UPPSALA
UNIVERSITET

UPTEC X16031
Examensarbete 30 hp
Januari 2017

Introducing quality assessment and efficient management of cellular thermal shift assay mass spectrometry data

Joakim Hellner

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Introducing quality assessment and efficient management of cellular thermal shift assay mass spectrometry data

Joakim Hellner

Recent advances in molecular biology has led to the discovery of many new potential drugs. However, difficulties with in situ analysis of ligand binding prevents quick advancement in clinical trials, which stresses the need for better direct methods. A relatively new methodology, called Cellular Thermal Shift Assay (CETSA), allows for detection of ligand binding in a cells natural environment and can be used in combination with Mass Spectrometry (MS) for readout. With help from the Pelago Bioscience team, I developed a pipeline for processing of CETSA MS data and a web based system for viewing the results. The system, called CETSA Analytics, also evaluates the results relevance and helps its users to locate information efficiently. CETSA Analytics is currently being tested by Pelago Bioscience AB as a tool for experimental data distribution.

Handledare: Dr. Daniel Martinez Molina
Ämnesgranskare: Prof. Jonas Bergquist
Examinator: Dr. Jan Andersson
ISSN: 1401-2138, UPTEC X16031

Populärvetenskaplig Sammanfattning

Läkemedelsframställning är idag en strikt kontrollerad process med höga krav på den nya substansens egenskaper. Kliniska studier är samtidigt väldigt dyra och många projekt tvingas till avslut i förtid om man har svårt att påvisa läkemedlets verkan. Ett av de mest problematiska momenten är att få fram tillräckliga bevis för att substansen har den önskade effekten i dess biologiska miljö. Analys av renat protein ger ofta en simplifierad version av verkligheten där många faktorer inte tas i beaktning, e. g. membrantransport och hjälpproteiner. Det finns därför ett starkt behov av nya direkta metoder som kan ersätta dagens alternativ, vilka ofta inkluderar dyra affinitets-prober.

Under de senaste åren har intresset stigit för en ny metod, där man detekterar en substans bindning till ett protein genom att studera komplexets värmetolerans. När en ligand binder till ett protein sker förändringar i dess struktur, vilka har direkt påverkan på komplexets stabilitet. Genom att kvantifiera proteiner i ett stegvis ökande temperaturintervall kan man således särskilja proteiner vilka bundit en ligand från de som förblivit opåverkade. Denna princip utnyttjas i metoden, vilken har namngivits Cellular Thermal Shift Assay (CETSA). Metodiken kan även utföras i samband med mass spectrometri (MS) under detektionsfasen, vilket tillåter storskaliga studier av hela proteom.

CETSA MS producerar stora dataset som ofta motsvarar närmare fem tusen proteiner. Utan tillräcklig teoretisk bakgrund, både inom dataanalys och proteinbiologi, kan resultatet vara svårtolkat och tidskrävande att gå igenom. Av denna anledning har jag, i sammarbete med Pelago Bioscience AB, utvecklat ett arbetsflöde för automatiserad analys som även utvärderar datakvalitén samt indikationer på ligandbinding. Detta möjliggör för rankning av resultatet, vilket effektiviserar tolkningsprocessen. För att underlätta åtkomsten av resultatet och slippa problematiken med olika plattformar, utvecklades även ett webbaserat system vid namn CETSA Analytics. CETSA Analytics lagrar all experimentell data ner till peptidnivå och hjälper användaren att utvärdera sitt resultat i ett användarvänligt gränssnitt.

Abbrevations

TE	Target Engagement
CETSA	Cellular Thermal Shift Assay
ITDR	Isothermal Dose Response
MS	Mass Spectrometry
TMT	Tandem Mass Tags
iTRAQ	isobaric Tags for Relative and Absolute Quantification
CSV	Comma Separated Values
SQL	Structured Query Language
PHP	Hypertext Preprocessor
HTML	HyperText Markup Language
CSS	Cascading Style Sheets

Contents

Chapter 1

Introduction

Advances in drug discovery has led us to many innovative therapies over the last couple of years. However, clinical trials have relatively low success rate due to difficulties with *in situ* analysis, which stressed the need for more direct methods that can work as an alternative to the expensive affinity probes [1].

One of the most prominent challenges in drug discovery is to ensure that the compound binds to its cognate target protein with sufficient affinity and specificity, a process commonly referred to as Target Engagement (TE). Interactions with proteins other than the intended target, so called off-targets, may potentially cause undesired effects. Such effects have to be considered, but producing conclusive results of TE *in situ* has proven difficult [1][2].

In recent years a new promising methodology has become increasingly popular, called Cellular Thermal Shift Assay (CETSA), which uses a heat pulse to provoke unfolding of proteins. Proteins that have bound a compound will unfold differently from those that have not, allowing identification of TE. The method can be combined with mass spectrometry (MS) for readout, which offers proteome-wide analysis of wanted and potentially unwanted proteins [2][3].

CETSA MS produce large datasets, which is why efficient processing workflows are essential. Estimating the results relevance can also be challenging, especially without sufficient theoretical knowledge on the studied system. This report describes the development of a data management system and the formation of a CETSA MS processing pipeline. The aim was to provide easy access and distribution of data in a user friendly environment, that can help to locate and interpret experimental results.

Chapter 2

Background Theory

2.1 Biophysical Stability of Proteins

The stability of a protein is highly dependent on its conformational stucture. When a protein's tertiary structure changes, the energy of the bonds between the amino acid chains will change with it. Some bonds will break or form, and others will experience minor shifts in energy induced by the new distance. This results in an overall shift in Gibbs free energy, and thus the protein's stability [5].

When a ligand binds to its cognate target protein, it induces a conformational shift and hence an increase or decrease in stability. By measuring this change, it is possible to distinguish between proteins that bind a ligand and those that do not [4].

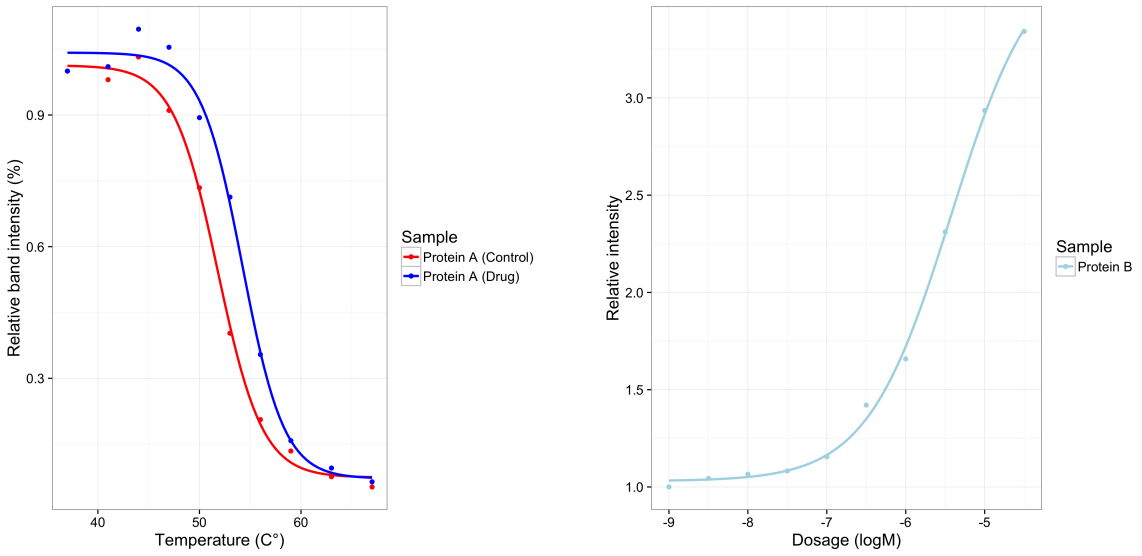
2.2 Cellular Thermal Shift Assay

Thermal shift assays are one of the most common methods used to study TE. It builds on the principle that changes in stability also affects the temperature at which the protein unfolds, and you can thus detect target engagement by comparing melting characteristics. An increased stability comes with higher resistance to heat-induced unfolding, and vice versa [6].

In 2013, an article titled *Monitoring Drug Target Engagement in Cells and Tissues Using the Cellular Thermal Shift Assay* was published in the Science journal, describing a promising new method called Cellular Thermal Shift Assay. As the name suggests, CETSA can be applied to intact cells, which allows for valuable TE analysis in the proteins biological environment. The method is currently becoming increasingly popular in a variety of different studies, since it also allows for studies of membrane transportation rates and downstream cellular events [1].

2.2.1 Experimental Setup

In a typical experiment, a treated and a control sample are aliquoted and heated to temperatures between 37-70 degrees and allowed to cool down. Soluble proteins are then separated from the precipitated fraction in a centrifugation step and quantified with western blot. The samples are plotted in consecutive order with an increasing temperature on the X-axis, showing a curve with a negative slope around the protein's melting temperature [2]. The resulting melt curve has a central role in CETSA analysis. The idea is to use the fraction of intact protein as an indicator of whether the protein has bound a ligand or not. Since the increased stability allows the complex to stay intact in higher temperature, it will show as a shift between the control and drug treated curve, see figure 2.1a.



(a) Melt curve

(b) Dose response

Figure 2.1: Example CETSA curves that indicate TE.

2.2.2 Isothermal Dose Response

In a CETSA study it is common to derive the Isothermal Dose Response (ITDR) of a protein, as complementary data to the standard melt curves. ITDR experiments follow an almost identical procedure except that, instead of varying the temperature, the dosage differs between the samples and the temperature is kept constant[2]. Substrate binding is recognized as a sudden increase in intensity around a certain concentration, as is shown in figure 2.1b.

2.2.3 Proteome-wide mass spectrometry CETSA

In MS, molecular compounds are broken up and accelerated in electromagnetic fields to determine its identity through its mass. The theory is that the movement of heavier compounds will be less affected by the fields and therefore hit the sensor in a different location or at a different time, depending on the type of MS instrument. Each impact registers as a peak, and by adding them together you can form a spectrum. The spectra can in turn be matched to molecular specific patterns and sequence information to determine their quantity and identity [3][7].

CETSA can be used in combination with MS readout for proteome-wide analysis. CETSA MS allows detection of directly and indirectly affected proteins, which in clinical trials can provide valuable information about a compounds wanted and unwanted effects. It can also be used to find the molecular origin of side-effects observed during pre-existing drugs therapies [3].

2.2.4 Isobaric Mass Tag Labeling

In the quantification step of CETSA MS, multiplexing of typically 8-10 temperatures or dosages can be enabled by using isobaric tandem mass tags (TMT10) or isobaric tags for relative and absolute quantification (iTRAQ). These strategies uses different isobaric tags to label the soluble fractions after the heat up phase, allowing tracking of the individual temperatures during MS-readout. Each replicate consequently require two labeling and MS sessions, one for the drug and one for the control sample [3].

Chapter 3

Data Reduction & Processing

3.1 Input Data

Two pipelines that automate the data processing have been developed in R [8], one for melt curves and one for dose responses. The starting points are the output files produced by Proteome Discoverer or MaxQuant [9][10]. These software are used to process raw data of dose responses or melt curves into txt-files, where peptides have been assigned to protein groups with a false discovery rate of less than 4%. This ensures that the poorest and general data has already been removed.

The cornerstone of the analysis are the intensities, representing the samples. Information about the master accession number, peptide sequence, modifications, description, identity score are also considered, while remaining columns are discarded. A total of four scripts are used to process melt curves and dose responses from both programs, since the file structure differs between them. For validating purposes all data considered stem from experiments conducted twice, thus represented by two datasets.

3.2 Normalization

Before further processing can be done, the data has to undergo a normalization step. This is necessary to avoid bias, since the intensities can vary between the different peptides due to its abundance in the sample. By dividing the peptides intensities with the value from its first sample, we transform the data to relative intensities that are easier to work with.

For the dose responses we also perform a second normalization for every column, to account for independent variance between samples that could have been caused by human or technical

errors, e.g. pipetting errors. Here we divide every sample intensity with the mean value of that column, i.e. the mean intensity for that particular sample. Since few proteins actually are affected by the substrate, i.e. the relative intensity will remain constant at one, this allows us to correct for possible errors without risking to alter correctly performed experiments. Even though it is not done here, this kind of normalization can also be conducted for melt curves. However, note that melt curves can look very different from case to case, and normalizing the columns can therefore lead to a more noticeable effect.

3.3 Curve Fitting

In order to facilitate comparison of curves, their inflection points are determined by applying a curve fitting model. This measurement, representing the point at which the curve changes from being concave to convex, is known as the EC50 value or the melting temperature (Tm) for dose responses and melt curves, respectively.

Both dose responses and melt curves have previously been shown to follow the pattern of a logistic function, which is why it can be advised to use the R package *Self-Starting Nls Four-Parameter Logistic Model*. It applies the formula shown in 3.1 and a non linear least square method to find the curve which best follows the intensities. Where A and B are the horizontal extreme values, to the left and right. Xmid represents the inflection point, and scale is a scaling parameter that reflects the steepness of the curve. The resulting coefficients and R^2 error value provides the measures to recreate the curve and an indication of how well the curve fits the points.

$$\frac{\boldsymbol{A} + (\boldsymbol{B} - \boldsymbol{A})}{\boldsymbol{1} + \boldsymbol{e}^{(\boldsymbol{xmid} - \boldsymbol{input})/\boldsymbol{scale}}} \tag{3.1}$$

3.4 Z-test

The curves representing the protein groups are determined by the mean value of their underlying peptide data. For the dose response data this concludes the processing, since no analysis has to be conducted between repeats. Melt curves, however, needs further processing to decide if the two treated samples differ from the controls.

As a first step, a two sided Z-test is performed to check if any possible shift can be explained with a normal distribution. The peptide data of both controls are pooled and compared to

corresponding data of the drug treated samples, individually. Their standard deviation and difference in Tm are used as determining factors to derive a p-value for each replicate. This value reflects the likeliness of observing the same shift by coincidence.

3.5 Shift Size

The p-values reflect significance of shifts but do not account for resolution. Consequently, very low p-values can be assigned to shifts too small to identify substrate binding. To strengthen the signal of true target engagement, a complementary measurement of the shift size is derived. The value is given by the optima searched for within the interval described by formula 3.2, also illustrated in figure 3.1. The margin of 0.1, removed from both sides of the interval, is intended to exclude the flattening sections where unproportionally large shifts can occur.

$$]\boldsymbol{maxmin(drug, control)} + \mathbf{0.1}, \boldsymbol{minmax(drug, control)} - \mathbf{0.1}[\tag{3.2}$$

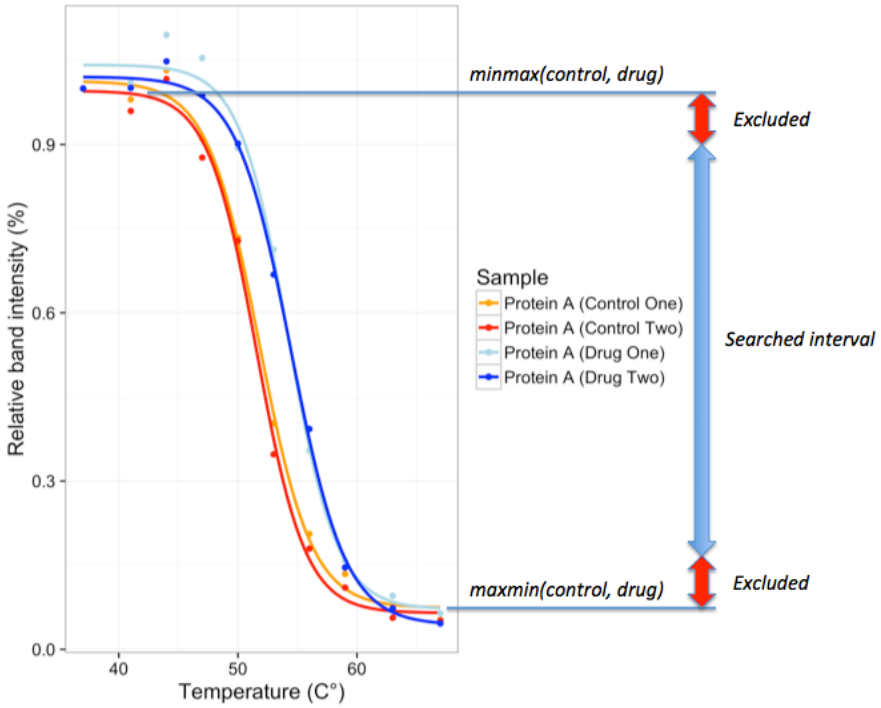
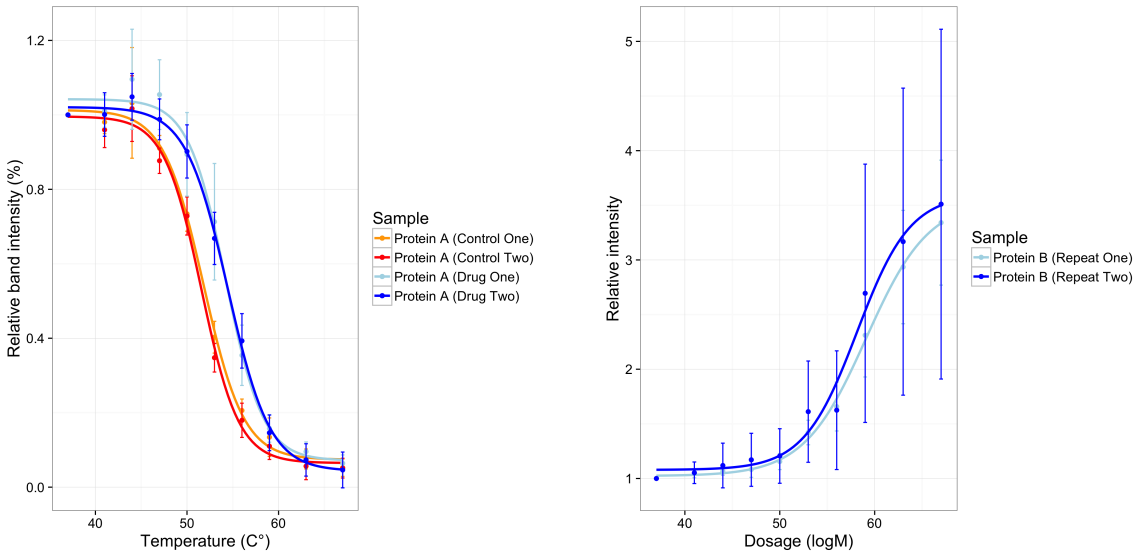


Figure 3.1: Maximum shift size is searched for within the interval indicated by the blue arrow.



(a) Melt curve

(b) Dose response

Figure 3.2: Example output curves, produced in the final step of the pipeline.

3.6 Plotting & Saving

As a final processing step, all curves representing protein groups are plotted and saved to a local directory. Peptide distribution, in form of a standard deviation, is indicated by error bars, as is shown in figure 3.2. For further details about a certain protein and its underlying data, one must open the tables that are saved as comma-separated values (CSV) files or, more conveniently, insert them into a database and use queries.

3.7 Passing Criteria

Not all data are fit, or even possible, to process. Rules have therefore been set to filter out data that, due to different reasons, have been deemed too poor. Table 3.1 shows a summary of the excluded data and the reasoning behind it.

Table 3.1: Excluded data during processing

Excluded	Reasoning
Peptides assigned to more than one protein group	This happens when an MS spectra can not be uniquely tied to single protein group, which in proteomics can be due to conserved regions or isoforms. Allowing these peptides to persist would provide false positives for the falsely listed groups, which may skew the result. Removing them, however, weakens the signal for the correctly listed one. This is always the tricky trade off between sensitivity and selectivity. Here the more selective approach is chosen.
Peptides with any normalized intensity above two *	An increase in intensity indicates that the protein suddenly becomes more abundant. In reality, this is close to impossible and is more likely to have its explanation in structural biology or a technical/human error. A cutoff is therefore placed to remove some data that does not make sense.
Peptides with data points not manageable by the curve fitting model	The model used can only work with data points that somewhat follow the pattern of a four variable logistic function, due to the formula used. Poor and fluctuating data may fall too far from the pattern and provoke an error. Removing this data simply allows us to discard data that would provide unreliable results. In addition, dose responses that show no signal, form constant lines and are thus also removed by this criteria.
*	*Only applied in the melt curve pipeline*

Chapter 4

Quality Assessment

Sorting out the most interesting parts in a proteome-wide dataset can prove challenging, especially if you do not know what to look for. By implementing a value that estimates the result's relevance, you can narrow the search field and thus save time as well as lower the bar for required theoretical knowledge. The solution used here is a threshold based score, derived from a number of parameters, summarized in table 4.1. As you will notice, some parameters differ between dose responses and melt curves, due to different characteristics of the method. Ligand binding in melt curves are for example indicated by shifts, thus p-value and shift size will play an important role. In dose responses, however, the most important factor is the max value, i.e. the response to the treatment. Exactly what thresholds are used and how points are assigned, is shown in table 4.2 and 4.3.

The final score given to a protein group does not only reflect the indication of ligand binding, but also the data quality. Ranks from A to D are set in a number of categories, based on their underlying value, to make the scoring easier to follow. The weights placed on the different ranks has been decided by testing the algorithm on a well studied test set and evaluate the result. The aim is to rank cases were a shift can be observed in the top, but still punish for poor quality enough to degrade data that is not trustworthy. In the ideal case we will get the shifts with sufficient data quality in the top, followed by high quality data showing no shift. This will bring unreliable shifts down in rank, which should give a good indication that they are to be treated with care.

Table 4.1: Parameters used to determine curve quality

Parameter	Description
Shift size**	Distance between the control and treated sample. The parameter is treated as a boolean value in the scoring process, meaning that its either a shift or its not, nothing in between. View section 3.5 for further details.
P-value**	Shift significance, based on the melting temperature of all involved peptides. The value is decided by a two sided Z-test, see section 3.4.
Max*	Max value of the protein function, which is the same as the B-value in formula 3.1. This indicates how much stability is gained by the treatment, for that particular protein group.
Common peptides	Number of peptides the samples have in common. This parameter reflects how robust the result is. An indication of ligand binding is more trustworthy if it can be shown for the same peptide in all samples, preferably several peptides.
Standard deviation	A measure to reflect the distribution of peptides. A low value indicates that the peptides of a particular sample follow the same pattern. The thresholds used for setting the score are determined by partioning a sorted test set into four equally large sections and capturing the border values.
R^2 error	The R^2 error is fetched from the curve fitting model that assembles the peptides to a protein group. It describes how well the points reflects the final curve, calculated as the sum of squares. Thresholds are set the same way as for the standard deviation parameter.
*	*Only used for dose responses*
**	*Only used for melt curves*

Table 4.2: Score thresholds for melt curve parameters

Measure	Value	Ranking	Score
P-value	< 0.05	A	40
	< 0.10	B	30
	≥ 0.10	C	10
	NaN	D	0
Shift Size	> 2	Y	30
	≤ 2	N	0
Common peptides	> 4	A	10
	4	B	8
	3	C	5
	≤ 2	D	0
Standard deviation	≤ 2.4	A	10
	≤ 2.8	B	8
	≤ 3.3	C	5
	> 3.3	D	0
R^2 error	≤ 0.032	A	10
	≤ 0.045	B	8
	≤ 0.060	C	5
	> 0.060	D	0

Table 4.3: Score thresholds for dose response parameters

Measure	Value	Ranking	Score
Max	Repeat 1 & Repeat 2 > 3	A	55
	Repeat 1 & Repeat 2 > 2	B	40
	Repeat 1 or Repeat 2 > 2	C	25
	Repeat 1 & Repeat 2 ≤ 2	D	0
Common peptides	> 4	A	15
	4	B	10
	3	C	5
	≤ 2	D	0
Standard deviation	≤ 0.48	A	15
	≤ 0.67	B	10
	≤ 0.89	C	5
	> 0.89	D	0
R^2 error	≤ 0.01	A	15
	≤ 0.019	B	10
	≤ 0.034	C	5
	> 0.034	D	0

Chapter 5

CETSA Analytics

Making the result accessible and easy to interpret had a high priority in this project. If data are to be viewed by multiple persons, it quickly becomes inefficient to pass files between local computers, and factors like computer experience, formats and operating system can become obstacles along the way. With this in mind, a web based system named CETSA Analytics was developed, which allows clients to access the data directly from their browsers. The systems structure is formed by a website and a Structured Query Language (SQL) database, both uploaded to a web hotel. New data can be uploaded to the server at any time from any computer, provided that you have access to the credentials. With the many styling options in a web environment its also possible to present the result in an appealing manner.

5.1 Database

Databases are key components in any system that repeatedly needs to locate information. Doing so in a large dataset can prove challenging, even more so if it lacks in structure. By inserting data into a database, you provide it with a solid structure that facilitate searching as well as storing of data.

The processing pipeline described in chapter 3 generates three related files, containing protein-, peptide- and quality data, of which the largest hold about one million rows. This makes the use of a MySQL database especially convenient. MySQL has high performance for data that are somehow related, i.e. have one or more attributes in common, and that scale well thanks to indexing [11]. The database can also be made accessible through the web programming language Hypertext Preprocessor (PHP), allowing queries to be determined and executed

in a web browser environment [12]. The structure used here consists of six tables, containing proteins, peptides and quality for dose responses and melting curves, separately. All which can be imported directly from the CSV files provided by R.

5.2 User Interface

The environment in which the user operates is commonly referred to as the Graphical User Interface (GUI) or just User Interface (UI). In CETSA Analytics the UI is built similarly to any website, with HyperText Markup Language (HTML), Cascading Style Sheets (CSS) and JavaScript [13][14][15]. The only exception is that it also include PHP code, which is required to communicate with the database and perform checks to maintain a level of security. What elements the site contains, e.g. tables, pictures and text, and their styling is decided by HTML and CSS. JavaScript is used to add scripts to the site that can provide convenient features, e.g. interactively show or hide elements.

5.2.1 Site architecture

The site uses a one tier layout, in the sense that no section is a subsection of another. Multiple tiers usually make sites easier to navigate, but may feel unnecessary with only a few different pages. There is, however, a navigation bar with topics to guide you to the right location. Figure 5.1 shows a complete picture of the site architecture.

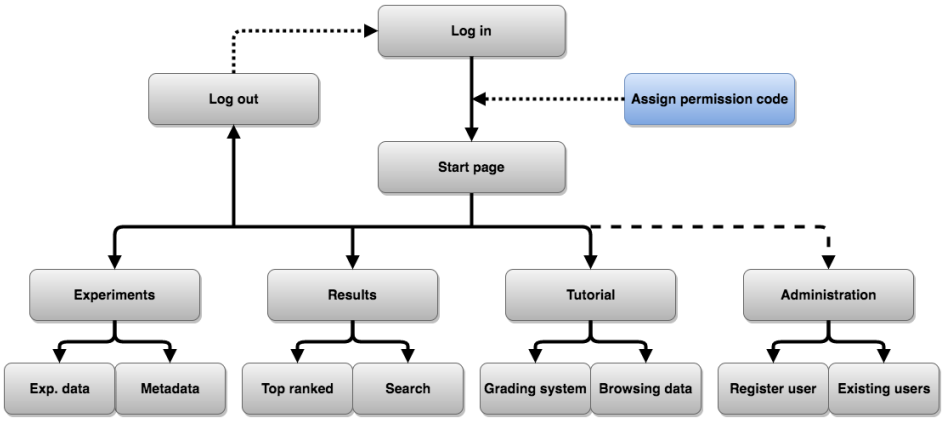


Figure 5.1: The site architecture, where dotted lines represent automatic procedures and semi drawn lines indicate sections only accessible with admin permissions.

5.3 Security

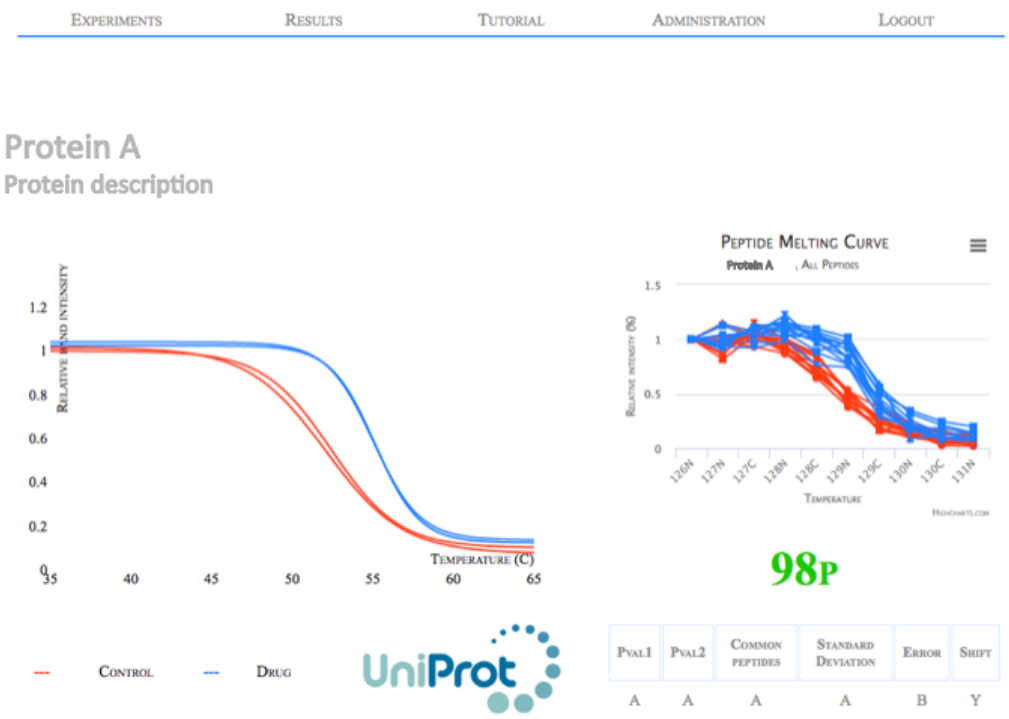
One of the drawbacks with utilizing a web based system is the ever present risk of security breaches. To develop a completely secure system is next to impossible, but you can lower the risk significantly by implementing some features. The most obvious and important security feature in CETSA Analytics is hashed passwords. It means that the passwords are encrypted with a key before stored in the database, making the information nonsense unless you can get a hold of both parts. The same key is then used again to decipher the passwords when called for during login attempts. During login, the system also checks for brute-force attempts, by storing time stamps in a table, and blocks the user if wrong password is given too many times within a time window. As a measure to ensure that clients only are able to browse their own data, there are also permission codes added to the data, only allowing a user matching the permission to view it.

5.4 Use Cases

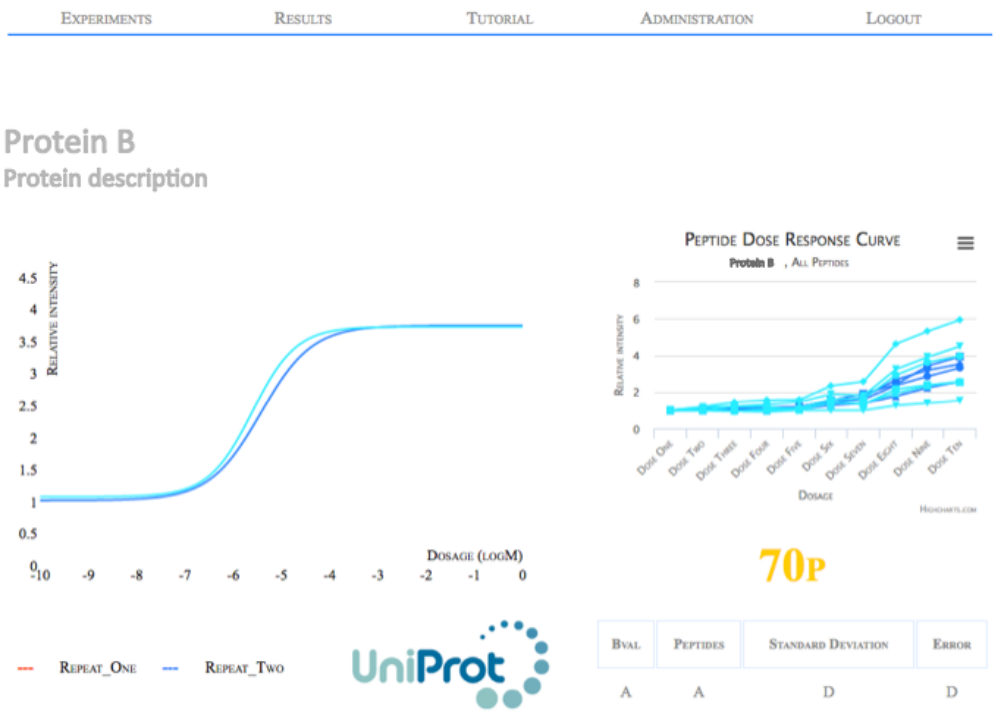
The main focus of CETSA Analytics is to allow easy access and interpretation of processed dose response and melt curve data. However, as is stated in section 5.2.1 it is also intended to have an administrative, experimental and tutorial section, of which the latter two are not yet fully implemented. To clarify how the system can be used more specifically in its current state, a number of use cases are demonstrated.

5.4.1 View detailed result data

The detailed result page is the core of the system, seen in figure 5.2. Every protein group present in the database can be viewed individually here, with score and underlying peptide data. The two graphs at the top of the page represent the protein group and all its peptides, colored to represent controls and treated samples, or repeat one and two for dose responses. A number is shown directly beneath the peptide graph, indicating the score, ranging up to one hundred. Here you also have the rankings the score is based on, as well as a link the uniprot page for the protein. By scrolling down you can view the underlying petide data for the individual samples, in form of tables and graphs. This allows you to follow exactly what peptides are present and how they are distributed.



(a) Melt curve



(b) Dose response

Peptide Details

Control One

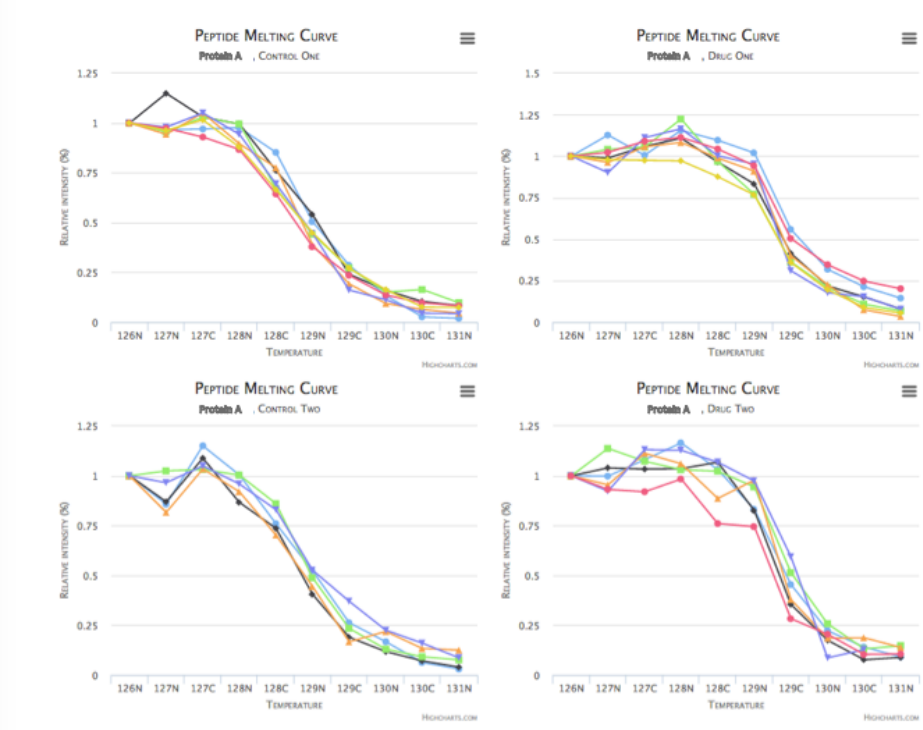
Sequence	Intensity One	Intensity Two	Intensity Three	Intensity Four	Intensity Five	Intensity Six	Intensity Seven	Intensity Eight	Intensity Nine	Intensity Ten	MeltingTemp	Mods
TPPEAIALCSSLLEYTPSSR	1	0.965626	0.969896	0.976517	0.851768	0.504364	0.285146	0.133389	0.0262723	0.0193209	53.3814	Carbamidomethyl [C9]
KDELYLNLVLEYVPETVYR	1	1.14872	1.02982	0.994786	0.760984	0.540103	0.240188	0.159799	0.103585	0.0832452	52.2753	
LSPLEACAHSFFDELR	1	0.955205	1.02887	0.994912	0.670791	0.441027	0.272034	0.150593	0.163772	0.0976226	51.5587	Carbamidomethyl [C7]
TSSFAEPGGGGGGGGGGPGGSASGPGGTGGGK	1	0.942904	1.05075	0.895416	0.775093	0.388722	0.191961	0.0939626	0.0638457	0.0450474	52.0163	
VTTVVATLGQGPER	1	0.979094	1.05024	0.942731	0.693985	0.445587	0.161101	0.11268	0.044955	0.0429468	51.9143	
DIKPQNLLVDPDTAVLK	1	0.973776	0.929113	0.866969	0.643381	0.378509	0.234854	0.136127	0.0969984	0.0819333	51.2224	
SQEVAYTDIK	1	0.963265	1.01557	0.878654	0.664528	0.448655	0.270862	0.162692	0.0758152	0.073995	51.8466	

Drug One

Control Two

Drug Two

(c) Peptide tables



(d) Peptide graphs

Figure 5.2: Detailed result page of CETSA analytics. a) & b) Example protein plot, combined peptide plot and score section of a protein group from a melt curve and dose response experiment, respectively. c) Example peptide table belonging to a protein group from a melt curve experiment. d) Example peptide plots belonging to a protein group from a melt curve experiment. Every sequence can be tracked individually by hovering over a series.

5.4.2 Perform searches

A most important feature when dealing with large datasets is the ability to search. This page can be accessed from *Results*, but require you to know the accession number or name of the protein group you are interested in. If you just provide the first two or three letters, all proteins starting with those will be listed. With this done you now have two options, you can either click directly on the group of interest to view its detailed result page, or you can add it to the preview. The preview is a smaller window at the bottom if the page that updates when you click the *add to preview* button. This will only show the protein curves, without any details, but allows you to add additional graphs next to it, in case you want to compare it with others. You are now allowed to perform a new search without removing the data in the preview, unless you decide to press the reset button. Figure 5.3 shows an example page, where a search has been executed and three proteins have been added to preview from previous searches.

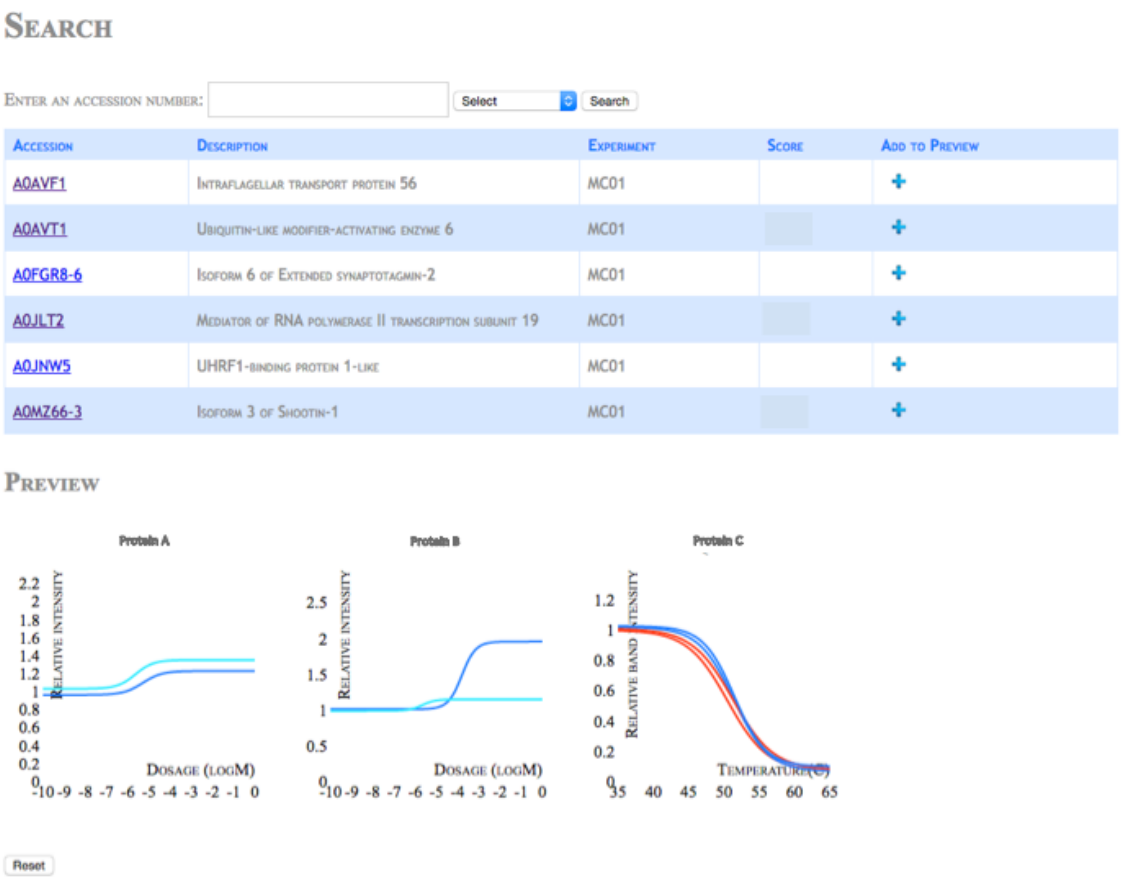


Figure 5.3: A search example, where two censured dose responses and one melt curve have been added to the preview. The scores of the searched proteins have been removed, with respect to the owner of the dataset.

5.4.3 View most relevant protein groups

The first page visited after conducting a new experiment is probably *Top Ranked*, found under *Results*. Here you have the option to view the best scoring protein groups of a particular experiment in descending order, see figure 5.4. The layout is similar to that of preview in the search section, in the sense that it only draws the protein curves. You can, however, click on any graph you find interesting to jump directly to its detailed result page.

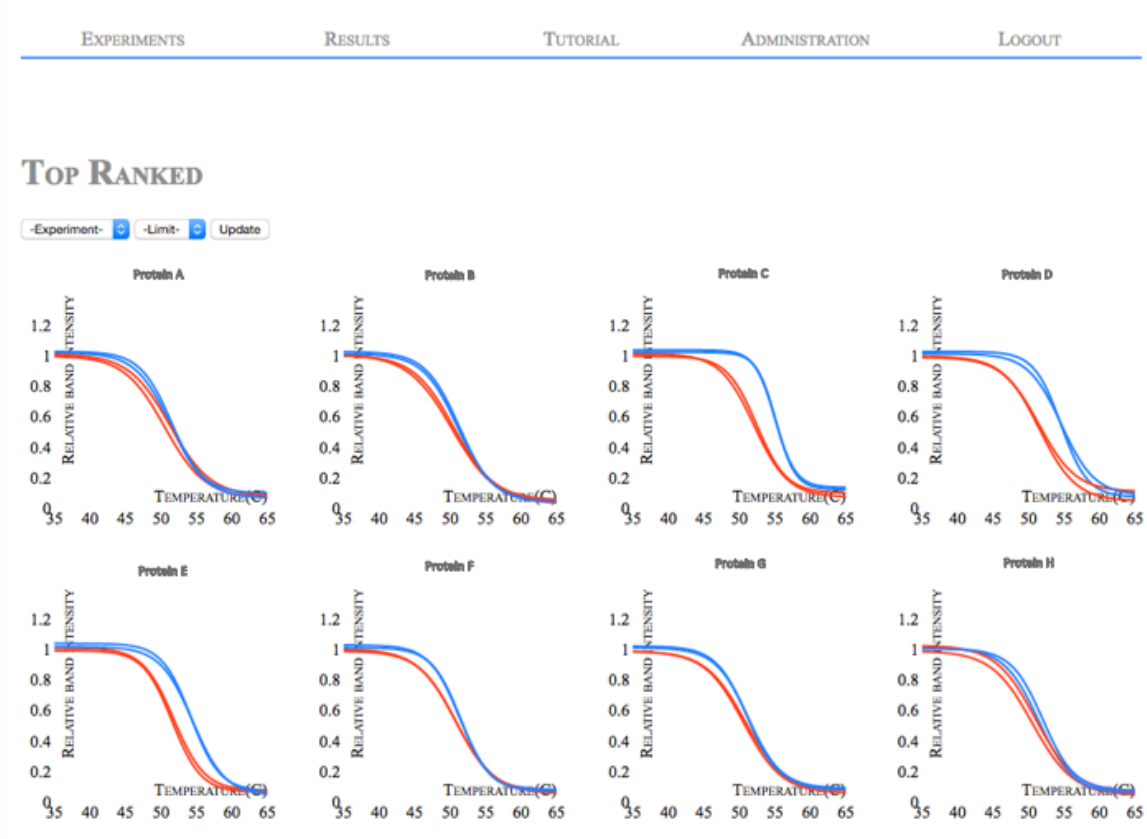


Figure 5.4: Best scoring proteins of a selected experiment.

5.4.4 Register

Registration of new accounts can only be done by an administrator. With such permissions you will see a topic named *Administration* in the navigation bar, under which you can find *Register new user*. Here you type in all the credentials, including permission code, and store it to the database, see figure 5.5. The permission code decides what data the user will be able to view, and must match with their datasets. If you later add more data for that particular client, you can visit *Existing users* to check what code they where given.

Figure 5.5: Registration page. Only accessible by admin users.

Chapter 6

Discussion

CETSA Analytics is currently being tested by Pelago Bioscience AB, as a tool to reach out to their clients with results of ordered experiments.

The system features searching and viewing of CETSA MS data of both melt curve and dose response character, and the experimental section is under construction. It runs on servers provided by One.com, which also oversees the storage of the SQL-database, and is fully available from any web-browser. Appropriate theoretical knowledge is still required, but future plans include development of a tutorial section.

We have yet to see how robust and resilient to stress the system is. It has so far only been tested for three simultaneous users, but it is likely to host more users in the future. If system crashes start occurring, it could be due to insufficient memory on the server or passages of code that tend to go into loops. A well developed error handling could in such cases help to pinpoint the problem and find the code that needs to be rewritten.

In its current state the system can draw the top 500 proteins in about 12 seconds and perform a search in under a second. This indicates that the bottleneck is the drawing of the curves, which is an issue with the processor on the server rather than the database. If 12 seconds feel unbearable, or more than 500 proteins is desired, it is possible to upgrade the account at one.com to gain access to more processor power.

As data accumulates in the database, we will also experience increased query times. This can be prevented by keeping data in the system for a limited time and then store it to a local backup instead. It is also possible to increase the performance by introducing a slave-master setup for the hard drives, where the drives have different designated tasks. However, this is something that would have to be done at the server side, by one.com, if not already implemented.

The processing pipelines will be evaluated, and possibly reworked, after comparison with alternative workflows from other organizations that practice CETSA. The grading system is especially likely to change over time, as we adjust the weights of the parameters. It would probably be for the better if the threshold based score in time can be replaced with a mathematical formula, but this has to be done carefully. The advantage with using a threshold base score is the robustness it brings. Outliers, consisting of very high or low values, are given the same score as other values surpassing our predefined criteria. A p-value of 10^{-17} are for example not given a better score than one of 10^{-3}. If we are to implement a continuous formula we have to carefully consider what will happen to those outliers and minimize their impact. However, if this can be achieved, a continuous formula would prevent the problems we are now facing with unrealistically big jumps in the score. For example, if our predefined threshold of a shift is set to one, we will have a 30 points difference between a shift of 0.99 and one of 1.01. This is of course not ideal.

Even though the system still lacks in some regards, it has proven decently accurate at picking out the most interesting results of an experiment. It is also, as far as we know, the only system that provide melt curve analysis and scoring that consider data at peptide level. The fact that it is based on peptide data allows for a robust analysis where it is easy to follow and evaluate the result, even in cases where the performance of the scoring algorithm is questionable. By the increased efficiency offered by the system in regards of distribution and interpretation, it will hopefully be a helpful tool in the struggle towards faster progression of ligand analysis.

Chapter 7

Acknowledgements

I thank my supervisor D. M. Molina [Pelago Bioscience AB] for valuable discussions throughout the project, J. Lengqvist [Karolinska Institute (KI)] for help with data interpretation and discussion about data quality, J. Bergquist [Uppsala University] for input about the quality assessment and for reviewing the project, and J. Andersson [Uppsala University] for reviewing my report and for the administrative work surrounding the project.

I am also grateful to M. Dabrowski [CEO at Pelago Bioscience AB] and the Pelago Bioscience team for giving me this opportunity and making it a pleasant experience.

Bibliography

[1] D. M. Molina and P. Nordlund, "The Cellular Thremal Shift Assay: A Novel Biophysical Assay for IN SItu Drug Target Engagement and Mechanistic Biomarker Studies", in *Annual Review of Pharmacology and Toxicology*, 56, 141-161, November 2015.

[2] D. M. Molina et al., "Monitoring Drug Target Engagement in Cells and Tissues USing the Cellular Thermal Shift Assay", in *Science*, 341, 84-87, July 2013

[3] M. M. Savitski et al., "Tracking cancer drugs in living cells by thermal profiling of the proteome", in *Science*, 346, 1255784, October 2014

[4] M. Vedadi et al., "Chemical screening methods to identify ligand that promote protein stability, protein crystallization and structure determination", in *Proc. Natl. Acad. Sci. U.S.A.*, 103, 15835-40, October 2006

[5] D. L. Nelson and M. M. Cox, "Principles of Biochemistry", 5th edition, 2008

[6] O. Fedorov et al., "A systematic interaction map of validated kinase inhibitors with Ser/Thr kinases", in *Proc. Natl. Acad. Sci. U.S.A.*, 104, 20523-8, December 2007

[7] H. Lodish et al., "Molecular Cell Biology", 6th edition, 2007

[8] R Core Team, "R: A Language and Environment for Statistical Computing", 2016, https://www.R-project.org/, [Accessed 10 November 2016]

[9] J. Cox, "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification", in *Nat Biotechnol*, November 2008

[10] Thermo Scientific, "Proteome Discoverer", 2016, https://www.thermofisher.com/, [Accessed 10 November 2016]

[11] MySQL AB, "MySQL", 2016, https://www.mysql.com/, [Accessed 10 November 2016]

[12] The PHP Group, "PHP: Hypertext Preprocessor", 2016, http://www.php.net/, [Accessed 10 November 2016]

[13] World Wide Web Consortium, "HTML: HyperText Markup Language", 2016, https://www.w3.org/, [Accessed 10 November 2016]

[14] World Wide Web Consortium, "CSS: Cascading Style Sheets", 2016, https://www.w3.org/, [Accessed 10 November 2016]

[15] B. Eich, "JavaScript"