NCEMS – CURE: Project Outline

2025-08-20 by Ian Sitarik

**Table of Contents**

[(I) Examining Two “Shot-Gun” Proteomics Experiments 2](#_Toc207009888)

[(I – 1) Extracting key information from the publications 2](#_Toc207009889)

[Q. What cleavage enzyme was used and what amino acids does it cut after? 2](#_Toc207009890)

[Q. Was the dataset labeled or label-free? 3](#_Toc207009891)

[Q. Were any specific chemical modifications or post-translational modifications (PTMs) mentioned by the authors? 4](#_Toc207009892)

[(II): Open Modification Search with Spectral Alignment Guided Engine (SAGE) 5](#_Toc207009893)

[(II – 1) How SAGE Works: Overview of the Pipeline 6](#_Toc207009894)

[(II – 2) Examine the peptide quantification file 6](#_Toc207009895)

[Q. Why do some peptides have only letters and others have numbers in brackets? 7](#_Toc207009896)

[Q. What is the range, mean, median, and mode of the peptide charges observed? 7](#_Toc207009897)

[Q. (Bonus) why do we ignore +1 charged peptide fragments? 7](#_Toc207009898)

[Q. Do any peptides map to more than 1 protein? 7](#_Toc207009899)

[Q. What is the difference between a q-value reported here and a p-value? 7](#_Toc207009900)

[Q. What are the units of the spectral intensity (if any)? 7](#_Toc207009901)

[(III) Calculate and Examine the Protein Abundance Distributions 7](#_Toc207009902)

[(III – 1) Filter the peptide tables 7](#_Toc207009903)

[(III – 2) Select and clean relevant columns for analysis 8](#_Toc207009904)

[(III – 3) Estimate Protein Abundances 8](#_Toc207009905)

[(III – 4) Compute Sample Statistics and Make Box and Violin Plots of each Sample Separately 8](#_Toc207009906)

[Q. Do you observe any noticeable differences in the distribution of protein abundances between technical replicates under the same condition? 9](#_Toc207009907)

[Q. How about between conditions? 9](#_Toc207009908)

[Q. What does this tell you about whether the treatment had an effect? 9](#_Toc207009909)

[Q. Is it possible that there is affect we cannot observe by examining this population level change? If so how? 9](#_Toc207009910)

[(IV): Differential Protein Abundance from LFQ Mass Spectrometry Data Datasets 9](#_Toc207009911)

[(IV – 1) Standard Scale Renormalization of Protein Abundance 9](#_Toc207009912)

[(IV – 2) Statistical testing (Welch’s t-test) 9](#_Toc207009913)

[(IV – 3) Ranking and multiple testing correction 10](#_Toc207009914)

[Q. What fraction of proteins had a significant change in their abundance? 10](#_Toc207009915)

[(IV – 4) Volcano plots 10](#_Toc207009916)

[Q. How many proteins have a significant change after treatment and a ? 10](#_Toc207009917)

[Q. How many proteins have a significant change after treatment and a ? 10](#_Toc207009918)

[Q. How many proteins have a significant change after treatment and a ? 10](#_Toc207009919)

[Q. How many proteins DO NOT have a significant change after treatment and a ? 10](#_Toc207009920)

[Q. How many proteins DO NOT have a significant change after treatment and a ? 10](#_Toc207009921)

[Q. How many proteins DO NOT have a significant change after treatment and a ? 10](#_Toc207009922)

[Q. which of these groups would you focus your efforts on investigating further to understand how the treatment is potentially affecting protein production and function? 10](#_Toc207009923)

# (I) Examining Two “Shot-Gun” Proteomics Experiments

Shotgun proteomics is a high-throughput strategy for identifying and quantifying proteins in complex biological samples. It works by enzymatically digesting proteins into peptides, separating those peptides via liquid chromatography, and analyzing them using tandem mass spectrometry (MS/MS). This approach allows researchers to characterize thousands of proteins in a single experiment without prior knowledge of the sample’s composition.

In this project, we will explore how protein abundance information can be extracted from public datasets in the [PRIDE Archive](https://www.ebi.ac.uk/pride/archive/projects/), using mass spectrometry-based shotgun proteomics. We focus on two biologically and technically distinct datasets:

* Soybean (*Glycine max*): [PXD023343](https://www.ebi.ac.uk/pride/archive/projects/PXD023343)
* Human (*Homo sapiens*): [PXD005187](https://www.ebi.ac.uk/pride/archive/projects/PXD005187)

These examples illustrate different use cases, labeling strategies, and experimental goals in proteomics.

## (I – 1) Extracting key information from the publications

In this section, you’ll extract important metadata from both the original publication and the experimental files associated with your dataset. This information will be used to build a \*.json\* configuration file required by SAGE to process the data.

[1] DOI: 10.1038/cddiscovery.2017.10, Des: Kathiriya JJ, Nakra N, Nixon J, Patel PS, Vaghasiya V, Alhassani A, Tian Z, Allen-Gipson D, Davé V. Galectin-1 inhibition attenuates profibrotic signaling in hypoxia-induced pulmonary fibrosis. Cell Death Discov. 2017 Apr 10;3:17010. eCollection 2017

[2] DOI: 10.3390/ijms22020920, Des: Cheng L, Min W, Li M, Zhou L, Hsu CC, Yang X, Jiang X, Ruan Z, Zhong Y, Wang ZY, Wang W. Quantitative Proteomics Reveals that GmENO2 Proteins Are Involved in Response to Phosphate Starvation in the Leaves of Glycine max L. Int J Mol Sci. 2021 22(2)

### Q. What cleavage enzyme was used and what amino acids does it cut after?

Why it matters*:* The enzyme determines the peptide cleavage rules which are crucial for understanding how a protein will fragment in the spectrometer.

*Where to look:* Check the methods section of the manuscript for sample preparation or digestion steps. Common entries include trypsin, Lys-C, chymotrypsin, or Glu-C. This information may also be in PRIDE metadata under “Sample Protocol” or “Digestion.”

#### Table 1: Common Cleavage Enzymes

| Enzyme | Cleaves After (Residues) | Cleavage Blocked By | Terminal Side of Cleavage | Notes |
| --- | --- | --- | --- | --- |
| Trypsin | K, R | P | C-terminal | Most commonly used; highly specific |
| Lys-C | K | — | C-terminal | Stable in high denaturant (e.g., urea) |
| Arg-C | R | — | C-terminal | Less specific than trypsin |
| Glu-C | E, sometimes D | — | C-terminal | Cleaves more broadly at high pH |
| Chymotrypsin | F, Y, W, L, M | P | C-terminal | Broader specificity; more missed cleavages likely |
| Asp-N | D | — | N-terminal | N-terminal cleavage; less common |
| Pepsin | Broad (especially F, L, E) | pH-dependent | C-terminal or mixed | Non-specific, active at low pH |
| Thermolysin | L, I, V, F, M | — | N-terminal | Metalloprotease, stable at high temperature |

### Q. Was the dataset labeled or label-free?

Why it matters*:* Labeling strategies (e.g., SILAC, TMT) enable multiplexed, highly accurate quantification, while label-free methods are simpler and more scalable. Knowing the approach clarifies the strengths, limitations, and comparability of the quantitative results.   
Where to look*:* Look for keywords like “TMT”, “iTRAQ”, or “LFQ” in the methods section or PRIDE metadata.

#### TMT-Labeled Data

TMT (Tandem Mass Tags) are chemical labels attached to peptides that allow multiplexed MS-based quantification using reporter ions. These are detected at MS2 or MS3 levels.

Use this when*:*

* The methods mention TMT10, TMT11, TMT16, or iTRAQ.
* Quantification is based on reporter ion intensities.

#### SILAC-Labeled Data

SILAC (Stable Isotope Labeling by Amino acids in Cell culture) incorporates heavy isotopes into amino acids (e.g., 13C615N4-Arg & 13C615N2-Lys) during protein synthesis, shifting precursor m/z but not fragment ions.

Use this when*:*

* The methods mention *SILAC*, *light/heavy*, or *metabolic labeling*.

#### Label-Free Quantification (LFQ)

LFQ uses *precursor ion intensity* without any chemical or isotopic labeling. It’s based on matching peptide features across runs.

Use this when*:*

* There is *NO mention of TMT, iTRAQ, or SILAC, sometimes the authors will explicitly state they used LFQ methods*.

### Q. Were any specific chemical modifications or post-translational modifications (PTMs) mentioned by the authors?

Why it matters*:* These modifications often change the mass and charge of a peptide and thus its mass spectra. Sometimes these modifications are intentional such as in the case of labeling experiments to test how protein abundances changes under different conditions.   
Where to look*:* Usually listed under *sample preparation* or *search parameters* in the methods section.

#### Static vs. Variable Modifications

In proteomics, modifications refer to changes in the mass of a residue or terminus caused by chemical labeling, post-translational modifications (PTMs), or experimental artifacts.

#### Static Modifications

A *static modification* is a change that occurs on every instance of a given residue or terminus.

Use static mods when:

* The modification was applied to all peptides during sample preparation
* You are using chemical labels like TMT or iTRAQ
* The mass shift is consistent and universal

#### Examples of common static mods:

| Modification | Residue | Mass Shift | When It Occurs |
| --- | --- | --- | --- |
| Carbamidomethylation | C | +57.0215 | During alkylation with iodoacetamide |
| TMT16 labeling | K, ^ | +304.207 | Chemical labeling (TMT 16plex) |
| iTRAQ 8plex | K, ^ | +304.2054 | iTRAQ chemical labeling |

#### Variable Modifications

A *variable modification* is optional — the search engine will consider both the modified and unmodified versions of the peptide. This allows detection of PTMs or partial labeling.

Use variable mods when:

* You expect the modification to occur on only some peptide
* You’re searching for biologically relevant PTMs
* You want to allow missed modifications (e.g., incomplete labeling)

#### Examples of common variable mods:

| Modification | Residue | Mass Shift | When It Occurs |
| --- | --- | --- | --- |
| Oxidation | M | +15.9949 | Spontaneous or regulated (PTM) |
| Deamidation | N, Q | +0.984 | Non-enzymatic or enzymatic (PTM) |
| Pyro-glutamate formation | ^Q, ^E | −17.0265 / −18.0106 | N-terminal loss of ammonia |
| Acetylation | ^ | +42.0106 | N-terminal (PTM) |
| SILAC heavy lysine (labeling) | K | +8.0142 | Metabolic labeling |
| SILAC heavy arginine | R | +10.0083 | Metabolic labeling |

# (II): Open Modification Search with Spectral Alignment Guided Engine (SAGE)

SAGE (Spectral Alignment Guided Engine) is an open-source, high-throughput software tool for database searching and quantification of mass spectrometry (MS/MS) proteomics data. Developed by the Lazar Lab at the NIH, SAGE is designed to scale efficiently across large datasets while maintaining high sensitivity and accuracy in peptide identification.

SAGE implements modern algorithmic strategies and leverages predictive scoring models, retention time prediction, and multi-level quantification schemes to identify and quantify peptides from tandem mass spectra with high confidence.

SAGE is particularly useful for:

* High-throughput proteomics datasets
* Label-free and isobaric tag quantification (e.g., TMT, SILAC)
* Studies requiring open or semi-open modification searches
* Peptide identification across multiple samples with alignment

## (II – 1) How SAGE Works: Overview of the Pipeline

1. Database Generation (In Silico Digestion)
   * The input file containing the protein sequences of every protein in the organisms proteome is digested virtually using user-defined cleavage rules and specified static/variable modifications
   * All possible peptide sequences are generated, filtered by length and mass
2. Spectral Preprocessing
   * Experimental spectra are filtered to remove low-intensity noise and spectra with insufficient peaks.
3. Scoring and Peptide Matching
   * SAGE matches experimental MS/MS spectra to the in silico-generated peptide fragments using scoring algorithms based on:
     1. Ion type matches (e.g., b/y ions)
     2. Fragment ion intensity
     3. Mass tolerance (in ppm or Da)
     4. Number of matched ions
   * It computes statistical scores to rank peptide-spectrum matches (PSMs).

* Retention Time (RT) alignment helps improve scoring by matching predicted vs. observed elution behavior, reducing false positives.

## (II – 2) Examine the peptide quantification file

Let us examine the output from SAGE for either the Soybean or Human samples. You can find them located in:

* data/Human\_SAGE\_results/processed\_sage\_results.tsv
* data/Human\_SAGE\_results/processed\_sage\_results.tsv

Each of these files contains one peptide fragment per row that was observed in the mass spectrometry experiments along with some important information regarding the peptide listed below.

SAGE Output File Description

|  |  |
| --- | --- |
| Column name | Description |
| peptide | The amino-acid sequence of the identified peptide (after in silico digestion, e.g., tryptic). |
| Charge | The precursor ion charge state observed in MS1 (e.g., +2, +3). |
| proteins | The protein accession(s) the peptide maps to, in UniProt-style format (e.g., sp|P12345|PROT\_NAME). |
| q\_value | The estimated false discovery rate (FDR)–adjusted confidence metric for the peptide-spectrum match (PSM), expressing the minimal FDR at which that identification would be accepted. |
| score | The primary peptide–spectrum match (PSM) score reflecting how well the experimental MS/MS spectrum agrees with the candidate peptide. |
| spectral\_angle | A spectral similarity metric (often the normalized spectral contrast angle) between the observed experimental MS/MS spectrum and a reference/predicted spectrum. |
| \*.mzML,.. | Quantitative intensity estimates (e.g., MS1 extracted ion currents or aggregated peptide-level abundances) for that peptide in the three biological/technical replicates of the LP condition |

### Q. Why do some peptides have only letters and others have numbers in brackets?

* TAEENLDR vs. EM[+15.9949]NDAAMFYTNR for example

### Q. What is the range, mean, median, and mode of the peptide charges observed?

### Q. (Bonus) why do we ignore +1 charged peptide fragments?

### Q. Do any peptides map to more than 1 protein?

### Q. What is the difference between a q-value reported here and a p-value?

### Q. What are the units of the spectral intensity (if any)?

# (III) Calculate and Examine the Protein Abundance Distributions

Now that you have examined the list of identified peptides that have been confidently matched to proteins of interest we need to obtain estimates of the protein level abundances. Here we will lay out a general procedure for this quantification so you can attempt it on your own using your preferred programming language. You are also free to use the python Jupyter notebooks/scripts located in the src/data/ folder.

## (III – 1) Filter the peptide tables

Explanation: Filters out low-confidence identifications. Only keeps peptides with q\_value ≤ 0.05, controlling peptide-level false discovery.

Why: Filtering by q-value ensures you are only using peptide measurements with statistical support propagating too many false positives would pollute protein-level inference.

Note: The threshold (0.05) is conventional; students can try stricter (0.01) or looser as an exercise to see how discovery changes.

## (III – 2) Select and clean relevant columns for analysis

Explanation: Constructs a subset of columns: retaining peptide, the original proteins annotation, q\_value, and all other columns except technical metadata (charge, score, spectral\_angle) which are not needed for quantification.

* The proteins field is in UniProt format like sp|P12345|PROT\_NAME;tr|P12345|PROT\_NAME;....
  + Multiple proteins are separated by a semi colon “;”
* For each peptide make a row copy for every protein it could possibly map to.

Why: Reduces clutter and focuses downstream computation on actual intensity data and needed identifiers.

## (III – 3) Estimate Protein Abundances

Explanation: Group all peptides assigned to each UniProt accession and collapses their intensities by taking the median intensity across peptides within each sample. You should end up with 6 estimates of the protein abundance per protein.

Why median? Median is robust to outliers (e.g., aberrant peptide measurements) and is a simple estimate of protein abundance from multiple peptide surrogates.

Note: Alternatives include mean, weighted mean (by peptide quality), or more sophisticated models (like MaxLFQ’s ratio-based inference). Students can compare results of median vs mean as an exploration. Some further reading on how complicated this abundance estimation can get: <https://www.nature.com/articles/nmeth.3901.pdf>

## (III – 4) Compute Sample Statistics and Make Box and Violin Plots of each Sample Separately

You should compute the following statistics for each sample:

1. The estimated protein abundance mean
2. The estimated protein abundance minimum and maximum
3. The estimated protein abundance median
4. The estimate protein abundance first and third quartiles

### Q. Do you observe any noticeable differences in the distribution of protein abundances between technical replicates under the same condition?

### Q. How about between conditions?

### Q. What does this tell you about whether the treatment had an effect?

### Q. Is it possible that there is affect we cannot observe by examining this population level change? If so how?

# (IV): Differential Protein Abundance from LFQ Mass Spectrometry Data Datasets

## (IV – 1) Standard Scale Renormalization of Protein Abundance

Explanation: Compute global mean and standard deviation separately for each group of samples (flattening all proteins × replicates).

* Z-scores each group independently: subtract mean, divide by standard deviation.

Why: This rescales each condition so that its overall distribution is standardized. It removes global scale differences that might come from batch effects or loading differences.

Caveat: Standardizing separately means the two groups are no longer directly on the same absolute scale, this could obscure true global shifts. There are other methods for normalizing in differential testing:

* Normalize all samples together (e.g., median normalization across all)
* Use ratio-based methods where differences are preserved without groupwise rescaling

## (IV – 2) Statistical testing (Welch’s t-test)

Explanation: Apply Welch’s t-test per protein, comparing the standardized abundance vectors between the two conditions.

Why Welch’s test? It does not assume equal variance between groups, a prudent choice for biological data where heteroskedasticity is common.

Note: There are only 3 replicates per group and degrees of freedom are low which typically leads to a high variance.

Note: Welch’s t-test is generally powerful when sample sizes are modest and distributions are approximately symmetric; the Mann–Whitney test is robust to non-normality but tests for difference in distribution location in a different way. It’s good practice to examine data (e.g., via violin/box plots) to decide if assumptions are reasonable.

## (IV – 3) Ranking and multiple testing correction

Explanation: Sorts proteins by raw significance (smallest p-value first) to prioritize candidates.

* Apply Benjamini–Hochberg false discovery rate (FDR) correction to control the expected proportion of false discoveries among proteins.

Why: Thousands of proteins may be tested; without correction, the number of false positives at p<0.05 would be unacceptably high. FDR balances discovery with error control.

Note: There are different FDR correction methods that rely on different assumptions. You should always check that the assumptions are not violated in your dataset before application of any statistical method.

### Q. What fraction of proteins had a significant change in their abundance?

## (IV – 4) Volcano plots

A volcano plot is a scatter plot that simultaneously displays both magnitude of change and statistical significance for each protein (or peptide) when comparing two conditions. It’s a rapid visual filter to highlight candidates that are both substantially and confidently different.

Axes of a volcano plot:

* X-axis: Fold change: This is usually reported as the, of the abundance ratio between the two conditions (positive means higher in treatment condition, negative means higher in control condition). It reflects effect size how much a protein’s abundance changes.
* Y-axis: from a statistical test (e.g., Welch’s t-test on replicate abundances). Higher values mean stronger statistical evidence against the null (i.e., more significant). Using the negative log makes small p-values (strong significance) appear at the top.

### Q. How many proteins have a significant change after treatment and a ?

### Q. How many proteins have a significant change after treatment and a ?

### Q. How many proteins have a significant change after treatment and a ?

### Q. How many proteins DO NOT have a significant change after treatment and a ?

### Q. How many proteins DO NOT have a significant change after treatment and a ?

### Q. How many proteins DO NOT have a significant change after treatment and a ?

### Q. which of these groups would you focus your efforts on investigating further to understand how the treatment is potentially affecting protein production and function?