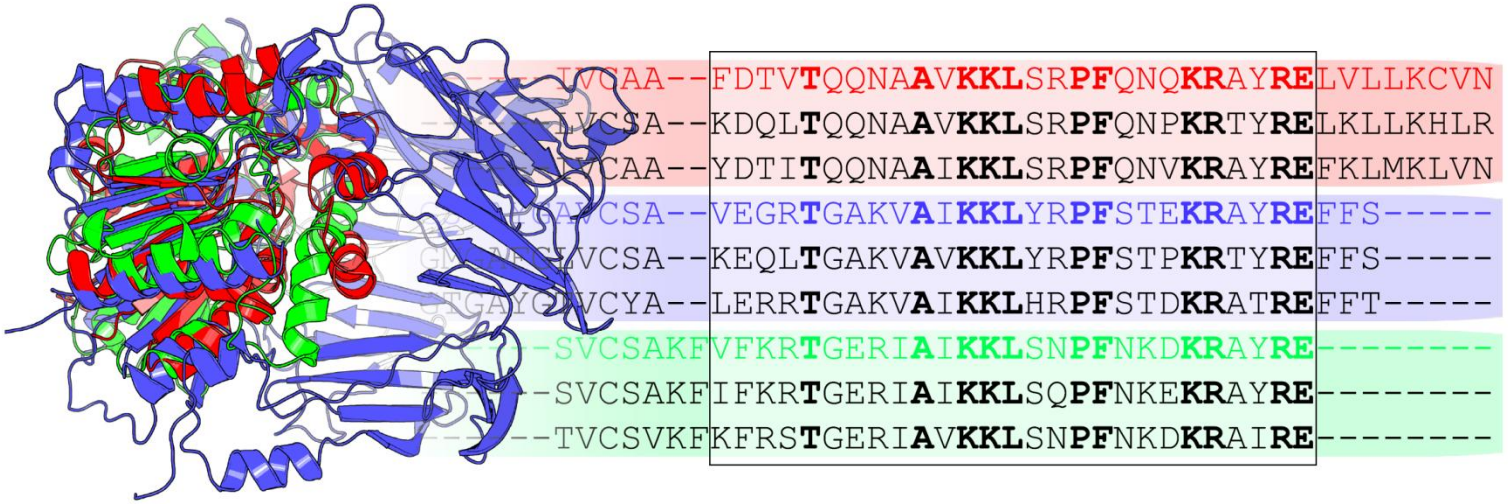
Mustguseal

Server for **Mu**ltiple **St**ructure-**Gu**ided **Se**quence **Al**ignment of Protein Families



https://biokinet.belozersky.msu.ru/mustguseal

Introductory presentation
by Dmitry Suplatov
d.a.suplatov@belozersky.msu.ru

Lomonosov Moscow State University
Moscow, Russia
October 1st, 2019

Multiple alignments as a tool in protein studies and engineering

Comparative bioinformatic analysis of functionally diverse protein families can be used to:

- study structure-function relationship in proteins;
- predict key residues to be mutated in order to produce more stable and functionally diverse proteins and enzymes;
- discover and characterize novel binding sites in protein structures.

Mustguseal

Multiple **St**ructure-**Gu**ided **Se**quence **Al**ignment

- a bioinformatic **protocol** to build large alignments of functionally diverse protein families;
- a **platform of five integrated servers** to provide a user-friendly web-based interface to the Mustguseal protocol and sister methods to further study the obtained multiple alignments;
- Mustguseal can be used to build a large alignment of the *selected protein families* or superimpose a diverse set of proteins representing a *superfamily*.

Mustguseal: the Aim

- to construct large alignments of functionally diverse protein families based on all available information about their structures and sequences in public databases;
- to automate complex bioinformatic procedures;
- to provide a freely-available user-friendly platform on the Internet for daily use in the laboratory practice.

Mustguseal: the Approach

- **structure similarity search** is implemented to collect remote evolutionary relatives, which are expected to represent different protein families;

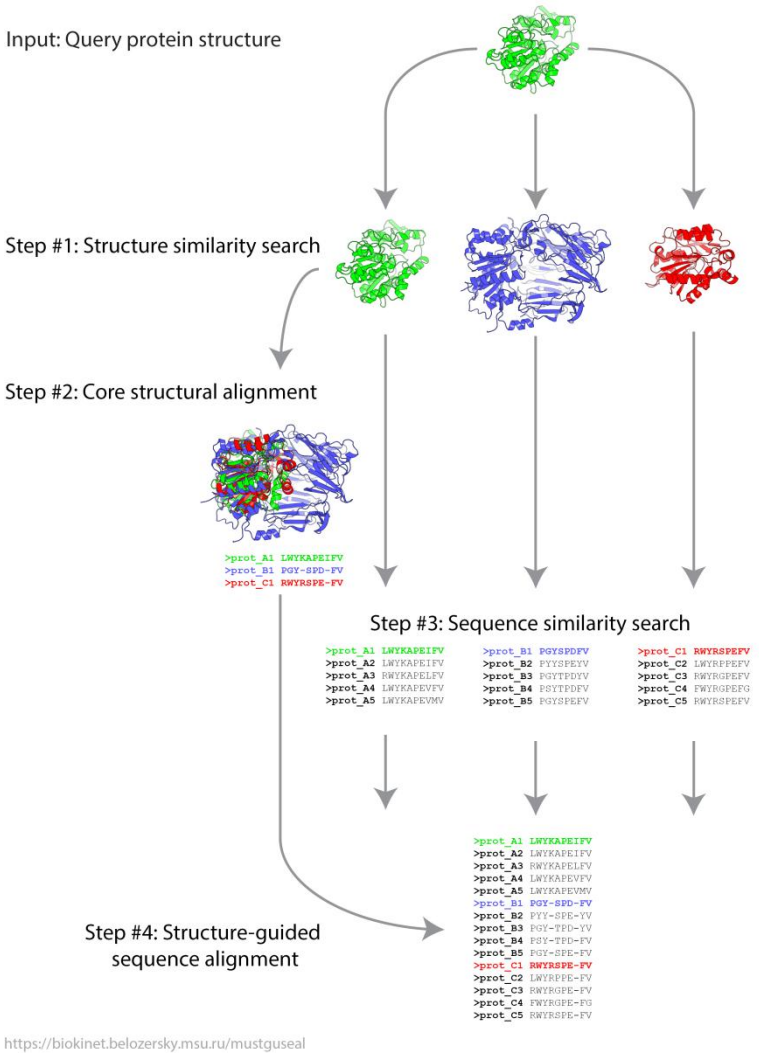
- **sequence similarity search** is implemented to collect close evolutionary relatives - members of the corresponding families;

- a combination of **structure and sequence alignment** procedures is then implemented to build the final multiple alignment;

- multiple alignments of **thousands** of protein sequences and structures can be automatically constructed using this public web-server.

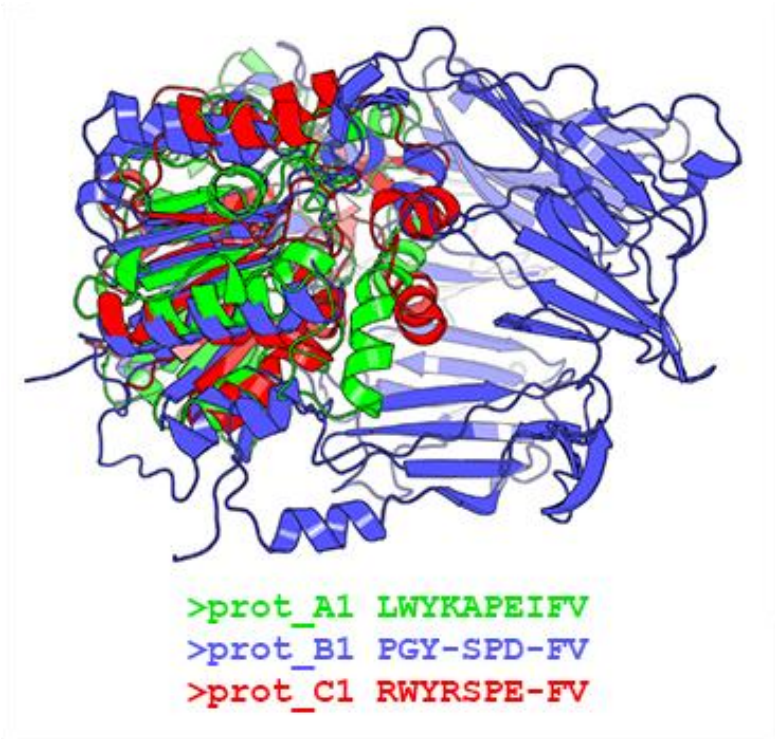
Mustguseal: the Input

Mode 1: Submit a query protein



Mustguseal: the Input

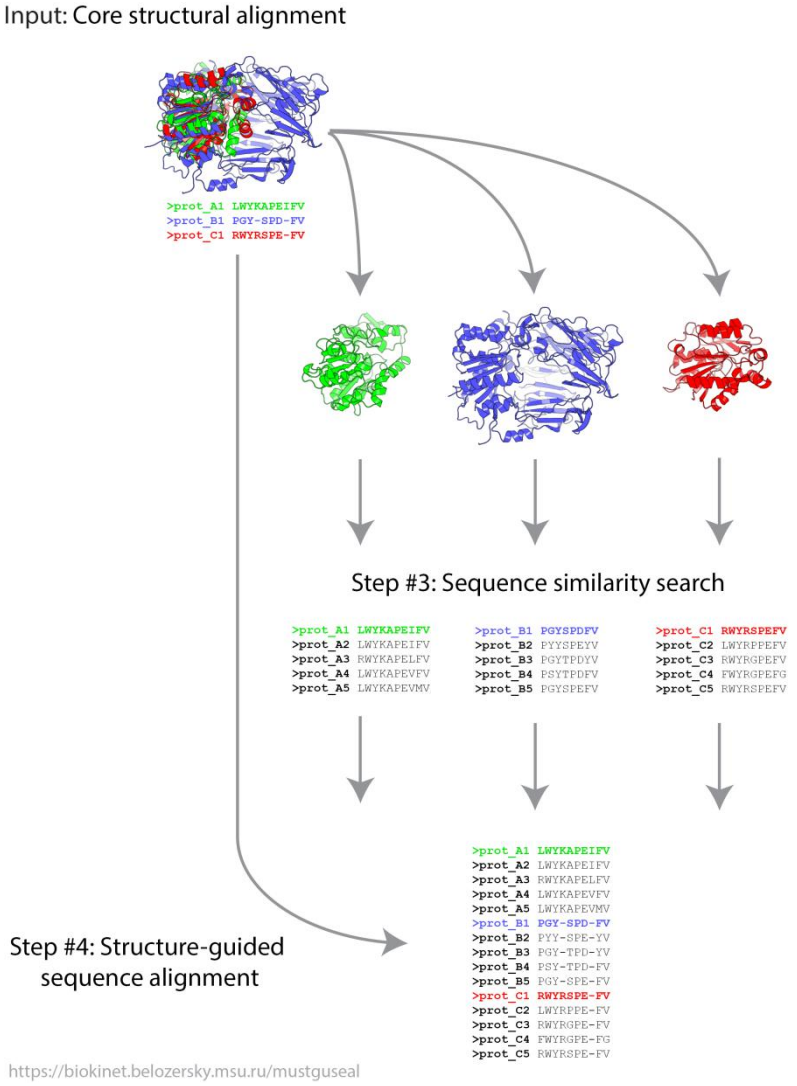
User-defined core structural alignment



- Mustguseal can be used to build an alignment of the *selected protein families* or superimpose a large collection of proteins representing a *superfamily*;
- The scope of the final alignment is defined by the **diversity of representative proteins in the core structural alignment** which can be created on-site or submitted by the user;
- A user-defined 3D-alignment can be built from a selected set of protein structures on a local computer or a supercomputer and then submitted to the Mustguseal server in Mode 2.

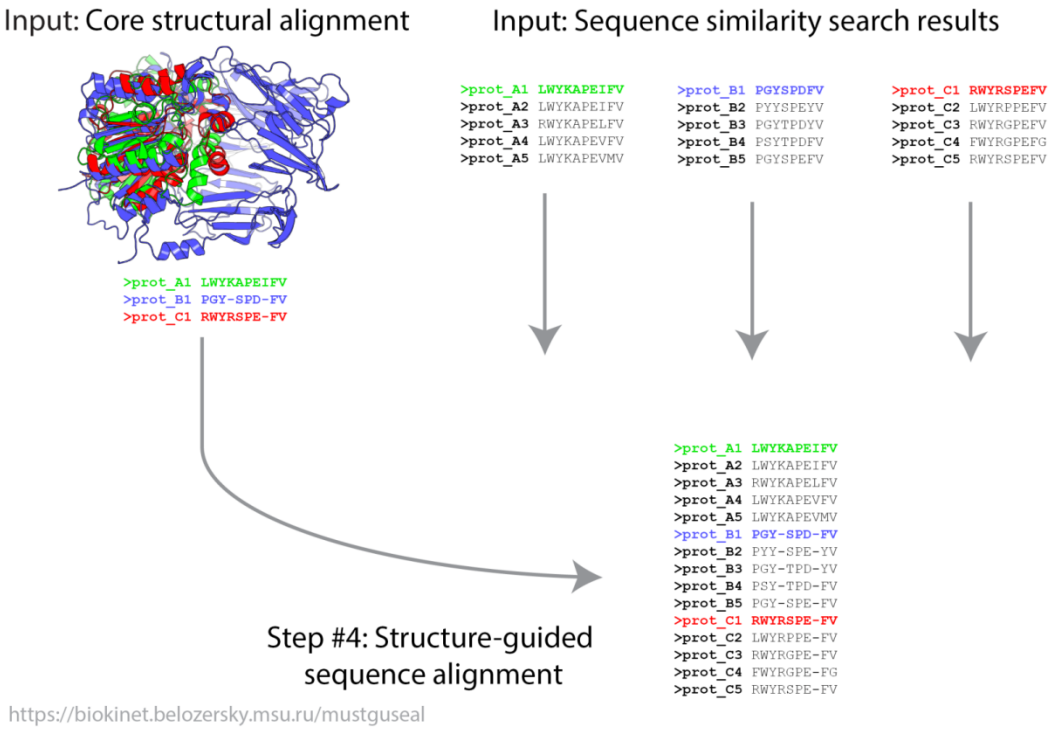
Mustguseal: the Input

Mode 2: Submit a core structural alignment



Mustguseal: the Input

Mode 3: Submit a core structural alignment and results of sequence similarity search



User may alter the results of sequence similarity search (i.e. choose different proteins or change the way sequences are being superimposed within each group) and submit all building blocks of the alignment as new task in Mode 3

Mustguseal: the Output

- **Download section**

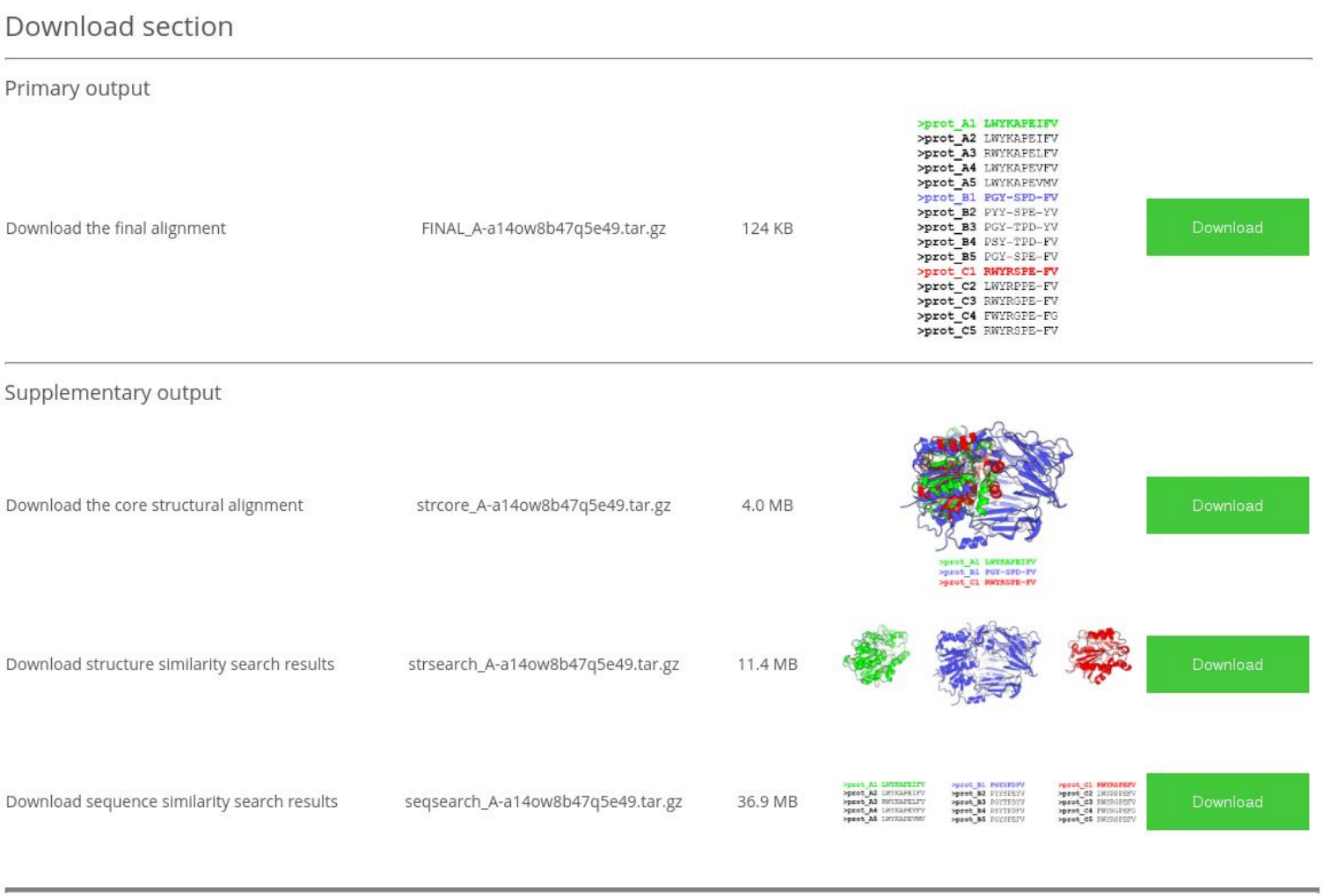
  Download the final alignment and supplementary output to your computer;

- **Analysis section**

  View basic alignment statistics and use interactive on-line tools for sequence and structure analysis.

Mustguseal: the Output

Download section



Mustguseal: the Output

Analysis section: Basic alignment statistics

Basic alignment statistics

General alignment statistics:

Total number of proteins in the final alignment: 5211

Total number of columns in the final alignment: 3831

Protein length statistics:

Protein length average:	311 aa
Protein length minimum:	223 aa
Protein length maximum:	463 aa

Alignment coverage statistics:

Number of columns with at most 0% of gaps:	13	(4% of the average protein length)
Number of columns with at most 5% of gaps:	230	(74% of the average protein length)
Number of columns with at most 30% of gaps:	253	(81% of the average protein length)
Number of columns with at most 50% of gaps:	275	(88% of the average protein length)

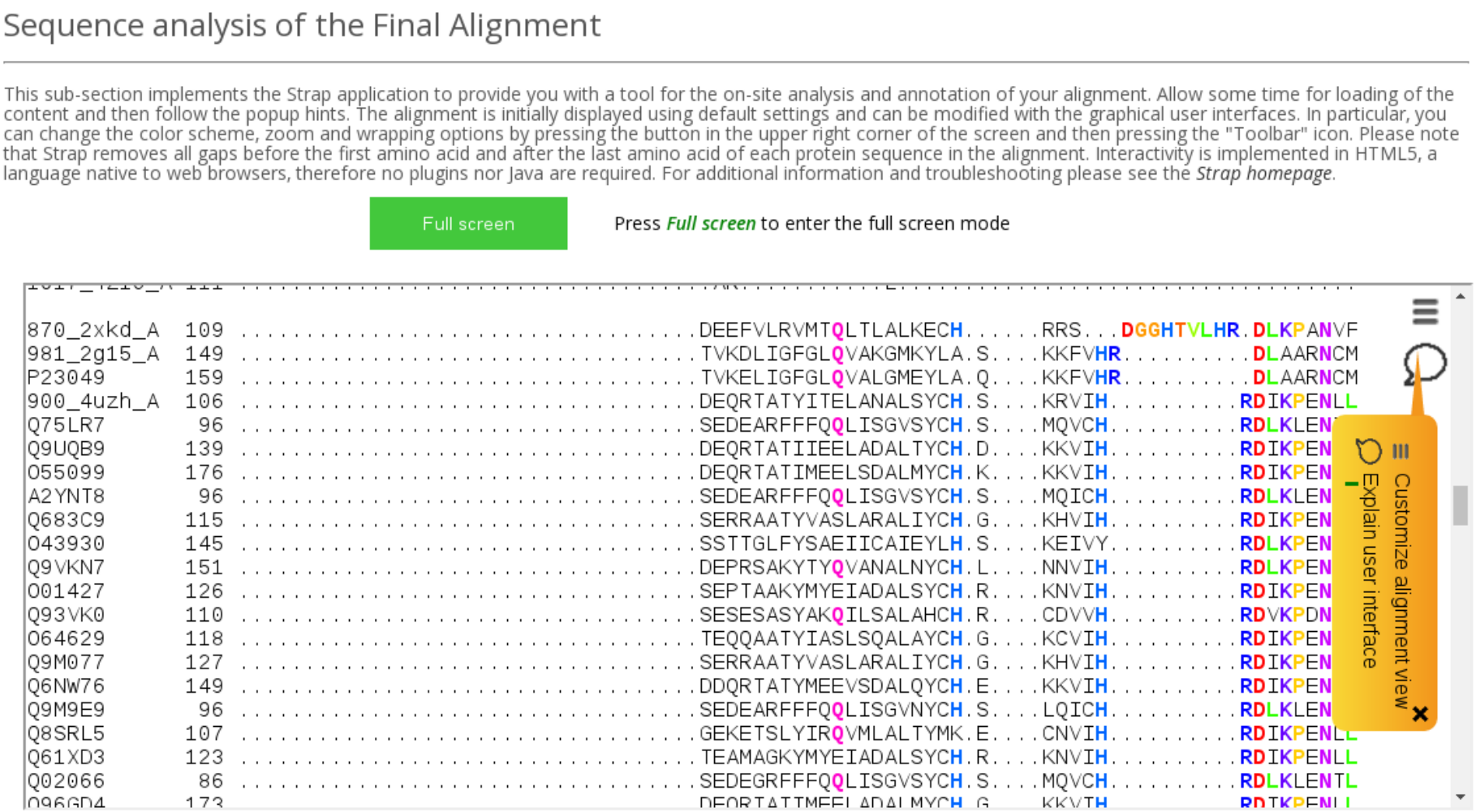
Column conservation statistics:

Number of columns with conservation index at least 100% :	0	(0% of the average protein length)
Number of columns with conservation index at least 95% :	3	(1% of the average protein length)
Number of columns with conservation index at least 75% :	14	(5% of the average protein length)
Number of columns with conservation index at least 50% :	69	(22% of the average protein length)

The conservation index for each column is calculated as the occurence of the most frequent amino acid

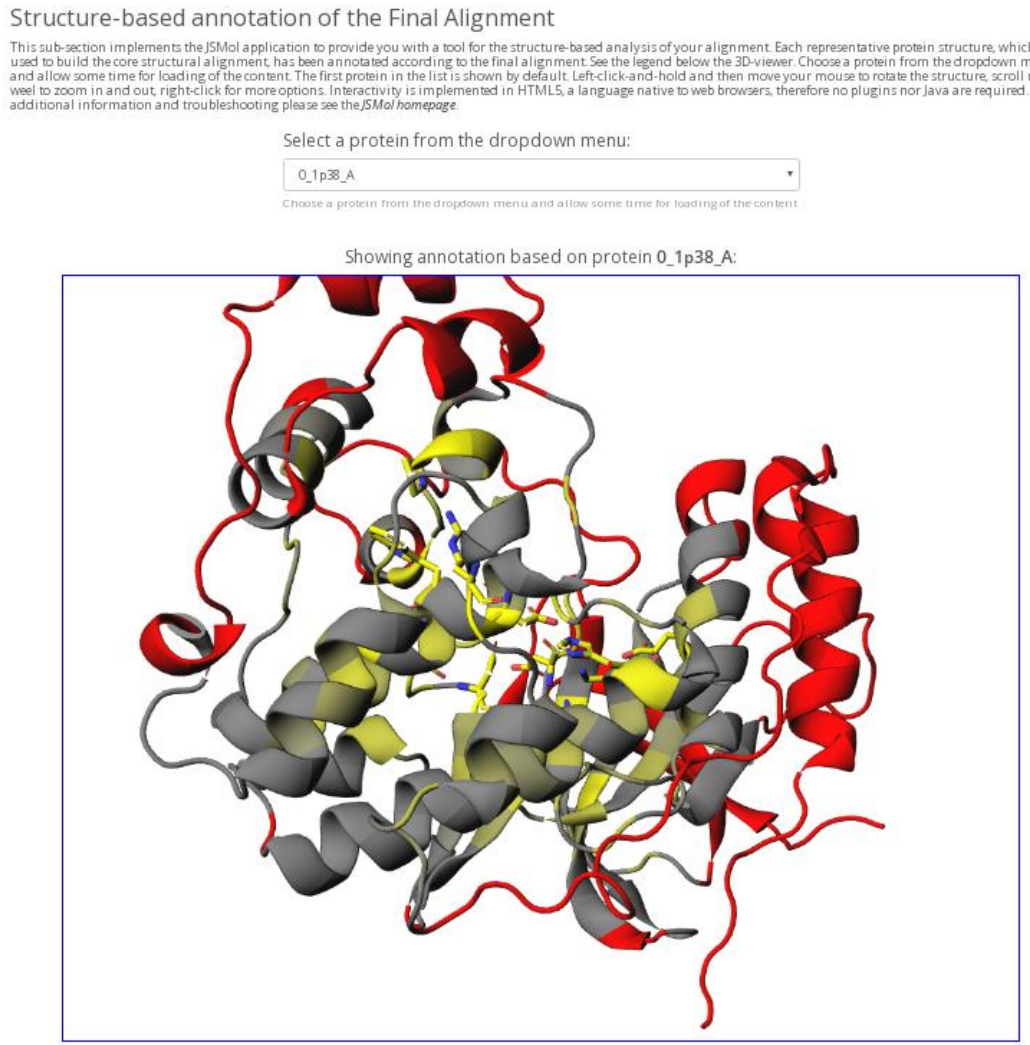
Mustguseal: the Output

Analysis section: Sequence analysis of the Final Alignment



Mustguseal: the Output

Analysis section: Structure-based annotation of the Final Alignment



The selected protein structures are annotated according to basic statistics of the final alignment:

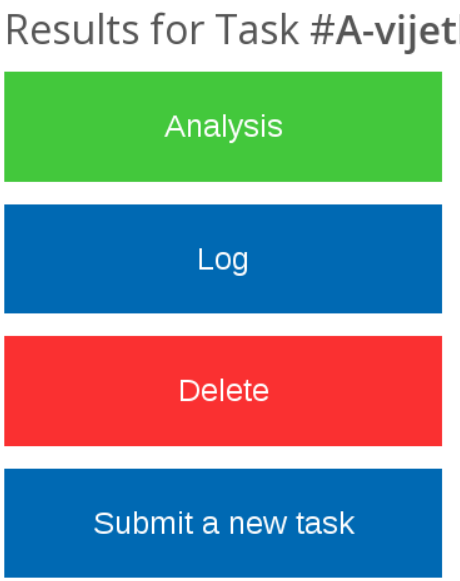
- The most conserved residues are colored in yellow and shown as sticks;
- The gradient paint of the protein backbone corresponds to sequence conservation in the corresponding position (yellow – highly conserved, grey – variable);
- Red paint corresponds to positions that contain more than 5% of gaps in the final alignment.

Mustguseal: File Sharing

Your task #A-rmw94xafb83ta8 has been submitted.
Redirecting ...

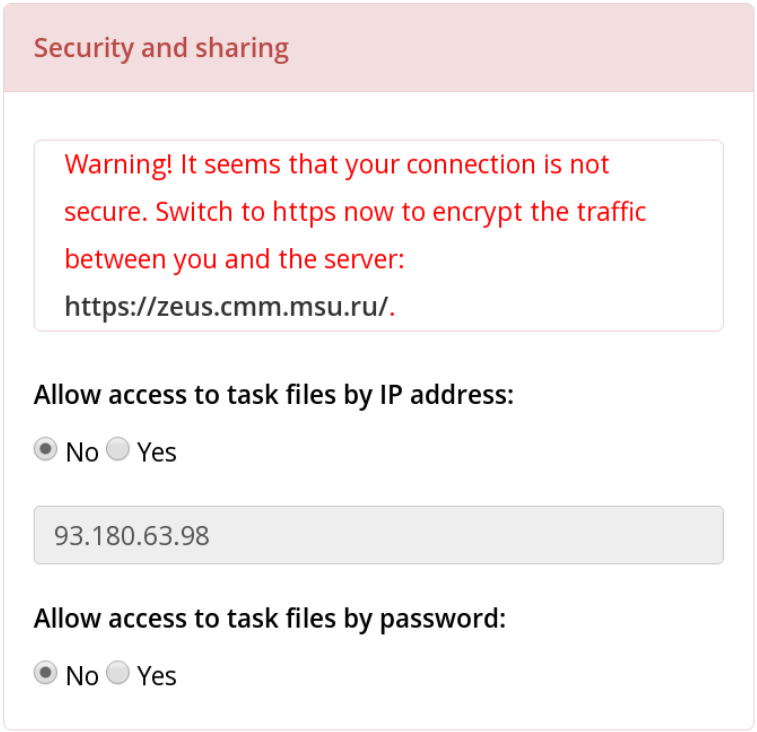
- At the time of submission a new task is assigned a unique 16-symbol access code – **TaskID**;
- TaskID can be used to access results on the Mustguseal server at any time;
- TaskID can be sent to a colleague to share the results.

Mustguseal: File Sharing



At any time the submission can be retracted by pressing the "**Cancel**”/“**Delete**” button at the top of every page - this will stop the task processing and delete the input data as well as any intermediate data and results that have been created.

Mustguseal: Security



Mustguseal offers the following security features to protect your data:

- Data transfer to and from the server can be **encrypted** with a signed certificate using HTTPS protocol;
- **IP-based authentication** can be installed to restrict access to the input data and the results;
- **Password-based authentication** can be installed to restrict access to the input data and the results.

The Mustguseal publication

Suplatov D.A. et al. (**2018**) *Bioinformatics*, 10.1093/bioinformatics/btx831.

Supplementary Data are available from the authors or at *Bioinformatics* online

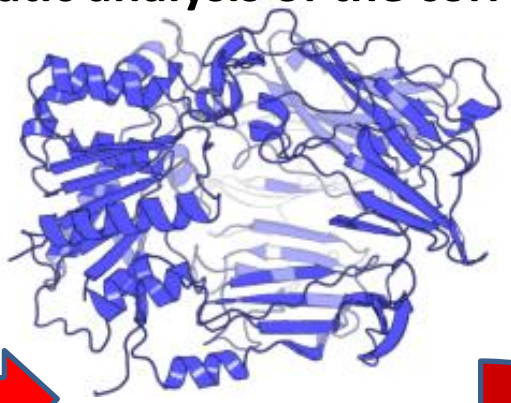
Please read the "Examples" section in the Supplementary Data file for a quick summary of the Mustguseal`s capabilities. Please read "The Protocol", "The Input Modes", and "The Parameters" sections in the Supplementary Data file for an overview of the parameters selection.

Advanced tools
to further study the Mustguseal alignment

The final alignment of protein families can be further submitted to sister services of Mustguseal to analyze **conserved**, **subfamily-specific** and **co-evolving** residues at studying a protein function and regulation, designing improved enzyme variants for practical application and selective modulators of enzyme functional properties.

The key concept

is to study the structure-function relationship of a particular protein by systematic bioinformatic analysis of the corresponding superfamily

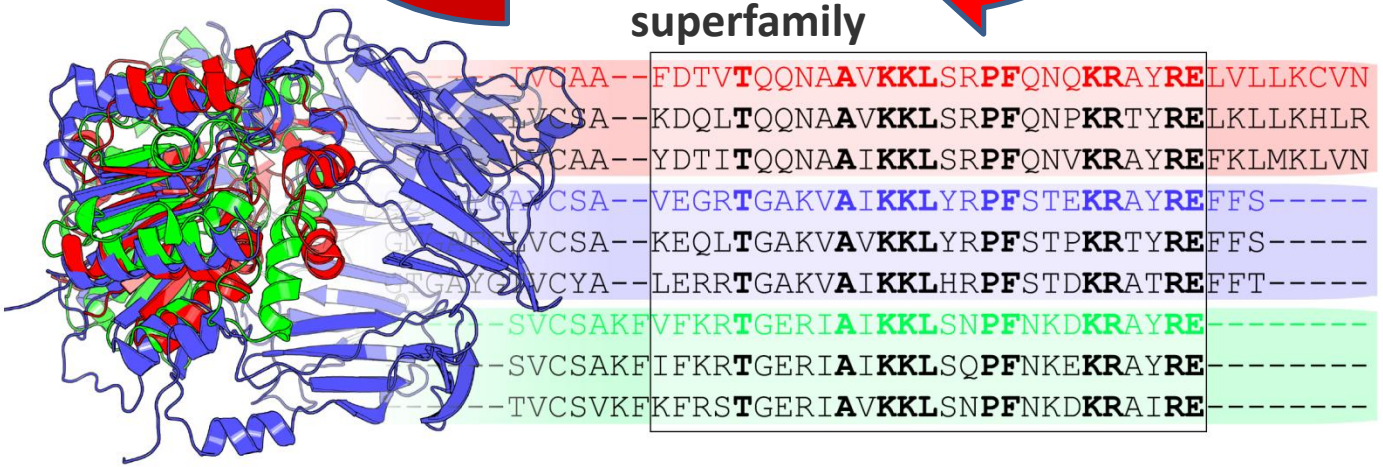


Protein of interest

Conclusions regarding the sequence/structure-function relationship in the particular protein

Collection and analysis of all the available sequence and structural data corresponding to the superfamily

Corresponding superfamily

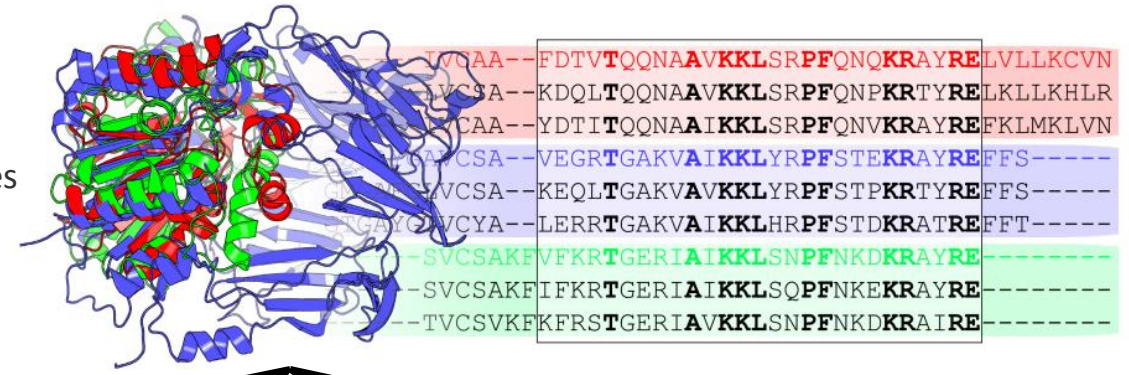


Open-access on-line platform for bioinformatic analysis in computational enzymology

Mustguseal

can automatically collect from public databases and align thousands of sequences and structures of proteins within a superfamily to produce a large structure-guided sequence alignment

Bioinformatics, 2018



Zebra

To identify and rank the subfamily-specific positions as determinants of functional diversity and binding specificity

J Biomol Struct Dyn. , 2014

pocketZebra

To identify and rank binding sites in proteins by functional significance and select particular positions in the structure that are important for selective binding of substrates and ligands

Nucleic Acids Research, 2014

visualCMAT

To predict and visualize correlated mutations/co-evolving residues in protein structures as a mechanism of allosteric communication, and a source of compensatory mutations for rational design

J Bioinform Comput Biol., 2018

Yosshi

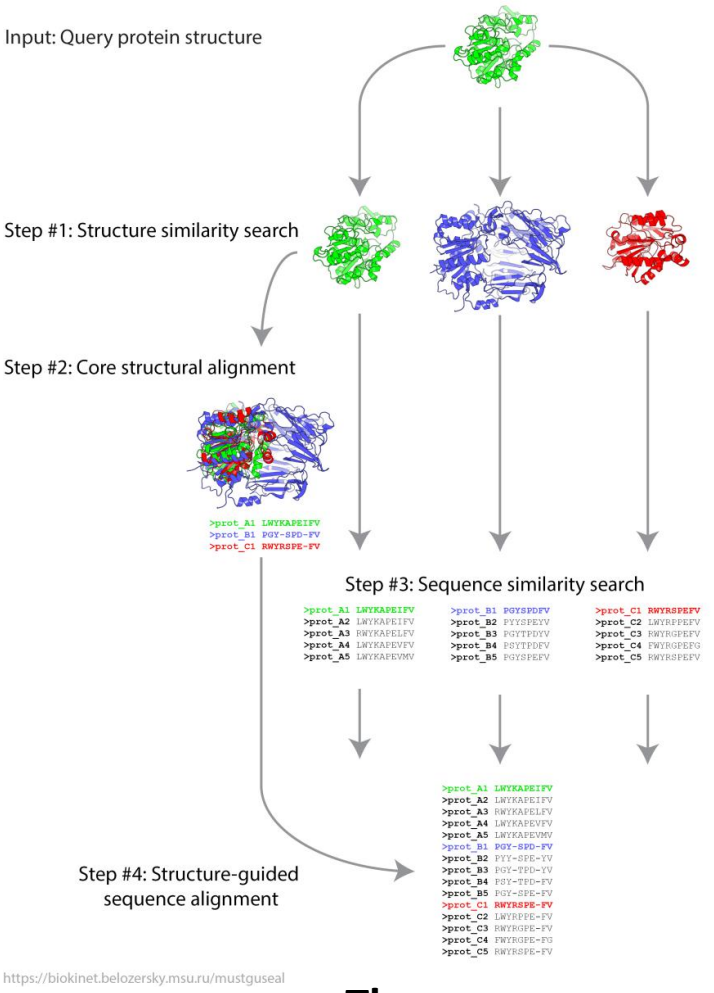
To systematically classify and study disulfide bonds in diverse protein families, and to assist at selecting hot-spots for disulfide engineering

Nucleic Acids Research, 2019

https://biokinet.belozersky.msu.ru/m-platform

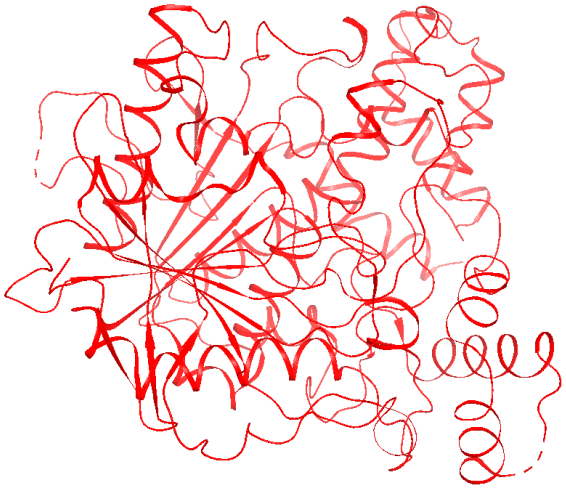
Automatic construction of a large structure-guided sequence alignment of your protein family by the Mustguseal

The input:
PDB structure of human acetylcholinesterase

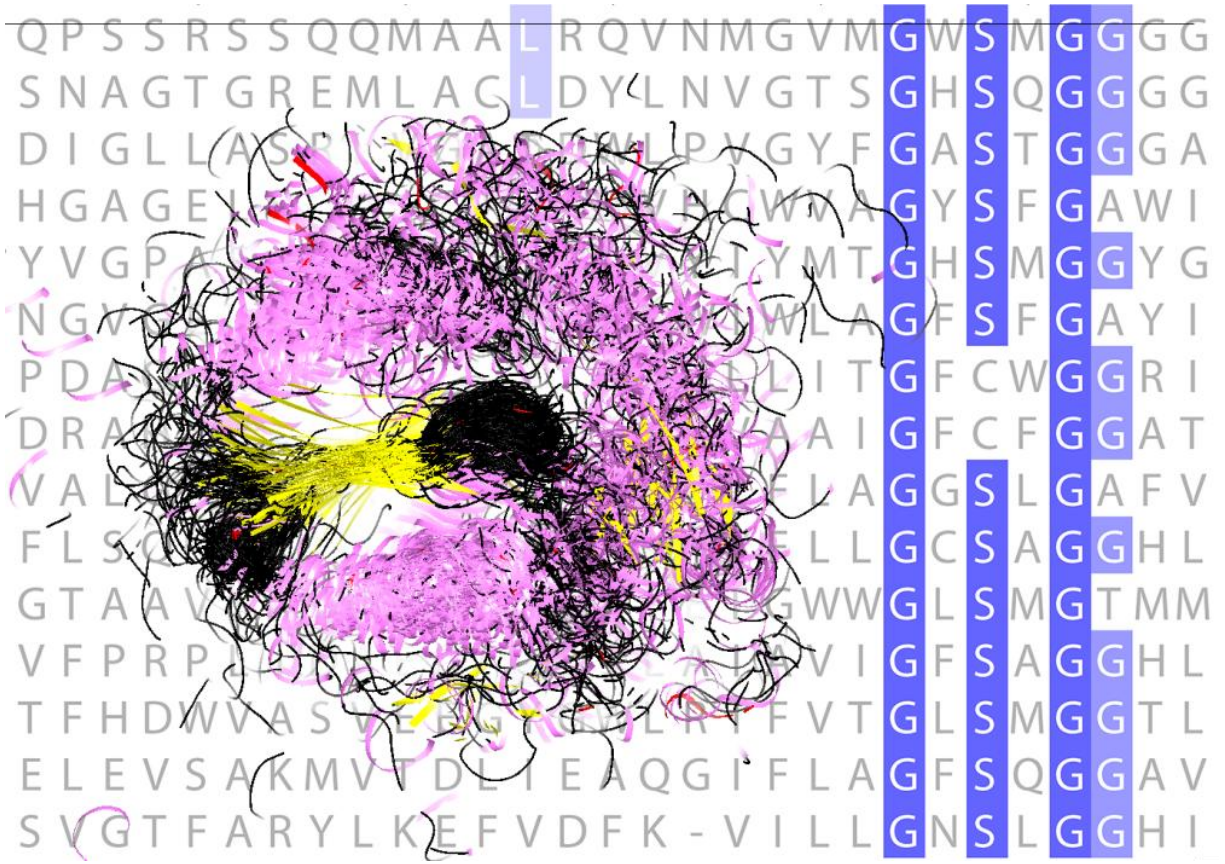


The process:
Automatic collection and alignment of all the available protein sequences and structures from public databases

Automatic construction of a large structure-guided sequence alignment of your protein family by the Mustguseal

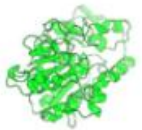


The input:
PDB structure of human acetylcholinesterase



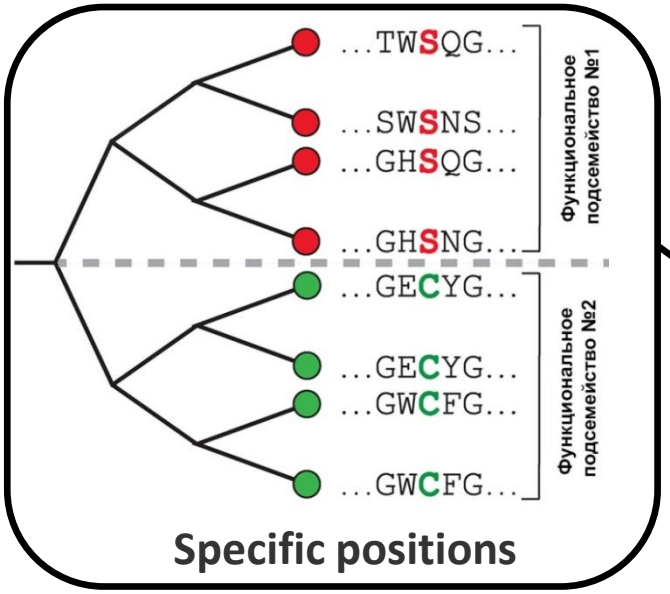
The output:
Structure guided-sequence alignment of human acetylcholinesterase and its homologs from the α/β-hydrolase superfamily

Submit the final Mustguseal alignment for further analysis

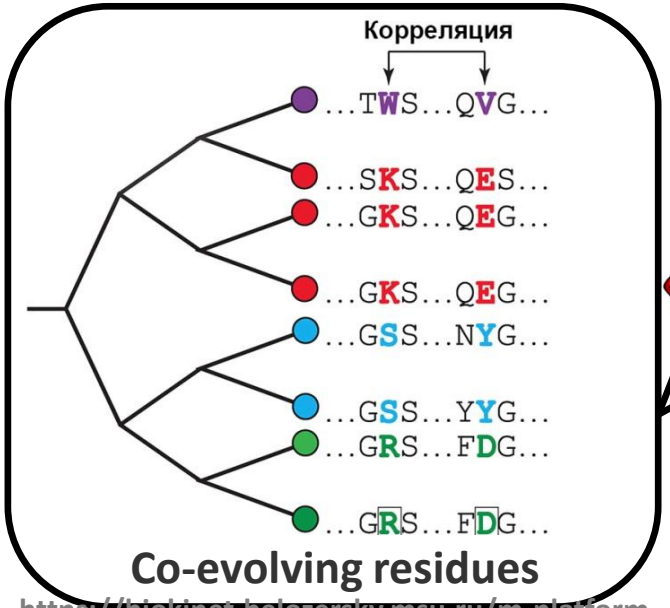
Submit the Final Mustguseal alignment	FINAL_A-ditxcb9jvnyuqz.fasta	12.0 MB		Submit to **Zebra**
Submit the Final Mustguseal alignment and a PDB structure of the representative protein	FINAL_A-ditxcb9jvnyuqz.fasta	12.0 MB		Submit to **Zebra**
				Submit to **pocketZebra**
	0_1u8f_O.pdb	166 KB		Submit to **visualCMAT**

- A new submission to Zebra, pocketZebra, visualCMAT, and Yosshi can be made directly from the Mustguseal Results page.

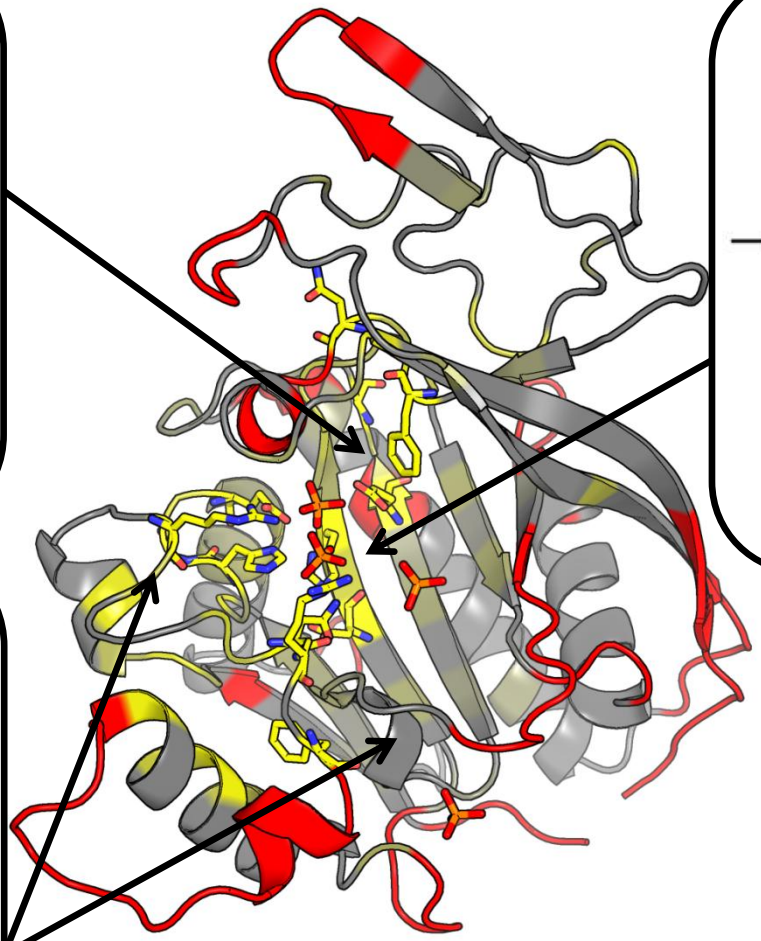
Annotation of the protein of interest according to the bioinformatic analysis of the superfamily

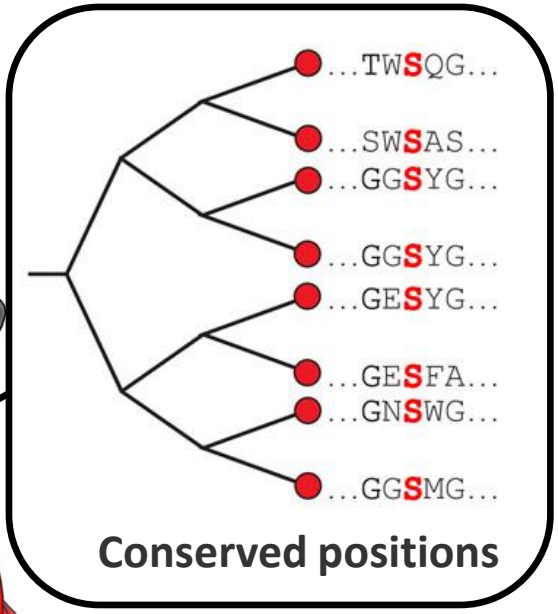


Specific positions



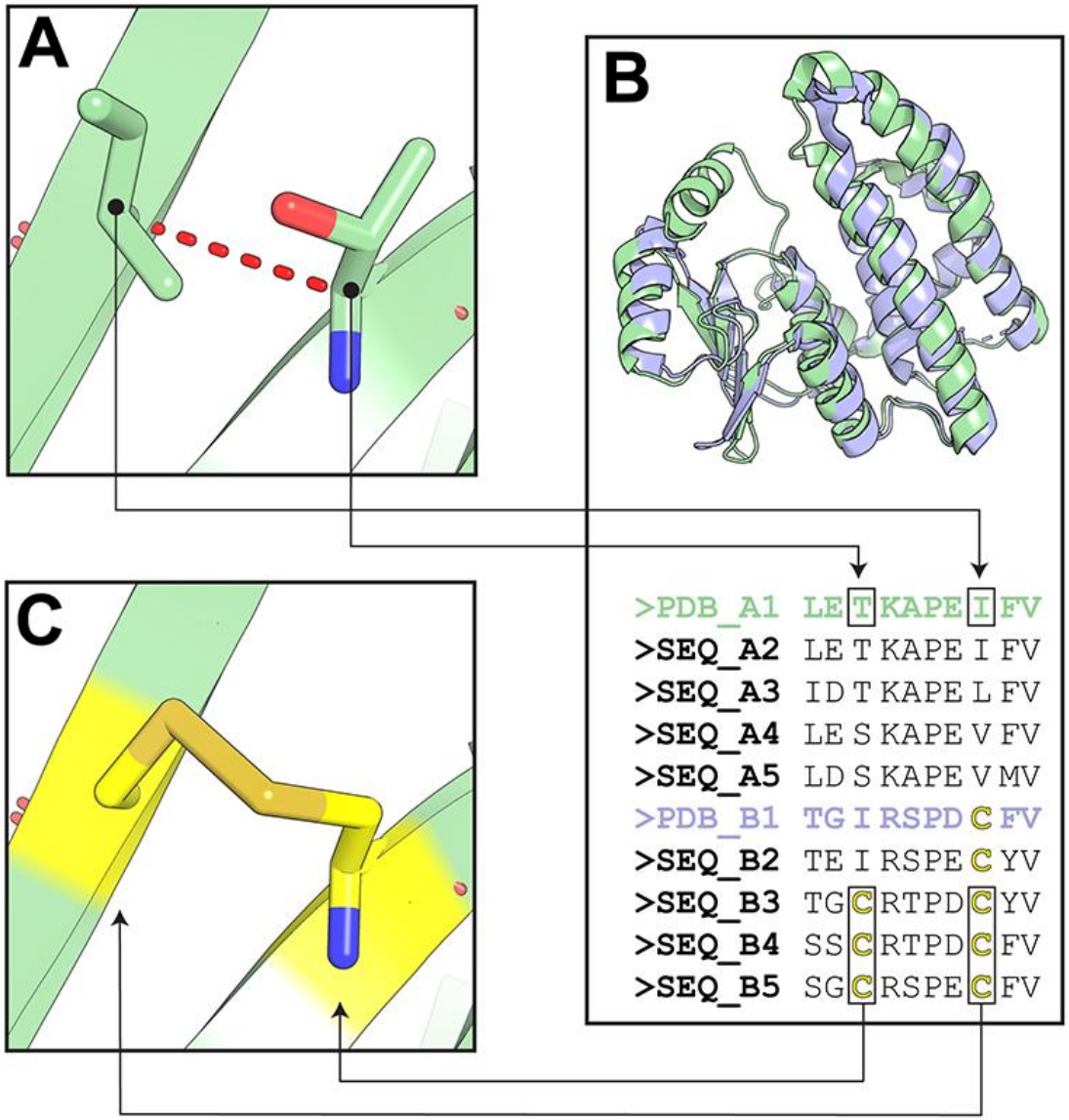
Co-evolving residues





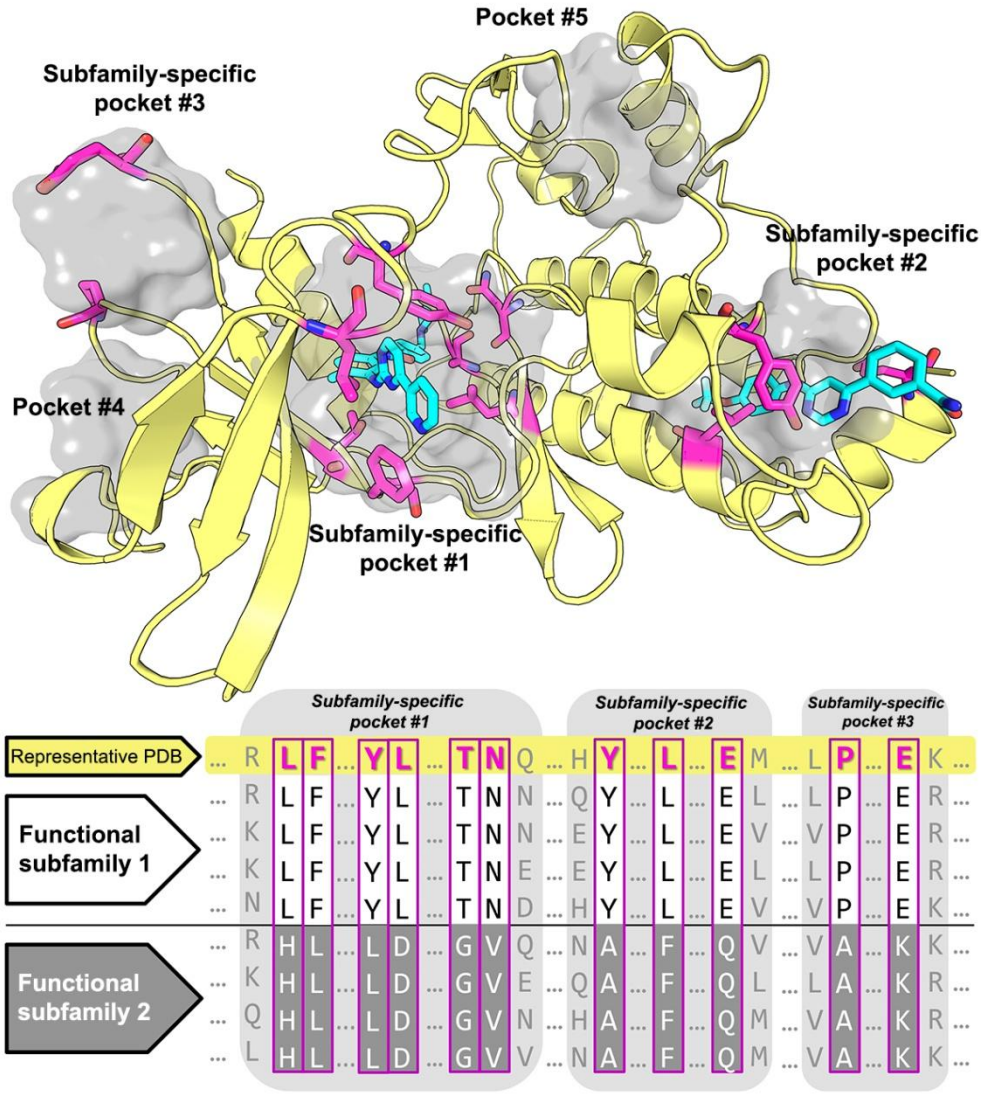
Conserved positions

Systematically classify and study disulfide bonds in diverse protein families



The **Yosshi** web-service can be used to systematically classify and study disulfide bonds in diverse protein families, and to assist at selecting hot-spots for disulfide engineering in the structure of your query protein. The "**YO**ur web-server for **S-S** bond **H**arvest**I**ng" is a new highly automated on-line tool for a systematic homology-driven analysis and engineering of disulfide bonds that can be easily used by a general biologist at a daily laboratory routine. The Yosshi facilitates a broader interpretation of disulfides not just as a factor of structural stability, but rather as a mechanism to implement diversity within a superfamily; Suplatov D., et al. (2019) *Nucl. Acids Res.*

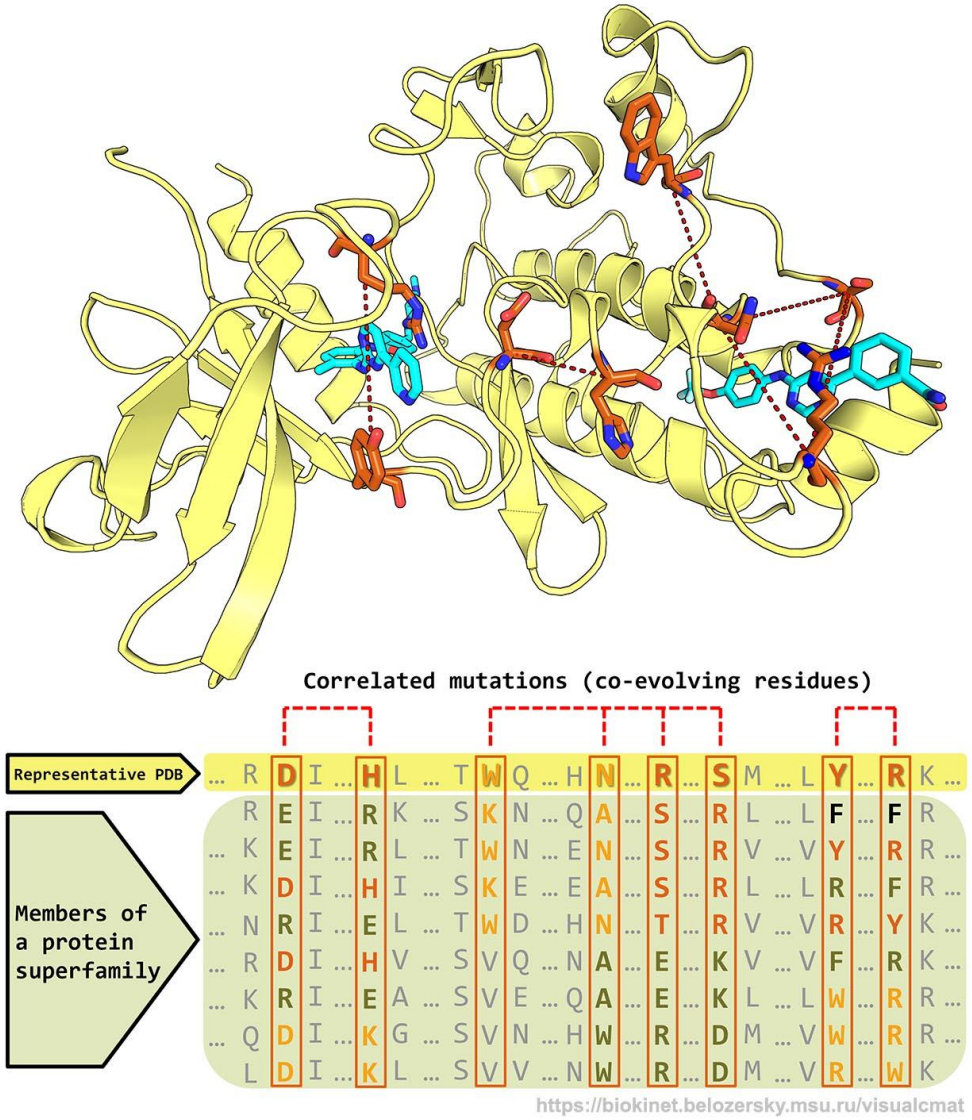
Identify and study the conserved and subfamily-specific positions



- The **Zebra** web-service can be used to identify conserved positions, which define properties common among functionally diverse protein families, as well as subfamily-specific positions responsible for functional diversity – to select hotspots for directed evolution or rational design experiments; Suplatov D., et al. (2014) *J.Biomol.Struct.Dyn*.

- The **pocketZebra** web-service can be used to identify and rank binding sites in proteins by functional significance and select particular positions in the structure that are important for selective binding of substrates/inhibitors/effectors; Suplatov D., et al. (2014) *Nucl. Acids Res.*

Predict and visualize the correlated mutations (co-evolving residues)



- The **visualCMAT** web-service can be used to predict and visualize correlated mutations/co-evolving residues in protein structures. The visualCMAT can be used to understand the relationship between structure and function and identify co-evolution patterns in protein superfamilies, implemented at selecting hotspots and compensatory mutations for rational design and directed evolution experiments to produce novel enzymes with improved properties, and employed at studying the mechanism of selective ligand's binding and allosteric communication between topologically independent sites in protein structures;
  Suplatov D., et al. (2018) *J Bioinform Comput Biol*.

Structure-based sequence alignments of functionally diverse protein families as a tool in protein engineering and drug discovery

Implementation of Mustguseal, Zebra, pocketZebra, and visualCMAT in the laboratory practice can help at studying a protein function and regulation, designing improved enzyme variants for practical application and selective modulators of enzyme functional properties.

Suplatov, D., Kirilin, E., & Švedas, V. (**2016**). Bioinformatic Analysis of Protein Families to Select Function-Related Variable Positions. In *Understanding Enzymes: Function, Design, Engineering, and Analysis* (pp. 351-385) Ed. Allan Svendsen. Pan Stanford. [link]

Suplatov, D., Voevodin, V., & Švedas, V. (**2015**). Robust enzyme design: Bioinformatic tools for improved protein stability. *Biotechnology journal*, 10(3), 344-355. [link]

Suplatov, D., & Švedas, V. (**2015**). Study of functional and allosteric sites in protein superfamilies. *Acta Naturae*, 7(4), 27, 34-45. [link]

Contacts

- The Mustguseal title page
  https://biokinet.belozersky.msu.ru/mustguseal
- Mustguseal Support
  d.a.suplatov@belozersky.msu.ru
- Collaboration
  vytas@belozersky.msu.ru
- Press to ask your question on-line

Ask for help

Lomonosov Moscow State University, the Main building