DOI 10.18699/SBB-2022-176

Classification of families of DNA-recognizing protein domains based on structural features of DNA-protein complexes

Panova V.[1]*, Baulin E.[2], Karyagina A.[3, 4, 5], Alexeevski A.[3, 6], Spirin S.[3, 6, 7]

[1] *Faculty of Bioengineering and Bioinformatics of Moscow State University, Moscow, Russia*

[2] *Institute of Mathematical Problems in Biology, RAS, Pushchino, Moscow Region, Russia*

[3] *Belozersky Institute of Moscow State University, Moscow, Russia*

[4] *Gamaleya National Research Center of Epidemiology and Microbiology, Moscow, Russia*

[5] *All-Russia Research Institute of Agricultural Biotechnology, Moscow, Russia*

[6] *Institute of System Studies, Moscow, Russia*

[7] *National Research University Higher School of Economics, Moscow, Russia*

* *verapanova877@gmail.com*

Key words: contact type, interaction class, interaction mode, miscellaneous

Motivation and Aim: DNA-protein interactions play a central role in such cellular processes as DNA replication, transcription, and repair [1]. Prediction of the interactions between DNA and protein is a significant but not trivial task [2]. Currently, 6637 DNA-protein 3D structures are available in open databases. Classification of structures of DNA-protein complexes makes it possible to effectively study the patterns of DNA-protein recognition. The protein domains are united in families, and domains of the particular family recognize DNA in a similar way.

This work aims to correct errors in the existing version of NPIDB ([3], https://npidb.belozersky.msu.ru/), to create a corrected classification of the structures of DNA-protein complexes, to describe structural features, and to make an internal classification of the most abundant families based on corrected data.

Methods and Algorithms: The classification is based on the principles from [5]. It classifies structures of protein domains with neighboring DNA double helix. Instead of domains allocated in protein chains according to the SCOP database whose support was discontinued in 2009, now the domains allocated according to the Pfam database [4] are classified. The classification is based on contacts between DNA and protein molecules. Hydrogen bonds, hydrophobic interactions, and water bridges are considered as contacts. The contact type is a pair of contacting structural elements. Protein structural elements are helices, beta-strands, and turn/unstructured segments, whileDNA structural elements are the major groove, the minor groove, and the sugar-phosphate backbone. The interaction mode for the domain and domain structure is defined as a list of contact types, and the interaction class for a family of domains is defined as the intersection of interaction modes of all domains of the family. If the interaction class is empty, the family is called *miscellaneous.*

Results and Conclusions: A set of Python programs determining interaction modes and interaction classes was written. Only protein structures in complexes with a double DNA helix (with at least 6 complementary pairs), solved by X-ray and CryoEM analysis with a resolution < 3 Å were considered.

The data were statistically processed, the numbers of structures and domains with each interaction mode and the number of families with each interaction class were obtained. The classification of structures of DNA-protein complexes was significantly improved.

Several particular families (including all six miscellaneous families) have been described in detail. The internal classification of families, based on the superposition of DNA-binding parts of a protein in the domain, was described. For all six miscellaneous families, the reasons why the domains of these families do not have common contact types were found.

References

1. Kulandaisamy A. et al. Dissecting and analyzing key residues in protein-DNA complexes. *J Mol Recognit.* 2018;31(4).
2. Gao M., Skolnick J. From nonspecific DNA-protein encounter complexes to the prediction of DNA-protein interactions. *PLoS Comput Biol.* 2009;5:e1000341.
3. Kirsanov D. et al. NPIDB: nucleic acid-protein interaction database. *Nucleic Acids Res.* 2013;41(D1):D517-D523.
4. Mistry J. et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 2021;49(D1):D412-D419.
5. Zanegina O. et al. An updated version of NPIDB includes new classifications of DNA-protein complexes and their families. *Nucleic Acids Res.* 2016;45(D1):D144-D153.