



**BRNO UNIVERSITY OF TECHNOLOGY**

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

**FACULTY OF INFORMATION TECHNOLOGY**

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

**DEPARTMENT OF COMPUTER SYSTEMS**

ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

**DEVELOPMENT OF WEB-BASED COMPUTATIONAL  
TOOL FOR ANALYSIS OF MICROCALORIMETRY DATA**

VÝVOJ WEBOVÉHO NÁSTROJE PRO ANALÝZU KALORIMETRICKÝCH DAT

**BACHELOR'S THESIS**

BAKALÁŘSKÁ PRÁCE

**AUTHOR**

AUTOR PRÁCE

**SAVA NEDELJKOVIĆ**

**SUPERVISOR**

VEDOUCÍ PRÁCE

**Ing. LENKA SUMBALOVÁ**

**BRNO 2018**

## **Zadání bakalářské práce**

Řešitel: **Nedeljković Sava**

Obor: Informační technologie

Téma: **Vývoj webového nástroje pro analýzu kalorimetrických dat**  
**Development of Web-Based Computational Tool for Analysis of Microcalorimetry Data**

Kategorie: Bioinformatika

### Pokyny:

1. Seznamte se s existujícími přístupy pro měření rozpustnosti proteinů s využitím kalorimetrie.
2. Seznamte se s metodami a algoritmy pro fitování experimentálních dat na vybrané modely. Zaměřte se na program Calfitter pro fitování kalorimetrických dat.
3. Navrhněte webovou aplikaci pro fitování kalorimetrických dat s využitím algoritmů použitých v programu Calfitter.
4. Navrženou webovou aplikaci implementujte a funkčnost ohodnoťte s využitím vhodných testovacích dat.
5. Zhodnoťte dosažené výsledky a diskutujte možnosti dalšího pokračování projektu.

### Literatura:

- Dr Chris M. Johnsson: Differential Scanning Calorimetry: Theory and practice
- Javier Sancho: The stability of 2-state, 3-state and more-state proteins from simple spectroscopic techniques
- Lyubarev AE, Kurganov BI: Modeling of irreversible thermal protein denaturation at varying temperature

Pro udělení zápočtu za první semestr je požadováno:

- Splnění bodů 1 a 2 zadání.

Podrobné závazné pokyny pro vypracování bakalářské práce naleznete na adrese

<http://www.fit.vutbr.cz/info/szz/>

Technická zpráva bakalářské práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap (20 až 30% celkového rozsahu technické zprávy).

Student odevzdá v jednom výtisku technickou zprávu a v elektronické podobě zdrojový text technické zprávy, úplnou programovou dokumentaci a zdrojové texty programů. Informace v elektronické podobě budou uloženy na standardním nepřepisovatelném paměťovém médiu (CD-R, DVD-R, apod.), které bude vloženo do písemné zprávy tak, aby nemohlo dojít k jeho ztrátě při běžné manipulaci.

Vedoucí: **Sumbalová Lenka, Ing.**, UIFS FIT VUT

Datum zadání: 1. listopadu 2017

Datum odevzdání: 16. května 2018

**VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ**  
Fakulta Informačních technologií  
Ústav informačních systémů  
602 00 Brno, Bžetěchova 2

doc. Dr. Ing. Dušan Kolář  
vedoucí ústavu

## Abstract

Proteins are organic molecules made up of amino acids which are present in all living organisms. They are the main building blocks of life. Study of protein stability has great application in pharmaceutical, biochemical and medical industries, which makes it in high demand. Stability of a protein and many other properties can be determined by observing its unfolding mechanisms. Differential Scanning Calorimetry (DSC) has proven to be a very precise method for this task. The main goal of this thesis is to improve functions of software CalFitter, which analyses data obtained from DSC experiments. The new functionality of the software enables it to automatically calculate initial parameters, which are needed for successful analyses.

## Abstrakt

Proteiny jsou organické molekuly složené z aminokyselin, které jsou přítomny ve všech živých organizmech. Jsou to hlavní stavební prvky života. Zkoumání stability proteinů má velké uplatnění ve farmaceutickém, biochemickém a zdravotnickém průmyslu, proto je o něj velký zájem. Stabilitu proteinu a mnoho dalších vlastností lze stanovit pozorováním jeho mechanismu rozkládání. Diferenciální skenovací kalorimetrie (DSC) se ukázala jako velmi přesná metoda pro tento úkol. Cílem této práce je zlepšit funkce softwaru CalFitter, který analyzuje data získaná z experimentů DSC. Nové funkce softwaru umožňují automaticky vypočítat počáteční parametry, které jsou potřebné pro úspěšnou analýzu.

## Keywords

Protein engineering, protein stability, protein unfolding, calorimetry, DSC

## Klíčová slova

Proteinové inženýrství, stabilita proteinů, rozkládání proteinů, kalorimetrie, DSC

## Reference

NEDELJKOVIĆ, Sava. *Development of Web-Based Computational Tool for Analysis of Microcalorimetry Data*. Brno, 2018. Bachelor's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Ing. Lenka Sumbalová

## Rozšířený abstrakt

Proteiny jsou hlavními stavebními kameny všech živých organismů. Plní nejdůležitější funkce nezbytné pro řádné fungování organismů. Bez proteinů by život, jak ho známe, neexistoval. Studium proteinů, porozumění jejich funkcí a vlastnostem je velmi důležité pro mnoho oborů, jako jsou například lékárnictví, lékařství a biotechnologie. Jednou z nejdůležitějších vlastností proteinů je stabilita. Stabilita určuje, jak se bude protein chovat v různých prostředích a jak ovlivní funkce živého organismu, ve kterém se nachází.

V dnešní době neexistuje mnoho přesných teoretických modelů pro stanovení stability a dalších vlastností proteinů. Z tohoto důvodu jsou vlastnosti proteinů obvykle určovány analýzou dat získaných z nějaké experimentální metody. Diferenciální skenovací kalorimetrie (DSC) se ukázala jako velmi praktická a přesná metoda pro získání relevantních dat, na základě kterých je možné vypočítat množinu parametrů popisující různé vlastnosti proteinu. Tato metoda má dlouhou historii zajímavých a užitečných aplikací v oblasti proteinového inženýrství, a také v analýze jiných oblastí biologie. Jejím cílem je charakterizovat konformační změny vyvolané změnou teploty v proteinech a jiných biologických molekulách.

Neexistuje mnoho programů pro analýzu dat získaných pomocí experimentů DSC. Toto byla hlavní motivace pro vývoj programu CalFitter. CalFitter automatizuje celý proces analýzy dat. V ideálním případě by program měl vypočítat všechno sám a interakce uživatele s programem by měla být minimální. Schopnost programu vypočítat hodnoty určitých parametrů potřebných pro analýzu je vyžadována z toho důvodu, aby je uživatel nemusel zadávat sám.

Cílem této bakalářské práce bylo přidat do softwaru nové funkce, které by umožnily automaticky vypočítat počáteční hodnoty parametrů pro analýzu dat. Hodnoty těchto parametrů CalFitter později vylepší metodou nelineární regrese. Ačkoli tato práce přispívá k již existujícímu projektu CalFitter, výsledky a skripty se mohou použít i samostatně za předpokladu, že by se ostatní potřebné výpočty provedly ručně.

Během DSC experimentu protein může přejít z biologicky aktivní konformací, tj. z nativního (N) stavu do neaktivní konformace, tj. denaturovaného (D) stavu. Tento proces se nazývá rozklad proteinů nebo také proteolýza. Rozklad je u DSC experimentu způsoben zvyšováním teploty. Přejod z N do D se může skládat z několika kroků. Každý krok může být reverzibilní nebo nereverzibilní. Aby se parametry vypočítaly, je potřeba znát, z kolika kroků se rozklad skládá a které kroky jsou reverzibilní či nereverzibilní. Tyto údaje je nutno zadat ručně. Skript, který je výsledkem této bakalářské práce, pak provede analýzu vstupních dat získaných z DSC experimentu. Během analýzy se nad těmito daty vykonají různé matematické operace, jako např. numerická integrace a hledání lokálního maxima. Výsledkem analýzy jsou parametry popisující rozklad proteinu. Na základě těchto parametrů je možné modelovat tj. zrekonstruovat původní data z DSC experimentu. Čím méně se modelovaná data liší od původních, tím jsou parametry lepší. Odchylka se může vypočítat např. metodou nejmenších čtverců. CalFitter potom provede fitování modelovaných dat, co by mělo zlepšit hodnoty parametrů. Aby fitování proběhlo úspěšně, je důležité, aby počáteční parametry byly dostatečně přesné.

Výsledný skript může vypočítat modelovací parametry rozkladu s 1, 2 nebo 3 kroky. Parametry jsou dostatečně přesné, aby se úspěšně mohlo provést fitování. Experimenty ukázaly, že čím je víc kroků rozkladů, tím je přesnost výsledků menší. Důvodem je to, že některé údaje nelze vypočítat na základě vstupních dat, a proto je potřeba některé parametry odhadnout. Tyto odhady nejsou vždy správné. Skript má větší úspěch s analýzou reverzibilních kroků než nereverzibilních, protože reverzibilní kroky lze jednodušeji modelovat a analyzovat.

Prostor pro budoucí vylepšení. Například výpočet počátečních parametrů a postup fitování by se mohli sloučit do jednoho procesu. Mohou být přidány také nové typy modelů rozkladu. Program CalFitter nicméně nabízí širokou škálu modelů a může být použit jako profesionální nástroj. O programu CalFitter byl publikován vědecký článek v časopise *Nucleic Acids Research* [29]. Tato práce byla vytvořena ve spolupráci s Loschmidtovými laboratoři Masarykovy univerzity.

# Development of Web-Based Computational Tool for Analysis of Microcalorimetry Data

## Declaration

Hereby I declare that this bachelor's thesis was prepared as an original author's work under the supervision of Ing. Lenka Sumbalová. The supplementary information was provided by Stanislav Mazurenko, Dr. and Mgr. Jan Stourač. All the relevant information sources, which were used during preparation of this thesis, are properly cited and included in the list of references.

.....

Sava Nedeljković

May 16, 2018

## Acknowledgements

I would like to thank Ing. Lenka Sumbalová, Stanislav Mazurenko, Dr. and Mgr. Jan Stourač for their helpful advice, introduction to the topic and providing me all the necessary materials and information.

# Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b>                         | <b>2</b>  |
| <b>2</b> | <b>Theoretical Background</b>               | <b>4</b>  |
| 2.1      | Biological Functions of Proteins . . . . .  | 4         |
| 2.2      | Protein Structure . . . . .                 | 6         |
| 2.3      | Protein Denaturation . . . . .              | 12        |
| 2.4      | Differential Scanning Calorimetry . . . . . | 16        |
| <b>3</b> | <b>Introduction to CalFitter</b>            | <b>21</b> |
| 3.1      | Denaturation Models Notation . . . . .      | 22        |
| 3.2      | Use Case . . . . .                          | 22        |
| <b>4</b> | <b>Modeling Parameters Calculation</b>      | <b>25</b> |
| 4.1      | Thermodynamic Analysis . . . . .            | 25        |
| 4.2      | Reversible Denaturation Model . . . . .     | 28        |
| 4.3      | Irreversible Denaturation Model . . . . .   | 31        |
| 4.4      | Multi Step Denaturation Modeling . . . . .  | 32        |
| <b>5</b> | <b>Implementation</b>                       | <b>34</b> |
| 5.1      | Working principle . . . . .                 | 35        |
| 5.2      | Results . . . . .                           | 40        |
| <b>6</b> | <b>Conclusion</b>                           | <b>42</b> |
|          | <b>Bibliography</b>                         | <b>44</b> |
| <b>A</b> | <b>User Manual</b>                          | <b>47</b> |
| A.1      | Input Data Format . . . . .                 | 47        |
| A.2      | Data Upload and Model Selection . . . . .   | 48        |
| A.3      | Run Initial Calculation Script . . . . .    | 48        |
| <b>B</b> | <b>DVD Contents</b>                         | <b>50</b> |

# Chapter 1

## Introduction

Proteins are the main building blocks of all living organisms. They fulfill the most important functions necessary for the proper functioning of organisms. Without proteins, life as we know it, would not exist. Studying proteins, understanding their functions and properties is very important for many disciplines such as pharmacy, medicine, biotechnologies etc [18].

One of the most important properties of a protein is its stability. It determines how the protein will behave in different environments and how it will affect functions of a living organism where protein resides. In biochemistry, stability is defined as physical property. It is mainly referred to as thermodynamic or kinetic stability and not as chemical stability. Stability determines the state of a protein (native or denatured)[24].

Nowadays there are no many precise theoretical models for determining stability and other properties of a protein. For this reason, protein's properties are usually determined by analyzing data obtained from some experimental method. Differential Scanning Calorimetry (DSC) has proven to be very practical and precise technique for obtaining relevant data which describe parameters for both thermodynamic and kinetic stability of a protein. By analyzing and processing these data we can get set of parameters that can many different protein properties [16].

There are no many programs for analyzing data obtained by DSC experiments. Usually, scientists have to calculate necessary parameters manually or by making a program which will perform the necessary calculations. This was the main motivation for developing software CalFitter. The main goal of CalFitter is to automate the whole process of data analysis. Everything should be calculated by this software and user interaction with the software should be minimal. In order to achieve this, the functionality of calculating initial parameters, which normally had to be specified by the user, was a crucial part of the development.

This bachelor thesis is done in cooperation with Loschmidt Laboratories (LL) of Masaryk University (MUNI). It continues previous development of CalFitter software, which I was also a part of. CalFitter is a newly developed program for curve fitting and computational analysis calorimetric data obtained from Differential Scanning Calorimetry measurements. It was developed by LL MUNI. The main goal of the thesis was to add new functionality to the software, which would enable it to automatically calculate initial values of parameters for data analysis. These parameters describe proteins properties. Parameters' values are



later improved by performing curve fitting. Although this thesis is contributing to already existing project CalFitter, results and scripts may be used individually, provided that other necessary calculations, which CalFitter provides, are done manually.

The CalFitter software has been published under a name „CalFitter: Web Server for Protein Thermal Denaturation Data Analysis“ in journal Nucleic Acids Research [29].

Chapter 2 briefly explains some of the most important properties of the protein, how they correlate and why are they important for this thesis. In Chapter 3 description of the basic functionality of the software, CalFitter is given. It is also explained how this thesis contributes to its further development. Chapter 4 gives theoretical background behind the calculations of initial values of modeling parameters, which are needed by the CalFitter software. Design approaches and implementation of necessary calculations are explained in Chapter 5. In Chapter 6 results and experiments are summarized and possible future improvements are discussed.

## Chapter 2

# Theoretical Background

### 2.1 Biological Functions of Proteins

Proteins are very important macromolecules. They perform essential functions throughout the systems of all living cells. Proteins cover a wide range of functions. For example, there are proteins that serve as structural elements, transportation channels, signal receptors, and transmitters, etc. They have many different active functional groups attached to them to help define their properties and functions [15]. We can classify them by their shape, size, internal structure and by their biological roles within the living organism. They fulfill many tasks, which are necessary for sustaining life. They are one of the most important organic chemicals and they have the key role in the life cycle of every living cell. We can roughly separate them based on their biological functions (among other properties) into several groups:

- Enzymes
- Hormones
- Structural Proteins
- Nutrient Proteins
- Defence Proteins
- DNA Associated Proteins
- Other Proteins

Enzymes are so-called catalysts, which means they increase the rate of chemical reactions that take place in living systems, but they are also consumed by the very same reaction. Most of the necessary chemical reactions in the body would not efficiently take place without enzymes. For example, one type of enzyme functions as an aid in digesting large protein, carbohydrate, and fat molecules into smaller molecules. Enzymes are the most varied and specialized proteins [9].

Proteins which control cellular activity and physiological processes are called hormones. They are in charge of many bodily functions and they enable cell signaling and communication [11]. One of the most important hormones is secretin. It stimulates the intestines and

pancreas in order to help digestive processes [19]. Another great example of protein hormones is insulin. It regulates sugar metabolism by controlling the concentration of glucose inside cells and thus has the most important role in the control of diabetes [19]. Growth hormone is also protein hormone.

Structural proteins provide support in our bodies and structure for cells. They are also responsible for the ability of our bodies to move. One of the most important structural protein is elastin. It has the ability to stretch in 2 dimensions and is mostly found in ligaments [30]. Another member of structural proteins is keratin, which is found in fingernails, hair, feathers, etc [8]. Collagen is also structural protein. It gives strength to various parts of a living organism [6]. Fibroin is found in silk and spider web.

Proteins can serve as a major source of nutrients and energy. Many various proteins are found in milk, egg whites, plant seeds (especially corn, wheat and rice) which makes them rich in energy and nutrient. If living organism consumes more protein than needed for sustaining life, an organism can use it as an additional source of energy or it can be used for creating fat if it's not used at all.

Some proteins form antibodies and are used for defense of the living organism. They can prevent illness, disease, and infection. They can even bind to specific viruses and bacteria to help protect the organism. They are part of the whole immune system and are in combination with other defensive mechanisms. Fibrinogen and thrombin are blood clotting proteins which stop the loss of blood when damage occurs to the vascular system.

A gene is a segment of a DNA molecule that contains the instructions needed to make a unique protein. All of our cells contain the same DNA molecules, but each cell uses a different combination of genes to build the particular proteins it needs to perform its specialized functions. But proteins are used to assist with the formation of new molecules by reading the genetic information stored in DNA. DNA associated proteins regulate chromosome structure during cell division or play a role in regulating gene expression. The creation of DNA could not happen without the action of these proteins.

There are numerous other proteins whose functions are exotic in nature and therefore, are not easily classified. There are many different ways that proteins contribute to the structure of living things. But they can also be very unhealthy even dangerous. There are a lot of people who suffer from various allergies. For example reaction to pollen allergy is caused by proteins on the surface of pollen. A lot of venoms and toxins are also proteins, one of the examples being protein based snake venom [24][4][5].

Because of such versatile functions of proteins, they are the goal of many scientific disciplines. Study of proteins' properties helps us better properties of life itself.

## 2.2 Protein Structure

Protein is made of very long strands of amino acids linked together in specific sequences. Those amino acids are small organic molecules that consist of an alpha carbon atom linked to an amino group, a carboxyl group, a hydrogen atom, and a variable component called a side chain. The side chain is also referred to as R group. Within a protein, multiple amino acids are linked together by peptide bonds, thereby forming a long chain. The largest group of amino acids have nonpolar side chains. Several other amino acids have side chains with positive or negative charges, while others have polar but uncharged side chains. The chemistry of amino acid side chains is critical to protein structure because these side chains can bond with one another to hold a length of protein in a certain shape or conformation. Charged amino acid side chains can form ionic bonds, and polar amino acids are capable of forming hydrogen bonds. Hydrophobic side chains interact with each other via weak van der Waals interactions. The vast majority of bonds formed by these side chains are noncovalent. Because of side chain interactions, the sequence and location of amino acids in a particular protein guides where the bends and folds occur in that protein [7].

The term structure when used in relation to proteins, takes on a much more complex meaning than it does for other molecules. The structure and shape of a protein are critical to its function because it determines how the protein will interact with other molecules. Protein structure depends on its amino acid sequence and local, low-energy chemical bonds between atoms in both the polypeptide backbone and in amino acid side chains. Protein structure plays a key role in its function, for example, if a protein loses its shape at any structural level, it may no longer be functional. Protein structures are very complex, and researchers have only very recently been able to easily and quickly determine the structure of complete proteins down to the atomic level. Protein structure is the three-dimensional arrangement of atoms in an amino acid-chain molecule. There are 4 different types of protein structure which determine how the protein gets its final shape or conformation. These structure types are called primary, secondary, tertiary, and quaternary structure [18][15].

### Primary Structure

The simplest level of protein structure is the primary structure. The primary structure is simply the linear sequence of amino acids in a polypeptide chain. It may be thought of as a complete description of all of the covalent bonding in a polypeptide chain or protein. The two ends of the polypeptide chain are referred to as the carboxyl terminus (C-terminus) and the amino terminus (N-terminus) based on the nature of the free group on each extremity [7][14].

The most common way to denote a primary structure is to write the amino acid sequence using the standard three-letter abbreviations for the amino acids. For example, Gly-Gly-Ser-Ala is the primary structure for a polypeptide composed of glycine, glycine, serine, and alanine, in that order, from the N-terminal amino acid (glycine) to the C-terminal amino acid (alanine). Diagram of pancreatic hormone insulin is shown in the Figure 2.1. Insulin has two polypeptide chains, A and B. Each chain has its own set of amino acids assembled in a particular order. For instance, the sequence of the A chain consists of 21 amino acids. It

starts with glycine at the N-terminus and ends with asparagine at the C-terminus. B chain consists of 30 amino acids and starts with phenylalanine and ends with threonine[7][14].

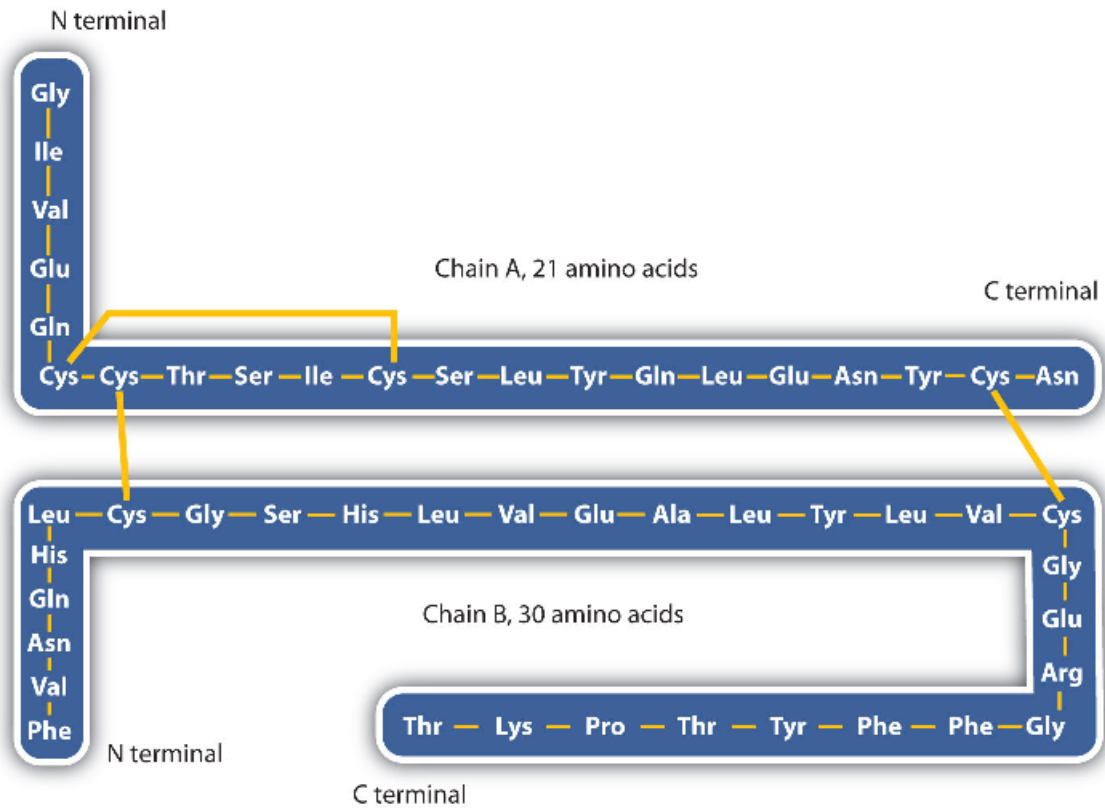


Figure 2.1: Insulin protein structure. Source: [31]

Protein primary structures can be directly sequenced, or inferred from DNA sequences. In other words, DNA sequences carry the information which makes polypeptides with a defined amino acid sequence. A change in a nucleotide sequence of the DNA sequence's coding region may lead to a different amino acid being added to the growing polypeptide chain, causing a change in protein structure and therefore its function. An average polypeptide is about 300 amino acids in length, and some genes encode polypeptides that are a few thousand amino acids long. While the amino acid sequence makes up the primary structure of the protein, the chemical and biological properties of the protein are very much dependent on the three-dimensional or tertiary structure [15].

## Secondary Structure

Secondary structure refers to the coiling or local folding of a polypeptide chain that gives the protein its three-dimensional shape. Secondary structures arise as hydrogen bonds form between local groups of amino acids in a region of the polypeptide chain. Secondary structure however rarely is present in the whole polypeptide chain. It is usually present just in a section of the chain[7][14].

There are two main types of secondary structures observed in proteins:

- Alpha Helix structure

- Beta Pleated Sheet

Alpha helix structure resembles a coiled spring and is secured by hydrogen bonding in the polypeptide chain. Simplified diagram of the alpha helix structure is shown in Figure 2.2. „The polypeptide backbone forms a repeating helical structure that is stabilized by hydrogen bonds between a carbonyl oxygen and an amine hydrogen. These hydrogen bonds occur at regular intervals of one hydrogen bond every fourth amino acid and cause the polypeptide backbone to form a helix. The most common helical structure is a right-handed helix with its hydrogen bonds parallel to its axis. The hydrogen bonds are formed between carbonyl oxygen and amine hydrogen groups of four amino acid residues away. There is an average of ten amino acid residues per helix with its side chains orientated outside of the helix. Different amino acids have different propensities for forming Alpha helix structure. Amino acids that prefer to adopt helical conformations in proteins include methionine, alanine, leucine, glutamate, and lysine.“ [18]

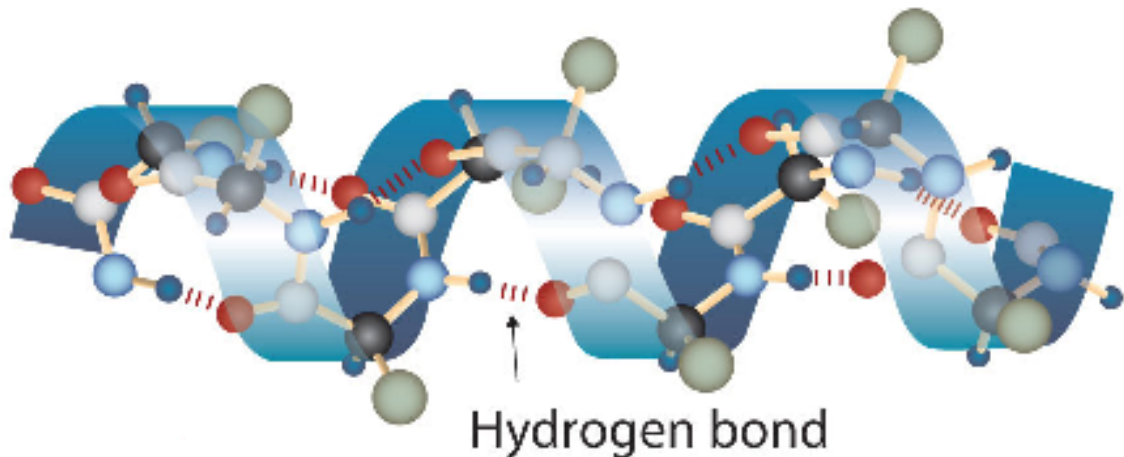


Figure 2.2: Secondary Alpha helix structure. Source: [31]

The second type of secondary structure in proteins is the Beta pleated sheet. This structure appears to be folded or pleated and is held together by hydrogen bonding between polypeptide units of the folded chain that lie adjacent to one another. Simplified diagram of the Beta pleated sheet is shown in Figure 2.3.

In beta-pleated sheet regions of the polypeptide backbone come to lie parallel to each other and are connected by hydrogen bonds. The hydrogen bonds are formed between the carbonyl oxygen and the amine hydrogen of amino acid in adjacent strands in a polypeptide, which means that the hydrogen bonds are inter-stand. Beta pleated sheet regions are more extended than an alpha helix. Hydrogen bonding in beta-strand can occur as parallel, anti-parallel or a mixture. Amino acid residues in beta-parallel configuration run in the same orientation. Beta pleated sheets never occur alone. They have to hold in place by other beta-pleated sheets. Amino acids which prefer to adopt beta-pleated sheet structures are isoleucine, valine, and threonine. This orientation is energetically favorable because of its slanted, non-vertical hydrogen bonds[7][14].

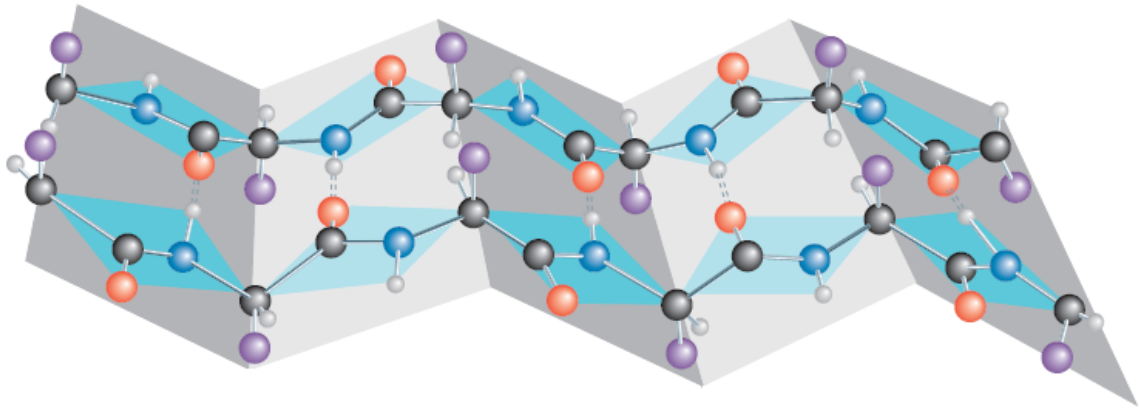


Figure 2.3: Secondary Beta pleated sheet structure. Source: [2]

### Tertiary Structure

Tertiary structure is a three-dimensional structure of the polypeptide chain of a protein, which is the overall shape of a polypeptide. It is the result of folding and refolding of secondary structured polypeptide upon itself. Simplified diagram of the Tertiary structure is shown in Figure 2.4. The alpha helixes and Beta pleated sheet are folded into a compact globular structure.

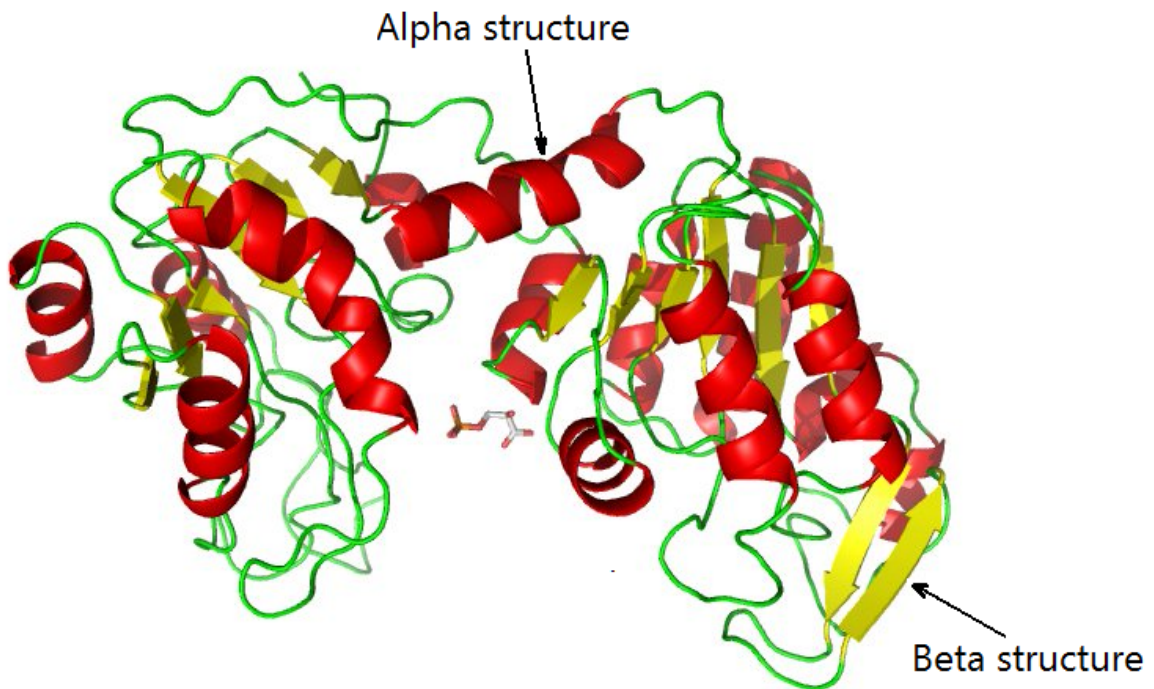


Figure 2.4: Tertiary structure. Source: [3]

Although the three-dimensional shape of a protein may seem irregular and random, it is influenced by many stabilizing forces due to bonding interactions between the side-chain groups of the amino acids. There are several types of bonds and forces that hold a protein in

its tertiary structure. A shape of a protein is highly influenced by hydrophobic interactions of amino acids.

Side chain, also known as R group, which is variable part of amino acids, can either have hydrophobic or hydrophilic properties. The amino acids with hydrophobic side chains attempt to avoid water and they seek to position themselves in the center part of the protein. The amino acids with hydrophilic side chains will on the other hand attempt to contact the water and thus will surround hydrophobic side chains, i.e. position themselves on the outer part of the protein. The final shape of the protein is stabilized by hydrogen bonding between side chains. Covalent bonding can also occur between side chains which come in close contact and thus create so-called disulfide bridge. Side chains with opposite charges which are in close contact can also create ionic bonds. Another factor which contributes to the stabilization of protein structure is van der Waals forces, which can occur between molecules [13].

### Quaternary Structure

Quaternary structure is also a three-dimensional structure. It consists of two or more polypeptide chains which have tertiary structure and are connected to one another by intermolecular interactions. Proteins made from a single polypeptide cannot have a quaternary structure. Each polypeptide chain is usually referred to as subunit. Such group of subunits then results in creating of one single operational unit, i.e. quaternary structure. Group of subunits is usually referred to as multimers, i.e. they consist of multiple polypeptide chains (subunits). Based on the number of subunits, multimer can be a dimer (created by exactly 2 subunits), a trimer (created by exactly 3 subunits), a tetramer (created by exactly 4 subunits), and a pentamer (created by exactly 5 subunits). Subunits can be all the same, in which case they will create a so-called homodimer. They can also be different and in such case will create a so-called heterodimer [19][7]. Simplified diagram of the Quaternary structure is shown in Figure 2.5.

Hemoglobin is an example of a protein with quaternary structure. Hemoglobin, found in the blood, is an iron-containing protein that binds oxygen molecules. It contains four subunits thus is a tetramer. Hemoglobin is also a heterodimer because subunits, from which it consists of are two alpha subunits and two beta subunits, i.e. they are not all the same[19].

Generally speaking, all subunits are assembled together by same interactions and forces which are also involved in tertiary structures, mostly weak interactions, such as hydrogen bonding, disulfide bridges, London dispersion forces, etc[5].

Overall the various forces and interactions stabilize the final shape of the protein complex. All four types of protein structures and how they are related is shown in Figure 2.6.



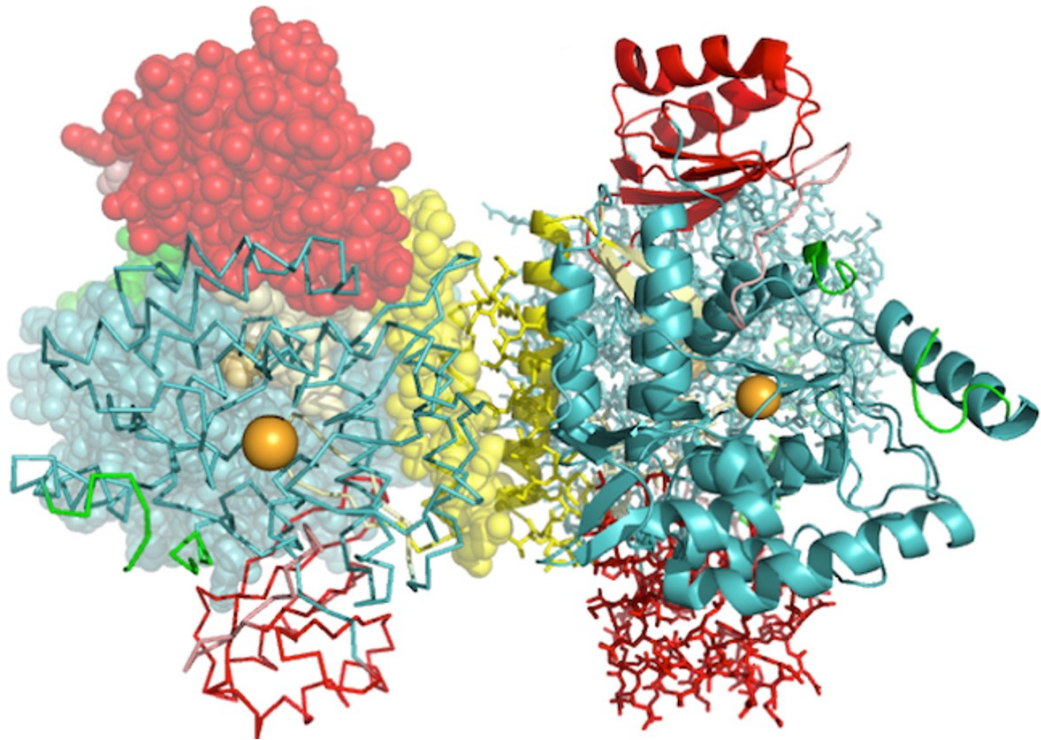


Figure 2.5: Quaternary structure. Source: [12]

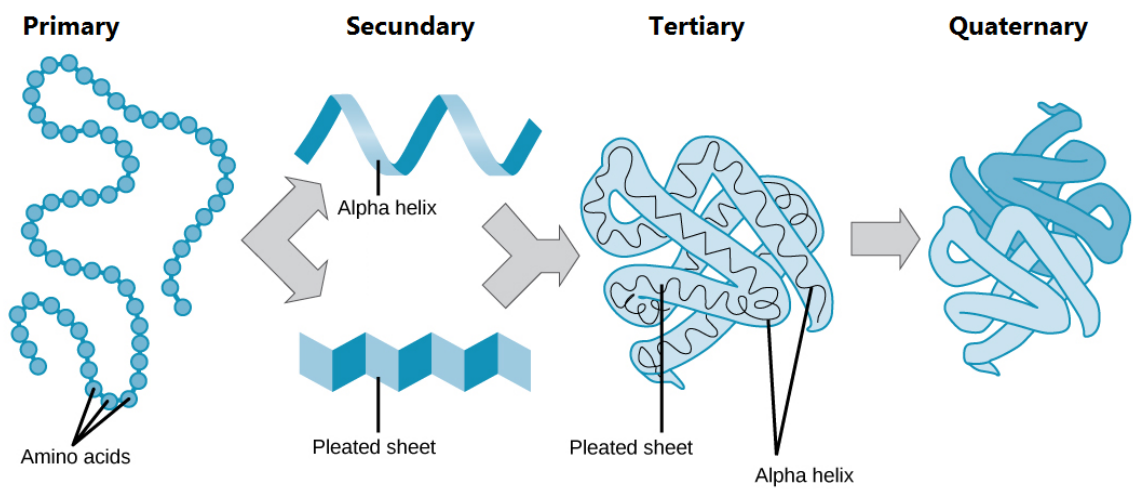


Figure 2.6: All 4 protein structures. Source: [18]

## 2.3 Protein Denaturation

Denaturation is a process which disrupts and modifies protein's structure. During this process, many weak chemical bonds and interactions inside protein's molecule tend to break. These bonds are responsible for proteins original structure, i.e. protein's native state. Disruption of these chemical bonds leads to a loss of protein's three-dimensional structure, i.e. protein loses its quaternary, tertiary and/or secondary structure. Peptide bonds among individual amino acids are however strong enough to resist denaturation factors. Therefore protein will always retain its primary (single dimensional) structure [26][21].

In case of quaternary structure disruption, intermolecular connections between subunits inside protein are broken, which may lead to separation of subunits into individual tertiary structures.

When it comes to tertiary structure disruption, chemical bonds of individual side chains are broken. Since there are 4 types of bonds among side chains (hydrophobic molecular interaction, ionic, covalent and hydrogen bonding), disruption can be caused by a various combination of agents and conditions. Side chains are separated into an alpha helix and beta-pleated secondary structures.

Secondary structure is disrupted in a way, that makes it lose regular repeating patterns of amino acid chains, thus it losses alpha helix or beta-pleated shape.

Disruption of protein's structure can also be referred to as protein unfolding. During unfolding, protein changes its state from native to denatured. Since denaturation highly affects protein's structure and protein structure are major factors in protein's biological, chemical and physical properties of a protein, denaturation is responsible for the loss of protein's original functionality. This then usually causes living cell, in which protein resides, to become biologically inactive or even dead. However, the death of a living cell can also initiate protein denaturation. Denaturation makes protein become insoluble. It decreases protein's surface tension and increases its viscosity. All of these changes make the cell change its propensities as well. A real-life everyday example of protein denaturation effect is high digestibility of cooked food. This specific property is caused by exposure of amino acid peptide bonds to digestive enzymes.

There are 2 types of denaturation:

- Irreversible
- Reversible

Irreversible denaturation is most common. During this process, a protein may undergo 1 or 2 state changes before it gets into a denatured state. The effects of denaturation cannot be reversed, so the protein will remain in the denatured state indefinitely. An example of a protein with irreversible denaturation is egg whites. The usual reason for lack of reversibility is a loss of biological functions during denaturation [21].

Denaturation is however in many cases reversible. In such cases, protein can go back to its original state and therefore return to its native state. This reverse process is referred

to as renaturation and can also be called refolding. It is usually initiated upon removal of initial denaturation factors and when conditions for proteins native state are established again. For example, if denaturation is caused by heating the protein, renaturation may in some cases be caused by cooling the same protein. Reversibility of a denaturation depends on many conditions. If the concentration of the unfolded protein is low, or if there are hydrophobic sequestering chaperones, there's more of a chance that the protein will exist in relative isolation for long enough to refold. Proteins that can undergo reversible denaturation are hemoglobin (red blood cell pigment for carrying oxygen), ribonuclease (nuclease which speeds up RNA reduction into smaller parts), etc [21].

There are many different factors that can cause denaturation. Some of the most common factors of denaturation are:

- Heat
- Physical pressure
- Presence of salt
- Change of pH level

Heat is the most relevant factor of denaturation to this bachelor's thesis because it is used in the experimental part of the thesis as part of Differential Scanning Calorimetry technique to collect necessary data.

When increasing temperature of a solution containing a protein by transferring heat to it will cause the increase of kinetic energy of the solution. Consequently, with increased kinetic energy molecular motion will also increase. Molecules inside the protein will increase the rate of their vibration. These vibrations may cause individual molecules to break out of weak chemical bonds, such as hydrophobic interactions, hydrogen bonds, etc. Temperature, at which protein properties start to change rapidly, is called melting temperature. Getting its value is very similar to the process of getting the melting temperature of a solid, hence the name similarity. Melting temperature of protein denaturation may be different for various proteins. In many cases melting temperature can be as low as 40 °C, which implies the dangers of high fever for human health. Heat-induced denaturation has the greatest contribution in sterilization by heat (heat increases temperature, high temperature causes denaturation, denaturation causes bacteria to die)[26].

An everyday real-life example of heat-induced denaturation can be observed in cooking egg whites. Before cooking, egg whites are transparent and they are in a liquid form. During cooking, they slowly start to harden and change color to white, which is the effect of protein denaturation. This is an example of an irreversible denaturation. Figure 2.7 illustrates how heat untangles individual protein structure and how that affects color and solubility of a frying egg. Of course, if a protein is kept at a low temperature, that will reduce the chances of heat-induced denaturation[32].

Applying physical force or pressure can also cause denaturation. When physical force such as shaking or fast stirring is applied to a solution containing a protein, individual protein structures will start to aggregate and thus causing the protein to denature. We can observe this phenomenon again in eggs. When we whip egg whites, they'll slowly change

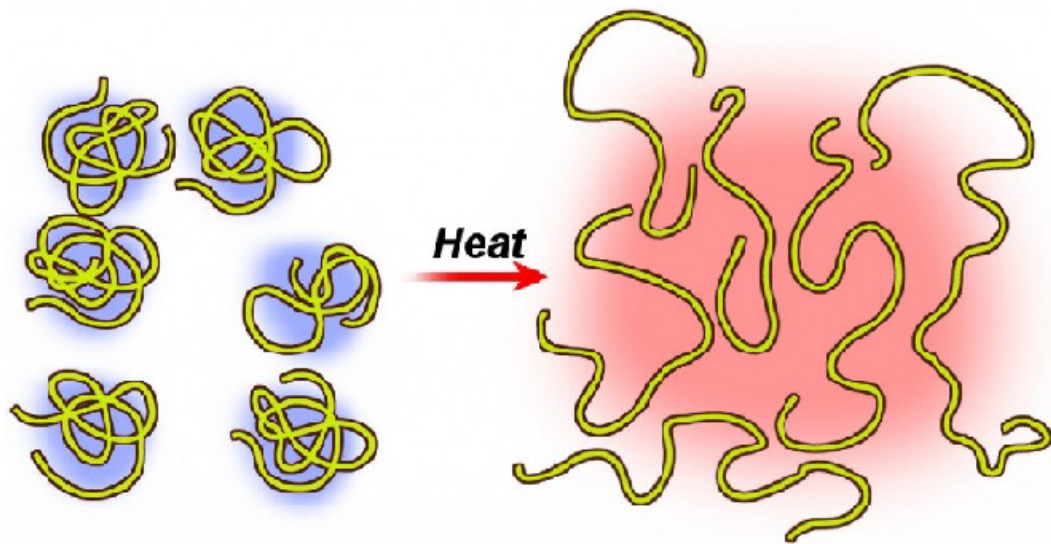


Figure 2.7: Heat induced denaturation diagram. Source [27]

their viscosity and color, they will have more foamy structure and white color. Another example is foam in the sea water, which is caused by various proteins inside the water.

Denaturation can be caused by the mere presence of salts, especially those form heavy metal. Because heavy metal salts are ionic they can react with ionic bonds in protein (which are often called salt bridges). This usually causes denaturation. „For example, silver nitrate is used to prevent gonorrhoea infections in the eyes of newborn infants. Silver nitrate is also used in the treatment of nose and throat infections, as well as to cauterize wounds“. Presence of other salts can have the opposite effect. For example, ammonium sulfate will increase protein melting temperature and thus stabilize the protein. Various bases, acids and some compounds such as alcohol can cause similar reactions like salt [25].

Changes in pH affect the chemistry of amino acid residues and can lead to denaturation. Hydrogen bonding often involves these side changes. Protonation of the amino acid residues (when an acidic proton of hydrogen attaches to a lone pair of electrons on a nitrogen) changes whether or not they participate in hydrogen bonding, so a change in the pH can denature a protein [22].

There are also many other factors that cause protein denaturation, but they are not the subject of this thesis.

Denaturation of proteins is a very common occurrence in nature, so studying it is of great significance for many scientific departments. Findings based on those studies allow us to characterize how the various proteins will interact with their surroundings in different conditions and thus help to better understand the protein properties.

## 2.4 Differential Scanning Calorimetry

Differential Scanning Calorimetry is highly precise and reliable but also a simple technique which measures temperatures and heat flow of observed substances. DSC is used in many different scientific fields like chemistry, biology, biochemistry, pharmacy, medicine, nanoscience, etc. It can be used for analysis of much different substance like proteins, oils, and fats, nucleic acids, etc [17].

DSC, as its name implies, belongs to calorimetry measuring techniques. Its main focus is analyzing thermal propensities of an observed sample and also measuring how various physical and chemical reactions occur along with the change of temperature against time. Every measured and recorded value is a function of time. In summary, DSC is used for observing and recording many heat-related properties of a given substance. Measurement is done in a controlled environment [17][1].

The first word of the technique (Differential) indicates, that differences in temperature or heat flow are recorded, not the absolute values. That means that there are always two substances which are observed. Usually, one substance serves as a reference to the other substance, i.e. we usually want to see how one substance's reactions compare to the familiar simple reactions of some other reference substance.

In this thesis, DSC is used to observe protein's properties and record data about its denaturation. One of the reasons why DSC was chosen is because the measuring apparatus was available at Loschmidt Laboratories. Another reason is that DSC is very simple compared to other popular techniques. For example, it relies solely on temperature and heat measurements, so there are no spectroscopic measurements like in many other procedures, which means that solutions containing protein don't have to be optically clear. DSC very effectively characterizes all thermodynamic transitions of a protein which are induced by a change of transferred heat. Data obtained from DSC experiment are very precise and are sufficient to characterize the observed protein, but crucial part of the whole experiment is correct data interpretation and analysis, which was the main goal of this thesis.

## Measurement Process

The calorimeter consists of 2 heating pans. 2 glass dishes should be filled with a specific liquid substance. A sample of a protein whose properties we want to observe should be prepared. Prepared Sample should be poured in one of the dishes. This dish will be referred as protein sample dish. The other dish will contain only the liquid substance which was added at the beginning of the experiment, nothing else should be added. This other dish will be referred to as reference dish. Both reference and protein sample dishes should be then placed on their own heating pan of the calorimeter.

Differential scanning calorimeter heats both pans and dishes on them at a constant rate, i.e. rate, will be a linear function of time. This rate should be specified before the experiment. It is referred to as scanning rate. For example, good value of scanning rate is 1 °C pre minute. This means that the temperature of both pans will increase by 1 degree over the course of 1 minute. Calorimeter will keep both pans at approximately the same temperature. Starting temperature is room temperature, i.e. 25 °C. Calorimeter will heat both pans at the same rate until they reach final temperature, which should be specified before the experiment. For example good value of final temperature is 100 °C. When both pans reach final temperature, calorimeter will cease to heat pans. Both dishes will then be left to cool to the room temperature. This whole process of transferring heat to the dishes and letting them cool off is referred to as the first run.

During the whole experiment, calorimeter will record heat flows going to both reference and protein sample dishes. It will also their temperature (which should be roughly the same). However, the whole point of the experiment of this thesis is to compare heat flows of two dishes. This is why only the difference between of transferred heat are reported at the end of the measurement. The principle can be explained as follows:

Since temperatures of both dishes must be roughly the same during the whole measurement, if a protein inside sample dish undergoes some physical or chemical change, it will either consume more or less heat than the reference dish. Undergoing reaction in the protein sample dish can either be endothermic (consume heat) or exothermic (produce heat). Type of reaction determines will determine if the protein sample dish take more heat (in case of endothermic reaction) or less heat (in case of exothermic reaction) than the reference cell. During one measurement, protein can undergo many different reactions.

Transferred heat is represented in a form of a heat capacity of substance in a corresponding dish. Heat capacity (sometimes known as thermal capacity) is defined as a physical quantity which is equal to the amount of heat needed to raise of an object by 1 degree. In the end, we are interested only in heat capacity of 1 mol of protein, that is why original heat capacity reading must be adjusted based on the amount of protein inside sample dish.

Sometimes the second measurement is done, right after the first measurement, i.e. first run, when both dishes have cooled to the room temperature. This new measurement is referred to as the second run, or sometimes also as reheating run. The whole point of the second run is to examine the final state of the protein after the first run since the first run will be influenced by the reactions of the first run.

The calorimeter is controlled by a computer. After the experiment, data are saved on the same computer. Figure 2.8 illustrates the DSC experiment process.

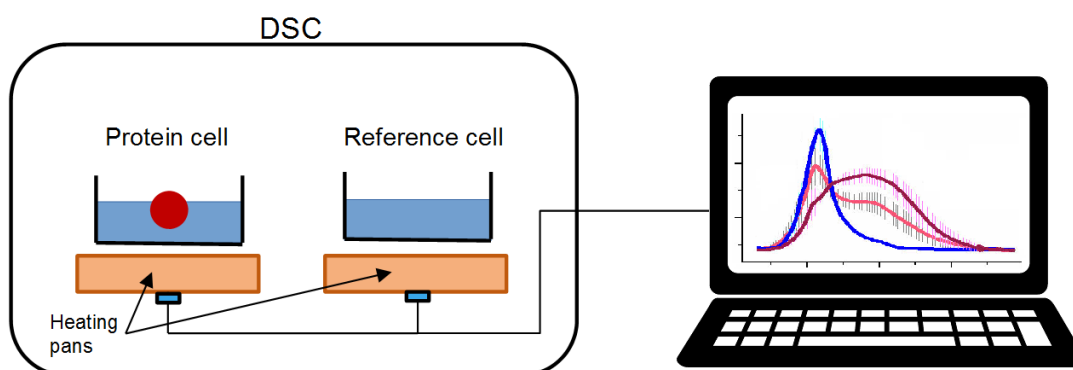


Figure 2.8: DSC process diagram.



## Output Data

Output data is saved on a computer in a form of spreadsheet document file. For each measurement, there should be 1 pair of values. This pair consists of temperature readings and difference of heat capacities between 2 dishes. That means that if only 1 measurement has been done, output spreadsheet document file should have 2 occupied columns. The output file does not contain information about final temperature and scanning rate. Those have to be filled in manually. There is also no information about temperature and heat capacity units. The calorimeter can be set to produce output data in specified units but output file will not contain any information about used units.

Usually, there will be multiple measurements on multiple samples with the same type of protein during one experiment. That is why usually output spreadsheet document file will have more than 1 pair of columns occupied. An example of the output spreadsheet file is illustrated in the Figure 2.9. In this example, there have been 3 measurements. Temperature values are stored in columns B, E, and H from row number 5. Heat capacities differences are stored in columns C, F and I, also from row number 5.

|    | B           | C                       | D | E           | F                       | G | H           | I                       |
|----|-------------|-------------------------|---|-------------|-------------------------|---|-------------|-------------------------|
| 1  | T[°C]       | $\Delta$ Cp [cal/T/mol] |   | T[°C]       | $\Delta$ Cp [cal/T/mol] |   | T[°C]       | $\Delta$ Cp [cal/T/mol] |
| 2  | T_f = 100°C | SR = 1 °C/min           |   | T_f = 100°C | SR = 1 °C/min           |   | T_f = 100°C | SR = 1 °C/min           |
| 3  |             |                         |   |             | *reheating              |   |             |                         |
| 4  |             |                         |   |             |                         |   |             |                         |
| 5  | 22.72175    | -5136.04286             |   | 23.03256    | -3956.72633             |   | 22.73577    | -2520.38295             |
| 6  | 23.22175    | -5136.54286             |   | 23.53256    | -3957.22633             |   | 23.23577    | -2520.88295             |
| 7  | 23.63723    | -5114.63284             |   | 23.95321    | -3939.46829             |   | 23.65424    | -2550.87649             |
| 8  | 24.05092    | -5093.63457             |   | 24.37569    | -3917.76531             |   | 24.071      | -2573.19474             |
| 9  | 24.46531    | -5070.36756             |   | 24.79185    | -3891.41426             |   | 24.48805    | -2585.20757             |
| 10 | 24.88281    | -5046.73983             |   | 25.20759    | -3866.29194             |   | 24.90563    | -2597.05656             |
| 11 | 25.30386    | -5024.20231             |   | 25.62324    | -3849.38603             |   | 25.32076    | -2609.91039             |
| 12 | 25.71978    | -4999.98656             |   | 26.03792    | -3833.62782             |   | 25.7354     | -2623.53936             |
| 13 | 26.13713    | -4979.91142             |   | 26.45473    | -3816.85382             |   | 26.15272    | -2630.27712             |
| 14 | 26.55433    | -4958.31122             |   | 26.8709     | -3805.95718             |   | 26.569      | -2645.15783             |
| 15 | 26.97031    | -4938.96101             |   | 27.2871     | -3792.54053             |   | 26.98477    | -2659.69588             |

Figure 2.9: DSC output data example.

Metadata about measurements which are in columns B, C, E, F, H and I from row number 1 to 3 have been manually entered into the spreadsheet. The name of each column is specified in the first row. Units are specified right beside them inside square brackets. The second row holds information about final temperature and scan rate values. The third row specifies if the corresponding pair of columns is from a second (reheating) run. In this case, all final temperature values are 100 °C, scan rate values are 1 °C per minute and the second column pair (E and F) represents the reheating run of a first column pair (B and C). These metadata are not mandatory but are recommended because they will make future referencing a lot easier.

Figure 2.10 shows us how the plots from example output data file look like.

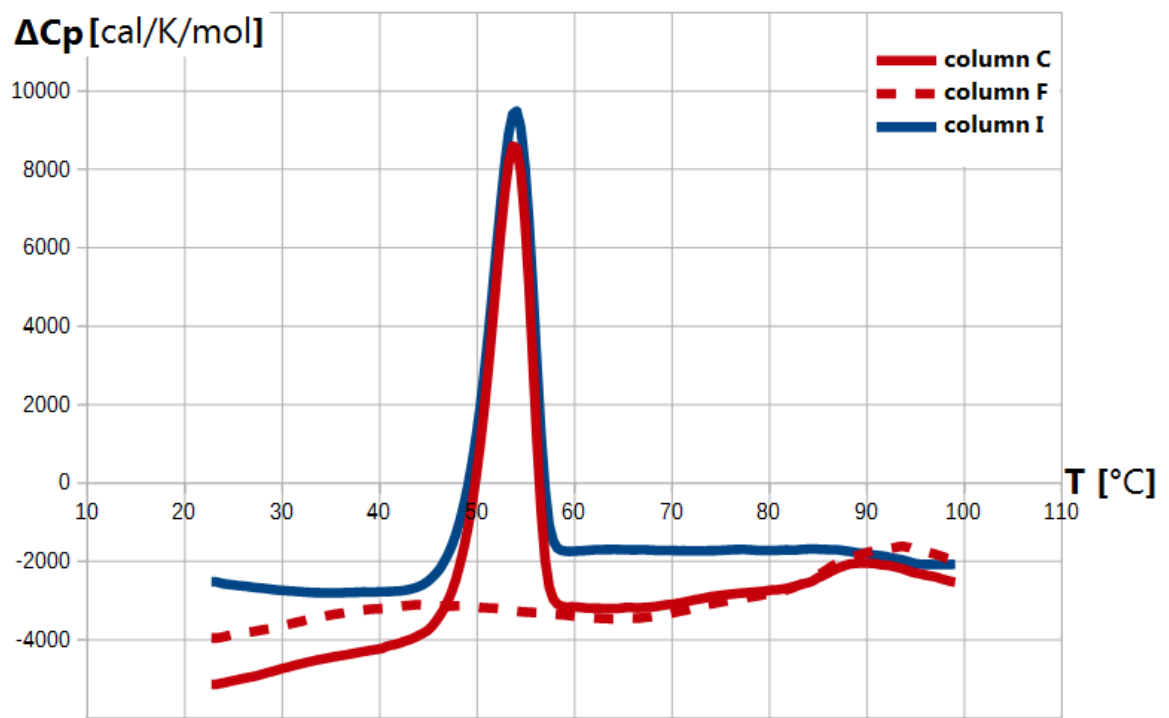


Figure 2.10: DSC output data example plot.

## Chapter 3

# Introduction to CalFitter

CalFitter is a software developed for analysis of calorimetric protein melting curves obtained from DSC measurements. It is developed as a web application. The back end (mathematical calculations) are implemented in Matlab. The graphical user interface has a form of a web page so no installation is required.

CalFitter has been designed to accept data formatted in a spreadsheet format as described in the previous section. Raw data from DSC measurements are displayed in a graph form in a similar way as it is shown in the Figure 2.10. The software offers to the user to select the type of denaturation which occurred during the DSC experiment and which should be modeled by the software. A user can then set values for different parameters which should describe the whole process of protein denaturation and therefore model as closely as possible the raw data, i.e. parameters' values are used to reconstruct the heat capacity measurements during the DSC experiment. Type of parameters and their number is defined by the specified type of observed denaturation. After the parameters' values have been set, CalFitter will display newly modeled data. A user can then adjust their values as he wishes in order to get the most precise modeled data as a result. Good evaluation of the modeled data is their visual representation. Similarities between plots of modeled data and raw data obtained from DSC experiment describe if the set parameters are precise enough, i.e. the more the plot of modeled data resembles the raw data plot, the higher the precision is. Important statistics information also displayed to give a numerical evaluation of modeled data. Such information is a sum of squared residuals, degrees of freedom, Akaike, and Bayesian information criteria. How those parameters' values can be estimated as described in the next chapter of this thesis. Up until now, it was required by the user to enter those values manually. The result of this thesis is set of programs which will calculate necessary values automatically based on the raw data obtained from DSC measurements. These programs will be integrated into the CalFitter software, which will automate the whole procedure.

When modeled data closely resemble original data, the fitting procedure can be initiated. Parameters' values are then improved as much as possible by performing fitting modeled data to the original data in the specified number of iteration. Usually, the higher the number of iterations should improve the fitting procedure.

My contribution to the development of the software prior to this thesis was a prototype of the web-based graphical user interface and the prototype of a protocol for communication

between the front and the back end. The authors and co-authors of this software (including me) are Mazurenko, S., Stourac, J., Kunka, A., Nedeljkovic, S., Bednar, D., Prokop, Z. and Damborsky, J.

### 3.1 Denaturation Models Notation

#### 1 step models:

- $N = D$ : single step reversible
- $N \rightarrow D$ : single step irreversible

#### 2 step models:

- $N = I = D$ : both steps are reversible
- $N = I \rightarrow D$ : first step is reversible and second step is irreversible
- $N \rightarrow I \rightarrow D$ : both steps are irreversible

#### 3 step models:

- $N = I_1 = I_2 \rightarrow D$ : first and second steps are reversible and third step is irreversible
- $N = I_1 \rightarrow I_2 \rightarrow D$ : first is reversible and second and third steps are irreversible
- $N \rightarrow I_1 \rightarrow I_2 \rightarrow D$ : all three steps are irreversible

Here **N** indicates natured state. **D** indicates denatured state. **I** indicates intermediate step in multi step denaturation. Symbol  $=$  indicates reversible step and symbol  $\rightarrow$  indicates irreversible step.

### 3.2 Use Case

A user should first upload the raw data obtained from DSC measurement process. He should then specify the units in which data are represented. After that modeling and fitting procedure can start. Raw data will be visualized in form of a plot (heat capacity vs. temperature) as shown in Figure 3.1. In the mentioned example figure there are 6 datasets displayed, each of which is represented by different color. Each dataset consists of main (represented by solid line) and reheating run (represented by dashed line). A user can exclude some datasets if he thinks they will not contribute to the modeling. For example, there can be a lot of noise in some of the datasets, indicating that DSC measurement hasn't been done correctly.

Next step should be selecting the correct type of denaturation. The user could determine the type of denaturation by visually examining the raw data. In this example, there is only one peak and it is present only in the first (main) run, which may indicate that there has been irreversible denaturation. After selecting the 1 step irreversible model user is asked to fill in parameters which define that mode as shown in Figure 3.2b. Modeled data will be automatically calculated and displayed on the same graph (represented by the thick

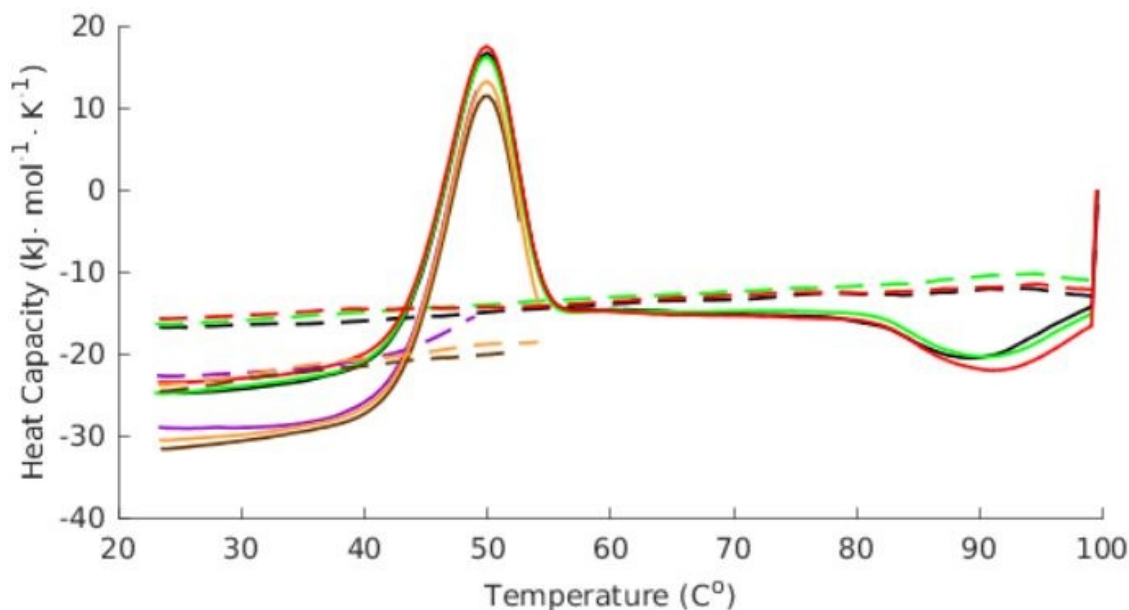
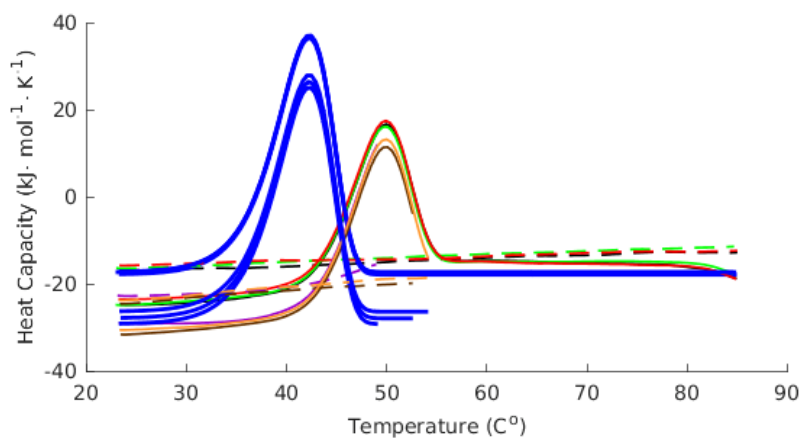


Figure 3.1: Raw DSC data visualized by CalFitter.

solid blue line) as raw data as shown in Figure 3.2a. Based on the graphs in this example, we can observe that modeling parameters are not very precise since modeled data doesn't quite resemble original raw data. Modeled peaks are bigger and their position is shifted in comparison to the original ones.



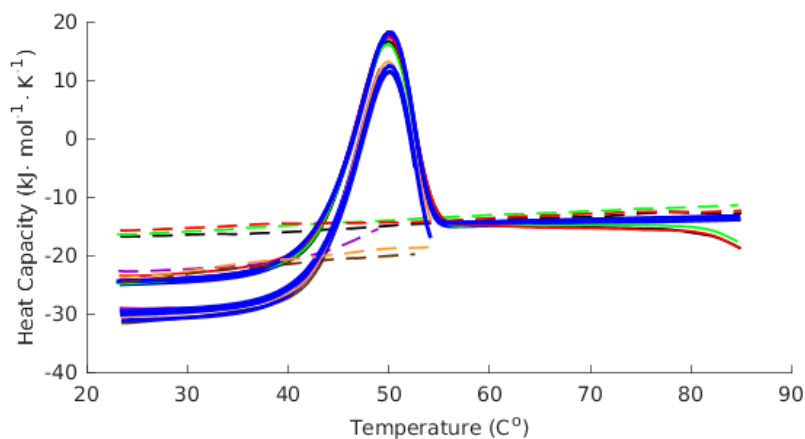
(a) Modeled and original data plots.

| N -> D                |       |         |                          |
|-----------------------|-------|---------|--------------------------|
|                       | value | CI 95 % | fixed                    |
| $E_{act}$             | 300   | N/A     | <input type="checkbox"/> |
| $T_{act}$             | 57    | N/A     | <input type="checkbox"/> |
| $\Delta H^\ddagger$   | 400   | N/A     | <input type="checkbox"/> |
| $\Delta C_p$          | 0     | N/A     | <input type="checkbox"/> |
| Additional parameters |       |         |                          |
|                       | value | CI 95 % | fixed                    |
| Slope                 | 0     | N/A     | <input type="checkbox"/> |

(b) Modeled parameters window.

After adjusting modeling parameters as shown in Figure 3.3b we can see that modeled and original raw data plots are pretty much the same, which is shown in Figure 3.3a.

Fitting procedure can be initiated after the user is satisfied with the precision of modeling parameters. CalFitter will then perform curve fitting based on the specified modeling parameters, and thus further improve them. In CalFitter fitting is implemented as an iterative process, where the number of iterations can be specified manually by the user. Success and time duration of fitting is influenced by the number of iterations but also by the preci-



(a) Modeled and original data plots.

| N -> D                |        |         |                          |
|-----------------------|--------|---------|--------------------------|
|                       | value  | CI 95 % | fixed                    |
| E <sub>act.</sub>     | 423.13 | ± 4.14  | <input type="checkbox"/> |
| T <sub>act.</sub>     | 62.23  | ± 0.2   | <input type="checkbox"/> |
| ΔH <sup>‡</sup>       | 408    | ± 10.35 | <input type="checkbox"/> |
| ΔC <sub>p</sub>       | 8.99   | ± 0.35  | <input type="checkbox"/> |
| Additional parameters |        |         |                          |
|                       | value  | CI 95 % | fixed                    |
| Slope                 | 0.02   | ± 0.01  | <input type="checkbox"/> |

(b) Modeled parameters window.

sion of modeling, that is why it's recommended to enter modeling parameters which are as precise as possible. The user can of course experiment with parameters' values, but since those values can be estimated automatically with pretty high precision, the goal of this thesis is to automate that process. Next chapter explains how it is done, and also how those parameters correlate with the denaturation.

## Chapter 4

# Modeling Parameters Calculation

The main advantage of DSC is that it helps us characterize proteins by observing how heat affects its thermodynamic stability[20]. These measurements give us complete energy signatures of protein unfolding, which can be studied and by performing mathematical modeling and curve fitting help us calculate protein's unfolding mechanisms[28].

Modeling, in general, is the process of applying specific mathematical equations on a set of parameters and input data in order to approximate or model some signal which is also the function of the same input data. In this thesis mathematical modeling is applied to temperature readings from DSC by using specific parameters so it can recreate the original heat capacity readings. Such modeled data are only approximations and are almost never a perfect representation of the original. That is why the process of curve fitting complements the modeling process. Curve fitting has a similar meaning to data modeling. In general, it means fitting equations which describe some signal (usually in form of a curve) to the raw data. In this thesis, we use curve fitting to improve modeling parameters so that the modeled curve is as close as possible to the original data from DSC.

### 4.1 Thermodynamic Analysis

In previous chapters, it was described that protein can undergo a transition between biologically active conformation, also known as the native state (N) and inactive conformation, commonly known as the denatured state (D). During the process of transition, we say that protein is unfolding. The transition can be caused by increasing temperature outside some specific range. Figure 4.1 shows how an example sigmoidal function which describes protein transition of states. At temperatures below 300 K protein is mostly in its native (N) state. At temperatures between 300 and 340 K it undergoes a transition and at temperatures above 340 K protein is mostly denatured (D). Lengths of blue and black arrows indicate the percentage of N and D states respectively[20].

In order to correctly estimate modeling parameters, we need to fully understand modeling equations. Protein unfolding can be explained as a series of steps, during which protein undergoes the transition of states. Equation 4.1 describes 1 step 2 state reversible protein unfolding process. Protein goes from state N to state D in one step defined by an equilibrium constant  $K$ .  $K$  is the equilibrium position between states, i.e. it describes relative

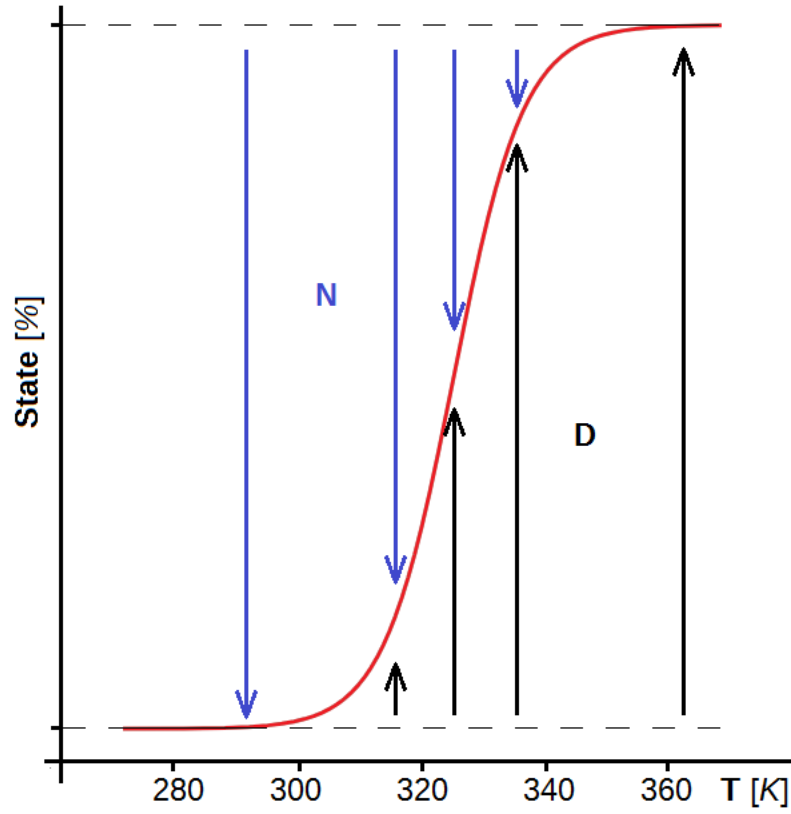


Figure 4.1: State change of a protein.

concentrations of N and D. It is the function of temperature[20]. For example in Figure 4.1 we can observe that equilibrium shifts from N to D as the temperature increases.



K is generally defined by equation 4.2 where  $\chi_N$  and  $\chi_D$  are relative concentrations of N and D state respectively.

$$K = \frac{\chi_N}{\chi_D} \quad (4.2)$$

Relationship between  $\chi_N$  and  $\chi_D$  is defined by equation 4.3:

$$\chi_N + \chi_D = 1 \quad (4.3)$$

In relation to Gibbs free energy, K can be derived from equation 4.4. Here R is gas constant, T is temperature and  $\Delta G$  is Gibbs free energy.

$$\Delta G = -\ln(K)RT \quad (4.4)$$



The well-known equation of Gibbs free energy is equation 4.5, where  $\Delta H$  is the change of enthalpy and  $\Delta S$  is the change of entropy,  $T$  is temperature.

$$\Delta G = \Delta H - T\Delta S \quad (4.5)$$

$\Delta H$  is defined by Kirchoff's law (equation 4.6).  $C_p$  is heat capacity obtained from DSC,  $T_N$  is the temperature below which protein is mostly in N state and  $T_D$  is the temperature above which protein is mostly in D state.

$$\Delta H = \int_{T_N}^{T_D} C_p \delta T \quad (4.6)$$

If we combine equations 4.2 and 4.4 we get equation 4.7:

$$K = \exp\left(-\frac{\Delta H}{RT} + \frac{\Delta S}{R}\right) \quad (4.7)$$

We can describe 1 step 2 state irreversible unfolding with equation 4.8. Here N and D have the same meaning as in equation 4.1.  $k$  represents the constant rate of irreversible transition and is defined by Arrhenius equation 4.11 which is described in [23].  $E$  is activation energy of irreversible step and  $T_f$  is temperature where  $k = 1$ .



$$k = \exp\left(-\frac{E_a}{R}\left(\frac{1}{T} - \frac{1}{T_f}\right)\right) \quad (4.9)$$

## 4.2 Reversible Denaturation Model

Reversible two state denaturation, described by equation 4.1 can be modeled by well known equation 4.10:

$$C_p^{modeled}(T) = B_0 + B_1T + \frac{K(T)}{K(T) + 1} \Delta C_p + \frac{K(T)}{(K(T) + 1)^2} \frac{\Delta H(T)^2}{RT^2} \quad (4.10)$$

$K(T)$  is defined by equation 4.11:

$$K(T) = \exp\left(-\frac{\Delta H}{RT}\left(1 - \frac{T}{T_m}\right) - \frac{\Delta C_p}{RT}(T - T_m - T \ln\left(\frac{T}{T_m}\right))\right) \quad (4.11)$$

as described in [28].

$\Delta H(T)$  is defined by equation 4.12:

$$\Delta H(T) = \Delta H + \Delta C_p(T - T_m) \quad (4.12)$$

as described in [28].

Now that we understand which equations are used to model single step reversible denaturation, we can estimate parameters which are used by those equations. Those parameters are:  $B_0$ ,  $B_1$ ,  $\Delta H$ ,  $\Delta C_p$  and  $T_m$ .

We can see that  $C_p^{modeled}$  is a function of temperature. This means that T is used from DSC measured and has the constant value, i.e.  $C_p^{modeled}$  is calculated for every temperature reading from DSC.

First parameters are  $B_0$  and  $B_1$ . These parameters describe the linear component of heat capacity readings from DSC. As it was described in the second chapter, during DSC measurements protein is dissolved in specific reference fluid. That means that both protein and reference fluid are heated simultaneously and thus they both contribute to heat capacity readings. Fortunately, reference fluid's heat capacity has linear character, as opposed to protein's heat capacity which has nonlinear character, similar to quadratic function. Linear component is easily identified in first few points of overall heat capacity reading. For first few temperature readings, protein doesn't start to unfold, which means that heat capacity is mostly influenced by reference fluid.  $B_0$  is the starting point and  $B_1$  is the slope value of the linear component.  $B_0$  is easily calculated as it is just the first heat capacity value. We can calculate  $B_1$  value by selecting first few (for example 5) heat capacity and corresponding temperature values. From those values, we can then construct straight line using the first-degree polynomial fitting function. Figure 4.2 illustrates an example how linear component can be constructed. We can then select any two points from the newly constructed line and calculate slope and intercept from their x and y coordinates by using well-known equations:

$$slope = \frac{y_2 - y_1}{x_2 - x_1}$$

$$intercept = y_1 - slope \cdot x_1$$

which will define the straight line as:

$$y = \text{slope} \cdot x + \text{intercept}$$

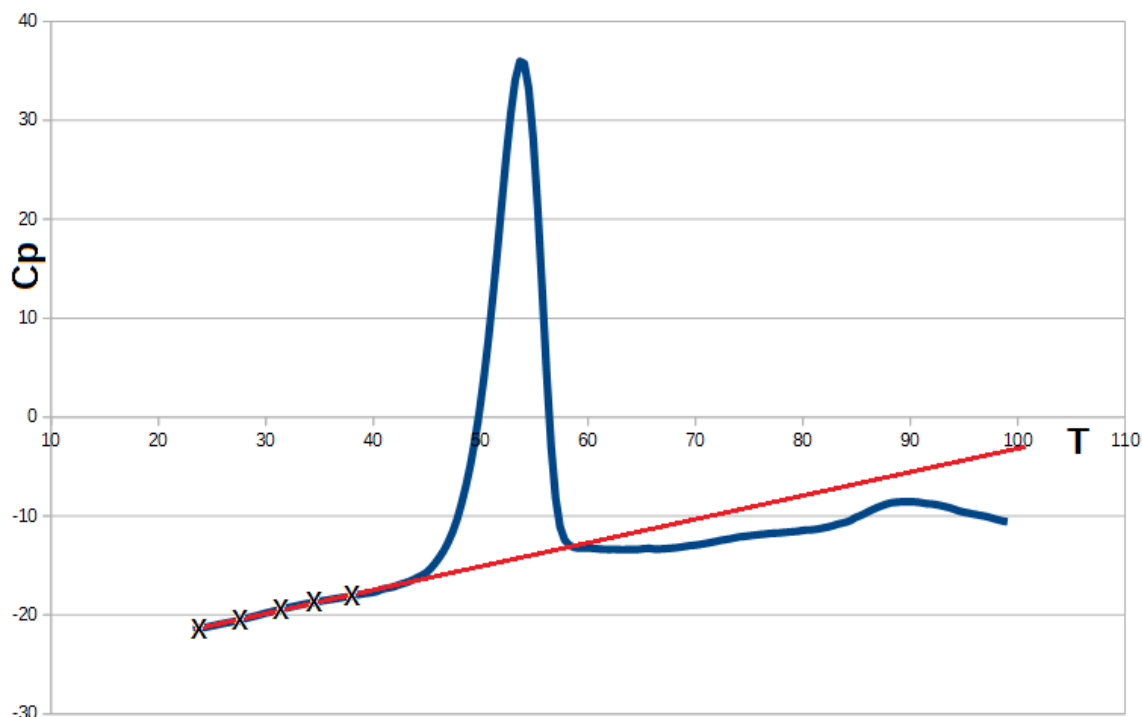


Figure 4.2: Linear component approximation.

After we calculate  $B_0$  and  $B_1$  we should subtract linear component from original heat capacity values. The result of this step is heat capacity just from observed protein. After subtraction from example in Figure 4.2 we will get graph shown in Figure 4.3. This will help us calculate next set of modeling parameters with greater precision.

Next parameter we need to calculate is  $\Delta H$  (change of enthalpy). It is mathematically defined by equation 4.6. In data obtained from DSC, it is represented as an area under a peak. In order to get its value, we need to numerically integrate heat capacity values over temperature from DSC readings.

$\Delta C_p$  (heat capacity change) can be calculated as a difference between last and first values of heat capacity from DSC readings. It is good practice to first calculate an average of first few and last few values and then perform the subtraction.

$T_m$  (peak or melting temperature) is the temperature where heat capacity has its maximal value. It can be read from DSC readings by firstly finding the maximal heat capacity value and then obtaining its corresponding temperature value.

Figure 4.4 illustrates how these parameters can be approximated from DSC readings.

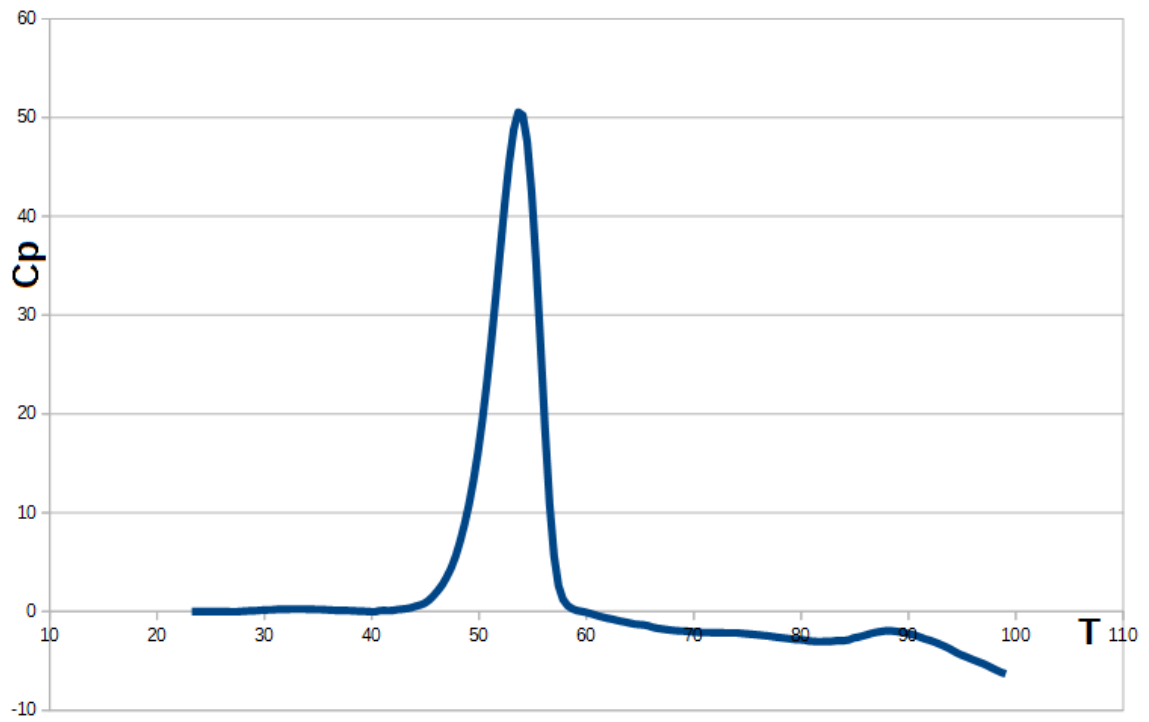


Figure 4.3: Heat capacity without linear component

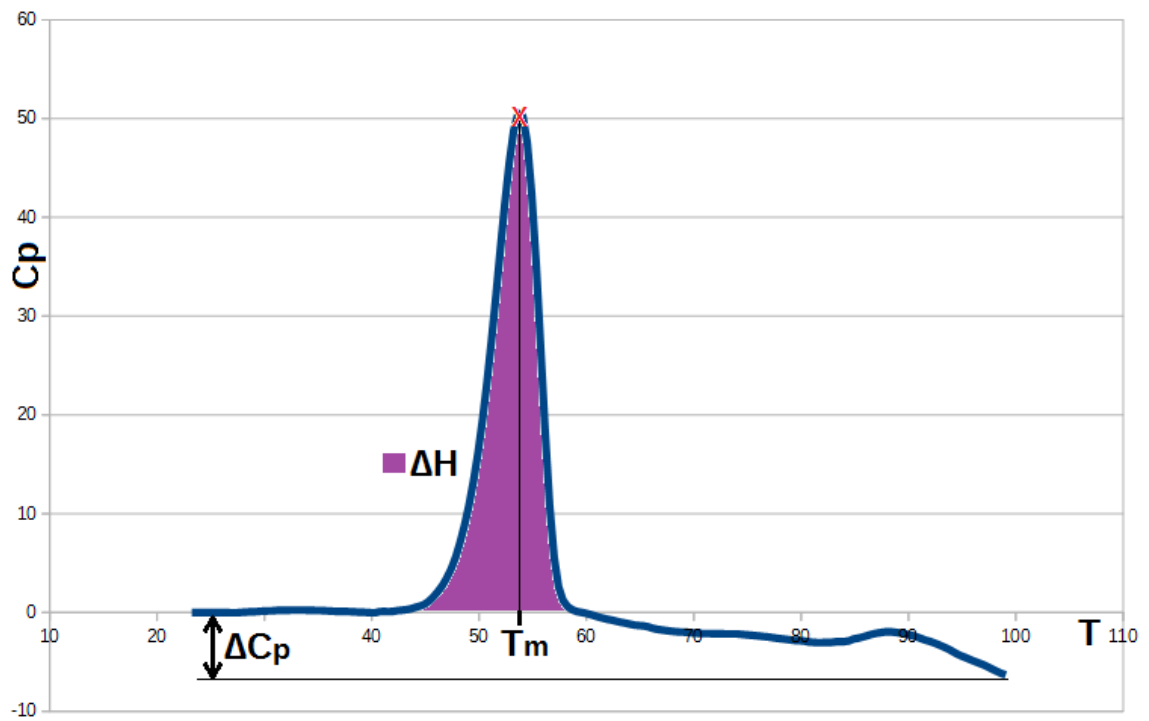


Figure 4.4: Reversible denaturation parameters' approximation.

### 4.3 Irreversible Denaturation Model

Equations described in this section have been obtained Irreversible two state denaturation, described by equation 4.8 can be modeled by well-known equation 4.13:

$$C_p^{modeled} = B_0 + B_1T + \frac{\Delta H(T)}{v}k(T)\chi_N(T) + (1 - \chi_N(T))\Delta C_p \quad (4.13)$$

as described in [23].

$v$  is the DSC scanning rate.  $k$  is defined by equation 4.14:

$$k = \exp\left(-\frac{E_a}{R}\left(\frac{1}{T} - \frac{1}{T_f}\right)\right) \quad (4.14)$$

as described in [23]

Kinetic behaviour of a model is described by differential equation 4.15:

$$\frac{\delta\chi_N}{\delta T} = -\frac{\chi_N(T)k(T)}{v} \quad (4.15)$$

as described in [23].

In the case of single step irreversible denaturation modeling we need to calculate these parameters:  $B_0$ ,  $B_1$ ,  $\Delta H$ ,  $\Delta C_p$ ,  $E_a$  and  $T_f$ . Parameters  $B_0$ ,  $B_1$ ,  $\Delta H$  and  $\Delta C_p$  are calculated exactly the same way like in case of single step reversible denaturation.

$E_a$  (energy of activation) is defined by equation 4.16.  $R$  is gas constant.  $T_m$  is melting temperature.  $C_p(T_m)$  is heat capacity value at melting temperature.

$$E_a = \frac{2.718RC_p(T_m)T_m^2}{\Delta H} \quad (4.16)$$

as described in [10].

Value of  $E_a$  can be approximated just with value of  $\Delta H$  since:

$$E_a \approx \Delta H \quad (4.17)$$

$T_f$  is derived from equation 4.18 and 4.19:

$$A = \exp\left(\frac{E}{RT_f}\right) \quad (4.18)$$

$$\ln\left(\frac{v}{T_m}\right) = \ln\left(\frac{AR}{R_a} - \frac{E}{R}\left(\frac{1}{T_m}\right)\right) \quad (4.19)$$

as described in [10].

Ultimately  $T_f$  can be calculated by using equation 4.20:

$$T_f = \frac{1}{\frac{R}{E_a} \ln\left(\frac{vE_a}{RT_m^2}\right) + \frac{1}{T_m}} \quad (4.20)$$

## 4.4 Multi Step Denaturation Modeling

An assumption was made that during one DSC measurement, protein can undergo a maximum of three denaturation steps. The reason for this is that data analysis algorithm described in this thesis can work with a maximum of 3 denaturation steps. Protein can have one of 4 structure types, but denaturation can only occur if the protein has secondary, ternary and quaternary structure. Denaturation steps can either be reversible or irreversible but only in a specific configuration. The irreversible step will never precede reversible step. For example denaturation R-R-I is possible but R-I-R is not possible (R indicates reversible step, I indicate irreversible step).

In order to model multi-step denaturation of a protein, each of the steps must be treated separately and consequently modeling parameters must be calculated separately for each step. A good indication that there has been more than one step in the presence of more than one peak in heat capacity measurements. In simple cases, each peak will be the result of 1 denaturation step. If such is the case, DSC measurements can be split into separate parts based on the number of peaks. Figure 4.5 illustrates how DSC measurements could be split. Each peak is then treated as a single step denaturation, modeling parameters are calculated and finally, the model is constructed. After all separate models have been constructed, they are added together resulting in a multi-step denaturation modeled heat capacity.

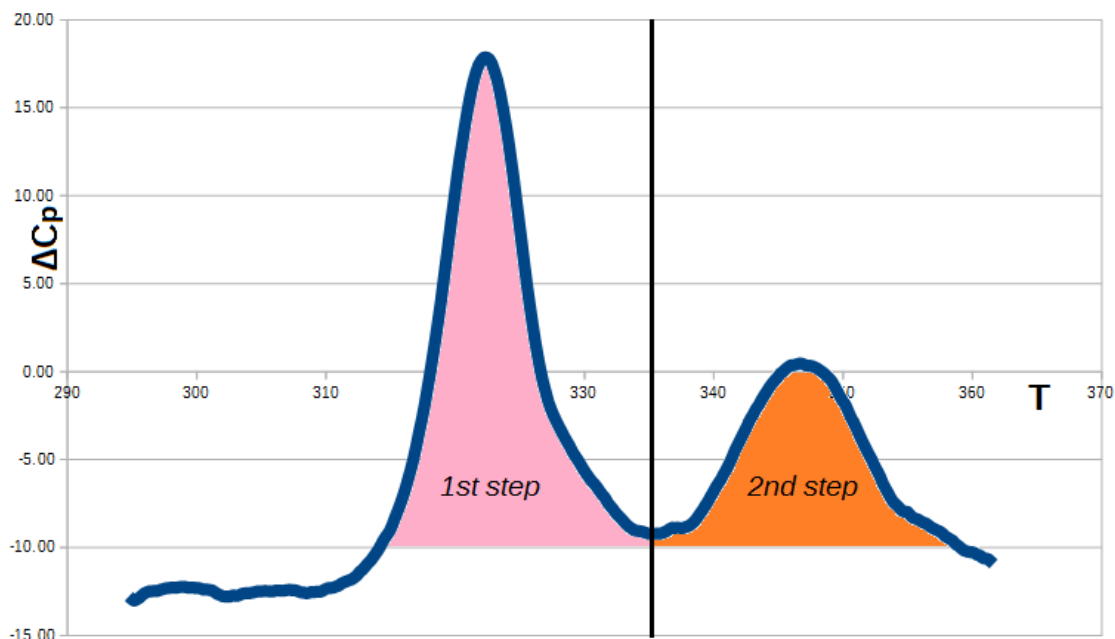


Figure 4.5: Signal splitting into separate peaks.

In some rare more complicated cases one peak can represent 2 steps of denaturation. This peak then must be treated in such a way as if it was made out of 2 smaller peaks. Each smaller peak is assumed to have its  $\Delta H$  value one half of the original peak. Maximal  $C_p$

value of each peak is also assumed to be one half of the original peak. Melting temperatures ( $T_m$ ) are assumed to spread evenly in a specific range. This range can be determined by horizontally cutting the original at the level of one half of maximal  $C_p$  value. Temperature values (x coordinates) where this cut intersects the peak define the temperature range. We can then evenly distribute  $T_m$  values of each smaller peak. All other parameters calculated based on these approximated values. Figure 4.6 illustrates how the  $T_m$  values are approximated.

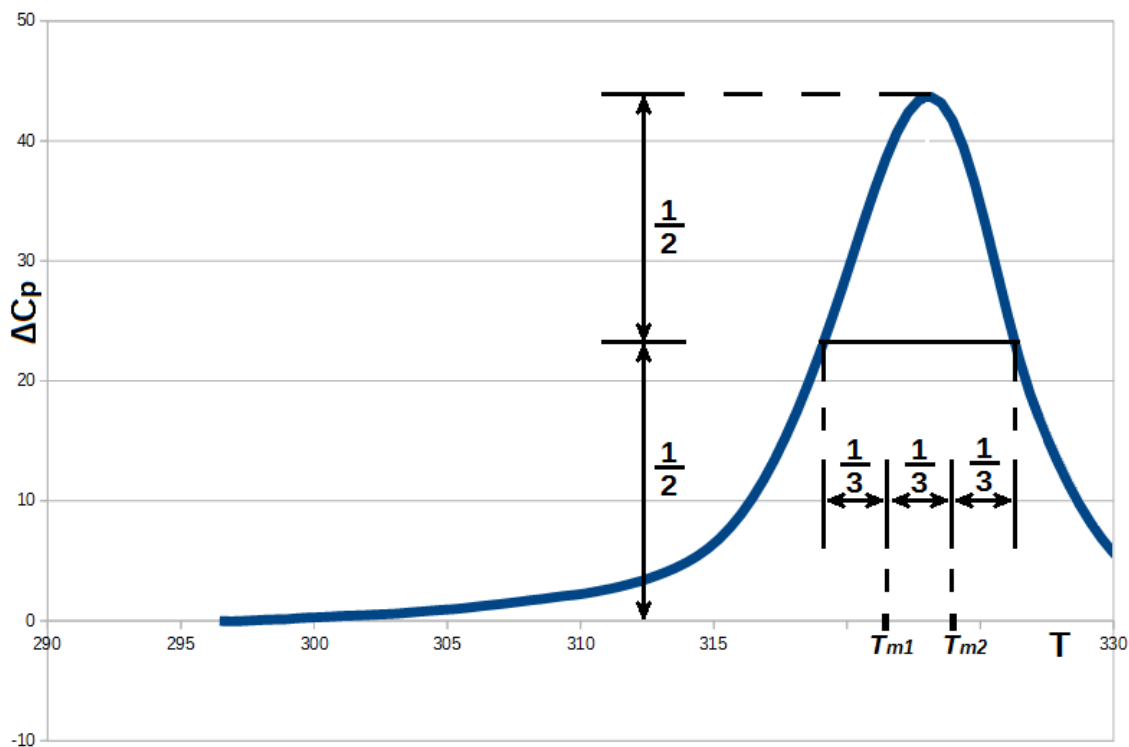


Figure 4.6: 2 step single peak  $T_m$  values approximation.

One peak can also represent 3 steps of denaturation. The process of parameter approximation is the same as previously described, except that the original peak is treated as it was made out of 3 smaller peaks.

ClaFitter is designed in such a way, that user must specify the number of steps, he thinks occurred during denaturation. The user can make this assumption by visually inspecting data obtained from DSC or he can experiment and increase or decrease the number of steps. The script for calculating initial parameters is also designed the same way. All the splitting and transformations and calculations are executed automatically.

## Chapter 5

# Implementation

The back end of the CalFitter has been programmed in Matlab. This is why the Matlab language was chosen for implementation of experimental part of this thesis. The result of this varies easy and simple integration of script for automatically calculation into existing CalFitter software. Matlab language is a good choice because it offers many necessary mathematical functions out of the box. A downside of Matlab is that it is slower than for example Python or C++ but implementing some of the functions which are offered by Matlab would be very hard and time-consuming.

All the necessary functions are written in 1 script called:

```
calculate_initial_parameters.m
```

This script contains all the necessary functions. It has several auxiliary functions and 1 interface function. This interface function is used for calling the script and it has the same name as the script file name. Interface function is declared in the following way:

```
calculate_initial_parameters(DSC, R, I)
```

It requires 2 parameters. First parameter called **DSC** is 2 column matrix. The first column must hold temperature readings and the second column must have heat capacity readings from DSC process. All values are represented as real numbers. Temperature reading must be expressed in  $K$  and heat capacity in  $kJ/mol/K$ . These units are chosen because CalFitter automatically converts input data to these units for internal manipulation. Second parameter **R** is the number of reversible steps the user thinks occurred during denaturation and final parameter **I** is the number of irreversible steps. The maximal number of steps is 3, which means that parameters **R** and **I** must satisfy expression:

$$R + I \leq 3$$

The script must be called either manually from Matlab interpreter or by another Matlab function.



## 5.1 Working principle

The script is invoked by calling its interface function, like most Matlab function. The first thing that the function does, is checking the validity of input parameters. It check's if the first parameter (**DSC**) is 2 column matrix. It also checks if the matrix has at least 10 rows, ensuring that there is enough data to work with. There is no particular significance in number 10, it can be easily changed in the code, but it is not recommended to lower this number. If there are less than 10 rows, results will probably not be correct and that is usually an indication, that DSC measurements have not been conducted correctly. Function finally checks if the second and third parameter, when added together, is less or equal than 3 since that is the upper limit of the number of steps. Parameter validation flow chart is illustrated in Figure 5.1.

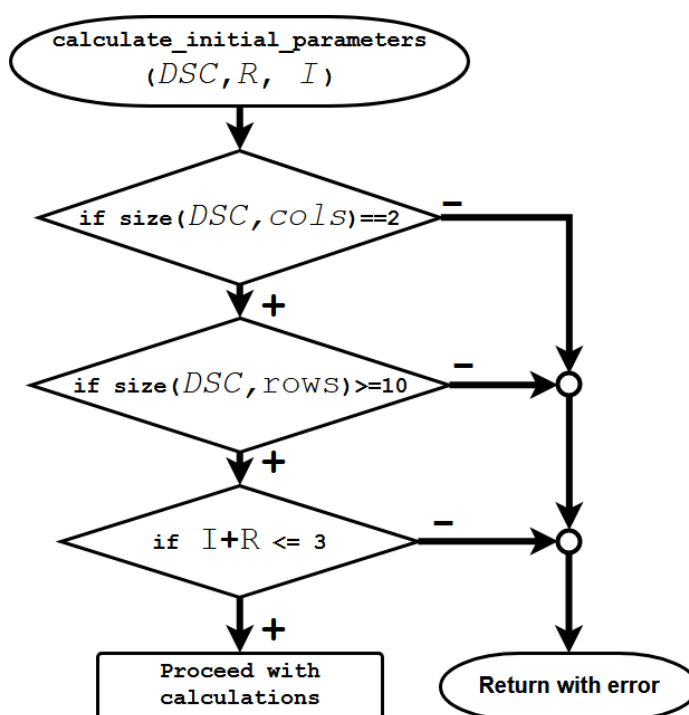


Figure 5.1: Validation of input parameters.

After the input parameters have been validated, the function then splits the first argument by its column into **X** values (temperature readings) and **Y** values (heat capacity readings) so that the later calculations can be performed more easily.

First parameters to be calculated is slope ( $B_1$  from equations 4.10 and 4.13). Due to its nature, as it was explained in the fourth chapter, it is mainly influenced by reference fluid. This means it is global and unique for the whole DSC measurement readings regardless of the number of denaturation steps nor their types. This is the reason why this parameter is calculated first and only once, as opposed to other parameters which need to be calculated for every step separately. In order to calculate slope we first need to fit first order polynomial function to the first 5 **X** and **Y** values. For this task Matlab built-in function `polyfit()`

is used. The slope is then calculated from the newly fitted straight line. This line is then subtracted from original  $\mathbf{Y}$ , thus removing the linear component from heat capacity readings.

Next step is peak detection.  $\mathbf{Y}$  values are used to identify peaks in heat capacity readings. Start and end position of each peak are extracted, so that DSC data can easily be split into separate peaks. This is done by iteratively examining each  $\mathbf{Y}$  value in combination with Matlab built-in function `findpeaks()`. After this step, we can compare the number of peaks found in DSC data and the number of steps, for which initial parameters should be calculated (I+R). If the number of peaks is the same as the number of steps, the function simply treats each peak as one step and calculates initial parameters based on step type (reversible or irreversible). If there are more peaks than specified steps, function disregards the excessive peaks and calculates initial parameters as if the number of peaks is the same as the number of steps like it was described in the previous case. If the number of steps is greater than the number of peaks, that means that 1 peak should be treated as a multi-step peak. If there are 2 peaks than the first peak is always chosen to be a multi-step peak with an exception 3 step R-R-I denaturation model, when the second peak is chosen. Experiments have shown that this approach gives best results, it doesn't have a great impact on the end result. Figure 5.2 how data is prepared for parameters calculation.

After the data has been pretreated parameters can be easily calculated. The function first calculates  $\Delta C_p$  parameter, as it is the same for both the reversible and irreversible step. Since the linear component has already been subtracted from heat capacity readings  $\Delta C_p$  is calculated as the last heat capacity ( $\mathbf{Y}$ ) value.

Next parameter is  $\Delta H$ . There are two methods how  $\Delta H$  could be calculated. First is the simple one, where the given peak's  $\mathbf{Y}$  values are numerically integrated over their corresponding  $\mathbf{X}$  values. Integration is implemented with Matlab built-in function `trapz()`. The second method is to integrate just first half of the peak and then multiply this value by 2. In the ideal case, the peak should be symmetrical. If that is the case both methods should give pretty much the same result. The problem with the first method is that it can sometimes be hard to precisely identify where in the peak has denaturation step ended, which is usually the case if the peak is not symmetrical. As  $\Delta H$  is defined as the area under the peak where denaturation occurs, If we would to integrate the whole peak, we would get a higher value. The second method somewhat fixes this problem by assuming that the peak should have been perfectly symmetrical. Neither method is perfect, as both have their own use cases. Experiments have shown that first method gives more precise results for the reversible step and the second is more precise for irreversible step.

$T_m$  is easily calculated by firstly finding the maximum value of peak's  $\mathbf{Y}$  value, finding its index and then using that index to find the corresponding  $\mathbf{X}$  value.

$E_a$  is calculated by simply solving Equation 4.16.  $T_f$  is calculated by solving Equation 4.20.

After this, all parameters have been calculated and function simply returns their values. If the peak has been chosen as a multi-step then additional calculation process is slightly

different. First,  $\Delta H$  value is divided by the number of steps. This new value will later be used for every step. Each step, however, needs to have different  $T_m$  value. Their values are evenly spread in the specific interval as it was described in the fourth chapter and illustrated in Figure 4.6. The last step's  $\Delta Cp$  value will have original peak's value, while all the previous steps will have their value set to 0. Lastly  $E_a$  and  $T_f$  are calculated for every irreversible step using previously modified  $\Delta H$  and newly calculated  $T_m$  and  $\Delta Cp$  values. Finally, the function returns calculated parameters. Flowchart of parameter calculation is shown in Figure 5.3. Before the function returns, all parameters' values are printed to the console (standard output) alongside with their names.

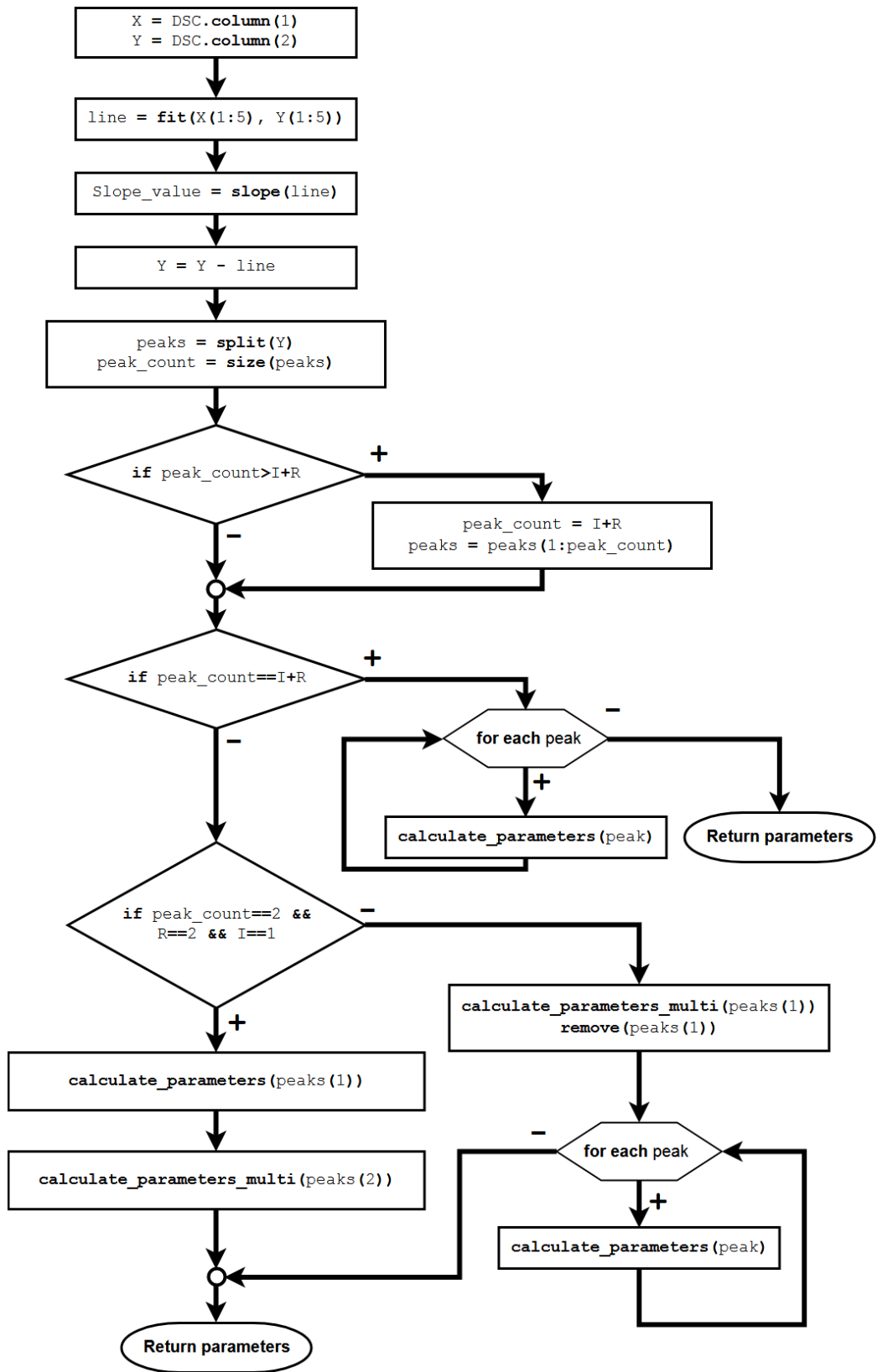


Figure 5.2: Data preparation.

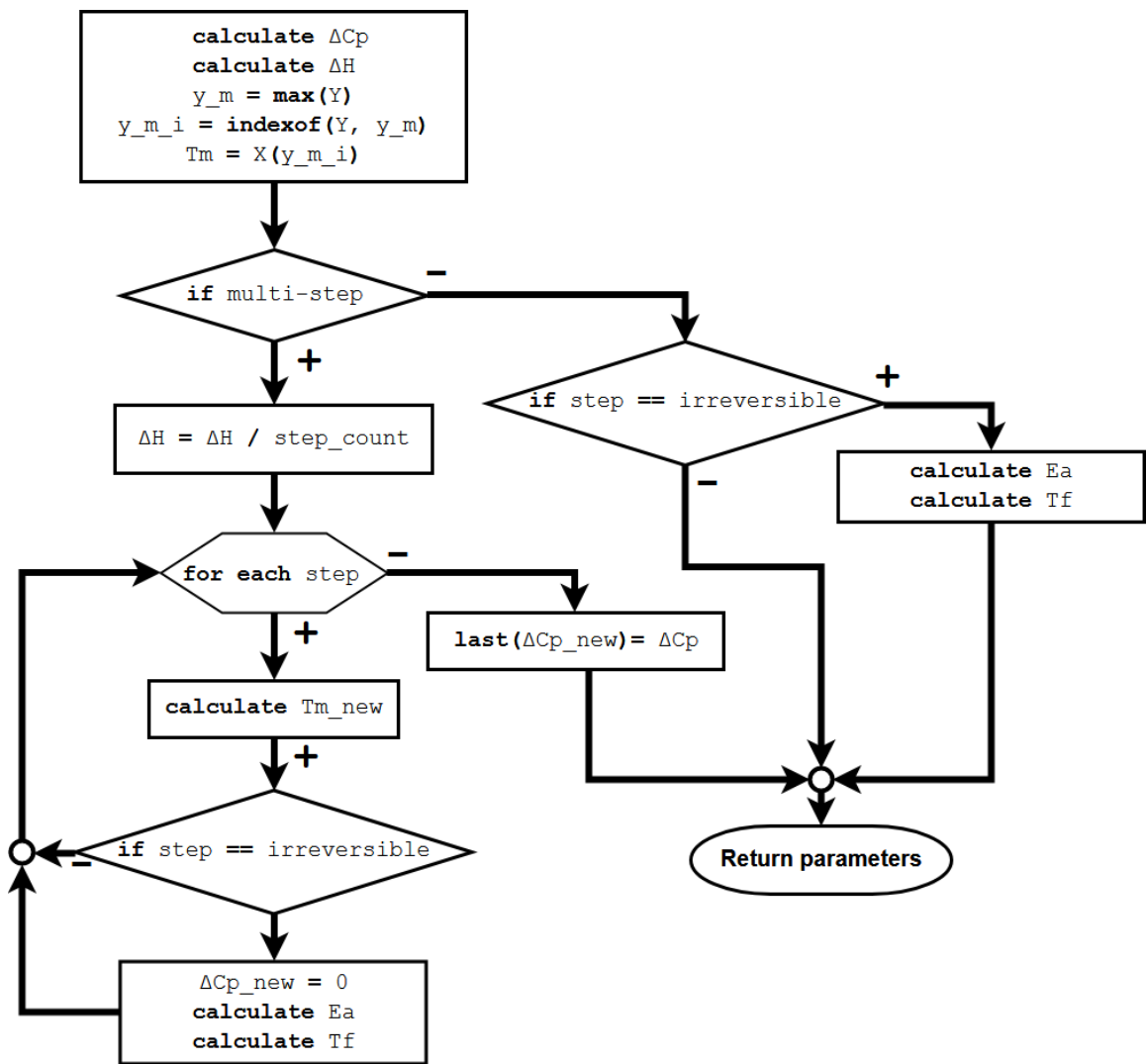


Figure 5.3: Parameter calculation.

## 5.2 Results

The script was run on file `resdatadataset_1.csv`. First 1 step reversible model has been selected ( $N = D$ ). Calculated parameters were:

$$T_m = 49.95 \quad \Delta H = 351.72 \quad \Delta C_p = 10.69 \quad \text{Slope} = -0.043$$

Figure 5.4 shows original and modeled data. Modeled data are shown in blue color, while original data are shown in black color. We can see that plots are roughly the same. Modeled peak is a bit higher and wider.

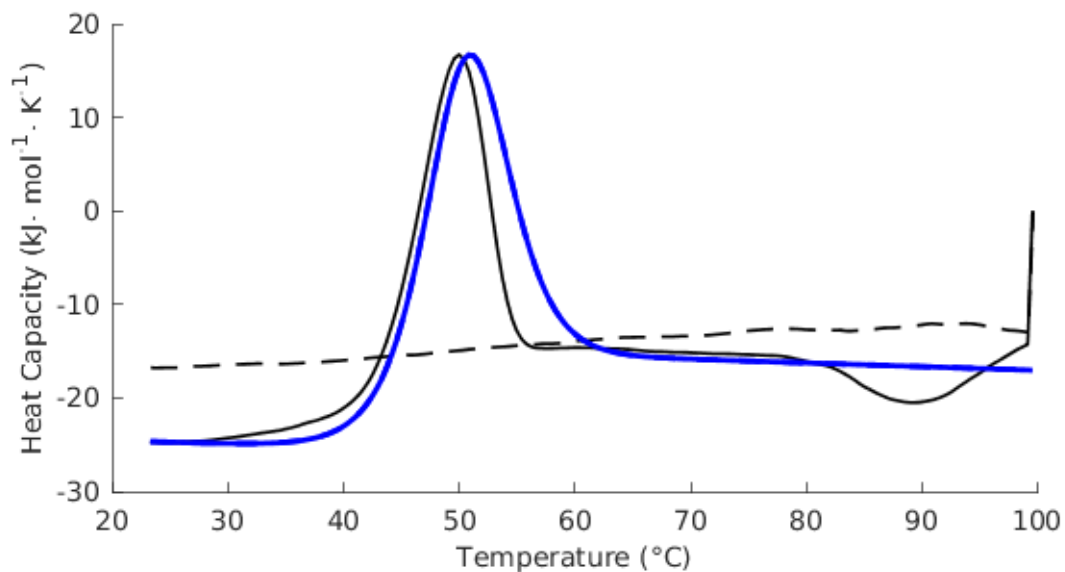


Figure 5.4: Original and modeled data plots for 1 step reversible model.

Later 1 step irreversible model was chosen ( $N \rightarrow D$ ). The same file was used. Calculated parameters were:

$$E_a = 467.63 \quad T_f = 58.92 \quad \Delta H = 448.80 \quad \Delta C_p = 10.69 \quad \text{Slope} = -0.043$$

Plots are shown in Figure 5.6. We

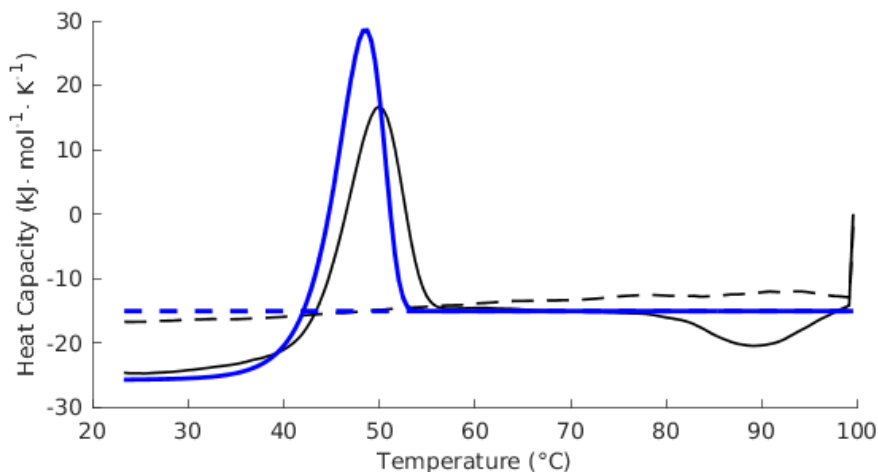


Figure 5.5: Original and modeled data plots for 1 step irreversible model.

Finally 2 step model with 1 reversible and 1 irreversible step was chosen ( $N = I \rightarrow D$ ). Calculated parameters were:

$$T_m = 48.28 \quad \Delta H = 224.40 \quad \Delta C_p = 5.34$$

$$E_a = 251.07 \quad T_f = 70.47 \quad \Delta H = 224.40 \quad \Delta C_p = 0.00 \quad \text{Slope} = -0.043$$

Plots for 2 step model with 1 reversible and 1 irreversible step are in Figure 5.6.

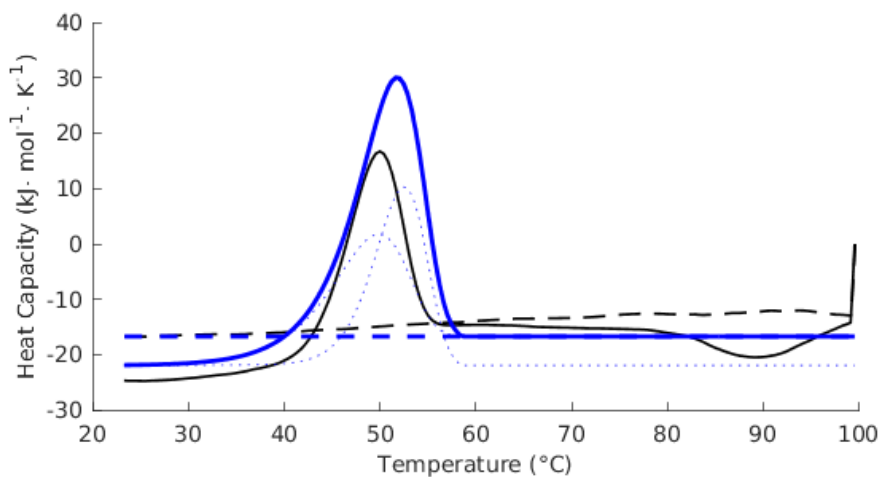


Figure 5.6: Original and modeled data plots for 1 step irreversible model.

It is apparent from the data, that the best results were for the 1 step reversible model and the worst for 2 step model.

## Chapter 6

# Conclusion

The main goal of this thesis was to contribute to the development of software CalFitter and to improve its functionality. CalFitter is a web server application that is used for thermodynamic data analysis of protein unfolding. In order to get the relevant data, protein first needs to undergo denaturation which is usually artificially caused by some experiment method. One of the best methods for this task is Different Scanning Calorimetry (DSC). This is why DSC was chosen to be a source of data for CalFitter. During DSC measurement protein is heated, which causes it to unfold. Relevant data like current temperature and heat capacity of a protein are being recorded. By analyzing these data, a lot of protein's properties like protein stability and many others can be determined.

At the time of writing this thesis, CalFitter offers the functionality of modeling heat capacity values of protein unfolding based on a set of parameters. Modeled values should be the same or as close as possible to the original data obtained from DSC experiment. The type of denaturation that occurred determines which set of parameters should be used. The simplified workflow is as follows: user uploads the data, he then chooses the type of denaturation from the ones which are supported by the CalFitter and enters set of parameters. CalFitter then uses these parameters and models heat capacity values. Finally, CalFitter could perform curve fitting on original data, after which parameters values will be significantly improved. This thesis improves CalFitter's by adding automatic initial parameters calculation functionality and thus eliminates the need for the user to enter those values by himself. Initial parameters values influence speed and performance of the fitting procedure. If parameters are off by a great value fitting procedure will either take a very long time or not be successful. If the parameters are correct from the start than curve fitting is not needed.

Initial parameters are calculated by analyzing DSC data. During one DSC experiment protein may undergo several steps of denaturation. The result of this thesis is a script that can calculate all the necessary parameters for every step of denaturation with great success. The script can calculate initial modeling parameters for 1, 2 or 3 step denaturation model, where step can either be reversible or irreversible. Experiments have shown that the precision of results decreases with increasing number of steps. The reason for this is that in some cases for multi-step denaturation models certain assumptions have to be made which may not be completely correct. Script also has greater success with reversible steps than with the irreversible steps, because reversible steps are easier to model and analyze.



There is a great potential for future improvements. For example, initial parameters calculation and fitting procedure could be joined together. New types of denaturation models can also be added. Nevertheless, CalFitter offers a wide range of models and it can be used as a professional tool. CalFitter has been published as a scientific article in the Nucleic Acids Research journal [29]. This thesis has been made in collaboration with Loschmidt Laboratories of Masaryk University.

# Bibliography

- [1] *Differential Scanning Calorimetry (DSC) Thermal Analysis*. Anderson Materials Evaluation, Inc.  
Retrieved from: <http://www.andersonmaterials.com/dsc.html>
- [2] Structure of Protein.  
Retrieved from:  
<https://drgpinstitute.wordpress.com/tag/%CE%B2-pleated-sheets/>
- [3] Tertiary Structure of Proteins - I.  
Retrieved from: <http://chemistry.umeche.maine.edu/MAT500/Proteins8.html>
- [4] Arthur M. Lesk : *Introduction to Protein Architecture: The Structural Biology of Proteins*. Oxford University Press. 2000.
- [5] Carl Branden, John Tooze: *Introduction to Protein Structure*. Garland Publishing. 1999.
- [6] Gloria A. Di Lullo, Shawn M. Sweeney, Jarmo Körkkö, Leena Ala-Kokko, James D. San Antonio : *Mapping the Ligand-binding Sites and Disease-associated Mutations on the Most Abundant Protein in the Human, Type I Collagen*. The Journal of Biological Chemistry. 2001.
- [7] Kenneth P. Murphy: *Protein Structure, Stability, and Folding. Methods in Molecular Biology*. Humana Press. 2001.
- [8] Thomas S. Argyris: *Keratins. Their Composition, Structure and Biosynthesis*. R. D. B. Fraser , T. P. MacRae , G. E. Rogers. The Quarterly Review of Biology 48, no. 2. 2005.
- [9] Alan Fersht: *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. World Scientific. 2017.
- [10] Alfonso Arroyo-Reyna; Salvador R. Tello-Solís; Arturo Rojo-Domínguez: *Stability parameters for one-step mechanism of irreversible protein denaturation: a method based on nonlinear regression of calorimetric peaks with nonzero DCp*. Analytical Biochemistry. 2004.
- [11] Anthony W. Norman, Gerald Litwack: *Hormones*. Academic Press. 1997.
- [12] Arturo, E.: A Step Toward Solving PKU.  
Retrieved from: <http://exelmagazine.org/article/solving-pku/>

- [13] Bailey, R.: Learn About the 4 Types of Protein Structure. 2017.  
Retrieved from: [thoughtco.com/protein-structure-373563](http://thoughtco.com/protein-structure-373563)
- [14] Bret A. Shirley: *Protein Stability and Folding, Theory and Practice, Methods in Molecular Biology*. Humana Press. 1995.
- [15] Campbell, N. A.: *Biology, 8th Edition*. Pearson Benjamin Cummings. 2018.
- [16] Charles H. Spink: *Methods in cell biology: Differential Scanning Calorimetry*. Elsevier BV. 2010.
- [17] Christopher M. Johnson: *Differential scanning calorimetry as a tool for protein folding and stability*. Archives of Biochemistry and Biophysics. 2013.
- [18] CNX, O.: OpenStax, Biology. 2018.  
Retrieved from:  
<http://cnx.org/contents/185cbf87-c72e-48f5-b51e-f14f21b5eabd@11.1>
- [19] Donald Voet, Judith G. Voet: *Biochemistry, 4th Edition*. Wiley. 2011.
- [20] Dr Chris M. Johnson: *Differential Scanning Calorimetry: Theory and practice*. Malvern Instruments Limited. 2014.
- [21] L. Mauer : *Encyclopedia of Food Sciences and Nutrition*. Purdue University. 2003.
- [22] Lars Konermann: *Protein Unfolding and Denaturants*. Wiley. 2012.
- [23] Lyubarev AE, Kurganov BI: *Modeling of Irreversible Thermal Protein Denaturation at Varying Temperature*. Bach Institute of Biochemistry, Russian Academy of Sciences. 1999.
- [24] M. Michael Gromiha: *Proteins: Structure, Function and Bioinformatics*. Elsevier, A Division of Reed Elsevier India, Pvt. Ltd.. 2010.
- [25] Markus J. Tamás, Sandeep K. Sharma, Sebastian Ibstedt, Therese Jacobson, Philipp Christen: *Heavy Metals and Metalloids As a Cause for Protein Misfolding and Aggregation*. Biomolecules. 2014.
- [26] Ophardt, C. E.: Virtual Chembook. 2003.  
Retrieved from: <http://chemistry.elmhurst.edu/vchembook/index.html>
- [27] Rajani, R.: Denaturing Proteins.  
Retrieved from: <https://cdn.thinglink.me/api/image/2923mMxKEVjjfxz3721EkhTqx6bZsySRbr16BQjNdroRxFffFR6DpfdJG12NpGPFTPBF5eVfSH1nazeWuk5320/320/scaledown>
- [28] Stanislav Mazurenko, Antonin Kunka, Koen Beerens, Christopher M. Johnson, Jiri Damborsky, Zbynek Prokop: *Exploration of Protein Unfolding by Modelling Calorimetry Data from Reheating*. Scientific Reports volume 7. 2014.
- [29] Stanislav Mazurenko, Jan Stourac, Antonin Kunka, Sava Nedeljković, David Bednar, Zbynek Prokop, Jiri Damborsky: *CalFitter: a web server for analysis of protein thermal denaturation data*. Nucleic Acids Research. 2018.

- [30] Suzanne M. Mithieux, Anthony S. Weiss: *Fibrous Proteins: Coiled-Coils, Collagen and Elastomers*. Advances in Protein Chemistry. 2005.
- [31] Timberlake, K.: *Introduction to Chemistry: General, Organic, and Biological*. 1999.
- [32] Yoshinori Mine, Tatsushi Noutomi, Noriyuki Haga: *Thermally induced changes in egg white proteins*. J. Agric. Food Chem. 1990.

# Appendix A

## User Manual

At the time of writing this thesis, initial parameters calculations are not integrated into CalFitter. This is why user manual is provided. It demonstrates how to use the result script in combination with CalFitter. In order for the script to work, Matlab must be installed on a local computer. CalFitter is implemented as a web server so internet connection is also required.

### Disclaimer

This user manual doesn't demonstrate every functionality of CalFitter. It just demonstrates how to use initial parameters calculation script in combination with CalFitter manually. For a complete CalFitter user manual, please visit:

```
loschmidt.chemi.muni.cz/calfitter/?action=help&
```

or

```
loschmidt.chemi.muni.cz/calfitter/?action=example&
```

### A.1 Input Data Format

DSC data should be stored inside **.csv** file. The file should have only 2 columns. The first column should have temperature values expressed in **K** and the second column should have heat capacity values expressed in **KJ/mol/K**.





The first cell in the first column should have id number of DSC measurements. The first cell in the second column should also have that same value (DSC id number). For the sake of simplicity, both values can be set to 1. The second cell in the first column should hold final temperature value which was obtained from the DSC measurements expressed in **K**. The second cell in the second column should have the value of the used scan rate during DSC measurements expressed in **K/min**. The rest of the cells in the first column should have temperature values from DSC measurements expressed in **K**. The rest of the cells in the second column should have heat capacity values expressed in **KJ/mol/K**. Figure [A.1a](#) shows a preview of such file. Sample files can also be located on the provided DVD under the path **/res/data/**

## A.2 Data Upload and Model Selection

Follow the instructions on:

`loschmidt.chemi.muni.cz/calfitter/?action=help&`

on how to upload the .csv file. After the file has been uploaded, data should be plotted in the graph area. In the top right corner of the CalFitter application, select desired model. Note that only some of the models are supported by the initial calculation scripts. Supported models are mentioned in the section 3.1. After the desired model has been selected, take note how many reversible and how many irreversible steps does the selected model consists of (see 3.1 for reference). In Figure A.1b supported models are marked in red.

|    | B  | C   | D |   |
|----|--|---|---|---|
| 1  |  1      |  1 |   | 1-step<br>▶ N -> D<br>▶ N = D               |
| 2  |  373.15 |  1 |   | ▶ N = D (Van't Hoff's)<br>N <=> D           |
| 3  | 296.41   | -24.65  |   | 2-step                                      |
| 4  | 296.83   | -24.71  |   | ▶ N -> I -> D<br>▶ N = I -> D               |
| 5  | 297.25   | -24.72  |   | N <=> I -> D                                |
| 6  | 297.67   | -24.72  |   | ▶ N = I = D<br>N -> I1 -> D; N -> I2 -> D   |
| 7  | 298.08   | -24.73  |   | 3-step                                      |
| 8  | 298.5  | -24.75  |   | ▶ N -> I1 -> I2 -> D<br>▶ N = I1 -> I2 -> D |
| 9  | 298.91   | -24.72  |   | N <=> I1 -> I2 -> D                         |
| 10 | 299.33   | -24.68  |   | ▶ N = I1 = I2 -> D                          |
| 11 | 299.75   | -24.65  |   |   |

(a) .csv file preview. Red arrows points to DSC id numbers. Blue arrow points to final temperature. Green arrow points to scan rate. (b) Model selection preview. Supported ones are marked with a red arrow.

## A.3 Run Initial Calculation Script

First copy the previously uploaded .csv file to the current directory. Then copy `calculate_initial_parameters.m` script to the same directory. Then Matlab interpreter. From the interpreter run the following command:

```
» DSC = csvread('dsc_file.csv');
```

where `dsc_file.csv` is the name of the .csv that has just been uploaded. Note that the path needs to be inside single quotation marks (`'`). Then run the following command:

```
» calculate_initial_parameters(DSC, R, I);
```

where **R** is the number of reversible steps and **I** number of irreversible steps of the selected denaturation model. The script will print initial parameters values alongside their names to the standard output of the Matlab interpreter. Figure A.2 show how commands should be executed and what the output can look like. In case some error occurs, an error message will be printed out instead of parameters. Finally, copy parameters' values to the model window inside CalFitter. Modeling data should automatically be plotted in the graph area. Additionally, the fitting can be executed by pressing the **Fit** button, which should improve initial parameters. Visual representation of how precise parameters are is the similarity between modeled data and original data plot. See Figures 3.2a, 3.2b, 3.3a and 3.3b for reference.

```
>> DSC = csvread('data.csv');  
>> calculate_initial_parameters(DSC, 1, 0);  
  
Tm= 49.957790; dH= 351.720600; dCp= 10.697779; dHvh= 365.340931;  
  
Slope= -0.043891;  
  
>> █
```

Figure A.2: Script's output.

# Appendix B

## DVD Contents

Path :

- *doc/*
- *doc/xnedel08.pdf*
- *doc/src/*
- *doc/src/\**
- *res/*
- *res/src/*
- *res/src/calculate\_initial\_parameters.m*
  
- *res/data/*
- *res/data/\*.csv*

Description :

- Documentation directory
- Bachelor's thesis text
  
- Source files for PDF
  
- Result of this thesis, script that calculates initial parameters
  
- .csv files containing data obtained from DSC measurements (in correct units)