Fully Automated Ancestral Sequence Reconstruction using FireProt^{ASR}

Rayyan Tariq Khan,[1,2] Milos Musil,[1,2,3] Jan Stourac,[1,2] Jiri Damborsky,[1,2,4] and David Bednar[1,2,4]

[1]Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic
[2]International Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic
[3]Department of Information Systems, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic
[4]Corresponding authors: *jiri@chemi.muni.cz*; *davidbednar1208@gmail.com*

Protein evolution and protein engineering techniques are of great interest in basic science and industrial applications such as pharmacology, medicine, or biotechnology. Ancestral sequence reconstruction (ASR) is a powerful technique for probing evolutionary relationships and engineering robust proteins with good thermostability and broad substrate specificity. The following protocol describes the setting up and execution of an automated FireProt^{ASR} workflow using a dedicated web site. The service allows for inference of ancestral proteins automatically, from a single protein sequence. Once a protein sequence is submitted, the server will build a dataset of homology sequences, perform a multiple sequence alignment (MSA), build a phylogenetic tree, and reconstruct ancestral nodes. The protocol is also highly flexible and allows for multiple forms of input, advanced settings, and the ability to start jobs from: (i) a single sequence, (ii) a set of homologous sequences, (iii) an MSA, and (iv) a phylogenetic tree. This approach automates all necessary steps and offers a way for novices with limited exposure to ASR techniques to improve the properties of a protein of interest. The technique can even be used to introduce catalytic promiscuity into an enzyme. A web server for accessing the fully automated workflow is freely accessible at *https://loschmidt.chemi.muni.cz/fireprotasr/*.

Basic Protocol: ASR using the Web Server FireProt^{ASR}

Keywords: ancestral sequence reconstruction • automation • protein engineering • protein evolution • thermostability

How to cite this article:
Khan, R. T., Musil, M., Stourac, J., Damborsky, J., & Bednar, D. (2021). Fully automated ancestral sequence reconstruction using FireProt^{ASR}. *Current Protocols*, *1,* e30. doi: 10.1002/cpz1.30

INTRODUCTION

Phylogenetic analysis of homologous protein sequences allows a user to statistically build and probe evolutionary relationships within the dataset by yielding a tree-like chart called a phylogenetic tree (Charleston, 2013; Liberles, 2007). The phylogenetic tree is made up of leaves (extant protein sequences that were present in the dataset) that are connected by binary branches. Similar sequences are placed on the same binary branches. All branching events yield a node (the point before a branch occurs), which corresponds to a common ancestor of the extant sequences. All the branches connect together as they

CURRENT PROTOCOLS
A Wiley Brand
Current Protocols e30, Volume 1
Published in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpz1.30

go back to the root node (a common ancestor of all sequences in the dataset). Ancestral sequence reconstruction (ASR) can be used to calculate the most probable ancestral protein sequences that may have existed at any or all nodes on the phylogenetic tree (Gaucher, 2007). Ancestral proteins generally tend to be more robust, thermostable, and expressible (Babkova et al., 2020; Watanabe, Ohkuri, Yokobori, & Yamagishi, 2006) than the extant proteins, with broader substrate specificity (Babkova, Sebestova, Brezovsky, Chaloupkova, & Damborsky, 2017; Gaucher, Govindarajan, & Ganesh, 2008; Wheeler, Lim, Marqusee, & Harms, 2016) and promiscuity (Chaloupkova et al., 2019); thus they can make excellent candidates for industrial applications. The whole process is tedious, as even the most basic iteration of the process requires the user to manually curate a protein dataset, construct a multiple sequence alignment (MSA), curate the alignment, construct the tree, and calculate sequences of ancestors.

To address these issues, we developed a web server with a simple and intuitive graphical user interface called FireProt^{ASR} (*https://loschmidt.chemi.muni.cz/fireprotasr/*). The

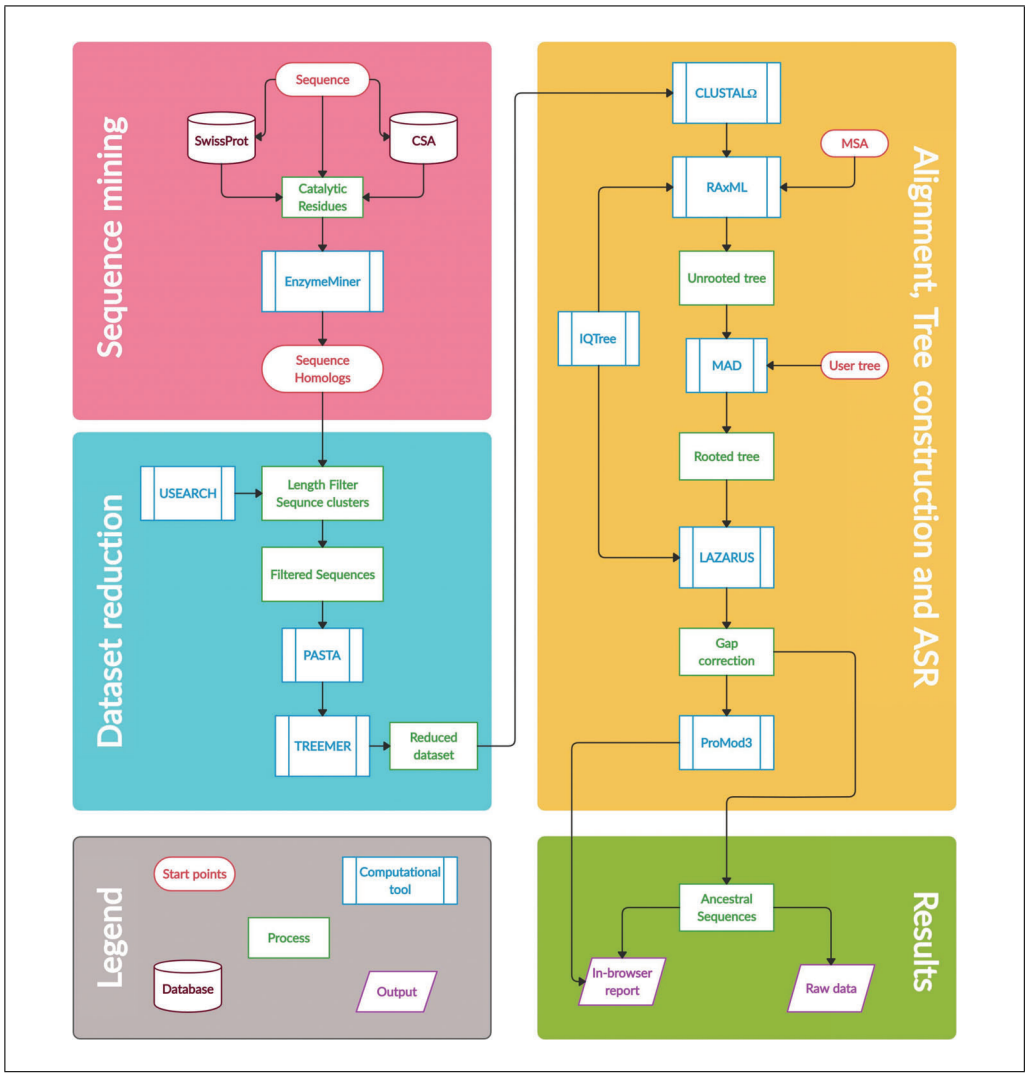

Figure 1 Schematic representation of FireProt^{ASR} workflow. The pipeline is made of several computational tools that allow for automated ancestral sequence reconstruction (ASR) from four different starting points: protein sequence, set of sequences, MSA, and phylogenetic tree. The workflow consists of four steps: "Sequence mining," "Dataset reduction," "Alignment, Tree construction, and ASR," and "Results download." The figure is colored as follows: starting points are in red, process parts are in green, databases are in brown, computational tools are in blue, and outputs are in purple.

web server allows users to easily run entire ASR protocols, without having much expertise in bioinformatics or evolutionary biology, from a single query protein sequence. Even though the workflow can be run automatically with default settings, experienced users can also use advanced settings for greater control over the procedure. Moreover, FireProt^{ASR} provides alternative entry points to the ASR pipeline; a user who has a curated protein dataset can simply start the ASR job from the protein dataset, or likewise for MSA and phylogenetic tree (Fig. 1).

The Basic Protocol describes the simplest way to perform a full phylogenetic analysis and ASR via the FireProt^{ASR} web server, describing all the mandatory steps and advanced settings options.

ANCESTRAL SEQUENCE RECONSTRUCTION USING THE WEB SERVER FireProt^{ASR}

BASIC PROTOCOL

The FireProt^{ASR} web server runs on a pipeline composed of multiple computational tools. While it is possible to use all these computational tools separately to perform phylogenetic analysis and ASR, our web server allows users to bypass the need to install and use these tools separately. The user has to follow a few simple steps, and all the results can then either be viewed in the browser or downloaded as raw files. This protocol guides a user through the few steps required to run and evaluate a job on the FireProt^{ASR} web server using haloalkane dehalogenase (DhaA). For a guided example, see Video 1: *https://loschmidt.chemi.muni.cz/data/fireprotasr/tutorial*. It also elaborates on the alternative starting points for the calculation. Finally, it explains how to explore and interpret the results in the browser, or downloaded as raw files. The full protocol takes about one day to complete; the first step of the calculation takes about one-third of the total calculation time.

Materials

Hardware

Any computer with internet access. There are no special requirements for the memory, space, or performance of the computer. However, low-performance computers may take more time when visualizing the results or editing selected ancestral sequences.

Software

An up-to-date web browser such as Google Chrome (*https://www.google.com/chrome/*), Mozilla Firefox (*https://www.mozilla.com/*), or Apple Safari (*https://www.apple.com/safari/*).
A text reader software application such as Notepad

Access FireProt^{ASR} tool

1. Navigate to the FireProt^{ASR} homepage (*https://loschmidt.chemi.muni.cz/fireprotasr/*). Copy and paste the haloalkane dehalogenase (DhaA) sequence (Štěpánková et al., 2013) in FASTA format (available at *https://www.rcsb.org/fasta/entry/4E46/display*) into the sequence field. Click on the "Validate" button to validate the sequence (Fig. 2). Click on the "Next" button to go to the "Settings of calculation" page.

   The default setting on the home page allows a user to start from a single protein sequence. The source field allows a user to select whether to type a sequence in FASTA format or plain text, or to upload a FASTA-format protein sequence file. Positive validation of a sequence using the "Validate" button will result in a checkmark next to the button. The entire Job Information field (Job title and Email) is optional, but it is highly recommended, as the calculation can take up to a day. It is possible to start from the three alternative start points of the workflow by clicking on "USER DATA" on the home screen

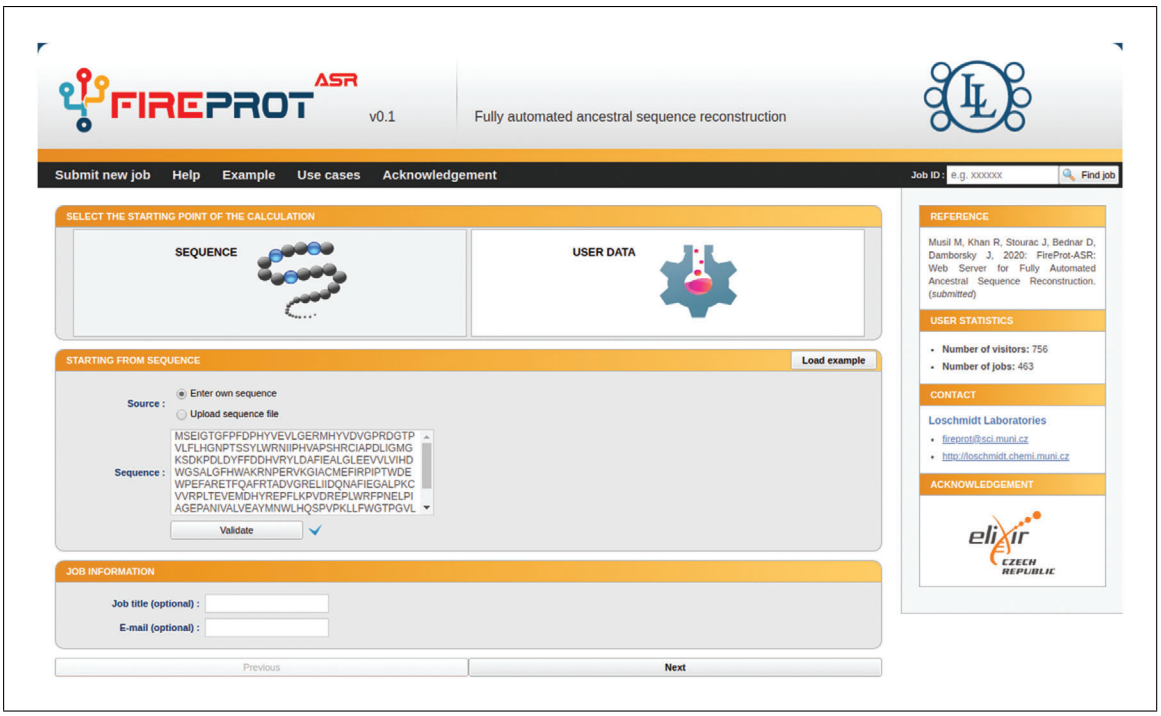

Figure 2 FireProt^{ASR} web server's homepage. This page allows users to either start from a single "SEQUENCE" or "USER DATA." The "SEQUENCE" should be provided in FASTA format. The "USER DATA" can be multiple homologous sequences, a custom-made MSA or a phylogenetic tree.

(Fig. 3). To start from a set of sequences selected by the user, select "Set of initial data" as a "Source," and upload a FASTA format file. To start from an MSA, select "Multiple-sequence alignment" as a source, and upload alignment in FASTA format. If you already have a phylogenetic tree, then in addition to uploading the MSA, also upload or paste a tree in NEWICK format into the tree field and mark if the tree is rooted or unrooted. The "USER DATA" option is meant for more experienced users who are able to manually curate the data according to their knowledge of the studied system. In any of the latter three cases, it is important to pick one protein sequence as a "Query" by naming it as such or choosing the query sequence in the pop-up window.

2. The "Settings of calculation" page (Fig. 4) shows a list of essential residues aggregated from the SwissProt database (Boutet et al., 2016) and/or the Mechanism and Catalytic Site Atlas M-CSA database (Ribeiro et al., 2018). Essential residue positions 41, 106, 107, 130, and 272 are pre-selected in the example case. These are catalytic residues of this enzyme. Essential residues help in mining similar proteins for the homolog dataset. See Critical Parameters and Troubleshooting for tips on selecting essential residues for other proteins.

   Catalytic and ligand or co-factor binding residues are the best residues to pick in this step, as they will help FireProt^{ASR} mine catalytically similar enzyme sequences. If the query protein is not an enzyme, it is recommended to select a few family-specific residues that will help FireProt^{ASR} mine similar enzymes. It is also possible to run a sequence without specifying anything in the "ESSENTIAL RESIDUES" field. In this case, a less specific homolog dataset will be mined. It is important to note that selecting too many residues may result in the job failing, as discussed in Critical Parameters and Troubleshooting.

3. Select the "Two-steps: manual elimination of the sequences from a phylogenetic tree enabled (experts)" option from the field labeled "Number of calculation steps" and click on the "Submit job" button. This will lead to a job status page. This step usually takes less than 6 hours to complete.

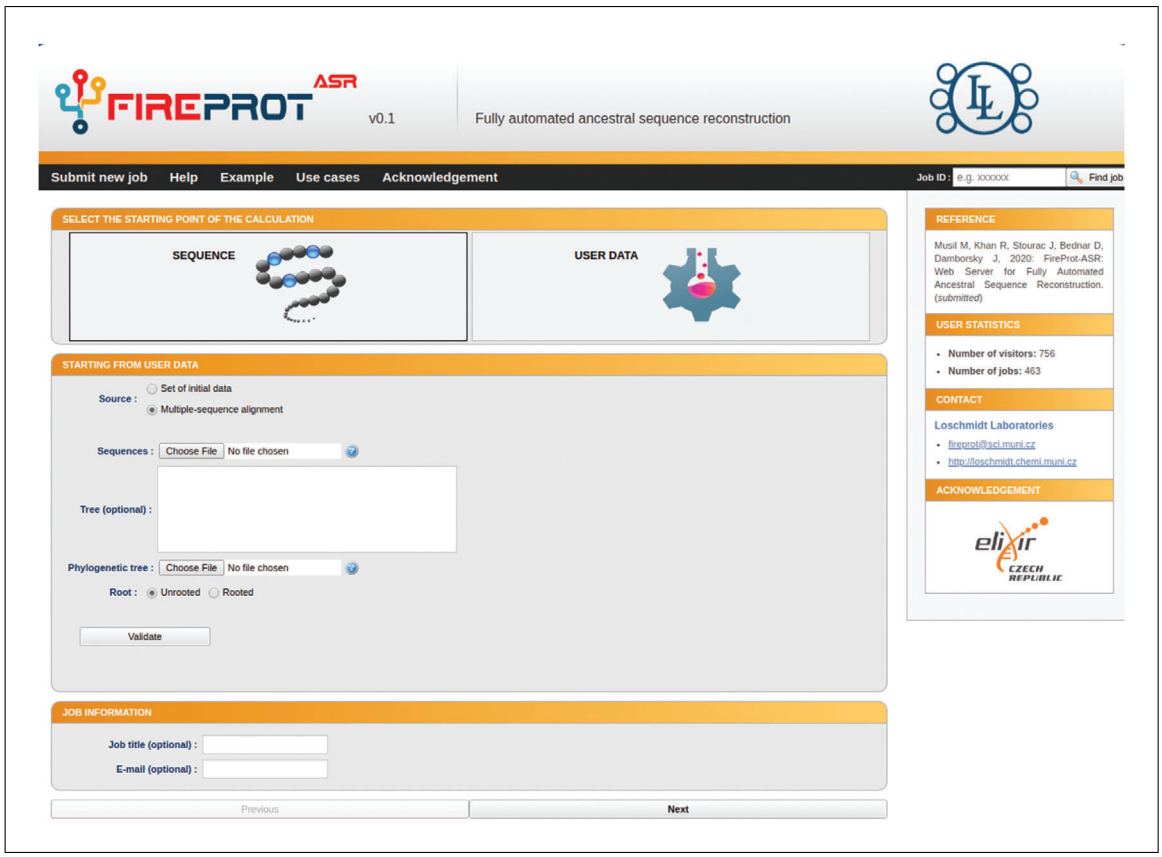

Figure 3 FireProt^{ASR} web server alternative starting points. The "USER DATA" option allows users to either start by uploading a set of sequences, an MSA, or a phylogenetic tree.

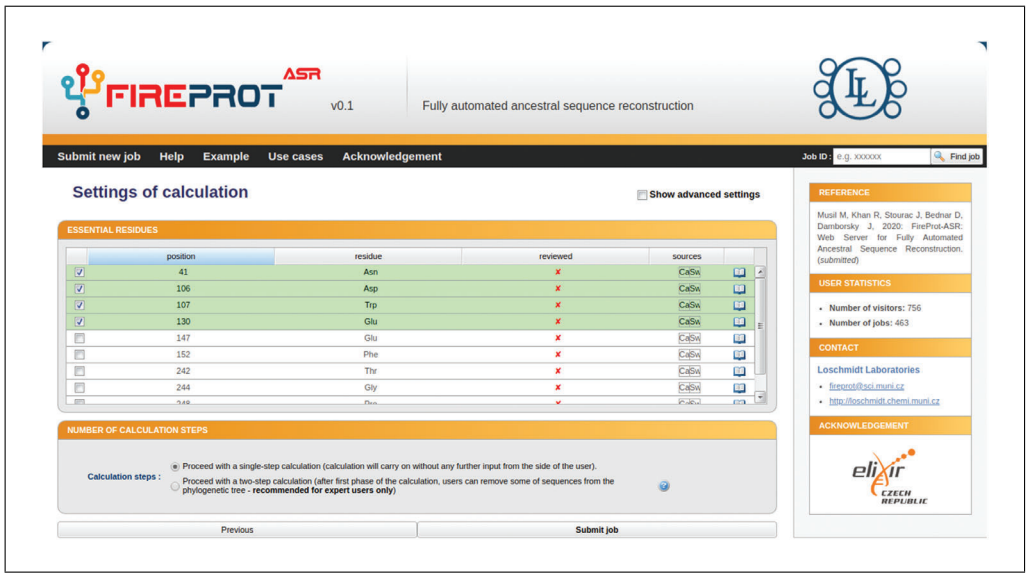

Figure 4 Settings of calculations page. The page allows users to select mined essential residues, define their own essential residues, and choose the number of calculation steps. The page also allows users to access the advanced settings.

Picking the "One-step: fully automated calculation does not require any further input from the user (default)" option will help you skip directly to step 6 in this protocol. This option is recommended for inexperienced users.

4. Access the Tree Pruning page once the first step of the calculation is complete (Fig. 5). Manipulate the visualized phylogenetic tree by clicking and dragging and/or zooming in or out using the mouse's scroll wheel. Click the "Submit" button to

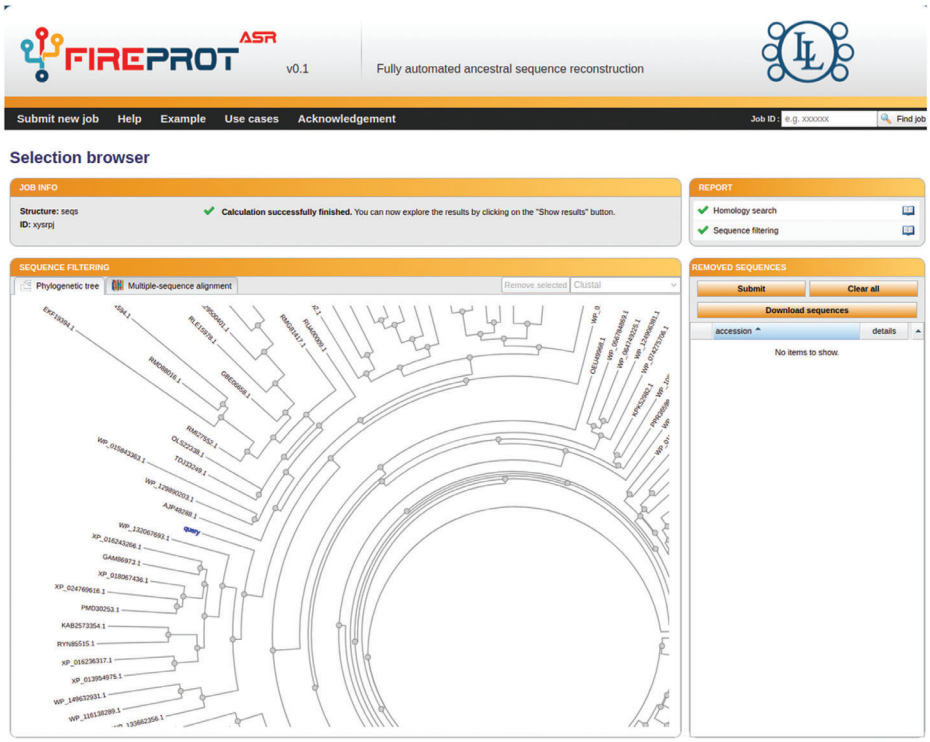

Figure 5 Selection browser page. The "Phylogenetic tree" tab in the "SEQUENCE FILTERING" section shows a tree constructed from the mined dataset. It is possible at this stage to prune the tree, before further calculations. The "query" sequence is noted in blue color.

progress to the second step of the calculation. This will lead to the "Results browser" page, with a "JOB INFO" tab; this tab will display the current status of the calculation. The second step of the calculation will generally take less than 18 hours to complete.

The phylogenetic tree shows the calculated evolutionary relationships among the filtered ≤150 protein sequences. Any subtree in the full phylogenetic tree that carries protein sequences that you do not want to keep for the next step of the calculation can be removed by clicking on the representative node and then clicking on "Remove whole subtree." Similarly, any single protein sequence that you want to remove from the next step of the calculation can be removed either by clicking on it in the "Phylogenetic tree" tab or by clicking on the check box next to the sequence in the "Multiple-sequence alignment" tab (Fig. 6), and then clicking on the "Remove selected" button. Any removed sequences can be restored back to the tree using the Removed Sequences field.

5. Once the job info page states "Calculation successfully finished" (Fig. 7), click on the "Show results" button to access the "Results browser" page.

   Alternatively, it is possible to click on the "Download archive" button and skip to step 17 of this protocol. You will also skip the on-site analysis available in the "Results browser" page.

6. The results browser (Fig. 8) page is composed of four panels: Viewer, Visualization settings, Mutations, and Sequence designer (from top to bottom, left to right). Use the Viewer panel to explore a 3D model of the query protein haloalkane dehalogenase (DhaA). You can use the Visualization settings panel to change the visualization style and/or visualization quality, as required.

7. Use the respective tabs in the Mutations panel to visualize the Phylogenetic tree as well as the MSA of the full protein dataset (including ancestral proteins).

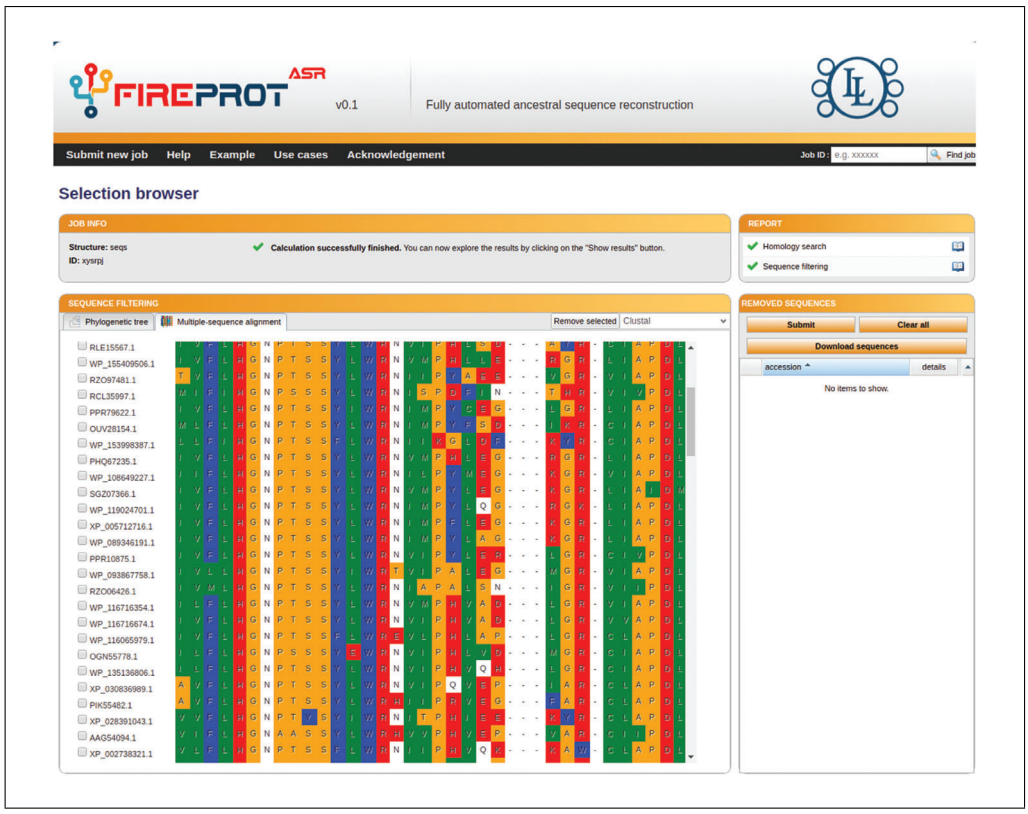

Figure 6 Multiple sequence alignment tab. This tab allows users to select and remove sequences from the MSA before the next step of the calculation.

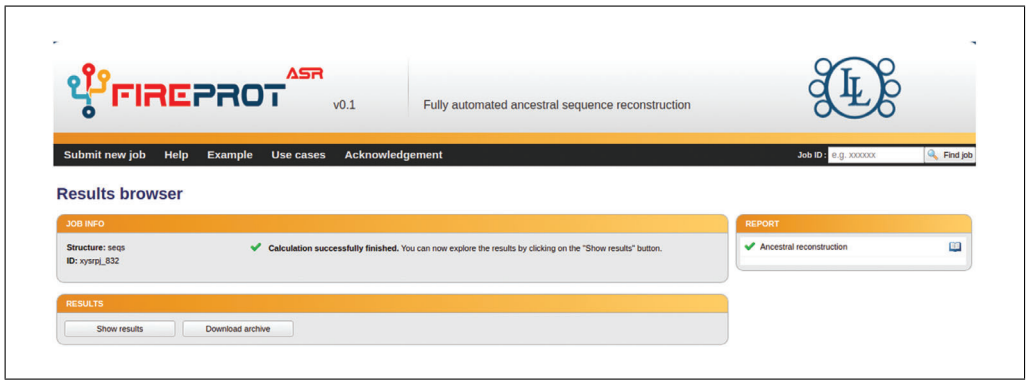

Figure 7 Job completion status page. This page allows users to view the results in the browser and download the complete set of results.

Visualizing differences between ancestral proteins and the query protein

FireProt^{ASR} allows users to visualize the differences between the query protein and a chosen ancestor. The substitutions can be imposed on to the 3D model of the Query protein. This can aid in the selection of ancestral proteins for further analysis.

8. Select the Phylogenetic tree tab in the Mutations panel and click on the most central ancestral node (Ancestral node 151) in the phylogenetic tree, also known as the root. Then, click on the "Show this ancestor" option in the drop-down menu. By default, all substitutions in the chosen ancestral protein sequence, when compared to the query protein sequence, will be highlighted in blue color on the 3D structure, in the "Viewer" panel.

Modifying an ancestral protein's sequence

In the case where the user needs to modify the inferred ancestral sequence according to prior knowledge (previous mutational experiments, different amino acids based on

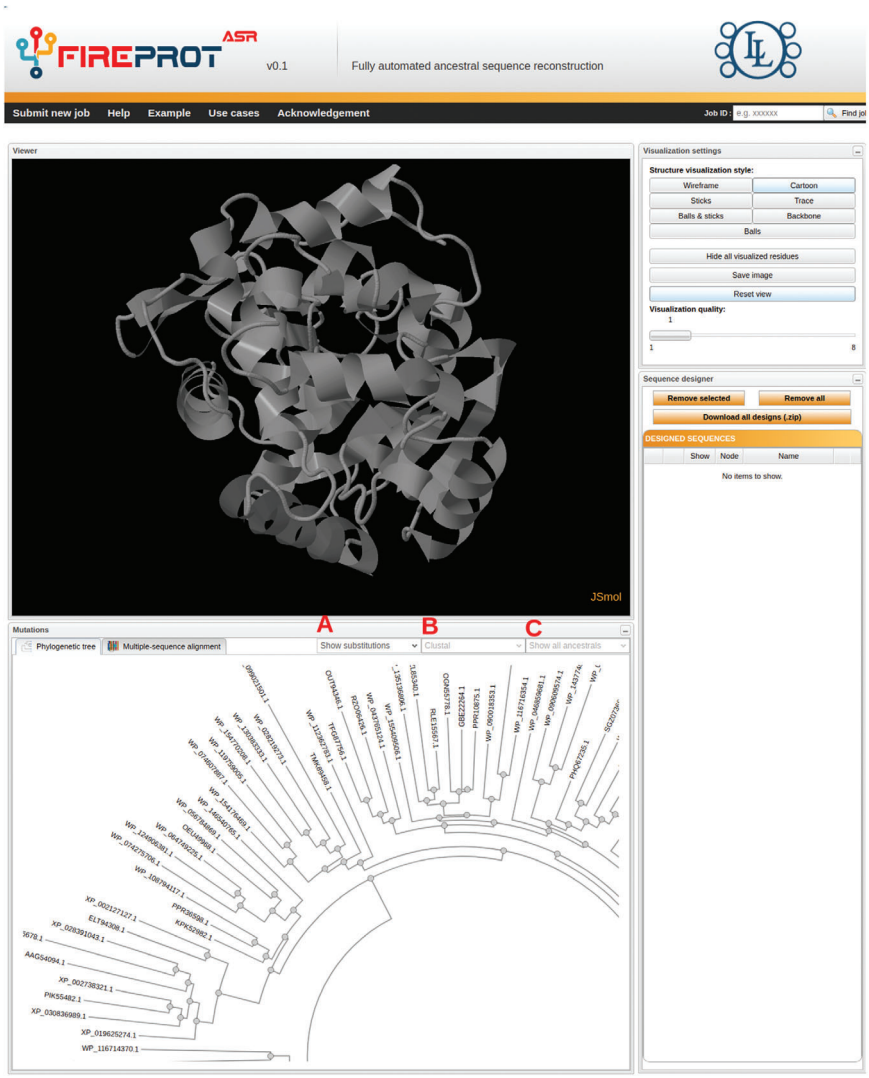

Figure 8 Results browser page. The four panels include "Viewer" for visualizing the query protein structure, "Visualization settings" for controlling the 3D view of the protein structure, "Mutations" for accessing the phylogenetic tree and the MSA, and "Sequence designer" for managing ancestral sequences. The drop-down menu labeled A is for projecting differences between the query protein sequence and a selected ancestral sequence, which will then be projected onto the model in the "Viewer" panel. The drop-down menus labeled B can be used to change the color scheme in the "Multiple-sequence alignment" tab. The drop-down menu labeled C can be used to switch between an alignment of either all ancestral sequences and selected ancestral sequences in the MSA of filtered sequences.

posterior probabilities of close homolog sequences, etc.), the following three steps will help with the manual curation.

9. Select the Phylogenetic tree tab in the Mutations panel and click on the most central ancestral node (Ancestral node 151) in the phylogenetic tree. Then, click on the "Show ancestral sequence" option in the drop-down menu to access the "Sequence information" window.

10. Horizontally scroll and locate the residue that you would like to change. Click on the posterior probability distribution bar of that residue to access a detailed window for the chosen residue.

    Residue positions that are in dull colors are recommended gaps.

11. Click on the bar in front of your chosen amino acid's three letter code, and then click the "Submit" button in this window. Once all residue positions have been modified in a similar manner, click on "Store sequence," give your modified ancestral design a name, and click on "OK." You will now see your sequence in the "Sequence designer" panel.

Viewing the MSA

12. Select the "Multiple-sequence alignment" tab in the Mutations panel to view the MSA of the entire sequence dataset. This alignment also contains all the ancestral protein sequences by default. To show an alignment of all the extant protein sequences with only the ancestral protein sequence that you selected in step 8, click on the drop-down menu labeled C in Figure 8 and click on "Show only selected" option.

13. You can change the color scheme of the MSA by clicking on the drop-down menu labeled B in Figure 8 and selecting Clustal (Sievers & Higgins, 2018), Taylor (Taylor, 1997), or Zappo (Procter et al., 2010) coloring schemes.

Downloading your results

14. Use the "Sequence designer" panel to remove or view the ancestral sequences you designed in steps 10-11. Click the "Download all designs (.zip)" button to download a compressed folder with all the results.

    It is possible to update designed sequences in the "Sequence designer" panel by clicking on the book icon and following the same procedure as in steps 11-13 of this protocol.

15. Unzip the compressed results file and navigate to the `ancestrals` folder (file path `results/fireprot/results/ancestrals`) to access all your ancestral sequences in a .fas format.

16. Navigate to the `user` folder (file path `results/fireprot/results/user`) to access all user-designed ancestral sequences in `.fas` format.

17. The results folder (file path `results/fireprot/results`) also contains the following files:

    i. `structure.pdb` presenting the same 3D protein structure that was visualized in step 6. Any program that supports the `.pdb` format can be used to explore this file.
    ii. `ancestrals.csv`, which contains an overview of all the results, including posterior probabilities, in an Excel sheet.
    iii. `bigMSA.fasta`, which is an MSA of all the extant proteins sequences and all the reconstructed ancestral protein sequences.
    iv. `msa.fasta`, which is an MSA of only the extant protein sequences that were used to reconstruct ancestral proteins.
    v. `tree.tre`, which contains the phylogenetic tree in Newick format ("Newick's 8:45" Tree Format Standard, Olsen, 1990). The file can be opened using a text editor, and it contains two versions of the tree, a tree with branch lengths, and the same tree as a cladogram, which includes the node numbers that are the same as the name of the ancestral sequences. Thus, to find the location of an ancestral node in a tree, use the node number of the ancestral sequence and search for it on the cladogram. Phylogenetic trees can be visualized using a tree visualizer, such as the Interactive Tree Of Life (ITOL; Letunic & Bork, 2019).

    The results generated from this protocol can be matched with precalculated results available at https://loschmidt.chemi.muni.cz/fireprotasr/?action=calculation&job=xysrpj_832&.

COMMENTARY

Background Information

Reconstructing ancestral proteins from a single query point is a multi-step process that requires significant experience in bioinformatics. There are several online servers that automate some parts of the process, but none of them are able to go through the whole process without substantial contribution from the user. Ancestors (Diallo, Makarenkov, & Blanchette, 2010), GRASP (Foley et al., 2019), and FASTML (Ashkenazy et al., 2012) are online web servers that allow users to generate ancestral sequences from an initial MSA and a phylogenetic tree (which are required as input).

The PhyloBot (Hanson-Smith & Johnson, 2016) web server makes the process slightly easier by only requiring a FASTA-format collection of orthologous protein sequences. Then, it performs MSA, phylogenetic tree building, and eventually ASR. The iPlant Discovery Environment (Matasci & McKay, 2013) is a web portal that allows users to access multiple bioinformatics tools online; thus, it is possible to conduct individual steps of the ASR process through the portal, but it is not automated and requires considerable skills and knowledge.

While the aforementioned online web servers are helpful, and they allow advanced users to generate ancestral sequences, they would still present a significant challenge to less skilled users who would have to generate an initial MSA and phylogenetic tree as inputs. The FireProt^{ASR} web server, presented in this protocol, is the only web tool that allows for a *fully* automatic selection of sequence homologs, construction of the MSA and rooted phylogenetic tree, inference of ancestral sequences, and gap reconstruction using only a single protein sequence as a starting point for the calculation.

Critical Parameters and Troubleshooting

• **Sequence validation failure**: This may be due to the fact that the uploaded or typed-in sequence is not in FASTA or plain text format. A proper format of the sequence will be ensured by its validation. Make sure that only one-letter codes for the 20 standard amino acids are used. One-letter codes for the non-standard amino acids, as well as the codes for unknown amino acid (X), and gaps (-), will invalidate the sequence.

• **Selecting protein sequence**: In the case where a set of sequences is used as a starting point for the calculation, proper selection of the query protein sequence is important, as it will determine sequences collected by the EnzymeMiner tool. Selecting a sequence with few homologs across various protein databases may result in the job failing. Although protein sequence selection is highly dependent on the reason for performing ASR, using a protein that is known to have a good number of orthologs in protein databases is important. In the case of smaller families, filters can be adjusted to a less strict setting. In the case where a set of sequences is uploaded, one of the sequences should be named "Query." If no protein sequence is named "Query," a popup window will let you choose one.

• **Selecting essential residues**: Proper selection of essential residues is a key step, as EnzymeMiner uses this information to mine the dataset for similar protein sequences. Preferably, select all essential residues that have well established catalytic properties (this will help filter out proteins that have similar sequences, but different catalytic activities). Selecting residues that are not conserved through the group of studied proteins may lead to overparameterization, which can result in mining of a very small dataset. It may even lead to the job failing due to no protein sequences being found. In such a case, reduce the number of selected residues to only those that are proven to be essential in databases or the literature, and try again.

• **No essential residues located**: In some sequences, no essential residues may be found. This may be due to the fact that there are no essential residues in the SwissProt or M-CSA database. It may also be possible that the query is a non-enzymatic protein. In the former case, reviewing the literature may yield potential essential residues. The SwissProt server API, which FireProt^{ASR} uses to search for essential residues in the SwissProt database, is prone to failure. In the latter case, it may be prudent to restart the job one more time, or visit the SwissProt and M-CSA databases directly and retrieve this information manually.

• **Calculation failure**: If the calculation fails in the second/last step of calculations (step 5 of Basic Protocol), it may be due to the fact that the final number of sequences in the dataset is less than 10. Selecting fewer essential residues on the "Settings of calculations" page could increase the number of sequences, but be careful about incorporating sequences from different families. In this case,

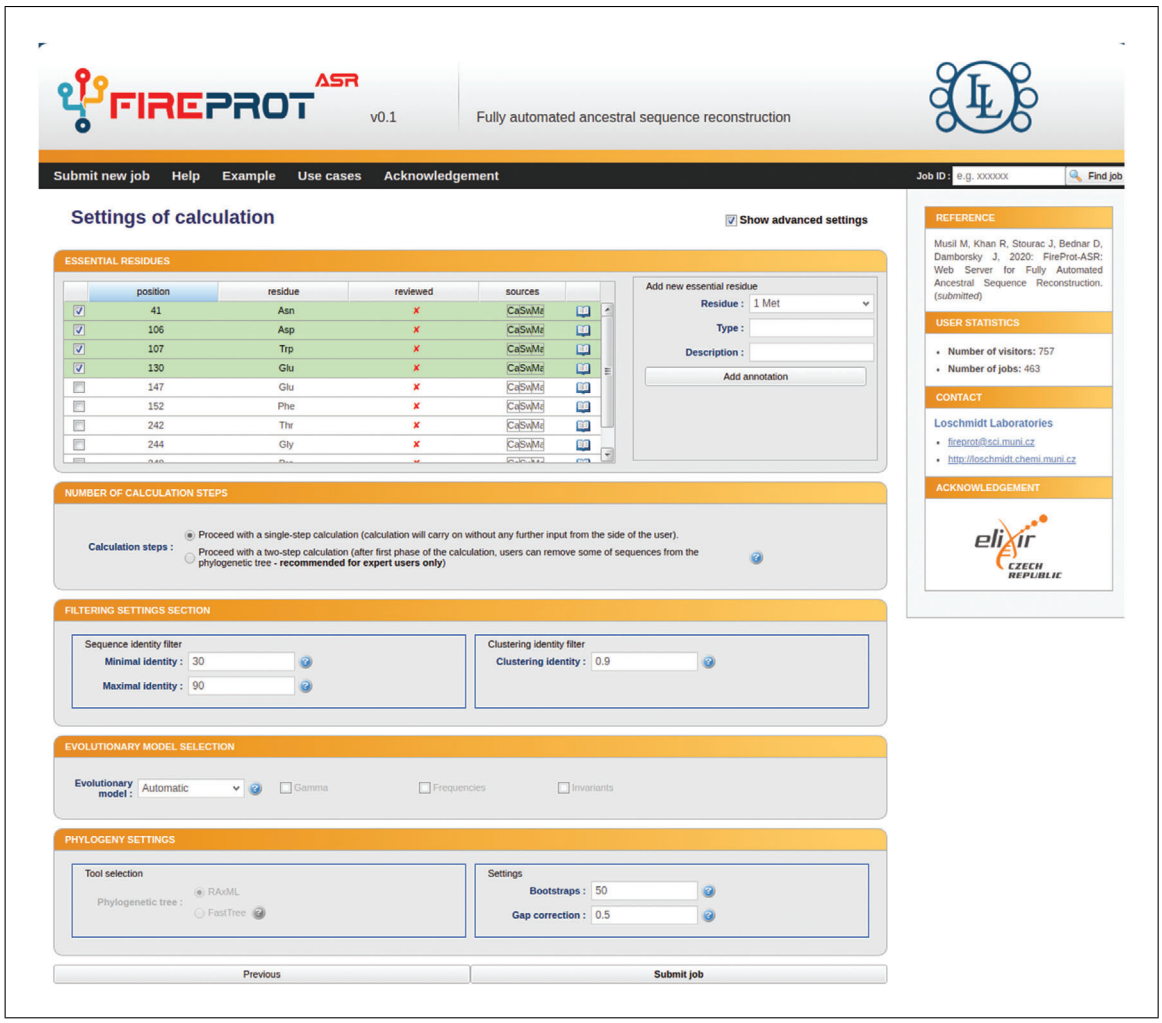

Figure 9 Advanced settings page. This page makes it possible to change the filtering parameters as well as the evolutionary models before the start of calculations.

checking the right position of all catalytic residues is necessary for proper function. Alternatively, providing extra protein sequences by yourself may counteract this issue.

• **3D model construction failure**: ProMod3 (Waterhouse et al., 2018), the tool that the FireProt^{ASR} pipeline uses to construct a 3D model of the query protein for the results viewer page, may sometimes fail. This is usually due to the lack of template protein structures in the PDB database. ASR is sequence-based analysis, and therefore is not dependent on 3D structure, which is used only for visualization purposes. All the ancestral sequences are inferred normally, and the result page is still usable with minor limitations.

Advanced Parameters

Checking the box on the top right corner of the "Settings of calculation" page allows access to the advanced settings (Fig. 9), which are discussed below.

• **Manual addition of an essential residue**: To manually add an essential residue, use the "Add new essential residue" field to select the residue and its position. You can also add a "Type" and "Description" for each residue you want to add. Once finished, click the "Add annotation" button. Once added to the list, check the box next to it in the list.

• **Filtering sequences based on sequence identity**: In the "FILTERING SETTINGS SECTION," the minimal and maximal sequence identity filter allows users to filter out any collected protein sequences in the dataset, with the percent identity to the query protein sequence above or below a set threshold. Increasing sequence identity will result in a less diverse dataset, while decreasing it significantly may result in a "noisy" dataset with homolog sequences that may not only be part of different families, but also from a different structural fold (particularly when a low number or no essential residues are set).

• **Using clustering identity filtering**: The clustering identity filter in the "FILTERING SETTINGS SECTION" determines the sequence similarity at which protein sequences will be clustered together. For example, if the value is set at 0.9, it means that all the protein sequences that share a sequence similarity at or above 90% will be in a single cluster. When FireProt^{ASR} reduces the mined dataset, it picks a single sequence from every cluster, thus ensuring maximal diversity in the final dataset.

• **Evolutionary models and substitution matrices**: The "EVOLUTIONARY MODEL SELECTION" section allows users to select evolutionary models and substitution matrices. These matrices score the probability of an amino acid mutation over time, and hence are important from the perspective of protein evolution. The default setting will let the IQ-TREE tool predict the best-fitting model for a particular job.

• **Bootstrap selection**: The "PHYLOGENY SETTINGS" section allows users to select the number of "Bootstraps" applied during tree construction. A high number of bootstraps may slightly increase the accuracy of the tree construction (by fine-graining the convergence of the final tree shape). However, more bootstraps will increase the time demands of the calculation. The section also allows users to select the "Gap correction" value. A higher value may result in a stricter gap correction, leading to ancestral sequences with fewer gaps.

Understanding Results

A compressed folder containing the output files that match the results of the job performed in the Basic Protocol can be accessed by navigating to *https://loschmidt.chemi.muni.cz/fireprotasr/?action=calculation&job=xysrpj_832&* and clicking on the "Download archive" button. If you click on "Show results," the same results will be opened in the visualization window on the FireProt^{ASR} results page.

FireProt^{ASR} provides three types of output: (i) phylogenetic tree, (ii) MSA, and (iii) ancestral sequences. The phylogenetic tree is a representation of the most probable evolutionary relationships among all the proteins in the dataset. The MSA shows the alignment of the sequences in the dataset. The ancestral protein sequences are the most probable protein sequences that may have existed at their respective ancestral nodes. Selecting the right ancestral sequence for further experimentation is highly dependent on the actual needs and requirements of the experimenter.

Time Considerations

The Basic Protocol takes about one day to finish calculations in the one-step version. The two-step version of the Basic Protocol takes approximately the same time, but requires the user to evaluate the preliminary unrooted tree after several hours and re-start the calculation. The first step of the workflow usually accounts for approximately one-third of the total calculation time.

Acknowledgments

This work was supported by the Czech Ministry of Education (CZ.02.1.01/0.0/0.0/16_026/0008451), the Czech Grant Agency (20-15915Y), the Technology Agency of Czech Republic (TH02010219), Brno University of Technology (FIT-S-20-6293), and the European Commission (814418 and 722610). Computational resources were supplied by the project "e-Infrastruktura CZ" (LM2018131) and ELIXIR (LM2018131).

Author Contributions

Rayyan Tariq Khan: Conceptualization, data curation, methodology, software, visualization, writing-original draft, writing-review & editing. **Milos Musil:** Conceptualization, data curation, formal analysis, methodology, software, visualization, writing-review & editing. **Jan Stourac:** Conceptualization, formal analysis, methodology, software, visualization, writing-review & editing. **Jiri Damborsky:** Conceptualization, funding acquisition, methodology, resources, software, supervision, writing-review & editing. **David Bednar:** Conceptualization, formal analysis, methodology, software, supervision, writing-review & editing.

Literature Cited

Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., & Pupko, T. (2012). FastML: A web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research*, *40*(W1), W580–W584. doi: 10.1093/nar/gks498.

Babkova, P., Dunajova, Z., Chaloupkova, R., Damborsky, J., Bednar, D., & Marek, M. (2020). Structures of hyperstable ancestral haloalkane dehalogenases show restricted conformational dynamics. *Computational and Structural Biotechnology Journal*, *18*, 1497–1508. doi: 10.1016/j.csbj.2020.06.021.

Babkova, P., Sebestova, E., Brezovsky, J., Chaloupkova, R., & Damborsky, J. (2017).

Ancestral haloalkane dehalogenases show robustness and unique substrate specificity. *ChemBioChem*, *18*(14), 1448–1456. doi: 10.1002/cbic.201700197.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A. J., … Xenarios, I. (2016). UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: How to use the entry view. In D. Edwards (Ed.), *Plant bioinformatics*, (pp. 23–54). New York, NY: Humana Press.

Chaloupkova, R., Liskova, V., Toul, M., Markova, K., Sebestova, E., Hernychova, L., … Damborsky, J. (2019). Light-emitting dehalogenases: Reconstruction of multifunctional biocatalysts. *ACS Catalysis*, *9*, 4810–4823. doi: 10.1021/acscatal.9b01031.

Charleston, M. (2013). Phylogeny. S. Maloy and K. Hughes (Eds.), *Brenner's encyclopedia of genetics*, (pp. 324–325).

Diallo, A. B., Makarenkov, V., & Blanchette, M. (2010). Ancestors 1.0: A web server for ancestral sequence reconstruction. *Bioinformatics*, *26*(1), 130–131. doi: 10.1093/bioinformatics/btp600.

Foley, G., Mora, A., Ross, C. M., Bottoms, S., Sützl, L., Lamprecht, M. L., … Bodén, M. (2019). Identifying and engineering ancient variants of enzymes using Graphical Representation of Ancestral Sequence Predictions (GRASP). *BioRxiv*, 2019–2012. doi: 10.1101/2019.12.30.891457.

Gaucher, E. A. (2007). Ancestral sequence reconstruction as a tool to understand natural history and guide synthetic biology: Realizing and extending the vision of Zuckerkandl and Pauling. In D. A. Liberles (Ed.), *Ancestral sequence reconstruction*, (pp. 20–33). Oxford, UK: Oxford University Press.

Gaucher, E. A., Govindarajan, S., & Ganesh, O. K. (2008). Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature*, *451*(7179), 704–707. doi: 10.1038/nature06510.

Hanson-Smith, V., & Johnson, A. (2016). PhyloBot: A web portal for automated phylogenetics, ancestral sequence reconstruction, and exploration of mutational trajectories. *PLoS Computational Biology*, *12*(7), e1004976. doi: 10.1371/journal.pcbi.1004976.

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Research*, *47*(W1), W256–W259. doi: 10.1093/nar/gkz239.

Liberles, D. A. (Ed.). (2007). *Ancestral sequence reconstruction*. Oxford, UK: Oxford University Pres.

Matasci, N., & McKay, S. (2013). Phylogenetic analysis with the iPlant discovery environment. *Current Protocols in Bioinformatics*, *42*(1), 6–13. doi: 10.1002/0471250953.bi0613s42.

Olsen, G. (1990). The" Newick's 8: 45" tree format standard. Available at http://evolution.genetics.washington.edu/phylip/newick_doc.html.

Phylogeny. (2013). 324–325.

Procter, J. B., Thompson, J., Letunic, I., Creevey, C., Jossinet, F., & Barton, G. J. (2010). Visualization of multiple alignments, phylogenies and gene family evolution. *Nature Methods*, *7*(3), S16–S25. doi: 10.1038/nmeth.1434.

Ribeiro, A. J. M., Holliday, G. L., Furnham, N., Tyzack, J. D., Ferris, K., & Thornton, J. M. (2018). Mechanism and Catalytic Site Atlas (M-CSA): A database of enzyme reaction mechanisms and active sites. *Nucleic Acids Research*, *46*(D1), D618–D623. doi: 10.1093/nar/gkx1012.

Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science*, *27*(1), 135–145. doi: 10.1002/pro.3290.

Štěpánková, V. (2013). Expansion of access tunnels and active-site cavities influences activity of haloalkane dehalogenases in organic cosolvents. Haloalkane dehalogenases in non-conventional reaction media. *ChemBioChem*, *14*(7), 79. doi: 10.1002/cbic.201200733.

Taylor, W. R. (1997). Residual colours: A proposal for aminochromography. *Protein Engineering*, *10*(7), 743–746. doi: 10.1093/protein/10.7.743.

Watanabe, K., Ohkuri, T., Yokobori, S. I., & Yamagishi, A. (2006). Designing thermostable proteins: Ancestral mutants of 3-isopropylmalate dehydrogenase designed by using a phylogenetic tree. *Journal of Molecular Biology*, *355*(4), 664–674. doi: 10.1016/j.jmb.2005.10.011.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., … Lepore, R. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research*, *46*(W1), W296–W303. doi: 10.1093/nar/gky427.

Wheeler, L. C., Lim, S. A., Marqusee, S., & Harms, M. J. (2016). The thermostability and specificity of ancient proteins. *Current Opinion in Structural Biology*, *38*, 37–43. doi: 10.1016/j.sbi.2016.05.015.