BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INFORMATION SYSTEMS
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

INTERACTIVE DATABASE FOR THE STORAGE AND MAINTENANCE OF THE BIOLOGICAL DATA

INTERAKTIVNÍ DATABÁZE PRO ÚSCHOVU A ÚDRŽBU BIOLOGICKÝCH DAT

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR
AUTOR PRÁCE

Bc. JURAJ ONDREJ DÚBRAVA

SUPERVISOR
VEDOUCÍ PRÁCE

Ing. MILOŠ MUSIL

BRNO 2021

Brno University of Technology
Faculty of Information Technology

Department of Information Systems (DIFS) Academic year 2020/2021

Master's Thesis Specification

23790

Student:	**Dúbrava Juraj Ondrej, Bc.**
Programme:	Information Technology
Field of study:	Bioinformatics and Biocomputing
Title:	**Interactive Database for the Storage and Maintenance of the Biological Data**
Category:	Biocomputing

Assignment:

1. Study issues dealing with protein stability.
2. Study the current state of data and the available sources of the protein stability data.
3. Obtain the data from the available sources (datasets, literature, ...).
4. Design a novel database of the protein stability data that would account for the problems connected with the storage and maintenance of the biological data, e.g. the inconsistencies between sequence and structural indexes.
5. Implement the database and its methods.
6. Implement interactive user interface that would allow for an advanced search in the available data.
7. (optional) Expand the database with a more sophisticated visualization of the biological data, e.g. the visualization of the sequence data and tertiary structure of the protein.

Recommended literature:

- Bava KA, et al. ProTherm, version 4.0: thermodynamic database for proteins and mutants. Nucleic Acids Research. 2004, 32, 120-121.
- Nair PS, Vihinen M. VariBench: A benchmark database for variations. Human mutation. 2012, 34, 1.
- Wang CY, et al. ProtaBank: A repository for protein design and engineering data. Tools for protein science. 2018, 27, 1118-1124.

Requirements for the semestral defence:

- First four items of the assignment.

Detailed formal requirements can be found at https://www.fit.vut.cz/study/theses/

Supervisor:	**Musil Miloš, Ing.**
Head of Department:	Kolář Dušan, doc. Dr. Ing.
Beginning of work:	November 1, 2020
Submission deadline:	May 19, 2021
Approval date:	October 27, 2020

Abstract

The main focus of this thesis is to develop a novel database of protein stability data that will focus on storage and maintenance of the experimental data. The result is a novel database FireProtDB, providing manually curated experimental data from all available sources, while presenting the data via implemented graphical user interface. The user interface provides all necessary information stored in the database with ability to search data using advanced search engine to create customized search queries targeting users seeking data for construction of the machine learning datasets.

Abstrakt

Cieľom tejto práce je vytvorenie novej databázy dát pre proteínovú stabilitu, ktorá bude udžiavať a poskytovať experimentálne dáta. Výsledkom práce je databáza FireProtDB, ktorá poskytuje manuálne overené experimentálne dáta z dostupných zdrojov a implementuje grafické užívateľské rozhranie, ktoré poskytuje dôležité informácie o dátach spoločne s možnosťou vyhľadávania umožňujúcim vytvárať dotazy na mieru a cieliacim na užívateľov, ktorí hľadajú dáta pre vytváranie dátových sád pre nástroje využívajúce strojové učenie.

Keywords

protein stability, mutations, database, machine learning, manual curation

Kľúčové slová

proteínová stabilita, mutácie, databáza, strojové učenie, manuálna kontrola

Reference

DÚBRAVA, Juraj Ondrej. *Interactive Database for the Storage and Maintenance of the Biological Data.* Brno, 2021. Master's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Ing. Miloš Musil

Rozšírený abstrakt

Proteíny dnes nachádzajú veľmi široké uplatenenie od výroby nových liečiv až po použitie v priemysle alebo poľnoshospodárstve. Väčšina proteínov sa však nevyvinula tak, aby bola schopná zniesť podmienky, ktoré sú vyžadované v ich biotechnologických aplikáciach. Jednou z najväčších prekážok ich použitia je nedostatočná stabilita pri vyšších teplotách.

Stabilné proteíny sa vytvárajú mutáciami reziduí v sekvencí proteínu. Skúmanie vplyvu takýchto mutácií sa najčastejšie určuje laboratórne, avšak laboratórne techniky sú typicky drahé a časovo náročné. Kvôli týmto problémom sa dnes čoraz častejšie používajú in silico prediktory, ktoré by pomohli určiť potenciálne zaujímavé mutácie za kratší čas. Čoraz väčšie uplatnenie v tejto oblasti nachádzajú rôzne metódy strojového učenia a v posledných rokoch už vzniklo niekoľko nástrojov využívajúce strojové učenie, ktoré sa snažia určiť vplyv mutácií na výslednú stabilitu proteínu. Avšak, výsledky takýchto nástrojov sú veľmi závislé od kvality a množstva dostupných dát na trénovanie a testovanie ich výkonnosti.

Kvalita experimentálnych stabilitných dát dnes nie je optimálna. V súčasnosti existújú tri hlavné zdroje experimentálnych dát, z ktorých je najznámejšou databáza ProTherm. Tento najväčší zdroj však trpí mnohými nedostatkami, akými sú nepresnosti v dátach, chýbajúce hodnoty, ktoré znemožňujú jednoduché použitie pre potreby vývojárov nástrojov. Problémom ProThermu je aj neaktuálnosť dát a jej neudržovanie. Preto v oblasti udržiadavania a správy stabilitných dát vzniká potreba pre zdroj dát, ktorý bude poskytovať kvalitné experimentálne dáta a umožňovať jednoduché vyhľadávanie v dátach pre potreby tvorby dátových sád pre potreby vývojárov aplikácií využívajúcich strojové učenie.

Táto práca sa zaoberá problémom návrhu a implementácie databázy stabilitných dát. Výsledkom práce je databáza FireProtDB, ktorá vznikla v spolupráci s odborníkmi z Loschmidtových laboratórií. Databáza obsahuje aktuálne dostupné dáta, ktoré sú dôkladne prefiltrované a očistené, tak aby neobsahovali nekorektné dáta. Databáza je orientovaná na výskumníkov, ktorí pracujú so stabilitnými dátami a taktiež na vývojárov predikčných nástrojov.

Implementáciu databázy FireProtDB tvorí webová aplikácia zložená z niekoľkých častí. Databáza ukladá stabilitné dáta rozšírené o sekvenčné a štruktúrne informácie, ktoré rozšíria pohľad na stabilitné dáta. Back-end aplikácie poskytuje dáta uložené v databáze vo forme REST API a implementuje najdôležitejšiu časť, ktorou je rozšírené vyhľadávanie v dátach na základe užívateľsky vytvorených dotazov. Užívatelia majú možnosť vyhľadávania pomocou full-textového vyhľadávania a tiež pokročilej možnosti tvorby dotazov. Pokročilé vyhľadávanie umožňuje tvorbu vlastných zložených dotazov, ktoré špecifikujú aké dáta sú požadované na základe použitia rôznych parametrov a kritérií. Ďalšou implementovanou časťou je grafické užívateľské rozhranie, ktoré je navrhnuté tak, aby spĺňalo typické prípady použitia. Databáza umožňuje prezeranie jednotlivých položiek na úrovni proteínu a mutácie a taktiež umožňuje užívateľom v dátach vyhľadávať za pomoci full-textového vyhľadávania a pokročilého vyhľadávania. Pokročilé vyhľadávanie je orientované na užívateľov, ktorí presne vedia, aké dáta potrebujú a umožňuje im vytvárať konfigurovateľné dotazy, ktoré využijú pri tvorbe dátových sád.

Interactive Database for the Storage and Maintenance of the Biological Data

Declaration

I hereby declare that this Master's thesis was prepared as an original work by the author under the supervision of Mr. Miloš Musil. I have listed all the literary sources, publications and other sources, which were used during the preparation of this thesis.

. . . .
Juraj Ondrej Dúbrava
May 17, 2021

Acknowledgements

I would like to thank my supervisor Ing. Miloš Musil and the whole team from Loschmidt Laboratories for their technical support, advices and suggestions.

Contents

Chapter 1

Introduction

Proteins are the key building properties of living organisms. Their use has a broad spectrum of applications in many fields such as medicine, drug discovery, agriculture, or industry. Many applications require for proteins to be able to function in harsh conditions. This ability is strongly connected with one of the protein's features, which is stability.

Protein stability determines the ability of the protein to withstand unfavorable conditions of the environment. Stability is influenced by mutations occurring in the protein sequence when one amino acid residue is replaced by another. These changes can lead to both better or worse stability, the stabilizing mutations are interesting from a usage perspective but also destabilizing mutations can be interesting, e.g. from the point of medicine. Therefore, discovering the mutations which would help to create more stable and more durable proteins is an important area of the research.

Stability is often measured in laboratory conditions, which is costly and time-consuming. In recent years, in silico methods found their use in the field of protein engineering. Many machine learning-based tools for protein stability prediction were developed to provide a less time-consuming method for determining the effect of mutations on protein stability. These methods could be very useful for selecting mutations with strong potential for further usage, but their bottleneck is the amount and the quality of available experimental data that can be used to train and test such tools. The current state of stability data is not optimal as researchers were not used to deal with systematic collection and management of the data. At this moment, there is no single central repository of stability data, data contains many errors and not much data from recent years are available in the repositories. Without higher amount of more recent data, machine learning tools cannot be further improved. Therefore, there is a need for a repository with data collected from recent literature, free of errors, and with additional features that would help to improve predictions provided by machine learning tools.

The main goal of this work was to design and develop a database of experimental thermostability data for a single-point mutations that would provide reliable data from existing sources and recently published experiments. Another goal was to create a database that could become a standard in the field of protein stability and that would be completed by interactive options for users to search, display and analyze the available data. The result of this work is a database called FireProtDB, which was created with cooperation with experts from Loschmidt Laboratories.

Chapter 2 provides a basic introduction to proteins such as types of proteins, the process of their creation, different structures and mutations occurring in the proteins. Chapter 3 presents the key information about protein stability, methods used for its measurement, and properties used to quantify protein stability. Chapter 4 includes information about currently available sources of protein stability data, description of data and existing issues. In the chapter 5, the main requirements for FireProtDB are presented and the structure of the designed model is described in more detail. Chapter 6 describes details of each part of the application structure and the last chapter 7 describes the overview of the achieved results.

Chapter 2

Proteins

Proteins are the main building blocks of all living organisms. They perform a majority of cell functions such as transportation of molecules from one location to another, enzymatic catalysis or regulation of cellular and physiological activities. However, proteins are not only important in living organisms, but they are very useful in the fields such as medicine, agriculture or industry. Because of their broad applicability, the study of the process of their creation and function is a very important subject of research. This chapter presents the basic biological introduction and describes information about types of proteins, the process of their formation, the structure of proteins and possible mutations.

2.1 Type of proteins

Proteins are biopolymers consisting of one or more polypeptide chains. The polypeptide chain is made of a sequence of amino acids connected with the peptide bond. Proteins perform many functions and each specific function is connected with their exact spatial conformation. Conformation is related to the primary structure, which is a sequence of amino acids connected in a certain order. Based on their function, proteins can be divided into several categories [40]:

- **Enzymes or catalytic proteins:** their function is to perform catalysis of chemical reactions, e.g DNA polymerases and ligases.
- **Structural proteins:** these proteins create building blocks of cells and fiber, e.g. keratin which is a building part of hair or nails.
- **Transport proteins:** they transport small molecules and ions in the organism, e.g. hemoglobin, that transports the oxygen in the blood circulation.
- **Contractile proteins:** they are the origin of movement in cells and fiber, e.g. actin or myosin.
- **Storage proteins:** they are used to store small molecules and ions, e.g. casein in the milk providing the necessary source of amino acids for newborn organisms.
- **Effector proteins:** they are used to transmit informational signals between cells, e.g. the well known protein insulin that regulates the amount of sugar in blood.
- et al.

2.2 Amino acids

Amino acids are derived from organic acids and they represent a group of molecules with one mutual property: all of them have a carboxyl (COOH) and an amino (NH_2) group. These two groups are bonded to one carbon atom called the α carbon. The difference between each amino acid is based on the sidechain. The sidechain of amino acid determines the chemical properties of the amino acid. A protein molecule is composed of the sequence of amino acids connected with the polypeptide bond that connects the carboxyl group of one amino acid and the amino group of another amino acid in the sequence. There are 20 different amino acids that can be divided into several groups according to the chemical properties of their side chain [40]:

- **Aliphatic side chain amino acids:** alanine (Ala), valine (Val), isoleucine (Ile) and leucine (Leu)
- **The acidic amino acids:** aspartate (Asp) and glutamate (Glu)
- **The amide-containing amino acids:** asparagine (Asn) and glutamine (Gln)
- **The sulfur-containing amino acids:** cysteine (Cys) and methionine (Met)
- **The basic amino acids:** lysine (Lys) and arginine (Arg)
- **The aromatic or hydroxyl-containing amino acids:** phenylalanine (Phe), tyrosine (Tyr), tryptophan (Trp), Histidine (His), serine (Ser) and threonine (Thr)
- **Containing secondary amin:** proline (Pro)

2.3 Protein synthesis

Proteins are created in the process called proteosynthesis. This process has 2 steps: transcription and translation.

- **Transcription:** the process represents the transcription of a certain part of the nucleotide sequence of DNA (gene) into a sequence of RNA. An important part of this process is an enzyme called RNA polymerase. To start the process of transcription, RNA polymerase has to first find the so-called promoter, which is a specific short sequence of nucleotides located before the sequence of nucleotides representing the gene itself. After that, the enzyme will bind to this site and transcription can begin. The process terminates when the polymerase finds the short sequence called the terminator of the transcription. The resulting RNA molecule is called mediator RNA (mRNA).
- **Translation:** in the process of translation, the information contained in the sequence of mRNA is being transferred into the polypeptide sequence of amino acids. The sequence of nucleotides in the RNA sequence is then read as triplets (so-called codons) of amino acids. Each triplet is translated into one of the twenty amino acids. With the triplets of amino acids, it is possible to create 64 different combinations, so one amino acid can be represented by several triplets. The final result of the translation is the sequence of amino acids that can be folded into protein.

2.4 Structure

Description of protein structure can be divided into four levels of organization [40]:

- **Primary structure**: a sequence of amino acids in the polypeptide chain, individual residues are linked via peptide bond.

- **Secondary structure**: it represents regular and recurring arrangements in space of neighbour amino acids in a polypeptide chain. The major conformations forming secondary structure are α-helices and β-structures. α-helix is a helix shaped secondary structure which is the most common structural motif in proteins. It is stabilized by the hydrogen bonds, which are nearly parallel to the long axis of the helix. The second mentioned β-structures include β-strands and β-sheets, where β-strands are smaller parts of the polypeptide chain that are almost fully extended. β-sheet consists of several β-strands. The stability of β-sheets is provided by the hydrogen bonds between neighboring β-strands.

- **Tertiary structure**: it represents three-dimensional conformation of folded polypeptide chain in space. The shape of the final tertiary structure is influenced by the chemical features of amino acids and their positions in the protein sequence.

- **Quaternary structure**: it represents spatial relationship between the polypeptide chains in protein molecule. This is related only to so-called oligomeric proteins consisting of multiple chains.

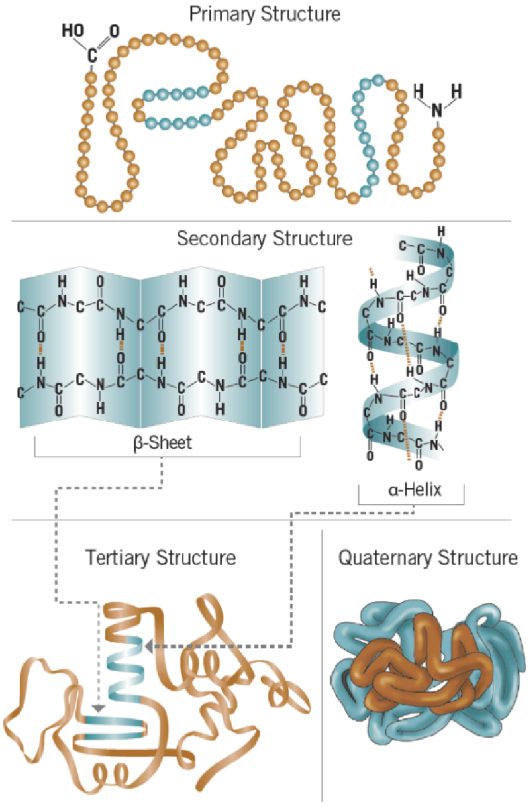


Figure 2.1: Primary, secondary, tertiary and quaternary structure of protein.[1]

2.5 Mutations

Mutations represent changes in the DNA occurring randomly or performed with a specific purpose. They are necessary for biological evolution because, without them, evolution would stop sooner or later. Changes in genetic information, which are not results of segregation or recombination of existing parts of genotype, are considered as mutations. According to the level of occurrence, mutations can be divided into three categories [2]:

- **Gene mutations:** change in the DNA represented by the change of nucleotide sequence on a specific position. They are also called point mutations and they are the most important from the predictive perspective.
- **Chromosome mutations:** structure of chromosome is changed.
- **Genome mutations:** number of chromosomes is changed.

Type of mutations

There are three main types of mutations [2]:

- **Substitution:** change of one or more nucleotides, length of the original protein sequence is not changed.
- **Insertion:** insertion of one or more nucleotides into original protein sequence, resulting sequence is longer.
- **Deletion:** deletion of one or more following nucleotides, shortens the length of original sequence.

In case that mutation occurs in the coding region of the protein, mutations can be divided into following groups [2]:

- **Synonymous mutations:** they are related to the so-called degeneration of genetic code, change of nucleotide does not have to affect the structure of the protein at all.
- **Non-synonymous mutations:** change of nucleotide in the codon leads to change of amino acid and possibly to the change in protein conformation.
- **Frameshift mutations:** deletion or insertion of a number of bases that is not a multiple of three. Usually introduces premature STOP codons in addition to high number of amino acid changes.
- **Nonsense mutations:** lead to change from amino acid to a STOP codon, resulting in premature termination of translation of the protein.

[1] Source: https://lubrizolcdmo.com/technical-briefs/protein-structure

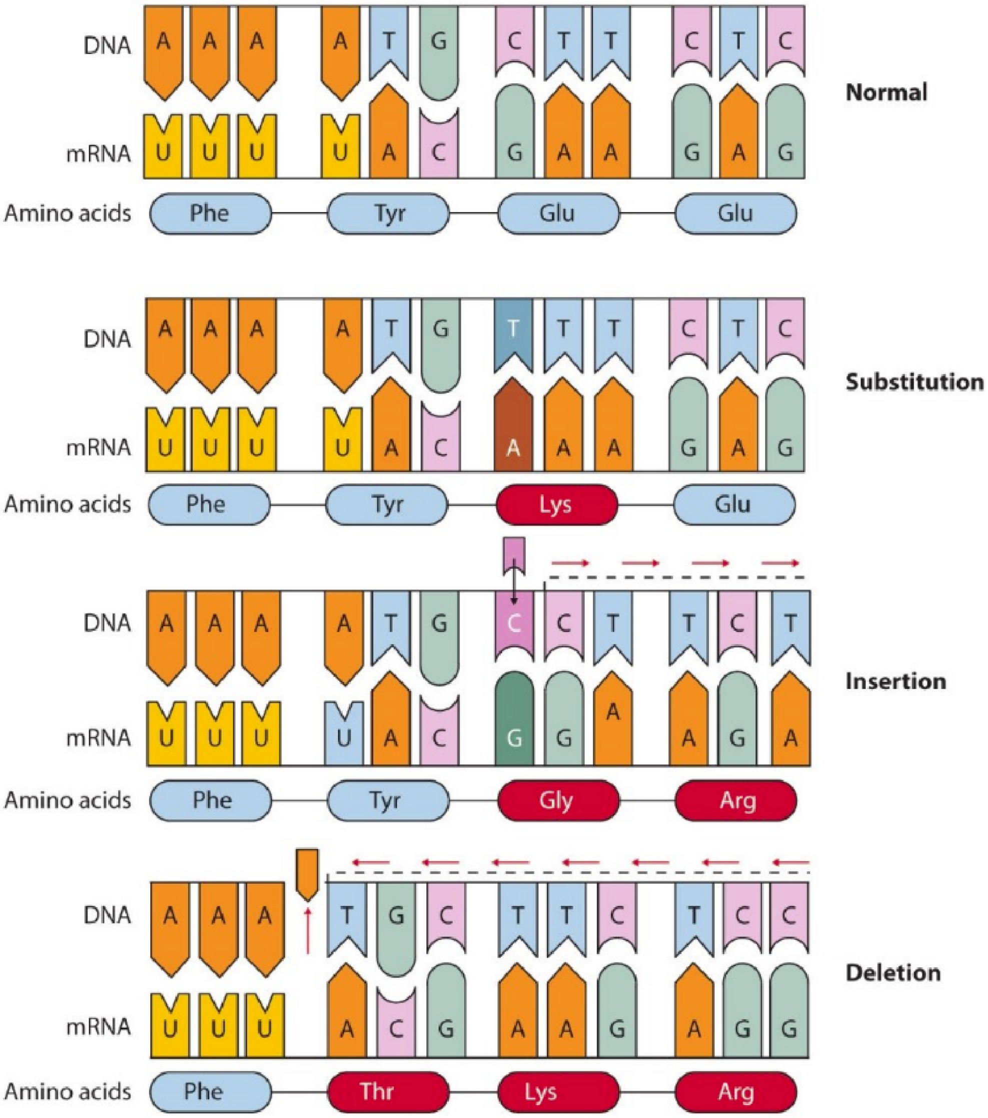


Figure 2.2: Types of mutations: substitution, insertion and deletion.[2]

[2] Source: https://alevelbiology.co.uk/notes/types-of-mutations

Chapter 3

Protein stability

Protein stability is one of the most important features of proteins and it can be described as a net balance of different forces that have different influence on stability. Their combination determines whether the protein will remain in its native state or it will denaturate to its unfolded form [18]. Stable proteins are able to withstand harsh conditions such as higher temperatures [4] or chemical properties of the environment. In many fields such as medicine, drug discovery, or agriculture there is a need to create more stable proteins as the proteins have generally evolved to function in mild conditions. Therefore, the study of the protein stability is of high interest in the field of protein engineering. Yet, this feature is still not well understood. This chapter presents two commonly used ways of quantifying protein stability and the most common methods used to measure it.

3.1 Protein stability representation

There are several standardized ways how to quantify protein stability: the two most common options are using Gibbs free energy and melting temperature, respectively. Those methods are described in more detail in the following sections.

Gibbs free energy

Gibbs free energy (G) is a thermodynamic potential which reflects the maximal amount of reversible work done by a thermodynamic system at constant pressure and temperature. It is defined by the following equation [37]:

$$G = H - TS, \tag{3.1}$$

where H represents the enthalpy, T is the temperature and S represents the entropy. In biology, value of the potential is often specified in Calories, but the official SI unit is Joule.

Using Gibbs free energy, protein stability is represented as a change of Gibbs free energy upon folding (ΔG), which means the energy required for protein to change from folded to unfolded state or vice versa. Change of the energy is defined by the following equation:

$$\Delta G = G_{folded} - G_{unfolded} \tag{3.2}$$

Mutations of amino acid residues have a significant effect on protein stability. Change of single amino acid residue in a protein sequence can significantly influence the forces that stabilize the protein. Mutations can lead to improving the stability, however they can also cause destabilization of the protein. Effect of the mutation is represented by the so-called change of Gibbs free energy upon mutation ($\Delta\Delta G$), which represents the difference between ΔG values of wild-type and mutated protein.

$$\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wild\_type} \tag{3.3}$$

Commonly used unit for value of $\Delta\Delta G$ is $kcal/mol$. If the value is calculated by equation 3.3, negative value of $\Delta\Delta G$ means stabilizing mutation. Because of non-standardized format of $\Delta\Delta G$, stabilizing mutation can also be represented by a positive value. This is caused by switching the wild-type and mutant.

Melting temperature

Melting temperature is another way how to represent protein stability. Equation defining melting temperature is following:

$$\Delta G_{folding}(T_m) = 0 \tag{3.4}$$

To explain the meaning of T_m, it is the temperature at which free energy of the folded and unfolded states is equal, and half of the population is unfolded, and the other half is folded. Value of ΔT_m indicates the change of melting temperature upon mutation. Therefore ΔT_m and $\Delta\Delta G$ are very similar, even though there is no simple way how to transform these values between each other.

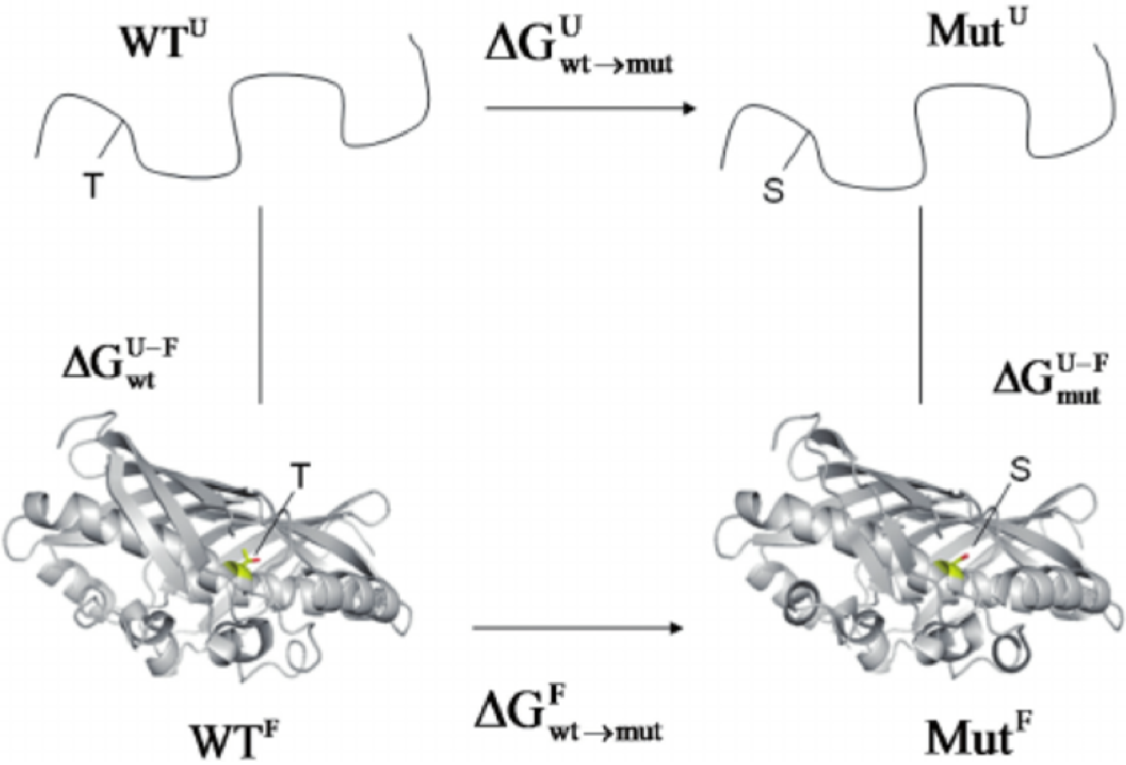


Figure 3.1: Thermodynamic cycle used for computation of $\Delta\Delta G$.[1]

[1] Source: https://bit.ly/36WiQ4h

3.2 Protein stability measurement

Measurement techniques for protein stability can be divided into two categories: methods used in laboratory conditions and predictive approaches making use of the computational resources. These techniques are described in more detail in the following sections.

Laboratory methods

Several laboratory methods exist to measure protein stability [18]. Results provided by these methods depend on exact conditions during the experiments and small differences are also caused when using a different technique. When comparing results from several measurements, only results obtained by the same method and with the same conditions during the experiment should be considered. The most commonly used techniques are:

- **Circular dichroism:** it is a method based on differential absorption of left and right circularly polarized light. Left-hand circular and right-hand circular polarized light represent two possible spin angular momentum states for a photon. Chiral molecules, which are optically active, will preferentially absorb one direction of the circularly polarized light. This technique is commonly used to determine the aspect of the secondary structure of proteins [1].

- **Differential scanning calorimetry (DSC):** it is a thermoanalytical method. This technique measures the difference between the amount of heat, which is required to increase the temperature of a sample, and a reference as a function of temperature [13]. Valuable thermodynamics information about proteins can be obtained by using this method.

- **Absorption spectroscopy:** this technique measures absorption of radiation as a function of wavelength or frequency, due to its interaction with the sample. Energy is absorbed by the sample and the absorption intensity varies as a function of frequency [12]. The method finds its usage as an analytical chemistry tool.

Other used techniques include nuclear magnetic resonance (NMR) spectroscopy, fluorescence (Fl) and others.

Predictive approaches

Protein stability does not have to be obtained using only laboratory methods. With the progress in the field of information technology and with more available computational resources, different approaches to calculate protein stability can be used. The most well-known approaches are force-field measurement and machine learning-based prediction tools.

Force-field measurement

Technique of force-field measurement is based on simulation of existing force-fields to calculate free energy of the protein of interest. Force-fields are simulating effects of physicochemical properties on protein structure. Calculation of such force-field is shown in the following simple example [18][28]:

In this example, free energy in the folded state is defined as

$$G_F = G_{hy} + G_{el} + G_{hb} + G_{vw} + G_{ss}, \tag{3.5}$$

where G_{hy} is hydrophobic, G_{el} is electrostatic, G_{hb} is hydrogen bonding, G_{vw} is van der Waals and G_{ss} is disulphide bonding free energy.

The hydrophobic and hydrogen bonding interactions are the two main contributors to the protein stability. Hydrogen bonds play a significant role in the process of creation of secondary structure. Their contribution to this process is mostly based on geometric information. The charged side chains of residues Lysine, Arginine, Histidine and Glutamic and Aspartic acid are the most important contributors to electrostatic interactions. Free energy in the unfolded state is defined as

$$G_U = G_{en} + G_{ne}, \tag{3.6}$$

where G_{en} is entropic free energy and G_{ne} is nonentropic free energy.

Machine learning

In recent years, machine learning has been used to solve different tasks in the field of protein engineering, including protein stability prediction. This approach could save a lot of time and money as it is less time-consuming than laboratory methods. Unfortunately, it is not yet accurate enough to fully replace the laboratory measurements.

The usage of machine learning can be helpful in the process of designing new stable proteins. To obtain a protein with the required parameters, a large number of mutations have to be designed and each of them require time-consuming laboratory measurement. Also, a significant portion of these mutations will be ineffective and will lead to decrease in stability. Therefore, there is a dire need to speed up this process and make it more efficient. With machine learning methods, interesting and promising mutations leading to improvement in stability could be chosen in a short amount of time and then further studied using laboratory methods.

The problem of protein stability determination is usually being solved by the supervised methods. Using this approach, methods require training and testing data. Each data sample have to be labelled by the effect of the mutation, so for each mutation, the corresponding class will be determined from the value of $\Delta\Delta G$ or ΔT_m. Designers of such tools also choose other different features to enrich experimental properties to improve their tools. These features can also have a very significant influence on stability. Machine learning models could be able to identify the underlying dependencies between these properties better than any other expert and it could lead to significantly better results. This step depends on the creators of the model, because they can choose different structural and sequential properties to explore potential connections between different features and protein stability that would influence the performance of the model. Several machine learning-based tools were recently developed to solve this task. Tools such as the AUTO-MUTE [25] and I-Mutant 2.0 and 3.0 [10] use support vector machines (SVM) or decision trees for prediction, which are the most common methods used in stability prediction tools.

The accuracy of the predictive tools is also based on the amount of available data, their diversity and quality. Those are the crucial issues in the field of protein stability. A crucial feature that the predictive tools need to deal with is the balance of the provided training data. Stability data are significantly imbalanced and stability datasets usually contain more destabilizing entries than stabilizing, which has a significant impact on the results of predictive models. Some algorithms could be very sensitive to this issue, so designers need to choose and try the suitable one. One of the strategies used to overcome the imbalance and overfitting is the ensemble strategy. It combines more than one classifier and results are produced by the combination of partial results from the classifiers. One of the representatives of such a strategy is the tool ELASPIC [41].

Despite the usage of different machine learning methods and techniques to overcome problems with the data, results from prediction tools cannot be considered very accurate because of the underlying character of the data they use. Many tools report very high accuracy of prediction and these results are often unreliable because prediction tools have to use small and unbalanced datasets due to the insufficient size of the available data. Huge improvement in the performance would be attained with higher amount of reliable experimental data, which size increase on very low pace. Another important issue connected with the reliability of the tools is the problem to reproduce the results. Many developers of the tools use their own data, which are not part of any stability database or dataset. These data are then not published with the results and researchers also usually do not tend to provide the data for storage in the stability databases. The results of such tools without provided training and testing data are hard to verify.

Chapter 4

State-of-the-art

Quality experimental data is the key property in the field of protein engineering. In the past, researchers usually did not tend to care much about collecting data from their experiments and maintaining them in some kind of repository for usage in the future. A lot of experimental measurements are just present in the scientific papers without additional validation and storage of provided results in some kind of database. With the increasing usage of methods that require a big amount of data, this issue has become more troublesome. Data are a necessary part of the creation of the machine learning-based tools for different types of problems in protein engineering, which also includes a prediction of protein stability. Without proper data storage with sufficient amount of high quality data, no significant improvement in development of the computational tools can be done.

The current state of the experimental stability data is not sufficient to improve the performance of the computational tools. Currently available sources of data suffer from having outdated and inconsistent data, which are not validated after their report. This chapter presents more detailed overview of the current sources of protein stability data and their properties.

4.1 ProTherm

ProTherm [5] represents a source of thermodynamic data for proteins and mutants. The first version was published in 1999 and it is, therefore, a pioneer among all other sources of such kind of data. The main goal to build ProTherm was to establish a uniform source of thermodynamic data where experts could find necessary data and also provide data from their experiments and extend the knowledge stored in the database.

The current version contains more than 25,000 entries from 740 unique proteins and it includes both single- and multiple-point mutations. ProTherm is used as the main source of data for many recently developed prediction tools, but it has not been updated since 2013. Therefore, it is hard to build better prediction tools with a lack of new data. Table 4.1 shows more detailed statistics of the data stored in ProTherm:

Data organization

ProTherm contains thermodynamic data, structural information, measuring methods, experimental conditions, reversibility of folding, and literature information for wild-type proteins and also for both single- and multi-point mutations.

Number of entries:	25,820
Number of unique proteins:	740
Number of single-point mutations:	12,561
Number of multi-point mutations:	2,876
Number of wild-types:	10,383

Table 4.1: Statistics of current ProTherm version

Database entry has a unique identifier and includes four different sections of information. Each section includes several values describing information about a certain topic.

The first section contains information about sequence and structure. It includes basic information such as protein name with its source, PIR [16] and Swiss-Prot [6] (UniProt [11]) identifiers to identify the protein. Other provided identifiers are PDB [36] identifiers for wild-type and mutant structures and also enzyme commission (EC) number. More specific information about protein includes length and molecular weight of the protein calculated from the sequence information that is given by PIR, secondary structure, accessible surface area and information about mutation such as position or type of the mutation.

The second section contains information related to experimental conditions used during the laboratory measurement. It includes information about temperature, the value of pH, type of the measurement, the method used in the experiment such as the ones described in 3.2. Other values provided in this section are buffers and ions with information about concentrations and additives.

The most important section is the one with thermodynamic information containing important values about stability. It includes important values about Gibbs free energy such as unfolding Gibbs free energy change when denaturant is present or not (ΔG and ΔG^{H_2O}) and difference of Gibbs free energy as an effect of a mutation in the presence and absence of denaturant ($\Delta\Delta G$ and $\Delta\Delta G^{H_2O}$). Information about melting temperature is also present if used in the experiment. Information includes values of (T_m) and change in melting temperature (ΔT_m). Other important parameters related to the measurement are van't Hoff enthalpy change (ΔH_{vH}), change of calorimetric enthalpy (ΔH_{cal}), the slope of denaturation curve (m), midpoint of denaturant concentration (C_m), heat capacity change (ΔC_p), folding reversibility and activity.

The last section includes important information about the source of the experimental data. It stores keywords, reference to the source represented by the name of the publication, authors, remarks and also related entries in the database.

Data problems

As mentioned previously, ProTherm is a major source of the thermodynamic data, but many inconsistencies were reported [15]. Usage of such data is then very time-consuming because of the additional filtering and validation process. Many datasets derived from ProTherm were created to overcome this issue and provide a reliable source of data. These subsets are much smaller in comparison with ProTherm and often overlap.

Critical issues that were reported by the users include:

- redundancy of the entries
- absence of $\Delta\Delta G$ value – many entries do not include this information at all or they include ΔT_m instead. Without one of the values, entry is useless for further usage.

- incorrect sign of $\Delta\Delta G$ – redundant data often contain disagreement in thermodynamic parameters, which includes the opposite sign of $\Delta\Delta G$. Such inconsistency will completely change the meaning of the value, and the user has to determine which one is correct.

- absence of intermediate state – data in ProTherm does not account for the existence of the intermediate state in the process of protein folding, so many entries can contain only part of the folding process data.

- missing or incorrect reference to identify the protein – many entries do not contain UniProt or Swiss-Prot identifiers. Without this information, there is no way how to identify the target protein, because protein sequence is not included. Some of the entries have also problems with presence of obsolete identifiers.

4.2 VariBench

VariBench [34] is a benchmark database for variation data. Its main purpose is to provide verified high-quality data used for performance evaluation of different computational methods which are used for predicting the effects of variations, or training newly developed predictive methods.

There are five different categories of variation datasets in VariBench in its current version:

- **Tolerance datasets:** these datasets contain information on whether missense variants are tolerated (i.e., benign) or not (functionally impaired) in proteins.

- **Protein stability datasets:** they contain experimentally studied effects of variations on protein stability.

- **Mismatch Repair Gene Variants datasets:** datasets contain mismatch repair variants with a known functional effect.

- **Variations Affecting Transcription Factor Binding**

- **Variations Causing Splice Site Aberrations**

Protein stability datasets

Twenty-two protein stability datasets are part of the VariBench database. Many of them are subsets derived from the ProTherm database. For each of them, VariBench provides similar information that includes the PDB identifier, position of the variation in PDB structure, PDB chain identifier, corresponding UniProt identifier, variation position in protein sequence, and change in the Gibbs free energy.

- **Potapov et al. [29]:** it consists of 2,156 single variations filtered for determined structures. Multiple $\Delta\Delta G$ change measurements for a single variation are replaced with the average value.

- **Khan and Vihinen [22]:** dataset contains 1,784 missense variations from 80 proteins with experimentally determined $\Delta\Delta G$ values. Experiments with $\Delta\Delta G$ values between 0.5 and -0.5 kcal/mol were labeled as neutral, experiments where $\Delta\Delta G \leq -0.5$ were labeled as stabilizing and entries with $\Delta\Delta G \geq 0.5$ were labeled as destabilizing. The dataset includes 931 destabilizing, 222 stabilizing and 631 neutral entries. Authors created this dataset to evaluate performance of several prediction tools.

- **Capriotti et al.[9]:** dataset with 1,615 single variations from 42 proteins with known three-dimensional structures and experimentally determined $\Delta\Delta G$ values. There are two subsets derived from this dataset, the first is S1615 and includes 1,615 variations from 42 proteins, the second one is S388 with 388 variations from 17 proteins and is a subset of S1615. Experiments in S388 were performed at physiological conditions (pH in range 6-8 and temperature in range of 20° - 40°C). Authors created these datasets for their needs during the development of the I-Mutant2.0 prediction tool. S1615 was used to train the model and S388 was a test set.

- **Guerois et al. [19]:** it represents a dataset composed of two subsets. The first contains 339 experimentally studied variants in 9 proteins and the second contains 625 entries with a single conservative variation in a monomeric protein studied with pH values in range from 6 to 8.

- **PON-Tstab dataset:** dataset containing 1,564 variations from 99 proteins, used for training and testing of PON-Tstab [43] prediction tool.

- **IMutant2.0 S2087,S1948:** datasets used for training and testing of I-Mutant2.0 prediction tool, S2087 contains 2,087 variants with sequence information and S1948 includes 1,948 variants with three-dimensional structures.

- **Saraboji S1791,S1396,S2204:** datasets created by Saraboji et al. [33], S1791 contains 1,791 variations with known PDB structure, S1396 contains 1,396 variations with thermal denaturation and S2204 includes 2,204 variations with chemical denaturation.

- **iPTREE-STAB S1859 dataset:** dataset with 1,859 single variations from 64 proteins used for iPTREE-STAB [20] prediction tool.

- **SVM-WIN31_SVM-3D12 S1681,S1634,S499 datasets:** datasets used in tools SVM-WIN31 and SVM-3D12. S1681 contains 1,681 substitutions from 58 proteins, S1634 contains 1,634 variations in 55 proteins with available PDB structures and S499 includes 499 additional variations [10].

- **PoPMuSiC-2.0 S2648 dataset:** 2,648 substitutions from 131 proteins used in PoPMuSiC-2.0 [14] prediction tool.

- **sMMGB S1109 dataset**: dataset used in sMMGB [44] tools containing 1,109 variations.

- **M47_M8 datasets:** it consist of S2760 dataset with 2,760 variations in 75 proteins and S1810 with 1,810 variants in 71 proteins. Datasets are used in M47 and M8 tools [42].

- **EASE-MM dataset:** dataset consisting of three datasets, S238 contains 238 variations and is a sub selection of I-Mutant2.0, S1676 contains 1,676 variants and S543 contains 543 variations in 53 proteins and represents a subset of PoPMuSiC-2.0 S2648. Dataset was used in EASE-MM [15] tool.

- **HoTMuSiC S1626 dataset:** dataset with 1,626 variations in 90 proteins used for HoTMuSiC [31] prediction tool.
- **SAAFEC dataset:** dataset consists of S1262 containing 1,262 variants in 49 proteins and S983 containing 983 variants in 42 proteins with known three-dimensional structure, used for SAAFEC [17] prediction tool.
- **STRUM dataset:** dataset consists of Q3421 with 3,421 variants with available structures and Q306 including 306 variants from 32 proteins. These sets were used in the STRUM [32] prediction tool.
- **Broom S605 dataset:** it is used in metapredictor created by [7] and including 605 variants in 60 proteins.
- **Automute dataset:** dataset consists of S1962 with 1,962 variants from Saraboji S2204, S1925 includes 1,925 selections from I-Mutatant2.0 S1948 and S1749 includes 1,749 variants selected from Saraboji S1971. Datasets were used in the Automute [25] prediction tool.
- **TP53 dataset:** it represents a dataset of 42 variants in TP53 protein, used in mCSM [27] prediction tool.
- **S^{sym} dataset:** dataset composed from 684 variants inserted in 357 structures [30].
- **PSTAB datasets:** it consists of set of six datasets: first contains 768 hot-spots of Alanine-scanning mutations [24], the second consists of 2,971 entries from ProTherm, the third dataset includes 2,154 mutations from Potapov, the fourth contains 1,005 variations from Guerois, fifth dataset includes 380 variations [23], and the last one consists of 1,210 substitutions [21].

4.3 ProtaBank

ProtaBank [38] is a repository for different types of protein engineering data. The database is not primarily focused only on stability data and a wide range of properties are provided, including properties of solubility, folding, binding, or activity. ProtaBank database provides the ability to query and analyze the data as well as submission of new data entries by the users.

The database gathers mutational data from different sources and approaches. It includes data obtained from computational and other types of rational design, saturation mutagenesis, directed evolution, and deep mutational scanning. The data comes from published literature and if new data should be stored into the database directly from users, the data source is validated before the submission.

Search and analysis options

The database provides a web interface for users to search in the data and analyze the features of single entries.

Searching can be done using several properties such as publication or study details (title, abstract, author), protein name, PDB identifier, UniProt accession number, or protein sequence string. No advanced options for searching according to more specific properties of the data are available, at least in its non–commercial version.

Search results are represented by the study entries related to the mutational data. After search, table with information about study names, protein names in the study, authors, and year of publication is displayed. Users can only use this information to filter data in the results table, no more specific options for filtering are available. Another searching strategy that can be used in the database is the option of BLAST [3] search based on the specified protein sequence. With this approach, data related to this sequence can be identified.

In the non–commercial version of ProtaBank, users do not have an option to download search results in any format. Without possibility for advanced search and filtration, the usage of the database is very limited for users seeking data for new prediction tools without hard manual search in the data.

Data analysis can be done on the level of study entry. Single entry includes more specific information about the study, protein sequence, and three-dimensional structure where mutational data can be mapped and shown. Mutational data are displayed in a table with sequence information, wild type amino acid, the position of mutation and mutated amino acid, the value of $\Delta\Delta G$ and units used in the measurement. Data from the table can be downloaded in several formats such as CSV or XLS, but no additional annotations are included, so user have to manually add annotations about related protein or structure.

Chapter 5

FireProtDB design

Chapter 4 presented an overview of the current state in the field of protein stability data. Availability, quality and amount of the experimental data obtained in the laboratory play an essential role when using in silico methods such as machine learning. Data obtained in the experiments are nowadays split among several resources without any defined format of the data. Data are suffering from many inconsistencies leading to existence of many manually filtered subsets of the data used specifically for construction of the predictive tools. The data acquisition from many resources with possibility of inclusion of the incorrect data and need to manually filter the results make these data sources hard to use. This chapter describes the design of the novel database FireProtDB, a database of manually curated protein stability data.

5.1 Requirements

Requirements for the FireProtDB are considered from the data and user perspective and they need to meet certain specified criteria. From the data perspective, database has to provide certain type of data with sufficient quality. Requirements for the data stored in the database have to incorporate following criteria and information:

- **reliability:** data should be without any errors, mismatches and duplicates.
- **protein information:** data should provide the most important information about protein such as protein sequence, PDB identifier, UniProt identifier and other useful information.
- **residue information:** information about single residues should be provided. Data should contain information about related protein sequence, conservation information, correlation with other residues, secondary structure and other structural information such as position in the protein tunnel, pocket or whether residue is catalytic.
- **mutation information:** information about residues mutations should contain mutated amino acid residue, position of mutation, values of predicted mutation and information about related experiments.
- **mutation experiment information:** information about experimental thermostability data should contain exact values of $\Delta\Delta G$ or ΔT_m, information about experiment conditions such as pH, half-life and other parameters.

- **dataset information:** information about related dataset should be available for mutation.

- **publication information:** information about related publication for experiments and datasets should be available.

Another important requirements are put on the database from the users perspective. They are focused on data presentation and ability of searching in the database. The criteria are following:

- **user interface:** database should provide a graphical user interface for displaying protein information (e.g. sequence, structure), information about mutation and its related experiments, publications, and datasets.

- **search:** database should provide full-text search option to search according to key features such as protein name, UniProt identifier, PDB identifier, or publication. For creating highly customized search queries, the advanced search should be provided to create search queries dynamically according to more specific features of the protein, mutation, related experiments, or dataset.

- **export and filtering:** database have to provide the ability to download search results in a specific format and filter them according to certain specific features.

- **API:** database should provide an API for users to obtain data with their own scripts.

5.2 Architecture

FireProtDB is designed as an application consisting from several layers to satisfy all requirements specified in 5.1.

Database stores all important information about proteins, mutations, thermodynamic experiments and other additional data. *Web service* will provide all necessary interactions with database and also searching ability. *Web server* will connect web service and application. All incoming requests will be handled by the server. *Web application* will provide graphical user interface to present data on several levels as well as ability to define custom search queries.

Database design

Database layer of FireProtDB is the core of the application. It is designed as a traditional relational database. The top of the database hierarchy is represented by a unique protein sequence with a UniProt identifier. The sequence is preferred to structure because of the availability of sequential data instead of tree-dimensional structures.

The sequence is represented by a string of amino acid residues on certain positions. Each position can be related to multiple mutations and each mutation can be assigned to multiple experimental measurements. Each measurement includes the most crucial information about measurement, information on whether the value was manually validated, and related publication if available.

Structural information includes one or more biological units related to certain protein sequences that identify biologically relevant quaternary structures of asymmetric units available in the PDB database. Additional structural and sequential annotations were computed for stored biological units. This information provides another useful data about mutations and can be utilized as features in predictive tools.

Following section will provide more detailed description of individual tables and their properties:

- **Table sequence:** table contains information about protein sequence. It stores a string of amino acids, UniProt identifier, name of the protein, organism and enzyme commission (EC) number.
- **Table uniprot:** table contains information about other related UniProt entries to certain protein sequence. It provides only a single value defining the relation of UniProt entry to protein sequence, whether it is obsolete or not.
- **Table interpro_families_domains:** table provides the type and name of InterPro families which are related to certain protein.
- **Table sequence_interpro:** table connects related protein sequences and InterPro families. It also stores the order of InterPro entry expressing the importance of this entry.
- **Table structures:** table stores information about protein structures such as PDB identifier, the method used to obtain structure and resolution of structure.
- **Table sequence_structures:** table connects protein sequence with all known determined structures.
- **Table biological_unit:** table stores representative biological unit for a specific sequence. It includes reference to the structures table in the form of a PDB identifier and also includes a unit number.
- **Table sequence_biological_unit:** table creates a connection between protein sequence and the biological unit, contains information about the old and new chain. New chain represents chain id used in HotSpot Wizard application and old chain represents chain id in the original PDB file.
- **Table hsw_jobs:** table contains information about HotSpot Wizard jobs that were used to calculate structural information related to the certain biological unit.
- **Table catalytic_pocket:** table provides information about catalytic pocket inside protein structure. It includes relevance, volume, and drugability values.
- **Table tunnels:** table stores information about tunnels inside catalytic pockets. The table includes the length of the tunnel, distance to the surface, its curvature, and throughput.
- **Table pocket_tunnels:** table creates a connection between a catalytic pocket and related protein tunnel and includes coordinates of starting position of the tunnel.
- **Table bottleneck:** table includes information about a bottleneck in protein tunnel. Data included in the table are radius, starting coordinates, and ball number of the bottleneck.

- **Table bottleneck_residues:** table creates a connection between bottleneck and residues in the bottleneck. Information about the presence of sidechain is also provided.

- **Table residues:** table contains information about residue belonging to a specific protein sequence. It includes a position of residue, amino acid, and its conservation.

- **Table residues_catalytic_pocket:** table creates a connection between catalytic pockets and certain residues inside them.

- **Table catalytic_annotations:** table provides information about catalytic residues. It includes reference to the residue, source, accession code, identity, description, type, and neighborhood.

- **Table structure_annotations:** table contains structural information about the residues defining the structure of the protein. The table contains information about related residue and the biological unit, b-factor value, the value of accessible surface area (ASA), type of secondary structure in which is residue presented, structure index representing position of residue in structure, new chain, and insertion code.

- **Table correlated_residues:** table contains information about correlated residues related to target residue. It includes the position of correlated residue, related sequence, and correlation consensus value.

- **Table mutations:** includes information about the mutation of specific residue. It stores mutation identifier, a reference to the residue, and mutant residue.

- **Table consensus_annotations:** table includes information about mutations leading back to consensus. It stores a reference to mutation, residue after mutation, frequency, and the ratio of the mutation.

- **Table stability_experiments:** table contains information about experiment related to certain mutation. It stores a reference to mutations table and values related to experiment settings such as scan rate, protein concentration, T offset, and pH. Another measurement-related information is the method and technique used to obtain values in the experiment and their details. The most important values are those related to protein stability such as $\Delta\Delta G$, ΔT_m,C_p, half-life or curating flag whether experimental data were manually verified in literature.

- **Table experiments_datasets:** table connects corresponding experiment with the dataset in which experiment occurs.

- **Table datasets:** table contains information about datasets with experiments. Table itself contains information about its name, version, and links to the dataset and homepage of the institution which constructed the dataset.

- **Table publications:** table contains information about publication in which the specific experiment was presented. It includes the title of the publication, name of the journal, volume, issue, year of publication, pages, DOI, and PMID identifiers.

- **Table authors:** table stores information about the dataset or publication authors with their first name, last name, and initials.

- **Table datasets_authors:** table connects related datasets and authors
- **Table authors_publications:** table connects related publications and authors

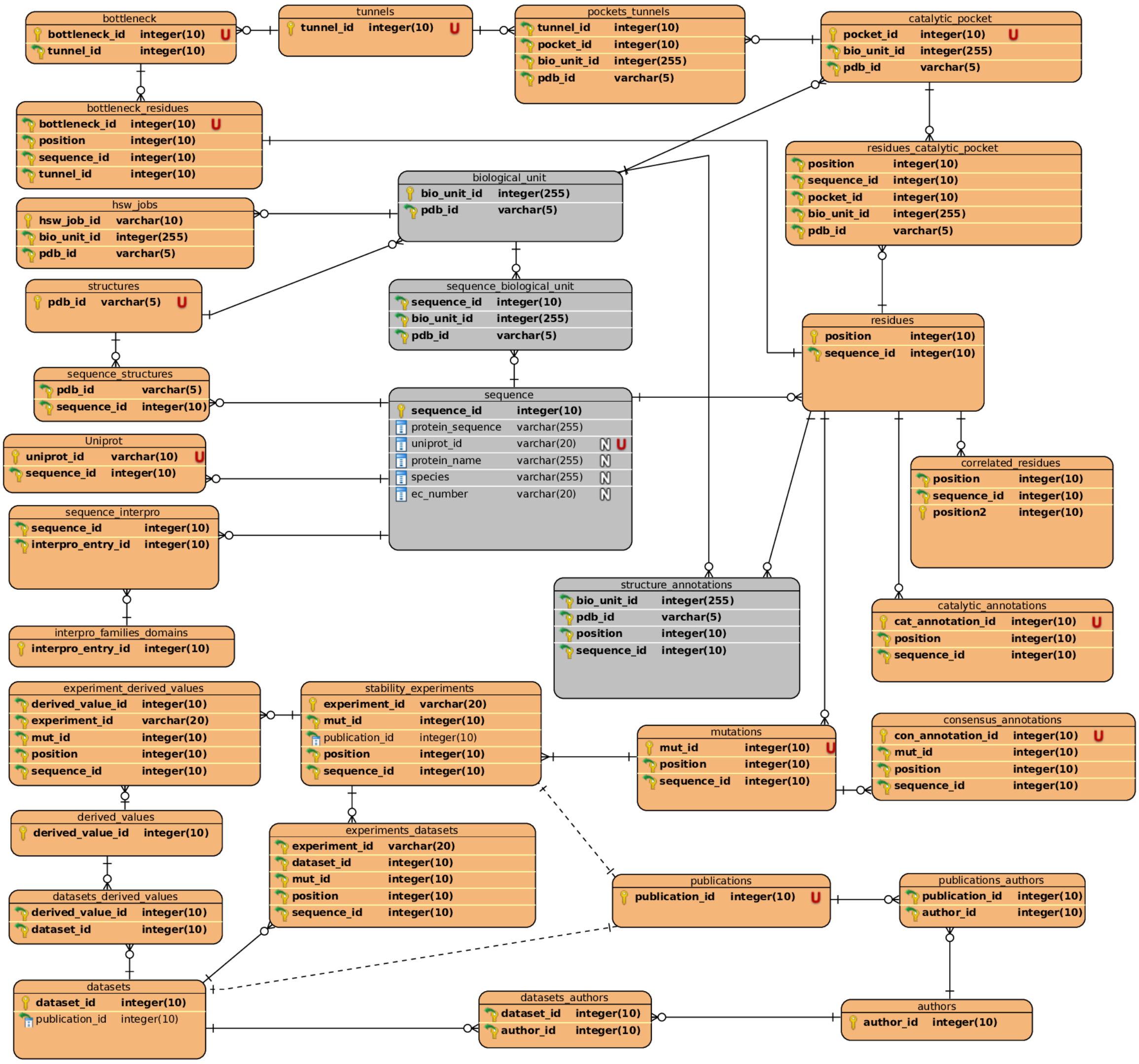


Figure 5.1: Entity relationship diagram of FireProtDB.

Figure 5.1 shows design of the FireProtDB as entity relationship diagram. Tables do not include attributes stored by each table because of the complexity of the database. For simplification, information stored in tables are described in the previous more detailed description of individual tables. From the design perspective, authors have to be connected with the dataset as well as with publications, because existing dataset do not have to be published and therefore no information about its authors would be available if it would be connected only via its publication.

5.3 Sources of data

Experimental and additional data in the FireProtDB were obtained from the following sources:

- **ProTherm database:** the main source of the stability data used in the FireProtDB. For purpose of FireProtDB, only single-point mutation experiments were chosen for further usage.

- **ProtaBank database:** database was searched for studies in which the stability data are present. Data from this source add experiments from more recent studies to enrich the data from ProTherm.

- **VariBench database:** database contains 22 different stability datasets which were obtained from this source. Provided datasets were used in training and testing of many existing prediction tools and in most cases consist of entries derived from ProTherm. Therefore, data from these datasets are primarily used to add information about dataset membership for ProTherm data.

- **External literature:** data from more recent literature sources obtained by the team in Loschmidt Laboratories.

- **Loschmidt Laboratories:** database also incorporates experimental data, which were obtained during experiments done by the team in Loschmidt Laboratories.

- **HotSpot Wizard:** tool for automated design of mutations and smart libraries [35]. This tool was used to obtain additional sequence and structure-related data such as catalytic pockets, tunnels, or residues located in the structure as it provides all necessary data in a single calculation. The additional features were added to provide more information about structure-function relationship and to be used as potential features for training of the machine learning models.

5.4 Problems with the data

Nowadays, there is a rapid growth of the biological data, but in general, this type of data often suffers from the various errors, inconsistencies, or unspecified data format. These issues require manual correction and validation of the data, which was also an issue while obtaining data for FireProtDB. Other problems arise from relationships between biological data, which have to be incorporated into the database design.

Incorrect data problems

The main source of issues related to the data inconsistency and errors was ProTherm, which is known for these problems. Therefore, data selected from this database have to meet certain requirements and several steps of preprocessing were done to overcome this issue.

In the first step, only single-point mutations were selected from all of the entries. The second step included a selection of mutations, which were not insertions or deletions. In the third step, entries including at least one of the SwissProt (UniProt) or PDB identifiers were selected. Without one of the identifiers, the target protein cannot be identified and such experimental entry is useless. The last step included the selection of the entries that had at least one of the values of $\Delta\Delta G$ or ΔT_m.

FireProtDB is a database of manually curated data, so each entry was further validated by the team of Loschmidt Laboratories from the original publications. After this manual validation, entries for which the original publication could be found were further checked for errors and then labeled as curated.

Design problems

The design of the database reflects problems related to the relationships between biological data. First of all, the protein sequence was chosen as a top-level representation instead of structures as the number of available sequences is much higher than the number of known structures. The sequence has to be then connected to a representative biological unit. Each unit can be related to one or more sequences and composed of multiple chains. This problem is solved with table *biological_unit*, where the representative biological units are stored and the connection table between sequence and the biological unit contains information about the chain. Biological units have to be connected to the PDB structure because table *biological_unit* stores only representative structure, but protein sequence can have more than one known structure. All structures are therefore stored in separate table *structures* that needs to be connected with *biological_unit* and also with table *sequence_structure*. This table makes a connection between protein and all related structures.

The next problem is related to the information about residue in the protein sequence. Residues in the FireProtDB are designed to store position that reflects the position in the protein sequence, but the data can contain positions that are determined as positions in the structure. This issue was connected especially with ProTherm entries, which use structure position index. Therefore, positions had to be transformed into their sequence counterpart using the Needleman–Wunch [26] algorithm to create global sequence alignment of sequence related to PDB structure and UniProt sequence entry. After this step, mutations were checked for correct positions in the alignment. The issue with the position in structure and sequence is related not only to the residues on mutated positions but to all residues. For the ability to visualize residues that are part of the structure, table *structure_annotations* is defined to provide pre-computed features of such residues and also mapping between position in sequence and structure. This table serves as a connection between the reference structure for protein and all residues that are part of the structure. All features in this table are currently computed only for one structure, that is the reference one, to provide ability for displaying residues in the tree-dimensional structure.

An important problem that can occur in the data is the incorrect UniProt identifier for identifying the target protein. In many cases, data contained UniProt identifiers that were obsolete and the UniProt database was already containing entries with updated identifiers. To provide also old identifiers for search purposes, an additional table UniProt was added to provide all known UniProt identifiers for the protein and their relationship with protein for entries that were replaced.

Experimental data are provided in the form of entries for individual measurements of certain mutations. Therefore, each mutation can be related to multiple experiments with different results and conditions. Table *stability_experiments* is used to store all thermostable data measured during the experiment. In some publications, experiments can be sometimes reported with an average or mean value of $\Delta\Delta G$ or ΔT_m from all the measurements. To capture this issue, table *derived_values* was created to store these values and their type. The table is connected to the table with the experiment and also to the datasets table.

5.5 Data statistics

From the available sources, 15,989 experimental entries were obtained, originating from 8,411 mutations. Provided experiments are from 242 different proteins representing 304 known structures. Number of experimental entries from individual sources is shown in the table 5.1.

Source	No. entries
ProTherm	8180
VariBench	5636
Literature & ProtaBank	2114
Loschmidt Laboratories	49

Table 5.1: FireProtDB data statistics with number of experimental entries and source.

Critical issue with the experimental stability data is their imbalance. The most of the experiments belong to destabilizing mutations and this issue is also reflected in the amount of such entries obtained for FireProtDB. Stabilizing entries are in minority and increase in their amount would help to create more balanced datasets. Data contains 46 % of destabilizing entries, where $\Delta\Delta G > 1$ or $\Delta T_m < -1$, while stabilizing entries, where $\Delta\Delta G < -1$ or $\Delta T_m > 1$, make up 10 % of all entries and remaining 44 % of entries, where $-1 \leq \Delta\Delta G \leq 1$ or $-1 \leq \Delta T_m \leq 1$, are considered as neutral.

Figure 5.2 shows a distribution of number of experimental entries in intervals of $\Delta\Delta G$ and ΔT_m values. Figures 5.3 and 5.4 show top eight proteins according to number of related experimental entries and number of entries for substitutions from and to each amino acid.

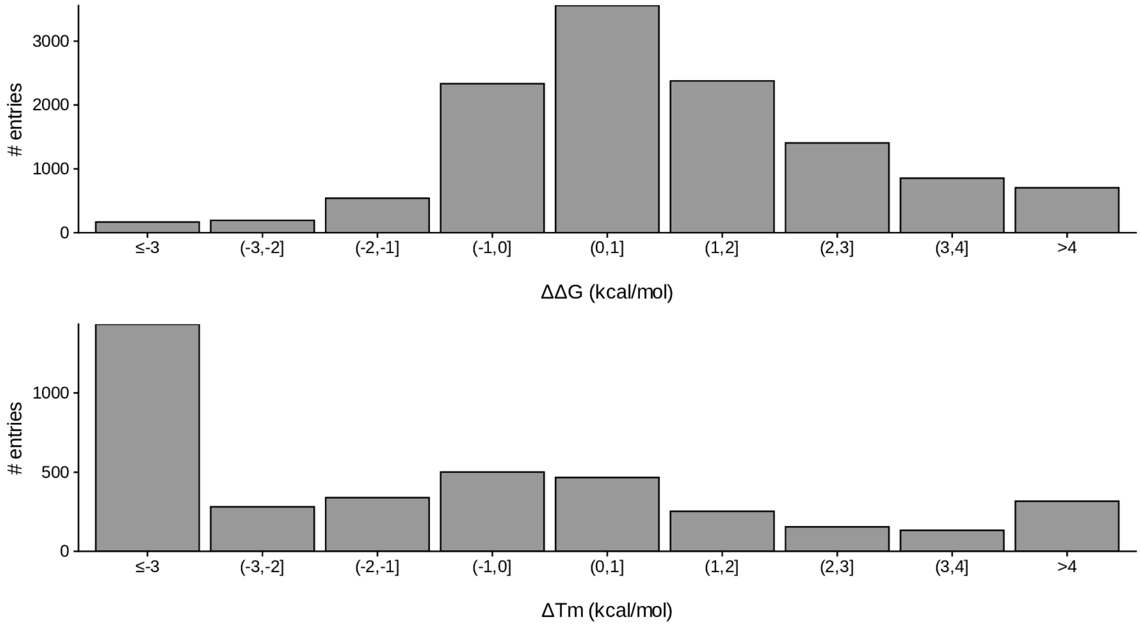


Figure 5.2: Histogram of $\Delta\Delta G$ and ΔT_m values.

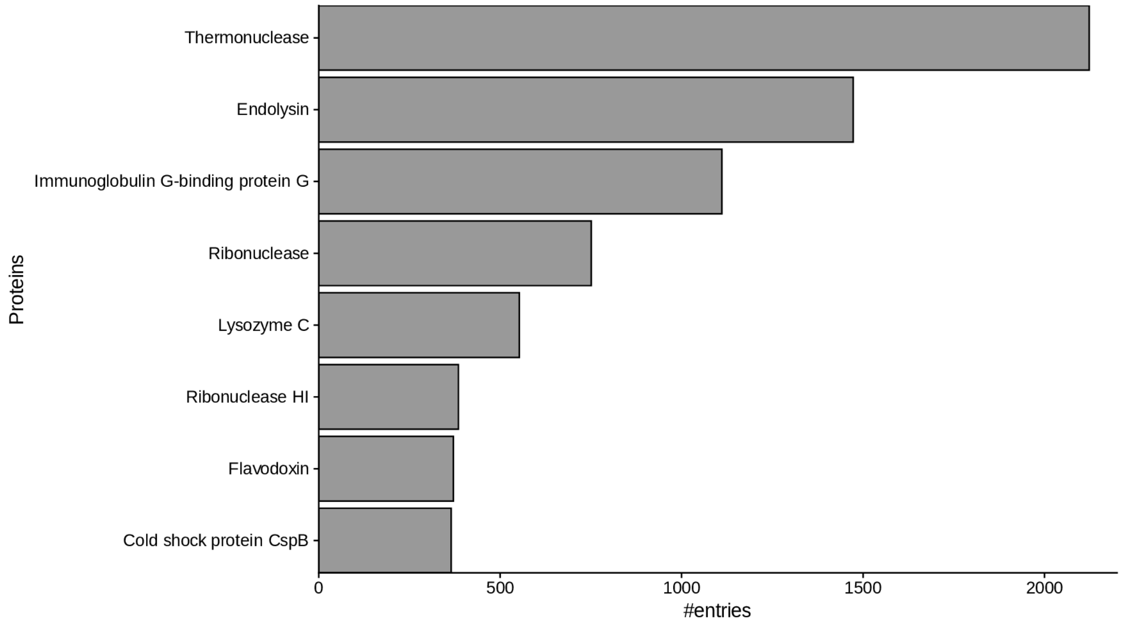


Figure 5.3: Top eight proteins according to number of belonging experiments.

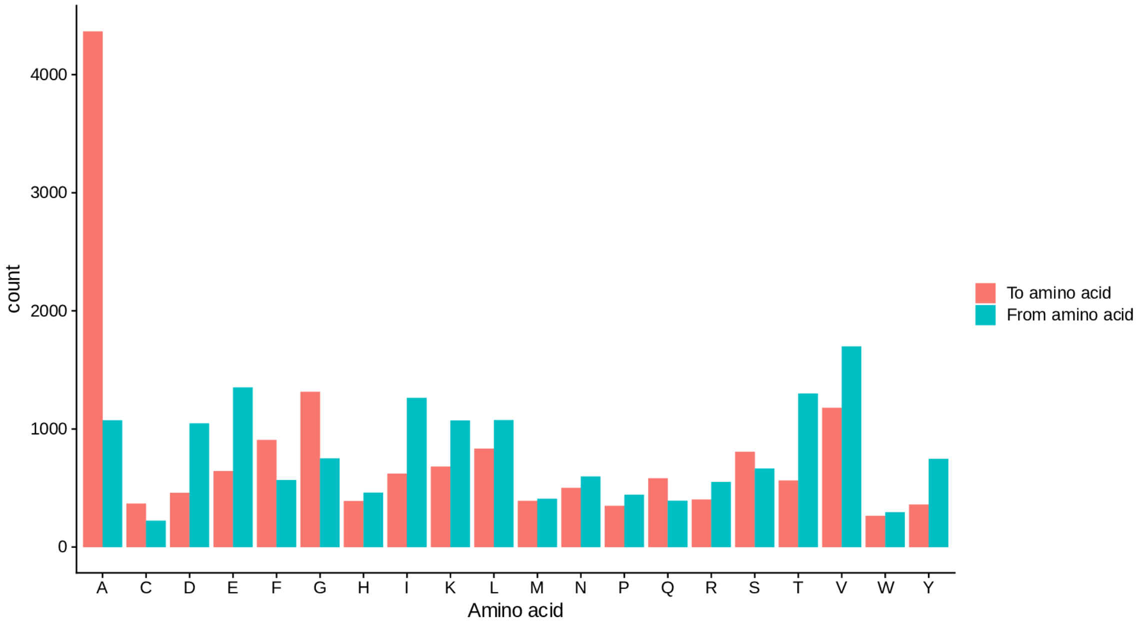


Figure 5.4: Number of experiments according to substitutions of each amino acid.

Chapter 6

FireProtDB implementation

FireProtDB consists of several layers: database, web service, web server and web application. Each of them has its specific function to ensure the functionality of FireProtDB. The database layer is implemented with the use of `MySQL` relational database to store all necessary biological data, and `Elasticsearch`[1] database is utilized to provide the ability to full-text search for certain data with the required speed. Web service is implemented in Java as HTTP application interface with the use of the `Spring Boot`[2] framework. Web server is implemented as a simple `Node.js`[3] server and serves as a bridge between web application and web service. Web application implements graphical user interface to provide all necessary information to the user and it is implemented as a single-page application in `TypeScript`[4] with the use of `React`[5] and `Redux`[6] library for state management. All parts of the application are provided as the standalone containers created and connected with the `Docker`[7] system.

6.1 Web service implementation

FireProtDB web service implements all required logic for the database communication. It also provides HTTP endpoints for serving the data according to the requests.

Communication with *MySQL* database is done via *Spring Data JPA* API which provides easier work with Java Persistence API (JPA). Communication with Elasticsearch database is done via *Spring Data Elasticsearch* API. Each table from the database has to be mapped to a certain entity for which the corresponding Spring Data repository has to exist. Repositories provide abilities to perform CRUD operations. Because of different API for Elasticsearch, different repositories for entities used for full-text search have to be implemented. The same entity then could have a JPA repository as well as an Elasticsearch repository if used for full-text searching.

[1] https://www.elastic.co/elasticsearch/

[2] https://spring.io/

[3] https://nodejs.org/en/

[4] https://www.typescriptlang.org/

[5] https://reactjs.org/

[6] https://react-redux.js.org/

[7] https://www.docker.com/

Search implementation

Searching is the most important feature of the FireProtDB web service. User requirements for FireProtDB specified in 5.1 define that the application needs to support full-text searching as well as advanced search with the ability to create customized search queries from pre-defined parameters.

There are several issues connected with the implementation of the search engine. Web service is designed as a RESTful API, therefore, an object representing a search query have to be designed to store information from the user. It has to support both search types, multiple presences of the same search parameters with different values, as well as nested queries. Because of these problems, two different objects were designed to overcome mentioned problems.

```
{
    "type":string,
    "key":string,
    "value":string,
    "additionalOptions": Array
}
```

Listing 6.1: Search expression object

```
{
    "type":string,
    "options": Array

}
```

Listing 6.2: Search operator object

Search expression object contains *type* field, which can only take on value `expr`. This specifies that an object represents an expression. *Key* field holds a value of the search parameter according to which user wants to search and *value* field specifies an exact value of a parameter. The last field *additionalOptions* is an array of certain options which are supported in query creation by some of the search parameters.

Search operator object has only two fields, *type* field specifies the type of logical operator used between objects in the options field. This field can obtain two different values: `AND` and `OR`. *Options* field in search operator object could be an array of search expression objects, operator objects, or a combination of both. The operator object is used when more than one search parameter is specified.

The final search request object cannot be represented only by one of the mentioned objects defining query. Results on the client side need to be sortable. However, to limit the amount of transferred data and provide enough speed, results are paginated. The user does not have all search data available, thus sorting needs to be done remotely. The request has to include information about the parameter used to sort the data and the order in which the sorting will be done. To overcome this issue, the final search request is defined as:

```
{
    "searchData": Object,
    "filter": Object
}
```

Listing 6.3: Search request object

Field *searchData* can be search expression object or search operator object. Filter field represents an object which has two fields: key defining name of parameter used for sorting and order in which the sorting is done.

The last issue connected with the search is what type of data should be returned as a result. There is a view from a mutational perspective and also a view on the individual experiments. To provide more information for a user, a view from the experimental level was chosen and results provide information about individual experiments. Creation of result entities could be slow as joining of several tables is required to obtain all of the provided information. To speed up the search and construction of the results, database contains pre-defined view *mutation_experiments_search*.

Type of search is defined in search endpoint parameter as well as page number and size that will be returned to the client.

Full-text search

Full-text search is implemented using Elasticsearch database, which provides fast search over indexed stored data. Data in the database are stored as JSON documents. To use it with the Spring Boot application, *Spring Data Elasticsearch* API is used to communicate with the database and perform search queries.

For this type of search, several entities were chosen for user to perform the search. They need to be stored individually in the Elasticsearch database as well as in the relational database. These entities include dataset, publication, structure, authors, protein and InterPro entries. Each entity has several attributes that were used to match the search query.

Input for the full-text search is a string query that can include several phrases delimited by white space. If the user chooses several phrases, they are split and each of them is individually used for search against each entity and its attributes.

Search is done in two phases. The first phase of search returns entities such as dataset or protein, but they cannot be used as a result. In the second phase, the corresponding experimental entities need to be found. As several phrases can lead to entities with the same experimental data, experiments need to be additionally filtered to include only unique entries.

Advanced search

Advanced search is implemented using *Spring Data JPA* API and special JPA Criteria API. *Spring Data JPA* provides usage simplification of JPA based repositories, so less code is needed. Advanced search has to support the creation of dynamic queries, which is done using Criteria API.

Model entity and corresponding repository is created for each table in the MySQL database to provide all CRUD operations with the database. Each repository has to extend *JpaSpecificationExecutor* to be able to support queries created by Criteria API.

The basis of advanced search is custom CriteriaBuilder class *SearchCriteriaBuilder*. This class contains CriteriaBuilder object, CriteriaQuery object and objects representing joins of specific tables represented by Java entities. Using CriteriaBuilder a query, which will return experiments objects, is created. The class contains methods that define and return objects representing joined entities. The most important is the *buildQuery* method, which creates query based on the provided Criteria API Predicate. The object of type Predicate represents part of an SQL query in the `WHERE` clause.

Final predicate, representing whole `WHERE` clause, required to build the query, is constructed dynamically from the search request. The user-defined search request is represented by a JSON object which is parsed and in this process, the predicate is being gradually built. As user has to be able to search according to several parameters of not only one entity, predicates for supported parameters are pre-defined.

Predicates are defined for only certain entities including mutation experiments, mutations, datasets, proteins and publications. Predicates can be easily created and defined for any entity to extend provided predicates. For supported parameters, an enumeration with their names that is in accordance with corresponding predicates is created. Each predicate takes a value of the searched parameter and optionally can take the value of additional options, which are supported only by certain parameters.

Most predicates are defined for mutation experiments entity. Experimental predicates include ones that check the presence of value or compare the value of certain parameter with the value specified by the user. If predicate compares the value of parameter with user-provided value, the string used as an input for predicate contains selected operator and value. They are then split and the predicate is created according to the specified operator. Predicates that check if the experiment is stabilizing, destabilizing or neutral also need to check the value of additional options. These options are that query is used on a mutational level, which means that if there is an experiment breaking the stability criteria, corresponding mutation and its experiments will be excluded from the results. The second option is that values of $\Delta\Delta G$ and ΔT_m parameters have to agree as the experiment can have values of these parameters indicating an opposite stability. For each type of additional option, a different query is created and if there is no additional option, another query is created.

Protein predicates include SQL LIKE type predicates for parameters such as name, protein sequence or species. They also include predicates for comparing the value of parameter with input value such as UniProt identifier. Another type of predicate is the presence of the protein in certain InterPro family. Input for such predicate is a string value containing operator `IN` or `NOT IN` and values of selected families delimited by a semicolon. For each operator, a corresponding query is defined.

Dataset predicates also include predicate with `IN` or `NOT IN` operator and values of selected datasets. Predicate then checks the presence of experiments in the selected datasets. Publication predicates include a check of the DOI number.

Each of the predicates is created based on the value of the *key* field in the search expression object. When the whole search request is parsed and the final predicate is created, the query is executed and the result is obtained as a list of mutation experiments.

Export implementation

Besides searching ability, very important ability of FireProtDB is the export of the data. Search results, as well as the whole database dump, can be exported. Existing stability databases do not provide an option for downloading search results, which is important for their users.

Search results can be exported as a CSV file and provide more detailed data. The database structure is very complex and obtaining all necessary information provided in a CSV file would require joining multiple tables.

To solve this problem and provide fast export, the underlying MySQL database contains a pre-defined view *mutation_experiments_summary* that combines all necessary tables.

FireProtDB also supports higher control over data with the ability to download a full dump of the database as an SQL file. The exported database dump also contains *mutation_experiments_summary* view definition for the users to use it as their starting point for additional filtering and construction of custom queries.

Dump of the database is implemented using Java library *mysql-backup4j*[8]. In the first step, the SQL dump file is created and then the file is compressed and sent to the client as a ZIP file.

6.2 Web application implementation

The web application is implemented with React library for the design of front-end applications. TypeScript, which is a superset of JavaScript with support of static types, is employed as a main language. For managing the state of the application, the Redux library was used.

The application is designed as a single page web application. With the use of React, the designer uses a declarative approach to declare the structure of the components, what they should display, and the state of the application. Data displayed by the components are based on the current state of the application. React manages the content of the components and if the state is changed, changes are projected to the appearance of the components. Changes are done dynamically without the need to communicate with the server which leads to the same URL. Managing the URL is done via `React Router`[9] library, which uses the history of the browser to manage the URL.

Application state management is handled by the Redux library. It supports the reusability of the components. The appearance of the components and provided logic are separated. Components are parameterized and they display data based on the parameters, which could be data or function references.

With the use of Redux, the application has one global state which is defined by the designer. Each component uses only part of the state. Redux require the definition of several abstraction parts: containers, reducers and components. *Containers* contain necessary logic and define which data from the state can be used by the certain component. They also provide functions capable of changing the state of the application for the components. *Reducers* are handling the events leading to the state changes and the state can be changed only by the reducers.

User interface

Web application implements a graphical user interface, which provides several views on the data in the database. The view is split into several pages: main page, protein page, mutation page, datasets page and search results page.

Main page provides the basic introduction to the FireProtDB and the basic overview of the data stored in the database. The overview is realized in the form of charts (fig. 6.1), that show statistics of the most important properties of the data. Chart with overall statistics of the database provides information about progress in amount of proteins, experiments or mutations through time. For users, the most important statistics related to stability data is data distribution according to values of $\Delta\Delta G$ and ΔT_m. Such overview is presented by individual histograms for both $\Delta\Delta G$ and ΔT_m. Histograms will provide very important initial information about underlying character of available experimental data. Statistics also show relationship between proteins and experimental data, top proteins chart provide top 8 proteins with the most experiments and the top InterPro families chart provide information about top 10 InterPro families with the most experiments. Users will get information

[8] https://github.com/SeunMatt/mysql-backup4j

[9] https://reactrouter.com/

that could help them to focus on specific proteins and families. The last charts are focused on providing the statistics of individual residues. Relationship between number of experiments that were performed to and from certain residue is provided as histogram with number of experiments for each single residue. Statistics that show relationship between individual residues are provided as a heat map, that contains number of substitutions for each pair of amino acid residues. Users have option to choose between the heat map with full amino acid alphabet or chart using reduced alphabet that was proposed in [8], where amino acids were clustered according to local structural features, which can help designers of prediction tools.

Besides providing useful initial overview, some of the charts provide advanced ability that allows users to search in the database only by clicking on selected data in the chart, e.g. if user is interested in exploration of experiments by value of $\Delta\Delta G$, then simple click on bar displaying number of experiments in specified range will perform search query in the database and show results with experimental data in the range. This ability is supported by charts displaying top proteins and InterPro families as well as histograms for $\Delta\Delta G$ and ΔT_m. This feature will be very helpful as a fast shortcut in initial data exploration, because users will not have to specify search query to obtain the data of interest.

Protein page (fig. 6.2 and fig. 6.3) provides data related to a certain protein. Provided data includes basic information about protein, sequential and structural features and also experimental data belonging to the protein.

Basic information provides the most important features to identify protein such as the name of the protein, species, UniProt identifier and EC number to link protein to the related databases for further analysis. Basic information is further enriched by the list of the InterPro families, to which the protein belongs.

Section with sequence features provides a more complex view of the features connected with the protein sequence. Data are displayed using the interactive ProtVista [39] component, which provides interactive data tracks. The component provides information about the secondary structure of certain parts of the protein, sites, a charge of amino acids, information about residue's presence in the catalytic pocket or the tunnel, information about b-factor, conservation and mutations. Every feature has its own track with the data. Data for secondary structure and sites are obtained via UniProt API[10] and data for the remaining features are constructed from the database data. Features are always related to the specific residues and they are displayed as graphical elements in the track at the position of the certain residue. Details about residue's data can be displayed in the tooltip.

Structural features section provides basic structural information about structure such as PDB identifier, the method and resolution with visualization of the structure. Because protein can have more related structures, user has the ability to choose currently displayed structure and its information. For visualization part, the third-party component PDBe Molstar[11] is used to display the three-dimensional structure of the protein based on the currently selected PDB identifier.

The main feature of visualization component is to provide ability to show user selected residues in the structure and analyze their properties. Selection of residues that are part of the structure is connected with issue that protein can have more determined structures, but positions and chain could not be valid for each of them. Database currently stores positions of structural residues only for one representative structure. If this representative structure is currently selected, user can select residues directly in the viewer or by using more advanced

[10] https://www.uniprot.org/help/api

[11] https://github.com/PDBeurope/pdbe-molstar

option. The advanced option is represented by the slider that provides the ability to select residues in the structure according to features including $\Delta\Delta G$, ΔT_m, conservation and b-factor values. Slider allows user to specify the range of values of selected feature and residues with values within the interval are then highlighted. This feature focuses on easy and fast analysis, users can explore location of such residues in very quickly. Besides selection and visualization abilities, slider also provides more advanced feature that is connected with search ability. Currently selected range of values for selected option could be used as a search query. It means that for currently viewed protein and selected range of values, search query will be created and performed. This features will be helpful for fast and interactive selecting of interesting experiments of certain protein.

The last section contains a table with basic information about experimental data for the current protein such as its curation, $\Delta\Delta G$ and ΔT_m values or publication. The position of each mutation can be also selected in the structure viewer to analyze it. In addition, mutated positions can be displayed also from ProtVista track containing mutational data.

Mutation page (fig. 6.4) provides information related to certain mutation and its experimental data. Basic information about mutation includes original and mutated residues, the position of the mutation and the protein where the mutation occurred. The most important are the experimental data determining the character of the mutation. Data are provided in the form of a table, where each row represents values of one experiment such as $\Delta\Delta G$ and ΔT_m, curation information, related datasets and publication or identifier of the experiment. Because some researchers use to perform several experiments and then aggregate obtained results, protein page contains table with derived values. Table include values of mean and standard deviation for related experiments.

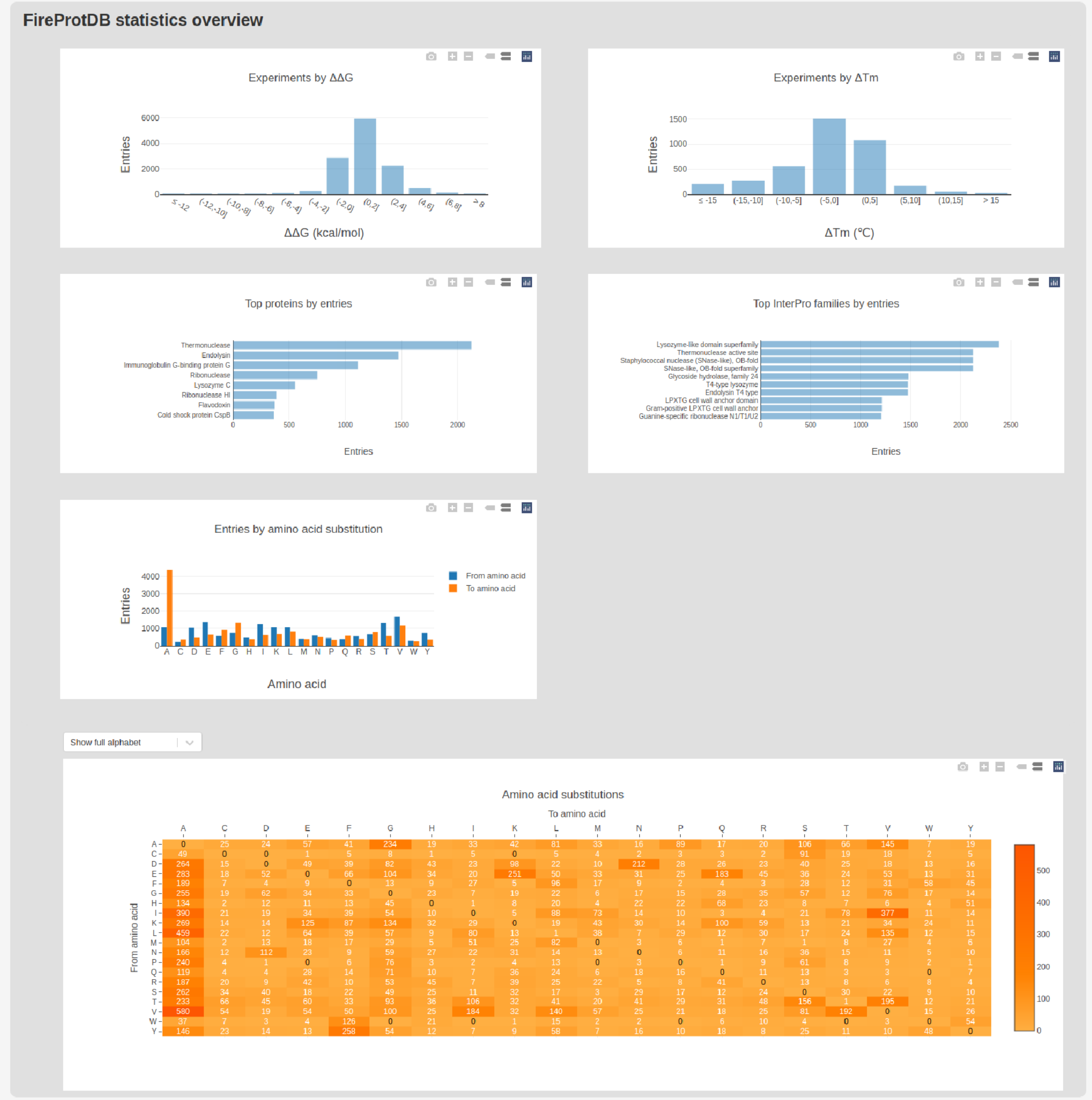
Dataset page (fig. 6.5) provides a view of the datasets in the database. It includes basic information about the dataset such as its name, version, related publication and its authors. Additionally, the page contains two charts with statistics. The first displays the stability character of the experimental data in the dataset, while the second one shows a comparison of the selected dataset in how many entries is overlapping with other datasets.

Search results page (fig. 6.6) provides results based on the search query. Results are shown as a table where each row represents information about one experiment. By default, the table displays information such as the name of the related protein, information about curation, mutation, $\Delta\Delta G$ and ΔT_m values and uses colours to determine the effect of the mutation, so the user can easily navigate in the table. Besides the previous features, the user can dynamically display additional information such as pH or protein concentration with a manual selection of the currently displayed features. From the results table, the user can easily navigate to the protein page or mutational page for selected protein or mutation. Very important feature that is missing in current stability databases, is ability to download search results. Here, user has the ability to export the results in machine readable CSV file. The exported file contains more specific information about experiments in the result table. Therefore, the search request that was used to obtain the current results, is sent to the server. The request is used to get the same experiments as provided in the table, but exported data are additionally enriched with further information related to the experiments, which are not provided in the table.

Besides providing search results in the table form, results page provides advanced ability to display charts with statistics related to the current results. Statistics can be fetched on request and provide similar statistics as charts on the main page. Charts provide quick overview of resulting data, user then can easily explore data distribution information and make further decisions what to look into in the next step of the analysis.

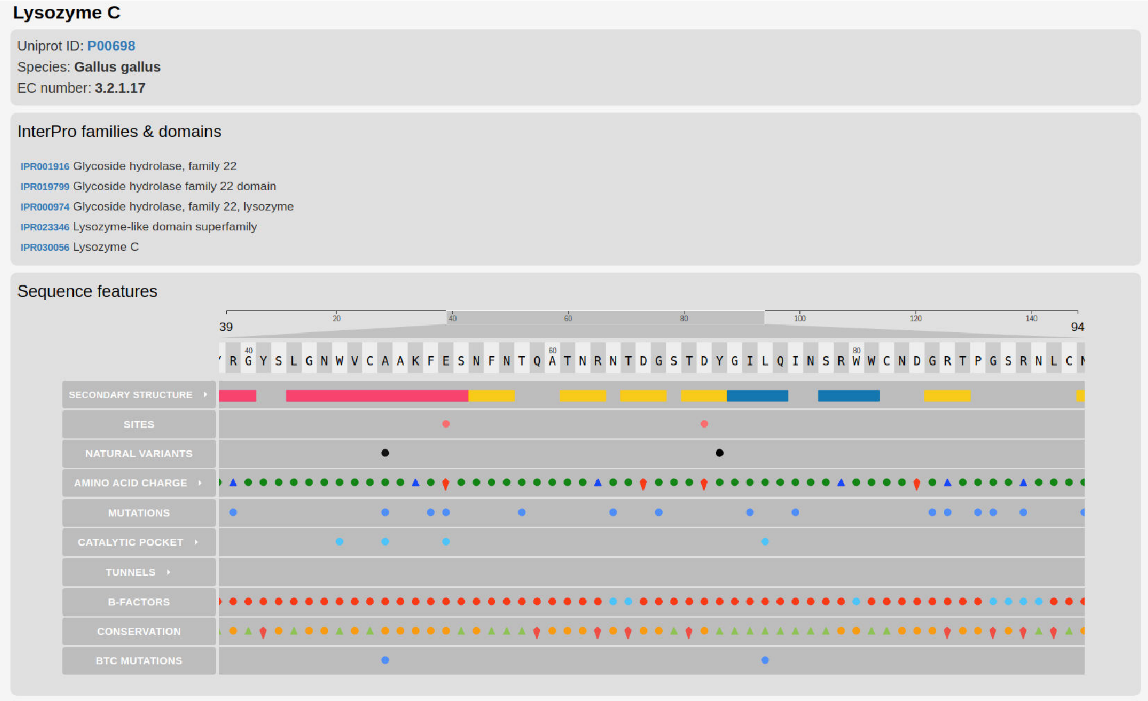
The most important part of the web application is the search component. Search is available on the top of the page and can be always accessed no matter what the current page is. FireProtDB search provides the ability for full-text searching and also the ability to create custom search query with advanced search. Phrases from full-text search or custom query are used to construct a part in the SQL query in `WHERE` clause. By default, the search component provides a text field to input phrases for full-text searching. Advanced search component is displayed after manual selection.

Advanced search component (fig. 6.7) provides the ability to add or remove search items dynamically and create a query from them. Users can create query only by connecting search items with logical connectives `AND` and `OR` and they can also use the support of brackets. With brackets usage, users can create complex predicates and connect them to construct a final query. Each search item represents one of the pre-defined parameters which can be searched in the database. According to the selected parameter, different field to provide target values of the parameter is provided. Certain parameters such as dataset name, organism or InterPro families support the selection of names, where the value of the parameter should be included or excluded, respectively. Several numerical parameters support the definition of operator used to define criteria for the value. Another parameter is either boolean or supports only the value definition of the specified parameter. After the query is defined, control of value types or missing values is done and an array of search items is parsed. From the array, the corresponding postfix notation is created and finally, the search object is created and send to the server. An important issue related to the search is storing the state of the search in the URL that is displayed on the results page. The URL contains the type of the search represented by the string and the search state is encoded using *base64* encoding. Search state can be then easily restored from URL and user can also share URL with an encoded search query that was used to obtain specific results.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	25	24	57	41	234	19	33	42	81	33	16	89	17	20	106	66	145	7	19
C	49	0	0	1	5	8	1	5	0	5	4	2	3	3	2	91	19	18	2	5
D	264	15	0	49	39	82	43	23	98	22	10	212	28	26	23	40	25	18	13	16
E	283	18	52	0	66	104	34	20	251	50	33	31	25	183	45	36	24	53	13	31
F	189	7	4	9	0	13	9	27	5	96	17	9	2	4	3	28	12	31	58	45
G	255	19	62	34	33	0	23	7	19	22	6	17	15	28	35	57	12	76	17	14
H	134	2	12	11	13	45	0	1	8	20	4	22	22	68	23	8	7	6	4	51
I	390	21	19	34	39	54	10	0	5	88	73	14	10	3	4	21	78	377	11	14
K	269	14	14	125	87	134	32	29	0	19	43	30	14	100	59	13	21	34	24	11
L	459	22	12	64	39	57	9	80	13	1	38	7	29	12	30	17	24	135	12	15
M	104	2	13	18	17	29	5	51	25	82	0	3	6	1	7	1	8	27	4	6
N	166	12	112	23	9	59	27	22	31	14	13	0	6	11	16	36	15	11	5	10
P	240	4	1	0	6	76	3	2	4	13	0	3	0	1	9	61	8	9	2	1
Q	119	4	4	28	14	71	10	7	36	24	6	18	16	0	11	13	3	3	0	7
R	187	20	9	42	10	53	45	7	39	25	22	5	8	41	0	13	8	6	8	4
S	262	34	40	18	22	49	25	11	32	17	3	29	17	12	24	0	30	22	9	10
T	233	66	45	60	33	93	36	106	32	41	20	41	29	31	48	156	1	195	12	21
V	580	54	19	54	50	100	25	184	32	140	57	25	21	18	25	81	192	0	15	26
W	37	7	3	4	126	0	21	0	1	15	2	2	0	6	10	4	0	3	0	54
Y	146	23	14	13	258	54	12	7	9	58	7	16	10	18	8	25	11	10	48	0

Figure 6.1: FireProtDB statistics on main page providing basic overview of data stored in the database.



Lysozyme C

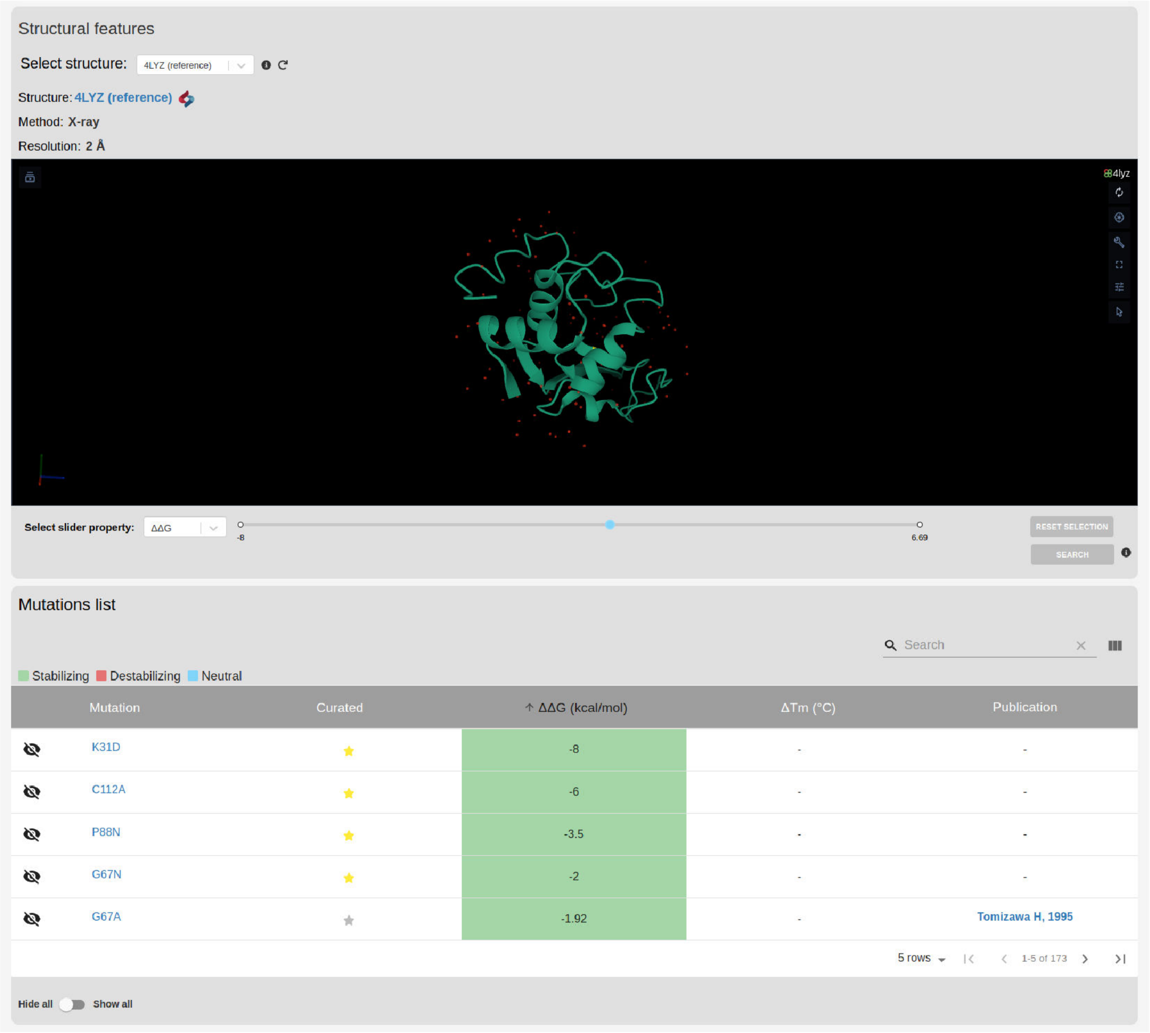
Uniprot ID: **P00698**
Species: **Gallus gallus**
EC number: **3.2.1.17**

InterPro families & domains

IPR001916 Glycoside hydrolase, family 22
IPR019799 Glycoside hydrolase family 22 domain
IPR000974 Glycoside hydrolase, family 22, lysozyme
IPR023346 Lysozyme-like domain superfamily
IPR030056 Lysozyme C

Sequence features

Figure 6.2: Basic information about protein with list of related InterPro families and sequential information displayed as a interactive ProtVista tracks with the data.



Structural features

Select structure: 4LYZ (reference)

Structure: 4LYZ (reference)

Method: X-ray

Resolution: 2 Å

Select slider property: ΔΔG

Mutations list

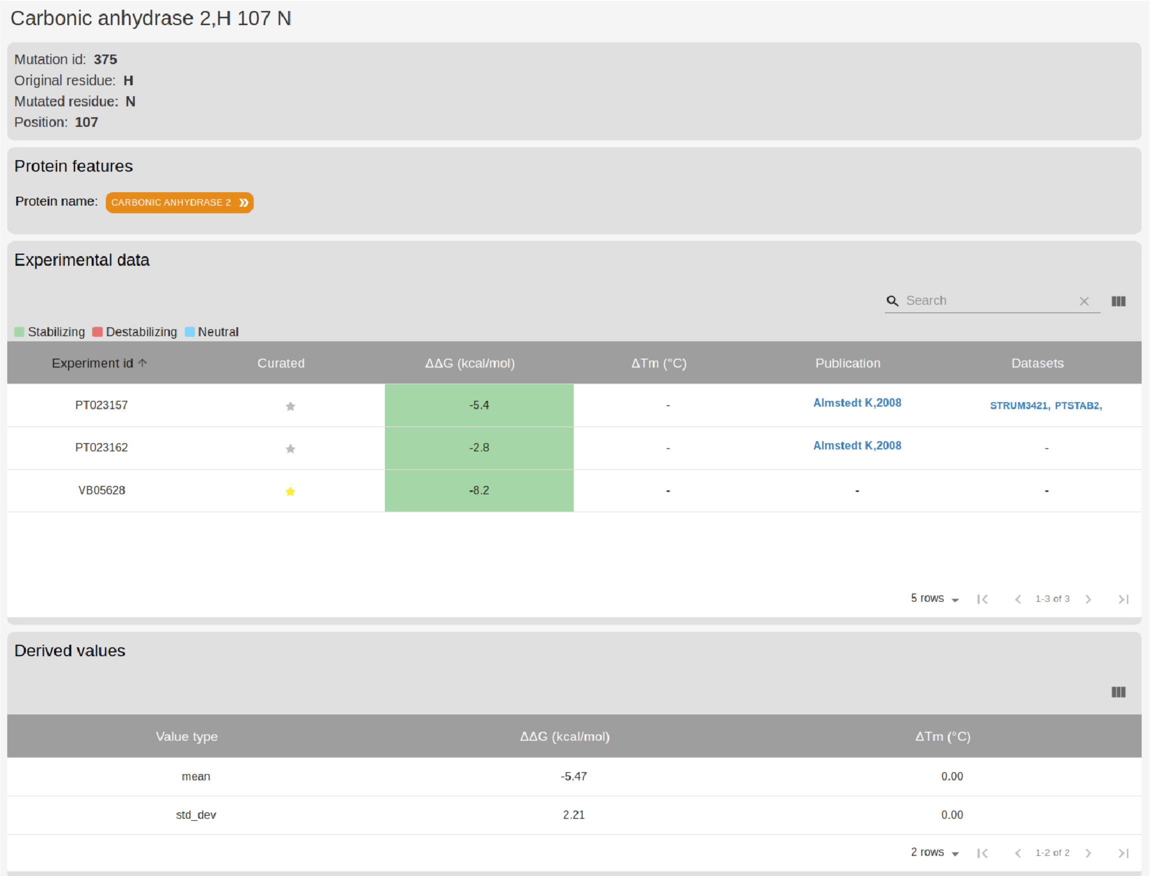
🟩 Stabilizing 🟥 Destabilizing 🟦 Neutral

	Mutation	Curated	↑ ΔΔG (kcal/mol)	ΔTm (°C)	Publication
	K31D	★	-8	-	-
	C112A	★	-6	-	-
	P88N	★	-3.5	-	-
	G67N	★	-2	-	-
	G67A	☆	-1.92	-	Tomizawa H, 1995

5 rows 1-5 of 173

Hide all Show all

Figure 6.3: Structural and mutational features of protein displaying basic information of selected structure with Molstar visualization component that allows user to select and view positions of interest in the structure and table with mutated positions and their properties.



Carbonic anhydrase 2,H 107 N

Mutation id: **375**
Original residue: **H**
Mutated residue: **N**
Position: **107**

Protein features

Protein name: CARBONIC ANHYDRASE 2 »

Experimental data

Search

🟩 Stabilizing 🟥 Destabilizing 🟦 Neutral

Experiment id ↑	Curated	ΔΔG (kcal/mol)	ΔTm (°C)	Publication	Datasets
PT023157	★	-5.4	-	Almstedt K,2008	STRUM3421, PTSTAB2,
PT023162	★	-2.8	-	Almstedt K,2008	-
VB05628	★	-8.2	-	-	-

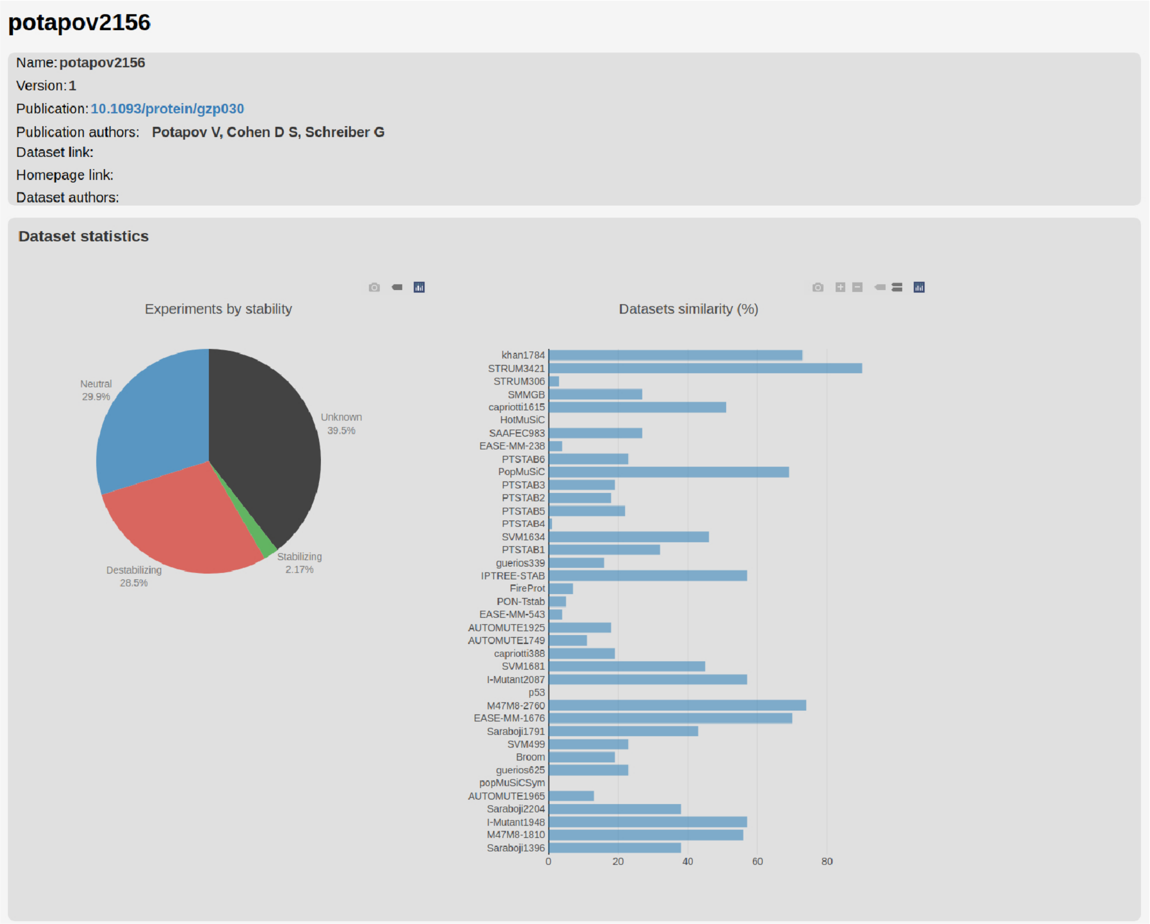
5 rows 1-3 of 3

Derived values

Value type	ΔΔG (kcal/mol)	ΔTm (°C)
mean	-5.47	0.00
std_dev	2.21	0.00

2 rows 1-2 of 2

Figure 6.4: Mutation page with basic information about mutation and list of all related experimental data.



potapov2156

Name: potapov2156

Version: 1

Publication: 10.1093/protein/gzp030

Publication authors: Potapov V, Cohen D S, Schreiber G

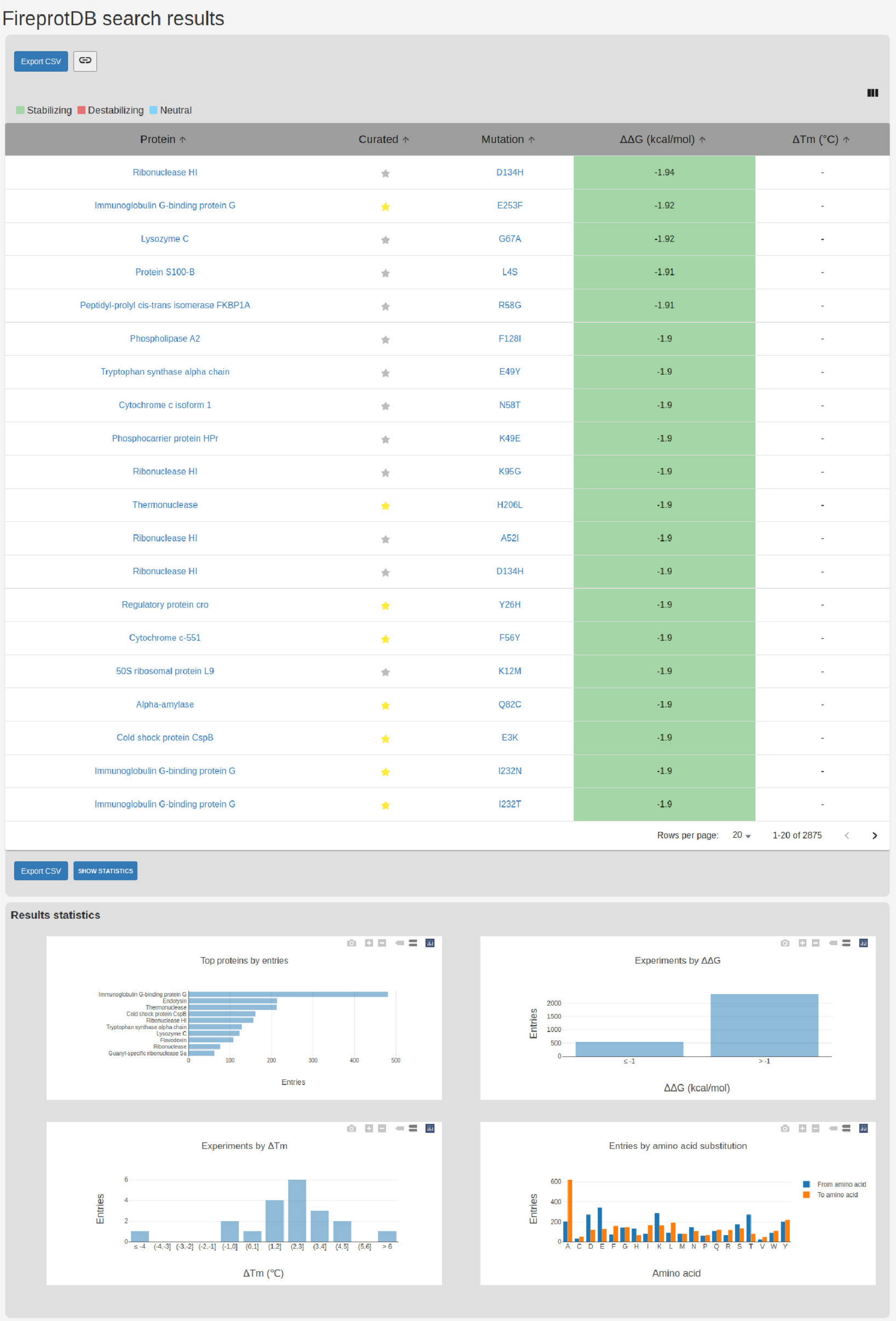
Dataset link:

Homepage link:

Dataset authors:

Dataset statistics

Figure 6.5: Dataset page provides basic information about dataset with additional statistics providing information about mutation types in the dataset and overlapping with another stored datasets.



FireprotDB search results

Export CSV

🟩 Stabilizing 🟥 Destabilizing 🟦 Neutral

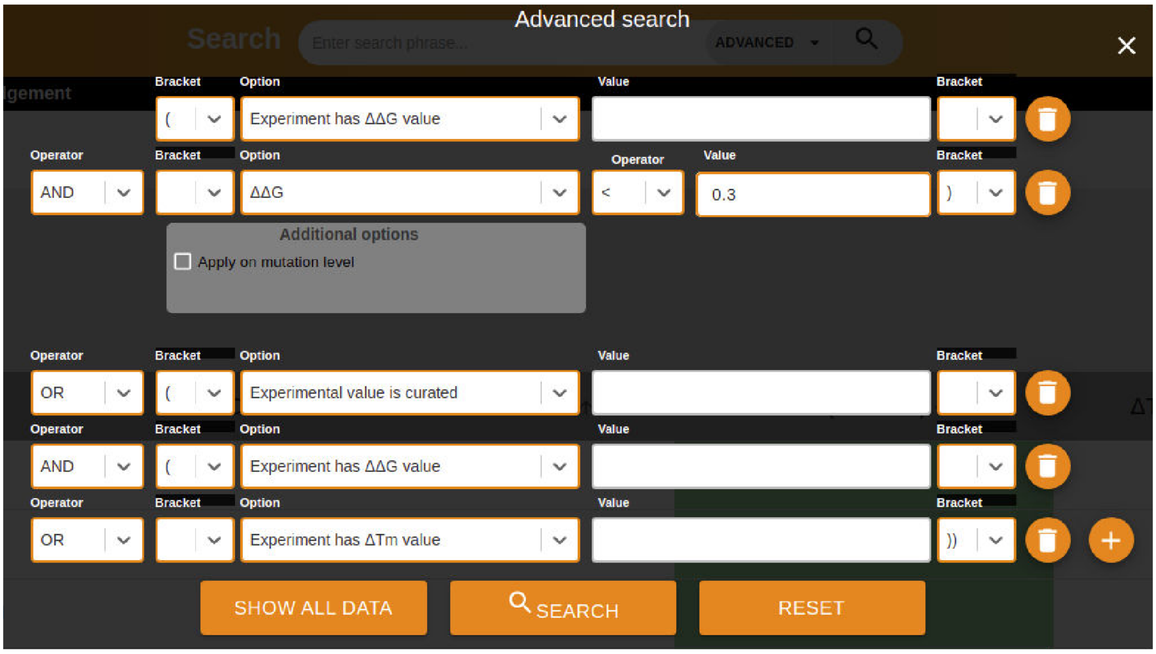
Protein ↑	Curated ↑	Mutation ↑	ΔΔG (kcal/mol) ↑	ΔTm (°C) ↑
Ribonuclease HI	☆	D134H	-1.94	-
Immunoglobulin G-binding protein G	★	E253F	-1.92	-
Lysozyme C	☆	G67A	-1.92	-
Protein S100-B	☆	L4S	-1.91	-
Peptidyl-prolyl cis-trans isomerase FKBP1A	☆	R58G	-1.91	-
Phospholipase A2	☆	F128I	-1.9	-
Tryptophan synthase alpha chain	☆	E49Y	-1.9	-
Cytochrome c isoform 1	☆	N58T	-1.9	-
Phosphocarrier protein HPr	☆	K49E	-1.9	-
Ribonuclease HI	☆	K95G	-1.9	-
Thermonuclease	★	H206L	-1.9	-
Ribonuclease HI	☆	A52I	-1.9	-
Ribonuclease HI	☆	D134H	-1.9	-
Regulatory protein cro	★	Y26H	-1.9	-
Cytochrome c-551	★	F56Y	-1.9	-
50S ribosomal protein L9	☆	K12M	-1.9	-
Alpha-amylase	★	Q82C	-1.9	-
Cold shock protein CspB	★	E3K	-1.9	-
Immunoglobulin G-binding protein G	★	I232N	-1.9	-
Immunoglobulin G-binding protein G	★	I232T	-1.9	-

Rows per page: 20 ▾ 1-20 of 2875 ‹ ›

Export CSV SHOW STATISTICS

Results statistics

Figure 6.6: Search results page with table containing experimental entries together with ability to display additional statistics for current results.



Operator	Bracket	Option	Operator	Value	Bracket
	(	Experiment has ΔΔG value			
AND		ΔΔG	<	0.3	)
OR	(	Experimental value is curated			
AND	(	Experiment has ΔΔG value			
OR		Experiment has ΔTm value			))

Additional options

□ Apply on mutation level

Figure 6.7: Advanced search provides users ability to dynamically create complex queries using brackets and logical operators from pre-defined options about experiments, dataset, protein and others.

Chapter 7

Conclusion

Nowadays, proteins find their use in a broad spectrum of applications in various fields of industry and medicine, where they often need to withstand harsh environmental conditions. To overcome such an issue, protein stability needs to be improved to make the usage of proteins possible. Design of more stable proteins is done by mutations of certain residues, but to obtain required results, large amount of mutations have to be designed. The determination of stability is usually done in laboratory using one of the commonly used methods, which is costly and time-consuming.

In the recent years, machine learning methods have been utilized in the field of protein engineering, which also applies on the problem of determining the protein's stability. Usage of machine learning methods could help with faster pre-selection of promising mutations that would be further studied in the laboratory. Crucial issue related to the usage of machine learning methods is the requirement of the high amount of reliable data. Without the data, machine learning tools cannot be trained to provide predictions with high accuracy.

Current situation with experimental stability data is not optimal. Experimental data are split among three major data sources: ProTherm, ProtaBank and VariBench and a large portion of data still remains only in the scientific literature without being stored in any database. The largest well-known source of experimental data, ProTherm, suffers from the presence of many inconsistencies and errors which were reported by the users and it has not been updated since 2013. The other two databases, ProtaBank and VariBench, are not primarily focused on the collection and maintenance of the stability data. Issues related to the data that are currently available have significant influence on the usage of the machine learning methods. The most important is the amount of the data, which did not increase in the sufficient way, and without new experimental data, tools cannot perform better. Other important issues are that users usually need to perform several steps of filtering and validation to obtain data with necessary quality. Current data sources also do not support advanced abilities for search and data export, which would significantly decrease the time required for data acquisition.

The main goal of this thesis was to design and implement a novel database of experimental thermostability data for single-point mutants, which would provide reliable data from the current sources and could possibly become a standardized source of the experimental protein stability data. The result of this thesis is FireProtDB database that was created in cooperation with experts from Loschmidt Laboratories. The database focuses on researchers that need to explore experimental data and also on users seeking the data for development of the machine learning tools. FireProtDB provides experimental data collected from several existing sources that were filtrated and validated. Data were also enriched by additional sequential and structural annotations that will be especially useful as features for machine learning models. The database provides a web interface that allows users to explore provided data on several levels with an advanced display of sequential features or three-dimensional structure view with the ability to show residues according to specific properties. An important feature of the FireProtDB is the implementation of the search engine that focuses on flexible selection of the data of interest. Database allows users to create complex custom search queries which will be useful for developers constructing training and testing datasets for their predictive tools. For easy processing, FireProtDB provides search results in machine readable CSV format. For more advanced users there is an ability to obtain a dump of the database and perform additional changes or perform more specific queries. The web interface of the database can be publicly accessed at `https://loschmidt.chemi.muni.cz/fireprotdb/` and more than 2,000 users have already visited and used the database. The database was already published in the scientific journal Nucleic Acids Research (impact factor 11.501). The journal article of the database is attached in appendix B. FireProtDB will be maintained by the team in the Loschmidt Laboratories to be an important and reliable source of protein stability data for scientific community.

Bibliography

[1] *Circular Dichroism* [online]. 2020 [cit. 2021-04-16]. Available at: https://chem.libretexts.org/@go/page/1761.

[2] Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M. et al. *Molecular Biology of the Cell., 6th Edition.* Garland Science, 2014. ISBN 978-0-815-34432-2.

[3] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research.* september 1997, vol. 25, no. 17, p. 3389–3402. DOI: 10.1093/nar/25.17.3389. Available at: https://doi.org/10.1093/nar/25.17.3389.

[4] Babkova, P., Sebestova, E., Brezovsky, J., Chaloupkova, R. and Damborský, J. Ancestral Haloalkane Dehalogenases Show Robustness and Unique Substrate Specificity. *ChemBioChem.* june 2017, vol. 18, no. 14, p. 1448. DOI: 10.1002/cbic.201700335. Available at: https://www.doi.org/10.1002/cbic.201700335.

[5] Bava, K., Gromiha, M., Uedaira, H., Kitajima, K. and Sarai, A. ProTherm, version 4.0: Thermodynamic Database for Proteins and Mutants. *Nucleic acids research.* february 2004, vol. 32, p. D120–1. DOI: 10.1093/nar/gkh082.

[6] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research.* january 2003, vol. 31, no. 1, p. 365–370. DOI: 10.1093/nar/gkg095. Available at: https://doi.org/10.1093/nar/gkg095.

[7] Broom, A., Jacobi, Z., Trainor, K. and Meiering, E. Computational tools help improve protein stability but with a solubility tradeoff. *Journal of Biological Chemistry.* july 2017, vol. 292, no. 35, p. 14349–14361. DOI: 10.1074/jbc.M117.784165. Available at: https://doi.org/10.1074/jbc.M117.784165.

[8] Caldararu, O., Mehra, R., Blundell, T. and Kepp, K. Systematic Investigation of the Data Set Dependency of Protein Stability Predictors. *Journal of Chemical Information and Modeling.* august 2020. DOI: 10.1021/acs.jcim.0c00591. Available at: https://www.doi.org/10.1021/acs.jcim.0c00591.

[9] Capriotti, E., Fariselli, P. and Casadio, R. A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics.* august 2004, vol. 20, p. 63–68. DOI: 10.1093/bioinformatics/bth928. Available at: https://doi.org/10.1093/bioinformatics/bth928.

[10] Capriotti, E., Fariselli, P. and Rossi, I. A three-state prediction of single point mutations on protein stability changes. *BMC bioinformatics.* february 2008, 9 Suppl

2, p. S6. DOI: 10.1186/1471-2105-9-S2-S6. Available at: https://www.doi.org/10.1186/1471-2105-9-S2-S6.

[11] Consortium, T. U. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research.* november 2020, vol. 49, D1, p. D480–D489. DOI: 10.1093/nar/gkaa1100. Available at: https://doi.org/10.1093/nar/gkaa1100.

[12] contributors, W. *Absorption spectroscopy — Wikipedia, The Free Encyclopedia* [online]. 2021 [cit. 2021-04-16]. Available at: https://en.wikipedia.org/w/index.php?title=Absorption_spectroscopy&oldid=1015615625.

[13] contributors, W. *Differential scanning calorimetry — Wikipedia, The Free Encyclopedia* [online]. 2021 [cit. 2021-04-16]. Available at: https://en.wikipedia.org/w/index.php?title=Differential_scanning_calorimetry&oldid=1011455611.

[14] Dehouck, Y., Grosfils, A., Folch, B., Gilis, D., Bogaerts, P. et al. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics.* august 2009, vol. 25, no. 19, p. 2537–2543. DOI: 10.1093/bioinformatics/btp445. Available at: https://www.doi.org/10.1093/bioinformatics/btp445.

[15] Folkman, L., Stantic, B. and Sattar, A. Feature-based multiple models improve classification of mutation-induced stability changes. *BMC genomics.* may 2014, 15 Suppl 4, S6. DOI: 10.1186/1471-2164-15-S4-S6. Available at: https://www.doi.org/10.1186/1471-2164-15-S4-S6.

[16] George, D. G., Dodson, R. J., Garavelli, J. S., Haft, D. H., Hunt, L. T. et al. The Protein Information Resource (PIR) and the PIR-International Protein Sequence Database. *Nucleic Acids Research.* january 1997, vol. 25, no. 1, p. 24–27. DOI: 10.1093/nar/25.1.24. Available at: https://doi.org/10.1093/nar/25.1.24.

[17] Getov, I., Petukh, M. and Alexov, E. SAAFEC: Predicting the Effect of Single Point Mutations on Protein Folding Free Energy Using a Knowledge-Modified MM/PBSA Approach. *International Journal of Molecular Sciences.* april 2016, vol. 17, no. 4, p. 512. DOI: 10.3390/ijms17040512. Available at: https://www.doi.org/10.3390/ijms17040512.

[18] Gromiha, M. M. *Protein Bioinformatics: From sequence to function.* Elsevier, 2010. ISBN 978-81-312-2297-3.

[19] Guerois, R., Nielsen, J. E. and Serrano, L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. *Journal of Molecular Biology.* 2002, vol. 320, no. 2, p. 369–387. DOI: 10.1016/S0022-2836(02)00442-4. Available at: https://www.sciencedirect.com/science/article/pii/S0022283602004424.

[20] Huang, L.-T., Gromiha, M. and Ho, S.-Y. IPTREE-STAB: Interpretable Decision Tree Based Method for Predicting Protein Stability Changes upon Mutations. *Bioinformatics.* march 2007, vol. 23, no. 10, p. 1292–1293. DOI: 10.1093/bioinformatics/btm100. Available at: https://www.doi.org/10.1093/bioinformatics/btm100.

[21] Kellogg, E., Leaver Fay, A. and Baker, D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins*. march 2011, vol. 79, p. 830–8. DOI: 10.1002/prot.22921. Available at: https://www.doi.org/10.1002/prot.22921.

[22] Khan, S. and Vihinen, M. Performance of protein stability predictors. *Human mutation*. june 2010, vol. 31, no. 6, p. 675—684. DOI: 10.1002/humu.21242. Available at: https://doi.org/10.1002/humu.21242.

[23] Kortemme, T. and Baker, D. A simple physical model for binding energy hot spots in protein–protein complexes. *Proceedings of the National Academy of Sciences*. 2002, vol. 99, no. 22, p. 14116–14121. DOI: 10.1073/pnas.202485799. Available at: https://www.doi.org/10.1073/pnas.202485799.

[24] Kortemme, T., Kim, D. E. and Baker, D. Computational Alanine Scanning of Protein-Protein Interfaces. *Science Signaling*. 2004, vol. 2004, no. 219, p. pl2–pl2. DOI: 10.1126/stke.2192004pl2. Available at: https://stke.sciencemag.org/content/2004/219/pl2.

[25] Masso, M. and Vaisman, I. AUTO-MUTE 2.0: A Portable Framework with Enhanced Capabilities for Predicting Protein Functional Consequences upon Mutation. *Advances in bioinformatics*. august 2014, vol. 2014, p. 2783–85. DOI: 10.1155/2014/278385. Available at: https://www.doi.org/10.1155/2014/278385.

[26] Needleman, S. B. and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*. july 1970, vol. 48, no. 3, p. 443–453. DOI: 10.1016/0022-2836(70)90057-4. Available at: https://www.sciencedirect.com/science/article/pii/0022283670900574.

[27] Pires, D. E. V., Ascher, D. B. and Blundell, T. L. MCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*. november 2013, vol. 30, no. 3, p. 335–342. DOI: 10.1093/bioinformatics/btt691. Available at: https://doi.org/10.1093/bioinformatics/btt691.

[28] Ponnunswamy, P. and Gromiha, M. On the Conformational Stability of Folded Proteins. *Journal of Theoretical Biology*. 1994, vol. 166, no. 1, p. 63–74. DOI: https://doi.org/10.1006/jtbi.1994.1005. Available at: https://www.sciencedirect.com/science/article/pii/S0022519384710058.

[29] Potapov, V., Cohen, M. and Schreiber, G. Assessing computational methods for predicting protein stability upon mutation: Good on average but not in the details. *Protein engineering, design and selection : PEDS*. july 2009, vol. 22, p. 553–60. DOI: 10.1093/protein/gzp030. Available at: https://doi.org/10.1093/protein/gzp030.

[30] Pucci, F., Bernaerts, K. V., Kwasigroch, J. M. and Rooman, M. Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics*. april 2018, vol. 34, no. 21, p. 3659–3665. DOI: 10.1093/bioinformatics/bty348. Available at: https://doi.org/10.1093/bioinformatics/bty348.

[31] Pucci, F., Bourgeas, R. and Rooman, M. Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC.

Scientific Reports. march 2016, vol. 6. DOI: 10.1038/srep23257. Available at: https://www.doi.org/10.1038/srep23257.

[32] Quan, L., Lv, Q. and Zhang, Y. STRUM: Structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics*. june 2016, vol. 32, no. 19, p. 2936–2946. DOI: 10.1093/bioinformatics/btw361. Available at: https://www.doi.org/10.1093/bioinformatics/btw361.

[33] Saraboji, K., Gromiha, M. and Ponnuswamy, M. N. Average assignment method for predicting the stability of protein mutants. *Biopolymers*. 2006, vol. 82.

[34] Sasidharan Nair, P. and Vihinen, M. *Human mutation*. january 2013, vol. 34, no. 1, p. 42—49. DOI: 10.1002/humu.22204. Available at: https://doi.org/10.1002/humu.22204.

[35] Sumbalova, L., Stourac, J., Martinek, T., Bednar, D. and Damborský, J. HotSpot Wizard 3.0: Web server for automated design of mutations and smart libraries based on sequence input information. *Nucleic Acids Research*. july 2018, vol. 46, W1, p. W356–W362. DOI: 10.1093/nar/gky417. Available at: https://www.doi.org/10.1093/nar/gky417.

[36] Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J. et al. Protein Data Bank (PDB): Database of Three-Dimensional Structural Information of Biological Macromolecules. *Acta Crystallographica Section D*. november 1998, vol. 54, 6 Part 1, p. 1078–1084. DOI: 10.1107/S0907444998009378. Available at: https://doi.org/10.1107/S0907444998009378.

[37] Voet, D., Voet, J. G. and Pratt, C. W. *Fundamentals of Biochemistry: Life at the Molecular Level, 3rd Edition*. John Wiley and Sons, Inc., 2008. ISBN 0470129301.

[38] Wang, C. Y., Chang, P. M., Ary, M. L., Allen, B. D., Chica, R. A. et al. ProtaBank: A repository for protein design and engineering data. *BioRxiv*. 2018. DOI: 10.1101/272211. Available at: https://www.biorxiv.org/content/early/2018/02/26/272211.

[39] Watkins, X., Garcia, L. J., Pundir, S., Martin, M. J. and Consortium, U. ProtVista: visualization of protein sequence annotations. *Bioinformatics*. march 2017, vol. 33, no. 13, p. 2040–2041. DOI: 10.1093/bioinformatics/btx120. Available at: https://doi.org/10.1093/bioinformatics/btx120.

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

[41] Witvliet, D., Strokach, A., Giraldo Forero, A. F., Teyra, J., Colak, R. et al. ELASPIC web-server: Proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity. *Bioinformatics*. january 2016, vol. 32, no. 10, p. 1589–1591. DOI: 10.1093/bioinformatics/btw031. Available at: https://www.doi.org/10.1093/bioinformatics/btw031.

[42] Yang, Y., Chen, B., Tan, G., Vihinen, M. and Shen, B. Structure-based prediction of the effects of a missense variant on protein stability. *Amino acids*. october 2012, vol. 44, p. 847–855. DOI: 10.1007/s00726-012-1407-7. Available at: https://www.doi.org/10.1007/s00726-012-1407-7.

[43] Yang, Y., Urolagin, S., Niroula, A., Ding, X., Shen, B. et al. PON-tstab: Protein Variant Stability Predictor. Importance of Training Data Quality. *International Journal of Molecular Sciences*. march 2018, vol. 19, no. 4, p. 1009. DOI: 10.3390/ijms19041009. Available at: https://www.doi.org/10.3390/ijms19041009.

[44] Zhang, Z., Wang, L., Gao, Y., Zhang, J., Zhenirovskyy, M. et al. Predicting folding free energy changes upon single point mutations. *Bioinformatics*. march 2012, vol. 28, no. 5, p. 664–71. DOI: 10.1093/bioinformatics/bts005. Available at: https://www.doi.org/10.1093/bioinformatics/bts005.

Appendix A

Content of DVD

Content of the attached DVD contains following directories or files:

- directory **/thesis** with source code of the thesis and text of the thesis
- directory **/fireprotdb-backend** with source code of the web service
- directory **/fireprotdb-frontend** with source code of the web application
- file **fireprotdb.sql** – database schema together with the data
- file **FireProtDB.pdf** – journal article published in *Nucleic Acids Research*
- file **FireProtDB_ERD.pdf** – Entity Relationship Diagram of the database
- file **docker-compose.yml** – definitions of Docker containers
- file **README.md** – usage guide for application

Appendix B

Journal paper

Paper with title FireProtDB: database of manually curated protein stability data was published in Nucleic Acids Research journal (impact factor 11.501) on November 9, 2020.

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-

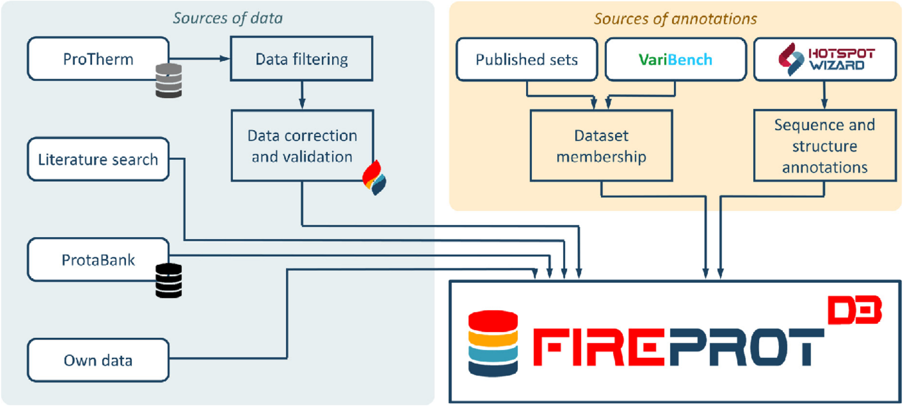


Figure 1. A schematic representation of the data comprising FireProt^{DB}. The primary source of data is filtered ProTherm (7). The FireProt data subset (18) was manually curated, compared to the source publications, and marked with the 'curated' flag. The publications from ProtaBank (12) and manual 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.

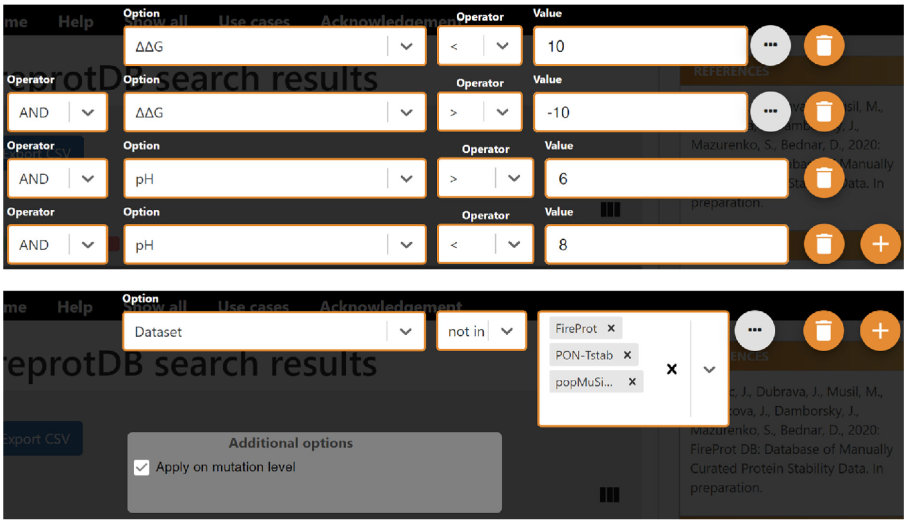
RESULTS

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

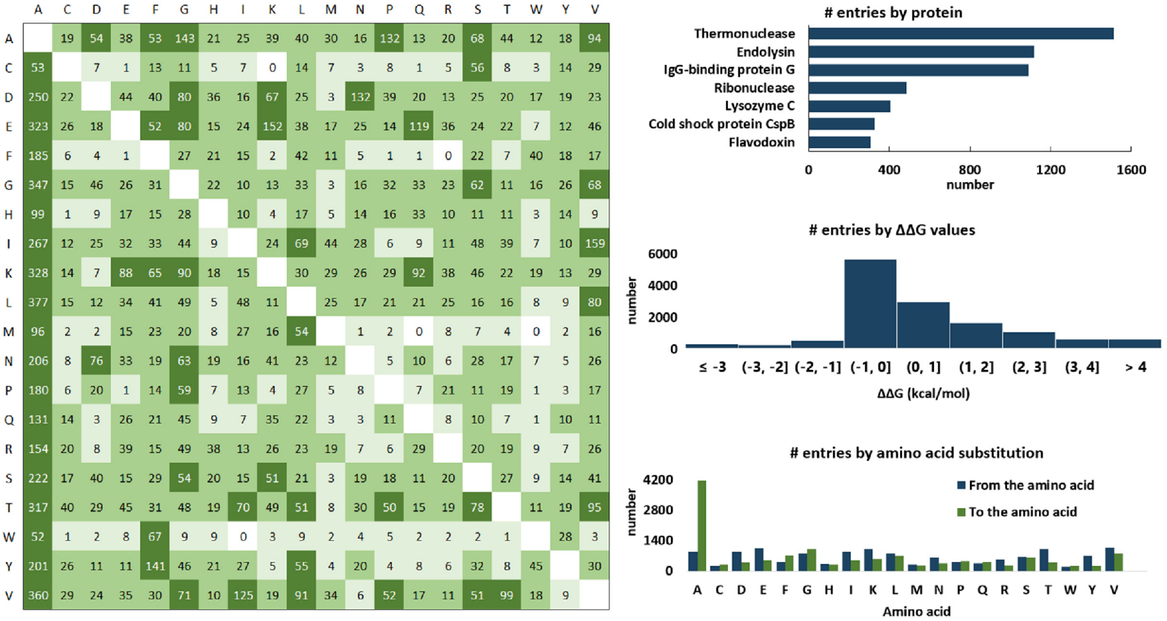
Option		
Dataset	not in	FireProt × PON-Tstab × popMuSi... ×

Additional options

☑ Apply on mutation level

me
Help
Show all
Use cases
Acknowledgement
eprotDB search results
Export CSV
REFERENCES
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 ($\Delta T_m \leftarrow 1$ or $\Delta\Delta G > 1$ kcal/mol), 14% stabilizing ($\Delta T_m >$ 1 or $\Delta\Delta G \leftarrow 1$ kcal/mol), and 43% considered neutral (–1 $\leq \Delta T_m \leq 1$ 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.