CRASP: software package for analysis of physicochemical parameters of aligned sequences of protein families

D.A. Afonnikov,

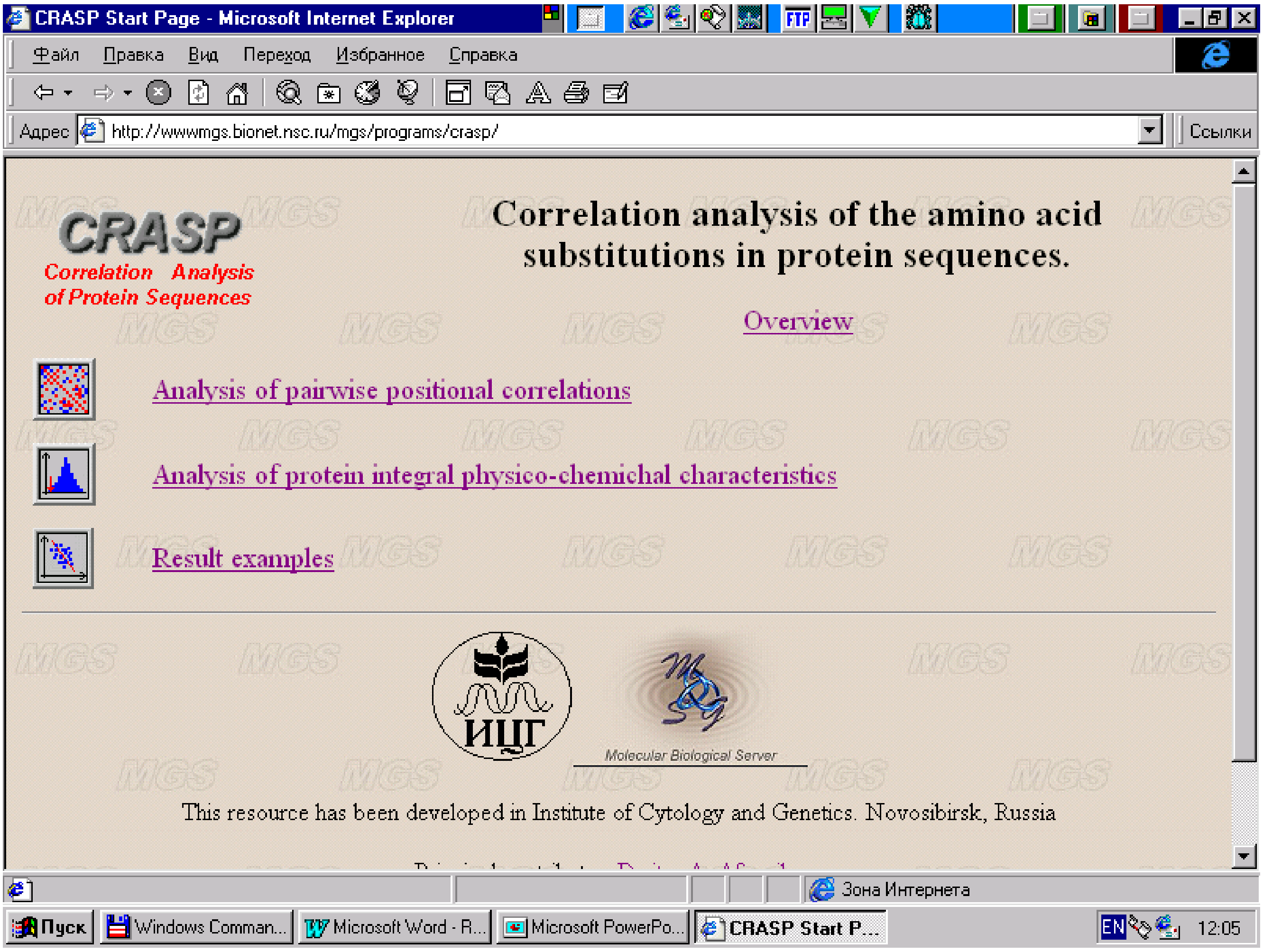
D.Yu. Oschepkov,

N.A. Kolchanov

Institute of Cytology and
Genetics SB RAS,
Novosibirsk, Russia

Main page of the package for correlation analysis of amino acid substitutions in protein sequences (*CRASP*).

Program is available at *http://wwwmgs.bionet.nsc.ru/mgs/programs/crasp/.*



CRASP Start Page - Microsoft Internet Explorer
Файл
Правка
Вид
Переход
Избранное
Справка
Адрес
http://wwwmgs.bionet.nsc.ru/mgs/programs/crasp/
Ссылки
CRASP
Correlation Analysis of Protein Sequences
ИЦГ
Molecular Biological Server
Зона Интернета
Пуск
Windows Comman...
Microsoft Word - R...
Microsoft PowerPo...
CRASP Start P...
EN
12:05

Correlation analysis of the amino acid substitutions in protein sequences.

Overview

Analysis of pairwise positional correlations

Analysis of protein integral physico-chemichal characteristics

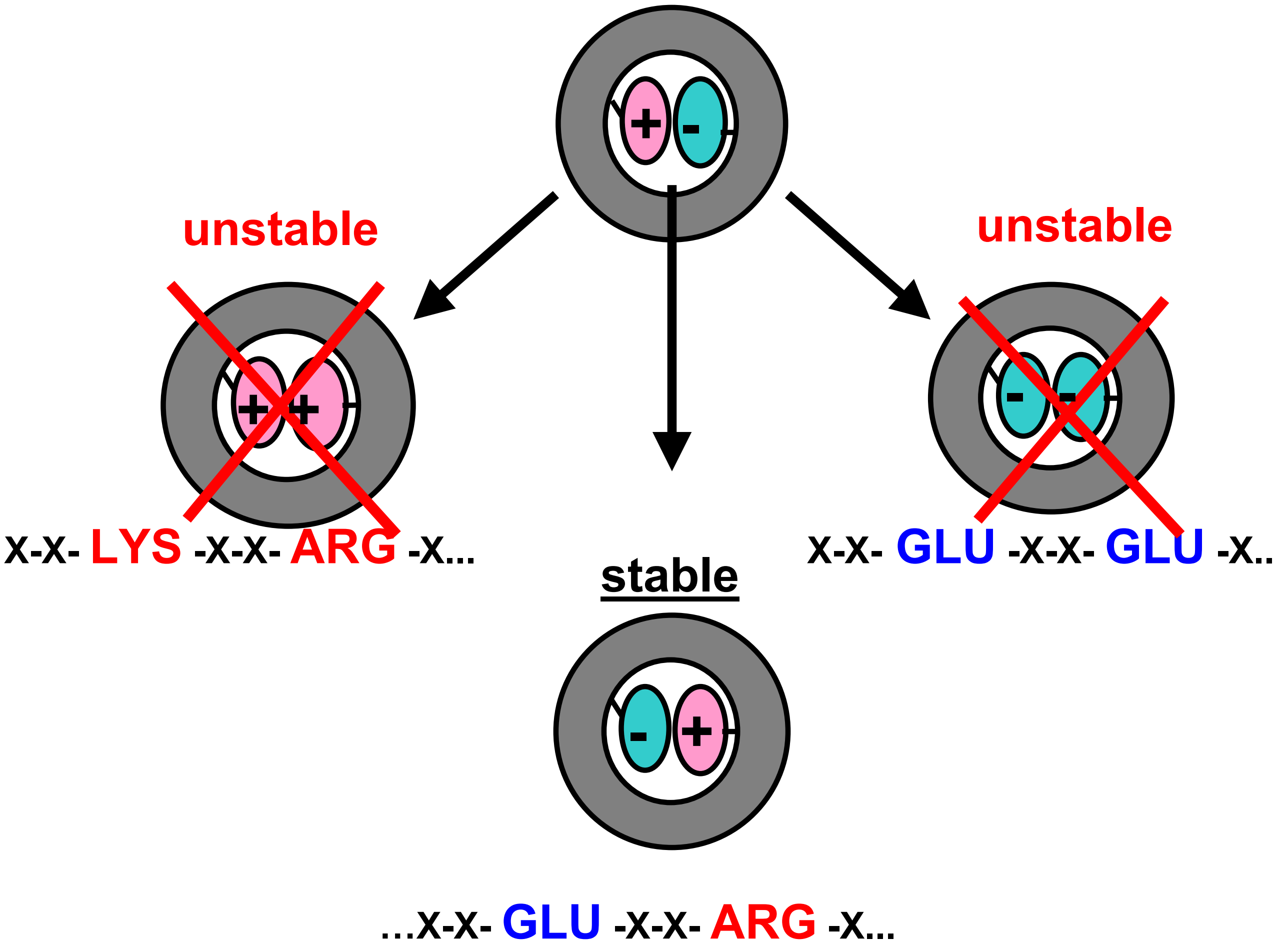
Result examples

This resource has been developed in Institute of Cytology and Genetics. Novosibirsk, Russia

AN EXAMPLE OF COMPENSATORY (relatively to the charge sign) AMINO ACID SUBSTITUTIONS AT A PAIR OF PROTEIN POSITIONS

```
LIVKSMDGAL      Analysis of multiple
STMECARLIT      alignment of
GTSDNSHQLI      sequences of
LIMKVVDGYA      a protein family
```

..X-X-X- LYS -X-X- ASP -X-X-X..



unstable

X-X- LYS -X-X- ARG -X...

unstable

X-X- GLU -X-X- GLU -X..

stable

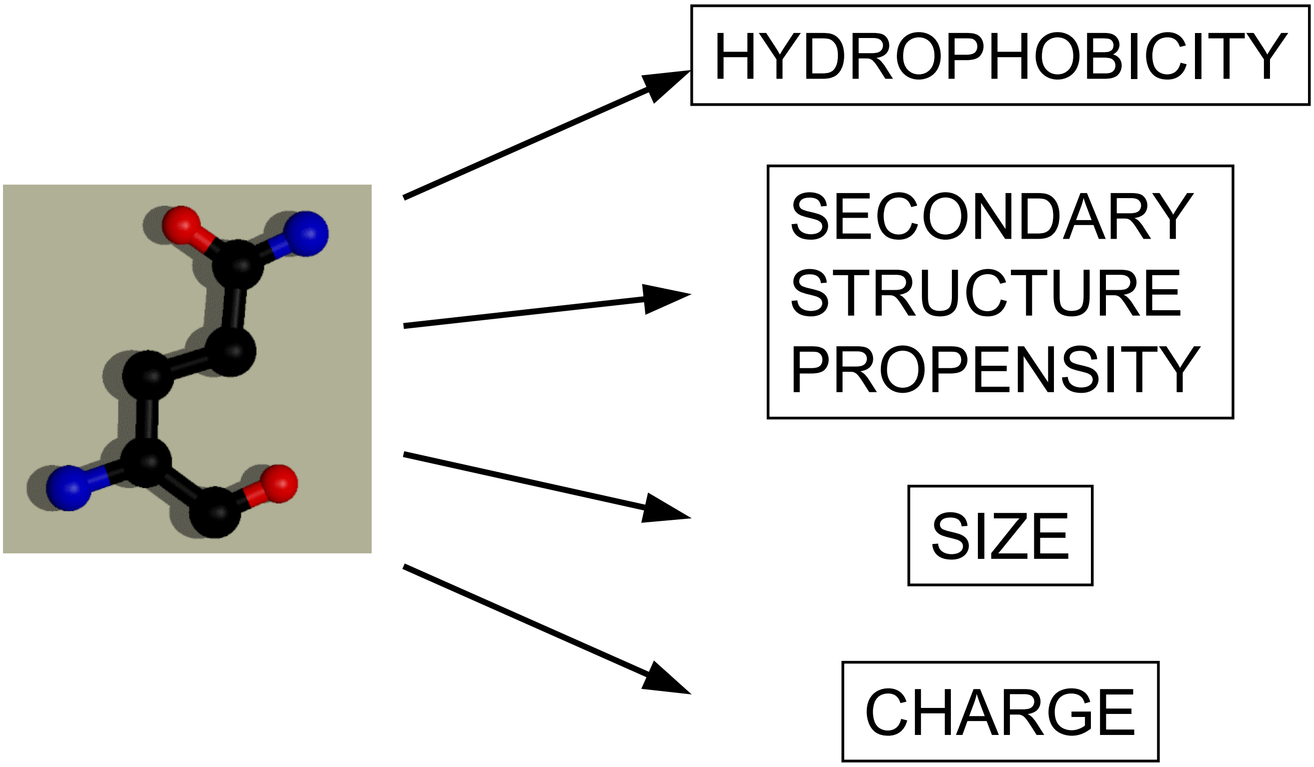
...X-X- GLU -X-X- ARG -X...

Main goals of the CRASP package analysis:

- detection of protein position pairs with co-adaptive residue substitutions;
- detection of protein integral characteristics which conservation (variability) is due to co-adaptive residue substitutions.

The importance of physico-chemical characteristics analysis.

The values of characteristics reflect specific interactions of residues:



HYDROPHOBICITY

SECONDARY STRUCTURE PROPENSITY

SIZE

CHARGE

THE SCHEME OF CORRELATION ANALYSIS IN FAMILY OF RELATED PROTEINS (CRASP PACKAGE)

Set of aligned protein sequences

```
ASDRTFNHKLIYRRMNL-
GAAQWCCV-RHNGTYKLS
DFGSSR---WWRQACHMP
DFGSSR---WWRQACHMP
```

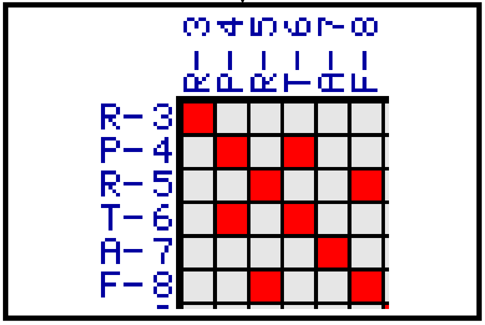
Selection of physico-chemical characteristic

```
+4.0;+6.2;-1.8;
+5.9;+3.2;+5.6;
-3.4;-2.1;+4.0;
-6.9;+2.3;+5.0;
```

Estimation of pairwise correlation coefficients between residue substitutions

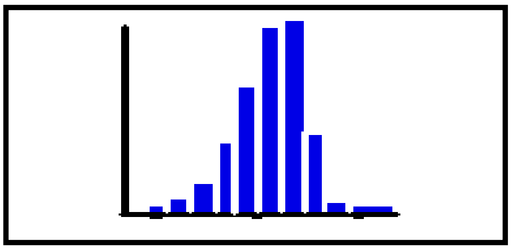
$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$$

Analysis of pairwise correlation coefficients, inferring conserved physico-chemical characteristic.



	R-3	P-4	R-5	T-6	A-7	F-8
R-3	🟥					
P-4		🟥		🟥		
R-5			🟥			🟥
T-6		🟥		🟥		
A-7					🟥	
F-8			🟥			🟥

Integral physico-chemical characteristic analysis

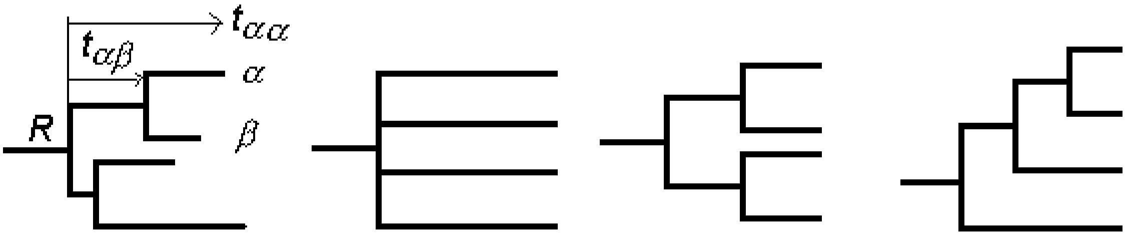


THREE PROBLEMS

- •The problem of evolutionary relationship of sequences
- •The problem of chained correlation
- •The problem of stability of correlation coefficient estimates

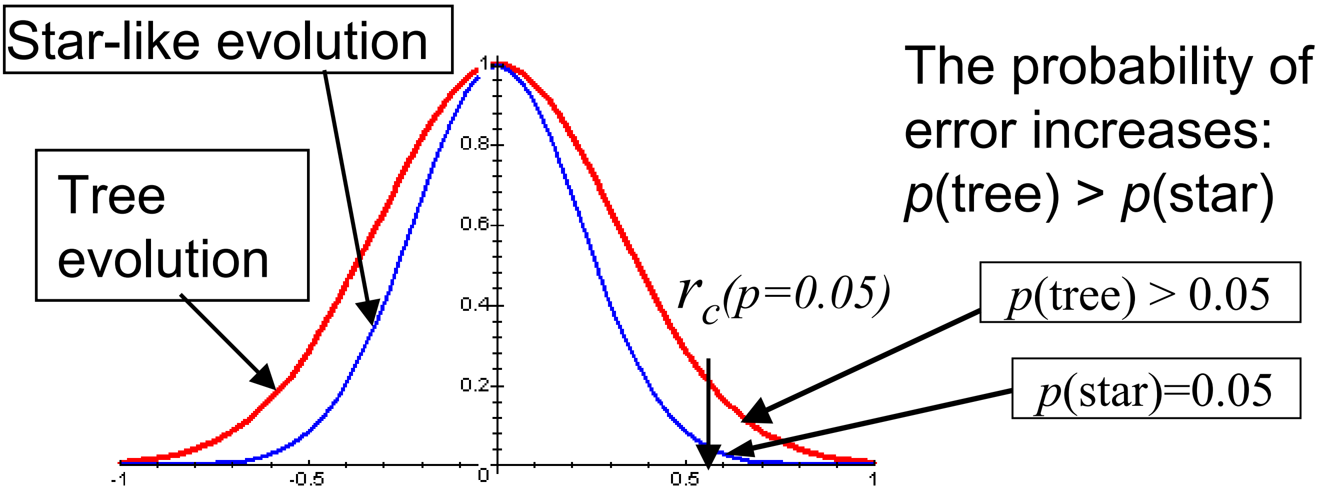
1. TAKING TO ACCOUNT EVOLUTIONARY DEPENDENCE OF ANALYSED SEQUENCES

Evolutionary dependencies viewed as phylogenetic trees:



Random Star-like Balanced Imbalanced

Distribution of correlation coefficient for independent positions: $D(r,\text{star})=1/(\text{N-1}) < D(r,\text{tree})$



Star-like evolution

Tree evolution

$r_c(p=0.05)$

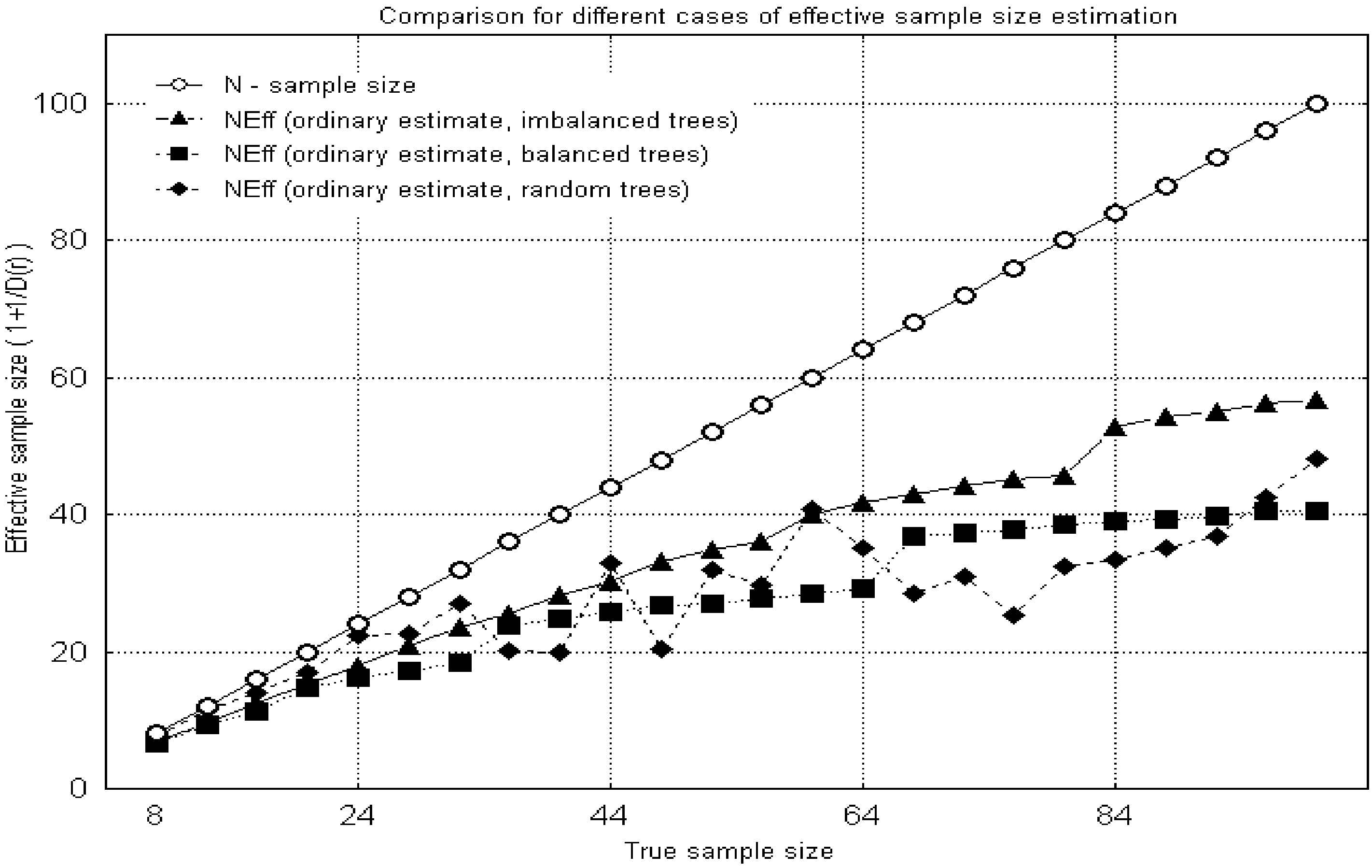
$p(\text{tree}) > 0.05$

$p(\text{star})=0.05$

The probability of error increases: *p*(tree) > *p*(star)

Testing: numerical simulation of evolution with independent sites. Sequence length=500. Sample size and tree topology were varied. For each topology and each sample size 1000 samples were generated.

Estimated parameter: N_{eff}=1+1/D(r). For independent sequences N_{eff}=N, for evolutionary dependent sequences $N_{eff} < N$



Wrong estimation of critical threshold for correlation coefficient (t_p – percentile of Student's distribution):

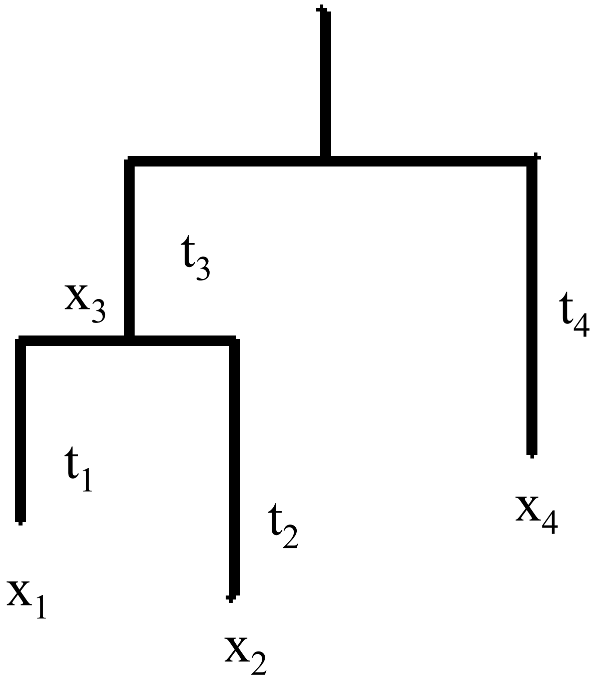
$$|r_c| = \sqrt{\frac{t_P}{t_P + N - 2}}$$

Two possible solutions:

-Numerical simulation to estimate true threshold (time consuming)

-Weighting sequences

Applied method: weighting according to Felsenstein J. (1985) *Am. Nat.*, **125**, 1-15.



1. Estimate values of parameter x at internal nodes of the tree (contrasts) on the basis of values of x at leaf nodes and Gaussian model of distribution.

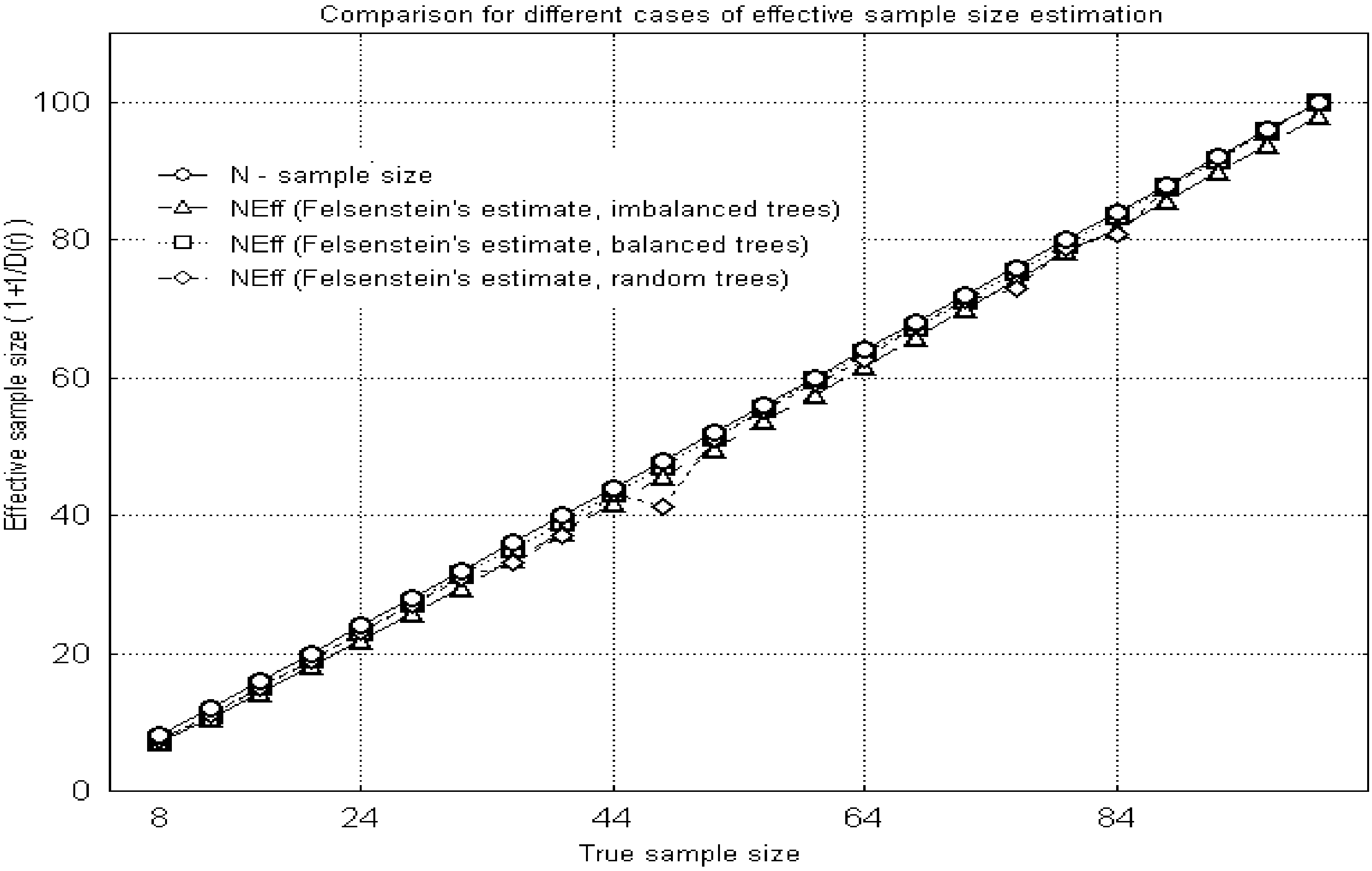
$$x_3 = (x_1 \cdot t_2) + x_2 \cdot t_1)/(t_1 + t_2)$$

$$t_3' = t_3 + t_1 t_2 /(t_1 + t_2)$$

$$D(t) \sim t$$

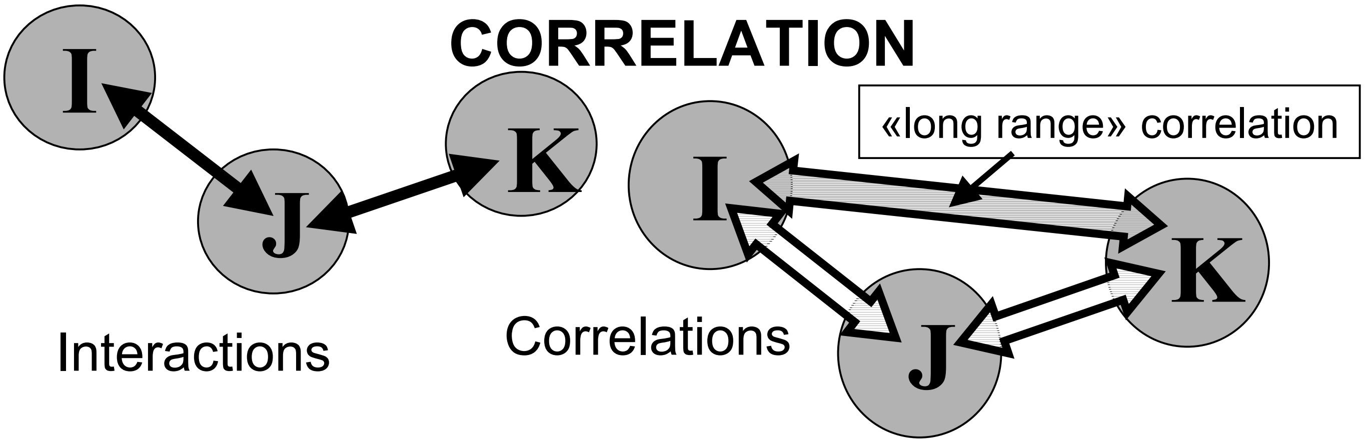
2. Estimate means, variances and correlation coefficients and dispersion for contrasts.

Testing: numerical simulation of evolution with independent sites. Apply weighting estimates for correlation coefficients.



Result: *D*(r,tree) ~ *D*(r,star). It is possible to select threshold r_c as for independent sequences. Weighting allows to choose r_c value the same as for independent sequences

2. THE PROBLEM OF CHAINED CORRELATION

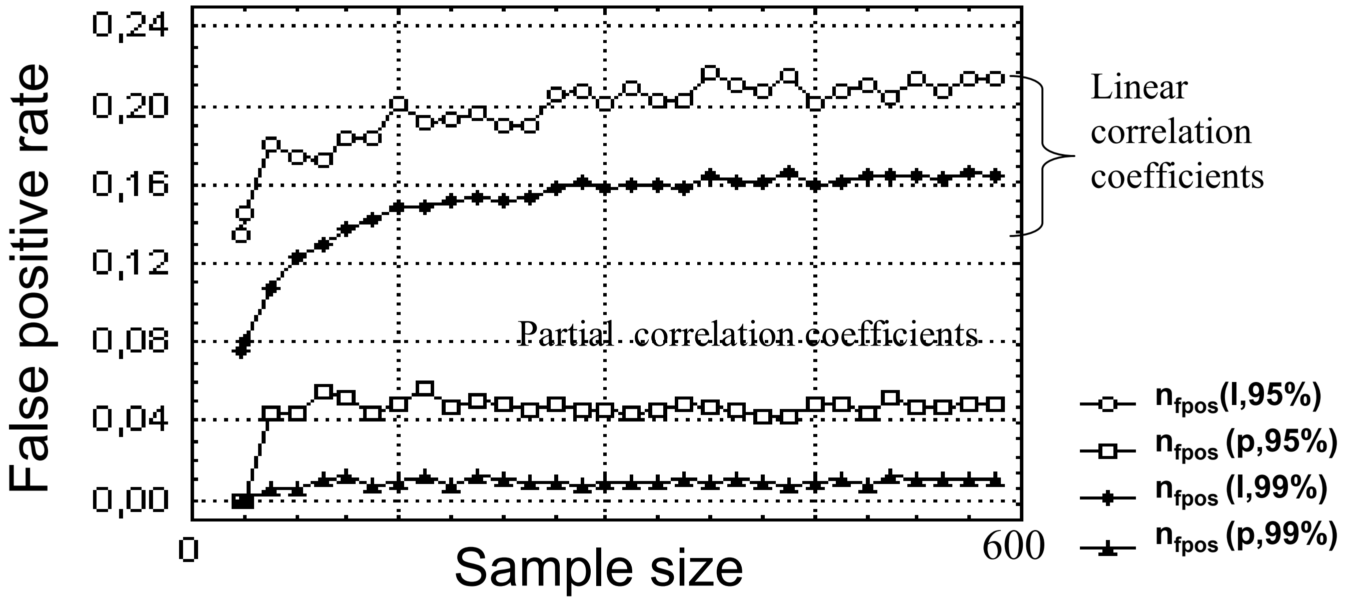


Possible solution: partial correlation coefficients

$$r_{ij\cdot k} = \frac{-a_{ij}}{\sqrt{a_{ii}a_{jj}}}, A = S^{-1}$$

Testing: numerical simulation with harmonically interacting residues, sequences are independent, Metropolis algorithm.

Estimated parameter: fraction of false positives (pairs with no interaction, but significant, at 95 and 99% levels, correlation) n_{fpos}.

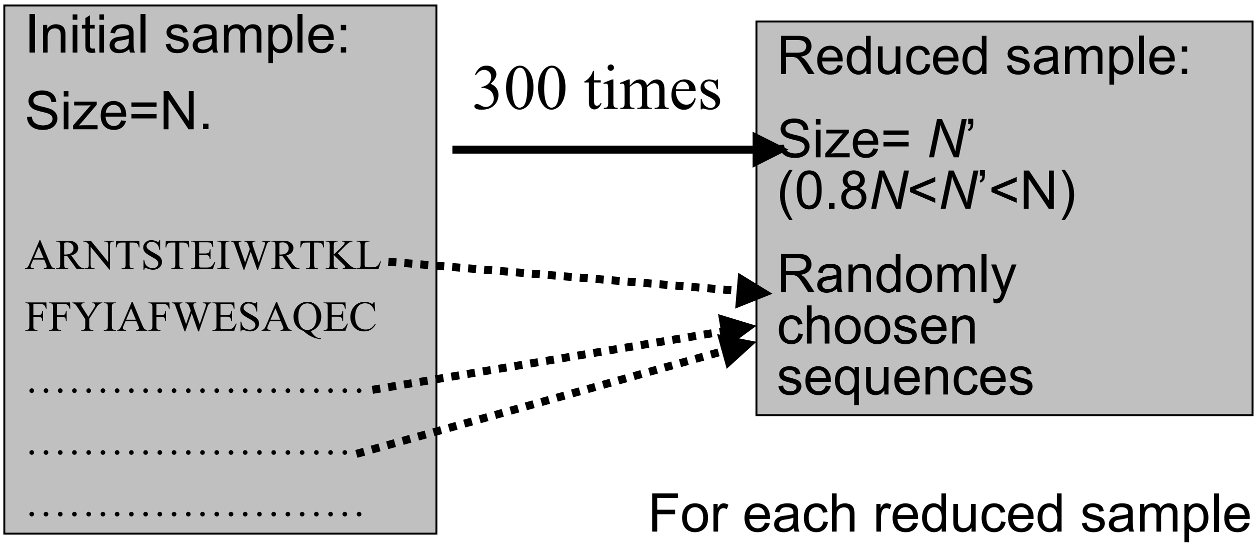


ESTIMATION THE STABILITY OF CORRELATION COEFFICIENTS

Resampling procedure (*N*' samples out of initial, 0.8*N*<*N*'<N)

Estimation the dispersion of ratio

$$rs = r_{ij} / \sqrt{1/(N'-1)}$$



Initial sample:

Size=N.

ARNTSTEIWRTKL

FFYIAFWESAQEC

....

....

....

300 times

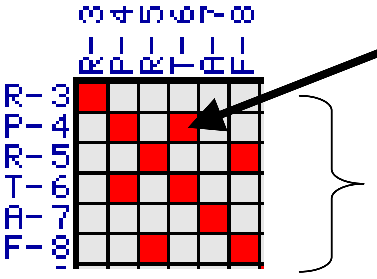
Reduced sample:

Size= *N*'

(0.8*N*<*N*'<N)

Randomly choosen sequences

For each reduced sample estimate r_{ij}



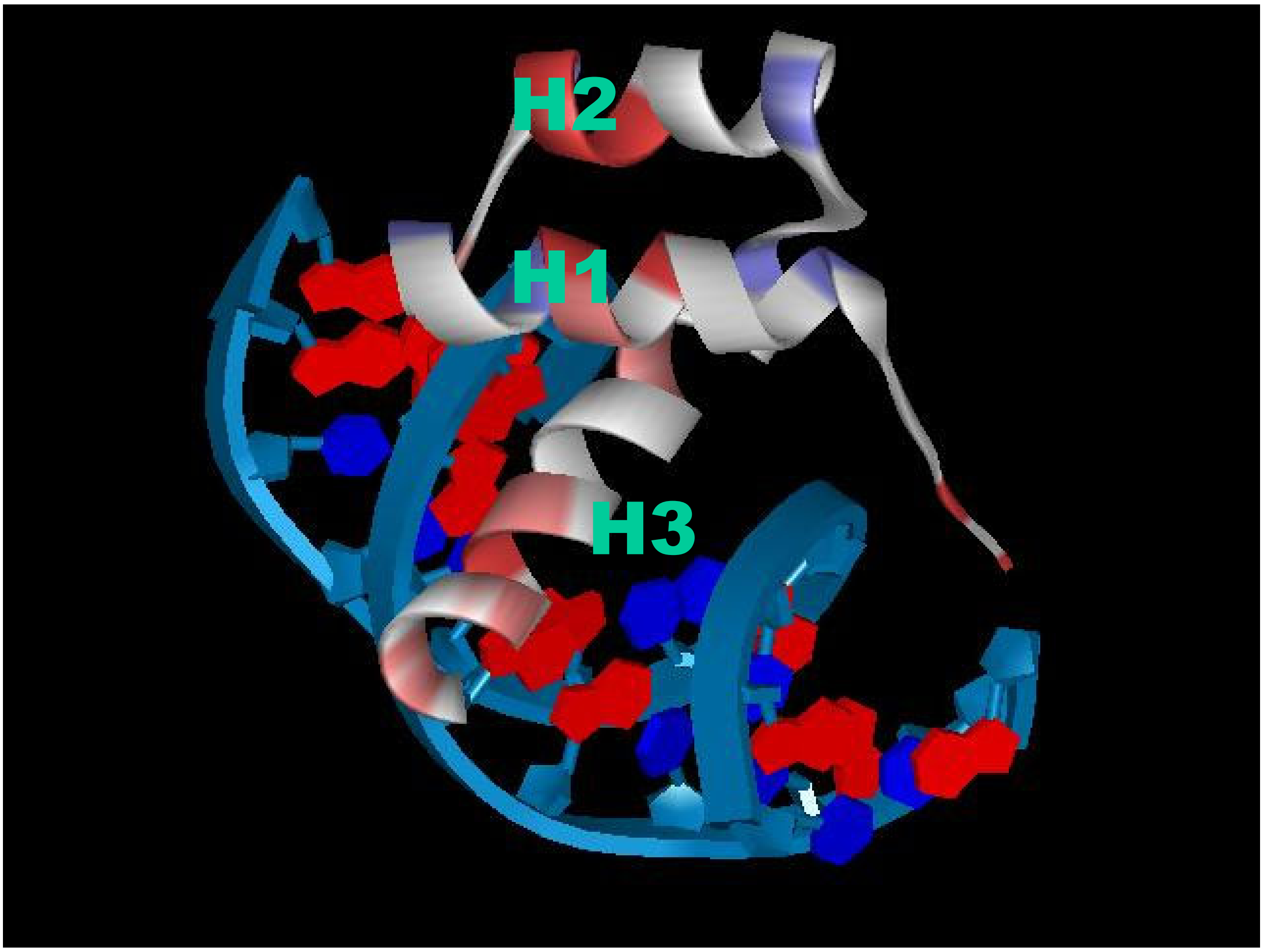
Estimate the dispersion of rs parameter using 300 r_{ij} values for each pair i,j.

5% pairs with highest dispersion of rs parameter considered as unstable and eliminated from analysis

APPLICATION FOR HOMEODOMAIN FAMILY ANALYSIS

372 sequences (source - Pfam), 47 positions.
Analysed characteristic - isoelectric point
Evolutionary tree estimated bu CLUSTALW program.

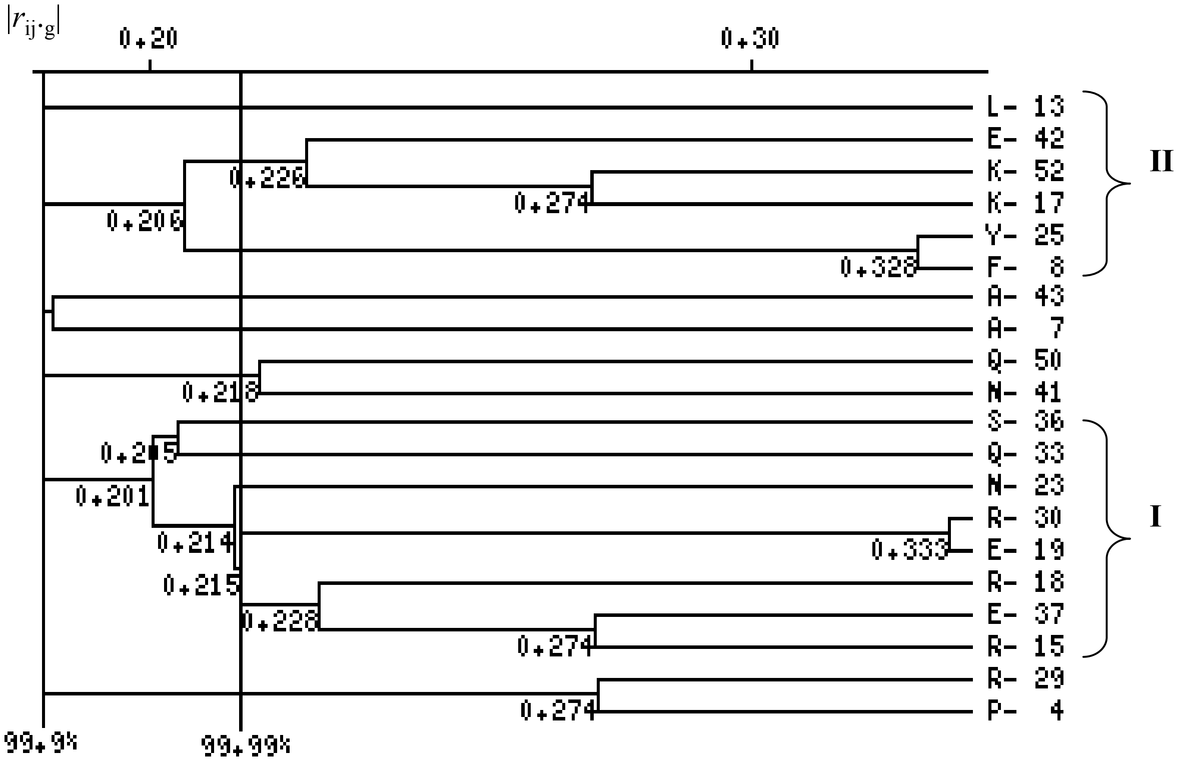
Spatial structure of homeodomain complex



CLUSTERING APPROACH TO DETECT GROUPS OF HIGHLY CORRELATED POSITIONS

The clustering of the sequence positions is performed with the distance measure dependent on the absolute value of correlation coefficient $|r_{ij}|$ between positions, both partial coefficients were used:

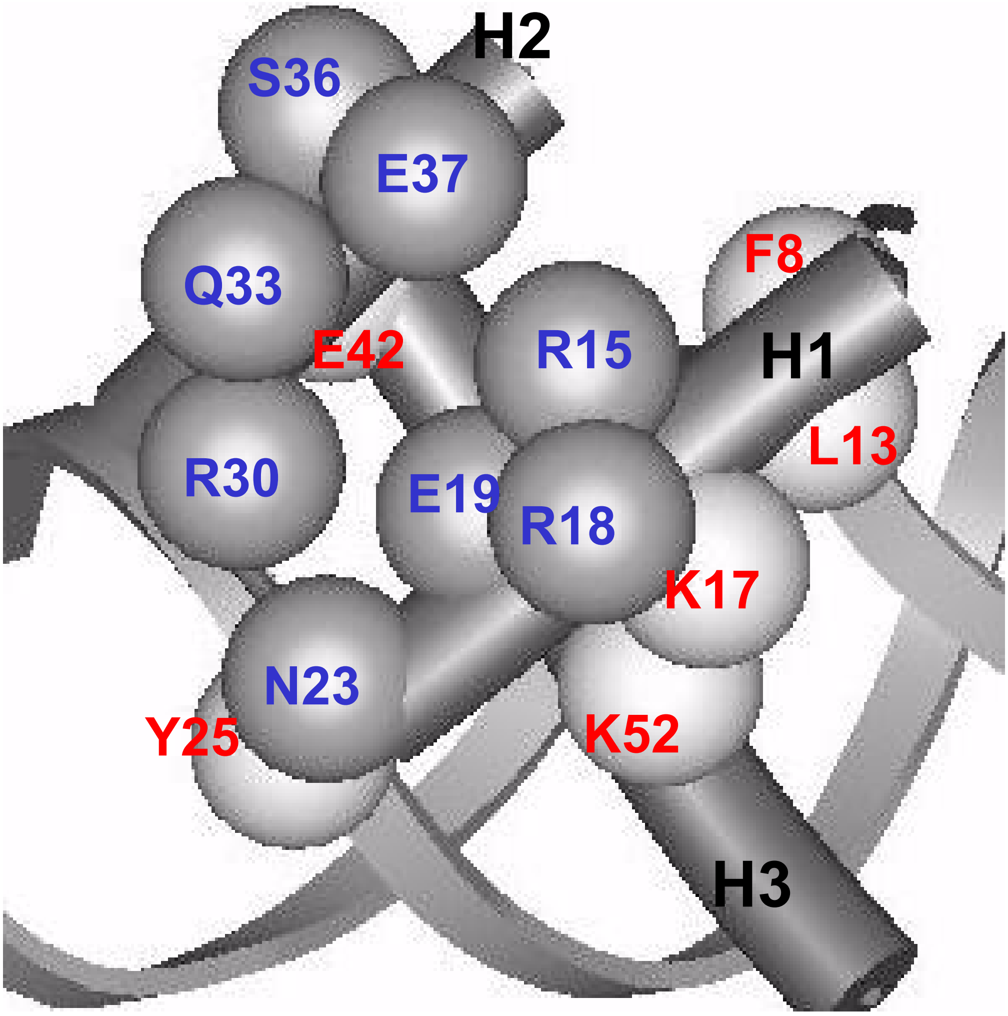
$$d_{ij} = 1 - |\, r_{ij} \,|$$



Two groups of positions have been determined group I and group II

SPATIAL LOCATION OF RESIDUES FROM GROUPS I AND II

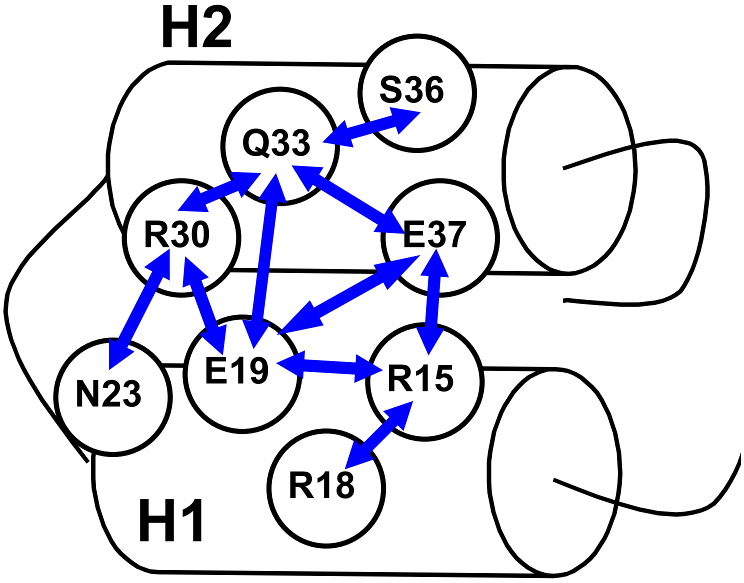
Cluster I residues are shown in dark grey and blue letters,
Cluster II residues are shown in light grey and red letters.



ANALYSIS OF POSITION FROM CLUSTER I

Position	$r_{ij \cdot g}$
R18-R15	-0.228
E19-R15	-0.215
N23-E19	-0.214
R30-E19	-0.333
R30-N23	-0.185
Q33-E19	-0.190
Q33-R30	-0.201
S36-Q33	-0.205
E37-R15	-0.274
E37-E19	-0.185
E37-Q33	-0.194

The values of significant correlation coefficients and schematic representation of relationships between residues in group I



Proposed conserved characteristic: net isoelectric point value (net charge)

$Q_I = pI_{15} + pI_{18} + pI_{19} + pI_{23} + pI_{30} + pI_{37} + pI_{33}$

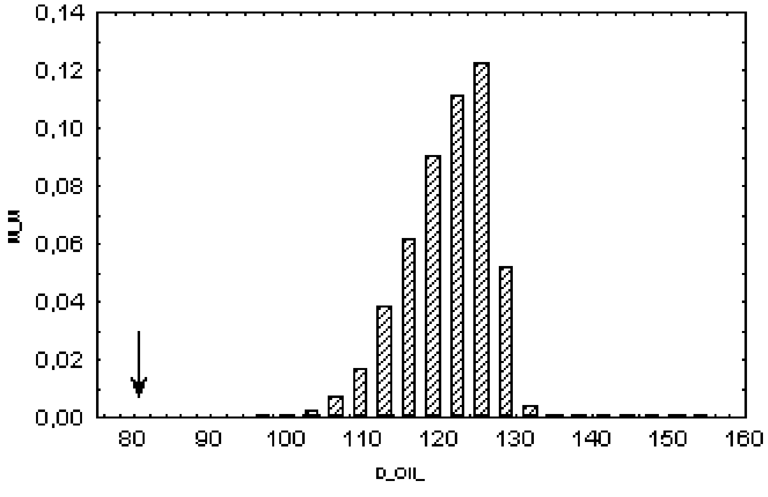
ANALYSIS OF Q_I CONSTANCY

Expected variance (in absense of correlations, c_i=1, $D(f_i)$)-positional dispersions

$$D_{\exp}(F) = \sum_{i=1}^{L} c_i^{\,2} D(f_i)$$

Comparison of observed Q_I variance with that expected for random samples and result of numerical simulations

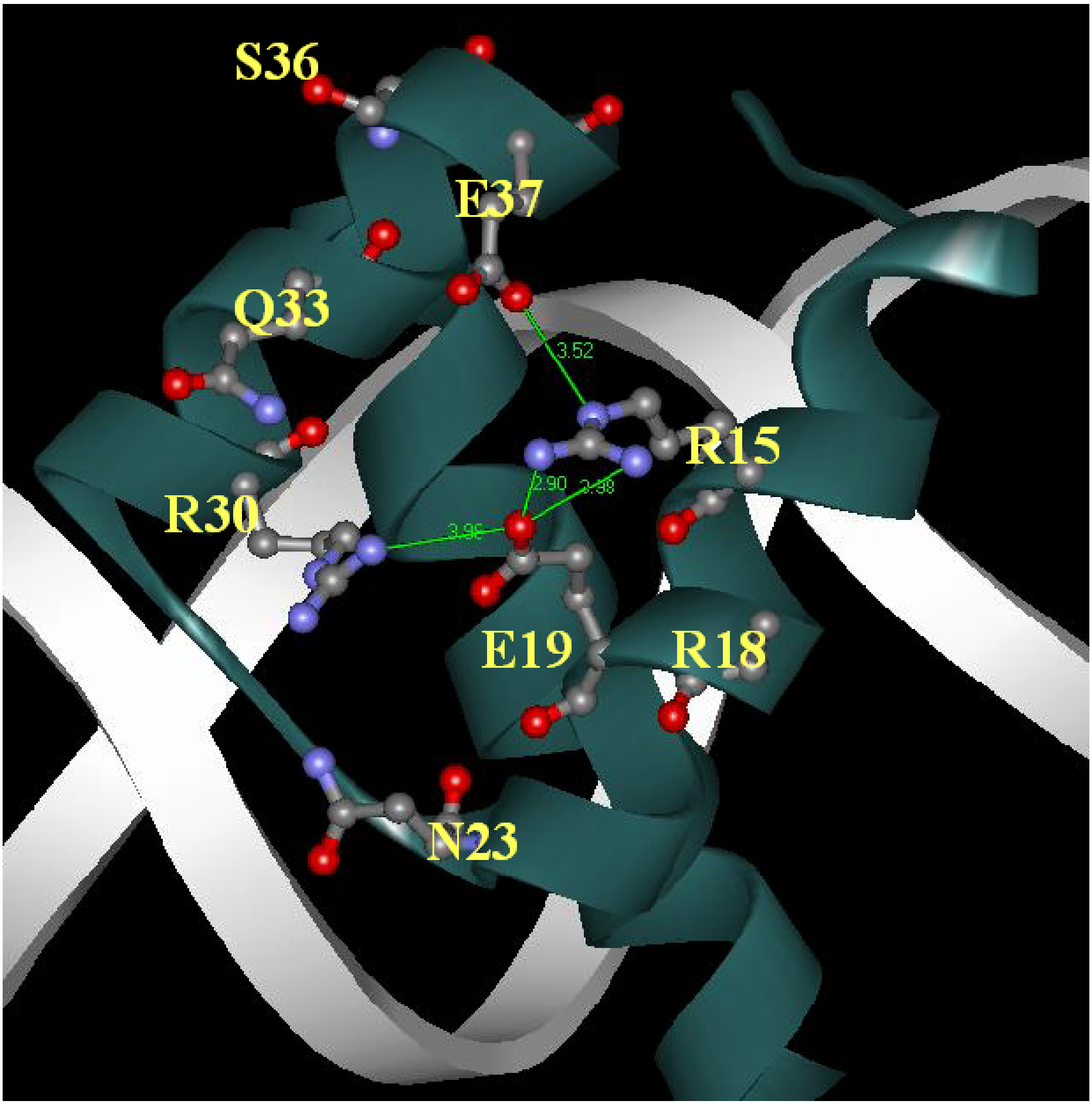
F	$D(F)$	$D_{EXP}(F)$	$D_{RND}(F)$, *mean*	$N(D_{RND}(F)>D(F))$
Q_I	*80.758*	*127.742*	*128.093*	*100000*



Distribution of D_{rnd} (F) in 100000 samples and the value of D(F) (arrow). Significance, estimated from the F-distribution of the Dexp/D ratio: P > 99%.

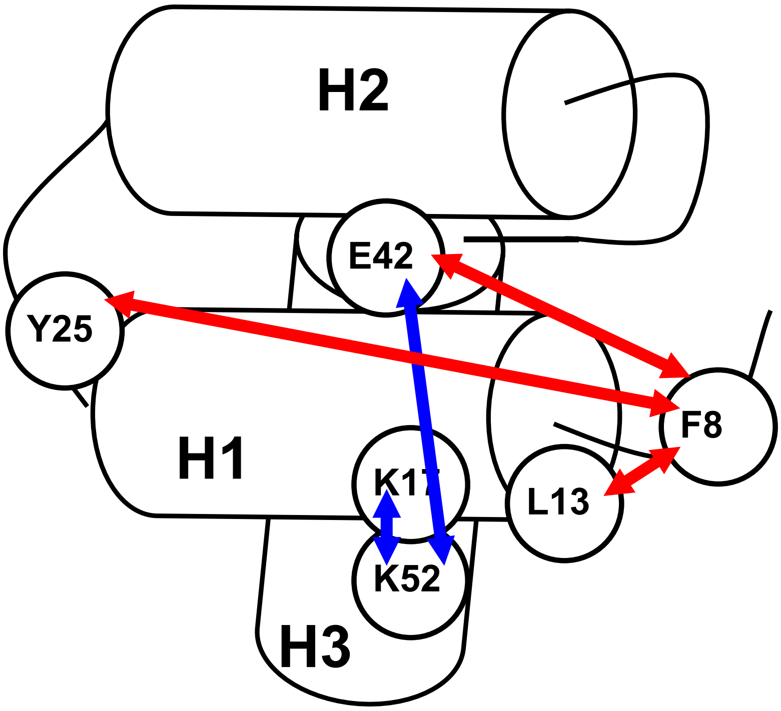
We may conclude, that Q_I is conserved due to co-adaptive substitutions.

SOME OF RESIDUES FROM GROUP I FORM SALT BRIDGES



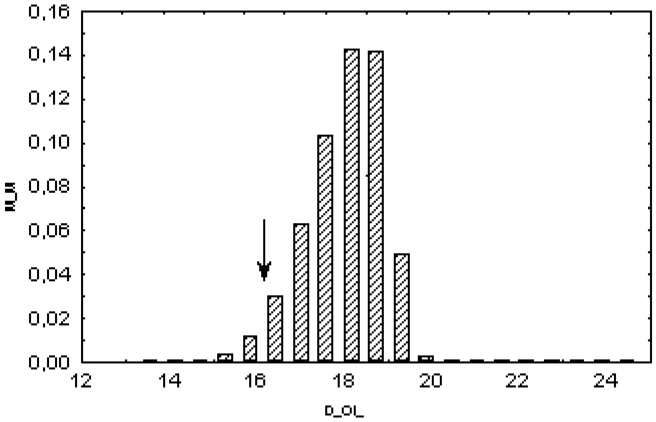
Salt bridges in 1HDC structure: R30–E19, E37–R15, and E19–R15.
Functional importance of Q1 characteristic: stabilization of H1-H2 helix packing.

ANALYSIS OF Q2 CHARACTERISTIC CONSTANCY



Proposed characteristic for cluster II positions: $Q_{II}=pI_{13}+pI_{17}+pI_{25}+pI_{42}+pI_{52}-pI_{8}$

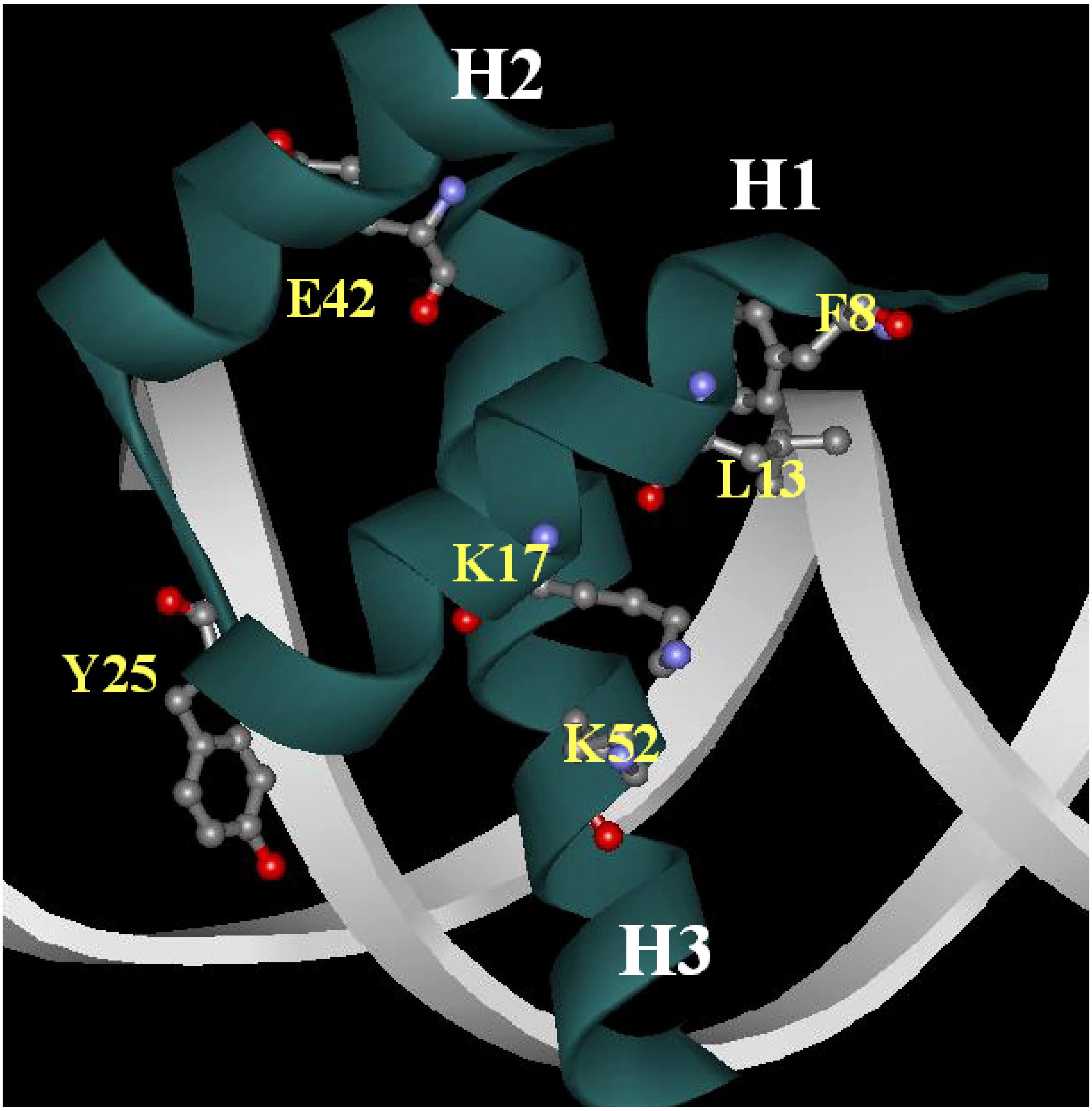
F	*D(F)*	$D_{EXP}(F)$	$D_{RND}(F)$, *mean*	$N(D_{RND}(F)>D(F))$
Q_{II}	16.181	18.939	18.996	98339



Distribution of D_{rnd} (F) in 100000 samples and the value of D(F) (arrow)

Significance, estimated from the F-distribution of the Dexp/D ratio: 99% > P > 95%.

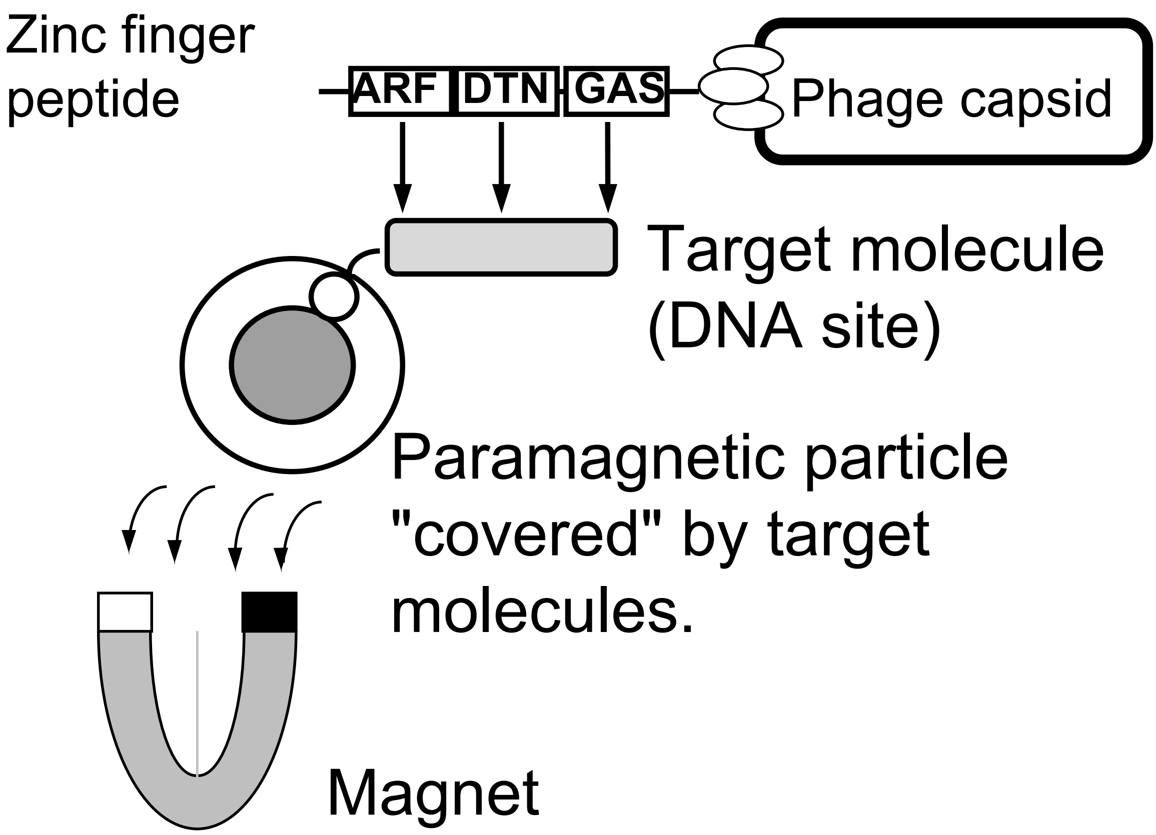
RESIDUES FROM GROUP II ARE CLOSE TO DNA BACKBONE



Proposed function: Interaction with DNA; providing for an appropriate DNA - protein orientation.

ANALYSIS OF SEQUENCE ALIGNMENTS FROM PHAGE DISPLAY EXPERIMENTS

Scheme of *in vitro* selection experiment.



Zinc finger peptide

ARF DTN GAS

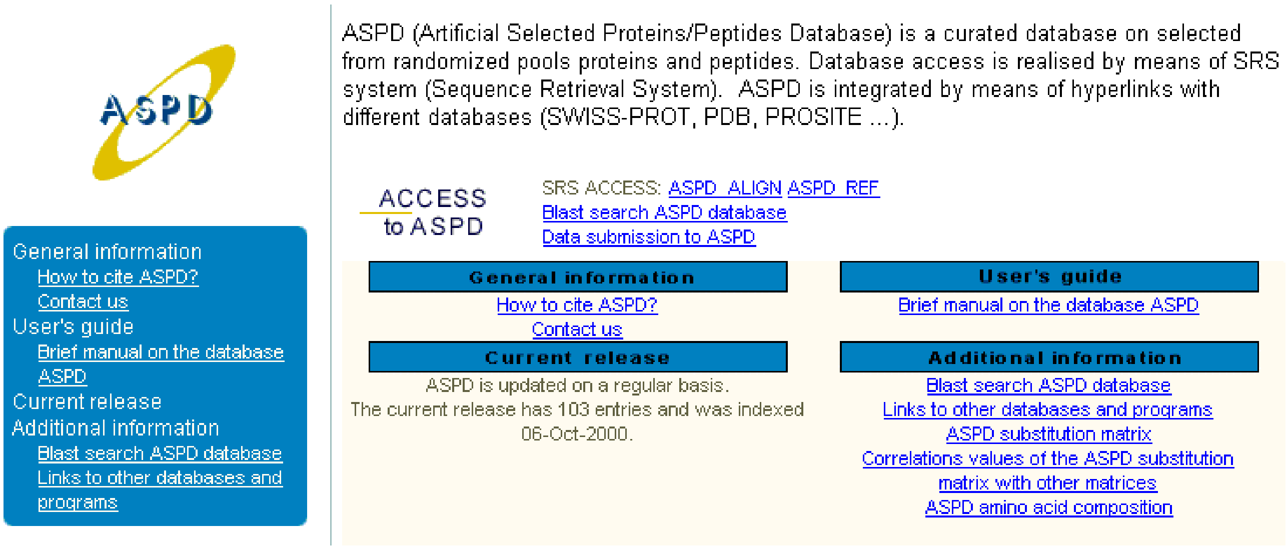
Phage capsid

Target molecule (DNA site)

Paramagnetic particle "covered" by target molecules.

Magnet

ASPD- Artificially Selected Peptides Database.



ASPD (Artificial Selected Proteins/Peptides Database) is a curated database on selected from randomized pools proteins and peptides. Database access is realised by means of SRS system (Sequence Retrieval System). ASPD is integrated by means of hyperlinks with different databases (SWISS-PROT, PDB, PROSITE ...).

ACCESS to ASPD

SRS ACCESS: ASPD_ALIGN ASPD_REF
Blast search ASPD database
Data submission to ASPD

General information
How to cite ASPD?
Contact us
User's guide
Brief manual on the database ASPD
Current release
Additional information
Blast search ASPD database
Links to other databases and programs

General information

How to cite ASPD?
Contact us

User's guide

Brief manual on the database ASPD

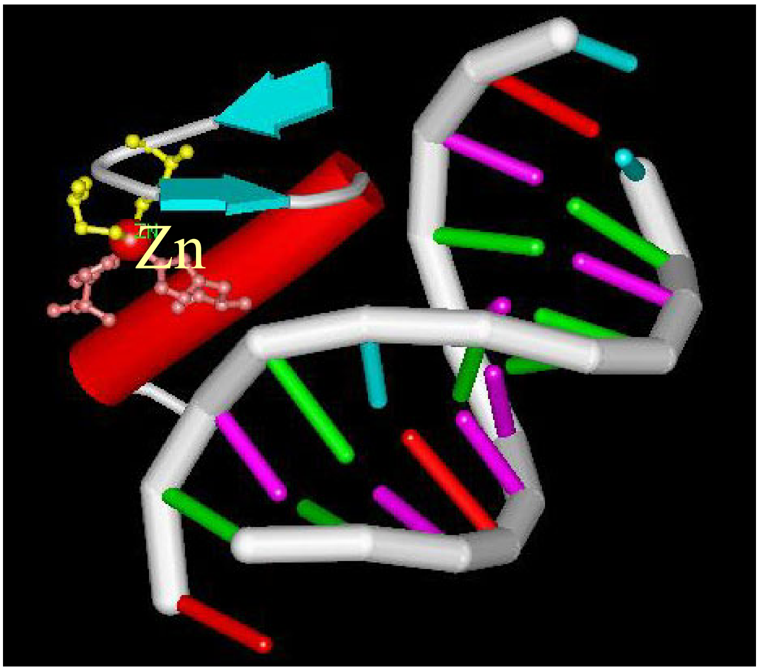
Current release

ASPD is updated on a regular basis.
The current release has 103 entries and was indexed 06-Oct-2000.

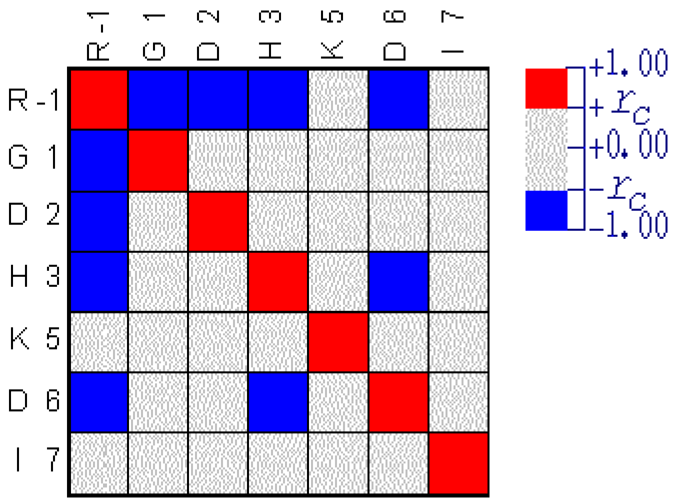
Additional information

Blast search ASPD database
Links to other databases and programs
ASPD substitution matrix
Correlations values of the ASPD substitution matrix with other matrices
ASPD amino acid composition

CORRELATION ANALYSIS OF C2H2 ZINC-FINGER



CO-ADAPTIVE SUBSTITUTIONS: ISOELECTRIC POINT VALUES



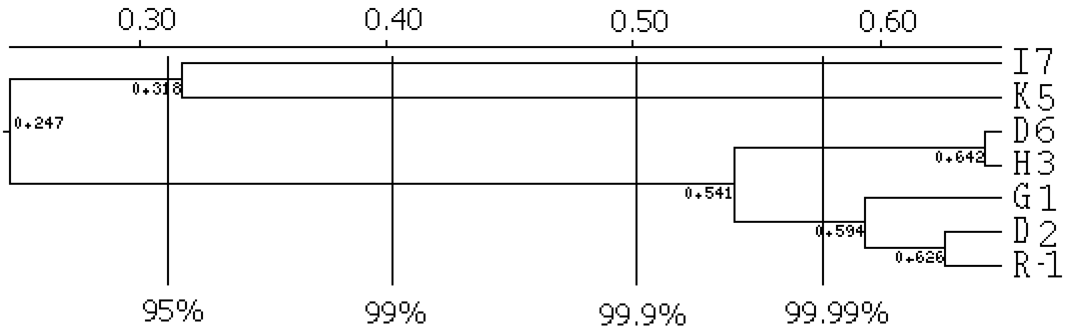
	R-1	G 1	D 2	H 3	K 5	D 6	I 7
R-1	🟥	🟦	🟦	🟦	▨	🟦	▨
G 1	🟦	🟥	▨	▨	▨	▨	▨
D 2	🟦	▨	🟥	▨	▨	▨	▨
H 3	🟦	▨	▨	🟥	▨	🟦	▨
K 5	▨	▨	▨	▨	🟥	▨	▨
D 6	🟦	▨	▨	🟦	▨	🟥	▨
I 7	▨	▨	▨	▨	▨	▨	🟥

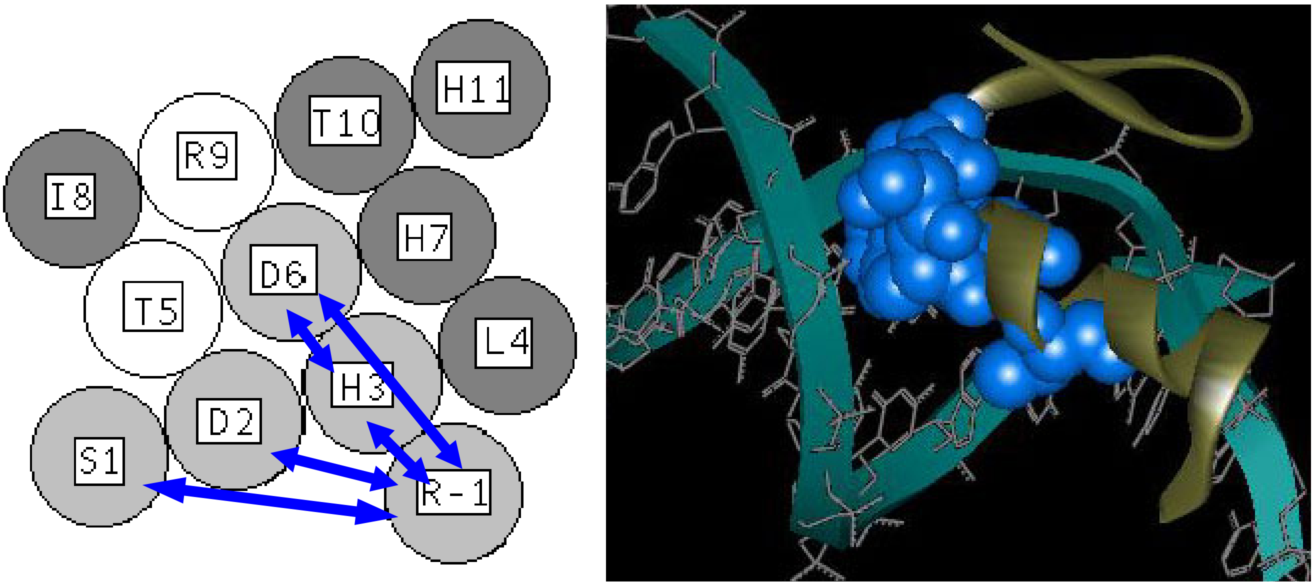
Partial Correlation $r_c = 0.40$
Isoelectric point

45 non-identical sequences from Choo & Klug experiments were analyzed. Position numbering is relative to α-helix first residue.

DETECTION AND ANALYSIS OF CORRELATION NETWORK

Hierarchical clustering diagram





a). Structure of the correlation network: alpha helical projection of residues. Cluster residues in light gray color. Invariant residues in dark gray color. All significant correlation are negative (blue).

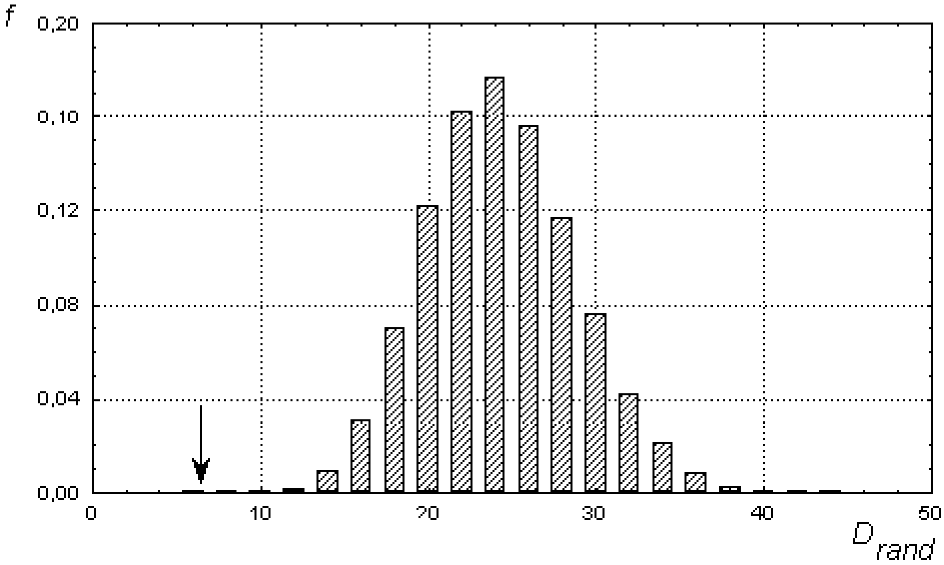
b) Spatial location of residues from detected cluster.

Proposed conserved integral characteristic:

$$Q=pI_{-1} + pI_{1} + pI_{2} + pI_{3} + pI_{6}.$$

ANALYSIS OF THE CONSTANCY OF PROPOSED CHARACTERISTIC

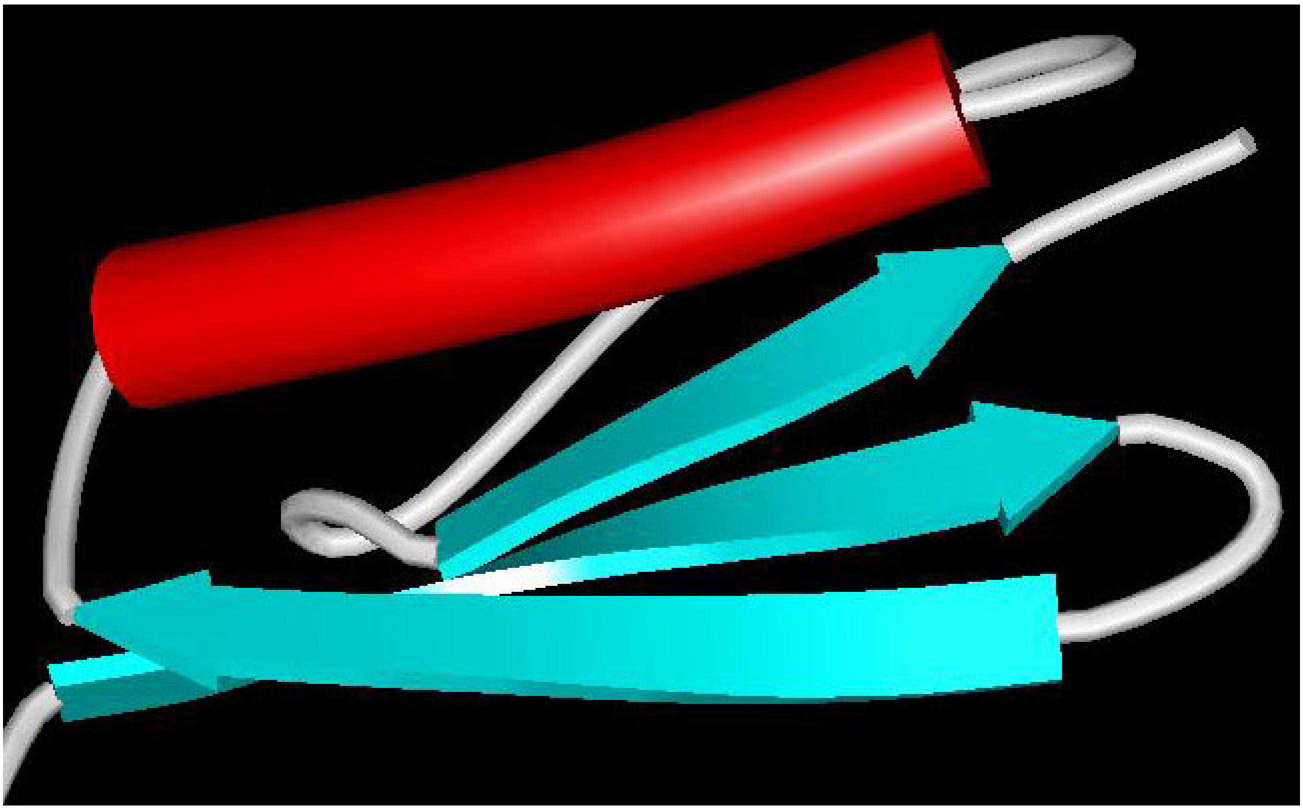
F	$D(F)$	$D_{EXP}(F)$	$D_{RND}(F)$, *mean*	$N(D_{RND}(F)>D(F))$
Q	6.5	24.91	24.4	100000



Distribution of D_{rnd} (F) in 100000 samples and the value of D(F) (arrow). Significance, estimated from the F-distribution of the Dexp/D ratio: P > 99%.

Possible role of the characteristic Q: unspesific electrostatic interaction with DNA , anchoring the helix into the major groove.

ANALYSIS OF THE Ig BINDING DOMAIN (PHAGE DISPLAY DATA)



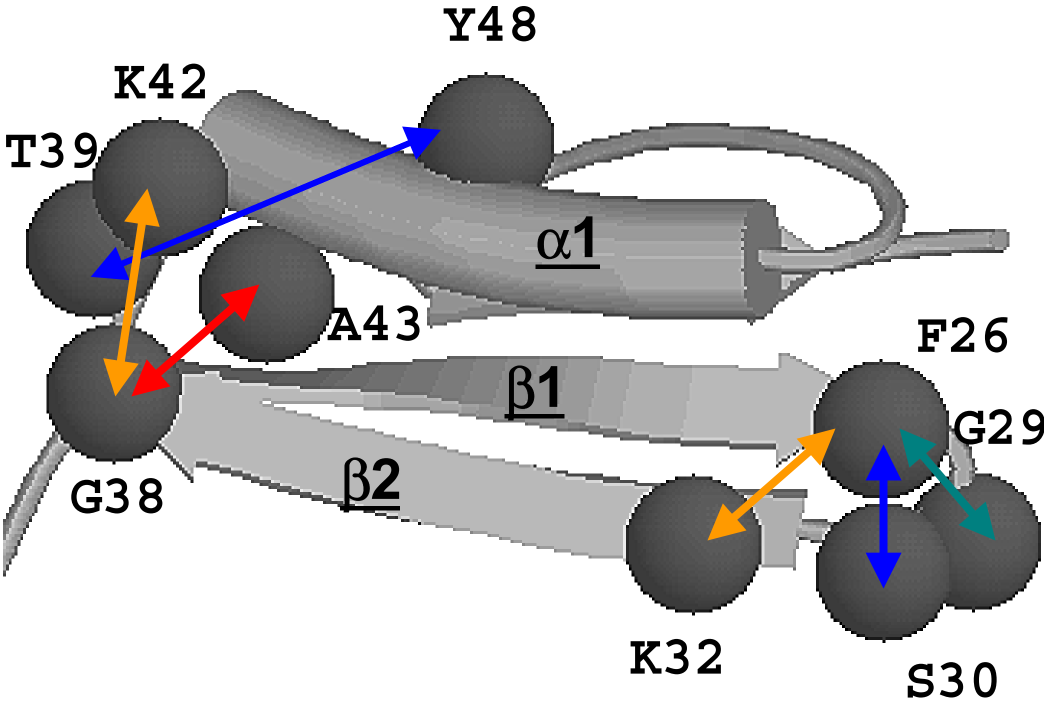
Sequences were selected by fast and correct folding. Two sequence alignments were analysed:

•Gu H., Yi Q., Bray S.T., Riddle D.S., Shiau A.K., Baker,D. Protein Sci. 1995. V. 4, P. 1108-1117.

•Kim D.E., Gu H., Baker D., Proc. Natl. Acad. Sci. USA, 1998. V. 95, P. 4982-4986.

RESULTS OF THE CORRELATION ANALYSIS OF THE Ig BINDING DOMAIN

Tested physico-chemical characteristics: side chain volume; isoelectric point; polarity; hydrophobisity.



↔ Isoelectric point, negative

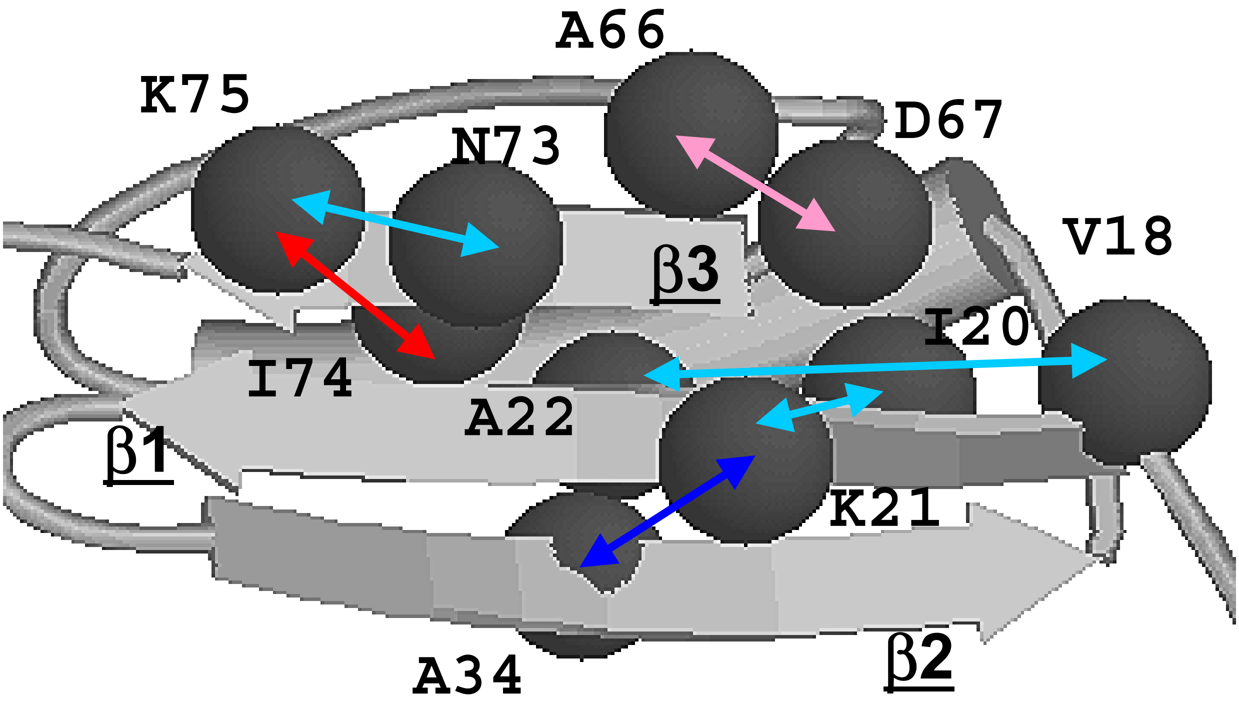
↔ Isoelectric point, positive

↔ Volume, negative

↔ Hydrophobicity, negative

↔ Hydrophobicity, positive

↔ Polarity, positive



Possible function of these interactions: stabilize protein fold, providing for the proper packing of secondary structure elements.

THE POSSIBLE ROLE OF CORRELATED NETWORKS IN PROTEINS.

- **mutational flexibility of the protein in the course of its molecular evolution**
- **the network could form a "collective protein position” subjected to the selective pressure ant reflecting global structural and functional features of proteins**

Acknowledgements:
Yuri Kondrakhin,
Igor Titov,
Vadim Valuev,
Dmitry Grigorovich