Protein Sequence Analysis Using the MPI Bioinformatics Toolkit

Felix Gabler,[1] Seung-Zin Nam,[1] Sebastian Till,[1] Milot Mirdita,[2]
Martin Steinegger,[2,3] Johannes Söding,[2] Andrei N. Lupas,[1]
and Vikram Alva[1,4]

[1] Department of Protein Evolution, Max Planck Institute for Developmental Biology, Tübingen, Germany
[2] Quantitative Biology and Bioinformatics, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany
[3] Present address: Department of Biology, Seoul National University, Seoul, South Korea
[4] Corresponding author: *vikram.alva@tuebingen.mpg.de*

The MPI Bioinformatics Toolkit (*https://toolkit.tuebingen.mpg.de*) provides interactive access to a wide range of the best-performing bioinformatics tools and databases, including the state-of-the-art protein sequence comparison methods HHblits and HHpred. The Toolkit currently includes 35 external and in-house tools, covering functionalities such as sequence similarity searching, prediction of sequence features, and sequence classification. Due to this breadth of functionality, the tight interconnection of its constituent tools, and its ease of use, the Toolkit has become an important resource for biomedical research and for teaching protein sequence analysis to students in the life sciences. In this article, we provide detailed information on utilizing the three most widely accessed tools within the Toolkit: HHpred for the detection of homologs, HHpred in conjunction with MODELLER for structure prediction and homology modeling, and CLANS for the visualization of relationships in large sequence datasets. © 2020 The Authors.

Basic Protocol 1: Sequence similarity searching using HHpred
Alternate Protocol: Pairwise sequence comparison using HHpred
Support Protocol: Building a custom multiple sequence alignment using PSI-BLAST and forwarding it as input to HHpred
Basic Protocol 2: Calculation of homology models using HHpred and MODELLER
Basic Protocol 3: Cluster analysis using CLANS

Keywords: CLANS • cluster analysis • HHpred • HMM • homology • profile hidden Markov models • sequence comparison • sequence similarity searches • structure prediction

How to cite this article:
Gabler, F., Nam, S.-Z., Till, S., Mirdita, M., Steinegger, M., Söding, J., Lupas, A. N., & Alva, V. (2020). Protein sequence analysis using the MPI bioinformatics toolkit. *Current Protocols in Immunology*, *72,* e108. doi: 10.1002/cpbi.108

CURRENT PROTOCOLS
A Wiley Brand

Current Protocols in Bioinformatics e108, Volume 72
Published in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpbi.108

INTRODUCTION

The structure, function, and evolution of new or uncharacterized proteins are routinely inferred based on their homology to proteins with experimentally characterized properties. Sequence searches are a common first step in this process, as sequence similarity is widely accepted as the best marker for substantiating homologous relationships. Over the years, many high-quality sequence search methods [e.g., BLAST (Altschul et al., 1997; Ladunga, 2017), HMMER (Potter et al., 2018; Prakash, Jeffryes, Bateman, & Finn, 2017), HHblits (Remmert, Biegert, Hauser, & Soding, 2011), HHpred (Soding, 2005; Steinegger et al., 2019)]; protein sequence and domain databases [SCOPe (Fox, Brenner, & Chandonia, 2014), ECOD (Cheng et al., 2014; Schaeffer, Liao, & Grishin, 2018), Pfam (Coggill, Finn, & Bateman, 2008; El-Gebali et al., 2019), RefSeq (O'Leary et al., 2016), UniProt (Pundir, Martin, O'Donovan, & The UniProt Consortium, 2016; The UniProt Consortium, 2019)]; and integrative Web resources [the EMBL-EBI Bioinformatics Web Services (Madeira et al., 2019; Madeira, Madhusoodanan, Lee, Tivey, & Lopez, 2019), the SIB Bioinformatics Resource Portal (Swiss Institute of Bioinformatics Members, 2016), National Center for Biotechnology Information Web Resources (NCBI Resource Coordinators, 2018; Gibney & Baxevanis, 2011; Yang, Derbyshire, Yamashita, & Marchler-Bauer, 2020)] have been developed to help researchers make meaningful inferences based on homology. Driven by our work at the interface of computational and experimental biology, we launched the MPI Bioinformatics Toolkit in 2005 to provide researchers in the life sciences with easy, web-based access to the best-performing bioinformatics tools and databases (Biegert, Mayer, Remmert, Soding, & Lupas, 2006). The Toolkit has been in continuous operation ever since, and we replaced the first version with an entirely new one built using more scalable and robust web technologies in 2017 (Alva, Nam, Soding, & Lupas, 2016; Zimmermann et al., 2018). The Toolkit currently includes 35 in-house and external tools for sequence similarity searching [e.g., PSI-BLAST (Altschul et al., 1997), HHblits, HHpred]; calculation of multiple sequence alignments [ClustalΩ (Sievers et al., 2011), T-Coffee (Notredame, Higgins, & Heringa, 2000)]; prediction of secondary structure and sequence features [Quick2D, PCOILS (Gruber, Soding, & Lupas, 2006), TPRpred (Karpenahalli, Lupas, & Soding, 2007)]; and sequence classification [CLANS (Frickey & Lupas, 2004), MMseqs2 (Mirdita, Steinegger, & Soding, 2019)].

Over the years, the Toolkit has established itself as an important resource for molecular biology research, mainly due to the sensitive sequence-comparison tools HHblits and HHpred, which, in many instances, can detect homologous relationships that are not readily recognized by other tools. A further strength of the Toolkit lies in the tight interconnection of the tools, allowing the results of one tool to be forwarded as input to others; for instance, the output of a PSI-BLAST search could be forwarded to ClustalΩ to obtain a multiple sequence alignment (MSA) of the identified matches or to MMseqs2 to obtain a reduced set filtered by pairwise sequence identity. Finally, our implementations of some external tools offer enhanced features, such as versions of the NCBI nonredundant (nr) database for PSI-BLAST that are clustered down to 30% (nr30), 50% (nr30), 70% (nr30), or 90% (nr90) sequence identity.

In this article, we describe detailed protocols for the application of the three most frequently used tools. Basic Protocol 1 describes how to use HHpred to search for remote homologs of a protein and make inferences about its domain composition, structure, function, and evolution. The Alternate Protocol describes the pairwise comparison mode of HHpred, which allows two protein sequences or MSAs to be compared with each other. The Support Protocol describes how to build a custom, high-quality MSA starting with a protein sequence and use it as input for HHpred. Basic Protocol 2 describes how to use HHpred in conjunction with MODELLER (Sali & Blundell, 1993) to build a three-dimensional (3D) structural model for a protein sequence of interest. Basic Protocol 3 describes the use of PSI-BLAST in conjunction with CLANS to detect distant homologs of a protein of interest and then visualize the relationships between the detected homologs. To demonstrate these protocols, we use as an example the experimentally uncharacterized FtsZ protein of the Asgard group archaeon *Prometheoarchaeum syntrophicum* strain MK-D1, which currently represents the closest cultured prokaryotic relative of eukaryotes (Imachi et al., 2020). In most bacteria, many archaea, all chloroplasts, and some mitochondria, with the latter two representing endosymbiosis-derived eukaryotic organelles, FtsZ forms filaments that assemble into a ring (Z-ring) at the future site of cell division (Lowe & Amos, 1998; Margolin, 2005; Szwedziak, Wang, Bharat, Tsim, & Lowe, 2014). Notably, eukaryotic tubulins, which polymerize to form microtubules, a major component of the cytoskeleton, are remotely homologous to FtsZ (Nogales, Downing, Amos, & Lowe, 1998). FtsZ and tubulins are GTPases that comprise an N-terminal GTP-binding domain with a highly conserved GGGTG(T/S)G motif associated with GTP binding and a C-terminal regulatory domain (Erickson, 1998). Strikingly, the pairwise sequence identity between FtsZ and tubulins is lower than 15%. Therefore, most sequence search methods fail to substantiate a homologous relationship between them. We note that the structure, function, and evolution of FtsZ and tubulins have been studied extensively, and that their evolutionary relatedness is also widely accepted (Erickson, 1998; Nogales et al., 1998). However, for instructional purposes, we will assume that the homology between them is unclear. In the following, we show how the Toolkit could be used to investigate the relationship between FtsZ and tubulins.

SEQUENCE SIMILARITY SEARCHING USING HHpred

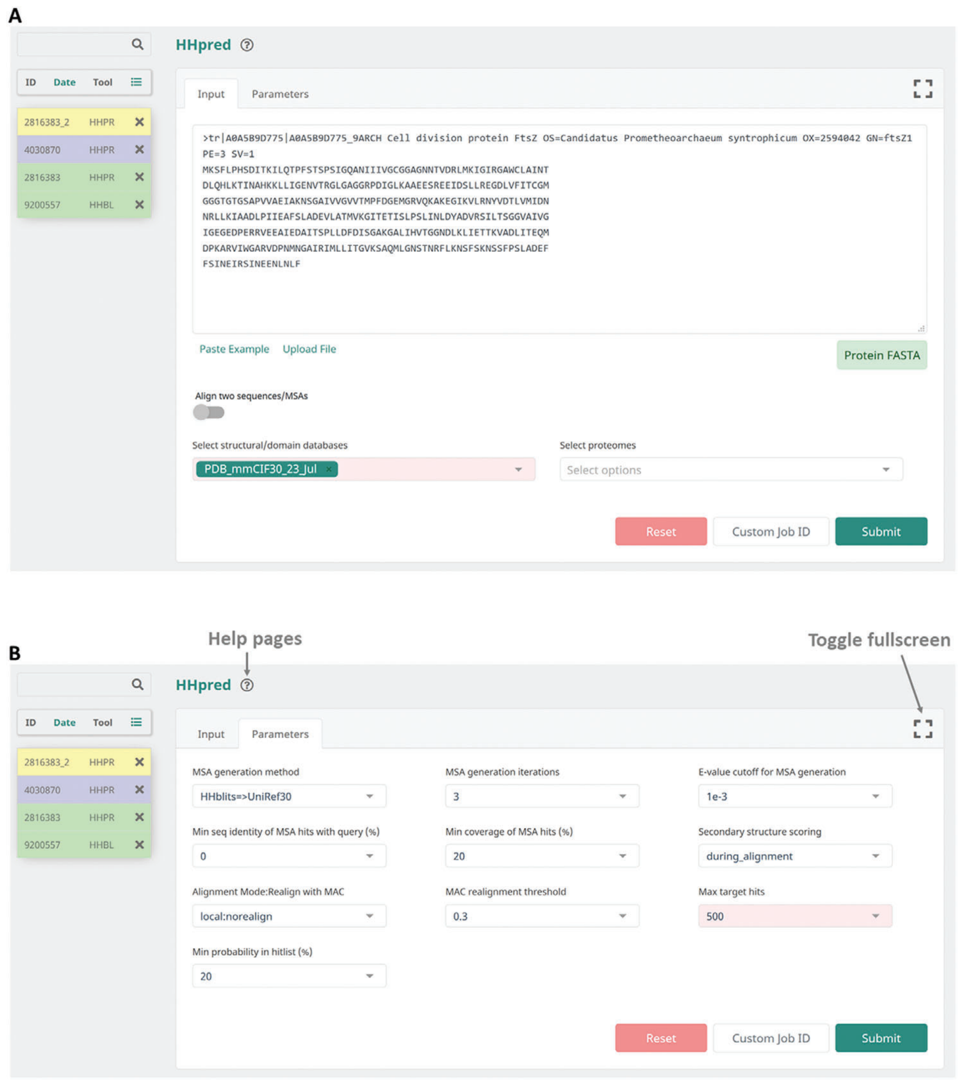
BASIC PROTOCOL 1

An almost ubiquitous first step in the characterization of a protein is the identification of functionally and structurally characterized homologs using BLAST (Altschul et al., 1997) or HMMER (Potter et al., 2018). Frequently, however, these search methods fail to detect statistically significant connections to characterized proteins. In many such cases, the more sensitive sequence search method HHpred (Steinegger et al., 2019), which is based on the comparison of profile hidden Markov models (HMMs), is able to establish connections to remotely homologous, characterized proteins. Starting from a single protein sequence, HHpred builds a multiple sequence alignment using HHblits (Steinegger et al., 2019) or PSI-BLAST (Altschul et al., 1997) and annotates the obtained alignment with the predicted secondary structure using PSIPRED (Jones, 1999). Next, this annotated alignment is converted to a profile HMM and compared to each profile HMM in user-selected target databases, which represent proteins of known structure or annotated protein families. Such databases are, for example, the Pfam (El-Gebali et al., 2019), CDD (Lu et al., 2020), and SMART (Letunic & Bork, 2018) domain databases; the SCOPe (Fox et al., 2014) and ECOD (Cheng et al., 2014) structural classification databases; the Protein Data Bank (Berman et al., 2000); and proteomes of several model organisms. Database HMMs are built using three iterations of HHblits over UniRef30 (Mirdita et al., 2017), which is a version of the UniRef sequence database (Suzek et al., 2015) clustered into groups of similar sequences at a length coverage of at least 80% and a maximum pairwise sequence identity of 30%. Like query HMMs, database HMMs include secondary structure information, either predicted by PSIPRED or assigned based on 3D structure by DSSP (Joosten et al., 2011; Kabsch & Sander, 1983). The inclusion of secondary structure information significantly increases the sensitivity of HHpred. The output of HHpred is a list of the closest homologs, with pairwise alignments.

Necessary Resources

Hardware

A desktop computer, a laptop, or a tablet with Internet access



ID	Tool
2816383_2	HHPR
4030870	HHPR
2816383	HHPR
9200557	HHBL

```
>tr|A0A5B9D775|A0A5B9D775_9ARCH Cell division protein FtsZ OS=Candidatus Prometheoarchaeum syntrophicum OX=2594042 GN=ftsZ1
PE=3 SV=1
MKSFLPHSDITKILQTPFSTSPSIGQANIIIVGCGGAGNNTVDRLMKIGIRGAWCLAINT
DLQHLKTINAHKKLLIGENVTRGLGAGGRPDIGLKAAEESREEIDSLLREGDLVFITCGM
GGGTGTGSAPVVAEIAKNSGAIVVGVVTMPFDGEMGRVQKAKEGIKVLRNYVDTLVMIDN
NRLLKIAADLPIIEAFSLADEVLATMVKGITETISLPSLINLDYADVRSILTSGGVAIVG
IGEGEDPERRVEEAIEDAITSPLLDFDISGAKGALIHVTGGNDLKLIETTKVADLITEQM
DPKARVIWGARVDPNMNGAIRIMLLITGVKSAQMLGNSTNRFLKNSFSKNSSFPSLADEF
FSINEIRSINEENLNLF
```

Align two sequences/MSAs

Select structural/domain databases: PDB_mmCIF30_23_Jul

Select proteomes: Select options

MSA generation method: HHblits=>UniRef30

MSA generation iterations: 3

E-value cutoff for MSA generation: 1e-3

Min seq identity of MSA hits with query (%): 0

Min coverage of MSA hits (%): 20

Secondary structure scoring: during_alignment

Alignment Mode:Realign with MAC: local:norealign

MAC realignment threshold: 0.3

Max target hits: 500

Min probability in hitlist (%): 20

Figure 1 Submission page of HHpred. The submission page of all tools within the Toolkit, including HHpred, contains two tabs: (**A**) ‘Input’ and (**B**) ‘Parameters.’ In the ‘Input’ tab, the amino acid sequence of FtsZ from *P. syntrophicum* in FASTA format (UniProt ID: A0A5B9D775) is shown as an example, and the target database is set to PDB_mmCIF30 (version of July 23, 2020). In the ‘Parameters’ tab, default values are set for all parameters, except ‘Max target hits’ (= 500).

Software

An up-to-date, JavaScript-enabled Web browser (preferably Google Chrome, Mozilla Firefox, or Apple Safari)

Input files

A protein sequence (in FASTA format or as plain text) or a multiple protein sequence (MSA) alignment (in FASTA, STOCKHOLM, or CLUSTAL format)

Submission page of HHpred

1. Navigate your Web browser to the submission page of HHpred at *https://toolkit.tuebingen.mpg.de/tools/hhpred*.

   The submission page of HHpred is organized into two tabs, an ‘Input’ tab (Fig. 1A) and a ‘Parameters’ tab (Fig. 1B). The ‘Input’ tab contains a large text box for pasting the query protein sequence or MSA, and drop-down lists for choosing the target profile HMM database(s). It also includes options for pasting an example protein sequence (‘Paste

Example'), uploading the input sequence as a file ('Upload File'), and activating the pairwise comparison mode ('Align two sequences/MSAs'). The 'Parameters' tab provides drop-down lists for customizing different input parameters (Fig. 1B). Options to access the help pages, toggle between windowed mode and full-screen mode, enter a custom job identifier, and submit a job are provided on both tabs.

2. Paste the amino acid sequence of your protein of interest (in FASTA format or as plain text) or an MSA (in FASTA, CLUSTAL, or STOCKHOLM format) into the large textbox (Fig. 1A). Alternatively, the input sequence or MSA can be uploaded using the 'Upload File' option. Follow the Support Protocol to build a custom MSA, starting with a protein sequence of interest.

   If you do not have the amino acid sequence of your protein of interest at hand, you can retrieve it from the protein database at NCBI (https://www.ncbi.nlm.nih.gov/protein) or the UniProt database (https://www.uniprot.org). The query sequence or MSA is validated as soon as it is pasted or uploaded, and an error message is displayed if it is not in one of the permitted formats. Upper- and lowercase letters, as well as the special characters for a gap ('.', '-') and stop codon (''), are allowed in the amino acid sequence. If your input sequence is longer than 2000 residues, we advise you to cut it into overlapping blocks of less than 2000 residues and search with these blocks separately. Generating MSAs for long sequences is computationally very expensive and might result in your jobs running for several hours.*

   In Figure 1A, we use the putative FtsZ protein of the archaeon P. syntrophicum as an example (UniProt ID: A0A5B9D775; NCBI ID: WP_147661771).

3. Select target profile HMM database(s) against which you wish to compare the query protein (Fig. 1A).

   The target profile HMM databases are organized into two different drop-down lists, one for structural and annotated sequence family databases and the other for proteomes of several archaeal, bacterial, and eukaryotic model organisms. Presently, up to four databases can be selected at a time from one or both drop-down lists. Detailed information on the databases is available in the help pages. The choice of target database primarily depends on the research question one is trying to address. To identify a homolog of known structure and function, an obvious first choice is the PDB_mmCIF70 or the PDB_mmCIF30 database. These are versions of the Protein Data Bank (PDB), a repository for all publicly available 3D structures of proteins, filtered for a maximum pairwise sequence identity of 70% (PDB_mmCIF70) or 30% (PDB_mmCIF30). To make inferences about function, evolutionary history, and domain architecture, searches can also be carried out against the expert-curated structural classification databases ECOD and SCOPe, both of which organize proteins of known structure into hierarchies of families, superfamilies, and folds based on their evolutionary history. For these databases, we offer versions that are filtered for a maximum pairwise sequence identity of 70% (ECOD_F70 and SCOPe70). Since annotated sequence family databases such as PfamA, SMART, and CDD include conserved domains, both with or without characterized function or structure, they can be beneficial for the inference of function. Finally, proteomes of model organisms can be searched to identify extremely divergent homologs. We update these target databases regularly and also include new ones. For instance, we recently included a profile HMM database comprising all manually curated viral proteins in the UniProt database (UniProt-SwissProt-viral70).

   For our example sequence, we will run four separate searches against four different target databases: PDB_mmCIF30 (Fig. 1A) to identify homologs of known structure and function, ECOD_F70 and Pfam-A to identify domains, and the proteome of Saccharomyces cerevisiae to identify its homologs in eukaryotes.

4. Customize input parameters in the 'Parameters' tab (Fig. 1B). The default values for the various parameters are set to yield the best results for most standard cases, and we recommend using them, at least in the initial steps of the analysis.

Detailed information on each parameter is available in the help pages. Upon selection of custom values for parameters, the corresponding drop-down lists are highlighted in light red, and a 'Reset' button for reloading the default values is introduced (Fig. 1B). The Toolkit caches the last used custom values and reloads them when a new submission is initiated. The sensitivity of HHpred relies heavily on the quality of MSAs. By default, three iterations of HHblits over the UniRef30 database are used to build the query MSA; if the input is an MSA, the number of iterations ('MSA generation iterations') is set to 0. In some instances, the depth of identified homologs is too low in UniRef30, which is filtered at 30% sequence identity. For instance, highly conserved proteins such as ribosomal proteins, ubiquitin, and heat shock proteins are poorly represented in UniRef30. In such cases, building the MSA with PSI-BLAST over nr70, which is a version of the nonredundant protein sequence database filtered for a maximum pairwise sequence identity of 70%, is recommended ('MSA generation method'). On the other hand, in cases where you know that your protein's orthologs are extremely divergent, you could increase the sensitivity of an HHpred search substantially by trying to include more remotely homologous sequences into the query MSA. You could, for instance, use more iterations of HHblits or PSI-BLAST, or make the MSA building criteria less stringent by increasing the 'E-value cutoff for MSA generation' (hits with an E-value better than this cutoff are used in the next search iteration or, in the last iteration, for building the query HMM). We note that a corrupted query alignment, typically resulting from the inclusion of non-homologous sequences, especially repetitive or low-complexity ones, is the main source of high-scoring false positives. If you suspect that the hits yielded by your search are false positives, make the MSA-building criteria more stringent by adjusting the values for 'E-value cutoff for MSA generation', 'Min seq identity of MSA hits with query', and 'Min coverage of MSA hits'. Finally, using an expert-built or expert-edited alignment as input may significantly increase the sensitivity and reliability of an HHpred search. Since scoring the secondary structure similarity of query and template sequences improves the sensitivity of HHpred searches significantly, 'Secondary structure scoring' is turned on by default.

For our example, we will set 'Max target hits', which controls how many matches will be displayed in the results, to 500, and use default values for all other input parameters (Fig. 1B).

5. Optionally, assign your job a custom identifier by entering one in the 'Custom Job ID' text field (Fig. 1). The identifier should contain at least two characters. If this text field is left empty, an identifier is assigned automatically.

   For our example jobs, we will let the Toolkit assign identifiers automatically.

6. Start your search by pressing the 'Submit' button.

   Upon submitting a job, a new tab that shows the current status of the job is appended. Also, an entry for the job is added to the job pane located on the left of the screen. This pane provides an overview of all jobs in the current session, allows their sorting by different criteria, gives access to individual jobs, and includes a search box to identify jobs. A job starts immediately or is queued depending on our compute cluster's actual load at the time of submission. If a previously completed job with identical input and parameters is found, an option to reload the results is offered. In the job pane, jobs are color-coded based on their current status: queued jobs are colored gray, running jobs yellow, completed jobs green, failed jobs red, and jobs with an identical copy in our database colored lavender. A new submission with modified parameters and target database(s) can be initiated from a running or completed job by simply switching to the 'Input' or 'Parameters' tabs.

HHpred search results

7. Typical HHpred searches take about 5 min to run through. However, searches involving long input sequences (>600 residues), large input MSAs, higher MSA generation iterations (4 or more), or multiple target databases could take hours to complete.

Upon successful completion, the status tab is removed from the job view, and five new tabs are appended: 'Results', 'Raw Output', 'Probability Plot', 'Query Template MSA', and 'Query MSA' (Fig. 2).

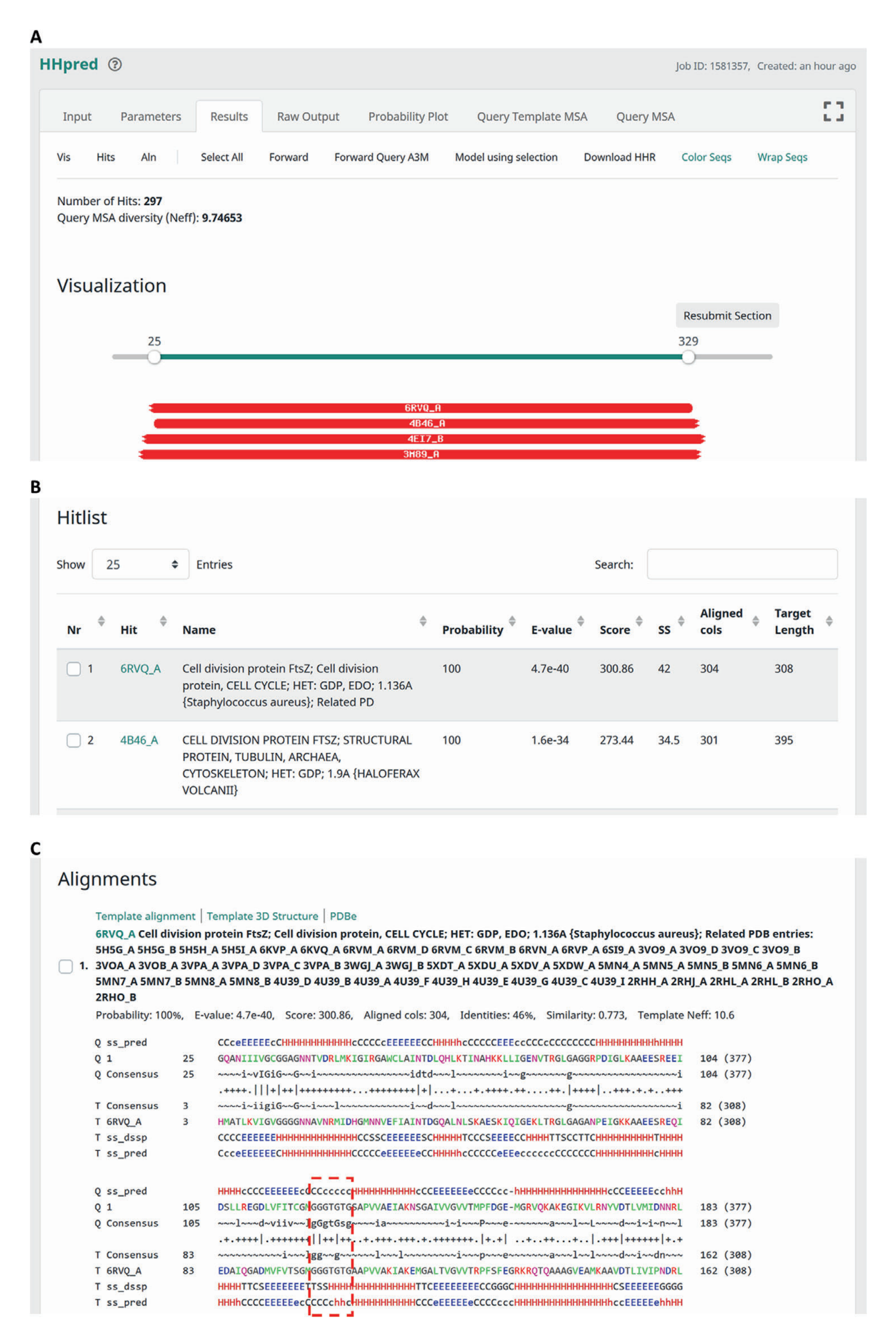
8. The 'Results' tab presents information on the detected matches in a user-friendly and interactive manner (Fig. 2).

The output is organized into three sections: 'Visualization', 'Hitlist', and 'Alignments' (Fig. 2). These sections can be accessed directly, without having to scroll to them, using the quick links ('Vis', 'Hits', and 'Aln') in the floating toolbar offered at the top of the tab. The number of detected matches and quantification for the diversity of the query MSA (Neff) is displayed directly above the 'Visualization' section. Neff ranges between 1 (for a single sequence) and 20 (for an extremely deep MSA). Additionally, the user is alerted if the query MSA contains too few sequences or if a signal peptide, coiled coils, intrinsically disordered regions, or transmembrane segments are detected.

The 'Visualization' section (Fig. 2A) shows the query sequence as a slider bar. The database matches are shown as horizontal bars underneath, indicating their coverage with respect to the query. The bars are color coded according to their significance from red (very significant) to orange, yellow, green, and cyan, to blue (less significant). In this section, only hits with a probability value of more than 40% are shown. Placing the mouse cursor over a bar shows a textual description of the match, and clicking on it takes one to the corresponding query-template alignment in the 'Alignments' section.

The 'Hitlist' section (Fig. 2B) provides a tabular listing of matches sorted by their probability of being a true positive (column 'Probability'). It includes columns with information on identifiers ('Hit'), descriptions ('Name'), E-values ('E-value'), raw scores ('Score'), secondary structure scores ('SS'), lengths of the aligned region ('Aligned Cols'), and lengths of the template ('Target Length'). The hits can be sorted by clicking on the column headers or filtered by a keyword using the search box above the table. Clicking on an index number in the leftmost column takes one to the corresponding query-template alignment in the 'Alignments' section. For the calculation of probability, the raw score ('Score') and the secondary structure score ('SS') are considered. The raw score is computed using the Viterbi HMM-HMM alignment and the secondary structure score from the alignment of secondary structure assignments between query and template, as provided by PSIPRED (3-state) or determined by DSSP (8-state). The E-value is the average number of false positives (wrong hits) with a score better than the one for the given template in the target database(s). While E-values close to 0 signify a very reliable hit, an E-value of 10 indicates that about 10 wrong hits are expected to be found in the database with a score at least this good. The P-value is the E-value divided by the number of sequences in the database. It is the probability that a wrong hit will score at least this well in a pairwise comparison. Unlike Probability, E-value and P-value are calculated without taking the secondary structure score into account.

The 'Alignments' section (Fig. 2C) provides pairwise alignments between query and template for all matches. Each entry starts with a row of hyperlinks. These always include a link to an alignment of the 100 most distinct database sequences used to generate the template HMM, and may also provide links to external resources, for example, in order to visualize the structure of the template (if the search was carried out over PDB_mmCIF30, PDB_mmCIF70, ECOD_F70, SCOPe70). The entry header then lists a description of the match and the scores for Probability, E-value, Score, Aligned cols, Identities, Similarity, and Template Neff (quantification for the diversity of the template MSA). The alignment between the query and the template itself is split into one or more blocks in which lines corresponding to the query are marked with 'Q' and the template with 'T'. The amino acid residues are colored based on their physicochemical properties. 'ss_pred' and 'ss_dssp' display secondary structure predicted by PSIPRED and assigned by DSSP, respectively. The three states predicted by PSIPRED are: H (α-helix; colored red), E (extended strand; blue), and C (residues not in H and E); upper-case and lower-case letters indicate high and low prediction confidence, respectively. The eight states assigned by DSSP are: H (α-helix; red), B (residue in isolated β-bridge), C (loop or irregular element), E (extended strand; blue),



HHpred

Job ID: 1581357, Created: an hour ago

Input | Parameters | Results | Raw Output | Probability Plot | Query Template MSA | Query MSA

Vis | Hits | Aln | Select All | Forward | Forward Query A3M | Model using selection | Download HHR | Color Seqs | Wrap Seqs

Number of Hits: **297**
Query MSA diversity (Neff): **9.74653**

Visualization

Resubmit Section

Hitlist

Show 25 Entries Search:

Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
☐ 1	6RVQ_A	Cell division protein FtsZ; Cell division protein, CELL CYCLE; HET: GDP, EDO; 1.136A {Staphylococcus aureus}; Related PD	100	4.7e-40	300.86	42	304	308
☐ 2	4B46_A	CELL DIVISION PROTEIN FTSZ; STRUCTURAL PROTEIN, TUBULIN, ARCHAEA, CYTOSKELETON; HET: GDP; 1.9A {HALOFERAX VOLCANII}	100	1.6e-34	273.44	34.5	301	395

Alignments

Template alignment | Template 3D Structure | PDBe

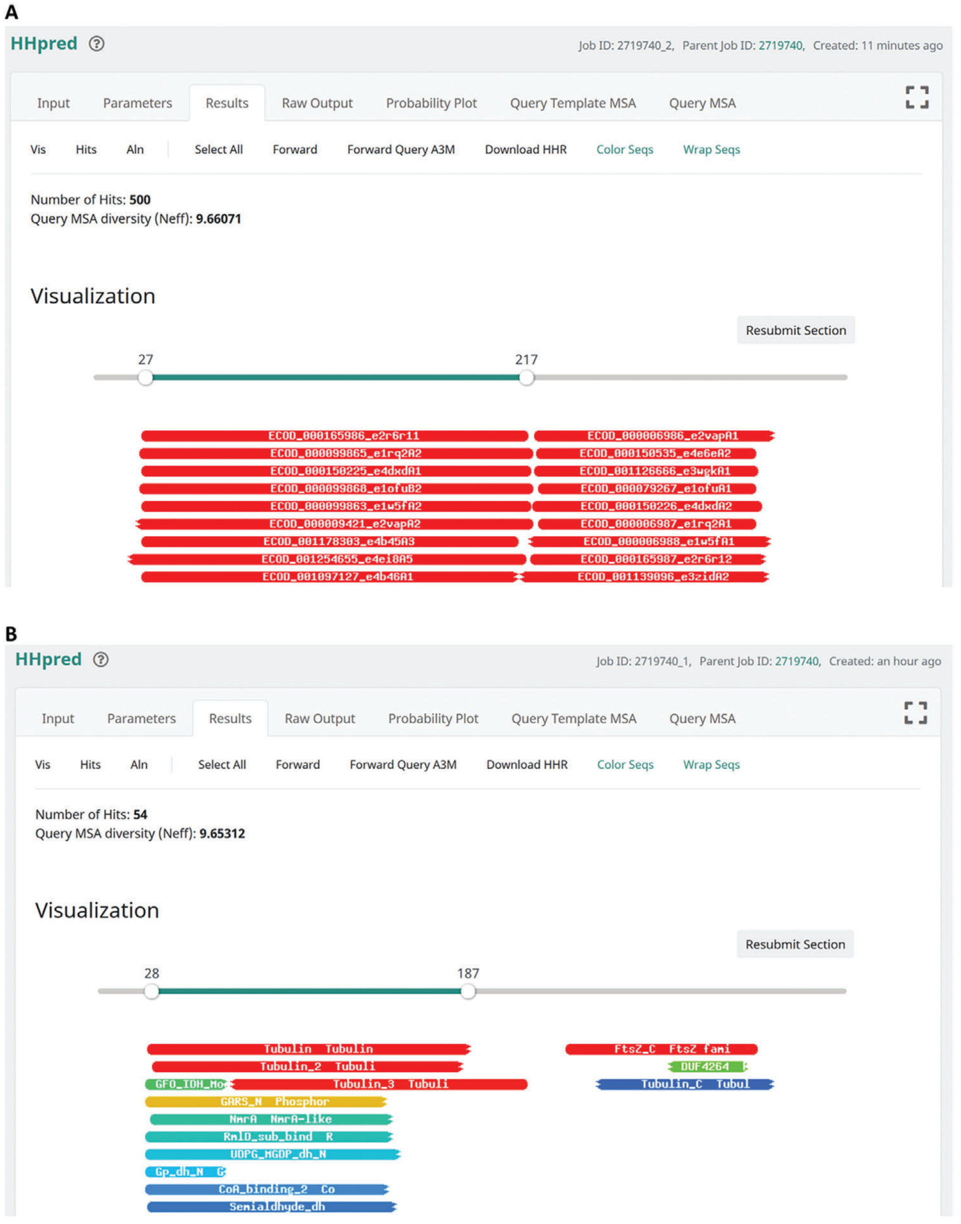
☐ 1. **6RVQ_A Cell division protein FtsZ; Cell division protein, CELL CYCLE; HET: GDP, EDO; 1.136A {Staphylococcus aureus}; Related PDB entries: 5H5G_A 5H5G_B 5H5H_A 5H5I_A 6KVP_A 6KVQ_A 6RVM_A 6RVM_D 6RVM_C 6RVM_B 6RVN_A 6RVP_A 6SI9_A 3VO9_A 3VO9_D 3VO9_C 3VO9_B 3VOA_A 3VOB_A 3VPA_A 3VPA_D 3VPA_C 3VPA_B 3WGJ_A 3WGJ_B 5XDT_A 5XDU_A 5XDV_A 5XDW_A 5MN4_A 5MN5_A 5MN5_B 5MN6_A 5MN6_B 5MN7_A 5MN7_B 5MN8_A 5MN8_B 4U39_D 4U39_B 4U39_A 4U39_F 4U39_H 4U39_E 4U39_G 4U39_C 4U39_I 2RHH_A 2RHJ_A 2RHL_A 2RHL_B 2RHO_A 2RHO_B**

Probability: 100%, E-value: 4.7e-40, Score: 300.86, Aligned cols: 304, Identities: 46%, Similarity: 0.773, Template Neff: 10.6

```
Q ss_pred           CCceEEEEEcCHHHHHHHHHHHHcCCCCcEEEEEECCHHHHhcCCCCCEEEccCCCcCCCCCCCCHHHHHHHHHHhHHHH
Q 1           25    GQANIIIVGCGGAGNNTVDRLMKIGIRGAWCLAINTDLQHLKTINAHKKLLIGENVTRGLGAGGRPDIGLKAAEESREEI   104 (377)
Q Consensus   25    ~~~~i~vIGiG~~G~~i~~~~~~~~~~~~~~~~~~~idtd~~~l~~~~~~~~i~~g~~~~~~~g~~~~~~~~~~~~~~~~~~i   104 (377)
                    .++++.|||+|++|+++++++++...++++++++|+|...+...+.++++.++....++.|++++|..+++.+.+..+++
T Consensus   3     ~~~~i~iigiG~~G~~i~~~l~~~~~~~~~~~~~~i~~d~~~l~~~~~~~~~~~~~~~~~~~~~~~g~~~~~~~~~~~~~~~~~~i   82 (308)
T 6RVQ_A      3     HMATLKVIGVGGGGNNAVNRMIDHGMNNVEFIAINTDGQALNLSKAESKIQIGEKLTRGLGAGANPEIGKKAAEESREQI   82 (308)
T ss_dssp          CCCCEEEEEEHHHHHHHHHHHHHHCCSSCEEEEEESCHHHHHTCCCSEEEECCHHHHTTSCCTTCHHHHHHHHHHTHHHH
T ss_pred           CcceEEEEEECHHHHHHHHHHHHCCCCCeEEEEEeCCHHHHhcCCCCCeEEeccccccCCCCCCCHHHHHHHHHHcHHHH

Q ss_pred           HHHHcCCCEEEEEEcCCCcccccHHHHHHHHHHHcCCEEEEEEeCCCCcc-hHHHHHHHHHHHHHHHcCCEEEEEcchhH
Q 1           105   DSLLREGDLVFITCGNGGGTGTGSAPVVAEIAKNSGAIVVGVVTMPFDGE-MGRVQKAKEGIKVLRNYVDTLVMIDNNRL   183 (377)
Q Consensus   105   ~~~l~~~d~viiv~~lgGgtGsg~~~~ia~~~~~~~~~~~~~~~i~i~~~P~~~e-~~~~~~a~~~l~~L~~~~d~~i~i~n~~l   183 (377)
                    .+.++++|.+++++++||++|++..+.+++.+++.+.+++++++.|+.+| ..+..++...+..|.+++|++++++|+.+
T Consensus   83    ~~~~~~~~~~~~~~~i~~~lgg~~g~~~~~~~l~~~l~~~~~~~~~~~~~i~~~p~~~e~~~~~~~~~a~~~l~~l~~~~d~~i~~dn~~~   162 (308)
T 6RVQ_A      83    EDAIQGADMVFVTSGMGGGTGTGAAPVVAKIAKEMGALTVGVVTRPFSFEGRKRQTQAAAGVEAMKAAVDTLIVIPNDRL   162 (308)
T ss_dssp          HHHHTTCSEEEEEEETTSSHHHHHHHHHHHHHHHTTCEEEEEEEECCGGGCHHHHHHHHHHHHHHHHHCSEEEEEEGGGG
T ss_pred           HHHhCCCCEEEEEecCCCCchhcHHHHHHHHHHHCCCeEEEEEeCCCCcccHHHHHHHHHHHHHHHHhccEEEEEehhHH
```

Figure 2 ‘Results’ tab of HHpred. The results yielded by an HHpred search are presented in an interactive format and organized into three sections: (**A**) ‘Visualization’, (**B**) ‘Hitlist’, and (**C**) ‘Alignments’. An HHpred search with FtsZ of *P. syntrophicum* over the PDB_mmCIF30 database (version of July 23, 2020, performed on September 3, 2020) yielded 297 matches, including several high-scoring matches to FtsZ proteins of archaea and bacteria as well as to eukaryotic tubulins. In the ‘Alignments’ sections, the conserved GTP-binding motif (GGGTG(T/S)G) is marked.



Job ID: 2719740_2, Parent Job ID: 2719740, Created: 11 minutes ago

Number of Hits: 500
Query MSA diversity (Neff): 9.66071

Visualization

Job ID: 2719740_1, Parent Job ID: 2719740, Created: an hour ago

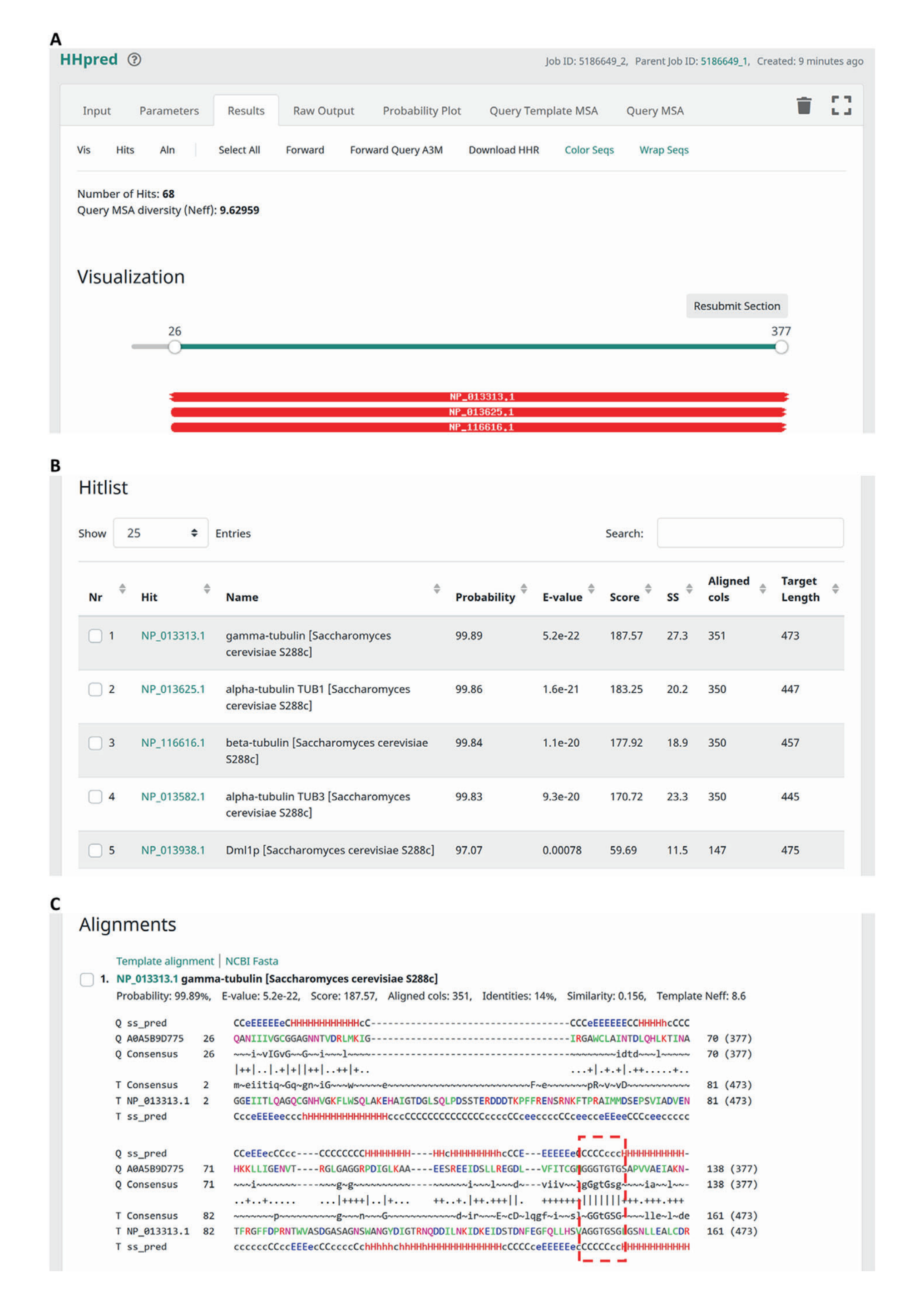
Number of Hits: 54
Query MSA diversity (Neff): 9.65312

Visualization

Figure 3 Results of HHpred searches with *P. syntrophicum* FtsZ over ECOD_F70 (**A**) and Pfam-A (**B**). The results indicated that our query protein consists of two domains, an N-terminal Tubulin/FtsZ family domain followed by an FtsZ-family C-terminal-like domain.

G (3_{10}-helix), I (π-helix), T (hydrogen-bonded turn), and S (bend). In the consensus sequences, upper- and lower-case characters indicate high (≥60%) and moderate (≥40%) conservation, respectively. The row between the query and template consensus sequences indicates the quality of the column-column match: '|' very good, '+' good, '·' neutral, '−' bad, and '=' very bad.

The search with our protein of interest, FtsZ of P. syntrophicum, over the PDB_mmCIF30 database (version of July 23, 2020, performed on September 3, 2020) yielded 297 matches (Fig. 2). The top six matches were to FtsZ proteins of archaea and bacteria and the FtsZ-like plasmid replication protein (RepX) from Bacillus cereus, all at a probability value of 100% and E-values better than 8.3e-30. Furthermore, the pairwise alignments showed the conservation of the GTP-binding motif (GGGTG(T/S)G), substantiating that our query protein is a member of the FtsZ family (Fig. 2C). The next best match was Tubulin alpha-1B of mammals at a probability value of 99.92% and E-value of 3.4e-24,



Number of Hits: **68**
Query MSA diversity (Neff): **9.62959**

Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	NP_013313.1	gamma-tubulin [Saccharomyces cerevisiae S288c]	99.89	5.2e-22	187.57	27.3	351	473
□ 2	NP_013625.1	alpha-tubulin TUB1 [Saccharomyces cerevisiae S288c]	99.86	1.6e-21	183.25	20.2	350	447
□ 3	NP_116616.1	beta-tubulin [Saccharomyces cerevisiae S288c]	99.84	1.1e-20	177.92	18.9	350	457
□ 4	NP_013582.1	alpha-tubulin TUB3 [Saccharomyces cerevisiae S288c]	99.83	9.3e-20	170.72	23.3	350	445
□ 5	NP_013938.1	Dml1p [Saccharomyces cerevisiae S288c]	97.07	0.00078	59.69	11.5	147	475

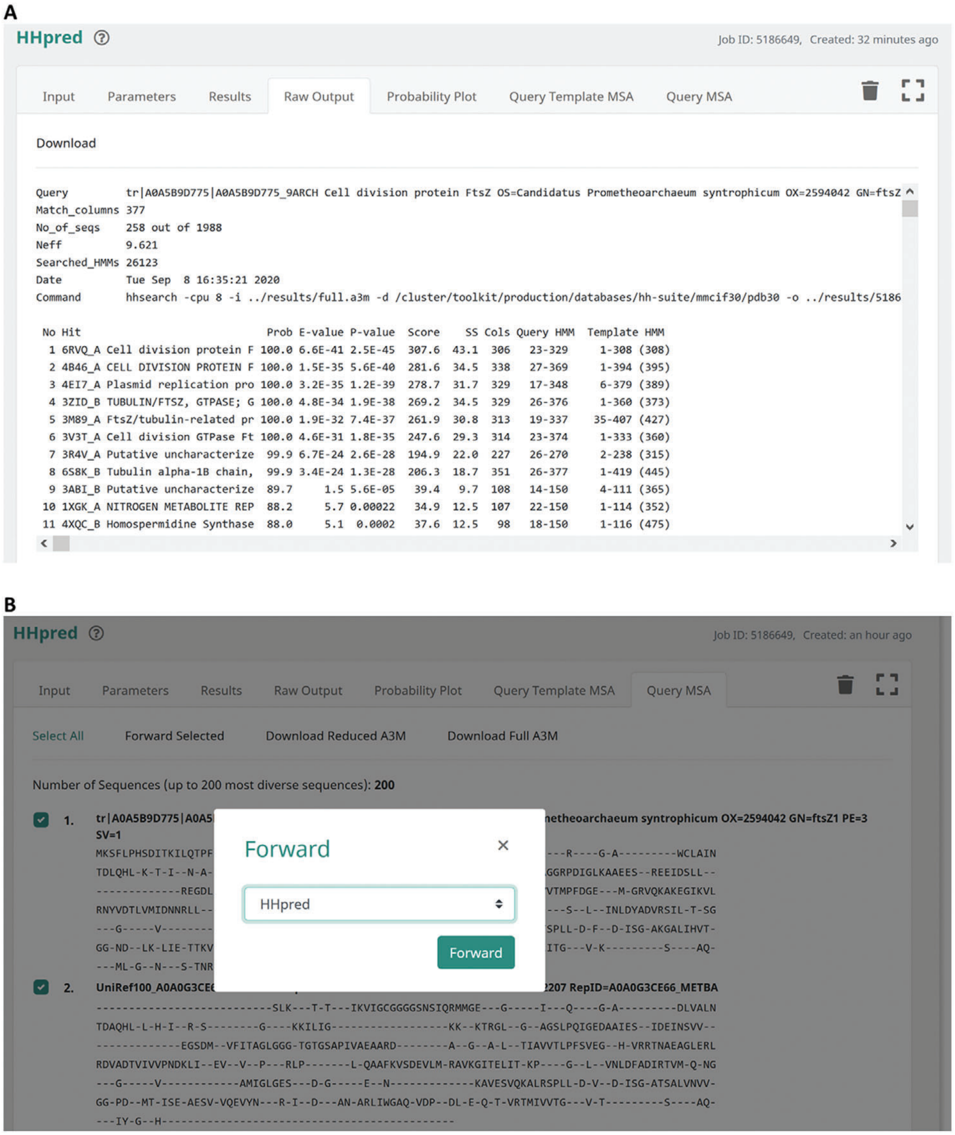
□ 1. NP_013313.1 **gamma-tubulin [Saccharomyces cerevisiae S288c]**
Probability: 99.89%, E-value: 5.2e-22, Score: 187.57, Aligned cols: 351, Identities: 14%, Similarity: 0.156, Template Neff: 8.6

```
Q ss_pred           CCeEEEEEeCHHHHHHHHHHHHcC-----------------------------------CCCeEEEEEECCHHHHhcCCC
Q A0A5B9D775    26  QANIIIVGCGGAGNNTVDRLMKIG-----------------------------------IRGAWCLAINTDLQHLKTINA   70 (377)
Q Consensus     26  ~~~i~vIGvG~~G~~i~~~l~~~~-----------------------------------~~~~~~~~idtd~~~l~~~~~   70 (377)
                    |++|..|.+|+||++|..++|+..                                   ...+|.+.+|.++.....+..
T Consensus      2  m~eiitiq~Gq~gn~iG~~~w~~~~~e~~~~~~~~~~~~~~~~~~~~~~~~~F~e~~~~~~~pR~v~vD~~~~~~~~~~~~   81 (473)
T NP_013313.1    2  GGEIITLQAGQCGNHVGKFLWSQLAKEHAIGTDGLSQLPDSSTERDDDTKPFFRENSRNKFTPRAIMMDSEPSVIADVEN   81 (473)
T ss_pred           CcceEEEeecccchHHHHHHHHHHHHHHcccCCCCCCCCCCCCCccccCCceeccccCCceecceEEeeCCCceecccc

Q ss_pred           CCeEEecCCcc----CCCCCCCCHHHHHHHH----HHcHHHHHHHHhcCCE---EEEEEecCCCCcccHHHHHHHHHHH-
Q A0A5B9D775    71  HKKLLIGENVT----RGLGAGGRPDIGLKAA----EESREEIDSLLREGDL---VFITCGMGGGTGTGSAPVVAEIAKN-  138 (377)
Q Consensus     71  ~~~i~~~~~~~~----~~~g~g~~~~~~~~~~----~~~~~~i~~~l~~~d~---viiv~~lgGgtGsg~~~~ia~~l~~-  138 (377)
                    ..+..+.....    ...|++++|..|+...    ++..+.|++.+++||.   ++++++|||||||+++.+++.+++
T Consensus     82  ~~~~~~~p~~~~~~~~~~~~g~~~~n~~~~G~~~~~~~~~d~ir~~~E~cD~lqgf~i~~sl~GGtGSG~~~~lle~l~de  161 (473)
T NP_013313.1   82  TFRGFFDPRNTWVASDGASAGNSWANGYDIGTRNQDDILNKIDKEIDSTDNFEGFQLLHSVAGGTGSGLGSNLLEALCDR  161 (473)
T ss_pred           ccccccCCccEEEecCCcccccCchHhhhchhHHhHHHHHHHHHHHHHHcCCCCceEEEEEecCCCCCccHHHHHHHHHHH
```

Figure 4 Results of an HHpred search with *P. syntrophicum* FtsZ over the proteome of *S. cerevisiae*. On September 3, 2020, this search returned a total of 68 matches, with tubulins and Dml1p being the best matches. In the 'Alignments' section, the conserved GTP-binding motif (GGGTG(T/S)G) is marked.

indicating that our query protein is homologous to eukaryotic tubulins. The proteins share an overall pairwise sequence identity of only ~14%, but both possess a highly conserved GTP-binding motif. The search against the ECOD_F70 and Pfam-A databases indicated that our query protein comprises two domains, an N-terminal Tubulin/FtsZ family GTPase domain followed by an FtsZ-family C-terminal-like domain (Fig. 3). The search against the proteome of S. cerevisiae yielded as best matches tubulins and Dml1p, a protein involved in the partitioning of mitochondria, all at probability values greater than 97% (Fig. 4).



```
Query        tr|A0A5B9D775|A0A5B9D775_9ARCH Cell division protein FtsZ OS=Candidatus Prometheoarchaeum syntrophicum OX=2594042 GN=ftsZ
Match_columns 377
No_of_seqs    258 out of 1988
Neff          9.621
Searched_HMMs 26123
Date          Tue Sep  8 16:35:21 2020
Command       hhsearch -cpu 8 -i ../results/full.a3m -d /cluster/toolkit/production/databases/hh-suite/mmcif30/pdb30 -o ../results/5186

 No Hit                             Prob E-value P-value  Score    SS Cols Query HMM  Template HMM
  1 6RVQ_A Cell division protein F 100.0 6.6E-41 2.5E-45  307.6  43.1  306   23-329     1-308 (308)
  2 4B46_A CELL DIVISION PROTEIN F 100.0 1.5E-35 5.6E-40  281.6  34.5  338   27-369     1-394 (395)
  3 4EI7_A Plasmid replication pro 100.0 3.2E-35 1.2E-39  278.7  31.7  329   17-348     6-379 (389)
  4 3ZID_B TUBULIN/FTSZ, GTPASE; G 100.0 4.8E-34 1.9E-38  269.2  34.5  329   26-376     1-360 (373)
  5 3M89_A FtsZ/tubulin-related pr 100.0 1.9E-32 7.4E-37  261.9  30.8  313   19-337    35-407 (427)
  6 3V3T_A Cell division GTPase Ft 100.0 4.6E-31 1.8E-35  247.6  29.3  314   23-374     1-333 (360)
  7 3R4V_A Putative uncharacterize  99.9 6.7E-24 2.6E-28  194.9  22.0  227   26-270     2-238 (315)
  8 6S8K_B Tubulin alpha-1B chain,  99.9 3.4E-24 1.3E-28  206.3  18.7  351   26-377     1-419 (445)
  9 3ABI_B Putative uncharacterize  89.7     1.5 5.6E-05   39.4   9.7  108   14-150     4-111 (365)
 10 1XGK_A NITROGEN METABOLITE REP  88.2     5.7 0.00022   34.9  12.5  107   22-150     1-114 (352)
 11 4XQC_B Homospermidine Synthase  88.0     5.1  0.0002   37.6  12.5   98   18-150     1-116 (475)
```

Number of Sequences (up to 200 most diverse sequences): **200**

```
☑ 1. tr|A0A5B9D775|A0A5                              etheoarchaeum syntrophicum OX=2594042 GN=ftsZ1 PE=3
     SV=1
     MKSFLPHSDITKILQTPF                              ---R----G-A---------WCLAIN
     TDLQHL-K-T-I--N-A-                              GGRPDIGLKAAEES--REEIDSLL--
     -------------REGDL                              VTMPFDGE---M-GRVQKAKEGIKVL
     RNYVDTLVMIDNNRLL--                              ---S--L--INLDYADVRSIL-T-SG
     ---G-----V--------                              SPLL-D-F--D-ISG-AKGALIHVT-
     GG-ND--LK-LIE-TTKV                              ITG---V-K---------S----AQ-
     ---ML-G--N---S-TNR
☑ 2. UniRef100_A0A0G3CE                              2207 RepID=A0A0G3CE66_METBA
     ---------------------------SLK---T-T---IKVIGCGGGGSNSIQRMMGE---G-----I---Q----G-A---------DLVALN
     TDAQHL-L-H-I--R-S--------G----KKILIG-----------------KK--KTRGL--G--AGSLPQIGEDAAIES--IDEINSVV--
     -------------EGSDM--VFITAGLGGG-TGTGSAPIVAEAARD--------A--G--A-L--TIAVVTLPFSVEG--H-VRRTNAEAGLERL
     RDVADTVIVVPNDKLI--EV--V--P---RLP-------L-QAAFKVSDEVLM-RAVKGITELIT-KP----G--L--VNLDFADIRTVM-Q-NG
     ---G-----V------------AMIGLGES---D-G-----E--N-------------KAVESVQKALRSPLL-D-V--D-ISG-ATSALVNVV-
     GG-PD--MT-ISE-AESV-VQEVYN---R-I--D---AN-ARLIWGAQ-VDP--DL-E-Q-T-VRTMIVVTG---V-T---------S----AQ-
     ---IY-G--H--------------------------------------------
```

Figure 5 ‘Raw Output’ and ‘Query MSA’ tabs of HHpred. The ‘Raw Output’ tab (**A**) provides access to the output file produced by HHpred in plain text format, whereas the ‘Query MSA’ tab (**B**) provides access to the MSA built by HHpred for the query. The latter also allows an MSA of all or just selected sequences to be forwarded as input to other tools.

9. The ‘Raw Output’ tab allows visualizing and downloading the raw output file yielded by an HHpred search (Fig. 5A). It is advisable to download and save this file for future reference.

10. The ‘Probability Plot’ tab displays a cumulative histogram of the hits and can be used to obtain a count of matches with probability values higher or lower than a given value.

11. The ‘Query Template MSA’ tab provides access to an MSA comprising the query sequence and sequences of all the obtained hits. It provides options to download the complete alignment (‘Download MSA’) or to forward the alignment to other tools (‘Forward Selected’), either completely (‘Select All’) or only for individually selected sequences.

12. The ‘Query MSA’ tab provides access to the MSA built by the HHpred server for the query (Fig. 5B). The tab displays the 200 most divergent sequences and allows an MSA of selected or all sequences to be forwarded to other tools (‘Forward Selected’). This tab also includes options for downloading this reduced query alignment or the full alignment in A3M format, a space-efficient format that we use internally to store alignments. Alignments in A3M format can be converted to FASTA using the FormatSeq tool offered within our Toolkit (*https://toolkit.tuebingen.mpg.de/tools/formatseq*).

ALTERNATE PROTOCOL

PAIRWISE SEQUENCE COMPARISON USING HHpred

The pairwise mode of HHpred allows the comparison of two sequences or MSAs. This is particularly useful when you wish to substantiate a homologous relationship between two proteins that you suspect to be homologous, compare proteins that do not exist in our profile HMM databases, or obtain an HMM-HMM based alignment of two distantly related proteins. HHpred builds MSAs for the two input sequences using HHblits or PSI-BLAST, assigns secondary structure using PSIPRED, and converts the annotated MSAs to profile HMMs. In the next step, it compares the computed HMMs and reports an alignment if a match is found that satisfies the cutoffs set in ‘Parameters’. For proteins that contain multiple homologous repeats or domains, it typically reports two or more alignments. For detailed information on using HHpred, please refer to Basic Protocol 1.

Necessary Resources

Same as for Basic Protocol 1

Submission page of HHpred

1. Navigate your Web browser to the submission page of HHpred at *https://toolkit.tuebingen.mpg.de/tools/hhpred*. If desired, click on the ‘Reset’ button to reload default values for input parameters.

2. Click on the switch labeled ‘Align two sequences/MSAs’, located below the query textbox, to activate the pairwise comparison mode of HHpred. A second sequence input textbox will be shown (Fig. 6).

3. Paste the amino acid sequences of your proteins of interest (in FASTA format or as plain text) or two MSAs (FASTA, CLUSTAL, or STOCKHOLM format) into the two textboxes. Alternatively, upload them using the ‘Upload File’ option.

   In Figure 6, we use the amino acid sequences of P. syntrophicum FtsZ (UniProt ID: A0A5B9D775; NCBI ID: WP_147661771) and human Tubulin alpha-1A (UniProt ID: Q71U36; NCBI ID: NP_001257328.1) as input.

4. Customize the input parameters in the ‘Parameters’ tab. Refer to step 4 of Basic Protocol 1 for more information on this.

   For our example, we will use default values for all input parameters.

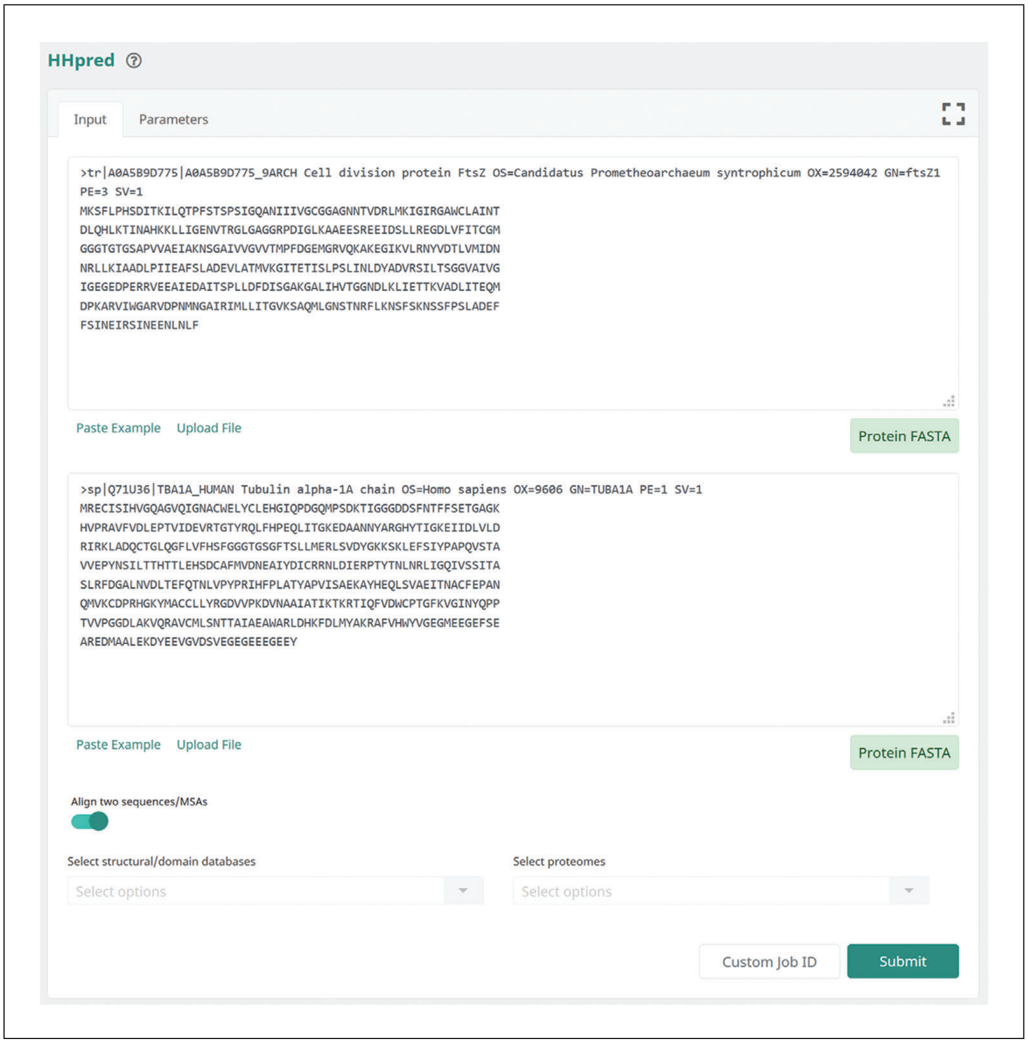
5. Optionally, assign your job a custom identifier by entering one in the ‘Custom Job ID’ text field.

6. Start your search by pressing the ‘Submit’ button.

HHpred search results

7. Typical pairwise comparisons take about 5 to 10 min. However, searches involving long input sequences (>600 residues) could take several hours to complete.

   Upon successful completion, the status tab is removed from the job view, and five new tabs are appended: ‘Results’, ‘Raw Output’, ‘Probability Plot’, ‘Query Template MSA’, and



```
>tr|A0A5B9D775|A0A5B9D775_9ARCH Cell division protein FtsZ OS=Candidatus Prometheoarchaeum syntrophicum OX=2594042 GN=ftsZ1
PE=3 SV=1
MKSFLPHSDITKILQTPFSTSPSIGQANIIIVGCGGAGNNTVDRLMKIGIRGAWCLAINT
DLQHLKTINAHKKLLIGENVTRGLGAGGRPDIGLKAAEESREEIDSLLREGDLVFITCGM
GGGTGTGSAPVVAEIAKNSGAIVVGVVTMPFDGEMGRVQKAKEGIKVLRNYVDTLVMIDN
NRLLKIAADLPIIEAFSLADEVLATMVKGITETISLPSLINLDYADVRSILTSGGVAIVG
IGEGEDPERRVEEAIEDAITSPLLDFDISGAKGALIHVTGGNDLKLIETTKVADLITEQM
DPKARVIWGARVDPNMNGAIRIMLLITGVKSAQMLGNSTNRFLKNSFSKNSSFPSLADEF
FSINEIRSINEENLNLF
```

```
>sp|Q71U36|TBA1A_HUMAN Tubulin alpha-1A chain OS=Homo sapiens OX=9606 GN=TUBA1A PE=1 SV=1
MRECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMPSDKTIGGGDDSFNTFFSETGAGK
HVPRAVFVDLEPTVIDEVRTGTYRQLFHPEQLITGKEDAANNYARGHYTIGKEIIDLVLD
RIRKLADQCTGLQGFLVFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFSIYPAPQVSTA
VVEPYNSILTTHTTLEHSDCAFMVDNEAIYDICRRNLDIERPTYTNLNRLIGQIVSSITA
SLRFDGALNVDLTEFQTNLVPYPRIHFPLATYAPVISAEKAYHEQLSVAEITNACFEPAN
QMVKCDPRHGKYMACCLLYRGDVVPKDVNAAIATIKTKRTIQFVDWCPTGFKVGINYQPP
TVVPGGDLAKVQRAVCMLSNTTAIAEAWARLDHKFDLMYAKRAFVHWYVGEGMEEGEFSE
AREDMAALEKDYEEVGVDSVEGEGEEEGEEY
```

Figure 6 Pairwise comparison mode of HHpred. Amino acid sequences of *P. syntrophicum* FtsZ (UniProt ID: A0A5B9D775) and human Tubulin alpha-1A (UniProt ID: Q71U36) are shown as example input.

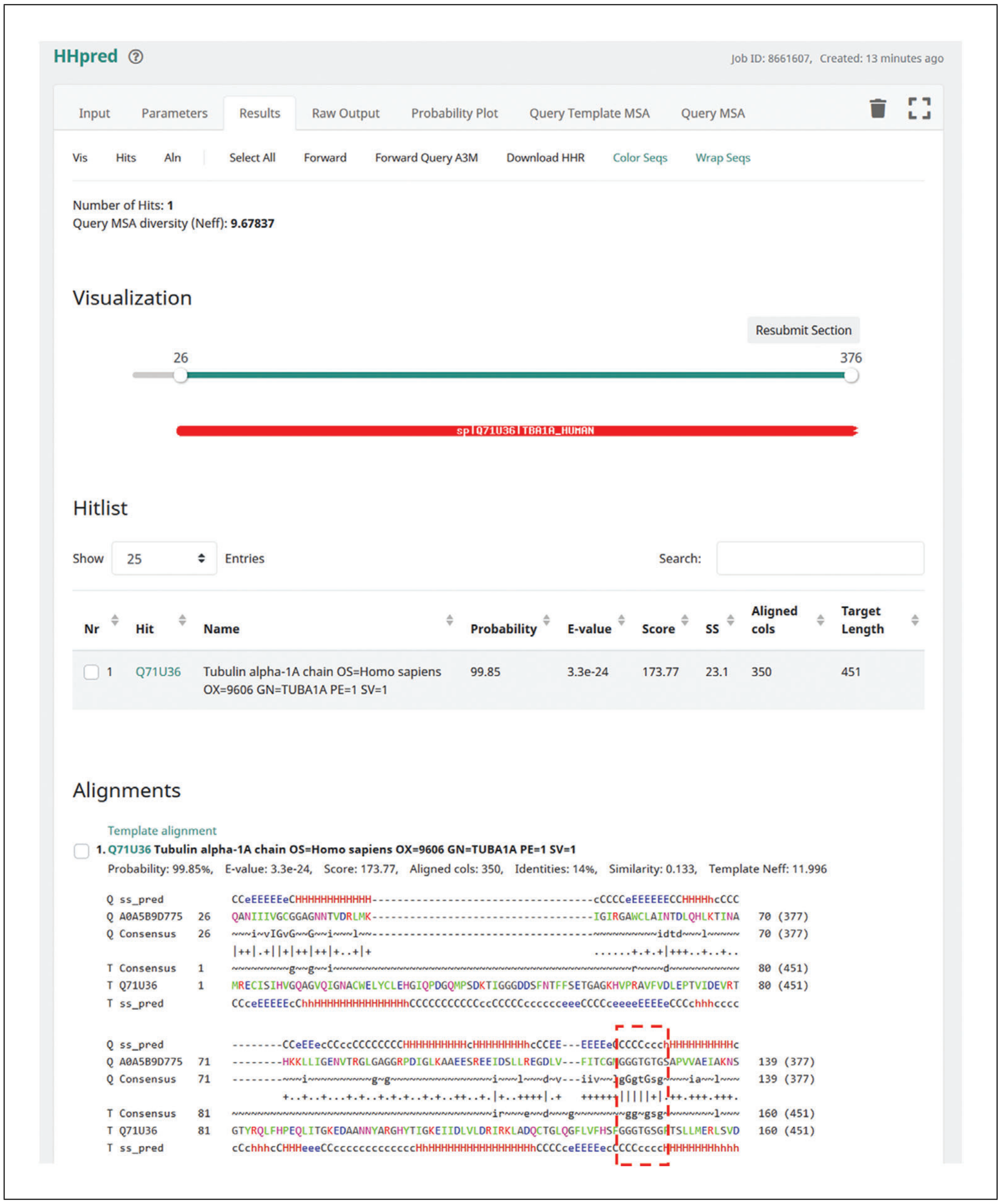
'Query MSA' (Fig. 7). For a detailed description of these tabs, please refer to steps 7-12 of Basic Protocol 1.

Figure 7 shows the outcome of comparing P. syntrophicum FtsZ with human Tubulin alpha-1A. In a comparison performed on September 3, 2020, HHpred matched them with a probability value of 99.85% and an E-value of 3.3e-24. Although the two sequences are very divergent and share a pairwise sequence identity of only 14%, the GTP-binding site is very clearly conserved in both sequences.

SUPPORT PROTOCOL

BUILDING A CUSTOM MULTIPLE SEQUENCE ALIGNMENT USING PSI-BLAST AND FORWARDING IT AS INPUT TO HHpred

The sensitivity of HHpred searches depends significantly on the quality of MSAs built for the query sequence. For building these MSAs, by default the HHpred server uses three iterations of HHblits over UniRef30 or allows using PSI-BLAST over nr70. Occasionally, however, the query MSAs may not be diverse enough or may be corrupted due to the inclusion of non-homologous sequences, resulting in no statistically significant matches or false positives, respectively. In such cases, using custom-built MSAs as input may significantly increase the sensitivity and reliability of an HHpred search. In the following, we show how to build a custom MSA using PSI-BLAST over a user-selected sequence database, such as the nonredundant protein sequence database (nr), UniProtKB/TrEMBL



HHpred ⓘ Job ID: 8661607, Created: 13 minutes ago

Input | Parameters | Results | Raw Output | Probability Plot | Query Template MSA | Query MSA

Vis | Hits | Aln | Select All | Forward | Forward Query A3M | Download HHR | Color Seqs | Wrap Seqs

Number of Hits: **1**
Query MSA diversity (Neff): **9.67837**

Visualization

Resubmit Section

Hitlist

Show 25 Entries Search:

Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	Q71U36	Tubulin alpha-1A chain OS=Homo sapiens OX=9606 GN=TUBA1A PE=1 SV=1	99.85	3.3e-24	173.77	23.1	350	451

Alignments

Template alignment
□ **1. Q71U36 Tubulin alpha-1A chain OS=Homo sapiens OX=9606 GN=TUBA1A PE=1 SV=1**
Probability: 99.85%, E-value: 3.3e-24, Score: 173.77, Aligned cols: 350, Identities: 14%, Similarity: 0.133, Template Neff: 11.996

```
Q ss_pred            CCeEEEEEeCHHHHHHHHHHHH-----------------------------------cCCCCeEEEEEECCHHHHhcCCC
Q A0A5B9D775   26  QANIIIVGCGGAGNNTVDRLMK-----------------------------------IGIRGAWCLAINTDLQHLKTINA   70 (377)
Q Consensus    26  ~~~i~vIGvG~~G~~i~~~l~~-----------------------------------~~~~~~~~~~idtd~~~l~~~~~   70 (377)
                   |++|.+||+|++|++|+..+|+                                   ......+.+.+|+++..+..+..
T Consensus     1  ~~~~~~~~~g~~g~~i~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~r~~~~d~~~~~~~~~~~   80 (451)
T Q71U36       1  MRECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMPSDKTIGGGDDSFNTFFSETGAGKHVPRAVFVDLEPTVIDEVRT   80 (451)
T ss_pred            CCceEEEEEcChhHHHHHHHHHHHHHHhCCCCCCCCCCccCCCCCcccccceeeCCCCceeeeEEEEeCCCchhhcccc

Q ss_pred            --------CCeEEecCCccCCCCCCCCHHHHHHHHHHcHHHHHHHHhcCCEE---EEEEeCCCCCccchHHHHHHHHHHc
Q A0A5B9D775   71  --------HKKLLIGENVTRGLGAGGRPDIGLKAAEESREEIDSLLREGDLV---FITCGMGGGTGTGSAPVVAEIAKNS  139 (377)
Q Consensus    71  --------~~~i~~~~~~~~~~g~g~~~~~~~~~~~~~~~~i~~~l~~~d~v---iiv~~lgGgtGsg~~~~ia~~l~~~  139 (377)
                           +..+..+...+.+..+.+.+..+.+..++..+.|+..++++|.+   ++++++|||||+||.++.+++.+++.
T Consensus    81  ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ir~~~e~~d~~~g~~~~~~~gg~gsg~~~~~~~~l~~~  160 (451)
T Q71U36      81  GTYRQLFHPEQLITGKEDAANNYARGHYTIGKEIIDLVLDRIRKLADQCTGLQGFLVFHSFGGGTGSGFTSLLMERLSVD  160 (451)
T ss_pred            cCchhhcCHHHeeeCCcccccccccccHhHHHHHHHHHHHHHHHhCCCCceEEEEecCCCCccccHHHHHHHHhhhh
```

Figure 7 Results of comparing *P. syntrophicum* FtsZ with human Tubulin alpha-1A. HHpred substantiated a homologous relationship between them at a probability value of 99.85% and an E-value of 3.3e-24. Although the two sequences share a pairwise sequence identity of only 14%, the pairwise alignment shows the clear conservation of the GTP-binding motif.

(uniprot_trembl), and UniProtKB/Swiss-Prot (uniprot_sprot), and forward the obtained alignment as input to HHpred.

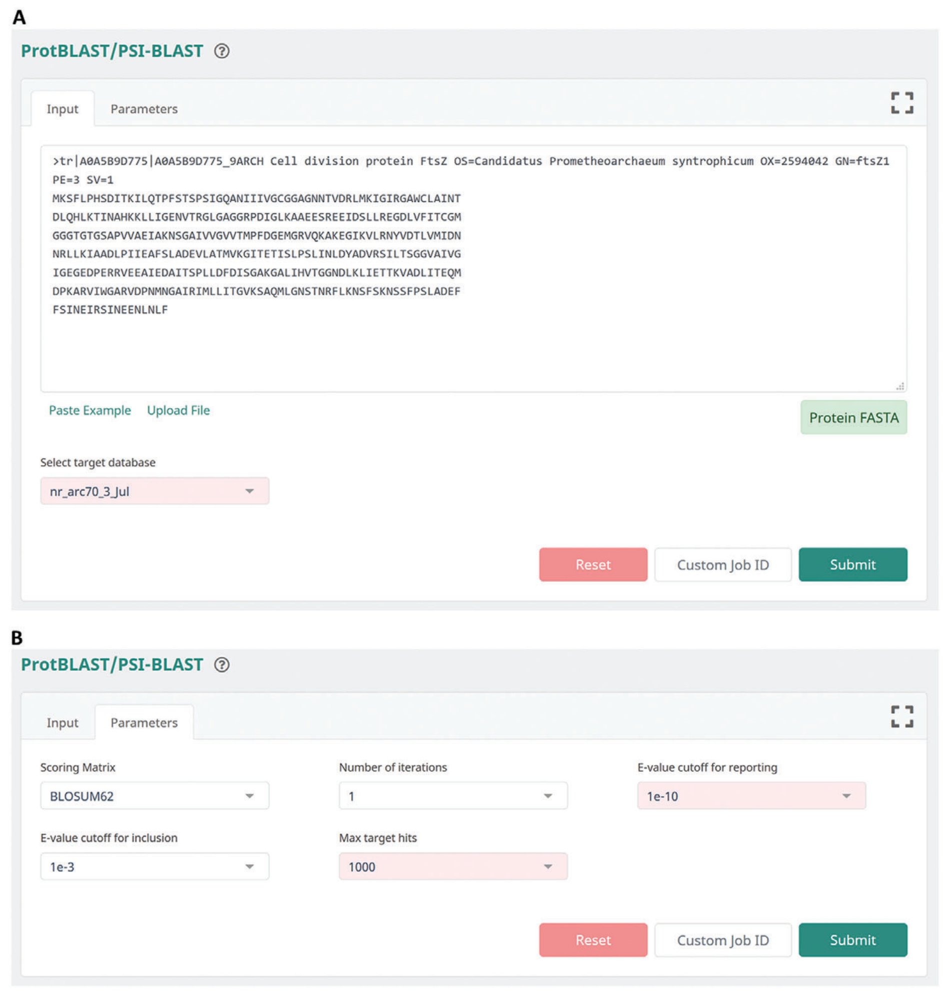
Necessary Resources

Same as for Basic Protocol 1

Submission page of PSI-BLAST

1. Navigate your Web browser to the submission page of PSI-BLAST at *https://toolkit.tuebingen.mpg.de/tools/psiblast*.

   The submission page of all tools within the Toolkit, including PSI-BLAST, is organized similarly to that of HHpred and includes two tabs, ‘Input’ and ‘Parameters’.



```
>tr|A0A5B9D775|A0A5B9D775_9ARCH Cell division protein FtsZ OS=Candidatus Prometheoarchaeum syntrophicum OX=2594042 GN=ftsZ1
PE=3 SV=1
MKSFLPHSDITKILQTPFSTSPSIGQANIIIVGCGGAGNNTVDRLMKIGIRGAWCLAINT
DLQHLKTINAHKKLLIGENVTRGLGAGGRPDIGLKAAEESREEIDSLLREGDLVFITCGM
GGGTGTGSAPVVAEIAKNSGAIVVGVVTMPFDGEMGRVQKAKEGIKVLRNYVDTLVMIDN
NRLLKIAADLPIIEAFSLADEVLATMVKGITETISLPSLINLDYADVRSILTSGGVAIVG
IGEGEDPERRVEEAIEDAITSPLLDFDISGAKGALIHVTGGNDLKLIETTKVADLITEQM
DPKARVIWGARVDPNMNGAIRIMLLITGVKSAQMLGNSTNRFLKNSFSKNSSFPSLADEF
FSINEIRSINEENLNLF
```

Select target database: nr_arc70_3_Jul

Scoring Matrix: BLOSUM62

Number of iterations: 1

E-value cutoff for reporting: 1e-10

E-value cutoff for inclusion: 1e-3

Max target hits: 1000

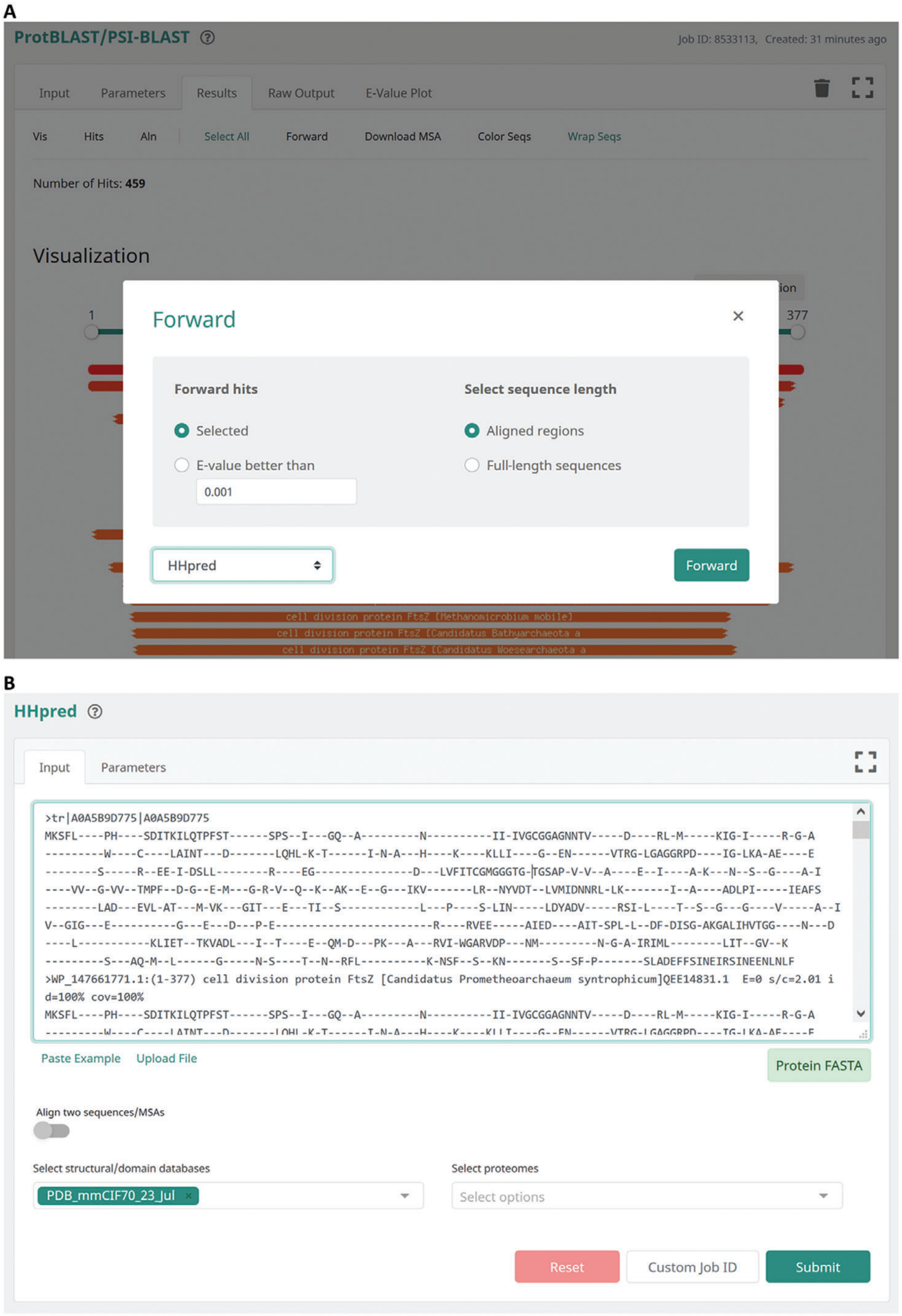
Figure 8 Submission page of PSI-BLAST. In the 'Input' tab (**A**), the amino acid sequence of FtsZ from *P. syntrophicum* in FASTA format (UniProt ID: A0A5B9D775) is shown as an example, and the target database is set to nr_arc70 (version of July 3, 2020). In the 'Parameters' tab (**B**), default values are set for all parameters, except 'E-value cutoff for reporting' (= 1e-10) and 'Max target hits' (= 1000).

2. Paste the amino acid sequence of your protein of interest (in FASTA format or as plain text) or an MSA (in FASTA, CLUSTAL, or STOCKHOLM format) into the large textbox (Fig. 8A). Alternatively, the input sequence or MSA can be uploaded using the 'Upload File' option.

   In Figure 8A, we use the amino acid sequence of FtsZ from P. syntrophicum in FASTA format as an example (UniProt ID: A0A5B9D775; NCBI ID: WP_147661771).

3. Select a target protein sequence database in the drop-down list over which you wish to build the alignment (Fig. 8A).

   The target sequence databases we currently offer include UniProtKB/TrEMBL (uniprot_trembl), UniProtKB/Swiss-Prot (uniprot_sprot), the protein sequences from the PDB (pdb_nr), and the nonredundant protein sequence database (nr). For the nr database, we also offer versions filtered (i) for sequence identity, such as nr30, nr50, and nr70; (ii) for taxonomy, such as nr_bac, which includes all bacterial sequences in nr, or nr_pro, which includes all bacterial and archaeal sequences; and (iii) for both sequence identity and taxonomy, such as nr_arc70, which includes all archaeal sequences in nr filtered for a maximum pairwise sequence identity of 70%. The nr and uniprot_trembl databases are not optimal for building deep, evolutionarily informative MSAs, as they contain too



Number of Hits: **459**

Forward

Forward hits

● Selected

○ E-value better than 0.001

Select sequence length

● Aligned regions

○ Full-length sequences

HHpred

Forward

```
>tr|A0A5B9D775|A0A5B9D775
MKSFL----PH----SDITKILQTPFST------SPS--I---GQ--A---------N----------II-IVGCGGAGNNTV-----D----RL-M-----KIG-I-----R-G-A
---------W----C----LAINT---D-------LQHL-K-T------I-N-A---H----K----KLLI----G--EN------VTRG-LGAGGRPD----IG-LKA-AE----E
--------S-----R--EE-I-DSLL--------R----EG--------------D---LVFITCGMGGGTG-TGSAP-V-V--A----E--I----A-K---N--S--G----A-I
----VV--G-VV--TMPF--D-G--E-M---G-R-V--Q--K--AK--E--G---IKV-------LR--NYVDT--LVMIDNNRL-LK-------I--A----ADLPI-----IEAFS
--------LAD---EVL-AT---M-VK---GIT---E---TI--S-----------L---P----S-LIN-----LDYADV-----RSI-L----T--S--G---G----V-----A--I
V--GIG---E----------G---E---D---P-E------------------------R----RVEE-----AIED----AIT-SPL-L--DF-DISG-AKGALIHVTGG----N---D
----L-----------KLIET--TKVADL---I--T----E--QM-D---PK---A---RVI-WGARVDP---NM---------N-G-A-IRIML--------LIT--GV--K
---------S---AQ-M--L------G-----N-S----T--N--RFL----------K-NSF--S--KN-------S--SF-P-------SLADEFFSINEIRSINEENLNLF
>WP_147661771.1:(1-377) cell division protein FtsZ [Candidatus Prometheoarchaeum syntrophicum]QEE14831.1  E=0 s/c=2.01 i
d=100% cov=100%
MKSFL----PH----SDITKILQTPFST------SPS--I---GQ--A---------N----------II-IVGCGGAGNNTV-----D----RL-M-----KIG-I-----R-G-A
---------W----C----LAINT---D-------LQHL-K-T------I-N-A---H----K----KLLI----G--EN------VTRG-LGAGGRPD----IG-LKA-AE----E
```

Align two sequences/MSAs

Select structural/domain databases: PDB_mmCIF70_23_Jul

Select proteomes: Select options

Figure 9 ‘Results’ tab of PSI-BLAST. On September 3, 2020, our search with *P. syntrophicum* FtsZ over nr_arc70 returned 459 matches (**A**); ‘E-value cutoff for reporting’ was set to 1e-10 and ‘Max target hits’ to 1000. An MSA of these matches can be forwarded as input to HHpred (**B**).

many closely related sequences, and searches over them predominantly return only such sequences. However, the nr70 database is a good choice for most standard cases, as it reduces redundancy while maximizing diversity. Nonetheless, nr70 might still be too redundant for certain proteins, with PSI-BLAST searches over it only returning closely related sequences. In such cases, we recommend using nr50, because it offers a good compromise between redundancy and diversity.

For our example sequence, we will run the search over the nr_arc70 to build the alignment (Fig. 8A).

4. Customize input parameters in the ‘Parameters’ tab (Fig. 8B).

   Detailed information on each parameter is available in the help pages. To build a deep alignment, one could make the search criteria less stringent by using permissive cutoffs for ‘E-value cutoff for reporting’ and ‘E-value cutoff for inclusion’, or use more search iterations (‘Number of iterations’). Contrarily, to build a focused alignment, use stringent cutoffs and just a single search iteration.

   For our example, we will set ‘E-value cutoff for reporting’ to 1e-10 and ‘Max target hits’ to 1000 (Fig. 8B).

5. Optionally, assign your job a custom identifier.

6. Start your search by pressing the ‘Submit’ button.

PSI-BLAST search results

7. Typical PSI-BLAST searches take about 3 min. However, searches involving long input sequences (>600 residues) over large databases such as nr could take much longer to complete.

   Upon successful completion, three new tabs are shown: ‘Results’, ‘Raw Output’, and ‘E-value Plot’. Similarly to the ‘Results’ page of HHpred, the ‘Results’ page of PSI-BLAST is also arranged into three sections: ‘Visualization’, ‘Hitlist’, and ‘Alignments’ (Fig. 9A).

   On September 3, 2020, our search yielded 459 matches (Fig. 9A).

Forwarding an alignment to HHpred

8. Inspect the obtained hits and unselect spurious ones.

   By default, all hits with an E-value better than the value set for the ‘E-value cutoff for inclusion’ are selected. To exclude seemingly non-homologous hits, inspect all pairwise alignments in the ‘Alignments’ section, especially those at the non-significant end, and unselect them. It is also often useful to exclude short alignments resulting from partial sequences in the databases.

9. Click on ‘Forward’ in the floating toolbar to forward an alignment of hits to HHpred. These can be the ones selected in step 8 (‘Selected’), or all hits satisfying an E-value cutoff (‘E-value better than’). In the ‘Forward’ modal, select ‘HHpred’ in the selection list at the bottom left corner and click on the ‘Forward’ button to send the alignment to HHpred (Fig. 9A).

   This will provide an entry to step 2 of Basic Protocols 1 and 2 (Fig. 9B).

CALCULATION OF HOMOLOGY MODELS USING HHpred AND MODELLER

BASIC PROTOCOL 2

The availability of 3D structures is extremely useful for the functional characterization of proteins. However, for many proteins, no experimental structures are available. Since proteins with recognizable sequence similarity generally also have quite similar 3D structures, a structure for a protein of interest can be modeled computationally from its sequence, based on homology to proteins of known structure. This approach is referred to as comparative modeling or homology modeling. In the following, we show how to use HHpred to select homologous templates of known structure for a query protein and how to extract and subsequently forward their alignment to MODELLER (Sali & Blundell, 1993), a popular program for homology modeling. Please refer to Basic Protocol 1 for detailed instructions on using HHpred.

Necessary Resources

Same as for Basic Protocol 1

Submission page of HHpred

1. Navigate your Web browser to the submission page of HHpred at *https://toolkit.tuebingen.mpg.de/tools/hhpred*.

2. Paste the amino acid sequence of your protein of interest (in FASTA format or as plain text) or an MSA (in FASTA, CLUSTAL, or STOCKHOLM format) into the large textbox. Alternatively, the input sequence or MSA can be uploaded using the ‘Upload File’ option. Follow Support Protocol to build a custom MSA, starting with a protein sequence of interest.

   We use the amino acid sequence of FtsZ from P. syntrophicum as an example (UniProt ID: A0A5B9D775; NCBI ID: WP_147661771); please refer to Figure 1A.

3. Select either PDB_mmCIF70 or PDB_mmCIF30 as the target database. If other target databases are included, the option for modeling is not offered.

   For our example, we will run the search over the PDB_mmCIF30 database (Fig. 1A).

4. Customize input parameters in the ‘Parameters’ tab. The default values for the various parameters are set to yield the best results for most standard cases, and we recommend using them, at least in the initial steps of the analysis.

   For our example, we will use default values for all input parameters, except ‘Max target hits’ = 500 (Fig. 1B).

5. Optionally, assign your job a custom identifier.

6. Start your search by pressing the ‘Submit’ button.

Selecting templates for modeling

7. On the ‘Results’ page of HHpred, analyze the obtained templates and select one or more as wished (Fig. 10A). Next, click on ‘Model using selection’ in the floating toolbar offered at the top of the tab to generate an alignment of the query and the selected templates. If a user clicks on ‘Model using selection’ without having selected any template, the first hit is used for modeling.

   Refer to Basic Protocol 1 for an explanation of the HHpred ‘Results’ page. The quality of the calculated homology model depends primarily on the selected templates. We recommend inspecting the probability values, E-values, identities, secondary structure scores, and alignments to identify the best templates. High-scoring templates with lengths similar to that of the query and few gapped columns in the pairwise query-template alignment generally represent the best templates.

   On September 3, 2020, our search yielded several high-scoring templates with the same length as our query. We selected the first five matches as templates for modeling (Fig. 10A).

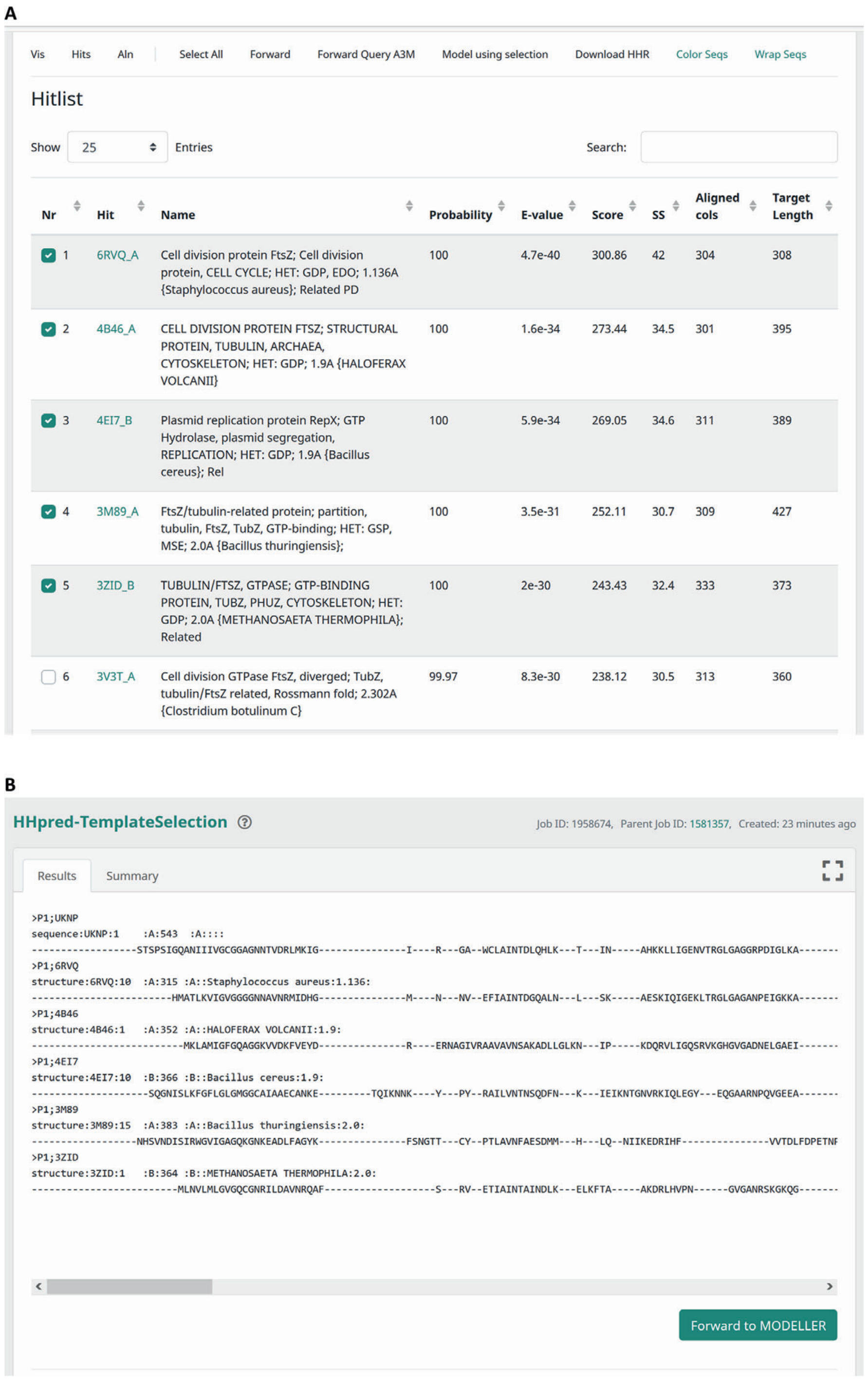
8. After clicking on ‘Model using selection’, the query-template alignment is displayed in PIR format in a new job view after a few minutes (Fig. 10B). Click on the ‘Forward to MODELLER’ button to send this alignment as input to MODELLER.

Running MODELLER

9. On the submission page of MODELLER, enter your MODELLER key (Fig. 11A).

   To obtain a key for MODELLER, please register at the following site: http://salilab.org/modeller/registration.shtml. MODELLER is available free of charge to academic users.

10. Optionally, enter a custom job identifier and click on the ‘Submit’ button to start the job.



	Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
☑	1	6RVQ_A	Cell division protein FtsZ; Cell division protein, CELL CYCLE; HET: GDP, EDO; 1.136A {Staphylococcus aureus}; Related PD	100	4.7e-40	300.86	42	304	308
☑	2	4B46_A	CELL DIVISION PROTEIN FTSZ; STRUCTURAL PROTEIN, TUBULIN, ARCHAEA, CYTOSKELETON; HET: GDP; 1.9A {HALOFERAX VOLCANII}	100	1.6e-34	273.44	34.5	301	395
☑	3	4EI7_B	Plasmid replication protein RepX; GTP Hydrolase, plasmid segregation, REPLICATION; HET: GDP; 1.9A {Bacillus cereus}; Rel	100	5.9e-34	269.05	34.6	311	389
☑	4	3M89_A	FtsZ/tubulin-related protein; partition, tubulin, FtsZ, TubZ, GTP-binding; HET: GSP, MSE; 2.0A {Bacillus thuringiensis};	100	3.5e-31	252.11	30.7	309	427
☑	5	3ZID_B	TUBULIN/FTSZ, GTPASE; GTP-BINDING PROTEIN, TUBZ, PHUZ, CYTOSKELETON; HET: GDP; 2.0A {METHANOSAETA THERMOPHILA}; Related	100	2e-30	243.43	32.4	333	373
☐	6	3V3T_A	Cell division GTPase FtsZ, diverged; TubZ, tubulin/FtsZ related, Rossmann fold; 2.302A {Clostridium botulinum C}	99.97	8.3e-30	238.12	30.5	313	360

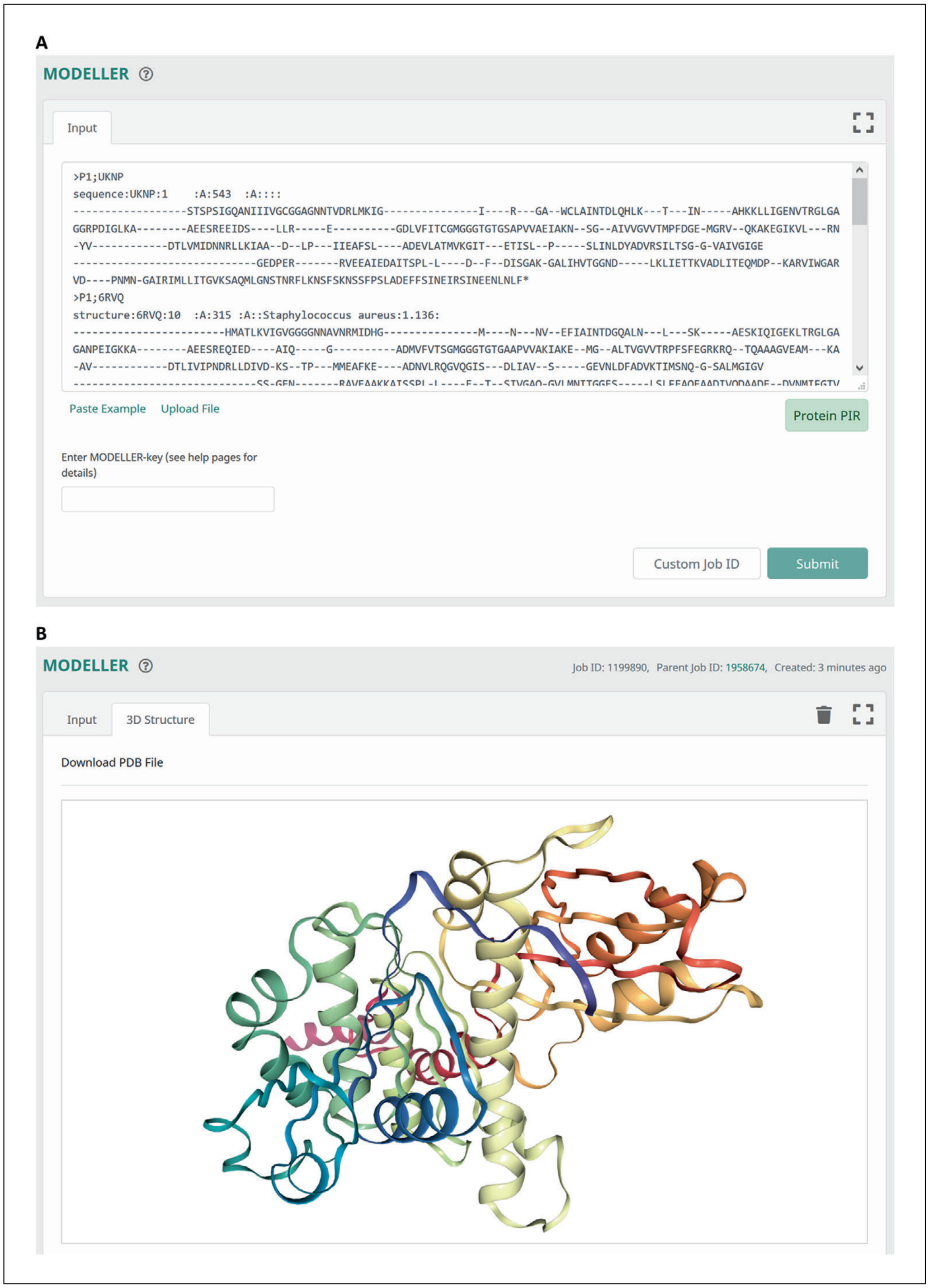
Job ID: 1958674, Parent Job ID: 1581357, Created: 23 minutes ago

```
>P1;UKNP
sequence:UKNP:1    :A:543  :A::::
------------------STSPSIGQANIIIVGCGGAGNNTVDRLMKIG---------------I----R---GA--WCLAINTDLQHLK---T---IN-----AHKKLLIGENVTRGLGAGGRPDIGLKA-------
>P1;6RVQ
structure:6RVQ:10  :A:315 :A::Staphylococcus aureus:1.136:
------------------------HMATLKVIGVGGGGNNAVNRMIDHG---------------M----N---NV--EFIAINTDGQALN---L---SK-----AESKIQIGEKLTRGLGAGANPEIGKKA-------
>P1;4B46
structure:4B46:1   :A:352 :A::HALOFERAX VOLCANII:1.9:
--------------------------MKLAMIGFGQAGGKVVDKFVEYD---------------R----ERNAGIVRAAVAVNSAKADLLGLKN---IP-----KDQRVLIGQSRVKGHGVGADNELGAEI-------
>P1;4EI7
structure:4EI7:10  :B:366 :B::Bacillus cereus:1.9:
--------------------SQGNISLKFGFLGLGMGGCAIAAECANKE---------TQIKNNK----Y---PY--RAILVNTNSQDFN---K---IEIKNTGNVRKIQLEGY---EQGAARNPQVGEEA-------
>P1;3M89
structure:3M89:15  :A:383 :A::Bacillus thuringiensis:2.0:
------------------NHSVNDISIRWGVIGAGQKGNKEADLFAGYK---------------FSNGTT---CY--PTLAVNFAESDMM---H---LQ--NIIKEDRIHF---------------VVTDLFDPETNF
>P1;3ZID
structure:3ZID:1   :B:364 :B::METHANOSAETA THERMOPHILA:2.0:
-------------------------MLNVLMLGVGQCGNRILDAVNRQAF-------------------S---RV--ETIAINTAINDLK---ELKFTA-----AKDRLHVPN------GVGANRSKGKQG-------
```

Figure 10 Selection of templates for modeling. An HHpred search with *P. syntrophicum* FtsZ over PDB_mmCIF30 (performed on September 3, 2020) returned several high-scoring matches. The first five hits were selected as templates for modeling (**A**), a query-template alignment was generated by clicking on ‘Model using selection’ (**B**), and forwarded to MODELLER.

11. Typical MODELLER jobs take less than 5 min to run through. In the resulting view, the modeled structure is displayed in the NGL Viewer application (Rose et al., 2018) for molecular visualization, and an option is also provided to download the atomic coordinates (Fig. 11B).

Figure 11B shows the structural model generated for FtsZ from P. syntrophicum.



```
>P1;UKNP
sequence:UKNP:1    :A:543  :A::::
-----------------STSPSIGQANIIIVGCGGAGNNTVDRLMKIG--------------I----R---GA--WCLAINTDLQHLK---T---IN-----AHKKLLIGENVTRGLGA
GGRPDIGLKA--------AEESREEIDS----LLR-----E----------GDLVFITCGMGGGTGTGSAPVVAEIAKN--SG--AIVVGVVTMPFDGE-MGRV--QKAKEGIKVL---RN
-YV-----------DTLVMIDNNRLLKIAA--D--LP---IIEAFSL----ADEVLATMVKGIT---ETISL--P-----SLINLDYADVRSILTSG-G-VAIVGIGE
----------------------------GEDPER-------RVEEAIEDAITSPL-L----D--F--DISGAK-GALIHVTGGND-----LKLIETTKVADLITEQMDP--KARVIWGAR
VD----PNMN-GAIRIMLLITGVKSAQMLGNSTNRFLKNSFSKNSSFPSLADEFFSINEIRSINEENLNLF*
>P1;6RVQ
structure:6RVQ:10  :A:315 :A::Staphylococcus aureus:1.136:
-----------------------HMATLKVIGVGGGGNNAVNRMIDHG--------------M----N---NV--EFIAINTDGQALN---L---SK-----AESKIQIGEKLTRGLGA
GANPEIGKKA--------AEESREQIED----AIQ-----G----------ADMVFVTSGMGGGTGTGAAPVVAKIAKE--MG--ALTVGVVTRPFSFEGRKRQ--TQAAAGVEAM---KA
-AV-----------DTLIVIPNDRLLDIVD-KS--TP---MMEAFKE----ADNVLRQGVQGIS---DLIAV--S-----GEVNLDFADVKTIMSNQ-G-SALMGIGV
----------------------------SS-GEN-------RAVEAAKKATSSPL-L----E--T--STVGAQ-GVLMNTTGGES-----LSLEEAQEAADTVQDAADE--DVNMIEGTV
```

Figure 11 Submission and output pages of MODELLER. The shown input (**A**) was generated as described in Basic Protocol 2, step 8. The 3D structural model built for *P. syntrophicum* FtsZ is shown in **B**.

BASIC PROTOCOL 3

CLUSTER ANALYSIS USING CLANS

Although HHpred is extremely powerful in detecting remote homologs of a protein of interest, it may occasionally not find any meaningful connections because the available target databases contain only a significantly filtered subset of known sequences. Since building profile HMMs is computationally expensive, we currently do not include profile HMMs of large sequence databases for HHpred. In the following, we show how the speed of PSI-BLAST and the power of all-against-all pairwise comparisons can be used to detect remote homologs. For this, we will exploit the observation that the non-significant part of PSI-BLAST searches frequently contains many biologically meaningful connections (indeed, for most proteins, most homologs have non-significant scores). These non-significant pairwise connections nevertheless collectively allow sequences to cluster within a larger sequence space, revealing family relationships. Here, we combine PSI-BLAST with our cluster analysis tool, CLANS (Frickey & Lupas, 2004), to illustrate this approach. CLANS is an implementation of the Fruchterman-Reingold graph-drawing algorithm. It represents protein sequences as points in a virtual 2D or 3D space and allows them to attract or repel each other in proportion to the statistical significance of their all-against-all pairwise comparison. CLANS then searches for the global energy minimum of this landscape of forces, yielding a map in which related sequences group together in connected clusters and unrelated ones drift to the periphery. Here, we will collect input sequences by exploiting the Toolkit implementation of PSI-BLAST, which allows changing the target database between iterations. We will start by building a high-quality PSI-BLAST profile on a small, focused database (nr_arc70) using a strict E-value cutoff (1e-10), then search for all potential homologs up to an E-value of 100 in a more comprehensive database (uniport_prot). We will then forward the obtained hits to the CLANS tool within the Toolkit to carry out an all-against-all sequence comparison. Finally, we will load the resulting CLANS file into the CLANS desktop application to visualize the relationships between the hits and identify groups of related sequences.

Necessary Resources

Same as for Basic Protocol 1, but additionally, the Java Runtime Environment or Java Development Kit (*https://www.oracle.com/java/technologies/javase-jre8-downloads.html* or *https://openjdk.java.net/install*) needs to be installed on the hardware, and at least 4 GB RAM is recommended

PSI-BLAST–Iteration 1

1. Navigate your Web browser to the submission page of PSI-BLAST at *https://toolkit.tuebingen.mpg.de/tools/psiblast*.

2. Paste the amino acid sequence of your protein of interest (in FASTA format or as plain text; Fig. 12A). Alternatively, the input sequence can be uploaded using the 'Upload File' option.

   In Figure 12A, we use the amino acid sequence of FtsZ from P. syntrophicum in FASTA format as an example (UniProt ID: A0A5B9D775; NCBI ID: WP_147661771).

3. Select a target database. For more information on this step, please refer to Support Protocol.

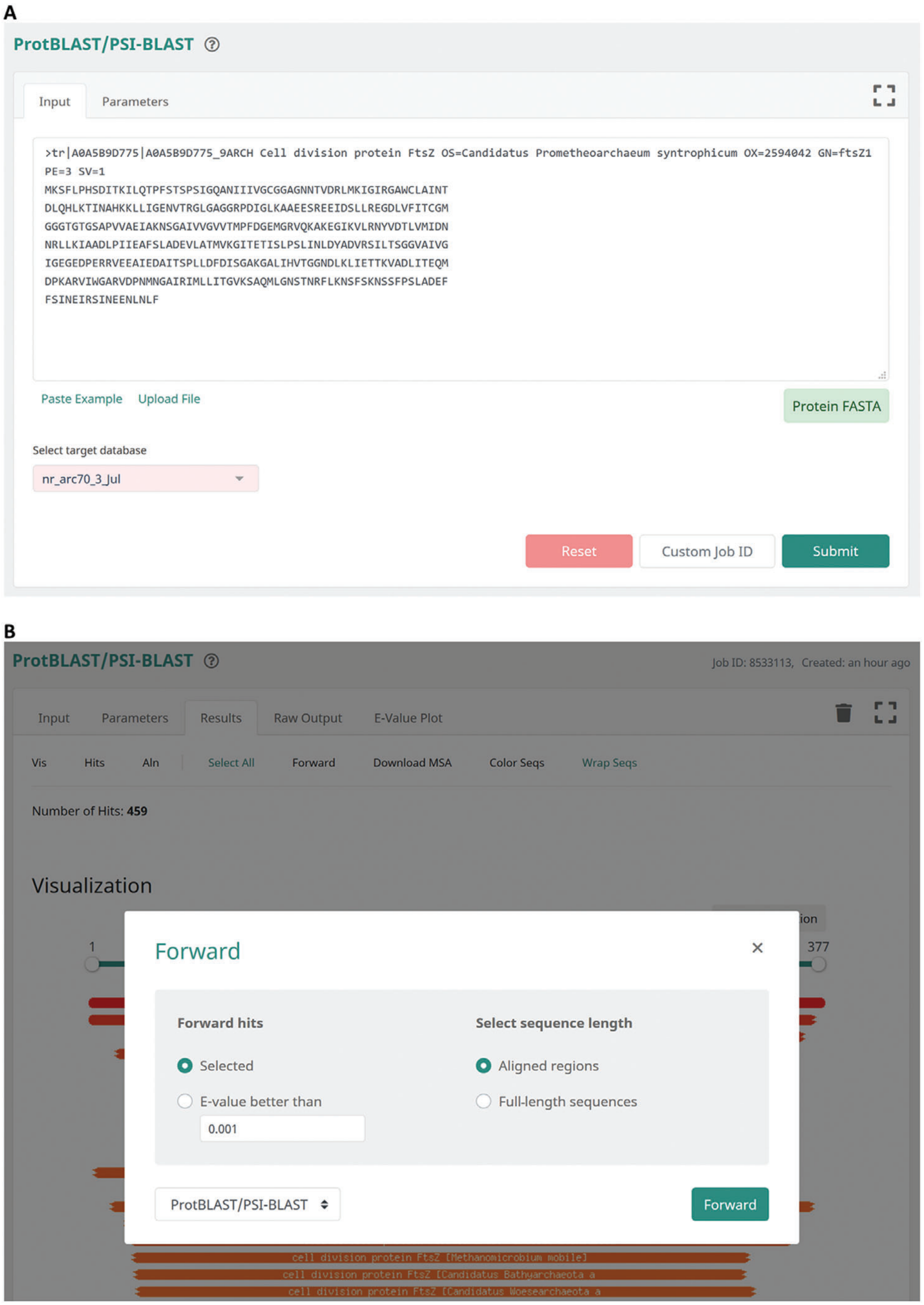
   For our example sequence, which originates from an archaeon, we will create a focused alignment to be used as the input for the second iteration of PSI-BLAST by searching for its homologs in nr_arc70.

4. Customize input parameters in the 'Parameters' tab. Detailed information on each parameter is available in the help pages.

   For our example, we will use a stringent E-value cutoff ('E-value cutoff for reporting' = 1e-10) to reduce the chances of obtaining spurious hits. We will set 'Max target hits' to 5000 and use default values for all other parameters.

5. Optionally, assign your job a custom identifier and click on the 'Submit' button to start the first iteration of PSI-BLAST.

   Typical PSI-BLAST searches take about 3 min. However, searches involving long input sequences (>600 residues) over large databases such as nr could take much longer to complete. Upon completion of the job, the 'Results' tab of PSI-BLAST will be shown.



```
>tr|A0A5B9D775|A0A5B9D775_9ARCH Cell division protein FtsZ OS=Candidatus Prometheoarchaeum syntrophicum OX=2594042 GN=ftsZ1
PE=3 SV=1
MKSFLPHSDITKILQTPFSTSPSIGQANIIIVGCGGAGNNTVDRLMKIGIRGAWCLAINT
DLQHLKTINAHKKLLIGENVTRGLGAGGRPDIGLKAAEESREEIDSLLREGDLVFITCGM
GGGTGTGSAPVVAEIAKNSGAIVVGVVTMPFDGEMGRVQKAKEGIKVLRNYVDTLVMIDN
NRLLKIAADLPIIEAFSLADEVLATMVKGITETISLPSLINLDYADVRSILTSGGVAIVG
IGEGEDPERRVEEAIEDAITSPLLDFDISGAKGALIHVTGGNDLKLIETTKVADLITEQM
DPKARVIWGARVDPNMNGAIRIMLLITGVKSAQMLGNSTNRFLKNSFSKNSSFPSLADEF
FSINEIRSINEENLNLF
```

Select target database: nr_arc70_3_Jul

Job ID: 8533113, Created: an hour ago

Number of Hits: **459**

Forward hits
- ● Selected
- ○ E-value better than 0.001

Select sequence length
- ● Aligned regions
- ○ Full-length sequences

ProtBLAST/PSI-BLAST

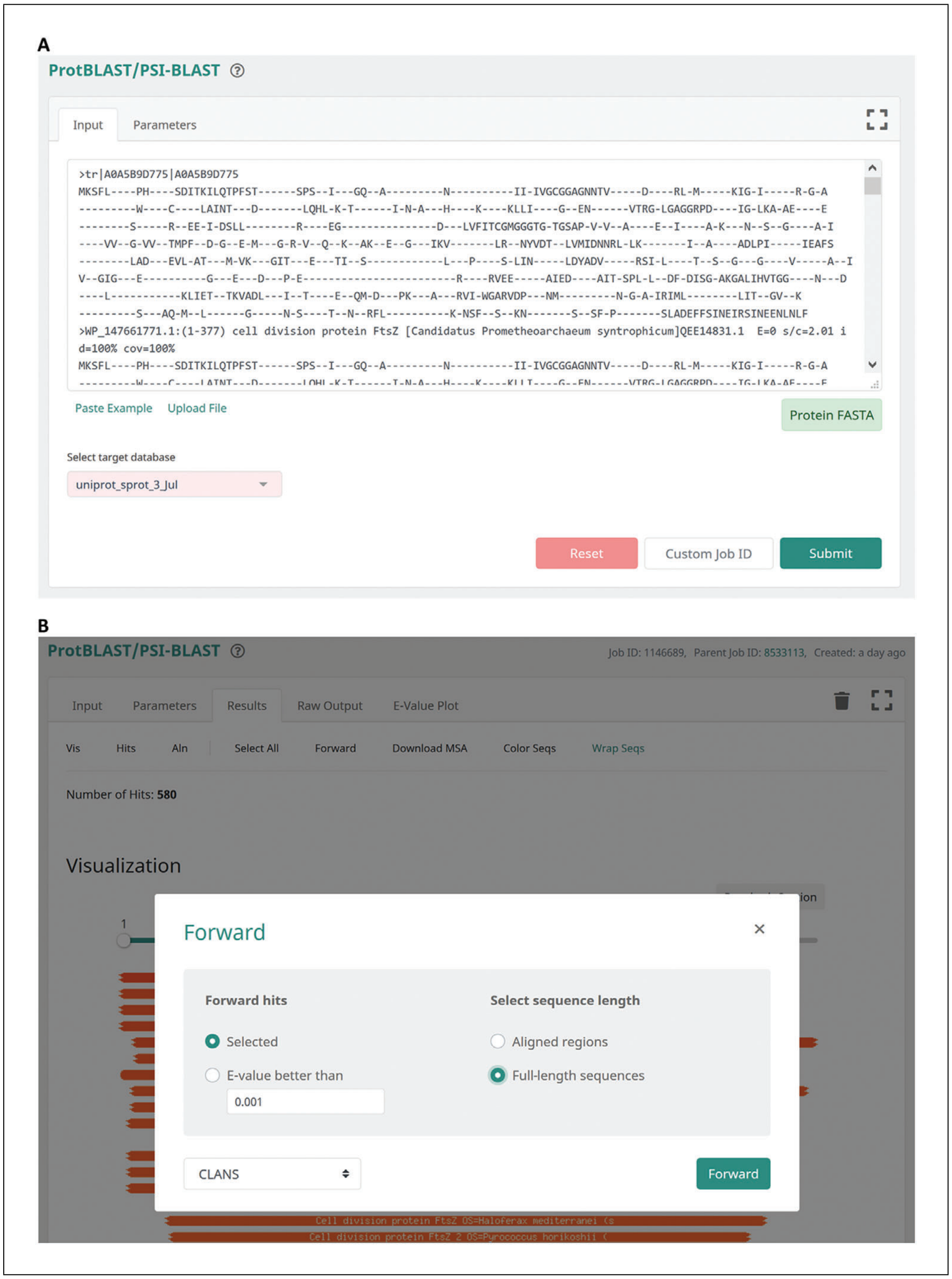
cell division protein FtsZ [Methanomicrobium mobile]
cell division protein FtsZ [Candidatus Bathyarchaeota a
cell division protein FtsZ [Candidatus Woesearchaeota a

Figure 12 ‘Input’ and ‘Results’ tabs of PSI-BLAST–iteration 1. In the first iteration of PSI-BLAST with *P. syntrophicum* FtsZ over nr_arc70 (**A**), 459 hits were identified; ‘E-value cutoff for reporting’ was set to 1e-10 and ‘Max target hits’ to 5000. An MSA of these hits was forwarded as input to PSI-BLAST (**B**).

PSI-BLAST–Iteration 2

6. To initiate the second iteration using an MSA of the obtained hits, forward these to PSI-BLAST by clicking on ‘Forward’ in the floating toolbar on the ‘Results’ page of PSI-BLAST (Fig. 12B). ‘PSI-BLAST’ is selected by default. Press the ‘Forward’ button to forward the MSA.

On September 3, 2020, our search yielded 459 matches (Fig. 12B).



```
>tr|A0A5B9D775|A0A5B9D775
MKSFL----PH----SDITKILQTPFST------SPS--I---GQ--A---------N----------II-IVGCGGAGNNTV-----D----RL-M-----KIG-I-----R-G-A
---------W----C----LAINT---D-------LQHL-K-T------I-N-A---H----K----KLLI----G--EN------VTRG-LGAGGRPD----IG-LKA-AE----E
--------S-----R--EE-I-DSLL--------R----EG---------------D---LVFITCGMGGGTG-TGSAP-V-V--A----E--I----A-K---N--S--G----A-I
----VV--G-VV--TMPF--D-G--E-M---G-R-V--Q--K--AK--E--G---IKV-------LR--NYVDT--LVMIDNNRL-LK-------I--A----ADLPI-----IEAFS
--------LAD---EVL-AT---M-VK---GIT---E---TI--S------------L---P----S-LIN-----LDYADV-----RSI-L----T--S--G---G----V-----A--I
V--GIG---E----------G---E---D---P-E------------------------R----RVEE-----AIED----AIT-SPL-L--DF-DISG-AKGALIHVTGG----N---D
----L-----------KLIET--TKVADL---I--T----E--QM-D---PK---A---RVI-WGARVDP---NM---------N-G-A-IRIML--------LIT--GV--K
---------S---AQ-M--L------G-----N-S----T--N--RFL----------K-NSF--S--KN-------S--SF-P-------SLADEFFSINEIRSINEENLNLF
>WP_147661771.1:(1-377) cell division protein FtsZ [Candidatus Prometheoarchaeum syntrophicum]QEE14831.1  E=0 s/c=2.01 i
d=100% cov=100%
MKSFL----PH----SDITKILQTPFST------SPS--I---GQ--A---------N----------II-IVGCGGAGNNTV-----D----RL-M-----KIG-I-----R-G-A
```

Select target database: uniprot_sprot_3_Jul

Job ID: 1146689, Parent Job ID: 8533113, Created: a day ago

Number of Hits: **580**

Forward

Forward hits
● Selected
○ E-value better than 0.001

Select sequence length
○ Aligned regions
● Full-length sequences

CLANS

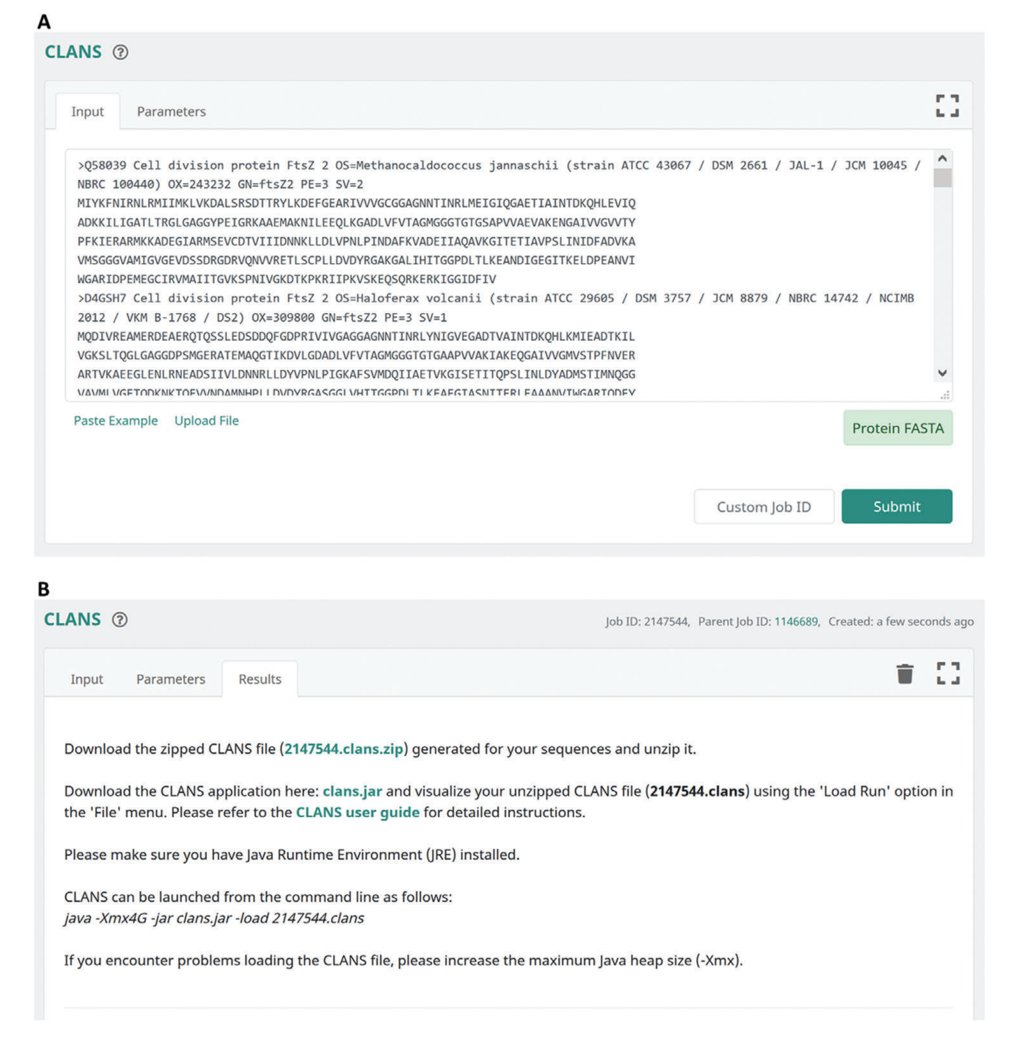
Forward
Cell division protein FtsZ OS=Haloferax mediterranei (s
Cell division protein FtsZ 2 OS=Pyrococcus horikoshii (

Figure 13 ‘Input’ and ‘Results’ tabs of PSI-BLAST–iteration 2. The second iteration using the forwarded MSA (see Basic Protocol 3, step 6) over uniprot_sprot (**A**) yielded 580 hits; ‘E-value cutoff for reporting’ was set to 100 and ‘Max target hits’ to 5000. Full-length sequences of these hits were forwarded as input to CLANS (**B**).

7. On the submission page of PSI-BLAST, select a target protein sequence database (Fig. 13A) and customize input parameters for the second iteration of PSI-BLAST.

   For our example, we will search the more comprehensive uniprot_sprot database (Fig. 13A). Since we wish to detect biologically interesting matches among lower-scoring sequences, we will set the ‘E-value cutoff for reporting’ to 100. We will use default values for all other parameters, except ‘Max target hits’ (= 5000).

8. Optionally, assign your job a custom identifier and click on the ‘Submit’ button to start the second iteration of PSI-BLAST.



```
>Q58039 Cell division protein FtsZ 2 OS=Methanocaldococcus jannaschii (strain ATCC 43067 / DSM 2661 / JAL-1 / JCM 10045 /
NBRC 100440) OX=243232 GN=ftsZ2 PE=3 SV=2
MIYKFNIRNLRMIIMKLVKDALSRSDTTRYLKDEFGEARIVVVGCGGAGNNTINRLMEIGIQGAETIAINTDKQHLEVIQ
ADKKILIGATLTRGLGAGGYPEIGRKAAEMAKNILEEQLKGADLVFVTAGMGGGTGTGSAPVVAEVAKENGAIVVGVVTY
PFKIERARMKKADEGIARMSEVCDTVIIIDNNKLLDLVPNLPINDAFKVADEIIAQAVKGITETIAVPSLINIDFADVKA
VMSGGGVAMIGVGEVDSSDRGDRVQNVVRETLSCPLLDVDYRGAKGALIHITGGPDLTLKEANDIGEGITKELDPEANVI
WGARIDPEMEGCIRVMAIITGVKSPNIVGKDTKPKRIIPKVSKEQSQRKERKIGGIDFIV
>D4GSH7 Cell division protein FtsZ 2 OS=Haloferax volcanii (strain ATCC 29605 / DSM 3757 / JCM 8879 / NBRC 14742 / NCIMB
2012 / VKM B-1768 / DS2) OX=309800 GN=ftsZ2 PE=3 SV=1
MQDIVREAMERDEAERQTQSSLEDSDDQFGDPRIVIVGAGGAGNNTINRLYNIGVEGADTVAINTDKQHLKMIEADTKIL
VGKSLTQGLGAGGDPSMGERATEMAQGTIKDVLGDADLVFVTAGMGGGTGTGAAPVVAKIAKEQGAIVVGMVSTPFNVER
ARTVKAEEGLENLRNEADSIIVLDNNRLLDYVPNLPIGKAFSVMDQIIAETVKGISETITQPSLINLDYADMSTIMNQGG
VAVMLVGETQDKNKTQEVVNDAMNHPLLDVDYRGASGGLVHTTGGPDLTLKEAEGTASNTTERLEAAANVTWGARTQDEY
```

Job ID: 2147544, Parent Job ID: 1146689, Created: a few seconds ago

Download the zipped CLANS file (2147544.clans.zip) generated for your sequences and unzip it.

Download the CLANS application here: clans.jar and visualize your unzipped CLANS file (**2147544.clans**) using the 'Load Run' option in the 'File' menu. Please refer to the CLANS user guide for detailed instructions.

Please make sure you have Java Runtime Environment (JRE) installed.

CLANS can be launched from the command line as follows:

java -Xmx4G -jar clans.jar -load 2147544.clans

If you encounter problems loading the CLANS file, please increase the maximum Java heap size (-Xmx).

Figure 14 ‘Input’ and ‘Results’ tabs of CLANS. The forwarded full-length sequences (see Basic Protocol 3, step 10) were used as input (**A**). The ‘Results’ tab of CLANS provides hyperlinks for downloading the computed CLANS file, the CLANS desktop application, and a user guide (**B**).

Forwarding sequences to CLANS

9. On the ‘Results’ page of PSI-BLAST, click on ‘Select All’ to select all obtained matches (Fig. 13B). By default, only matches with E-values lower than the ‘E-value cutoff for inclusion’ are selected.

   On September 3, 2020, our search yielded 580 matches (Fig. 13B).

10. Click on ‘Forward’ to send sequences of the obtained matches to CLANS. In the ‘Forward’ modal, select ‘CLANS’ in the drop-down list at the bottom left corner, select ‘Full-length sequences’, and click on the ‘Forward’ button (Fig. 13B).

11. On the submission page of CLANS, click on the ‘Submit’ button to start the job (Fig. 14A).

Visualizing sequence relationships in CLANS

12. Depending on the number and length of the input sequences, the CLANS server needs up to several hours to perform all-against-all pairwise PSI-BLAST comparisons. The output page of CLANS offers hyperlinks to the computed raw file in ZIP format, which contains the all-against-all pairwise similarity scores as measured

by PSI-BLAST *P*-values, the CLANS application (`clans.jar`), and a user guide (Fig. 14B).

13. Download the zipped raw file and extract it.

    On the Linux command line, the raw file can be uncompressed using the unzip command:

    ```
    $ unzip filename.clans.zip
    ```

    To extract the raw file under Windows OS, right-click it and select 'Extract All'.

14. Download the CLANS application and launch it. To run CLANS, you will need to have a Java Runtime Environment (JRE) or a Java Development Kit (JDK) installed. On the Linux command line, CLANS can be launched using the following command:

    ```
    $ java [-Xmx4G] -jar clans.jar [-load <filename.clans>]
    ```

    The optional parameter '`-load`' specifies the path of the CLANS file computed by the Toolkit. If it is omitted, the CLANS GUI is launched with an empty map. The CLANS file can be loaded subsequently using File > Load Run. For loading and displaying large maps, CLANS might require large amounts of memory. The memory available to CLANS can be increased using the '`-Xmx`' parameter. For instance, using '`-Xmx8G`' directs the Java Virtual Machine to allocate a memory of 8G for CLANS.

    Alternatively, CLANS can also be launched by double-clicking the CLANS jar file (`clans.jar`); however, this does not allow increasing the memory available to CLANS.

15. In the cluster map loaded within the CLANS application, protein sequences are shown as dots (nodes). Initially, they are placed randomly in a 3D space. Lines (edges) connecting the dots can be shown by selecting 'show connections' in the bottom panel. They reflect the significance of the sequence similarity between the sequences; the darker a line, the higher the sequence similarity.

    To generate a publication-quality image, we recommend using a 2D space instead of the default 3D space (Misc > Cluster in 2D).

16. Click on 'Start run' in the bottom panel to start a clustering run.

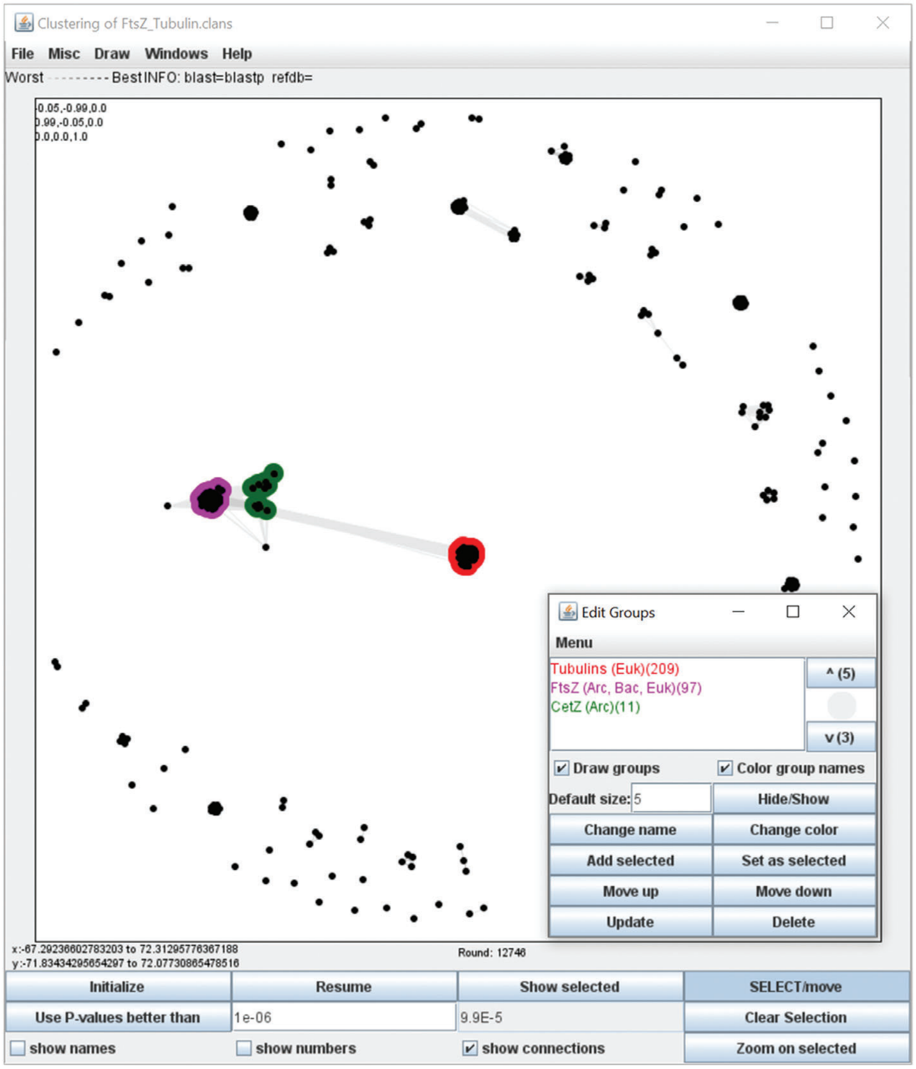
    Dots move around iteratively in the virtual 2D or 3D space based on the force vectors resulting from all pairwise interactions, until they find their equilibrium position. Then, the overall movement of the dots with respect to each other becomes negligible. Maps containing a few hundred sequences typically find their equilibrium after a few thousand iterations (rounds), whereas maps containing more than 5000 sequences may need more than 10,000 iterations. In the equilibrated map, groups of similar sequences come to lie together, forming tightly connected clusters. To speed up clustering, we recommend hiding the edges (unselect 'show connections' in the bottom panel) and redrawing the map less frequently (set Misc > Only draw every Nth round to a higher number).

17. Analyze the substructure of the map by varying the *P*-value cutoff used for clustering. To choose a new *P*-value cutoff, click on 'Stop' in the bottom panel to pause the clustering run, enter a value in the 'Use *P*-values better than' text field (e.g., 1e-06 or 0.000001), click on the 'Use *P*-values better than' button, and click on 'Resume'.

    Upon choosing a stringent P-value cutoff, all pairwise connections with a higher P-value are discarded and not considered for clustering. Consequently, previously compact clusters are partitioned into multiple sub-clusters.

    For our example, we will use a P-value cutoff of 1e-06.

18. Analyze the obtained clusters, annotate them, and produce a publication-quality image of the cluster map.



Clustering of FtsZ_Tubulin.clans

Worst --------- Best INFO: blast=blastp refdb=

-0.05,-0.99,0.0
0.99,-0.05,0.0
0.0,0.0,1.0

x:-67.29236602783203 to 72.31295776367188
y:-71.83434295654297 to 72.07730865478516

Round: 12746

Edit Groups

Tubulins (Euk)(209)
FtsZ (Arc, Bac, Euk)(97)
CetZ (Arc)(11)

☑ Draw groups ☑ Color group names

Default size: 5

Use P-values better than 1e-06 9.9E-5

□ show names □ show numbers ☑ show connections

Figure 15 Cluster map of FtsZ and Tubulin sequences. The raw CLANS file obtained in Basic Protocol 3, step 13 was loaded, clustered, and annotated in the CLANS desktop application. In the map, dots represent sequences, and line coloring reflects sequence similarity; the darker a line, the higher the similarity. Tubulins are colored red, FtsZ magenta, and CetZ green.

To obtain information on a group of sequences, click on 'select/MOVE' in the bottom panel, drag the mouse over a region of interest to select a group of sequences, and click on 'Show selected'.

To color a group of sequences, select them and use Windows > Edit Groups. Next, in the 'Edit Groups' window, click on 'Add selected' to add them to a group, select 'Draw groups', and click on 'Update'. In this window, options to change the shape, size, or color of dots within a group are also provided.

To generate an image of the annotated cluster map, turn on the antialiasing mode (Draw > Antialiasing) for nicer graphics, resize the CLANS window as desired, take a screenshot, and paste the captured image into an image editor (e.g., GIMP). Alternatively, the map can be saved as a PDF or PostScript file using File > Print view (select the 'Print to file' option). A clustering session can be saved to a file and loaded from a file using File > Save Run and File > Load Run, respectively.

For detailed instructions on using CLANS, please refer to the user guide offered for download on the 'Results' page of CLANS.

In the map obtained on September 3, 2020, FtsZ (colored magenta) and Tubulin (red) sequences formed connected, central clusters (Fig. 15), additionally including a cluster of FtsZ-like euryarchaeal proteins (CetZ; green). All other sequences are located at the periphery of the map and are not connected to these clusters. In a situation in which possible evolutionary connections have not yet been analyzed (as they have been in the FtsZ/Tubulin superfamily), this cluster map could have formed the basis for computational and experimental investigations on the assumption of homology.

COMMENTARY

Background Information

Two proteins are said to be homologous if they descended from a common ancestor. Generally, homologous proteins have a similar structure and, depending on the degree of divergence, similar functions, cellular localization, or ligands. Since homology of proteins offers a rich source of functional and structural information, the inference of homology has become an essential tool in molecular biology research and underpins the use of model organisms to study biological processes. The divergent evolution of proteins from hypothetical ancestral forms is generally inferred from the similarity of modern representatives. These comparisons are usually made with sequence data because sequence space is essentially infinite and convergence by chance is, therefore, improbable. In contrast, the number of folded conformations available to the polypeptide chain is limited. Hence, unrelated proteins tend to converge on similar structural solutions, especially at the subdomain level. Also, sequence data are easier to obtain than structural data and, thus, more plentiful by orders of magnitude. Over the years, many different sequence comparison methods have been developed. They achieve different levels of sensitivity, depending on the amount of information they incorporate. Methods that compare individual protein sequences, such as BLAST, are the least sensitive, as they use only the information from the pairwise comparison of two sequences, scored by a global substitution matrix. An additional level of sensitivity is achieved by methods that compare sequence profiles to sequences, such as the iterated version of BLAST, PSI-BLAST. Profiles record the frequencies of the 20 amino acids for each column of an MSA and therefore include family-specific information for the query sequence. Profile-to-profile comparison methods, such as COMPASS (Sadreyev, Tang, Kim, & Grishin, 2009), provide an additional improvement by using family-specific information for both sequences being compared. Incorporation of position-specific insertion and deletion frequencies into profiles yields profile hidden Markov models (HMMs). Methods based on HMM-to-HMM comparison, such as HHpred, are currently our most sensitive tools in the detection of sequence similarity.

Understanding Results

The protocols presented here are meant to allow non-expert users to identify distant homologs to their protein of interest and evaluate potential structural similarities. Since the inference of homology becomes progressively more difficult with decreasing pairwise sequence identity, we focus here on such difficult, divergent protein pairs. The region between 20% and 35% pairwise sequence identity has been named the ‘twilight zone’, and the region below it the ‘midnight zone’ (Rost, 1999); for almost all proteins, most homologs are in these zones. Hence, progress in sequence search tools has been mainly measured by their ability to substantiate homology far into the midnight zone. HHpred, the main tool discussed here, is generally acknowledged as the currently best-performing tool for sequence comparisons. Nevertheless, there are instances where it gives scores indicative of homology to proteins not actually related to the query. It is therefore important to evaluate the plausibility of its results based on a few simple guidelines.

- **Check probability and E-value:** The probability value reported by HHpred for a match to be a true positive is the most important criterion to infer if a match is homologous to the query or is just a high-scoring chance hit. When it is greater than 95%, evolutionary relatedness is highly likely. Typically, one should give a match serious consideration if it has a probability value >50%, or it has a probability value >30% and is among the top three hits. The E-value is an alternative measure of statistical significance. It is the number of chance hits with a score better than the one for the given match that is expected to be found in the target database. The lower the E-value, the more significant the match is. Unlike the true-positive probability, the HHpred E-value does not take secondary structure similarity into account. Thus, it is a less sensitive measure than the probability. Consequently, even when the E-value is ~1, matches can be significant by the probability criterion.

• **Check secondary structure similarity:** If the secondary structure of the query and match is substantially different, the match is probably a false positive.

• **Check relationships among top hits:** If several of the top matches are homologous to each other, for instance, when they are members of the same SCOPe superfamily or ECOD homology level, then their likelihood of being homologous to the query is very high.

• **Check if homology is biologically suggestive:** Does the database hit have a function you would also expect for your query? Does it come from an organism that is likely to contain a homolog of your query protein?

• **Check for possible conserved motifs:** Most homologous pairs of proteins will have at least one (semi-)conserved motif in common. You can identify such putative (semi-)conserved motifs by inspecting HHpred alignments for clusters of three or more well-matching columns (marked with a '|' sign in the row between the query and template consensus sequences) and also by matching consensus sequences. Some false positive matches may have high scores due to possessing an amino acid composition similar to that of the query. In such cases, the alignments tend to be long and lack conserved motifs. You could also scan the alignments for motifs known to be involved in enzymatic function or binding of ligands, such as the GTP-binding motif discussed in this report.

• **Check query and template alignments:** A corrupted query or template alignment is the main source of high-scoring false positives. The two most common sources of corruption in an alignment are (1) non-homologous sequences, especially repetitive or low-complexity sequences in the alignment, and (2) non-homologous fragments at the ends of the aligned database sequences. Inspect the query and template MSAs for the presence of spurious sequences. In fact, the HHpred server displays an alert message when coiled-coil, transmembrane, or low-complexity segments are detected in the query.

• **Check if you can reproduce the results with other parameters:** For instance, if you expect the query to be globally homologous to the putative homolog, you could re-run the search using the global alignment mode instead of the local one. You could turn off secondary structure scoring if you suspect that the match between the query and template was scored highly because of a chance similarity of their PSIPRED-predicted or DSSP-determined secondary structures. You can also run the query over other databases to check if similar matches are returned.

Acknowledgments

We would like to thank Andre Noll and Johannes Wörner for their contributions to the development of the Toolkit. This work was supported by institutional funds of the Max Planck Society. Open access funding enabled and organized by Projekt DEAL.

Author Contributions

Felix Gabler: Software; writing-review & editing. **Seung-Zin Nam:** Software. **Sebastian Till:** Software. **Milot Mirdita:** Software. **Martin Steinegger:** Software. **Johannes Söeding:** Software. **Andrei Lupas:** Funding acquisition; writing-review & editing. **Vikram Alva:** Conceptualization; project administration; software; supervision; visualization; writing-original draft; writing-review & editing.

Literature Cited

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. doi: 10.1093/nar/25.17.3389.

Alva, V., Nam, S. Z., Soding, J., & Lupas, A. N. (2016). The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Research*, *44*(W1), W410–415. doi: 10.1093/nar/gkw348.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. doi: 10.1093/nar/28.1.235.

Biegert, A., Mayer, C., Remmert, M., Soding, J., & Lupas, A. N. (2006). The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic Acids Research*, *34*(Web Server issue), W335–339. doi: 10.1093/nar/gkl217.

Cheng, H., Schaeffer, R. D., Liao, Y., Kinch, L. N., Pei, J., Shi, S., … Grishin, N. V. (2014). ECOD: An evolutionary classification of protein domains. *PloS Computational Biology*, *10*(12), e1003926. doi: 10.1371/journal.pcbi.1003926.

Coggill, P., Finn, R. D., & Bateman, A. (2008). Identifying protein domains with the Pfam database. *Current Protocols in Bioinformatics*, *23*, 2.5.1–2.5.17. doi: 10.1002/0471250953.bi0205s23.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., … Finn, R. D. (2019). The Pfam protein families database in 2019.

Nucleic Acids Research, *47*(D1), D427–D432. doi: 10.1093/nar/gky995.

Erickson, H. P. (1998). Atomic structures of tubulin and FtsZ. *Trends in Cell Biology*, *8*(4), 133–137. doi: 10.1016/s0962-8924(98)01237-9.

Fox, N. K., Brenner, S. E., & Chandonia, J. M. (2014). SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, *42*(Database issue), D304–D309. doi: 10.1093/nar/gkt1240.

Frickey, T., & Lupas, A. (2004). CLANS: A Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, *20*(18), 3702–3704. doi: 10.1093/bioinformatics/bth444.

Gibney, G., & Baxevanis, A. D. (2011). Searching NCBI databases using Entrez. *Current Protocols in Bioinformatics*, *34*, 1.3.1–1.3.25. doi: 10.1002/0471250953.bi0103s34.

Gruber, M., Soding, J., & Lupas, A. N. (2006). Comparative analysis of coiled-coil prediction methods. *Journal of Structural Biology*, *155*(2), 140–145. doi: 10.1016/j.jsb.2006.03.009.

Imachi, H., Nobu, M. K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., … Takai, K. (2020). Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature*, *577*(7791), 519–525. doi: 10.1038/s41586-019-1916-6.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, *292*(2), 195–202. doi: 10.1006/jmbi.1999.3091.

Joosten, R. P., te Beek, T. A., Krieger, E., Hekkelman, M. L., Hooft, R. W., Schneider, R., … Vriend, G. (2011). A series of PDB related databases for everyday needs. *Nucleic Acids Research*, *39*(Database issue), D411–D419. doi: 10.1093/nar/gkq1105.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–2637. doi: 10.1002/bip.360221211.

Karpenahalli, M. R., Lupas, A. N., & Soding, J. (2007). TPRpred: A tool for prediction of TPR-, PPR- and SEL1-like repeats from protein sequences. *BMC Bioinformatics*, *8*, 2. doi: 10.1186/1471-2105-8-2.

Ladunga, I. (2017). Finding homologs in amino acid sequences using network BLAST searches. *Current Protocols in Bioinformatics*, *59*, 3 4 1–3 4 24. doi: 10.1002/cpbi.34.

Letunic, I., & Bork, P. (2018). 20 years of the SMART protein domain annotation resource. *Nucleic Acids Research*, *46*(D1), D493–D496. doi: 10.1093/nar/gkx922.

Lowe, J., & Amos, L. A. (1998). Crystal structure of the bacterial cell-division protein FtsZ. *Nature*, *391*(6663), 203–206. doi: 10.1038/34472.

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., … Marchler-Bauer, A. (2020). CDD/SPARCLE: The conserved domain database in 2020. *Nucleic Acids Research*, *48*(D1), D265–D268. doi: 10.1093/nar/gkz991.

Madeira, F., Madhusoodanan, N., Lee, J., Tivey, A. R. N., & Lopez, R. (2019). Using EMBL-EBI Services via Web Interface and Programmatically via Web Services. *Current Protocols in Bioinformatics*, *66*(1), e74. doi: 10.1002/cpbi.74.

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., … Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, *47*(W1), W636–W641. doi: 10.1093/nar/gkz268.

Margolin, W. (2005). FtsZ and the division of prokaryotic cells and organelles. *Nature Reviews Molecular Cell Biology*, *6*(11), 862–871. doi: 10.1038/nrm1745.

Mirdita, M., Steinegger, M., & Soding, J. (2019). MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics*, *35*(16), 2856–2858. doi: 10.1093/bioinformatics/bty1057.

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Soding, J., & Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research*, *45*(D1), D170–D176. doi: 10.1093/nar/gkw1081.

NCBI Resource Coordinators. (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *46*(D1), D8–D13. doi: 10.1093/nar/gkx1095.

Nogales, E., Downing, K. H., Amos, L. A., & Lowe, J. (1998). Tubulin and FtsZ form a distinct family of GTPases. *Nature Structural Biology*, *5*(6), 451–458. doi: 10.1038/nsb0698-451.

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, *302*(1), 205–217. doi: 10.1006/jmbi.2000.4042.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. doi: 10.1093/nar/gkv1189.

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., & Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Research*, *46*(W1), W200–W204. doi: 10.1093/nar/gky448.

Prakash, A., Jeffryes, M., Bateman, A., & Finn, R. D. (2017). The HMMER Web Server for Protein Sequence Similarity Search. *Current Protocols in Bioinformatics*, *60*, 3 15 11–13 15 23. doi: 10.1002/cpbi.40.

Pundir, S., Martin, M. J., O'Donovan, C., & The UniProt Consortium. (2016). UniProt tools. *Current Protocols in Bioinformatics*, *53*, 1.29.21–21.29.15. doi: 10.1002/0471250953.bi0129s53.

Remmert, M., Biegert, A., Hauser, A., & Soding, J. (2011). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, *9*(2), 173–175. doi: 10.1038/nmeth.1818.

Rose, A. S., Bradley, A. R., Valasatava, Y., Duarte, J. M., Prlic, A., & Rose, P. W. (2018). NGL viewer: Web-based molecular graphics for large complexes. *Bioinformatics*, *34*(21), 3755–3758. doi: 10.1093/bioinformatics/bty419.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering*, *12*(2), 85–94. doi: 10.1093/protein/12.2.85.

Sadreyev, R. I., Tang, M., Kim, B. H., & Grishin, N. V. (2009). COMPASS server for homology detection: Improved statistical accuracy, speed and functionality. *Nucleic Acids Research*, *37*(Web Server issue), W90–W94. doi: 10.1093/nar/gkp360.

Sali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, *234*(3), 779–815. doi: 10.1006/jmbi.1993.1626.

Schaeffer, R. D., Liao, Y., & Grishin, N. V. (2018). Searching ECOD for homologous domains by sequence and structure. *Current Protocols in Bioinformatics*, *61*(1), e45. doi: 10.1002/cpbi.45.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., … Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*, 539. doi: 10.1038/msb.2011.75.

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, *21*(7), 951–960. doi: 10.1093/bioinformatics/bti125.

Steinegger, M., Meier, M., Mirdita, M., Vohringer, H., Haunsberger, S. J., & Soding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, *20*(1), 473. doi: 10.1186/s12859-019-3019-7.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., & UniProt, C. (2015). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, *31*(6), 926–932. doi: 10.1093/bioinformatics/btu739.

Swiss Institute of Bioinformatics Members. (2016). The SIB Swiss Institute of Bioinformatics' resources: Focus on curated databases. *Nucleic Acids Research*, *44*(D1), D27–D37. doi: 10.1093/nar/gkv1310.

Szwedziak, P., Wang, Q., Bharat, T. A., Tsim, M., & Lowe, J. (2014). Architecture of the ring formed by the tubulin homologue FtsZ in bacterial cell division. *Elife*, *3*, e04601. doi: 10.7554/eLife.04601.

The UniProt Consortium. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515. doi: 10.1093/nar/gky1049.

Yang, M., Derbyshire, M. K., Yamashita, R. A., & Marchler-Bauer, A. (2020). NCBI's conserved domain database and tools for protein domain analysis. *Current Protocols in Bioinformatics*, *69*(1), e90. doi: 10.1002/cpbi.90.

Zimmermann, L., Stephens, A., Nam, S. Z., Rau, D., Kubler, J., Lozajic, M., … Alva, V. (2018). A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *Journal of Molecular Biology*, *430*(15), 2237–2243. doi: 10.1016/j.jmb.2017.12.007.