New automated classification of DNA-protein complexes in Nucleic acids – Protein Interaction Database

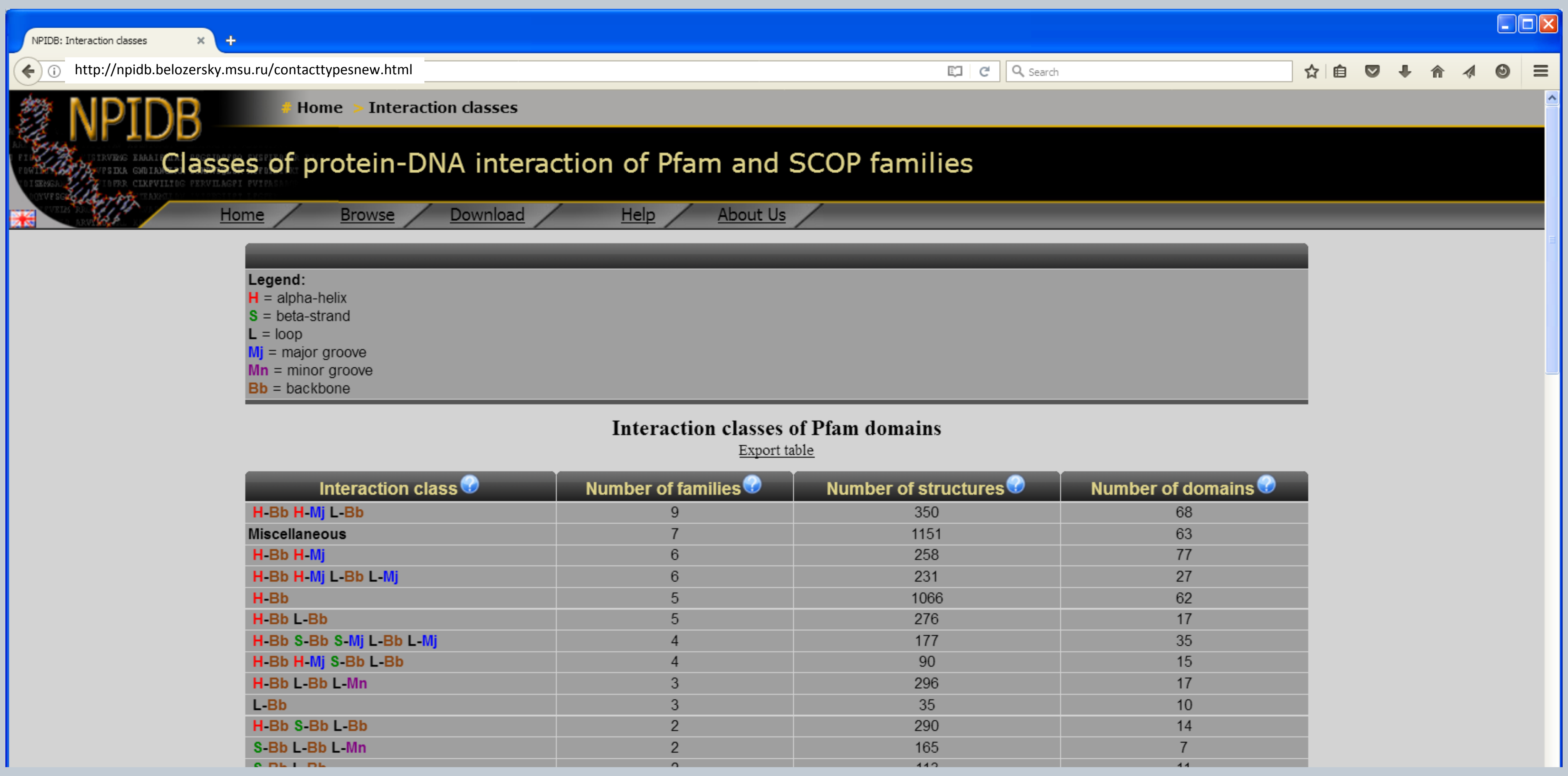
E.F. Baulin[1], O. N. Zanegina[2], A.S. Karyagina[2,3,4], A.V. Alexeevski[2,5], S.A. Spirin[2,5,6*]

[1] Institute of Mathematical Problems in Biology, Puschino, Russia
[2] A.N. Belozersky Institute of Physico-Chemical Biology, M.V. Lomonosov Moscow State University.
[3] N.F. Gamaleya Institute of Epidemiology and Microbiology, Moscow, Russia
[4] Institute of Agricultural Biotechnology, Moscow, Russia
[5] Scientific Research Institute for System Studies, Moscow, Russia
[6] Higher School of Economics, Moscow, Russia
*Corresponding author, e-mail: sas@belozersky.msu.ru

Introduction

Systematization of protein-DNA complexes can be useful for understanding mechanisms of protein-DNA interaction. As the number of known structures rises, it is important to develop an automatic classification of new protein-DNA complexes. We use a contact-based approach for our classification.

Comparison of structures within protein families allow to select conserved features. Thus we developed a classification also for families of complexes. These families are defined according to the Pfam families for the protein parts of the complexes.


NPIDB: Interaction classes
http://npidb.belozersky.msu.ru/contacttypesnew.html
Search
NPIDB
Home > Interaction classes
Home
Browse
Download
Help
About Us

Classes of protein-DNA interaction of Pfam and SCOP families

Legend:
H = alpha-helix
S = beta-strand
L = loop
Mj = major groove
Mn = minor groove
Bb = backbone

Interaction classes of Pfam domains

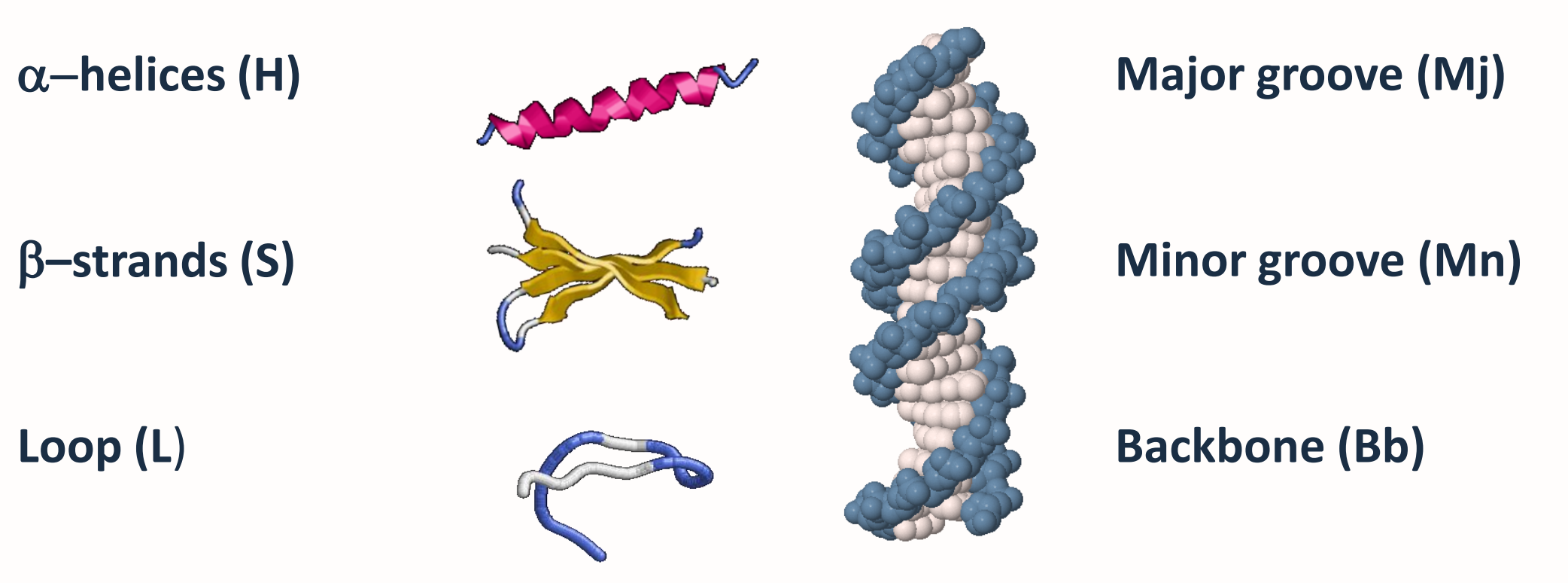
Export table

Interaction class	Number of families	Number of structures	Number of domains
H-Bb H-Mj L-Bb	9	350	68
Miscellaneous	7	1151	63
H-Bb H-Mj	6	258	77
H-Bb H-Mj L-Bb L-Mj	6	231	27
H-Bb	5	1066	62
H-Bb L-Bb	5	276	17
H-Bb S-Bb S-Mj L-Bb L-Mj	4	177	35
H-Bb H-Mj S-Bb L-Bb	4	90	15
H-Bb L-Bb L-Mn	3	296	17
L-Bb	3	35	10
H-Bb S-Bb L-Bb	2	290	14
S-Bb L-Bb L-Mn	2	165	7

Principles

The suggested classification is based on interacting structural elements of protein and DNA molecules.

Pairs of interacting structural elements



H – Mj	S – Mj	L – Mj
H – Mn	S – Mn	L – Mn
H – Bb	S – Bb	L – Bb

We consider not whole protein molecules but Pfam protein domains. Structures of the domains contacting with DNA were extracted from PDB by tools of NPIDB [1], our database of protein-nucleic acid structures.

We analyzed protein-DNA contacts in 11564 structures. The protein domains from these structures belong to 324 Pfam families. Each structure can be characterized by its set of pairs of interacting structural elements. We call this set the **interaction mode** of the structure.

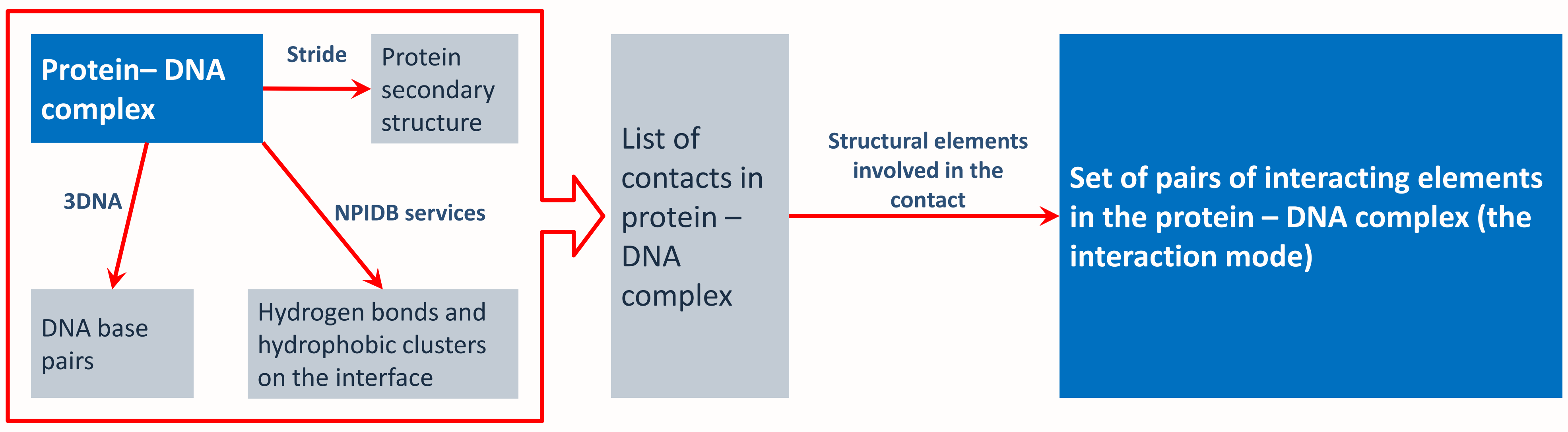
Typically within one Pfam family of DNA-binding protein domains several interaction modes are found. We developed a procedure of automatic Pfam family classification. 77 families containing three or more different proteins were classified onto 29 **interaction classes**.

The previous version of the classification, which is based on SCOP instead of Pfam, is described in our paper [2].

Methods

For each complex we determined all hydrogen bonds and hydrophobic contacts between the DNA molecule and the protein domain. Each contact was assigned to the pair of interacting structure element of protein (α-helix, β-strand, loop) and DNA (the major groove, the minor groove, the backbone).

We characterize each protein–DNA complex by its **interaction mode**, which is the set of pairs of interacting elements.



Protein– DNA complex

Stride

Protein secondary structure

3DNA

NPIDB services

DNA base pairs

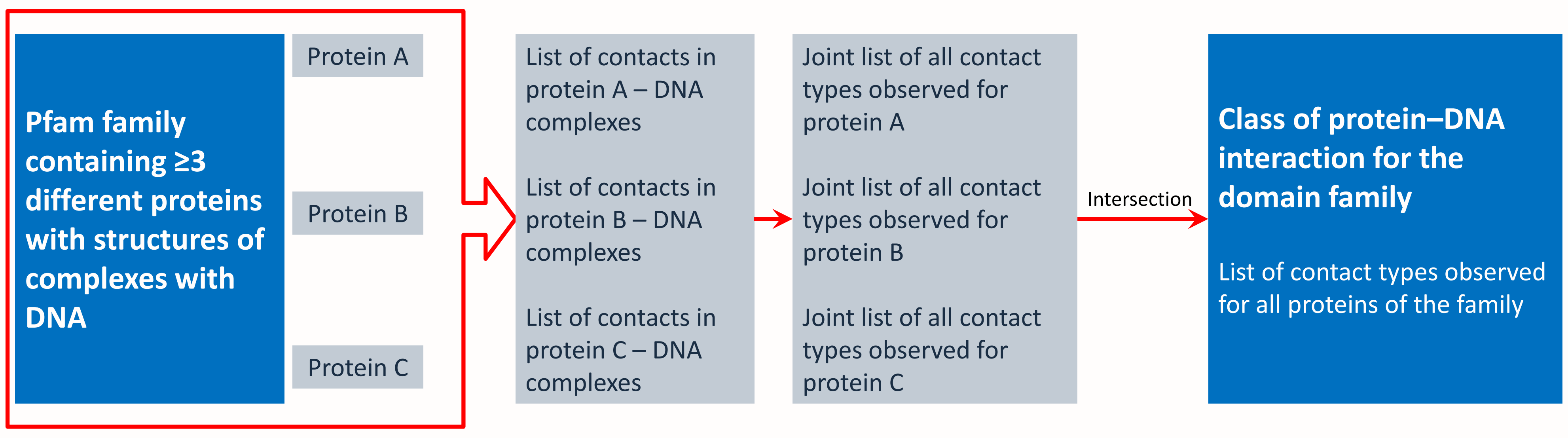
Hydrogen bonds and hydrophobic clusters on the interface

List of contacts in protein – DNA complex

Structural elements involved in the contact

Set of pairs of interacting elements in the protein – DNA complex (the interaction mode)

Interaction class for a Pfam family **is** defined as a list of those pairs of interacting elements that are presented in all protein domains of the family, each pair in at least one structure of each domain.



Pfam family containing ≥3 different proteins with structures of complexes with DNA

Protein A

Protein B

Protein C

List of contacts in protein A – DNA complexes

List of contacts in protein B – DNA complexes

List of contacts in protein C – DNA complexes

Joint list of all contact types observed for protein A

Joint list of all contact types observed for protein B

Joint list of all contact types observed for protein C

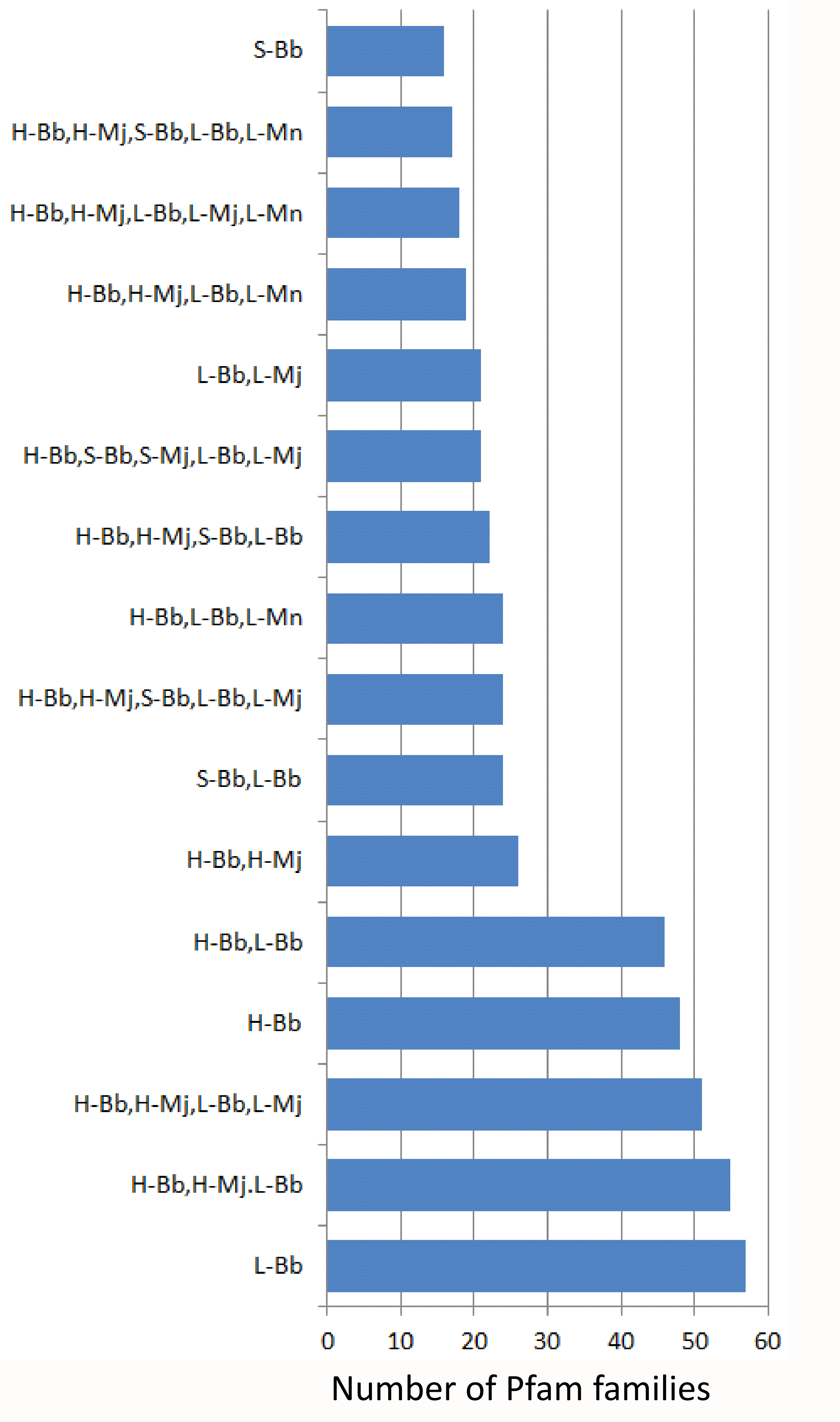
Intersection

Class of protein–DNA interaction for the domain family

List of contact types observed for all proteins of the family

Results

Most represented interaction modes of protein–DNA complexes

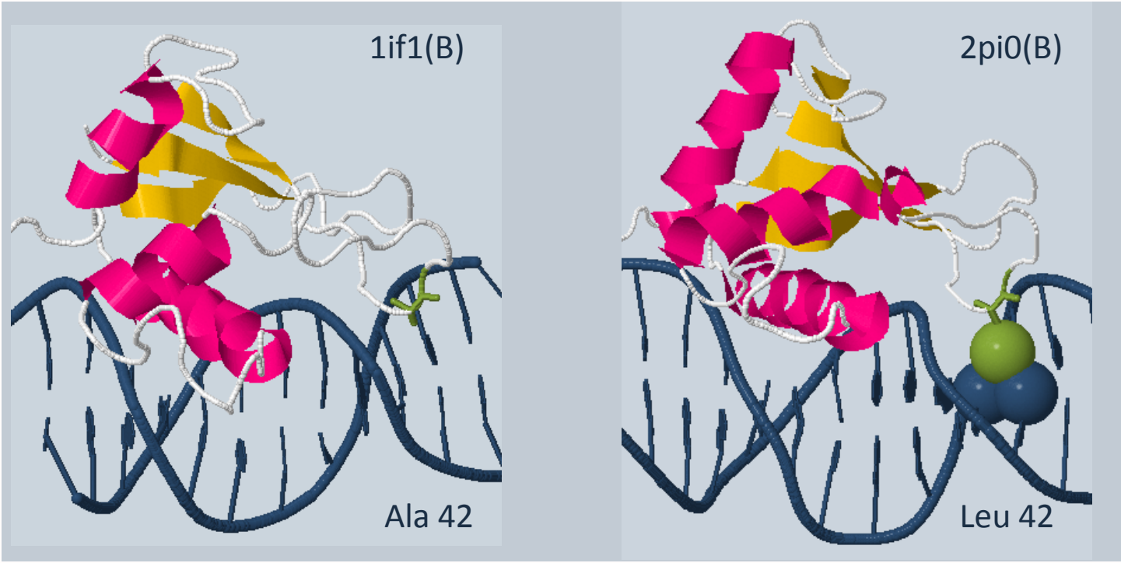


Interaction classes with more than 3 families

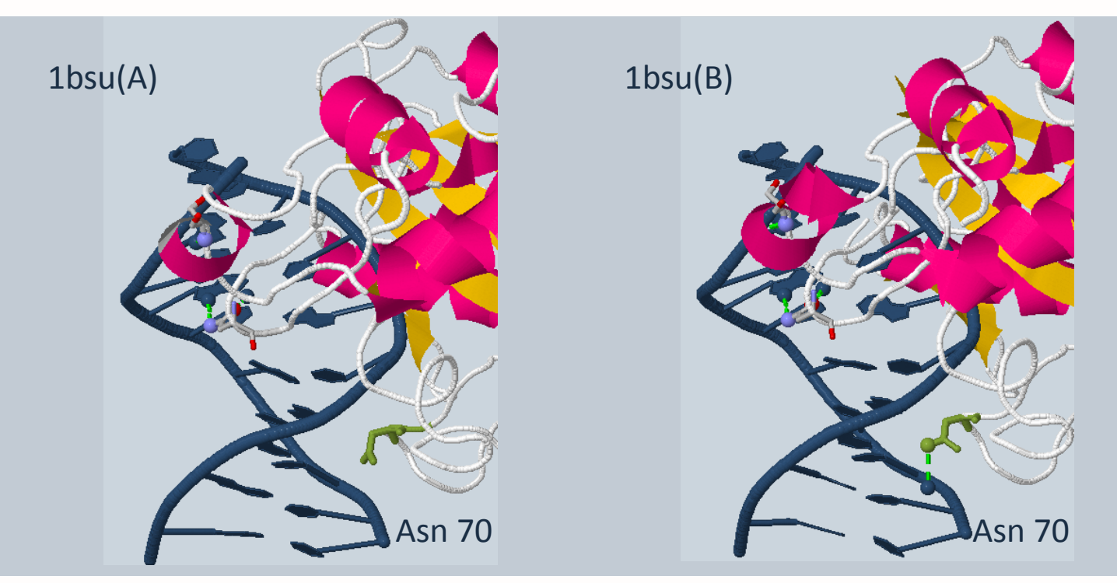
Interaction class	Families	# structures
H-Bb,H-Mj,L-Bb	zf-C4, TetR_N, Ets, GerE, MarR_2, IRF, HTH_18, HTH_Crp_2, E2F_TDP	350
Miscellaneous	TAL_effector, IMS, RNA_pol_Rpb1_1, DNA_pol_B_exo1, RHD_dimer, UvrD_C, CBFD_NFYB_HMF	1151
H-Bb,H-Mj	Homeobox, HLH, SRF-TF, Myb_DNA-binding, bZIP_1, bZIP_Maf	258
H-Bb,H-Mj,L-Bb,L-Mj	HTH_3, LacI, Zn_clus, GntR, Pou, Rep_3	231
H-Bb	Histone, RNA_pol_Rpb1_5, TFIIB, TFIIF_beta, XPG_N	1066
H-Bb,L-Bb	IMS_HHH, MethyltransfD12, UvrD-helicase, Topo_C_assoc, HHH_6	276
H-Bb,S-Bb,S-Mj,L-Bb,L-Mj	LAGLIDADG_1, MH1, T_Ag_DNA_bind, MazE_antitoxin	177
H-Bb,H-Mj,S-Bb,L-Bb	Trans_reg_C, Arg_repressor, HSF_DNA-bind, Penicillinase_R	90

http://npidb.belozersky.msu.ru/

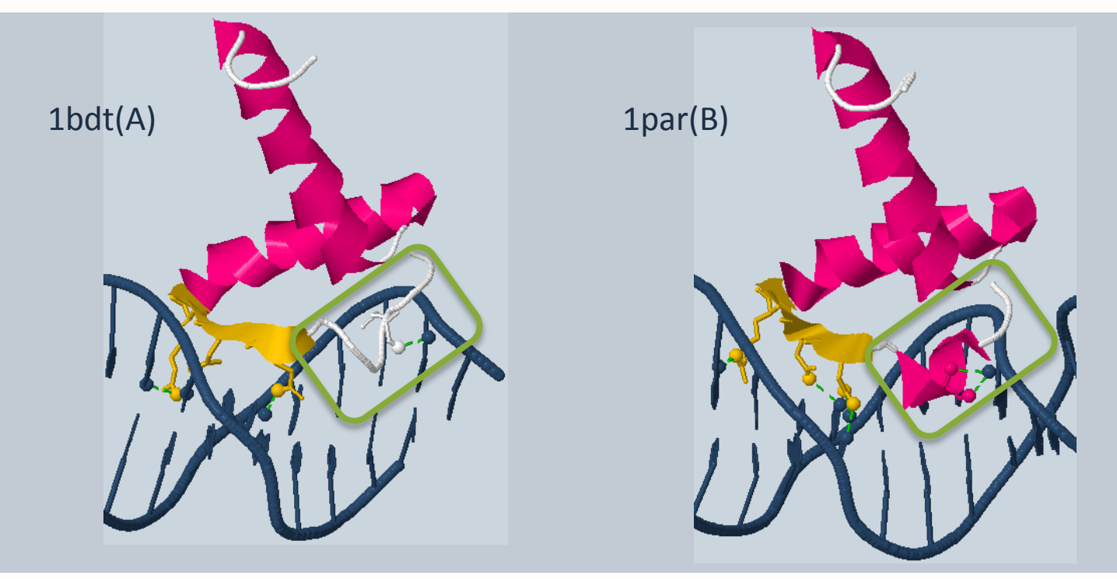
Variations of DNA recognition within one protein family



Interferon regulatory factor family:
replacement Ala → Leu
causes an additional hydrophobic contact.



Restriction endonuclease EcoRV family:
side chain mobility
causes “intermittent” contact (L – Mn).



Arc/Mnt-like phage repressors family:
variation of protein secondary structure
brings an influence on the interaction group.

Causes of DNA recognition variety:
- protein sequence differences;
- molecules flexibility;
- artifacts in secondary structure determination.

References

1. Kirsanov D.D, Zanegina O.N, Aksianov E.A, Spirin SA, Karyagina A.S, Alexeevski A.V. (2013) NPIDB: Nucleic acid-Protein Interaction DataBase. *Nucleic Acids Res.* **41**: D 517-523.

2. Zanegina O., Kirsanov D., Baulin E., Karyagina A., Alexeevski A., Spirin S. (2016). An updated version of NPIDB includes new classifications of DNA-protein complexes and their families. *Nucleic Acids Res.* **44**: D144–153.

Acknowledgements

This work was supported by the Russian Science Foundation, grant no. 14-50-00029