# Tasks for training in structures of DNA-protein complexes

# IRTG GRK1384 "Enzymes and Multienzyme Complexes acting on Nucleic Acids" Offspring Meeting, Moscow June 28, 2013

Andrei Alexeevski [aba@belozersky.msu.ru](mailto:aba@belozersky.msu.ru)

Sergei Spirin [sas@belozersky.msu.ru](mailto:sas@belozersky.msu.ru)

Dmity Kirsanov [ddk@belozersky.msu.ru](mailto:ddk@belozersky.msu.ru)

Evgeniy Aksianov [evaksianov@belozersky.msu.ru](mailto:evaksianov@belozersky.msu.ru)

**Create a file for recording your results, observations and figures.**

Recommended file name is **<Your last name>.doc**

**(1) Analysis of protein-DNA (or RNA) contacts with NPIDB services (hydrogen bonds, water-mediated contacts, hydrophobic clusters).**

Analysis of specific contacts between Protein and DNA in the structure containing Zinc-finger C2H2 domain using NPIDB

**Methods: data base NPIDB**

a) Choose one of the complexes **(**1JK2, 1A1H, 1P47, 1A1G, 1LLM, 1ZAA, 1A1J, 1A1F, 1A1I, 1A1L, 1F2I**)** in the NPIDB (<http://npidb.belozersky.msu.ru>).

**HELP:** *Use the Search form on the main page of DB.*

b) Which SCOP-domains does this complex contain? Write down all these domains using its name and range of aa’s in the structure.

**For example***: Homeobox, 1apl, 244-350:A*

**HELP:** *Use the “SCOP-domains” section on the page of complex***.**

c) Determine one SCOP-domain in analyzed PDB entry that forms the most number of contacts with DNA. Select this domain for the analysis (*Underline it in your previous list*).

**HELP**: *Use the “Sequences” section on the page of complex*.

d) Find all specific contacts (H-bonds, W-bridges, hydrophobic clusters) that selected domain forms with DNA. Write down them using form:

---------------------------------------------------------------------

**Hydrogen bonds**:

ARG133 --- G4

…

**Water-bridges**:

ASP103 --- W4001 --- C54

…

**Hydrophobic clusters**:

Cluster 1: (LEU44, PHE65 --- T46, T47)

…

---------------------------------------------------------------------

**HELP**: *Run “Interaction Service” from the page of complex.*

**(2) Reconstruction of neighborhood of a complex in crystal using PyMol. Analysis of contacts forming the crystal.**

PDB file 1mnm represents X-ray structure of the yeast homeodomain repressor protein MATalpha2 and the MADS-box transcription factor MCM1 bound to DNA. The structure contains two MATalpha2 subunits (chains C, D) and two MCM1 subunits (chains A, B). One protein molecule contains a fragment (>10aa) that cannot preserve such conformation in a solution.

Find this fragment and explain why is was possible to solve this fragment by X-ray and why contacts of this protein molecule with DNA seems, do not reflect actual contacts in native complexes.

**Methods: Pymol**

a) Load PDB entry by Pymol

**HELP:** *Menu PlugIn => PDB loading service*

b) Create an appropriate view.

Recommended view: cartoon model, color by chains, grey background.**HELP:** *Use the menu in the main window: A – “action”, S – “Show”, H – “Hide”, C – “color”  
Use command* bg\_color <color> *in command line window to change backgroun*d

c) Find questionable fragment, and write down its coordinates. Make selection for this fragment

Remember, that a polypeptide chain not contacting with the protein globule or other molecules is flexible!   
**HELP:** *Useful possibilities:*

* *clicking on atom allows to determine chain, residue number etc.*
* *command*select my\_fragment, resi 105-127 and chain A

*creates new selection named “my\_fragment” and adds it to the list shown in the right panel of the main window*

d) Reconstruct neighborhood of the chosen fragment in the whole crystal

**HELP:** *Use command*

symexp ppp, 1mnm, my\_fragment, 5.0

*here   
 ppp – prefix for neighboring objects*

*1mnm – an object to which the command will be applied*

*my\_fragment – the selection in the neighborhood of which other objects in the crystal are searched*

*5.0 – distance threshold in angstroms*

e)Watch and describe the result in the protocol. Add an image from Pymol.

f) \* Describe in the protocol how DNA oligomers are arranged in the crystal. Add figures.

**HELP:** *Delete neighboring objects:*  delete ddd\* (*delete everything with names started by ddd)*

*Create view of DNAs only. (One of possibilities is: menu Display => Sequence mode – select chains, Display => Sequence, in the main window choose appropriate chains, rename selection (sel) to dna; all: hide everything; dna: show cartoon )*

*Command* symexp ddd, 1mnm, dna, 30.0

*Watch the result, describe and make nice figure.*

**(3) Superimposition of structures with PDBeFold. Analysis of differences between structures of a free protein and of the same protein in complex with DNA. Using pre-calculated superimpositions in NPIDB.**

**Task 1. Prediction of DNA-interacting region of the protein by structural similarity search**

EXAMPLE.PDB is a structure of DNA-interacting protein, solved without DNA. The task is to find potential DNA-interacting region of the surface of this protein.

**Methods: Server PDBeFold, Pymol**To solve the task you should find the PDB file with complex of DNA with a protein which structure is similar to the structure of given protein. The hypothesis is that these similar proteins interact with DNA by similar regions.

a) Find a structure of DNA- protein complex, similar to given example.

*Use PDBeFold program; use Google to find this server. PDBeFold is like BLAST, but for 3D structures instead of sequences. It allows to create pairwise or multiple superpositions and to search structures similar to your query in whole PDB bank.*

*Create pairwise superpositions using search against PDB. For the every hit PDB ID is available. You can read PDB header of the hit on PDB web site (enter PDB ID in the query field and hit Display files* ***→*** *PDB Header. All macromolecules of PDB entry (protein and DNA chains) are listed in start of the header. Find the structure of protein in complex with DNA.*

*Save ID and title of the found complex and parameters of superposition: RMSD, length of the alignment and percent of identical residues.*

b) Download the superposition and the sequence alignment.

*The left numbers on the result page are links to the detailed description of the every hit. To download superposed query and the hit file check “whole entries” option and press “view superposed”. JMol will start and you can save PDB file using its menu (press right mouse button on the structure, select first item and than – “View \_ssm\_viewer\_file1” - .gz file will be downloaded).*

*PDBeFold renames protein and DNA chains in structures. For example, if the query was a structure with one chain, named C, and target has also one chain with name C, the query will be renamed to A and the target to B in superposed file. To determine which chain in superposed file corresponds to query and target – calculate the number of Cα-atoms in them. For example, in PyMol it can be done by the command like*

select a\_backbone, chain a and name ca

*for the query and superposed file. Numbers of Cα-atoms in identical chains are equal, obviously.*

c) Which amino acids of the query are potentially bind to DNA?

*Let chain A is a query, chain B is a target, C and D are DNA chains.*

sel example, chain a

sel hit, chain b

sel dna, chain c+d

hide everything, all

show cartoon, example or hit or dna

center example

*We can simply find amino acids from the query, which are near to DNA atoms of the hit in superposition.*

sel cnt, example within 5 of dna

show spheres, cnt

*Now you can pick on selected atoms and find amino acids, which are they from. To determine their amino acids use*

hide spheres, all

sel cnt\_ca, name ca and byres cnt

show spheres, cnt\_ca

*Really, conformations of the side chain of the protein in complex with DNA can differ from those of the single protein. So, if you see that amino acids 25, 26, 28, 29 and 30 of example are near to DNA of the target, the possible hypotheses is that 27 is also possible to contact with DNA.*

*One of the more accurate method to fined cnt set is follows. Determine which atoms of target are in contact with DNA*

sel hit\_atm, hit within 5 of dna

*and Cα-atoms of those amino acids*

sel hit\_ca, name ca and byres hit\_atm

*The correspondent Cα-atoms in query are*

sel example\_ca, (example within 1 of hit\_ca) and name ca

*and all atoms of those amino acids are*

sel example\_cnt, byres example\_ca

show line, example\_cnt

d) Save potentially DNA-interacted regions of the example in the protocol. Find correspondent regions in the alignment. Are sequences of example and hit regions similar?

**Task 2.** Examine and describe in protocol conservation of relative conformations of DNA and DNA-binding protein within SCOP family.

**Methods: NPIDB**

**HELP:** *Browse SCOP families. Chose one-by-one several families marked with orange stars. For these families superimposition of DNA-protein complexes are available. Watch superimpositions and write down your impressions.*

**Task 3.** Transcription activator–like effectors (TALe) are promising proteins for engineering highly specific DNA cleavage proteins (TALe + ENdonuclease = TALEN) due to their module structure. One module consists of two alpha-helixes and two aa turn between them. These two aa’s determine the specificity of the recognition of just one DNA base pair. Compare different modules of LANe for PDB file 3v6t. Which parts change conformation (particularly, are conformations of backbone of all turns the same or not?) depending on variable aa’s?

**Methods: PDBeFold and/or Pymol  
HELP:** Find a way to superpose separated modules!

**(4) Finding ‘conserved’ water molecules by comparative analysis of structures (NPIDB, wLake service). How to interpret water in PDB structures.**

Resolving water molecules by X-ray analysis is more complicated problem than atoms of large molecules.

First, waters can be found only if in the crystal water molecules occupy the same position relative to protein or DNA in all or almost all crystallographic cells. This can be if a water molecule is bonded to a protein molecule by mainly hydrogen bonds. Thus, waters are frozen in the crystal and are believed to retain for a relatively long time in the same position in a solution.

Second, there is a freedom during PDB model reconstruction from X-ray experimental data. Interpretation of an electron density condensation as either a water or just a noise depends on parameters of a computer program.

Third, reliability of a water molecule depends, of cause, on the resolution. Typically, water molecules are not included into PDB file if the resolution is worse than 2.5 Å. In a case of extremely high resolution ( < 1.2 Å), waters are often endowed with occupancy coefficient, which indicated merely the fraction of proteins (or DNA) containing water molecule in this position.

Position of a water molecule in a PDB entry is much more reliable if in two or more independently solved structures waters are found approximately in the same place. These waters may be found in superimposed structures of related proteins as clusters of waters from several separate entries.

**Task 1.** Find putatively pre-existent in free restriction endonuclease BglII water molecules, which are involved into interaction with DNA.

**Methods: NPIDB and Jmol.**

a) Find page of the “restriction endonuclease BglII” SCOP family in NPIDB. Find which ones of family members are in complex with DNA and which ones – no.

b) In Jmol window, show water molecules that are in the same position IN ALL STRUCTURES.

**HELP:** *Open ”Water clusters” “…pivot.txt” and find in the table numbers of such waters.*

*Show selected waters by cliquing on their identifiers.*

c) Show DNA atoms with which selected water molecules form hydrogen bonds.

**HELP:** *open Jmol consol by right cliquing on the background in Jmol window.*

*Names of waters defined in NPIDB are the same: wat25 etc.*

*Select and show needed DNA atoms, using consol commands:*

select within(3.5, watNNN) and DNA cpk 200

color cpk

*Here NNN is a number of water cluster*

*3.5 is the threshold for the distance between a donor and an acceptor in a hydrogen bond, usually used for detecting H-bonds in PDB files.*

d) Examine DNA atom types, write down your conclusions in the protocol.

**(5) Finding specific DNA motifs in 3D complexes using 3D-footprint DB services**

GATC is well-known site which is methylated by Dam methyltraspherases in a number of bacterial genomes. According REBASE ([http://rebase.neb.com](http://rebase.neb.com/)) GATC is also cleavage site for more than one hundred restriction enzymes.

**Task 1.** Find PDB structures of protein-DNA complexes such that a protein interacts with the GATC site, and describe the proteins functions.

**Methods: 3D-footprint DB (**[**http://floresta.eead.csic.es/3dfootprint/**](http://floresta.eead.csic.es/3dfootprint/)**)**

**HELP:** *Put GATC into appropriate form for search. The result contains LOGO for each hit. This LOGO is based on computation of energy of interactions between a base pair and protein. Small height of the letter reflects weak interaction of the bases pair with the protein. On the page of a hit there are LOGO and so called Positional Specific Scoring Matrix (PSSM) . PSSM is also computed on the base DNA-protein contacts in the structure. It is used for search of similar sites.*

*“Interface signature” shows aa’s contacting with the DNA. See also “protein sequence” in which residue numbers of this residues are written.*

**Task 2.** 3D-footprint data base reports similarities in protein-DNA interaction sometimes even for not homologous protein. Find such example within GATC specific proteins, inspect this similarity and write down your opinion in the protocol.

**Methods: 3D-footprint DB (**[**http://floresta.eead.csic.es/3dfootprint/**](http://floresta.eead.csic.es/3dfootprint/)**), NPIDB Jmol or Pymol.**

### HELP: *find a hit on the page of which the “Dendrogram of similar interfaces*[*?*](http://floresta.eead.csic.es/3dfootprint/help.html#cluster)*” is presented. Choose a pair of protein-DNA complexes with putatively similar interfaces with DNA. Open these complexes, it is appropriate to use direct links to NPIDB and visualization services there. Create appropriate view of protein–DNA interfaces. Compare structures and describe found similarities in the protocol.*

**(6) Mistakes in PDB files and how to find them using Electron Density Server.**

PDB model is reconstructed from X-ray data by several step procedure using computer programs. Each step contains underlying potential problems. This is why experts in X-ray call 3D structure ‘a PDB model of a protein’. ‘Experimental’ electron density which may be shown by computer programs is partially based on experiments (so called, phases are taken from the model – PDB file, and so called, factors? are taken from experiment).

**Task 1.** The structure of LANe protein in complex with DNA was solved with good resolution 1.85 Å. Does it mean that coordinates of all atoms presented in PDB entry 3V6T are correctly derived from the experimental data?

**Methods: Electron Density Server (EDS,** [**http://eds.bmc.uu.se/eds/**](http://eds.bmc.uu.se/eds/)**), Pymol.**

a) Find aa’s and nucleotides with insufficient electron density (Real-space R-value >0.2). Choose one for further analysis.

**HELP:** *Open “Real-space R-value”(RSR) plot. RSR greater than 0.2 indicates poor electron density around amino acid residue (or nucleotide, or water molecule). Consequent aa’s with bad RSR have more chances to be a mistake than a single aa.*

b) Download electron density map from EDS.

**HELP:** *Save electron density map in your directory (link Download, Maps). Format “O”, map is called “2mFo-DFc”* . *Unpack map to 3v6t.omap file.*

c) Load structure 3V6T from PDB. Make selection containing (i) chosen residue, (ii) neighborhood of it.

**HELP:** *Pymol => plugin => Loading service. Commands for selections:*

select rrr, resi <NNN> and chain <X>

*NNN is a residue number, X is chain id*

select neib, byres(rrr around 5)

d) Upload electron density map (ED map).

**HELP:** load 3v6t.omap, map

e) Make appropriate view of the selection

**HELP:***Hide everything*

*Show rrr by sticks*

*Show neib by lines*

*Center rrr (Action => center)*

f) Create electron density (ED) map around the selection. Adjust the level for ED map to be shown. If too low level is chosen (for example, 0.5 sigma), then noise ED, probably, will appear; if too high (2.5 sigma) then, probably, too less atoms will be surrounded by ED. Try to find the golden mean!

**HELP:** isomesh new\_map, map, 1.5, rrr, 2.0

*new\_map is a name of a map that will be created around selection rrr*

*map is name of the map of all structure, it was defined previousely*

*1.5 is a level of electron density, which will be shown; is measured in standard deviation from an average electron density*

*1.0 means that ED will be extended in 1 angstrom alone each axis.*

g) Save final image in the protocol and comment the result.

**Task 2.** Find a mistake within four structures of homeodomains in complex with DNAs: 1akh chain A (yeast MatA1), 1akh chain B (yeast Matalpha2), 1au7 chain A (rat Pit1), 1b72 (human HoxB1). The mistake concerns conserved asparagines that are homologs (occupies the same position) of Asn120 in 1akh\_a. Look at hydrogen bonds formed by Asn’s with the DNAs.

**Methods: PDBeFold, Pymol**

a) Superimpose all structures

b) Create appropriate view of Asn’s in all structures, color by element. Find and show DNA bases, contacting with Asn’s.

c) Watch donor-acceptors forming hydrogen bonds

d) Write down comments in the protocol.