

Sequence analysis

Mustguseal: a Server for Multiple Structure-Guided Sequence Alignment of Protein Families

D.A. Suplatov¹, K.E. Kopylov², N.N. Popova³, V.I.V. Voevodin^{3,4}, V.K. Švedas^{1,2*}

¹Belozersky Institute of Physicochemical Biology, ²Faculty of Bioengineering and Bioinformatics, ³Faculty of Computational Mathematics and Cybernetics, ⁴Research Computing Center of the Lomonosov Moscow State University, Vorobjev hills 1-40, Moscow 119991, Russia

The Supplementary materials are provided in the order they are being referred to in the Main text

The Protocol

The Mustguseal protocol contains four major steps (see Fig. 1 in the Main text). First, the structure similarity search using the popular SSM algorithm (Krissinel and Henrick, 2004) is implemented for the submitted query protein to collect evolutionarily remote relatives that are expected to represent different protein families (Step 1). The selection of hits is based on the percentage of secondary structure equivalences (by default at least 70% of the query have to make at least 70% of the target). The discovered proteins are filtered by sequence identity to select the non-redundant core collection of at most N_{max} representative proteins as follows (the value of N_{max} depends on the input mode, see The Input Modes below). The initial set of proteins collected by the structure similarity search is clustered at the 95% pairwise sequence identity threshold, and one representative member is automatically selected from each cluster by the CD-HIT tool (Fu et al., 2012). If the number of selected representative proteins exceeds N_{max} they are further clustered at the 90% pairwise sequence identity threshold, then, if necessary, at 85%, 80%, etc., with a 5% decrement. If the number of representative proteins collected by applying the 40% sequence identity threshold exceeds N_{max} , then the remaining proteins are ranked by the pairwise structure similarity Q-score with the query structure, which is provided by the SSM, and the first N_{max} structures (i.e., the most structurally similar to the query) are finally selected. The 40% cut-off is the smallest sequence identity threshold permitted by the CD-HIT tool and marks the edge of the “safe zone” of a protein sequence comparison (Rost, 1999).

Superposition of the collected structures of representative proteins by the MATT algorithm on Step 2 is used to create the core structural alignment (Menke et al., 2008). MATT searches for compatible pairs of fragments and permits structural allowances such as twists and translations, and thus demonstrates good performance in aligning distant relationships and length variations (Kalaimathy et al., 2011).

Then, on Step 3, each representative protein is independently used as a query to run a sequence similarity search (Altschul et al., 1990; Vouzis and Sahinidis, 2011) and collect evolutionarily close relatives – members of the corresponding families – from either the Swiss-Prot database or Swiss-Prot+TrEMBL databases. By default, at most 500 proteins are collected in each sequence similarity search. Filters are further applied within each sequence collection to eliminate redundant entries and outliers: first, the sequence length filter is applied – i.e., by default, proteins which differ by more than 20% in length from the representative protein, used as a query to run the sequence similarity search, are removed to eliminate too small and too large protein sequences, which can correspond to incomplete or incorrect database entries, and thus to decrease the number of columns which are highly gapped in the alignment (see a special case when this filter should be released in The Parameters below); then, similar sequences are removed by the redundancy filter – i.e., by default, all sequences are clustered at the 95% pairwise sequence identity threshold and one representative member is automatically selected in each cluster by the CD-HIT tool; finally, the dissimilarity filter is applied – i.e., by default, sequences sharing less than 0.5 bit score per column with the query representative protein are removed to eliminate too distant proteins which can cause errors during the sequence alignment (as described in Fischer et al., 2008 using the software by Söding et al., 2005). The protein sets which were collected from independent sequence similarity searches (i.e., with different representative proteins used as queries) are compared for redundant entries, by default, at the 95% pairwise sequence identity threshold, and each such entry is preserved only in one protein set that corresponds to a representative protein with the largest pairwise sequence identity to that entry (i.e., this guarantees that there will be no redundant protein sequences in the final alignment). The sequences within each set are then independently aligned by the MAFFT routine (Katoh and Standley, 2013).

Finally, on Step 4, the alignment of sequences of the evolutionarily remote relatives is created by using the structural alignment of the representative proteins as a guide. During this step the superimpositions built by MATT and MAFFT on Steps 2 and 3 do not change. Columns of gaps are inserted into individual sequence alignments in a way that their total lengths become equal and the superposition of the representative proteins in the merged sequence alignments matches their superimposition in the core structural alignment.

The Input Modes

The Mode 1 is the default, fully automated, and the easiest way to obtain an alignment by submitting PDB and chain IDs of a query protein. Large structure-guided sequence alignments of functionally diverse protein families that include thousands of proteins basing on all available information about their structures and sequences in public databases can be constructed using the Mustguseal web-server in Mode 1 (see the corresponding Examples presented below). Modes 2 and 3 provide an opportunity to customize an alignment for a particular purpose – e.g., the user can submit a custom core structural superimposition to build a multiple structure-guided sequence alignment of only selected protein families and their closest homologs in Mode 2, or edit the automatically created sequence alignment sets and then use Mode 3 to merge them together.

The Mustguseal web-server

The on-line Mustguseal tool limits the size of the core collection (i.e., N_{max} – the number of representative protein structures) to optimize the performance at Steps 2 (i.e., the structural alignment) and 3 (i.e., the sequence similarity search) of the protocol. By default, the value of N_{max} in Mode 1 (i.e., the Swiss-Prot database is selected to perform the sequence similarity search) is set to 32 structures. If the Swiss-Prot+TrEMBL databases were selected by the user, then the N_{max} is reduced to 16 structures due to a higher computational cost of sequence similarity searches in the TrEMBL database, which are performed for each representative protein (see The Protocol above). In Mode 2 the user can submit a custom core structural alignment of at most 64 proteins to perform automatic sequence similarity searches in the Swiss-Prot database, or at most 32 proteins, if the Swiss-Prot+TrEMBL databases were selected. There are no limitations on the size of a final alignment in Modes 1 and 2. Finally, in Mode 3 the user can submit a custom core structural alignment limited to at most 150 proteins, and the corresponding sequence alignment sets of at most 15 000 proteins in total.

Technical documentation describing the input format requirements to operate in Modes 1, 2, and 3, as well as the guidelines for compatibility with other on-line tools are provided in the on-line user manual at <https://biokinet.belozersky.msu.ru/mustguseal-input>. Technical documentation describing the Mustguseal output in Modes 1, 2, and 3 is provided in the on-line user manual at <https://biokinet.belozersky.msu.ru/mustguseal-results>.

The Parameters

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 (i.e., in Mode 1) or submitted by the user (i.e., in Modes 2 or 3). In Mode 1 the user can include more (or less) evolutionarily distant relatives in the alignment by requesting higher (or lower) percentage of secondary structure equivalences between query and target at the structure similarity search. No particular values for these parameters can be recommended in advance: they should be selected based on the user's research objective, structural organization of the protein family of interest, and availability of the corresponding data in public databases. The 90% and 90% setup (i.e., at least 90% of the query have to make at least 90% of the target) can be used to collect proteins from the PDB database with a highly conserved structure but originating from different organisms. The minimum percentage of secondary structure equivalences can be set to 70% and 70% to incorporate more evolutionarily remote and functionally diverse proteins into the alignment which still share a sufficient structural similarity to produce a meaningful superimposition. The 70%-70% thresholds for secondary structure equivalences are used by default in the Mustguseal web-server and in the PDBeFOLD web-server which also implements the SSM algorithm to search for structurally similar proteins in the PDB database (Krissinel and Henrick, 2004). A particular example implementing these thresholds is discussed below (see Examples). The users should note that at least some level of similarity may be found in almost any pair of protein structures picked at random from the PDB database. Therefore, a decrease of the two thresholds to a very low values (e.g. 30% and 30%) may help to identify all available members of a superfamily of interest, but increases the probability to collect unrelated (i.e., not homologous) proteins which cannot be reasonably compared by the MATT structural alignment. Finally, non-symmetrical thresholds can be used to collect proteins with a different domain composition. E.g., sialidase from *Trypanosoma rangeli* (PDB 1MZ6) and Human sialidase Neu2 (PDB

2F0Z) both contain a structurally similar catalytic domain assigned to the GH33 family of the CAZy classification (Lombard et al., 2013). However, in the first case the GH33 catalytic domain is attached to a CBM40-like lectin domain within a single polypeptide chain, and the second protein contains only the catalytic domain. The two homologous proteins share 42% and 77% secondary structure equivalences, respectively (i.e., 77% of the Human protein's structure makes only 42% of the trypanosomal protein's structure) due to an additional lectin domain in the sialidase from *Trypanosoma rangeli*. Consequently, to collect the 1MZ6 structure when using the 2F0Z structure as a query in Mode 1 the structure similarity search thresholds should be set to at most 77% for the query and at most 42% for the target.

The sequence similarity search parameters can also be changed. The user can choose to perform sequence similarity searches either in the Swiss-Prot database or deal with the much larger dataset basing on Swiss-Prot+TrEMBL databases. The pre-calculated non-redundant sets of each database clustered at the 100% (nr100), 95% (nr95) and 80% (nr80) sequence identity thresholds are actually being used by the server to accelerate the sequence similarity searches. If the redundancy filter threshold T is set by the user within a range $95\% < T \leq 100\%$, then the nr100 database set is used; for $80\% < T \leq 95\%$ – the nr95 database set; finally, for $T \leq 80\%$ – the nr80 database set. E.g., if the user-defined redundancy filter threshold is set to 90%, then the nr95 set of the selected databases (i.e., Swiss-Prot or Swiss-Prot+TrEMBL) is used to run a sequence similarity search, at most 500 sequences are collected and further processed by the sequence length filter, the redundancy filter at the 90% sequence identity threshold, and the dissimilarity filter (see the Protocol above). If the protein family/superfamily of interest has a limited representation in the PDB database and therefore the desired diversity cannot be achieved by the structure similarity search alone, then the dissimilarity filter threshold can be decreased to 0.25 bit score per column (i.e., the possibility of using this particular value of the filter to construct alignments of protein families was shown by Fischer et al., 2008, see the discussion of the Benchmark sets as a Supplementary material to that paper) accompanied by a decrease of the redundancy filter threshold to 80% (to implement a more diverse nr80 database set) and an increase of the maximum number of proteins to be collected from each sequence similarity search to 1000 – to incorporate more diverse proteins in the sequence alignment. Setting the dissimilarity filter threshold parameter to lower values can result in selecting too distant proteins which can cause errors at the MAFFT sequence alignment stage. Finally, the sequence length filter should be released, if the representative protein structures correspond to protein fragments or proteins which are represented by a larger precursor sequences in the Swiss-Prot/TrEMBL databases. E.g., the neuraminidase A (NanA) from *Streptococcus pneumoniae* contains lectin, catalytic, and membranophylic domains within a single polypeptide chain, but the PDB structure 2YA8 contains only the catalytic domain of NanA. Consequently, if the 2YA8 structure is included into the core structural alignment (and, consequently, further used as a query to run a sequence similarity search), then the sequence length filter should be set to 121% (or to any larger value) as the full length of the NanA's polypeptide chain is 221% of the length of the catalytic domain presented in the PDB 2YA8.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W., and Lipman,D.J. (1990). Basic local alignment search tool. *J.Mol.Biol.*, **215**(3), 403-410

Fischer,J.D., Mayer,C.E., and Söding,J. (2008). Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*, **24**(5), 613-620.

The Mustguseal web-server

- Fu,L., Niu,B., Zhu,Z., Wu,S., and Li,W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28(23)**, 3150-3152.
- Kalaimathy,S., Sowdhamini,R., and Kanagarajadurai,K. (2011). Critical assessment of structure-based sequence alignment methods at distant relationships. *Brief. Bioinformatics*, **12(2)**, 163-175.
- Katoh,K. and Standley,D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30(4)**, 772-780.
- Krissinel,E. and Henrick,K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.*, **60(12)**, 2256-2268.
- Lombard,V., Golaconda Ramulu,H., Drula,E., Coutinho,P.M. and Henrissat,B. (2013) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.*, **42(D1)**, D490-D495
- Menke,M., Berger,B., and Cowen,L. (2008). Matt: local flexibility aids protein multiple structure alignment. *PLoS Comput Biol*, **4(1)**, e10.
- Rost,B. (1999). Twilight zone of protein sequence alignments. *Protein engineering*, **12(2)**, 85-94.
- Söding,J., Biegert,A., and Lupas,A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33(S2)**, W244-W248.
- Vouzis,P.D. and Sahinidis,N.V. (2011). GPU-BLAST: using graphics processors to accelerate protein sequence alignment. *Bioinformatics*, **27(2)**, 182-188.

Continued on the next page

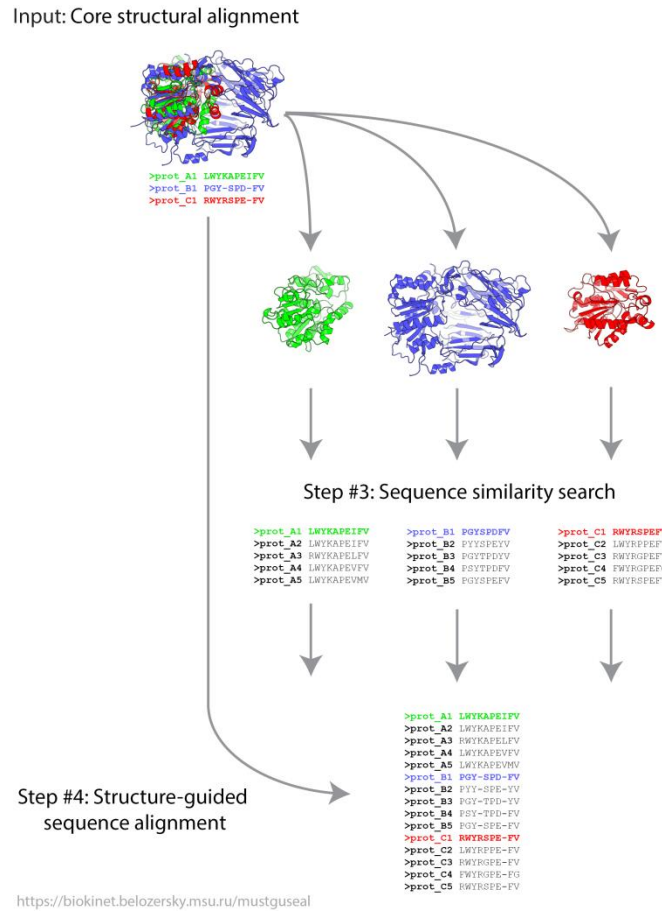


Fig. S1. The outline of the Mustguseal protocol in Mode 2.

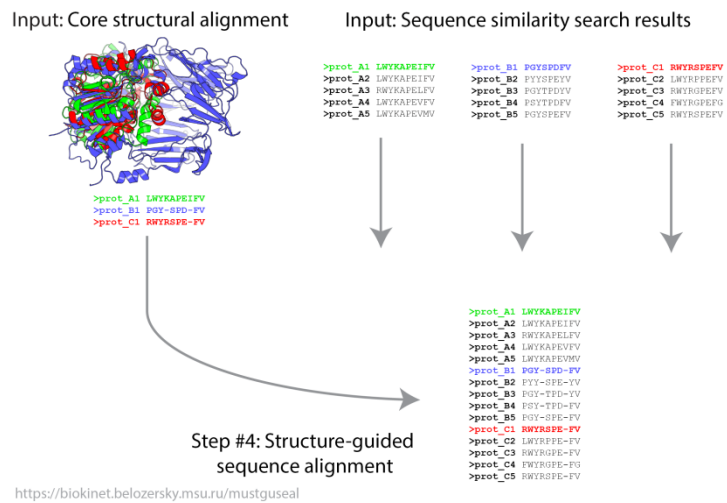


Fig. S2. The outline of the Mustguseal protocol in Mode 3.

Download section

This section provides links to the primary output (the Final alignment) and supplementary output (core structural alignment, structure similarity search results, and sequence similarity search results). All files are packed in 'tar.gz' archives. To extract files from a 'tar.gz' archive use the command `tar xzf file.tar.gz` in Linux and in Widows use a free 7-zip tool.

Primary output

Download the final alignment	FINAL_A-ditxcb9jvnyuqz.tar.gz	952 KB	<pre> >spout_A1 LKWKARSLPV >spout_A2 LKWKARSLPV >spout_A3 LKWKARSLPV >spout_A4 LKWKARSLPV >spout_A5 LKWKARSLPV >spout_B1 FVY-SFD-FV >spout_B2 FVY-SFD-FV >spout_B3 FGI-TSD-FV >spout_B4 FGI-TSD-FV >spout_B5 FGI-SFD-FV >spout_C1 RWKRSR-FV >spout_C2 LKWRSS-FV >spout_C3 RWKRSR-FV >spout_C4 FVRSSE-FV >spout_C5 RWKRSR-FV </pre>	Download
------------------------------	-------------------------------	--------	---	----------

Supplementary output

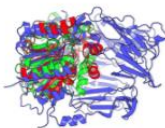

Download the core structural alignment	strcore_A-ditxcb9jvnyuqz.tar.gz	2.2 MB		Download
Download structure similarity search results	strsearch_A-ditxcb9jvnyuqz.tar.gz	6.1 MB		Download
Download sequence similarity search results	seqsearch_A-ditxcb9jvnyuqz.tar.gz	33.2 MB	<pre> >spout_A1 LKWKARSLPV >spout_A2 LKWKARSLPV >spout_A3 RWKRSR-FV >spout_A4 LKWKARSLPV >spout_A5 LKWKARSLPV >spout_B1 RWKRSR-FV >spout_B2 RWKRSR-FV >spout_B3 RWKRSR-FV >spout_B4 RWKRSR-FV >spout_B5 RWKRSR-FV </pre>	Download

Fig. S3. Example of the Download section at the Mustguseal Results Page. This section provides links to the primary output (the final alignment) and the supplementary output (core structural alignment, structure similarity search results, and sequence similarity search results).

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 occurrence of the most frequent amino acid

Fig. S4. Example of the Basic alignment statistics at the Mustguseal Analysis Page.

Sequence analysis of the Final Alignment

This sub-section implements the Strap application to provide you with a tool for the on-site analysis and annotation of your alignment. Allow some time for loading of the content and then follow the popup hints. The alignment is initially displayed using default settings and can be modified with the graphical user interfaces. In particular, you can change the color scheme, zoom and wrapping options by pressing the button in the upper right corner of the screen and then pressing the "Toolbar" icon. Please note that Strap removes all gaps before the first amino acid and after the last amino acid of each protein sequence in the alignment. Interactivity is implemented in HTML5, a language native to web browsers, therefore no plugins nor java are required. For additional information and troubleshooting please see the [Strap homepage](#).

Full screen
Press **Full screen** to enter the full screen mode

870_2xkd_A	109	DEEFVLRVMTQLTLALKECH.....	RRS...DGGHTVLHR.DLKPANVF					
981_2g15_A	149	TVKDLIGFGLQVAKGMKYLA.S.....	KKFVHR.....DLAARNCM					
P23049	159	TVKELIGFGLQVALGMEYLA.Q.....	KKFVHR.....DLAARNCM					
900_4uzh_A	106	DEQRTATYITELANALSYCH.S.....	KRVIH.....RDIKPENLL					
Q75LR7	96	SEDEARFFFQLISGVSYCH.S.....	MQVCH.....RDLKLEN					
Q9UQB9	139	DEQRTATIIIEELADALTYCH.D.....	KKVIH.....RDIKPEN					
O55099	176	DEQRTATIMEELSDALMYCH.K.....	KKVIH.....RDIKPEN					
A2YNT8	96	SEDEARFFFQLISGVSYCH.S.....	MQICH.....RDLKLEN					
Q683C9	115	SERRAATYVASLARALYCH.G.....	KHVIH.....RDIKPEN					
O43930	145	SSTTGLFYSAEIIICAIEYLH.S.....	KEIVY.....RDLKLEN					
Q9VKN7	151	DEPRSAKYTVANALNYCH.L.....	NNVIH.....RDLKLEN					
O01427	126	SEPTAAKMYEIDADALSYCH.R.....	KNVIH.....RDIKPEN					
Q93VK0	110	SESESASYAKQILSALAHCH.R.....	CDVVH.....RDVKPDN					
O64629	118	TEQQAATYIASLSQALAYCH.G.....	KCVIH.....RDIKPEN					
Q9M077	127	SERRAATYVASLARALYCH.G.....	KHVIH.....RDIKPEN					
Q6NW76	149	DDQRTATYMEEVSDALQYCH.E.....	KKVIH.....RDIKPEN					
Q9M9E9	96	SEDEARFFFQLISGVNYCH.S.....	LQICH.....RDLKLEN					
Q8SRL5	107	GEKETSLYIRVMLLALTYMK.E.....	CNVIH.....RDIKPENLL					
Q61XD3	123	TEAMAGKYMYEIDADALSYCH.R.....	KNVIH.....RDIKPENLL					
Q02066	86	SEDEGRFFFQLISGVSYCH.S.....	MQVCH.....RDLKLENTL					
O966D4	173	DEQRTATIMEELADALMYCH.G.....	KKVIH.....RDTKPEMLI					

☰ Explain user interface

Fig. S5. Example of the on-line sequence analysis of the final alignment at the Mustguseal Analysis Page. The final alignment can be viewed and studied on-line using the integrated Strap utility (Gille et al., 2014).

Gille,C., Föhling,M., Weyand,B., Wieland,T., and Gille,A. (2014). Alignment-Annotator web server: rendering and annotating sequence alignments. *Nucleic Acids Res.*, **42**: W3-6

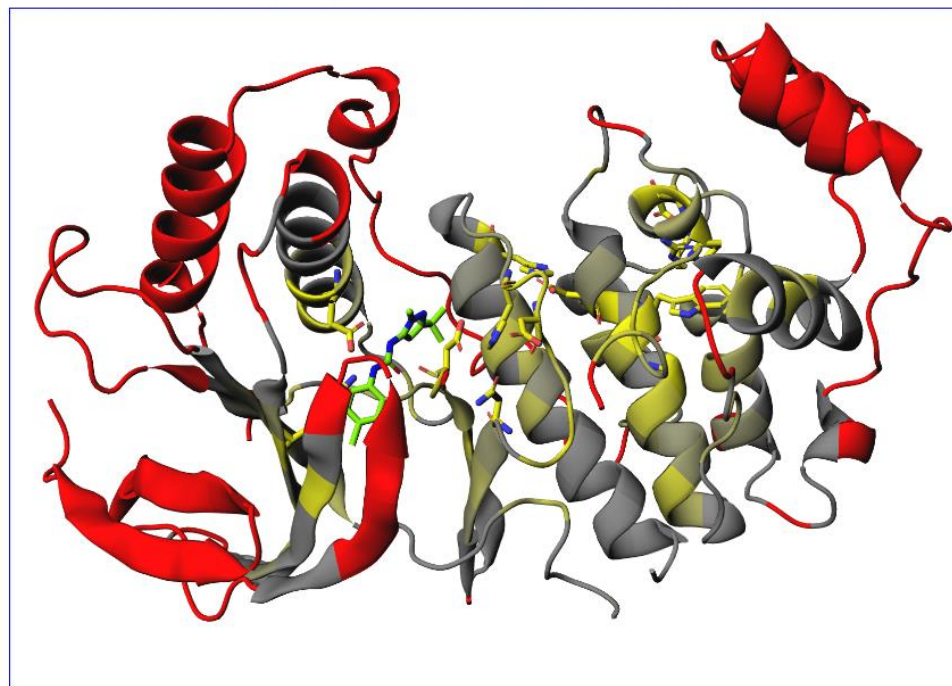
Structure-based annotation of the Final Alignment

This sub-section implements the JSMol application to provide you with a tool for the structure-based analysis of your alignment. Each representative protein structure, which was used to build the core structural alignment, has been annotated according to the final alignment. See the legend below the 3D-viewer. Choose a protein from the dropdown menu and allow some time for loading of the content. The first protein in the list is shown by default. Left-click-and-hold and then move your mouse to rotate the structure, scroll mouse wheel (or Shift + Leftclick + Mouse Up/Down) to zoom in and out, right-click for more options. Interactivity is implemented in HTML5, a language native to web browsers, therefore no plugins nor Java are required. For additional information and troubleshooting please see the *JSMol homepage*.

Select a protein from the dropdown menu:

Choose a protein from the dropdown menu and allow some time for loading of the content

Showing annotation based on protein 0_1kv1_A:



Set the PDB viewer window size: 420x300, 840x600, 1260x900, 1680x1200.

Fig. S6. Example of the on-line structure-based analysis of the final alignment at the Mustguseal Analysis Page. The selected protein structures are annotated according to basic statistics of the final alignment. Amino acid residues that are at least 95% conserved in the final alignment are colored in yellow and shown as sticks. The gradient paint of the protein backbone corresponds to sequence conservation in a corresponding position of the multiple alignment quantified by Shannon entropy (yellow – highly conserved, grey – variable). Red paint highlights positions in protein structures which are aligned to columns with more than 5% of gaps. The interactivity of the structure-based annotation is provided by the integrated JSMol utility (Hanson et al., 2013).

Hanson,R., Prilusky,J., Renjian,Z., Nakane,T., and Sussman,J. (2013). JSMol and the Next-Generation Web-Based Representation of 3D Molecular Structure as Applied to Proteopedia. *Isr. J. Chem.*, **53(3-4)**, 207-216

Examples

The Mustguseal web-server provides an opportunity to construct alignments of the selected protein families, and is capable of automatically collecting and superimposing a large set of related proteins within a superfamily. The running time of a Mustguseal task and the size of a final alignment will depend on the particular input, parameter setup, and availability of data in the PDB, Swiss-Prot, and TrEMBL databases. The Mode 1 is the default, fully automated, and the easiest way to obtain the alignment by submitting PDB and chain IDs of a query protein. It also takes more time to complete because all steps of the Mustguseal protocol are executed to collect and align the related sequences and structures from the selected databases. The structure similarity search in Mode 1 works more efficiently, if the requested structural similarity thresholds are higher, i.e., within a range 70-100%, and the Step 1 will take longer when the percentage of secondary structure equivalences is set to 30-40%. All pairwise comparisons that were once created during the structure similarity search in Mode 1 are hashed into a PostgreSQL-controlled database to be re-used in the consequent searches. When a new task is submitted in Mode 1 with the same query structure, the Step 1 (i.e., the structure similarity search) takes only a few seconds to complete because the results are not re-computed but restored from the database, which is hosted on a very fast solid-state drive. This provides an opportunity for the user to refine the alignment by submitting a new task with the same query but different parameters and getting the results significantly faster. The user can choose to perform sequence similarity searches either in the Swiss-Prot database or deal with the much larger dataset basing on Swiss-Prot+TrEMBL databases. The GPU-compatible version of BLAST is used to accelerate sequence similarity searches (Vouzis and Sahinidis, 2011). The redundancy filter threshold has a direct impact on the speed of a sequence similarity search – a value below 80% is the fastest option, and 100% is the slowest option – because a pre-calculated non-redundant sets of Swiss-Prot and TrEMBL databases are actually being used by the server, and the nr80 database set is smaller in size compared to the nr100 database set (see The Parameters above). Runtime of a task submitted in Mode 2 is on average at least two times faster than in Mode 1 because the time consuming Steps 1 and 2 (i.e., the structure similarity search and construction of the structural alignment) are skipped. A task submitted in Mode 3 takes between several seconds to several minutes. Examples of the running time with details on how to reproduce them are provided below. The exact number of collected sequences and structures in the discussed examples may vary due to continuous database updates.

It takes **25 minutes** to automatically collect and align a non-redundant set of **225 proteins** from PDB and Swiss-Prot databases corresponding to different functional families within the Glutathione S-Transferases (GST) superfamily. To reproduce this example, submit PDB 2GST (chain A) of the class Mu glutathione S-transferase in Mode 1 and leave other parameters to the default values. Alternatively, the user can submit a custom-built core structural alignment in Mode 2 to construct a structure-guided sequence alignment of only the selected protein families and their closest homologs. Submit PDB codes corresponding to proteins of the nine classes of the GST superfamily 1ev4 (class Alpha), 1f2e (class Beta), 1jlv (class Delta), 1fw1 (class Zeta), 1ljr (class Theta), 1lbk (class Pi), 1gwc (class Tau), 1axd (class Phi), and 1eem (class Omega) to the MATT structural alignment web-server (<http://matt.cs.tufts.edu>, “1ev4:A 1f2e:A 1jlv:A 1fw1:A 1ljr:A 1lbk:A 1gwc:A 1axd:A 1eem:A”) and then submit this superimposition (the `alignment.fasta_aln` file) to the Mustguseal server in Mode 2 with the default parameters. It takes **15 seconds** in Mode 2 to automatically collect and

align a non-redundant set of **173 proteins** from the PDB and Swiss-Prot databases corresponding to the selected families of the GST superfamily. A similar alignment was used to discuss functional promiscuity of the GST enzymes (Suplatov et al., 2014). The server is capable of constructing multiple alignments of functionally diverse families that include thousands of proteins basing on all available information about their structures and sequences in public databases. Submit PDB 3K3I (chain A) of the human p38 α MAP Kinase as a query in Mode 1. Select the UniProtKB/Swiss-Prot+TrEMBL option for the sequence similarity search. Leave other parameters to the default values. It takes **60 minutes** to automatically collect and align a non-redundant set of **2134 sequences and structures** of MAP Kinases including families P38, JNK, ERK, CDK, and others (i.e., PDB structures of 16 proteins with at least 70% pairwise secondary structure equivalences with the query and at most 40% pairwise sequence identity with each other, and 2118 protein sequences from Swiss-Prot and TrEMBL with at most 95% pairwise sequence identity).

We further discuss in details a construction of multiple structure-guided sequence alignment of **proteins from the Fold Type IV PLP-dependent enzymes superfamily**. To reproduce this example submit PDB 5CE8 (chain A) of the Branched-chain Aminotransferase (BCAT) from thermophilic archaea *Thermoproteus uzoniensis* (TUZN) as a query in Mode 1. TUZN was selected as a query because this protein presented the primary interest and was available for experimental studies (Bezsudnova et al., 2016). Select the UniProtKB/Swiss-Prot+TrEMBL option for the sequence similarity search and set the “Maximum number of sequences to collect” to 1000 to collect more proteins from the sequence database. Leave other parameters to default values. On the first step of the protocol the structure similarity search was used to automatically select 15 representative proteins which shared at least 70% secondary structure equivalences with the submitted query protein and at most 40% pairwise sequence identify with each other. The selected proteins were superimposed to create the core structural alignment. Each representative protein was independently used as a query to run sequence similarity search in the selected databases and create 15 sequence alignments which were finally merged according to the core structural alignment. The final superimposition contained a non-redundant set of **5211 proteins** (i.e., PDB structures of 15 proteins with at least 70% pairwise secondary structure equivalences with the query and at most 40% pairwise sequence identity with each other, and 5196 protein sequences from Swiss-Prot and TrEMBL with at most 95% pairwise sequence identity). It took **60 minutes** to construct this alignment **in the fully automatic Mode 1**. The basic alignment statistics, which was automatically calculated and presented at the Analysis page, indicated that 230 columns of the alignment (74% of the average length of proteins in the alignment) contained at most 5% of gaps, suggesting good alignment coverage and availability of information for further analysis. The structure-based annotation of the final alignment, which was also available at the Analysis page, clearly indicated that the PLP Fold type IV is highly conserved in the aligned proteins and the active site area is the most conserved by sequence among homologs (Fig. S7). In particular, three amino acid residues – the Schiff base-forming Lys150, as well as Arg54 and Glu184 residues (numbering as in PDB 5CE8) located in the cofactor-binding subsite are conserved in at least 95% of aligned sequences, what is in agreement with the published mechanism of the PLP-dependent catalysis (Boyko et al., 2016; Phillips, 2015; Toney, 2011). At the same time it was noted that the loop 98-112 (numbering as in PDB 5CE8) which is a part of the catalytic site of the enzyme is poorly aligned among homologs (Fig. S8, A).

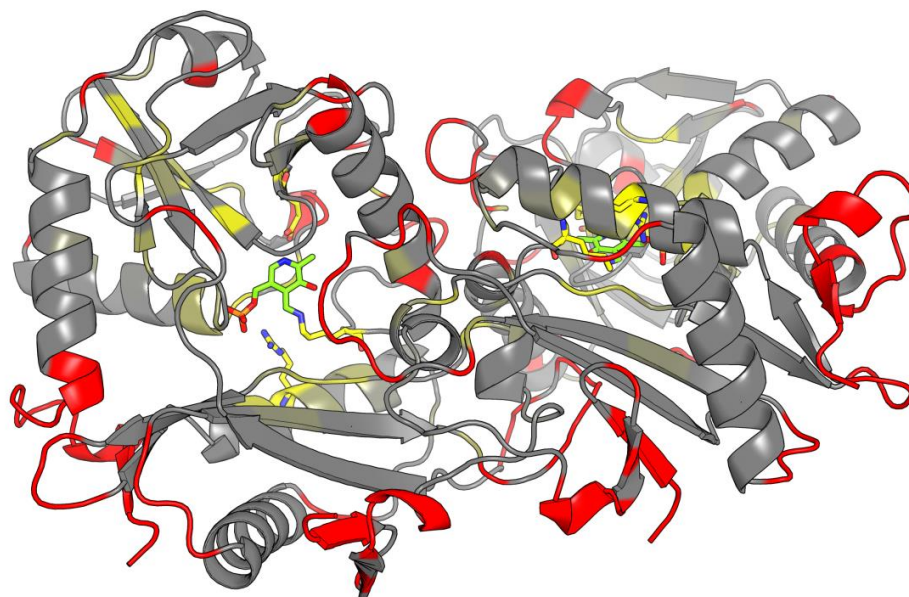


Fig. S7. Structure-based annotation of the final alignment of 5211 Fold Type IV PLP-dependent enzymes mapped on the PDB 5CE8 (generated by the Mustguseal web-server). Amino acid residues that are at least 95% conserved in the final alignment are colored in yellow and shown as sticks. The gradient paint of the protein backbone corresponds to sequence conservation in a corresponding position of the multiple alignment quantified by Shannon entropy (yellow – highly conserved, grey – variable). Red paint highlights positions in the protein structure which are aligned to columns with more than 5% of gaps. The PLP cofactor is shown as sticks and colored in chartreuse.

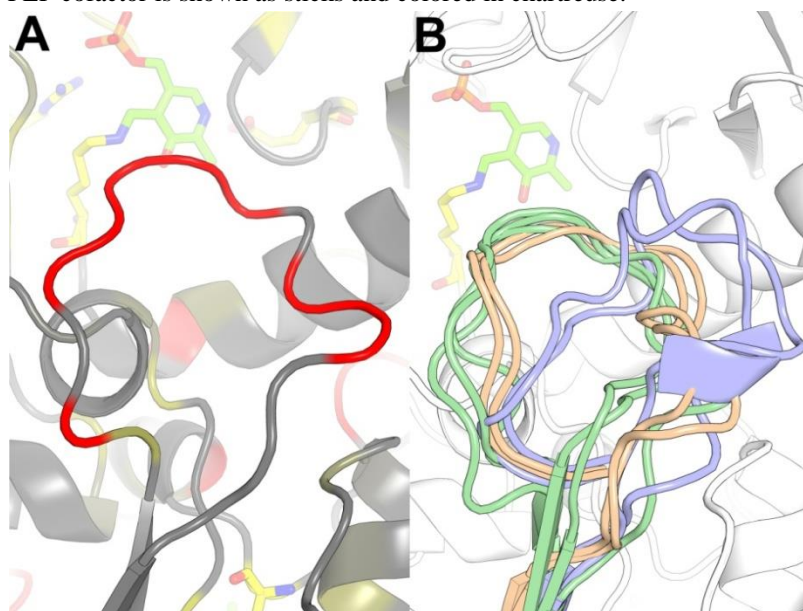


Fig. S8. Structural analysis of the substrate-binding loop 98-112 (numbering as in PDB 5CE8). (A) The structure-based annotation of the final alignment of 5211 Fold Type IV PLP-dependent enzymes mapped on the PDB 5CE8 (generated by the Mustguseal web-server). See the legend in Fig. S7. (B) Structural alignment of selected representative enzymes with different functional annotations. The orientation of the loop 98-112 is similar in closely related BCATs (green) and DAAs (wheat), but has a significantly different structural organization in more distant relatives, e.g., R-stereoselective transaminases (blue). The PLP cofactor is shown as sticks and colored in chartreuse and the conserved Schiff base-forming Lys150 is shown as sticks and colored in yellow.

The core structural alignment pack and the structure similarity search results pack were downloaded from the Mustguseal Results page and studied – i.e., the file `strcore/strcore.fasta` contains the sequence representation of the structural alignment, the folder `strcore/aligned_pdb/` contains superimposed PDB files of all representative proteins, and the file `strsearch/superimpose.list` contains the results of pairwise structural comparisons of each representative protein with the query protein and their functional annotations from PDB. It was shown that the loop 98-112 has a similar structural organization in other BCATs and closely related D-amino acid aminotransferases (DAAs), but a significantly different structural organization in more distant relatives, e.g., R-stereoselective transaminases (Fig. S8, B), and this local diversity has a direct implication to the catalytic mechanism (e.g., Skalden et al., 2015). The purpose of our study (Bezsudnova et al., 2016) was to understand the unique substrate specificity of the BCAT from *Thermoproteus uzoniensis* (i.e., TUZN). Consequently, six proteins (BCATs and DAAs) with a similar structural organization of the loop 98-112 were selected, aligned using a locally installed MATT software (alternatively, the MATT structural alignment web-server at <http://matt.cs.tufts.edu> or the PROMALS3D structural alignment web-server at <http://prodata.swmed.edu/promals3d> can be used to align selected protein structures on-line), and submitted to the Mustguseal web-server in Mode 2 with the same parameters for the sequence similarity search which were used for the initial submission in Mode 1. It took **13 minutes** to process this task in Mode 2 and produce a multiple structure-guided sequence alignment of a non-redundant set of **3120 proteins**. The bioinformatic analysis of a qualitatively similar alignment helped to identify subfamily-specific positions in the loop 98-112 and in strands $\beta 7$ and $\beta 8$ that determine the unique substrate specificity profile of TUZN (see Bezsudnova et al., 2016).

The Mustguseal web-server can automatically construct large structure-guided sequence alignments of functionally diverse protein families that include thousands of proteins basing on all available information about their structures and sequences in public databases. A systematic analysis of protein families and superfamilies can help at studying the molecular mechanisms of protein action and regulation, designing improved variants of enzymes/proteins, and selective ligands to modulate their functional properties (Suplatov et al., 2016; Suplatov et al., 2015; Pleiss, 2014; De Juan et al., 2013; Kourist et al., 2010; Rausell et al., 2010). We hope the Mustguseal web-server will facilitate further studies of the structure-function relationship in protein families and superfamilies by providing an easy-to-use interface for the construction of multiple alignments of structurally and functionally diverse proteins for a common use in a laboratory practice to assist experimental research.

Bezsudnova,E.Y., Stekhanova,T.N., Suplatov,D.A., Mardanov,A.V., Ravin,N.V., and Popov,V.O. (2016). Experimental and computational studies on the unusual substrate specificity of branched-chain amino acid aminotransferase from *Thermoproteus uzoniensis*. *Arch. Biochem. Biophys.*, **607**, 27-36.

Boyko,K.M., Stekhanova,T.N., Nikolaeva,A.Y., Mardanov,A.V., Rakitin,A.L., Ravin,N.V., Bezsudnova,E.Y. and Popov,V.O. (2016). First structure of archaeal branched-chain amino acid aminotransferase from *Thermoproteus uzoniensis* specific for l-amino acids and R-amines. *Extremophiles*, **20(2)**, 215-225.

De Juan,D., Pazos,F., and Valencia,A. (2013). Emerging methods in protein co-evolution. *Nat. Rev. Genet.* Genetics, **14(4)**, 249.

Phillips,R.S. (2015). Chemistry and diversity of pyridoxal-5'-phosphate dependent enzymes. *BBA - Proteins and Proteomics*, **1854(9)**, 1167-1174.

The Mustguseal web-server

- Pleiss, J. (2014). Systematic Analysis of Large Enzyme Families: Identification of Specificity- and Selectivity-Determining Hotspots. *ChemCatChem*, **6**, 944-950.
- Rausell, A., Juan, D., Pazos, F., and Valencia, A. (2010). Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc. Natl. Acad. Sci. U.S.A.*, **107**(5), 1995-2000.
- Skalden, L., Thomsen, M., Höhne, M., Bornscheuer, U.T. and Hinrichs, W. (2015). Structural and biochemical characterization of the dual substrate recognition of the (R)-selective amine transaminase from *Aspergillus fumigatus*. *FEBS J*, **282**(2), 407-415.
- Suplatov, D., Kirilin, E., Takhaviev, V., and Švedas, V. (2014). Zebra: a web server for bioinformatic analysis of diverse protein families. *J. Biomol. Struct. Dyn.*, **32**(11), 1752-1758
- Suplatov, D., Kirilin, E., and Švedas, V. (2016). Bioinformatic Analysis of Protein Families to Select Function-Related Variable Positions. In Svendsen, A. (ed.) *Understanding Enzymes*, Pan Stanford, pp. 351-385.
- Suplatov, D., Voevodin, V., and Švedas, V. (2015). Robust enzyme design: Bioinformatic tools for improved protein stability. *Biotechnology J.*, **10**(3), 344-355.
- Toney, M.D. (2011). Controlling reaction specificity in pyridoxal phosphate enzymes. *BBA - Proteins and Proteomics*, **1814**(11), 1407-1418.
- Vouzis, P.D., and Sahinidis, N.V. (2011). GPU-BLAST: using graphics processors to accelerate protein sequence alignment. *Bioinformatics*, **27**(2), 182-188.

Advanced tools to study the Mustguseal alignment

The final alignment can be further submitted to sister services of Mustguseal – *Zebra*, *pocketZebra* and *visualCMAT* web-servers for analysis of conserved, subfamily-specific and co-evolving residues to help at studying mechanisms of protein action and regulation, designing enzymes with improved properties for practical application and selective modulators of activity of the wild-type proteins. You can make a submission to *Zebra*, *pocketZebra*, and *visualCMAT* directly from this page. Choose the server and the data to be submitted and press the "Submit" button. Your data will be automatically uploaded from the Mustguseal platform to the selected server and you will be redirected to a corresponding submission page. Press the "Submit" button on that page to start the analysis with the default settings. Would you wish to customize your submission to *Zebra*, *pocketZebra*, and *visualCMAT* (e.g., use a different PDB file or change the default settings) you can make a submission to the selected server manually. Learn more about these advanced tools to study the Mustguseal alignment: [\[link\]](#).

Your final Mustguseal alignment contains 5211 proteins.



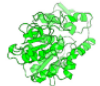
Submit the Final Mustguseal alignment	FINAL_A-97559z7bebdwtz.fasta	20.0 MB		Submit to Zebra
Submit the Final Mustguseal alignment and a PDB structure of the representative protein	FINAL_A-97559z7bebdwtz.fasta	20.0 MB		Submit to Zebra
	0_5ce8_A.pdb	146 KB		Submit to pocketZebra
				Submit to visualCMAT

Fig. S9. A new submission to *Zebra*, *pocketZebra*, and *visualCMAT* web-services can be made directly from the Mustguseal Results Page.