NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science
Bachelor's Programme "Applied Mathematics and Informatics"

Research Project Report on the Topic:

Developing Accurate False Discovery Rate Control Methods

Fulfilled by:

Student of the Group БПМИ2310
Palienko Anastasiya Feodorovna

Assessed by the Project Supervisor:

Attila Kertesz-Farkas
Head of Research Lab
Faculty of Computer Science, HSE University

Moscow 2025

Contents

Annotation

Think of identifying new protein fragments as of solving a puzzle without looking at the picture on the box. This is actually what scientists face with de novo peptide sequencing — a method that identifies unknown peptides (protein building blocks) through mass spectrometry, without using existing databases as references. However, the possibility of incorrect findings makes it difficult to guarantee the accuracy of these annotations. But how can we ensure that the identifications are reliable?

The aim of this study is to develop precise false discovery rate (FDR) control methods for de novo sequencing. This is crucial in proteomics, where applications vary from detecting cancer cells to environmental researches.

Keywords

De novo peptide sequencing, tandem mass spectrometry, false discovery rate (FDR), peptide spectrum matches (PSMs)

1 Introduction

1.1 Research dictionary

- **Tandem Mass Spectrometry (MS/MS)**
  - an analytical technique that involves two stages of mass measurement with a step in between. Identifies and quantifies proteins and peptides by analyzing the unique fragmentation patterns of precursor ions, enhancing specificity and sensitivity in complex mixtures.
- **False discovery rate (FDR)**
  - statistical metric used to estimate the proportion of incorrectly identified peptides among all identified peptides. Ensures the reliability of the results by controlling the number of false positives.
- **FDR control**
  - a group of statistical methods that sets a threshold to control the rate of false positive results, thereby maintaining a desired level of confidence in the accepted findings.
- **De novo sequencing**
  - a method used in mass spectrometry to determine the amino acid sequence of a peptide without prior knowledge of its sequence

1.2 Problem statement

According to numerous studies and scientific articles, in case of de novo sequencing, newly identified peptides can be put into a reference database. The false discovery rate (FDR) associated with these de novo peptides is typically estimated through an additional database search. However, this approach has its pitfalls, such as the possibility of data correlation issues leading to biased algorithm. The purpose of this research is to assess reliability of the described methodology.

1.3 Subject area

The object of the study is the assessment of existing false discovery rate (FDR) control methods and the development of more accurate techniques for de novo sequencing. The subject

of this research is the application of various statistical approaches to improve the reliability of peptide identifications, focusing on reducing bias in the verification processes.

2 Literature Review

2.1 Peptide annotation & FDR

When an annotation method is applied on some spectral data, the result annotations are arranged from the highest score to the least likely accurate ones. However, not all spectra can be annotated with high confidence, and the remaining annotations are often most likely inaccurate. This is where the problem of determining how to precisely convey confidence in the annotations arises, and to evaluate the results of tandem mass spectrometry methods, we need a statistical approach. This is where the false discovery rate (FDR) comes from.

In terms of peptide annotations, the FDR measures the proportion of incorrect spectrum annotations (false discoveries) among a set of hypothesis tests (total annotation which has matching scores greater than or equal to the threshold t). The goal is to determine t for the desired level α:

$$FDR(t) = \mathbb{E}\left[\frac{D(t)}{T(t)} \,\middle|\, T(t) > 0\right] \cdot P(T(t) > 0), \tag{1}$$

where $D(t)$ represents the number of incorrect peptide spectrum matches (PSMs) with scores greater than or equal to t, and $T(t)$ is the number of correct PSMs with scores above t. This equation calculates the expected proportion of false discoveries (FDR) at a threshold t; multiplication by $P(T(t) > 0)$ adjusts the FDR based on the likelihood of having any true positives. In practice, FDR is often estimated using a method called the "decoy approach," where a set of decoy sequences (which are known to be false) is used to estimate the number of false positives, so then it can be calculated as:

$$\text{FDR} = \frac{\text{Number of False Positives}}{\text{Number of True Positives} + \text{Number of False Positives}}$$

2.2 De novo sequencing

In the field of proteomics, de novo sequencing became a game-changing approach that provided an alternative to traditional database-driven peptide identification methods. Unlike conventional techniques that rely on preexisting references, de novo sequencing enables the inference

of full-length or partial peptide sequences directly from experimental tandem mass spectrometry (MS/MS) data.

The main goal of a standard mass spectrometry procedure is to accurately identify peptides and proteins by matching observed spectra with theoretical predictions. Conventional database search methods involve evaluating candidate peptides against received spectra, leading to selection of the best peptide spectrum match (PSM). However, this approach has its limitations. It is clear that the method requires prior knowledge of the potential peptides present, missing unexpected options brought about by some mutations or modifications. In addition, the computational resources needed for comprehensive database searches are substantial, making the process time-consuming. De novo sequencing, on the other hand, overcomes these problems by enabling peptide identification only from experimental data. This method is beneficial in scenarios where broad genomic information is not accessible, as it can reveal novel peptide sequences that would otherwise go undetected.

2.3 PEAKS Online platform

The sensitivity is especially critical for immune peptidomics applications due to the difficulty of obtaining large samples and the complexity of the peptides. In this case, "sensitivity" actually refers to the capacity of mass spectrometry techniques to identify peptides with low abundance. To address this, recent research has suggested that the combination of different data acquisition strategies, such as Data Independent or Data Dependent (DIA / DDA) and different analysis techniques might improve the sensitivity of MS-based immune peptidomics. Data Independent Acquisition – mass spectrometry strategy, in which the most intense precursor ions are selected for fragmentation based on their intensity in the initial scan; Data Dependent – a strategy where all precursor ions within a certain mass range are fragmented, providing a comprehensive dataset of all peptides in a sample.

Presented in the article, the PEAKS streamlined platform combines three computational : spectral library search, database search, and de novo sequencing, which can be performed separately or used together in a workflow. Then a new spectral library with the list of peptides found using those techniques was constructed, followed by a final search. The final search of the whole dataset was performed against this study and is aimed at reconfirming the identified peptides and providing a unified global FDR. The false discovery rate of identified PSMs is calculated using a target-decoy approach, in which decoy peptides and spectra are generated by randomly permuting the peptide sequences and the corresponding fragment ions.

According to the authors, de novo peptides can be added to a new database as a result of de novo sequencing, and the FDR of the peptides can then be estimated through a further database search.

2.4 Target-Decoy Approach for Mass Spectrometry-Based Proteomics

To begin, a composite data set is created. Starting by obtaining a database of 'target' protein sequences corresponding to the analysed protein mixture. Next, a 'decoy' database is constructed to mimic the general characteristics of the target database; sequences must be properly labeled to distinguish them from target sequences in search results. Additionally, the ideal decoy database must meet the following criteria:

1 Distribution of amino acid: The decoy to target protein sequences should have a similar amino acid composition.

2 Protein length distribution: The lengths of the decoy proteins should match the length distribution in the target protein list.

3 Protein count: The decoy database should contain a similar number of proteins as the target database.

4 Number of predicted peptides: The decoy database should generate a comparable number of predicted peptides as the target database.

5 No shared peptides: There should be no predicted peptides in common between the target and decoy sequence lists.

Tandem mass spectrometry (MS/MS) spectra are typically searched against a single composite target-decoy database to identify peptide-spectrum matches (PSMs). Since decoy sequences do not exist in the biological sample and are artificially generated, all matches to the decoy database are considered false positives (FP). These matches provide a way to calculate the false positive rate in the dataset.

The FDR is estimated using the formula:

$$\text{FDR} = \frac{2 \times \text{Decoy hits}}{\text{Target hits} + \text{Decoy hits}}$$

The factor of 2 in the formula accounts for the assumption that the distribution of false positives between the target and decoy databases is equal.

2.5 Database-searching approach

Database searching is an essential element of large-scale proteomics. In general, this method can be described as a pattern recognition exercise that involves finding the entries in the database that are most similar to the query spectrum. The similarity is assessed by different statistical measures, for example, the correlation coefficient.

Most algorithms are based on concepts that were firstly used in SEQUEST – a MS/MS data analysis software used for protein identification. SEQUEST evaluates protein sequences from a database to generate a list of candidate peptides that can be derived from each, allowing accurate identification by comparison with experimental tandem mass spectra.

To determine a match between a spectrum and sequence, four basic approaches are used:

1 Descriptive algorithms predict how the peptides fragment in tandem on a mass spectrometer, and the predictions are then compared to experimental spectra to assess the quality of the match.

2 Interpretative approaches are based manually interpreting a partial sequence from an MS/MS and adding it to a database search with additional analysis of matches between the sequence and the spectrum

3 Stochastic models use probability-based approaches to predict the spectra generation and peptide fragmentation patterns by inferring fragment ion match probabilities from training data of spectra of known sequences.

4 Statistical models establish a relationship between MS/MS spectra and sequences, using this structure to determine the probability of peptide identification.

In combination, these approaches form the basis of modern database search techniques, ensuring precise identification of peptides in proteomics.

3 Research methodology

3.1 Prerequisites

3.1.1 Proteomics identification database: PXD004452

The dataset used in this research project is derived from a proteomic analysis of the HeLa – immortalized cell line of human origin, specifically from cervical cancer cells. Originated from

Homo sapiens, HeLa cells are known for their durability, enabling comprehensive applications in various scientific fields. Used cells are particularly distinguished for their genetic modifications, which include an anomalous number of chromosomes, ranging from 76 to 80 (compared to the normal diploid number of 46 in human cells).

In Project PXD004452, samples were processed using a combination of Lys-C and trypsin digestion, ensuring efficient protein breakdown into peptides. After that, peptides were subjected to a high-resolution, non-reversed-phase peptide separation at high pH, which led to the appearance of many fractions. These fractions were subsequently analyzed by short online chromatographic separations and fast peptide sequencing using Orbitrap tandem mass spectrometry. The described approach made it possible to identify an enormous number of peptides and proteins, providing a deep understanding of the HeLa cell proteome. The Data Processing Protocol involved analyzing the raw LC–MS/MS data using MaxQuant version 1.5.3.6 with the Andromeda Search engine, searching against the complete human Uniprot database to ensure accurate identification of peptides and proteins. The explained proteomics strategy resulted in the identification of about 584,000 unique peptide sequences and 14,200 protein isoforms, corresponding to approximately 12,250 protein-coding genes.

The dataset provides an opportunity for developing new false discovery rate control methods or validating already existing ones. Its inclusion of various post-translational modifications makes it an ideal candidate for evaluating the effectiveness of various statistical approaches in handling false positives. As a result, it is directly relevant to the subject of my research.

3.1.2 Crux tool and Tide searching engine

An important method that will be used again and again in this research is the database-searching approach. The main ideas of the method have already been described above, but in this part we will focus on specific used algorithms.

The Crux Mass Spectrometry Analysis Toolkit is a project that aims to provide scientists a set of analytical tools for interpreting protein mass spectrometry data. The updated Crux v2.0 incorporates Comet and Tide, two search engines that perform SEQUEST-style database searches. The Crux software offers different methods of confidence validation, including false positive rate and calculation of accurate p-values for the Tide search engine.

In this research, the Tide searching engine is used. It separates the search into two phases: peptide indexing, where the index of peptides is stored on disk before any scoring, and the actual search. These steps make Tide-index approach fast and memory efficient by reducing computation

if searching the same database several times is needed.

3.1.3 Q-values

In addition to the PXD004452 dataset, tailor.assign-confidence.target was used for this study. It was obtained with a database-searching method and the results were validated via target-decoy analysis and has 591950 annotations. To ensure that annotations in tailor.assign-confidence.target are correct, additional data filtering is required. The following paragraph explains the q-value method, which sorts sequences by how confident we are in their accuracy.

In statistical analysis, the p-value is a key concept. A p-value represents the probability of obtaining a test result as extreme as the observed one, assuming the null hypothesis – the default assumption that there is no real effect or no difference between groups – is true. In contrast, q-value is used to control the false discovery rate in multiple hypothesis testing (i.e., "Is this peptide a true match?"), which is especially important in peptide annotation. Unlike p-values, which only indicate the probability of observing such data under the null hypothesis, a q-value of 0.05 directly tells us the FDR at which we can trust the result. This helps balance sensitivity (finding real peptides) and specificity (avoiding false positives).

The pseudocode for calculating q-values within the target-decoy approach is presented in Algorithm 1. Or a more general description of the method:

1 Compute p-values for each peptide spectrum match (PSM), which indicate how likely a match is to be false under the null hypothesis.

2 Sort the p-values from smallest to largest.

3 Estimate the FDR at each threshold. This is done by comparing the number of accepted PSMs to the expected number of false positives.

4 Assign a q-value to each PSM, representing the minimum FDR threshold at which it is still accepted.

Returning to our tailor file, the code below produces a sorted data frame that may be used to evaluate other data. Here we set a threshold of 0.01:

```
df = pd.read_csv("tailor.assign-confidence.target.txt", delimiter="\t")
df = df[df['tdc q-value'] < 0.01]
```

Algorithm 1 Q-value calculation using Target-Decoy Analysis

Require: List of spectra with annotations and their scores:

$$\langle s_1, h_{1j}, c_1 \rangle, \ldots, \langle s_n, h_{nj}, c_n \rangle$$

Where h – the identification of the peptide hypothesis assigned to the i-th spectrum; c – numerical value associated with the i-th spectrum that quantifies the quality or confidence of the peptide annotation

Ensure: Q-values computed for each spectrum:

$$\langle s_1, h_{1j}, c_1, q_1 \rangle, \ldots, \langle s_n, h_{nj}, c_n, q_n \rangle$$

Sort spectra in decreasing order based on annotation scores
Initialize counters (# = number of annotations):
$\#Target \leftarrow 0$
$\#Decoy \leftarrow 0$
for each spectrum $i = 1$ to n **do** ▷ n = total number of analyzed spectra
    if h_{ij} is a target peptide **then**
        $\#Target \leftarrow \#Target + 1$
    else
        $\#Decoy \leftarrow \#Decoy + 1$
    end if
    $q_i \leftarrow \frac{\#Decoy}{\#Target}$ ▷ Compute q-value
end for
for each spectrum $i = n - 1$ to 1 **do** ▷ Ensure q-values are non-decreasing
    if $q_{i+1} < q_i$ **then**
        $q_i \leftarrow q_{i+1}$
    end if
end for

3.1.4 Data processing

The next step involved aligning sequences from `pepnet_PXD004452` to `tailor.assign-confidence.target.txt` identifications. This involved merging all TSV files from the PXD004452 dataset into a single DataFrame. This is achieved using the provided bellow `merge_tsv_files` function. It takes a directory path as input and reads all files, then adds a file_key column, which is derived from the filename by removing the extension. Then, using regular expressions, function extracts the scan number from the TITLE.

```
def merge_tsv_files(directory):
    '''
    Args: directory (str): Path to the directory containing TSV files.
    Returns: pd.DataFrame: Merged DataFrame containing all TSV data.
    '''
    all_files = [f for f in os.listdir(directory) if f.endswith(".tsv")]
    df_list = []

    for f in all_files:
```

```
        df = pd.read_csv(os.path.join(directory, f), sep='\t')
        df["file_key"] = f[:-4]
        df["scan"] =
        ↪  df["TITLE"].str.extract(r"scan=(\d+)").astype(float).fillna(-1).astype(int)
        df_list.append(df)

    return pd.concat(df_list, ignore_index=True) if df_list else pd.DataFrame()
```

For example, row in `20150410_QE3_UPLC9_DBJ_SA_46fractions_Rep1_1` from `pepnet_PXD004452` looks like:

TITLE	20150410_QE3_UPLC9_DBJ_SA_46fractions_Rep1_1.5.5. File:"", NativeID:"scan=5"
DENOVO	RFDNR
Score	0.9957
PPM Difference	-7.9395256
Positional Score	[0.99997246, 0.9999999, 0.999995, 0.9995622, 0.9980884]

The `merge_with_metadata` is used to preprocess metadata and merge it with another data based on two key columns: file_key and scan. The file column in the metadata DataFrame is preprocessed to create a file_key; this is done by removing the .mzML extension from the filename (if present) using Python's os.path.basename and string slicing.

```
def merge_with_metadata(merged_df, df):
    '''
    Args: merged_df (pd.DataFrame): DataFrame containing merged TSV data.
          df (pd.DataFrame): Metadata DataFrame containing 'file' column.
    Returns: pd.DataFrame: DataFrame resulting from an inner join on 'file_key' and
    ↪  'scan'.
    '''
    df = df.copy()
    df["file_key"] = df["file"].apply(lambda x: os.path.basename(x)[:-5] if
    ↪  x.endswith(".mzML") else x)
    return merged_df.merge(df, on=["file_key", "scan"], how="inner")
```

After running the code, `merged_df` will contain De Novo search results from `pepnet_PXD004452` files (DENOVO sequence, Score, PPM Difference, Positional Score as shown above) and some additional information from tailor.assign-confidence.target.txt (Xcorr score, Q-values, etc.).

It also can be seen that some sequences from the Tailor file contain square brackets ([ ]). This appears due to certain mass shifts resulting from post-translational modifications or chemical adducts. For example, `K[14.0157]STGGK[42.0106]APR` means that the first lysine (K) has a mass shift of +14.0157 Da (dalton unit), likely due to methylation, while the second one is acetylated, corresponding to a mass shift of +42.0106 Da. There is no point in going into detail about the

various modifications, because this information is not essential for the study. For further analysis, the brackets and everything in between was simply removed:

```
merged_df['updated sequence'] = merged_df['sequence'].str.replace(r'\[.*?\]', '',
↪  regex=True)
```

The next step is to filter incorrectly identified sequences and put them to `only_incorrect_df` data base. The purpose of this research is to study accurate methods for estimating FDR, so it is important to focus on exploring false discoveries.

3.2 Experimental Data Analysis

3.2.1 FASTA format

The FASTA format – industry standard in bioinformatics – is a text-based format for representing either nucleotide sequences or amino acid (protein) sequences. Due to its simplicity, FASTA representation makes it simple to use text-processing tools to parse and analyze sequences. In the target-decoy approach, decoy sequences are often generated by reversing or shuffling target sequences. FASTA format makes it easy to manipulate sequences programmatically to create decoys.

A sequence begins with a greater-than symbol ('>') followed by its description. The next lines contain sequence representation, where nucleotides or amino acids are denoted by single-letter codes. The following code snippet illustrates the implementation for storing FASTA peptide sequences in `output_fasta_file.txt` file."

```
def convert_to_fasta(output_path, df, column_name):
    with open(output_path, 'w') as output_file:
        for index, row in df.iterrows():
            identifier_line = f">{row['file']}|{row['scan']}|{row['charge']}\n"
            sequence_line = row[column_name] + "\n"

            output_file.write(identifier_line)
            output_file.write(sequence_line)

    return(f"FASTA file saved!")
```

The function uses file name, scan number and charge state of the precursor ion (the ion that was fragmented to produce the MS/MS spectrum) as a description. In case of incorrect annotations from `PXD004452`, first 3 annotations from generated FASTA file look like:

```
>/hdd/data/PXD004452/20150410_QE3_UPLC9_DBJ_SA_46fractions_Rep1_40.mzML|570|3
KRNQNSQISTEK
```

>/hdd/data/PXD004452/20150410_QE3_UPLC9_DBJ_SA_46fractions_Rep1_40.mzML|980|3
GHVQDPNDRR
>/hdd/data/PXD004452/20150410_QE3_UPLC9_DBJ_SA_46fractions_Rep1_40.mzML|1011|2
KSTGGKAPR

3.2.2 XCorr score

The output TDA files contain XCorr (cross-correlation) score value. The xcorr function originally appeared in signal processing and quantified the similarity of two signals by comparing one signal to time-shifted versions of the other, measuring their alignment at different lags. In mass spectrometry, this principle adapts to compare theoretical peptide ion masses with observed spectral peaks across m/z shifts. If the PSM is correct, it is expected that the theoretical peptide and observed ions align on the m/z axis; otherwise, only a small number of them may align if there is a shift in the m/z axis of either. The xcorr estimator takes advantage of this by using local m/z shifts as background, effectively distinguishing true matches from random alignments.

Here is how XCorr score calculated by a fast SEQUEST Cross Correlation Algorithm:

$$\text{XCorr} = x_0 \cdot y', y' = y_0 - \frac{\sum_{\tau=-75,\tau\neq 0}^{+75} y_\tau}{150} \tag{2}$$

Here, x_0 represents the theoretical mass spectrum of a peptide from the database, while y denotes the acquired (experimental) mass spectrum at different m/z offsets, τ. When $\tau = 0$, y corresponds to the original acquired mass spectrum. The equation of y' subtracts the mean of 150 shifted versions of the experimental spectrum (excluding $\tau = 0$) from the original acquired spectrum, effectively reducing background noise and random correlations by penalizing unmatched annotations. These shifts are carried out within a window of ± 75 m/z bins, capturing a wide range of possible misalignments. Since the dot product measures similarity, this approach ensures that only meaningful peptide matches receive high XCorr scores. However, it is important to note that the xcorr function is uncalibrated, meaning that a good score for one spectrum may not be a good one for another.

3.2.3 Msconvert & valid indexing

Msconvert is a command-line utility by ProteoWizard for converting between various mass spectrometry data formats, including RAW, mzML and mgf. Initially, a file named `only_incorrect.tsv` was created containing just two columns: file and scan, which list the incorrectly annotated files and their corresponding scan numbers. Next, a directory `answers` was

generated using msconvert and a custom bash script, containing data files that represent the raw spectra of particular targeted scans.

```
file        scan
/PXD004452/20150410_QE3_UPLC9_DBJ_SA_46fractions_Rep1_40.mzML        569
/PXD004452/20150410_QE3_UPLC9_DBJ_SA_46fractions_Rep1_40.mzML        979
/PXD004452/20150410_QE3_UPLC9_DBJ_SA_46fractions_Rep1_40.mzML        1010
```

A key challenge in this process was that msconvert saves filtered spectra using the name of the original file. As shown in the example above, since multiple scans from the same file needed to be extracted, this default behavior would have led to overwriting. To get around this problem, some renaming manipulations were implemented, as you can see in the script below.

```
#!/bin/bash

TSV_FILE="/tear/PXD004452/only_incorrect.tsv"
INPUT_DIR="/tear"
OUTPUT_DIR="./answers"

mkdir -p "$OUTPUT_DIR"

TOTAL_FILES=$(tail -n +2 "$TSV_FILE" | wc -l)
PROCESSED_FILES=0

tail -n +2 "$TSV_FILE" | while IFS=$'\t' read -r name scan; do
    INPUT_FILE="$INPUT_DIR$name"
    FILTER="index $scan"

    msconvert "$INPUT_FILE" --mgf -o "$OUTPUT_DIR" --filter "$FILTER"

    ORIGINAL_MGF="$OUTPUT_DIR/$(basename "$name" .mzML).mgf"
    OUTPUT_NAME="$(basename "$name" .mzML)_scan$((scan + 1)).mgf"
    OUTPUT_FILE="$OUTPUT_DIR/$OUTPUT_NAME"

    if [[ -f "$ORIGINAL_MGF" ]]; then
        mv "$ORIGINAL_MGF" "$OUTPUT_FILE"
    else
        echo "Error: File $ORIGINAL_MGF not found!" >&2
    fi
```

```
    PROCESSED_FILES=$((PROCESSED_FILES + 1))
    echo "Processed files: $PROCESSED_FILES of $TOTAL_FILES"
done

echo "All files processed!"
```

Next, to run the database-search method on the data, it is necessary to combine all the files from the answers folder into a single large `mgf` file. This is done using a fairly simple command in the terminal:

```
find answers -name "*.mgf" -exec cat {} + > concatenated_incorrect.mgf
```

3.2.4 PPM Difference

In mass spectrometry, PPM stands for parts per million. It is a unit of measurement utilized to express the mass resolution or mass accuracy of the mass spectrometer; ppm is used to calculate the error between an observed mass and the theoretical value. Since the theoretical value is taken as fact, it is the divisor in the corresponding formula:

$$\text{PPM Difference} = \frac{\text{Theoretical m/z value} - \text{Experimental m/z value}}{\text{Theoretical m/z value}} \times 10^6, \tag{3}$$

where m/z stands for mass-to-charge ratio: $\text{m/z} = \frac{\text{mass number}}{\text{charge number}}$

A large PPM difference indicates that the annotation of peptides is likely incorrect. Therefore, to remove possible false matches, an additional filter was applied to the data, selecting only sequences where the absolute value of PPM difference is less than or equal to 10.

```
only_incorrect_df_ppm = only_incorrect_df[abs(only_incorrect_df['PPM Difference']) <=
↪  10]
```

3.2.5 Edit distance

We can determine whether a peptide sequence is correct based on a perfect match. But sometimes the algorithm is wrong by only a small number of characters, and such sequences can be considered separately. Edit distance as a metric of similarity between two strings helps in such cases. It quantifies the difference between two sequences by measuring the minimum number of single-character edits required to transform one string into another. The Levenshtein distance

specifically allows for three types of edits: insertion, deletion, and substitution.

The Levenshtein distance between two strings s_1 and s_2, with lengths $|s_1|$ and $|s_2|$, is defined using head(x) to denote the first character of a string x and tail(x) to represent the string without its first character:

$$\text{lev}(s_1, s_2) = \begin{cases} |s_1| & \text{if } |s_2| = 0, \\ |s_2| & \text{if } |s_1| = 0, \\ \text{lev}(\text{tail}(s_1), \text{tail}(s_2)) & \text{if } \text{head}(s_1) = \text{head}(s_2), \\ 1 + \min \begin{cases} \text{lev}(\text{tail}(s_1), s_2) & \text{– } \textit{deletion} \\ \text{lev}(s_1, \text{tail}(s_2)) & \text{– } \textit{insertion} \\ \text{lev}(\text{tail}(s_1), \text{tail}(s_2)) & \text{– } \textit{substitution} \end{cases} & \text{otherwise} \end{cases}$$

Consider the peptide sequences $s_1 = \text{QRLENSQLSTEK}$ and $s_2 = \text{KRNQNSQISTEK}$.

To calculate the Levenshtein distance between these two sequences, we can visualize some necessary edits:

1 Change 'Q' in s_1 to 'K': **Q**RLENSQLSTEK → **K**RLENSQLSTEK

2 Change 'L' in s_1 to 'N': KR**L**ENSQLSTEK → KR**N**ENSQLSTEK

3 Change 'E' in s_1 to 'Q': KRN**E**NSQLSTEK → KRN**Q**NSQLSTEK

4 Change 'L' in s_1 to 'I': KRNQNSQ**L**STEK → KRNQNSQ**I**STEK

Thus, the Levenshtein distance $\text{lev}(s_1, s_2) = 4$.

The next step involves calculating the distance for different peptides. But first, let's generate reversed decoy sequences: leave the first and the last characters in the sequence of correct annotation (from tide file) and do a reversal of all characters between them.

```
def create_decoy(sequence):
    if len(sequence) < 2:
        return sequence
    return sequence[0] + sequence[-2:0:-1] + sequence[-1]
```

Another method of generating decoy sequences is still leaving the first and the last characters but randomly shuffling everything in between:

```
def create_random_decoy(sequence):
    if len(sequence) < 2:
        return sequence
    shuffled = list(sequence[1:-1])
    random.shuffle(shuffled)
    return sequence[0] + ''.join(shuffled) + sequence[-1]
```

Using the code below, the edit distance between the DENOVO, Tailor (correct) and decoy sequences is calculated.

```
import Levenshtein as lev

merged_df.loc[:, 'edit distance pepnet-tailor'] = merged_df.apply(
    lambda row: lev.distance(row['DENOVO'], row['updated sequence']), axis=1
)

merged_df.loc[:, 'edit distance tailor-decoy'] = merged_df.apply(
    lambda row: lev.distance(row['updated sequence'], row['decoy (reversed)']),
    ↪  axis=1
)

merged_df.loc[:, 'edit distance tailor-decoy-shuffled'] = merged_df.apply(
    lambda row: lev.distance(row['updated sequence'], row['decoy (shuffled)']),
    ↪  axis=1
)
```

4 Results

The approach under investigation involves taking de novo peptide sequences and subjecting them to an additional database search to estimate the false discovery rate. One of the concerns is whether this approach introduces bias – that is, whether it systematically "forces" false matches by correlating de novo identifications with the database search results.

Under ideal (unbiased) conditions, one would expect that false peptides, being truly random or non-existent, would be preferentially grouped with lower scores, while any true positives (or even high-quality false hits) would have noticeably higher scores. However, if the algorithm is biased, it will tend to produce inflated scores for certain de novo peptides even when they are incorrect; the search engine may assign them scores that are artificially high, effectively "forcing" false matches. In the plots comparing XCorr for target and decoy searches, a biased algorithm would result in significant overlap between the two. This means that the histogram (or density plot) for decoy peptides would spread into the higher-scoring range, similar to the target peptides.

Alternatively, if the algorithm "forces" false matches, one might notice that false peptides receive scores that are artificially close to the scores of the target peptides (even after applying filtering criteria e.g., PPM difference $\leqslant 10$ or edit distance cutoffs).

Previously the `only_incorrect_df` was created to contain all incorrect annotations. Further analysis is based on applying various filtering methods to this dataframe.

4.1 Only incorrect annotations: first search

After running the `convert_to_fasta` function for `only_incorrect_df`, the created FASTA file is subjected to a database-searching approach, where mass spectrometry data is searched against both a target database of known protein sequences and a decoy database of reversed or shuffled sequences. Two output files are produced by this process:
`wrong-pepnet-peptides.tide-search.target`, which contains the identified target peptides, and `wrong-pepnet-peptides.tide-search.decoy`, which contains decoy sequences used for error estimation.

Extracting xcorr scores from `wrong-pepnet-peptides.tide-search.target` and `wrong-pepnet-peptides.tide-search.decoy`, the distribution can be visualized. As can be seen, the plots for target and decoy distributions have a significant overlap. This overlap occurs because the incorrect annotations initially do not contain distinguishing features that would allow for clear separation; instead, both target and decoy sequences exhibit similar scoring characteristics, resulting in their nearly identical distributions. In other words, since both target and decoy peptides are essentially "random" in terms of accuracy, their score distributions behave this way.

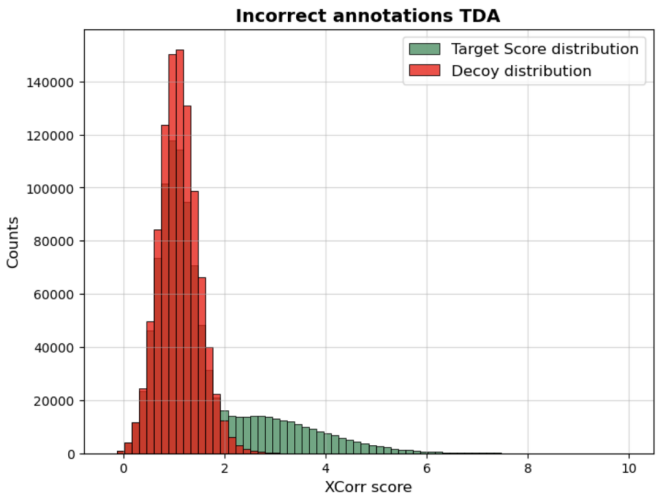


Figure 4.1: Distribution of XCorr scores of incorrect annotations

4.2 Only incorrect annotations: second search

To estimate the FDR of de novo peptides, additional database search is required. The Crux Tide approach was applied to the FASTA file (incorrect de novo peptides) and the RAW file (mass spectra data). The database search attempts to determine which peptide sequences best explain the observed fragmentation spectra in the RAW data, and the best match is chosen based on a scoring function XCorr. As mentioned before, Tide search produces two files: `target` and `decoy`.

Target database results:

1 Contains matches between real peptides from FASTA file and spectra from the RAW file

2 The sequences in this file are the incorrect de novo peptides that were extracted earlier

Decoy database results:

1 Contains matches between randomized or shuffled peptides (decoys) and spectra from the RAW file

2 Decoy sequences are not real peptides, but are generated to estimate false positive matches and the false discovery rate

The distributions of XCorr scores can be visualized again. As can be seen from Figure 3.3 (a), the overlap between the target and decoy plots compared to previous search (Figure 3.2) is quite small.

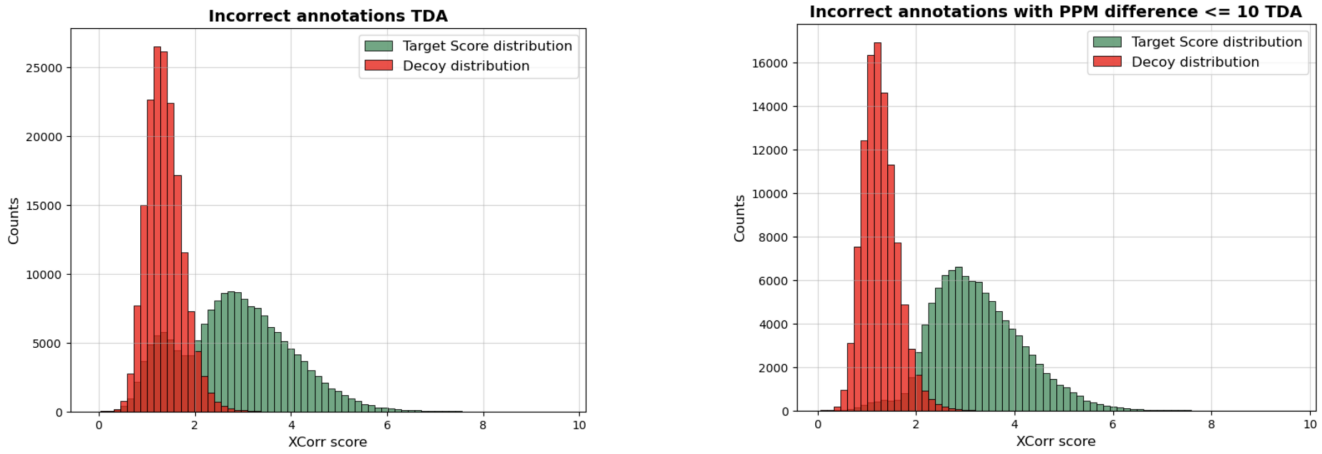
The decoy distribution is mostly at low XCorr scores, meaning that the search engine does not often mistake random sequences for valid peptides. The target distribution mostly has higher XCorr scores, showing that real (though incorrect de novo) sequences still produce significantly better matches than decoy ones. However, the distributions exhibit a small region of overlap: this means that some false positives (decoys) received similar scores to true targets, indicating potential misassignments.

4.3 PPM difference filtering

This section focuses on data where the ppm difference is less than or equal to 10.

The steps described above are now repeated: creating new FASTA file, but now for `only_incorrect_df_ppm` data frame, `mgf` file with raw data, and then running the database-searching algorithm on these two files. The distributions of the obtained XCorr values are presented in Figure 3.3 (b). It is easy to see that the plot is not very different from the previous one (Figure 3.3 (a)), except that overlap is no more presented. This is due to the fact that selected

false positives were closer to the theoretical mass and thus more likely to be mistaken as correct. This suggests that after filtering, incorrect annotations with better ppm values behaved more like true targets, reducing similarity between target and decoy scores. Since sequences with PPM Difference > 10 cannot be true, the algorithm on filtered data "uncertainly" identifies such sequences as true, which is why the overlap is reduced.



(a) Without PPM Difference filtration

(b) With PPM Differences $\leqslant 10$

Figure 4.2: Distributions of XCorr scores for incorrect annotations

4.4 Edit distance filtering

First, let's plot the edit distance distribution for Pepnet (DENOVO) sequences and Tailor ones. Note that the y-axis is presented in logarithmic scale to enhance interpretation of the results.

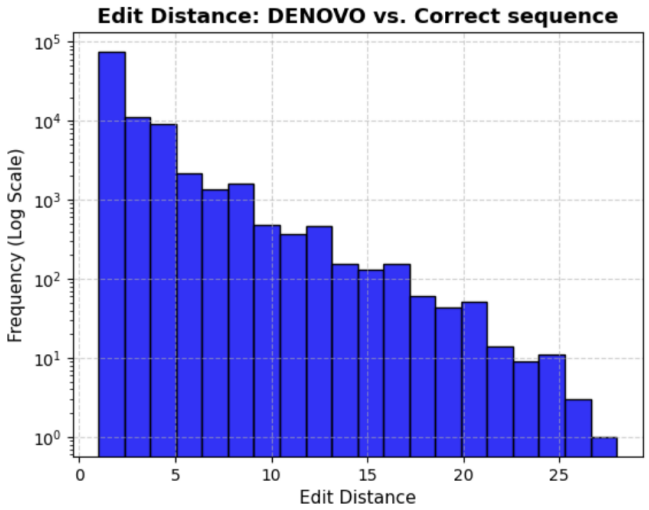
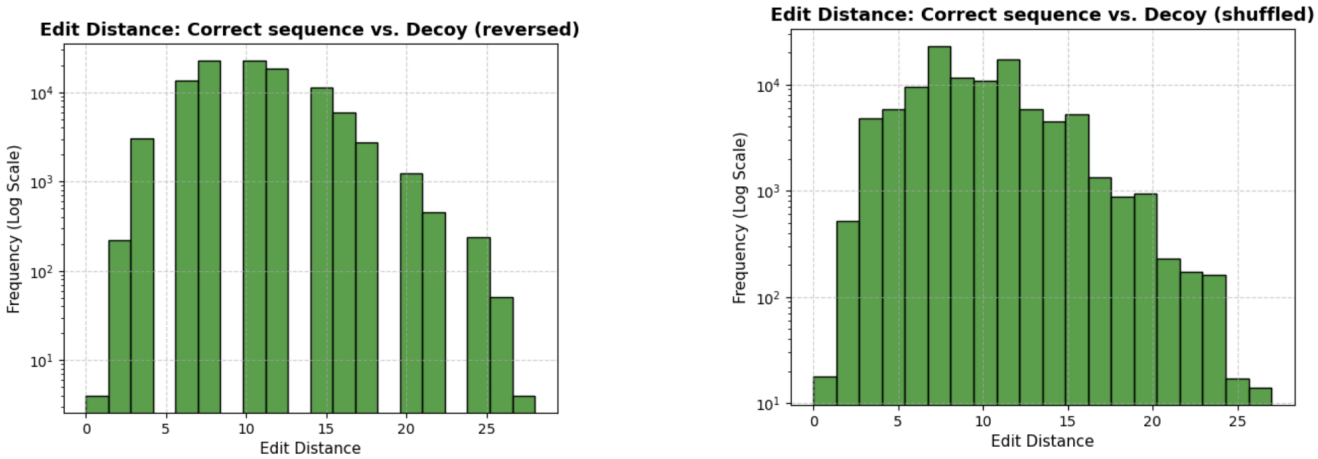


Figure 4.3: Distribution of XCorr scores of incorrect annotations

Now let's look at the distributions of edit distance between Tailor and different types of decoy sequences on Figure 3.4. Overall the distributions are similar, except that the left plot

has gaps in some values. This happens because the reversed sequence tends to produce specific patterns of differences that are not evenly distributed across all possible edit distances. The right plot (Figure 3.4, (b)) shows that random shuffling does not completely destroy all internal relationships and creates a greater variety of results.

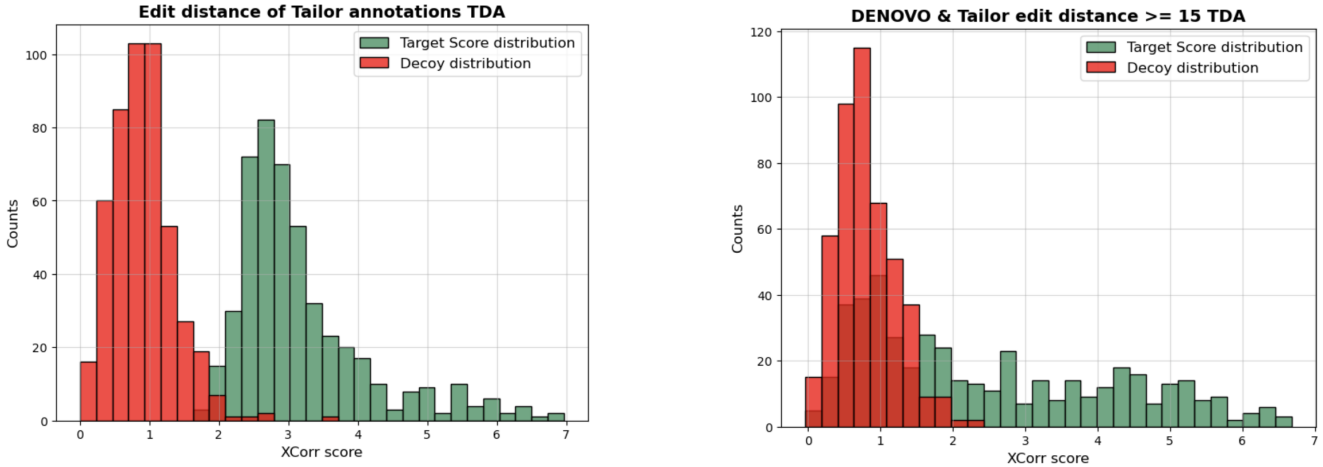


(a) Decoy (reversed)-Correct sequences

(b) Decoy (shuffled)-Correct sequences

Figure 4.4: Edit distance distribution between different sequences

Lets select annotations where edit distance between Pepnet and Tailor sequences $\geqslant 15$ (Table 6.1 in the Appendices). Intuitively this means that the sequence detected by the DENOVO algorithm is very far from the truth. This process results in a new dataset, consisting of 478 annotations. Now, repeating the algorithm with the launch of the database searching approach on FASTA and mgf files, it is important to pay attention to two plots (Figure 4.5): not only the distribution of DENOVO values, but also Tailor.



(a) Tailor FASTA file

(b) PepNet FASTA file

Figure 4.5: Distribution of XCorr scores with edit distance filtration for Tailor and PepNet sequences

In the left histogram (Figure 4.5 (a)), the FASTA file contains sequences that are correct and well-validated. Even if the edit distance between the de novo annotations and the Tailor sequences is large, the searching algorithm can still find reliable matches, and this results in a higher Corr score distribution. If the database contains sequences far from the true peptides, the scoring algorithm finds matches that are less reliable and result in lower XCorr scores. That is why the significant overlap in Figure 4.5 (b) can be seen.

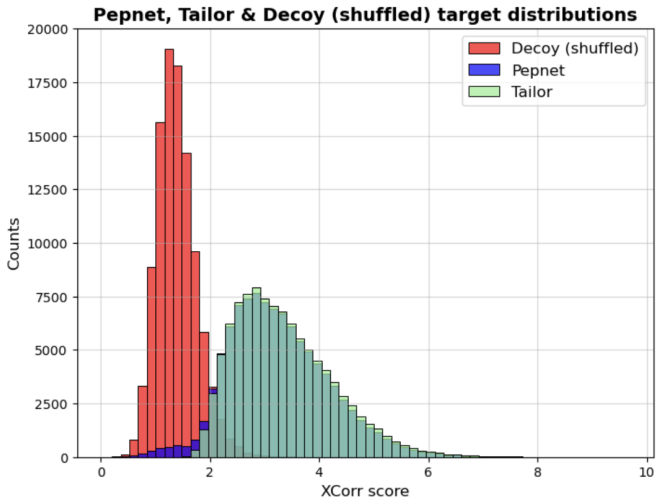


Figure 4.6: Distribution of XCorr scores of incorrect annotations

To check the distributions of Pepnet, Tailor and Decoy (shuffled) sequences, a database-searching approach was run on only incorrect annotations. In this case, only filtering by qvalues and ppm difference was taken into account. In Figure 4.6 can be seen the distribution of XCorr scores only for target files. It is clear that the behavior of Pepnet and Tailor sequences is similar, while Decoy bins are at a much lower xcorr value; this is explained by the fact that Decoy has artificial, non-existent sequences, so the algorithm is very uncertain about determining them as correct.

5 Conclusion

Through rigorous experimental analysis and the application of various statistical and computational methods, this study demonstrates that FDR estimation using secondary database search remains valid, addressing the core problem associated with potential algorithmic bias. Various filtering approaches were tested, including extracting data with ppm difference and edit distance within certain limits. Importantly, the algorithm did not demonstrate any tendency to force matches to incorrect annotations, nor did it interfere with accurate FDR boundary determination. However, despite the approval of the method, we still lack an understanding of why no correlation between the data appears during the operation of such an algorithm. Future research will aim to expand validation approaches to further enhance the accuracy and reliability of FDR control in proteomics, supporting critical applications in biomedical diagnostics and beyond.

Appendices

You can find implemented code for the research project on GitHub: **FDR Analysis for Tandem Mass Spectrometry** by coranmiel.

Table 5.1: Peptides with DENOVO, correct sequences Edit distance $\geqslant 15$

DENOVO sequence	Correct sequence	Decoy (shuffled)
RVLGQLHGGPSSCSATGTNR	EIQTRSASPSNIKAQFR	ESITRNSPQIAAFQSKR
APQEAETAEREEESDPPPEEEK	APQEVEEDDGRSGAGEDPPMPASR	AEDPESAAGSEQPMEDGVDRGPPR
HLRPEHLDSKEKHLLNGV	QKYPHLQVIGGNVVTAAQAK	QPLKQGAVHVNVGAITQYAK
RGLYYYQQPPPPSSSSSRPLLR	RAYETMAGAGVPFSANGRPLASGIR	RSTNAMYPVARELAAIAGGPFSGGR
RLRCCGGDNTPLVCNGLAHGVASGSR	MSLMSTATFLTSKDEGLKATTTDVR	MVKADTDAMTFSSLTLTTSGLETKR
SSPRRPDDTGDCCVAYGGVEVVR	MSSCPHVSPAGILCVADQCHGLR	MHLHGALSIVVASGPCSDQPCCR
KAEEELSSPLSLLLLLSCTACHHESFSR	KPVRVPAEPQTMSVLISCTTCHSEFPSR	KVETSIASCTPVVHSSFPPTLMPQCERR
SALFAQLNQGESLTHALK	RATGRPRCDLLQAMPR	RCLPPQDMTALGRRAR
SALFAQLNQGESLTHALK	RATGRPRCDLLQAMPR	RDTRCLQGLRAPMPAR
GNNDLLTLDQQQPPDVLTQHLK	DPALPTLLNPKTLPSGLDDYPR	DPAPLLSGLTYLPPKLDNPTDR
DGVALTALDQDHVAVLGSPLAASK	QNVHMLNKGIQAGNLEIVNGAK	QEMGHNNANGNAGKVQLILVIK
RTLPLFDMPTTTALPWSEGDLK	LVSDEMVVELIEKNLETPLCK	LESMPELNVEVLVTLEIKDCK
DYDPPGELPPAQSVVVPSSLLLALRR	GGKPGLTIQAPQLEVSVPSANIEGLEGK	GVGQSSEIGLQPLEELPTIGVAKAPGNK
DLQQQYLQLDQQLTELHPLK	NNLTILQRYMSSKIPAVTYPK	NPPSSIYLYKRATQMLVNTIK
ATDFFSHHPDPLTTTDEVSLCCSK	ATDFFGGILMNNTTDVEMAAIDCDK	AEFVTDCDINDTTMALMIFDGGANK
VGTCLLEETTVNKEHEHQGVVLAAANNLK	VGTDLLEEEITKFEEHVQSVDIAAFNKI	VKEIETAEFIDVEAQTLVGESNDLHFKI
KKTYYALQGTTEVQQSSGLHDQAAVVLFK	LKPYFLTDGTGTVTPANASGINDGAAAVVLMK	LDKTNGLSVLAMAYGFVPAGNAGTVPIADTTK
QLQNNLPPQNPVVYYLLLK	LLEAQIASGGVVDPVNSVFLPK	LVGSPSDAALVPEQVGNILFVK
KQARALNNVPLSDDPK	LETNLHLLGATGIEDR	LHETGLLNIGDTELAR
CLYRRLDVGAVVLGHVLGGR	RYLEGTGSVAGVVLLPEGHKK	RGESLKLAGPGGLVVHEVYTK
LHGLNLNYNNNLCGNYYYRGPK	IHTGEKPYECNECGNAFYVKAR	IGNCYNAKPGVYACKHTEEFER
LEALSSAHLLDQQRANNFR	ITVNSSSLSQDDKINKTYR	ITDKKVDSNQTNSLSYISR
EFWRRHHLLLLVAMSDAKK	HLEEGRLYPPLNTIRDVSLK	HGVLYDNERPTLILRPSLEK
ELEAHVDQLTEMAAVMRK	LVKGHAYSVTGAKQVNYR	LANKGVVQYYVGTAKSHR
NNQFEALLQYADPVSAQHAK	NQVVHVPLEAYDSSLPLIK	NLHIQSPSELDALVYVVPK
TGASYYTSHEHHHLDSGSSSVVGLLKK	TGASYYGEQTLHYIATNGESAVVQLPK	THQLQGTVAIPLSYSGTVAGAEYYENK
PGGLLLGDVAPNFEADTTVGR	KGGPGPASARPSESKEMTGAR	KEGSGAAEGGPPKSTMSPRAR
PGGLLDPFSEYCDDEPPEAALARRAYAAAR	QFRPPMGGYCIEFPAGLIDDGETPEAAALR	QDEAPAMRPLYPLCTADGFIAFEGGPIEGR
PNSPVLLEDPVLCALAK	FGFEIIETKKNYYK	FYKTGEEIKNYFIK
GGVLLLGDVAPNFEANTTVGR	LAGGNDVGIFVAGVLEDSPAAK	LNPFVAEDLAVAGDAVIGSGGK
FFDLEPPDTEADLQFRPR	MLLEASTKPEMSTVINNTR	MIAKTEPESSNTVMLNTLR
DWNPKTPPTFLECLAENCVGYCGADLK	SANPAVSKDFSSHDEINNYLQQIDQLK	SDLLSSVYASIPQFIQAHNEDDNKNQK
VAEDLHGLEDLHGRFSLSK	IGEHTPSALAIMENANVLAR	IHALEALMPNNASTIGVEAR
SALFAQLNQGESLTHALK	RATGRPRCDLLQAMPR	RRGLLMQCRDTAAPPR
RTTFACGGSSQTSVNARSADVLGR	TNHLVTVEGGWPQFGVGAEICAR	TGVQGACVIFVWLGENAEGTPHR
AWVWDTHADFADECPKPELLALR	KETELLGSFSKNESVPEVEALLAR	KESEEALGNLLPAVEFKVSSTLER
HLEVELLGDQYGNLLHLYER	AGSHGRRSPGGGSEANGLALVSGFK	AGGPRNGHGLERSSSAVSLAGGFGK
LPLVQYEVNFNNGLECQAYVK	GYYEVTPPTLVQTQVEGGATLFK	GLGTPQTYQEVLEYATGTVFPVK
APTLVESEELLPSGLDHPVFPK	QLFVLAGSAEEGVMTPELAGVIK	QAIVMFLGSGALPEVVTEEGLAK
TLNPDSDGAYADVVAAALLSGDK	GPAVGIDLGTTYSCVGVFQHGK	GDCFGVVPGVLGQSAGIHTYTK
LGDEEDVALDDSLPDFDDLAAFF	LLAPDMFTESDDMFAAYFDSAR	LADDFDFALDASMMYPAFTESR
SLGSVQAPPPPESDREMMSVRAAAGASR	SLGSVQAPSYGARPVSSAASVYAGAGGSGSR	SSASGSSVGAVASVGPGPRSGYALAASYQGR
VTHAHATVNPYGVHHLHVGGGR	VTHAVVTVPAYFNDAQRQATK	VTHAADTQNAAPTFVQVVRYK
AGPHYVNLQQGGGTRFFDGTGGK	VPSGFYVGSYVHGAMQSPSLHK	VFYAVGYQHGSSMSHGSPLVPK
TNFAPCPPPPPRRPPGLAAHTTQPPPFFK	TNFVQPMPGEGLRPSLPTQAHTTQPTPFK	TPPPHQMQEVPGTPLTTFRLPNTSGAFQK
SALFAQLNQGESLTHALK	RATGRPRCDLLQAMPR	RLLARAPPDMQTGRCR
RLASESLNEYEAQTPPRLPEVTK	DPGPPRPPAGATQDEELQGSPLSRK	DGRPEAPPSPLGQDEPRLAGSPQTK
RPGRSPQLFPAFVDQR	NGQTREHALLAYTLGVK	NTQYLALRVATLHEGGK
FSLPDSVVYLAQVDQSPK	KGSITSVQAIYVPADDLTD	KSALVAYITTPQSDIDGVD
RSSSSAYECDEELNALLAAVLGR	RTMMACGGSIQTSVNALSADVLGR	RGALVMSMGNVSITCLDSAQTAGR
GTYAEGPRRSKDDESGQLLPHWK	SAYALGGLGSGICPNRETLMDLSTK	SAGSLIRGTADLTGEYGCPLLSMNK
MEAAGFTLSSSHPLCPPMLTDAR	MAAVAATAAAKGNGGGGGRAGAGDASGTR	MAAADVARGAAGKGGASGTAATAGGNGGR
AGYLLPLEGPGLTTTESR	TVSVSVPPHGGGALPRCR	TPAGGSVVHGSRPPVLCR
AWPWTHHTAFEELELVHPPLAEE	AWVWNTHADFADECPKPELLAIR	AWHWKDAEVFPCLADETAIPLNR
TRPFFDDLLQDELQHLK	DAKNLIPMDPNGLSDPYVK	DDMANINPLYPPDVSKLGK
LYNPLPEQLVCDDDNPSVVHLFFTTTLEGR	YLNPLQTDAAENNVCDINSVHGLFATGTIEGR	YSAITCLLQGEAAGNHDTNTEVLPVFDNGINR
VYDTTTNTVTAAEHHESEHYKKF	VYDEETTDLLAHWNEAYHFINK	VAWEYNITDTEHFAYLNHELDK
GFGLSLLQNSEYFSEGR	MAAISNTVQFLEEYCK	MFLAETQEVAICNYSK
RSVAPSEVMMPSDLQLPDVVSDPDK	IMRPTDVPDQGLLCDLLWSDPDK	ICPGVPLWDRQMDLPLDDTLSDK
TGASDATGSLWGSSMTETVAEELVVSLLK	STVVEFSNKDASEINSEQDKENSLIK	SKESNNSDETEQEDVNISSLAVIKFK
LPSPSSAGLLPEEEEVFEEK	SLPIVFDEFVDMDFGTGAVK	SVGLDMIEDVDFFATVFGPK
HVLGKQGEETQGTDQSSSSSARRR	HVIGLQMGSNRGASQAGMTGYGRPR	HASNYGGRQVSMITPGQMGRALGGR
SVVPSHHPPHPAAVVVVPSPK	SRVLFPLGLGHAAEYVRPR	SVAPLAYLFRRVHLGEGPR
SALFAQLNQGESLTHALK	RATGRPRCDLLQAMPR	RLTCGAAQPDMRPLRR
DSHEFHALNGMLYNLPGLR	AEEPLAAVTPAVGLDLGMDTR	ALGALEATTPMADVLEPVGDR
TYYVHRPTAGQLTPALLHSR	RLFNVDRHVGMAVAGLLADAR	RALDDVRLHGAAFVNMVGLAR

Continued on next page

DENOVO sequence	Correct sequence	Decoy (shuffled)
LDPWPPPPPPLPPPLLLLLRR	SLPTPAVLLSPTKEPPPLLAKPK	SPLPPPLPEKKPLALTVPATSLK
DNAPHRLHPLLLLLDPPRRR	RQGGLGPIRIPLLSDLTHQISK	RIQPDLQRGHLPSTLLGGIISK
MVSAHLAQPVTAVLLATASLQTEK	ALQGPDSVPPGVVDAIYGALRTLR	ATVPQIDPLSAGGDAVPRGLLYVR
VVDDVPAGFTPPLLVLVVVCPVLR	VLGDVPGACTPVVPTRIPAIVSLSR	VAARDTVCTPISGPIGLPVSVVPLR
LLYFRPDTATPLSDLDHLADLATQK	ILYLINQGEHLGTTEATEAFFAMTK	IENLTGLFTYAGIHAQALFTEMETK
VNEMLLDGFAFTFLK	GVAFVGGFSYADVLGSAK	GASGVSGFFVLGDAAVYK
KSSCVTQVGLLESVYEMFRK	RQLGFEEVSELQGDPGYPLGR	RLQGGGGQPEPSDYFEELVLR
STLHTQAASLPYTALTAWSALNK	AVVHGILMGVPVPFPIPEPDGCK	AVGDMVELCPIVPVGGHPPPFIK
SSHADHPDVATMLNLLALVYRDQDK	IISVFYTVVTPLLNPVIYSLRNK	IFPLLVLYNSVVRTVYTIPNISK
CMALAQLLVEENFPALALHR	GAGSYTIMVLFADQATPTSPIR	GFVSGSDPIMPAAITTQYLTAR
PMPGDGPVVAMVWEGLNVVK	VLLQDFTGIPAMVDFAAMR	VMDLFIGMAQVPALAFDTR
LYQLEYYLSHFAPEYFGLLAADGVVLAAER	LYQVEYAMEAIGHAGTCLGILANDGVLLAAER	LLELLHIAENVIQGGAYTAAMLGVEDACGYAR
MQLAGAADETKEQREEEL	RTNYDPGALQESVPSQSGK	RSNDEATGSYGQVLPSPQK
PHPPPPPEEPPEEEAEEEE	GVPHPEDDHSQVEGPESLR	GSVQDHHEDELPSPGPEVR
GYHEEQEEEEEELVDPLTTVR	YGLQDSDEEEEEHPSKTSTKK	YETSQKGKPSEDEDTEESLHK
ERNTDQASMEDNTAAQK	MKQDLEDASNKAEEER	MEEKNDQEAALDESKR
EESVEELEEEMEEEETEETTPPSPMS	GNSVEELEEMDSQDAEMTNTTEPMDHS	GMDNTEMESEAEPENDTSQVMETHDLS
LQSMLEEELRHLEEADDLHQTGSSSK	LQSMLGSLRPAHLGPCSDGHYQSASGQK	LGGSHMHPQLQSGCARDPYQLSLSSAGK
GGLMRPLVVGEVGGPRRPGPPGGGGLK	LALRGPAGPMGLTGRPGPVGPPGSGGLK	LGGLRLGGGPGPSATMVPALPPRGPGGK
FREMLPFAVVGSDHEYQVDGK	MARPPVPGSVVVPNWHESAEGK	MVWVVGVRPAEPEHGASNSPPK
NFSVHLSAGDDDTTGECRLLLKKEEPPSSK	NMSVHLSPCFRDVQIGDIVTVGECRPLSK	NVVHVRSCVRDLTPILCMPFDQSSEGIGK
LYLELVQPDLPPDNNKVGNNVTTK	IYITLTGVHQVPTENVQVHFTER	INQHPELTFHVVEVYITQTTVGR
DGWAHLNGSLFWTNNLPPDGGEPK	EAEEVFERIFGDPNSHMPFKSK	EEEIGMFFPHPSAEVDKNSFRK
RPHLEVQLLGDEYGNLLHLYER	QLEASFARTVNKEYPGLADPVFR	QGFANVPKTLVYADRFEPAELSR
RRELGNAYVVVVSTFEPNK	NIREQQILTLFRENFEK	NRTEQELLRNFIEIFQK
VATLHDQGTAQWADLSSEFYLR	IPQSHIQQICETILTSGENLAR	ICTIQHEQSNAGPILSTEILQR
VVQPEPVVSSLFSPSPRTAARLRR	VVGAWDPTVSVEEVRPQITALVRK	VTAVSVTQLAVVWEPDRPGEIVRK
LDNNTTVVLLSATMPSDVLEVTKK	LVQLYAVVSEEPIYIVTEYMSK	LYPVMEVSIAYQEYSEILVTVK
LLQMDEQTLGEELSYDLVLSKPSK	MCLAADVPLIESGTAGYLGQVTTIKK	MLCSTGTYPALKELGDVAQTAIVGIK
HLDCSSNLLETLLGNPSLLGYGTTLPLWRR	HLDCNSNLLETIPPELAGMESLELLYLRR	HPRYASTENLPLELGILSMNLEECLLDLR
HDENVLQWLNAMDELGLPK	QFGFIVLTTSAGIMDHEEAR	QAFIDEELVGMITASTGHFR
QLEGLSVSFLQHHLHLEPLNEQLTTLEK	QLEGLVSAQVQLCRSNLELMHTVVHAAR	QLTLHELALSVQCSAVLGVQNRHVAMER
CTECFPKKLPFCQECDPGDVDDVFLPKPFR	CTEDMTEDELREFFSQYGDVMDVFIPKPFR	CGQPVDELRDFDFEPKTFYITVMEEDSFMR
MYEQSNEDLPLAEQSSK	NSKAGSGGKSQITWDNPK	NPDKSSKTGIAWQGSNGK
RGCESEENLPEHHHEQDLDDACEAASTER	SEDADRCTLPEHESPSQDISDACEAESTER	SSAQASDDCREDETEHSETDESEPAICPLR
DDDQRREEESLEELKDDEK	EVESLRMQLLDYQAQSDEK	ELALRSVQSYDLEDQMEQK
EEEETLEEEGEGSTEPEPRRPPPPP	QGTTETIEEVEAEQEEEAGSTAPEPR	QEPEEEETQTIEAGVSEPTAGATEER
GLALQHPGTEVLLGTDSLPNLHK	QEIAAARAADALLKAVAASSVAEK	QAAAAISLVSAAAAVAKAEDELRK
VGFTNPNGGGEEETPLDLQAPTSSSERRR	VGGTNPNGGSFEEVLSSTAHASAQSLAGGSRR	VLPQGVGGETFSGAGGHASSSTANRESANSLR
KSTYSLSSLLELPNEVSQSHDAYFNEVSHH	CHLQEAFSHPILVRSGCHNKISSIGYWK	CLKHHNGAWEIRYISISSVQLCHFSGPK
YGDLSHNRLLDYYLSFDK	KPRQHSGDHENLMNVPSDK	KRQDNHSEDPMLSVNPGHK
GGPPGPVPPPPPLPLGSLLFSLQK	AAPVGPVGPTPTVLPMGAPVPRPR	AGPGGVLPPVVATMPTPVPARPPR
ERPLGLPAFYTPGGGGGYLLVSEGGSSLL	IRAGGAGVPAFYTPTGYGTLVQEGGSPIK	IAFSTGGYGGETQPYGPRVVTPLAIGAGK
GVMTTGELRGFTGSEADVVKPPHHGEEFF	GVMTTGSDIGFSVIQAHDVQNPLHFGEMK	GMDGGFLPGQSVISDTIVHMTFNHAEQVK
NNQFEALLQYADPVSAQHAK	NQVVHVPLEAYDSSLPLIK	NYLVQDPLSVHPLAVIESK
LSSLPACVPALELAQVPR	QAVTPPGLQEAINDLVKK	QGNIAELVPLKVQPTDAK
GAYGTFGQDLSMTFNHANGLTLVSR	NDPGHHIIEDMWLGVTVASQGPAGR	NHDGVMIASWPGPEDITVALQHGGR
TTTGVGEDVGGSVQAAQDDVRNPLHFWMK	GVMTTGSDIGFSVIQAHDVQNPLHFGEMK	GGGVIHITAFPSVSDENQHMLVMGDFTQK
HLEVQLLGDEYGNLLHLYER	AGSHGRRSPGGGSEANGLALVSGFK	AGLVGPESNFGSGGSAHGRSAGLRK
KSKQQQVVFQQGQGPTTEEVTK	LNSNTQVVLLSATMPSDVLEVTK	LPENDAVTSTLSTQNMSVLVLVK
GKYEGEVALPFAEEELGK	ERITSEAEDLVANFFPK	ENSPERILFAFADETVK
ASLPLLLAAYAYLPPAPDDLTATTT	APGSLVISAYAVCPDITATVTPDLK	APVAVTLVAGDLTDPYSPITCASIK
DGTTHETSLELFMYLNEVAGK	DPQEKTLSIFMEYMPGGSIK	DMIQPEFEIGSKPMSGYLTK
NVCQGNQVQTQLLTLLLEYG	NVCQTCLLDLEYGLPIQVR	NDPICQLELQTVCGYLLVR
AREDEYENLFTMLVELPR	TSLYEIPAVSSSSFFEEFGK	TSGSVEELFPYSFEFSSAIK
THVTGLTTLTSGGSQEEESLNNNWLLKK	TVPPAVTGITFLSGGQSEEEASINLNAINK	TNNGAIIGEVVSESELTPGSATFAILQPNK
TAEMLLEELGVKAEYYYQGLDDAAFNNL	VGTDMLEEQITAFEDYVQSMDVAAFNKI	VEETEQFDVQKGDLAMYDVAMAISTNFI
VVSRRCGVEQSLHLLSR	LYMKSLLKIFAWATLR	LKLALLAISFMYKTWR
KLDELYRDHEKKLLLNYGTWRK	KIENAMKTGGSVLLQNLLETLAPGLK	KTEEVPGLTNAGMLQLLGSNKILALK
MHGGPPSVMGLPLLK	AINQQTGAFVEISR	AVFEIASNGTQIQR
LQPTSHQPCQLDPPGGGRPPPPPLLSQPGK	LQGAASHVPEGFDPTGPAGLGRPTPGLSQGPGK	LGPGQGFDQAPLASGTLTPPPHRSAPGGVEGGK
MSSNPPVMVYVPDNFPSKPYANNFTTPPK	VKGTNEDMVFRGNIDNNTPYANSFTPPIK	VNNYDPVSRIGMPKFGDTTFNANPTIENK
EGSLHLPLLSPLAGFNTACVTK	HASPILPITEFSDIPRRAPK	HTPIEPIAFAPSILRPRDSK
SKKPALVLGPVSLETTTPYYTYSSSK	SKVPVLDEGLTSVETYTPAIRANDNK	SIAKTLSYPTGEDLTNNVEVRDPVAK
DAWERLPVDAQQLTHRDAVR	SLELFFATSQNNRGEHLVDGK	SLFRLFEESHQTLDNAVNGGK
LGLRPNNPGFFEDLEPFR	RGHVFEESQVAGTPMFVVK	RGFGAVSVVETMEPFQHVK
RGRPGGGGLLSCTCTPGFELSPPGEER	SPLDPDSGLLSCTLPNGFGGQSGPEGER	SGGGGELNFCETPQGSSDLPLGSLPPDR
LSCDDAELLDDLDLNMELLLNLPK	ICEDSAEILDSDVLDRPLNIMIPK	ILSEIIIEDLDDCVMDNPPLSARK
DPEDEQLEPELLPEEPGTK	HFVDSELECNDVVLFWR	HDLLVFVEFVWEDSCNR
KLEGDSTDLSDQLAELLQAQSELL	VIAMEKAEAETTLAEVMPILEAAK	VAEMKELTVTPIAEALAAAEMIEK
DASLVGSDAAFFPSDGFLLLLAK	DRVTVVFSTVFKDDDDVVIGK	DVTKFRDVTDSIVGDDVVFVK
LREEEEDGQDDDAAAAALEEK	SGYRIDFYFDENPYFENK	SIPFREDGYDFFYNYNEK
SSSHEPWEALETPLLPHQK	DLILQKGITQNALDYMKK	DLGQALIQIYKKTDNMLK
EPREPTVYTSPPDKEDYMTDLR	GEAIKYLTEALQSISELELEDK	GEEQDTLALSAIIEELLYEKSK
VTQDLLVGTDGCGRLPPRTTK	VTQLDPKEEEVSLQGINTRK	VDQESPREGKLVIEQTNTLK
TSNLSSSHQLDDDFPLDDGGHEERELLLR	TSNLSENCHLYEESPQPIGSLGHDADLRR	TSLRLLPHYELIANSSPQDCDGEGEHSNR
SVDPDLDQQKTDKRREQALPELCMEGK	SVDPDSPAEASGLRAQDRIVEVNGVCMEGK	SCQEDVEMNEGSGSARDRVDPALAVIVPGK

Continued on next page

DENOVO sequence	Correct sequence	Decoy (shuffled)
ENGSSSSQPDDEGELDTLGAD	SSSMSSLTGAYTSGIPSSSR	SSSTASSSLSMIGPYSGTSR
AFLEAQQEALEELELESPK	FGAGGGAAGAVLGIDNVCELAAR	FEIGGAVGAAVGALGDAANLCGR
DLGLYEELVPPSEAAALVGDEEEK	DLGLPTEAYISVEEVHDDGTPTSK	DPAEGHYGLLDTVTPVEDIESSTK
KKHPVEEEDHQLPHDSNNNDDLYR	AGLQVYNKCWKFEHCNFNDVTTR	ANFVFTDQGCNHLWKVKNCTEYR
RVEHDQSYSQAGLTETEWTSGSSK	QKGMMRPNASQPGGAKDSVNGTMAR	QGSDRNAKSPMGQKAATMVMGGNPR
ARAYGPGLEPTGDMVK	LKGEATVSFDDPPSAK	LSVDKPFPDGTASAEK
VGDLEEEVEQQELQQHEEPTVVVRSHTANK	DFQPSRSTAQQELDGKPASPTPVIVASHTANK	DKSGSNQHDRAAASFPISTQTPATQVPLEVPK
ETLDNSQGAYQEAFDLSK	SAAGGGGGSAGGGAGGAKTSKGSSK	SGATAGAGKSKSSGGAAGGSGGGGK
LLARDPQQEDMEELLELL	NQDRFISTLKLQIEDLK	NLDSLRIEFQQLTIKDK
AETNGGGDAVCEGGFFPNLLENNK	AEGSDVANAVLDGADCIMLSGETAK	AATSDADEECVDMALLGAGNGVSIK
FGHHGNEEEEPPVNHLLLER	ENPETEEDVGPVVQHIYELR	EPDVEGVEIHQNVTYLEPER
DDPPALLVVDLGGERPPCHHYYYC	QTPSSAALTAAAVAAPPHCPGGSASPSSSK	QPSPPSTPSAHASGGAAPTSAACAASVLSK
MTPDDLAERDLPPVVELEEEEAK	LEGDLTGPSVGVEVPDVELECPDAK	LECVTSPVDEPLDDELAEGGGVPVK
SHLVYYEEESEEDPPPPPTRRRPPSSSEK	VPLVAPEDLRDDIIENAPTTHTEEYSGEEK	VITNIVTAPEGEAERLDHYEPPLESDDTEK
FKEDGFDCGVVALLQLVECELLPPFFEE	KVSDEFDCGVVFEEVREDEAVLPVFEEK	KCEFVEFSVPEVADEVEDVLFDEERVGK
RVLGQLHGGPSSCSATGTNR	EIQTRSASPSNIKAQFR	EINFQPSIATRASSKQR
LNRFVHLQAGQCGNQLGTK	LLQVNTGAKEQLFFEAPR	LKVTFELPAELNGQQFAR
DGGRLRHEEPLPPSSLLPGGK	GWRLPEYTVTQESGPAHRK	GRGQPAEPTVTEWSRYLHK
KHLSQLSVAEDDEESLLGTHFLVGK	SARFTPTHAFLATWEQVGAYEEVK	STRALGHFAATEWEPYVVTEQFAK
PPMVDESGVFLYTTSNHMK	EAHLPPGAMAAVGLSWEECK	EWHVGMAALPCPAELEGASK
EGGEDRRDLDLTSPPGLPLDLLDR	VSVSGSCKVASSPASSQSTPVKETVR	VSPCASSVSSAVQESSVVTPKSGTKR
KGHPTEGHELVDSVLDVVR	SDLEAKEGEVLDELSLLGR	SLLEAELLLEGVGEDDSKR
NFYDEHEELTNLTPQQLLDLR	EAVSVGTKDLPTVQTGDIPPLSGVK	EGGTLQTPLVDDPVSIAPVKGSVTK
EPGNVLQNEEGETTSHLMGMFYR	NDEKMNEVMNLAHTFLQNFCR	NMEQNLAFLDNTENHFKVCMR
DMDPLNDNLATLLHQSSDK	GGSCPLMPDKPLSANVPNDK	GLPNMVNCPGPPSAKSLDDK
PQLNWDLPSDLEEYVHR	DHMKSVIPSDGPSVACVKK	DSSIVGCKVPVKMDSPAHK
DDDLDDLGTGLGPDGDLPCLALLR	GDLPDPIGTGLDPDWSAIAATQCR	GALGDDAIPDPQCDPIGSALTWTR
AELSGPVYLLLLLQDLQEELCK	VELKSFLEVLDGKIDDLHDFR	VIHLFFSLKELDKVDEDDGLR
GFVDPYFEDNAGVLDNNK	QDEPIDLFMIEIMEMK	QFIMPEEMEIIDDMLK
DSQPTQTFLLLKREEEKLDDLLNNDDALL	DSQKPTSPLQSAGDHLEEELDLLLNLDAPIK	DLLGLNQLSEPSLLQDAHPSADLTPIKEDEK
VLQHYEESDKGEELGPGNVQK	FLYMLMEYVPGGELFSYLR	FMGYELMLPGFLYSEYLVR
EKPAQDPDYQLPSPSSSPPAAPGGGRR	EKPAQDPLYDVPNASGGQAGGPQRPGR	ESGRQQPAPLAPYVPKAGQDGPNGDGR
HEDLNTDQENLVGTHDAPLR	HMEAQRAEENIRSLMSTEK	HESMTLSEIANERQEARMK
LSSEHVEGDEDDLPSPEHAWWK	DRVAILVDDMADTCGTICHAADK	DTDGDRVCVLIAHTMDICADAAK
CTQDLTGSLCEEVEVVVEEGLENSPPSSTK	SAMCLTGSPQEQGVSVVSEEGLENSAPESASR	SEGLPSVSQSGPLESAGNAEVCSQEVESMTAR
PGDTLTELLETPTTSELEAEHQR	QHMSSIKEETPTCAVSVQKQGK	QQKPEVHGSTAKMCIESTSQVK
LKEETELLLETAVEQLDDRAAGGTSKK	ISTEVGITNVDLSTVDKDQSIAPKTTR	ISNDLTKIKVEADTVTSTDIQSTGVPR
LDKEKELLEGELLLEQLDGGKDMLGR	IKEETEIIEGEVVEIQIDRPATGTGSK	IVVIGIAEQGETEDESTKEETIRGPIK
RVLGQLHGGPSSCSATGTNR	EIQTRSASPSNIKAQFR	ERISKSQTIASQPANFR
LVGQLHGGPSSSSATMTR	KQVLHGKDFQVDCK	KCGDFDQVKLVHQK
RSDHDELVEHYK	QSSTATSSSSTGGSIR	QSSSSTGSAGSTTISR
LNSHMDALHLGSQANR	DGVKVPTTLAEYCVK	DKCAEPVTTGVYLVK
MRMLVHLQAGQCGNQLGAK	GFAYPSELKAHEAKHASGR	GHHPAAYASFSKLGEAKER
VDHGKTEDFPGDLDPPSYADLGK	IHTPEQGSPSLGQNWSWGNRK	IRSNPGLSQEWQGHTPWGNSK
LFVDSSRRSSQEDDEHHQPPPSLN	FIVDGWHEMDAENPLHQPSPSLNK	FQMHPLVELWSGSEDNPHNDIPAK
SLGMDGGRRLMQPYCQLLASAAQEEGQK	SVGPLQYASHMEPQVCLHASEAQEGLQK	SQEGMPLSQAEQYVGLPALACQHHVESK
LTAECCRGDLTPQELLKDGQLFFF	ITAEDCTMEVTPGAEIQDGRFNLFK	ILGIFDFEAEATTEMTGCQNRVPDK
RLPLNCEQMLLLKPLDGLTDELK	AVAEPAHRTAPAGGGHPGATPHLAGRPR	ATAPAHGAHRPPPGGVHLPGAATRGEAR
EQPEKEEPLPYYPPQLQNNTLLR	QQNVDQAVATLQGEGLSVTGTVCHVGK	QVASHLGNQTGGQLTVCGDTAEVQVVK
DLHGSNPPPHHFYYPPLSR	DLYANTVLSGGTTMYPGIADR	DGGSYMIVLATTGADLTPYNR
KRRPTFLEEDDQAFSVAER	GFGIGSEILKNLSQVAMRCR	GGVCKLFSMLEANRIQSIGR
DLNFSLHHDHHDHPLGGGLL	DLYANTVLSGGTTMYPGIADR	DLDTATYLSINGAYPTGGVMR
QDLSESSFVLLELPPFFEDDEPKK	EDGKDSEFAAIENLPGFELLSDREK	EESRILDAEFADFPGKEDEGNLLSK
KTKPNDELELKGTQGNWALLEEEFNSDLK	IVKPNGEKPDEFESGISQALLELEMNSDLK	IQMGESILFSKGANNLEDKESVPPDELLEK
HHEEEEYVPAEELSSFLFT	WGDAGAEYVVESTGVFTTMEK	WYVGVGAEDESGATVETTFMK
LCMGLLDKEKEPAEQAAMAMTAAS	LCYVALDFEQEMATAASSSSLEK	LTSLAVADSCYLQAESMEASEFK
LEGDSTDLNDQLAELQAQLAELK	ELNVCREQLLEREEEIAELK	EREEELNRELVAECQLIELK
KQQEELHLQLLTQQQAGKPQPK	AQLAAQLGLTQTQVKIWFQNKR	AGLFTALQLKKAQTQQNWVQIR
KPVGECKDDHHSSSSSRGGLLLLGK	KPVGEVHSQFSTGHANSPCTIIIGK	KSACGIVNGTGSIISPEHHTPFVQK
MRDFMALSPTAAANFTRR	SSLVTPSISKEEILESSK	SSESTISVLKSEPIESLK
GRPAPGFHHGDGPGVSVQELMLPASK	SSEETIQPKEGDIPKAPEETIQSK	SPIIESDEPEPQGSQKEIKETTAK
PRDVSSVELLMNNHQGLK	LAYELYTEALGIDPNNIK	LAEILTAEPNIGLNYYDK
LDLYLYPETPHHEELEPP	KEETFTPMPSPYYMELTK	KLMTSPYFEPETPMYTEK
RSEEQLTVAEELEESTEEAAFGTPK	AYHEQLTVAEITNACFEPANQMVK	AEVNPITQATNQEFYAHCMALVEK
SLYVNNVDYDATAEQLEAAFHGGGGHYR	FQDTSQYVCAELQALEQEQRQIDGR	FGQLSQEYAVIQCEREDQQADQLTR
VDLLTLAGTPPAAYEHCRREEEK	IIEEAPATIATPAVFEHMEQCAVK	IIVEAMHPEFEIAEAPCQTTAVAK
EHGLDPAGGYVGKEVLWEK	SVLGPVSTGPPPVNKPEMR	SVPKGPEGVVSNPPLTMPR
GNMNLDDDDAMAMMF	DPGMGAMGGMGGGMGGGMF	DPGMMGGMGAGGGGGMMGF
AYHEQLSVAEEGVACFECGLLVGGR	FEFTPGRDTAEGVSQELISAGLVDGR	FAELEGAEDTVGGSIRSLDVTQFGPR
HQLNFNLPSDLEEYVHR	DHMKSVIPSDGPSVACVKK	DPSIKKVHMAVGSSVCPDK
LEQQVSSMLLVGSQQTDDGEALVVR	LEQQVPVNQVFGQDEMIDVIGVTK	LVQDGQVQPVNQEFVGITMIEVDK
LEGDSDELSDQLAELQAQLAELK	AFEILDRGPLSLDNTLYQVEK	ALDPDFYIQTEELSLVNRGLK
RDDLFYTSPPPPPQQLVMSK	VESRDKLPQPVQPDPVSHCK	VPPHKRPSDEQLSPVCVQDK
EKPYPPSAVSSSSSSAVGGSHHSPLGK	EKKPGDGEVSPSTEDAPFQHSPLGK	EGTGPKPGVDPEDPSAQSLSEHFKK
EAAAAAPPGAPPPPPPHVVHCPFTSPPAK	AEAAAAPTVAPGPAQPGHVSPTPATTSPGEK	AATVSEAQGPAPGPAAEPPGSHTAVPTPATK
APPRPPGREEEFPLDDQQVEEQPGSS	FYSRHKTMNFMSLEGTDTIEPNSK	FFTGKRITTDSSHENPLESMNYMK
VGDVMRPQGEVHLEEEEEEFFEDRRHSDAR	FLNEHPGGEEVLLEQAGVDASESFEDVGHSSDAR	FVSLGAELDGSVADDESFHVGEEAENLGPQSEHR

Continued on next page

DENOVO sequence	Correct sequence	Decoy (shuffled)
HGEVCPAGWKPGSDTLKNDVQK	KPDNQSLGHRGRRPSGPDGAAR	KRGQRPGDPAGLDPSNAHRGSR
HGEVCPAGWKPGSDTLKNDVQK	KPDNQSLGHRGRRPSGPDGAAR	KHGDQGPADRPRRNGSGSPALR
SNSNSDRTSLHEAMEQQSLSLSK	GSPGSQPEQVTQRPEEGKESLSK	GSSPEGEEPSLPRQQEVGKSQTK
DSGPLGTEFNTGFSSEVK	FGIVTSSAGTGTTEDTEAK	FGTTADTTSIAEEVTGSGK
LNPHLQAGQCGNQLGAK	KFAISIYLSEVSLQK	KELLASQIFYSIVSK
HVSPAGAAVGLPLDEDEAK	VDTKAAGSGELGVTMKGPK	VEAMSGLPGATKGVKDTGK
SLGNVVHPLDVVDGGQDQSK	FASAGDDGIVVVWNAQTGEK	FENGAVGTIVDVQAGSWDAK
FNSQNSNSSVLEVVSEKPLSER	HFDSYIETALDGRKESEALVK	HTSAVGSDYKEEREFDIALLK
AGCGGGCGEELLTCCLLLLQK	ADCPTMEAQTTLTTNDIVISK	AVESALTMNTIPDCTDQTITK
RMPGDKPLLEDVTTEDEEMGPK	NVLSETPAICPPQNTENQRPK	NITQPERNSQACNVEPPPLTK
LATTAGPYMETDSETTTLMFCP	ECLAVMESYFNENQYPDEAK	ELMYDEYSPNACFAVEQNEK
DVNVTEDFFSSEPLPTTTQQQR	ESESAPGDFSLSVKFGNDVQHFK	ESSDFHFDGSSPFLVKENQVAGK
LPSSSRRSSGLYDLVSSA	YLSSGIATSHSAKPPTHK	YATLSKIHPSSSHPATGK
RVLGQLHGGPSSCSATGTNR	EIQTRSASPSNIKAQFR	ENKPIFSSQQARTAISR
DHQFGGAHHVHHHFSFSPLEPTPPPPR	DHQNGSMAAVNGHTNSFSPLENNVKPRK	DNNPVQPNNFSLSHNKAMGVGRESAHTK
ARGHWFPKPPVGTKMLVEK	RISHEGSPVKPVAIREFQK	RKGPAVEIPQIVFSHRESK
RVTLLTQAGATGGGGGTSGDSSK	KSIGILSPGVALGMAGSAMSSK	KSSPSMLSGGASLAGMGVIIAK
HTSFGVASVESSSGEAFHVGK	ELKEDSLWSAKEISNNDK	ESDNWKLSDKINELESAK
YYTGVVNNAEEVLLKCDPNNPYDK	YNTAADALAALVDAAASAPQMDVSKTK	YADAATNAPDLVSVAAATLDQAMAKSK
SSDSESSHHHHSNPPPPVLLSSSNPWKR	SSQASQNRHSMEISPPVLISSSNPTAAAR	SMVQAPNSALASISPRNAISSTHEPSQSR
KSSYSLDDVLELTPSNFNR	WLVALGSAKACLTDSRTQK	WGTLQCSLVTALDRKAASK
KLYSSSDDVLELTVSNFNR	KIYPTVNCQPLGMISLMK	KIYTGIMQCLVSNPMPLK
LYQEVENASVDAFKGREK	HHISSGTITSKEEKTEEK	HEITKHESSEESIGTKTK
PTQVSSSLSPQGTLTVMAPMPK	IFPENDKQQASPSCPKNIK	ISKPQPFNIENSAKQDCPK
LAELEEFLNGPDDAHLQQVGDR	RDFIATLEAEAFDDVVGETVGK	REGDVALDAEDFAVITTEGFVK
RRAAESNALLVQQGGLHEEELLR	RVVEDTTAIDVQVGLLYEEGVRK	RDTGEEDVQVTYVILRLVEAVGK
VGLLSSYGEMPGDEQQAMMEK	VNIAFNYDMPEDSDTYLHR	VPTSNEYAMDDDYHLINFR
TLGRRDGYHEMLESSVLNNK	LFLLNSVGQEMSRCKTSIR	LEMRQSSKFNTVSIGLLCR
GSGNNEQLLWESLMLLLGSYSSP	ISIENEQLVIGSYSQPSDSWDK	IPWIGIENESLQDSSQSYSDVK
KSLTEYPSEDLTQPATSAK	TPPLGPMPNSDIDLSNLER	TGDDPLELMLPNISNSPPR
GFWEVLDLEVDEGPPCYHGDSDLQLDR	LFMHALKMDPDFVDALTEFGIFSEEDK	LELTEMHKDDFLSFGIAFEVMDAFPDK
QFFFKQTTEDQDPPFQQQTSTDPPGGDPFK	GADPFKGDPFQNDPFAEQQTTSTDPFGGDPFK	GSDEPAGKQDDGFQTAPPPFTFPFTQFGNDDK
DLDLLFEEELAEALELDELWAALDVVESLK	DIDILIVRENTEGEYSSLEHESVAGVVESLK	DIVELETLSIGASESEVGYHENILVEVDRSK
TEEGGESVEEEEEELLLLELYEVNFLFF	GSDSPAADVEIEYVTEEPEIYEPNFIFFK	GNADVDFEYVSYEIPPAFFIPTSEEIEEK
VENMDDDLDQPELNLFVLSDSSSLEDLVVK	DLPQTMDQIQDQFNDLVISDGSSLEDLVVK	DDIQFIPNVLVDSQLQVDLSESQDMDGTLK
TEYYDEVPSEPPEPPPPPPPPPP	SLTYDEVISFVPPPLDQEEMES	SQEEPVIMLELDSTDPEVFPYS
EGGREEVTTERATEVEEVQR	KGFAELQTDMTDLTKELNR	KKMDLDTQTFLLGETENAR
SLLKPVLEAPPESELDQGGSSEEEALDPR	SILKPSTPIPPQEGEEVGESSEEQDNAPK	SVPGEQEAGSPEQPPEESNETDPKIISLK
LTGSSASEEAEGVALGEAPDHSYESLR	IGDCTDLTVQDHESSTTEREEIAR	ISLDEGETVSTITRDTEHEDAQCR
LLVCEDLDCEENGPVCQR	AKSEENQGDNSSENGNGKEK	ANENSNGSQGSEGNKDEEKK
LEECDYPPLPSDHYSEELR	TLASEDIPDLPPGGGLDCKSAR	TKDPLDADGLCLPIPGAESGSR
KKDQTEHVPDDDLLEETGQGSAAELSTLAK	TGAGGDGLARPEDDLPDLENGQGSAAEISTIAK	TPRGADANSAGGEGTDPADEELQAGLSIIGDLK
MLAERESALDAEEEERRVVTAK	SVQACLAKASEGASSESLLSVPGQK	SPESGALAGSVQLSQVALESSCAKK
LRPPGDDVELLGEEEGEVVYDLADHPDFY	GLENLTLLDLSCNPEITDAGIGYLFSFR	GTLINSDSPLDYCGGFLEEILLNLFATR
DNYVPELVLLEDEEEMAAPVDK	ENLLIGSTSYVEEEMPQIETR	ETNLMTISGEQIEEYVELPSR
EVRELLAQYQQQQSQASADSTSR	KQLQMQIAYQQQSQEAEKEEK	KKLQEEAIQQEMQQQSYAEQK
NRDFVHFDDTSEPPTESLRK	LETTLNGAHSTSEGPAKPKSSR	LSEHSTKPNSAAPESGLTGKTR
AQQEKPEGYTPHSEQQQEELAAAAKK	KDNGEFSHHDLAPALDTGTTEEDRLK	KLAGTTFHGADLPDLHDRENETSEDK
AAGLEYNTQLLHGSPNPHCYLLLEEEQQDR	LAGIENQSLDQTPQSHSSEQIQAIKEEEEEK	LSEHLAEQQAPDSQGETESIEEEQSIQNKIK
YREEELELDDFPQSLK	SINEKFAGSAGWEGTESLK	SASTAWGFNKEISLGEGEK
ADLHAQSMNFYHDDDATTDLAK	VHNDAQSFDYDHDAFLGAEEAK	VASDYNGHDFAFEDAEHADLQK
TELLLEEEEQEELSEMQQLYAQDD	QEPLGSDSEGVNCLAYDEAIMAQQDR	QIDEEDPLSNAACDEYQGMQGSAVLR
KPLQGGSDDGGGGCRYDDEALFQQADR	QEPLGSDSEGVNCLAYDEAIMAQQDR	QDVIAAECSYSQNPMEGDLGLAQDER
AAKPCGGAASSESLQQPPEELLLR	FPNRPQMVKISKLPSDFTVPK	FSRMSIVKDVNTPQLFPPPKK
GGGEEEVEQQQMQEELVQQQEYVGFFPR	VAEDLESEGLMAEEVQAVQQQEVYGMMPR	VVEMGMEYAQDVEEAEQMVQAEGSPLQLR
RVLGQLHGGPSSCSATGTNR	EIQTRSASPSNIKAQFR	ERSPASSQAINKFTIQR
YKPVVNCCCLHPYFNER	REAEAMKATEDGTPYDPYK	RDPEAYYATEEAMTGDKPK
ESAHFQPVTAYSVAAEVK	DLPFVEEIKEGECQVK	DVCEIGEELEPVKFQK
SRSSASPTPAPSSPPAAPWPFGLR	GKHGFQVGLFPGHCVELINQK	GLVHEQHKFNGPQGIVCLGFK
APEANSHPVQLQTQLLQK	GPVSRTFVLAGSADSPELGK	GRAGVFGLDSPTVESPSLAK
EYNYHHEALSSAALLQQAALGDSN	SVEMHHEALSEALPGDNVGFNVK	SDSGGEMVHEPVNAELVFALNHK
AACQDMRLPPCDQLWSPPPGGQQR	NRQPPDSGPMCDLLWSDPQPQNGR	NQGLDMPRSLCQPPWDQNSDGPPR
GPYGGPYDPPGGHHHHGPPTHQYYWHALEK	RAFVHWYVGEGMEEGEFSEAREDMAALEK	RGMAADLVEMFEVSEREGFEEWEYGAAHK
EFHAPLLLDEDGVHELVK	KPVCPILGGTVMPNKTVR	KITVPGLVPTVPGCNKMR
LSDALAVEDDAVVVPPVVLEPPPPLWDDKK	LSDALAVEDDQVAPVPLNVVETSSSVRERK	LVSLSPESDVLDAAVRDVTPAVNRVESQEK
FAMVAKEDDLHVLLLSSEEEETEAANHHHK	FAMVAPDVQIEDGKGTILISSEEGETEANNHK	FEGDNTGEKTVILHNAAMEPVSEDEIIQASGK
TTELELFSRLEMEEQQQHHK	GTGENQFLTQQPAIITSIMGNGR	GEGNMAFQGNLSTPIQQTITIGR
HAVSDGSLLDSLDEDEDLLLLLNNNRRR	HKEMALKLEALHLEAICEANIETWMK	HHETEMAIAAEACLLNLKELIMWKEK
RRRRRDPNLEEEERLSLDDDEESSPEEQAK	APSEIDPRENPDLACLQSIIFDEERSPEEQAK	AEERRPIPLQEEQCSADDPSPENIDIELFSAK
SYYEDVVSEEPPPSSVQQPPPPP	SLTYDEVISFVPPPLDQEEMES	SEDLEYIEQSPVFLVPMPETDS
ELDPEDVELLGEADEELLL	EIQMDSPMLLADLPDLQDP	EPMQDDMISLQDLPLDLAP
LLRHELTNLSNVDVETQSGK	VCLVHPDVKWGPGKSQMTR	VSKHCQVWKGPGLMTVDPR
ARGGNWQPTEELLLKDTAESTYY	ARVANPSGNLTETYVQDRGDGMYK	AMSRRGVNTPVGQLTGYNDYDAEK
MKQQEMQDNESLLLLLQQLL	EVYNKENLFNSLNYDVAAKK	EVLFKLAEDNKNYYASNVNK
EPGPRPVFVAPCADVVVR	SLVKQLERGEASVVDLK	SLVVQELGDEVKSRALK
FQSSHHPTDLTSLDEYVER	GALSKGSESLTLMFSHEDQK	GSLKLTGQLEHFSAMSSEDK
TPLSEAGPQSPCSLEGVELK	REAEKSEDSSGAAGLSGLHR	RSGELEELGHDSSGAASKAR

Continued on next page

DENOVO sequence	Correct sequence	Decoy (shuffled)
YQDSFSAGALGPPSSEGGLLSQQR	ATLSSTSGLDLMSESGEGEISPQR	AETLISGPLLGSSEMEGSQTSDSR
LLAMSTNGLQPSSDSSSDEREEESER	AIAMSLGQDIPMDQRAESPEEVACR	ADAPSAEVCSGIIMPDMQREQAELR
SLYGLLLGDWDDEEEEVVDDLEK	TFEHVTSEIGAEEAEEVGVEHLLR	TSEEETHLFAVEEEGHEVILVAGR
PLADPDSFVLVEEEVVCRLEALK	KQILGSSSSGKFFCLYTEEFASK	KTQSGSEFSESLYFILSCFGKAK
QVRDVHLQAGQCGNQLGAK	YLQESLLKENMQKDLGK	YLGSKLNDQEKLEQMLK
VSTKHHHPSTCTFNSFFAASSSDDSEDD	LSESGFHMVACSSTGTCAFASSTDQSEDK	LESEGDFDTSCMQTSSASGVTASACHFSK
CPENAFGMPNNAFFLHHVR	EGDYVLFHHEGGVDVGDVDAK	EHGVVAGGDDDDLVFVYGEHK
LTHGMTYDPAPDLNPPSYADLGK	IHTPEQGSPSLGQNWSWGNRK	INHPLTQSGSPSWWGNEGRQK
KAELLDDEKVAAVVAPLTTGYTVK	ISQILFMFLVGLSILANTFVPK	IVLFSAILMTNIPLQFLSGFVK
KVWEGDPDVCCLALVAANEERFTTPPTTK	KVDAQSSAGEEDVLLSKSPSSLSANIISSPK	KVDSESQPAVSASKNLSEDSGSASLILISPK
KASPLPPAPAPGTSYLVSPLTGEK	TFQSPGVILSYLQNVSLSLPSK	TQLIVPLSSLGQSLPYNFSVSK
EASSGAGAAAAAAAAALNNAVR	DCLSLAAAIKACHTLK	DASTHIACLLALCAKK
LYCLELLCLLPSSVDDDDEEESGGPPPCC	GFTMIGEHSIYCTVNNDEGEWSGPPPECR	GNDYHEPGITETIFGECMGPWPCSSENVR
RPKGVNLPPPHDPQGPPPAGPPPQLR	SKSPPKVPIVIQDDSLPAGPPPQIR	SVPPIPVSGKSIPIDALQPPPQDKR
PGEDEPLHALVTANTMENVKK	RALEEPGPAADPTAFQGPWAR	RGTAAQFEDAPGPALAWEPPR
VDNTGSDVEEAVADALKK	CAVSEAAIILNSCVEPK	CISEAPCVSENAVIALK
WGAHAQVFEEEEEEVQQMMAK	EADIDGDGQVNYEEFVQMMTAK	EGEDNEDGYVADTVFQAQMIMK
TFEECDDDDGDSSQGGPAFERFLSLLGEK	AGQSSEEEMYNNEEAGPAFEEFLSLIGEK	AENEGESEQYSAEENFMPEFGSGLLEAIK
HLHSQVTVQLNSAEQELK	STLGDLDTVAGLEKELSNAK	SLLADVEKLGLSEDTANTGK
HLVDEVQNLLGKQNCDQFEK	KTVHELDDAEDRAGMAETALNK	KDELAMTEARNTGVDAHLADEK
SSESEREESSVVKCNKEPEDDDTTVR	EVGRPMCMTSVQMSNKEPEDMITGER	EEVTMEGGTDQISEMKMRMCVPPSNR
KKEEVDDDDLDDLDLEPDDDLLAPGVDLSR	NHEEEMNALRGQVGGEINVEMDAAPGVDLSR	NMNDAGLEGEVSGEMRIDHNAEPLVVEQGAR
RHEEEDLDLKEDEDLKVEMDAAPGVDSLR	NHEEEMNALRGQVGGEINVEMDAAPGVDLSR	NENDGEMGVEIEMVAELHSLDNAQGGPARVR
SVPPPLQSSPFKPDDSSSGFAD	KLSEFDVEMSMREDVYQR	KVEVDMRMELQSFEDYSR
HNNEEEEEDLLVPLLEPEEPEAAPGVDSLR	NHEEEMNALRGQVGGEINVEMDAAPGVDLSR	NMDNSVNGMGVGIAAPEEVLEHQLDEARGER
NLEEEAGVGADALDLREALSK	QVEEAERLKQSAEEQAQAR	QSELAEQVEKAQEEQARAR
NHEESYDLDEHLNLLLNNEMAAAGGVDLSR	NHEEEMNALRGQVGGDVNVEMDAAPGVDLSR	NDVGPDMDENEHMLVSALNGRVEGEAGAQVR
SVTETPGAQEAVPKKFGGPQPPGTG	SVTEQGAELSNEERNLLSVAYK	SVLGELSVLAENEEARQNYTSK
PSTEPLLAATGSPAAVPPEK	ISTIRTHASASLYLLMR	IRLLATMSIHTYASLSR
APQVTESLESSELVTTCQAETKGGVK	KIYLDIIHTYMEVHATVYGSSTK	KTISESMYDVLIITYAYGHTVHK
NHEEESDLLHHNRLHLVNEMAAAGGDDSSR	NHEEEMNALRGQVGGEINVEMDAAPGVDLSR	NVANGEMSHRLGGGENQEPVAIEEVLDAMDR
EHEEEEEELLLLLDLELGGYDAAPGDDLSR	NHEEEMNALRGQVGGEINVEMDAAPGVDLSR	NGLGDVAMNVGPGAHEADLERSMVEIEQNER
MTEELDPLTTDLPEEDNDESSYEAFVR	MTEEEVEMLVAGHEDSNGCINYEAFVR	MYECTAVHDVGFGNMEAIEEVEESLNR
NHEEEMDNLLHHNLHLVVQMAAAGGDLSSR	NHEEEMNALRGQVGGEINVEMDAAPGVDLSR	NGVGEGARMADINPEDMHQNGEEELSVLVAR
DALEAALADAEQRGELALK	WVELTAIVSTWLAVSSK	WLSEASTVLWSVTVAIK
ATSSGLQCDSWSDGQALLCDLLNETEEPGK	ATDGSLQSEDWALNMEICDIINETEEGPK	AITNNLDEESLSGGDDAIPEMCQTEEIWK
VTPVCHRTCKPEYAYGSGGSPPK	KPDNVPKCDEILMEEIKDYK	KDPPNIEEDLECYDVIMKKK
DGDDLHHGGLFGDYFSAYYDGYGFGSSDR	MRRGAYGGGYGGYDDYNGYNDGYGFGSDR	MDDSYYNGYGGFDGDYRAGRYGNYGGGGR
LADDRRHLANEDDNEEEDDLSKK	ALENDPDCRHVIPMNPNTDDLFK	ANDHDLDMNIFNCDLPVPPRTEK
VVAAPCRPGTEPVPEEPDTVLQSETLK	GAKGDPGAIGAPGKTGPVGPAGPAGKPGPDGLR	GGAPAGIGPGTPAVDPLAAPKGGKPKDPGGGGR
NNHEEECLLEVNNLLKLDEMMAAPAVDLSR	KNHEEEMNALRGQVGGDVNVEMDAAPGVDLSR	KDNQMNLMEHLDRGGVEGDEAVPAAEVNSVGR
ENHEEMSRTLENLLHKLNEMMSAPAADLSR	KNHEEEMLALRGQTGGDVNVEMDAAPGVDLSR	KVQGAELTHEAEANLGNRVGSLMVGDEPMDDR
SYEYEEAAKKKEKHEPVVMMMDAAPVDSLR	KNHEEEMNALRGQVGGEINVEMDAAPGVDLSR	KGIQVGGNDNEVERAVENESAAGMMHPLDELR
RPPRERRNLDDPYDPPPPQEEEER	IPRDVRDTVLEPYADPYYDYEIER	IDYYYRDDIPVEDPARPETLEYVR
FEDADDDAHAMAHHHHLEEEEVVR	FEDEEAQAVYWHSSAHIMGEAMER	FDSWGEEVAAEHQAHYEMSAMEIR
RPLPEDVQTVEPVTEVQSLYR	ASLEAAIADAEQRGELAIKDANAK	AASAIAAKNQEAELLIAGEDADRK
GGVVADLDRLPVVGELLYSSNEEEVVGAAK	AGVVANDAGDRVTPAVVAYSENEEIVGLAAK	AVADPLVEGVGANATENDVEGVVAARSAYIK
SLLLLAEEDEKATDDNPEKEKPPE	EAGSALLALQQTALQEDQENINPEK	EQGEDEENNLQAQASLIALATPLQK
EVYGMVYCPDHSLYYSELGGLSEQLLRR	EVDPLVYNMSHEDPGNVSYSEIGGLSEQIR	EYNIGVPSLESLVYINPHQGMDGVDEESSR
YASLCVQNGLVPLVEPEELLDGDHDLK	TLPAMHFVDHSLQVVRLDSCRPGFGK	TVQGFSDARMVSVPFLDPLLRHCGHK
EEVGPLSVEEEEDEEPALDTKEVEGDR	EEVGNISILQENDFGDFGMDDREIMR	EVDIIGELMFDGQGMIDSENDNRFER
DALEAALADAEQRGELALK	WVELTAIVSTWLAVSSK	WSSATISEVWTALVLVK
EGLEAALADAEQRGELALK	WVELTAIVSTWLAVSSK	WWSSLVVAAIVTTSELK
RVLGQLHGGPSSCSATGTNR	EIQTRSASPSNIKAQFR	ENKSSQISQIPRAFTAR
VRVHGPGLQGGTTNKKNQFTVETR	RLGAAALVPEAQDSQVTSTKSPTVR	RVSVLGDTAPVSKTALQSAEAPQTR
RHFWQQDDPQSSWDRVK	TQLYHAEIDALYKDLTAK	TDAQYATYLADIHELLKK
VCREVASVEEGLLQQEVSS	MNLASEPQEVLHIGSAHNR	MSLLAGQHEIAEHSNPNVR
FALVAPVEAEEDSGNVNGKK	TSELATLSQPPRSATPPAR	TPSPPSALRLAPASEQTTR
LTHGQTFDTANPENPPSYADLGK	IHTPEQGSPSLGQNWSWGNRK	INGPWRTSQLHGEWNSQPSGK
SSHLMVSQASQLLQQQQQQQLR	LENIGMEPLEKLEVTSKVLTTK	LGELSLMNTTKVELIEKPEVTK
NLYHSEAFSLNRFDAEEAK	EAGGRGSLHPAAGPGTAFPSPGR	EGHAAAFRPPPGGSLGSPGGTAR
VYRQTNLMNLDQAFSVAER	KLLSKEPSPPIDEVINTPR	KNVPLDEPSILPSPKEITR
NLYHSEAFSLNRFDAEEAK	EAGGRGSLHPAAGPGTAFPSPGR	EAAFGLPGSAPSRTGHPGGPAGR
VEFDRLELELPGQYDGR	ERISALNLQIEEEKNK	ENKEQILNEALSREIK
AADLNTETEREEKEKTFPLFKTSSEDYG	AADLNGDLTATREEFTAFLHPEEFEHMK	ATLGAHRMFAELLETEDPEFFNEHADTK
ERWEEELEAFLLGGPPPEEEEGHHVL	ERTAADELEAFLGGGAPGGRHPGGGDYEEL	EDEGAPPEGYLFEGAHRDRALATEGGGGGL
YAPLGLGGAAEEAEALEEPQQLPDTWHHLK	EVGEVSVLVNNAGVVSGHHLLECPDELIER	ESLVNHVHGVESVIDVEALNPGVLEEGCLR
ELPEMLHQTLQELEQLLLNNNDDHTT	TEHQVPSSVSSPDDAMVSPLKPAPKMTR	TMSHDTSPASPVVPQEPVKALSSMPKDR
TVDLEDAEEAVELVQYAYFK	ITIPDNYTIEVEAENGDGVIK	IINTYVEEPDEDINTGGAIVK
KGAQEPELEEVQLQSK	RPRELEDAQAGSGTIGR	RILEQTEPADGRAGSGR
RRDEPNVPPQPPNDDATR	QCGQVAAAAAAQPPASHGPER	QPPAAAAEAGPSQVACAGHQR
SVEVPPPPPSEEEEEESESLSK	LSTTPSPTSSLHEDGVEDFRR	LRHFSSSGSDEDPEVTPTTLR
TGSSAQEEDSGVALGSAPDHSYESLR	GSPQAQTVRRSPSADQSLEDSPSK	GSPQSPSRPESDTVAASSRQLQDK
VAQLSSDYHEDEEELLEEEEE	VSNISADDMLYTETDLEESMDK	VAIYEEEDDTTLMDLSDNSSMK
EETELLEGQVVQLQLDRQGSAGSSK	ETEEIKDLEEQLTEGQIAANEALK	ETILDENLAKLATEIEEAGQEEQK
SSLRDPESDDEQHWK	AGGGGTGRADDGGGGGGFHAR	AGFRGGGTAGGDGGHGDAGGR
QGRPPLVLSEQQEEGFTVVTFR	GVASDGGAVRLVAQEWFRVSSQR	GWAASQVDGGRQLVSESAVFRVR

Continued on next page

DENOVO sequence	Correct sequence	Decoy (shuffled)
VAQNSPSVENLQTSQAEQAK	FHNSRWMVAGKADPEMPK	FGAENPPDWRVMKHSAMK
SMMQDREDESLLCTGESGAGK	ASVKNYEGMIDNYKSQVMK	AKNSYVYVSIKMDEMNQGK
TGYGELDGNAGERELSLK	CRHFIGLMQMIEGMR	CMGQRIEMFGHIMLR
LPTGFSDEELEEHHQSLK	CELSKNSDIEQSSDSKVK	CVNESKQEKSSSDILSDK
DNHEDTASAADEVVAQGEEETNNRVVCTK	HDNEDTASASEGSNMIGTEETNFDRGYIK	HINNEYSEADSGNTGERTDSMATFIEGDK
EGFSFLGSSDSEFFFFYGYSDDDQQSSSDK	ATEDFLGSSSGYSSEDDYVGYSDVDQQSSSSR	ATYSLYSDSFDQDQEDSDGSSYSSSVGVSGER
KPEEYSDEEDDGEEEPPVRQNPEYK	FVSENKNLPIENTTDCLSTMASVCR	FPNEETTSSTVKANILSLCNVMDCR
LEAEAELEALETDHHDLF	TLDALIETESKRSAIFK	TDSTIKARFSLEALEIK
EGRRDGETSHLLEEEEESQEQDEEEEDVK	CGSGPVHISGQHLVAVEEDAESEDEEEEDVK	CGEVEEQDLSSGHEGVVHSEIEVEDAEDAPK
DRLFNEEEPPFHHLNGGNENLLVVKK	DRVALSNMNVIDRKPYPDDENLVEVK	DNLDVERLPEKNVVIMYSNDPVRDAK
VNEGSEHLPTCLLLAAPPKPSLVGSK	TEVTGDHIPTPQDLPQRKPSLVASK	THPVIPQKETTLDSDPPGLVRQSAK
FQSSHHPTDLTDLDQYVER	GLQNLAYQLGLDESREMTR	GLERLSYLTQNEDMQAGLR
LPLTPPQMQQLLQEETLPR	NQLATANKMITSVLEKEASK	NEQESKVALTLMTAKINSAK
LESSSSRSSSKPSGLLSSTPTTTSAAGLQK	MQIPKHISIEDITATSTSTTGTSHLVK	MAIGPSTHTQTHETIKSTLDSVTISIK
KQAQKPSGANQQRGAFVELDK	VQADQVDVKLPEGHLPEGAGLK	VKEPEDGGLDLLQVPAQAVGHK
CEDCGKPLSLKDDDLSDDDLDLPLNDER	CEDCGKPLSIEADDNGCFPLDGHVLCRK	CLPDFADIPCGLGSDHKLEVGDCNERCK
KGFVVQVLSDDFEELGDQGGERK	VIDRFDEGEDGEGDFLVVGSIRK	VDDSGGVIEGVDDLFIFRGEREK
RLLLLEEEEEEEQSGLSGQEGEDCTLSVYR	NRVGDTITVAAWNHSASMEEEGEDCTLSVYR	NIVCAYEDLDSEVSGWMEGAETSRTNVAHTR
LDCDNLNQYYLQQGGGPDKK	YSAAAQYKFFIVACDPPQK	YVQADAFACKIFYQASPPK
GSHHTLPHELEEHEHHCSVVESSDLAPGLR	SGTHTLPVESGDMKGSFALSFPVESDVAPIAR	SMTSLDTKVEGFSIDPLSGASEPGVPAFAVHR
EGMTVESAMLPLECQYLNK	RQVDPEFADMITVQEFCK	REETADIMCFDQPVVFQK
THSFGVASVESSSGMAFHVGK	VVADLSCVGDEYIAALGGAGGK	VVGGECDGYDLAASGAILAGVK
EGAYALVEELVPPPLDTGYDPP	EGNYAIVANVESMDYDPLVVK	ESYEMILYVVPDVNAVNAGDK
LCYVMPGSSDQLLLVTAAASSQQK	AAFQPEANPSHLTLNTALVESEDL	AHLVFTAANLSLDPTEPEEASNQL
AAVPSGASTGLYEALELRDDDK	ALGSAIEYTIENVFESAPNPR	ALPIEISNVFNYPTAAEGSER
AKLVDELEWELAQVDPK	IDAGKTEPSWKINPIWK	ITSAWPEIKNWKGDIPK
LALRRREEEAQHHVSGLSDD	IEVHQEEVVAEVHVSTVEER	IAVSEHHVQEVVVEEEEVTR
YYLEFETTEEKK	ALSSDSILSPAPDAR	ASASPAIPLSSDDLR
SRQGPAETGGQGQPQGPGLR	VPITWLQGKRESMSCR	VEPLWTGQSIMCSRKR
KGGSEQPDDDLLKEEGGEEEELLRR	KVSMGKPDPLRDSGTDDQEEEPLER	KEDMDSLKGEPETVDPRLEDPGQSR
THEAELVEGEDHTYCLR	GTENLQRAQAEVLQSVR	GAVLNQTEAVSQELRQR
TEHLNDGGGRRREGGNGGGYYK	DVILDDLSLTGEKMSDIYVK	DVIDDLIVLYMDLKGTSESK
VLAVDQENEHLMEDYEK	FKPGESFGGETSNSGDPHK	FSPHGEETFDPGKGNSSGK
GDDTQTGEGEEQGEPQPPSTTYEEWDLQKK	DGDTQTDAGGEPDSLGQQPTDTPYEWDLDKK	DLSDPEYDGDGWTGLAQQDTQDDTPPTGKEK
RVLGQLHGGPSSCSATGTNR	EIQTRSASPSNIKAQFR	EFSQTSISIAPRKQANR
LLHVNGFDGEGGEEDPQAAR	GEGERKDIVSSSMPRPNR	GMPPENDSKVRSSIEGRR
QELNSWQHELHQLR	RVDSITAAGGEGPFPTSR	RIVGDTSEAPFPGASGTR
ERDSGVASVVSSSMAAFHVGK	RGTGGVDTAAVGGVFDVSNADR	RVGAGDASTGFGGVNTADDVVR
HGLLVPNNETDQELQHLR	TYAYLFSHPSRMPVYPK	TRPLYYFHMAPSPYVSK
TRSSEESEERLLLPPPDSYAQR	SSSTGNLLDKDDLAIPPPDYGAASR	SALYNIASGDGLPASTPPDDSDKLR
HLPNELEEELSELVLESRR	SSLVSSLYKVIQEPQSEWR	SYSSEQVEPKIVSLQLSWR
KRLEEEESLLEEEEEEEERDAAPPLLAER	KNHEEEMNALRGQVGGDVNVEMDAAPGVDLSR	KAVGNDMNDGLALNQEDVVGREVMAESGPHER
RFFADEVLDETLQVDDFDDYK	DNRVVYGGGAAEISCALAVSQEADK	DAAQASVDNGGCGRVIEYVSALAEK
RRRRPSTSSEPEEMEPEEGPPPPVESVASR	TDPAGLSSPHLPGTSSAAPDLEGPEFPVESVASR	TSSPDEAPGPALAGSSGSTDFLLSPPEPHEVAVR
RLQQEEVDLLVDLDHQR	TLVGPSELPTASAVAPGPGTGAR	TTSAETPGGVAVSLLPAGPGAPR
GLVDESQQAYQEAFELSK	ARADGGGTESRPVLRYSK	ARGGSDRYAVSEGTPRLK
KSSPEHEEAAPPPPPPPEETAEAAAR	AEPTATMDDMALPPPPPELLSDQQK	AQETDQLPLPTMMADPPEPAPLSDK
ELSDDESEEVVDVLETQVEEGATTNALR	EAMCPGVSGEDSSLLLATQVEGQATNLQR	EGTQPNSLATQVAVLELQGAEDCSMGLSR
SEVFPDSPEEPFADATLPHLKK	RQLPFRGDEGIFEESFIEER	RDFEGEEGEIRILFSPQEFR
HHLELDLDGYEVPSLSEEAADDDDLVDLSK	NHEEEMNALRGQVGGDVNVEMDAAPGVDLSR	NGPSVDEDELGMRNMALHANVVDEGEQAVGR
HHLELSLELHFPPPEPLETHHHSSAAMSS	VHIEIGPDGRVTGEADVEFATHEDAVAAMSK	VDMAGHDTVVEHAREDVEIGIATFGSPAEAK
SQRWDAAGGETQEAAAAAAAAAR	RGSQKSTDSPGADAELPESAAR	RDSKPATPSAAESAGEGSLQDR
LQGLEEEGDDADPALESSK	TMIVHDDVESEPAMTPSK	TVMEDVAIPMEDTSSPHK
KEEELQLTNQAREEEERDLLLEEESMK	QVCEQLIQSHMARYTAILNQIPSHSSSIR	QSHRTILIIVYEHLSQNISCQAPQASMSR
LPAAGRDSTRGPDGLLLPPPSSSLR	SAGHGRDSDKRPSLGLAPGGLAVVGR	SKVGGRGALDGDAGRGLSPSAPLHVR
TNHLGHTGYLNTVTVSVDGSLCASGGK	RNTPHRGSSAGGGGSGAAAATAATAGGQHR	RHAAATRQTGNAPGGAASGGAAGSTSGHGR
SLPHLPVVSTPSSLPPPSSVK	YASIILFALQDTKISEWK	YAKSSILQALDIFIEWTK
GNFGGSFAGSFGGAGGHAVGVAR	IYTKTGDKGFSSTFTGER	IYTTTDGSFTKGEKGFSR
GPLLRLPHPPPPPTPLPCQQQLL	ARGIKPSPAPPPYTPPTHVLQTQI	ATTPAQGTLPHPQSVPKRIPPYPI
RAEHLSVESEEEHPHVVAAGGLVLGK	HISQISVAEDDDESLLGHLMIVGKK	HGVESLQSLKMEADDHVIIGIDSLK
AALVDMDPPKPGKVEEFSDYPPHGR	MSMGRVTPGQLMSYIQLFKNNLK	MPRVMLQFSLKGNSQTYNMGILK
DLLVFYGSQTGTAEEFANR	RAVLMGGSALSPAAVISHER	RVEAHAVAIGMLGASSSPLR
RFLFLYLGVVEPPDDLSSSSFDDDYDFFQR	RSAAEMYGSVTEHPSPSPLLSSSFDLDYDFQR	RVMSSSTQSYYSPAGDSALDEHPELSFDPFLR
HLSELSVAEDDDESLLGHLMLVGK	KMFLITNSPSSFVDKGMSYIVGK	KSIKITYVMFFNPGVLMGSSSDK
HLSQLSVAEDDDHLLHHLMLVGK	KMFLITNSPSSFVDKGMSYIVGK	KGVFSDNSFGIMSPLSKVYTIMK
LVGPEELGVTEAGFGADLGHEK	LTDQPPLVQAIFSGDPEEIR	LIPEGIVSQDQELPDFAPTR
QFTFGNLALLDDEDAQEDGVALPLPGVV	CPITTKLVLDEDEETKEPLVQVHR	CEDLVLTVPIEKKEHTDTQLPVER
MELGETRPDVLLQTFLDDTFPGDK	YGLQAGHNWFIISMQWWQQWK	YSMWQIWQWQGWQLAINFHGK
QVELRHPAVGTLQAHQEALK	DAAGIKVGAHAITAVPPPLQDK	DQTGVHPVAIPDLAIAAGPKAK
KRGFAFVTFDDHDDVDK	IGNWNEDVYLEEELMK	INELNGDEMVYLEEWK
KRRPEQYPPVPVVAVGSVSK	LRGISTKPVYIPEVELNHK	LVLTEGNVRIYKSHEPPIK
QLHCCHLASLQELVDPEAPQK	WQKKGGQPPGTAESKPDSQPQK	WGQPSQKQPSGKPEGAQTPKDK
YPGHNTTYFLPGEGNVPLLLEEPK	LGKDPNTYFIVGTAMVYPEEAEPK	LIEGVKFTAEMGDPPPNTAYEVYK
FLPNSSSSRHDRHDNNYFFFGPDLCGGGCK	LFPNSLDQTDMHGDSEYNIMFGPDICGPGTK	LFEQPCTNDDHSDGGGLDIYMGNTMIPPSFK
KFNGEDLDTLSPTLGFDLK	QMIKEAFAGDDVIRDFLK	QDFEGLRIAAVFDDIKMK
VEHELSEGDVATAAAAALASAATK	ATVLESEGTRESAINVAEGKK	AVLEAERVTSEKTGNGSIAEK
LPPPPVAFPDFSYCLLKR	RSSPAAFINPPIGTVTPALK	RPTPFVPTISLNGSIPAAAK

Continued on next page

DENOVO sequence	Correct sequence	Decoy (shuffled)
NLQQQYLQLDEQLTELHPLK	NNLTILQRYMSSKIPAVTYPK	NPASPYSIKLMLQITYTVNRK
GLEEEELLGGHLLELEECVHHNV	MEGAAWPGAGTGELLWDVHSHVVR	MALLTGGVDGAGEHVHPSVWAEWR
ELELASEEHVLFKEELELCR	EIEMASEERPPAQALEIMMGLK	ELPMLAREEPIAMSAIQEMGEK
RRRREEEEEEEETTPSETTKY	KKPFMLDEEGDTQTEETQPSETK	KTLETEFKPQTMQEESDDGTPEK
ERFVQDPGTPGDYVEVVR	GLDEEATPGTPGDPARPPASK	GGTEPGEAPLDTPPSADRPAK
LLGFEREEEEEELDEAEK	VVPGQFDDADSSDSENRDLK	VQDDPLDDDVANSGSRFESK
KGDTEGVDGTLTSNVADSPR	NLEQYNKLDQDLNEVK	NYEQKDLNEDNQLVLK
QRGSSNGNEGSLERREESTLK	KSLDKDPLLLSGTHVMEGSGR	KGTVLDLKPGGLDLSMSSHER
FTGSGEDLPFGFEDDLCCVCLK	FTGSFDDDPDPHRDPYGEEVDR	FSTGEDDVRPDYPGHFDPEDDR
SNLEEEEKLLEASGSNNPPPESSYSGGLC	KSPATPQAKPDGVTATAADEEEDEYSGGLC	KKQAESYDGEPPLDSTDATVAAEGTEGPAC
EGQRGGEAFVELESEDEVK	IPDPEAVKPDDWDEDAPAK	IPDDVAKAEAPEDWDPDPK
AYLEEPPPEPMETSLDSSEMAK	SRQAAGQTAMSPTESNKSSTTSK	SKSEQTRAMSPSATASSTTGQNK
ARYPPPEEVAHESAEPYAK	ITLQNIPSQTAPGFTAEMK	IQGPQETNAITFSALTPMK
SHEAELVEGENHTYCLR	SLYGGFVEWQMGEQADGK	SFGGWADLYGEMEQGVQK
KGEKREEAAAEEEEEAALGGDDEEK	RSASPDDDLGSSNWEAADLGNEERK	RDADSGDEWNESRANDEPGLASSLK
PPTNDTHPLCCPSTTETGKNTAK	KNSITEISDNEDDLLEYHRR	KIISDEEYTENLNLDHRDSR
VQLDVLEEEEPDPPHPDESYDDEELHDPR	VQIPVSRPDPEPVSDNEEDSYDEEIHDPR	VSDRESPSPQEVHYDVIDDPENDEEIPPR
PSTTSSSAALLSSSSSSSSPDEESESQK	DSGSDTASAIIPSTTPSVDSDDESVVK	DTDTGSIDESASDDVTSPSVPSIVSAK
EGRDKPVTDAENCHLAR	EAADRADGAAPGVASRNAVAG	EAGAAGRSNPADAAVAARDVG
DHLWVENDAYPGTDRTEGVK	AFFQVEDSLELSFQGKDDVK	ADLFEFGQSDFQSVLVKDEK
KTSSHEQEALNDLPELSNCENFQK	LSTQSNSNNIEPARTAGGSGLARAASK	LGGLAIRATSETAQNSSNAGRAPSNSK
KVDKLEELDEENEAALENGLK	AVKEEGQDPDEIGIELEATSKK	AEEVEELKGIEIADDTPGQKSK
ASEQQELQELEEQLEEEESARLK	VEVYADADEILQEEIKEYKGYGR	VEVDLQEAYGKIEEGKYDEYIAR
YKGTDEWGGHPESEETVTR	TSSLPNHSEPDHDTDAGLER	TLEHSEGDPSHLNDPATDSR
DDEESGVPGPPPEAPKGPGEEEEESEKGGK	MEEESGAPGVPSGNGAPGPKGEGERPAQNEK	MGNAERAGEGEGNGQEAEEPGVKPSPPSGPK
VVAPAEEEEVEPAPLPRSEEEEEEEEERR	VVVPATEEEAEVDEFPTDGEMSAQEEDRRK	VTTASEDPAMAEEDVFDREQVGEVREEEPK
LTEGDVELQLNDEEGQSEVPEKPPR	MGQILGKIMMSHQPQPQEEQSPQR	MIGQQEQEHGQIQSPLPPQSMMKR

References

1. Lihua Sun et al. “A streamlined platform for analyzing tera-scale DDA and DIA mass spectrometry data enables highly sensitive immunopeptidomics”. In: *Nature Communications* (2022). Published: 07 June 2022. URL: https://www.nature.com/articles/s41467-022-30867-7.

2. Kaiyuan Liu et al. “Accurate de novo peptide sequencing using fully convolutional neural networks.” In: *Nature Communications* (2023). Published: 02 December 2023. URL: https://doi.org/10.1038/s41467-023-43010-x.

3. A. Kertesz-Farkas. “Introduction to Computational Mass Spectrometry”. Higher School of Economics, 2022.

4. Joshua Elias and Steven Gygi. “Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry”. In: *Nature Methods* 4 (2007), pp. 207–214. Published: 2007. URL: https://doi.org/10.1038/nmeth1019.

5. Attila Kertesz-Farkas, Beáta Reiz, Michael P Myers, Sándor Pongor et al. "Database Searching in Mass Spectrometry Based Proteomics". In: *Current Bioinformatics* 7.2 (2012), pp. 221-230. Published: May 2012. DOI: 10.2174/157489312800604354.

6. Frank Acquaye, Attila Kertesz-Farkas, William Stafford Noble et al. "Efficient Indexing of Peptides for Database Search Using Tide". In: *bioRxiv* (2022). Published: October 2022. DOI: 10.1101/2022.09.30.510396. License: CC BY 4.0.

7. Rovshan G. Sadygov et al. "Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book". In: *Nature Methods* 1.3 (2004), pp. 195–202. Published: December 2004. DOI: 10.1038/nmeth725. PMID: 15789030.

8. Joshua E. Elias and Steven P. Gygi. “Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics”. In: *Methods in Molecular Biology*(2010), pp. 55–71. Published: 2010. DOI: 10.1007/978-1-60761-444-9_5. PMCID: PMC3434995 (NIH Public Access).