BRNO UNIVERSITY OF TECHNOLOGY

Fakulta informačních technologií

DOCTORAL THESIS

Brno, 2020 Ing. MILOŠ MUSIL

BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

DEPARTMENT OF INFORMATION SYSTEMS
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

COMPUTATIONAL DESIGN OF STABLE PROTEINS
AUTOMATIZOVANÝ NÁVRH STABILNÍCH PROTEINŮ

DOCTORAL THESIS
DIZERTAČNÍ PRÁCE

AUTHOR AUTOR PRÁCE	Ing. Miloš Musil
ADVISOR VEDOUCÍ PRÁCE	Doc. Ing. Jaroslav Zendulka, CSc.

BRNO 2020

ABSTRACT

Stable proteins are utilized in a vast number of medical and biotechnological applications. However, the native proteins have mostly evolved to function under mild conditions inside the living cells. As a result, there is a great interest in increasing protein stability to enhance their utility in the harsh industrial conditions. In recent years, the field of protein engineering has matured to the point that enables tailoring of native proteins for specific practical applications. However, the identification of stable mutations is still burdened by costly and laborious experimental work. Computational methods offer attractive alternatives that allow a rapid search of the pool of potentially stabilizing mutations to prioritize them for further experimental validation. A plethora of the computational strategies was developed: i) force-field-based energy calculations, ii) evolution-based techniques, iii) machine learning, or iv) the combination of several approaches. Those strategies are usually limited in their predictions to less impactful single-point mutations, while some more sophisticated methods for prediction of multiple-point mutations require more complex inputs from the side of the user. The main aim of this Thesis is to provide users with a fully automated workflow that would allow for the prediction of the highly stable multiple-point mutants without the requirement of the extensive knowledge of the bioinformatics tools and the protein of interest.

FireProt is a fully automated workflow for the design of the highly stable multiple-point mutants. It is a hybrid method that combines both energy- and evolution-based approaches in its calculation core, utilizing sequence information as a filter for robust force-field calculations. FireProt workflow not only detects a pool of potentially stabilizing mutations but also tries to combine them together while reducing the risk of antagonistic effects.

FireProt$^{\text{ASR}}$ is a fully automated workflow for ancestral sequence reconstruction, allowing users to utilize this protein engineering strategy without the need for the laborious manual work and the knowledge of the system of interest. It resolves all the steps required during the process of ancestral sequence reconstruction, including the collection of the biologically relevant homologs, construction of the rooted tree, and the reconstruction of the ancestral sequences and ancestral gaps.

HotSpotWizard is a workflow for the automated design of mutations and smart libraries for the engineering of protein function and stability. It allows for a wider analysis of the protein of interest by utilizing four different protein engineering strategies: i) identification of the highly mutable residues located in the catalytic pockets and tunnels, ii) identification of the flexible regions, iii) calculation of the sequence consensus, and iv) identification of the correlated residues.

FireProt$^{\text{DB}}$ is a database of the known experimental data quantifying a protein stability. The main aim of this database is to standardize protein stability data, provide users with well-manageable storage, and allow them to construct protein stability datasets to use them as training sets for various machine learning applications.

KEYWORDS

protein engineering, protein stability, stability prediction, multiple-point mutants, ancestral sequence reconstruction

ABSTRAKT

Stabilní proteiny nacházejí široké uplatnění v řadě medicínských a biotechnologických aplikacích. Přírodní proteiny se vyvinuly tak, aby fungovaly převážně v mírných podmínkách uvnitř buněk. V důsledku toho vzniká zájem o stabilizaci proteinů za účelem jejich širšího uplatnění také v průmyslovém prostředí. Obor proteinového inženýrství se v posledních letech rozvinul do úrovně umožňující modifikovat proteiny pro různá využití, ačkoliv identifikace stabilních mutací je stále zatížená drahou a časově náročnou experimentální prací. Výpočetní metody se proto uplatňují jako atraktivní alternativa, která dovoluje prioritizovat potenciálně stabilizující mutace pro laboratorní práci. Během posledních let bylo vyvinuto velké množství výpočetních strategií: i) výpočty energie pomocí silových polí, ii) evoluční metody, iii) strojové učení a iv) kombinace více přístupů. Spolehlivost a využití nástrojů jsou často limitovány predikcí pouze jednobodových mutací, které mají malý dopad na stabilitu proteinů, zatímco sofistikovanější metody pro predikci multibodových mutací vyžadují větší množství práce na straně uživatele. Hlavním záměrem této práce je poskytnout uživatelům plně automatizované metody, umožňující návrh vysoce stabilních vícebodových mutantů bez potřeby pokročilých znalostí bioinformatických nástrojů a zkoumaného proteinu. V této práci jsou prezentovány následující nástroje a databáze:

FireProt je plně automatizovaná metoda pro návrh stabilních vícebodových mutantů z kategorie tzv. hybridních přístupů. Ve svém výpočetním jádře spojuje jak energetické tak i evoluční metody, přičemž evoluční informace jsou užívány především jako filtry pro časově náročné výpočty energií. Kromě detekce potenciálně stabilizujících mutací se FireProt rovněž snaží spojit tyto mutace do jednoho vícebodového mutanta s minimalizací rizika vzniku antagonistických efektů.

FireProt$^{\mathbf{ASR}}$ je plně automatizovaná platforma pro rekonstrukci ancestrálních sekvencí, která dovoluje uživatelům využít tuto strategii bez nutnosti velkého objemu manuální práce a hluboké znalosti zkoumaného proteinu. FireProt^{ASR} řeší všechny kroky ancestrální rekonstrukce, včetně sběru biologicky relevantních sekvencí, konstrukce zakořeněného fylogenetického stromu a rekonstrukce ancestrálních sekvencí.

HotSpotWizard je nástroj pro návrh mutací a mutačních knihoven za účelem zlepšení stability a aktivity zkoumaných proteinů. Nástroj dovoluje provést i širší analýzu za využití čtyř různých strategií běžně používaných v oboru proteinového inženýrství: i) identifikace evolučně variabilních pozic v blízkosti katalytických kapes a tunelů, ii) identifikace pohyblivých regionů, iii) výpočet sekvenčního konsensu a iv) identifikace korelovaných pozic.

FireProt$^{\mathbf{DB}}$ je databáze dostupných experimentálních dat popisujících stabilitu proteinů. Hlavním účelem této databáze je standardizovat data v oblasti proteinové stability, poskytnout uživatelům platformu k jejich snadnému ukládání a umožnit intuitivní vyhledávání, které by mohly být využité k trénování nových nástrojů s využitím technik strojového učení.

KLÍČOVÁ SLOVA

proinové inženýrství, stabilita proteinů, predikce stability, vícebodové mutace, ancestrální rekonstrukce sekvencí

MUSIL, Miloš. *Computational design of stable proteins*. Brno, 2020, 123 p. Doctoral thesis. Brno University of Technology, Fakulta informačních technologií, Department of Information Systems. Advised by Doc. Ing. Jaroslav Zendulka, CSc.

Typeset by the `thesis` package, version 4.00; `http://latex.feec.vutbr.cz`

DECLARATION

I declare that I have written the Doctoral Thesis titled “Computational design of stable proteins” independently, under the guidance of the advisor and using exclusively the technical references and other sources of information cited in the thesis and listed in the comprehensive bibliography at the end of the thesis.

As the author I furthermore declare that, with respect to the creation of this Doctoral Thesis, I have not infringed any copyright or violated anyone’s personal and/or ownership rights. In this context, I am fully aware of the consequences of breaking Regulation § 11 of the Copyright Act No. 121/2000 Coll. of the Czech Republic, as amended, and of any breach of rights related to intellectual property or introduced within amendments to relevant Acts such as the Intellectual Property Act or the Criminal Code, Act No. 40/2009 Coll., Section 2, Head VI, Part 4.

Brno . . . .

author’s signature

ACKNOWLEDGEMENT

I would like to give my thanks to my supervisors Jaroslav Zendulka, Jiří Damborský and David Bednář for their expertise and continuous support during the years of my endless studies. Furthermore, thanks to my colleagues from Loschmidt Laboratories, especially to Jan Štourač, who provided me with a lot of help for most of my projects and Jaroslav Bendl, who dragged me into dark waters of scientific research. Finally, my gratitude is dedicated to my family, friends and my wife that had to learn a bit of the science againts her will.

Contents

List of Figures

List of Tables

1 Introduction

Proteins are the building blocks of every living organism, where they perform a wide variety of functions, including DNA replication, catalysis of metabolic reactions, responding to the stimuli, and transporting molecules between different parts of the living structures [1]. They consist of one or more long chains of amino acid residues connected by peptide bonds. The sequence of the amino acids in the protein determines its structure and function. Therefore, mutations leading to amino acid alteration are the driving force of evolution at the molecular level.

Over time, Nature has developed a remarkable diversity of biochemical reactions vital to the continuing evolution of living organisms and the preservation of life. These biochemical reactions scale from the simple one-step degradation processes to more complex pathways employing several different proteins. The recent advances of the next-generation sequencing, together with the steady growth of the computational resources and advances in bioinformatics have allowed wider access to these naturally evolved processes and their utilization in various medical, industrial and biotechnological applications. Furthermore, protein engineering has matured to the point that enables tailoring of native proteins for specific practical applications, thus overcoming the limitations of the native variants that have evolved to function in mild conditions [38].

As a result, the ability to understand what drives the protein folding, its function, and other characteristics is crucial for further advances in the field of protein engineering as the mutations introduced into a modified protein can affect it in many different ways. Only a small portion of the mutations will have a beneficial impact on the protein characteristics, considering its intended purpose in the specific medical or industrial applications. Some of the mutations can influence protein stability, while others will affect its solubility, activity, expression yields, or ability to fold into the 3D structure and create more complex quaternary structures by interacting with other molecules. Both positive and harmful effects can be observed by introducing mutations into the sequence of the protein of interest, and in many cases, there is an apparent trade-off between some of the characteristics of the proteins [2, 3, 4]. As a result, mutation improving protein stability can harm its function and vice versa. Thus, it is necessary to analyze a large number of mutations to obtain the variant most suitable for its intended use.

This Thesis focuses mainly on the aspect of protein stability as one of the main characteristics that determine the usability of the natural biochemical reactions in the harsh environment of the medical and industrial applications. Stable proteins are able to withstand extreme temperatures, acidic or basic pH, or an unfavorable effect of organic solvents and proteases [6]. Furthermore, stable proteins are often distinguished by higher half-life, making them easier to transport and store for later use [8]. As a result, there is a high interest in increasing protein stability, and many different methods were designed over the years to accomplish such a task.

In the ideal case, the saturation mutagenesis would be applied to evaluate every possible mutation on every position of the engineered protein. However, such search space would be enormous, and the experimental evaluation laborious and costly. Therefore, there rises a need for effective and precise computational methods to predict protein stability. To satisfy this goal, a number of in silico tools have been developed recently. Unfortunately, due to the limited reliability and potential antagonistic effect between individual mutations, only single-point mutations with an almost negligible effect on protein stability are usually predicted in the existing tools. Such mutations typically enhance the stability of the target proteins only mildly, while higher stabilization can be achieved by engineering multiple-point mutants [17].

1.1 Objectives of the Thesis

The main aim of this Thesis is to develop new methods that would allow for the design of highly stable multiple-point mutants, and it presents several possible solutions. FireProt is a hybrid method that combines several different computational approaches into a single workflow, allowing for a more robust and reliable construction of the stable multiple-point mutants. The second solution, FireProt^{ASR}, is based on natural evolution and the observation that the ancestral proteins were significantly more stable than their extant counterparts. Finally, HotSpotWizard is presented as a tool that can be utilized to highlight potentially interesting residues in the protein, where mutations could have a positive impact not only on the stability but also on other protein characteristics. The new database FireProt^{DB} is introduced as a possible solution for a current troubling situation surrounding the storage and management of the existing data obtained from the laboratory measurements of the protein stability. Such a compilation of manually curated data is very much needed for future development of reliable predictive tools based on machine learning.

The main goals of this Thesis are:

- to analyze the physico-chemical forces that participate in the increase of protein stability
- to construct a reliable protein stability dataset that could be used for the validation of the existing tools and force-fields and for the training of the methods based on machine learning

- to develop, integrate and thoroughly validate a hybrid workflow for an automated design of the stable multiple-point mutants
- to resolve the algorithmic and technical problems connected with the automatization of the ancestral sequence reconstruction with the primary focus on improvement of the proteins' thermal stability
- to develop, integrate and validate a fully automated workflow for ancestral sequence reconstruction

1.2 Organization of the Thesis

This Thesis is organized as follows. With regards to the interdisciplinary nature of the work described in **Chapter 1**, **Chapter 2** is devoted to the introduction into the biological background of the problematics of the protein stability engineering. The main aim of **Chapter 3** is to acquaint the reader with the different methods and strategies that are viable for protein stabilization. **Chapter 4** then focuses on the available experimental data, and **Chapter 5** describes the current state-of-the-art. Lastly, **Chapter 6** provides a deeper understanding of the problematics of the ancestral sequence reconstruction as the means for protein stabilization. These six chapters try to establish the theoretical basis of protein stability prediction but also critically discuss practical applications of the described methods together with their advantages and their most common pitfalls. The practical part of this Thesis presents a limited selection of the achieved results. **Chapter 7** summarizes the conducted research and published manuscripts dealing with the problematics of the protein stability. Four published manuscripts corresponding to three developed tools and one database are attached at the end of this thesis. **Appendix A** describes FireProt, the hybrid workflow for the design of the stable multiple-point mutants. **Appendix B** is focused on FireProt$^{\mathrm{ASR}}$, an automatized workflow for the ancestral sequence reconstruction. **Appendix C** is devoted to the FireProt$^{\mathrm{DB}}$, a novel database for the storage and maintenance of the protein stability data. Protein engineering software HotSpotWizard is lastly described in the **Appendix D**. Finally, the results are concluded in **Chapter 8**.

2 Protein stability

Protein stability is one of the key properties determining protein's applicability under harsh conditions. The stable protein is able to withstand extreme temperatures [5], acidic or basic pH, or unfavourable effects of organic solvents and proteases [6]. Furthermore, stable proteins are usually positively correlated with expression yields [7] and their half-life [8]. As a result, there is a great interest in increasing protein stability to enhance its utility in various medical, biotechnological, and industrial applications.

Stability is strongly connected with proteins' conformation and can be qualified as the net balance of various intramolecular interactions and conformational entropy [9]. Those forces determine whether a protein will stay in its native folded conformation. They can be strengthened or disrupted by introducing mutations into the protein of interest. In this chapter, various physical and biochemical forces will be described together with the mechanisms of protein folding and well-established metrics for the protein stability quantification.

2.1 Stability of the folded protein

In 1969, Cyrus Levinthal stated that, because of the high number of degrees of freedom in an unfolded polypeptide chain, folding of the protein from its primary to the tertiary structure cannot occur randomly [26]. Based on his estimation, if we consider a relatively small protein of 100 amino acids with only three allowed conformations per residue and a sampling time of only 0.1 ps per conformation, folding of such a protein could demand as long as $5 * 10^{34}$ seconds. Therefore, if the protein were to find its stable folded conformation by a simple random trial, this process would take longer than is the age of our universe, while in reality, the protein folding usually occurs in the matter of microseconds up to several minutes in the case of complex proteins.

While Levinthal's paradox contradicts the possibility of random folding, Afinsen's thermodynamic hypothesis further supports this theory by proving that for a globular protein in their standard physiological environment, the native structure is determined only by the protein's amino acid sequence [27]. This hypothesis shows that the process of protein folding not only cannot be random, but it is also deterministic, meaning that at the same environmental conditions, the native structure of the protein is defined only by sequence of amino acids in the polypeptide chain.

Both Levinthal's and Afinsen's claims have acknowledged the existence of powers governing protein folding. Those powers can be distinguished on covalent and non-covalent interactions, together with the factor of conformational entropy. Covalent bonds are very strong and stable under standard environmental conditions. Covalent interactions are mostly created by sharing the valence electrons between atoms in the polypeptide chain. Covalent interactions are, therefore, the main forces governing the creation of the protein's primary structure. Non-covalent interactions are significantly weaker, and they play a fundamental role in the construction of proteins secondary, tertiary, and quaternary structures. Non-covalent interactions are electrostatic, polar, and non-polar [9].

Electrostatic interactions are long-range charge-charge interactions between charged residues (Arginine, Lysine, Glutamine, Asparagine and also Histidine in low pH). The strength of those interactions decreases with r^2 according to Coulomb's law. Furthermore, electrostatic interactions are strongly dependant on the environment as they are influenced by the pH of solvent, salt concentration, and permittivity. Solvents with higher permittivity, such as water, shields the charged residues from each other in the exposed, solvent-accessible regions of the protein.

Polar interactions can be divided into hydrogen bonds and aromatic interactions. Polar residues (Serine, Threonine, Aspartic acid, Cysteine, Tryptophan, Tyrosine, and Histidine) can share hydrogen attached to an electronegative atom with hydrogen acceptor. This usually occurs at a distance of about 3 Ångström (Å). Hydrogen bonds are also the main driving forces for the formation of secondary structures. Aromatic interactions are attractive forces between aromatic rings of the aromatic residues (Phenylalanine, Tryptophan, Tyrosine, and Histidine) governed by their π electrons. With aromatic interactions, the distance of the center of mass is about 5 Å. Polar interactions are crucial in the governing of the formation of the secondary structures.

Non-polar interactions are responsible for the creation of tertiary structures. Van der Waals interactions are weaker, short-ranged attractive and repulsive forces between all the atoms in the protein molecule. However, their effect decreases fast with distance and they are negligible beyond 5 Å. Tertiary structure is also influenced by hydrophobic effect due to the unfavourable entropy of the water molecules ordered around hydrophobic residues (Phenylalanine, Proline, Methionine, Leucine, Isoleucine, Valine, and Alanine). These residues have a tendency to aggregate, forming a hydrophobic core of the protein, and exclude water molecules. Hydrophobic effect leads to favourable increase of hydrogen bonding between water molecules and minimizes the area between water and non-polar residues.

Conformational entropy is associated with a number of conformations of the proteins structure. It is a major contributor to the energetic stabilization of the denatured state and therefore acts as a countering force to the sum of electrostatic, polar and non-polar interactions (Figure 2.1). The conformational entropy yielded by the random coils is significantly higher than the entropy gain given by the secondary structures such as α-helixes and β-sheets. Due to this reason, proteins with higher number of secondary elements are usually more stable than the ones with the high concentration of the random coils.

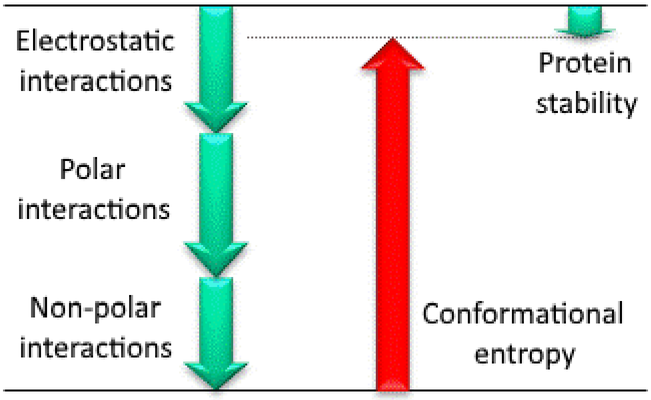


Fig. 2.1: Major forces influencing protein stability. Protein stability is given as a difference of the conformational entropy and the sum of the electrostatic, polar and non-polar interactions. (adapted from [10])

2.2 Mechanisms of protein folding

Several different mechanisms of protein folding were designed to explain the process in which non-covalent interactions transform the polypeptide chain into a complex tertiary structure [12]. First, the nucleation-growth model presumed the continuous growth of the tertiary structure from the initial nucleus of the local secondary structure. However, this model was dismissed as it did not account for folding intermediates. In response, several other models were designed (Figure 2.2):

Framework model: the secondary structure is folded first and is followed by the coalescence of the secondary structural units to the structure of the native protein.

Hydrophobic collapse model: the polypeptide initially forms secondary structures representing localized regions of predominantly hydrophobic residues. Due to the polypeptide's contact with the molecules of water, thus creating intense thermodynamic pressure, those regions are then aggregated into a tertiary conformation with a hydrophobic core.

Nucleation-condensation model: secondary and tertiary structure is formed in parallel as the formation of the tertiary structure is catalyzed by the folding of the initial nucleus – a small segment of the protein with correctly folded secondary structure. However, this initial nucleus is stable only in the presence of approximately correct tertiary structure interactions.

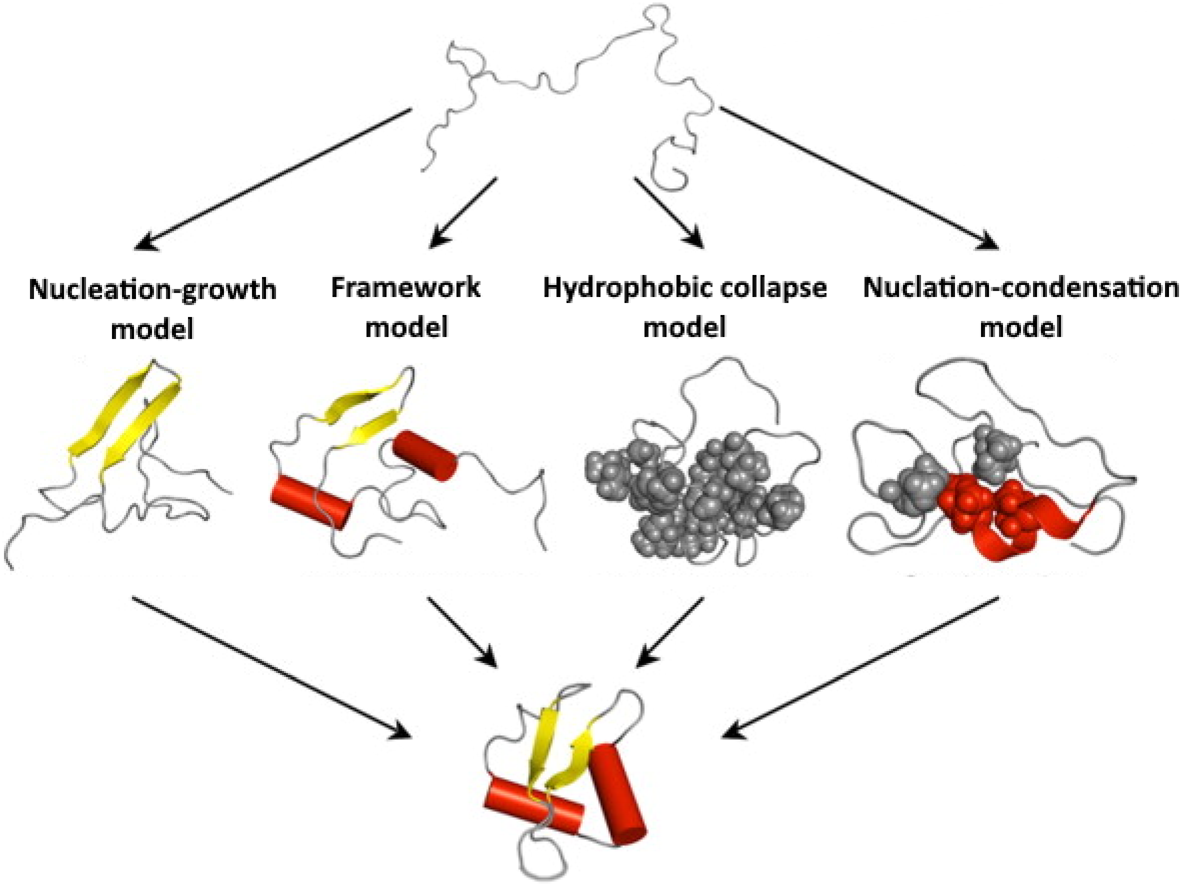


Fig. 2.2: The visualization of the various suggested mechanisms of protein folding. (adapted from [12])

2.3 Protein stability quantification

In the field of protein engineering, there are several ways how to quantify the protein stability. The two most common are melting temperature and Gibbs free energy [9].

2.3.1 Gibbs free energy

Gibbs free energy (G) is a thermodynamic potential that can be used to calculate the maximum of reversible work performed by a thermodynamic system at a constant temperature and pressure. It is defined as

$$G = H - TS,$$

where H is the enthalpy, T is the temperature, and S stands for the entropy. The official SI unit is Joule, however, in biology, Calories are often used instead. The stability of the protein is generally represented by the change in the Gibbs free energy upon folding (ΔG), which means the difference between free energies of the folded and unfolded state of the protein.

$$\Delta G = G_{folded} - G_{unfolded},$$

Finally, if we are interested in the effect that the given amino acid mutation has on protein stability, we measure the so-called change of Gibbs free energy upon mutation ($\Delta\Delta G$), which is the difference between ΔG of the mutated and wild-type protein.

$$\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wild-type}$$

The most commonly used unit is kcal/mol, and in this scenario, the negative value of $\Delta\Delta G$ indicates a stabilizing mutation. However, the format of $\Delta\Delta G$ is not standardized, and therefore in some studies, mutant and wild-type can be switched, meaning that the improvement of the protein stability will be noted by the positive sign. The computation of $\Delta\Delta G$ is based on the thermodynamic cycle captured in Figure 2.3.

2.3.2 Melting temperature

A second way, how to quantify protein stability is the melting temperature (T_m). The definition of melting temperature is

$$\Delta G_{folding}(T_m) = 0$$

In other words, the temperature at which free energy of the unfolded and folded states is equal, and half of the population is unfolded, and the other half is folded. Similarly to the Gibbs free energy, ΔT_m indicates the change of melting temperature upon mutation. While there is a strong correlation between Gibbs free energy and melting temperature (Pearson correlation coefficient is approximately 0.71 [109]), the transformation between the two is not exactly linear, and therefore there is no simple equation allowing for the estimation of $\Delta\Delta G$ based on the values of ΔT_m and vice versa.

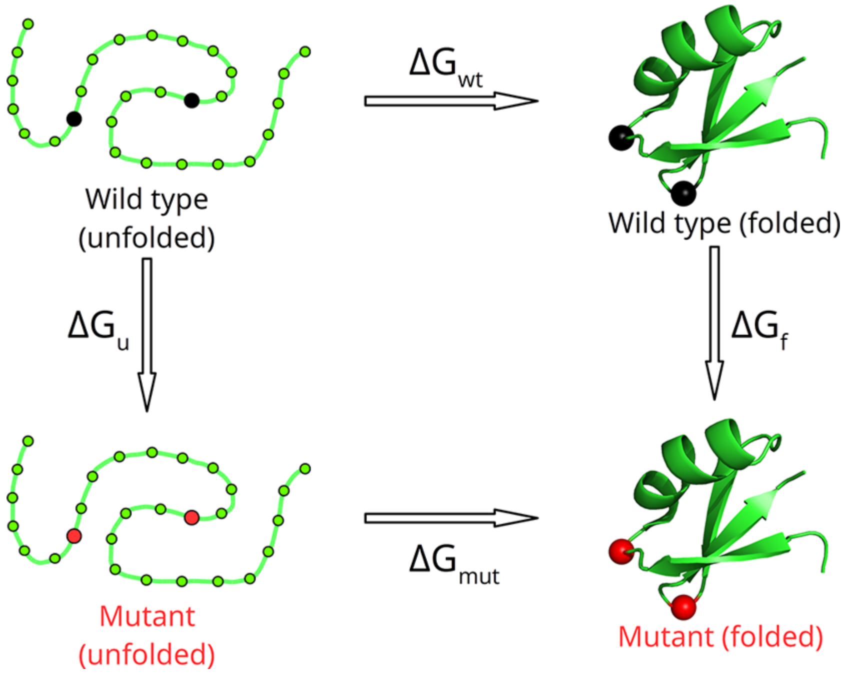


Fig. 2.3: Thermodynamic cycle commonly utilized for the computation of $\Delta\Delta G$. The change of the Gibbs free energy upon mutation is estimated as a difference of the Gibbs free energy upon folding of the wild-type and mutant protein, respectively. In the figure, the respective mutation sites have been coloured in black for wild-type and red for the mutant protein. [11]

2.4 Laboratory measurements

Considering the laboratory techniques, there are several ways how to measure protein stability [9]. The results provided by the individual methods will differ not only based on the used experimental conditions (salinity, pH of the buffer, temperature ramp) but in a smaller scale also on the selected method itself. Therefore, only the measurements obtained with the same method and experimental conditions should be utilized for the comparison of the experimental results [11].

Differential scanning calorimetry: is thermoanalytical technique, where the difference in the amount of heat required to increase the temperature of a sample and reference is measured as a function of temperature. It is one of the most widely used methods for studying the thermodynamics of protein unfolding (Figure 2.4a).

Circular dichroism: is based on the circularly polarized light. There is a differential absorption of left- and right-handed light and therefore left-hand and right-hand circular polarized light represents two possible spin angular momentum states of a photon. This phenomenon is exhibited in the absorption bands of optically active chiral molecules. Circular dichroism is commonly used to investigate the stability of secondary structure of proteins (Figure 2.4b).

Absorption spectroscopy: is based on the measure of the absorption of radiation as a function of frequency or wavelength, due to its interaction with a sample. The sample absorbs energy and the intensity of the absorption varies as a function of frequency. It finds its use in the analytical chemistry.

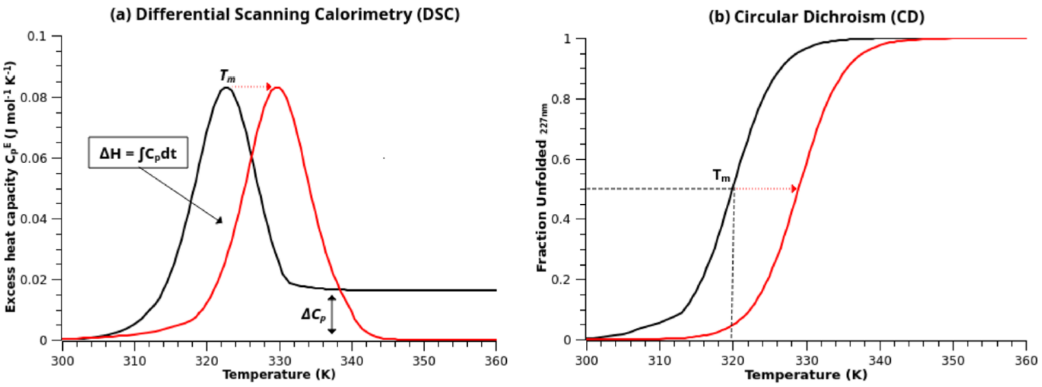


Fig. 2.4: Representative experimental methods to quantify protein stability. Curves for hypothetical wild-type protein are shown in black, while improved variant exhibiting higher stability is visualized in red. (a) Differential scanning calorimetry curve. T_m is the midpoint of the transition, ΔC_p is the difference between the pre- and post-transition baselines, and ΔH is the area under the curve between the pre- and post-transition baselines. (b) Circular dichroism curve. Following the change of molar ellipticity at a specific wavelength over a wider temperature range monitors the change in secondary structure of and unfolding protein. The midpoint of the sigmoid curve is related to T_m of the protein. [11]

In recent years, new experimental techniques allowed for faster and cheaper measurements of the proteins' characteristics. However, for large-scale analyses such as saturation mutagenesis of each possible mutation in the protein of interest, the laboratory experiments still represent an unattainable solution. Therefore, there is a need for less expensive in silico approaches.

2.5 Force-field calculations

With the widespread of information technology and the constant growth of the available computational resources, it is now possible to evaluate protein stability without the need for the use of highly laborious experimental methods. This is viable by calculating the free energy using the existing force-fields, which simulates the physico-chemical effects occurring in the protein structure. A simple example of force-field is such as bellow [13, 14]:

In our example force-field, the folded state free energy is calculated as a sum of individual atomic forces in the proteins tertiary structure as

$$G_F = G_{hy} + G_{el} + G_{hb} + G_{vw} + G_{ss},$$

where G_{hy}, G_{el}, G_{hb}, G_{vw} and G_{ss} are hydrophobic, electrostatic, hydrogen bonding, van der Waals and disulphide bonding free energies. Hydrophobic free energy can be estimated from solvent accessibility. It is calculated as

$$G_{hy} = \sum \Delta\sigma_i [A_i(folded) - A_i(unfolded)],$$

where i stands for different atom types, and $A_i(folded)$ and $A_i(unfolded)$ represent the accessible surface area (ASA) of each atom type in the folded and the unfolded state of the protein structure, respectively. Finally, $\Delta\sigma_i$ are the atomic solvation parameters. The significant contributors to electrostatic interactions are the charged side chains of the residues Lysine, Histidine, Arginine, and Glutamic and Aspartic acid. It has been observed that the electrostatic interactions are strongly context-dependent, where electrostatic interactions on the protein surface are generally contributing less than 1 kcal/mol, and buried ones are contributing around 3 kcal/mol to protein stability [18]. Hydrogen bonds are one of the main participants in the creation of secondary structures. Their contributions are calculated mostly based on their geometric information. One of the approaches was described with a program HBPLUS [19]. Van der Waals energies can be computed utilizing Lennard-Jones potential [20] as

$$G_{vw} = \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}},$$

where $A_{ij} = \epsilon_{ij}^*(R_{ij}^*)^{12}$, $B_{ij} = 2\epsilon_{ij}^*(R_{ij}^*)^6$, $R_{ij}^* = (R_i^* + R_j^*)$, $\epsilon_{ij}^* = \epsilon_i^*\epsilon_j^*$. The indexes i and j are the indexes of the individual atoms and R^* and ϵ^* are the van der Waals radius and well depth, respectively.

Finally, the site-directed mutagenesis experiments established that one disulphide bond to protein stability is approximately 2.3 kcal/mol [21]. The unfolded state free energy is calculated from entropic and non-entropic free energies as

$$G_U = G_{en} + G_{ne},$$

where G_{en} represents entropic and G_{ne} nonentropic free energies. The free energy of the unfolded protein is strongly connected with the size of the protein, and it was estimated that one residue adds approximately 1.2 kcal/mol [22].

3 Computational approaches for prediction of protein stability

In the ideal case, saturation mutagenesis of each possible mutation would be carried by the rigorous experimental validation. However, in most projects, such validation is close to impossible due to the costly and laborious nature of those experiments. Considering a standard protein consisting of approximately 300 amino acids, this leaves us with over 5,000 single-point mutations. Furthermore, single-point mutations often provide an almost negligible effect on protein stability ($< 2kcal/mol$) [15, 16], and therefore combining several stabilizing mutations is typically required to procure a significant improvement of protein stability [17]. Unfortunately, the additive effect of stabilizing mutations is not guaranteed as synergistic or antagonistic effects can occur between any subset of stabilizing single-point mutations. Mutations are considered synergistic if their combined effect on protein stability is notably higher than the sum of the individual mutations, while the antagonistic effect means the exact opposite. The synergistic effect usually appears due to the creation of a new physico-chemical interaction such as a salt bridge between anionic carboxylate and cationic ammonium or a disulphide bridge between two cysteine residues. On the other hand, the antagonistic effect disturbs some of the newly introduced interactions or creates clashes between the side chains of the mutated or original residues. This, for example, can be easily observed when several mutations are designed to fill the same space in the structure of the protein, filling the void each by itself, however being unable to fit in if combined. This could either damage protein stability or even completely prevent it from a successful folding.

In most cases, antagonistic effects are not easily detectable, and therefore further experimental validation is needed. With only 100 potentially stabilizing mutations, close to 5,000 experiments would have to be performed to evaluate all possible double-point mutants, and this number is exponentially increasing with each added mutation. As a result, there is an ever-growing need for fast and accurate computational methods that would allow for rapid evaluation of the potentially stabilizing mutations, and serve as a reliable tool for the prioritization of mutations for the rigorous laboratory experiments. In general, the computational methods for the prediction of the effect of mutations on protein stability can be divided into four categories [11]:

Force-field methods relying on the calculation of the $\Delta\Delta G$ based on the models of molecular mechanics.

Phylogenetic analysis utilizing the evolutionary information contained in the set of homolog sequences.

Machine learning methods constructing a computational model based on the stability data provided by previous experimental validation.

Hybrid methods and meta-predictors combining several of the previous approaches or several different methods of a single approach together to obtain more robust and reliable results.

3.1 Principles of methods based on force-field calculations

In silico design of the stable proteins based on the calculation of the energy force-fields is deeply rooted in our current state of knowledge of the physico-chemical properties of the individual amino acids and their description by molecular mechanic force-fields. Therefore, these calculations do not rely on the availability of the diverse, high-quality experimental data. In general terms, a force-field is a description of all bonded and non-bonded interactions in the protein of interest [38, 39]. These interactions are captured in the energy-field equation used to estimate the potential energy of a molecular system [40]. The most accurate methods in this category are the free energy methods, relying on molecular dynamics (MD), or Metropolis Monte Carlo simulations. Unfortunately, those methods require a tremendous amount of computational power and are viable only for a limited number of mutations or smaller, less expensive systems of interest [41]. A number of heuristic approaches were created over the last decades to overcome this bottleneck, however huge analysis is still viable only with the use of simulation-independent stability predictors that can be divided into three categories [42, 43]:

Physical effective energy functions (PEEFs) are closely related to classical molecular mechanic force-fields, which allow for a fundamental analysis of the molecular interactions [40]. The individual terms of the energy-field equations are calculated via the simplification of the known physical laws and are still burdened by high computational demands reaching from hours up to several days for a single mutation. However, similarly to the molecular dynamics methods, they are versatile, accurate, and capable of predicting the behaviour of the enzymes under non-standard conditions such as non-physiological pH, non-standard salinity, or elevated temperature [54].

Statistical effective energy functions (SEEFs) are viable for rapid analysis as they can predict changes in stability over the entire sequence space of an average-sized enzyme in a matter of minutes [44, 45]. Compared to PEEFS, terms used in the SEEFs energy-field equations are derived from curated data sets of available experimental protein structures projected into several stability descriptors. An effective potential can be then extrapolated for every descriptor distribution and utilized as a part of the overall energy function [44, 46]. SEEFs do not explicitly model physical molecular interactions and are strongly dependent on the folded protein structures' availability and diversity [43].

Empirical effective energy functions (EEEFs) represent a bridge between PEEFs and SEEFs as they include both physical and statistical terms in their energy-field equations, which are weighted and parametrized to match experimental data [42, 43]. The thermodynamic data used in the derivation of terms typically originate from mutational experiments conducted under standard conditions. As a result, EEEFs provide a reasonable compromise between computational demands and the accuracy of the free energy function [49]. A major drawback of EEEFs is that their applicability is restricted to the environmental conditions under which the experimental data used for the parametrization were acquired [50, 51].

Even though force-field-based calculations are currently considered the most powerful tool for predicting the effect of mutations on protein stability, the accuracy of the energy functions is still suboptimal due to insufficient conformational sampling, imbalances in force-fields, and the problems connected with the existing data sets [50, 55, 56, 57, 58, 59]. The computation of $\Delta\Delta G$ is based on the thermodynamic cycle, and therefore it requires modelling the folded and unfolded states of both wild-type and mutant protein [32, 41]. The value of $\Delta\Delta G$ is then established as the difference between both folded states with several issues reported for various energy functions. All energy functions are known to overestimate hydrophobicity and tend to favour nonpolar mutations as the stabilizing ones [28, 29, 23]. PEEFs often underestimate the stabilization provided by the buried polar residues as they overestimate the energetic cost of unsatisfied salt bridges and hydrogen bonds in the protein core [37, 60, 61]. The estimation of the conformational and solvent-related entropy is also imprecise. The inability of force-field methods to account for entropy-driven contributions can be partially resolved by utilizing evolutionary-based approaches inside the more robust hybrid workflows [35, 36, 62, 63]. Another shortcoming comes with the prediction of the multiple-point mutants as most stability predictors have been parametrized using only a single-point mutant datasets. As a result, the predictive power for the multiple-point mutants is limited for most of the existing force-fields [64, 65]. This shortcoming can also be attributed to the insufficient conformational sampling of the folded state, especially in the case of mutations introducing large-scale backbone movements into the mutant protein structure [66]. In PEEFs and EEEFs, such movements are simulated by the utilization of the rotamer libraries to the fixed protein backbones, thereby reducing computational demands while providing comparable precision for the predictions of the single-point mutations [59]. However, this approach does not stand in the case of the multiple-point mutants and multistate designs. Therefore, flexible backbone sampling techniques [55, 67, 68, 69], generating conformational ensembles and utilizing energetically more favourable conformations, are required. Finally, the accuracy of the force-field methods is strongly dependent on the quality of the available tertiary structure. Their applicability for the proteins without resolved tertiary structure is given by the reliability of the structure modelling tools and the similarity of the closest sequence homology with a known tertiary structure. Furthermore, structures obtained by X-ray crystallography ($> 90\%$ proteins in PDB database [25]) do not necessarily reflect the global energy minimum of the native state of the protein in its natural environment [70] and may, in some cases, be misleading starting point for a comprehensive prediction of protein stability [51, 71].

3.2 Principles of methods based on phylogenetic analysis

A phylogenetic or evolutionary analysis are methods that take advantage of the information hidden in the set of homolog sequences. The evolutionary approach's main advantage is that those methods do not require tertiary structure and are therefore viable for the majority of known protein sequences (about 200 million of sequences in UniProt [24] compared to 100 thousand structures in the PDB database [25]). The only limitation in its applicability occurs in the families with the low representation of sequences in the database. However, with the rise of the next-generation sequencing methods, this limitation slowly mitigates as the number of sequences in the databases almost doubles every three years. The two most widely used phylogeny-based methods are consensus design and ancestral sequence reconstruction, both built on top of the reasonably-sized set of homolog sequences.

Consensus design (CD) starts by building a compact multiple-sequence alignment (MSA) using a small number of homolog sequences ranging between a dozen and a few hundred. This MSA allows for a computation of every amino acid's frequency distribution in each position in the sequence alignment [83]. Positions, where one or just a few amino acids are significantly more prevalent than others, are conserved as those residues changed only sparsely during evolution. CD's core assumption is that conserved positions are somehow crucial for the function of the protein (stability, activity, protein folding, etc.), and the most frequent amino acid at the given position is more likely to be stabilizing [83, 84, 85, 86, 87, 88]. CD can be utilized when amino acid in the designed sequence differs from the most dominant ones in those conserved regions. This residue's mutation to the dominant amino acid suggested by evolution often leads to a non-negligible improvement of protein's thermal stability. It has been observed that high levels of sequence diversity in the MSA can interfere with the preservation of catalytic activity in the designed proteins, particularly if the MSA contains both prokaryotic and eukaryotic sequences [84, 89]. On the other hand, including only closely related homologs might introduce an evolutionary bias that prohibits CD from discovering more thermostable variants [88]. In recent studies, the proportions of neutral and destabilizing CD mutations have been estimated to be 10 and 40%, respectively [83, 87]. In 2012, Sullivan was able to increase the proportion of correctly identified stabilizing mutations to 90% by discarding mutations of the residues with high statistical correlations to other positions in the MSA [84]. This would suggest an inability of the CD analysis to account for any synergic or antagonistic effects. The second possible weakness comes from an apparent phylogenetic bias when the MSA is dominated by a small number of highly similar subfamilies [85, 94]. If tertiary structure for the protein of interest is available, the CD can be further refined by utilizing information about an active site, secondary structures, and intramolecular contacts or by analyzing molecular fluctuations based on crystallographic B-factors or MD simulations [90, 91].

Ancestral sequence reconstruction (ASR) is a probabilistic method that explores the deep evolutionary history of homolog sequences to reassemble protein's evolutionary trajectory [93, 96]. The method was initially developed to study molecular evolution. ASR is able to unearth sequences of the long-extinct genes and organisms from which the current ones evolved and is, therefore, an invaluable tool in the field of evolutionary biology. ASR has also been shown to be a very effective strategy for thermostability engineering [33, 34, 5, 95] and for improving other protein's characteristics such as specificity, activity, or expression rates. Similarly to CD, ASR starts with the MSA's construction from the set of relevant homolog sequences. However, while CD relies on the simple analysis of the conservation of amino acids on the individual positions in the sequence alignment, ASR goes much further by considering evolutionary information depicted by the phylogenetic tree. Two main algorithms, maximum-likelihood [97, 98] (ML) and Bayesian inference [99] (BI) were designed to interfere with ancestral sequences from MSA and phylogenetic tree. Over the years, many tools were built to make those algorithms accessible to the broad scientific community. However, several crucial steps in the calculation of ASR were not yet resolved in a satisfactory way that would allow for a fully automatized inference of the ancestral proteins, i.e., selection of the biologically relevant subset of homolog sequences, rooting of the phylogenetic tree and the reconstruction of the ancestral gaps. This limits the ASR's applicability as the method requires an in-depth knowledge of the biological system of interest and necessary bioinformatics tools together with the abysmal amount of manual work. ASR approach, its methods, advantages, and shortcomings will be further discussed in the **chapter**

6. Furthermore, **appendix B** describes a novel solution for a fully automatized calculation of the ancestral sequences. FireProt^{ASR} utilizes several filters to obtain homolog sequences with the same or similar protein function and more than ten bioinformatics tools and databases to proceed with the ancestral reconstruction using a single protein sequence as a sole input of the calculation. This allows ASR to be utilized by a wide scientific community without prior knowledge of the method and required bioinformatics tools.

3.3 Principles of the methods based on machine learning

In recent years, machine learning has become one of the most dominant approaches in predicting protein stability [72, 73, 74] and many other fields reaching far above the limited scope of protein engineering applications. The popularity of machine learning methods comes mostly from their ability to construct computational systems without being explicitly programmed. Statistical techniques are used to analyze training data sets and recognize patterns that might be difficult to detect, given the limitations of human knowledge and cognitive abilities. The system based on the machine learning approach can be trained either with or without supervision. Both find their utilization in the field of protein engineering. In the supervised approaches, the system is given a set of training inputs and the expected outputs in the form of labels indicating each input's correct classification. Those methods are well-suitable for training predictive systems. On the other hand, unsupervised approaches are mostly implemented for tasks involving data clustering.

As the machine learning systems are constructed during the learning process, they do not require a full understanding of the mechanistic principles underpinning the target function. This advantage shines, especially in situations where there is a severe gap in human knowledge-base, and therefore expert construction of the predictive systems is not entirely possible. Machine learning can also expand existing systems by discovering previously unrecognized features, patterns, and relationships hidden in the training dataset. Furthermore, machine learning methods are very flexible because any characteristic extracted from the data can be used as a feature if it improves the prediction accuracy, i.e., minimizes the prediction error. Moreover, machine learning is also much less time demanding than other methods because once the model has been constructed using the training data, predictions can be obtained at an almost instant rate.

However, the reliability of the machine learning approaches strongly depends on the quality and size of the training data set. The weights representing the relative importance of the individual features and the relationships between them are based on the provided experimental observations. Consequently, it is crucial to use high-quality experimental data with high consistency of experimental measurements and wide diversity when training and testing machine learning methods. The size and balance of the training dataset must also be considered. A modest dataset with only a few hundreds of cases might be too small to establish useful descriptors during learning. Additionally, lower diversity of the training data usually leads to a higher risk of overtraining and, therefore, losing its ability to generalize on a new, previously unknown data. In such cases, the weights assigned to the individual descriptors tend to be influenced by over-representing some of the descriptors in the training data, while other features with high informational value are under-estimated or omitted entirely. Unbalanced training datasets with substantial differences in the individual prediction categories' size could also lead to erroneous predictions. For example, a training dataset in which more than two-thirds of the mutations are stated as deleterious would mislead the predictor to classify most mutations as deleterious because of the prevalence of such mutations during the learning. Some methods, namely support vector machines and random forests, are known to be more resistant to overfitting caused by unbalanced datasets [75, 76, 77], while decision trees and standard neural networks are particularly sensitive. If the dataset is not sufficiently sized for the manual balancing by cutting part of the mutations out of the training set, this problem can be partially addressed using cost-sensitive matrices [78], which penalize the system more strictly for misclassifying mutations that are sparsely represented in the training set. Some oversampling techniques such as SMOTE [137] or ADASYN [138] can be also utilized.

In parallel to the issue of the construction of the high-quality training data set, there arises the problem of model validation. In the ideal scenario, the validation data should also be balanced and utterly independent of the data used for training. However, due to the limited amount of experimental data, this scenario is often hard to reach. In bioinformatics, especially in the prediction of the effect of mutations on protein stability, it has become a common practise to use k-fold cross-validation as a standard method to validate the performance of the newly developed tools. This method entails randomly partitioning the original dataset into k subsets, using k - 1 subsets for the system's training, and the last random subset is left for the following validation. This process is then performed for each of the k subsets. The main argument of the utilization of cross-validation instead of splitting the data into independent training and testing datasets is that the available set of experimental data is often too small to support such a division without compromising the model's ability to identify the essential patterns and relationships. However, combining unbalanced datasets with the random aspect of k-fold cross-validation further increases the risk of overestimating the system's accuracy on the general data [79]. Therefore, cross-validation is often no longer accepted as a means of validation of the bioinformatics tools. This is particularly problematic in protein stability, where the construction of the sizeable, high-quality training dataset is impossible due to the lack of experimental data.

In summary, machine learning is a powerful approach that allows for detecting the previously unknown dependencies and interactions in the protein molecules. However, the utilization of the machine learning approaches in the predictions of the protein stability currently suffers from the overestimation of the accuracy of the existing machine learning-based tools due to the usage of the k-fold cross-validation as the method for their validation. This disadvantage is partially mitigated by using less vulnerable methods, such as random forests, and the cost-sensitive matrices. However, the issue of the availability of high-quality experimental data still stands, as described in **chapter 4**. Finally, a possible solution for this issue is presented in the **appendix C**, describing the novel protein stability database FireProt$^{\mathrm{DB}}$.

3.4 Meta-predictors and principles of the methods based on hybrid approach

Methods based on the hybrid approaches cannot be considered a singular tool but more as a combination of several different methods, tools, and computational strategies. Those methods are usually more robust and provide users with mostly reliable results as the hybrid methods usually incorporate both energy- and evolution-based approaches into their workflows, utilizing their strengths and mitigating their shortcomings.

The analysis of the highly conserved regions and the residues that show a high correlation with one or more other residues in the MSA (correlated residues are usually changing together during evolution) is a starting point for most of the hybrid methods [36, 23, 63]. This comes from the presumption that the conserved or highly correlated residues are somehow crucial for the correct function of the target protein, and therefore mutations designed on those positions would be at high risk of damaging some of the characteristics of the proteins. Conserved regions are often clustered around active sites, while the evolutionary correlation of two or more residues suggests an important intramolecular interaction. For this reason, hybrid approaches often exclude those positions from further calculation, making the mutational space safer and, at the same time, reducing the computational demands. Furthermore, it was previously proven that evolution-based and force-field methods are complementary in many proteins as there is only a partial overlap of the stabilizing mutations designed by force-fields and evolution [62]. This complementarity might be in part caused by the inability of the energy-based methods to correctly classify the charge changing mutations due to their weak implementation in the current force-fields and by the inability to estimate the effect of mutation on the unfolded state of the protein. As a result, hybrid methods are able to identify potentially stabilizing mutations that would be omitted by using only energy- or evolution-based approaches.

Due to the higher complexity and robustness of the hybrid methods, these methods are often viable not only for predicting the effect of single-point mutations but also for significantly more stable multiple-point mutants. In general, multiple-point mutants are unattainable by the tools based on a singular approach, as there is a high risk of undesired antagonistic effects. However, this issue is tackled in hybrid methods such as PROSS [36] and our novel FireProt strategy [23].

Finally, meta-predictors are the special subset of the hybrid methods that combine the results of several different tools into one consensual prediction using the simple majority voting or utilizing some form of weights. Those predictors are usually more accurate than their components. However, they lack the complexity and robustness of the real hybrid workflows. The problematics of the hybrid methods and the current state-of-the-art will be further discussed in the **chapter 5**.

4 Data sets for protein stability

The accuracy and reliability of the computational methods, especially those utilizing one of the machine learning techniques, depends strongly on the size, structure, and quality of the chosen training and validation datasets. Up to this date, the primary source of validation data for protein stability engineering is the ProTherm database [47]. ProTherm is the most extensive freely available database of thermodynamic parameters such as $\Delta\Delta G$, ΔT_m, and ΔC_p. It currently contains about 26,000 entries representing both single- and multiple-point mutants of 740 unique proteins. Although ProTherm is the most common stability data source, it suffers from high redundancy, missing data, and serious inconsistencies. Particularly troubling are differences in the pH values at which the thermodynamic parameters were determined, redundant entries with non-agreeing data, and strikingly even disagreements about the signs of $\Delta\Delta G$ values. ProTherm also neglects intermediate states' existence, and therefore some data might be considering only one part of the folding pathway [72]. To overcome those problems, the data must be filtered and manually corrected to construct reliable training and validation datasets. More detailed statistics of the ProTherm database are noted in Table 4.1.

Tab. 4.1: Statistics of the ProTherm database.

Number of entries	25,820
Number of proteins	1,045
Unique proteins	740
Proteins with mutants	311
Single-point mutations	12,561
Double-point mutations	1,744
Multiple-point mutations	1,132
Wild-type	10,383

Several subsets of the ProTherm database have been derived and utilized to train and validate a vast scale of prediction tools. One of the most popular is the freely available PopMuSiC dataset [101], which contains 2,648 mutations extracted directly from the ProTherm database. However, this dataset is unbalanced as only 568 of its mutations are classified as stabilizing, while 2,080 are classified as destabilizing. Furthermore, 755 of its 2,648 mutations have observed $\Delta\Delta G$ values ranging in the interval from -0.5 to 0.5. Mutations with such inconclusive $\Delta\Delta G$ cannot be considered either stabilizing or destabilizing because the average experimental error in $\Delta\Delta G$ measurements was established to be 0.48 kcal/mol [102]. Additionally, the data extracted from ProTherm are insufficiently diverse: around 20% of the Pop-

MuSiC dataset comes from a single protein, and ten proteins (out of 131 represented in the dataset) account for half of the available data. Further inspection of the data reveals that mutations to more hydrophobic residues located on the protein surface tend to be stabilizing, whereas mutations that increase the hydrophilicity in the protein core are usually destabilizing. Consequently, most computational tools are likely to identify mutations that increase surface hydrophobicity as stabilizing even though such designs often fail to fold properly because of poor protein solubility [37].

Some predictive tools use alternative data sets derived from ProTherm or PopMuSiC for training and validation. The most common benchmarking data set utilized for independent validation is S350 [101], which contains 90 stabilizing and 260 destabilizing mutations in 67 unique proteins. However, this data set is still unbalanced and relatively small for comprehensive evaluation. The recently published PopMuSiC$^{\text{sym}}$ dataset [103] tries to address the issue of unbalanced data by including 342 mutations inserted into the mutant proteins. A comparative study conducted using this dataset showed a bias of the existing tools toward destabilizing mutations, as they performed significantly worse on the set of inverse mutations. Because of the overlaps of the mutations in training and validation datasets, the individual tools' results can be overestimated. Even the new derivatives of the ProTherm database do not solve the problems arising from the available data size and structure. Therefore, there is an urgent need for new experimental data, particularly on the side of stabilizing mutations. Moreover, it would be of immense help for the future development of predictive tools to proceed with the standardization of the stability data, e.g., a unified definition of $\Delta\Delta G$ as a subtraction of the ΔG values for the mutant and the wild-type as mentioned in **chapter 2**.

Until the new unbiased datasets arise, measuring the predictive tools' accuracy based only on the amount of correctly classified mutations is deemed insufficient. Instead, the Matthews correlation coefficient (MCC) can be utilized for binary classification, as it was designed as a balanced measure that is usable even for datasets with a significant difference in the sizes of individual classes [77]. Similarly, when binary predictions are utilized as a filtration step in the hybrid approaches, metrics like sensitivity, specificity, and precision might be useful. When numerical measures are considered, the linear correlation between the predicted and experimental values can be estimated using the Pearson correlation coefficient (PCC) and the average error established as the root-mean-square error (RMSE). Finally, the bias of the computational tools can be estimated as the sum of $\Delta\Delta G$ for the direct and inverse mutations, according to Thiltgen and Goldstein [65]. Critical evaluation of the existing tools using the S350 dataset revealed that the PCC ranges from 0.29 to 0.81, with an average RMSE of about 1.3 kcal/mol.

One of the main aims of this Thesis is to suggest a standardized method for storing and further managing the protein stability data. FireProt$^{\mathrm{DB}}$ is a novel database providing users with free access to available experimentally validated data while resolving the issues accompanying the original ProTherm database. All the included data are checked for their correctness, and the database is completed with an interactive search engine and annotations obtained from Uniprot and other databases. FireProt$^{\mathrm{DB}}$ also contains structural information obtained from the HotSpotWizard calculations (**appendix D**). The main goal of the database is to provide community with the clean reliable set of experimental data that would allow for a better validation of the existing tools same as for the continuous development of the predictive tools utilizing the machine learning techniques. Further information about FireProt$^{\mathrm{DB}}$ is provided in the **appendix C**.

5 Available computational tools

In recent years, a plethora of tools for predicting the effect of mutations on protein stability has been developed. Most of those tools belong to one of the four categories: i) tools based on the free energy calculations, ii) tools utilizing information obtained from proteins evolution, iii) tools employing mathematical statistics and machine learning techniques, and iv) meta-predictors and hybrid methods, combining several different tools or strategies into one robust workflow. The general concept of those approaches was previously described in **chapter 3**. This chapter will compare some of the well-known computational tools, their strengths and weaknesses, and recommended applications. Tables 5.1, 5.2, 5.3, 5.4, included at the end of each of the sections, provide a list of the selected software tools for each approach.

5.1 Software tools based on the energy calculations

Software tools utilizing force-fields as the means for the prediction of the effect of mutations are based on either molecular modelling of the physical interactions between the atoms in the tertiary structure of the protein of interest (PEEFs), using methods of mathematical statistics (SEEFs), or the combination of both (EEEFs). The border between the three is not exactly sharp as even tools that are considered to be PEEFs are utilizing some statistical approximations, while some SEEFs can also use a small number of physical descriptors. Therefore, the division of the force-field methods into the three categories is very obscure, with many overlaps and inconsistencies. Furthermore, EEEF is not always recognized in the published literature. As a result, the division of the tools suggested in this Thesis should be taken with the grain of salt.

In most of the published works, the Rosetta suite [59] is considered to be a state-of-the-art for predicting the effect of mutations on protein stability. It is one of the most versatile software packages for macromolecular modelling and related tasks. It consists of several modules, including stability predictions, molecular simulations, and ab-initio modelling. Two of the modules are applicable for *in silico* design of the stable proteins. Rosetta Design is a more general module for protein design engineering that is able to reflect their predicted stability in physically detached Rosetta energy units. Those are automatically converted into well-interpretable $\Delta\Delta G$ values in the newest versions of the Rosetta suite. Secondly, ddg_monomer is a stand-alone module built on top of the Rosetta Design that was parametrized specifically for predicting $\Delta\Delta G$ values and protein stability [54]. Finally, the Rosetta suite is also supplemented with a wide range of force-fields and protocols, allowing users to adjust their calculations based on the protein of interest and available computational resources.

Several other computational tools can be assigned to the PEEFs category. The ERIS software [104] utilizes its own Medusa force-field that incorporates a side-chain packing algorithm and backbone relaxation method. Similarly, the Concoord/Poisson-Boltzmann surface area method [105] (CC/PBSA) employs the Concoord program [107] to rapidly generate alternative protein conformations to sample available conformation space and the energy function calculated by GROMACS force-field [106] is then averaged over the generated structural ensemble.

Some tools simply fit the force-field equations using the values derived from the available experimental data instead of estimating individual terms' values in the equation by performing calculations based on the Newtonian physics. One of the most popular representatives of the SEEFs category is the PopMuSiC method. In PopMuSiC [45], the force-field equation is constructed using thirteen physical and biochemical terms with approximate values derived from databases of known protein structures. A similar approach can be found in other statistical tools such as DMutant [108] or HotMuSiC [109]. HotMuSiC differs from its predecessor (PopMuSiC) mainly because it was parametrized, especially for estimating ΔT_m instead of $\Delta\Delta G$ as the correlation between the two is only -0.71 [109]. In HotMuSiC, the previous force-field equation was expanded using five temperature-dependent potentials based exclusively on the data extracted from mezostable and thermostable proteins.

Finally, some of the tools balance their prediction accuracy with the time demands by using both physically calculated and statistically derived terms in their force-field equations. CUPSAT [134] is employing the atom and torsion angles potentials derived from the tertiary structures obtained from PISCES [135]. However, Boltzmann's energy values are then predicted from the radial pair distribution of amino acid atoms, and the Gaussian apodization function is applied to assign favourable energy values for the neighbouring orientations of the observed torsion angles combinations. FoldX suite [49] can also be included in the EEEFs category.

While PEEFs provide more reliable results in general, in the majority of cases, SEEFs still perform reasonably well compared with most machine learning methods and are orders of magnitude faster than PEEFs. Therefore, SEEFs and EEEFs seem to be an acceptable trade-off between predictive power and computational demands, primarily when utilized as filters to prioritize the mutations in hybrid workflows.

5.2 Software tools based on the phylogenetics

The main advantage of phylogeny-based methods is that they do not require high-resolution protein structure and, therefore, can be applied to any protein with

Tab. 5.1: Software tools for the prediction of the effect of mutations on protein stability utilizing force-field based approach.

Method	**Model**	**Input**	**Output**	**Mutations**
PoPMuSiC [45]	SEEF	Structure	$\Delta\Delta G$	Sigle
FoldX [49]	EEEF	Structure	$\Delta\Delta G$	Sigle
CUPSAT [134]	Atom potentials Torsion angles	Structure	$\Delta\Delta G$	Sigle
Rosetta [59]	PEEF	Structure	$\Delta\Delta G$	Sigle/multiple
ERIS [104]	PEEF	Structure	$\Delta\Delta G$	Sigle
CC/PBSA [105]	PEEF	Structure	$\Delta\Delta G$	Sigle
DMutant [108]	Amino acid potentials Torsion angles	Structure	$\Delta\Delta G$	Sigle
SDM [142]	SEEF	Structure	$\Delta\Delta G$	Sigle
HotMuSiC [109]	SEEF	Structure	ΔT_m	Sigle
STRUM [144]	SEEF	Structure	$\Delta\Delta G$	Sigle
AUTO-MUTE [143]	SEEF/ML	Structure	Binary/$\Delta\Delta G$	Sigle

known amino acid sequence and the sufficient amount of sequence homologs in the databases. Although phylogeny-based methods are well-established in protein thermostability engineering, the influence of individual mutations suggested by the evolution is hard to quantify, and not all mutations will move the protein characteristics in a desirable way. Only about 50% of mutations identified by evolution-based approaches are truly stabilizing, but many of them will rather positively affect protein solubility or activity [86]. Phylogeny-based methods, especially consensus design, are therefore mainly utilized as filters during core calculations of hybrid workflows or as components of predictive tools for hotspot identification. Consensus design is available in several bioinformatics packages such as 3DM [30], VectorNTI [124], EMBOSS [123], and HotSpotWizard [125]. No stand-alone tools are available as the implementation of consensus design as it is extremely simple. On the other hand, there are many tools dealing with the problematics of ASR using either maximum-likelihood (FastML [127], RAxML [126], and Ancestors [128]) or Bayesian inference (HandAlign [129] and MrBayes [130]) methods. Both of those groups of methods are well-established in the scientific community.

A significant limitation of those methods is that most of the tools require users to upload their own MSA and phylogenetic tree. Constructing these input data is the most crucial and demanding step of the entire process as the ASR is strongly dependent on the initial set of homolog sequences, their alignment, and the topology of the resulting phylogenetic tree. To obtain reliable predictions, the initial set of homolog sequences must be manually curated to identify a reasonably-sized subset of biologically relevant sequences. The initial set of sequences if usually obtained using word search approaches such as BLAST [80], profile-based search methods such as position-specific iterative BLAST [81], or hidden Markov models utilized in HMMER [82, 131]. Simple sequence identity is not the best measure of the homologs' relevance as the sequences selected for the ASR analysis should be sufficiently diverse, while maintaining the same or at least similar biological function. Furthermore, the MSA and the tree's topology requires to be manually inspected, and the rooting of the tree is usually done by selecting the appropriate outgroup (sequence or a group of sequences that are the most distant from the other sequences in the set). This is uneasy to do in the automatized manner in eukaryotic organisms and close to impossible in the prokaryotic proteins due to the high occurrence of the horizontal gene transfers. Lastly, the reconstruction of the ancestral gaps is completely omitted in most of the available tools. For all those reasons, the most laborious part of the ASR analysis is left in users' hands and requires them to attain a deep understanding of the bioinformatics tools and the biological knowledge of the system of interest. This Thesis attempts to address these issues, providing users with a fully automated workflow. FireProt$^{\mathrm{ASR}}$ handles all parts of the ancestral reconstruction, including the search for the homolog sequences, dataset reduction, construction of the rooted phylogenetic tree, and the reconstruction of the ancestral sequences together with the identification of the ancestral gaps. The problematic parts of the calculation were resolved by utilizing some novel techniques: the homolog search was improved by using filters checking for the similarity in the protein function, the rooting of the phylogenetic tree is done via the recent minimal ancestral deviation algorithm [161] and the novel algorithm for the ancestral gaps reconstruction was designed to replace the most commonly used Fitch's algorithm [162]. Further information about FireProt$^{\mathrm{ASR}}$ workflow is provided in the **appendix B**.

5.3 Software tools based on the machine learning

Predictive tools based on machine learning techniques are very common as the machine learning approach does not require comprehensive knowledge of the physical and biochemical forces acting within proteins' tertiary structure. Predictions are therefore based exclusively on the available experimental data. The most popular machine learning tools are utilizing support vector machines (e.g., I-Mutant [110], EASE-MM [72], MuStab [73], and MuPro [111]) or random forests (e.g., PROTS-RF [113] and ProMaya [112]), which are known to be comparatively more resistant to overtraining when using unbalanced training datasets. Due to the nature of the available protein thermostability data (see **chapter 4**), neural networks are very

Tab. 5.2: Software tools for the prediction of the effect of mutations on protein stability utilizing evolutionary information.

Method	**Model**	**Input**	**Output**	**Mutations**
HotSpotWizard [125]	CA	Seq/struct	hotspots	Single/multiple
FastML [127]	ML	MSA+tree	Sequences	Single
RAxML [98]	ML	MSA	Phylogeny	Single
MLGO [145]	ML	MSA+tree	Seq+phylogeny	Single
Ancestors [128]	ML	MSA+tree	Seq+PP	Single
HandAlign [129]	BA	MSA+tree	Seq+PP+phylogeny	Single
TreeTime [146]	BA	MSA+tree	Seq+PP+phylogeny	Single
PAML [97]	ML	MSA+tree	Seq+PP+phylogeny	Single
PhyloBot [147]	ML	MSA	Seq+PP+phylogeny	Single
MaxAlike [148]	ML	MSA+tree	Seq+PP+seq. logo	Single

sparsely employed for engineering protein stability as they are highly sensitive to the quality and size of the training data.

In the past years, some of the more recent approaches, such as deep learning, have been applied to the diverse problems in genome and protein engineering. Deep learning was successfully utilized to predict the effect of mutations on human health (e.g., DANN [114]) and predict protein secondary structures (e.g., SSREDNs [115]). However, the applicability of deep learning for protein stability prediction is still very limited as it suffers from the shortcomings of the standard neural networks. Deep learning is prone to overfitting because of the added layers of abstraction that increase the network ability to model rare dependencies, thus resulting in a loss of generality. This issue can be partially addressed by using regularization methods such as Ivakhnenko's unit prunning [116, 117]. However, this does not entirely negate problems arising from the insufficient size of training datasets as deep learning has very stringent requirements on the training data. As a result, deep learning is still very sparsely applied to predict protein stability (e.g., TopologyNet [118]).

The reliability and robustness of the computational tools based on machine learning can be improved by combining several different models into a single multi-agent system. This approach was utilized in tools such as ELASPIC [136] and MAESTRO [119]. In MAESTRO, traditional neural networks are combined with support vector machines, multiple linear regression, and limited statistical potentials. The outputs of the individual methods are then averaged into a single consensual prediction. Furthermore, in such tools, machine learning can also be applied to train the arbiter to decide how to combine the outputs of the individual methods and take their weights, balance, and the relative strengths of each method under consideration based on the type of the predicted mutation.

It is difficult to compare individual computational tools based on the results presented in the primary sources as most of them were evaluated by authors using different validation datasets. Thus, the tool's performance is usually biased toward particular proteins and mutation types, causing its reported accuracy to be overestimated. Therefore, there is a need for independent comparative studies. The critical evaluations reported by Potapov et al. [48], Kellog et al. [59], and Khan and Vihinen [120] revealed that methods based on PEEF calculations systematically outperform tools utilizing solely the machine learning techniques or statistical potentials when tested on the independent dataset. Furthermore, it was shown by Pucci et al. [103] and Usmanova et al. [121] that the tools based on the machine learning methods tend to be more biased, and their reported accuracies are significantly overestimated. Finally, Montanucci et al. [122] estimated that the upper bound of the Pearson's correlation coefficient is about 0.8, and the lower bound of the RMSE is 1 kcal/mol for the most commonly used protein stability datasets. Those findings are not in agreement with many of the results reported by the original publications. The upper bound is also deemed to change in the future with the increase in the available experimental data's size and diversity.

Tab. 5.3: Software tools for the prediction of the effect of mutations on protein stability utilizing machine learning.

Method	**Model**	**Input**	**Output**	**Mutations**
EASE-MM [72]	SVM	Sequence	$\Delta\Delta G$	Single
MuStab [73]	SVM	Sequence	Binary	Single
ProMaya [112]	RF	Sequence	$\Delta\Delta G$	Single
mCSM [149]	Graph based	Sequence	$\Delta\Delta G$	Single
ELASPIC [136]	SVM+HMM	Structure	$\Delta\Delta G$	Single/multiple
MuPro [111]	SVM	Seq./Struct.	$\Delta\Delta G$	Single
I-Mutant2.0 [110]	SVM	Seq./Struct.	$\Delta\Delta G$	Single
TopologyNet [118]	Deep learning	Structure	$\Delta\Delta G$	Single
PROTS-RF [113]	RF	Structure	$\Delta\Delta G$	Single
MAESTRO [119]	Multi-agent system	Structure	$\Delta\Delta G$	Single/multiple
IPTREE-STAB [74]	Decision tree	Sequence	Binary	Single
INPS-MD [150]	Sup. vec. regression	Sequence	$\Delta\Delta G$	Single
iStable [151]	SVM	Structure	$\Delta\Delta G$	Single
Prethermut [152]	SVM+RF	Structure	$\Delta\Delta G$	Single/multiple

5.4 Software tools based on hybrid approaches

Hybrid methods make predictions by combining information obtained from several fundamentally different approaches. As a result, they offer improved robustness and reliability compared to individual computational tools. This allows for the multiple-point mutants to be designed with the minimal risk of combining mutations with potentially antagonistic effects. Consequently, the hybrid approaches are of interest for many research groups; however, only three tools are currently available to the scientific community.

The Framework for Rapid Enzyme Stabilization by Computational Libraries [63] (FRESCO) was the first of the hybrid methods available to the users as a set of individual tools and scripts. Therefore, its usage requires advanced knowledge of the bioinformatics tools. FRESCO initially selects a pool of potentially stabilizing mutations based on the predictions obtained from FoldX and Rosetta and filters out all residues located in close proximity to the active sites. Disulfide bridges are designed by dynamic disulfide discovery using snapshots from MD simulations and subsequently evaluated utilizing the set of geometric criteria. Furthermore, short MD simulations predict changes in backbone flexibility upon mutation to remove faulty designs with unreasonable features that are expected to destabilize the protein. FRESCO approach is not fully automated as it suggests only a pool of the potentially stabilizing single-point mutations that have to be subjected to further experimental validation. This experimental validation greatly reduces the risk of false positives but requires a substantial effort from the users.

Protein Repair One-Stop-Shop [36] (PROSS) is an automated web-based protein stabilization platform. Similarly to FRESCO, PROSS workflow starts with a Rosetta design calculation in which the residues positioned closely to the protein's active and binding sites are excluded from the further analysis. A position-specific substitution matrix is analysed to govern the design process away from the amino acids that are rarely observed in the sequence homologs [132]. Rosetta's computational mutation scanning tool [133] is used to search the remaining pool of potential amino acid mutations. Finally, Rosetta's combinatorial sequence design tool is utilized to find an optimal combination of stabilizing mutations, and an energy function is applied that favours amino acid identities based on their frequency in the MSA. This hybrid approach allows for the mutants' design containing some neutral or even slightly destabilizing mutations, while taking into account potential epistatic effects [31].

FireProt [23, 35] is another fully automated web-based protein stabilization platform that combines both energy- and evolution-based approaches to design thermostable multiple-point mutants. FireProt workflow includes 16 computational tools and databases and utilizes both sequence and structure information in the process. The utilization of the evolutionary information in the process of the calculation prohibits the mutations of the potentially important residues, while also reducing its time demands. Two force-field based methods, FoldX and Rosetta are employed to increase the reliability of the predicted potentially stabilizing mutations. Finally, pair-wise calculation of the potentially stabilizing mutations using a simple graph based algorithm is utilized to reduce the risk of introducing the antagonistic effects into the designed mutant protein. Detailed description of the FireProt method can be found in the **appendix A**.

In summary, hybrid methods represent the next step in predicting protein stability as their robustness and complexity allow for the construction of significantly more stable multiple-point mutants, while maintaining reasonable computational demands. Those methods often utilize evolutionary information as filters for removing potentially deleterious mutations in the conserved and correlated regions of the proteins of interest, thus lowering the risks of antagonistic effects and further increasing the required speed. Finally, hybrid methods can be further expanded by the predictions of protein solubility or catalytic activity.

Tab. 5.4: Software tools for the prediction of the effect of mutations on protein stability based on the hybrid or other non-standard methodology.

Method	**Model**	**Input**	**Output**	**Mutations**
		Hybrid methods		
FireProt [23, 35]	Evolution+energy	Structure	Mutations+$\Delta\Delta G$	Multiple
PROSS [36]	Evolution+energy	Structure	Mutations	Multiple
FRESCO [63]	Evolution+energy	Structure	Mutations	Multiple
		Other methods		
pStab [139]	Equilibrium thermodynamics	Structure	Unfolding curves	Charges
Encom [140]	Normal mode analysis	Structure	$\Delta\Delta G$	Single
Neemo [141]	Residue interaction networks	Structure	$\Delta\Delta G$	Single

6 Ancestral sequence reconstruction

ASR was already established as one of the two main phylogeny-based methods that are often utilized for the task of protein thermostability engineering in the **chapter 3**. Furthermore, it will be tackled again in the results part of this Thesis as one of the included original papers, FireProt^{ASR}, describing a computational tool built on top of the ASR strategy. However, this article does not include any information about practices utilized in the process of ancestral reconstruction as they are well-known to the scientific community interested in evolutionary biology and the related fields. This chapter aims to provide the basic theoretical knowledge required for the correct usage of ASR and interpretation of the results. More specifically, the algorithms used for constructing the phylogenetic tree will be described together with the means for finding the root of the newly constructed tree and the ancestral sequence reconstruction itself. Possible solutions for the selection of the biologically relevant subset of homolog sequences will not be described as they overstep the main scope of this Thesis and up to this date, the biologically relevant subsets of sequences are mostly produced by laborious manual work requiring in-depth knowledge of evolutionary biology and the biological systems of interest.

6.1 Construction of the phylogenetic tree

The construction of the phylogenetic tree starts with the sequence alignment of the known homolog sequences. From the MSA, it is possible to identify conserved regions the same as the evolutionary patterns that occurred during the course of protein development (correlated mutations, the similarity of the conserved regions, and speciation). In general, there are three basic approaches on how to construct a phylogenetic tree: i) methods based on distances, ii) methods based on characters, and iii) methods based on probabilities.

6.1.1 Methods based on the distances

These methods are built on top of the idea that the distance of the homolog sequences in the evolution is equal to the number of mutations that have to be included to transform from one homolog to another. The primary representative of this method is Neighbour-joining (NJ) [153] that is based on the additive distance matrix. This square matrix captures the number of differences between individual sequences and can be easily constructed from the MSA.

The algorithm starts with constructing the phylogenetic tree with the star-like topology, where all homolog sequences are connected into one central node. In the second step, the search algorithm identifies two terminal nodes with the closest distance to each other that are also the most distant from the other sequences in the initial tree. This selection is represented by the function:

$$min[D(a,b) - u(a) - u(b)],$$

where $D(a,b)$ is the distance of the nodes A and B; $u(a)$ and $u(b)$ represent the average distance of the nodes A and B to the remaining nodes, respectively. In the following step, nodes A and B are connected into a newly formed node U and the evolutionary distances between the nodes A, B and U are estimated as:

$$L(A,U) = \frac{[D(a,b) + u(a) - u(b)]}{2}; L(B,U) = \frac{[D(a,b) + u(b) - u(a)]}{2}$$

Finally, the new distance matrix is established by replacing nodes A and B with delegated node U and the distance between all remaining nodes and newly formed node U is calculated as:

$$D(c,u) = [D(a,c) + D(b,c) - D(a,b)]/2$$

The whole process is then repeated until the whole tree is constructed. NJ method is swift and provides reliable results for all the trees with the additive distance matrices. However, it can be successfully utilized even without satisfying this condition. The alternative for the NJ is Unweighted Paired Group Method with Arithmetic mean (UPGMA) [154]. The distance methods, more specifically NJ approach is currently utilized in the FastTree algorithm [155].

6.1.2 Methods based on characters

Methods based on characters try to omit the need to derive the distance matrix by constructing the phylogenetic tree directly from the MSA, thus preserving all the information contained in the sequence alignment. The method is split into two parts – big and small parsimony problem. Small parsimony problem evaluates the constructed tree with its parsimony score, while big parsimony problem searches for the best possible topology.

The most basic solution for the small parsimony problem is the Sankoff algorithm [156], which requires the scoring matrix to assign values for all possible character substitutions. Each terminal and non-terminal node is represented by the vector of the length corresponding to the number of characters in the approved alphabet (four for the nucleotides, twenty for the set of standard amino acids). In this vector, the character corresponding to the real character on the given position of the terminal sequence is tagged with zero, and the scores of the remaining characters are set to infinite. The algorithm then moves from the terminals to the root of the tree, and the score for each field St of the non-terminal vector is calculated as:

$$S_t(parent) = min_i\{S_i(left) + \delta_{it}\} + min_j\{S_j(right) + \delta_{jt}\},$$

where S_i and S_j represent the current value of the given character in the vector of the child node and δ_{it}, and δ_{jt} shows the price of the substitution between the child and parental node. This represents the minimal score of the parsimony of the field S_t, considering the values in the vector of its child nodes. Once the algorithm reaches the root of the tree, it moves back from the root to the terminal nodes, assigning the characters to the internal nodes based on the parsimony score of their parent.

Big parsimony problem tries to find the best topology of the phylogenetic tree that would provide the lowest parsimony score in the vector of its root. In general, this is an NP-hard problem, and therefore some heuristics such as Nearest Neighbour interchange and tree cutting and re-grafting are required [157]. Furthermore, the character-based methods do not consider the different lengths of the branches and the molecular clock. For this reasons, no actively used tools are currently utilizing character-based methods.

6.1.3 Methods based on probability

Probabilistic methods try to establish a model that would be the most likely representation of the provided data. This approach's main representative is a maximum-likelihood method, which can be further divided into tiny, small, and big likelihood problem. Tiny likelihood problem evaluates the tree's internal nodes and estimates a total likelihood for a given tree. The topology of the tree and the distances of the branches have to be already assigned. If those conditions are met, the Felsenstein algorithm can be applied.

Felsenstein algorithm [158] is based on dynamic programming. It starts by assigning a vector to all terminal and non-terminal nodes. In the terminal node, the individual fields' value is set to zero, if the character in the field does not correspond to the character in the sequence alignment and one if otherwise. The values of the internal nodes are calculated as:

$$L_{S_k}(k) = [\sum_{S_i} P_{S_k S_i}(t_i) * L_{S_i}(i)] * [\sum_{S_j} P_{S_k S_j}(t_j) * L_{S_j}(j)]$$

where $P_{S_k S_i}$ represents the probability of the given substitution between nodes S_k and S_i, with the evolutionary distance between those two nodes being t_i and

$L_{S_i}(i)$ contains a current probability for a given character in the node S_i. The value of the root of the phylogenetic tree is then calculated as:

$$L = \sum_{S_0} P_{S_0} * L_{S_0}(0)$$

In the second step, the algorithm proceeds from the tree's root to its terminal nodes and assigns the internal nodes' values based on its maximum-likelihood. The tiny likelihood problem serves only for the evaluation of the already existing trees. The small likelihood problem estimates the branch lengths of the tree with the maximum likelihood of L.

A small likelihood problem utilizes the hill-climbing algorithm to increase the maximum likelihood L of the given tree. In the beginning, the initial branch lengths are assigned at random, and in the following steps, those values are adjusted iteratively until the algorithm climbs to its maximum.

The big likelihood problem represents the last step in this probabilistic approach as it tries to identify the topology of the tree itself. This is done by gradually adding new branches into the phylogenetic tree, and the small likelihood problem is repeatedly calculated after each iteration. After several new branches are added into the tree, the tree is cut and re-grafted. For the tree with n terminal nodes, the total amount of $2n^2 - 9n + 8$ different tree topologies is evaluated. Unlike maximum parsimony, maximum likelihood approach considers different lengths of the branches for a cost of very high computational demands. Probabilistic methods are the most common and are utilized in most of the existing tools, such as RAxML [98], PAML [97], or MaxAlike [148].

6.2 Rooting of the phylogenetic trees

Except for the UPGMA algorithm, majority of the previously described algorithms produce the phylogenetic tree in its unrooted form. This contradicts the general idea of the phylogenetics and the ancestral sequence reconstruction as they presume the existence of the common ancestor of all the homolog sequences in the phylogenetic tree. Therefore, it is required to root the phylogenetic tree before using it for the ancestral sequence reconstruction. In this section, three different methods of rooting will be described.

6.2.1 Outgroup rooting

The outgroup rooting [159] is the most instinctive approach in evolutionary biology and is utilized in most phylogeny-based applications. Unlike the other two methods, outgroup rooting requires existing knowledge of the system of interest to place the root in the tree's correct position. This is done by manual selection of a so-called outgroup – a sequence or a small group of sequences known to be more distantly related than all the other sequences in the phylogenetic tree. The root of the tree is then placed between the outgroup and the rest of the tree. This method is more resistant in the cases where there are different evolutionary rates between individual species in the tree (i.e., rodent lineage is evolving faster than humans). However, automation of this approach is close to impossible as it relies on the expert knowledge introduced into the calculation. Furthermore, the outgroup selection is more straightforward in the trees containing only eukaryotic organisms as the topology of the tree well-reflects the natural pace of the evolution. However, in the prokaryotes, the situation is much less conclusive due to the frequent occurrence of the horizontal gene transfers in bacterial life.

6.2.2 Midpoint rooting

Unlike outgroup rooting, midpoint does not require any expert knowledge as it attempts to root the tree in its middle point [160]. This is done by calculating the distances between all terminal nodes' pairs and selecting the longest one. The root is then placed exactly half-way between these two terminal nodes. The midpoint rooting is a viable strategy for the trees with a constant evolutionary rate. However, it can easily misbehave if the evolutionary rates in the tree are not reasonably balanced. Therefore, the usage of the outgroup rooting is preferable to the midpoint, but it can still be a viable option for more closely related trees or if the outgroup cannot be established.

6.2.3 Minimal ancestor deviation

The midpoint algorithm stands strong only under the assumption of a strict molecular clock. However, this assumption is false for most of the evolutionary trees. In practical applications, midpoint deviates from the actual position of the ancestral root node. Minimal ancestor deviation algorithm [161] tries to evaluate the midpoint criterion's deviations for all possible root positions and all pairs of terminal nodes in the unrooted tree. The algorithm considers each branch as a possible root position. The pairwise relative deviation is defined as:

$$r_{bc,a} = [\frac{2d_{ab}}{d_{bc}} - 1] = [\frac{2d_{ac}}{d_{bc}} - 1],$$

where d_{ab} is a distance between nodes a and b. For two terminal nodes b and c and ancestral node a, the distances to the ancestor are d_{ab} and d_{ac}. Based on the midpoint criterion, both should be equal to $\frac{d_{bc}}{2}$. Branch ancestor deviation for a putative root in a branch a, b connecting adjacent nodes a and b of the unrooted tree is then defined as the root-mean-square of the pairwise relative deviations. In general terms, a minimal ancestor deviation algorithm tries to place the ancestral root into the position where there is the lowest deviation of the ancestral root from the midpoint of the given branch for all the branches in the unrooted tree.

As in the case of the midpoint rooting, minimal ancestor deviation is a mathematical approach and therefore does not require any knowledge of the system of interest. However, unlike midpoint, it can also be utilized for the trees without a strict molecular clock. Finally, it was proven that the accuracy of minimal ancestor deviation is comparable to the outgroup rooting in eukaryotic systems and is superior for prokaryotic organisms where outgroup rooting is hard to establish due to the occurrence of horizontal gene transfers [161].

6.3 Inference of the sequence ancestors

In general, there are three main algorithms usable for the ancestral sequence reconstruction: i) maximum parsimony, ii) maximum-likelihood, and iii) Bayesian inference. The ancestral inference process with the use of the maximum parsimony and maximum-likelihood approaches was previously described in the **section 6.1** as the inference of the ancestral characters is a part of the small parsimony and tiny likelihood problems, respectively. Bayesian inference is a probabilistic approach, similar to maximum-likelihood, that combines the tree's prior probability with the likelihood data to produce posterior probability distribution on the given trees. Bayesian inference utilizes the Markov Chain Monte Carlo method [163], which can be described in three steps. At first, the stochastic algorithm proposes a new state for the Markov chain. Next, the probability of this state to be correct is calculated. Finally, a new random variable from the interval $(0, 1)$ is proposed. If this new value is lower than the acceptance probability, the new state is accepted, and the Markov chain updated accordingly. The whole process is then repeated.

Each method of ancestral sequence reconstruction has its advantages and shortcomings. Maximum parsimony is a straightforward approach that provides a simple interpretation for a given set of homolog sequences as the ancestral states are reconstructed to include as few changes across the sequences as possible. This is exhibited by the smallest number of the evolutionary steps that have to be carried out to explain the data. The simplicity of this method allowed its usage with limited computational resources in the past; however, it has been overshadowed by statistically more consistent probabilistic approaches in recent years. Its obsolescence was also pushed forward by its inability to consider the branches' lengths in the phylogenetic tree, which yields erroneous observations accumulated with the addition of more sequences into the tree [164].

Maximum-likelihood is currently the most dominant approach in evolutionary biology as it is able to calculate the length of the branches in the phylogenetic tree. Furthermore, it also considers the probability of each tree explaining the given homolog sequences based on the suggested model of evolution. This means that the substitution rates for amino acids and nucleotides are taken into account, leading to more realistic evolutionary relationships. However, maximum-likelihood is computationally very expensive, and to explore all the possible trees comprehensively is out of reach for bigger sets of homolog sequences. Finally, same as the maximum parsimony, the maximum-likelihood is unable to account for the phylogenetic uncertainty in the prediction of the ancestral gaps.

Compared to the previous approaches, Bayesian inference is able to incorporate complex models of evolution, and it quantifies the uncertainties in the data. It has also been recommended as a possible solution for the bias of probabilities in more distant ancestors as they have been systematically overestimated by the maximum-likelihood methods [95, 165]. However, Bayesian inference tends to compute ancestral sequences with considerably lower posterior probabilities, which sometimes leads to the loss of the ancestors' biological relevance [100]. On the other hand, it is more computationally effective than MP and ML methods.

In conclusion, there is no optimal method of ancestral reconstruction as each of them comes with their shortcomings. However, the maximum parsimony method is losing its relevance with the continuous growth of the computational resources.

6.4 Reconstruction of ancestral gaps

Reconstruction of gaps in ancestral sequences is one of the most crucial issues in the process of ancestral sequence reconstruction as the insertions and deletions cannot be treated in the same way as the standard characters during the inference of the ancestral states. This problem was not yet resolved in a robust way, and the ancestral gaps are usually included in the ancestral sequences by laborious manual curation. So far, only a few algorithms were suggested to deal with this issue in an automated manner. The most common is the algorithm based on Fitch's parsimony [162]. This approach is composed of two steps. In the beginning, the algorithm assigns vectors of the length of the sequences in the MSA to all the terminal and internal nodes. Fields in the vector of the terminal nodes are then filled with either 0 or 1, which signifies the gap in the given position in the MSA of the sequence corresponding to the terminal node. The algorithm then moves from the terminals to the root of the tree, and the fields in the vectors of the internal nodes get their values assigned by the following rules:

- If the value of the field on the corresponding position is 0 for both its left and right children, the result is 0.
- If the value of the field on the corresponding position is 1 for both its left and right children, the result is 1.
- If the value of the field on the corresponding position is 0 for left child and 1 for right child or otherwise, the result is X.
- If the value of the field on the corresponding position is X for one of its children, but not for the other, this second value is adapted.
- If the value of the field on the corresponding position is X for both children, X is assigned also to the parent.

Once the algorithm reaches the root of the tree, the values in the root vector are updated in the way that there is no uncertainty in any position. All the fields with the value X are compared with their posterior probabilities. If some of the characters on this position have higher posterior probability than the specified threshold, this field is set to 0, signifying character in this position. In other cases, the field is tagged as a position with the gap. In the second step, the algorithm moves from the root to the terminal nodes, and any uncertainties are removed by replacing them with the value from the vector of their parental node.

There are several issues connected with the ternary nature and its negligence of the evolutionary distances. As a result, lonely branches and smaller subtrees have a similar impact on the final decision as the well-resolved branches. This is especially notable when the decision is influenced by one short branch connected to the vast levelled subtree. A possible solution for this issue is provided in the form of the algorithm described in the **appendix B**, describing FireProt$^{\text{ASR}}$ method. This method not only considers the length of the branches but also tracks the evolutionary distances during the course of the evolution.

7 Research summary

This chapter summarizes the research that was conducted in connection with the main topic of this thesis, i.e. the development of the *in silico* tools that can be employed to design stable protein structures. Four original publications describing three tools and one database: FireProt, FireProt^{ASR}, FireProt^{DB}, and HotSpotWizard 2.0 are included. In this chapter, abstracts and contributions for each of the individual publications are presented, while the full versions of the forementioned publications can be found in the appendix. A brief list of the research published by the author that is not mentioned in this thesis is attached at the end of this chapter.

7.1 FireProt

MUSIL M, STOURAC J, BENDL J, BREZOVSKY J, PROKOP Z, ZENDULKA J, MARTINEK T, BEDNAR D, DAMBORSKY J. FireProt: Web Server for Automated Design of Thermostable Proteins. *Nucleic Acids Research.* 2017, 45(W1), W393-W399.

Author participation: 60%
Journal impact factor: 11.501 (Q1)

Author contribution

Designing and performing most of the computational experiments, analysing the data, writing the manuscript, implementing most of the software code.

Abstract

There is a continuous interest in increasing proteins stability to enhance their usability in numerous biomedical and biotechnological applications. A number of in silico tools for the prediction of the effect of mutations on protein stability have been developed recently. However, only single-point mutations with a small effect on protein stability are typically predicted with the existing tools and have to be followed by laborious protein expression, purification, and characterization. Here, we present FireProt, a web server for the automated design of multiple-point thermostable mutant proteins that combines structural and evolutionary information in its calculation core. FireProt utilizes sixteen tools and three protein engineering strategies for making reliable protein designs. The server is complemented with interactive, easy-to-use interface that allows users to directly analyze and

optionally modify designed thermostable mutants. FireProt is freely available at http://loschmidt.chemi.muni.cz/fireprot.

7.2 FireProt^{ASR}

MUSIL M, KHAN RT, BEIER A, STOURAC J, KONEGGER H, DAMBORSKY J, BEDNAR D. FireProt-ASR: Web Server for Fully Automated Ancestral Sequence Reconstruction. *Briefings in bioinformatics*. 2020, 0, 1-11. *(available in early access)*

Author participation: 60%
Journal impact factor: 8.990 (Q1)

Author contribution

Designing and conducting most of the experiments, analysing the data, writing the manuscript, designing and developing a novel algorithm for ancestral gaps reconstruction, implementing most of the software code.

Abstract

There is a great interest in increasing proteins' stability to widen their usability in numerous biomedical and biotechnological applications. However, native proteins cannot usually withstand the harsh industrial environment, since they are evolved to function under mild conditions. Ancestral sequence reconstruction is a well-established method for deducing the evolutionary history of genes. Besides its applicability to discover the most probable evolutionary ancestors of the modern proteins, ancestral sequence reconstruction has proven to be a useful approach for the design of highly stable proteins. Recently, several computational tools were developed, that make the ancestral reconstruction algorithms accessible to the community, while leaving the most crucial steps of the preparation of the input data on users' side. FireProt^{ASR} aims to overcome this obstacle by constructing a fully automated workflow, allowing even the unexperienced users to obtain ancestral sequences based on a sequence query as the only input. FireProt^{ASR} is complemented with an interactive, easy-to-use web interface and is freely available at https://loschmidt.chemi.muni.cz/fireprotasr/.

7.3 FireProt^{DB}

STOURAC J, DUBRAVA J, MUSIL M, HORACKOVA J, DAMBORSKY J, MAZURENKO S, BEDNAR D. FireProt-DB: Database of Manually Curated Protein Stability Data. *Nucleic Acids Research.* 2021, 49, D319-D324.

Author participation: 20%
Journal impact factor: 11.501 (Q1)

Author contribution

Defining the requirements for the project, defining data standardization, designing structure of the database, collecting and cleaning the initial data, contributing to the writing of the manuscript.

Abstract

The majority of naturally occurring proteins have evolved to function under mild conditions inside the living organisms. One of the critical obstacles for the use of proteins in biotechnological applications is their insufficient stability at elevated temperatures or in the presence of salts. Since experimental screening for stabilizing mutations is typically laborious and expensive, in silico predictors are often used for narrowing down the mutational landscape. The recent advances in machine learning and artificial intelligence further facilitate the development of such computational tools. However, the accuracy of these predictors strongly depends on the quality and amount of data used for training and testing, which have often been reported as the current bottleneck of the approach. To address this problem, we present a novel database of experimental thermostability data for single-point mutants FireProt^{DB}. The database combines the published datasets, data extracted manually from the recent literature, and the data collected in our laboratory. Its user interface is designed to facilitate both types of the expected use: (i) the interactive explorations of individual entries on the level of a protein or mutation and (ii) the construction of highly customized and machine learning-friendly datasets using advanced searching and filtering. The database is freely available at https://loschmidt.chemi.muni.cz/fireprotdb.

7.4 HotSpotWizard 2.0

BENDL J, STOURAC J, SEBESTOVA E, VAVRA O, MUSIL M, BREZOVSKY J, DAMBORSKY J. HotSpotWizard 2.0: Automated Design of Site-specific Mutations and Smart Libraries in Protein Engineering. *Nucleic Acids Research.* 2016, 44(W1), W479-W487.

Author participation: 15%
Journal impact factor: 11.501 (Q1)

Author contribution

Designing and developing one of the computational modules, performing its validation, analysingthe data, contribution on the writing of the paper.

Abstract

HotSpot Wizard 2.0 is a web server for automated identification of hot spots and design of smart libraries for engineering proteins' stability, catalytic activity, substrate specificity and enantioselectivity. The server integrates sequence, structural and evolutionary information obtained from 3 databases and 20 computational tools. Users are guided through the processes of selecting hot spots using four different protein engineering strategies and optimizing the resulting library's size by narrowing down a set of substitutions at individual randomized positions. The only required input is a query protein structure. The results of the calculations are mapped onto the protein's structure and visualized with a JSmol applet. HotSpot Wizard lists annotated residues suitable for mutagenesis and can automatically design appropriate codons for each implemented strategy. Overall, HotSpot Wizard provides comprehensive annotations of protein structures and assists protein engineers with the rational design of site-specific mutations and focused libraries. It is freely available at http://loschmidt.chemi.muni.cz/hotspotwizard.

7.5 Other original publications

- BENDL J, MUSIL M, ZENDULKA J, DAMBORSKY J, BREZOVSKY J. PredictSNP2: a Unified Platform for Accurately Evaluating SNP Effects by Exploiting the Different Characteristics of Variants in Distinct Genomic Regions. *PLoS Computational Biology.* 2016, 12, e1004962.

  Author participation: 35%
  Journal impact factor: 4.428 (Q1)

- MUSIL M, KONEGGER H, HON J, BEDNAR D, DAMBORSKY J. Computational Design of Stable and Soluble Biocatalysts. *ACS Catalysis.* 2018, 9, 1033-1054.

Author participation: 35%
Journal impact factor: 12.350 (Q1)

- BEERENS K, MAZURENKO S, KUNKA A, MARQUES S, HANSEN N, MUSIL M, CHALOUPKOVA R, WATERMAN J, BREZOVSKY J, BEDNAR D, PROKOP Z, DAMBORSKY J. Evolutionary Analysis as a Powerful Complement to Energy Calculations for Protein Stabilization. *ACS Catalysis*. 2018, 8, 9420-9428.

Author participation: 15%
Journal impact factor: 12.350 (Q1)

- KHAN RT, MUSIL M, STOURAC J, DAMBORSKY J, BEDNAR D. Fully Automated Ancestral Sequence Reconstruction using FireProt-ASR. *Current protocols in bioinformatics*. 2021. *(accepted for publication)*

Author participation: 40%
Journal impact factor: 9.630 (Q1)

- PLANAS-IGLESIAS J, MARQUES S, PINTO G, MUSIL M, STOURAC J, BEDNAR D, DAMBORSKY J. Computational Design of Enzymes for Biotechnological Applications. *Biotechnology advances*. 2021. *(accepted for publication)*

Author participation: 20%
Journal impact factor: 10.744 (Q1)

8 Concluding remarks

Stable proteins are utilized in various medical and biotechnological applications. However, native proteins have evolved to function in very mild conditions. Therefore, there is an increasing interest in improving protein stability by introducing mutations into the sequences of modern proteins. However, the saturation mutagenesis of all possible mutations is still far out of reach for many academic laboratories, creating the need for fast and reliable computational approaches. In the recent years, a plethora of computational tools was designed to deal with such a task, falling into one of the three main categories: i) tools based on force-field calculations, ii) tools utilizing the evolutionary information extracted from the set of homolog sequences, and iii) models built on top of the existing experimental data with the use of the modern machine learning methods.

The steady growth of the computational resources allowed for a comprehensive analysis of the mutational space, while the accuracy of stability-predicting methods is currently well-sufficient for the prioritization of experimentally validated mutations. Thus, *in silico* approaches are reducing the need for expensive and laborious laboratory experiments. However, most of the existing methods are viable only for predicting the single-point mutations with only a negligible effect on protein stability, while the construction of the multiple-point mutants is more complicated due to the possible occurrence of the antagonistic effects.

In this Thesis, several computational tools were presented to deal with designing stable multiple-point mutants. FireProt is a fully automated hybrid workflow that combines both energy- and evolution-based approaches in its calculation core. The tool utilizes sequence information, such as conservation and correlation of the amino acids in the MSA, as an initial filter to exclude those risky regions from the further calculation. Force-field approaches are then employed to select a pool of the potentially stable single-point mutations, which are then combined while eliminating most of the antagonistic effects by evaluating all the mutations' pairs. The second approach, FireProt$^{\mathrm{ASR}}$, is based on the idea that the ancestral proteins were significantly more stable than their extant counterparts. It is a fully automated workflow that allows users to utilize ancestral sequence reconstruction for their proteins without the deep knowledge of the essential bioinformatics tools and the biological system. FireProt$^{\mathrm{ASR}}$ deals with all steps of the ancestral reconstruction, including the search for the biologically relevant homolog sequences, construction of the MSA and phylogenetic tree, rooting of the tree without the need to specify its outgroup and finally the reconstruction of the ancestral sequences together with the identification of the ancestral gaps.

As the introduction of the stabilizing mutations into the protein structure often causes deterioration of other protein properties, the protein engineering tool HotSpotWizard was designed to add another level of abstraction. HotSpotWizard allows observing the protein by many different criteria, including its conservation and flexibility. Moreover, it provides the visualization of the sites and tunnels that are crucial for the function of the protein of interest. Stabilizing mutations designed by other methods can be analyzed in the HotSpotWizard tool to consider their position within a tertiary structure and the distance of those mutations from the sites essential for protein function. Such an analysis can unearth mutations that could (while stabilizing) compromise proteins activity and other properties, and therefore removing such a mutation could lead to the safer design of the engineered variant.

Finally, the work presented in this Thesis takes a stance on the current unsatisfactory situation surrounding the storage and management of the experimental data that are crucial for the training and validation of the computational tools based on the machine learning approaches. FireProt$^{\mathrm{DB}}$ is a comprehensive database of a protein stability data, supplemented with a sophisticated search engine and expanded by various annotations from the sequence and structural databases.

In conclusion, this Thesis presents a set of methods that aim to ease the engineering of highly stable multiple-point mutants, while providing users with a further analysis of the designed protein by considering other factors such as protein flexibility and location of the functional sites. Furthermore, it aims to simulate further improvement of the protein stability predictors by providing the research community with easy access to reliable experimental data.

In the future, the plan is to utilize the new high-quality dataset that was compiled for FireProt$^{\mathrm{DB}}$ to train a novel machine learning-based predictor of the effect of mutations on protein stability. This novel predictor would not be just a simple implementation of some of the standard machine learning techniques (e.g., SVM, RF), but rather a more complex multi-agent system that would focus more deeply on the mutations that are hard to predict by the existing predictors such as charge changing mutations located on the protein surface.

Bibliography

[1] Whitford D. *Proteins: Structure and function.* Wiley. 2005, ISBN 978-0-471-49894-0.

[2] Kurahashi R, Tanaka SI, Takano K. *Activity-stability trade-off in random mutant proteins.* J. Biosci. Bioeng. 2019, 128, 405-409.

[3] Siddiqui KS. *Defying the activity–stability trade-off in enzymes: taking advantage of entropy to enhance activity and thermostability.* Crit. Rev. Biotech. 2015, 37, 309-322.

[4] Yu H, Dalby PA. *Exploiting correlated molecular-dynamics networks to counteract enzyme activity–stability trade-off.* PNAS. 2018, 15, E12192-E12200.

[5] Babkova P, Sebestova E, Brezovsky J, Chaloupkova R, Damborsky J. *Ancestral Haloalkane Dehalogenases Show Robustness and Unique Substrate Specificity.* ChemBioChem 2017, 18, 1448-1456.

[6] Polizzi KM, Bommarius AS, Broering JM, Chaparro-Riggers JF. *Stability of biocatalysts.* Curr. Opin. Biotechnol. 2007, 11, 220-225.

[7] Ferdjani S, Ionita M, Roy B, Dion M, Djeghaba Z, Rabiller C, et al. *Correlation between thermostability and stability of glycosidases in ionic liquid.* Biotechnol. Lett. 2011, 33, 1215-1219.

[8] Gao D, Narasimhan DL, Macdonald J, Brim R, Ko M-C, Landry DW, et al. *Thermostable variants of cocaine esterase for long-time protection against cocaine toxicity.* Mol. Pharmacol. 2009, 75, 318-323.

[9] Gromiha MM. *Protein bioinformatics.* Elsevier. 2010, ISBN 978-81-312-2297-3.

[10] Bhu V. *THERMODYNAMICS AND IMFs IN PROTEIN STABILITY [online].* cit. 14. 10. 2020, `http://biochem-vivek.tripod.com/id23.html`

[11] Musil M, Konegger H, Hon J, Bednar D, Damborsky J. *omputational design of stable and soluble biocatalysts.* ACS Catalysis. 2018, 9, 1033-1054.

[12] Nickson AA, Clarke J. *What lessons can be learned from studying the folding of homologous proteins?.* Methods. 2010, 52, 38-50.

[13] Eisenberg D, McLachlan AD. *Solvation energy in protein folding and binding.* Nature. 1986, 319, 199-203.

[14] Ponnuswamy PK, Gromiha MM. *On the conformational stability of folded proteins.* J. Theor. Biol. 1994, 1, 63-74.

[15] Wijma HJ, Floor RJ, Janssen DB. *Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability.* Curr. Opin. Struct. Biol. 2013, 23, 588-594.

[16] Gumulya Y, Reetz MT. *Enhancing the thermal robustness of an enzyme by directed evolution: least favorable starting points and inferior mutants can map superior evolutionary pathways.* ChemBioChem. 2011, 12, 2502–2510.

[17] Bommarius AS, Paye MF. *Stabilizing biocatalysts.* Chem. Soc. Rev. 2013, 42, 6534–6565.

[18] Barlow DJ, Thornton JM. *Ion-pairs in proteins.* J. Mol. Biol. 1983, 168, 867-885.

[19] McDonald IK, Thornton JM. *Satisfying Hydrogen Bonding Potential in Proteins.* J. Mol. Biol. 1994, 238, 777-793.

[20] Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, et al. *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules.* J. Am. Chem. Soc. 1995, 117, 5179-5197.

[21] Thornton JM. *Disulphide bridges in globular proteins.* J. Mol. Biol. 1981, 151, 261-287.

[22] Kauzmann W. *Sulfur in proteins.* Elsevier. 1959, ISBN 978-0-12-395705-4.

[23] Musil M, Stourac J, Bendl J, Brezovsky J, Prokop Z, Zendulka J, Martinek T, Bednar D, Damborsky J. *FireProt: Web Server for Automated Design of Thermostable Proteins.* Nucleic Acids Res. 2017, 45, W393-399.

[24] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, et al. *UniProt: the universal protein knowledgebase.* Nucleic Acids Res. 2004, 32, D115-D119.

[25] Sussman JL, Lin D, Jiang J, Manning NO, Prilusky J, RItter O, Abola EE. *Protein Data Bank (PDB): Database of Three-Dimensional Structural Information of Biological Macromolecules.* Acta Cryst. 1998, D54, 1078-1084.

[26] Levinthal C. *How to Fold Graciously.* Spect. in Biol. Sys. 1969, 22-24.

[27] Anfinsen CB. *Principles that Govern the Folding of Protein Chains.* Science 1973, 181, 223-230.

[28] Tokuriki N, Stricher F, Serrano L, Tawfik DS. *How Protein Stability and New Functions Trade Off.* PLoS Comput. Biol. 2008, 4, e1000002.

[29] Arabnejad H, Dal Lago M, Jekel PA, Floor RJ, Thunnissen AMWH, Terwisscha van Scheltinga AC, Wijma HJ, Janssen DB. *A Robust Cosolvent-Compatible Halohydrin Dehalogenase by Computational Library Design.* Protein Eng, Des. Sel. 2017, 30, 175-189.

[30] Kuipers RK, Joosten HJ, Berkel WJH, Leferink NGH, Rooijen E, Ittmann E, Zimmeren F, Joschens H, et. al. *3DM: Systematic Analysis of Heterogeneous Superfamily Data to Discover Protein Fuctionalities.* Proteins: Struct., Funct., Bioinf. 2010, 78, 2101-2113.

[31] Goldenzweig A, Fleishman SJ. *Principles of Protein Stability and Their Application in Computational Design.* Annu. Rev. Biochem. 2018, 87, 105-129.

[32] Hansen N, Gunsteren WF. *Practical Aspects of Free-Energy Calculations: A Review.* J. Chem. Theory Comput. 2014, 10, 2632-2647.

[33] Nguyen V, Wilson C, Hoemberger M, Stiller JR, Agafonov RV, Kutter S, English J, Theobald DL, Kern D. *Evoluionary Drivers of Thermoadaptation in Enzyme Catalysis.* Science 2017, 355, 289-294.

[34] Risso Va, Gavira JA, Gaucher EA, Sanchez-Ruiz JM. *Phenotypic Comparisons of Consensus Variants versus Laboratory Resurrections of Precambrian Proteins.* Proteins: Struct., Funct., Genet. 2014, 82, 887-896.

[35] Bednar D, Beerens K, Sebestova E, Bendl J, Khare S, Chaloupkova R, Prokop Z, Brezovsky J, Baker D, Damborsky J. *FireProt: Energy- and Evolution-Based Computational Design of Thermostable Multiple-Point Mutants.* PLoS Comput. Biol. 2015, 11, e1004556.

[36] Goldenzweig A, Goldsmith M, Hill SE, Gertman O, Laurino P, Ashani Y, Dym O, Unger T, Albeck S, Prilusky J, Lieberman RL, et. al. *Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability.* Mol. Cell 2016, 63, 337-346.

[37] Broom A, Jacobi Z, Trainor K, Meiering EM. *Computational Tools Help Improve Protein Stability but with a Solubility Tradeoff.* J. Biol. Chem. 2017, 292, 14349-14361.

[38] Modarres HP, Mofrad MR, Sanati-Nezhad A. *Protein Thermostability Engineering.* RSC Adv. 2016, 6, 115252-115270.

[39] Pace CN, Scholtz JM, Grimsley GR. *Forces Stabilizing Proteins.* FEBS Lett. 2014, 588, 2177-2184.

[40] Lazaridis T, Karplus M. *Effective Energy Function for Protein Structure Prediction.* Curr. Opin. Struct. Biol. 2000, 10, 139-145.

[41] Seeliger D, Groot BL. *Protein Thermostability Calculations Using Alchemical Free Energy Simulations.* Biophys. J. 2010, 98, 2309-2316.

[42] Guerois R, Nielsen JE, Serrano L. *Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More than 1000 Mutations.* J. Mol. Biol. 2002, 320, 369-387.

[43] Mendes J, Guerois R, Serrano L. *Energy Estimation in Protein Design.* Curr. Opin. Struct. Biol. 2002, 12, 441-446.

[44] Dehouck Y, Gilis D, Rooman M. *A New Generation of Statistical Potentials for Proteins.* Biophys. J. 2006, 90, 4010-4017.

[45] Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. *PoPMuSiC 2.1: A Web Server for the Estimation of Protein Stability Changes upon Mutation and Sequence Optimality.* BMC Bioinf. 2011, 12, 151.

[46] Liu H. *On Statistical Energy Functions for Biomolecular Modeling and Design.* Quant. Biol. 2015, 3, 157-167.

[47] Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A. *ProTherm and ProNIT: Thermodynamic Databases for Proteins and Protein-Nucleic Acid Interactions.* Nucleic Acids Res. 2006, 34, D204-206.

[48] Potapov V, Cohen M, Schreiber G. *Assessing Computational Methods for Predicting Protein Stability upon Mutation: Good on Average but Not in the Details.* Protein Eng., Des. Sel. 2009, 22, 553-560.

[49] Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. *The FOldX Web Server: An Online Force Field.* Nucleic Acids Res. 2005, 33, W382-388.

[50] Kepp KP. *Towards a Golden Standard for Computing Globin Stability: Stability and Structure Sensitivity of Myoglobin Mutants.* Biochim. Biophys. Acta, Proteins Proteomics 2015, 1854, 1239-1248.

[51] Christensen NJ, Kepp KP. *Accurate Stabilities of Laccase Mutants Predicted with a Modified FoldX Protocol.* Protocol. J. Chem. Inf. Model. 2012, 52, 3028-3042.

[52] MacKerell AD, BashFord D, Bellott M, Dunbrack RL, Evanseck JD, et. al. *All-Atom Empirical Potentials for Molecular Modeling and Dynamics Studies of Proteins.* J. Phys. Chem. B. 1998, 102, 3586-3616.

[53] Oosternbrink C, Villa A, Mark AE, Gunsteren WF. *A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS FOrce-Field Parameter Sets 53A5 and 53A6.* J. COmput. Chem. 2004, 25, 1656-1676.

[54] Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, et. al. *THe Rosetta All-Atom Energy Function for Macromolecular Modeling and Design.* J. Chem. Theory Comput. 2017, 13, 3031-3048.

[55] Davey JA, Damry AM, Euler CK, Goto NK, Chica RA. *Prediction of Stable Globular Proteins Using Negative Design with Non-Native Bakcbone Ensembles.* Structure 2015, 23, 2011-2021.

[56] Conchúir S, Barlow KA, Pache RA, Ollikainen N, Kundert K, O'Meara MJ, Smith CA, Kortemme T. *A Web Resource for Standardized Benchmark Datasets, Metrics, and Rosetta Protocols for Macromolecular Modeling adn Design.* PLoS One 2015, 10, e0130433.

[57] Trainor K, Broom A, Meiering EM. *Exploring the Relationships between Protein Sequence, Structure and Solubility.* Curr. Opin. Struct. Biol. 2017, 42, 136-146.

[58] Das R. *Four Small Puzzles that Rosetta Doesn't Solve.* PLoS One 2011, 6, e20044.

[59] Kellog EH, Leaver-Fay A, Baker D. *Role of Conformational Sampling in Computing Mutation-Induced Changes in Protein Structure and Stability.* Proteins: Struct., Funct., Genet. 2011, 79, 830-838.

[60] Bush J, Makhatadze GI. *Statistical Analysis of Protein Structures Suggests That Buried Ionizable Residues in Proteins Are Hydrogen Bonded or Form Salt Bridges.* Proteins: Struct., Funct., Genet. 2011, 79, 2027-2032.

[61] Stranges PB, Kuhlman BA. *Comparison of Successful and Failed Protein Interface Designs Highlights the Challenges of Designing Buried Hydrogen Bonds.* Protein Sci. 2013, 22, 74-82.

[62] Beerens K, Mazurenko S, Kunka A, Marques SM, Hansen N, Musil M, Chaloupkova R, Waterman J, Brezovsky J, Bednar D, Prokop Z, Damborsky J. *Evolutionary Analysis Is a Powerful Complement to Energy Calculations for Protein Stabilization.* ACS Catal. 2018, 8, 9420-9428.

[63] Wijma HJ, Floor RJ, Jekel PA, Baker D, Marrink SJ, Janssen DB. *Computationally Designed Libraries for Rapid Enzyme Stabilization.* Protein Eng., Des. Sel. 2014, 27, 49-58.

[64] Wickstrom L, Gallicchio E, Levy RM. *The Linear Interaction Energy Method for the Prediction of Protein Stability Changes Upon Mutation.* Proteins: Struct., Funct., Genet. 2012, 80, 111-125.

[65] Thiltgen G, Goldstein RA. *Assessing Predictors of Changes in Protein Stability upon Mutation Using Self-Consistency.* PLoS One 2012, 7, e46084.

[66] Bub O, Rudat J, Ochsenreither K. *FoldX as Protein Engineering Tool: Better Than Rnadom Based Approaches?.* Comput. Struct. Biotechnol. J. 2018, 16, 26-33.

[67] Barlow KA, Conchúir ÓS, Thompson S, Suresh P, Lucas JE, Heinonen M, Kortemme T. *Flex DdG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation.* J. Phys. Chem. B. 2018, 122, 5389-5399.

[68] Ludwiczak J, Jarmula A, Dunin-Horkawicz S. *Combining Rosetta with Molecular Dynamics (MD): A Benchmark of the MD-Based Ensemble Protein Design.* J. Struct. Biol. 2018, 203, 54-61.

[69] Davis IW, Arendall WB, Richardson DC, RIchardson JS. *The Backrub Motion: How Protein Backbone Shrugs when a Sidechain Dances.* Structure 2006, 14, 265-274.

[70] Fan H, Mark AE. *Relative Stability of Protein Structures Determined by X-Ray Crystallography or NMR Spectroscopy: A Molecular Dynamics Simulation Study.* Proteins: Struct., Funct., Genet. 2003, 53, 111-120.

[71] Kuzmanic A, Pannu NS, Zagrovic B. *X-Ray Refinement Significantly Underestimates the Level of Microscopic Heterogeneity in Biomolecular Crystals.* Nat. Commun. 2014, 5, 3220.

[72] Folkman L, Stantic B, Sattar A, Zhou Y. *EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models.* J. Mol. Biol. 2016, 428, 1394-1405.

[73] Teng S, Srivastava AK, Wang L. *Sequence Feature-Based Prediction of Protein Stability Changes upon Amino Acid Substitutions.* BMC Genomics 2010, 11, S5.

[74] Huang LT, Gromiha MM, Ho SY. *IPTREE-STAB: Interpretable Decision Tree Based Method for Predicting Protein Stability Changes upon Mutations.* bioinformatics 2007, 23, 1292-1293.

[75] Liaw A, Wiener M. *Classification and Regression by RandomForest.* R. News 2002, 2, 18-22.

[76] Breiman L. *Random Forests.* Mach. Learn. 2001, 45, 5-32.

[77] Boughorbel S, Jarray F, El-Anbari M. *Optimal Classifier for Imbalanced Data Using Matthews Correlation Coefficient Metric.* PLoS One 2017, 12, e0177678.

[78] Ling CX, Sheng VS. *Cost-Sensitive Learning and the Class Imbalance Problem.* In Encyclopedia of Machine Learning. Sammut, C., Springer. New York, 2007.

[79] Rao R, Fung G, Rosales R. *On the Dangers of Cross-Validation. An Experimental Evaluation.* In Proceedings of the 2008 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics: Philadelphia. 2008, 588-596.

[80] Altschul SF, Gish W, Miller W, Mayers EW, LIpman DJ. *Basic Local Alignment Search Tool.* J. Mol. Biol. 1990, 215, 403-410.

[81] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. *Gapped BLAST and PSI-BLAST: A new Generation of Protein Database Search Programs.* Nucleic Acids Res. 1997, 25, 3389-3402.

[82] Remmert M, Biegert A, Hauser A, Soding J. *HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment.* Nat. Methods 2012, 9, 173-175.

[83] Steipe B, Schiller B, Pluckthun A, Steinbacher S. *Sequence Statistics Reliably Predict Stabilizing Mutations in Protein Domain.* J. Mol. Biol. 1994, 240, 188-192.

[84] Sullivan BJ, Nguyen T, Durani V, Mathur D, Rojas S, THomas M, Syu T, Magliery TJ. *Stabilizing Proteins from Sequence Statistics: The Interplay of Conservation and Correlation in Triosephosphate Isomerase Stability.* J. Mol. Biol. 2012, 420, 384-399.

[85] Lehmann M, Kostrewa D, Wyss M, Brugger R, D['Arcy A, Pasamontes L, Loon AP. *From DNA Sequence to Improved Functionality: Using Protein Sequence Comparisons to Rapidly Design a Thermostable Consensus Phytase.* Protein Eng., Des. Sel. 2000, 13, 49-57.

[86] Magliery TJ. *Protein Stability: Computation, Sequence Statistics, and New Experimental Methods.* Curr. Opin. Struct. Biol. 2015, 33, 161-168.

[87] Porebski BT, Buckle AM. *Consensus Protein Design.* Protein Eng., Des. Sel. 2016, 29, 245-251.

[88] Jackel C, Bloom JD, Kast P, Arnold FH, Hilvert D. *Consensus Protein Design without Phylogenetic Bias.* J. Mol. Biol. 2010, 399, 541-546.

[89] Goyal VD, Magliery TJ. *Phylogenetic Spread of Sequences Data Affects Fitness of SOD1 Consensus Enzymes: Insights from Sequence Statistics and Structural Analyses.* Proteins: Struct., Funct., Genet. 2018, 86, 609-620.

[90] Vazquez-Figueroa E, Chaparro-Riggers J, Bommarius AS. *Development of a Thermostable Glucose Dehydrogenase by a Structure-Guided Consensus Concept.* ChemBioChem 2007, 8, 2295-2301.

[91] Parthasarathy S, Murthy MR. *Protein Thermal Stability: Insights from Atomic Displacement Parameters (B Values).* Protein Eng., Des. Sel. 2000, 13, 9-13.

[92] Cole MF, Gaucher EA. *Exploiting Models of Molecular Evolution to Efficiently Direct Protein Engineering.* J. Mol. Evol. 2011, 72, 193-203.

[93] Hochberg GKA, Thornton JW. *Reconstructing Ancient Proteins to Understand the Causes of Structure and Function.* Annu. Rev. Biophys. 2017, 46, 247-269.

[94] Aerts D, Verhaeghe T, Joosten HJ, Vriend G, Soetaert W, Desmet T. *Consensus Engineering of Sucrose Phosphorylase: The Outcome Reflects the Sequence Input.* Biotechnol. Bioeng. 2013, 110, 2563-2572.

[95] Trudeau DL, Kaltenbach M, Tawfik DS. *On the Potential Origins of the High Stability of Reconstructed Ancestral Proteins.* Mol. Biol. Evol. 2016, 33, 2633-2641.

[96] Wheeler LC, Lim SA, Marqusee S, Harms MJ. *The Thermostability and Specificity of Ancient Proteins.* Curr. Opin. Struct. Biol. 2016, 38, 37-43.

[97] Yang Z. *PAML: A Program Package for Phylogenetic Analysis by Maximum Likelihood.* Bioinformatics 1997, 13, 555-556.

[98] Stamatakis A. *RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models.* Bioinformatics 2006, 22, 2688-2690.

[99] Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. *Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology.* Science 2001, 294, 2310-2314.

[100] Eick GN, Bridgham JT, Anderson DP, Harms MJ, Thornton JW. *Robustness of Reconstructed Ancestral Protein Functions to Statistical Uncertainty.* Mol. Biol. Evol. 2016, 34, 247-261.

[101] Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M. *Fast and Accurate Predictions of Protein Stability Changes upon Mutations Using Statistical Potentials and Neural Networks: PoPMuSiC-2.0.* Bioinformatics 2009, 25, 2537-2543.

[102] Khatun J, Khare SD, Dokholzan NV. *Can Contact Potentials Reliably Predict Stability of Proteins?.* J. Mol. Biol. 2004, 336, 1223-1238.

[103] Pucci F, Bernaerts KV, Kwasigroch JM, Rooman M. *Quantification of Biases in Predictions of Protein Stability Changes upon Mutations.* Bioinformatics 2018, 34, 3659-3665.

[104] Yin S, Ding F, Dokholyan NV. *ERIS: An Automated Estimator of Protein Stability.* Nat. Methods 2007, 4, 466-467.

[105] Benedix A, Becker CM, Groot BL, Caflisch A, Bockmann RA. *Predicting Free Energy Changes Using Structural Ensembles.* Nat. Methods 2009, 6, 3-4.

[106] Pronk S, Pall S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, Shirts MR, Smith JC, Kasson PM, Spoel D, Hess B, Lindahl E. *GROMACS 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit.* Bioinformatics 2013, 29, 845-854.

[107] Groot BL, Aalten DM, Scheek RM, Amadei A, Vriend G, Berendsen HJC. *Prediction of Protein Conformational Freedom from Distance Constraints.* Proteins: Struct., Funct., Genet. 1997, 29, 240-251.

[108] Hoppe C, Schomburg D. *Prediction of Protein Thermostability with a Direction- and Distance-Dependent Knowledge-Based Potential.* Protein Sci. 2005, 14, 2682-2692.

[109] Pucci F, Bourgeas R, Rooman M. *Predicting Protein Thermal Stability Changes upon Point Mutations Using Statistical Potentials: Introducing HoTMuSiC.* Sci. Rep. 2016, 6, 23257.

[110] Capriotti E, Fariselli P, Casadio R. *I-Mutant2.0: Predicting Stability Changes upon Mutation from the Protein Sequence or Structure*. Nucleic Acids Res. 2005, 33, W306-310.

[111] Cheng J, Randall A, Baldi P. *Prediction of Protein Stability Changes for Single-Site Mutations Using Support Vector Machines*. Proteins: Struct., Funct., Genet. 2006, 62m 1125-1132.

[112] Wainreb G, Wolf L, Ashkenazy H, Dehouck Y, Ben-Tal N. *Protein Stability: A single Recorded Mutation Aids in Predicting the Effects of Other Mutations in the Same Amino Acid Site*. Bioinformatics 2011, 27, 3286-3292.

[113] Li Y, Fang J. *PROTS-RF: A Robust Model for Predicting Mutation-Induced Protein Stability Changes*. PLoS One 2012, 7, e47247.

[114] Quang D, Chen Y, Xie X. *DANN: A Deep Learning Approach for Annotating the Pathogenicity of Genetic Variants*. Bioinformatics 2015, 31, 761-763.

[115] Wang Y, Mao H, Yi Z. *Protein Secondary Structure Prediction by Using Deep Learning Method*. Know.-Based Syst. 2017, 118, 115-123.

[116] Ivakhnenko AG. *Polynomial Theory of Complex Systems*. IEEE Trans. Syst., Man, Cybern. 1971, SMC-1, 364-378.

[117] Bengio Y, Boulanger-Lewandowski N, Pascanu R. *Advances in Optimizing Recurrent Networks*. IEEE International Conference on Acoustics, Speech and Signal Processing 2013, 8624-8628.

[118] Cang Z, Wei GW. *TopologyNet: Topology Based Deep Convolutional and Multi-Task Neural Networks for Biomolecular Property Predictions*. PLoS Comput. Biol. 2017, 13, e1005690.

[119] Laimer J, Hofer H, Fritz M, Wegenkittl S, Lackner P. *MAESTRO - Multi Agent Stability Predictor upon Point Mutations*. BMC Bioinf. 2015, 16, 116.

[120] Khan S, Vihinen M. *Performance of Protein Stability Predictors*. Hum. Mutat. 2010, 31, 675-684.

[121] Usmanova DR, Bogatyreva NS, Arino BJ, Eremina AA, Gorshkova AA, Kanevskiy GM, Lonishin LR, Meister AV, et. al. *Self-Consistency Test Reveals Systematic Bias in Programs for Prediction Change of Stability upon Mutation*. Bioinformatics 2018, 34, 3653-3658.

[122] Montanucci L, Martelli PL, Ben-Tal N, Fariselli PA. *Natural Upper Bound to the Accuracy of Predicting Protein Stability Changes upon Mutations.* Bioinformatics 2019, 35, 1513-1517.

[123] Rice P, Longden I, Bleasby A. *EMBOSS: The European Molecular Biology Open Software Suite.* Trends Genet. 2000, 16, 276-277.

[124] Lu G, Moriyama EN. *Vector NTI, a Balanced All-in-One Sequence Analysis Suite.* Briefings Bioinf. 2004. 5, 378-388.

[125] Bendl J, Soutrac J, Sebestova E, Vavra O, Musil M, Brezovsky J, Damborsky J. *HotSpot Wizard 2.0: Automated Design of Site-Specific Mutations an Smart Libraries in Protein Engineering.* Nucleic Acids Res. 2016, 44, W479-487.

[126] Stamatakis A. *RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenetics.* Bioinformatics 2014, 30, 1312-1313.

[127] Ashkenazzy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, Pupko T. *FastML: A Web Server for Probabilistic Reconstruction of Ancestral Sequences.* Nucleic Acids Res. 2012, 40, W580-584.

[128] Diallo AB, Makarenkov V, Blanchette M. *Ancestors 1.0: A Web Server for Ancestral Sequence Reconstruction.* Bioinformatics 2010, 26, 130-131.

[129] Westesson O, Barquist L, Holmes I. *HandAlign: Bayesian Multiple Sequence Alignment, Phylogeny and Ancestral Reconstruction.* Bioinformatics 2012, 28, 1170-1171.

[130] Ronquist F, Teslenko M, Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. *MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space.* Syst. Biol. 2012. 61, 539-542.

[131] Finn RD, Clements J, Eddy SR. *HMMER Web Server: Interactive Sequence Similarity Searching.* Nucleic Acids Res. 2011, 39, W29-37.

[132] Altschul SF, Gertz EM, Agarwala R, Schaffer AA, Yu YK. *PSI-BLAST Pseudocounts and the Minimum Description Length Principle.* Nucleic Acids Res. 2009, 37, 815-824.

[133] Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, De Matos C, Myers CA, Kamisetty H, Blair P, Wilson IA, Baker D. *Optimization of Affinity, Specificity and Function of Designed Influenza Inhibitors Using Deep Sequencing.* Nat. Biotechnol. 2012, 30, 543-548.

[134] Parthiban V, Gromiha MM, Schomburg D. *CUPSAT: Prediction of Protein Stability Upon Point Mutations.* Nucleic Acids Res. 2006, 34, W239-242.

[135] Wang G, Dunbrack RL. *PISCES: a Protein Sequence Culling Server.* Bioinformatics 2003, 12, 1589-1591.

[136] Witvliet DK, Strokach A, Giraldo-Forero AF, Teyra J, Colak R, Kim PM. *ELASPIC Web-Server: Proteome-Wide Structure-Based Prediction of Mutation Effects on Protein Stability and Binding Affinity.* Bioinformatics 2016, 32, 1589-1591.

[137] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. *SMOTE: Synthetic Minority Over-Sampling Technique.* J. of Artific. Intel. Res. 2002, 16, 321-357.

[138] He H, Bai Y, Garcia EA, Li S. *ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning.* IJCNN 2008, 1322-1328.

[139] Gopi S, Devanshu D, Krishna P, Naganathan AN. *pStab: prediction of stable mutants, unfolding curves, stability maps and protein electrostatic frustration.* Bioinformatics 2017, 43, 875-877.

[140] Frappier V, Chartier M, Najmanovich RJ. *ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability.* Nucleic Acids Res. 2015, 43, W395-W400.

[141] Giollo M, Martin AJM, Walsh I, Ferrari C, Tosatto SCE. *NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation.* BMC Genomics 2014, 15, S7.

[142] Pandurangan AP, Ochoa-Montano B, Acher DB, Blundell TL. *SDM: a server for predicting effects of mutations on protein stability.* Nucleic Acids Res. 2017, 45, W229-W235.

[143] Masso M, Vaisman II. *AUTO-MUTE 2.0: A Portable Framework with Enhanced Capabilities for Predicting Protein Functional Consequences upon Mutation.* Adv. in Bioinf. 2014, 278385.

[144] Quan L, Lv Q, Zhang Y. *STRUM: structure-based prediction of protein stability changes upon single-point mutation.* Bioinformatics. 2016, 32, 2936-2946.

[145] Hu F, Lin Y, Tang J. *MLGO: phylogeny reconstruction and ancestral inference from gene-order data.* BMC Bioinformatics. 2014, 15, 354.

[146] Sagulenko P, Puller V, Neher RA. *TreeTime: Maximum-likelihood phylodynamic analysis.* Virus Evolution. 2018, 4, vex042.

[147] Hanson-Smith V, Johnson A. *PhyloBot: A Web Portal for Automated Phylogenetics, Ancestral Sequence Reconstruction, and Exploration of Mutational Trajectories.* PLoS Comp. Biol. 2016, e1004976.

[148] Menzel P, Stadler PF, Gorodkin J. *maxAlike: maximum likelihood-based sequence reconstruction with application to improved primer design for unknown sequences.* Bioinformatics. 2011, 27, 317-325.

[149] Pires DEV, Ascher DB, Blundell TL. *mCSM: predicting the effects of mutations in proteins using graph-based signatures.* Bioinformatics. 2014, 30, 335-342.

[150] Savojardo C, Fariselli P, Martelli PL, Casadio R. *INPS-MD: a web server to predict stability of protein variants from sequence and structure.* Bioinformatics. 2016, 16, 2542-2544.

[151] Chen CW, Lin J, Chu YW. *iStable: off-the-shelf predictor integration for predicting protein stability changes.* BMC Bioinformatics. 2013, 14, S5.

[152] Tian J, Wu N, Chu X, Fan Y. *Software Predicting changes in protein thermostability brought about by single- or multi-site mutations.* BMC Bioinformatics. 2010, 11, 370.

[153] Saitou N, Nei M. *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Mol. Biol. and Evol. 1987, 4, 406-425.

[154] Sokal M. *A statistical method for evaluating systematic relationships.* Univ. of Kansas Sci. Bulletin. 1958, 1409-1438.

[155] Price MN, Dehal PS, Arkin AP. *FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments.* PLoS One. 2010, 5, e9490.

[156] Sankoff, D. *Simultaneous solution of the RNA folding, alignment and protosequence problems.* SIAM J. Appl. Math. 1985, 45810–825.

[157] Takahashi K, Nei M. *Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used.* Mol. Biol. Evol. 2000, 17, 1251-8.

[158] Felsenstein J. *Maximum-likelihood estimation of evolutionary trees from continuous characters.* Am. J. Hum. Genet. 1973, 25, 471-492.

[159] Maddison WP, Donoghue MJ, Maddison DR. *Outgroup analysis and parsimony.* Syst. Zool. 1984, 33, 83–103.

[160] Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. *Phylogenetic inference.* Molecular Systematics. 1996, ISBN 978-0878932825.

[161] Tria FDK, Landan G, Dagan T. *Phylogenetic rooting using minimal ancestor deviation.* Nature Ecol. Evol. 2017,1,0193.

[162] Fitch WM, Margoliash E. *Construction of phylogenetic trees.* Science. 1967, 155, 279-284.

[163] Bolstad WM. *Understanding Computational Bayesian Statistics.* Wiley. 2010, ISBN 0-470-04609-0.

[164] Felsenstein J. *Cases in which parsimony or compatibility methods will be positively misleading.* System. Zoology. 1978, 27, 401-410.

[165] Williams PD, Pollock DD, Blackburne BP, Goldstein RA. *Assessing the accuracy of ancestral protein reconstruction methods.* PLoS Comput. Biol. 2006, 2, e69.

List of symbols, quantities and abbreviations

MSA	Multiple-sequence alignment
ASR	Ancestral sequence reconstruction
ML	Maximum-likelihood
MSA	Multiple-sequence alignment
BI	Bayesian inference
CD	Consensus design
CA	Conservation analysis
PP	Posterior probabilities
SVM	Support vector machines
RF	Random forest
HMM	Hidden Markov Model
Å	Ångström
G	Gibbs free energy
T_m	Melting temperature
PCC	Pearson correlation coefficient
MCC	Matthews correlation coefficient
ASA	Accessible surface area
MD	Molecular dynamics
PEEF	Physical effective energy functions
SEEF	Statistical effective energy functions
EEEF	Empirical effective energy functions
PDB	Protein Data Bank
RMSE	Root-mean-square error
UPGMA	Unweighted Paired Group Method

List of appendices

A Original publication I: FireProt

MUSIL M, STOURAC J, BENDL J, BREZOVSKY J, PROKOP Z, ZENDULKA J, MARTINEK T, BEDNAR D, DAMBORSKY J. FireProt: Web Server for Automated Design of Thermostable Proteins. *Nucleic Acids Research*. 2017, 45(W1), W393-W399.

Author participation: 60%

Journal impact factor: 11.501 (Q1)

Published online 26 April 2017 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* ***W393–W399***
doi: 10.1093/nar/gkx285

FireProt: web server for automated design of thermostable proteins

Milos Musil[1,2,3,†], Jan Stourac[1,3,†], Jaroslav Bendl[1,2,3], Jan Brezovsky[1,3], Zbynek Prokop[1,3], Jaroslav Zendulka[2,4], Tomas Martinek[1,2,4], David Bednar[1,3,*] and Jiri Damborsky[1,3,*]

[1]Loschmidt Laboratories, Department of Experimental Biology, Masaryk University, Brno, Czech Republic, [2]Department of Information Systems, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic, [3]International Centre for Clinical Research, St. Anne's University Hospital Brno, Brno, Czech Republic and [4]Centre of Excellence IT4Innovations, Technical University Ostrava, Ostrava

Received February 11, 2017; Revised April 02, 2017; Editorial Decision April 10, 2017; Accepted April 11, 2017

ABSTRACT

There is a continuous interest in increasing proteins stability to enhance their usability in numerous biomedical and biotechnological applications. A number of *in silico* tools for the prediction of the effect of mutations on protein stability have been developed recently. However, only single-point mutations with a small effect on protein stability are typically predicted with the existing tools and have to be followed by laborious protein expression, purification, and characterization. Here, we present FireProt, a web server for the automated design of multiple-point thermostable mutant proteins that combines structural and evolutionary information in its calculation core. FireProt utilizes sixteen tools and three protein engineering strategies for making reliable protein designs. The server is complemented with interactive, easy-to-use interface that allows users to directly analyze and optionally modify designed thermostable mutants. FireProt is freely available at http://loschmidt.chemi.muni.cz/fireprot.

INTRODUCTION

Proteins are widely used in numerous biomedical and biotechnological applications. However, naturally occurring proteins cannot usually withstand the harsh industrial environment, since they are mostly evolved to function at mild conditions (1). Protein engineering has revolutionized the utilization of naturally available proteins for different industrial applications by improving various protein features such as stability, activity or enantioselectivity to surpass their natural limitations. Protein stability is generally strongly correlated with its expression yield (2), half-life (3), serum survival time (4) and performance in the presence of denaturing agents (5). Thus, stability is one of the key determinants of proteins applicability in biotechnological processes.

In the ideal case, the saturation mutagenesis would be applied to evaluate every possible mutation on every position of the engineered protein (6). However, such a search space would be enormous and the experimental evaluation can delay the design of truly thermostable protein for months or even years. Therefore, there are demands for effective and precise predictive computation of protein stability. To satisfy this goal a number of *in silico* tools have been developed recently. Some of these tools such as EASE-MM (7), I-Mutant (8) or mCSM (9) are based on machine learning techniques. Others are using so-called energetic functions. These programs can be further categorized into two groups. The first group utilizes a physical effective energy function for simulating the fundamental forces between atoms and is represented by the programs like Rosetta (10) and Eris (11). The second group is based on statistical potentials for which the energies are derived from frequencies of residues or atom contacts reported in the datasets of experimentally characterized protein mutants, e.g. PopMuSiC (12) and FoldX (13). However, due to the potentially antagonistic effect of mutations, only single-point mutations are usually predicted *in silico* and have to be followed by laborious and costly protein expression, purification and characterization. Single-point mutations typically enhance the melting temperature of target proteins by units of degree (3,14). A much higher degree of stabilization can be achieved by constructing multiple-point mutants (15). We have recently developed the FireProt (16), combining energy- and evolution-based approaches for reliable design of stable multiple-point mutants. The protocol includes several preceding filters that accelerate the calculation by omitting potentially deleterious mutations. FireProt is currently

[*]To whom correspondence should be addressed. Tel: +420 5 4949 3467; Fax: +420 5 4949 6302; Email: jiri@chemi.muni.cz; Website address: http://loschmidt.chemi.muni.cz/fireprot (type of web server: computational workflow)
Correspondence may also be addressed to David Bednar. Email: 222755@mail.muni.cz
[†]These authors contributed equally to the work as first authors.

available only in a stand-alone format and requires extensive experience in bioinformatics to carry out all necessary steps of the work flow. Currently, we are aware of only one server for design of stable multiple-point mutants - PROSS (17), utilizing Rosetta modeling and phylogenetic sequence information in its computation core.

Here, we present a web version of FireProt for the automated design of thermostable proteins. FireProt integrates sixteen computational tools and utilizes both sequence and structural information. FireProt web server provides users with thermostable proteins, constructed by three distinct strategies: (i) evolution-based approach, utilizing back-to-consensus analysis; (ii) energy-based approach, evaluating change in free energy upon mutation and (iii) combination of both evolution-based and energy-based approaches. In our view, it is very important to have this integrated approach, since phylogenetic analysis enables identification of the mutations stabilized by entropy, which cannot be predicted by force field calculations (Beerens *et al.*, under review). The server allows users to include preferred mutations into the thermostable protein, to generate corresponding structures and sequences for gene syntheses. Compared to the previously published FireProt protocol (16), minimum effort and no bioinformatics knowledge is required from users to calculate and analyze the results. Furthermore, all input parameters and computational protocols were optimized to minimize otherwise highly time demanding procedure. The server was complemented with a graphical interface allowing users to directly analyze the protein of interest and design multiple-point mutants.

MATERIALS AND METHODS

The basic workflow of FireProt strategy is outlined in Figure 1. In order to design a highly reliable thermostable multiple-point mutant, a protein defined by the user is annotated using several prediction tools and databases (Phase 1). With this knowledge in hand, energy- and evolution-based approach is applied to assemble a list of potentially stabilizing single-point mutations (Phase 2). Finally, three multiple-point mutants are generated in an additive manner, while removing potentially antagonistic effects of mutations (Phase 3).

Phase 1: Annotation of the protein

Initially, the user is requested to specify the protein structure, either by providing its PDB ID or by uploading a user-defined PDB file. The biological assembly of the target protein is then automatically generated by the MakeMultimer tool (http://watcut.uwaterloo.ca/tools/makemultimer/). Sequence homologs are obtained by performing a BLAST search (18) against the UniRef90 database (19), using the target protein sequence as an input query. Identified homologs are then aligned with the query protein using USEARCH (20), while sequences whose identity with the query is below or above the user defined thresholds (default: 30 and 90%) are excluded from the list of homologs. The remaining sequences are clustered using UCLUST (20), with a 90% identity threshold to remove close homologs. The cluster representatives are sorted based on the BLAST query coverage and by default, the first 200 of them are used to create a multiple sequence alignment with Clustal Omega tool (21). The multiple sequence alignment is used to: (i) estimate the conservation coefficient of each residue position in the protein based on the Jensen–Shannon entropy (22); (ii) identify correlated positions employing a consensual decision of the OMES (23), MI (24), aMIc (25), DCA (26), SCA (27), ELSC (28), McBASC (29) and (iii) analyze amino acid frequencies at individual positions within the protein.

Phase 2: Prediction of single-point mutations

In accordance with the original FireProt protocol, potentially stabilizing single-point mutations are identified via two separate branches: one relying on the estimation of the change of free energy upon mutation and second utilizing back-to-consensus approach.

The first, energy-based approach is employing FoldX and Rosetta tools that performed best on our testing dataset. Preceding filters accelerate the calculation by omitting potentially deleterious mutations. Prior to the identification of the single-point mutations itself, the target protein structure is amended and minimized. FoldX protocol is utilized to fill in the missing atoms in the residues and patched structure is consequently minimized with Rosetta minimization module. Conserved and correlated positions are immediately excluded from further analysis. It was observed that functional and structural constraints in proteins generally lead to the conservation of amino acid residues (30–33). Similarly, correlated residues ordinarily help to maintain protein function, folding or stability (34–36). Mutations conducted on these positions are therefore considered unsafe by current FireProt strategy, even though there is certainly a space for more sophisticated treatment of correlated positions, which will be further developed in future versions of FireProt server.

The remaining positions are subjected to saturation mutagenesis by using FoldX tool. Mutations with predicted ddG over given threshold (default: –1 kcal/mol) are steered away and rest is forwarded to Rosetta calculations. Finally, the mutations predicted by Rosetta as strongly stabilizing (default cut-off: –1 kcal/mol) are tagged as potential candidates for the design of the multiple-point mutants.

A high time demands of Rosetta analysis were one of the most excruciating issues with the original FireProt protocol. Even with the application of filters over 100 mutations was usually left for precise, but slow, Rosetta calculations. For this reason, we have evaluated several force fields and Rosetta protocols with the newly assembled dataset containing 1573 mutations from ProTherm database (37) and HotMuSiC dataset (38). Based on the results of the evaluations, the best trade-off between the time requirements and precision was selected. With Rosetta protocol 3, we have achieved more than tenfold increase in calculation speed while preserving high prediction accuracy. Details on dataset construction and protocols evaluation can be found in the Supplement 1 (Supplementary Tables S1–S5).

The second approach is based on the information obtained from multiple sequence alignment. The most common amino acid in each position of protein sequence often

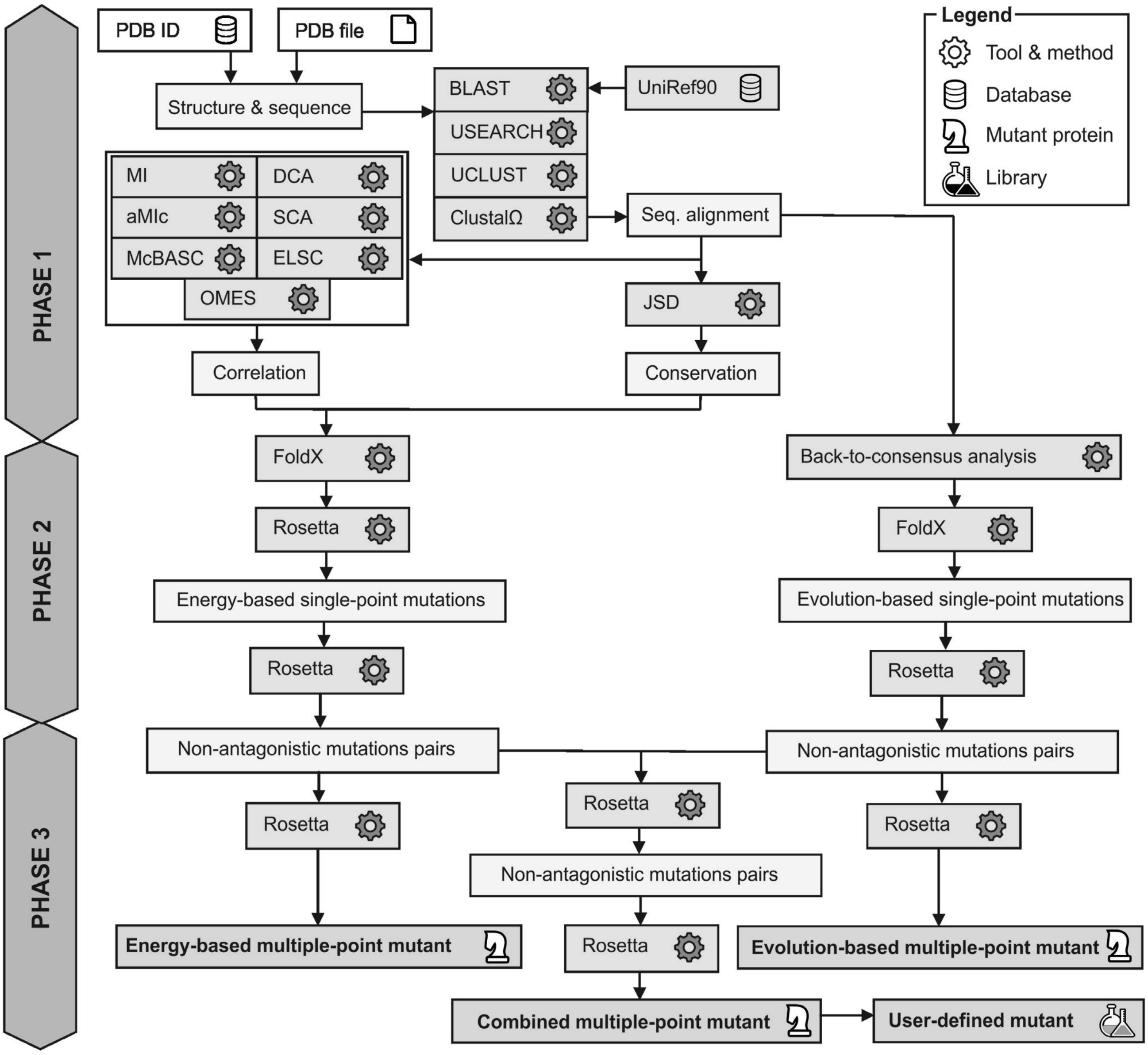

PHASE 1
PHASE 2
PHASE 3
PDB ID
PDB file
Structure & sequence
BLAST
USEARCH
UCLUST
ClustalΩ
UniRef90
Seq. alignment
MI
DCA
aMIc
SCA
McBASC
ELSC
OMES
JSD
Correlation
Conservation
FoldX
Rosetta
Energy-based single-point mutations
Rosetta
Non-antagonistic mutations pairs
Rosetta
Energy-based multiple-point mutant
Back-to-consensus analysis
FoldX
Evolution-based single-point mutations
Rosetta
Non-antagonistic mutations pairs
Rosetta
Evolution-based multiple-point mutant
Rosetta
Non-antagonistic mutations pairs
Rosetta
Combined multiple-point mutant
User-defined mutant
Legend
Tool & method
Database
Mutant protein
Library

Figure 1. Workflow of FireProt strategy.

provides a non-negligible effect on protein stability (39–42). Therefore, FireProt implements majority and frequency ratio approach to identify mutations at positions where the wild-type amino acid differs from the most prevalent one. By default, the single out mutations are located in the positions where the consensus residue is present in at least 50% of all analyzed sequences (majority method) or where consensus residue frequency is 40% and is at least five times more frequent than the wild-type amino acid (frequency ratio method). These thresholds were chosen in accordance to the previously published HotSpot Wizard method (43). Selected mutations are evaluated by FoldX and the stabilizing ones are listed as candidate mutations for the engineering of multiple-point mutant.

Phase 3: Design of thermostable protein

In total, three protein designs are provided by FireProt strategy. The first design includes only the mutations from energy-based approach, the second contains the mutations suggested by the evolution-based approach and the third is the combination of both. Naturally, because of potentially antagonistic effects between individual mutations, we cannot combine individual mutations blindly.

To avoid possible clashes, FireProt strategy is trying to minimize antagonistic effects by utilizing Rosetta. In the first step, all pairs of single-point mutations within the range of 10 Å are evaluated separately for energy- and evolution-based approach. Once change in free energy is obtained for all residue pairs, FireProt starts to introduce them into the multiple-point mutant in the order based on their predicted stability, excluding the mutations that are colliding with already included mutations. Algorithm stops once there are no mutations left or the stabilizing effect of analyzed pair drops below defined threshold.

Upon the completion of previous step, procedure is repeated this time considering only the pairs between the mutations chosen for the construction of energy- and evolution-based mutants. Finally, structures of all three mutants are modeled using the Rosetta protocol 16.

DESCRIPTION OF THE WEB SERVER

Input

The only required input to the web server is a tertiary structure of the protein of interest, provided either as a PDB ID or a user-defined PDB file. The user can then choose a predefined biological unit generated by the MakeMultimer tool or manually select chains for which the calculation should be performed. The calculations can be configured in either basic or advanced mode.

In the basic mode, user is allowed to change the setting of BLAST search and alignment construction. The advanced mode expands the list of modifiable parameters by the ones connected with: (i) the identification of consensus residues by majority and frequency ratio approach, (ii) the thresholds used by FoldX and Rosetta prediction tools and (iii) the decision threshold employed in the consensual analysis of correlated positions. Advanced mode allows expert users to fine-tune the parameters of calculation according to studied systems. However, the presented default values are optimized to provide reliable results for most of the systems and we therefore do not advice their change in the general scenarios.

Output

Upon submission, a unique identifier is assigned to each job to track the calculation and the 'Results browser' informs the user about the status of the individual steps in the FireProt workflow (Figure 2B). Once the job is finished, users can either directly download the results in the .zip archive or navigate themselves into the 'Results page' for further analysis. The 'Results page' is intuitively organized into several panels as described below.

Protein visualization. The wild-type and the mutant structure is interactively visualized in the web browser (Figure 2D) utilizing the Jsmol applet (http://wiki.jmol.org/index.php/JSmol). Users can switch between different protein visualization styles and also highlight selected amino acids in the protein structure. Residues that were included into energy-based mutant are colored in orange, evolution-based mutations are in blue and all other residues are in gray. User selected residues that were not part of any mutant are underlined in red.

Mutant overview. The 'Mutant overview' panel is organized into four tabs (Figure 2A). The first three tabs provide information about mutations included into combined, energy-based and evolution-based mutant. The checkbox, allowing users to visualize the chosen residues in Jsmol applet, can be found in each row together with all data relevant for a given computational approach. The last tab contains the list of all residues in the wild-type structure. While 'wild-type' tab is active, the wild-type structure is visualized in Jsmol applet instead of the mutated one and the user is allowed to introduce user-defined mutations into multiple-point mutant via the 'plus' icon in the last column.

General information. The 'FireProt protocol design' panel provides users with general information about the target protein and the designs constructed by FireProt strategy, such as a number of mutations and estimated change in free energy (Figure 2C).

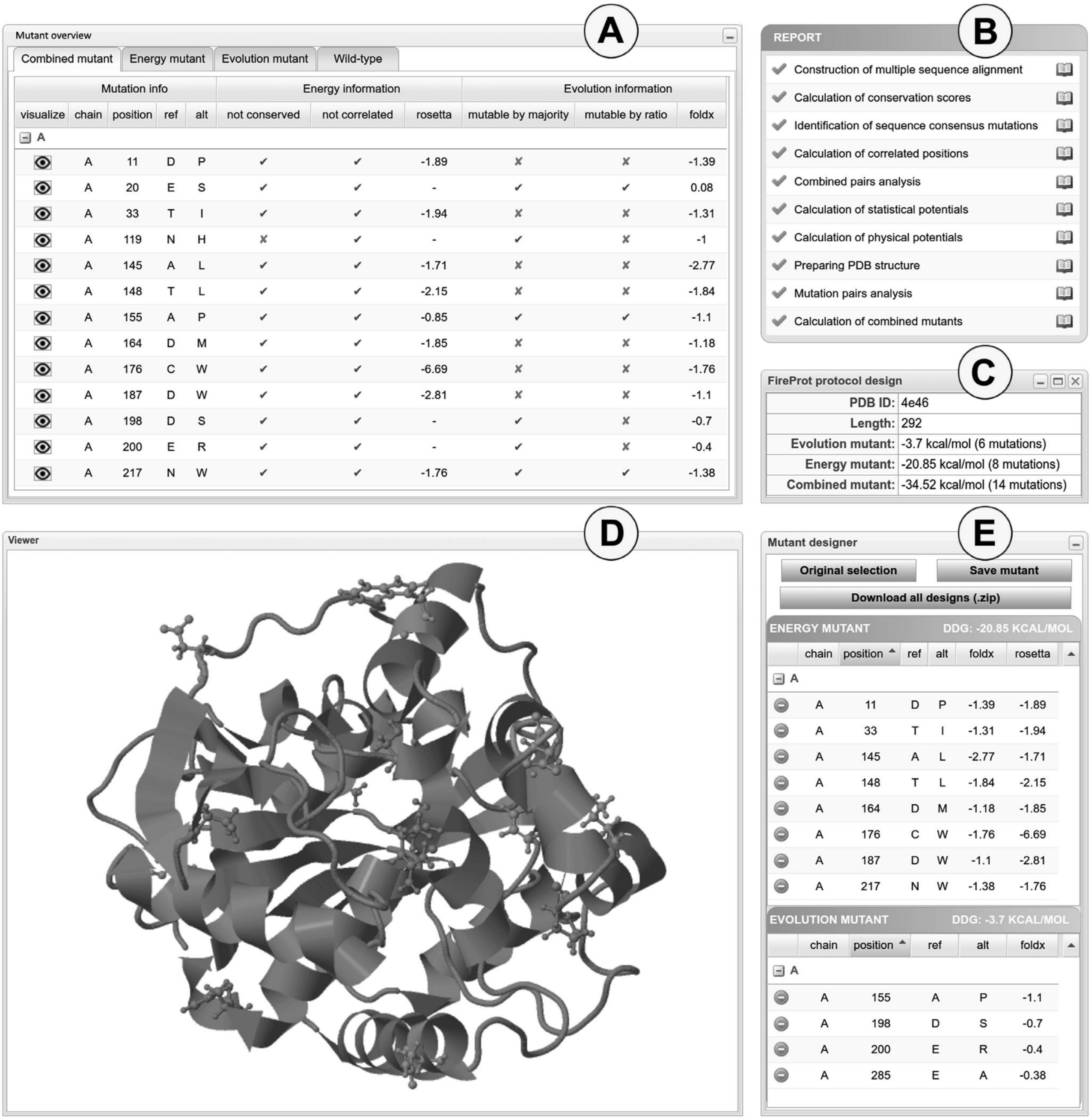
Mutant designer. The 'Mutant designer' panel allows the user to design own multiple-point mutant by managing mutations divided into energy- and evolution-based subset. If all mutations in the subset have their predicted energy values assigned, a total change in Gibbs free energy is immediately estimated assuming simple additivity. Users can also generate an amino acid sequence from the designed multiple-point mutant that combines mutations included into energy- and evolution-based subsets. All prepared designs can be downloaded in one .zip archive (Figure 2E).

EXPERIMENTAL VALIDATION

The original FireProt strategy was experimentally verified with three proteins (haloalkane dehalogenase DhaA, PDB ID 4E46; γ-hexachlorocyclohexane dehydrochlorinase LinA, PDB ID 3A76; and fibroblast growth factor 2, PDB ID 4OEE) and provided respective stabilization of proteins $\Delta T_\mathrm{m} = 25$, 21 and 15°C (Table 1). The original protocol was modified to enable fully automated calculation at the reasonable time, while maintaining high prediction accuracy (Supplementary Table S6). Prediction of eight multiple-point mutants using this modified protocol was validated using the data of FRESCO (44) and identified mutations were compared with another online protein stabilization tool PROSS (17). FireProt and PROSS showed similar predictive power, correctly identifying 29 and 20 potentially stabilizing positions, respectively (Supplementary Table S7).

CONCLUSIONS AND OUTLOOK

FireProt is a web server that provides users with a one-stop-shop solution for the design of thermostable multiple-point mutant proteins. In comparison with the standalone FireProt strategy (16), all default parameters and computational protocols were optimized to increase the calculation speed, while maintaining the prediction accuracy. The designs produced by the FireProt workflow were experimentally verified and thus users can obtain highly reliable thermostable proteins with minimal experimental effort. The server is complemented by an easy-to-use graphical interface that allows users to interactively analyze individual mutations selected as a part of energy- or evolution-based approach together with the ability to design their own multiple-point mutants on top of our robust strategy.



Mutation info					Energy information			Evolution information		
visualize	chain	position	ref	alt	not conserved	not correlated	rosetta	mutable by majority	mutable by ratio	foldx
⊟ A										
👁	A	11	D	P	✓	✓	-1.89	✗	✗	-1.39
👁	A	20	E	S	✓	✓	-	✓	✓	0.08
👁	A	33	T	I	✓	✓	-1.94	✗	✗	-1.31
👁	A	119	N	H	✗	✓	-	✓	✗	-1
👁	A	145	A	L	✓	✓	-1.71	✗	✗	-2.77
👁	A	148	T	L	✓	✓	-2.15	✗	✗	-1.84
👁	A	155	A	P	✓	✓	-0.85	✓	✓	-1.1
👁	A	164	D	M	✓	✓	-1.85	✗	✗	-1.18
👁	A	176	C	W	✓	✓	-6.69	✗	✗	-1.76
👁	A	187	D	W	✓	✓	-2.81	✗	✗	-1.1
👁	A	198	D	S	✓	✓	-	✓	✗	-0.7
👁	A	200	E	R	✓	✓	-	✓	✗	-0.4
👁	A	217	N	W	✓	✓	-1.76	✓	✓	-1.38

- ✓ Construction of multiple sequence alignment
- ✓ Calculation of conservation scores
- ✓ Identification of sequence consensus mutations
- ✓ Calculation of correlated positions
- ✓ Combined pairs analysis
- ✓ Calculation of statistical potentials
- ✓ Calculation of physical potentials
- ✓ Preparing PDB structure
- ✓ Mutation pairs analysis
- ✓ Calculation of combined mutants

PDB ID:	4e46
Length:	292
Evolution mutant:	-3.7 kcal/mol (6 mutations)
Energy mutant:	-20.85 kcal/mol (8 mutations)
Combined mutant:	-34.52 kcal/mol (14 mutations)

ENERGY MUTANT DDG: -20.85 KCAL/MOL

	chain	position ▲	ref	alt	foldx	rosetta
⊟ A						
⊖	A	11	D	P	-1.39	-1.89
⊖	A	33	T	I	-1.31	-1.94
⊖	A	145	A	L	-2.77	-1.71
⊖	A	148	T	L	-1.84	-2.15
⊖	A	164	D	M	-1.18	-1.85
⊖	A	176	C	W	-1.76	-6.69
⊖	A	187	D	W	-1.1	-2.81
⊖	A	217	N	W	-1.38	-1.76

EVOLUTION MUTANT DDG: -3.7 KCAL/MOL

	chain	position ▲	ref	alt	foldx
⊟ A					
⊖	A	155	A	P	-1.1
⊖	A	198	D	S	-0.7
⊖	A	200	E	R	-0.4
⊖	A	285	E	A	-0.38

Figure 2. FireProt's graphical user interface showing the results obtained for the haloalkane dehalogenase DhaA (PDB ID: 4e46). (**A**) The 'Mutant overview' panel provides a list of mutations introduced into protein structure. (**B**) The 'Report' panel shows the status of calculation in the individual steps of the computational pipeline. (**C**) The 'Protocol design' panel provides general information about FireProt designs. (**D**) The JSmol ´Viewer´ allows interactive visualization of the protein. (**E**) The 'Mutant designer' panel enables manual adjustment of a new combined mutant.

Table 1. Experimental validation of FireProt strategy

Protein PDB ID	Energy-based mutations	Evolution-based mutations	ΔT_{m} [°C]
4E46	8	3	+25
3A76	4	3	+21
4OEE	4	2	+15

The automation of the whole procedure makes the process of the design of thermostable proteins accessible to users without any prior expertise in bioinformatics since it eliminates the need to select, install and evaluate tools, optimize their parameters, and interpret intermediate results. However, the energy-based approach of the FireProt strategy depends on the quality of provided protein structure and therefore the prediction accuracy might be compromised in the case of low-resolution structures or homology models.

In the future, we plan to implement new strategies such as a design based on the analysis of correlated positions that would contribute to the construction of the final combined mutant, elimination of highly flexible regions and introduction of disulfide bridges. Also, we plan to equip FireProt with several new filters, e.g. exclusion of the amino acids located in the close neighborhoods of the active sites or the ones participating in oligomerization.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank the developers of PROSS Adi Goldenzweig and Dr Sarel Fleishman (Weizmann Institute of Science, Israel) for providing the data set for validation purposes. Computational resources were supplied by the Ministry of Education, Youth and Sports of the Czech Republic under the Projects CESNET (Project No. LM2015042) and CERIT-Scientific Cloud (Project No. LM2015085) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

FUNDING

This research and open access charge was funded by Ministry of Education, Youth and Sports of the Czech Republic from the National Programme of Sustainability II [LQ1605, LO1214]; European Regional Development Fund [ELIXIR-CZ LM2015047]; Grant Academy of the Czech Republic [16-06096S]; Brno University Technology [FIT-S-17-3964].

Conflict of interest statement. None declared.

REFERENCES

1. Modarres,H.P., Mofrad,M.R. and Sanati-Nezhad,A. (2016) Protein thermostability engineering. *RSC Adv.*, **6**, 115252–115270.
2. Ferdjani,S., Ionita,M., Roy,B., Dion,M., Djeghaba,Z., Rabiller,C. and Tellier,C. (2011) Correlation between thermostability and stability of glycosidases in ionic liquid. *Biotechnol. Lett.*, **33**, 1215–1219.
3. Wijma,H.J., Floor,R.J. and Janssen,D.B. (2013) Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr. Opin. Struct. Biol.*, **23**, 588–594.
4. Gao,D., Narasimhan,D.L., Macdonald,J., Brim,R., Ko,M.C., Landry,D.W., Woods,J.H., Sunahara,R.K. and Zhan,C.G. (2009) Thermostable variants of cocaine esterase for long-time protection against cocaine toxicity. *Mol. Pharmacol.*, **75**, 318–323.
5. Polizzi,K.M., Bommarius,A.S., Broering,J.M. and Chaparro-Riggers,J.F. (2007) Stability of biocatalysts. *Curr. Opin. Chem. Biol.*, **11**, 220–225.
6. Gray,K.A., Richardson,T.H., Kretz,K., Short,J.M., Bartnek,F., Knowles,R., Kan,L., Swanson,P.E. and Robertson,D.E. (2001) Rapid evolution of reversible denaturation and elevated melting temperature in a microbial haloalkane dehalogenase. *Adv. Synth. Catal.*, **343**, 607–617.
7. Folkman,L., Stantic,B., Sattar,A. and Zhou,Y. (2016) EASE-MM: sequence-based prediction of mutation-induced stability changes with feature-based multiple models. *J. Mol. Biol.*, **428**, 1394–1405.
8. Capriotti,E., Fariselli,P. and Casadio,R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, W306–W310.
9. Pires,D.E., Ascher,D.B. and Blundell,T.L. (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335–342.
10. Kellogg,E.H., Leaver-Fay,A. and Baker,D. (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins*, **79**, 830–838.
11. Yin,S., Ding,F. and Dokholyan,N.V. (2007) Modeling backbone flexibility improves protein stability estimation. *Structure*, **15**, 1567–1576.
12. Dehouck,Y., Grosfils,A., Folch,B., Gilis,D., Bogaerts,P. and Rooman,M. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, **25**, 2537–2543.
13. Guerois,R., Nielsen,J.E. and Serrano,L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.
14. Gumulya,Y. and Reetz,M.T. (2011) Enhancing the thermal robustness of an enzyme by directed evolution: least favorable starting points and inferior mutants can map superior evolutionary pathways. *ChemBioChem*, **12**, 2502–2510.
15. Bommarius,A.S. and Paye,M.F. (2013) Stabilizing biocatalysts. *Chem. Soc. Rev.*, **42**, 6534–6565.
16. Bednar,D., Beerens,K., Sebestova,E., Bendl,J., Khare,S., Chaloupkova,R., Prokop,Z., Brezovsky,J., Baker,D. and Damborsky,J. (2015) FireProt: energy- and evolution-based computational design of thermostable multiple-point mutants. *PLoS Comput. Biol.*, **11**, e1004556.
17. Goldenzweig,A., Goldsmith,M., Hill,S.E., Gertman,O., Laurino,P., Ashani,Y., Dym,O., Unger,T., Albeck,S., Prilusky,J. *et al.* (2016) Automated structure- and sequence-based design of proteins for high bacterial expression and stability. *Mol. Cell*, **63**, 337–346.
18. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
19. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B. and Wu,C.H. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.
20. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
21. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Soding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.*, **7**, 539.
22. Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.
23. Kass,I. and Horovitz,A. (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins*, **48**, 611–617.
24. Korber,B.T.M., Farber,R.M., Wolpert,D.H. and Lapedes,A.S. (1993) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 7176–7180.
25. Lee,B.C. and Kim,D. (2009) A new method for revealing correlated mutations under the structural and functional constraints in proteins. *Bioinformatics*, **25**, 2506–2513.
26. Weigt,M., White,R.A., Szurmant,H., Hoch,J.A. and Hwa,T. (2008) Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 67–72.
27. Lockless,S.W. and Ranganathan,R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
28. Dekker,J.P., Fodor,A., Aldrich,R.W. and Yellen,G. (2004) A perturbation-based method for calculating explicit likelihood of

evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, **20**, 1565–1572.

29. Valencia,A. (2003) Multiple sequence alignments as tools for protein structure and function prediction. *Compar. Funct. Genomics*, **4**, 424–427.
30. Benner,S.A. and Gerloff,D. (1991) Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. *Adv. Enzyme Regul.*, **31**, 121–181.
31. Brenner,S. (1988) The molecular evolution of genes and proteins: a tale of two serines. *Nature*, **334**, 528–530.
32. Cooperman,B.S., Baykov,A.A. and Lahti,R. (1992) Evolutionary conservation of the active site of soluble inorganic pyrophosphatase. *Trends Biochem. Sci.*, **17**, 262–266.
33. Howell,N. (1989) Evolutionary conservation of protein regions in the protonmotive cytochrome b and their possible roles in redox catalysis. *J. Mol. Evol.*, **29**, 157–169.
34. Gobel,U., Sander,C., Schneider,R. and Valencia,A. (1994) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317.
35. Neher,E. (1994) How frequent are correlated changes in families of protein sequences? *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 98–102.
36. Taylor,W.R. and Hatrick,K. (1994) Compensating changes in protein multiple sequence alignments. *Protein Eng.*, **7**, 341–348.
37. Bava,K.A., Gromiha,M.M., Uedaira,H., Kitajima,K. and Sarai,A. (2004) ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.*, **32**, D120–D121.
38. Pucci,F., Bourgeas,R. and Rooman,M. (2016) Predicting protein thermal stability changes upon point mutations using statistical potentials: introducing HoTMuSiC. *Scientific Rep.*, **6**, 23257.
39. Amin,N., Liu,A.D., Ramer,S., Aehle,W., Meijer,D., Metin,M., Wong,S., Gualfetti,P. and Schellenberger,V. (2004) Construction of stabilized proteins by combinatorial consensus mutagenesis. *Protein Eng. Des. Select.*, **17**, 787–793.
40. Lehmann,M., Loch,C., Middendorf,A., Studer,D., Lassen,S.F., Pasamontes,L., van Loon,A.P. and Wyss,M. (2002) The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng.*, **15**, 403–411.
41. Pey,A.L., Rodriguez-Larrea,D., Bomke,S., Dammers,S., Godoy-Ruiz,R., Garcia-Mira,M.M. and Sanchez-Ruiz,J.M. (2008) Engineering proteins with tunable thermodynamic and kinetic stabilities. *Proteins*, **71**, 165–174.
42. Sullivan,B.J., Nguyen,T., Durani,V., Mathur,D., Rojas,S., Thomas,M., Syu,T. and Magliery,T.J. (2012) Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability. *J. Mol. Biol.*, **420**, 384–399.
43. Bendl,J., Stourac,J., Sebestova,E., Vavra,O., Musil,M., Brezovsky,J. and Damborsky,J. (2016) HotSpot wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering. *Nucleic Acids Res.*, **44**, W479–W487.
44. Floor,R.J.1, Wijma,H.J., Colpa,D.I., Ramos-Silva,A., Jekel,P.A., Szymański,W., Feringa,B.L., Marrink,S.J. and Janssen,D.B. (2014) Computational library design for increasing haloalkane dehalogenase stability. *ChemBioChem*, **15**, 1660–1672.

B Original publication II: FireProt^{ASR}

MUSIL M, KHAN RT, BEIER A, STOURAC J, KONEGGER H, DAMBORSKY J, BEDNAR D. FireProt-ASR: Web Server for Fully Automated Ancestral Sequence Reconstruction. *Briefings in bioinformatics.* 2020, 0, 1-11. *(available in early access)*

Author participation: 60%

Journal impact factor: 8.990 (Q1)

OXFORD

Briefings in Bioinformatics, 00(00), 2020, 1–11

doi: 10.1093/bib/bbaa337
Problem Solving Protocol

FireProt^{ASR}: A Web Server for Fully Automated Ancestral Sequence Reconstruction

Milos Musil, Rayyan Tariq Khan, Andy Beier, Jan Stourac, Hannes Konegger, Jiri Damborsky and David Bednar

Corresponding author: David Bednar, Department of Experimental Biology and RECETOX, Loschmidt Laboratories, Faculty of Science, Masaryk University, 611 37 Brno, Czech Republic. Tel.: +420 605 143 394. E-mail: davidbednar1208@gmail.com

Abstract

There is a great interest in increasing proteins' stability to widen their usability in numerous biomedical and biotechnological applications. However, native proteins cannot usually withstand the harsh industrial environment, since they are evolved to function under mild conditions. Ancestral sequence reconstruction is a well-established method for deducing the evolutionary history of genes. Besides its applicability to discover the most probable evolutionary ancestors of the modern proteins, ancestral sequence reconstruction has proven to be a useful approach for the design of highly stable proteins. Recently, several computational tools were developed, which make the ancestral reconstruction algorithms accessible to the community, while leaving the most crucial steps of the preparation of the input data on users' side. FireProt^{ASR} aims to overcome this obstacle by constructing a fully automated workflow, allowing even the unexperienced users to obtain ancestral sequences based on a sequence query as the only input. FireProt^{ASR} is complemented with an interactive, easy-to-use web interface and is freely available at https://loschmidt.chemi.muni.cz/fireprotasr/.

Key words: ancestral sequence reconstruction; ancestral enzymes; evolution; phylogeny-based analysis; protein stability

Introduction

Proteins are widely used in numerous biomedical and biotechnological applications. Native proteins have mainly evolved under mild intracellular conditions [1]. Therefore, their applicability is often limited in the harsh industrial environments characterized by inhospitable temperature, extreme pH, high pressure or the presence of organic co-solvents. As a result, there is a continuous interest in increasing protein stability. New approaches in the field of protein engineering, such as fluorescence-activated cell sorting and microfluidics, have widened the throughput of

Milos Musil is a bioinformatician at Loschmidt Laboratories, Masaryk University. He carries out his doctoral thesis at the Brno University of Technology, designing and implementing bioinformatics tools for the automatized design of stable proteins.
Rayyan Tariq Khan is a doctoral candidate at Loschmidt Laboratories, Masaryk University. His work is focused on ancestral sequence reconstruction, experimental evolution and design of bioinformatics tools.
Andy Beier is doing his postdoc in protein engineering at the Loschmidt Laboratories, Masaryk University. His main responsibilities are the mutagenesis, production and detailed biochemical and biophysical characterization of enzymes and the development of an ultra-high-throughput assay for dehalogenases.
Jan Stourac is a bioinformatician at Loschmidt Laboratories, Masaryk University. He carries out his doctoral thesis at the Faculty of Informatics, Masaryk University, focusing on the design and implementation of the bioinformatics tools for the analysis of protein tunnels.
Hannes Konegger is a former MSCA fellow, examined the biochemical basis of evolutionary- and structure-based protein engineering methods. He recently turned his field of interest towards microbial ecology and applied bioenergetics.
Jiri Damborsky is a professor of Biochemistry at Masaryk University and group leader at International Clinical Research Center at St. Ann's Teaching Hospital. He is interested in development of software tools for computational enzyme design.
David Bednar is a team leader at Loschmidt Laboratories, Masaryk University. His team is focused on molecular modelling, bioinformatics and development of software for protein engineering.

Submitted: 14 August 2020; **Received (in revised form):** 12 October 2020

directed evolution experiments. However, saturation mutagenesis of all positions and systematic re-combinations of many single-point mutations of the protein of interest is often out of reach.

In the past decades, various computational methods were designed to unburden costly and laborious experimental work by narrowing down the search space for potential stabilizing mutations. Most of those methods can be assigned to one of the three categories: (i) machine learning, (ii) force-field-based predictions and (iii) molecular evolution. Each category has its advantages and shortcomings [2]. Machine-learning methods are able to unearth hidden features and dependencies overreaching the current state of expert knowledge, while still struggling with the insufficient size, quality and diversity of the experimental data, essential for training and validation of statistically significant models. Force-field-based approaches are a robust solution for the prediction of protein stability; however, they rely on the high-resolution protein structures that are available for only a small fraction of the known proteins. Evolution-based approaches do not suffer from these limitations due to the rapid growth of the sequence databases. However, this continuous growth widens the search space and increases noise in the data, requiring laborious and time-demanding manual corrections from the side of the user with expert knowledge of the system of interest. Inexperienced user may not therefore utilize evolution-based methods effectively to obtain accurate and reliable results.

The two most widely used evolution-based methods for stability engineering are ancestral sequence reconstruction (ASR) and consensus design. Both methods start with the multiple-sequence alignment (MSA) of the set of relevant homolog sequences. Consensus design relies on the simple analysis of the conservation of the amino acids on the individual positions in the sequence alignment. As a result, it cannot account for the coevolution of the residues located in the sites responsible for the protein's activity [3] and is utilized mostly as a part of the hybrid workflows [4, 5]. In comparison, ASR goes much further by also considering evolutionary information depicted by the phylogenetic tree. This inclusion of the evolutionary distances inscribed into the phylogenetic tree is mostly negligent at the positions with low Shannon entropy; however, the discrepancies grow stronger with noisy MSA [6]. ASR is a probabilistic method that explores the deep evolutionary history of homolog sequences to reassemble protein's evolutionary trajectory [7]. ASR is able to unearth sequences of the long-extinct genes and organisms from which the current ones evolved and is, therefore, an invaluable tool in the field of evolutionary biology [8, 9]. ASR has also been shown to be a very effective strategy not only for thermostability engineering [10, 11], but also for improving other protein's characteristics such as specificity [12], activity, or expression [13]. Furthermore, ASR was previously proven to be an effective strategy for the stabilization of prokaryotic proteins [10, 11], as well as for the improvement of significantly more complex eukaryotic proteins such as cytochrome P450 [14, 15]. Two main algorithms, maximum-likelihood [16, 17] (ML) and Bayesian inference [18] (BI) were designed to infer ancestral sequence from MSA and phylogenetic tree. Many tools were built over the years to make those algorithms accessible to the community. However, the requirement of the MSA of carefully selected homologs and the rooted phylogenetic tree are still huge limiting steps for the general use of ASR method by the non-expert users.

FireProt^{ASR} addresses those limitations by introducing one-stop-shop solution for the ancestral sequence reconstruction. It covers all steps of ancestral inference including search for homolog sequences, selection of the biologically relevant subset of the sequences, construction of the multiple-sequence alignment, construction and rooting of the phylogenetic tree and finally the ancestral inference with the use of ML. Our computational workflow is fully automated and removes the need for extensive expert knowledge of the system of interest as well as employed bioinformatics tools. Furthermore, a novel algorithm based on the localized weighted back-to-consensus analysis was utilized to resolve an issue of the ancestral gaps reconstruction. Assembled workflow and developed web server were thoroughly validated using: (i) *in-house* laboratory experiments, (ii) detailed comparison with three previously published studies and (iii) a large number of proteins representing structurally and functionally different families. FireProt^{ASR} does not require installation and settings of any software packages as the method is implemented in the interactive web interface freely available at: https://loschmidt.chemi.muni.cz/fireprotasr/.

Methods

Workflow description

The basic workflow of the FireProt^{ASR} method is outlined in Figure 1. To infer ancestral sequences representing all ancestral nodes of the evolutionary tree in a fully automated way, a set of biologically relevant homologous sequences must be collected from genomic databases and reduced to a suitable size (Phase 1). With the initial set of homologous sequences in hand, several state-of-the-art methods are utilized to construct a multiple-sequence alignment and a phylogenetic tree, which are then used to support the inference of ancestral nodes and reconstruction of ancestral gaps (Phase 2). The FireProt^{ASR} workflow requires no user intervention beyond providing a query sequence and (in the case of enzymes) selecting catalytic residues used to identify a biologically relevant set of homologous sequences. However, it is also possible to start a calculation with a user-defined initial set of homologous sequences, MSA, or even a phylogenetic tree instead of a single sequence, thus skipping the first phase of the calculation.

Phase 1: collection of the initial set of homologous sequences

The query sequence of the target protein in plain text or FASTA format is the only input required from the side of the user. Once the query sequence has been uploaded to the server and checked for validity, searches for the catalytic residues are performed automatically using SwissProt [19] and the Catalytic Site Atlas [20]. The user can also specify the catalytic residues by themselves if no/incorrect catalytic residues are found. Once the catalytic residues and query sequence have been specified, an in-house tool called EnzymeMiner [21] is used to collect an initial set of homologous sequences. EnzymeMiner first performs two rounds of PSI-BLAST [22] against the NCBI nr database [23] and then filters out all sequences lacking the designated catalytic residues, thereby ensuring the biological relevance of the remaining homologs. EnzymeMiner searches can yield up to tens of thousands of homologous sequences for large families. If no catalytic residues were selected or provided by the user, BLAST [24] will be used instead of EnzymeMiner, to obtain an initial set of homologous sequences with potentially lower quality.

Next, the FireProt^{ASR} reduces the set of homologous sequences to the required number, which is set to 150 sequences by default. Several filters are applied during this process. First, all homologs

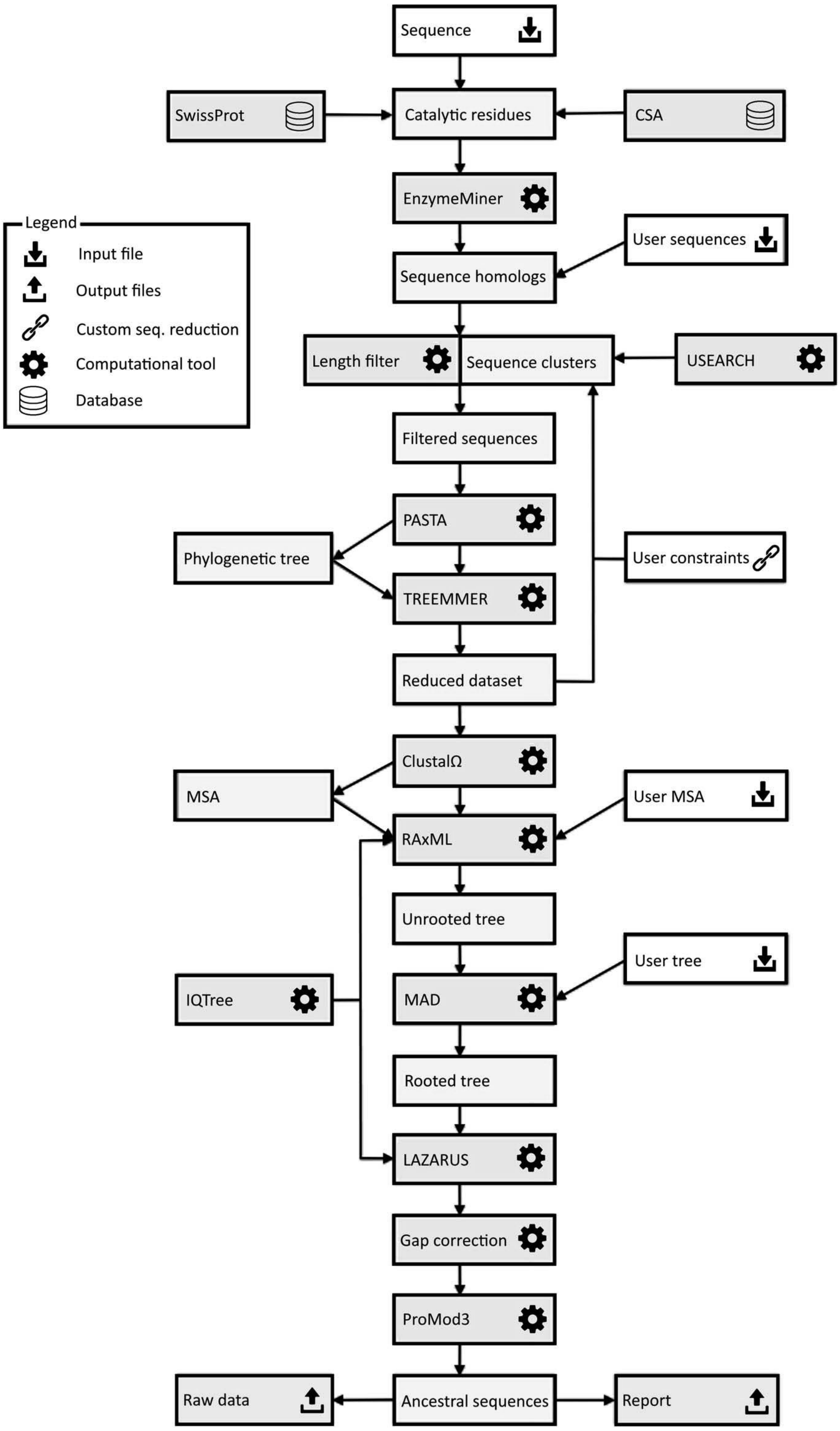


Figure 1. Workflow diagram for the FireProt^{ASR} method. The workflow has two phases: (1) collection of the initial set of homologous sequences and (2) ancestral sequence reconstruction. Colour coding: yellow denotes intermediate results and blue denotes computational tools. Grey and green denote inputs and outputs of the calculations, respectively.

with sequence lengths 20% higher or lower than that of the query sequence are excluded from the initial set. This sequence length normalization is done to remove potential outliers that could lead to a construction of a noisy MSA with many gaps. Second, all homologs whose sequence identity to the query falls outside a certain range are removed from the initial set. By default, the upper and lower similarity limits are set to 90 and 30%, respectively. This step ensures that the phylogenetic tree is unbiased towards the query sequence while removing distant homologs that would degrade the quality of the sequence alignment. Third, USEARCH [25] is used to cluster the remaining sequences with 90% sequence identity, and a single sequence is randomly selected from each cluster.

Applying these filters produces a diverse set containing hundreds to thousands of homologous sequences. An initial phylogenetic tree is quickly constructed with the PASTA software suite [26], using MAFFT [27] and the swift neighbour-joining algorithm implemented in FastTree 2.0 [28]. The resulting phylogenetic tree is then forwarded to Treemmer [29], which iteratively prunes leaves from the input tree until a specific number of leaves remains, while minimizing the loss of genetic diversity. The pruned tree is then displayed to the user via the interactive user interface, allowing the user to choose to exclude selected branches or even whole subtrees of the phylogenetic tree from further calculations.

Phase 2: ancestral sequence reconstruction

In the second phase, the ancestral sequences are inferred from the initial set of up to 150 homologs approved by the user. To begin with, a new MSA is constructed from the reduced set of homologous sequences. For this task, ClustalΩ [30] is utilized by default, but other methods will be available in upcoming versions of FireProt^{ASR}. For inference of the final phylogenetic tree, the best-fitting evolutionary matrix must be selected. This is done using one of the modules of the IQTREE package [31]. Alternatively, if the user prefers a specific evolutionary matrix for the biological system of interest, the appropriate model and all the relevant modifiers can be specified manually when setting up the calculation.

The evolutionary model and its parameter settings along with the MSA are then forwarded into RAxML [17], which is used to construct a robust phylogenetic tree. By default, fifty bootstraps are performed at the start of the maximum-likelihood search; since no outgroup is provided, the resulting phylogenetic tree is unrooted. Automated outgroup sequence selection is not straightforward, especially for prokaryotic proteins due to the high frequency of horizontal gene transfers. Rooting of the tree is therefore performed using a minimal ancestor deviation algorithm, which was shown to achieve comparable levels of accuracy to outgroup rooting in trees describing the evolution of eukaryotes, and to surpass both outgroup and midpoint rooting in the case of prokaryotes [32].

The MSA constructed with ClustalΩ, the selected evolutionary model, and the rooted phylogenetic tree from RAxML are used as inputs for the Lazarus method [33], which is implemented using the PAML software package [16]. The Lazarus method was re-implemented for FireProt^{ASR} to enable calculations to be performed without specifying outgroup. Consequently, ancestral sequences of all ancestral nodes are parsed from their posterior probabilities and provided to users in separate files in FASTA format. Additionally, BLASTp [24] is used to search for a template in the PDB database [34], and a model structure of the query sequence is constructed by homology modelling using the ProMod3 program [35]. This model is shown in the web interface to allow users to visualize the differences between the query sequence and the selected ancestor.

Finally, due to the large number of undesirable ancestral gaps inserted into ancestral sequences by Lazarus, a novel algorithm for ancestral gap reconstruction was designed for use in FireProt^{ASR}. This algorithm is based on the principle of localized weighted back-to-consensus because consensus analysis has proven to be an effective approach for increasing proteins' thermal stability [36–38]. To begin with, each terminal node of the phylogenetic tree is assigned a binary vector of length equal to the length of the corresponding sequence in the MSA. Each position in this vector is assigned a value of −1 or 1, indicating the presence of a gap or standard amino acid, respectively, at the corresponding position of the relevant sequence. On moving from the terminals towards the root of the tree, the probability of a gap in ancestral node A_n at position i is calculated as $A_{n_i} = \frac{A_{k_i} * t_1 * + A_{l_i} * t_2}{t_1 + t_2}$, where A_k, A_l are the child nodes of A_n and t_1, t_2 are the evolutionary distances between A_n and its child nodes. Taking t_3 to be the evolutionary distance between A_n and its parental node, its value can be updated based on the values of t_1 and t_2 as follows: $t_{3\_new} = t_3 + \frac{t_1 + t_2}{2}$. This new value is computed before proceeding with the calculation for the parental node; its use increases the relative impact of well-branched subtrees and therefore limits the impact of lone sequences and small subtrees compared to that of well-represented ones. Finally, ancestral sequences are reconstructed based on the scores in the corresponding vector. Positions with values lower than 0 are assigned as gaps, and the remaining amino acids are selected based on their posterior probabilities as estimated by Lazarus. The nature of inconclusive positions with scores in the interval $<-0.1, 0.1>$ is determined based on the frequencies of gaps in the global alignment and the state of the parental node. To include the ancestral gap, frequencies of gaps in the global alignment should reach over 60%, or over 40% if the ancestral gap is present in the parental node sequence. The model case for a single position in the sequence alignment is shown in Figure 2.

Experimental validation

The workflow was experimentally validated using haloalkane dehalogenases as a model enzyme. This enzyme was selected as a typical representative of the α/β superfamily, counting over 100 000 proteins. The sequence of the haloalkane dehalogenase DhaA (UniProt ID P0A3G2) was used as the sole input for the calculation. Six different ancestral sequences were selected and experimentally characterized.

Chemicals and growth media

1-bromobutane and LB medium were purchased from Sigma-Aldrich Co. (St. Louis, MO, USA). IPTG was purchased from Duchefa Biochemie B.V. (Haarlem, The Netherlands). All chemicals used in this work were of analytical grade.

Expression in *Escherichia coli* BL21 (DE3)

Escherichia coli Dh5α cells were obtained from Invitrogen and *Escherichia coli* BL21 (DE3) from New England Biolabs. The genes for the ancestral dehalogenases were synthesized and subcloned into the expression vector pET21b. The generated plasmids were transformed into chemo-competent *E.coli* BL21 (DE3) cells. Obtained colonies were used to prepare precultures by inoculation into 10 ml of LB medium (with 100 μg/ml ampicillin) followed by overnight incubation at 37°C and 180 rpm. For expression of each variant, 1 l of LB medium supplemented with 100 μg/ml ampicillin was inoculated with 5 mL of the appropriate pre-culture (1/200). The flasks were incubated at 37°C and 180 rpm until OD_{600} 0.6–0.8 was reached, then incubated at 20°C for 30 min. β-D-1-thiogalactopyranoside (IPTG, 0.2 mM) was then added for induction, and the culture was incubated at 20°C and 180 rpm overnight. Finally, the culture was harvested by centrifugation at 4500 × g, 4°C for 15 min, after which the cell pellets were frozen at −80°C until further use.

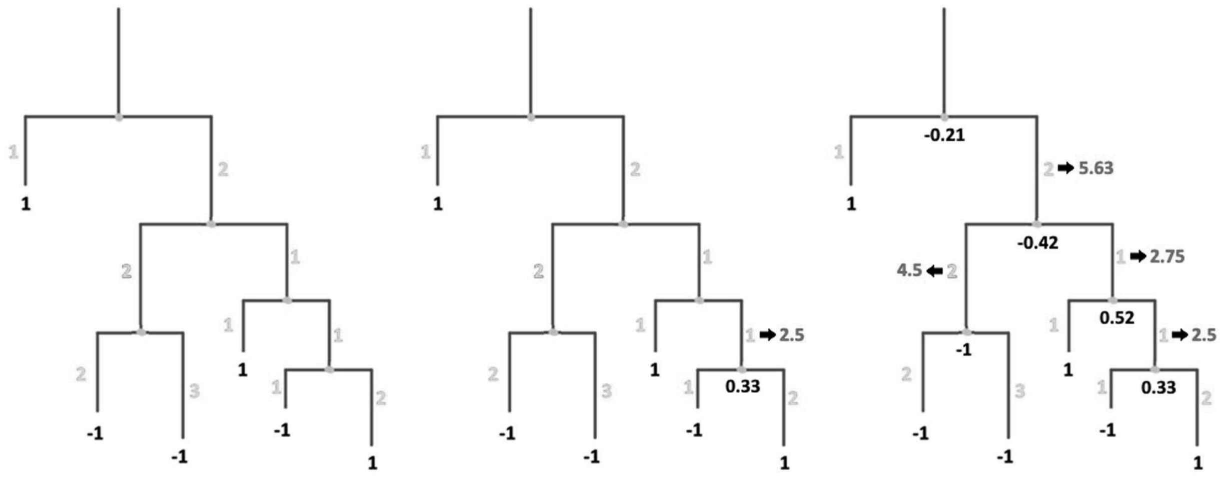


Figure 2. Ancestral gaps reconstruction algorithm. Green colour denotes the initial branch lengths of the phylogenetic tree. Black numbers indicate the values of the vectors of the terminal and the ancestral sequences at the given position in the multiple sequence alignment. Red values show the modified branch lengths that are updated after the calculation of the underlying ancestral node.

Protein purification

The cell pellets were suspended in 50 ml of equilibration buffer (20 mM phosphate buffer pH 7.5 containing 0.5 M NaCl and 10 mM imidazole) and disrupted by sonication with a Hielscher UP200S ultrasonic processor (Hielscher, Germany) four times for 4 min each. Disrupted cells were centrifuged at 13 000 × *g* and 4°C for 1 h (Laborzentrifugen, Germany). The crude extract was then collected, filtered and loaded onto a Ni-NTA Superflow Cartridge (Qiagen, Germany) in equilibration buffer. Unbound and weakly bound proteins were washed out using increasing imidazole concentrations. The target enzyme was eluted with purification buffer containing 300 mM of imidazole. The eluted protein was dialyzed three times overnight against 50 mM of phosphate buffer (pH 7.5), after which its purity was checked by SDS–polyacrylamide gel electrophoresis (SDS–PAGE). About, 15% polyacrylamide gels were stained with Instant Blue (Fluka, Switzerland). Protein concentrations were determined by NanoDrop (Sigma-Aldrich, USA). The enzymes were lyophilized using a vacuum pump system for long-term storage.

Circular dichroism (CD) spectroscopy

CD spectra were recorded at 20°C using a spectropolarimeter Chirascan (Applied Photophysics, United Kingdom). Data were collected from 190 to 260 nm, at 100 nm/min with a 1-s response time and 1-nm bandwidth using a 0.1-cm quartz cuvette. Each spectrum shown is the average of five individual scans and was corrected for absorbance caused by the buffer. Collected CD data were expressed in terms of the mean residue ellipticity (Θ_{MRE}), which was calculated using the equation:

$$\Theta_{\mathrm{MRE}} = \frac{\Theta_{\mathrm{obs}} \cdot M_{\mathrm{w}} \cdot 100}{n \cdot c \cdot l}$$

where Θ_{obs} is the observed ellipticity in degrees, M_{w} is the protein molecular weight, n is number of residues, l is the cell path length, c is the protein concentration (0.2 mg/ml) and the factor 100 originates from the conversion of the molecular weight to mg/dmol.

Thermal denaturation

Thermal unfolding was followed by monitoring the ellipticity at 224 nm over the temperature range of 20–94°C, with a resolution of 0.1°C at a heating rate of 1°C/min. Recorded thermal denaturation curves were roughly normalized to represent signal changes between approximately 1 and 0 and fitted to sigmoidal curves using Origin 6.1 (OriginLab Corporation, USA). The melting temperature (T_m) was evaluated as the midpoint of the normalized thermal transition.

Enzymatic haloalkane dehalogenase activity

Dehalogenation activity was assayed using the colorimetric method of Iwasaki *et al.* [49]. The release of halide ions was analyzed spectrophotometrically at 460 nm using an Eon microplate reader (BioTek, USA) after reaction with mercuric thiocyanate and ferric ammonium sulfate. The reactions were performed at 37°C in 25-ml Reacti Flasks closed with Mininert Valves. The reaction mixtures consisted of 10 ml 100 mM glycine buffer (pH 8.6) and 10 μl of the substrate 1-bromobutane. Reactions were initiated by adding the enzyme to a final concentration of 0.01 (DhaA 172Loc), 0.0065 (DhaA 172Glob), 0.0052 (DhaA 230Glob), 0.028 (DhaA 238Loc) or 0.014 mg/ml (DhaA 238Glob). Reactions were monitored by withdrawing 1 ml of samples from the reaction mixture after 0, 5, 10, 15, 20 and 30 min. The samples were immediately mixed with 0.1 ml of 35% nitric acid to stop the reaction. Dehalogenation activities were quantified as rates of product formation over time. Each activity was measured in three independent replicates.

Enzymatic luciferase activity

Luminescence activity measurements were performed with a FLUOstar OPTIMA Microplate reader (BMG Labtech, Germany) using coelenterazine as the substrate at 37°C. A 25 μl of sample of purified enzyme at a concentration of about 1 mg/ml was placed into a microtiter plate well. After baseline collection for 10 s, the luminescence reaction was initiated by adding 225 μl of 8.8 μM coelenterazine in reaction buffer (100 mM potassium phosphate buffer, pH 7.5). Luminescence was recorded for 72.5 s, and each sample was measured in at least three independent experiments. The areas of the resulting luminescence intensity peaks in relative luminescence units (RLU) were converted into values in units of RLU/mg/s.

Results

Web server input

The only required input to the web server is a query sequence of the target protein in plain text or FASTA format. Alternatively, one can upload a FASTA file containing an initial set of sequence homologs or a multiple sequence alignment (MSA). Rooted and unrooted phylogenetic trees in the standard Newick format can also be provided. When performing calculations in basic mode, only the table containing the essential residues is available to the user. Essential residues are identified automatically by searching in SwissProt [19] and mCSA [20]. However, the initial selection can be changed by the user. The default values and settings of individual computational tools are optimized to provide reliable results for most systems. Operating in advanced mode expands the list of modifiable parameters to include those related to: (i) the thresholds of the homolog identity filters and sequence clustering, (ii) selection of the evolutionary model and (iii) construction of the phylogenetic tree. Advanced mode allows experts to fine-tune the calculation's parameters based on the studied biological system, which may be useful when dealing with particularly small or large protein families.

Selection and reduction

Upon submission, a unique identifier is assigned to each job to track the calculation. The 'calculation browser' informs the user about the status of the individual steps in the ancestral sequence reconstruction workflow. Once the first phase of the job is finished, the initial phylogenetic tree is displayed to the user using a strongly updated adaptation of PhyloTree library (Figure 3A) [39], together with the table of removed sequences (Figure 3B). By clicking on the individual leaves of the phylogenetic tree, the user can exclude selected sequences from future calculations. Furthermore, whole subtrees can be removed by choosing this option in the menu of the selected ancestral node. The MSA of the homologous sequences can be also visualized by switching to the multiple sequence alignment tab. This mode is intended for the expert users with the greater knowledge of the system of interest as it allows for the removal of the noise and outliers from the initial set of homolog sequences. If the expert mode is utilized, it is recommended to exclude the sequences that do not share the function similar to the query protein or that cause a significant disturbance in the MSA.

Web server output

The calculation's progress can be tracked in the 'calculation browser' similarly to the selection step. Once finished, users can either download the results in the zipped archive directly from the calculation page or navigate to the 'Result page' for further analysis. The 'Result page' is organized into several panels allowing users to interactively visualize and design ancestral enzymes.

Protein visualization

The homology model of the query protein predicted by ProMod3 is interactively visualized in the web browser using the JSmol applet [40] (Figure 3D). Users can switch between different visualization styles such as backbone, wireframe or cartoon and change the quality of the visualized structure. It is also possible to visualize the differences between the query and the selected ancestral sequence on the modelled protein structure: substitutions and deletions are shown in blue and red, respectively, while insertions are indicated by regions between red and yellow residues.

Ancestral tree panel

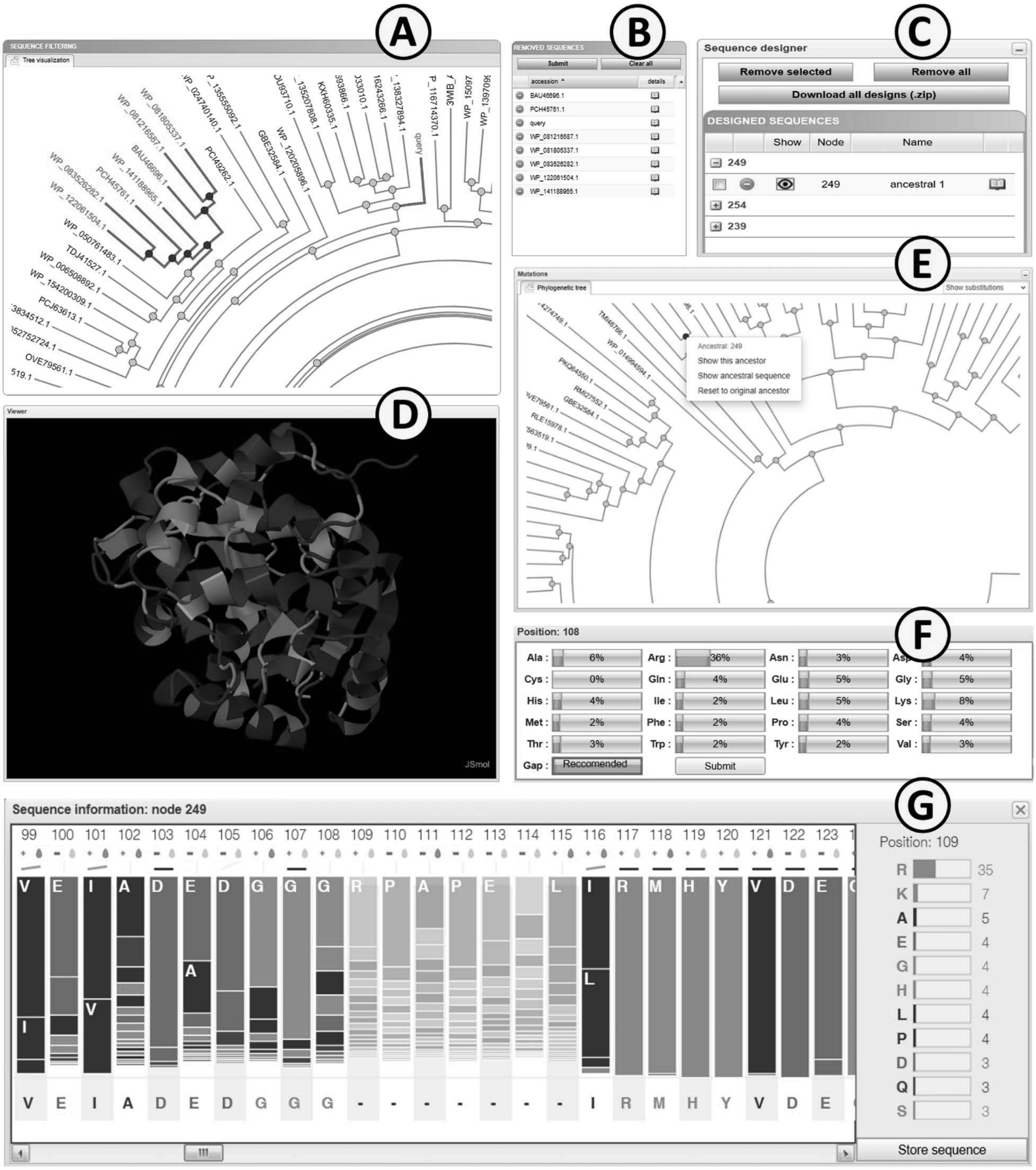
The 'ancestral panel' shows the final phylogenetic tree constructed by RAxML [17] along with further information about the precalculated ancestral sequences (Figure 3E). By selecting any of the ancestral nodes, it is possible to either (i) visualize the differences between a wild-type protein and the selected ancestor node on the protein structure or (ii) open a new window providing an overview of the posterior probabilities for individual amino acids in the sequence of the selected ancestor (Figure 3G). Posterior probabilities are shown in the bar-styled sequence logo together with the percentages for each considered amino acid, and each bar is expanded with information about the charge and hydrophobicity of the most probable amino acids. The bar representation was in part derived from the SequenceLogo library [41]. The user can edit the ancestral sequence and store it as a new user-defined ancestor (Figure 3F). This option is useful for the experts with more in-depth knowledge of the system of interest and allows to force some specific mutations, e.g., the mutations with the previously known effect on proteins stability, into the constructed ancestral sequence. It can also be used to bring some biological insight into the positions with noisy posterior probabilities. Furthermore, the ancestral sequences' MSA can be visualized in the multiple sequence alignment tab for further analysis.

Sequence designer

The 'Sequence designer' panel allows users to manage and edit user-defined ancestral sequences. Additionally, new sequences can be created by modifying existing custom ancestors (Figure 3C). Differences between the query sequence and custom ancestors can also be visualized on the protein structure in this panel. All prepared designs can be downloaded in one zipped archive together with the original ancestors and the structure prepared by homology modelling.

Web server experimental validation

In one of our previous studies, we have presented experimental characterizations of six inferred ancestral proteins from haloalkane dehalogenase subfamily II [10]. Relative to their contemporary counterparts, these ancestral proteins exhibited higher thermal stability (by 8–24°C), improved yields and broadened substrate specificity. Those ancestral sequences were reconstructed by clustering an initial set of homologous sequences that was reduced by inspection in the sequence-editing program BioEdit [42]. A multiple sequence alignment was then manually curated using a structure-guided alignment of eight proteins from HLD-II and poorly conserved regions were removed from the alignment. The topology of the phylogenetic tree was optimized by subtree pruning and re-grafting, and the tree's root was established using outgroup selected on the basis of expert judgement. Finally, the ancestral sequences and positioning of gaps were refined by manual inspection.



REMOVED SEQUENCES

Submit | Clear all

accession	details
BAU46696.1	📖
PCH45761.1	📖
query	📖
WP_081216587.1	📖
WP_081805337.1	📖
WP_083526282.1	📖
WP_122061504.1	📖
WP_141188965.1	📖

Sequence designer

Remove selected | Remove all

Download all designs (.zip)

DESIGNED SEQUENCES

		Show	Node	Name	
⊟ 249					
□	⊖	👁	249	ancestral 1	📖
⊞ 254					
⊞ 239					

Position: 108

Ala :	6%	Arg :	36%	Asn :	3%	Asp	4%
Cys :	0%	Gln :	4%	Glu :	5%	Gly :	5%
His :	4%	Ile :	2%	Leu :	5%	Lys :	8%
Met :	2%	Phe :	2%	Pro :	4%	Ser :	4%
Thr :	3%	Trp :	2%	Tyr :	2%	Val :	3%
Gap :	Reccomended	Submit					

Sequence information: node 249

99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123
V	E	I	A	D	E	D	G	G	G	-	-	-	-	-	-	-	I	R	M	H	Y	V	D	E

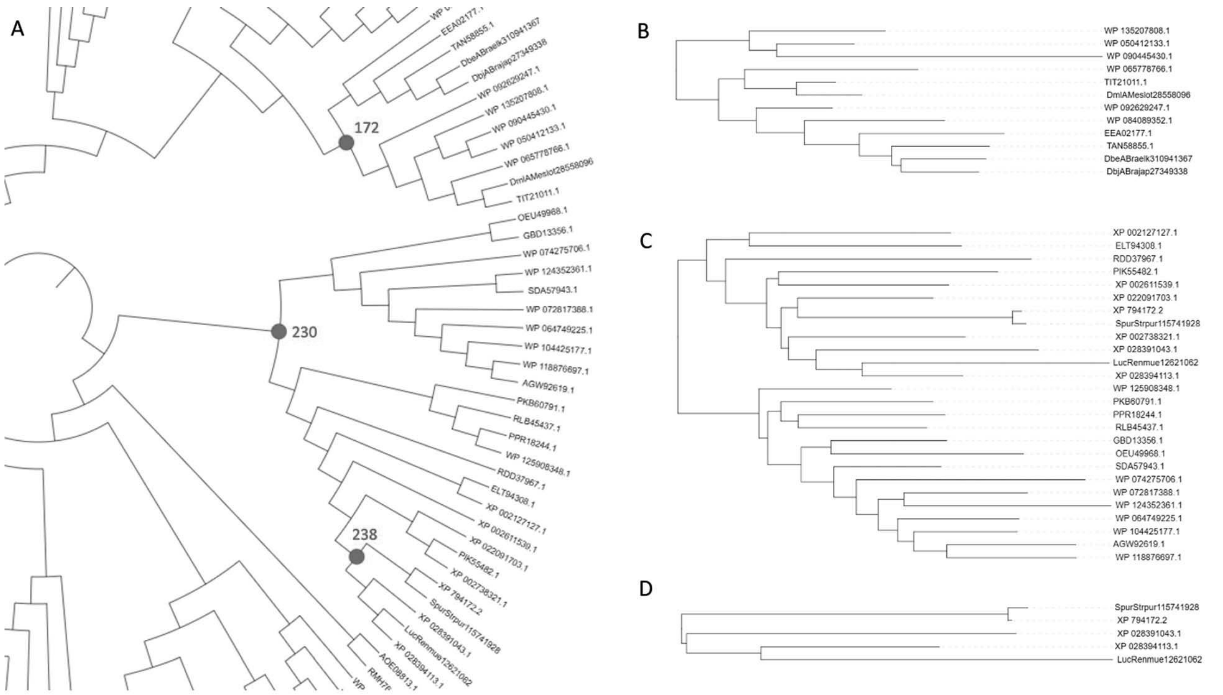
Position: 109

R	35
K	7
A	5
E	4
G	4
H	4
L	4
P	4
D	3
Q	3
S	3

Store sequence

Figure 3. The FireProtASR graphical user interface showing results obtained for the haloalkane dehalogenase DhaA (UniProt ID P0A3G2, PDB ID 4E46). (**A**) The sequence-filtering panel allows users to exclude selected branches from the calculation. (**B**) The reduction table shows the list of removed sequences. (**C**) The sequence designer allows users to download and edit ancestral sequences. (**D**) The JSmol viewer provides interactive protein visualization. (**E**) The mutations panel contains all designed ancestral sequences in the ancestral tree. (**F**) The edit window enables amino acid substitutions at individual positions. (**G**) The sequence information window shows detailed information on selected ancestral sequences.

As part of the validation of FireProtASR, we tried to replicate these results by using the sequence of haloalkane dehalogenase DhaA (UniProt ID P0A3G2) as the only input query. All steps of the calculation, including homologous sequence selection, multiple sequence alignment construction, phylogenetic rooting and ancestral reconstruction were carried out automatically. Three pairs of ancestral sequences were selected, each pair containing one 'global' and one 'local' ancestral node (Figure 4A). Global ancestor (Glob) represents ancestral sequence obtained directly from the fully automated workflow, while local ancestor



E

```
query        ---------------MSEIGTGFPFDPHYVEVLGERMHYVDVGPRD---GTPVLFLHGNPTSSYLWRNIIPHVAPSHRCIAPDLIGMGKSDK-PDLDYFFDDHVRYLDAFIEALGLEE-V  100
BTC_analysis ---------------MSEIGTGFPFDPHYVEVLGERMHYVDEGPRD---GTPVLFLHGNPTSSYLWRNIIPHVAPSHRCIAPDLIGFGKSDK-PDLDYFFDDHVRYLDAFIEALGLEE-T  100
230Global    ---------------MTSKTADMPHPRKRVSVLDSEMSYVDTGE-----GDPVVFLHGNPTSSYLWRNVIPHVEPLGRCLAPDLIGMGDSGKSPDGSYRFVDHARYLDAWFDALGLRENV  100
230Local     MSRITVILGHTMSSTSLTTSADVWHKRKRVDVLDSEMSYVDTGPGQ---GDTVVFLHGNPTSSYLWRNVIPHVEPLARCLAPDLIGMGRSGKLPNGSYRFVDHYRYLSAWFDAVDLPEKV  117
238Local     ------MVSHPQSTRPLITADEWWGKCKKINVLDSYMSYYDSEPHNLSGKNTVVFLHGNPTSSYLWRNIIPHVEPIARCLAPDLIGMGRSGKLGNGSYRFLDHYKYLSAWFDALKLPEKI  114
238Global    -----MTVSASQSSTSLVTADEWWGKCKRVNVLDSEMSYYDSDPHQTHSKNTVVFLHGNPTSSYLWRNIIPHVEPLARCLAPDLIGMGRSGKLPNGSYRFVDHYRYLSAWFDAVNLPEKI  115
172Local     ------------------MQEMTAVARSQVPVLDSTMAYREAGAPD---GPTVLFLHGNPTSSYIWRNIIPHVAPVARCIAPDLIGFGQSGK-PDIDYRFLDHVRYLDAFIETLGLED-F  97
172Global    ------------------MTAQFAVSLRHVPVLGSTMAYVEAGAPD---GPTVLFLHGNPTSSYIWRNIIPHVAPVARCIAPDLIGFGQSGK-PDIDYRFTDHVRYLDAFIETLGLSD-V  97
                                          : **.. * * :          *:**********:***:**** *  **:******:* *.*  : .* * ** :**.*::::: * :

query        VLVIHDWGSALGFHWAKRNPERVKGIACMEF-IRPIPTWDEWPE--FARETFQAFRTADVGRELIIDQNAFIEGALPKCVVRPLTEVEMDHYREPFLKPV-DREPLWRFPNELPIAGE-P  215
BTC_analysis VLVIHDWGSALGFHWAARNPERVKGIACMEF-IRPIPTWDEWPE--FARETFQAFRTADVGRELIIDQNAFIEGALPKCVVRPLTEVEMDHYREPFLKPV-DREPLWRFPNELPIAGE-P  215
230Global    TLVIHDWGSALGFHWAHRHPERVKGIVYMEAIVQPL-TWDDWPE--AAREIFQALRSP-AGEEMVLEKNLFVERILPGSILRKLSEEEMDAYRRPFAEPGESRRPTLTWPREIPIEGE-P  215
230Local     TLVIHDWGSGLGFHWAHRHPERVKAIVYMEALVATVKTWDDFPE--AARNIFQAMRSP-AGEEMVLEKNLFVERLLPSSIMRELSEEEMDAYRKPFKEPGESRRPTLTWPREIPIEGDGP  234
238Local     TIVCHDWGSGLGFHWCHEHPDRVKAIVHMESVVQPIRSWETWPD--IARDIFQALRSD-AGEEMILKKNFFIERLLPSSIMRKLRPEEMDAYREPFKEKGESRRPMLTWPREIPIEGEGP  231
238Global    TIVCHDWGSGLGFHWCHQHPDRVKAIVHMESIVGTVPSWDEFPE--IARDIFQALRSD-AGEEMVLKKNLFVERLLPSSIMRKLSEEEMDAYREPFKEPGESRRPTLTWPREIPIEGEGP  232
172Local     FLVAQDWGTALAFHLAARHPQRVLGLAFMEF-IRPMPSWDDFHQRPQARELFRQFRTPGEGEKLILEDNVFIEKVLPNSVLRRLSDEEMAAYRAPFPTPA-SRKPILRLPRELPIAGE-P  214
172Global    FLVAQDWGTALAFHLAARHPERVRGLAFMEF-IRPMPTWDDFHQTPQARETFRKFRTPGVGEKLILEDNVFIERVLPGSVVRKLSDEEMNAYRAPFPTPA-SRKPVWRFPNELPIAGE-P  214
              :* :***:.*.** . .:*:** .:. **  :  : :*: : :   **: *: :*:   *.::::..* *:*  **  .::* *   **  ** **      .*.*    *.*:** *: *

query        ANIVALVEAYMNWLHQS-PVPKLLFWGTPGVLIPPAEAARLAESLPNCKTVDIGPGLHYLQEDNPDLIGSEIARWLPALHHH-----  296
BTC_analysis ANIVALVEAYMNWLHQS-PVPKLLFWGTPGVLIPPAEAARLAESLPNCKTVDIGPGLHYLQEDNPDLIGSEIARWLPALHHH-----  296
230Global    ADVVEIVEAYSQWLQQS-SIPKLFINAEPGAILT-GRQREFCRTWPNQQEVTV-PGIHFIQEDSPDEIGQAIADFLSSLRS------  293
230Local     QDVVEIVEAYSKWLSQSESIPKLFINADPGFFST-GAIRKFCKNWPNQQTVTV-KGLHFLQEDSPDEIGEAIREFLEELLSKESSK-  318
238Local     EDVLKIVSDYHAFMKASEDLPKLFIEADPGFFSK--AITEIAKQWPNTKVVKV-KGLHFLQEDSPDEIGNAIKEFLEKVFKTNS---  312
238Global    QDVVEIVNAYHKWLKQSENIPKLFIEADPGFFSK--AIRKVTKNWPNHKTVKV-KGLHFLQEDSPDEIGKAIKEFLEELLK------  310
172Local     ADVYALLEADHRALRAS-TYPKLLFSGDPGALISPAVAREFAAGLKNCRLVNLGPGAHYLQEDHPDRIGREIADWIAELAAANAAAE  300
172Global    ADVYAMLEAAHAALAAS-TYPKLLFSATPGALISPALAQQFAASLKHCKLVNLGPGAHYLQEDHPDAIGREIADWIAEIAAERAA--  298
              :: ::.     :  *    ***:: . ** :         ..      : : * :  * *::*** ** **  *  ::  :
```

Figure 4. Results provided by the FireProtASR workflow using haloalkane dehalogenase DhaA as an input query. (**A**) Phylogenetic tree of the HLD-II constructed by the FireProtASR strategy with indicated three global ancestors reconstructed within this study. (**B**) Phylogenetic tree for the local ancestor of the ancestral node 172. (**C**) Phylogenetic tree for the local ancestor of ancestral node 230. (**D**) Phylogenetic tree for the local ancestor of ancestral node 238. (**E**) Multiple sequence alignment comparing the query sequence with the suggested ancestral sequences and the result of the back-to-consensus analysis.

(Loc) was constructed by carrying out FireProt^{ASR} workflow for a second time using only the sequences included in the subtree beneath the selected ancestral node. Local ancestor therefore represents a root of a phylogenetic tree constructed from only the sequences most relevant to the selected ancestral node. Node 238 (Figure 4D) is an ancestor of only five leaves and was selected because of its close proximity to luciferase and dehalogenase, providing a fair comparison to the previously published ancestors. Similar comparison can be also achieved with node 172 (Figure 4B), having several stable dehalogenases in its progeny. Finally, node 230 (Figure 4C) was highlighted as a more distant ancestor of both luciferase and dehalogenase. No pruning, curation or re-grafting was performed in the process. Selected ancestral sequences were then subjected to the experimental validation. MSA of the query protein, selected ancestors, and the sequence provided by executing back-to-consensus analysis is attached in Figure 4E.

Although the selected sequences have high implied sequence similarity (92–97%) with the inferred ancestors, experimental validation showed that the ancestors' thermal stability was 20–26°C higher than that of wild-type DhaA (Table 1). The ancestral proteins also exhibited high expressibility, solubility, yields and

Table 1. Characteristics of reconstructed and experimentally characterized ancestral haloalkane dehalogenases

Protein code	Expression (% of total protein)	Solubility (%)	Yield (mg/l)	T_m (°C)	HLD act. (μmol/mg·s)	LUC act. (RLU/mg·s)
DhaA wt	17	83.1	91.1	50.56 ± 2.4	0.032 ± 0.0059	n.a.
DhaA 172Loc	23	85.5	74.9	71.60 ± 0.7	0.038 ± 0.0002	1.41 ± 0.26
DhaA 172Glob	21	65.2	88.2	70.04 ± 1.5	0.061 ± 0.0045	n.a.
DhaA 230Loc	20	n.d.	n.d.	n.d.	n.d.	n.d.
DhaA 230Glob	23	84.8	108.5	72.14 ± 0.4	0.061 ± 0.0118	n.a.
DhaA 238Loc	23	63.2	74.9	70.36 ± 0.6	0.014 ± 0.0021	353.5 ± 14.58
DhaA 238Glob	19	83.3	94.4	76.19 ± 0.2	0.030 ± 0.0012	3.18 ± 0.33

Notes: DhaA, haloalkane dehalogenase from *Rhodococcus rhodochrous* NCIMB 13064; wt, wild type; Loc, ancestral protein inferred from local alignment; Glob, ancestral protein inferred from global alignment; T_m, melting temperature; HLD act., haloalkane dehalogenases activity; LUC act., luciferase activity; n.d., not determined due to poor solubility of this protein; n.a., not active under tested conditions.

catalytic activity. Moreover, inference based on both haloalkane dehalogenases and luciferases led to the discovery of the very interesting enzyme ancHLD-Rluc, which exhibits dual dehalogenase and monooxygenase activity. This experimental validation provides direct experimental evidence of the good functionality and reliability of the fully automated version of FireProt^{ASR}.

Additionally, results obtained using FireProt^{ASR} were thoroughly and quantitatively compared to three previously published experimental studies. For this purpose, Euclidean distance [43], and the Subtree prune and regraft distance [44] were calculated to compare the trees obtained from the FireProt^{ASR} and published literature. The two trees were also graphically compared using the Jaccard index utilizing ColorBrewer [45] scheme. Detailed comparison of all three experimental studies with the results produced by FireProt^{ASR} server is attached in Supplementary Data 1–3, available online at https://academic.oup.com/bib. Finally, the robustness and reliability of the FireProt^{ASR} server was tested using 60 diverse proteins from various protein families (see Supplementary Data 4 available online at https://academic.oup.com/bib).

Discussion

ASR has been shown to be a very effective strategy for the protein thermostability engineering and as such was implemented in various computational tools using maximum-likelihood (FastML [46], RaxML [17], Ancestors [47]) or Bayesian inference (HandAlign [48], MrBayes [18]) methods. However, a significant limitation of those methods is that they require complex input data to be uploaded by the users. Those requirements are reaching from a simple set of homolog sequences to the MSA or even rooted phylogenetic tree, leaving the most crucial and laborious parts of the calculation in the hands of the users. Non-expert users without the deep knowledge of the bioinformatics tools and the system of interest are therefore hindered from the successful use of the ASR method.

FireProt^{ASR} is a web server that aims to provide users with one-stop-shop solution for the ancestral sequence reconstruction. FireProt^{ASR} requires minimal input from the users, and the whole calculation can be processed from a single protein sequence, set of homolog sequences, MSA and phylogenetic tree. All steps of the calculation, including the search for biologically relevant homolog sequences, dataset reduction and the ancestral reconstruction are automated. Moreover, a novel algorithm based on localized weighted back-to-consensus analysis is implemented to resolve an issue with ancestral gap reconstruction. FireProt^{ASR} web server is also complemented by an easy-to-use web interface that allows users to interactively analyze sequences of the individual ancestral nodes together with the ability to design their own ancestral sequences based on the posterior probabilities of the existing nodes.

The robustness and reliability of the results produced by the FireProt^{ASR} workflow was evaluated by experimental characterization of six ancestral sequences of haloalkane dehalogenase from HLD-II subfamily. With the exception of the local variant of the ancestral node 230, all designed ancestral sequences are soluble and also retain high expressibility and yields on the levels comparable to the DhaA wild type. However, the thermal stability has increased by over 20°C and global variants 172 and 230 have also increased the HLD activity by two-fold. Increase in HLD activity cannot be observed in the constructed local variants that utilize smaller subsets of homolog sequences, and thus only a limited amount of evolutionary information. This would encourage the usage of the global variants for the design of highly stable and active proteins. However, more focused view using a localized variants of the ancestral nodes can provide some useful results as can be observed in the local variant of the node 238 that shows both dehalogenase and monooxygenase activity. High thermal stabilization was also achieved in those variants.

Finally, the results provided by the FireProt^{ASR} web server are consistent with the designs presented in the published literature as the fully automatized designs obtained by FireProt^{ASR} method maintain high sequence similarity (>90%) with the manually designed and curated ancestors. Finally, the comprehensive analysis of approximately 60 different proteins from various protein families have proven the robustness and reliability of the presented method.

The full automation of the FireProt^{ASR} method eliminates the need to select, install and evaluate individual tools, optimize their parameters and interpret intermediate results. Together with its general applicability for a wide range of protein families, FireProt^{ASR} makes the procedure of ancestral reconstruction accessible to the users without any prior expertise in bioinformatics, and the intuitive web interface allows for a further analysis utilizing both sequence and structural information.

Key Points

- FireProt^{ASR} is a web service for a fully automated design of stable proteins using ancestral sequence reconstruction and is accompanied by an interactive and easy-to-use interface.

- FireProt^{ASR} allows users to utilize ancestral reconstruction without prior knowledge of the necessary bioinformatics tools and the biological system.
- The robustness and reliability of the FireProt^{ASR} method were thoroughly tested by both laboratory experiments and by comparing predictions with the results published in scientific literature.
- Laboratory characterization of the ancestral designs showed up to 26°C improvement in thermostability and some of the proteins poses even dual catalytic activity.

Data availability

All data validating the robustness and accuracy of our service are available in the Supplementary materials 1-4. Web service and tutorials are freely available at https://loschmidt.chemi.muni.cz/fireprotasr/.

Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics*.

Funding

Czech Ministry of Education (CZ.02.1.01/0.0/0.0/17_043/0009632, 857560, 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 (720776, 814418, 722610) and Marie Curie@MUNI (CZ.02.2.69/0.0/0.0/19_074/0012727). Computational resources were supplied by the project 'e-Infrastruktura CZ' (LM2018140) and ELIXIR (LM2015047). This project has received funding from the European Union's Horizon 2020 research and Innovation programme. The article reflects the author's view and the Agency is not responsible for any use that may be made of the information it contains.

References

1. Modarres HP, Mofrad MR, Sanati-Nezhad A. Protein thermostability engineering. *RSC Adv* 2016;**6**:115252–70.
2. Musil M, Konegger H, Hon J, *et al.* Computational design of stable and soluble biocatalysts. *ACS Catal* 2019;**9**: 1033–54.
3. Hendrikse NM, Charpentier G, Nordling E, *et al.* Ancestral diterpene cyclases show increased thermostability and substrate acceptance. *FEBS J* 2018;**285**:4660–73.
4. Musil M, Stourac J, Bendl J, *et al.* FireProt: web server for automated design of thermostable proteins. *Nucleic Acids Res* 2017;**45**:W393–9.
5. Goldenzweig A, Goldsmith M, Hill SE, *et al.* Automated structure- and sequence-based design of proteins for high bacterial expression and stability. *Mol Cell* 2016;**63**:337–46.
6. Risso VA, Gavira JA, Gaucher EA, *et al.* Phenotypic comparisons of consensus variants versus laboratory resurrections of Precambrian proteins. *Proteins Struct Funct Bioinforma* 2014;**82**:887–96.
7. Hochberg GKA, Thornton JW. Reconstructing ancient proteins to understand the causes of structure and function. *Annu Rev Biophys* 2017;**46**:247–69.
8. Bickelmann C, Morrow JM, Du J, *et al.* The molecular origin and evolution of dim-light vision in mammals. *Evolution* 2015;**69**:2995–3003.
9. Hobbs JK, Prentice EJ, Groussin M, *et al.* Reconstructed ancestral enzymes impose a fitness cost upon modern bacteria despite exhibiting favourable biochemical properties. *J Mol Evol* 2015;**81**:110–20.
10. Babkova P, Sebestova E, Brezovsky J, *et al.* Ancestral haloalkane dehalogenases show robustness and unique substrate specificity. *Chembiochem* 2017;**18**:1448–56.
11. Risso VA, Gavira JA, Sanchez-Ruiz JM. Thermostable and promiscuous Precambrian proteins. *Environ Microbiol* 2014;**16**:1485–9.
12. Wheeler LC, Lim SA, Marqusee S, *et al.* The thermostability and specificity of ancient proteins. *Curr Opin Struct Biol* 2016;**38**:37–43.
13. Zakas PM, Brown HC, Knight K, *et al.* Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. *Nat Biotechnol* 2017;**35**:35–7.
14. Bart AG, Harris KL, Gillam EMJ, *et al.* Structure of an ancestral mammalian family 1B1 cytochrome P450 with increased thermostability. *J Biol Chem* 2020;**295**:5640–53.
15. Gumulya Y, Baek J-M, Wun S-J, *et al.* Engineering highly functional thermostable proteins using ancestral sequence reconstruction. *Nat Catal* 2018;**1**:878–88.
16. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 2007;**24**:1586–91.
17. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinforma Oxf Engl* 2014;**30**:1312–3.
18. Ronquist F, Teslenko M, van der Mark P, *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 2012;**61**:539–42.
19. Boeckmann B, Bairoch A, Apweiler R, *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003;**31**:365–70.
20. Ribeiro AJM, Holiday GL. Mechanism and catalytic site atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res* 2018;**46**:618–23.
21. Hon J, Borko S, Stourac J, *et al.* EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. *Nucleic Acids Res* 2020;**48**:W104–9.
22. Altschul SF, Madden TL, Schaffer AA, *et al.* PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**17**:3389–402.
23. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2016;**44**:D7–19.
24. Camacho C, Coulouris G, Avagyan V, *et al.* BLAST+: architecture and applications. *BMC Bioinforma* 2009;**10**:421.
25. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinforma Oxf Engl* 2010;**26**:2460–1.
26. Mirarab S, Nguyen N, Guo S, *et al.* PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *J Comput Biol* 2015;**22**:377–86.
27. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;**30**:772–80.
28. Price MN, Dehal PS, Ap A. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;**5**(3):e9490. doi: 10.1371/journal.pone.0009490.

29. Menardo F, Loiseau C, Brites D, *et al*. Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics* 2018;**19**:164.
30. Sievers F, Wilm A, Dineen D, *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal omega. *Mol Syst Biol* 2011;**7**:539.
31. Nguyen L-T, Schmidt HA, von Haeseler A, *et al*. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;**32**:268–74.
32. Tria FDK, Landan G, Dagan T. Phylogenetic rooting using minimal ancestor deviation. *Nat Ecol Evol* 2017;**1**:193.
33. Hanson-Smith V, Kolaczkowski B, Thornton JW. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Mol Biol Evol* 2010;**27**:1988–99.
34. Sussman JL, Lin D, Jiang J, *et al*. Protein data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr D Biol Crystallogr* 1998;**54**:1078–84.
35. Biasini M, Schmidt T, Bienert S, *et al*. OpenStructure: an integrated software framework for computational structural biology. *Acta Crystallogr D Biol Crystallogr* 2013;**69**: 701–9.
36. Amin N, Liu AD, Ramer S, *et al*. Construction of stabilized proteins by combinatorial consensus mutagenesis. *Protein Eng Des Sel* 2004;**17**:787–93.
37. Lehmann M, Loch C, Middendorf A, *et al*. The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng Des Sel* 2002;**15**: 403–11.
38. Sullivan BJ, Nguyen T, Durani V, *et al*. Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability. *J Mol Biol* 2012;**420**:384–99.
39. Shank SD, Weaver S, Kosakovsky Pond SL. Phylotree.Js—a JavaScript library for application development and interactive data visualization in phylogenetics. *BMC Bioinformatics* 2018;**19**:276.
40. Hanson RM, Prilusky J, Renjian Z, *et al*. JSmol and the next-generation web-based representation of 3D molecular structure as applied to Proteopedia. *Isr J Chem* 2013;**53**:207–16.
41. Maguire E, Rocca-Serra P, Sansone S-A, *et al*. Redesigning the sequence logo with glyph-based approaches to aid interpretation. In: *Proceedings of EuroVis 2014 Short Paper, IEEE Visualization and Graphics Technical Committee (IEEE VGTC)* 2014.
42. Kirmani S. A user friendly approach for design and economic analysis of standalone SPV system. *Smart Grid Renew Energy* 2015;**06**:67–74.
43. de Vienne DM, Aguileta G, Ollier S. Euclidean nature of phylogenetic distance matrices. *Syst Biol* 2011;**60**:826–32.
44. Bordewich M, Semple C. On the computational complexity of the rooted subtree prune and Regraft distance. *Ann Comb* 2005;**8**:409–23.
45. Harrower M, Brewer CA. ColorBrewer.Org: an online tool for selecting colour schemes for maps. *Cartogr J* 2003;**40**:27–37.
46. Ashkenazy H, Penn O, Doron-Faigenboim A, *et al*. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res* 2012;**40**:W580–4.
47. Diallo AB, Makarenkov V, Blanchette M. Ancestors 1.0: a web server for ancestral sequence reconstruction. *Bioinforma Oxf Engl* 2010;**26**:130–1.
48. Westesson O, Barquist L, Holmes I. HandAlign: Bayesian multiple sequence alignment, phylogeny and ancestral reconstruction. *Bioinforma Oxf Engl* 2012;**28**:1170–1.
49. Iwasaki I, Utsumi S, Ozawa T. New colorimetric determination of chloride using mercuric thiocyanate and ferric ion. *Bulletin of the Chemical Society of Japan* 1952;**25**(3):226.

C Original publication III: FireProt$^{\mathsf{DB}}$

STOURAC J, DUBRAVA J, MUSIL M, HORACKOVA J, DAMBORSKY J, MAZURENKO S, BEDNAR D. FireProt-DB: Database of Manually Curated Protein Stability Data. *Nucleic Acids Research.* 2021, 49, D319-D324.

Author participation: 20%
Journal impact factor: 11.501 (Q1)

Published online 9 November 2020 *Nucleic Acids Research, 2021, Vol. 49, Database issue* ***D319–D324***
doi: 10.1093/nar/gkaa981

FireProt^{DB}: database of manually curated protein stability data

Jan Stourac[1,2,†], Juraj Dubrava[1,3,†], Milos Musil[1,2,3], Jana Horackova[1], Jiri Damborsky[1,2], Stanislav Mazurenko[1,*] and David Bednar[1,2,*]

[1]Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Masaryk University, Brno, Czech Republic, [2]International Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic and [3]Department of Information Systems, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic

Received August 14, 2020; Revised September 18, 2020; Editorial Decision October 09, 2020; Accepted October 12, 2020

ABSTRACT

The majority of naturally occurring proteins have evolved to function under mild conditions inside the living organisms. One of the critical obstacles for the use of proteins in biotechnological applications is their insufficient stability at elevated temperatures or in the presence of salts. Since experimental screening for stabilizing mutations is typically laborious and expensive, *in silico* predictors are often used for narrowing down the mutational landscape. The recent advances in machine learning and artificial intelligence further facilitate the development of such computational tools. However, the accuracy of these predictors strongly depends on the quality and amount of data used for training and testing, which have often been reported as the current bottleneck of the approach. To address this problem, we present a novel database of experimental thermostability data for single-point mutants FireProt^{DB}. The database combines the published datasets, data extracted manually from the recent literature, and the data collected in our laboratory. Its user interface is designed to facilitate both types of the expected use: (i) the interactive explorations of individual entries on the level of a protein or mutation and (ii) the construction of highly customized and machine learning-friendly datasets using advanced searching and filtering. The database is freely available at https://loschmidt.chemi.muni.cz/fireprotdb.

INTRODUCTION

Proteins play essential roles in many biotechnological and biomedical applications, where they are often subjected to extreme environments, e.g. elevated temperatures or the presence of various salts. However, naturally occurring proteins have mostly evolved to function in the mild environmental conditions, and therefore their applicability is limited in the industrial applications. For this reason, protein engineers generally aim to improve protein stability, and thermostability is one of their primary targets (1) as it is correlated with serum survival time (2), half-life (3), expression yield (4) and activity in the presence of denaturants (5). A reliable assessment of the effect of a mutation on protein stability is often performed experimentally. Extensive experimental screening, however, is slow and costly, prompting the use of *in silico* approaches for the pre-selection of promising mutations. These methods are usually based on one of the three principles: (i) free energy calculations, (ii) phylogenetics or (iii) machine learning. With the recent advances in artificial intelligence, tool developers increasingly resort to the third group of methods. However, the accuracy of the machine learning-based predictors is still severely limited by the lack of high-quality data (6). Experimental characterizations are usually not capable of producing large amounts of data, and the majority of these measurements are scattered in the scientific literature. Thus, there is a strong demand for systematic collection, validation, and organization of such data in a database.

Two attempts have been made to establish a systematic and extensive collection of thermostability data so far. The first and largest database is the Thermodynamic Database for Proteins and Mutants–ProTherm (7). It was first released in 1999 with the aim to collect experimentally determined thermodynamic parameters for wild-type proteins

To whom correspondence should be addressed. Tel: +420 605 143 394; Email: davidbednar1208@gmail.com
Correspondence may also be addressed to S. Mazurenko. Email: mazurenko@mail.muni.cz

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Website address: https://loschmidt.chemi.muni.cz/fireprotdb.

and their mutants from the published literature. Its latest version contains >25 000 entries from 740 proteins, and it serves as the primary source of protein stability data for the development of new predictors. However, ProTherm was last updated in 2013 so the database is already out-of-date. Moreover, several critical issues have been reported, such as inaccurate annotations or wrong signs of values (6,8–10). This makes ProTherm even more difficult to use as time-demanding manual filtering and validation steps are required to confirm the values in the original articles. This manual filtering led to the construction of many different, often overlapping, subsets with corrected values and occasionally new data. Some of these derivative datasets were deposited to the VariBench database (11) without any attempts to reintegrate the changes into ProTherm or create an improved database. This changed in 2018 when ProtaBank (12) was released. This database aims to collect a wide range of protein engineering data such as thermostability, activity, expression, binding and several others. The developers imported all the data from ProTherm, yet they did not seem to perform any manual curation. Therefore, the critical issues listed above were not resolved. And while ProtaBank enriched the ProTherm data with recent experimental studies, the database does not offer any advanced searching and filtering capabilities, at least in its non-commercial version. This makes the data extraction and processing tedious by necessitating many manual steps and hindering the application of such data-driven methods as machine learning.

To overcome these limitations, we established the FireProt^{DB} database that holds manually curated thermostability data for single-point mutants. The database contains the data available in ProTherm, ProtaBank, and our extensive manual literature search. Its user-friendly interface allows easy and interactive browsing through the experimental data and provides links to the corresponding UniProt and PDB entries. Moreover, advanced searching and filtering capabilities, the ability to download the data in a simple table format, and meticulous labelling of data entries used for training and testing of published tools prompt the further application of machine learning.

MATERIALS AND METHODS

Database architecture and data model

The top-level entity of the FireProt^{DB} database is a unique protein sequence entry with the assigned UniProt ID (13). Protein sequences were preferred to structures due to the broader availability of the former. Each sequence is a string of amino acids in specified positions. Multiple mutations can be assigned to a single position, and each mutation can be evaluated by multiple measurements and derived values. The measurements represent the experimental values of the Gibbs free energy changes upon mutation ($\Delta\Delta G$) or changes in melting temperatures (ΔT_m). The derived values stand for averages or medians of multiple measurements for a particular mutation. Each measurement is also accompanied by a curation flag that indicates whether the value was manually validated against the original publication to guarantee its correctness. Furthermore, each measurement and derived value can be assigned to multiple published datasets to promote accurate validation and benchmarking of computational tools.

From the structural point of view, each sequence can have one or more assigned biological units that denote biologically relevant quaternary structures of asymmetric units stored in the PDB database (14). For representative biological units, the HotSpot Wizard 3.0 (15) calculation was executed to compute additional sequential and structural annotations. These annotations can help with the analysis of selected mutations and serve as pre-calculated features applicable in machine learning models.

Stability data acquisition and curation

FireProt^{DB} is composed of the data from four sources: the ProTherm database, the ProtaBank database, manual mining of the scientific literature, and data collected in our laboratory (Figure 1). The primary data source was ProTherm. Due to the multiple problems mentioned in the introduction, we followed several filtering steps. In the first step, we retained only those entries that met the following four criteria: (i) they have a single-point mutation; (ii) the mutation is not an insertion or deletion; (iii) the protein has a SwissProt accession code and/or a PDB identifier; (iv) the entry includes a measured $\Delta\Delta G$ and/or ΔT_m. Secondly, we performed a validity check of SwissProt accession codes and updated obsolete entries. ProTherm references mutations by their structure index, i.e., the residue number in the structure, which in many cases does not match their sequence index, i.e. the position in the sequence. To overcome this issue, we used a similar approach as in PDBSWS (16): use the Needleman-Wunsch algorithm (17) to construct the global sequence alignment of sequences extracted from PDB and UniProt entries and map the mutations onto the UniProt sequences. In the next step, we confirmed that the reported wild-type amino acids are in the correct positions in the structures and unified the reported units. Finally, we matched the data with the manually curated entries in the FireProt dataset (18), updated the values, and marked them as ‘curated’.

In addition to ProTherm, we explored the studies reported in the ProtaBank database, extracted the thermostability data, and integrated them into our database. We also performed a manual literature search using stability-based keywords such as ‘protein stability’, ‘thermostability’, ‘free energy upon mutation’, ‘protein stabilization’. We mined the recent scientific articles reporting mutants with measured stability data and contacted the authors of the publications when the relevant data were not available in the article. All such entries were marked as ‘curated’ as we extracted them directly from the original publications. Finally, we reviewed the thermostability data collected in our lab throughout the last few years and added them to the database. We perform experimental protein characterization in our protein engineering projects on a regular basis, and measuring protein stability is an essential part of such characterization. In total, the three sources led to a significant enlargement of the data size by 62% in terms of all the entries. The number of curated entries more than dou-

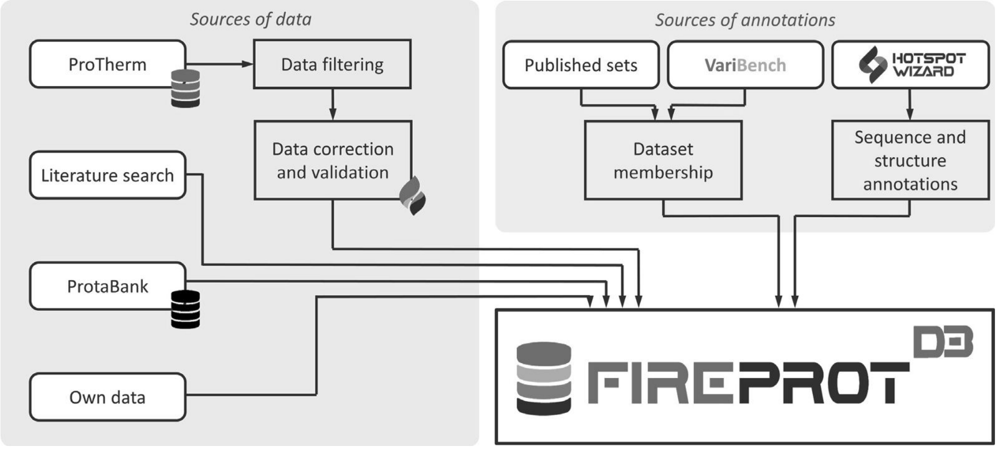


Figure 1. A schematic representation of the data comprising FireProt^{DB}. The primary source of data is filtered ProTherm (7). The FireProt data subset (18) was manually curated, compared to the source publications, and marked with the 'curated' flag. The publications from ProtaBank (12) and manual literature search were also used to deposit the data. Each mutation in the deposited data was annotated according to its membership in the published datasets and those deposited on VariBench (11). The HotSpot Wizard 3.0 (15) annotation tool was applied to each protein entry with a known tertiary structure.

bled compared to the previously collected cleaned FireProt subset of ProTherm.

Dataset assignment

In the second acquisition step, we collected 40 datasets from the VariBench database (11) and literature (18), which were used previously for training or testing of existing predictors. Since all these datasets are at least partially derived from ProTherm, we could label each measurement in FireProt^{DB} by its membership in the datasets. These labels are particularly useful for the comparison of new prediction models to the existing tools. This task is usually done by the performance evaluation of predictors on a dataset that is entirely independent of the training and test sets used for the development of the tools. Since the dataset construction is often laborious and consists of a manual data processing, the possibility to directly exclude the data present in given datasets significantly simplifies and speeds up the construction process.

Calculation of additional annotations

To provide our users with a more advanced description of their proteins of interest, we enriched the database by several important sequence- and structure-related information. These calculations were performed by HotSpot Wizard 3.0 (15), which is currently the only tool capable of deriving all these features in a single calculation (19) and provides machine-readable results. HotSpot Wizard was executed on a representative biological unit of each protein and provided the annotations for a structure, such as the residues located in protein pockets and tunnels, and a sequence, such as catalytic residues, evolutionary conservation scores, back-to-consensus mutations, and correlated pairs. These annotations can be helpful for a better understanding of structure-function relationships as well as for generating features for machine learning.

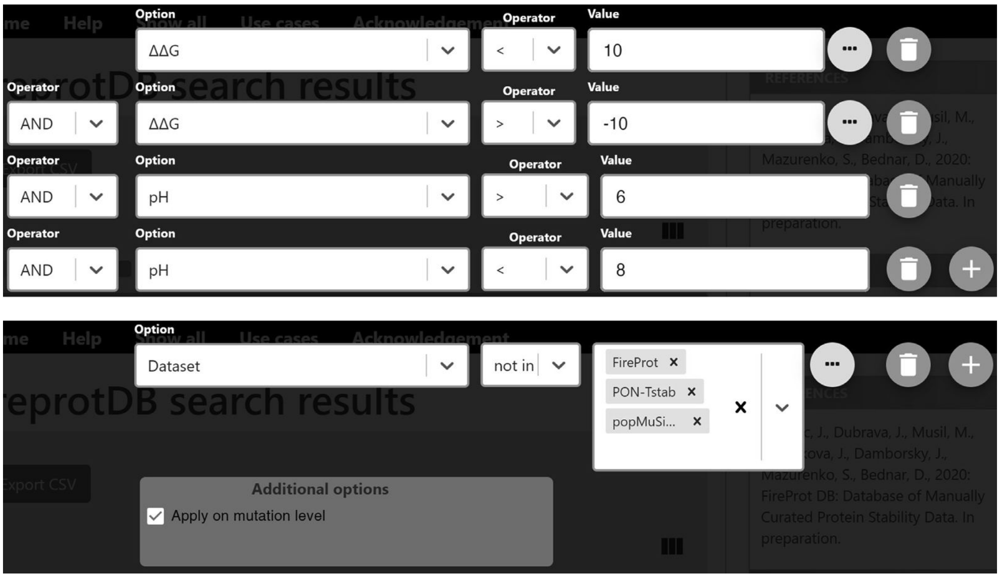
RESULTS

Web interface

The web interface was designed for both types of expected users—protein chemists and software developers. Protein chemists are often looking for the thermostability evidence for their protein of interest, and they will benefit from its interactivity and details pages with additional information. Machine learning experts and bioinformaticians will be more interested in advanced filtering capabilities facilitating the process of construction of highly customized datasets for the training or assessment of various predictors. The entry point to the database is the search form, which allows browsing in two major ways: (i) a simple full-text search for querying the database using protein name, UniProt accession codes, PDB identifiers, protein names, publications, authors or organisms and (ii) an advanced search allowing the users to construct complex rules based on the relational algebra and all available database fields. The latter is one of the key features of FireProt^{DB} as it facilitates the construction of highly customized datasets needed for the development of new predictors.

Once the user clicks on the 'Search' button, they are redirected to the page with the result table. This table contains a list of available experiments, their basic annotations, and measured values. The table is paginated to eliminate possible performance issues and allows further interactive filtering of displayed values. The user can then easily export the search results in the CSV format using the 'Export' button at the top or the bottom of the page.

Clicking on a mutation name leads to a page with a more detailed view, showing all the data entries and datasets that include the selected mutation. Clicking on a protein name leads to a page providing the basic information such as UniProt accession code, organism and Enzyme Commission number, as well as detailed annotation of secondary structure, catalytic sites, natural variants and amino acid charges derived from UniProt database using interactive



Operator	Option	Operator	Value
	ΔΔG	<	10
AND	ΔΔG	>	-10
AND	pH	>	6
AND	pH	<	8

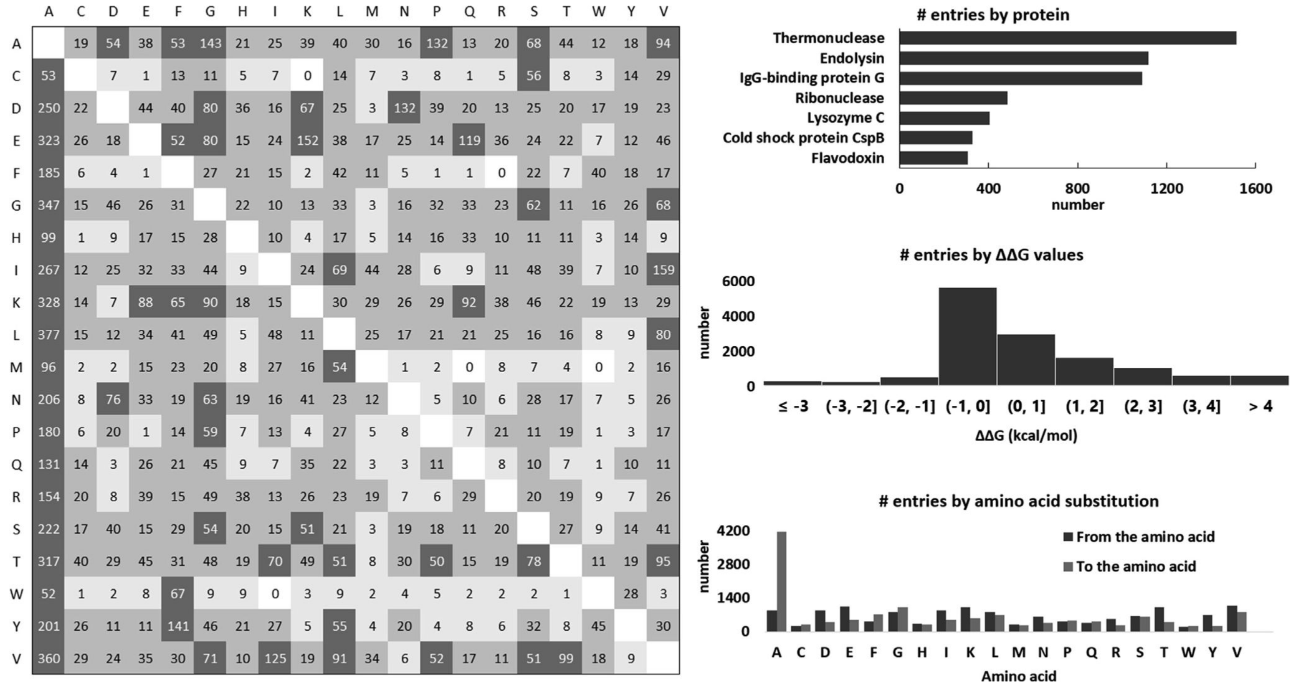
Option		
Dataset	not in	FireProt × PON-Tstab × popMuSi... ×

Additional options

☑ Apply on mutation level

me Help Show all Use cases Acknowledgement
FireprotDB search results
Export CSV
REFERENCES
c, J., Dubrava, J., Musil, M., ova, J., Damborsky, J., Mazurenko, S., Bednar, D., 2020: FireProt DB: Database of Manually Curated Protein Stability Data. In preparation.

Figure 2. Examples of filtering protocols in FireProt^{DB}. **Top**: The request filters out the data collected at extreme pH or with extreme $\Delta\Delta G$ values, resulting in >3500 data points left. **Bottom**: An example of excluding all the mutations that appear in PopMuSiC, FireProt, or PON-Tstab datasets.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	W	Y	V
A		19	54	38	53	143	21	25	39	40	30	16	132	13	20	68	44	12	18	94
C	53		7	1	13	11	5	7	0	14	7	3	8	1	5	56	8	3	14	29
D	250	22		44	40	80	36	16	67	25	3	132	39	20	13	25	20	17	19	23
E	323	26	18		52	80	15	24	152	38	17	25	14	119	36	24	22	7	12	46
F	185	6	4	1		27	21	15	2	42	11	5	1	1	0	22	7	40	18	17
G	347	15	46	26	31		22	10	13	33	3	16	32	33	23	62	11	16	26	68
H	99	1	9	17	15	28		10	4	17	5	14	16	33	10	11	11	3	14	9
I	267	12	25	32	33	44	9		24	69	44	28	6	9	11	48	39	7	10	159
K	328	14	7	88	65	90	18	15		30	29	26	29	92	38	46	22	19	13	29
L	377	15	12	34	41	49	5	48	11		25	17	21	21	25	16	16	8	9	80
M	96	2	2	15	23	20	8	27	16	54		1	2	0	8	7	4	0	2	16
N	206	8	76	33	19	63	19	16	41	23	12		5	10	6	28	17	7	5	26
P	180	6	20	1	14	59	7	13	4	27	5	8		7	21	11	19	1	3	17
Q	131	14	3	26	21	45	9	7	35	22	3	3	11		8	10	7	1	10	11
R	154	20	8	39	15	49	38	13	26	23	19	7	6	29		20	19	9	7	26
S	222	17	40	15	29	54	20	15	51	21	3	19	18	11	20		27	9	14	41
T	317	40	29	45	31	48	19	70	49	51	8	30	50	15	19	78		11	19	95
W	52	1	2	8	67	9	9	0	3	9	2	4	5	2	2	2	1		28	3
Y	201	26	11	11	141	46	21	27	5	55	4	20	4	8	6	32	8	45		30
V	360	29	24	35	30	71	10	125	19	91	34	6	52	17	11	51	99	18	9	

Figure 3. An overview of the data deposited to FireProt^{DB}. **Left**: The table shows the total number of each substitution pair with the wild type amino acids in rows, mutant amino acids in columns, and the coloring according to the thresholds of 1 (light green), 10 (medium green) and 50 (dark green) entries for the corresponding substitution. **Right**: Histograms showing the top seven proteins by their UniProt IDs, the $\Delta\Delta G$ values, and the cumulative number of amino acid substitutions.

ProtVista tracks (20). This page also contains a list of all known biological units and a table with all experimental measurements.

Search queries

Several types of search queries may be of interest to the users. The first one relates to data filtering by values (10). Typically, software developers filter out the data collected at extreme pH (<6 or >8) due to changes in charged states for ionizable residues. The entries with large absolute $\Delta\Delta G$ or ΔT_m are also sometimes excluded due to likely higher measurement errors, and also because dramatic changes to the stability may indicate significant structural alterations to the wild type, which may become a problem for structure-based features. The second type is relevant for benchmarking of a newly designed predictor against the existing tools or creating a meta predictor. In either case, one usually needs to derive a data subset that has not been used by the existing predictors for training. The main reason is the robust performance estimate, which is typically overoptimistic for these sets (6). Two corresponding examples of such filtering protocols are shown in Figure 2.

Database dump

For the users requesting even higher control over the data and filtering capabilities, we offer the possibility to download the complete dump of the database in the SQL format. This data file can be easily imported to any modern MariaDB server, version 10.2, and higher. Since the database structure is complex and any custom query requires joining of multiple tables, the dump also contains a pre-defined view 'mutation_experiments_summary'. The summary combines all the tables and provides the data in a similar structure as the CSV export from the user interface. This view or its definition can serve as a useful starting point for additional filtering or creating custom queries.

Data statistics

Currently, FireProt^{DB} contains 13274 entries for 237 proteins (Figure 3), from which 8189 measurements originated from ProTherm. The remaining 5085 entries were added from our literature search (18%), publications from ProtaBank (28%), VariBench (53%), and our own records (1%). In total, 43% entries are destabilizing mutations ($\Delta T_{\rm m} \leftarrow 1$ or $\Delta\Delta G > 1$ kcal/mol), 14% stabilizing ($\Delta T_{\rm m} > 1$ or $\Delta\Delta G \leftarrow 1$ kcal/mol), and 43% considered neutral ($-1 \leq \Delta T_{\rm m} \leq 1$ or $-1 \leq \Delta\Delta G \leq 1$ kcal/mol). The database also includes annotations for 40 various published datasets derived from ProTherm, deposited to VariBench (11), or available in the corresponding articles and web servers. As far as enzymes are concerned, those collected in the database cover the first six EC classes, three of which by >40% on the second level.

DISCUSSION

The availability of large high-quality datasets is one of the critical requirements for the advancement of machine learning-based *in silico* predictors. While some promising high-throughput experimental methods have been released recently (21,22), their validation is still ongoing, and protein stability experiments are still time-consuming and expensive. Building training and testing datasets is hindered by the data being hidden in the original articles, generating a strong demand for their systematic mining, collection, validation, and homogenization. The existing databases are not fulfilling all the requirements as ProTherm is outdated and contains incorrect data, and ProtaBank does not provide advanced search and export tools and is partly commercial.

FireProt^{DB} is a novel database for experimental thermostability data of protein single-point mutants. It consists of the data manually extracted from ProTherm, articles from ProtaBank, new data obtained by mining the recent literature, and the data collected in our laboratory. The database is accessible via a user-friendly graphical web interface allowing the users to search and browse the data interactively. Moreover, all the entries are annotated to indicate whether they belong to the already published datasets. These annotations, combined with the advanced searching and filtering capabilities, make FireProt^{DB} a valuable data resource for machine learning developers interested in constructing highly customized datasets.

In the future, we will improve our searching queries and employ automatic text-mining machine learning-based approaches (23–25) to accelerate literature mining and data collection, which will be followed by manual curation. We will also prepare an interactive form for data submissions by the users. Finally, we will extend the set of automatically generated features for mutations and add sequence similarity filtering to improve the data usability by the community of engineers applying machine learning to predict changes in protein stability.

FUNDING

Czech Ministry of Education, Youth and Sports [LQ1605, LM2015047, LM2018121, 02.1.01/0.0/0.0/18_046/0015975 to J.D.]; Operational Programme Research, Development and Education project MSCAfellow@MUNI [CZ.02.2.69/0.0/0.0/17_050/0008496 to S.M.]; Brno University of Technology [FIT-S-20-6293 to M.M.]; CETOCOEN EXCELLENCE Teaming 2 project supported by Horizon2020 of the European Union [857560 to J.D.]; Czech Science Foundation [20-15915Y to D.B.]. Funding for open access charge: Czech ministry of Education, Youth and Sports [LM2015047].

Conflict of interest statement. None declared.

REFERENCES

1. Modarres,H.P., Mofrad,M.R. and Sanati-Nezhad,A. (2016) Protein thermostability engineering. *RSC Adv.*, **6**, 115252–115270.
2. Gao,D., Narasimhan,D.L., Macdonald,J., Brim,R., Ko,M.-C., Landry,D.W., Woods,J.H., Sunahara,R.K. and Zhan,C.-G. (2009) Thermostable variants of cocaine esterase for long-time protection against cocaine toxicity. *Mol. Pharmacol.*, **75**, 318–323.
3. Wijma,H.J., Floor,R.J. and Janssen,D.B. (2013) Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr. Opin. Struct. Biol.*, **23**, 588–594.
4. Ferdjani,S., Ionita,M., Roy,B., Dion,M., Djeghaba,Z., Rabiller,C. and Tellier,C. (2011) Correlation between thermostability and stability of glycosidases in ionic liquid. *Biotechnol. Lett.*, **33**, 1215–1219.
5. Polizzi,K.M., Bommarius,A.S., Broering,J.M. and Chaparro-Riggers,J.F. (2007) Stability of biocatalysts. *Curr. Opin. Chem. Biol.*, **11**, 220–225.
6. Musil,M., Konegger,H., Hon,J., Bednar,D. and Damborsky,J. (2019) Computational design of stable and soluble biocatalysts. *ACS Catal.*, **9**, 1033–1054.
7. Kumar,M.D.S., Bava,K.A., Gromiha,M.M., Prabakaran,P., Kitajima,K., Uedaira,H. and Sarai,A. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. *Nucleic Acids Res.*, **34**, D204–D206.
8. Pucci,F., Bernaerts,K.V., Kwasigroch,J.M. and Rooman,M. (2018) Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics*, **34**, 3659–3665.
9. Folkman,L., Stantic,B., Sattar,A. and Zhou,Y. (2016) EASE-MM: sequence-based prediction of mutation-induced stability changes with feature-based multiple models. *J. Mol. Biol.*, **428**, 1394–1405.

10. Mazurenko,S. (2020) Predicting protein stability and solubility changes upon mutations: data perspective. *Chem. Cat. Chem.*, **12**, doi:10.1002/cctc.202000933.
11. Sasidharan Nair,P. and Vihinen,M. (2013) VariBench: a benchmark database for variations. *Hum. Mutat.*, **34**, 42–49.
12. Wang,C.Y., Chang,P.M., Ary,M.L., Allen,B.D., Chica,R.A., Mayo,S.L. and Olafson,B.D. (2018) ProtaBank: a repository for protein design and engineering data. *Protein Sci.*, **27**, 1113–1124.
13. The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
14. Jefferson,E.R., Walsh,T.P. and Barton,G.J. (2006) Biological units and their effect upon the properties and prediction of protein-protein interactions. *J. Mol. Biol.*, **364**, 1118–1129.
15. Sumbalova,L., Stourac,J., Martinek,T., Bednar,D. and Damborsky,J. (2018) HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information. *Nucleic Acids Res.*, **46**, W356–W362.
16. Martin,A.C.R. (2005) Mapping PDB chains to UniProtKB entries. *Bioinformatics*, **21**, 4297–4301.
17. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
18. Musil,M., Stourac,J., Bendl,J., Brezovsky,J., Prokop,Z., Zendulka,J., Martinek,T., Bednar,D. and Damborsky,J. (2017) FireProt: web server for automated design of thermostable proteins. *Nucleic Acids Res.*, **45**, W393–W399.
19. Sequeiros-Borja,C.E., Surpeta,B. and Brezovsky,J. Recent advances in user-friendly computational tools to engineer protein function. *Brief. Bioinform.*, doi:10.1093/bib/bbaa150.
20. Watkins,X., Garcia,L.J., Pundir,S., Martin,M.J. and UniProt Consortium (2017) ProtVista: visualization of protein sequence annotations. *Bioinformatics*, **33**, 2040–2041.
21. Bunzel,H.A., Garrabou,X., Pott,M. and Hilvert,D. (2018) Speeding up enzyme discovery and engineering with ultrahigh-throughput methods. *Curr. Opin. Struct. Biol.*, **48**, 149–156.
22. Matreyek,K.A., Starita,L.M., Stephany,J.J., Martin,B., Chiasson,M.A., Gray,V.E., Kircher,M., Khechaduri,A., Dines,J.N., Hause,R.J. *et al.* (2018) Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.*, **50**, 874–882.
23. Naderi,N. and Witte,R. (2012) Automated extraction and semantic analysis of mutation impacts from the biomedical literature. *BMC Genomics*, **13**, S10.
24. Witte,R. and Baker,C.J.O. (2007) Towards a systematic evaluation of protein mutation extraction systems. *J. Bioinform. Comput. Biol.*, **5**, 1339–1359.
25. Wei,C.-H., Harris,B.R., Kao,H.-Y. and Lu,Z. (2013) tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, **29**, 1433–1439.

D Original publication IV: HotSpotWizard 2.0

BENDL J, STOURAC J, SEBESTOVA E, VAVRA O, MUSIL M, BREZOVSKY J, DAMBORSKY J. HotSpotWizard 2.0: Automated Design of Site-specific Mutations and Smart Libraries in Protein Engineering. *Nucleic Acids Research*. 2016, 44(W1), W479-W487.

Author participation: 15%
Journal impact factor: 11.501 (Q1)

Published online 12 May 2016 *Nucleic Acids Research, 2016, Vol. 44, Web Server issue* ***W479–W487***
doi: 10.1093/nar/gkw416

HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering

Jaroslav Bendl[1,2,3,†], Jan Stourac[1,†], Eva Sebestova[1], Ondrej Vavra[1], Milos Musil[1,2], Jan Brezovsky[1,3,*] and Jiri Damborsky[1,3,*]

[1]Loschmidt Laboratories, Department of Experimental Biology and Research Centre for Toxic Compounds in the Environment RECETOX, Masaryk University, 625 00 Brno, Czech Republic, [2]Department of Information Systems, Faculty of Information Technology, Brno University of Technology, 612 66 Brno, Czech Republic and [3]International Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic

Received February 21, 2016; Revised April 22, 2016; Accepted May 03, 2016

ABSTRACT

HotSpot Wizard 2.0 is a web server for automated identification of hot spots and design of smart libraries for engineering proteins' stability, catalytic activity, substrate specificity and enantioselectivity. The server integrates sequence, structural and evolutionary information obtained from 3 databases and 20 computational tools. Users are guided through the processes of selecting hot spots using four different protein engineering strategies and optimizing the resulting library's size by narrowing down a set of substitutions at individual randomized positions. The only required input is a query protein structure. The results of the calculations are mapped onto the protein's structure and visualized with a JSmol applet. HotSpot Wizard lists annotated residues suitable for mutagenesis and can automatically design appropriate codons for each implemented strategy. Overall, HotSpot Wizard provides comprehensive annotations of protein structures and assists protein engineers with the rational design of site-specific mutations and focused libraries. It is freely available at http://loschmidt.chemi.muni.cz/hotspotwizard.

INTRODUCTION

The development of tailor-made enzymes for industrial applications is facilitated by understanding the molecular mechanisms of protein function. However, despite significant advances in recent decades, it is not yet clear how a protein's sequence encodes its function (1,2). Traditional directed evolution circumvents this problem by using repeated rounds of random mutagenesis and screening of large sequence libraries to explore the mutational landscape and find proteins with desired properties (2–5). This approach has the advantage of requiring no prior knowledge of the protein's structure or understanding of its structure–function relationships (6), but necessitates the laborious and costly screening of very large libraries (4). The efficiency of directed evolution experiments can be significantly improved by creating smaller, higher quality libraries that are more likely to yield positive results. Such 'smart' libraries can be generated by focusing mutagenesis on a limited number of 'hot spot' positions that are likely to affect the property of interest, or by selecting a limited set of substitutions (1–5).

The optimal strategy for identifying hot spots depends on the property being targeted. Catalytic properties such as activity, specificity and stereoselectivity are often related to amino acid residues that mediate substrate binding, transition-state stabilization or product release (7,8). Such residues can be identified using tools for predicting and analyzing enzyme-ligand interactions (9–11) or detecting binding pockets or access tunnels (12–14). Strategies for improving protein stability include rigidification of flexible sites, cavity-filling, tunnel engineering, consensus and ancestral mutation methods, or redesigning of surface charges (15–17). While hot spots for some of these strategies can be identified straightforwardly using a single computational tool (18), others require multi-step analyses or the use of molecular modelling methods (19). Having obtained a set of promising sites for manipulating the desired property, the next challenge is to draw up a list of allowed substitutions at individual positions. This can be done by considering the amino acid distribution at the corresponding positions in sequence homologs (20,21), by using reduced sets of amino acids with either specific desired physicochemical properties or a balanced set of these properties (22,23), or on the basis of the predicted effects of specific substitutions on the protein's properties (24,25). Finally, an appropriate degen-

[*]To whom correspondence should be addressed. Tel: +420 5 4949 3467; Fax: +420 5 4949 6302; Email: jiri@chemi.muni.cz
Correspondence may also be addressed to Jan Brezovsky. Email: brezovsky@mail.muni.cz
[†] These authors contributed equally to the work as first authors.

erate codon covering the specified set of amino acids must be selected for each targeted position. Ideally, these codons should exhibit minimal amino acid bias and minimize the frequency of premature stop codons (26). Several tools are available to facilitate this task and to calculate the size of the designed library (27).

Here, we present HotSpot Wizard 2.0, a web server for the automated identification of hot spots and design of smart libraries for engineering protein stability, enzymatic activity, substrate specificity and enantioselectivity. Compared to its predecessor (28), HotSpot Wizard 2.0 introduces several major improvements, extending the scope and quality of its analyses. It implements four different established protein engineering strategies, enabling the user to selectively target sites affecting the protein's stability and catalytic properties. Users can easily select suitable substitutions for individual hot spots based on predictions of tolerated amino acids or amino acid distributions in sequence homologs, and suitable degenerate codons for these substitutions can be designed automatically via the HotSpot Wizard interface. A new graphical user interface provides an intuitive and comprehensive overview of the results of the analysis, allowing users to think directly about the obtained designs. The resulting pipeline of twenty integrated tools and three databases represents a unique one-stop solution that makes library design accessible even to users with no prior knowledge of bioinformatics.

MATERIALS AND METHODS

The workflow of HotSpot Wizard is outlined in Figure 1. In order to explore the mutational landscape and find the most promising mutagenesis targets, a protein selected by the user is annotated using several prediction tools and databases (Phase 1). With this knowledge in hand, four protein engineering strategies are used to identify suitable hot spots for improving desired protein properties (Phase 2). Finally, suitable substitutions and appropriate degenerate codons are proposed for each selected hot spot, enabling the design of a smart library (Phase 3).

Phase 1: annotation of the protein

The first step in the workflow requires the user to specify the protein structure of interest, either by providing its PDB ID or by uploading a suitable PDB file. If possible, the biological assembly of the target protein is automatically generated by the MakeMultimer tool (http://watcut.uwaterloo.ca/tools/makemultimer), and information about 'essential residues' directly involved in catalysis or binding is obtained from the Catalytic Site Atlas (29) and UniProtKB/Swiss-Prot (30) databases. The DSSP algorithm (31) is then used to assign the protein's secondary structure, and its accessible surface area is computed using the Shrake and Rupley algorithm (32) with BioJava (33). The average B-factors are computed for the protein's amino acid residues (34). The raw B-factor values are accompanied by residue rankings ranging from 1–100%; rankings of 1–25%, 26–75% and 76–100% indicate high, moderate and low levels of relative structural flexibility, respectively. Protein pockets are then identified with Fpocket (35). For each chain, the pocket containing the greatest number of essential residues is identified as the catalytic pocket. If there are two or more pockets that satisfy this criterion, a decision is made according to the Fpocket score. Having identified the putative catalytic pockets, their centers of mass are determined and used as starting points to identify access tunnels with CAVER (36). Sequence homologs of the target protein are then obtained by performing a BLAST (37) search against the UniRef90 (38) database, using the target protein sequence as a query. All identified homologs are aligned with the query protein using USEARCH (39). By default, sequences whose identity with the query is below 30% or above 90% are excluded from the list of homologs. The remaining sequences are then clustered using UCLUST (39), with a 90% identity threshold to remove close homologs. The cluster representatives are sorted based on the BLAST query coverage and by default, the first 200 of them are used to create a sequence data set. A multiple sequence alignment of the resulting sequence data set is created with Clustal Omega (40) and used to (i) estimate the conservation of each position in the protein based on the Jensen–Shannon entropy (41); (ii) identify correlated positions using an ensemble of the MI (42), aMIc (43), OMES (44), SCA (45), DCA (46), McBASC (47) and ELSC (48) methods; (iii) predict the tolerated amino acids at each position in the protein sequence using RAPHYD (see Supplementary Data 1); and (iv) analyze amino acid frequencies at individual positions within the protein. The conservation scores are used to assign mutability values to individual residues. To facilitate interpretation, these values are divided into three groups: values of 1–3, 4–5 and 6–9 indicate low, moderate and high mutability, respectively.

Phase 2: identification of mutagenesis hot spots

Based on the comprehensive annotation of the target protein, four protein engineering strategies are used to identify different types of hot spots: (i) functional hot spots, (ii) stability hot spots based on structural flexibility, (iii) stability hot spots based on sequence consensus and (iv) correlated hot spots. Some examples illustrating the use of these strategies to engineer selected properties in 12 different proteins (34,49–62) are shown in Figure 2. Functional hot spots correspond to highly mutable residues located in the catalytic pockets or tunnels connecting these pockets with the bulk solvent. Residues located in close proximity to the active site have been identified as good mutagenesis targets for engineering activity, enantioselectivity and substrate specificity (52,63,64). To prevent mutagenesis at positions that are indispensable for protein function, all essential residues are designed immutable and thus excluded from the list of potential hot spots. Supplementary Data 2 shows that HotSpot Wizard provides a significantly greater proportion of viable mutants than random mutagenesis. Stability hot spots are identified by analyzing structural flexibility and sequence consensus. The former approach aims to rigidify flexible protein regions by mutating residues with high average B-factors (34). B-factor provides a metric for flexibility which is due in part to inherent flexibility of the macromolecule, but also includes stabilizing/destabilizing energy from packing in the crystal lattice. The rationale for targeting these flexible residues is that they have relatively

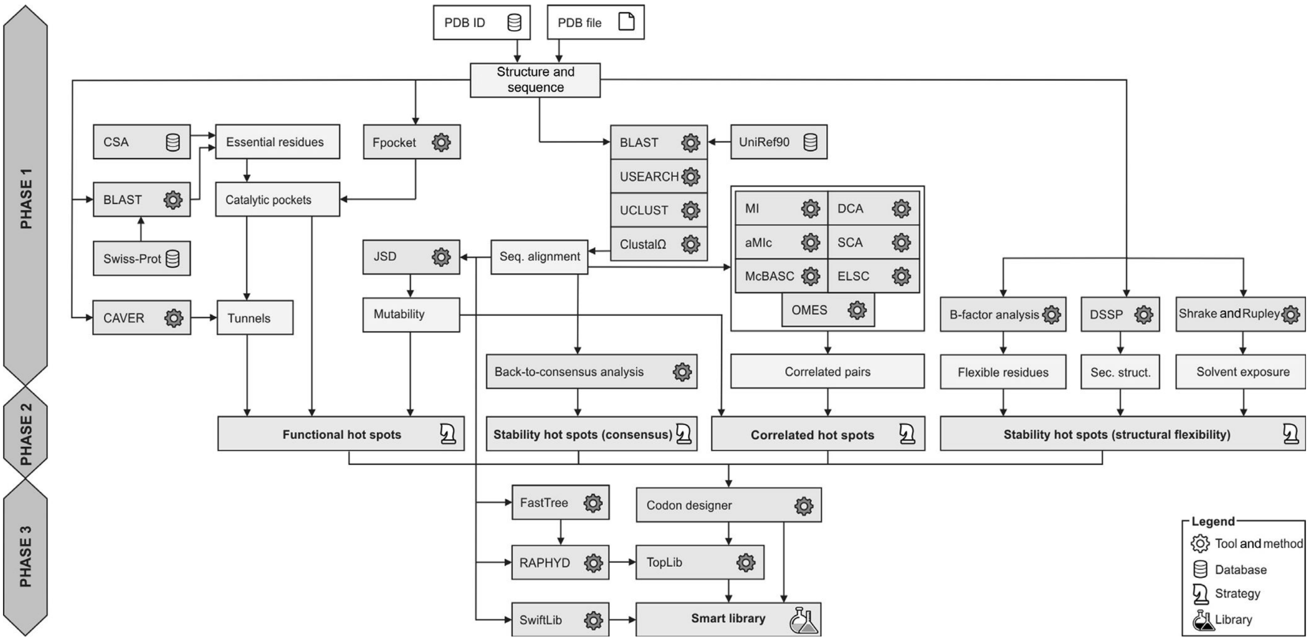


Figure 1. Workflow of HotSpot Wizard.

few contacts with neighbors, so their substitution can produce more interactions (34,54,55). In contrast, the sequence consensus protocol implements majority and frequency ratio approaches, both of which suggest mutations at positions where the wild-type amino acid differs from the most prevalent amino acid (i.e. the consensus residue) at a given position in the multiple sequence alignment. The assumption that the most common amino acid is likely to be stabilizing has proven to be very successful at creating more stable proteins (56–58,65). By default, if the consensus residue is present in at least 50% of all analyzed sequences, the corresponding position is identified as a hot spot in the majority approach. The frequency ratio approach has a less strict criterion for the consensus residue's frequency – the default value is 40%, but it must also be at least five times more frequent than the wild-type residue as a hot spot. The final strategy involves searching for coordinated changes of the amino acids at two separate positions within the protein. Such pairs of positions are referred to as correlated hot spots, and arise when one amino acid substitution has an unfavorable effect that is compensated for by a second mutation of a residue that is located in close structural proximity to the first. This second, correlated mutation typically helps to maintain protein function, stability or folding (66). Methods developed for identifying correlated pairs have revealed mutations responsible for modulating substrate specificity (67), enantioselectivity (68) and mutagenesis targets for stability engineering (69). The identification of correlated positions in HotSpot Wizard is based on an ensemble of seven prediction tools. Each tool generates a raw score for each pair of residues in the protein that measures the pair's degree of correlation. The mean and standard deviation of the degrees of correlation for all pairs of residues in the protein are then calculated and the raw scores are converted into Z-scores, which measure the number of standard deviations by which each pair's raw score deviates from the mean. Based on the work of Martin *et al.* (70), a pair is considered to be correlated if its average Z-score ≥ 3.5 and both of its positions have at least a moderate degree of mutability – by definition, highly conserved positions cannot co-evolve (71).

Phase 3: design of the smart library

The efficiency of directed evolution experiments can be improved by focusing mutagenesis on a limited number of hot spots, but also by restricting the number of allowed substitutions at individual positions using appropriate codons (20–25). For each protein engineering strategy, HotSpot Wizard provides a way to prioritize amino acids at the randomized positions (Table 1) and identifies degenerate codons encoding all desired amino acids with the minimum redundancy and the smallest possible ratio of stop codons. Alternatively, the SwiftLib tool (73) can be used to calculate optimal degenerate codons while keeping the library diversity within the specified limits (the default 10 000). Although the resulting library may not necessarily fully cover the desired set of amino acids, the probability of omitting the important amino acids is relatively low as their weights are set according to selected prioritization method (e.g. based on amino acid distributions in sequence homologs). For both approaches, the most common metrics, such as expected coverage or library size, are computed with TopLib (72).

DESCRIPTION OF THE WEB SERVER

Input

The only required input to the web server is a tertiary structure of the query protein, provided either as a PDB ID or a PDB file. The user can then choose a predefined biological

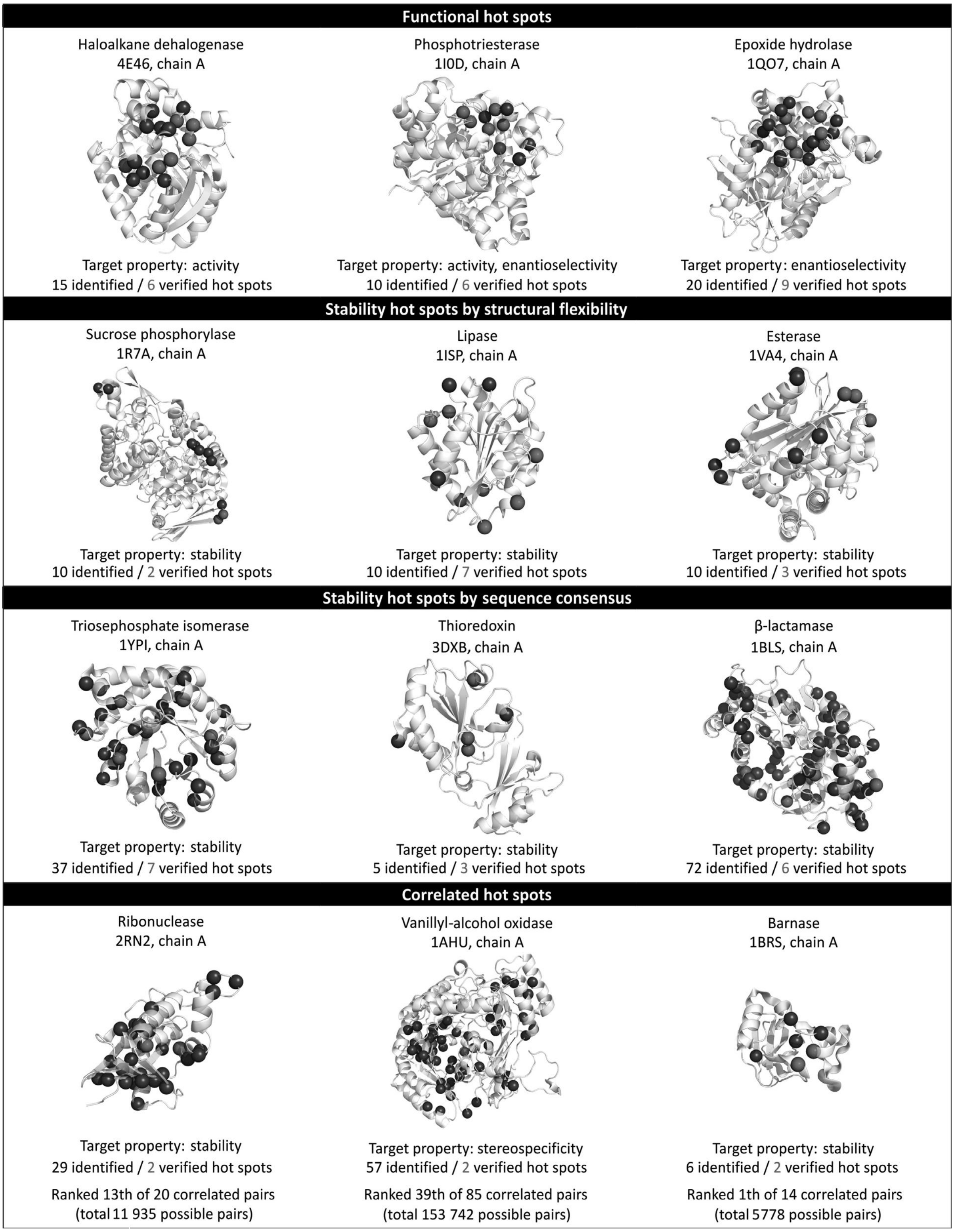

Functional hot spots
Haloalkane dehalogenase
4E46, chain A
Target property: activity
15 identified / 6 verified hot spots
Phosphotriesterase
1I0D, chain A
Target property: activity, enantioselectivity
10 identified / 6 verified hot spots
Epoxide hydrolase
1QO7, chain A
Target property: enantioselectivity
20 identified / 9 verified hot spots
Stability hot spots by structural flexibility
Sucrose phosphorylase
1R7A, chain A
Target property: stability
10 identified / 2 verified hot spots
Lipase
1ISP, chain A
Target property: stability
10 identified / 7 verified hot spots
Esterase
1VA4, chain A
Target property: stability
10 identified / 3 verified hot spots
Stability hot spots by sequence consensus
Triosephosphate isomerase
1YPI, chain A
Target property: stability
37 identified / 7 verified hot spots
Thioredoxin
3DXB, chain A
Target property: stability
5 identified / 3 verified hot spots
β-lactamase
1BLS, chain A
Target property: stability
72 identified / 6 verified hot spots
Correlated hot spots
Ribonuclease
2RN2, chain A
Target property: stability
29 identified / 2 verified hot spots
Ranked 13th of 20 correlated pairs
(total 11 935 possible pairs)
Vanillyl-alcohol oxidase
1AHU, chain A
Target property: stereospecificity
57 identified / 2 verified hot spots
Ranked 39th of 85 correlated pairs
(total 153 742 possible pairs)
Barnase
1BRS, chain A
Target property: stability
6 identified / 2 verified hot spots
Ranked 1th of 14 correlated pairs
(total 5778 possible pairs)

Figure 2. Some notable applications of the four protein engineering strategies implemented in the HotSpot Wizard web server.

Table 1. Methods for selecting substitutions at hot spot positions identified using the four different protein engineering strategies

Selection mode	Availability in strategies	Description
Amino acid frequency	FUNC, FLEX	suggests amino acid residues fulfilling the criterion of minimal frequency in the multiple sequence alignment
Mutational landscape	FUNC, FLEX	suggests amino acid residues fulfilling the criterion of minimal probability of preservation of protein function
Sequence consensus	CONS	suggests amino acid residues fulfilling the criteria of at least one of approaches implemented in sequence consensus strategy: (i) majority approach or (ii) frequency ratio approach
Correlated positions	CORREL	suggests amino acid residues fulfilling the criterion of minimal frequency of co-occurrence with some other specific residue from coupled position
Manual	ALL	manual selection of amino acid residues

FUNC – Analysis of functional hot spots; FLEX – Analysis of stability hot spots/structural flexibility approach; CONS – Analysis of stability hot spots / sequence consensus approach; CORREL – Analysis of correlated hot spots

unit generated by the MakeMultimer tool or manually select chains for which the calculation should be performed. The calculations can be configured in either basic or advanced mode. Basic mode directs the user's attention to the most important parameters, providing an overview of the identified essential residues and highlighting the main parameters involved in the identification of pockets and tunnels. The designation of essential residues is a key step in the functional strategy because these residues are excluded from the list of potential hot spots and are also used to detect catalytic pockets and access tunnels. The user should therefore inspect the automatically generated list of essential residues and correct it if necessary. If no essential residues are detected, the user should specify them manually. In basic mode, the user can specify three parameters: (i) the probe radius, which is used in pocket identification and defines the minimum radius of an alpha sphere in a pocket (default 2.8 Å); (ii) the minimum probe radius, which defines the minimum radius of a putative tunnel (default 1.4 Å); and (iii) the clustering threshold, which determines how the hierarchically clustered tunnels are cut and thus affects the number of tunnels that can be identified (default 3.5 Å). Advanced mode allows expert users to fine-tune parameters of individual calculations in the pipeline to achieve more specialized objectives.

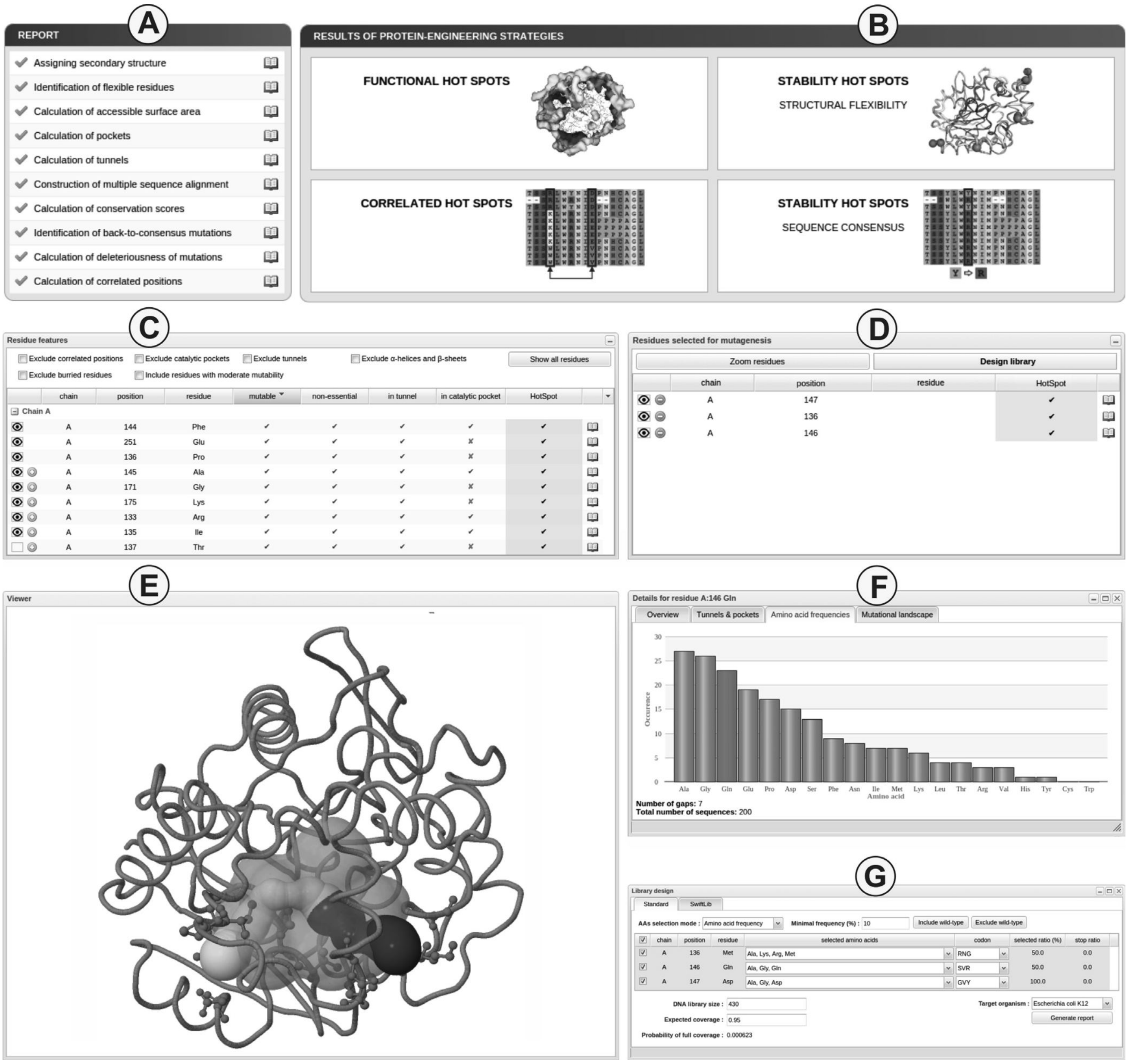
Output

Upon submission, a unique identifier is assigned to each job to track the calculation. The 'Results browser' panel provides information on the status of individual steps in the computational pipeline (Figure 3A). Once the job is finished, the navigation panel provides links to the results obtained using each of the four different protein engineering strategies (Figure 3B). The result pages for each strategy are all organized in the same way, which is described below.

Residue features. The 'Residue features' panel lists all of the identified hot spots together with information relevant to the selected protein engineering strategy (Figure 3C). Several checkboxes can be found at the top of this panel, allowing users to reduce the list of hot spots by applying additional criteria such as excluding buried residues, correlated positions or residues forming a catalytic pocket. The 'Show all residues' button enables users to inspect any residue of the target protein and possibly select hot spots based on their own criteria. Importantly, a pop-up window containing detailed information about a given residue is displayed after clicking the 'book' icon in the last column of the table. Users can visualize individual residues within the protein structure by selecting the 'eye' icon in the first column, and can add residues to the list of mutagenesis hot spots by clicking the 'plus' icon in the second column. All selected mutagenesis hot spots listed in the 'Residues selected for mutagenesis' panel (Figure 3D) can be used for designing a smart library by clicking the 'Design library' button.

Residue details. The information in the 'Residue details' panel is organized into several tabs (Figure 3F): (i) 'Overview', which provides basic information on the residue's characteristics such as its mutability, average B-factor and secondary structure; (ii) 'Annotations', describing the residue's function (only available for essential residues); (iii) 'Tunnels and Pockets', which lists the pockets and/or tunnels of which the residue is a part; (iv) 'Sequence consensus', listing potential consensus mutations for a given position; (v) 'Amino acid frequencies', providing the distribution of amino acids in the corresponding column of the multiple sequence alignment; (vi) 'Mutational landscape', quantifying the probability of preservation of protein function for individual substitutions at a given site; and (vii) 'Correlated positions', listing all positions correlated with the site in question.

Design of smart library. The 'Library design' panel allows the user to select a set of substitutions and design degenerate codons for systematic mutagenesis of the selected positions (Figure 3G). An automatic method for prioritizing amino acids suitable for the chosen protein engineering strategy will be pre-selected. The panel contains two tabs, each corresponding to one library optimization mode. In the 'Standard mode', users can manually define their own set of required substitutions for individual positions if they so desire. After any change in the list of amino acids, HotSpot Wizard automatically identifies the most suitable codons covering all desired amino acids with the lowest possible redundancy, and the library size corresponding to the specified expected coverage. The parameters of the library can be modified interchangeably, allowing the user to adjust the final library based on its size or preferred degree of its coverage. In the 'SwiftLib mode', users specify the maximum acceptable library diversity and the method reports the op-



REPORT

- ✓ Assigning secondary structure
- ✓ Identification of flexible residues
- ✓ Calculation of accessible surface area
- ✓ Calculation of pockets
- ✓ Calculation of tunnels
- ✓ Construction of multiple sequence alignment
- ✓ Calculation of conservation scores
- ✓ Identification of back-to-consensus mutations
- ✓ Calculation of deleteriousness of mutations
- ✓ Calculation of correlated positions

Residue features

□ Exclude correlated positions □ Exclude catalytic pockets □ Exclude tunnels □ Exclude α-helices and β-sheets [Show all residues]

□ Exclude burried residues □ Include residues with moderate mutability

	chain	position	residue	mutable	non-essential	in tunnel	in catalytic pocket	HotSpot
Chain A								
	A	144	Phe	✓	✓	✓	✓	✓
	A	251	Glu	✓	✓	✓	✘	✓
	A	136	Pro	✓	✓	✓	✘	✓
	A	145	Ala	✓	✓	✓	✓	✓
	A	171	Gly	✓	✓	✓	✘	✓
	A	175	Lys	✓	✓	✓	✘	✓
	A	133	Arg	✓	✓	✓	✓	✓
	A	135	Ile	✓	✓	✓	✓	✓
	A	137	Thr	✓	✓	✓	✘	✓

Residues selected for mutagenesis

[Zoom residues] [Design library]

chain	position	residue	HotSpot
A	147		✓
A	136		✓
A	146		✓

Number of gaps: 7
Total number of sequences: 200

Library design

Standard | SwiftLib

AAs selection mode : Amino acid frequency **Minimal frequency (%) :** 10 [Include wild-type] [Exclude wild-type]

☑	chain	position	residue	selected amino acids	codon	selected ratio (%)	stop ratio
☑	A	136	Met	Ala, Lys, Arg, Met	RNG	50.0	0.0
☑	A	146	Gln	Ala, Gly, Gln	SVR	50.0	0.0
☑	A	147	Asp	Ala, Gly, Asp	GVY	100.0	0.0

DNA library size : 430
Expected coverage : 0.95
Probability of full coverage : 0.000623
Target organism : Escherichia coli K12
[Generate report]

Figure 3. HotSpot Wizard's graphical user interface, showing results obtained for the haloalkane dehalogenase LinB (PDB ID: 1CV2). (**A**) The 'Report' panel shows the status of the calculations in the individual steps of the computational pipeline. (**B**) Results obtained using the four protein engineering strategies. (**C**) The 'Residue features' panel, which provides an overview of the identified hot spots. (**D**) The 'Residues selected for mutagenesis' panel, which presents a user-adjustable list of residues representing targets for mutagenesis. (**E**) The JSmol viewer allows interactive visualization of the protein and the identified tunnels and pockets. (**F**) The 'Residue details' pop-up window, which provides comprehensive information on the residue's annotations, organized under several tabs. (**G**) The 'Library design' panel, which shows the list of substitutions and appropriate codons for randomization of selected positions.

timal combination of codons with the minimal redundancy of amino acids. However, this efficiency is often achieved at the price of omitting some of desired amino acids with lower weights. The initial amino acid weights derived from the selected prioritization scheme can be changed by selecting the 'Edit amino acid weights'. Additionally, users can request multiple solutions and thus inspect also the solutions which are considered as less optimal by the method, but may better meet the users' needs. Finally, users can generate a nucleotide sequence from the designed amino acid sequence based on the codon usage of selected organism (default is *Escherichia coli*) with the European Molecular Biology Open Software Suite (EMBOSS) Backtranseq tool (74).

Protein visualization. The protein of interest is interactively visualized in the web browser using the JSmol applet (http://wiki.jmol.org/index.php/JSmol). Users can display individual amino acid residues as well as identified tunnels and pockets (Figure 3E). The hot spot residues are colored in red, residues in tunnels and pockets in yellow and all other residues in grey.

Structural features. The main characteristics of all pockets and access tunnels are presented in the 'Pockets' and 'Tunnels' panels, respectively. These panels allow users to visualize individual pockets and tunnels in the structure and to open a pop-up window showing a list of all the residues comprising the chosen structural feature.

CONCLUSIONS AND OUTLOOK

HotSpot Wizard 2.0 is a web server for the automatic identification of hot spots and the design of site-specific mutations and mutant libraries for engineering protein stability, catalytic activity, substrate specificity and enantioselectivity. The server provides a unified interface allowing users to apply four well-established protein engineering strategies that combine structural, functional and evolutionary information to identify suitable positions for mutagenesis. Moreover, HotSpot Wizard integrates several schemes for automatic prioritization of mutations and codon optimization for selected hot spot positions to facilitate the design of smart libraries. The automation of the multi-step procedure makes the process of library design accessible to users without expertise in bioinformatics because it eliminates the need to select, install and evaluate tools, optimize their parameters, perform conversions between different data formats, and interpret intermediate results.

In the future, we plan to implement a protocol for structure prediction based on homology modeling, extending the applicability of HotSpot Wizard to proteins for which no experimental structure is yet available. Additionally, we aim to assess other established protein engineering strategies and, if they prove suitable, to develop new modules so they can be added to the server's portfolio of methods.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to express many thanks to Dr Antonin Pavelka (Masaryk University, Brno, Czech Republic) for valuable discussions and Dr Yuval Nov (University of Haifa, Haifa, Israel) for kindly providing the source code of TopLib. Uwe Bornscheuer (University Greifswald), Marco Fraaije (Groningen University) and Moshe Goldsmith (Weizmann Institute of Science) are sincerely acknowledged for constructive comments on the tool. MetaCentrum and CERIT-SC are acknowledged for providing access to computing and storage facilities [LM2015085 and LM2015042].

FUNDING

Ministry of Education of the Czech Republic [LO1214, LQ1605, LM2015055 and LM2015047]; Grant Agency of the Czech Republic [GA16-06096S]; European Commission REGPOT [316345]; Horizon 2020 Research Infrastructure ELIXIR-EXCELERATE [676559]; Brno University of Technology [FIT-S-14-2299 to M.M.]. Funding for open access charge: Grants from Czech Ministry of Education [LO1214, LQ1605, LM2015055 and LM2015047].
Conflict of interest statement. None declared.

REFERENCES

1. Romero,P.A. and Arnold,F.H. (2009) Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.*, **10**, 866–876.
2. Currin,A., Swainston,N., Day,P.J. and Kell,D.B. (2015) Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem. Soc. Rev.*, **44**, 1172–1239.
3. Cheng,F., Zhu,L. and Schwaneberg,U. (2015) Directed evolution 2.0: improving and deciphering enzyme properties. *Chem. Commun. (Camb.)*, **51**, 9760–9772.
4. Lutz,S. (2010) Beyond directed evolution–semi-rational protein engineering and design. *Curr. Opin. Biotechnol.*, **21**, 734–743.
5. Acevedo-Rocha,C.G., Reetz,M.T. and Nov,Y. (2015) Economical analysis of saturation mutagenesis experiments. *Sci. Rep.*, **5**, 10654.
6. Lo Surdo,P., Walsh,M.A. and Sollazzo,M. (2004) A novel ADP- and zinc-binding fold from function-directed in vitro evolution. *Nat. Struct. Mol. Biol.*, **11**, 382–383.
7. Denard,C.A., Ren,H. and Zhao,H. (2015) Improving and repurposing biocatalysts via directed evolution. *Curr. Opin. Chem. Biol.*, **25**, 55–64.
8. Bornscheuer,U.T., Huisman,G.W., Kazlauskas,R.J., Lutz,S., Moore,J.C. and Robins,K. (2012) Engineering the third wave of biocatalysis. *Nature*, **485**, 185–194.
9. Xie,Z.-R. and Hwang,M.-J. (2015) Methods for predicting protein-ligand binding sites. *Methods Mol. Biol.*, **1215**, 383–398.
10. Yuan,Y., Pei,J. and Lai,L. (2013) Binding site detection and druggability prediction of protein targets for structure-based drug design. *Curr. Pharm. Des.*, **19**, 2326–2333.
11. Lavecchia,A. and Di Giovanni,C. (2013) Virtual screening strategies in drug discovery: a critical review. *Curr. Med. Chem.*, **20**, 2839–2860.
12. Sebestova,E., Bendl,J., Brezovsky,J. and Damborsky,J. (2014) Computational tools for designing smart libraries. *Methods Mol. Biol.*, **1179**, 291–314.
13. Brezovsky,J., Chovancova,E., Gora,A., Pavelka,A., Biedermannova,L. and Damborsky,J. (2013) Software tools for identification, visualization and analysis of protein tunnels and channels. *Biotechnol. Adv.*, **31**, 38–49.
14. Zhang,Z., Li,Y., Lin,B., Schroeder,M. and Huang,B. (2011) Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics*, **27**, 2083–2088.
15. Bommarius,A.S. and Paye,M.F. (2013) Stabilizing biocatalysts. *Chem. Soc. Rev.*, **42**, 6534–6565.
16. Wijma,H.J., Floor,R.J. and Janssen,D.B. (2013) Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr. Opin. Struct. Biol.*, **23**, 588–594.
17. Yu,H. and Huang,H. (2014) Engineering proteins for thermostability through rigidifying flexible sites. *Biotechnol. Adv.*, **32**, 308–315.
18. Folkman,L., Stantic,B., Sattar,A. and Zhou,Y. (2016) EASE-MM: Sequence-based prediction of mutation-induced stability changes with feature-based multiple models. *J. Mol. Biol.*, **428**, 1394–1405.
19. Bednar,D., Beerens,K., Sebestova,E., Bendl,J., Khare,S., Chaloupkova,R., Prokop,Z., Brezovsky,J., Baker,D. and Damborsky,J. (2015) FireProt: Energy- and evolution-based computational design of thermostable multiple-point mutants. *PLoS Comput. Biol.*, **11**, e1004556.
20. Reetz,M.T. and Wu,S. (2008) Greatly reduced amino acid alphabets in directed evolution: making the right choice for saturation mutagenesis at homologous enzyme positions. *Chem. Commun. (Camb)*, **43**, 5499–5501.
21. Jochens,H. and Bornscheuer,U.T. (2010) Natural diversity to guide focused directed evolution. *Chembiochem*, **11**, 1861–1866.

22. Pines,G., Pines,A., Garst,A.D., Zeitoun,R.I., Lynch,S.A. and Gill,R.T. (2015) Codon compression algorithms for saturation mutagenesis. *ACS Synth. Biol.*, **4**, 604–614.
23. Reetz,M.T., Kahakeaw,D. and Lohmer,R. (2008) Addressing the numbers problem in directed evolution. *Chembiochem*, **9**, 1797–1804.
24. Goldsmith,M. and Tawfik,D.S. (2013) Enzyme engineering by targeted libraries. *Methods Enzymol.*, **523**, 257–283.
25. Chaparro-Riggers,J.F., Polizzi,K.M. and Bommarius,A.S. (2007) Better library design: data-driven protein engineering. *Biotechnol. J.*, **2**, 180–191.
26. Gaytán,P., Contreras-Zambrano,C., Ortiz-Alvarado,M., Morales-Pablos,A. and Yáñez,J. (2009) TrimerDimer: an oligonucleotide-based saturation mutagenesis approach that removes redundant and stop codons. *Nucleic Acids Res.*, **37**, e125.
27. Nov,Y. (2014) Probabilistic methods in directed evolution: library size, mutation rate, and diversity. *Methods Mol. Biol.*, **1179**, 261–278.
28. Pavelka,A., Chovancova,E. and Damborsky,J. (2009) HotSpot Wizard: a web server for identification of hot spots in protein engineering. *Nucleic Acids Res.*, **37**, W376–W383.
29. Furnham,N., Holliday,G.L., de Beer,T.A.P., Jacobsen,J.O.B., Pearson,W.R. and Thornton,J.M. (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.*, **42**, D485–D489.
30. UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
31. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
32. Shrake,A. and Rupley,J.A. (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.*, **79**, 351–371.
33. Prlić,A., Yates,A., Bliven,S.E., Rose,P.W., Jacobsen,J., Troshin,P.V., Chapman,M., Gao,J., Koh,C.H., Foisy,S. *et al.* (2012) BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics*, **28**, 2693–2695.
34. Reetz,M.T., Carballeira,J.D. and Vogel,A. (2006) Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angew. Chem. Int. Ed Engl.*, **45**, 7745–7751.
35. Le Guilloux,V., Schmidtke,P. and Tuffery,P. (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, **10**, 168.
36. Chovancova,E., Pavelka,A., Benes,P., Strnad,O., Brezovsky,J., Kozlikova,B., Gora,A., Sustr,V., Klvana,M., Medek,P. *et al.* (2012) CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput. Biol.*, **8**, e1002708.
37. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
38. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B., Wu,C.H. and UniProt Consortium. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.
39. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
40. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
41. Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.
42. Korber,B.T., Farber,R.M., Wolpert,D.H. and Lapedes,A.S. (1993) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 7176–7180.
43. Lee,B.-C. and Kim,D. (2009) A new method for revealing correlated mutations under the structural and functional constraints in proteins. *Bioinformatics*, **25**, 2506–2513.
44. Kass,I. and Horovitz,A. (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins*, **48**, 611–617.
45. Lockless,S.W. and Ranganathan,R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
46. Weigt,M., White,R.A., Szurmant,H., Hoch,J.A. and Hwa,T. (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 67–72.
47. Olmea,O., Rost,B. and Valencia,A. (1999) Effective use of sequence correlation and conservation in fold recognition. *J. Mol. Biol.*, **293**, 1221–1239.
48. Dekker,J.P., Fodor,A., Aldrich,R.W. and Yellen,G. (2004) A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, **20**, 1565–1572.
49. Pavlova,M., Klvana,M., Prokop,Z., Chaloupkova,R., Banas,P., Otyepka,M., Wade,R.C., Tsuda,M., Nagata,Y. and Damborsky,J. (2009) Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate. *Nat. Chem. Biol.*, **5**, 727–733.
50. Gopal,S., Rastogi,V., Ashman,W. and Mulbry,W. (2000) Mutagenesis of organophosphorus hydrolase to enhance hydrolysis of the nerve agent VX. *Biochem. Biophys. Res. Commun.*, **279**, 516–519.
51. Watkins,L.M., Mahoney,H.J., McCulloch,J.K. and Raushel,F.M. (1997) Augmented hydrolysis of diisopropyl fluorophosphate in engineered mutants of phosphotriesterase. *J. Biol. Chem.*, **272**, 25596–25601.
52. Reetz,M.T., Wang,L.-W. and Bocola,M. (2006) Directed evolution of enantioselective enzymes: iterative cycles of CASTing for probing protein-sequence space. *Angew. Chem. Int. Ed Engl.*, **45**, 1236–1241.
53. Reetz,M.T., Torre,C., Eipper,A., Lohmer,R., Hermes,M., Brunner,B., Maichele,A., Bocola,M., Arand,M., Cronin,A. *et al.* (2004) Enhancing the enantioselectivity of an epoxide hydrolase by directed evolution. *Org. Lett.*, **6**, 177–180.
54. Cerdobbel,A., De Winter,K., Aerts,D., Kuipers,R., Joosten,H.-J., Soetaert,W. and Desmet,T. (2011) Increasing the thermostability of sucrose phosphorylase by a combination of sequence- and structure-based mutagenesis. *Protein Eng. Des. Sel.*, **24**, 829–834.
55. Jochens,H., Aerts,D. and Bornscheuer,U.T. (2010) Thermostabilization of an esterase by alignment-guided focussed directed evolution. *Protein Eng. Des. Sel.*, **23**, 903–909.
56. Sullivan,B.J., Nguyen,T., Durani,V., Mathur,D., Rojas,S., Thomas,M., Syu,T. and Magliery,T.J. (2012) Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability. *J. Mol. Biol.*, **420**, 384–399.
57. Pey,A.L., Rodriguez-Larrea,D., Bomke,S., Dammers,S., Godoy-Ruiz,R., Garcia-Mira,M.M. and Sanchez-Ruiz,J.M. (2008) Engineering proteins with tunable thermodynamic and kinetic stabilities. *Proteins*, **71**, 165–174.
58. Amin,N., Liu,A.D., Ramer,S., Aehle,W., Meijer,D., Metin,M., Wong,S., Gualfetti,P. and Schellenberger,V. (2004) Construction of stabilized proteins by combinatorial consensus mutagenesis. *Protein Eng. Des. Sel.*, **17**, 787–793.
59. Akasako,A., Haruki,M., Oobatake,M. and Kanaya,S. (1997) Conformational stabilities of Escherichia coli RNase HI variants with a series of amino acid substitutions at a cavity within the hydrophobic core. *J. Biol. Chem.*, **272**, 18686–18693.
60. van den Heuvel,R.H.H., Fraaije,M.W., Ferrer,M., Mattevi,A. and van Berkel,W.J.H. (2000) Inversion of stereospecificity of vanillyl-alcohol oxidase. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 9455–9460.
61. Killick,T.R., Freund,S.M. and Fersht,A.R. (1998) Real-time NMR studies on folding of mutants of barnase and chymotrypsin inhibitor 2. *FEBS Lett.*, **423**, 110–112.
62. Encell,L.P., Friedman Ohana,R., Zimmerman,K., Otto,P., Vidugiris,G., Wood,M.G., Los,G.V., McDougall,M.G., Zimprich,C., Karassina,N. *et al.* (2012) Development of a dehalogenase-based protein fusion tag capable of rapid, selective and covalent attachment to customizable ligands. *Curr. Chem. Genomics*, **6**, 55–71.
63. Reetz,M.T., Bocola,M., Carballeira,J.D., Zha,D. and Vogel,A. (2005) Expanding the range of substrate acceptance of enzymes: combinatorial active-site saturation test. *Angew. Chem. Int. Ed Engl.*, **44**, 4192–4196.
64. Morley,K.L. and Kazlauskas,R.J. (2005) Improving enzyme properties: when are closer mutations better? *Trends Biotechnol.*, **23**, 231–237.
65. Lehmann,M., Loch,C., Middendorf,A., Studer,D., Lassen,S.F., Pasamontes,L., van Loon,A.P.G.M. and Wyss,M. (2002) The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng.*, **15**, 403–411.

66. de Juan,D., Pazos,F. and Valencia,A. (2013) Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, **14**, 249–261.
67. Kuipers,R.K.P., Joosten,H.-J., Verwiel,E., Paans,S., Akerboom,J., van der Oost,J., Leferink,N.G.H., van Berkel,W.J.H., Vriend,G. and Schaap,P.J. (2009) Correlated mutation analyses on super-family alignments reveal functionally important residues. *Proteins*, **76**, 608–616.
68. Nobili,A., Tao,Y., Pavlidis,I.V., van den Bergh,T., Joosten,H.-J., Tan,T. and Bornscheuer,U.T. (2015) Simultaneous use of in silico design and a correlated mutation network as a tool to efficiently guide enzyme engineering. *Chembiochem*, **16**, 805–810.
69. Wang,C., Huang,R., He,B. and Du,Q. (2012) Improving the thermostability of alpha-amylase by combinatorial coevolving-site saturation mutagenesis. *BMC Bioinformatics*, **13**, 263.
70. Martin,L.C., Gloor,G.B., Dunn,S.D. and Wahl,L.M. (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, **21**, 4116–4124.
71. Fodor,A.A. and Aldrich,R.W. (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins*, **56**, 211–221.
72. Nov,Y. (2012) When second best is good enough: another probabilistic look at saturation mutagenesis. *Appl. Environ. Microbiol.*, **78**, 258–262.
73. Jacobs,T.M., Yumerefendi,H., Kuhlman,B. and Leaver-Fay,A. (2015) SwiftLib: rapid degenerate-codon-library optimization through dynamic programming. *Nucleic Acids Res.*, **43**, e34.
74. Li,W., Cowley,A., Uludag,M., Gur,T., McWilliam,H., Squizzato,S., Park,Y.M., Buso,N. and Lopez,R. (2015) The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.*, **43**, W580–W584.

E CD content

Attached CD contains full versions of the showcased manuscripts together with their supplementary materials. Furthermore, it contains other research papers published by the author.

```
/.............................................................root directory
├── FireProt
│   ├── fireprot_manuscript.pdf
│   ├── force_field_evaluation.xlsx
│   ├── threshold_estimation.xlsx
│   └── tools_comparison.xlsx
├── FireProtASR
│   ├── fireprotasr_manuscript.pdf
│   ├── evaluation_1.docx
│   ├── evaluation_2.docx
│   ├── evaluation_3.docx
│   └── job_success_ratio.xlsx
├── FireProtDB
│   └── fireprotdb_manuscript.pdf
├── HotSpotWizard 2.0
│   ├── hotspotwizard_manuscript.pdf
│   ├── modes_description.docx
│   ├── raphyd.docx
│   └── mutagenesis_effect.pdf
└── other_manuscripts
    ├── predictsnp2.pdf
    ├── computational_design_review.pdf
    ├── evolutionary_analysis_complement.pdf
    ├── fireprotasr_protocol.pdf
    └── enzyme_engineering_review.pdf
```