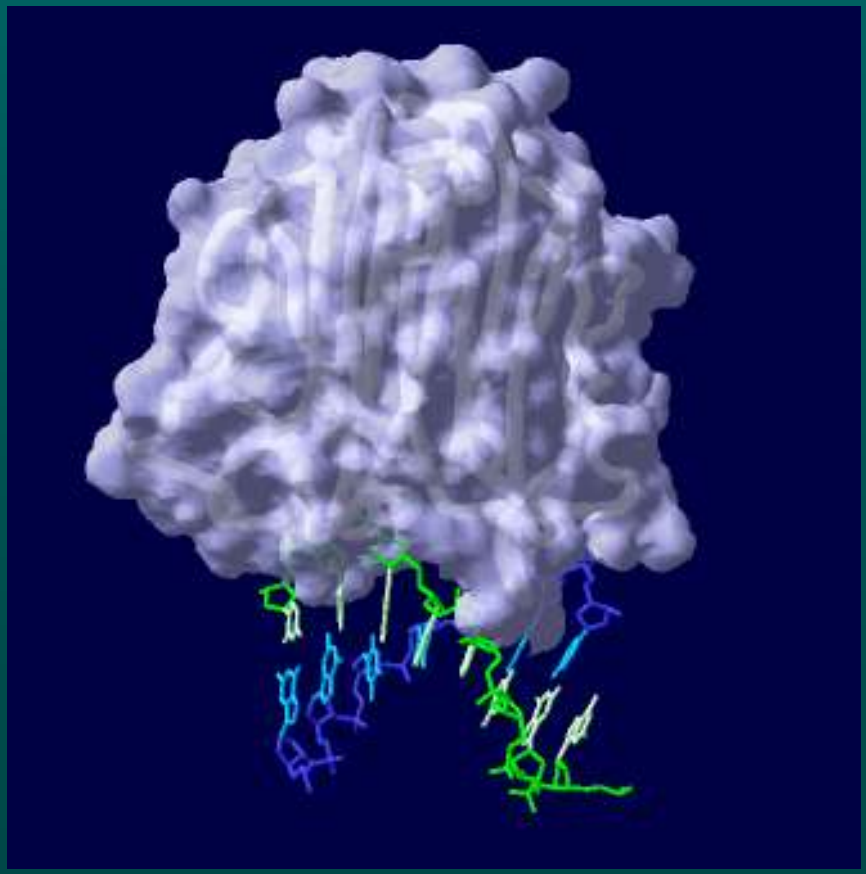
	ss	ds
Dinucleotide	16	10
Trinucleotide	64	32
Tetranucleotide	256	136

- **Data sufficient for:**

	Dinucleotide models	Trinucleotide models
X-ray	+	
NMR	+	
Gel electrophoresis	+	
Nucleosome positioning		+
DNaseI digestion	+	+

DNA bendability

- DNase I (E.C. 3.1.21.1)
  - sequence independent
  - bends DNA
  - digestion frequency proportional to bendability

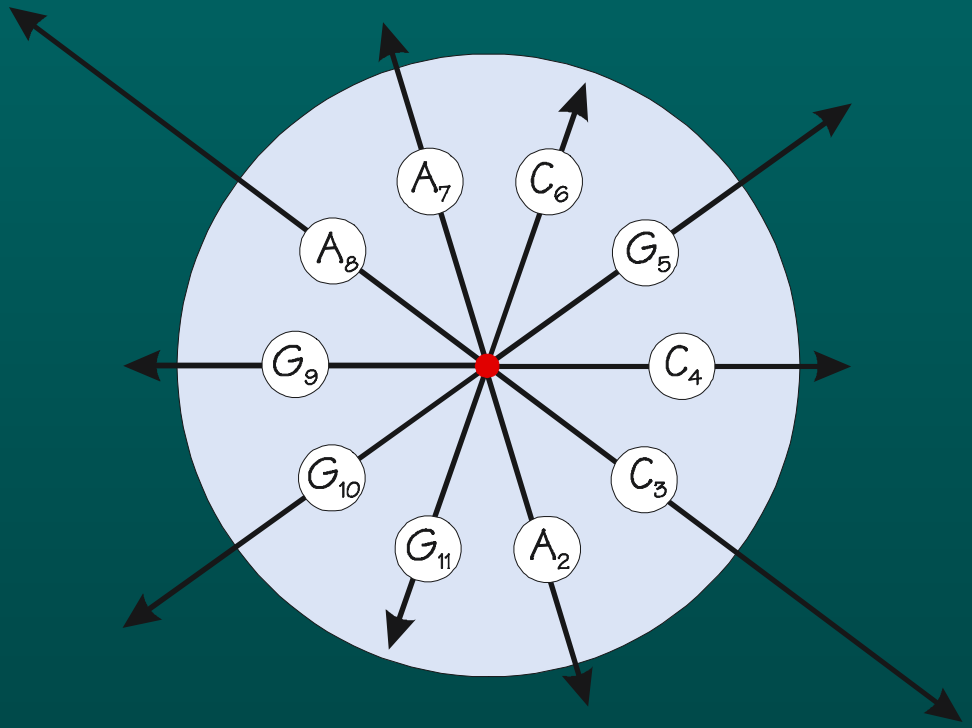


Bendability plot

Trinucleotide bendability (a.u.)

AAA/TTT	0.1	CAG/CTG	9.6
AAC/GTT	1.6	CCA/TGG	0.7
AAG/CTT	4.2	CCC/GGG	5.7
AAT/ATT	0.0	CCG/CGG	3.0
ACA/TGT	5.8	CGA/TCG	5.8
ACC/GGT	5.2	CGC/GCG	4.3
ACG/CGT	5.2	CTA/TAG	7.8
ACT/AGT	2.0	CTC/GAG	6.6
AGA/TCT	6.5	GAA/TTC	5.1
AGC/GCT	6.3	GAC/GTC	5.6
AGG/CCT	4.7	GCA/TGC	7.5
ATA/TAT	9.7	GCC/GGC	8.2
ATC/GAT	3.6	GGA/TCC	6.2
ATG/CAT	8.7	GTA/TAC	6.4
CAA/TTG	6.2	TAA/TTA	7.3
CAC/GTG	6.8	TCA/TGA	10.0

Bendability symmetrical in most DNA segments



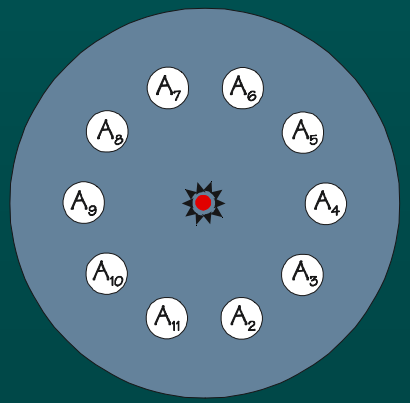
B=7.3

$$B = \frac{1}{n} \sum b_n$$

Genomic DNA:

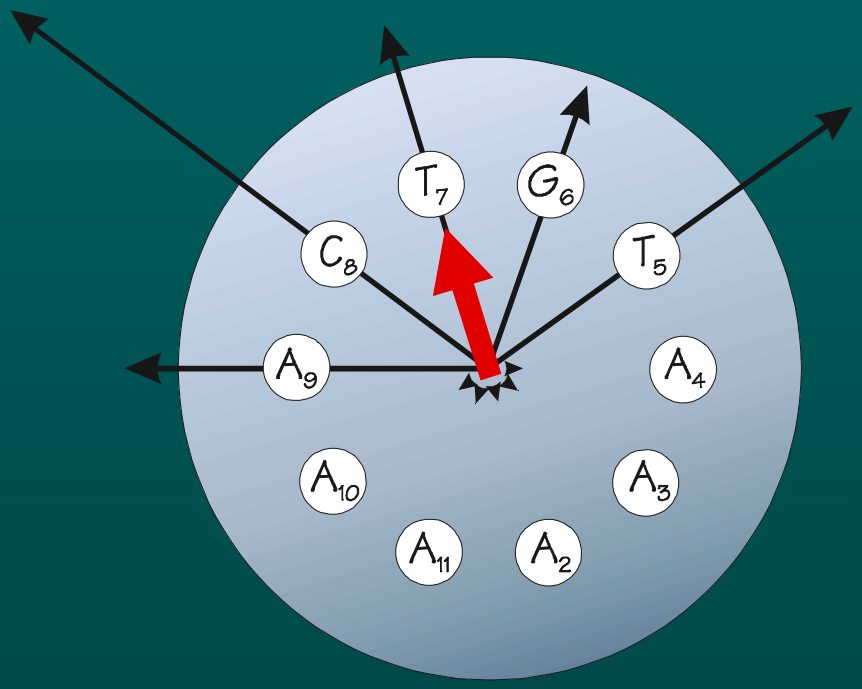
average ~ 5.0

s.d. ~ 0.5



B=0.0

Bendability assymetrical in curved segments



H=24.3

$$\vec{H} = \frac{1}{n} \sum \vec{b}_n$$

Genomic DNA:

average ~ 5.0

s.d. ~ 2-3

Assymetry is a measure of predicted curvature

Prediction quality

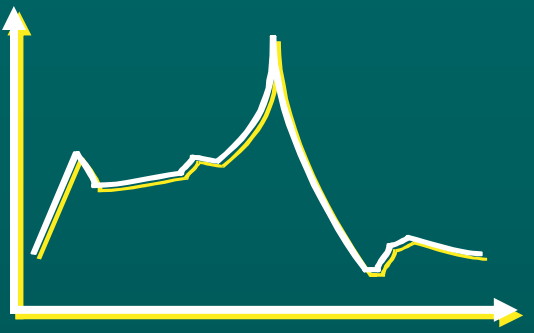
Sequence motif	PAGE	X-rays	NMR	Nucleos.	DnaseI
Curved DNA (deg./h.t.)					
$(aaaattttgc)_n$	26.2	6.9	18.3	13.7	17.4
$(aaaattttcg)_n$	21.0	3.8	3.8	17.7	17.2
$(tctcaaaaaacgcgaaaaaaccggaaaaaagc)_n$	27.1	8.2	16.7	17.1	15.9
$(ccgaaaaagg)_n$	14.7	6.8	13.1	23.3	20.2
$(tctctaaaaaatatataaaaa)_n$	27.8	3.0	7.5	10.9	18.5
(ggcaaaaaac)n	26.8	12.0	20.1	20.4	19.5
ccaaaaatgtcaaaaaataggcaaaaaatgcc	26.0	6.4	15.7	19.6	20.5
Straight DNA (deg./h.t.)					
$(atctaatctaacacaacaca)_n$	0.8	0.5	2.7	1.2	0.8
actacgttaaatctatcaccgcaagggataaa	10.4	5.5	4.9	5.9	7.8
actacgttaaatctatcaccacaagggataaa	11.0	5.5	3.4	6.2	8.1
$(a)_n$ - poly-A	0.0	0.0	0.0	0.0	0.0
$(ttttaaaacg)_n$	1.5	7.1	14.5	10.7	4.2
$(ttttaaaagc)_n$	1.7	0.8	16.0	16.3	9.6
$(aaaaactctctaaaaactctcgggccctagaggggccc)_n$	27.1	3.1	5.9	7.5	8.3

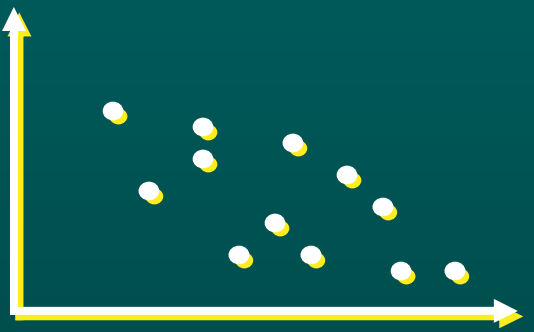
Parametric representation

ATGACGTAATAATGC . . .
(SEQUENCE)

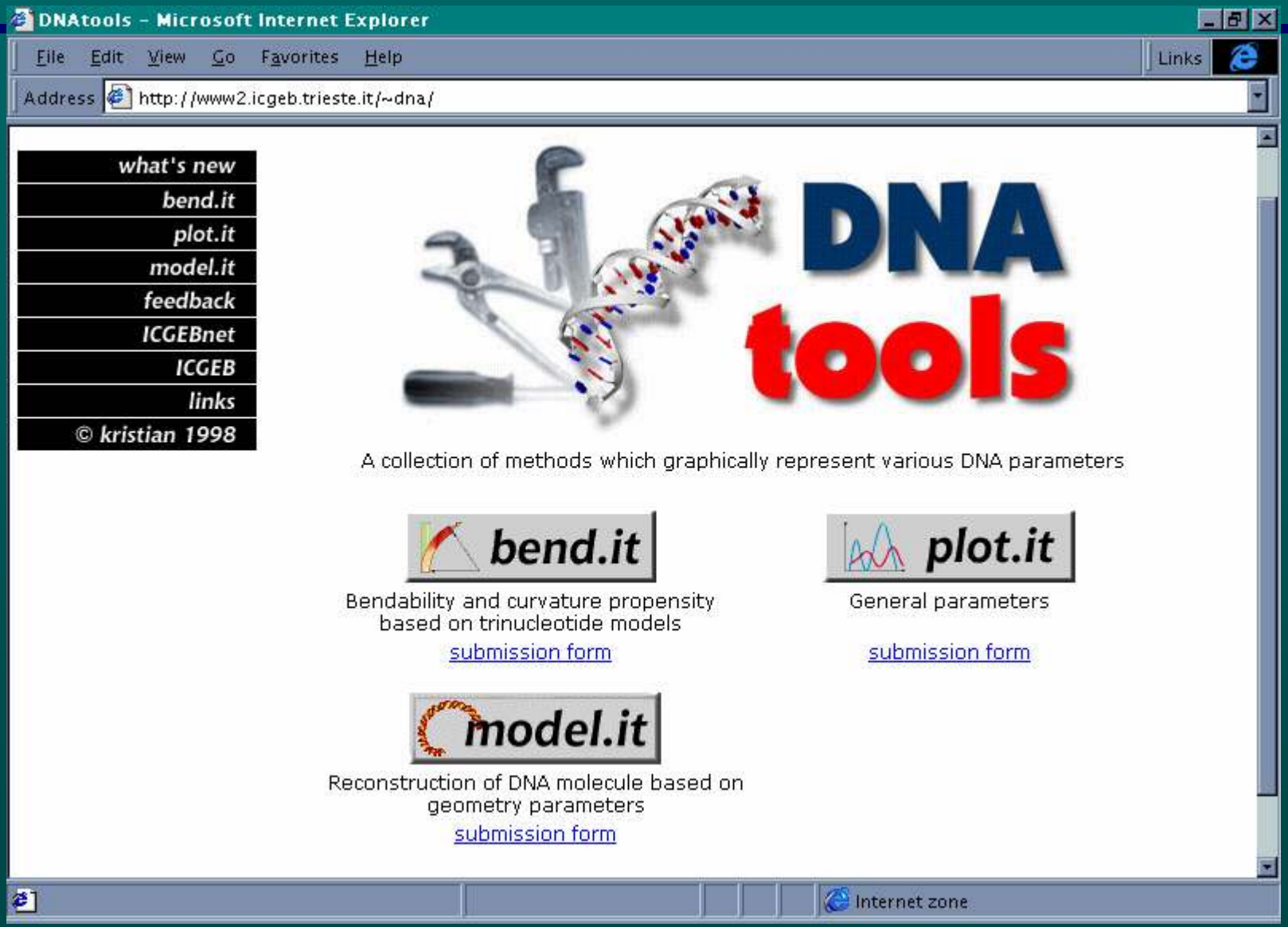
AAA 0.1
AAC 1.6
. . .
(PARAMETER SET)

Prediction methods





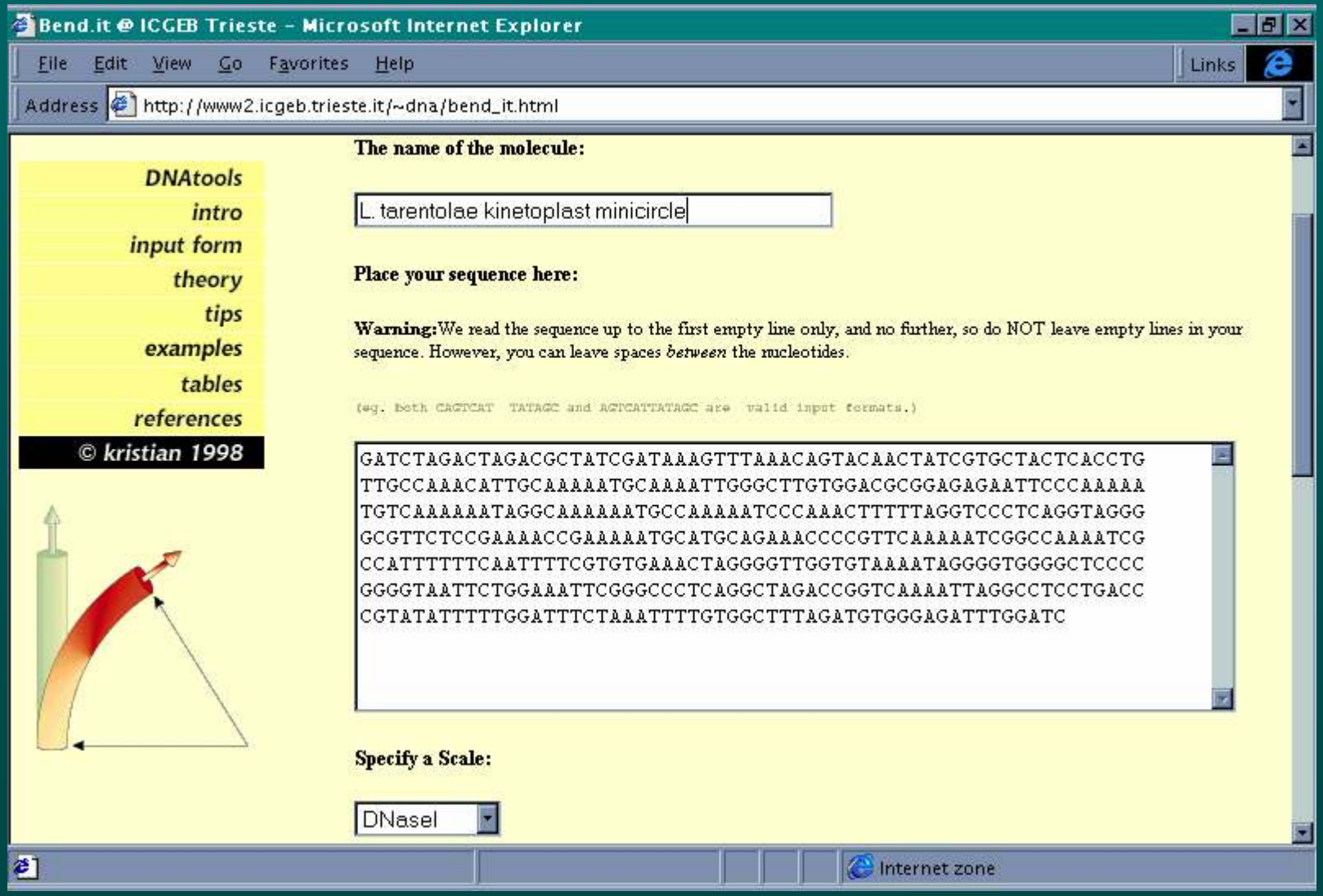
DNAtools - http://www.icgeb.trieste.it/dna



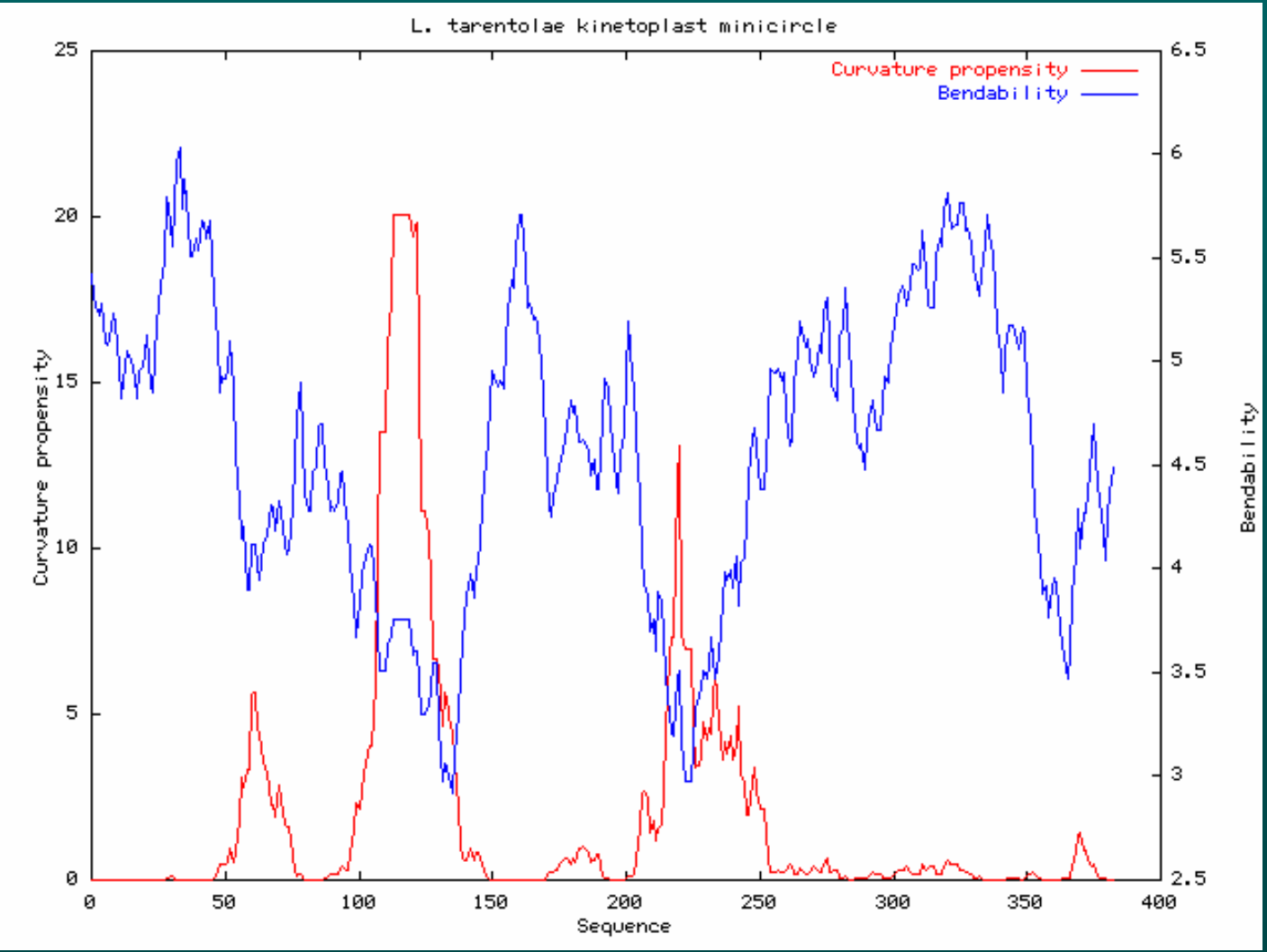
plot.it – General parameters

1 - G+C content
2 - bendability (Brukner et al.)
3 - bendability (Sarai et al.)
4 - trinuc. roll from nucleosome (Satchwell et al.)
5 - consensus (Brukner and Satchwell) bendability
6 - Watson-Crick interaction energy (Lewis-Sankey)
7 - ΔG of B ↔ A form transition (Aida et al.)
8 - ΔG of B ↔ Z form transition (Hartman et al.)
9 - B -> A form transition propensity (Ivanov et al.)
10 - ΔH of B ↔ coil transition (SantaLucia et al.)
11 - ΔG of B ↔ coil transition (SantaLucia et al.)
12 - ΔS of B ↔ coil transition (SantaLucia et al.)
13 - ΔH of B ↔ coil transition (Sugimoto et al.)
14 - ΔG of B ↔ coil transition (Sugimoto et al.)
15 - ΔS of B ↔ coil transition (Sugimoto et al.)
16 - ΔH of B ↔ coil transition (Breslauer et al.)
17 - ΔG of B ↔ coil transition (Breslauer et al.)
18 - ΔS of B ↔ coil transition (Breslauer et al.)
19 - roll (Bansal et al.)
20 - roll (Bolshoy et al.)
21 - roll (DeSantis et al.)
22 - roll (Gorin et al.)
23 - roll (Uljanov and James)
24 - tilt (Bansal et al.)
25 - tilt (Bolshoy et al.)
26 - tilt (DeSantis et al.)
27 - tilt (Gorin et al.)
28 - tilt (Uljanov and James)
29 - twist (Bansal et al.)
30 - twist (Bolshoy et al.)
31 - twist (DeSantis et al.)
32 - twist (Gorin et al.)
33 - Complexity

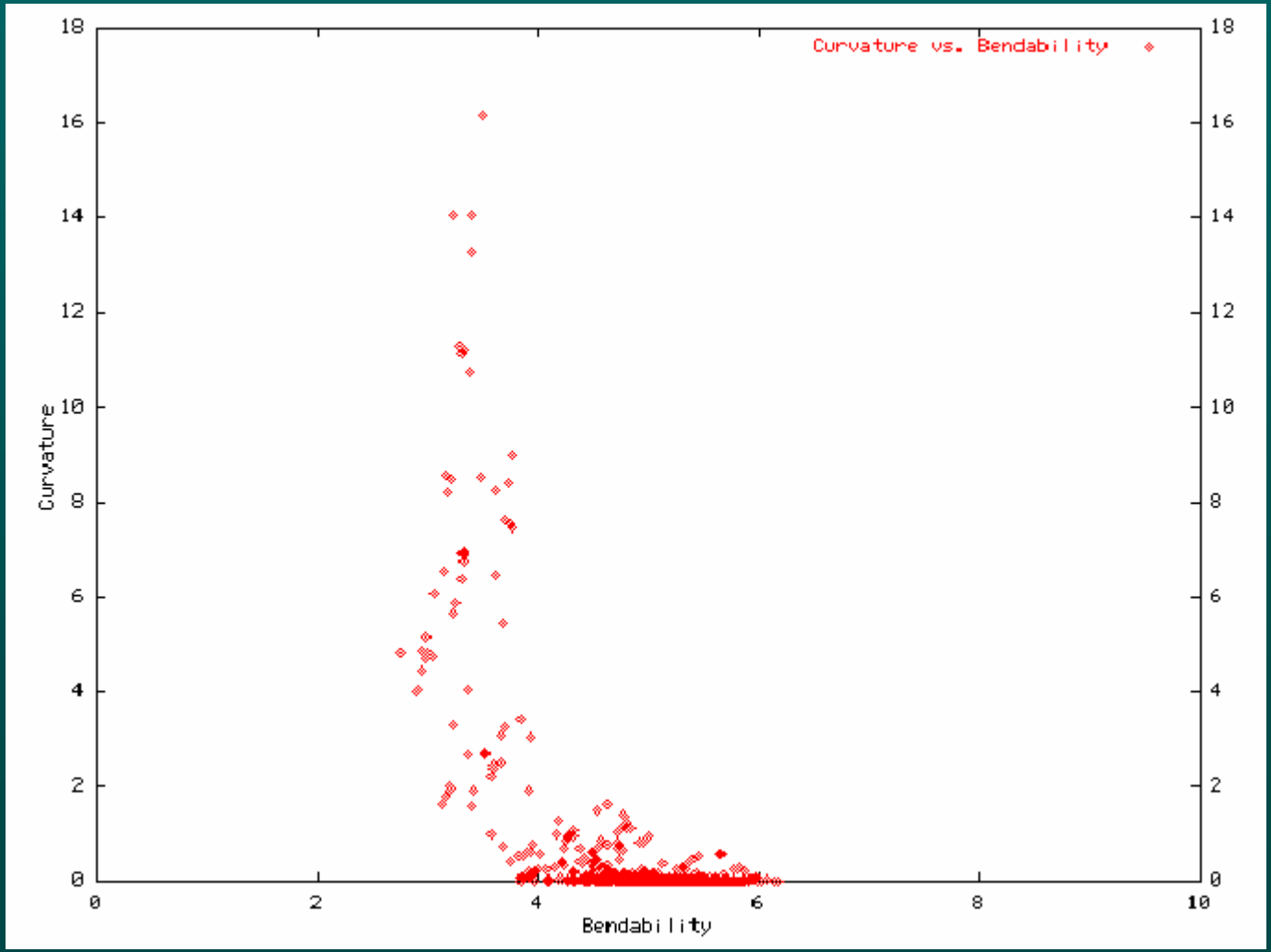
bend.it – Curvature & bendability



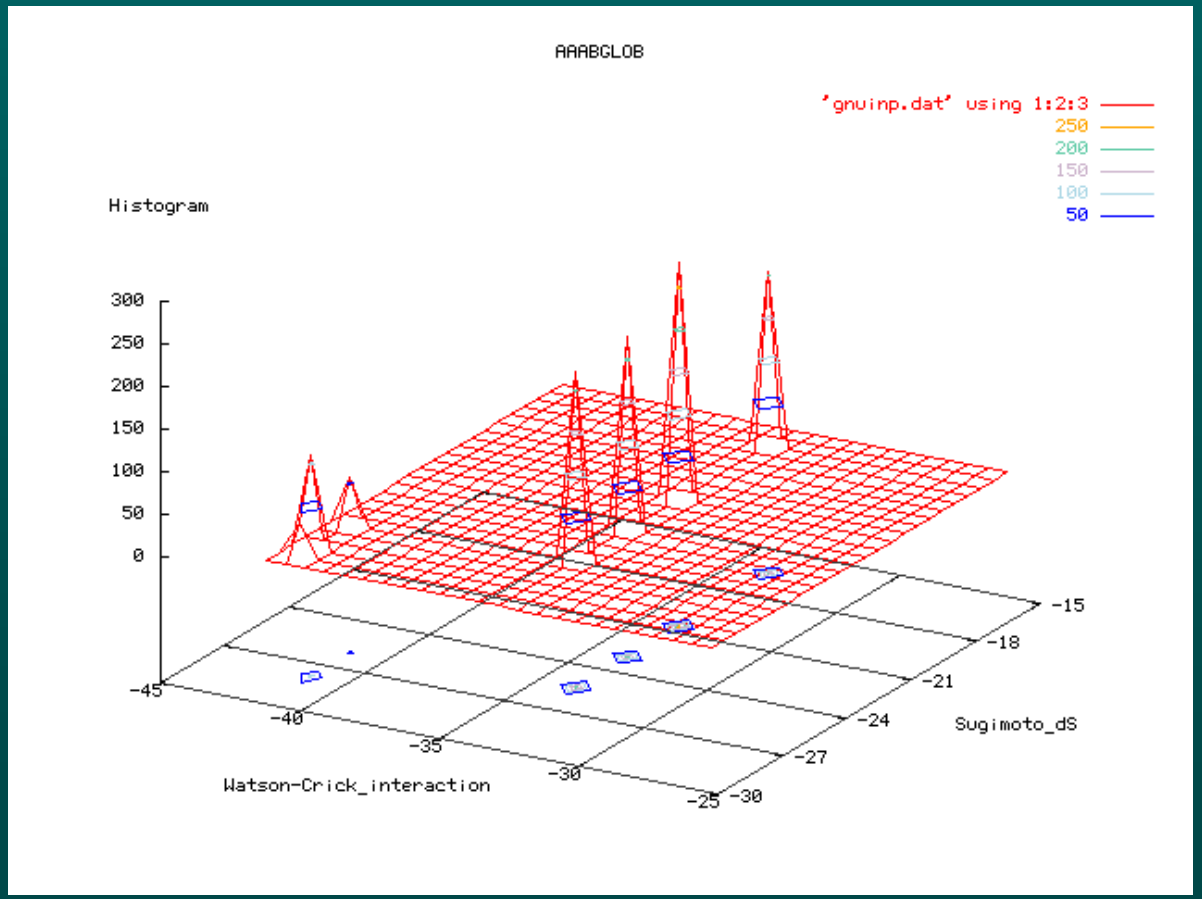
bend.it - 1D parametric plot



bend.it - 2D parametric plot



Histogram (3D) plot for long DNA

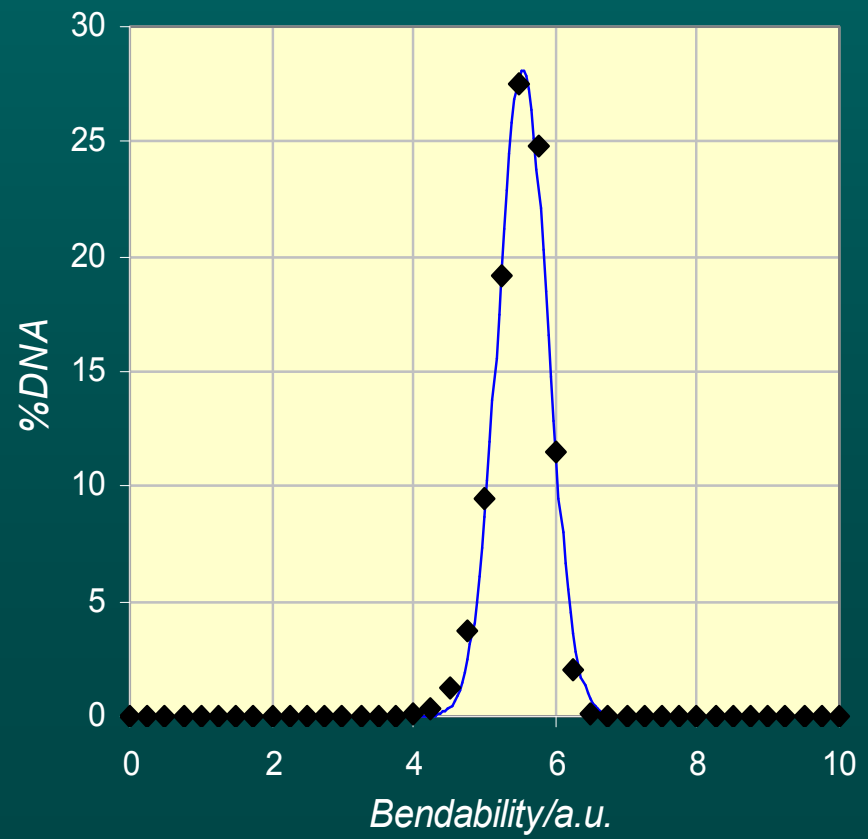


bend.it –Curvature models

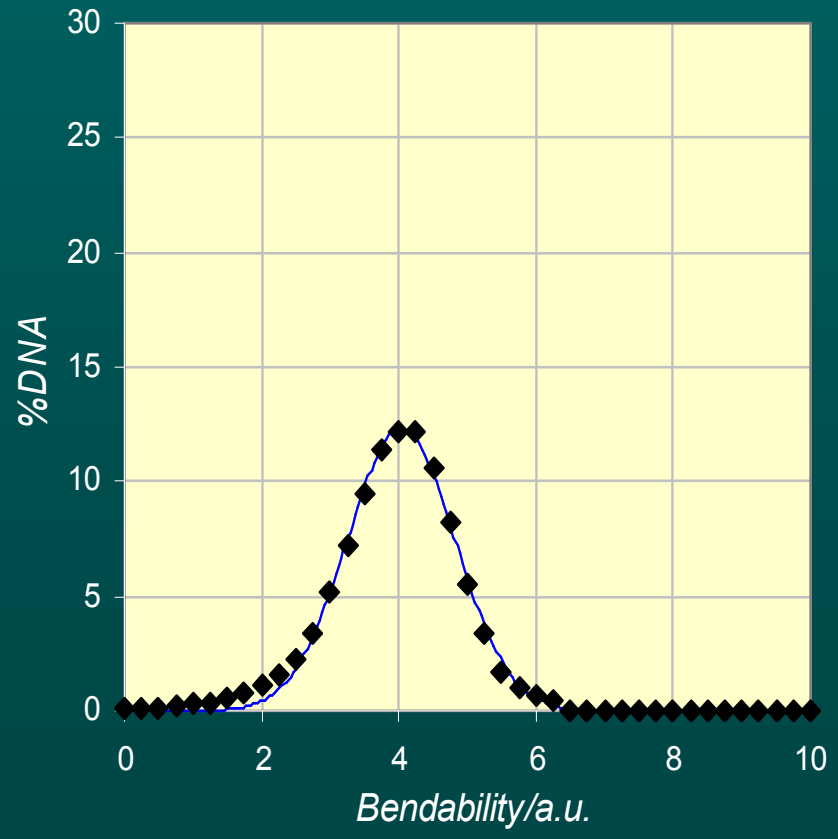
- Dinucleotide models
  - X-ray (Olson et al)
  - NMR (Ulyanov and James)
  - Electrophoresis (Bolshoy)
- Trinucleotide models
  - DNase I (Brukner et al)
  - Nucleosome (Travers et al)

Bendability distributions

- *M. tuberculosis*

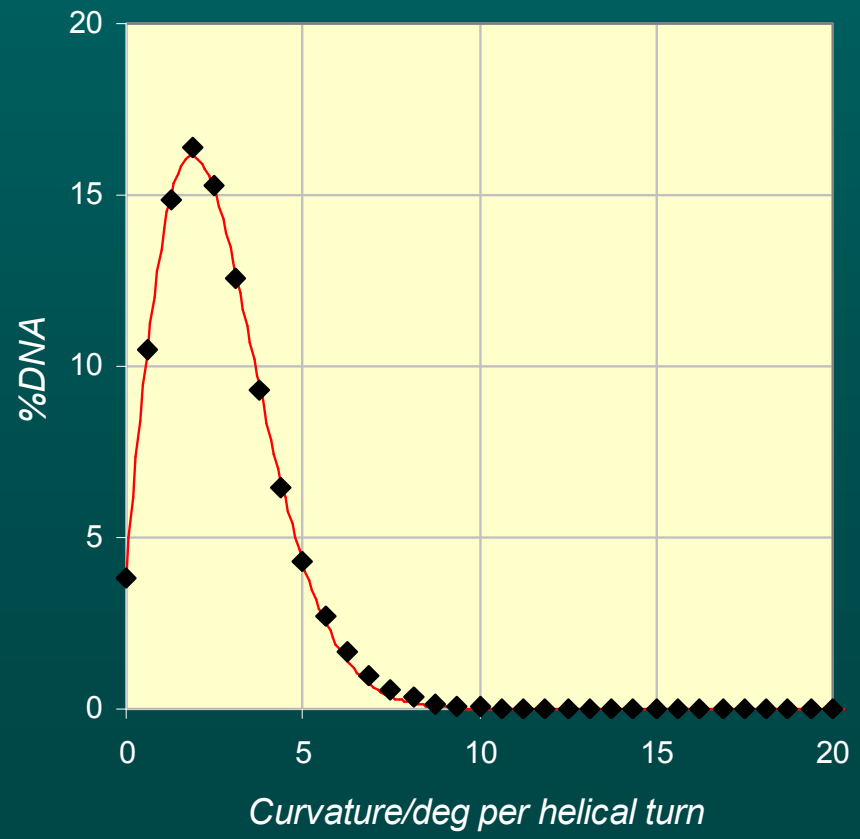


- *P. falciparum* chr II

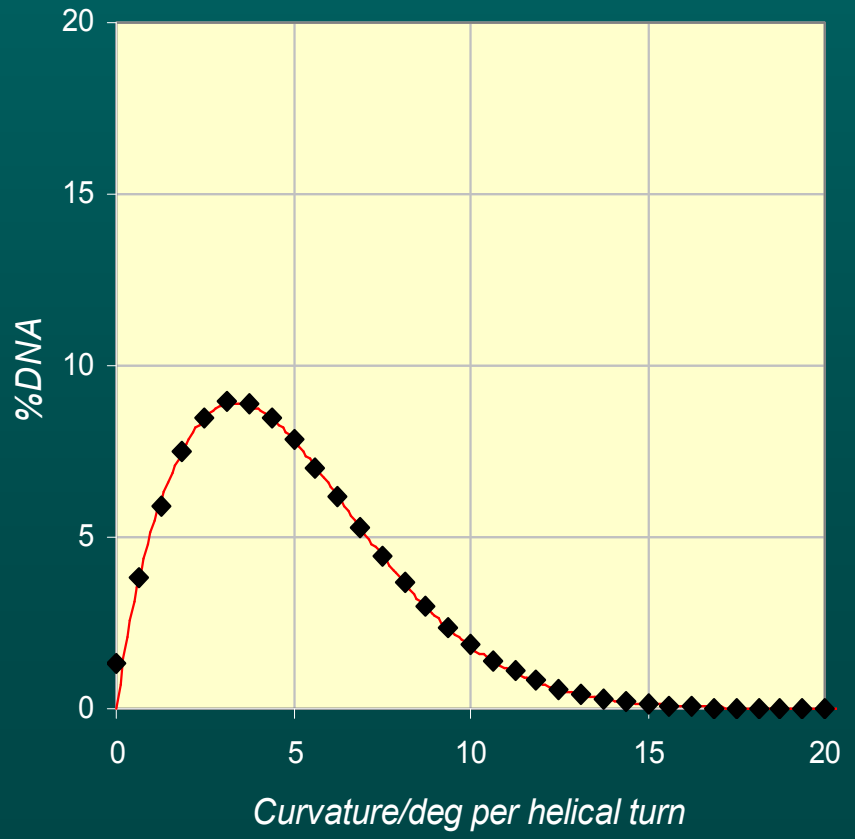


Curvature distributions

- *Aeropyrum pernix*

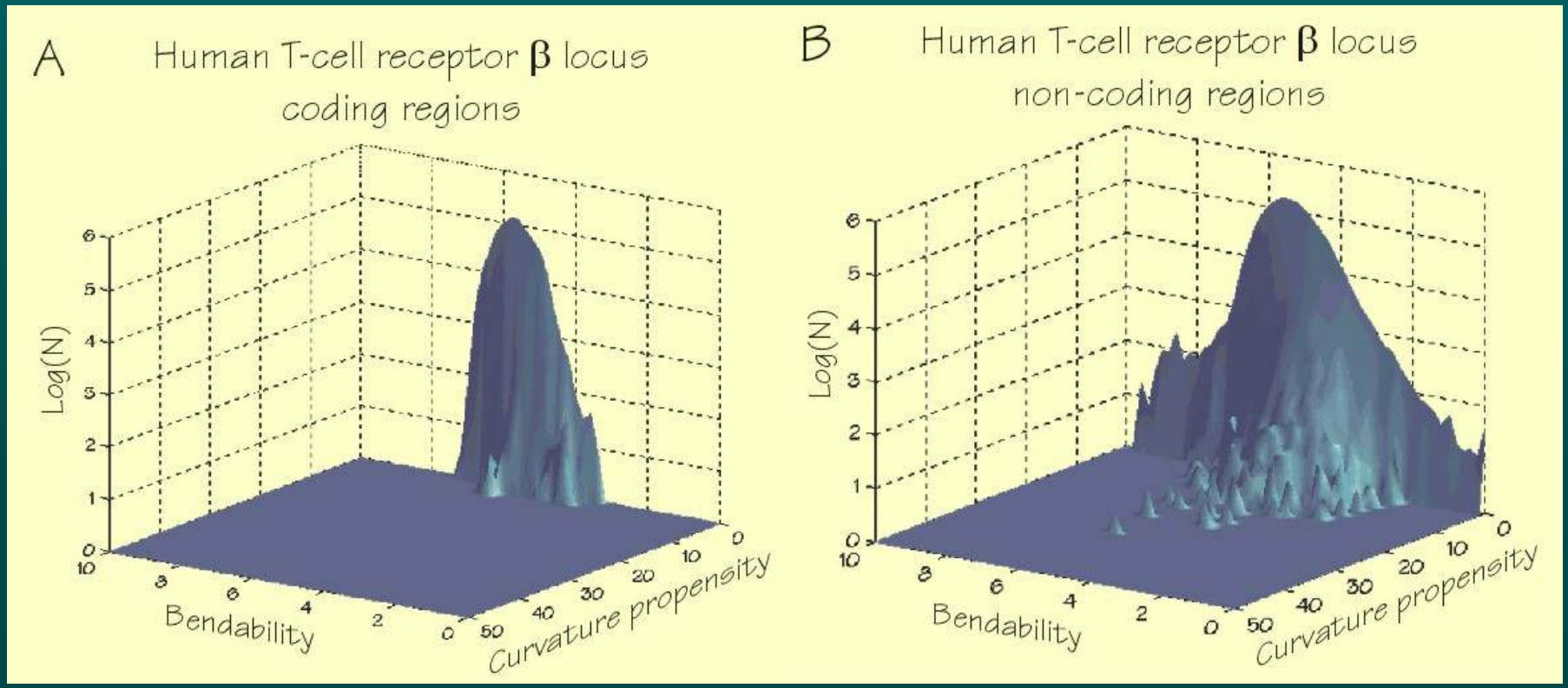


- *C. elegans* chromosome I



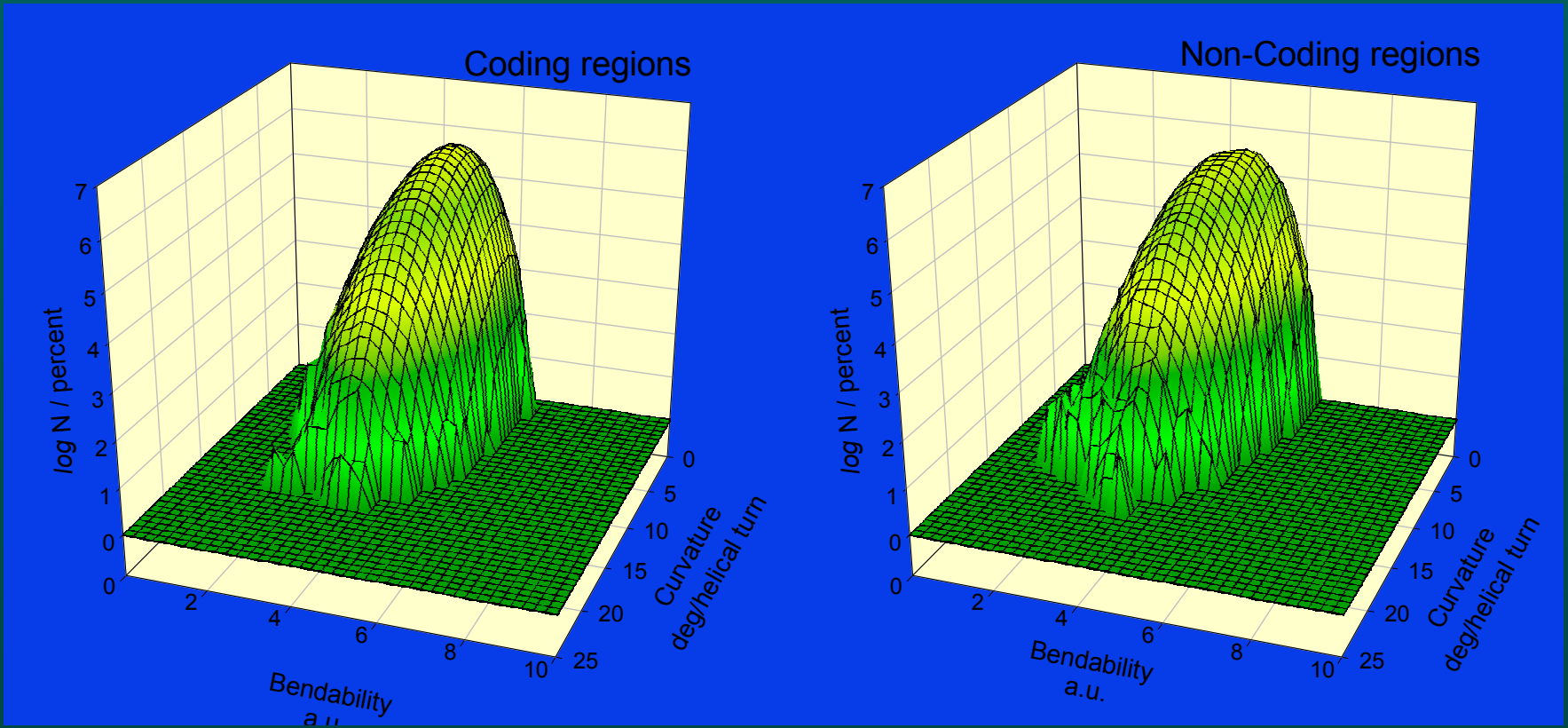
Coding vs. non-coding regions

- Human T-cell receptor locus



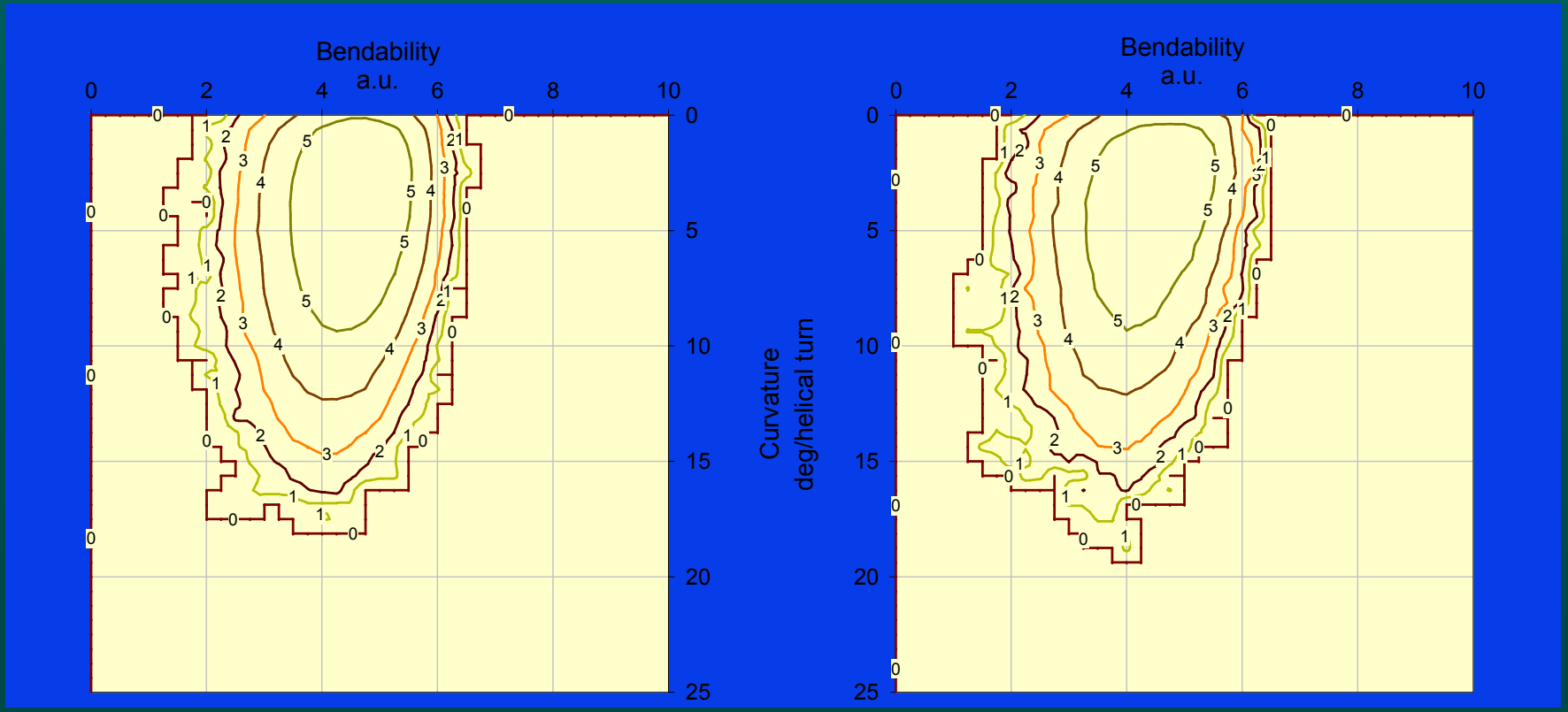
Coding vs. non-coding regions

- *H. influenzae*



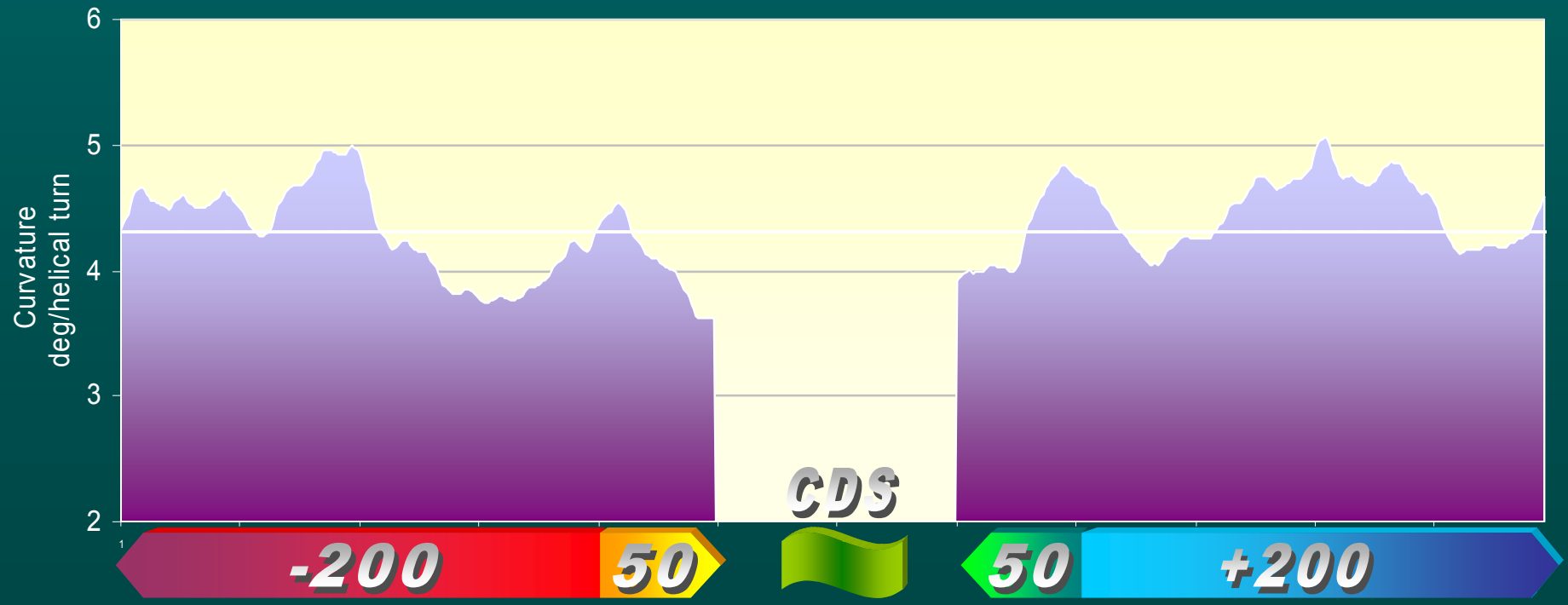
Coding vs. non-coding regions

- *H. influenzae*

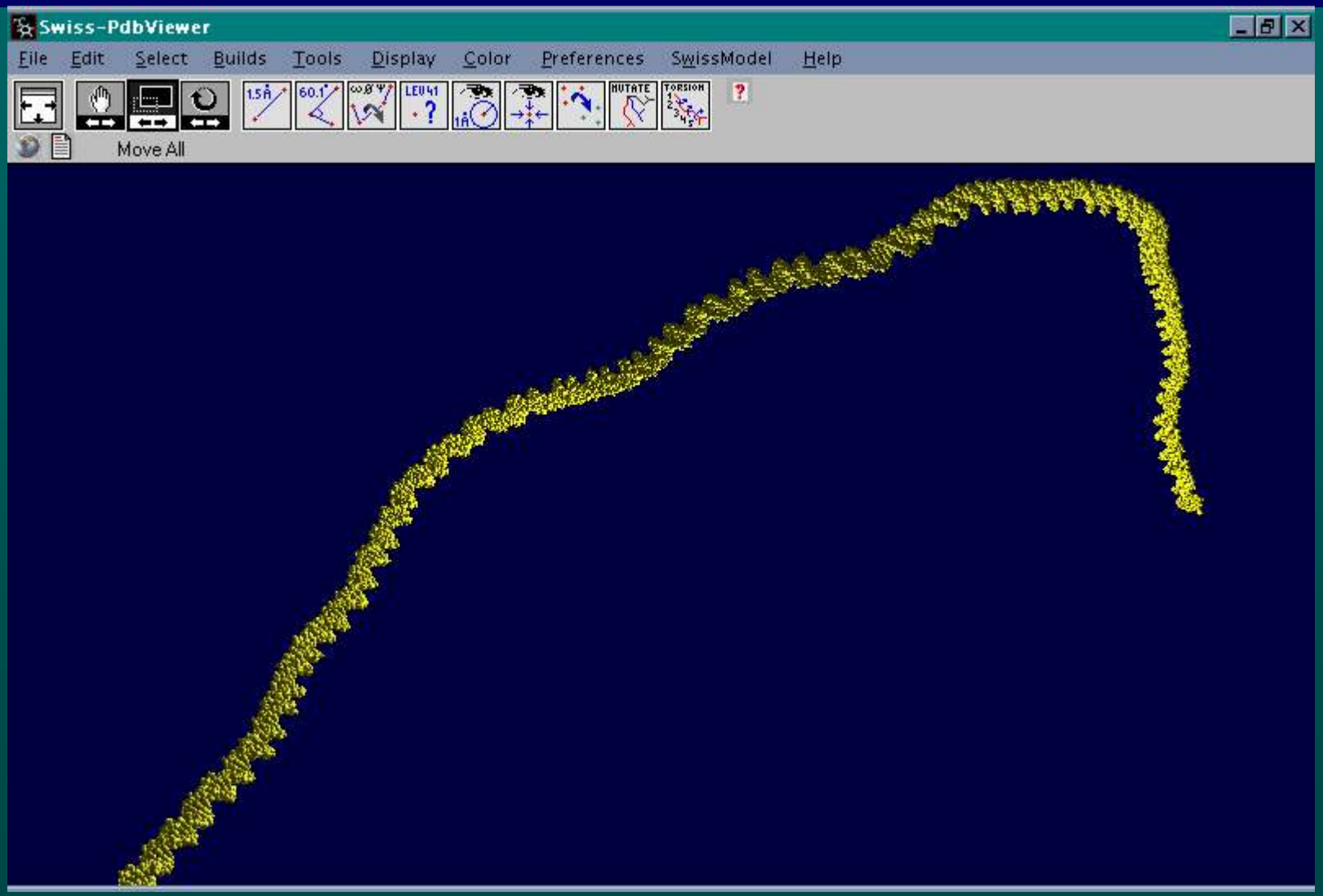


Positional preference for curvature

- Yeast chromosome I



model.it – molecular modelling

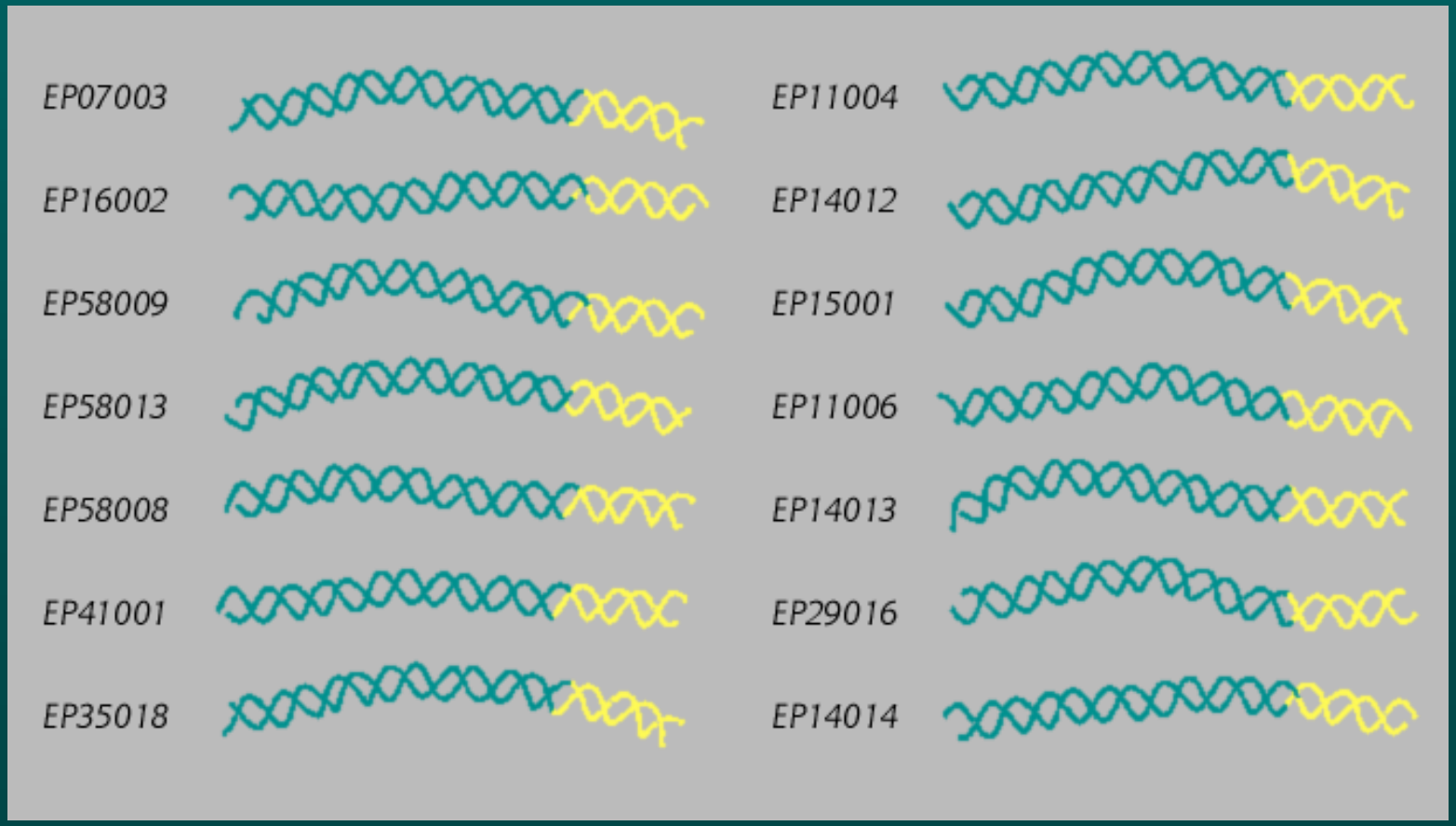


model.it - Methods

- A-DNA
- B-DNA
- Curved DNA
  - Dinucleotide parameters
    - X-ray, NMR, Electrophoresis
  - Trinucleotide parameters
    - DNase I, Nucleosome

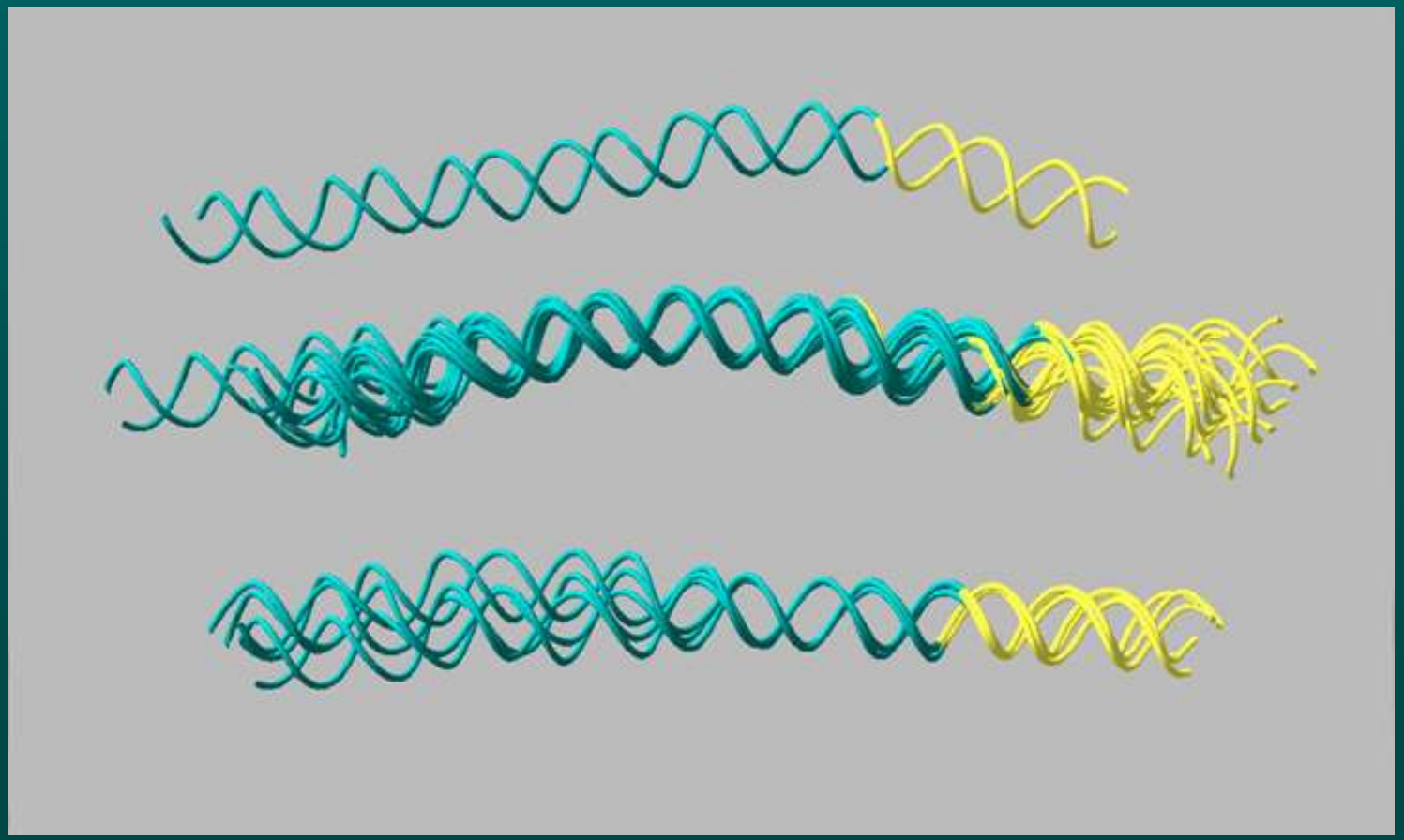
Molecular modeling

- *Zea mays* promoter regions



Molecular modeling

- *Zea mays* promoter regions



Future directions

- Experimentally determine biological importance of DNA curvature
- Improve prediction accuracy
- Extract maximum possible information from genomes

Thanks to

- SBASE: János Murvai, Kristian Vlahovicek, László Kaján, Mircea Pacurar
- PRIDE: Oliviero Carugo, Zoltán Gáspári,
- CX, DPX: Alessandro Pintar, Oliviero Carugo
- DNA-tools: Ivan Brukner, Andrei Gabrielian, Kristian Vlahovicek
- Server design: Mircea Pacurar, Kristian Vlahovicek