Introduction to Bioinformatics

Pázmány Péter Catholic University
Faculty of Information Technology
Fall Semester, 2016/17

Core operations II

Aggregation, projecctions

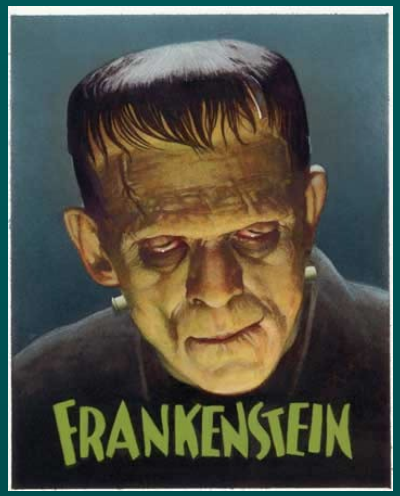
Sándor Pongor

This lecture
Aggregation, projections

- Aggregation
  - Numbers, vectors
  - Sequences
  - Assembly problems
  - Projections
- Projections (1D plots etc)
  - Amino acid property plots
  - Nucleic acid property plots

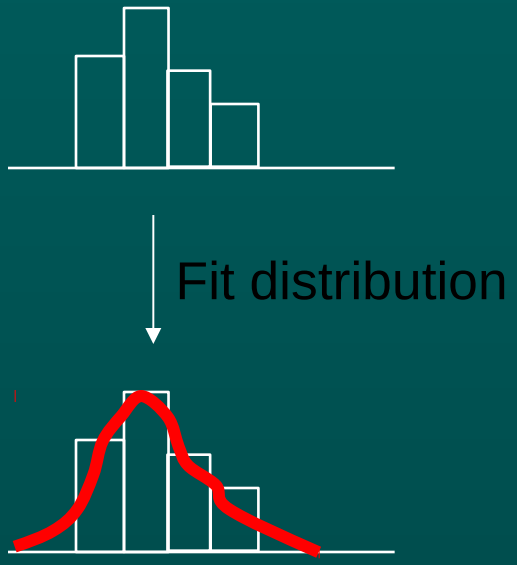
Core Operations 2

- Aggregation taSKS



Aggregation of numbers

- A Numerical aggregation operations
  - Sum
  - Average
  - Median
  - Minimum
  - Maximum
  - Stdev*…
- B Diagrammatic aggregation: Histograms and distributions

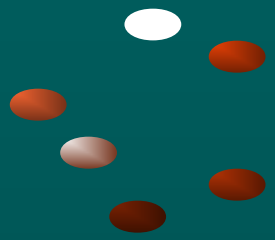


*When we calculate standard deviation, we automatically suppose that the distribution is normal (Gaussian, bell curve). This is often not true, but we still do this as a first approximation….

Aggregation of sequences

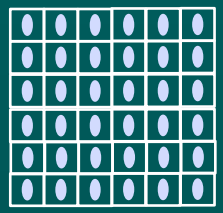
- 1) Aggregation by links
  - By similarity/distance links → distance matrix, heat map
  - → similarity group
  - → cladogram, tree
  - By functional (context) links → pathways
- Aggregation by common motif
  - Multiple alignment
  - Consensus descriptions of multiple alignment

Aggregation by links



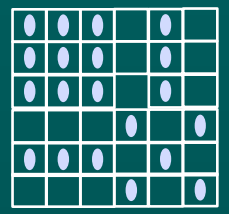
Arbitrary group of sequences

Calculate similarity scores

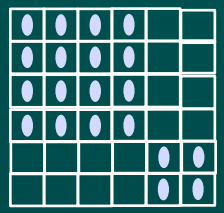


Distance matrix (similarity matrix)

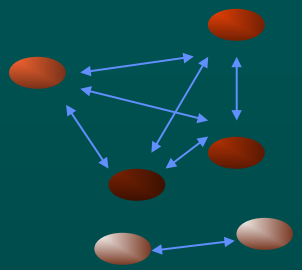
Delete scores below threshold



Rearrange (place most similar items near)

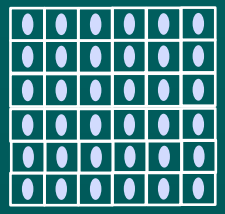


Heat map (coloured by scores)



Similarity group or neighborhood

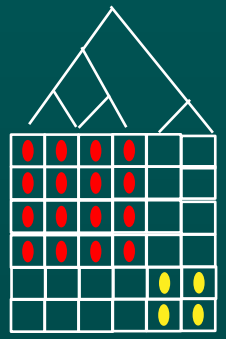
Aggregation by links 2



Distance matrix
(similarity matrix)

Phylogenetics

Tree or cladogram

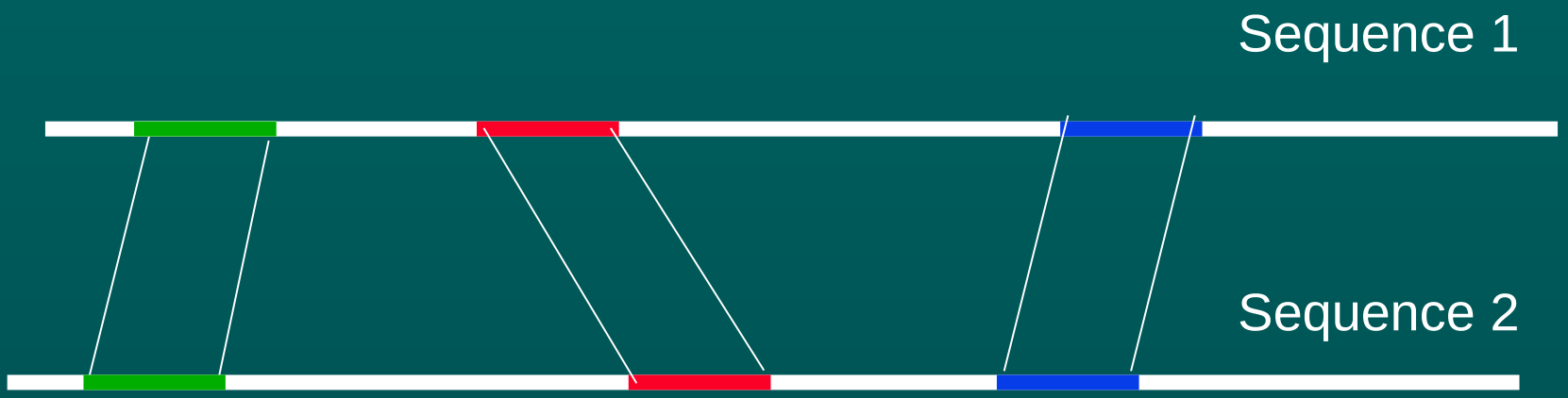


Heat map + tree

Agregation in analysis and identification...

- Biological objects are large and complex (genomes, proteomes, metagenomes, pathway data, etc.)
- Often, measuring instruments can only collect data on small pieces (next generation sequencing reads, peptide spectra in proteomics)
- Computational analysis of small fragments is accurate.
- There is one general trick: We divide a complex object into simple parts (like characteristic motifs), identify individual parts by simple numerical means, and then AGGREGATE the results.
- Not elegant, but works, even with very complex problems (forensic fingerprints)

Aggregating local sequence similarities



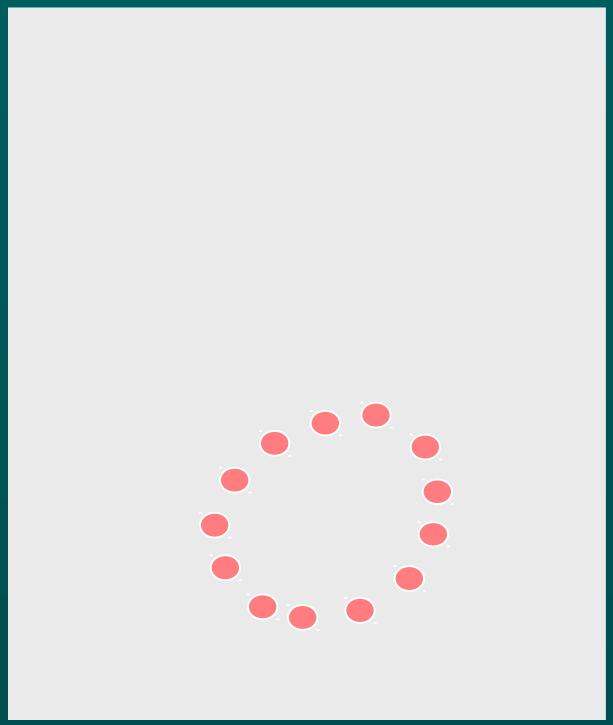
- Are these two sequences related by evolution? (are they homologous?) Only probabilistic answers...
- We need aggregate scores, i.e. probabilities for finding combinations by chance...

BLAST

Examples for aggregation in bioinformatics

- Single proteins, genes: constructing protein/gene similarity from local similarities (BLAST) Inferring homolgy.
- Proteomics: Constructing protein similarities from peptide fragment similarities. Inferring protein presence.
- Genomics1: Aggregating a long sequence from short reads (next generation sequencing). Inferring a genome.
- Genomics2: Putting protein similarities together into pathways.
- Metagenomics: Inferring a microbial community from species similarities.

The human mind is good at aggregating noisy signals according to structures



Psychology of vision.

- Contour recognition principles
- In bioinformatics, computers do this *in an abstract space of data,* and *without human intuition*.
- → Filtering, search space reduction is useful when designing bioinformatics tools.

Core operation 3

Projections (1D plots)

We assign a number to a position in a sequence…

What numbers do we plot

- A property of an amino acid/nucleotide. I.e. a value stored in a lookup-table.
- A value calculated from the sequence or from the associated 3D structure (a „window”)
- A value determined by experiment: The sequencing quality of the position. Number of reads “hitting” a position

Simple programs 1

Hydrophobicity Plot

Prediction of hydrophobic and hydrophilic regions in a protein

Hydrophobicity/Hydrophilicity Values

	Fauchere & Pliska	Kyte & Doolittle	Hopp & Woods	Eisenberg
R	-1.37	-4.50	3.00	-2.53
K	-1.35	-3.90	3.00	-1.50
D	-1.05	-3.50	3.00	-0.90
Q	-0.78	-3.50	0.20	-0.85
N	-0.85	-3.50	0.20	-0.78
E	-0.87	-3.50	3.00	-0.74
H	-0.40	-3.20	-0.50	-0.40
S	-0.18	-0.80	0.30	-0.18
T	-0.05	-0.70	-0.40	-0.05
P	0.12	-1.60	0.00	0.12
Y	0.26	-1.30	-2.30	0.26
C	0.29	2.50	-1.00	0.29
G	0.48	-0.40	0.00	0.48
A	0.62	1.80	-0.50	0.62
M	0.64	1.90	-1.30	0.64
W	0.81	-0.90	-3.40	0.81
L	1.06	3.80	-1.80	1.06
V	1.08	4.20	-1.50	1.08
F	1.19	2.80	-2.50	1.19
I	1.38	4.50	-1.80	1.38

Hydrophobicity Plot

- Sum amino acid hydrophobicity values in a given window
- Plot the value in the middle of the window
- Shift the window one position

$$\langle H_i \rangle = \frac{1}{2k+1} \sum_{n=i-k}^{i+k} H_n$$

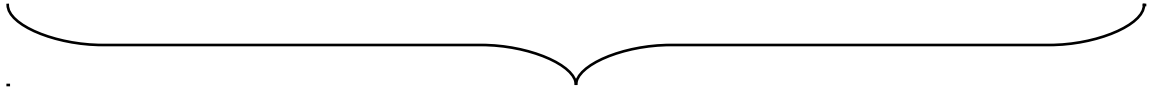
Large H → hydrophobic, e.g. membrane bound segments

Sliding Window Approach

- Calculate property for first sub-sequence
- Use the result (plot/print/store)

I L I K E I R

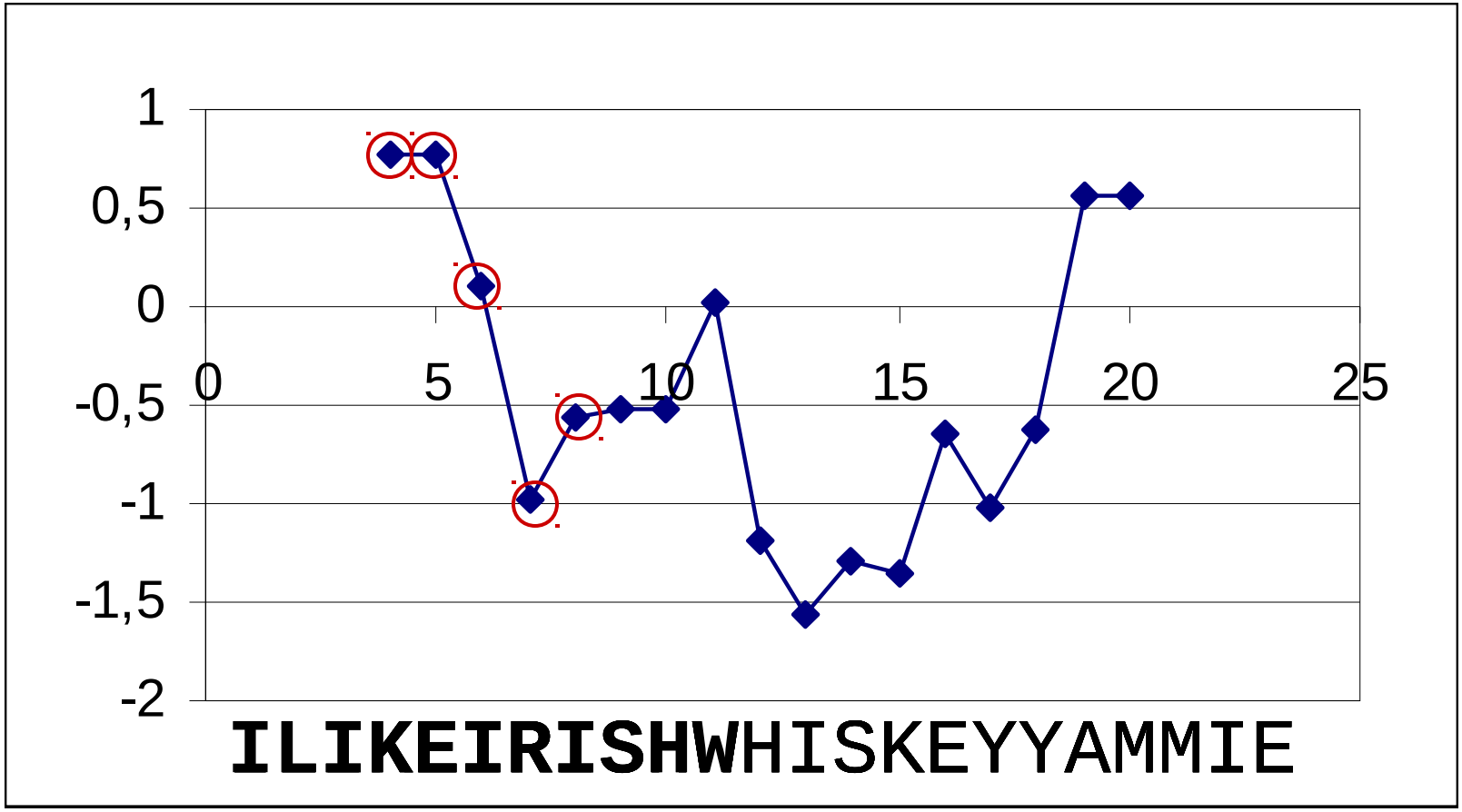
4.50 + 3.80 + 4.50 - 3.90 - 3.50 + 4.50 - 4.50



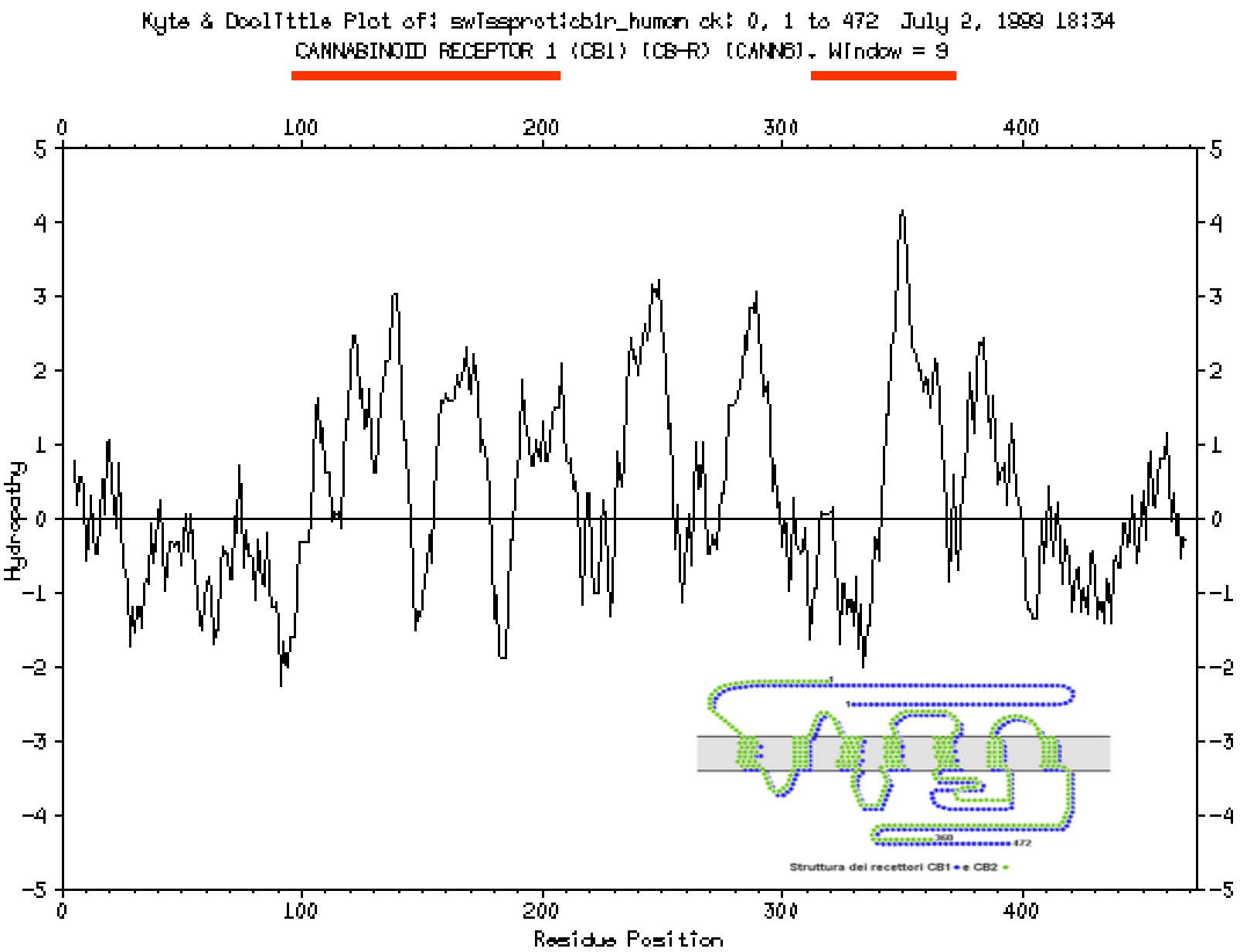
5.4 / 7 = 0.77

- Move to the next position in the sequence

Hydrophobicity Plot

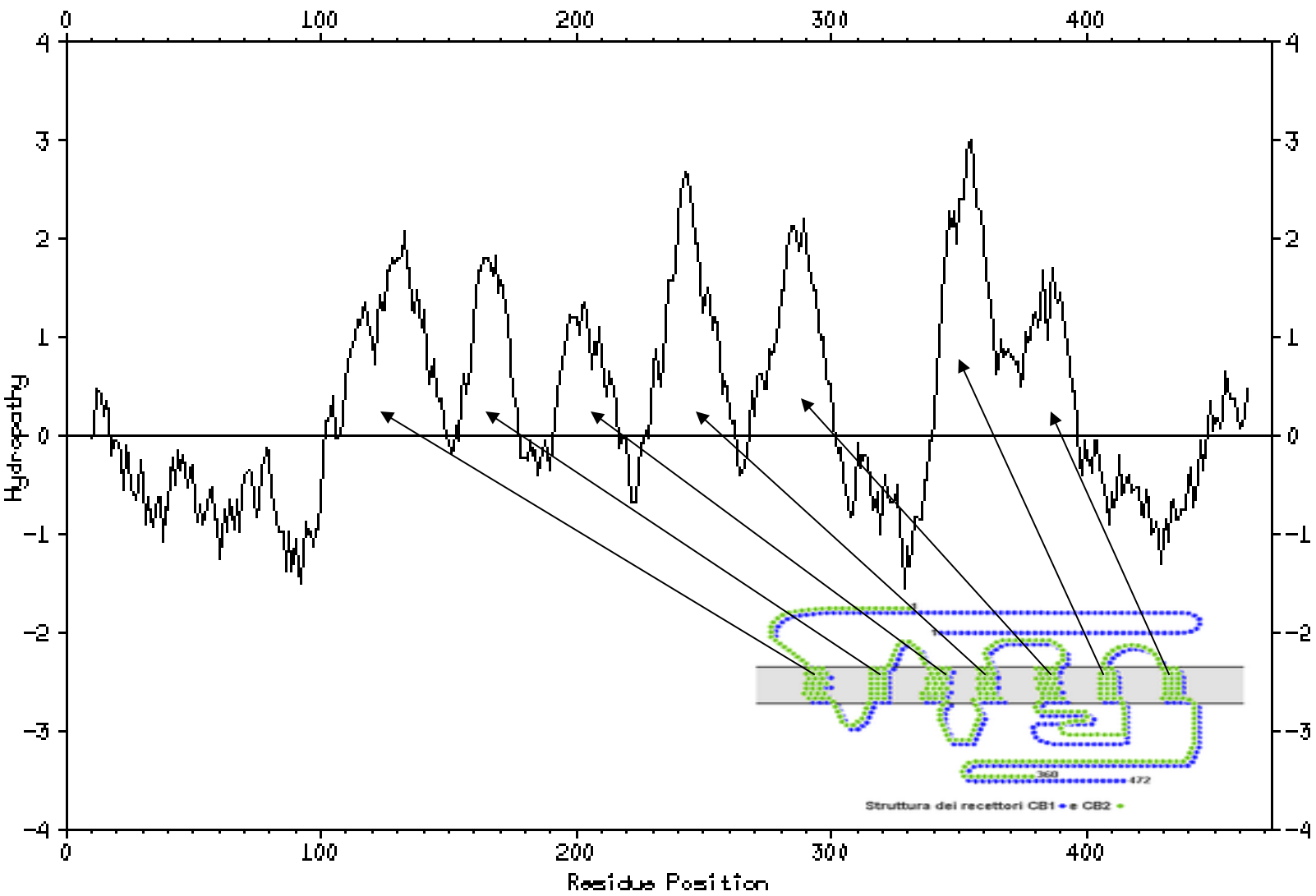


Large H → hydrophobic, e.g. membrane bound segments


Kyte & Doolittle Plot of: swissprot:cb1r_human ck: 0, 1 to 472 July 2, 1999 18:34
CANNABINOID RECEPTOR 1 (CB1) (CB-R) (CANN6). Window = 9
Hydropathy
Residue Position
0
100
200
300
400
5
4
3
2
1
0
-1
-2
-3
-4
-5
Struttura dei recettori CB1 e CB2

Kyte & Doolittle Plot of: swissprot:cb1r_human ck: 0, 1 to 472 July 2, 1999 18:32

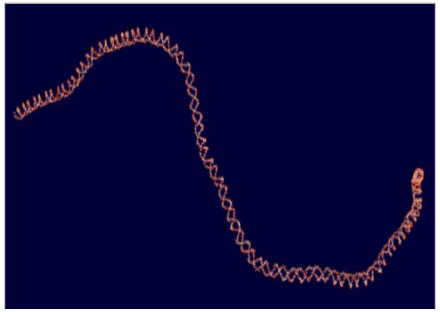
CANNABINOID RECEPTOR 1 (CB1) (CB-R) (CANN6), WIndow = 19



Simple programs 2

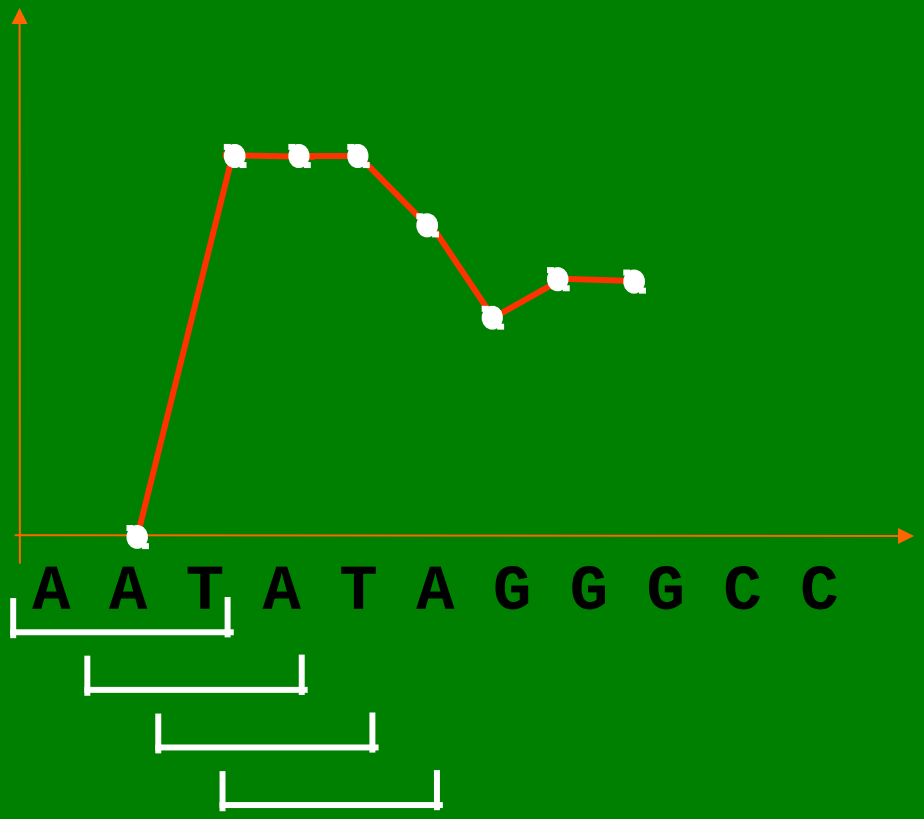
DNA bending plot

Prediction bent regions in DNA

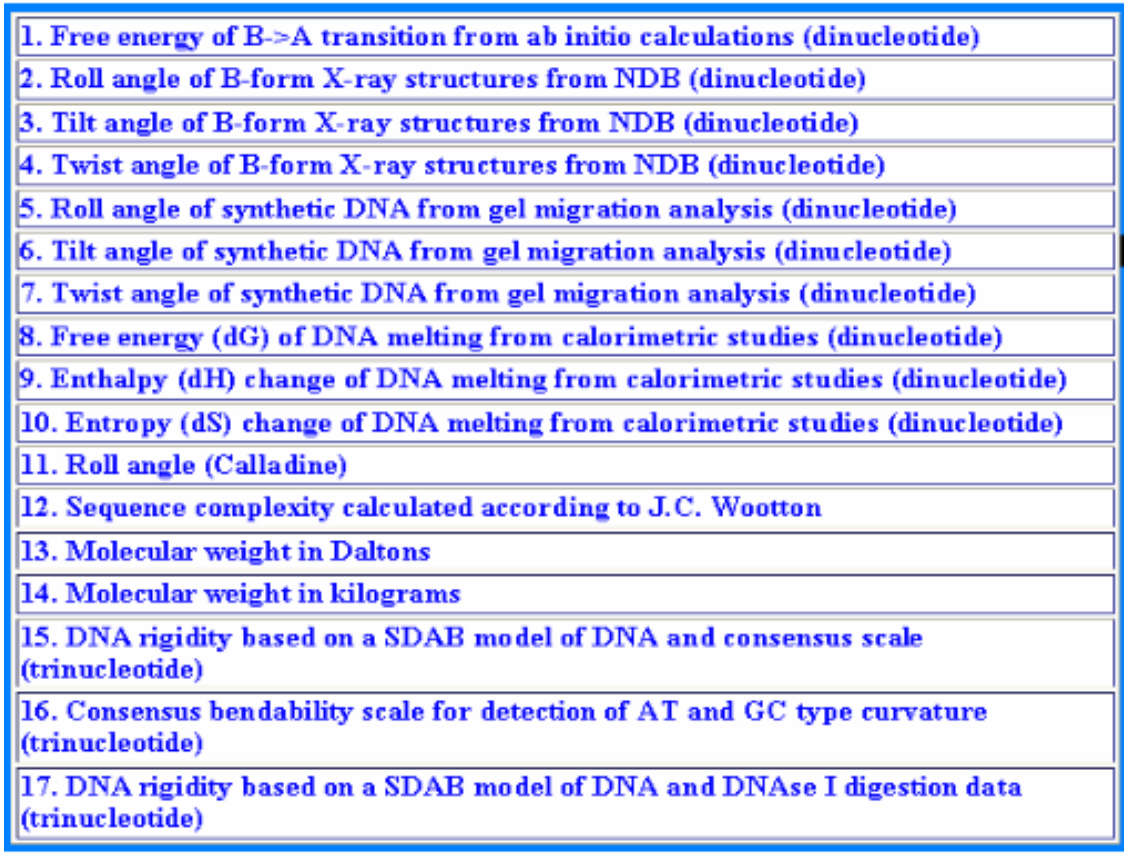


DNA bending plot

Trinucleotide	bending parameter (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



Many DNA property plots are possible

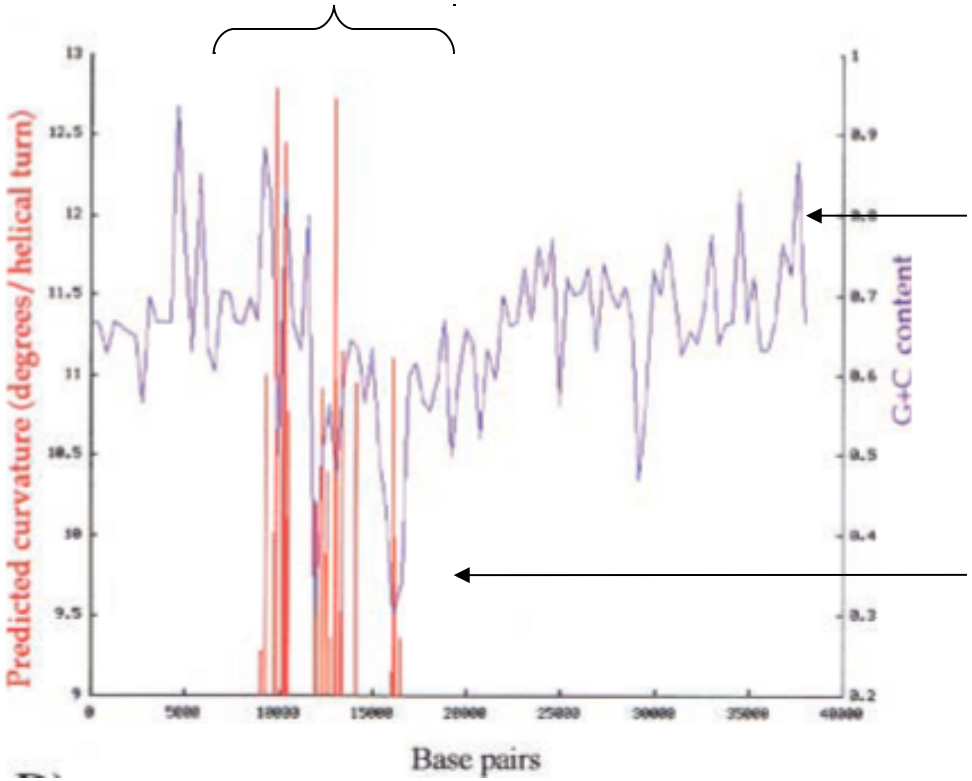


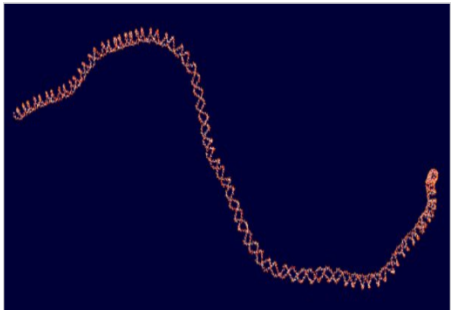
The goal: visual identification of “potentially interesting” regions in large DNA sequences

Prediction of bent regions in DNA

Bent region!

A strange chromo-somal region in Leishmania – analyzed with a www server





http://pongor.itk.ppke.hu/?q=bioinfoservices

What you should know

- Aggregation of numbers and vectors
- Aggregation of sequences by functional and similarity links
- Aggregation by similarity into distance matrices, heat maps, trees

- Projection: Numerical annotation of sequences, window sliding
- Projection: Hydrophobicity plots
- Projection: DNA bending plots