**Title: PTMKB – A Database for Post Translational Modifications Compiled by dbPTM**

**Abstract**

Current methods to detect Post Translational Modifications (PTMs) for a protein has made spectacular developments in recent years, including mass spectrometry and nuclear magnetic resonance spectroscopy. However, such techniques are riddled with either long experimentation times, high costs, or result inaccuracies. Therefore, more recent PTM detections have revolved around using artificial intelligence to predict where PTMs could occur, but such models have the caveat to deal with an exponentially large number of features without prior knowledge, including the protein sequences which cannot be fed to machine learning models normally without some encoding methods. We present PTMKB, a database housing PTMs acquired from dbPTM and provides log-odd frequencies of amino acids in protein sequences for each PTM.

**Table of Contents**

1. Introduction
2. Home Page
   * Search Functionality
   * Protein Information Display
   * Highlighted PTMs
   * Interactive Features
3. Documentation Page
   * Supported PTMs
   * REST API Support
4. Download Page
5. Conclusion
6. References

**1. Introduction**

Today, methods for detecting post-translational modifications (PTMs) have seen remarkable advancements in recent years, particularly with techniques like mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy. These methods offer high sensitivity and specificity, allowing researchers to identify and characterize a wide array of modifications. However, they come with significant drawbacks, including lengthy experimentation times, high costs, and potential inaccuracies in quantification and interpretation.

Mass spectrometry, while a powerful tool for PTM analysis, often requires extensive sample preparation and prior enrichment of modified proteins, complicating the workflow. Similarly, NMR spectroscopy, although capable of providing detailed structural information, is generally limited to smaller proteins due to sensitivity constraints. These challenges have prompted researchers to explore alternative approaches for PTM detection.

Recent developments in artificial intelligence (AI) have opened new avenues for predicting PTMs. Machine learning models can analyze vast datasets to identify potential sites of modification based on patterns in protein sequences. However, these models face significant challenges, particularly when dealing with the exponential growth of features inherent in protein sequences. Raw protein sequences cannot be directly fed into machine learning algorithms; they require effective encoding methods to transform them into a suitable format for analysis. This need for robust feature representation highlights the complexity of accurately predicting PTMs without extensive prior knowledge.

To address these challenges, we introduce PTMKB, a comprehensive database dedicated to housing information on PTMs sourced from the dbPTM database. PTMKB offers a unique resource by providing log-odds frequencies of amino acids in protein sequences associated with each PTM. This feature allows researchers to assess the likelihood of specific modifications occurring at particular sites based on empirical data, thereby enhancing predictive modeling efforts. By integrating this information, PTMKB aims to facilitate more accurate predictions of PTMs and support further investigations into their biological roles. We offer the usage of PTMKB through a website interface as well as support for REST API, allowing users to programmatically utilize PTMKB to its fullest extent.

**2. Website**

The website portion of PTMKB allows users to view proteins gathered by dbPTM and gain additional insights thanks a to a combination of information from dbPTM, UniProtKB, and our calculations.

**Home Page**

The home page is presented with a simple search bar, guiding users to search for proteins based on a protein’s UniProt Identifier or UniProt Accession Number. As the user types in their desired identifier, a list of suggestions pops up, allowing the user to select whatever protein they may want to analyze. If the queried identifier does not exist in the current database, the user is promptly notified.

Upon a successful identifier query, external information is fetched from UniProtKB using its REST API functionalities. The list of information gathered from UniProtKB includes the following:

* + **Protein Name**
  + **Gene**
  + **Species**
  + **Protein Sequence**
  + **Sequence Length**
  + **Function**

Besides the additional information, the protein sequence itself is displayed in a separate, scrollable window below. This sequence shows where in the sequence a PTM has been observed based on evidences. These individual amino acids are separately highlighted. Hovering over a highlighted amino acid shows the type of PTM that has occurred on the amino acid, along with what position the amino acid is in the sequence. When the user clicks on the highlighted amino acid, more information is shown exclusively about the type of PTM that has occurred. The following information is displayed:

* + **PTM**
  + **Position**
  + **Evidence(s)**
  + **Log-Odd Probability at Position**
  + **Additive Score**
  + **Multiplicative Score**
  + **\*-Multiplicative Score**

The first three are provided by dbPTM. The rest of the values were calculated using the sequences provided by dbPTM. The criteria for calculating the values is as follows: the window where PTM occurs contains an additional 10 upstream and 10 downstream amino acids, giving a total sequence length of 21. These sequences are stacked and transposed for calculation. The values are calculated for each PTM, by taking the count of each amino acid over the amount of proteins recorded for that PTM, along each position in the window. This calculates the frequency of an amino acid appearing at a position relative to the site of modification, as per the evidences recorded by dbPTM. Finally, as the values are extremely minute in magnitude, they are passed through a natural log operator to contain the magnitude to a scale that is easier to interpret. These tables that house log-off probabilities exist for each PTM separately.

From these log-odd probabilities, scores are calculated. For each PTM, its window is captured from the protein sequence (21 amino acids), and using the table respective to the PTM occurring on the amino acid, the values are picked based on each amino acid’s type and relative position, creating a vector of log-odd probabilities. These values are used to calculate the **Additive**, **Multiplicative**, and **\*-Multiplicative** scores. Here is how each score is calculated:

* + **Additive:** sum all of the log-odd values, ignoring infinity values.
  + **Multiplicative:** multiply only the window where the vector does not contain any infinity or 0 values.
  + **\*-Multiplicative:** Adjusted Multiplicative score, this takes the probability computed using Multiplicative score formula, make the probability positive, and take another natural log of the positive probability, and finally add it as a reciprocal over 1.

Along with the protein sequence is a list of PTMs recorded by dbPTM, all displayed as active checkboxes. Clicking on the checkboxes makes the respective highlighted amino acids appear/disappear. This allows the user to only look for certain PTMs in protein sequences while ignoring the rest.

**Documentation**

The documentation page of the website primarily deals with the type of PTMs collected from dbPTM, and how the user can programmatically fetch information from the database using REST API functionalities. The user is able to fetch the list of PTMs recorded by dbPTM. The user is also able to fetch the details of the PTM including the type of PTM, where in the protein sequence it occurs, and evidences, if any, for that PTM being observed in research. Finally, the user can also fetch the log-odd probability tables constructed for each PTM.

**Download Page**

The user has the option to select a PTM and download its log-odd probability table in multiple file formats, including JSON, CSV, PNG, PDF, or SVG.

**5. Conclusion**

PTMKB is an innovative database designed to enhance the understanding of PTMs through a user-friendly website and REST API. Through a combination of dbPTM and UniProtKB integrations, PTMKB provides detailed insights into protein sequences including the identification and statistical analysis of PTMs using log-odd probabilities. The platform facilitates effective visualization of PTM occurrences and supports researchers with tools to explore and download valuable data.

**6. References**

* [UniProt API Documentation](https://www.uniprot.org/help/programmatic_access)
* [dbPTM Database](https://awi.cuhk.edu.cn/dbPTM/)